FACULTAD DE INGENIERÍA Y CIENCIAS AGROPECUARIAS

DISEÑO E IMPLEMENTACIÓN DE ALGORITMO DE PREDICCIÓN EN BASE AL MUESTREO DE DATOS DE UNA RED DE SENSORES PARA LA EMPRESA ROSAPRIMA

Autor

Luis Fernando Díaz Quishpe

Año

2018

FACULTAD DE INGENIERÍA Y CIENCIAS AGROPECUARIAS

DISEÑO E IMPLEMENTACIÓN DE ALGORITMO DE PREDICCIÓN EN BASE AL MUESTREO DE DATOS DE UNA RED DE SENSORES PARA LA EMPRESA ROSAPRIMA

Trabajo de Titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en Redes y Telecomunicaciones

Profesor guía

MSc. Carlos Marcelo Molina Colcha

Autor

Luis Fernando Díaz Quishpe

Año

2018

DECLARACIÓN DEL PROFESOR GUÍA

“Declaro haber dirigido este trabajo a través de reuniones periódicas con el estudiante orientando sus conocimientos y competencias para un eficiente desarrollo del tema elegido y cumpliendo con todas las disposiciones vigentes que regulan los trabajos de titulación”

_____

Carlos Marcelo Molina Colcha
Magister en Tecnologías de la Información y Comunicación (TIC)
CI: 170962421-5

DECLARACIÓN DEL PROFESOR CORRECTOR

“Declaro haber corregido este trabajo a través de reuniones periódicas con el estudiante orientando sus conocimientos y competencias para un eficiente desarrollo del tema escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los trabajos de titulación”

_____

Jorge Wilson Granda Cantuña
Magister en Ingeniería Eléctrica
CI: 170859418 -7

DECLARACIÓN DE AUTORÍA DEL ESTUDIANTE

“Declaro que este trabajo es original, de mi autoría, que se han citado las fuentes correspondientes y que en su ejecución se respetaron las disposiciones legales que protegen los derechos de autor vigentes”

_____

Luis Fernando Díaz Quishpe
C.I.: 1720944949

AGRADECIMIENTOS

Agradezco a mi padre por todo el apoyo que me ha brindado, por su amor y confianza depositada en mí.

Agradezco a las personas que me han brindado con el tiempo y conocimiento para el desarrollo de este trabajo de titulación, una mención especial al Sr. Diego Vega y al Sr Ruben Vasquez por su ayuda con las validaciones en el manejo de base de datos de la empresa ROSAPRIMA CIA. LTDA.

Finalmente, un agradecimiento muy especial Ing. Carlos Molina por la orientación que me proporciono a lo largo de la elaboración del proyecto, por su determinación y con fianza en mi proyecto

Luis Fernando Díaz

RESUMEN

Hoy las técnicas de Business Intelligence, son una gran herramienta dentro de las empresas, ya que están han permitido que todos los recursos del área de TI sean aprovechados al máximo de su capacidad. Sobre todo, la información que la empresa genera a cada segundo de manera interna y externa a tomado un rol muy importante al momento de tomar decisiones.

La minería de datos, una herramienta muy útil dentro de los BI ha permitido, que la información cumpla un rol muy importante dentro de la empresa. Esta toma los datos almacenados en grandes bodegas y mediante el uso de algoritmos matemáticos analiza el comportamiento de datos, buscando padrones repetitivos los cuales aportan para la toma de decisiones, o también nos brinda la posibilidad de mostrar datos estadísticos en ventas, producción e incluso en la mejora de áreas de la empresa

Dentro de este proyecto, se demostrará como los datos obtenidos de diferentes bases, pueden ser representados en un algoritmo, de tal manera que una vez aplicado nos brinde una predicción

ABSTRACT

Today Business Intelligence techniques are a great tool within enterprises because they have allowed them to take advantage of all their technological resources at the maximum of its capacities. Above all, the internal and external information that a business generates every second have an important role in the decision making of the company.

Data mining, a useful tool within BI, has allowed information to become the most important asset of the companies. It takes data saved in storages and using arithmetic algorithms analyzes the behavior of data, searching for repetitive patterns that will make easier the decision making process. It also brings the possibility of showing statistics of sales, productivity and even better areas of improvement within company.

In this project, it well demonstrates how data obtained from different databases can be represented in one algorithm which once it’s applied could give a prediction as the output.

ÍNDICE

ÍNDICE DE FIGURA

INTRODUCCIÓN

Diariamente las empresas generan diferente tipo de información, muchas veces esta es interna de conocimiento propio debido a la actividad diaria que se realiza dentro del negocio. Además, se genera información externa que no conocemos, pero está allí presente, cuyo objetivo es dar un uso significativo como: investigación, examinación, y planificación ayudando así a la empresa a la toma de decisiones

La evolución de los diferentes componentes del área de TI, especialmente los de equipos de cómputo y transmisión, han permitido que la información se administre de mejor manera, es decir mejorando la manera de transmisión y la forma en la cual esta va a ser almacenada. Esto se dio gracias a los siguientes factores:

1. Evolución de los sistemas de almacenamiento, y de bajo costo
2. Tecnologías de transmisión a gran velocidad
3. Mejoras de hardware y software dentro de los centros de computo
4. Evolución de los sistemas gestores de bases de datos

A pesar de las diferentes mejoras, las empresas no suelen sacar el mayor provecho a la información, la mayoría solo se dedica a almacenarla sin ningún fin. Pero si esta fuese analizada y procesada puede brindar un sin número de soluciones las cuales ayudarían, a que la empresa pueda tomar decisiones muy importantes.

La minería de Datos una herramienta tecnológica que usa los datos como materia prima principal, para poder generar conocimiento. Es decir, esta cuando el usuario toma los datos y les da un significado en especial, este se convierte en información

ALCANCE

El alcance de este trabajo de titulación es diseñar e implementar un algoritmo de predicción, utilizando herramientas que parten de datos de diferentes fuentes y con el fin de almacenar, analizar información, y transmitir (estructurados, no estructurados y semiestructurados) de los sensores meteorológicos ubicados en una de las fincas, y los datos de producción para procesar estos dentro de un servidor y así elaborar un análisis estadístico el cual nos brinde una proyección de producción de Rosas para la venta de Rosas a nivel nacional o Internacional.

Además, el sistema permitirá:

- Compartición de información estadística con el usuario
- Generar planificación de producción
- Proyecciones climatológicas
- Compartición de información en tiempo real

JUSTIFICACIÓN

El clima es uno de los factores más importantes para la producción de rosas y en Ecuador este es muy variable, cuando es frio o lluvioso puede afectar a los sembríos. Hoy en día existen muchas formas de obtener datos meteorológicos los cuales nos permiten predecir el clima, pero este no ayuda a predecir la producción. La minería de datos es una herramienta, ya que al trabajar con una serie que grandes empresas usan para la recolección de datos y así generar una estadística de producción, para generar ventas más seguras, además brinda la opción de que esta información pueda ser compartida con el usuario para la verificación de cuando el mercado puede ser más viable. Además de que es información en tiempo real beneficia tanto al productor, al vendedor y al comprador

OBJETIVO GENERAL.

Diseñar e implementar un sistema de minería de datos, el cual permita abstraer datos meteorológicos y de producción, permitiendo buscar padrones repetitivos para obtener conocimiento de cuantos tallos podrán salir a la venta, tomando en cuenta los factores principales como el clima y la producción.

OBJETIVOS ESPECÍFICOS.

Diseñar una topología de red para empresa la cual permitirá evidenciar, que los datos de meteorológicos y de producción serán ubicados en un servidor principal el cual se encontrara en la nube, en plataformas de Amazon.

Analizar el uso de los diferentes protocolos de comunicación, para la transmisión de datos y su respectivo almacenamiento.

Diseñar un algoritmo el cual permita calcular un valor estadístico de producción y ventas.

Implementar un servidor en la nube el cual permita al usuario visualizar información en tiempo real de producción y ventas.

1. CAPITULO I. MARCO TEÓRICO

Introducción

Dentro de este capítulo se procederá con la revisión de los diferentes conceptos que componen Bussines Intelligence, ya que la minería de datos es un parte muy importante dentro de estos sistemas. También se revisarán los conceptos más importantes que componen el Data Mining (DM), ya que estos son más complejos al momento de la generación de un algoritmo

1.1. Concepto Bussines Intelligence (BI)

Se define como el conjunto o combinación de tecnologías, herramientas, o procesos los cuales permiten a la empresa transforma los datos almacenados en información, y esta su vez convertirla en conocimiento. Una vez que se obtiene el conocimiento este ayudara a la toma de decisiones, con el cumplimiento de objetivo y optimización de recursos.[1]

En la FIG1 se puede visualizar como la combinación de técnicas y herramientas o procesos van convirtiendo los datos en oportunidades a la empresa

Figura 1. Definición de IB.
Adaptado de: **(Uceda, 2013)**

[1] (Oracle, ¿Qué es Business Intelligence?, 2012, pág. 1)

1.1.1. Características de BI

Para que un sistema de BI cumpla con su funcionalidad correctamente de cumplir con 5 características fundamentales. En la Fig2, se puede observar las características que debe cumplir los sistemas de BI.

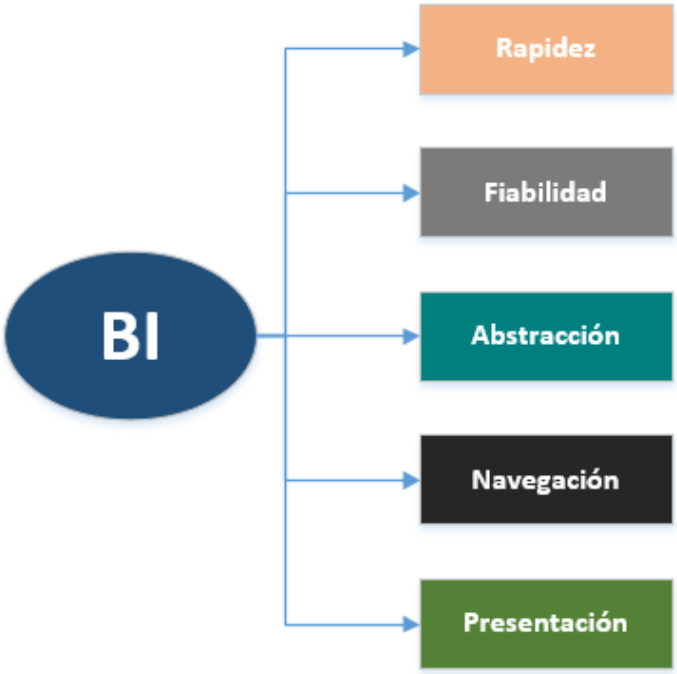


Figura 2. Características BI
Adaptado de: (Oracle, 2012)

- **Rapidez:** Modelo de datos o conocida también como la capa lógica, que brinda la flexibilidad para facilitar las respuestas necesarias
- **Fiabilidad:** información de calidad, integra, y estandarizada
- **Abstracción:** brindar respuestas a todo tipo de problemas sin importar su complejidad
- **Navegación:** transformación de la información simple a conocimiento complejo
- **Presentación:** Interpretación de la información.

1.1.2. Importancia de implementar BI

- **Carencia de Información:** la empresa por lo general solo almacena datos, pero no la convierte en información, por ello es recomendable guardar todo tipo de datos internos o externos, en aplicaciones o gestores de bases de datos. Esto permitirá obtener una gran ventaja frente a la competencia ya que, al profundizar en nuestros datos, permitirá encontrar diferentes patrones repetitivos o comportamientos, y así encontrar respuestas a nuestros problemas.
- **Fragmentación:** las empresas suelen trabajar con aplicaciones diferentes para cada uno de los departamentos, ya que carecen de una visión global es decir no tienen las herramientas de BI para integrar fuentes de datos de origen diferente o de diferente tipo. Esto se convierte en una gran limitación para que puedan tomar decisiones importantes, ya que no posee los recursos necesarios a la mano.
- **Manipulación Manual:** es un proceso lento y costoso tanto en lo económico como en hora hombre, poca confianza en la elaboración de informes ya que esto puede ser sujetos a errores.

1.1.3. Ventajas de utilizar BI

- Detección patrones, tendencias, oportunidades, y riesgos las cuales podrían ser utilizadas como ventajas
- Estandarización de procesos, lo cual permite a la empresa eliminar un sin número de procesos manuales, para la generación de la información.
- Simplifica el acceso a la información permitiendo así ahorrar el tiempo en la generación de informes o reportes.
- Ofrece el manejo de KPIs (Key performance indicator), para obtener un control de rendimiento.

- La toma de decisiones se encuentra fundamentada con información fiable y precisa
- En Ocasiones el uso de los BI permite detectar fraudes y padrones delictivos.

1.1.4. Arquitectura BI

La Fig3 muestra, como se encuentra conformada los elementos del BI, para después ser analizados.

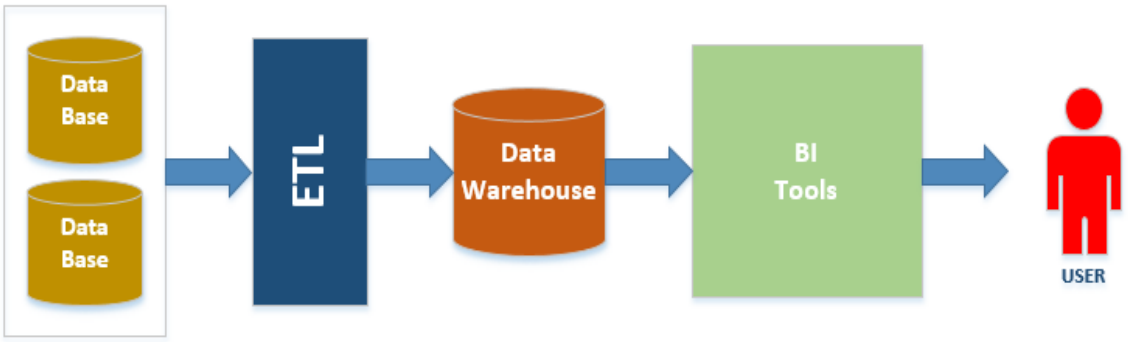


Figura 3. Arquitectura BI
Adaptado de: (Conesa & Curto, 2011)

1.1.5. Multiple Data Sources

Las Múltiples Fuentes de Datos, se conoce como el conjunto de sistemas operacionales, sistemas departamentales o sistemas de fuentes externas, donde se encuentran almacenados los datos dependiendo su origen.

La FIG4 nos describe cuales son los diferentes sistemas que funcionan para dar origen a los datos

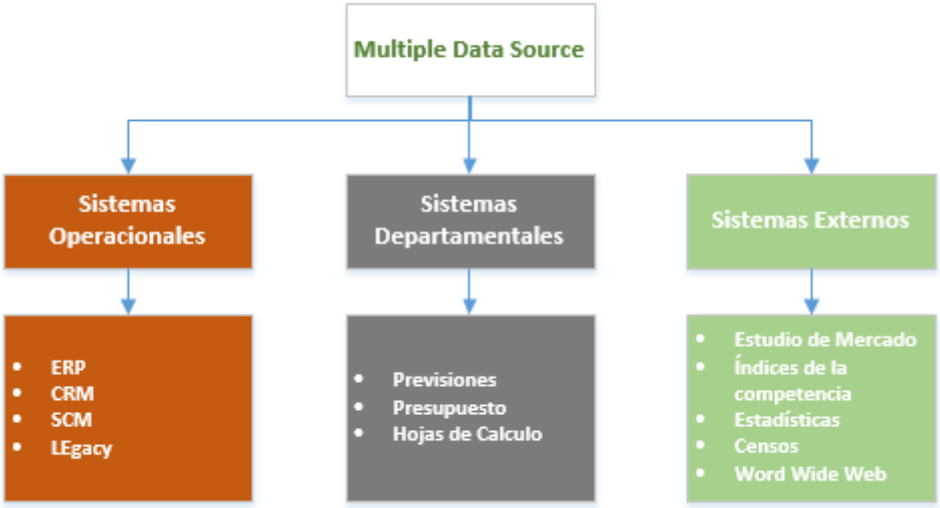


*Figura 4***.** Múltiples fuentes de Datos
Adaptado de: (SENA, s.f.)

Además, estos datos almacenados deben cumplir con ciertas características para ser analizados o procesados y así garantizar la calidad dentro de ellos como, por ejemplo:

- Precisión
- Integridad
- Coherencia
- Totalidad
- Validez
- Disponibilidad
- Accesibilidad

1.1.6. ETL (Extract, Transform, Load)

Se define como el proceso de extracción, transformación y carga de los datos

- **Extract:** proceso en el cual los datos son extraídos desde el sistema de origen, para analizarlos y así verificar que estos cumplan con la

estructura con la cual se los desea interpretar, en caso de no cumplir estos son rechazados. Además, se encarga de convertir los datos en un formato para la transformación. Se debe tener en cuenta que en este proceso los datos deben estar totalmente organizados, ya que estos se fusionan con los datos provenientes de otra fuente.

- **Transform:** dentro de este proceso se deben aplicar normas y reglas que aplica la empresa es decir estos pueden ser declarados, excepciones, o restricciones al momento de ser cargados. Hay que asegurar que estos sean declarativos, independientes, claros, intangibles. Y sobre todo que cumplan una finalidad útil para el negocio.
  Se debe tomar en cuenta que solo se debe utilizar columnas necesarias omitiendo especialmente las que tengan valores nulos.
- **Load:** en este proceso los datos son cargados en el sistema de destino es decir serán almacenados dentro de un Data warehouse, tomando en cuenta los requerimientos de la empresa. En algunos casos se toma en cuenta que los datos pueden ser sobrescritos cuando sea necesario especialmente cuando estos son antiguos, y en otros casos como con datos nuevos solo se resumen las transacciones y se almacenan en un espacio considerado.
  Además, se debe tomar en cuenta el tipo de carga que se desea realizar existen dos tipos:
  - **Acumulación Simple:** resume todas las transacciones que se ejecutaron en un tiempo seleccionado y transportar solo el resultado como una única transacción hacia el Data WareHouse.
  - **Rolling:** almacena información resumida a diferentes niveles ordenada de una manera jerárquica, en orden cronológico o incluso en la magnitud almacenada.

No importa cual proceso se utilice, se debe en este proceso se debe tener mucho en cuenta que esta interactúa con el destino, y es por ello por lo que se debe tener en cuenta las restricciones como, por ejemplo:

- Valores únicos
- Campos Obligatorios
- Integridad Referencial
- Rango de Valores

1.1.7. Data Warehouse

Gran repositorio de información, en el cual se puede acceder a manipular grandes volúmenes de datos, los cuales son originados de diferentes fuentes y diferente naturaleza[2]

Como se puede observar en la FIG5, se puede ver que data DW se compone de varios sistemas de bases datos unidos en uno solo.

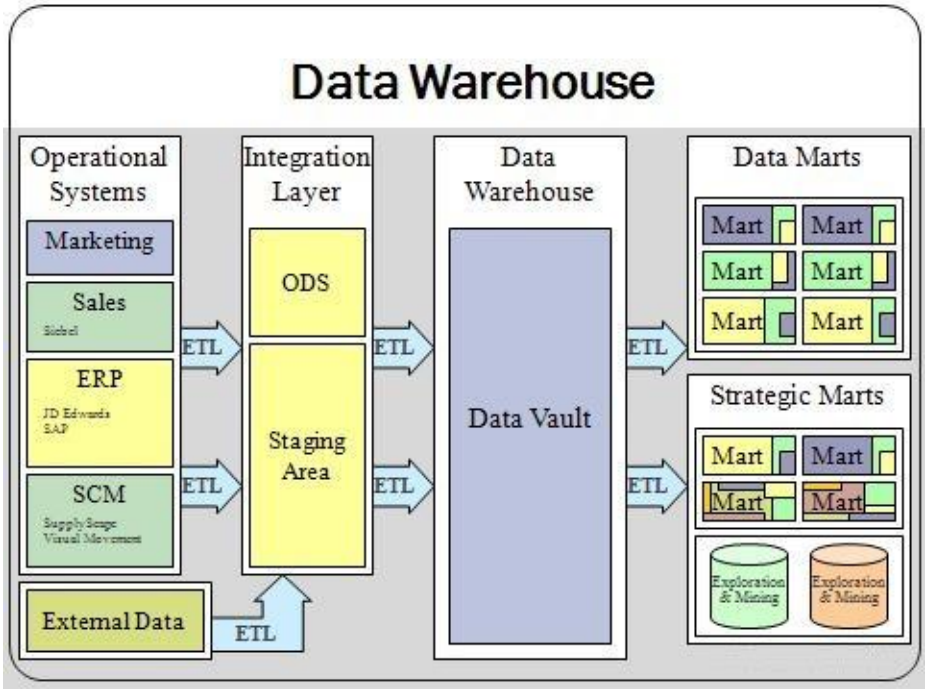


Figura 5. Estructura Data Warehouse
Adaptado de: (Microsoft, 2015)

[2] (Rosado & Rico, 2010, págs. 321-326)

1.2. Minería de datos

1.2.1. KDD

Knowledge Discovery in Databases o el Descubrimiento de Conocimiento en bases de datos, es un término que empezó a utilizar se al principio de los años 90, con el fin de realizar búsquedas amplias de alto nivel con métodos específicos referentes a la minería de datos[3]. Se debe tener en cuenta que el descubrimiento es una inducción de conocimiento, el cual no se encuentra supervisado que utiliza dos procesos

- La búsqueda de regularidades en los datos de partida
- Formulación de leyes que la describan

Para entrar a trabajar a profundidad con minería de datos se debe tomar en cuenta el verdadero significado de descubrimiento:

- **Descubrimiento:** se define como la observación y explicación mediante el uso de recolección de datos, formulación de hipótesis, con el fin de diseñar experimentos para explicar los hallazgos dentro de estas y realizar comparaciones junto a las de otros investigadores.
- **Descubrimiento del conocimiento:** se denomina como la extracción de información no trivial, la cual está implícita dentro de un almacén de datos, pero esta es previamente desconocida y de suma utilidad a la empresa

Sin embargo, el concepto KDD, sigue siendo muy ambiguo por lo cual se debe analizar las principales ideas o componentes del descubrimiento de conocimiento dentro de cualquier sistema:

---

[3] (Rosado & Rico, 2010, págs. 321-326)

- **Lenguaje de alto nivel:** cuando se descubre el conocimiento, se encuentra en un leguaje de alto nivel, este es intangible desde un punto de vista humano. Pero dentro del KDD, se realizan representaciones a bajo nivel generadas como las conocidas redes neuronales, generalmente utilizados en la minería de datos.
- **Precisión:** el descubrimiento es la representación del contenido de la base de datos, es un reflejo de la realidad, es decir no puede ser perfecta y llena de contenido no deseado. es decir, no todos los datos serán útiles para adquirir conocimiento y el grado de certeza medirá el crédito que el sistema o el usuario le pueda dar a este; tomando en cuenta que, si no cumple con la certeza requerida en los patrones descubiertos, no se podrá conocer como conocimiento.
- **Interés:** numerosos padrones pueden ser extraídos de una base de datos, pero el usuario solo considera como conocimiento aquellos que se vean interesantes tomando en cuenta los criterios de él, pero el padrón más importante o de interés debe ser nuevo, útil y no debe ser trivial
- **Eficiencia:** para considerar que el proceso de descubrimiento sea eficientemente debe estar en una computadora. Para considerar que un algoritmo es eficiente se debe tomar en cuenta diferentes factores como el tiempo de ejecución y el espacio de memoria sean los adecuados ya que estos crecen de manera polinomial al momento que ingresan los datos de entrada.

El KDD nace o tiene sus inicios en el aprendizaje automático o la estadística, pero hay componentes que lo hace diferente como por ejemplo su objetivo es el de descubrir conocimiento útil, relevante, valido y nuevo de un fenómeno o actividad mediante el uso de algoritmos ya que por el uso de creciente de los datos.

Otro aspecto para considerar para dentro del descubrimiento de conocimiento es la interacción entre la máquina y el ser humano, esta debe ser dinámica y

flexible ya que esto influye al momento de presentar los resultados de manera visual o de una manera muy clara, ya que el resultado debe ser interesante y que su calidad no se encuentre afectada por las grandes cantidades de información o por el ruido presente en los datos.

En la FIG6 se puede evidenciar las diferentes tareas que el KDD, nos ayuda al momento de aplicarlo en los sistemas de minería de datos

Figura 6. Descripción del KDD
Adaptado de: (Microsoft, 2015)

1.2.2. Proceso KDD

El KDD cumple una serie de pasos dentro del proceso interactivo e iterativo, los cuales son:

- **Conocimiento Relevante:** es paso de considera como el desarrollo y entendimiento del aplicativo, buscando el objetivo del usuario final. Para ello se debe tomar en cuenta factores como los cuellos de botella, conocer las partes susceptibles del procesamiento automático. Además, se debe tomar en cuenta cuales son los objetivos criterios de rendimiento que el proceso exige, para obtener simplicidad y precisión en el resultado del conocimiento que se pudo extraer.

- **Datos Objetivo:** paso en el cual se crea un conjunto de datos, en el cual se selecciona un subconjunto ya sea de variables o ejemplo de lo que se desea descubrir. Para ello se considera la homogeneidad, variación del tiempo de los datos estrategias para muestras y grado de dependencia.

- **Procesado:** consiste en eliminar el ruido, es decir usa estrategias para manejar valores nulos y procede a normalizar los datos.

- **Transformación y Reducción:** consiste en la acción de buscar en los datos las características de utilidad dependiendo su objetivo final, además realiza una reducción de variables y proyecta en los datos un espacio de búsqueda más fácil al momento de encontrar una solución. Este paso marca la diferencia dentro del proceso ya que requiere un conocimiento del problema y una buena intuición y así garantiza el éxito o fracaso de la minería de datos

- **Selección del sistema Minería de Datos:** se escoge el sistema que se va a utilizar ya que se debe utilizar para la clasificación, regresión, agrupamiento de conceptos, y las detecciones de desviaciones

- **Elección algoritmos de minería de datos:** se escoge el algoritmo acorde a la necesidad

- **Minería de Datos:** es la búsqueda del conocimiento y la representación del mismo, el éxito de esta depende de la ejecución de los pasos anterior por parte del usuario.

- **Interpretación del Conocimiento:** los resultados obtenidos dependen mucho de otros factores tales como definición de medidas de interés de conocimiento, filtración automática, técnicas de visualización, para facilitar la valoración de resultados o la búsqueda de los mismos.

- **Conocimiento descubierto:** paso en el cual se incorpora al sistema documentado solo la parte de interés, dentro de este también incluye la revisión y resolución a las inconsistencias.

A continuación, FIG7 se puede observar el diagrama de como el KDD va cumpliendo su proceso

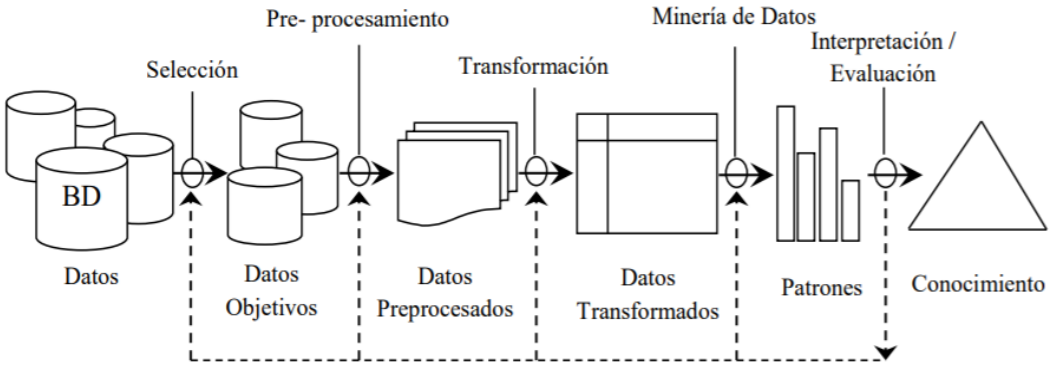


Figura 7. Proceso del KDD
Adaptado de: (Beltran, s.f.)

1.2.3. Fases KDD

El proceso KDD, básicamente se encarga de la extracción no trivial del conocimiento, tomando en cuenta que este se encuentra escondido y es sumamente utilitario, partiendo de un conjunto grande de datos. En el proceso de KDD es compone de 6 estados:

- **Recolección de Datos**

  En esta primera fase de KDD, tomando en cuenta que son sucesivas son capaces de extraer la validez y la utilidad del conocimiento tomando de origen la información original. Para ello se debe tomar en cuenta que la información:

  - Este en bases de datos
  - Que sean internas o externas
  - Las fuentes deben ser transaccionales

Posteriormente se realiza el análisis, el cual ser mucho más fácil si esta es unificada accesible y fuera del trabajo transaccional, en esta fase se toma en cuenta para la minería de datos los siguientes aspectos:

Fuente

- OLAP u OLTP
- ROLAP o MOLAP
- Data warehouse

Usuario

- **Picapedreros:** su función es la generación de informes periódicos, evolución de parámetros, control de valores.
- **Exploradores:** encuentran nuevos patrones o valores significativos utilizando técnicas de minería de datos

Cuando se realiza la recolección externa de información, puede ser extraída de diferentes fuentes como

- Documentos Gráficos impresos o digitales
- Datos compartidos con otros tipos de industrias
- Datos de marketing y mercadeo de la competencia
- Bases de Datos externas compradas

- **Selección, limpieza, y transformación de datos**

  Cuando se trabaja con un sin número de datos posiblemente existan algunos que se encuentren erróneas o inconsistentes y deben ser eliminados a esto se le denomina como limpieza, mientras que los datos irrelevantes son seleccionados con métodos estadísticos.

  Para determinarlos datos que van a ser seleccionados se debe tomar en cinta las siguientes acciones:

  - Outliers
    - Al momento que se implementa un algoritmo se debe tomar en cuenta que son muy robustos a datos anómalos y deben ser ignorados
    - Filtración (eliminar o reemplazar) la columna; se debe tomar que esta es una acción realmente

compleja, ya que una columna puede depender de otra. Es aconsejable reemplazarla por otra más discreta.

- Filtrar fila: al momento de sesgar los datos, suelen aparecer datos erróneos ya que estos suelen estar relacionados con casos especiales.
- Reemplazar valores: dependiendo el algoritmo el valor nulo puede o no puede ser tomado en cuenta. Así que este procedimiento depende mucho de los valores que del outlier,
- Discretizar: se usa valores discretos en vez de valores continuos

- Missing Values
  - Ignorar: usar algoritmos robustos a datos faltantes
  - Filtrar la columna: al existir columna dependiente con datos irrelevantes por lo cual no es recomendable eliminar, es mejor reemplazar por una columna booleana.
  - Filtrar la fila: cuando los datos son sesgados, suelen aparecer datos faltantes relacionados con casos especiales
  - Reemplazar un valor por medidas: predicción a partir de otros datos, utilizando técnicas de minería de datos.
  - Segmentar: se clasifican de una manera ordenada la lista de los datos, que tiene disponibilidad, y se asigna un modelo diferente segmentos para luego recombinarlas.

- **Minería de Datos.**

  Al momento de maneja una gran cantidad de datos, se debe tomar en cuenta mucho las técnicas de como aprendizaje automático y estadísticas so son de aplicación directa por las siguientes razones:

- Los datos se encuentran almacenados dentro de un disco y estos no pueden ser escaneados
- El muestro no siempre es compatible con los algoritmos
- La dimensionalidad es alta, es decir muchos datos
- Imperfección en los datos

Dentro de la minería de datos se debe tener en cuenta cuales son los padrones, que se desean descubrir o interpretar, así que se debe tomar en cuenta lo siguientes recomendaciones

- El explorador deduce que tipo de patrón quiere descubrir, una vez que los datos ya fueron recolectados, tomando en cuenta que estos deben ser de gran interés
- La técnica de minería de datos que se desea utilizar será definida por el tipo de conocimiento a extraer

En los sistemas de minería de datos son los encargados de elegir el algoritmo más adecuado para determinar un tipo de padrón deseado.

- **Evaluación y Validación**

  Dentro de la minería de datos se generan una o ms hipótesis del modelamiento de los datos, así que para validar el uso de un modelo se debe tomar en cuenta la evolución de hipótesis dentro de dos fases

  - Primera Fase: la precisión de un modelo de ser representado dentro de un banco de ejemplos independientes, para su comprobación y para ser utilizado.
  - Segunda Fase: se recomienda organizar un plan piloto con el modelo seleccionado, para obtener más fiabilidad en modelo utilizado.

- **Interpretación y difusión**

La interpretación del modelo en algunas ocasiones es trivial pero generalmente requiere un proceso de implantación:

- El modelo requiere de una implementación
- El modelo debe ser descriptivo y representativo

- Al ser utilizado por muchos usuarios este modelo requiere difusión y debe ser expresado de una manera comprensible para el uso de toda la organización.

- **Actualización y Monitorización**

Se consideran como los procesos de manteamiento del modelo

- **Actualización:** un modelo que sea válido puede sufrir cambios de contexto
- **Monitorización:** cada cierto tiempo los datos son nuevos, así que consiste en realizar revalidaciones, con el objetivo de detectar actualizaciones dentro del modelo

1.2.4. Minería de Datos

- **Concepto**

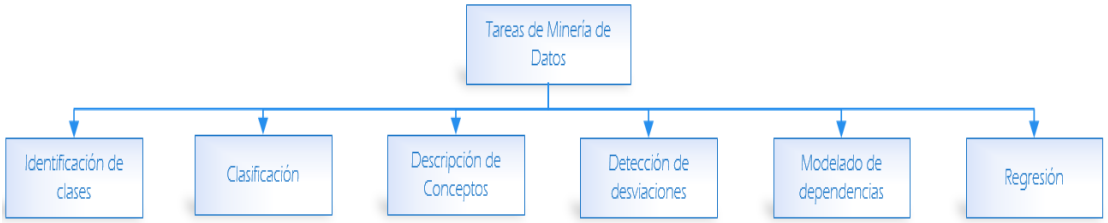
  La minería de datos es una técnica que posibilita la explotación de los datos extraídos de un banco de información el cual no puede ser detectado a simple vista. La minería de datos combina una serie de técnicas semiautomáticas de inteligencia artificial, bases de datos, estadística y visualización gráfica. Se encarga de descubrir tendencias, relaciones comportamientos atípicos, relaciones, padrones y trayectorias ocultas, con el de crear un proceso de toma de decisiones usando el conocimiento[4].

  Su nombre hace analogía a una montaña, ya que dentro de ella oculto entre piedras y tierra se encuentran diamantes de gran valor, que mediante el uso de técnicas de minería son extraídos para aprovecharlos de mejor manera.

- **Tareas de Minería de Datos**

  Como se puede evidenciar en la Fig8, la minería de datos pone a su disposición un sin número de tareas

[4] (Beltran, s.f.)



FIigura 8. Tareas de Minería de Datos
Adaptado de: (Beltran, 2014)

- **Identificación de Clase:** se identifica el conjunto de categorías o conjuntos para poder describir a los datos. Se debe tomar en cuenta que estos pueden ser exclusivo y exhaustivos mutuamente, o también pueden constituirse en una representación jerárquica o permitir solapamientos.
- **Clasificación:** es la tarea con la habilidad de adquirir funciones que clasifica los elementos de un dato a varias clases predefinidas.
- **Descripción de conceptos:** en esta tarea se busca encontrar un método que permita describir compactamente un subconjunto de datos, se toma mucho en cuenta que hay métodos muy sofisticados que involucran reglas de compactación, técnicas para visualización multivariada y descubrimiento de relación funcionales entre las variables. Generalmente este tipo de técnicas suelen ser usadas en el análisis en forma interactiva y en generación de reportes
- **Detección de Desviaciones:** tarea en la cual se detecta los cambios de mayor importancia en los datos, tomando en cuenta valores actuales y pasados, además sirve como un filtro de grandes volúmenes de datos que son menos interesantes, una de las dificultades que presenta es el determinar la desviación para dar significado a los datos de interés
- **Modelo de dependencias:** aquí se define un modelo para describir las tendencias significativas entre variables. Existen dos tipos, el estructural que suele ser de manera gráfica y las variables son

dependientes una de la otra, y a nivel cuantitativo el cual describe la robustez de dependencias y utiliza escalas numéricas.

- **Regresión:** adquiere una función la cual se encarga de clasificar los elementos de un dato a una variable de predicción la cual obtiene un valor real.

1.2.5. Proceso de Minería de Datos

Para dar inicio a la minería de datos, el primer paso en la identificación de los datos, así que para ello se debe imagina cuales son los datos requeridos, además verificar su ubicación y como conseguirlos. Obtenidos los datos se preparan, ubicándolos en una base de datos en un formato adecuado o a su vez construir un data warehouse. Este es el paso más complicado de la minería de datos. Con los datos en el formato adecuado se seleccionan los más esenciales y se elimina lo innecesario.

Para proceder con el análisis de los datos dentro de la minería de datos, se recomienda en tener idea cual es el interés que se desea averiguar, las herramientas requeridas y como proceder. Tras el uso de las herramientas la cual fue seleccionada o creada por nosotros se debe saber cómo los resultados o padrones van a ser interpretados, pasa así obtener facilidad de extraer solo los necesarios. Una vez realizada la inspección de resultados hay que identificar cuáles será s las acciones a tomar, discutir y pensar en los procedimientos que se llevaran a cabo.

Ya implementadas las herramientas se evalúa observando los resultados, beneficios y los costos para reevaluar el procedimiento completo, mientras se realiza este procedimiento los datos pueden haber cambiado, así que se debe tener en cuenta la disponibilidad de nuevas herramientas y permitirá planificar el ciclo de la minería que lo siga.

Básicamente la minería de datos no es más que un proceso en el cual involucra ajustar modelos normalmente de tipos estadísticos, tomando en cuenta que puede existir ruido o errores dentro del mismo. Además, también se puede establecer padrones generalmente son de tipo probabilístico o determinístico.

Figura 9. Proceso de Minería de Datos

Adaptado de: (Beltran, 2014)

- **Procesamiento de los Datos**

  El formato almacenado de los datos en la base o en el data warehouse, nunca es el idóneo, incluso no pueden ser utilizados en los algoritmos. Mediante el procesado se filtran los datos incorrectos, no válidos y desconocidos, y según la necesidad, para así seleccionar el algoritmo que se acople a las necesidades donde se obtendrán las muestras o los valores posibles.

- **Selección de Características**

  En este paso, reduce los datos, tomando en cuenta las variables más influyentes dentro del problema, tonando en cuenta que esta no debe modifica ni perjudicar la calidad del modelo del conocimiento adquirido en el proceso de minería de datos.
  Los métodos para utilizar deben cumplir con las siguientes dos características:

- Basados en la elección de mejores atributos para la solución del problema
- Búsqueda de variables independientes con el uso de test de sensibilidad, algoritmos heurísticos o de distancia

- **Extracción de Conocimiento o Algoritmos de Aprendizaje**

  El modelo del conocimiento se obtiene mediante el uso de técnicas de minería de datos, los cuales son representados por padrones de comportamiento de los valores en las variables o por relaciones de asociación., existen casos en los cuales se pueden usar varias técnicas sin embargo esto con lleva a que los datos tengan un procesado diferente.

- **Evaluación y Validación**

  Ya obtenido el modelo, se procede a realizar su respetiva validación, comparando los resultados con los objetivos planteados, para poder arrojar nuestras propias conclusiones y determinar si fueron satisfactorias. Cuando se obtienen diferentes modelos es por el uso de diferentes técnicas así que se debe realizar una validación uno por uno para ver cuál cumple con la solución al problema. Si el modelo no alcanza con lo esperado se debe repetir uno a uno los pasos anteriores.

1.3. Técnicas y Métodos de Minería de datos.

La minería de datos solo se encarga de crear modelos os cuales son predictivos y descriptivos según el problema propuesto. Pero se debe tomar en cuenta que un modelo predictivo responde a datos futuros, mientras que un modelo descriptivo menciona las características entre los datos y sus características.[5]

[5] (Rosado & Rico, 2010)

Se caracterizan por predecir un valor de un atributo, denominado como etiqueta, de un conjunto de datos. Una vez que aparece dicha etiqueta se busca una la relación entre la etiqueta y una serie de atributos con el fin de predecir datos cuya etiqueta es desconocida. A esta también se la conoce como aprendizaje supervisado él trabaja en dos fases:

- Entrenamiento: construya modelos de un subconjunto de datos con etiquetas conocidas
- Pruebas: realiza pruebas del modelo con el resto de los datos.

Cuando el aplicativo no es lo suficientemente robusto o con el potencial requerido para una solución predictiva, se toca recurre a los métodos denominados como de Descubrimiento o no supervisados, los cuales se encargan de encontrar tendencias o patrones en datos actuales es decir no usan datos históricos. Su objetivo final es buscar un beneficio para el negocio a partir de estos.

1.3.1. Taxonomía de las Técnicas de minería de datos

Clasificación de las técnicas de aprendizaje

- **Interpolación:** es cuando una función es continua sobre varias dimensiones.
- **Predicción secuencial:** cuando las observaciones se encuentran ordenadas de manera secuencial el aplicativo predice el siguiente valor consecutivo.
- **Aprendizaje supervisado:** las observaciones poseen un valor a la clase que pertenecen, y aprende este clasificador
- **Aprendizaje no supervisado:** no tiene clases asociadas, en si su objetivo es encontrar las regularidades en los datos de cualquier tipo.
- **Abducción o aprendizaje analítico:** el objetivo es explicar la evidencia con respecto a una constante.

En la figura 10, se visualiza un mapa conceptual el cual nos explica de mejor manera como se compone la taxonomía de las diferentes técnicas de minería de datos

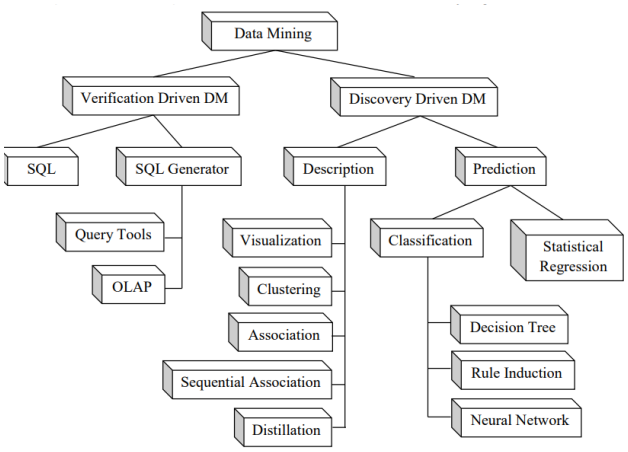


Figura 10. Taxonomía de Técnicas de Minería de Datos

Adaptado de: (Calderon, 2013)

1.3.2. Clasificación algoritmos predictivos y descriptivos.

Los algoritmos son herramientas de minería de datos, a continuación, en la Fig11 visualizaremos su clasificación

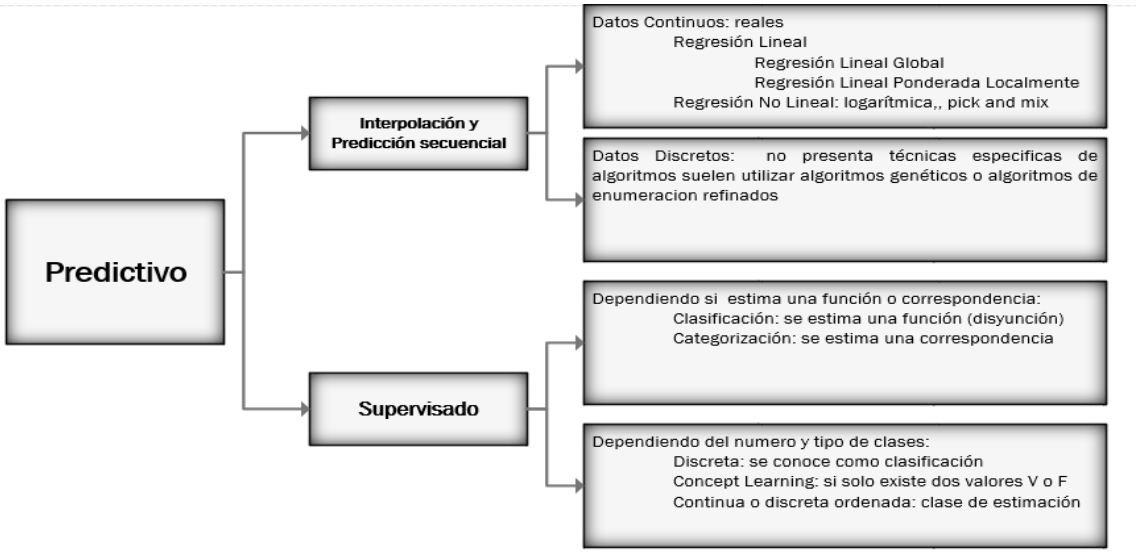


Figura 11. Clasificación de Algoritmos Predictivos

Adaptado de: (Beltran, 2014)

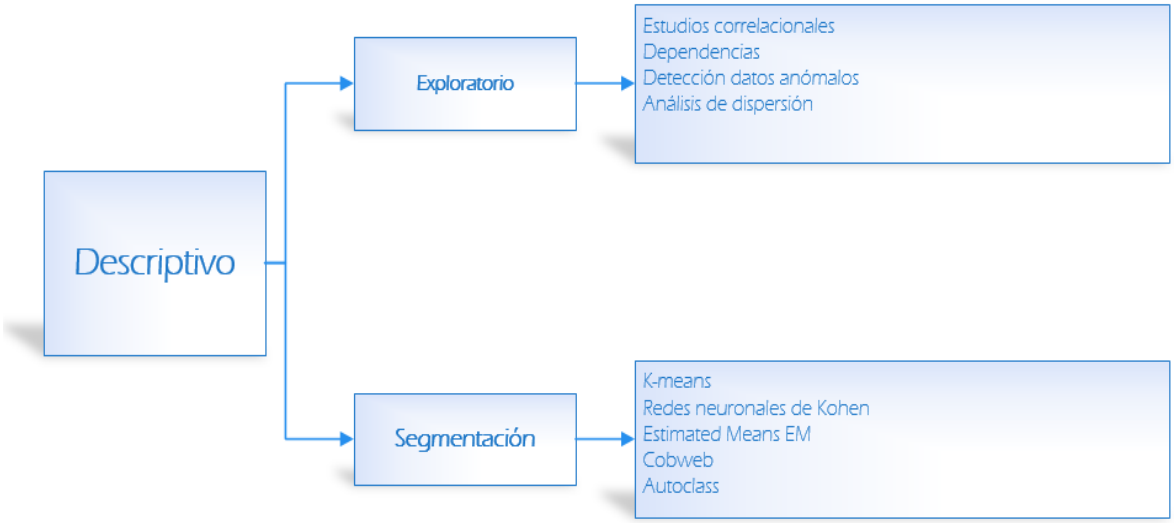


Figura 12. Clasificación de Algoritmos Descriptivos

Adaptado de: (Beltran, 2014)

1.3.3. Técnicas no supervisadas y descriptivas

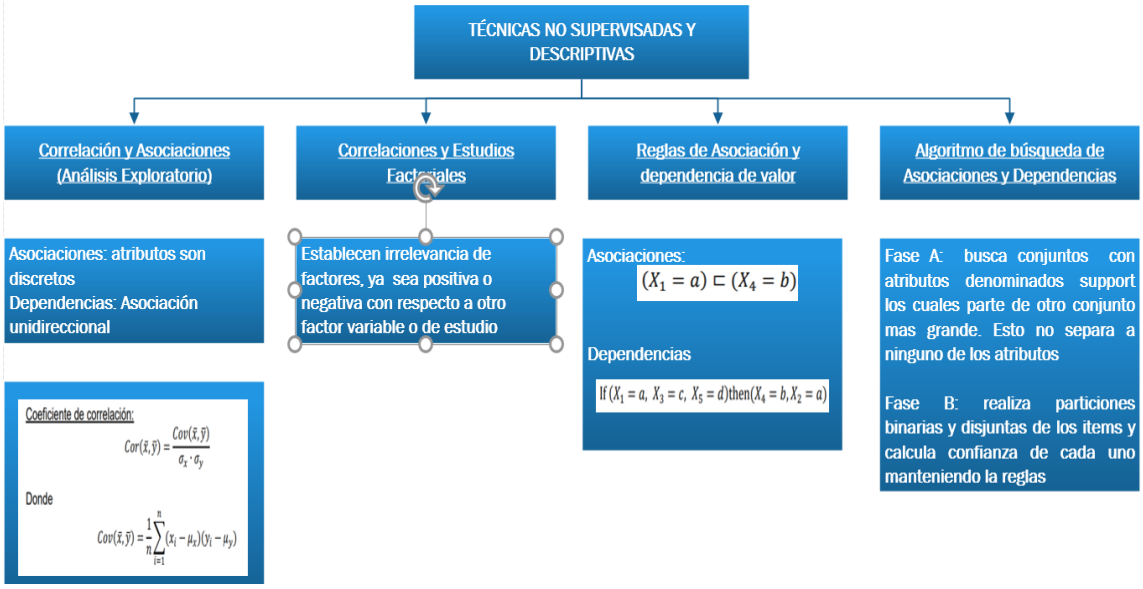


Figura 13. Clasificación de Técnicas Supervisadas y Descriptivas

Adaptado de: (Beltran, 2014)

1.3.4. Técnicas supervisadas y predictivas

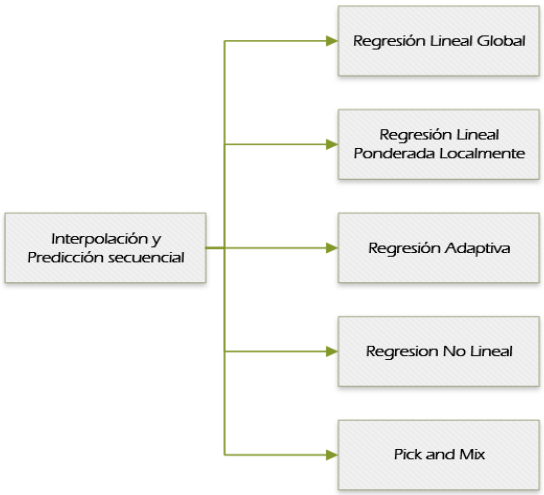


Figura 14. Clasificación de Predicción Secuencial

Adaptado de: (Beltran, 2014)

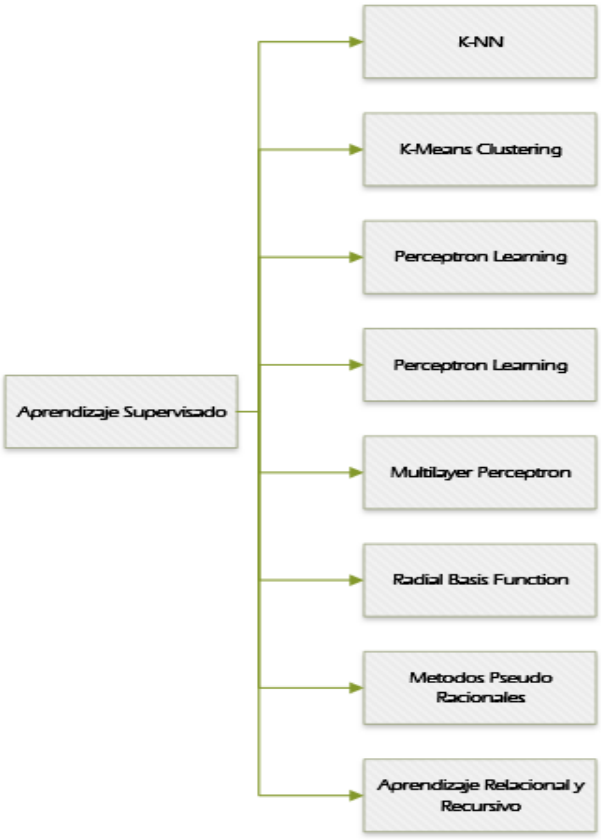


Figura 15. Clasificación de Predicción Secuencial

Adaptado de: (Beltran, 2014)

1.3.5. Métodos de Minería de Datos

El objetivo principal del uso de métodos de minería de datos tiene como meta la predicción de datos escondidos y desconocidos, además la descripción de patrones, tomando en cuenta que se pueden aplicar en diferentes criterios, y así clasificar a los sistemas de minería de datos y los sistemas de aprendizaje inductivo en computadoras entre los cuales están:

- Dependiendo el objetivo del aprendizaje
  - **Clasificación:** clasifica los datos en clases predefinidas
  - **Regresión:** convierte los datos en valores de una función de predicción
  - **Agrupación de conceptos:** agrupa los datos en conjuntos
  - **Compactación:** busca la manera de compactar la descripción de los datos
  - **Modelado de dependencias:** dependencia entre las variables de los datos
  - **Detección de desviaciones:** búsqueda de desviaciones importantes con recto a un valor
- Dependiendo de la tendencia del problema
  - **Sistemas conexionistas:** Redes neuronales
  - **Sistemas Evolucionistas:** Algoritmos genéticos
  - **Sistemas Simbólicos**
- Dependiendo el lenguaje utilizado
  - **Lógica de proposiciones**
  - **Lógica de predicados**
  - **Estructurada**
  - **No simbólicas**

A continuación, se detallará los diferentes métodos de minería de datos los cuales nos servirá para poder representar el conocimiento

- **Clustering**

  Denominada también como segmentación, identifica la tipologías o grupos de elementos los cuales guardan cierta similitud entre ellos y diferencia a grupos externos. Para llegar a las diferentes tipologías o grupo que existen dentro de una base de datos, este tipo de métodos requiere una serie de herramientas como la entrada de información debe ser colectiva y segmentada, y esta mismo representara a los valores concretos para cada uno de los elementos en un momento del tiempo de una serie de variables o con respecto al comportamiento del tiempo en cado de los elementos.

  Estas herramientas se basan en técnicas estadísticas es decir emplean algoritmos matemáticos, generan reglas y redes neuronales al momento de tratar registros; y como resultado el tratamiento de la información se presentará diferentes grupos con variables con valores representativos

- **Association Pattern Discovery**

  Se los denomina también como Asociación, su función es establecer posible relaciones o correlaciones entre acciones o sucesos aparentemente independientes, con el fin de reconocer la apuración o acción de un suceso que puede inducir o generar otros,

  Al igual que el anterior posee fundamentos estadísticos como el análisis de correlación y variación.

- **Sequential Pattern Discovery**

  Identifica las ocurrencias de acciones en el tiempo, puede desencadenar otras posteriormente. En este caso el tiempo es un variable crítica e imprescindible al momento de analizar la información.

- **Patern Marching**

  Asocian de una señal de información entrante con información ya guardada con la cual comparte similitud. Estas habitualmente son usadas en los procesadores de texto. Dentro de la minería de datos nos permitirá identificar problemas e incidencias de las posibles soluciones a buscar.

Por lo general estas se sustentan en técnicas de redes neuronales y algoritmos matemáticos.

- **Forecasting**

  Establece un comportamiento hacia el futuro más problema tomando en cuenta la evolución de pasado y del presente, Esta hace el uso en el tratamiento de series temporales.

- **Simulación**

  Básicamente se lo usa para investigaciones científicas como herramientas de diseño y producción, sometiendo a una amplia serie de condiciones normales y extremas. No solo permite ajustar el diseño, sino que también establecerá márgenes y límites del funcionamiento.

- **Optimización**

  Ha tenido in uso en la resolución de problemas con logística de distribución y a la gestión de stock en los negocios, determinando parámetros teóricos a partir de experimentos de investigación científica. En si busca solucionar los problemas de maximización o minimización de funciones de una serie de variables m encontrando valores de satisfacen las condiciones al máximo y reduciendo los costes.

- **Clasificación**

  Se encarga de agrupar las herramientas las cuales asignan un elemente perteneciente a un grupo o clase. Esto trabaja con establecimiento de clases en función a los valores de las diferentes variables y permite asignar un grado de discriminación o influencia.

  Para este tipo de herramientas la predicción o la evaluación donde normalmente se aplican técnicas numéricas, establecen a cada elemento un valor dependiente de los valores de la variable. Se debe tomar en cuenta que usa técnicas matemáticas, análisis de discriminante, análisis de variaciones, sistemas de conocimiento e inducción a reglas.

1.3.6. Técnicas de Minería de Datos

El uso cotidiano del análisis de datos se ha dirigido a la verificación ha permitido reemplazarlo por un enfoque dirigido al conocimiento, la diferencia de ambos esta que en el segundo descubre la información sin necesidad de formular una hipótesis anteriormente. El algoritmo de la minería de datos tiene la habilidad de reconocer fácilmente los patrones en los datos, por tal motivo la técnica resulta mucho más eficiente que un análisis dirigido a la verificación al momento de explorar datos que provienen de repositorios grandes y de alta complejidad.

Los algoritmos de minería de datos, como se mencionó anteriormente se clasifican en dos grupos: los supervisados o predictivos y los supervisados o descubrimiento del conocimiento.

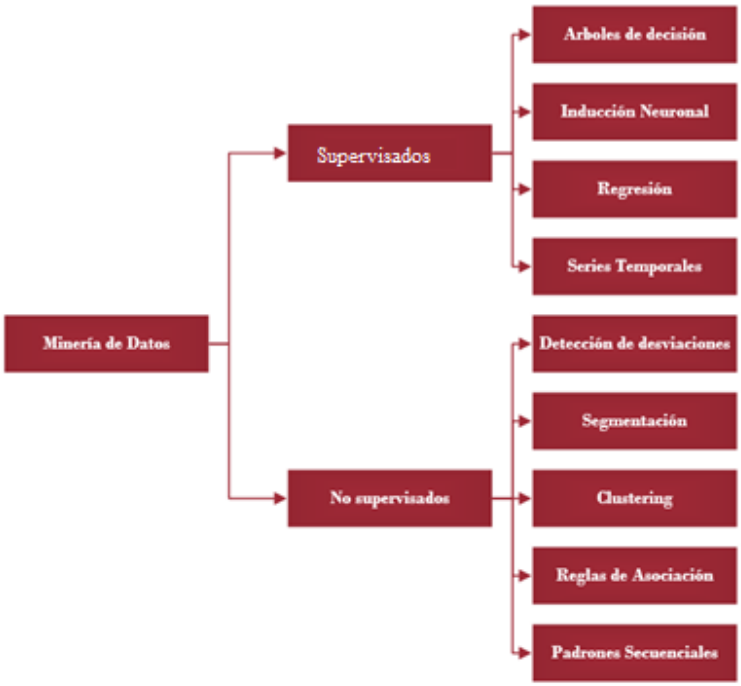


Figura 16. Técnicas de Minería de Datos
Adaptado de: (Beltran, 2014)

Cuando se usa algoritmos dentro de la minería de datos, es necesario ejecutar una serie de actividades anteriormente ya que se deben preparar los datos de

entrada ya que no pueden tener el formato adecuado o pueden tener ruido, así que es por ello por lo que también se debe interpretar y evaluar los resultados que se obtienen.

Las técnicas más utilizadas dentro de la minería de datos son:

- **Métodos Estadísticos**

  La estadística es la herramienta principal de la minería de datos, ya que facilita al manejar grandes volúmenes de dato numéricos y así no poner a duda su efectividad al momento de obtener los resultados y además permite poseer un amplio conjunto de modelos de análisis para cubrir el tratamiento de poblaciones y series de datos,

  - **Anova:** analiza la varianza, que existe en las diferencias de medidas de una o más variables continuas en poblaciones de diferentes grupos
  - **Ji Cuadrado:** resalta la hipótesis de independencia de valores
  - **Componentes Principales:** reduce el tamaño o número de variables que están siendo observadas a un número menor de variables artificiales tomando en cuenta que se conserva mayor parte de la información sobre la varianza entre las variables.
  - **Análisis de clúster:** clasifica una población en un cierto número determinado de grupos y busca la semejanza y diferencias de perfiles que existen dentro de los componentes de una población.
  - **Análisis de discriminante:** clasificación de individuos en grupos que anteriormente fueron establecidos, y permite encontrar la regla de calificación de los elementos de cada grupo para poder identificar las variables que definan al grupo.
  - **Regresión Lineal:** es la relación que existe de una variable dependiente con una independiente, se asume que la relación es línea.
  - **Regresión Logística:** trabaja con variables discretas y requiere que todas las variables sean lineales.

- **Arboles de decisión**

  Se consideran como herramientas las cuales se emplean para descubrir reglas y relaciones mediante la subdivisión o ruptura de la información de manera sistemática que se encuentra contenida en el grupo de los datos. Esta técnica se caracteriza por ser continua y solo se detiene cuando no encuentra diferencias que tenga un significado dentro de las variables de predicción.

  Esta técnica tiene forma de árbol la raíz viene a ser el conjunto integro de datos, mientras tanto que los subconjuntos y los subsubcojuntos vienen a ser las ramas además existe un conjunto denominado como nodo que es donde realiza las particiones

- **Reglas de asociación**

  Realiza un análisis el cual se encarga de buscar coincidencias, realizando correlaciones o co-ocurrencias en los sucesos dentro de la base de datos, aplicando reglas de SI y ENTONES

- **Redes de neuronas artificiales**

  Se consideran como una nueva técnica de analizar la información, ya que, a diferencia de las técnicas tradicionales, estas son capaces de detectar, aprender patrones, y características de los datos en los datos. Su comportamiento es parecido a la del cerebro humano es decir aprende de la experiencia el pasado aplicando el conocimiento.

  Estas se construyen a través de niveles o capas compuesta por nodos o neuronas que son capaces de aprender de dos maneras de forma supervisada y no supervisada.

- **Algoritmos genéticos**

  Estos algoritmos matemáticos se encargan de representar los elementos como cromosomas en un aspecto evolutivo para alcanzar la estructura y la composición más adecuada para las áreas de supervivencia, al considerar un proceso evolutivo de búsqueda y optimización de cada uno de los elementos está dispuesto a mutaciones o cambio afectándolos.

Además, estos usan muchas técnicas biológicas de reproducción como la mutación o el cruce para ser utilizadas en la solución de todo tipo de problemas de búsqueda y optimización.

- **Lógica difusa**
  Esta técnica permite un tratamiento probabilístico, permitiendo crear barreras difusas entre los grupos categorizados de un colectivo o de los diferentes elementos.
- **Series temporales**
  Estudia a las variables a través del tiempo, partiendo desde el bajo supuesto que no va a producir cambios estructurales, para realizar las predicciones. Suelen utilizar ciclos, tendencias y estacionalidades, que se diferencian cada uno por el tiempo que abarca,

1.4. Modelos de Gestión

En los años 70 las empresas alrededor del mundo buscaban la manera de actualizar el hardware dentro de sus empresas con el fin de que sus equipos trabajen a mayor velocidad, sin embargo, con el paso de los años 80 las empresas decidieron poner mayor interés en el desarrollo de software, y desde el año 90 se han concentrado más en la gestión de servicios.

El objetivo de la gestión de servicios es conseguir una actividad que madure día a día, poniendo en práctica cada la teoría que surgen cotidianamente en el ámbito de empresarial. La Gestión de IT ha ido creciendo inimaginablemente con lo que ha llevado a convertirlo en un estándar denominado ITIL, con el fin que las empresas desarrollen sus propicios marcos de Gestión de Servicios IT.[6]

Los modelos de Gestión son considerados como una base para el desarrollo, para el desarrollo de procesos que no están descritos dentro de ITIL. A continuación, se mencionarán algunos modelos de Gestión TI desarrollados en

[6] (Oriente, 2014)

los últimos tiempos: Para la implementación de nuestro servicio, no enfocaremos en el sistema de gestión ITIL, ya que esta se encuentra orientada a las buenas prácticas dentro del área de TI.

1.5. Sistema de gestión ITIL

A ITIL se lo conoce como un código para las buenas prácticas dirigidas para el cumplimiento de objetivos médiate el uso de un enfoque sistemático del servicio de IT dentro de los procesos, procedimientos y establecimiento de estrategias para la gestión de la infraestructura.[7]

Entre los principales objetivos de usar ITIL dentro de un Sistema de Gestión TI son:

- Brindar calidad dentro de la gestión
- Incrementar la eficiencia
- Procesos de negocio e Infraestructura debidamente alineados
- Disminuir los riesgos vinculados a los servicios de TI
- Crecimiento del negocio.

Figura 17. Ciclo de Vida ITIL
Adaptado de: (Oriente, 2014)

[7] (Oriente, 2014)

1.5.1. Estrategia del servicio

El objetivo principal de la estrategia del servicio es definir la forma en la cual se va a representar los planes y patrones del servicio que se va a implementar con el fin de alcanzar a cumplir los objetivos del negocio de la empresa.

Además, posee otros objetivos secundarios los cuales se pueden resultar:

- Identificación de servicios y clientes
- Identificación de oportunidades que el servicio brindara
- Modelamiento de servicios provisionales
- Coordinación y documentación de activos del servicio
- Optimización del rendimiento del servicio

Para lograr que la estrategia del servicio sea exitosa se compone de un proceso de 3 pasos:

- **Gestión del Portafolio**
  Asegura que todos los servicios están totalmente bien definidos y estén correctamente orientados al cumplimiento del objetivo de la empresa es decir observar las actividades del diseño, la transición y operación brinden algún valor a la empresa. Además, que proporciona un proceso para decidir los servicios que quiere ser proporcionados en la empresa tomando en cuenta:

  - Nivel de riesgo
  - Necesidades del negocio
  - Estrategia para responder a cambios
  - Control de servicios
  - Análisis de servicio viables y obsoletos

- **Gestión Financiera de servicios IT**

  Su objetivo es encontrar el equilibrio que existe entre la calidad y el coste del servicio, además también de mantener el equilibrio entre el servicio y el usuario.

- **Gestión de Relación con el negocio**

  Esta se encarga de la identificación de necesidades del usuario con el fin de asegurar que el servicio cumpla con estas, tomando en cuenta los cambios dentro de la empresa fijándose en el tiempo y las circunstancias También debe fijar el nivel de satisfacción del cliente con el servicio manteniendo la comunicación constante, con el fin de buscar mejoras.

1.5.2. Diseño del Servicio

El diseño de un servicio IT consiste en el manejo de la practicas, procesos y políticas del área de TI, facilitando la operación de los servicios tomando en cuenta la calidad en el que este se entrega dicho servicio. Diseñar un servicio IT de manera efectiva con lleva solamente a realizar mejoras mínimas durante el ciclo de vida de este.

- **Coordinación del diseño.**

  Cuando se plantean meta y objetivos se debe asegurar que estos se cumplan en diseño de servicio, siempre debe mantener un único punto de coordinación y control para las diferentes actividades y procesos.

  Un diseño de servicio se encarga de gestionar la información, arquitectura, tecnología y valores de control para cumplir los requerimientos solicitados. Con el fin de realizar cambios dentro de la infraestructura física y lógica de la empresa coordinando planificaciones. Para no afectar a lo servicio existentes.

- **Catálogo de Servicios**

  Es la fuente que abarca la información de los servicios que están ejecutándose, el cual sirve como herramienta para las personas

autorizadas, dentro de este se detallan las características más importantes de los servicios a implementar o ejecutándose tomando en cuenta las políticas definidas para los mismos. Este debe estar disponible todo el tiempo para el personal de TI para brindar soporte

- **Nivel de servicio**

  Es el encargado de controlar y asegurar que los procesos de la gestión de servicios del área de IT cumpliendo los acuerdos de nivel operativo OLAs, además verifica que los incidentes o requerimientos de soporte sean los más apropiados para el cumpliendo de objetivos.

  Además, no solo se encarga del control si no de la documentación, monitoreo, medición y reporte de los niveles de servicios de IT, para proveer y promover las acciones correctivas de ser necesarios. Sirve como una herramienta de comunicación entre el usuario y el servicio.

  Este debe ser medible para obtener un feedback o respuesta por parte del usuario evaluando la satisfacción del usuario y la operabilidad del mismo.

- **Suministradores**

  Asegura que el servicio de IT cumpla con los requisitos del negocio, con el fin de cumplir las necesidades de la empresa y del usuario final tomando en cuenta la evaluación de los SLAs acordados en el nivel de servicio.

- **Disponibilidad**

  Un servicio de IT debe estar disponible 365x7x24, ya que este debe cumplir con el nivel de servicio requerido, este es un factor medible dentro de los SLAs, ya que la mayoría de los servicios por implementarse o implementados pueden causar un gran impacto dentro de la empresa.

- **Capacidad**

  Un servicio de IT y la infraestructura IT deben tener las características necesarias para su funcionamiento dentro de la empresa tomando en cuenta los cambios futuros que se puedan dar dentro de la misma. Este también se ajusta dentro del área de servicio ya que al no cumplir con la

capacidad tanto de servicio como de infraestructura con lleva a las incidencias relacionadas con el rendimiento del mismo.

- **Continuidad del servicio IT**
  Analiza los riesgos que pueden afectar dentro de la continuidad de funcionamiento del servicio de IT, el área debe proveer niveles de servicios que aseguren el mismo para no generar un alto impacto a la empresa.
- **Seguridad de la información**
  Se encarga de la alineación de la seguridad del área de IT, tomando en cuenta que la información de ser confidencial, integra y debe estar disponible siempre para el funcionamiento del servicio. Aquí se definen las reglas o políticas para la divulgación de ala información ya que se debe proteger con los intereses de la empresa es decir que la información, sistemas y comunicación no debe sufrir daño alguno.

1.5.3. Transición del Servicio

Es la guía para desarrollo y mejora de las diferentes capacidades que van a ser introducidos o implementadas dentro de los nuevos servicios o dentro de los servicios ya existentes, con el fin planificar y administrar los cambios dentro de los mismo de manera eficiente y efectiva.

- **Planificación y transición del Servicio**
  Coordina y planifica las transiciones de los requerimientos del servicio.
- **Administración de activos y configuraciones del servicio**
  Los servicios que se encuentran activos deben ser controlados de manera apropiada, y la información tanto del servicio IT como de la infraestructura deben estar totalmente disponibles y debe detallar como se encuentran configurados.
- **Administración de cambios**
  El ciclo de vida de los cambios debe ser controlado permitiéndole así la implementación con el menor impacto en el área IT

- **Administración de Versiones**
  Es el control de la implementación, pruebas y la puesta a producción de versiones mejoradas o con nuevas funcionalidades las que fueron solicitadas por la empresa. Siempre tomando en cuenta la protección de los servicios ya existentes.
- **Gestión del Conocimiento**

Sirve para compartir los diferentes puntos de vista, ideas, y experiencias del servicio con el fin de asegurar que estén disponibles en lugar y en el momento exacto.

1.5.4. Operación del Servicio

Día a día las empresas dependen más y más de las tecnologías para llevar acabo sus funciones, con el fin de cumplir las necesidades que el negocio requiere, Con frecuencia el departamento de IT no considera que los objetivos de la empresa que son parte de él, si este se considera a sí mismo como un proveedor. ITIL considera que el área de TI debe estar trabajando a la par con empresa, es decir integrando al departamento con la empresa y así trabajar a la par en el cumplimiento de objetivos.

En una empresa los usuarios requieren mucho de los servicios de TI, y el área busca entregar un servicio el cual sea constante y estable, incluyendo una disponibilidad 365x24x7. Hay que tomar en cuenta que al momento de gestiona en base a los procesos y actividades lógicas con el fin de cumplir un objetivo, al utilizar estos procesos nos proporciona las siguientes ventajas:

- Un proceso se encarga de describir la forma de llegar al objetivo
- Cada proceso debe estar defino por un input y output, los cuales juntos a otros procesos permitirá cumplir los objetivos de ambas partes
- La organización se compone de procesos, y todos estos procesos pueden ser vigilados uno a uno

- Las personas asignadas al manejo de los procesos son eficientes, eficaces y muestran resultados al ejecutar los procesos.
- Divide las respetabilidades para evitar conflictos internos.

Los procesos básicos que se aplica dentro de la operación son:

- **Gestión de Eventos**
  Se encarga de la gestión de los eventos del servicio dentro de su ciclo de vida ITL, para detectar los diferentes cambios de estado que son considerados como importantes.
- **Gestión de Incidencias**
  Busca que la restauración del servicio se tan rápida con el fin de minimizar los impactos que afecten a la empresa, tomando en cuenta los niveles de servicio.
- **Gestión de Requerimientos**
  El objetivo principal es llevar el control de las solicitudes o requerimientos por parte de los usuarios, para mantener la satisfacción de los mismos.
- **Gestión de Problemas**
  Tomar el control de todos los problemas presentes dentro del sistema, minimizando el impacto de las incidencias.
- **Gestión de Accesos**
  Administra los privilegios de los usuarios, y otorga los permisos dependiendo su necesidad o puesto dentro de la empresa.

1.5.5. Mejora continua del Servicio

El objetivo principal es crear una alineación de todos los servicios de IT, con el fin de modificarlos o cambiarlos de acuerdo con las necesidades de la empresa creando ciclos de vida de estrategia, diseño, transición y operación de los mismos.

Para mejorar un servicio dentro de una empresa de debe tener en cuenta los siguientes pasos:

- **Definir lo que se quiere cambiar:** analizar cuál va a ser la estrategia de mejora es decir la visión, la necesidad de la empresa, metas y procesos
- **Definir lo que realmente se va a cambiar:** cuando ya se ha planificado una estrategia y un diseño, y dicha información esta lista se procede a crear un ciclo de vida.
- **Obtención de datos: en** una empresa los datos pueden ser obtenidos de diferentes formas, esta puede ser manual o automática, estos deben estar acorde a las metas y objetivos de la empresa.
- **Procesamiento de los Datos:** una vez los datos recolectados y procesados se alinean con los factores críticos de la empresa.
- **Análisis de datos e Información**

  Dentro del análisis, la empresa busca le cumplimiento de los SLA, identificación de problemas, cumplimiento de disponibilidad d y si esta de acorde a la capacidad solicitada, cual es el plan de contingencia a seguir entre las más importantes.
- **Uso y presentación de la información**

  Cuando se obtiene mucha información esta debe tener un formato específico, ya que esta debe ser filtrada para ser utilizada en diferentes momentos del ciclo de vida del servicio.
- **Acciones Correctivas**

  Al conocimiento obtenido se le puede dar diferentes usos; con el fin de mejorar y corregir los servicios o procesos que abarca el área de IT aplicando el ciclo PDCA

2. Capito II. Levantamiento de Información

Introducción

Dentro de este capítulo se analizará la infraestructura actual de la empresa, con el fin de verificar como se encuentra actualmente y así verificar si cumple con los requisitos necesarios para implementar el servicio dentro de la misma. Además, con el uso del estándar ANSI/TIA942 como guía, podremos verificar cuales son las fallas dentro de la misma

2.1. Subsistema de Arquitectura

La empresa dispone de un área dedicada al alojamiento de todos los equipos que componen el data center. Además, posee una separación entre el cableado eléctrico y el cableado de la red mediante el uso de canaleta metálica y canaleta plástica

2.1.1. Ubicación del Data Center

El data center en el cual vamos a instalar nuestro servicio se encuentra ubicado en la ciudad de Cayambe, en la finca principal ROSAPRIMA R1. Este Ocupa un espacio físico de 3 de ancho, 4 de largo y 6 de Alto en una estructura de paredes de ladrillo y cemento

Figura 18. Ubicación Rosaprima R1

Adaptado de: Google Maps, 2017

2.1.2. Puertas de acceso y techo

La puerta del data center en la finca cumple con los requisitos establecidos tiene acceso desde el exterior, además es extraíble con apertura hacia afuera, también cuenta con seguridad para que solo el personal autorizado tenga acceso, sus dimensiones son altura de 2,13 metros de alto y un acho de 1 metro

Figura 19. Acceso Data Center Rosaprima R1

2.1.3. Iluminación

Dentro del data center se cuenta con un mínimo de 8 lux de iluminación medido horizontalmente y 380 lux como máximo verticalmente además esta posee una alimentación eléctrica independiente cumpliendo así lo recomendado por la norma ANSI/TIA-942

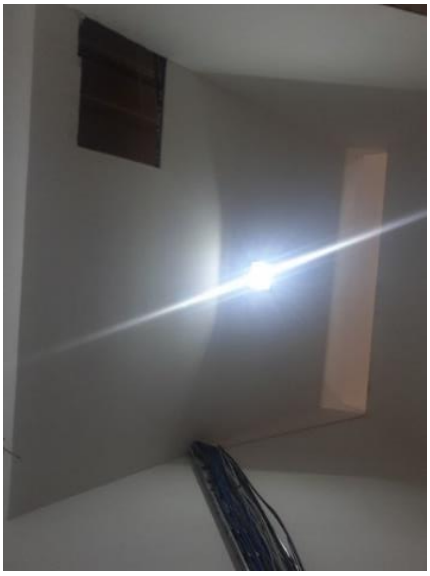


Figura 20. Densidad de Iluminación

2.1.4. Piso y Techos Falsos

Dentro de la finca no trabajamos con pisos falso, todo el cableado que llega al Data Center por un techo falso de Gypsum, el cual puede soportar el peso de las canaletas y de las tuberías las cuales cruza por la oficina aproximadamente de 2.4 Kpa.

2.2. Subsistema de Telecomunicaciones

Dentro de este subsistema se procederá a revisar cuales son los elementos infraestructura de telecomunicaciones, el cual se encarga de manejar todos los servicios informáticos de la empresa principalmente de la finca.

Además, se especificará cuáles son las conexiones principales, las cuales van a permitir trabajar dentro de la empresa, con el acceso a la red WAN, donde se va a encontrar nuestro servidor de data mining

2.2.1. Topología Física de la Empresa

La Topología de la empresa se encuentra compuesta por los diferentes dispositivos de comunicación, ordenadores, incluso posee una serie de Firewalls los cuales permite mantener una mayor seguridad de la información o datos de la empresa. Además, mediante una VPN contratada a la empresa Telconet.

Para una rápida visualización en la FIG19, se puede evidencia un modelo diseñado en Microsoft Visio de cómo se encuentra estructurada la empresa actualmente.

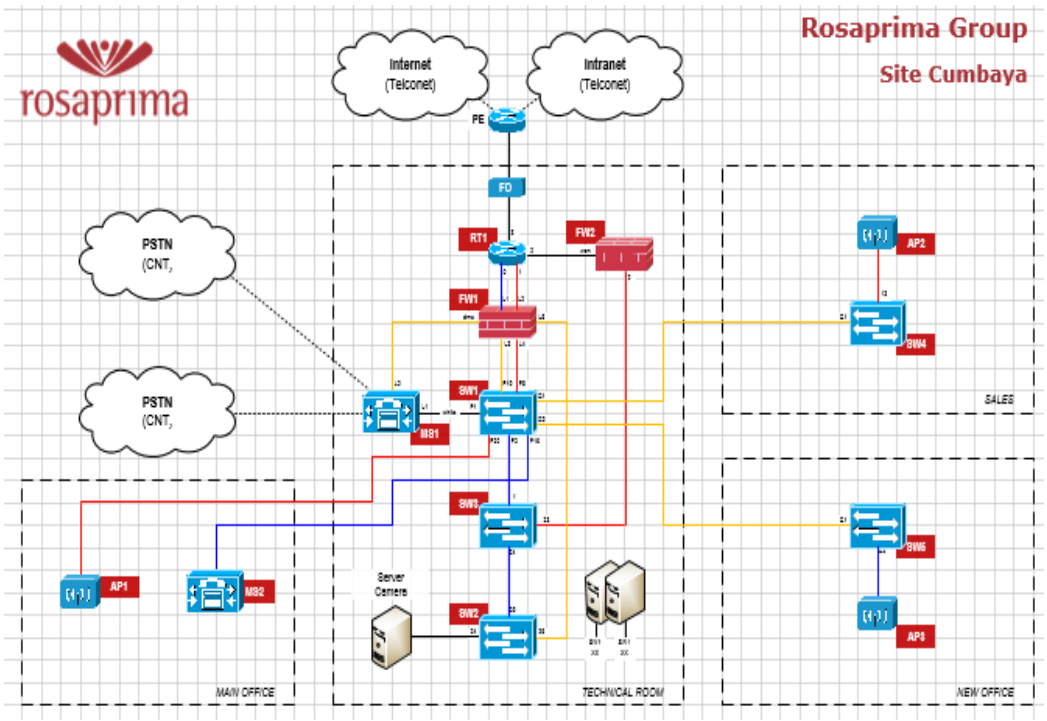


Figura 21. Topología Finca R1

Adaptado de: Rosaprima,2016

2.2.2. Análisis de la Topología de la Empresa

- **Cuarto de Telecomunicaciones (TR)**

Espacio en el cual se alberga el cableado de equipos provenientes del exterior, además se combina el MDA y HDA.

- **Cuarto de Equipos (ER)**

Lugar en donde se alberga todo alberga los equipos y servicios de telecomunicaciones, para distribuir a las diferentes áreas de trabajo o departamentos., Se debe considerar que estos equipos

- **Área de distribución principal (MDA)**

Se ubica junto en al ER, es aquí donde se encuentra la distribución principal del cableado es decir aquí se están routers, switches, centrales telefónicas, equipos SAN, además los equipos por medio el cual el proveedor nos brinda el servicio de internet. El data center se encuentra en un punto céntrico que no excede las longitudes del cableado, y en total mente cerrado ya que se debe manejar con cuidado los equipos de alimentación eléctrica

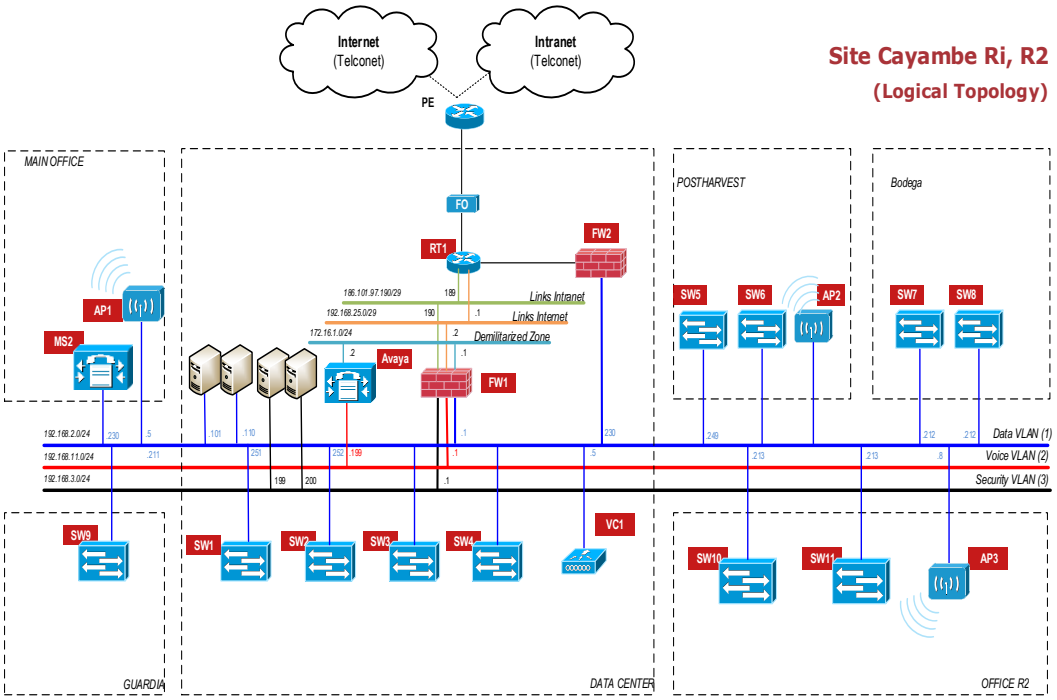
- **Área de distribución horizontal (HDA)**

Área de distribución del cableado y equipos activos para la finca. Aquí se encuentran los racks tomando en cuenta que se separan los de Fibra óptica con los UTP, también están los switchs y los pach panels con el fin de reducir el uso de cable UTP.

- **Área de Distribución de Equipos (EDA)**

Lugar donde se encuentran conectados los equipos finales es decir computadores, servidores, y equipos de telecomunicaciones.

2.2.3. Topología Lógica.



RT1 Router Cisco 800
FW1 Firewall Fortinet Fortigate 80C
FW2 Firewall Polycom VBP 4555
SW1 Switch Cisco 2960(SW_CAYAMBE_1)
SW2 Switch Cisco 2960(SW_CAYAMBE_2)
SW3 Switch TP-LINK
SW4 Switch Cisco SG200
SW5 Switch Cisco SF300
SW6 Switch TP-LINK
SW7 Switch Cisco 2960
SW8 Switch TP-LINK
SW9 Switch D-LINK 16 PUERTOS
SW10 Switch Cisco 2960
SW11 Switch TP-LINK
AP1 Aruba Access Point IAP- 205-RW
AP2 Cisco Access Point IAP- 205-RW
AP3 Cisco Access Point IAP- 205-RW
MS1 Avaya IP S05
MS2 Polycom HDX 7000

Figura 22. Topología Lógica Rosaprima
Adaptado de: Rosaprima,2016

2.2.4. Administración del Cableado Estructurado

La empresa últimamente ha realizado varios cambios en la infraestructura, el área de IT aprovecho de esto para realizar un recableado de la red, es decir un

cambio categoría 5 a 6 con el fin de obtener una mejor tasa de transmisión de los datos; con un tiempo de duración estimado de 10 años. Por lo cual se lo maneja de una manera ordenada y estructurada aplicando estándares como el ANSI/TIA, ya que son necesarios al momento de montarlos en el rack con el fin de manejar un control unificado de todas las instalaciones del data center.

- **Rack**

Esta construido de metal, con el fin de alijar los servidores y equipos de telecomunicaciones que son necesarios para la empresa. Según el estándar ANSI/TIA recomienda que el hardware debe ser compatible con este ya que debe cumplir con altura de 1.75m de alto y 600mm de ancho, y se debe considerar una separación de 15 cm entre el rack y el hardware instalado con el fin de permitir conexiones dentro del mismo

A continuación, en la FIG21, se puede evidenciar como actualmente se encuentra conformado el rack de la empresa

Figura 23. Rack Data Center Rosaprima R1

2.3. Subsistema de Eléctrico

2.3.1. Energía

La fuente de energía principal de la Finca Rosaprima R1, es proporcionada por la empresa EMELNORTE, al encontrarse en un lugar apartado del pueblo, los cortes de energía suelen ser repentinos y prolongados, es por ello que se hace uso de un generador de energía ubicado dentro de la misma.

De acuerdo con la empresa proveedora de energía, el voltaje que proporciona a las instalaciones es de 122 V de corriente con una frecuencia de 60 Hz.

Figura 24. Emelnorte proveedora de Electricidad en Rosaprima R1
Adaptado de: Emelnorte,2010

2.3.2. UPS

Aparte de obtener energía por parte la compañía de eléctrica de la ciudad de Cayambe, el data center se encuentra dispone de una serie de UPS, los cuales son accionados al momento del corte de energía

Como se puede evidencia en la Fig22, la empresa utiliza UPS APC 550 para proteger y mantener los equipos durante un corte de energía

Figura 25. UPS APC 550 usado en Finca Rosaprima R1
Adaptado de: APC,2016

2.3.3. PDU

Para conectar el gran número de equipos que disponemos dentro del Data center, se usó una serie de regletas las cuales brindan varias tomas eléctricas para los diferentes equipos, sin embargo, no brindan protección a sobretensiones, estas están instaladas tanto en el rack como en el piso.

2.3.4. Generador

Este equipamiento completo nos permite suministrar de energía alterna para los equipos tanto de computo como de telecomunicaciones en caso de presentar posibles cortes energéticos por parte de la empresa proveedora del servicio. Además, este se encuentre diseñado para suministrar corrientes armónicas, debido a que se trabaja con una serie UPS, evitando la sobrecarga térmica

La empresa dispone de un generador que funciona a base de Diesel, el cual entra en funcionamiento cual un corte se genera, se lo puede visualizar en Fig23.

Figura 26. Generador Instalado en Rosaprima R1
Adaptado de: Rosaprima,2016

2.4. Subsistema de Mecánico

2.4.1. Sistema de Aire Acondicionado

El data center mantiene un ducto de ventilación, por el cual permite que el calor no se concentre dentro de él, sin embargo, no tiene un sistema de aire acondicionado para poder trabajar normalmente

2.5. Tablas de fallas

Con el fin de mejorar la infraestructura perteneciente al área de IT, se realizó un levantamiento de información en los diferentes subsistemas con el fin encontrar las diferentes fallas presentes, con el fin de proponer de mejoras tomando en cuenta el uso de normas y estándares.

A continuación, en la Tab1 se pueden verificar los errores que se presentaron.

Tabla 1
Fallas ubicadas dentro del Data Center

Subsistema	**Fallas**		
Subsistema Eléctrico	Acceso al panel de monitoreo	Cortes de Luz Constante	Conexiones a tierra
Subsistema Telecomunicaciones	Etiquetas en Cableados	Etiquetas servidores	Separación de equipos
Subsistema De Arquitectura	Carencia de piso Falso	Puerta no es deslizable	seguridad en la puerta de acceso
Subsistema Mecánico	Tuberías de Ventilación	Carencia de sistema de enfriamiento	

3. CAPITO III. DISEÑO E IMPLEMENTACIÓN DEL SERVICIO

Introducción

Antes de proceder con el diseño y la implementación de nuestro servicio se realizará un análisis algunos casos de estudios ya implementados y funcionales, para tomarlos como referencia al momento de crear el nuestro. A continuación, en la Tab2 se presentará un cuadro comparativo de los diferentes gestores de bases de datos con el fin de seleccionar.

Tabla 2
Cuadro Comparativo Gestores de Base de Datos

DBMS	Características	Ventajas	Desventajas
DB2	IBM Integración de XML Arquitectura Relaciona	Multiplataforma Elimina tareas rutinarias Uso menor de recursos de hardware	No es robusto a diferencia de otros
MySQL	Licenciado y Libre	Fácil manejo y aprender Multiplataforma Código Abierto Rapidez en operaciones Fácil configuración	Triggers básicos Conversiones de datos Privilegios en tablas manuales
Oracle	El uso de arrays en la herencia de tablas	Soporte técnico	Con una mala configuración los resultados pueden ser
SQLite	Libre Relacional Lenguajes de programación lo incluyen en módulos y bibliotecas	Multiplataforma Compatibilidad en leguajes de programación	Tamaño limita de 2 TB Asignación de valores individuales Portabilidad
InterBase	Licenciado Arquitectura única procedimientos y triggers potentes	Compatibilidad de Sistemas operativos Copias de Seguridad	No permite realizar particiones
SQL Server	Licenciado Microsoft Herramientas de análisis e Integración Recuperación de datos rápida y eficaz Portabilidad	Transaccional Escalabilidad Seguridad Entorno grafico Soporta procedimientos almacenados	Es pagado Uso de recursos computacionales como memoria RAM

Se tomará como guía la documentación y herramientas Microsoft SQL Server, tomando en cuenta que sus características, ventajas y desventajas implementación de Minería de Datos con herramientas Microsoft

3.1. Requisitos para instalar un servidor de minería de Datos

- **Hardware:** en la TAB2 se mencionarán cuáles son los requisitos de hardware que Microsoft recomienda tener como mínimo en el servidor físico para crear el modelo de minería de datos.

Tabla 3
Requisitos de Hardware

Hardware	
RAM	512 MB
Procesador	AMD OPTERON
	AMD ATHLON
	INTEL XEON soporte EM64T
	INTEL PENTIUM IV soporte EM64T
	INTEL iX
Disco Duro	6 GB en adelante
Monitor	VGA (800 x 600)
RED	xEthernet (TPC/IP, VIA)

- **Software:** en la TAB3 se mencionarán cuáles son los requisitos de software que Microsoft recomienda tener como mínimo en el servidor físico para crear el modelo de minería de datos.

Tabla 4
Requisitos de software

Software		
Sistema Operativo	Windows 8	Standart
	Windows 8	Professional
	Windows 8	Enterprise
	Windows 10	Standart
	Windows 10	Professional
	Windows 10	Enterprise
	Windows Server 2012	Standart
	Windows Server 2012	Essencial
	Windows Server 2012	Datacenter
	Windows Server 2012	Foundation

	Windows Server 2012 R2	Standart
	Windows Server 2012 R2	Essencial
	Windows Server 2012 R2	Datacenter
	Windows Server 2012 R2	Foundation
	Windows Server 2016	Standart
	Windows Server 2016	Essencial
	Windows Server 2016	Datacenter
Componentes	Framework	4.6
	Actualización	KB2919355

3.2. Implementación de Minería de Datos con herramientas Microsoft

En minería de datos como primer paso consiste en la definición del problema planteado y como usar los mismos datos para dar una posible solución al mismo. Para ello se debe tomar en cuenta las diferentes estrategias que ayudaran al momento de implementar la minería de datos:

- **Pronóstico:** al momento de trabajar una gran cantidad de datos almacenados se debe tomar en cuenta cual es la carga de información que reside dentro del servidor además el tiempo de inactividad del mismo. Esto con el fin de realizar cálculos en las ventas y de predicciones a futuro de la empresa
- **Riesgo y Probabilidad:** Determinar un punto de equilibrio, el cual simule diferentes escenarios de riesgo y asigne probabilidades de diagnóstico para la elección de clientes
- **Recomendaciones:** determinación de un cierto número de producto producido el cual pueda ser vendido en el mercado.

- **Búsqueda de secuencias:** analizar cuáles fueron los principales artículos seleccionados por los clientes con mayor frecuencia con el fin de predecir nuevos posibles eventos.
- **Agrupación:** la creación de grupos de productos, clientes y eventos con el fin de analizar y crear predicción de afinidades.

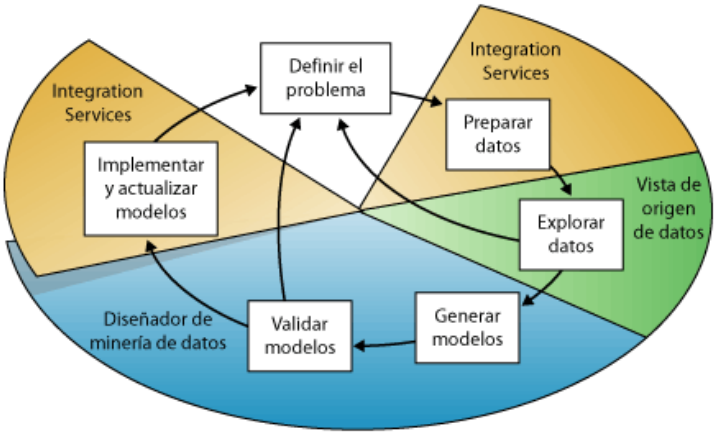


Figura 27. Ciclo de Vida de Minería de Datos con SQL server

Adaptado de: Microsoft,2016

3.3. Casos de Estudio Referenciales

3.3.1. Ejemplo 1: Análisis modelo predictivo para que entidades Bancarias entreguen créditos a personas naturales.

Definición del Problema

Una entidad bancaria realiza entrega de tarjetas de crédito para sus clientes con el fin de emitir nuevos préstamos o para dar un seguimiento a los mismos. Para que esta sea una institución sea competitiva debe adquirir nuevos clientes, sin embargo, se limita al incumplimiento de pago.

Con el fin de maximizar las ganancias el banco utiliza una tarjeta de crédito de puntación para que el cliente o el listado de clientes cumpla a tiempo el pago

de sus deudas. La tarjeta es utilizada como un modelo predictivo, con el fin de proyectar los valores de incumplimiento.

Preparación y Exploración de Datos

Los datos que se identificados para obtener la información de los clientes y realizar el análisis predictivo de riesgo son:

- Histórico de créditos bancarios
- Histórico de pagos y créditos de la superintendencia de bancos
- Datos demográficos de tercero

Para ello dicho banco debe trabajar con los siguientes requerimientos para realizar la explotación de los datos:

- Tratar con valores típicos y faltantes
- Preparar las variables continuas y categorizar las variables
- Identificar y resolver las variables especialmente las que se encuentren correlacionadas.

Generación del modelo

- El administrador de base de datos o el desarrollador es quien selecciona el modelo a usar, este puede ser una solución Microsoft o puede ser proveniente de terceros, con el fin de encontrar el algoritmo indicado para predicción de riesgo de pago de créditos.

Validar el modelo

- Para validar el modelo se debe tomar en cuenta que este es un gran conjunto de datos y por ello se debe probar desde lo más pequeño.
- El modelo debe presentar una regresión lógica por defecto

Implementación y Actualización del Modelo

- Una vez validado el modelo se procede con la implementación del mismo en la base de datos
- Continuamente el banco realizará monitoreo a las tarjetas de los clientes ya que estas son las que ayudan a la evolución del modelo
- Realizar revalidaciones al modelo bajo desarrollo, permitirá al banco crear nuevas versiones.

Dentro de este ejemplo, se puede denotar de una manera más sencilla de las diferentes acciones que se toman en cada uno de los pasos del procedimiento Microsoft.

3.3.2. Ejemplo 2: Análisis de modelo de predicción de Florícolas en Colombia con respecto al clima.

Definición del Problema

Las florícolas de Colombia requieren encontrar un modelo de proyección o predicción de rosas la cuales le permita pre visualizar los volúmenes de producción generados en un día con el fin de aplicar a nivel administrativo y gerencial tomando en cuenta las variaciones climatológicas y las fechas del calendario, y predecir cuales son las condiciones para un buen cultivo

Preparación y Exploración de Datos

Los datos que van a ser utilizados para realizar este modelo

- Variedad
- Calidad
- Fragancia
- Color
- Temperatura

Para las florícolas trabajar con los siguientes requerimientos para realizar la explotación de los datos:

- Variabilidad de la Temperatura
- Grados – Días (Promedio entre temperatura máxima y mínima)
- Efectos ambientales (calentamiento global, Efecto de invernadero)

Generación del modelo

- Para la generación de este modelo, el investigado utilizo como herramienta Microsoft Excel, mediante el uso de fórmulas construir graficas de producción tomado en cuenta diferentes variables que existen dentro de él.
- Para la generación del modelo se toma mucho en cuenta en la producción del día anterior, además para analizar la curva va contando el número de días en los cuales la planta creció

Validar el modelo

- Para validar el modelo el autor realiza una comparación de crecimiento de la planta de las diferentes variedades que fueron seleccionadas para el estudio en una cierta cantidad de días, y además el porcentaje de la variación de la curva. Tomando en cuenta que el uso del clima aun lado.
- El modelo indica cuanto corte de tallos se puede producir en un día, tomando en cuenta la producción del día anterior, crecimiento de la curva t el porcentaje de éxito de cosecha

Implementación y Actualización del Modelo

- El modelo no requiere implantación ya que este trabaja en Excel, y puede ser mostrado en una hoja de cálculo normal. Sin embargo, se toma en cuenta que dicho programa dispone de add-ins más sofisticados para Data Mining.

Dentro de este ejemplo de denota como buscar patrones repetitivos mediante el uso de gráficas y formulas mediante el uso de datos ingresados manualmente. Sin embargo, Excel en este ejemplo no fue explotado a su mayor capacidad ya que al sr una herramienta Microsoft también dispone de más herramientas para realizar la búsqueda de patrones repetitivitos.

3.4. Diseño e Implantación del servicio para ROSAPRIMA

3.4.1. Descripción de la Topología para el servicio de Minería de Datos

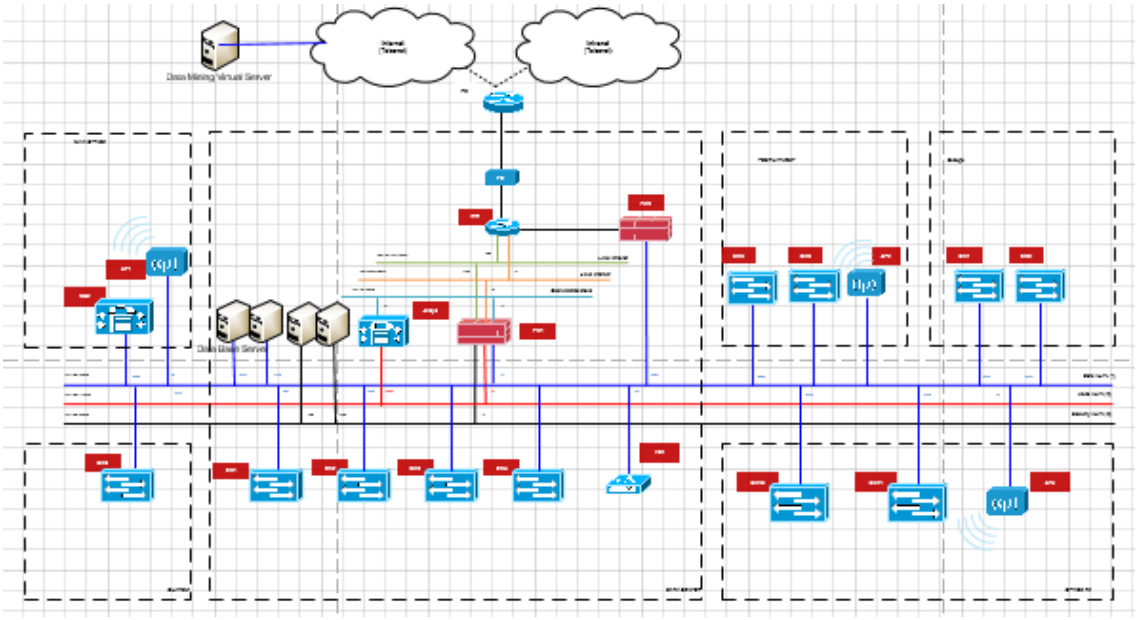


Figura 28. Topología del Actual en la Empresa Rosaprima

Como se evidencia en la FIG28, en la sucursal de Cayambe la empresa dispone de su servidor de Base de Datos el cual se encarga de albergar los datos de producción. En cada sucursal se dispone se ha instalado un router de la empresa proveedora de servicio de internet TELCONET el cual proporciona la salida al mundo exterior, y mediante una VPN dentro de esta nos permite tener comunicación con todas las sucursales.

Además, se dispone de un servidor virtualizado dentro de las plataformas AMAZON, al cual también tenemos acceso mediante una VPN, y alberga el

software necesario que se encargara de realizar las tareas de minería de datos y generación de algoritmo.

3.4.2. Descripción del Servidor de minería de datos

En TAB4, se podrá visualizar cuales son las principales características la cual se procederá a realizar nuestro modelamiento de minería de datos.

Tabla 5
Requisitos de Hardware

Servidor de Base de Datos	
Procesador	Intel Core i3, 3.10 GHZ
RAM	8 GB
Disco Duro	500 GB
Sistema Operativo	Windows Server 2016 Standart x64 bits
Servidor de Base de Datos	SQL server 2017
	Visual Studio 2017
	Sybase Central 16.0
Red	Fastethernet, TPC/IP
Video	VGA (1024 x 768)

3.4.3. Definición del Problema

Hace 20 años en la Ciudad de Cayambe nace ROSAPRIMA, con tan solo 3 hectáreas de invernaderos dedicados a la producción y exportación de rosas, con el tiempo ha crecido de manera impresionante; actualmente con 4 fincas y más de 94 hectáreas dedicadas a su labor social, ha permitido generar más de 1000 plazas de empleo. Además, ROSAPRIMA ha logrado ofertar al mercado más de 150 variedades las cuales cumplen con lo los mejores estándares de calidad, permitiéndoles así ser reconocida con los mejores galardones de la industria florícola y cumpliendo con la satisfacción del cliente. Esto se debe por que cuenta con el mejor personal, totalmente especializado en todas sus áreas,

y siempre sus instalaciones cuentan con la mejor tecnología de vanguardia para todas sus labores.

El clima y la producción de las rosas son dos factores importantes para que la empresa siga adelante, por ello dispone de una estación meteorológica; la cual ayuda de una manera poco eficiente para predecir el número de tallos que se producirán ese día; pero si relacionamos datos estructurados, no estructurados y semiestructurados de la estación meteorológica y producción, dentro de un data sistema de minería podremos realizar cálculos estadísticos de producción obteniendo un número más acertados de tallos para venta tanto nacional como internacionalmente

- **¿Qué se está buscando?, ¿Qué tipos relaciones se busca?**
  La empresa está buscando en implementar un servicio de minería de datos, el cual nos proporcione una predicción de producción tomando en cuenta la relación que existe entre las variables de datos de tallos cortados, datos meteorológicos y enfermedades, con el fin de obtener una proyección tanto de flor de exportación como de flor nacional
- **¿Qué está intentando resolver o mejorar dentro de la empresa?**
  Mediante el uso de un algoritmo proporcionado por Microsoft SQL, implementado dentro de un sistema de minería de datos, obtener un reporte de tallos cortados, para la venta tanto internacional como
- **¿Qué predicciones desea obtener a partir de la minería de datos?**
  A través de la minería de datos la empresa busca:
    - Proyección de tallos cortados en el día dependiendo las variaciones climatológicas y la presentación de enfermedades en la planta
    - Proyección de numero de tallos que saldrán a la venta tanto en el mercado nacional como en el internacional
    - Proyección de presencia de enfermedades en la planta con respecto al clima

- **¿Qué resultado o atributo desea predecir?**
  El resultado que se desea obtener es la predicción de numero de tallos que van a ser cortados tomando en cuenta el clima y la presencia de enfermedades.
- **¿Qué tipos de datos tiene?, ¿Están relacionados?,**
  Dentro de la empresa se manejan los datos de producción, enfermedades, y climatológicas. Están se encuentran relacionas

3.4.4. Preparación de Datos

La Base de Datos de toda la empresa es demasiado grande, compuesta de diferentes tipos de datos, tablas y relaciones.

- **Configuración de Conexiones ODBC a las Bases de Datos.**
  - **Configuración Base de Datos Principal de la empresa**

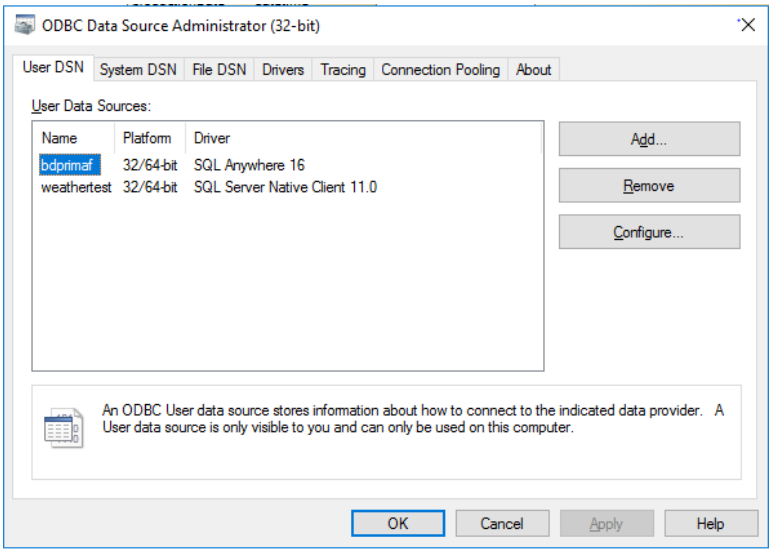


Figura 29. Configuración de Conexiones ODBC a la Base de Rosaprima

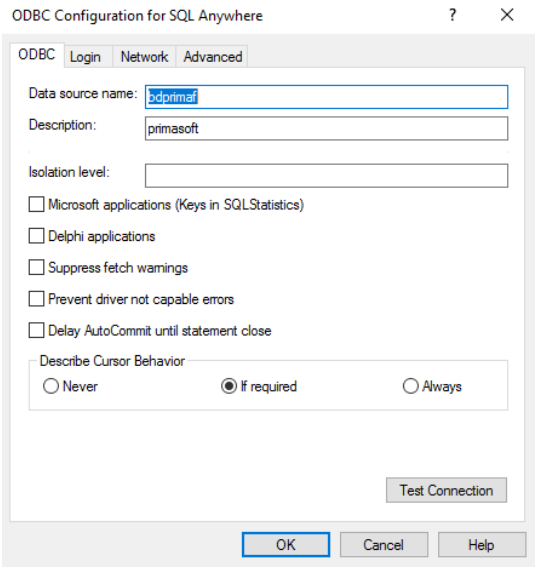


Figura 30. Origen y Descripción de la Base de Datos

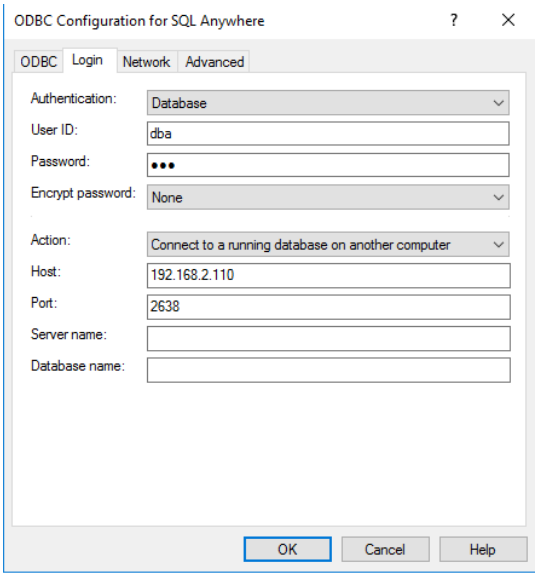


Figura 31. Parámetros de Conexión a la Base de Datos Rosaprima

- **Limpieza de Información**

Con las conexiones ODBC, ya se tienen acceso a las bases de datos tanto a la empresarial como a la local. Cuando el problema fue plantado se debe tener en cuenta los elementos o variables que van a servir para nuestro análisis:

  - Producción
  - Enfermedades
  - Clima

Dentro de la base de datos empresarial se tiene los datos de producción y enfermedades, pero para poder obtener solo los datos realmente necesarios a tomar en cuenta se realizan consultas o querys. Con el fin de obtener las tablas y columnas que conformaran el Data Warehouse.

El gestor de base de datos que la empresa es Sybase Central 16.0 o también conocido como SQL Anywhere. Para ingresar a la necesitamos la dirección o nombre del servidor, puerto de comunicación y credenciales de autenticación.

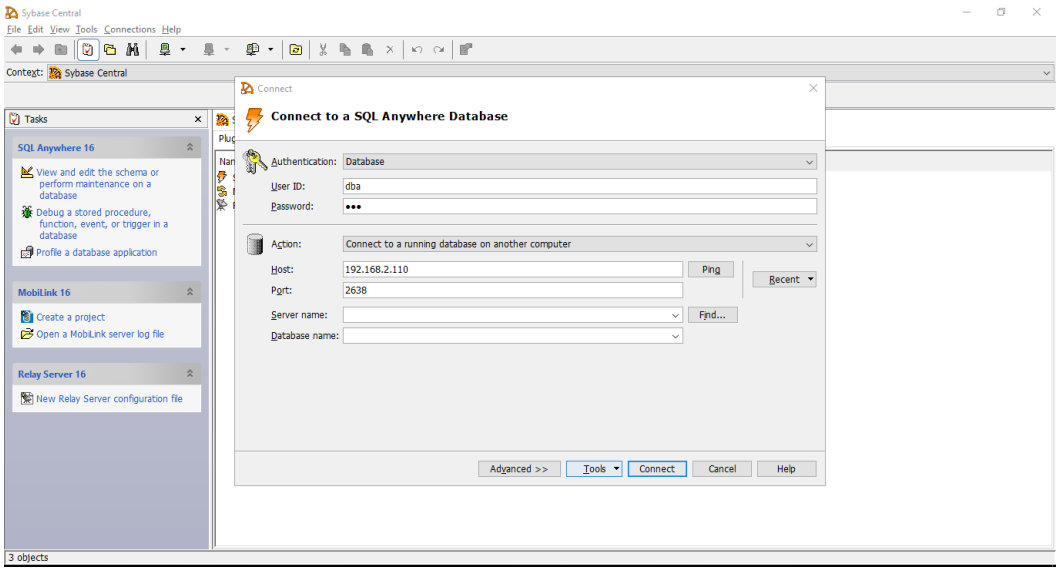


Figura 32. Ingreso a Sybase Central, Base de Datos de Rosaprima

Una vez dentro de la base de datos, ir a la sección de tablas, dar clic derecho y seleccionar la opción VIEW DATA IN INTERACTIVE SQL, se desplegará una ventana en la cual se podrán realizar las consultas. Las cuales nos permitirá realizar la limpieza de datos y solo obtener lo mas importantes.

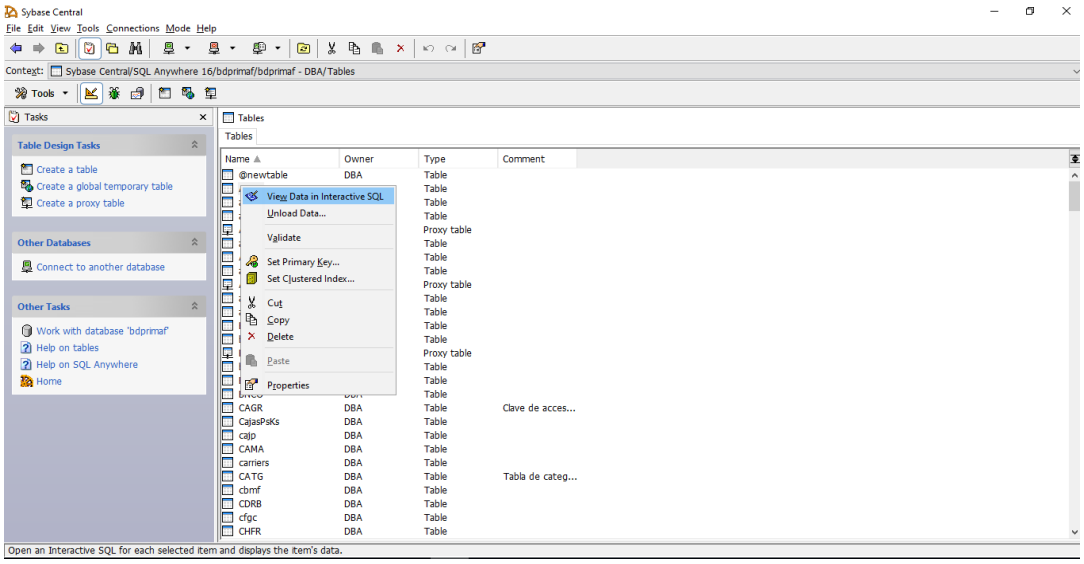


Figura 33. Ingreso al Gestor de Querys

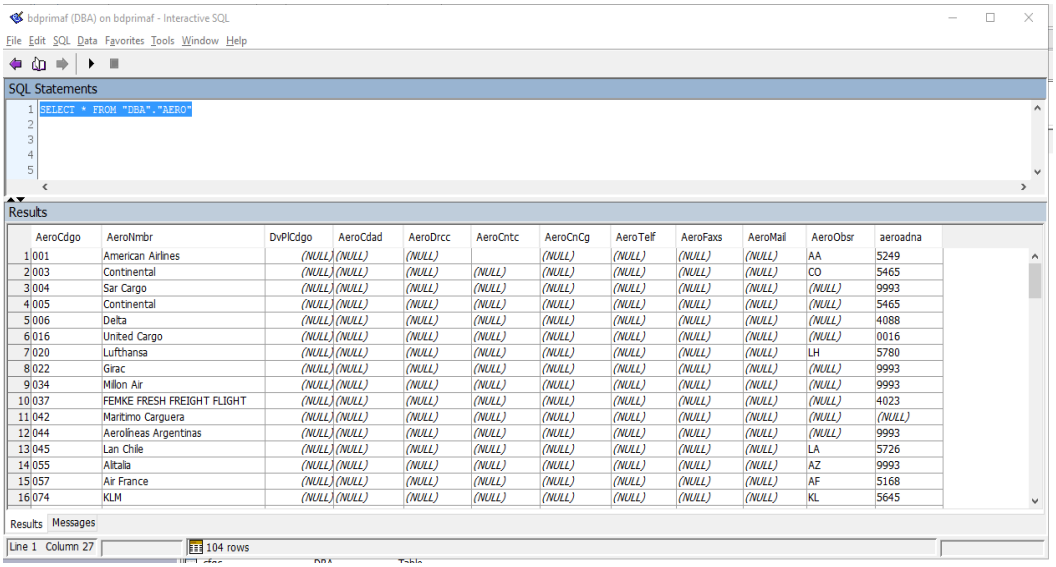


Figura 34. Dashboard para realizar consultas

Antes de realizar las consultas de los datos, requiere realizar un análisis de las variables, tablas y columnas que son requerida para el armado del Data Warehouse

Producción

Id de producción: identificador

Fecha: el día que se realizó el corte.

Finca: Lugar de donde se obtendrán los datos

Nombre: nombre de la variedad

Cantidad: número de tallos cortados ese día

- **Enfermedades**

  id de enfermedad: identificador

  fecha: fecha de registro de enfermedad

  Variedad: a qué tipo de flor afecto

  Nombre: nombre de la enfermedad

  Cantidad: número de tallos afectados

- **Clima**

  Id Clima: identificador

  Clima máximo: punto más alto de calor

  Clima mínimo: punto más bajo de frio

  Promedio: valor promedio entre minio y máximo

  Horas de sol: número de horas que el sol brilla

  Precipitación: humedad

Una vez realizado el análisis vamos a proceder con la limpieza y con la preparación de la información que vamos a necesitar para poder implementar nuestro Data Warehouse es decir mediante el uso de consultas o querys vamos a tomar solo las columnas más importantes para la realización de predicciones.

- **Consulta para producción**

```
select gclt.finca as finca, date(mllsfcha) as fecha, tpfl.tpflnmbr as variedad, sum(mlls.mllsnmro * mlls.mllsfnca) as tallos
from mlls
inner join tpfl
on tpfl.tpflcdgo = mlls.tpflcdgo
inner join gclt
on gclt.gcltcddgo = mlls.gcltcddgo
where date(mllsfcha) = date(today())
group by finca, fecha, variedad;
```

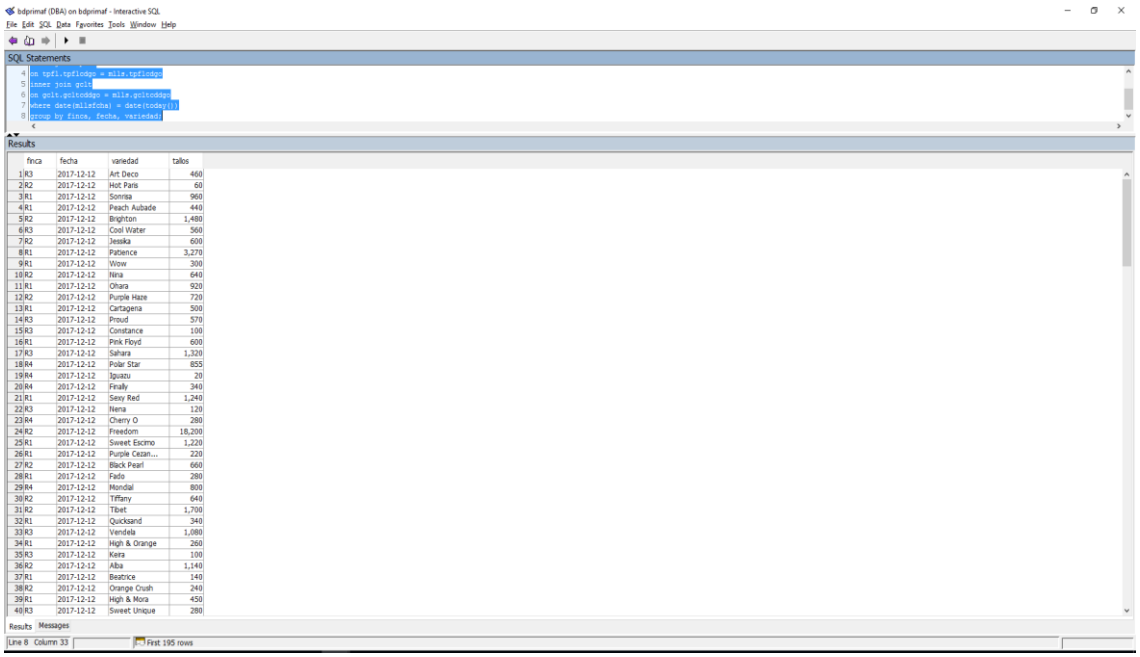


Figura 35. Resultados Consulta Producción

- **Consulta Enfermedades**

```
select bdga.bdgafnca as finca, tpfl.tpflnmbr as variedad ,ocrr.ocrrnmbre as enfermedad, date(rnacfcha) as fecha, sum(rnac.rnacnum) as cantidad
from rnac
JOIN ocrr ON ocrr.ocrrcdgo = rnac.ocrrcdgo
join TPFL on TPFL.tpflcdgo = rnac.tpflcdgo
join BDGA on BDGA.bdgacdgo = rnac.BdgaCdgo
where bdga.bdgafnca = 'R1'
and date(rnac.rnacfcha) = date(today())
group by bdga.bdgafnca, tpfl.tpflnmbr, ocrr.ocrrnmbre, date(rnacfcha)
```

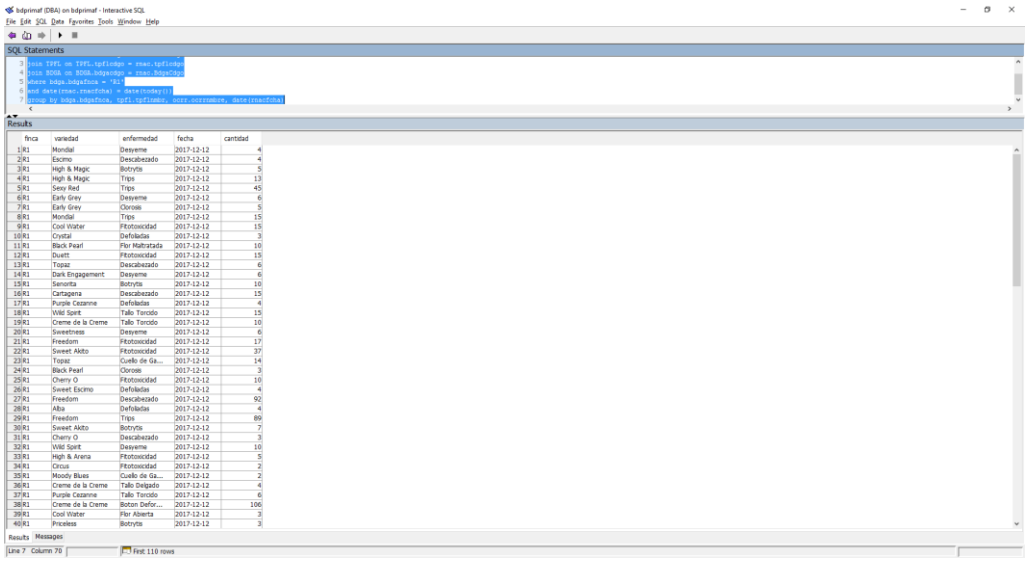


Figura 36. Resultado Consulta enfermedades

- **Creación del Data Warehouse**

En este punto se utilizan a ser los campos que ayudaran a la generación del modelo del DW; de la gran DB que la empresa dispone solo se toma en cuenta o se abstraerán los datos que se obtuvieron de las consultas o querys, que realmente brindaran una pauta al modelo para obtener el conocimiento requerido. Mediante el uso de POWER BUILDER se modelarán las tablas que conformarán el DW como se puede visualizar en la siguiente FIG.36.

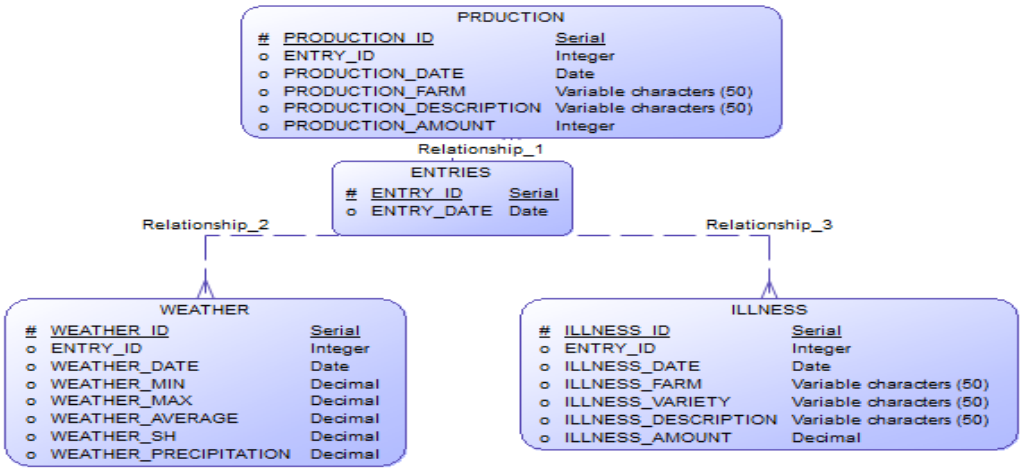


Figura 37. Modelo Físico del Dataware House

Una vez creado el modelo de la base de datos, POWER DESIGNER nos brinda la posibilidad de generar un script, para ejecutarlo como query dentro del SQL SERVER 2017.

Para crear el DW ejecutamos el MICROSOFT SQL SERVER MANAGEMENT STUDIO, ingresar el usuario y el password del administrador de la DB.

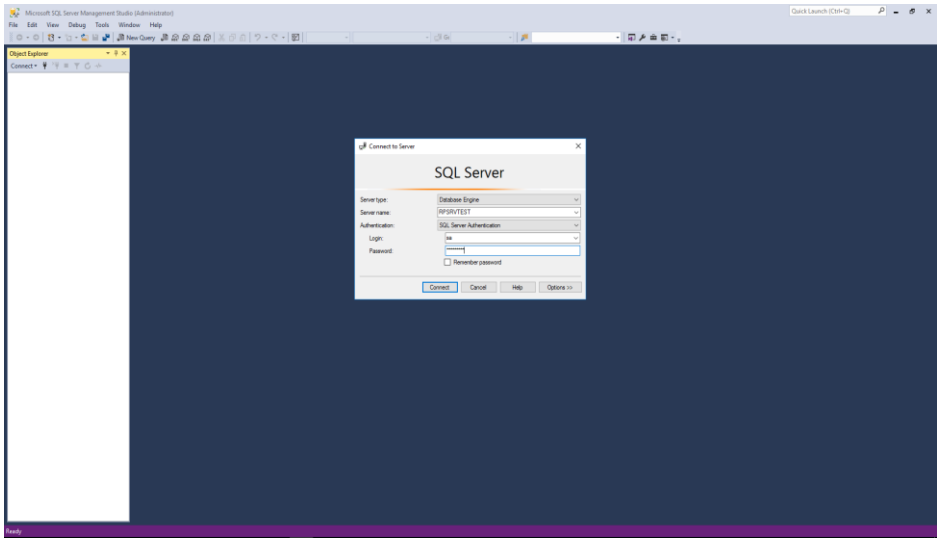


Figura 38. Ingreso al Gestor de Base de Datos

Una vez ingresado dar clic derecho sobre la carpeta Databases y seleccionamos la opción NEW DATABASE. En la ventana que se despliega asignamos un nombre y damos clic en OK

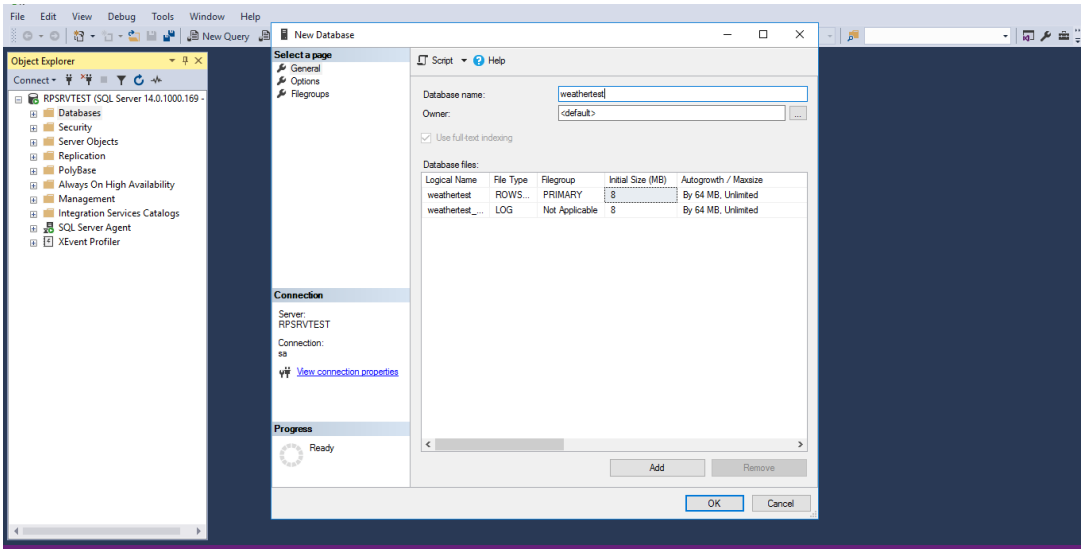


Figura 39. Creación del Dataware House

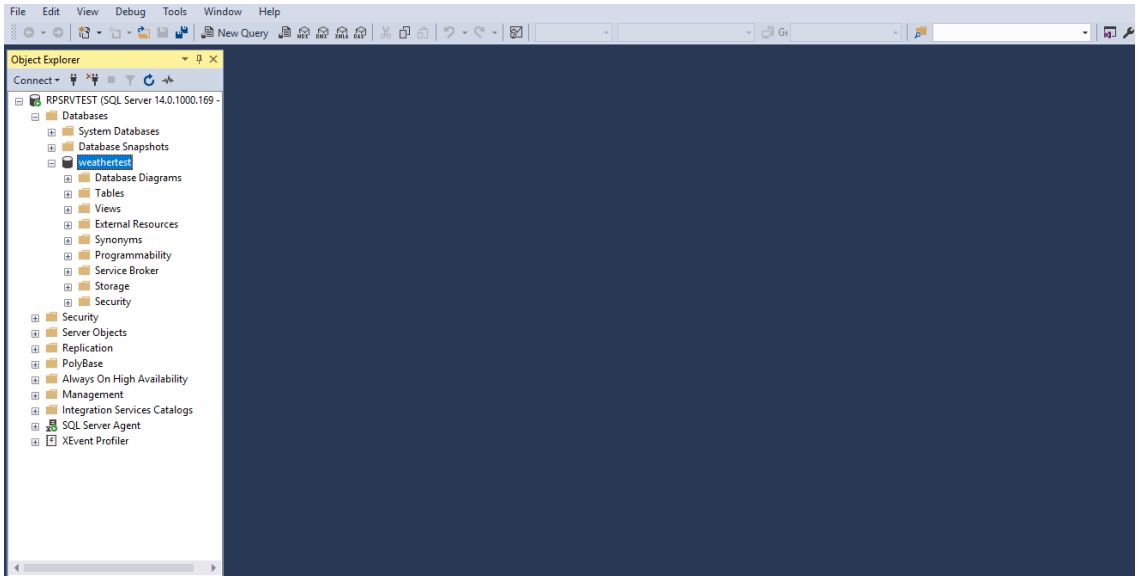


Figura 40, Visualización del Data Warehouse

Para crear las tablas que van a ser utilizados dentro del DW, se utilizara el script que genero el POWER DESIGNER, este puede ser visualizado con el block de nota, se copia todo el texto. A continuación, se muestra el script a utilizar:

```
/*==============================================================
===*/
/* DBMS name:      Microsoft SQL Server 2012                    */
/* Created on:     3/1/2018 11:47:24                            */
/*==============================================================
===*/

if exists (select 1
  from sys.sysreferences r join sys.sysobjects o on (o.id = r.constid and o.type
= 'F')
  where r.fkeyid = object_id('ILLNESS') and o.name =
'FK_ILLNESS_RELATIONS_ENTRIES')
alter table ILLNESS
  drop constraint FK_ILLNESS_RELATIONS_ENTRIES
go
```

```
if exists (select 1
  from sys.sysreferences r join sys.sysobjects o on (o.id = r.constid and o.type
= 'F')
  where r.fkeyid = object_id('PRDUCTION') and o.name =
'FK_PRDUCTIO_RELATIONS_ENTRIES')
alter table PRDUCTION
  drop constraint FK_PRDUCTIO_RELATIONS_ENTRIES
go

if exists (select 1
  from sys.sysreferences r join sys.sysobjects o on (o.id = r.constid and o.type
= 'F')
  where r.fkeyid = object_id('WEATHER') and o.name =
'FK_WEATHER_RELATIONS_ENTRIES')
alter table WEATHER
  drop constraint FK_WEATHER_RELATIONS_ENTRIES
go

if exists (select 1
          from  sysobjects
          where  id = object_id('ENTRIES')
          and   type = 'U')
  drop table ENTRIES
go

if exists (select 1
          from  sysindexes
          where  id    = object_id('ILLNESS')
          and   name  = 'RELATIONSHIP_3_FK'
          and   indid > 0
          and   indid < 255)
  drop index ILLNESS.RELATIONSHIP_3_FK
go
```

```
if exists (select 1
           from  sysobjects
          where  id = object_id('ILLNESS')
           and   type = 'U')
   drop table ILLNESS
go

if exists (select 1
           from  sysindexes
          where  id    = object_id('PRDUCTION')
           and   name  = 'RELATIONSHIP_1_FK'
           and   indid > 0
           and   indid < 255)
   drop index PRDUCTION.RELATIONSHIP_1_FK
go

if exists (select 1
           from  sysobjects
          where  id = object_id('PRDUCTION')
           and   type = 'U')
   drop table PRDUCTION
go

if exists (select 1
           from  sysindexes
          where  id    = object_id('WEATHER')
           and   name  = 'RELATIONSHIP_2_FK'
           and   indid > 0
           and   indid < 255)
   drop index WEATHER.RELATIONSHIP_2_FK
go
```

```
if exists (select 1
           from  sysobjects
          where  id = object_id('WEATHER')
           and   type = 'U')
  drop table WEATHER
go

/*==============================================================
===*/
/* Table: ENTRIES                                               */
/*==============================================================
===*/
create table ENTRIES (
  ENTRY_ID            numeric             identity,
  ENTRY_DATE          datetime            null,
  constraint PK_ENTRIES primary key nonclustered (ENTRY_ID)
)
go

/*==============================================================
===*/
/* Table: ILLNESS                                               */
/*==============================================================
===*/
create table ILLNESS (
  ILLNESS_ID          numeric             identity,
  ENTRY_ID            int                 null,
  ILLNESS_DATE        datetime            null,
  ILLNESS_FARM        varchar(50)         null,
  ILLNESS_VARIETY     varchar(50)         null,
  ILLNESS_DESCRIPTION  varchar(50)        null,
  ILLNESS_AMOUNT      decimal             null,
```

```
   constraint PK_ILLNESS primary key nonclustered (ILLNESS_ID)
)
go

/*==============================================================
===*/
/* Index: RELATIONSHIP_3_FK                                     */
/*==============================================================
===*/
create index RELATIONSHIP_3_FK on ILLNESS (
ENTRY_ID ASC
)
go

/*==============================================================
===*/
/* Table: PRDUCTION                                             */
/*==============================================================
===*/
create table PRDUCTION (
   PRODUCTION_ID        numeric              identity,
   ENTRY_ID             int                  null,
   PRODUCTION_DATE      datetime             null,
   PRODUCTION_FARM      varchar(50)          null,
   PRODUCTION_DESCRIPTION varchar(50)        null,
   PRODUCTION_AMOUNT    int                  null,
   constraint PK_PRDUCTION primary key nonclustered (PRODUCTION_ID)
)
go
```

```
/*==============================================================
==
/* Index: RELATIONSHIP_1_FK                                     */
/*==============================================================
===*/
create index RELATIONSHIP_1_FK on PRDUCTION (
ENTRY_ID ASC
)
go
/*==============================================================
==
/* Table: WEATHER                                               */
/*==============================================================
==
create table WEATHER (
   WEATHER_ID           numeric              identity,
   ENTRY_ID             int                  null,
   WEATHER_DATE         datetime             null,
   WEATHER_MIN          decimal              null,
   WEATHER_MAX          decimal              null,
   WEATHER_AVERAGE      decimal              null,
   WEATHER_SH           decimal              null,
   WEATHER_PRECIPITATION decimal             null,
   constraint PK_WEATHER primary key nonclustered (WEATHER_ID)
)
Go
```

En el MICROSOFT SQL MANAGEMENT STUDIO dar clic derecho sobre la base de datos creada y seleccionar la opción de NEW QUERY y a continuación se desplegará un espacio en blanco en el cual se pegará el texto del script y se lo ejecutará,

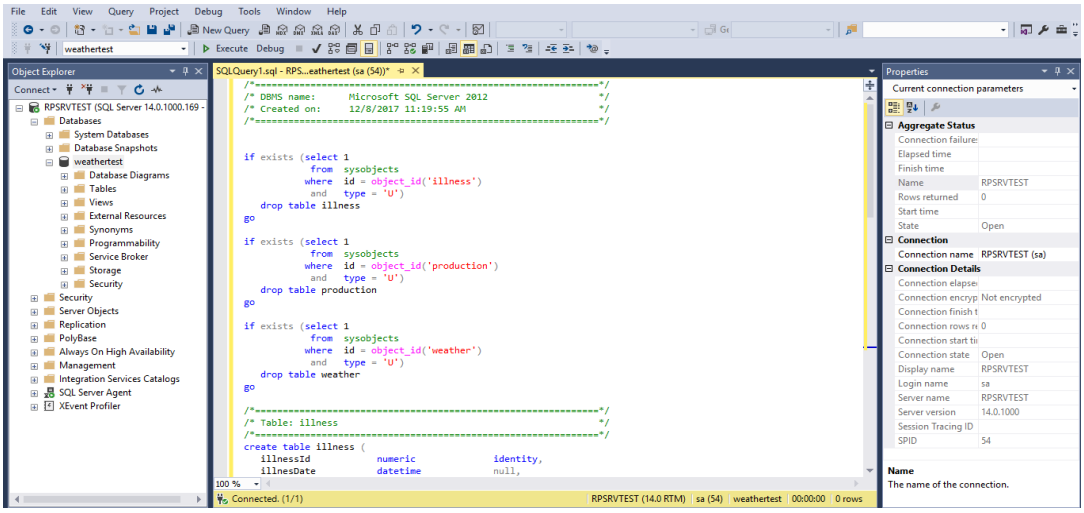


Figura 41. Ejecución de Script del Data WareHouse

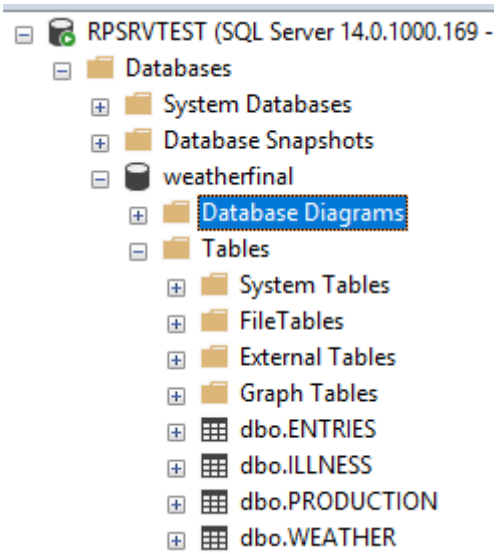


Figura 42. Resultados de la Ejecución de Query

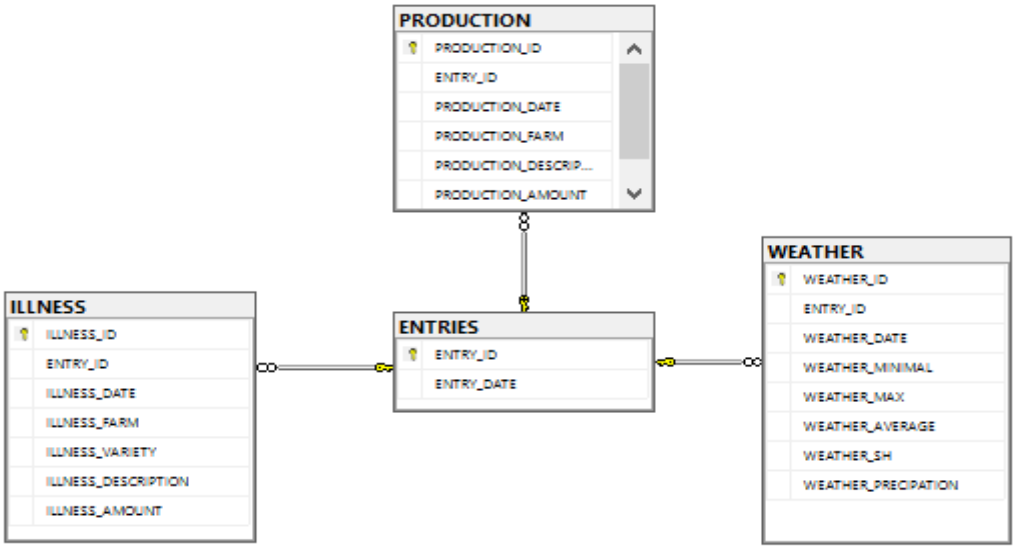


Figura 43. Relación de tablas dentro del Data Warehouse

3.4.5. Exploración de Datos

Dentro de este proceso se realizará la exploración de los datos que van a ser extraídos de las diferentes tablas de las bases de datos y de la estación climatológica de la empresa. Dentro de este procedimiento se realiza el ETL la extracción, la transformación y la carga de los datos provenientes de las diferentes bases hacia el Data Warehouse. Para ello se utilizará.

Para realizar este procedimiento Microsoft tiene herramientas que permite realizar Integrations Services, es decir herramientas que permite manejar los datos de diferentes fuentes y agruparlos en un todo para realizar las diferentes técnicas de minería de datos. En este caso Microsoft dispone SQL Data Tool Server permite integral un numero de base de datos mediante los diferentes tipos de conexiones existentes. Este caso la conexión de la base de datos con ODBC.

Se debe tomar en cuenta que los datos de la Tabla de ENTRIES no puede ser abstraídos. Ya que estos deben ser ingresados manualmente y se consideran una variable dentro del entorno del algoritmo. Los datos que deben ser ingresados en la tabla son de tipo DATE.

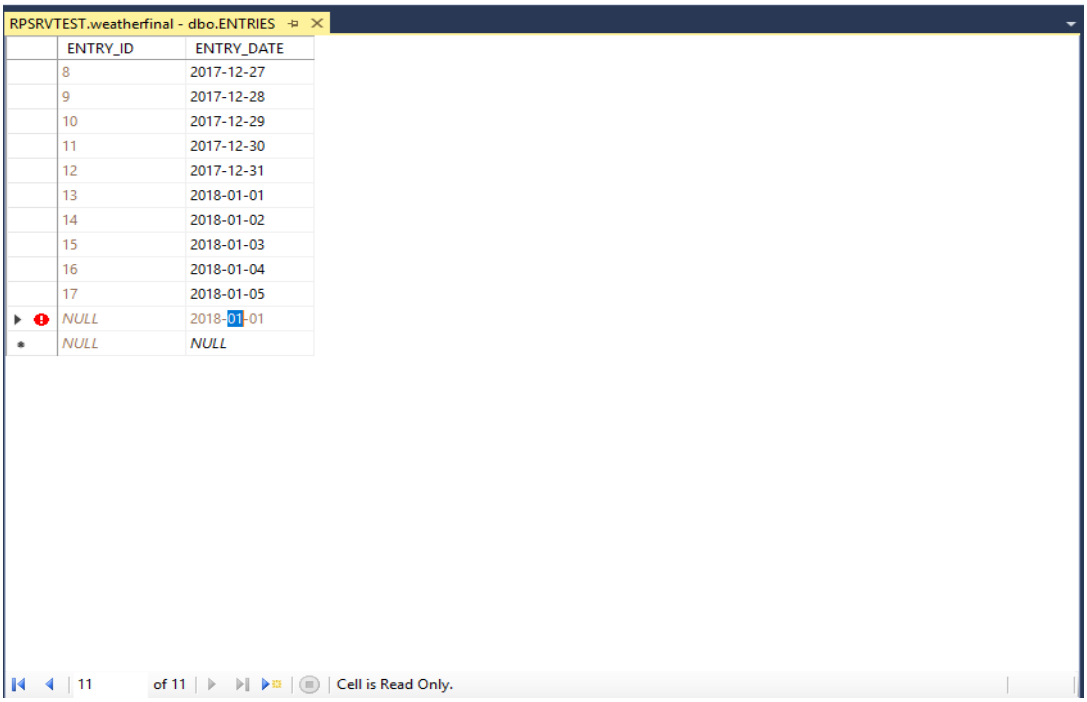
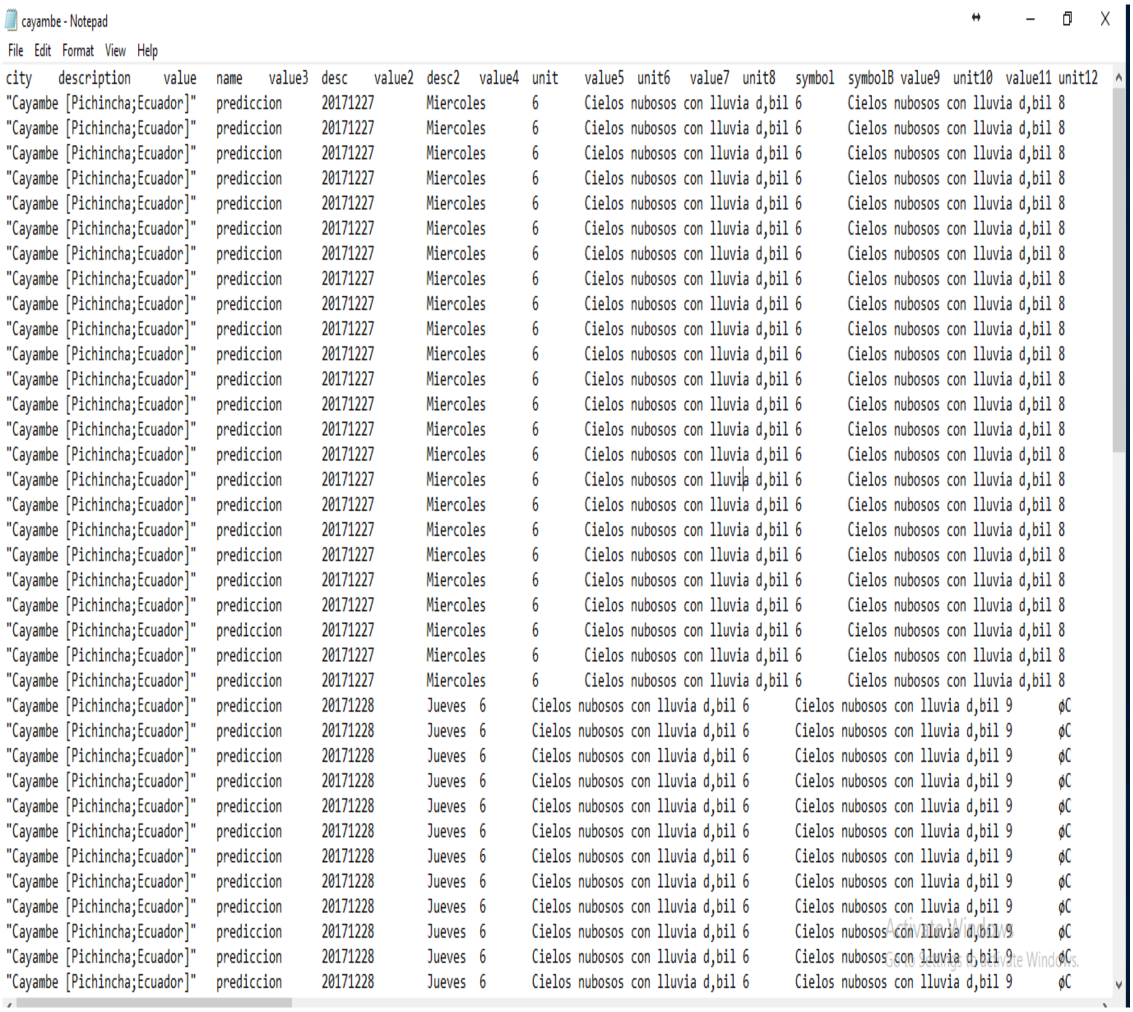


Figura 44. Ingreso de datos en la tabla ENTRIES

Además, los datos que se obtienen de estación meteoróloga proporcionan un documento plano con su propio formato, sin embargo, con un bloc de notas se lo puede visualizar de una manera poco ordenada



```
cayambe - Notepad
File Edit Format View Help
city	description	value	name	value3	desc	value2	desc2	value4	unit	value5	unit6	value7	unit8	symbol	symbolB	value9	unit10	value11	unit12
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171227	Miercoles	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	8
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
"Cayambe [Pichincha;Ecuador]"	prediccion	20171228	Jueves	6	Cielos nubosos con lluvia d,bil	6	Cielos nubosos con lluvia d,bil	9	øC
```

Figura 45. Datos obtenidos de Estación Meteorológica

Se recomienda que los datos proporcionado por la estación climatológica, sean enviados a un archivo de Excel, para obtener una mejor vista de cada uno de columnas que vamos a necesitar y deshacer las columnas que no se va a utilizar

city	description	date	temp min (C°)	temp max (C°)	precipitation (mm)	sun hours	avg
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	13
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	13
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	14
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	15
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	15
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	15
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	14
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	13
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	12
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	11
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171227	6	8	5,4	3	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	11
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	12
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	13
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	13
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	13
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	12
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	11
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171228	6	9	3,3	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	10
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	12
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	16
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	17
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	18
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	18
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	17
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	16
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	15
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	14
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	12
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	11
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	9
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	8
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	8
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	7
Cayambe [Pichincha;Ecuador]	prediccion	20171229	3	6	0	2	6
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	6
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	7
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	10
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	14
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	16
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	15
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	12
Cayambe [Pichincha;Ecuador]	prediccion	20180101	6	5	1,6	1	10
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	10
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	9
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	10
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	16
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	17
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	15
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	11
Cayambe [Pichincha;Ecuador]	prediccion	20180102	5	8	3	1	11

Figura 46. Datos Estación Meteorológica

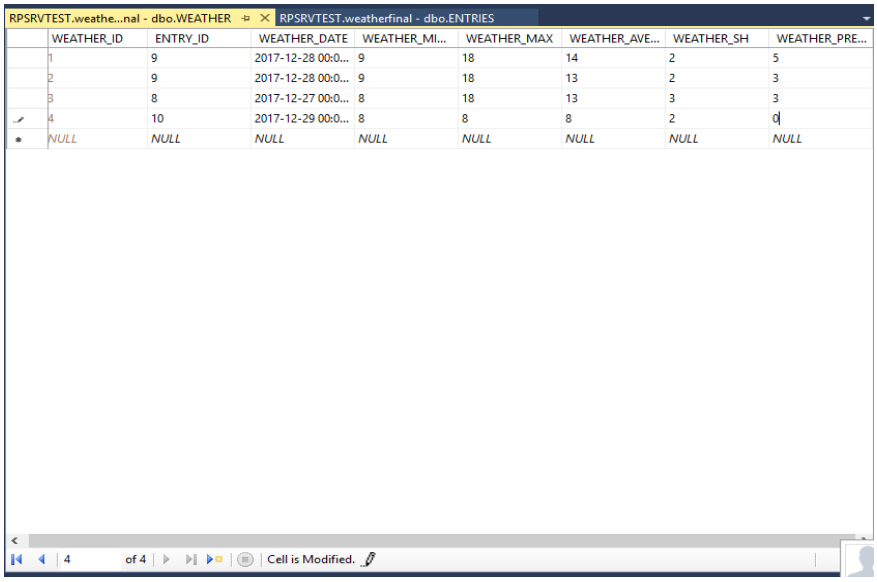


Figura 47. Ingreso de datos a la tabla del Clima

- **ETL de los Datos para la tabla de Producción**

Para realizar la extracción de datos, se debe tener la tabla de ENTRIES poblada de datos, ya que esta tabla es la principal y se encuentra relacionada con las demás. Como primer paso es poder abstraer el campo ENTRY_ID hacia la tabla de producción.

1. Abrir SQL SERVER DATA TOOL, en archivo ir a NEW, PROJECT, se despliega una ventana, y seleccionar INTEGRATION SERVICES PROJECT, y asignar un nombre y dar clic en OK

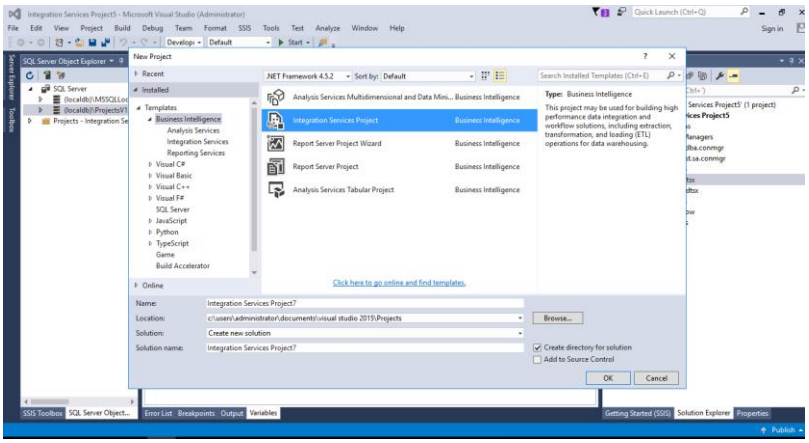


Figura 48. Creación de Proyecto de Integration Services para producción

2. En el lado izquierdo, en el SOLUTION EXPLORER, en la carpeta CONNECTION MANAGERS, dar clic derecho en NEW CONNECTION MANAGER, se desplegará una ventana mostrando el tipo de conexiones en este caso ODBC.

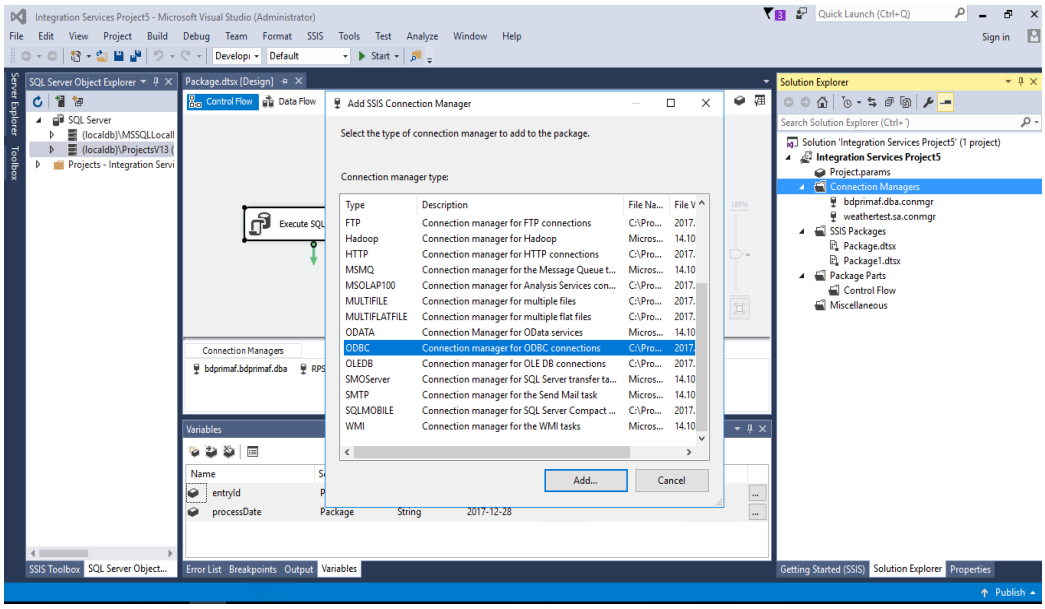


Figura 49. Tipo de Conexiones a la Base de Datos para producción

Se desplegará una ventana que muestra las conexiones existentes, para crear dar clic en NEW y en la casilla USE USER OR SYSTEM DATA SOURCE NAME nos mostrara las bases con conexiones ODBC que podemos utilizas, ingresar es USER NAME y el PASSWOR.

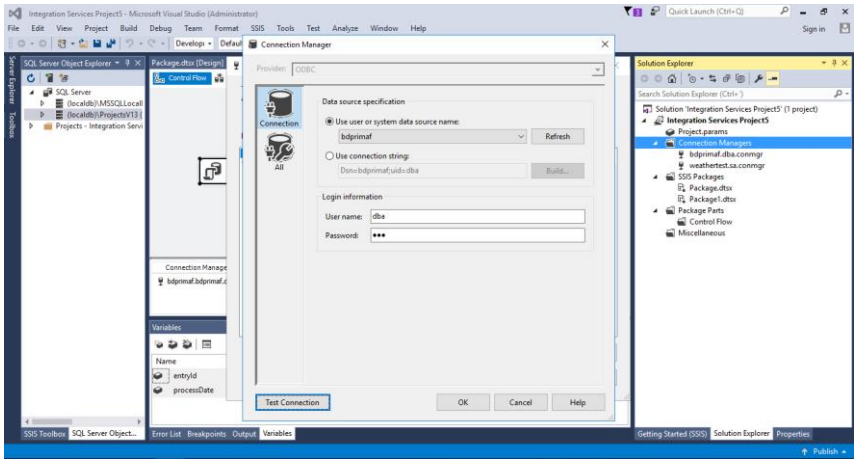


Figura 50. Configuración Conexión a la Base de Datos para producción

3. Para verificar la conexión realizar un TEST CONNECTION, dando clic en el botón

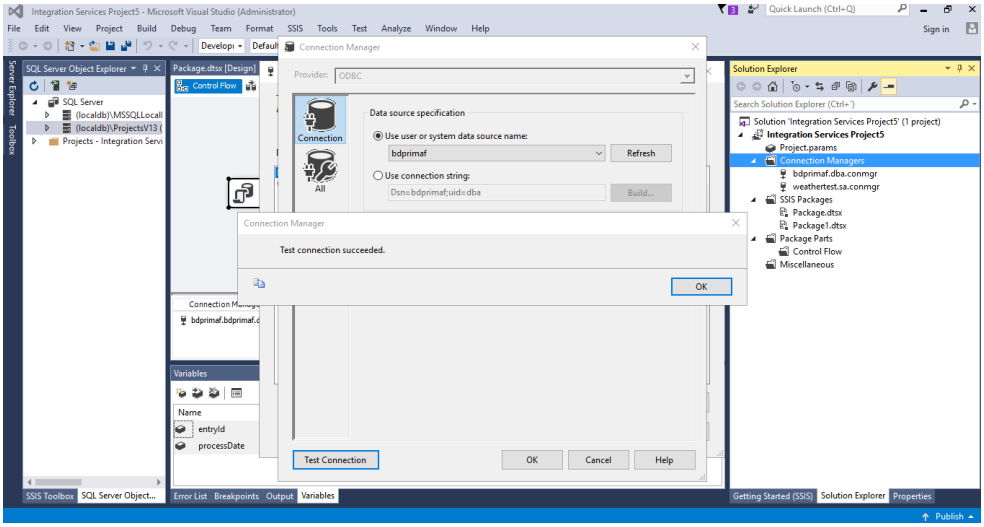


Figura 51. Test de Conexión a la Base de Datos para producción

Este procedimiento se realiza para las dos bases de datos, la principal de la empresa BDPRIMAF de donde se extraer la información y la WEATHERFINAL, la cual almacenara los datos.

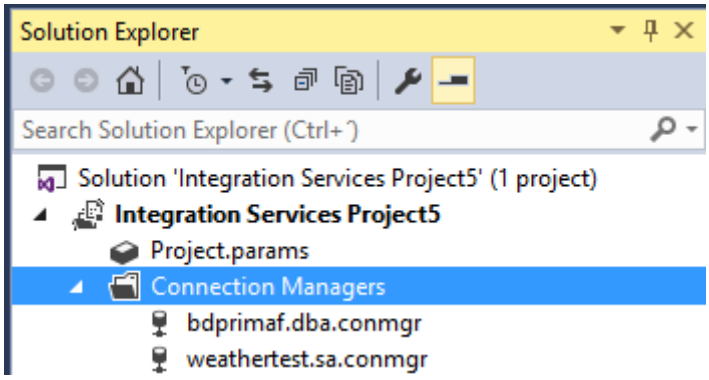


Figura 52. Conexiones a las Bases de Datos para producción

4. Dentro del SOLUTION EXPLORER, en la carpeta SSIS PACKAGES, dar clic derecho y seleccionar NEW SSIS PACKAGE, se creará un paquete y mostrara un DASHBOARD en blanco y el TOOLBOX, con los elementos o herramientas los cuales permitirán realizar el proceso ETL

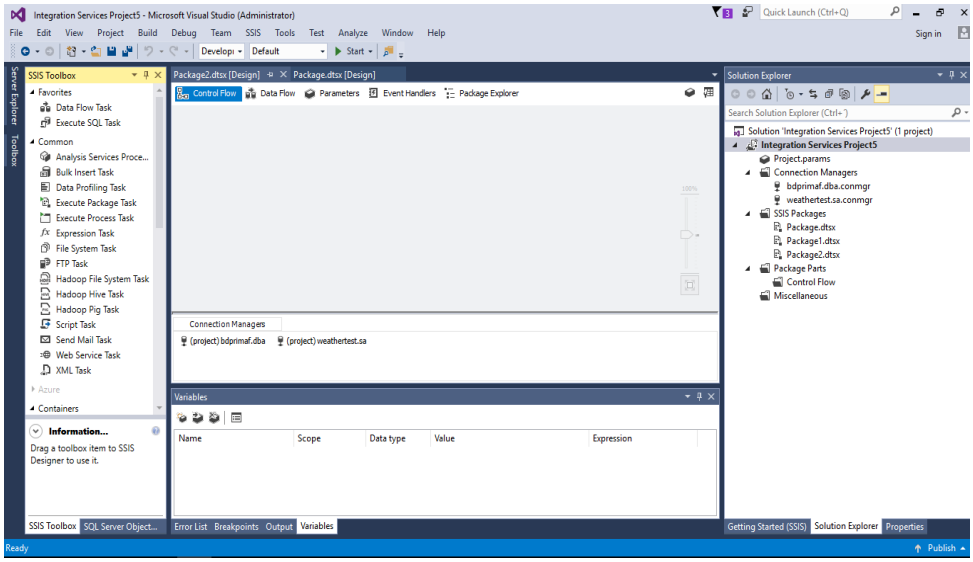


Figura 53. Dashboard de un Package para producción

5. Dentro del Toolbox, vamos a encontrar dos elementos los cuales no permitirá extraer la información de la tabla ENTRIES y la base de datos de la empresa

EXECUTE SQL TASK: conocido también como procedimiento almacenado desde un paquete, dicha tarea puede estar conformada por una única o varias declaraciones, variables o parámetros de SQL, que se ejecutan secuencialmente con el fin de modificar datos, tablas e incluso vistas.

En este caso el EXCUTE SQL TASK requiere de dos, una la cual se extraerá la información, y la otra que se encargara de procesar y de convertirla, para asignar en el campo correspondiente en las diferentes tablas.

Name	Scope	Data type	Value	Expression
entryId	Package	Int32	0	
processDate	Package	String	2017-12-28	

Figura 54. Declaración de Variables para Execute SQL TASK para producción

Una vez declarada las variables, se procede a configurar las propiedades generales del task como se puede visualizar en la FIG 55.

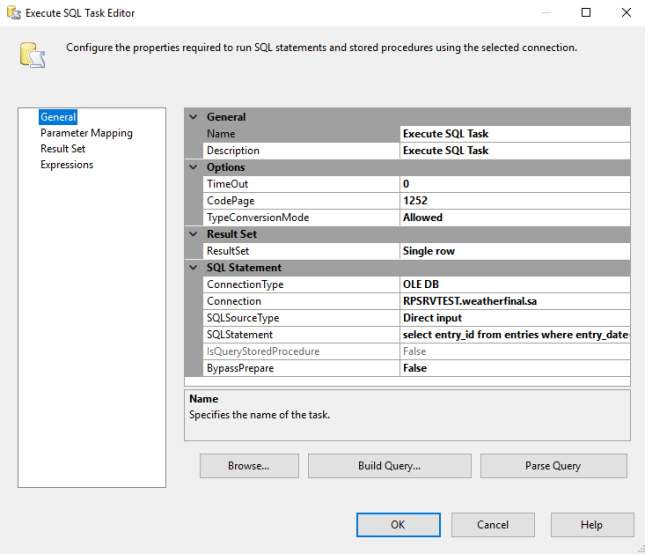


Figura 55. Configuraciones de Generales del Excute SQL Task para producción

Dentro de la configuración del task se debe ejecutar un SQL query el cual nos permitirá extraer los valores de la tabla ENTRIES para poder asignarlo en la relación con las otras tablas. El query es el siguiente:

select entry_id from entries where entry_date=?

Se le asigna un valor ? a la variable para que esta pueda ser llamada a las otras tablas relacionas con la tabla principal, ya que en las configuraciones de PARAMETER MAPPING al valor se le asigna la variable que se le asigno, la dirección que tendrá el tipo de dato, se configura lo valores:

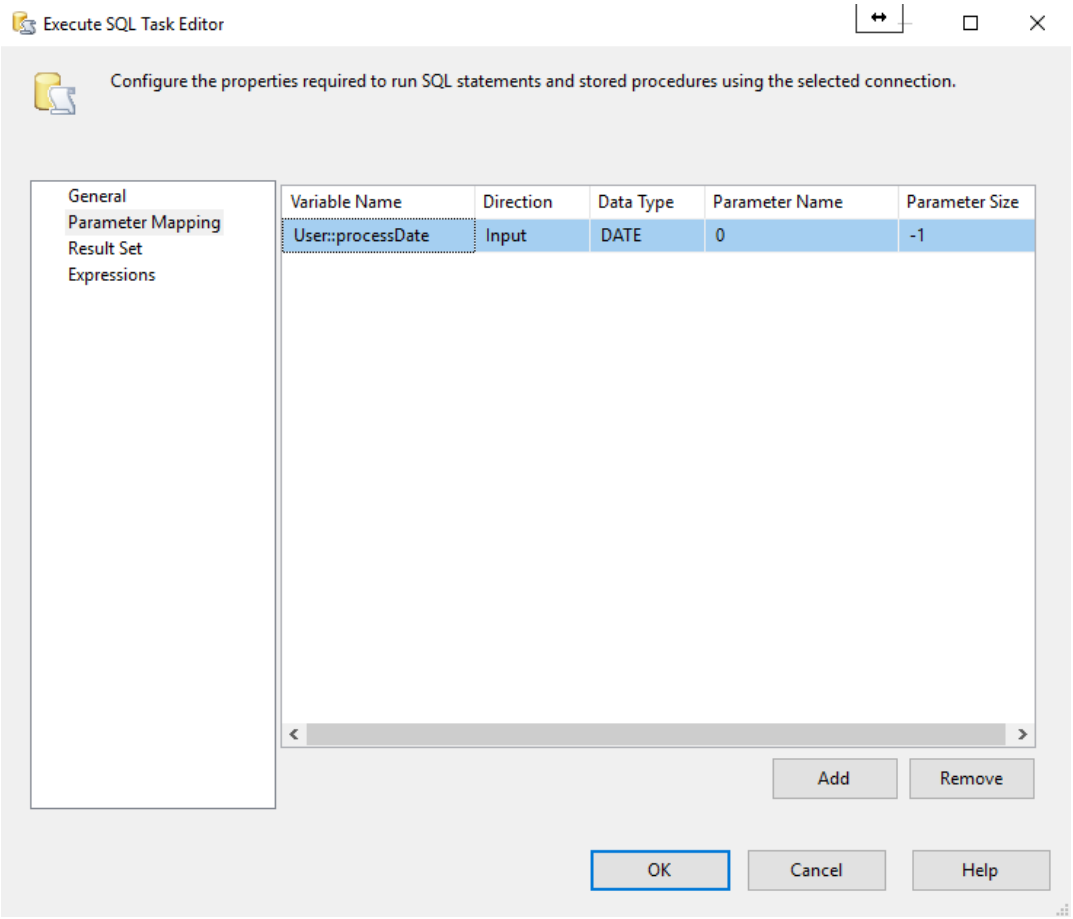


Figura 56. Configuración PARAMETER MAPPING para producción

Dentro del Result Set, se configura el valor que deseamos recibir y la variable que la contiene

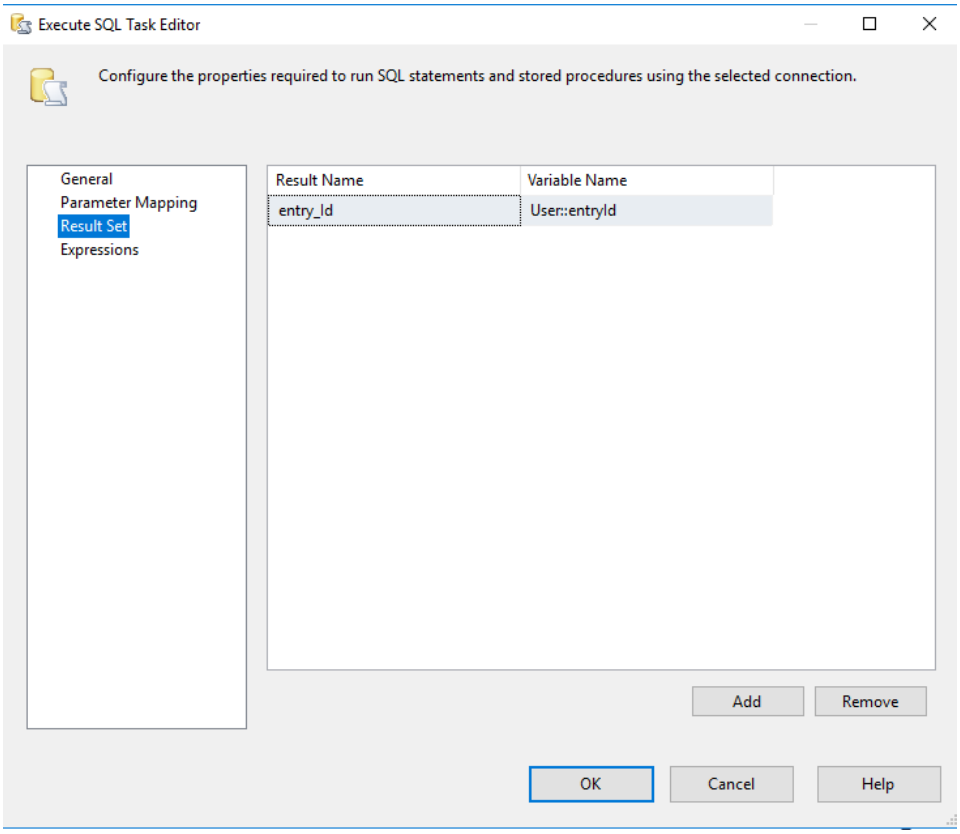


Figura 57. Configuración de Resulta Set para producción

Data Flow task: dentro de este se encapsula las tareas de extracción transformación y carga de los datos.

Dentro del caso el Data Flow task se compone:

OLE DB SOURCE: abarca el CONNECTION MANGER, en este caso será la base de origen BDPRIMAF con sus respectivas credenciales de acceso, el query utilizado para tener los datos de la base de datos de origen, y una vez ejecutado el mismo mostrara las columnas que lo conforman.

En la preparación de los datos ya se utilizó los querys o consultas para tener la información que se necesitaba para nuestro algoritmo de minería de datos, este es el mismo solo que se le agrega el parámetro ? el cual permite abstraer realizar el llamado a los datos de la tabla principal ENTRIES

select ? as entry_id, gclt.finca as finca, date(mllsfcha) as fecha, tpfl.tpflnmbr as variedad, sum(mlls.mllsnmro * mlls.mllsfnca) as tallos

```
from mlls
inner join tpfl
on tpfl.tpflcdgo = mlls.tpflcdgo
inner join gclt
on gclt.gcltcddgo = mlls.gcltcddgo
where date(mllsfcha) = date(?)
group by finca, fecha, variedad;
```

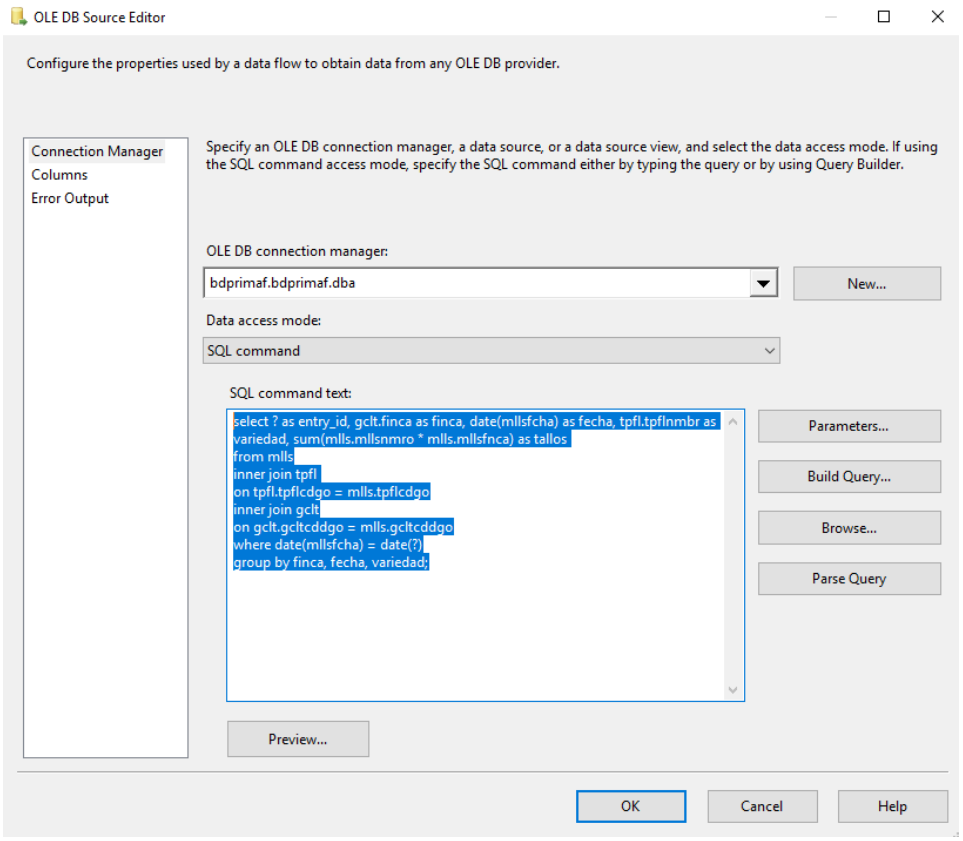


Figura 58. Configuración de OLE DB Source para producción

Cuando el query ejecutado, se encuentra ejecutado sin fallas, en la opción de COLUMS muestra las columnas de la base de datos de origen incluyendo el campo que se encuentra relacionado con la tabla de ENTRIES.

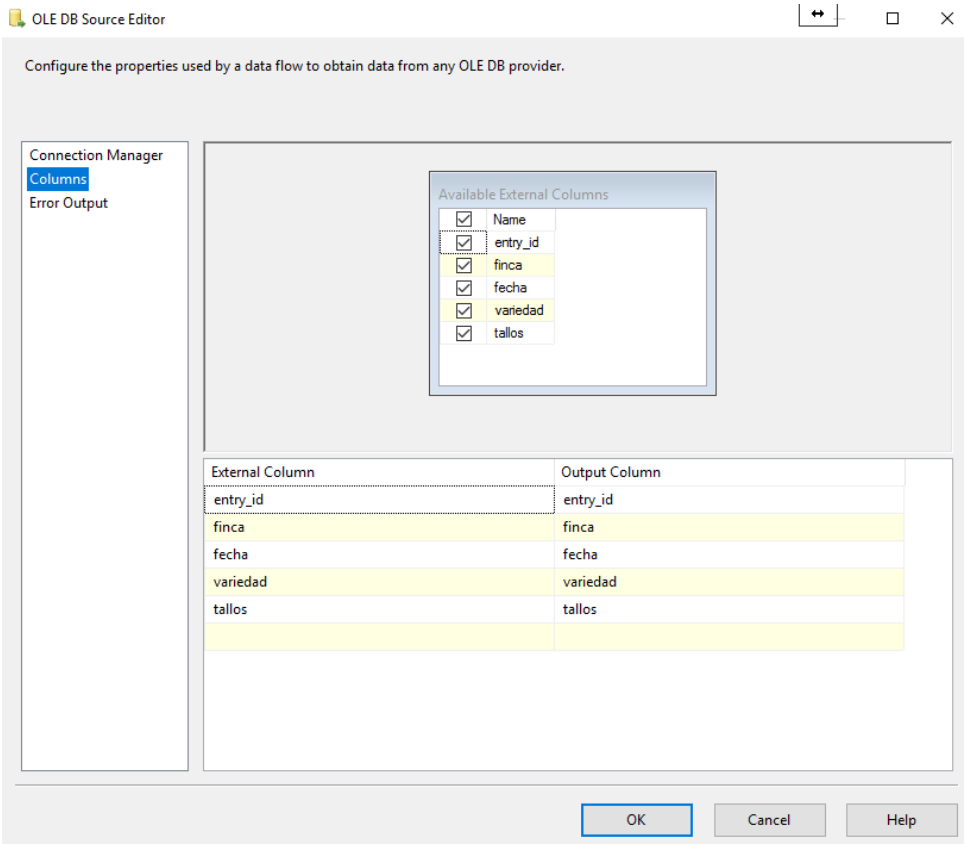


Figura 59. Muestra de Columnas para producción

DATA CONVERSION: generalmente este elemento nos permite realizar cambios en los tipos de datos ya que los diferentes gestores de bases de datos en ocasiones no suelen compatibles. En este caso no se realiza ningún cambio ya que las conexiones ODBC y OLE son compatibles entre ellas. Además, SQL y Sybase Central son dos gestores de base de datos que utilizan los mismos tipos de datos.

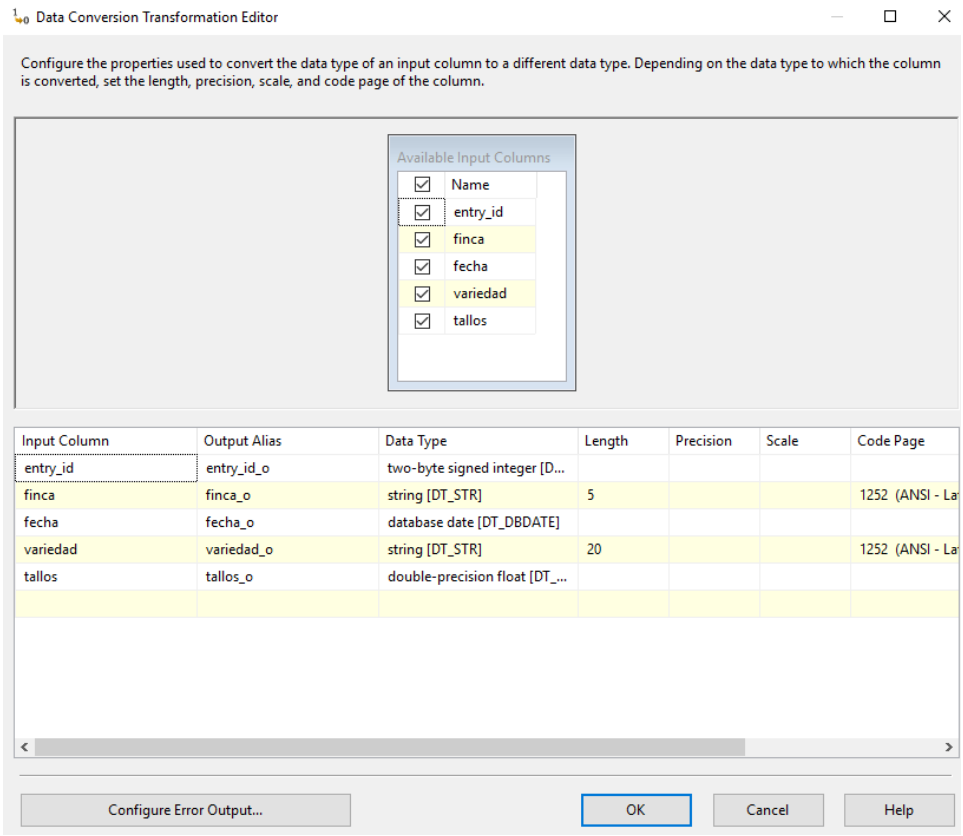


Figura 60. Data Conversion para producción

OLE DB Destination: dentro de elemento se debe conectar a la base de datos de destino WEATHERFINAL, con sus respectivas credenciales de acceso, en DATA ACCESS MODE como primera y más recomendable la opción TABLE o VIEW y selecciona la tabla a la cual nosotros vamos a cargar los datos en este caso sería a la tabla de producción, este procedimiento también se lo puede realizar mediante query, pero con llevaría un gran tiempo.

Como paso final realizar un test de conexión a la base de datos y a la tabla.

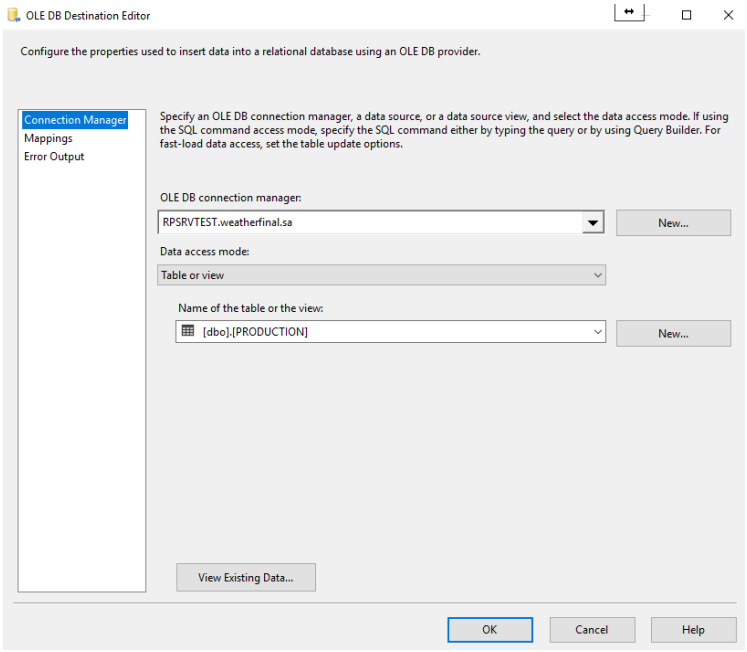


Figura 61. OLE DB Destination para producción

En la opción de Mappings se van a unir los campos de la tabla de la base de origen con los del destino arrastrando cada uno al perteneciente.

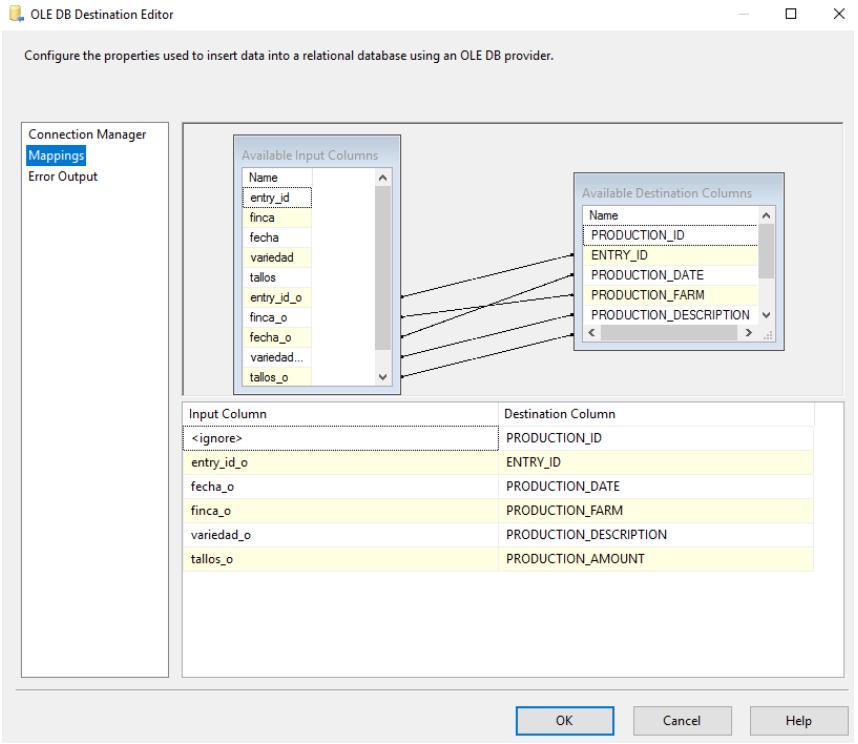


Figura 62. Mapping de Campos para producción

6. Una vez terminado, esto se realiza un deploy de Data Flow tas y sus elementos, en caso de realizarse exitosamente este se mostrará con vistos verdes, después se ejecuta el deploy de package completo y se mostrará de misma manera, por último, vamos al gestor de la base de datos y verificamos que los datos estén cargados.

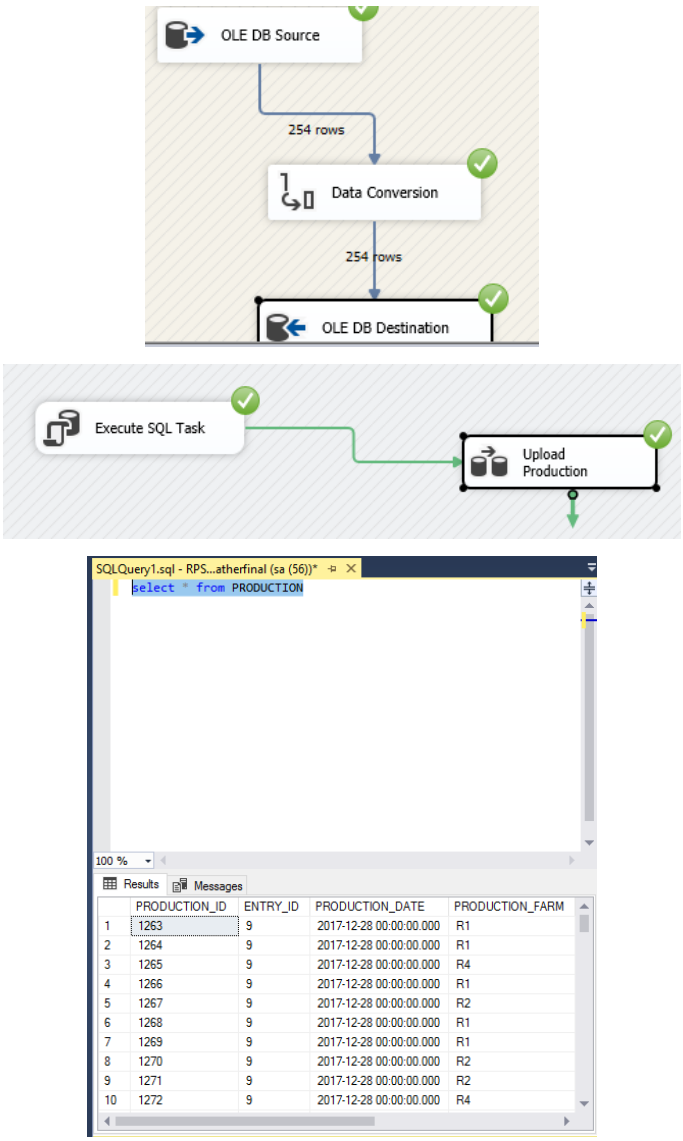


Figura 63. Deploy para producción ETL completo

El proceso anterior se debe realizar también en la tabla en la tabla de enfermedades tomando en cuenta que en el OLE DB source se debe caminar el SQL query utilizado, y la tabla destino.

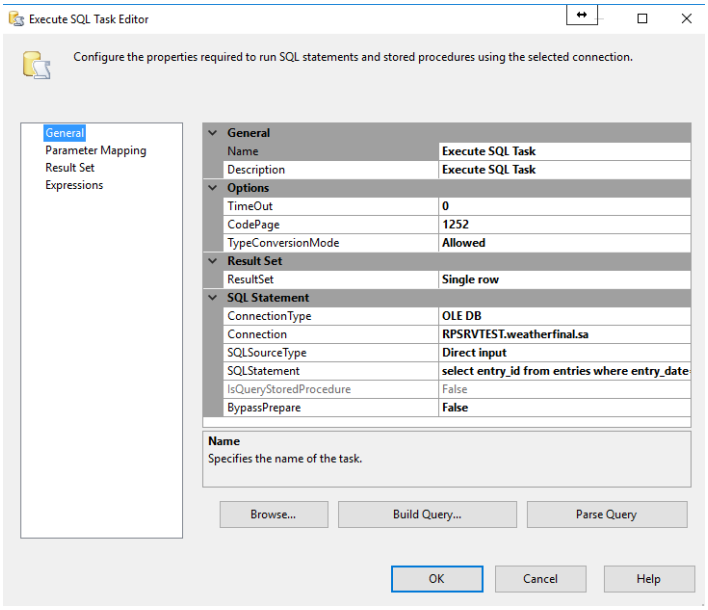


Figura 64. Configuraciones de Generales del Excute SQL Task para enfermedades

select ? as entry_id, bdga.bdgafnca as finca, tpfl.tpflnmbr as variedad
,ocrr.ocrrnmbre as enfermedad, date(rnacfcha) as fecha, sum(rnac.rnacnum) as cantidad from rnac
JOIN ocrr ON ocrr.ocrrcdgo = rnac.ocrrcdgo
join TPFL on TPFL.tpflcdgo = rnac.tpflcdgo
join BDGA on BDGA.bdgacdgo = rnac.BdgaCdgo
where bdga.bdgafnca = 'R1'
and date(rnac.rnacfcha) = date(?)
group by bdga.bdgafnca, tpfl.tpflnmbr, ocrr.ocrrnmbre, date(rnacfcha)

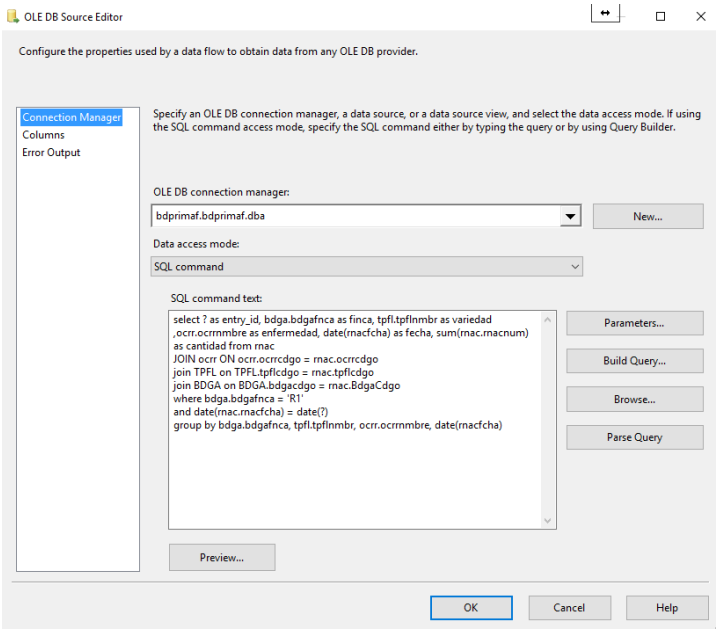


Figura 65. Configuración de OLE DB Source para enfermedades

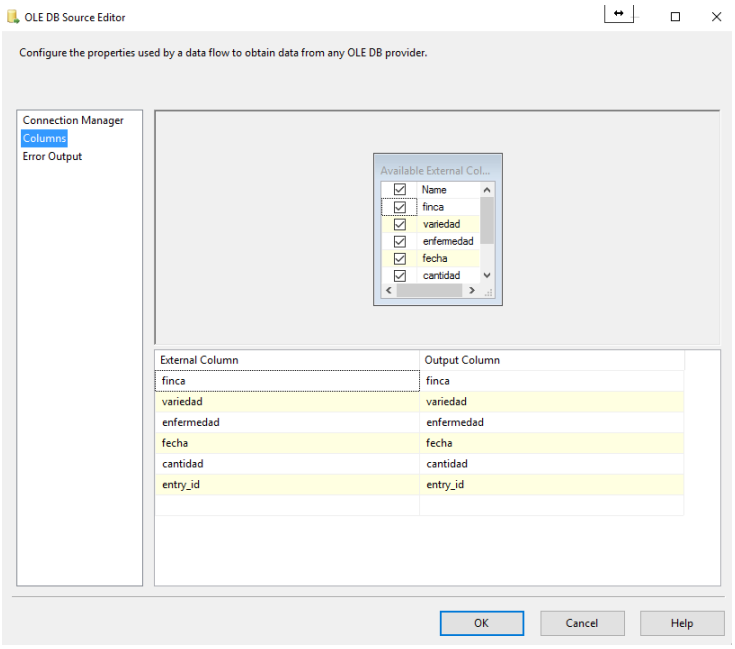


Figura 66. Columns para enfermedades

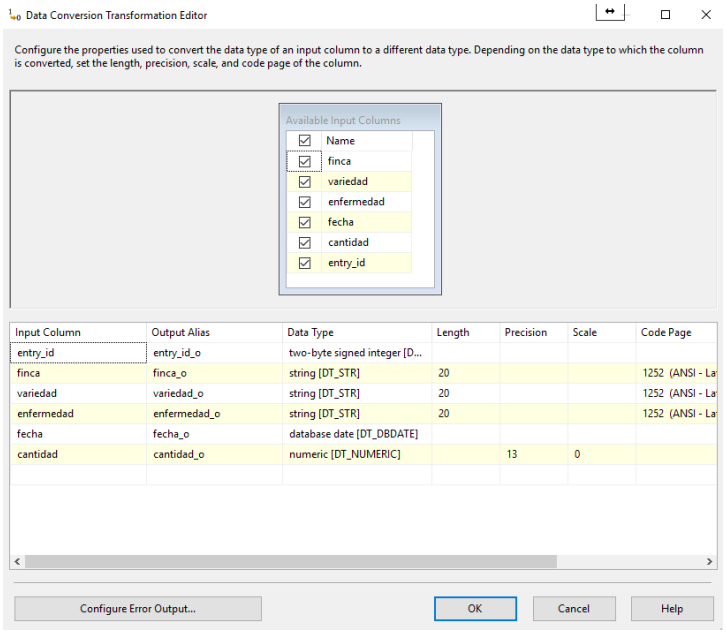


Figura 67. Data Conversion para enfermedades

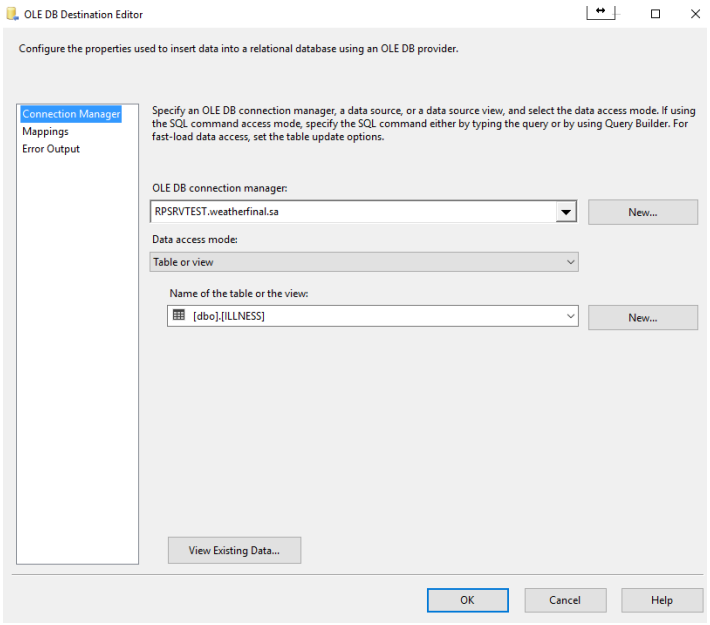


Figura 68. OLE DB Destination para enfermedades

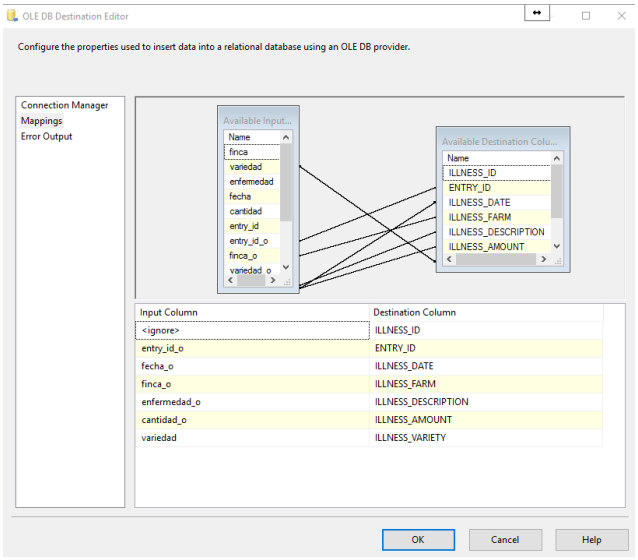


Figura 69. Mapping de Campos para enfermedades

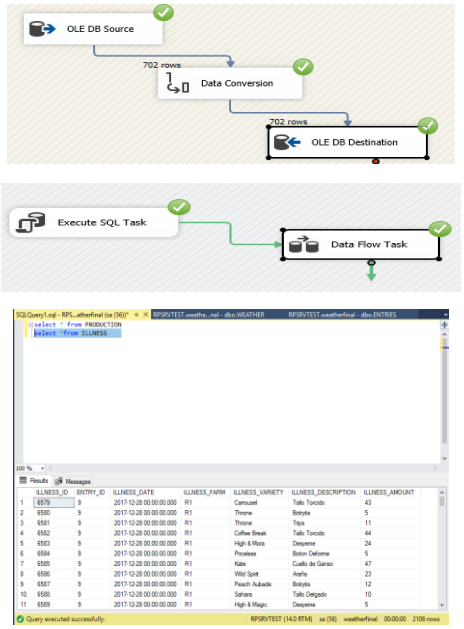


Figura 70. Deploy para producción ETL completo

3.4.6. Generación de Modelos de Minería de Datos

Para crear un proyecto de minería de datos de datos, se seguirá utilizando el SQL SERVER DATA TOOL, como primer paso iremos a File, clic en new Project, se desplegará la ventana, seleccionaremos ANALYSIS SERVICES MULTIDIMENTIONAL AND DATA MINING BUSINESS INTELLIGENCE se asignará un nombre y por último dar clic en OK

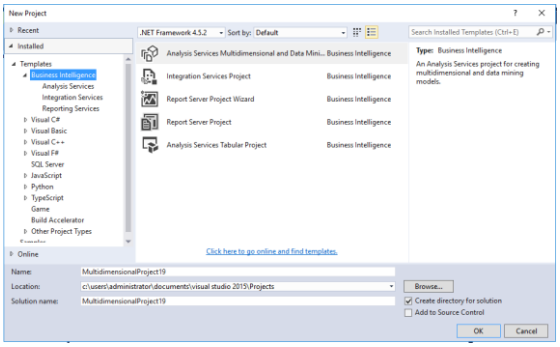


Figura 71. Creación proyecto para minería de datos

El siguiente paso es crear la conexión a la base de datos, así que en la venta del Explorer solution, dar clic derecho sobre la carpeta data source se abrirá un asistente, dar clic en siguiente, y mostrara dos opciones la una para crear un nueva y otra para usar una existente, en este caso se escoge la segunda ya que anterior mente se creó las conexiones a las bases de datos

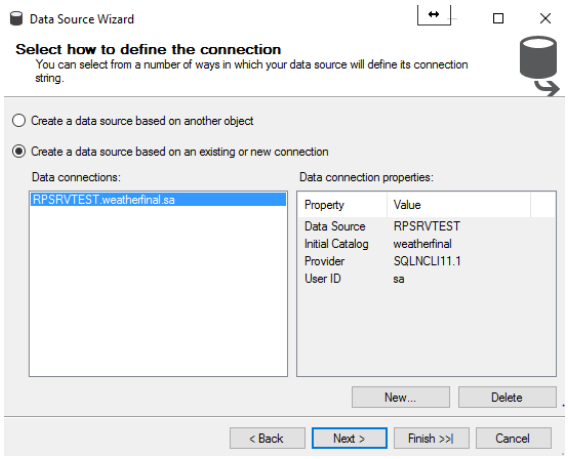


Figura 72. Creación de Conexión a la base de datos

A continuación, se debe asignar un user y un password para ingresar a la base de datos este puede ser asignado manualmente o también se lo puede heredar de las conexiones anteriores, las dos opciones son validas

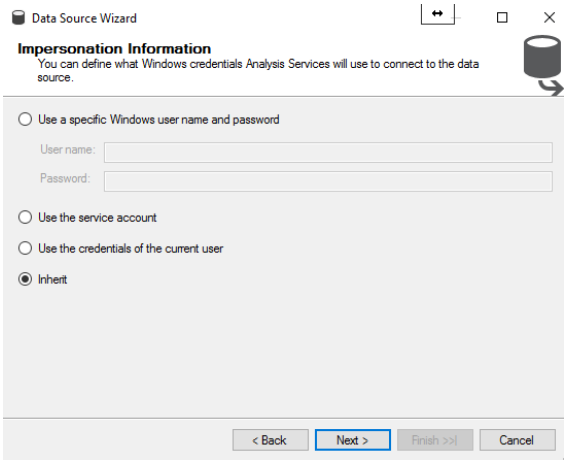


Figura 73. Configuración de Acceso a Origen de la base de datos

A continuación, se debe crear, la vista de las tablas, dar en solution explorer dar clic derecho sobre Datawiew Source Views, con asistente asignamos la base de origen y seleccionamos las tablas que deseamos analizar

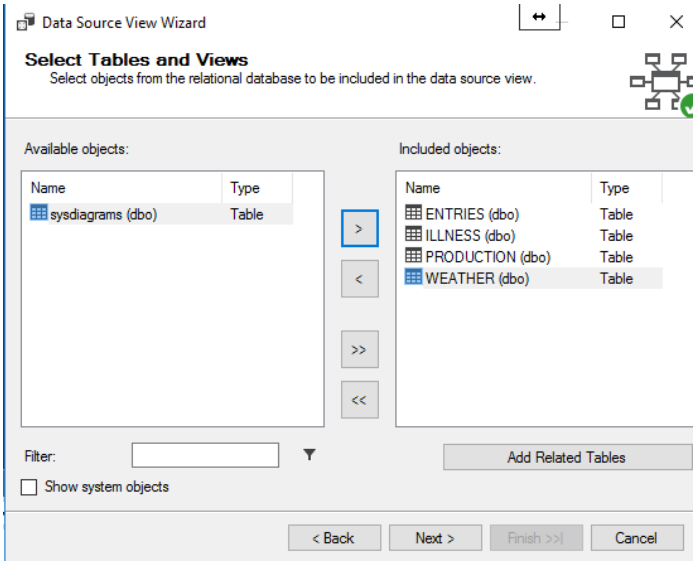


Figura 74. Selección de tablas

Después aparecerá, una ventana mostrando la relación del DW.

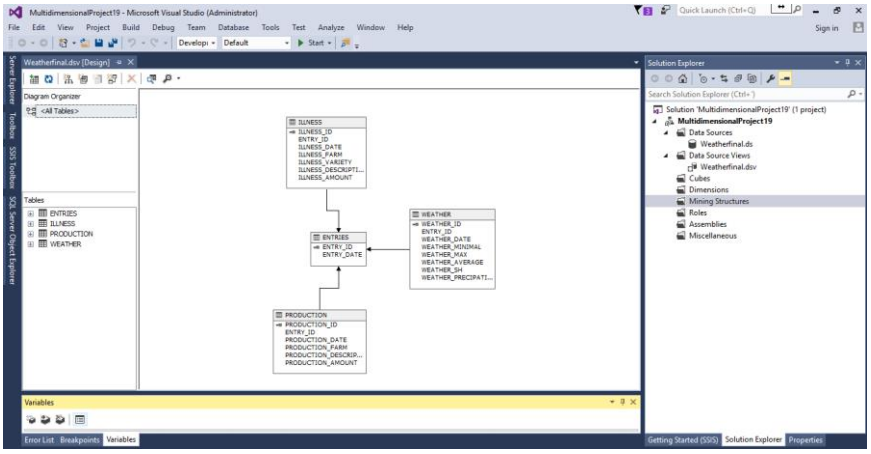
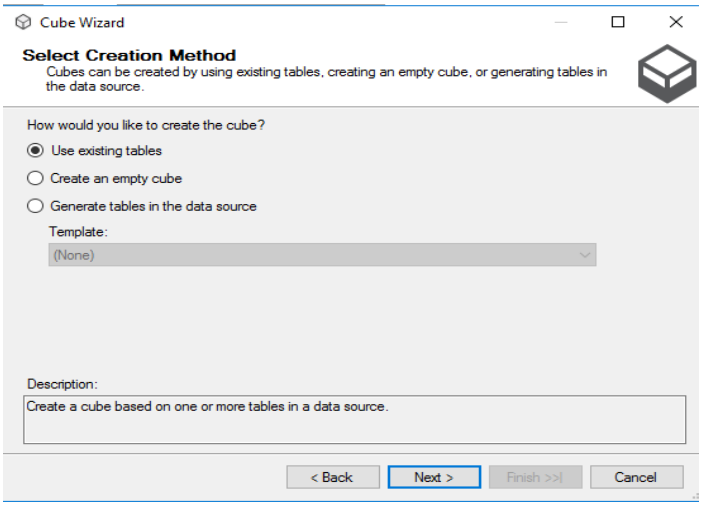


Figura 75. **Vista de Relación del DW**

La mejor manera de trabajar con minería de datos en Microsoft es utilizar cubos OLAP con el fin de tener dimensiones y medidas de la información que es abstraída de diferentes. Para crear un cubo dar clic en CUBES, seleccionar la opción de NEW CUBE, y seleccionar la opción de crear uno con tablas existentes



Fiigura 76. Creación de un cubo con tablas existentes

Después se procederá a mostrar el cuadro de la conexión a la base de datos, y muestra las tablas disponibles, se procederá a seleccionar las tablas que serán parte del grupo de las medidas del mismo.

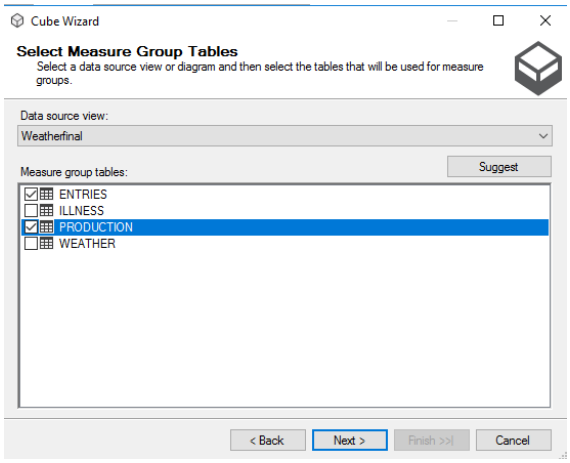


Figura 77. Selección de Tablas que conforman el cubo

A continuación, el asistente mostrara cuales van a ser las medidas del cubo, es decir cuáles van a ser las tablas de donde se analizará la información y realizara la predicción

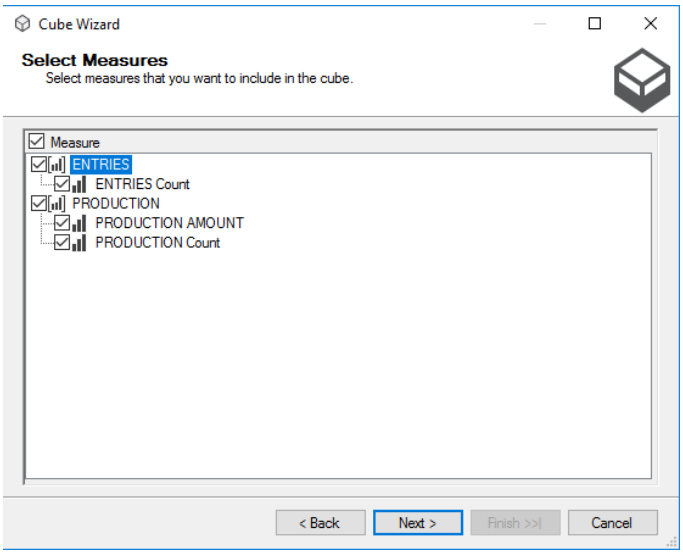


Figura 78. Medidas del Cubo

a dimensión es la parte más importarte de un cubo ya que de esta depende el análisis a realizar, esta es la tabla principal o de donde se obtendrá los patrones más importantes.

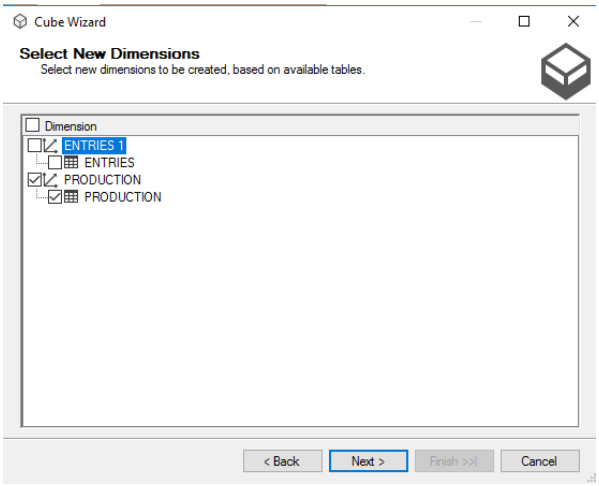


Figura 79. Dimensión del Cubo

Como paso final para crear el cubo, el asistente solicita que se le asigne un nombre.

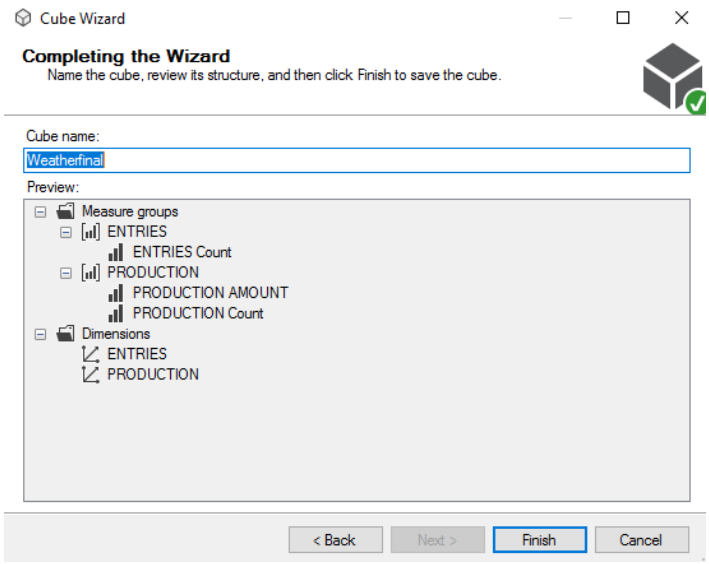


Figura 80. Asignación de nombre para el cubo

Una vez creado el cubo, se procederá a ver si este contiene errores en caso de tener errores este no ejecutara al momento de realizar el proceso de minería, se recomienda bien escoger bien las medidas y las dimensiones del cubo OLAP. Dar clic sobre el cubo creado y seleccionar la opción de PROCESS, Se desplegarán dos mensajes, dar clic en sí, y por último dar escoger en RUN, si este está bien generado mostrada dos ventanas con un visto verde en la esquina.

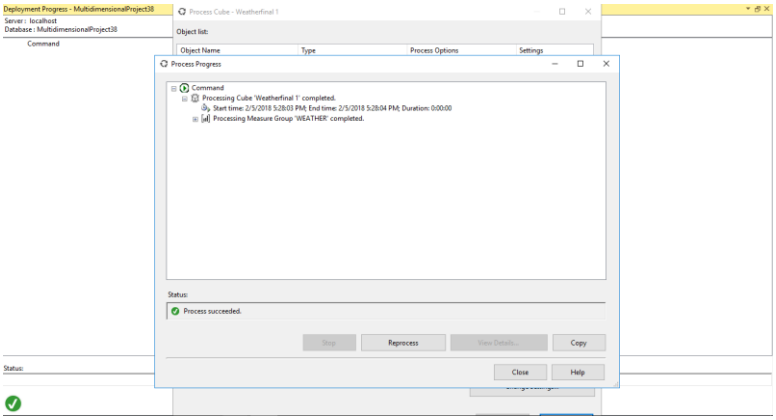


Figura 81. Process del Cubo

Como paso final se creará el mining structure, ir al explorer solution en la capeta con el mismo nombre dar clic derecho, crear uno nuevo. Se abrirá una ventana del asistente dar clic, en siguiente y se mostrará las opciones de usar la base de datos relaciona o un cubo seleccionamos la primera opción

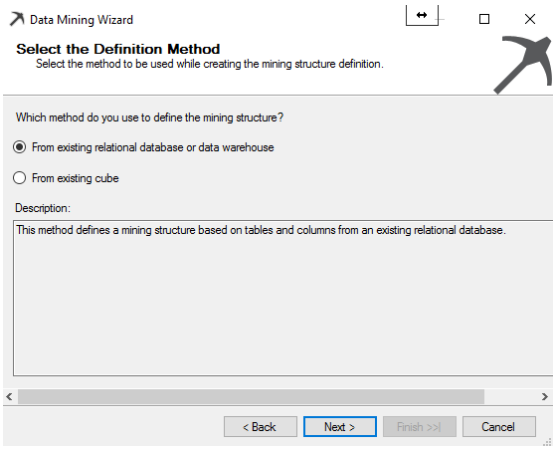


Figura 82. Configuración del Data Minig Structure

Como siguiente paso se debe seleccionar el algoritmo que se va a utilizar, tomando en cuenta cuales son los valores que realmente necesitamos predecir y tomando en cuenta los factores

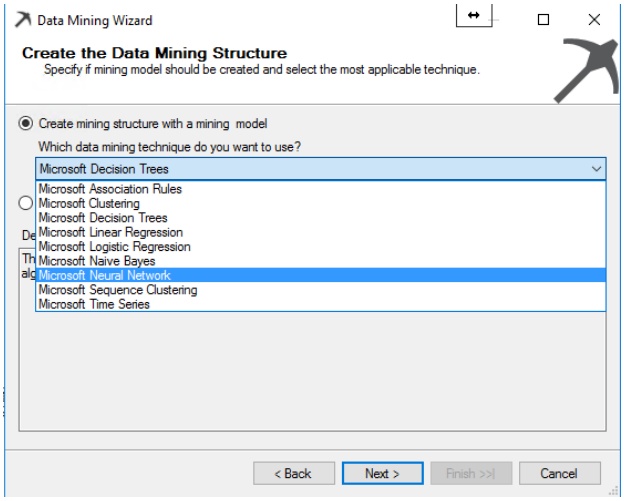


Figura 83. Listado de Algoritmos

En este punto se analiza, cual es mejor algoritmo que se va a utilizar para predecir la cantidad de rosas que se pueden producir tomando en cuenta. Las variables y recomendaciones que ofrece Microsoft

Según los pasos anteriores, se fue definiendo poco a poco van a ser los datos que se iban a utilizar y cuál debe ser el producto final de la implementación del algoritmo.

La empresa rosa prima busca:

- Predecir el numero tallos que se puede producir
- Predecir cuando la flor está más propensa a una enfermedad
- Predecir los cambios climatológicos.

Elegir un algoritmo por tarea

Con el fin de ayudarle a seleccionar un algoritmo para su uso con una tarea específica, la tabla siguiente proporciona sugerencias para los tipos de tareas para las que se usa normalmente cada algoritmo.

Ejemplos de tareas	Algoritmos de Microsoft que se pueden usar
Predecir un atributo discreto: Marcar los clientes de una lista de posibles compradores como clientes con buenas o malas perspectivas. Calcular la probabilidad de que un servidor genere un error en los próximos 6 meses. Clasificar la evolución de los pacientes y explorar los factores relacionados.	Algoritmo de árboles de decisión de Microsoft Algoritmo Bayes naive de Microsoft Algoritmo de clústeres de Microsoft Algoritmo de red neuronal de Microsoft
Predecir un atributo continuo: Pronosticar las ventas del año próximo. Predecir los visitantes del sitio a partir de tendencias históricas y estacionales proporcionadas. Generar una puntuación de riesgo a partir de datos demográficos.	Algoritmo de árboles de decisión de Microsoft Algoritmo de serie temporal de Microsoft Algoritmo de regresión lineal de Microsoft
Predecir una secuencia: Realizar un análisis clickstream del sitio web de una empresa. Analizar los factores que dan como resultado errores en el servidor. Capturar y analizar secuencias de actividades durante las visitas de pacientes externos, para formular las prácticas recomendadas en las actividades comunes.	Algoritmo de clústeres de secuencia de Microsoft
Buscar grupos de elementos comunes en las transacciones: Usar el análisis de la cesta de la compra para determinar la posición del producto. Sugerir a un cliente la compra de productos adicionales. Analizar los datos de una encuesta a los visitantes a un evento, para descubrir qué actividades o stands estaban correlacionados con el fin de programar actividades futuras.	Algoritmo de asociación de Microsoft Algoritmo de árboles de decisión de Microsoft
Buscar grupos de elementos similares: Crear grupos de pacientes con perfiles de riesgo en función de atributos como datos demográficos y comportamientos. Analizar usuarios mediante patrones de búsqueda y compra de productos. Identificar servidores con características de uso similares.	Algoritmo de clústeres de Microsoft Algoritmo de clústeres de secuencia de Microsoft

Figura 84. Recomendaciones Microsoft de Algoritmos

Para este proyecto realizaremos la validación de series temporales, ya que puede pronosticar cual es la producción en un determinado tiempo, además permite realizar las validaciones con los históricos que se encuentran almacenados

3.4.7. Validación de los Modelo

Series Temporales

Dentro de este tipo de algoritmos la variable de tiempo es muy esencial, Microsoft utiliza la metodología ARIMA con el fin de realizar sus predicciones mediante la revisión del marco teórico, identificación de variables relevantes, y la especificación de forma funcional permitiendo que los datos temporales de una variable permitan el estudio de las características de una estructura probabilística subyacente esto también se lo conoce como ECONOMETRÍA DE SERIES TEMPORALES.

- **Proceso estocástico**
  Se denomina así a una sucesión de variables ordenadas aleatoriamente denotadas como $\mathbf{Y_t}$ donde t toma cualquier valor de -∞ y ∞.

$$Y_{-5}, y_{-4}, y_{-3}, y_{-2}, \ldots\ldots\ldots y_3, y_4$$ *Ecuación1. Variables Yt*

- **Serie Temporal y proceso estocástico**
  Se deduce que una serie temporal es una muestra concreta con valores. El análisis de las series temporales parte de los datos que la componen infiriendo características probabilísticas subyacentes.
- **Estacionariedad de un proceso**
  La utilización de modelos ARIMA como estrategia de predicción de series temporales sólo tiene sentido si las características observadas en la serie permanecen en el tiempo.
  - **Proceso estocástico estacionario en sentido fuerte.**
    Cada una de las variables $\mathbf{Y_t}$ que configuran un proceso estocástico tendrán su propia función de distribución con sus correspondientes momentos. Cada conjunto de variables tendrá su correspondiente función de distribución conjunta y sus funciones de distribución marginales. Habitualmente, conocer

esas funciones de distribución resulta complejo de forma que, para caracterizar un proceso estocástico, basta con especificar la media y la varianza para cada $\mathbf{Y_t}$ y la covarianza para variables referidas a distintos valores de t:

$$E[Y_t] = \mu_t$$

$$\sigma_t^2 = Var(y_t) = E[y_t - \mu_t]^2$$ *Ecuación2. Media, Varianza y Covarianza de*

$$\gamma_{t,s} = Cov(Y_t, Y_s) = E[(y_t - \mu_t)(y_s - \mu_s)]$$

Yt

Un proceso estocástico es estacionario en sentido estricto o fuerte si las funciones de distribución son invariantes con respecto a un desplazamiento en el tiempo es decir considerando que t, t+1, t+2, ..., t+k reflejan períodos sucesivos:

$$F(Y_t, Y_{t+1}, ..., Y_{t+k}) = F(Y_{t+m}, Y_{t+1+m}, ..., Y_{t+k+m})$$ *Ecuación3. Desplazamiento de t*

- **Proceso estocástico estacionario en sentido débil**

  Un proceso estocástico es débilmente estacionario sí. Las probabilidades matemáticas de las variables aleatorias no dependen del tiempo, son constantes:

$$E[Y_t] = E[Y_{t+m}] \quad \forall \ m$$ *Ecuacion4. Y y t son constantes*

Las varianzas tampoco dependen del tiempo (y son finitas):

$$Var[Y_t] = Var[Y_{t+m}] \neq \infty \quad \forall \ m$$ *Ecuación5. Y y t son finitas*

Las covarianzas entre dos variables aleatorias del proceso correspondientes a períodos distintos de tiempo (distintos valores de t) sólo dependen del lapso de tiempo transcurrido entre ellas:

$$Cov(Y_t, Y_s) = Cov(Y_{t+m}, Y_{s+m}) \quad \forall \ m$$ *Ecuacion6. Covarianza de Variables*

De esta última condición se desprende que, si un fenómeno es estacionario, sus variables pueden estar relacionadas linealmente entre sí, pero de forma que la relación entre dos variables sólo depende de la distancia temporal k transcurrida entre ellas.

- **Definición informal de un proceso estacionario**

De una manera informal, diremos que un proceso es estacionario cuando se encuentra en equilibrio estadístico, en el sentido de que sus propiedades no varían a lo largo del tiempo

- **Ruido Blanco**

  Un ruido blanco es una sucesión de variables aleatorias con esperanza nula, varianza constante, y covarianzas nulas para distintos valores de t.

- **Modelos autorregresivos AR(p)**

  Definimos un modelo AR como aquel en el que la variable endógena de un período t es explicada por las observaciones de ella misma correspondientes a períodos anteriores (parte sistemática) más un término de error ruido blanco. El orden del modelo expresa el número de observaciones retasadas de las series temporal analizada que intervienen en la ecuación.

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + a_t$$ *Ecuación7. AR (1)*

La expresión genérica de un modelo autorregresivo, de un AR(p) donde p es el número observaciones a analizar de sería la siguiente:

$$Y_t = \phi_0 + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + a_t$$ *Ecuación8. Autorregresiva de AR(p)*

Esta forma funcional se acompaña de una serie de restricciones conectadas con importantes hipótesis analíticas:

- El proceso no debe ser anticipante (hipótesis de recursividad temporal); lo que quiere decir que los valores de una variable en un momento t no dependerán de los que esta misma tome en t+j.
- La correlación entre una variable y su pasado va reduciéndose a medida que nos alejamos más en el tiempo (proceso ergódico)
- La magnitud de los coeficientes está limitada en valor absoluto: así, por ejemplo, en el caso de un AR (1), el coeficiente autorregresivo de un proceso estocástico estacionario ha de ser inferior a 1 en valor absoluto; en el caso de un Ar (2), es la suma de los dos coeficientes la que no puede exceder la unidad. Estas restricciones expresadas en los coeficientes conectan con las propiedades de estacionariedad del proceso analizado o, dicho de

otro modo: sólo los modelos cuyos coeficientes respetan una serie de condiciones (que dependen del orden "p" del modelo) representan procesos estocásticos estacionarios y, por tanto, tienen utilidad analítica.

- **Operador Retardos**

  El operador retardo $\mathbf{L^p}$ aplicado al valor $\mathbf{Y_t}$ de una determinada serie devuelve el valor de esa serie retardado "p" observaciones, es decir:

$L^p Yt = Y_{t-p}$ *Ecuación9*. Ecuación de Retardo

- **Polinomio de Retardos**

  Un polinomio de retardos de orden "p" $\phi_p(L)$ se compone de una sucesión de "p" operadores de retardos con sus respectivos coeficientes:

$$\phi_p(L) = 1 - \phi_1 L - \phi_2 L^2 - ..... - \phi_p L^p$$ *Ecuación10. Polinomio de Retardos*

Los analistas pueden evaluar características relevantes del proceso estocástico que se está modelizando estudiando las propiedades matemáticas del polinomio de retardos.

- **Modelo de medias móviles MA(q)**

  Un modelo de los denominados de medias móviles es aquel que explica el valor de una determinada variable en un período t en función de un término independiente y una sucesión de términos de error, de innovaciones correspondientes a períodos precedentes, convenientemente ponderados.

$$Y_t = \mu + a_t + \theta_1 a_{t-1} + \theta_2 a_{t-2} + ... + \theta_q a_{t-q}$$ *Ecuación11. Medias Móviles*

que de nuevo puede abreviarse utilizando el polinomio de retardos (como en el caso de los modelos AR):

$$Y_t = \theta_q(L) a_t + \mu$$ *Ecuación12. Retardos MA(q)*

En realidad, un modelo de medias móviles puede obtenerse a partir de un modelo autorregresivo sin más que realizar sucesivas sustituciones:

$$Y_t = \phi Y_{t-1} + a_t \rightarrow Y_{t-1} = \phi Y_{t-2} + a_{t-1} \rightarrow$$
$$Y_t = a_t + \phi a_{t-1} + \phi^2 Y_{t-2} \rightarrow .....$$
$$.....Y_t = a_t + \phi a_{t-1} + \phi^2 a_{t-2} + \phi^3 a_{t-3} + ... + \phi^j a_{t-j} +$$

Ecuación13. Autorregresiva móvil

Validación de Series Temporales en Rosaprima

Para poder validar los resultados que se van a obtener, se van a tomar como muestra un conjunto de datos de las diferentes tablas que componen la data warehouse.

- **Validación de Producción**

  Para realizar la validación de algoritmo de series tomara como muestra las variedades de color las cuales son más populares en el mercado.

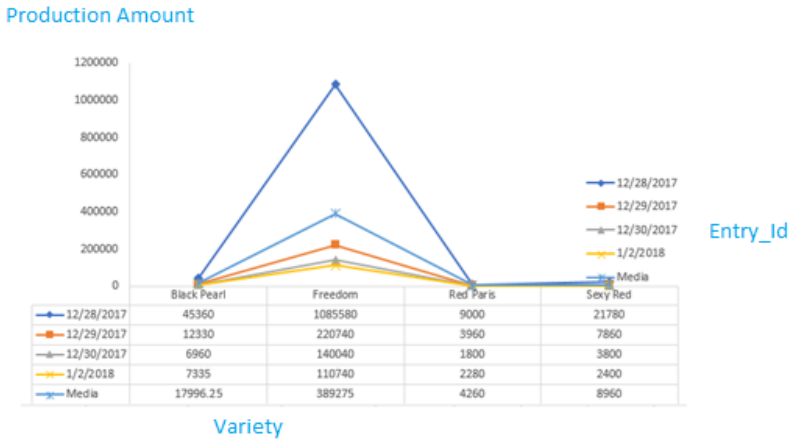


Figura 85. Gráfica Actual del Muestreo de Producción

Cálculo de la Media de Producción

$$\boldsymbol{M = \frac{Y1 + Y2 + Y3 + Y4 \ldots\ldots\ldots\ldots.}{t}}$$

$$\frac{45360 + 12330 + 6960 + 7335}{4} = 17996.25$$

Media de Black Pearl

$$\frac{1085580 + 220740 + 140040 + 110740}{4} = 17133.75$$

Media de Freedom

$$\frac{9000 + 3690 + 1800 + 2280}{4} = 4192.5$$

Media de Red Paris

$$\frac{21780 + 7860 + 3800 + 2400}{4} = 8960$$

Media Sexy Red

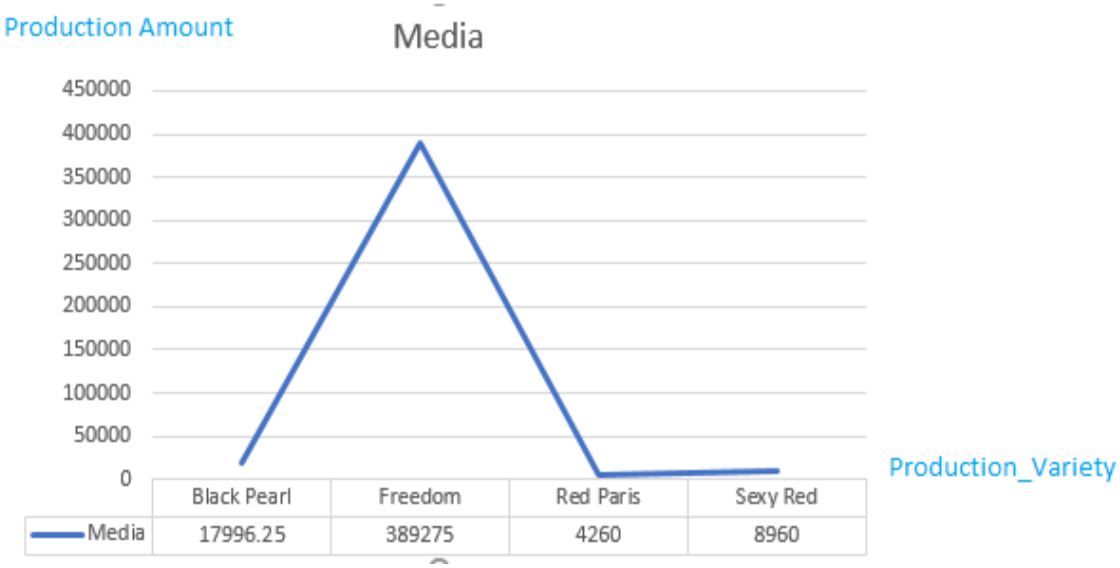


Figura 86. Representación de la Media de Producción

Calculo de la Media Móvil de Producción (Serie Temporal)

$$\boldsymbol{M = \frac{\Sigma Y + Yx}{t}}$$

$$\frac{45360 + 12330 + 6960 + 7335 + 3945}{4} = 18982.50$$

Media Móvil Black Pearl

$$\frac{1085580 + 220740 + 140040 + 110740 + 58820}{4} = 403980$$

Media Móvil Freedom

$$\frac{9000 + 3690 + 1800 + 2280 + 960}{4} = 4027.50$$

Media Móvil Red Paris

$$\frac{21780 + 7860 + 3800 + 2400 + 2020}{4} = 9465$$

Media móvil Red Paris

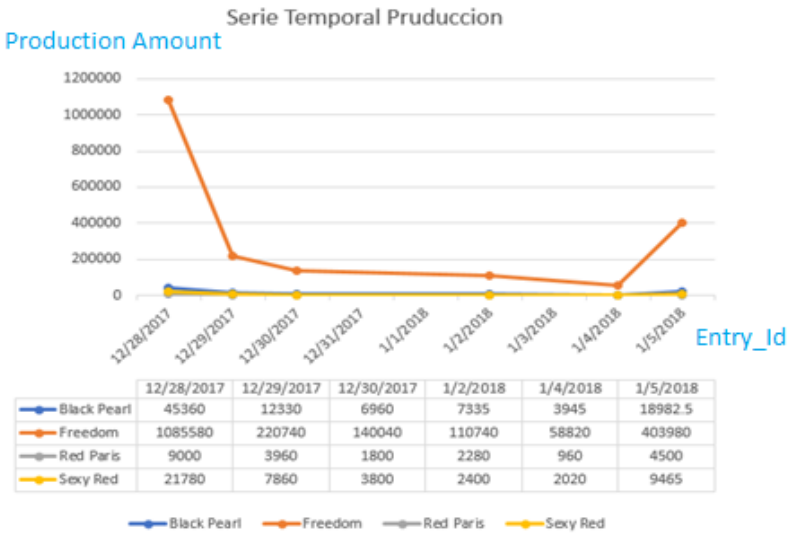


Figura 87. Serie Temporal predicción del muestreo

En la FIG81 se puede evidenciar que en el muestreo tomado del data Warehouse aplicando las ecuaciones de serie temporal. Como se puede ver para el 5 de enero se obtuvo un crecimiento en las variedades utilizadas. Sin embargo, hay que tomar en cuenta que para el modelo de serie temporal completo depende de todas las variedades y puede existir un crecimiento, decrecimiento, o mantenerse la producción.

Una vez analizado el funcionamiento matemático de las Series Temporales en Microsoft, se procede a implementado en el Microsoft Visual Studio Data tool

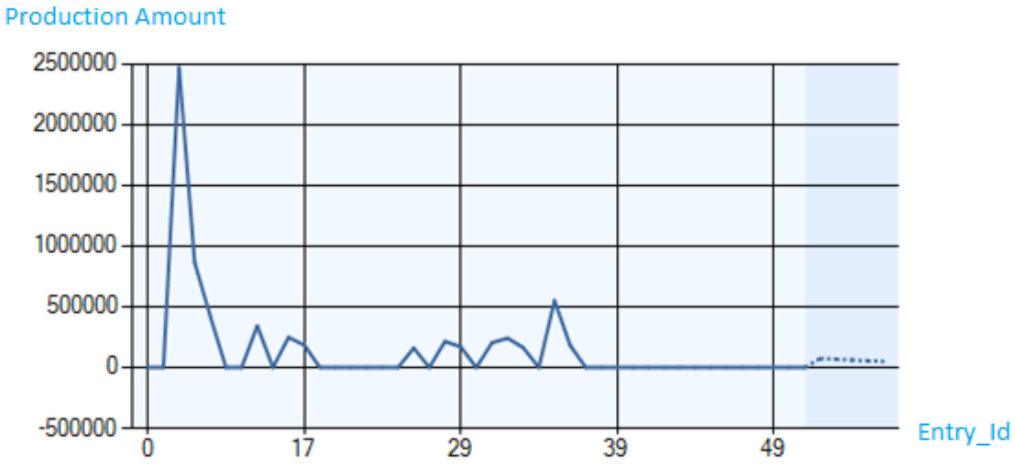


Figura 88. Resultados Serie Temporal Producción

3.4.8. Implementación del Modelo

Para que el usuario tenga acceso, debe ingresar de forma remota al servidor del aplicativo empresarial, para ello debe ir REMOTE DESKTOP CONNECTION, ingresar al servidor 10.0.1.21, el cual se encuentra en Amazon. Dentro de este el usuario debe tener configurados los ODBCs explicados anteriormente tanto para el ingreso a la base de datos principal y a la base del Data Warehouse.

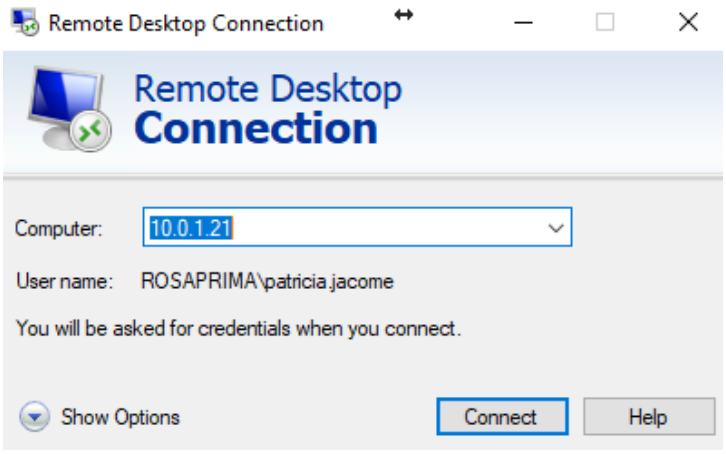


Figura 89. Ingreso al Servidor de Amazon

Para mantener la seguridad del ingreso de los usuarios cada uno tiene acceso, mediante credenciales que se encuentran vinculadas al Active Directory de la empresa.

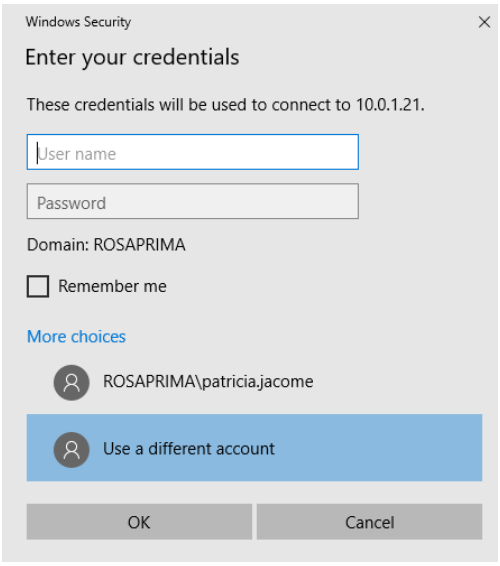


Figura 90. Ingreso de Credenciales al servidor en la Nube

Una vez ingresado al servidor vía remoto, en el escritorio se desplegará el escritorio en el cual se encontrará un archivo de Excel, este archivo se encuentra configurado para que se conecte al DW.

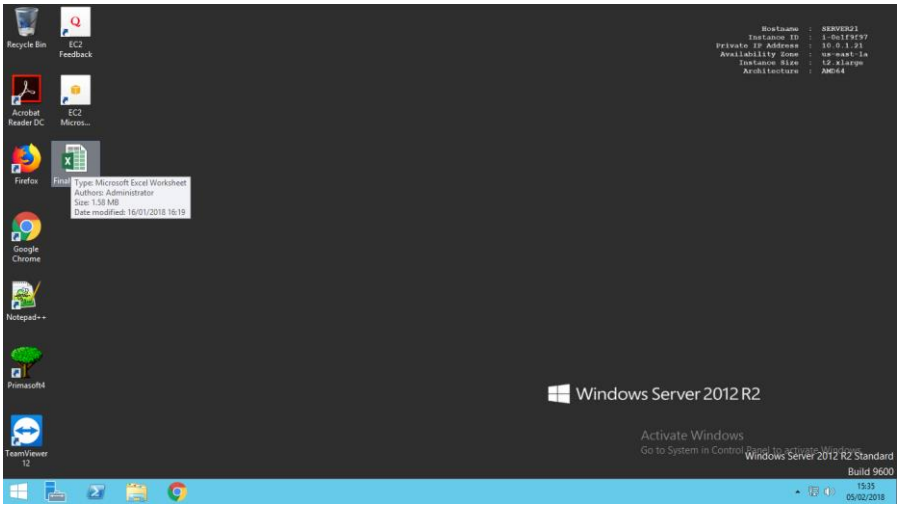


Figura 91. Escritorio y archivo de Excel

Dentro del archivo, se encontrarán varias pestañas u hojas de trabajo las cuales muestran cada una de las tablas del DW. El usuario como primer paso debe actualizar las conexiones a la base de datos. Debe ir al menú DATA y dar clic en REFRESH ALL con el fin de actualizar las conexiones y los datos ingresados

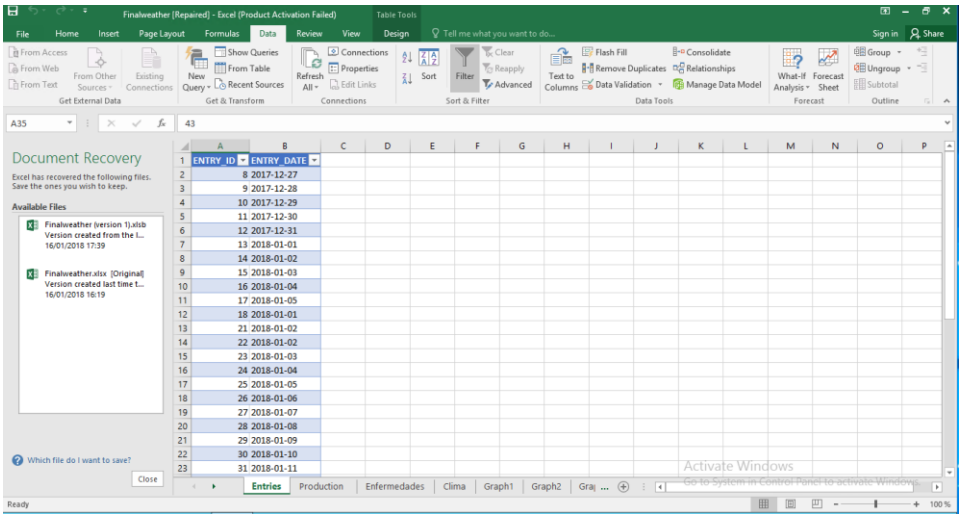
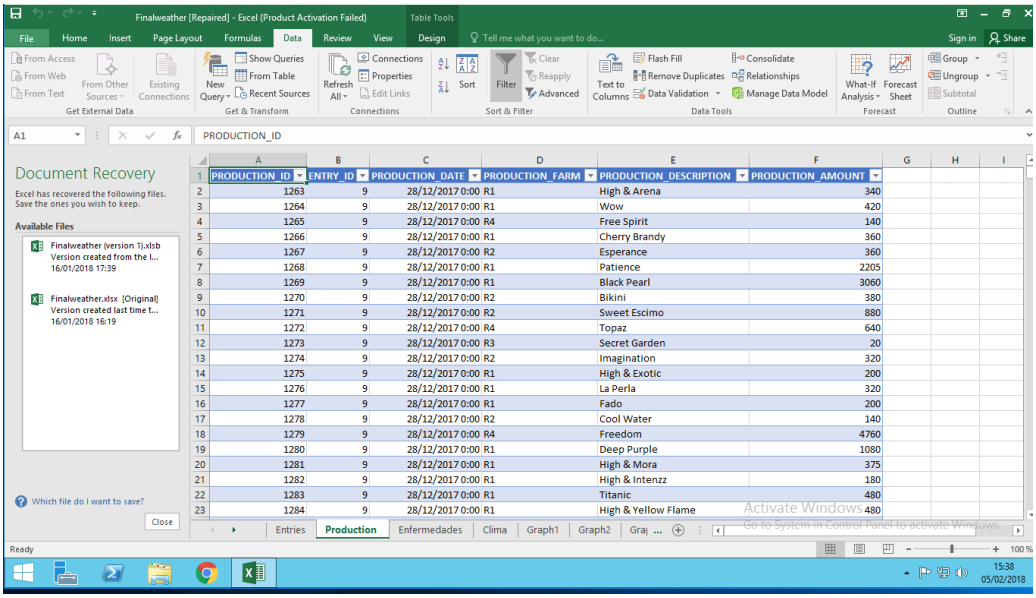
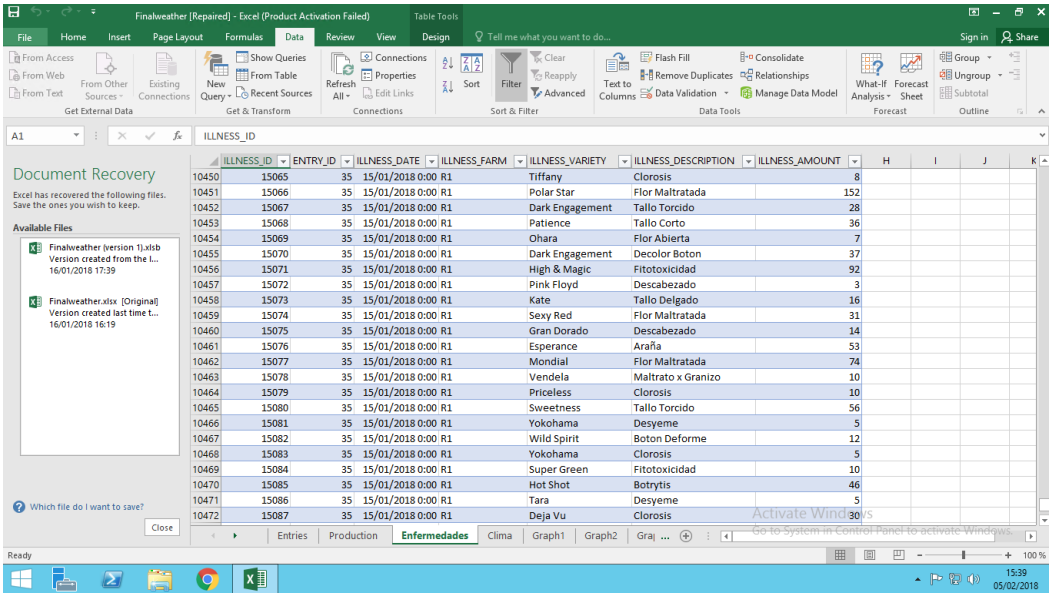


Figura 92. Hoja de Tabla Entries



	PRODUCTION_ID	ENTRY_ID	PRODUCTION_DATE	PRODUCTION_FARM	PRODUCTION_DESCRIPTION	PRODUCTION_AMOUNT
2	1263	9	28/12/2017 0:00	R1	High & Arena	340
3	1264	9	28/12/2017 0:00	R1	Wow	420
4	1265	9	28/12/2017 0:00	R4	Free Spirit	140
5	1266	9	28/12/2017 0:00	R1	Cherry Brandy	360
6	1267	9	28/12/2017 0:00	R2	Esperance	360
7	1268	9	28/12/2017 0:00	R1	Patience	2205
8	1269	9	28/12/2017 0:00	R1	Black Pearl	3060
9	1270	9	28/12/2017 0:00	R2	Bikini	380
10	1271	9	28/12/2017 0:00	R2	Sweet Escimo	880
11	1272	9	28/12/2017 0:00	R4	Topaz	640
12	1273	9	28/12/2017 0:00	R3	Secret Garden	20
13	1274	9	28/12/2017 0:00	R2	Imagination	320
14	1275	9	28/12/2017 0:00	R1	High & Exotic	200
15	1276	9	28/12/2017 0:00	R1	La Perla	320
16	1277	9	28/12/2017 0:00	R1	Fado	200
17	1278	9	28/12/2017 0:00	R2	Cool Water	140
18	1279	9	28/12/2017 0:00	R4	Freedom	4760
19	1280	9	28/12/2017 0:00	R1	Deep Purple	1080
20	1281	9	28/12/2017 0:00	R1	High & Mora	375
21	1282	9	28/12/2017 0:00	R1	High & Intenzz	180
22	1283	9	28/12/2017 0:00	R1	Titanic	480
23	1284	9	28/12/2017 0:00	R1	High & Yellow Flame	480

Figura 93. Hoja de Tabla de Producción



	ILLNESS_ID	ENTRY_ID	ILLNESS_DATE	ILLNESS_FARM	ILLNESS_VARIETY	ILLNESS_DESCRIPTION	ILLNESS_AMOUNT
10450	15065	35	15/01/2018 0:00	R1	Tiffany	Clorosis	8
10451	15066	35	15/01/2018 0:00	R1	Polar Star	Flor Maltratada	152
10452	15067	35	15/01/2018 0:00	R1	Dark Engagement	Tallo Torcido	28
10453	15068	35	15/01/2018 0:00	R1	Patience	Tallo Corto	36
10454	15069	35	15/01/2018 0:00	R1	Ohara	Flor Abierta	7
10455	15070	35	15/01/2018 0:00	R1	Dark Engagement	Decolor Boton	37
10456	15071	35	15/01/2018 0:00	R1	High & Magic	Fitotoxicidad	92
10457	15072	35	15/01/2018 0:00	R1	Pink Floyd	Descabezado	3
10458	15073	35	15/01/2018 0:00	R1	Kate	Tallo Delgado	16
10459	15074	35	15/01/2018 0:00	R1	Sexy Red	Flor Maltratada	31
10460	15075	35	15/01/2018 0:00	R1	Gran Dorado	Descabezado	14
10461	15076	35	15/01/2018 0:00	R1	Esperance	Araña	53
10462	15077	35	15/01/2018 0:00	R1	Mondial	Flor Maltratada	74
10463	15078	35	15/01/2018 0:00	R1	Vendela	Maltrato x Granizo	10
10464	15079	35	15/01/2018 0:00	R1	Priceless	Clorosis	10
10465	15080	35	15/01/2018 0:00	R1	Sweetness	Tallo Torcido	56
10466	15081	35	15/01/2018 0:00	R1	Yokohama	Desyeme	5
10467	15082	35	15/01/2018 0:00	R1	Wild Spirit	Boton Deforme	12
10468	15083	35	15/01/2018 0:00	R1	Yokohama	Clorosis	5
10469	15084	35	15/01/2018 0:00	R1	Super Green	Fitotoxicidad	10
10470	15085	35	15/01/2018 0:00	R1	Hot Shot	Botrytis	46
10471	15086	35	15/01/2018 0:00	R1	Tara	Desyeme	5
10472	15087	35	15/01/2018 0:00	R1	Deja Vu	Clorosis	30

Figura 94. Hoja de Tabla Enfermedades

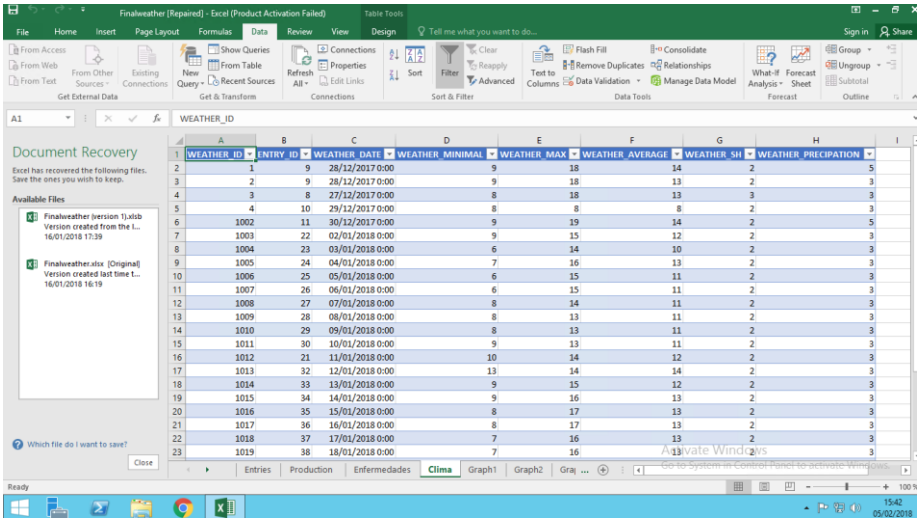


Figura 95. Hoja de Tabla Clima

Además, también contiene hojas la cuales está conectadas a los cubos OLAP, creados dentro del Microsoft Studio Data Tool, cuyo objetivo es mostrar como se encuentran las gráficas de producción, enfermedades y el clima actualmente para después realizar la proyección

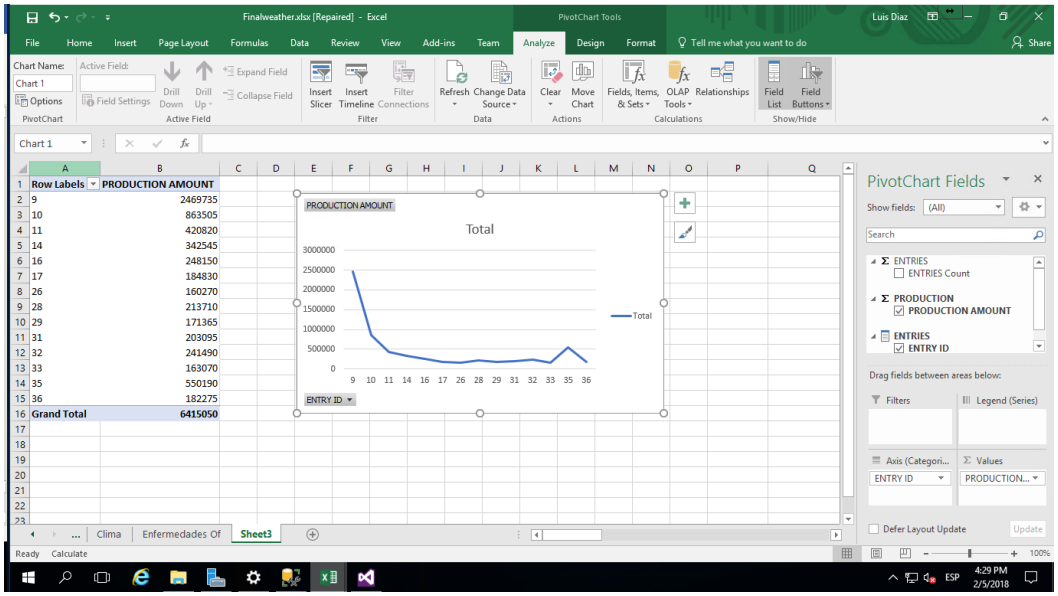


Figura 96. Representación Actual de Producción

A continuación de la gráfica actual, se procederá a insertar la proyección realizada dentro del Microsoft SQL Data Tool, con el fin de que esta pueda ser interpretada por el usuario final.

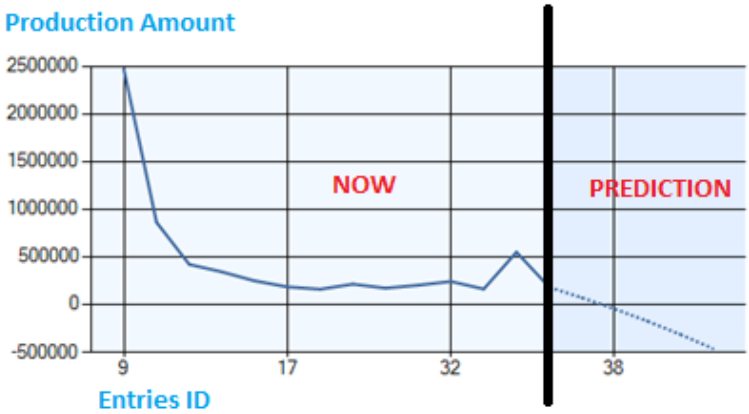


Figura 97. Proyección de Producción

Por ejemplo, en la FIG 97 se puede observar en la gráfica actual que la producción de la flor ha ido decreciendo hasta el ENTRY_ID: 35. La proyección muestra que a partir de dicho registro la producción ira decayendo hasta no obtener producción, incluso generando pérdidas. Pero es necesario el saber el porqué de este decrecimiento. Por ellos se realiza también el análisis de las otras dos tablas enfermedades y clima.

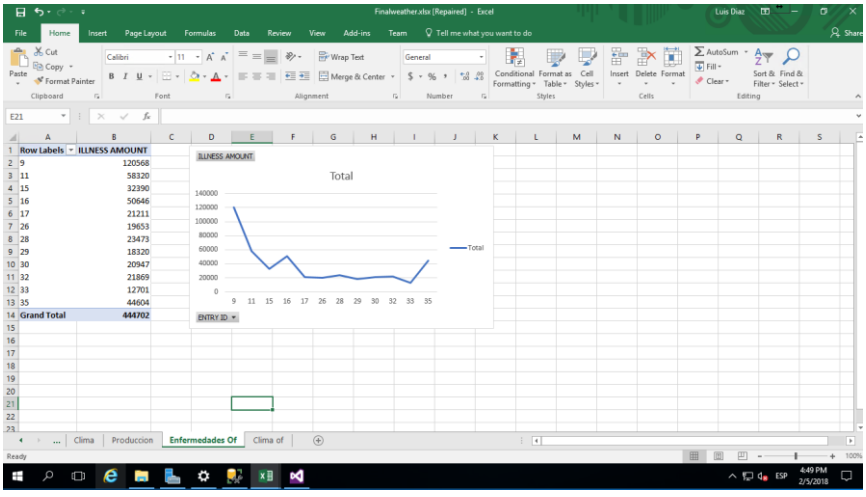


Figura 98. Representación Actual de Enfermedades

En la Fig.98 se muestra la hoja donde se graficaron los datos actuales de las enfermedades presentes en la producción, mostrando cual fue la cantidad afectada en las diferentes fechas.

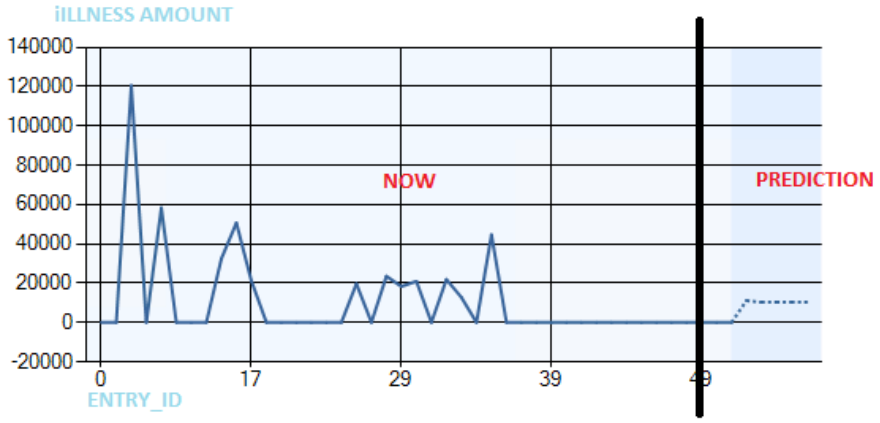


Figura 99. Proyección de Enfermedades

Al parecer la aparición de enfermedades es un factor muy bajo para afectar la producción en las rosas como se puede visualizar en la Fig. 99, Hasta el ENTRY_ID; 49, las enfermedades se mantuvieron bajas, pero su proyección denota que puede que estas pueden crecer.

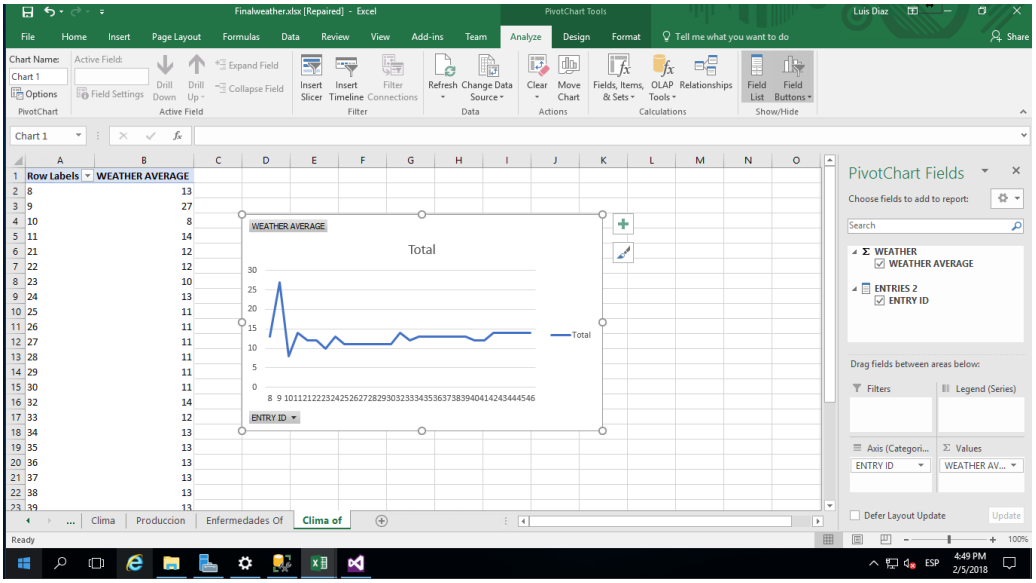


Figura 100. Representación Actual del Clima

En la FIG100, se evidencia la tabla climatológica, para obtener estos datos se utilizó el promedio entre la temperatura máxima y la temperatura mínima de un cierto día almacenado en la tabla de ENTRY.

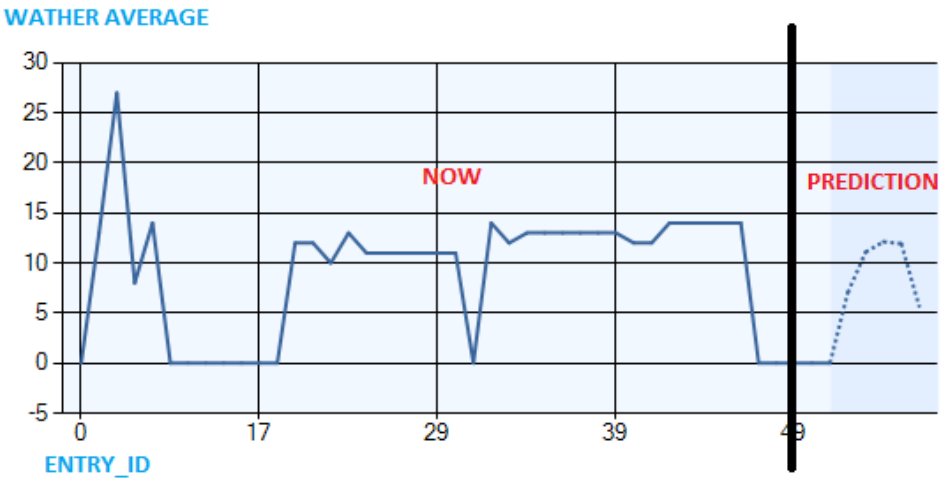


Figura 101. Proyección Clima

En la Fig101, muestra que el clima es muy variante, a partir del ENTRY_ID:49 el clima obtendrá una pequeña variación.

Como análisis general se puede constatar que la producción y las enfermedades dependen mucho del clima, ya que si este es muy elevado se obtendrá un mayor número de tallos, pero las enfermedades se harán más presentes, Mientras cuando el clima tiende a ser templado este mantiene un equilibrio entre la producción y las enfermedades. Y cuando es frio la producción baja ya que este suele quemar más los pétalos y la flor tiende a ser rechazada para su venta y exportación

4. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones

La aplicación de Business Intelligence es una herramienta muy poderosa dentro de cualquier empresa y brinda la posibilidad de crecer mediante el uso de su propia información tanto interna como externa proporcionada por los clientes o proveedores de la misma

El uso de un servidor en la nube virtualizado con lleva un ahorro en la infraestructura de la empresa, sin embargo, al manejar una gran cantidad de datos, estos suelen trabajar con lentitud e incluso suele presentar cortes de comunicación, además puede presentar congelamientos del gestor de la base de datos al momento de realizar las consultas para obtener los datos.

La minería de datos busca dar respuesta a varias incógnitas dentro de una empresa, años atrás los datos imposibilitaba el manejo de grandes volúmenes de datos con la aparición de nuevo software, mediante el cual se ha aumentado su velocidad de procesado y precisión en cálculos. Las empresas utilizan la minería de datos como una herramienta de última generación con el fin de buscar tendencias no visibles dentro de sus datos para tomar decisiones rentables.

El uso de algoritmos en la minería de datos evita que el usuario realice cálculos matemáticos estadísticos o la búsqueda de patrones repetitivos dentro de la base de datos, brindado así tener valores más ciertos de producción e incluso ver cuál es la variedad más afectada tomando en cuenta las enfermedades y el clima

4.2. Recomendaciones

Al manejar grandes volúmenes de información dentro de la nube, se requiere un gran ancho de banda para evitar los cortes de comunicación con el servidor, congelamiento y lentitud en las consultas, e incluso caída del servidor mismo, se recomienda contratar con el ISP un plan el cual pueda solventar a la empresa sin la presencia de fallos.

El uso de un servidor virtualizado con lleva ahorro a empresa sin embargo puede presentar una debilidad si este no se encuentra activo todo el tiempo, por lo cual se recomienda tener un servidor físico, dentro del cual se encuentre una réplica de la base de datos y del data warehouse.

Al momento de manejar base de datos se deber mucho cuidado, ya que son datos reales de la empresa, y pueden ser utilizados en otros procesos activos, por lo cual es necesario tener una base de pruebas o una réplica de la misma en el servidor back.

Para evitar la pérdida de información o la sustracción de información, se recomienda tener un back up de la misma ejecutada un periodo de tiempo predeterminado, además almacenarlo en un servidor seguro o en un disco seguro, sin embargo, si desea almacenar en la nube el uso VPN dentro de esta beneficia la seguridad de la data.

REFERENCIAS

Aguirre G, Andrade H, Maldonado D, y Ureta L. (2006). *BI Business Intelligence* (Tesis de Maestría): Universidad del Litoral, Guayaquil, Ecuador. Recuperado el 3 Octubre del 2017 de http://www.msig.espol.edu.ec/recursos/1.business_intelligence_resumen.pdf.

Arevalillo J. (2008). *Data Mining con Arboles de Decisión* [PDF file].. Madrid, España: Universidad Complutense de Madrid. Recuperado el 30 de Octubre del 2017 de https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartin-slides.pdf.

Cano J. (2014). *Business Intelligence: Competir con la Información.* (1.*ed.). Barcelona, España: ESADE

Espinosa R. (2009). *¿Qué es business Intelligence?* [PDF file]. Madrid, España: Rincón de BI. Recuperado: 5 de noviembre del 2017 de https://churriwifi.wordpress.com/2009/11/02/1-que-es-business-intelligence/an

Ferrari A. y Russo M. (2015). *The Definitive Guide to DAX. Business Intelligence with Microsoft Excel, SQL server Analysis Services, and PowerBI* (1*.ed.). Seatle, United States: Microsoft Press.

Ferrari A. y Russo M. (2016). *Introducing Microsoft Power BI.* (1*.ed.). Seatle, United States: Octal Publishing

Han J y Kamber M. (2006). *Data Mining: Concepts and Techniques.* (2.*ed.). San Francisco, United States: Morgan Kaufmann

Hughes R. (2012). *Agile Data Warehousing Project Management.* (1.* ed.). Massachusset, United States: Elsevier

Ibermática. (2007). *Business Intelligence . El conocimiento Compartido* [PDF file]. Madrid , España. Recuperado el 3 de diciembre del 2017 https://churriwifi.files.wordpress.com/2009/11/business-intelligence-ibermatica.pdf

Microsoft. (2017), *Conceptos de Minería de Datos.* Recuperado 6 de octubre del 2017. https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts

Oracle. (2012). *¿Qué es Business intelligence?* [PDF file]. Recuperado el 5 de Octubre del 2017 de http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/317529_esa.pdf.

Root R y Mason C. (2012). *Pro SQL server 2012 BI Solutions*. (1.*ed.). NY, United States: Springe.

Rosado A y Rico Dewar. (2010). *Inteligencia de Negocios: Estado del Arte/Business Intelligence: State fo the Art* (Informe IEEE). Universidad Tecnológica de Pereira, Pereira, Colombia

Whitee K. (2010). *Microsoft Business Intelligence*. (1.* ed.). Indianapolis, United States: Wiley Publising INC.

ANEXOS

Glosario

- **BI:** Business Intelligence
- **IT:** Information Techonology
- **KPI:** Key performance indicator
- **ERP:** Enterprise Resource Planning
- **CRM:** Customer Relashionship Management
- **SCM:** Supply Chain Management
- **ETL:** Extract. Transforation, Load
- **DW:** Data Warehouse
- **KDD:** Knowledge Discovery in Databases
- **DB:** Data Dase
- **OLAP:** On Line Analytical Processing
- **OLTP:** On Line Transaction Processinh
- **ROLAP:** Relational On Line Analytical Processing
- **MOLAP:** Multidimentional On Line Analytical Processing
- **SQL:** Structured Query Language
- **KPA:** Kilopascal
- **TR:** Telecommunications Room
- **MDA:** Main Distribution Area
- **HDA:** Horizontal Distribution Area
- **ER:** Equipment Room
- **EDA:** Equipment Distribution Area
- **RT:** Router
- **FW:** Firewall
- **SW:** Switch
- **AP:** Access Point

- **ANSI/TIA:** American National Stardards Institute/ Telecommunication Industy Association
- **UPS:** Uninterruptible Power Supply
- **PDU:** Power Distribution Unit
- **BDM** : Business Development Methodology
- **ASL: Aplication Services Library**
- **CobiT:** Control objectives information and related Technology
- **eTOM** enhanced Telecom Operation Map
- **IPW:** Introducing Process orientes Working methods
- **IMM:** IT Management Model
- **ISM:** Integrated Service Management
- **MSP:** Managerial Step by Step plan
- **MIP:** Managing the information Provision
- **ITPM:** T Process Model
- **ISPL:** Information Services Procurement Library
- **MOF:** Microsoft Operation Framework
- **RPM:** Recursive Process Management
- **SIMA:** Standart Integrated Manage Approach
- **ITIL:** IT Infrastructure Library
- **SLA:** Service Level Agreement
- **OLA:** Operation Level Agreement
- **PDCA:** Planification, Do, Check, Act
- **SSIS:** SQL Server Integration Services

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

PREDICCIÓN METEOROLÓGICA

INTELIGENCIA EN REDES DE COMUNICACIONES

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

INTRODUCCIÓN

En esta práctica se pretende obtener un sistema real de predicción meteorológica utilizando técnicas de aprendizaje automático para obtener modelos de caracterización y predicción, empleando como herramienta la plataforma Weka. En concreto se pretende desarrollar tres modelos de predicción:

Predicción de temperatura a 1 hora
Predicción de temperatura a 24 horas
Predicción de condiciones meteorológicas a 24 horas

Para poder desarrollar este sistema, se ha tenido que usar una información adicional que serán los datos necesarios para el aprendizaje automático. Estos datos han sido obtenidos a través del METAR (Meteorological Actual Report) donde se proporcionan los datos en los siguientes campos:

- HoraCET
- Temperatura (grados Fahrenheit)
- Punto de rocío (grados Fahrenheit)
- Humedad
- Presión (pulgadas)
- Visibilidad
- Dirección del viento
- Velocidad del viento (millas por hora)
- Velocidad de ráfagas de viento (millas por hora)
- Precipitación
- Eventos
- Condiciones

Una vez conocidos los datos con los que se va a trabajar, se necesita realizar un preprocesado exhaustivo de los datos, no solo para poder tenerlos en el formato adecuado (.arff), sino también como parte fundamental para obtener buenos resultados y así poder introducirlos en la plataforma Weka que dispone de varios algoritmos para la posterior predicción.

MINERÍA DE DATOS

¿Qué es la minería de datos?

"*La minería de datos comprende una serie de técnicas, algoritmos y métodos cuyo fin es la explotación de grandes volúmenes de datos de cara al descubrimiento de información previamente desconocida y que pueda ser empleada como ayuda a la toma de decisiones.***"**

La minería de datos es un proceso analítico diseñado para explorar grandes volúmenes de datos estructurada y almacenada en bases de datos, con el objeto de descubrir patrones y modelos de comportamiento o relaciones entre diferentes variables. Por esto, la minería de datos se utiliza como herramienta de análisis y descubrimiento de conocimiento a partir de datos de observación o de resultados de experimentos. Las

fases de la minería de datos son siempre las mismas, independientemente de la técnica específica de extracción de conocimiento usada.

Estas fases son:

- ***Filtrado de datos***.-

El formato de los datos no suele ser el adecuado para ser usados por los distintos algoritmos, por lo que se deberán filtrar para eliminar datos inválidos, así como ajustarlos a las necesidades del programa. A este filtrado se le suele referir como preprocesado de los datos. También permite la segmentación, es decir, agrupar a todos los objetos parecidos entre sí en base a las propiedades comunes. Las técnicas utilizadas para segmentación son *clustering clásico* y *segmentación neuronal*.

- ***Selección de variables***.-

A continuación del preprocesado, normalmente, se sigue teniendo una cantidad muy elevada de variables, por lo que habrá que elegir las variables que sean más influyentes para nuestro estudio sin que ello conlleve perder información.

- ***Extracción del conocimiento.-***

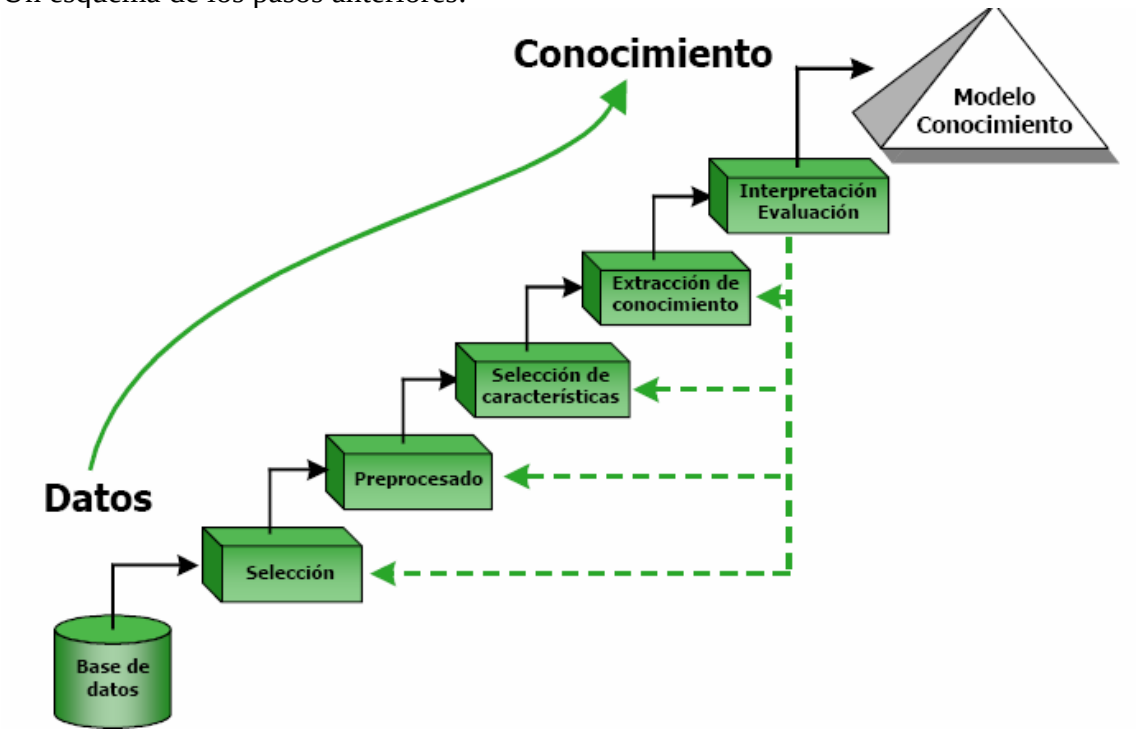
Para conseguir un modelo de conocimiento, se tiene que partir de un algoritmo que represente los patrones observados de las variables o relaciones de asociación entre dichas variables. Con estos algoritmos se puede extraer un modelo de conocimiento bastante preciso con la realidad dependiendo del algoritmo utilizado.

- ***Interpretación y evaluación.-***

Partiendo del modelo, se tiene que comprobar que dicho modelo nos devuelve resultados correctos, es decir, validar dichos resultados. Si la validación no es satisfactoria, habría que repetir los pasos anteriores para obtener los resultados deseados.

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

Un esquema de los pasos anteriores:



Para el desarrollo de esta práctica se seguirán los pasos anteriores para alcanzar una predicción adecuada.

PREDICCIÓN METEOROLÓGICA

- ***PREPROCESADO DE DATOS (Filtrado de Datos)***

Esta es la parte más importante y tediosa de la práctica ya que hay que trabajar con un gran volumen de información y transformarlo en un conjunto más pequeño, manejable y adecuado sin que con ello perdamos información sobre el tema a estudiar.

- Eliminación datos inválidos

Lo primero que se ha realizado es la eliminación de los datos que no eran válidos para el posterior análisis. Estos datos podían ser inválidos tanto por su tipo, estar en una variable nominal y ser de tipo numérico, como también por contener caracteres o símbolos que la plataforma WEKA no entiende (p.ej: hay que sustituir símbolos como “-”, “N/A”, “9999.0”, etc por “?”).

- Agrupación de Datos

Una vez concluida la labor de eliminar los datos inválidos, se observo que la cantidad de información era muy elevada así que se decidió agrupar los datos en subgrupos para poder realizar un análisis separado de cada subgrupo. La decisión fue agruparlos en estaciones, dividiendo así el grueso de los datos en 4 grupos (primavera, verano, otoño, invierno), cada uno de

ellos de 3 meses. De esta forma se pudo reducir el conjunto total de más de 120.000 entradas en 4 subconjuntos de unas 30.000, que era mucho más asequible de manejar.

- Eliminación de filas irrelevantes

Al observar los datos se podía ver que había bastantes filas de información que no disponían de información, en especial en los campos que se iban a considerar importantes, por lo que se tomo la decisión de eliminar dichas filas directamente. Estas eran las que no estuvieran completas (malformadas), y las que no tuviesen información en los campos, Temperatura, Rocío, Presión y Condiciones. Así también pudimos reducir aún más nuestro conjunto de datos, de forma adecuada.

- Formato adecuado (.arff)

Para poder usar la plataforma WEKA los datos tienen que estar presentes de una forma adecuada que es separada por comas para cada variable, y luego debe poseer una cabecera que indique que variable es y de que tipo, tanto nominal como numérico. A su vez, para la realización de esta práctica, como el objetivo es la predicción de tres casos, había que incluir datos para estos casos, para que se pudiese aprender también de los datos anteriores. Esto se hizo añadiendo 3 variables nuevas que son temperatura+1hora, temperatura+24horas y condiciones+24horas, que sólo son la replicación de los datos en los lugares adecuados. Una vez obtenidos los datos de forma correcta se le añade la siguiente cabecera:

@relation NOMBRE_RELACION
@attribute r1 real
@attribute r2 real ...
...
@attribute i1 integer
@attribute i2 integer
…
@attribute s1 {v1_s1, v2_s1,…vn_s1}
@attribute s2 {v1_s1, v2_s1,…vn_s1}
…
@data

- Eliminación de Datos por falta de memoria física

A pesar que WEKA es una herramienta muy potente, al introducir todos los datos no fue capaz de procesarlos por falta de memoria físca, por lo que se tuvo que tomar la decisión de reducir el conjunto de datos. La opción elegida finalmente fue reducir el conjunto a la mitad, tomando una fila de cada dos. De esta forma se consigue reducir los datos y se pierde poca información ya que como los datos están tomados cada media hora, suelen estar bastante correlados a los datos tanto anterior como posterior, por lo que

se sigue teniendo información sobre ese instante de tiempo, a pesar de que para la predicción interesa la correlación para ser más exactos.

Una vez realizados los pasos anteriores, los datos ya están preprocesados y preparados para su análisis.

➢ *SELECCIÓN DE VARIABLES*

Este apartado es útil no sólo para reducir un poco más el conjunto de datos con los que trabajamos sino porque no todos los datos de los que se dispone son relevantes para el conocimiento. Para nuestro conjunto de datos partimos de 18 variables:

Año, Mes, Día, Hora, Temperatura (grados Fahrenheit), Punto de rocío (grados Fahrenheit), Humedad, Presión (pulgadas), Visibilidad, Dirección del viento, Velocidad del viento (millas por hora), Velocidad de ráfagas de viento (millas por hora), Precipitación, Eventos, Condiciones, Temperatura+1, Temperatura+24, Condiciones+24.

Al haber tantas variables el estudio se hace mucho más complejo innecesariamente ya que no todas las variables proporcionan la misma información para la predicción de la temperatura y condiciones, por lo que en nuestro caso se ha optado reducir el número de variables a 7 siendo las elegidas las siguientes:

- Temperatura
- Punto de Rocío
- Presión
- Condiciones
- Temperatura+1H
- Temperatura+24H
- Condiciones+24H

➢ ***ALGORITMOS (Extracción del Conocimiento)***

Conjuntive Rule

Con este algoritmo se crean un conjunto de reglas a partir de las cuales se trata de predecir. Las regla es una regla conjuntiva sencilla, es decir, realiza la conjunción de los antecedentes del atributo a predecir.

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

- ***Decision Tree***

En este caso, se crea un árbol de decisión para predecir, siendo los nodos intermedios los atributos de ejemplos presentados, as ramas sus posibles valores y las hojas los resultados.

- ***REPTree***

Es un algoritmo basado en un árbol de decisión que aprende mediante decisión rápida. Construye un árbol de decisión usando la información de varianza y lo poda usando como criterio la reducción del error. Solamente clasifica valores para atributos numéricos una vez. Los valores que faltan se obtienen partiendo las correspondientes instancias.

- ***KStar***

KStar es un clasificador basado en instancias, esto significa que la clasificación de una instancia está basada en la clasificación de instancias de entrenamiento similares, determinadas por alguna función de similitud. Se diferencia de otros aprendizajes basados en lo mismo en que usa una función de distancia basada en entropía.

- ***J48***

Es un algoritmo basado en un árbol de decisión que implementa un algoritmo extendido del ID3. El ID3 es un algoritmo de aprendizaje por inducción que pretende modelar los datos mediante un árbol, llamado árbol de decisión.

- ***Multilayer Perceptron***

El perceptron multicapa es una red neuronal con varias capas ocultas de neuronas que utiliza como función de aprendizaje la propagación hacia atrás. Tiene dos variables que son el número de neuronas y el número de capas con las que hay que jugar para obtener el resultado adecuado.

- ***LeastMedSq***

Este algoritmo implementa una regresión lineal “least median square” para calcular los coeficientes de la función con la que estima el parámetro.

- ***Linear Regression***

Este tipo de algoritmos devuelve una función, es decir una regresión lineal que servirán para predecir valores numéricos.

- ***IB1***

Esta basado en los K vecinos más próximos, y en concreto, este algoritmo predice según el vecino mñas cercano.

- ***IBK***

Este algoritmo es el genérico de K vecinos, y según el número de vecinos clasifica y predice. Variando el número de vecinos se obtienen resultados diferentes.

- ***Bagging***

Es un algoritmo tipo META, esto quiere decir que está basado en el tipo de clasificador que se utilice. Los resultados dependerán mucho de si el calsificador es bueno o no para nuestro modelo, no sólo del algoritmo.

- **Conjunto de Validación**

Los resultados anteriores han sido probados con los diversos algoritmos y clasificadores para obtener un resultado adecuado de predicción para los tres casos estudiados. Esto se ha completado enseñando al algoritmo con los datos de los años 1996 hasta el 2003 y luego se han validado dichos resultados con los datos obtenidos en el año 2004. En la plataforma WEKA esto se hace incluyendo el conjunto de validación con un porcentaje en el “Test set”, que para nuestro caso es aproximadamente el último 11% de los datos son los que corresponden con los datos obtenidos del año 2004.

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

- **RESULTADOS**

TEMPERATURA +1

- **LeastMedSq:**

=== Predictions on test split ===

inst#,	actual,	predicted,	error
1	60.8	61.289	0.489
2	62.6	61.06	-1.54
3	39.2	40.871	1.671
4	44.6	43.473	-1.127
5	35.6	36.783	1.183
6	32	31.142	-0.858
7	42.8	48.869	6.069
8	30.2	33.082	2.882
9	62.6	64.528	1.928
10	39.2	36.602	-2.598
11	44.6	48.418	3.818
12	50	49.12	-0.88
13	37.4	36.903	-0.497
14	44.6	42.284	-2.316
15	42.8	42.736	-0.064
16	60.8	58.827	-1.973
17	39.2	37.548	-1.652
18	57.2	64.624	7.424
19	33.8	33.98	0.18
20	37.4	38.671	1.271
21	44.6	45.488	0.888
22	42.8	41.671	-1.129
23	42.8	43.647	0.847
24	42.8	41.573	-1.227
25	51.8	51.642	-0.158

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9628
Mean absolute error	1.7647
Root mean squared error	2.5332
Relative absolute error	24.3491 %
Root relative squared error	27.3172 %
Total Number of Instances	825

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

- **Linear Regression:**

Time taken to build model: 51.96 seconds

=== Predictions on test split ===

inst#,	actual,	predicted,	error
1	60.8	60.789	-0.011
2	62.6	60.511	-2.089
3	39.2	40.314	1.114
100	57.2	56.75	-0.45
101	50	46.541	-3.459
102	33.8	38.953	5.153
103	28.4	29.129	0.729
104	44.6	44.316	-0.284
105	46.4	46.345	-0.055
106	57.2	58.929	1.729
107	37.4	37.97	0.57
108	33.8	32.445	-1.355
109	44.6	46.283	1.683
110	46.4	44.384	-2.016
111	39.2	40.74	1.54
112	48.2	46.339	-1.861
113	48.2	48.199	-0.001
114	41	41.589	0.589
115	60.8	57.069	-3.731
116	44.6	45.777	1.177
117	41	46.415	5.415
118	50	48.205	-1.795
119	44.6	46.562	1.962
120	48.2	48.104	-0.096

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9637
Mean absolute error	1.7345
Root mean squared error	2.4697
Relative absolute error	23.9325 %
Root relative squared error	26.6331 %
Total Number of Instances	825

- **IBK: (3 vecinos)**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9025
Mean absolute error	2.3617
Root mean squared error	4.2349
Relative absolute error	32.5867 %
Root relative squared error	45.6688 %
Total Number of Instances	825

- **Bagging:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9613
Mean absolute error	1.8111
Root mean squared error	2.5529
Relative absolute error	24.9898 %
Root relative squared error	27.53 %
Total Number of Instances	825

- **AdditiveRegression:**

=== Classifier model (full training set) ===

Additive Regression
ZeroR model
ZeroR predicts class value: 46.43916700040055
Base classifier weka.classifiers.trees.DecisionStump
10 models generated.
Model number 0
Decision Stump

Classifications
TemperaturaF <= 47.3 : -6.3702306303607035
TemperaturaF > 47.3 : 8.394434484748052
TemperaturaF is missing : -7.091199429371925E-14

Model number 1
Decision Stump

Classifications
TemperaturaF <= 59.9 : -0.9697806755051356
TemperaturaF > 59.9 : 10.345837767187945
TemperaturaF is missing : -1.6440371068698353E-15

Model number 2
Decision Stump

Classifications
TemperaturaF <= 36.5 : -6.240289149745015
TemperaturaF > 36.5 : 1.0671996763828488
TemperaturaF is missing : 4.980604842656943E-15

Model number 3
Decision Stump

Classifications
TemperaturaF <= 47.3 : 1.7796396457023036
TemperaturaF > 47.3 : -2.3451377633187307
TemperaturaF is missing : -3.5257991812338305E-15

Model number 4
Decision Stump

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

Classifications
TemperaturaF <= 40.1 : -2.4464677754609885
TemperaturaF > 40.1 : 0.8425649068753854
TemperaturaF is missing : 1.4274168995821396E-15

Model number 5
Decision Stump

Classifications
TemperaturaF <= 68.9 : -0.1607556076701166
TemperaturaF > 68.9 : 7.509437112437094
TemperaturaF is missing : -1.507508278904985E-15

Model number 6
Decision Stump

Classifications
TemperaturaF <= 47.3 : 0.8001474318121924
TemperaturaF > 47.3 : -1.0544022005223155
TemperaturaF is missing : 4.758056331940466E-16

Model number 7
Decision Stump

Classifications
TemperaturaF <= 52.7 : -0.6191821942457815
TemperaturaF > 52.7 : 2.0237202656374946
TemperaturaF is missing : 1.4216071622219331E-16

Model number 8
Decision Stump

Classifications
TemperaturaF <= 29.299999999999997 : -4.756999426142775
TemperaturaF > 29.299999999999997 : 0.11778616910148311
TemperaturaF is missing : -8.074349270001139E-17

Model number 9
Decision Stump

Classifications
TemperaturaF <= 47.3 : 0.7085658285580569
TemperaturaF > 47.3 : -0.9337196360856332
TemperaturaF is missing : -2.7359120323829113E-16

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9363
Mean absolute error	2.5516
Root mean squared error	3.2543
Relative absolute error	35.207 %
Root relative squared error	35.0939 %
Total Number of Instances	825

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

- **REPTree:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9363
Mean absolute error	2.5516
Root mean squared error	3.2543
Relative absolute error	35.207 %
Root relative squared error	35.0939 %
Total Number of Instances	825

- **M5P:**

M5 pruned model tree:
(using smoothed linear models)

```
TemperaturaF <= 47.3 : LM1 (4259/30.732%)
TemperaturaF >  47.3 :
|  TemperaturaF <= 56.3 : LM2 (2137/36.237%)
|  TemperaturaF >  56.3 :
|  |  TemperaturaF <= 65.3 : LM3 (807/40.829%)
|  |  TemperaturaF >  65.3 :
|  |  |  TemperaturaF_+24 <= 70.7 : LM4 (203/32.519%)
|  |  |  TemperaturaF_+24 >  70.7 :
|  |  |  |  TemperaturaF <= 70.7 :
|  |  |  |  |  Nivel_de_Rocio <= 40.1 : LM5 (8/375.661%)
|  |  |  |  |  Nivel_de_Rocio >  40.1 : LM6 (14/26.736%)
|  |  |  |  TemperaturaF >  70.7 : LM7 (63/24.192%)
```

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9629
Mean absolute error	1.7636
Root mean squared error	2.5015
Relative absolute error	24.3342 %
Root relative squared error	26.9753 %
Total Number of Instances	825

- **Decision Table:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.9486
Mean absolute error	2.2084
Root mean squared error	2.9288
Relative absolute error	30.4712 %
Root relative squared error	31.5841 %
Total Number of Instances	825

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

TEMPERATURA +24

- **IBK: (K=3)**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.816
Mean absolute error	4.1707
Root mean squared error	5.4265
Relative absolute error	57.2344 %
Root relative squared error	58.435 %
Total Number of Instances	825

- **IBK: (K=5)**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.8323
Mean absolute error	3.9695
Root mean squared error	5.159
Relative absolute error	54.4743 %
Root relative squared error	55.5549 %
Total Number of Instances	825

- **IBK: (K=20)**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.8398
Mean absolute error	3.8559
Root mean squared error	5.0375
Relative absolute error	52.9151 %
Root relative squared error	54.2468 %
Total Number of Instances	825

- **Decisión Table:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.8409
Mean absolute error	3.8927
Root mean squared error	5.0248
Relative absolute error	53.4201 %

Root relative squared error 54.1096 %
Total Number of Instances 825

- **Linear Regression:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient 0.8419
Mean absolute error 3.899
Root mean squared error 5.0114
Relative absolute error 53.5068 %
Root relative squared error 53.9652 %
Total Number of Instances 825

- **Additive Regression:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient 0.8184
Mean absolute error 4.2344
Root mean squared error 5.3317
Relative absolute error 58.1095 %
Root relative squared error 57.4147 %
Total Number of Instances 825

- **REPTree:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient 0.8447
Mean absolute error 3.7941
Root mean squared error 4.978
Relative absolute error 52.0664 %
Root relative squared error 53.6059 %
Total Number of Instances 825

- **M5P:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient 0.8564
Mean absolute error 3.722
Root mean squared error 4.7938
Relative absolute error 51.0779 %
Root relative squared error 51.6218 %

Total Number of Instances 825

- **Bagging:**

=== Evaluation on test split ===
=== Summary ===

Correlation coefficient	0.8564
Mean absolute error	3.6849
Root mean squared error	4.7932
Relative absolute error	50.5681 %
Root relative squared error	51.6157 %
Total Number of Instances	825

CONDICIONES +24

- **Decision Table**

=== Run information ===

Scheme: weka.classifiers.rules.DecisionTable -X 1 -S 5
Relation: Datos_Invierno
Instances: 7491
Attributes: 7
TemperaturaF
Nivel_de_Rocio
PresionIn
Conditions
TemperaturaF_+1
TemperaturaF_+24
Conditions_+24
Test mode: split 89% train, remainder test

=== Confusion Matrix ===

```
  a  b  c  d  e  f  g  h  i  j  k  l  m  n  o  p  q  r  s  t  u  <-- classified as
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  a = Granizo_Leve
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  b = Nubes_de_Polvo
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  c = Caidas_de_Granizo
  0  0  0 237  0  0  0 113  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  d = Despejado
  0  0  0  4  0  0  0  10  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  e = Bruma
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  f = Fuertes_Lluvia
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  g = Tormenta
  0  0  0 70  0  0  0 213  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  h = Nublado
  0  0  0 40  0  0  0  75  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  i = Nubes_Dispersas
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  j = Tormentas_y_Lluvia
  0  0  0  1  0  0  0  13  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  k = Lluvia
  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  l = Lluvia_Leve_Abundante
  0  0  0  5  0  0  0  16  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  m = Niebla
```

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

```
 0  0  0  5  0  0  0 19  0  0  0  0  0  0  0  0  0  0  0  0  0 |  n = Lluvia_Leve
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  o = Abundantes_Lluvia
 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0 |  p = Nieve
 0  0  0  0  0  0  0  2  0  0  0  0  0  0  0  0  0  0  0  0  0 |  q = Nevada_Leve
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  r = Abundante_Nieve
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  s = Niebla_Leve
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  t = Abundante_Niebla
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0 |  u = Nube_en_Embudo
```

- **Conjunctive Rule:**

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         450               54.5455 %
Incorrectly Classified Instances       375               45.4545 %
Kappa statistic                          0.2684
Mean absolute error                      0.0595
Root mean squared error                  0.172
Relative absolute error                 91.1737 %
Root relative squared error             95.5345 %
Total Number of Instances              825
```

- **Decision Table:**

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         469               56.8485 %
Incorrectly Classified Instances       356               43.1515 %
Kappa statistic                          0.3002
Mean absolute error                      0.0556
Root mean squared error                  0.1699
Relative absolute error                 85.2404 %
Root relative squared error             94.3548 %
Total Number of Instances              825
```

- **REPTree:**

```
=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances         450               54.5455 %
Incorrectly Classified Instances       375               45.4545 %
Kappa statistic                          0.2854
Mean absolute error                      0.0548
Root mean squared error                  0.1731
Relative absolute error                 83.955  %
Root relative squared error             96.1123 %
Total Number of Instances              825
```

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	Granizo_Leve
0	0	0	0	0	Nubes_de_Polvo
0	0	0	0	0	Caidas_de_Granizo
0.74	0.284	0.657	0.74	0.696	Despejado
0.071	0.005	0.2	0.071	0.105	Bruma
0	0	0	0	0	Fuertes_Lluvia
0	0	0	0	0	Tormenta
0.625	0.354	0.48	0.625	0.543	Nublado
0.096	0.046	0.25	0.096	0.138	Nubes_Dispersas
0	0	0	0	0	Tormentas_y_Lluvia
0.071	0.004	0.25	0.071	0.111	Lluvia
0	0	0	0	0	Lluvia_Leve_Abundante
0	0.005	0	0	0	Niebla
0.042	0.005	0.2	0.042	0.069	Lluvia_Leve
0	0	0	0	0	Abundantes_Lluvia
0	0	0	0	0	Nieve
0	0	0	0	0	Nevada_Leve
0	0	0	0	0	Abundante_Nieve
0	0	0	0	0	Niebla_Leve
0	0	0	0	0	Abundante_Niebla
0	0	0	0	0	Nube_en_Embudo

- **Part:**

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	398	48.2424 %
Incorrectly Classified Instances	427	51.7576 %
Kappa statistic	0.2209	
Mean absolute error	0.0529	
Root mean squared error	0.1961	
Relative absolute error	80.9924 %	
Root relative squared error	108.8957 %	
Total Number of Instances	825	

Conditions = Nublado AND
TemperaturaF_+24 > 39.2 AND
PresionIn > 29.92 AND
TemperaturaF_+1 > 41 AND
Nivel_de_Rocio > 41 AND
TemperaturaF_+24 > 46.4 AND
TemperaturaF <= 46.4: Nublado (3.0/1.0)

Conditions = Nublado AND
TemperaturaF_+24 > 39.2 AND
PresionIn > 29.92 AND
TemperaturaF_+1 > 41 AND
Nivel_de_Rocio <= 41 AND
PresionIn <= 30.01: Despejado (3.0/1.0)

Number of Rules : 965

Time taken to build model: 303.53 seconds

- **J48:**

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	436	52.8485 %
Incorrectly Classified Instances	389	47.1515 %
Kappa statistic	0.2779	
Mean absolute error	0.0521	
Root mean squared error	0.1872	
Relative absolute error	79.8231 %	
Root relative squared error	103.9396 %	
Total Number of Instances	825	

- **BayesNet:**

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	447	54.1818 %
Incorrectly Classified Instances	378	45.8182 %
Kappa statistic	0.2761	
Mean absolute error	0.0559	
Root mean squared error	0.1729	
Relative absolute error	85.7338 %	
Root relative squared error	95.9989 %	
Total Number of Instances	825	

=== Classifier model (full training set) ===

Bayes Network Classifier
not using ADTree
#attributes=7 #classindex=6
Network structure (nodes followed by parents)
TemperaturaF(5): Conditions_+24
Nivel_de_Rocio(3): Conditions_+24
PresionIn(5): Conditions_+24
Conditions(21): Conditions_+24
TemperaturaF_+1(6): Conditions_+24
TemperaturaF_+24(5): Conditions_+24
Conditions_+24(21):
LogScore Bayes: -64863.95048839756
LogScore BDeu: -63401.000022161694

LogScore MDL: -69225.2190543383
LogScore ENTROPY: -65482.667599969674
LogScore AIC: -66321.66759996966

Time taken to build model: 1.75 seconds

- **KStar:**

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	489	59.2727 %
Incorrectly Classified Instances	336	40.7273 %
Kappa statistic	0.3565	
Mean absolute error	0.0534	
Root mean squared error	0.1634	
Relative absolute error	81.9064 %	
Root relative squared error	90.7228 %	
Total Number of Instances	825	

- **IB1:**

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	Class
0	0	0	0	0	Granizo_Leve
0	0	0	0	0	Nubes_de_Polvo
0	0	0	0	0	Caidas_de_Granizo
0.649	0.282	0.629	0.649	0.639	Despejado
0.071	0.012	0.091	0.071	0.08	Bruma
0	0	0	0	0	Fuertes_Lluvia
0	0	0	0	0	Tormenta
0.406	0.28	0.431	0.406	0.418	Nublado
0.191	0.13	0.193	0.191	0.192	Nubes_Dispersas
0	0	0	0	0	Tormentas_y_Lluvia
0	0.018	0	0	0	Lluvia
0	0	0	0	0	Lluvia_Leve_Abundante
0.143	0.024	0.136	0.143	0.14	Niebla
0.083	0.034	0.069	0.083	0.075	Lluvia_Leve
0	0.002	0	0	0	Abundantes_Lluvia
0	0.002	0	0	0	Nieve
0	0.002	0	0	0	Nevada_Leve
0	0	0	0	0	Abundante_Nieve
0	0	0	0	0	Niebla_Leve
0	0	0	0	0	Abundante_Niebla
0	0	0	0	0	Nube_en_Embudo

=== Evaluation on test split ===
=== Summary ===

Correctly Classified Instances	370	44.8485 %

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

Incorrectly Classified Instances	455	55.1515 %
Kappa statistic	0.1911	
Mean absolute error	0.0525	
Root mean squared error	0.2292	
Relative absolute error	80.492 %	
Root relative squared error	127.2649 %	
Total Number of Instances	825	

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

➢ CONCLUSIONES (INTERPRETACIÓN Y EVALUACIÓN):

Una vez visto los resultados de las diferentes técnicas y algoritmos de aprendizaje sobre nuestros datos, hay que analizar lo observado y obtener conclusiones. Para ello, primeramente se identificarán los algoritmos que mejor resultado han dado para cada una de las tres variables a estudiar. A continuación se compararán los resultados entre variables para ver cuál puede ser predicha con mayor precisión y porqué. Finalmente se tratará de explicar las causas de estas diferencias. Hay que decir que todos los análisis se han hecho con un mismo fichero de datos (*Invierno.arff*), ya que el análisis es análogo para el resto, aunque los resultados pueden variar, dependiendo de cómo de buenos sean los datos para el modelo escogido.

A la vista de los resultados se puede concluir que los algoritmos que mejor resultado han dado son:

Condiciones (+24h):

Algoritmo	Acierto (%)	Fallo (%)
KStar	59.27	40.73
Decision Table	56.85	43.15
REPTree	54.55	45.45
Conjunctive Rule	54.55	45.45
Bayes Net	54.18	46.82
J48	52.85	48.15
PART	48.24	51.76
IB1	44.85	55.15

Temperatura (+24h):

Algoritmo	Error Relativo (%)
Bagging	50.57
M5P	51.08
REPTree	52.07
IBK (K=20)	52.91
Linear Regression	53.51
Decisión Table	53.42
IBK (K=5)	54.47
Additive Regression	58.11

ALVARO GALÁN SANCHEZ 100021823
JULIO DANIEL PÉREZ ORR 100025227

Temperatura (+1h):

Algoritmo	Error Relativo (%)
Linear Regression	23.93
M5P	24.33
LeastMedSq	24.35
Bagging	24.99
Decisión Table	30.47
IBK (K=3)	32.59
REPTree	35.21
Additive Regression	35.21

Lo primero que hay que destacar es lo elevado del error en la mayoría de los casos, puesto que suele acercarse al 50%. Esto puede ser debido a muchos factores, pero principalmente será debido a una mala elección tanto de reglas en los algoritmos aplicados, como de atributos seleccionados para el tratamiento de datos.

Esto quiere decir que al haber reducido el número de atributos, seguramente habrá varios de los eliminados que influyan de manera considerable en la predicción de las variables deseadas. Así pues, también puede ser debido a que, aunque la elección de atributos se hizo pensando que eran los más importantes, puede que varios de estos atributos estén altamente correlados, por lo que la aportación de información de los atributos sea casi nula introduciendo ruido a nuestro análisis, y por tanto un elevado error. Otra de las posibles causas, son los propios algoritmos. Se han analizado algoritmos de muchos tipos (basados en funciones como regresiones, en los K vecinos más próximos, algoritmos basados en clasificadores, etc.), pero puede ser que la elección tanto del propio algoritmo, como de los parámetros que lo componen no haya sido la más correcta, dando lugar a un resultado poco óptimo.

A pesar del elevado nivel de error se pueden sacar algunas conclusiones acerca de los datos y las predicciones obtenidas. Lo primero que hay que destacar es que la predicción de la temperatura una hora más tarde es mucho más precisa que la de la temperatura al día siguiente, lo que indica que efectivamente los atributos escogidos son más adecuados para predecir la temperatura en una hora. Además si nos fijamos en el algoritmo que da la menor tasa de error se trata de la regresión lineal, lo que indica que los datos están relacionados, de algún modo, de forma lineal. Sin embargo, para predecir la temperatura al día siguiente, por lo visto lo mejor es un algoritmo basado en el clasificador utilizado.

Por último, analizando los resultados de la predicción de las condiciones al día siguiente, se ve que al igual que en el caso de las temperaturas, el error es muy elevado, probablemente porque el sesgo hecho a los datos haya truncado una continuidad de los datos y no se correspondan bien unos datos con sus siguientes o predecesores. Además, es probable que el agrupamiento de condiciones realizado también haya influido de

forma negativa, puesto que creará más confusión acerca de cuando ocurrirán ciertas condiciones, como nublado, lluvias o situaciones menos generales como niebla.

MODELO DE PROYECCION DE LA PRODUCCION DE ROSAS, BASADO EN LAS CURVAS DE CRECIMIENTO DE LAS PLANTAS.

JUAN JOSE VILA ARBOLEDA

UNIVERSIDAD DE LA SALLE
FACULTAD DE ADMINISTRACION DE EMPRESAS AGROPECUARIAS
BOGOTA D.C.
2009

MODELO DE PROYECCION DE LA PRODUCCION DE ROSAS, BASADO EN LAS CURVAS DE CRECIMIENTO DE LAS PLANTAS.

JUAN JOSE VILA ARBOLEDA

Trabajo de grado presentado como requisito para optar el titulo de Administrador de Empresas Agropecuarias.

Director
GUSTAVO CORREA ASSMUS
Ingeniero Ambiental.

UNIVERSIDAD DE LA SALLE
FACULTAD DE ADMINISTRACION DE EMPRESAS AGROPECUARIAS
BOGOTA D.C.
2009

DIRECTIVAS

RECTOR: Hno. Carlos Gabriel Gómez Restrepo F.S.C.

VICERECTOR ACADEMICO: Dr. Mauricio Fernández Fernández.

VICERECTOR DE PROMOCION YDESARROLLO HUMANO:
Hno. Carlos Alberto Pabon Meneses.

VICERECTOR DE INVESTIGACION Y TRANSFERENCIA:
Hno. Manuel Cancelado Jiménez.

SECRETARIO: Dr. Patricia Inés Ortiz Valencia.

DECANO DE LA FACULTAD DE CIENCIAS AGROPECUARIAS:
Dr. Luis Carlos Villamil Jiménez.

DECANO DEL PROGRAMA:
Dr. Héctor Horacio Murcia Cabra.

TABLA DE CONTENIDO

Pag.

INDICE DE ANEXOS

Pag.

INDICE DE TABLAS

Pag.

INDICE DE GRAFICAS

Pag.

1 ANTECEDENTES DEL ESTUDIO

1.1 PLANTEAMIENTO DEL PROBLEMA Y JUSTIFICACION

Un modelo fenológico[1] es aquel que permite predecir el tiempo en que ocurre un evento en el desarrollo de un organismo, la fenología hace parte de la meteorología que investiga las variaciones atmosféricas de la vida animal y de las plantas.

El comportamiento de las plantas dentro del invernadero es diferente dependiendo del sitio donde estén situadas dentro de éste, ya que el calor acumulado no es igual en todo el área del invernadero este calor acumulado se conoce como tiempo fisiológico, o en una forma mas técnica como grados-día.

Debido a la situación actual de las empresas floricultoras del país donde la gran mayoría de estas empresas manejan sus modelos de predicción con base en sus datos históricos, al igual debido a las diferentes variaciones del clima y las distintas fechas en el calendario no son una buena base para la toma de decisiones de estos cultivos. Por ello se ha tomado la decisión de investigar y desarrollar un modelo de proyección de rosas a futuro, con el cual las empresas del sector puedan administrar y desarrollar de una manera más acertada sus producciones. El otro motivo que dio origen para el desarrollo de esta investigación es la necesidad que tiene las empresas floriculturas de obtener de manera acertada sus volúmenes de producción para así mismo dar a conocer su oferta real a las comercializadoras para que estas puedan desempeñar sus estrategias de ventas adecuadamente.

De acuerdo con lo anterior, se estudiaron las curvas de crecimiento de las diferentes variedades a las cuales se les quiso hacer dicha predicción, tomando el comportamiento de las plantas para crear el modelo de predicción que arrojara los resultados esperados.

La necesidad que tienen las empresas floricultoras colombianas para encontrar o desarrollar un modelo de predicción o proyección de rosas, que les permita anticipar los volúmenes de producción e implementar las diferentes estrategias administrativas y gerenciales que se manejan dentro de estas empresas de una manera mas acertada es de gran importancia, ya que debido a las variaciones del clima y las distintas fechas en el calendario por las diferentes fiestas que se realizan a nivel nacional o mundial; la manera en que se manejan estos datos actualmente en estas empresas no son una buena base para la toma de decisiones de estos cultivos.

---

[1] Parte de la meteorología que investiga las variaciones atmosféricas en su relación con la vida de animales y las plantas.

El modelo que se está trabajando en Colombia, está diseñado por el tiempo fisiológico de las plantas, esto quiere decir la temperatura que manejan las rosas o grados-día como se le conoce, pero como hay diferentes variaciones en el clima estos modelos no son tan acertados; además de requerir un alto costo de mano de obra y tiempo.

Por lo anterior, se propone desarrollar para las empresas floricultoras del país un nuevo modelo de proyecciones, que obtenga mejores resultados para que estas empresas puedan tomar decisiones oportunas en bienestar de las mismas y del país. El modelo que se presentará está basado en estudios que se realizarán a la planta para determinar su comportamiento y dependiendo de su comportamiento se construirá dicho modelo que arroja los resultados de las predicciones de producción en rosas.

Pregunta de investigación.
¿Un modelo basado en curvas de crecimiento permite gestionar de manera más precisa la producción de rosas?

1.2 MARCO TEORICO Y ESTADO DEL ARTE

En COLOMBIA hay 6.544 hectáreas de cultivos bajo invernadero que producen más de 50 tipos de flor, como rosa, pompón, alstroemeria, clavel, statice, gerbera y tropicales, entre otras especies; un área pequeña, si se compara con otras actividades agropecuarias del país (Asocolflores, 2005). Las rosas modernas (híbridos de té) por lo general son triploides o tetraploides, altamente vigorosas, presentan usualmente una flor única por tallo y cumplen con ciertas características como: tallo largo entre 50 y 90 cm. follaje verde brillante, flores de apertura lenta, colores vivos, buena conservación en florero, resistencia a plagas y enfermedades, altos rendimientos por metro cuadrado y la posibilidad de ser cultivadas a temperaturas no muy elevadas. Esto permite que sean utilizadas en programas extensivos de flor de corte bajo invernadero (Bastidas et al., 2000). Debido a las variaciones del clima, las fechas calendario no son una buena base para la toma de decisiones de manejo del cultivo de la rosa, por lo que se ha venido implementando en los cultivos de flores en Colombia el uso de curvas de crecimiento y la técnica de grados-día, con el fin de predecir con más exactitud el desarrollo de los estadios fenológicos de las plantas y en consecuencia el momento del corte de la flor. Con este trabajo no se refuta la definición o el cálculo de la técnica de grados-día, más bien se propone la evaluación de la producción
de forma directa en campo. Donde el objetivo es determinar el comportamiento de las variedades de rosa ANNA´S, JADE, LIPSTICK, LIGHT ORLANDO, MUSTANG y TINEKE. De acuerdo a la curva de crecimiento que desarrolle cada una de las variedades en el estudio que se realizo en el campo.
“Con el despertar de la inteligencia humana, el hombre hubo de aprender a servirse de la plantas y hubo a empezar también a acumular información respecto a estas, es

cuando aquí el hombre empieza a pensar en ser productivo en terrenos mas pequeños"[2]. El hombre a través de estudios e investigaciones ha demostrado que puede ser más productivo en unidades cada vez mas pequeñas ya que conoce cada vez acerca del tipo de producción que maneja, y su vez aprende mas de su cultivo en una forma completa no solo productiva si no también ambiental, comercial y laboral. "Las rosas son arbustos leñosos con hojas compuestas que brotan en disposición espiral sobre los tallos con respecto a la flor principal. Los brotes o tallos tienen generalmente algunas hojas labiales en la base. La clasificación se basa en el número de flores por inflorescencia, su tamaño, la longitud de los brotes y la forma de la planta"[3]

La clasificación taxonómica de las rosas, dice que pertenecen a la familia Rosaceae, cuyo nombre científico es Rosa sp.

Algunas características para la flor cortada que se utilizan en los tipos híbridos de té presentan largos tallos y atractivas flores dispuestas individualmente o con algunos capullos laterales, de tamaño mediano o grande.

La clasificación de las rosas se realiza según la longitud del tallo, existen pequeñas variaciones en los criterios de clasificación, orientativamente se detallan a continuación:

- Calidad EXTRA: 90-80 cm.
- Calidad PRIMERA: 80-70 cm.
- Calidad SEGUNDA: 70-60 cm.
- Calidad TERCERA: 60-50 cm.
- Calidad CORTA: 50-40 cm.

A. Fragancia

Uno de las características más apreciadas de los rosales es su aroma. Estos alcanzan distintos matices, a limón, afrutado, almizcle, té o su característico olor a rosas.

B. Color

Rosales modernos hay de casi todos los colores. En los Rosales antiguos hay una gama de color más reducida.

El color azul sigue siendo una leyenda ya que los que hay son lilas pálidos.

---

[2] Tomado de: FULLER, Harry. CAROTHERS, Zane. PAYNE, Willard. BALBACH, Margaret. Botánica. 5° ed. 1999. 10 p.

[3] Tomado de: PIZANO, Marta. Cultivo de rosas bajo invernadero: Características botánicas. Hortitecnica Ltda. ed. 2001. 9 p

C. Tipos de flores

- Flores sencillas: 4 a 7 pétalos.
- Flores semidobles: 8-14 pétalos.
- Flores dobles: 15-20 pétalos.
- Flores muy dobles: más de 30 pétalos.

Las plantas son organismos dinámicos que al igual que los animales, poseen la propiedad protoplasmática fundamental de irritabilidad y resultan afectadas en múltiples formas, por consiguiente por los factores cambiantes a sus alrededores, tales como el clima, la temperatura, las propiedades del suelo, Aspectos sanitarios, plagas enfermedades entre otros aspectos que influyen.

La variabilidad de la temperatura dentro de un invernadero determina el índice del desarrollo en los cultivos de flores es decir determina el tiempo para la floración, las temperaturas son variables a lo largo y ancho del invernadero lo que significa que el tiempo para la floración también es variable dentro de este. Esta floración es más rápida en los puntos mas calientes dentro del invernadero, que en los puntos límites donde la concentración de calor es menor. Ya que en los puntos centrales donde hay más calor las plantas acumulan sus grados-día, de una forma mas rápida que en los puntos donde se obtiene menos concentración de calor.

Monroy, Pérez, Cure en la revista de ingeniería de la Universidad de los Andes (2002), realizaron un estudio el cual define, que el comportamiento de varios puntos dentro de un invernadero es diferente, debido a que la concentración de calor y por lo tanto de grados-día, es variable según la zona que se evalué en un invernadero.

En el estudio realizado se observa el resultado de evaluar tres zonas diferentes dentro de un invernadero, para cinco fincas de la Sabana de Bogotá. En la tabla se puede identificar la cantidad de grados-día que se acumuló en cada zona dentro de un número de días justificable, y otra columna en la cual se observa un promedio de grados-día acumulados por un día.

TABLA 1 DISTRIBUCION DE TEMPERATURA EN UN INVERNADERO.

FINCA	GRADOS DÍA			DIAS			PROMEDIO		
	CAMINO	MITAD	BORDE	CAMINO	MITAD	BORDE	CAMINO	MITAD	BORDE
F1	920	950	935	91	97	99	10,1099	9,79381	9,44444
F2	925	940	970	107	108	112	8,64486	8,7037	8,66071
F3	1070	1177	1151	121	139	129	8,84298	8,46763	8,92248
F4	1054	1050	1000	105	107	109	10,0381	9,81308	9,17431
F5	919	975	1026	92	97	108	9,98913	10,0515	9,5

(Fuente de: Monroy, Pérez, Cure. Revista Universidad de los Andes, volumen 14. (2002). 42 p.)

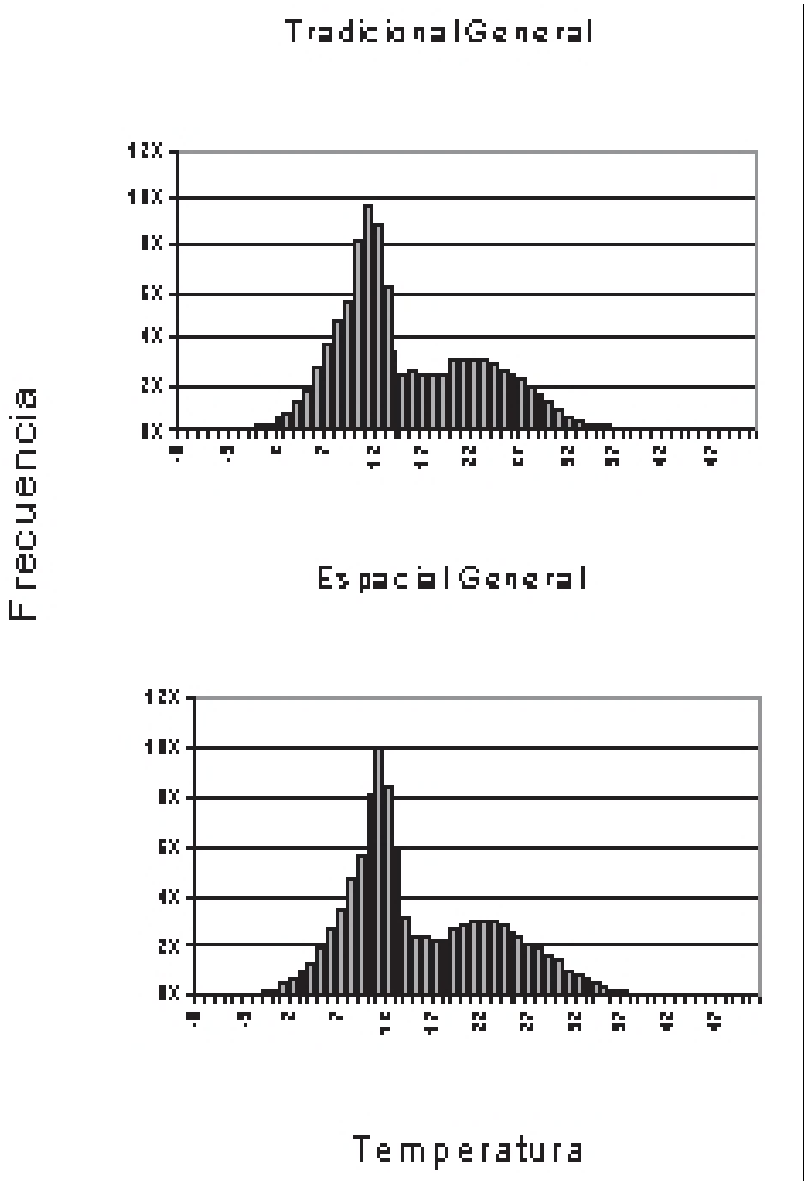
La tabla da a entender que la temperatura en los diferentes puntos dentro de un invernadero es variable; se puede notar que las diferencias entre las tres zonas de los invernaderos son notorias por lo anterior se puede observar que una rosa en la mitad del invernadero necesita menos días para acumular toda su energía que una flor que se encuentra en el borde del invernadero. Otros factores que pueden afectar la acumulación de los grados-día son las razones climáticas, o porque el plástico del invernadero no es el adecuado o simplemente se encuentra sucio y no permite la entrada óptima de los rayos de luz solar.

La opción que más convendría para mantener la temperatura óptima dentro del cultivo de rosas, estaría en implementar tecnología de primera mano, para que esta por medio de censores mantenga la temperatura ideal dentro del cultivo, pero llevar a que esta opción se cumpla es difícil en las fincas productoras de la Sabana de Bogotá por el alto costo de la infraestructura tecnológica.

La otra opción estaría en buscar el diseño y los accesorios ideales para un invernadero que nos lleve a mantener la temperatura buscada o una temperatura cercana a la que se necesita.

De acuerdo con Monroy, Pérez, Cure. En la revista de la Universidad de los Andes (2002), dan a conocer, un estudio que se realizó en cinco fincas de la Sabana de Bogotá, para encontrar si existían diferencias entre los tipos de invernaderos tradicionales y espaciales. Según los resultados a continuación se ve claramente que no existe una gran diferencia a la hora de conservar calor dentro de los dos tipos de invernaderos. La diferencia de temperatura-hora entre los invernaderos tradicionales y espaciales es similar.

GRAFICA 1 ESTUDIO DE VARIABILIDAD EN LA TEMPERATURA[4]



[4] Tomado de: Monroy, Pérez, Cure. Revista universidad de los Andes, volumen 14. (2002). 40 p.

Al observar el gráfico anterior se puede definir que las temperaturas son muy similares y que no existen grandes diferencias.

El estudio establece que entre las horas de la 5:00 PM y las 6:00 AM no hay diferencias entre la temperatura promedio en el interior de los dos tipos de invernaderos. Pero durante las horas del día se encontró que el invernadero espacial obtuvo un grado centígrado de temperatura más respecto a los invernaderos tradicionales.

La intensidad lumínica es otro factor que es importante para tener en cuenta en estos tipos de producciones, estudios realizados dentro de la zona muestran resultados de la cantidad de luz recibida dentro un invernadero y este promedio de luz recibida esta cerca al 70%, pero ésta puede variar debido a la limpieza que se tenga sobre éste o a la calidad del tipo de plástico que se use para llevar a cabo el buen funcionamiento de las plantas dentro del invernadero.

En Colombia existen mas de 7142 hectáreas cubiertas bajo invernadero, de las cuales aproximadamente 6500 hectáreas han estado dedicadas al cultivo de flores de corte.
El área cubierta por invernaderos en el territorio Colombiano dedicada al cultivo de flores, se encuentra distribuida en un 92% en la Sabana de Bogotá, 6% en la zona de Rionegro Antioquia, 1% en el antiguo Caldas y 1 % en la zona del Valle del Cauca.[5]

La Sabana de Bogotá está ubicada a 2600 metros sobre el nivel del mar, por lo tanto su clima hace parte del piso térmico frío, su temperatura media anual es aproximadamente 14° centígrados, la máxima es de 20° centígrados y la mínima es de 5° centígrados. Su precipitación media anual es de 1013 mm, la presión atmosférica de 752 milibares y su humedad relativa anual es de 72%.[6]

En Colombia los invernaderos se han caracterizado por tener dimensiones similares, independiente del modelo estructural o de su uso, las cuales varían muy poco. Su altura mínima está entre 2 y 3 metros, el ancho de la nave se encuentra entre 6.5 y 7 metros, el largo de la nave entre 60 y 70 metros y la pendiente longitudinal entre el 4 y el 8 por ciento.

En cuanto a materiales de cobertura todos los modelos utilizan el polietileno como cerramiento por su bajo costo, su poco peso comparándolos con el peso de los materiales estructurales y porque sus especificaciones técnicas permiten la entrada necesaria de luz y calor.

---

[5] Tomado de: Republica de Colombia. Ministerio de Agricultura y Desarrollo Rural. Estadísticas 2001

[6] Tomado de: htpp://www.atarraya.org/agenda/Bogota.html

La industria de la floricultura Colombiana se ha desarrollado con poca información local sobre las respuestas de las estructuras de los invernaderos a los diferentes aspectos esperados como el tipo de cultivo, maquinaria, equipos entre otras y a los eventos inesperados como borrascas, granizadas, hundimientos, y de mas aspectos que se puedan presentar en una eventualidad.[7]

Es por esto que los invernaderos no han evolucionado por tecnologías o desarrollo humano, en Colombia han evolucionado los invernaderos según las experiencias y las necesidades que han tenido los floricultores que utilizan estos invernaderos para obtener sus producciones.

Debido a este proceso de obtener un buen invernadero, en Colombia se habla de cuatro tipos de invernaderos.

A. Invernadero Tradicional: Es el tipo de invernadero más utilizado en Colombia, debido a su bajo costo, y aunque se le han realizado modificaciones a través del tiempo, continúa en vigencia su topología constructiva. Su geometría se caracteriza por su cercha a dos aguas y su ventilación cenital fija. Este tipo de invernaderos no es muy alto, ya que las columnas o postes aumentan los costos, igualmente ocurre con la cercha ya que la longitud comercial de los elementos, da forma a la estructura.

B. Invernadero Espacial: El invernadero de estructuras espaciales se caracteriza por tener un gran espacio interior libre ya que el plástico se encuentra suspendido en guayas que están sostenidas por postes de concreto. Como no tienen cercha y poseen pocos parales tienen una gran luminosidad y es más fácil el trabajo en su interior, los operarios se pueden subir en las guayas para hacer la limpieza del plástico, debido a su gran altura tiene buena ventilación, y su mantenimiento es mínimo.

C. Invernadero Semitúnel: Los invernaderos semitúnel se caracterizan por la forma circular de la cercha, y por estructura totalmente metálica con soportes de tubos de hierro galvanizado. Poseen gran capacidad para el control de los factores climáticos. Presentan gran resistencia a vientos y su instalación es rápida por estar compuesto de estructuras prefabricadas. Su elevada altura proporciona y facilita la circulación de aire.

[7] Tomado de: ACUÑA, Fabio. ORTIZ, Diana Marcela. Invernaderos la experiencia iberoamericana: Estructuras de invernaderos: la experiencia en Colombia, 1° ed. Almería España (2004). 83-102 p.

D. Invernadero Colgante: Este invernadero se caracteriza por tener los limatones desplazados media nave con respecto a los invernaderos tradicionales. De esta forma no existe cercha como tal, sino listones que se unen a los limatones por medio de tornillos y entre si mediante un ángulo de acero que también sirve como soporte para el canal del desagüe.

Al realizar el modelo de proyecciones las empresas del sector podrán contar con una herramienta útil con la que desarrollarán una mejor labor estratégica dentro de la empresa.

Las empresas floricultoras se verán beneficiadas con este modelo ya que pueden pronosticar volúmenes o cantidades de tallos cosechados en diferentes épocas del año, dado a esto las empresas pueden planear y establecer los volúmenes deseados para satisfacer a los clientes en el tiempo en que éstos las requieran como las fiestas más importantes del mundo, también es importante encontrar una variable que indique en qué épocas se deben hacer las podas para obtener dichos volúmenes deseados y poder entregarlos a tiempo a las comercializadoras para que éstas realicen sus estrategias de ventas.

Algunas de las estrategias que el modelo de proyecciones podría ayudar a definir o complementar serían aspectos como de mano de obra, entrega a tiempo de producción, tiempo de operarios en las diferentes labores y horas extras; son algunos de los aspectos que el modelo de proyecciones puede ayudar a mejorar y ser mas preciso dentro de la empresa.

A. Mano de obra. La mano de obra, que se necesita dentro de una empresa productora de rosas es variable, ya que los volúmenes en las producciones varían, depende del mes y según las ventas que estén programadas. La tabla de proyecciones muestra en cantidades los volúmenes que se obtendrán a futuro. De acuerdo con esto se puede calcular cuántas personas de tiempo fijo se necesitan para suplir todas las necesidades dentro de la empresa, y cuántas personas adicionales en los meses donde las producciones son muy altas, es recomendable para este tipo de unidades productivas mantener actualizada una base de datos la cual diga qué personas están dispuestas a aceptar los cargos temporales.

B. La entrega a tiempo de las producciones, Para el área de poscosecha sería útil conocer estos volúmenes en la producción ya que con estos datos entrarían a organizar su personal para cumplir a tiempo con el pedido que se ha pactado con la comercializadora. La tabla indica qué volumen y qué día se obtendrá dicha

producción, así el área de poscosecha puede estar más organizada y cumplir con todos los objetivos que se trazan para alcanzar las metas dentro de la empresa.

C. El tiempo de los operarios. Teniendo en cuenta que se conocen los volúmenes de producción que están por salir, se puede controlar a los operarios en sus tiempos para que cumplan todas las labores dentro del cultivo, como el corte de la rosas, el control de algunas enfermedades, el arreglo de las podas y el control sanitario dentro del cultivo. Al igual en el área de poscosecha se puede controlar el tiempo de las operarias que realizan la función de boncheo y empaque, esta herramienta puede colaborar para ayudar a organizar algunos procesos del cultivo y permitirle a la empresa ser más productiva.

D. Horas extras. Las horas extras que se pueden dar dentro del cultivo al obtener altas producciones y éstas tengan que ser procesadas para que las flores no tengan ningún cambio ni sufran daños en sus hojas o en su tallo. Al conocer la producción y el tiempo que gastan los operarios para cumplir con sus metas de trabajo se podrían calcular el número de horas extras, sin exceder de las ya estipuladas por el código sustantivo de trabajo.

1.3 OBJETIVOS

- OBJETIVO GENERAL:

Diseñar un modelo que apoye la predicción de cosechas y la gestión de un cultivo a nivel productivo y laboral.

- OBJETIVOS ESPECIFICOS:

- Realizar pruebas de significancía para determinar si el número de muestras es acertado o se tendría que ajustar este número, para que las curvas de crecimiento, sean más precisas.
- Ofrecer a las empresas del sector una herramienta para que puedan desempeñar una mejor labor en sus estrategias gerenciales y laborales.
- Diseñar la metodología que explique el uso indicado para el óptimo funcionamiento del modelo de proyección de rosas.
- Investigar qué tipo de invernadero sería el mas adecuado para desarrollar la practica de cultivos de rosas desde el punto tecnológico y monetario.

1.4 METODOLOGIA

El estudio que se llevará a cabo es objetivo, ya que está basado en datos y fenómenos reales y un sentido práctico que se desarrollará en la construcción del modelo de proyección del cual se obtendrán los resultados esperados. El tipo de estudio es objetivo ya que a partir de algunos datos históricos sobre el tema y alguna idea innovadora, se puede realizar una investigación y llegar a obtener resultados que ayuden al bienestar de la sociedad.

Al comprender este fenómeno que ocurre dentro de los invernaderos y entender que todas las plantas no se desarrollan o crecen de la misma manera, se realizó un estudio[8] para poder observar el comportamiento de las plantas, y así empezar a construir el modelo de proyecciones de rosas a futuro, el cual es una herramienta necesaria para las empresas del sector de la floricultura, el estudio se realizó para las variedades de rosa (JADE, LIGHT ORLANDO, MUSTANG, TINEKE, ANNA´S y LIPSTICK)

El método que se utilizará es analítico o deductivo, ya que en cada fase en que se realizará la investigación está compuesta por pequeños aspectos, los cuáles serán sometidos a varios estudios rigurosos con el fin de identificar los diferentes resultados los cuales permitirán que la matriz del modelo de proyecciones de rosas funcione correctamente.

La investigación se llevará a cabo en dos pasos: un estudio realizado en el campo para determinar la curva de crecimiento de cada variedad y así poder observar cual es el comportamiento de cada variedad, y el segundo paso que será analizar estos resultados, para poder desarrollar el modelo de proyecciones el cual arrojará los pronósticos de las variedades que se estudiaran previamente.

El primer paso de la investigación será desarrollar un estudio en el campo, y este se divide en tres pasos para poder llegar a la construcción del modelo que entregará los resultados de los pronósticos de rosas a futuro.

Para las seis variedades, a las cuales se les realizará el estudio se marcarán 50 pinch[9] o cortes, para determinar la curva de crecimiento a cada variedad, y así mismo el comportamiento de cada una de estas.

[8] Diseñado por el autor en colaboración con el Ingeniero Agrónomo William Cruz.
[9] Corte que se le realiza a la flor al momento de su cosecha o desecho.

Las variedades que fueron sometidas al estudio se escogieron porque en este momento los pronósticos que se llevan dentro de la empresa no son muy acertados, ya que se encontraban entre un 70 y 75 por ciento de acierto en las proyecciones.

Las variedades de rosa a estudiar fueron, (JADE, LIGHT ORLANDO, MUSTANG, TINEKE, ANNA´S, LIPSTICK).
Se desarrollará un formato debidamente diseñado para llevar todas las anotaciones que darán los resultados para desarrollar el modelo de pronósticos de rosas.

El formato estará diseñado para que muestre los resultados del comportamiento que tienen las variedades en un ciclo; el ciclo que va desde el pinch hasta el día en que sean cosechadas. La tabla indicará el número de días en que se demora el ciclo de cada uno de los tallos que se marcaron y cuál será el porcentaje en que se divide cada uno de los días en que se distribuye la curva de crecimiento de cada variedad.

El formato está diseñado para obtener el porcentaje de éxito de la cosecha y el porcentaje de pérdida de esta misma. Este porcentaje de pérdida en la cosecha esta representado por los tallos no activados, tallos ciegos y tallos perdidos. Los tallos no activados son los cuales al realizarles el pinch o corte, no brota la yema que se convertirá después en flor. Los tallos ciegos son los que su brote si es activado pero en algún momento del proceso se detiene su crecimiento y no se desarrolla la flor. Y los tallos perdidos son los que por diferentes razones se les pierde la marca que los distingue de los demás tallos y no se les puede seguir haciendo el seguimiento.

De este formato se obtendrá la curva de crecimiento, ya que el formato indicará los dias transcurridos de las plantas entre el día de su pinch hasta su cosecha, al igual el número de veces que se repite un día y su porcentaje y se podrá realizar una gráfica la cual demuestra el comportamiento de la producción durante su etapa de desarrollo.

Al obtener los anteriores datos, la investigación pasará a desarrollar su segunda parte que se refiere al desarrollo o construcción del modelo de proyecciones, que es el paso más importante ya que de este modelo se obtendrán los resultados, para tomar decisiones gerenciales y laborales más acertadas dentro de las empresas.

El modelo se construirá por medio del programa de Excel. El modelo funcionaría con tres datos importantes que se recolectaron durante el primer paso, Los tres datos ya mencionados son los que alimentan al modelo de proyecciones para que éste arroje los resultados de cada variedad que se desee pronosticar, estas tres variables son las que componen la formula que pronostica los volúmenes de rosas y son las siguientes; establecer el porcentaje de éxito en la cosecha de la variedad o las variedades que se deseen proyectar, la segunda variable es hallar una curva de crecimiento para las variedades que deseen ser proyectadas esta curva de distribución es la variable más importante ya que nos muestra en valores los días en que se varia reflejada la

proyección, y la tercer variable es recoger en la poscosecha el dato diario de una producción real, para que esta sea proyectada, ya que el modelo de proyección inicia con una producción que se allá obtenido dentro del sistema productivo.

Variables que pronostican los volúmenes de una producción.

A. Porcentaje de éxito en la cosecha: este porcentaje se obtiene al realizar un estudio de campo, dicho porcentaje proviene del número de plantas que se le haga un seguimiento o estudio, este porcentaje es la diferencia que hay entre las plantas cosechadas sobre el número total de plantas con las que se inicio el seguimiento. es decir depende del número de plantas a las cuales se les realizó el pinch (corte que se realiza a la flor al momento de cosecharla) y compararlas con el número de rosas que se cosecharon.

B. Curva de crecimiento: el crecimiento de las rosas está determinado por los grados día, estos grados día dependen de la temperatura que se registre dentro de un invernadero; esta temperatura es la que le da a la planta la energía para desarrollar su crecimiento, pero la temperatura no es igual en todos los puntos dentro de un invernadero; es decir que la temperatura de las camas del centro de un invernadero es mayor que la temperatura de las camas que se sitúan a los limites de éste. Al igual sucede con las plantas, las plantas que estén situadas en el centro del invernadero serán cosechadas en un menor tiempo que las que se encuentran en los límites de el, siendo cortadas en el mismo, por este fenómeno se habla de una curva de crecimiento, la cual indicará en qué porcentajes (cantidad de tallos) y el número de días (ciclo de la variedad) en que se distribuirá el ciclo de una variedad.

C. Producción real: está variable esta definida por una producción ya establecida dentro del sistema productivo es decir, que sea un dato histórico.

Al haber obtenido las tres variables, que sustentan el por que de la investigación, se realizo el segundo paso la construcción de la matriz de proyección de rosas a futuro.

La matriz está diseñada para predecir volúmenes de tallos de rosas a futuro, dependiendo de una producción inicial, que se haya realizado dentro del sistema productivo y sea de la variedad que se va a predecir. Y de los resultados de la curva de crecimiento que presentó las variedades a predecir.
Esta matriz se diseñó para que se pueda utilizar en todas las épocas del año y poder tener un mejor control del cultivo y la empresa. El modelo está diseñado para que cada variedad pueda insertar los datos de sus curvas de crecimiento y pueda ser pronosticada de una manera mas acertada.
El modelo se realizó por medio del programa Excel, es un programa elaborado por conocimientos previos del investigador.

2. ANÁLISIS DEL MODELO DE PREDICCIÓN Y MANEJO DEL CULTIVO DE ROSAS

El modelo de predicción y manejo del cultivo de rosas, es un estudio realizado en la Sabana de Bogotá, para diseñar una herramienta a las empresas floricultoras del país ya que al aplicar esta herramienta dentro de sus sistemas productivos, estos tendrán un mejor rendimiento operativo y financiero.

2.1 MODELO DE PREDICCIÓN Y MANEJO DE CULTIVOS DE ROSAS

Este modelo de predicción y manejo de cultivos de rosas, basado en el estudio de grados-día, es una herramienta predictiva que proporciona una información valiosa a los productores de rosa colombianos. Al contar con herramientas que permitan un control del clima al igual que el enfriamiento y calentamiento de los invernaderos, se obtendría un mejor resultado en los estudios que se quieran desarrollar en un futuro.

El modelo de proyecciones basado en los grados-día, es una herramienta que permite obtener datos a un tiempo futuro, pero estos datos que entrega dicho modelo no son tan acertados ya que en este modelo no se está contando con un tiempo anticipado para poder predecir las producciones dentro de estos cultivos, es decir que para poder predecir los datos de una temporada,(desde que se realiza el pinch a la rosa, hasta que sea cosechada), no se cuenta con el tiempo suficiente para darles a las comercializadoras de rosas los datos acertados de las proyecciones. Para poder entregar los datos de las producciones a tiempo a las comercializadoras de rosa, se tendría que tomar datos históricos de las temperaturas de los años anteriores. Pero estos datos no son tan confiables ya que las temperaturas año tras año son diferentes debido al calentamiento global o a los distintos fenómenos como los gases del efecto invernadero.

Para llegar a poder predecir volúmenes de tallos de rosas es necesario tener en cuenta dos factores importantes, los cuales están determinados por el tipo de estudio que se implementó para determinar el crecimiento o comportamiento de las plantas dentro del proyecto ya sea modelo de grados-día, o el modelo por curvas de crecimiento, el otro factor que determina las producciones de rosas son los tallos ciegos (los tallos ciegos, son aquellos que al realizar el corte no se desarrolla ninguna planta). Y esto se define por un porcentaje de pérdida en la cosecha.
Por lo anterior el modelo de predicción y manejo de cultivos de rosas, no tiene gran porcentaje de acierto en sus datos.

2.2 DESCRIPCIÓN TEÓRICA DE SUS COMPONENTES

2.2.1 Qué son los grados-día

Se entiende por grados-día a la diferencia algebraica, expresada en grados, entre la temperatura media de un día del año, y la temperatura de referencia 0 ° centígrado. Estos grados día son acumulados por una estación o periodo, (semanas, meses, años) y en cualquier momento durante el cual, el total puede ser usado como índice del efecto pasado de la temperatura sobre alguna cantidad. El crecimiento de las plantas está determinado dentro de un umbral para que las plantas tengan un óptimo funcionamiento, este umbral está marcado entre 5,3° centígrados y 30° centígrados, si la temperatura se mantiene entre estas medidas las plantas se desarrollarán en perfecto condiciones.[10]

2.2.2 Los cambios en el clima

El clima a lo largo de la existencia mundial siempre ha variado, el problema del cambio climático es que en el último siglo el ritmo de estas variaciones se ha acelerado de manera anómala, a tal grado que afecta ya la vida planetaria. Al buscar la causa de esta aceleración, algunos científicos encontraron que existe una relación directa entre el calentamiento global o cambio climático y el aumento de las emisiones de gases de efecto invernadero (GEI), provocado principalmente por las sociedades industrializadas.

"Un fenómeno que preocupa al mundo es el calentamiento global y su efecto directo, el cambio climático, que ocupa buena parte de los esfuerzos de la comunidad científica internacional para estudiarlo y controlarlo, porque, afirman, que pone en riesgo el futuro de la humanidad.

¿Por qué preocupa tanto? Destacados científicos coinciden en que el incremento de la concentración de gases efecto invernadero en la atmósfera terrestre está provocando alteraciones en el clima.

Coinciden también en que las emisiones de gases efecto invernadero (GEI) han sido muy intensas a partir de la Revolución Industrial, momento a partir del cual la acción del hombre sobre la naturaleza se hizo intensa. "[11]

---

[10]Tomado de: Dirección nacional de meteorología.
htpp ://www.meteorologia.com.uy/glosario_g.htm, Bogota, 2007. 1p

[11] Tomado de: Microsoft ® Encarta ® 2006. © 1993-2005 Microsoft Corporation. Reservados todos los derechos.

A. El efecto invernadero: “El efecto invernadero es un fenómeno natural que permite la vida en la Tierra. Es causado por una serie de gases que se encuentran en la atmósfera, provocando que parte del calor del sol que nuestro planeta refleja quede atrapado manteniendo la temperatura media global en +15º centígrados, favorable a la vida, en lugar de -18 º centígrados, que resultarían nocivos. Así, durante muchos millones de años, el efecto invernadero natural mantuvo el clima de la Tierra a una temperatura media relativamente estable y permitía que se desarrollase la vida. Los gases invernadero retenían el calor del sol cerca de la superficie de la tierra, ayudando a la evaporación del agua superficial para formar las nubes, las cuales devuelven el agua a la Tierra, en un ciclo vital que se había mantenido en equilibrio. Durante unos 160 mil años, la Tierra tuvo dos periodos en los que las temperaturas medias globales fueron alrededor de 5º centígrados más bajas de las actuales. El cambio fue lento, transcurrieron varios miles de años para salir de la era glacial. Ahora, sin embargo, las concentraciones de gases invernadero en la atmósfera están creciendo rápidamente, como consecuencia de que el mundo quema cantidades cada vez mayores de combustibles fósiles y destruye los bosques y praderas, que de otro modo podrían absorber dióxido de carbono y favorecer el equilibrio de la temperatura. “[12] Ante ello, la comunidad científica internacional ha alertado de que si el desarrollo mundial, el crecimiento demográfico y el consumo energético basado en los combustibles fósiles, siguen aumentando al ritmo actual, antes del año 2050 las concentraciones de dióxido de carbono se habrán duplicado Con respecto a las que había antes de la Revolución Industrial. Esto podría acarrear consecuencias funestas para la viva planetaria.

B. El calentamiento global: “Es el aumento de la temperatura de la Tierra debido al uso de combustibles fósiles y a otros procesos industriales que llevan a una acumulación de gases invernadero (dióxido de carbono, metano, óxido nitroso y clorofluorocarbonos) en La atmósfera. Se sabe que el dióxido de carbono ayuda a impedir que los rayos infrarrojos escapen al espacio, lo que hace que se mantenga una temperatura relativamente cálida en nuestro planeta (efecto invernadero). Sin embargo, el incremento de los niveles de dióxido de carbono puede provocar un aumento de la temperatura global, lo que podría originar importantes cambios climáticos con graves implicaciones para la productividad agrícola.”[13]

[12-13] Tomado de:
Microsoft ® Encarta ® 2006. © 1993-2005 Microsoft Corporation. Reservados todos los derechos.

2.3 CALCULO DE GRADOS-DÍA

El método de grados-día, se basa en el análisis de un denominado valor de temperatura base, que es el punto en el cual debe existir un equilibrio dinámico entre el objeto o sistema en estudio y el ambiente.

Según Monroy, Cure, Pérez (2003) para ajustar los datos por medio del modelo de grados-día, se toman los datos originales de temperatura durante 24 horas, se establece un rango que va entre 5,3 y 30° centígrados, los umbrales inferior y superior respectivamente, en este rango las plantas de rosa tendrán un buen funcionamiento fisiológico y por fuera de estos umbrales, las plantas no tendrán un desarrollo significativo. Se hace el promedio para obtener la lectura diaria, que sumada a través del tiempo, representa el acumulado de la temperatura.

En las áreas de fenología y desarrollo de cultivos, el concepto de unidad calórica, medida en grados-día de crecimiento ha mejorado ampliamente en la descripción y predicción de los eventos fonológicos, la ecuación que se desarrolla para determinar los grados-día es la siguiente.

GDC= (Tmax+Tmin) /2-Tbase.

Donde:
GDC, es el cálculo de los grados-día.

Tmax, es la temperatura máxima registrada en el día.

Tmin, seria la temperatura mínima que se dio en el día.

Tbase, que es la temperatura cuando el proceso no progresa es decir 0 ° centígrado.

Es decir que si la temperatura mayor registrada en un día es de 16° centígrados y la temperatura menor registrada fue de 6° grados centígrados, los grados-dia acumulados de ese día se calculan con la formula ya establecida.

GDC= (Tmax+Tmin) /2-Tbase.

GDC= (16+6) /2-0.

GDC= 11

Los grados dias de crecimiento que acumuló la planta durante las 24 horas del día serian 11° grados-día.

2.4 RESULTADOS DEL MODELO GRADOS-DÍA

Para dar justificación al modelo de grados-día, se realizó un estudio en las fincas de la Sabana de Bogotá en el cual se estudiaron cuatro variedades de rosa, el desarrollo de estas variedades de rosa no es igual ya que todas las variedades de rosa tienen un crecimiento fisiológico diferente.

Lo anterior demuestra que hay que realizar un estudio de comportamiento, grados-día o de cualquier otro modelo que se desee implementar dentro de estos cultivos, para las diferentes variedades de rosa que requiera o a las cuales se necesite realizar las proyecciones a futuro.

Las distintas variedades que se estudiaron en el modelo de predicción y manejo en el cultivo de rosas, basado en los grados-dias de las plantas, se demuestran en unas tablas las cuales muestran el número de dias en que se desarrollan las plantas dependiendo de la temperatura promedio que se registre en la temporada.
Para determinar cuantos grados-día necesita cada variedad para que cumpla su ciclo, se realizara un estudio en el cual se le diseñó un seguimiento a cada variedad para obtener el resultado de cuántos dias necesitan las variedades para acumular todos sus grados-día.

Al encontrar el resultado de los dias y la cantidad de grados-día que requieren las plantas para cumplir con todo su desarrollo, diseñaron una tabla la cual a través de una temperatura promedio en una temporada predice las producciones de una variedad.
Al realizar el seguimiento de la variedad Alsmeer Gold, se encontró que la planta necesita de 717 grados-día para cumplir con su ciclo de desarrollo, en la Sabana de Bogotá. El día en que esta planta sea cosechada depende de la temperatura promedio que se registre en la temporada donde la planta acumulará todo el calor hasta que cumpla con todo su desarrollo.
Para la variedad Alsmeer Gold, los resultados fueron los siguientes.

TABLA 2. GRADOS-DÍA ACUMULADOS.

Temp. promedio	Numero de dias a cosecha a partir de:			
	corte	brote	arroz	color
13	93	85	46	19
14	82	75	41	17
15	74	67	37	15
16	67	61	33	14
17	61	56	30	13
18	56	51	28	12
grados-día	717	651	357	150

(Fuente de: Monroy, Pérez, Cure. Revista universidad de los Andes, volumen 15. (2003). 20 p.

La tabla da a conocer que la variedad Alsmeer Gold necesita de 717 grados-día para que cumpla con todo su ciclo de desarrollo y pueda ser cosechada, el día en que sea cosechada esta la planta depende de la temperatura promedio que se registre dentro de la temporada, (desde que se le realiza el pinch a la rosa hasta que esta es cosechada).

Es decir que si la temperatura promedio registrada durante el tiempo en que la planta cumple con todo su ciclo de desarrollo, acumula los 717 grados-dias, es de 15° Centígrados, el número de dias en que se obtendra la cosecha de esta planta es de 74 dias, la planta necesitara de 67 dias desde que esta presente su brote hasta que sea cosechada, cuando la planta se encuentre en el estado arroz y presente la temperatura promedio de 15° centígrados le faltaria 37 dias para que la planta cumpla con todo su ciclo de desarrollo, y presentando esta misma temperatura la planta solo necesitaria de 15 dias en pasar de boton color a ser cosechada.

Los cuatro estados de la flor, que se muestran en la tabla son los estados donde la planta sufre los cambios mas notorios entre todos los estados que se tienen para diferenciar los cambios que se obtienen en el desarrollo de las rosas.

3. DESARROLLO DEL MODELO DE PROYECCIONES, BASADO EN LAS CURVAS DE CRECIMIENTO DE LAS PLANTAS

Este modelo se desarrolló para que las empresas productoras de rosas de la Sabana de Bogotá, contaran con una herramienta con la cual, dichas empresas podrán desarrollar diferentes tareas administrativas y gerenciales ya que con esta herramienta obtendrán los cálculos de sus producciones y a si mismo podrán organizar sus labores productivas y administrativas a un mejor nivel. El estudio se realizó en dos partes, la primera consistió en un estudio de campo para determinar el comportamiento de las rosas dentro de esta zona y la segunda parte, en la cual se desarrolló la construcción del modelo de proyecciones.

El modelo de proyecciones es una herramienta modificable, ya que cada empresa que desee implementar este modelo tiene que elaborar su estudio de campo ya que cada empresa o invernadero tiene situaciones y elementos diferentes. Por está razón al elaborar éste estudio el modelo de proyecciones se ajusta a las condiciones de cada empresa.

Con este modelo de proyecciones también se pueden realizar algunos cálculos para determinar producciones sobre la capacidad instalada que maneja cada una de las empresas que implemente el modelo ya que al obtener el porcentaje de éxito en la cosecha se pueden programar las diferentes técnicas como podas o cortes para obtener grandes producciones, dependiendo de la capacidad instalada de las empresas o los pedidos de rosa que se requieran realizar en los meses donde la demanda de rosas es mayor y las empresas puedan obtener mayores utilidades dentro de esta.

3.1 DEFINICIÓN TEÓRICA DE LAS CURVAS DE CRECIMIENTO EN LAS PLANTAS

Debido a las variaciones del clima y otros factores que inciden para tomar decisiones acertadas en las empresas floricultoras, se ha venido implementando dentro de estos cultivos el uso de curvas de crecimiento y la técnica de grados-día, con el fin de predecir con más exactitud el desarrollo de los estados fenológicos de las plantas. Y así mismo, adelantarse a las tareas que se necesitan para manejar un cultivo de rosas.

Las curvas de crecimiento, es un seguimiento que se les realiza a las plantas para definir cuál es el comportamiento de su desarrollo. Es decir que, en las condiciones normales de los cultivos que se tienen dentro de cada empresa, se trata de encontrar cuál es el comportamiento del desarrollo que muestra cada variedad de rosas distribuidas dentro del invernadero, y así poder determinar las curvas de crecimiento para cada variedad que desee ser proyectada a futuro.

La curva de crecimiento es el tiempo (días) en que se demora la planta en completar todo su desarrollo o que acumula toda su energía (grados-día), El ciclo de desarrollo de las plantas se cuenta desde que se realiza el pinch, al tallo de rosas hasta que la flor es cosechada. La curva de crecimiento es la distribución de los días, en que se demora la muestra (número de rosas estudiadas) en completar su ciclo de desarrollo, la distribución se mide en porcentaje para observar el comportamiento de cada variedad.

Es decir, que en una muestra de una variedad "X" de cinco plantas, cada una representaría el 20%, al realizar el seguimiento de dicha muestra se tomará el número de dias en que se demora las plantas en completar todo su ciclo. Si al final de realizar el estudio a dicha muestra se encuentra que la planta A necesitó de 70 dias para desarrollarse, la planta B requirió de 71 dias, la C y la E necesitaron 72 dias y la D exigió 73 dias para cumplir su desarrollo, se podría concluir que el comportamiento de la cosecha para la variedad a la cual se le realizo el estudio presentaría su distribución de la siguiente forma.

El primer 20% de la producción se obtendría el día 70, igual porcentaje se conseguirá en el día 71, el pico de la producción se dará en el día 72 con el 40% de la producción, y el día 73 se obtendrá el 20% faltante.

De esta manera se puede observar el comportamiento de las variedades de rosas y poderlas proyectar a futuro, para tener un mejor manejo administrativo y gerencial dentro de las empresas.

Para poder observar cómo es el comportamiento de cada variedad y así poder analizarlo dentro de una gráfica, se realizo un estudio de campo. El día 11 de noviembre de 2007, mientras se cosechaba la producción, se marcaron cincuenta tallos de cada variedad para poder determinar el comportamiento de cada variedad y así poder proyectarlas en el modelo. Al realizar el corte para cosechar una rosa, de ese mismo corte se inicia la activación para producir un nuevo tallo o rosa.

Se realizó un seguimiento diario las primeras semanas, para observar cuál era el comportamiento de las variedades, a las cuales se les realizo el estudio y así determinar su comportamiento y a su vez desarrollar la curva de crecimiento exacta de cada una de estas variedades.

Se diseñó un formato debidamente formulado en el programa Excel (Anexo B). El cual al introducir los datos del estudio que se realizó previamente en campo, dé a conocer cuál es la distribución de cada curva de crecimiento para así poder proyectarlas dentro del modelo de proyecciones.

3.2 RECOLECCIÓN DE DATOS Y DESARROLLO DE LA CURVA DE CRECIMIENTO

Al realizar la recolección de datos se diseñó un formato (Anexo A), en el cual se encuentran los datos de cada variedad desde el día en que se les realizó el pinch hasta el día en el cual fueron cosechados, pasando por algunos momentos importantes en las diferentes etapas del desarrollo de una rosa.

Al obtener los resultados del primer formato, y adquirir los datos del día en que se les realizó el pinch, y el día en que fueron cosechadas las plantas que lograron completar todo su ciclo de desarrollo, se construyó un segundo formato (Anexo B) que se consistió en una tabla formulada en el programa Excel, para determinar la distribución de la curva de crecimiento de cada variedad. Este formato muestra el día en que se le realizó el pinch a cada rosa, hasta el día en el cual estas rosas fueron cosechadas, es necesario insertar los datos de pérdidas de tallos ya sean tallos ciegos o tallos a los cuales se les perdió el rastro y no se les pudo hacer todo el seguimiento.

Al insertar los datos dentro del formato es importante, tener en cuenta que los tallos ciegos sí inciden en el comportamiento de la planta ya que estos tallos dan el resultado del éxito de la cosecha que es otra de las variables para el funcionamiento de este modelo. Los tallos perdidos no inciden en el resultado final ya que éstos no cuentan porque no se sabe cuál es el destino final del tallo pero disminuye las probabilidades de obtener datos claros y acertados, lo recomendable es realizar un seguimiento en el cual al marcar estos tallos se note, y desarrollar comunicación dentro del personal del cultivo para obtener toda la información posible sobre el comportamiento en el campo de la variedades de rosas a estudiar.

3.3 FORMULA DEL MODELO DE PROYECCIONES

Al alcanzar los datos del primer pasó y averiguar cuál es el comportamiento de cada variedad a proyectar, se encaminó a realizar el segundo paso de esta investigación; está compuesto por la construcción del modelo de proyecciones, este modelo se construyó en el programa Excel y está compuesto por tres variables. De estas tres variables depende la cosecha de una producción de rosas, y estas tres variables componen la siguiente formula.

Cosecha = (PA * % de EC) * CdC

PA= producción anterior.

% de EC= porcentaje de éxito en la cosecha

CdC= curva de crecimiento

La cosecha de una producción de rosas depende de estos tres factores, ya que los factores climáticos o las labores agronómicas que se llevan dentro de estos cultivos están ya evaluadas en el estudio de las curvas de comportamiento ya que este paso se tiene que realizar en las condiciones normales de un cultivo.

A. Producción anterior.: Para llegar a proyectar una producción de rosas a futuro es necesario tener en cuenta la producción anterior de la variedad. Esta producción original dada por el cultivo en las condiciones normales es la que se proyectará, los dias en que se obtendrán los resultados de esta producción depende de la distribución que se haya calculado previamente en la curva de crecimiento de la variedad que se va a proyectar. Esta información es fácil de acceder ya que en todas las empresas floricultoras se llevan los datos históricos de las producciones que se han logrado dentro de ésta. Todas las empresas, cuentan con formatos o programas diferentes para registrar y guardar estos datos de producción. Se recomienda tener un formato o programa en el que se registren las producciones obtenidas diariamente (Anexo C), si se necesitan datos como producciones semanales o mensuales, depende de cómo lo quiera manejar cada empresa. Se necesitan los datos diarios de las producciones ya que el modelo de proyecciones fue diseñado para los 365 dias del año.

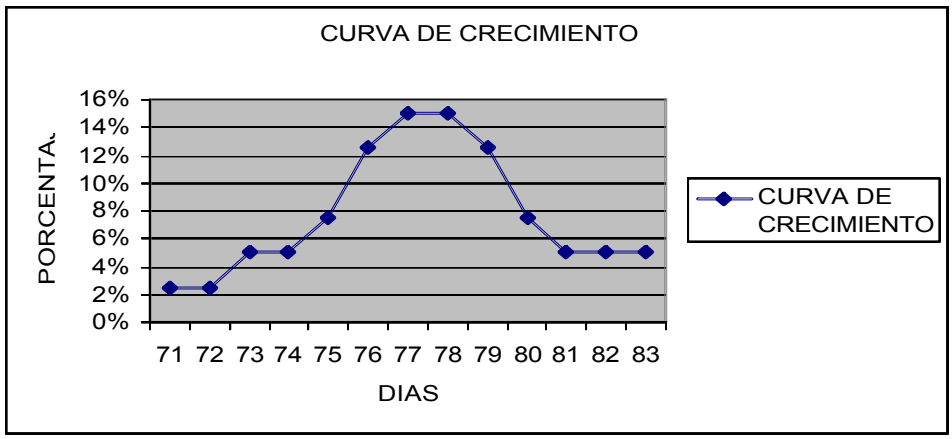
B. Porcentaje de éxito en la cosecha: El formato de recolección de datos (Anexo B) también está diseñado para arrojar el dato de éxito en la cosecha. Ya que al haber realizado el seguimiento en campo de las variedades que se estudiaron, se puede observar cuántos tallos de la muestra fueron cosechados y cuántos de estos tallos no se activaron o se quedaron en tallos ciegos. Al insertar la información dentro del formato de recolección de datos (Anexo B), ya sea los datos de la fecha de recolección o los tallos ciegos o no activados el programa de Excel por medio de las diferentes formulas que lo componen, entregará automáticamente el resultado del porcentaje de éxito en la cosecha. Este resultado lo presenta el formato en la parte inferior en la casilla de color azul, como se puede observar en el recuadro de este anexo.

C. Construcción de las curvas de crecimiento como variable para el modelo de proyecciones: Como se ha relatado anteriormente la curvas de crecimiento son la distribución de los días en que se demora la muestra (número de rosas estudiadas) en completar su ciclo de desarrollo, y al haber insertado todos los datos estudiados en campo en el formato de recolección (Anexo B), en las columnas 6 (% DÍA) y 7 (DISTRIBUCION) se encontrarán los porcentajes y los dias en que la variedad estudiada se distribuye.

Al observar los resultados y teniendo un poco de conocimiento del programa Excel, se pueden aplicar estos resultados para construir una gráfica y observar el

comportamiento que expresa cada variedad para asi poder proyectarla en el modelo de proyecciones

GRAFICA 2 CURVA DE CRECIMIENTO.



Fuente: Elaborada por el autor

Con esta curva de crecimiento se puede observar cómo es el comportamiento de la planta. Esta gráfica indica qué día y en que porcentaje se distribuyó la muestra de rosas estudiadas. Se puede observar que la muestra se distribuyó en doce dias obteniendo su pico de producción entre los dias 77 y 78.
Para poder realizar las proyecciones de rosas se ha tomado como variable estas curvas de crecimiento, ya que al comprender cómo se comportan estas plantas dentro del cultivo se puede realizar por medios matemáticos una matriz que arroje los resultados de las proyecciones de rosas.

3.4 CONSTRUCCIÓN DEL MODELO DE PROYECCIONES, BASADO EN LAS CURVAS DE CRECIMIENTO

Para la construcción del modelo de proyecciones, se desglosó la fórmula expuesta anteriormente, en el programa Excel y las variedades estudiadas fueron proyectadas. Como se ha estando relatando el modelo de proyecciones necesita de tres variables para que arroje los resultados de las proyecciones. Las tres variables que se analizaron anteriormente, se estudiaron para poder construir el modelo de proyecciones.

A continuación se presenta un ejemplo real con la variedad de rosa JADE, para entender cómo se construyó y como funciona este modelo.

- Producción anterior.

Como ya se sabe esta producción fue dada por el cultivo en condiciones normales, estas producciones establecidas se distribuirán de acuerdo a la curva de crecimiento y así serán proyectadas. En el (Anexo C) se encuentra la producción de la variedad JADE de los dias 3, 4 y 5 de enero, y esta fue la siguiente:
Enero 3= 480 rosas.
Enero 4= 580 rosas.
Enero 5= 520 rosas.

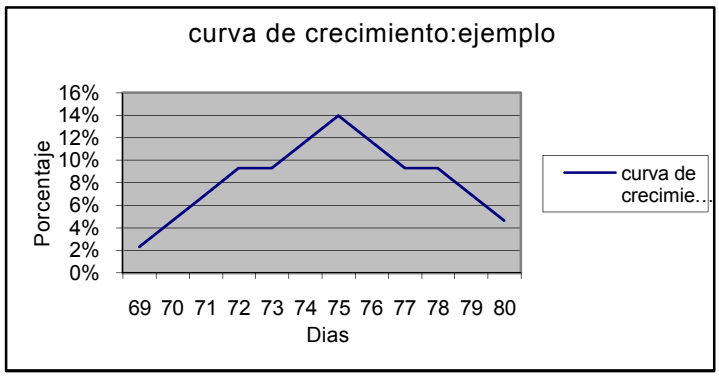
- ❖ Porcentaje de éxito en la cosecha.

Para hallar el porcentaje de éxito de la cosecha, se deben introducir los resultados de la investigación realizada en campo dentro del formato de recolección de datos (Anexo B). Para entender este proceso se registraron los datos obtenidos por la variedad JADE, y se registraron los siguientes resultados, que se pueden observar en el (Anexo D). Para la variedad JADE el porcentaje de éxito en la cosecha fue de 86%, lo anterior quiere decir que de los 50 tallos a los cuales se les realizó el seguimiento 43 tallos fueron cosechados, los 7 restantes se quedaron en tallos ciegos o simplemente no fueron activados.

- ❖ Curvas de crecimiento.

Para diseñar la curva de crecimiento de cada variedad y encontrar el comportamiento de cada variedad se deben introducir los datos de la investigación efectuada en el campo en el formato de recolección de datos (Anexo B). Para el ejemplo que se está realizando con la variedad JADE que arrojó los resultados que se pueden observar en el (Anexo D), se diseñó la siguiente gráfica para observar como es el comportamiento de esta variedad.

GRAFICA 3 EJEMPLO.



Fuente: Elaborada por el autor.

Como se puede observar en el (Anexo D) o en la gráfica anterior, la variedad tiene una duración entre 69 y 80 dias de cosecha obteniendo su pico más alto en el día 75 con un porcentaje del 14% sobre la producción a proyectar. Al obtener el resultado de las tres variables y haber entendido de dónde proviene cada uno de estos se elaboró la

siguiente tabla, de la cual se obtendrán los resultados de las proyecciones de las variedades que se deseen estudiar.

TABLA 3 MODELO DE PROYECCIONES.

VARIEDAD	N° de invernadero o Bloque					
Semana	1 (2007)					
Día			3	4	5	
Produccion			480	580	520	

Éxito de la cosecha	86%

Distribucion dias	69*	70	71	72	73	74	75	76	77	78	79	80
Porcentaje	2%	5%	7%	9%	9%	12%	14%	12%	9%	9%	7%	5%

Proyeccion / Dias	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	18-Mar	19-Mar	20-Mar	21-Mar	22-Mar	23-Mar	24-Mar	25-Mar
03-Ene	8	21	29	37	37	50	58	50	37	37	29	21		
04-Ene		10	25	35	45	45	60	70	60	45	45	35	25	
05-Ene			9	22	31	40	40	54	63	54	40	40	31	22
Total Proyectado	8	31	63	94	113	135	158	173	160	136	114	96	56	22

Fuente: elaborado por el autor

Al introducir los datos en el modelo tal como se muestra en el anterior ejemplo, éste automáticamente arroja el dato de las proyecciones por medio de la fórmula que se dio anteriormente.

Cosecha= (PA * % de EC) * CdC

Según los dias en que se ha distribuido en la curva de crecimiento, se puede observar que la producción del 3 de enero de 2007, empieza a ser cosecha nuevamente entre el 12 de marzo y el 23 de marzo, que son los dias en que la curva de crecimiento tuvo su comportamiento 69 y 80 dias, de duración.

La fórmula que se diseñó para la producción del día 3 de marzo en el programa fue la siguiente.

Cosecha 12 Marzo= [(480 * $86%$) * 2%] [14]

Cosecha 12 Marzo= 8,256 rosas.

---

* Entre el día 3 de Enero han transcurrido 69 dias, que es donde empieza a repicar la producción el 12 de Marzo.

[14] El signo “$” se utiliza en el programa Excel para dar un valor fijo a los diferentes números que se utilicen en una fórmula.

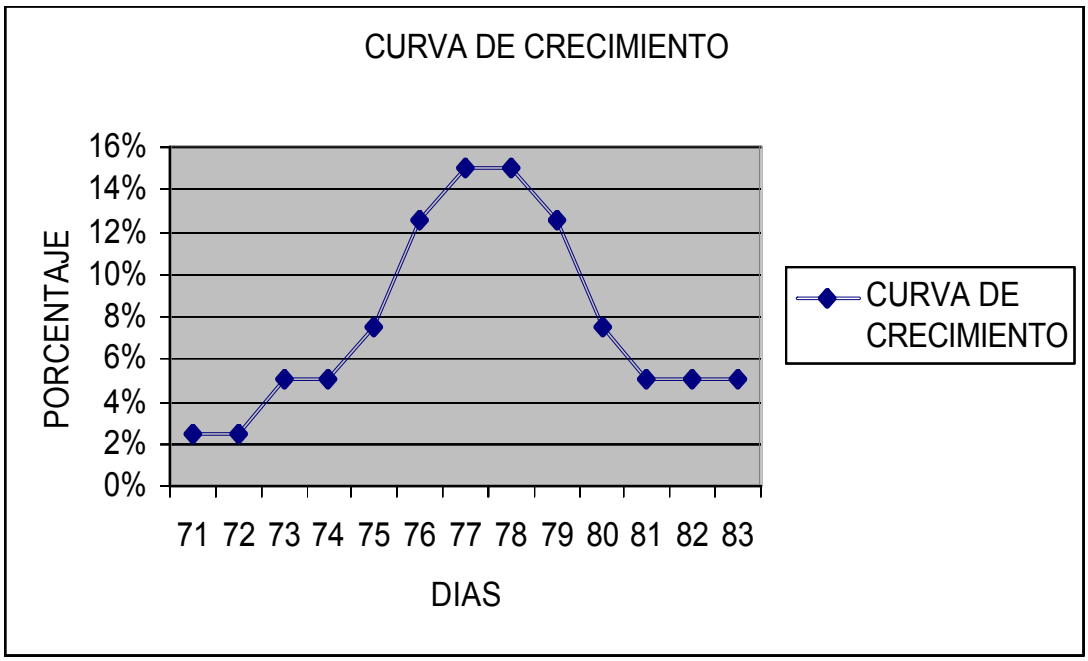
Toda la tabla está diseñada para que arroje los datos de acuerdo a los valores que se obtengan de las tres variables con las cuales funciona el modelo de proyecciones.

3.5 RESULTADOS DEL ESTUDIO

Como resultado de toda la investigación se obtuvo el modelo de proyecciones (Anexo 3), basado en las curvas de crecimiento.

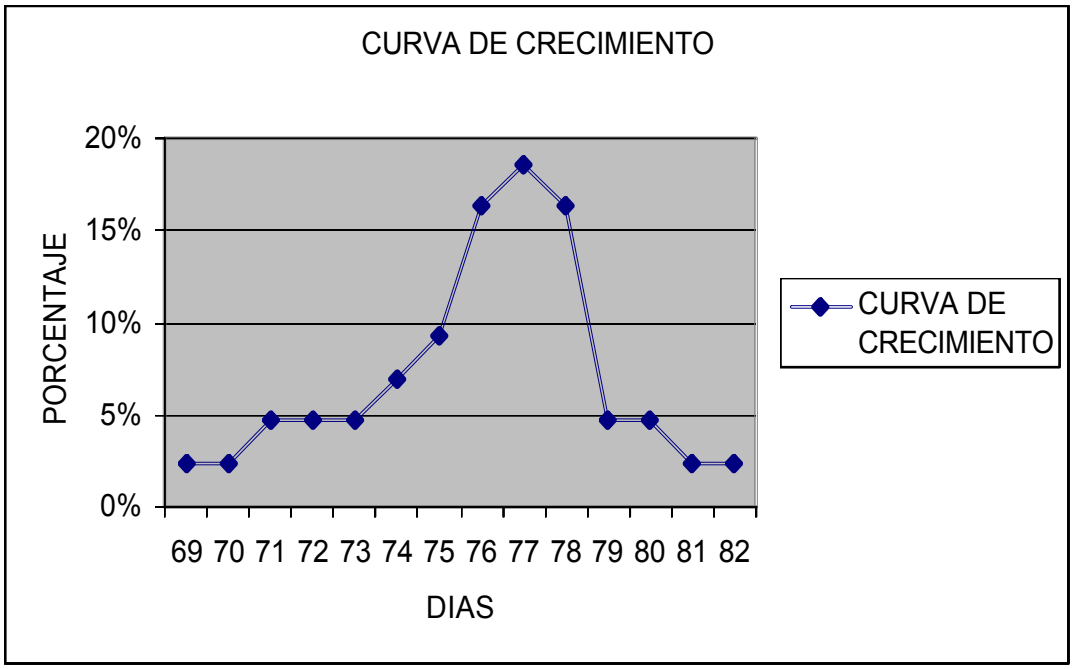
Durante el estudio realizado se obtuvieron los resultados de las cinco variedades de rosa estudiados.
Esto resultados son diferentes para cada variedad ya que todas no tienen el mismo comportamiento.

GRAFICA 4 RESULTADOS VARIEDAD: TINEKE.



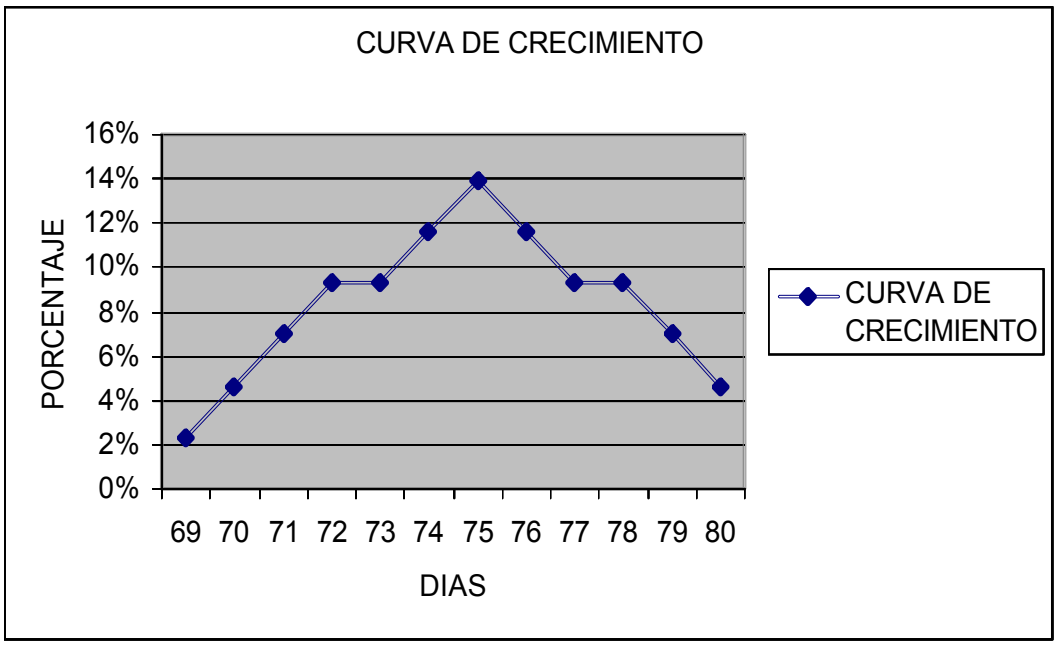
Fuente: Elaborada por el autor

GRAFICA 5 RESULTADOS VARIEDAD: MUSTANG.



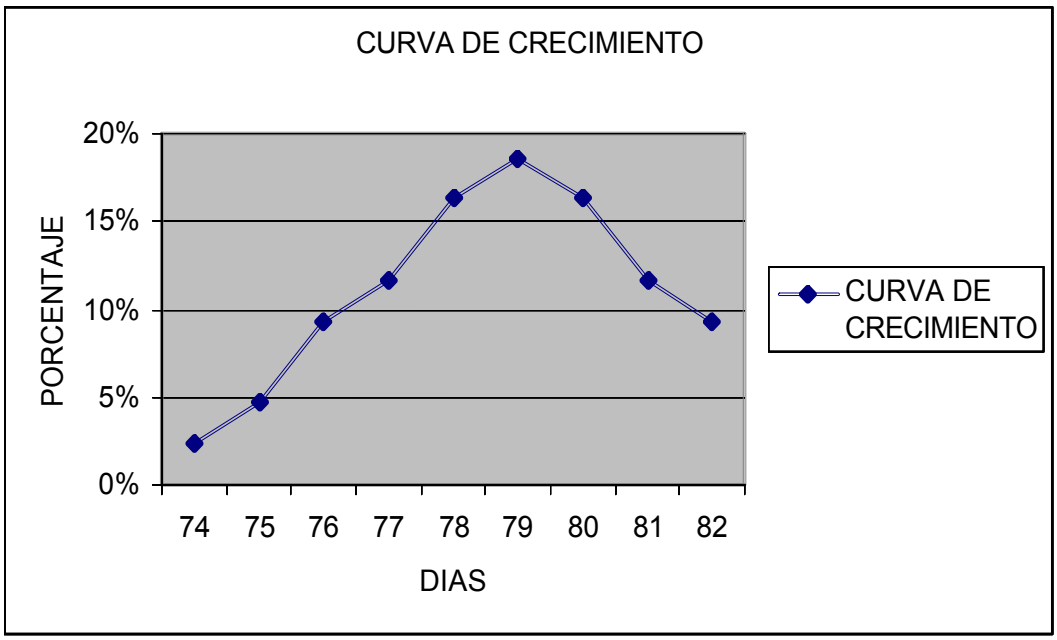
Fuente: Elaborada por el autor

GRAFICA 6 RESULTADOS VARIEDAD: JADE



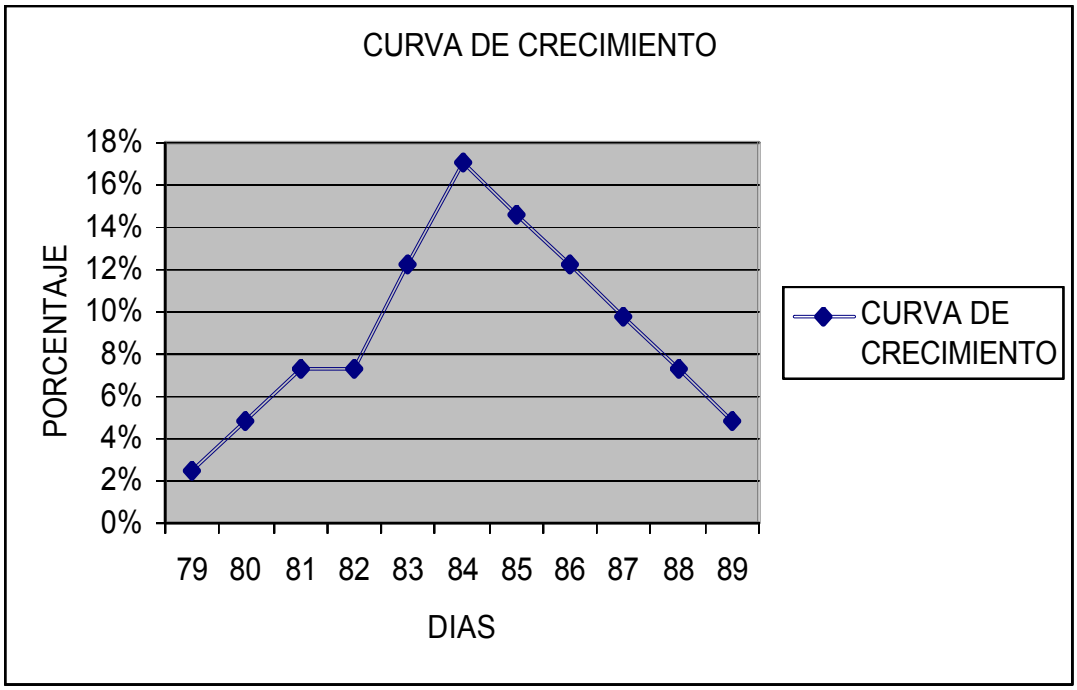
Fuente: Elaborada por el autor

GRAFICA 7 RESULTADOS VARIEDAD: LIGHT ORLANDO



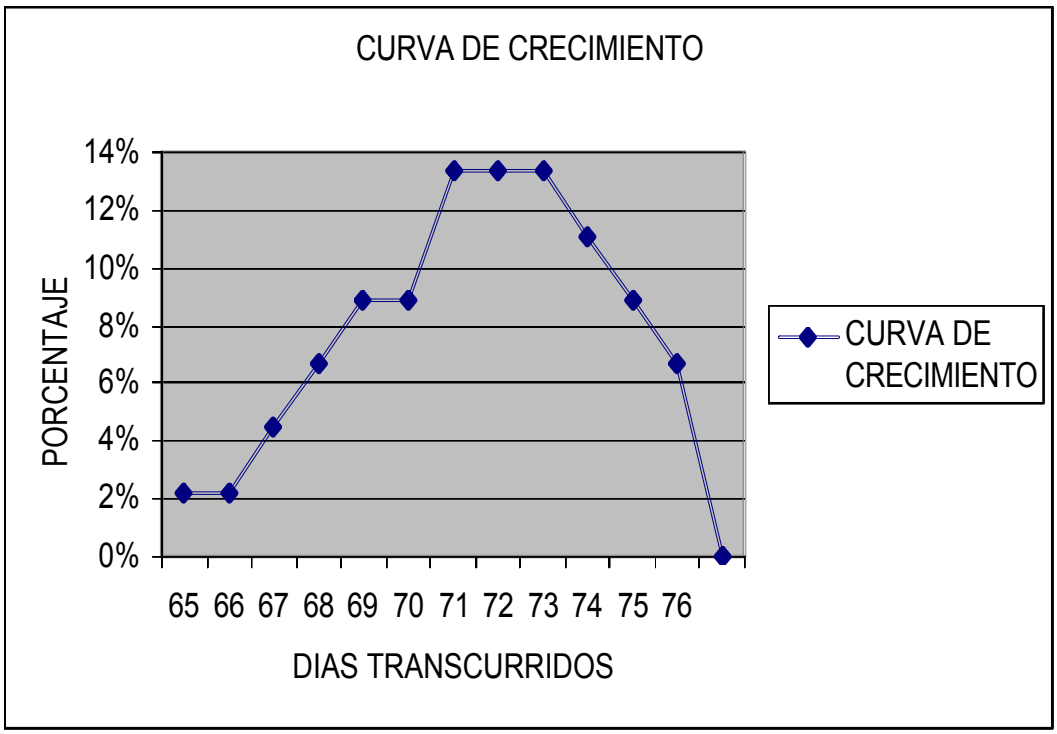
Fuente: Elaborada por el autor

GRAFICA 8 RESULTADOS VARIEDAD: LIPSTICK.



Fuente: Elaborada por el autor

GRAFICA 9 RESULTADOS VARIEDAD: ANNAS



Fuente: Elaborada por el autor

Al observar los resultados de las curvas de crecimiento del estudio realizado se puede apreciar que todas las variedades de rosas tienen un comportamiento diferente. Se puede diferenciar que los dias en que se cosechan las plantas de cada una de las variedades y los porcentajes en que se reparten en esta distribución son diferentes para cada variedad que se estudio.

Para dar un uso excelente del modelo de proyecciones es necesario que al realizar el estudio de campo se tomen todas las observaciones requeridas para lograr obtener un óptimo resultado de las curvas de crecimiento y al proyectar las producciones de la empresa éstas sean más exactas y de gran ayuda.

Al realizar las proyecciones con los resultados que se obtuvieron del estudio realizado en campo, se pudo expresar que los porcentajes de acierto de estas proyecciones se mantuvieron en un rango entre el 85,29% y el 117,29%, los resultados evaluados fueron las producciones desde la semana 13 hasta la semana 19 del 2008. Al término de estas

semanas de evaluación el porcentaje de acierto en las variedades que fueron evaluadas se obtuvo un resultado de éxito del 101,04%.
En la evaluación que se desarrolló para observar la efectividad del modelo, se realizó una tabla para demostrar el resultado de las producciones reales de estas semanas comparadas con los resultados que arrojó el modelo de proyecciones.
Para dar un ejemplo de donde se obtienen los resultados que se van a presentar, se presenta tabla (proyección de la variedad Tineke durante la semana 13), y este dato se comparo o se evalúo con el resultado real que se obtuvo en esas semanas en la empresa (Anexo E).

TABLA 4 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 13 del 26 de marzo al 1 de abril.

variedad	Producción real semana 13	Proyecciones de la semana 13	%de acierto.
anna´s	12908	12607	97,66%
jade	3696	3447	93,26%
light orlando	5554	5745	103,44%
lipstick	8204	7392	90,10%
mustang	3864	3681	95,26%
tineke	6440	6099	94,70%
total	40666	38971	95,83%

Fuente: Elaborado por el autor.

TABLA 6 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 14 del 2 de abril al 8 de abril.

variedad	Producción real semana 14	Proyecciones de la semana 14	%de acierto.
anna´s	14964	13872	92,70%
Jade	4984	4251	85,29%
light orlando	5628	6197	110,11%
Lipstick	7336	6942	94,62%
mustang	5060	5863	115,86%
Tineke	6776	5944	87,72%
Total	44748	43069	96,24%

Fuente: Elaborado por el autor.

TABLA 7 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 15 del 9 de abril al 15 de abril.

variedad	Producción real semana 15	Proyecciones de la semana 15	%de acierto.
anna´s	17360	16775	96,63%
Jade	5852	6344	108,41%
light orlando	8792	8985	102,20%
Lipstick	5740	6733	117,29%
mustang	9324	9045	97,007%
Tineke	9268	9220	99,48%
Total	56336	57102	101,35%

Fuente: Elaborado por el autor.

TABLA 8 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 16 del 16 de abril al 22 de abril.

variedad	Producción real semana 16	Proyecciones de la semana 16	%de acierto.
anna´s	22596	21201	93,82%
Jade	8008	8117	101,36%
light orlando	14252	14563	102,18%
Lipstick	9016	9360	103,81%
mustang	12376	13137	106,15%
Tineke	10920	11412	104,51%
Total	77168	77790	101,22%

Fuente: Elaborado por el autor.

TABLA 9 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 17 del 23 de abril al 29de abril.

variedad	Producción real semana 17	Proyecciones de la semana 17	%de acierto.
anna´s	22792	24772	108,68%
Jade	13496	12168	90,16%
light orlando	18060	17429	96,58%
Lipstick	11424	10870	95,15%
mustang	14644	12639	86,31%
Tineke	13636	13085	95,96%
Total	94052	90963	96,71%

Fuente: Elaborado por el autor.

TABLA 10 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 18 del 30 de abril al 6 de mayo.

variedad	Producción real semana 18	Proyecciones de la semana 18	%de acierto.
anna´s	13580	12579	92,62%
Jade	4632	5033	108,66%
light orlando	7852	7465	95,07%
Lipstick	6048	6950	114,91%
mustang	5217	4523	86,69%
Tineke	9584	8155	85,08%
Total	46913	44705	95,29%

Fuente: Elaborado por el autor.

TABLA 11 Resultado de las proyecciones del modelo VS la producción real del cultivo, durante la semana 19 del 7 de mayo al 13 de mayo.

variedad	Producción real semana 19	Proyecciones de la semana 19	%de acierto.
anna´s	12572	13448	106,96%
Jade	3332	2985	89,59%
light Orlando	5124	4572	89,22%
Lipstick	4592	5125	111,61%
mustang	4003	3408	85,13%
Tineke	6745	5888	87,29%
Total	36368	36380	100,003%

Fuente: Elaborado por el autor

Los resultados proyectados se evaluaron con los datos obtenidos dentro de la empresa en dichas semanas (Anexo E), Estos datos fueron prestados por la base de datos de la poscosecha de la empresa en donde están las producciones obtenidas de todas las variedades que se cultivan.

TABLA 12 Porcentaje de acierto del estudio, transcurridas las 7 semanas de evaluar las proyecciones del modelo VS las producciones reales que se obtuvieron en las respectivas semanas.

Variedad	Promedio del acierto al completar las 7 semanas de evaluación.
Annas	95,83%
Jade	96.24%
Light Orlando	101,35%
Lipstick	101,22%
Mustang	96,05%
Tineke	93,53%
Total acierto del estudio	97,37%

Fuente: Elaborado por el autor.

El modelo de proyecciones en algunos resultados es mayor del 100%, ya que esta trabajando con un margen de casi el 17% que es permisible en la cosecha de esta clase de cultivos. Este 17% se puede ajustar para que en otros estudios disminuya y obtenga un margen de error mas corto pero con el mismo porcentaje de acierto en las producciones ya que este fue del 97,37%, al entregar una proyección en la producción total de las 7 semanas evaluadas de 391.980 Rosas, comparada con la producción total del cultivo durante esas semanas que fue de 396.251 Rosas. Al aumentar el número de muestras en el estudio realizado en campo, se podría realizar con mayor exactitud la construcción de las curvas de crecimiento, para que estas curvas permitan que el desarrollo del modelo de proyecciones sea preciso y se pueda disminuir el margen del 17% que se presento en este estudio.

En conclusión, las proyecciones realizadas por el modelo después de 7 semanas de evaluación el resultado promedio se sitúo en 97,37%, las proyecciones aumentaron positivamente en mas del 15%, ya que se estaba trabajando dentro de la empresa con un porcentaje de acierto del 75 %. pero al observar que los resultados dentro de estas semanas de evaluación los resultados están en un rango entre 85,05% y 117,91%, esta gran diferencia da a entender que las curvas de crecimiento no están mostrando el verdadero comportamiento de las plantas en algunos momentos, es necesario realizar con más cuidado el estudio de campo ya que el comportamiento de las plantas puede ajustarse mejor para que este rango sea mas corto pero que tenga la misma efectividad en el resultado ya que la producción real de la 7 semanas fue de 396.251 rosas y el total de rosas proyectadas fue de 391.980 rosas para un acierto del 98,92%. Lo que

demuestra que el modelo tiene alta efectividad, pero se tendría que realizar el estudio de campo mas estrictamente para obtener mejores resultados en las curvas de crecimiento y así mismo los resultados de las proyecciones obtendrían un margen menor al 17% que se desarrollo en este estudio.

4 INTERPRETACIÓN ADMINISTRATIVA DEL USO DEL MODELO DE PROYECCIONES

Las empresas del sector floricultor al implementar el modelo de proyecciones, obtendrán un acierto más elevado sobre sus producciones de rosas a futuro. Construyendo y analizando algunos indicadores de gestión sobre las diferentes tareas que se llevan a cabo dentro de estos cultivos tanto en el área de campo como en el área de poscosecha, se puede mejorar en el aspecto administrativo y mejorar los recursos con los que cuenta cada empresa.

Como el modelo de proyección de rosas es una herramienta que pronostica las producciones de rosas de cada empresa, obteniendo estos resultados la empresa también puede realizar pronósticos en las áreas laborales y comerciales. Con ayuda de los indicadores de gestión se pueden programar diferentes tareas aplicando los principios administrativos de planear controlar dirigir y ejecutar, al tener claro el funcionamiento del modelo y ayudarlo con los principios administrativos se puede ser mas eficiente en los diferentes sectores de la empresa, al igual el modelo sirve de ayuda para programar la compra de materia prima, los dias de despacho de rosas, como también puede desempeñarse como una guía para la contratación de personal, entre otras tareas que se pueden llegar a programar obteniendo las producciones de rosas anticipadas.

4.1 DESPACHO DE PRODUCCIÓN DE ROSAS

El despacho en las producciones de rosas es un paso importante ya que se tiene que cumplir con todos los requisitos que plantean las comercializadoras para la entrega de rosas y por supuesto las producciones que se han planeado meses antes para la venta.

Las comercializadoras piden las producciones con varios meses de anticipación para que éstas puedan cumplir con sus estrategias de ventas y puedan salir de estas producciones rápido ya que en algunos meses puede ser difícil su venta.
Para poder realizar los despachos de estas producciones de rosas es necesario contar con un cuarto frió, ya que este se requiere para almacenar dichas rosas mientras se cumple con el total del pedido que se pactó con la comercializadora meses atrás. El cuarto frió debe tener el tamaño necesario para poder almacenar una alta producción dentro de cada empresa, es decir el área de almacenamiento debe estar relacionado con la productividad de rosas en el campo.

Con el modelo de proyección de rosas a futuro, se pueden observar algunas de estas variables para que así se pueda tener una perspectiva mejor y se pueda cumplir con las entregas a tiempo del producto. Al observar la cantidad de rosas producidas en un tiempo futuro, la empresa puede ponerse de acuerdo con la comercializadora, para

decidir que dias y en que horario se puede recoger la cantidad de rosas producidas que pactó meses atrás, ya que la empresa al conocer su producción se puede anticipar a todas las tareas que se realizan en estas unidades productivas así mismo la empresa puede estar mejor organizada y se puede contar con mas tiempo para manejar mejor una situación si se llegara a presenta algún imprevisto.

4.2 MANO DE OBRA

La mano de obra de los cultivos de rosas, depende del área que se ha sembrado para obtener dichas producciones en cada empresa. Como en los cultivos de rosas se dan los picos de producción que son las fiestas más importantes nacionales y mundiales, como el día de san Valentín y el día de la madre, en estos dias las producciones de rosas en los cultivos son mucho mas altas que las que se pueden obtener en otros meses o dias del año, por eso se habla de que en los cultivos de rosas hay que contratar mano de obra a largo y a corto plazo.

El modelo de proyección de rosas a futuro, muestra cual es la cantidad y en que meses o dias se registran los picos de sobreproducción y los meses o dias en que estas producciones mantienen constantes. Lo mas recomendable es calcular los tiempos estimados que gastan los operarios en cumplir con todas las tareas que se requieren en todas las labores para la producción de rosas. En los meses donde estas producciones son constantes, al evaluar a los operarios para obtener tiempos estimados, se puede calcular cuantas personas se necesitan contratar para realizar todas las tareas del cultivo. Y así mismo poder calcular cuanto es la mano de obra extra que se necesita para los picos de sobreproducción.

Al haber realizado los seguimientos a los operarios y obtener los tiempos estimados de cuanto tiempo requieren para realizar sus actividades, se construyen los indicadores de gestión. Que determinen cuantas personas se necesitan para cubrir la producción o capacidad instalada de la empresa o cuantas personas se necesitan para cubrir una producción constante, así mismo estos indicadores ayudan a controlar las funciones de todo el personal.

Área de cultivos o invernaderos.

A. tiempo estimado en que un operario, se demora en cortar una caja de rosas. [15]

[15] La caja de rosas contiene 24 unidades.

Tiempo de corte = tiempo (minutos) / Caja de rosas (24 unidades)

Tiempo de corte = X minutos / rosa.

B. El tiempo en que un operario demora en llevar las cajas de rosas desde los invernaderos hasta la sala de poscosecha. Este indicador depende en el numero de cajas que pueda llevar el transporte en un viaje, es decir el tipo de transporte que se tiene en cada empresa ya que los modelos de los carros de carga son diferentes en todas las empresas.

Tiempo de recorrido = Tiempo (minutos) / Número de cajas transportadas.[16]

Tiempo de recorrido = X minutos / cajas.

Este indicador, puede variar ya que en una empresa floricultura se pueden tener más de un invernadero, por lo anterior hay que tomar este indicador desde cada invernadero que se tenga en la empresa hasta la sala de poscosecha.

Área de poscosecha.

A. Calcular el tiempo en que se demora un operario en bonchear[17] una docena de rosas.
B. Tomar el tiempo en que se demora un operario en realizar un arreglo de rosas.[18]
C. Tiempo en que se demora un operario en trasladar una caja de rosas arregladas hasta el cuarto frió.
D. Calcular el tiempo en que se demora el operario en arreglar y limpiar su sitio de trabajo.
E. El tiempo estimado en que se demora un operario en llevar todos los residuos hasta el área donde se depositan las basuras o donde se pone el material vegetal para ser convertido en materia orgánica.

Se recomienda elaborar los indicadores de gestión para todas las labores directas en la producción de rosas, para determinar cuantas personas se necesitan dentro del el área de cultivo y poscosecha.

[16] El número de cajas transportadas depende de la capacidad que tenga el medio de transporte que se utilice en cada empresa para esta labor.
[17] La poncheada esta determinada por un arreglo de una docena de rosas.
[18] El arreglo de rosas esta compuesto por 24 unidades.

4.3 INVENTARIOS Y MATERIA PRIMA

Al haber establecido el modelo de proyecciones a la empresa, se puede obtener una mejor rotación en los inventarios como también en los pedidos de materia prima.

El objetivo que se traza en este punto es observar las producciones mensuales de rosas y así poder calcular y tener mejor y más organizado los inventarios en bodegas, ya que al poder observar estas producciones se pueden realizar los pedidos de materia prima sin tener que llenar a tope estas bodegas o no que se puedan quedar sin materia prima para sacar las producciones de rosas.

5. CONCLUSIONES

- El modelo de proyecciones es una herramienta de ayuda para las empresas floricultoras del país ya que esta da a conocerlos volúmenes de producción que puedan obtener estas empresas, conocer este volumen de producción es importante ya que las empresas pueden anticipar todas sus tareas y labores que se presentan dentro de estos cultivos.
- Las variedades estudiadas obtuvieron un porcentaje de acierto del 97,37% en las producciones totales proyectadas, pero al observar el estudio semanal en las variedades estas mostraron que tiene un margen de mas o menos 17% ya que en algunos casos y para algunas variedades se obtuvieron sobreproducciones.
- El modelo de proyecciones mostro un aumento de mas del 15% en las proyecciones semanales, y mostro un acierto de 99% para la producción total al finalizar las 7 semanas de evaluación.
- Para las variedades JADE, LIGHT ORLANDO y LIPSTICK, el estudio fue excelente ya que el porcentaje de acierto en las proyecciones aumento significativamente, con lo cual se concluye que las curvas de crecimiento para estas variedades se construyeron con poco margen de error.
- En algunas semanas y para algunas variedades los porcentajes de acierto se encontraron por debajo del 90% o se obtuvieron sobreproducciones, lo que indica que se tiene que realizar el estudio de campo cuidadosamente para que al desarrollar las curvas de crecimiento se obtenga un resultado con un menor margen de error y no se obtengan sobreproducciones.
- Dar a conocer a las empresas floriculturas del país una herramienta, que les permita conocer las producciones de manera anticipada para que puedan desarrollar sus estrategias administrativas y gerenciales de una forma adecuada.
- Esta herramienta facilita las labores administrativa, ya que se puede tener un mejor flujo en las producciones de rosas y también en obtener un excelente flujo en el manejo de inventarios.
- Al realizar una evaluación de acuerdo con las proyecciones de las 7 semanas evaluadas contra la producción real obtenida en campo durante estas 7 semanas se puede definir que este resultado es excelente ya que la producción real total fue de 396.251 rosas y la producción total proyectada por el modelo de proyección se situó en 391.980 rosas. Para una efectividad del 98,92%.

6. BIBLIOGRAFIA

FULLER, Harry. CAROTHERS, Zane. PAYNE, Willard. BALBACH, Margaret. Botánica: Historia, 2° ed. (1999). 10p.

PIZANO, Marta. Cultivos de rosas bajo invernadero: Características botánicas, 1° ed. Bogota: Hortitecnica Ltda. 2001. 9 p.

ACUÑA, Fabio. VALERA, Diego. AVENDAÑO, Juan Carlos. Invernaderos la experiencia iberoamericana: Estructuras de invernaderos: la experiencia en Colombia, 1° ed. Almería España (2004). 83-102 p.

MONROY, Néstor. PEREZ, Ignacio. CURE, José Ricardo. Estudio de la variabilidad en el clima y la producción de rosas en la sabana de bogota. Nº 14 Noviembre de 2001; p. 38-43

MONROY, Néstor. PEREZ, Ignacio. CURE, José Ricardo. Modelo de predicción y manejo del cultivo de rosas. Nº 15 Mayo de 2002; p.18-22

RODRIGUEZ, Wbeimar. FLOREZ, Víctor. Comportamiento fonológico de tres variedades de rosas rojas en función de la acumulación de temperatura. En: Agronomía Colombiana. V. 24, Nº 2; p.

Republica de Colombia. Ministerio de Agricultura y Desarrollo Rural. Estadísticas 2001.

Asocolflores (Asociación Colombiana de Cultivadores de Flores). (Consulta 10 Ene. 2008). Disponible en: www.colombianflowers.com/info/info_plana.php

INFOAGRO. (Base de datos en línea). (Consulta 31 Ene. 2008). Disponible en <http://www.infoagro.com/flores/flores/rosas2.htm

ROSALES. (Base de datos en línea). (Consulta 31 Ene. 2008). Disponible en <http://articulos.infojardin.com/rosales/historia-rosa-cultivo-rosa.htm

BOGOTA. (Ubicación). (consulta 3 Mayo. 2008). Disponible en <htpp://www.atarraya.org/agenda/Bogota.html

DIRECCION NACIONAL DE MEEOROLOGIA. (Consulta 20 de Marzo. 2008). Disponible en <http://www.meteorologia.com.uy/glosario.g.htm

Tabla 5. Proyeccion de la variedad tineke en la semana 13

VARIEDAD	Bloque o ivernadero.

Semana	semana 1 (2007)							semana 2 (2007)							semana 3 (2007)						
Día	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Produccion Anterior	0	0	1576	1260	1260	1652	1044	820	1248	1204	1003	876	1008	987	987	876	765	923	1005	989	765

Éxito cosecha	80%

Distribucion Dias	71	72	73	74	75	76	77	78	79	80	81	82	83
Porcentaje	3%	3%	5%	5%	8%	13%	15%	15%	13%	8%	5%	5%	5%

Proy / Días	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	18-Mar	19-Mar	20-Mar	21-Mar	22-Mar	23-Mar	24-Mar	25-Mar	26-Mar	27-Mar	28-Mar	29-Mar	30-Mar
31 DE DICIEMBRE 0	0	0	0	0	0	0	0	0	0	0	0	0	0						
1 DE ENERO 07		0	0	0	0	0	0	0	0	0	0	0	0	0					
2 DE ENERO 07			38	38	63	63	101	164	189	189	164	101	63	63	63				
3 DE ENERO 07				30	30	50	50	81	131	151	151	131	81	50	50	50			
4 DE ENERO 07					30	30	50	50	81	131	151	151	131	81	50	50	50		
5 DE ENERO 07						40	40	66	66	106	172	198	198	172	106	66	66	66	
6 DE ENERO 07							25	25	42	42	67	109	125	125	109	67	42	42	42
TOTAL			38	68	124	183	266	386	509	619	705	690	598	491	378	234	158	108	42

Proyeccion / Días	04-Ene	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	18-Mar	19-Mar	20-Mar	21-Mar	22-Mar	23-Mar	24-Mar	25-Mar	26-Mar	27-Mar	28-Mar	29-Mar	30-Mar	31-Mar	01-Abr	02-Abr	03-Abr	04-Abr	05-Abr	06-Abr
7 DE ENERO 07								20	20	33	33	52	85	98	98	85	52	33	33	33						
8 DE ENERO 07									30	30	50	50	80	130	150	150	130	80	50	50	50					
9 DE ENERO 07										29	29	48	48	77	125	144	144	125	77	48	48	48				
10 DE ENERO 07											24	24	40	40	64	104	120	120	104	64	40	40	40			
11 DE ENERO 07												21	21	35	35	56	91	105	105	91	56	35	35	35		
12 DE ENERO 07													24	24	40	40	65	105	121	121	105	65	40	40	40	
13 DE ENERO 07														24	24	39	39	63	103	118	118	103	63	39	39	39
TOTAL								20	50	92	136	196	299	428	537	620	642	631	593	526	418	290	179	115	80	39

Proyeccion / Días	12-Mar	13-Mar	14-Mar	15-Mar	16-Mar	17-Mar	18-Mar	19-Mar	20-Mar	21-Mar	22-Mar	23-Mar	24-Mar	25-Mar	26-Mar	27-Mar	28-Mar	29-Mar	30-Mar	31-Mar	01-Abr	02-Abr	03-Abr	04-Abr	05-Abr	06-Abr	07-Abr	08-Abr	09-Abr	10-Abr	11-Abr	12-Abr	13-Abr
14 DE ENERO 07															24	24	39	39	63	103	118	118	103	63	39	39	39						
15 DE ENERO 07																21	21	35	35	56	91	105	105	91	56	35	35	35					
16 DE ENERO 07																	18	18	31	31	49	80	92	92	80	49	31	31	31				
17 DE ENERO 07																		22	22	37	37	59	96	111	111	96	59	37	37	37			
18 DE ENERO 07																			24	24	40	40	64	105	121	121	105	64	40	40	40		
19 DE ENERO 07																				24	24	40	40	63	103	119	119	103	63	40	40	40	
20 DE ENERO 07																					18	18	31	31	49	80	92	92	80	49	31	31	31
TOTAL															24	45	79	115	175	274	378	460	530	555	558	538	479	362	251	166	110	70	31

SEMANA 13, DEL 25 DE MARZO AL 31 DE MARZO	
TOTAL SEMANA 13	6099

919	938	898	879	854	810	800

ANEXO A

Formato: Seguimineto del crecimiento de las plantas.

Variedad:_____

ESTADO / TALLO	DÍA PINCH	DÍA ACTIVACION	BOTON ARROZ	BOTON GARBAZO	BOTON COLOR	DÍA COSECHA
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						
26						
27						
28						
29						
30						
31						
32						
33						
34						
35						
36						
37						
38						
39						
40						
41						
42						
43						
44						
45						
46						
47						
48						
49						
50						

Fuente: Elaborado por el autor.

ANEXO B

FORMATO DE RECOLECCION DATOS

NA= No Activado
C= Ciego
P= Perdido
= Cantidad de C;NA;P
= Éxito de la cosecha
= Perdida de la cosecha

VARIEDAD / N° TALLOS	TINEKE		BLOQUE 3	N° TALLOS/ DIA	% DIA
	DIA PINCH	DIA CORTE	DIAS TRANS		
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28					
29					
30					
31					
32					
33					
34					
35					
36					
37					
38					
39					
40					
41					
42					
43					
44					
45					
46					
47					
48					
49					
50					

Tallos no cultivados	Porcentaje de éxito	Porcentaje de perdida
0	0,00	0,00

Fuente: Elaborado por el autor.

DISTRIBUCION

ANEXO C
PRODUCCION REAL MES ENERO

Bloque	Variedad	Fecha 02/01/2007	03/01/2007	04/01/2007	05/01/2007	06/01/2007	07/01/2007	08/01/2007	09/01/2007	10/01/2007	11/01/2007	12/01/2007	13/01/2007	14/01/2007	15/01/2007	16/01/2007
01																
02	ANNA	1540	812	1148	1204	1323	1678	1200	1176	1498	1596	1372	1456	1567	1345	1234
	JADE	520	480	580	520	860	300	580	640	620	500	560	740	440	476	644
	Total	2060	1292	1728	1724	2183	1978	1780	1816	2118	2096	1932	2196	2007	1821	1878
03	ANNA	1428	728	700	809	1023	984	678	701	912	807	840	763	624	453	342
	TINEKE	1576	1260	1260	1652	1044	820	1248	1204	1003	876	1008	987	987	876	765
	Total	3004	1988	1960	2461	2067	1804	1926	1905	1915	1683	1848	1750	1611	1329	1107
04	LIPSTICK	2934	2087	2758	2138	1567	896	868	1122	1376	1234	952	868	812	1036	1009
	Total	2934	2087	2758	2138	1567	896	868	1122	1376	1234	952	868	812	1036	1009
05																
06	LIGHT ORLANDO	1176	560	756	924	1090	1148	1237	987	765	678	745	896	1092	1129	905
	Total	1176	560	756	924	1090	1148	1237	987	765	678	745	896	1092	1129	905
07	MUSTANG	784	420	420	588	728	328	592	509	417	434	560	644	924	978	1016
	Total	784	420	420	588	728	328	592	509	417	434	560	644	924	978	1.016
Total		9.958	6.347	7.622	7.835	7.635	6.154	6.403	6.339	6.591	6.125	6.037	6.354	6.446	6.293	5.915

Bloque	Variedad	17/01/2007	18/01/2007	19/01/2007	20/01/2007	21/01/2007	22/01/2007	23/01/2007	24/01/2007	25/01/2007	26/01/2007	27/01/2007	28/01/2007	29/01/2007	30/01/2007	31/01/2007	Total
01																	
02	ANNA	1546	1543	1484	567	654	680	1512	1879	2128	2245	2376	1987	1675	1487	1456	20.149
	JADE	476	728	840	784	588	728	728	1036	1260	1008	1428	1120	924	1372	1372	8.460
	Total	2022	2271	2324	1351	1242	1408	2240	2915	3388	3253	3804	3107	2599	2859	2828	28.609
03	ANNA	465	576	765	876	745	644	698	734	897	1009	978	1148	1278	1356	1434	11.792
	TINEKE	923	1005	989	765	1768	1567	1656	1967	1675	1456	1654	2043	2212	2567	2909	16.566
	Total	1388	1581	1754	1641	2513	2211	2354	2701	2572	2465	2632	3191	3490	3923	4343	28.358
04	LIPSTICK	952	1456	1596	1288	1344	1652	2380	1876	1428	1652	1316	1176	1788	2567	2869	21.657
	Total	952	1456	1596	1288	1344	1652	2380	1876	1428	1652	1316	1176	1788	2567	2869	21.657
05																	-
06	LIGHT ORLANDO	1189	1148	1512	896	1176	1148	1148	1876	2072	2772	1820	2492	1932	3612	2996	14.088
	Total	1189	1148	1512	896	1176	1148	1148	1876	2072	2772	1820	2492	1932	3612	2996	14.088
07	MUSTANG	937	789	854	868	997	1272	1788	1432	1680	1102	1428	1904	1978	2380	2072	9.342
	Total	937	789	854	868	997	1.272	1.788	1.432	1.680	1.102	1.428	1.904	1.978	2.380	2.072	9.342
Total		6.488	7.245	8.040	6.044	7.272	7.691	9.910	10.800	11.140	11.244	11.000	11.870	11.787	15.341	15.108	102.054

Fuente: Base de datos Flores el Pino ltda.

ANEXO D

FORMATO DE RECOLECCION DATOS:EJEMPLO

NA= No Activado
C= Ciego
P= Perdido
= Cantidad de C;NA;P
= Éxito de la cosecha
= Perdida de la cosecha

VARIEDAD / N° TALLOS	JADE		BLOQUE 2	N° TALLOS/ DIA	% DIA	DISTRIBUCION
	DIA PINCH	DIA CORTE	DIAS TRANS			
1	08/11/2006	26/01/2007	79	1	2%	69
2	08/11/2006	NA		2	5%	70
3	08/11/2006	22/01/2007	75	3	7%	71
4	08/11/2006	22/01/2007	75	4	9%	72
5	08/11/2006	24/01/2007	77	4	9%	73
6	08/11/2006	25/01/2007	78	5	12%	74
7	08/11/2006	23/01/2007	76	6	14%	75
8	08/11/2006	24/01/2007	77	5	12%	76
9	08/11/2006	17/01/2007	70	4	9%	77
10	08/11/2006	18/01/2007	71	4	9%	78
11	08/11/2006	C		3	7%	79
12	08/11/2006	23/01/2007	76	2	5%	80
13	08/11/2006	17/01/2007	70			
14	08/11/2006	26/01/2007	79			
15	08/11/2006	22/01/2007	75			
16	08/11/2006	25/01/2007	78			
17	08/11/2006	C				
18	08/11/2006	22/01/2007	75			
19	08/11/2006	24/01/2007	77			
20	08/11/2006	21/01/2007	74			
21	08/11/2006	27/01/2007	80			
22	08/11/2006	19/01/2007	72			
23	08/11/2006	20/01/2007	73			
24	08/11/2006	21/01/2007	74			
25	08/11/2006	24/01/2007	77			
26	08/11/2006	20/01/2007	73			
27	08/11/2006	21/01/2007	74			
28	08/11/2006	26/01/2007	79			
29	08/11/2006	19/01/2007	72			
30	08/11/2006	C				
31	08/11/2006	23/01/2007	76			
32	08/11/2006	20/01/2007	73			
33	08/11/2006	19/01/2007	72			
34	08/11/2006	16/01/2007	69			
35	08/11/2006	18/01/2007	71			
36	08/11/2006	19/01/2007	72			
37	08/11/2006	20/01/2007	73			
38	08/11/2006	C				
39	08/11/2006	21/01/2007	74			
40	08/11/2006	25/01/2007	78			
41	08/11/2006	22/01/2007	75			
42	08/11/2006	23/01/2007	76			
43	08/11/2006	23/01/2007	76			
44	08/11/2006	C				
45	08/11/2006	21/01/2007	74			
46	08/11/2006	18/01/2007	71			
47	08/11/2006	25/01/2007	78			
48	08/11/2006	NA				
49	08/11/2006	22/01/2007	75			
50	08/11/2006	27/01/2007	80			

Tallos no cultivados	Porcentaje de éxito	Porcentaje de perdida
7	86,00	14,00

Fuente: Elaborado por el autor

ANEXO E

PRODUCCIONES REALES DE LAS SEMANA 11 A LA 21.

	variedad																
Semana	Aalsmmer gold	Annas	Avant Garde	Jade	Kiko	Light orlando	Lipistck	Maaike	Movie star	Mustang	Papaya	Ravel	Tineke	Tressor 2000	Vendela	Verdi	TOTAL
11	2856	8428	1020	2660	5516	4424	5376	4956	3220	2884	4116	5740	6552	9212	4732	3500	75192
12	3808	14056	1640	2716	8008	6552	6188	6412	4088	4508	5684	8736	9072	11508	5628	4625	103229
13	5488	12908	1440	3696	8764	5544	8204	5152	4900	3864	6916	11172	6440	14476	7112	3625	109701
14	7560	14964	1760	4984	13188	5628	7336	8008	6132	5060	7560	14954	6776	13636	6188	5950	129684
15	9240	17360	3240	5852	14000	8792	5740	25732	7588	9324	9744	17780	9268	15456	4648	6325	170089
16	13384	22596	4944	8008	18760	14252	9016	32004	10164	12376	16110	25368	10920	21280	7672	12073	238927
17	18256	22792	5156	13496	17136	18060	11424	45444	8456	14644	20104	21868	13636	37324	12516	15500	295812
18	17248	13580	2572	4632	8036	7852	6048	25200	5740	5217	11228	13356	9584	24416	10752	11550	177011
19	7672	12572	1856	3332	4760	5124	4592	10528	4368	6636	6188	9128	7658	19516	7224	7200	118354
20	5936	9843	2448	2908	5096	6378	4328	8848	5208	6243	6300	9576	11872	20020	7308	6025	118337
21	3276	7692	1800	3087	4284	4967	4122	4648	3696	6082	4396	6916	9548	11816	5768	5650	87748
TOTAL	94724	156791	27876	55371	107548	87573	72374	176932	63560	76838	98346	144594	101326	198660	79548	82023	1624084

Vuelo

Vuelo	Origen	Destino	Salida	Regreso	Clase	
EQ133	Quito	Manta	vie, 23 mar, 21:00	vie, 23 mar, 21:50	Económica	USD 203.79
EQ132	Manta	Quito	dom, 25 mar, 18:00	dom, 25 mar, 18:50	Económica	

Total remote payment: USD 203.79

Referencias de reserva

Referencia web:	26900073633 Revisar y administrar
Referencia del proveedor para SITA Res:	MYQ7N
Detalles del billete:	Boleto electrónico
Pasajero adulto:	Mr Luis Diaz Mrs Maria Díaz Mrs Katia Puga

Detalles del billete

Boleto electrónico:	2692134471116
Pasajero adulto	Mr Luis Diaz
Vuelo:	EQ133 Económica vie, 23 mar 21:00 Quito Manta
Vuelo:	EQ132 Económica dom, 25 mar 18:00 Manta Quito
Boleto electrónico:	2692134471117
Pasajero adulto	Mrs Maria Díaz
Vuelo:	EQ133 Económica vie, 23 mar 21:00 Quito Manta
Vuelo:	EQ132 Económica dom, 25 mar 18:00 Manta Quito
Boleto electrónico:	2692134471118
Pasajero adulto	Mrs Katia Puga
Vuelo:	EQ133 Económica vie, 23 mar 21:00 Quito Manta
Vuelo:	EQ132 Económica dom, 25 mar 18:00 Manta Quito

TENDENCIAS DE
BI Business Intelligence

LIC. LUIS ALFONSO CUTRO

TEMARIO

PRÓLOGO

Vivimos en la sociedad de la información. Gracias a Internet y al desarrollo de los sistemas de información en las empresas, sus directivos pueden acceder a mucha más información, de más calidad y con mayor rapidez. El potencial que ello ofrece para mejorar la toma de decisiones y para guiar a las empresas hacia la consecución de sus objetivos es enorme. Sin embargo, muchos directivos se enfrentan a la paradoja de que "cada vez tienen más información y menos tiempo para analizarla".

Para enfrentar estos problemas, en los últimos años han surgido una serie de técnicas que facilitan el procesamiento avanzado de los datos. La idea clave es que los datos contienen más información oculta de la que se ve a simple vista. El verdadero valor de la información se revela cuando a partir de ella somos capaces de descubrir conocimiento. Este es el verdadero objetivo de la Business Intelligence.

INTRODUCCIÓN

En la actualidad, en cualquier organización se hace necesario la toma decisiones.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y en otras fuentes de almacenamiento.

Surge aquí la necesidad de conjugar los distintos ficheros y bases de datos de manera que se pueda utilizarlos para extraer conclusiones.

La estructuración de los datos no es sencilla y esto se agrava cuando los diferentes ficheros o bases de datos se encuentran en sistemas informáticos y soportes diferentes. Lo razonable sería recoger los datos (información histórica) en un sistema separado y específico. Nace el Data Warehouse (DWH), Almacén o Bodega de Datos, con la necesidad de unificar los distintos ficheros y bases de datos para poder comprenderlos. Por ello, se necesita de tecnologías que sirvan de guía para comprender el contenido de las Bases de Datos.

QUÉ ES BUSINESS INTELLIGENCE

Business Intelligence es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios.

También se define a Business Intelligence como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales (internos y externos a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLTP / OLAP, alertas...) o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

En definitiva, una solución BI completa permite:

- Observar: ¿qué está ocurriendo?.
- Comprender: ¿por qué ocurre?.
- Predecir: ¿qué ocurriría?.
- Colaborar: ¿qué debería hacer el equipo?.
- Decidir: ¿qué camino se debe seguir?.

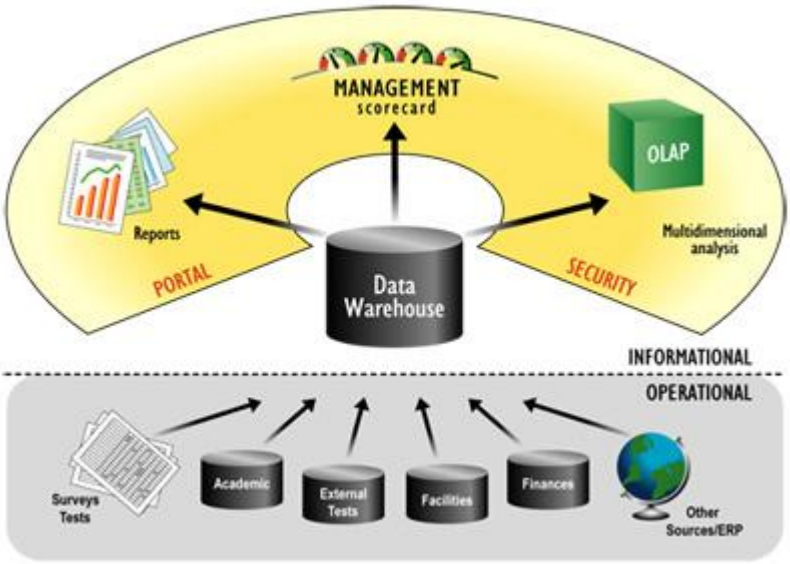
COMPONENTES DE BUSINESS INTELLIGENCE

Los principales productos de Business Intelligence que existen hoy en día son:

- Cuadros de Mando Integrales (CMI).
- Sistemas de Soporte a la Decisión (DSS).
- Sistemas de Información Ejecutiva (EIS).

Por otro lado, los principales componentes de orígenes de datos en el Business Intelligence que existen en la actualidad son:

- Data Mart.
- Data Warehouse.



BENEFICIOS QUE PROPORCIONA A LA ORGANIZACIÓN

Entre los beneficios más importantes que BI proporciona a las organizaciones, vale la pena destacar los siguientes:

- Reduce el tiempo mínimo que se requiere para recoger toda la información relevante del negocio, ya que la misma se encontrará integrada en una fuente única de fácil acceso.
- Automatiza la asimilación de la información, debido a que la extracción y carga de los datos necesarios se realizará a través de procesos predefinidos.
- Proporciona herramientas de análisis para establecer comparaciones y tomar decisiones.
- Cierra el círculo que hace pasar de la decisión a la acción.
- Permite a los usuarios no depender de reportes o informes programados, porque los mismos serán generados de manera dinámica.
- Posibilita la formulación y respuesta de preguntas que son claves para el desempeño de la empresa.
- Permite acceder y analizar directamente los indicadores de éxito.
- Se pueden identificar cuáles son los factores que inciden en el buen o mal funcionamiento de la empresa.

- Se podrán detectar situaciones fuera de lo normal.
- Se encontrarán y/o descubrirán cuáles son los factores que maximizarán el beneficio.
- Permitirá predecir el comportamiento futuro con un alto porcentaje de certeza, basado en el entendimiento del pasado.
- El usuario podrá consultar y analizar los datos de manera sencilla.

TENDENCIAS DE BUSINESS INTELLIGENCE

Analizando las tendencias tecnológicas actuales y las nuevas tendencias estratégicas, la Inteligencia de Negocios (BI, Business Intelligence) podrá beneficiarse de estas tendencias para ampliar aún más, su impacto en las Organizaciones. Algunas de estas son:

Cómputo Colaborativo. El trabajo colaborativo y relacional propiciado por las redes sociales en las Empresas fomentará que las nuevas funcionalidades de BI, como la de poder hacer anotaciones o comentarios compartidos en los informes, faciliten el mejor entendimiento y socialización de los mismos. Adicionalmente el poder acceder a estos datos no estructurados desde las plataformas de BI para detectar oportunidades y optimizar las decisiones, será uno de los grandes retos para los proveedores de este tipo de soluciones.

Computación móvil. El creciente aprovechamiento de las redes celulares para potenciar la computación en cualquier parte y en cualquier momento, facilitará el uso de los mensajes de datos inteligentes en el momento justo, mejorando así todo el proceso de gestión comercial y gerencial, pues una vez llegue el mensaje de

alerta al móvil, se podrá visualizar en él, en tiempo real el informe completo y actualizado, haciendo más efectiva su labor con el Cliente o donde quiera que se este llevando a cabo la presentación.

Visualización de datos. Las cada vez más sofisticadas herramientas para la visualización de datos (en formatos que van más allá de las simples imágenes estáticas) permitirán presentar información de forma clara y efectiva. Este concepto de visualización de datos ha estado directamente relacionado con las tecnologías de BI desde sus orígenes y obtener mayor interactividad y claridad de mediante la visualización será uno de los objetivos de las soluciones analíticas de los próximos años.

Análisis predictivos. Las soluciones de Inteligencia de Negocios permitirán también aprovechar otras tendencias del sector, como son los análisis predictivos o de performance, que tratan de aplicar disciplinas matemáticas y de minería de datos (estadística, reconocimiento de modelos o inteligencia artificial) para tratar de encontrar patrones y tendencias en los datos que permitan mejorar el desempeño en áreas como proyecciones de consumo, prevención de fraudes, tendencias de compra o ideas para nuevos productos. Gartner predice que para el 2011, las Empresas que implementen aplicaciones de Performance Management (CPM) para soportar una cultura organizacional basada en él, tendrán mejores resultados que sus pares en un 30%.

Análisis Dinámico de datos. Uno de los más importantes retos de la Inteligencia de Negocios es cómo poder realizar análisis sobre grandes volúmenes de datos, con origen diverso y hacerlo muy rápidamente, según vayan llegando. Los próximos avances en esta tendencia permitirán que nuevos análisis sean viables en sectores muy diversos como seguimiento y control al tráfico vehicular, mercados financieros globalizados, seguimiento médico a pacientes geográficamente dispersos, monitoreo de ecosistemas y medio ambiente global, además de agilizar las tradicionales de análisis global y en línea de mercados, ventas, clientes, compras, proveedores, etc.

Aplicaciones compuestas. Las nuevas arquitecturas orientadas a servicios web se irán integrando cada vez más con la tradicional plataforma de BI, habilitando de este modo una proliferación de nuevas aplicaciones que pueden ser creadas o integradas directamente por los usuarios de una manera sencilla, flexible y orientadas a la Web, entre estas nuevas aplicaciones figurarán informes de fuentes de datos RSS, Twitter, Redes Sociales, etc. Esta tendencia aumentará los

próximos años, a medida que más usuarios se sientan cómodos creando aplicaciones perfectamente adaptadas a sus necesidades a partir de componentes.

Cloud computing. El movimiento Web 2.0 también ha traído conceptos como el de cloud computing, un modelo que ofrece servicios accesibles en cualquier momento a través de Internet, de forma que el lugar donde se encuentra un usuario o dispositivo es irrelevante. Las tecnologías de base que soportan el cloud computing (virtualización, automatización, estándares abiertos e informática basada en la Red) permiten a los centros de datos corporativos actuar con la eficiencia de Internet. La relación del cloud computing con la inteligencia de negocio aumenta a medida que los usuarios se hacen cada vez más móviles, disponen de varios dispositivos preparados para Internet y necesitan acceder a su información independientemente del lugar en que se encuentren. Algunos ejemplos de esta tendencia son el Elastic Compute Cloud (EC2) de Amazon, la Platform as a Service (PaaS) de Google, y la Azure platform de Microsoft, para solo citar algunos de ellos.

Interfaces Multitouch. Esta tecnología se popularizó gracias a la película Minority Report y al lanzamiento del iPhone de Apple. Se trata de una interfaz que dispone de una pantalla multitáctil que ofrece nuevas e innovadoras capacidades de interacción (como ampliar o reducir la pantalla con un simple gesto del dedo). Las herramientas de BI también aprovecharán las ventajas de esta tecnología para simplificar la navegación o manipular dispositivos móviles, por ejemplo. Con la llegada del movimiento ecológico y del cloud computing, la creciente proliferación de dispositivos móviles cada vez más inteligentes, la capacidad de combinar aplicaciones para crear una aplicación compuesta o mashup y la constante evolución de las tecnologías de visualización y modelización predictiva, se han abierto nuevos horizontes en la aplicación estratégica del análisis de negocios.

PROVEEDORES DE SISTEMAS BI

En el mercado de sistemas de BI existen varios proveedores, entre los que se destacan los siguientes:

Provee software integrado para la planificación, scorecarding y business intelligence. Información adicional está disponible en: www.cognos.com.

Compañía que ofrece software comprensivo y servicios para atender las necesidades únicas de las empresas. Información adicional está disponible en: www.sap.com.

Provee aplicaciones de software para las empresas que operan en tiempo real. Información adicional está disponible en: www.peoplesoft.com

Corporación que además de bases de datos provee varias aplicaciones de software de BI para las mismas. Información adicional está disponible en: www.oracle.com.

División de base de datos de IBM que provee bases de datos y aplicaciones para producir informes y consultas. Información adicional está disponible en: http://www-306.ibm.com/software/data/db2bi/

INTRODUCCIÓN A IBM COGNOS BI, LA SUITE DE BUSINESS INTELLIGENCE DE IBM

IBM Cognos 8 BI es una de las suites de Business Intelligence más utilizadas. Es un software bastante completo, y a la vez manejable, y uno de los líderes del mercado de BI.

Las aplicaciones principales se utilizan desde un portal web que controla el servidor de Business Intelligence, que es el corazón de la herramienta.

Este portal recibe el nombre de **Cognos Connection** y desde el mismo, siempre por web, se accede a opciones de administración del entorno y de los servicios, a las diferentes aplicaciones que provee Cognos, a la estructura de carpetas en que se organizan los informes, a los cuadros de mando, y a otros complementos que se pueden integrar en el portal.

Cada aplicación está orientada a cubrir un tipo de necesidades que suelen darse en entornos de este tipo. La mayoría se maneja 100% desde el explorador web, tanto para desarrollar o diseñar informes, eventos y métricas como para consultarlos o realizar tareas de análisis.

Estas son las principales herramientas que proporciona la suite:

- IBM Cognos Query Studio.
- IBM Cognos Report Studio.
- IBM Cognos Analysis Studio.
- IBM Cognos Event Studio.
- IBM Cognos Metric Studio.
- IBM Cognos Powerplay Transformer.
- IBM Cognos Framework Manager.
- IBM Cognos Planning.

- IBM Cognos TM1.

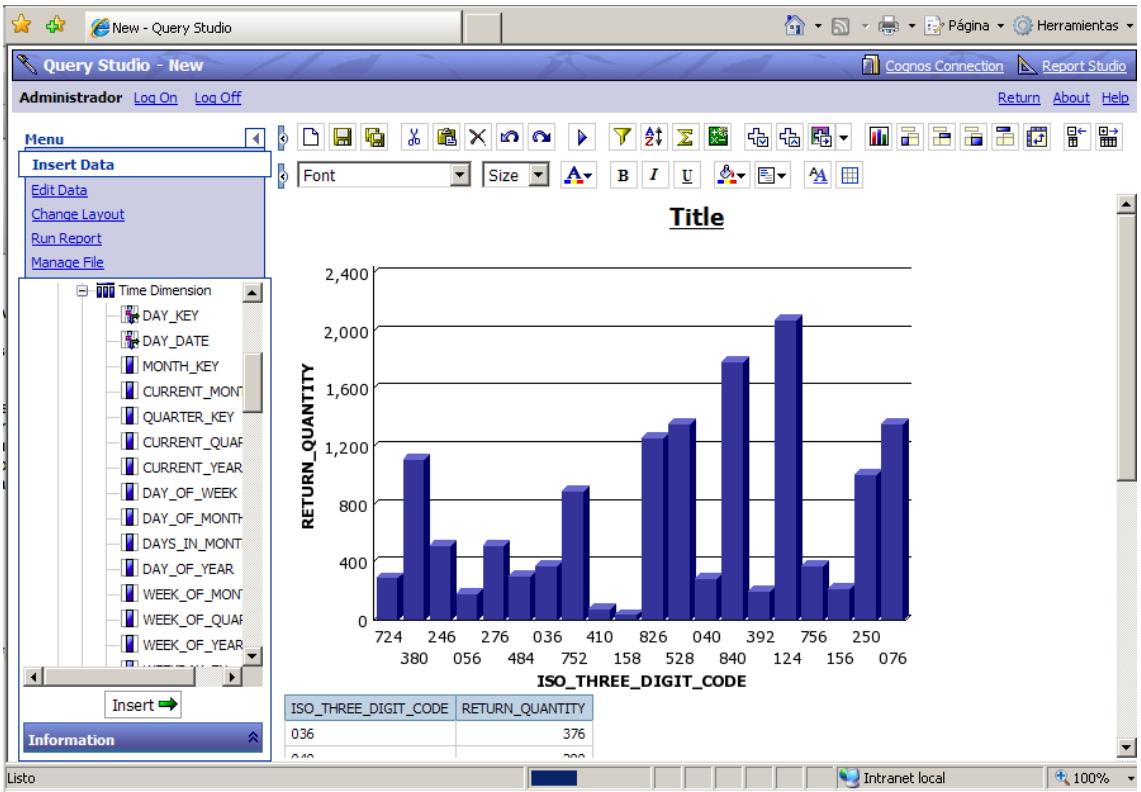
IBM COGNOS QUERY STUDIO

Es la herramienta más simple, y se podría considerar más bien como un complemento. Su objetivo es permitir al usuario realizar consultas sencillas para resolver rápidamente cuestiones puntuales que le puedan surgir.

Permite acceder a la misma estructura de datos que utilizan las otras herramientas, tanto si se trata de un modelo relacional como si la estructura es dimensional.

Con Query Studio se puede crear un informe en segundos arrastrando campos desde el explorador de datos hasta el área de diseño de informes. Permite también aplicar filtros, ordenaciones, operaciones de agrupación de datos e incluso crear gráficas. También tiene opciones de formateo, aunque bastante limitadas.

Donde está más limitado es precisamente en la aplicación de formato al informe, y en la creación de campos calculados complejos, utilización de parámetros y otras muchas opciones más avanzadas para las que se ha de utilizar Report Studio.



IBM COGNOS REPORT STUDIO

Es la aplicación principal para la creación de informes. Se asemeja a Query Studio, pero es mucho más completa.

A la izquierda muestra un explorador de objetos desde el que se puede accede a la estructura de datos, y a otros objetos insertables en los informes. A la derecha se encuentra el área de diseño del informe, donde se pueden arrastrar estos objetos e ir componiendo así la estructura.

Estos objetos pueden ser de diferentes tipos: origen de datos, datos específicos del informe y herramientas de diseño. Cada objeto que se incrusta en el informe tiene sus propiedades configurables, y mediante estas se puede llegar a un nivel muy alto de personalización, tanto en los datos que se muestran como en el diseño del formato.

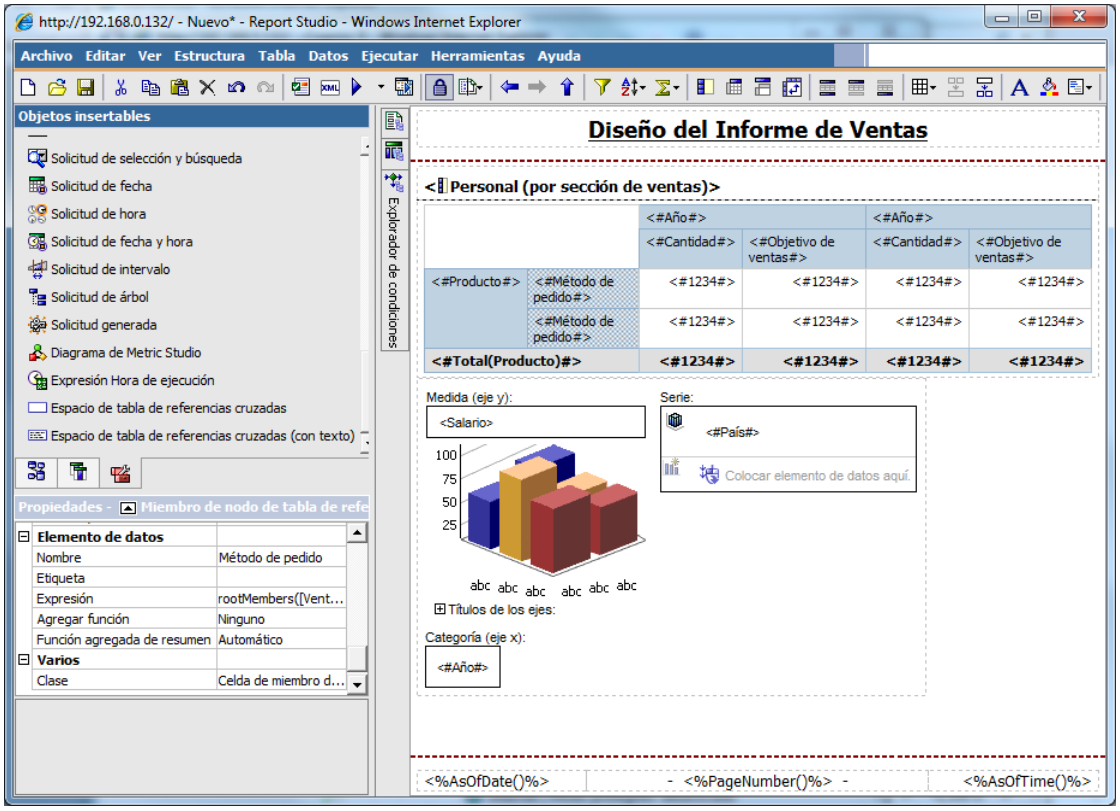
se puede trabajar tanto con estructuras de datos relacionales como con estructuras dimensionales, sólo hay que tener en cuenta que en función del tipo de origen existen diferencias en cuanto a las propiedades aplicables a los datos, e incluso en cuanto al comportamiento en el área de diseño. Aunque no es obligatorio hacerlo así, para mostrar datos de estructura dimensional, lo más apropiado es utilizar informes de tipo crosstab. Se puede elegir entre varios tipos de estructura básica para los informes.

Existen diferentes tipos de gráficas, e incluso mapas que se pueden incluir en los informes, mostrar de manera individualizada o guardar para formar parte de un cuadro de mando que se mostraría en el portal.

Las opciones de utilización de parámetros y prompts son también bastante completas, aunque la manera en que se definen no es muy intuitiva y resulta un tanto engorrosa.

Como en todos estos tipos de herramientas, se pueden definir filtros, ordenar, agrupar y trabajar con agregados, crear subtotales, campos calculados, formateado condicional. También se puede habilitar el drill up, y drill down, y utilizar drill through.

Las consultas a orígenes operacionales las realiza con SQL y para los modelos dimensionales utiliza MDX. Las consultas resultantes pueden visualizarse e incluso editarse y modificarse directamente.



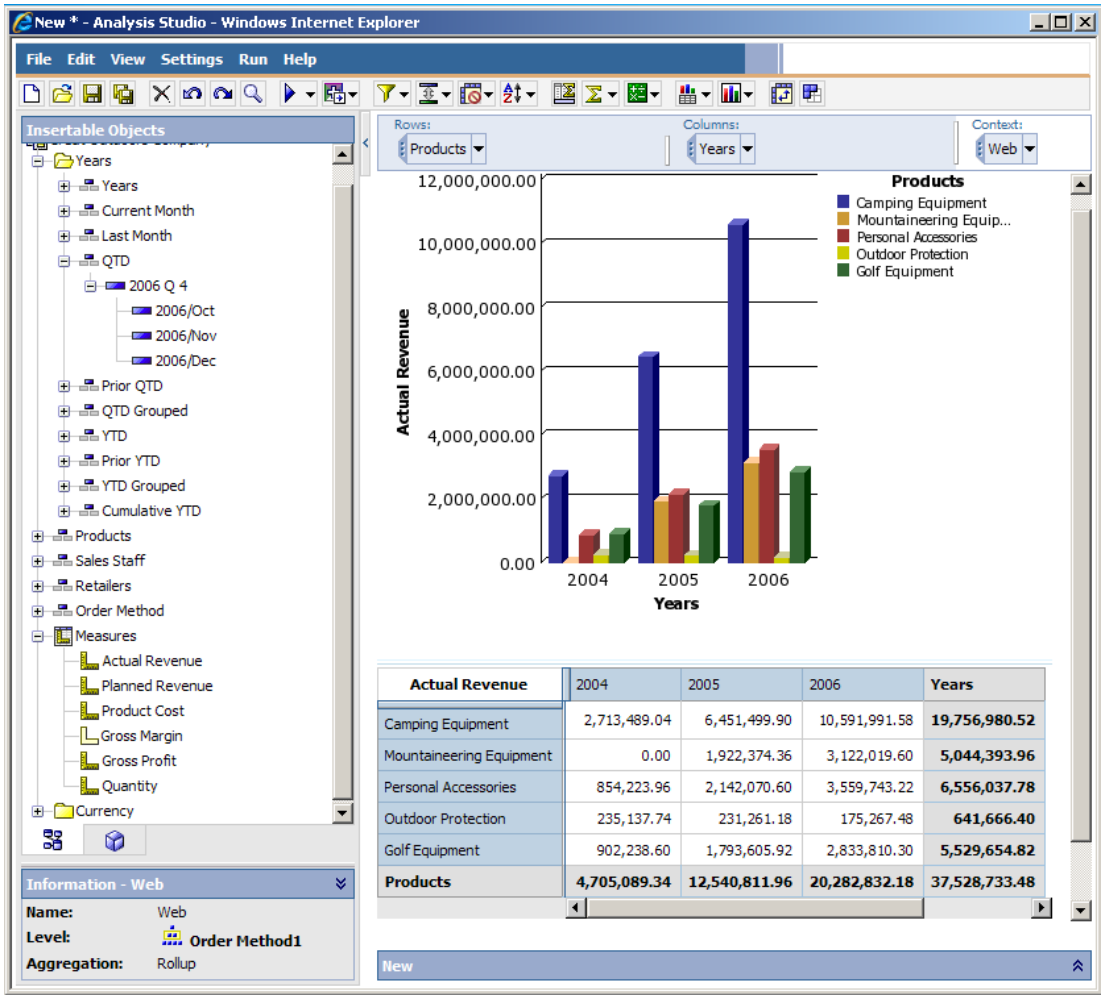
IBM COGNOS ANALYSIS STUDIO

Análisis Studio tiene una función similar a Query Studio, pero para orígenes multidimensionales. Esta herramienta permite la navegación por estructuras multidimensionales como cubos OLAP, que no necesariamente han de ser de Cognos. También puede atacar a orígenes de datos relacionales, siempre que estén modelados dimensionalmente desde Framework Manager.

El objetivo principal de este software es permitir que el analista de negocio pueda 'navegar' por los datos cargados en las estructuras dimensionales sin depender del soporte del área de TI. Utilizando Analysis Studio un usuario de negocios puede realizar análisis complejos y comparativos de datos para descubrir tendencias, riesgos y oportunidades.

El área de trabajo es similar a la de Query Studio y Report Studio y las opciones estándar para la creación de informes son muy similares a las de Query Studio, con aplicación de filtros, ordenaciones, operaciones de agrupación de datos, creación de gráficas, etc.

Además ofrece funcionalidades más orientadas a orígenes dimensionales, como la navegación con drill up / drill down, o la creación de filtros de contexto.



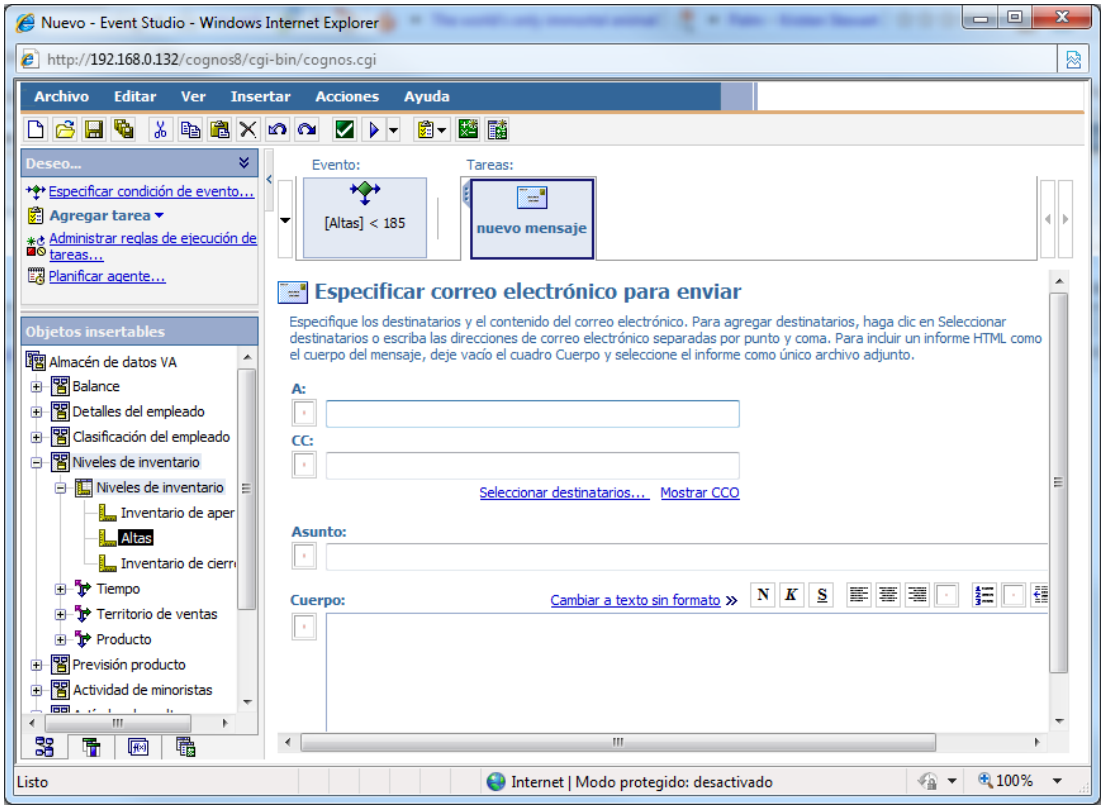
IBM COGNOS EVENT STUDIO

Esta es la herramienta de notificaciones y alertas de la suite.

Con Event Studio se crean agentes que van chequeando los datos o las KPI's definidas, y detectan eventos importantes para el negocio cuando se alcanzan determinados valores o se cumplen ciertas condiciones.

En ese momento la herramienta pasa a ejecutar las acciones o tareas que se hayan asociado a los eventos. Puede ejecutar o distribuir informes, generar emails, comunicarse con otras herramientas de software, ejecutar jobs u otros agentes, llamar a procedimientos almacenados de bases de datos, o incluso llamar a un Web Service.

Una vez construidos los agentes de Event Studio, con el mismo entorno de Scheduler que se utiliza para planificar la ejecución automatizada de informes, se programa y se controla la ejecución periódica de los mismos.



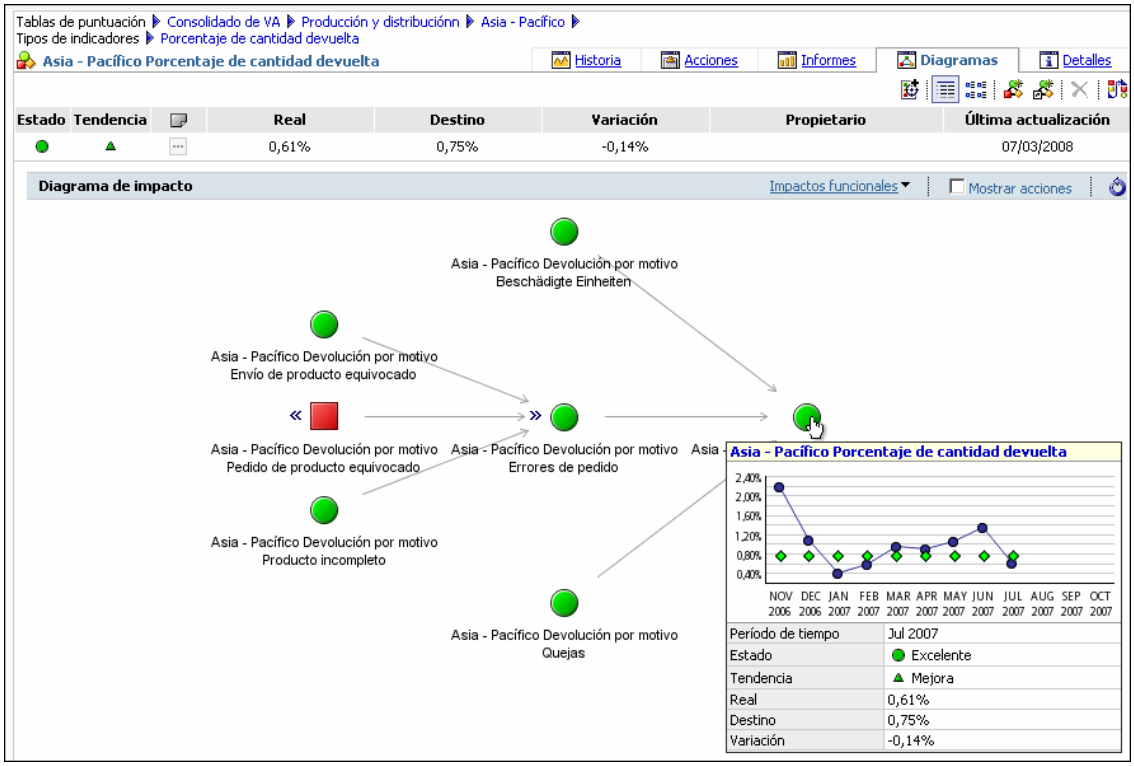
IBM COGNOS METRIC STUDIO

Es la herramienta que se utiliza para la construcción de métricas y cuadros de mando.

Con Metric Studio se definen los KPI o Indicadores Clave del Rendimiento del negocio, se organizan y relacionan entre ellos, se asocian a diferentes perfiles, y se monitorizan, permitiendo así comparar en todo momento objetivos frente a rendimiento, y definir acciones automatizadas, como notificaciones en caso de desviaciones.

Con estas métricas se construyen cuadros de mando que permiten a nivel operativo monitorizar el rendimiento frente a los objetivos, y a nivel estratégico 'mapear' la estrategia corporativa y facilitar su transmisión a todos los niveles de la organización.

Las métricas se pueden construir a partir de diferentes orígenes de datos, tales como cubos OLAP, bases de datos relacionales, hojas de cálculo, ficheros de texto, e incluso valores informados manualmente, y la herramienta dispone de asistentes para facilitar la construcción de las métricas y los cuadros de mando.

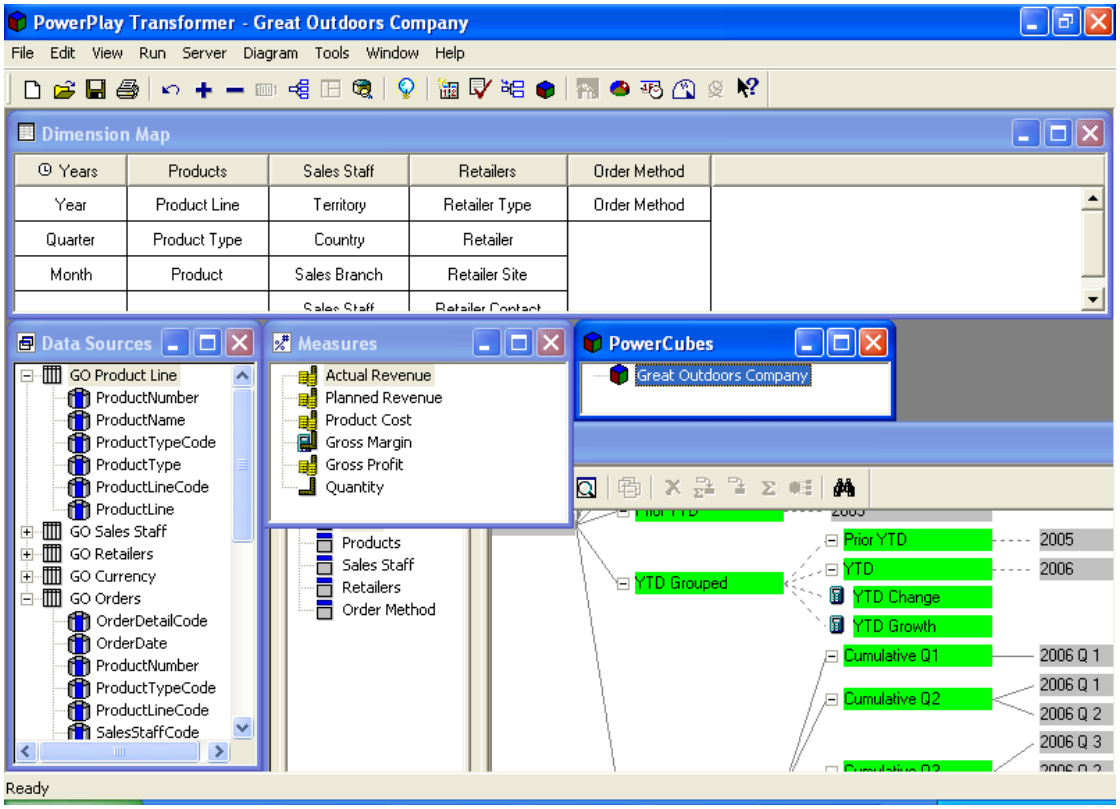


IBM COGNOS POWERPLAY TRANSFORMER

Las herramientas mencionadas anteriormente permiten acceder a cubos OLAP como origen de datos. IBM Cognos PowerplayTransformer es la herramienta que permite construir cubos OLAP, los llamados IBM Cognos PowerCube. Aunque las herramientas de reporting de Cognos pueden trabajar en ROLAP y atacar bases de datos relacionales, para realizar tareas analíticas lo más eficiente suele ser utilizar un cubo OLAP como origen de datos, es decir, trabajar en MOLAP. Con un volumen de datos controlado los tiempos de respuesta en la utilización de los informes pueden ser mucho mejores.

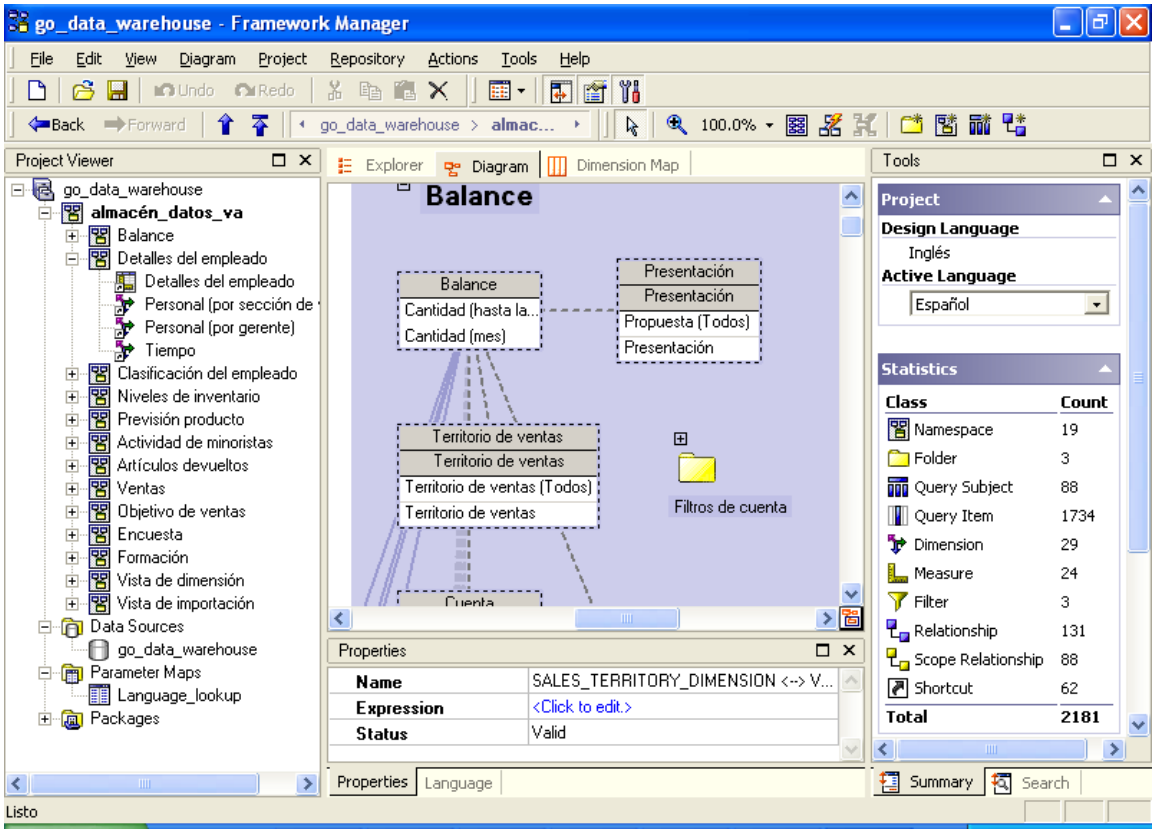
Con Powerplay se definen los orígenes de datos, se modeliza la estructura multidimensional que va a conformar el cubo, se valida, y se procede a la construcción del mismo.

Esta herramienta se instala en modo cliente, no se accede por web. Se utiliza de manera independiente para generar los cubos, que es lo que atacan las otras herramientas web.



IBM COGNOS FRAMEWORK MANAGER

Esta es la herramienta que se utiliza para la construcción de los metadatos necesarios para que todas las demás funcionen. No suele mencionarse como producto porque es la que utiliza el área técnica para crear los paquetes de metadatos que las herramientas de reporting consultan para mostrar al usuario de negocios una estructura inteligible, y permitir crear un árbol de navegación que acaba consultando los datos de los sistemas origen.


go_data_warehouse - Framework Manager
File Edit View Diagram Project Repository Actions Tools Help
Undo Redo
Back Forward
go_data_warehouse > almac...
100.0%
Project Viewer
go_data_warehouse
almacén_datos_va
Balance
Detalles del empleado
Detalles del empleado
Personal (por sección de
Personal (por gerente)
Tiempo
Clasificación del empleado
Niveles de inventario
Previsión producto
Actividad de minoristas
Artículos devueltos
Ventas
Objetivo de ventas
Encuesta
Formación
Vista de dimensión
Vista de importación
Data Sources
go_data_warehouse
Parameter Maps
Language_lookup
Packages
Explorer
Diagram
Dimension Map
Balance
Balance
Cantidad (hasta la...
Cantidad (mes)
Presentación
Presentación
Propuesta (Todos)
Presentación
Territorio de ventas
Territorio de ventas
Territorio de ventas (Todos)
Territorio de ventas
Filtros de cuenta
Properties
Name
SALES_TERRITORY_DIMENSION <--> V...
Expression
<Click to edit.>
Status
Valid
Properties
Language
Tools
Project
Design Language
Inglés
Active Language
Español
Statistics
Class
Count
Namespace 19
Folder 3
Query Subject 88
Query Item 1734
Dimension 29
Measure 24
Filter 3
Relationship 131
Scope Relationship 88
Shortcut 62
Total 2181
Summary
Search
Listo

CONCLUSIONES

El uso de las herramientas BI tiene muchas ventajas, pero a la vez, contiene ciertas limitaciones relacionadas con la cultura de las organizaciones que los encargados de implantarlas deben tomar en consideración al momento de planificar e implantar un sistema de BI en sus empresas. Como toda herramienta, el BI tiene la posibilidad de influenciar o impactar positiva o negativamente ciertas áreas funcionales de la empresa.

La literatura evidencia que un creciente número de empresas está aplicando una variedad de tecnologías de BI para desarrollar ventajas competitivas. Lo cierto es que el futuro del e-business es el mundo del BI estratégico por lo que toda la información que la empresa posee debe ser integrada para asistir eficientemente a la toma de decisiones. Por esto, al igual que todo proceso, su implantación debe ser planificada y controlada, ya sea por un IT o un MIS. De esta forma se garantiza que la empresa está enfocada en los propósitos que persigue con estas herramientas, así como cumplir con los requerimientos tecnológicos a corto y largo plazo.

En cuanto a las tendencias futuras de estas herramientas, es la opinión del que redacta, que en la medida que las empresas descubran otros usos para los datos que se generan en la organización y éstos se integren más eficientemente con otras herramientas que se utilizan en las empresas como ERP, CRM, etc., otros mercados se identificarán y nuevos clientes internos y externos para esos datos se desarrollarán.

Asimismo, nuevos desarrollos en los sistemas de bases de datos como el "grid computing", prometen nuevas alternativas para atender las necesidades de información de las empresas del futuro. En cualquier caso, la Internet será el medio preferido por las empresas para generar, producir y acceder esta información.

Ahora más que nunca, la Inteligencia de Negocios sigue proporcionando la información que necesitan los directivos para que la empresa pueda seguir su rumbo en un entorno tan competitivo como el actual. Aprovechando las nuevas tendencias, el BI puede ayudar a todo tipo de organizaciones, pequeñas, medianas y grandes, a capitalizar estratégicamente las oportunidades de negocios y responder a los retos con mayor rapidez.

BIBLIOGRAFÍA

- IBM Enterprise Analytics for the Intelligent e-Business, IBM Press, USA, 2001.
- Business Intelligence. Técnicas de análisis para la toma de decisiones estratégicas, Elizabeth Vitt, Michael Luckevich, Stacia Mister. MMcGRAW – HILL / INTERAMERICANA DE ESPAÑA, S.A.U.
- Business Intelligence. COMPETIR CON INFORMACIÓN, Josep Lluís Cano.

ARTÍCULOS

- Adelman, S., "Data Strategy Introduction." DMDirect, November 2001.
- Brath & Peters, "Information Visualization for Business".
- Imhoff, C. Ph.D. "Why Open Source BI, Data Integration, and Data Warehousing Solutions are Gaining in Acceptance", Intelligent Solutions, Inc., June 2009.
- Cavazos, E. GARTNER: Business Intelligence es la prioridad en el 2009, Business Intelligence, Tech & Biz, 2009.
- Future. DM Review, January 2005, pp. 40-43.

Business Intelligence: conceptos y actualidad

Autor: Jorge Alfredo Medina Soto
Nueva economía, internet y tecnología

06-2005

Resumen

Desde principios de los 90's, las aplicaciones de BI han evolucionado dramáticamente en muchas direcciones, debido al crecimiento exponencial de la información. El motivo de este documento es dar una panorama general y sobre todo actualizado, de todo aquello que envuelve Business intelligence dentro de las organizaciones y su manera de evolucionar a través del tiempo. Las aplicaciones de Business Intelligence (BI) son herramientas de soporte de decisiones que permiten en tiempo real, acceso interactivo, análisis y manipulación de información crítica para la empresa.

1. Introducción

Históricamente, la tecnología de Business Intelligence ha encontrado lugar en dos niveles primarios: entre los altos ejecutivos quienes necesitan obtener información estratégica y entre los administradores de la línea de negocios que son responsables del análisis táctico. Estas tradicionales actividades de soporte a la decisión son importantes, pero ellos solamente muestran superficialmente el potencial de la inteligencia de negocios dentro de la empresa., involucrando quizá el 5% de los usuarios y el 10% de los datos disponibles (Information Builders, 2005).

Desde principios de los 90's, las aplicaciones de BI han evolucionado dramáticamente en muchas direcciones, debido al crecimiento exponencial de la información. Desde reportes operacionales generados por mainframes, modelación estadística de campañas publicitarias, ambientes OLAP multidimensionales para analistas así como dashboards y scorecards para ejecutivos. Las compañías empiezan a demandar mas formas de analizar y realizar reportes de datos.

Las inversiones en aplicaciones empresariales, tales como planeación de recursos (ERP) y administración de la relación con el cliente (CRM), han resultando en una enorme cantidad de datos dentro de las organizaciones. Estas organizaciones ahora quieren apalancar estas inversiones y usar la información para ayudarles a tomar mejores decisiones, se más ágiles con organización y tener una mayor comprensión de cómo correr sus negocios.

Por ellos mucha pequeña y mediana empresa está adoptando BI para ayudarles a poner en marcha sus negocios.

El corazón de Business Intelligence es la habilidad de una organización para acceder y analizar la información, y entonces explotar su ventaja competitiva. En la era digital, las capacidades que ofrece Business Intelligence será la diferencia entre el éxito y el fracaso.

2. Breve historia del BI

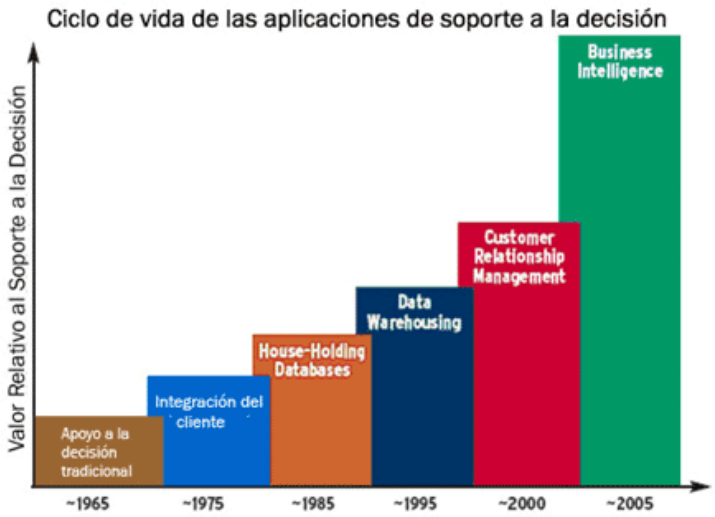
En un tiempo, las organizaciones dependían de sus departamentos de sistemas de información para proporcionarles reportes estándar y personalizados. Esto ocurrió en los días de los mainframes y minicomputadoras, cuando la mayoría de los usuarios no tenía acceso directo a las computadoras. Sin embargo, esto comenzó a cambiar en los años 70's cuando los sistemas basados en servidores se convirtieron en la moda.

Aún así estos sistemas eran usados principalmente para transacciones de negocios y sus capacidades de realizar reportes se limitaba a un número predefinido de ellos. Los sistemas de información se sobrecargaban y los usuarios tenían que esperar por días o semanas para obtener sus reportes en caso que requirieran reportes distintos a los estándares disponibles.

Con el paso del tiempo, fueron desarrollados los sistemas de información ejecutiva (EIS, por sus siglas en inglés), los cuales fueron adaptados para apoyar a las necesidades de ejecutivos y administradores. Con la entrada de la PC, y de computadoras en red, las herramientas de BI proveyeron a los usuarios de la tecnología para crear sus propias rutinas básicas y reportes personalizados.

La figura 1 muestra una breve reseña histórica de cómo fue desarrollándose lo que ahora se conoce como Business Intelligence, también se puede observar la manera en que las aplicaciones relacionadas al soporte de decisiones han ido evolucionando con el paso del tiempo.

Figura 1. Ciclo de vida de las aplicaciones de soporte a la decisión



3. Definición de Business intelligence

Las aplicaciones de Business Intelligence (BI) son herramientas de soporte de decisiones que permiten en tiempo real, acceso interactivo, análisis y manipulación de información crítica para la empresa. Estas aplicaciones proporcionan a los usuarios un mayor entendimiento que les permite identificar las oportunidades y los problemas de los negocios. Los usuarios son capaces de acceder y apalancar una vasta

cantidad de información y analizar sus relaciones y entender las tendencias que últimamente están apoyando las decisiones de los negocios. Estas herramientas previenen una potencial pérdida de conocimiento dentro de la empresa que resulta de una acumulación masiva reinformación que no es fácil de leer o de usar. (CherryTree & Co., 2000)

4. Importancia de BI en las organizaciones

El exceso de información no es poder, pero el conocimiento si lo es. Con demasiada frecuencia, la transformación y el análisis de toda la información y los datos que las propias compañías generan se convierte en un verdadero problema y, por lo tanto, la toma de decisiones se vuelve desesperadamente lenta.

Las tecnologías de BI intentan ayudar a las personas a entender los datos más rápidamente a fin de que puedan tomar mejores y más rápidas decisiones y, finalmente, mejorar sus movimientos hacia la consecución de objetivos de negocios. Los impulsores claves detrás de los objetivos de BI son incrementar la eficiencia organizacional y la efectividad. Algunas de las tecnologías de BI apuntan a crear un flujo de datos dentro de la organización más rápido y accesible. Por otro lado, novedosas tecnologías de BI toman un enfoque más agresivo redefiniendo los procesos existentes con otros nuevos, mucho más estilizados que eliminan gran cantidad de pasos o crean nuevas capacidades.

En una reciente encuesta realizada por Gartner, BI fue catalogado en el número 2 en la lista de prioridades tecnológicas de los CIO para el 2005, después de ubicarse en el lugar número 2 en el año 2004.

Debido a este nuevo énfasis en BI, el mercado de herramientas software de BI alrededor del mundo creció un 7.7 % en 2004, basado en estimaciones preliminares del mercado compuesto.

El crecimiento en 2004 fue conducido por el alto desempeño de vendedores específicos, incluyendo Cognos y Microsoft. El ranking no cambio respecto al año 2003 tal y como se esperaba. Los tres mayores vendedores de herramientas de BI en el mercado global, según datos de Gartner son:

Proveedor	Posición en el mercado compartido
Business Objects	1
SAS Institute	2
Cognos	3

Tabla 1. Mayores proveedores de herramientas de BI
Fuente: Gartner Dataquest (Febrero 2005)

5. Tipos de productos de BI

Las herramientas de software de BI son usadas para acceder a los datos de los negocios y proporcionar reportes, análisis, visualizaciones y alertas a los usuarios. La gran mayoría de las herramientas de BI son usadas por usuarios finales para acceder, analizar y reportar contra los datos que más frecuentemente residen en data warehouse, data marts y almacenes de datos operacionales. Los desarrolladores de

aplicaciones usan plataformas de BI para desarrollar y desplegar aplicaciones (las cuales no son consideradas herramientas de BI). Ejemplos de una aplicación de BI son las aplicaciones de consolidación financiera y presupuestos.

Actualmente el mercado de herramientas de BI se encuentra constituido de dos subsegmentos: suites de BI empresarial (EBIS, por sus siglas en inglés) y plataformas de BI. La mayoría de las herramientas de BI, como las desarrolladas por los vendedores mencionados en la tabla 1, son BI empresarial y plataformas de BI.

Gartner Dataquest (2005) realizó un pronóstico a cinco años, basado en una estimación preliminar de tamaño del mercados y una revisión de los inhibidores e impulsores, llegando a la conclusión de que el total de mercado de herramientas de BI proyecta un crecimiento de $ 2.5 billones en 2004 a $ 2.9 billones en 2009, con una tasa de crecimiento anual de 7.4%.

6. Contrastes: BI empresarial Vs. Plataformas

Tiedrich (2003), menciona que las plataformas de BI son ambientes de desarrollo de aplicaciones, comúnmente ofrecen un lenguaje de codificación como Visual Basic y otros lenguajes para la creación de aplicaciones personalizadas. Además en su:

Ventajas.	Desventajas
Aplicaciones personalizadas.	Complejidad en el desarrollo de aplicaciones
Alta funcionalidad analítica.	

Tabla 2. Ventajas y desventajas de las plataformas de BI
Fuente: Gartner Dataquest (Junio, 2003)

Las plataformas de BI se usan cuando hay una necesidad de analizar aplicaciones complejas con muchos cálculos (por ejemplo, rentabilidad de un producto) o para crear aplicaciones amigables para usuarios ocasionales.

En cambio las herramientas de BI empresarial, contienen una funcionalidad estándar. Una vez que una o más fuente de datos es mapeado por las herramientas de suites de BI empresarial (EBIS, por sus siglas en inglés), la funcionalidad toma vida. A pesar de que algunas herramientas contienen algunas facilidades de codificación, crear aplicaciones a la medida es un desafío.

Según lo dicho por Tiedrich (2003), consultor de Gartner, las EBIS contiene las siguientes ventajas y desventajas.

Ventajas.	Desventajas
Implementación más sencilla.	Funcionalidad menos analítica
Funcionalidad estándar.	Poca facilidad de personalización

Tabla 3. Ventajas y desventajas de *Business Intelligence* Empresarial
Fuente: Gartner Dataquest (Junio, 2003)

Los EBIS son usualmente utilizados cuando hay muchos usuarios de diversos niveles de habilidad técnica, cada uno con requerimientos de reportes y vistas que son menos analíticos (por ejemplo, reportes administrativos o análisis de variantes simples).

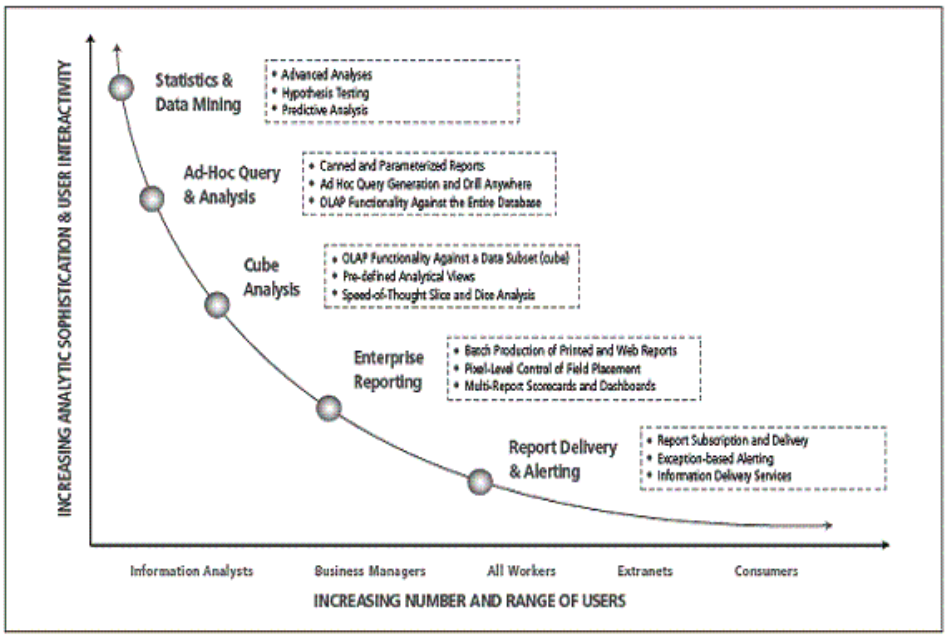
7. Tecnologías de BI

Durante el periodo formativo, las compañías han descubierto activamente nuevas maneras de usar sus datos para apoyar la toma de decisiones, realizar una optimización de procesos y realizar reportes operacionales. Y durante esta era de invenciones, los vendedores de tecnología de BI han construidos nichos de software para implementar cada nuevo patrón de aplicaciones que las compañías inventan. Estos patrones de aplicación resultan en productos de software centrados exclusivamente en cinco estilos de BI (Microstrategy, 2002), tales como:

a) Reporte empresarial. Los reportes escritos son usados para generar reportes estáticos altamente formateados destinados para ampliar su distribución con mucha gente.
b) Cubos de análisis. Los cubos basados en herramientas de BI son usados para proveer capacidades analíticas a los administradores de negocios.
c) Vistas Ad Hoc Query y análisis. Herramientas OLAP relacionales son usadas para permitir a los expertos visualizar la base de datos y ver cualquier respuesta y convertirla en información transaccional de bajo nivel.
d) Data mining y análisis estadísticos. Son herramientas usadas para desempeñar modelado predictivo o para descubrir la relación causa efecto entre dos métricas.
e) Entrega de reportes y alertas. Los motores de distribución de reportes son usados para enviar reportes completos o avisos a un gran número de usuarios, dichos reportes se basan en suscripciones, calendarios, etc

Hasta este punto, las grandes empresas han tenido que comprar diferentes conjuntos de herramientas de BI a distintos vendedores, con cada herramienta dirigida a una nueva aplicación de BI y cada una de ellas dando al usuario funcionalidad en solo uno de los estilos de BI.

Una manera de ver estos estilos de BI es dar lugar a un espacio de dos dimensiones (figura 2) donde el eje vertical representa la sofisticación e interactividad del proceso analítico y el eje horizontal representa la escala, o el tamaño de la población de usuarios. Es entonces cuando se pueden localizar los 5 estilos de BI dentro del cuadrante.



La siguiente tabla muestra las tecnologías que son usadas para Business Intelligence y las cuales entran dentro de los cinco estilos mencionados anteriormente.

Tecnologías de BI
Servidores de base de datos relacional.
Servidores de base de datos OLAP
Data Warehouses
Data Marts
Transformación de datos y herramientas de limpieza
Herramientas de reportes y vistas
Herramientas de análisis y exploración
Herramientas de visualización de datos
Herramientas de Data Mining
Scorecards, portales, y *dashboards*
Hojas de calculo
Herramientas de predicción y modelación
Sistemas de alertas y notificaciones
Aplicaciones analíticas

Tabla 4. Tecnologías usadas en *Business Intelligence*
Fuente: Lokken (2001)

8. BI Operacional

Para mantener el ritmo de competencia, las empresas cada vez demandan Business Intelligence a nivel operacional, análisis incrustados dentro de los procesos para manejar excepciones y tomar decisiones en tiempo real.

Algunos usuarios corporativos que están implementando técnicas como herramientas provenientes de vendedores como SAS Institute Inc., Information Builders Inc. y Cognos Inc.

SAS, Information Builders y Cognos son un grupo del número creciente de vendedores que están creando Business intelligence, según Keith Gile, an analyst at Forrester Research Inc.

"Los negocios quieren dar mayor valor agregado a los datos, no solo al datawarehouse. Muchas de las decisiones en tiempo real que necesitan ser tomadas deben de ser hechas mientras los procesos ocurren, por ejemplo, mientras el consumidor esta en el teléfono o cuando un paciente está siendo tratado" dijo Gile.

La siguiente figura muestra claramente el cuadro completo de BI empresarial y en que parte se encuentra situado el BI operacional.

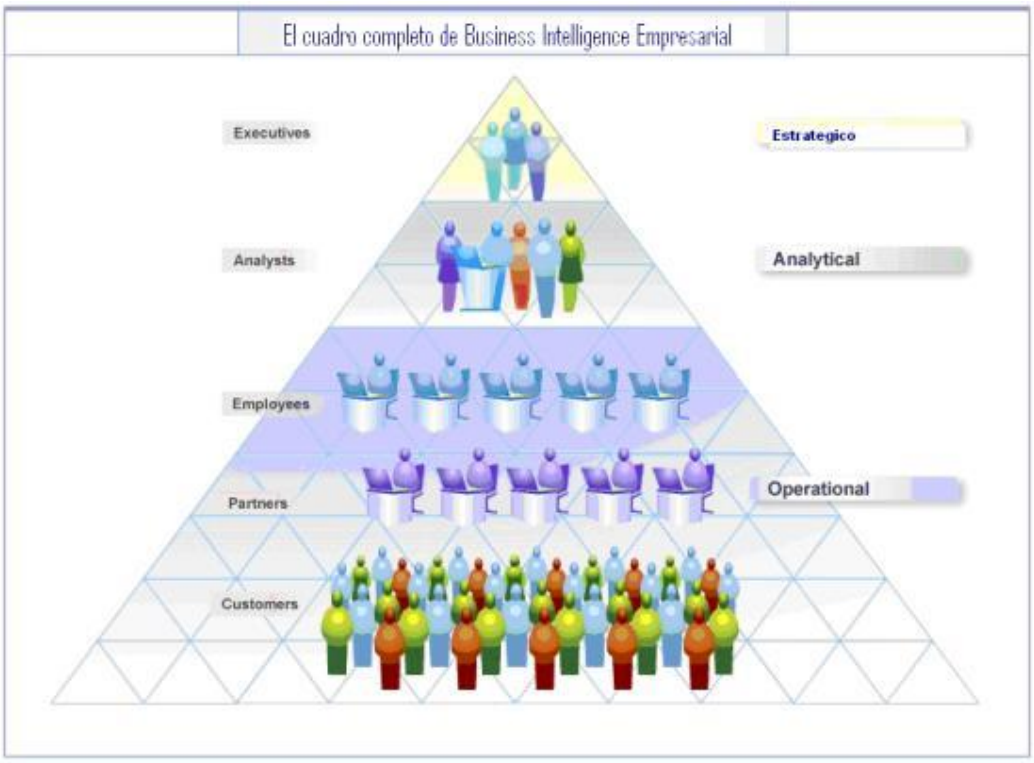


Figura 3. Business Intelligence Operacional
Fuente: Information Builders (Consultado en junio 2005)

9. Factores críticos de éxito.

Lokken (2001) menciona que todos los sistemas de BI tienen un número crítico de factores de éxito en común, ya que ellos:

a) Proveen acceso a datos adecuados. Sin organizar los datos, es difícil lograrlo.

b) Incrementan la habilidad de los usuarios para entender los resultados. Saturar a las personas de números en estos días crea más problemas que los que resuelven. Diez años atrás el problema era obtener los datos; pero hoy en día tiene que ver más con el manejo de ellos.

c) Incrementan el entendimiento de los negocios por parte de los usuarios. Conocer que es lo que los datos dicen es algo bueno, pero en la actualidad es necesario saber que hacer con ellos. Este conocimiento es difícil de construir dentro de una pieza de software.

d) Ayudan a comunicar los hallazgos y tomar acciones. Es raro que un individuo pueda ejecutar cualquier cosa significativa dentro de una organización sin involucrar a otros.

Los cinco factores críticos de éxito de negocios que se deben de considerar al elegir un EBIS son:

a) Minimizar los costos totales de propiedad.

b) Apuntar hacia oportunidades de ROI altos.

c) Apalancar la arquitectura de datos existente.

d) Conocer los requerimientos de los usuarios finales.

e) Asegurar al máximo la escalabilidad y capacidad de realización.

En la actualidad BI debe estar dirigido a estos cinco aspectos y ayudar a simplificar todo el mar de datos para los usuarios. Por ello, el éxito de BI nunca es un accidente; cuando las compañías lo alcanzan logran los siguientes beneficios:

a) Toman mejores decisiones con una asombrosa velocidad y confianza.

b) Dinamizan sus operaciones

c) Reducen los ciclos de vida de sus productos.

d) Maximizan el valor de las líneas de producto y anticipan nuevas oportunidades.

e) Hacer un mejor y más enfocado marketing mejorando las relaciones con los clientes y proveedores por iguales.

Sin embargo las organizaciones deben de entender y dirigir correctamente 10 desafíos críticos para el éxito de BI (Atre, 2003). Los proyectos de BI fallan debido que:

1. Las empresas fallan en reconocer que los proyectos de BI son iniciativas de negocios interorganizacionales, y en entender dichas iniciativas difieren de las típicas soluciones independientes.
2. Existe la falta de compromisos por parte de los sponsors (los cuales tienen autoridad en la empresa.
3. Se tiene poca disponibilidad de los representantes de negocios.
4. Hay ausencia de un personal disponible y habilidoso.
5. Existe un mal concepto del software de BI.
6. No trabajan bajo una estructura detallada.

7. No existe un análisis del negocio o estandarización
8. No existe una apreciación del impacto que causan los datos de mala calidad en la rentabilidad del negocio.
9. No se entiende la necesidad del uso de un meta datos.
10. Demasiada confianza métodos y herramientas no alineadas.

10. Riesgos de BI

Basta con decir que el uso apropiado de las herramientas de BI puede marcar la diferencia entre la vida y la muerte de muchas empresas, entre el estancamiento y el crecimiento, entre los resultados opacos y el desempeño financiero sobresaliente, entre el servicio impersonal y de mala calidad y el excelente servicio al cliente personalizado, y entre la relación óptima con los proveedores y la pérdida de los beneficios de trabajar con ellos y con otros socios de negocios. Por todo ello BI es importante. (Tiedrich, 2003)

Como riesgo, el riesgo que se corre no es demasiado hablando propiamente de evaluar las necesidades reales de BI en la empresa y entonces seleccionar el proveedor más apropiados y sus productos, así como su implementación.

El mayor riesgo tecnológico es que la tecnología está cambiando rápidamente. Naturalmente, las nuevas tecnologías tienen algo de riesgo hasta que son probadas completamente. Por ejemplo, el uso de la tecnología móvil para BI ha sido adoptado muy lentamente.

Dos de los más importantes riesgos son la habilidad de los vendedores para cumplir y últimamente, su viabilidad, lo cual es algo que hay que considerar.

Algunos de los grandes riesgos relacionados con el uso de las herramientas de BI están basados en los datos. Los datos que son usados no son transformados apropiadamente. Debido a que en el ámbito de los negocios las empresas muy frecuentemente escogen sus propias herramientas de BI, una empresa puede terminar con múltiples herramientas, así como múltiples data marts con datos que no están claramente definidos o con meta datos que no son compatibles. Esto puede inducir a tener diferentes conclusiones acerca de los mismos datos.

11. Cuadrantes mágicos de Business intelligence Empresarial

Un cuadrante mágico fue una herramienta analítica creada y promovida por la empresa Gartner y la cuál muestra una representación gráfica del mercado compartido en un determinado periodo de tiempo. Los Cuadrantes Mágicos de Gartner proporcionan a las empresas un medio para identificar y diferenciar a los proveedores de servicios del sector de las tecnologías de la información.

Según define Gartner, los líderes en los cuadrantes mágicos son aquellos fabricantes de software que operan bien hoy día, tienen una visión clara de la dirección del mercado y desarrollan activamente las competencias necesarias para mantener su posición de líderes en el mercado.

A continuación aparecen los 2 cuadrantes mágicos proporcionados por Gartner con fecha de noviembre de 2004. El primero de ellos es referente a las plataformas de Business Intelligence y en segundo lugar para las suites de Business intelligence

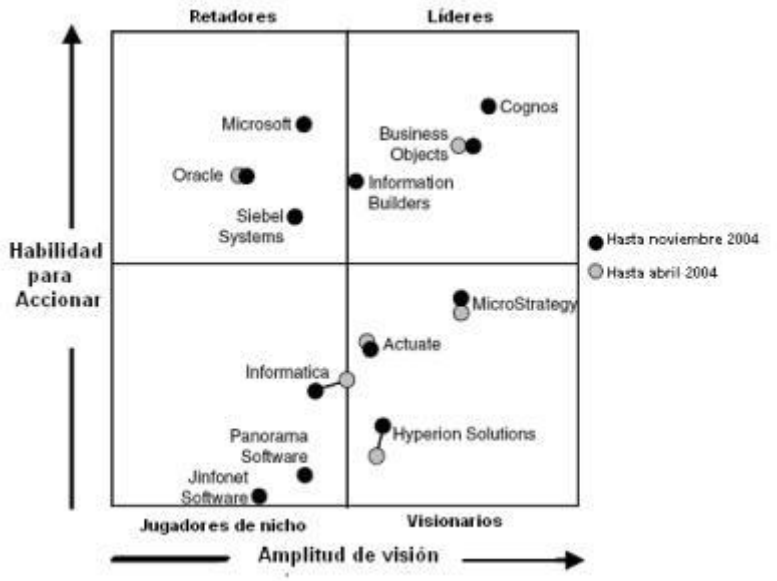
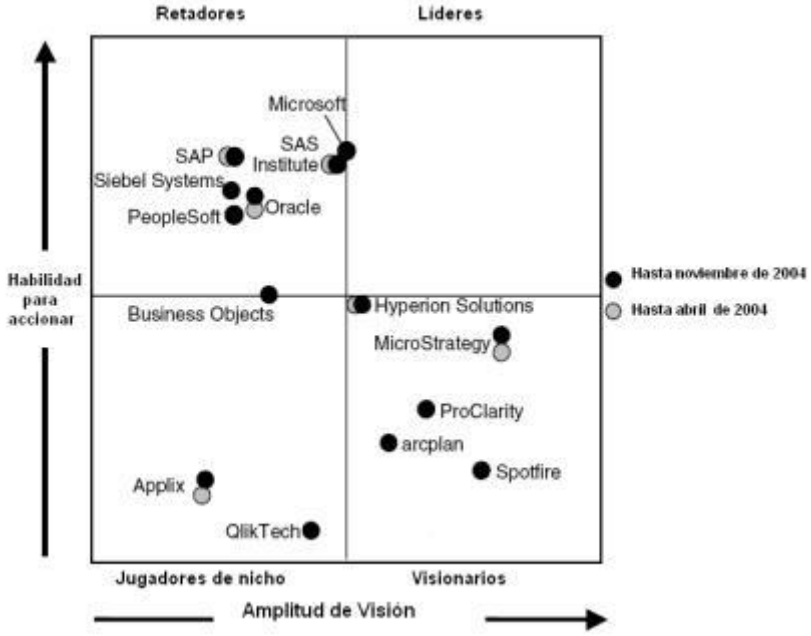


Figura 4. Cuadrante mágico de las plataformas de BI
Fuente: Gartner Research (Noviembre 2004)



La manera de interpretarlo según los especialista es de las siguiente manera los que figuran en el cuadrante principal pueden ofrecer un gran servicio prácticamente a cualquier cliente. Otros podrían ser empresas que abasteciesen nichos de mercado, por ello las notas tratan sobre los nichos de cada una de las empresas y describen los 'puntos favorables' de todas ellas.

En este caso, se está hablando principalmente de grandes clientes corporativos. Un Cuadrante Mágico no deja de ser potencialmente útil para pequeñas y medianas empresas (PYMES), pero éstas posiblemente

tengan que calibrar aspectos adicionales como, por ejemplo, 'el modo en que se dicho proveedor concreto se pondría en contacto conmigo'."

El cuadrante mágico debe tomarse como una herramienta y no como una guía específica de acción. En el caso de Business Intelligence Empresarial, el gran visionario hasta noviembre de 2004 es COGNOS. (Gartner, 2004)

Para Gartner, las empresas visionarias son aquellas que presentan un enfoque claro sobre la dirección del mercado y que orientan sus esfuerzos en este sentido, y que todavía pueden optimizar sus servicios. La consultora define el CPM como las metodologías, métricas, procesos y sistemas utilizados para monitorizar y gestionar el rendimiento de una empresa.

Las soluciones CPM de Cognos, que alinean la ejecución con la estrategia corporativa, se basan en: la solidez de Cognos Enterprise BI Series, la herramienta de Business Intelligence más completa de la industria: Cognos Enterprise Planning Series, su plataforma de planificación, de presupuestos, modelado y previsiones; y Cognos Metrics Manager, la solución para cuadro de mandos más robusta y flexible del mercado. (Cognos, 2005)

12. Hype cycle de Business intelligence Empresarial (Gartner, 2004)

Este ciclo también fue definido por Gartner para modelar la introducción y el desarrollo de nuevas tecnologías.

El Hype Cycle es un gráfico que mide a las diversas tecnologías según un ciclo de vida. Tiene como etapas el “disparador tecnológico” (cuando aparece el concepto en el mercado), “el pico de expectativa inflada” (cuando se habla mucho del concepto, pero está poco aplicado), “el valle de la desilusión” (cuando la herramienta está por debajo de lo que se esperaba de ella), “la pendiente de tolerancia” (el camino hacia la madurez) y el “plateau de productividad” (cuando alcanza la madurez).

Desde que Gartner publicó el primer Hype Cycle de BI, en diciembre de 2001, han ocurrido algunos cambios

BI basado en ERP descendió al valle de la desilusión. Sin embargo, es probable que ascienda al plateau de productividad. CRM Analítico (aCRM) mantiene su lugar y sus vendors están mudando su atención a otras áreas. Corporate Performance Management (CPM), que emergió el último año, está escalando rápidamente hacia el pico de expectativas infladas. Por su alto impacto en los procesos administrativos, es probable que la adopción masiva sea un proceso relativamente lento. Business Activity Monitoring (BAM) es otra tendencia que escala rápidamente hacia el pico de las expectativas infladas. Las plataformas BI, EBIS (Enterprise BI Suite), OLAP y los reportes de producción permanecen estables.

El Hype Cycle de BI muestra claramente que la innovación tecnológica precede a las aplicaciones. BI mobile puede resurgir como algo completamente diferente. Lo mismo puede ocurrir con BI Web Services, Distribución de BI basada en XML y BI colaborativo.

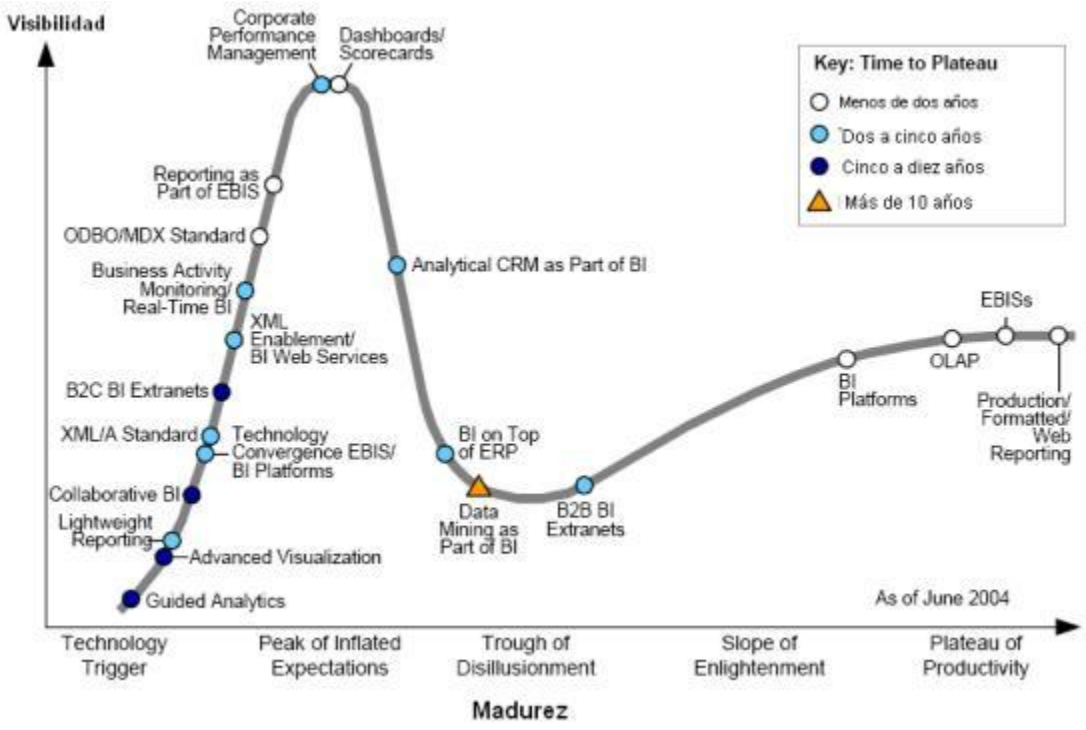


Figura 6. Hype cycle de BI empresarial
Fuente: Gartner Research (Noviembre 2004)

13. Conclusiones

Las organizaciones deben usar BI para apalancar las inversiones realizadas en años previos en aplicaciones empresariales que han derivado en el uso de enormes cantidades de datos; así de esta manera BI valida, mide y maneja nuevas oportunidades e inversiones en nuevos negocios.

Business Intelligence posiciona a una compañía para generar el mayor valor de las líneas de negocios existentes y anticipar nuevas oportunidades. Asimismo, los sistemas de Business intelligence le pueden ayudar a la compañía a reducir los ciclos de desarrollo de productos, agilizar operaciones, afinar campañas de marketing y mejorar relaciones con clientes y proveedores, todo lo cual significa menores costos y mayores márgenes de utilidad.

Con Business Intelligence, la compañía puede analizar tendencias que representan oportunidades nuevas e importantes y anticipar problemas potenciales y hacer ajustes antes de que se conviertan en un problema.

En la era digital, las capacidades que ofrece Business Intelligence será la diferencia entre el éxito y el fracaso.

Agrociencia
ISSN: 1405-3195
agrocien@colpos.mx
Colegio de Postgraduados
México

Aguado-Rodríguez, G. Javier; Quevedo-Nolasco, Abel; Castro-Popoca, Martiniano; Arteaga-Ramírez, Ramón; Vázquez-Peña, M. Alberto; Zamora-Morales, B. Patricia
PREDICCIÓN DE VARIABLES METEOROLÓGICAS POR MEDIO DE MODELOS ARIMA
Agrociencia, vol. 50, núm. 1, enero-febrero, 2016, pp. 1-13
Colegio de Postgraduados
Texcoco, México

PREDICCIÓN DE VARIABLES METEOROLÓGICAS POR MEDIO DE MODELOS ARIMA

METEOROLOGICAL VARIABLES PREDICTION THROUGH ARIMA MODELS

G. Javier **Aguado-Rodríguez**[1*], Abel **Quevedo-Nolasco**[1], Martiniano **Castro-Popoca**[1], Ramón **Arteaga-Ramírez**[2], M. Alberto **Vázquez-Peña**[2], B. Patricia **Zamora-Morales**[3]

[1]Hidrociencias. Campus Montecillo. Colegio de Postgraduados. 56230. Montecillo, Estado de México. (aguado.graciano@colpos.mx), (anolasco@colpos.mx), (mcastro@colpos.mx). [2]Irrigación. Universidad Autónoma Chapingo. 56230. Chapingo, Estado de México. (arteagar@correo.chapingo.mx), (mvazquezp@correo.chapingo.mx). [3]Instituto Nacional de Investigaciones Forestales, Agrícolas y Pecuarias. Km. 13.5. Carretera Los Reyes-Texcoco. 56250. Texcoco, Estado de México (zamora.patricia@inifap.gob.mx).

Resumen

La predicción de las variables meteorológicas se aplica en la agricultura al predecir el consumo de agua de las plantas para planear la lámina de riego. En esta investigación se elaboró un programa para realizar la predicción de la temperatura, radiación solar, evapotranspiración de referencia y humedad relativa con modelos autorregresivos integrados de media móvil (ARIMA) y se probó la efectividad del programa para realizar la predicción en condiciones de alta y baja precipitación. Los periodos de predicción evaluados fueron en marzo y en junio de 2013 en tres estaciones meteorológicas automáticas (EMAS) del Servicio Meteorológico Nacional (SMN). El análisis de los resultados indicó que la predicción de las variables meteorológicas con modelos ARIMA fue mejor que con la predicción persistente en el periodo con condiciones de baja precipitación (marzo).

Palabras clave: Prónostico, R Statictics, tiempo real.

Introducción

Hay grandes progresos en el desarrollo y las aplicaciones de la predicción del clima a mediano plazo y su predicción estacional (Vitart *et al.* 2012). Los algoritmos de predicción automáticas más usados son con base en el suavizado exponencial o modelos autorregresivos integrados de media móvil (ARIMA) (Hyndman y Khandakar, 2008). Box y Jenkins (1976) desarrollaron la metodología clásica que emplea las series de tiempo para

* Autor responsable ❖ Author for correspondence.
Recibido: septiembre, 2014. Aprobado: agosto, 2015.
Publicado como ARTÍCULO en Agrociencia 50: xxx-xxx. 2016.

Abstract

Meteorological variables prediction is applied in agriculture to predict water uptake of plants for planning irrigation depths. In the present study a program was made for the prediction of temperature, solar radiation, reference evapotranspiration and relative humidity by means of autoregressive integrated mobile media models. The effectiveness of the program was tested for prediction under high and low rainfall conditions. The prediction periods evaluated were in March and in June, 2013, in three automatic meteorological stations (EMAS) of the National Meteorological Service (SMN). The analysis of results indicated that the prediction of meteorological variables with ARIMA models was better than with persistent prediction in the period with low rainfall conditions (March).

Key words: Prediction, R Statistics, real time.

Introduction

There is great progress in the development and applications of medium term weather prediction and seasonal climate (Vitart *et al.*, 2012). The most frequently used automatic prediction algorithms are based on the softened exponential or autoregressive integrated mobile media models (ARIMA) (Hyndman and Khandakar, 2008). Box and Jenkins (1976) developed the classic methodology that uses the time series for generating models such as the autoregressive mobile media model (ARMA) or also the ARIMA model for obtaining predictions.

Karl *et al.* (2000) found an increment in the global warming rate using the time series of global mean

generar modelos como el autoregresivo de media móvil (ARMA) o también el modelo ARIMA para obtener predicciones.

Karl *et al*. (2000) reportaron un aumento en la tasa de calentamiento global usando la serie de tiempo de la temperatura media global indicada por Quayle *et al*. (1999), por medio del análisis de valores mensuales de temperatura y con modelos ARMA. Reikard (2009) investigó la predicción de la radiación solar en intervalos de tiempos de 5 min hasta varias horas y aunque los datos exhibieron variabilidad no lineal debido a la nubosidad, en la mayoría de las pruebas se obtuvieron los mejores resultados usando los modelos ARIMA. Pulido (2002) propuso estimar la demanda de agua en las próximas 24 h en un sistema de distribución de agua para riego usando modelos ARIMA y otros modelos. Para predecir la lluvia del monzón de verano en la India, Chattopadhyay y Chattopadhyay (2010) identificaron un modelo ARIMA como adecuado, pero el modelo de redes neuronales autorregresivas (ARNN) proporcionó mejores predicciones, mientras que Narayanan *et al*. (2013) usaron modelos ARIMA para predecir las lluvias antes del monzón en el oeste de la India.

Debido a que los modelos ARIMA son una herramienta para realizar predicción de series de tiempo univariadas, en esta investigación se propuso elaborar un programa de cómputo que calcule la predicción en tiempo real de variables meteorológicas usando modelos ARIMA y probar su efectividad en condiciones de baja y alta precipitación.

Materiales y Métodos

Para esta investigación se usó una computadora con procesador de 2.2 GHz, 2 GB de memoria RAM y sistema operativo Windows 7®. En la computadora se instaló: el programa de cómputo MySQL Server®, que es un gestor de bases de datos para almacenar información (Korhonen *et al.*, 2008); Microsoft Visual Studio 2010® que es un conjunto completo de herramientas de desarrollo para la generación de aplicaciones Web ASP.NET, Servicios Web XML, aplicaciones de escritorio y aplicaciones móviles (Randolph *et al.,* 2010); MySQL Conector Net 6.3.5® que es un conector del programa Microsoft Visual Studio 2010® con MySQL Server (Kofler, 2005); R Statistics 2.15.3®, un paquete de cómputo estadístico (Dalgaard, 2008); librerías 'rcom' y 'rscproxy'del programa R Statistics 2.15.3 (conectores del programa R Statistics 2.15.3 con Microsoft Visual Studio 2010); y la librería 'forecast' del programa R Statistics 2.15.3, se usó para la estimación y predicción de los modelos ARIMA.

temperature indicated by Quayle *et al.* (1999), using the analysis of monthly values of temperature and with ARMA models. Reikard (2009) investigated the prediction of solar radiation in 5 min time intervals for various hours, and although the data exhibited non-linear variability due to cloudiness, in most of the tests best results were obtained using the RIMA models. Pulido (2002) proposed the estimation of water demand in the next 24 h in a water distribution system for irrigation using ARIMA and other models. To predict rainfall of the summer monsoon in India, Chattopadhyay and Chattopadhyay (2010) identified an ARIMA model as adequate, but the autoregressive neuronal network model (ARNN) provided better predictions, while Narayanan *et al.* (2013) used ARIMA models to predict rainfall prior to the monsoon in western India.

Because the ARIMA models are a tool used for univariate weather prediction, the present investigation was made with the purpose of elaborating a computer program that calculates prediction in real time of meteorological variables using ARIMA models and testing its effectiveness under low and high rainfall conditions.

Materials and Methods

The present investigation used a computer with a processor of 2.2 GHz, 2 GB of RAM memory and Windows 7® operative system. The following programs were installed: MySQL Server®, which is an administrator of data bases for storing information (Korhonen *et al.*, 2008); Microsoft Visual Studio 2010®, which is a complete set of development tools for the generation of applications of Web ASP.NET, XML Web Services, desktop and mobile applications (Randolph *et al.*, 2010); MySQL Connector Net 6.3.5® which is a connector of the program Microsoft Visual Studio 2010® with MySQL Server (Kofler, 2005); R Statistics 2.15.3®, computer statistical package (Dalgaard, 2008); 'rcom' and 'rscproxy' libraries of the program R Statistics 2.15.3 (connectors of the program R Statistics 2.15.3 with Microsoft Visual Studio 2010); and the 'forecast' library of the program R Statistics 2.15.3, which was used for the estimation and prediction of the ARIMA models.

To store meteorological information, a data base was made integrated with two data tables, in the program MySQL Server (Figure 1). The first data table was called 'station' and was used to store the information of each meteorological station, and for each station an identifier is required of station, latitude, longitude, altitude and name, and the primary key is the station identifier. The second data table, called 'elemhoraria' was used

Para almacenar información meteorológica se elaboró una base de datos integrada con dos tablas de datos, en el programa MySQL Server (Figura 1). La primera tabla de datos se denominó 'estación' y fue usada para guardar la información de cada estación meteorológica, y por cada estación se requiere un identificador de estación, latitud, longitud, altitud y nombre, y la llave primaria es el identificador de estación. La segunda tabla de datos, denominada 'elemhoraria', se usó para almacenar la información de los datos meteorológicos a nivel horario de estaciones meteorológicas; los datos almacenados en esta tabla son: fecha y hora, evapotranspiración (ET_0 en mm), velocidad del viento (VELS en m/s), precipitación (mm), radiación solar (RADSOL en W/m^2), temperatura media (TEMP en °C), humedad relativa (HR en %), y un identificador de estación de la cual provienen los datos; la llave primaria es la unión de los datos de fecha e identificador de estación.

En la Figura 1 se observa que una estación puede tener muchos registros a nivel horario y muchas estaciones pueden tener datos meteorológicos para una hora en particular.

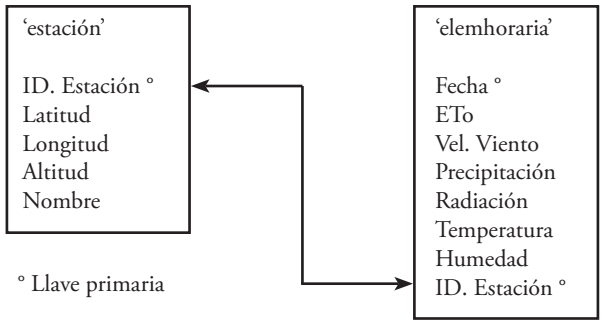


Figura 1. Base de datos.
Figure 1. Data base

Datos meteorológicos

Para comprobar la bondad predictiva de los modelos ARIMA, se usaron datos de tres estaciones meteorológicas automáticas (EMAS) del Servicio Meteorológico Nacional, México, para el 2013. Las EMAS fueron: ENCB. II del IPN, ubicada en 19° 29' 55" N, 99° 08' 43" O y altitud de 2240 m; Acolman, ubicada en 19° 38' 05" N, 98° 54' 42" O y altitud de 2269 m; Chapingo, ubicada en 19° 29' 39" N, 98° 53' 19" O y altitud de 2260 m.

En las EMAS para este estudio hay datos continuos a nivel horario de cinco variables meteorológicas en dos periodos: el primero es del 7 de marzo de 2013 a las 16:00 h y el 17 de marzo de 2013 a las 15:00 h; el segundo es del 16 de junio de 2013 a las 16:00 h y 26 de junio de 2013 a las 15:00 h. Las variables meteorológicas obtenidas de las EMAS fueron: velocidad del viento (m/s), precipitación (mm), radiación solar (W/m^2), temperatura media (°C), humedad relativa (%). Además se calculó la evapotranspiración de referencia (ET0) por el método de Penman Monteith (Allen, 2006) con los datos anteriores.

Modelos ARIMA

Según Pankratz (1983), los modelos ARIMA sirven para predecir series simples (de una sola variable), en los que las predicciones de los modelos ARIMA están basadas sólo en valores pasados de la variable a predecir. Los modelos ARIMA se pueden usar para hacer predicciones a corto plazo porque la mayoría de ellos ponen mayor énfasis en el pasado reciente que en el pasado distante; se aplican a variables discretas o continuas, aunque el tiempo debe ser igualmente espaciado y en intervalos discretos;

to store the information of the meteorological data at the time level of meteorological stations; the data stored in this table are: date and time, evapotranspiration (ET_0 in mm), wind velocity (VELS in m/s), rainfall (mm), solar radiation (SOLRAD in W/m^2), mean temperature (TEMP in °C), relative humidity (RH in %), and an identifier of the station from which the data is from: the primary key is joining the data of date and station identifier.

In Figure 1, it is observed that a station can have many records at the hourly level and many stations can have meteorological data for a particular hour.

Meteorological data

To test the predictive goodness of the ARIMA models, data were used from three automatic meteorological stations (EMAS) of the National Meteorological Service, Mexico, for 2013. The EMAS considered were as follows: ENCB. II of the IPN, located at 19° 29' 55" N, 99° 08' 43" W and altitude of 2240 m; Acolman, located at 19° 38' 05" N, 98° 54' 42 W and altitude of 2269 m; Chapingo, located at 19° 29' 39" N, 98° 53' 19" W and altitude of 2260 m.

In the EMAS for this study there are continuous data at the hourly level of five meteorological variables in two periods: the first is of March 7, 2013 at 16:00 h and March 17, 2013 at 15:00 h; the second is of June 16 , 2013 at 16:00 h and June 26, 2013 at 15:00 h. The meteorological variables obtained from the EMAS were as follows: wind velocity (m/s), rainfall (mm), solar radiation (W/m^2), mean temperature (°C), relative humidity (%). In addition, reference evapotranspiration (ET0) was calculated by the Penman Monteith method (Allen, 2006) with the above data.

ARIMA Models

According to Pankratz (1983), the ARIMA models serve to predict simple series (of a single variable), in which the

son útiles para predecir series de datos que contienen variación estacional (u otras variaciones periódicas), incluyendo aquellas con patrones estacionales cambiantes; requieren como mínimo alrededor de 50 observaciones; se aplican sólo a series de datos estacionarios, y una serie de tiempo estacionaria tiene una media, varianza y función de autocorrelación constantes a través del tiempo (Pankratz, 1983).

El requisito de una serie de tiempo estacionaria puede parecer enteramente restrictiva, pero la mayoría de las series no estacionarias en la práctica se pueden transformar a una serie estacionaria a través de una transformación llamada "diferenciar", la cual es una operación relativamente simple que envuelve el cálculo de cambios sucesivos en los valores de las series de datos. Los cambios en la serie de datos se conocen como (ω_t) y se obtienen con la ecuación $\omega_t=z_t - z_{t-1}$, donde z representa los valores de la serie de datos. Con las diferencias se construye una nueva serie diferente de la serie original, y una "diferencia" es cuando la media de una serie de datos cambia con el tiempo. Es posible "diferenciar" más de una sola vez para obtener una serie estacionaria. Al ya tener una serie estacionaria, se realiza la búsqueda por un buen modelo ARIMA y consiste en: identificación, estimación, diagnóstico del modelo; y si el modelo es adecuado se realiza la predicción (Pankratz, 1983).

Descripción del procedimiento de la librería Forecast para estimar el modelo ARIMA

Según Hyndman *et al*. (2013), un obstáculo común al usar modelos ARIMA para predecir es que el proceso de selección del orden es generalmente considerado subjetivo y difícil de aplicar. Por tanto, se elaboró la librería Forecast para elegir el orden del modelo de manera automática, y donde los algoritmos son aplicables a ambos, datos estacionales y no-estacionales.

Para Hyndman *et al*. (2013) un proceso ARIMA(p,d,q) no-estacional está dado por:

$$(1-B^d)y_t = c + \phi(B)y_t + \phi(B)\varepsilon_t$$

donde $\{\varepsilon_t\}$ es un proceso de ruido blanco con media cero y varianza σ^2, B es el operador de retraso, y $\phi(z)$ y $\phi(z)$ son polinomios de orden p y q, respectivamente. Para asegurar casualidad e invertibilidad se asume que $\phi(z)$ y $\phi(z)$ no tienen raíces para $|z|<1$. Si $c\neq 0$, hay un polinomio implícito de orden d en la función de predicción. El proceso estacional ARIMA (p,d,q) $(P,D,Q)_m$ está dado por

$$(1-B^m)^D (1-B)^d y_t = c + \phi (B^m) \phi(B) y_t + \Theta(B^m) \theta(B) \varepsilon_t$$

predictions of the ARIMA models are based only on past values of the variable for prediction. The ARIMA models can be used to make short term predictions because most of them place more emphasis on the recent past than on the distant past; they are applied to discrete or continuous variables, although time should be equally spaced and in discrete intervals; they are useful for predicting data series that contain seasonal variation (or other periodic variations), including those with changing seasonal patterns; they require a minimum of 50 observations; it is applied only to series of stationary data, and a series of stationary time has a mean, variance and function of autocorrelation that are constant through time (Pankratz,1983).

The requirement of a stationary time series may seem totally restrictive, but most of the non-stationary series in practice can be transformed into a stationary series through a process called "differentiation", which is a relatively simple operation that involves the calculation of successive changes in the values of the data series. The changes in the data series are known as (ω_t) and are obtained with the equation $\omega_t=z_t-z_{t-1}$, where z represents the values of the data series. With the differences a new series is constructed, different from the original, and a "difference" is when the mean of a series of data changes with time. It is possible to "differentiate" more than just once to obtain a stationary series. When a stationary series is obtained, a good ARIMA model is sought and consists of : identification, estimation, diagnostic of the model, and if the model is adequate the prediction is made (Pankratz, 1983).

Description of the procedure of the Forecast library for estimating the ARIMA model

According to Hyndman *et al.* (2013), a common obstacle when using ARIMA models for prediction is that the selection process of the order is generally considered subjective and difficult to apply. Therefore, the Forecast library was made to select the order of the model automatically, and where the algorithms are applicable to both stationary and non-stationary data.

For Hyndman *et al.* (2013), a non-stationary ARIMA process (p,d,q) is obtained by:

$$(1-B^d)y_t = c + \phi(B)y_t + \phi(B)\varepsilon_t$$

where $\{\varepsilon_t\}$ is a white noise process with mean zero and variance σ^2, B is the delay operator, and $\phi(z)$ and $\phi(z)$ are polynomials of order p and q, respectively. To insure causality and invertibility, it is assumed that $\phi(z)$ and $\phi(z)$ do not have roots for $|z|<1$. If $c\neq 0$, there is an implicit polynomial of d order in the prediction function. The seasonal process ARIMA$(p,d,q)(P,D,Q)_m$ is obtained as follows:

donde $\phi(z)$ y $\phi(z)$ son polinomios de orden P y Q respectivamente, cada uno no conteniendo raíces dentro del círculo unitario. Si c≠0, hay un polinomio implícito de orden *d*+*D* en la función de predicción.

La tarea principal de la librería Forecast que Hyndman *et al*. (2013) realizan en la predicción automática del modelo ARIMA, es seleccionar un apropiado orden de modelo y son los valores de *p, q, P, Q, D, d*. Si *D* y *d* conocidos. Los órdenes *p, q, P* y *Q* se pueden seleccionar por medio de un criterio de información como el Criterio de Información de Akaike (AIC):

$$AIC = -2\log(L) + 2(p+q+P+Q+k)$$

donde k=1 si c≠0 y 0 de otra manera, y L es la probabilidad maximizada del modelo ajustado a los datos diferenciados $(1-B^m)^D (1-B)^d y_t$.

Para propósitos de predicción Hyndman *et al*. (2013) indican que es mejor hacer tan pocas diferencias como sea posible. Para datos no estacionales Hyndman *et al*. (2013) consideran modelos ARIMA(*p,d,q*) donde *d* es seleccionada basándose en el test de raíces unitarias sucesivas KPSS (Kwiatkowski *et al*., 1992). El método prueba los datos para una raíz unitaria; si el resultado de la prueba es significativa, se prueban los datos diferenciados para para una raíz unitaria; y así sucesivamente.

Para datos estacionales, en la librería Forecast se consideran modelos ARIMA $(p,d,q)\ (P,D,Q)_m$ donde m es la frecuencia estacional y D=0 o D=1, dependiendo de una prueba extendida de Canova-Hansen (Canova and Hansen, 1995). Después de seleccionar *D* se elige *d* aplicando el test de raíces unitarias sucesivas KPSS a los datos estacionales diferenciados (si *D*=1) o a los datos originales (si *D*=0).

Estimación de la predicción

Con Microsoft Visual Studio 2010 se desarrolló una aplicación ejecutable (.exe), con la cual se realizan funciones para la predicción de las variables meteorológicas. Sin embargo, antes de describir dichas funciones es importante remarcar que la mayoría de las EMAS tienen la opción de descargar la información que registran y guardarla en archivos de texto. Por lo anterior, se realizó la primera función para extraer los datos de las variables meteorológicas almacenados en archivos de texto (de cada EMAS) y guardarlos en la base de datos. En la base de datos se almacenan los datos meteorológicos a nivel horario de distintas EMAS y se organizan por fecha y por identificador de EMA; los datos obtenidos se guardan en la tabla de datos 'elemhoraria'. Con esta función también se calcula la evapotranspiración de referencia

$$(1-B^m)^D (1-B)^d y_t = c + \phi(B^m)\phi(B)\, y_t + \Theta(B^m)\theta(B)\,\varepsilon_t$$

where $\phi(z)$ and $\phi(z)$ are polynomials of order P and Q, respectively, neither one containing roots within the unitary circle. If c≠0, there is an implicit polynomial of order *d*+*D* in the prediction function.

The principal task of the Forecast library that Hyndman *et al.* (2013) carry out in the automatic prediction of the ARIMA model is to select an appropriate order of model and they are the known values of *p, q, P, Q, d.* If *D* and *d* are known. The orders *p, q, P,* and *Q* can be selected by means of a criterion of information such as the Akaike Information Criterion (AIC):

$$AIC = -2\log(L) + 2(p+q+P+Q+k)$$

where k=1 if c≠0 and 0 otherwise, and L is the maximized probability of the model fitted to the differentiated data $(1-B^m)^D (1-B)^d y_t$.

For purposes of prediction, Hyndman *et al.* (2013) point out that it is better to make the fewest differences possible. For non-seasonal data, Hyndman *et al.* (2013) consider ARIMA (*p,d,q*) models where *d* is selected based on the test of successive unitary roots KPSS (Kwiatkowski *et al.*, 1992). The method tests the data for a unitary root; if the result of the test is significant, differentiated data are tested for a unitary root; and so on.

For seasonal data, in the Forecast library ARIMA$(p,d,q)\ (P,D,Q)_m$ models are considered where m is the seasonal frequency and D = 0 or D = 1, depending on an extended test of Canova-Hansen (Canova and Hansen, 1995). After selecting *D*, *d* is selected applying the test of successive unitary roots KPSS to the differentiated seasonal data (if *D =1*) or to the original data (if *D*=0).

Estimation of the prediction

With Microsoft Visual Studio 2010 a usable application was developed (.exe), with which functions are made for the prediction of the meteorological variables. However, before describing these functions it is important to emphasize that most of the EMAS have the option of downloading the information they record and saving it in text files. Therefore, a first function was made for extracting the data of the meteorological variables stored in text files (of each EMAS) and storing them in the data base. The meteorological data are stored in the data base at the schedule level of different EMAS and are organized by date and identifier of EMA; the data obtained are stored in the data table 'elemhoraria'. Its function is also used to calculate reference

por el método de Penman Monteith (Allen, 2006) y el resultado se almacena en la misma tabla de datos.

Cuando termina la primera función se tienen los promedios de las variables meteorológicas a nivel horario en la base de datos. Con la segunda función se genera una serie de tiempo por cada variable meteorológica de las tres EMAS, y la serie de tiempo así generada tiene 60 datos. Cada dato de la serie de tiempo consiste en el promedio de dos horas; por ejemplo, si para el día 16/06/2013 a las 16:00, 17:00, 18:00 y 19:00 h el promedio de temperatura fue 22, 23.7, 24.7 y 24.8 °C respectivamente, y el día 21 de junio de 2013 a las 14:00 y 15:00 h el promedio de temperatura fue 16.2 y 15.6 °C, respectivamente, entonces la serie de tiempo tendrá los valores 22.85, 24.75, …, 15.9 °C, con un total de 60 datos. Las series de tiempo generadas se almacenan en un archivo de texto creado automáticamente con extensión '.txt' para cada EMA (Figura 2); el archivo tiene seis columnas (una columna por cada variable meteorológica) y 61 filas. La primer fila contiene los nombres de las variables meteorológicas; sin embargo, sólo se analizaron cuatro variables. La primera variable está en la primer columna y tiene los datos de evapotranspiración de referencia (mm), en la cuarta columna están los datos de radiación solar (W/m^2), en la quinta columna están los datos de la temperatura (°C), y en la sexta columna están los datos de humedad relativa (%).

Al terminar la segunda función se tienen las series de tiempo para cada variable meteorológica necesaria para realizar la predicción. Con la tercera función se realiza la predicción. En el primer paso de la tercera función se establece una conexión entre el programa R Statistics 2.15.3 con Microsoft Visual Studio 2010; el lenguaje de programación fue C#. Después se envía un comando al programa R Statistics 2.15.3 para hacer el ajuste de las series de tiempo, de cada variable meteorológica, a un modelo ARIMA usando la función "auto.arima" de la librería Forecast (Hyndman *et al.*, 2013); después de lo anterior, ya con el modelo ARIMA estimado automáticamente, se envía un comando al programa

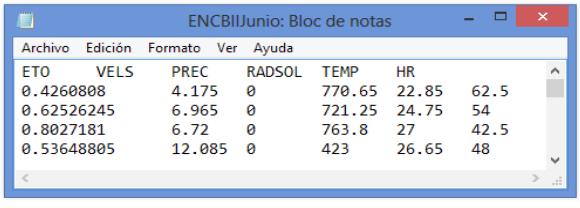


Figura 2. Contenido del archivo de texto con series de tiempo de variables meteorológicas para la EMA ENCBII para el periodo de junio.

Figure 2. Content of the text file with time series of meteorological variables for the EMA ENCBII for the period of June.

evapotranspiration by the Penman Monteith method (Allen, 2006) and the result is stored in the same data table.

When the first function is completed, the averages of the meteorological variables are obtained at the hour level in the data base. The second function is used to generate a time series for each meteorological variable of the three EMAS, and the resulting time series contains 60 data. Each data of the time series consists of the average of two hours. For example, if for day 16/06/2013 at 16:00, 17:00, 18:00 and 19:00 the average temperature was 22, 23.7, 24.7 and 24.8 °C, respectively, and on June 21 of 2013 at 14:00 and 15:00 h the average temperature was 16.2 and 15.6 °C, respectively, then the time series will have the values 22.85, 24.75, …, 15. °C, with a total of 60 data. The generated time series are stored in a text file created automatically with extension '.txt' for each EMA (Figure 2). The file has six columns (one column for each meteorological variable) and 61 days. The first row contains the names of the meteorological variables; however, only four variables are analyzed. The first variable is in the first column and has the data of reference evapotranspiration (mm), the fourth column has the data of solar radiation (W/m^2), the fifth column contains the data of temperature (1C) and the sixth column includes the data of relative humidity (%).

When the second function is finished, we have obtained the time series for each meteorological variable necessary for carrying out the prediction. The prediction is obtained with the third function. In the first step of the third function a connection is established between the program R Statistics 2.15.3 and Microsoft Visual Studio 2010; the programming language was C#. Then a command is sent to the program R Statistics 2.15.3 to fit the time series of each meteorological variable to an ARIMA model using the function "auto.arima" of the Forecast library (Hyndman *et al.*, 2013). Next, using the automatically estimated ARIMA model, a command is sent to the program R Statistics 2.15 to make the prediction of the next 60 elements in the time series.

The function "auto.arima" of the Forecast library (Hyndman *et al.*, 2013) gives back the best ARIMA model. However, the function "auto.arima" requires arguments such as a univariate time series, the order of the first difference "d" (if it is not put in, the function "auto.arima" selects a value according to the KPSS test), the order of the first seasonal difference "D" (if it is not put in, the function "auto.arima" calculates it), the maximum value for *p, q, P, Q,* the initial value of *p, q, P, Q* (optional), if the time series is stationary, if the time series is seasonal, among other optional data.

For the function "auto.arima" the options were specified of maximum value of *p* equal to 5, maximum value of *q* equal to 5, and the seasonal option equal to "TRUE", because otherwise the search for non-seasonal models is restricted. The last function consists of saving the predictions obtained with the function

R Statistics 2.15 para hacer la predicción de los 60 elementos siguientes en la serie de tiempo.

La función "auto.arima" de la librería Forecast (Hyndman *et al.*, 2013) regresa el mejor modelo ARIMA. No obstante, la función "auto.arima" requiere argumentos como una serie de tiempo univariada, el orden de la primera diferencia "d" (si no se coloca, la función "auto.arima" elige un valor de acuerdo con la prueba KPSS), el orden de la primera diferencia estacional "D" (si no se coloca, la función "auto.arima" lo calcula), el máximo valor para *p, q, P, Q*, el valor inicial de *p, q, P, Q* (opcional), si la serie de tiempo estacionaria, si la serie de tiempo es estacional, entre otros datos opcionales.

Para la función "auto.arima" se especificaron las opciones valor máximo de *p* igual a 5, valor máximo de *q* igual a 5, y la opción estacional igual a "TRUE" porque de lo contrario se restringe la búsqueda para modelos no-estacionales. La última función consiste en guardar las predicciones obtenidas con la función "auto.arima". Para lo anterior, se crean archivos de texto con extensión '.txt' y en estos se almacenan las predicciones; el nombre del archivo se crea con la unión del nombre de la EMA, el nombre de la variable y la palabra PRED al final (para indicar que es predicción). En la Figura 3 se muestra el archivo generado de la estación ENCBII para la variable temperatura del aire.

En el archivo texto generado por la predicción (Figura 3), en la primera línea, la predicción del promedio de la variable entre las 16:00 y 17:00 h el 21 de junio de 2013, en la segunda línea está la predicción del promedio de la variable entre las 18:00 y 19:00 h el 21 de junio de 2013, y en la última fila del archivo está la predicción del promedio de la variable entre las 14:00 y 15:00 h del día 26 de junio de 2013.

Resultados y Discusión

La precisión de la predicción en los dos periodos fue evaluada. En el primero, los datos observados entre 7 de marzo de 2013 a las 16:00 h hasta el 12 de marzo de 2013 a las 15:00 h, se usaron para generar la serie de tiempo, y los datos observados entre 12 de marzo de 2013 a las 16:00 y 17 de marzo de 2013 a las 15:00 h se usaron para compararlos con los datos estimados con el modelo de predicción. Después se evaluó la precisión de la predicción en el segundo periodo. Los datos observados entre el16 de junio de 2013 a las 16:00 h y el 21 de junio de 2013 a las 15:00 h se usaron para generar la serie de tiempo. Los datos observados entre 21 de junio de 2013 a las 16:00 h y el 26 de junio de 2013 a las 15:00 h se usaron para compararlos con los datos de las predicciones estimadas.

"auto.arima". To carry this out, text files are created with extension '.txt' where the predictions are stored. The name of the file is created with the combination of the name of the EMA, the name of the variable and the word PRED at the end (to indicate that it is prediction). Figure 3 shows the name of the file generated of the station ENCBII for the variable air temperature.

In the text file generated by the prediction (Figure 3), in the first line, the prediction of the average of the variable between 16:00 and 17:00 h on June 21 of 2013, in the second line is the prediction of the average of the variable between 18:00 and 19:00 h on June 21 of 2013, and in the last row of the file is the prediction of the average of the variable between 14:00 and 15:00 h of June 26 of 2013.

Results and Discussion

The precision of the prediction in the two periods was evaluated. In the first, the data observed between March 7 of 2013 at 16:00 h until March 12 of 2013 at 15:00 h were used to generate the time series, and the data observed between March 12 of 2013 at 16:00 and March 17 of 2013 at 15:00 h were used to be compared with the data estimated with the prediction model. Next, the precision of the prediction in the second period was evaluated. The data observed between June 16 of 2013 at 16:00 h and June 21 of 2013 at 15:00 were used to generate the time series. The data observed between June 21 of 2013 at 16:00 h and June 26 of 2013 at 15:00 h were used to compare with the data of the estimated predictions.

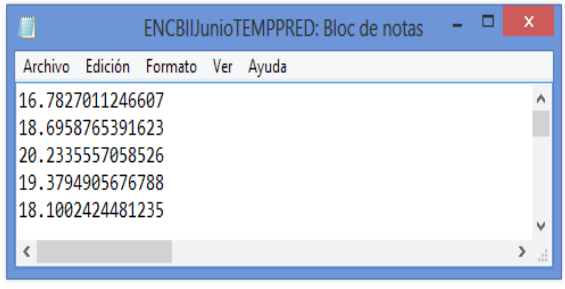


Figura 3. Archivo generado con datos de la predicción estimada con la función "auto.arima" para la EMA ENCBII en el periodo de junio y la variable meteorológica temperatura del aire (°C).

Figure 3. File generated with data of the prediction estimated with the function "auto.arima" for the EMA ENCBII in the period of June and the meteorological variable air temperature (°C).

En las Figuras 4 y 5 se graficaron los datos observados con los datos estimados de las variables meteorológicas: humedad relativa (%), temperatura del aire (°C), radiación solar (W/m^2) y evapotranspiración de referencia (mm). Los datos tomados como observados de evapotranspiración de referencia fueron calculados con el método de Penman Monteith (Allen, 2006) y usando los datos observados de las demás variables meteorológicas.

In Figures 4 and 5 the observed data were graphed with the estimated data of the meteorological variables: relative humidity (%), air temperature (°C), solar radiation (W/m^2) and reference evapotranspiration (mm). The data obtained as observed from reference evapotranspiration were calculated with the method of Penman Monteith (Allen, 2006) and using the observed data of the other meteorological variables.

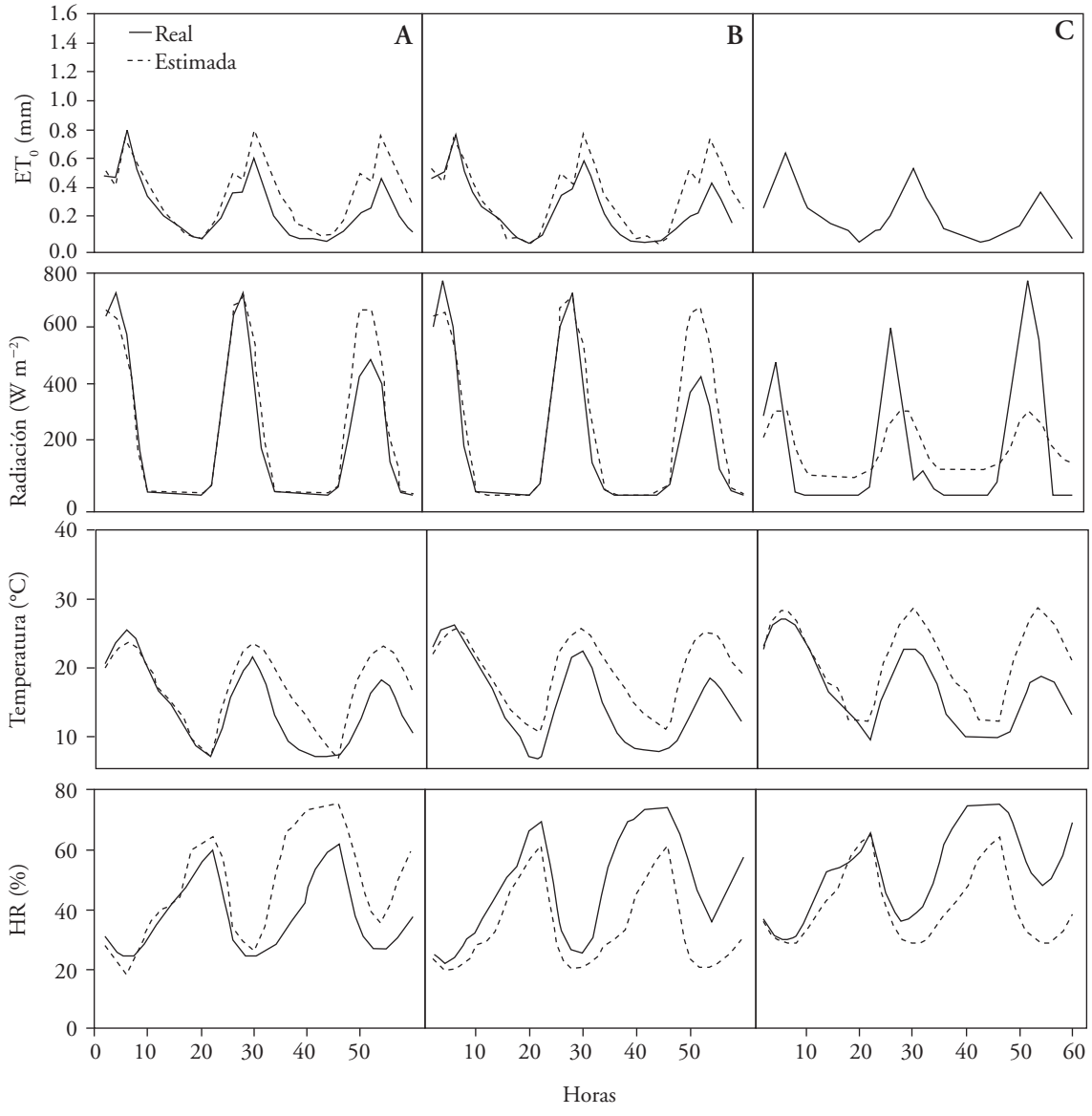


Figura 4. Predicción y variables observadas 60 h hacia adelante para la EMA Acolman (A), Chapingo (B), y ENCB. II del IPN (C) en el primer periodo (marzo).

Figure 4. Prediction and variables observed 60 h ahead for the EMA Acolman (A), Chapingo (B), and ENCB. II of the IPN (C) in the first period (March).

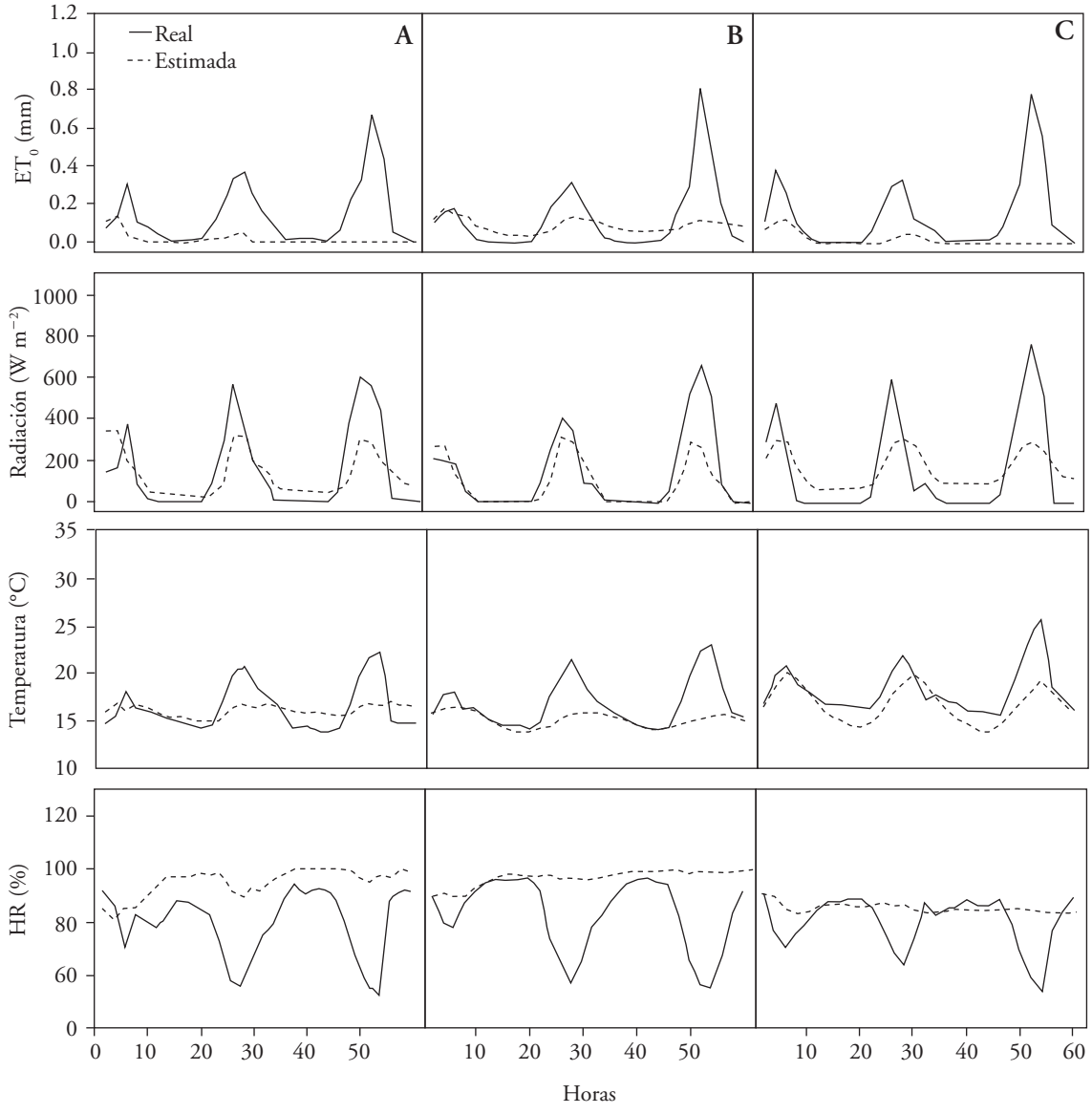


Figura 5. Predicción y variables observadas 60 h hacia adelante para la EMA Acolman (A), Chapingo (B), y ENCB. II del IPN (C) en el segundo periodo (junio).

Figure 5. Prediction and variables observed 60 h forward for the EMA Acolman (A), Chapingo (B), and ENCB II of the IPN (C) in the second period (June).

Los modelos ARIMA obtenidos para cada estación y variable meteorológica están en el Cuadro 1.

Para una serie de tiempo dada $\{y_n\}$, la predicción persistente se obtiene al colocar y(n+1)=y(n), que implica que el promedio de la variable para la siguiente hora es igual al promedio de la variable en la hora actual (Kavasseri *et al*., 2009).

Para comparar la bondad predictiva de los modelos ARIMA con la predicción persistente se calcularon

The ARIMA models obtained for each station and meteorological variable are found in Table 1.

For a given time series $\{y_n\}$, the persistent prediction is obtained by placing y(n+1)=y(n), which implies that the average of the variable for the next hour is equal to the average of the variable in the present hour (Kavasseri *et al*., 2009).

To compare the predictive goodness of the ARIMA models with the persistent prediction,

Cuadro 1. Modelos ARIMA para las series de tiempo de las variables meteorológicas evapotranspiración (ET0), humedad relativa (HR), radiación solar (RADSOL) y temperatura del aire (TEMP).

Table 1. ARIMA models for the time series of the meteorological variables evapotranspiration (ET0), relative humidity (RH), solar radiation (SOLRAD) and air temperature (TEMP).

Estación	Variable	ϕ_1	ϕ_2	θ_1	θ_2	θ_3	Φ_1	Θ_1	ARIMA $(p, d, q)(P, D, Q)_{12}$
					Periodo de marzo				
Acolman	ET0	0.38					−0.58		(1, 0, 0)(1, 1, 0)
Chapingo	ET0	0.68							(1, 0, 0)(0, 1, 0)
ENCBII	ET0								Sin ajuste
Acolman	HR			0.86	0.69	0.45	−0.67		(0, 0, 3)(1, 1, 0)
Chapingo	HR	1.11	−0.35				−0.28		(2, 0, 0)(1, 1, 0)
ENCBII	HR			−0.44				−0.45	(0, 1, 1)(0, 1, 1)
Acolman	RADSOL	0.81	−0.26				−0.72		(2, 0, 0)(1, 1, 0)
Chapingo	RADSOL	0.40	−0.25				−0.67		(2, 0, 0)(1, 1, 0)
ENCBII	RADSOL			−0.27					(0, 0, 1)(0, 1, 0)
Acolman	TEMP							−0.61	(0, 1, 0)(0, 1, 1)
Chapingo	TEMP	1.45	−0.62	−0.43			0.97	−0.50	(2, 0, 1)(1, 0, 1)
ENCBII	TEMP			0.76					(0, 0, 1)(0, 1, 0)
					Periodo de junio				
Acolman	ET0	1.63	−0.87	−1.68	0.78		0.73		(2, 1, 2)(1, 0, 0)
Chapingo	ET0	1.19	−0.51	−0.96			0.57		(2, 1, 1)(1, 0, 0)
ENCBII	ET0			0.22			0.60		(0, 1, 1)(1, 0, 0)
Acolman	HR	0.34					0.68		(1, 1, 0)(1, 0, 0)
Chapingo	HR	0.35					0.38		(1, 1, 0)(1, 0, 0)
ENCBII	HR	0.72		−0.50	-0.42		0.36		(1, 1, 2)(1, 0, 0)
Acolman	RADSOL	1.23	−0.55				0.83		(2, 0, 0)(1, 0, 0)
Chapingo	RADSOL	0.62		0.52			0.89	−0.39	(1, 0, 1)(1, 0, 1)
ENCBII	RADSOL			1.02	0.50		0.80	−0.37	(0, 0, 2)(1, 0, 1)
Acolman	TEMP	1.38	−0.63				0.64		(2, 0, 0)(1, 0, 0)
Chapingo	TEMP	1.66	−0.91	−1.67	0.72		0.29		(2, 1, 2)(1, 0, 0)
ENCBII	TEMP	0.41		0.04	-0.35	-0.56	0.95	−0.75	(1, 1, 3)(1, 0, 1)

mediciones del error y del cuadrado medio del error (MSE). Cadenas y Rivera (2007) mencionan que si el valor observado en el tiempo t es y_t y Ft es la predicción para el mismo tiempo, entonces el error se define como $e_t=y_t-F_t$, y el cuadrado medio del error es:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

En nuestro estudio el valor observado del promedio de la temperatura del aire entre las 16:00 y 17:00 del 12 de marzo, 2013, fue 20.7 °C y el valor obtenido con la predicción del modelo ARIMA para el mismo tiempo fue 19.96 °C, por lo cual el MSE del modelo ARIMA, de ese tiempo fue de 0.547. El valor de la predicción persistente de ese mismo tiempo fue

measurements of the error and of the mean square of the error (MSE) were calculated. Cadenas and Rivera (2007) point out that is the value observed in the time t is y_t and Ft is the prediction for the same time, then the error is defined as $e_t = y_t - F_t$, and the mean square of the error is:

$$MSE = \frac{1}{n}\sum_{t=1}^{n} e_t^2$$

In our study the observed value of the average of air temperature between 16:00 and 17:00 of March 12, 2013, was 20.7 °C and the value obtained with the prediction of the ARIMA model for the same time was 19.96 °C, thus the MSE of the ARIMA model of that time was 0.547. The value of the persistent

14.95 (el valor observado del promedio de la temperatura del aire entre las 14:00 y 15:00 h del 12 de marzo de 2013); por lo tanto, el valor del MSE del modelo persistente, de ese tiempo fue de 33.06. La misma operación se realizó para las 40 h hacia adelante (20 tiempos hacia delante).

El valor MSE se calculó en tiempos de 5 en 5 hacia adelante para todas las variables (5, 10, 15, 20, etc.) hasta que el valor del MSE obtenido con la predicción de los modelos ARIMA (MSE_A) fuera mayor que el valor del MSE obtenido con la predicción del modelo de persistencia (MSE_B). En la variable temperatura del aire del periodo de marzo de la estación Acolman se encontró que hasta 15 tiempos hacia adelante el valor del MSE_A (2.531) fue menor que el del MSE_B (10.309); y a los 20 tiempos el valor del MSE_A (11.204) fue mayor que el del MSE_B (10.422). Para comparar los errores se calculó un porcentaje de mejoría de la predicción del modelo ARIMA con respecto a la predicción con el modelo persistente.

El porcentaje de mejoría de la predicción del modelo ARIMA con respecto a la predicción con el modelo persistente (PM) se calculó como sigue:

$$PM = 100 - \frac{MSE_A * 100}{MSE_P}$$

donde MSE_A es el MSE del modelo ARIMA, MSE_P es el MSE del modelo persistente.

En la variable temperatura del aire del periodo de marzo de la estación Acolman y 15 tiempos hacia adelante, se encontró que el valor del PM fue 75.4 %, lo cual indica que el modelo ARIMA se desempeña 75.4 % mejor que el modelo persistente hasta los 15 tiempos hacia adelante (30 h). De la misma manera se realizó el cálculo del PM para todas las variables meteorológicas en ambos periodos (marzo y junio) y para las tres EMAS (Cuadro 2).

Dos aspectos importantes en un esquema de predicción son: 1) que tan bien un modelo retiene su precisión sobre el horizonte de predicción y, 2) que tan robusto es el esquema para la elección del horizonte de predicción (Kavasseri *et al.*, 2009). Para observar el primer aspecto se realizaron predicciones de los valores de las variables meteorológicas hacia adelante hasta que el PM fuera menor que cero. Cuando

prediction of this same time was 14.95 (the observed value of the average of the air temperature between 14:00 and 15:00 h of March 12, 2013); therefore, the value of the MSE of the persistent model of this time was 33.06. The same operation was carried out for the 40 h ahead (20 times ahead).

The MSE value MSE was calculated from 5 in 5 times ahead for all of the variables (5, 10, 15, 20, etc) until the value of the MSE obtained with the prediction of the ARIMA models (MSE_A) was higher than the value of the MSE obtained with the prediction of the persistence model (MSE_B). In the air temperature variable of the period of March of the Acolman station, it was found that up to 15 times ahead the value of MSE_A (2.531) was lower than that of the MSE_B (10.309); and at 20 times the value of MS_A (11.204) was higher than that of MS_B (10.422). To compare the errors, a percentage of improvement of prediction was calculated of the ARIMA model with respect to the prediction with the persistent model.

The percentage of improvement of the prediction of the ARIMA model with respect to the prediction with the persistent model (PM) was calculated as follows:

$$PM = 100 - \frac{MSE_A * 100}{MSE_P}$$

where MSE_A is the MSE of the ARIMA model, MSE_p is the MSE of the persistent model.

In the variable of air temperature of the period of March at the Acolman station and 15 times ahead, it was found that the value of PM was 75.4 %, which indicates that the ARIMA model performs 75.4 % better than the persistent model as far as 15 times ahead (30 h). Thus, the calculation of the PM was made for all of the meteorological variables in both periods (March and June) and for the three EMAS (Table 2).

Two important aspects in a prediction plan are: 1) how well a model retains its precision over the prediction horizon, and 2) how robust is the plan for the selection of the prediction horizon (Kavasseri *et al.*, 2009). To observe the first aspect, predictions were made of the values of the meteorological variables ahead until the PM was less than zero. When the values of PM are less than zero, it indicates that

Cuadro 2. Porcentaje de mejoría de la predicción del modelo ARIMA con respecto a la predicción con el modelo persistente (PM) para las estaciones meteorológicas Acolman (A), Chapingo (B), y ENCBII (C), los tiempos hacia adelante (T) y en los periodos de marzo y junio.

Table 2. Percentage of improvement of the prediction of the ARIMA model with respect to the precision with the persistent model (PM) for meteorological stations Acolman (A), Chapingo (B), and ENCBII (C), the times ahead (T) and in the periods of March and June.

		PM (%)							
		Marzo			Junio			Promedio	
Variable	T	A	B	C	A	B	C	Marzo	Junio
ET0	20	51.4	55.7	----	−153.3	−42.6	−56.1	53.6	−84.0
HR	15	65.7	27.7	49.5	−408.8	−425.3	−135.3	46.7	−323.1
RADSOL	30	83.3	64.9	65.0	19.1	13.1	23.8	71.0	18.7
TEMP	15	75.4	33.0	21.7	−42.0	−140.8	4.9	43.4	−59.3

los valores de PM son menores que cero, indica que el MSE_P fue mejor que el MSE_A (la predicción con el modelo persistente fue mejor que con el modelo ARIMA). Para observar el segundo aspecto se eligieron dos periodos: el periodo de marzo con muy poca precipitación (menos de tres eventos de precipitación y menos de 1 mm en total), y el periodo de junio en donde se presentan más de 20 eventos de precipitación (más de 8 mm en total).

El modelo ARIMA predice mejor que el persistente más de 15 tiempos hacia adelante para las variables ET0, HR, RADSOL y TEMP en el periodo de marzo (Cuadro 2). Para el periodo de junio se encontró que el modelo persistente fue mejor que el ARIMA para las variables ET0, HR y TEMP. Una probable razón de esto es que la precipitación afecta las demás variables meteorológicas y puede cambiar el comportamiento de una serie temporal.

Conclusiones

El uso de software de computadora y modelos ARIMA permite al investigador estimar la predicción de variables meteorológicas automáticamente y en tiempo real. Sin embargo, los resultados indican que, en promedio, en el periodo de marzo con muy poca precipitación (menos 1 mm), la predicción con los modelos ARIMA fue mejor que la predicción con el modelo persistente en: 53.6 % en evapotranspiración de referencia hasta 20 tiempos hacia adelante (40 h); 46.7 % en humedad relativa hasta 15 tiempos hacia adelante (30 h); 71 % en radiación solar hasta 30 tiempos hacia adelante (60 h); 43.4% en temperatura

the MSE_P was better than the MSE_A (the prediction with the persistent model was better than with the ARIMA model). To observe the second aspect, two periods were selected: the period of March with very little precipitation (less than three rainfall events and less than 1 mm total) and the period of June in which more than 20 rainfall events occurred (more than 8 mm total).

The ARIMA model predicts better than the persistent model more than 15 times ahead for the variables ET0, RH, SOLRAD and TEMP in the period of March (Table 2). For the period of June it was found that the persistent model was better than the ARIMA model for the variables ET0, RH and TEMP. A possible reason for this is that rainfall affects the other meteorological variables and can change the behavior of a time series.

Conclusions

The use of computer software and ARIMA models allows the investigator to estimate the prediction of meteorological variables automatically and in real time. However, the results indicate that on the average, in the period of March with very little rainfall (less than 1 mm), prediction with the ARIMA models was better than prediction with the persistent model with: 53.6 % in reference evapotranspiration by as much as 20 times ahead (40 h); 46.7 % in relative humidity up to 15 times ahead (30 h); 71 % in solar radiation up to 30 times ahead (60 h); 43.4 % in air temperature as much as 15 times ahead (30 h). In the period of June, the predictions obtained with the

del aire hasta 15 tiempos hacia adelante (30 h). En el periodo de junio, las predicciones obtenidas con el modelo persistente fueron mejores que con el modelo ARIMA.

Literatura Citada

Allen, R. G. 2006. Evapotranspiración del cultivo: guías para la determinación de los requerimientos de agua de los cultivos. FAO 56: 89-173

Box, G. E. and G. M. Jenkins. 1976. Time Series Analysis: Forecasting and Control. Revised Ed., Holden-Day, San Francisco. pp: 469-471

Cadenas, E., and W. Rivera. 2007. Wind speed forecasting in the south coast of Oaxaca, Mexico. Renew. Energy 32: 2116-2128.

Canova, F. and B. E. Hansen. 1995. Are seasonal patterns constant over time? a test for seasonal stability. J. Bus. Econ. Stat. 13: 237-252.

Chattopadhyay, S., and G. Chattopadhyay. 2010. Univariate modelling of summer-monsoon rainfall time series: comparison between ARIMA and ARNN. C. R. Geoscience 342: 100-107.

Dalgaard, P. 2008. Introductory Statistics with R. Second Edition. Springer Science Business Media, LLC. New York, NY, USA. 364 p.

Hyndman, R. J. and Y. Khandakar. 2008. Automatic time series forecasting: The forecast package for R. J. Stat. Software 27: 1-22.

Hyndman R. J., G. Athanasopoulos, S. Razbash, D. Schmidt, Z. Zhou, Y. Khan and C. Bergmeir. 2013. Forecasting functions for time series and linear models. R package version 4.06.http://cran.r-project.org/web/packages/forecast. (Accesed: June 2013).

Karl, T. R., R. W. Knight, and B. Baker. 2000. The record breaking global temperatures of 1997 and 1998: Evidence for an increase in the rate of global warming?. Geophys. Res. Lett. 27: 719-722.

Kavasseri, R. G., and K. Seetharaman. 2009. Day-ahead wind speed forecasting using f-ARIMA models. Renew. Energy 34(5): 1388-1393.

persistent model were better than with the ARIMA model.

—End of the English version—

Kofler, M. 2005. The Definitive Guide to MySQL 5. David Kramer. Third edition. APRESS. New York, NY, USA. 172 p.

Korhonen, K., F. Donadini, P. Riisager, and L. J. Pesonen. 2008. GEOMAGIA50: an archeointensity database with PHP and MySQL. Geochem. Geophys. Geosyst. 9: 1-14.

Kwiatkowski, D., P. C. Phillips, P. Schmidt, and Y. Shin. 1992 Testing the null hypothesis of stationarity against the alternative of a unit root. J. Econom. 54: 159-178.

Narayanan, P., A. Basistha, S. Sarkar, and S. Kamna. 2013. Trend analysis and ARIMA modelling of pre-monsoon rainfall data for western India. C. R. Geoscience 345: 22-27.

Pankratz, A. 1983. Forecasting With Univariate Box-Jenkins Models Concepts and Cases. Ed., John Wiley & Sons. United States. pp. 3-19.

Pulido-Calvo, I., J. Roldán, R. López-Luque, and J. C. Gutiérrez-Estrada. 2002. Técnicas de predicción a corto plazo de la demanda de agua. Aplicación al uso agrícola. Ing. del Agua 9: 319-331.

Quayle, R. G., T. C. Peterson, A. N. Basist, and C. S. Godfrey. 1999. An operational near-real-time global temperature index. Geophys. Res. Lett. 26: 333-335.

Randolph, N., D. Gardner, M. Minutillo, and C. Anderson. 2010. Professional Visual Studio 2010. Wrox. Wiley Publishing, Inc, Indianápolis, Indiana. 1177 p.

Reikard, G. 2009. Predicting solar radiation at high resolutions: A comparison of time series forecasts. Sol. Energy 83: 342-349.

Vitart, F., A. W. Robertson, and D. L. Anderson. 2012. Proyecto de predicción subestacional a estacional: tendiendo un puente entre el tiempo y el clima. Boletín de la OMM 61: 23-28.

INTELIGENCIA DE NEGOCIOS: ESTADO DEL ARTE
BUSINESS INTELLIGENCE: STATE OF THE ART

RESUMEN

La Inteligencia de Negocios BI (Business Intelligence) es una herramienta bajo la cual diferentes tipos de organizaciones, pueden soportar la toma de decisiones basadas en información precisa y oportuna; garantizando la generación del conocimiento necesario que permita escoger la alternativa que sea más conveniente para el éxito de la empresa. La investigación comienza con la definición y aplicaciones de BI; además se muestran trabajos relevantes en algunas de las herramientas para hacer BI, como son Data Warehouse (Bodega de Datos), Olap (Cubos Procesamiento Analítico en Línea), Balance Scorecard (Cuadro de Mando) y Data Mining (Minería de Datos).

PALABRAS CLAVES: Bodega de Datos, Gobernabilidad, Minería de Datos.

ABSTRACT

Business Intelligence BI is a tool, below different kind organizations, supports decisions making processes, based in an exact and accurate information; guarantying the production of the needed knowledge that lets to choose the most appropiate option for the company success. The investigation begins with the BI definition and applications; by addition shows definitions and relevant BI investigations tools, like Data Warehouse, Olap, Balance Scorecard and Data Mining.

KEYWORDS: *Data Warehouse, Governance, Data Mining*

ALVEIRO ALONSO ROSADO GOMEZ
Ingeniero de Sistemas
Especialista en Gestión de Proyectos Informáticos
D.I 88.283.394
aarosadog@ufpso.edu.co
alveiro@hotmail.com
Tel: 097-5613560 / 3153190561
Docente Tiempo Completo
Universidad Francisco de Paula Santander Ocaña – Colombia
Grupo de Investigación en Teleinformática y Desarrollo de Software (GITYD).
D en Colciencias. U.F.P.S.O.

Dewar Willmer Rico Bautista
Ingeniero de Sistemas. MCC (c)
Docente Tiempo Completo. UFPSOcaña
dwricob@ufpso.edu.co,
ing_dewar@yahoo.com
Grupo de Investigación en Teleinformática y Desarrollo de Software (GITYD).
D en Colciencias. U.F.P.S.O.
Grupo Investigación Ciencias Computacionales (CICOM).
B en Colciencias. UniPamplona

1. INTRODUCCIÓN

En la actualidad la gran mayoría de las organizaciones cuenta con un sistema de información que soporta gran parte de las actividades diarias propias del sector de negocios en donde se esté desempeñando, este sistema puede ser sencillo o robusto todo depende de las exigencias del negocio, con el transcurso del tiempo estas aplicaciones llegan a tener la historia de la organización, los datos almacenados en las bases de datos, pueden ser utilizados para argumentar la decisión que se quiera tomar.

Un estudio realizado en Europa por *Information Builders Ibéric* mostró el costo que tiene la falta de sistemas de toma de decisiones en las organizaciones, según estos datos, el empleado europeo medio pierde una media de 67 minutos diariamente buscando información de la compañía, lo que equivale a un 15,9% de su jornada laboral. Para una organización de 1.000 empleados que gane unos 50.000 euros al día esto equivale a 7,95 millones de euros al año de salario perdido, todo ello por la búsqueda de información para tomar una decisión. (Zúmel 2008)

Fecha de Recepción: Enero 26 de 2010
Fecha de Aceptación: Marzo 25 de 2010

El poder competitivo que puede tener una empresa se basa en la calidad y cantidad de la información que sea capaz de usar en la toma de de decisiones; mediante la implementación de Inteligencia de Negocios se proporcionan las herramientas necesarias para aprovechar los datos almacenados en las bases de datos de los sistemas transaccionales para utilizar la información como respaldo a las decisiones, reduciendo el efecto negativo que puede traer consigo una mala determinación.

La investigación comienza con la definición de BI, sus aplicaciones; adicionalmente se muestran conceptos y trabajos relevantes en algunas de las herramientas para hacer BI, como son *Data Warehouse* (Bodega de Datos), Olap (Cubos Procesamiento Analítico en Línea), *Balance Scorecard* (Cuadro de Mando) y *Data Mining* (Minería de Datos).

2. DESARROLLO

2.1 INTELIGENCIA DE NEGOCIOS

La inteligencia de negocios se define como la habilidad corporativa para tomar decisiones. Esto se logra mediante el uso de metodologías, aplicaciones y tecnologías que permiten reunir, depurar, transformar datos, y aplicar en ellos técnicas analíticas de extracción de conocimiento (Parr 2000), los datos pueden ser estructurados para que indiquen las características de un área de interés (Stackowiak et al. 2007), generando el conocimiento sobre los problemas y oportunidades del negocio para que pueden ser corregidos y aprovechados respectivamente. (Ballard et al. 2006)

Implementar herramientas de BI dentro de la organización permite soportar las decisiones que se toman; al nivel interno ayuda en la gestión del personal (Sharma et al. 2009) y del lado externo produce ventajas sobre sus competidores (Maureen 2009). Existen ocasiones en las cuales no se pueden lograr todos los beneficios que tiene BI; debido al proceso que lleva consigo implementar un proyecto de estas características, se puede cometer errores en la definición del planteamiento de las necesidades de conocimiento de la empresa; el no determinar la magnitud de los problemas de información a solucionar generalmente repercute en el fracaso del proyecto.

En la actualidad se está planteando un concepto nuevo llamado *Agile BI Governance*, el cual propone, arquitecturas, métodos y herramientas necesarios para implantar una infraestructura para BI. Esta definición, combina conceptos de *IT Governance*, Manifiesto Ágil y *Data Governance*, para lograr un alcance que contemple las diferentes unidades de negocio, y soporte el proceso estratégico de obtención de valor del *Business Intelligence* en la empresa. Permite conocer cómo controlar un sistema de estas características, qué políticas debo aplicar, qué métodos de control tengo que poner en marcha y cómo debo gobernar los sistemas de BI. (Fernández 2008)

Agile BI Governance establece 4 valores básicos, pero dependiendo de cada organización puede incluir los que vayan en relación con su propia estrategia.

- Adaptabilidad Continúa. La incertidumbre y el cambio continuo son el estado natural de los sistemas de toma de decisiones, pero parece ser que muchas organizaciones aún no son conscientes de ellos. En este tipo de proyectos siempre se está cambiando el punto de vista analítico.
- Trabajo Conjunto. El usuario operativo del software ha de ser parte activa dentro de los grupos de IT que desarrollan los sistemas de BI.
- Jerarquías Flexibles. Los grupos de trabajo dentro del *Agile BI Governance* deberán estar estructurados con jerarquías flexibles que fomenten el intercambio de información.
- Personas Antes que Procesos. Priorizar la entrega de la información a las personas que controlan los procesos y no tanto en definir los procesos que han de controlar las personas. (Fernández 2008)

2.2 *DATA WAREHOUSE*

Es el proceso de extraer datos de distintas aplicaciones (internas y externas), para que una vez depurados y especialmente estructurados sean almacenados en un depósito de datos consolidado para el análisis del negocio. Requiere una combinación de metodologías, técnicas, hardware y los componentes de software que proporcionan en conjunto la infraestructura para soportar el proceso de información (Stackowiak et al. 2007). La estructura que se defina debe reflejar las necesidades y características del negocio, sus departamentos, equipos de trabajo y directivos[1] , esto permitirá responder a interrogantes generados al tratar de tomar las decisiones (Witten 2000) y con el tiempo se va convirtiendo en la memoria corporativa (Wang 2009); describiendo el pasado y el presente de la empresa. *Data Warehouse* desglosa, resume, ordena y compara, pero no descubre, ni predice. (Flores 2004)

Para la construcción de un *Data Warehouse* se establecen tres etapas; la primera está dedicada a examinar el esquema Entidad Relación de la base de datos operacional, generando los esquemas multidimensionales candidatos.

La segunda etapa, consiste en recoger los requisitos de usuario por medio de entrevistas, para obtener información acerca de las necesidades de análisis de estos, y la tercera etapa, contrasta la información obtenida en la segunda etapa, con los esquemas multidimensional candidatos formados en la primera etapa generando así, una solución que refleja los requisitos de usuario (Zenaido 2008).

Por otra parte implementar una solución de este tipo, ocasiona un costo que no todas las organizaciones están dispuestas a pagar (debido a sus capacidades de inversión), es por eso que los promotores del proyecto dentro de la empresa deben persuadir a los directivos y compañeros de trabajo, una buena alternativa de hacerlo es mediante el uso de técnicas administrativas, que permitan conocer a los directivos como se puede establecer el retorno de la inversión del proyecto equiparando inversión contra beneficios. (Arturo 2001)

[1] MicroStrategy. 2002. The 5 Styles of Business Intelligence: INDUSTRIAL-STRENGTH BUSINESS INTELLIGENCE, 2002. Disponible en http://www.innovacons.com/docs/5_styles.pdf

Al ser un depósito de datos consolidado para el análisis del negocio necesita tomar datos de distintas fuentes, Internas y Externas (Stackowiak et al. 2007), y como las características de las empresas son diferentes la cantidad de registros almacenados en algunas de ellas puede llegar a ser de proporciones exponenciales; es por esta razón que se necesita de procesos que optimicen los tiempos de extracción, transformación y transferencia de los datos del sistemas de información a la fuente de datos esto se logra implementando técnicas incrementales que mediante el uso de *Snapshots y Triggers*, se encarguen de sacar, transformar y transferir los registros que existen en el sistema de información a la fuente de datos. (Flores 2003)

El uso de *Data Warehouse* es tan amplio que llega a diferentes tipos de organizaciones y distintos temas de interés, puede ser implementado con conceptos Administrativos, en la administración; ayuda en la identificación de elementos de cambio que definan una nueva manera de hacer negocios, en donde la competencia debe estar orientada a trabajar no sólo de forma aislada, sino en colaboración con los diversos grupos de interés o actores de la industria, buscando referencias diferenciadoras para alcanzar el éxito (Romero 2002), en empresas petroquímicas; incrementa la exactitud y precisión en la toma de decisiones con un 93.9% en la rentabilidad (Silva 2009), en la Web; optimiza búsqueda Web de metadatos con características semi-inteligentes y también suministra el soporte necesario para crear comunidades de colaboración científica (Luna et al. 2008) (Ameur et al. 2006), en transformadores de potencia; almacenando, la monitorización del estado del flujo de energía (Mariño et al. 2004).

2.3 OLAP

El procesamiento analítico en línea permite obtener acceso a datos organizados y agregados de orígenes de datos empresariales[2], organiza subconjuntos de datos con una estructura multidimensional de manera que represente un significado especial o responda a una pregunta en particular [3] [4] . (Roussel 2006)

Estas herramientas soportan el análisis interactivo de la información de resumen, soportando muchas tareas de agrupación de datos que no pueden realizarse empleando las facilidades básicas de agregación y agrupamiento (Silberschatz et al. 2006)

2.3.1 TIPOS DE SISTEMAS OLAP

Tradicionalmente, este sistema se clasifica según las siguientes categorías:

- ROLAP. Implementación que almacena los datos en un motor relacional. Típicamente, los datos son detallados, evitando las agregaciones y las tablas se encuentran normalizadas.
- MOLAP. Esta implementación almacena los datos en una base de datos multidimensional. Para optimizar los tiempos de respuesta, el resumen de la información es usualmente calculado por adelantado.
- HOLAP (Hybrid OLAP). Almacena algunos datos en un motor relacional y otros en una base de datos multidimensional[5] .

Al igual que *Data Warehouse*, OLAP también es aplicable a un amplio rango de temas diferentes, uno de ellos es en Bases de Datos espaciales proporcionando características necesarias para los sistemas de tipo geográfico; como hechos, dimensiones, miembros, niveles, jerarquías, operaciones de navegación, operaciones de consolidación y comportamiento del clima (Abril 2007) (Bernier et al. 2009). También se utiliza el almacenamiento MOLAP y ROLAP, para generar índices que mejoran los tiempos de accesos a las consultas de manera que los tiempos de entrega de la información demore el menor tiempo posible (Tamayo 2006). Otra de las aplicaciones es en la educación al ser aplicado en ambientes de aprendizaje proporcionando las dimensiones y los indicadores necesarios para hacer la definición de un modelo de evaluación académica (Cockbaine 2004).

2.4 CUADRO DE MANDO INTEGRAL

El cuadro de mando integral (*Balanced Scorecard*) es una herramienta que permite alinear los objetivos de las diferentes áreas o unidades con la estrategia de la empresa y seguir su evolución[6] . El uso que se le puede dar a un Cuadro de Mando Integral es tan diverso que se puede contemplar autoevaluaciones del personal (Martínez 2008), hasta la definición de conceptos netamente organizacionales como son; la misión, la política de calidad; plan de comunicación, imagen corporativa, acciones de formación, catálogo de servicios; la confección de una cartera de clientes y la realización de acciones para conocer mejor sus opiniones y preferencias, así como para personalizar la presentación

[2] Microsoft Developer Network, Trabajar con procesamiento analítico en línea (OLAP), Julio 2006, disponible en http://msdn.microsoft.com/es-s/library/ms175367(SQL.90).aspx

[3] MicroStrategy, Análisis OLAP, disponible en http://www.microstrategy.com.ar/Solutions/5Styles/olap_analysis.asp

[4] Glosario.net, Tecnología OLAP, Octubre 2006, http://tecnologia.glosario.net/terminos-tecnicos-internet/tecnolog%EDa-olap-1579.html

[5] Wikipedia, OLAP, http://es.wikipedia.org/wiki/OLAP

[6] Ibermatica, Business Intelligence, 2006, disponible en http://www.ibermatica.com/ibermatica/publicaciones/BusinessIntelligence.pdf/download

de la oferta de servicios para los clientes más importantes (Villalbía et al. 2005) (Matilla 2007). En fin la ejecución de un cuadro de mando es tan amplia y generosa que puede llegar a cambiar la forma en que se presta un servicio en entidades públicas (Peters et al. 2007) (Weir et al. 2009).

2.5 DATA MINING

Es el proceso de Seleccionar, Explorar, Modificar, Modelizar[7] y valorar grandes cantidades de datos con el objetivo de descubrir conocimiento (Pérez 2006). El proceso debe ser automático o semi-automático. Los modelos hallados deben ser significativos demostrando cierto patrón o regla de comportamiento[8]. Las aplicaciones más utilizadas son las que necesitan algún tipo de predicción. Por ejemplo, cuando una persona solicita una tarjeta de crédito, la compañía emisora quiere predecir si la persona, clasifica con el perfil identificado de usuarios morosos (Silberschatz et al. 2006).

La minería de datos, permite la gestión en tiempo real de manera eficaz, es una herramienta aplicable a cualquier tipo de empresa. Una amplia gama de compañías puede tener aplicaciones exitosas con ella (Angeles et al. 2010).

Beneficios asociados a la minería de datos (López 2004): Incremento de los resultados como consecuencia del aumento de la cuota de mercado; Fidelización de la clientela dada una mejor respuesta a sus requerimientos; Mejora del rendimiento; Reducción del factor riesgo; Optimización de las estrategias y toma de decisiones y Optimización de la gestión, maximizando rentabilidades.

La aplicación de la Minería de Datos, además de permitir el descubrimiento del conocimiento en general, también se utiliza Biología soporta las investigaciones en la rama biológica, como herramienta insustituible para enfrentar la avalancha de datos que producen esta clase de proyectos (Febles 2002), en la Web Semántica; convierte la información en conocimiento que está distribuida en la web, proporcionando a las computadoras una mayor capacidad para gestionar y recuperar dichos datos (Rodríguez 2006), en las Redes de computadores; mediante la recolección de información acerca de los factores que impactan sobre la infraestructura de seguridad para descubrir la información relevante que ayude a tomar decisiones para corregir y/o mejorar las infraestructura de seguridad (Rojas 2005), en la Educación; haciendo seguimiento en los procesos de auto aprendizaje (Radenkovic et al. 2009).

La Minería de Datos permite hacer simulación del comportamiento humano; en procesos forenses; establece los diferentes escenarios de ocurrencia de accidentes (Parhizi et al. 2009); en procesos biométricos; con el reconocimiento de emociones faciales (Yang et al. 2009), en los tribunales ayuda a determinar la culpabilidad de los sospechosos de delito; basado en información histórica de otros delincuentes (Han-Wei 2009) (Chen et al. 2004). En la medicina; con la determinación de la dosificación ideal de medicamentos (Razali 2009), y en el tratamiento de trastornos del habla (Danubianu 2009), compartir información (Houston et al. 1999) y pre-decisión temprana del cáncer de mama (Bellaachia 2006); también con Minería de Datos se puede predecir la violencia domestica (Joelving 2009).

3. CONCLUSIONES

- BI, proporciona una manera rápida y efectiva de recopilar, abstraer, presentar, formatear y distribuir la información de sus fuentes de datos corporativos, permitiendo a los profesionales de la empresa, tanto dentro como fuera de la organización, visualizar y analizar datos precisos sobre las actividades fundamentales del negocio y utilizarlos para mejorar la toma de decisiones y la planificación estratégica. (Zúmel 2008)
- Una nueva forma de implementar BI dentro de las organizaciones, es la utilización de *BI Governance*, que combina las técnicas de BI, con el manifiesto Ágil, *IT Governance* y *Data Governance*, y este conjunto de teorías da como resultado el proceso de administración y seguimiento a la implantación de un proyecto de BI.
- En esta investigación se conocieron nuevas y diferentes formas de complementar el trabajo con *Data Warehouse*, optimizando los tiempos de transferencia de los datos del sistema transaccional a la bodega de datos, acompañar el proceso de montaje de un *Data Warehouse* y un estudio del retorno de la inversión.
- La principal enseñanza que se establece con este trabajo es la enorme gama de posibilidades que ofrece BI y sus herramientas, aquí se mostraron casos diferentes en los cuales se puede aplicar BI, en organización de diferentes sectores, con diferentes formas de trabajar, soportadas por sistemas de información particulares a cada una de ellas y con distintos contenidos en sus bases de datos. BI se establece como el siguiente paso a seguir para poner a las empresas en un nivel competitivo.

4. LISTA DE REFERENCIAS

[1] Abril D., Pérez J. 2007. Estado actual de las tecnologías de bodega de datos y OLAP aplicadas a bases de datos espaciales, Abril 2007. Disponible en http://www.scielo.org.co/pdf/iei/v27n1/v27n1a08.pdf

[7] Se trata de un neologismo que se usa con el significado de 'crear un modelo teórico' (de algo).

[8] MicroStrategy. 2002. *The 5 Styles of Business Intelligence: INDUSTRIAL-STRENGTH BUSINESS INTELLIGENCE*, 2002. Disponible en http://www.innovacons.com/docs/5_styles.pdf

[2] Ameur A., Yankovski V., Enroth S., Spjuth O. y Komorowski J.2006. Databases and ontologies The LCB Data Warehouse, 2006, disponible en http://bioinformatics.oxfordjournals.org/cgi/reprint/22/8/1024

[3]Angeles L., Maria y Santilan G., Angelica. Mineria de datos, concepto, caracteristicas, estructura y aplicaciones Disponible en: http://www.ejournal.unam.mx/rca/190/RCA19007.pdf

[4] Arturo L., Carmona C. 2001. GUÍA PARA OBTENER EL RETORNO A LA INVERSIÓN EN PROYECTOS DE DATA WAREHOUSE, disponible en http://copernico.mty.itesm.mx/phronesis/mty/tmp/ITESMMTY2002175.pdf

[5] Ballard, C., Abdel-Hamid A., Frankus R., Hasegawa F., Larrechart J., Leo P. y Ramos J. 2006. Improving Business Performance Insight with Business Intelligence and Business Process Management. Disponible en http://www.redbooks.ibm.com/redbooks/pdfs/sg247210.pdf

[6] Bellaachia A., Guven E.2006. Predicting Breast Cancer Survivability Using Data Mining Techniques, 2006. Disponible en http://www.siam.org/meetings/sdm06/workproceed/Scientific%20Datasets/bellaachia.pdf

[7] Bernier E., Gosselin P., Badard T., Bédard Y.2009. Easier surveillance of climate-related health vulnerabilities through a Web-based spatial OLAP application, 2009, disponible en http://www.ij-healthgeographics.com/content/8/1/18

[8] Cockbaine J., Casas I.2004. UN METAMODELO OLAP PARA LA EVALUACIÓN DE AMBIENTES TEL, Noviembre 2004, disponible en http://www.scielo.cl/pdf/rfacing/v13n1/art02.pdf

[9] Chen H., Chung W., JieXu J., Yi Qin G., Chau M.2004. Crime Data Mining: A General Framework and Some Examples, 2004. Disponible en http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.58.5929&rep=rep1&type=pdf

[10] Danubianu M., Socaciu Y.2009. Does Data Mining Techniques Optimize the Personalized Therapy of Speech Disorders?, 2009. Disponible en http://jacs.usv.ro/getpdf.php?issue=5&paperid=52

[11]Febles J., González A.2002. Aplicación de la minería de datos en la bioinformática, 2002. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352002000200003&lang=pt

[12] Fernández J., Mayol E. y Pastor J. 2008. Agile Business Intelligence Governance: Su justificación y presentación. Disponible en http://www.uc3m.es/portal/page/portal/congresos_jornadas/congreso_itsmf/Agile%20Business%20Intelligence%20Governance.pdf

[13] Fernández J. 2008. Los 4 Valores del Agile BI Governance. Disponible en http://sistemasdecisionales.blogspot.com/2008/01/los-4-valores-del-agile-bi-governance.html

[14] Flores A. 2003. Uso de triggers y snapshots como técnica incremental en el proceso de extracción, transformación y transferencia de datos (ETT) en un Data Warehouse, Junio 2003. Disponible en http://copernico.mty.itesm.mx/phronesis/mty/tmp/ITESMMTY2003356.pdf

[15] Flores R. 2004. Aplicación de Minería de Datos en un Ambiente Universitario, Diciembre 2004. Disponible http://copernico.mty.itesm.mx/phronesis/mty/tmp/ITESMMTY2005515.pdf

[16] Han-Wei L.2009. An Entity Sui Generis in the WTO: Taiwan's WTO Membership and Its Trade Law Regime, 2009. Disponible en http://www.jiclt.com/index.php/jiclt/article/view/90/89

[17] Houston A., Chen H., Hubbard S., Schatz B., Ng T., Sewell R.,Tolle K.1999. Medical Data Mining on the Internet: Research on a Cancer Information System, 1999. Disponible en http://www.icadl.org/intranet/papers/Medical-99.pdf

[18] Joelving F.2009. Data-Mining Medical Records Could Predict Domestic Violence, 2009. Disponible en http://www.wired.com/wiredscience/2009/09/domestic-abuse-prediction/

[19] López R., Daniel. Del conocimiento tácito al dato explícito. 2004. Disponible en http://www.redcientifica.com/doc/doc200405180600.html

[20] Luna E., Ambriz H., Nungaray J., Álvarez F., Mondragón J. 2008. DISEÑO DE UN MODELO SEMI-INTELIGENTE DE BÚSQUEDA DE METADATOS EN LA WEB, APLICADO A SISTEMAS DATA WAREHOUSING, Noviembre 2008. Disponible en http://www.scielo.cl/pdf/ingeniare/v16n3/art04.pdf

[21] Matilla M., Chalmeta R. 2007. Metodología para la Implantación de un Sistema de Medición del Rendimiento Empresarial, 2007. Disponible en http://www.scielo.cl/pdf/infotec/v18n1/art16.pdf

[22]Mariño P., Poza F., Ubeira M., Machado F. 2004. Sistema de Adquisición y Almacenamiento de Datos para Monitorización del Estado de Transformadores de Potencia, 2004. Disponible en http://www.scielo.cl/scielo.php?pid=S0718-07642004000200017&script=sci_arttext

[23] Martínez A., Martínez V.2008. Modelo de evaluación y diagnosis de excelencia en la gestión, basado en el cuadro de mando integral y el modelo EFQM de excelencia. Aplicación a las cajas rurales, 2008. Disponible en http://dspace.upv.es/xmlui/bitstream/handle/10251/3791/tesisUPV2909.pdf

[24] Maureen L., Fernández V. 2009. La gestión del valor de la cartera de clientes y su efecto en el valor global de la empresa: diseño de un modelo explicativo como una herramienta para la toma de decisiones estratégicas de marketing. Disponible en http://eprints.ucm.es/8064/1/T29976.pdf

[25] Parhizi Sh., Shahrabi J., Pariazar M.2009, A New Accident Investigation Approach Based on Data Mining Techniques, 2009. Disponible en http://www.scialert.net/pdfs/jas/2009/731-737.pdf

[26] Parr, O. 2000. Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship Management. Disponible en http://books.google.com.co/books?id=L3w0loZrcU0C&printsec=frontcover&dq=Data+Mining+Cookbook#v=onepage&q=&f=false
[27] Pérez C., Santin D.2006. Data Mining Soluciones con Enterprise Miner, Alfaomega Ra - Ma, 2006.
[28] Peters D., Ahmed Noor D., Singh L., Kakar F., Hansen P., BurnhamI G.2007. Un cuadro de mando para los servicios de salud del Afganistán, Febrero 2007. Disponible en http://www.scielosp.org/pdf/bwho/v85n2/v85n2a13.pdf
[29] Radenkovic B., Despotovic M., Bogdanovic Z., Barac D.2009. Creating adaptive environment for e-learning courses, 2009. Disponible en http://jios.foi.hr/index.php/jios/article/view/107
[30] Razali M., Ali S.2009. Generating Treatment Plan in Medicine: A Data Mining Approach, 2009. Disponible en http://www.scipub.org/fulltext/ajas/ajas62345-351.pdf
[31] Rodríguez K., Ronda R. 2006. El web como sistema de información, Febrero 2006. Disponible en http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352006000100008&lang=pt
[32] Rojas L., Carlos J. 2005. Uso de la Minería de Datos con Fines Predictorios de la Infraestructura de Seguridad de Redes, 2005. Disponible en http://copernico.mty.itesm.mx/phronesis/mty/tmp/ITESMMTY2005533.pdf
[33] Romero A. 2002. Modelo para el diseño de estrategias competitivas haciendo uso de una herramienta de Data Warehouse, Abril 2002. Disponible en http://copernico.mty.itesm.mx/phronesis/mty/tmp/ITESMMTY2002215.pdf
[34] Roussel G. 2006. Decision support systems serving the company: the secrets to a successful project, 2006. Disponible en http://www.symtrax.com/en/WhitePaper/EN_WhitePaper_SolutionsBI_SQ.pdf
[35] Sharma S., Sharma J. y Devi A. 2009. Corporate Social Responsibility: The Key Role of Human Resource Management, Disponible en http://www.saycocorporativo.com/saycoUK/BIJ/journal/Vol2No1/article9.pdf
[36] Silberschatz A., Korth H., Sudarshan S.2006. Fundamentos de Base de Datos, McGraw-Hill, Madrid, España, 2006
[37] Silva P., Silva R. 2009. Asimilación del Almacén de Datos en las Organizaciones Corporativas Petroquímicas, 2009. Disponible en http://www.scielo.cl/pdf/infotec/v20n2/art06.pdf
[38] Stackowiak, R. Rayman J. Greenewald R. 2007. Oracle Data Warehousing and Business Intelligence Solutions. Disponible en http://books.google.com.co/books?id=Gxy6_drRWRgC&dq=%22Oracle+Data+Warehousing+and+Business+Intelligence+Solutions%22&printsec=frontcover&source=bn&hl=es&ei=W0uJSqmqGsqltgewwtjnDA&sa=X&oi=book_result&ct=result&resnum=4#v=onepage&q=&f=false
[39] Tamayo M., Moreno F.2006. Análisis del modelo de almacenamiento MOLAP frente al modelo de almacenamiento ROLAP, Diciembre 2006, http://www.scielo.org.co/pdf/iei/v26n3/v26n3a16.pdf
[40] Villalbía J., Guixa J., Casasa C., Borrella C., Durana J., Artazcoza L., Camprubía E., Cusía M., Rodríguez-Montuquína P., Armengola J., Jiménez G.2005. El Cuadro de Mando Integral como instrumento de dirección en una organización de salud pública, 2005. Disponible en http://scielo.isciii.es/pdf/gs/v21n1/originales_breves2.pdf
[41] Wang J. 2009. Encyclopedia of Data Warehousing and Mining. Disponible en http://books.google.com.co/books?id=CJqnVVejkP8C&pg=PA1468&dq=Encyclopedia+of+Data+Warehousing+and+Mining#v=onepage&q=&f=false
[42] Weir E., d'Entremont N., Stalker S., Kurji K., Robinson V.2009. Applying the balanced scorecard to local public health performance measurement: deliberations and decisions, 2009. Disponible en http://www.biomedcentral.com/1471-2458/9/127
[43] Witten I., Frank E. 2000. Data mining parctical machine learning tools and techniques with java implementations, Academic Press, San Francisco, United States of America, 2000, p 371
[44] Yang Y., Wang G., Kong H.2009. Self-Learning Facial Emotional Feature Selection Based on Rough Set Theory, 2009. Disponible en http://downloads2.hindawi.com/journals/mpe/volume-2009/802932.pdf
[45] Zenaido L., Sánchez Z. 2008. Metodología para el diseño conceptual de almacenes de datos. Disponible en http://dspace.upv.es/xmlui/handle/10251/2506
[46] Zúmel, P.2008. Gestión del rendimiento, Noviembre 2008. Disponible en http://www.gestiondelrendimiento.com/Articulos/010/gdr010.pdf

URL

[47] Microsoft Developer Network, Trabajar con procesamiento analítico en línea (OLAP), Julio 2006. Disponible en http://msdn.microsoft.com/es-es/library/ms175367(SQL.90).aspx
[48] MicroStrategy, Análisis OLAP. Disponible en http://www.microstrategy.com.ar/Solutions/5Styles/olap_analysis.asp
[49] MicroStrategy. 2002. The 5 Styles of Business Intelligence: INDUSTRIAL-STRENGTH BUSINESS INTELLIGENCE, 2002. Disponible en http://www.innovacons.com/docs/5_styles.pdf
[50] Glosario.net, Tecnología OLAP, Octubre 2006, http://tecnologia.glosario.net/terminos-tecnicos-internet/tecnolog%EDa-olap-1579.html
[51] Ibermatica, Business Intelligence, 2006. Disponible http://www.ibermatica.com/ibermatica/publicaciones/BusinessIntelligence.pdf/download
[52] Wikipedia, OLAP. Disponible en http://es.wikipedia.org/wiki/OLAP