MAESTRÍA EN GERENCIA DE SISTEMAS Y TECNOLOGÍAS DE INFORMACIÓN

TÍTULO DEL TRABAJO

DATA MINING Y ANÁLISIS DE DATOS DEL PROCESO DE ADMISIÓN A LA EDUCACIÓN SUPERIOR EN ECUADOR.

Trabajo de Titulación presentado en conformidad a los requisitos establecidos para optar por el título de Magister en Gerencia de Sistemas y Tecnologías de Información.

Profesor guía
Ing. Jaime Vinueza Trujillo, MBA

Autor
Ing. Eddy Armas

Año
2014

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 escogido y dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación"

Ing. Jaime Vinueza, MBA

CI: 1716028509

DECLARACIÓN DEL ESTUDIANTE

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

Ing. Eddy Armas

CI: 1711715803

DEDICATORIA

Dedico este trabajo a mi familia, a mi hijo Juan Francisco por ser la fortaleza constante de mi alma y mi razón de vivir, a mi esposa Karol por su paciencia y comprensión durante todo este ciclo de estudios. A mis padres por haberme entregado lo mejor de ellos para mi superación personal y formarme como un hombre de bien.

RESUMEN

El Sistema Nacional de Nivelación y Admisión (SNNA) tiene como objetivo "garantizar la igualdad de oportunidades, la meritocracia, transparencia y acceso a la educación superior"[1]. El proceso de admisión de los aspirantes a las Instituciones de Educación Superior (IES) públicas se compone de varios subprocesos que se ejecutan en la siguiente secuencia:

- Inscripción a través del portal web de la institución.
- Aplicación del Examen Nacional para la Educación Superior (ENES).
- Postulación de los aspirantes a sus carreras de preferencia.
- Asignación de cupos en función del puntaje obtenido y los cupos ofertados por las IES.

Este proceso se encuentra informatizado a través de una aplicación web que permite la recolección inicial de datos de los aspirantes y además se utilizan aplicaciones informáticas desarrolladas internamente para los subprocesos que se ejecutan en cada aplicación del ENES. Por lo tanto se explorarán los datos del proceso de admisión en la búsqueda de patrones que permitan generar conocimiento para la toma de decisiones.

Existen varios marcos de referencia para el desarrollo de proyectos de minería de datos, sin embargo, para la ejecución del presente proyecto se utilizó el estándar de facto CRISP-DM (Cross Industry Standard Process for Data Mining), cuyo enfoque se basa en un modelo de desarrollo rápido.

Este marco de referencia inicia con el entendimiento del negocio o actividad de la organización que es considerada la fase más importante de la metodología y que comprende varias actividades orientadas a la comprensión de los objetivos de la organización desde el punto de vista institucional y a la familiarización con el conocimiento que la organización desea obtener.

[1] Objetivos del Sistema Nacional de Nivelación y Admisión (www.snna.gob.ec)

La siguiente fase del marco de referencia es la comprensión de los datos y consiste principalmente en la recolección de los datos que se desea analizar, la descripción e identificación de la calidad de los mismos. Luego, la etapa de preparación de datos es la que mayor cantidad de tiempo y esfuerzo requiere y consiste en la preparación de los datos para las etapas de modelamiento, evaluación de resultados y su posterior despliegue.

ABSTRACT

The main objective of the "Sistema Nacional de Nivelación y Admisión" (SNNA) is to ensure equal opportunities, meritocracy, transparency and access to higher education". The admission process of applicants to public higher education Institutions (IES) is composed of multiple sub process that runs in the following sequence:

- Registration through SNNA web site
- Application of the national exam (ENES).
- Nomination of candidates to their careers of choice.
- Allocation of quotas based on the score obtained and quotas offered by IES.

The admission process is implemented through a web application that allows the initial collection of applicant´s data, and also use computer applications developed internally for each sub process running on each application of the ENES. Therefore the data of the admissions process will be explored in search for patterns to generate knowledge for decision making.

There are several methodologies for the development of data mining projects, however, CRISP-DM standard (Cross Industry Standard Process for Data Mining) was used for the development of this project, whose approach is based on a rapid development model.

This methodology begins with business understanding that is considered the most important phase of the methodology and includes several activities aimed to understand organization objectives from an institutional point of view and adjustment with the knowledge that the organization wants to obtain.

The next phase of the methodology is data understanding which consists of data recollection for analyze, describe and identify quality of them. At this time,

data preparation stage require large amount of time and effort to prepare data for steps of modeling, assessment of results and their subsequent deployment.

ÍNDICE

Capítulo 1 Fundamento teórico 1

1.1. Antecedentes 1

1.2. Introducción 1

1.3. Técnicas de Análisis y Minería de Datos 3

1.4. Metodologías de proyectos de minería de datos 16

1.4.1. Cross-Industry Standard Process for Data Mining (CRISP-DM)... 16

1.4.2. Knowledge Discovery in Databases (KDD) 17

1.4.3. Sample, Explore, Modify, Model, Assess (SEMMA) 19

1.4.4. Comparación de las metodologías presentadas 19

1.4.5. Definición de la metodología a utilizar en el proyecto 21

Capítulo 2 Comprensión del negocio 22

2.1 Comprensión del negocio 23

2.1.1 Determinación de los objetivos de negocio 24

2.1.1.1 Objetivo General 25

2.1.1.2 Objetivos específicos 25

2.2 Valoración de la situación actual 26

2.3 Determinación de los objetivos del proyecto de minería 27

2.4 Elaboración del plan del proyecto 27

2.4.1. Cronograma 29

Capítulo 3 Comprensión de los datos 30

3.1 Recolección de datos iniciales 31

3.2 Atributos relevantes 35

3.3 Descripción de datos iniciales 37

3.3.1 Cantidad de datos 37

3.4 Verificación de la calidad de los datos 37
3.4.1 Perfilamiento de datos y manejo de excepciones 38
Capítulo 4 Preparación de los datos 46
4.1 Selección de los datos 47
4.1.1 Selección de registros 47
4.1.2 Selección de atributos 47
4.1.3 Inclusión / Exclusión de datos 48
4.2 Limpieza de los datos 49
4.2.1 Reporte de limpieza de los datos 49
4.3 Construcción de datos 51
4.3.1 Derivación de atributos 51
4.3.2 Generación de registros 54
4.3.3 Resumen del proceso 54
Capítulo 5 Modelamiento de datos 55
5.1. Selección de la técnica de modelado 56
5.1.1. Descripción de las técnicas seleccionadas 56
5.2. Generación del plan de prueba 57
5.3. Construcción de los modelos de minería de datos 57
5.3.1. Modelo en base a árboles de decisión 58
5.3.2. Modelo en base a Naive Bayes 59
5.3.3. Modelo en base a Clúster 60
5.3.4. Modelo en base a red neuronal 61
5.3.5. Resumen de modelos construidos 62
Capítulo 6 Evaluación de resultados 63
6.1. Valoración de resultados 64
6.1.1. Resultados del modelo de árboles de decisión 64
6.1.2. Resultados del modelo de Naive bayes 67

6.1.3. Resultados del modelo de Clustering 69
6.1.4. Resultados del modelo de red neuronal 70
6.1.5. Comparación de los modelos construidos 71
6.1.6. Aplicación del modelo seleccionado 72
6.2. Revisión del proceso 81
Capítulo 7 Despliegue de resultados 82
7.1. Plan de despliegue 82
7.2. Plan de monitoreo y mantenimiento 83
Conclusiones y recomendaciones 84
Referencias 86
Anexos 89

ÍNDICE DE FIGURAS

Figura 1: Técnicas de minería de datos 4
Figura 2: Conjunto de instancias 5
Figura 3: Algoritmos por técnica 6
Figura 4: Clustering 8
Figura 5: Árbol de decisión 10
Figura 6: Modelo Naive bayes 11
Figura 7: Red neuronal 12
Figura 8: Fases de la metodología CRISP-DM 17
Figura 9: Fases del modelo KDD 18
Figura 10: Fases del proceso SEMMA 19
Figura 11: Comprensión del negocio 22
Figura 12: Objetivos del SNNA 25
Figura 13: Cronograma del proyecto 29
Figura 14: Comprensión de los datos 30
Figura 15: Universo de aspirantes por subproceso 31
Figura 16: Atributo "género" 39
Figura 17: Atributo "edad" 39
Figura 18: Atributo "región" 40
Figura 19: Atributo "provincia" 41
Figura 20: Atributo "sector" 42
Figura 21: Atributo "estado_civil" 43
Figura 22: Atributo "discapacidad" 43
Figura 23: Atributo "unidad_educativa" 44
Figura 24: Atributo "tipo_unidad_educativa" 45
Figura 25: Resumen perfilamiento de datos 45
Figura 26: Preparación de los datos 46
Figura 27: Repositorio de datos 48
Figura 28: Selección de atributos 48
Figura 29: Filtrado de datos 50
Figura 30: Renombre de atributos 51

Figura 31: Modelado de datos 55
Figura 32: Parámetros del algoritmo árboles de decisión 59
Figura 33: Parámetros del algoritmo Naive Bayes 60
Figura 34: Parámetros del algoritmo de clustering 61
Figura 35: Parámetros del algoritmo red neuronal 61
Figura 36: Resumen de modelos generados – nivel socioeconómico 62
Figura 37: Evaluación de resultados 63
Figura 38: Divisiones del nodo raíz del árbol 64
Figura 39: Árboles de decisión – atributos socioeconómicos (notas => 848) 65
Figura 40: Árboles de decisión – atributos socioeconómicos (notas<= 681) 66
Figura 41: Naive Bayes – atributos socioeconómicos 67
Figura 42: Naive Bayes - Distinción de rangos 68
Figura 43: Clústeres 69
Figura 44: Detalle de Clústeres 70
Figura 45: Resultado del algoritmo red neuronal 71
Figura 46: Precisión de los modelos de minería de datos 72
Figura 47: Probabilidad de predicción 72
Figura 48: Modelo para variable TIPO_UED 73
Figura 49: Red de dependencias para NOTA_EVAL 74
Figura 50: Red de dependencias para NOTA_VERBAL 75
Figura 51: Red de dependencias para NOTA_ABSTRACTO 75
Figura 52: Red de dependencias para NOTA_LOGICO MATEMATICO 76
Figura 53: Naive Bayes – atributos UED 77
Figura 54: Resumen de resultados – NOTA_EVAL 78
Figura 55: Resumen de resultados – NOTA_ABSTRACTO 79
Figura 56: Resumen de resultados – NOTA_LOGICO_MATEMATICO 80
Figura 57: Resumen de resultados – NOTA_VERBAL 80
Figura 58: Despliegue de resultados 82

ÍNDICE DE TABLAS

Tabla 1: Comparación de varios algoritmos 14
Tabla 2: Comparación de metodologías para proyectos de minería de datos . 20
Tabla 3: Datos necesarios para la inscripción 33
Tabla 4: Atributos demográficos 36
Tabla 5: Atributos socio-económicos 36
Tabla 6: Grupos socio económicos 52
Tabla 7: Estructura de datos para atributos socioeconómicos 58
Tabla 8: Estructura de datos para atributos de unidades educativas 58
Tabla 9: Resumen de datos - UED 73
Tabla 10: Rangos de valores para NOTA_EVAL 76

Capítulo 1 Fundamento teórico

1.1. Antecedentes

Desde la masificación de las computadoras y el Internet a nivel mundial el volumen de la información ha tenido un crecimiento exponencial. En nuestra vida diaria cada uno de nosotros generamos gran cantidad de datos que están siendo analizados por las organizaciones con las que interactuamos, por ejemplo la empresa eléctrica, de agua, de teléfono, los supermercados, etc. Generalmente las organizaciones almacenan todos estos datos recolectados en sus sistemas de bases de datos, que luego serán analizados para obtener información y generar conocimiento útil para las mismas organizaciones.

De todo este proceso de recolección de datos cada vez más surgen varias necesidades para las propias organizaciones. Una de ellas es la necesidad cada vez más evidente de procesar grandes volúmenes de información en cortos períodos de tiempo.

En el caso de las instituciones públicas no es la excepción y esta información debería ser utilizada para obtener conocimiento que permita la toma de decisiones, pero sobre todo que permita la generación de una política pública adecuada para el beneficio de la sociedad en general.

1.2. Introducción

La minería de datos se puede definir como el proceso de descubrir conocimiento a partir de la identificación de patrones en grandes volúmenes de datos. En los últimos años este campo de la computación ha sufrido varios cambios debido principalmente al aumento de datos no estructurados que se generan desde las redes sociales, blogs y sitios de Internet. Además, la capacidad de procesamiento de grandes volúmenes de datos se ha vuelto cada

vez más crítica y por ello han surgido nuevas tecnologías de procesamiento de información como por ejemplo el procesamiento en memoria.

Para descubrir este conocimiento es necesario utilizar diferentes técnicas de los campos del aprendizaje automático y la estadística. El aprendizaje automático es un proceso de inducción de conocimiento y hoy en día tiene una gran variedad de aplicaciones como por ejemplo la detección de fraude en el uso de tarjetas de crédito, los diagnósticos médicos, el análisis de ADN, la robótica y muchas aplicaciones más.

Las técnicas de minería de datos contemplan el uso de algoritmos de aprendizaje automático que entre los más principales destacan los árboles de decisión, las redes bayesianas y los algoritmos de clusterización.

Los árboles de decisión realizan predicciones basándose en las relaciones entre las columnas del conjunto de datos y modelan las relaciones como series de divisiones en forma de árbol en valores específicos. Las redes bayesianas calculan la probabilidad de la relación entre todas las columnas de entrada y de predicción. Los algoritmos de clusterización permiten la agrupación de objetos homogéneos entre si y heterogéneos en relación con otros grupos.

Para llevar a cabo este proyecto se revisará los modelos de procesos y metodologías más adecuadas para el desarrollo de proyectos de minería de datos. Tres de las principales son: CRISP-DM, KDD y SEMMA.

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) es actualmente la guía de referencia más utilizada para el desarrollo de proyectos de minería de datos y se compone de 6 fases que serán descritas en la sección 1.4.1.

La metodología KDD (Knowledge Discovery in Databases) de acuerdo a (Fayyad, Piatesky-Shapiro, & Smyth, 1996), es el proceso mediante el cual se

descubre conocimiento mediante la identificación de patrones válidos de información dentro de un gran volumen de datos Este modelo de proceso se compone de 9 fases.

El modelo de procesos SEMMA (Sample, Explore, Modify, Model, Assess) que tiene su origen en el fabricante SAS y se compone de 5 etapas definidas las cuales son descritas en la sección 1.4.3.

1.3. Técnicas de Análisis y Minería de Datos

La minería de datos combina una gran cantidad de datos almacenados en uno o varios repositorios, las habilidades analíticas de los analistas de información y su entendimiento del negocio para tratar de descubrir patrones de información que serán la base de modelos que producirán nuevo conocimiento.

Según (Gartner, Inc.) la minería de datos es "el proceso de descubrimiento de nuevas relaciones, patrones y tendencias dentro de una gran cantidad de datos, y el uso de tecnologías de reconocimiento, técnicas estadísticas y matemáticas".

Según (Pérez López, 2007), la clasificación inicial de las técnicas de minería de datos se distinguen entre técnicas no supervisadas (descriptivas) y técnicas supervisadas (predictivas). La figura 1 muestra esta clasificación.

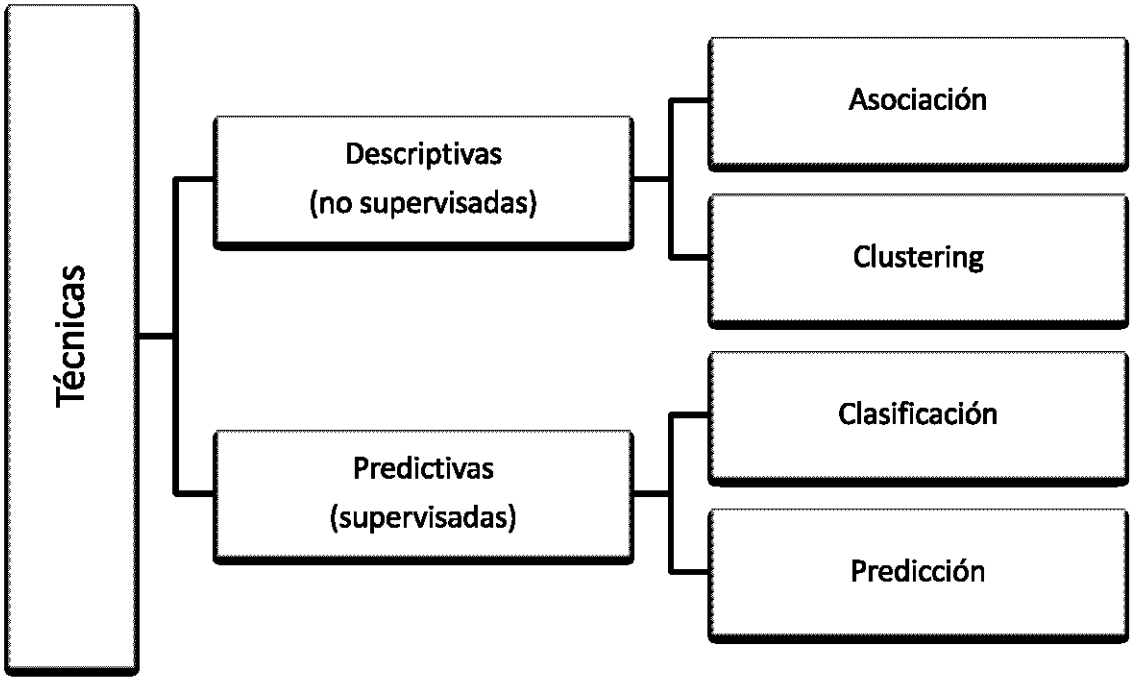


Figura 1: Técnicas de minería de datos

Las técnicas **descriptivas o no supervisadas** son utilizadas, por lo general, en situaciones en las que los patrones de datos no son conocidos, con el fin de entender y clasificar los objetos de estudio antes de aplicar otras teorías. El proceso de modelado de datos se lleva a cabo en un conjunto de ejemplos de los que no se tiene información sobre las categorías existentes en esos ejemplos. El modelo debe ser capaz de detectar los patrones de datos para construir estas categorías.

Las técnicas **predictivas o supervisadas** son utilizadas cuando se tiene un conocimiento previo del contenido en los datos. El objetivo de las técnicas supervisadas es el de crear un modelo de datos capaz de predecir el valor correspondiente a cualquier objeto de entrada válido después de haber visto una serie de ejemplos con los datos de entrenamiento. Este tipo de aprendizaje puede llegar a ser muy útil en problemas de investigación biológica y computacional.

Asociación:

La técnica de asociación detecta automáticamente las reglas que relacionan 2 o más atributos observando si la frecuencia de aparición de los valores determinados para los atributos seleccionados es relativamente alta. Estos modelos se usan especialmente para realizar recomendaciones.

Un ejemplo de esta técnica es cuando se desea identificar si los clientes de un supermercado compran crema de leche cada vez que compran frutas, así la próxima vez se puede sugerir a los clientes que compran frutas una promoción con algún producto de crema de leche.

Clustering (segmentación / agrupamiento):

Partiendo de un conjunto de instancias, esta técnica permite la agrupación de objetos homogéneos entre sí y heterogéneos en relación con otros grupos. Algunas aplicaciones de esta técnica pueden ser la segmentación de estudiantes que se gradúan de la educación general básica de acuerdo al conjunto de unidades educativas por región, etc.

	A_1	...	A_i	...	A_n
I_1	x_1^1	...	x_i^1	...	x_n^1
...	...	...	...	...	...
I_j	x_1^j	...	x_i^j	...	x_n^j
...	...	...	...	...	...
I_N	x_1^N	...	x_i^N	...	x_n^N

Figura 2: Conjunto de instancias

Clasificación:

La clasificación es una técnica supervisada que permite encontrar propiedades comunes entre un conjunto de datos y encasillarlos en diferentes clases. El objetivo de estas técnicas es desarrollar una descripción para cada clase utilizando las características disponibles en los datos. Luego, estas descripciones son utilizadas para clasificar nuevos datos.

Predicción:

El objetivo de esta técnica es predecir los valores de una variable continua a partir del cambio o evolución de otra variable continua que generalmente puede ser el tiempo, por ejemplo, se puede predecir el número de clientes a partir de los resultados de varios meses o años anteriores.

A continuación se presentan los principales algoritmos relacionados con las diferentes técnicas de aprendizaje automático descritas anteriormente:

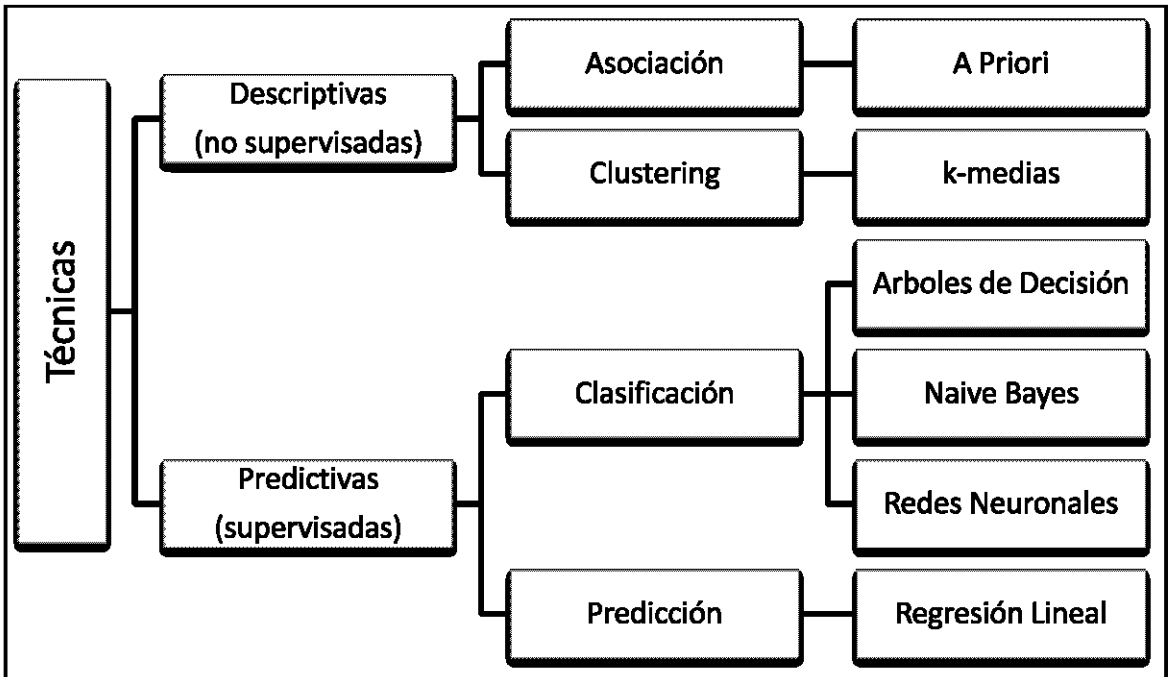


Figura 3: Algoritmos por técnica

Algoritmo A-Priori

Se utiliza para la generación de reglas de asociación sobre un conjunto de datos. Se basa en el conocimiento previo de conjuntos frecuentes, para reducir el espacio de búsqueda y aumentar la eficiencia.

Dado un conjunto de datos, el algoritmo intenta encontrar subconjuntos que tienen en común por lo menos un número mínimo de instancias. El algoritmo a-priori tiene un enfoque de abajo hacia arriba (bottom up) en donde subconjuntos frecuentes se extienden un elemento a la vez y grupos de candidatos son examinados contra los datos. El algoritmo termina cuando ya no se encuentran más extensiones.

Se parte de un conjunto de instancias T= $\{i_1, i_2, \ldots, i_m\}$

El primer criterio de selección de reglas del algoritmo "A priori" es la precisión o confianza, dada por el porcentaje de veces que instancias que cumplen el antecedente cumplen el consecuente, pero el segundo es el soporte, dado por el número de instancias sobre las que es aplicable la regla.

Algoritmo k-medias

Este algoritmo requiere que se especifique el número de clusters a obtener (k). Entonces, de manera randómica se seleccionan k puntos como los centros de los clusters que inicialmente no tienen ningún miembro. Todas las instancias son asignadas al clúster con centro más cercano y cuando todas las instancias han sido asignadas, se tendrán k clusters basados en los k centros originales pero estos centros ya no serán los verdaderos centros. Luego, el centro de las instancias en cada nodo es recalculado. Estos centros son tomados como los

nuevos puntos centrales de sus respectivos clusters. Esta iteración continúa hasta que los puntos centrales de los clusters se han estabilizado.

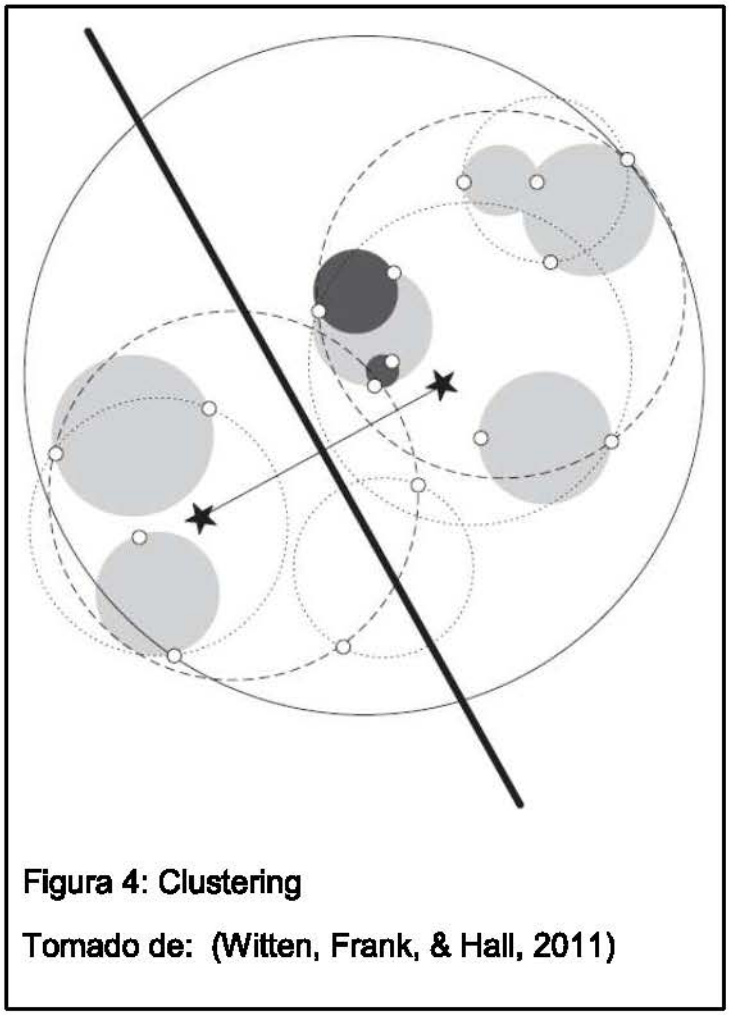


Figura 4: Clustering

Tomado de: (Witten, Frank, & Hall, 2011)

Arboles de decisión

Por lo general este algoritmo representa el modelo de datos construido como un árbol invertido con la raíz en la parte superior del árbol y sus ramas hacia abajo. Comparado con otras técnicas, esta puede resultar más fácil de interpretar sus resultados. La meta del algoritmo es crear un modelo de clasificación para predecir el valor del atributo destino (también llamado "etiqueta") en base a un conjunto de atributos de entrada. Cada nodo hoja del

árbol representa un valor del atributo destino de acuerdo a los valores de los atributos de entrada, representados por la ruta desde la raíz hasta la hoja.

Los árboles de decisión son generados por particionamiento recursivo. El particionamiento recursivo significa dividir en varias ocasiones los valores de los atributos. En cada recursión el algoritmo sigue los siguientes pasos:

- Se selecciona un atributo a ser dividido. Con una correcta elección de atributos a dividir se puede generar un árbol de decisión útil. El atributo es seleccionado de acuerdo al criterio de selección especificado en los parámetros del algoritmo.

- Las instancias en el conjunto de datos son ordenadas dentro de subconjuntos, un subconjunto por cada valor en caso de ser un atributo nominal o varios subconjuntos disjuntos para rangos de valores en caso de ser atributos numéricos.

- Se retorna un árbol con una rama para cada subconjunto. Cada rama tiene un subárbol descendiente o un valor producido por la aplicación recursiva del algoritmo.

En general la recursión se detiene cuando todas las instancias tienen en mismo valor de destino o se cumplen las siguientes condiciones:

- Hay poco número de instancias en el subárbol actual

- Ningún atributo alcanza el umbral establecido. Esto se puede ajustar utilizando el parámetro de ganancia mínima.

- Se alcanza la profundidad máxima. Esto se puede modificar con el parámetro de profundidad máxima.

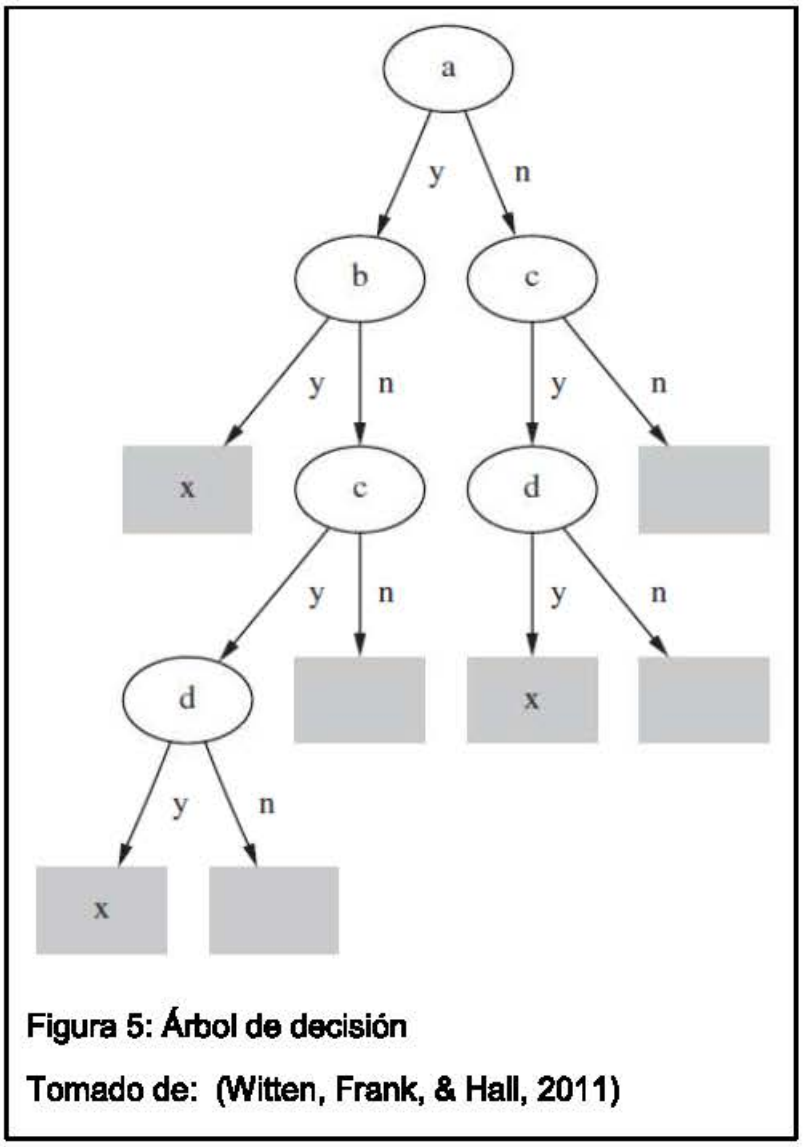


Figura 5: Árbol de decisión

Tomado de: (Witten, Frank, & Hall, 2011)

Naive Bayes

Este algoritmo de clasificación probabilística se basa en la independencia de sucesos, es decir, se asume que el valor de una característica particular de una instancia es independiente al valor de cualquier otra de sus características, por ejemplo, si una fruta es de color amarillo, tiene forma redonda y su diámetro es de 5 cm., entonces la fruta puede ser considerada como una naranja. En este ejemplo cada una de las características de la fruta contribuye de forma independiente a la probabilidad de que la fruta sea una naranja.

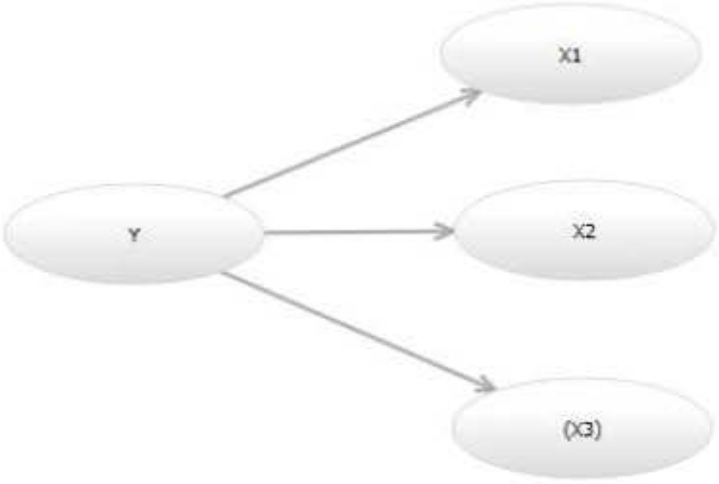


Figura 6: Modelo Naive bayes

Tomado de: **(BayesServer)**

Redes Neuronales

Históricamente las redes neuronales surgen para solucionar problemas tratando de simular la forma en que trabaja el cerebro. Hoy son generalmente vistas como poderosas técnicas de modelización.

Una red neuronal típica está construida por varias neuronas organizadas en capas para crear una red. Cada neurona puede verse como un elemento de procesamiento que se ocupa de una parte simple de la tarea que trata de resolver. Las conexiones entre las neuronas dan a la red la habilidad para aprender los patrones y las interrelaciones en los datos.

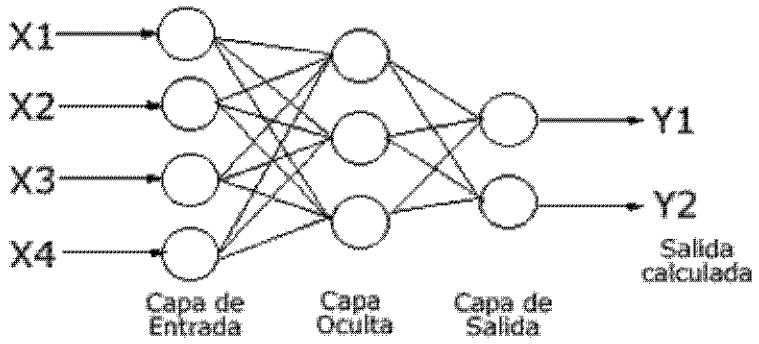


Figura 7: Red neuronal

Tomado de: (Universidad Nacional de Colombia)

La figura 7 ilustra una red neuronal simple (red de Perceptrón multicapa). La capa de entrada corresponde a las variables predictores (inputs). La capa de salida (output) contiene el campo pronosticado. La capa oculta (pueden ser varias) tiene un número de neuronas donde los resultados se combinan desde la capa anterior. Todas las neuronas en una capa de la red están conectadas a todas las neuronas en la siguiente capa.

Mientras la red neuronal aprende las relaciones entre los datos y los resultados se suele decir que está aprendiendo. Una vez se ha entrenado por completo, podemos suministrar a la red datos nuevos y ver así qué decisión toma basándose en su experiencia.

Regresión Lineal (RL)

La regresión lineal es uno de los modelos estadísticos más conocidos. La técnica básica de RL sirve para pronosticar una variable cuantitativa (Dependiente) a partir de una serie de predictores (Independientes) también cuantitativos. Sin embargo, las variables Cualitativas también se pueden incluir creando variables falsas en la base de datos. La RL asume que hay una

relación de tipo lineal entre la variable resultado y las variables que pronostican.

Métodos de selección de variables:

La selección del método permite especificar cómo se introducen las variables independientes en el análisis (predictoras). Utilizando distintos métodos se pueden construir diversos modelos de RL a partir del mismo conjunto de variables. Para introducir todas las variables independientes en un sólo paso se selecciona el método de introducción.

El método de eliminación hacia atrás, incluye en el modelo todas las variables predictoras y en cada paso se elimina la variable que no supera la respectiva prueba de hipótesis.

El método hacia adelante considera una regresión lineal simple que incluye a la variable predictora que da la correlación más alta con la variable dependiente (respuesta). Paso a paso se van incluyendo una a una otras variables predictoras según la prueba de hipótesis.

El método paso a paso es una modificación del método hacia adelante, donde una variable que ha sido incluida en el modelo en un paso previo, puede ser eliminada posteriormente.

Pasos para la revisión de un modelo de RL:

- Elegir el método de selección de variables
- Verificar si las variables independientes superan las prueba de hipótesis.
- Verificar la colinealidad entre las variables independientes (predictoras). Se recomienda emplear VIF «variance inflation factor».
- Verificar cuales coeficientes superaron la prueba de hipótesis. Aunque el coeficiente no supere la prueba, es necesario incluirlo en el modelo.

- Revisar el coeficiente de determinación, también llamado R cuadrado.
- Intentar interpretar los coeficientes, aunque en muchos casos no es posible encontrar interpretación
- ANOVA: Variación explicada vs. Variación no explicada

En la tabla 1 se proporciona un resumen de los algoritmos más importantes.

Tabla 1: Comparación de varios algoritmos

Nombre del algoritmo	**Descripción**	**Se usa en**
Asociación	Crea reglas que describen qué artículos es probable que aparezcan juntos en una transacción.	Análisis de la cesta de compras
Clústeres	Identifica relaciones en un conjunto de datos que no podría extraer lógicamente mediante la observación casual. Usa técnicas iterativas para agrupar los registros en clústeres que contengan características similares.	Detectar categorías
Árboles de decisión	Realiza predicciones basándose en las relaciones entre las columnas del conjunto de datos y modela las relaciones como series de divisiones en forma de	Clasificar Estimación

	árbol en valores específicos. Admite la predicción de atributos discretos y continuos.	
Naive Bayes	Encuentra la probabilidad de la relación entre todas las columnas de entrada y de predicción. Este algoritmo es útil para generar rápidamente modelos de minería de datos para descubrir relaciones. Admite sólo atributos discretos o discretizados. Trata todos los atributos de entrada como independientes.	Analizar influenciadores clave
Red neuronal	Analiza datos complejos de entrada o problemas empresariales para los que hay disponible una cantidad significativa de datos de aprendizaje pero de los que no se pueden derivar reglas fácilmente con otros algoritmos. Puede predecir varios atributos.	Este algoritmo se puede usar para clasificar atributos discretos y la regresión de atributos continuos.
Regresión lineal	Si existe una dependencia	Para crear un modelo

	lineal entre la variable de destino y las variables que se examinan, encuentra la relación más eficiente entre el destino y sus entradas. Admite la predicción de atributos continuos.	que utilice este algoritmo se puede crear una estructura y, a continuación, agregar manualmente un modelo.

Adaptado de: (Microsoft MSDN)

1.4. Metodologías de proyectos de minería de datos

Se ha revisado bibliografía acerca de las diferentes metodologías y modelos de procesos existentes para llevar a cabo proyectos de minería de datos y se han identificado 3 de las más utilizados, las cuales son: Cross Industry Standard Process for Data Mining (CRISP-DM); Knowledge Discovery in Databases (KDD); Sample, Explore, Modify, Model, Assess (SEMMA).

A continuación se va a realizar una breve descripción de cada uno de estos modelos:

1.4.1. Cross-Industry Standard Process for Data Mining (CRISP-DM)

Fue concebida a finales de 1996 por un consorcio de empresas (DaimlerChrysler, SPSS, NCR) y es actualmente la guía de referencia más utilizada para el desarrollo de proyectos de minería de datos. Su proceso se compone de 6 fases:

- Comprensión del negocio: se debe tener un claro entendimiento del negocio para fijar los objetivos del proyecto de minería de datos.
- Comprensión de los datos: en base a los objetivos de negocio se deben formar hipótesis sobre la información oculta en los datos.

- Preparación de los datos: incluye la selección, limpieza y transformación de los datos.
- Modelado: aquí se realiza la selección de técnicas de modelado y calibración de sus parámetros.
- Evaluación: el modelo es evaluado para verificar que cumple los objetivos del proyecto.
- Implantación: difusión del conocimiento obtenido del proceso de minería de datos.

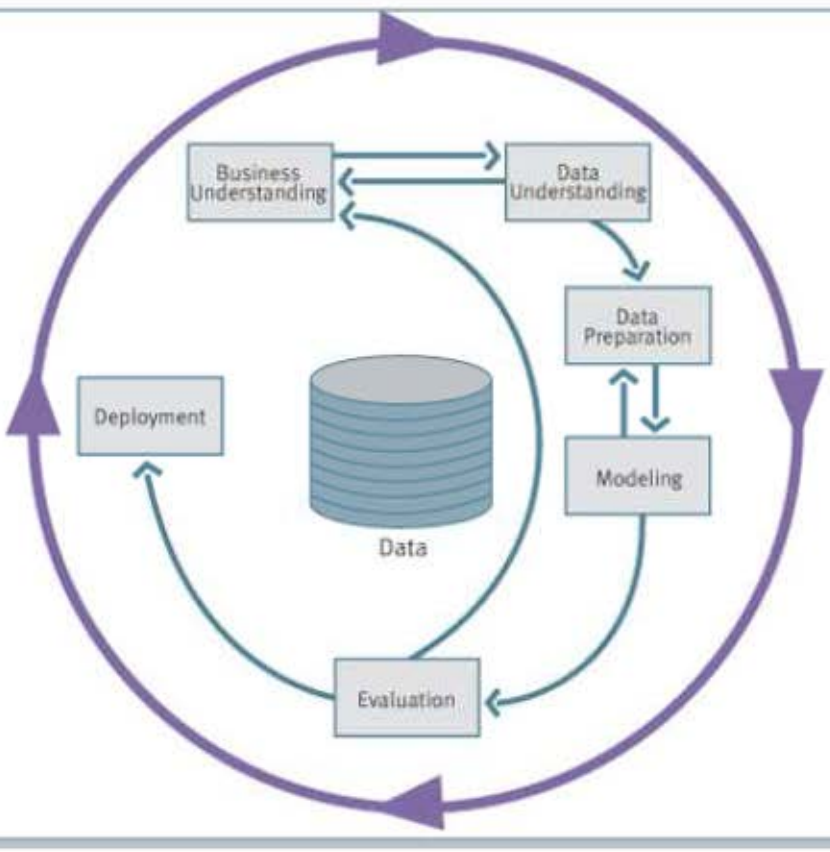


Figura 8: Fases de la metodología CRISP-DM

Tomado de: (Shearer, 2000)

1.4.2. Knowledge Discovery in Databases (KDD)

Tiene sus inicios en el año 1996. Es el proceso mediante el cual se descubre conocimiento mediante la identificación de patrones válidos de información

dentro de un gran volumen de datos (Fayyad, Piatesky-Shapiro, & Smyth, 1996). Este modelo de proceso se compone de 9 fases:

- Entendimiento del dominio de aplicación: en donde se identifican las metas del proceso de minería de datos desde el punto de vista de la organización.
- Creación del conjunto destino de datos: selección del conjunto de datos donde se va a realizar el descubrimiento.
- Limpieza y pre-procesamiento: uso de técnicas de tratamiento de campos de datos faltantes.
- Reducción y protección de datos: tareas operativas de transformación de datos en función de las metas del proceso de minería
- Definición de la tarea de minería de datos: se define el algoritmo de minería de datos que se utilizará para alcanzar las metas del proceso de minería.
- Data Mining: búsqueda de patrones válidos de datos
- Interpretación: visualización de los patrones encontrados
- Uso del conocimiento descubierto: documentación y reporte hacia los interesados

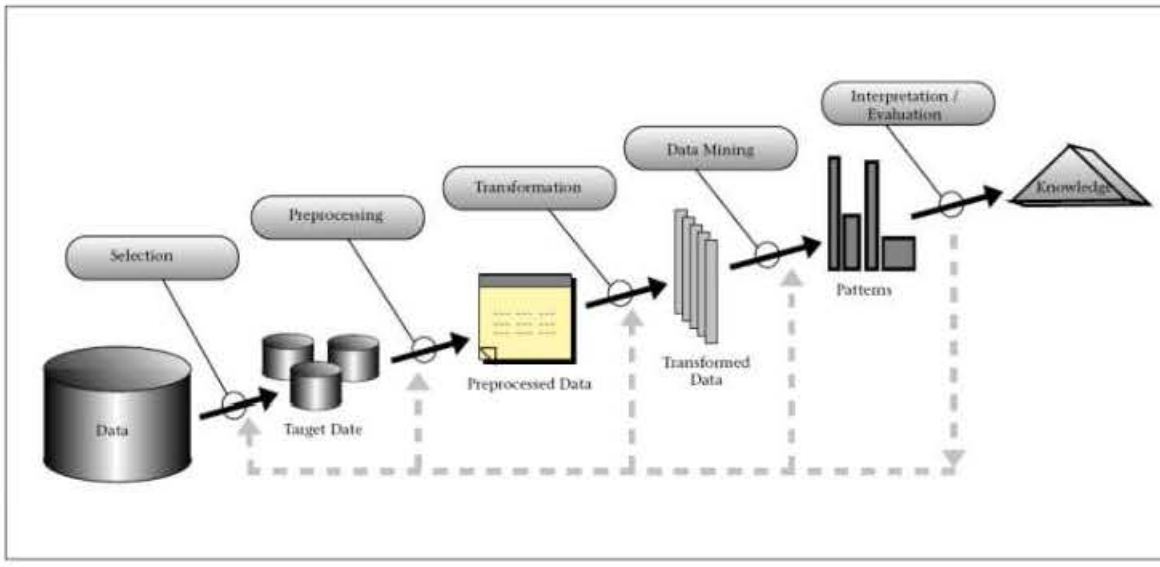


Figura 9: Fases del modelo KDD

Tomado de: (Fayyad, Piatesky-Shapiro, & Smyth, 1996)

1.4.3. Sample, Explore, Modify, Model, Assess (SEMMA)

En este modelo de minería de datos se definen las siguientes fases:

- Muestreo: de una gran cantidad de datos se extrae una pequeña porción con información significativa y que pueda ser manipulada fácilmente. Analizando una muestra representativa en lugar de todo el volumen de datos, se reduce el tiempo de procesamiento requerido para para descubrir información importante.
- Exploración: Esta etapa consiste en la exploración de los datos mediante la búsqueda de tendencias con el fin de obtener la comprensión y las ideas.
- Modificación: Esta etapa consiste en la modificación de los datos mediante la creación, selección y transformación de las variables para ajustar el proceso de selección de datos.
- Modelamiento: Esta etapa consiste en el modelamiento de los datos de tal forma que permita a la herramienta de software buscar automáticamente combinaciones de datos para predecir resultados de forma confiable.
- Valoración: Esta etapa consiste en la evaluación de la utilidad y la fiabilidad de los resultados del proceso de minería de datos.

Figura 10: Fases del proceso SEMMA

1.4.4. Comparación de las metodologías presentadas

Uno de los criterios de comparación de estas metodologías son las fases del proceso de minería de datos de cada modelo. A continuación se presenta un cuadro comparativo de las fases de cada uno de los procesos.

Tabla 2: Comparación de metodologías para proyectos de minería de datos		
CRISP-DM	**KDD**	**SEMMA**
Comprensión del negocio	aprendizaje del dominio de aplicación	
Comprensión de los datos	Creación del conjunto destino de datos	Muestreo
Preparación de los datos	Limpieza y pre procesamiento de datos	Exploración
	Reducción y proyección de datos	Modificación
Modelamiento	Determinación de la tarea de Data mining	Modelamiento
	Determinación del algoritmo de Data Mining	
	Minería de datos	
Evaluación	Interpretación	Valoración
Despliegue	Uso del conocimiento descubierto	

Adaptado de (Moine, Gordillo, & Haedo, 2011)

1.4.5. Definición de la metodología a utilizar en el proyecto

Del análisis efectuado en el punto anterior, se confirma que la metodología CRISP-DM es más completa que los modelos KDD y SEMMA y se ha convertido en un estándar de facto por ser una de las más aplicadas en proyectos de minería de datos.

KDD y CRISP-DM comienzan el proceso de minería de datos por el análisis del negocio, en cambio SEMMA inicia con un muestreo de los datos.

KDD y CRISP-DM finalizan el proceso de minería de datos con el despliegue o uso del conocimiento descubierto. Esta fase no es incluida en el modelo SEMMA.

En los modelos KDD y SEMMA se plantean solo las fases de un proyecto de minería de datos, sin llegar al detalle de las actividades que deben ejecutarse. CRISP-DM especifica con mayor detalle cada una de las fases del proceso.

Por tales razones, la metodología a utilizar para el proyecto de minería de datos del SNNA es CRISP-DM.

Capítulo 2 Comprensión del negocio

Esta etapa consiste en tener un claro entendimiento de la organización que permita fijar las metas del proyecto de minería de datos. Se revisan los objetivos generales y específicos de la organización así como su situación actual. Se trata de identificar las necesidades a resolver con el análisis de datos. En base a los objetivos de negocio se deben formar hipótesis sobre la información oculta en los datos.

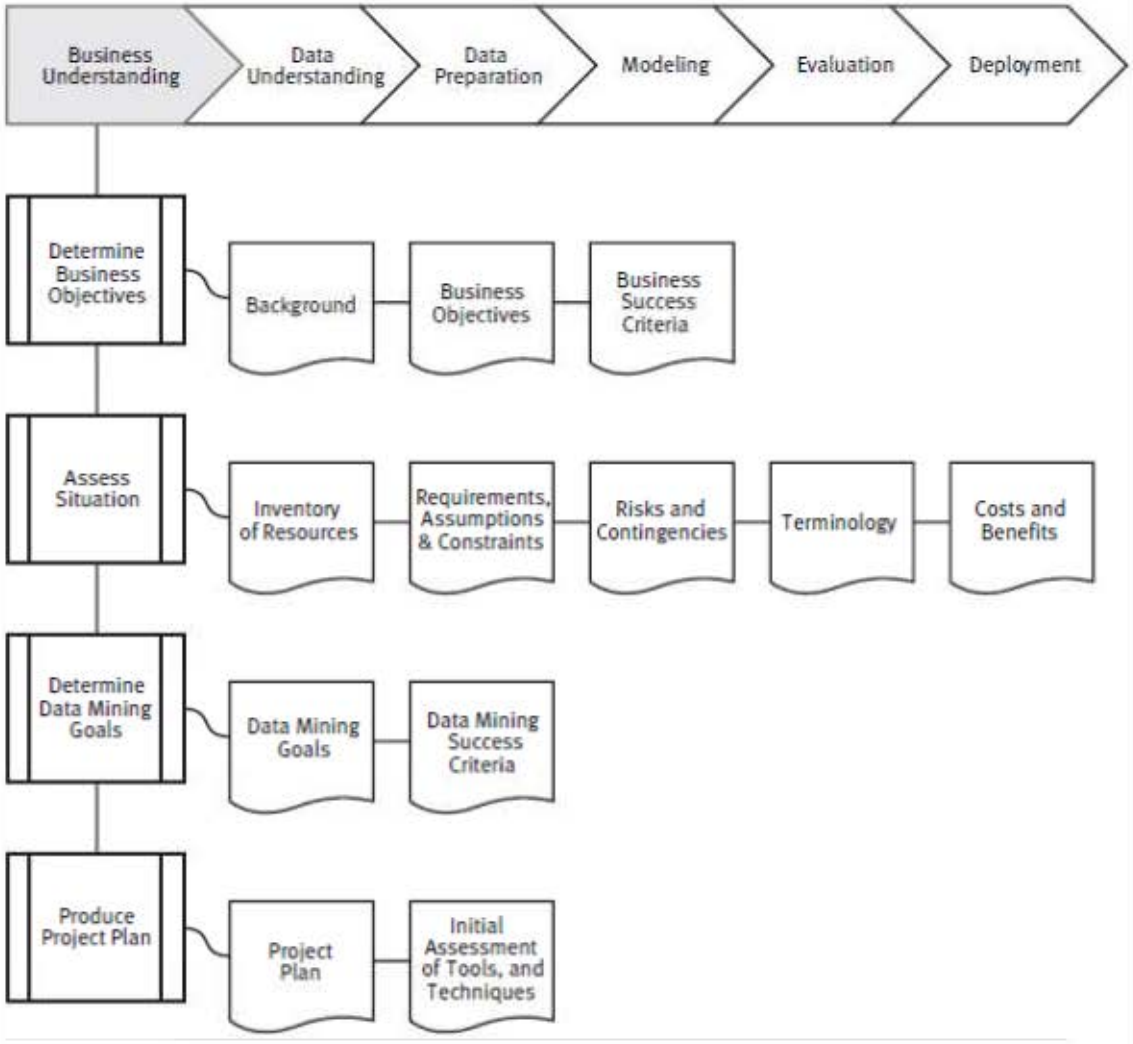


Figura 11: Comprensión del negocio

Tomado de: (Chapman, y otros, 2000)

2.1 Comprensión del negocio

El Sistema Nacional de Nivelación y Admisión (SNNA, 2013) tiene como objetivo regular el ingreso a las Instituciones de Educación Superior (IES) públicas para "garantizar la igualdad de oportunidades, la meritocracia, la transparencia acerca del acceso a la educación superior". Al considerarse un proyecto emblemático para el desarrollo de la educación superior del país y tomando en cuenta que hasta el momento se han ejecutado 4 aplicaciones a nivel país del Examen Nacional de la Educación Superior, se hace imprescindible realizar el análisis de los datos y descubrimiento de información a través de diferentes variables para que le permita a la organización tomar decisiones oportunas en base a la información que almacena el sistema informático del SNNA.

El proceso de admisión de los aspirantes a ingresar a las Instituciones de educación superior públicas se enmarca en la siguiente normativa:

Ley Orgánica de Educación Superior (LOES, 2010) en sus artículos:

Art. 81.- Sistema de Nivelación y Admisión.- El ingreso a las instituciones de educación superior públicas estará regulado a través del Sistema de Nivelación y Admisión, al que se someterán todos los y las estudiantes aspirantes.

Art. 77.- Las IES establecerán programas de becas o ayudas económicas que apoyen al 10% del número de estudiantes.

Art. 78.- La SENESCYT definirá el concepto de beca.

Art. 183.- Funciones de la Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación. Es función de la Senescyt "Diseñar, implementar, administrar y coordinar el Sistema Nacional de Información de la Educación Superior del Ecuador, y el Sistema de Nivelación y Admisión;"

Además, en base al reglamento del SNNA (Senescyt, 2013) a través del cual se establece el proceso que el aspirante debe seguir para conseguir su ingreso en las instituciones de educación superior públicas, una vez concluido el

bachillerato, a fin de realizar los estudios correspondientes en los niveles de formación técnica, tecnológica superior y de grado hasta el tercer nivel, mediante la realización de un examen de aptitud y la superación de las distintas modalidades de los cursos de nivelación.

Reglamento general a la LOES

Art. 3.- El SNNA tendrá dos componentes:

•Admisión.- que será permanente y establecerá un sistema nacional unificado de inscripciones, evaluación y asignación de cupos en función al mérito.

•Nivelación.- tomará en cuenta la heterogeneidad en la formación del bachillerato y/o las características de las carreras universitarias.

Disposición Transitoria Quinta.-

•Obligación del período académico de nivelación, organizado por las IES.

•Examen de evaluación de conocimientos con fines de exoneración del período de nivelación, organizado por las IES.

2.1.1 Determinación de los objetivos de negocio

A continuación se describen los objetivos generales y específicos de la organización, una descripción de la situación actual y las posibles preguntas acerca de la organización que se quiere responder a través del análisis de datos.

Figura 12: Objetivos del SNNA

Tomado de: (Senescyt, 2011)

2.1.1.1 Objetivo General

Garantizar la igualdad de oportunidades, la meritocracia, transparencia y acceso a la educación superior.

2.1.1.2 Objetivos específicos

•Diseñar, implementar y administrar un Sistema de Admisión, que potencie la pertinencia de la oferta académica, una adecuada ocupabilidad de las vacantes, que sea equitativo y meritocrático, basado en la aplicación de pruebas estandarizadas debidamente validadas.

•Diseñar y financiar el Sistema de Nivelación impartido por IES públicas que garantice la igualdad de oportunidades y compense las asimetrías formativas antes del ingreso a las carreras.

2.2 Valoración de la situación actual

El proceso de Admisión del SNNA se encuentra informatizado a través de una aplicación web que permite la recolección inicial de datos de los aspirantes y además se utilizan aplicaciones informáticas desarrolladas localmente para los subprocesos que se ejecutan en cada aplicación del ENES (Examen Nacional para la Educación Superior), pero actualmente el análisis de estos datos se lo hace a partir de reportes y filtros en hojas de cálculo. El tipo de análisis de datos que se realiza en la actualidad consiste principalmente en operaciones de hoja de cálculo como filtros y tablas dinámicas para la obtención de resultados numéricos y gráficos sobre estos resultados. Uno de los principales inconvenientes de este método es el volumen de datos que se requiere analizar ya que la cantidad de registros se acerca a los 150.000 por cada proceso de aplicación del ENES y con cada registro de datos compuesto por cerca de 200 campos, lo que dificulta el trabajo con esta cantidad de datos.

Otro inconveniente es que la información es analizada por cada proceso de aplicación de la prueba ENES de forma independiente, por ejemplo, en una hoja de cálculo se analizan los datos del proceso ENES del 1er. período académico del 2012 mientras en otra hoja de cálculo se analizan los datos del 2do. Período académico del 2012 y no se ha podido realizar un análisis consolidado de información de todos los procesos ENES ejecutados hasta la actualidad por el SNNA.

Además, al momento no se cuenta con una herramienta para la realización de análisis estadístico de la información que permita generar indicadores sobre los procesos ENES efectuados hasta el momento.

Por lo expuesto anteriormente se requiere contar con un mecanismo que le permita a la Institución realizar un análisis más efectivo de los datos que se generan en cada proceso del ENES.

Esto hace que sea imprescindible el uso de herramientas y técnicas de inteligencia de negocios con la finalidad de poder explotar los datos obtenidos en cada proceso y generar indicadores que permitan reforzar las decisiones políticas y faciliten el mejoramiento continuo del proceso.

2.3 Determinación de los objetivos del proyecto de minería de datos

Uno de los objetivos del proyecto es descubrir cómo aplicar los algoritmos de aprendizaje supervisado para descubrir las relaciones entre los atributos de los aspirantes a la educación superior.

Otro de los objetivos de este proyecto es construir un modelo de clasificación de datos que permita descubrir las relaciones entre los atributos socio-económicos de los aspirantes a la educación superior y los resultados de su examen de admisión.

2.4 Elaboración del plan del proyecto

A continuación se define el plan del proyecto y la secuencia de actividades a realizar durante el resto del proyecto, incluyendo la selección de herramientas y técnicas.

De acuerdo a la metodología seleccionada CRISP-DM, se deben realizar las siguientes tareas durante todo el proyecto:

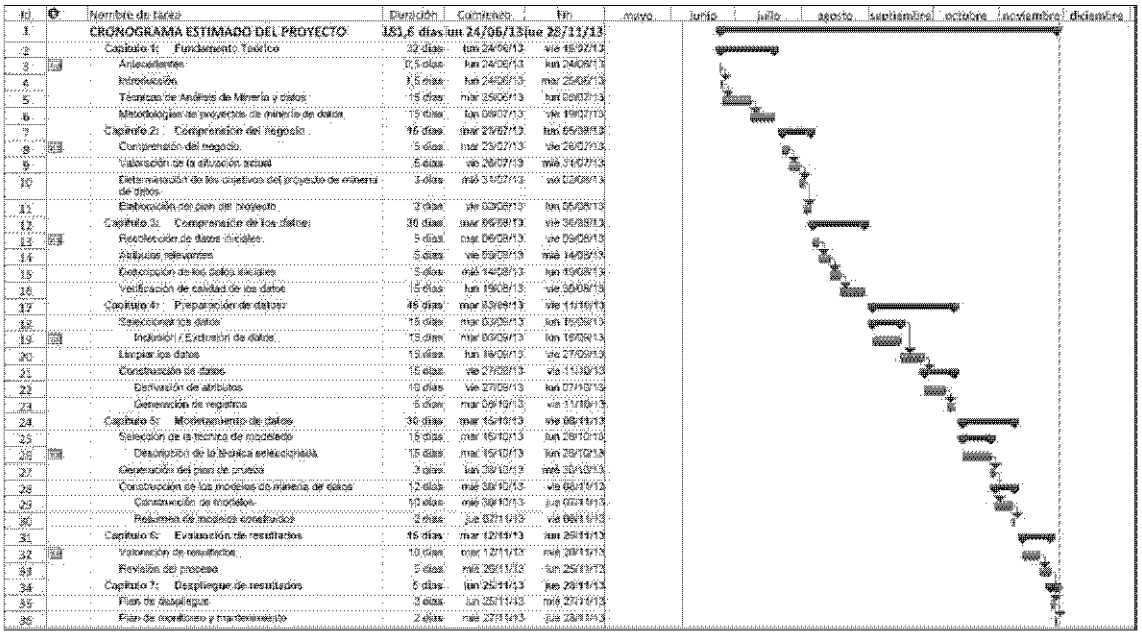
- Comprensión del negocio: Esta etapa consiste en tener un claro entendimiento del negocio que permita fijar las metas del proyecto de minería de datos. Se revisan los objetivos generales y específicos de la organización así como su situación actual. Se trata de identificar las necesidades a resolver con el análisis de datos. En base a los objetivos

de negocio se deben formar hipótesis sobre la información oculta en los datos.

- Comprensión de los datos: Esta etapa consiste principalmente en la recolección de los datos que se desea analizar, la descripción e identificación de la calidad de los mismos.
- Preparación de los datos: Esta etapa consiste en definir un conjunto de datos que contenga los atributos considerados candidatos para estimar el valor de la variable que se va a predecir.
- Modelado: En esta etapa se debe elegir la técnica de modelamiento de datos que se va a utilizar sobre el conjunto de datos definido en la etapa anterior. El objetivo es descubrir la relación del conjunto de datos y el atributo que se desea predecir. Las principales actividades en esta etapa son, la selección de la técnica de modelamiento, y la generación del plan de pruebas y la construcción del modelo de datos.
- Evaluación: En esta etapa del proyecto se evalúa la precisión y generalidad del modelo de datos. Debe evaluarse el cumplimiento de los objetivos del proyecto de minería de datos. En este punto debe determinarse si es necesario que se realicen iteraciones adicionales o se debe proseguir con la etapa de despliegue.
- Implantación: En esta etapa se realiza la difusión del conocimiento obtenido del proceso de minería de datos. La actividad más relevante es la generación del reporte final del proyecto.

2.4.1. Cronograma

En función de las actividades definidas en el punto anterior se estima un tiempo de 180 días calendario para la ejecución del proyecto y para el desarrollo de todas las actividades contempladas en la metodología.



Id	Nombre de tarea	Duración	Comienzo	Fin
1	CRONOGRAMA ESTIMADO DEL PROYECTO	151,5 días	lun 24/06/13	lue 28/11/13
2	Capítulo 1: Fundamento Teórico	12 días	lun 24/06/13	vie 05/07/13
3	Antecedentes	0,5 días	lun 24/06/13	lun 24/06/13
4	Introducción	1,5 días	lue 24/06/13	mar 25/06/13
5	Técnicas de Análisis de Minería y datos	15 días	mar 25/06/13	lun 08/07/13
6	Metodologías de proyectos de minería de datos	15 días	lun 08/07/13	vie 19/07/13
7	Capítulo 2: Comprensión del negocio	16 días	mar 23/07/13	lun 05/08/13
8	Comprensión del negocio	5 días	mar 23/07/13	vie 26/07/13
9	Valoración de la situación actual	5 días	vie 26/07/13	mié 31/07/13
10	Determinación de los objetivos del proyecto de minería de datos	3 días	mié 31/07/13	vie 02/08/13
11	Elaboración del plan del proyecto	2 días	vie 02/08/13	lun 05/08/13
12	Capítulo 3: Comprensión de los datos	30 días	mar 06/08/13	vie 06/09/13
13	Recolección de datos iniciales	5 días	mar 06/08/13	vie 09/08/13
14	Atributos relevantes	5 días	vie 09/08/13	mié 14/08/13
15	Descripción de los datos iniciales	5 días	mié 14/08/13	lun 19/08/13
16	Verificación de calidad de los datos	15 días	lun 19/08/13	vie 06/09/13
17	Capítulo 4: Preparación de datos	46 días	mar 03/09/13	vie 11/10/13
18	Seleccionar los datos	15 días	mar 03/09/13	lun 16/09/13
19	Inclusión / Exclusión de datos	15 días	mar 03/09/13	lun 16/09/13
20	Limpiar los datos	15 días	lun 16/09/13	vie 27/09/13
21	Construcción de datos	15 días	vie 27/09/13	vie 11/10/13
22	Derivación de atributos	10 días	vie 27/09/13	lun 07/10/13
23	Generación de registros	5 días	mar 08/10/13	vie 11/10/13
24	Capítulo 5: Modelamiento de datos	30 días	mar 15/10/13	vie 08/11/13
25	Selección de la técnica de modelado	15 días	mar 15/10/13	lun 28/10/13
26	Descripción de la técnica seleccionada	15 días	mar 15/10/13	lun 28/10/13
27	Generación del plan de prueba	3 días	lun 28/10/13	mié 30/10/13
28	Construcción de los modelos de minería de datos	12 días	mié 30/10/13	vie 08/11/13
29	Construcción de modelos	10 días	mié 30/10/13	jue 07/11/13
30	Resumen de modelos construidos	2 días	jue 07/11/13	vie 08/11/13
31	Capítulo 6: Evaluación de resultados	15 días	mar 12/11/13	lun 25/11/13
32	Valoración de resultados	10 días	mar 12/11/13	mié 20/11/13
33	Revisión del proceso	5 días	mié 20/11/13	lun 25/11/13
34	Capítulo 7: Despliegue de resultados	5 días	lun 25/11/13	lue 28/11/13
35	Plan de despliegue	2 días	lun 25/11/13	mié 27/11/13
36	Plan de supervisión y mantenimiento	2 días	mié 27/11/13	jue 28/11/13

Figura 13: Cronograma del proyecto

Capítulo 3 Comprensión de los datos

Esta etapa consiste principalmente en la recolección de los datos que se desea analizar, la descripción e identificación de la calidad de los mismos. En la descripción de los datos se especifica el total de registros de datos así como el número de campos por registro y el significado de cada campo. En la verificación de calidad de los datos se revisa la consistencia de los mismos en relación a valores válidos permitidos para cada campo.

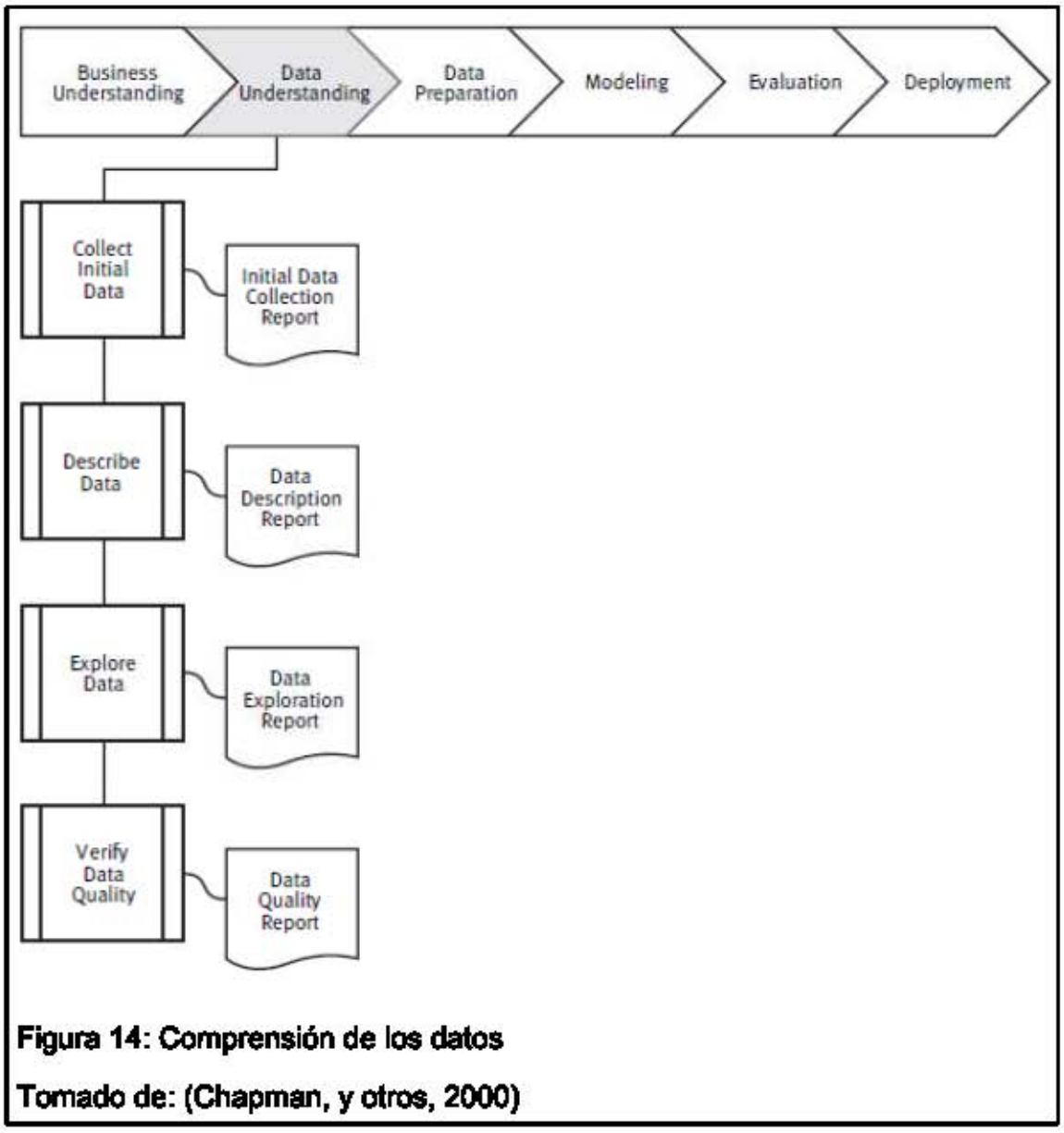


Figura 14: Comprensión de los datos

Tomado de: (Chapman, y otros, 2000)

3.1 Recolección de datos iniciales

Los datos para el análisis se obtienen en cada uno de los pasos del proceso de admisión de los aspirantes a ingresar a las universidades públicas. Estos datos son registrados es la base de datos transaccional del sistema informático utilizado para el proceso de admisión. La figura 15 muestra la cantidad de aspirantes que participan conforme el proceso avanza hasta la fase de asignación de cupos.

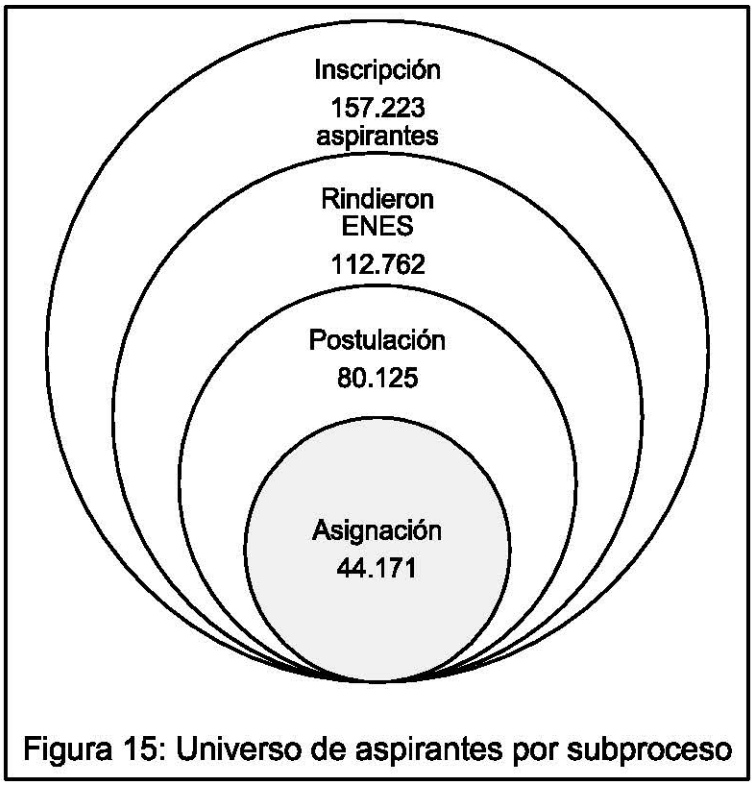


Figura 15: Universo de aspirantes por subproceso

El proceso de Inscripción a través de la plataforma web es el que mayor cantidad de registros a nivel de usuario produce. Generalmente en cada proceso se estima que se inscriban aproximadamente 150.000 aspirantes.

Los aspirantes inscritos al examen deben presentarse el día de la aplicación a nivel nacional presentando su comprobante de inscripción. Por lo general no todos los aspirantes que se Inscriben se presentan a rendir el examen y se estima que un 70% de los aspirantes que se inscriben se presentan efectivamente a rendir el examen.

De todos los aspirantes que rindieron el examen no todos pueden postular a una carrera debido a que se toma en cuenta la nota obtenida. Para el examen del 28 de septiembre pasado, si la nota es mayor a 600 puntos entonces el aspirante puede postular por una carrera. Además, se presentan casos en los que los aspirantes no están conformes con la nota obtenida y prefieren rendir un próximo ENES con la meta de mejorar la nota obtenida. En este proceso ENES, el 51% de los aspirantes que se inscribieron participaron en el proceso de postulación.

Una vez ejecutado el proceso de asignación de cupos, la cantidad de aspirantes que efectivamente consiguieron el cupo que escogieron es del 55% de aspirantes que postularon. Para el caso del presente análisis el universo final de aspirantes que obtuvieron cupo en este proceso es de aproximadamente 44.171 personas.

3.1.1 Inscripción

Todos los aspirantes que van a ingresar a estudiar una carrera en una Institución de Educación Superior pública deben inscribirse a través de la plataforma web que el SNNA ha creado. Una vez que el aspirante ha creado su cuenta de usuario debe inscribirse registrando sus datos personales y datos de ubicación.

Los requisitos para inscribirse al ENES definidos por el (portal SNNA) son:

- Ser bachiller o estar cursando el tercer año de bachillerato.
- Cédula de ciudadanía.
- Fotografía tamaño carné con fondo blanco, digital en formato JPG de hasta 100 kb.
- Tener una cuenta de correo electrónico activa.
- En el caso de aspirantes extranjeros el documento habilitante es el pasaporte o carnet de refugiado.

Datos necesarios para la inscripción

Para la inscripción se le pide al aspirante que ingrese al sistema la siguiente información:

Tabla 3: Datos necesarios para la inscripción

DATOS PERSONALES	RESIDENCIA
• Número de Cédula • Nombres • Apellidos • Fotografía tamaño carné • Fecha de nacimiento (edad) • Nacionalidad • N° carné CONADIS (si tiene alguna discapacidad)	• Provincia • Cantón • Parroquia • Dirección del domicilio
DATOS DE CONTACTO	**COLEGIO DE PROCEDENCIA**
• Teléfono • Número telefónico celular • Dirección de correo electrónico	• Nombre • Provincia • Cantón • Parroquia • Fecha de graduación o fecha estimada para obtención del título de bachiller • Título o Acta de Grado (opcional)

Tomado de: (portal SNNA)

3.1.2 Encuesta de Contexto

Todos los aspirantes que vayan a rendir el ENES deben ingresar a su cuenta de usuario en el sistema informático del SNNA (www.snna.gob.ec) para completar la encuesta de contexto, la cual recolecta la información socio económica del aspirante. Ver el Anexo 2

3.1.3 Aplicación y calificación del ENES

Una vez que el ENES ha sido aplicado a nivel nacional, las pruebas son calificadas de forma automática y luego resultados son publicados en la cuenta de cada aspirante.

3.1.4 Postulación

Para el proceso de Postulación es necesario haber realizado antes el proceso de Inscripción y haber rendido el examen de admisión correspondiente a la fecha del mismo período. En este proceso el sistema permite al aspirante elegir 5 carreras de su preferencia. Al seleccionar cualquiera de los registros el aspirante puede elegir el NIVEL de estudios que desea (tercer nivel, nivel técnico o nivel superior tecnológico) en donde el tercer nivel es realizado por las Universidades y el nivel técnico o superior tecnológico lo dan los Institutos. Además se ingresará:

Nombre de la carrera: en este campo se despliegan varias carreras relacionadas con el texto ingresado y el aspirante debe elegir su opción de preferencia.

IES: este campo muestra la Institución de Educación Superior que está ofertando cupos para la carrera seleccionada

Campus: en este campo se despliega el Campus (Cuidad) donde la carrera está habilitada.

3.1.5 Asignación

El proceso de asignación de cupos definido por el (portal SNNA) determina que los cupos son distribuidos en función de:

- El puntaje obtenido en el ENES
- El número de cupos reportados por las instituciones de educación superior
- El Orden de selección de las opciones de carrera.

El sistema informático ha sido programado para que el proceso de asignación de cupos se realice de la siguiente manera:

Las notas son ordenadas de mayor a menor en estricto orden, el sistema irá ubicando a los aspirantes de acuerdo al número de cupos reportados por las instituciones de educación superior y de las PREFERENCIAS señaladas por los aspirantes. Así, si en una institución existen 200 cupos disponibles en la carrera xyz, modalidad presencial en el campus N, obtendrán un cupo los 200 aspirantes MEJOR PUNTUADOS en el ENES que seleccionaron entre sus opciones esta carrera, siempre iniciando en la primera opción.

Para obtener un cupo en las carreras de alta sensibilidad social (medicina y educación) es necesario obtener en el ENES un puntaje mayor o igual a 800 puntos.

3.2 Atributos relevantes

A continuación se obtienen datos relevantes en cada uno de los subprocesos que se ejecutan en admisión:

La tabla 4 muestra los atributos demográficos de los aspirantes:

Tabla 4: Atributos demográficos	
Nombre	**Descripción**
GENERO	Masculino o Femenino
EDAD	Edad del aspirante
PROVINCIA_NACIMIENTO	Nombre de la Provincia de nacimiento del aspirante
CANTON_NACIMIENTO	Nombre del Cantón de nacimiento del aspirante
PARROQUIA_NACIMIENTO	Nombre de la Parroquia de nacimiento del aspirante
PROVINCIA_RESIDENCIA	Nombre de la Provincia de residencia del aspirante
CANTON_RESIDENCIA	Nombre del Cantón de residencia del aspirante
PARROQUIA_RESIDENCIA	Nombre de la Parroquia de residencia del aspirante
AREA_RESIDENCIA	Urbana o Rural
ESTADO_CIVIL	Estado civil del aspirante
PROVINCIA_UED	Nombre de la Provincia de la Unidad Educativa del aspirante
CANTON_UED	Nombre del Cantón de la Unidad Educativa del aspirante
PARROQUIA_UED	Nombre de la Parroquia de Unidad Educativa del aspirante
NOMBRE_UED	Nombre de Unidad Educativa del aspirante
TIPO_UED	Tipo de Unidad Educativa del aspirante
REGIMEN_UED	Costa o Sierra
SECTOR_ZONA	Zonas de planificación territorial

La tabla 5 presenta los atributos socio-económicos de los aspirantes:

Tabla 5: Atributos socio-económicos	
Nombre	**Descripción**
NIVEL_EDUCATIVO_JH	Nivel educativo del jefe de hogar
ACTIVIDAD_JH	Actividad laboral del jefe de hogar
MATERIAL_EXTERIOR_VIVIENDA	Adobe, caña, ladrillo, etc.
MATERIAL_PISO_VIVIENDA	Cerámica, duelo, tierra, etc.
AGUA	Fuente de suministro de agua
TIPO_SERVICIO_HIGENICO	Conectado a red pública de alcantarillado, pozo ciego, etc.
DORMITORIOS	Número de dormitorios en la vivienda
TIPO_VIVIENDA	Propia, arrendada, etc.
TIENE_CELULAR	¿Tiene servicio de telefonía celular?
TIENE_COMPUTADOR	¿Tiene computador personal?
TIENE_INTERNET	¿Tiene servicio de Internet?
TV_PAGADA	¿Tiene servicio de televisión pagada?
RED_SOCIAL_1	¿Tiene cuenta en Facebook?
TOTAL_CELULARES	Número de teléfonos celulares

Estos atributos han sido consolidados en una tabla de base de datos para facilitar el análisis de la información.

3.3 Descripción de datos iniciales

3.3.1 Cantidad de datos

Considerando que hasta el momento se han ejecutado 5 aplicaciones del ENES a nivel nacional y que cada aplicación tiene en promedio 150.000 aspirantes inscritos, la cantidad de tablas, columnas y registros es alta. Por tal motivo se ha considerado realizar el análisis de un subconjunto de datos correspondiente a una sola aplicación del ENES. Esto significa que el análisis se realizará sobre un conjunto de datos aproximado de 44.000 registros y 31 columnas. Este mismo estudio se podría aplicar sobre todo el conjunto de datos si el SNNA lo considera pertinente.

Toda la información del proceso se almacena en una única base de datos, lo que facilitará la tarea de preparación de datos.

3.4 Verificación de la calidad de los datos

De acuerdo a (Olson, 2003) los datos son de calidad si satisfacen los requerimientos de uso de los mismos, es decir, la calidad depende mucho del uso que se quiera dar a los datos. Para ello los datos deben ser precisos, relevantes, completos y confiables.

El proceso de inscripción en el SNNA se lo realiza a través de un formulario web que tiene validaciones en todos los campos numéricos y tipo texto donde el aspirante debe ingresar información y el resto de campos son elegidos de varios catálogos del sistema.

Los procesos de calificación, postulación y asignación almacenan datos en el sistema de acuerdo al resultado del ENES, la elección del aspirante y las reglas

de negocio definidas en el sistema, por lo que la calidad de estos datos se mantendrá para todos los registros.

El proceso de encuesta de contexto se compone de un formulario web conformado de múltiples campos tipo texto, numéricos, de selección, listas, etc., y aquí es donde podemos encontrar varios campos vacíos debido a que algunos de ellos no son obligatorios como por ejemplo ¿qué idioma habla como segunda lengua? Además, existen sub-preguntas que deben ser contestadas dependiendo de la respuesta a la pregunta que la antecede. Esto hace también que existan varios campos vacíos.

3.4.1 Perfilamiento de datos y manejo de excepciones

Según (Oracle) el perfilamiento de datos es la evaluación de la calidad de los datos de un sistema u organización

A continuación se presenta el análisis de los atributos más relevantes con la finalidad de entender su contenido y describir las principales reglas de validación utilizadas en los diferentes campos (atributos). La herramienta utilizada para desarrollar esta actividad es Data Quality Services de Microsoft SQL Server 2012.

Atributo: género

Este atributo contiene solo 2 posibles valores, F para indicar género femenino y M para masculino. En la verificación de calidad se reemplazan estos valores iniciales por su significado completo.

Dato de entrada	Dato de salida
F	Femenino
M	Masculino

Figura 16: Atributo "género"

Atributo: edad

Este atributo es de tipo decimal. En la verificación de calidad todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

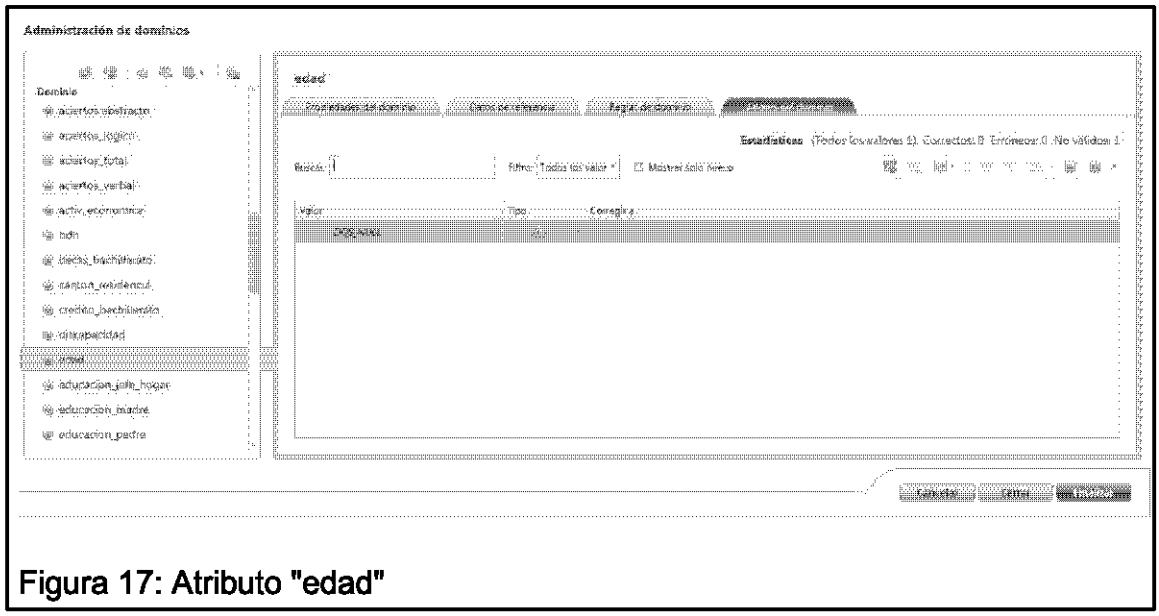


Figura 17: Atributo "edad"

Atributo: Región de residencia:

Este atributo contiene los nombres de las 4 regiones geográficas que tiene el Ecuador (costa, sierra, oriente, galápagos). En la verificación de calidad todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

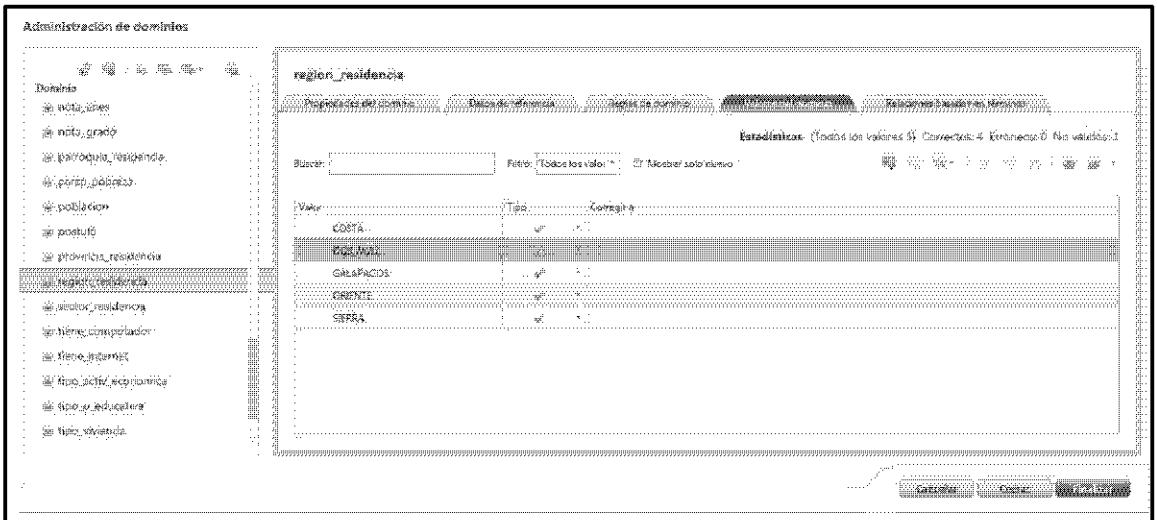


Figura 18: Atributo "región"

Atributo: Provincia de residencia:

Este atributo contiene los nombres de las 24 provincias que tiene el Ecuador. En la verificación de calidad todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

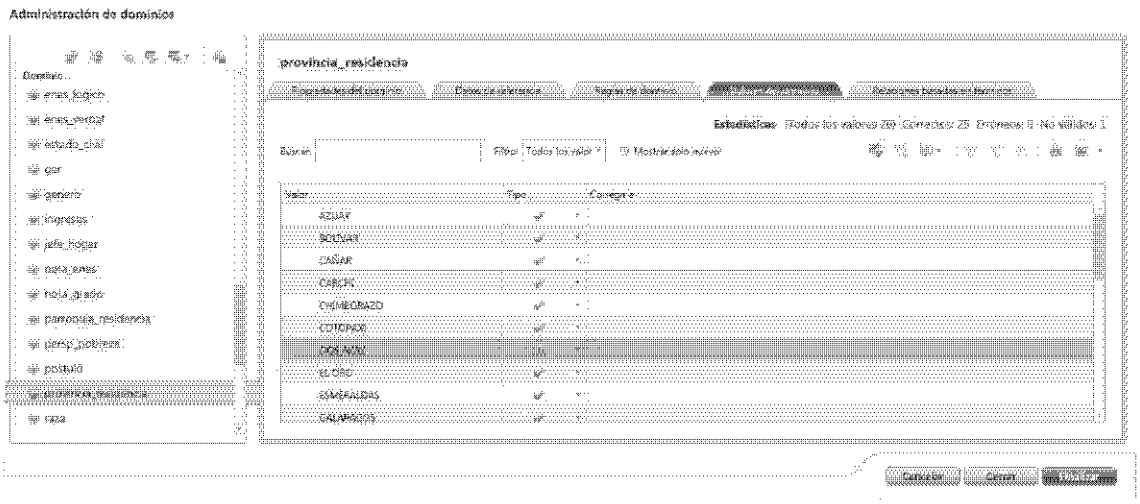


Figura 19: Atributo "provincia"

Atributo: Cantón de residencia:

Este atributo contiene los nombres de todos los cantones que tiene el Ecuador. En la verificación de calidad todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

Atributo: Parroquia de residencia:

Este atributo contiene los nombres de todas las parroquias que tiene el Ecuador. En la verificación de calidad todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

Atributo: Sector de Residencia:

Este atributo contiene el sector donde reside actualmente el aspirante, los valores posibles son U para indicar que se trata de sector urbano y R para el sector rural. En la verificación de calidad se corrigen estos valores iniciales con su significado completo en forma descriptiva.

Dato de entrada	Dato de salida
U	Urbano
R	Rural

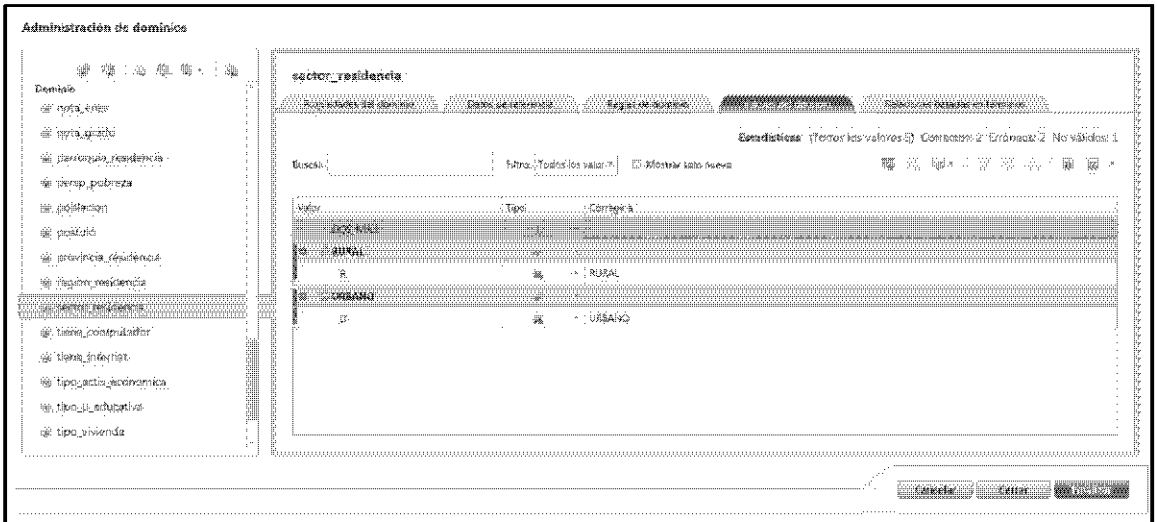


Figura 20: Atributo "sector"

Atributo: Estado civil

Este atributo contiene el estado civil del aspirante. En la verificación de calidad se reemplazan los valores iniciales por su significado completo.

Dato de entrada	Dato de salida
C	Casado(a)
D	Divorciado(a)
S	Soltero(a)
U	Unión libre
V	Viudo

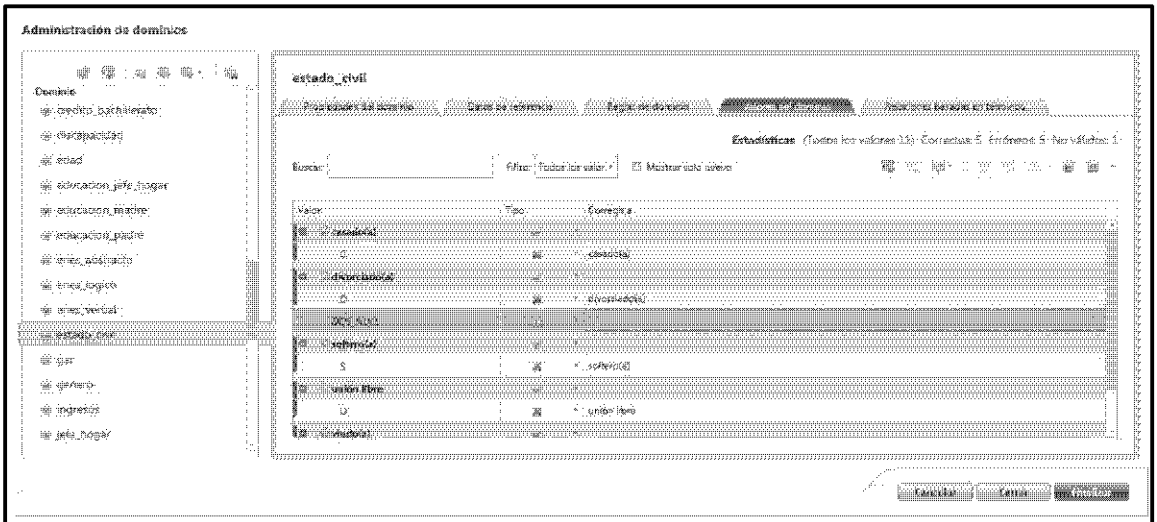


Figura 21: Atributo "estado_civil"

Atributo: Discapacidad

Este atributo contiene solo 2 posibles valores, S para indicar que el aspirante tiene algún tipo de discapacidad y N si no tiene discapacidad. En la verificación de calidad se reemplazan estos valores iniciales por su significado completo.

Dato de entrada	Dato de salida
N	No
S	Sí

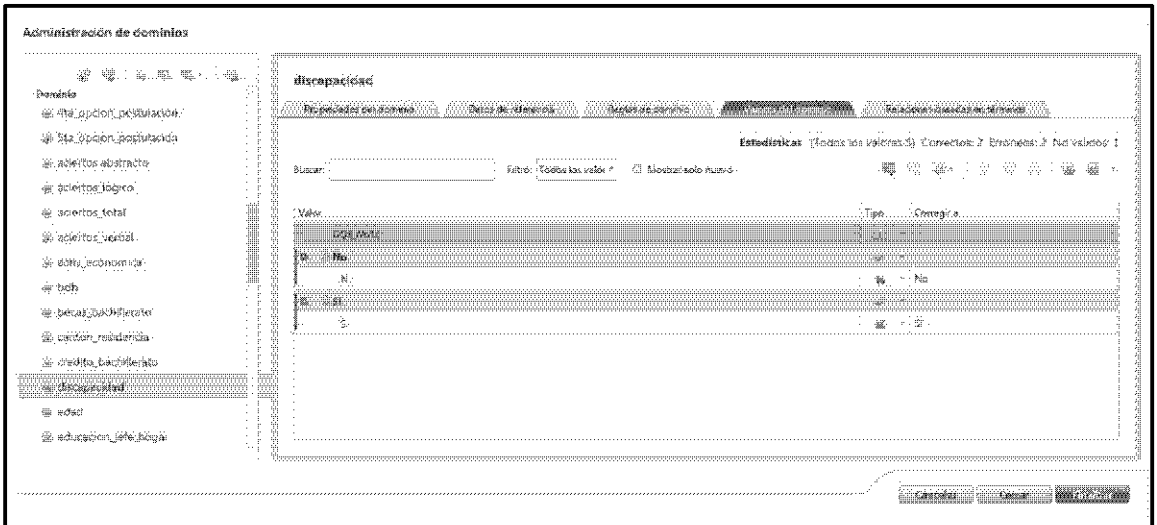


Figura 22: Atributo "discapacidad"

Atributo: Unidad Educativa:

Este atributo contiene los nombres de las unidades educativas (colegios) de la cual provienen los aspirantes. En la verificación de calidad se revisan principalmente diferencias producidas por signos de puntuación o mala digitación. En este caso de ejemplo se tienen 2 valores similares para el campo "unidad_educativa". Lo que se hace es tomar el valor correcto del campo como referencia para corregir todos los valores incorrectos encontrados dentro de este mismo campo. Todos los valores incorrectos encontrados son corregidos automáticamente al valor correcto. Por ejemplo:

Dato de entrada	Dato de salida
DR. CAMILO GALLEGOS DOMINGUEZ	DR CAMILO GALLEGOS DOMINGUEZ
(con punto)	(sin punto)

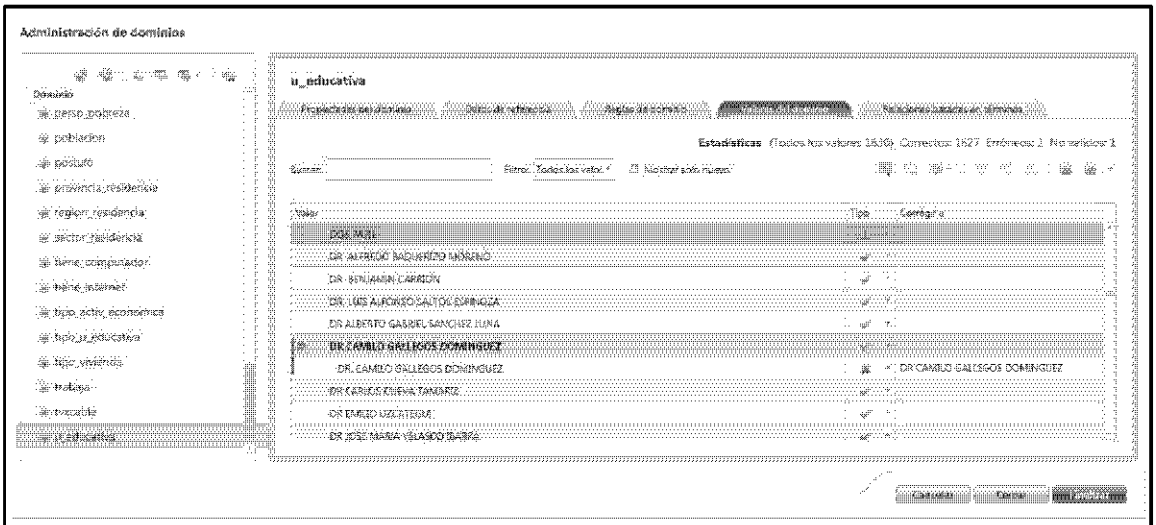


Figura 23: Atributo "unidad_educativa"

Atributo: Tipo unidad educativa

Este atributo contiene el tipo de unidad educativa de la cual proviene el aspirante. En la verificación de calidad se verifican que los valores válidos sean [fiscal, fisco misional, municipal, particular]. Todos los valores nulos de este campo son marcados como inválidos y serán omitidos del resultado final.

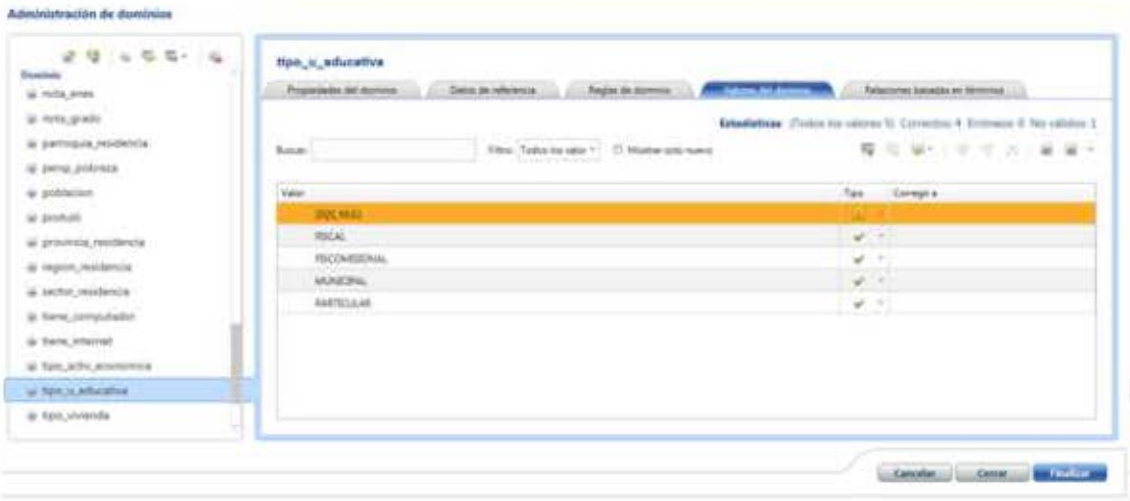


Figura 24: Atributo "tipo_unidad_educativa"

A continuación se presenta el resultado del perfilamiento de datos

Dominio	Valores corregidos		Valores sugeridos		Integridad
parroquia_residencia	551	(0 %)	557	(0 %)	
sector_residencia	115534	(100 %)	0	(0 %)	
estado_civil	0	(0 %)	0	(0 %)	
discapacidad	115819	(100 %)	0	(0 %)	
u_educativa	1210	(1 %)	1140	(1 %)	
tipo_u_educativa	0	(0 %)	0	(0 %)	
educacion_madre	0	(0 %)	0	(0 %)	
educacion_padre	0	(0 %)	0	(0 %)	
tipo_vivienda	0	(0 %)	0	(0 %)	
tiene_computador	0	(0 %)	0	(0 %)	
tiene_internet	0	(0 %)	0	(0 %)	
tvcable	0	(0 %)	0	(0 %)	

Estadísticas de origen

Registros:	115819	
Registros corregidos:	17164	(15 %)
Registros sugeridos:	981	(1 %)

Figura 25: Resumen perfilamiento de datos

Capítulo 4 Preparación de los datos

Esta etapa consiste en definir un conjunto de datos que contenga los atributos considerados candidatos para estimar el valor de la variable que se va a analizar. Para ello se realizan tareas de selección de datos, limpieza y construcción de datos adicionales. El conjunto de datos construido en esta etapa debe tener el formato adecuado requerido para la etapa de modelamiento en el posterior capítulo.

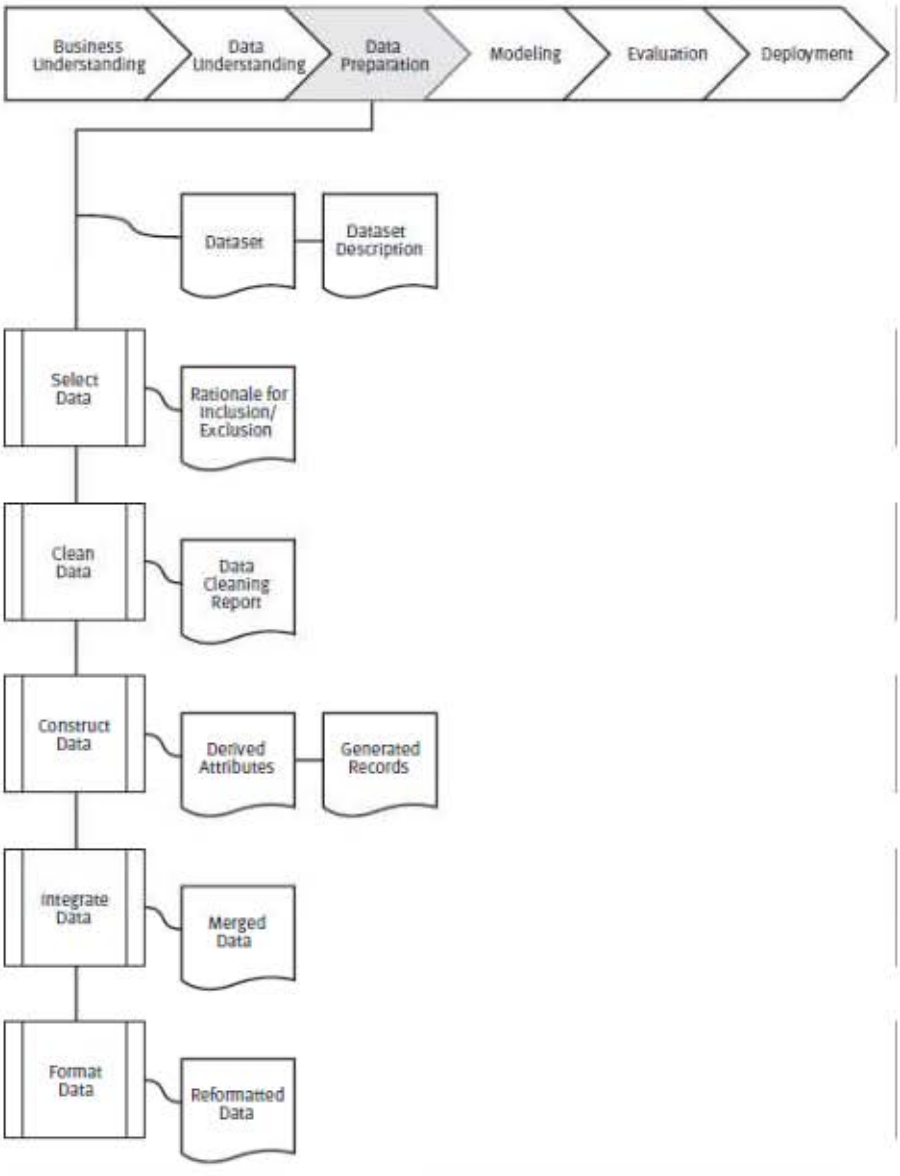


Figura 26: Preparación de los datos

Referencia: (Chapman, y otros, 2000)

4.1 Selección de los datos

En base al conocimiento de los datos, realizado en el capítulo anterior, se procederá a seleccionar los datos para el análisis de acuerdo a los siguientes criterios:

4.1.1 Selección de registros

El presente análisis se limitará a los datos correspondientes al proceso ENES del 28 de septiembre del 2013 que son aproximadamente 112.000 aspirantes los que han rendido el examen y han llenado su encuesta de contexto. Datos obtenidos del (portal SNNA).

4.1.2 Selección de atributos

Los atributos más relevantes para el análisis son los que corresponden a:

- Datos de procedencia del aspirante
- Datos de la evaluación del aspirante
- Datos de la encuesta de contexto

A continuación se muestra la construcción del repositorio de datos que se utilizará en el análisis, utilizando la herramienta RapidMiner. Todos estos atributos han sido juntados en una sola tabla para facilitar el análisis de la información como se puede observar en el flujo diseñado de la Figura 27.

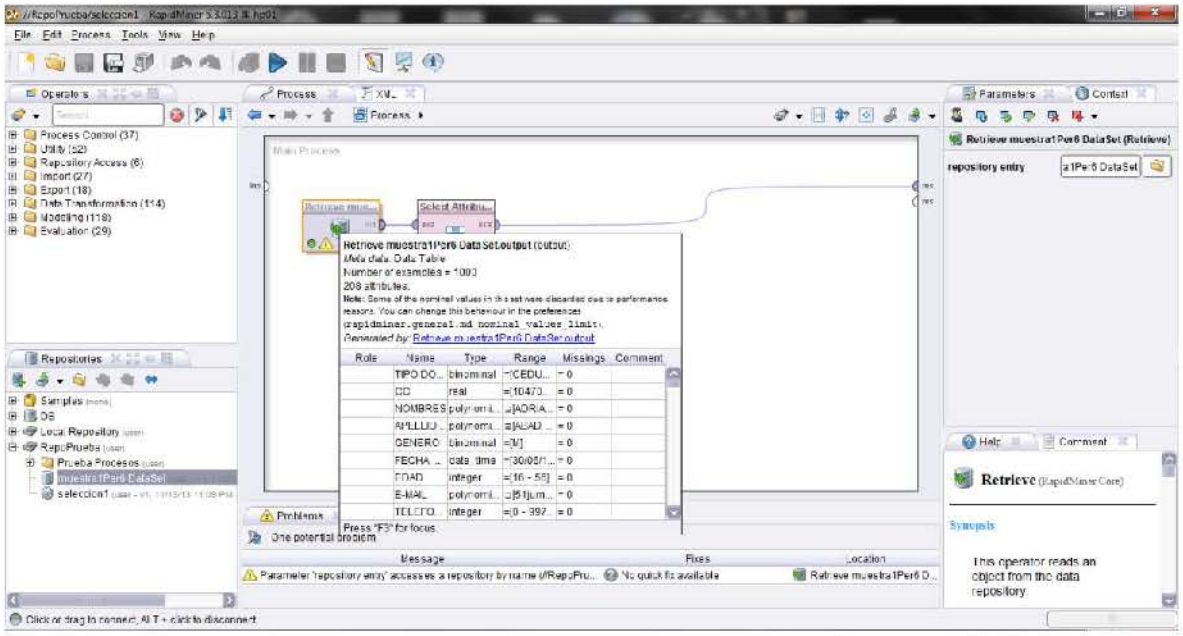


Figura 27: Repositorio de datos

4.1.3 Inclusión / Exclusión de datos

Una vez que se tiene el repositorio de datos se procede a la selección de los campos a utilizar en el análisis como se constata en la Figura 28.

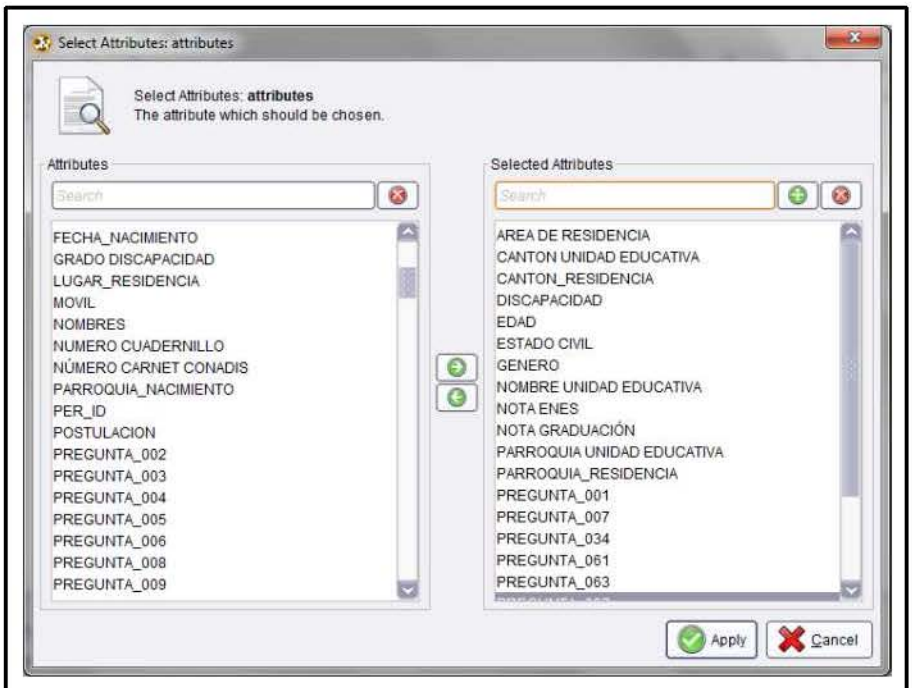


Figura 28: Selección de atributos

4.2 Limpieza de los datos

Para el caso del SNNA como se mencionó anteriormente la recolección inicial de datos se la realiza por medio de formularios web que tienen validaciones en la mayoría de sus campos, sin embargo, de la revisión de los datos se ha encontrado que hay atributos que no tienen valores en un porcentaje del 15%.

4.2.1 Reporte de limpieza de los datos

4.2.1.1 Filtrado de datos

Después del examen de los datos del repositorio, se determina que existen registros que tienen uno o más de sus atributos con campos vacíos. Por ejemplo, de los "Datos se Residencia" se observa que existen valores vacíos para los campos:

- Región Residencia
- Sector Zona
- Régimen UED

Se determina un filtro de los registros que tengan uno o varios de sus atributos vacíos y no se los considera dentro del análisis, por lo tanto aquí se aplica la técnica de exclusión.

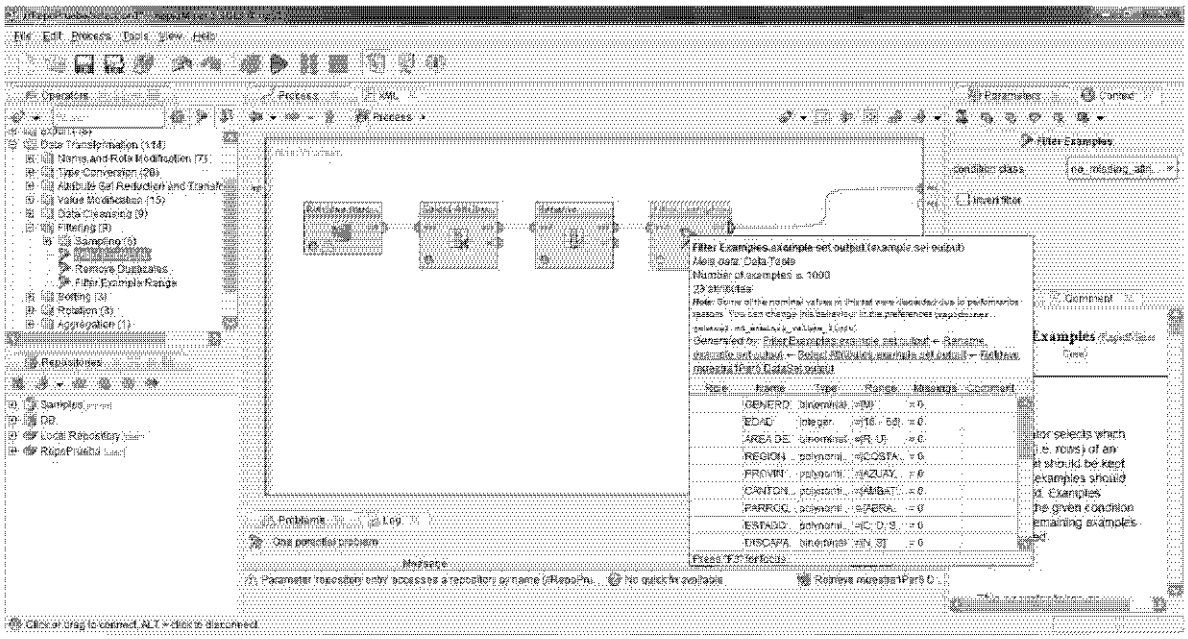


Figura 29: Filtrado de datos

En definitiva, solo se consideran para el análisis los registros que tengan todos sus atributos con datos.

4.2.1.2 Renombre de atributos

Algunos atributos de la encuesta de contexto necesitan ser renombrados para mejorar la descripción de los datos dentro del análisis.

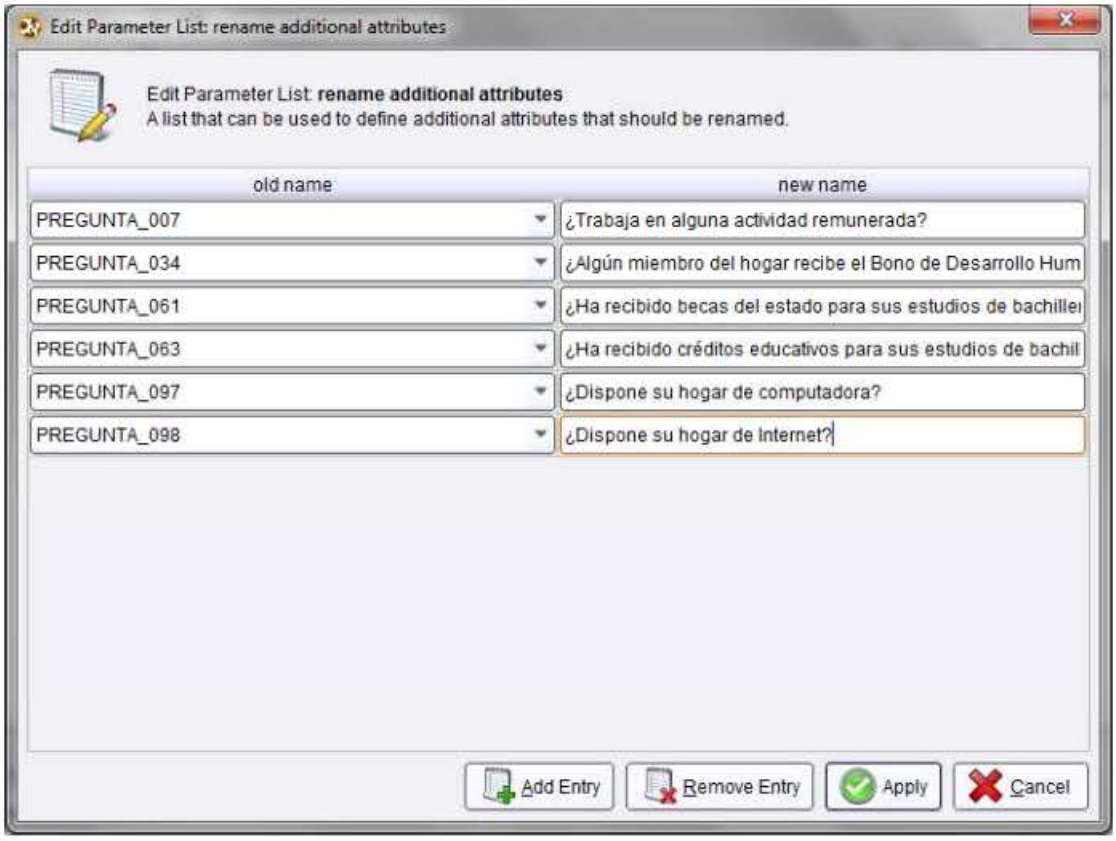


Figura 30: Renombre de atributos

4.3 Construcción de datos

4.3.1 Derivación de atributos

Se necesita definir el estrato socioeconómico al que pertenece el aspirante. Para ello se va a tomar como referencia la metodología utilizada por el Instituto Nacional de Estadística y Censos. (Instituto Nacional de Estadística y Censos (INEC), 2013)

De acuerdo al INEC se definen 5 estratos socioeconómicos en base al puntaje obtenido en diferentes parámetros de evaluación. Los grupos definidos son:

Tabla 6: Grupos socio económicos	
Grupos	Puntaje
A	De 845 a 1000 puntos
B	De 696 a 845 puntos
C+	De 535 a 696 puntos
C-	De 316 a 535 puntos
D	De 0 a 316 puntos

Tomado de: (Instituto Nacional de Estadística y Censos (INEC), 2013)

A continuación se presenta las principales características de los estratos definidos por INEC:

Nivel A:

- Todos los hogares de este nivel cuentan con servicio de internet.
- La mayoría de estos hogares tienen computadora de escritorio y/o portátil
- El Jefe de Hogar tiene un nivel de instrucción superior y un número considerable alcanza estudios de post grado.
- Los jefes de hogar del nivel A se desempeñan como profesionales científicos, intelectuales, miembros del poder ejecutivo, de los cuerpos legislativos, personal del directivo de la Administración Pública y de empresas privadas.

Nivel B:

- La mayoría de los hogares de este nivel cuentan con servicio de internet.
- La mayoría de estos hogares tienen computadora de escritorio y/o portátil

- El Jefe del Hogar tiene un nivel de instrucción superior.
- Una parte importante de los jefes de hogar del nivel B se desempeñan como profesionales científicos, intelectuales, técnicos y profesionales del nivel medio

Nivel C+:

- Una cantidad importante de los hogares de este nivel cuentan con servicio de internet.
- Buena parte de estos hogares tienen computadora de escritorio y/o portátil.
- El Jefe del Hogar tiene un nivel de instrucción de secundaria completa.
- Los jefes de hogar del nivel C+ se desempeñan como trabajadores de los servicios, comerciantes y operadores de instalación de máquinas y montadores.

Nivel C-:

- Una baja cantidad de hogares tienen computadora de escritorio y tienen acceso a Internet.
- El Jefe del Hogar tiene un nivel de instrucción de primaria completa
- Los jefes de hogar del nivel C- se desempeñan como trabajadores de los servicios y comerciantes, operadores de instalación de máquinas y montadores y algunos se encuentran inactivos.

Nivel D:

- Por lo general estos hogares no tienen acceso directo a Internet ni tampoco poseen computadora de escritorio o portátil
- El Jefe del Hogar tiene un nivel de instrucción de primaria completa
- Los jefes de hogar del nivel D se desempeñan como trabajadores no calificados, trabajadores de los servicios, comerciantes, operadores de

instalación de máquinas y montadores y algunos se encuentran inactivos.

De acuerdo al propio INEC, se debe recalcar que esta estratificación no tiene nada que ver ni guarda relación con indicadores de pobreza o desigualdad. Son dos mecanismos, dos objetivos y dos metodologías distintas para clasificar a los hogares.

4.3.2 Generación de registros

No se considera necesario generar nuevos atributos ni registros adicionales para el caso de estudio presente

4.3.3 Resumen del proceso

Dentro del proceso de preparación de datos se han ejecutado las siguientes actividades:

- Definición del repositorio de datos
- Selección de atributos relevantes
- Renombre de atributos
- Filtrado de registros
- Construcción de atributos nuevos

Capítulo 5 Modelamiento de datos

Esta es la principal etapa del proyecto y consiste en elegir la técnica de modelamiento de datos que se va a utilizar sobre el conjunto de datos definido en la etapa anterior. El objetivo es descubrir la relación del conjunto de datos y el atributo que se desea predecir. Las principales actividades en esta etapa son, la selección de la técnica de modelamiento, la generación del plan de pruebas y la construcción del modelo de datos.

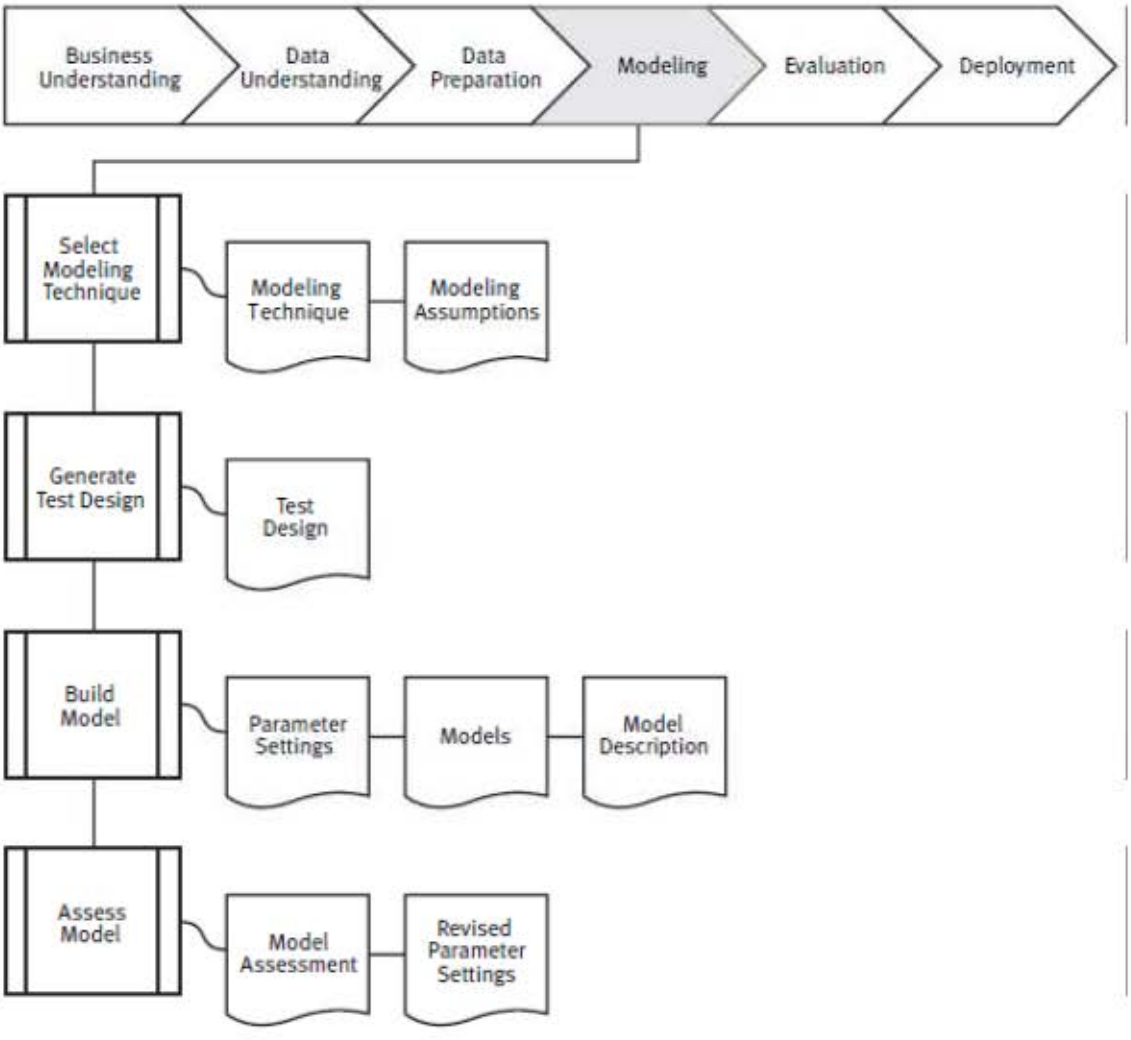


Figura 31: Modelado de datos

Tomado de: (Chapman, y otros, 2000)

5.1. Selección de la técnica de modelado

Las técnicas de clasificación predicen variables discretas en base a otros atributos del conjunto de datos. Las técnicas de clusterización dividen el conjunto de datos en grupos de datos que tienen propiedades similares.

Las técnicas de predicción trabajan sobre variables continuas pero el conjunto de datos definido en el capítulo anterior contiene atributos discretos, por lo que no es factible el uso de este tipo de técnicas.

Las técnicas de asociación buscan la creación de reglas de asociación y el ejemplo más común es el análisis de la cesta de compras. El conjunto de datos definido contiene un único registro por aspirante, por lo que no es posible crear reglas de asociación para el análisis que se pretende realizar.

Por tal motivo las técnicas de modelado de datos a utilizar en el presente trabajo son las técnicas de clasificación y de clusterización. La herramienta utilizada para crear los modelos de datos es Microsoft Analysis Services de la suite de Microsoft SQL Server 2012.

5.1.1. Descripción de las técnicas seleccionadas

Para la técnica de clasificación se van a utilizar los algoritmos de árboles de decisión, naive bayes y redes neuronales. Para la técnica de clusterización se utilizará el algoritmo k-medias. Se intentó probar otros algoritmos pero la estructura de datos definida no permite la aplicación de algunos algoritmos como la regresión lineal pues requiere que los datos de entrada deben ser continuos y no discretos.

5.2. Generación del plan de prueba

El objetivo de este paso es diseñar las pruebas que van a utilizarse para probar la validez de los resultados arrojados por el modelo que luego va a ser implementado.

Para los cuatro modelos de datos que se van a construir la prueba consiste en dividir el total de registros en dos grupos, generados en forma aleatoria: El primer grupo es el conjunto de entrenamiento (training set), que contendrá aproximadamente el 70% del conjunto total; y el segundo grupo es el conjunto de validación (testing set) que será utilizado para validar los grupos detectados por el algoritmo.

5.3. Construcción de los modelos de minería de datos

A continuación se van a crear varios modelos de minería de datos con el objetivo de analizar el atributo "nota de evaluación" (NOTA_EVAL) y luego determinar qué modelo se ajusta más al objetivo del proyecto de minería de datos.

Se va a realizar el análisis de los atributos socioeconómicos de los aspirantes y su relación con la nota del examen ENES. Además, se va a realizar el análisis de los atributos de la unidad educativa de donde proviene el aspirante y su relación con los tres grupos de preguntas que componen el ENES (razonamiento lógico, razonamiento matemático, razonamiento abstracto). Para ello, el primer paso es crear las estructuras de minería de datos que van a ser utilizadas por los diferentes modelos. La tabla 7 muestra todos los atributos definidos en la estructura para el análisis de los atributos socioeconómicos y que serán considerados como entradas de los diferentes algoritmos utilizados. La tabla 8 muestra todos los atributos definidos en la estructura de datos para el análisis de los atributos de las unidades educativas.

Tabla 7: Estructura de datos para atributos socioeconómicos

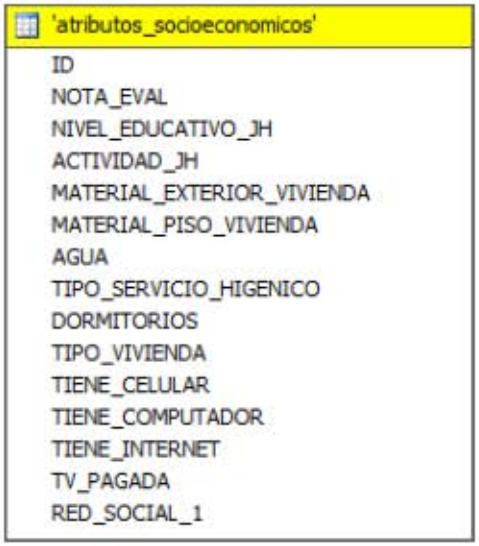
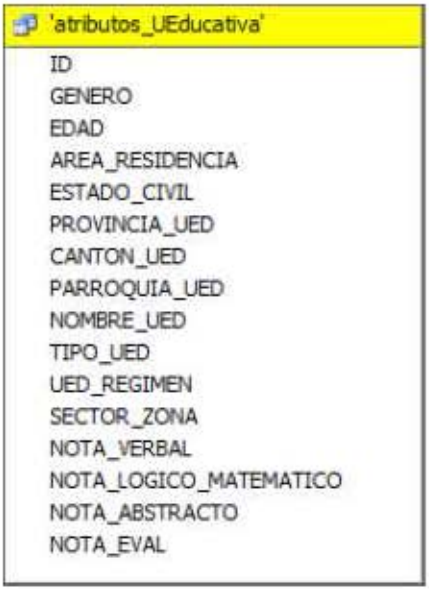


Tabla 8: Estructura de datos para atributos de unidades educativas



5.3.1. Modelo en base a árboles de decisión

Del conjunto de datos se selecciona la columna que se va a analizar (NOTA_EVAL) y las columnas de entrada. Para el análisis, las columnas de entrada son los atributos socio económicos que tiene el aspirante.

La figura 32 muestra los parámetros requeridos por el algoritmo. Se utilizarán los parámetros por definidos por defecto.

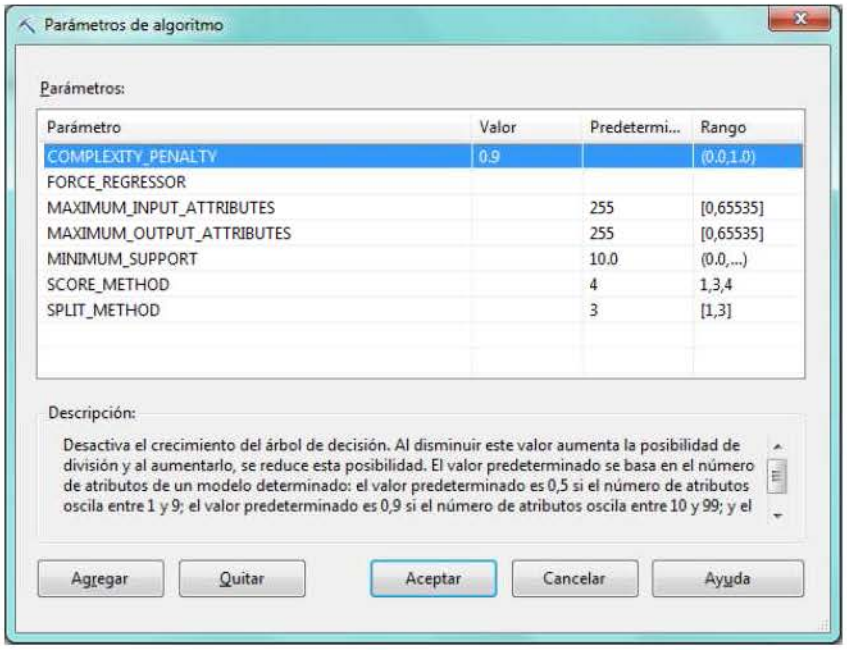


Figura 32: Parámetros del algoritmo árboles de decisión

El conjunto de datos es dividido automáticamente en un conjunto de entrenamiento y en un conjunto de pruebas. El motor de minería de datos usará el conjunto de entrenamiento para entrenar el modelo de minería de datos y el conjunto de pruebas para probar la precisión del modelo.

5.3.2. Modelo en base a Naive Bayes

A continuación se construirá el modelo de clasificación utilizando el algoritmo de Naive Bayes. La figura 33 muestra los parámetros requeridos por este algoritmo. Se utilizarán los parámetros definidos por defecto.

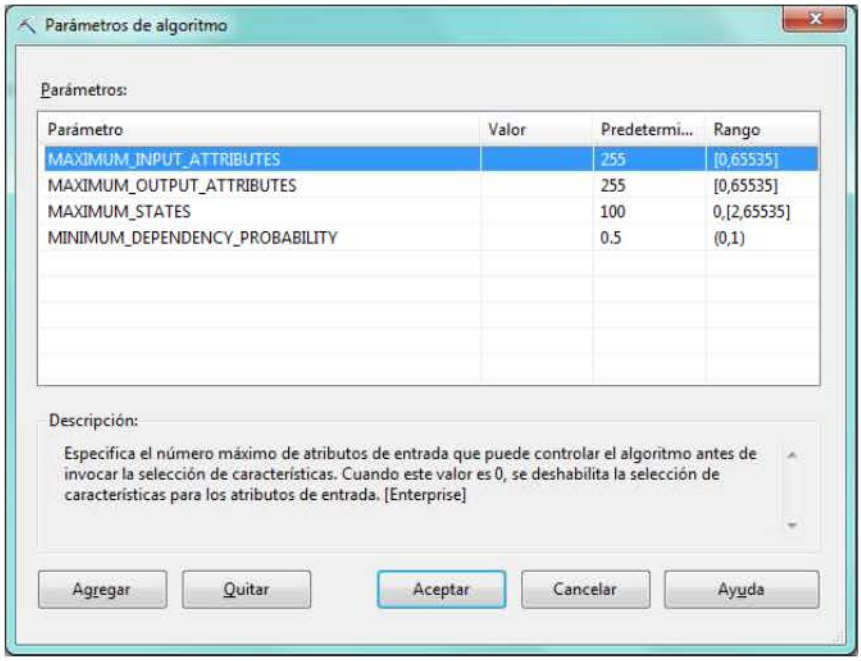


Figura 33: Parámetros del algoritmo Naive Bayes

5.3.3. Modelo en base a Clúster

A continuación se construirá el modelo de clasificación utilizando la técnica de clúster. Se selecciona el algoritmo de agrupación en clústeres k-mediana por ser el más utilizado, sin embargo, el resto de métodos disponibles devuelven resultados similares. La figura 34 muestra los parámetros requeridos por el algoritmo. Se utilizarán los parámetros definidos por defecto.

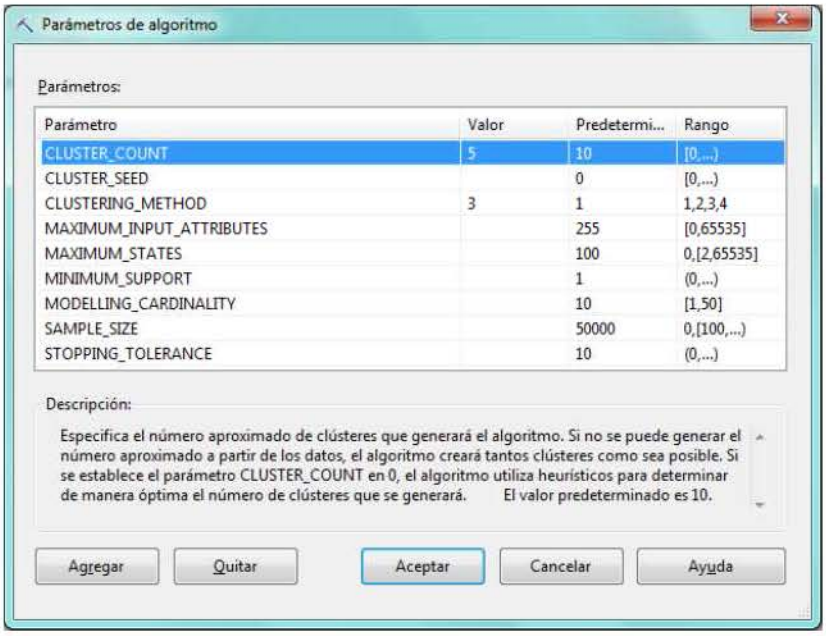


Figura 34: Parámetros del algoritmo de clustering

5.3.4. Modelo en base a red neuronal

A continuación se construirá el modelo de clasificación utilizando la técnica de red neuronal. La figura 35 muestra los parámetros requeridos por el algoritmo. Se utilizarán los parámetros definidos por defecto.

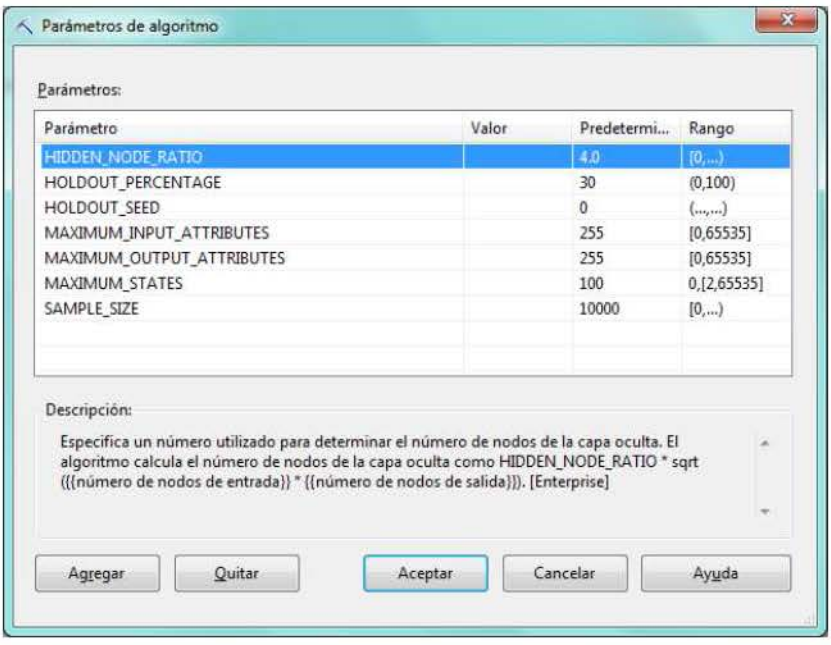


Figura 35: Parámetros del algoritmo red neuronal

5.3.5. Resumen de modelos construidos

A continuación se presenta en la figura 36 el resumen de los modelos creados sobre la estructura de minería de datos definida para el análisis de los atributos socioeconómicos. Todos los modelos tienen definido como variable de análisis el atributo NOTA_EVAL.

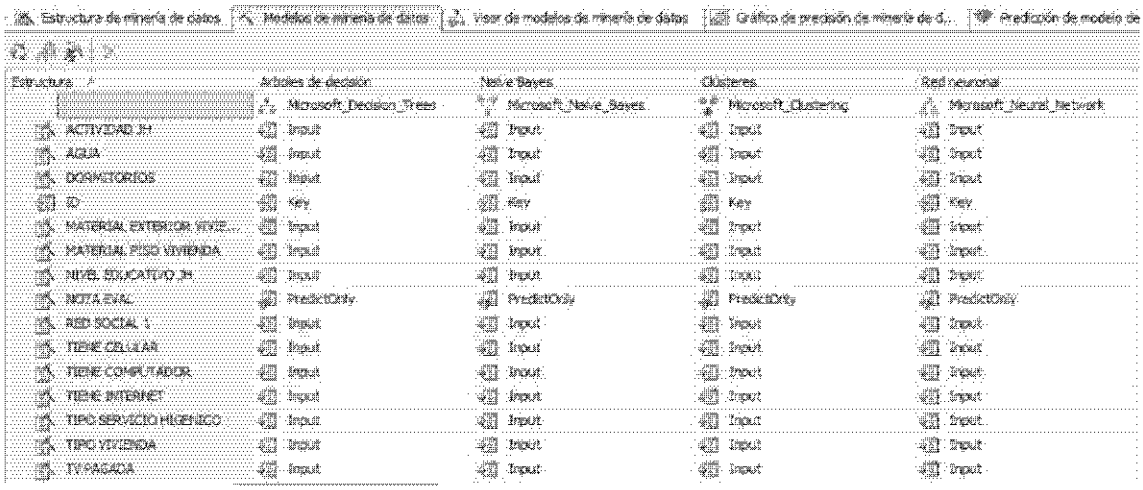


Figura 36: Resumen de modelos generados – nivel socioeconómico

En el siguiente capítulo se evaluará cuál de estos modelos alcanza una mayor probabilidad de predicción y en base a esta técnica se realizará el análisis de la variable "tipo de unidad educativa" (tipo_UED) de la que provienen los aspirantes y su relación con el examen de admisión.

Capítulo 6 Evaluación de resultados

En esta etapa se evalúan los resultados del modelo de datos desde la perspectiva de la organización. Se analiza el nivel en el que el resultado se acerca a la necesidad de la organización y se trata de determinar si por alguna razón el modelo no es eficiente. Esta etapa de evaluación también cubre otros resultados de minería de datos que no necesariamente están relacionados con los objetivos originales pero que podrían revelar información para las direcciones futuras. Las principales actividades en esta etapa son la valoración de resultados y la revisión del proceso.

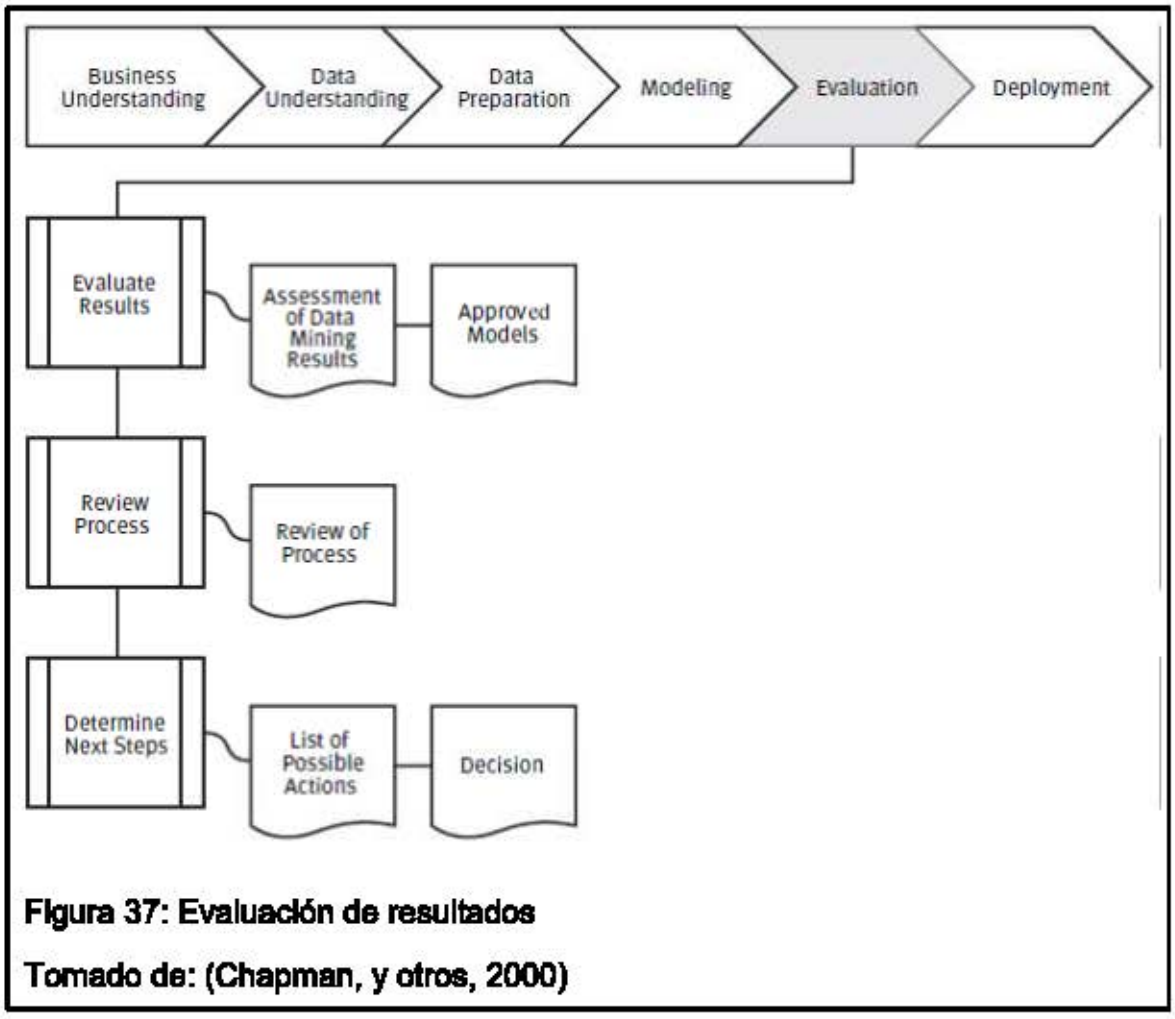


Figura 37: Evaluación de resultados

Tomado de: (Chapman, y otros, 2000)

6.1. Valoración de resultados

Uno de los objetivos de la organización es "garantizar la igualdad de oportunidades, la meritocracia, la transparencia acerca del acceso a la educación superior".

El objetivo del proyecto de data mining definido en la etapa de comprensión del negocio del capítulo 2 fue "Descubrir las relaciones entre los atributos socio-económicos de los aspirantes a la educación superior y los resultados de su examen de admisión"

A continuación se presentan los resultados de los cuatro modelos de datos creados y la comparación de los mismos con respecto a la precisión de resultados. La herramienta utilizada para el análisis de datos es Microsoft Analysis Services.

6.1.1. Resultados del modelo de árboles de decisión

El nodo raíz del árbol de decisión que representa la nota del examen de admisión ENES, ha sido dividido de forma automática en 5 rangos:

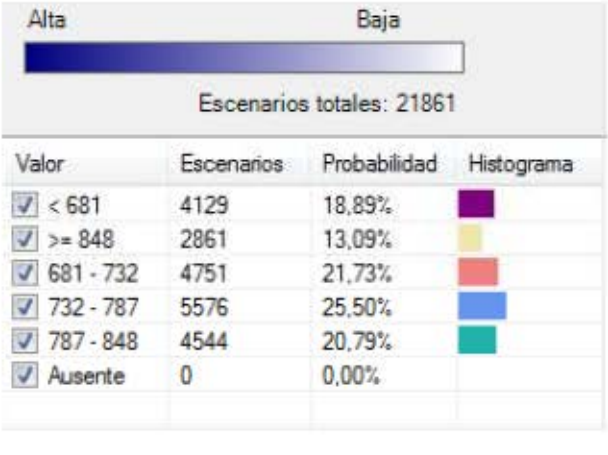


Figura 38: Divisiones del nodo raíz del árbol

La densidad de las hojas del árbol está representada por el color de las mismas. Mientras más oscuro es el color mayor es la densidad de la población de la hoja. El siguiente árbol de decisión resalta los atributos socioeconómicos de los aspirantes que obtuvieron notas mayores o iguales a 848 puntos.

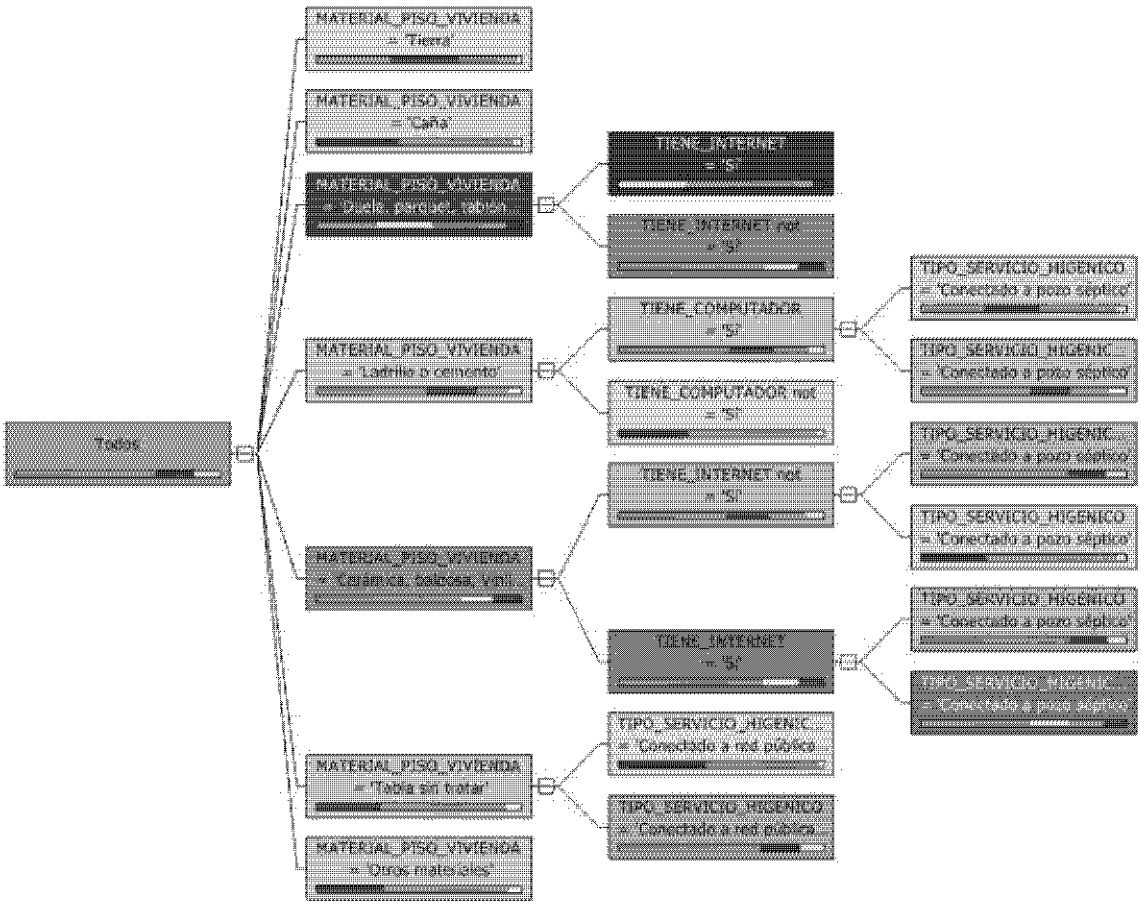


Figura 39: Árboles de decisión – atributos socioeconómicos (notas => 848)

Cada uno de los rangos en el nodo raíz representa un subconjunto de instancias de acuerdo a su probabilidad de ocurrencia.

El subconjunto que mayor probabilidad de ocurrencia presenta (25,50%) es el que tiene una nota de examen entre 732 y 787 puntos. En cambio el subconjunto con puntaje >=848 puntos tiene una probabilidad del 13,09%.
El siguiente es el árbol de decisión para el rango de notas menores a 681 puntos.

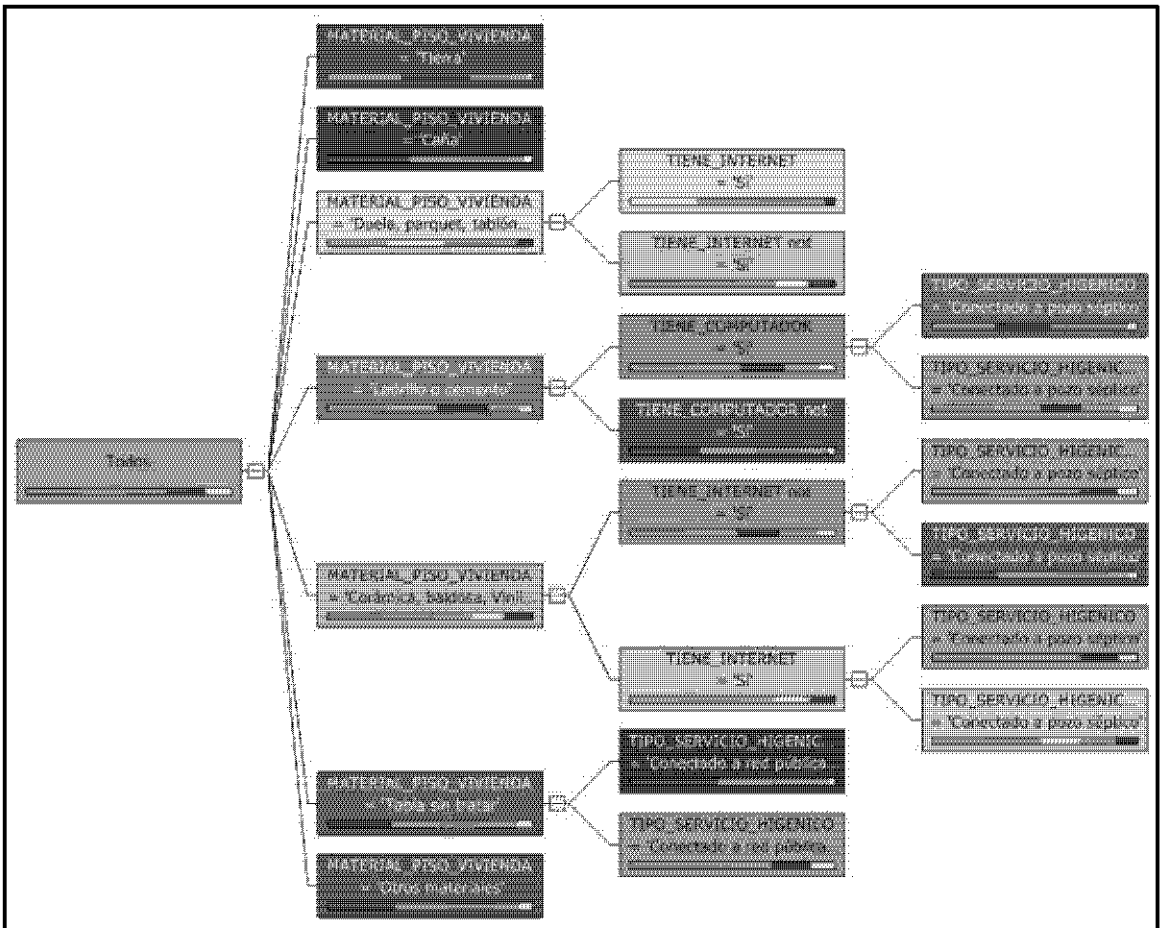


Figura 40: Árboles de decisión – atributos socioeconómicos (notas<= 681)

El siguiente nivel de ramas en el árbol está dado por el atributo de entrada "MATERIAL_PISO_VIVIENDA". Si analizamos el subconjunto de aspirantes con puntaje <=681 puntos se puede observar que las ramas más significativas del árbol están dadas por los aspirantes cuya vivienda es de tierra o caña. Revisando el grupo de los aspirantes con puntaje >=844 puntos se observa que el material del piso de su vivienda es de duela, parquet o tablón. Los siguientes

niveles de ramas en el árbol están dados principalmente por el atributo "TIENE_INTERNET" y "TIENE_COMPUTADOR".

6.1.2. Resultados del modelo de Naive bayes

A continuación se revisan los resultados obtenidos por el modelo de clasificación en base al algoritmo de Naive Bayes. De la misma forma que el modelo anterior la variable a analizar es la nota de evaluación.

Este algoritmo permite clasificar la información de acuerdo a la probabilidad de distribución de cada atributo y diferenciar varios perfiles de información como se muestra en la figura 41.

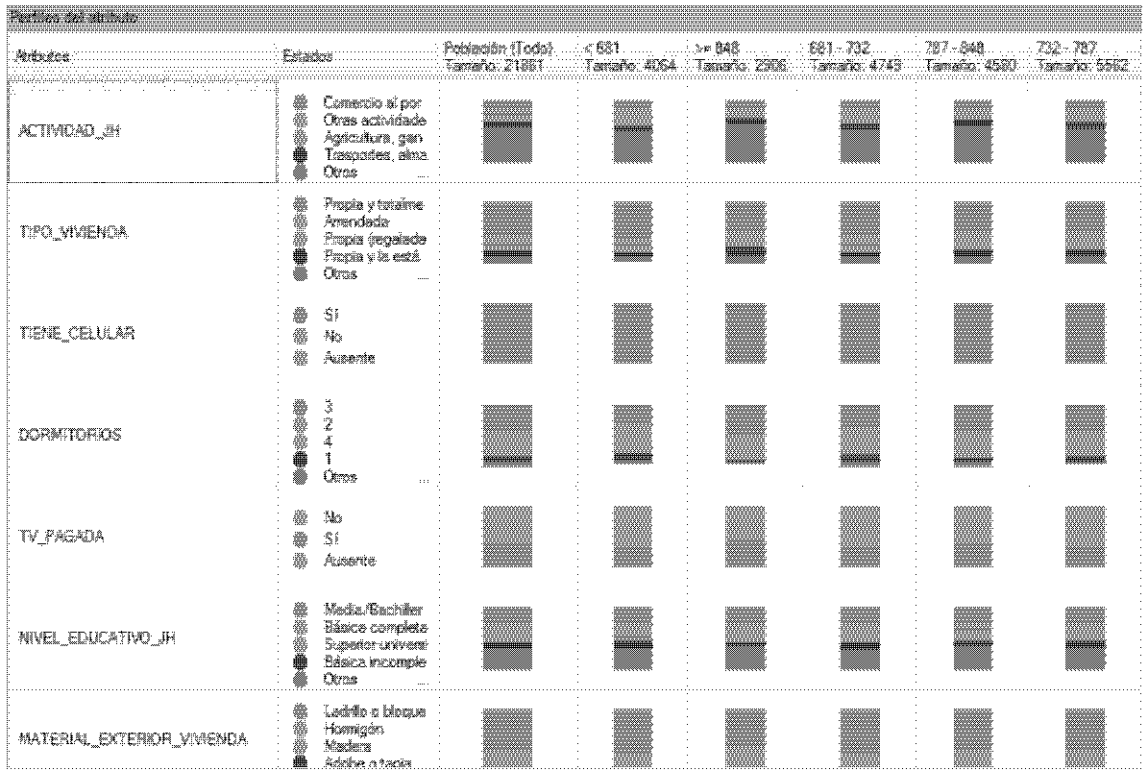


Figura 41: Naive Bayes – atributos socioeconómicos

Se puede observar que se han conformado 5 rangos de notas (iguales al modelo de árboles de decisión). Cada uno de estos rangos contiene información de los atributos de entrada. Los rangos más interesantes de

análisis son las notas mayores o iguales a 848 y las notas menores a 681 puntos.

El modelo permite hacer un contraste entre dos de los rangos establecidos. En la figura 42 se observa el contraste entre los rangos de interés.

Puntuaciones de distinción para < 681 y >= 848

Atributos	Valores	Favorece < 681	Favorece >= 848
TIENE_INTERNET	Sí		
TIENE_INTERNET	No		
TIENE_COMPUTADOR	Sí		
TIENE_COMPUTADOR	No		
TIPO_SERVICIO_HIGENICO	Conectado a red pública de alcantaril...		
MATERIAL_PISO_VIVIENDA	Ladrillo o cemento		
TIPO_SERVICIO_HIGENICO	Conectado a pozo séptico		
MATERIAL_PISO_VIVIENDA	Duela, parquet, tablón o piso flotante		
AGUA	De la red pública		
NIVEL_EDUCATIVO_JH	Superior universitario completa (titula...		
MATERIAL_EXTERIOR_VIVIENDA	Hormigón		
ACTIVIDAD_JH	Agricultura, ganadería, silvicultura y c...		
MATERIAL_PISO_VIVIENDA	Cerámica, baldosa, Vinil o mármol		
AGUA	De Pozo		
NIVEL_EDUCATIVO_JH	Básica completa		
DORMITORIOS	3		
NIVEL_EDUCATIVO_JH	Post grado completo (titulado)		
AGUA	Otra fuente por tubería		
TIPO_VIVIENDA	Propia y la está pagando		
MATERIAL_PISO_VIVIENDA	Tabla sin tratar		
NIVEL_EDUCATIVO_JH	Superior universitaria incompleta (incl...		
MATERIAL_EXTERIOR_VIVIENDA	Madera		
TV_PAGADA	No		
TV_PAGADA	Sí		
MATERIAL_EXTERIOR_VIVIENDA	Ladrillo o bloque		
TIPO_SERVICIO_HIGENICO	Conectado a pozo ciego		
DORMITORIOS	1		
NIVEL_EDUCATIVO_JH	Básica incompleta		
NIVEL_EDUCATIVO_JH	Media/Bachillerato incompleta		

Figura 42: Naive Bayes - Distinción de rangos

Como se puede observar, las condiciones socio-económicas de estos dos grupos de aspirantes son muy definidas. Mientras los aspirantes que obtienen notas superiores a los 848 puntos cuentan en su hogar con servicio de Internet, computador, servicio de agua potable, piso de madera o cerámica, vivienda propia y el nivel de educación del jefe de hogar está entre superior universitario y postgrado. En cambio, para los aspirantes cuya nota de evaluación es menor

a 681 puntos, su hogar carece de servicio de Internet, no poseen computador, el piso de su vivienda es de cemento, el servicio de agua proviene de otras fuentes que no es la red normal de agua potable y el nivel de educación del jefe de hogar esta entre básico incompleto y básica completa.

6.1.3. Resultados del modelo de Clustering

A continuación se revisan los resultados obtenidos por el modelo de clustering en base al algoritmo de k-medias. De la misma forma que el modelo anterior la variable a analizar es la nota de evaluación.

Este algoritmo permite la agrupación de objetos homogéneos entre sí y heterogéneos en relación con otros grupos como se muestra en la figura 43.

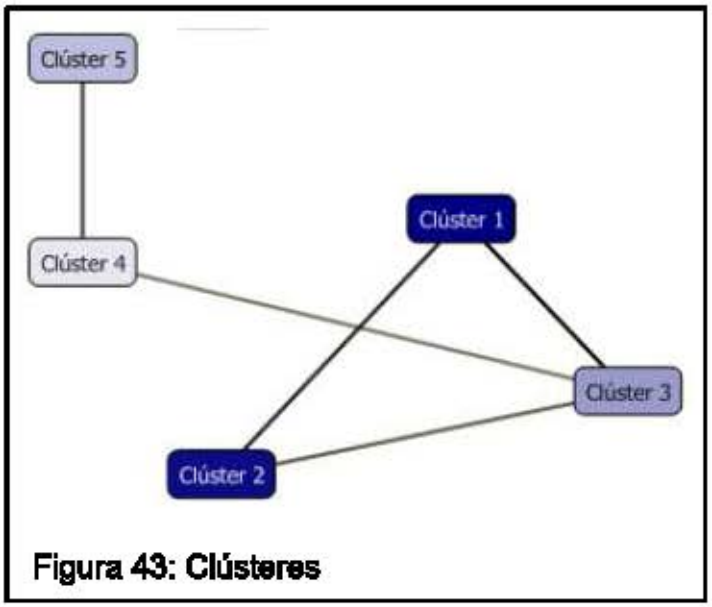


Figura 43: Clústeres

De los resultados obtenidos vemos que el algoritmo generó 5 rangos de notas que difieren de los dos modelos vistos anteriormente. Si analizamos el rango de notas mayor o igual a 856, los clústeres más significativos son el clúster 1 y 2. Las líneas indican la vinculación entre los diferentes clústeres.

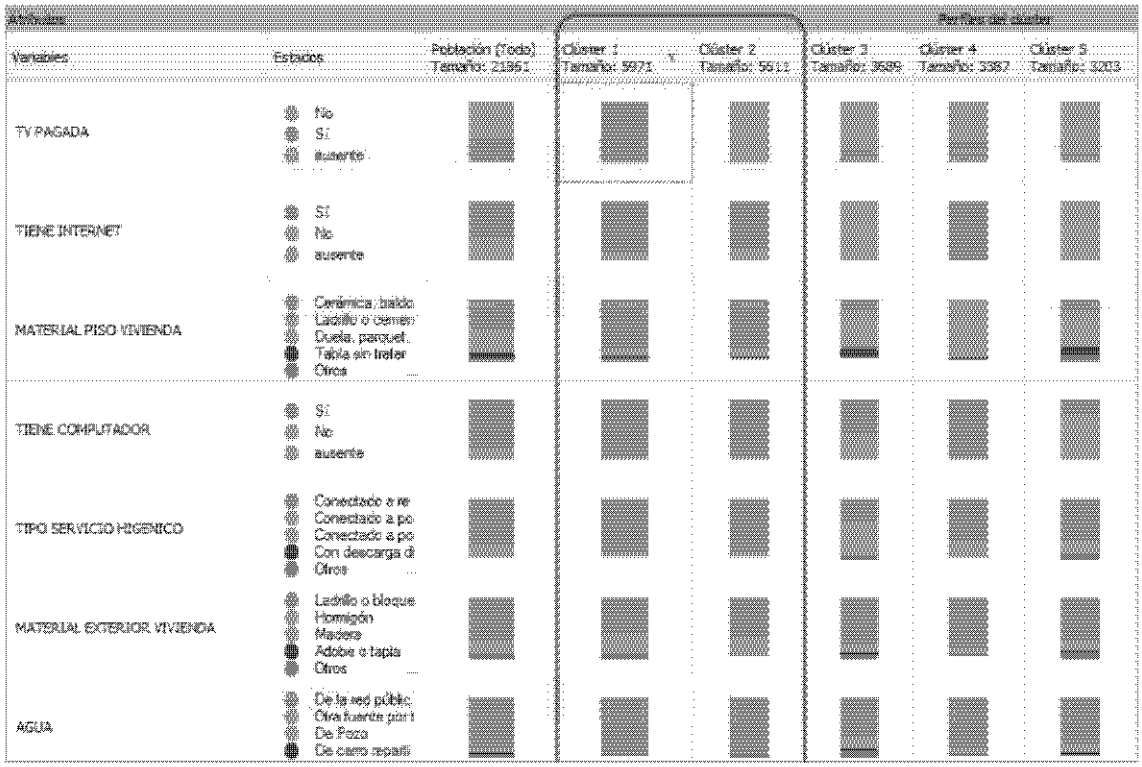


Figura 44: Detalle de Clústeres

Observando las características de los clústeres 1 y 2 se puede verificar que la mayor cantidad de aspirantes cuya nota de evaluación supera los 856 puntos se encuentran en estos grupos. Su hogar cuenta con servicio de televisión pagada, servicio de Internet, computador, etc., es decir, de acuerdo a (Instituto Nacional de Estadística y Censos (INEC), 2013) poseen las características socio-económicas de los estratos A o B.

6.1.4. Resultados del modelo de red neuronal

A continuación se revisan los resultados obtenidos por el modelo de red neuronal. De la misma forma que el modelo anterior la variable a analizar es la nota de evaluación.

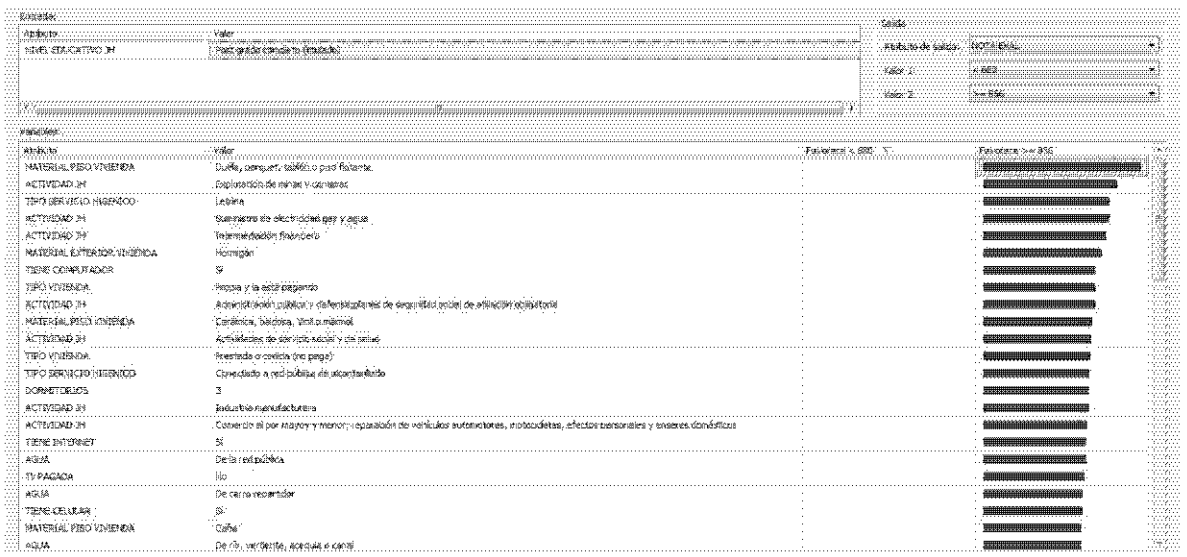


Figura 45: Resultado del algoritmo red neuronal

La figura 45 muestra el resultado de analizar dos valores de atributo a predecir. Si por ejemplo elegimos como atributos de entrada al "nivel educativo del jefe de hogar" con el valor de "Postgrado completo" se puede observar que el valor del resto de atributos favorece a los aspirantes cuyo hogar posee como material del piso de su vivienda a madera o cerámica, y cuenta con servicio de Internet y computador.

6.1.5. Comparación de los modelos construidos

Una vez que se han construido y ejecutado los cuatro modelos revisados, se procede a examinar el gráfico de precisión de los modelos de minería de datos.

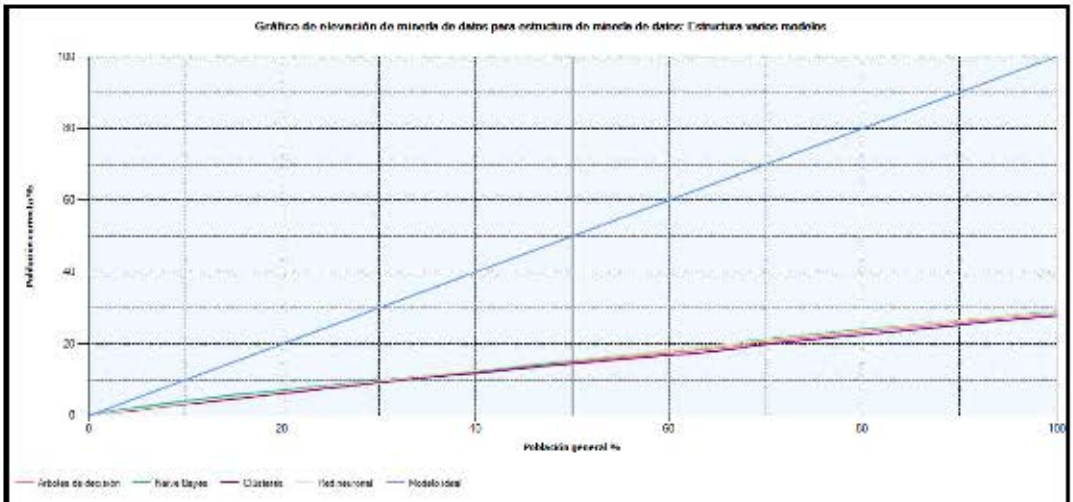


Figura 46: Precisión de los modelos de minería de datos

Serie, Modelo	Puntuación	Población correcta	Probabilidad de predicción
Arboles de decisión	0,30	14,98%	27,76%
Naive Bayes	0,31	15,41%	34,89%
Clústeres	0,29	14,54%	30,89%
Red neuronal	0,30	15,64%	31,22%
Modelo ideal		50,00%	

Figura 47: Probabilidad de predicción

El resultado de la comparación muestra que el modelo de mayor probabilidad de predicción es el basado en el algoritmo de naive bayes. Sin embargo, hay que resaltar que los resultados observados en los diferentes algoritmos son muy similares, lo que permite tener certeza a la hora de definir los resultados del proyecto de minería de datos.

6.1.6. Aplicación del modelo seleccionado

Una vez que se ha definido cuál es el modelo más adecuado para continuar con el proceso de minería de datos. Se procederá al análisis de la variable "tipo de unidad educativa" (TIPO_UED) de la que provienen los aspirantes y su relación con el examen de admisión.

En base a la estructura de datos definida en la sección 5.3, se construye el modelo de minería de datos para el análisis de la variable TIPO_UED. La figura 48 muestra el modelo construido en base al algoritmo naive bayes.

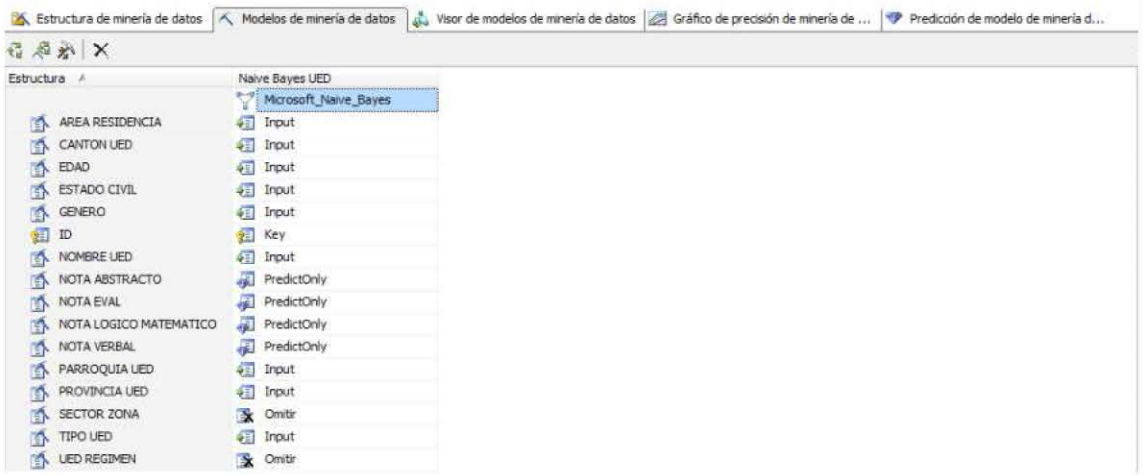


Figura 48: Modelo para variable TIPO_UED

La tabla 9 muestra el conjunto de datos sobre el que trabajará este modelo y muestra la cantidad de unidades educativas por tipo y la cantidad de aspirantes por tipo de unidad educativa que obtuvieron cupo en la universidad.

Tabla 9: Resumen de datos - UED

Tipo de Unidad Educativa (UED)	Cantidad de UED	Aspirantes x UED que obtuvieron cupo
Fiscal	1.160	28.379
Particulares	1.072	12.673
Fisco misional	178	2.568
Municipales	35	438
TOTAL	2.445	44.058

De los datos presentados se observa que a nivel país hay una relación 1:33 entre unidades educativas municipales y fiscales, es decir, por cada UED municipal hay 33 UED fiscales.

De este conjunto de datos, el 30% de los mismos se utiliza internamente por la herramienta para probar el modelo de minería de datos y el restante 70% de

los datos se utiliza para entrenar el modelo. Por ello, el universo de instancias que utiliza el modelo es de 30.840 aspirantes elegidos de forma aleatoria.

En la figura 49 se muestra el resultado de procesar el modelo de datos creado para el análisis de las UED. El nodo en color verde muestra la variable a predecir y los nodos en color amarillo muestran las variables más significativas para la predicción. Las líneas resaltadas representan los vínculos más fuertes entre las variables del modelo.

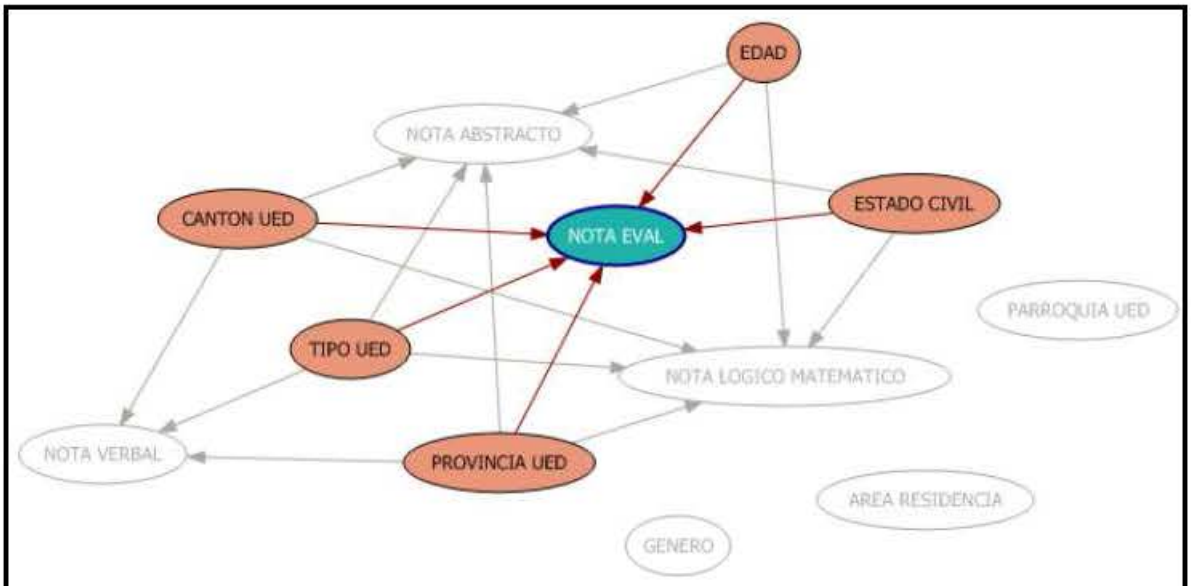


Figura 49: Red de dependencias para NOTA_EVAL

En este caso vemos que la variable "NOTAL EVAL" tiene vínculos más fuertes con la provincia, el cantón y el tipo de unidad educativa, además con la edad y el estado civil del aspirante. La nota del examen tiene tres componentes que son conformados por las preguntas que miden la aptitud verbal, la aptitud abstracta y la aptitud lógica matemática.

La figura 50 muestra que la nota de la aptitud verbal tiene mayor vinculación con el tipo de unidad educativa, la provincia y el cantón al que pertenece.

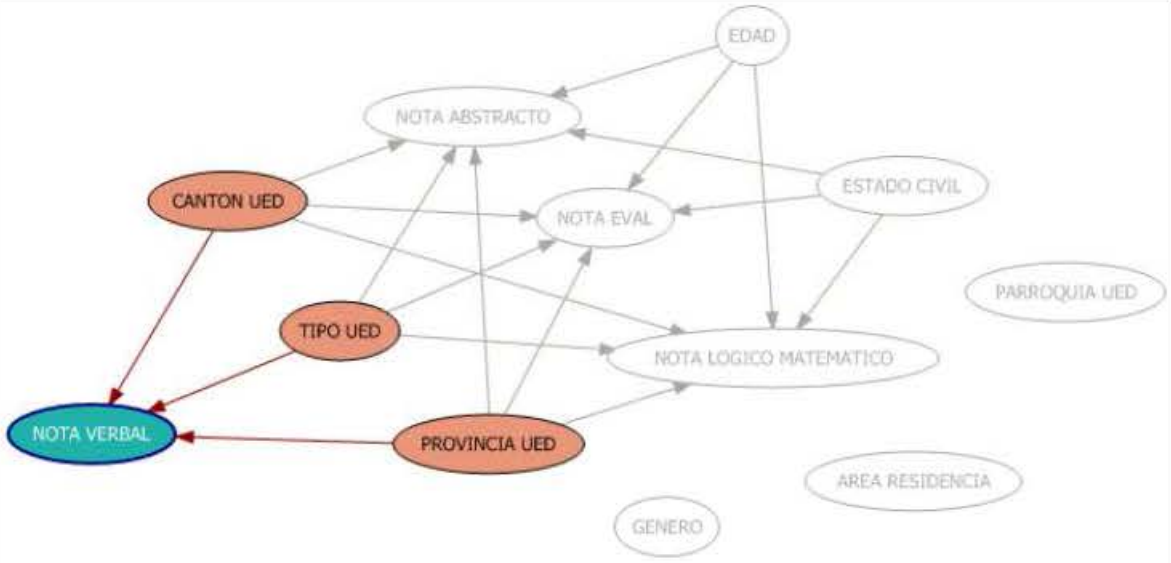


Figura 50: Red de dependencias para NOTA_VERBAL

La figura 51 muestra que la nota de la aptitud abstracta tiene mayor vinculación con el tipo de unidad educativa, la provincia y el cantón al que pertenece, además la edad y el estado civil del aspirante.

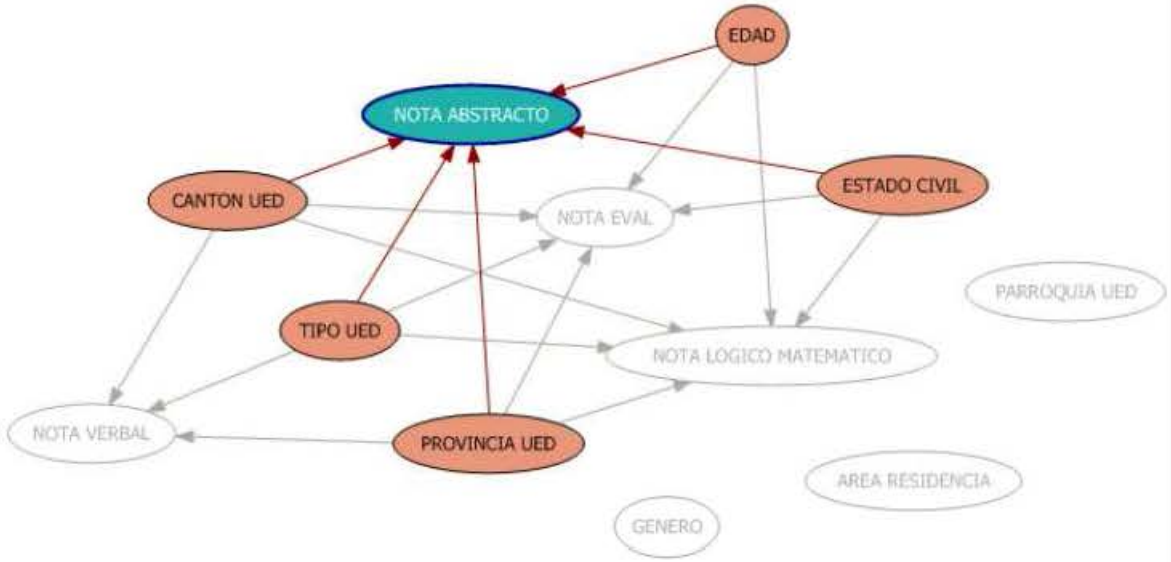


Figura 51: Red de dependencias para NOTA_ABSTRACTO

La figura 52 muestra que la nota de la aptitud lógica-matemática tiene mayor vinculación con el tipo de unidad educativa, la provincia y el cantón al que pertenece, además la edad y el estado civil del aspirante.

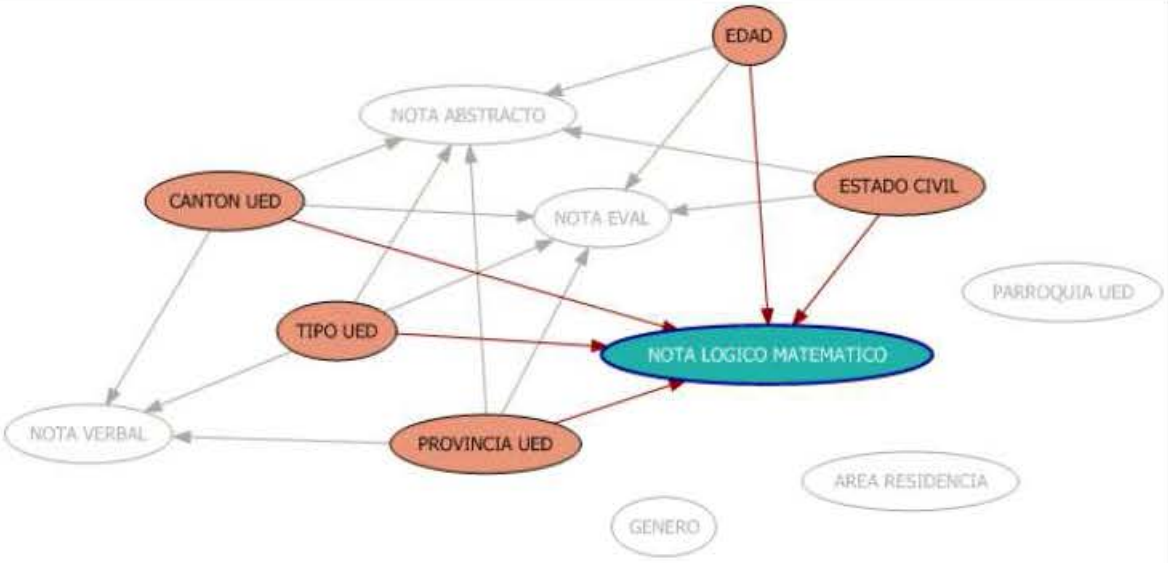


Figura 52: Red de dependencias para NOTA_LOGICO MATEMATICO

En la tabla 10 se resumen los porcentajes de aspirantes de acuerdo a los rangos de notas creados automáticamente por el modelo.

Tabla 10: Rangos de valores para NOTA_EVAL		
Rango de notas	Cantidad de aspirantes	Porcentaje
< 684	6810	22%
684 <= x < 734	7015	23%
734 <= x < 793	7807	25%
793 <= x < 852	5782	19%
>=852	3426	11%
TOTAL	30840	100%

En la figura 53 se muestran los grupos de aspirantes de acuerdo a su nota global en el examen de admisión y los atributos de la unidad educativa de la que provienen.

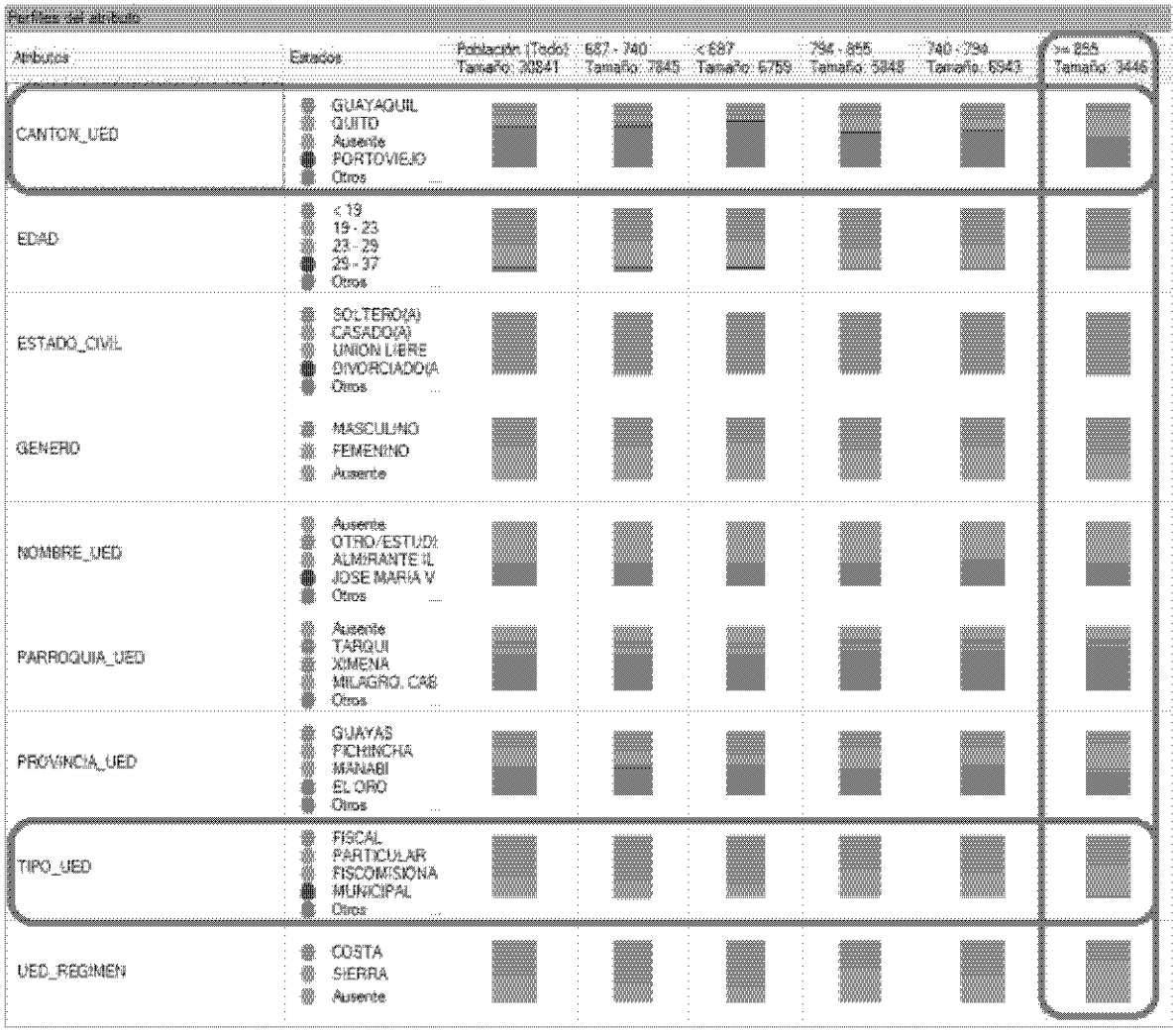


Figura 53: Naive Bayes – atributos UED

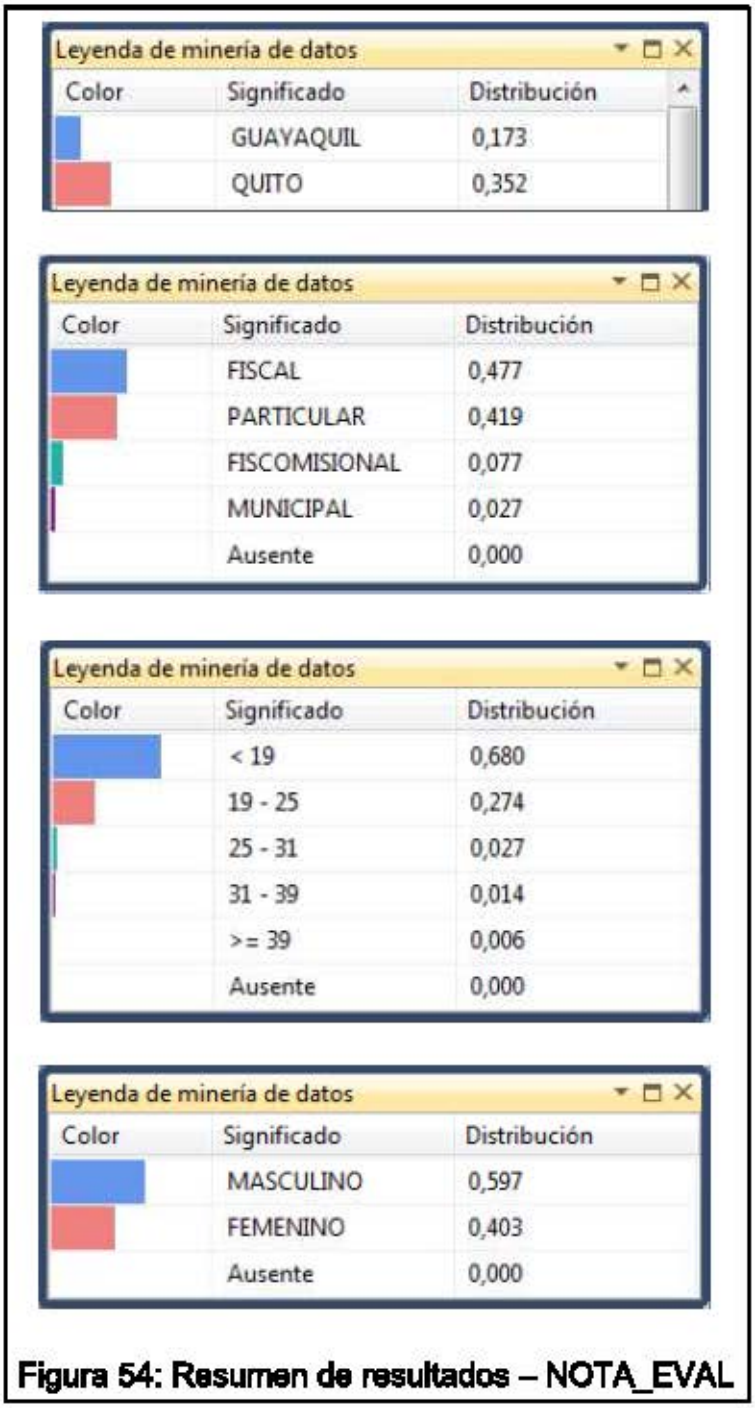


Figura 54: Resumen de resultados – NOTA_EVAL

En la figura 54 se resalta que de los 3.446 aspirantes con notas altas (>=852 puntos) en el examen de admisión, el 35% provienen de unidades educativas de la ciudad de Quito, seguidos por los aspirantes de la ciudad de Guayaquil con un 27%, el 47% vienen de colegios fiscales y el 41% vienen de colegios particulares. El 68% tiene menos de 19 años de edad, el 59% es de género masculino y el 40% femenino. El 96% son solteros.

Esta es la tendencia que se presenta para los 3 componentes del examen de admisión (aptitud abstracta, lógica-matemática y verbal).

La figura 55 muestra que el número de aspirantes con notas mayores a 831 puntos en aptitud abstracta es de 7.757. El 25% provienen de Quito, seguidos por los aspirantes de la ciudad de Guayaquil con un 21%. Además, provienen de colegios fiscales en un 54% y de particulares en un 36%

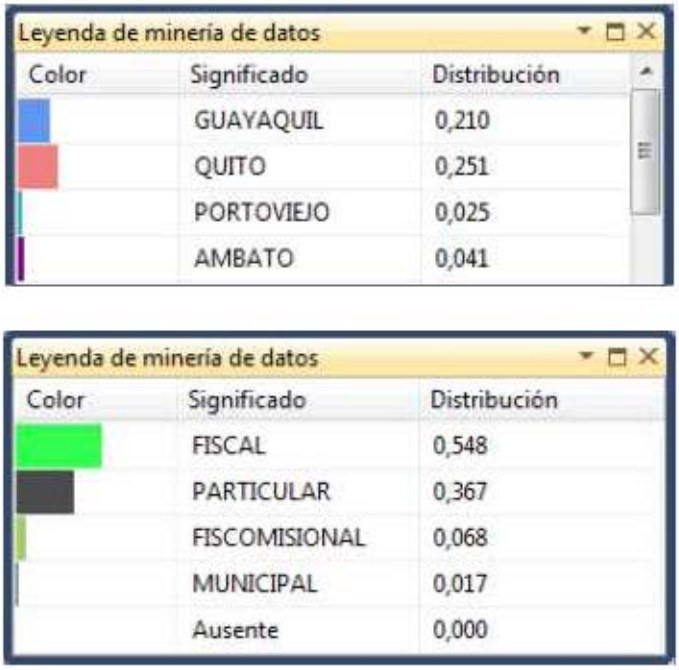


Figura 55: Resumen de resultados – NOTA_ABSTRACTO

En la figura 56 se observa que los aspirantes con mejor nota en aptitud lógica-matemática provienen de Quito y representan el 30%, seguido por los aspirantes de la ciudad de Guayaquil con un 19%. Además, provienen de colegios fiscales en un 51% y de particulares en un 38%.

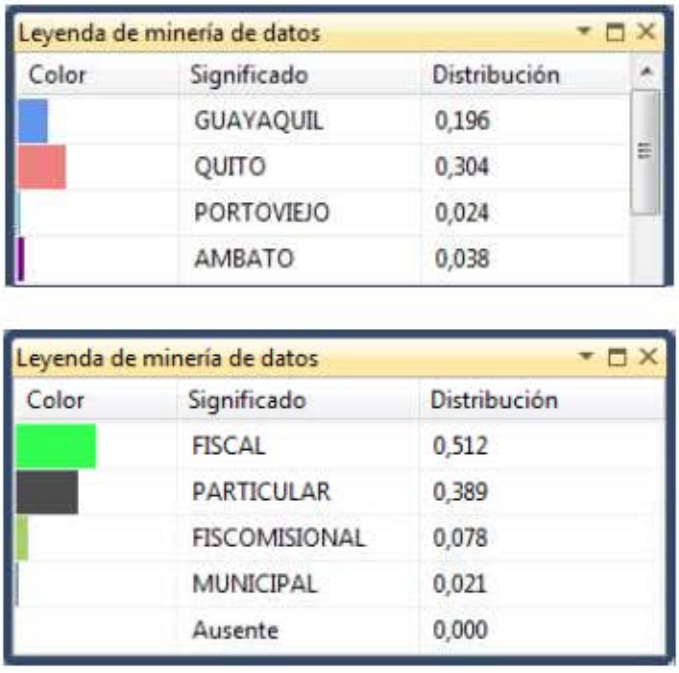


Figura 56: Resumen de resultados – NOTA_LOGICO_MATEMATICO

La figura 57 muestra que los aspirantes con mejor nota en aptitud verbal provienen de Quito y representan el 27%, seguido por los aspirantes de la ciudad de Guayaquil con un 21%. Además, provienen de colegios fiscales en un 48% y de particulares en un 41%.

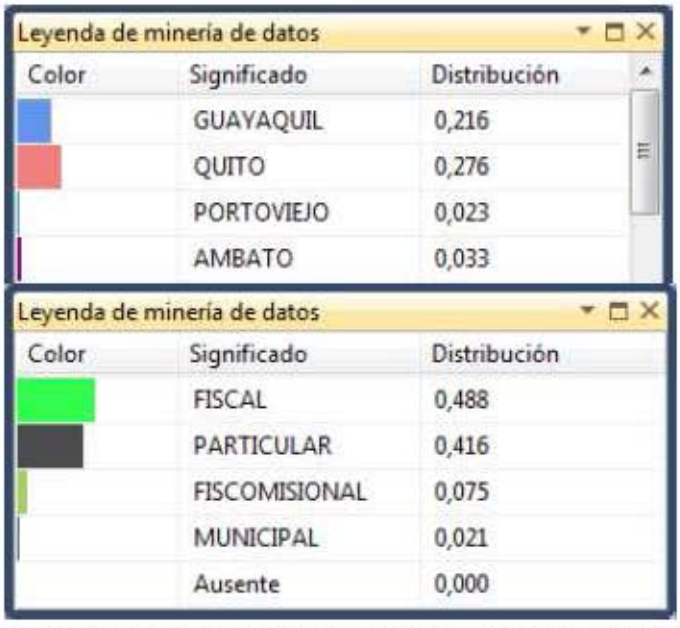


Figura 57: Resumen de resultados – NOTA_VERBAL

6.2. Revisión del proceso

Durante la ejecución del presente proyecto se ha utilizado como referencia las actividades definidas dentro de la metodología CRISP-DM. En el caso particular de este proyecto no fue necesario ejecutar absolutamente todas las tareas definidas dentro de la metodología sino que se consideró trabajar únicamente en aquellas que daban aporte al proyecto. La mayor carga de trabajo se concentró en la fase de preparación de los datos.

A pesar de que el conjunto de atributos definido para este proyecto ha sido seleccionado minuciosamente de una base de datos que contiene más de 300 campos, es posible que se hayan omitido de forma no intencional algunos atributos importantes que puedan aportar información útil para este análisis.

Para el cálculo de los estratos socio-económicos se tomó como referencia la metodología presentada por el INEC, sin embargo, no todos los dominios definidos en la metodología pudieron aplicarse por falta de tal información. Sin embargo, se considera que el cálculo realizado se ajusta a la realidad.

Es recomendable que se ejecuten nuevas iteraciones en cada uno de las etapas para seguir afinando los resultados obtenidos

Capítulo 7 Despliegue de resultados

En esta etapa se toman los resultados de la etapa anterior y se define un plan de despliegue que permita aplicar estos resultados en la organización.

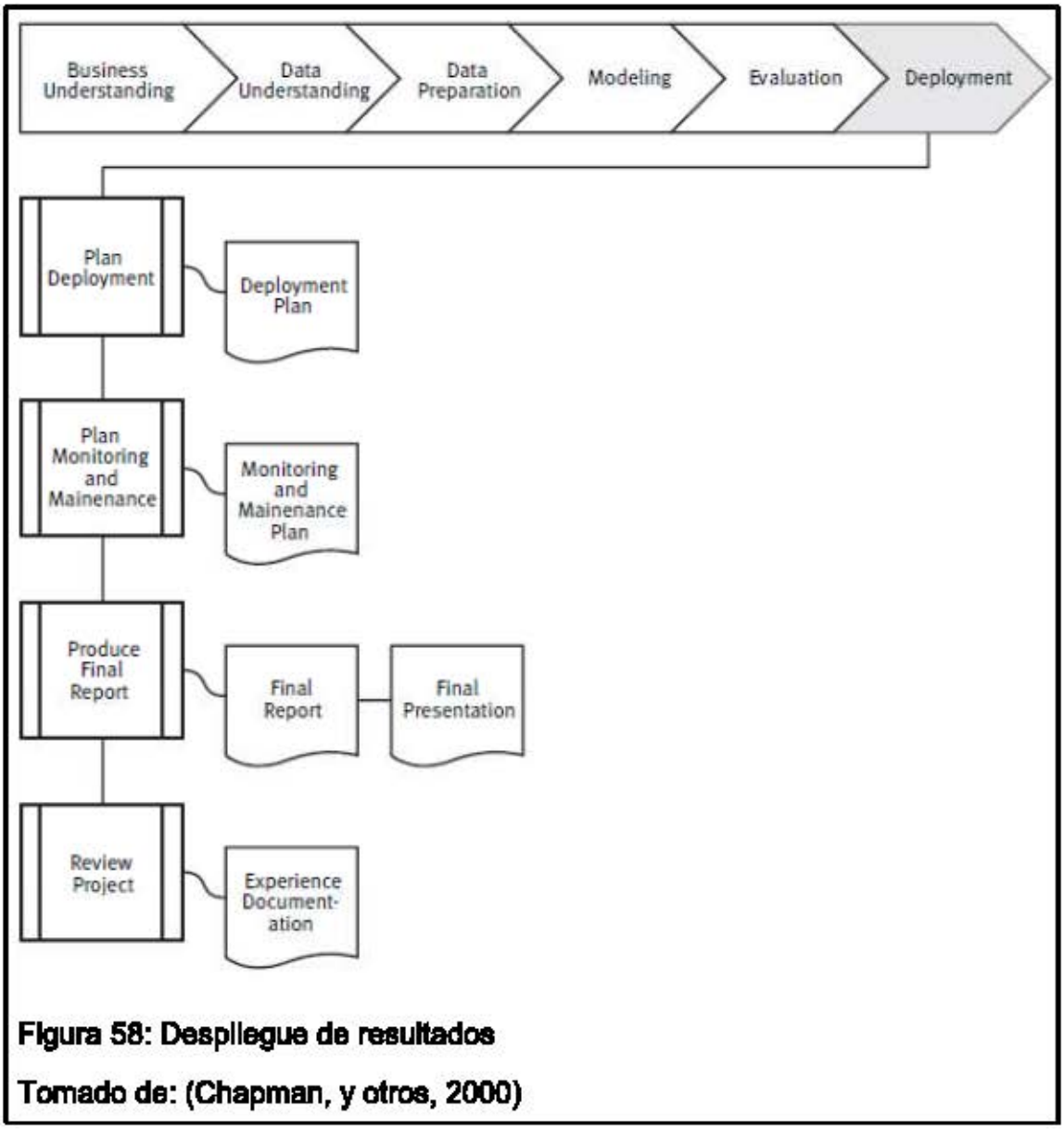


Figura 58: Despliegue de resultados

Tomado de: (Chapman, y otros, 2000)

7.1. Plan de despliegue

Para el presente proyecto el plan de despliegue de resultados consiste en la presentación de los resultados del presente proyecto al Director Ejecutivo del CTT-ESPE-CECAI (o a su delegado) en donde se presentará la importancia

para la organización de trabajar en este tipo de proyectos. Luego, se deberán esperar las debidas autorizaciones para poder iniciar un proyecto de análisis de datos cuyo entregable final sea un proceso que pueda ser llevado de forma continua dentro de la organización.

Este proyecto deberá contemplar el análisis de herramientas de minería de datos para determinar si es más conveniente para la Institución trabajar con herramientas open source o propietarias. También, es necesario considerar la incorporación de personal técnico con perfil específico de analista de información. Además se deberá considerar la adquisición de nuevo hardware específico para estas actividades de análisis ya que por el volumen de información que genera el proyecto hay tareas que requiere gran cantidad de recursos de hardware como por ejemplo la limpieza de datos.

7.2. Plan de monitoreo y mantenimiento

Debido al tipo de proyecto de minería de datos, se considera que no es necesario definir un plan de monitoreo y mantenimiento.

Conclusiones y recomendaciones

La metodología CRISP-DM seleccionada para el desarrollo del proyecto es muy adecuada debido al nivel de detalle que ofrece en cada una de sus tareas. En el caso particular de este proyecto no fue necesario ejecutar todas las tareas definidas dentro de la metodología sino que se consideró trabajar únicamente en aquellas que daban aporte al proyecto. La mayor carga de trabajo se concentró en la fase de preparación de los datos, sobretodo en la definición del estrato social al que pertenece cada aspirante.

Con relación a la técnica de aprendizaje automático utilizada se observa que Naive Bayes es la más adecuada ya que permite clasificar a los aspirantes de acuerdo a sus atributos socio-económicos, con una mayor probabilidad de predicción. Este análisis fue comparado con la metodología utilizada por el Instituto Nacional de Estadística y Censos para la estratificación socio-económica y se encontró que no existe diferencia significativa, por lo tanto los resultados de clusterización obtenidos en el proceso de minería de datos se consideran válidos.

En base a los resultados obtenidos de la minería de datos se observa que los aspirantes que se presentan a rendir en examen de evaluación muestran diferentes niveles de preparación académica. Esta diferencia de preparación puede deberse a varios factores como por ejemplo:

1. Grupo socio-económico al que pertenece el aspirante: del análisis de los datos se concluye que los aspirantes que pertenecen a los grupos socio-económicos A y B (descritos en la sección 4.3) son los que obtienen el mayor puntaje en el examen de evaluación y por ende son los que obtienen cupo en las carreras de su elección. La preparación académica del jefe de hogar es un factor importante para definir el rendimiento académico de sus hijos. De los resultados se ha demostrado que los aspirantes que obtuvieron notas altas en el examen de admisión

pertenecen a hogares donde el jefe de hogar tiene educación de tercer nivel y/o postgrados.

2. Acceso a herramientas tecnológicas: de los grupos analizados se observa que los aspirantes que obtienen notas mayores a 850 puntos en el ENES tienen en su hogar un computador con servicio de Internet (77%). El acceso a las tecnologías de información es uno de los factores importantes que influyen en el rendimiento académico de los aspirantes. El Internet sin duda alguna es la mayor Referencia de información y consultas a nivel académico para los aspirantes.

3. Tipo de unidad educativa donde estudió el aspirante. La observación de los resultados muestra que los aspirantes que obtienen notas mayores a 850 puntos en el ENES, provienen de unidades educativas fiscales (48%) y particulares (41%). Las unidades educativas municipales y fisco misionales representan el grupo más pequeño de aspirantes con notas altas.

4. Sectores demográficos. Las ciudades que concentran aspirantes con notas mayores a 850 puntos en el ENES son Quito (35%) y Guayaquil (17%), que son las ciudades de mayor desarrollo a nivel país.

Debido a esto se puede concluir que, en la actualidad, no todos los aspirantes llegan en igualdad de condiciones a rendir el examen de admisión a la educación superior, por lo que las políticas de estado deberían reforzarse en la educación general básica.

Definitivamente se recomienda implementar dentro de la organización este tipo de proyectos y que se puedan tener procesos continuos de análisis de información para una acertada toma de decisiones basada en información confiable.

Referencias

BayesServer. (n.d.). www.bayesserver.com. Retrieved from Classification: http://www.bayesserver.com/Techniques/Classification.aspx

Bishop, C. (2006). Pattern Recognition and Machine Learning. New York: Springer.

Chapman, P. (., Clinton, J. (., Kerber, R. (., Khabaza, T. (., Reinartz, T. (., Shearer, C. (., & Wirth, R. (. (2000). CRISP-DM 1.0, Step-by-step data mining guide. The CRISP-DM consortium.

Fayyad, U., Piatesky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence.

Gartner, Inc. (s.f.). IT Glossary. Obtenido de http://www.gartner.com/it-glossary

Instituto Nacional de Estadística y Censos (INEC). (2013). Encuesta de Estratificación del Nivel Socioeconómico. Obtenido de http://www.inec.gob.ec/estadisticas/?option=com_content&view=article&id=112&Itemid=90&

KDnuggets. (s.f.). Algorithms for Data Mining (Nov 2011). Obtenido de http://www.kdnuggets.com/polls/2011/algorithms-analytics-data-mining.html

LOES. (12 de octubre de 2010). Ley Orgánica de Educación Superior. Ley Orgánica de Educación Superior. Quito, Pichincha, Ecuador: Registro Oficial.

Microsoft. (s.f.). MSDN Library. Obtenido de http://msdn.microsoft.com/es-es/library/ms175595.aspx

Ministerio de Educación, Cultura y Deporte - España. (s.f.). Observatorio Tecnológico. Obtenido de 20Q. Inteligencia Artificial Divertida: http://recursostic.educacion.es/observatorio/web/ca/internet/recursos-online/291-sandra-miranda-esteban

Mitchell, T. (1997). Machine Learning. McGraw-Hill Science/Engineering/Math.

Moine, J., Gordillo, S., & Haedo, A. (2011). Análisis comparativo de metodologías para la gestión de proyectos de minería de datos. XVII Congreso argentino de ciencias de la computación.

Molina, J., & Herrero, J. (s.f.). TÉCNICAS DE ANÁLISIS DE DATOS. Madrid, España.

Olson, J. (2003). Data Quality. San Francisco, CA: Elsevier.

Oracle. (s.f.). Oracle. Obtenido de http://www.oracle.com/technetwork/es/documentation/317527-esa.pdf

Pérez López, C. (2007). Minería de datos: técnicas y herramientas. Paraninfo.

Rupnik, R., & Jaklič, J. (2009). The Deployment of Data Mining into Operational Business Processes. En Data Mining and Knowledge Discovery in Real Life Applications (pág. 438). Vienna, Austria: Julio Ponce and Adem Karahoca.

SAS. (s.f.). SAS Enterprise Miner. Obtenido de SEMMA: http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html

Senescyt. (diciembre de 2011). Presentación SNNA. Obtenido de http://www.educacionsuperior.gob.ec/wp-content/uploads/downloads/2012/07/SNNA_PRESENTACION.pdf

Senescyt. (12 de marzo de 2013). REGLAMENTO DEL SISTEMA NACIONAL DE NIVELACION Y ADMISION. REGLAMENTO DEL SISTEMA NACIONAL DE NIVELACION Y ADMISION. Registro Oficial 910.

Shearer, C. (2000). The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing.

Sisorg. (s.f.). QData. Obtenido de http://qdata.sisorg.com.mx/definicion.html

SNNA. (2013). Objetivos - Sistema Nacional de Nivelación y Admisión. Obtenido de http://www.snna.gob.ec/wp-content/themes/institucion/snna_objetivos.php

SNNA. (s.f.). Proceso - Sistema Nacional de Nivelación y Admisión. Obtenido de http://www.snna.gob.ec/wp-content/themes/institucion/procesodeadmision.php

Tutorial introduccion a las Redes Neuronales. (s.f.). Obtenido de http://www.redes-neuronales.com.es/tutorial-redes-neuronales/tutorial-redes.htm

Universidad Nacional de Colombia. (s.f.). http://www.virtual.unal.edu.co. Obtenido de Aprendizaje automático: http://www.virtual.unal.edu.co/cursos/ingenieria/2001832/lecciones/cap_4/intro_rna.htm

Witten, I., Frank, E., & Hall, M. (2011). Data Mining, Practical Machine Learning Tools and Techniques. Elsevier.

ANEXOS