FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

ANÁLISIS DE LAS AMENAZAS PROVENIENTES DE INTERNET QUE COMPROMETEN LA SEGURIDAD DE UNA RED EMPRESARIAL, CON EL USO DE TÉCNICAS DE MINERÍA DE DATOS.

AUTOR

ISRAEL JOSÉ NOLIVOS ARMAS

AÑO

2019

FACULTAD DE INGENIERÍA Y CIENCIAS APLICADAS

ANÁLISIS DE LAS AMENAZAS PROVENIENTES DE INTERNET QUE COMPROMETEN LA SEGURIDAD DE UNA RED EMPRESARIAL, CON EL USO DE TÉCNICAS DE MINERÍA DE DATOS.

"Trabajo de Titulación presentado en conformidad con los requisitos establecidos para optar por el título de Ingeniero en redes y telecomunicaciones"

Profesor guía

Msc. William Eduardo Villegas Chiliquinga

Autor

Israel José Nolivos Armas

Año

2019

DECLARACIÓN DEL PROFESOR GUÍA

"Declaro haber dirigido el trabajo, Análisis de las amenazas provenientes de internet que comprometen la seguridad de una red empresarial, con el uso de técnicas de minería de datos, a través de reuniones periódicas con el estudiante Israel José Nolivos Armas, en el semestre 201920, 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".

_____

William Eduardo Villegas Chiliquinga
Magister en redes de comunicaciones
CI: 1715338263

DECLARACIÓN DEL PROFESOR CORRECTOR

"Declaro haber revisado este trabajo, Análisis de las amenazas provenientes de internet que comprometen la seguridad de una red empresarial, con el uso de técnicas de minería de datos, del estudiante Israel José Nolivos Armas, en el semestre 201920, dando cumplimiento a todas las disposiciones vigentes que regulan los Trabajos de Titulación".

_____

Milton Neptalí Román Cañizares
Magister en gerencia de redes y telecomunicaciones
CI: 0502163447

DECLARACIÓN DE AUDITORIA DEL ESTUDIANTE

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

_____

Israel José Nolivos Armas
CI: 1719434449

AGRADECIMIENTO

Agradezco a mi familia por su total apoyo durante este tiempo, y a la universidad de las Américas por brindarme las herramientas necesarias para mi formación.

DEDICATORIA

Dedico este trabajo a mis padres que gracias a su apoyo han hecho posible mi formación y culminación de esta etapa.

RESUMEN

Las redes empresariales a nivel mundial se han visto afectadas por incidentes de seguridad provenientes de Internet, aprovechándose de las vulnerabilidades y falencias de seguridad que una red con acceso a Internet pueda tener. Por esta razón se propone analizar las amenazas provenientes desde el mundo que comprometen la seguridad de una red empresarial con el uso de técnicas de minería de datos.

La información se recolectará de equipos de seguridad obtenida de diversas fuentes (logs de herramientas y equipos de seguridad, sniffers), para posteriormente organizarla y subirla a un gestor de base de datos, con el fin de analizarla.

Por medio del uso de técnicas de minería de datos se logrará modelar la información obtenida desde lo más general hasta lo más específico. Mediante el descubrimiento de patrones, se identificará: ataques, tipos de amenazas y redes con mayor vulnerabilidad, con el fin de evaluar resultados y proponer técnicas de mitigación en un tiempo efectivo.

ABSTRACT

Business networks worldwide have been affected by security incidents coming from the Internet, taking advantage of security vulnerabilities and failures that a network with Internet access may have. For this reason, it is proposed to analyze the threats coming from the world that compromise the security of a business network with the use of data mining techniques.

The information will be collected from security equipment obtained from various sources (logs, sniffers), to later organize it and upload it to a database manager, in order to analyze it.

Through the use of data mining techniques it will be possible to model the information obtained from the most general to the most specific. Through the discovery of patterns, we will identify: attacks, types of threats and networks with greater vulnerability, in order to evaluate results and propose mitigation techniques in an effective time.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1. Capitulo I. Introducción

Los incidentes de seguridad provenientes de Internet a nivel mundial han afectado en su mayoría a redes empresariales, aprovechándose de las vulnerabilidades y fallas de seguridad que una red con acceso a Internet pueda tener. En la actualidad las organizaciones están expuestas a múltiples amenazas, muchas de estas ya conocidas y fácilmente identificables, mientras que otras surgen a partir de nuevos intentos de ataques; y a la falta de personal capacitado en las compañías. Bajo esta premisa se afirma que las amenazas pueden ser identificadas y mitigadas, pero jamás se eliminarán en su totalidad. (Tarazona, 2006).

Las fallas de seguridad están relacionadas tanto al factor técnico como humano, ya que muchas veces un equipo mal configurado, puertos abiertos, falta de seguridad en las estaciones de trabajo, o un usuario mal capacitado puede convertirse en un blanco de ataque. (Tarazona, 2006). Las amenazas se pueden catalogar dependiendo su criticidad:

- Amenazas que no comprometen la seguridad de la red.
- Amenazas que puede comprometer la pérdida total o parcial de la información.

Mediante herramientas de minería de datos y con métodos de exploración, se busca identificar los ataques con mayor incidencia sobre las redes empresariales con el fin de proponer técnicas de mitigación en un tiempo efectivo, en base a los resultados obtenidos.

Para el desarrollo de este proyecto se recolectará información de varias fuentes principalmente de equipos de seguridad implementados en redes empresariales, (historial de log) y reportes de sniffer en busca de identificar vulnerabilidades lo que representará el núcleo del proyecto.

Utilizando métodos de descubrimiento de datos se buscará encontrar patrones repetitivos, potencialmente interesantes tanto de forma manual o automática dependiendo de la técnica a utilizar, estas puedes ser: técnicas de predicción o hipótesis creadas por el usuario.

Las técnicas de minería de datos, con el fin del descubrimiento de datos siguen los siguientes pasos:

- Clasificación
- Predicción
- Agrupamiento
- Resumen
- Análisis de datos

1.1 Objetivo General

Identificar cuáles son los ataques y vulnerabilidades que afectan a las redes empresariales y proponer técnicas de mitigación en un tiempo efectivo.

1.2 Objetivos Específicos

- Identificar las vulnerabilidades de las redes empresariales con el uso de modelos de minería de datos y presentar los resultados.
- Proponer técnicas de mitigación de ataques en tiempo efectivo en base a los resultados de la aplicación de minería de datos.
- Evaluar resultados (tipo de ataques, técnicas de mitigación, vulnerabilidades en la red)

1.3 Alcance

Analizar las amenazas provenientes de internet que comprometen la seguridad de una red empresarial con el uso de técnicas de minería de datos. Mediante el descubrimiento de patrones, se identificarán los ataques y se propondrán técnicas de mitigación en un tiempo efectivo.

1.4 Justificación

En la actualidad los ataques centralizados desde el Internet hacia redes públicas empresariales han incrementado exponencialmente, debido a fallas de seguridad y a la explotación de sus vulnerabilidades, con el fin de obtener recursos o indisponer los servicios.

La finalidad de este proyecto es analizar cuáles son los ataques y vulnerabilidades que mayormente afectan a este tipo de redes e identificar medidas que permitan mitigar en un tiempo efectivo, con el uso de técnicas de mitigación de datos.

2. Capitulo II. Marco Teórico

En este capítulo se describe a detalle los fundamentos y teoría a utilizarse durante el desarrollo del análisis de las amenazas provenientes de internet que comprometen la seguridad de una red empresarial, con el uso de técnicas de minería de datos.

2.1 Incidencia de seguridad

Un incidente de seguridad se define como una irregularidad o un hecho imprevisto que puede comprometer la integridad, confidencialidad y disponibilidad de los datos, redes o recursos informáticos de una compañía. (Marrero, 2019).

Como se observa en la figura 1, la información es el bien más importante para las compañías, por ello, en gestión de seguridad es primordial garantizar la integridad, confidencialidad y disponibilidad de la información dentro de las redes empresariales.

Figura 1. Seguridad de la información.

Tomado de (Wordpress, 2013).

Los incidentes de seguridad dirigen sus ataques a obtener, modificar o corromper información de importancia para una compañía, indisponer un servicio o hacer uso de algún recurso informático sin previa autorización. (Marrero, 2019).

Estos incidentes se dan debido al aprovechamiento de brechas de seguridad encontradas en las redes empresariales, comúnmente son gestionados por piratas informáticos que en ingles se los conoce como *hackers,* y por programas sin verificación instalados por los usuarios finales. Por lo cual la necesidad de identificar y responder de forma rápida y sistemática a este tipo de amenazas, con el fin de minimizar el impacto sobre la red, así como los datos de la organización.

Según la norma ISO 27035, un evento de seguridad es una ocurrencia identificada en el estado de un sistema, servicio de red, indicando una posible violación de la

seguridad de la información, política o falla de los controles, o una situación previamente desconocida que puede ser relevante para la seguridad. (Ecucert, s.f).

2.1.1 Criticidad

Para determinar la criticidad de un incidente se catalogan en 5 grupos los cuales son: 1 Bajo/Nulo, 2 Medio, 3 Alto, 4 Muy alto y 5 Critico. Como se observa en la figura 2, se asigna un nivel y un nombre a cada elemento, con el fin de identificar la criticidad de cada uno.

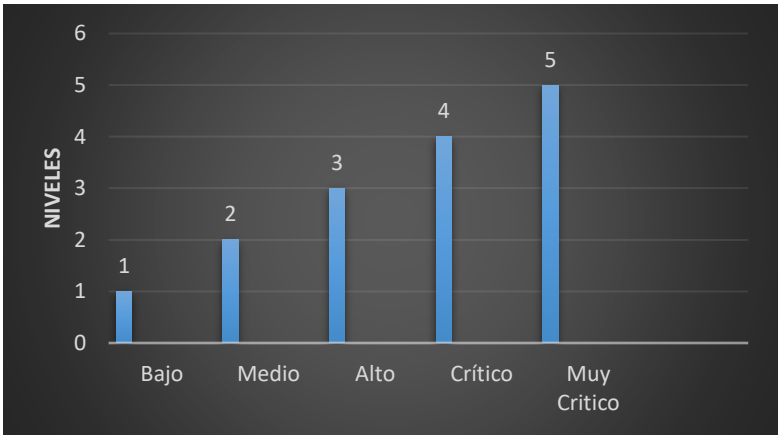


Figura 2. Niveles de Criticidad.

Adaptado de (CCN, 2018).

El nivel de criticidad es una referencia para determinar acciones a tomar en busca de soluciones, y esta se cataloga dependiendo del impacto que el incidente tenga sobre la infraestructura de red, se toma como referencia factores como: indisponibilidad de la red, experiencia del equipo de seguridad o impacto económico para la organización. Este criterio puede cambiar tanto para elevar el nivel de criticidad como para bajarlo de categoría, y depende, si durante el proceso de gestión la amenaza aumenta o cesa su incidencia sobre la red. (Centro Criptográfico Nacional [CCN], 2018).

2.1.1.1 Nivel bajo / Nulo

Este nivel se refiere a las amenazas conocidas, es decir que se tiene constancia que afectaron en algún momento la red o están incidiendo en ese instante, pero que los controles de seguridad extendidos en la red están operando normalmente y permiten contrarrestar dicha amenaza. Por esta razón su impacto es bajo o nulo. (CCN, 2018).

Las redes empresariales deben estar preparadas para responder a este tipo de amenazas, puesto que son muy comunes, tanto con personal capacitado como con controles de seguridad que limiten el impacto del incidente.

2.1.1.2 Nivel Medio

Este nivel cataloga a las amenazas que afectaron en algún momento la red o están incidiendo en ese instante, y tiene un impacto limitado tanto en la infraestructura informática como en la información no critica para la organización. Como por ejemplo equipos de usuarios. (CCN, 2018).

2.1.1.3 Nivel Alto

Se considera con este nivel a las amenazas que generan un impacto considerable a la red, afectando de esta forma la confidencialidad, disponibilidad e integridad en la infraestructura informática, como en la información no critica para una organización. Como por ejemplo equipos de usuarios. (CCN, 2018).

2.1.1.4 Nivel Muy Alto

El impacto de este nivel es considerado alto, ya que afecta a equipos centralizados de una organización, impactando de esta forma la confidencialidad, disponibilidad e integridad de la infraestructura informática, así como a la información crítica de una

organización. Como por ejemplo equipos de seguridad lógica, perimetral, autenticación, etc. (CCN, 2018).

2.1.1.5 Nivel Crítico

Se cataloga con este nivel a las amenazas que tienen un impacto muy considerable en la infraestructura de red y afecta a la continuidad y reputación del negocio, perjudicando seriamente la confidencialidad, disponibilidad e integridad en la infraestructura informática y la información crítica para la organización. (CCN, 2018).

Estos incidentes se caracterizan por causar degradación o indisponibilidad de los servicios tanto para usuarios internos como externos, puede causar la perdida de información de forma permanente, además que puede implicar cargos criminales por el acceso no autorizado, esto depende de la normativa de caga país. (CCN, 2018).

2.1.2 Clases y Tipos de Incidentes

Para la correcta categorización de los incidentes de seguridad se toma en cuenta varios factores como:

- Connotación legal.
- Impacto del incidente en relación a imagen pública o buen nombre de la organización.
- Número de equipos, servicios o redes afectadas.
- Criticidad de los servicios afectados.
- Tipo y origen de la amenaza.

Los incidentes de seguridad se clasifican de forma general, ya que un incidente puede tener varios tipos, por lo que es necesario tener en cuenta los factores anteriormente mencionados. (CCN, 2018).

2.1.2.1 Ataques

Los ataques son de tipo dirigido o mutilación.

Ataque Dirigido: Se cataloga de esta forma, cuando las víctimas (personas u organizaciones) son deliberadamente escogidas con el fin de introducirse en la red y sustraer información valiosa de sus equipos informáticos. (CCN, 2018). Los ataques dirigidos son considerados como las amenazas más importantes para la seguridad de las organizaciones, ya que pueden dañar la reputación e imagen, así como causar daños económicos elevados. (Spam, DoS, Inyección SQL, Spear Phishing, etc.)

Ataque Mutilación: Modifica el contenido de las páginas web, aprovechándose de las falencias en la seguridad de la red en los sistemas de alojamiento. Los atacantes buscan modificar el contenido de la página web, insertando links a sitios maliciosos, agregando información de publicidad el atacante, o desprestigiar el nombre de la compañía propietaria del sitio web. (CCN, 2018).

2.1.2.2 Código Malicioso

Los ataques pueden ser de tipo Infección única o infección extendida. Es un software cuyo propósito es introducirse en un equipo, servidor o elemento de red con el fin de dañarlo. El código malicioso puede afectar a un solo dispositivo o a un grupo completo de equipos o sistemas dentro de la red, la penetración de este tipo de ataque representa una violación a la seguridad de la red implementada en ese momento. (Caballo de Troya, Script, Ransomware, etc) (CCN, 2018).

2.1.2.3 Denegación de servicio (DoS)

Los ataques pueden ser de tipo exitosa y no exitosa.

Exitosa: El objetivo es indisponer un servicio aprovechando las vulnerabilidades de seguridad dentro de la red, o habitualmente mediante el envío de gran cantidad de datos a un servicio en particular, evitando de esta forma que peticiones auténticas puedan ser despachadas e incapacitando los sistemas. (DoS, DDoS, Sabotaje, etc) (CCN, 2018).

No exitosa: El ataque fue eliminado o controlado parciamente por los sistemas de seguridad en la red, y la perdida de información no fue relevante para los intereses de la compañía. (CCN, 2018).

2.1.2.4 Acceso no Autorizado

Los ataques son del tipo acceso no autorizado y pérdida de datos. El agresor consigue acceso físico o remoto a los servidores, datos, aplicaciones o algún equipo informático dentro de la red, este puede ser originado tanto de forma interna como externa. La pérdida de datos también se contempla por el robo o pérdida de equipamiento con información crítica para la organización como: calves, diagramas de red, listas de usuarios, documentación de sistemas y aplicaciones, etc. (CCN, 2018).

2.1.2.5 Pruebas y reconocimientos

Los ataques son del tipo pruebas no autorizadas y monitoreo. Se determina como cualquier actividad sospechosa que busca obtener acceso o información de equipos de red, puertos abiertos, aplicaciones, servicios, datos, etc para su posterior uso o intento de acceso fraudulento. En cuanto a monitoreo se refiere a los eventos registrados por equipos de seguridad perimetral implementados en la organización (firewall, IDS, IPS, etc) que no entren en alguna categoría previamente establecida. (CCN, 2018).

2.1.2.6 Daños físicos

Un usuario sin previa autorización logra obtener acceso físico a los equipos y modifica o daña los mismos. Comúnmente este tipo de ataques lo gestionan elementos internos a la compañía, pero también lo pueden realizar externos. (CCN, 2018).

2.1.2.7 Obtención de información

Ataques más sofisticados, los cuales buscan obtener información estratégica en ámbitos de seguridad, con el propósito de identificar vulnerabilidades. (Ingeniería social, Sniffing, escaneo de puertos, etc.) (CCN, 2018).

2.1.2.8 Priorización y tiempo

Para brindar atención apropiada a los incidentes de seguridad se debe establecer una prioridad, con el fin de atender los requerimientos urgentes y los aplazables según la necesidad. (Ministerio de Tecnologías de la Información y las Comunicaciones [MINTIC], 2016).

Para lo cual se definen algunas variables las cuales permitirá medir los incidentes.

- Criticidad
- Criticidad de Impacto
- Impacto Actual
- Impacto Futuro

2.1.2.9 Nivel de Prioridad

La criticidad se determina dependiendo del impacto sobre la red o sistemas informáticos afectados. Como se observa en la tabla número 1, para determinar el

nivel de prioridad se toma en cuenta el o los sistemas que el incidente se seguridad a afectado y se asigna un valor de acuerdo a su criticidad. (MINTIC, 2016).

Tabla 1.

Nivel de criticidad

Nivel de criticidad	Valor	Definición
Inferior	0,10	Sistemas no críticos, como estaciones de trabajo.
Bajo	0,25	Sistemas dependientes de una sola localidad.
Medio	0,50	Sistemas dependientes de varias localidades.
Alto	0,75	Sistemas del área de tecnología, usuarios con funciones críticas.
Superior	1,00	Sistemas críticos.

Adaptado de (MINTIC, 2016).

2.1.2.10 Impacto Actual

Se determina por la cantidad de daño que ha logrado realizar el incidente de seguridad al instante de su detección. (MINTIC, 2016).

2.1.2.11 Impacto Futuro

Se determina por la cantidad de daño que ha logrado realizar el incidente, si este no es mitigado o eliminado en su totalidad. Como se puede observar en la tabla número 2, el nivel de impacto se determina por la cantidad de daño que un incidente de seguridad ha podido causar y se asigna un valor de acuerdo a su criticidad. (MINTIC, 2016).

Tabla 2.

Nivel de impacto

Nivel de Impacto	Valor	Definición

Inferior	0,10	Impacto leve, en uno de los componentes de cualquier sistema de información.
Bajo	0,25	Impacto moderado, en uno de los componentes de cualquier sistema de información.
Medio	0,50	Impacto alto, en uno de los componentes De cualquier sistema de información.
Alto	0,75	Impacto moderado, en uno o más componentes de más de un sistema de información.
Superior	1,00	Impacto alto, en uno o más componentes de más de un sistema de información.

Adaptado de (MINTIC, 2016).

La prioridad del incidente de seguridad se obtiene gracias a las variables anteriormente definidas con la siguiente formula. (MINTIC, 2016).

Nivel Prioridad = (Impacto actual * 2,5) + (Impacto futuro * 2,5) + (Criticidad del Sistema * 5) como se puede observar en la tabla número 3. (MINTIC, 2016).

Tabla 3.

Nivel de prioridad

Nivel de prioridad	**Valor**
Inferior	00,00 – 02,49
Bajo	02,50 – 03,74
Medio	03,75 – 04,99
Alto	05,00 – 07,49
Superior	07,50 – 10,00

Adaptado de (MINTIC, 2016).

2.1.3 Tiempo de respuesta

El tiempo de respuesta es una medida que indica el tiempo máximo en que un incidente de seguridad debe ser atendido, estos tiempos dependen de la criticidad y el impacto que tiene el incidente sobre la red. Este tiempo no hace referencia al

tiempo de solución de un incidente, ya que esto varía dependiendo del caso. Como se describe en la tabla número 4, el tiempo de respuesta está ligado a la prioridad del incidente de seguridad, y este a su vez, se determina por la criticidad y el impacto del incidente. (MINTIC, 2016).

Tabla 4.

Tiempo de respuesta

Nivel de prioridad	**Tiempo de respuesta**
Inferior	3 horas
Bajo	1 hora
Medio	30 minutos
Alto	15 minutos
Superior	5 minutos

Adaptado de (MINTIC, 2016).

2.2 Red empresarial

La infraestructura de red empresarial se define como el conjunto de redes LAN y WAN, así como dispositivos de red como routers y switches. Las redes empresariales se han convertido en una herramienta fundamental para el desarrollo de las compañías, ya que a través de estas gestionan sus procesos empresariales buscando la permanencia y competitividad del negocio. En la actualidad las redes han crecido considerablemente en busca de alcanzar niveles óptimos de diseño y flexibilidad en la red, con el fin de adaptarse a las nuevas exigencias del mercado. (Istnetgroup, 2018).

La evolución de la tecnología ha hecho que las organizaciones cambien la forma tradicional de administrar sus redes y ha convertido a su departamento de TI en pieza clave del negocio. La virtualización, redes definidas por sofware por sus siglas en inglés SDN, redes inalámbricas, redes inteligentes, computación en la nube del inglés *cloud computing*, son algunos de los avances tecnológicos que han redefinido la forma de administrar las redes empresariales, simplificando el manejo y gestión

de las mismas. Por otro lado, la convergencia de red es el presente de las redes empresariales, ya que se aprovecha la capacidad de la red con el propósito de soportar nuevos servicios y aplicaciones por la infraestructura existente. (Istnetgroup, 2018).

2.3 Minería de datos

La minería de datos se define como un grupo de técnicas y tecnologías que permiten analizar gran cantidad de información, con el fin de hallar tendencias y patrones repetitivos que logren el descubrimiento de datos ocultos, mediante el análisis matemático de los datos. El análisis tradicional de datos no permite detectar estos patrones y tendencias, debido a la gran cantidad de información y porque que las relaciones pueden resultar demasiado complejas para el ojo humano. Para el análisis de datos existen modelos de minería de datos, estos se definen por las tendencias y patrones esperados estos son: (Github, 2019).

- Previsión
- Riesgo y Probabilidad
- Recomendaciones
- Secuencias
- Agrupación

SAS Institute define el concepto de minería de datos como el proceso de seleccionar, explorar, modificar, modelizar y valorar gran cantidad de datos con el objetivo de descubrir patrones desconocidos que puedan ser utilizados como ventaja comparativa respecto a los competidores. Este proceso es resumido con las siglas SEMMA. La figura 3 ilustra las fases del proceso de minería de datos. (Software and Analytics Solutions Institute [SAS], 2017).

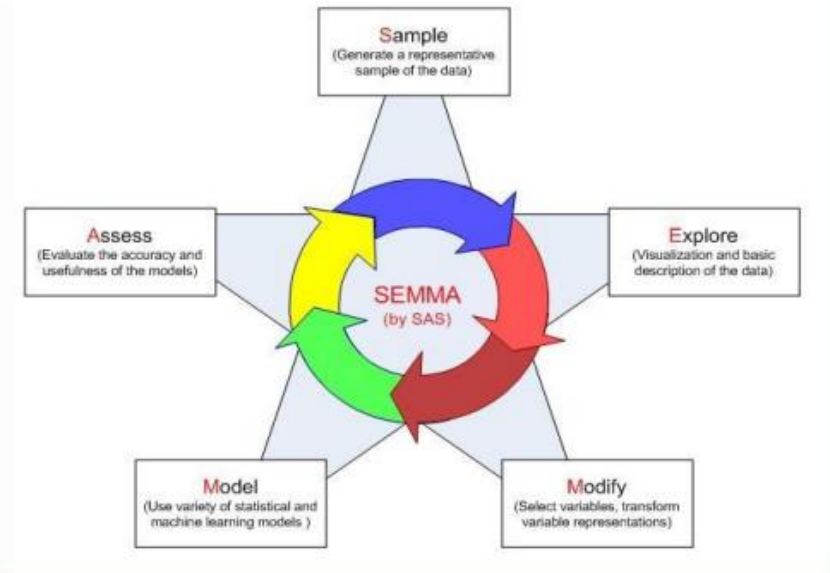


Figura 3. SEMMA.

Tomado de (Bavirisetty, 2015).

El desarrollo de un modelo de minería de datos responde a una pregunta sobre qué información se espera obtener de los datos entregados, y se busca responder a esta pregunta con la implementación de un modelo. Para lo cual se debe seguir los siguientes pasos. (Github, 2019).

2.3.1 Definir el problema

El primer paso en minería de datos es definir el problema, con el fin de buscar una respuesta al problema planteado, mediante el análisis de los datos entregados. En este paso se define el ámbito del problema, los objetivos específicos que se busca alcanzar mediante la minería de datos y las métricas a evaluarse. (Github, 2019).

2.3.2 Preparar los datos

En este paso se trabaja con los datos previamente recolectados, mediante la consolidación y filtrado de información; es decir se limpian los datos entregados con

el propósito de discriminar la información relevante de la irrelevante para este trabajo. Como se observa en la figura 4. (Github, 2019).

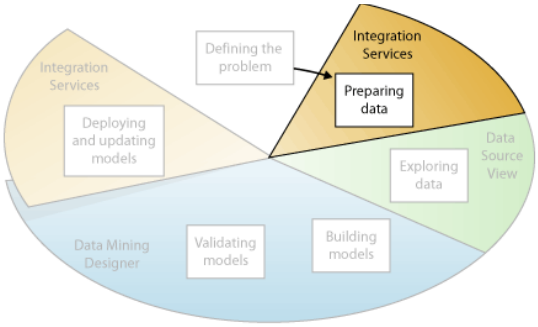


Figura 4. Preparación de datos.

Tomado de (Github, 2019).

El filtrado de información es uno de los pasos más importantes dentro del proceso de minería de datos, puesto que mucha de la información puede ser errónea, estar incompleta, desactualizada o almacenada en diferentes formatos. Por esta razón los datos incompletos, erróneos y la información que parece ser independiente, en realidad puede estar estrechamente relacionada, y puede influir de manera inesperada en los resultados del estudio. Es por esto que antes de empezar con este trabajo es necesario identificar estos problemas y buscar el mecanismo para corregirlos. (Github, 2019).

La depuración de los datos no solamente implica quitar datos erróneos, eliminar datos no relevantes para el análisis, o proponer nuevos datos en los campos incompletos, sino también buscar relaciones entre estos datos y determinar los campos más adecuados para el análisis, con el fin descubrir información oculta y patrones de conducta. (Github, 2019).

2.3.3 Consolidación de datos

ETL se define como el proceso de integración de datos, el cual consta de tres pasos: extraer, transformar y cargar, se utiliza para unir datos de varias fuentes con el fin

de crear una base de datos consolidada, la cual puede ser de tipo almacén de datos del inglés *data warehouse*, cubos olap, archivo excel, etc. Como se puede observar en la figura 5. (SAS, 2017).

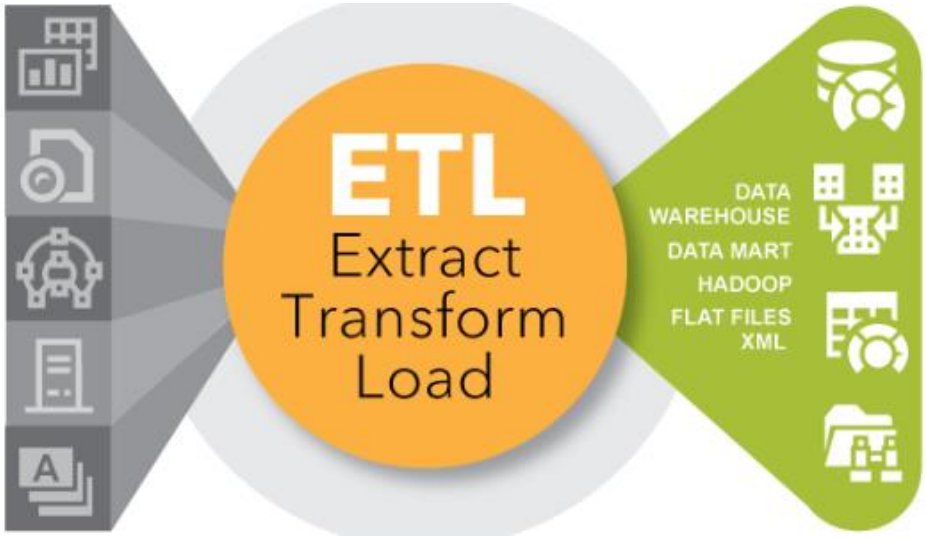


Figura 5. ETL.

Tomado de (SAS, 2017).

2.3.4 Almacén de Datos

Del inglés *data warehouse* es una base de datos que alberga gran cantidad de información, se caracteriza por reunir información de varias fuentes, para luego ser procesadas mediante un ETL. Se caracteriza por ser una estructura integrada, lo que implica que las inconsistencias entre los diversos sistemas de donde se toma la información deben ser eliminados, con el fin de organizar la información en distintos niveles de detalle para adecuarse a los requerimientos del usuario. (Sinnexus, s.f).

Para el desarrollo de esta tesis no se tomará esta alternativa, puesto que no se cuenta con demasiada información la cual no requiere nivel de detalle.

2.3.5 Cubos Olap (procesamiento analítico en línea)

Es una base de datos multidimensional capaz de procesar gran cantidad de datos simultáneamente, gracias a que está formada de vectores, permite que cada dimensión cuente con información específica y permite ser comparada con las demás estructuras del cubo. (Tecnología-informática, 2019).

Esta estructura es muy utilizada por organizaciones que cuentan con gran cantidad de información y requieren realizar consultas de forma rápida, debido a su estructura de tres dimensiones. Para el desarrollo de la tesis no se optará por esta solución, ya que no se cuenta con gran cantidad de información para el análisis y no es necesario gestionar los datos en dimensiones para las consultas.

2.3.6 Explorar los datos

Una vez identificados los datos relevantes para el análisis, se pone en marcha el estudio. Esta fase quizá sea la más sensible durante el proceso de minería de datos, puesto que se deberá escoger una muestra lo suficientemente relevante para el desarrollo del análisis, además del tipo de variables y modelos que se adapten a la naturaleza del estudio. (Lopez, 2019).

Además, es necesario modificar los datos, creando, transformando y seleccionando variables, estas determinan modelo a utilizar en el desarrollo del análisis. Las modificaciones no siempre se realizan al principio del análisis, sino también se ejecutan durante el desarrollo del mismo, puesto que los datos pueden cambiar con el tiempo, haciendo que el entorno de trabajo sea dinámico e interactivo. (Wordpress, 2011).

2.3.7 Modelos

Existen diferentes tipos de técnicas de minería de datos, las cuales se emplean dependiendo del resultado esperado, sin embargo, se pueden utilizar varias técnicas para analizar la misma tarea, con el propósito de obtener resultados diferentes e incluso más de uno. Para ello se tienen: técnicas de clasificación, regresión, segmentación, asociación y análisis de secuencias. (Github, 2019).

- Clasificación: Consiste en la predicción de uno o varios atributos, tomando como fuente el conjunto de datos total o parcialmente. Entre las técnicas más conocidas están: árbol de decisión y red neuronal.
- Regresión: Predice uno o más atributos numéricos, basándose en el conjunto de datos. La técnica de regresión lineal es la más común en este tipo de técnicas.
- Segmentación: El objetivo es agrupar información con propiedades similares, tanto en grupos como clústeres. Las técnicas más comunes son: Análisis de grupos del inglés *clustering* y red neuronal.
- Asociación: El objetivo es correlacionar diferentes atributos dentro de un mismo grupo de datos, con el fin de obtener información no relacionada inicialmente. Algunas técnicas importantes son: análisis de asociación, patrones secuenciales y series temporales.
- Análisis de Secuencias: hace referencia a los eventos o episodios que se repiten frecuentemente.

2.3.8 Generar modelos

En esta fase se procesa los datos, luego de haberlos obtenido, limpiado, filtrados, ordenado, modificado, etc. Mediante el uso de un software especializado se analiza de forma completa la información ingresada, mediante la combinación de datos se busca coincidencias, relaciones y se logra predecir de una forma confiable los resultados buscados. (Wordpress, 2011).

Para la generación de modelos es necesario crear una estructura de minería de datos, para lo cual se debe escoger las columnas que serán utilizadas para el procesamiento de datos. Las herramientas de minería de datos utilizan algoritmos matemáticos y estadísticos concretos para su análisis, lo que genera información adicional a la ingresada en la estructura de datos, esta información servirá para extraer patrones e identificar relaciones entre sí. (Github, 2019).

2.3.9 Árbol de decisiones

Se identifica como una técnica de clasificación robusta, que permite tomar decisiones en base a sus resultados, los mismos que son fácil de interpretar. Esta técnica permite evaluar los resultados en tres parámetros diferente que son: probabilidad, que muestra las posibilidades de los resultados mostrados, decisión, muestra la elección a tomarse en base a los resultados, y terminales, muestra el resultado definitivo de la consulta. La predicción mediante el árbol de decisiones muestra información relacionada antes no conocida, lo que ayuda a tomar decisiones en base a los resultados obtenidos. (Lucidchart, 2019).

2.3.10 Análisis de grupo (Clustering)

Análisis de grupos del inglés *Clustering* es una técnica de minería de datos del tipo agrupación, el proceso consiste en agrupar la información en grupos de datos similares. Por medio de relaciones entre las variables de cada grupo de datos, mide la similitud entre sí y organiza la información en clases con datos similar, pero de diferentes orígenes. Mediante la aplicación de esta técnica se logra simplificar los datos, reduciendo la información original y mostrando en segmentos fáciles de interpretar.

2.3.11 Clasificador Bayesiano (Naive Bayes)

Clasificador Bayesiano del inglés *Naive Bayes* es una técnica de clasificación la cual permite predecir resultados mediante la utilización de modelos matemáticos. Clasificador Bayesiano busca asociaciones y relaciones dentro de un conjunto de datos con el propósito de predecir el comportamiento de un atributo, para la ejecución de este modelo de minería de datos es necesario asignar un dato clave, este sirve como fuente de predicción y las relaciones encontradas están en función de esta variable. (Blogspot, 2009).

2.3.12 Explorar y validar los modelos

En este proceso se evalúa el correcto funcionamiento de los modelos, antes de colocarlos en un entorno de producción. Al gestionar un modelo de minería de datos se crean varias configuraciones, las cuales deben ser puestas a prueba y verificar la mejor opción, la misma deberá resolver el problema plateado inicialmente. (Github, 2019).

Para esto se debe evaluar exhaustivamente las variables que formaran parte del modelo seleccionado, esta depende del tipo de información ingresada. (Wordpress, 2011).

2.3.13 Implementar y actualizar los modelos

El propósito de este punto es realizar un análisis de resultados, con el fin de confirmar si la información obtenida es coherente con la información ingresada y con el planteamiento del problema. Una vez que los modelos de minería de datos entran en producción, la herramienta puede brindar variada información, esto depende de la necesidad del usuario. (Github, 2019 y Wordpress, 2011).

Finalmente, el modelo implementado debe ser actualizado periódicamente, ya que las variables ingresadas pueden volverse obsoletas y poco significativas para el análisis con el pasar del tiempo. (Lopez, 2019).

3. CAPÍTULO III. RECOLECCIÓN Y ANÁLISIS DE INFORMACIÓN

Para el desarrollo de esta tesis se ordena los datos recolectados en su formato original, el propósito es filtrar la información con el fin de extraer los datos relevantes para el análisis y eliminar datos inconsistentes o inexistentes y colocarlos en una estructura fácil de interpretar. Para lo cual se utiliza el proceso ETL, mediante la herramienta Microsoft Visual Studio con la funcionalidad *Analysis Services*.

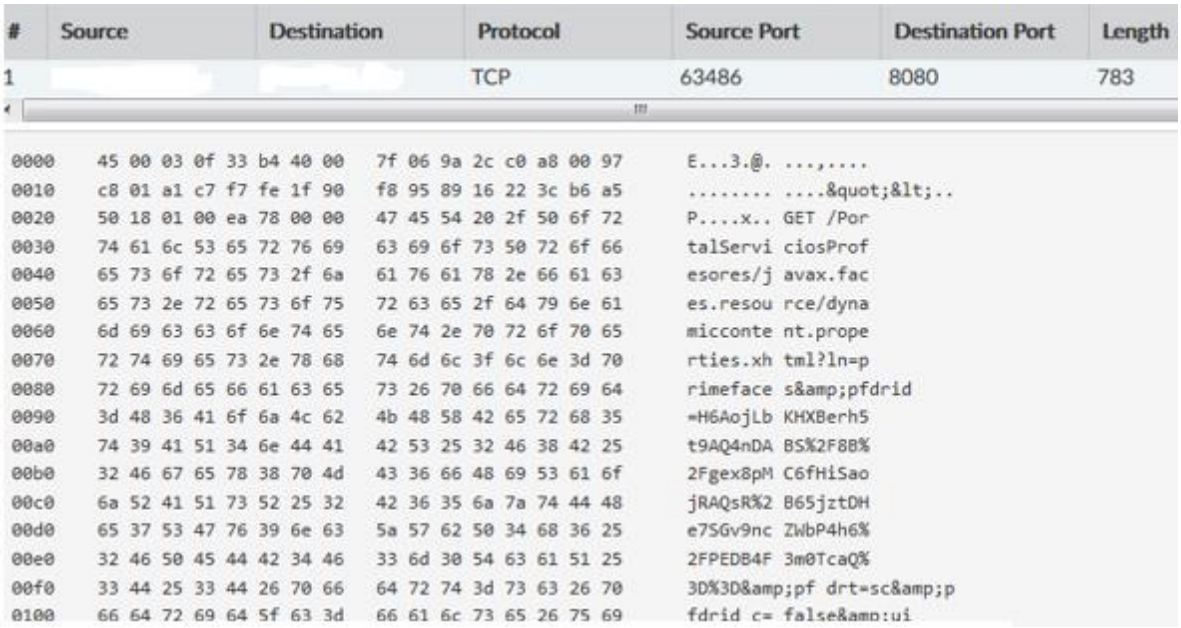
3.1 Recolección de Información

Las fuentes de información incluida en la fase de recolección son logs de equipos de seguridad perimetral, enfocados en detectar los mecanismos de intento de intrusión en la red y reportes de sniffers (analizador de red), dirigidos a identificar las vulnerabilidades de los equipos por medio de puertos abiertos, equipos sin protección, etc.

3.1.1 Logs de equipos de seguridad

Se obtienen los logs de varios equipos de seguridad perimetral que actualmente están en producción dentro de redes empresariales, estos, mediante niveles de confianza permiten o deniegan el acceso interno o externo a equipos o sistemas dentro de la red, y almacenan información de los incidentes que fueron detectados y considerados como intrusión o intento de intrusión a la red. Mediante la revisión de estos logs se analiza la forma en que un atacante intenta acceder a la red e identificar el propósito del ataque.

En la figura 6, los logs proporcionan gran cantidad de información sobre el incidente en un formato difícil de imprentar. La primera columna representa información de numeración de filas, la segunda columna la dirección IP origen, la tercera dirección IP destino, la columna número 4 es la más importante ya que es donde se registra toda la información requerida para el análisis. Aquí se identifica el origen del incidente, el puerto, el servicio, e información propia del equipo. La mayoría de esta información no es relevante para el análisis, por esta razón es necesario filtrar los datos relevantes y colocarlos en un formato amigable, ya que el log no es fácil de interpretar a simple vista.



#	Source	Destination	Protocol	Source Port	Destination Port	Length
1			TCP	63486	8080	783

```
0000   45 00 03 0f 33 b4 40 00   7f 06 9a 2c c0 a8 00 97   E...3.@. ...,....
0010   c8 01 a1 c7 f7 fe 1f 90   f8 95 89 16 22 3c b6 a5   ........ ....&quot;&lt;..
0020   50 18 01 00 ea 78 00 00   47 45 54 20 2f 50 6f 72   P....x.. GET /Por
0030   74 61 6c 53 65 72 76 69   63 69 6f 73 50 72 6f 66   talServi ciosProf
0040   65 73 6f 72 65 73 2f 6a   61 76 61 78 2e 66 61 63   esores/j avax.fac
0050   65 73 2e 72 65 73 6f 75   72 63 65 2f 64 79 6e 61   es.resou rce/dyna
0060   6d 69 63 63 6f 6e 74 65   6e 74 2e 70 72 6f 70 65   micconte nt.prope
0070   72 74 69 65 73 2e 78 68   74 6d 6c 3f 6c 6e 3d 70   rties.xh tml?ln=p
0080   72 69 6d 65 66 61 63 65   73 26 70 66 64 72 69 64   rimeface s&amp;pfdrid
0090   3d 48 36 41 6f 6a 4c 62   4b 48 58 42 65 72 68 35   =H6AojLb KHXBerh5
00a0   74 39 41 51 34 6e 44 41   42 53 25 32 46 38 42 25   t9AQ4nDA BS%2F8B%
00b0   32 46 67 65 78 38 70 4d   43 36 66 48 69 53 61 6f   2Fgex8pM C6fHiSao
00c0   6a 52 41 51 73 52 25 32   42 36 35 6a 7a 74 44 48   jRAQsR%2 B65jztDH
00d0   65 37 53 47 76 39 6e 63   5a 57 62 50 34 68 36 25   e7SGv9nc ZWbP4h6%
00e0   32 46 50 45 44 42 34 46   33 6d 30 54 63 61 51 25   2FPEDB4F 3m0TcaQ%
00f0   33 44 25 33 44 26 70 66   64 72 74 3d 73 63 26 70   3D%3D&amp;pf drt=sc&amp;p
0100   66 64 72 69 64 5f 63 3d   66 61 6c 73 65 26 75 69   fdrid_c= false&amp;ui
```

Figura 6. Intento de instrucción.

3.1.2 Reportes de sniffers

Mediante el software especializado en escaneo de puertos Nmap versión 7.70 se obtiene reportes de vulnerabilidad en las redes empresariales con acceso a Internet y que utilizan IPs públicas para el efecto, obteniendo información importante como

el tipo de equipo, sistema operativo, servicios corriendo sobre la red, entre otros. Como se puede observar en la figura 7.

```
22/tcp   open   ssh        (protocol 2.0)
| fingerprint-strings:
|   NULL:
|_    SSH-2.0-axXYs
| ssh-hostkey:
|   1024 85:82:70:de:39:90:b9:ab:cf:88:d9:0e:04:dd:77:e9 (RSA)
|_  256 3b:b1:7d:0b:0a:42:a0:c7:6e:8a:8f:f5:d9:eb:90:96 (ED25519)
113/tcp  closed ident
443/tcp  open   ssl/http   Fortinet security device httpd
| http-methods:
|_  Supported Methods: GET HEAD OPTIONS
|_http-server-header: xxxxxxxx-xxxxx
|_http-title: Site doesn't have a title (text/html).
| ssl-cert: Subject: commonName=FG200D4614812365/organizationName=Fortinet/stateOrProvinceName=California/countryName=US
| Issuer: commonName=support/organizationName=Fortinet/stateOrProvinceName=California/countryName=US
| Public Key type: rsa
| Public Key bits: 1024
| Signature Algorithm: sha1WithRSAEncryption
| Not valid before: 2014-12-31T09:50:21
| Not valid after:  2038-01-19T03:14:07
| MD5:   6981 8b62 b5fd f8ea 7ad6 68de 2e86 491a
|_SHA-1: 12e9 40c3 81f7 defe 7b3d 8053 d686 74b3 3b6d 0550
|_ssl-date: TLS randomness does not represent time
1443/tcp open   tcpwrapped
1 service unrecognized despite returning data. If you know the service/version, please submit the following fingerprint at
```

Figura 7. Reporte de vulnerabilidad.

El análisis de este reporte permitirá conocer las vulnerabilidades más comunes que afectan a las redes empresariales, sus principales debilidades y las probabilidades de ataques a las cuales están expuestas. Además, el reporte debe ser depurado puesto que la gran mayoría de la información no es relevante para el estudio.

3.2 Análisis de información

Se considera relevante toda la información que permita entender la naturaleza del ataque, como son: dirección IP origen, dirección IP destino, puerto origen, puerto destino, protocolo, servicio, dispositivo, sistema operativo, tipo de incidente, entre otros. Además, se puede agregar información que se crea importante para el análisis y que no consta en el log como la prioridad del incidente, el cual de detalla en esta tesis en el capítulo 1.1.3.1 y la forma en que se detectó el incidente.

- IP origen: IP desde donde se genera la amenaza, dependiendo del incidente el origen puede provenir de una o varias IPs.
- IP destino: IP objetivo del ataque. (equipo o servicio)
- Puerto Origen: Puerto desde donde se genera la amenaza
- Puerto Destino: Puerto por el cual el atacante está teniendo acceso a la red.
- Protocolo: Protocolo de transporte de datos puede ser TCP o UDP.
- Servicio: Protocolo corriendo sobre la red.
- Dispositivo: Objetivo de la amenaza generalmente suelen ser servidores, routers, dispositivos finales, etc.
- Sistema operativo: Sistema operativo del dispositivo afectado.
- Tipo de incidente: Dependerá de la vulnerabilidad detectada, por ejemplo, OpenDns.
- Origen: Hace referencia al país o región desde donde se gestiona la amenaza.
- Prioridad del incidente: Puede ser inferior, bajo, medio, altos y superior dependiendo el caso.
- Detección: Para el desarrollo de esta tesis, se tendrá únicamente dos fuentes de detección que son open port y vulnerability.
- Acceso: Se determina por como el atacante toma control o acceso de los recursos de red.

Debido a que la información se ha tomado de diferentes fuentes se utiliza la herramienta Microsoft Visual Studio con las características Analysis Services, con el propósito de crear un ETL que unifique la información, desarrollando un formato único, el que reúne los datos más importantes de cada una de las fuentes, de esta manera se obtiene la mayor y mejor información disponible para el análisis. Adicionalmente se emplea la herramienta de base de datos Microsoft SQL Server para almacenar la información obtenida mediante el ETL.

3.2.1 Construcción ETL

Los datos se obtuvieron de diferentes fuentes y en formatos distintos, por lo cual la consolidación de la información se realiza mediante un ETL. De esta forma se obtiene toda la información en una misma base de datos y con un formato en específico.

3.2.1.1 Extracción

Este proceso se refiere a la recolección, limpieza y accesibilidad de la información, con el fin que los datos estén disponibles para ser utilizados. (Pearlman, 2019).

Los datos obtenidos para el desarrollo de este proyecto están en un formato plano .txt, por lo cual es necesario cambiar a un formato plano que permita mostrar la información en forma de tablas, .csv (*comma-separated values*). Este formato logra separar los datos en forma de columnas cuando hay una separación de por medio, esta puede ser punto, punto y coma o espacio y las filas cuando existe una nueva línea de texto. Como se puede observar en la figura 8.

Completedx	Initxiating	Complneted	Initiatilng	Scxanning	Discoverñed	Dimscovered	Comxpleted	Initiatingw	Scaxnning	Completsed	Initiatfing	Retrying	Initiateeing
Ping	Parallel	Parallel	SYN	host-157-100	open	open	SYN	Service	2	Service	OS	OS	Traceroute
Scan	DNS	DNS	Stealth	(157.100.1.22	port	port	Stealth	scan	services	scan	detection	detection	at
at	resolution	resolution	Scan	[1000	111/tcp	22/tcp	Scan	at	on	at	(try	(try	9:48
9:47	of	of	at	ports]	on	on	at	9:47	host-157-100	9:48	#1)	#2)	
0.92s	1	1	9:47				9:47			6.06s	against	against	
elapsed	host.	host.					3.03s			elapsed	host-157-100		
(1	at	at					elapsed			(2			
total	9:47	9:47					(1000			services			
hosts)		0.01s					total			on			
		elapsed					ports)			1			

Figura 8. Extracción de información.

Mediante la utilización de la herramienta Microsoft Visual Estudio con las características Analysis Services, se automatiza el acceso a la fuente de información, con el fin de que cualquier tipo de log pueda ser ingresado y analizado. Por lo cual se crea la ubicación de la información, esta puede estar almacenada localmente como puede ser tomada directamente de los equipos de

seguridad que generan el log. Para el desarrollo de esta tesis los datos están almacenados localmente. Como se puede observar en la figura 9.

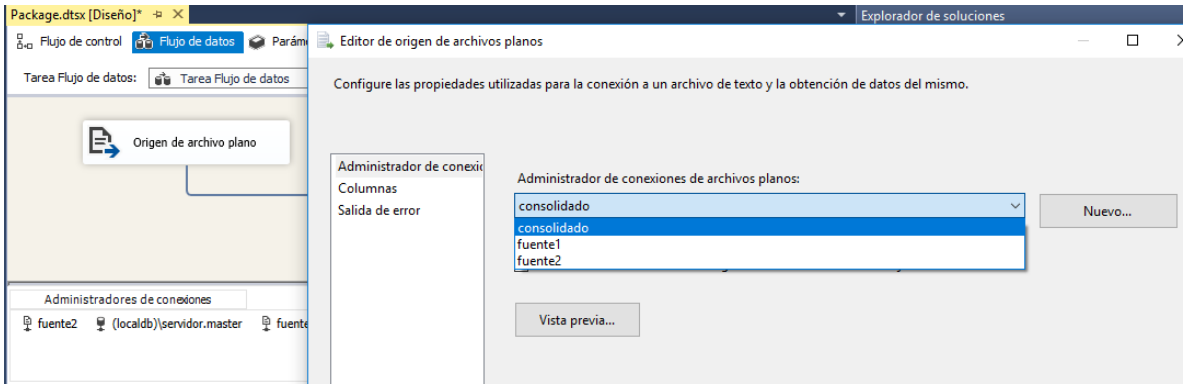


Figura 9. Consolidación de las fuentes de información mediante el ETL.

3.2.1.2 Transformación

La fase de transformación se refiere a la modificación de los datos con el propósito de enmárcalos en un formato funcional para el análisis, para ello elimina datos duplicados y define qué información es considerada para el análisis y cual es desechada. Además, la verificación de información es un proceso importante dentro de la transformación de los datos, puesto que permite identificar información similar, que puede estar o no relacionada con los datos escogidos para el análisis. Esto permite seguir filtrando la información hasta obtener solo lo necesario. (Pearlman, 2019).

Una vez conocido el origen de la información, lo datos son extraídos para transformarlos en un formato libre de información irrelevante, mediante un proceso de ETL. En este proceso se discrimina la información importante de la intrascendente, las columnas se eligen en base a las características del análisis, por lo cual esta fase es muy importante para lograr identificar los ataques y las vulnerabilidades que afectan a una red empresarial. Como se puede observar en la figura 10

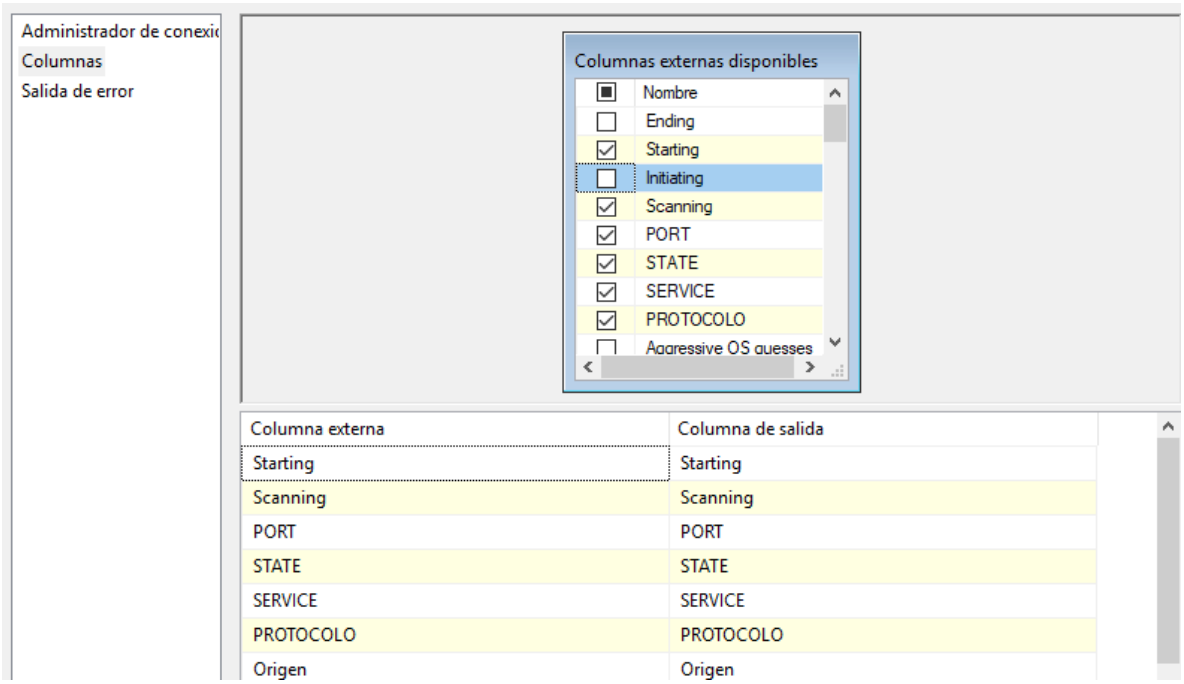


Figura 10. Proceso de escoger la información relevante.

3.2.1.3 Carga

En esta fase como su nombre lo indica se trata de cargar los datos anteriormente extraídos en su nuevo formato y ubicación. La carga de información se la realiza de dos formas carga completa o carga incremental. (Pearlman, 2019).

Para el desarrollo de esta tesis la carga es incremental, es decir que cuando exista una nueva fuente de información o se modifique la ya existente, esta puede ser ejecutada en el ETL y cargada en la base de datos destino para su análisis.

El proceso ETL después de la transformación de la información y eliminación de datos irrelevantes, es almacenado en una base de datos tipo OLE DB, a través del gestor de base de datos Microsoft SQL Server, esta información es la fuente de datos para el análisis de minería de datos mediante la herramienta Rapidminer. Como se puede observar en la figura 11 y 12 respectivamente.

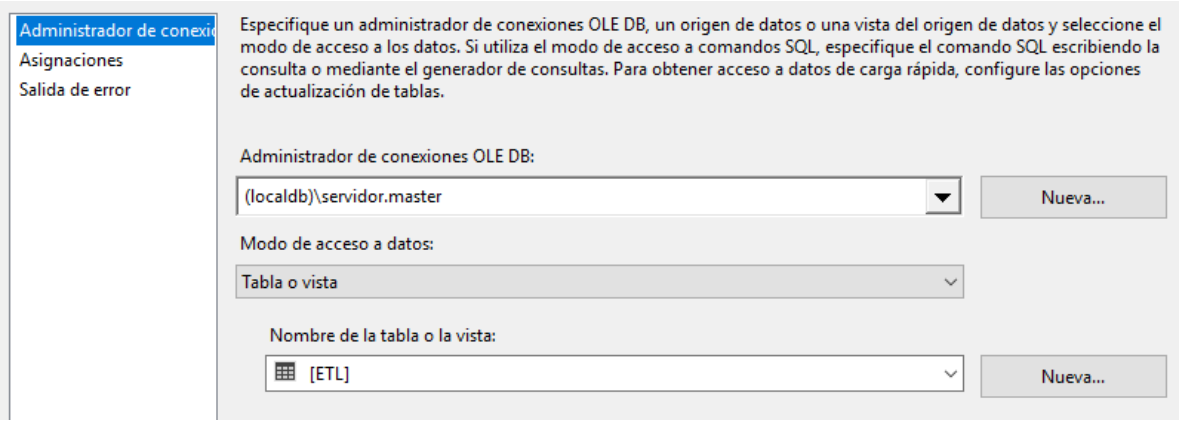


Figura 11. Asignación de base de datos y creación de tabla.

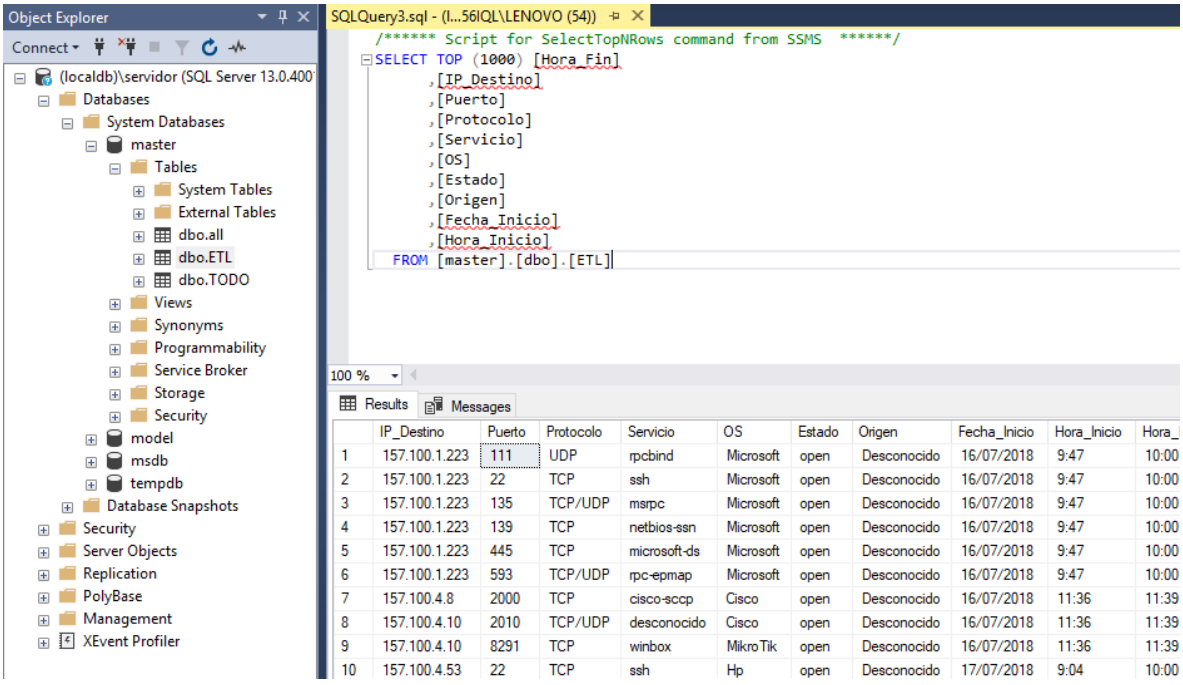


Figura 12. Visualización de la información extraída del ETL en SQL server.

4. CAPÍTULO IV. TÉCNICAS DE MINERÍA DE DATOS

Las técnicas de minería de datos son un conjunto de algoritmos y cálculos matemáticos que ayudan a crear modelos a partir de gran cantidad de información ingresada. Mediante el análisis matemático es posible descubrir patrones y

tendencias que no son detectables mediante técnicas tradicionales de exploración. (Chunga, 2013).

4.1 Metodología

Para el desarrollo de este trabajo se emplean tres técnicas de minería de datos, las cuales son: Árbol de decisiones, Análisis de grupos del inglés *Clustering* y Clasificador bayesiano del inglés *Naive Bayes*, debido a que son las técnicas enfocadas en obtener valores de salida. Para lo cual se utiliza la herramienta de minería de datos RapidMiner, puesto que es una herramienta de acceso libre (Open Source) que brinda información gráfica y estadística muy completa. El objetivo es comparar los resultados de la ejecución de estas técnicas de minería de datos y sacar conclusiones que permitan proponer técnicas de mitigación.

4.1.1 Dato Clave

Para aplicar las técnicas de minería de datos se debe declarar una variable principal o dato clave, las predicciones y relaciones obtenidas al ejecutar un modelo de minería de datos están ligadas a esta variable. Es por esto que al cambiar de dato clave se obtiene resultados totalmente diferentes.

4.2 Árbol de decisiones

Para la ejecución de este modelo se conecta con la base de datos OLE DB creada mediante el ETL con la herramienta de minería de datos Rapidminer, y se toma diferentes datos clave con el fin de obtener datos relacionados y de esta forma poder tomar una decisión. Los datos clave escogidos para el análisis son: puerto, origen y prioridad, debido a que es la información más relevante dentro de los datos obtenidos. El objetivo es buscar información oculta que se relacione entres estos parámetros. Para asignar una variable como clave es necesario cambiar el tipo de

atributo a tipo label. Como se puede observar en las figuras 13 y 14 respectivamente.

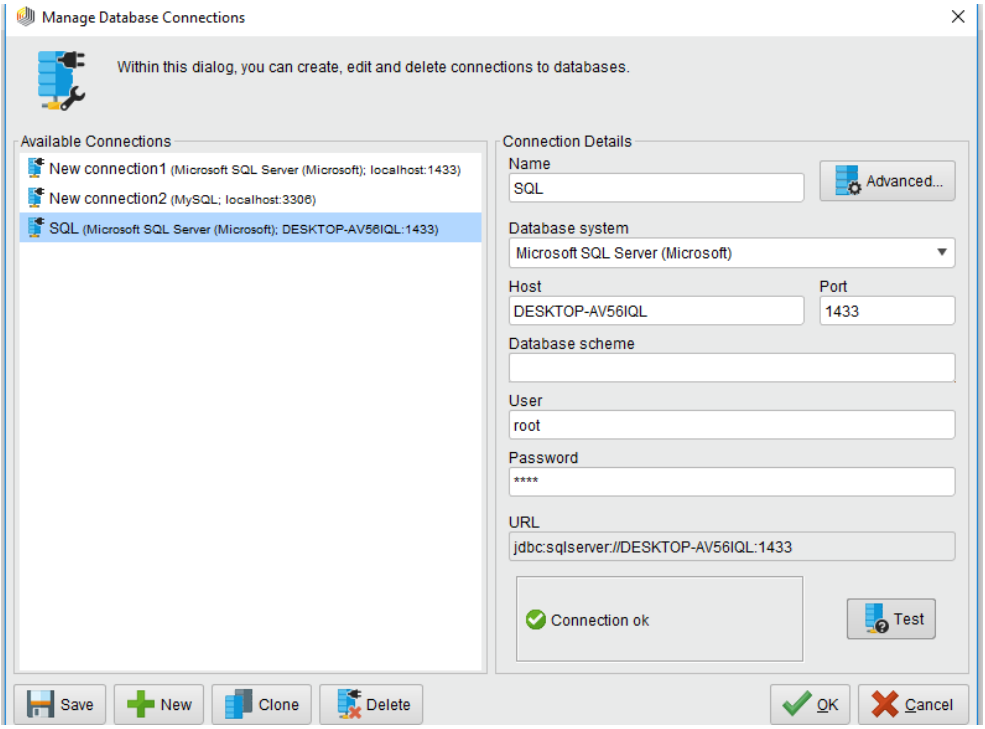


Figura 13. Conexión Rapidminer con la base de datos SQL.

Debido a que Rapidminer toma toda la información de la base de datos es necesario indicarle que atributo es el dato clave y asignarle el tipo label. Como se puede observar en la figura 14.

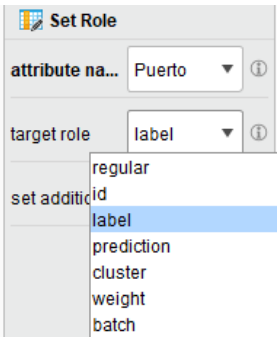


Figura 14. Asignación dato clave.

El dato clave puerto es escogido con el fin de entender cuál es el comportamiento de este dato en relación a los demás atributos dentro de la base de datos. Es decir, que información está relacionada entre sí y poder determinar mediante la predicción cuando un puerto es más o menos vulnerable dentro de un entorno empresarial, además de información adicional de importancia que no es conocida.

Al aplicar la técnica de árbol de decisiones se obtienen información relacionada entre el tipo de protocolo utilizado en el ataque, el tipo de acceso y la probabilidad de que puerto se utiliza para la intrusión, además de información relacionada con el incidente. Como se puede observar en la figura 15.

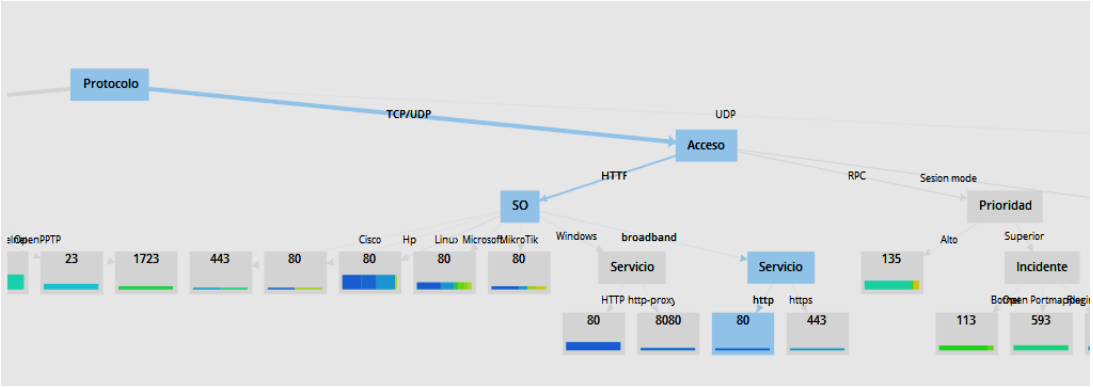


Figura 15. Probabilidad puerto vulnerable.

Al modificar el dato clave se obtiene información totalmente diferente, en este caso se toma la columna origen como dato clave, se espera obtener información de los países con mayores incidencias registradas y el tipo de ataques.

Al aplicar la técnica de minería de datos se obtiene información relacionada entre el sistema operativo del equipo vulnerable, la prioridad del incidente, el protocolo y la

probabilidad de que un país sea el origen de un ataque. Como se puede observar en la figura 16.

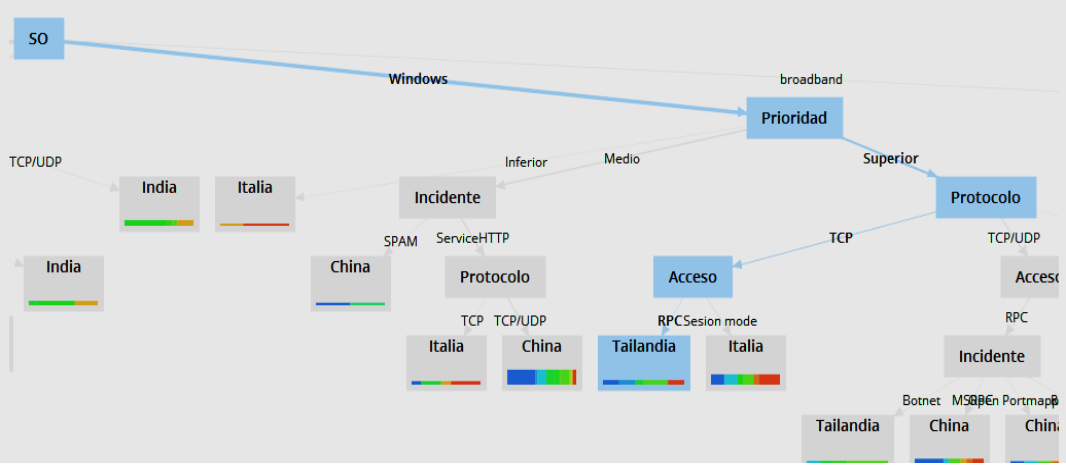


Figura 16. Probabilidad país origen de las amenazas.

Para el ejemplo se toma como dato clave el atributo origen, mediante el cual se espera obtener información que muestre el tipo de sistema operativo del dispositivo vulnerable y la forma en que el atacante toma control del mismo.

Al utilizar como dato clave el tipo de prioridad se encuentra una relación previamente desconocida, entre el sistema operativo, prioridad del incidente y el medio por el cual se toma control de este dispositivo. Como se observa en la figura 17.

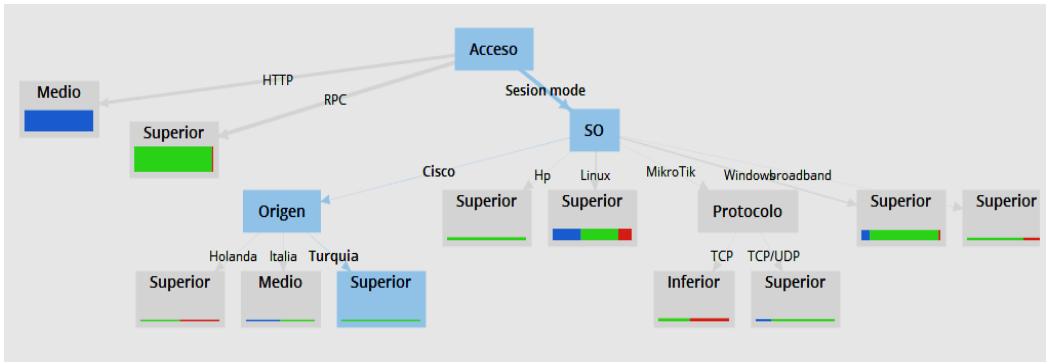


Figura 17. Probabilidad prioridad del incidente.

4.3 Análisis de grupos (Clustering)

Para la aplicación de esta técnica no es necesario escoger un dato clave, ya que toma la información global de la base de datos, y mediante algoritmos estadísticos logra agrupar información tomando algunos parámetros como: media, máximo, mínimo, etc. Existe una desventaja con esta técnica y es que, al utilizar argumentos estadísticos, mucha de la información se pierde y al final se obtiene un resumen de los datos ingresados.

Debido a que en esta técnica no es necesario cambiar los atributos de los campos, se ingresa directamente la información contenida en la base de datos OLE DB creada mediante el ETL y conectada al operador clustering. Como se puede observar en la figura 18.

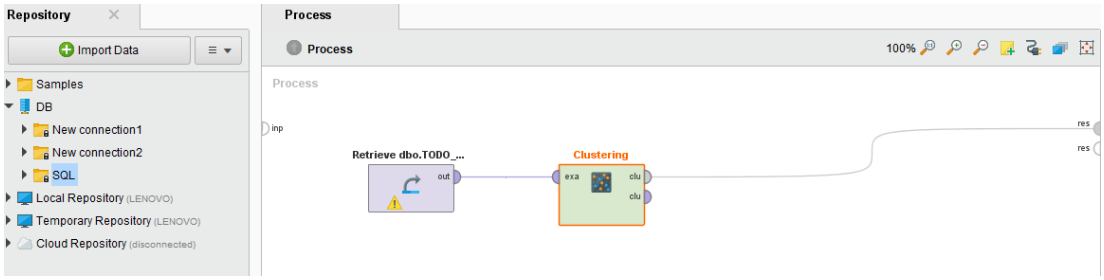


Figura 18. Clusteres Rapidminer.

Al ejecutar esta técnica de minería de datos, obtenemos gran cantidad de información relacionada al tipo de dato y dependiendo del caso la media, el máximo o la sumatoria de las variables. Tal como se puede observar en la figura 19.

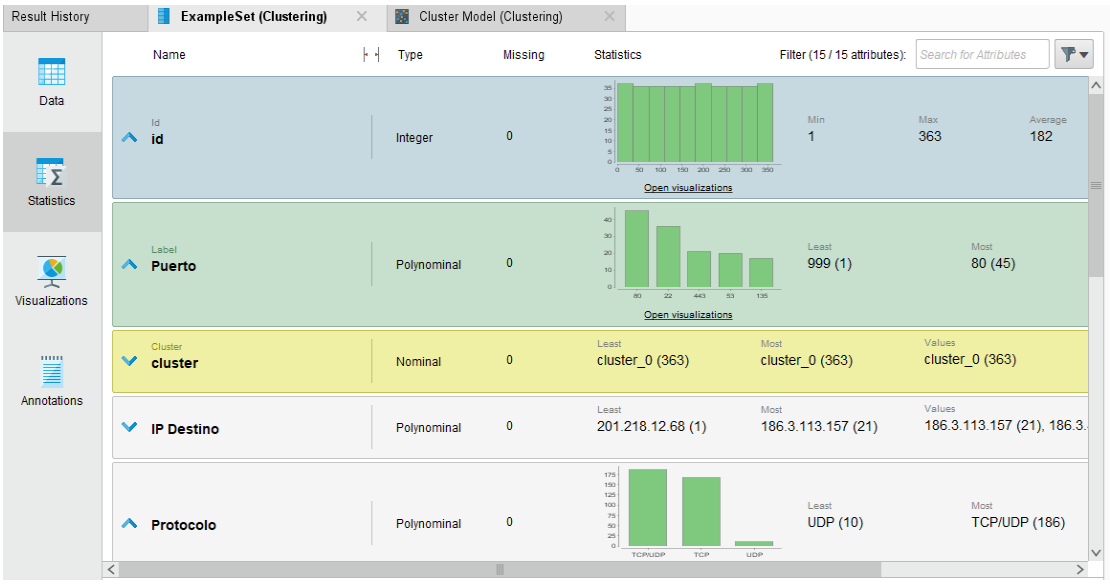


Figura 19. Puertos más y menos vulnerables.

Esta información está construida y relacionada manualmente por medio de gráficos, con la facilidad de agregar variables como sea necesario y de esta forma obtener información vital para el análisis. Como se puede observar en la figura 20.

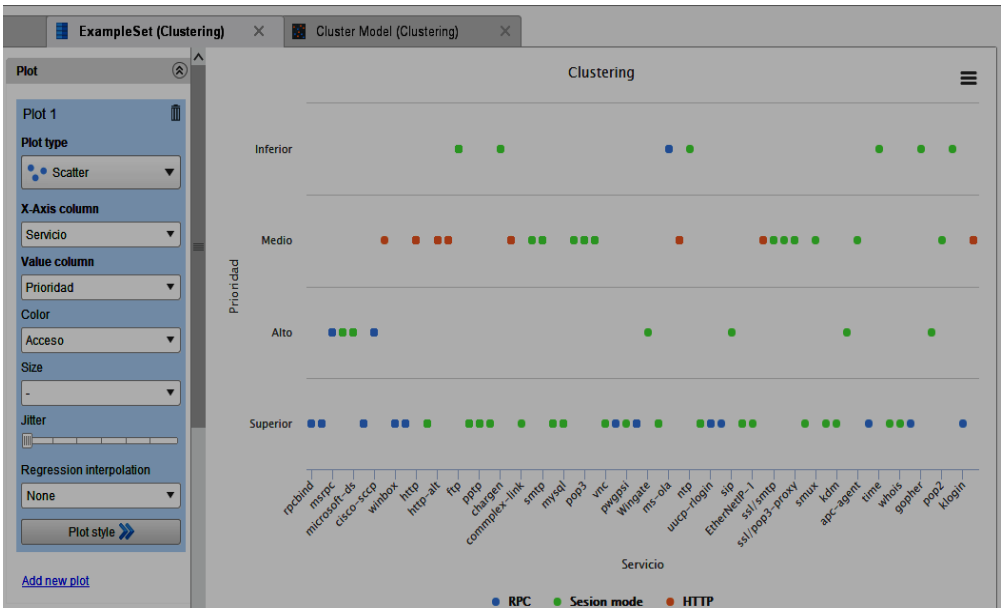


Figura 20. Incidentes relacionados con la prioridad del incidente y el tipo de acceso.

4.4 Clasificador Bayesiano (Naive Bayes)

Para la implementación de esta técnica de minería de datos, se toma la información de la base de datos OLE EB obtenida mediante el ETL. Debido a que esta técnica requiere un dato calve para su funcionamiento, se asigna el atributo de tipo label al dato clave. Al igual que en la técnica árbol de decisiones se utiliza las columnas puerto, origen y prioridad como datos claves y permite entender el comportamiento de las amenazas en el tiempo, gracias a la predicción.

Al utilizar el puerto como dato clave se espera obtener un dato de predicción sobre los puertos con mayor probabilidad de ataque, en base a una relación entre los datos dependientes. La información que se obtiene es la esperada, muestra la probabilidad porcentual que un puerto sea mayor o menor vulnerable a ataques, la técnica de minería de datos toma todos los parámetros de la tabla para definir esta probabilidad, además brinda información gráfica sobre el resultado de la obtenido, el cual permite realizar comparaciones con los demás atributos de la tabla. Como se puede observar en las figuras 21 y 22 respectivamente.

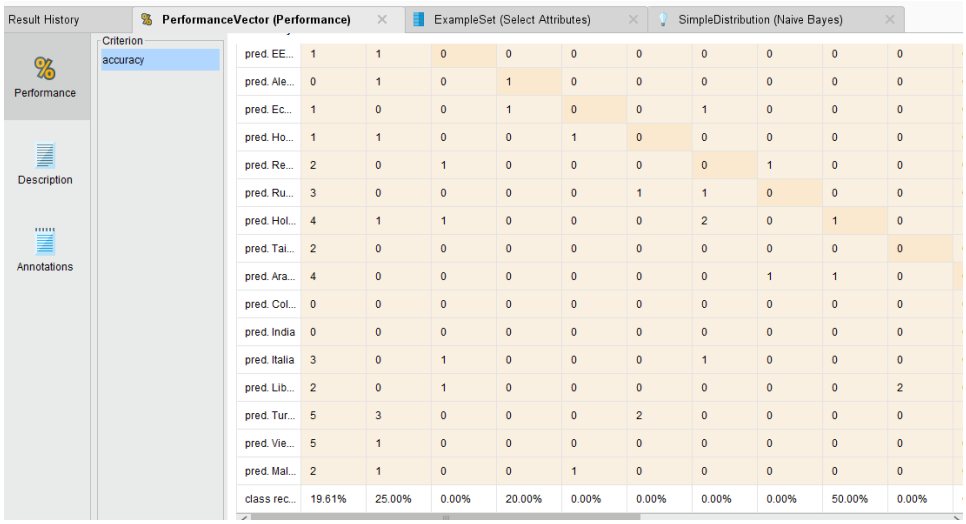


Figura 21. Probabilidad porcentual que un puerto pueda ser vulnerable.

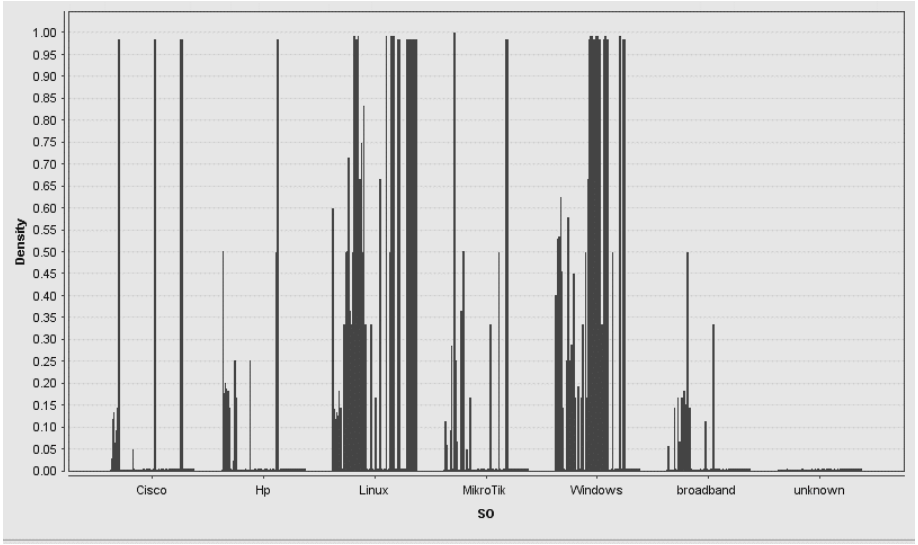
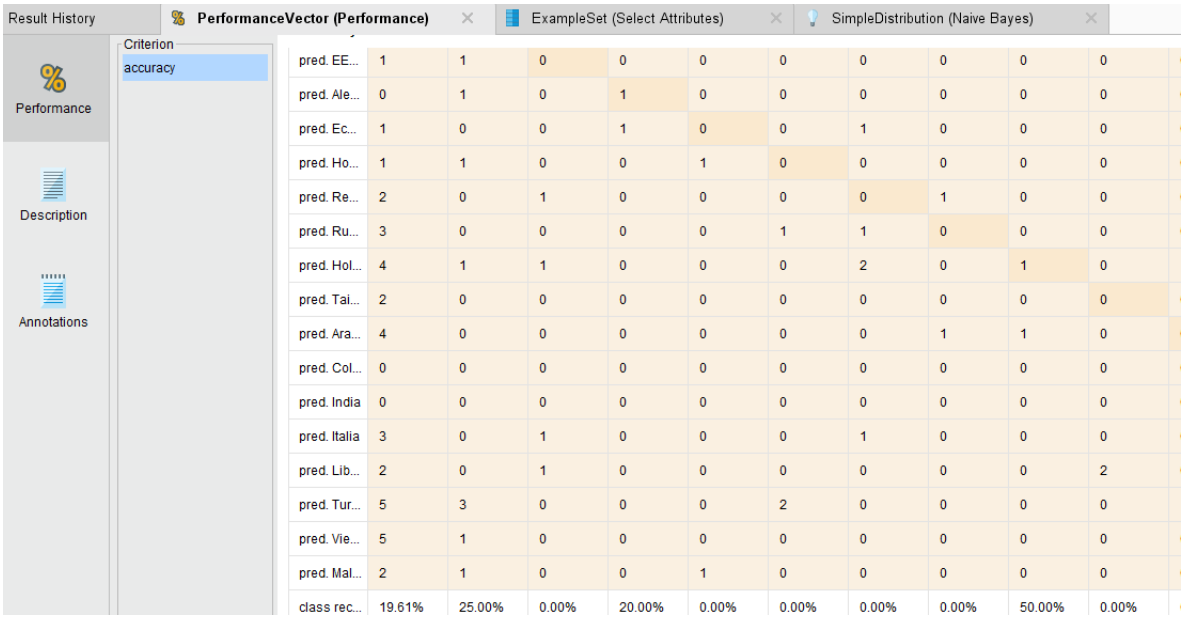


Figura 22. Probabilidad porcentual que un sistema operativo sea vulnerable.

Se obtiene una predicción la cual varia en base al atributo asignado, en este caso se ingresa como dato clave la variable origen, obteniendo información probabilística de los países con mayor tendencia a ser origen de los ataques. Como se puede observar en la figura 23.



pred. EE...	1	1	0	0	0	0	0	0	0	0
pred. Ale...	0	1	0	1	0	0	0	0	0	0
pred. Ec...	1	0	0	1	0	0	1	0	0	0
pred. Ho...	1	1	0	0	1	0	0	0	0	0
pred. Re...	2	0	1	0	0	0	0	1	0	0
pred. Ru...	3	0	0	0	0	1	1	0	0	0
pred. Hol...	4	1	1	0	0	0	2	0	1	0
pred. Tai...	2	0	0	0	0	0	0	0	0	0
pred. Ara...	4	0	0	0	0	0	0	1	1	0
pred. Col...	0	0	0	0	0	0	0	0	0	0
pred. India	0	0	0	0	0	0	0	0	0	0
pred. Italia	3	0	1	0	0	0	1	0	0	0
pred. Lib...	2	0	1	0	0	0	0	0	0	2
pred. Tur...	5	3	0	0	0	2	0	0	0	0
pred. Vie...	5	1	0	0	0	0	0	0	0	0
pred. Mal...	2	1	0	0	1	0	0	0	0	0
class rec...	19.61%	25.00%	0.00%	20.00%	0.00%	0.00%	0.00%	0.00%	50.00%	0.00%

Figura 23. Probabilidad porcentual que un país pueda ser origen de una amenaza.

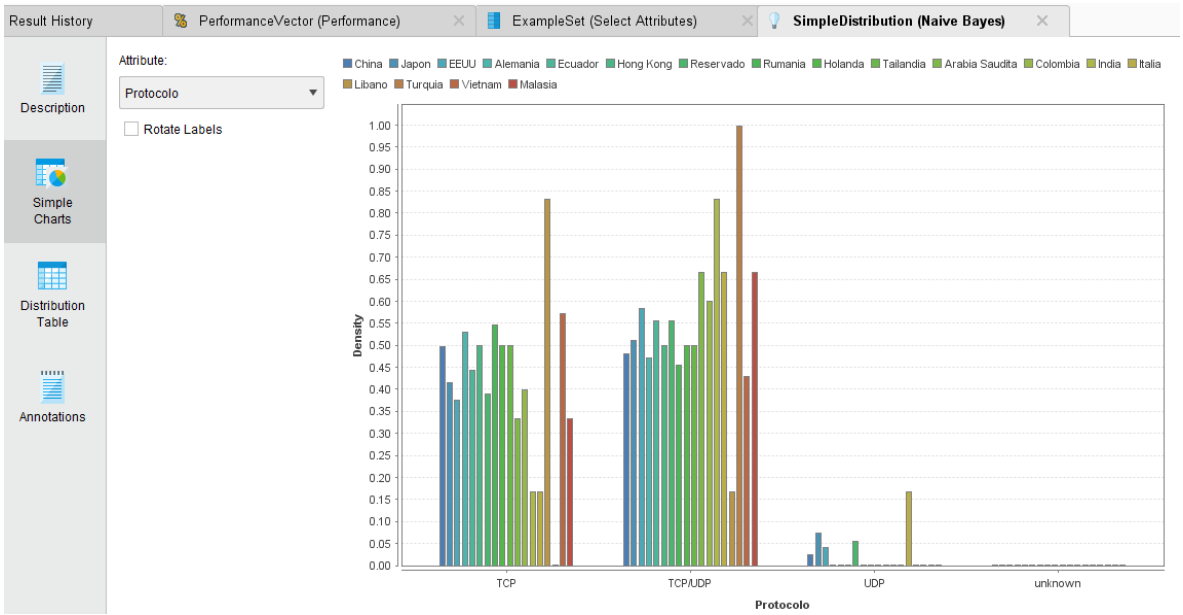


Figura 24. Vulnerabilidad por tipo de acceso y país.

4.5 Análisis de resultados

Tras implementar los modelos de minería de datos, se resume el análisis en dos grupos; El primero, la implementación del algoritmo árbol de decisiones y análisis de grupos para el estudio de puertos abiertos supone una gran alternativa ya que se logra obtener información relacional muy valiosa para el análisis; y el segundo, la técnica clasificador bayesiano resulta compleja de interpretar, pero con el correcto análisis muestra predicciones que ayudaran a concluir con las técnicas de mitigación a emplearse.

Para el desarrollo de esta tesis se ha considerado evaluar tres técnicas de minería de datos, utilizando los mismos atributos como variables en cada una, el objetivo es comparar los resultados y descubrir datos relacionados que permitan entender el comportamiento de las amenazas y vulnerabilidades que afectan a una red empresarial, mediante la predicción, decisión y agrupación para de esta forma proponer técnicas de mitigación.

4.5.1 Modelo árbol de decisiones

El análisis mediante el atributo puerto nos brinda gran cantidad de información, ya que la mayoría de datos se relacionan con este atributo; por lo cual se considerada importante excluir algunos datos que no son relevantes para el análisis, como son la fecha, hora de inicio y hora de fin del incidente, además de la dirección IP destino.

De esta forma se obtiene información relacionada de diferentes tipos de atributos, siendo el principal el protocolo, seguido del tipo de acceso y la prioridad. Debido a que el grafico de árbol de decisiones es demasiado extenso el análisis se lo realiza por partes, para el ejemplo se toma datos una ramificación de información relacionada entre el protocolo, la prioridad, el tipo de acceso, el modo se sesión, hasta llegar al incidente y puerto.

Mediante la lectura de la figura 25 se evidencia que cuando el protocolo utilizado para la intrusión es TCP, la incidencia es del tipo alto y el modo de acceso es sesión, lo más probable es que los equipos afectados sean de marca Cisco, HP, Linux o Windows, mediante los puertos 445 y 139 que corresponden a una incidencia del tipo Netbios.

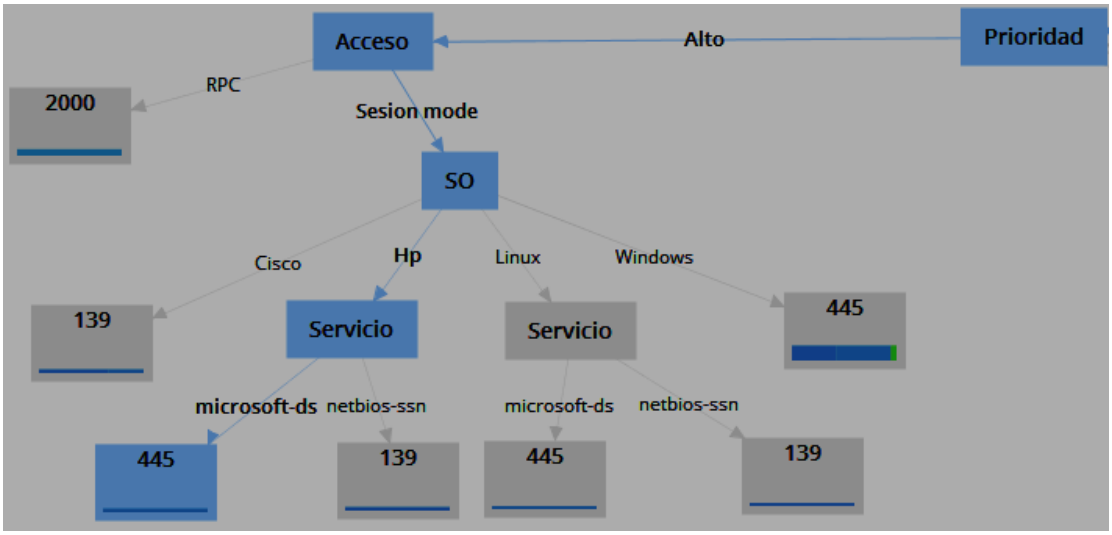


Figura 25. Ramificación árbol de decisiones prioridad alta.

En el grafico 26 se evidencia dos tipos de datos, el primero, cuando el protocolo es de tipo TCP y la prioridad es inferior, lo más probable es que los puertos vulnerables sean el 123, 21 y 2383, y que las vulnerabilidades sean Ntpversion, OPENTFTP o EternalBlue. El Segundo, cuando el protocolo es de tipo TCP, la prioridad es medio y el tipo de acceso es HTTP, la probabilidad es que los puertos 8000 y 3389 sean vulnerables. Como se puede observar en la figura 26.

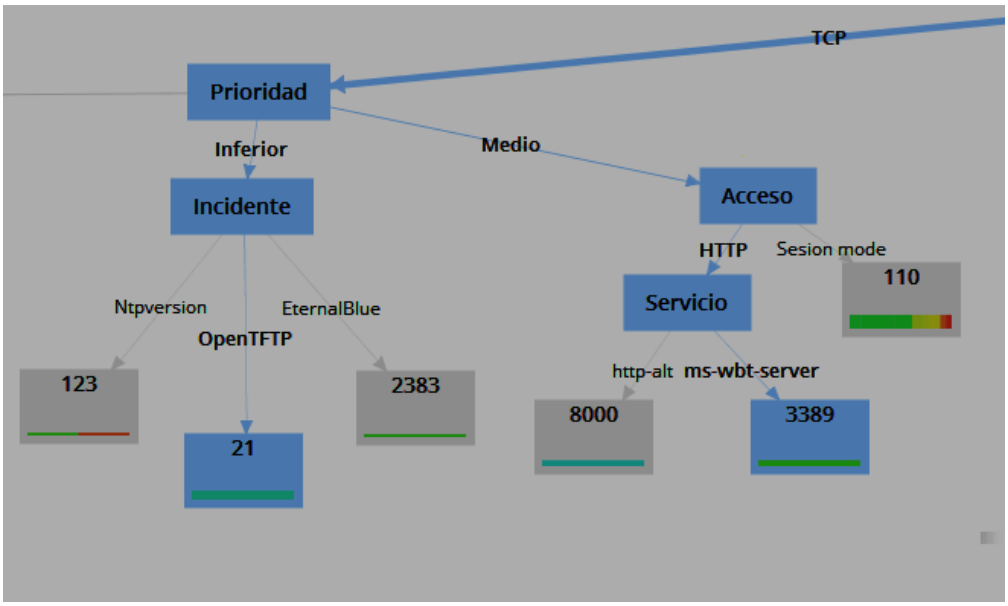


Figura 26. Ramificación árbol de decisiones prioridad inferior media.

En la figura 27 cuando el protocolo es de tipo TCP, la prioridad superior y el tipo de acceso RPC, lo más probable es que los puertos vulnerables sean el 22, 23 y 8291 correspondientes a un incidente del tipo acceso remoto.

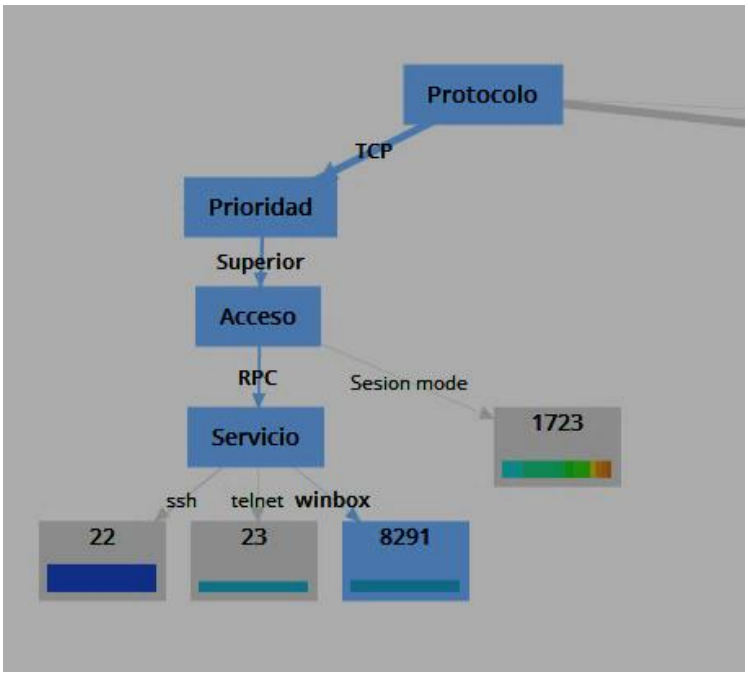


Figura 27. Ramificación árbol de decisiones prioridad superior.

El la figura 28 se observa información determinante para el análisis, mediante el acceso HTTP todos los sistemas operativos son vulnerables, lo que indica una gran probabilidad que este sea el modo de acceso más común al momento de irrumpir en una red, y probablemente los puertos que utilicen para la intrusión sean el 80 y el 443, por medio del servicio http o https.

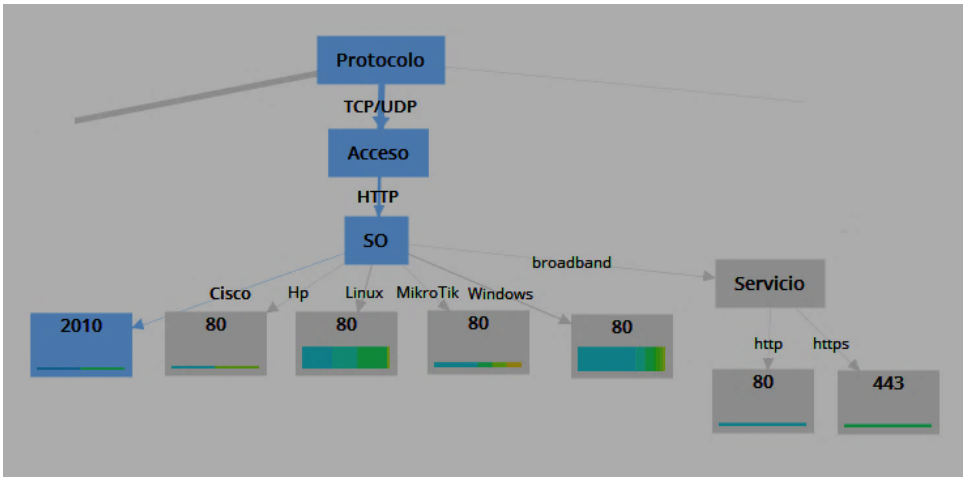


Figura 28. Ramificación árbol de decisiones acceso HTTP.

El análisis de la figura 29 proporciona la información más importante hasta el momento, ya que indica los incidentes más críticos que una red empresarial está expuesta. Cuando la amenaza utiliza los puertos TCP/UDP mediante el acceso RPC, la probabilidad más alta es que los puertos 113, 541 y 593 estén comprometidos, y que los incidentes sean del tipo botnet, Rlogin o Openmapper, con la prioridad más alta de atención.

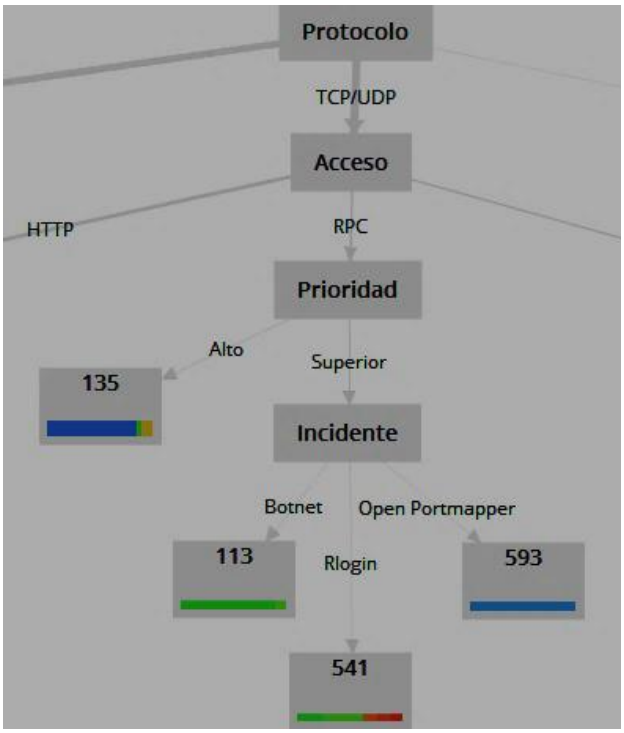


Figura 29. Ramificación árbol de decisiones acceso RPC.

La figura 30 proporciona información mediante el modo de acceso UDP, existe la probabilidad que se deba dar la prioridad más alta (superior) ya que los servicios utilizados para esta intrusión pueden desembocar en un incidente del tipo botnet o DoS.

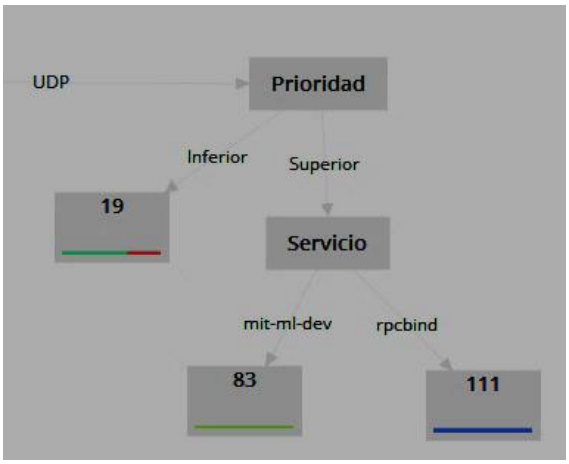


Figura 30. Ramificación árbol de decisiones protocolo UDP.

Como se puede observar en la tabla 5, después del análisis se obtiene la siguiente información: Los incidentes que utilizan RPC como medio de acceso, mediante los protocolos TCP, UDP o TCP/UDP son los que más riego llevan a la red. Por otro lado, los incidentes con prioridad medio y que utilizan HTTP como medio de acceso son los más comunes, y afectan a todos los dispositivos por igual, estos incidentes podrían o no comprometer la seguridad de la red.

Tabla 5.
Árbol de decisiones – Dato clave puerto

Prioridad	**Acceso**	**Definición**
Superior	RPC	Incidentes con mayor riego a la integridad de la seguridad Botnet, Rlogin y DoS.
Alto	Sesión	Incidente de riesgo considerable, afecta a equipos de las marcas como: Cisco, HP, Linux o Windows.
Medio	HTTP	Son los incidentes más comunes, pueden o no afectar la seguridad de la red. ServiceHTTP.
Inferior	Sesión	Incidentes menos comunes y con riego bajo. Ntpversion, OPENTFTP o EternalBlue.

Al aplicar la técnica de árbol de decisiones con el dato clave origen se obtiene demasiada información basura, por lo cual se excluyen los siguientes datos: fecha, hora de inicio y hora de fin, dirección IP y tipo de acceso. Estos datos no aportan información relevante para el análisis e incrementan datos inservibles al análisis. Al aplicar la variable origen como dato clave, la técnica de árbol de decisiones toma como información principal el sistema operativo de los dispositivos, seguidos por la prioridad del incidente los cuales están relacionados directamente con el origen de la amenaza.

Mediante el análisis de la figura 31 se identifica que existe una probabilidad alta que los ataques hacia equipos de marca Cisco provengan de China. Mientras que cuando la marca es HP, depende de la prioridad. Los países que más atacan a este traficante son: China, Japón y Alemania.

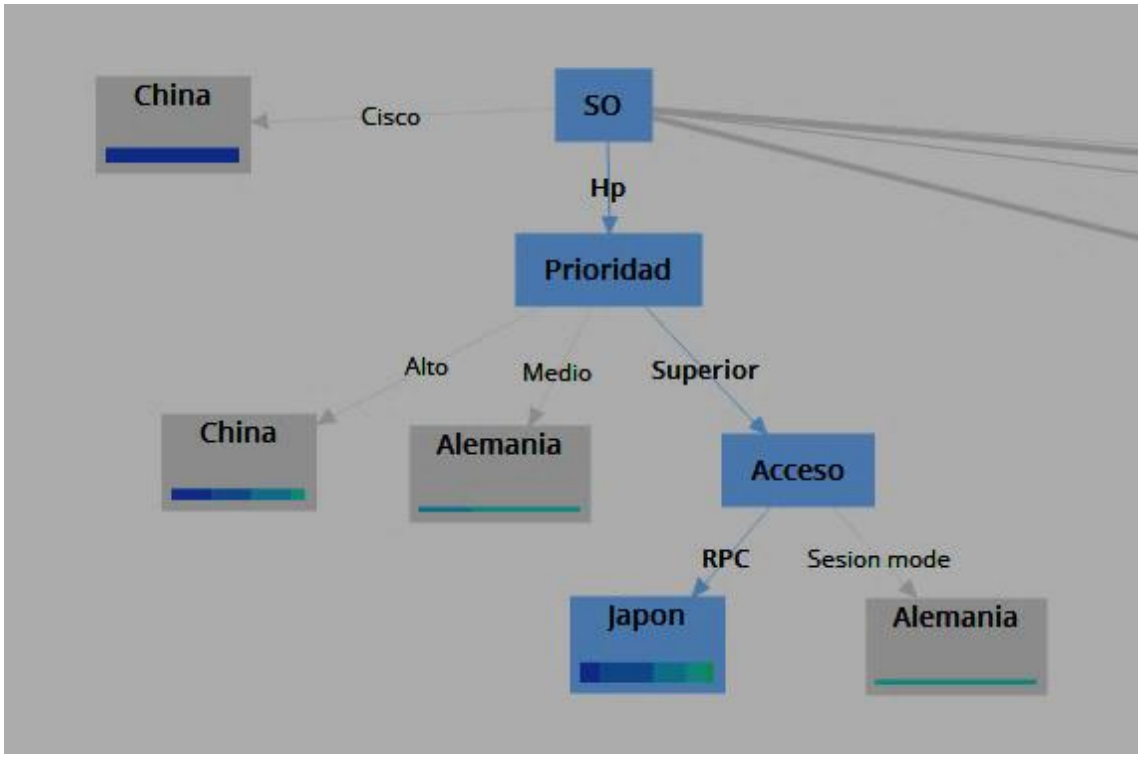


Figura 31. Ramificación árbol de decisiones sistema operativo HP y Cisco.

La figura 32 proporciona información acerca del sistema operativo Linux, el origen de las amenazas está dirigido por la prioridad del incidente. Para los incidentes con prioridad superior el país origen probablemente sea china o Japón, depende del protocolo. Para los incidentes con prioridad alta existe una gran probabilidad que el país origen de la amenaza sea China. Los incidentes de prioridad media, los países origen de las amenazas pueden ser Hong Kong o China la probabilidad depende del protocolo utilizado. Finalmente, si la prioridad es inferior existe la posibilidad que las amenazas provengan de China o su origen sea reservado.

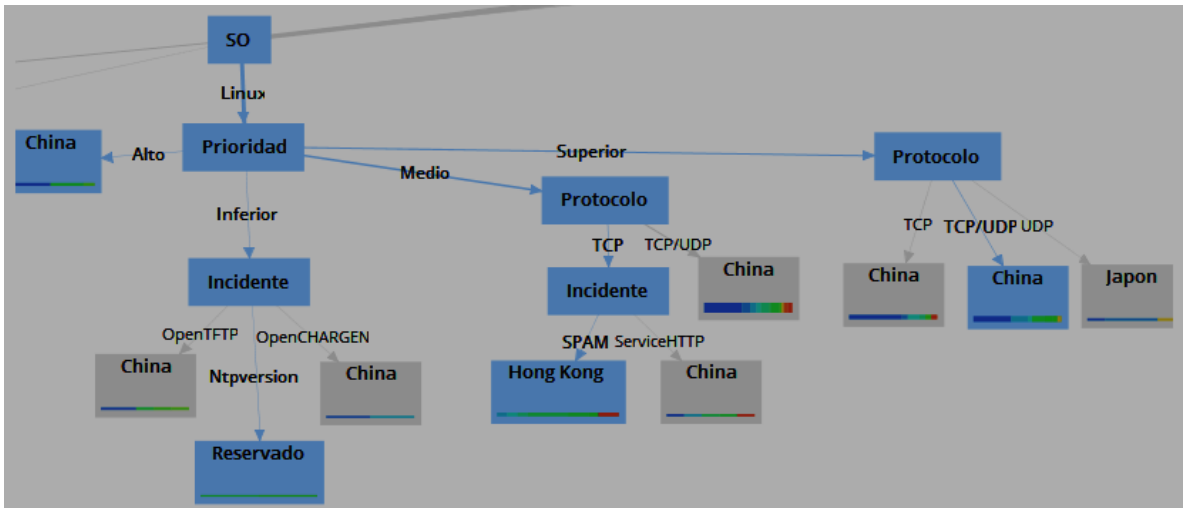


Figura 32. Ramificación árbol de decisiones sistema operativo Linux.

La figura 33 muestra que tanto los sistemas operativos MikroTik, Windows y Broadband tiene una probabilidad alta que el origen de las amenazas sea China.

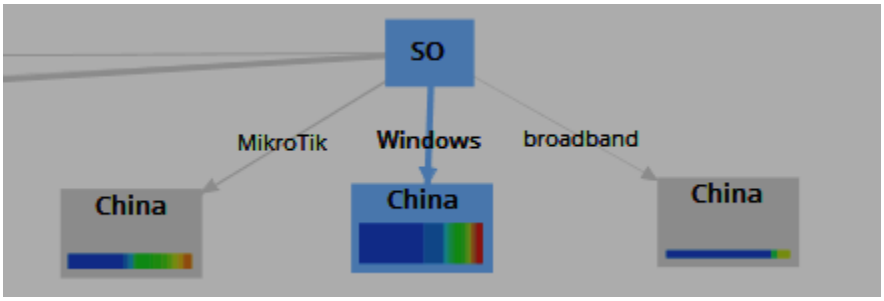


Figura 33. Ramificación árbol de decisiones sistema operativo Windows.

El país con mayor número de incidentes registrados para la muestra es China, seguido de Japón, Alemania y Hong Kong. Mientras que el sistema operativo más vulnerable es Windows.

Se toma a la prioridad como dato clave, ya que se espera obtener información del tipo de ataques más recurrentes dentro de una red empresarial.

Al analizar la figura 34 se obtiene información relacionada entre el modo de acceso, el sistema operativo y el origen. Para lo cual se determina que cuando el acceso es mediante HTTP, la probabilidad es que se trate de un incidente de prioridad media, cuando el acceso es RPC la probabilidad es que el incidente sea de prioridad superior. Mientras que cuando el acceso es mediante sesión la criticidad depende del sistema operativo, si la marca es Cisco o HP probablemente el incidente sea de prioridad alta, mientras que si las marcas son; Windows, Linux, MikroTik o Broadband el incidente puede ser de tipo superior.

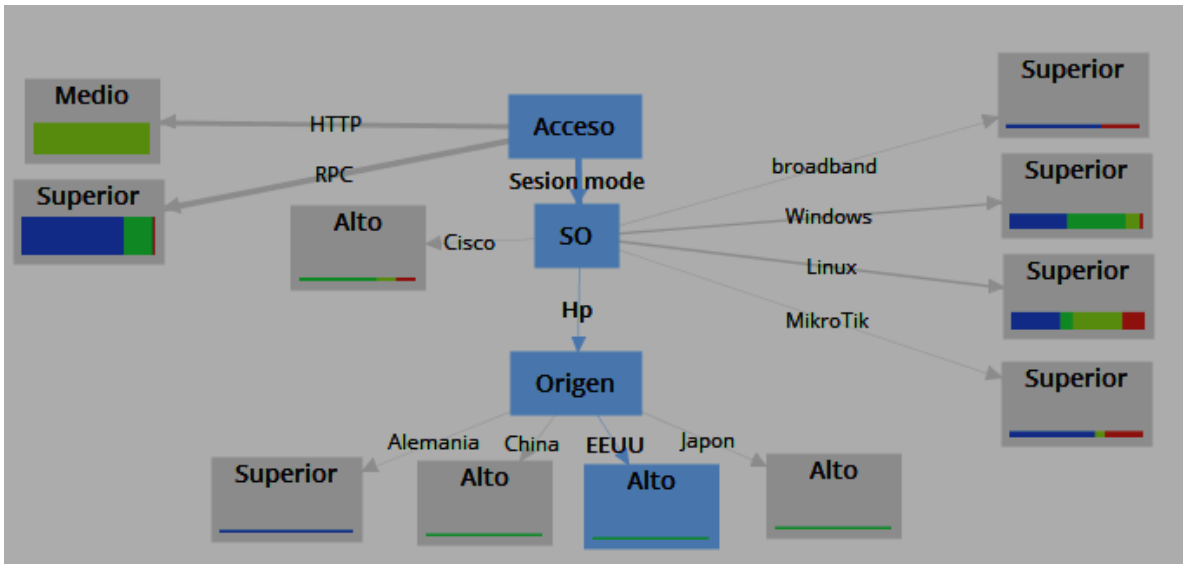


Figura 34. Ramificación árbol de decisiones nivel de prioridad según el SO y el país.

El análisis del dato clave prioridad indica que los sistemas operativos más

vulnerables son Linux, Mikrotik, Windows y Broadband con prioridad superior mientras que los sistemas operativos Cisco y HP tienen una prioridad Alta.

4.5.2 Análisis de Grupos (Clústeres)

La técnica de minería de datos Clústeres no requiere de un dato clave para la revisión, por lo cual analiza el total de datos ingresados, obteniendo la siguiente información.

Mediante el análisis de los resultados obtenidos se puede identificar que los puertos comúnmente atacado son el puerto 80 y 22, mientras que el menos atacado es el puerto 543. Como se puede observar en la figura 35.

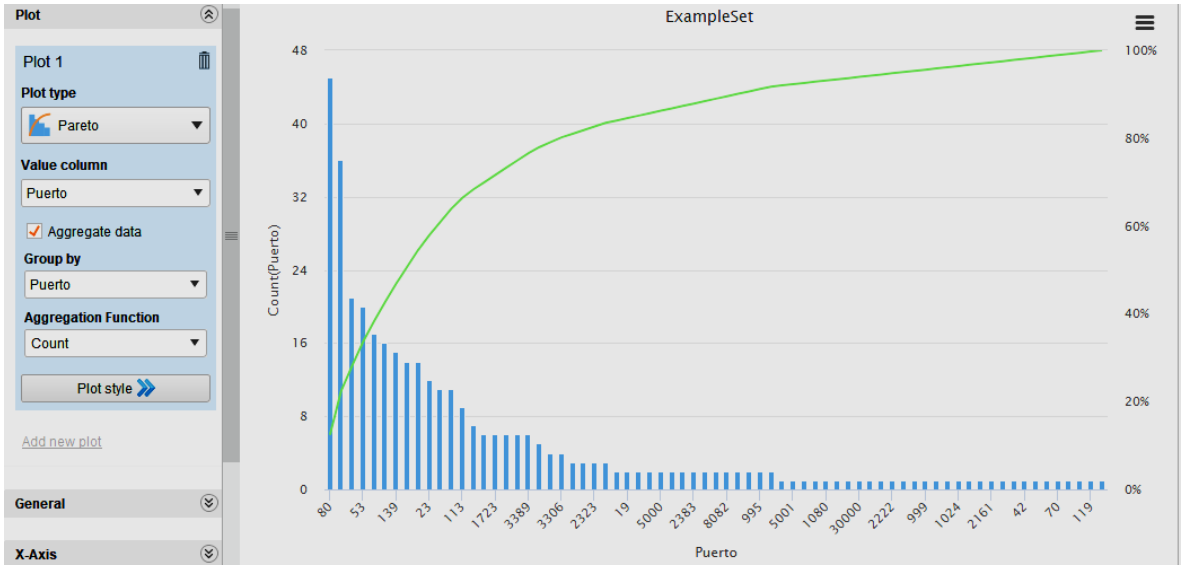


Figura 35. Puertos más y menos vulnerables.

Se evidencia que los ataques pueden venir de los protocolos TCP/UDP y es menos probable que se utilice el protocolo UDP para los ataques. Como se puede observar en la figura 36.

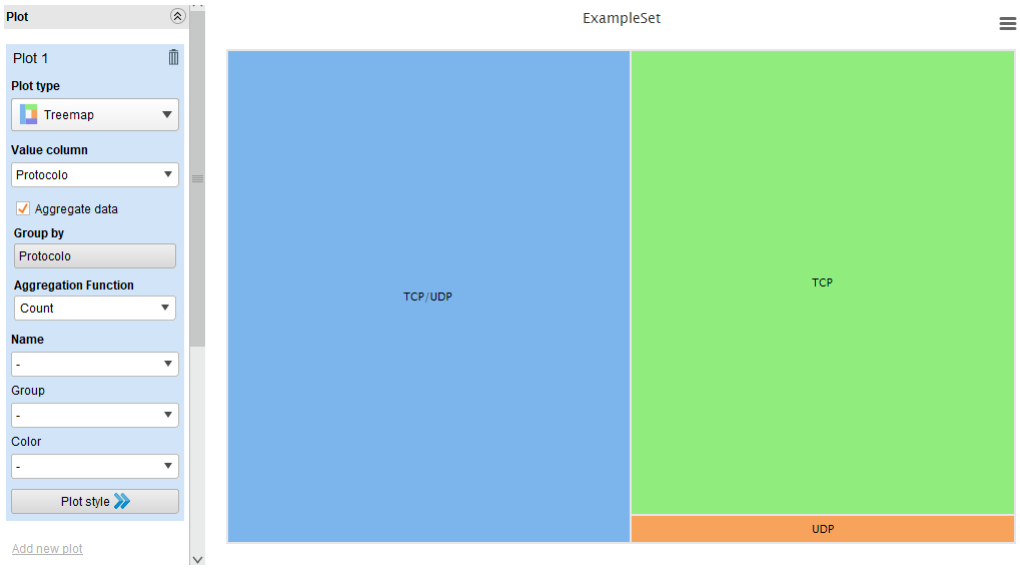


Figura 36. Protocolo utilizado para el ataque.

En la figura 37, el servicio más vulnerable es SSH, sin embargo, los servicios HHTP y HTTPS son los más frecuentes en presentar vulnerabilidades.

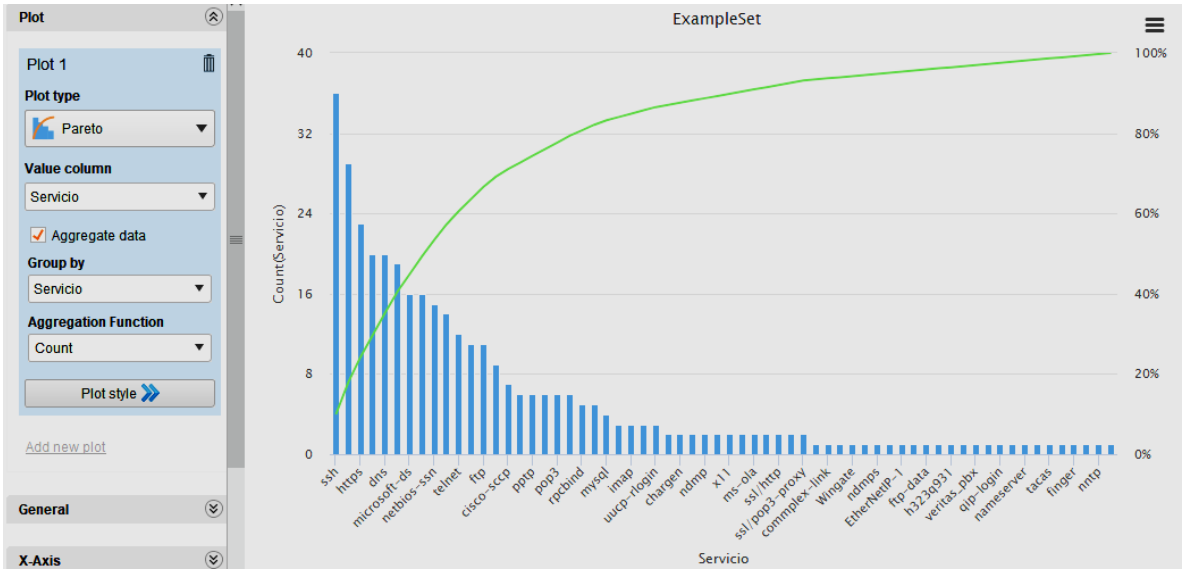


Figura 37. Servicios utilizados para la intrusión.

Según la figura 38, se puede concluir que el sistema operativo más vulnerable a ataques es Windows, mientras que el sistema operativo con menor probabilidad de ataques es Cisco.

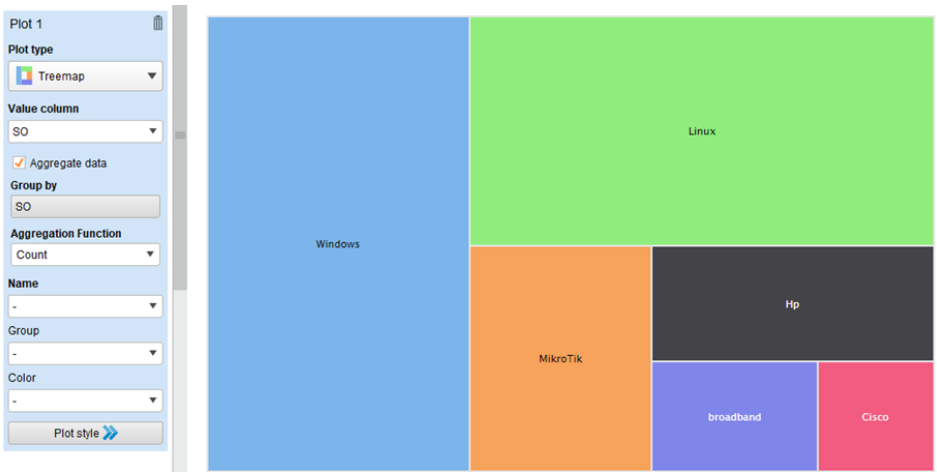


Figura 38. Sistemas operativos más vulnerables.

En la figura 39, se observa que China es el país con mayor probabilidad de ataques, seguido de Japón, mientras que Turquía representa como el origen menos probable.

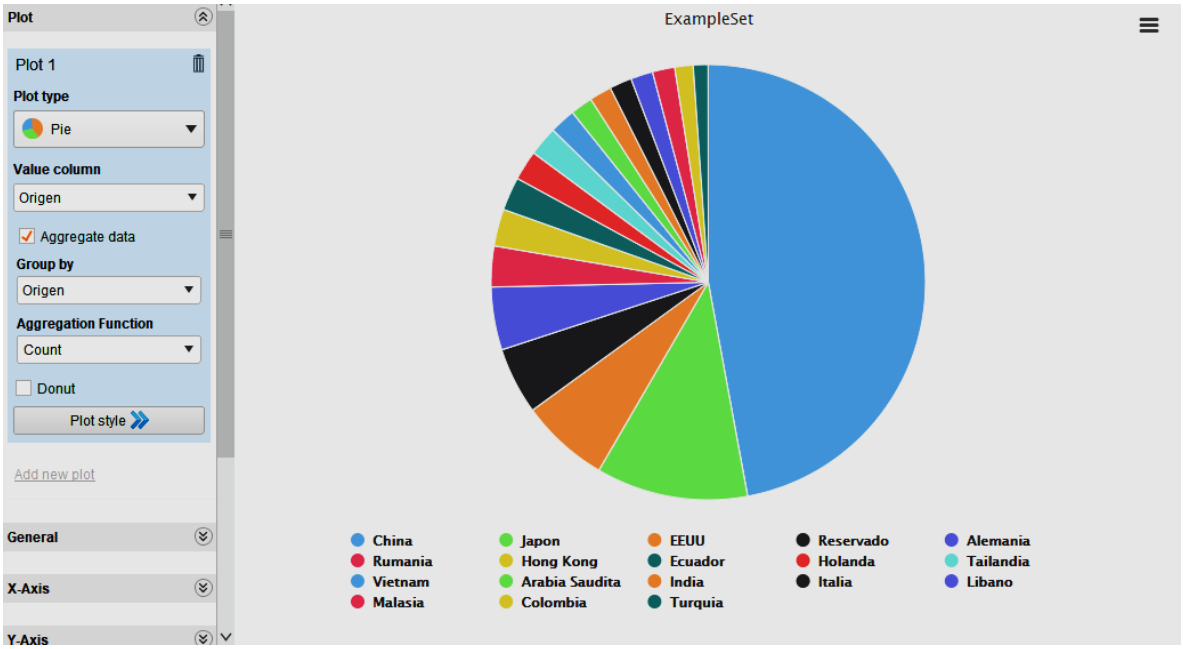


Figura 39. Países provenientes de las amenazas.

La figura 40 muestra información de los incidentes con mayor probabilidad de ataque, la cual tiene relación con el dato servicio. Como se puede observar el

incidente con mayor probabilidad de ataque es Services HTTP, mientras que el servicio menos usado es OpenProxy.

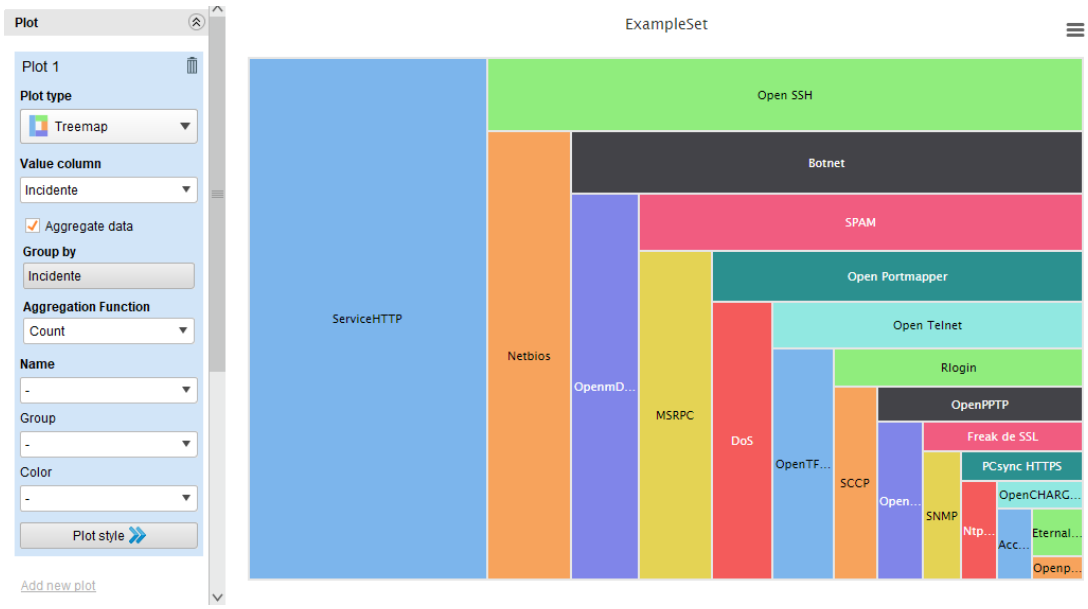


Figura 40. Incidentes más vulnerables.

Los incidentes con prioridad superior son los más registrados, seguidos por los incidentes de prioridad medio, incidentes de prioridad alta y finalmente inferior. Como se puede observar en la figura 41.

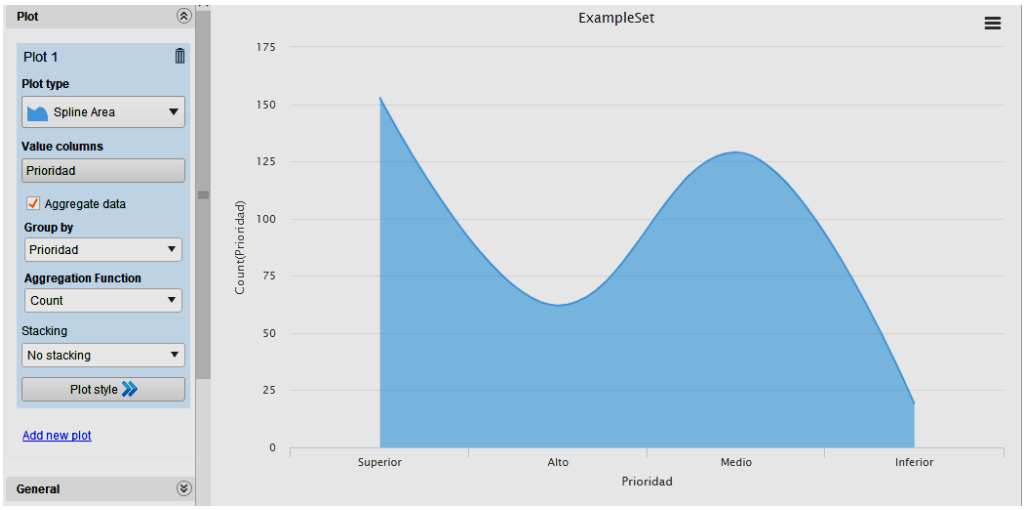


Figura 41. Prioridad de los incidentes registrados.

4.5.3 Análisis Clasificador Bayesiano (Naive Bayes)

La información obtenida al aplicar puerto como dato clave es diversa, pero se puede afirmar que el puerto 22 es el más vulnerables en una red, seguidos por el puerto 80 con un 71,43% de probabilidad, el puerto 21 con un 66,67% de probabilidad, el puerto 135 con un 40% de probabilidad, el puerto 8080 con un 25% de probabilidad y finalmente el puerto 53 con un 16,67% de probabilidad. Como se puede observar la figura 42.

	true 111	true 22	true 135	true 80	true 554	true 8000	true 8080	true 21	true 53
pred. 111	0	0	0	0	0	0	0	0	0
pred. 22	1	10	0	0	0	3	0	0	4
pred. 135	0	0	2	3	5	0	2	0	0
pred. 139	0	0	0	0	0	0	0	0	0
pred. 445	0	0	2	1	0	0	0	0	0
pred. 593	0	0	0	0	0	0	0	0	0
pred. 2000	0	0	0	0	0	0	0	0	0
pred. 2010	0	0	0	0	0	0	0	0	0
pred. 8291	0	1	0	0	0	0	0	0	0
pred. 23	0	0	0	0	0	0	0	0	0
pred. 80	0	0	0	0	0	0	0	0	0
pred. 554	0	0	0	0	0	0	0	0	0
pred. 8000	0	0	0	0	0	0	0	0	0
pred. 119	0	0	0	0	0	0	0	0	0
pred. 543	0	0	0	0	0	0	0	0	0
class rec...	0.00%	90.91%	40.00%	71.43%	0.00%	0.00%	25.00%	66.67%	16.67%

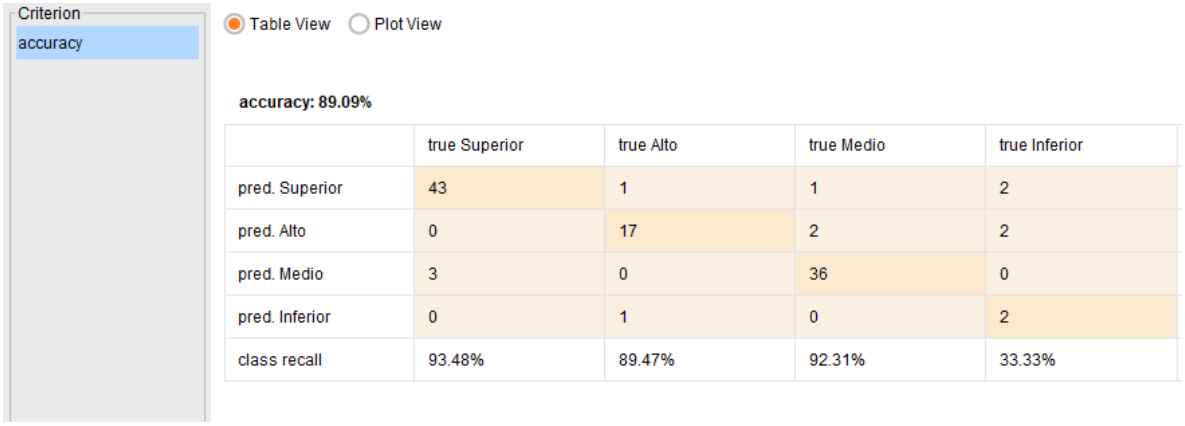
Figura 42. Probabilidad porcentual puerto más vulnerable.

La probabilidad en términos porcentuales indica que los países de los cuales pudieran provenir una amenaza son: China 62,75%, Japón 41,67%, Alemania 40% y Ecuador 33,3%. Como se puede observar en la figura 43.

	true China	true Japon	true EEUU	true Ale...	true Ecu...	true Hon...	true Res...	true Ru...	true Hola...
pred. Chi...	32	0	0	0	2	0	0	0	1
pred. Ja...	1	5	2	2	0	0	0	0	0
pred. EE...	0	1	0	0	0	0	0	0	0
pred. Ale...	1	1	0	2	0	0	0	0	0
pred. Ec...	1	0	0	0	1	0	0	0	0
pred. Ho...	0	1	0	1	0	0	0	0	0
pred. Re...	2	0	1	0	0	1	0	1	0
pred. Ru...	2	0	0	0	0	1	2	0	0
pred. Hol...	0	0	1	0	0	0	2	0	0
pred. Tai...	0	0	0	0	0	0	0	2	0
pred. Ara...	2	0	0	0	0	0	0	0	1
pred. Col...	0	0	0	0	0	0	0	0	0
pred. India	1	0	0	0	0	0	0	0	0
pred. Vie...	2	1	0	0	0	0	0	0	0
pred. Mal...	1	0	0	0	0	0	0	0	0
class rec...	62.75%	41.67%	0.00%	40.00%	33.33%	0.00%	0.00%	0.00%	0.00%

Figura 43. Probabilidad porcentual país origen de las amenazas.

Los incidentes tienen el 93,48% de probabilidad de ser de orden superior, el 92,31 de ser de orden Medio, el 89, 47% de ser de orden Alto y el 33,33% de ser inferior. Según el grafico 44.



	true Superior	true Alto	true Medio	true Inferior
pred. Superior	43	1	1	2
pred. Alto	0	17	2	2
pred. Medio	3	0	36	0
pred. Inferior	0	1	0	2
class recall	93.48%	89.47%	92.31%	33.33%

Figura 44. Probabilidad porcentual prioridad del incidente.

4.5.4 Comparación de resultados

Mediante el análisis de las tres técnicas de minería de datos árbol de decisiones, análisis de grupo y clasificador bayesiano, y utilizando los mismos parámetros en la ejecución se obtiene información de probabilidad y predicción relacionadas entre sí.

Al ejecutar como dato clave el atributo puerto se obtiene información diversa, por un lado, la técnica análisis de grupo muestra los puertos más y menos vulnerables utilizando la función matemática máximo y mínimo, además de los servicios afectados por medio de estos puertos. De esta forma existe la probabilidad que los puertos más vulnerables sean el puerto 80, 8080 y 22 mientras que los menos vulnerables sean los puertos 543, 119 y 70. Al aplicar la técnica clasificador bayesiano se obtiene la probabilidad porcentual de que un resultado sea más o menos probable de suceder, para esto analizando el dato clave puerto se obtiene que los puertos más probables de ser atacados son: el puerto 22 con un 90% de probabilidad, seguido por el puerto 80 con el 71,43%, mientras que el menos vulnerable es 53 con el 16.67%. Árbol de decisiones proporciona información de predicción en base a varios parámetros, como son: prioridad, modo de acceso, S.O, puerto, servicio, ect. Mediante la combinación de estos parámetros es posible predecir diversos escenarios sobre qué puertos pueden ser más o menos vulnerables.

El análisis del dato clave origen mediante la técnica análisis de grupos se obtiene que el país con mayor probabilidad de ataques es China mientras que el menos probable es Turquía. La técnica análisis clasificador bayesiano se obtiene que existe la probabilidad porcentual del 62.75% que China sea origen del ataque, por otro lado, con un 33% la probabilidad de ataque viene de Ecuador. Mediante la implementación de la técnica árbol de decisiones se observa que el país más propenso a gestionar ataques es China, para lo cual utiliza la combinación de varios parámetros como prioridad, S.O, modo de acceso, etc.

Se evalúa el dato clave prioridad mediante la técnica análisis de grupos verificando que los incidentes comúnmente son de orden superior, seguidos por prioridad media, alto y finalmente inferior. Clasificador bayesiano muestra información porcentual de la prioridad de los incidentes obteniendo 93.48% de probabilidad que el incidente sea de orden superior, 92.31% medio, 89.47% alto y 33.33% inferior. En cuanto a árbol de decisiones la prioridad del incidente generalmente será superior y las demás prioridades dependerán de parámetros como S.O, modo de acceso, origen, etc.

4.5.5 Resumen

Mediante la comparación de las tres técnicas de minería de datos sobre el dato clave puerto se obtiene el mismo resultado, el servicio por el cual el atacante logra el acceso a una red empresarial es HTTP mediante el puerto 80. Lo que indica que la mayor parte de amenazas detectadas dentro de una red empresarial son atribuidas al usuario final, por desconocimiento o falta de capacitación.

El análisis del dato clave origen muestra que el país más propenso a ser origen de los ataques es China, mientras que no existe una claridad en el país menos propenso a ser origen de ataques, ya que podría originarse de diversas fuentes.

El incidente más recurrente dentro de una red empresarial es de orden superior, seguida por medio, alto y finalmente inferior, lo que indica que la mayoría de los incidentes registrados deben ser tratados en los 5 primeros minutos de ser detectados, puesto que pueden suponer una amenaza critica para la organización.

Tabla 6.

Comparación de resultados

Técnica	**Puerto**	**Origen**	**Prioridad**
Análisis de grupos	80, 22	China, Japón	S,M,A,I
Clasificador Bayesiano	80, 22	China, Japón	S,M,A,I
Árbol de decisiones	80, 443	China, Japón	S,M,A,I

Nota. S = Superior; A = Alto; M = Medio e I = Inferior

4.6 Técnicas de mitigación

- Los incidentes de seguridad son inherentes a las redes empresariales, por lo cual es necesario establecer la forma de gestionarlos, para lo cual se recomienda proveer un tiempo de respuesta y asignar una prioridad a cada incidente. Se recomienda: incidentes con prioridad superior 5 minutos, alto 15 minutos, medio 30 minutos, bajo 1 hora e inferior 3 horas.
- Debido a que el servicio más comprometido es HTTP y HTTPS mediante los puertos 80 y 443 que corresponde a un incidente del tipo ServicesHTTP, es decir un intento de instrucción mediante el acceso web, lo que quiere decir que la amenaza proviene dentro de una red empresarial es el usuario final y la falta de seguridad en la red. Se recomienda la instalación de un equipo de seguridad perimetral, con el objetivo de restringir el acceso a ciertos sitios que puedan comprometer la seguridad de la red. Además, crear niveles que limiten el acceso a internet a los usuarios evitando de esta forma ataques por desconocimiento.

- La prioridad de los incidentes se da por el nivel en que la seguridad de la red se pude ver comprometida, por lo cual es necesario aplicar listas de acceso que denieguen el paso a los puertos que pudieran comprometer la seguridad de la red, especialmente los puertos con prioridad superior y acceso RPC (113, 541 y 593) los cuales son incidentes del tipo Botnet o DoS. Los puertos con prioridad inferior o media dependen del criterio del administrador ya que algunos puertos pueden ser utilizados por algún equipo o aplicación específica dentro de la red empresarial.
- Algunos servicios son necesarios en las redes empresariales, tales como: Telnet, SSH, DNS, etc. Estos pueden resultar como focos de ataques por ser puertos conocidos, se plantea el uso de puertos alternativos como una solución al problema de vulnerabilidad.
- Dado que el sistema operativo más vulnerable es Windows se recomienda mantener instaladas las actualizaciones del software, así como los parques de seguridad que el sistema operativo requiera. De esta manera se mitigará los ataques a este sistema operativo.
- Los usuarios finales suelen ser blanco de ataques, ya sea por falta de capacitación o poca seguridad en sus dispositivos conectados a Internet, por esta razón se recomienda impedir el acceso a dispositivos USB a las estaciones de trabajo, mediante la instalación de un antivirus licenciado. Además, de la capacitación constante al personal sobre las amenazas provenientes de internet y sus consecuencias.

5. CONCLUSIONES Y RECOMENDACIONES

5.1 Conclusiones

Los incidentes de seguridad son amenazas centralizadas provenientes de Internet que afectan la seguridad de las redes empresariales causando pérdidas de información, daño del equipamiento informático o indisponibilidad de la red, lo que conlleva perjuicio económico a las empresas. Por esta razón es necesario identificar los incidentes que mayormente afectan a estas redes con el fin de proponer técnicas de mitigación en un tiempo efectivo.

Mediante el análisis y comparación de las técnicas de minería de datos utilizando la herramienta Rapidminer, se logra identificar cuáles son los ataques más comunes que afectan a las redes empresariales y las vulnerabilidades a las cuales están expuestas, mediante el análisis de los logs recolectados de equipos de seguridad se identifica que las amenazas más comunes a las cuales están expuestas las redes empresariales son *ServicesHTTP* y son introducidas a la red por medio del usuario final, mediante el acceso a sitios no confiables o a través del *clickeo* en banners de publicidad, etc. Convirtiéndose de esta forma en la vulnerabilidad más importante dentro de una red empresarial, por desconocimiento o falta de capacitación.

La técnica árbol de decisiones es la técnica que mejor se adapta al estudio realizado, puesto que muestra información desglosada en base a los parámetros que intervienen en el análisis, y de esta forma permite tomar una decisión en base a gran cantidad de información obtenida, reduciendo significativamente la tasa de error. Por otro lado, la técnica más amigable al usuario es análisis de grupos ya que muestra de forma gráfica y explicita los resultados del análisis y permite al usuario obtener una visión rápida del estado de los datos ingresados para el análisis.

La mayoría de intentos de intrusión identificados provienen del servicio HTTP, lo que alerta la necesidad de restringir o limitar el acceso a Internet a los usuarios,

además, los puertos abiertos representan una brecha de seguridad importante ya que dependiendo de la criticidad del incidente puede comprometer la seguridad de la red.

Las vulnerabilidades dentro de las redes empresariales son identificadas mediante los puertos abiertos dentro de la infraestructura de red, por falencias en la seguridad implementada, así como por un mal uso del equipamiento informático por parte del usuario final, es por esta razón la necesidad de mantener al personal capacitado en el correcto uso del equipamiento asignado, para evitar intrusiones inesperadas dentro de la red.

Los incidentes de seguridad son innatos en las redes empresariales, es por esto que eliminarlos por completo es imposible. Mitigar las incidencias de seguridad en un tiempo efectivo es primordial, puesto que se evita la perdida de información y pérdidas económicas en las empresas.

La evaluación de los resultados obtenidos al aplicar las técnicas de minería de datos empleadas para el descubrimiento de información ha permitido proponer de mejor manera las técnicas de mitigación a emplearse dependiendo del equipo afectado, así como del tipo de incidente. Para el caso de los equipos se identifica que el sistema operativo más vulnerable es Windows por lo cual se plantea la necesidad de tener equipos licenciados y actualizados a su última versión, en cuanto a los incidentes las técnicas son diversas desde colocar ACL en los equipos que impidan el paso de determinados puertos o la utilización de puertos alternativos en servicios conocidos.

5.2 Recomendaciones

El ETL usado para la transformación de información es escalable, lo que permite agregar nuevos logs de equipos de seguridad diferentes a los ingresados para este análisis, mediante la importación de la información al ETL.

Existen diversos software de minería de datos en el mercado, siendo uno de los más versátiles Rapidminner además de ser de código libre, sin embargo, se recomienda aplicar minería de datos mediante la herramienta SQL Server Analysis Services con el fin de simplificar la extracción de información entre la base de datos y la aplicación SQL. Para el estudio no se contempló esta herramienta ya que requiere licenciamiento para su completo funcionamiento.

El uso de diferentes técnicas de minería de datos para evaluar los resultados es una alternativa importante para tener mayor detalle de las vulnerabilidades de seguridad que estén afectando la red, para este estudio se han tomado las técnicas más utilizadas para la predicción y clasificación, sin embargo, existen otras técnicas que pudieran aplicarse para la evaluar resultados.

REFERENCIAS

Bavirisetty, M. (2015). *Advanced Analytics - Frameworks, Platforms and Metholodologies*. Recuperado el 30 de mayo de 2019, de https://www.slideshare.net/MohanBavirisetty/advanced-analytics-frameworks-platforms-and-metholodologies-v-10

BI para todos. (2017). Crear ETL Básica - Cargar archivo plano. Recuperado el 13 de mayo de 2019, de https://www.youtube.com/watch?v=y_WrDeR2qFY

Blogspot. (2009). Algoritmo *Naive Bayes*. Recuperado el 22 de mayo de 2019, de http://algoritmosmineriadatos.blogspot.com/2009/12/algoritmo-naive-bayes.html

CIBSEGC. (s.f). Gestión de los incidentes de seguridad. Recuperado el 21 de marzo de 2019, de http://e-forma.kzgunea.eus/mod/book/view.php?id=9929

CCN. (2017). Criterios comunes para la Gestión de Incidentes de Seguridad en el Esquema Nacional de Seguridad (ENS). Recuperado el 25 de marzo de 2019, de https://www.ccn-cert.cni.es/publico/seriesCCN-STIC/series/800-

CCN. (2018). Esquema nacional de seguridad gestión de *ciber* incidentes. Recuperado el 05 de abril de 2019, de https://www.ccn-cert.cni.es/series-ccn-stic/800-guia-esquema-nacional-de-seguridad/988-ccn-stic-817-gestion-de-ciberincidentes/file.html

Chunga, J. (2013). La minería de datos. Recuperado el 22 de mayo de 2019, de https://www.infotecarios.com/la-mineria-de-datos/#.XS58wOtKjIU

Cisco Systems. (2013). Infraestructura de redes empresariales Cisco ONE: la base automatizada y centrada en las aplicaciones para la empresa moderna. Recuperado el 01 de mayo de 2019, de https://www.cisco.com/c/dam/global/es_es/assets/pdf/en-04_en-white-paper_wp_cte_es.pdf

Cisco Systems. (2013). La infraestructura de redes empresariales Cisco ONE hace posible la transformación empresarial. Recuperado el 01 de mayo de 2019, de https://www.cisco.com/c/dam/global/es_es/assets/pdf/en-04_zk-it-simplicity__wp_cte_es.pdf

Ecured. (s.f). *Clustering*. Recuperado el 22 de mayo de 2019, de https://www.ecured.cu/Clustering/Esquema_Nacional_de_Seguridad/817-Gestion_incidentes_seguridad/817-Gestion_incidentes_seguridad-ago12.pdf

GestioPolis. (s.f). ¿Qué es *Data Mining*? Recuperado el 28 de abril de 2019, de https://www.gestiopolis.com/que-es-data-mining/

Github. (2019). Conceptos de minería de datos. Recuperado el 22 de mayo de 2019, de https://github.com/MicrosoftDocs/sql-docs.es-es/blob/live/docs/analysis-services/data-mining/data-mining-concepts.md

López, J. (2017). Minería de datos. Recuperado el 21 de abril de 2019, de https://economipedia.com/definiciones/mineria-de-datos.html

Lucidchart. (2019). Qué es un diagrama de árbol de decisión. Recuperado el 28 de abril de 2019, de https://www.lucidchart.com/pages/es/que-es-un-diagrama-de-arbol-de-decision

Marrero, L. (2017). Incidencias de seguridad, en materia de protección de datos. Recuperado el 21 de marzo de 2019, de https://ofiseg.wordpress.com/2012/04/09/que-es-una-incidencia-de-seguridad-en-materia-de-proteccion-de-datos/

Microsoft. (2019). Conceptos de minería de datos. Recuperado el 15 de abril de 2019, de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017

Microsoft. (2019). Algoritmos de minería de datos (*Analysis Services*: Minería de datos). Recuperado el 22 de mayo de 2019, de https://docs.microsoft.com/es-

es/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-server-2017

MINTIC. (2016). Guía para la Gestión y Clasificación de Incidentes de Seguridad de la Información. Recuperado el 29 de marzo de 2019, de https://www.mintic.gov.co/gestionti/615/articles5482_G21_Gestion_Incidentes.pdf

Núñez, F., Zavaleta, I., Felipe, A., Meléndez, J. (2017). Aplicación de técnicas de minería de datos para la tipificación de enfermedades cardiovasculares en alumnos universitarios. Recuperado el 1 de mayo de 2019, de https://www.uaeh.edu.mx/scige/boletin/huejutla/n11/a1.html

Pearlman, S. (2019). ¿En qué consiste un proceso de ETL (Extraer, Transformar y Cargar)? Recuperado el 16 de mayo de 2019, de https://es.talend.com/resources/what-is-etl/

Powerdata. (2013). Procesos ETL: Definición, Características, Beneficios y Retos. Recuperado el 07 de mayo de 2019, de https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312584/procesos-etl-definici-n-caracter-sticas-beneficios-y-retos

SAS. (2019). ETL Qué es y por qué es importante. Recuperado el 07 de mayo de 2019, de https://www.sas.com/es_ar/insights/data-management/what-is-etl.html

Sinnexus. (s.f). *Datawarehouse*. Recuperado el 16 de mayo de 2019, de https://www.sinnexus.com/business_intelligence/datawarehouse.aspx

Tecnoligia&Informatica. (2019). Sistemas OLAP Análisis empresarial. Cubos y tipos de OLAP. Recuperado el 16 de mayo de 2019, de https://tecnologia-informatica.com/sistemas-olap-cubos/

Wordpress. (2011). Algoritmos TDIDT aplicado al Análisis de suelo. Recuperado el 01 de mayo de 2019, de

https://yoshibauco.wordpress.com/2011/03/07/metodologia-para-desarrollo-de-proyectos-de-mineria-de-datos-2da-parte/

Wordpress. (2013). Objetivos de la seguridad informática. Recuperado el 20 de abril de 2019, de https://infosegur.wordpress.com/2013/11/10/objetivos-de-la-seguridad-informatica/