MINERÍA DE DATOS APLICADA EN DETECCIÓN DE INTRUSOS

Diego Vallejo P.
Bancolombia. Medellín, Colombia
dvallejo@bancolombia.com.co

Germán Tenelanda V.
HLB Fast & ABS Auditores. Medellín, Colombia
german.tenelanda@fastauditores.com

(Tipo de Artículo: **Reflexión**. Recibido el 25/11/2011. Aprobado el 25/04/2012)

RESUMEN
Con base a los fundamentos y técnicas de la minería de datos se pueden diseñar y elaborar modelos que permiten encontrar comportamientos clandestinos de fácil detección a simple vista como lo es la información no evidente -desconocida a priori y potencialmente útil- en referencia a hechos determinados. En particular la utilidad de la minería de datos en esta área radica en una serie de técnicas, algoritmos y métodos que imitan la característica humana del aprendizaje: ser capaz de extraer nuevos conocimientos a partir de las experiencias. La minería de datos posee características como: análisis de grandes volúmenes de información, generación de comportamientos que no son fácilmente perceptibles, depuración de datos para toma de decisiones. Estas características pueden ser de vital importancia para ser aplicadas en la seguridad de la información a través de la detección de intrusos. En la actualidad la seguridad de la información es uno de los grandes retos que tiene el mundo, y en especial, la detección de anomalías en los registros de acceso de los diferentes sistemas de información. Con esta aplicabilidad resulta un método básico y muy eficiente de poder prevenir intrusiones. Se centra el campo de en la detección de intrusos al nutrir el proceso de seguimiento de los acontecimientos que ocurren en la red informática, seguido del análisis de los mismos; con el fin de detectar los factores que ponen en peligro la confidencialidad, integridad, disponibilidad y no repudio de los datos. En el presente trabajo se pretende mostrar el aporte a la seguridad de la información de la minería de datos en el contexto de la detección de intrusos.

Palabras clave
Ataques, ciber-terrorismo, datos, denegación de servicios, fuga de datos, intrusiones, minería de datos, modelo, modelado, predicción, vulnerabilidades.

DATA MINING APPLIED FOR INTRUSION DETECTION

ABSTRACT
Based on the fundamentals and techniques of data mining can design and develop models to find illegal behavior easy to detect with the naked eye as it is not obvious information-priori unknown and potentially useful, in reference to particular facts. In particular, the usefulness of data mining in this area lies in a range of techniques, algorithms and methods that mimic the human characteristic of learning: ability to extract new knowledge from experience. Data mining has features such as analysis of large volumes of information, generation of behaviors that are not easily discernible, treatment of data for decision making. These features can be of vital importance to be applied in information security through intrusion detection. At present the information security is one of the great challenges facing the world, and especially the detection of anomalies in the access logs of different information systems. With this basic method applied is very efficient and able to prevent intrusions. It focuses in the field of intrusion detection to nurture the process of monitoring the events occurring in the network, followed by analysis of the same, with a view to identifying the factors that threaten the confidentiality, integrity, availability and non-repudiation of data. In the present work aims to show the contribution to the information security of data mining in the context of intrusion detection.

Keywords
Attacks, cyberterrorism, data, data mining, denial of service, intrusions, vulnerabilities, leakage, model, modeling, prediction.

FOUILLE DE DONNEES APPLIQUE DANS DETECTION DES INTRUS

RÉSUMÉ
D'après les fondations et techniques de la fouille de données on peut se concevoir et réaliser des modèles qui permettent de trouver des conduites clandestines faciles de trouver à première vue comme l'information que n'est pas évident –inconnu a priori et qui est potentiellement utile- par rapport à des faits spécifiques. En particulier l'utilité de la fouille de données dans ce domaine réside dans un ensemble de techniques, algorithmes et méthodes qui imitent la caractéristique humaine de l'apprentissage : être capable d'extraire des nouvelles connaissances à partir des expériences. La fouille de données a des caractéristiques comme : l'analyse des grandes volumes d'information, la création des conduites qui ne sont pas percevables facilement, épuration des donnes pour la prise de décisions. Ces caractéristiques peuvent être très importants pour leur appliquer dans la sécurité de l'information à travers de la détection des intrus. Actuellement la sécurité de l'information est un des grands défis dans le monde entier, et particulièrement, la détection d'anomalies dans les registres d'accès des différents systèmes d'information. Avec cette applicabilité on a une méthode de base et très efficace pour prévenir des intrusions. On est centré sur le domaine de la détection des intrus à partir d'alimenter le processus de suivi des événements qui passent dans le réseau informatique, aussi avec l'analyse d'eux ; avec l'intention de détecter les facteurs qui mettent en danger la confidentialité, l'intégrité, la disponibilité et la non répudiation des données. Dans ce travail on prétendre de montrer l'apport à la sécurité de l'information de la fouille de données sur le contexte de la détection des intrus.

Mots-clés
Attaques, cyber-terrorisme, données, dénégation de services, fuite de données, intrusions, fouille de données, modèle, modelage, prédiction, vulnérabilités.

1. INTRODUCCIÓN

Actualmente, los datos son el activo más importante para cualquier persona o empresa y por lo tanto se deben proteger porque son vulnerables en el ambiente de las redes locales y en internet. Es posible que existan individuos o software malicioso (*malware*) que intenten llegar a esos datos valiéndose de distintos medios y generar consecuencias indeseables como imprevisibles. Los vándalos informáticos aprovechan las vulnerabilidades y atacan a los sistemas computarizados, el control de flujo de navegación en la red y el intercambio de correo, entre otros.

Muchas de estas batallas de bits pasan desapercibidas, no se percibe el ataque, no se es consciente de tener software malicioso latente en el computador o que sea un zombi manipulado por alguien en cualquier lugar del mundo. Todo esto puede pasar desapercibido, por lo tanto, es necesario detectar las anomalías que se presenten en los registros de acceso, lo cual es posible mediante técnicas de detección de intrusos, analizando aquellos accesos que pongan en peligro la confidencialidad, integridad, disponibilidad y no repudio de los datos.

Al registrar los datos del acceso a los sistemas y almacenarlos para posterior análisis, se crea gran volumen de datos que, a simple vista, no son fáciles de analizar y de correlacionar entre sí. En este trabajo se utiliza la minería de datos para buscar información no trivial que se encuentre oculta o dispersa en ellos, es decir, se exploran los datos para descubrir la interconexión e interrelación y poder obtener la información oculta [1-3].

La herramienta que se utiliza para la minería de datos es IBM SPSS Modeler [4]; una aplicación que permite elaborar modelos predictivos de forma rápida e intuitiva sin necesidad de programación y que descubre patrones y tendencias en datos estructurados o no estructurados. Los datos con los que se trabaja se toman de los utilizados en The Third International Knowledge Discovery and Data Mining Tools Competition [5] y fueron preparados y dirigidos por los laboratorios de MIT Lincoln. Esta base de datos contiene un conjunto estándar que incluye una amplia variedad de intrusiones simuladas en un entorno de red militar de la fuerza aérea de Norteamérica y son la recopilación de siete semanas de tráfico TCP de red.

De este análisis se espera descubrir la utilidad de la minería de datos en la exploración, a fin de encontrar y predecir ataques que no son detectados por antivirus, cortafuegos o sistemas de detección de intrusos.

2. DESCRIPCIÓN DE LA PROBLEMÁTICA

2.1 Causas del Problema

Las redes de cómputo locales unidas a la Internet han facilitado la comunicación de las personas y empresas, pero a la vez ponen en riesgo el activo más importante: los datos. Esta facilidad de intercambiar información multiplica la capacidad de los ataques y promueve a usuarios maliciosos y crackers a buscar objetivos vulnerables, como las aplicaciones no actualizadas —sistemas operativos, bases de datos—, sistemas infectados con virus a través de correos electrónicos, navegación por páginas web, redes de datos empresariales, descargas de datos, ejecución de servicios inseguros o puertos abiertos. Por lo tanto, es necesario crear “alarmas” que ayuden a notificar a los administradores y jefes de seguridad de la información a fin de responder oportunamente a estas amenazas. Esas alarmas son llamadas sistemas de detección de intrusos.

2.2 Síntomas y signos

Los ataques son todas las acciones que violan el sistema de seguridad computacional, afectando la confidencialidad, integridad, disponibilidad o no repudio y pueden presentar los siguientes signos verificables: (a) interrupción, el recurso se vuelve no disponible, (b) intercepción, “alguien” no autorizado consigue acceso a un recurso y (c) modificación, además de la intercepción es capaz de manipular los datos. Todos estos signos se podrían manifestar con lentitud y desaparecer archivos y datos o hacer que los periféricos funcionen incorrectamente.

Sin la utilización de herramientas especiales se pueden presentar otros signos no identificables u ocultos, como escanear puertos, buscar puertos abiertos y tomar los de utilidad, ataques de autenticación, el atacante suplanta a una persona que tiene autorización; explotación de errores, los desarrollos computacionales presentan fallas o agujeros de seguridad, ataques de denegación de servicios, consisten en saturar un servidor con múltiples solicitudes hasta dejarlo fuera de servicio.

2.3 Consecuencias

Las consecuencias de los ataques informáticos se podrían clasificar en:

- Datos dañados: la información que no contenía daños pasa a tenerlos.
- Denegación de servicios —Denial of Service-DoS— servicios que deberían estar disponibles no lo están.
- Fuga de datos —Leakage—: los datos llegan a destinos a los que no deberían llegar.
- Sabotaje informático: daños realizados por empleados descontentos.
- Pornográfica: una fuente económica que mueve mucho dinero.
- Ciber-terrorismo: organizaciones criminales la utilizan con fines terroristas.

En todo el mundo los periódicos y los noticieros dan cuenta de las consecuencias: Republica Dominicana,

el grupo Anonymus amenaza con atacar páginas del Gobierno; Ecuador, Anonymus reveló datos de 45 mil policías; México, el narcotráfico secuestra a hacker, con el objeto de clonar tarjetas de crédito y robar datos personales de internet; Brasil, ciber-ataques paralizan sitios del Gobierno brasileño que fueron atacados por tercer día consecutivo por piratas cibernéticos; Reino Unido, miembros de un grupo denominado TeaMp0isoN filtró información privada del ex primer ministro británico Tony Blair así como varios números de teléfono y direcciones personales de responsables políticos del país; España, el denominado grupo Anonymous bloqueo las páginas oficiales de cuerpo nacional de policía, o el servicio público de empleo Estatal; Colombia, un ataque informático que colapsó la página web de la Registraduría donde se iban a publicar los resultados de las legislativas de marzo de 2010, al parecer tuvo como origen direcciones IP del Ministerio de Defensa, la Policía y el Departamento Administrativo de Seguridad.

2.4 Perspectivas de solución

Partiendo de una base de datos con captura de conexiones realizada para el The Third International Knowledge Discovery And Data Mining Tools Competition [5] y basados en la metodología CRISP-DM [6] aplicada en minería de datos, se pretende mostrar mediante diferentes técnicas de modelado, basadas en cálculos estadísticos, cómo la aplicación de esta técnica sirve para analizar y procesar grandes volúmenes de datos de una manera predictiva, con miras a entregar resultados de manera oportuna, eficiente, eficaz y confiable y entregando información que ayude en la toma de decisiones en la prevención de posibles intrusiones.

3. METODOLOGÍA Y HERRAMIENTAS

Metodología CRoss-Industry Standard Process for Data Mining CRISP-DM [6]. Metodología lanzada en el año 1996, por la comisión Europea (Figura 1), es abierta y sin un propietario específico. Puede ser desarrollada sobre cualquier herramienta de minería de datos. Es un proceso viable y repetible que permite plasmar las experiencias de análisis para luego proceder a replicarlas. Ayuda en la planeación y ejecución de modelos de minería de datos.

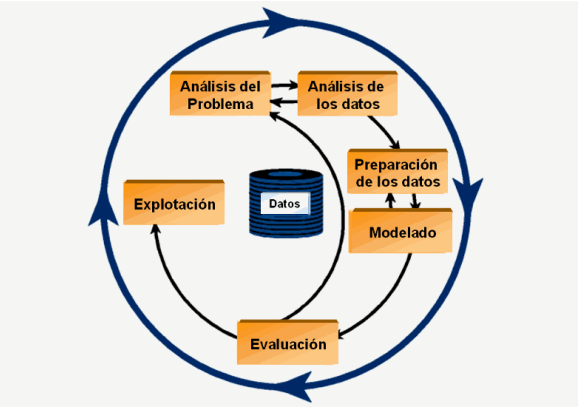

Fig. 1: Metodología CRISP-DM

3.1 Procedimiento

1. **Análisis o entendimiento del negocio.** Incluye la comprensión de los objetivos y requerimientos del proyecto desde una perspectiva empresarial, con el fin de convertirlos en objetivos técnicos y en una planificación.

2. **Análisis o entendimiento de los datos.** Comprende la recolección inicial de datos, en orden a que sea posible establecer un primer contacto con el problema, identificando la calidad de los datos y estableciendo las relaciones más evidentes que permitan establecer las primeras hipótesis.

3. **Preparación de los datos.** incluye las tareas generales de selección de datos a los que se va a aplicar la técnica de modelado (variables y muestras), limpieza de los datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

4. **Modelado.** se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos específico.

5. **Evaluación de resultados.** Se evalúa el modelo, no desde el punto de vista de los datos, sino del cumplimiento de los criterios de éxito del problema.

6. **Explotación o despliegue de resultados.** Normalmente los proyectos de minería de datos no terminan en la implantación del modelo sino que se debe documentar y presentar los resultados de manera comprensible en orden a lograr un incremento del conocimiento. Además, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados.

3.2 Comparación de herramientas

Con el paso del tiempo han surgido herramientas para la explotación de datos, con el fin de abarcar un mercado que ha tenido gran auge en los últimos años y brindando soluciones sobre las fuertes demandas de las empresas. Para la aplicación de minería de datos de acceso gratuito o licencia libre se puede mencionar: MiningMart, Orange, TariyKDD, ARMiner y WEKA; por el lado comercial: Cart, SAS Enterprise Miner, Tiberius, Kxen, IBM SPSS Modeler, que es la herramienta de minería que se aplica para realizar este artículo. En la Figura 2 se muestra un comparativo entre algunas de estas herramientas.

3.3 Herramientas seleccionadas

El software de minería de datos utilizado es IBM SPSS Modeler [4], que es un área de trabajo de minería de datos que permite elaborar modelos predictivos de forma rápida e intuitiva, sin necesidad de programación. Descubre patrones y tendencias en datos estructurados o no estructurados mediante una

interfaz visual soportada por análisis avanzado. Modela los resultados y reconoce factores que influyen en ellos. Así, se puede aprovechar las oportunidades y atenuar los riesgos. Para éste estudio se utilizó la versión 14.2.

Característica	Modeler	SAS Enterprise Miner	Tariykdd	Weka
Licencia libre	No	No	Si	Si
Requiere conocimientos avanzados	No	No	No	No
Acceso a SQL	Si	No	Si	Si
Multiplataforma	No	Si	Si	Si
Requiere bases de datos especializadas	No	---	No	No
Métodos de máquinas de soporte vectorial	Si	Si	No	Si
Métodos bayesianos	Si	---	No	Si
Puede combinar modelos	Si	Si	No	Si (no resulta muy eficiente)
Modelos de clasificación	Si	Si	Si	Si
Implementa arboles de decisión	Si	Si	Si	Si
Modelos de regresión	Si	Si	No	Si
Clusterin y agrupamiento	Si	Si	No	Si
Interfaz amigable	Si	Si	Si	Si
Permite visualización de datos	Si	Si	Si	Si

Fig. 2: Comparación de herramientas de Minería de Datos

4. TÉCNICA DE MODELADO PARA MINERÍA DE DATOS

Las técnicas de modelado se basan en el uso de algoritmos. Existen tres clases principales de técnicas de modelado y la herramienta utilizada, IBM SPSS Modeler, ofrece varios ejemplos de cada uno: clasificación, asociación, segmentación.

4.1 Los modelos de clasificación

Utilizan el valor de uno o más campos de entrada para predecir el valor de uno o más resultados o campos de destino. Algunos ejemplos de estas técnicas son:

- Modelo de árbol de clasificación y regresión C&R: genera un árbol de decisión que permite pronosticar o clasificar observaciones futuras. El método utiliza la partición reiterada para dividir los registros de prueba en segmentos minimizando las impurezas en cada paso, un nodo se considera “puro” si el 100% de los casos del nodo corresponden a una categoría específica del campo objetivo. Los campos de entrada y objetivo pueden ser continuos (numéricos) o categóricos (nominal, ordinal o marca). Todas las divisiones son binarias (sólo crea dos subgrupos).

- Modelo QUEST: proporciona un método de clasificación binario para generar árboles de decisión; está diseñado para reducir el tiempo de procesamiento necesario para realizar los análisis de C&RT y reducir la tendencia de los métodos de clasificación de árboles para favorecer a las entradas que permitan realizar más divisiones. Los campos de entrada pueden ser continuos (numéricos), sin embargo el campo objetivo debe ser categórico. Todas las divisiones son binarias.

- El modelo CHAID: genera árboles de decisión utilizando estadísticos de chi-cuadrado para identificar las divisiones óptimas. A diferencia de los nodos C&RT y QUEST, CHAID puede generar árboles no binarios, lo que significa que algunas divisiones generarán más de dos ramas. Los campos de entrada y objetivo pueden ser continuos (numéricos) o categóricos. CHAID exhaustivo es una modificación de CHAID que examina con mayor precisión todas las divisiones posibles.

- El modelo C5.0: genera un árbol de decisión o un conjunto de reglas. El modelo divide la muestra basándose en el campo que ofrece la máxima ganancia de información en cada nivel. El campo objetivo debe ser categórico. Se permiten varias divisiones en más de dos subgrupos.

- Los modelos lineales predicen un destino continuo tomando como base las relaciones lineales entre el destino y uno o más predictores.

- La regresión lineal: es una técnica estadística común utilizada para resumir datos y realizar pronósticos ajustando una superficie o línea recta que minimice las discrepancias existentes entre los valores de salida reales y los pronosticados.

- La regresión logística: es una técnica estadística para clasificar los registros en función los valores de los campos de entrada. Es análoga a la regresión lineal pero toma un campo objetivo categórico en lugar de uno numérico.

- El modelo Regresión de Cox: permite crear un modelo de supervivencia para datos de tiempo hasta el evento en presencia de registros censurados. El modelo produce una función de supervivencia que pronostica la probabilidad de que el evento de interés se haya producido en el momento dado (t) para valores determinados de las variables de entrada.

- El modelo Red bayesiana: permite crear un modelo de probabilidad combinando pruebas observadas y registradas con conocimiento del mundo real para establecer la probabilidad de instancias. El nodo se centra en las redes Naïve Bayes aumentado a árbol (TAN) y de cadena de Markov que se utilizan principalmente para la clasificación.

4.2 Modelos de Asociación

Estos modelos encuentran patrones en los datos en los que una o más entidades, como eventos, compras o atributos, se asocian con una o más entidades. Los modelos construyen conjuntos de reglas que definen estas relaciones. Aquí los campos de los datos pueden funcionar como entradas y destinos. Podría encontrar estas asociaciones manualmente, pero los algoritmos de reglas de asociaciones lo hacen mucho más rápido, y pueden explorar patrones más complejos. Los modelos inducción de reglas generalizado son:

- Inducción de reglas generalizado GRI: es capaz de encontrar las reglas de asociación existentes en los datos. Por ejemplo, los clientes que compran cuchillas y loción para después del afeitado también suelen comprar crema de afeitado. GRI extrae reglas con el mayor nivel de contenido de información basándose en un índice que tiene en cuenta tanto la generalidad (soporte) como la precisión (confianza) de las reglas. GRI puede gestionar entradas numéricas y categóricas, pero el objetivo debe ser categórico.

- El modelo A priori: extrae un conjunto de reglas de los datos y destaca aquellas reglas con un mayor contenido de información. A priori ofrece cinco métodos diferentes para la selección de reglas y utiliza un sofisticado esquema de indización para procesar eficientemente grandes conjuntos de datos. En los problemas de mucho volumen, A priori se entrena más rápidamente que GRI, no tiene un límite arbitrario para el número de reglas que puede retener y puede gestionar reglas que tengan hasta 32 precondiciones. A priori requiere que todos los campos de entrada y salida sean categóricos, pero ofrece un mejor rendimiento ya que está optimizado para este tipo de datos.

- El modelo CARMA: extrae un conjunto de reglas de los datos sin necesidad de especificar campos de entrada ni de objetivo. A diferencia de A priori y GRI, el nodo CARMA ofrece configuraciones de generación basadas en el soporte de las reglas (soporte para el antecedente y el consecuente), no sólo en el soporte de antecedentes. Esto significa que las reglas generadas se pueden utilizar en una gama de aplicaciones más amplia, por ejemplo, para buscar una lista de productos o servicios (antecedentes) cuyo consecuente es el elemento que se desea promocionar durante esta temporada de vacaciones.

- El modelo Secuencia: encuentra reglas de asociación en datos secuenciales o en datos ordenados en el tiempo. Una secuencia es una lista de conjuntos de elementos que tiende a producirse en un orden previsible. Por ejemplo, es probable que un cliente que compra una cuchilla y una loción para después del afeitado compre crema para afeitar la próxima vez que vaya a comprar. El nodo Secuencia se basa en el algoritmo de reglas de asociación de CARMA, que utiliza un método de dos pasos para encontrar las secuencias.

4.3 Modelos de Segmentación

Dividen los datos en segmentos o conglomerados de registros que tienen patrones similares de campos de entrada. Como sólo se interesan por los campos de entrada, los modelos de segmentación no contemplan el concepto de campos de salida o destino. Ejemplos de modelos de segmentación son:

- El modelo K-medias: agrupa conjuntos de datos en grupos distintos (o conglomerados). El método define un número fijo de conglomerados a los que asigna los registros de forma iterativa y ajusta los centros hasta que no se pueda mejorar el modelo. En lugar de intentar pronosticar un resultado, los modelos k-medias utilizan el proceso de aprendizaje no supervisado para revelar los patrones del conjunto de campos de entrada.

- El modelo Kohonen: genera un tipo de red neuronal que se puede usar para conglomerar un conjunto de datos en grupos distintos. Cuando la red se termina de entrenar, los registros que son similares se deberían presentar juntos en el mapa de resultados, mientras que los registros que son diferentes aparecerían aparte. Puede observar el número de observaciones capturadas por cada unidad en el nugget de modelo para identificar unidades fuertes. Esto le proporcionará una idea del número apropiado de conglomerados.

- El modelo Bietápico: es un método de conglomerado de dos pasos. El primer paso es hacer una única pasada por los datos para comprimir los datos de entrada de la fila en un conjunto de sub-conglomerados administrable. El segundo paso utiliza un método de conglomerado jerárquico para fundir progresivamente los sub-conglomerados en conglomerados cada vez más grandes. El bietápico tiene la ventaja de estimar automáticamente el número óptimo de conglomerados para los datos de entrenamiento. Puede tratar tipos de campos mixtos y grandes conjuntos de datos de manera eficaz.

- El modelo Detección de anomalías: identifica casos extraños, o valores atípicos, que no se ajustan a patrones de datos “normales”. Con este nodo, es posible identificar valores atípicos aunque no se ajusten a ningún patrón previamente conocido o no se realice una búsqueda exacta.

5. CASO DE ESTUDIO

5.1 Descripción

Debido al alto nivel de conectividad de los equipos de cómputo, hoy tenemos acceso a gran cantidad de datos e información de diferentes fuentes y desde de cualquier lugar del mundo. El conocimiento está al alcance de personas intentado realizar buenas acciones o intrusos intentando llegar a los datos para causar daño.

Antes se necesitaban conocimientos profundos de computadores, de redes de computadores, conexiones y otros para realizar un ataque de intrusión; hoy en día a través de consultas en un navegador o a un clic, se puede obtener información y herramientas que permiten realizar intrusiones. Entonces, la detección de intrusos es necesaria porque:

- Existen numerosas amenazas.
- En la red circula tráfico normal y malicioso.
- Es necesario un esquema de monitoreo para eventos maliciosos en la infraestructura de las compañías.
- Una sola intrusión puede ocasionar un impacto catastrófico.

¿Cómo se detectan las intrusiones? Para este estudio se toma una base de datos de evaluación de detección de intrusos que fue utilizada en la Competición Internacional de Datos KDD de 1999, que incluye una amplia variedad de intrusiones simuladas en un entorno militar típico de la Fuerza Aérea de Estados Unidos y que contiene ataques de intrusión clasificados en cuatro tipos de ataque:

- **Denegación de servicios (DoS- Denial of service).** El objetivo es tratar de detener el funcionamiento de la red, la máquina o el proceso, de tal forma que un servicio o recurso sea inaccesible a los usuarios legítimos [7].

- **R2L (Remote to local).** Acceso no autorizado desde una máquina remota.

- **U2R (User to root).** Acceso no autorizado mediante escalamiento de privilegios de una cuenta de usuario autorizado hasta llegar a superusuario.

- **Indagación y exploración (Probing).** Este ataque escanea las redes, en busca de vulnerabilidades mediante la recolección de información, tal como, direcciones IP validas, servicios, sistemas operativos y otros.

Esta simulación se centró en la selección y clasificación de paquetes TCP/IP. Una conexión TCP/IP es una secuencia de paquetes de datos con un protocolo definido hacia y desde una dirección IP. Para este caso los paquetes fueron clasificados como normales y tipo de ataque detectado. La base de datos contiene 41 variables independientes en cada registro, con los cuales se describen diferentes características de cada conexión y 24 tipos de ataque, con 14 tipos adicionales en los datos de test [5], como se observa en las Tablas 1-4.

TABLA 1. Tipos de ataques

Ataque	Descripción	Tipo
Back	Ataque contra el servidor web Apache cuando un cliente pide una URL que contiene muchas barras.	DoS
Land	Envío de TCP/SYN falso con la dirección de la víctima como origen y destino, causando que se responda a sí mismo continuamente.	DoS
Neptune	Inundación por envíos de TCP/SYN en uno o más puertos.	DoS
Pod	Ping de la muerte: manda muchos paquetes ICMP muy pesados.	DoS
Teardrop	Usa el algoritmo de fragmentación de paquetes IP para enviar paquetes corruptos a la víctima.	DoS
Smurf	El atacante envía un ping, que parece proceder de la víctima, en *broadcast* a una tercera parte de la red, donde todos los host responderán a la víctima.	DoS
Ftp_write	Usuario FTP remoto crea un archivo .rhost y obtiene un login local.	R2L
Guess_passwd	Trata de adivinar la contraseña con telnet para la cuenta de visitante	R2L
Imap	Desbordamiento remoto del búfer utilizando el puerto imap.	R2L
Multihop	Escenario de varios días donde el atacante primero accede a una máquina que luego usa como trampolín para atacar a otras máquinas.	R2L
Phf	Script CGI que permite ejecutar comandos en una máquina con un servidor web mal configurado.	R2L
Spy	Analizador de protocolos LAN por la interfaz de red.	R2L
Warezclient	Los usuarios descargan software ilegal publicado a través de FTP anónimo por el warezmaster.	R2L
Warezmaster	Subida FTP anónima de Warez (copias ilegales de software).	R2L
Buffer_overflow	Desbordamiento de la pila del búfer.	UR2
Loadmodule	Ataque furtivo que reinicia la IFS para un usuario normal y crea un shell root.	UR2
Perl	Establece el id de usuario como root en un script de perl y crea un shell de root.	UR2
Rootkit	Escenario de varios días donde un usuario instala componentes de un rootkit.	UR2
Ipsweep	Sondeo con barrido de puertos o mandando pings a múltiples direcciones host.	Probing
Nmap	Escaneo de redes mediante la herramienta nmap.	Probing
Portsweep	Barrido de puertos para determinar qué servicios se apoyan en un único host.	Probing
Satan	Herramienta de sondeo de redes que busca debilidades conocidas.	Probing

TABLA 2. Atributos básicos de las conexiones TCP

Atributo	Descripción	Tipo
Duration	Longitud (número de segundos) de la conexión.	Continuo
Protocol_type	Tipo de protocolo (tcp...)	Discreto
Service	Tipo de servicio de destino (HTTP, Telnet, SMTP…)	Discreto
Src_bytes	Número de bytes de datos de fuente a destino	Continuo
Dst_bytes	Número de bytes de datos de destino a la fuente.	Continuo
Flag	Estado de la conexión (SF, S1, REJ…)	Discreto
Land	1 si la conexión corresponde mismo host/puerto; 0 de otro modo.	Discreto
Wrong_fragment	Número de fragmentos erróneos.	Continuo
Urgent	Número de paquetes urgentes.	Continuo

TABLA 3. Atributos especiales

Atributo	Descripción	Tipo
Hot	Número de indicadores “hot”.	Continuo
Num_failed_logins	Número de intentos de acceso fallidos.	Continuo
Logged_in	1 si acceso exitoso 0 de otro modo	Discreto
Num_compromised	Número de condiciones “sospechosas”.	Continuo
Root_shell	1 si se obtiene superusuario para acceso a root; 0 de otro modo.	Discreto

Su_attempted	1 si se intenta el comando “su root”; 0 de otro modo.	Discreto
Num_root	Número de accesos a root.	Continuo
Num_file_creations	Número de operaciones de creación de ficheros.	Continuo
Num_shells	Número de Shell prompts.	Continuo
Num_access_files	Número de operaciones de control de acceso a ficheros.	Continuo
Num_outbound_cmds	Número de comandos de salida en una sesion ftp.	Continuo
Is_hot_login	1 si el login pertenece a la lista “hot”; 0 de otro modo.	Discreto
Is_guest_login	1 si el acceso es un “guest” ‘login; 0 de otro modo.	Discreto

TABLA 4. Atributos con ventana de 2 segundos

Atributo	Descripción	Tipo
Count	Número de conexiones a la misma máquina que la conexión actual en los dos últimos segundos	Continuo
Atributos que se refieren a las conexiones de mismo host		
Serror_rate	Porcentaje de conexiones que tienen errores “SYN”.	Continuo
Rerror_rate	Porcentaje de conexiones que tienen errores “REJ”.	Continuo
Same_srv_rate	Porcentaje de conexiones con el mismo servicio.	Continuo
Diff_srv_rate	Porcentaje de conexiones con diferentes servicios.	Continuo
Srv_count	Número de conexiones al mismo servicio que la conexión actual en los dos últimos segundos	Continuo
Atributos que se refieren a las conexiones de mismo servicio		
Srv_serror_rate	Porcentaje de conexiones que tienen errores “SYN”.	Continuo
Srv_rerror_rate	Porcentaje de conexiones que tienen errores “REJ”.	Continuo
Srv_diff_host_rate	Porcentaje de conexiones a diferentes hosts.	Continuo

6. RESULTADOS

Para el caso de estudio se seleccionó la metodología CRISP-DM [6] y se desarrollaron las siguientes etapas:

1. **Análisis o entendimiento del negocio.** Partiendo de los datos de la Base de Datos KDD se pretende predecir la repetición del ataque realizado en un periodo de tiempo dado, a través de definición de reglas creados para nuestro fin.

2. **Análisis o entendimiento de los datos.** Cuando se introducen los datos en Modeler, la herramienta permite representar diferentes tipos de tablas, información de tipo estadística y gráficas, que relacionan atributos de los datos allí contenidos.

   En las Figuras 3 y 4 se observa información que se puede obtener del nodo Auditar datos: Campo, Gráfico de muestra, Medida, Mínimo, Máximo, Media, Desviación Estándar, Asimetría, entre otros, los cuales sirven para hacer una visualización inicial de los datos. Se puede observar en este caso que el número de registros que se tiene en la base de datos es de 311.029 registros.

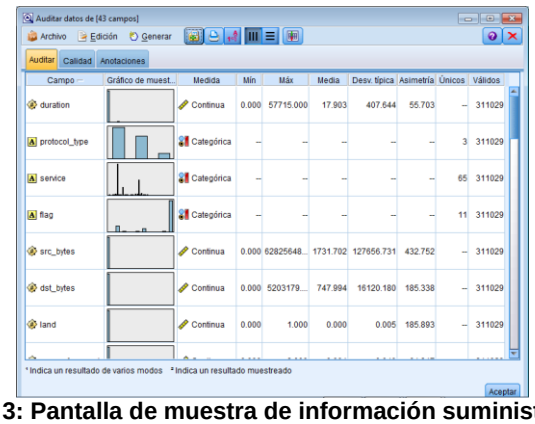

Fig. 3: Pantalla de muestra de información suministrada por el Nodo Auditar Datos

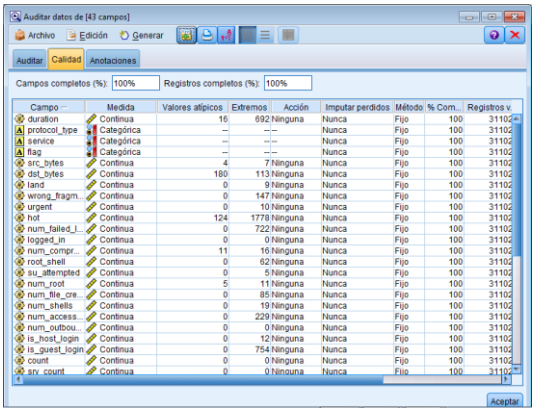

Fig. 4: Pantalla de muestra de información suministrada por el Nodo Auditar Datos

La Figura 5 muestra los diferentes campos estadísticos que se pueden obtener para hacer los diferentes cálculos en los modelos a evaluar.

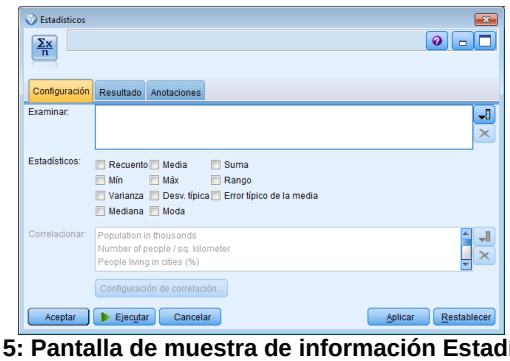

Fig. 5: Pantalla de muestra de información Estadística

La Figura 6 muestra información tomada desde “Modeler”, en la que se puede observar la tabla de entrada de datos, registro por registro de la Base de Datos de prueba.

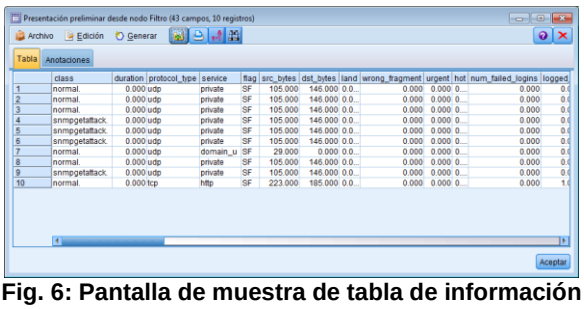

Fig. 6: Pantalla de muestra de tabla de información

En la Figura 7, se observa cual es la frecuencia con la que aparece cada tipo de ataque en los datos de prueba. Los ataques de denegación de servicio (DoS), aparecen casi todos en muchas ocasiones ('back' 1.098 veces, 'pod' 87 veces, 'teardrop' 12 veces o 'land' con sólo 9 apariciones), destacando los que se basan en inundación que sobresalen ampliamente del resto, con las 58.001 apariciones de 'neptune' y 'smurf' con 164.091 conexiones, siendo el ataque más frecuente de todos. Los ataques de tipo R2L y U2R están contenidos dentro de un paquete de datos y suelen consistir en una única conexión; se observa que casi todos o no aparecen (como 'ftp_write', 'multihop', 'perl', 'buffer_overflow', 'rootkit' y 'spy') o constan de una única aparición (como 'guess_passwd', 'phf', 'imap', 'loadmodule', y 'warezmaster').

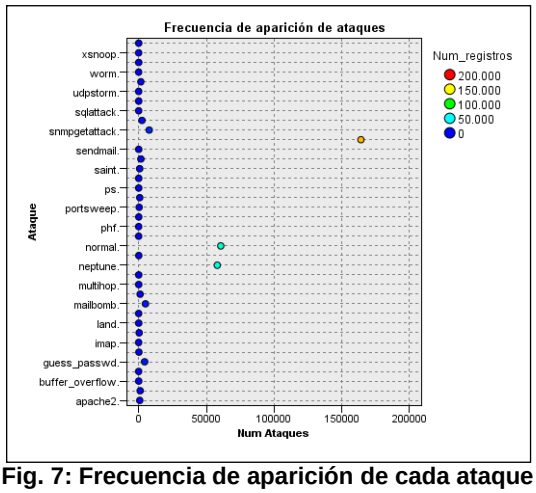

Fig. 7: Frecuencia de aparición de cada ataque

Aquellos ataques que no aparecen en la muestra inicial de datos los modelos de minería de datos no pueden predecirlos, dado que no se tienen características mínimas para que el modelo los procese.

En la Tabla 5, se puede apreciar que los ataques de sondeo (probing) tienen varias ocurrencias cada uno: 'satan' (1.633), 'portsweep' (354), 'ipsweep' (306), 'nmap' (84). También, se puede apreciar que las conexiones normales son muchas (60.593), pero no llegan a ser las más frecuentes, ganándole en número los ataques de inundación (Smurf 164.091).

TABLA 5. Número de Ataques por Tipo

Ataque	Cantidad
Smurf	164.091
Normal	60.593
Neptune	58.001
Snmpgetattack	7.741
Mailbomb	5.000
Guess_passwd	4.367
Snmpguess	2.406
Satan	1.633
Warezmaster	1.602
Back	1.098
Mscan	1.053
Otros ataques: apache2, processtable, saint, portsweep, ipsweep, httptunnel	3.444

En la Figura 8 y en la Tabla 7 se puede ver que el tipo de ataque depende mucho del tipo de protocolo que se use, como todos los ataques '*smurf*' usan '*icmp*', los '*teardrop*' '*udp*' y los de tipo '*back*', '*tcp*', por lo que este atributo es muy importante para aumentar la calidad de la predicción del modelo.

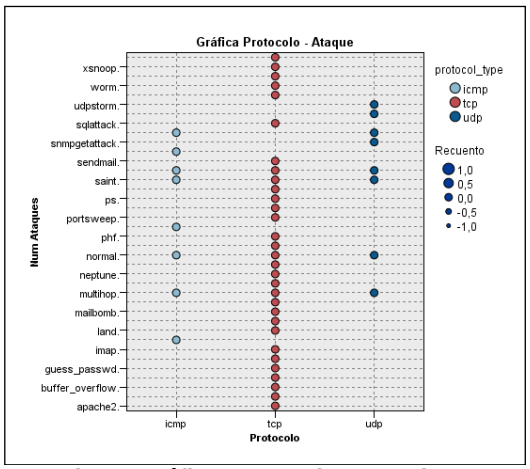

Fig. 8: Gráfica protocol_type - class

TABLA 7. Tipos de ataque

Protocolo	Ataque	Registros
Icmp	Smurf.	164.091
Icmp	Normal.	378
Icmp	Ipsweep.	306
Icmp	Saint.	102
Icmp	Pod.	87
Icmp	Snmpguess.	3
Icmp	Satan.	1
Icmp	Multihop.	1
Udp	Normal.	16.097
Udp	Snmpgetattack.	7.741
Udp	Snmpguess.	2.403
Udp	Satan.	413
Udp	Saint.	27
Udp	Teardrop.	12
Udp	Multihop.	8
Udp	Udpstorm.	2
Tcp	Neptune.	58.001
Tcp	Normal.	44.118
Tcp	Mailbomb.	5.000
Tcp	Guess_passwd.	4.367
Tcp	Warezmaster.	1.602
Tcp	Satan.	1.219
Tcp	Back.	1.098
Tcp	Mscan.	1.053
Tcp	Apache2.	794
Tcp	Processtable.	759
Tcp	Saint.	607
Tcp	Portsweep.	354
Tcp	Httptunnel.	158
Tcp	Nmap.	84
Tcp	Buffer_overflow.	22
Tcp	Sendmail.	17
Tcp	Named.	17
Tcp	Ps.	16
Tcp	Xterm.	13
Tcp	Rootkit.	13
Tcp	Multihop, xlock, land, xsnoop, ftp_write, perl, loadmodule, phf, worm, sqlattack, imap.	45

En la Figura 9 se observa que la mayoría de los ataques 'teardrop' tienen un número de fragmentos erróneos (campo Wrong_fragment) superior al resto de ataques.

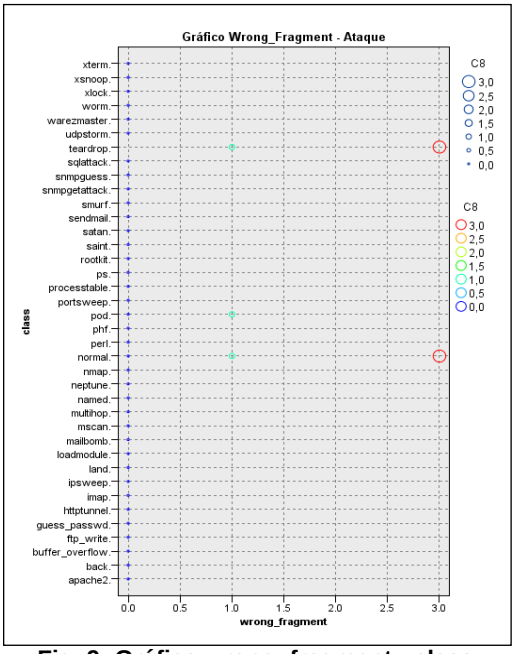

Fig. 9: Gráfica wrong_fragment - class

Con los atributos especiales también se ve cómo se distribuyen los datos. Todos los ataques 'warezclient' y 'back' tienen un 'logged_in' exitoso (Figura 10).

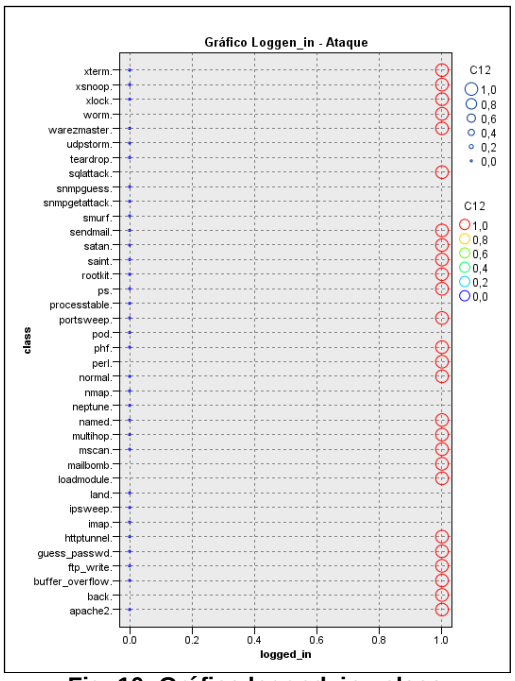

Fig. 10: Gráfica logged_in - class

En la Figura 11, respecto a los atributos basados en tiempo de "mismo host"', se puede observar que los ataques de tipo 'smurf' y 'satan' presentan un número elevado de conexiones a la misma máquina que una conexión actual en los dos últimos segundos. Esto sirve para verificar qué algunos ataques de Probing, como 'satan', escanean los puertos con un intervalo de tiempo mucho mayor de dos segundos. Así también se puede afirmar que los atributos de "mismo host" van a ser de gran utilidad para detectar este tipo de ataques.

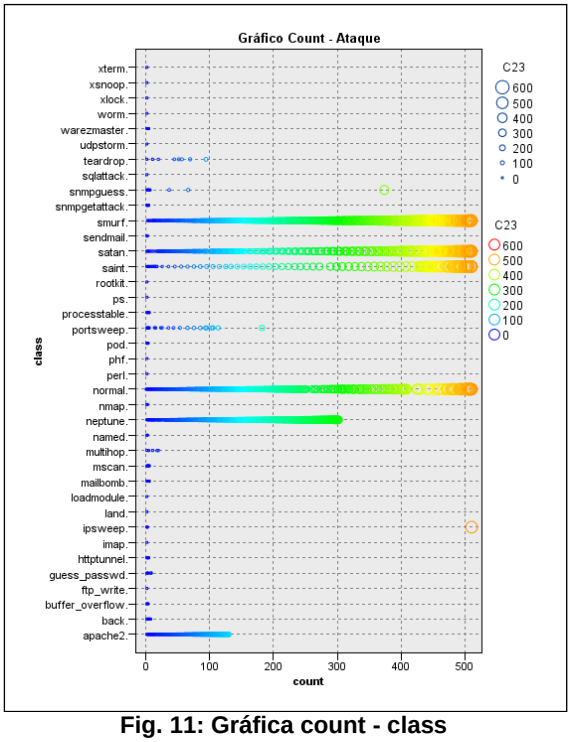

Fig. 11: Gráfica count - class

De esta misma manera, se pueden hacer con muchos de los atributos restantes visualizaciones, donde algunos aportan mucha información simplemente mirando las gráficas y con otros no será tan fácil de ver simple vista. Así se intuye cómo puede influir un atributo en la clasificación que hará el modelo de predicción resultante, y decidir si merece la pena su utilización respecto al costo computacional que suponga incluirlo.

3. **Preparación de los datos.** En términos generales durante esta etapa o fase, como su nombre lo indican, se preparan los datos que van a ser trabajados, para ello en algunos casos es necesario crear variables nuevas, cambiar los valores de continuos a nominales, o a tipo marca, etc. Este tipo de operaciones con los campos y/o con los registros son los que nos permiten obtener la muestra óptima para continuar con la siguiente etapa.

4. **Modelado.** En esta etapa se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos (Figura 12).

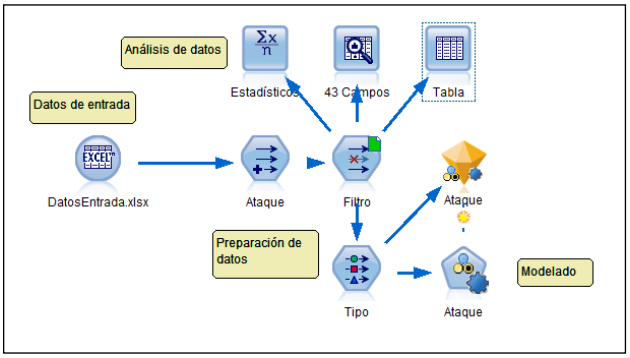

Fig. 12: Ambiente de la Aplicación IBM SPSS Modeler para desarrollo de Modelos de Minería de Datos

Algunos de los modelos evaluados en este caso son (Figura 13): El modelo CHAID (Figuras 14-16), el C 5.0 (Figuras 17-19) y el modelo árbol C&RT (Figuras 20-22).

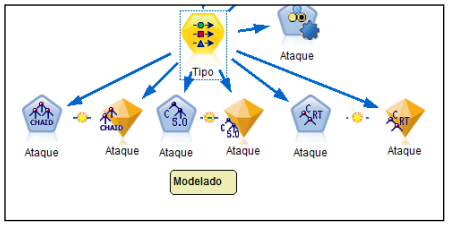

Fig. 13: Modelos evaluados

Modelo CHAID. A continuación mostramos gráficamente algunos de los análisis que la herramienta permite generar.

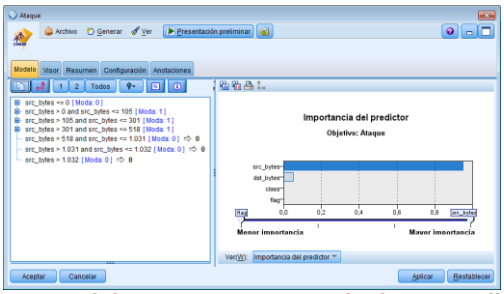

Fig. 14: Modelo CHAID. Importancia de un predictor

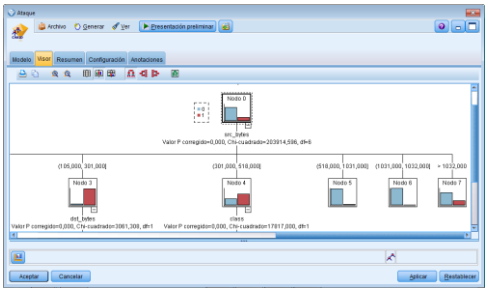

Fig. 15. Modelo CHAID. Visor de reglas generadas

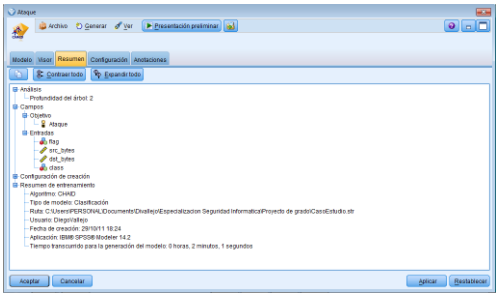

Fig. 16: Modelo CHAID. Resumen

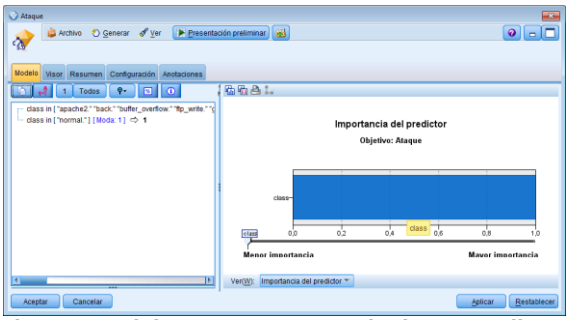

Fig. 17: Modelo C 5.0. Importancia de un predictor

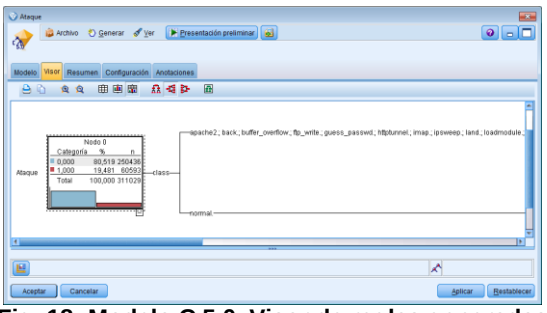

Fig. 18: Modelo C 5.0. Visor de reglas generadas

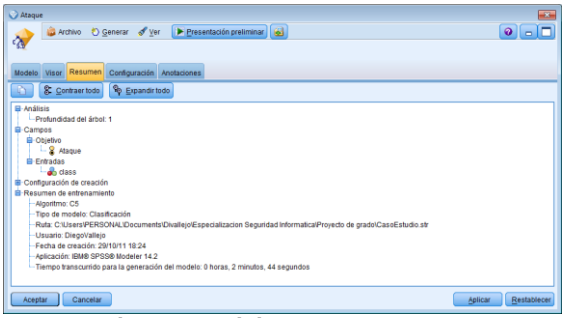

Fig. 19: Modelo C 5.0. Resumen

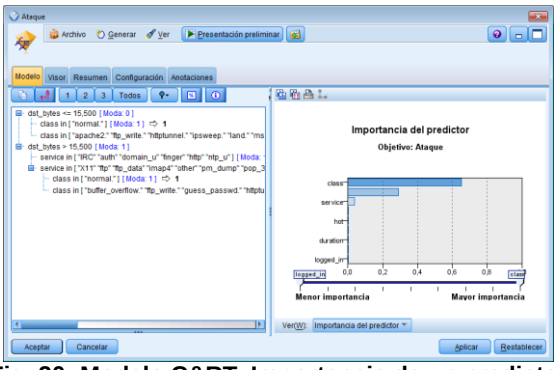

Fig. 20: Modelo C&RT. Importancia de un predictor

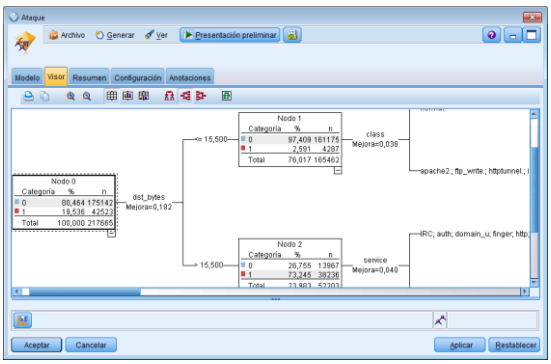

Fig. 21: Modelo C&RT. Visor de reglas generadas

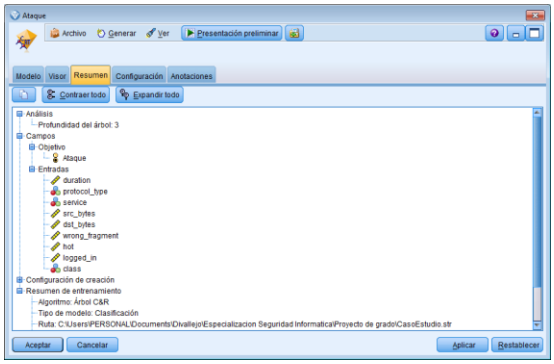

Fig. 22: Modelo C&RT. Resumen

De la información generada por cada uno de los modelos es pueden comenzar a diseñar reglas y patrones que indican cuál es el modelo que más puede adecuarse a la necesidad que se tiene.

5. **Evaluación de resultados.** Luego de evaluar cada uno de los modelos según la necesidad esperada, se procede a revisar los resultados entregados por el modelo, es de esta manera como se obtiene un mejor acercamiento a los datos entregados. En nuestro caso el modelo que mejor comportamiento tiene en evaluar la variable "Ataque" vs las otras variables de entrada fue el modelo CHAID.

6. **Explotación o despliegue de resultados.** Se puede pasar a realizar el despliegue de los resultados, en el cual se detallan cada uno de los puntos que el modelo ayuda a predecir. En la Figura 23 se observa un detalle de la información que suministra Modeler respecto al modelo utilizado.

Campos
Objetivo
Ataque
Entradas
class
duration
protocol_type
service
src_bytes
dst_bytes
wrong_fragment
hot
logged_in
flag
Configuración de creación
Utilizar los datos en particiones: falso
Resumen de entrenamiento
Ruta: D:\divallejo\Especializacion Seguridad Informatica\Proyecto de grado\CasoEstudio.str
Usuario: dvallejo
Fecha de creación: 27/10/11 03:58 PM
Aplicación: PASW® Modeler 14
Modelos planificados: 4
Modelos finalizados: 4
Modelos descartados según los resultados finales: 1
Modelos que no se han podido generar o puntuar: 0
Modelos no finalizados debidos a una interrupción: 0
Tiempo transcurrido para la generación del modelo: 0 horas, 3 minutos, 11 segundos
Detalles del modelo
C5 1
Análisis
Profundidad del árbol: 1
Campos
Objetivo
Ataque
Entradas
class
Configuración de creación
Utilizar los datos en particiones: falso
Calcular importancia de predictor: falso
Calcular puntuaciones brutas de propensión: falso
Calcular puntuaciones de propensión ajustada: falso
Utilizar ponderación: falso
Tipo de resultados: Árbol de decisión
Agrupar simbólicos: falso
Utilizar aumento: falso
Efectuar validación cruzada: falso
Modo: Simple
Favorecer: Precisión
Ruido esperado (%): 0
Utilizar costes de clasificación errónea: falso

Resumen de entrenamiento
Algoritmo: C5
Tipo de modelo: Clasificación
Ruta: D:\divallejo\Especializacion Seguridad Informatica/Proyecto de grado\CasoEstudio.str
Usuario: dvallejo
Fecha de creación: 27/10/11 03:55 PM
Aplicación: PASW® Modeler 14
Tiempo transcurrido para la generación del modelo: 0 horas, 2 minutos, 55 segundos
Árbol C&R 1
Análisis
Profundidad del árbol: 3
Campos
Objetivo
Ataque
Entradas
duration
protocol_type
service
src_bytes
dst_bytes
wrong_fragment
hot
logged_in
class
Configuración de creación
Utilizar los datos en particiones: falso
Calcular importancia de predictor: verdadero
Calcular puntuaciones brutas de propensión: falso
Calcular puntuaciones de propensión ajustada: falso
Utilizar frecuencia: falso
Utilizar ponderación: falso
Niveles por debajo del raíz: 5
Modo: Experto
Número máximo de sustitutos: 5
Cambio mínimo en la impureza: 0,0
Medida de impureza para objetivos categóricos: Gini
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Podar árbol: verdadero
Utilizar regla de error típico: falso
Probabilidades previas: Basadas en datos de entrena.
Corregir previas por costes de clasificación errónea: falso
Utilizar costes de clasificación errónea: falso
Resumen de entrenamiento
Algoritmo: Árbol C&R
Tipo de modelo: Clasificación
Ruta: D:\divallejo\Especializacion Seguridad Informatica/Proyecto de grado\CasoEstudio.str
Usuario: dvallejo
Fecha de creación: 27/10/11 03:55 PM
Aplicación: PASW® Modeler 14
Tiempo transcurrido para la generación del modelo: 0 horas, 2 minutos, 55 segundos
CHAID 1
Análisis
Profundidad del árbol: 2
Campos
Objetivo
Ataque
Entradas
flag
src_bytes
dst_bytes
class
Configuración de creación
Utilizar los datos en particiones: falso
Calcular importancia de predictor: verdadero
Calcular puntuaciones brutas de propensión: falso
Calcular puntuaciones de propensión ajustada: falso
Continuar entrenando modelo existente: falso
Utilizar frecuencia: falso
Utilizar ponderación: falso
Niveles por debajo del raíz: 5
Alfa para división: 0,05
Alfa para fusión: 0,05
Épsilon para convergencia: 0,001

Número máximo de iteraciones para la convergencia: 100
Utilizar corrección de Bonferroni: verdadero
Permitir división de categorías fusionadas: falso
Método de chi-cuadrado: Pearson
Criterios de parada: Utilizar porcentaje
Número mínimo de registros en rama parental (%): 2
Número mínimo de registros en rama filial (%): 1
Utilizar costes de clasificación errónea: falso
Resumen de entrenamiento
Algoritmo: CHAID
Tipo de modelo: Clasificación
Ruta: D:\divallejo\Especializacion Seguridad Informatica\Proyecto de grado\CasoEstudio.str
Usuario: dvallejo
Fecha de creación: 27/10/11 03:55 PM
Aplicación: PASW® Modeler 14
Tiempo transcurrido para la generación del modelo: 0 horas, 2 minutos, 55 segundos

Fig. 23: Resultados entregados por IBM PASS MODELER

7. CONCLUSIONES

La minería de datos es el proceso de ahondar en los datos para detectar patrones y relaciones ocultos; emplea una orientación empresarial clara y potentes tecnologías analíticas para explorar rápida y concienzudamente montañas de datos y extraer de ellas la información útil y aplicable que se necesita.

Es precisamente por lo expresado anteriormente que pudimos determinar que la minería de datos basada en una metodología adecuada, puede ser muy útil en el proceso de exploración de datos, toda vez que mediante tecnologías analíticas y procesos estadísticos nos permitió generar reglas a partir de datos históricos de capturas, para generar reglas y patrones que permiten predecir intrusiones.

A la fecha se están realizando investigaciones que permiten la detección de intrusos en tiempo real utilizando técnicas de minería de datos y análisis estadísticos [8, 9].

8. AGRADECIMIENTOS

A Sergio Gutiérrez Bonnet, Presidente de la Empresa Informese IBM SPSS Modeler. A Dougglas Hurtado Carmona, M.Sc. Ingeniería de Sistemas y Computación, asesor de nuestro artículo. A Francisco Antonio Ruiz Escobar, Director Dirección de Cumplimiento, Bancolombia.

9. REFERENCIAS

[1] M. S. Shin & K. J. Jeong. "Alert Correlation Analysis in Intrusion Detection". Proceedings of the Second international conference on Advanced Data Mining and Applications ADMA'06, pp. 1049-1056, 2006.

[2] M. Xue & C. Zhu. "Applied Research on Data Mining Algorithm in Network Intrusion Detection". Proceedings International Joint Conference on Artificial Intelligence JCAI '09, pp. 275-277, 2009.

[3] M. Castellano & G. B. de Grecis. "Applying a Flexible Mining Architecture to Intrusion Detection". Proceedings Second International Conference on Availability, Reliability and Security, ARES 2007, pp. 845-852, 2007.

[4] IBM. SPSS Modeler. Online [Jun. 2011].

[5] KDD Cup 1999 Data. Online [Jun. 2011].

[6] P. Chapman et al. "CRISP-DM 1.0: Step-by-step data mining guide". SPSS Inc., 2000.

[7] Wikipedia. "Ataque de denegación de servicio". Online [Jun. 2011].

[8] L. Wenke et al. "Real Time Data Mining-based Intrusion Detection". DARPA Information Survivability Conference and Exposition, pp. 89-101, 2000.

[9] L. Zenghui & L. Yingxu. "A Data Mining Framework for Building Intrusion Detection Models Based on IPv6". Proceedings of the 3rd International Conference and Workshops on Advances in Information Security and Assurance ISA '09, pp. 608-628, 2009.