UNIVERSIDAD NACIONAL DE INGENIERÍA

FACULTAD DE INGENIERÍA ECONÓMICA, ESTADÍSTICA Y CIENCIAS SOCIALES
ESCUELA PROFESIONAL DE INGENIERÍA ESTADÍSTICA

Design of an Statistical Systems for Classifying of Financial Fraud Using Neural Networks

TESIS

Para optar el Título Profesional de Ingeniero Estadístico

Por la modalidad de Tesis

Elaborado por:

Huamani Gonzales L. Dayana

LIMA – PERÚ
2016

Dedicatoria

A mis padres Rafael Huamani Valencia y
Adriana Gonzales Tapia por creer siempre en mí y
enseñarme a nunca rendirme, a mi hermana Diana
por sus consejos cuando los necesité y su apoyo incondicional.

A la memoria de mi primo Willy Sandoval, quien fue
mi mejor amigo y del que siempre me sentiré
orgullosa.

Agradecimientos

A la Universidad Nacional de Ingeniería, por brindarme la oportunidad de ser una profesional, a la escuela de Ingeniera Estadística por permitirme desarrollar habilidades para mi desarrollo profesional y personal.

A todos mis profesores por el conocimiento impartido todos estos años.

A mi profesor del curso de Elaboración de Tesis, Richard Fernández Vásquez por enseñarme a trabajar con disciplina y constancia, por guiarme desde el primer día y mostrarme mis errores para corregirlos.

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. 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 poder ser capaz de extraer nuevos conocimientos a partir de las experiencias. Estas características pueden ser de vital importancia para ser aplicadas en la seguridad de la información a través de la detección de intrusos. En el presente trabajo se pretende mostrar el aporte a la seguridad de la información utilizando la Superficie de Respuesta como una técnica que con la experiencia del investigador podrá precaver las nuevas modalidades de robo de información.

Palabras clave

Denegación de servicios, ANN, intrusiones, minería de datos, modelo, predicción.

ABSTRACT

Based on the fundamentals and techniques of data mining, models can be designed and elaborated that allow finding clandestine behaviors that are easily detected by the naked eye. In particular the usefulness of data mining in this area lies in a series of techniques, algorithms and methods that mimic the human characteristic of learning to be able to extract new knowledge from experiences. These characteristics can be of vital importance to be applied in information security through the detection of intruders. This paper aims to show the contribution to the information security using the Response Surface as a technique that with the experience of the investigator can prevent the new modalities of information theft.

Keywords

Denial of services, ANN, intrusions, data mining, model, prediction.

INDICE

[i]Capítulo I

1. Antecedentes:

1.1 Investigaciones

DETECCIÓN DE INTRUSOS EN REDES DE TELECOMUNICACIONES IP USANDO MODELOS OCULTOS DE MARKOV
Enero -2009

En los primeros capítulos se abordan los Sistemas de Detección de Intrusos, sus características, los tipos existentes, las arquitecturas y los mecanismos de detección más convencionales, así como los ataques de red más comunes y una metodología de intrusión ampliamente aceptada.
Estudia como forma alternativa los principios de la teoría de los Modelos Ocultos de Markov, sus características y se describen los procesos de entrenamiento, decodificación y evaluación.
Posteriormente, recoge la teoría de los capítulos anteriores y define un esquema para detectar intrusiones de tráfico en redes de telecomunicaciones IP usando Modelos Ocultos de Markov, describe el proceso de entrenamiento, las pruebas realizadas y los resultados obtenidos, finalmente presenta las conclusiones del proyecto de investigación y las recomendaciones para trabajo futuro a partir de este nuevo esquema.
Como resumen, describe las dificultades ocurridas en el desarrollo de la investigación, a pesar de tener buenos resultados en la

detección de intrusos en el tráfico de redes, no se puede generar el modelo en tiempo real por la gran cantidad de información procesada.
Se observó mediante experiencia del investigador que el conjunto de trazas de IDEVAL es, según la búsqueda realizada y varios autores, la única fuente disponible para realizar pruebas sobre sistemas de detección de intrusos de acceso público.

DETECTING EVOLUTIONARY FINANCIAL STATEMENT FRAUD
Agosto-2010

En el presente estudio se examinan las técnicas de minería de datos como la regresión, árboles de decisión, redes neuronales y las redes Bayesianas como una ayuda para identificar el fraude en los estados financieros. La eficacia de estos métodos (y sus limitaciones) se examinaron sobre todo porque los nuevos esquemas de fraude en los estados financieros se adaptan a las técnicas de detección. Luego explora un marco auto- adaptativo (basado en un modelo de superficie de respuesta) con conocimiento del tema para detectar fraude en los estados financieros. Los autores concluyen sugiriendo que, en una era con fraudes financieros evolutivos, los mecanismos de detección de fraude automatizado asistido por ordenador serán más eficaces y eficientes con el conocimiento del dominio especializado.
En los tres primeros puntos el autor proporciona una visión general de las técnicas de detección existentes basadas en la minería para detectar fraudes financieros existentes y su tendencia, luego se propone un nuevo marco evolutivo para detectar el fraude de estados financieros. Se revisó la aplicación de la regresión, árboles de decisión, redes neuronales y las redes de creencias bayesianas analizándose la eficacia y sus limitaciones en cuanto a las tecnologías. Se encontraron los principales inconvenientes que presentan estos modelos y para tener mayor conocimiento acerca del tema se estudió la historia y tendencia a través del tiempo del fraude financiero. Acontinuación, sugerimos un marco que aborda el problema cuando el fraude financiero emergente es evolutivo. La sección 6 concluye el artículo con una breve discusión sobre las ideas cosechó y posible investigación futura. El autor no descarta que exista un mejor método que el RSM.

MÉTODO PARA LA DETECCIÓN DE INTRUSOS MEDIANTE REDES NEURONALES BASADO EN LA REDUCCIÓN DE CARACTERÍSTICAS
2010

La aplicación de técnicas basadas en Inteligencia Artificial para la detección de intrusos (IDS), fundamentalmente las redes neuronales artificiales (ANN), están demostrando ser un enfoque muy adecuado para paliar muchos de los problemas abiertos en esta área. Sin embargo, el gran volumen de información que se requiere cada día para entrenar estos sistemas, junto con la necesidad exponencial de tiempo que requieren para asimilarlos, dificulta enormemente su puesta en marcha en escenarios reales.
El trabajo presente propone un método basado en la aplicación de una técnica para la reducción de características, denominada Análisis de Componentes Principales (PCA), asegurando que la pérdida de información sea mínima y, en consecuencia, disminuyendo la complejidad del clasificador neuronal y manteniendo estables los tiempos de entrenamiento. Para validar la propuesta se ha diseñado un escenario de prueba mediante un IDS basado en ANN. Los resultados obtenidos a partir de las pruebas realizadas demuestran la validez de la propuesta y acreditan las líneas futuras de trabajo.

MODELO DE DETECCIÓN DE INTRUSIONES EN SISTEMAS DE RED, REALIZANDO SELECCIÓN DE CARACTERÍSTICAS CON FDR Y ENTRENAMIENTO Y CLASIFICACIÓN CON SOM1
Septiembre – 2012

Los Sistemas de Detección de Intrusos comerciales actuales clasifican el tráfico de red, detectando conexiones normales e intrusiones, mediante la aplicación de métodos basados en firmas; ello conlleva problemas pues solo se detectan intrusiones previamente conocidas y existe desactualización periódica de la base de datos de firmas. En este artículo se evalúa la eficiencia de un modelo de detección de intrusiones de red propuesto, utilizando métricas de sensibilidad y especificidad, mediante un proceso de simulación que emplea el dataset NSL-KDD DARPA, seleccionando de éste las características más relevantes con FDR y entrenando una red neuronal que haga uso de un algoritmo de aprendizaje no supervisado basado en mapas auto-organizativos, con el propósito de clasificar el tráfico de la red en conexiones normales y ataques, de forma automática.

Modelo propuesto El modelo comprende tres fases: entrenamiento, clasificación y cálculo de métricas de desempeño. Para su aplicación se implementaron varios escenarios de simulación variando la cantidad de características por evaluar en las dos primeras fases; para ello se priorizó la escogencia de las características mediante su razón discriminante. A continuación, se aplica el algoritmo de reducción de características FDR y se seleccionan las características por orden de relevancia y, por último, se realiza el entrenamiento del SOM, lo cual implica una normalización, creación de la estructura de datos, inicialización del mapa, entrenamiento del mismo y un etiquetado de los datos.

Se procede por último a clasificar los datos, generando una estructura de datos que contiene tanto el etiquetado de la nueva data como el etiquetado predictivo a partir del cálculo de las BMU basado en el mapa creado en la fase de entrenamiento.

En la fase final se calculan las métricas de desempeño; para ello se recorre la estructura de datos generada en la fase anterior, calculando falsos positivos, verdaderos positivos, falsos negativos y verdaderos negativos, los cuales permiten determinar las métricas de sensibilidad y especificidad que van a indicar la eficiencia del modelo planteado.

A SURVEY OF INTRUSION DETECTION TECHNIQUES IN CLOUD

Enero- 2013

En este trabajo, examinamos diferentes ataques que afectan la disponibilidad, confidencialidad e integridad de los recursos y servicios en la nube. Las propuestas que incorporan Intrusion Detection Systems (IDS) y sistemas de prevención de intrusiones (IPS) en la nube son examinados. Recomendamos IDS / IPS posicionamiento en el entorno de la nube para lograr la seguridad deseada en las redes de próxima generación.

Capítulo II

2. PLANTEAMIENTO DEL PROBLEMA

2.1 Descripción del problema

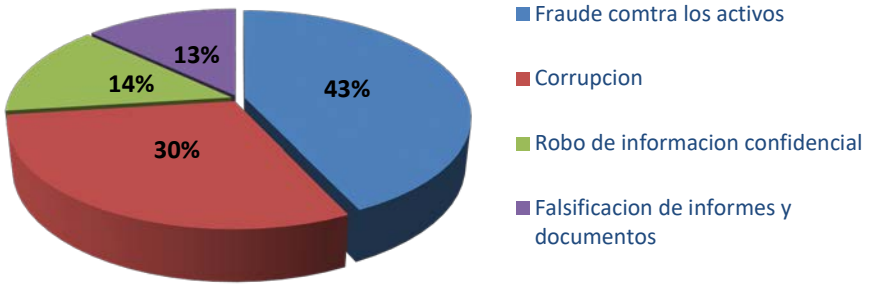
La economía peruana se ha caracterizado estos últimos años por el crecimiento sostenido, siendo identificado como fuente de oportunidades de negocio.

Como consecuencia las empresas nacionales se han afianzado y varios grupos económicos peruanos ya se han expandido no solo en Perú, sino a sus alrededores, las inversiones internacionales también han llegado a consolidarse en todas las zonas del Perú.

El crecimiento y la inversión trae consigo ciertos riesgos, las empresas se adecuan a las especificaciones que trae cada mercado y esto ha llevado a considerar el fraude y la corrupción como uno costo más al hacer negocios.

Revisando investigaciones concernientes al fraude financiero, se detectó que en el 2012 el Perú era una de las regiones con mayor índice de fraude y esto no ha cambiado en los últimos años. Se identificó cuáles eran los casos de fraude más reportados a través de una aproximación realizando una encuesta a diferentes empresas nacionales e internacionales radicadas en Perú.

Grafico N° 2.1
Casos de fraude reportados



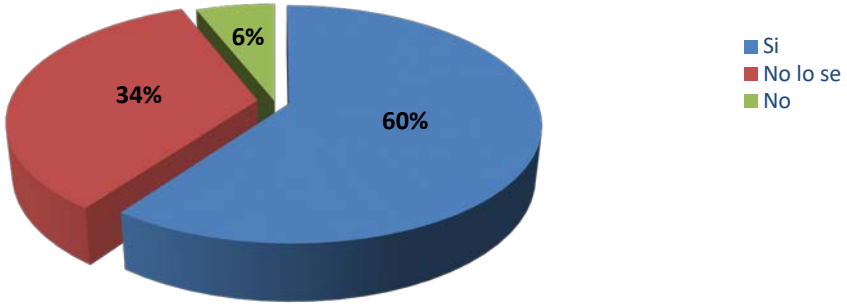
FUENTE: Informe de fraude en el Perú-KPMG. 2012

Como se observa, las empresas enfrentan todo tipo de fraudes, pero en la presente tesis nos centraremos en el robo de información confidencial que ocupa el tercer puesto en los casos de fraude más reportados.

Con la mayor dependencia de la tecnología para la administración de los negocios, se incrementa el riesgo de que a través de la manipulación de datos o programas informáticos se busque defraudar a una Compañía.

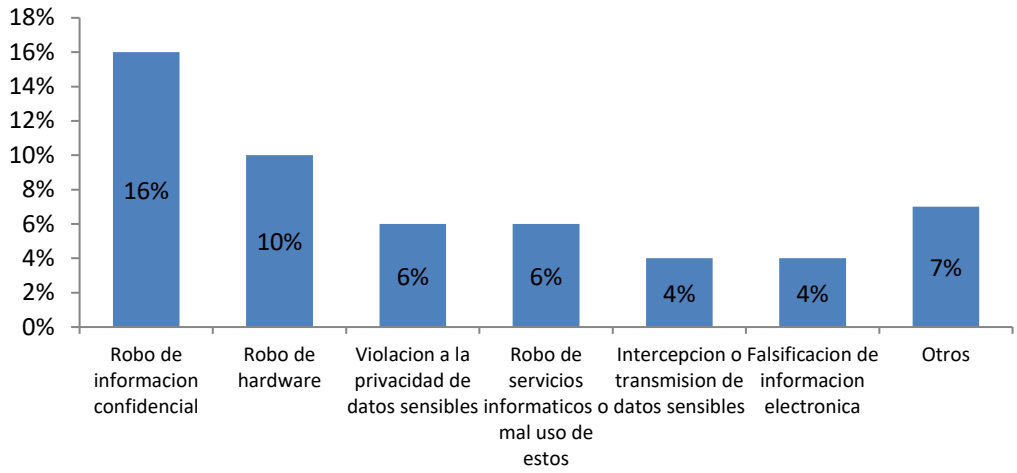
El siguiente gráfico refleja que tanto saben las empresas si fue víctima de un fraude informático y las modalidades detectadas:

Grafico N° 2.2
¿ Sabe Ud. si su organizacion fue victima de fraude informatico en el año 2011?



FUENTE: Informe de fraude en el Perú-KPMG. 2012

Grafico N° 2.3
Modalidades de delitos informáticos



FUENTE: Informe de fraude en el Perú-KPMG. 2012

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 ya que son fuente de pérdida constante de dinero porque se comparte información con usuarios no deseados que hacen mal uso de dicha información.

La literatura revisada concluye que existen modelos de segmentación que encuentran patrones en los datos en los cuales una o más entidades, como eventos, compras o atributos, se asocian con una o más entidades. Estos 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. Sin embargo, así como hemos desarrollado métodos para detectar anomalías cuando se comparte información los usuarios o hackers encuentran la manera de burlar estos patrones, generando nuevos comportamientos para la intrusión en los sistemas de información.

2.2 Formulación del problema

2.2.1 Problema general

¿Cuáles son los efectos de incluir previamente el modelo de superficie de respuesta que selecciona las variables independientes para optimizar la clasificación entre las conexiones malignas y conexiones normales frente a la aplicación directa del modelo de Redes neuronales en los datos simulados de las USAF de EEUU en el periodo de 1999?

2.2.2 Problemas específicos

- ¿Cuál es el método de optimización para encontrar los valores de las variables independientes que producen valores deseables en la respuesta?
- ¿Cuáles son las variables independientes que producen valores deseables en la respuesta?
- ¿El modelo de Redes Neuronales que usa variables óptimas obtenidas por el modelo de superficie de respuesta tiene mayor especificidad que el modelo de Redes Neuronales?

2.3 OBJETIVOS DE LA INVESTIGACIÓN

2.3.1 Objetivo general

Demostrar que el modelo de superficie de respuesta que selecciona las variables independientes para optimizar la clasificación entre las conexiones malignas y conexiones normales tiene una mejor predicción frente a la aplicación directa del modelo de ANN en los datos simulados de las USAF de EEUU en el periodo de 1999.

2.3.2 Objetivos específicos

- Estudiar el modelo de primer orden y de segundo orden.
- Encontrar cuáles de las variables independientes principales están asociados a la optimización de la variable dependiente (conexión maligna/conexión normal).

- Predecir las conexiones malignas y conexiones normales a través del modelo ANN usando variables seleccionadas por el RSM.

2.4 HIPÓTESIS

2.4.1 HIPOTESIS GENERAL

El modelo generado con RSM que selecciona las variables óptimas para la generación del modelo de Redes Neuronales tiene un coeficiente de Gini mayor a 96% comparado con el modelo de Redes Neuronales con metodología clásica.

2.4.2 Hipótesis específicas

- El modelo de segundo orden tiene un R-cuadrado ajustado mayor al 60% a comparación del modelo de primer orden.
- Las variables wrong_fragment, protocol_type,service, land y count son significativas en el modelo de Segundo orden.
- El ratio de especificidad es mayor para el modelo de Redes Neuronales con uso del RSM.

2.5 JUSTIFICACIÓN

El aumento en el volumen y variedad de información, el auge del internet y otras tecnologías han traído consigo un aumento en los esquemas fraudulentos. Actualmente empresas bancarias como BCP, IBK, BBVA sufren fraude para clonación de tarjetas de crédito, las empresas de telefonía e instituciones del gobierno como MIT, se ven afectadas por personas que ingresan a los sistemas de seguridad de las redes donde buscan los puntos más débiles para poder obtener información confidencial o causar daño en los sistemas operativos con el fin de obtener otro tipo de beneficio.

Conociendo el contexto actual, el presente estudio pretende de manera exploratoria mostrar el aporte del RSM en la detección de intrusos a los usuarios de seguridad de información para producir alertas con el fin de detectar los factores que ponen en peligro la confidencialidad, integridad, disponibilidad de la información.

Según DmitryBestuzhev, director del equipo de Investigación y Análisis para América Latina de la compañía rusa Seguridad e Informática KasperskyLab, ciudadanos y organizaciones de México, Venezuela y Perú son víctimas de entre el 26 % y 36 % de los ataques en la red**,** delitos que incluyen robo de información financiera y personal, ciber espionaje, sabotaje, eliminación de datos o daños a la reputación corporativa. Así mismo con el creciente auge de la conectividad en internet, se estima que Perú por el alto índice en generadores de malware llegaría en 10 años a ser el segundo país latinoamericano con problemas de intrusos en las redes.

Por lo tanto, este documento pretende ayudar a los usuarios de seguridad de información ya sea de bancos, instituciones del gobierno (SUNAT, Indecopi,etc) en la mejora en los procesos para detectar intrusos en la red.

Capítulo III

3. MARCO TEORICO

3.1 ANTECEDENTES

DETECCIÓN DE INTRUSOS EN REDES DE TELECOMUNICACIONES IP USANDO MODELOS OCULTOS DE MARKOV
Enero -2009

En los primeros capítulos se abordan los Sistemas de Detección de Intrusos, sus características, los tipos existentes, las arquitecturas y los mecanismos de detección más convencionales, así como los ataques de red más comunes y una metodología de intrusión ampliamente aceptada.

Estudia como forma alternativa los principios de la teoría de los Modelos Ocultos de Markov, sus características y se describen los procesos de entrenamiento, decodificación y evaluación.

Posteriormente, recoge la teoría de los capítulos anteriores y define un esquema para detectar intrusiones de tráfico en redes de telecomunicaciones IP usando Modelos Ocultos de Markov, describe el proceso de entrenamiento, las pruebas realizadas y los resultados obtenidos, finalmente presenta las conclusiones del proyecto de

investigación y las recomendaciones para trabajo futuro a partir de este nuevo esquema.

Como resumen, describe las dificultades ocurridas en el desarrollo de la investigación, a pesar de tener buenos resultados en la detección de intrusos en el tráfico de redes, no se puede generar el modelo en tiempo real por la gran cantidad de información procesada.

Se observó mediante experiencia del investigador que el conjunto de trazas de IDEVAL es, según la búsqueda realizada y varios autores, la única fuente disponible para realizar pruebas sobre sistemas de detección de intrusos de acceso público.

DETECTING EVOLUTIONARY FINANCIAL STATEMENT FRAUD

Agosto-2010

En el presente estudio se examinan las técnicas de minería de datos como la regresión, árboles de decisión, redes neuronales y las redes Bayesianas como una ayuda para identificar el fraude en los estados financieros. La eficacia de estos métodos (y sus limitaciones) se examinaron sobre todo porque los nuevos esquemas de fraude en los estados financieros se adaptan a las técnicas de detección. Luego explora un marco auto- adaptativo (basado en un modelo de superficie de respuesta) con conocimiento del tema para detectar fraude en los estados financieros. Los autores concluyen sugiriendo que, en una era con fraudes financieros evolutivos, los mecanismos de detección de fraude automatizado asistido por ordenador serán más eficaces y eficientes con el conocimiento del dominio especializado.

En los tres primeros puntos el autor proporciona una visión general de las técnicas de detección existentes basadas en la minería para detectar fraudes financieros existentes y su tendencia, luego se propone un nuevo marco evolutivo para detectar el fraude de estados financieros. Se revisó la aplicación de la regresión, árboles de decisión, redes neuronales y las redes de creencias bayesianas analizándose la eficacia y sus limitaciones en cuanto a las tecnologías. Se encontraron los principales inconvenientes que presentan estos modelos y para tener mayor conocimiento acerca del tema se estudió la historia y tendencia a través del tiempo del fraude financiero. Acontinuación, sugerimos un marco que aborda el problema cuando el

fraude financiero emergente es evolutivo. La sección 6 concluye el artículo con una breve discusión sobre las ideas cosechó y posible investigación futura. El autor no descarta que exista un mejor método que el RSM.

MÉTODO PARA LA DETECCIÓN DE INTRUSOS MEDIANTE REDES NEURONALES BASADO EN LA REDUCCIÓN DE CARACTERÍSTICAS

2010

La aplicación de técnicas basadas en Inteligencia Artificial para la detección de intrusos (IDS), fundamentalmente las redes neuronales artificiales (ANN), están demostrando ser un enfoque muy adecuado para paliar muchos de los problemas abiertos en esta área. Sin embargo, el gran volumen de información que se requiere cada día para entrenar estos sistemas, junto con la necesidad exponencial de tiempo que requieren para asimilarlos, dificulta enormemente su puesta en marcha en escenarios reales.
El trabajo presente propone un método basado en la aplicación de una técnica para la reducción de características, denominada Análisis de Componentes Principales (PCA), asegurando que la pérdida de información sea mínima y, en consecuencia, disminuyendo la complejidad del clasificador neuronal y manteniendo estables los tiempos de entrenamiento. Para validar la propuesta se ha diseñado un escenario de prueba mediante un IDS basado en ANN. Los resultados obtenidos a partir de las pruebas realizadas demuestran la validez de la propuesta y acreditan las líneas futuras de trabajo.

MODELO DE DETECCIÓN DE INTRUSIONES EN SISTEMAS DE RED, REALIZANDO SELECCIÓN DE CARACTERÍSTICAS CON FDR Y ENTRENAMIENTO Y CLASIFICACIÓN CON SOM1

Septiembre – 2012

Los Sistemas de Detección de Intrusos comerciales actuales clasifican el tráfico de red, detectando conexiones normales e intrusiones, mediante la aplicación de métodos basados en firmas; ello conlleva problemas pues solo se detectan intrusiones previamente conocidas y existe desactualización periódica de la base de datos de firmas. En este artículo se evalúa la eficiencia de un modelo de detección de intrusiones de red propuesto, utilizando métricas de sensibilidad y

especificidad, mediante un proceso de simulación que emplea el dataset NSL-KDD DARPA, seleccionando de éste las características más relevantes con FDR y entrenando una red neuronal que haga uso de un algoritmo de aprendizaje no supervisado basado en mapas auto-organizativos, con el propósito de clasificar el tráfico de la red en conexiones normales y ataques, de forma automática.

Modelo propuesto El modelo comprende tres fases: entrenamiento, clasificación y cálculo de métricas de desempeño. Para su aplicación se implementaron varios escenarios de simulación variando la cantidad de características por evaluar en las dos primeras fases; para ello se priorizó la escogencia de las características mediante su razón discriminante. A continuación, se aplica el algoritmo de reducción de características FDR y se seleccionan las características por orden de relevancia y, por último, se realiza el entrenamiento del SOM, lo cual implica una normalización, creación de la estructura de datos, inicialización del mapa, entrenamiento del mismo y un etiquetado de los datos.

Se procede por último a clasificar los datos, generando una estructura de datos que contiene tanto el etiquetado de la nueva data como el etiquetado predictivo a partir del cálculo de las BMU basado en el mapa creado en la fase de entrenamiento.

En la fase final se calculan las métricas de desempeño; para ello se recorre la estructura de datos generada en la fase anterior, calculando falsos positivos, verdaderos positivos, falsos negativos y verdaderos negativos, los cuales permiten determinar las métricas de sensibilidad y especificidad que van a indicar la eficiencia del modelo planteado.

A SURVEY OF INTRUSION DETECTION TECHNIQUES IN CLOUD

Enero- 2013

En este trabajo, examinamos diferentes ataques que afectan la disponibilidad, confidencialidad e integridad de los recursos y servicios en la nube. Las propuestas que incorporan IntrusionDetectionSystems (IDS) y sistemas de prevención de intrusiones (IPS) en la nube son examinados. Recomendamos IDS / IPS posicionamiento en el entorno de la nube para lograr la seguridad deseada en las redes de próxima generación.

3.2 BASES TEORICAS

3.2.1 TECNICAS PREVIAS

I. METODOLOGIA DE SUPERFICIE DE RESPUESTA

Definición:

La metodología de superficie de respuesta, o MSR, es una colección de técnicasmatemáticas y estadísticasútiles en el modelado y el análisis de problemas en los que una respuesta de interés recibe la influencia de diversas variables y donde el objetivo es optimizar esta respuesta. Por ejemplo, suponga que un ingeniero quiere encontrar los niveles de temperatura (X1) y presión (X2) que maximicen el rendimiento (y) de un proceso. El rendimiento del proceso es una función de los niveles de la temperatura y la presión, por ejemplo

$$y = f(x_1, x_2) + \varepsilon$$

Donde ε representa el ruido o error observado en la respuesta y. si la respuesta esperada se denota por $E(y) = f(x_1, x_2) = \eta$, entonces a la superficie representada por $\eta = f(x_1, x_2)$ se le llama **superficie de respuesta**.

Por lo general la superficie de respuesta se representa gráficamente como en la Figura 3.1, donde η se grafica contra los niveles de $x_1 \, y \, x_2$. Para ayudar a visualizar la forma de una superficie de respuesta, con frecuencia se grafican los contornos de la superficie de respuesta, como se muestra en la Figura 3.2.En la gráfica de contorno se trazan las líneas de respuesta constante en el plano $x_1 \, y \, x_2$. Cada contorno corresponde a una altura particular de la superficie de respuesta. También se ha visto antes la utilidad de las gráficas de contorno.

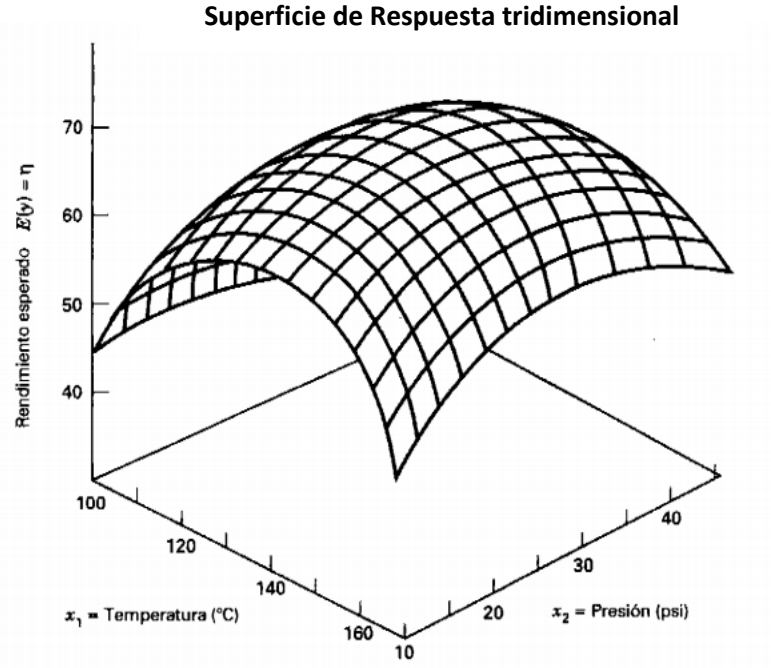
En la mayoría de los problemas de MSR, la forma de la relación entre la respuesta y las variables independientes es desconocida. Por lo tanto, el primer paso de la MSR es encontrar una aproximación adecuada de la verdadera relación funcional entre y y

el conjunto de variables independientes. Por lo general se emplea un polinomio de orden inferior en alguna región de las variables independientes. Si la respuesta está bien modelada por una función lineal de las variables independientes, entonces la función de aproximación es el modelo de primer orden

$$y = \beta_{0+}\beta_1$$

Figura N° 3.1

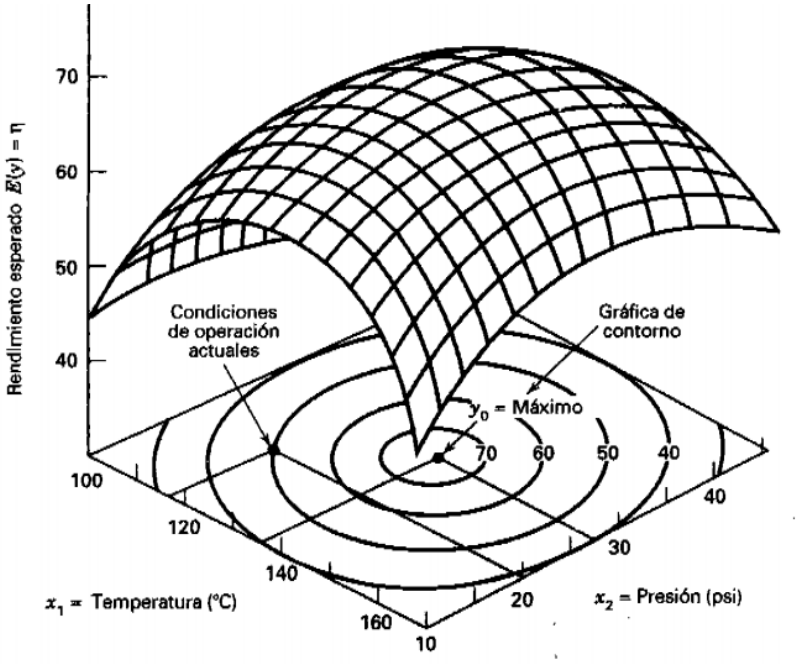
Superficie de Respuesta tridimensional



FUENTE: Montgomery. Diseño y análisis de experimentos. 2da edición.

Figura N° 3.2

Gráfico de contorno de una Superficie de Respuesta



FUENTE: Montgomery. Diseño y análisis de experimentos. 2da edición.

Si hay curvatura en el sistema, entonces debe usarse un polinomio de orden superior, tal como el modelo de segundo orden

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i<j} \sum \beta_{ij} x_i x_j + \varepsilon$$

En casi todos los problemas de MSR se usa uno de estos modelos, o ambos. Desde luego, es probable que un modelo polinomio sea una aproximación razonable de la verdadera relación funcional en el espacio completo de las variables independientes, pero para una región relativamente pequeña suelen funcionar bastante bien.

El método de mínimos cuadrados, se usa para estimar los parámetros de los polinomios de aproximación. Después se realiza el análisis de la superficie de respuesta utilizando la superficie ajustada. Si la superficie ajustada es una aproximación adecuada de la verdadera función de la respuesta, entonces el análisis de la superficie ajustada será un equivalente aproximado del análisis del sistema real. Los parámetros del modelo pueden estimarse de manera más eficiente cuando se emplean los diseños experimentales apropiados para recolectar los datos. Los diseños

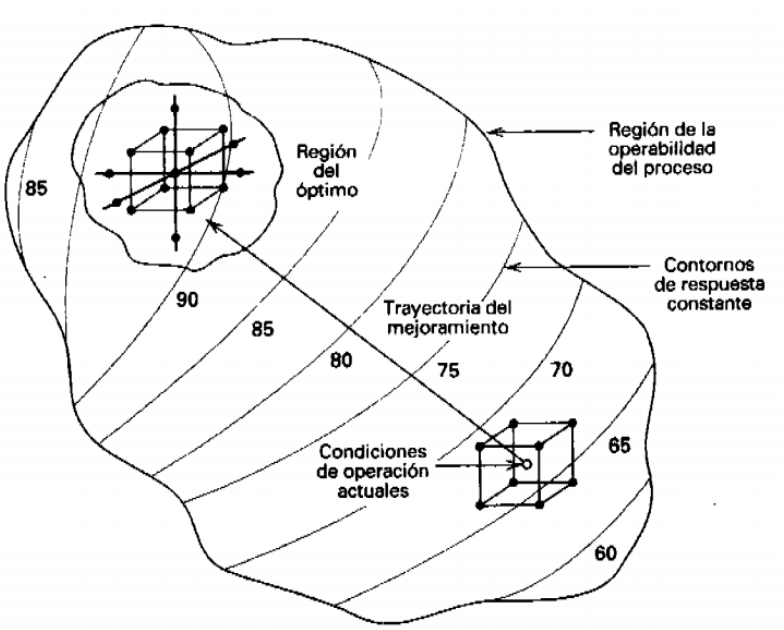
para ajustar superficies de respuesta se denominan diseños de superficie de respuesta.

La MSR es un procedimiento secuencial. Muchas veces, cuando se está en un punto de la superficie de respuesta que esta apartado del optimo, como en el caso de las condiciones de operación actuales de la Figura 3.3,el sistema presenta una curvatura moderada y el modelo de primer orden será apropiado. El objetivo en este caso es llevar al experimentador de manera rápida y eficiente por la trayectoria del mejoramiento hasta la vecindad general del óptimo. Una vez que se ha encontrado la región del óptimo, puede emplearse un modelo más elaborado, como el de segundo orden, y llevarse a cabo un análisis para localizar el óptimo. En la Figura 3.3sepuede ver que el análisis de una superficie de respuesta puede considerarse como “el ascenso a una colina”, donde la cima de esta representa el punto de la respuesta máxima.

Si el verdadero optimo es un punto de respuesta mínima, entonces la situación puede considerarse como “el descenso de un valle”.

El objetivo último de la MSR es determinar las condiciones de operación óptimas del sistema o determinar una región del espacio de los factores en la que se satisfagan los requerimientos de operación.

Figura N° 3.3
Carácter secuencial del Modelo de Superficie de Respuesta



FUENTE: Montgomery. Diseño y análisis de experimentos. 2da edición.

Polinomio de primer orden: Generalmente se desconoce la relación entre la respuesta y las variables independientes, por ello requerimos un modelo que aproxime la relación funcional entre Y y las variables independientes. Si la respuesta se describe adecuadamente por una función lineal de las variables independientes se utiliza el modelo de primer orden:

Los parámetros del modelo se estiman mediante el método de mínimos cuadrados. Una vez que se tienen los estimadores se sustituyen en la ecuación y obtenemos el modelo ajustado:

Donde la matriz X puede escribirse alternativamente como X = [1 : D], con D la matriz de combinaciones de niveles de los factores, denominada matriz de diseño. Si la matriz X es de rango completo, entonces el estimador de β obtenido por el método de mínimos cuadrados es que es, de hecho, el mejor estimador lineal de β) y la matriz de varianzas-covarianzas de b viene dada por

Este modelo se utiliza cuando queremos estudiar el comportamiento de la variable de respuesta únicamente en la región y cuando no conocemos la forma de la superficie. La forma de la función f que determina la relación entre los factores y la variable respuesta es, en general, desconocida, por lo que el

primer objetivo de la RSM consiste en establecer experimentalmente una aproximación apropiada de la función f. Para ello, se propone un modelo de ecuación, generalmente poli nómico, en los k factores X1, X2, ..., Xk y se selecciona un conjunto de tratamientos sobre los que realizar las observaciones experimentales, que se utilizarán tanto para obtener estimaciones de los coeficientes en el modelo propuesto (por ejemplo, a través del método de mínimos cuadrados) como para obtener una estimación de la variación del error experimental (para lo que es necesario tener al menos 2 observaciones por cada tratamiento). Se realizan, entonces, contrastes sobre las estimaciones de los parámetros y sobre el ajuste del modelo y si el modelo se considera adecuado, puede utilizarse como función.

II. REDES NEURONALES

Definición:

Un sistema de computación compuesto por un gran número de elementos simples, elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas.

Las Redes neuronales artificiales son redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico.

VENTAJAS QUE OFRECEN LAS REDES NEURONALES

Debido a su constitución y a sus fundamentos, las redes neuronales artificiales presentan un gran número de características semejantes a las del cerebro. Por ejemplo, son capaces de aprender de la experiencia, de generalizar de casos anteriores a nuevos casos, de abstraer características esenciales a partir de entradas que representan información irrelevante, etc. Esto hace que ofrezcan numerosas ventajas y que este tipo de tecnología se esté aplicando en múltiples áreas. Entre las ventajas se incluyen:

a) **Aprendizaje Adaptativo**. Capacidad de aprender a realizar tareas basadas en un entrenamiento o en una experiencia inicial.

b) **Auto-organización**. Una red neuronal puede crear su propia organización o representación de la información que recibe mediante una etapa de aprendizaje.

c) **Tolerancia a fallos**. La destrucción parcial de una red conduce a una degradación de su estructura; sin embargo, algunas capacidades de la red se pueden retener, incluso sufriendo un gran daño.

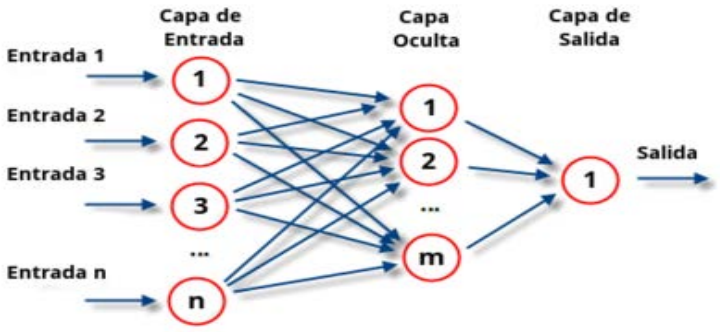
d) **Operación en tiempo real**. Los cómputos neuronales pueden ser realizados en paralelo; para esto se diseñan y fabrican máquinas con hardware especial para obtener esta capacidad.

e) **Fácil inserción dentro de la tecnología existente**. Se pueden obtener chips especializados para redes neuronales que mejoran su capacidad en ciertas tareas. Ello facilitará la integración modular en los sistemas existentes.

ELEMENTOS BASICOS:

Figura N° 3.4

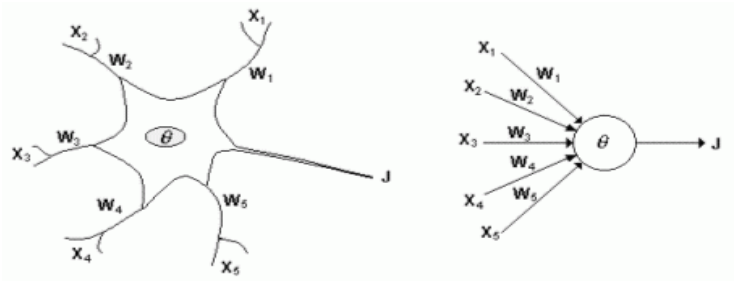
Elementos del proceso en el Modelo de Redes Neuronales



FUENTE: Elaboración propia.

La misma está constituida por neuronas interconectadas y arregladas en tres capas (esto último puede variar). Los datos ingresan por medio de la “capa de entrada”, pasan a través de la “capa oculta” y salen por la “capa de salida”. Cabe mencionar que la capa oculta puede estar constituida por varias capas.

Figura N° 3.5
Comparación entre una neurona biológica y artificial



FUENTE: Elaboración propia.

A. Función de entrada (input function):

La neurona trata a muchos valores de entrada como si fueran uno solo; esto recibe el nombre de entrada global. Por lo tanto, ahora nos enfrentamos al problema de cómo se pueden combinar estas simples entradas (ini1, ini2, ...) dentro de la entrada global, gini. Esto se logra a través de la función de entrada, la cual se calcula a partir del vector entrada. La función de entrada puede describirse como sigue:

inputi= (ini1•wi1)* (ini2•wi2)*... (inin•win)

Dónde:

*: Operador apropiado
n: número de entradas a la neurona
Ni: neurona.
Wi: peso.

Los valores de entrada se multiplican por los pesos anteriormente ingresados a la neurona. Por consiguiente, los pesos que generalmente no están restringidos cambian la medida de influencia que tienen los valores de entrada. Es decir, que permiten que un gran valor de entrada tenga solamente una pequeña influencia, si estos sonlo suficientemente pequeños. Algunas de las funciones de entrada más comúnmente utilizadas y conocidas son:

a) **Sumatoria de las entradas pesadas:** es la suma de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$\sum_{j=1}^{n}(n_{ij}w_{ij})$$

b) **Productoria de las entradas pesadas:** es el producto de todos los valores de entrada a la neurona, multiplicados por sus correspondientes pesos.

$$\prod_{j=1}^{n} n_{ij}w_{ij},$$

c) **Máximo de las entradas pesadas:** solamente toma en consideración el valor de entrada más fuerte, previamente multiplicado por su peso correspondiente.

$$\max(n_{ij}w_{ij})\ j=1,\ldots,\ n$$

Función de activación (activationfunction)

Una neurona biológica puede estar activa (excitada) o inactiva (no excitada); es decir, que tiene un estado de activación. Las neuronas artificiales también tienen diferentes estados de activación; algunas de ellas solamente dos, al igual que las biológicas, pero otras pueden tomar cualquier valor dentro de un conjunto determinado.

La función de activación calcula el estado de actividad de una neurona; transformando la entrada global (menos el umbral Θ_i) en un valor (estado) de activación, cuyo rango normalmente va de (0 a 1) o de (-1 a 1). Esto es asi, porque una neurona puede estar totalmente inactiva (0 o -1) o activa (1).

Las funciones de activación, es una función de la entrada global gin_i menos el umbral (Θ_i). Las funciones de activación más comúnmente utilizadas son: Función lineal, Función sigmoidea, Función tangente hiperbólica.

B. Función de salida (output function).

El último componente que una neurona necesita es la función de salida. El valor resultante de esta función es la salida de la neurona i(outi); por ende, la función de salida determina que valor se transfiere a las neuronas vinculadas. Si la función de activación está por debajo de un umbral determinado, ninguna salida se pasa a la neurona subsiguiente. Normalmente, no cualquier valor es permitido como una entrada para una neurona, por lo tanto, los valores de salida están comprendidos en el rango [0, 1] o [-1, 1]. También pueden ser binarios {0, 1} o {-1, 1}.

Dos de las funciones de salida más comunes son:

- Ninguna: este es el tipo de función más sencillo, tal que la salida es la misma que la entrada. Es también llamada función identidad.

- Binaria $\begin{cases} 0, & si\ \mathrm{act_i} \geq \varepsilon_i \\ 1, & de\ lo\ contrario \end{cases}$

  Donde ε_i es el umbral

C. MECANISMOS DE APRENDIZAJE

Se ha visto que los datos de entrada se procesan a través de la red neuronal con el propósito de lograr una salida. También se dijo que las redes neuronales extraen generalizaciones desde un conjunto determinado de ejemplos anteriores de tales problemas de decisión. Una red neuronal debe aprender a calcular la salida correcta para cada constelación (arreglo o vector) de entrada en el conjunto de ejemplos. Este proceso de aprendizaje se denomina: proceso de entrenamiento o acondicionamiento. El conjunto de datos (o conjunto de ejemplos) sobre el cual este proceso se basa es, por ende, llamado: conjunto de datos de entrenamiento.

Si la topología de la red y las diferentes funciones de cada neurona (entrada, activación y salida) no pueden cambiar durante el aprendizaje, mientras que los pesos sobre cada una

de las conexiones si pueden hacerlo; el aprendizaje de una red neuronal significa: adaptación de los pesos.

En otras palabras, el aprendizaje es el proceso por el cual una red neuronal modifica sus pesos en respuesta a una información de entrada. Los cambios que se producen durante el mismo se reducen a la destrucción, modificación y creación de conexiones entre las neuronas. En los sistemas biológicos existe una continua destrucción y creación de conexiones entre las neuronas. En los modelos de redes neuronales artificiales, la creación de una nueva conexión implica que el peso de la misma pasa a tener un valor distinto de cero. De la misma manera, una conexión se destruye cuando su peso pasa a ser cero.

Hay dos métodos de aprendizajes importantes que pueden distinguirse, Aprendizaje supervisado y Aprendizaje no supervisado.

Aprendizaje supervisado.

El aprendizaje supervisado se caracteriza porque el proceso de aprendizaje se realiza mediante un entrenamiento controlado por un agente externo (supervisor, maestro) que determina la respuesta que debería generar la red a partir de una entrada determinada. El supervisor controla la salida de la red y en caso de que ésta no coincida con la deseada, se procederá a modificar los pesos de las conexiones, con el fin de conseguir que la salida obtenida se aproxime a la deseada.

Redes Neuronales Artificiales en la detección de intrusos. Desde la propuesta realizada por Denning, del primer modelo de detección de intrusos, han aparecido un sinnúmero de IDS queaplican diferentes técnicas que han variado desde métodos basados en conocimientos, hasta métodos de la estadística clásica y para el aprendizaje automático.

Las ANN constituyen una de las técnicas que ha presentado amplias ventajas en su aplicación para la detección de intrusos. Las redes neuronales han demostrado ser potentes clasificadores con grandes capacidades de generalización y aprendizaje que presentan características que hacen muy

factible su aplicación en los IDS tal y como se expresa en la tabla 3.1.

TABLA 3.1

Características de las ANN	Ventajas en la detección de intrusiones
Forma de representación del conocimiento muy apropiada para problemas de clasificación.	En los IDS se trabaja precisamente con un problema de clasificación, posibilitando tomar ventajas de la facilidad de representación del conocimiento.
Alta tolerancia a errores, siendo capaces de clasificar datos que presentan ruidos o están incompletos.	Esta característica es de gran importancia sobre todo en los NIDS y NNIDS que analizan paquetes TCP/IP que pueden haber sufrido algunas modificaciones por problemas en la red.
Devuelven un valor numérico que da idea del nivel de seguridad de la clasificación realizada.	Da una idea más clara al administrador acerca de la decisión a tomar.
Pueden clasificar datos desconocidos.	Brinda la posibilidad de detectar ataques de los cuales aún no se tiene conocimiento.

Las ANN tratan de captar la esencia de procesos biológicos y aplicarla a nuevos modelos de computación. La Neurona Artificial es un modelo simplificado de neurona biológica. La interconexión entre neuronas decide el flujo de información en la red y, junto con los pesos y las funciones de salida de cada neurona definen el comportamiento global de la red neuronal artificial. Esto hace que las redes neuronales estén formadas por un gran número de elementos de cómputo lineales y no lineales (neuronas) complejamente interrelacionados y organizados en capas.

3.2.2 TEORIA DE TECNICA APLICADA

Marco para Detectar el fraude financiero adaptativo

Similar al procedimiento de clasificación tradicional, en general se consideran dos etapas en nuestro marco (Graf. 2). En la primera etapa, las variables externas e internas relevantes que diferencian a las industrias, las condiciones económicas, la elección de la administración, consideraciones de tiempo y cualesquiera otros factores que tienen el potencial para formar el conocimiento del dominio se seleccionan y experimentaron. En la segunda etapa, los datos financieros de la empresa en cuestión se analizan basados en este conocimiento del dominio aprendió de la etapa anterior.

Cierta estrategia de detección se forma en consecuencia y los datos se analizan adicionalmente usando técnicas de minería de datos. Para hacer realidad el modelo general descrito anteriormente y para coger la dinámica del mundo real y los posibles nuevos métodos para cometer fraude de estados financieros, se propone un marco de aprendizaje adaptativo (Fig. 3).

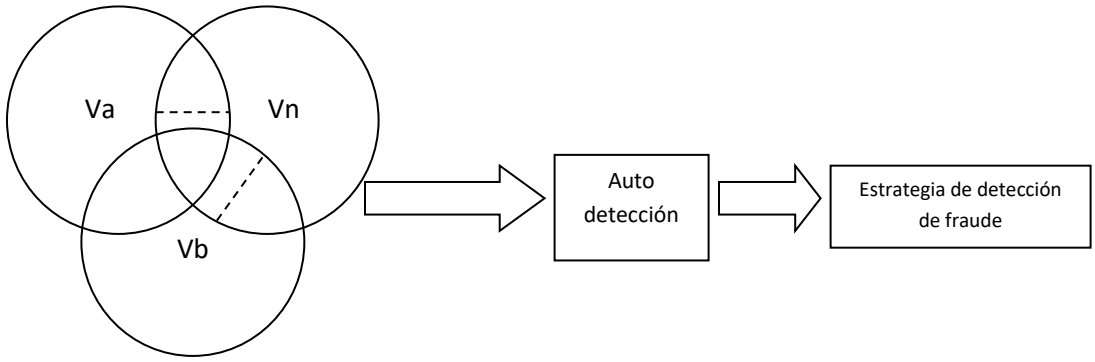
Consideramos, pero no se limitan a, los parámetros exógenos descritos en la literatura existente, como la estructura de capital, las condiciones, las opciones, las actitudes de gestión, etc.

El mecanismo funciona de la siguiente manera: en base a las circunstancias económicas externas e internas, la administración selecciona su acción sobre la conveniencia o no de cometer fraude en los estados financieros al cierre del ejercicio. Datos y declaraciones de preocupación financieros son auditados y examinados por una unidad de detección de fraudes. Resultantes informes de auditoría se evalúan más y aprendidas por un módulo de auto-adaptable para recoger los patrones y tendencias de cada empresa en diferentes industrias relevantes. Mientras tanto, un módulo de detección de fraude adaptativo sigue evolucionando con parámetros exógenos para descubrir patrón desconocido, pero posible de fraude en los estados financieros. Los nuevos descubrimientos también fueron evaluados y aprendieron a preparar la base de conocimientos para la futura detección de fraudes.

Para aumentar la relevancia de la detección y reducir la complejidad computacional, también proponemos la selección de características de adaptación que se adapta dominio específico para elegir los parámetros adecuados para las empresas con el medio ambiente financieras internas y externas similares. Una vez que se seleccionan los parámetros relevantes, optamos por una metodología adecuada y la técnica de minería de datos para detectar el fraude financiero que evoluciona. Como se discutió en la sección anterior, ninguna técnica de detección FSF basada en la minería de datos único es perfecta y cada uno de ellos está sujeto a sus propias desventajas.

Proponemos metodología de superficie de respuesta para la construcción de la base con el fin de encontrar la técnica de detección basada en la minería de datos correcta.

Figura N° 3.6
Aumento de relevancia de la detección y reducción de complejidad en clasificación de fraude



FUENTE: Detecting evolutionary financial statement fraud. Wei Zhou

DETECCIÓN DE FRAUDE FINANCIERO ADAPTATIVA CON RSM

La metodología de superficie de respuesta, es un método para construir aproximaciones globales al comportamiento del sistema basado en resultados calculados en varios puntos en el espacio de diseño, mediante un ajuste natural se puede estimar las relaciones entre las variables y las técnicas de fraude en los estados financieros. RSM proporciona estadísticamente validados modelos predictivos que pueden ser manipulados para encontrar la probabilidad de diferentes formas de posibles fraudes de los estados financieros.

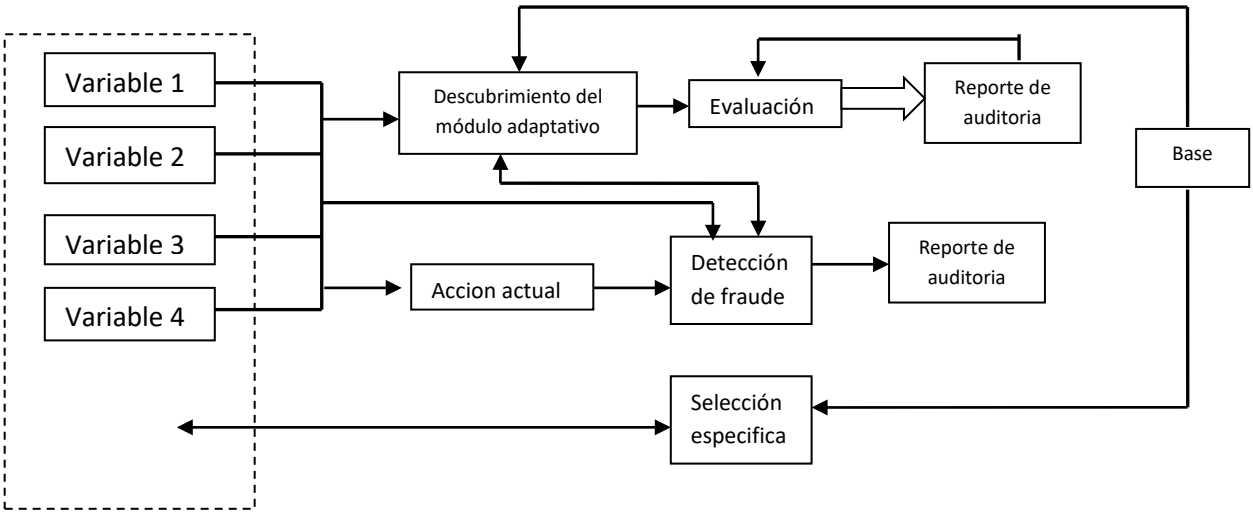
RSM se aplica ampliamente en diversas situaciones en las que el rendimiento de salida o un servicio de calidad, que se llama respuesta, está influenciada por una lista de varias variables de entrada, que pueden o no estar completa y reciben el nombre de variables independientes. Las variables independientes están sujetos al control del diseñador experimento y la aproximación de la relación entre la variable respuesta y las variables independientes se puede visualizar por RSM, que consta de tres factores.

La estrategia experimental para explorar el espacio de las variables independientesmodelización estadística empírica para desarrollar una relación que sea aproxima y adecuada entre la variable respuesta y las variables independientes.

El método de optimización para encontrar los valores de las variables de proceso que producen valores deseables de respuesta.

Figura N° 3.7

Nueva metodología con superficie de respuesta



FUENTE: Detecting evolutionary financial statement fraud. Wei Zhou

Graf 7

En teoría, el modelo es adecuado entre la variable respuesta Y y las variables independientes x1, x2, x3,...,xnque puede ser construido como Y = f (x1, x2,....., xn) + e, donde la forma de la cierto tiempo real función de respuesta f es desconocida y puede ser complejo. ε, que por lo general incluye el error de medición de la respuesta y otro ruido imprevisible, es un término aleatorio que representa la variabilidad no ha caído en la función de respuesta. Si suponemos que tiene una distribución normal con media cero, Y = E (y) = f (x1, x2, ..., xn). Con la forma de la verdadera función de

respuesta f mantuvo desconocido, el diseñador tiene que aproximarse a ella y además utilizar para localizar la posible respuesta con variables independientes descubiertos.

La selección de características se puede implementar para identificar una lista de variables relevantes que el diseñador puede utilizar más para construir la superficie de respuesta. Una vez que tenemos una empresa de que se trate, que son capaces de encontrar la posibilidad de que ciertos fraudes de los estados financieros en base a la estimación de superficie de respuesta. A continuación, seleccionamos las técnicas de minería de datos que se adapten al perfil que hemos aprendido de la etapa anterior.

RSM proporciona herramientas estadísticas para el análisis de los datos históricos y la selección de las variables dirigidas a una mejor predicción. Los objetivos para el uso de RSM en el contexto de la detección de fraude en los estados financieros son para encontrar la respuesta óptima y para entender cómo los cambios de respuesta en una dirección dada por el ajuste de las variables de diseño [17]. Cuando hay limitaciones en los datos de diseño, entonces la variable de selección y el diseño experimental tiene que cumplir con los requisitos de las restricciones. En general, la superficie de respuesta se puede visualizar gráficamente y el gráfico, si en menos de tres dimensiones, es útil para navegar a través de la superficie de respuesta para alcanzar el resultado deseado.

En una forma sencilla, una función f (x1, x2) se representó frente a los niveles de x1 y x2, y este gráfico tridimensional se forma una parcela de superficie de respuesta Graf. 7 Esta figura muestra un marco de aprendizaje adaptativo para detectar fraudes de los estados financieros evolutivos utilizando un método de superficie de respuesta Graf. 8 Se procede primeras variables relevantes de selección que podría ser o bien los tres parámetros en el modelo CMA, los parámetros en el modelo de un 3C [19], una mezcla de ambos modelos, o cualquier otro parámetro que tienen alguna relación causal con el estado financiero fraudes. Las variables seleccionadas se indican cómo$\boldsymbol{V_a; V_b; V_c; \ldots; V_n}$ respectivamente, por lo que la probabilidad de cometer fraude en los estados financieros en forma de k puede ser descrito como la ecuación (2) que está sujeto a las ecuaciones (3) y 4 (4). Suponemos que la primera n formas de fraude han sido descubiertos y que el número de formas posibles es infinito, de tal manera que:

$$\boldsymbol{P(FSF_k) = f(V_a; V_b; V_c; \ldots; V_n)}$$

Que está sujeto a:

$$\sum_{k=1}^{n} P(FSF_k) < 1$$

3.2.3 TERMINOLOGIA BÁSICA

a) **IDS**

IntrusionDetectionSystem, es un programa de detección de accesos no autorizados a un computador o a una red.

El IDS suele tener sensores virtuales (por ejemplo, un sniffer de red) con los que el núcleo del IDS puede obtener datos externos (generalmente sobre el tráfico de red). El IDS detecta, gracias a dichos sensores, las anomalías que pueden ser indicio de la presencia de ataques y falsas alarmas.

b) FAMILIA DE PROTOCOLOS TCP/IP

La familia de protocolos de Internet es un conjunto de protocolos de red en los que se basa Internet y que permiten la transmisión de datos entre computadoras.

En ocasiones se le denomina conjunto de protocolos TCP/IP, en referencia a los dos protocolos más importantes que la componen, que fueron de los primeros en definirse, y que son los dos más utilizados de la familia:TCP (Transmission Control Protocol), Protocolo de Control de Transmisión, eIP (Internet Protocol), Protocolo de Internet.

C) **INTRUSIÓN**

Se puede entender por intrusión a una violación de la política de seguridad del sistema.Intrusiones se pueden producir de varias formas: atacantes que acceden a los sistemas desde Internet, usuarios autorizados del sistema que intentan ganar privilegios adicionales para los cuales no están autorizados y usuarios autorizados que hacen un mal uso de los privilegios que se les han asignado.

D) MINERÍA DE DATOS

La minería de datos es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiado dato.

E) FACTORES

Son las condiciones del proceso que influencian la variable de respuesta. Estos pueden ser cuantitativos ocualitativos.

F) RESPUESTA

Es una cantidad medible cuyo valor se ve afectado al cambiar los niveles de los factores. El interés principal es optimizar dicho valor.

G) FUNCIÓN DE RESPUESTA

Al decir que un valor de respuesta Ydepende de los niveles x1, x2, ...xkde kfactores, ξ1, ξ2,... ξk, estamos diciendo que existe una función matemática de x1, x2, ... xkcuyo valor para una combinación dada de los niveles de los factores corresponde a Y, esto es Y=f(x1, x2, ...xk).

H) PROTOCOLO DE COMUNICACIONES

Los protocolosson un conjunto de pautas que posibilitan que distintos elementos que forman parte de un sistema establezcan comunicaciones entre sí, intercambiando información.

Los protocolos de comunicación instituyen los parámetros que determinan cuál es la semántica y cuál es la sintaxis que deben emplearse en el proceso comunicativo en cuestión. Las reglas fijadas por el protocolo también permiten recuperar los eventuales datos que se pierdan en el intercambio.

Si nos centramos en las computadoras, el protocolo de comunicación determina cómo deben circular los mensajes dentro de una red. Cuando la circulación de la información se desarrolla en Internet, existen una serie de protocolos específicos que posibilitan el intercambio.

Los protocolos de comunicación en Internet más importantes son TCP (cuyas siglas pueden traducirse como Protocolo de Control de Transmisión) e IP (Protocolo de Internet). Su acción conjunta (TCP/IP) posibilita el enlace entre todos los equipos que acceden a la red.

POP, SMTP y HTTP son otros protocolos vinculados a Internet, que los usuarios suelen utilizar a diario aunque no lo adviertan ni sepan cómo funcionan. Estos protocolos permiten navegar a través de los sitios web, enviar correo electrónico, escuchar música online, etc.

i) TCP

El Protocolo de Control de Transmisión es, como se explica anteriormente, uno de los elementos básicos de Internet. Su creación data del periodo comprendido entre los años 1973 y 1974 y se adjudica al ingeniero Vinton Gray Cerf y al investigador Robert ElliotKahn.

Entre las utilidades de este protocolo de comunicación se encuentra la creación de conexiones entre diversos programas presentes en una red de datos para llevar a cabo un flujo de información. Gracias a su aplicación en un caso tal, queda garantizado que los datos lleguen a destino sin errores y ordenados de la misma forma en la cual se hallaban antes de ser enviados. Además, el TCP ofrece la posibilidad de reconocer cada aplicación del resto, gracias al uso de los puertos.

Cuando se realiza una comunicación a través de Internet, por ejemplo, el router simplemente debe ocuparse del envío de datos pero no de realizar un monitoreo de los mismos, dado que de esto se encarga el TCP, que también se conoce con el nombre de capa de transporte, entre la aplicación y el protocolo de Internet (IP).

j) IP

Este protocolo de comunicación es mucho más conocido por los usuarios de Internet, aunque sólo los expertos sepan en profundidad de qué se trata realmente. La función del IP, que se encuentra en la capa de red, es permitir la comunicación en dos direcciones, en destino u origen, para que sea posible la transmisión de datos a través de un protocolo no orientado a conexión que envía paquetes conmutados por medio de diferentes redes físicas que han sido enlazadas con anterioridad siguiendo la norma OSI.

Capítulo IV

4. METODOLOGIA

4.1 TIPO, NIVEL Y DISEÑO DE INVESTIGACIÓN

4.11 Tipo de investigación:

El tipo de investigación desarrollado es aplicativa y cuantitativa ya que se usan datos simulados en contextos reales y se aplica una nueva metodología para una mejor clasificación.

4.12 Nivel de investigación:

El nivel aplicado será exploratoria, descriptiva y predictiva, ya que se implementa una nueva metodología, se hace un análisis descriptivo de las variables y mediante el aprendizaje de ANN se clasificará las conexiones normales y malignas.

4.2 DISEÑO MUESTRAL

4.2.1 POBLACION EN ESTUDIO

Este conjunto de datos fue simulado en un típico entorno militar de las fuerzas áreas de Estados Unidos-USAF en el año 1999. Se generó con el propósito de ser utilizado en el tercer Concurso Internacional de descubrimiento de conocimiento y herramientas de minería de datos, que se realizó en conjunto con KDD-99 La Quinta Conferencia

Internacional sobre el descubrimiento de conocimiento y minería de datos.

La tarea de la competencia era construir un detector de intrusión en la red, un modelo predictivo capaz de distinguir entre malas conexiones, llamadas intrusiones o ataques y conexiones normales.

4.2.2 FUENTE DE INFORMACION

Los Laboratorios Lincoln crearon un entorno para adquirir un volcado de datos TCP durante nueve semanas, en una red de área local (LAN) que simulaba la típica red de las Fuerzas Aéreas de EE.UU salpicada con múltiples ataques. El conjunto bruto de datos de entrenamiento, obtenidos durante las primeras 7 semanas, ocupaban cerca de cuatro gigabytes, lo que equivale aproximadamente a cinco millones de registros de conexión.

4.2.3 DEFINICION DE VARIABLES

La base de datos contiene 41 variables independientes, con los cuales se describen diferentes características de cada conexión y 24 tipos de ataque, con 14 tipos adicionales en los datos de test que son tipos de ataque no identificados.
Para identificar la variable dependiente se clasificó los tipos de ataque como: normales y tipo de ataque detectado.

El número de registros que se tiene en la base de datos es de 311 029.

Tipos de ataque:

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
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.	R2L
loadmodule	Ataque furtivo que reinicia la IFS para un usuario normal y crea un shell de root.	R2L
perl	Establece el id de usuario como root en un script de perl y crea un shell de root.	R2L
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 de 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

4.3 PROCEDIMIENTO

4.3.1 Análisis o entendimiento del negocio.

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.

4.3.2 Análisis o entendimiento de los datos.

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: interrupción, el recurso se vuelve no disponible, intercepción, “alguien” no autorizado consigue acceso a un recurso y 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.

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.

4.3.3 Preparación de los datos.

Tenemos a disposición los siguientes conjuntos de datos:

- kddcup.data.gz: Datos de entrenamiento originales (743 MBdescomprimido).
- kddcup.data_10_percent.gz: Subconjunto del 10% de los datos de entrenamiento (75 MB descomprimido).

Se debe tener en cuenta que el software R a pesar de manejar grandes conjuntos de datos es necesario partir la data en dos muestras de entrenamiento y test.

De esta forma se dispone de un conjunto de 494021 instancias de datos de entrenamiento y 311029 instancias de datos de test corregidos, que siguen siendo demasiado grandes como para poder trabajar con ellos.

El conjunto que más limita es el de datos de entrenamiento, pues con él se tiene que construir el modelo de clasificación, lo que consume mucha memoria. El proceso de clasificación de los datos de test suele ser mucho más rápido.
Probando con diferentes tamaños de conjuntos de datos se han elegido finalmente un subconjunto de 20585 instancias de datos de entrenamiento y 10034 de test, pues con estos conjuntos se pueden utilizar casi todos los algoritmos de minería de datos.
La elección de estos sub-subconjuntos se ha hecho muestreando los datos (cogiendo por ejemplo uno de cada 24 en el caso de los de entrenamiento), para tener unas muestras heterogéneas.

4.3.4 Modelado.

Para el modelado se pretende usar el software R- studio por ser gratuito y altamente potente en uso de grandes volúmenes de datos y SAS por su aplicabilidad en el diseño de experimentos.

Redes neuronales-Superficie de respuesta

- Encontrar las probabilidades de intrusión
- Maximizar la probabilidad a través del método ascendiente
- Encontrar las variables que maximizan la probabilidad
- Construir el modelo de redes neuronales con las variables seleccionadas.
- Evaluar el modelo en los datos de entrenamiento

Redes neuronales

- Encontrar las probabilidades de intrusión
- Maximizar la probabilidad a través del método ascendiente
- Encontrar las variables que maximizan la probabilidad
- Construir el modelo de redes neuronales con las variables seleccionadas.
- Evaluar el modelo en los datos de entrenamiento

Gantt

Fecha	Actividades	Agosto			Septiembre				Octubre				Noviembre			
		3	4	5	1	2	3	4	1	2	3	4	1	2	3	4
16-ago	Definir el Tema de investigación															
22-ago	Análisis crítico de tesis relacionadas al tema de investigación.															
29-ago	Búsqueda de antecedentes nacionales e internacionales															
05-sep	Situación problemática, Delimitación del problema y Formulación del problema.															
12-sep	Objetivos, Hipótesis y Justificación. 4º Trabajo: Descripción del problema, formulación del problema, Objetivos, Hipótesis y Justificación.															
19-sep	Marco teórico															
26-sep	Definición de variables, operacionalización de variables, matriz de consistencia y metodología															
10-oct	Revisión de Metodología															
18-oct	Revisión general de tesis															
24-oct	Análisis descriptivo de las variables															
25-oct	Resultados de la técnica a usar															
31-oct	Revisión de resultados															
08-nov	Recomendaciones para mejorar el Título, Problema de Investigación, Objetivos, Hipótesis, Conclusiones y Recomendaciones.															
21-nov	Revisión de Bibliografía															

Costo y Presupuesto:

	Descripción	Monto
Bienes	Materiales de escritorio: Lapicero, folder, engrampadoras etc	15 Soles
	consumo de energía en Computador PC, Laptop	10 Soles
Servicio	Fotocopia de material bibliográfico	20 Soles
	adquisición de material académico	50 Soles
	impresión	13 Soles
	movilidad	9 Soles
	Consumo de alimentos	10 Soles
total		**127 Soles**

Capítulo V

5. Resultados

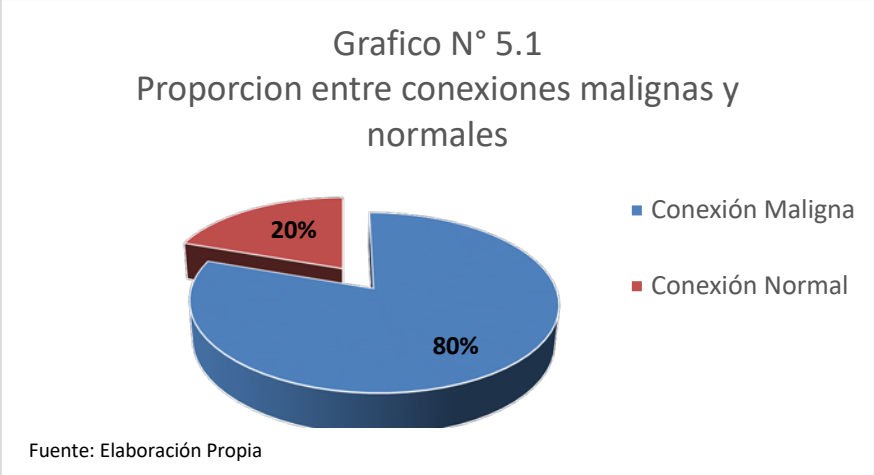
5.1 Análisis descriptivo

5.1.1 Análisis Univariado

Examinando la variable dependiente conexión, se puede identificar dos tipos de conexiones maligna y normal, que esta desbalanceado debido a la proporción 80/20 que posee (Grafico N° 5.1)

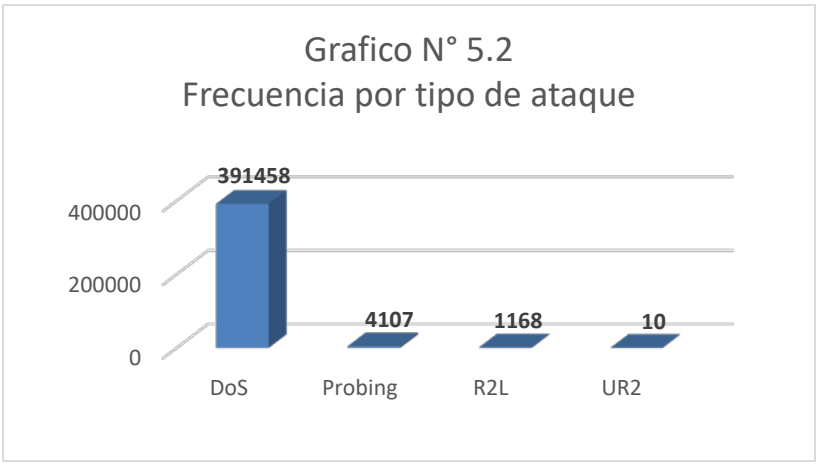
Tabla 5.1

Tipo de Conexión	%
Conexión maligna	80.31%
Conexión normal	19.69%
Total general	**100.00%**

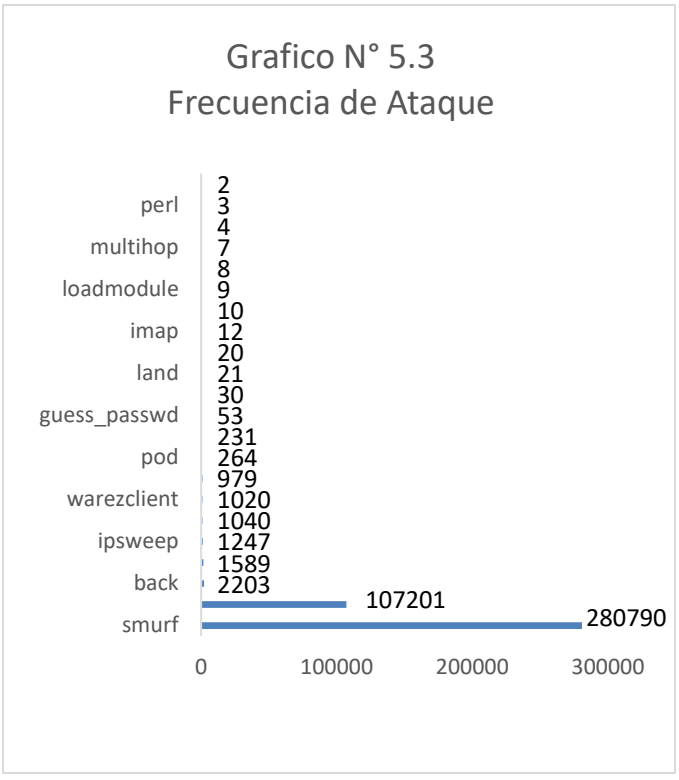


Fuente: Elaboración Propia

Por otro lado, los tenemos 4 tipos de ataque que producen conexiones malignas, siendo las más frecuente el tipo DoS cuyo 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. (Grafico N° 5.2), seguida del probing y por ultimo las menos frecuentes que son los ataques R2L y UR2. Dentro de los tipos de ataque más frecuentes tenemos los smurf (280790) y neptune (107201) ver gráfico N° 5.3



Fuente: Elaboración Propia

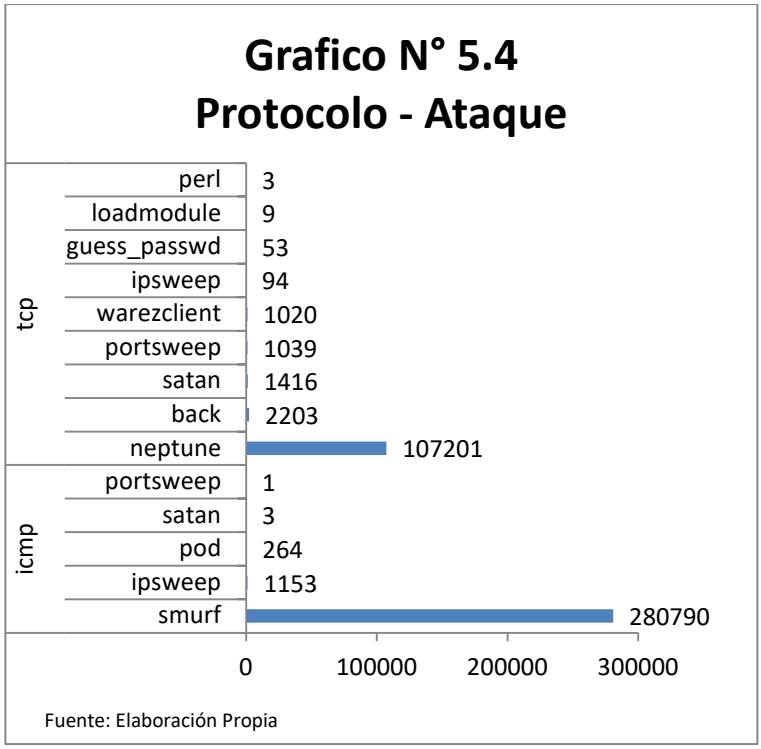


Fuente: Elaboración Propia

Tabla 5.2

Ataque	Frecuencia
smurf	280790
neptune	107201
back	2203
satan	1589
ipsweep	1247
portsweep	1040
warezclient	1020
teardrop	979
pod	264
nmap	231
guess_passwd	53
buffer_overflow	30
land	21
warezmaster	20
imap	12
rootkit	10
loadmodule	9
ftp_write	8
multihop	7
phf	4
perl	3
spy	2
Total general	**396743**

En el grafico N° 5.4 se puede ver que existe una posible relación entre la variable del tipo de ataque con la del tipo de protocolo que se use. Por lo tanto se incluirá esta variable para aumentar la precisión del modelo predictivo.



5.1.2 Análisis de significancia:

Para mejorar el análisis entre las variables se verificará si existe una relación entre la variable dependiente e independientes.

Se realizará una diferencia de medias, usando la prueba T-studenty la prueba de Levene para la igualdad de varianzas para evaluar sila variable dicotómica y continua proviene de muestras independientes.

Solo se tomaron en cuenta las variables continuas que son normales. Para aquellas que no lo son se realizará una prueba no paramétrica para la diferencia de medianas.

TABLA 5.3 **Prueba de muestras independientes**

	Prueba de Levene para la igualdad de varianzas		Prueba T para la igualdad de medias		
	F	Sig.	t	gl	Sig. (bilateral)
src_bytes	7,261	,007	2,866	197009	,004
			2,887	148530,770	,004
dst_bytes	882,549	,000	-22,957	197009	,000
			-22,830	161707,806	,000
count	256479,862	,000	801,390	197009	,000
			811,182	102362,114	,000
srv_count	397034,598	,000	482,503	197009	,000
			488,441	101633,915	,000
dst_host_same_src_port_rate	57466,167	,000	344,060	197009	,000
			345,947	167895,996	,000
dst_host_srv_diff_host_rate	16417,095	,000	-109,393	197009	,000
			-109,053	183259,141	,000
dst_host_count	517844,133	,000	313,004	197009	,000
			309,429	104986,378	,000

Como se observa en la Tabla 5.3 la Hipótesis nula que afirma que las medias en las dos poblaciones son iguales, es rechazada y por tanto podemos concluir que las medias pertenecen a diferentes poblaciones.

TABLA 5.4 **Correlaciones**

	Y	src_bytes	dst_bytes	count	srv_count	dst_host_count	dst_host_same_src_port_rate	dst_host_srv_diff_host_rate
Y	1	-,006**	,052**	-,875**	-,736**	-,576**	-,613**	,239**
src_bytes	-,006**	1	,000	-,012**	-,009**	-,023**	,011**	,016**
dst_bytes	,052**	,000	1	-,050**	-,043**	-,045**	-,040**	,015**
count	-,875**	-,012**	-,050**	1	,956**	,529**	,828**	-,259**
srv_count	-,736**	-,009**	-,043**	,956**	1	,446**	,891**	-,219**
dst_host_count	-,576**	-,023**	-,045**	,529**	,446**	1	,322**	-,513**
dst_host_same_src_port_rate	-,613**	,011**	-,040**	,828**	,891**	,322**	1	-,101**
dst_host_srv_diff_host_rate	,239**	,016**	,015**	-,259**	-,219**	-,513**	-,101**	1

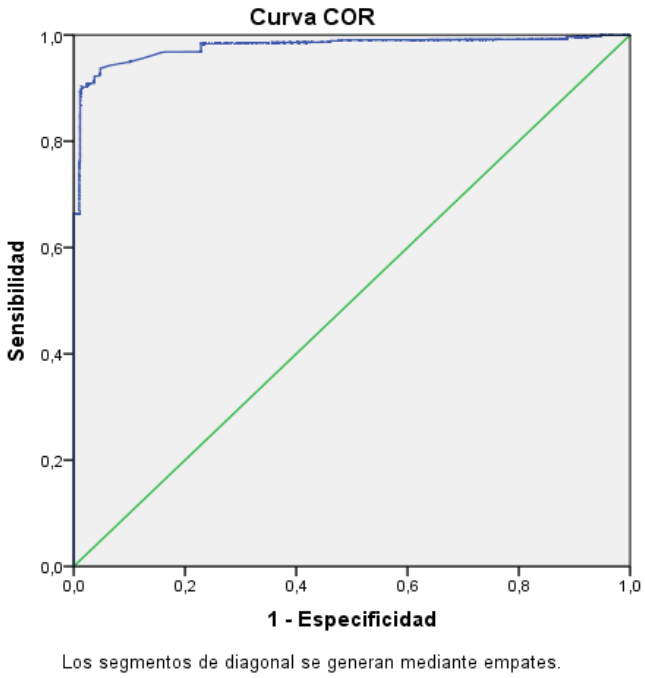
Se realizó un análisis de correlación y se encontró que la variable dependiente esta correlacionada a las variables count,srv_count, dst_host_count, dst_host_same_src_port_rate y dst_host_srv_diff_host_rate.

5.2 Análisis de Redes Neuronales

Debido a que existe correlación entre variables, reduciremos su cantidad usando el IV que comparará el poder predictivo de cada variable.

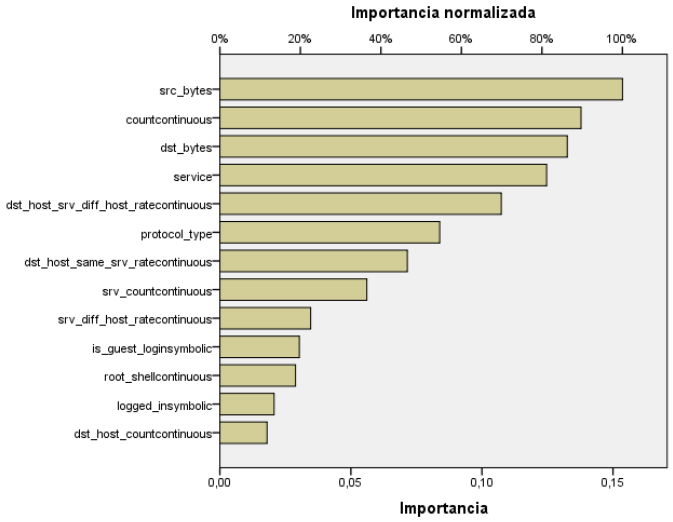
Variable	Information Value
Count	**8.31%**
srv_count	**7.19%**
dst_bytes	**5.75%**
src_bytes	**5.41%**
service_ecr	**4.57%**
dst_host_same	**4.37%**
src_port_rate	**4.37%**
logged_in	**4.08%**
protocol type icm	**3.62%**
service_http	**3.56%**
dst_host_count	**2.90%**
dst_host_srv_diff	**2.66%**
f_host_rate	**2.66%**
flag_S0	**1.39%**
protocol_type_tcp	**1.12%**
flag_SF	0.41%
service_priv	0.28%

5.2.1 Curva ROC



Para un numero de 15 variables, el área bajo la curva es AUC=0.977, siendo el coeficiente de Gini=0.95

5.2.2 Importancia de variables



Acerca de la bondad del modelo

Prueba onmibus

Pruebas omnibus sobre los coeficientes del modelo

		Chi cuadrado	gl	Sig.
Paso 5	Paso	.448	1	.503
	Bloque	260635.649	75	0.000
	Modelo	260635.649	75	0.000

Resumen del modelo

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
5	12448,996[a]	.734	.978

a. La estimación ha finalizado en el número de iteración 20 porque se han alcanzado las iteraciones máximas. No se puede encontrar una solución definitiva.

R cuadrado de cox y snell indica que el 73% de la variable clasificación es explicada por las variables seleccionadas.

5.3 Análisis de Superficie de Respuesta

5.3.1 Modelo de segundo orden

Análisis de Varianza

Fuente	GL	SC Ajust.	MC Ajust.	Valor F	Valor p
Modelo	10	2852098613	285209861	0.82	0.634
Lineal	4	1248790678	312197670	0.89	0.530
protocolo	1	154008100	154008100	8.44	0.002
flag	1	148693636	148693636	0.43	0.010
loged	1	475697910	475697910	14.36	0.001
rango_count	1	470391032	470391032	9.35	0.050
Interacción de 2 factores	6	1603307935	267217989	0.77	0.628
protocolo*flag	1	517403262	517403262	1.48	0.278
protocolo*loged	1	142468096	142468096	13.41	0.001
protocolo*rango_count	1	162078361	162078361	0.46	0.526
flag*loged	1	133310116	133310116	0.38	0.002
flag*rango_count	1	164685889	164685889	0.47	0.523
loged*rango_count	1	483362210	483362210	1.38	0.292
Error	5	1746202733	349240547		
Total	15	4598301346			

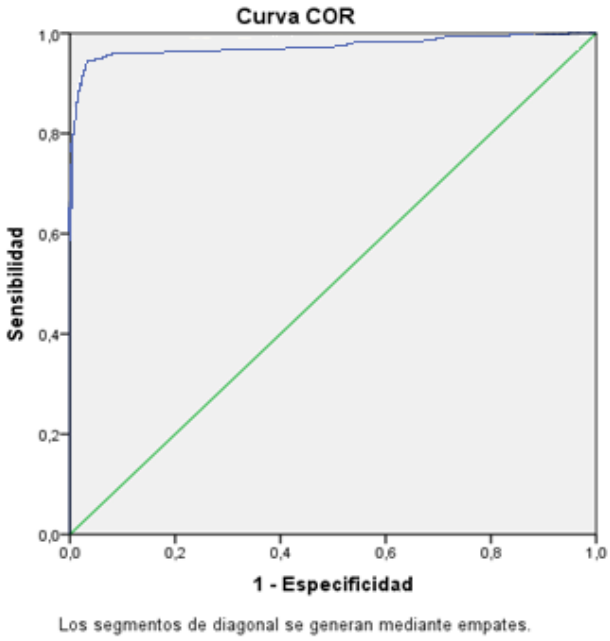
R-cuadrado ajustado.
62.03%

5.3.2 Modelo de primer orden

Término	Efecto	Coef	EE del coef.	Valor T	Valor p
Constante		5940	4362	1.36	0.201
protocolo	-6205	-3103	4362	10.71	0.002
flag	-6097	-3049	4362	9.70	0.005
loged	-10905	-5453	4362	11.25	0.017
rango_count	10844	5422	4362	11.24	0.015

R-cuadrado ajustado.
60.16%

Seleccionamos el modelo de segundo orden ya que se obtiene un mayor indicador.



6. Conclusiones

La comparación de los modelos de redes Neuronales, usando y no usando la superficie de respuesta previamente no fue clara y contundente; al determinar que el modelo de Redes Neuronales sin el uso de superficie de respuesta obtuvo un Gini=95% mientras que el modelo que usó las variables protocolo, flag, loged, rango_count como parte de optimizar el la conexión maligna obtuvo un gini=98%.

Se determinó que el modelo de segundo orden es el más adecuado, ya que se comparó el R- ajustado en ambos modelos y obtuvo un 62,03% de predicción de la variable conexión maligna, que es justamente la variable que queremos maximizar.

Se obtuvo que las variables protocolo, flag, loged y rango_count son las variables que optimizan las conexiones malignas en el modelo de primer orden, posteriormente con el metodo de ascenso mas pronunciado realizaremos el calculo del modelo de segundo orden en el cual se

obtuvieron como variables significativas a protocolo, flag, loged,rango_count , protocolo*flag, protocolo*loged, loged*rango_count y flag_loged

7. Recomendaciones:

El modelo de Superficie de Respuesta no trajo consigo un aumento significativo del coeficiente de Gini por lo tanto, para el esfuerzo hecho no es recomendable usar este método ya que también es difícil su implementación para predecir en tiempo real si un usuario es intruso.

Se recomienda también usar métodos para reducir variables, ya que muchas de ellas están relacionadas y crean modelos con altos coeficientes de Gini.

El método utilizado podría ayudar a otro tipo de investigaciones donde el aporte del investigador sea de vital importancia.

Bibliografía

[1] Damián Jorge Matich. (2001). Redes Neuronales: Conceptos Básicos y Aplicaciones. Dirección URL <https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf>

[2] Enrique López González. (2004). DETECCION DE INTRUSOS EN AUDITORIA INFORMATICA UTILIZANDO SISTEMAS BASADOS EN REGLAS DIFUSAS Y PROCESOS DE APRENDIZAJE CON ALGORITMOS GENETICOS. Dirección URL <http://sicodinet.unileon.es/dpi2001-0105/>

[3] Shwarz, John. La importancia de la seguridad en materia empresarial.Dirección URL http://www.symantec.com.

[4] Concepto de Mineria de datos. Microsoft Direccion URL <https://msdn.microsoft.com/es-es/library/ms174949.aspx>

[5] Protocolo de comunicación. JulianPerez Porto y Ana Gardey . Publicado en 2015. Dirección URL <http://definicion.de/protocolo-de-comunicacion/>

[6] Diseño y analisis de experimentos. Montgomery. 2da Edicion. DireccionUrl<https://wwwyyy.files.wordpress.com/2013/02/disec3b1o-de-experimentosmontgomery.pdf>

[7] Machine Learning Repository KDD Cup 1999. DireccionUrl<https://archive.ics.uci.edu/ml/datasets/KDD+Cup+1999+Data>

[8] Machine Learning Repositor. DireccionUrlhttps://archive.ics.uci.edu/ml/machine-learning-databases/kddcup99-mld/

i