Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Académico Profesional de Ingeniería de Sistemas

Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias

TESINA

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Carol Maribel ÑAUPAS CARAZA

ASESOR

Jorge Luis ZAVALETA CAMPOS

Lima, Perú

2016

Referencia bibliográfica

Ñaupas, C. (2016). *Minería de datos aplicada a la detección de fraude electrónico en entidades bancarias*. [Tesina de pregrado, Universidad Nacional Mayor de San Marcos Facultad de Ingeniería de Sistemas e Informática, Escuela Académico Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMATICA
PROGRAMA DE ACTUALIZACIÓN PROFESIONAL 2014-II

Acta de Sustentación de Tesina

Siendo las 18:50hrs Del día..27...de Mayo del año 2016, se reunieron los docentes designados como miembros de Jurado de la Tesina, presidido por el Mg. Frank Edmundo, Escobedo Bailón, el Msc. Juan, Gamarra Moreno (Miembro) y la Ing. María Rosa, Dámaso Ríos (Miembro) para la sustentación de la Tesina intitulada: "MINERÍA DE DATOS APLICADA A LA DETECCIÓN DE FRAUDE ELECTRÓNICO EN ENTIDADES BANCARIAS". Por la Srta. Bach, CAROL MARIBEL ÑAUPAS CARAZA; para optar el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesina, el Presidente invitó al graduando a dar respuesta a las preguntas establecidas por los Miembros de Jurado.

El graduando en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros de Jurado, el graduando obtuvo la nota de....16..... (En letras)..dieciseis

A continuación el Presidente del Jurado el Dr. Frank Edmundo, Escobedo Bailón, declara al graduando **Ingeniero de Sistemas.**

Siendo las 1930. horas, se levantó la sesión. Asesoramiento

.....
Presidente
Dr. Frank Edmundo Escobedo Bailón

.....
Miembro
Msc. Juan, Gamarra Moreno

.....
Miembro
Ing. María Rosa, Dámaso Ríos

DEDICATORIA:

Para mis padres y familia por el apoyo incondicional que siempre me han dado.

AGRADECIMIENTOS

- A mi familia y amigos que he conocido a lo largo de mi vida personal y profesional y han dejado gratos recuerdos y enseñanzas en mi.
- A todas las personas que me apoyaron en desarrollo del presente proyecto.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

MINERÍA DE DATOS APLICADA A LA DETECCIÓN DE FRAUDE ELECTRÓNICO EN ENTIDADES BANCARIAS

Autor: ÑAUPAS CARAZA, Carol Maribel

Asesor: ZAVALETA CAMPOS, Jorge Luis

Grado: Ingeniero de Sistemas e Informática

Titulo: Tesina, para optar el Título Profesional de Ingeniero de Sistemas

Fecha: Enero 2016

RESUMEN

El fraude es unos de los principales riesgos a los que se enfrentan las entidades bancarias y los canales electrónicos son el principal blanco a los ataques de los defraudadores, es por ello que se deben adoptar diversas medidas para prevenirlo y tomar decisiones en tiempo real.

Debido a que los avances tecnológicos están permitiendo a las empresas gestionar grandes volúmenes de datos, resulta de gran valor analizarlos, producir información y descubrir conocimientos a partir de ellos, para que sean utilizados estratégicamente en la toma de decisiones.

El presente trabajo propone a través de la aplicación de un proceso de descubrimiento de conocimientos en bases de datos, la generación de un modelo automático que permite clasificar las transacciones de la Banca por internet y de la Banca Móvil de personas naturales de una entidad financiera, como fraudulentas o íntegras, mediante la aplicación de técnicas de Minería Predictiva basada en arboles de Clasificación.

Palabras Clave: Descubrimiento de Conocimiento en Bases de Datos, Minería de Datos Predictiva, Arboles de Clasificación.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMÁTICA

ESCUELA ACADÉMICO PROFESIONAL DE INGENIERÍA DE SISTEMAS

DATA MINING APPLIED TO THE DETECTION FRAUDE ON ELECTRONIC BANKING

Autor: ÑAUPAS CARAZA, Carol Maribel

Asesor: ZAVALETA CAMPOS, Jorge Luis

Grado: Ingeniero de Sistemas e Informática

Titulo: Tesina, para optar el Título Profesional de Ingeniero de Sistemas

Fecha: Enero 2016

ABSTRACT

The Fraud is one of the main risks that banks face and electronic channels are the main target of the attacks of fraudsters, which is why we must take various measures to prevent and make decisions in real time.

Because technological advances are enabling companies to manage large volumes of data, it is of great value analyze, produce information and knowledge discovery from them, to be strategically used in decision -making.

This paper proposes through the implementation of a process of knowledge discovery in databases , generating an automatic model to categorize transactions of Internet Banking and Mobile Banking individuals of a financial institution , as fraudulent or integrity , by applying Predictive Mining techniques based classification trees .

Keywords: Knowledge Discovery in Databases, Predictive Data Mining, Classification Trees.

ÍNDICE DE FIGURAS:

ÍNDICE DE TABLAS

ÍNDICE DE CONTENIDOS

INTRODUCCION

El desarrollo de las nuevas tecnologías hace que las entidades financieras avancen a la par con ellas y ofrezcan a sus clientes diversos medios y aplicaciones que les faciliten realizar sus transacciones financieras, sin embargo esto es atractivo no solo para los clientes sino también para los defraudadores pues hace que cada vez existan más maneras de cometer fraudes.

Cada día crece más la cantidad de usuarios que prefieren realizar sus operaciones mediante los canales electrónicos, tales como la banca por internet y la banca móvil, en vez de ir a una oficina, por la facilidad y ahorro de tiempo que les proporcionan, además que esta tendencia es también promovida por las entidades bancarias para migrar la mayor cantidad de transacciones que se realizaban en oficinas hacia los canales electrónicos.

Según informe de prensa de ASBANC, Asociación de Bancos del Perú, la banca móvil es el canal de atención bancaria que más ha crecido en el 2015. En el periodo enero-septiembre del 2015, mostró una expansión de 6,585% a tasa anual, al registrar S/ 1,633 millones en operaciones monetarias. El siguiente canal con mejor desempeño fue el de banca por internet, que aumentó 67.81% a S/. 167,977 millones en el periodo de análisis *[ASBANC 2015]*.

Sin embargo este crecimiento y facilidades tienen que ir acompañadas con la seguridad, lo cual obliga a las entidades financieras a ofrecer canales más seguros y emplear técnicas para combatir el fraude.

Es por ello que el ámbito de este proyecto contemplará el uso de la minería de datos para la detección de fraudes en las transacciones realizadas en los canales de banca por internet y banca móvil de entidades financieras, proporcionando una herramienta o modelo, que permita obtener conocimiento, clasificando las transacciones como fraudulentas o integras por medio de la aplicación de algoritmos de Minería de Datos Predictiva.

Esta tesina está organizada en 5 capítulos. En capítulo 1 se detalla el planteamiento metodológico de la investigación, empezando por la definición del problema, sus antecedentes, los objetivos de la investigación, la delimitación y alcance del estudio. En el capítulo 2 se describen los conceptos claves para el desarrollo y comprensión de la solución propuesta. En el capítulo 3 Estado del Arte se describen las principales técnicas existentes y trabajos previos de estudio para la solución al problema. En el capítulo 4 se enfoca en la aplicación de la metodología seleccionada y el desarrollo de la solución para la clasificación de transacciones en una entidad financiera y se analizará los resultados obtenidos. En el capítulo 6 se presentan las conclusiones y recomendación a partir de los resultados obtenidos.

CAPITULO I. PLANTEAMIENTO METODOLÓGICO

1.1 Antecedentes del problema

El fraude es uno de los principales riesgos en el sistema bancario y las cifras a las que ascendió en el 2015 fueron de US$ 5 millones. Entre los fraudes más comunes se encuentran la suplantación de identidad y la clonación de tarjetas *[La Republica 2015]*.

Si bien la tecnología ha permitido aumentar la seguridad en el acceso a los canales electrónicos mediante claves de acceso, uso de dispositivos adicionales de autenticación como tokens, tarjetas coordenadas y clave SMS, las entidades bancarias están obligadas a ofrecer canales más seguros y emplear técnicas no solo para prevenir sino también para detectar el fraude en tiempo real y tener la capacidad de reaccionar oportunamente frente a alguna operación fraudulenta.

Ante esta problemática cada banco cuenta con un área dedicada exclusivamente a la prevención del fraude buscando reducir las pérdidas sufridas. Los sistemas de prevención han sido bastante útiles para la detección en línea pero resultan insuficientes pues muchas veces no logran detectar el cambiante comportamiento del defraudador que busca no ser atrapado y se las ingenia para mezclarse entre los patrones de compra habituales de los clientes pasando desapercibidos dentro de la herramienta de monitoreo de las transacciones. Los datos anteriores dan una idea del impacto que tiene el fraude en el sector bancario, además del costo por mantenimiento del área.

El impacto del fraude no solo afecta a la institución, afecta también a sus clientes, además que en muchas ocasiones una operación puede ser calificada como fraudulenta deteniendo alguna transacción verdadera lo que puede provocar malestar en el cliente.

En tal sentido, se propone realizar un estudio, que dada las características de una transacción y el historial de transacciones fraudulentas obtenidas en el pasado, permita clasificarla como fraudulenta o integra, mediante la obtención de un

modelo de aprendizaje, a partir de la base de datos de transacciones de una entidad financiera, y la aplicación automatizada de algoritmos de minería de datos.

1.2 Definición o formulación del problema

El delito de fraude contra el sistema financiero constituye un problema para los bancos y sus clientes, dado que ocasionan pérdida económica, pérdida de imagen y desconfianza de los clientes. Esto puede implicar disminución de las operaciones, así como de crecimiento y de expansión.

Las entidades financieras han reforzado sus estrategias de seguridad para prevenir y mitigar el riesgo de ser víctima de fraude ya sea en contra suya o de sus clientes. Una de las estrategias es implementar controles que permitan descubrir y prevenir estos delitos y en consecuencia minimizar los daños que podrían ocasionar, a pesar de los constantes esfuerzos el riesgo no se minimiza completamente.

Una vez identificados los riesgos de fraude a los que se enfrentan las entidades financieras, resulta imprescindible utilizar herramientas informáticas, que permitan identificar dentro de miles o millones de transacciones y registros, patrones de comportamiento que son inusuales y/o que corresponden a actividades potencialmente fraudulentas.

Detectar las situaciones de fraude conlleva a evitar daños, proteger la reputación, los activos corporativos e incrementar la confianza por parte de los clientes. En tal sentido, se propone realizar un estudio, que dada las características de una transacción y el historial de transacciones fraudulentas obtenidas en el pasado, permita clasificarla como fraudulenta o integra, mediante la obtención de un modelo de aprendizaje, a partir de la base de datos de transacciones de una entidad financiera, y la aplicación automatizada de algoritmos de minería de datos.

1.3 Objetivos

1.3.1 Objetivo General

Implementar un modelo basado en técnicas de Minería de Datos que permitirá clasificar las transacciones realizadas en los canales de banca por internet o banca móvil como fraudulentas o integras, por medio de la aplicación de un proceso de descubrimiento de conocimientos en bases de datos, mediante la aplicación de algoritmos arboles de clasificación.

1.3.2 Objetivos Específicos

- Revisar las principales técnicas de Minería de Datos para la detección de fraude.

- Definir modelos de características generales de fraude (patrones), en base a los datos de las transacciones realizadas en los canales electrónicos, banca por internet y banca móvil, y con estos hacer predicciones de futuros ingresos de transacciones que pudieran ser sospechosas de fraude.

- Implementar un modelo que permita validar el método propuesto y permita clasificar una transacción como fraudulenta o integra.

1.4 Justificación

El objeto de este estudio consiste en el análisis de las características de las transacciones financieras realizadas en canales electrónicos, para clasificarlas como fraudulenta o íntegra, mediante la obtención de conocimiento, a través de la aplicación automatizada de algoritmos de minería de datos dentro de un proceso de descubrimiento de conocimiento en bases de datos.

Las técnicas consideradas en el desarrollo de este trabajo permitirán identificar patrones de comportamiento y detectar anomalías en los mismos, los cuales son perjudiciales para el cliente y la entidad financiera, considerando para el caso de

estudio el fraude por uso de canales electrónicos como la banca por internet y la banca móvil, para los cuales se pueden registrar operaciones en diferentes horarios que pueden ser realizadas o no por el cliente.

La detección de Fraude no es un tema trivial, las metodologías usadas por los defraudadores no son las mismas de hace algunos años; cuando las entidades identifican un patrón de comportamiento, los defraudadores ya están pensando en otras alternativas, es por ello que las soluciones deben ir actualizándose conforme a los datos actuales.

También se busca cumplir con el reglamento de la Superintendencia de Banca y Seguro (SBS), el cual tiene por finalidad reforzar la administración, expedición y seguridad de las tarjetas. Entre los aspectos que son de especial interés para la prevención del fraude financiero, a través de canales electrónicos, están los Artículos 16 y 17 *[SBS 2013]* que dictan medidas relativas al monitoreo de transacciones y la protección de datos de las tarjetas de crédito y débito a través del cumplimiento del PCI Data Security Standard (PCI DSS).

Existen diversas herramientas o productos comerciales en el mercado para la detección de fraude electrónico, que pueden adquirir las instituciones financieras y realizar un proceso de adaptación y configuración consumiendo un tiempo considerable, representando altos costos.

Por esta razón se propone como solución a tal problemática un modelo que permita clasificar transacciones electrónicas financieras en la banca por internet y banca móvil como fraudulentas o integras, por medio de la aplicación de un proceso de descubrimiento de conocimientos en base de datos, procurando favorecer considerablemente los procesos de gestión del área de seguridad de las entidades bancarias.

1.5 Hipótesis

Se plantea que el estudio propuesto ayude a identificar operaciones fraudulentas realizadas en los canales electrónicos con un alto grado de asertividad, de modo que ayude a identificar el fraude en entidades bancarias permitiendo tomar acciones oportunas al momento de que este se presente.

1.6 Alcance del estudio

1.6.1 Delimitación del problema

El alcance de estudio del presente trabajo entra en acción cuando el fraude está por ser cometido por el defraudador, es decir para esto él ya cuenta con la información del cliente de acceso a los canales, los cuales ha podido obtener previamente por las diversas modalidades de robo de información que serán descritas en los siguientes capítulos.

Los canales en los que se basará el estudio son la banca por internet y la banca móvil, ya que a pesar de que constantemente se toman diversas medidas de seguridad para prevenir el fraude, estos canales siguen siendo unos de los blancos principales de los defraudadores.

Entre los inconvenientes que pudieran presentarse en el desarrollo de este trabajo, se encuentra lo relacionado a la criticidad y confidencialidad de la información, datos relacionados con cuentas financieras, clientes y sus hábitos de consumo. Para tal fin, durante las etapas de preparación, limpieza y selección de variables, se determinará que los datos tomar sean los necesarios y que no sean críticos ni identificadores del cliente, y que aporten mayor ganancia de información, sin afectar la investigación.

1.6.2 Aspectos de estudio que comprende el problema

- El estudio consistirá en la aplicación de un proceso de descubrimiento de conocimientos, con la finalidad de generar un modelo que permitirá la clasificación de transacciones electrónicas financieras realizadas en la banca por internet y banca móvil como fraudulentas o integras, y se determinará si el modelo puede resolver el problema propuesto o se adapta a la solución.

- El proceso de descubrimiento de conocimiento involucra la aplicación iterativa de las fases de selección y preparación de los datos, la aplicación de los algoritmos y técnicas de aprendizaje de minería de datos, y la interpretación de los resultados obtenidos para dar significado a los patrones encontrados.

- Los algoritmos a utilizar durante el desarrollo de este modelo pertenecen a las técnicas de arboles de clasificación, para el caso de estudio se aplicará el algoritmo de C4.5.

- Para determinar si la solución a desarrollar resuelve el problema, se realizará la validación de la misma, comprobando que las conclusiones que arroja son válidas en términos de exactitud.

1.6.3 Propuesta Metodológica

La propuesta que se plantea en este trabajo es el estudio de las técnicas de minería de datos predictiva aplicadas en la explotación de la información histórica transaccional para definir patrones de fraude y la información en línea en el acceso a los canales de estudio, para que al momento de realizarse una operación si se cumplen ciertos patrones de comportamiento se pueda clasificar una transacción como fraudulenta o integra.

El proyecto se desarrollará siguiendo cada uno de los pasos de la metodología KDD Descubrimiento de Conocimiento en Bases de Datos, la cual se detallara en los posteriores capítulos.

1.7 Organización de la tesina

Esta tesina está organizada de la siguiente manera:

Capítulo I: Planteamiento Metodológico

En este capítulo se ha descrito el problema, sus antecedentes, delimitación y alcance, también los objetivos a cumplir, la justificación del caso de estudio, y cual será el universo donde se analizará y aplicara la solución, también se mencionara las técnicas a utilizar para el desarrollo de la solución.

Capítulo II: Marco Teórico

En este capítulo se describen todos los conceptos que se necesitan conocer y que intervienen directamente en el desarrollo del trabajo.

Capítulo III. Estado del Arte Metodológico

En este capítulo se tratará las soluciones existentes y trabajos previos sobre mismo ámbito de estudio.

Capítulo IV. Desarrollo de la solución

En este capítulo se justificará la metodología seleccionada, se describirá el desarrollo de la solución aplicada y los resultados obtenidos.

Capítulo V. Conclusiones y recomendaciones

En este capítulo se verán el análisis de los resultados, y las conclusiones obtenidas del estudio, así como las recomendaciones.

CAPITULO II. MARCO TEÓRICO

A continuación se detallará los principales conceptos teóricos que se aplicarán en el presente trabajo:

2.1 FRAUDE ELECTRONICO

Se conoce como fraude electrónico, al uso indebido o manipulación fraudulenta de elementos informáticos de cualquier tipo, líneas de comunicaciones, información mecanizada, que posibilita un beneficio ilícito. Toda conducta fraudulenta realizada a través o con la ayuda de un sistema informático, por medio del cual alguien trata de obtener un beneficio.

2.1.1 Tipos de fraude

Los tipos de fraude más comunes son:

a) Software espía

Mediante esta modalidad, el delincuente utiliza un software espía (programas que se instalan en el computador sin autorización) que permite monitorear las actividades del usuario de dicho computador (por ejemplo páginas que visita, tipo de información que busca, etc.) desde otro computador remoto, e incluso la información que escribe en su teclado y los contenidos de sus correos electrónicos.

b) Obtención de datos personales o Phishing

Consiste en la obtención fraudulenta de datos personales de los clientes con el fin de realizar posteriormente operaciones no autorizadas con sus cuentas. Para ello los estafadores remiten correos electrónicos hacia los clientes de una entidad financiera, utilizando el logo y los colores característicos de dicha entidad. En estos correos, solicitan que el cliente acceda a una página web en la cual se le pedirá que ingrese su número de tarjeta, clave secreta u otros datos personales, para ello incluyen avisos

sobre promociones u ofertas en caso el cliente complete sus datos. Luego de obtenida esta información, los delincuentes realizan operaciones no autorizadas con las cuentas de los clientes afectados. *[SBS 2015]*.

c) Pharming

Consiste en redireccionar al usuario hacia un URL fraudulento sin que este se entere. Cuando un usuario trata de acceder a un URL, la dirección fraudulenta lo lleva hacia sitio fraudulento donde se le presenta una pantalla (similar a la original) en la cual ingresa su user y password. El sitio fraudulento responde que hay error en user y/o password y que se debe intentar de nuevo. Cuando el usuario reintenta, es direccionado al sitio legítimo.

d) Vshing

Es una práctica criminal fraudulenta en donde se hace uso del Protocolo Voz sobre IP (VoIP) y la ingeniería social para engañar personas y obtener información personal. El término es una combinación del inglés "voice" (voz) y phishing.

- El criminal llama a números telefónicos aleatorios.
- Cuando la llamada es contestada, una grabación le informa al cliente que debe llamar a un número telefónico específico para comunicase con la entidad bancaria.
- Cuando la víctima llama a este número, le contesta una grabación que le indica al "cliente" que su cuenta necesita ser verificada y le solicita información financiera (usualmente números de tarjeta, usuarios y claves).
- De esta manera el delincuente tiene toda la información necesaria para realizar operaciones fraudulentas por canales como banca por teléfono o internet.

e) Smishing

Es una práctica fraudulenta en donde se hace uso de los mensajes de texto de los celulares y la ingeniería social para engañar a personas y obtener información personal y financiera. El término es la combinación entre SMS (mensajes de texto a través de telefonía celular) y Phishing.

Puede ser de dos tipos:

- El delincuente envía en un mensaje de texto con una dirección de una página fraudulenta de internet a través de la cuál recogen la información del usuario.
- El delincuente envía un mensaje de texto con un número telefónico falso que aparenta ser el Call Center del banco. Cuando la persona llama, capturan su información.

f) Key Logger

El delincuente utiliza herramientas de software o hardware que permiten grabar el texto que escribe una persona en su teclado. En el caso del software, el key logger captura todo lo que escribe el usuario y lo envía a una dirección de correo electrónico configurado por el delincuente. Estos programas se instalan y funcionan de manera 'invisible' (no se da cuenta el usuario).

En el caso del hardware, existen unos dispositivos que se conectan al computador y graban en una memoria interna el texto tecleado.

g) Ingeniería Social

La ingeniería social es la práctica de obtener información confidencial a través de la manipulación de usuarios legítimos. Con esta técnica, el ingeniero social se aprovecha de la tendencia natural del hombre a confiar en la gente, engañarles para romper los procedimientos normales de seguridad y manipularles para realizar acciones o divulgar información sensible.

2.2 KDD DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

KDD (Knowledge Discovery in Databases) es una metodología genérica para encontrar información en un gran conjunto de datos y con ello generar conocimiento. Se define como un proceso no trivial de extracción de información a partir de los datos, la cual se encuentra presente de forma implícita, previamente desconocida y potencialmente útil para el usuario o para el negocio *[Fayyad 1996]*.

El objetivo principal de esta metodología es automatizar el procesamiento de los datos, permitiendo a los usuarios dedicar más tiempo a las tareas de análisis y al descubrimiento de relaciones entre los datos.

El KDD es un proceso que consta de una serie de etapas consecutivas, y funciona de forma iterativa e interactiva. Iterativa, ya que es posible regresar desde cualquier etapa a una anterior para ajustar los parámetros o supuestos previos, e interactiva pues el usuario experto del negocio tiene que estar presente para aportar con su conocimiento en la preparación de los datos y en la validación de los resultados que se obtengan durante el proceso. En la figura 1 se muestran las etapas del KDD *[Fayyad 1996]*.

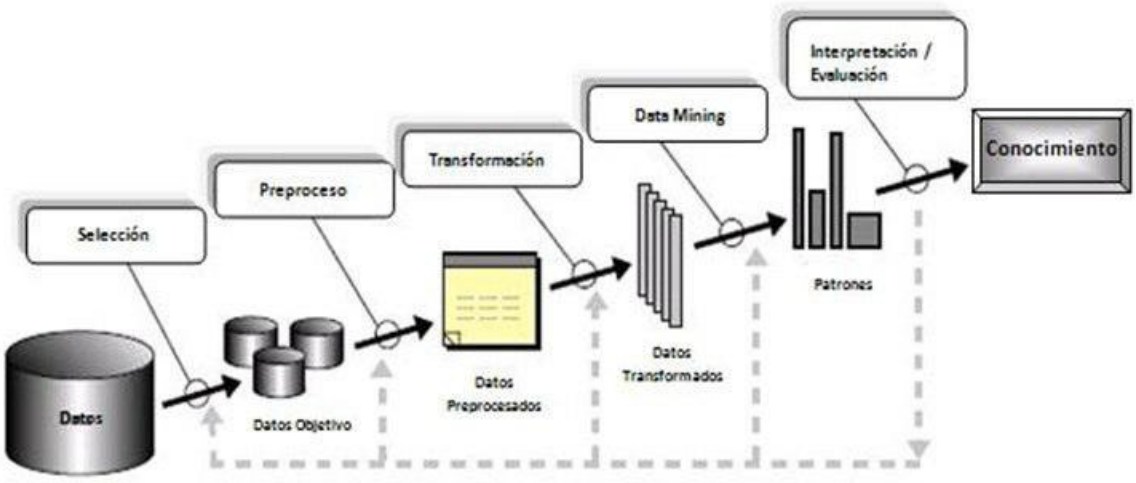

Figura 1. Proceso de KDD. Fuente: (Fayyad 1996)

Las etapas de este proceso KDD son las siguientes:

- **Selección de Datos**

  Consiste en la extracción de los datos relevantes o de interés al área de análisis del almacenamiento de los datos, eligiendo las variables más determinantes en el problema.

  La selección de datos puede ser de forma horizontal, en el sentido de que sólo se eligen instancias completas representativas del total de los datos disponibles.
  En el caso que se realice una selección vertical, es decir, de atributos, la idea es seleccionar los atributos más relevantes en base a algún criterio o al problema en particular.

- **Limpieza e Integración de Datos**

  Es el procesamiento de tratamiento de los datos ruidosos, erróneos, faltantes irrelevantes, y la integración de múltiples fuentes de datos en una única fuente.

- **Transformación de los Datos**

  La transformación consiste en consolidar los datos en formas apropiadas para ser introducidos en el algoritmo de minería. Este paso, tiene como una de sus tareas la construcción de nuevos atributos, por medio de la aplicación de alguna operación o función a los atributos originales, solo en los casos en que estos últimos no aporten suficiente poder predictivo por si solos.

- **Minería de Datos**

  Es el paso esencial donde se aplican diversos métodos para extraer patrones, en este caso, como algoritmos de minería se aplicarán técnicas de aprendizaje para obtener un modelo de conocimiento, el cual representa patrones de comportamiento observados en los valores de las variables del problema.

Es la etapa de descubrimiento en el proceso de KDD, paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados. Es en esta etapa, donde se determina la relevancia y calidad de la data preprocesada. La selección de un algoritmo acorde al problema va a ser determinante en la validez total del modelo.

Todas las técnicas de modelación tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los parámetros óptimos para la técnica de modelación es un proceso iterativo y se basa exclusivamente en los resultados generados.

- **Interpretación y Evaluación**

  En la fase de interpretación y evaluación, se procede a la validación del modelo obtenido, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. Si el modelo no alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar un nuevo modelo *[Vallejos 2006]*.

- **Conocimiento.**

  Aplicación del conocimiento descubierto.

2.3 TECNICAS DE MINERÍA DE DATOS

La Minería de Datos o Data Mining es una etapa del KDD (Knowledge Discovery in Databases - Descubrimiento de conocimiento en Bases de Datos), y consiste en un conjunto de técnicas de múltiples disciplinas tales como: tecnología de bases de datos, estadística, aprendizaje, reconocimiento de patrones, visualización de datos, obtención de información, procesamiento de imágenes y de señales y análisis de datos *[Fayyad 1996]*.

La minería de datos ofrece un rango de técnicas que permiten identificar casos sospechosos, basados en modelos *[Santamaría 2010]*. Estos modelos se pueden clasificar en:

- **Modelos de datos inusuales**

  Estos modelos, pretenden detectar comportamientos raros en un dato respecto a su grupo de comparación, o con el mismo, por ejemplo la consignación de altas sumas de dinero en efectivo. Para este caso, se puede emplear técnicas de análisis de Agrupamiento (Clustering), seguido de un análisis de detección de Outlier.

- **Modelos de relaciones inexplicables**

  A través de este tipo de modelos, se desea encontrar relaciones de registros que tienen iguales valores para determinados campos, resaltando el hecho que la coincidencia de valores debe ser auténticamente inesperado, desechando similitudes obvias.

- **Modelos de características generales de Fraude**

  Con estos modelos se pretende, una vez detectado ciertos casos, hacer predicciones de futuros ingresos de transacciones sospechosas.
  Para estas predicciones usualmente se emplean técnicas de regresión, arboles de decisión y redes neuronales.

La minería de datos se puede dividir en dos clases: Descriptiva y Predictiva (clasificación) como se presenta en la Figura 2 *[Santamaría 2010]*.

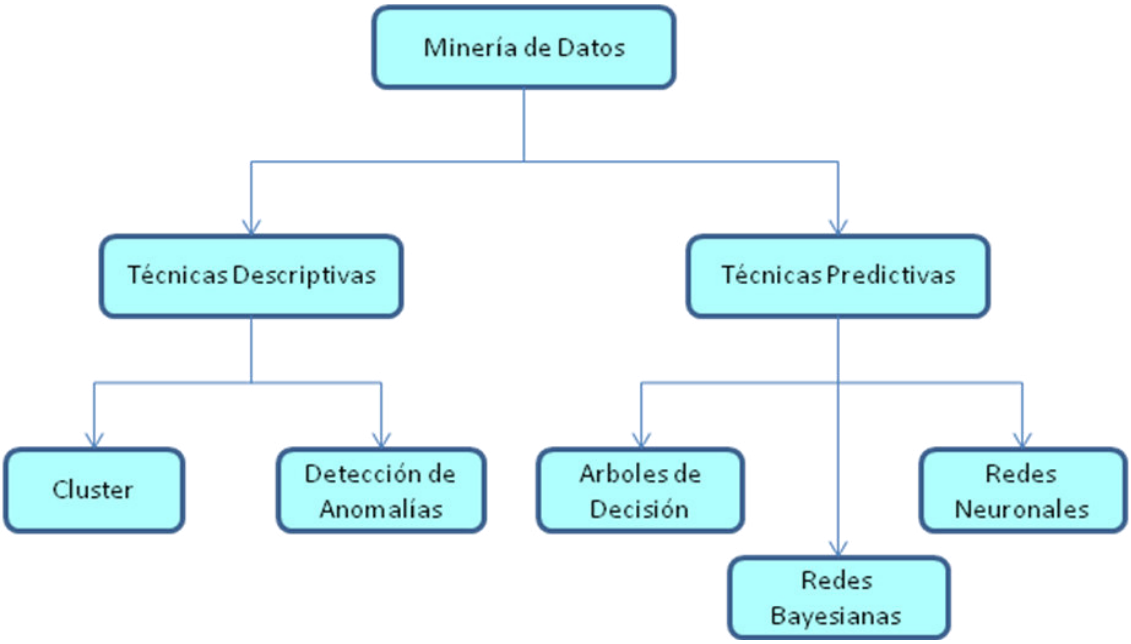

Figura 2. Taxonomía de Técnicas de Minería de Datos. Fuente: (Santamaría 2010)

2.3.1 Técnicas de Minería Descriptiva:

El objetivo de este tipo de minería, es encontrar patrones (correlaciones, tendencias, grupos, trayectorias y anomalías) que resuman relaciones en los datos *[Chen 1996]*. Dentro de las principales técnicas descriptivas encontramos:

a) Detección de Anomalías(Outlier):

La meta principal en la detección de Anomalías es encontrar objetos que sean diferentes de los demás, estos objetos son conocidos como Outlier.

La detección de anomalías también es conocida como detección de desviaciones, porque objetos anómalos tienen valores de atributos con una desviación significativa respecto a los valores típicos esperados.

Aunque los Outlier son frecuentemente tratados como ruido o error en muchas operaciones, tales como Clustering, para propósitos de detección

de fraude, son una herramienta valiosa para encontrar comportamientos atípicos en las operaciones que un cliente realiza en una entidad financiera.

Las técnicas actuales de detección de Outlier se clasifican en:

- **Técnicas basadas en Modelos.** Se basan en el campo de la estadísticas; dada la premisa de conocer la distribución de los datos.

- **Técnicas basadas en proximidad.** Esta técnica se fundamenta en el manejo de distancias entre objetos, entre mayor sea la distancia del objeto respecto a los demás, éste es considerado como un Outlier.

- **Técnicas basadas en densidad.** Se hace uso de la estimación de densidad de los objetos, para ello, los objetos localizados en regiones de baja densidad, y que son relativamente distantes de sus vecinos se consideran anómalos.

Este método de minería de datos, generalmente es de aprendizaje no supervisado, ya que en la mayoría de los casos, para ello se asigna una calificación a cada instancia que refleja el grado con el cual la instancia es anómala.

b) Clustering:

El análisis de cluster es un proceso que divide un grupo de objetos, de tal forma que los miembros de cada grupo son similares de acuerdo a alguna métrica. El agrupamiento de acuerdo a la similitud, es una técnica muy poderosa, la clave para esto es trasladar alguna medida intuitiva de similitud dentro de una medida cuantitativa.

Las técnicas de clustering son utilizadas comúnmente para hacer segmentación. Las técnicas de segmentación permiten identificar claramente el comportamiento de un grupo de casos que difiere de otros

grupos o conjuntos. Se emplea en el reconocimiento de patrones, en el descubrimiento de perfiles de clientes, entre otros.

Los algoritmos de cluster funcionan con una metodología basada en la construcción inicial de un gran cluster, y luego la subdivisión del mismo hasta encontrar grupos de muestras muy cercanas, otros por el contrario, parten asumiendo que cada registro es un cluster, y luego empiezan a agrupar registros hasta que se consolidan cluster no superpuestos más grandes. *[Santamaría 2010]*.

2.3.2 Técnicas De Minería Predictiva

El objetivo de este tipo de minería, es predecir/clasificar el valor particular de un atributo basado en otros atributos. El atributo a predecir es comúnmente llamado "clase" o variable dependiente, mientras que los atributos usados para hacer la predicción se llaman variables independientes. *[Tan 2005]*

Dentro de las principales técnicas predictivas encontramos:

a) Árboles de decisión

Un árbol de decisión es un conjunto de condiciones organizadas en una estructura jerárquica, de tal manera que la decisión final a tomar, se puede determinar siguiendo las condiciones que se cumplen desde la raíz del árbol hasta sus hojas. Se utilizan comúnmente cuando se necesitan detectar reglas del negocio que puedan ser fácilmente traducidas al lenguaje natural o SQL, o en la construcción de modelos predictivos.

Existen dos tipos de árboles:

- **Los arboles de clasificación**, mediante los cuales un registro es asignado a una clase en particular, reportando una probabilidad de pertenecer a esa clase.
- **Los árboles de regresión**, que permiten estimar el valor de una variable numérica objetivo.

El funcionamiento general de un árbol se basa en la aplicación de premisas que pueden ser cumplidas, o no, por un registro; el registro pasa a través del árbol de premisa en premisa hasta que se evalúa totalmente o hasta que encuentra un nodo terminal.

Los caminos que describe el árbol para llegar a los nodos terminales, representan el conocimiento adquirido y permiten la extracción de reglas de clasificación de la forma IF-THEN.

Según el tema de estudio, los arboles pueden crecer tanto que resultan difíciles de interpretar, o muy cortos que arrojan respuestas obvias o insuficientes. La mayoría de los algoritmos y herramientas en el mercado permiten la configuración de los parámetros como el tamaño mínimo de nodos, dado que cada uno de los nodos de árbol corresponden a una pregunta sobre una variable especifica, los arboles de decisión no pueden descubrir reglas que impliquen relaciones entre variables.

Entre los principales algoritmos de aprendizaje de arboles de decisión se encuentran:

- **CART** propuesto por Breinman. Se basa en el lema "divide y vencerás", son métodos que construyen arboles binarios basados en el criterio de partición GINI y que sirven para clasificación como para regresión. La poda se basa en una estimación de la complejidad del error.

- **ID3.** Propuesto por Quintlan en 1986, es considerado el árbol de decisión más simple, usa la ganancia de información como criterio de separación. El árbol crece hasta encontrar un nodo final. No emplea procedimientos de poda, ni manejo de valores perdidos.

- **C4.5**. Es la evolución del ID3, presentado por Quinlan en 1993. Usa como criterio de separación el radio de ganancia.

b) Redes Neuronales

Las redes neuronales consisten en "neuronas" o nodos interconectados que se organizan en capas. Por lo general los modelos neuronales constan de tres capas: de entrada, oculta y de salida. Cada neurona evalúa los valores de entrada, calcula el valor total de entrada, compara el total con el mecanismo de filtrado (valores de umbral), y en seguida determina su propio valor de salida. El comportamiento complejo se modela conectando un conjunto de neuronas. El aprendizaje o capacitación" ocurre modificando la fuerza de conexión o los parámetros que conectan las capas. Las redes neuronales se acondicionan con muestras adecuadas de la base de datos.

Las redes neuronales aprenden en forma supervisada o no supervisada. En la modalidad supervisada, la red neuronal intenta predecir los resultados para ejemplos conocidos, compara sus predicciones con la respuesta objetivo y aprende de sus errores. Las redes neuronales supervisadas se emplean para predicción, clasificación y modelos de series históricas. Las redes supervisadas crean sus propias descripciones y validaciones de clase y trabajan exclusivamente a partir de los patrones de datos. El aprendizaje no supervisado es eficaz para la descripción de datos, pero no para la predicción de resultados.

Las redes neuronales se ven afectadas por tiempos prolongados de aprendizaje. Debido a que actúan como una caja negra, algunos analistas empresariales no confían en ellas.

Las redes neuronales se utilizan generalmente para identificar patrones de comportamiento, el uso más común que tienen las redes neuronales es en la detección de fraude. Esta técnica es altamente utilizada en modelos predictivos basados en análisis históricos.

Entre más grande sea una red, es decir, más capas ocultas posea o mayor número de nodos, la complejidad de la ecuaciones matemáticas que se

deben resolver al interior del nodo de salida se aumenta excesivamente, lo que hace prácticamente imposible entender su funcionamiento o explicar el resultado. Las redes se utilizan en casos en que el resultado es más importante que el "como", dado que constituyen modelos no lineales que no producen reglas.

Para lograr un buen funcionamiento de las redes es importante realizar un buen entrenamiento, el cual consiste, de manera general, en la asignación de los pesos que debe tener cada variable de entrada con el fin de lograr la mejor aproximación. En la construcción o utilización de una red se deben preparar cuidadosamente los conjuntos de datos a utilizar, por ejemplo, en una red no se utilizan valores categóricos, solo numéricos, por lo que para aquellas variables categóricas como: país, ciudad, etc., se debe asignar un número por cada valor posible "variables Dummy" *[Santamaría 2010]*.

Entre los modelos más utilizados en redes neuronales se encuentran:

- **Perceptrón Multicapa (MLP) o Hacia Adelante (Feedforward).**

  Es el modelo más estudiado y usado en la industria. Un MLP es una red conformada por una capa de entrada, una o varias capas ocultas, una salida y una función de transferencia en cada nivel. Se caracterizan por tener una conexión completa entre capas sucesivas, es decir, cada nodo en una capa está totalmente conectado solo a todos los nodos en las capas adyacentes.

- **Hopfield.**

  Son un tipo especial de redes, capaces de guardar recuerdos o patrones como el cerebro, no tienen una arquitectura de capas, sino por el contrario, es una sola capa de neuronas completamente interconectadas, en las cuales hay bucles de retroalimentación entre las neuronas.

- **Mapas Auto-organizados**

  (Kohonen's Self-organizing Maps -SOM). Son modelos de redes neuronales para la reducción de dimensiones y agrupación de datos, con el fin de visualizar similitudes entre patrones.

c) Redes de Creencia Bayesiana

La clasificación Bayesiana se basada en el teorema estadístico de Bayes, el cual provee un cálculo para la probabilidad a posteriori. De acuerdo al teorema de Bayes, si H es una hipótesis, tal que, el objeto X pertenece a la clase C, entonces la probabilidad que la hipótesis ocurra es:
*P (X|H) = (P(X|H) * P(H)) / P(X).*

Una Red de Creencia Bayesiana provee una representación grafica de las dependencias entre un conjunto de atributos. Se compone principalmente de dos elementos:

- **Un grafo acíclico** que codifica la dependencia de relaciones entre un conjunto de variables.

- **Una tabla de probabilidad** asociada a cada nodo para su nodo padre inmediato.

En una Red de Creencia Bayesiana, para cada nodo *X*, existe una tabla de probabilidad condicional, en la cual se especifica la probabilidad condicional de cada valor de *X*, para cada posible combinación de los valores se sus padres (distribución condicional *P(x|padre(x))*). La estructura de la red puede ser definida o ser inferida desde los datos. Para propósitos de clasificación uno de los nodos puede definirse como nodo "clase". La red puede calcular la probabilidad de cada alternativa de "clase" *[Santamaría 2010].*

En la Tabla 1, se presenta un breve resumen de las tareas, metas y técnicas de Minería más utilizadas en la Detección de Fraude *[Santamaría 2010]*.

Tarea	Meta	Técnica de Minería
Encontrar Datos inusuales	Detectar registros con valores anormales. Detectar múltiples ocurrencias de valores. Detectar relaciones entre registros.	Análisis de Anomalías
Identificar Relaciones Inexplicables	Determinar perfiles. Determinar registros duplicados. Detección de registros con referencias de valores anormales. Detectar relaciones indirectas entre registros. Detectar registros con combinaciones de valores anormales.	Análisis de Cluster Análisis de Cluster y Anomalías Análisis de Relaciones Asociación
Características generales de Fraude	Encontrar criterios, tales como reglas. Calificación de transacciones sospechosas.	Modelos Predictivos

Tabla 1. Técnicas de Minería de Datos. Fuente: (Santamaría 2010)

2.4 ARBOLES DE DECISION

Un Árbol de Decisión es un modelo de predicción construido por un algoritmo de aprendizaje a partir de un conjunto de ejemplos de entrenamiento y evaluación, donde el modelo obtenido sirve para representar una serie de condiciones sucesivas para la resolución de un problema.

El Árbol de Decisión puede recibir de entrada valores discretos como continuos. Cuando los valores de entrada son discretos, el problema a resolver se denomina Clasificación (es el caso más común de aplicación) y cuando los valores de entrada son continuos, el problema a resolver se denomina regresión.

Datos de entrada:

Los datos de entrada pueden ser numéricos o no numéricos. Para distinguir entre ambos tipos generalmente se usan los términos datos cuantitativos y datos cualitativos.

Los datos cuantitativos son los que se utilizan como entrada para el algoritmo de aprendizaje que genera un Árbol de Decisión. Estos se distinguen en datos discretos y datos continuos. Los datos discretos son aquellos en los que el número posible de valores es un número finito y los datos continuos (numéricos) son aquellos que resultan de un número finito de posibles valores que se pueden asociar a los puntos de una escala continua, en un rango de valores sin espacios ni interrupciones.

Existe otra clasificación común de los datos denominada uso de niveles de medición: nominal u ordinal. Los datos nominales consisten exclusivamente en nombres. Etiquetas, categorías o clases que no pueden ordenarse de acuerdo a un esquema.

Elementos de un Árbol de Decisión:

Los elementos de un Árbol de Decisión son los siguientes: raíz, nodos, ramas y hojas *[Russell y Norvig 2004]*.

El **nodo raíz** y los nodos internos del árbol corresponden a una prueba del valor de una de las propiedades y **las ramas nodo** son identificadas mediante los posibles valores de la prueba. En los **nodos hoja** del árbol se especifica el valor que hay que producir en el caso de alcanzar dicha hoja.

Algoritmos de Árboles de Decisión:

La mayoría de los algoritmos utilizados para construir un árbol son variaciones de uno genérico llamado "Greedy algorithm" que básicamente va desde la raíz hacia abajo (Top-Down) buscando de manera recursiva los atributos que generan el mejor árbol hasta encontrar el óptimo global con una estructura de árbol lo más simple posible.

Los algoritmos más conocidos son el ID3, el C4.5 (propuesto por Quinlan en 1993), C5.0 y CART (Classification And Regression Trees). La diferencia entre los 3 primeros y CART es la posibilidad de que este último puede obtener valores reales como resultados versus valores discretos en el resto de los modelos. Las principales características de estos algoritmos es su capacidad de procesar un gran volumen de información de manera eficiente y que pueden manejar el ruido (error en los valores o en la clasificación de estos) que pudiese existir en los datos de entrenamiento. En general los distintos modelos se diferencian en el algoritmo que clasifica los distintos atributos del árbol y la eficiencia de estos de obtener un mejor árbol que sea los más simple posible.

2.4.1 Algoritmo ID3 (Iterative Dichotomies 3)

Este algoritmo fue propuesto por J. Ross Quinlan en 1975 en su libro "Machine learning". Básicamente ID3 construye un árbol de decisión (DT) desde un set fijo de "ejemplos", el DT generado se usa para clasificar futuros ejemplos. Cada ejemplo tiene varios atributos que pertenecen a una clase (como los valores sí o no). Los nodos de "hoja" del árbol (leaf nodes) contienen el nombre de la clase, mientras que los nodos "no-hoja" son los nodos de decisión donde cada uno de ellos (cada rama) corresponde un

posible valor del atributo. Cada nodo de decisión es una prueba del atributo con otro árbol que comienza a partir de él.

El algoritmo ID3 utiliza el criterio de la "ganancia de información" para decidir que atributo va en cada nodo de decisión. Esta medida estadística mide que tan bien un atributo divide los ejemplos de entrenamiento en cada una de las clases seleccionando aquella con más información (información útil para separar). Para definir "ganancia de información" primero debemos definir el concepto entropía que básicamente corresponde a la cantidad de incertidumbre en un atributo.

Dada una colección S (data sobre la cual se calcula la entropía que cambia con cada iteración) con c posibles resultados:

$$Entropía(S) = \sum -p_x \log_2 p_x$$

Donde:

- S: es una colección de objetos
- p_x: es la probabilidad de los posibles valores
- x: las posibles respuestas de los objetos.

Cuando $Entropía(S) = 0$, entonces el set *S* está perfectamente clasificado, es decir, que todos los elementos de *S* están en la misma clase (si es 1 es que están clasificados en forma aleatoria). De la definición anterior se desprende que la "ganancia de información" (Information Gain "IG") corresponde a la diferencia en entropía entre antes de haber separado la data *S* por un atributo versus después de hacerlo, es decir, en cuanto se redujo la incertidumbre en el set *S* después de dividirla en el atributo "A":

$$IG(S,A) = Entropía(S) - \sum \left(\frac{|S_v|}{|s|}\right) Entropía(x)$$

Donde S_v es el set de *S* para el cual el atributo *A* tiene el valor *v*. Los elementos $|S_v|$ y $|S|$ corresponden al número de observaciones en S_v y *S* respectivamente.

2.4.2 Algoritmo C4.5

Este algoritmo fue desarrollado por Ross Quinlan en 1993 y básicamente es una versión avanzada del algoritmo ID3 en el que se incluyen las siguientes capacidades o ventajas *[Quinlan 1993]*:

- Manejo de valores continuos y discretos: Para manejar atributos continuos lo que hace el algoritmo es crear un umbral para después dividir el atributo entre aquellos que están sobre y bajo el umbral. Esta característica es fundamental para este estudio donde la mayoría de los valores son continuos y sus umbrales pueden ser de alta relevancia.

- Tiene la capacidad de manejar valores de atributos faltantes: En el caso de un atributo faltante el algoritmo usa una ponderación de valores y probabilidades en vez de valores cercanos o comunes. Esta probabilidad se obtiene directamente de las frecuencias observadas para esa instancia, por lo que se puede decir que el algoritmo C4.5 usa la clasificación más probable calculada como la suma de los pesos de las frecuencias de los atributos.

- Es capaz de generar un set de reglas que son mucho más fáciles de interpretar para cualquier tipo de árbol.

- Este algoritmo construye un gran árbol y lo concluye con una "poda" de las ramas de manera de simplificarlo de manera de generar resultados más fáciles de entender y haciéndolo menos dependiente de la data de prueba.

Seudocódigo del algoritmo C4.5

Este recibe como argumentos de entrada un conjunto de atributos no clasificadores (R), un atributo clasificador o clase (C) y un conjunto de datos de entrenamiento (S) y como salida, genera un modelo clasificador que se puede representar como un Árbol de Decisión.

Función C4.5 (R: Conjunto de atributos no clasificadores
C: Atributo clasificador
S: Conjunto de datos de entrenamiento) retorno un Árbol de Decisión;

Inicio
Si S esta vacio, entonces
Retorna un nodo único con Valor Error;
Si todos los registros de S tienen el mismo valor para el atributo clasificador, entonces
Retorna un nodo único con dicho valor;
Si R está vacío, entonces
Retorna un nodo único con el valor más frecuente del atributo clasificador
Si R no está vacío, entonces
D ← atributo con mayor Proporción de Ganancia {D, S} entre los atributos de R;
Sean {dj| j= 1,2,…n} los valores del atributo D;
Sean {Sj| j= 1,2,…n} los subcojuntos de S correspondientes a los valores del Dj respectivamente;
Retorna un árbol con el nodo raíz denominado como D y con las ramas denominadas como d1, d2,…, dn
Que descienden respectivamente a los arboles
C4.5 {R-{D}, C, S1}, {R{D},C,S2}, …, {R-{D},C, Sn};

En cada nodo el algoritmo debe decidir cual prueba elegir para particionar los datos.

Los tipos de pruebas propuestas por Quinlan para C4.5 son tres:

- Prueba "estándar" para atributos discretos, obtenidos como resultado y una rama para cada valor posible del atributo.

- Prueba basada en un atributo discreto, en donde los valores posibles son asignados a un número variable de grupos con un resultado posible para cada grupo, en lugar de uno para cada valor.

- Si un atributo *A* tiene valores numéricos continuos, se realiza una prueba binaria con resultados *A<=L y A>L*, para lo que se debe determinar el valor limite de *L*.

Estas pruebas se evalúan de igual forma, observando el resultado de la ganancia de información.

Para el caso de estudio de la presente tesina el algoritmo aplicado es el C4.5 (en la herramienta Rapidminer es el componente Decision Trees).

2.5 HERRAMIENTA RAPIDMINER

RapidMiner (anteriormente, YALE, Yet Another Learning Environment) es un programa informático para el análisis y minería de datos. Permite el desarrollo de procesos de análisis de datos mediante el encadenamiento de operadores a través de un entorno gráfico.

RapidMiner proporciona más de 500 operadores orientados al análisis de datos, incluyendo los necesarios para realizar operaciones de entrada y salida, preprocesamiento de datos y visualización. También permite utilizar los algoritmos de otras herramientas.

Características:

- Desarrollado en Java
- Multiplataforma
- Representación interna de los procesos de análisis de datos en ficheros XML.
- Permite el desarrollo de programas a través de un lenguaje de script.
- Puede usarse de diversas maneras:
  - A través de un GUI
  - En línea de comandos
  - En batch (lotes)
  - Desde otros programas a través de llamadas a sus bibliotecas
- Extensible.
- Incluye gráficos y herramientas de visualización de datos.

CAPITULO III. ESTADO DEL ARTE METODOLÓGICO

El problema del fraude ha sido estudiado en diversos aspectos, mediante técnicas, algoritmos y soluciones de software que se ofrecen en el mercado.

Entre las principales soluciones se encuentran:

3.1 TRABAJOS DE INVESTIGACION SIMILARES

3.1.1 Un algoritmo genético para la detección de fraude electrónico en tarjetas de debito en el Perú [Lavado 2013]

Titulo: "Un algoritmo genético para la detección de fraude electrónico en tarjetas de debito en el Perú".
Autor: Luis Lavado Napaico.
Trabajo de investigación. Universidad Nacional Mayor de San Marcos. Facultad de Ingeniería de Sistemas e Informática.

Resumen:

Este trabajo presenta el modelado y diseño de un algoritmo genético para obtener las reglas más representativas de compra de los tarjetahabientes dentro del universo de datos transaccionales de un banco.

Propone la utilización de un modelo heurístico basado en el comportamiento transaccional de los clientes y la determinación de los patrones de desviación que sean catalogadas como sospechosas empleando técnicas de algoritmos genéticos. El algoritmo genético sigue el enfoque iterativo Rule Learning (IRL) donde la solución global está formada por las mejores reglas de una serie de ejecuciones sucesivas.

De las pruebas experimentales que realizaron obtuvo una precisión de 95.5% en el canal de internet y 95.8 para el canal de punto de venta POS. Este

trabajo concluye que el empleo de la estrategia de algoritmo evolutivo tiene una aceptable exactitud en la predicción.

3.1.2 Segmentación para detección de transacciones inusuales en tarjeta crédito *[Rosas - Uribe 2012]*

Titulo: "Segmentación para detección de transacciones inusuales en tarjeta crédito".
Autor: Blanca Inés Rojas Peña, María Alejandra Uribe Acosta.
Trabajo de investigación. Universidad Nacional de Colombia - Sede Medellín.

Resumen:

En este artículo las autoras proponen realizar la segmentación de transacciones inusuales con tarjetas de crédito, es decir de las operaciones que no corresponden al perfil transaccional del cliente, de acuerdo a las características de cada uno de ellas, con el fin de identificar agrupaciones. Su objetivo principal es identificar patrones de fraude, que sean implementados mediante alertas por cumplimiento de reglas, con el fin de ayudar a los expertos de negocio a examinar y verificar más fácilmente los resultados obtenidos para apoyar la toma de decisiones.

La técnica utilizadas en este estudio son: análisis de conglomerados o clustering, para generar grupos de transacciones con características similares, luego se aplicó la técnica de análisis de discriminante para evaluar la pertenencia de los datos a los grupos generados y por último se uso la técnica de arboles de decisión para interpretación los resultados y construir las reglas que definen los patrones de comportamiento del defraudador. Los datos que utilizaron en este estudio son transacciones fraudulentas con tarjetas de crédito de una entidad financiera.

3.1.3 Modelo de Detección de fraude basado en el Descubrimiento de reglas de clasificación extraídas de una red neuronal *[Santamaría 2010]*

Titulo: "Modelo de Detección de fraude basado en el Descubrimiento de reglas de clasificación extraídas de una red neuronal".
Autor: Wilfredy Santamaría Ruiz.
Tesis de grado para optar por el titulo de Magister en Ingeniería de Sistemas y Computación. Universidad Nacional de Colombia.

Resumen:

En este trabajo el autor presenta un modelo de detección de fraude empleando las técnicas de minería de datos tales como redes neuronales y extracción simbólica de reglas de clasificación a partir de una red neuronal entrenada, que ayude a los expertos del negocio en la toma de decisiones.
El caso de estudio lo aplico sobre un conjunto de datos de una organización que realiza el envío y pago de remesas, y el objetivo es diseñar un modelo que permita identificar patrones ligados a la detección de fraude.

3.2 SOLUCIONES ALTERNATIVAS

3.2.1 Concentrador de Información de Fraude [ASBANC 2015]

Solución tecnológica que se alimenta, en tiempo real, de los sistemas de monitoreo transaccional de las entidades participantes con las transacciones de los tarjetahabientes que presentan consumos no reconocidos, a fin de integrar esta información en un DataMart asociado a un motor de reglas automatizadas que monitorea el comportamiento de fraude del sistema financiero.

Objetivo:

El objetivo principal de este sistema es detectar e informar en tiempo real los puntos comunes donde se podría haber comprometido información para su

posterior investigación, y lugares donde la delincuencia materializa el fraude; cuenta con una interfaz gráfica, donde se puede visualizar alertas y nuevos comportamientos de fraude.

Utiliza la funcionalidad de definición de reglas adaptivas, manejo de score de riesgo, aplicación de redes neurales, generación alertas por correo, etc. para lograr cumplir en forma satisfactoria los objetivos.

Ventajas:

- Detección de lugares de compromiso de información en tiempo real a fin de poder tomar acciones y evitar el uso de las tarjetas comprometida.
- Detección de zonas, comercios y tipos de comercios de riesgo.
- Contar con una fuente de información del comportamiento de fraude del sistema financiero para alimentar las reglas de monitoreo de las entidades participantes.
- Reducción del Riesgo de fraude y pérdidas.
- Contar con indicadores que permitan comparar a cada entidad versus todo el sistema financiero.

Desventajas:

- Solo sirve como fuente de información del comportamiento del fraude, patrones de fraude.
- No detecta o clasifica si en tiempo real si una transacción es fraudulentas o integra.

CAPITULO IV. DESARROLLO DE LA SOLUCIÓN O DEL ESTUDIO

A continuación se propone una solución al problema planteado en esta tesina, en función de los objetivos de la investigación, considerando los pasos del Proceso de Descubrimiento de Conocimientos KDD, explicada en el capítulo anterior, desde la limpieza e integración de los datos hasta la interpretación y evaluación de los resultados, haciendo más énfasis en la etapa de Minería de Datos, pues en esta se aplicará la técnica seleccionada.

La población estará compuesta por todas las operaciones realizadas en los canales de Banca por Internet y Banca Móvil de personas naturales, las cuales se encuentran registradas en una Base de Datos DB2 del Sistema IBM Mainframe AS 4700 (contiene información de hasta 3 meses).

En la fase de limpieza e integración de Datos se seleccionará, con el apoyo del experto del negocio de seguridad y prevención de fraudes, la muestra de transacciones a emplear en el trabajo, las cuales se integrarán en una única fuente de datos, seleccionando solo los atributos o columnas requeridos.
También se integrará a la fuente de datos, diversos atributos de interés relevante al presente estudio, esto con la finalidad de preparar el conjunto de datos de entrenamiento y validación.

Seguidamente, se transformaran los datos, y se exportará la información hacia un archivo plano, con la finalidad de tener una primera vista del conjunto de casos de estudio.
Una vez obtenido el archivo se cargará en el Sistema DataMining para iniciar con la fase de Minería de Datos y aplicar los algoritmos correspondientes a las técnicas seleccionadas.
Durante la fase de Minería de datos se aplicara las técnicas de: Discretización, Normalización, Arboles de Decisión, mediante la aplicación el algoritmo C4.5.
Finalmente, en la fase de interpretación, se analizará los resultados del proceso aplicado para determinar que transacciones fueron detectadas como fraudulentas, y cuáles son las reglas que determinaron su calificación como fraudulenta o íntegra.

4.1 APLICACIÓN DE LA METODOLOGÍA SELECCIONADA

4.1.1 Selección de las fuentes de datos

Teniendo en cuenta que el problema del negocio ya fue identificado en los capítulos anteriores, se prosigue con la descripción de los datos a utilizar y del procesamiento que es necesario realizar a cada uno.

Se inició el proceso considerando las tablas de registros de LOG de transacciones electrónicas, contenidas en la Base de Datos DB2 del Sistema IBM Mainframe AS 4700. El conjunto de datos proviene del Log transaccional de todas las operaciones realizadas en los canales de Banca por Internet y la Banca Móvil de personas naturales, las cuales conformarán el universo de operaciones. Por cada operación realizada se toma los datos requeridos para la evaluación y se completa con información almacenada en las bases de datos del sistema.

A continuación se describen las fuentes de información utilizadas:

a) Catalogo Transacciones fraudulentas:

Contiene las transacciones consideradas como posibles afectas a fraudes, identificadas como riesgosas por el experto del área de fraudes, las cuales serán tomadas como muestra de todo el universo de datos.

Transacción	Descripción	Limite diario
T1	Envío de Efectivo Móvil	S/ 1,500
T2	Transferencia a cuentas de terceros	S/ 3,000
T3	Pago tarjeta de crédito de terceros	S/ 3,000
T4	Transferencias a cuentas de otros bancos	S/ 3,000
T5	Pago de tarjeta de crédito de otro banco	S/ 3,000
T6	Transferencias el exterior	S/ 3,000
T7	Transferencias nacionales	S/ 3,000

Tabla 2. Catalogo de transacciones fraudulentas. Fuente: (Elaboración propia)

b) Log de operaciones:

Contiene la información de la transacción realizada en el canal, fecha y hora de la operación, canal en que se realizó, transacción realizada, importe de la operación, cuenta cargo de la operación (pudiendo ser cuenta o tarjeta), cuenta destino (pudiendo ser cuenta, tarjeta, código interbancario, documento de identidad, puede variar según la transacción realizada), tarjeta de ingreso al canal.

4.1.2 Limpieza e integración de datos

Durante esta fase, se realizó una serie de sentencias e instrucciones para depurar las operaciones electrónicas de las transacciones que no son afectas al fraude y se seleccionó una muestra de 4,980 operaciones del universo total de operaciones registradas.

De la tabla de LOG de Transacciones actual se tomará solo los campos requeridos para la solución, cuya estructura inicial es la siguiente:

Variable	Descripción	Tipo
CODTRANS	Código de la transacción realizada.	Alfanumérico(2)
CTACARGO	Número de cuenta cargo	Alfanumérico(20)
CTAABONO	Número de cuenta Abono	Alfanumérico(20)
MONEDA	Moneda de la operación	Alfanumérico(3)
IMPORTE	Importe de la operación	Numérico (15,2)
CANAL	Canal en que se realizo la operación	Alfanumérico(4)
TARJETA	Número de tarjeta de acceso al canal	Alfanumérico(16)
FECHAOPER	Fecha de operación. Formato AAAA-MM-DD	Alfanumérico(10)
HORAOPER	Hora de operación. Formato hh.mm.ss	Alfanumérico(8)

Tabla 3. Campos y Atributos de Log de Transacciones. Fuente: (Elaboración propia)

La funcionalidad de esta fuente de datos consiste en servir como respaldo de las transacciones electrónicas, que pueden ser usadas para consultas y análisis en sistemas de gestión gerencial y de soporte a los sistemas de gestión de reclamos.

4.1.3 Transformación de los Datos

En esta fase se calculó y consolidó los valores de los nuevos atributos sugeridos por los criterios del experto, para ello se aplicó una serie de operaciones sobre los atributos originales.

Se elimino las variables o atributos como TARJETA, CTACARGO, CTAABONO, por representar campos con información personal y fueron reemplazados por valores calculados a partir de la aplicación de reglas de asociación y clasificación.

Los campos importe y Moneda para su mejor interpretación se trasladaron a Nuevos Soles, los montos de las operaciones que correspondían a otras monedas diferentes a la moneda nacional, multiplicando por su correspondiente tipo de cambio. Luego el importe se traslado a una variable nominal que solo indicara si el monto excede o no el limite diario por defecto.

Para finalizar esta fase, se procedió a generar el archivo de datos, mediante la selección de los atributos más relevantes, y registros de información más determinantes según el conocimiento del experto en el área bancaria.

Las variables NroOperDia, MayLimDia, MayPromTrans, IndMayPromTot y ImpAcumMayLim, son variables calculadas para lo cual se aplicaron una serie de instrucciones sobre el archivo base.
Al atributo “Sospechoso” se le asignó los posibles ‘S’ o ‘N’, basándose en el registro de las operaciones clasificadas como fraudulentas, proporcionado por el experto de seguridad.

Tabla 4. Estructura de la tabla LOGTRANS luego de la fase de Transformación de Datos.

Variable	Descripción	Tipo
Transacción	Código identificador de la operación/transacción realizada.	Alfanumérico(2)
NroOperDia	Numero de operaciones en el día. Valores S/N.	Numérico(3)
MayLimDia	Indica si la transacción es mayor al límite por defecto diario. Valores S/N.	Alfanumérico(1)
ImpAcumMayLim	Indica si la suma de importes de todas las operaciones del cliente ha superado el límite por defecto diario. Valores S/N.	Alfanumérico(1)
BenefAnt	Indica si el cliente realiza más de una operación al mismo beneficiario en el día. Valores S/N.	Alfanumérico(1)
Sospechoso	Indica si la transacción es sospechosa de fraude o No. Valores S/N.	Alfanumérico(1)

Tabla 4. Transformación de Datos. Fuente: (Elaboración propia)

4.1.4 Minería de Datos

Esta es la etapa del proceso de KDD, en la cual se genera el modelo de conocimiento, aplicando la técnica de minería de datos seleccionada.

Para el caso de detección de fraudes se eligió las técnicas de minería predictiva, esto en base a la comparación de todas las técnicas explicadas en los capítulos anteriores y análisis de otros trabajos desarrollados en cuanto a la detección de fraudes.

Por lo que se plantea generar un modelo de conocimiento que permite predecir ciertos comportamientos ante la ocurrencia de nuevas situaciones.

La técnica aplicada en esta etapa fue el de Arboles de Clasificación.

- **Algoritmo de árboles de decisión - Algoritmo C4.5**

  Este modelo de predicción tiene la función es representar y categorizar una serie de condiciones que ocurren de forma sucesiva para la resolución de un problema. Se aplico el algoritmo de arboles de clasificación (Decision tree) de RapidMiner, que es una implementación del algoritmo C4.5, uno de los más populares de minería de datos.

4.2 IMPLEMENTACIÓN SEGÚN EL CASO DE ESTUDIO

El dataset considerado para el caso de estudio es una muestra de 4980 operaciones reales, realizadas por los usuarios de los canales de la Banca por internet y la Banca móvil de personas naturales.

Se consideraron 7 tipos de operaciones, según la lista de transacciones afectas al fraude proporcionadas por el experto del negocio:

- Envío de Efectivo Móvil
- Transferencia a cuentas de terceros
- Pago tarjeta de crédito de terceros
- Transferencias a cuentas de otros bancos
- Pago de tarjeta de crédito de otro banco
- Transferencias el exterior
- Transferencias nacionales

ExampleSet (4980 examples, 1 special attribute, 6 regular attributes)

Row No.	Sospechosa	Funcion	Canal	NroOperDia	BenefAnt	MayLimDia	ImpAcumM...
1	N	T2	I	1	1	N	N
2	N	T2	I	2	1	N	N
3	N	T1	I	2	1	N	N
4	N	T2	I	1	1	N	N
5	N	T2	I	1	1	N	N
6	N	T2	I	1	1	N	N
7	N	T2	I	1	1	N	N
8	N	T2	I	2	2	N	N
9	N	T2	I	2	2	N	N
10	N	T2	I	1	1	N	N
11	N	T2	I	1	1	N	N
12	N	T2	I	1	1	N	N
13	N	T2	I	1	1	N	N
14	N	T2	I	1	1	N	N
15	N	T2	I	1	1	N	N
16	N	T2	I	1	1	N	N
17	N	T2	M	1	1	N	N
18	N	T2	I	2	1	N	N
19	N	T2	I	2	1	N	N
20	N	T2	I	1	1	N	N

Figura 3. Datos de operaciones Fuente: (Elaboración propia).

label								
Sospechosa	Binominal	0	Least S (75)		Most N (4905)		Values N (4905), S (75)	
Funcion	Polynominal	0	Least T7 (2)		Most T2 (3844)		Values T2 (3844), T4 (448), ...[5 more]	
Canal	Binominal	0	Least M (2089)		Most I (2891)		Values I (2891), M (2089)	
NroOperDia	Integer	0	Min 1	Max 29	Average 1.714		Deviation 2.594	
BenefAnt	Integer	0	Min 1	Max 5	Average 1.098		Deviation 0.398	
MayLimDia	Binominal	0	Least S (135)		Most N (4845)		Values N (4845), S (135)	
ImpAcumMayLim	Binominal	0	Least S (357)		Most N (4623)		Values N (4623), S (357)	

Figura 4. Estadísticas de los Datos. Fuente: (Elaboración propia)

4.2.1 Modelado

Se aplico el algoritmo DecisionTree en Rapidminer para desarrollar un modelo predictivo que identifique los atributos que mejor explican la clase "Sospechoso" para el DataSet de muestra.

El Árbol de Decisión se creó usando el operador Decision Tree basado en el algoritmo C4.5 de Quinlan.

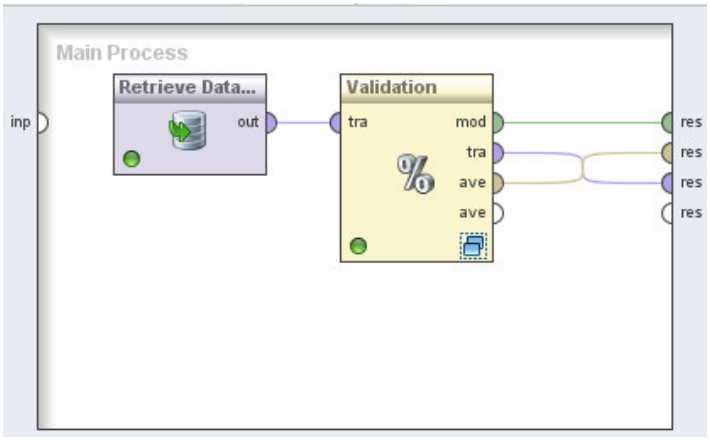

Figura 5. Modelado de Algoritmo Decision Tree C4.5- Proceso principal.
Fuente: (Elaboración propia)

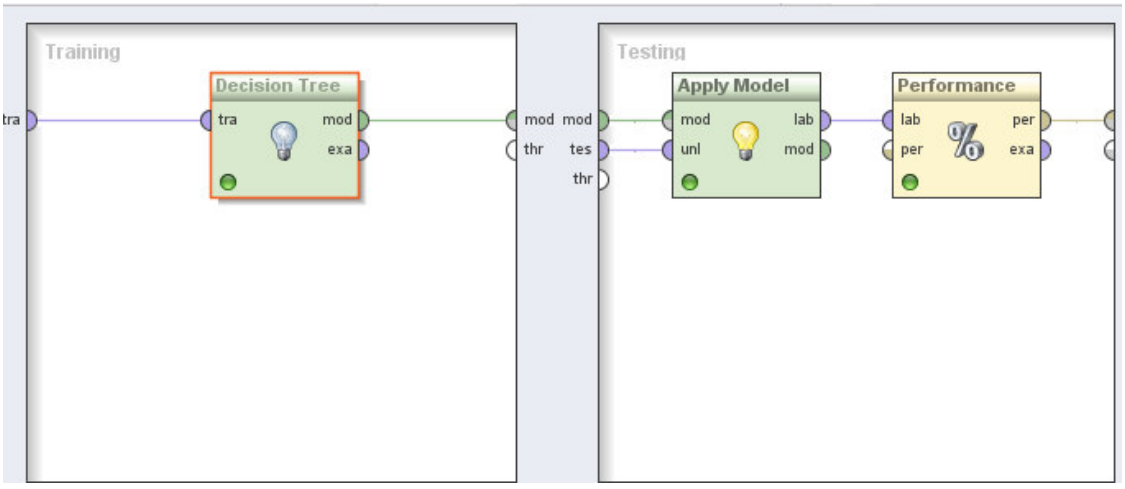

Figura 6. Modelado de Algoritmo Decision Tree C4.5- Subproceso-Validation.
Fuente: (Elaboración propia)

Parámetros del Modelo:

- Criterion: especifica el criterio de selección de atributos y de divisiones numéricas (ganancia de información, índice gini, precisión, proporción de ganancia).
- Minimal Size for Split: tamaño mínimo de divisiones que se pueden dar en cada nodo.
- Minimal leaf size: tamaño mínimo de la hoja.
- Minimal gain: la ganancia mínima que debe lograrse con el fin de producir una división.
- Maximal depth: la profundidad máxima del árbol.
- Number of prepruning: el número de nodos alternativos probados cuando la técnica de la poda evitaría una división.
- No preprunig: las reglas de poda se aplican luego de cada iteración.
- Prenuning: las reglas de poda basada en el criterio correspondiente después de generar el árbol.

Determinar los valores óptimos de este modelo no es una labor fácil de realizar, por lo que después de iterar una serie de valores se llego a los siguientes valores.

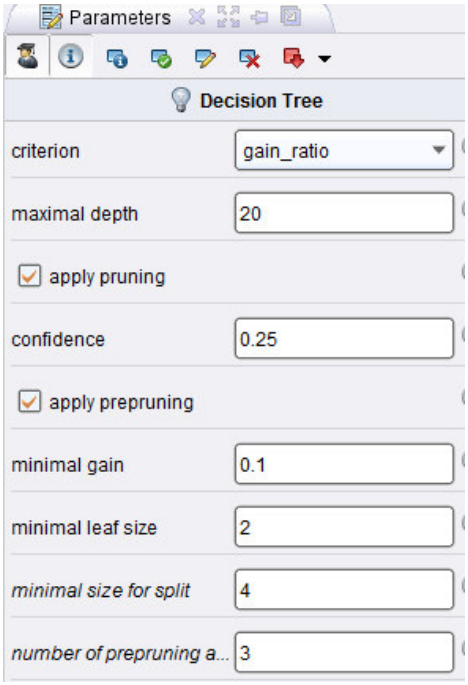

Figura 7. Parámetros Decision Tree. Fuente: (Elaboración propia)

Árbol de Decisión generado:

Luego de correr el modelo se tiene el siguiente árbol resultado para el cálculo de la clase "Sospechoso":

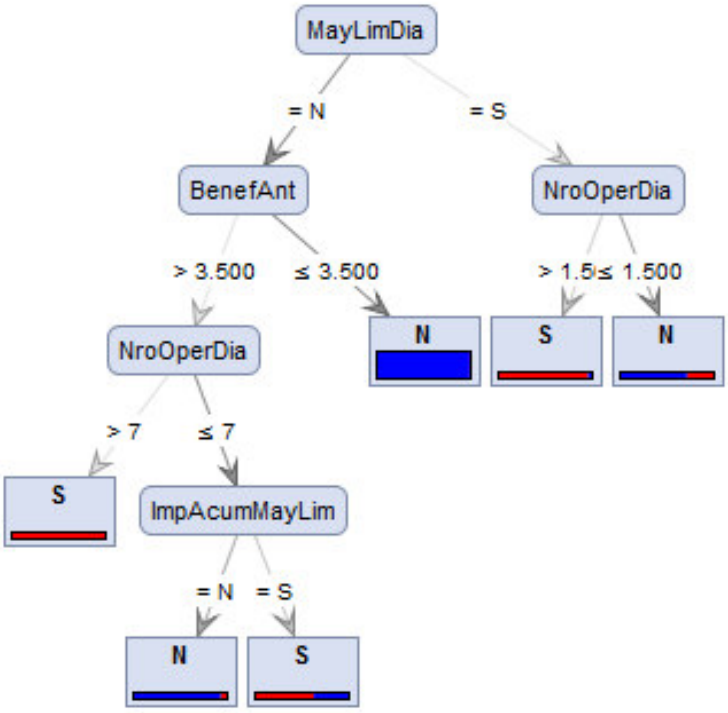

Figura 8. Árbol de Decisión del modelo. Fuente: (Elaboración propia)

Algoritmo Resultado:

```
MayLimDia = N
|   BenefAnt > 3.500
|   |   NroOperDia > 7: S {N=0, S=2}
|   |   NroOperDia ≤ 7
|   |   |   ImpAcumMayLim = N: N {N=17, S=1}
|   |   |   ImpAcumMayLim = S: S {S=7, N=4}
|   BenefAnt ≤ 3.500: N {N=4806, S=8}
MayLimDia = S
|   NroOperDia > 1.500: S {S=28, N=1}
|   NroOperDia ≤ 1.500: N {N=77, S=29}
```

4.2.2 Resultados Obtenidos

Luego de ejecutar la definición del proceso se obtiene los siguientes resultados: se crearon 3 atributos especiales adicionales:

- prediction respuesta del modelo
- confidence_Yes Probabilidad de Yes
- confidence_No Probabilidad de No

Estos atributos presentan los valores que se muestran en la siguiente tabla:

ExampleSet (4980 examples, 4 special attributes, 6 regular attributes) Filter (4,980 / 4,980 exam

Row No.	Sospechosa	prediction(S...	confidence(...	confidence(...	Funcion	Canal	NroOperDia	BenefAnt	MayLimDia	ImpAcumM
1	N	N	0.998	0.002	T2	I	1	1	N	N
2	N	N	0.998	0.002	T2	I	2	1	N	N
3	N	N	0.998	0.002	T1	I	2	1	N	N
4	N	N	0.998	0.002	T2	I	1	1	N	N
5	N	N	0.998	0.002	T2	I	1	1	N	N
6	N	N	0.998	0.002	T2	I	1	1	N	N
7	N	N	0.998	0.002	T2	I	1	1	N	N
8	N	N	0.998	0.002	T2	I	2	2	N	N
9	N	N	0.998	0.002	T2	I	2	2	N	N
10	N	N	0.998	0.002	T2	I	1	1	N	N

Figura 9. Indicadores de confianza y predicción. Fuente: (Elaboración propia)

Se debe observar que la suma de las confianzas es 1.0 y que la predicción depende de la confianza.

Rendimiento del modelo

Para calcular el desempeño del modelo se aplicó la medida de evaluación de la Matriz de confusión:

Matriz de confusión

La tabla 1 representa las posiciones que comúnmente muestra una matriz de confusión:

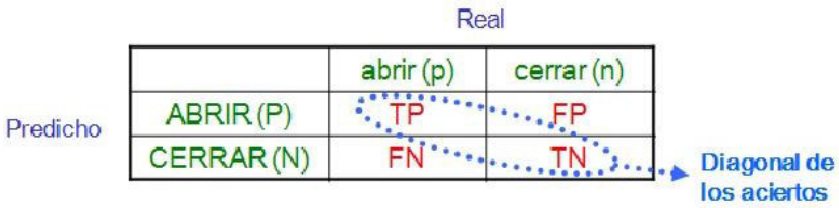

Figura 10. Matriz de Confusión ejemplo.

- TP: son los casos que pertenecen a la clase y el clasificador los definió en esa clase.
- FN: son los casos que si pertenecen a la clase y el clasificador no los definió en esa clase.
- FP: son los casos que no pertenecen a la clase pero el clasificador los definió en esa clase.
- TN: son los casos que no pertenecen a la clase y el clasificador definió que no pertenecen a esa clase.
- Accuracy: es la proporción del número total de predicciones que son correctas, se determina utilizando la siguiente ecuación:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Matriz de confusión Resultado del caso de estudio

Table View Plot View

accuracy: 99.02% +/- 0.34% (mikro: 99.02%)

	true N	true S	class precision
pred. N	4900	44	99.11%
pred. S	5	31	86.11%
class recall	99.90%	41.33%	

Figura 11. Matriz de Confusión resultado. Fuente: (Elaboración propia)

El campo accuracy indica que el 99.02% de registros de la muestra fueron clasificados correctamente según las reglas definidas como resultado del algoritmo.

CAPITULO V. CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- Se analizaron las principales técnicas de minería de Datos y resultado del estudio y comparación de éstas, se concluye que las técnicas predictivas resultan eficientes para descubrir conocimientos y permiten inferir como una variable o atributo puede incidir en otros.

- Se analizó las características de las diferentes transacciones financieras del caso de estudio, y con la utilización de técnicas de minería de datos fue posible identificar los atributos y reglas principales que permiten determinar un comportamiento sospechoso de fraude y un comportamiento habitual de un cliente.

- Se obtuvo un modelo de aprendizaje que permite clasificar las transacciones financieras de la Banca por Internet y la Banca Móvil como fraudulentas o íntegras con un 99.02% de exactitud, resultado de la aplicación de las diferentes reglas que conforman el Árbol de Decisión el cual fue resultado del análisis de la relación de los atributos más destacados de la transacción financiera. Con esto se puede concluir que la técnica predictiva de Árboles de Decisión a pesar de ser una técnica sencilla proporciona un alto grado de exactitud para identificar fraudes.

5.2 RECOMENDACIONES

- Se recomienda combinar o aplicar métodos de minería de datos que puedan complementarse y ayuden a mejorar el desempeño y así obtener mejores resultados.

- El modelo resultante de este trabajo de investigación se debe complementar creando una interfaz de comunicación entre los sistemas de la entidad bancaria y el modelo, para que cada nueva transacción que se realice en los canales de banca por internet y banca móvil, se validen contra este modelo y se pueda emitir o clasificar una transacción como fraudulenta o integra en línea.

- Debido a que se evidencia que solo un porcentaje muy bajo de las transacciones resultan sospechosas, se puede concluir que aplicar modelos de minería de datos a la información para la detección de Fraude, podrían obtener como resultado que todas las transacciones son íntegras o tener una baja probabilidad de ser sospechosas.

- Debido a que las modalidades de fraude están cambiando constantemente, es importante que las reglas y atributos definidos para el modelo resultado de la investigación, se vayan actualizando a la par, para lo cual es importante la participación de los expertos del negocio para la aplicación de nuevas reglas y atributos que deben considerarse.

REFERENCIAS BIBLIOGRÁFICAS.

[1] Ale 2005. Juan Ale, 2006. “Introducción a Data Mining”.

[2] Chen 1996. Chen, Ming Syan, Jiawei Han and Phillip Yu, 1996. “Data Mining: An overview from Database Perspective”.

[3] Fayyad 1996. Fayyad, G. Piatetsky-Shapiro and P. Smyth. 1996. “From data mining to Knowledge Discovery in Databases: an overview”.

[4] Han 2006. Jiawei Han, Micheline Kamber and Jian Pei, 2006. “Data Mining: Concepts and Techniques”.

[5] Lavado 2013. Luis Lavado Napaico. “Un algoritmo genético para la detección de fraude electrónico en tarjetas de debito en el Perú”.
Universidad Nacional Mayor de San Marcos. Facultad de Ingeniería de Sistemas e Informática.

[6] Martinez - 2012.Clemente Martinez. “Aplicación de técnicas de minería de datos para mejorar el proceso de control de gestión en ENTEL”.
Tesis para optar al grado de magíster en gestión de operaciones memoria para optar al título de ingeniero civil industrial. Universidad de Chile.

[7] Pang 2001. S. N. Pang, D.Kim and S.Y. Bang, 2001. “Fraud detection using support vector machine ensemble”.

[8] Perez 2005. M. Perez, 2005. “Fraud detection using support vector machine ensemble”.

[9] Quintlan 1990. J.R. Quintlan, 1990. “Induction of Decision Trees”.

[10] Quintlan 1993. J.R. Quintlan, 1993. “C4.5: programs for machine learning”.

[11] Robles 2003. V. Robles, 2003. “Clasificación Supervisada basada en Redes Bayesianas. Aplicación en Biología Computacional”.

[12] Russell y Norvig 2004. Russell, S., Norvig, P. “Inteligencia artificial: un enfoque moderno”.

[13] Santamaría 2010. Wilfredy SantaMaria Ruiz. “Modelo de Detección de fraude basado en el Descubrimiento de reglas de clasificación extraídas de una red neuronal”. Tesis de grado para optar por el titulo de Magister en Ingeniería de Sistemas y Computación. Universidad Nacional de Colombia.

[14] Tan 2005. Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2005. “Introduction to Data Mining”.

[15] Vallejos 2006. Sofía Vallejos, 2006. “Minería de Datos. Diseño y Administración de Datos”.

Sitios Web

[16] ASBANC 2015.
Portal WEB ASBANC. 2015. Asociación de Bancos del Perú. [Internet], [30 de Noviembre 2015].
Disponible en:
www.asbanc.com.pe/Informes%20de%20Prensa/TRANSACCIONES_POR_CANAL_Septiembre_2015.pdf

[17] ASBANC 2015.
Portal WEB ASBANC. 2015. Asociación de Bancos del Perú. [Internet], [En línea].
Disponible en:
http://www.asbanc.com.pe/Paginas/Servicios/Servicios.aspx?idservicio=5

[18] La Republica 2015.
Sitio Web del Diario La Republica. 2015. [Internet], [18 de Setiembre 2015].
Disponible en:
http://larepublica.pe/impresa/economia/704275-fraudes-en-sistema-bancario-suman-us-5-millones-en-lo-que-va-del-ano-dice-asbanc

[19] SBS 2013.
Portal WEB SBS. 2013. Superintendencia de Banca y Seguros. [Internet], [30 de Octubre 2013].
Disponible en:
https://intranet1.sbs.gob.pe/IDXALL/FINANCIERO/DOC/RESOLUCION/PDF/6523-2013.R.PDF

[20] SBS 2015.
Portal WEB SBS. 2015. Superintendencia de Banca y Seguros. [Internet], [En línea].
Disponible en:
http://www.sbs.gob.pe/usuarios/categoria/consejos-para-realizar-operaciones-financieras-por-internet/1431/c-1431