INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN
MAESTRÍA EN SISTEMAS COMPUTACIONALES

DETECCION DE COMPORTAMIENTO SOSPECHOSO CON APRENDIZAJE AUTOMATICO EN PERSONAL OPERATIVO BANCARIO

T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRO EN SISTEMAS COMPUTACIONALES

PRESENTA:
ALFONSO VELÁZQUEZ CAPULEÑO

DIRECTORES DE TESIS:
M.I. LUIS ARMANDO CÁRDENAS FLORIDO

LA PAZ, BAJA CALIFORNIA SUR, MÉXICO, JUNIO 2021.

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de La Paz
División de Estudios de Posgrado e Investigación

La Paz, B.C.S., 08/junio/2021

DEPI_MSC/031/2021

ASUNTO: Autorización de impresión

C. ALFONSO VELÁZQUEZ CAPULEÑO
ESTUDIANTE DE LA MAESTRÍA EN
SISTEMAS COMPUTACIONALES,
P R E S E N T E .

Con base en el dictamen de aprobación emitido por el Comité Tutorial de la Tesis denominada: **"DETECCIÓN DE COMPORTAMIENTO SOSPECHOSO CON APRENDIZAJE AUTOMÁTICO EN PERSONAL OPERATIVO BANCARIO"**, mediante la opción de tesis (Proyectos de Investigación), entregado por usted para su análisis, le informamos que se **AUTORIZA** la impresión

A T E N T A M E N T E
Excelencia en Educación Tecnológica®

JUAN PABLO MORALES ÁLVAREZ,
JEFE DE LA DIV. DE ESTUDIOS DE POSGRADO E INV.

INSTITUTO TECNOLÓGICO DE LA PAZ
DIVISIÓN DE ESTUDIOS DE POSGRADO
E INVESTIGACIÓN

c.c.p. Depto. de Servicios Escolares
c.c.p. Archivo.

*JPMA/icl**

Boulevard Forjadores de B.C.S. #4720,
Col. 8 de Octubre 1ra. Sección, C.P. 23080,
La Paz, B.C.S.
Tels. (612) 121-04-24,
email: depi_paz@tecnm.mx
tecnm.mx | lapaz.tecnm.mx

México 2021 Año de la Independencia

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

Instituto Tecnológico de La Paz
División de Estudios de Posgrado e Investigación

DICTAMEN DEL COMITÉ TUTORIAL

La Paz, B.C.S., **04/JUNIO/ 2021**

C. JUAN PABLO MORALES ALVAREZ,
JEFE DE LA DIVISIÓN DE ESTUDIOS DE POSGRADO E INVESTIGACIÓN,
P R E S E N T E.

Por medio del presente, enviamos a usted dictamen del Comité Tutorial de tesis para la obtención del grado de Maestro, con los siguientes datos generales:

No. de Control M18310007	Nombre ALFONSO VELÁZQUEZ CAPULEÑO
Maestría en:	SISTEMAS COMPUTACIONALES
Título de la tesis: DETECCIÓN DE COMPORTAMIENTO SOSPECHOSO CON APRENDIZAJE AUTOMÁTICO EN PERSONAL OPERATIVO BANCARIO	
DICTAMEN: Se autoriza el trabajo de investigación, en virtud de que realizó las correcciones correspondientes conforme a las observaciones planteadas por este Comité Tutorial.	

A t e n t a m e n t e.
El Comité Tutorial

DR. MARCO ANTONIO CASTRO LIERA

MSC. ILIANA CASTRO LIERA

MATI. LUIS ARMANDO CÁRDENAS FLORIDO

c.c.p. Coordinador de la Maestría.
c.c.p. Departamento de Servicios Escolares.
c.c.p. Estudiante.

ITLP-DEPI-RTT-08 ***Rev.1***

Boulevard Forjadores de B.C.S. #4720,
Col. 8 de Octubre 1ra. Sección, C.P. 23080,
La Paz, B.C.S.
Tels. (612) 121-04-24,
email: depi_paz@tecnm.mx
tecnm.mx | lapaz.tecnm.mx

México 2021 Año de la Independencia

CARTA CESIÓN DE DERECHOS

La presente se extiende en la Ciudad de La Paz, B.C.S. El día 10 del mes junio del año 2021, el que suscribe Alfonso Velázquez Capuleño estudiante del Programa de **Maestría en Sistemas Computacionales** con número de control M18310007, manifiesta que es autor (a) intelectual del presente trabajo de Tesis bajo la dirección de MATI. Luis Armando Cárdenas Florido y cede los derechos del trabajo intitulado DETECCIÓN DE COMPORTAMIENTO SOSPECHOSO CON APRENDIZAJE AUTOMÁTICO EN PERSONAL OPERATIVO BANCARIO, en forma NO EXCLUSIVA, al Tecnológico Nacional de México/Instituto Tecnológico de la Paz para su reproducción total o parcial en cualquier medio con fines académicos, científicos y culturales, así como para su publicación electrónica del texto completo para difusión y consulta.

Los usuarios de la información no deben reproducir el contenido textual, gráficas o datos del trabajo sin el permiso expreso del autor y/o director del trabajo. Este puede ser obtenido escribiendo a la siguiente dirección Boulevard Forjadores de B.C.S. #4720, Col. 8 de Octubre 1ra. Sección C.P. 23080, La Paz, B.C.S. Si el permiso se otorga, el usuario deberá dar el agradecimiento correspondiente y citar la fuente del mismo.

Alfonso Velázquez Capuleño

Dedicatoria

Dedico este esfuerzo a todos aquellos que de una manera u otra han contribuido a que este documento pueda salir a la luz. No quiero ser injusto con nadie porque todos, absolutamente todos, con los que he tenido una conversación a lo largo de mi vida me han alimentado para ser lo que soy.

Agradecimientos

Tienen un papel preponderante en mis agradecimientos mis padres que con sus métodos particulares y tan distintos me han hecho lo que soy como persona. Les doy gracias a ellos desde lo más profundo de mi corazón.

Resumen

Con los veloces avances en la tecnología se logran optimizar tiempos y costos en los sectores productivos. La tecnología está presente en nuestros centros de trabajo y en nuestros hogares casi por igual. El teléfono celular, siendo el medio de acceso más popularizado, nos permite tener a la mano dichos avances.

El problema identificado, objeto de esta investigación, es que no existen herramientas tecnológicas que permitan confirmar un comportamiento sospecho de algún colaborador, del sector financiero mexicano, con la antelación suficiente para evitar un fraude.

Las nuevas tecnologías y tendencias en el procesamiento de información, como lo es el aprendizaje automático, pueden ser aprovechadas para resolver algunos de los problemas que aquejan a las empresas del sector financiero; como es el caso del fraude interno.

Con el apoyo de datos históricos es posible aplicar algoritmos de Minería de datos con la finalidad de obtener patrones de comportamiento de los empleados bancarios que permitan generar alertas tempranas e ir persuadiendo a los empleados sospechosos de persistir en dichos comportamientos anómalos.

En este trabajo de investigación usamos la metodología de extracción de conocimiento KDD (*Knowledge Discovery in Databases*) para desarrollar una herramienta práctica y de uso universal por el tipo de datos que usa para resolver algunos de los patrones que siguen los empleados fraudulentos en el sector financiero mexicano.

A través de la metodología KDD, desde la selección de información y hasta la obtención de modelos y reglas se consigue obtener un modelo que, bajo ciertas circunstancias, ayuda a detectar un comportamiento sospechoso que puede terminar en un fraude. Es, por consiguiente, el primer paso hacia la automatización del análisis de comportamiento de la intención para cometer un ilícito.

Abstract

With the rapid advances in technology, times and costs are optimized in the productive sectors. Technology is present in our workplaces and in our homes almost equally. The cell phone, being the most popular mean of access, allows us to have these advances at hand.

The problem identified, which is the object of this investigation, is that there are no technological tools that warn us about suspicious behavior of a collaborator, from the Mexican financial sector, sufficiently in advance to avoid fraud.

New technologies and trends in information processing, such as machine learning, can be used to solve some of the problems that afflict companies in the financial sector; as is the case with internal fraud.

With the support of historical data, it is possible to apply data mining algorithms in order to obtain behavior patterns of bank employees to generate early alerts and persuading suspected employees to persist in such anomalous behaviors.

In this research work we use a KDD (Knowledge Discovery in Databases) knowledge extraction methodology to develop a practical tool to solve some of the patterns those fraudulent employees follow in the financial sector. Mexican.

Through the KDD methodology it is possible to obtain a model that, under certain circumstances, helps to detect suspicious behavior that may end in fraud. It is, therefore, the first step towards automating the behavior analysis of the intention to commit an offense.

Índice general

Índice de figuras

Índice de tablas

Índice de programas

Capítulo 1. Motivación y delimitación del problema

En este capítulo se abordará de manera específica la motivación que llevó a la selección y desarrollo del presente trabajo. El problema detectado está en el contexto financiero mexicano; un sector que en cierta forma es cerrado para el público en general. Y es cerrado porque así conviene mantenerlo pues genera inestabilidad para la vida normal de un país cuando los acontecimientos de fraudes en el sector bancario son expuestos ante la opinión pública.

Se detallan, además del problema localizado, el planteamiento de la hipótesis y el planteamiento de los objetivos para poder verificarla y validarla.

1.1 Introducción

En el ámbito productivo es muy importante tener control sobre las actividades y acciones de los colaboradores de una empresa. Los administradores de la empresa (es decir, los altos mandos y los mandos medio de la estructura organizacional) deben esforzarse por la productividad de sus subordinados. También deben buscar que sus colaboradores no usen sus habilidades y conocimientos para hacerle un daño a la empresa. Esta intención de control se intensifica cuando el empleado se desenvuelve en una empresa del sector financiero.

Una empresa (de cualquier tamaño y de cualquier género) está expuesta a dos clases de amenazas, las externas y las internas. Las amenazas externas están dadas por cualquier actor ajeno a la empresa que interactúe o no en las actividades que son el objeto principal (o razón de ser). Por su parte las amenazas internas están caracterizadas por todos los agentes (humanos y tecnológicos) que participan en las actividades desde dentro de la empresa.

Hoy en día existen muchas iniciativas tecnológicas para detectar amenazas externas, desde soluciones de hardware como de software, e incluso soluciones híbridas. Estas iniciativas pueden ser accesibles (en tecnología y/o presupuesto) para su implementación. Se debe poner énfasis en que no es el mismo caso para las amenazas internas.

Hay una razón muy simple para que la oferta de soluciones de detección de amenazas internas sea tan reducida: cada empresa es un objeto único. Su unicidad la determina la combinación de su cultura empresarial, sus procesos operativos, su grado de madurez, el sector en el que se desenvuelve, el presupuesto disponible y la gente que forma parte de la empresa (dentro de las principales características). Es de esperarse que con estas variables (cada una con grados de complejidad altos) se tengan muchas combinaciones posibles que se traducen en escenarios (o casos) sobre los cuales se debe tener atención.

La variable más complicada para intentar hacer una predicción de comportamiento es la relacionada con el comportamiento humano. En el sector financiero, probablemente más que en otros sectores, se vuelve imperioso detectar, con la mayor antelación posible, un comportamiento sospechoso en cualquiera de sus empleados.

Con la tecnología actual (hablando de la extracción de conocimiento de un conjunto de datos) se presume posible buscar patrones de comportamiento de los empleados bancarios. Integrar en un solo esfuerzo la información que generan los empleados, las capacidades de procesamiento de un equipo de cómputo y los algoritmos de aprendizaje automático permite pensar que es posible identificar los comportamientos sospechosos con cierta antelación. Esto es importante pues siempre será mejor actuar de manera proactiva para disuadir una conducta.

La tecnología necesaria para hacer este procesamiento de datos ya está al alcance la mano pues son varias las plataformas de software como los lenguajes de programación que permiten obtener el conocimiento necesario. Estas plataformas y lenguajes de programación están disponibles para arquitecturas de hardware que también están al alcance de prácticamente todos los presupuestos; esto último porque muchos de ellos se encuentran bajo la oleada del software libre.

Se requiere experiencia en la metodología de extracción de conocimiento, pero es muy importante tener conocimiento suficiente en el contexto del sector financiero para poder explotar con el mayor provecho posible los datos que se tengan a la mano. Las empresas financieras no

usan software “de caja”, generalmente tienen sus desarrollos propios y esto genera el problema de la no estandarización de los datos (ni en formato ni en detalle).

1.2 Contexto histórico social del objeto de estudio

Esta tesis se desarrolla en el ámbito productivo financiero mexicano. La empresa objeto de estudio tiene una empresa financiera con casi 20 años de operación. Los productos financieros que ofrece a sus clientes son el Crédito y la captación de Ahorros (a la vista y a plazo fijo). Se encuentra operando en nueve estados del noroeste del país.

Esta entidad financiera mexicana es regulada y vigilada por distintas entidades públicas. Dentro de las más importantes, y visibles, están la CNBV (Comisión Nacional Bancaria y de Valores), Banco de México y vigilada por la CONDUSEF (Comisión Nacional para la Protección y Defensa de los Usuarios de los Servicios Financieros). Por la oferta de productos financieros y el modelo de negocio con el que opera pueden existir algunos otros actores, algunos de ellos son de alcance regional por la ubicación geográfica en la que se desempeña.

Se ha utilizado información, con la autorización correspondiente, de la operación de esta empresa para el desarrollo del presente trabajo.

1.3 Contexto histórico de la tecnología

La tecnología que se aprovecha en esta investigación es el Aprendizaje Automático (conocido como *Machine Learning* en la literatura tecnológica).

Para algunos, los orígenes del aprendizaje automático se remontan a 1950 cuando Alan Turing publica un artículo titulado Computación e Inteligencia, en donde plantea lo que ahora conocemos como la Prueba de Turing. Esta es una prueba de habilidad de una máquina de mostrar un comportamiento inteligente similar al de un humano. La misma no evalúa el

conocimiento de la máquina en cuanto a su capacidad de responder preguntas correctamente, solo se toma en cuenta la capacidad de ésta de generar respuestas similares a las que daría un humano. Derivado de que no plantea un mecanismo de aprendizaje, este hito en la historia, no es considerado por muchos como parte de la historia [1].

El evento que es mayormente aceptado como el inicio de la historia del aprendizaje automático es el ocurrido en 1952 cuando Arthur Samuel escribe el primer programa para computadora capaz de aprender. El software era simplemente un programa que jugaba a las damas y que podía aprender de sus errores partida tras partida.

Para 1957 Fran Rosenblatt diseña el Perceptrón, una red neuronal en hardware para reconocimiento de caracteres. El propósito era el de explicar y modelar las habilidades de reconocimiento de patrones de los sistemas visuales biológicos.

Pasan muchos años en donde no se registran avances importantes en la materia y es hasta 1979 cuando unos estudiantes de la Universidad de Stanford, diseñan un carro capaz de moverse autónomamente por una habitación evitando obstáculos.

Tan solo dos años después (en 1981) Gerald DeJong crea el concepto de Aprendizaje Basado en Experiencia, haciendo que un equipo de cómputo analice información de entrenamiento y cree una regla general que le permita descartar información no importante.

Es en 1985 cuando Terry Sejnowski inventa NetTalk, un software que aprende a pronunciar palabras de la misma manera que lo haría un niño.

Ya en los inicios de la década de los 90s del siglo pasado los científicos de la época empiezan a crear programas que analicen grandes cantidades de datos y saquen conclusiones, o aprendan, de los resultados. Es así que en 1996 el equipo de cómputo llamado Deep Blue de IBM vence una partida de ajedrez a Gary Kaspárov, campeón del mundo vigente, aunque al final Kaspárov ganó 3 partidas más, derrotando a Deep Blue. Para mayo de 1997 se vuelven a enfrentar, pero

esta vez con una nueva versión de computador llamado Deeper Blue, esta vez se jugaron 6 partidas siendo el vencedor el computador.

Después de ese gran avance, la historia del aprendizaje automático tiene que esperar hasta 2006 para registrar que Geoffrey Hinton presenta el concepto de Deep Learning o aprendizaje profundo. Con este concepto se explicaron los nuevos algoritmos que permiten que los equipos de cómputo distingan diversos objetos y textos tanto en imágenes como en videos.

Es en el año 2010 cuando el dispositivo Kinect (consola de videojuegos) de Microsoft es capaz de reconocer 20 características del cuerpo humano a una velocidad de 30 veces por segundo.

Los avances se suceden de manera inmediata y en 2011 el ordenador Watson de IBM vence a dos inteligentes concursantes en la tercera ronda del concurso estadounidense de preguntas y respuestas Jeopardy.

En 2012 se crea GoogleBrain por Jeff Dean de Google y Andrew Ng profesor de la Universidad de Stanford. El propósito de este proyecto fue de crear una red neuronal utilizando toda la capacidad de infraestructura de Google para detectar patrones en vídeos e imágenes.

Es en 2012 cuando los laboratorios Google X, ahora llamado solamente X, desarrollan un algoritmo de aprendizaje automático que puede navegar de forma autónoma por los videos de Youtube para identificar los videos que contienen gatos.

Es hasta 2014 que un programa de computadora logra convencer a más del 30% de los jueces que era genuinamente humano; esto como parte de la comprobación de la prueba de Turing (planteada en 1950). Se trata de un chatbot (robot programado para charlas online) que obedece al nombre de Eugene Goostman, el programa fue capaz de convencer al 33% de los jueces que participaron en la prueba de que estaban chateando con un niño ucraniano de 13 años.

Facebook, en el año 2014, desarrolla DeepFace, un algoritmo de software que puede reconocer individuos en fotos al mismo nivel que los humanos.

Y es en 2015 que se tienen muchos eventos importantes en materia de aprendizaje automático. Entre ellos Amazon lanza su propia plataforma de aprendizaje automático o Machine Learning. Por su parte Microsoft crea el kit de herramientas para el aprendizaje de máquinas distribuidas, que permite la distribución eficiente de problemas de aprendizaje automático en múltiples computadoras. También Google aporta a los avances entrenando un agente conversacional de inteligencia artificial, que no solo puede interactuar convincentemente con humanos como un servicio de soporte técnico, sino también discutir la moralidad, expresar opiniones y responder preguntas generales basadas en hechos. En el mismo año OpenAI es creada; esta es una compañía de investigación de inteligencia artificial sin fines de lucro que tiene como objetivo promover y desarrollar inteligencia artificial amigable de tal manera que beneficie a la humanidad en su conjunto (entre sus fundadores se encuentra Elon Musk, el mismo de Tesla y SpaceX). Debido a los grandes avances obtenidos en el área de Machine Learning e inteligencia artificial señalados en este año, más de 3,000 investigadores de estas áreas, respaldados por Stephen Hawking, Elon Musk y Steve Wozniak, firman una carta abierta advirtiendo del peligro de las armas autónomas que seleccionan y atacan objetivos sin intervención humana.

Ya en el 2016 el algoritmo de inteligencia artificial de Google vence a un jugador profesional en el juego de mesa chino Go, que es considerado el juego de mesa más complejo del mundo y es muchas veces más difícil que el ajedrez. El algoritmo desarrollado por Google DeepMind logró ganar cinco juegos de cinco en la competencia de Go.

Para el 2017 OpenAI entrena chat bots o agentes conversacionales, que inventan su propio lenguaje para cooperar y lograr su objetivo de manera efectiva. Poco después, Facebook también capacitó exitosamente a agentes para negociar e incluso mentir. En este mismo año, un algoritmo desarrollado también por OpenAI derrota a los mejores jugadores en partidos 1 contra 1 del juego en línea Dota 2.

Hoy en día, con esta investigación, se aprovecha el momento de madurez con el que cuenta el Machine Learning como una rama de la Inteligencia Artificial [1].

1.4 Antecedentes

Un comportamiento es una forma de actuar o proceder bajo unas determinadas condiciones. Un sospechoso suele ser una persona que inspira sospecha, es una persona que brinda fundamentos para hacer un mal juicio de su conducta o de sus acciones [2].

Cuando nos referimos a comportamiento sospechoso estamos haciendo alusión a una forma de proceder no correcta o adecuada. Con esto señalamos que desconfiamos de alguien por conjeturas fundadas en apariencias o indicios de una verdad documentada [3].

Para analizar y desvirtuar la sospecha sobre una persona, se deben analizar una serie de eventos o evidencias (variables) hasta llegar a un juicio certero; y con eso confirmar si ese comportamiento sospechoso en verdad era un comportamiento de culpabilidad. Es muy cierto que un comportamiento sospechoso de un empleado generalmente termina convirtiéndose en un fraude [4].

Desde la perspectiva de la investigación podemos mencionar que existen algunos trabajos técnicos publicados por científicos de todo el mundo que abordan técnicas de *Machine Learning* en favor de la detección de fraudes (o comportamientos sospechosos) dentro de los ámbitos informáticos.

Uno de los primeros esfuerzos es un análisis de tráfico en las redes de cómputo realizado en China (*Comprehensive analysis of network traffic data* - Yuantian Miao) donde describe el proceso de análisis para etiquetar mensajes de tráfico en una red de cómputo. El proceso comienza con un pre procesamiento, para definir un etiquetado previo de los mensajes en el tráfico de la red, que tiene como objetivo facilitar el análisis en tiempo real de los mensajes que se generan en la red de cómputo. Con esto, el análisis en tiempo real es más eficiente en términos de tiempo necesario para la clasificación de los paquetes que circulan por una red de cómputo. En dicha investigación se implementan seis distintos algoritmos: *Machine Learning Algorithms*

-Naïve Bayes, *Decision Tree*, *1-Nearest Neighbor*, *Random Forest*, *Support Vector Machine* y *H2O*. Los mejores algoritmos, en términos de eficacia al distinguir tráfico malicioso, fueron el *Random Forest* y *Nearest Neighbor*. Se señala en el reporte técnico que la segunda fase de este trabajo técnico es la de mejorar este proceso para hacer el etiquetamiento en tiempo real, que sin duda es una debilidad de esta técnica [5].

Un Segundo trabajo de investigación localizado es sobre la correlación de información en el tráfico de una red de cómputo (*Network Traffic Classification Using Correlation Information* - Jun Zhang). En este trabajo de investigación el método utilizado está basado en el algoritmo del vecino más cercano que, de acuerdo al artículo, mostró un rendimiento de clasificación superior. Aquí se propone un enfoque no paramétrico para la clasificación del tráfico, que puede mejorar el rendimiento de la clasificación de manera efectiva al incorporar información correlacionada en el proceso de clasificación. Se demuestra, con su método, que el rendimiento de la clasificación de tráfico puede mejorarse significativamente incluso con muy pocas muestras de entrenamiento (cosa que es la parte sobresaliente de este trabajo). La debilidad se enfoca en que sólo abarca tráfico del tipo TCP dejando fuera UDP; y consideran para una siguiente fase el abarcar este último protocolo [6].

En el ámbito nacional se tiene el trabajo de identificación de ataques en redes de cómputo utilizando redes neuronales artificiales (Javier Alberto Carmona Troyo). En este trabajo se analizó un tipo de ataque informático a una red de cómputo usando comando de NMAP y un comando ping simulando un ataque de denegación de servicio (DDoS). Esta técnica simula uno de los ataques más recurrentes. Se mencionan varios experimentos, con diferentes capas ocultas en la red neuronal, y se llega a la conclusión de que a partir de las siete capas ocultas los resultados son los mismos (o tienden a ser los mismos). El entrenamiento de la red es supervisado; y se llegó a un modelo a implementar, pero no queda claro, en las conclusiones, como puede ser aplicado para la detección de ataques On Line. Se identifica que el ataque DDoS es de los que mayores avances tiene y para los cuales existen muchas soluciones [7].

Estos trabajos se enfocan en amenazas externas, argumento que valida la necesidad de hacer investigaciones para amenazas internas.

1.5 Definición del problema científico

En la actualidad no se cuenta con trabajos científicos documentados en la línea de investigación correspondiente a confirmar un comportamiento sospechoso de empleados bancarios en productos financieros de ahorro (amenazas internas).

Si este tipo de investigación existe corresponde a trabajos particulares de las empresas del sector financiero y estarán enfocados a situaciones y realidades muy específicos; es decir que no son escalables o estandarizados para ser replicados en otras empresas del mismo sector.

En contraparte, si se cuenta con varios algoritmos de extracción de conocimiento para usar bajo diversas situaciones, que puede aumentar el éxito de un trabajo en este sentido.

1.6 Planteamiento de la hipótesis

Por todo lo expuesto anteriormente se pretende confirmar que, basado en una metodología de extracción de conocimiento, como la KDD, combinada con el uso de algoritmos de Minería de Datos es posible obtener los patrones de comportamiento de los empleados que presentan comportamiento sospechoso en un ámbito financiero cuando usan los sistemas informáticos de la empresa para atender a clientes que tienen cuentas de ahorro en México.

1.7 Objetivo general

Identificar fraudes bancarios informáticos a partir de patrones de comportamiento sospechoso descubiertos con técnicas de aprendizaje automático.

1.8 Objetivos específicos

- Identificar variables, dentro del contexto informático bancario, que determinan un comportamiento sospechoso de un empleado.
- Identificar patrones de comportamiento a partir de las variables identificadas y de los casos documentados de empleados fraudulentos sobre los datos históricos que se tengan disponibles.
- A partir de los patrones de comportamiento sospechoso identificados, detectar a un empleado (o grupo de empleados) con un perfil de empleado fraudulento lo más oportunamente posible.

Capítulo 2. Técnicas actuales de detección de fraude financiero

Para poder llegar a la descripción formal de la solución que se le ha dado al problema localizado y explicar los resultados obtenidos se requiere entender y repasar algunos temas desde el punto de vista teórico.

2.1 Marco teórico

En este capítulo se aborda un apartado del fraude y en específico del fraude en el sector financiero mexicano. Se requiere recopilar la teoría de la metodología de extracción de conocimiento KDD para justificar las fases de la selección y depuración de la información que ha sido utilizada en este trabajo de investigación. La Minería de Datos la debemos entender en su aspecto general para saber la manera en que está constituida y las herramientas que nos ofrece para analizar datos. Sobre este último aspecto se hace mención específica a dos algoritmos: Árboles de decisión y K-vecinos; estos son los algoritmos sobre los cuales descansan los resultados obtenidos.

Al inicio de este trabajo se habló de la definición de comportamiento sospechoso; mencionábamos que un comportamiento sospechoso deriva, en la mayor parte de las ocasiones, en un fraude [2].

Por lo anteriormente mencionado, en este trabajo desde la perspectiva subjetiva, se definirá el concepto formal (y a una profundidad suficiente) de fraude. Posteriormente se proporciona una breve descripción de la Metodología de Extracción de Conocimiento KDD y conceptos relacionados a la Minería de Datos. El marco teórico lo cierra la explicación de los algoritmos de Árboles de decisión y de Agrupación (k vecinos más cercanos) que sustentan este escrito.

2.2 Sobre el fraude

2.2.1 Concepto de fraude

Del latín *fraus*, un fraude es una acción que resulta contraria a la verdad y a la rectitud. El fraude se comete en perjuicio contra otra persona o contra una organización (como el Estado o una empresa).

El concepto de fraude está asociado al de estafa, que es un delito contra el patrimonio o la propiedad. Consiste en un engaño para obtener un bien patrimonial, haciendo creer a la persona o la empresa que paga que obtendrá algo que, en realidad, no existe. Algunas palabras consideradas sinónimos que están asociadas al fraude son: engaño, timo, falsificación, mentira y dolo entre otras más [2].

Con lo anterior queda muy claro que estamos abordando un tema no grato para alguna de las partes implicadas en algún tipo de relación laboral y que adicionalmente deriva en una acusación.

2.2.2 Concepto de fraude en el ramo financiero

Los fraudes financieros son acciones que una persona realiza con el fin de obtener un beneficio propio a costa de dañar la economía de otra. La mayoría de los defraudadores buscan conseguir los datos de los clientes para realizar una acción ilícita [2].

En un ámbito financiero hacer una transacción es común. Una transacción es un depósito, cobrar un cheque, retirar dinero de un cajero automático, solicitar un crédito, pagar con la tarjeta de crédito, realizar compras en línea y ahorrar, entre otras tantas. Desde la perspectiva de un cliente bancario se corre el riesgo de ser víctima de un fraude. Hoy en día, los cuentahabientes (o clientes bancarios) se apoyan en los medios informáticos que las instituciones financieras ponen a disposición para realizar sus transacciones.

En México, estos medios informáticos, van tomando fuerza con el paso del tiempo. Sin embargo, aún no están totalmente generalizados en la población bancarizada (que se estima está en un 30 a 40% del total de la población económicamente activa) [8].

Esto significa que las sucursales bancarias siguen siendo un elemento de atención de clientes muy fuerte y arraigado en el país. Aún existen muchas sucursales disponibles en México (poco más de 17 mil. En este canal de atención (la Sucursal bancaria) se requiere fuerza humana (empleados bancarios) para la atención de los clientes. Los empleados bancarios son los que ejecutan las transacciones en los sistemas informáticos que los clientes les solicitan.

El presente trabajo se enfoca, de manera muy particular, en este último aspecto pues el presente trabajo consiste en determinar el momento en que los empleados bancarios dejan de hacer solo su trabajo y rebasan los límites éticos para cometer fraude en detrimento de los clientes y de la empresa en la que laboran; incluso en detrimento de algunas otras instituciones financieras con las que se tiene contacto.

No importa que tan madura pueda estar la estrategia de controles antifraude que tenga implementada la empresa financiera; y tampoco las campañas de concientización que se lleven a cabo al interior de la empresa. Es muy común, por su parte, la fuga de información y el factor clave ahí es el empleado bancario [9].

2.3 Casos de fraude famosos

Existen infinidad de casos de fuga de información famosa a nivel internacional (no acabaríamos señalando todos los casos). Muchos de esos casos no están documentados de manera pública; sin embargo, de entre los casos más significativos y relevantes de fuga de información encontramos los siguientes:

- Renault fue víctima del espionaje industrial cuando sus altos directivos vendieron información confidencial sobre su vehículo eléctrico, investigación en la que habían invertido más de 4.000 millones de euros [10].
- 81 estaciones de gasolina de Miami y Florida robaron información de las tarjetas de crédito de sus clientes [11].
- La empresa CardSystems (procesador de operaciones de tarjetas de crédito) permitió el robo de información de las 200.000 tarjetas de crédito de usuarios que realizaron compras por Internet y en persona en los Estados Unidos [12].
- Bancolombia abrió un juicio de cárcel en contra de un cajero de una de sus Sucursales. El empleado aprovechando su puesto clonó tarjetas de débito que después vaciaba para hacer transferencias a su propia cuenta [13].
- Banco de México, en el 2018, sufrió un robo millonario de aproximadamente 300 millones de pesos al hacer transferencias a cuentas y retirarlos en efectivo en la red de sucursales de diferentes bancos nacionales [14].
- Para ejemplificar un fraude perpetrado por el personal interno podemos remontarnos al año 2014 cuando el banco FICREA fue intervenido por los malos manejos del personal interno de nivel directivo que llevó a la empresa a la quiebra y con ello a los ahorradores que estaban asociados con la empresa. Aproximadamente seis mil clientes fueron víctimas de esta situación. El modus operandi de ese fraude fue que las arcas de la empresa financiera generalmente estaban vacías porque los créditos que otorgaban los hacían a través de empresas aledañas al grupo de socios de la financiera. Dichos créditos no fueron bien otorgados y por consiguiente no se pudo recuperar el capital [15].

Debe considerarse que existe una cantidad muy grande (incuantificable y no pública) de fraudes en México. Estos fraudes a cuentahabientes son de cantidades menores que no llegan a ser famosos en la prensa.

2.4 Metodología de extracción de conocimiento

Se cuenta con un conjunto de datos que más adelante detallaremos; y se cuenta también con algoritmos que nos ayudarán a analizar dichos datos. Requerimos entonces, de una metodología que nos permita llevar la extracción del conocimiento como un proceso [16].

Esta metodología de extracción del conocimiento es la llamada KDD. Con esta metodología vamos a buscar hacer eficientes los esfuerzos de modelado y comprobación de la hipótesis.

El descubrimiento de conocimiento en bases de datos es básicamente un proceso automático en el que se combinan descubrimiento y análisis. El proceso consiste en extraer patrones en forma de reglas o funciones, a partir de los datos, para que el usuario los analice. Esta tarea implica generalmente realizar un procesamiento previo a los datos, hacer minería de datos y presentar resultados. La metodología KDD se puede aplicar en diferentes dominios, entre ellos el determinar perfiles o comportamientos sospechosos (con potencial de fraude) [17].

El procesamiento previo de los datos puede implicar la implementación de una Bodega de Datos que permita mantener organizados los datos a analizar.

En este proceso es necesario tener claro el problema, su contexto y los resultados que se piensan obtener. Aunque se trata, en muchas ocasiones, de descubrir conocimiento es necesario tener una idea de lo que se puede encontrar, sobre todo para que el alcance no se haga indefinido y se caiga en un proceso infinito.

2.4.1 Proceso KDD

El proceso KDD, que se muestra en la *Figura 1*, es interactivo e iterativo. Involucra numerosos pasos con la intervención del usuario en la toma de muchas decisiones.

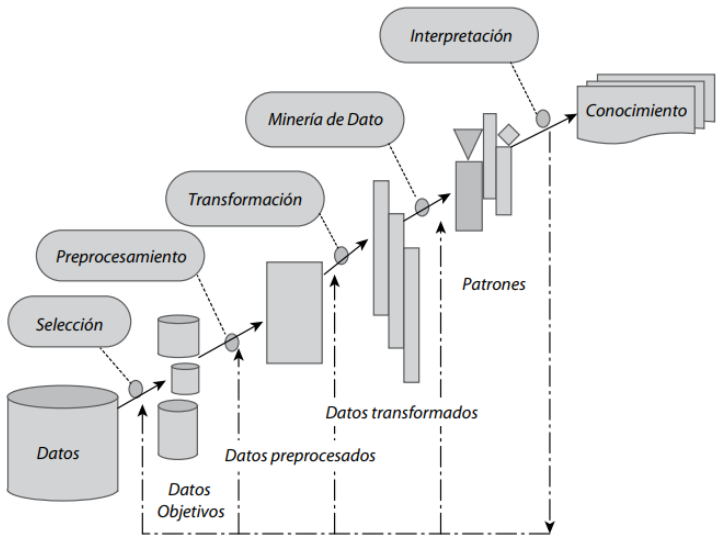

Figura 1. Fases del KDD

Se resume en las siguientes etapas:

- Selección.
- Preprocesamiento/limpieza.
- Transformación/reducción.
- Minería de datos.
- Interpretación/evaluación.

2.4.1.1 Etapa de selección

En la etapa de selección se debe identificar el conocimiento relevante y prioritario para definir las metas del proceso KDD. Desde el punto de vista del usuario final, se crea un conjunto de datos objetivo, seleccionando todo el conjunto de datos o una muestra representativa de este, sobre el cual se realiza el proceso de descubrimiento. La selección de los datos varía de acuerdo con los objetivos que se persiguen.

2.4.1.2 Etapa de procesamiento previo / limpieza

En la etapa de procesamiento previo / limpieza se analiza la calidad de los datos, se aplican operaciones básicas como la remoción de datos ruidosos, se seleccionan estrategias para el manejo de datos desconocidos, datos nulos, datos duplicados y técnicas estadísticas para su reemplazo. En esta etapa, es de suma importancia la interacción con algún experto en el área en la que se está trabajando (en caso de no contar con experiencia suficiente en el contexto del problema) [17].

Los datos ruidosos son valores que están significativamente fuera del rango de valores esperados; se deben principalmente a errores humanos, a cambios en el sistema, a información no disponible a tiempo y a fuentes heterogéneas de datos. Los datos desconocidos son aquellos a los cuales no les corresponde un valor en el mundo real y los datos perdidos son aquellos que tienen un valor que no fue capturado. Los datos nulos son datos desconocidos que son permitidos por los sistemas SGBDR (Sistemas Gestores de Bases de Datos Relacionales). En el proceso de limpieza todos estos valores se ignoran, se reemplazan por un valor por omisión, o por el valor más cercano, es decir, se usan métricas de tipo estadístico como media, moda, mínimo y máximo para reemplazarlos cuando son relevantes para el análisis.

2.4.1.3 Etapa de transformación / reducción

En la etapa de transformación / reducción de datos, se buscan características útiles para representar los datos dependiendo de la meta del proceso. Se utilizan métodos de reducción de dimensiones o de transformación para disminuir el número efectivo de variables bajo consideración o para encontrar representaciones invariantes de los datos [17].

Los métodos de reducción de dimensiones pueden simplificar una tabla de una base de datos horizontal o verticalmente. La reducción horizontal implica la eliminación de tuplas idénticas como producto de la sustitución del valor de un atributo por otro de alto nivel, en una jerarquía definida de valores categóricos o por la discretización de valores continuos (por ejemplo, edad

por un rango de edades). La reducción vertical implica la eliminación de atributos que son insignificantes o redundantes con respecto al problema, como la eliminación de llaves, la eliminación de columnas que dependen funcionalmente (por ejemplo, edad y fecha de nacimiento). Se utilizan técnicas de reducción como agregaciones, compresión de datos, histogramas, segmentación, discretización basada en entropía, muestreo, entre otras.

2.4.1.4 Etapa de Minería de Datos

El objetivo de la etapa Minería de Datos es la búsqueda y descubrimiento de patrones insospechados y de interés, aplicando tareas de descubrimiento como clasificación, *clustering*, patrones secuenciales y asociaciones entre otras [17].

Las técnicas de minería de datos crean modelos que son predictivos o descriptivos. Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, dependientes o clases, usando otras variables denominadas independientes o predictivas. Entre las tareas predictivas están la clasificación y la regresión. Los modelos descriptivos identifican patrones que explican o resumen los datos; sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos. Entre las tareas descriptivas se cuentan las reglas de asociación, los patrones secuenciales, los *clustering* y las correlaciones [17].

Por lo tanto, la selección de un algoritmo de Minería de Datos incluye la selección de los métodos por aplicar en la búsqueda de patrones en los datos, así como la decisión sobre los modelos y los parámetros más apropiados, dependiendo del tipo de datos (categóricos, numéricos) por utilizar.

2.4.1.5 Etapa de interpretación / evaluación de datos

En la etapa de interpretación / evaluación, se interpretan los patrones descubiertos y posiblemente se retorna a las anteriores etapas para posteriores iteraciones. Esta etapa puede

incluir la visualización de los patrones extraídos, la remoción de los patrones redundantes o irrelevantes y la traducción de los patrones útiles en términos que sean entendibles para el usuario. Por otra parte, se consolida el conocimiento descubierto para incorporarlo en otro sistema para posteriores acciones o, simplemente, para documentarlo y reportarlo a las partes interesadas; también para verificar y resolver conflictos potenciales con el conocimiento previamente descubierto [17].

2.5 Bodega de datos

En el apartado de la Metodología de Extracción de Conocimiento se abordó la posibilidad de necesidad / conveniencia sobre el uso de una Base de Datos que nos ayudara a consolidar la información que vamos recabando para el análisis de los datos y partir de ahí para generar un conocimiento. Esta Base de Datos es conocida como la Bodega de Datos (o *Dataware House* en la literatura técnica del tema) [18].

Una Bodega de Datos tiene capacidades superiores de almacenamiento y procesamiento de datos que una Base de Datos convencional.

Se hace un paréntesis para señalar que en este texto distinguiremos el término dato de información. Entendiendo que un dato solo es la representación de alguna cosa o concepto como puede ser el nombre de una persona o una temperatura leída desde algún termómetro (sin importar, por el momento la forma de la lectura de este dato); y por el otro lado nos referiremos a información cuando un dato o un conjunto de datos han sido interpretados y han generado un conocimiento de alguna índole sobre un contexto en común a los datos [19].

Ahora bien, los datos generados son almacenados en bases de datos, al igual que en dispositivos de distintos tipos (celulares, aparatos médicos, dispositivos de la industria, etc) y también en discos desconectados de la internet (como USB, CD, DVD, etc). Lo anterior puede redituar en un enorme volumen de datos que podrían, incluso, exceder la capacidad natural de análisis del ser humano para poder sacarle beneficios directos a los mismos. Sin embargo,

sabemos que estos datos representan un gran valor por lo que tienen “oculto” y que podemos explotar con la ayuda de la misma tecnología.

Este concepto (la Bodega de datos) es fundamental para dar pie al concepto de Minería de Datos. Una definición que detalla, con una buena precisión, lo que significa este término, es “Una bodega de datos es una herramienta empresarial utilizada como una solución informática que consolida los datos de diferentes fuentes de una entidad, institución o negocio ya sean de bases de datos, archivos planos o de otros sistemas del negocio, extrayéndolos, transformarlos en el caso de ser necesario y almacenarlos en un solo repositorio...” [17].

Si bien la empresa objeto de estudio está en el ámbito empresarial el concepto de Bodega de Datos tiene un objetivo de investigación para este trabajo.

Para alimentar la Bodega de Datos se utiliza la técnica del ETL (*Extraction – Transformation - Load*) que es un conjunto de procesos sistematizados y organizados para poder Extraer, Transformar y Cargar los datos en la Bodega de Datos. Esta técnica en realidad es un procesamiento de datos y como tal requiere que se haga todo un análisis, diseño e implementación del mismo. “La tarea de un diseñador de procesos de ETL involucra: (1) analizar las fuentes de datos existentes para encontrar la semántica oculta en ellas y (2) diseñar el flujo de trabajo que extraiga los datos desde las fuentes, repare sus inconsistencias, los transforme en un formato deseado, y, finalmente, los inserte en la bodega de datos…”. Esta actividad puede ser tan compleja como la cantidad de fuentes de datos y su integridad. Podríamos, inclusive, mencionar que se debe hacer un análisis específico para cada problema a resolver [17].

En la *Figura 2* se pueden identificar gráficamente los conjuntos de procesos sistematizados y organizados de los que hemos hablado.

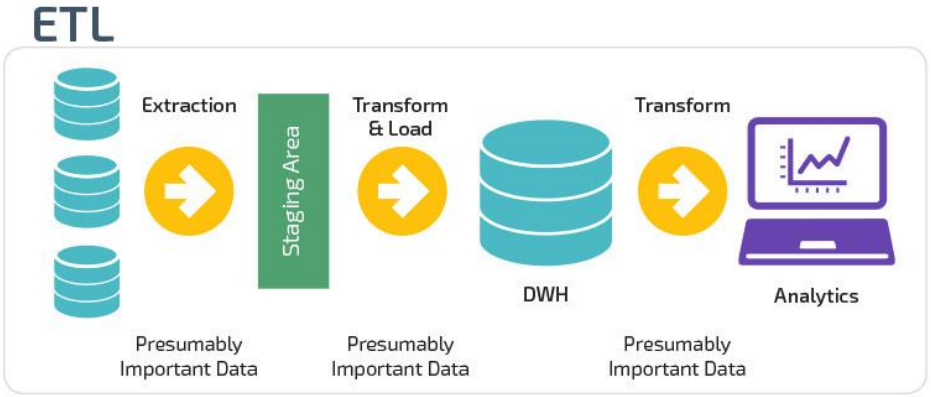

Figura 2. Diagrama de proceso ETL

El procesamiento ETL sirve para darle a la Bodega de datos sus cuatro características fundamentales, que son: Integrado, Temático, de Tiempo Variante y No volátil. A manera de resumen podemos comentar lo siguiente para cada uno de ellos [17]:

- Integrado, significa las inconsistencias que en algunas ocasiones contienen los sistemas operacionales que alimentan la Bodega de datos; estas deben ser eliminadas para formar una estructura consistente.
- Temático, significa que los datos deben ser organizados por temas que estén dentro del contexto de la generación de conocimiento que se tenga como objetivo. Con esto también se debe señalar que algunos datos de los sistemas operacionales podrían quedar fuera del procesamiento ETL si no forman parte del contexto del objetivo a cumplir.
- De Tiempo variante, que significa que en la Bodega de datos podemos analizar tendencias. En otras palabras, esto significa que en la Bodega de datos se puede analizar la situación en diversos momentos del tiempo y en el sistema operacional solo podemos ver la situación del momento actual.
- No volátil, que significa que la Bodega de datos está hecha para poder ingresar datos y hacer consultas de estos; no está permitido hacer actualizaciones de los datos una vez almacenados.

Las Bodegas de datos, como se puede intuir, son un eslabón esencial en todo lo que se refiere a la generación de conocimiento.

2.6 Minería de Datos

Como ya lo mencionábamos en apartados anteriores la Minería de Datos y la Metodología de Extracción de Conocimiento suelen confundirse y tratarse por igual. Si bien es cierto que algunos de sus términos o elementos llegan a ser muy semejantes no podemos olvidar que uno es una metodología y el otro son un conjunto de técnicas, aunque en ambos casos persiguen el mismo objetivo [17].

La Minería de Datos tiene como objetivo analizar los datos para extraer conocimiento. Este conocimiento puede ser en forma de relaciones, patrones o reglas inferidas de los datos, o bien en forma de una descripción más concisa (resumen). Estas relaciones o resúmenes constituyen el modelo de los datos analizados. Existen muchas formas diferentes de representar los modelos y cada una de ellas determina el tipo de técnica que puede usarse para inferirlos.

Las técnicas de minería de datos utilizan métodos para tratar la alta dimensionalidad de los datos en conjunto con algoritmos pertenecientes al ámbito de la inteligencia artificial, así como métodos matemáticos y estadísticos que juntos permiten poder realizar búsquedas de patrones, secuencias o comportamientos sistemáticos que pongan de manifiesto interrelaciones entre los datos o que sirvan para predecir comportamientos futuros. Estas técnicas son muy variadas, pues no todas son aplicables en cualquier conjunto de datos ni a todo procedimiento de extracción de información. En general, cualquiera que sea el problema para resolver, podemos decir que no existe una única técnica para solucionarlo y posiblemente el abanico de técnicas que comprende el campo de la minería de datos, nos permita hacer visible diferentes realidades de nuestro conjunto de datos [17].

En la práctica, cuando aplicamos los algoritmos de la Minería de datos, se pueden generar dos tipos de modelos: los predictivos y los descriptivos [19].

Los modelos predictivos pretenden estimar valores futuros o desconocidos de variables de interés, que se denominan variables objetivo, usando otras variables de la base de datos, a las que se conocen como variables independientes. Por su parte los modelos descriptivos identifican patrones que explican o resumen los datos, es decir, sirven para explorar las propiedades de los datos examinados, no para predecir nuevos datos.

2.6.1 Taxonomía de técnicas de Minería de datos

Se define como taxonomía a la estructura organizacional o clasificación que se tiene de las diferentes técnicas y algoritmos dentro de la Minería de datos. Lo anterior se muestra a detalle en la *Figura 3*:

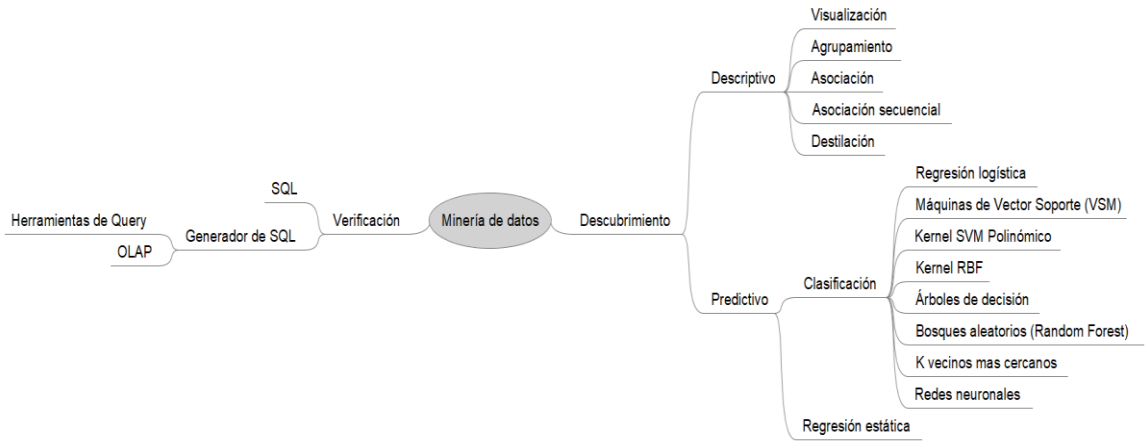

Figura 3. Taxonomía de técnicas de Minería de datos

Se cuenta con dos grandes tipos de descubrimiento en la Minería de Datos: la verificación y el descubrimiento. En este último rubro se encuentran las clasificaciones descriptivas y predictivas. Nos enfocaremos, en este trabajo, a la parte de la clasificación ya que tenemos una clase (fraudulento) y dos posibles valores en cada uno de ellos.

2.6.2 Descubrimiento predictivo: Clasificación

La clasificación es una subcategoría del aprendizaje supervisado en la que el objetivo es predecir las etiquetas de clases categóricas (discreta, valores no ordenados, pertenencia a grupo) de las nuevas instancias, basándonos en observaciones pasadas [17].

Hay dos tipos principales de clasificaciones:

- Clasificación Binaria: Es un tipo de clasificación en el que tan solo se pueden asignar dos clases diferentes (por ejemplo, cero o uno).
- Clasificación Multiclase: En este tipo de clasificación se pueden asignar múltiples categorías a las observaciones.

El siguiente ejemplo (que se muestra en la *Figura 4*) es altamente representativo de una clasificación binaria. Tenemos dos clases: círculos y cruces; y dos características: X1 y X2. El modelo es capaz de encontrar la relación entre las características de cada punto de datos y su clase, y de establecer una línea de separación entre ellos, de forma que cuando se le facilitan nuevos datos, puede estimar la clase a la que pertenece dadas sus características.

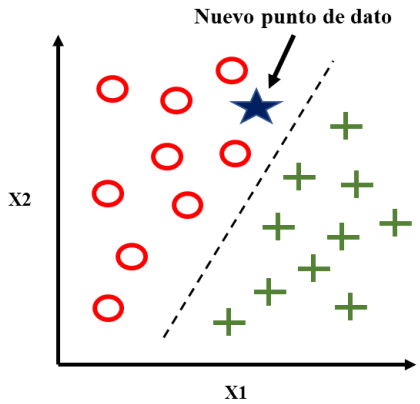

Figura 4. Ejemplo clásico de Clasificación

En este caso, el nuevo punto de dato cae en el subespacio de círculos y, por consiguiente, el modelo predecirá que su clase es un círculo.

Aquí es importante indicar que no todos los modelos de clasificación serán útiles para separar adecuadamente las diferentes clases de un conjunto de datos. Algunos algoritmos no convergerán al aprender los pesos del modelo si las clases no pueden separarse por una frontera de decisión lineal.

Algunos de los casos más típicos se representan en las siguientes figuras que forman parte de la *Figura 5*:

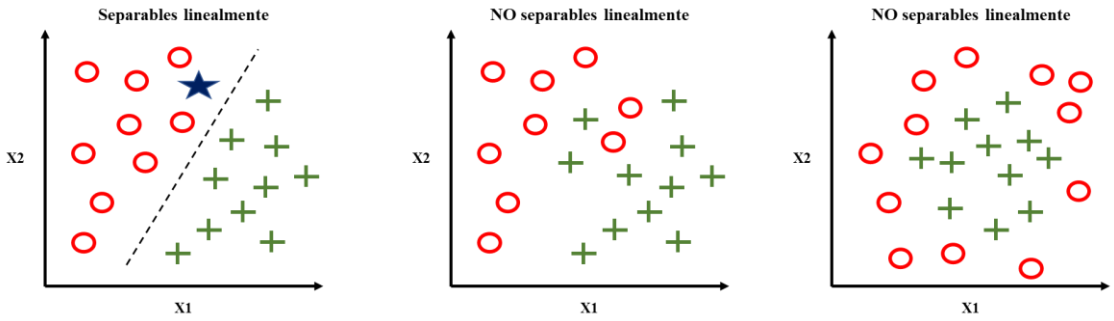

Figura 5. Diferentes casos de Clasificación

Esto último es lo que origina que existan diversos algoritmos y que sea importante conocerlos a detalle pues su uso depende en gran medida del tipo de problema que necesitemos abordar.

2.6.2.1 Algoritmos de Árbol de Decisión

Los Árboles de decisión son modelos comprensibles y proposicionales. Por modelo debemos entender que lo que se construye es un modelo, hipótesis o representación de la regularidad existente en los datos. Por su parte, que sea comprensible significa que se pueden representar de manera simbólica o de conjunto de condiciones por lo cual son de fácil entendimiento para los humanos. Y proposicional significa que se restringe a una sola tabla de datos y que no establece relaciones entre más de una fila a la vez y tampoco en más de un atributo a la vez [17].

Los algoritmos de árbol de decisión desglosan el conjunto de datos mediante la formulación de preguntas hasta conseguir el fragmento de datos adecuado para hacer una predicción.

Basado en las características de los datos de entrenamiento, el árbol de decisión “aprende” una serie de factores para inferir las etiquetas de clase de los ejemplos.

El nodo de comienzo es la raíz del árbol (o nodo de decisión), y el algoritmo dividirá de forma iterativa el conjunto de datos en la característica que contenga la máxima ganancia de información, hasta que los nodos finales (hojas o resultados) sean puros. Lo anterior se ejemplifica gráficamente en la *Figura 6*.

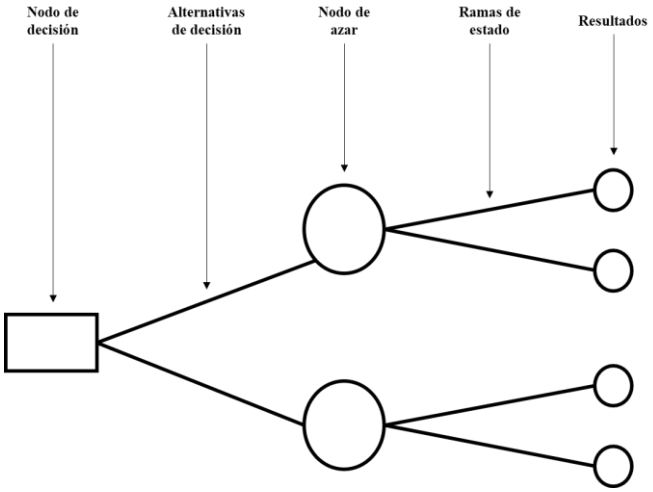

Figura 6. Estructura clásica de un árbol de decisión

Los Árboles de decisión cuentan con hiper parámetros que los definen. Son los siguientes:

a) Máxima profundidad: La máxima profundidad es la mayor longitud desde la raíz a las hojas. Una gran profundidad puede causar sobreajuste, y pequeña profundidad puede causar sub ajuste. Para evitar sobreajuste, ‘podaremos’ el árbol de decisión estableciendo un hiper parámetro con la máxima longitud.

$$Max\ Profundidad = K\ \rightarrow como\ máximo\ 2^k\ hojas$$

b) Máximo número de muestras: Cuando cortamos un nodo, se podría tener el problema de conseguir 99 muestras en uno de los cortes y una muestra en el otro, lo que sería un mal uso de los recursos. Para evitarlo, podemos establecer un máximo para el número de muestras que

permitimos para cada hoja. Esto se puede especificar como un entero o como un número flotante. Un pequeño número de muestras caerá en sobreajuste, mientras que un gran número de muestras caerá en sub ajuste.

c) Mínimo número de muestras: Análogo al anterior, pero con valores mínimos.

d) Máximo número de características: Muy frecuentemente tendremos muchas características (columnas) para construir un árbol. En cada corte, tendremos que hacer revisar todo el conjunto de datos en cada una de las características, lo que puede ser muy costoso. Una solución a este problema es limitar el número de características que se buscan en cada corte. Si este número es suficientemente alto, es probable que encontramos una buena característica entre aquellas que buscamos (aunque pueda no ser la perfecta). Sin embargo, si no es tan alto como el número total de características, la velocidad de los cálculos se elevará de manera significativa.

2.6.2.2 Los K-vecinos Más Cercanos (K Nearest Neighbors o KNN)

Los K-vecinos más cercanos, o KNN, pertenecen a un tipo especial de modelos de Minería de datos que se llaman frecuentemente “algoritmos perezosos”. Reciben este nombre porque no aprenden cómo discriminar el conjunto de datos con una función optimizada, en su lugar memorizan el conjunto de datos.

El nombre también se refiere a la clase de algoritmos llamados “no paramétricos”. Estos son algoritmos basados en instancias (ejemplos o eventos), que se caracterizan por memorizar el conjunto de datos de entrenamiento, y el aprendizaje perezoso es un caso particular de estos algoritmos, asociados con coste computacional cero durante el aprendizaje.

El algoritmo. El proceso que sigue el algoritmo es:

1. Escoge el número de los k vecinos y la distancia.

2. Encuentra el k vecino más cercano de la muestra que se pretende clasificar.
3. Asigna la etiqueta de clase por votación mayoritaria.

Gráficamente se puede representar como lo muestra la *Figura 7*.

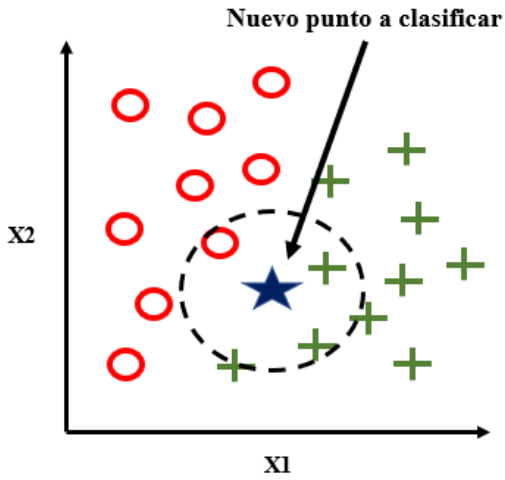

Figura 7. Ejemplo genérico de K-vecinos

El algoritmo encuentra las k instancias (ejemplos o eventos) que son más cercanas al punto que se quiere clasificar, basando sus predicciones en la distancia métrica. La principal ventaja es que se adapta a los nuevos datos de entrenamiento, al ser un algoritmo basado en la memoria. La desventaja es que el coste computacional se incrementa linealmente con el tamaño de los datos de entrenamiento.

Puntos a tener en cuenta:

- Si el algoritmo se enfrenta a un bucle, preferirá los vecinos con menor distancia a la muestra de clasificación. Si la distancia es similar, KNN elegirá la etiqueta de clase que esté primero en el conjunto de datos.
- Es fundamental elegir el valor k correcto para tener un buen balance entre el sobreajuste y el sub ajuste.

- También es crucial establecer una distancia métrica apropiada. Normalmente se usa la distancia 'Minkowski' que es una generalización de las distancias "Euclídea" y "Manhattan". Esta distancia, la Minkowski, se define en (1) de la siguiente manera:

$$(1) \qquad d\left(x^{(i)}, x^{(j)}\right) = \sqrt[p]{\sum_{k} \left| x_k^{(i)} - x_k^{(j)} \right|^p}$$

2.6.2.3 Matriz de confusión

Una herramienta muy útil para poder evaluar el desempeño de un algoritmo de aprendizaje automático es la matriz de confusión. Con esta herramienta se busca informar la manera en que se distribuyen los errores con algún algoritmo de clasificación [17].

Para los problemas más sencillos de clasificación donde tenemos solamente dos clases, como el que nos ocupa en esta investigación con una clase positiva (fraudulento representada con un uno) y una negativa (no fraudulento representada con un cero), la matriz de confusión es de dimensionalidad 2x2.

Se muestra, en la *Tabla I*, un resumen de la conformación de dicha matriz de confusión y los datos que nos ofrece después de que un algoritmo ha actuado en un conjunto de datos.

Tabla I - Matriz de confusión

		Clasificador	
		Negativos	**Positivos**
Valores reales	**Negativos**	a	b
	Positivos	c	d

Cada uno de los elementos de la Matriz de confusión, descrita en la *Tabla 1*, tienen los siguientes significados:

- **a** es el número de predicciones **correctas** de que un caso es **negativo**.

- **b** es el número de predicciones **incorrectas** de que un caso es **positivo**, es decir, la predicción es positiva cuando realmente el valor tendría que ser negativo. A estos casos también se les llama errores de tipo I.
- **c** es el número de predicciones **incorrectas** de que un caso es **negativo**, es decir, la predicción es negativa cuando realmente el valor tendría que ser positivo. A estos casos también se les llama errores de tipo II.
- **d** es el número de predicciones **correctas** de que un caso es **positivo**.

Una matriz de confusión ideal es aquella que genera valores cero fuera de la diagonal principal; es decir que los elementos b y c son cero. Esto es precisamente porque estos elementos son los que señalan cuantos casos no pudieron ser clasificados de manera adecuada después del proceso de entrenamiento.

La descripción de cada uno de los elementos de la matriz se indica a continuación:

- La Exactitud (*Ac*, del inglés *Accuracy*) es la proporción del número total de predicciones que fueron correctas:

$$Ac = \frac{a + d}{a + b + c + d}$$

- La Razón de Verdaderos Positivos (*TPrate*, del inglés *True Positive Rate*), a veces también denominada *Recall*, es la proporción de casos positivos que fueron correctamente identificados:

$$TPrate = \frac{d}{c + d}$$

- La Razón de Falsos Positivos (*FPrate*, del inglés *False Positive Rate*) es la proporción de casos negativos que han sido incorrectamente clasificados como positivos:

$$FPrate = \frac{b}{a + b}$$

- La Razón de Verdaderos Negativos (*TNrate*, del inglés *True Negative Rate*) es la proporción de casos negativos que fueron correctamente identificados:

$$TNrate = \frac{a}{a + b}$$

- La Razón de Falsos Negativos (*FNrate*, del inglés *False Negative Rate*) es la proporción de casos positivos que fueron incorrectamente clasificados como negativos:

$$FNrate = \frac{c}{c + d}$$

- La Precisión (*P*, del inglés *Precision*) es la proporción de casos predichos positivos que fueron correctos:

$$P = \frac{d}{b + d}$$

Frecuentemente son utilizados también los términos siguientes:

- Sensibilidad (*Se*, del inglés *Sensitivity*) como sinónimo de *TPrate* porque es la capacidad del clasificador de ser "sensible" a los casos positivos. Se debe hacer notar que 1-Se = FNrate
- Especificidad (*Sp*, del inglés *Specificity*) como sinónimo de *TNrate* porque puede dar una medida de la especificidad del test para marcar los casos positivos. Se debe hacer notar que 1-Sp = FPrate

Si un clasificador puede variar determinados parámetros puede lograrse incrementar los TP a costa de incrementar los FP o viceversa. En otras palabras, se desea una alta sensibilidad con una gran especificidad (o equivalente a una reducida FPrate).

Derivado de lo anterior, podemos indicar que los indicadores de Exactitud, Razón de Verdaderos Positivos y Razón de Verdaderos Negativos deben tender al 100% para indicar que el algoritmo tiene un desempeño óptimo. Por el otro lado también es posible señalar que los indicadores de Razón de Falsos Positivos y Falsos Negativos deben tender al 0% para señalar que el algoritmo está en su mejor nivel de desempeño.

Capítulo 3. Aspectos metodológicos

La metodología de extracción de conocimiento KDD ha dado la pauta a ejecutar de manera organizada la presente investigación. En este capítulo se habla de la forma en que está generado el conjunto de datos, la ventana de tiempo en donde se encuadran estos datos y la forma en que actúan los algoritmos seleccionados para extraer conocimiento que lleve a la confirmación de la tesis planteada.

3.1 Datos del proyecto

Estamos ante un proyecto del tipo investigación-acción. Este tipo de proyectos son característicos por la coexistencia entre los aspectos cognoscitivos y una intención de conseguir efectos objetivos y medibles [20].

La parte cognitiva está dada por el objetivo de conocer la manera en que las variables seleccionadas logran describir el comportamiento de empleados bancarios fraudulentos. Y la parte objetiva y medible está dada por el modelo obtenido para poder hacer una clasificación de los empleados conforme van generando acciones (o transacciones) para atender a sus clientes.

La presente investigación está encuadrada en el tipo de estudio retrospectivo (al analizar información ya registrada en los sistemas informáticos de la empresa objeto de estudio) [20].

La presente investigación se lleva a cabo en una empresa del sector financiero mexicano localizada en el noroeste del país. No es de las empresas más grandes, del sector, en el país sin embargo tiene características básicas que lo hacen elegible para la investigación:

- Tiene poco más de 18 años operando en México.
- Tiene varios productos financieros (Crédito, Ahorro, Inversión a plazo fijo y venta de seguros).
- Tiene una plantilla aproximada de 350 empleados activos trabajando en sus 28 Sucursales.

- Tiene varios canales de atención, el principal es la Sucursal, pero cuenta también con Corresponsales Financieros haciendo transacciones en línea.
- Cuenta con segregación de funciones (perfiles operativos) para sus distintas transacciones financieras.
- Además de almacenar sus transacciones de negocio, de unos años a la fecha también cuenta con registros de auditorías en las cuales almacena fecha, hora, sucursal, empleado, tipo de transacción y monto, entre otros.
- Cuenta con un esquema operativo de límites permitidos por perfil, con lo cual tiene implementadas autorizaciones (por niveles superiores).

Los trabajos de la presente investigación se han llevado a cabo entre el segundo semestre del 2019 y primer semestre del 2020.

En la empresa objeto de estudio existen áreas operativas y de staff que manejan información confidencial sobre fraudes realizados a la institución. Se habló con los responsables de cada área involucrada en este aspecto para identificar los datos que pudieran servir para esta investigación.

Se identificó y recopiló información en archivos Excel. Las áreas aportaron algunos datos adicionales para enriquecer los datos a analizar, todos en el mismo formato. El listado completo de datos recibidos es:

1. Listado de empleados (activos y no activos a lo largo de la historia).
2. Registro de faltas de los últimos 3 años (2017, 2018 y 2019).
3. Registro de anticipo de sueldos de los últimos 3 años (2017, 2018 y 2019).
4. Registro de actas administrativas de los últimos 3 años (2017, 2018 y 2019).
5. Casos de fraude documentados (del 2010 al 2019).

Aunque la empresa está activa desde el 2001 no se cuenta con registros de fraude documentados de los años anteriores a los señalados entre paréntesis.

Para completar el conjunto de datos inicial se extrajo información de las bases de datos centrales de la empresa. Se enlistan los datos obtenidos:

a. Accesos de los empleados a los sistemas informáticos (del 2013 al 2019).
b. Operaciones históricas del módulo de Crédito (del 2001 al 2019).
c. Operaciones históricas del módulo de Ahorro (del 2001 al 2019).
d. Autorizaciones o excepciones a reglas de negocio (del 2001 al 2019).
e. Pistas de auditoria (del 2014 al 2019).

3.2 Desarrollo de la investigación

A continuación, se detallan los pasos ejecutados y los resultados parciales obtenidos en cada uno de ellos como parte de la descripción del proceso adoptado.

3.2.1 Etapa de selección de datos

La primera etapa de la metodología del descubrimiento de conocimiento es la selección de los datos a ser analizados. La primera de las actividades que se describe en el presente trabajo es la identificación de los datos con los cuales se contaba.

3.2.1.1 Fuentes de datos disponibles

La empresa objeto de estudio cuenta con registros electrónicos que permiten tener a disposición una selección amplia de datos para esta investigación. Las fuentes de datos son diversas, entre ellas tenemos:

- Modelos de datos en donde se mantienen recopilados todos los Clientes de la empresa aún y cuando ya no tengan algún producto financiero contratado y activo;
- Modelos de datos con todos los cambios en los datos generales (dirección, teléfono, correo, estado civil, etc) de todos los Clientes a lo largo de la historia;

- Modelos de datos con el detalle de las transacciones de todos los clientes en todos los productos financieros que ofrece la empresa. Los productos de Crédito son los más antiguos con los que cuenta la empresa, los productos de Ahorro se comenzaron a ofertar en el año 2012 y los productos de Seguros se comenzaron a ofertar en 2015;
- Modelos de datos con el detalle de los datos personales (nombre, género, edad, dirección, teléfono, correo electrónico) de todos los Empleados de la empresa;
- Modelos de datos con el detalle de los datos personales (nombre, género, edad, tipo de parentesco, dirección, teléfono, correo electrónico) de los familiares en primera y segunda línea del Empleado a de la empresa;
- Modelos de datos con el registro histórico de cambios de perfil (privilegios en los sistemas) durante la permanencia del empleado en la empresa. Estos cambios de perfil se pueden dar por promociones en la empresa o bien para cubrir a otro compañero de manera temporal (vacaciones, incapacidad, renuncia repentina, etc);
- Modelos de datos con el registro histórico de todas las autorizaciones que hacen los usuarios facultados para eliminar "candados" del proceso de crédito. Un candado es una regla de negocio que el sistema aplica durante el proceso del Crédito (solicitud, contratación o recuperación) y que un empleado facultado puede eliminar después de un análisis de los casos particulares que los generan. Por ejemplo, el sistema cuenta con un candado para no otorgar préstamo a un Cliente con una edad superior a los 60 años, pero si un Analista de Crédito analiza el caso y considera que para una persona en específico se puede omitir esta regla entonces elimina el candado y el sistema registra esta autorización;
- Bases de datos en Excel con los adelantos de nómina que piden los Empleados cuando tienen necesidades económicas (disponibles para los años 2018 y 2019);
- Bases de datos en Excel con las Actas Administrativas a las que se hacen acreedores los Empleados por presentar comportamientos no adecuados al reglamento interno de la empresa (disponibles para los años 2017 y 2018);

- Bases de datos en Excel con el registro de las faltas del personal de Sucursales. En estos documentos se registra el motivo de las faltas. De manera general podemos mencionar varios temas personales y señalar que un motivo es la acumulación de retardos a la hora de entrada (disponibles para los años 2017, 2018 y 2019);
- Bases de datos en Excel con el registro de las incidencias (fraudes) localizados por el área de auditoría (disponibles para todos los años comprendidos entre el 2011 y el 2019, incluidos ambos);
- Archivos de texto plano con información técnica de las operaciones que hacen los Empleados para atender a los Clientes. Estos archivos no tienen una estructura estándar y se requiere de cierta programación para poder hacer un análisis adecuado de su contenido.

3.2.1.2 Entrevistas con las áreas usuarias

Como parte de la recopilación de datos, para su posterior selección, están disponibles entrevistas con los actores principales involucrados cuando se trata de procesamiento de fraudes al interior de la empresa. En específico se entablan conversaciones con los responsables de las áreas de Auditoría y Recursos Humanos (proveedoras de información).

Ellos dan a conocer los mecanismos con los cuales se identifica un fraude interno. Se enlistan las mecánicas señaladas:

- Al momento de hacer auditorías internas presencial y tratar de "cuadrar" la documentación impresa existente con los registros electrónicos se identifican incidencias que llevaban a un comportamiento sospechoso.
- Otros Empleados usaron el canal anónimo de reportes de comportamiento sospechoso de algún compañero para que fuera investigado;

- Clientes se acercaron a la Sucursal que le correspondía para aclarar alguna situación de sus pagos y que no veían reflejados en sus estados de cuenta de adeudos (para el caso del Crédito);
- Las menos de las veces familiares de algún Cliente se acercaban a la Sucursal a pedir dinero de las cuentas de Ahorro de un familiar recién fallecido y no existía dinero que supuestamente el familiar declaró estaba disponible.

Como se puede apreciar, ninguno de los mecanismos señalados apunta al uso de la tecnología para tratar de minimizar, a través de la prevención, la existencia de estas situaciones.

3.2.1.3 Análisis inicial de la empresa objeto de estudio

Cabe señalar que la empresa objeto de estudio, a pesar de contar con tecnología propia, no cuenta con software o conocimientos técnico de los integrantes de su equipo de Sistemas en temas de Minería de Datos, Machine Learning y/o Inteligencia Artificial.

En la página de la empresa, objeto de estudio, se obtuvieron los productos financieros ofertados a sus Clientes. Con estos productos se realiza una revisión detallada para identificar los incidentes detectados y documentados en donde están involucrados los Empleados de la empresa (en la Base de datos en Excel).

Se identificó el producto financiero más afectado (Crédito por medio de la Cobranza) y algunos otros aspectos del negocio (como lo es el manejo del efectivo de la Bóveda de la Sucursal para uso personal).

Las estadísticas completas se muestran en la *Tabla II*, y entre otras cosas se observa que el Crédito es el proceso de negocio que más se afecta por medio del fraude interno. En el 90% de los 93 casos documentados se trata de dinero que recuperan los cobradores que operan en campo y que no reportan en la caja de la Sucursal.

Para esta situación la empresa objeto de este estudio ha implementado diversos reportes y procedimientos con la intención de minimizar los efectos del fraude interno. A finales del año 2018 se hizo la contratación de personal del área de auditoría (hasta ese año existía un solo auditor en toda la empresa). Con dicha medida se consiguió separar las Sucursales en 4 sectores, uno por cada auditor, y con ello aumentar la cantidad de auditorías al año por cada Sucursal. Las auditorías pasaron de 0.75 a 4 por Sucursal al año y surtieron un buen efecto pues las incidencias ocurridas en el 2019 bajaron de manera drástica como se puede apreciar en el cuadro de estadísticas mostrado anteriormente.

Tabla II - Casos de fraude reportados

Producto financiero	2011	2012	2013	2014	2015	2016	2017	2018	2019	Total
Crédito	15	16	13	12	7	12	8	8	2	**93**
Ahorro				1				1		**2**
Seguros										**0**
Uso de efectivo						3	1	1		**5**
Falsedad en documentación					1	1	1			**3**
Total	**15**	**16**	**13**	**13**	**8**	**16**	**10**	**10**	**2**	**103**

Por consiguiente, se determinó enfocar el presente trabajo a los incidentes que tuvieran relación con las cuentas de ahorro de los Clientes. Estas operaciones son especiales porque existen variantes en la operación que permiten que los empleados con conocimiento del proceso puedan aplicar algunas técnicas de defraudación (por temas de seguridad no hablaremos aquí a detalle de estos mecanismos ni de los vectores de defraudación que fueron usados).

Es un hecho que cualquier cuenta de ahorro con saldo disponible se vuelve un activo codiciado por el empleado fraudulento.

Dicho lo anterior, los productos de ahorro son el foco de esta investigación; y en específico la extracción de dinero de dichas cuentas sin el consentimiento de los Clientes.

Como lo comentamos anteriormente, la empresa sí lleva un registro histórico de fraudes cometidos a lo largo de la historia. Ya mencionamos que el registro de incidentes en este archivo Excel es 100% manual. Dicho registro se ha estado normalizando y estandarizando con el transcurso de los años.

Actualmente se registran varios datos del incidente como son: la fecha de ocurrencia, la Sucursal, el nombre del Empleado, el monto estimado del incidente y en general detalles de la operación que han encontrado los auditores. El archivo contiene muchos datos que son narrativas de los sucesos; son datos no estructurados y en muchos casos el modus operandi de la incidencia es el mismo por lo cual se identifica un copiar-y-pegar entre las distintas ocurrencias registradas.

3.2.2 Etapa de preprocesamiento y limpieza de datos

Una vez que se cuenta con el contexto de la empresa y que también se cuenta con una buena cantidad de datos relacionados al problema a resolver a través de la investigación de esta tesis se procede a la organización de estos para comenzar la fase del preprocesamiento.

Definitivamente la mejor forma de organizar los datos existentes es a través de una base de datos. En esta investigación el uso de una base de datos está determinado para realizar el análisis de los datos existentes y obtener un modelo de comportamiento sospechoso de los empleados para anticiparse a los fraudes internos.

3.2.2.1 Creación de una Bodega de Datos

Para organizar los datos recopilados se ha creado una base de datos que toma el rol de Bodega de Datos. Con esta herramienta es más sencillo manejar los datos a través de comandos con lenguaje SQL.

La base de datos está implementada en SQL Server® en su versión 2019 estándar para aprovechar los módulos referentes a la minería de datos. Con este software se podrían cumplir dos objetivos iniciales con un solo esfuerzo: a) organizar los datos; y b) obtener una primera vista gráfica de los datos recolectados (una imagen dice más que mil palabras).

Con todas las fuentes de datos disponibles se crea un modelo de datos. Se tomó como base la tabla de empleados. Las tablas relacionadas a la de empleados se muestran en la *Tabla III*.

Tabla III - Tablas relacionadas a la tabla de empleados

Datos personales de los empleados	Eventos administrativos de los empleados	Bitácoras de movimientos	Fraudes documentados
1) Empresas 2) Áreas 3) Puestos 4) Sucursales 5) Estado civil 6) Edades 7) Escolaridades	1) Anticipos de sueldo 2) Actas administrativas 3) Faltas 4) Motivos de faltas	1) Log de operaciones de sistemas internos 2) Tipos de operaciones de sistemas internos 3) Log de eliminación de candados operativos 4) Tipos de candados operativos 5) Log de créditos relacionados 6) Tipos de parentescos	1) Desviaciones 2) Tipos de desviaciones

Con lo mencionado anteriormente se llegó a un modelo de datos relacional que se muestra en la *Figura 8*.

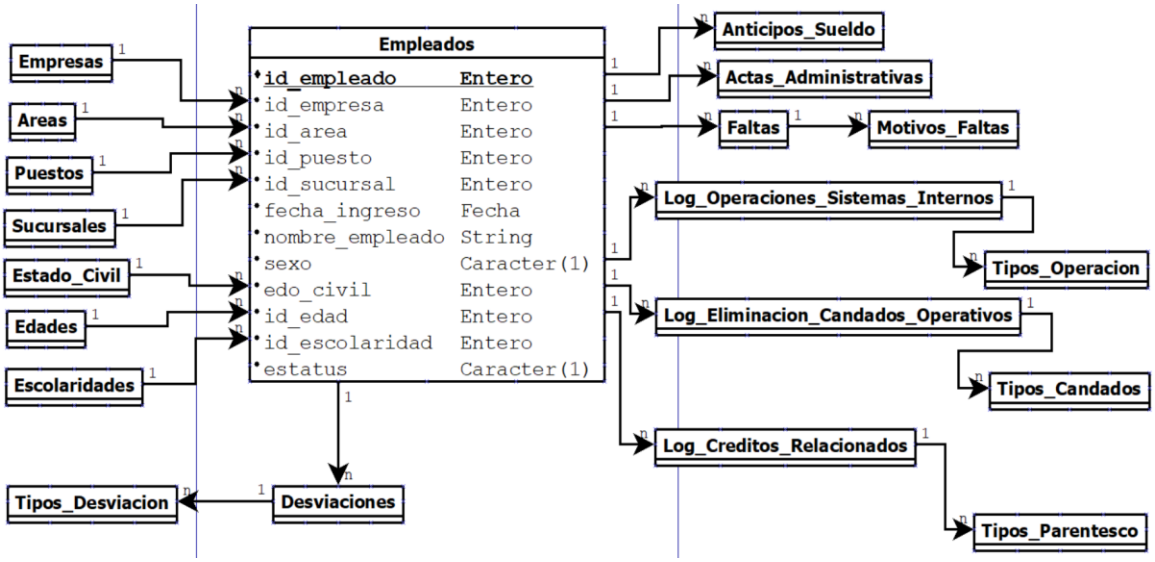

Figura 8 - Modelo de datos para análisis de información

3.2.2.2 Preprocesamiento de los datos

El preprocesamiento de los datos consiste en una separación manual del contenido de los archivos de fuentes de datos originales disponibles para poder poblar el modelo de datos.

También manualmente se realizan mecanismos de ordenamiento y eliminación de duplicados en cada una de las nuevas tablas; por ejemplo, para la obtención de los distintos

catálogos; y de separación de campos. Desde *SQL Server*® se alimenta de datos, manualmente, a cada tabla resultante.

En este punto, donde la actividad de carga de datos no es constante y/o recurrente, no se usan procesos ETL.

3.2.3 Limpieza de los datos

La limpieza de datos es el acto de descubrimiento y corrección o eliminación de registros de datos erróneos de una tabla o base de datos. El proceso de limpieza de datos permite identificar datos incompletos, incorrectos, inexactos, no pertinentes, etc. y luego substituir, modificar o eliminar estos datos sucios. Después de la limpieza, la base de datos podrá ser compatible con otras bases de datos similares en el sistema [19].

Podría sonar lógico, pero más vale aclarar, que una base de datos con datos sucios, puede llevar a conclusiones erróneas o incompletas por la falta de calidad. Para que un dato cumpla con la calidad deseada debe ser exacta. El ser exacta conlleva ser integra, consistente y densa.

Los datos para ser íntegros deben tener entereza; y la entereza se logra con la corrección de datos que presentan anomalías. Las anomalías pueden darse por muchos factores. Los datos, para ser íntegros, también deben cumplir con ser válidos; y son válidos si los datos satisfacen las restricciones de integridad como lo son el reflejo de un dato en una tabla distinta.

Los datos originales presentaban muchas inconsistencias que suelen existir en sistemas viejos que van evolucionando con nuevas funcionalidades o bien mantenimientos constantes a las mismas y que no se reflejan en las estructuras de datos legadas. Dentro de las deficiencias más sobresalientes se pueden mencionar:

- Los Empleados tienen más de un número de empleado en la base de datos. En las bases de datos disponibles se identificaron 3 números distintos: a) el que tiene asignado en la intranet como parte del catálogo central de empleados controlado por

el área de Recursos Humanos, b) el que tiene registrado en la base de datos de ahorro y con el cual tiene asignado sus privilegios para operar los productos financieros de ahorro; y c) el número de empleado que tiene asignado en la base de datos central que unifica todas las operaciones de la empresa.

- Existen muchos números de operaciones. Prácticamente cada módulo operativo del sistema central de la empresa cuenta con un número secuencial que tiene el objetivo de hacer única la transacción. Este dato, en muchos de los módulos, se ha reciclado con el paso del tiempo y entonces genera que varias transacciones tengan el mismo número después de cierto tiempo.
- No todas las claves de operaciones son categóricas o existen desde el inicio de las operaciones de la empresa, originando con esto que en algunos casos se tengan datos nulos que no permiten identificar o clasificar en algún grupo la operación

Las reglas de integridad del Modelo de datos, cuando fueron habilitadas, no lograron hacer que el modelo fuera exacto e íntegro. Con todo lo expuesto anteriormente, se puede intuir, que en ningún momento se consiguió obtener un gráfico con los datos alimentados en el Modelo de Datos. El resultado de lo anteriormente expuesto lleva a tomar una decisión drástica dentro de la investigación y consiste en buscar información complementaria (en caso de existir) o bien generarla a partir de los datos con los que se cuentan ya que la base de datos inicial no es suficiente para abordar la investigación.

3.2.4 Etapa de transformación y reducción de datos

Como contexto de esta etapa baste recapitular que la información que se ha obtenido hasta el momento no es suficiente para la investigación en curso. Una serie de anomalías en la misma no permite dar por terminada la fase de preprocesamiento y limpieza de datos.

En la *Tabla IV* se pueden observar, a manera de resumen, los datos seleccionados con su formato origen y el formato final producto de la transformación.

Tabla IV - Datos seleccionados con formato original y final

Dato	Formato original	Formato final (después de la transformación)
Empresa	Cadena de texto con el nombre de la empresa. Por ejemplo “Empresa 1”	Categorizado por números. Por ejemplo: 1, 2, 3, etc
Área	Cadena de texto con el nombre del área en la que se desenvuelve un empleado o los distintos nombres con los que se conoce a una sola área. Por ejemplo “Crédito” o “Crédito y Cobranza”	Categorizado por números. Por ejemplo: 1, 2, 3, etc
Puesto	Cadena de texto con el nombre del puesto del empleado incluyendo mismo puesto con faltas de ortografía. Por ejemplo “Analista de Credito” o “analista crédito”	Categorizado por números. Por ejemplo: 1, 2, 3, etc
Sucursal	Cadena de texto con el nombre de la sucursal. Por ejemplo “La Paz”	Categorizado por números. Por ejemplo: 1, 2, 3, etc
Estado civil	Cadena de texto con los distintos estados civiles incluyendo errores tipográficos. Por ejemplo “Viudo” o “Viudo con hijos”	Categorizado por números. Por ejemplo: 1, 2, 3, etc
Edades	Número que representa la edad al momento de la contratación. Por ejemplo 25	Categorizado por rangos de edad al momento del análisis de una incidencia de fraude
Escolaridad	Cadena de texto con la descripción de una escolaridad. Incluyendo precisiones con el grado de estudios. Por ejemplo “Bachillerato” o “Bachillerato trunco”	Categorizado por números. Por ejemplo: 1, 2, 3, etc

Esto significa que en la fase de transformación y reducción de datos se inicia un nuevo ciclo de la metodología de extracción de conocimiento (que la misma metodología tiene contemplada como una mejora continua). Así, se requiere de una nueva selección de datos y sus subsecuentes tratamientos para procesamiento previo y limpiar. En este punto es un poco confuso el proceso pues con los conocimientos adquiridos de los datos de la empresa ha permitido mezclar las etapas con la intención de optimizar tiempos en su tratamiento. A esta nueva etapa se le llamará en lo sucesivo extracción de información adicional.

3.2.4.1 Extracción de información adicional

Junto con el área de Sistemas de la empresa se verificó que es posible extraer información adicional, con un formato específico, de la base de datos central. La intención es cruzar distintas

tablas para garantizar que los datos se pueden relacionar y con ello tener integridad que nos lleve a conclusiones certeras.

Con las entrevistas correspondientes, y después de varios cruces de datos, se ha formado un conjunto de datos adicionales. Se tomaron técnicas de validación de información manuales para confirmar, por medio de las pantallas de los sistemas, que los datos pertenecen a las transacciones seleccionadas. Los conjuntos de datos acordados consisten en lo siguiente:

- Base de datos en archivo plano con las actualizaciones a los números telefónicos de los todos los Clientes. Se incluyeron a todas las sucursales y a todos los usuarios sin filtrar a los perfiles seleccionados para esta tesis. El período del tiempo de la información fue del 01/Ene/2018 al 30/Jun/2019, es decir 18 meses;
- Base de datos en archivo plano con todas las transacciones de retiro de dinero en efectivo de las cuentas de ahorro de los Clientes. Se incluyeron a todas las sucursales y a todos los usuarios sin filtrar a los perfiles seleccionados para esta tesis. El período del tiempo de la información fue del 01/Ene/2018 al 30/Jun/2019, es decir 18 meses;
- Base de datos en archivo plano con todas las transacciones de transferencias entre cuentas del "mismo banco" que un empleado ha realizado, es decir aquellas en donde el dinero no salió de la empresa y solo cambió de cuenta de un cliente a otro. Se incluyeron a todas las sucursales y a todos los empleados y sin filtrar a los perfiles de Sucursal seleccionados para esta tesis. El período del tiempo de la información fue del 01/Ene/2018 al 30/Jun/2019, es decir 18 meses;
- Base de datos en archivo plano con los números de cuenta de ahorro de los empleados, es decir en donde el empleado sea también un cliente de la empresa. Se incluyeron a todas las sucursales y a todos los usuarios sin filtrar a los perfiles seleccionados para esta tesis. Esta consulta fue sin límite histórico para poder obtener las cuentas, incluso, de empleados que ya no laboran en la empresa;

Se debe aclarar, e insistir, que la información señalada en los puntos anteriores no es la extracción de datos de una tabla. Se trata de información relacionada y obtenida con el cruce de

llaves de distintas tablas. Para llegar a este punto se debe tener un conocimiento específico del Modelo de Datos de la empresa para hacer la extracción adecuada de datos.

La selección del período de extracción de la información está basada en el volumen de datos disponible para su análisis. Se busca que el volumen sea manejable para hacer experimentos desde una laptop y que en el período seleccionado existan casos documentados del tipo de fraude que se está buscando.

La Empresa, al momento del inicio de la tesis, contaba con 120 mil cuentas de ahorro (con igual cantidad de Clientes) y el promedio de transacciones por mes en este tipo de cuentas es de 6. Estamos hablando de poco más de 720 mil transacciones (o registros) por mes. En este conteo no se registran operaciones adicionales como los mantenimientos a los datos de las cuentas (como cambios en el número de teléfono o de domicilio).

3.2.4.2 Detalle de los datos seleccionados

Se comenzó a separar los datos con los cuales comenzar a trabajar para crear un Modelo de Datos, que hiciera las veces de Bodega de Datos, y pasar a las siguientes etapas de esta investigación. Los datos se agruparon para formar tres conjuntos de datos diferentes con sus correspondientes perspectivas de la operación:

- Transferencias entre cuentas del mismo banco de los Empleados de la Empresa; es decir, una transferencia de dinero que sucede desde la cuenta de ahorro de un empleado que también es cliente del banco y que lleva como destino la cuenta de ahorro de otro empleado que también es cliente del mismo banco.
- Registros de auditorías de sistemas (en ocasiones llamadas huellas de auditoría) por el mantenimiento al número telefónico de un Cliente.
- Disposiciones de dinero en efectivo, ocurridos en las Sucursales, desde las cuentas de ahorro de los Clientes.

La descripción detallada del primer conjunto de datos (nombre del campo y descripción), denominado transferencias entre cuentas del mismo banco, se muestran en la *Tabla V*. Para el período seleccionado (consistente en 18 meses históricos) se formó un archivo con un total de 396 registros.

Tabla V - Datos para Transferencia bancaria entre cuentas del mismo banco

No	Nombre del campo	Descripción
1	**fecha_creacion_transferencia**	Es la fecha, en formato dd/mm/yyy en la que se realiza la operación
2	**num_empleado**	Es el número de empleado que ejecuta la transacción en el sistema. Está relacionado con el catálogo de empleados de la empresa que se aloja en la intranet
3	**id_solicitante_empleado**	Es el número de empleado dueño de la cuenta de Ahorro, es decir también es un empleado-cliente
4	**nombre_solicitante_empleado**	Es el nombre completo del empleado-cliente que pide la transacción, es decir el que recibe los recursos económicos de la transacción
5	**rfc_solicitantante_empleado**	Es el Registro Federal de Contribuyentes del empleado-cliente. En el sistema es la llave para localizar a un empleado
6	**cuenta_empleado**	Es la cuenta destino de los recursos financieros involucrada en la operación. El sistema permite que los clientes tengan más de una cuenta de un mismo Producto financiero, por eso es importante identificar la cuenta que está involucrada en la operación
7	**descripción**	Este campo se usa para señalar el tipo de producto financiero que está involucrado
8	**pk_tranferencia_id**	Es un número secuencial consecutivo para marcar como única una transacción
9	**cuenta_origen**	Es el número de la cuenta desde la cual salen los recursos financieros. Esta es una cuenta perteneciente a alguno de los productos financieros de ahorro
10	**nombre_cliente_origen**	Es el nombre del cliente, que puede ser o no un empleado de la empresa, desde la cual se hace la extracción de los recursos financieros. En este caso es una cuenta de un producto financiero asociado al ahorro
11	**monto**	Es la cantidad de dinero por la cual se hizo la transacción
12	**cuenta_destino**	Es el número de la cuenta a la cual llegan los recursos financieros. Esta es una cuenta perteneciente a alguno de los productos financieros de ahorro
13	**movimiento_id**	Es un número secuencial en la tabla de movimientos para hacer único el movimiento
14	**usuario_creo_transaccion**	Es el número de empleado que ejecuta la transacción en el sistema. Está relacionado con el catálogo de empleados de la empresa que se aloja en el servidor central de la empresa
15	**nombre_usuario_creo_transaccion**	Es el nombre del empleado que ejecuta la transacción en el sistema. Está relacionado con el catálogo de empleados de la empresa que se aloja en el servidor central de la empresa
16	**no_empleado_creo_transaccion**	Es el número de empleado que ejecuta la transacción en el sistema. Está relacionado con el catálogo de empleados de la empresa y que se registran en el sistema de ahorro
17	**deposito_id**	Es un número secuencial en la tabla de movimientos del tipo depósito para hacer único el movimiento
18	**disposicion_id**	Es un número secuencial en la tabla de movimientos del tipo disposición para hacer único el movimiento
19	**spei**	Es un indicador que señala si la operación fue una transferencia de tipo *SPEI* (null para no y 1 para si)
20	**fecha_termino_usuario**	Es la fecha, en formato dd/mm/yyy en la que se termina la operación. En el pasado algunas transacciones se hacian manualmente y podia terminar en una fecha diferente a la que se iniciaba

Por su parte, la descripción detallada de los datos disponibles para el segundo conjunto de datos (nombre del campo y descripción) consistente en las auditorías a mantenimientos de números telefónicos de los Clientes, se muestran en la *Tabla VI*. Para el período seleccionado (consistente en 18 meses históricos) se formó un archivo con un total de 1,119 registros.

Para finalizar con este apartado, se hace mención que el detalle de los datos disponibles para el tercero y último de los conjuntos de datos (nombre del campo y descripción) referente a las

operaciones de disposición de dinero en efectivo desde las cuentas de ahorro se encuentran en la *Tabla VII*. Para el período seleccionado (consistente en 18 meses históricos) se formó un archivo con un total de 7,907 registros.

Tabla VI - Datos de auditorías al mantenimiento de número telefónico del cliente

No	Nombre del campo	Descripción
1	**fecha**	Es la fecha, en formato dd/mm/yyy en la que se realiza la operación
2	**clave**	Es la clave que indica el tipo del mantenimiento. Se tienen identificados los siguientes: 48 – Modificación de persona 176 – Actualización de número celular 177 – Actualización de número de teléfono
3	**operacion**	Es la descripción del tipo de operación y está relacionada al campo "clave" de este mismo archivo de datos
4	**id_solicitante**	Es el número de Cliente que solicita la transacción en el sistema
5	**cliente**	Es el nombre completo del Cliente que solicita la transacción en el sistema
6	**dato_actual**	Es el número de teléfono que permanece en la base de datos de la Empresa una vez que se termina con éxito la transacción
7	**dato_modificado**	Es el número de teléfono que fue modificado (viejo) en la base de datos de la Empresa una vez que se termina con éxito la transacción
8	**nombre_empleado**	Es el nombre del Empleado que ejecuta la transacción en el sistema
9	**puesto**	Es el puesto (relacionado al perfil) del Empleado que ejecuta la transacción
10	**cel_empleado**	Es el número de celular que el Empleado que ejecuta la transacción tiene registrado en la base de datos de empleados
11	**sucursal**	Es la Sucursal desde la cual se ejecuta la transacción
12	**observaciones**	En este archivo de datos se ocupa este campo para señalar que tipo de teléfono ha sido modificado (teléfono de casa o celular) y bien dado de alta

Como se puede observar, las perspectivas que se abordan con estos conjuntos de datos van desde el manejo de dinero electrónico (transferencias entre cuentas), manejo de efectivo (disposición desde la Sucursal) y mantenimiento a los datos sensibles del Cliente (teléfono celular) para identificar el comportamiento sospechoso de un Empleado.

Tabla VII - Datos para disposición de efectivo desde una cuenta

No	Nombre del campo	Descripción
1	**fecha_creacion_transferencia**	Es la fecha, en formato dd/mm/yyy en la que se realiza la operación
2	**num_empleado**	Es el número de Empleado que ejecuta la transacción en el sistema. Está relacionado con el catálogo de empleados de la Empresa
3	**id_solicitante_empleado**	Es el número de Empleado dueño de la cuenta de Ahorro, es decir también es un Empleado-Cliente
4	**nombre_solicitante_empleado**	Es el nombre completo del Empleado-Cliente que pide la transacción, es decir el que recibe los recursos económicos de la transacción
5	**rfc_solicitantante_empleado**	Es el Registro Federal de Contribuyentes del Empleado-Cliente. En el sistema es la llave para localizar a un Empleado
6	**cuenta_empleado**	Es la cuenta destino de los recursos financieros de la operación.
7	**descripcion**	Es la descripción del producto financiero involucrado en la operación
8	**pk_disposiciones_id**	Número secuencial único en la tabla de movimientos del tipo disposición
9	**monto**	Es el monto económico de la operación
10	**forma_pago_id**	Es el canal por el cual se hizo la disposición del dinero. Se tienen: 1 – Corresponsal Financiero 2 – Cheque 3 – Efectivo 4 – Depósito 27 – Transferencia 31 – Transferencia *SPEI* 32 – Traspaso Ahorro
11	**forma_pago**	Nombre de la forma de pago. Está relacionado a la forma_pago_id
12	**banco_id**	Es el identificador del elemento contable que registra la transacción.
13	**cinsncorto**	Es el nombre corto del banco_id
14	**nombre_corto**	En todos los casos la operación fue realizada desde la caja de una Sucursal. El contenido del campo es "CAJA SUCURSAL"
15	**cheque**	Las transacciones de este archivo fueron en efectivo; este campo está vacío
16	**movimiento_id**	Es un número secuencial que identifica a la transacción como única
17	**aviso_id**	Nulo por default. Con dato cuando una transacción genere una alerta.
18	**transaccion_id**	Todas las transacciones tienen un identificador formado por la concatenación de varios campos que forman el transaccion_id
19	**tarjeta_operativa_id**	Es el número de tarjeta que está involucrada en la transacción si el Cliente la presenta. Varios registros del archivo contienen valor nulo.
20	**app**	La app genera transacciones para los Clientes
21	**transaccion_version_id**	Es el número de versión del software con la cual se hizo la transacción
22	**usuario_creo_transaccion**	Es el número de Empleado que ejecuta la transacción en el sistema.
23	**nombre_usuario_creo_transaccion**	Nombre del Usuario que ejecuta la transacción en el sistema.
24	**no_empleado_creo_transaccion**	Número de Empleado que ejecuta la transacción en el sistema.
25	**fecha_termino_usuario**	Es la fecha de baja de un Empleado. Una fecha futura en este campo indica la fecha en la que vencen sus privilegios en el Sistema de la Empresa

3.2.4.3 Estructura del nuevo conjunto de datos

La siguiente actividad consiste en generar un formato estándar que permita integrar a los tres conjuntos de datos, cada uno con sus propios tipos de movimientos, en un solo conjunto de datos final. Esto tiene como objetivo poder procesar las tres perspectivas que aportan los conjuntos de datos al mismo tiempo.

Después de analizar los nuevos datos obtenidos de manera individual y con un enfoque global se define la estructura que se detalla en la *Tabla VIII* y que genera el nuevo conjunto de datos.

Tabla VIII - Datos del conjunto final integrado

No	Nombre del campo	Descripción
1	**id_operacion**	Es el tipo de operación que forma el conjunto de datos. Se tienen los siguientes tipos: 1 – Transferencias 2 – Auditoría de teléfonos 3 – Disposiciones de efectivo
2	**fecha_transaccion**	Fecha en la que se generó la operación
3	**num_empleado_origen**	Es el número de Empleado que genera la transacción
4	**id_cliente_origen**	Es el número de Cliente que es dueño de la cuenta sobre la que se está haciendo la transacción
5	**cuenta_origen**	Es el número de cuenta sobre la que se está haciendo la transacción. En la operación de "Transferencia" es la cuenta desde la cual se disponen los recursos
6	**cuenta_destino**	Es el número de cuenta a la que llegan los recursos de una transacción. En algunas operaciones se tiene significados diferentes: dato_modificado – para las auditorías de cambios en los teléfonos de los Clientes forma_pago_id – para las disposiciones de dinero de las cuentas de los productos de ahorro
7	**monto**	Es el importe de la transacción
8	**usuario_transaccion**	Es el Empleado que ejecutó la transacción
9	**num_empleado_transaccion**	Se llena con el Número de Emplead en el caso de que el Cliente que hizo la transacción también es Empleado
10	**fraude**	Es un indicador de si la transacción fue realizada por un Empleado registrado como fraudulento dentro de la Empresa

Para el período seleccionado (siendo consistente con los conjuntos de datos individuales, se compone de 18 meses históricos) se formó un archivo con un total de 8,302 registros ya integrando los tres tipos de movimientos.

Para este segundo ciclo de preparación de información, la limpieza de datos fue un poco más sencilla pues ya se habían detectado deficiencias en los datos originales con la anterior extracción de información. No fue necesario eliminar registros o complementar datos por

encontrarse datos nulos e inconsistentes. Este tipo de limpieza fue integrado en la extracción de información al hacer los cruces adecuados entre las tablas originales.

La fase de limpieza fue un poco particular pues en realidad se trató de ajustar uno de los datos, que permitiera complementar la estructura y generar un solo conjunto de datos. El dato ajustado es `cuenta_destino`.

La cuenta destino es el número de cuenta a la que llegan los recursos de una transacción del tipo transferencia. Derivado de que estamos integrando transacciones del tipo mantenimiento a números telefónicos y disposiciones de dinero se tuvo que ajustar este dato a significados diferentes según el tipo de operación. Con esto podemos señalar que en el caso de los mantenimientos de los teléfonos de los Clientes se integró el dato que fue modificado; es decir el número telefónico que se tenía registrado en la base de datos antes de hacer el mantenimiento. Y para el caso de las transacciones del tipo disposiciones se consideró el identificador de la forma de pago; que en nuestro caso es el tipo de producto financiero de la familia del ahorro.

Con este ajuste se cuenta con todo el grupo de datos completos. Y es con este conjunto de datos con el que se hace el análisis a través de algoritmos de Minería de datos.

3.2.5 Etapa de Minería de Datos

La etapa de Minería de datos es una de las partes esenciales de este trabajo pues es el uso de esta tecnología, aplicada a un grupo de datos, el que permite generar un modelo de comportamiento de empleados en el sector bancario.

La cantidad de registros con la que cuenta el conjunto de datos utilizado para los experimentos de esta investigación permitió que la infraestructura necesaria no requiriera de características especiales (y costosas para replicar el experimento en otras empresas del sector).

3.2.5.1 Arquitectura técnica de la solución

Se busca en todo momento que los requisitos tecnológicos sean accesibles para muchas empresas del sector y de las dimensiones que tiene la empresa objeto de estudio. Lo anterior para que el presente trabajo pueda ser fácilmente tomado como base para una implementación sencilla, económica y exitosa.

La arquitectura está formada por los tres componentes básicos de una solución informática: hardware, software y comunicaciones.

El hardware con el que se desarrolló la investigación está compuesto por un equipo personal móvil (laptop) con las características descritas en la *Tabla IX*. Y como se puede confirmar es un equipo de alta gama dentro del rubro de equipo personal, pero es de destacar que no se trata de un servidor de altas prestaciones.

Tabla IX - Características del Hardware usado en la investigación

Característica	Descripción
Fabricante:	Acer
Modelo:	Nitro AN515-51
Procesador:	Intel® Core™ i7-7700HQ CPU @ 2.80GHz
Memoria instalada:	16.0 GB
Sistema operativo:	Windows 10 (64 bits)

El software con el que se hizo la implementación de los algoritmos de Minería de datos se describe en la *Tabla X* (en estricto orden alfabético). Se puede confirmar que, con excepción del software base (Sistema Operativo) todo fue implementado con software libre que es relativamente sencillo obtener.

El lenguaje de programación seleccionado para poder implementar los algoritmos es Python apoyado en los intérpretes VS Code y Spyder; todo dentro de un marco de trabajo basado en Anaconda. El señalamiento de las dos versiones distintas de Python es derivado de que mientras se desarrollaba la investigación el fabricante liberó una versión más nueva. Sin embargo, al

revisar las compatibilidades no se encontró argumento para poder hacer la actualización correspondiente.

Tabla X - Características del Software usado en la investigación

Software	Utilidad en este trabajo
Anaconda navigator versión 3	Anaconda es una distribución libre y abierta de los lenguajes *Python* y *R*, utilizada en ciencia de datos, y aprendizaje automático. Esto incluye procesamiento de grandes volúmenes de información, análisis predictivo y cómputo científico. El uso de este *framework* es la organización de los experimentos con volúmenes de datos y scripts realizados en *Python* y *R*.
Dia v0.97.2	Es un editor de diagramas de tipo *Open Source*. Este software fue utilizado en este trabajo para realizar diversos diagramas (entre ellos el de la Bodega de Datos y la Taxonomia del *Machine Learning*).
Excel 2019 (pero puede ser desde versión 2010)	*Microsoft Excel* es una hoja de cálculo desarrollada por *Microsoft* para *Windows*, *macOS*, *Android* e *iOS*. Cuenta con cálculo, herramientas gráficas, tablas calculares y un lenguaje de programación macro llamado *Visual Basic* para aplicaciones. Este software fue utilizado para la organización e intercambio de la información inicial con los responsables de las áreas de la Institución.
Microsoft Analysis Services	Es una herramienta, desarrollada por *Microsoft*, para el procesamiento analitico y mineria de datos. Trabaja acoplado a *Microsoft SQL Server*. En esta investigación fue usado para diagramar los modelos resultantes, asi como para un entendimiento más rápido de la inter relación de variables.
Microsoft SQL Server 2019	*Microsoft SQL Server* es un sistema de gestión de base de datos relacional, desarrollado por la empresa *Microsoft*. El lenguaje de desarrollo utilizado es *Transact-SQL*, una implementación del estándar *ANSI* del lenguaje *SQL*, utilizado para manipular y recuperar datos, crear tablas y definir relaciones entre ellas. La Bodega de datos implementada en esta investigación fue implementada en este manejador de Bases de Datos.
Python versión 3.7.0 y versión 3.8.1	*Python* es un lenguaje de programación interpretado cuya filosofia hace hincapié en la legibilidad de su código. Se trata de un lenguaje de programación multiparadigma, ya que soporta parcialmente la orientación a objetos, programación imperativa y, en menor medida, programación funcional. Las dos versiones utilizadas obedecen a que al iniciar el trabajo la versión más estable era la 3.7.0 pero al avanzar en los trabajos se hizo necesario integrar algunas librerias de tratamiento de algoritmos de *Machine Learning* más depurados en la versión 3.8.1.
Spyder 4.0.1	Es un entorno de desarrollo integrado multiplataforma de código abierto para programación cientifica basada en *Python*. Los primeros análisis de correlación de datos de esta investigación, obtenidos de scripts escritos en *Python*, fueron ejecutados en esta plataforma.
VS Code versión 1.41.1	*Visual Studio Code* es utilizado para poder ejecutar los scripts y librerias de *Analisys Services* de *Microsoft* para el análisis gráfico de los datos. Los gráficos de correlación de variables usados en este trabajo provienen del uso de este software.

El complemento de la arquitectura, el componente de comunicaciones, es una conexión a internet. Dicha conexión fue usada para la instalación y actualización de las librerías necesarias en la arquitectura. Los experimentos, para su ejecución, no requirieron de la conexión a internet pues todo se hacía localmente en el hardware.

3.2.5.2 Selección del algoritmo apto al problema

El primer problema al iniciar la experimentación es la selección del algoritmo que pueda representar de la mejor manera el problema a resolver. El algoritmo más apto es aquel que tenga los mejores resultados basados en los valores que arrojan los indicadores al ejecutar cada uno de los algoritmos al mismo grupo de datos. En este caso, los indicadores, son los descritos por la matriz de confusión y el desempeño de eficiencia del algoritmo.

Se comienza la descripción del proceso revisando el programa en Python, en *Programa 1* se puede revisar el código fuente completo, que evalúa el desempeño de varios algoritmos (al mismo tiempo) sobre un solo conjunto de datos (incluso, sobre un mismo grupo de datos de entrenamiento).

```
#!/usr/bin/python3

#cargar librerias
import pandas
from pandas.plotting import scatter_matrix

import matplotlib.pyplot as plt
from sklearn import model_selection
from sklearn.metrics import classification_report
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix
from sklearn.tree import DecisionTreeClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB

filename = "DataSet_AVC_20191127.csv"
names = ['id_operacion', 'dias_trx', 'num_empleado', 'mismo_origen_destino', '
id_monto', 'dias_antig_fecha_trx', 'fraudulento']
dataset = pandas.read_csv(filename,names=names)

print(dataset.shape)
print(dataset.head(20))
```

```
print(dataset.describe())
print(dataset.groupby('fraudulento').size())

dataset.plot(kind='box', subplots=True, layout=(2,7), sharex=False, sharey=Fal
se)
plt.show()

dataset.hist()
plt.show()

scatter_matrix(dataset)
plt.show()

array = dataset.values
X = array[:,0:6]   #columnas 0,1,2,3,4,5 (dimensiones)
Y = array[:,6]     #columna 6 (indicador de empleado fraudulento)

validation_size = 0.20
seed = 73
scoring = 'accuracy'
X_train, X_validation, Y_train, Y_validation = model_selection.train_test_spli
t(X,Y,test_size=validation_size,random_state=seed)
models = []
models.append(('Logistic Regression',LogisticRegression(solver='liblinear',mul
ti_class='auto')))
models.append(('Linear Discriminant Analysis',LinearDiscriminantAnalysis()))
models.append(('K-Nearest Neighbors',KNeighborsClassifier()))
models.append(('Desicion Tree Classifier',DecisionTreeClassifier()))
models.append(('Gaussian Naive Bayes',GaussianNB()))
models.append(('Support Vector Machine',SVC(gamma='auto')))

results = []
names = ["LR","LDA","KNN","DTC","NB","SVC"]

for name, model in models:
    kfold = model_selection.KFold(n_splits=10,random_state=seed)
    cv_results = model_selection.cross_val_score(model,X_train,Y_train,cv=kfol
d,scoring=scoring)
    results.append(cv_results)
    msg = "Method %s: mean %f std(%f)"%(name,cv_results.mean(),cv_results.std(
))
    print(msg)

fig = plt.figure()
```

```
fig.suptitle('Comparacion de Algoritmos')
ax = fig.add_subplot(111)
plt.boxplot(results)
ax.set_xticklabels(names)
plt.show()
```

Programa 1 - Programa en Python para evaluación de conjunto de datos

La primera parte del *Programa 1* consiste en las librerías necesarias y la carga del conjunto de datos. Referente a esto se puede observar lo siguiente:

- Las librerías utilizadas son:
  - `pandas` para los cálculos y manejo de matrices necesarias en el procesamiento de datos; y
  - `matplotlib` para representar gráficamente las características y confrontación de las variables entre sí.
- Se importan librerías para los siguientes algoritmos de aprendizaje automático:
  - Árboles de decisión
  - K-vecinos
  - Regresión logística
  - Análisis de discriminantes lineales
  - Máquina de vectores de soporte
  - Naive Bayes
- Se importan librerías para obtener las siguientes métricas:
  - Matriz de confusión
  - Desempeño de eficiencia
  - Reporte de clasificación
- La carga del conjunto de datos que se ha determinado adecuada para esta investigación y que está almacenado en el archivo “DataSet_AVC_20191127.csv”

La segunda parte del *Programa 1* se mostrará con la ejecución del mismo pues para fines de entendimiento es lo más práctico.

- La ejecución del programa, desde la línea de comandos es la siguiente:

Lo primero que sucede es la ejecución del comando “print(dataset.head(20))” que muestra los primeros veinte registros del conjunto de datos y esto se hace para verificar que ha quedado cargada de manera adecuada el conjunto de datos. Lo anteriormente descrito se puede revisar a detalle en la *Figura 9*.

```
------------------
PS D:\Tesis> python Tesis_v00_01.py
(8302, 7)
    id_operacion  dias_trx  num_empleado  mismo_origen_destino  id_monto  dias_antig_fecha_trx  fraudulento
0              1       898         25873                     0         1                  1173            0
1              1       893         29309                     0         1                   806            0
2              1       893         13398                     0         1                  2689            0
3              1       893          1849                     0         1                  4163            0
4              1       893         25873                     0         1                  1168            0
5              1       696         14455                     0         2                  2310            1
6              1       895         16743                     0         1                  2252            0
7              1       887         33684                     0         1                   368            0
8              1       888         25862                     0         1                  1166            0
9              1       895         25626                     0         1                  1195            0
10             1       889         35380                     0         1                   216            0
11             1       696         14455                     0         2                  2310            1
12             1       664         14455                     0         2                  2278            1
13             1       673         14455                     0         2                  2287            1
14             1       893         25862                     0         1                  1171            0
15             1       376         25783                     0         3                   662            0
16             1       887          2126                     0         1                  4124            0
17             1       468         25783                     0         3                   754            0
18             1       895         25603                     0         1                  1201            0
19             1       893         36472                     0         1                   108            0
```

Figura 9 - Primeros 20 registros del conjunto de datos

El comando “print(dataset.describe())”, del *Programa 1*, muestra las primeras estadísticas del conjunto de datos. Estas son (por cada variable o campo): la cantidad de registros con dato (para identificar datos perdidos), la media, la desviación estándar, el valor mínimo, los percentiles 25%, 50% y 75% así como el valor máximo. Lo descrito anteriormente se puede revisar estos datos a detalle en la *Figura 10*.

```
       id_operacion     dias_trx  num_empleado  mismo_origen_destino     id_monto  dias_antig_fecha_trx  fraudulento
count   8302.000000  8302.000000   8302.000000           8302.000000  8302.000000           8302.000000  8302.000000
mean       2.904842   487.402915  20194.568176              0.010359     1.601662           1332.056854     0.060467
std        0.425773   275.049395  10537.369549              0.101256     0.620660           1320.418447     0.238365
min        1.000000     1.000000     31.000000              0.000000     1.000000           -973.000000     0.000000
25%        3.000000   243.000000  12843.000000              0.000000     1.000000            295.000000     0.000000
50%        3.000000   478.000000  22462.000000              0.000000     2.000000            942.500000     0.000000
75%        3.000000   749.000000  28244.000000              0.000000     2.000000           2032.750000     0.000000
max        3.000000   898.000000  38411.000000              1.000000     3.000000           6247.000000     1.000000
```

Figura 10 - Estadísticas generales del conjunto de datos

El siguiente comando “print(dataset.groupby(‘fraudulento’).size)”, del *Programa 1*, genera la estadística señalando la cantidad de registros para cada valor de la característica `fraudulento` en el conjunto de datos. Esto último se puede ver a detalle en la *Figura 11*.

```
fraudulento
0    7800
1     502
dtype: int64
-----------------
```

Figura 11 - Estadística de la característica "fraudulento"

La siguiente instrucción ejecutada por el *Programa 1* genera un gráfico que se muestra en la *Figura 12*. Este es un gráfico de caja (también llamado gráfico de bigotes) y se emplea como representación gráfica de variables cuantitativas. Este gráfico permite resumir, describir y analizar aspectos generales y particulares de un indicador. En este gráfico quedan ilustrados los datos centrales, datos adyacentes y datos raros (atípicos y extremos, si los hubiera). Su preferencia se debe a que es, simultáneamente, una herramienta sencilla y rigurosa de exploración y análisis de una distribución cuantitativa y, porque, además, permite establecer, en el mismo gráfico, comparaciones entre grupos [21].

En la primera caja (de izquierda a derecha de la *Figura 12*) se puede ver el dato `id_operación`. Este dato no genera una caja porque contiene valores discretos. Así vemos que el menor número de muestras está en el valor uno (círculo) y la mayor cantidad de las muestras (línea) están en el valor 3.

Los campos `mismo_origen_destino` y `fraudulento` (cuarta y última caja, respectivamente, de izquierda a derecha de la *Figura 12*) se analizan de la misma manera ya que ambos son campos con datos discretos y sus correspondientes registros solo contienen dos valores posibles.

Por otro lado, también identificamos que el campo `dias_antig_fecha_trx` (sexta caja de izquierda a derecha de la *Figura 12*) tiene la mayor cantidad de sus datos entre los valores

poco superior a 0 (cero) y 2,000 (dos mil). Algunos datos fuera del percentil 95 (llamados *outliers*) están en el rango de los 5,000 (cinco mil) y 6,000 (seis mil).

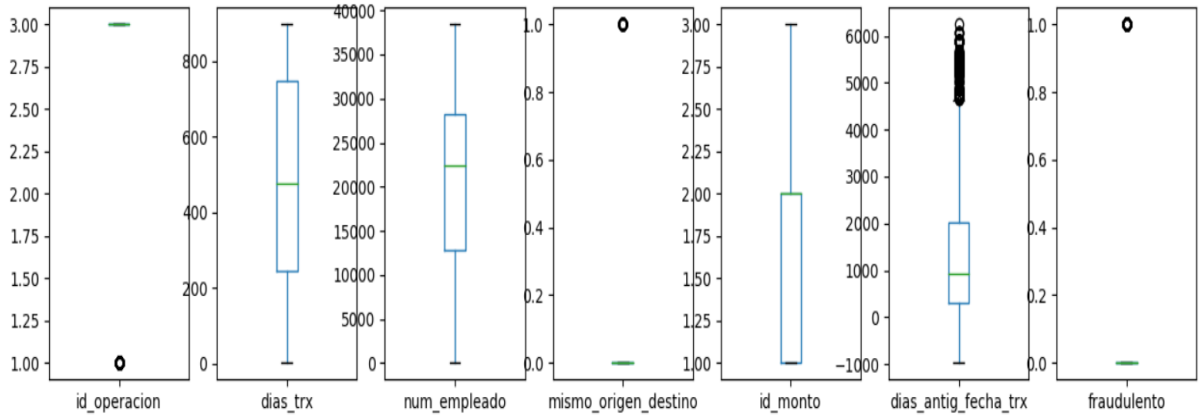

Figura 12. Diagrama de Caja de cada dato

El análisis sobre los campos `dias_trx`, `num_empleado` e `id_monto` (segunda, tercera y quinta caja, respectivamente, de izquierda a derecha de la *Figura 12*) es más sencillo de analizar pues la caja nos indica que prácticamente no tienen *outliers* y que los rangos del percentil 95 están perfectamente establecidos.

Adicionalmente, y siguiendo con el análisis de las variables que conforman el conjunto de datos, se utiliza una forma complementaria y distinta de ver la distribución de los valores, en cada uno de los campos del conjunto de datos; y ésta es por medio de gráficos en forma de histogramas.

Al ejecutar la siguiente línea de comando del programa nos muestra precisamente esta forma de visualización. Los histogramas nos ayudan a identificar los valores que tienen cada una de las tuplas del conjunto de datos.

Para el presente trabajo (y con fines de una mayor comprensión del conjunto de los datos utilizados), se ha preparado el *Programa 1* para separar cada variable y que pueda ser mostrado en un gráfico distinto e independiente. Para poder hacerlo así se usan los gráficos del tipo histograma. Esto, lo podemos ver representado en la *Figura 13*.

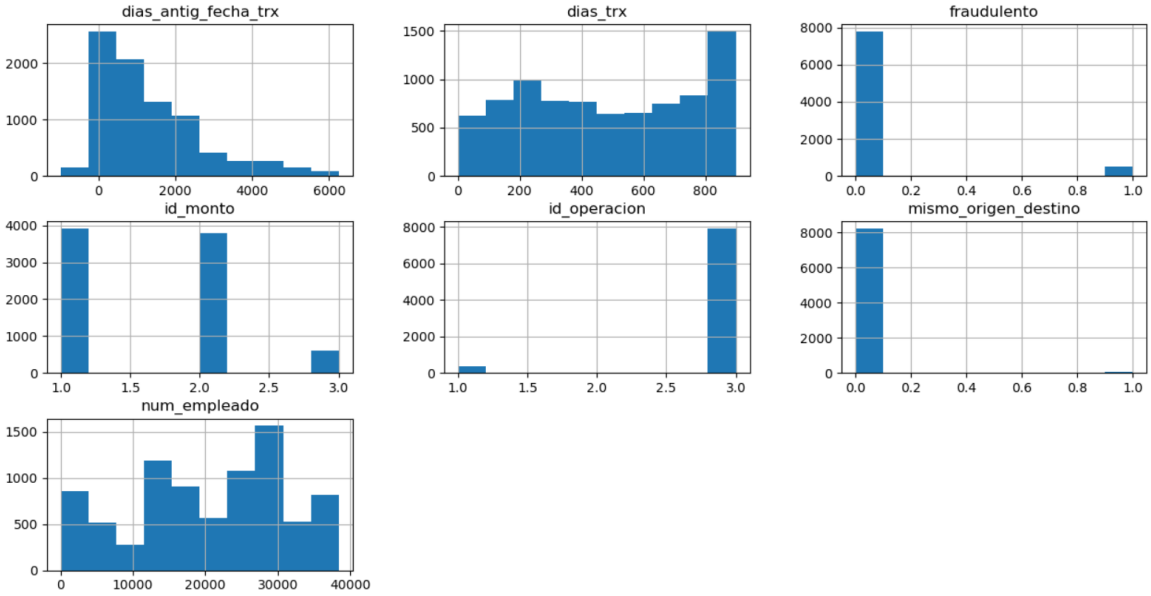

Figura 13. Histogramas de los datos del grupo de datos

La forma de leer, pero sobre todo interpretar, cada uno de los gráficos mostrados en la *Figura 13* es la siguiente:

- `dias_antig_fecha_trx` tiene una gran cantidad de tuplas (o eventos) con un valor de cero y conforme avanza el tiempo (el eje de las "x") se ve que existen menos tuplas. Los valores más grandes de esta variable es el seis mil.
- De la misma forma, la variable `dias_trx`, tiene una distribución más uniforme de los valores en sus tuplas. Presenta unos picos de mil y mil quinientas tuplas en los días doscientos y ochocientos respectivamente.
- Por otro lado, la variable `id_monto` se discretizó para este trabajo. En el histograma correspondiente se identifica en que clasificación se tienen las distintas tuplas del conjunto de datos.
- La variable `num_empleado` expresa su naturaleza al ser un dato secuencial que la Institución le asigna a cada empleado.
- En el caso de la variable `id_operacion`, la variable `mismo_origen_destino` y la variable `fradulento`, tienen la característica de ser variables discretas, se

puede identificar de una manera un poco más sencilla los valores más populares que conforman el conjunto de datos usados en este trabajo.

El *Programa 1* genera un tercer gráfico al momento de ejecutarse que tiene la característica de mostrar, en una sola matriz, todos los histogramas de las variables confrontados entre sí. Esto se puede ver a detalle en la *Figura 14*.

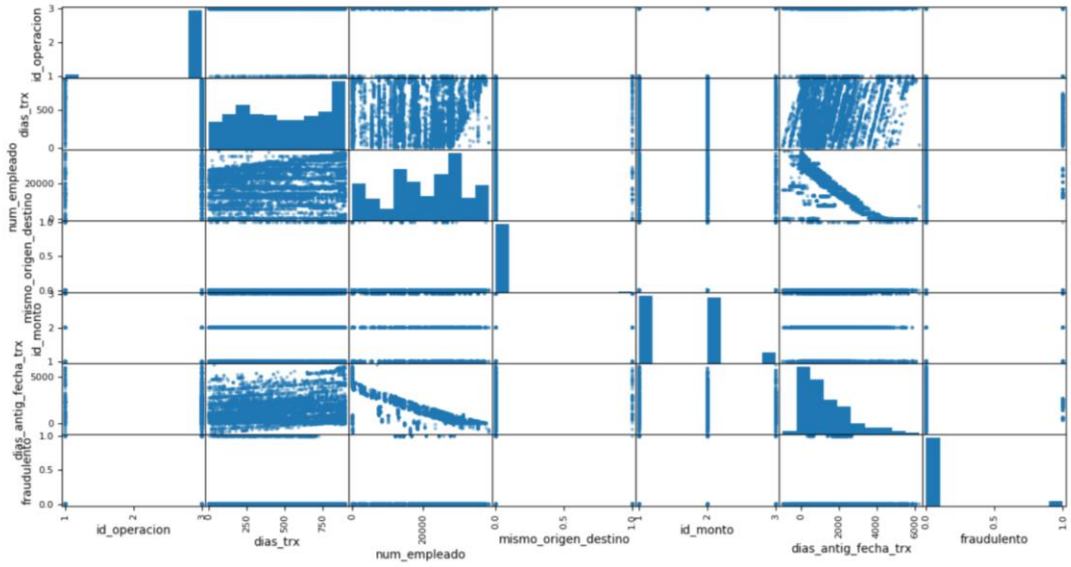

Figura 14. Matriz de histogramas confrontados

Donde se puede mostrar una matriz de confrontación de cada uno de los campos contra todos los demás. Con esta confrontación de campos se puede determinar, de una manera gráfica, la correlación que existe entre los distintos datos.

Se debe tomar la intersección renglón-columna para saber la forma en que están correlacionados los campos entre sí. Esta tabla la podemos ver como una matriz, entonces, como se puede apreciar, en la diagonal principal se confronta un campo contra sí mismo. La gráfica que se obtiene en la diagonal principal de la *Figura 14* es, y se corresponde, con el histograma del campo.

En la matriz de correlación de campos podemos identificar que no existe una correlación evidente entre los campos que conforman el conjunto de datos. Esto es bueno porque significa que no existe un dato mandatorio en el conjunto de datos y el conjunto de datos es apto para su uso en la investigación.

La siguiente tarea del *Programa 1* es determinar el desempeño de todos los algoritmos que forman parte de la evaluación de esta investigación. Los resultados arrojados se pueden ver en la *Figura 15*.

```
------------------
Method Logistic Regression: mean 0.939316 std(0.006907)
Method Linear Discriminant Analysis: mean 0.935250 std(0.007445)
Method K-Nearest Neighbors: mean 0.993977 std(0.002935)
Method Desicion Tree Classifier: mean 0.996085 std(0.001203)
Method Gaussian Naive Bayes: mean 0.940972 std(0.007908)
Method Support Vector Machine: mean 0.957386 std(0.005759)
------------------
```

Figura 15 - Desempeño de los algoritmos evaluados

El valor “mean”, que arroja el *Programa 1*, es el porcentaje de exactitud al procesar una nueva muestra después de haber sido entrenado el modelo.

Tabla XI - Tabla de desempeño de algoritmos evaluados

Algoritmo	Exactitud	Margen de error
Árboles de decisión (*DTC*, *Decisión Tree Classifier* en inglés)	99.61%	0.12%
K vecinos (*KNN*, *K-nearest Neighbors* en inglés)	99.40%	0.29%
Máquinas de Vectores de Soporte (*SVM*, *Support Vector Machine* en inglés)	95.74%	0.58%
Gauss Naive Bayes (*NB*, *Gaussian Naives Bayes* en inglés)	94.10%	0.79%
Regresión lineal (*LR*, *Logistic Regression* en inglés)	93.93%	0.69%
Análisis de discriminante lineal (*LDA*, *Linear Discriminant Analysis* en inglés)	93.53%	0.74%

En la *Tabla XI* se muestran los algoritmos evaluados con sus valores de desempeño, en específico la exactitud, ordenados de mayor a menor para poder identificar de mejor manera el desempeño de cada uno de ellos.

El resultado que se muestra en la *Tabla XI* es uno de los indicadores que lleva el mayor peso en las conclusiones de este trabajo. Lo anteriormente señalado se deriva de que es el primer indicio del camino que se debe seguir en obtener un valor agregado del conjunto de datos.

Al avanzar con la ejecución del *Programa 1* podemos observar el gráfico que se muestra en la Figura 16. Este es un gráfico con el formato de un diagrama de caja. Este gráfico sirve para confrontar en un solo lugar, y de manera visual, los resultados arrojados por todos los algoritmos evaluados y con el mismo conjunto de datos. La escala de exactitud está acoplada para mostrar lo mejor posible a cada algoritmo. Vemos que los algoritmos de Árboles de decisión y de *k-vecinos* son los que presentan los mejores resultados. En ambos casos su predicción se encuentra por arriba del 99%.

Y es en este punto en donde el *Programa 1* termina su ejecución.

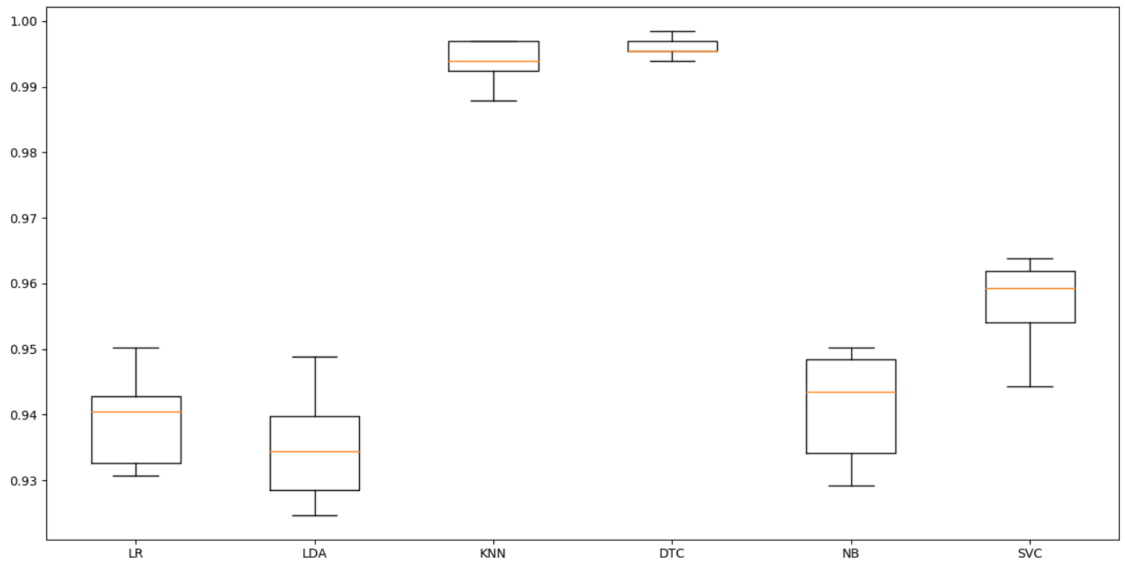

Figura 16. Diagrama de Caja comparando todos los algoritmos

Haciendo un resumen de lo expuesto hasta este momento y en continuidad con el análisis de los datos obtenidos se identifica que la variable número de empleado sobra en el conjunto de datos. Este es un dato que hace única a una tupla del tipo empleado. Por consiguiente, se excluye esta variable ya que tenemos una conclusión como la que sigue: "si el empleado tiene el número de empleado x entonces es un empleado fraudulento", conclusión totalmente incorrecta.

Se hicieron diferentes ejecuciones del *Programa 1* con la finalidad de revisar la consistencia de los resultados arrojados. Tras varios experimentos (ejecuciones del programa), excluyendo en cada uno de ellos diferentes variables, se obtuvieron diferentes resultados.

En la *Tabla XII* es en donde se muestra el resultado obtenido de cada uno de los experimentos mencionados. Cada columna de la *Tabla XII* es un experimento realizado y cada uno de ellos considera (o no) algún (o algunos) campo(s) con respecto al conjunto de datos original, como se detalla a continuación:

- Experimento sin considerar la variable `mismo_origen_destino`.
- Experimento sin considerar la variable `num_empleado` y considerando la variable `monto` que no se encontraba discretizada, es decir, en formato numérico puro (continuo).
- Experimento sin considerar la variable `num_empleado` y si considera la variable `monto` discretizado en nueve capas distintas.
- Experimento sin considerar la variable `num_empleado` y si considera la variable `monto` discretizado en diez capas distintas.

En la *Tabla XII* se pueden ver, en color verde los mejores resultados, y en color rojo los peores resultados. Con lo anteriormente expuesto se puede identificar, que también desde esta perspectiva, los algoritmos de Árboles de decisión y de K-vecinos son los más recomendables (aptos) para ahondar en la experimentación. Basado en esta conclusión es que se comienza la ejecución de los dos algoritmos (Árboles de decisión y k-vecinos) sobre el conjunto de datos

final con el objetivo de obtener un modelo final que ayude a confirmar la hipótesis de esta investigación.

Tabla XII - Tabla comparativa de desempeño por algoritmo y conjunto de datos

Método	Elemento	Layout original	Layout sin "mismo_origen_destino"	Layout sin "num_empleado" y "monto" no discretizado	Layout sin "num_empleado" y "monto" discretizado (9 capas)	Layout sin "num_empleado" y "monto" discretizado (10 capas)
Method Logistic Regression	mean	0.939316	0.939165	0.939165	0.939617	0.939768
	std()	0.006907	0.006472	0.006472	0.006754	0.006569
Method Linear Discriminant Analysis	mean	0.93525	0.934799	0.94022	0.936153	0.935551
	std()	0.007445	0.005913	0.007442	0.006151	0.006668
Method K-Nearest Neighbors	mean	0.993977	0.993977	0.959945	0.982231	0.982231
	std()	0.002935	0.002935	0.00763	0.004895	0.004895
Method Desicion Tree Classifier	mean	0.996085	0.995784	0.977112	0.980575	0.976058
	std()	0.001203	0.001312	0.897606	0.002816	0.004832
Method Gaussian Naive Bayes	mean	0.940972	0.940069	0.897606	0.930733	0.930281
	std()	0.007908	0.007057	0.008194	0.0071	0.007227
Method Support Vector Machine	mean	0.957386	0.954977	0.940822	0.952718	0.952115
	std()	0.005759	0.005889	0.006394	0.005686	0.005465
Multi Layer Perceptron	mean			0.939918	0.882087	0.89205
	std()			0.006652	0.116375	0.045424

Capítulo 4. Análisis de resultados

Para poder confirmar la hipótesis de que es posible identificar empleados del sector financiero que presentan un comportamiento sospechoso usando técnicas de aprendizaje automático se ha llegado a la conclusión de que los dos algoritmos que mejores resultados ofrecen para el conjunto de datos obtenido de la empresa objeto de estudio son los Árboles de decisión y el de k-vecinos.

En ambos casos estamos hablando de algoritmos enfocados a la clasificación de datos para formar reglas de comportamiento y formar grupos con características similares que definen a un empleado con comportamiento sospechoso.

En la *Figura 17* se muestra, solo de manera demostrativa y como punto de referencia inicial, el resultado del primer ejercicio con el algoritmo de Árboles de decisión. Este árbol se optimiza con este trabajo.

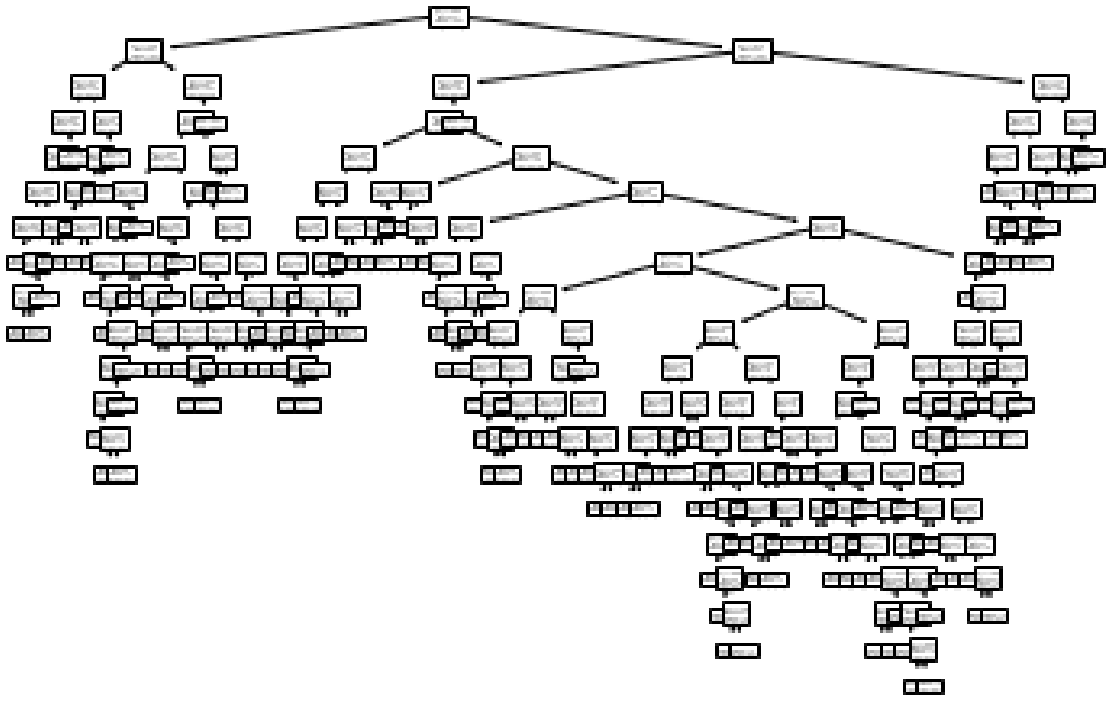

Figura 17. Primer árbol de decisión

Se trata de un árbol no óptimo por su tamaño (cantidad de reglas a administrar) y por consiguiente se desecha su uso, pero trabaja con la matriz de confusión (descrita su estructura a detalle en la sección 2.6.2.3 de esta misma investigación). En la *Tabla XIII* se puede observar los valores con los que cuenta la Matriz de confusión del algoritmo de Árboles de decisión para el conjunto de datos.

Tabla XIII - Matriz de confusión del algoritmo Árboles de decisión del conjunto original

		Clasificador	
		Negativos	Positivos
Valores reales	Negativos	1,507	58
	Positivos	41	55

Con la Matriz de confusión se pueden calcular los cinco indicadores de desempeño sobre los cuales se basará la decisión de una correcta implementación. Estos indicadores se pueden ver en la *Tabla XIV*, y están calificados con un código de colores. Los valores adecuados se marcan en color verde, los valores aceptables se marcan en color amarillo y los valores no aceptables se marcan en color rojo.

Tabla XIV - Indicadores con el algoritmo de Árboles de decisión del conjunto original

Concepto	Variable	Valor %
Exactitud:	Ac	94.04%
Razón de verdaderos positivos:	TPrate	57.29%
Razón de falsos positivos:	FPrate	3.71%
Razón de verdaderos negativos:	TNrate	96.29%
Razón de falsos negativo:	FNrate	42.71%

Para la investigación objeto de este trabajo, tres de los indicadores son adecuados (color verde). De la misma forma, dos de los indicadores se consideran no aceptables ya que están fuera de los parámetros para considerarlos aceptables (color rojo). Los indicadores que preocupan son los verdaderos positivos y los falsos negativos cuya importancia es alta derivado de que el algoritmo clasifica a pocos empleados que si son fraudulentos (verdaderos positivos) y al mismo tiempo clasifica como fraudulentos a muchos empleados que no lo son (falsos negativos).

Con los resultados anteriores se concluye que se debe trabajar con más detalle en el grupo de datos para encontrar mejores indicadores. Esto último provoca que se continúen los experimentos con la herramienta gráfica *Analysis Services*.

En la *Tabla XV* se pueden revisar los umbrales para los indicadores de desempeño del tipo verdadero.

Tabla XV - Clasificación para indicadores Verdaderos

	Adecuado	Aceptable (análisis)	No aceptable
Exactitud (Ac) % Verdaderos Positivos % Verdaderos Negativos	**Más del 85%**	**Entre 70% y 85%**	**Menor del 70%**

Por su parte en la *Tabla XVI* se pueden revisar los umbrales para los indicadores de desempeño del tipo falso.

Tabla XVI - Clasificación de indicadores Falsos

	Adecuado	Aceptable (análisis)	No aceptable
% Falsos Positivos % Falsos Negativos	**Menos del 10%**	**Entre 10% y 30%**	**Mayor del 30%**

Sobre los criterios anteriores es sobre los que se basarán los resultados y conclusiones de la presente investigación.

4.1 Resultados

Para efectos de resumen de lo expuesto hasta el momento, se menciona que, de un conjunto grande de datos, que hizo disponibles la empresa objeto de estudio, se hizo una depuración para usar solo aquellos que pudieran representar un valor en su procesamiento. Este proceso viene de una fase intensa de limpieza de datos.

La lógica utilizada para la creación de este grupo de datos es la siguiente:

- Solo se analizan operaciones realizadas en las cuentas de débito de los Clientes
- El conjunto de datos contiene operaciones realizadas entre el 01/Ene/2017 y el 30/Jun/2019; es decir 18 meses de operación
- A este conjunto de datos se le llamó “DATASET_20191127”

Cabe señalar que a partir de esta parte del proceso se utiliza la herramienta *SQL Server* 2019 con los módulos de Analítica de datos instalados (*Analysis Services*) y configurados para hacer mejores representaciones gráficas de los resultados. Con estas representaciones gráficas es como se puede entender e interpretar de una manera más sencilla los resultados obtenidos.

4.1.1 Estructura de datos utilizada

Inicialmente se muestra, en la *Figura 18*, la estructura del conjunto de datos:

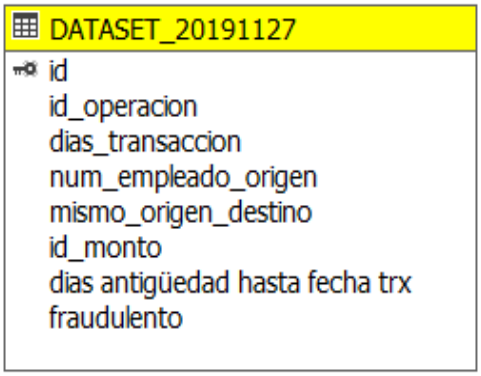

Figura 18 - Estructura del conjunto de datos

Aunque ya se ha tratado el tema de los datos en secciones anteriores, se hace una remembranza junto con algunos puntos a considerar que se convierten en relevantes por la herramienta que es usada para su tratamiento.

Se parte del hecho de que los algoritmos de aprendizaje automático requieren de dos tipos de variables. Por un lado, las variables de entrada que son las características que definen una situación u objeto; y por el otro las variables de salida que son los valores a predecir. A las

primeras variables (de entrada) se les denomina independientes y a las segundas (de salida) dependientes. En esta investigación se llegó a determinar que existen seis variables independientes (que son las características que definen a un empleado) y una variable independiente (que es la que, a través de los algoritmos de aprendizaje automático se desea calcular su valor siendo este último una clasificación de si un empleado tiene perfil de fraudulento o no a partir de las características que lo definen).

Es necesario comentar que existe una séptima variable independiente, el `id` que a su vez es el valor único de cada registro, que fue generado por la necesidad de hacer completamente único cada registro. Esto es porque las características de los registros en el conjunto de datos llegaban a repetirse en valor entre los distintos empleados. Esta técnica (agregar este valor secuencial único) ayuda a los algoritmos a operar de una mejor manera en el tratamiento de su información.

Se hace un resumen con la descripción de cada uno de los campos, y sus consideraciones particulares para cada caso:

- `id`. Esta es una variable independiente en el conjunto de datos. Es un dato que fue generado para su procesamiento por los algoritmos matemáticos, es decir, no está en la Base de Datos original. Su uso tiene como objetivo hacer único a cada uno de los registros en el conjunto de datos.
- `id_operacion`. Esta es una variable independiente en el conjunto de datos. Es un dato que representa el tipo de operación que se realiza en la Institución objeto de estudio. Para este conjunto de datos se están considerando dos tipos distintos de transacciones, la clave usada está entre paréntesis:
    - (1) para Transferencias entre cuentas
    - (3) para las Disposiciones de una cuenta, es decir, retiros de dinero de la cuenta
- `dias_transaccion`. Esta es una variable independiente en el conjunto de datos. Es la cantidad de días que han transcurrido a partir del 01/Ene/2017 a la fecha en que fue generada la transacción. Esto quiere decir que, si una transacción fue generada,

por ejemplo, el 02/Ene/2017 en este campo tendrá el número 1 pues solo transcurrió un día con referencia al 01/Ene/2017. Este dato fue creado, es decir no está de esta manera en la Base de Datos original, para poder tener un dato continuo que represente el paso del tiempo.

- `num_empleado_origen`. Esta es una variable independiente en el conjunto de datos. Es el número de empleado, que se tiene registrado en el sistema, para la persona que hace la transacción.
- `mismo_origen_destino`. Esta es una variable independiente en el conjunto de datos. Es un indicador con dos posibles valores, la clave usada está entre paréntesis:
    - (0) significa que el empleado que hizo la transacción la hizo sobre una cuenta que NO esté a su nombre
    - (1) significa que el empleado que hizo la transacción la hizo sobre una cuenta que SI está a su nombre
- `id_monto`. Esta es una variable independiente en el conjunto de datos. Es un dato que contiene el monto de la transacción. Para hacer manejable el dato se generaron los siguientes rangos para los distintos montos, la clave usada está entre paréntesis:
    - (1) cuando la operación es por un monto menor a los 500 pesos
    - (2) cuando la operación es por un monto que está entre los 500 y los 8,000 pesos
    - (3) cuando la operación es por un monto superior a los 8,000 pesos
- `dias_antiguedad_hasta_fecha_trx`. Esta es una variable independiente en el conjunto de datos. Es la cantidad de días de antigüedad que tiene un empleado en la empresa cuando se realiza la transacción que se está analizando.
- `fraudulento`. Esta es la variable dependiente en el conjunto de datos. Este dato no está de esta forma en la Base de Datos original. A partir de un control manual que se tiene dentro de la empresa objeto de estudio se definió un valor binario para identificar si una transacción está relacionada a un empleado fraudulento o no:

- (0) para indicar que la transacción NO está relacionada con un empleado identificado como fraudulento dentro de la Institución objeto de estudio.
- (1) para indicar que la transacción SI está relacionada con un empleado identificado como fraudulento dentro de la Institución objeto de estudio.

4.1.2 Análisis con el algoritmo de Árboles de decisión

Se indica en este apartado los resultados del procesamiento que se tuvo del conjunto de datos con el algoritmo de Árboles de decisión. La configuración de los datos para ejecutar el algoritmo se detalla en la *Figura 19*.

Estructura	DATASET_20191127_AD
	Microsoft_Decision_Trees
Dias Antigüedad Hasta Fecha Trx	Input
Dias Transaccion	Input
Fraudulento	Predict
Id Monto	Input
Id Operacion	Input
Id	Key
Mismo Origen Destino	Input
Num Empleado Origen	Omitir

Figura 19 - Configuración del conjunto de datos para el algoritmo de Árboles de decisión

Se identifican, en la figura anterior, las variables configuradas como independientes (*Input*) y la variable dependiente (*Predict*) que en este caso es la clase fraudulento. Con los primeros experimentos se dejó fuera (Omitir) una variable (el número de empleado que origina la transacción) que no era relevante en el proceso (como lo hemos comentado a lo largo de este escrito).

El algoritmo generó un árbol de seis niveles de profundidad, como se muestra en la *Figura 20*, en este punto de la investigación se vuelve más importante la profundidad del árbol derivado de que veníamos de uno altamente complejo.

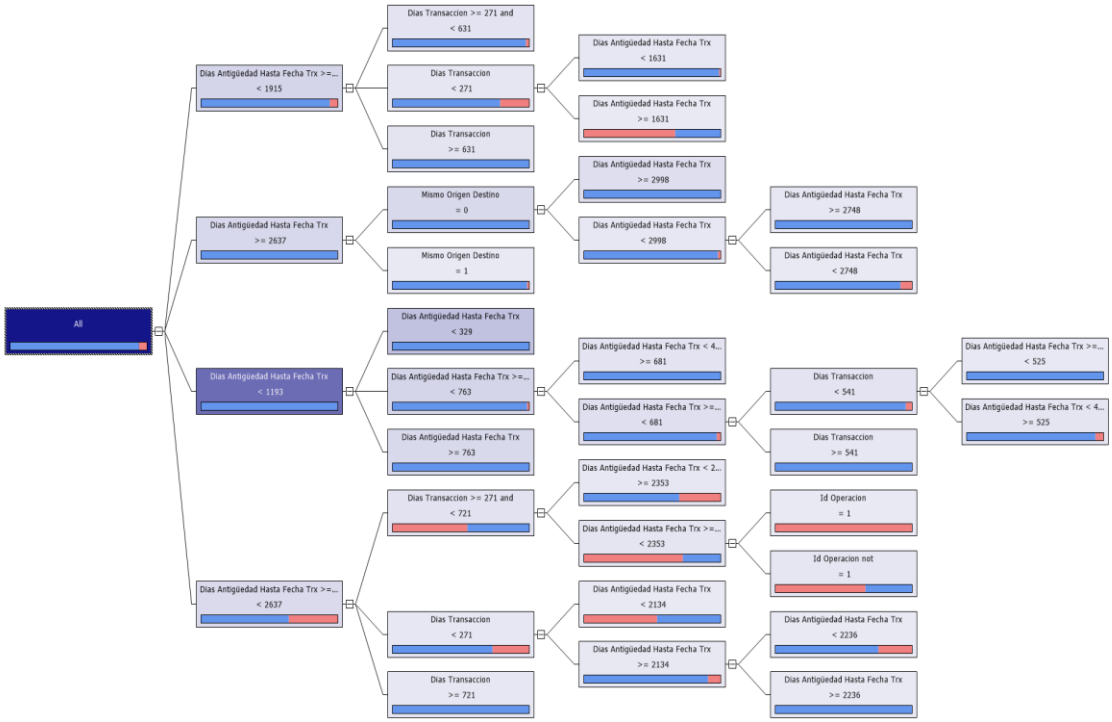

Figura 20. Árbol completo de grupo de datos final

En la *Figura 21* se puede ver a detalle las primeras estadísticas del grupo de datos que arroja el proceso de análisis.

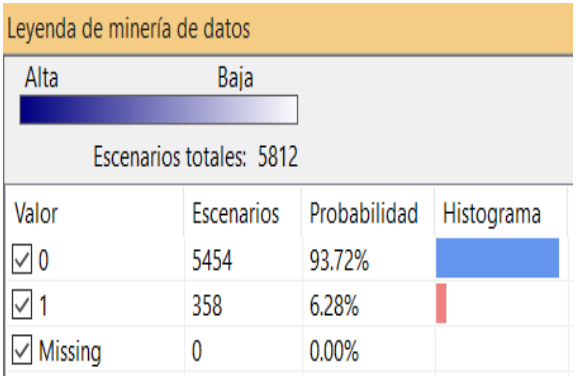

Figura 21. Árboles de decisión, estadísticas iniciales

Vemos que el conjunto de datos está compuesto por 5,812 casos (transacciones o escenarios) de los cuales el 6.16% corresponden a casos marcados como fraudulentos. También podemos

observar que no existen transacciones que estén sin una categorización (*missing* o perdido) en el conjunto de datos.

En la *Figura 22* se puede ver de manera visual que la variable más importante (de más peso) es "Días Antigüedad Hasta Fecha Trx" (primer nivel del árbol):

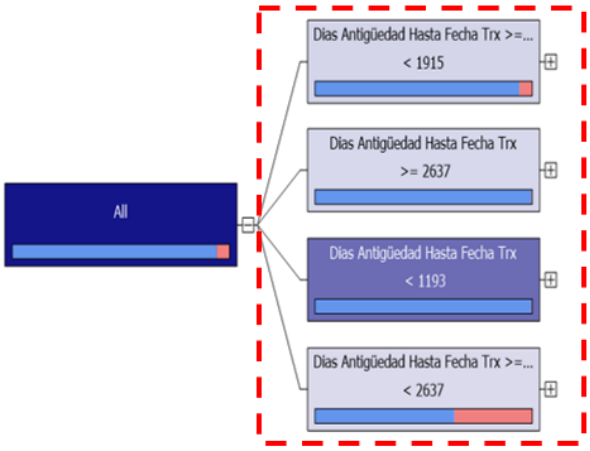

Figura 22. Árbol de decisión, primer nivel del árbol

Y es, a partir de este nivel, que se puede ir identificando la distribución de los casos marcados como fraudulentos en el conjunto de datos.

La rama que el algoritmo señala como de más peso (y que se puede identificar de manera visual por tener un color más fuerte) es en donde se lleva el conteo del acumulado de días de antigüedad de un empleado hasta la fecha en que es realizada la transacción, y que para este caso tiene un valor menor a 1,193 días. Lo anteriormente descrito, se puede ver con detalle en la *Figura 23*.

Sin embargo, vemos que esta rama solo se concentran 18 casos marcados como fraudulentos; y esto solo representa el 5.03% (18/358) del total de casos.

Haciendo la misma analogía e incluyendo la totalidad de las ramas del árbol de decisión se llega a los resultados mostrados en la *Tabla XVII*.

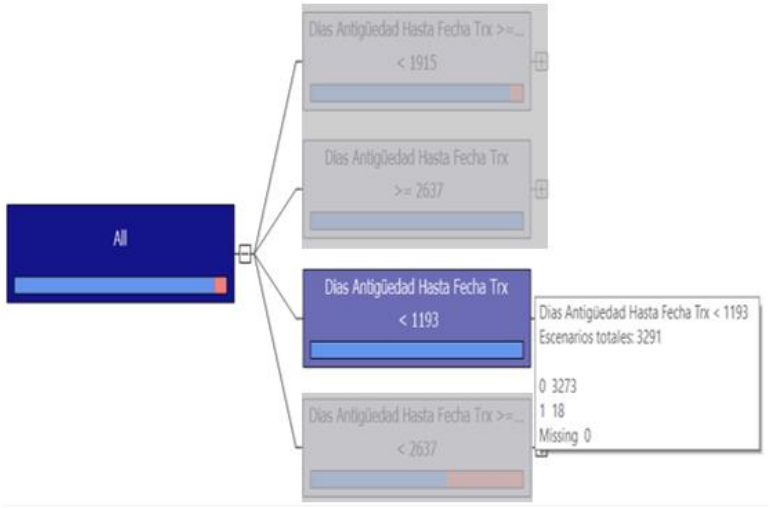

Figura 23. Árbol de decisión, rama de más peso

Tabla XVII - Análisis de casos fraudulentos por rama (Árboles de decisión)

Rama	Probabilidad	Casos donde fraudulento es 1	Casos donde fraudulento es 0	Días antiguedad hasta la fecha de la transacción
1	36.22%	277	488	>= 1915 and < 2637
2	6.33%	59	878	>= 1193 and < 1915
3	0.73%	4	815	>= 2637
4	0.58%	18	3273	< 1193

Si hacemos la selección de las dos primeras ramas, mostradas en la *Tabla XVII*, que son las que tienen mayor probabilidad de contener casos fraudulentos, estamos abarcando el 93.85% de la muestra de casos fraudulentos; y delimitamos los días de antigüedad al rango que va de 1,193 a 2,637. Conviene, entonces, hacer el análisis detallado, para la obtención de reglas sobre estas dos ramas.

Comenzando con la primera rama, se expandirán a todos los niveles disponibles y se hará el análisis en las hojas finales. Solo se tomará en cuenta a las hojas finales que presenten casos marcados como fraudulentos. Lo anteriormente expuesto se puede ver a detalle en la *Figura 24*, y se debe aclarar que se sigue un estricto orden de aparición en el gráfico que arroja el proceso de análisis.

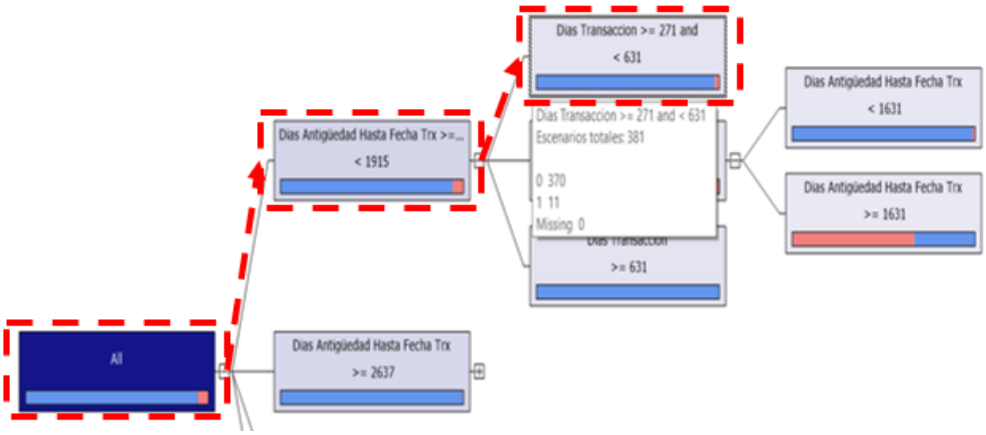

Figura 24. Árbol de decisión, análisis de rama completa

Las reglas que se cumplen en el nodo resaltado en la *Figura 24* son:

1) Días de antigüedad del empleado a la fecha de la transacción entre 1,193 y 1,194 días; y

2) Días transcurridos desde el 01/Ene/2017 a la fecha de ejecución entre 271 y 630 días.

Son 11 casos marcados fraudulentos y la probabilidad que arroja el algoritmo es de 2.91%.

Tabla XVIII - Árbol de decisión, tabla de probabilidades

Número	Probabilidad	Casos donde fraudulento es 1	Casos donde fraudulento es 0	Días transcurridos hasta la fecha de la transacción	Días antiguedad hasta la fecha de la transacción	Operación	Mismo Origen que Destino
1	99.99%	47	0	>= 271 and <721	>= 2,061 and	1	NA
2	66.66%	46	23	<271	>= 1,631 and	NA	NA
3	65.50%	131	69	>= 271 and <721	>= 2,061 and	<> 1	NA
4	53.03%	35	31	<271	>= 1,915 and	NA	NA
5	30.57%	55	125	>= 271 and <721	>= 1915 and < 2061 or >= 2353 and < 2637	NA	NA
6	25.00%	9	27	<271	>= 2,134 and	NA	NA
7	8.90%	4	41	NA	>= 2,637 and	NA	0
8	6.32%	18	267	<541	>= 417 and <471 or >= 525 and <681	NA	NA
9	2.91%	11	370	>= 271 and <631	>= 1,193 and	NA	NA
10	1.32%	2	151	<271	>= 1,193 and	NA	NA

En la *Tabla XVIII* se muestra el resumen de los 10 casos existentes en el árbol (ordenados del más probable al menos probable).

Al continuar con el análisis se identifica como la variable importante a la antigüedad del empleado cuando se realiza una transacción. Si se realiza una línea del tiempo se obtiene u resultado como el que se muestra en la *Tabla XIX*.

Tabla XIX - Árbol de decisión, línea del tiempo

		Línea del tiempo (días) y rangos que considera el variable "Días de Antiquedad hasta la fecha de transacción"															
Número	**Días antiguedad hasta la fecha de la transacción**	**0**	**417**	**471**		**525**	**681**		**1,193**	**1,631**	**1,915**	**2,061**	**2,134**	**2,236**	**2,353**	**2,637**	**2,748**
1	>= 2,061 and <2,353																
2	>= 1,631 and <1,915																
3	>= 2,061 and <2,353																
4	>= 1,915 and <2,134																
5	>= 1915 and < 2061 or >= 2353 and < 2637																
6	>= 2,134 and <2,236																
7	>= 2,637 and <2,748																
8	>= 417 and <471 or >= 525 and <681																
9	>= 1,193 and <1,915																
10	>= 1,193 and <1,631																
Línea del tiempo resultante:		417			54			512									

En la *Tabla XIX* se identifican muy pocos espacios, en la totalidad de la línea de tiempo planteada, en los que no tiene influencia la variable antigüedad del empleado hasta la fecha de la transacción. La forma de leer la tabla es que la variable en mención está marcada con color gris en los espacios de la línea del tiempo donde tiene influencia.

Para obtener un análisis más completo se deben considerar otras perspectivas, como lo es el gráfico de elevación de minería de datos (que se muestra en la *Figura 25*), en donde se compara el modelo ideal que se puede obtener con el proceso ejecutado y comparado con el modelo que se ha obtenido.

Al comparar, en el punto de Población General (que es el eje de las x en el gráfico de elevación) en el punto que marca el rango del 50%, se tiene una Población correcta del 50% (es decir, sin desviación entre el modelo ideal y el modelo obtenido) y una probabilidad de predicción del 99.98%. Con lo anteriormente descrito se puede mencionar que se tiene una muy alta probabilidad de predicción acertada.

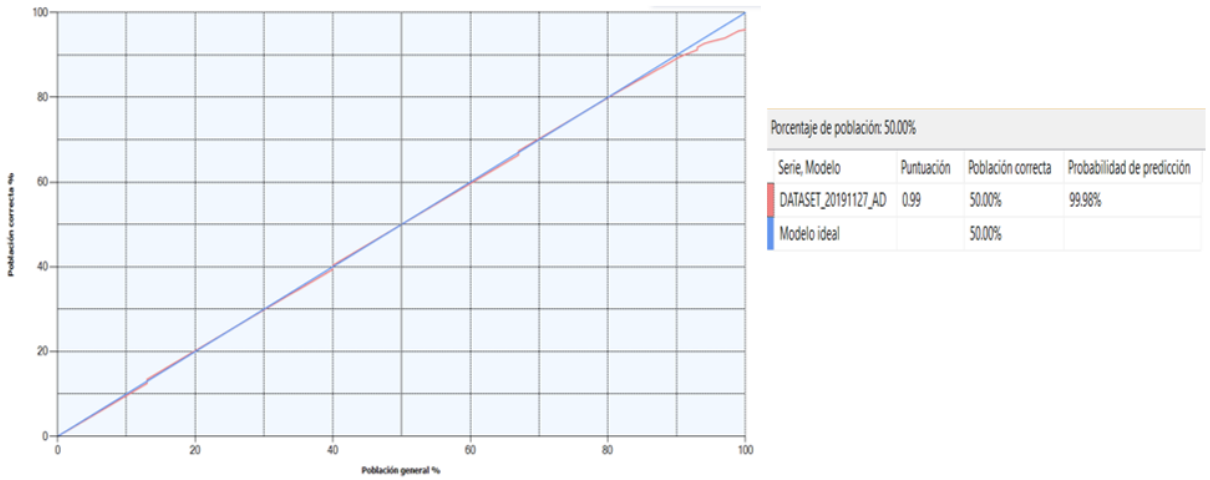

Figura 25. Árbol de decisión, gráfico de elevación

Con la misma lógica se hacen distintos cortes basados en los datos que arroja la gráfica de la *Figura 25* (que se muestran en la *Tabla XX*). Estos cortes son intervalos de 5 unidades, para el rango que va del 50% al 100%, que es en donde se identifica que comienza a degradarse la probabilidad de predicción.

Analizando los resultados vertidos en la *Tabla XX*, se identifica un punto de inflexión en el 79% del modelo ideal; ahí es donde comienza una degradación importante de la probabilidad hasta llegar al 53.03% con el modelo ideal del 100%.

Tabla XX - Árbol de decisión, tabla de probabilidades de corrección

Modelo ideal	Población correcta	Probabilidad de corrección
50%	80.00%	99.98%
55%	54.82%	99.98%
60%	59.64%	99.98%
65%	64.46%	99.98%
70%	70.24%	99.95%
75%	75.06%	99.95%
79%	78.92%	98.68%
80%	79.88%	97.09%
85%	84.50%	97.09%
90%	89.16%	93.68%
95%	93.17%	66.66%
100%	95.98%	53.03%

Cuando el análisis nos lleva a revisar la matriz de confusión de este nuevo conjunto de datos vemos los resultados reflejados en la *Tabla XXI*.

Tabla XXI - Matriz de confusión - Árboles de decisión del conjunto de datos final

		Clasificador	
		Negativos	Positivos
Valores reales	Negativos	2,290	44
	Positivos	56	100

Tabla XXII - Indicadores - Árboles de decisión del conjunto de datos final

Concepto	Variable	Valor %
Exactitud:	Ac	95.98%
Razón de verdaderos positivos:	TPrate	64.10%
Razón de falsos positivos:	FPrate	1.89%
Razón de verdaderos negativos:	TNrate	98.11%
Razón de falsos negativo:	FNrate	35.90%

Donde, por los umbrales definidos para esta investigación, se ve rendimiento superior al del conjunto de datos original mostrados en la *Tabla XIV*, lo cual lo hace adecuado y aceptable (con análisis) en tres de los cinco indicadores. En contraparte se ve que aún se tiene un rendimiento no suficientemente bueno para clasificar Verdaderos Positivos y Falsos Negativos. Sin embargo, estos indicadores no aceptables, muestran un mejor desempeño que el que mostraron en el conjunto original.

4.1.3 Análisis con el algoritmo K-vecinos

Se indica en este apartado los resultados del procesamiento que se tuvo del conjunto de datos final con el algoritmo de los K-vecinos. Para efectos de contextualizar este apartado se hace mención de que la configuración de los datos para ejecutar el algoritmo. Dicha configuración se muestra a detalle en la *Figura 26*.

Estructura	DATASET_20191127_CL
	Microsoft_Clustering
Dias Antigüedad Hasta Fecha Trx	Input
Dias Transaccion	Input
Fraudulento	Predict
Id Monto	Input
Id Operacion	Input
Id	Key
Mismo Origen Destino	Input
Num Empleado Origen	Omitir

Figura 26 - Configuración del conjunto de datos para el algoritmo de los K-vecinos

En la *Figura 26* se identifican las variables del conjunto de datos que están configuradas como variables independientes (*Input*) y la variable dependiente (*Predict*). Con los primeros experimentos se dejó fuera (Omitir) una variable (el número de empleado que origina la transacción) que no era relevante en el proceso.

El algoritmo generó un gráfico con diez grupos (o clúster) distintos, señalando con las tonalidades más fuertes (oscuras) sobre el color azul, los más importantes. También marca las diferentes relaciones entre los grupos, unas más fuertes que otras, también identificadas con distintos tonos de grises, como se muestra en la *Figura 27*.

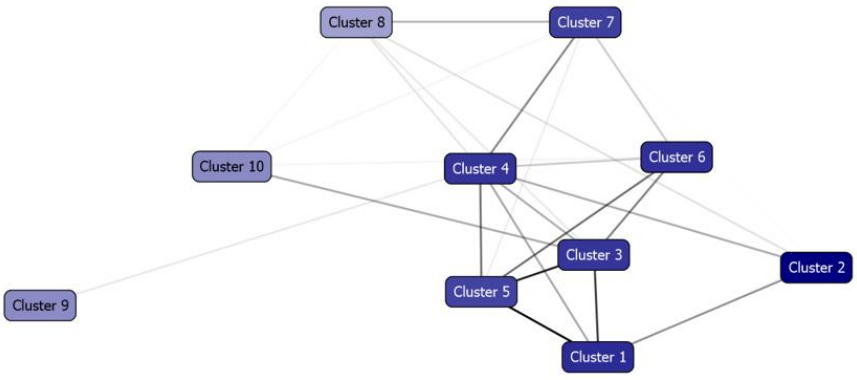

Figura 27. K-vecinos, grupos generados por el algoritmo

En la *Figura 27* se visualiza la configuración general de los grupos generados por el algoritmo sobre el conjunto de datos. El peso o importancia en el grupo lo define la intensidad del color con el que se ilumina el grupo. Y adicionalmente se puede determinar fácilmente la relación que existe entre todos los grupos del conjunto de datos.

Al colocar como el parámetro más importante del gráfico, con las facilidades de la herramienta gráfica, a la variable dependiente (`fraudulento`) con el valor de uno (que es una transacción calificada como fraudulenta) se genera el gráfico de la *Figura 28* que nos marca al grupo 9 como el más importante

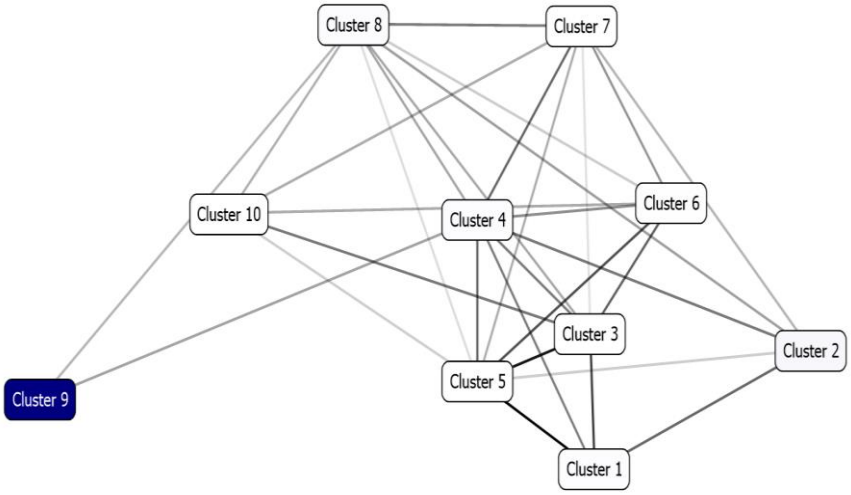

Figura 28. K-vecinos, grupo 9

Al avanzar en el análisis se genera la gráfica de la Figura 29 donde se marcan las relaciones más importantes del grupo 9, con la variable fraudulento en valor uno. El gráfico señala que los grupos 4 y 8 son los que mayor relación tienen con el grupo 9.

La misma herramienta gráfica nos permite visualizar el perfil del grupo, junto con todos los demás grupos generados. El perfil completo del grupo se ve a detalle en la *Figura 30*.

Cuando se revisa a detalle la *Figura 30* se confirma que el grupo 9 es el más importante con referencia a la variable dependiente (`fraudulento`) marcada con el valor de 1. En el primer

renglón de la matriz, de dicha *Figura 30*, se identifica a la variable dependiente y se observa que ese renglón se cruza con cada grupo distinto. En el cruce donde se unen la variable dependiente `fraudulento` y el grupo nueve se ve la distribución de casos donde `fraudulento` es uno. Sin duda alguna es el grupo que contiene el mayor número de casos.

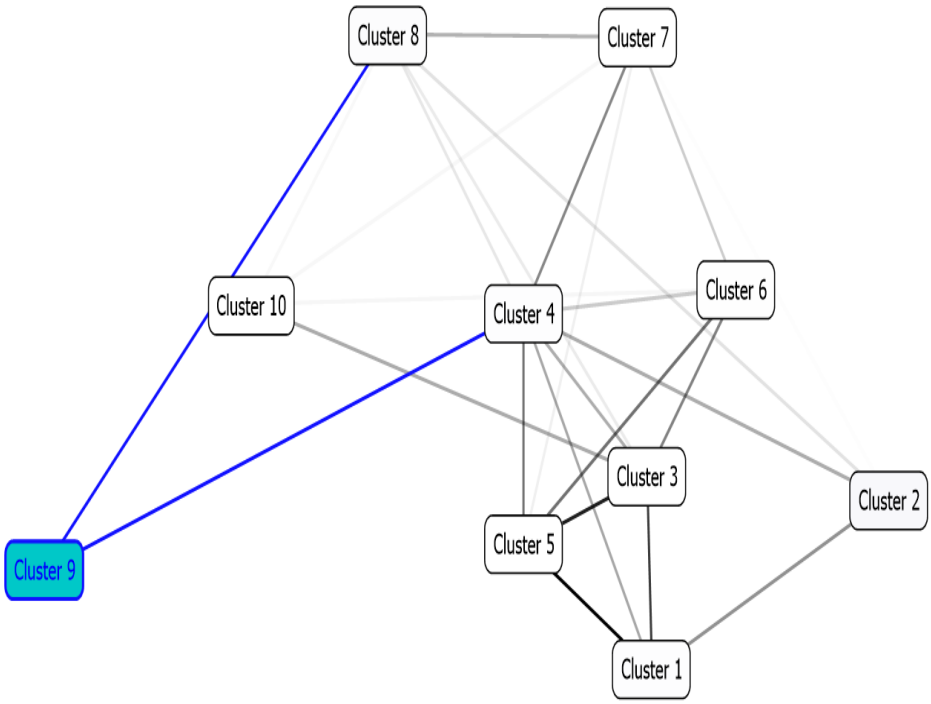

Figura 29. K-vecinos, grupo 9 y sus relaciones con otros grupos

Los detalles de la composición de este grupo nueve son los siguientes (en referencia a la *Figura 30*):

- Distribución:
  - Casos donde la variable Fraudulento es cero: 23.6%
  - Casos donde la variable Fraudulento es uno: 76.4%
- Las características (reglas) que determinan la composición del grupo, son:
  - Fraudulento = 1
  - Dias Antigüedad Hasta Fecha Trx entre 1,697 y 2,644
  - Id Operación = 1

- Dias Transaccion entre 29 y 856
- Id Monto = 2
- Mismo Origen Destino = 0

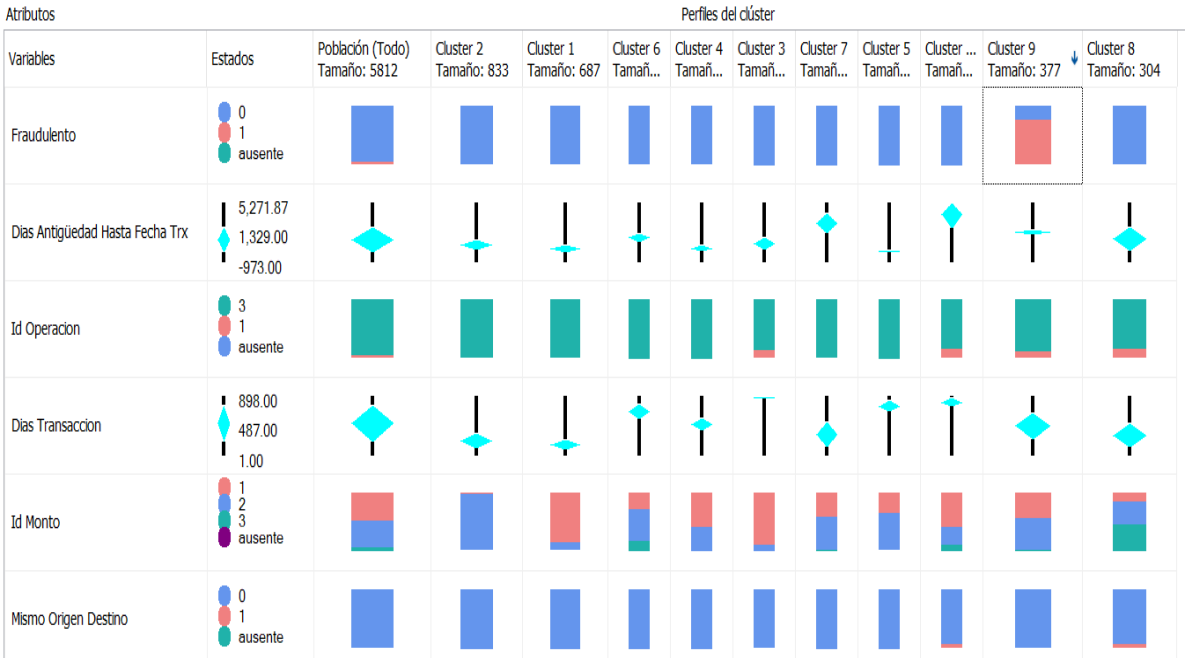

Figura 30. K-vecinos, perfil de todos los grupos

Por otro lado, el detalle de la composición del grupo ocho es la siguiente (en referencia a la *Figura 30*):

- Distribución:
  - Casos donde la variable Fraudulento es cero: 99.4%
  - Casos donde la variable Fraudulento es uno: 0.6%
- Las características (reglas) que determinan la composición del grupo, son:
  - Dias Antigüedad Hasta Fecha Trx <= -973
  - Id Monto = 3
  - Mismo Origen Destino = 1
  - Id Operación = 1
  - Dias Transaccion entre 1 y 659
  - Fraudulento = 0

Finalmente, el detalle de la composición del grupo cuatro es la siguiente:

- Distribución:
  - Casos donde la variable Fraudulento es cero: 99.6%
  - Casos donde la variable Fraudulento es uno: 0.4%
- Las características (reglas) que determinan la composición del grupo, son:
  - Dias Transaccion entre 270 y 659
  - Dias Antigüedad Hasta Fecha Trx entre -72 y 1,032
  - Id Monto = 1
  - Fraudulento = 0
  - Id Operación = 3
  - Mismo Origen Destino = 0

Revisando a detalle la composición y las características de los tres grupos más sobresalientes, se descartan los grupos cuatro y ocho por contar con un bajo número de casos con la variable `fraudulento` igual a uno.

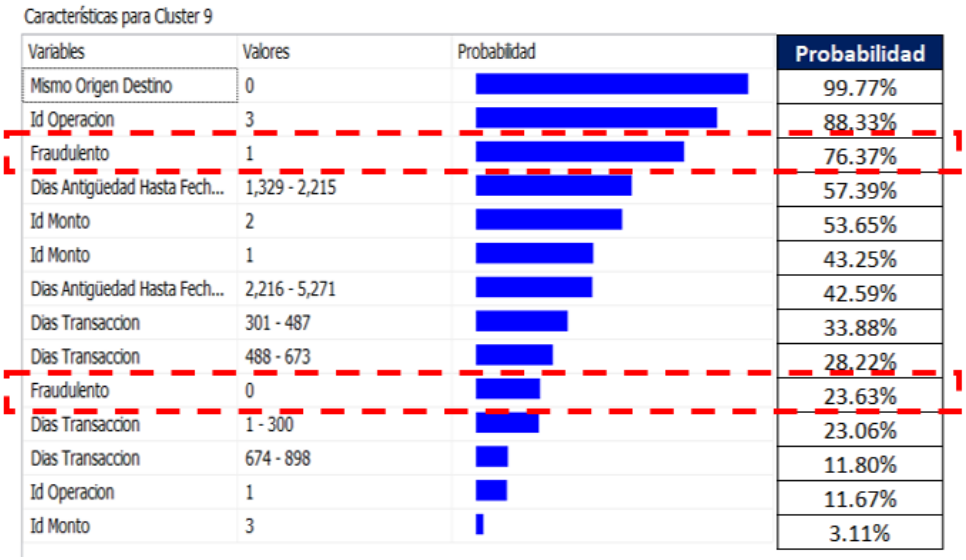

Figura 31. K-vecinos, características del grupo 9

Basado en lo comentado en los párrafos anteriores, se avanza en el detalle y se inicia la observación más detenida sobre las características del grupo nueve. Para este análisis detallado

se hace uso del gráfico mostrado en la *Figura 31* en donde se muestran las probabilidades de pertenencia al grupo para cada una de las variables del conjunto con sus valores probables.

Cuando usamos una opción de la herramienta para hacer un comparativo entre las variables que representan un mayor peso en la composición del grupo nueve podemos observar lo que se muestra en la *Tabla XXIII*.

Tabla XXIII - K-vecinos, análisis individual del grupo 9

Variables	Valores	Favorece Cluster 9	Favorece Complemento de Cluster 9
Fraudulento	1	100	
Fraudulento	0		100
Id Operacion	3		1.775
Id Operacion	1	1.775	

La *Tabla XXIII* indica que para tener mayor certeza en la determinación de que un caso es fraudulento se debe acompañar del tipo de operación en uno. En otra circunstancia, es decir, con el tipo de operación igual a 3 se puede catalogar, con mayor certeza, como fraudulento igual a cero (es decir, no es fraudulento).

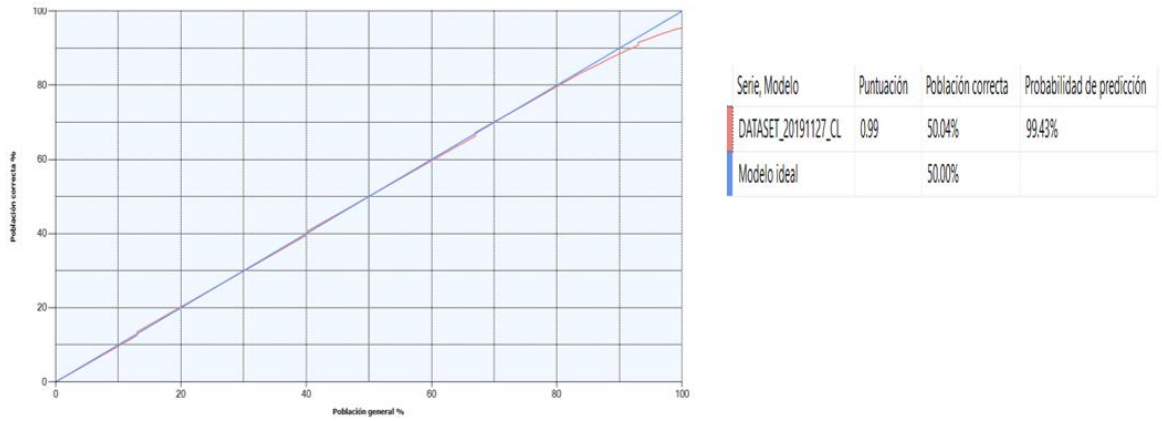

Figura 32. K-vecinos, gráfico de elevación

Se sigue con el análisis el grupo nueve desde otra perspectiva, como lo es el gráfico de elevación de minería de datos, en donde se compara el modelo ideal con el modelo obtenido. En la *Figura 32* se puede observar el gráfico de elevación de minería de datos.

Al comparar, en la *Figura 32*, el punto de Población General en el rango del 50%, se tiene una Población correcta del 50% (es decir, una desviación mínima comparado con el modelo ideal) y una probabilidad de predicción del 99.43% (que es alta).

Tabla XXIV - K-vecinos, tabla de probabilidades de corrección

Modelo ideal	Población correcta	Probabilidad de corrección
50%	50.04%	99.43%
55%	54.82%	99.42%
60%	59.64%	99.35%
65%	64.38%	99.05%
70%	70.16%	98.56%
75%	74.82%	98.36%
80%	79.64%	97.56%
85%	84.26%	92.23%
90%	88.47%	76.30%
95%	92.73%	58.33%
100%	95.46%	50.01%

Cuando hacemos la distinción del grupo nueve, para distintos puntos del eje de la población correcta se obtiene una tabla de resultados como la que se muestra en la *Tabla XXIV*.

El punto de quiebre, donde se separan de manera significativa el modelo ideal del modelo de predicción y que se muestra en la *Tabla XXIV*, se da a la altura del 85% del Modelo ideal. En este punto la probabilidad de corrección decae demasiado y a partir de ahí no se recupera.

La matriz de clasificación, de la misma herramienta gráfica, nos arroja valores que son colocados dentro de la Matriz de confusión. Con el cálculo correspondiente se obtienen los resultados mostrados en la *Tabla XXV*.

Tabla XXV - Matriz de confusión - K-vecinos del conjunto de datos final

		Clasificador	
		Negativos	Positivos
Valores reales	Negativos	2,325	94
	Positivos	19	52

A partir de la Matriz de confusión detallada en la *Tabla XXV* se realiza el cálculo de los indicadores de desempeño del algoritmo con el conjunto de datos final. Los resultados obtenidos se muestran en la *Tabla XXVI*.

Tabla XXVI - Indicadores - K-vecinos del conjunto de datos final

Concepto	Variable	Valor %
Exactitud:	Ac	95.46%
Razón de verdaderos positivos:	TPrate	73.24%
Razón de falsos positivos:	FPrate	3.89%
Razón de verdaderos negativos:	TNrate	96.11%
Razón de falsos negativo:	FNrate	26.76%

En esta *Tabla XXVI* se identifica un desempeño aceptable, en lo general, pues en su conjunto muestran un resultado aceptable. Aún los indicadores más difíciles de tratar en este trabajo, como lo son los Verdaderos Positivos y los Falsos Negativos muestran un resultado aceptable bajo un análisis específico.

4.1.4 Confrontación de resultados

Se muestra la Tabla *XXVII* con la comparativa del desempeño de los indicadores después de someterlos a los algoritmos señalados en este trabajo.

Tabla XXVII - Confrontación de indicadores de ambos algoritmos

Concepto	Variable	Arboles de decisión	K-vecinos
Exactitud:	Ac	95.98%	95.46%
Razón de verdaderos positivos:	TPrate	64.10%	73.24%
Razón de falsos positivos:	FPrate	1.89%	3.89%
Razón de verdaderos negativos:	TNrate	98.11%	96.11%
Razón de falsos negativo:	FNrate	35.90%	26.76%

Al analizar uno a uno los indicadores de desempeño se tienen los siguientes comentarios:

- Exactitud, el rendimiento es adecuado en ambos algoritmos. Desde el punto de vista numérico es prácticamente el mismo resultado entre ambos algoritmos, con una mínima mejora para los Árboles de decisión.
- Razón Verdaderos Positivos, hay una mejora importante en el algoritmo de los K-vecinos con respecto al algoritmo de los Árboles de decisión. Es importante señalar

que el desempeño pasa de no aceptable a aceptable haciendo el análisis con el algoritmo de los K-vecinos.

- Razón de Falsos Positivos, el rendimiento es adecuado y se cuenta con una buena aproximación al óptimo para los dos algoritmos utilizados. Existe una mejora sensible en el desempeño del algoritmo de Árboles de decisión.
- Razón de Verdaderos Negativos, el rendimiento es adecuado, y se destaca que es el indicador con el mejor desempeño en ambos algoritmos utilizados.
- Razón de Falsos Negativos, hay una mejora importante en el algoritmo de los K-vecinos con respecto al algoritmo de los Árboles de decisión. Es importante señalar que el desempeño pasa de no aceptable a aceptable haciendo el análisis con el algoritmo de los K-vecinos.

Capítulo 5. Conclusiones y recomendaciones

5.1 Conclusiones

Después de exhaustivas selecciones de datos y pruebas de rendimiento de varios algoritmos de aprendizaje automático se llega a la conclusión de que si podemos hacer uso de estas técnicas para identificar patrones de comportamiento sospechoso en un empleado (o en un grupo de empleados) del sector financiero mexicano.

De manera particular para el problema de clasificación de empleados fraudulentos con el grupo de datos obtenido, se ajustan mejor los de Árboles de decisión y de los K-vecinos.

El algoritmo de los Árboles de decisión es fácil de implementar y las reglas que arroja son claras para su implementación. Por el otro lado, los resultados que arroja el algoritmo de los K-vecinos son fáciles de interpretar por la forma gráfica que puede tomar el modelo resultado.

Ahora bien, partiendo de la gráfica de confrontación de rendimiento de los algoritmos; se identifica un mejor rendimiento del algoritmo de K-vecinos (sobre el de Árboles de decisión) al tener mejores estadísticas en lo referente a:

- Razón de Verdaderos Positivos (mayor calidad al clasificar casos fraudulentos)
- Razón de Falsos Negativos (menor error al clasificar casos no fraudulentos)

Para complementar el análisis final de esta investigación se debe ir un poco más allá porque las diferencias en el rendimiento entre uno y otro algoritmo son pequeñas, como se acaba de describir.

Para complementar las conclusiones se muestran, en la Tabla XXVIII, una comparativa con los puntos a favor y en contra de ambos algoritmos confrontados en este trabajo.

Con lo señalado en los párrafos anteriores determinamos que el algoritmo de Árboles de decisión entrega un árbol bien definido y con una clasificación muy específica. Este algoritmo puede utilizarse con un cierto factor de éxito para tratar de predecir e identificar a un empleado fraudulento en la empresa objeto de estudio. Pero no es el algoritmo que tuvo el mejor desempeño en la investigación.

Tabla XXVIII - Resumen de elementos a favor y en contra de los algoritmos analizados

Algoritmo	Elementos a favor	Elementos en contra
Árboles de decisión	o Elimina el uso de la variable Monto. o Aunque clasifica, elimina también, la variable días transacción. o Marca la poca relevancia de la variable tipo de transacción. o Marca la poca relevancia de la variable Mismo origen. o La variable de mayor peso (días de antigüedad hasta la fecha de la transacción) abarca prácticamente todo el espectro de la línea de tiempo. Solo con el análisis detallado se pudo identificar.	o Se degrada rápidamente la calidad de la predicción, con respecto al modelo ideal, a partir del uso del 79% de las muestras disponibles. o Rendimiento insuficiente para clasificar Verdaderos Positivos y Falsos Negativos.
K-vecinos	o Logra aislar muy bien en un sólo grupo la variable dependiente facilitando el entendimiento y la focalización del análisis. o Con la herramienta de distinción del grupo se pueden aislar los casos fraudulentos, de los no fraudulentos, dentro del mismo grupo. o Está clara la regla que se debe implementar.	o Se degrada rápidamente la calidad de la predicción, con respecto al modelo ideal, a partir del uso del 85% de las muestras disponibles. o Rendimiento insuficiente (pero cercano al Aceptable) para clasificar Verdaderos Positivos y Falsos Negativos.

La conclusión final de esta investigación es que el algoritmo de los K-vecinos genera el grupo que se requiere para aislar las operaciones que se requieren utilizar para confirmar la hipótesis y que indica que es posible identificar comportamiento sospechoso de empleados operativos en el sector financiero. Adicionalmente, y apoyando la conclusión final, se debe destacar que en el grupo principal generado por el algoritmo de los K-vecinos aparecen operaciones que no forman parte de lo que se busca localizar (comportamiento no sospechoso) y aun así es posible aislar las condiciones para detectar las operaciones en un ambiente real.

5.2 Recomendaciones

A lo largo de la investigación se encontraron una serie de dificultades para ir avanzando con los experimentos y su análisis. La mayor parte de las áreas de oportunidad son derivadas del

crecimiento orgánico que ha experimentado la empresa objeto de análisis. Esto mismo les sucede a muchas empresas de diferentes giros.

Una ventaja que tiene esta empresa objeto de estudio es contar con desarrollos propios; y al mismo tiempo se convierte en un reto. Lo anterior se deriva de que el paso del tiempo trae nuevas corrientes tecnológicas y no siempre, en empresas que nacen pequeñas y se van fortaleciendo con el tiempo, se toman las medidas para sentar las bases de esos crecimientos futuros.

Esto origina una serie de recomendaciones que se pueden poner en marcha y que tienen como objetivo apoyar en el robustecimiento de ser más seguras en su operación detectando comportamientos sospechosos del personal interno.

Para que la empresa pueda contar con un mejor y más eficiente mecanismo de detección de fraudes, de manera preventiva, se recomienda llevar a cabo las siguientes acciones:

- Implementar una plataforma, sencilla de operar, que permita:
    - Registrar eventos fraudulentos de los empleados en donde queden relacionados estos últimos;
    - Asociar a los eventos fraudulentos las transacciones involucradas;
    - Manejo adecuado de las fechas (tanto del evento fraudulento, como del descubrimiento del mismo);
    - Señalar a los clientes afectados;
    - Mantener un control estricto de este tipo de eventos.
- Mantener homologado el Modelo de Datos operativo para relacionar los diferentes sistemas que operan (esto por el crecimiento orgánico natural de la Institución).
- Automatizar el proceso de extracción y procesamiento de los datos para actualizar sus valores y mejorar con esto la precisión de las predicciones.
- Dentro de la automatización del proceso hacer un análisis periódico para identificar la viabilidad de agregar otras variables derivado del crecimiento de canales de atención a clientes, así como productos financieros que suceden con el tiempo.
- Para contrarrestar la falta de precisión al evaluar Falsos Negativos (que son los que más pueden preocupar) se pueden implementar alertas que lleven a un análisis

focalizado por un área auditora. Es decir que el señalamiento de un comportamiento sospechoso lo haga el algoritmo implementado pero que haya un seguimiento puntual del caso para tomar una decisión. Esto último en el período de tiempo que el proceso actual va mejorando en su desempeño y alcanza los niveles óptimos.

5.3 Trabajos futuros

El tema de los fraudes internos en el sector financiero tomará cada vez más relevancia por la apertura de canales tecnológicos para operar y por la operación a distancia (cada vez más común). Un trabajo a futuro es madurar este modelo obtenido con algunas otras variables y con la finalidad de hacer la predicción en tiempo real. El objetivo final es hacer eficiente el procesamiento y que se permita detectar una amenaza en el mismo momento en que esta se está materializando. Un objetivo mayor es hacer que el modelo se convierta en un estándar, o al menos en la base, de una plataforma con mayores capacidades y alcances que pueda ser implementada en otras entidades y que les genere valor en su operación.

Puede considerarse el tomar el presente trabajo en estudiantes de nivel maestría o incluso doctorado para darle el avance, y madurez, anteriormente señalado.

Bibliografía

[1] aprendeIA, [En línea]. Available: https://aprendeia.com/historia-de-machine-learning/. [Último acceso: 2021].

[2] Definición de fraude, [En línea]. Available: https://definicion.de/fraude/. [Último acceso: 2021].

[3] Real Academia de la Lengua Española, [En línea]. Available: https://dle.rae.es/. [Último acceso: 2021].

[4] La ley de Benford: ¿aprender a defraudar o a detectar fraudes?, [En línea]. Available: http://blog.kleinproject.org/?p=1634&lang=es. [Último acceso: 2021].

[5] Y. Miao, Z. Rua, L. Pan, J. Zhang y Y. Xiang, *Comprehensive analysis of network traffic data,* nº e4181, pp. 30-46, 2018.

[6] Z. Jun, X. Yang, W. Yu, Z. Wanlei, X. Yong y G. Yong, *Network traffic classification using correlation information,* vol. 24, nº 1, pp. 104-118, 2013.

[7] J. A. Carmona Troyo y otros, Identificación de ataques en redes de cómputo utilizando redes neuronales artificiales, 2017.

[8] Inversor LATAM, [En línea]. Available: https://inversorlatam.com/segun-cifras-del-inegi-solo-el-47-de-los-mexicanos-tienen-una-cuenta-bancaria/. [Último acceso: 2021].

[9] B. A. Doroteo Valdéz y otros, Modelo de concientización en la prevención de la fuga de información, 2010.

[10] Híbridos y eléctricos, [En línea]. Available: https://www.hibridosyelectricos.com/articulo/actualidad/fabricante-camiones-electricos-nikola-acusado-supuesto-fraude/20201110172148039785.html. [Último acceso: 2021].

[11] Diario Las Américas, [En línea]. Available: https://www.diariolasamericas.com/florida/detectan-dispositivos-robos-tarjetas-gasolineras-del-estado-n3067514. [Último acceso: 2021].

[12] La Nación, [En línea]. Available: https://www.lanacion.com.ar/tecnologia/mercado-negro-de-tarjetas-de-credito-en-internet-nid723688/. [Último acceso: 2021].

[13] El país, [En línea]. Available: https://www.elpais.com.co/judicial/asi-fue-la-investigacion-sobre-fraude-en-bancolombia.html. [Último acceso: 2021].

[14] Aristegui Noticias, [En línea]. Available: https://aristeguinoticias.com/1605/mexico/robo-a-bancos-de-mexico-via-transferencias-fue-por-300-mdp-banxico/. [Último acceso: 2021].

[15] El Economista, [En línea]. Available: https://www.eleconomista.com.mx/sectorfinanciero/Las-140-personas-que-no-veran-la-justicia-por-fraude-de-Ficrea-20190803-0020.html. [Último acceso: 2021].

[16] O. Rodríguez, Conceptos báiscos de Minería de datos, 2019.

[17] J. Hernández Orallo y otros, Introducción a la minería de datos, 2005.

[18] D. E. Roldán Pinzón, Diseño de una guía general para construir una bodega de datos, 2015.

[19] L. A. Cárdenas Florido, Minería de datos, 2018.

[20] J. Orión, Herramientas para la supervisión digital (cubos de información), 2018.
[21] E. Ballesteros Doncel, Estadística descriptiva univariante mediante el gráfico de caja y bigotes, 2015.

Glosario

A

Ahorro a la vista. Es dinero de cliente bancario, resguardado por un banco, que está disponible para su uso todo el tiempo; por ejemplo, el dinero guardado en una cuenta con tarjeta de débito que puede ser usada en cajeros automáticos o comercios.

Ahorro a plazo fijo. Es dinero de cliente bancario, resguardado por un banco, que no está disponible de manera inmediata. El banco, a cambio del congelamiento de este dinero, paga un rendimiento superior al dinero guardado a la vista.

Algoritmo. Es una serie ordenada de instrucciones, pasos o procesos que llevan a la solución de un determinado problema.

B

Banco de México®. Organismo autónomo mexicano que regula los medios de pago.

C

CNBV. Comisión Nacional Bancaria y de Valores; es la institución gubernamental que regula la operación del sector financiero.

Colaborador. Persona que está contratada (es empleado) de una empresa.

CONDUSEF. Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros.

Cultura empresarial. Es el conjunto de formas de actuar, de sentir y de pensar que se comparte entre los miembros de una organización y son los que identifican a la empresa ante sus clientes, proveedores y todos los que conocen de su existencia.

E

Entrenamiento. Preparación para perfeccionar el desarrollo de una actividad.

I

Ilícito. Que no está permitido por la ley o no es conforme a lo moral.

O

Opinión pública. Manera de pensar que es común a la mayoría de las personas acerca de un asunto.

P

Patrón de comportamiento. Forma constante de pensar, sentir, reaccionar físicamente y actuar en determinada situación.

S

Sector financiero mexicano. Grupo de empresas del ramo financiero mexicano que están certificadas y normadas por las autoridades competentes.

SGBDR. Sistema Gestor de Base de Datos Relacional.

Sistema informático. Componentes de software y de hardware que interactúan armónicamente para atender las transacciones financieras de los clientes.

SOFIPO. Sociedad Financiera Popular, es un tipo de empresa financiera mexicana cuyo objetivo es atender a clientes específicos (mayormente en el sector primario productivo) en el país.

SPEI. El Sistema de Pagos Electrónicos Interbancario (SPEI) es la infraestructura de pagos del Banco de México que permite a sus participantes (bancos, casas de bolsa, sofipos y otras entidades financieras reguladas) enviar y recibir pagos entre sí para poder brindar a sus clientes finales el servicio de transferencias electrónicas en tiempo real.

T

Taxonomía. Clasificación u ordenación en grupos de cosas que tienen unas características comunes.

Transacción. Operación financiera.