Universidad Nacional Autónoma de México

programa de posgrado en ciencias de la computación

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

Inteligencia Artificial

Detección de transferencias bancarias fraudulentas usando técnicas de aprendizaje de máquinas

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Ciencia e Ingeniería de la Computación

PRESENTA:

Luis García Rodríguez

TUTOR:
Gibran Fuentes Pineda
IIMAS

Ciudad Universitaria, CDMX, Marzo de 2022

En memoria de ***Ruth****,*

Que con su incondicional amor me acompañó en este maravilloso camino, siendo motivo, inspiración, brújula, cómplice y soporte.

"¡Qué maravilla es poder sentirte aunque no estás!
Siempre supe que en el viento te podías quedar."
Alejandro Filio

A ***Sebastián****,* ***Íker*** *y* ***Nadia****,*

Como un incentivo de superación.

Cada éxito comienza siendo un sueño, que aunque parezca lejano cuando establecemos metas, planificamos y trabajamos duro es alcanzable.

Disfruten su trayecto siempre con sus objetivos en mente.

Agradecimientos

A la UNAM, mi alma máter que me ha permitido alcanzar un nuevo peldaño académico y de esta forma fortalecer mi trayectoria profesional. En particular al Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas y al programa de Posgrado en Ciencia e Ingeniería de la Computación ya que me brindó la oportunidad de adentrarme en el mundo de la inteligencia artificial.

Al Dr. Gibran Fuentes Pineda que guió el presente trabajo de investigación aportando en todo momento sus conocimientos y experiencia. Sus consejos y asesoría han sido de gran valor.

A las Dras. Helena Montserrat Gómez Adorno y Rocío Aldeco Pérez así como al Dr. José Antonio Neme Castillo ya que con sus valiosos comentarios y aportaciones enriquecieron y permitieron mejorar la calidad de la versión final del presente trabajo de investigación.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Luis García Rodríguez. Ciudad Universitaria, CDMX, Marzo de 2022

Resumen

El creciente aumento en el uso de medios electrónicos y la accesibilidad al internet y dispositivos de cómputo han favorecido la adopción de los sistemas bancarios en modalidad software como servicio (como lo son las aplicaciones de banca por internet) para realizar transferencias electrónicas de dinero. Esto a su vez ha traído consigo un aumento en la comisión de transferencias fraudulentas, problema que afecta económicamente tanto a los usuarios de la banca por internet como a los bancos o instituciones financieras, estas últimas también se ven afectadas en su reputación.

El análisis de datos y las técnicas de aprendizaje de máquinas sirven como apoyo en la contención del problema toda vez que pueden ser utilizadas para aprender los patrones del comportamiento transaccional de los usuarios al ejecutar transferencias legítimas y de los defraudadores electrónicos al realizar transferencias fraudulentas.

En el presente trabajo, a partir de la información transaccional de una aplicación de banca por internet se propone un proceso metodológico para realizar la integración de la información originada en diversas fuentes de datos hacia una sola base de datos relacional.

A partir de la base de datos integrada se genera un archivo con el conjunto de características continuas y categóricas seleccionadas y con las cuales se ejecuta un análisis exploratorio de los datos en donde resaltan hallazgos que guían hacia la generación de características extendidas de tiempo y estadísticas.

Las características extendidas de tiempo se basan en las marcas de tiempo de eventos como el inicio de sesión de la banca por internet, el registro de cuentas destino, la captura de datos de la transferencia, la ejecución de la transferencia y el cierre de sesión de la banca por internet.

Las características extendidas de tipo estadístico contienen conteo de transacciones, montos mínimos, máximos, promedio y desviación estándar calculadas bajo tres ventanas de operación: transaccionalidad global, transaccionalidad anual y transaccionalidad mensual.

Finalmente, se realiza una selección de modelos de aprendizaje de máquinas con los cuales se ejecutan las fases de entrenamiento, validación y ajuste de hiperparámetros y prueba. Se analizan las métricas adecuadas para el problema de naturaleza desequilibrada y se diseña una serie de experimentos orientados en evaluar el desempeño de los modelos propuestos.

Con base en la experimentación propuesta y los resultados obtenidos se detectan al menos dos modelos (LightGBM y XGBoost) que alcanzan niveles satisfactorios en las métricas propuestas, logrando de esta manera alcanzar los objetivos propuestos.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La puesta en marcha de los servicios financieros a través de medios electrónicos como internet, telefonía y dispositivos móviles por parte de los bancos y en general de las instituciones financieras, constituye un requerimiento operativo primordial. Dichos medios representan canales cada vez más indispensables para brindar servicios así como para captar clientes y optimizar el flujo de fondos hacia las instituciones.

La digitalización de los servicios financieros se ha visto acelerada debido a la crisis sanitaria generada por la pandemia por COVID-19 extendida desde finales de 2019 y hasta la actualidad. La necesidad del aislamiento físico de la población durante este periodo ha forzado al uso de canales electrónicos a un número de usuarios de servicios financieros que anteriormente únicamente utilizaba canales (sucursales, ATMs) y medios de pago físicos (efectivo, cheque, vales en papel, etc.).

En Agur et al. [2], reportan incrementos promedio del 8 % y 15 % en las transacciones electrónicas durante 2018 y 2019 (años previos a la pandemia) en Economías Emergentes y en Desarrollo y establecen como oportunidad de expansión la consolidación de los servicios financieros digitales. En México, específicamente BBVA reporta un incremento de 55 % en las transacciones digitales durante 2020 [3].

En contraparte, estos canales conllevan costos operativos, administrativos y técnicos de alta especialización para las instituciones, además de que representan un riesgo de seguridad toda vez que a nivel mundial los ataques cibernéticos han ido en franco incremento con el pasar de los años.

La Asociación de Examinadores de Fraude Certificados (Association of Certified Fraud Examiners - ACFE) es una organización internacional que se encarga de generar guías, material instructivo y proveer entrenamiento para la prevención del fraude así como de cuantificar los daños generados por todo tipo de fraudes, entre ellos, los fraudes cibernéticos sobre servicios financieros.

La ACFE define al fraude como: “cualquier acto intencional o deliberado de privar a otro de bienes o dinero por astucia, engaño u otros actos injustos” [4]; y en su Reporte a las Naciones 2020 [5], indica que el fraude sobre servicios bancarios y financieros representa un costo medio de $100,000.00 USD por evento, mientras que en la región de América Latina y el Caribe el costo medio se incrementa al doble, siendo México y

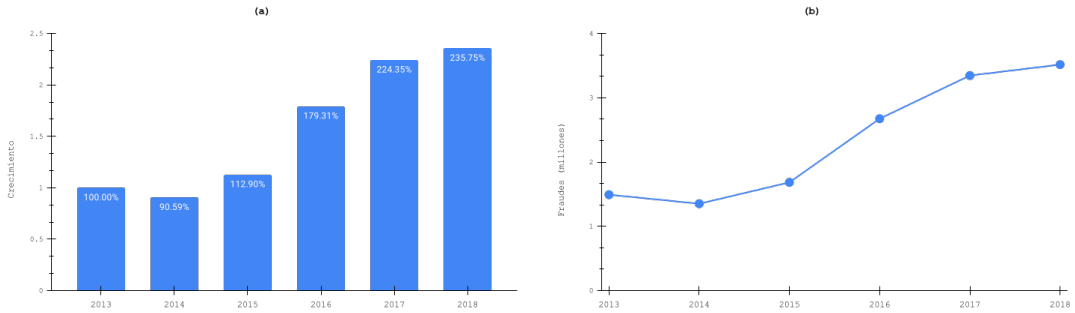

Figura 1.1: (a) Crecimiento anual de fraudes en el sistema financiero mexicano entre 2013 y 2018. (b) Cantidad de fraudes en el sistema financiero mexicano entre 2013 y 2018. Fuente CONDUSEF [1].

Brasil los países con mayor incidencia.

El reporte Nilson de diciembre de 2020 informa que las pérdidas por fraudes con tarjetas (incluyendo crédito, débito y prepago) alcanzan los $28,650,000,000.00 USD a nivel global [6].

En México, con base en cifras de la Comisión Nacional para la Protección y Defensa de los Usuarios de Servicios Financieros (CONDUSEF) la incidencia de fraudes cibernéticos sobre servicios financieros en canales electrónicos ha incrementado alrededor de 21 % en promedio por año [1]. La figura (1.1) ilustra el aumento de fraudes sobre transacciones electrónicas en México para el periodo de 2013 a 2018.

Se trata de una problemática que afecta monetariamente tanto a los usuarios de los servicios financieros como a las instituciones que otorgan dichos servicios.

1.1. Planteamiento del problema

La detección de fraudes sobre transacciones electrónicas se puede segmentar principalmente en las siguientes categorías:

- Transacciones de Tarjeta de Crédito.
- Transacciones de Comercio Electrónico.
- Transacciones a través de Banca por Internet.
- Transacciones a través de Banca Móvil.

Se trata de una problemática ampliamente abordada desde la comunidad científica y académica desde hace varios años, principalmente en su categoría de transacciones de tarjeta de crédito siendo también esta categoría la de mayor afectación. La siguiente

categoría mayormente abordada sobre todo en años recientes es la correspondiente a las transacciones de comercio electrónico.

En el presente proyecto de investigación se abordará la categoría de transacciones a través de banca por internet. Este tipo de transacciones tiene la particularidad de que la transferencia de fondos entre las cuentas se realiza de manera inmediata. Lo anterior hace que una vez ejecutada la transferencia sea muy difícil recuperar el dinero.

Para las diversas categorías de la detección de fraudes sobre transacciones electrónicas se han propuesto variadas soluciones. A continuación un listado con algunos de los modelos de aprendizaje de máquinas propuestos para solucionar esta problemática.

- Modelos de Aprendizaje Supervisados.
  - Árboles de Decisión [7] [8] [9] [10].
  - Bosques Aleatorios [7] [8] [11] [12] [13].
  - Máquinas de Soporte Vectorial [7] [8] [13].
  - Regresión Logística [7] [8].
  - K Vecinos más cercanos [14]
- Modelos de Aprendizaje No Supervisados.
  - Bosque de Aislamiento [15] [16] [17]
  - Factor de Valor Atípico [15] [17]
  - DBSCAN [18]
- Redes Neuronales Artificiales [19] [20].
  - Recurrentes [21] [22]
  - Convolucionales [23]

Si bien, cada uno de estos modelos ha aportado mejoras en la mitigación de los daños causados por los defraudadores, la realidad es que cada año se sigue presentando un crecimiento en la incidencia de fraudes, como se verá más adelante debido a la evolución en las técnicas y comportamiento del fraude.

Actualmente se cuenta con capacidades de cómputo suficientemente adecuadas para el procesamiento de información con grandes volúmenes que aunada a las ventajas que otorgan las redes neuronales artificiales han permitido mejorar sustancialmente el rendimiento de las tareas de aprendizaje de máquinas como en el caso de la clasificación y detección de eventos de fraude.

En contrapunto se tiene el hecho de que las técnicas desarrolladas por los defraudadores también han ido evolucionando con el paso del tiempo al grado de ser capaces de simular de manera muy eficiente el comportamiento de las operaciones legítimas.

Lo anterior obliga a mirar el problema desde diferentes puntos de vista en búsqueda de nuevas y mejores soluciones para contrarrestar las nuevas maneras de realizar fraudes.

1.2. Objetivos

A continuación se describe el objetivo general del presente trabajo de investigación así como los objetivos específicos.

1.2.1. General

Implementar el proceso de integración y limpieza de datos, ingeniería y selección de características, selección y evaluación de modelos de aprendizaje de máquinas en un sistema de banca por internet del sector financiero mexicano. Lo anterior bajo la hipótesis de que con el tratamiento, características y modelo adecuado es posible alcanzar niveles de detección de transferencias fraudulentas similares al estado del arte.

1.2.2. Específicos

Los objetivos específicos se listan a continuación:

- Preparar y consolidar la información disponible en fuentes previamente identificadas, en una sola base de datos.
- Investigar y analizar el estado del arte como referencia en cuanto a metodología, técnicas, modelos y resultados.
- Realizar un análisis exploratorio de los datos en busca de diferenciar distribuciones, detectar patrones de comportamiento y posibles valores atípicos.
- Analizar, seleccionar y ajustar los atributos de dominio para integrar los modelos de aprendizaje de máquinas.
- Seleccionar y evaluar modelos de aprendizaje de máquinas con base en el estado del arte.
- Experimentar con modelos supervisados.
- Experimentar con modelos no supervisados.

1.3. Retos en la detección de fraudes

La detección automática de fraudes es uno de los problemas recurrentes a los que se enfrentan empresas y organizaciones del sector financiero y comercio electrónico, entre otras, pues implica múltiples afectaciones tanto para las propias empresas como para sus clientes.

Las repercusiones no son únicamente económicas, también afectan otros factores intangibles como la confianza, credibilidad y reputación de las empresas incidiendo en

detrimento del valor de las marcas. Estos factores son difícilmente cuantificables de manera inmediata, sin embargo, en el mediano plazo se ven reflejados en situaciones como la pérdida de clientes y la dificultad para captar nuevos, la cancelación de cuentas y retiro de fondos, la disminución de la transaccionalidad lo que conlleva a la baja en el cobro de comisiones, etc.

Desde el punto de vista técnico también existen diversos retos a los cuales se debe hacer frente. Los defraudadores cibernéticos generalmente están varios pasos adelante de las instituciones financieras, evolucionando sus técnicas de delinquir [24]. Al tratarse de eventos aislados dentro de cientos de miles suele ser muy difícil detectarlos a tiempo para contrarrestar los daños.

1.3.1. Disponibilidad de conjuntos de datos

Uno de los grandes retos que se deben afrontar desde la investigación científica al abordar la problemática de la detección de fraudes sobre transacciones electrónicas con técnicas de aprendizaje de máquinas es la escasez de conjuntos de datos públicos reales. Específicamente para las variantes de transacciones a través de banca por internet y banca móvil a la fecha de elaboración del presente trabajo de investigación no se han encontrado conjuntos de datos públicos. En el caso de las variantes de transacciones con tarjeta de crédito y transacciones de comercio electrónico aunque son escasos, sí existen.

Para el caso de transacciones con tarjeta de crédito existen los conjuntos de datos alemán [25] y europeo [26]. Para el caso de las transacciones de comercio electrónico existe el conjunto de datos IEEE-CIS Fraud Detection (Kaggle) [27].

1.3.2. Integración de características

Los sistemas informáticos de las empresas financieras (bancos, casas de bolsa, hipotecarias, financiadoras, etc.) suelen ser ecosistemas complejos en los que cada módulo se especializa en responsabilidades específicas, por citar un ejemplo, los sistemas de autenticación suelen ser transversales (brindan servicio) a diversos aplicativos funcionales como podrían ser las bancas por internet y móvil. En dichos sistemas se administran las credenciales de los usuarios. Por otro lado los contratos y las cuentas suelen ser gestionados en otro módulo denominado sistema nuclear bancario. Esta modularización se lleva a cabo así, por razones de arquitectura, seguridad, continuidad de negocio, etc. En la figura (1.2) se muestra un diagrama de componentes a alto nivel que representa un entorno de banca por internet en el cual se puede apreciar la modularización de un sistema de banca por internet.

Lo anterior, genera que al momento de definir un proyecto de detección de fraude los atributos o características requeridas para integrar los modelos de aprendizaje de máquinas se encuentren dispersas entre diversos medios, bases de datos relacionales y no relacionales, archivos de bitácoras de los diversos sistemas, reportes operativos entre áreas y hacia las autoridades, informes, etc.

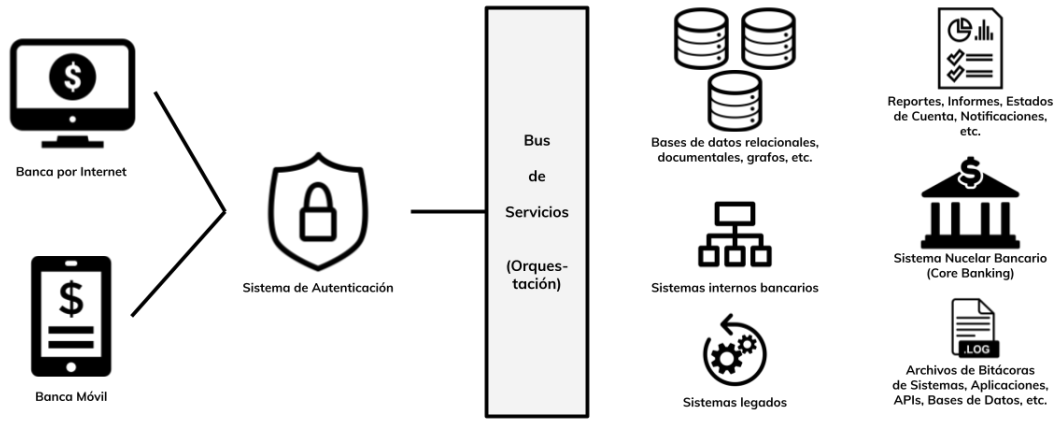

Figura 1.2: Variedad de sistemas y componentes tecnológicos que intervienen en el funcionamiento y generación de información de un sistema de banca por internet.

Por lo cual, la integración de la información suele ser un proyecto por sí mismo que además requiere de perfiles expertos en la gestión de datos y en el dominio de negocio [28].

1.3.3. Naturaleza desequilibrada del problema

Como se mencionó anteriormente la detección de fraudes es una problemática en la cual los eventos de interés se cuentan a razón de uno entre cientos de miles (1:100,000) [7] llegando incluso en algunos casos a contarse en proporción de uno entre millones (1:1,000,000). Muchas de las técnicas de aprendizaje de máquinas llegan a presentar una disminución importante en su rendimiento cuando se trata de un problema de naturaleza desequilibrada, esto debido a que los algoritmos no logran diferenciar los patrones de comportamiento de los eventos de fraude debido a que son muy escasos.

1.3.4. Evolución de las técnicas y comportamiento de fraudes

Los defraudadores cibernéticos cuentan con una amplia gama de posibilidades para la ejecución del fraude, desde técnicas de ingeniería social hasta sofisticadas técnicas de intrusión en los sistemas informáticos son utilizadas para alcanzar el objetivo, simular o ejecutar transacciones apócrifas.

Por otro lado, las técnicas avanzadas para detectar los fraudes al ser producto en muchas ocasiones de la ciencia e investigación suelen ser públicas lo que favorece su conocimiento por parte de los defraudadores. También es claro que los defraudadores modifican continuamente las técnicas y procedimientos de ataque [29] [30] [24] [31].

1.4. Metodología

Se cuenta con la información transaccional de un sistema de banca por internet para el periodo comprendido entre enero de 2015 hasta diciembre de 2019.

La base de datos contiene información anonimizada de las transferencias realizadas durante el periodo especificado: clientes, usuarios, montos, cuentas origen y destino, etiqueta de fraude, marca de tiempo del ingreso a la sesión de banca por internet, marca de tiempo de captura de la transferencia y marca de tiempo de registro de la cuenta destino.

Las fuentes de la información son las siguientes:

- Base de datos relacional con la información estructurada de: clientes, usuarios, cuentas bancarias y registro de transacciones.
- Archivos de bitácora operacional (texto plano) cuya información de relevancia son las marcas de tiempo de los eventos de registro de usuario (*login*), confirmación y autorización de las transacciones.
- Informes de seguimiento de reportes de fraude con dictaminación.

Debido a lo anterior, el procedimiento metodológico por ejecutar es el siguiente:

1. Generación de nuevas relaciones en la base de datos para incluir la información de la marca de tiempo de los eventos de inicio de sesión (*login*) así como la confirmación y autorización de las transacciones.
2. Extracción de la información de relevancia de los archivos de bitácora para su integración en las nuevas relaciones de la base de datos.
3. Análisis y selección de los atributos susceptibles de integrar el modelo de datos.
4. Definición de un conjunto de modelos de aprendizaje de máquinas (supervisados, no supervisados y redes neuronales artificiales) con base en el análisis e investigación del estado del arte.
5. Ejecución, evaluación y ajuste de los atributos y modelos seleccionados.

1.5. Contribuciones de la investigación

A continuación se describen las contribuciones del presente trabajo de investigación.

1.5.1. Análisis de temporalidad

El análisis exploratorio de los datos así como los algoritmos de clasificación seleccionados, buscan detectar cambios en los patrones de comportamiento de los fraudes a través del tiempo. En específico, se cuenta con información transaccional de un sistema de banca por internet para los años 2015 al 2019. En dicho rango se perciben cambios de comportamiento que los algoritmos deben ser capaces de detectar. El objetivo siempre será mejorar el rendimiento de los algoritmos.

1.5.2. Análisis de estacionalidad

Además del comportamiento secuencial a través del tiempo, se realiza un análisis enfocado en detectar posibles comportamientos estacionales de los eventos de fraude. Esto es, se buscan épocas del año en las cuales se incremente el número de eventos de fraudes y se analizan los diversos años para determinar si se trata de un patrón de comportamiento.

1.5.3. Análisis del comportamiento

La selección inicial de características que incluye atributos de negocio como los números de las cuentas origen y destino, el monto de la transacción y la fecha de la transacción. Al analizar dicho conjunto de características se determinó la integración de características extendidas con el objetivo de alcanzar una mejor separabilidad de las clases.

Entre las características extendidas propuestas se incluyen los diferenciales de tiempo entre los momentos de captura de la transferencia y su autorización, el ingreso a la sesión de banca por internet (*login*) y la autorización de la transacción así como el registro de la cuenta destino y la autorización de la transacción.

Estas características modelan el comportamiento de los usuarios dentro de la sesión de banca por internet en la cual se realizan transferencias. Se analiza el impacto de estos atributos de comportamiento en el rendimiento de los diversos modelos de clasificación para la detección de fraudes.

1.6. Estructura de la tesis

El presente trabajo se encuentra organizado en 6 capítulos, a continuación se describe brevemente el contenido de cada uno de ellos.

- **Capítulo 1**. Se describe el entorno de las transacciones electrónicas, se específica el concepto de fraude, se presentan datos numéricos sobre el crecimiento continuo de los fraudes electrónicos, se plantean los retos en la detección de fraudes.

- **Capítulo 2**. Se realiza una revisión acerca del estado del arte de los modelos de aprendizaje de máquinas para la clasificación binaria de fraudes sobre transacciones electrónicas.
- **Capítulo 3**. Se describen las técnicas de aprendizaje de máquinas supervisadas y no supervisadas así como de las redes neuronales artificiales para la clasificación de fraudes bancarios.
- **Capítulo 4**. Se describe detalladamente la metodología para la integración y el tratamiento de la fuentes de información, se presenta el análisis exploratorio de datos, se describe el proceso de selección de atributos, se presentan las métricas propuestas y se describen los experimentos propuestos.
- **Capítulo 5**. Se presentan los resultados obtenidos para la serie de experimentos de aprendizaje de máquinas propuestos.
- **Capítulo 6**. Se discuten los resultados y se establecen las conclusiones respecto al proceso de detección de fraudes bancarios.

Capítulo 2

Revisión del estado de arte

Como se mencionó anteriormente, diversas técnicas de aprendizaje de máquinas se han utilizado para afrontar el problema de detección de fraudes bancarios a través de medios electrónicos, principalmente en su variante de transacciones con tarjeta de crédito.

Si bien, el presente trabajo de investigación se enfoca en la variante de transacciones a través de banca por internet, existe un número reducido de trabajos de investigación científica enfocados específicamente en esta variante. La figura (2.1) muestra de forma comparativa el crecimiento en la generación de artículos científicos para (a) aprendizaje de máquinas y aprendizaje profundo y (b) banca por internet y tarjeta de crédito.

Las similitudes en cuanto al proceso de ejecución de las transacciones así como el flujo de información entre ambas variantes, permiten adaptar los modelos propuestos para la detección de fraudes con tarjeta de crédito hacia la detección de transferencias fraudulentas a través de banca por internet.

Respecto a los trabajos clásicos de aprendizaje de máquinas cabe señalar que si bien, existen investigaciones de 15 años o más de antigüedad [32] [33] [34] [35] [36] [37] [38], los más recientes presentan mejoras en cuanto al rendimiento de las métricas propuestas [9] [11] [39] y generalmente se presentan comparativas en las cuales se demuestran ventajas y desventajas de diversos modelos [7] [8] [15] [17].

Por otro lado, desde un enfoque costo - beneficio en comparativa con modelos de aprendizaje profundo encuentran motivos para continuar vigentes como menores tiempos de implementación en las fases de entrenamiento, validación, prueba y puesta en marcha en ambientes productivos [40]; menor poder de cómputo consumido [41] [42] y rendimiento razonablemente conveniente [39].

Respecto a los trabajos basados en aprendizaje profundo cabe señalar que para aquellas instituciones en las cuales la cantidad de transacciones y flujo de información se clasifican como datos masivos (*big data*), estos modelos son más adecuados [43]. Debido a las capacidades de procesamiento que presentan incluyendo la ventaja intrínseca de que dichos modelos no padecen un decaimiento en el rendimiento de las métricas sino todo lo contrario, a mayor cantidad de información mejor es el rendimiento reportado [44].

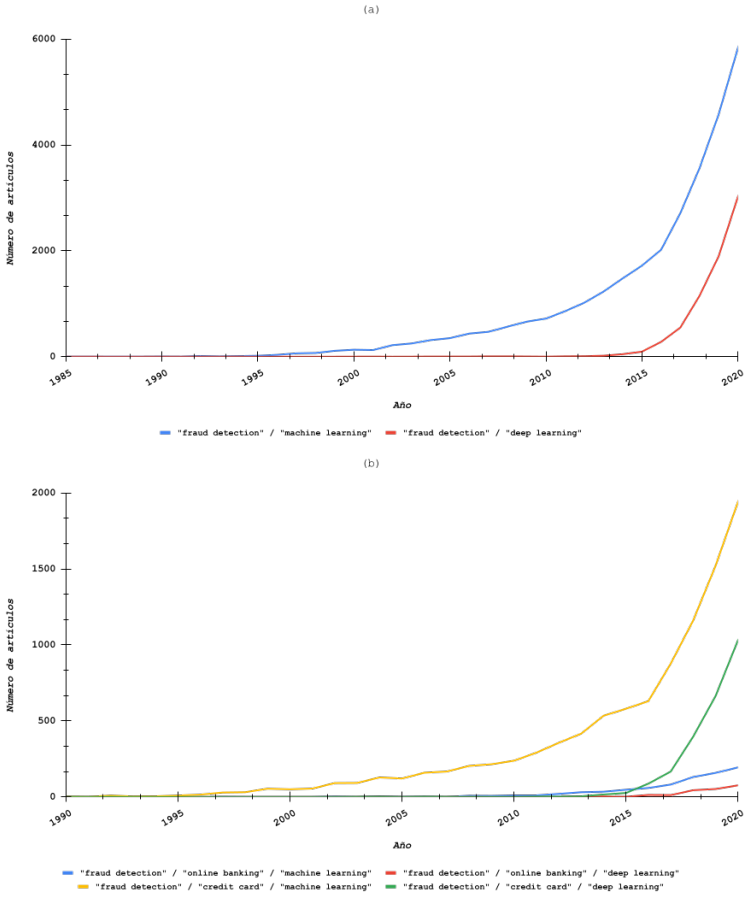

Figura 2.1: Evolución en la generación de artículos científicos sobre detección de fraudes, comparativa entre (a) aprendizaje de máquinas y aprendizaje profundo y (b) banca por internet y tarjeta de crédito. Información basada en búsquedas en Google Scholar.

Otra seria desventaja de los modelos de aprendizaje profundo es la capacidad expresiva o la capacidad de los modelos para ser interpretados y explicados [45]. Un tema relevante cuando se requiere comprender las causas raíz de la problemática, conocer la influencias de las diversas características en el modelo así como explicar las sutiles diferencias en cuanto al comportamiento de las técnicas de los defraudadores digitales [29] [24].

A continuación haremos un recorrido por diversos trabajos relacionados que han logrado resultados sobresalientes en cuanto al rendimiento alcanzado en las métricas utilizadas.

Husejinovic [9] realiza un ejercicio comparativo de técnicas de aprendizaje supervisado para abordar la detección de fraudes en tarjetas de crédito. Si bien no se especifica la información incluida en el estudio son relevantes para esta investigación los resultados obtenidos para el modelo implementado de árboles de decisión (*Decision Tree*).

Obtiene valores de 1.000 para la clase 1 (transacciones legítimas) y 0.927 para la clase 0 (transacciones fraudulentas) en la métrica precisión, demostrando una alta capacidad de inferencia para la clase positiva, es decir, las transacciones fraudulentas.

Para la métrica Sensibilidad (*recall - sensitivity*) que es complementaria de la precisión se obtienen valores de 1.000 para la clase 1 y 0.778 para la clase 0, este último valor indica que 0.222 de las instancias positivas reales fueron inferidas como negativas.

Finalmente obtiene los valores 0.999 para la clase 1 y 0.745 para la clase 0 al evaluar la Curva Precisión - Sensibilidad (*PRC - Precision Recall Curve*).

En dicho estudio el rendimiento más alto para la clasificación de transacciones fraudulentas se alcanzó con la implementación del algoritmo C4.5 de árboles de decisión.

Devi et al. [11] exploran en detalle las características del algoritmo Bosque Aleatorio (*Random Forest*) y su estrecha relación con los Árboles de Decisión C4.5, en particular, los beneficios reportados por el esquema de entrenamiento *"bagging"*. En dicho entrenamiento se fragmenta el conjunto de datos en un número determinado de bolsas y se realizan entrenamientos simultáneos para posteriormente establecer los pesos de cada fragmento de datos y finalmente obtener un resultado global con base en los datos de prueba.

Los autores configuran los hiperparámetros para establecer un enfoque de pesos sensible al costo de tal manera que se pueda tener una efectiva detección de fraudes contemplando el desequilibrio entre las clases.

Se proponen como métricas de evaluación del rendimiento *F-measure*, *Gmean* y AUC, alcanzando valores de 0.76815, 0.82298 y 0.778 respectivamente en promedio para ambas clases. En este caso la selección de métricas está orientada a ponderar la naturaleza desequilibrada del problema. *F-measure* obtiene un promedio ponderado de las métricas precisión y la sensibilidad, *G-mean* es un promedio geométrico y AUC establece la capacidad de clasificación del algoritmo.

La considerable baja en los valores de las métricas es debido a que las métricas seleccionadas operan sobre ambas clase y buscan dar un equilibrio a las métricas en las que se basan.

Para su experimentación utilizaron un par de conjuntos de datos públicos (alemán

[25] y australiano [46]) de transacciones con tarjeta de crédito alojados en ambos casos en el repositorio de conjuntos de datos de la Universidad de California Irvine [47].

Uno de los algoritmos relativamente recientes (2016) que ha sido probado en una diversidad de problemas con resultados exitosos es el Potenciamiento del Gradiente (*Gradient Boosting*). En particular la implementación XGBoost [48] [49] ha sido utilizada en numerosos concursos de aplicación de modelos de aprendizaje de máquinas para la resolución de problemas de datos tales como los propuestos por el portal Kaggle [50]. En dichos concursos XGBoost ha demostrado su contribución a los algoritmos basados en árboles de decisión y ha sido llamado un modelo extremo a extremo con capacidades para entregar resultados de frontera.

Al respecto Shimin et al. [39] proponen la implementación de XGBoost para la clasificación binaria de eventos de fraude sobre un sistema de transacciones de comercio electrónico. Cuentan con la base de datos IEEE-CIS [27] y realizan una comparativa entre los algoritomos Bayes Ingenuo (*Naïve Bayes*), Regressión Logistica, GBDT y XGBoost. Las métricas propuestas fueron ROC_AUC y Exactitud (*accuracy*) obteniendo como valores más altos 0.942 y 0.976 (en promedio para ambas clases) respectivamente XGBoost.

Mientras que Zhang et al. [51], utilizando la misma base de datos [27] realizan una comparativa entre los algoritmos Regresión Logística, Máquinas de Vectores de Soporte (SVM), Bosque Aleatorio (*Random Forest*) y Aumento de Gradiente (XGBoost). Las métricas propuestas para la evaluación de los modelos nuevamente son ROC_AUC y Exactitud (*accuracy*) obteniendo los mejores resultados XGBoost, 0.952 y 0.981 (en promedio para ambas clases) respectivamente.

Entre los algoritmos más utilizados para la clasificación binaria de transacciones fraudulentas por medios electrónicos se encuentra Máquinas de Vectores de Soporte (SVM), en años recientes se ha incluido constantemente en estudios comparativos junto a técnicas como Árboles de Decisión y Bosque Aleatorio. En [51] se compara su rendimiento respecto a Regresión Logística, Potenciamiento del Gradiente y Bosque Aleatorio, si bien, no obtiene los mejores resultados, se trata de resultados considerablemente aceptables (0.907 ROC_AUC, 0.958 Exactitud) sobre todo pensando en la relación costo - beneficio.

Mientras que Rtayli et al [52] proponen un enfoque integrado en el cual se utiliza SVM junto con la técnica de Reducción Recursiva de Características (*Recursive Feature Elimination - RFE*) [53]. Para disminuir el problema del desequilibrio entre las clases se echa mano de la Técnica de Sobremuestreo Sintético de Minorías (*Synthetic Minority Oversampling Technique - SMOTE*) [54] con el objetivo de balancear la clase de transacciones fraudulentas respecto a la clase de transacciones legítimas. La optimización de hiperparámetros es resuelta utilizando la técnica de validación cruzada GridSearchCV [55].

Las métricas utilizadas para medir el rendimiento del modelo son Sensibilidad (*recall*), Precisión y F_1, para el caso del conjunto de datos [26] se obtienen los siguientes valores 1.000 (sensibilidad), 0.970 (precisión) y 0.999 (F_1).

Como se observa la selección de modelos varía entre regresión logística, árboles

de decisión, bosque aleatorio, máquinas de vectores de soporte y potenciamiento del gradiente. Respecto a las métricas de rendimiento seleccionadas destacan sensibilidad, precisión, ROC_AUC y F_1. Destaca el uso combinado de métricas para evaluar el rendimiento de los modelos propuestos.

Capítulo 3

Aprendizaje de máquinas

La analítica de datos asistida por métodos estadísticos es una práctica longeva, con el surgimiento del cómputo y las base de datos relacionales cobró mayor relevancia. Ha sido en los últimos quince años con el surgimiento del cómputo en la nube, la generación de masiva de información y las grandes capacidades de cómputo que los métodos de aprendizaje de máquinas han tenido un resurgimiento. Las aplicaciones de estos métodos en diversos ámbitos tanto en la academia como en la industria se encuentran en auge.

En el caso particular de la detección de fraudes en transacciones bancarias sobre medios electrónicos se han realizado estudios encontrando ventajas e inconvenientes con diversas técnicas de aprendizaje de máquinas.

3.1. Aprendizaje de máquinas supervisado

Las técnicas de aprendizaje de máquinas supervisado aprovechan el conocimiento que se tiene sobre los ejemplos de transacciones legítimas y fraudulentas para construir los modelos de aprendizaje. Una de las problemáticas más marcadas en la detección de fraudes sobre transacciones financieras es la estrechez o similitud que existe entre los ejemplos de las transacciones legítimas y las fraudulentas. Esto debido a que los defraudadores digitales se encuentran constantemente mejorando sus técnicas, adaptándose a los mecanismos de seguridad implementados por las instituciones financieras y a los patrones de ejecución de transacciones legítimas de los usuarios [29].

Por lo anterior, contar con la información etiquetada supone una ventaja al momento de abordar estos problemas. En contraparte, se conoce que algunos algoritmos de aprendizaje supervisado requieren de un nivel alto de equilibrio entre las clases y algunos otros simplemente no operan adecuadamente en casos de información desequilibrada [56] [57].

A continuación se revisarán algunos de los algoritmos de aprendizaje de máquinas supervisados utilizados para abordar la tarea de la clasificación de transacciones fraudulentas en medios electrónicos.

3.1.1. Árboles de decisión (*Decision tree*)

Los árboles de decisión son algoritmos sustentados en la teoría de grafos y más específicamente en los árboles binarios, son ampliamente utilizados en el análisis de datos de una amplia variedad de contextos debido a su capacidad interpretativa y factibilidad en cuanto a requerimientos computacionales.

Los árboles de decisión como algoritmo de aprendizaje supervisado tienen utilidad para las tareas de clasificación y regresión con sutiles diferencias en sus implementaciones.

Entre las implementaciones más utilizadas de árboles de decisión se encuentran ID3 (*Iterative Dichotomizer 3*) [58] que implementan la clasificación utilizando entropía.

La Ecuación (3.1) define la entropía que utiliza la incertidumbre de la información como medida de evaluación, requiere que tanto las características independientes como la dependiente o predictora sean de tipo categórico. Utiliza el producto de la probabilidad de clase x_i por el logaritmo en base 2 de la probabilidad de la misma clase [61].

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \tag{3.1}$$

C4.5 [59] es una mejora de ID3, permite que las características independientes sean categóricas o continuas así como también implementa la tarea de regresión.

CART (Classification And Regression Trees) [60] que como su nombre lo indica implementa clasificación y regresión con la opción de utilizar la entropía o el índice Gini de pureza de la información como método de evaluación de las particiones.

La Ecuación (3.2) define al índice Gini en dónde p_i representa la frecuencia de la clase i dentro de un conjunto de datos de tamaño J. El índice Gini ($I_G(p)$) determina la pureza para la clase i [61].

$$\mathrm{I}_G(p) = 1 - \sum_{i=1}^{J} p_i^2 \tag{3.2}$$

La idea base del algoritmo es la separación del conjunto de datos en particiones binarias (aunque también es posible un mayor número de particiones), cada una de estas particiones es evaluada para determinar cual resulta en la mejor decisión, para determinar las regiones específicas que delimitaran dichas particiones (frontera de decisión) el algoritmo debe calcular todos los umbrales factibles y seleccionar el mejor de ellos.

Dadas las características de CART antes mecionadas, es seleccionado para integrar la fase de experimentación del presente trabajo de investigación. Dentro de los hiperparámetros relevantes del modelo CART se encuentra los siguientes:

- **criterion**: Especifica el criterio de evaluación a utilizar entropía o índice Gini.

- **max_depth**: Máxima profundidad del árbol. Se utiliza como estrategia de control de sobre ajuste. Se le conoce como estrategia de pre poda (*pre pruning*) debido a que se aplica previo a la construcción del árbol.
- **max_features**: Número máximo de características independientes a utilizar.
- **ccp_alpha**: Poda de Complejidad de Costo Mínimo. Se utiliza como estrategia de control de sobre ajuste. Se le conoce como estrategia de post poda (*post pruning*) debido a que se aplica posterior a la construcción del árbol.

Entre las ventajas de los árboles de decisión se encuentran las siguientes:

- **Alto nivel de interpretabilidad**. Ya sea de manera gráfica o a través de los valores de los nodos resulta sencillo realizar una interpretación.
- **Bajo nivel de pre procesamiento de los datos**. Dado que los árboles de decisión pueden trabajar con características categóricas y continuas por igual no es necesario una procesamiento excesivo de los datos.

Entre las desventajas de los árboles de decisión es que suelen presentar una alta sensibilidad al conjunto de datos de entrenamiento por lo tanto suelen mantener un nivel de sesgo bajo que podría observarse como una ventaja. Esta situación también tiene un efecto sobre la tasa de error respecto a los datos del conjunto de prueba generando en ocasiones un sobre ajuste (*overfitting*).

3.1.2. Bosque aleatorio (*Random forest*)

El bosque aleatorio (*Random forest*) es un algoritmo propuesto por Leo Breiman que al estar basado en los árboles de decisión tiene el objetivo de conservar las ventajas de este algoritmo además de mejorar sus áreas de oportunidad, en específico, la alta sensibilidad al conjunto de datos de entrenamiento.

La intuición de la que parte el diseño del bosque aleatorio es que se producen mejoras significativas en la exactitud de las tareas de clasificación al entrenar un ensamble de árboles de decisión y permitirles votar por la clase más popular [62].

Bajo este concepto el término bosque encuentra su origen en el hecho de generar un ensamble desde la fase de entrenamiento, es decir, se entrena una cantidad determinada de árboles de decisión, en la práctica se sugiere establecer esta cantidad a partir de 100.

Mientras que el término aleatorio viene dado por el hecho de que cada uno de los árboles de decisión es entrenado con un subconjunto aleatorio del conjunto de datos original. Estos subconjuntos aleatorios tienen la característica de ser generados bajo un enfoque de muestreo con reemplazo (*bootstrapping*) lo cual significa que una observación podría repetirse dentro de un mismo subconjunto así como entre diferentes subconjuntos.

Reforzando el carácter aleatorio del algoritmo el número de características independientes también es determinado de esta manera, es decir, no se utilizan todas las

características independientes disponibles en cada subconjunto de entrenamiento, sino que bajo la especificación de un hiperparámetro se utiliza únicamente un proporción aleatoria de ellas.

La siguiente fase del proceso consiste en establecer un mecanismo de votos para obtener un resultado unificado a partir de los resultados unitarios de cada uno de los árboles entrenados. En el caso de la tarea de clasificación que es la que nos compete, el mecanismo de voto se define como la categoría en la que la mayoría de los árboles clasificaron al dato.

Respecto a los hiperparámetros por definir para el algoritmo de bosque aleatorio se encuentra una base proveniente de árboles de decisión como son: el criterio de evaluación, la profundidad máxima y el número máximo de características, además de incluir entre otros:

- **n_estimators**: Define el número de árboles que integrarán el bosque, como se mencionó anteriormente es recomendable experimentar a partir de 100.
- **bootstrap**: Especifica si se utilizará la estrategia de muestreo por reemplazo o no. Se recomienda activarlo.

Entre las ventajas del bosque aleatorio se cuentan las siguientes:

- **Fácil entrenamiento**. La fase de entrenamiento del bosque aleatorio resulta
- **Reducción del sobre ajuste**. Al ser un algoritmo de ensamble con un mecanismo de votos permite reducir significativamente la sensibilidad al conjunto de datos de entrenamiento.
- **Estabilidad**. Su rendimiento es homogéneo a través de los diferentes conjuntos de datos así como con conjuntos de datos de gran volumen.

3.1.3. Potenciamiento del gradiente (*Gradient boosting*)

En general las técnicas de potenciamiento (*boosting*) utilizan un ensamble de predictores o clasificadores "base", este ensamble de clasificadores es conformado en una estructura de comité cuyo objetivo es que su decisión final mejore significativamente el rendimiento individual de cada clasificador [63].

Al igual que el bosque aleatorio, el potenciamiento del gradiente es un algoritmo de ensamble basado en los árboles de decisión, creando un conjunto de instancias determinado y ejecutando la fase de entrenamiento de forma múltiple. La intuición sobre los ensambles es conservar las ventajas que ofrece el algoritmo base y a través de la multiplicidad proponer mecanismos que mejoren sustancialmente posibles desventajas que pudiera llegar a presentar dicho algoritmo base.

En el caso de los árboles de decisión, como se mencionó anteriormente, una de sus cualidades positivas es un bajo sesgo durante la fase de entrenamiento. En contraparte

una cualidad negativa es que presenta una alta varianza en la fase de prueba; esta cualidad es la que propone mejorar el potenciamiento del gradiente.

Como vimos anteriormente en el bosque aleatorio los predictores (árboles de decisión) se entrenan de forma independiente para finalmente establecer un mecanismo de votos de la totalidad de predictores. Por su parte, en el potenciamiento del gradiente el entrenamiento se lleva a cabo de forma secuencial con la particularidad de que cada nuevo predictor (*weak learner*) toma en cuenta los resultados del anterior con el objetivo de mejorar su rendimiento.

Entre los hiperparámetros relevantes para el potenciamiento del gradiente se encuentran los siguientes:

- **n_estimators**: Define el número de predictores que integrarán el ensamble, como se mencionó anteriormente es recomendable experimentar a partir de 100.
- **lerning_rate**: Controla el ritmo con el que el modelo aprende.
- **max_depth**: Máxima profundidad del árbol. Se utiliza como estrategia de control de sobre ajuste. Se le conoce como estrategia de pre poda (*pre pruning*) debido a que se aplica previo a la construcción del árbol.

Entre las principales características del potenciamiento del gradiente se cuentan las siguientes:

- **Alta volumen y dimensionalidad**. El potenciamiento del gradiente cuenta con la capacidad de manejar de manera adecuada grandes volúmenes de datos así como grandes espacios de características conservando su rendimiento.
- **Alto rendimiento**. Mantiene un alto rendimiento para la tarea de clasificación.

3.1.4. Máquinas de vectores de soporte (*SVM - Support vector machines*)

Las máquinas de vectores de soporte o *SVM* por sus siglas en inglés son una familia de algoritmos de aprendizaje de máquinas supervisado diseñadas por Vladimir Vapnik para las tareas de clasificación y regresión [64].

En el caso de la clasificación binaria la idea central del algoritmo es la de encontrar un hiperplano capaz de separar las categorías especificadas por las etiquetas del conjunto de datos.

Se parte del supuesto que los datos son linealmente separables y podría existir más de un hiperplano factible. A partir de estos supuestos, en primera instancia el algoritmo requiere encontrar el hiperplano óptimo.

El hiperplano óptimo se define como aquel que conserva la mayor distancia hacia las observaciones de ambas categorías. En este sentido el procedimiento es el siguiente:

1. Detectar las dos observaciones más cercanas entre sí pertenecientes a diferentes categorías (vectores de soporte).
2. Calcular la distancia euclidiana entre ambas observaciones así como la recta imaginaria que los une.
3. Calcular una recta perpendicular equidistante a ambas observaciones (hiperplano óptimo).
4. Calcular los hiperplanos paralelos al hiperplano óptimo tocando a los vectores de soporte, estos hiperplanos están definidos como los márgenes.

Adicionalmente existen estrategias enfocadas en aportar ajustes al funcionamiento del algoritmo, a continuación repasaremos algunas de ellas.

3.1.4.1. Estrategia de margen duro (*Hard margin*)

De manera ideal se busca maximizar el tamaño del margen, es decir, la distancia entre el hiperplano óptimo y los hiperplanos generados a partir de los vectores de soporte, a esta versión del algoritmo se le conoce como margen duro.

El problema con el margen duro es que este es poco flexible, es sensible a valores atípicos (*outliers*) elevando en estos casos la tasa de error.

3.1.4.2. Estrategia de margen suave (*Soft margin*)

Una mejora al algoritmo es la versión de margen suave en la cual se incluye un hiperparámetro definido como C cuyo valor incide directamente en el tamaño del margen. A mayor valor de C, menor será el margen y a menor valor en C mayor será el margen.

En la práctica, al igual que cualquier otro hiperparámetro el valor de C requiere experimentación para establecer el valor adecuado a cada conjunto de datos.

3.1.4.3. Estrategia del *kernel*

Dado que existe una amplia variedad de conjuntos de datos que de origen no son linealmente separables, el algoritmo de máquinas de vectores de soporte integra la estrategia del *kernel*.

El aumento de la dimensionalidad se realiza a través de mapear el conjunto de datos original hacia un espacio de mayor dimensión a través de una función no lineal [65]. Estas funciones suelen ser polinomiales que obtienen las nuevas dimensiones a partir de la combinatoria de las características originales con potencias mayores a uno y funciones de base radial que generan el efecto de la campana de Gauss en las características seleccionadas. En ambos casos el efecto que se logra es alcanzar la separabilidad lineal.

Entre los hiperparámetros relevantes para la configuración del algoritmo se encuentran los siguientes:

- **C**: Parámetro de regularización, activa la funcionalidad de margen suave.
- **kernel**: Activa la estrategia del *kernel* con opciones a utilizar las funciones polinomial, sigmoide, lineal y rbf.
- **class_weight**: permite especificar de antemano el nivel de balance de las clases.

Una de las ventajas de las máquinas de vectores de soporte es la siguiente:

- **Alta dimensionalidad**. SVM cuenta con la capacidad de manejar conjuntos de datos con grandes espacios de características.
- •

3.2. Aprendizaje de máquinas no supervisado

Las técnicas de aprendizaje de máquinas no supervisado no cuentan a priori con el valioso conocimiento que significan las características predictoras (etiquetas) para las técnicas de aprendizaje de máquinas supervisados.

De tal manera que éstos algoritmos utilizan las características independientes para aprender su estructura, relaciones y patrones ocultos con el objetivo de reconocer sus diferencias. También son especialmente adecuados para detectar anomalías dentro de los conjuntos de datos.

Las tareas abordadas por los algoritmos de aprendizaje no supervisado son:

- **Agrupación (*Clustering*)**. Utilizan el aprendizaje para separar los datos en grupos de estrecha similutud, de tal forma que las muestras de un mismo grupo contengan características similares mientras que las muestras en diferentes grupos contengan características diferenciadas.
- **Detección de anomalías**. Al contar con la capacidad de detectar la estructura y patrones subyacentes de los datos también es capaz de discriminar valores atípicos.
- **Reducción de la dimensionalidad (*Dimensionality reduction*)**. Son técnicas enfocadas en analizar las características desde el punto de vista de su similutud y valor por aportar en un modelo predictivo, el objetivo es encontrar la máxima representatividad con el mínimo de características.

A continuación revisaremos algunos algoritmos de aprendizaje de máquinas no supervisado utilizados para abordar el problema de la detección de fraudes en transacciones bancarias sobre medios electrónicos.

3.2.1. Bosque de aislamiento (*Isolation forest*)

Los algoritmos diseñados para la tarea de detección de anomalías suelen partir de un enfoque de descripción detallada de la estructura de los datos "normales" para a partir de ahí discriminar los valores atípicos o anomalías; sin embargo, el bosque de aislamiento propone un enfoque diferenciado en el cual directamente se aíslan las anomalías partiendo de dos ventajas cuantitativas que poseen las anomalías:

- Se trata de la minoría de las observaciones
- Contienen valores diferenciables respecto a las observaciones "normales".

Es decir, bosque de aislamiento parte de la intuición de que las anomalías son pocas y diferentes, lo que las convierte en altamente susceptibles al aislamiento respecto a las observaciones normales [66].

Metodológicamente las similitudes con *Random forest* son diversas, se trata de un algoritmo de ensamble, basado en árboles como clasificadores base, cada árbol se entrena con un subconjunto aleatorio de los datos generado a través de la técnica de muestreo con reemplazo (*bootstrapping*).

Los árboles de aislamiento se fundamentan en la especificación de que las anomalías son aquellas observaciones cuya longitud de trayecto es la más corta, dado que las observaciones normales son aquellos puntos cuyo trayecto es más profundo en el árbol [66].

Entre las ventajas proporcionadas por el bosque de aislamiento se encuentran las siguientes:

- **Bajo costo computacional**. A diferencias de otros algoritmos de detección de anomalías basados en el cálculo de la distancia o de la densidad, el bosque aleatorio elimina la complejidad de dichos cálculos.
- **Complejidad lineal y bajos requerimientos de memoria**.
- **Capacidad de escalamiento**. Cuenta con la capacidad de manejar grandes cantidades de observaciones y conjuntos de datos de alta dimensionalidad.

Entre los hiperparámetros relevantes del bosque de aislamiento, se encuentran:

- **n_estimators**: Define el número de árboles que integrarán el bosque, como se mencionó anteriormente es recomendable experimentar a partir de 100.
- **max_samples**: Especifica el número de observaciones por utilizar durante la fase de entrenamiento de cada árbol.
- **contamination**: Establece la tasa de anomalías esperadas dentro del conjunto de datos.

3.2.2. Factor de valor atípico local (*LOF - Local outlier factor*)

LOF es un algoritmo de detección de anomalías cuya propuesta principal es la asignación de grado de anomalía (*outlier factor*) a cada observación del conjunto de datos. El grado de anomalía es local debido a que su cálculo se restringe a un subconjunto conformado por observaciones cercanas [67].

Una de las marcadas diferencias entre LOF y otros algoritmos orientados a la detección de anomalías es el hecho de que para LOF el concepto de anomalía no es una propiedad binaria.

El concepto de agrupación (*cluster*) construido a partir de la cercanía de las observaciones únicamente cobra relevancia como paso preliminar para el cálculo del grado de anomalía.

Entre los hiperparámetros relevantes de LOF se encuentran:

- **n_neighbors**: Número de vecinos a considerar para el cálculo del factor de anomalía.
- **algorithm**: Se utiliza para determinar el algoritmo a utilizar para el cálculo de los vecinos más cercanos.
- **metric**: Se utiliza para especificar la métrica a utilizar para calcular la distancia.
- **contamination**: Establece la tasa de anomalías esperadas dentro del conjunto de datos.

Entre las ventajas del factor de valor atípico local están las siguientes:

- **No se asume una distribución en particular**. A diferencia de otros algoritmos en los cuales su correcto funcionamiento asume que los datos siguen una distribución en particular para LOF este no es un inconveniente.
- **Probabilidad basada en la densidad**. A diferencia de otros algoritmos de detección de anomalías basados en la distancia LOF se basa en la densidad lo cual permite reducir la tasa de anomalías falsas.

3.3. Métricas de evaluación

Derivado de la naturaleza desequilibrada del problema de la detección de fraudes sobre transferencias electrónicas, la selección de las métricas del desempeño de los modelos de aprendizaje de máquinas requiere particular cuidado debido a que la asimetría entre las clases dificulta la adecuada identificación de la tasa real de fraudes [68]. Al tratarse de una tarea de clasificación binaria se cuenta con una amplia variedad de métricas basadas en la matriz de confusión, a continuación se describirán las métricas seleccionadas para evaluar los modelos de aprendizaje de máquinas.

		Clase Inferida		
		P	N	Instancias
Clase real	P	TP	FN	m_P
	N	FP	TN	m_N
Inferencias		e_P	e_N	m

Tabla 3.1: Estructura de la matriz de confusión

3.3.1. Matriz de confusión

La matriz de confusión es una herramienta utilizada para la evaluación de modelos de aprendizaje de máquinas abordando tareas de clasificación. Se trata de una matriz cuadrada cuyas dimensiones están dadas por el número de clases o categorías, en el caso del problema de clasificación binaria de la detección de fraudes en transferencias de banca por internet la matriz de confusión tiene tamaño 2×2.

El valor de la matriz de confusión se encuentra en el registro simultáneo de la información generada por los predictores o modelos de aprendizaje de máquinas y a la par de la información real, recordando que la tarea de clasificación es abordada principalmente a través de modelos de aprendizaje de máquinas supervisados.

La estructura de la matriz de confusión se ilustra en la tabla 3.1.

Donde:

- **TP (*True Positive* - Verdaderos Positivos)**. Instancias positivas en la realidad y que fueron inferidas como positivas por el modelo de aprendizaje de máquinas.
- **FN (*False Negative* - Falsos Negativos)**. Instancias positivas en la realidad, pero que fueron inferidas como negativas por el modelo de aprendizaje de máquinas.
- **FP (*False Positive* - Falsos Positivos)**. Instancias negativas en la realidad, pero que fueron inferidas como positivas por el modelo de aprendizaje de máquinas.
- **TN (*True Negative* - Verdaderos Negativos)**. Instancias negativas en la realidad y que fueron inferidas como negativas por el modelo de aprendizaje de máquinas.

El término "confusión" viene del hecho de contrastar la información real con la información inferida, permitiendo de esta manera detectar el error de clasificación (la confusión) del modelo de aprendizaje de máquinas [69].

A partir de la matriz de confusión se genera una serie de métricas orientadas en ponderar diferentes aspectos de la clasificación. La selección de la o las métricas adecuadas para cada caso de estudio depende en primer lugar del conocimiento del dominio. En ocasiones se requiere dar mayor peso o valor a la clase positiva y viceversa. Como se mencionó anteriormente, también existen métricas que son más adecuadas para problemas de clasificación equilibrada, algunas otras pensadas en ponderar adecuadamente los problemas de naturaleza desequilibrada.

A continuación se especifican las métricas seleccionadas específicamente para el problema de la clasificación binaria de la detección de fraudes de banca por internet.

3.3.1.1. Precisión

Esta métrica también conocida como Valor Predictivo Positivo - *Positive Predictive Value (PPV)* tiene por objetivo determinar la tasa o porcentaje de las instancias inferidas como positivas que en la realidad son positivas [68].

$$Precision = \frac{TP}{TP + FP} \tag{3.3}$$

En nuestro caso, la clase positiva identifica a las transferencias fraudulentas, por lo tanto, esta métrica es particularmente importante.

3.3.1.2. Sensibilidad (*Recall*)

Esta métrica también conocida como Sensibilidad (*Sensitivity*), Tasa de Aciertos (*Hit Rate*) o Tasa de Verdaderos Positivos - *True Positive Rate (TPR)* tiene por objetivo determinar la tasa o porcentaje de las instancias positivas reales inferidas correctamente [68].

$$Recall = \frac{TP}{TP + FN} \tag{3.4}$$

3.3.1.3. ROC AUC

Esta métrica se compone de la Curva de la Característica del Operador del Receptor - *Receiver Operator Characteristic (ROC) curve* que es una métrica para la clasificación binaria que pondera la relación de la sensibilidad *(TPR)* respecto a la tasa de falsos positivos - *False Positive Rate (FPR)* a través de diversos puntos en el umbral comprendido entre 0 y 1. El Área Bajo la Curva - *Area Under the Curve (AUC)* representa el valor numérico de la curva (gráfica) y determina la capacidad de clasificación del modelo bajo inspección.

3.3.1.4. F_1

Esta métrica también conocida como Promedio Armónico de la Precisión y la Sensibilidad tiene como objetivo combinar ambas métricas que operan sobre las instancias positivas para crear un promedio "*armónico*" a partir de ellas [70].

$$F_1 score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} = \frac{2TP}{2TP + FP + FN} \tag{3.5}$$

Desde el punto de vista de negocio la *Precisión* obtiene la tasa de transferencias fraudulentas detectadas en relación a las transferencias legítimas inferidas incorrectamente como transferencias fraudulentas, entre mayor sea el valor de la precisión menor es el grado del error, este error tienen un impacto no monetario ya que significa clientes molestos por obtener transferencias legítimas bloqueadas; mientras que la *Sensibilidad* obtiene la tasa de transferencias fraudulentas detectadas en relación a las transferencias fraudulentas inferidas incorrectamente como legítimas, entre mayor sea el valor de la sensibilidad menor es el grado del error, este error tiene un impacto económico ya que significa transferencias fraudulentas no detectadas cuyo monto eventualmente habrá que absorber.

Por lo anterior, F_1 cobra relevancia en la detección de transferencias fraudulentas de banca por internet, ya que al representar un promedio de la precisión y la sensibilidad permite mantener un equilibrio y buscar el modelo que minimice ambas de forma simultánea.

3.3.1.5. F_β

Esta métrica esta basada en F_1 *score* y su aportación es la capacidad de ponderar los valores de la precisión y la sensibilidad a través del parámetro β que determina la cantidad de veces que se considera más importante la sensibilidad respecto a la precisión.

$$F_\beta = (1 + \beta^2) \cdot \frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \tag{3.6}$$

En el caso de la detección de transferencias fraudulentas en banca por internet, si bien, ambas métricas son importantes se puede optar por dar mayor peso a la sensibilidad dado que su impacto es económico y de corto plazo.

Para los experimentos realizados se consideró un valor de 0.5 para β.

3.3.1.6. MCC

El Coeficiente de Correlación de Mathews (*Mathews Correlation Coefficient*) es una métrica desarrollada específicamente para tareas de clasificación binaria cuyas clases presentan un nivel de desequilibrio mayor. El aporte de esta métrica es su capacidad de

generar una ponderación equilibrada independientemente de la cantidad de muestras en cada clase dentro de los datos. [71]

$$MCC = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \tag{3.7}$$

Capítulo 4

Metodología y diseño experimental

En este capítulo se describe detalladamente el procedimiento metodológico ejecutado para integrar la información de las diversas fuentes de datos involucradas en una sola base de datos relacional. Al cierre de la fase de integración de datos se realiza un análisis de la características integradas con el objeto de determinar los alcances de la fase de limpieza de los datos. Posteriormente se realiza el análisis exploratorio de datos para realizar un reconocimiento de la información contenida en la base de datos integrada. Finalmente, se realiza la selección de características en dónde se determina qué datos se incluirán en los modelos de aprendizaje de máquinas propuestos para la fase de experimentación.

1. Integración de fuentes de información
    - Análisis de las fuentes de datos.
    - Extracción e integración de datos desde fuentes no estructuradas.
    - Análisis de compatibilidad de datos.
    - Diseño de modelo de datos integrado.
2. Limpieza de datos
    - Búsqueda de datos nulos o vacíos, estrategia de llenado o descarte.
    - Análisis de datos categóricos faltantes y estrategia de llenado.
3. Análisis exploratorio de datos
    - Análisis estadístico de datos continuos.
    - Análisis de las correlaciones entre las características seleccionadas.
    - Búsqueda de patrones en los datos.
    - Análisis de las distribuciones de las características.
4. Selección de características

- Con base en el análisis exploratorio seleccionar o descartar características para integrar los modelos de aprendizaje de máquinas.
- Determinar si se deben incorporar datos extendidos.

4.1. Integración de fuentes de datos

La integración de fuentes de datos es una actividad analítica para la cual se requiere un perfil profesional que integre una serie de conocimientos técnicos y de negocio. Por la parte técnica, infraestructura de sistemas (topologías de redes, configuración y acceso a servidores), entornos de ejecución de sistemas y aplicaciones (administración de servidores, ejecución y configuración de servidores de aplicaciones), instalación, configuración y acceso a bases de datos, etc. Por otro lado se requiere un profundo conocimiento del negocio y la interacción de las diversas áreas operativas y administrativas en el ecosistema de componentes, aplicaciones y sistemas que soportan la operación de la banca por internet.

En el caso particular del entorno de banca por internet en el cual se basa el presente trabajo de investigación se integró información de tres diferentes tipos de fuentes de datos:

1. Bases de datos relacionales de tres sistemas bancarios.
2. Los archivos en formato de texto simple de la bitácora transaccional de la aplicación de banca por internet.
3. Los archivos de informe de dictamen de investigación de reportes de fraude.

A continuación se describe cada una de las fuentes de datos seleccionadas para el proceso de integración especificando la naturaleza de la información contenida.

4.1.1. Base de datos de la banca por internet

La base de datos de las aplicaciones que conforman el ecosistema de banca por internet y móvil se encuentra implementado en un sistema de gestión de bases de datos relacionales y cuenta con 153 relaciones (tablas), sin embargo, para los fines del presente trabajo de investigación requerimos ocho relaciones. La Figura A.1 muestra el diagrama entidad - relación de la selección de relaciones de la base de datos de la banca por internet por utilizar incluyendo los atributos de cada relación, mientras que la Tabla 4.1 contiene la descripción funcional de cada relación.

La información seleccionada de esta base de datos es la siguiente:

- Número de cliente.
- Identificador del usuario de banca por internet (*login*)

Relación	Descripción
CAT_CLIENT_TYPE	Catálogo de tipos de clientes, define si se trata de un cliente persona física o moral (empresa).
TBL_CLIENT	Información del cliente, en específico requerimos del número de cliente, el cual es el identificador del cliente y es originado en el sistema nuclear del banco (core bancario), dentro de la base de datos de la banca por internet se replica dicho número y se mantiene como llave única del dato.
CAT_ACCOUNT_TYPE	Catálogo de tipo de cuenta bancaria, define el tipo de producto bancario con el cual se realizan las transacciones.
TBL_ACCOUNT	Información de la cuenta bancaria, en particular utilizamos el número de cuenta, el cual es el identificador de la cuenta y es originado en el sistema nuclear del banco (core bancario), dentro de la base de datos de la banca por internet se replica dicho número y se mantiene como llave única del dato. Mantiene una relación n:1 con TBL_CLIENT, es decir, un elemento de TBL_CLIENT puede estar relacionado con n elementos en TBL_ACCOUNT
CAT_TRANSACTION_STATUS	Catálogo de estatus de las transacciones, define si la transacción fue ejecutada con éxito, cancelada o rechazada. En particular nos interesan las transacciones ejecutadas con éxito.
TRX_TRANSACTION	Información de las transacciones, esta relación funciona como base para una serie de transacciones como transferencias bancarias, pagos de servicios, pagos a tarjeta de crédito, pagos de impuestos del SAT y pagos a la tesorería de la Ciudad de México. En particular nos interesan las transferencias bancarias. Esta relación proporciona el número de cuenta origen y el monto que son los datos comunes a todos los tipos de transacciones. Mantiene una relación n:1 con TBL_ACCOUNT, es decir, un elemento de TBL_ACCOUNT puede estar relacionado con n elementos de TRX_TRANSACTION.
TRX_BANK_TRANSFER	Información de las transferencias bancarias, extiende de TRX_TRANSACTION e incorpora el número de cuenta destino. Mantiene una relación 1:1 con TRX_TRANSATION, es decir, cada elemento en TRX_BANK_TRANSFER está relacionado exactamente con 1 elemento en TRX_TRANSACTION.
TBL_USER	Información del usuario de banca por internet, incorpora el identificador del usuario en el sistema de banca por internet (*login*). Mantiene una relación n:1 con TBL_CLIENT, es decir, un cliente puede tener uno o más usuarios. De manera práctica, los clientes tipo persona física están restringidos a sólo un usuario, mientras que los clientes de tipo moral (empresas) sí pueden tener n usuarios,

Tabla 4.1: Relaciones de la base de datos de la banca por internet por utilizar

Relación	Descripción
TBL_LOGIN	Información de las credenciales de acceso a las diversas aplicaciones del ecosistema bancario. Incorpora el identificador del usuario, el cual es el mismo que se tiene en la base de datos de la banca por internet, por lo cual se podrá realizar el enlace utilizando dicho dato como criterio de relación. Se contempla la incorporación de las marcas de tiempo de los eventos de inicio y fin de sesión dentro del sistema de banca por internet que se obtendrá de las bitácoras.

Tabla 4.2: Relación de la base de datos del sistema integral de autenticación.

- Número de cuenta origen.
- Número de cuenta destino.
- Folio de la transacción.
- Monto de la transferencia.
- Fecha de captura de la transferencia.

4.1.2. Base de datos del sistema integral de autenticación

Es común entre las instituciones del sistema financiero contar con un sistema integral de autenticación, a través del cual se proporciona cohesión a los datos de autenticación de los diversos sistemas y aplicaciones del banco. Desde el punto de vista de los clientes del banco esto resulta benéfico ya que pueden mantener un conjunto de credenciales (identificador, palabra e imagen secreta y token o contraseña de una sola vez) para todas las aplicaciones que requiera acceder (banca por internet, banca móvil, banca telefónica, etc).

En este sistema se almacena el registro de los eventos de inicio y fin de sesión de diversas aplicaciones bancarias entre ellas la banca por internet.

Esta base de datos también se encuentra implementado en un sistema de gestión de bases de datos relacionales y para el proceso de integración de información únicamente requerimos de una relación. La Figura A.2 muestra el diagrama entidad - relación de la relación seleccionada de la base de datos del sistema integral de autenticación por utilizar incluyendo sus atributos, mientras que la Tabla 4.2 contiene la descripción funcional de la relación.

El mecanismo para establecer un vínculo entre el sistema integral de autenticación y la banca por internet es a través del identificador del usuario de la banca por internet que existe en ambos sistemas y cuyo origen es la banca por internet.

Una vez vinculada la información es necesario generar un procedimiento que calcule la cercanía entre el par de eventos inicio y fin de sesión con una o más transferencias, por regla, las transferencias deben estar dentro de los límites de inicio y fin de sesión, de esta forma al integrar la información del sistema integral de autenticación se podrá asignar el registro de inicio y fin de sesión a cada transferencia.

La información seleccionada de esta base de datos es la siguiente:

Relación	Descripción
TRX_PREREGISTER	Información de las cuentas origen y destino registradas por el cliente. Esta información es validada al momento de ejecutar las transferencias, otra restricción es que debe transcurrir al menos media hora posterior al registro para poder realizar la transferencia. Incorpora los números de cuenta de las cuentas origen y destino, con lo cual se podrá realizar el enlace de la información utilizando dicho dato como criterio de relación. También incorpora la marca de tiempo del registro de las cuentas.

Tabla 4.3: Relación de la base de datos del sistema de identificación transaccional

- Marca de tiempo del evento de inicio de sesión en la banca por internet.
- Marca de tiempo del evento de fin de sesión en la banca por internet.

4.1.3. Base de datos del sistema de identificación transaccional

Se trata de un sistema interno bancario en el cual se concentra información referente a la variedad de transacciones que se pueden realizar en la institución bancaria, ya sea por medios físicos o electrónicos. La naturaleza de la información es variada, sin embargo, para fines del proceso de integración de información es concerniente la información específica de las transacciones registradas por los clientes, para lo cual únicamente requerimos una relación. La Figura A.3 muestra el diagrama entidad - relación de la relación seleccionada de la basae de datos del sistema de identificación transaccional por utilizar incluyendo sus atributos, mientras que la Tabla 4.3 contiene la descripción funcional de la relación.

El mecanismo para establecer el vínculo entre el sistema de identificación transaccional y la banca por internet es la cuenta origen, dato originado en el sistema nuclear del banco y que también existe en ambos sistemas. Se filtran únicamente los registros de cuentas destino en estatus “Activo”.

La información seleccionada de esta base de datos es la siguiente:

- Cuenta origen.
- Cuenta destino.
- Marca de tiempo del registro de cuenta destino.

4.1.4. Informe de dictamen de investigación de reportes de fraude

El proceso para determinar que una transacción es fraudulenta requiere la intervención de diversas áreas internas de la institución financiera y requiere varios días y en ocasiones semanas para establecer el dictamen final.

De manera resumida el proceso de dictamen es el siguiente:

1. El usuario se comunica al centro de atención telefónica para informar de una transacción no reconocida.

2. El centro de atención telefónica levanta el reporte y genera un identificador para el seguimiento del usuario así como seguimiento del proceso interno.
3. De forma automática es enviada una notificación a las áreas de Banca Electrónica, Seguridad de la Información y Gestión de Riesgos.
4. Cada una de las áreas comienza una investigación dentro de su área de competencia acerca de la legitimidad de la transacción.
5. El área de gestión de riesgos integra la información de las investigaciones por área y emite el dictamen final.

El dictamen es un archivo electrónico en formato PDF en el cual se identifica a la transacción por el folio de la transacción generado durante la ejecución, y al cliente se le identifica por el número de cliente y el identificador de sesión de banca por internet (*login*).

Se cuenta con los archivos electrónicos de dictamen de investigación de transferencia fraudulenta de 783 reportes para el periodo comprendido entre 2015 y 2019, de los cuales fueron dictaminados como fraude) 269.

El vínculo entre la base de datos de la banca por internet y el archivo de dictamen de investigación de reportes de fraude es el folio de la transacción que garantiza la unicidad.

La información seleccionada de esta fuente de datos es la siguiente:

- Resultado del dictamen [Transacción Legítima - Transacción Fraudulenta].

Se contempla ejecutar un procedimiento para integrar esta información en la base de datos de la banca por internet con fines de tenerlo disponible para el proceso de detección de fraudes con técnicas de aprendizaje de máquinas.

4.1.5. Archivo de la bitácora transaccional de la banca por internet

La base de datos de la banca por internet contiene la información de las transferencias ejecutadas, sin embargo, la información detallada de la navegación del usuario dentro de la sesión de la banca por internet no es almacenada dentro de ninguna relación de la base de datos. Lo anterior debido principalmente a las especificaciones de rendimiento que deben cumplir un canal digital, se requiere que los tiempos de respuesta dentro de la sesión de banca por internet sean "inmediatos", es decir, en el orden de milisegundos. Sin embargo, existen datos como las marcas de tiempo de los eventos de navegación que son relevantes para la detección de fraudes.

Los archivos de la bitácora transaccional son generados por el software contenedor de aplicaciones (*AppServer*) en un formato de texto simple. La información dentro de los archivos se escribe un evento por línea de manera sincrónica y simultánea para todos los usuarios que acceden a sesiones de banca por internet.

La Figura 4.1 muestra un ejemplo de la información que se puede encontrar dentro de la bitácora transaccional para un sólo usuario. Se trata de diversos eventos como:

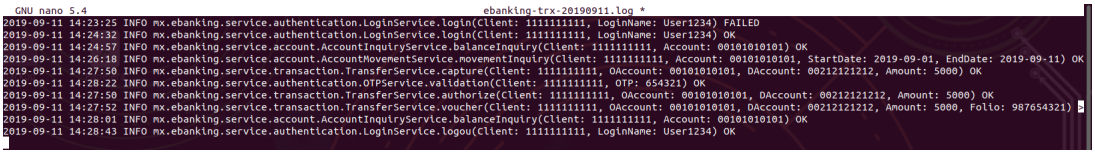

Figura 4.1: Ejemplo ilustrativo de la información contenida en los archivos de la bitácora transaccional.

- Inicio de sesión (*LoginService.login*)
- Consulta de saldo (*AccountInquiryService.balanceInquiry*)
- Consulta de movimientos (*AccounMovementService.movementInquiry*)
- Captura de transferencia (*TransferService.capture*)
- Validación de la contraseña de una sola vez (*OTPService.validation*)
- Autorización de la transacción (*TransferService.authorization*)
- Generación del comprobante de la trasferencia (*TransferService.voucher*)
- Consulta de saldo (*AccountInquiryService.balanceInquiry*)
- Cierre de sesión (*LoginService.logout*)

En todos los casos se incluye la marca de tiempo del evento y el identificador del cliente que ejecuta el evento. Adicionalmente, cada evento incluye información particular como puede ser:

- El identificador de usuario, en los casos del inicio y fin de sesión.
- El número de cuenta, en los casos de las consultas de saldo y movimientos.
- Los números de cuenta origen y destino así como el monto, en los casos de los eventos relacionados a las transferencias.

La Figura A.4 muestra el diagrama entidad - relación diseñado para integrar la información de la bitácora transaccional mientras que la Tabla 4.4 contiene la descripción funcional de la nueva relación.

Por lo anterior, se contempla el desarrollo de un componente de software especializado en la lectura de los archivos de la bitácora transaccional y extracción de la información relevante hacia una relación dentro de la base de datos.

La información seleccionada de esta fuente de datos es la siguiente:

- Marca de tiempo del evento de captura de la transferencia.

Relación	Descripción
TRX_LOGBOOK	Registro de eventos transcurridos dentro de los diversos componentes que conforman la aplicación de la banca por internet. Se registran los flujos de información a través de las capas de acceso a datos, servicios (transaccionalidad y lógica de negocio) e interfaz gráfica de usuario.

Tabla 4.4: Relación de la base de datos de la bitácora transaccional.

4.1.6. Modelo de datos integrado

Resultado del análisis de las diversas fuentes de información se realiza la integración de las relaciones seleccionadas ya existentes en dichas fuentes así como las nuevas relaciones diseñadas específicamente para el proyecto de investigación de detección de fraudes. La Figura 4.2 presenta el modelo de datos integrado.

A partir del modelo de datos integrado y tomando como punto de partida la base de datos de la banca por internet se crean las sentencias SQL necesarias para crear las relaciones diseñadas para extender la base de datos. Mediante un proceso de importación de datos se carga la información específica de cada relación en la base de datos integrada.

4.1.7. Conjunto de datos

De forma preliminar a partir de la base de datos integrada se diseñó una consulta en lenguaje SQL para generar el archivo con el conjunto de datos. La Tabla 4.5 contiene la descripción de las características incluidas en la primer versión del archivo del conjunto de datos que es utilizado como base para el proceso de limpieza de datos y un primer análisis exploratorio de datos.

Atributo	Tipo	Descripción
ORIGIN_ACCOUNT	Categórico	Es el identificador de la cuenta bancaria de donde salen los fondos de la transferencia.
DESTINATION_ACCOUNT	Categórico	Es el identificador de la cuenta bancaria hacia donde se transmiten los fondos de la transferencia.
AMOUNT	Continuo	Monto de la transferencia.
ACC_OPENING_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra el alta de la cuenta origen en el sistema central del banco.
PREREG_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra la cuenta destino como opción de transferencia en la aplicación de la banca por internet.
LOGIN_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra el inicio de la sesión en la banca por internet.
TRX_AUTHORIZATION_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra la autorización de la transferencia.

Continua en la siguiente página

Tabla 4.5 – *Continua de la página previa*

Atributo	Tipo	Descripción
LOGOUT_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra la finalización de la sesión en la banca por internet.
TRX_YEAR	Categórico	Año en el que se registró la transacción.
TRX_MONTH	Categórico	Mes en el que se registró la transacción.
TRX_DAY	Categórico	Día del mes en el que se registró la transacción.
TRX_HOUR	Categórico	Hora del día en el que se registró la transacción.
FRAUD	Categórico	Identificador de la categoría de las transferencias [0 - Legítimas, 1 - Fraudulentas].

Tabla 4.5: Selección inicial de características.

4.2. Limpieza de datos

Como se mencionó anteriormente, el primer tipo de fuente de datos está compuesto por tres diferentes bases de datos relacionales correspondientes a los modelos de datos de la banca por internet, el sistema integral de autenticación y el sistema de identificación transaccional. Dentro de las ventajas que supone basarse en sistemas de información estructurada y sobre todo sistemas regulados por una autoridad es que la información suele contar con un nivel de completitud bastante alto, por lo cual las labores de limpieza de datos suelen ser reducidas. A continuación se comentan las acciones ejecutadas en cada uno de estos modelos de datos.

4.2.1. Base de datos de la banca por internet

En una primera exploración a través de consultas de lenguaje SQL se detectan transacciones con información omitida en la columna monto, sin embargo, se detecta también que en todos los casos corresponde a transacciones cuyo estatus final es “No ejecutada” o “Cancelada”.

El estatus “No ejecutada” corresponde a las transacciones que por latencia en la red, ya sea por falla en las comunicaciones entre los sistemas bancarios internos o por desconexión por parte del cliente la transacción no se culminó.

El estatus “Cancelada” corresponde a las transacciones que fueron capturadas sin embargo, en el momento de ingresar la contraseña de una sola ocasión (OTP) el usuario decidió cancelar la operación.

En ambos casos, transferencias “No ejecutadas” y “Canceladas” carecen de relevancia para el objeto de la presente investigación, por lo anterior, se filtraron únicamente las transacciones correspondientes al estatus “Transacción exitosa” obteniendo un total de 87,009 transferencias.

En todos los casos correspondientes al estatus “Transacción exitosa” la información relevante se encuentra correctamente informada, es decir, no existen datos vacíos, nulos o incorrectos.

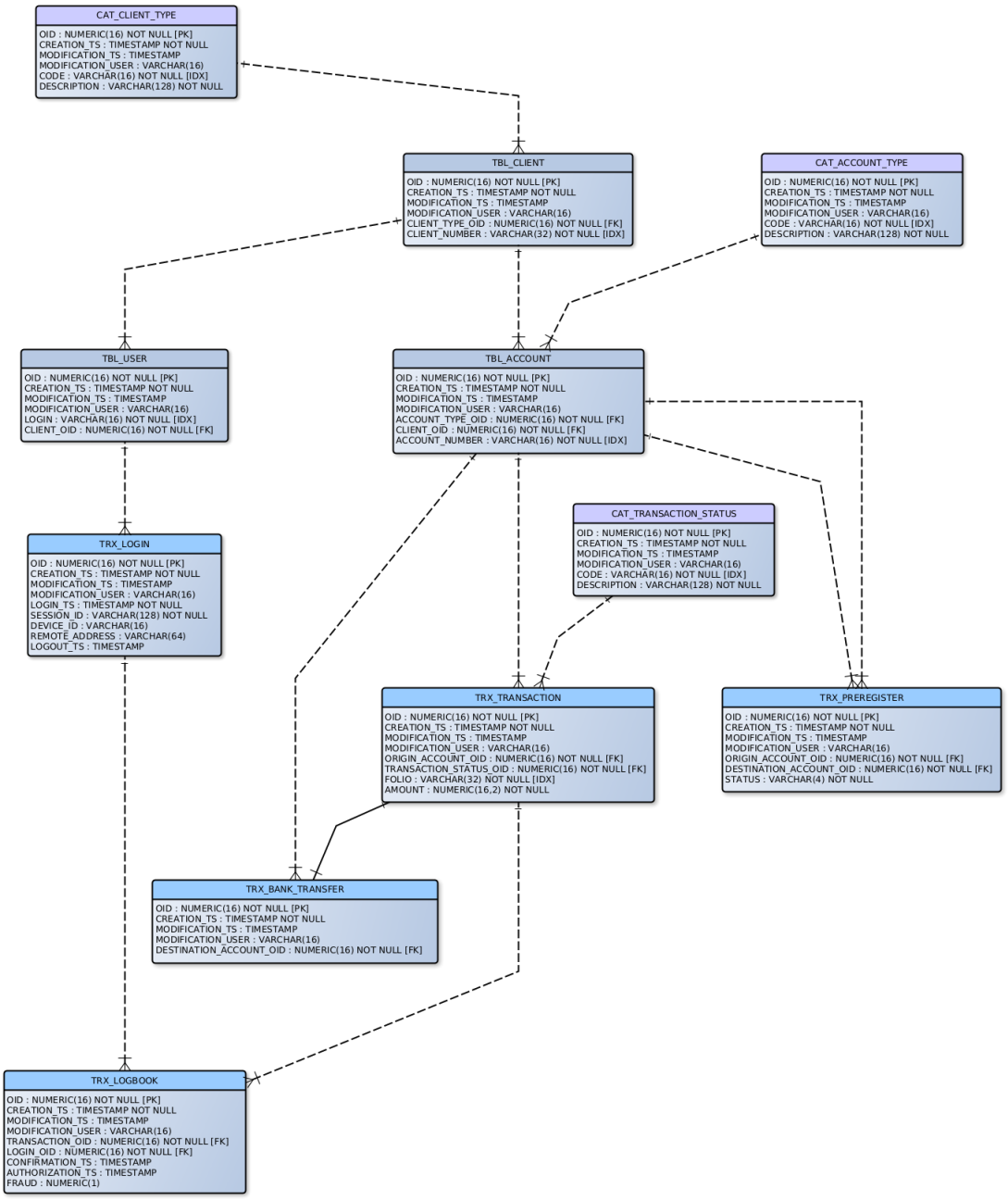

Figura 4.2: Diagrama Entidad - Relación del modelo de datos de la banca por internet extendido con la información de los dictámenes de investigación de reportes de fraude y la bitácora transaccional.

4.2.2. Base de datos del sistema integral de autenticación

Como se mencionó anteriormente, en esta base de datos se almacena la información correspondiente al inicio y fin de sesión de diferentes aplicaciones del banco, por lo anterior el filtrado de la información correspondiente a la banca por internet cobra relevancia. A continuación se describe el procedimiento ejecutado:

1. Diseño y ejecución de consulta SQL para obtener únicamente la información la correspondiente a la banca por internet.
2. Verificar la completitud de la información. En este punto es relevante especificar los datos requeridos.
   - Identificador del usuario de banca por internet (*login*)
   - Marca de tiempo del evento de inicio de sesión en la banca por internet
   - Marca de tiempo del evento de fin de sesión en la banca por internet.
3. Exportar la información en un archivo de texto simple en formato de datos separados por comas (CSV) ya que este formato es compatible con los mecanismos de importación de información de la mayoría de los gestores de bases de datos relacionales.
4. Importar la información seleccionada dentro del modelo de datos integrado para la detección de fraudes en transferencias bancarias.

4.2.3. Base de datos del sistema de identificación transaccional

Existen ciertas reglas de negocio a tomar en cuenta al momento de ejecutar el proceso de limpieza y filtrado de la operación. A continuación se describen las reglas de negocio por considerar:

- Al crear un registro de cuenta destino se le asigna el estatus “Pendiente de activación”, transcurridos treinta minutos del registro, se asigna el estatus “Activa sin uso” y al ejecutarse la primer transacción se establece el estatus “Activo”.
- Un registro de cuenta destino que no ha sido utilizado en un periodo de 6 meses es asignado con el estatus “Bloqueado por inactividad”
- Cuando la cuenta origen o destino está bajo investigación de reporte de fraude o se encuentra en la lista negra de las autoridades al registro de cuenta destino se le asigna el estatus “Bloqueado por la autoridad”.
- Cuando el usuario elimina el registro de cuenta destino se asigna el estatus “Cancelado”.

A continuación se describe el procedimiento ejecutado para la limpieza de la información:

1. Diseño y ejecución de consulta SQL para obtener los registros de cuenta destino correspondiente a los estatus: "Activo", "Bloqueado por inactividad" y "Bloqueado por la autoridad".
2. Verificar la completitud de la información. En este punto es relevante especificar los datos requeridos.
   - Número de cuenta origen
   - Número de cuenta destino
   - Marca de tiempo del evento de registro de cuenta destino
3. Exportar la información en un archivo de texto simple en formato de datos separados por comas (CSV) ya que este formato es compatible con los mecanismos de importación de información de la mayoría de los gestores de bases de datos relacionales.
4. Importar la información seleccionada dentro del modelo de datos integrado para la detección de fraudes en transferencias bancarias.

4.2.4. Informe de dictamen de investigación de reportes de fraude

El archivo con el dictamen de la investigación de reportes de fraudes se genera a partir de una plantilla en la cual se incluyen el número de cliente, el número de cuenta (origen), el identificador del usuario de banca por internet y el número de folio generado automáticamente en la banca por internet.

En el lapso de 2015 a 2019 se detecta que dicha plantilla ha sufrido dos modificaciones, es decir, existen tres versiones de la plantilla.

Esta información al provenir de una fuente de datos tipo archivo (PDF en este caso) se considera información no estructurada. Reforzando este estatus, el resultado del dictamen (transacción legítima o fraude) no es almacenado en alguna base de datos.

En este caso el procedimiento de limpieza consiste en tomar en cuenta únicamente los archivos correspondientes al dictamen de transferencia fraudulenta e ingresar el valor correspondiente dentro de una nueva columna en el modelo de datos de la banca por internet.

4.2.5. Archivo de bitácora transaccional de la banca por internet

Existen eventos generados a partir de la operación de la banca por internet cuya información no es almacenada en un modelo de datos relacional, debido a las condiciones operativas propias de un canal electrónico transaccional.

A continuación se mencionan algunos ejemplos de esta situación:

- Información detallada de intentos fallidos de inicio de sesión, únicamente se almacena la cantidad de eventos fallidos con el objetivo de bloquear la sesión al alcanzar el umbral establecido, sin embargo, las marcas de tiempo, identificador de la sesión, identificador de usuario, dirección IP origen y otros datos de relevancia no se conservan.

- Registro de navegación dentro de la banca por internet. Esto permitiría realizar un análisis de los patrones de conducta de los usuarios dentro de sus sesiones de banca por internet, con el objetivo de determinar cuando esos patrones son modificados abruptamente y establecer un mecanismo de alerta.

- Captura de la información de una transferencia. En la mayoría de los sistemas de banca por internet un paso previo a la ejecución de una transacción es la captura de los datos de la misma, en un paso previo a ingresar las credenciales del usuario (incluyendo la contraseña de una sola vez) se presenta ante el usuario la información capturada y se solicita una última verificación por parte del usuario. La marca de tiempo de este evento, suele ser un dato que no se almacena en el modelo de datos, sin embargo, al igual que el punto anterior, contiene información relacionada con los patrones de comportamiento de los usuarios.

El punto común de esta información es que a pesar de no ser almacenada en una fuente de datos estructurada, generalmente si es almacenada en los archivos de bitácora de la aplicación y del software que actúa como servidor de aplicaciones, que a pesar de ser una fuente no estructurada, suele contener la información en patrones reconocibles.

A continuación se describe el procedimiento ejecutado para la limpieza de la información:

1. Diseño y codificación de procedimiento para obtener las líneas de los archivos de bitácora correspondientes a los eventos de inicio y fin de transferencia. Este procedimiento está enlazado con otro, responsable de extraer de las líneas seleccionadas la información correspondiente a las marcas de tiempo de la captura de la transferencia.

2. Verificar la completitud de la información.

3. Exportar la información en un archivo de texto simple en formato de datos separados por comas (CSV) ya que este formato es compatible con los mecanismos de importación de información de la mayoría de los gestores de bases de datos relacionales.

4. Importar la información seleccionada dentro del modelo de datos integrado para la detección de fraudes en transferencias bancarias.

4.3. Análisis exploratorio de datos (EDA)

Una vez integradas las fuentes de información en un sólo modelo de datos y ejecutadas las tareas de limpieza y preparación de los datos es necesario explorar la información en busca de detectar patrones de comportamiento, tendencias, distribuciones, valores atípicos y en general hallazgos sobre la información que puedan guiar o conducir la experimentación con técnicas de aprendizaje de máquinas.

En Tukey et al. [72], define al análisis exploratorio de datos como un trabajo de detective basado en números y gráficos. Esencialmente es el primer contacto con los datos, en el cual conoceremos y aprenderemos que historias tienen por describirnos esos datos.

4.3.1. Naturaleza desequilibrada del problema

En primera instancia se busca verificar el grado de desequilibrio que presentan los datos, para lo cual la Figura 4.3a muestra una gráfica de barras con la relación entre la cantidad de transferencias respecto a su clase (Legítimas - Fraudes) en todo el conjunto de datos. En dicha gráfica se aprecia que el nivel de desequilibrio se encuentra en el orden de 3:1,000 (0.0031), es decir, de cada mil transferencias tres son fraudulentas.

Desde el punto de vista estadístico se confirma la naturaleza altamente desequilibrada del problema, mientras que desde el punto de vista de negocio se aprecia una alta tasa de comisión de fraudes, en comparación con la información publicada en [7] en donde se reporta un orden de 1:100,000 para la comunidad europea.

La Figura 4.3b vuelve a mostrar la relación entre la cantidad de transferencias respecto a su clase con el agregado que en esta ocasión se segmenta por año.

Con esta gráfica se busca verificar el crecimiento en la transaccionalidad así como el comportamiento en la comisión de transferencias fraudulentas a través del tiempo.

El crecimiento de la transaccionalidad se da en la siguiente relación: 18.98 % de 2015 a 2016, 21.99 % de 2016 a 2017, 26.92 % de 2017 a 2018 y 23.17 % de 2018 a 2019. Estos datos verifican una tendencia creciente en la adopción de canales digitales como medios transaccionales.

Respecto al comportamiento en la comisión de fraudes se observa un orden de 2.5:1,000 para el 2015, 2.6:1,000 para 2016, 2.6:1,000 para 2017, 3.2:1,000 para 2018 y 3.7:1,000 para 2019, es decir, se aprecia un ascenso en la comisión de fraudes a partir de 2018.

4.3.2. Exploración de la característica continua monto

El monto de las transferencias es una de las características relevantes por conocer en detalle y contrastar las tendencias entre las clases legítima y fraudulenta.

La Figura 4.4 muestra la distribución del monto por clase a través de un diagrama de cajas. Se aprecia que en ambas clases se distribuyen en rangos similares de montos,

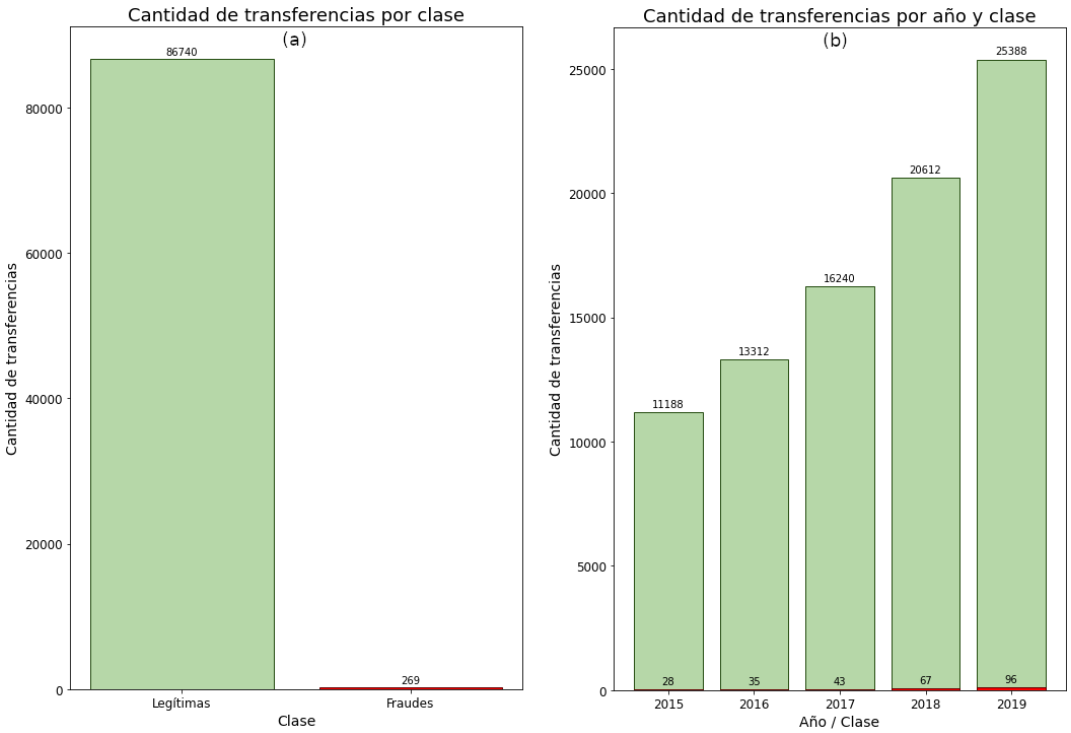

Figura 4.3: (a) Relación entre la cantidad de transferencias y la clase [Legítimas — Fraudes]. (b) Relación entre la cantidad de transferencias y la clase por año.

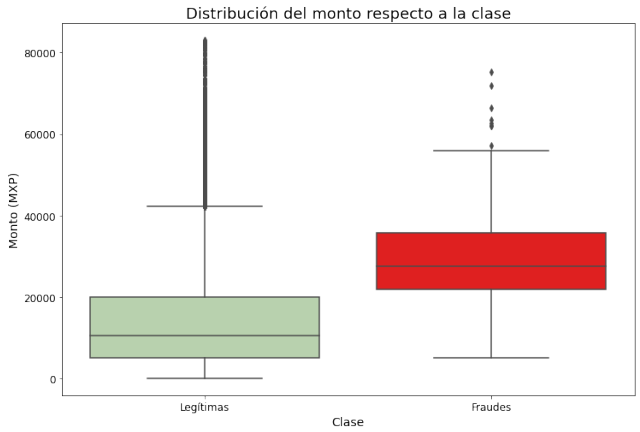

Figura 4.4: Diagrama de caja con la distribución del monto respecto a su clase

estando contenidas las transferencias fraudulentas dentro del rango de las legítimas, es decir, el rango de estas últimas tiene mayor amplitud.

Debido a la naturaleza desequilibrada del problema, se aprecia una cantidad considerable de valores atípicos en la clase de transferencias legítimas. Mientras que en la clase de fraudes los datos se distribuyen de forma más compacta y existen pocos valores atípicos. Las medianas de ambas clases se encuentran considerablemente distanciadas (legítimas 10,492.49, fraudulentas 27,465.10)

La Figura 4.5 muestra la dispersión de los montos de las transferencias a través del tiempo. Se puede apreciar que las transferencias fraudulentas ocupan el mismo espacio que las transferencias legítimas, particularmente a partir de 2018 lo cual dificulta su detección. El año de 2017 se podría considerar un año atípico en ese sentido al contar con un alto índice de transacciones fraudulentas por encima del monto máximo de las operaciones legítimas.

Las gráficas de dispersión mostradas en las Figuras 4.6 se enfocan en rangos temporales diferentes.

En el caso de la Figura 4.6a la gráfica muestra la transaccionalidad por mes del año, esto con el objeto de identificar patrones estacionales en la comisión de trasferencias fraudulentas. En el mes de marzo se nota un incremento general de la transaccionalidad, tanto en transacciones legítimas como fraudulentas, en el mes de abril se nota un decremento en las transferencias legítimas y un aumento en las transferencias fraudulentas. En el caso de los meses de noviembre y diciembre, también se aprecia un patrón de incremento en las transacciones fraudulentas.

Por otro lado, la Figura 4.6b muestra la transaccionalidad dentro del horario operativo de la banca por internet. En este caso se aprecia una tendencia de comisión de transferencias fraudulentas entre las 6 y 8 am. Durante el resto de las horas se aprecia un patrón estable a excepción de las 18 horas en donde se incrementa la comisión de fraudes y las 22 horas en donde disminuye notablemente.

Con el objetivo de explorar la temporalidad y los cambios de comportamiento en la

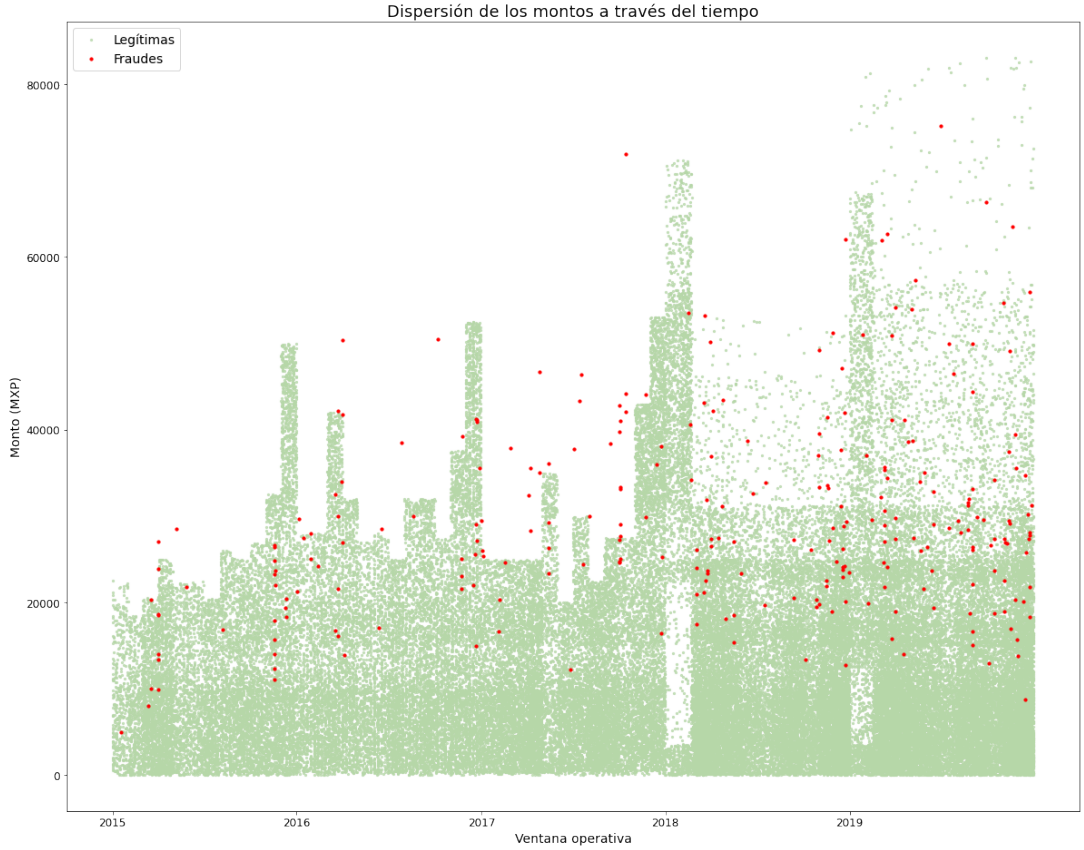

Figura 4.5: Gráfica de dispersión de los montos respecto a los años de operación

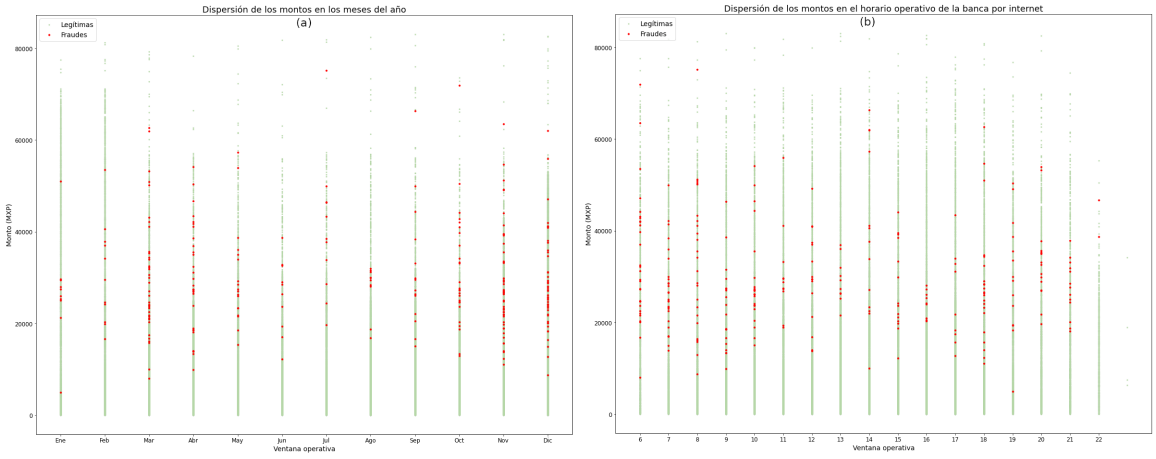

Figura 4.6: Gráfica de dispersión de (a) los montos respecto al mes del año, (b) los montos respecto a las horas operativas de la banca por internet

transaccionalidad se propone extender las características continuas generando bloques de información correspondientes a la transaccionalidad global, la transaccionalidad por año y la transaccionalidad por mes.

Las características continuas de transaccionalidad extendidas se describen en la Tabla 4.6.

Atributo	Tipo	Descripción
HIST_TRX_COUNT	Continuo	Contador histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_AVG	Continuo	Monto promedio histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_SUM	Continuo	Suma de los montos del histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_MIN	Continuo	Monto mínimo histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_MAX	Continuo	Monto máximo histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar histórica de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_VARIANCE	Continuo	Monto de la varianza histórica de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_COUNT	Continuo	Contador anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_AVG	Continuo	Monto promedio anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_SUM	Continuo	Suma anual de los montos de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_MIN	Continuo	Monto mínimo anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_MAX	Continuo	Monto máximo anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_VARIANCE	Continuo	Monto de la varianza anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_COUNT	Continuo	Contador anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_AVG	Continuo	Monto promedio anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_SUM	Continuo	Suma anual de los montos de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_MIN	Continuo	Monto mínimo anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_MAX	Continuo	Monto máximo anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_VARIANCE	Continuo	Monto de la varianza anual de las transferencias realizadas por el cliente con la cuenta origen.

Tabla 4.6: Características continuas de transaccionalidad extendidas.

4.3.3. Exploración de características continuas de tiempo

Además del monto de las transferencias se cuenta con las marcas de tiempo de los eventos alta de cuenta origen, registro de cuenta destino, inicio de sesión en la banca por internet, captura de la transferencia, autorización de la transferencia y fin de sesión en la banca por internet como características continuas.

En las gráficas de la Figura 4.7 se muestran las distribuciones de las seis características continuas de tiempo respecto a la clase. Salvo sutiles detalles las distribuciones son altamente similares.

Derivado de esta situación, se propone extender las características continuas de tiempo a través de generar los diferenciales de las marcas de tiempo, utilizando como marca base de comparación la marca de tiempo de la autorización de la transferencia, que es el momento en el cual se ejecuta la transferencia.

Las características continuas de tiempo extendidas se describen en la Tabla 4.7.

Atributo	Tipo	Descripción
TRX_ACC_OPENING_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de captura y autorización de la transacción.
TRX_PREREG_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de registro de cuenta destino y autorización de la transacción.
TRX_LOGIN_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de ingreso a la sesión de la banca por internet y autorización de la transacción.
TRX_AUTHORIZATION_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de apertura de la cuenta y autorización de la transacción.
TRX_LOGOUT_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de autorización de la transacción y la fecha de finalización de la sesión de banca por internet.

Tabla 4.7: Características continuas de tiempo extendidas.

En la Figura 4.8 se muestran las distribuciones de las características continuas de tiempo extendidas. El diferencial entre la apertura de la cuenta origen y la ejecución de la transferencia muestra una distribución muy similar entre las clases de transferencias legítimas y fraudulentas. Mientras que se aprecia que los diferenciales entre el registro de la cuenta destino y el inicio de sesión respecto a la ejecución de la transferencia tienen marcadas diferencias en la distribución entre las clases de transferencias legítimas y fraudulentas, siendo estas últimas notablemente más compactas y ubicándose al nivel del primer cuartil de las transferencias legítimas. Los diferenciales entre la captura de la transferencia y el cierre de sesión de la banca por internet respecto a la ejecución de la transferencia para la clase de transferencias fraudulentas muestran mayor amplitud en su distribución pero ubicándose entre los dos primeros cuartiles respecto a las transferencias legítimas. En conjunto se confirma su integración como características predictoras para los modelos de aprendizaje de máquinas.

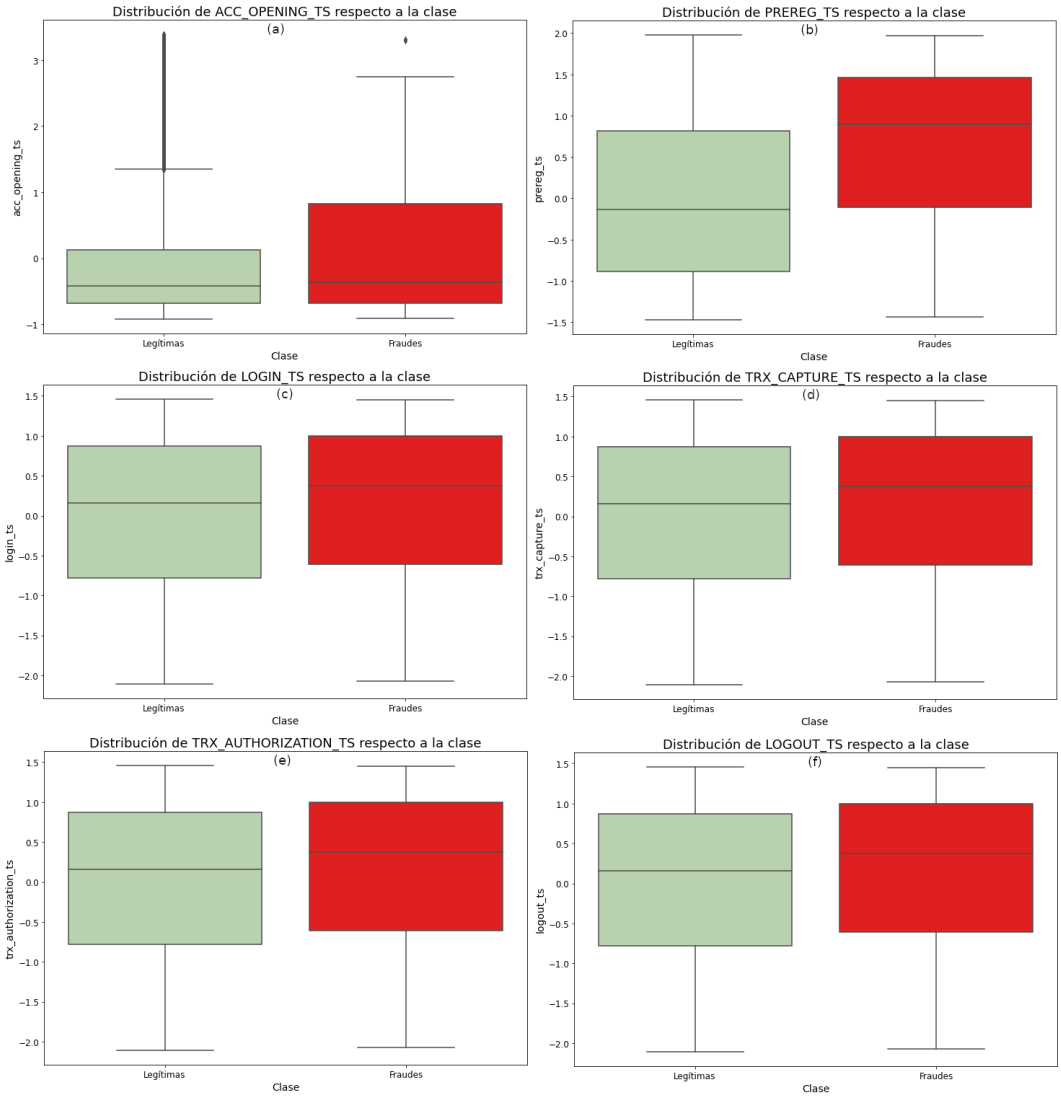

Figura 4.7: Distribución de la marca de tiempo de (a) apertura de la cuenta origen, (b) registro de la cuenta destino, (c) inicio de sesión en la banca por internet, (d) captura de la transferencia, (e) autorización de la transferencia y (f) fin de sesión de la banca por internet respecto a la clase.

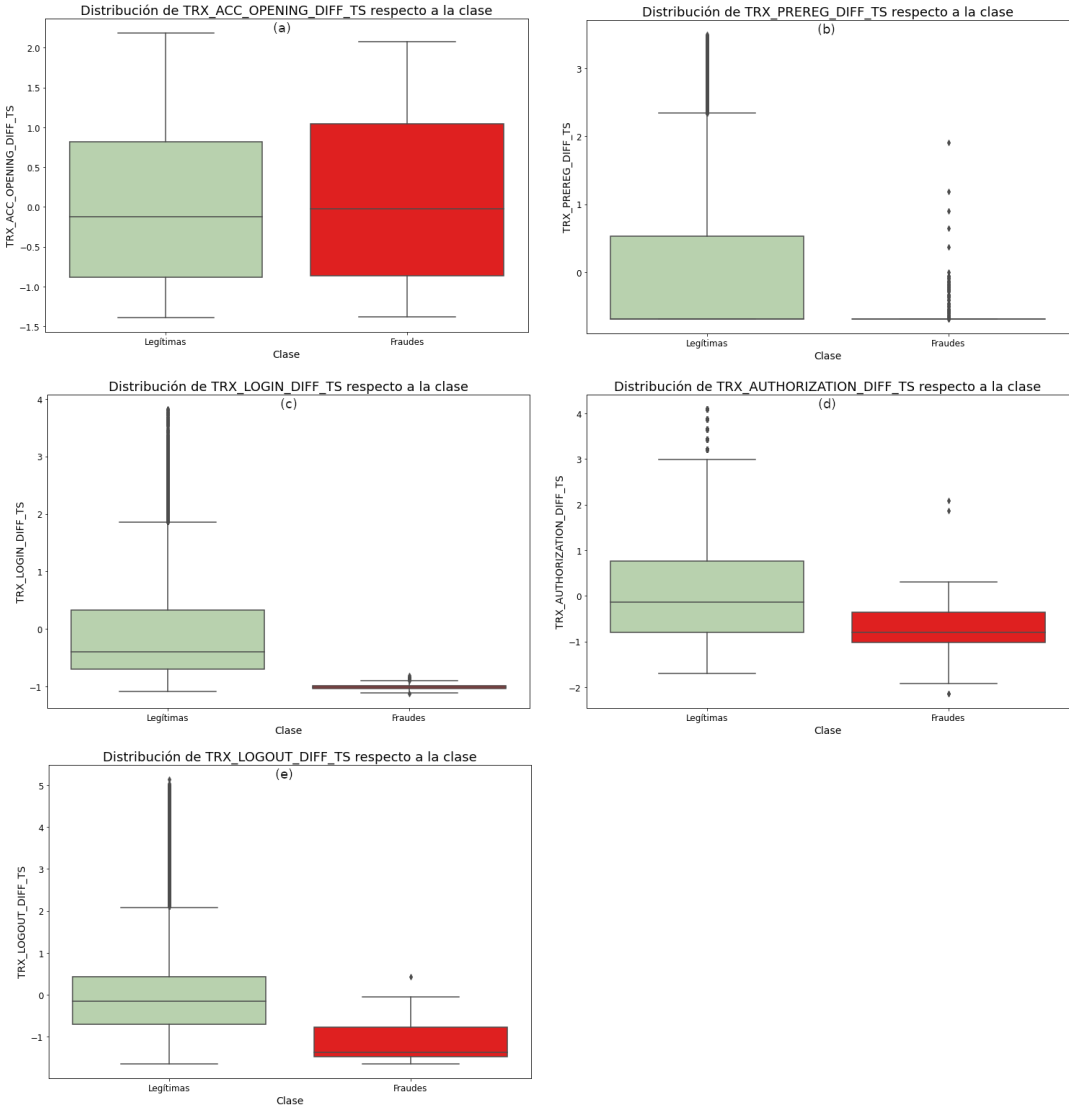

Figura 4.8: Distribución del diferencial de tiempo entre la autorización de la transferencia y (a) apertura de la cuenta origen, (b) registro de la cuenta destino, (c) inicio de sesión en la banca por internet, (d) captura de la transferencia, (e) autorización de la transferencia y (f) fin de sesión de la banca por internet respecto a la clase respecto a la clase

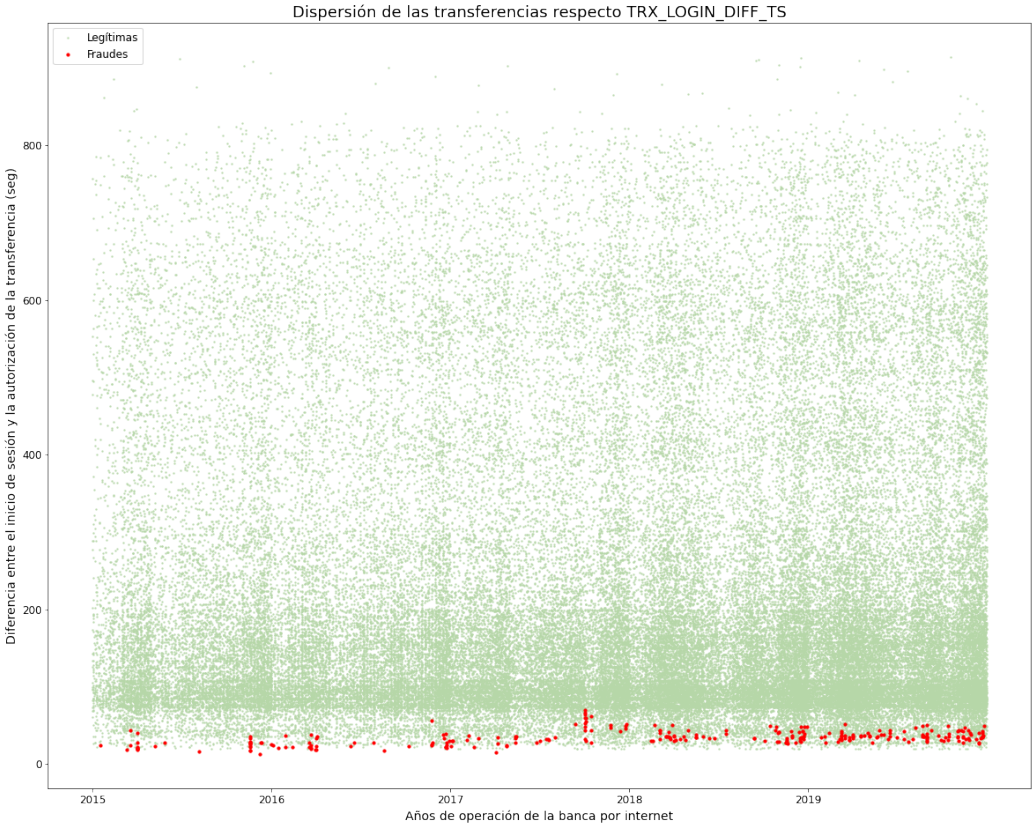

Figura 4.9: Dispersión de las transferencias, relación del diferencial entre el inicio de sesión de la banca por internet y la autorización de la transferencia.

Es importante explorar la relación entre las marcas de tiempo del evento de inicio de sesión en la banca por internet y la ejecución de la transferencia ya que un comportamiento esperado por parte de los defraudadores es ejecutar la transferencia dentro del mínimo lapso entre ambos eventos. En una transacción legítima se espera incurrir en tiempos derivados de la navegación dentro de la aplicación, como podría ser la carga inicial de la información del cliente y sus cuentas, la navegación a través de las opciones de menú y la carga de las pantallas de captura y confirmación de la transferencia, por lo cual una transferencia no podría se ejecutada de forma inmediata.

La Figura 4.9 muestra la dispersión de las transferencias en la relación diferencial entre las marcas de tiempo de inicio de sesión de la banca por internet y la autorización de la transferencia. En esta gráfica se aprecia que mientras las transferencias legítimas se dispersan en un espacio temporal entre 20 y 914 segundos con una mediana de 146 segundos y un tercer cuartil en 277 segundos, las transferencias fraudulentas se dispersan en un espacio temporal entre 13 y 70 segundos con una mediana en 34 segundos y un tercer cuartil en 39 segundos.

Figura 4.10: Dispersión de las transferencias, relación del diferencial entre la captura y la autorización de la transferencia.

La relación del diferencial entre las marcas de tiempo de la captura y la ejecución de la transferencia mantiene una cota superior de 30 segundos, es decir, una vez capturados los datos de la transferencia se presenta la pantalla de confirmación en la que aparecen los datos capturados sin opción a modificarlos, únicamente como opción de verificación y se solicita la captura de la contraseña de una sola vez (OTP); dicha pantalla tiene una vigencia de 30 segundos.

La Figura 4.10 muestra la dispersión de las transferencias en la relación diferencial entre las marcas de tiempo de la captura y la ejecución de la transferencia. Se aprecia que las transacciones legítimas se dispersan entre 4 y 30 segundos con una mediana en 11 segundos y un tercer cuartil en 15 segundos, mientras que las transacciones fraudulentas se dispersan en un lapso de tiempo entre 2 y 21 segundos con una mediana en 8 segundos y un tercer cuartil en 10 segundos.

La Figura 4.11 muestra la dispersión de las transferencias en la relación diferencial entre las marcas de tiempo de la ejecución de la transferencia y el cierre de sesión de la banca por internet. Se aprecia que las transacciones legítimas se dispersan entre 1

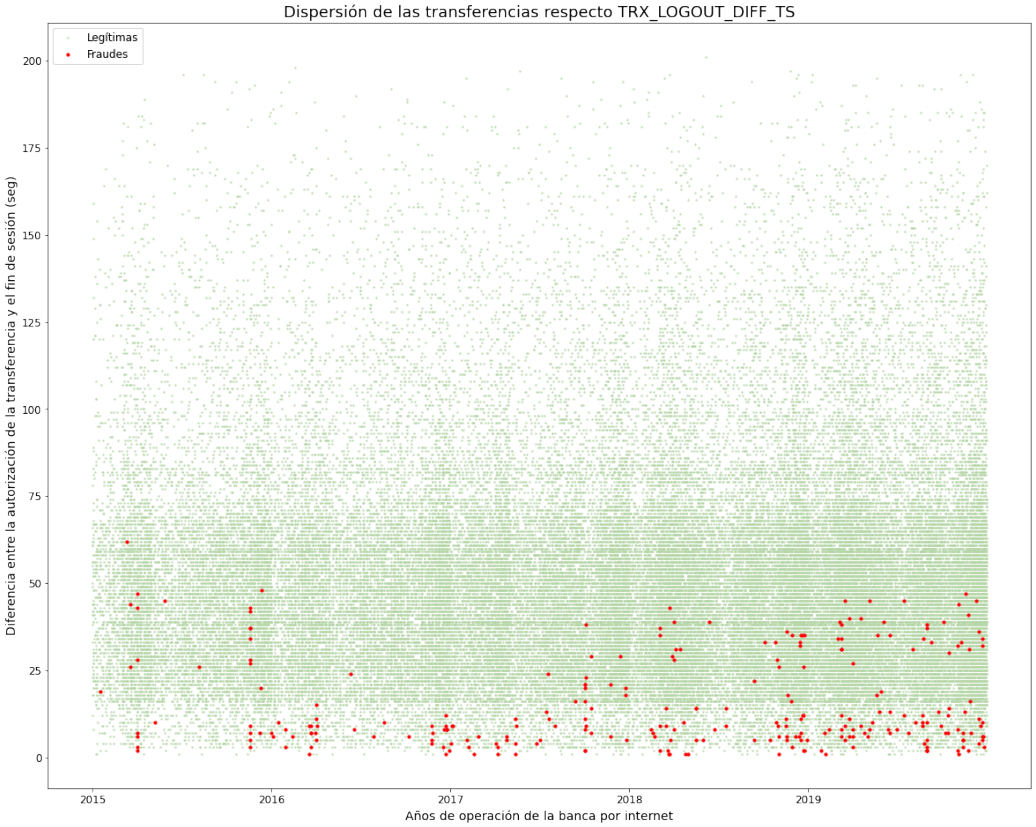

Figura 4.11: Dispersión de las transferencias, relación del diferencial entre la autorización de la transferencia y el fin de la sesión.

y 201 segundos con una mediana en 45 segundos y un tercer cuartil en 62 segundos, mientras que las transacciones fraudulentas se dispersan en un lapso de tiempo entre 1 y 62 segundos con una mediana en 9 segundos y un tercer cuartil en 27 segundos.

4.3.4. Correlación entre características continuas

El análisis de correlación tiene como objetivo determinar la magnitud y tipo de correlación existente entre las características integradas. En el presente trabajo de investigación se utilizó el coeficiente de correlación de Pearson [73] como métrica de la relación estadística de las características continuas. En la Figura 4.12 se muestra el mapa de calor sobre la matriz de correlación para las características continuas y la clase.

Entre los puntos por resaltar se encuentran las correlaciones con magnitud 1.00 entre las marcas de tiempo de los eventos de inicio de sesión, captura y ejecución de la transacción y el cierre de sesión. Esto indica que dichas características se encuentran

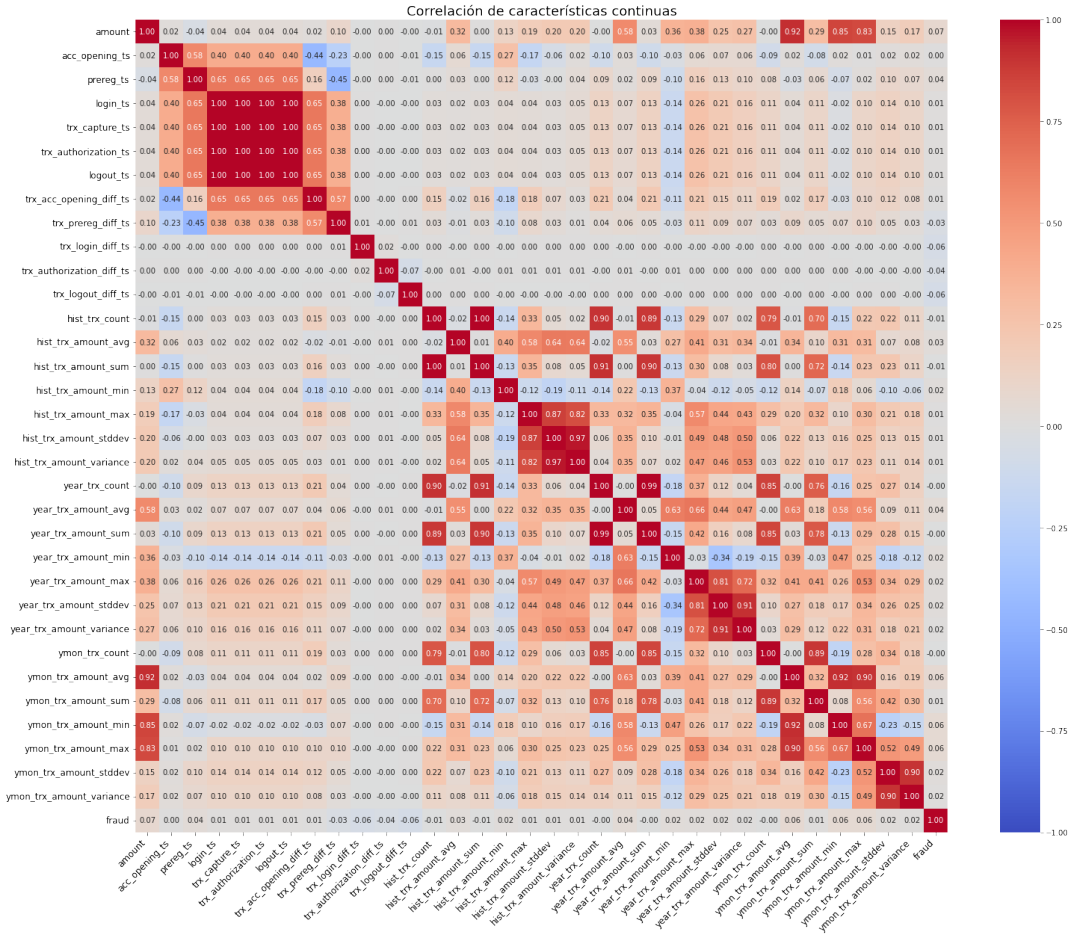

Figura 4.12: Mapa de calor de la correlación (Pearson) de las características continuas.

completamente correlacionadas de forma positiva, por lo tanto, se puede seleccionar sólo una de ellas, implicando esto el descarte de las otras tres; en este caso se selecciona el uso de la marca de tiempo del inicio de sesión.

Respecto a las características continuas transaccionales extendidas resalta la alta magnitud en la correlación entre la desviación estándar y la varianza del monto para los tres bloques de información, transaccionalidad global, anual y mensual. Por lo tanto, también se puede seleccionar una característica de cada bloque implicando el descarte de la otra. En este caso se seleccionan las desviaciones estándar descartando las varianzas.

Es relevante analizar en detalle la correlación entre las características continuas respecto a la clase, dado el objetivo de predecir este valor. La Figura 4.13 muestra que la mayor correlación positiva se da respecto al monto (0.07). Mientras que la mayor correlación negativa se da respecto al diferencial de la marca de tiempo del evento de fin de sesión(-0.06). Por otro lado, se observa una magnitud de correlación muy baja entre la marca de tiempo de la apertura de la cuenta (0.00), la suma de montos anuales

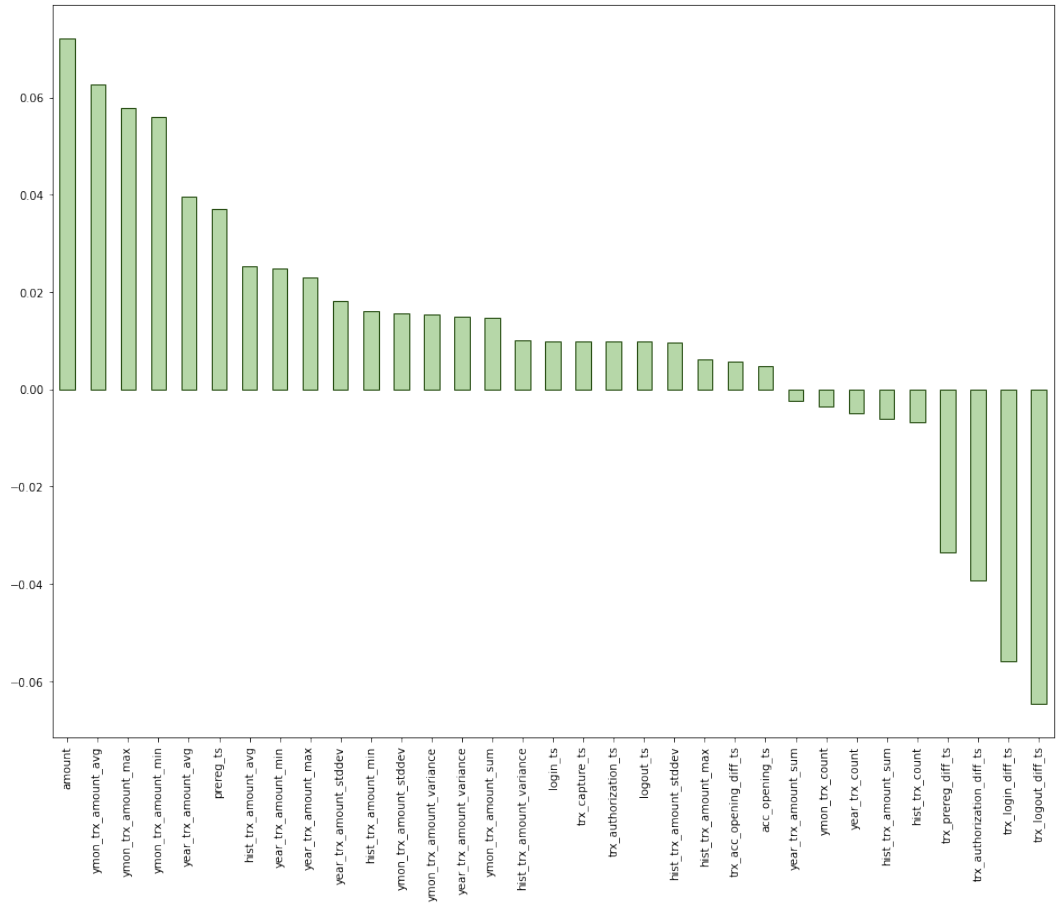

Figura 4.13: Correlación de las características continuas respecto a la clase.

(0.00) y el conteo de montos mensuales respecto a la clase (0.00).

4.4. Selección final de características

Tomando como base la selección preliminar de características descrita en la sección Modelo de datos integrado y el análisis de la sección Análisis exploratorio de datos (EDA) se realiza la selección final de características.

La Tabla 4.8 muestra la selección final de características especificando el tipo de dato y una breve descripción del mismo.

Atributo	Tipo	Descripción
ORIGIN_ACCOUNT	Categórico	Es el identificador de la cuenta bancaria de donde salen los fondos de la transferencia.
DESTINATION_ACCOUNT	Categórico	Es el identificador de la cuenta bancaria hacia donde se transmiten los fondos de la transferencia.
AMOUNT	Continuo	Monto de la transferencia.
PREREG_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra la cuenta destino como opción de transferencia en la aplicación de la banca por internet.
LOGIN_TS	Continuo	Fecha en formato UNIX (POSIX) en la que se registra el inicio de la sesión en la banca por internet.
TRX_YEAR	Categórico	Año en el que se registró la transferencia.
TRX_MONTH	Categórico	Mes en el que se registró la transferencia.
TRX_DAY	Categórico	Día del mes en el que se registró la transferencia.
TRX_HOUR	Categórico	Hora del día en el que se registró la transferencia.
TRX_ACC_OPENING_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de captura y autorización de la transferencia.
TRX_PREREG_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de registro de cuenta destino y autorización de la transferencia.
TRX_LOGIN_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de ingreso a la sesión de la banca por internet y autorización de la transferencia.
TRX_AUTHORIZATION_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de apertura de la cuenta y autorización de la transferencia.
TRX_LOGOUT_DIFF_TS	Continuo	Diferencial de tiempo en segundos entre las fechas de autorización de la transferencia y la fecha de finalización de la sesión de banca por internet.
HIST_TRX_COUNT	Continuo	Contador histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_AVG	Continuo	Monto promedio histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_SUM	Continuo	Suma de los montos del histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_MIN	Continuo	Monto mínimo histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_MAX	Continuo	Monto máximo histórico de las transferencias realizadas por el cliente con la cuenta origen.
HIST_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar histórica de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_COUNT	Continuo	Contador anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_AVG	Continuo	Monto promedio anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_SUM	Continuo	Suma anual de los montos de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_MIN	Continuo	Monto mínimo anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_MAX	Continuo	Monto máximo anual de las transferencias realizadas por el cliente con la cuenta origen.
YEAR_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_COUNT	Continuo	Contador anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_AVG	Continuo	Monto promedio anual de las transferencias realizadas por el cliente con la cuenta origen.

Continua en la siguiente página

Tabla 4.8 – *Continua de la página previa*

Atributo	Tipo	Descripción
YMON_TRX_AMOUNT_SUM	Continuo	Suma anual de los montos de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_MIN	Continuo	Monto mínimo anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_MAX	Continuo	Monto máximo anual de las transferencias realizadas por el cliente con la cuenta origen.
YMON_TRX_AMOUNT_STDDEV	Continuo	Monto de la desviación estándar anual de las transferencias realizadas por el cliente con la cuenta origen.
FRAUD	Categórico	Identificador de la categoría de las transferencias [0 - Legítimas, 1 - Fraudulentas].

Tabla 4.8: Selección final de características.

4.5. Estrategia de partición de los datos

La detección de transferencias fraudulentas en banca por internet es un problema que evoluciona a través del tiempo en diferentes aspectos como pueden ser:

- **Nivel de transaccionalidad**. Cantidad y frecuencia de las transferencias realizadas por usuario.
- **Montos de la transferencias**. El monto de las transferencias realizadas por usuario también se modifican con el paso del tiempo.
- **Origen y Destino de las transferencias**. A través del tiempo se registran nuevas cuentas destino y algunas otras dejan de ser utilizadas.

Los patrones de comportamiento se modifican, se crean nuevos y algunos otros desaparecen; todo lo anterior aplica tanto para las transferencias legítimas como para las fraudulentas. Las transferencias fraudulentas suelen tener sutiles diferencias en cuanto a su comportamiento respecto de las legítimas, sin embargo, tienden a imitar los patrones legítimos con el objeto de pasar desapercibidas el mayor tiempo posible.

Por lo anterior, en la presente experimentación se propone una estrategia de separación o partición de los datos de forma cronológica, optando por utilizar tres particiones.

1. **Entrenamiento**. Este fragmento de los datos es utilizado para ejecutar la fase de entrenamiento en donde se calculan los pesos y sesgos de cada uno de los modelos y se compone de la información correspondiente al periodo de 2015-01-01 hasta 2018-12-31.
2. **Validación**. Este fragmento de la información es utilizado para validar los diferentes conjuntos de hiperparámetros preseleccionados para cada uno de los modelos de aprendizaje de máquinas con el objeto de seleccionar el modelo que mejor

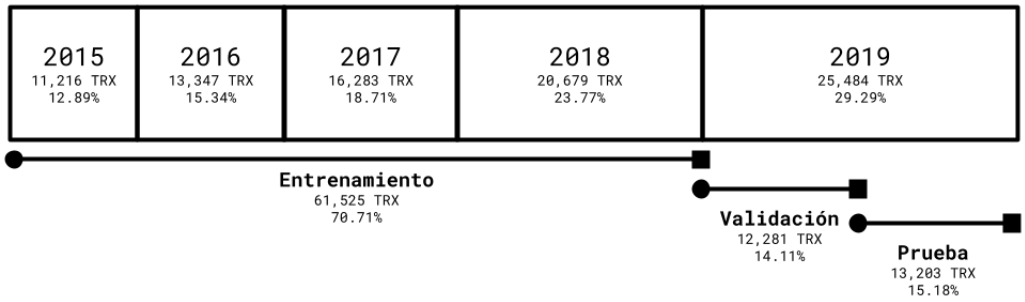

Figura 4.14: Estrategia de partición del conjunto de datos de la banca por internet

rendimiento presente en las métricas seleccionadas. Se compone de la información correspondiente al periodo de 2019-01-01 hasta 2019-06-30.

3. **Prueba**. Este fragmento de la información es utilizado para generar las métricas de evaluación finales de los experimentos. Se compone de la información correspondiente al periodo de 2019-07-01 hasta 2019-12-31.

Los objetivos de esta estrategia de partición de los datos son las siguientes:

1. **Modelar el comportamiento evolutivo**. La información en orden cronológico promueve el aprendizaje de la evolución de los patrones de comportamiento de las transferencias fraudulentas respecto a la legítimas.
2. **Evitar sesgos por patrones obsoletos**. Por la misma naturaleza evolutiva del problema existen patrones que se convierten en obsoletos a través del tiempo.

En la Figura 4.14 se ilustra la estrategia de partición de los datos.

Capítulo 5

Resultados experimentales

En este capítulo se describen los experimentos realizados sobre el conjunto de datos de transferencias de banca por internet con los modelos de aprendizaje de máquinas seleccionados. Se presentan y discuten los resultados obtenidos.

La codificación de los experimentos se desarrolló utilizando el lenguaje de programación de propósito general Python [74], debido a que cuenta con diversas bibliotecas para el tratamiento de datos, cálculos numéricos e implementación de modelos de aprendizaje de máquinas. Principalmente, se utilizaron las bibliotecas NumPy [75], Pandas [76], SciKit-Learn [77], XGBoost [49], LightGBM [78].

A manera de notas preliminares:

- Los valores reportados para las métricas propuestas corresponden a la fase de pruebas.
- En las tablas de resultados las columnas 2 a la 5 muestran los valores de la matriz de confusión.
- En las tablas de resultados las columnas 6 a la 11 muestran los valores obtenidos en las métricas seleccionadas para el problema de clasificación de transferencias fraudulentas.
- Para fines de una mejor compresión de los resultados, la Tabla 5.1 presenta el diccionario de algoritmos de aprendizaje de máquinas por utilizar.

La selección de estos algoritmos está basada en los resultados especificados en la Revisión del estado de arte así como en la descripción de los mismos establecida en Aprendizaje de máquinas.

5.1. Experimento 1

Se toma el conjunto de características descrito en la Tabla 4.5 con la particularidad de que se omite el uso de las características categóricas correspondientes a la cuenta origen y cuenta destino.

ID	Modelo
LR	Logistic Regression
KN	KNearest Neighbors
DT	Decision Tree
RF	Random Forest
SVM	Support Vector Machines
XGB	XGBoost
LGB	LightGBM
IF	Isolation Forest

Tabla 5.1: Diccionario de algoritmos de aprendizaje de máquinas utilizados en los experimentos.

Modelo	TN	FP	FN	TP	Precisión	Sensibilidad	ROC AUC	F_1	F_β	MCC
LR	10561	2585	3	54	0.020462	0.947368	0.875365	0.040059	0.094175	0.123079
KN	13128	18	56	1	0.052632	0.017544	0.508087	0.026316	0.020243	0.027975
DT	11774	1372	24	33	0.023488	0.578947	0.737291	0.045144	0.101041	0.100903
RF	11845	1301	23	34	0.025468	0.596491	0.748763	0.048851	0.108765	0.108198
SVM	12648	498	35	22	0.042308	0.385965	0.674041	0.076256	0.147059	0.117329
XGB	13138	8	55	2	0.200000	0.035088	0.517240	0.059701	0.042017	0.082171
LGB	11606	1540	17	40	0.025316	0.701754	0.792304	0.048870	0.110619	0.118090
IF	13099	47	56	1	0.020833	0.017544	0.506984	0.019048	0.018116	0.015217

Tabla 5.2: Resultados experimento 1

El motivo principal para omitir dichas características es el hecho de que al ser categóricas requieren un tratamiento específico de binarización. A saber, una técnica como *One Hot Encoding* [79] en este caso resultaría en un conjunto adicional de alrededor de 17,000 características. El dato anterior es el número aproximado de valores diferentes contenidos en conjunto entre ambas características.

La situación descrita en el párrafo anterior tiene los siguientes inconvenientes:

- Aumento proporcional de los cálculos requeridos, incidiendo en el uso de los recursos computacionales (tiempo de procesamiento y memoria).
- El espacio de características que generarían podría llevar al conjunto de datos a caer en la maldición de la dimensión [80].

Respecto a las características categóricas de tiempo (año, mes, día y hora de la transferencias) son tratadas bajo la técnica de binarización *One Hot Encoding*.

5.1.1. Resultados

La Tabla 5.2 muestra los resultados del *Experimento 1*.

5.1.2. Discusión

Como se puede observar en la Tabla 5.2 el espacio de características propuesto para este experimento carece en su conjunto de elementos que permitan alcanzar un nivel de satisfactorio de separación de los datos en las clases de transferencias legítimas y fraudulentas.

El modelo que obtiene mejor resultado para la precisión es XGBoost, sin embargo, es un valor muy bajo (0.200000) este efecto es debido a que los valores para los verdaderos positivos (TP) y falsos positivos (PF) son en ambos casos bajos (2 y 8 respectivamente). Desde el punto de vista de negocio se trata de una detección muy baja de fraudes.

En cuanto a la sensibilidad el mejor resultado lo obtiene la regresión logística (0.947368) ya que detecta correctamente (TP) 54 de las 57 transferencias fraudulentas. Esto podría interpretarse como un buen resultado, sin embargo, recordando que la sensibilidad y la precisión son complementarias también es bueno analizarlas en conjunto. En este caso la regresión logística obtiene el valor más bajo para precisión (0.020462), ya que su valor para los falsos positivos es excesivamente alto. Desde el punto de vista de negocio esta sería una solución inviable ya que resultaría en un exceso de transferencias legítimas bloqueadas.

Por otro lado ROC AUC indica que la regresión logística (0.875365) es el modelo que mejor capacidad de clasificación tiene, seguido de LightGBM (0.792304). En ambos casos tiene la problemática del excesivo valor para los falsos positivos.

Respecto a F_1 el mejor rendimiento lo obtiene SVM (0.076256) seguido de XGBoost, aunque ambos siguen siendo muy bajos. Se trata de los rendimientos más equilibrados entre la precisión y la sensibilidad.

F_β otorgando mayor peso a la sensibilidad respecto a la precisión vuelve a ser SVM el modelo con mejor rendimiento (0.147059) seguido en esta ocasión por LightGBM (0.110619).

El coeficiente de correlación de Mathews pondera todos los valores de la matriz de confusión y se especializa en problemas de naturaleza desequilibrada. La regresión logística es el modelo con mejor rendimiento (0.123079) seguido de LightGBM (0.118090).

En términos generales la regresión logística alcanza una tasa alta de verdaderos positivos (fraudes correctamente clasificados), sin embargo, lo logra a costa de una excesiva tasa de falsos positivos (transferencias legítimas bloqueadas), lo cual desde el punto de vista de negocio entorpece la operación de la banca por internet. DecisionTree, RandomForest y LightGBM alcanzan tasas de verdaderos positivos en un rango similiar entre ellos pero aún muy por debajo de lo deseado. El caso que podría encontrarse más equilibrado es el de SVM, sin embargo, su tasa de verdaderos positivos es muy baja.

Modelo	TN	FP	FN	TP	Precisión	Sensibilidad	ROC AUC	F_1	F_β	MCC
LR	10834	2312	5	52	0.021997	0.912281	0.868205	0.042957	0.100309	0.125932
KN	13142	4	57	0	0.000000	0.000000	0.499848	0.000000	0.000000	-0.001146
DT	12131	1015	37	20	0.019324	0.350877	0.636834	0.036630	0.079177	0.066754
RF	11966	1180	30	27	0.022370	0.473684	0.691962	0.042722	0.094077	0.087339
SVM	12730	416	38	19	0.043678	0.333333	0.650844	0.077236	0.143288	0.110812
XGB	13128	18	56	1	0.052632	0.017544	0.508087	0.026316	0.020243	0.027975
LGB	12235	911	30	27	0.028785	0.473684	0.702193	0.054271	0.115780	0.103204
IF	12976	170	56	2	0.005848	0.017544	0.502306	0.008772	0.012531	0.002674

Tabla 5.3: Resultados experimento 2

5.2. Experimento 2

Partiendo del conjunto de características especificado para el para el Experimento 1 se extiende el espacio de características definido complementándolo con el conjunto de características continuas transaccionales extendidas especificado en la Tabla 4.6. El objetivo es aprender patrones de comportamiento transaccional dento de las ventanas de tiempo definidas (mensual, anual y global).

5.2.1. Resultados

La Tabla 5.3 muestra los resultados del *Experimento 2*.

5.2.2. Discusión

En este caso es notable que el espacio de características continuas transaccionales extendidas no sólo no mejora la capacidad de clasificación de los diferentes modelos seleccionados, sino que en todos los casos reduce el rendimiento de los mismos bajo las métricas propuestas.

En este experimento nuevamente XGBoost es el modelo que obtiene mejor resultado para la precisión (0.052632), sin embargo, este valor además de ser muy bajo si lo analizamos respecto al número de verdaderos positivos (1) es completamente descartable. Desde el punto de vista de negocio se trata de una detección muy baja de fraudes.

Para la métrica sensibilidad el mejor resultado nuevamente lo obtiene la regresión logística (0.912281) detectando correctamente (TP) 52 de las 57 transferencias fraudulentas. Al igual que en el Experimento 1, al analizar en conjunto las métricas sensibilidad y precisión (0.021997) se observa que el valor obtenido para los falsos positivos es excesivamente alto. Desde el punto de vista de negocio esta sería una solución inviable ya que resultaría en un exceso de transferencias legítimas bloqueadas.

En cuanto a la capacidad de clasificación ROC AUC indica que la regresión logística (0.868205) es el modelo que mejor desempeño muestra, seguido de LightGBM (0.702193)

Modelo	TN	FP	FN	TP	Precisión	Sensibilidad	ROC AUC	F_1	F_β	MCC
LR	12907	239	2	55	0.187075	0.964912	0.973366	0.313390	0.526820	0.420670
KN	13131	15	45	12	0.444444	0.210526	0.604693	0.285714	0.235294	0.303883
DT	13109	37	21	36	0.493151	0.631579	0.814382	0.553846	0.598007	0.555941
RF	13144	2	14	43	0.955556	0.754386	0.877117	0.843137	0.787546	0.848475
SVM	13116	30	19	38	0.558824	0.666667	0.832192	0.608000	0.641892	0.608533
XGB	13143	3	11	46	0.938776	0.807018	0.903395	0.867925	0.830325	0.869896
LGB	13134	12	4	53	0.815385	0.929825	0.964456	0.868852	0.904437	0.870136
IF	13030	116	57	0	0.000000	0.000000	0.495588	0.000000	0.000000	-0.006199

Tabla 5.4: Resultados experimento 3

y Random Forest (0.691962). Los tres casos tienen la misma problemática, excesivo valor para los falsos positivos.

F_1 en donde se ponderan de manera conjunta la precisión y la sensibilidad el mejor rendimiento lo obtiene SVM (0.077236) seguido de LightGBM (0.054271) en esta ocasión, aunque ambos siguen siendo muy bajos. Se trata de los rendimientos más equilibrados entre la precisión y la sensibilidad mostrando un incremento en la detección de verdaderos positivos (19 y 27 respectivamente).

F_β otorgando mayor peso a la sensibilidad respecto a la precisión vuelve a ser SVM el modelo con mejor rendimiento (0.143288) seguido en esta ocasión por LightGBM (0.115780).

Para el coeficiente de correlación de Mathews los mejores rendimientos los obtienen regresión logística (0.125932) y SVM (0.110812).

Desde el punto de vista de negocio ninguno de los modelos es viable ya que aquellos que tienen mejor detección de fraudes (verdaderos positivos) incurren en una excesiva tasa de transacciones legítimas bloqueadas (falsos positivos) así como una relación con las transacciones fraudulentas no bloqueadas (falsos negativos) alta, este último dato incide en una afectación económica inmediata.

5.3. Experimento 3

Para el *Experimento 3* el espacio de características definido en el *Experimento 2* es usado como base y complementado con el conjunto de características continuas de tiempo extendidas definidas en la Tabla 4.7. La intuición sobre este experimento es que se puede dotar a los modelos con información acerca de la evolución en el comportamiento transaccional, tanto legítimo como fraudulento. Las marcas de tiempo y los diferenciales de tiempo de los eventos seleccionados dotarían de capacidad de clasificación a los modelos de aprendizaje de máquinas.

5.3.1. Resultados

La Tabla 5.4 muestra los resultados del *Experimento 3*.

5.3.2. Discusión

En este caso es notable que el espacio de características extendido con las características continuas de tiempo permite mejorar el rendimiento general de los modelos seleccionados. Principalmente destaca el rendimiento de LightGBM con la más alta tasa de verdaderos positivos, una muy baja tasa de falsos negativos y una tasa baja y aceptable de falsos negativos.

El modelo que obtiene mejor resultado para la precisión es Random Forest (0.955556) seguido por XGBoost, en ambos casos la relación entre los fraudes detectados (43 y 46) respecto a los fraudes no detectados (14 y 11) es aceptable aunque aún se puede mejorar. Desde el punto de vista de negocio es deseable disminuir la cantidad de falsos negativos y elevar la cantidad de verdaderos positivos.

En cuanto a la sensibilidad el mejor resultado lo obtiene la regresión logística (0.964912) ya que detecta correctamente (TP) 55 de las 57 transferencias fraudulentas. El problema con la regresión logística es que su precisión es de las más bajas (0.187075) debido a que continua con un valor alto para los falsos positivos. el siguiente modelo con mejor desempeño para la sensibilidad es LightGBM (0.929825) y que además tuvo una precisión alta (0.815385). Desde el punto de vista de negocio la regresión logística no puede considerarse viable debido a su baja precisión. Por otro lado, LightGBM, XGBoost y Random Forest sí serían viables debido a su equilibrio entre precisión y sensibilidad.

ROC AUC muestra que la regresión logística (0.973366) es el modelo que mejor capacidad de clasificación tiene, seguido de LightGBM (0.964456), XGBoost (0.903395) y Random Forest (0.877117).

Desde la perspectiva de F_1 el mejor rendimiento lo obtiene LightGBM (0.868852) seguido por XGBoost (0.867925) y Random Forest (0.843137), aunque ambos siguen siendo muy bajos. Se trata de los rendimientos más equilibrados entre la precisión y la sensibilidad. Los tres modelos se ratifican como candidatos viables desde el punto de vista de negocio oportunidades de mejora. Por otro lado, esta métrica descarta a la regresión logística (0.313390) al ser una ponderación integral de la precisión y la sensibilidad.

En la perspectiva de F_β otorgando mayor peso a la sensibilidad respecto a la precisión LightGBM (0.904437) se posiciona con el mejor rendimiento, seguido por XGBoost (0.830325). En este caso se nota un disminución en el rendimiento de la métrica en comparación de F_1, este efecto es debido a las mejoras que aún se pueden alcanzar en cuanto a los fraudes no detectados (falsos negativos) y las transacciones legítimas bloqueadas (falsos positivos).

El coeficiente de correlación de Mathews reitera a LightGBM (0.870136), XGBoost (0.869896) y Random Forest (0.848475) como opciones viables de implementación.

Desde el punto de vista de negocio el modelo de mayor viabilidad es LightGBM que mantiene un alto valor de detección de fraudes (53 de 57), un valor adecuadamente bajo para los fraudes no detectados (4) y el valor por mejorar son las transacciones legítimas bloqueadas (12).

Modelo	TN	FP	FN	TP	Precisión	Sensibilidad	ROC AUC	F_1	F_β	MCC
LR	13122	24	38	19	0.441860	0.333333	0.665754	0.380000	0.350554	0.381475
KN	13146	0	57	0	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000
DT	13126	20	36	21	0.512195	0.368421	0.683450	0.428571	0.390335	0.432343
RF	13143	3	24	33	0.916667	0.578947	0.789360	0.709677	0.625000	0.727625
SVM	13144	2	54	3	0.600000	0.052632	0.526240	0.096774	0.064378	0.176842
XGB	13145	1	15	42	0.976744	0.736842	0.868383	0.840000	0.774908	0.847816
LGB	13139	7	4	53	0.883333	0.929825	0.964646	0.905983	0.920139	0.905865
IF	13100	46	57	0	0.000000	0.000000	0.498250	0.000000	0.000000	-0.003894

Tabla 5.5: Resultados experimento 4

5.4. Experimento 4

Se integran las características categóricas correspondientes a la cuenta origen y cuenta destino bajo un tratamiento de binarización con la técnica *Entity Embedding* [81]. El objetivo de ejecutar este procedimiento es capturar los patrones de relación entre ambas cuentas en conjunto respecto al comportamiento de las transferencias bancarias.

Sabemos que los usuarios de la banca por internet deben registrar las cuentas destino hacia las cuales realizarán transferencias, la intuición para incluir estas características es que existen patrones de comportamiento entre los pares de cuentas que fortalecen la relación haciéndolas de confianza (transferencias legítimas). Cuando no existen esos patrones o relaciones las cuentas se vuelven de menor confianza (transferencias fraudulentas). También existe el caso de los registros de cuentas nuevas dadas de alta por cuentas origen con diversas relaciones de confianza o por cuentas origen de baja confianza.

Al incluir estas características se busca aprender todos estos patrones e integrarlos al espacio previamente experimentado con el objetivo de mejorar el rendimiento de los modelos.

5.4.1. Resultados

La Tabla 5.5 muestra los resultados del *Experimento 4*.

5.4.2. Discusión

En este caso se logran niveles de clasificación bastante adecuados en los modelos XGBoost y LightGBM alcanzando un rendimiento superior este último ya que mantiene la tasa más alta de verdaderos positivos manteniendo la tasa de falsos positivos más baja y una tasa de falsos positivos adecuadamente baja. XGBoost alcanza una tasa de verdaderos positivos únicamente superada por LightGBM, mientras que su tasa de falsos positivos mejora el rendimiento a la correspondiente de LightGBM.

El modelo que obtiene mejor resultado para la precisión es XGBoost(0.976744)

seguido por Random Forest (0.916667) y LightGBM (0.883333). Destaca el hecho de que Random Forest obtiene una mejor precisión de LightGBM a pesar de que detecta menos transferencias fraudulentas, 33 frente a las 53 de LightGBM. Lo anterior debido a que Random Forest y XGBoost mantienen un valor bajo para los falsos positivos. Desde el punto de vista de negocio es deseable aumentar la cantidad de fraudes detectados tanto en RandomForest como en XGBoost, mientras que LightGBM debe disminuir las transacciones legítimas bloqueadas.

En cuanto a la sensibilidad el mejor resultado lo obtiene LightGBM (0.929825) ya que detecta correctamente (TP) 53 de las 57 transferencias fraudulentas. El siguiente modelo con mejor desempeño para la sensibilidad es XGBoost (0.736842) con un rendimiento considerablemente lejano. Desde el punto de vista de negocio ambos modelos serían viables con oportunidades de mejora.

ROC AUC muestra que LightGBM (0.964646) es el modelo que mejor capacidad de clasificación tiene, seguido de XGBoost (0.868383) ambos basados en el potenciamiento del gradiente.

Desde la perspectiva de F_1 el mejor rendimiento lo obtiene LightGBM (0.905983) seguido por XGBoost (0.840000). En ambos casos mantienen rendimientos equilibrados entre la precisión y la sensibilidad. Los dos modelos se ratifican como candidatos viables desde el punto de vista de negocio oportunidades de mejora.

En la perspectiva de F_β otorgando mayor peso a la sensibilidad respecto a la precisión LightGBM (0.920139) se posiciona con el mejor rendimiento marcando una considerable distancia respecto a XGBoost (0.774908). En el caso de LightGBM mejora ligeramente el rendimiento de la métrica en comparación con F_1, este efecto es debido a que los valores alcanzados para la detección de fraudes, fraudes legítimos bloqueados y fraudes no detectados son altamente aceptables con menores oportunidades de mejora.

El coeficiente de correlación de Mathews reitera a LightGBM (0.905865) como la opción más viable de implementación seguido por XGBoost (0.847816).

Desde el punto de vista de negocio también es mejor opción LightGBM dado que se minimiza la afectación económica al permitir la ejecución de únicamente cuatro fraudes, el costo en cuanto transferencias legítimas bloqueadas incorrectamente es bajo, únicamente siete, lo cual lo hace manejable y como ya se mencionó alcanza la tasa más alta de fraudes detectados correctamente (53).

5.5. Experimento 5

Se toma como base el *Experimento 4* e incorpora la técnica de Desplazamiento o movimiento del umbral con el objetivo de mejorar el rendimiento de los modelos.

El Desplazamiento o movimiento del umbral consiste en representar la curva generada entre las métricas Tasa de Falsos Positivos (*FPR - False Positive Rate*) y Tasa de Falsos Negativos *FNR - False Negative Rate* a través de diversos umbrales en el rango entre cero y uno. Posteriormente se calcula el área bajo la curva y se propone un umbral óptimo en el cual las métricas seleccionadas mejoren su rendimiento.

Modelo	TN	FP	FN	TP	Precisión	Sensibilidad	ROC AUC	F_1	F_β	MCC
LR	13122	24	38	19	0.441860	0.333333	0.665754	0.380000	0.350554	0.381475
KN	13146	0	57	0	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000
DT	13126	20	36	21	0.512195	0.368421	0.683450	0.428571	0.390335	0.432343
RF	13110	36	11	46	0.560976	0.807018	0.902140	0.661871	0.741935	0.671198
SVM	13144	2	54	3	0.600000	0.052632	0.526240	0.096774	0.064378	0.176842
XGB	13143	3	11	46	0.938776	0.807018	0.903395	0.867925	0.830325	0.869896
LGB	13139	7	4	53	0.883333	0.929825	0.964646	0.905983	0.920139	0.905865
IF	13100	46	57	0	0.000000	0.000000	0.498250	0.000000	0.000000	-0.003894

Tabla 5.6: Resultados experimento 5

Una técnica propuesta para encontrar el umbral de manera semi automática es calcular la métrica *G-mean* o Media geométrica y utilizar su valor como nuevo umbral. Otra técnica es visualizar la gráfica de la curva *FPR - FNR* y proponer un nuevo valor basado en el desplazamiento de dicha gráfica.

5.5.1. Resultados

La Tabla 5.6 muestra los resultados del *Experimento 5*.

5.5.2. Discusión

La técnica de desplazamiento del umbral en la mayoría de los casos permite alcanzar una ligera mejoría en el rendimiento de los modelos, la excepción notable es LightGBM que si bien ya no mejora su rendimiento tampoco decae. El mayor efecto de mejora del rendimiento lo alcanzan Random Forest y XGBoost, sin embargo, en términos generales XGBoost se mantiene más equilibrado en sus tasas de falsos positivos y falsos negativos, siendo la tasa de falsos positivos la que marca la diferencia entre ambos modelos. En ambos casos aunque mejoran sus métricas no alcanzan el rendimiento de LightGBM.

El modelo que obtiene mejor resultado para la precisión es XGBoost(0.938776) seguido por LightGBM (0.883333). XGBoost mejora ligeramente su rendimiento respecto al experimento anterior mientras que LightGBM mantiene el mismo rendimiento. Desde el punto de vista de negocio es deseable aumentar la cantidad de fraudes detectados tanto en XGBoost, mientras que LightGBM debe disminuir las transacciones legítimas bloqueadas.

En cuanto a la sensibilidad el mejor resultado lo obtiene LightGBM (0.929825) ya que detecta correctamente (TP) 53 de las 57 transferencias fraudulentas. Los siguientes modelos con mejor desempeño para la sensibilidad son XGBoost (0.807018) y Random Forest (0.807018) con un rendimiento considerablemente distante. Desde el punto de vista de negocio LightGBM y XGBoost son viables con ligeras oportunidades de mejora. Random Forest requiere mejorar su precisión.

ROC AUC muestra que LightGBM (0.964646) es el modelo que mejor capacidad de clasificación tiene, seguido de XGBoost (0.903395) y Random Forest (0.902140).

Desde la perspectiva de F_1 el mejor rendimiento lo obtiene LightGBM (0.905983) seguido por XGBoost (0.867925). En ambos casos mantienen rendimientos equilibrados entre la precisión y la sensibilidad. Los dos modelos se ratifican como candidatos viables desde el punto de vista de negocio oportunidades de mejora.

En la perspectiva de F_β otorgando mayor peso a la sensibilidad respecto a la precisión LightGBM (0.920139) se posiciona con el mejor rendimiento marcando una considerable distancia respecto a XGBoost (0.830325). En el caso de LightGBM mejora ligeramente el rendimiento de la métrica en comparación con F_1, este efecto es debido a que los valores alcanzados para la detección de fraudes, fraudes legítimos bloqueados y fraudes no detectados son altamente aceptables con menores oportunidades de mejora. En el caso de XGBoost disminuye su rendimiento respecto F_1 pero lo mejora respecto al anterior experimento.

El coeficiente de correlación de Mathews reitera a LightGBM (0.905865) como la opción más viable de implementación seguido por XGBoost (0.869896) que también mejora en esta métrica respecto al experimento anterior.

Desde el punto de vista de negocio también es mejor opción LightGBM dado que se minimiza la afectación económica al permitir la ejecución de únicamente cuatro fraudes, el costo en cuanto transferencias legítimas bloqueadas incorrectamente es bajo, únicamente siete, lo cual lo hace manejable y como ya se mencionó alcanza la tasa más alta de fraudes detectados correctamente (53).

Capítulo 6

Conclusiones

La detección de transferencias fraudulentas en un sistema de banca por internet es una problemática en auge en los últimos años. Aunque actualmente existen soluciones basadas en aprendizaje de máquinas y aprendizaje profundo que logran paliar en gran medida la situación una característica importante del problema es que es evolutivo. Los defraudadores actualizan sus técnicas y simulan el comportamiento de las transferencias legítimas.

Ejecutar el proceso de integración de las diversas fuentes de datos en un solo modelo de datos relacional permite tener un primer contacto con la información disponible y realizar una primera selección de características.

El análisis exploratorio de los datos arroja hallazgos importantes en cuanto a las relaciones entre las características y los patrones de comportamiento de los defraudadores. Los análisis de temporalidad y estacionalidad permitieron detectar patrones del comportamiento transaccional a través del tiempo. De esta forma se delineó la estrategia de partición del conjunto de datos que a su vez incidió en las fases de entrenamiento, validación y pruebas.

Por otro lado, el análisis exploratorio de los datos también permitió detectar la necesidad de integrar características extendidas debido a que, con el espacio de características original no se encontraban elementos en las distribuciones de los datos o en sus relaciones que permitieran detectar patrones susceptibles de generar criterios para separar adecuadamente las clases.

Un primer conjunto de características extendidas basado en la transaccionalidad en tres ventanas operativas (global, anual y mensual) no aportó los elementos suficientes para marcar una diferencia sustentable.

Un segundo conjunto de características extendidas esta vez basado en el comportamiento de los usuarios dentro de la sesión de la banca por internet plasmado a través de diferenciales de tiempo de diversos eventos permitió visualizar patrones que sí marcaban diferencia, específicamente en los eventos de inicio de sesión, registro de la cuenta destino, captura de la transferencia y cierre de sesión.

La selección de métricas para evaluar los modelos es un punto de relevancia debido a que se deben tomar en cuenta las necesidades del negocio. En este caso es de suma

importancia alcanzar una alta tasa de verdaderos positivos (fraudes detectados correctamente). También es importante mantener la tasa de falsos positivos (transacciones legítimas inferidas como fraudes) en un nivel mínimo ya que representa una afectación en la reputación de la institución. La tasa de falsos negativos es necesario llevarla al mínimo (lo más cercana a cero y con una ligera prioridad sobre la tasa de falsos positivos) ya que representa una afectación económica.

Del conjunto de modelos de aprendizaje de máquinas seleccionados se destacan los rendimientos alcanzados por LightGBM y XGBoost ambos basados en el algoritmo de potenciamiento del gradiente y utilizando por diseño ensambles de árboles de decisión como clasificadores internos.

La técnica de desplazamiento de umbral permite alcanzar ligeras mejorías en los rendimientos de la mayoría de los modelos experimentales, siendo la excepción LightGBM que antes de dicho procedimiento ya alcanzaba el mejor rendimiento global.

6.1. Trabajo futuro

Derivado de la investigación y experimentación realizadas para el presente trabajo así como de los resultados obtenidos, se vislumbran las siguientes líneas de investigación de cara a extender el trabajo a futuro:

- **Redes Neuronales Profundas**. En los años recientes los avances alcanzados en el rendimiento de las redes neuronales profundas en diversas tareas y casos de estudio han sido notables y la detección de fraudes bancarios no es la excepción. Por lo tanto, es interesante estudiar entre otras, las siguientes opciones:
  - **Redes Neuronales Recurrentes con capa LSTM**. Este tipo de redes cuenta con la capacidad de detectar patrones sobre secuencias de información, en particular de origen fueron diseñadas para atender el análisis de secuencias de tiempo; adicionalmente, las capas LSTM (*Long Short Term Memory*) las dotan con la capacidad de "recordar" patrones lejanos en el tiempo. En ese sentido, es de particular interés explorar su aplicación en el problema evolutivo de la detección de fraudes en transferencias de banca por internet.
  - **Redes Neuronales Autoencoders**. Los *autoencoders* son redes neuronales que permiten por un lado compactar la información a su mínima expresión conservando la representatividad, para posteriormente con base en la información compactada recuperar la información en su dimensión original. Este comportamiento es interesante explorarlo en el problema de detección de fraudes bajo la perspectiva modelar el comportamiento de las transacciones legítimas para después contrastarlo con la información de las transacciones fraudulentas.

  - **Redes Neuronales de Grafos**. Este tipo de redes han destacado por su capacidad de modelar relaciones y detectar patrones complejos. Su eventual exploración requerirá realizar transformaciones de la información con el objeto de describirla en formato de grafos.

- **Características de ubicación**. Para el presente estudio no se contó con características categóricas sobre la ubicación de las cuentas, información que sería relevante estudiar con el objeto de detectar patrones de comportamiento desde la perspectiva de la región o zona de la comisión del fraude.

- **Transferencias de Empresas**. El conjunto de datos utilizado en la presente investigación contiene únicamente transferencias cuya cuenta origen pertenece a una persona (física), a futuro se puede extender el estudio a cuentas origen pertenecientes a empresas ya que se espera un comportamiento transaccional diferente.

- **Características de biometría del comportamiento**. Si bien, en el presente proyecto se incluyeron características del comportamiento de los usuarios como los diferenciales entre las marcas de tiempo de inicio de sesión, registro de la cuenta destino, captura de los datos de la transferencia y el cierre de la sesión respecto a la ejecución de la transacción, a futuro se puede extender el estudio con datos de biometría del comportamiento, como el tiempo entre pantallas, cantidad de clics en la sesión o toques a la pantallas, etc.

Apéndice A

Diagramas entidad relación complementarios

A.1. Modelo de datos de la banca por internet

En la figura (A.1) se muestra la selección de relaciones provenientes del modelo de datos de la banca por internet.

A.2. Modelo de datos del sistema integral de autenticación

En la figura (A.2) se muestra la selección de relaciones provenientes del modelo de datos del sistema integral de autenticación.

A.3. Modelo de datos del sistema de identificación transaccional

En la figura (A.3) se muestra la selección de relaciones provenientes del modelo de datos del sistema de identificación transaccional.

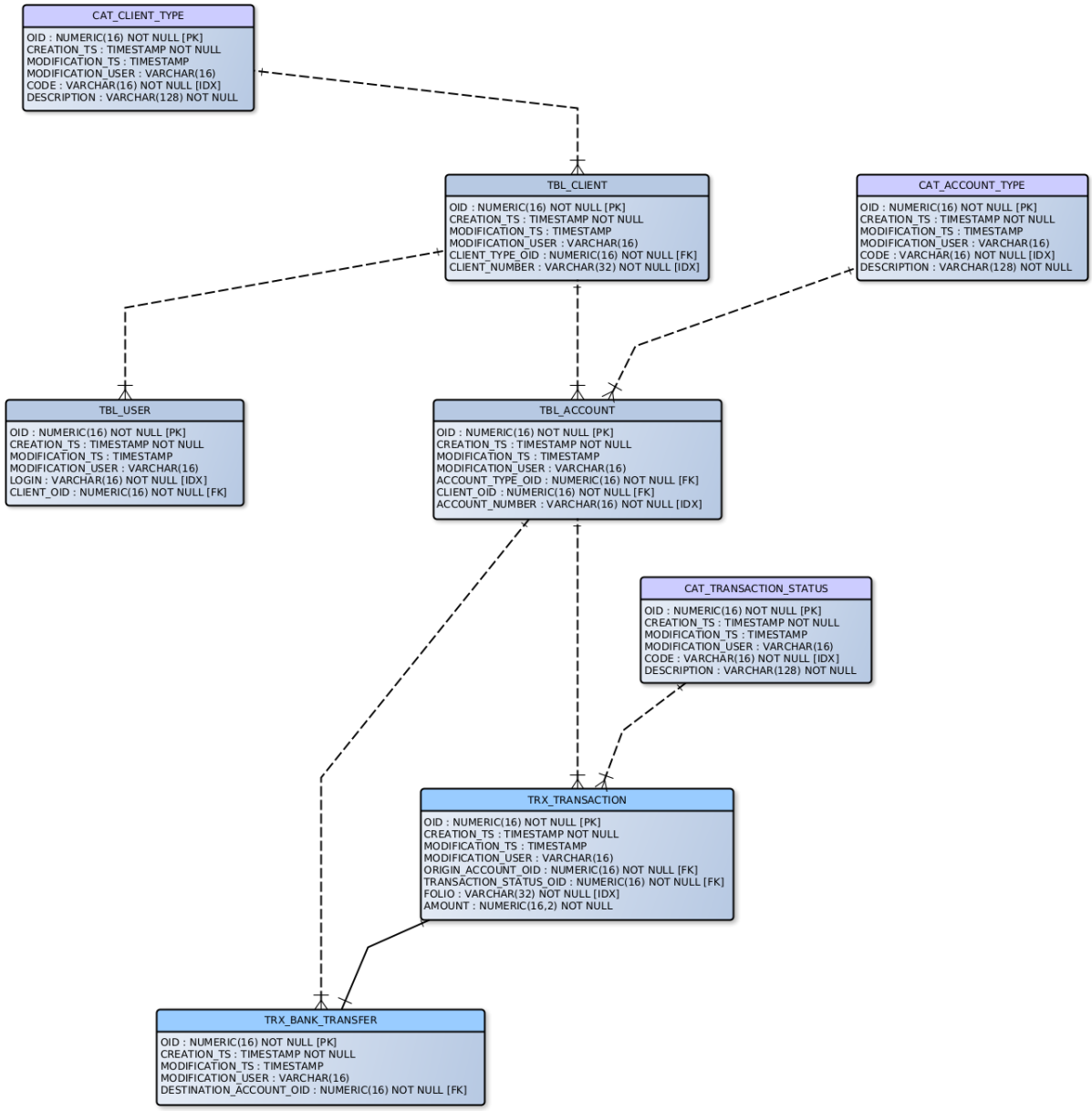

Figura A.1: Diagrama Entidad - Relación del modelo de datos del sistema de banca por internet. Relaciones seleccionadas para el proceso de integración de fuentes de datos.

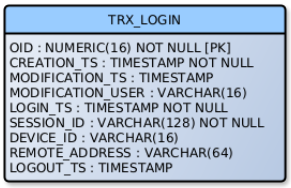

Figura A.2: Diagrama Entidad - Relación del modelo de datos del sistema integral de autenticación. Relación seleccionada para el proceso de integración de fuentes de datos.

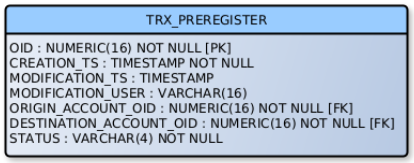

Figura A.3: Diagrama Entidad - Relación del modelo de datos del sistema de identificación transaccional. Relación seleccionada para el proceso de integración de fuentes de datos.

A.4. Modelo de datos diseñado para la bitácora transaccional

En la figura (A.4) se muestra el modelo de datos diseñado para integrar la información de la bitácora transaccional.

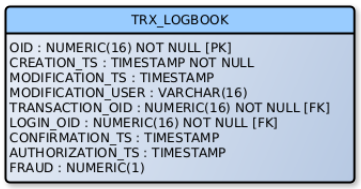

Figura A.4: Diagrama Entidad - Relación diseñado para integrar la información de la bitácora transaccional en el modelo de datos de la banca por internet.

Apéndice B

Análisis exploratorio de datos complementario

B.1. Exploración complementaria del monto

En la figura B.1 se muestra la distribución de los montos de las transferencias respecto a la clase por año. En lo general se detecta el mismo patrón observado en la gráfica global, en donde las transferencias legítimas se distribuyen en un rango más amplio y con una cantidad considerable de valores atípicos, mientras que la distribución de las transferencias fraudulentas es más compacta y siempre dentro del rango de las legítimas.

La figura B.2 muestra la dispersión de las transferencias respecto a (a) la fecha de apertura de la cuenta origen y (b) la fecha de registro de la cuenta destino. En ambos casos, se nota el efecto del tiempo en las cuentas más antiguas en las transacciones legítimas. En el caso de la fecha de apertura de la cuenta origen el patrón de comportamiento de las transferencias fraudulentas es muy similar al de las transferencias legítimas, es decir, se comenten fraudes con cuentas origen antigüas y recientes por igual. Mientras que en el caso de la fecha de registro de la cuenta destino se aprecia un patrón que indica que entre más antigüo sea el registro de la cuenta destino se comenten menos transferencias fraudulentas.

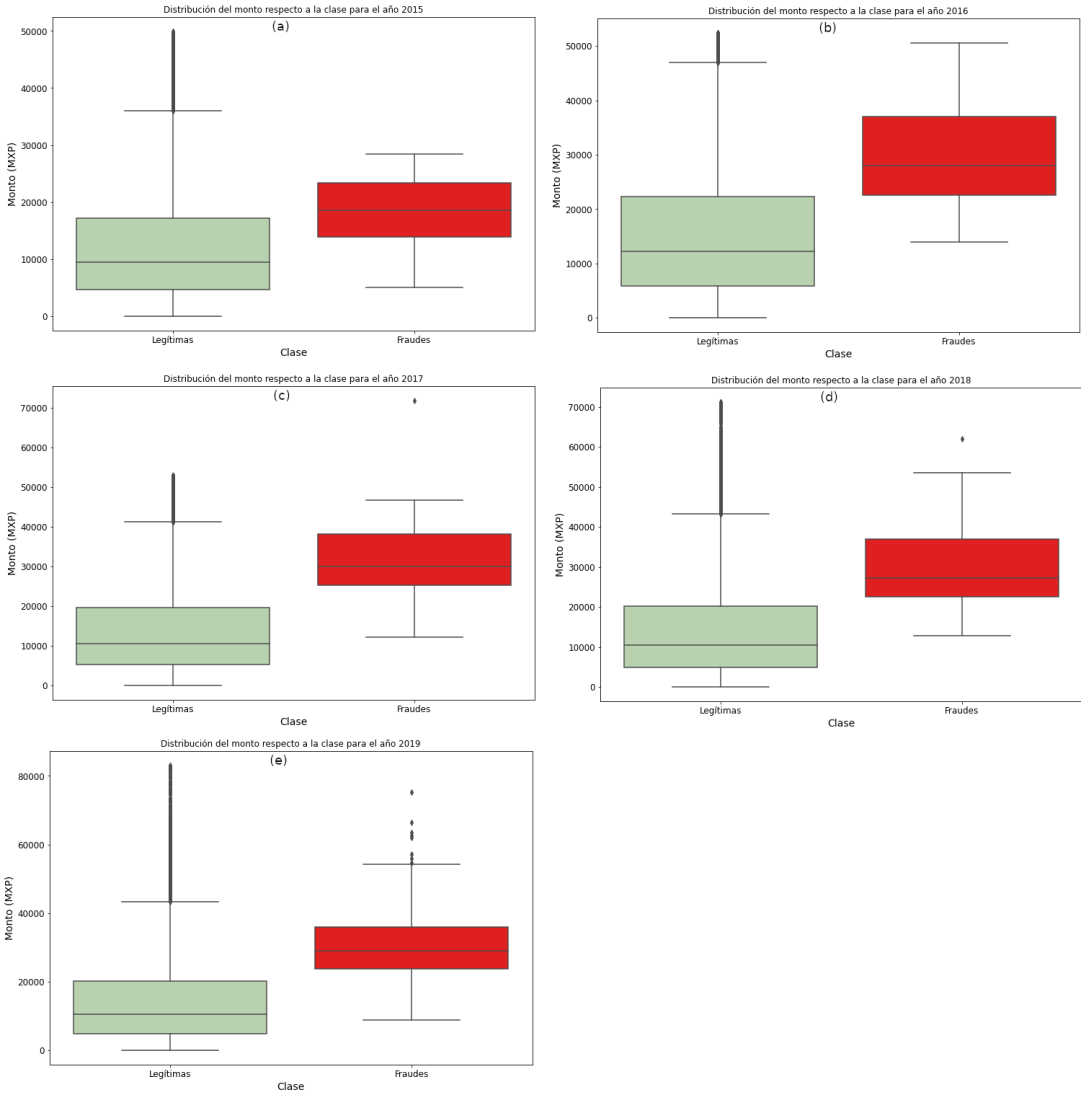

Figura B.1: Distribución del monto de las transferencias respecto a su clase para el año (a) 2015, (b) 2016, (c) 2017, (d) 2018 y (e) 2019.

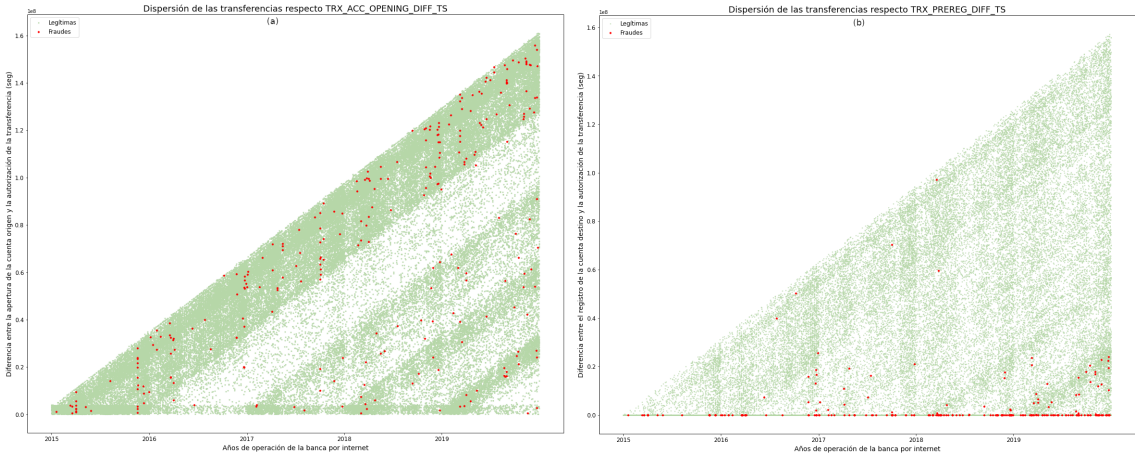

Figura B.2: Gráfica de dispersión de (a) los montos respecto a fecha de apertura de la cuenta, (b) los montos respecto a la fecha de registro de la cuenta destino

Bibliografía

[1] "Portal de fraudes financieros en México 2018," https://www.gob.mx/cms/uploads/attachment/file/400983/PORTAL_DE_FRAUDES_FINANCIEROS_vers7.pdf, accessed: 2021-08-02. xiii, 2

[2] I. Agur, S. Martinez Peria, and C. Rochon, "Digital financial services and the pandemic: Opportunities and risks for emerging and developing economies." *International Monetary Fund Special Series on COVID-19*, vol. Transactions 1, pp. 1–13, 2020, seminare Maurey-Schwartz (1975-1976). 1

[3] "BBVA méxico realizó casi 1,500 millones de transacciones digitales durante el año de la pandemia," https://www.bbva.com/es/mx/bbva-mexico-realizo-casi-1500-millones-de-transacciones-digitales-durante-el-ano-de-la-pandemi accessed: 2021-08-02. 1

[4] "ACFE," https://www.acfe.com/, accessed: 2021-05-12. 1

[5] ACFE, "Report to the nations. 2020 global study on occupational fraud and abuse," ACFE, Tech. Rep., 2020. 1

[6] T. Nilson report, "The Nilson report (2020)," The Nilson report, Tech. Rep., 2020. 2

[7] E.-A. Minastireanu and G. Mesnita, "An analysis of the most used machine learning algorithms for online fraud detection," *Informatica Economica*, vol. 23, pp. 5–16, 2019. 3, 6, 11, 42

[8] K. Navanshu and S. Y. Sait, "Credit card fraud detection using machine learning models and collating machine learning models," *International Journal of Pure and Applied Mathematics*, vol. 118, No. 20, pp. 825–838, 2018. 3, 11

[9] A. Husejinovic, "Credit card fraud detection using naive bayesian and c4.5 decision tree classifiers," *Periodicals of Engineering and Natural Sciences*, vol. 8, No. 1, pp. 1–5, 2020. 3, 11, 13

[10] R. Jain, G. Bhupesh, and S. Dubey, "A hybrid approach for credit card fraud detection using rough set and decision tree technique," *International Journal of Computer Applications*, vol. 139, No. 10, pp. 1–6, 2016. 3

[11] D. Devi, S. K. Biswas, and B. Purkayastha, "A cost-sensitive weighted random forest technique for credit card fraud detection," *10th international conference on computing, communication and networking technologies (ICCCNT)*, vol. 2019, pp. 1–6, 2019. 3, 11, 13

[12] M. S. Kumar, V. Soundarya, S. Kavitha, E. S. Keerthika, and E. Aswini, "Credit card fraud detection using random forest algorithm," *3rd International Conference on Computing and Communications Technologies (ICCCT)*, vol. 2019, pp. 1–6, 2019. 3

[13] Y. K. Saheed, M. A. Hambali, M. O. Arowolo, and Y. A. Olasupo, "Application of ga feature selection on naive bayes, random forest and svm for credit card fraud detection," *2020 International Conference on Decision Aid Sciences and Application (DASA)*, vol. 2020, pp. 1091–1097, 2020. 3

[14] A. Kannagi, J. Gori Mohammed, S. Sabari Giri Murugan, and M. Varsha, "Intelligent mechanical systems and its applications on online fraud detection analysis using pattern recognition k-nearest neighbor algorithm for cloud security applications," *Materials Today: Proceedings*, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2214785321031485 3

[15] J. Hyder and S. Naaz, "Credit card fraud detection using local outlier factor and isolation forest," *International Journal of Computer Sciences and Engineering*, vol. 7, Issue 4, pp. 1060–1064, 2019. 3, 11

[16] S. Ounacer, H. Ait-El-Bour, Y. Oubrahim, M. Y. Ghoumari, and M. Azzouazi, "Using isolation forest in anomaly detection: the case of credit card transactions," *Periodicals of Engineering and Natural Sciences (PEN)*, vol. 6, No.2, pp. 394–400, 2018. 3

[17] V. Vijayakumar, N. S. Divya, P. Sarojini, and K. Sonika, "Isolation forest and local outlier factor for credit card fraud detection system," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, Issue 4, pp. 261–265, 2020. 3, 11

[18] O. Ayano and S. O. Akinola, "A multi-algorithm data mining classification approach for bank fraudulent transactions," *African Journal of Mathematics and Computer Science Research*, vol. 10, No 1, pp. 5–13, 2017. 3

[19] C. Mishra, D. L. Gupta, and R. Singh, "Credit card fraud identification using artificial neural networks," *International Journal of Computer Systems*, vol. 04, Issue 07, pp. 5–13, 2017. 3

[20] P. Raghavan and N. E. Gayar, "Fraud detection using machine learning and deep learning," *International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, vol. 2019, pp. 334–339, 2019. 3

[21] S. Wang, C. Liu, X. Gao, H. Qu, and W. Xu, "Session-based fraud detection in online e-commerce transactions using recurrent neural networks," *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, vol. 2017, pp. 1–16, 2017. 3

[22] I. Benchaji, S. Douzi, and B. El-Ouahidi, "Credit card fraud detection model based on lstm recurrent neural networks," *Journal of Advances in Information Technology*, vol. 12, No. 2, pp. 113–118, 2021. 3

[23] Z. Zhang, X. Zhou, X. Zhang, L. Wang, and P. Wang, "A model based on convolutional neural network for online transaction fraud detection," *Security and Communication Networks*, vol. 2018, pp. 1–9, 2018. 3

[24] N. Laleh and M. A. Azgomi, "A taxonomy of frauds and fraud detection techniques," in *International Conference on Information Systems, Technology and Management*. Springer, 2009, pp. 256–267. 5, 6, 13

[25] H. Hofmann, "Statlog (german credit data) data set," 1994, datos obtenidos de UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data). 5, 14

[26] K. M. L. Group, "Credit card fraud detection. Anonymized credit card transactions labeled as fraudulent or genuine," 2018, datos obtenidos de Kaggle, https://www.kaggle.com/mlg-ulb/creditcardfraud. 5, 14

[27] K. Machine Learning Group, "IEEE-CIS fraud detection (Kaggle)," 2019, datos obtenidos de Kaggle, https://www.kaggle.com/c/ieee-fraud-detection/data. 5, 14

[28] M. Maiti, D. Vuković, A. Mukherjee, P. D. Paikarao, and J. K. Yadav, "Advanced data integration in banking, financial, and insurance software in the age of covid-19," *Software: Practice and Experience*, 2021. 6

[29] G.-D. Bobric, "The evolution of cyber fraud in the past decade," in *Proceedings of the 20 th European Conference on Cyber Warfare and Security*, 2021, pp. 44–51. 6, 13, 17

[30] A. K. Usman and M. H. Shah, "Critical success factors for preventing e-banking fraud," *The Journal of Internet Banking and Commerce*, vol. 18, no. 2, pp. 1–14, 2012. 6

[31] G. Cabanes, Y. Bennani, and N. Grozavu, "Unsupervised learning for analyzing the dynamic behavior of online banking fraud," in *2013 IEEE 13th International Conference on Data Mining Workshops*. IEEE, 2013, pp. 513–520. 6

[32] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical science*, vol. 17, no. 3, pp. 235–255, 2002. 11

[33] Y. Kou, C.-T. Lu, S. Sirwongwattana, and Y.-P. Huang, "Survey of fraud detection techniques," in *IEEE International Conference on Networking, Sensing and Control, 2004*, vol. 2. IEEE, 2004, pp. 749–754. 11

[34] R.-C. Chen, S.-T. Luo, X. Liang, and V. C. Lee, "Personalized approach based on svm and ann for detecting credit card fraud," in *2005 International Conference on Neural Networks and Brain*, vol. 2. IEEE, 2005, pp. 810–815. 11

[35] A. H. Sung and Q. Liu, "Behaviour mining for fraud detection," *Journal of research and practice in Information Technology*, vol. 39, no. 1, pp. 3–18, 2007. 11

[36] D. Yue, X. Wu, Y. Wang, Y. Li, and C.-H. Chu, "A review of data mining-based financial fraud detection research," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*. Ieee, 2007, pp. 5519–5522. 11

[37] A. Shen, R. Tong, and Y. Deng, "Application of classification models on credit card fraud detection," in *2007 International conference on service systems and service management*. IEEE, 2007, pp. 1–4. 11

[38] R. Bhowmik, "Data mining techniques in fraud detection," *Journal of Digital Forensics, Security and Law*, vol. 3, no. 2, p. 3, 2008. 11

[39] S. Lei, K. Xu, Y. Huang, and X. Sha, "An xgboost based system for financial fraud detection," *E3S Web of Conferences*, vol. 2020, p. 785–794, 214. 11, 14

[40] N. K. Chauhan and K. Singh, "A review on conventional machine learning vs deep learning," in *2018 International Conference on Computing, Power and Communication Technologies (GUCON)*. IEEE, 2018, pp. 347–352. 11

[41] L. F. W. Anthony, B. Kanding, and R. Selvan, "Carbontracker: Tracking and predicting the carbon footprint of training deep learning models," *arXiv preprint arXiv:2007.03051*, 2020. 11

[42] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021. 11

[43] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Information Fusion*, vol. 42, pp. 146–157, 2018. 11

[44] X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *IEEE access*, vol. 2, pp. 514–525, 2014. 11

[45] D. V. Carvalho, E. M. Pereira, and J. S. Cardoso, "Machine learning interpretability: A survey on methods and metrics," *Electronics*, vol. 8, no. 8, p. 832, 2019. 13

[46] J. R. Quinlan, "Statlog (australian credit approval) data set," 1987, datos obtenidos de UCI Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/statlog+(australian+credit+approval). 14

[47] U. of California Irvine, "University of california irvine," 2021, http://archive.ics.uci.edu/ml. 14

[48] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*, vol. 2018, p. 785–794, 2016. 14

[49] "XGBoost," https://github.com/dmlc/xgboost, accessed: 2021-07-23. 14, 59

[50] "Kaggle," https://www.kaggle.com/, accessed: 2021-07-23. 14

[51] Y. Zhang, J. Tong, Z. Wang, and F. Gao, "Customer transaction fraud detection using xgboost model," *2020 International Conference on Computer Engineering and Application (ICCEA)*, vol. 2020, p. 554–558, 2020. 14

[52] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on svm-recursive feature elimination and hyper-parameters optimization," *Journal of Information Security and Applications*, vol. 2020, p. 1–15, 2020. 14

[53] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*. IEEE, 2007, pp. 429–435. 14

[54] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002. 14

[55] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization." *Journal of machine learning research*, vol. 13, no. 2, 2012. 14

[56] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016. 17

[57] H. He and Y. Ma, "Imbalanced learning: foundations, algorithms, and applications," 2013. 17

[58] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986. 18

[59] ——, *C4.5: programs for machine learning*. Elsevier, 2014. 18

[60] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017. 18

[61] S. Tangirala, "Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 2, pp. 612–619, 2020. 18

[62] L. Breiman, "Random forests," *UC Berkeley TR567*, 1999. 19

[63] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006. 20

[64] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*, 1992, pp. 144–152. 21

[65] F. Friedrichs and C. Igel, "Evolutionary tuning of multiple svm parameters," *Neurocomputing*, vol. 64, pp. 107–117, 2005. 22

[66] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 eighth ieee international conference on data mining*. IEEE, 2008, pp. 413–422. 23

[67] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, p. 93–104, may 2000. [Online]. Available: https://doi.org/10.1145/335191.335388 24

[68] S. Mittal and S. Tyagi, "Performance evaluation of machine learning algorithms for credit card fraud detection," in *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2019, pp. 320–324. 24, 26

[69] A. Luque, A. Carrasco, A. Martín, and A. de las Heras, "The impact of class imbalance in classification performance metrics based on the binary confusion matrix," *Pattern Recognition*, vol. 91, pp. 216–231, 2019. 25

[70] D. Powers, "Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation," *Mach. Learn. Technol.*, vol. 2, 01 2008. 26

[71] D. Chicco, N. Tötsch, and G. Jurman, "The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData mining*, vol. 14, no. 1, pp. 1–22, 2021. 27

[72] J. W. Tukey *et al.*, *Exploratory data analysis*. Reading, Mass., 1977, vol. 2. 42

[73] P. Good, "Robustness of pearson correlation," *Interstat*, vol. 15, no. 5, pp. 1–6, 2009. 52

[74] "Python," https://www.python.org/, accessed: 2021-09-23. 59

[75] "NumPy," https://numpy.org/, accessed: 2021-09-23. 59

[76] "Pandas," https://pandas.pydata.org/, accessed: 2021-09-23. 59

[77] "SciKit-Learn," https://scikit-learn.org/stable/, accessed: 2021-09-23. 59

[78] "LightGBM," https://lightgbm.readthedocs.io/en/latest/, accessed: 2021-07-23. 59

[79] C. Seger, "An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing," 2018. 60

[80] M. Köppen, "The curse of dimensionality," in *5th Online World Conference on Soft Computing in Industrial Applications (WSC5)*, vol. 1, 2000, pp. 4–8. 60

[81] C. Guo and F. Berkhahn, "Entity embeddings of categorical variables," *arXiv preprint arXiv:1604.06737*, 2016. 65