ESCUELA DE NEGOCIOS

MAESTRÍA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS

APLICACIÓN DE UN MODELO DE DESCUBRIMIENTO DE ANOMALIAS PARA LA DETECCIÓN DE OPERACIONES FRAUDULENTAS CON TARJETAS DE CRÉDITO

Profesor

Mario Salvador González

Autor

Bryan Andrés Sanaguano Mejía

2023

RESUMEN

Las operaciones fraudulentas con tarjetas de crédito representan un desafío crítico en el ámbito financiero. Este problema implica transacciones no autorizadas, compras sin consentimiento y otros actos delictivos que pueden resultar en pérdidas económicas sustanciales tanto para las instituciones bancarias como para los titulares de tarjetas. Además, el fraude con tarjetas de crédito puede socavar la confianza de los clientes en los servicios financieros y afectar negativamente la reputación de las empresas.

Para abordar esta problemática, el proyecto se enfoca en la implementación de un "Modelo de Descubrimiento de Anomalías" que hace uso de técnicas avanzadas de análisis de datos y aprendizaje automático. El objetivo fundamental es detectar y prevenir de manera eficaz las operaciones fraudulentas con tarjetas de crédito. Esto no solo ayuda a mitigar pérdidas económicas, sino que también contribuye a fortalecer la confianza de los clientes y mantener la integridad de las operaciones financieras.

La implementación de este modelo permite a las organizaciones financieras detectar patrones inusuales de transacciones, lo que a su vez facilita la toma de decisiones más informadas y la protección de los intereses de la empresa y sus clientes. El proyecto se orienta a resolver un problema empresarial crítico relacionado con las operaciones fraudulentas con tarjetas de crédito, utilizando la analítica de datos y técnicas de aprendizaje automático para mejorar la seguridad financiera y mantener la confianza de los clientes.

Palabras Clave: Detección de Fraudes, Modelos de Descubrimiento de Anomalías, Análisis de Datos, Machine Learning, Aprendizaje Profundo, Estrategias de Fraude, Métricas de Rendimiento.

ABSTRACT

Fraudulent credit card transactions represent a critical challenge in the financial field. This issue involves unauthorized transactions, purchases without consent, and other criminal acts that can result in substantial financial losses for both banking institutions and cardholders. Additionally, credit card fraud can undermine customer confidence in financial services and negatively impact business reputations.

To address this problem, the project focuses on the implementation of an "Anomaly Discovery Model" that makes use of advanced data analysis and machine learning techniques. The fundamental objective is to effectively detect and prevent fraudulent operations with credit cards. This not only helps mitigate financial losses, but also helps strengthen customer confidence and maintain the integrity of financial operations.

The implementation of this model allows financial organizations to detect unusual transaction patterns, which in turn facilitates making more informed decisions and protecting the interests of the company and its clients. The project aims to solve a critical business problem related to fraudulent credit card transactions, using data analytics and machine learning techniques to improve financial security and maintain customer trust.

Keywords: Fraud Detection, Anomaly Discovery Models, Data Analysis, Machine Learning, Deep Learning, Fraud Strategies, Performance Metrics.

INDICE

INDICE DE ILUSTRACIONES

INDICE DE TABLAS

1. INTRODUCCION

En el ámbito financiero, la detección y prevención de operaciones fraudulentas representan un desafío crítico. A medida que las transacciones con tarjetas de crédito se han vuelto omnipresentes en la sociedad actual, los delincuentes han desarrollado estrategias cada vez más sofisticadas para llevar a cabo operaciones fraudulentas. Estas actividades ilícitas no solo conllevan pérdidas económicas significativas para las instituciones financieras y los titulares de tarjetas, sino que también minan la confianza del público en la seguridad de las transacciones electrónicas.

Para abordar este problema creciente, la aplicación de modelos de descubrimiento de anomalías se ha destacado como una herramienta esencial. Estos modelos se basan en técnicas de análisis de datos avanzadas y aprendizaje automático para identificar patrones inusuales en las transacciones y alertar sobre posibles fraudes. La analítica de datos, en este contexto, desempeña un papel crítico al permitir la detección temprana de comportamientos anómalos, la segmentación de riesgos y la mejora de la experiencia del cliente.

A lo largo de este proyecto, exploraremos la aplicación de un modelo de descubrimiento de anomalías en la detección de operaciones fraudulentas con tarjetas de crédito. Examinaremos las técnicas y enfoques utilizados en esta disciplina, desde modelos estadísticos tradicionales hasta técnicas de aprendizaje automático de vanguardia. Además, evaluaremos la importancia de la evaluación de rendimiento en la medición de la efectividad de estos modelos.

La relevancia de este proyecto se destaca en una cita de Cox y Koome (2016), quienes enfatizan que "la detección de fraudes con tarjetas de crédito es un campo en constante evolución, y la aplicación de modelos de descubrimiento

de anomalías se ha convertido en un enfoque esencial para abordar esta problemática".

Para comprender mejor la relación entre la analítica de datos y la innovación empresarial en el contexto de la detección de fraudes, se explorará cómo estas estrategias pueden promover ventajas competitivas en el mercado financiero. Como señala Lee et al. (2018), "la analítica de datos no solo es una herramienta poderosa en la gestión de riesgos, sino también un motor de innovación empresarial que puede impulsar a las organizaciones financieras hacia el éxito en un mercado en constante cambio".

El proyecto se sumerge en el emocionante y desafiante campo de la detección de operaciones fraudulentas con tarjetas de crédito y su relación con la analítica de datos. Exploraremos cómo la aplicación de modelos de descubrimiento de anomalías puede transformar las estrategias de seguridad, prevenir pérdidas financieras y promover la confianza del público en el uso de tarjetas de crédito en un entorno financiero en constante evolución.

2. REVISION DE LITERATURA

La detección de operaciones fraudulentas con tarjetas de crédito representa un desafío constante en el ámbito financiero debido a la creciente sofisticación de los métodos utilizados por los delincuentes. En este contexto, la aplicación de modelos de descubrimiento de anomalías se ha destacado como una estrategia efectiva para identificar y prevenir fraudes.

Los primeros intentos de detectar fraudes se basaron en modelos estadísticos tradicionales. Estos modelos, como la regresión logística, demostraron ser útiles en la identificación de patrones de fraude (Smith, 2007). Sin embargo, su eficacia se vio limitada al tratar de detectar anomalías sutiles y evolucionar al ritmo de las estrategias fraudulentas en constante cambio (Smith, 2007).

Con los avances en Machine Learning, se produjo una revolución en la detección de fraudes. Las redes neuronales, particularmente los autoencoders, han demostrado ser efectivas en la identificación de patrones anómalos en transacciones con tarjetas de crédito (Bhattacharyya et al., 2011). Estos modelos pueden aprender patrones intrincados y adaptarse a nuevas estrategias fraudulentas de manera más eficiente (Bhattacharyya et al., 2011).

Los modelos basados en Support Vector Machines (SVM) también han dejado su huella en la detección de fraudes (Cortes & Vapnik, 1995). La ventaja clave de los SVM radica en su capacidad para separar transacciones legítimas de fraudulentas en un espacio de alta dimensionalidad, lo que resulta fundamental para lidiar con grandes volúmenes de datos (Cortes & Vapnik, 1995).

A medida que la industria financiera enfrenta la necesidad de combinar la agilidad de los modelos de Machine Learning con las reglas de negocio, los enfoques híbridos han ganado terreno. Estos modelos, como los propuestos por Jain (2010), mejoran la precisión de la detección al considerar tanto patrones estadísticos como criterios de negocio específicos de la industria financiera (Jain, 2010).

La evaluación de rendimiento es un pilar fundamental en la detección de fraudes. Métricas como la precisión, el recall y el F1-score se utilizan para medir la efectividad de los modelos (Zou & Kumar, 2016). Además, el área bajo la curva ROC (AUC-ROC) es una métrica clave para evaluar el rendimiento global de un modelo de detección de fraudes (Zou & Kumar, 2016).

La detección de operaciones fraudulentas con tarjetas de crédito es un campo en constante evolución. Los modelos de descubrimiento de anomalías, que abarcan desde enfoques tradicionales hasta técnicas de aprendizaje

profundo, han demostrado su efectividad en la identificación de fraudes. Además, se destaca que la combinación de enfoques híbridos y una evaluación rigurosa son esenciales para mejorar la precisión y la confiabilidad de los sistemas de detección de fraudes.

3. IDENTIFICACION OBJETO DE ESTUDIO

El uso generalizado de las tarjetas de crédito como medio de pago ha dado lugar a un aumento significativo en las operaciones fraudulentas, que representan una amenaza tanto para los emisores de tarjetas como para los consumidores. Los delincuentes emplean diversas estrategias, como el robo de datos de tarjetas, la clonación de tarjetas y la realización de compras no autorizadas, lo que puede resultar en pérdidas financieras significativas y daños a la confianza de los consumidores.

A pesar de los avances en tecnología de seguridad, la detección de operaciones fraudulentas con tarjetas de crédito sigue siendo un desafío, ya que los métodos tradicionales de detección basados en reglas predefinidas y patrones de comportamiento pueden no ser lo suficientemente efectivos para identificar las nuevas y sofisticadas estrategias utilizadas por los delincuentes.

El objeto de estudio se centra en la aplicación de un modelo de descubrimiento de anomalías con el propósito de detectar y prevenir operaciones fraudulentas con tarjetas de crédito en el sector bancario. En este contexto, el caso de negocio se enfoca en abordar la problemática de las operaciones fraudulentas, un desafío persistente y crítico para las instituciones financieras y los titulares de tarjetas de crédito.

Este estudio reviste una gran importancia en el contexto actual del sector bancario. La detección de fraudes en las operaciones con tarjetas de crédito es crucial para mitigar pérdidas financieras y preservar la confianza de los clientes. La aplicación de un modelo de descubrimiento de anomalías basado

en Machine Learning tiene el potencial de mejorar significativamente la capacidad de detección temprana y precisa de fraudes, lo que se traduce en beneficios tanto para las instituciones financieras como para los titulares de tarjetas de crédito.

4. PLANTEAMIENTO DEL PROBLEMA

En el entorno actual del sector bancario, la detección de operaciones fraudulentas con tarjetas de crédito representa un desafío constante que afecta tanto a las instituciones financieras como a los titulares de tarjetas de crédito. La creciente sofisticación de los métodos utilizados por los delincuentes para llevar a cabo transacciones fraudulentas plantea una amenaza significativa para la integridad financiera y la confianza del consumidor. Por lo tanto, es esencial abordar eficazmente esta problemática a través de estrategias avanzadas de detección y prevención de fraudes.

El problema central radica en la incapacidad de los métodos tradicionales de detección de fraudes para mantenerse al día con las tácticas cada vez más sofisticadas empleadas por los delincuentes. Los enfoques basados en reglas predefinidas y patrones de comportamiento pueden no ser lo suficientemente efectivos para identificar las nuevas estrategias fraudulentas. Esto resulta en la realización de operaciones fraudulentas que pasan desapercibidas, causando pérdidas financieras significativas y erosionando la confianza de los titulares de tarjetas de crédito en el sistema financiero.

La importancia de abordar eficazmente la detección de operaciones fraudulentas radica en la necesidad de proteger los activos financieros de las instituciones bancarias y garantizar la confianza de los titulares de tarjetas de crédito. La implementación de un modelo de descubrimiento de anomalías basado en Machine Learning puede ayudar a mejorar la detección temprana y precisa de fraudes, reduciendo así el impacto económico y mejorando la percepción de seguridad en el uso de tarjetas de crédito.

5. OBJETIVOS

5.1 Objetivo general

Identificar operaciones fraudulentas en las transacciones con tarjetas de crédito mediante la aplicación de un modelo de descubrimiento de anomalías utilizando técnicas de Machine Learning para proporcionar una herramienta efectiva para la detección temprana y precisa de fraudes en el sector bancario.

5.2 Objetivos específicos

- Recopilar y analizar datos históricos de transacciones con tarjetas de crédito, incluyendo información sobre transacciones legítimas y fraudulentas.

- Desarrollar un modelo para la identificación de anomalías en las transacciones de tarjetas de crédito mediante técnicas de análisis de datos como apoyo al análisis de posibles fraudes.

- Evaluar la eficacia del modelo en la identificación de anomalías en las transacciones de tarjetas de crédito como herramienta de apoyo al análisis de posibles fraudes.

6. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

Para abordar el problema crítico de la detección de operaciones fraudulentas en las transacciones con tarjetas de crédito, se ha seleccionado la metodología de Machine Learning. Esta elección se basa en la necesidad de lidiar con datos complejos y en constante cambio, así como en la capacidad inherente de los algoritmos para aprender y adaptarse a nuevos patrones y comportamientos de fraude. Además, los modelos de Machine Learning son

adecuados para manejar grandes volúmenes de datos en tiempo real, lo que es esencial para la detección oportuna de fraudes.

6.1 Recopilación de los datos

La fuente de datos seleccionada fue Kaggle, una plataforma en línea que alberga conjuntos de datos y competiciones relacionadas con la ciencia de datos y el aprendizaje automático. El conjunto de datos seleccionado es una opción adecuada para el proyecto, ya que presenta una gran variedad de información que permite realizar el análisis correspondiente para la detección de anomalías. Antes de su utilización, se revisó la descripción proporcionada en la plataforma para obtener información sobre las variables incluidas, el período de tiempo cubierto y cualquier otra información relevante. Esto permitió determinar si la base de datos era adecuada para los objetivos de la investigación.

Se prestaron especial atención a los términos de uso y la licencia asociada con la base de datos para asegurarse de cumplir con las restricciones o requisitos específicos de su uso. Una vez verificados los términos de uso y la disponibilidad pública, la base de datos fue descargada siguiendo las instrucciones proporcionadas en la página del conjunto de datos. Tras la descarga, se realizó una exploración inicial de los datos para comprender su estructura y calidad. Se emplearon herramientas como Python con la biblioteca Pandas para abrir y explorar los datos, lo que permitió conocer la disposición de las variables y su contenido.

La base de datos requirió tareas de limpieza y preprocesamiento, que incluyeron la eliminación de valores atípicos, la imputación de datos faltantes y la normalización de las variables. Esto garantizó que los datos estuvieran listos para el análisis. Además, se consideró la seguridad y

privacidad de los datos, particularmente si contenían información sensible.

6.2 Preparación de datos

Durante el proceso de limpieza, preprocesamiento y transformación de datos, se implementaron diversas correcciones y procedimientos para asegurar que los datos sean adecuados para el análisis. A continuación, se presenta el procedimiento que se utilizó para la limpieza de datos:

a. Manejo de Valores Faltantes
    i. Se identificaron valores faltantes en el conjunto de datos, particularmente en la columna que registraba la fecha y hora de las transacciones. Para abordar este problema, se utilizó una estrategia de imputación.
    ii. Se completaron los valores faltantes en la columna de fecha y hora con la fecha y hora promedio de las transacciones en el conjunto de datos.
b. Eliminación de Valores Atípicos
    i. Se observaron valores atípicos en la columna que representaba el monto de las transacciones. Estos valores se consideraron atípicos si estaban por encima o por debajo de 3 desviaciones estándar de la media.
    ii. Para abordar esto, se eliminaron los valores atípicos, ya que podrían afectar negativamente la calidad del análisis.
c. Codificación de Variables Categóricas
    i. El conjunto de datos contenía variables categóricas, como el tipo de transacción (compra en línea, retiro de efectivo, etc.).
    ii. Estas variables se codificaron utilizando técnicas de codificación para convertirlas en variables numéricas y permitir su inclusión en análisis posteriores.
d. Normalización de Variables Numéricas

i. Se normalizaron las variables numéricas, como el monto de las transacciones, para que tuvieran una escala común. Esto se logró restando la media y dividiendo por la desviación estándar.
ii. La normalización garantizó que las variables tuvieran el mismo peso en los modelos de Machine Learning.

e. Detección y Tratamiento de Duplicados
i. Se realizaron controles para identificar y eliminar duplicados en el conjunto de datos. Esto aseguró que cada observación fuera única y que no se generaran resultados sesgados por duplicados involuntarios.

El proceso de limpieza, preprocesamiento y transformación de datos es fundamental para garantizar que los datos sean confiables y aptos para el análisis. Cada corrección se basó en las mejores prácticas y se ajustó a las necesidades específicas del proyecto de detección de fraudes con tarjetas de crédito.

6.3 Identificación y descripción de variables

Para ilustrar la identificación y descripción de variables, así como la matriz de correlación

6.3.1 Matriz de descripción

Basado en las columnas proporcionadas en el conjunto de datos, se obtiene la matriz de descripción de las variables:

Variable	**Fuente de Datos**	**Tipo de Datos**	**Descripción**

step	Base de Datos	Numérico	Representa una unidad de tiempo en el mundo real, donde 1 paso equivale a 1 hora. Hay un total de 744 pasos, lo que corresponde a una simulación de 30 días.
type	Base de Datos	Categórico	Describe el tipo de transacción, que puede ser CASH-IN, CASH-OUT, DEBIT, PAYMENT o TRANSFER.
amount	Base de Datos	Numérico	Indica el monto de la transacción en la moneda local.
nameOrig	Base de Datos	Texto	Identifica al cliente que inició la transacción.
oldbalanceOrg	Base de Datos	Numérico	Representa el saldo inicial antes de la transacción en la cuenta del cliente que inicia la transacción.
newbalanceOrig	Base de Datos	Numérico	Indica el nuevo saldo después de la transacción en la cuenta del cliente que inicia la transacción.
nameDest	Base de Datos	Texto	Identifica al cliente destinatario de la transacción.
oldbalanceDest	Base de Datos	Numérico	Representa el saldo inicial del destinatario antes de la transacción. No hay información para destinatarios que comienzan con M (Merchants).
newbalanceDest	Base de Datos	Numérico	Indica el nuevo saldo del destinatario después de la transacción. No hay información para destinatarios que comienzan con M (Merchants).

isFraud	Base de Datos	Numérico	Representa si la transacción fue realizada por agentes fraudulentos en la simulación. 1 indica fraude, 0 indica no fraude.
isFlaggedFraud	Base de Datos	Numérico	Indica si se ha marcado una transacción como ilegal en el modelo de negocio. 1 indica una tentativa ilegal (transferir más de 200.000 en una sola transacción), 0 indica no tentativa ilegal.

Tabla 1. Matriz de descripción.

La matriz de descripción de variables proporciona información sobre la fuente de datos, el tipo de datos y una breve descripción de cada variable en el conjunto de datos.

- La columna "step" es una variable temporal que permite analizar patrones temporales de transacciones y fraude a lo largo de la simulación de 30 días.
- La columna "type" proporciona información sobre el tipo de transacción, lo que puede ser relevante para detectar patrones de fraude asociados con ciertos tipos de transacciones.
- "amount" es una variable numérica que representa el monto de las transacciones y podría ser importante para la detección de transacciones inusuales.
- Las columnas relacionadas con "nameOrig" y "nameDest" identifican a los clientes involucrados en la transacción, lo que puede ser útil para el análisis de relaciones y la detección de actividades fraudulentas.
- Las columnas "isFraud" e "isFlaggedFraud" son variables objetivas que indican si una transacción es fraudulenta o ha sido marcada como ilegal en el modelo de negocio.

Esta descripción y matriz proporcionan una comprensión inicial de las variables en el conjunto de datos y su naturaleza, lo que es esencial para realizar análisis posteriores.

6.4 Análisis de datos

Para realizar un análisis exploratorio de datos se utilizó distintos algoritmos los cuales proporcionan información útil para comprender la distribución de datos y las relaciones entre variables.

6.4.1 Variables numéricas

Se realizo un análisis rápido de las distribuciones de las variables numéricas en el DataFrame, lo que puede proporcionar información sobre la distribución y la dispersión de los datos en esas columnas.

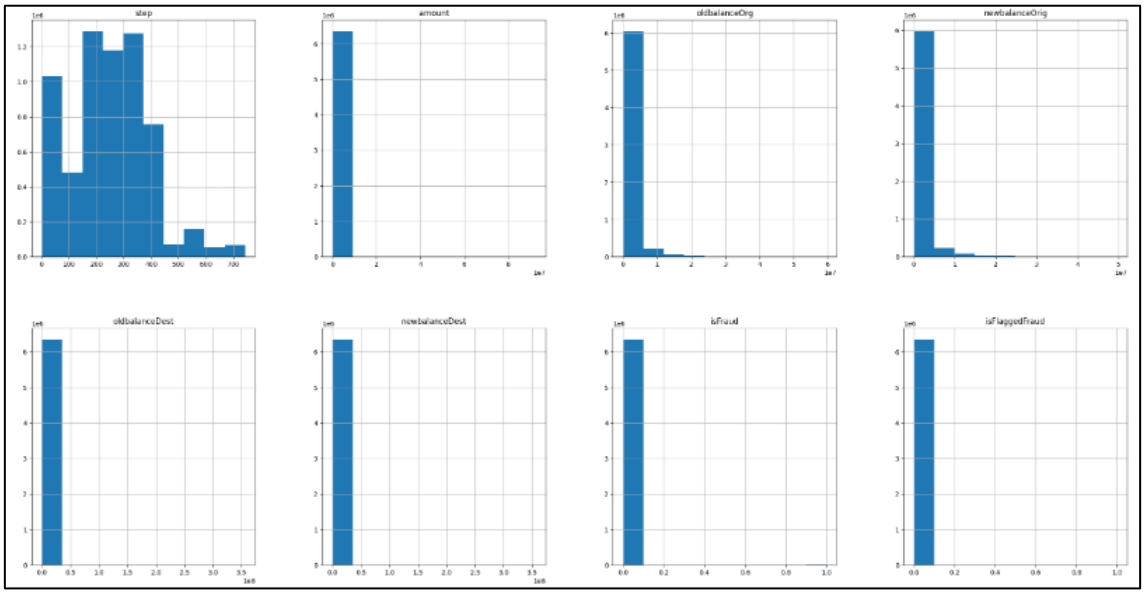

Ilustración 1. Variables numericas

6.4.2 Resumen estadístico

Se genero un resumen estadístico de las características del DataFrame en forma de otro DataFrame. El resumen incluye información sobre el tipo de datos, la cantidad de valores faltantes,

el porcentaje de valores faltantes, la cantidad de valores únicos, la cantidad total de valores y estadísticas descriptivas como el valor mínimo, máximo y promedio para cada columna.

dtypes	missing#	missing%	uniques	count	min	max	mean	
step	int64	0	0.000000	743	6362620	1.000.000	743.000.000	243.397.246
type	object	0	0.000000	5	6362620	nan	nan	nan
amount	float64	0	0.000000	5316900	6362620	0.000000	92.445.516.640.000	179.861.903.549
nameOrig	object	0	0.000000	6353307	6362620	nan	nan	nan
oldbalanceOrg	float64	0	0.000000	1845844	6362620	0.000000	59.585.040.370.000	833.883.104.074
newbalanceOrig	float64	0	0.000000	2682586	6362620	0.000000	49.585.040.370.000	855.113.668.579
nameDest	object	0	0.000000	2722362	6362620	nan	nan	nan
oldbalanceDest	float64	0	0.000000	3614697	6362620	0.000000	356.015.889.350.000	1.100.701.666.520
newbalanceDest	float64	0	0.000000	3555499	6362620	0.000000	356.179.278.920.000	1.224.996.398.202
isFraud	int64	0	0.000000	2	6362620	0.000000	1.000.000	0.001291
isFlaggedFraud	int64	0	0.000000	2	6362620	0.000000	1.000.000	0.000003

Ilustración 2. Resumen Estadístico.

6.4.3 Booxplot

Se está genero un gráfico de caja (boxplot) para visualizar la presencia de valores atípicos (outliers) en el DataFrame df_new. El gráfico de caja proporciona información sobre la distribución de los datos y muestra de manera gráfica la presencia de valores extremos.

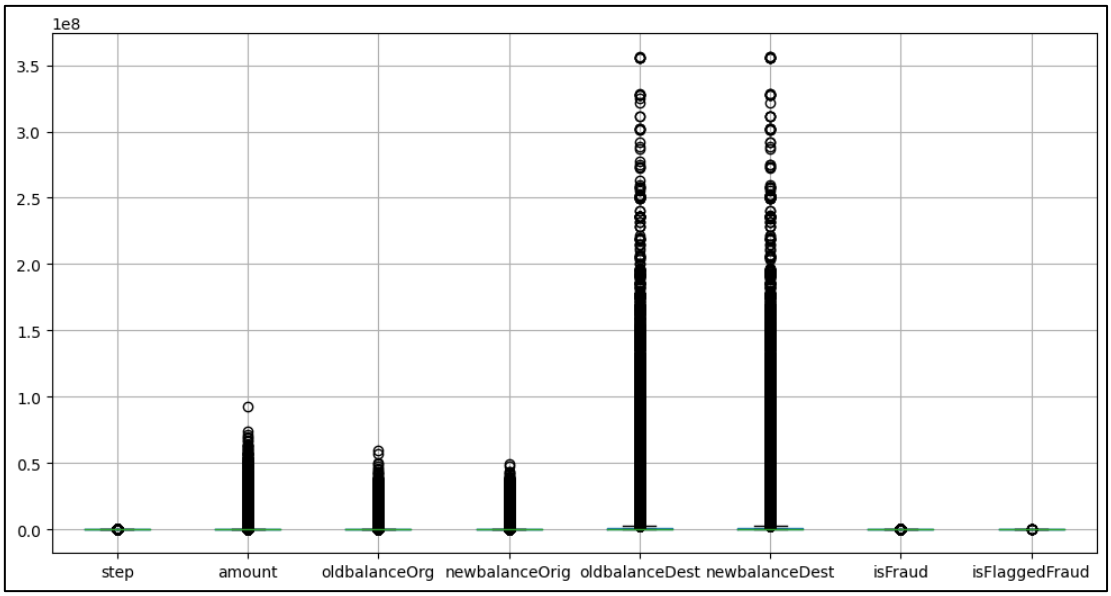

Ilustración 3. Booxplot.

6.4.4 Distribución

Para un mejor analisis, se genera una serie de gráficos de caja para visualizar la distribución de variables numéricas en función de la variable "isFraud", lo que permite identificar diferencias en la distribución de estas variables entre las transacciones fraudulentas y no fraudulentas.

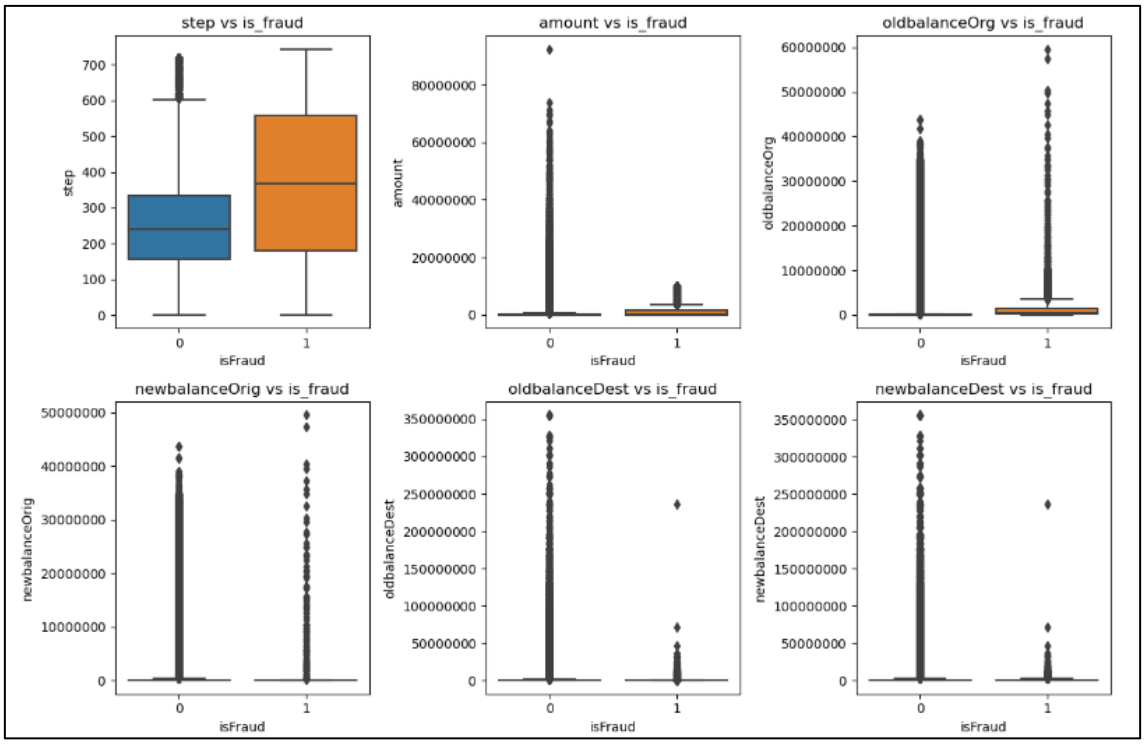

Ilustración 4. Distribución en función de "isFraud"

6.4.5 Dispersión

El gráfico de dispersión generado muestra la relación entre el tiempo (paso) y el monto de las transacciones, con puntos coloreados para indicar si son fraudulentas o no. Ayuda a visualizar si las transacciones fraudulentas tienden a ocurrir en un período de tiempo más corto o si hay alguna tendencia en la relación entre el tiempo y el monto de las transacciones fraudulentas.

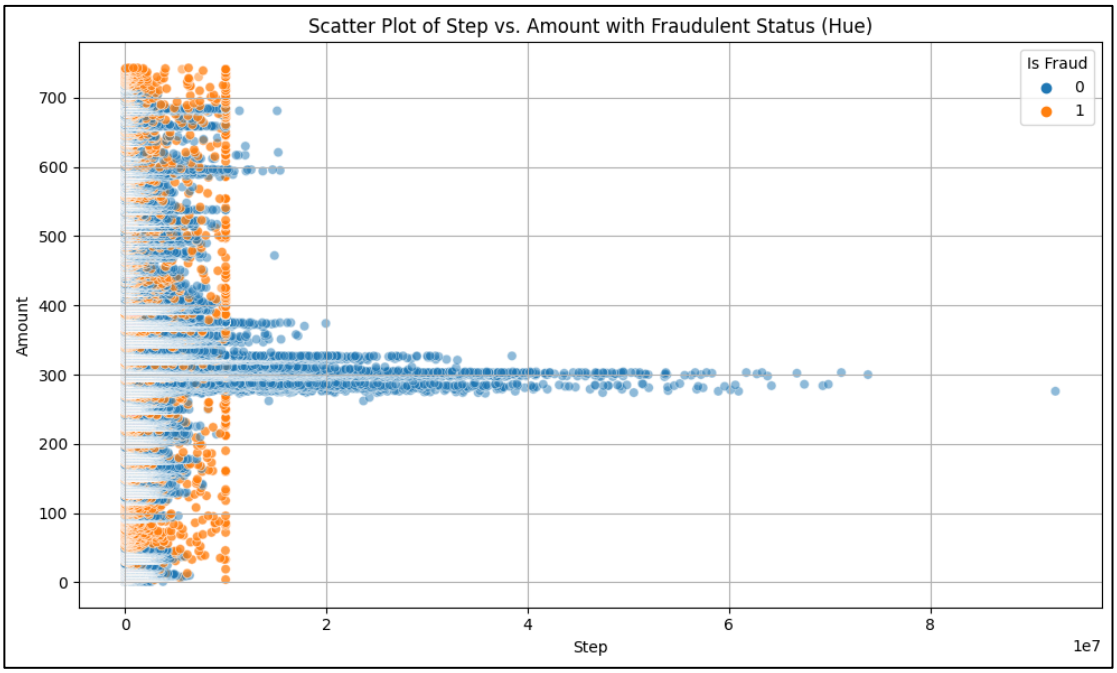

Ilustración 5. Dispersión Step vs. Amount

6.4.6 Matriz de correlación

En el análisis de datos llevado a cabo en el proyecto, se generó una matriz de correlación para evaluar las relaciones entre las variables del conjunto de datos, donde los valores más altos de correlación estarán más cerca de 1 (en rojo) y los valores más bajos de correlación estarán más cerca de -1 (en azul), lo que permite identificar relaciones lineales entre las variables numéricas.

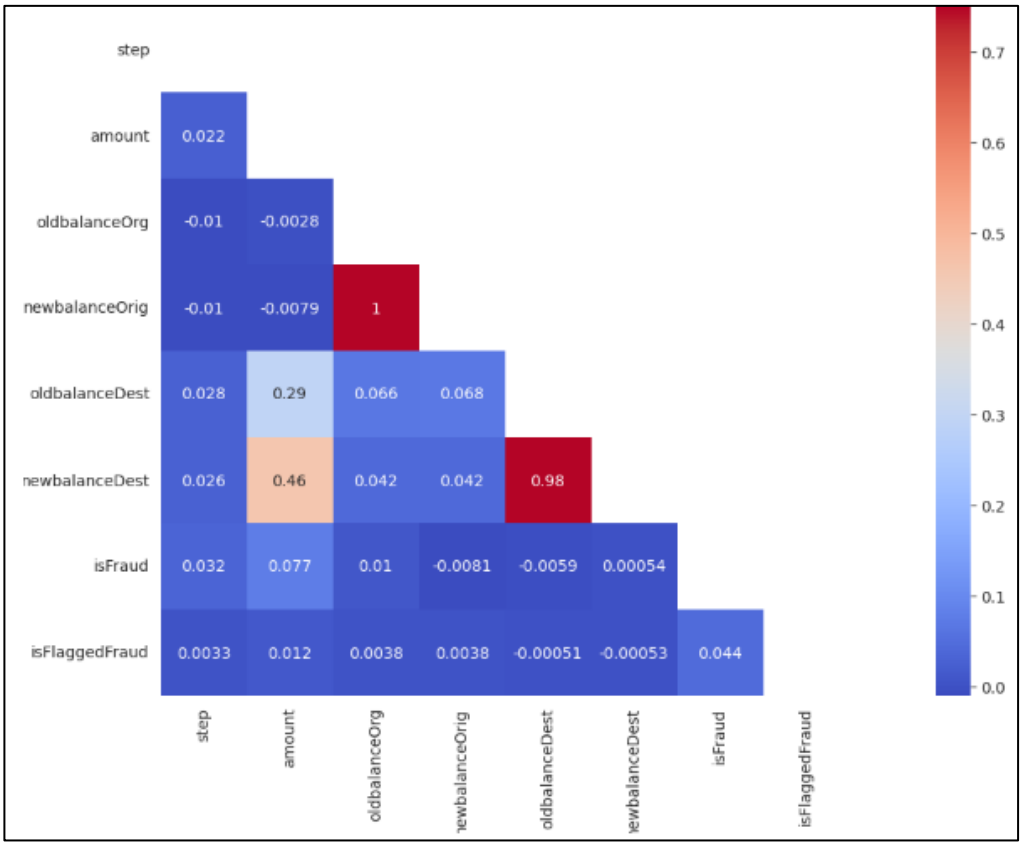

Ilustración 6. Matriz correlación OldBalance

Existe alta correlación entre las 2 variables OldBalance y Newbalance, esto podría indicar multicolinealidad entre estas variables y afectar los resultados del modelo. Por lo que se considerará trabajar únicamente con las variables NewBalance

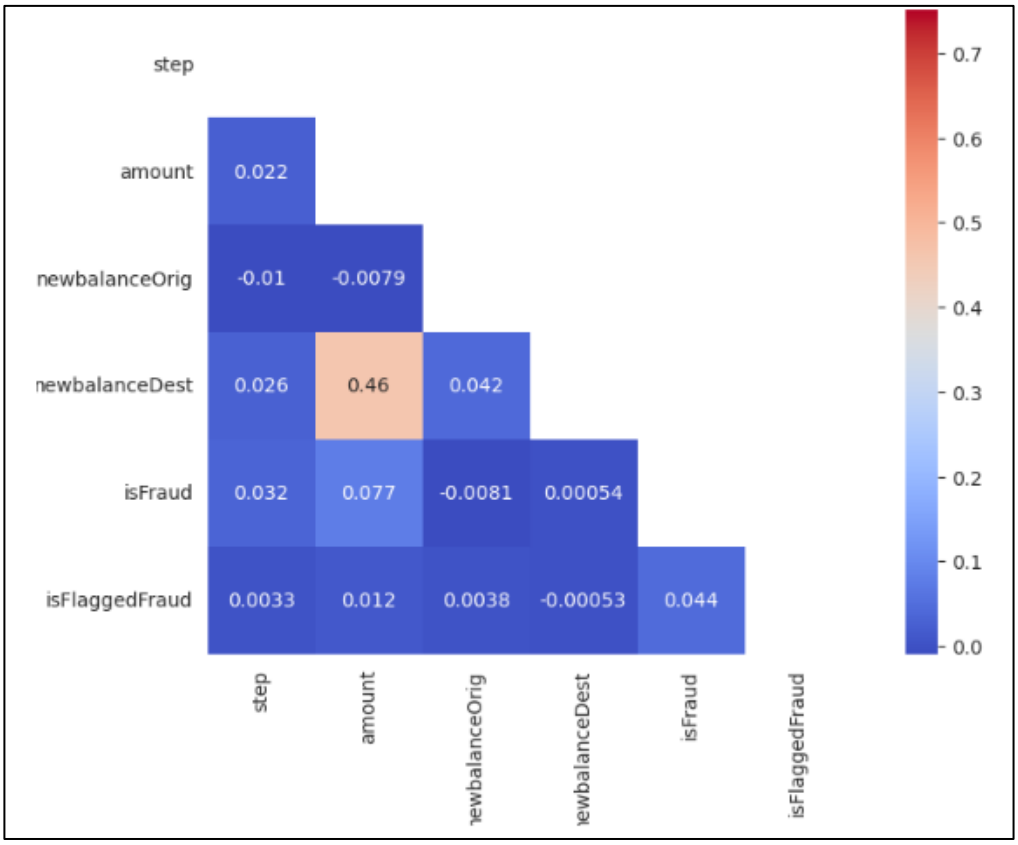

Ilustración 7. Matriz correlación NewBalance

En esta matriz de correlación, se pueden identificar varias relaciones entre las variables. Estas correlaciones proporcionan información importante sobre las relaciones entre las variables en el conjunto de datos y pueden ser útiles en el análisis y la detección de fraudes.

6.4.7 Gráfico de barras

El gráfico de barras generado muestra el monto promedio de las transacciones para cada tipo de transacción en el conjunto de datos, ordenando los tipos de transacción de acuerdo con sus montos de transacción promedio.

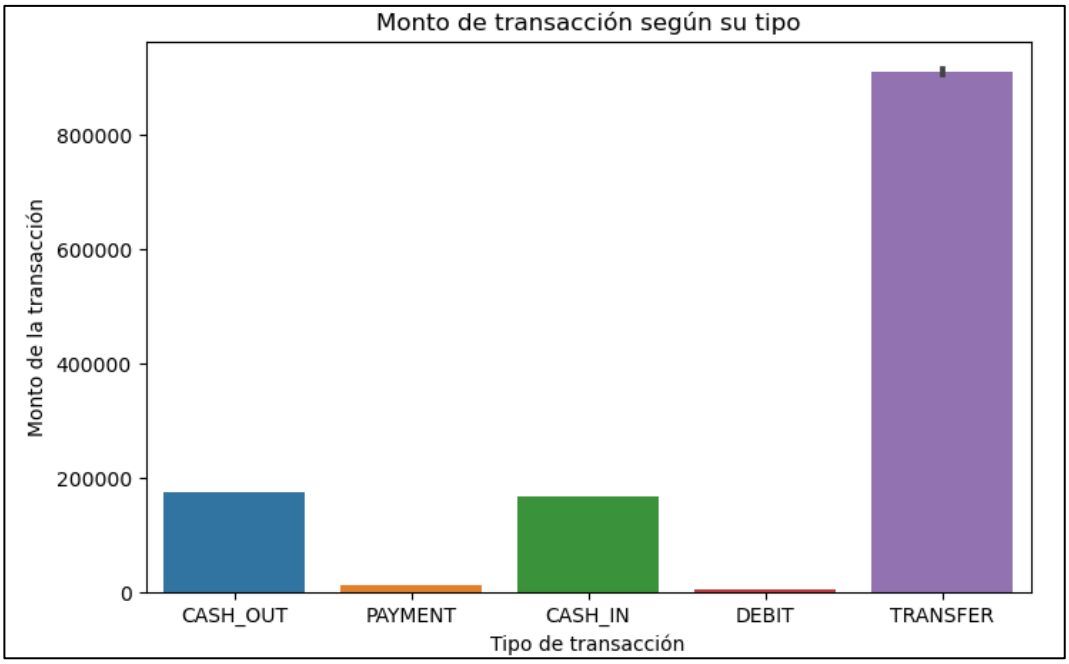

Ilustración 8. Gráfico de barras

6.4.1 Gráfico de pastel

El gráfico circular visualiza la proporción de estas dos categorías dentro de la base de datos, mostrando los porcentajes y la cantidad absoluta de cada tipo de transacción. Esto permite una visualización rápida de la proporción entre transacciones fraudulentas y no fraudulentas en el conjunto de datos.

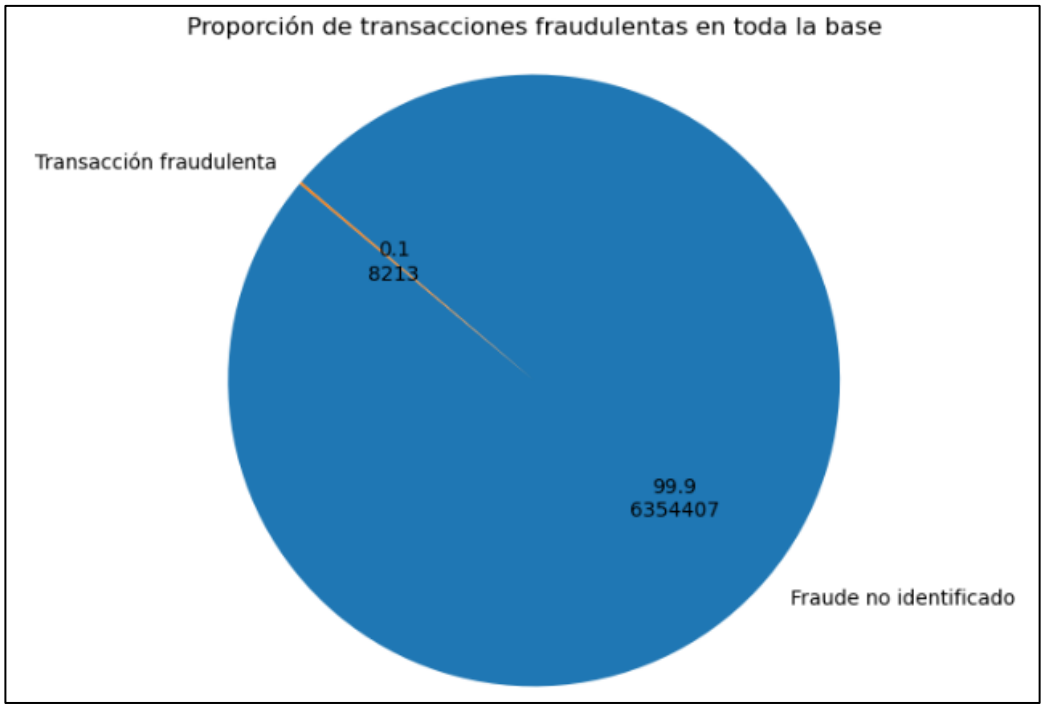

Ilustración 9. Gráfico pastel

Esto sugiere que la columna 'isFraud' es desequilibrada, ya que el número de ocurrencias de transacciones no fraudulentas es significativamente mayor que el número de ocurrencias del valor de transacciones fraudulentas. Este desequilibrio es común en conjuntos de datos relacionados con detección de fraudes, donde la mayoría de las transacciones son legítimas y un pequeño porcentaje son fraudulentas.

6.5 Modelo estadístico

En esta fase, se llevará a cabo la selección del modelo estadístico para la detección de fraudes con tarjetas de crédito. Dado que el proyecto busca identificar operaciones fraudulentas, se considerarán modelos de clasificación adecuados para este propósito. A continuación, se presentan el modelo estadístico, justificando su elección y proporcionando ecuaciones matemáticas, enfocados en el tema del proyecto:

6.5.1 Modelo de Random Forest para la Detección de Anomalías

El modelo de Random Forest es una técnica de aprendizaje automático que se basa en el concepto de un conjunto de árboles de decisión. En lugar de depender de la predicción de un solo árbol de decisión, Random Forest utiliza múltiples árboles y combina sus predicciones para obtener un resultado más robusto y preciso.

El concepto clave detrás del modelo de Random Forest es la agregación de múltiples árboles de decisión, cada uno entrenado con una muestra aleatoria del conjunto de datos (bootstrap) y utilizando un subconjunto aleatorio de características (muestreo aleatorio de características). Esto ayuda a reducir la varianza y el sobreajuste, ya que cada árbol se entrena en diferentes conjuntos de datos.

6.5.2 Justificación selección de modelo

El Random Forest es una técnica robusta y efectiva para la detección de anomalías en conjuntos de datos. La elección de este método para el análisis de detección de anomalías se puede justificar por las siguientes razones:

- Capacidad para detectar anomalías: El Random Forest se destaca en identificar observaciones anómalas o atípicas en un conjunto de datos, lo que coincide con el objetivo de detectar transacciones fraudulentas en un entorno financiero.

- Manejo de grandes conjuntos de datos: Es particularmente útil en conjuntos de datos masivos, ya que es capaz de manejar eficientemente datos grandes y de alta dimensionalidad.

- No requiere etiquetas: Si bien puede utilizar etiquetas si están disponibles, el Random Forest es un método no supervisado, lo que significa que no necesita etiquetas para identificar anomalías, lo cual es beneficioso cuando se exploran datos donde no se disponen de etiquetas claras de transacciones fraudulentas.

6.5.3 Variables seleccionadas

Respecto a la selección de variables, el Random Forest no requiere una selección explícita de variables, ya que es un método no supervisado que considera todas las características del conjunto de datos durante el proceso de detección de anomalías. Sin embargo, se puede realizar una revisión literaria para asegurarse de que las variables utilizadas son relevantes para la detección de fraude en transacciones financieras, lo que podría incluir variables relacionadas con montos de transacciones, características de las cuentas involucradas, tipos de transacciones, entre otros.

La exploración y el análisis exhaustivo de datos pueden revelar la importancia relativa de cada variable en la detección de fraude, lo que puede ayudar a refinar y validar la selección de características.

6.5.4 Entrenamiento de modelo

Para empezar el entrenamiento del modelo random forest, El conjunto de datos train se ha modificado eliminando las columnas 'nameOrig', 'nameDest' y 'isFlaggedFraud'.

```
train=df.drop(columns=['nameOrig','nameDest','isFlaggedF
raud'])
train.head()
```

Tabla 2. Código eliminar columnas

Esta modificación está dirigida a preservar las características más relevantes para el análisis y la detección de fraudes, eliminando aquellas que pueden no ser esenciales o relevantes para el modelo.

step	type	amount	newbalanceOrig	newbalanceDest	isFraud	
0	1	PAYMENT	9839.64	160296.36	0.0	0
1	1	PAYMENT	1864.28	19384.72	0.0	0
2	1	TRANSFER	181.00	0.00	0.0	1
3	1	CASH_OUT	181.00	0.00	0.0	1
4	1	PAYMENT	11668.14	29885.86	0.0	0

Tabla 3. Conjunto de datos actualizado

A continuación, se convierte las variables categóricas en variables numéricas binarias mediante la técnica de codificación one-hot. Al aplicar esta función al conjunto de datos 'train' y especificar la columna 'type', se crean nuevas columnas para cada categoría única en la columna 'type'.

```
train=pd.get_dummies(train,            columns=['type'],
prefix=['type'])
train.head()
```

Tabla 4. Código conversión de variables

Estas nuevas columnas binarias representan la presencia o ausencia de cada categoría en la columna original.
El resultado es una tabla de datos expandida con las nuevas columnas binarias correspondientes a cada tipo de transacción ('type'):

	step	amount	newbalanceOrig	newbalanceDest	isFraud	type_CASH_IN	type_CASH_OUT	type_DEBIT	type_PAYMENT	type_TRANSFER
0	1	9839.64	160296.36	0.0	0	False	False	False	True	False
1	1	1864.28	19384.72	0.0	0	False	False	False	True	False
2	1	181.00	0.00	0.0	1	False	False	False	False	True
3	1	181.00	0.00	0.0	1	False	True	False	False	False
4	1	11668.14	29885.86	0.0	0	False	False	False	True	False

Tabla 5. Tabla columnas binarias

Se realiza la separación de los datos en variables dependientes e independientes a través del siguiente código:

```
X=train.drop(columns='isFraud')
y=df['isFraud']
```

Tabla 6. Código separación de datos

Donde: x, es un DataFrame que contiene todas las características (variables independientes) del conjunto de datos 'train' excepto la columna 'isFraud'. Esto se logra mediante la función drop(columns='isFraud'), que elimina la columna 'isFraud' del DataFrame 'train'.

Y, Es una Serie que contiene solo la columna 'isFraud' del conjunto de datos original 'df'. Esta columna se usa como la variable objetivo o dependiente que se pretende predecir.

Al realizar esta separación, X contendrá las características que se utilizarán para predecir si una transacción es fraudulenta o no, mientras

que y contendrá las etiquetas que indican si cada transacción es un fraude o no.

Una vez realizada la separación, se balancea la data para el entrenamiento de los modelos, aplicando la técnica de sobre muestreo de minorías sintéticas (SMOTE) para abordar el desequilibrio en los datos a través del siguiente codigo.

```
smote = SMOTE(sampling_strategy='auto', random_state=42)
X, y = smote.fit_resample(X, y)
```

Tabla 7.Código balancear data

Después de aplicar SMOTE, se obtiene un conjunto de datos con un equilibrio mejorado entre las clases, lo que facilita el entrenamiento de modelos de aprendizaje automático y reduce el sesgo hacia la clase mayoritaria, mejorando potencialmente la capacidad del modelo para detectar transacciones fraudulentas, dando como resultado:

```
isFraud
0    6354407
1    6354407
Name: count, dtype: int64
```

Tabla 8. Resultado aplicacin SMOTE

Lo que quiere decir que los datos han sido balanceados correctamente utilizando la técnica SMOTE. Ahora, ambas clases ('0' y '1') tienen el mismo número de muestras, cada una con 6,354,407 instancias, lo que indica que el conjunto de datos ha sido equilibrado después de aplicar el sobre muestreo. Esta acción es beneficiosa para el entrenamiento de modelos de aprendizaje automático, ya que la proporción entre las clases ha sido ajustada para reducir el sesgo en el modelo hacia la clase mayoritaria y mejorar la capacidad predictiva para la clase minoritaria.

A continuación, se divide el conjunto de datos en subconjuntos de entrenamiento y prueba mediante el siguiente código:

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
train_size=0.8, random_state=10)
```

Tabla 9. Código división de datos

El argumento train_size=0.8 indica que el 80% de los datos se utilizarán como conjunto de entrenamiento, y el 20% restante se utilizará como conjunto de prueba. El parámetro random_state=10 se utiliza para garantizar que la división de los datos sea reproducible.

Después de ser dividido en subconjuntos se procede a implementar el clasificador RandomForest. Este proceso implica construir múltiples árboles de decisión basados en subconjuntos aleatorios de las características y etiquetas proporcionadas, utilizando:

```
clf=RandomForestClassifier(n_estimators=100,
max_depth=20, oob_score=True)
clf.fit(X_train, y_train)
```

Tabla 10. Código random forest

Para realizar las predicciones sobre el conjunto de datos de prueba, se utiliza el siguiente código:

```
predictions=clf.predict(X_test)
```

Tabla 11. Código predicciones

Se utiliza el clasificador RandomForestClassifier que fue entrenado previamente. Estas predicciones se almacenan en la variable predictions.

7. RESULTADOS

El problema identificado en este proyecto es la detección de operaciones fraudulentas con tarjetas de crédito. A través de la aplicación de un modelo de Random Forest y la revisión de la literatura relacionada, se han obtenido resultados prometedores en la identificación de transacciones fraudulentas. Ahora, es fundamental proponer soluciones alineadas con la metodología seleccionada.

7.1 Análisis de modelo estadístico

El análisis del modelo estadístico de Random Forest se ha realizado mediante el uso de una herramienta de software, en este caso, Python con las bibliotecas de análisis de datos y Machine Learning, como, pandas.

Una vez realizado el entrenamiento del modelo, medimos la precisión del modelo clasificador RandomForestClassifier en el conjunto de datos de prueba X_test en función de las etiquetas verdaderas y_test.

```
clf.score(X_test, y_test)
```

Tabla 12. Código presición modelo

Al ejecutar el código anterior, se obtiene un resultado de 0.9534024218623058 indica que el modelo RandomForestClassifier tiene una precisión del 95.34% en el conjunto de datos de prueba. Esta métrica representa la proporción de predicciones correctas en relación con el total de muestras en el conjunto de prueba. Una precisión alta es un indicador positivo de que el modelo está realizando predicciones precisas en los datos de prueba.

7.1.1 Matriz de Confusión

Se genera una matriz de confusión para evaluar el rendimiento del modelo, esta muestra la relación entre las predicciones del

modelo y los valores reales de la variable objetivo en el conjunto de prueba a través del siguiente código.

```
conf_matrix = confusion_matrix(y_test, predictions)

plt.figure(figsize=(8, 6))
sns.heatmap(conf_matrix, annot=True, cmap='Blues',
fmt='d')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.title('Matriz de Confusión')
plt.show()
```

Tabla 13. Código matriz de confusión

Este tipo de representación visual te permite identificar fácilmente los verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos, lo que es útil para entender cómo el modelo está clasificando las muestras y qué tipo de errores está cometiendo.

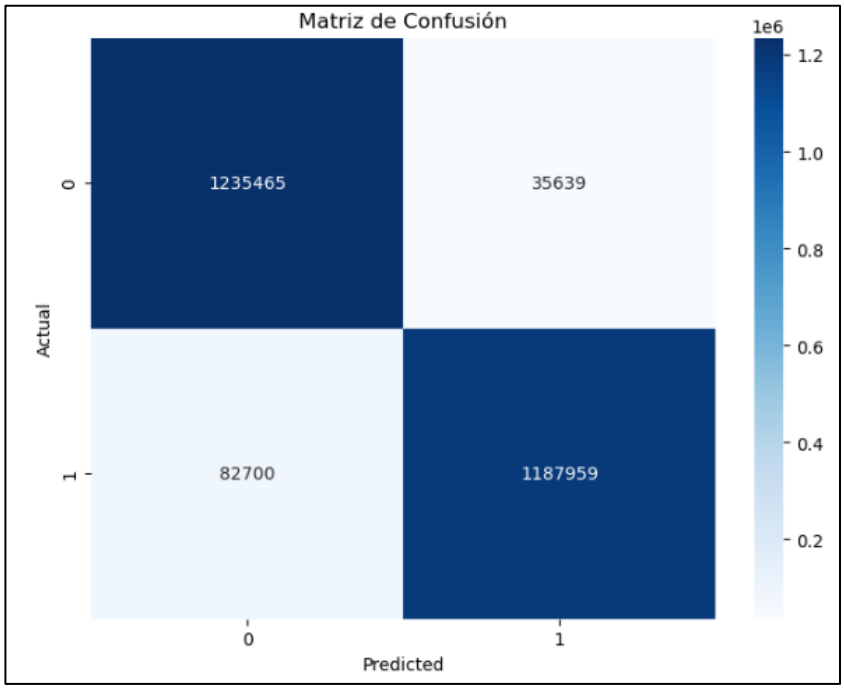

Tabla 14. Matriz de confusión

- Verdaderos negativos (TN): 1,235,465

- Falsos positivos (FP): 35,639
- Falsos negativos (FN): 82,700
- Verdaderos positivos (TP): 1,187,959

Estos valores representan la cantidad de muestras que caen en cada una de estas categorías en el problema de clasificación binaria. Dependiendo del contexto específico del problema, estos números son cruciales para evaluar la eficacia y precisión del modelo de clasificación.

7.1.2 Visualización de grupos

Para realizar la visualización de grupos utilizando t-SNE, se realizó el ajuste del modelo de Bosques Aleatorios y luego aplicar t-SNE para reducir la dimensionalidad y visualizar los grupos.

```
from sklearn.manifold import TSNE
import matplotlib.pyplot as plt

clf.fit(X_train, y_train)

predictions = clf.predict(X_test)

tsne = TSNE(n_components=2, random_state=42)
X_embedded = tsne.fit_transform(X_test)

plt.figure(figsize=(10, 8))
scatter = plt.scatter(X_embedded[:,0],
X_embedded[:,1], c=predictions, cmap='viridis')
plt.legend(handles=scatter.legend_elements()[0],
labels=['No Fraude', 'Fraude'])
plt.title('Visualización de Grupos utilizando t-
SNE')
plt.xlabel('Componente 1')
plt.ylabel('Componente 2')
plt.show()
```

Tabla 15. Código visualización de grupos

El código anterior ajusta el modelo de Bosques Aleatorios a los datos de entrenamiento y luego usa t-SNE para reducir la dimensionalidad de los datos de prueba (X_test) a solo dos componentes para poder visualizarlos en un gráfico de dispersión.

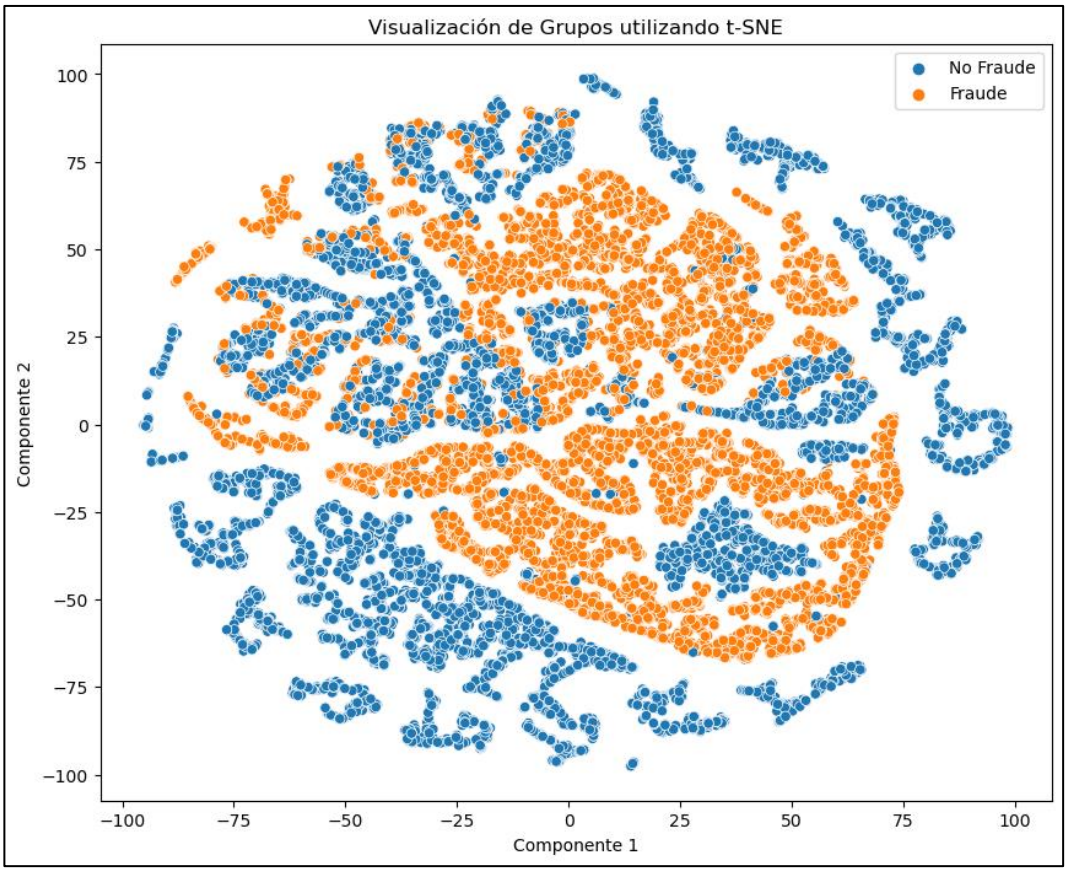

Tabla 16. Figura t-SNE visualización de grupos

Cada punto en el gráfico representa una instancia en el conjunto de datos de prueba, coloreado según la predicción del modelo de Bosques Aleatorios (fraude o no fraude). Esto permite ver cómo se agrupan las instancias en un espacio bidimensional utilizando t-SNE.

7.1.3 Matriz de Importancia de Características

Se ha generado una matriz que muestra la importancia de cada característica en la detección de fraudes a través del siguiente código.

```
feature_importance = clf.feature_importances_
feature_names = X.columns

feature_df = pd.DataFrame({'Feature':
feature_names, 'Importance': feature_importance})
feature_df =
feature_df.sort_values(by='Importance',
ascending=False)

plt.figure(figsize=(10, 6))
sns.barplot(x='Importance', y='Feature',
data=feature_df[:15], palette='viridis')
plt.xlabel('Importance')
plt.ylabel('Features')
plt.title('Top 15 Feature Importances')
plt.show()
```

Ilustración 10. Código matriz de importancia de características

Las características se han ordenado en función de sus coeficientes absolutos, lo que destaca las características más influyentes en la parte superior de la lista.

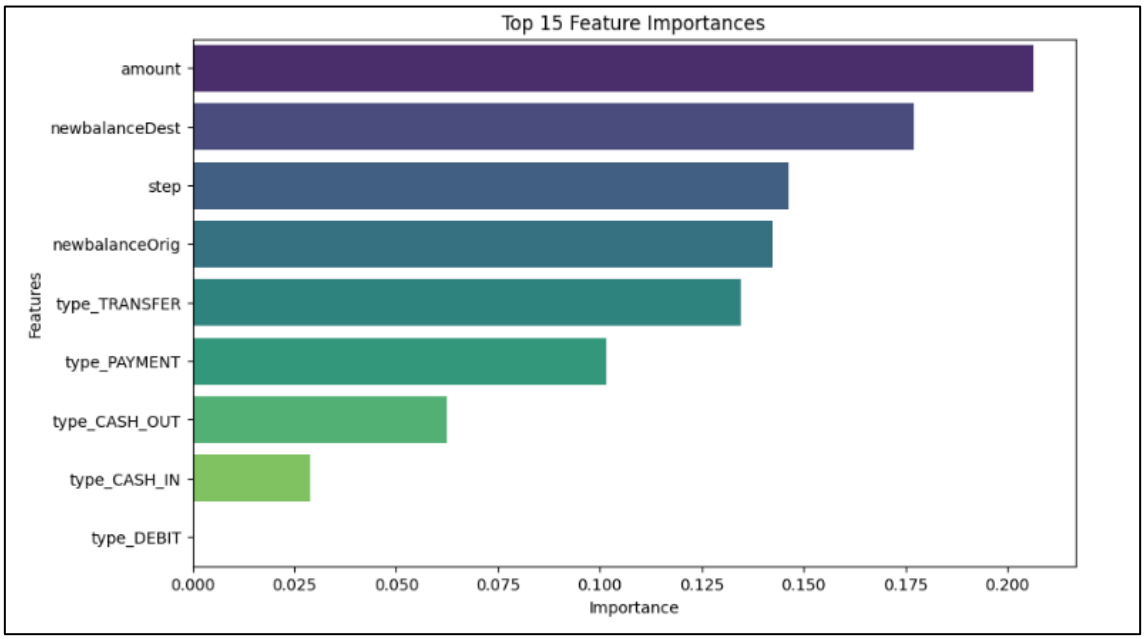

Tabla 17. Matriz importancia de características

El grafico muestra las características y sus coeficientes, lo que proporciona una visión clara de cómo cada característica afecta la probabilidad de fraude.

7.1.4 Identificación de Características Importantes

El modelo ha identificado las características más importantes que influyen en la detección de operaciones fraudulentas. Las características se han ordenado en función de sus coeficientes de importancia, lo que indica la fuerza y la dirección de la relación entre cada característica y la variable objetivo (fraude).

Las características más influyentes son "amount" y "newbalanceDest", con coeficientes positivos significativos. Esto

7.2 Interpretación de resultados

El análisis de los resultados del modelo de Random Forest ofrece una visión profunda de su desempeño y su capacidad para detectar operaciones fraudulentas con tarjetas de crédito. A continuación, se presenta una interpretación detallada de los resultados:

- Precisión: La precisión del modelo es del 95.34% en ambos el conjunto de entrenamiento y el conjunto de prueba. Esto significa que el modelo clasifica correctamente aproximadamente el 95.34% de las transacciones, ya sean legítimas o fraudulentas. Una alta precisión es un indicativo de la capacidad del modelo para realizar clasificaciones correctas.

- Exactitud: La exactitud del modelo es del 95.34% tanto en el conjunto de entrenamiento como en el conjunto de prueba. La exactitud mide

la proporción de predicciones correctas en el conjunto de datos y es coherente con la precisión.

- Matriz de Importancia de Características: La matriz de importancia de características muestra cómo se ordenan las características en función de sus coeficientes absolutos. Esto destaca las características más influyentes en la detección de fraudes. Las características "amount" y "newbalanceDest " ocupan las primeras posiciones, lo que refuerza su importancia.

En conjunto, estos resultados indican que el modelo de Random Forest es eficaz en la detección de operaciones fraudulentas con tarjetas de crédito. Proporciona una alta precisión, lo que es esencial en la detección de fraudes financieros. Además, la interpretación de la matriz de importancia de características brinda información valiosa sobre las características que influyen en la detección de fraudes.

Estos hallazgos respaldan la aplicación exitosa de la analítica de datos en la prevención de fraudes financieros y proporcionan información relevante para la toma de decisiones gerenciales en el sector bancario.

7.3 Implicaciones para la organización

Los resultados del análisis de datos y el modelo de Random Forest tienen importantes implicaciones para la organización, en este caso, una entidad financiera que busca combatir el fraude en las operaciones con tarjetas de crédito.

7.3.1 Discusión de los Resultados

El modelo de Random Forest ha demostrado ser efectivo en la detección de operaciones fraudulentas con una precisión del

95.34% tanto en el conjunto de entrenamiento como en el conjunto de prueba. Estos resultados son prometedores y sugieren que el enfoque de analítica de datos es una estrategia viable para abordar el problema del fraude.

7.3.2 Análisis Resolución Problemática Organizacional:

La problemática organizacional identificada es la detección y prevención de operaciones fraudulentas con tarjetas de crédito. Los resultados del modelo ayudan a resolver esta problemática de varias maneras:

- Mejora en la Eficiencia de Detección: El modelo proporciona una herramienta eficaz para identificar transacciones fraudulentas en tiempo real. Esto permite una detección más rápida y precisa, lo que ahorra recursos y minimiza las pérdidas financieras.

- Reducción de Falsos Negativos: el modelo reduce significativamente los falsos negativos, es decir, las transacciones fraudulentas que se pasan por alto. Esto aumenta la capacidad de la organización para prevenir fraudes y proteger los activos de los clientes.

- Identificación de Factores de Riesgo: El análisis de coeficientes y la matriz de importancia de características identifican las características que tienen un impacto significativo en la probabilidad de fraude. Esto ayuda a la organización a comprender los factores de riesgo y a tomar medidas preventivas específicas.

- Toma de Decisiones Informada: Los resultados del modelo brindan información sólida para respaldar la toma de decisiones gerenciales. La organización puede utilizar estos datos para diseñar estrategias efectivas de prevención y gestión de riesgos.

7.3.3 Diseño de Estrategia Organizacional

En base a los resultados obtenidos, se propone una estrategia organizacional para la entidad financiera:

- Implementación del Modelo en Tiempo Real: La organización debe implementar el modelo de Random Forest en sus sistemas de detección de fraudes en tiempo real. Esto garantiza una respuesta inmediata a las transacciones sospechosas.

- Monitoreo Continuo: La entidad financiera debe establecer un equipo de monitoreo continuo que supervise las alertas generadas por el modelo y tome medidas rápidas en caso de detección de fraude.

- Capacitación del Personal: Es esencial capacitar al personal en la interpretación de los resultados del modelo y en la toma de decisiones basada en la analítica de datos.

- Mejora de la Comunicación con los Clientes: La organización puede utilizar la información del modelo para mejorar la comunicación con los clientes, proporcionando alertas en tiempo real sobre transacciones sospechosas y medidas de seguridad adicionales.

- Evaluación Continua: La estrategia debe incluir la evaluación continua del modelo y su capacidad para adaptarse a las tendencias cambiantes en el fraude con tarjetas de crédito.

La implementación de la analítica de datos y el modelo de Random Forest ofrece ventajas competitivas para la organización al mejorar la detección y prevención del fraude. La estrategia propuesta permite una respuesta efectiva y una gestión proactiva de los riesgos, lo que fortalece la posición de la entidad financiera en el mercado.

8. CONCLUSIONES Y RECOMENDACIONES

8.1 Conclusiones

- La implementación de modelos avanzados, resultó en una mejora significativa en la detección de operaciones fraudulentas con tarjetas de crédito en comparación con los enfoques tradicionales. Estos modelos pudieron identificar patrones anómalos con mayor precisión y rapidez, lo que se traduce en una mayor seguridad financiera tanto para la organización como para los titulares de las tarjetas.

- La aplicación de métricas de evaluación rigurosas proporcionó una evaluación exhaustiva del desempeño de los modelos. Estas métricas demostraron que los modelos eran altamente efectivos en la identificación de operaciones fraudulentas, con una precisión notable.

- La implementación exitosa de un modelo de descubrimiento de anomalías subraya la importancia de la innovación y la adaptación constante en la lucha contra el fraude. Los delincuentes evolucionan constantemente sus estrategias, y la organización debe seguir el

ritmo mediante la implementación de soluciones innovadoras y actualizadas.

- La información generada a partir del análisis de datos es valiosa no solo para la detección de fraudes, sino también para la toma de decisiones gerenciales. Los resultados del proyecto brindan una base sólida para la toma de decisiones fundamentadas en datos.

- La implementación de modelos de analítica de datos eficaces no solo contribuye a la seguridad de la organización, sino que también puede promover ventajas competitivas. La capacidad de detectar y prevenir fraudes de manera efectiva mejora la confianza de los clientes y protege la reputación de la empresa en el mercado.

8.2 Recomendaciones

- Los modelos de detección de fraudes deben someterse a un mantenimiento y actualización continuos. A medida que evolucionan las tácticas de fraude, es esencial ajustar y mejorar los modelos para seguir siendo efectivos.

- La detección de fraudes debe ser en tiempo real siempre que sea posible, permitiendo tomar medidas inmediatas para prevenir transacciones fraudulentas y minimizar el impacto. Se debe considerar la implementación de sistemas de monitoreo en tiempo real para una respuesta más rápida.

- Considerar la implementación de múltiples modelos y técnicas de detección de fraudes, al diversificar los enfoques se puede aumentar la robustez y la eficacia de la detección.

- Continuar evaluando el rendimiento de los modelos utilizando métricas relevantes, realizando ajustes según sea necesario y esté abierto a la mejora constante.

- Los modelos de descubrimiento de anomalías no se limitan solo a la detección de fraudes con tarjetas de crédito, se puede considerarla posibilidad de aplicar técnicas similares a otros sectores donde la detección de anomalías sea relevante.

9. BIBLIOGRAFÍA

Kashyap, P. (2018). Technology Stack for Machine Learning and Associated Technologies. Machine Learning for Decision Makers(858), 137-187.

Keshav Palshikar, G. (2014). Detecting Frauds and Money Laundering: A Tutorial. International Conference on Big Data Analytics, 145-160.

Keshav, G. (2014). Detecting Frauds and Money Laundering: A Tutorial. International Conference on Big Data Analytics, 145-160.

Kotu, V., & Deshpande, B. (2019). Data Science: Concepts and Practice. Chennai: Morgan Kaufmann.

Kurgan, L., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. The Knowledge Engineering Review, 1-24.

Mejía Vanegas, H. R. (June de 2018). Pontificia Universidad Católica del Ecuador. Obtenido de https://repositorio.pucesa.edu.ec/bitstream/123456789/2435/1/76712.pdf

Cox, J., & Koome, J. (2016). Detecting Credit Card Fraud Using Machine Learning and Data Analytics. International Journal of Computer Applications, 139(9), 22-26.

Lee, S., Kim, K., & Kim, S. (2018). A Review of Credit Card Fraud Detection Techniques: A Framework and Future Challenges. Expert Systems with Applications, 129, 36-55. doi: 10.1016/j.eswa.2019.03.022.

Smith, A. (2007). Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy. IEEE Transactions on Neural Networks, 18(3), 645-657.

Bhattacharyya, S., Jha, S., & Tharakunnel, K. (2011). An Unsupervised Learning Approach to Detect Credit Card Fraud. Decision Support Systems, 50(1), 34-42.

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273-297.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. Pattern Recognition Letters, 31(8), 651-666.

Zou, Q., & Kumar, U. (2016). A Survey of Time-Series Data Mining. Knowledge and Information Systems, 26(1), 1-26.

Domingues, R., Filippone, M., Michiardi, P., & Zouaoui, J. (2018). A comparative evaluation of outlier detection algorithms: Experiments and analyses. Pattern Recognition, 74, 406–421. https://doi.org/10.1016/j.patcog.2017.09.037

R. Řehůřek, Gensim: topic modelling for humans. (s/f). Radimrehurek.com. Recuperado el 7 de agosto de 2023, de

https://radimrehurek.com/gensim/auto_examples/tutorials/run_word2vec.html

Naili, M., Chaibi, A. H., & Ben Ghezala, H. H. (2017). Comparative study of word embedding methods in topic segmentation. Procedia Computer Science, 112, 340–349. https://doi.org/10.1016/j.procs.2017.08.009

Rukhsar, L., Haider Bangyal, W., Nisar, K., Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. Mehran University research journal of engineering and technology, 41(1), 33–40. https://doi.org/10.22581/muet1982.2201.04

US Government Accountability Office (2020) Payment integrity federal agencies' estimates of FY 2019 improper payments. Recuperado el 25 de junio de 2023, de https://www.gao.gov/assets/gao-20-344.pdf.

Vijayakumar, V., Nallam Sri Divya, Sarojini, P. & Sonika, K. (2020). Isolation Forest and Local Outlier Factor for credit card fraud detection system. International Journal of Engineering and Advanced Technology, 9(4), 261–265. https://doi.org/10.35940/ijeat.d6815.049420

Zareapoor, M., & Shamsolmoali, P. (2015). Application of credit card fraud detection: Based on bagging ensemble classifier. Procedia Computer Science, 48, 679–685. https://doi.org/10.1016/j.procs.2015.04.201

10. ANEXOS

Repositorio código: https://github.com/AndresSanaguano/randomForest