# Detección de fraude bancario en Colombia mediante el análisis de grafos

Presentado para obtener el título de

**MAGISTER EN MATEMÁTICA APLICADA Y CIENCIAS DE LA COMPUTACIÓN**

Brayan Steven Calderon Adames

Dirección:

Juan Felipe Romero

Universidad del Rosario

Escuela de Ingeniería, Ciencia y Tecnología

Maestría en matemática aplicada y ciencias de la computación

## AGRADECIMIENTOS

Quisiera expresar mi más sincero agradecimiento a mi familia y colegas, cuyo apoyo incondicional ha sido fundamental a lo largo de mi maestría. Asimismo, estoy profundamente agradecido con mi director de tesis por el invaluable soporte y motivación brindados en este importante proyecto. Un agradecimiento especial a Kevin Beltrán, cuya asistencia ha sido crucial para la culminación exitosa de este trabajo. Sus contribuciones no solo han sido clave para alcanzar este logro, sino que también han enriquecido enormemente mi experiencia académica. Gracias a todos por ser parte de este significativo viaje.

## RESUMEN

Este proyecto se enfoca en desarrollar un sistema de puntuación de riesgo para los empleados de una entidad financiera, con el objetivo de mitigar el fraude interno. Para ello, se han implementado técnicas avanzadas de grafos, las cuales han demostrado ser cruciales en la identificación de relaciones complejas entre empleados y clientes. Estos grafos han sido fundamentales para capturar información vital y consistente, permitiendo la detección eficaz de anomalías en las interacciones entre estas partes.

Además, se ha integrado el uso de modelos de Machine Learning en el proyecto, lo que ha facilitado la creación de algoritmos predictivos. Estos modelos ofrecen la capacidad de prever posibles incidentes de fraude interno, lo que a su vez permite tomar medidas proactivas en la mitigación de riesgos. En resumen, la aplicación de estas metodologías computacionales ha resultado ser extremadamente valiosa, no solo para establecer controles de primera línea eficientes, sino también para desarrollar sistemas predictivos capaces de identificar potenciales defraudadores dentro de la organización financiera.

Palabras clave: Fraude, riesgo, machine learning, grafos, patrones

## A**BSTRACT**

This project focuses on developing a risk scoring system for employees of a financial entity, aimed at mitigating internal fraud. To achieve this, advanced graph techniques have been implemented, proving to be crucial in identifying complex relationships between employees and clients. These graphs have been fundamental in capturing vital and consistent information, enabling effective detection of anomalies in interactions between these parties.
Furthermore, the integration of Machine Learning models into the project has facilitated the creation of predictive algorithms. These models provide the capability to foresee potential internal fraud incidents, thereby allowing for proactive risk mitigation measures. In summary, the application of these computational methodologies has proven to be extremely valuable, not only in establishing efficient frontline controls but also in developing predictive systems capable of identifying potential fraudsters within the financial organization.

Keywords: Fraud, risk, machine learning, graphs, patterns

# TABLA DE CONTENIDO

# LISTA DE TABLAS

# Capítulo 1
# INTRODUCCIÓN

En el dinámico mundo de las ciencias de datos, el análisis de redes se ha establecido como un área critica de investigación, especialmente en los entornos organizacionales donde las relaciones y las conexiones de los datos son muy importantes, ya que estas tienen un gran impacto en el rendimiento y la eficiencia. Este proyecto, situado en la intersección de la ciencia de las redes, el aprendizaje automático y el análisis organizacional, explora la estructura y las dinámicas de las redes de empleados mediante el uso de grafos transaccionales.

El análisis de redes, en su esencia, se ocupa de como los individuos o entidades dentro de un sistema estén interconectados, estas conexiones pueden tomar muchas formas, desde comunicaciones simples hasta transacciones complejas. Estos vínculos en una organización pueden ser muy variados y reflejar tanto las estructuras de reporte y colaboración formales como las relaciones informales, el flujo de información y las dinámicas de influencia. Al representar a los empleados como nodos y sus interacciones como aristas en un grafo transaccional, se puede comenzar a visualizar y cuantificar estas redes complejas.

Este proyecto no se limita a la visualización de la red; también analiza cuantitativamente sus características. Los nodos (empleados) que cumplen funciones importantes en una red se identifican utilizando métricas de análisis de redes como la centralidad de grado, la cercanía y la intermediación. Estos nodos altamente conectados o centralmente ubicados con frecuencia actúan como importantes centros de comunicación o influencia; por lo tanto, es esencial identificarlos para comprender cómo se difunde la información dentro de la organización y cómo se pueden optimizar los procesos internos.

La extracción de incrustaciones de los nodos es un paso más allá en el análisis. Esta estrategia reduce la complejidad de la red al convertir las estructuras de red multidimensional en un espacio de características más fácil de manejar. Este paso es crucial para aplicar técnicas de aprendizaje automático a los datos de la red porque convierte la información de la red, que es intrínsecamente no euclidiana, en un formato adecuado para el modelado predictivo.

Varios modelos de aprendizaje automático se utilizan para explorar y predecir patrones en la red después de transformar la información de la red. Por otro lado, la regresión logística multiclase proporciona un modelo interpretable que puede identificar relaciones lineales entre clases y características. Sin embargo, los árboles de decisión ofrecen una mayor flexibilidad porque pueden modelar relaciones no lineales sin transformar datos extensamente. Por último, pero no menos importante, XGBoost, que es conocido por su rendimiento superior en una variedad de tareas de aprendizaje automático, se utiliza para manejar la complejidad inherente de los datos y mejorar la precisión de las predicciones.
Los modelos arrojaron resultados que ofrecen valiosas perspectivas sobre la red. Las métricas de rendimiento, tales como precisión, Recall y F1-score, proporcionan información sobre la efectividad de los modelos para clasificar y predecir roles y patrones en la red. También, estos modelos pueden ser útiles para identificar los factores principales que afectan a la creación de conexiones y comunidades dentro de la empresa, lo cual es fundamental para la toma de decisiones y la gestión estratégica.

# Capítulo 2
# OBJETIVOS

## 1.1 Objetivo general:

Generar un score de riesgo a los funcionarios que realizan colocación por los distintos canales habilitados por la entidad financiera con el fin de identificar anomalías para mitigar el fraude en la colocación, deterioro del portafolio y comerciales con bajo rendimiento.

## 1.2 Objetivos específicos:

1. Construir una red que sirva para identificar los productos asociados a cada cliente y como puede existir una relación entre más productos.

2. Construir una red de empleados y productos que sirva para identificar la relación entre los productos, clientes y empleado que ayudo a la colocación de dicho producto.

3. Detectar anomalías en el comportamiento de la colocación de los empleados en los productos de los clientes.

4. Construir un modelo de detección de anomalías usando como covariables la información de fraude y del grafo de empleados para poder predecir el fraude interno y predecir aquellos empleados que pueden tener un alto nivel de riesgo.

# Capítulo 3

# PROBLEMA Y JUSTIFICACIÓN

El fraude financiero en Colombia existe y es una realidad que las empresas están tratando de mitigar, entonces; ¿Qué se está haciendo para mejorar las cifras de fraude financiero en Colombia? ¿Qué metodologías se están usando para la detección del fraude financiero? ¿Cuántas perdidas monetarias se pueden mitigar al detectar el fraude financiero? Estas son las preguntas que se plantearon para poder iniciar con la investigación y finalmente preguntarse si al implementar metodologías más sofisticadas y que vayan orientadas al comportamiento particular de cada uno de los clientes de una entidad bancaria se puede evitar el fraude financiero y así mejorar el revenue de la organización.

**Justificación:**

Ya existen metodologías para la detección de fraude financiero las cuales son usados en la exploración de los datos y así encontrar puntos alejados o extraños al comportamiento natural de ellos ¿qué pasa con esta metodología? Muchas no reflejan comportamientos estacionales de los clientes ni la realidad diaria, así que esto no tiene una confiabilidad demasiado alta para una detección temprana del fraude, ni tampoco tiene la oportunidad de adaptarse ya que las metodologías de fraude interna son muy poco comunes y el fraude en este sector no es algo que se visualice a grandes rasgos.

Según encuestas hechas por compañías como PWC sobre el fraude financiero en Colombia las compañías tienen pleno conocimiento sobre el riesgo y los costos asociados a esta problemática.


**Figura 1.** Encuesta de crimen y fraude económico de PWC

# Capítulo 4

# ESTADO DEL ARTE

1. **“Internal auditing and fraud**” por the institute of internal Auditors (2009) [1]:

El fraude abarca una amplia gama de actos ilegales que involucran engaño o tergiversación intencional, según la definición del Instituto de Auditores Internos (IIA) IPPF. Este acto no depende de amenazas de violencia y puede ser perpetrado por individuos u organizaciones con el objetivo de obtener dinero, propiedades o servicios, evitar pagos, o asegurar una ventaja personal o comercial. Otra definición enfatiza que el fraude implica actos u omisiones diseñados para engañar a otros, resultando en pérdidas para la víctima y ganancias para el perpetrador. Es importante destacar que el término "fraude" puede tener connotaciones legales y también puede referirse a acciones específicas consideradas como corrupción.

2. **“Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019”** por Khaled Gubran Al-Hashedi, Pritheega Magalingam (2021) [2]:

El texto aborda la revolución tecnológica en Internet y destaca la influencia del comercio electrónico y la transferencia de dinero. Se mencionan tecnologías y modelos como la minería de datos, con enfoques estadísticos, de aprendizaje automático y de inteligencia artificial, utilizados para detectar fraudes financieros. Entre los modelos específicos se mencionan Naive Bayes, máquinas de soporte vectorial (SVM), y regresión logística, empleados en la identificación de actividades anómalas en transacciones pasadas.

Se enfatiza que estas técnicas enfrentan desafíos debido a la constante evolución de nuevos métodos por parte de estafadores, así como el impacto de tecnologías emergentes como la criptomoneda. La revisión abarca estudios sobre costos anuales de fraude financiero en EE. UU. y el Reino Unido, evidenciando pérdidas significativas. Además, se subraya la importancia de la revisión integral, que cubre áreas como tipos de fraude, pros y contras de las técnicas de minería de datos, conjuntos de datos y métricas de evaluación.

3. **“Fraud detection system_ A survey”** por Abdallah, Zainal (2016) [3]:

En este texto, se describe un enfoque de detección de anomalías utilizado por Sistemas de Detección de Fraudes (FDS), que se basa en métodos de perfilado conductual. Este método modela los patrones de comportamiento de individuos y monitorea desviaciones de la norma para detectar posibles fraudes. Los FDS basados en anomalías son utilizados en diversas áreas de fraude y tienen la capacidad de detectar fraudes novedosos. El texto también menciona la clasificación de los métodos de detección de anomalías en tres tipos: no supervisada, semisupervisada y supervisada.

En cuanto al aprendizaje supervisado, se destaca que implica el uso de un conjunto de datos etiquetado como "fraude" y "no fraude" para entrenar un clasificador. Aunque es el enfoque más común, tiene limitaciones como la dificultad de recopilar etiquetas en grandes conjuntos de datos y la ambigüedad en las etiquetas. Se mencionan varios algoritmos de aprendizaje supervisado, como redes neuronales artificiales, vecinos más cercanos, árboles,

regresión logística, Naive–Bayes y máquinas de vectores de soporte (SVM). Además, se señala que el aprendizaje no supervisado y semisupervisado se utilizan para superar las limitaciones del aprendizaje supervisado en algunos casos.

4. **"Support Vector Machine"** por Alka Rani, Nishant K. Sinha (2022) [4]:

El texto destaca que SVM (Support Vector Machine) es un algoritmo de aprendizaje automático supervisado utilizado para clasificación y regresión. Se deriva de la teoría estadística del aprendizaje. La idea fundamental de las SVM es encontrar el hiperplano adecuado para la clasificación, utilizando vectores de soporte que representan puntos de cada clase presentes en el margen.

5. **"Chapter 12 - Learning in Big Data: Introduction to Machine Learning"** por Khadija El Bouchefry PhD, Rafael S. de Souza PhD (2020) [5]:

El Bosque Aleatorio (RF), es un enfoque de clasificación en conjunto que aprovecha la votación mayoritaria de múltiples árboles de decisión para realizar predicciones de clases. En este método, se construyen varios árboles al seleccionar aleatoriamente un conjunto predefinido de variables para realizar divisiones en cada nodo, utilizando el método de bagging. Bagging genera conjuntos de datos de entrenamiento para cada árbol mediante muestreo con reemplazo, donde el número de ejemplos seleccionados es igual al número de ejemplos. RF implementa el índice Gini para determinar los umbrales óptimos de división de los valores de entrada para clases específicas. El índice Gini proporciona una medida de la heterogeneidad de clases dentro de los nodos hijos en comparación con el nodo padre.

6. **"Explainable fraud detection of financial statement data driven by two-layer knowledge graph"** por Siqi Cai**,** Zhenping Xie (2024) [6]:

El fraude en los estados financieros, caracterizado por la manipulación ilícita de información corporativa, ha sido responsable de considerables pérdidas económicas y quiebras empresariales. Los métodos tradicionales de detección, como la revisión manual de estados financieros, han demostrado ser ineficaces debido a la sofisticación en la ocultación de fraudes. En este contexto, la investigación se ha volcado hacia el desarrollo de modelos inteligentes que permitan una detección más efectiva.

En el ámbito de la detección de fraudes en estados financieros, se han empleado diversos enfoques, desde métodos estadísticos hasta técnicas más avanzadas como reglas de asociación, árboles de decisión y modelos de aprendizaje profundo, como las redes neuronales convolucionales (CNN). A pesar de la efectividad demostrada por muchos de estos métodos, una limitación importante radica en su falta de explicabilidad, lo cual podría no satisfacer completamente los requisitos regulatorios que demandan claridad en la evaluación de riesgos y la detección de fraudes.

# MARCO TEÓRICO

## Fraude:

Dentro de la gran cantidad de definiciones existentes sobre fraude la cual trata de abordar el término como una irregularidad o acto ilegal que se caracteriza por un engaño o cambio intencional. The institute of internal auditors que es el ente que lidera la profesión interna en Colombia define el fraude como: cualquier acto ilegal caracterizado por el engaño, el encubrimiento o la violación de la confianza, cabe aclarar que no son dependientes de la violencia o fuerza física. Los fraudes son perpetrados por personas, partidos u organizaciones para obtener algún dinero, bien o servicio, para evitar el pago o perdida del servicio o asegurar una ventaja personal y/o comercial [1].

## Fraude financiero:

Se refiere a una actividad ilegal o engañosa que se realiza en el contexto de una transacción financiera con el fin de obtener beneficios financieros de forma ilegal o injusta.
Dentro de las posibilidades de fraude pueden existir falsificación de documentos, el uso indebido de información privilegiada, la manipulación de los mercados financieros, la malversación de fondos, la apropiación indebida de activos, etc.
El fraude financiero puede perjudicar tanto a individuos, empresas, inversores y a la economía en general, causando daños reputacionales y financieros [1].

## Clasificación del fraude financiero:

### Fraude de crédito:

El fraude de crédito se refiere a la práctica de obtener crédito mediante la presentación de información falsa o engañosa, con el fin de obtener financiamiento sin la capacidad de pagar el mismo. También puede referirse a la manipulación o el uso no autorizado de tarjetas de crédito o información de cuentas para obtener beneficios financieros de manera fraudulenta.

### Fraude de inversión:

Es un esquema del fraude el cual se basa en engaños utilizados para convencer a los inversores para que inviertan su dinero en oportunidades que prometen altos rendimientos y poco o ningún riesgo, sin embargo, estas ‘oportunidades’ suelen ser esquemas piramidales o formas fraudulentas que no tienen ninguna intención de generar beneficios al inversor, en su mayoría los beneficiados son los estafadores.

### Fraude de tarjeta de crédito:

El fraude de crédito es una forma de robo o engaño que involucra el uso de información personal para realizar compras o extracción de dinero por medio de cuentas de crédito, este tipo de fraude causa un impacto muy grande a la víctima ya que podría traer consecuencias como el daño del historial crediticio o pérdida de fondos.

**Fraude de seguros:**

El fraude de seguros se refiere a cualquier acto de reclamación del seguro por medio de engaños o la intención de pagos fraudulentos, este tipo de fraudes pueden ser ejecutadas por solicitantes como proveedores, lo que se conoce de este tipo de fraudes es que anualmente les cuesta a las aseguradoras miles de millones en pérdidas.
En este documento nos centraremos en el fraude interno.

**Fraude interno:**

Se refiere a acciones fraudulentas cometidas por los empleados o directivos de una organización financiera. Estas actividades son preocupantes y difíciles de detectar ya que los involucrados tienen conocimientos de los sistemas y procesos bancarios que les otorgo una mayor ventaja para proceder y ocultar las acciones ilícitas.

**Tipos de fraude interno:**

Fraude en cuentas de clientes: Acceso y uso no autorizado de las cuentas de los clientes, las cuales pueden ser usadas para realizar transferencias no autorizadas, retiros de dinero, etc.
Uso indebido de recursos de la entidad financiera: Uso de los recursos de la entidad financiera de manera inapropiado para beneficio personal, muchas veces sin que la entidad financiera tenga conocimiento o de algún tipo de aprobación.
Manipulación contable: Alteración de las cifras en los libros contables para dar certeza o para mostrar algún tipo de salud financiera falsa.
Préstamos bancarios: Creación o aprobación de préstamos a orígenes ficticios para dar beneficios a terceros (cómplices) o personales.

Factores que contribuyen al fraude interno

Alguno de los factores que contribuyen que existan fraudes internos dentro de las entidades financieras son la falta de controles y detecciones tempranas, esto puede incrementar el riesgo de fraude dentro de las organizaciones.

Cultura organizacional: un ambiente laboral donde no se promueva la ética, profesionalismo y la integridad puede conducir a comportamientos fraudulentos.
Falta de capacitación: La falta de capacitación en las reglas y políticas que mitiguen el fraude puede incurrir en errores por no poder identificarlo o incluso entrar a ser directamente responsable.

**Consecuencias del fraude interno:**

Pérdidas financieras: las pérdidas financieras pueden estar encaminadas en dos sectores, una de ellas son las pérdidas directas que puede ocurrir debido al robo del dinero por cualquiera de los medios mencionados anteriormente o está la ruta indirecta que puede deberse a multas o sanciones.
Daño a la reputación: La confianza en las entidades bancarias es una parte muy fundamental en la industria financiera y que se conozcan casos no detectados y no solucionados dentro de los clientes y/o accionistas puede dañar gravemente la imagen, lo cual incurre en la pérdida de clientes, es decir, perdida de capital y movimiento de dinero.

Consecuencias legales: dependiendo del caso se pueden tener sanciones o cargos criminales a las personas involucradas.

Toda la información presentada en esta sección del documento es tomada de [18]

**Prevención y detección:**

Con la ayuda de controles y metodologías modernas es posible llegar a detectar anomalías en los comportamientos financieros, movimientos y demás cosas inusuales en los principales actores del fraude interno dentro de los bancos.
Una de las metodologías que pueden llegar a usarse son los grafos ya que estos ayudan a detectar relaciones no lineales entre cuentas, movimientos, transacciones, registros, empleados, directores y demás datos que puedes usarse para la construcción de la red [6].

**Teoría de grafos:**

Los grafos son definidos como estructuras matemáticas que son usadas para describir relaciones entre entidades, estas estructuras, aunque muy poco conocidas se encuentran en todas partes, por ejemplo, las redes de comunicación se encuentras construidas con grafos las cuales ayudan a enrutar la información, en este ejemplo los celulares móviles podrían ser los nodos del grafo y los enlaces o bordes son las conexiones entre ellos.
La teoría de grafos ha tomado mayor relevancia en la época actual ya que gracias a los avances en la tecnología y al desarrollo de múltiples algoritmos es posible identificar propiedades y mediciones en las relaciones de los objetos, también los modelos matemáticos han avanzado para ayudar a entender mejor el comportamiento de este método.

**Algunas propiedades de los grafos:**

Orden: el orden de un grafo es el número de vértices que contine y el tamaño del grafo es el número de bordes o enlaces.

Grados: Los grados de un vértice son el número de enlaces adyacentes a él. Los vecinos de un vértice en un grafo es un subconjunto del vértice inducido por todos los vértices.

Los vecinos de un grafo: Los vértices v de un grafo h es un subconjunto de h, compuesto de los vértices adyacentes de v y todos los enlaces conectados al vértice adyacente.

**Ejemplo de un grafo:**

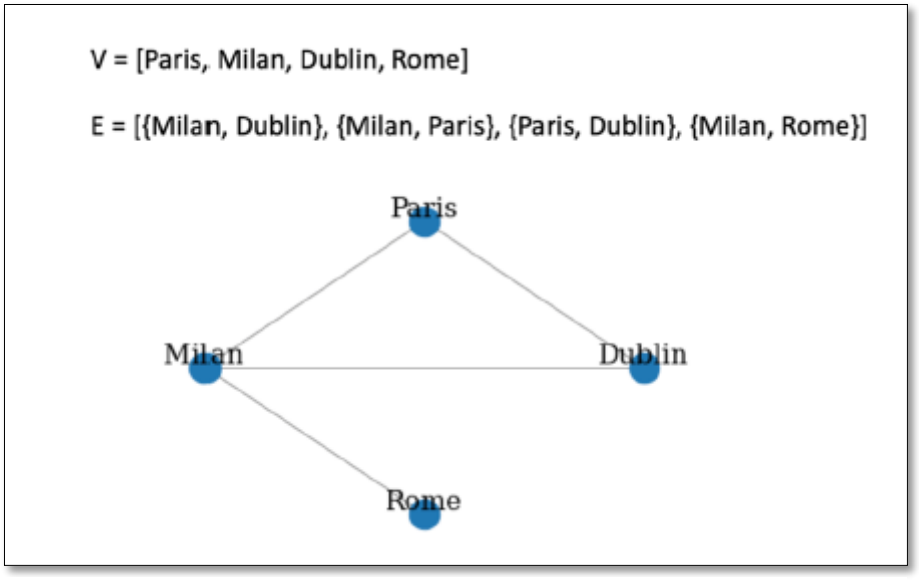


**Figura 2.** Ejemplo de grafo "Graph Machine Learning"

**Tipos de grafos:**

- Grafos dirigidos (Digraphs):

Un grafo G es definido como g = {V,E} donde V son los vértices, es decir un conjunto de nodos V = {v1,v2,v3,..,vn} y E son un conjunto de conexiones, es decir, E = {(Vk, Vw), (Vi,Vj)} es un conjunto de pares que crean una conexión entre nodos.

Como los conjuntos de cada elemento de E son pares ordenados, impone la dirección de la conexión, los bordes (Vk, Vw) significa que el nodo de Vk va a Vw, es una cuestión diferente los bordes (Vw, Vk), para este caso cambia el sentido Vw va a Vk a este tipo de relación se le conoce a Vw como cabeza (Head) y Vk como cola (Tail).

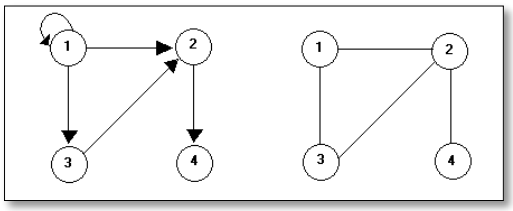


**Figura 3.** Ejemplo de grafo dirigido "Graph Machine Learning"

- **Multígrafo:**

Esta es muy sencilla de dar teniendo un conocimiento previo de lo que es un grafo, los multígrafos no es más sino una generalización de los grafos a lo cual permite tener múltiples conexiones, eso quiere decir que múltiples enlaces pueden estar conectados al mismo nodo inicial o final.
un multígrafo G es definido como G = (V, E) donde V es el conjunto de nodos y E es el conjunto de enlaces que están dirigidos a esos nodos. Para el caso de multígrafos se

denominaría multígrafo dirigido si E es un conjunto múltiple de pares ordenados, si E es un multiconjunto de dos conjuntos estaría denominado como multígrafo no dirigido.

**Ejemplo de multígrafo dirigido:**

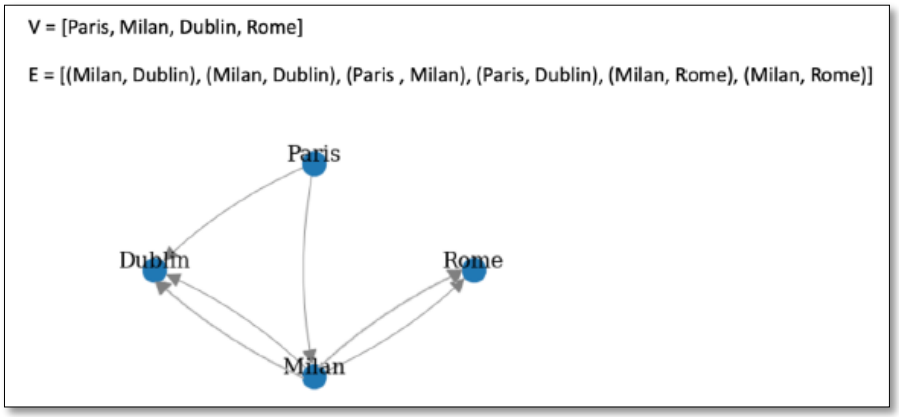


**Figura 4.** Ejemplo de multígrafo dirigido "Graph Machine Learning"

- **Grafos Ponderados:**

un grafo ponderado por la arista, denotado como (G=(V,E,w)), donde (V) es un conjunto de nodos, E es un conjunto de aristas, y (w: E → R) es una función ponderada que asigna a cada arista (e ∈ E) un peso representado como un número real. Asimismo, un grafo ponderado por el nodo, (G=(V,E,w)), se define con (V) como conjunto de nodos, E como conjunto de aristas, y \(w: V → R) como una función ponderada que asigna a cada nodo (v ∈ V) un peso expresado como un número real.

La principal diferencia entre las estructuras de los grafos dirigidos y los multígrafos son:

- Un multígrafo permite múltiples aristas entre los mismos nodos sin restricción de dirección.
- El dígrafo especifica la dirección para cada una de las aristas, estableciendo relaciones unidireccionales entre los nodos.

**Grado:**

Los grados en teoría de grafos se definen como la cantidad de enlaces que están conectados a un vértice o nodo, los grados son una medida importante ya que proporcionan información de cómo se encuentra la conexión de un grafo, de esto salen dos tipos de grado:

Debido a la presencia de la dirección dentro de los bordes se extiende la definición de los grados dentro de la teoría de grafos.

- Grado de entrada: Representa la cantidad de enlaces de un vértice a otro, esto quiere decir que da información de cuantos enlaces llegaron a un nodo, este tipo de medidas se encuentra en los grafos dirigidos o más conocidos como grafos orientados.

- Grado de salida: El grado de salida de un grafo dirigido es el número de aristas o enlaces que salen de un nodo y apuntan a otro.

Toda la información presentada en esta sección del documento es tomada de [7]

**Nod2vec:**

Para poder comprender el algoritmo nod2vec hay que conocer el algoritmo del cual está fuertemente inspirado ese es Word2vec. El Word2vec es una red neuronal poco profunda de dos capas que está construido para la reconstrucción de contextos lingüísticos de palabras, el sentido del modelo es crear representaciones de palabras en un vector dado un corpus de texto, las palabras se colocan en un espacio de incrustación de manera que si comparten un mismo contexto dentro del corpus las palabras se aproximaran entre sí.

Nod2vec es un algoritmo de incrustaciones de nodos que calcula representaciones vectoriales del nodo basándose en recorridos aleatorias del grafo, el barrido se muestrea basándose en paseos aleatorios, el algoritmo entra por una red neuronal de una sola capa oculta, la red neuronal calcula la probabilidad de encontrarse un nodo bajo una caminata aleatoria y en función de la aparición de los diferentes nodos [9].

**Caminata aleatoria:**

Una caminata aleatoria simula un recorrido por el grafo en el que las relaciones de esté, se eligen al azar, en una caminata clásica cada relación tiene la misma probabilidad de ser escogida, esta probabilidad no se ve alterada por los nodos visitadas con anterioridad [8].

DeepWalk aborda este propósito generando caminos aleatorios en el grafo, conocidos como "walks", recorriendo múltiples veces el grafo desde todos los nodos con una longitud específica. Este enfoque captura la estructura del grafo para traducirla a un espacio dimensional nuevo, buscando que nodos cercanos en el grafo resulten en vectores cercanos en dicho espacio. Estos caminos se utilizan como entrada para el modelo word2vec, mediante el descenso por gradiente, se aprenden los vectores en la nueva dimensión que maximizan la probabilidad de ocurrencia de cada nodo en sus respectivos caminos, dados los nodos previos [12].

**Embeddings:**

Los embeddings son una técnica de procesamiento de lenguaje natural que convierte las palabras en vectores matemáticos, el almacenamiento es menor y se preserva la información del grafo en sus vectores, haciendo que en cuestiones de procesamiento sea mucho más eficientes [10].

Los embeddings tienen como objetivo representar cada nodo de un grafo como un vector de características en un espacio de dimensiones d, siguiendo el mismo principio propuesto por

word2vec. Similar a cómo word2vec estima la probabilidad de ocurrencia de una palabra en una frase basándose en palabras cercanas, estas técnicas buscan estimar la probabilidad de que un nodo exista en un camino del grafo, dados los nodos anteriores en dicho camino [12].

**Normalización:**

La normalización de datos es una técnica muy utilizada en el procesamiento de datos, especialmente en el campo de aprendizaje automático y análisis de datos, para modificar la escala de las características o variables de manera que tengan un rango común o distribución especifica.

**Arboles de decisión:**

Es un modelo de aprendizaje automático supervisado utilizado para clasificación como para regresión, su estructura es similar a un diagrama de flujo donde cada nodo interno representa una decisión basada en ciertas características, cada rama representa el resultado de esa decisión y cada nodo hijo u hoja representa la predicción [5].

**Modelos logísticos:**

Es un método estadístico que se utiliza para predecir la probabilidad de una variable dependiendo de la categoría, es especialmente utilizada en datos de clasificación binaria, aunque también se puede utilizar en clasificación multiclase.

**Modelos de Boosting como Xgboost:**

Los modelos de Boosting como Xgboost son técnicas de aprendizaje automático usadas para la construcción de modelos predictivos robustos y precisos. Los Boosting son métodos de ensamblaje que mejoran el rendimiento del modelo combinando múltiples modelos débiles para crear un modelo fuerte y robusto [5].

**Otra definición de XGboost:**

XGBoost, o Gradient Boosting Extremo, se configura como un algoritmo de aprendizaje supervisado fundamentado en árboles de decisión, y destaca por su implementación eficiente del proceso de boosting, elevando así su capacidad predictiva. Reconocido como uno de los algoritmos más efectivos en Machine Learning, XGBoost ha alcanzado prominencia al demostrar consistentemente resultados superiores en competiciones. Este método se distingue por su diseño que aprovecha la capacidad de cómputo, permitiendo una ejecución paralelizable que utiliza eficazmente múltiples núcleos, resultando en un entrenamiento notablemente rápido.

Una característica destacada de XGBoost y otros métodos basados en árboles de decisión es su capacidad para proporcionar estimaciones de la importancia de las características. Evalúa la relevancia de cada característica mediante puntuaciones, considerando más relevantes aquellas características utilizadas con mayor frecuencia en la construcción de los árboles de decisión. En términos de la evolución de los algoritmos de Ensamble, XGBoost se posiciona en el estado

del arte, representando una avanzada contribución a este campo, como se ilustra en la figura adjunta [17].

**Ensamble:**

Las técnicas de ensamble es una forma de combinar modelos individuales para mejorar la estabilidad y la predicción de los modelos de machine learning. Se menciona que este tipo de métodos mejora el rendimiento de la predicción combinando lo mejor de múltiples modelos de aprendizaje automático, cada una de las predicciones se combinan para obtener una única predicción [17].

# Capítulo 5
# METODOLOGÍA

Los datos utilizados en este trabajo son tomados de los sistemas de información de la entidad financiera, por lo cual son de carácter confidencial.

**Recopilación de Datos:** Para este punto del proyecto se decidió hacer una búsqueda por coincidencia en los datos financieros de la entidad para encontrar caracteres que podrían servir para la construcción de los grafos señalados, esta metodología fue una búsqueda recursiva dentro de los metadatos almacenados y se extrajeron para realizar una exploración inicial.

**Instrumentos de Recopilación de Datos:** Dentro de la entidad se tienen organizada la información de la entidad financiera en una especie de repositorio central, donde de manera sencilla se puede hacer una búsqueda por nombramiento de la información necesaria para el estudio, se realizó un pequeño desarrollo en Google Script para organizar y clasificar los datos requeridos y así poder sintetizar todo en una Hoja de Excel que servirá para describir los campos de las tablas extraídas.

Este proceso servirá para identificar de manera inicial donde se pueden encontrar las fuentes y como se pueden relacionar entre ellas, esto de manera lógica teniendo en cuenta que dentro de sus descripciones están las llaves primarias y foráneas de cada tabla lo cual es útil para general esas uniones y relacionar la información de manera dimensional.

**Variables:** Las variables que se van a usar dentro del estudio son campos que describen a los empleados que generan colocaciones de crédito a los clientes, es decir, son empleados que están dirigidos directamente al Front de la organización financiera, a partir de este punto se localizaron variables que tuvieron correlación con los movimientos de estas personas. Se hablan también de variables transaccionales, las cuales indican movimientos, fechas, saldos y demás información sensible para del cliente y por último variables que describen los productos de los clientes, estos pueden ser créditos, cuentas bancarias, tarjetas, servicios de inversión, etc.

**Plan de Análisis:** Se comenzará con un análisis de la metodología dando como entendimiento que fuentes son necesarias para la construcción de los grafos, esto da como resultado unas tablas que serán utilizadas para la unificación de información que puede estar correlacionadas entre sí y puede mostrar patrones, anomalías, etc. La construcción formara una parte fundamental de la metodología ya que se debe tener en cuenta que covariables son importantes y es aquí donde el análisis, la lógica y la decisión de que se debería tener en el proyecto es importante, cuando se tengan las fuentes unificadas se analizara la relación lineal y no lineal de los empleados con los productos y con los clientes para poder detectar comportamientos y características que no son tan fáciles de observar a primera vista.
Después de tener este primer análisis se comprenderá mucho mejor la información y se podrá tener una mejor visual del comportamiento de los empleados y esto ayudará a la construcción de las firmas de las oficinas los cual tendrá como resultado un análisis de que tan probable será encontrar en una oficina un empleado que tena una calificación de fraude financiero.

**Aspectos Éticos:** Para esta parte ética lo que se usara es el habeas data la cual nos permitirá manipular la información de los clientes con aprobación de ellos para el uso de metodologías y análisis de datos personales.

**Logística de la Investigación:** Lo primero que se debe realizar es medir hasta dónde van los resultados del proyecto, encontrar las fuentes necesarias para la unificación de datos, desarrollo de la metodología de grafos y medición de los resultados de grafos, extracción de los vectores que servirán para el modelamiento, predicción o clusterización de la información para dar señalar el fraude interno dentro de la organización bancaria

**Limitaciones y Delimitaciones:** Las limitaciones que se han encontrado hasta el momento es la manipulación de la información ya que se presentan millones de productos dentro de la cantidad enorme de clientes existentes en la entidad financiera esto produce un gran gasto en procesamiento dentro de los valores en nube que pueden ser perjudiciales para el gasto interno.

También surgieron demoras en los entendimientos y unificación de las fuentes de fraude dentro de la organización bancaria, debido a que este tipo de problemáticas no tienen antecedentes claros y grandes evidencias que demuestren una etapa de fraude. El fraude financiero es una problemática muy grande de detectar y poder detener, entonces si se considera que el fraude interno no es tan común esto hace que sea un poco más difícil la tarea.

**Documentación:** Los primero es la búsqueda de la información requerida por el estudio dentro del conjunto de metadatos y el entendimiento de los datos. Segmentar, Transformar y unificar la información encontrada.

# Capítulo 6
# RESULTADOS Y DISCUSIÓN

Los resultados que se han obtenido dentro del proyecto vienen constituidos dentro de los objetivos específicos generados al inicio de la investigación y se mostraran punto a punto con resultados tangibles para que se pueda generar una evaluación lógica del proyecto.

**Exploración de los datos iniciales:**
Para la construcción de la dimensión de empleados que es la tabla maestra que contiene toda la información de los empleados la cual se constituye por:

- Números de identificación.
- Tipo de identificación.
- Registro del empleado.
- Oficina en la que trabaja.
- Título del trabajo desempeñado.
- Fecha de actualización del dato
- Vigencia

Se interesa para el estudio conocer el fraude interno en cuestión de colocación de crédito y movimiento irregulares de dinero a través de las cuentas se tomaron como título de trabajo los empleados que tienen una interacción directa con los clientes de la entidad financiera.

**Tabla 1.** Títulos trabajo de la entidad bancaria.

| *Título del trabajo* | *Cantidad personas* |
|---|---|
| Admin personas | 1553 |
| Administrador de recursos | 1724 |
| Informador | 1463 |
| Personas coordinadora | 1024 |
| **Cantidad empleados** | **5764** |

Construir una red que sirva para identificar los productos asociados a cada cliente y como puede existir una relación entre más productos.
Para esta construcción se necesitaron todas las fuentes transaccionales de la entidad financiera, las cuales generar una relación entre el cliente y los productos que se encuentran asociados a este, esta unificación de datos al final de cuentas arroja información histórica de los clientes con un conjunto de movimientos financieros que pueden llegar a ser útiles para detectar

patrones, anomalías y establecer un comportamiento tanto del cliente como de empleados que ayudaron a colocar algún producto.

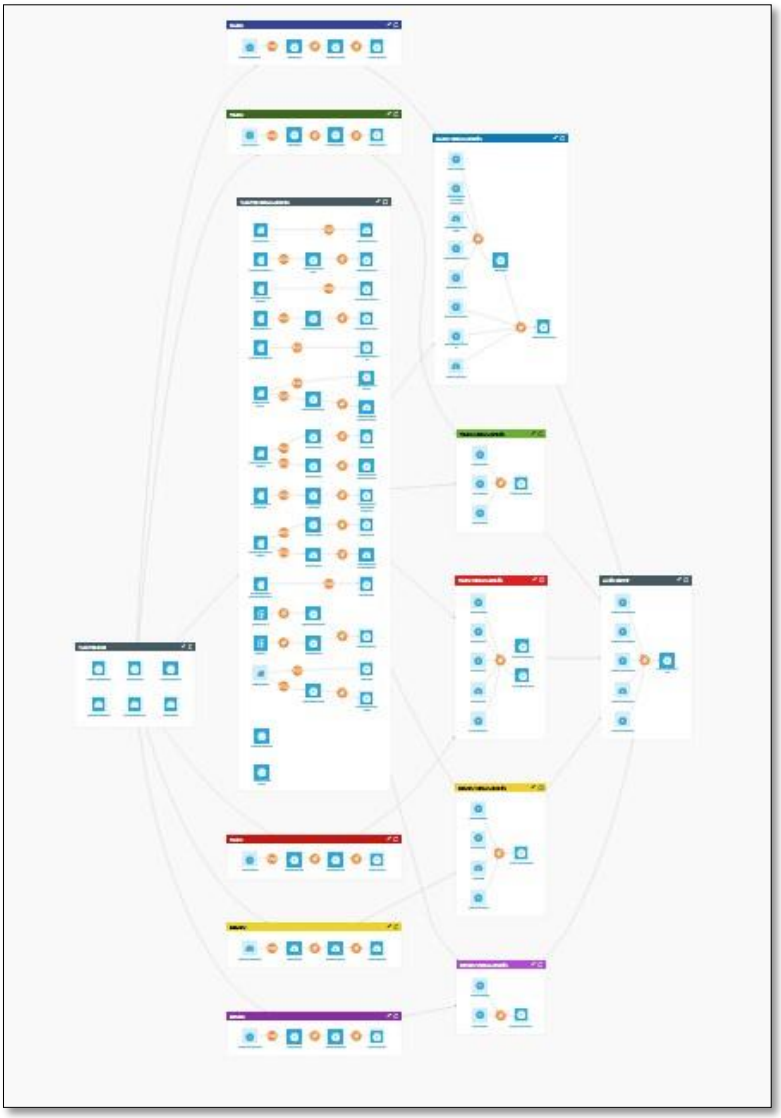

**Figura 5.** Esquema primario de la unificación de empleados

Construir una red de empleados y productos que sirva para identificar la relación entre los productos, clientes y empleado que ayudo a la colocación de dicho producto.

Para este objetivo lo que se tuvo en cuenta es la construcción de una dimensión de empleados con bases ingestadas, curadas y estandarizadas dentro de la entidad financiera que muestren covariables de los empleados que tienen un contacto directo con los clientes, más específicamente a sus productos y son aquellos que pueden generar colocación de créditos.
A este de construcciones se les llama dimensionales ya que tienen como objetivo unificar y centralizar la información específica de un segmento determinado en este caso la información de los empleados como; número de identificación, cuentas bancarias, fechas de ingreso,

ubicación donde trabaja, como se relaciona con el cliente, numero de productos que ayudo a colocar, saldos, movimientos, etc.

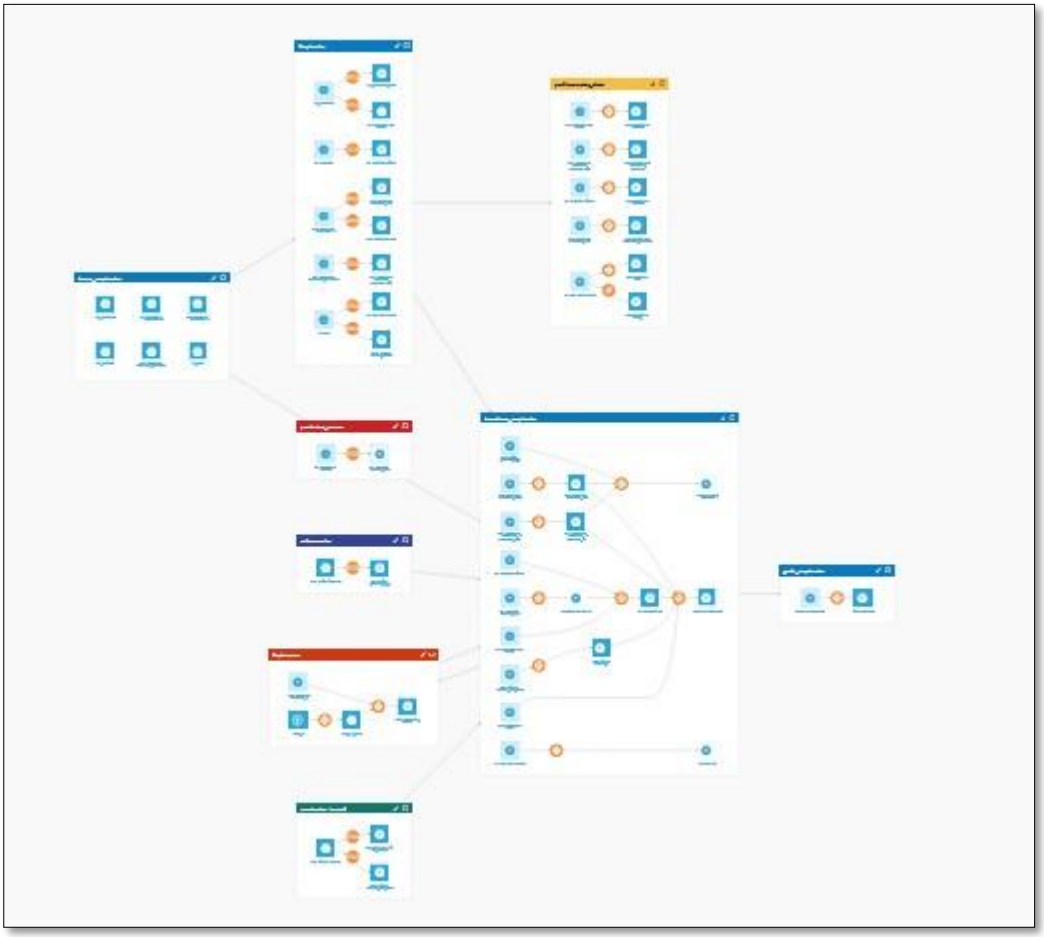

**Figura 6.** Esquema final del grafo de empleados

Construir el grafo que servirá como primera vista de la interacción de los empleados con los clientes y como estos se relación, teniendo en cuenta que el medio más cercano es por medio de los productos financieros.
En la siguiente sección se explicará a detalle cual fue la metodología y de qué forma se realizó la construcción a partir de la información dimensional mostrada en las figuras 4 y 5 lo cual fue utilizado para unificar la información de los empleados de la organización financiera y las transacciones que tiene esta por medio de la unificación a través del tiempo de sus registros financieros.

Las métricas y los exploratorios que se muestran en esta sección son con la ayuda de algoritmos y librerías como Networkx, nod2vec y Graphviz.

**Propiedades:**

**Captura de la información local y global en los grafos:** Los embeddings están diseñados para capturar la información local de un grafo esto quiere decir que encierra la información de las relaciones de los nodos y sus vecinos más cercanos, como también la información global que capturas las generalidades del grafo, esto permite que se vea la representación de un nodo y el papel que desempeña dentro del grafo.

**Conservación de las vecindades:** Se muestra la relación de si un nodo esta cercano a otro nodo dentro del grafo este tendrá la misma cercanía en un espacio de Embedding, es muy importante ya que remarcara la distancia entre los nodos cuando se vea representado en un espacio vectorial.

**Flexibilidad de dimensiones:** Los embeddings permiten las elecciones de dimensiones, haciendo que la captura de la información sea mucho más sencilla y ayude en términos de procesamiento de información.

**Métricas:**

**Estructura del grafo completo:**

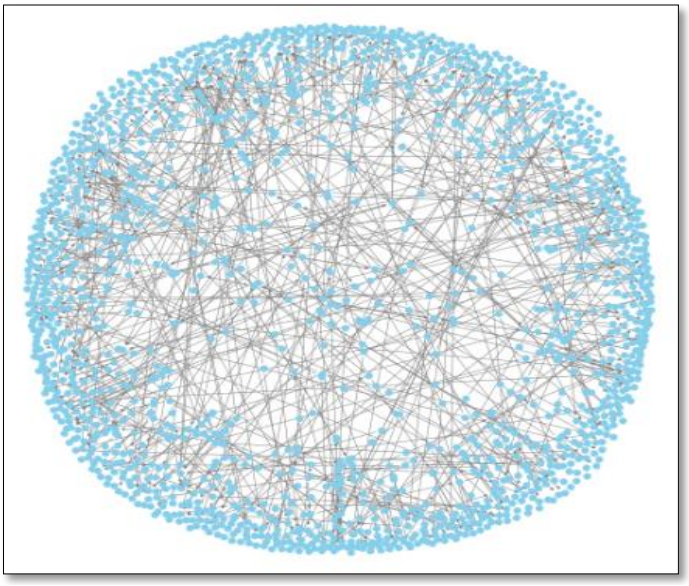

**Figura 7.** Grafo de empleados y clientes estructura completa

**Exploratorios del grafo:**

```
Graph Summary:
Number of nodes : 100962
Maximum degree : 2411
Minimum degree : 1
Average degree : 3.6319011113092055
Median degree : 1.0

Graph Connectivity
Strongly Connected Components : 97857
Weakly Connected Components : 129

Graph Distance

Graph Clustering
Transitivity : 0.006940419089131072
Average Clustering Coefficient : 0.021885348449784493
```

**Figura 8.** Exploratorios del grafo

Después de crear la dimensión de empleados la cual es útil para poder observar ciertas características de cada una de las personas que trabajan dentro de la organización financiera, las cuales pueden traer tanto los productos que contienen dentro del banco, números de cuenta, identificaciones, fechas, datos de contacto y más información útil y necesaria como dimensional, se cruzó la información transaccional para poder filtrar y señalar esos movimientos de dinero entre los empleados y cuentas asociadas, lo cual género que cuando se convirtiera la información dimensional a tipo grafo se encontraran más de 100 mil nodos en un periodo de tiempo de alrededor de un año de transacciones y en este un grado máximo de 2400 que no es otra cosa que el número de enlaces conectados a un nodo, esto nos indica que existen nodos altamente conectados, cuando se saca el promedio para observar la conexión alrededor de todos los nodos dentro del grafo este no da un estimado de 4 enlaces en promedio por cada nodo a grandes rasgos dejando evidencia de la cantidad de transacciones que pasan de una entidad a otra y la cantidad de enlaces generados por cada nodo en promedio.

**Tabla 2.** Métricas del Grafo grados de entrada y salida

| **Graph** | **out Degree** | **in Degree** |
|---|---|---|
| Number of nodes | 100963 | 100963 |
| Máximum Degree | 26 | 110 |
| Mínimum Degree | 0 | 0 |
| Average Degree | 5.74 | 5.74 |

| | | |
|---|---|---|
| Median Degree | 5.0 | 3.0 |

La tabla numero 2 está indicando las métricas iniciales para los grados tanto de entrada como de salida de los nodos los cuales se encuentran un total de 100963 nodos y un máximo de conexiones de salida de 26 lo cual puede ser un poco pequeño para la cantidad de nodos existentes esto quiere decir que no hay mucha interacción del nodo principal para generar enlaces con otros nodos que tenga alrededor pero para los nodos de entrada hay un máximo de 110 indicando que existen muchos enlaces de llega para los otros nodos, lo otro que puede destacar es el bajo promedio de conexiones entre los nodos para la cantidad de nodos existentes es por ello que se toma como iniciativa generar un subgrafo con los nodos mayormente conectados para evidenciar el comportamiento transaccional entre los empleados y el cliente.

4. Se genero un subgrafo con los nodos más conectados para poder observar de mejor manera las relaciones entre los empleados y los clientes ya sea que estos no estén conectados o no exista una adyacencia por medio de los enlaces o los grados.

Los nodos más conectados según la teoría de grafos son aquellos nodos que tienen un alto número de enlaces con otros nodos dentro del grafo, estas relaciones son realmente importantes porque permiten generar visualizaciones como en la figura 12 donde se resalta y se puede llegar a interpretar la relación de un nodo con otro así no se presente una conexión directa.

**Exploratorios del subgrafo:**

```
GetGraphMetrics(subgraph)

Graph Summary:
Number of nodes : 4079
Maximum degree : 110
Minimum degree : 1
Average degree : 11.483206668301055
Median degree : 9.0

Graph Connectivity
Strongly Connected Components : 1285
Weakly Connected Components : 65

Graph Distance

Graph Clustering
Transitivity : 0.10525872382011492
Average Clustering Coefficient : 0.12089968929688476
```

**Figura 9.** Exploratorios del Subgrafo

Para el subgrafo se seleccionaron los nodos que tuvieran más de 4 enlaces y menos de 400 enlaces entre sus nodos para poder encontrar las relaciones entre las transacciones, es por ello que el número total de nodos para este grafo son 4079 lo cual es bastante pequeño si se compara con el grafo observado con anterioridad, aunque se debe resaltar que las métricas en las conexiones mejoraron ya que para esta ocasión el numero promedio de grados o enlaces es de 12 con una mediana de 9 teniendo en cuenta que el anterior solamente llegaba a un promedio de 5, esto debido a la baja cantidad de transacciones generadas entre las cuentas de los empleados a cuentas desconocidas.

El subgrafo tiene unas conexiones fuertes de 1285 nodos bien conectados.

Ahora bien, se hablara de la transitividad la cual es una métrica que indica la relación entre los nodos en un grafo y como estos se conectan entre sí, es bueno tener un grafo altamente relacionado pero para esta ocasión la métrica marca un 0.10 lo cual es bueno sabiendo que el grafo tiene como nodos los empleados del banco y los clientes, si este tuviera una alta transitividad diría que hay una relación o una conexión entre las cuentas de los empleados y los clientes lo cual a nivel de riesgo y fraude sería perjudicial para la entidad bancaria porque se presentarían gran cantidad de estos casos con una métrica así.

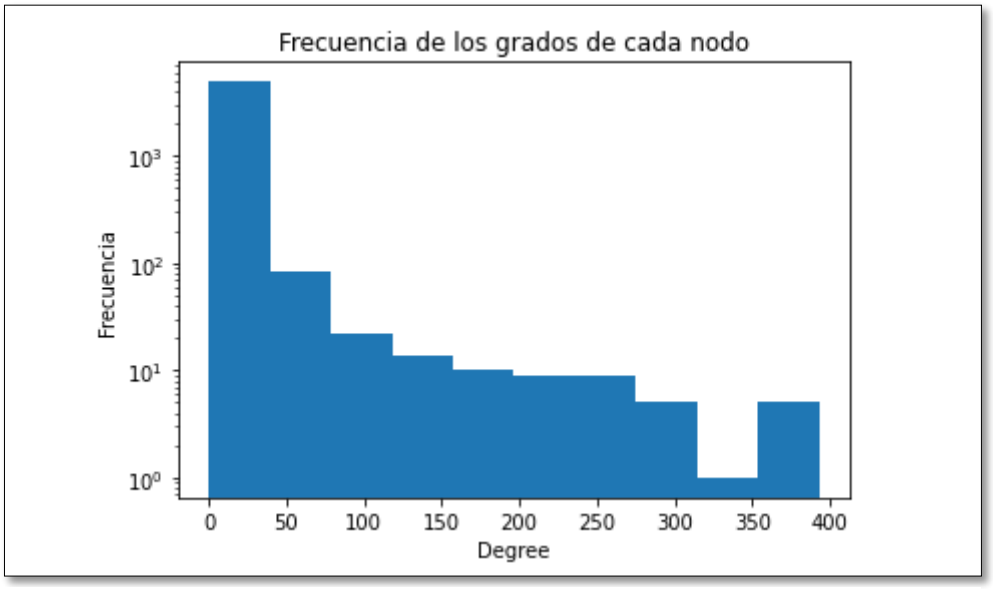


**Figura 10.** Frecuencia de los grados en el subgrafo

La figura numero 9 indica la frecuencia que se presenta en cada uno de los grados esto muestra la cantidad de enlaces que están en cada uno de los nodos, ejemplo existe una gran cantidad de nodos con una frecuencia baja de conexiones y una cantidad un poco menor que contiene muchos enlaces entre nodos por eso se tomó este subgrafo porque a pesar de contener una gran proporción de conexiones bajas contenía otras que estaban altamente conectadas y así poder formar un grafo que generara una conexión entre clientes y empleados y así encontrar anomalías en el segmento de empleados que están en cara al cliente y generan la colación de créditos.

**Tabla 3.** Métricas del Subgrafo grados de entrada y salida

| Graph | out Degree | in Degree |
|---|---|---|
| Number of nodes | 5085 | 5085 |
| Máximum Degree | 39 | 393 |
| Mínimum Degree | 0 | 0 |
| Average Degree | 8.24 | 8.24 |
| Median Degree | 8.0 | 4.0 |

Como se mostró en la tabla anterior donde se evidenciaban las métricas para los grados de entrada y de salida y se evidencio que existía un promedio bajo en los enlaces del nodo, las métricas del subgrafo son mucho más interesantes ya que muestran una conectividad mucho más alta, generando interacción entre los nodos de empleados y clientes.

**Subgrafo de empleados (Conexión entre empleados y clientes)**

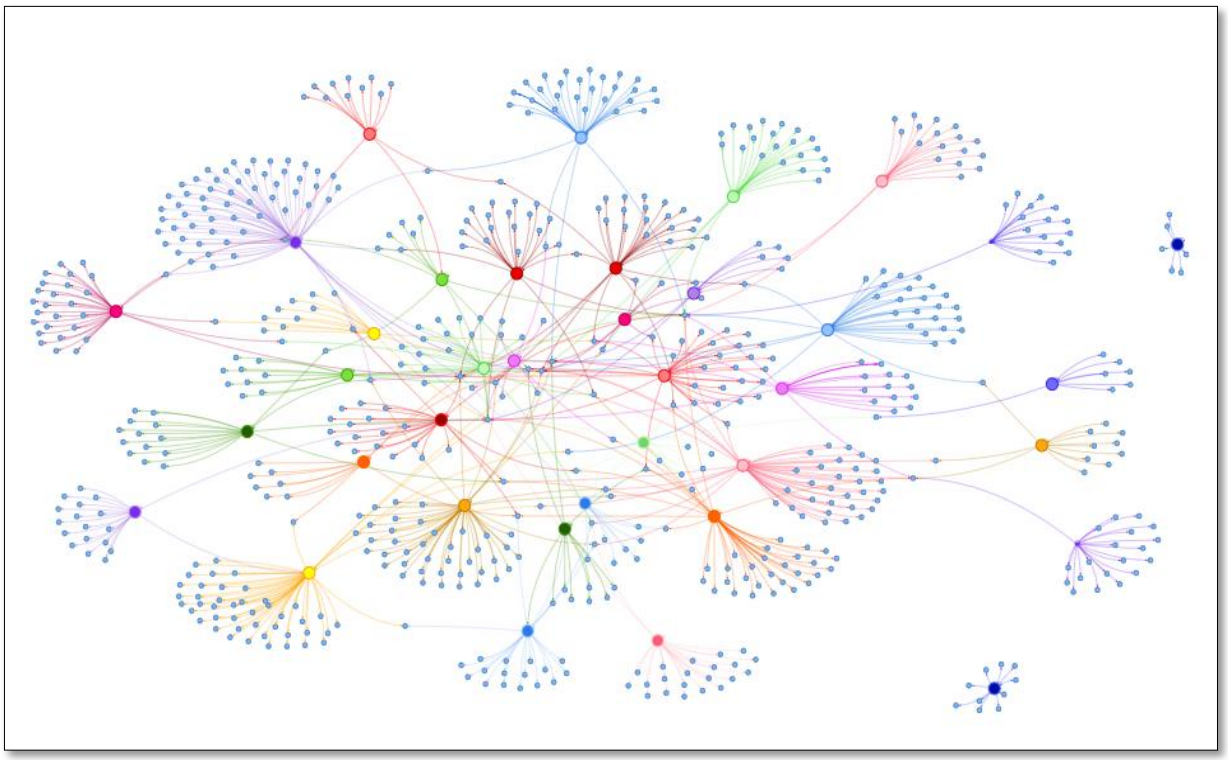

**Figura 11.** Subgrafo

Este es el subgrafo de empleados y clientes el cual es un subgrafo complejo con múltiples nodos y enlaces, este contiene características importantes las cuales son:
Los nodos centrales, estos nodos son los que tienen un grado mayor dentro de toda la visualización y se ven representados por los colores más vivos, los cual indica una importancia alta en el grafo ya que estos permiten la alta conexión entre los nodos de toda la red.
Los colores generados dentro de la visualización del grafo representan la conectividad entre los nodos más cercanos y los que tienen ciertas características comunes es por ello por lo que la tonalidad entre colores es similar para aquellos que están estrechamente relacionados entre sí con el resto del grafo.
La conectividad de los nodos varía entre cada para de clúster generado en la visualización ya que algunos parecen tener una alta conectividad dentro de las aglomeraciones de nodos presentes y otras parecen estar aisladas.
Los enlaces que tienen diferentes colores coinciden con la relación entre cada uno de los nodos conectado, también se puede apreciar el grosor de la arista la cual es un indicativo de la fuerza en la conexión de los nodos o el peso entre ellos.
La estructura del grafico permite intuir que es un grafo no uniforme y más jerárquico en donde los nodos los nodos con un tamaño mayor juegan un papel importante en la conectividad del grafo.
Existe un aislamiento de algunos nodos dentro de la visualización, esto indica la desconexión de nodos que podrían no ser funcionales del grafo indicando que no existe una relación de conectividad con los nodos centrales dentro de la red.
Este es un grafo dirigido, no se aprecia por la gran cantidad de nodos existentes, pero se muestra alta conectividad, la entrada y salida con flechas señaladas por cada nodo.

**RESULTADOS DEL MODELAMIENTO:**

Para la etapa de modelamiento se utilizó los embeddings que son las representaciones vectoriales densas y de baja dimensión provenientes del grafo construido anteriormente lo cual son representaciones de los nodos que se evidenciaron en la figura 10, se añadieron etiquetas provenientes de fraude las cuales fueron adicionadas por medio de cálculos aritméticos que dan como resultado si el empleado esta señalado de fraude o no, gracias a esta etiqueta es posible tener un modelo supervisado.
Los modelos que se generaron fueron:

- Decision Tree
- Random Forest
- XGboost

Se utilizo para cada uno de los modelos la metodología de optimización de hiperparametros con la librería de Hyperopt la cual es flexible y muy utilizada en el aprendizaje automático porque ayuda al ajuste de los hiperparametros de los modelos y a mejorar su rendimiento.
Se uso dentro de cada modelo una optimización para los parámetros correspondientes, se debe tener en cuenta que como son datos de transacciones y lo que se pretende es predecir el posible fraude los datos etiquetados, estos datos están altamente desbalanceados ya que encontrar fraude no es una tarea tan sencilla y también se debe tener en cuenta el uso de las métricas de rendimiento que no se vean afectadas por datos desbalanceados.

```
from sklearn.ensemble import RandomForestClassifier

model_random = RandomForestClassifier('class_weight': 'balanced_subsample',
                                      'max_depth': 20,
                                      'max_features': 1,
                                      'min_samples_leaf': 0.12707132804308563,
                                      'min_samples_split': 0.21117166910289642,
                                      'n_estimators': 53)
                                      max_evals = 200)

md_random_ajus = model_random.fit(x_train_scaler, y_train)
```

En la figura 12 se puede observar la perdida en cada una de las iteraciones de optimización del algoritmo lo que busca este es la minimización de perdida en el scoring que se le coloco para este caso se usó balanced_accuracy que es una métrica de precisión para clasificación donde tiene en cuenta esas bases desbalanceadas, al final de las iteraciones podemos observar una convergencia en los resultados al final de la optimización.

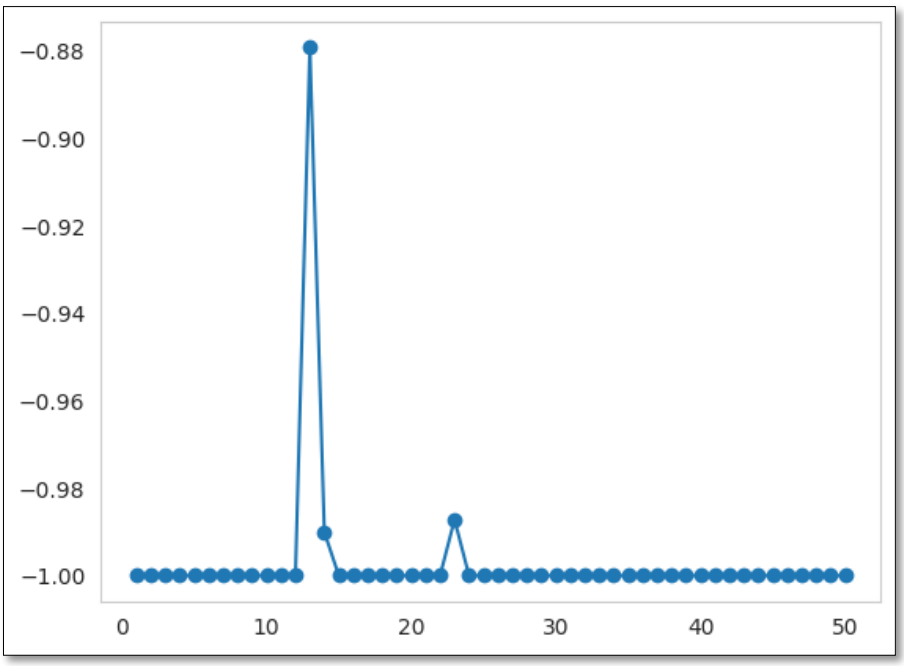


**Figura 12.** Perdida en optimización de hiperparametros

$$BALANCED\ ACCURACY = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

- TP son los Verdaderos Positivos.
- FN son los Falsos Negativos.
- TN son los Verdaderos Negativos.
- FP son los Falsos Positivos.

En la figura 12 podemos visualizar que después de la iteración numero 50 estos datos mantienen una constante lo cual indica que después de este número de iteraciones el modelo no cambia y no es posible minimizar la perdida más, si se hace un cambio en los hiperparametros para aumentarlos es posible que se caiga en un error de sobreajuste.

**Resultados de las métricas de los modelos:**

En la etapa de validación se probaron los distintos modelos siempre variando sus hiperparametros para llegar al mejor resultado posible, teniendo en cuenta que en cada evaluación de los hiperparametros se usó la metodología de Cross Validation para evaluar el rendimiento de los modelos utilizando la métrica de balanced_accuracy y también muy útil para poder evitar el sobreajuste, junto con este se utiliza la optimización bayesiana para minimizar la función objetivo y encontrar la menor perdida dentro de cada uno de los modelos, en este caso tomando el negativo de la precisión balanceada.

Hay que tener en cuenta que la librería tomada fue Hyperopt que utiliza el algoritmo de Tree of Parzen Estimator la cual modela la función objetivo con un proceso de construcción de árboles basándose en los resultados de las anteriores corridas.
se usaron árboles de clasificación los cuales tienen parámetros que ayudan en este tipo de ocasiones las cuales son tener datos altamente desbalanceados, como se puede observar en la **tabla 4** los resultados para el Decision Tree en la métrica de roc_auc_score fue de 0.82, esta medida es muy importante ya que mide la capacidad del modelo de poder distinguir entre clases y se tomará en cuenta para la elección del modelo que se pondrá productivo.

Para el siguiente modelo que es aquel que esta descrito en la figura 11 el cual es el Random Forest la cual es una iteración constante del Decision Tree mejorando ciertos parámetros como el peso en las clases dándole un mayor valor a la etiqueta menos frecuente en los datos para así poder balancear en la clasificación del dato, el Random Forest no es más que una construcción de árboles aleatorios los cuales promedian sus resultado, en esta ocasión se tuvo un resultado en la métrica de roc_auc_score del 0.91 indicando que el modelo pudo clasificar de manera correcta las etiquetas de las distintas clases.

El último modelo es un XGboost es un modelo altamente eficiente y que generalmente da muy buenos resultados para una etapa de entrenamiento arroja un 0.92 en la métrica de roc_auc_score que es muy buen resultado en la evaluación, en las otras métricas de rendimiento fue muy útil y dando un indicativo que en precisión es el mejor modelo a escoger.

**Tabla 4**. Métricas de los modelos

| | Decision Tree | Random forest | Xgboost |
|---|---|---|---|
| balanced_accuracy_score | | | |
| train | 0,8287 | 0,8974 | 0,8012 |
| test | 0,7753 | 0,9136 | 0,8423 |
| recall_score | | | |
| train | 0,7995 | 0,9680 | 1,0 |

| | | | |
|---|---|---|---|
| test | 0,7968 | 0,9504 | 1,0 |
| f1_score | | | |
| train | 0,8876 | 0,9826 | 1,0 |
| test | 0,8855 | 0,9738 | 0,9981 |
| roc_auc_score | | | |
| train | 0,8287 | 0,8974 | 0,9232 |
| test | 0,7753 | 0,9136 | 0,8423 |
| matthews_corrcoef | | | |
| train | 0,1875 | 0,8974 | 0,8891 |
| test | 0,1459 | 0,3775 | 0,8259 |

Estos son las métricas de los modelos en una etapa de resultados donde se evidencian variaciones respecto a la etapa de entrenamiento, teniendo en cuenta que estos fueron datos que para el modelo fueron nuevos y nunca se corrieron dentro de sus estructuras dieron muy buenos resultados tanto en sus precisiones como en la capacidad de clasificación para los datos desbalanceados, aunque hay que tener en cuenta que en casi todos los modelos se encontraron disminuciones en las métricas de test no fueron variaciones tan altas y es algo esperado al predecir datos nunca antes vistos por los modelos.

Se escoge el modelo XGboost por su alta eficiencia y rapidez en la clasificación y predicción de los datos.

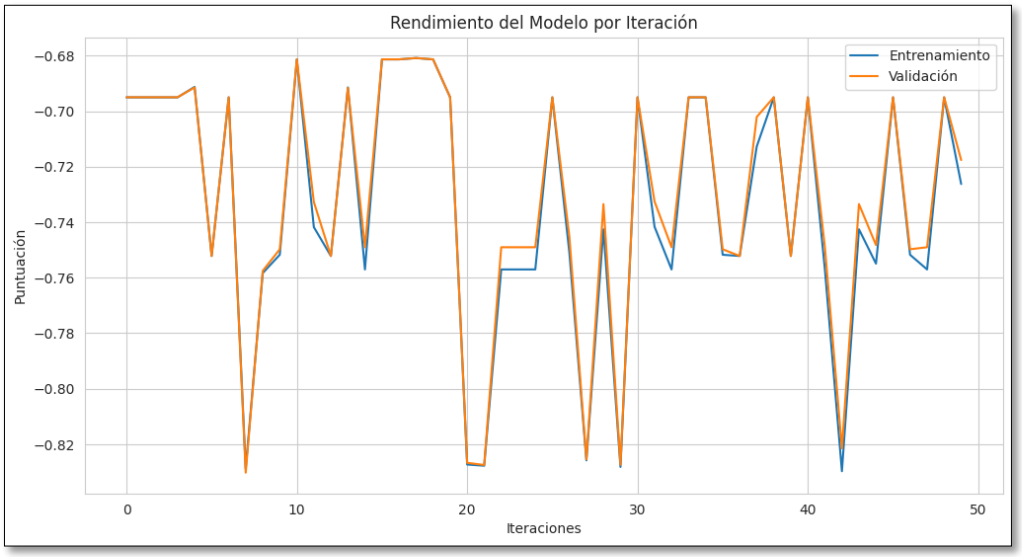


**Figura 13.** Evolución modelo en fase de entrenamiento

Dentro de la evolución del algoritmo cuando se optimizo con el método Bayesiano se ve un comportamiento muy cercano de validación respecto a cómo se va comportando a lo largo de las iteraciones, cuando se realizó el testeo del modelo con datos desconocidos que nunca había visto el algoritmo este tuvo muy buenos resultados.

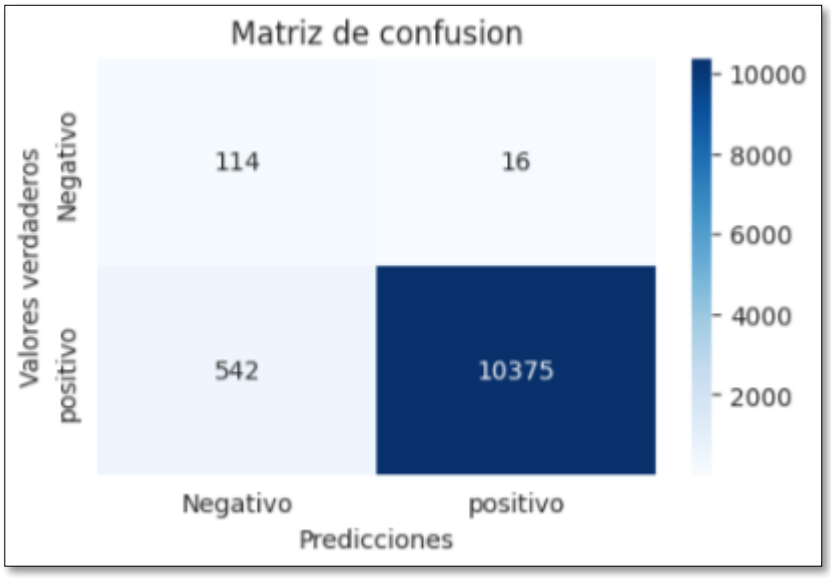


**Figura 14.** Perdida en optimización de hiperparametros

Hay una gran aproximación a la predicción de clasificación de los verdaderos positivos en el modelo de Random Forest donde este solamente se ha equivocado en la clasificación de 542 datos y 16 de los que serían personas o empleados que cometieron fraude según el etiquetado de datos.

# Capítulo 7

# CONCLUSIONES

En la etapa final del proyecto, tras una exhaustiva evaluación de riesgo, se ha establecido una estructura categorizada que revela insights cruciales sobre los posibles empleados de cara al cliente que generarían riesgo de fraude a la organización. Este análisis se centró en los niveles de riesgo entre los empleados de la entidad financiera, basándose en la probabilidad inicialmente no confirmada de un 1.32% de incidencias, como se puede ver es muy pequeña la cantidad de estimaciones hechas para los empleados.

- Alto riesgo (2.01% de los empleados): Este grupo identificado como el más propenso a comportamientos fraudulentos, merece una atención inmediata y detallada. La segmentación no representa un segmento significativo, pero las pérdidas financieras pueden llegar a ser mucho más grandes.

- Riesgo moderado: Este grupo es la parte gris dentro de los resultados y son aquellos casos que salen como falsos positivos, pero definitivamente no se pueden descartar de forma tan sencilla, se volverían casos de alta sospecha donde se priorizaría una investigación para aquellos empleados con alta sospecha. Esto permitirá enfocar esfuerzos y recursos de manera estratégica y eficiente para una gestión de riesgos más efectiva.

- Riesgo bajo o nulo: Aunque hay empleados que no fueron considerados con evidencias de riesgo de fraude, se subraya la importancia de recalibración periódica de variables para evitar una transacción de riesgo hacia las categorías de los empleados con riesgo más elevada. La vigilancia es clave para la mejora de los resultados.

Este estudio concluye que la cantidad de fraude interno dentro de la organización financiera fue catalogada con muy buenos resultados y aunque la cantidad de empleados etiquetados dentro del riesgo alto no es muy grande se deben considerar estrategias para la detección de fraude, por lo tanto, es bueno considerar en reevaluar y recalibrar las estrategias de gestión de riesgos. Esta reorientación estratégica no solo ayudara a alinear los riesgos con los umbrales de tolerancia de la entidad financiera, sino que también ayudara a minimizar las perdidas proyectadas en el fraude interno.

Se debe tener en cuenta que con este proyecto no se pretende solucionar todas las problemáticas de fraude que existen en la entidad financiera, los pasos a seguir son ajustar el modelo escogido y los datos para un paso productivo y generar monitoreos para activar la gestión de riesgos de fraude interno.

# REFERENCIAS

[1] The institute of internal auditors (2009). Internal Auditing And Fraud.

[2] Khaled Gubran Al-Hashedi, Pritheega Magalingam (2021). Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019.

[3] Abdallah, Zainal (2016). Fraud detection system_ A survey. https://www.sciencedirect.com/science/article/abs/pii/S1084804516300571

[4] Alka Rani, Nishant K. Sinha (2022). Support Vector Machine. https://www.sciencedirect.com/topics/computer-science/support-vector-machine.

[5] El Bouchefry PhD, S. de Souza PhD. (2020). Chapter 12 - Learning in Big Data: Introduction to Machine Learning. https://www.sciencedirect.com/science/article/abs/pii/B9780128191545000230

[6] Siqi Cai, Zhenping Xie, explainable fraud detection of financial statement data driven by two-layer knowledge graph (2024). https://www.sciencedirect.com/science/article/abs/pii/S0957417423036308

[7] Claudio Stamile, Aldo Marzullo, Enrico Deusebio. (2021). Graph Machine Learning. Packt Publishing Ltd.

[8] Neo4j, Inc. (2023) Nod2vec. neo4j. https://neo4j.com/docs/graph-data-science/current/machine-learning/node-embeddings/node2vec/

[9] Cohen Elior. (Apr 16, 2018). node2vec: Embeddings for Graph Data. Towards Data Science. https://towardsdatascience.com/node2vec-embeddings-for-graph-data-32a866340fef

[10] Tomaz Bratanic. (Aug 16, 2021). Complete guide to understanding Node2Vec algorithm. Towards Data Science. https://towardsdatascience.com/complete-guide-to-understanding-node2vec-algorithm-4e9a35e5d147

[11] Chris McCormick (19 Apr 2016). Word2Vec Tutorial - The Skip-Gram Model. Mccormickml. http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

[12] Irene. (April 23, 2020). Node Embeddings: DeepWalk & Node2Vec. wutheringgraphs https://wutheringgraphs.wordpress.com/2020/04/23/node-embeddings-deepwalk-node2vec/

[13] Ballesteros Jaime. (2021). Exploración de modelos transaccionales para recomendaciones de ítems [Grado en Ingeniería Informática, Universidad Autónoma de Madrid]. https://repositorio.uam.es/bitstream/handle/10486/698176/enriquez_ballesteros_jaime_tfg.pdf?sequence=1

[14] PWC. (2022, mayo). Encuesta Global de Crimen y Fraude 2022, PwC Colombia. https://www.pwc.com/co/es/publicaciones/encuesta-crimen-fraude-economico.html

[15] Chapman & Hall (2017). CRC Data Mining and Knowledge Discovery Series. Taylor & Francis Group.

[16] Dutta, K. y Perry, J. (2006). A tale of tails: An empirical analysis of loss distribution models for estimating operational risk capital. Federal Reserve Bank of Boston, Working Paper No. 06-13.

[17] Cortez Samuel. (2022). Introducción a los Métodos de Ensamble y al Algoritmo de XGBoost: Caso Práctico. https://medium.com/@oscars.cortezmo/introducci%C3%B3n-a-los-m%C3%A9todos-de-ensamble-y-al-algoritmo-de-xgboost-caso-pr%C3%A1ctico-e8cb0d58394b

[18] Eafit. (Sin fecha). ¿QUE ES FRAUDE? https://www.eafit.edu.co/escuelas/administracion/consultorio-contable/Documents/A%20FRAUDE.pdf

# LISTA DE FIGURAS