UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

Resolución: RPC-SO-32-No.536-2023-CES

PROYECTO DE TITULACIÓN EN OPCIÓN AL GRADO DE MAGÍSTER

Título del proyecto:
Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos
Línea de Investigación:
Ciencias de la ingeniería aplicadas a la producción, sociedad y desarrollo sustentable
Campo amplio de conocimiento:
Tecnologías de la Información y la Comunicación (TIC)
Autor/a:
Santiago David Lucio Cruz
Tutor/a:
Renato Mauricio Toasa Guachi **Mario Ruben Pérez Cargua**

Quito – Ecuador

2025

APROBACIÓN DEL TUTOR

Yo, Mario Ruben Pérez Cargua portador de la C.I: 0603251984 en calidad de Tutor del trabajo de investigación titulado: Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos.

Elaborado por: Santiago David Lucio Cruz, de C.I: 0502733371, estudiante de la Maestría: Big Data y Ciencia De Datos, de la **UNIVERSIDAD TECNOLÓGICA ISRAEL (UISRAEL)**, como parte de los requisitos sustanciales con fines de obtener el Título de Magister, me permito declarar que luego de haber orientado, analizado y revisado el trabajo de titulación, lo apruebo en todas sus partes.

Quito D.M., marzo de 2025

Firmado electrónicamente por:
MARIO RUBEN PEREZ CARGUA

Firma

APROBACIÓN DEL TUTOR

Yo, Renato Mauricio Toasa Guachi portador de la C.I: 1804724167 en calidad de Tutor del trabajo de investigación titulado: Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos.

Elaborado por: Santiago David Lucio Cruz, de C.I: 0502733371, estudiante de la Maestría: Big Data y Ciencia De Datos, de la **UNIVERSIDAD TECNOLÓGICA ISRAEL (UISRAEL)**, como parte de los requisitos sustanciales con fines de obtener el Título de Magister, me permito declarar que luego de haber orientado, analizado y revisado el trabajo de titulación, lo apruebo en todas sus partes.

Quito D.M., marzo de 2025

Firmado electrónicamente por:
RENATO MAURICIO
TOASA GUACHI

_____

Firma

DEDICATORIA

El presente trabajo va dedicado primero a Dios, por haberme puesto en el camino correcto para poder culminar este peldaño en mi vida profesional. A mis padres que siempre guían mi camino en cada paso que doy, dándome el respaldo necesario para cada día salir adelante. A mi esposa e hijas quienes son mi motivación y energía para cada día ser una mejor persona y un profesional de excelencia y de esta manera poder cubrir su camino y guiarlas para que superen mi camino recorrido.

AGRADECIMIENTO

A los catedráticos de la Universidad Tecnológica Israel por el conocimiento impartido durante las horas de clase, orientándome para la consecución de este logro alcanzado.

A mis hermanos que con su ejemplo y consejos me animaron a enrumbarme nuevamente en este camino que está llegando a su fin.

DECLARACIÓN DE AUTORIZACIÓN POR PARTE DEL ESTUDIANTE

Yo, Santiago David Lucio Cruz con C.I: 0502733371, autor/a del proyecto de titulación denominado Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos. Previo a la obtención del título de Maestría En Big Data y Ciencia De Datos.

1. Declaro tener pleno conocimiento de la obligación que tienen las instituciones de educación superior, de conformidad con el Artículo 144 de la Ley Orgánica de Educación Superior, de entregar el respectivo trabajo de titulación para que sea integrado al Sistema Nacional de Información de la Educación Superior del Ecuador para su difusión pública respetando los derechos de autor.
2. Manifiesto mi voluntad de ceder a la Universidad Tecnológica Israel los derechos patrimoniales consagrados en la Ley de Propiedad Intelectual del Ecuador, artículos 4, 5 y 6, en calidad de autor@ del trabajo de titulación, quedando la Universidad facultada para ejercer plenamente los derechos cedidos anteriormente. En concordancia suscribo este documento en el momento que hago entrega del trabajo final en formato impreso y digital como parte del acervo bibliográfico de la Universidad Tecnológica Israel.
3. Autorizo a la SENESCYT a tener una copia del referido trabajo de titulación, con el propósito de generar un repositorio que democratice la información, respetando las políticas de prosperidad intelectual vigentes.

Quito D.M., marzo de 2025

Firmado electrónicamente por:
SANTIAGO DAVID LUCIO CRUZ

_____

Firma

Tabla de contenidos

Índice de tablas

Índice de figuras

INFORMACIÓN GENERAL

Contextualización del tema

En la actualidad se comete una amplia variedad de tipos de fraude, donde nos encontramos con fraudes telefónicos, fraudes electrónicos por las redes electrónicas, fraudes personales, y otros tantos, que van a afectar el flujo de caja de las organizaciones, en la última década se ha incrementado el fraude electrónico el que ha ido a provocar un colapso mundial, ya que este problema no discrimina si la organización es grande o pequeña sólo le interesa cometer el delito, nos centraremos en transferencias electrónicas de dinero, a través de portales web o aplicaciones móviles que involucran directamente a las instituciones financieras que las proporcionan, que han facilitado la vida de las personas agilizando su día a día, mejorando la comodidad y accesibilidad del usuario a este cambiante mundo tecnológico, pero se ven involucrados indirectamente cuando sucede algún tipo de fraude cualquiera que este sea. Las instituciones financieras que carecen de un sistema de prevención de fraude se convierten en vulnerables a daños financieros deteriorando la confianza entre los clientes.

Los bancos han visto un aumento en el fraude, ya que los empleados robaron credenciales para enviar dinero a cuentas personales. Estos eventos no solo han causado daños financieros reales, sino que también han dañado seriamente la imagen de las instituciones. Los clientes afectados en ciertos escenarios dan aviso a los entes de control poniéndolos en alerta e incluyendo en su radar a esta institución, lo que conlleva a inspecciones generales que examinan las medidas de seguridad existentes y, simplemente el hecho de que esto suceda, las actividades pueden dañar la confianza entre los clientes y acarrear fuga de clientes a otras instituciones con seguridades más robustas

La iniciativa empleará datos simulados abiertamente accesibles, como el conjunto de datos encontrado en la “Plataforma Nacional de Datos Abiertos” del Perú en la sección de “Transferencias de fondos y asignaciones financieras” comúnmente utilizados para transparencia y tienen una licencia abierta y de acceso público. Estos datos serán ofuscados incluyendo datos no reales con el objetivo de obtener información para el análisis.

En este escenario, la aplicación de algoritmos de autoaprendizaje para la prevención del fraude pudiera presentar una respuesta a esa seguridad que las instituciones pueden estar buscando. A diferencia de las técnicas convencionales, que generalmente dependen de las pautas establecidas, los marcos de autoaprendizaje pueden evolucionar y mejorar con información y tendencias nuevas. Este punto es extremamente importante ya que en un

ambiente donde las estafas están a la orden del día se vuelve necesaria la implementación de un mecanismo de aprendizaje automático puede detectar irregularidades en las actividades de transacción, distinguiendo entre transacciones auténticas y aquellas potencialmente fraudulentas.

Problema de investigación

La problemática de la investigación se enfoca en la importancia de desarrollar un potente sistema de detección de fraude que permita la identificación de transacciones posiblemente sospechosas y permita tomar acciones inmediatas de las mismas en instituciones financieras, ya que, el ser víctima de estos eventos fraudulentos no solo pudieran resultar en pérdidas económicas, sino que también pudieran llegar a dañar la reputación de la institución y por ende pueden llevar a intervenciones regulatorias por parte de los entes de control.

El no contar con un proceso de detección de fraude en esta era digital, es una desventaja predominante en la seguridad de una institución. Los métodos manuales y tradicionales que usualmente se usan para la detección de transacciones fraudulentas, están demostrando cada día su ineficacia frente a el avance de las maneras sofisticadas que están utilizando los estafadores hoy en día. Adicional a esto, las personas utilizan con más frecuencia diferentes medios digitales como transferencias bancarias, incrementando el volumen de transacciones de las mismas y haciendo prácticamente imposible la detección de una transacción fraudulenta en medio de tanta información o data generada.

Este proyecto de investigación se va a enfocar en la implementación de un modelo predictivo basado en técnicas de aprendizaje automático que pueda ser capaz de analizar el comportamiento de las transacciones, y poder identificar transacciones que necesiten ser intervenidas o revisadas por los responsables de un área específica, para así determinar si una operación es legítima o posiblemente fraudulenta o que necesite confirmarse. El modelo a ser aplicado debería ser lo suficientemente inteligente para poder adaptarse a nuevas actualizaciones de fraude a medida que vayan siendo desarrolladas, y lo suficientemente eficiente como para poder trabajar en un ambiente donde las decisiones deben tomarse en medio de cantidades enormes de información

Objetivo general

Realizar el análisis del proceso de Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos.

Objetivos específicos

- Contextualizar los fundamentos teóricos sobre algoritmos de aprendizaje que se acoplen a los indicadores definidos y en base a este, estudiar el comportamiento histórico de las transacciones de clientes.
- Diagnosticar la situación actual referente a fraude en las transacciones de clientes.

- Realizar el análisis de fraude aplicando un modelo de aprendizaje automático
- Evaluar la propuesta mediante el criterio de especialistas en Big Data y Ciencias de Datos.

Vinculación con la sociedad y beneficiarios directos:

La implementación del proyecto " Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos" puede resultar necesario para poder solventar o mitigar todas las necesidades que se incrementan día a día en las instituciones financieras. Vivimos en tiempos en los que las constantes olas de fraude utilizando cada día más y mejores técnicas para realizarlas, están causado una pérdida de confianza entre los clientes y un posible retroceso en el uso de la tecnología, en especial por personas que no cambian al paso acelerado en el que lo realiza la tecnología, esto tarde o temprano resultará en una fuga significativa de capital y en el deterioro de la reputación institucional.

Este proyecto tiene como objetivo analizar un modelo que permita elevar la confianza de los clientes al analizar sus transacciones históricas diagnosticando su situación frente al fraude, alertando mediante un dashboard histórico que identifique las fases de aprendizaje del algoritmo, con información previa y posterior mostrando indicios sobre posibles fraudes y por ende mejorando la seguridad de las transacciones. La implementación del modelo a analizar en un entorno real puede proteger eficazmente a los clientes contra fraudes y por ende las instituciones no solo protegerían sus intereses financieros sino también consolidarían su posición en el mercado.

Esta iniciativa a p arte de mejorar la seguridad y por ende mejorar la confianza de los clientes, apoyará al mismo tiempo a los objetivos de desarrollo sostenible(ODS), concretamente el ODS 8: Trabajo decente y Crecimiento Económico al mejorar la seguridad y la confianza del tratamiento financiero, lo cual es un factor crítico en una expansión económica consistente y permanente.

Esta iniciativa respalda el ODS 9: Industria, Innovacion e Infraestructura ya que promueve la innovación a través de tecnologías avanzadas como el aprendizaje automático para prevenir el fraude y mejora el marco técnico de la organización.

En última instancia, apoyamos el ODS 16: Paz Justicia e Instituciones Sólidas, promoviendo la transparencia, la justicia y el establecimiento de compañías financieras confiables. Todas estas orientaciones enfatizan la importancia y el impacto de las iniciativas en la realización de la región fiscal y el mundo de ODS (ONU, 2015).

CAPÍTULO I: DESCRIPCIÓN DEL PROYECTO

1.1. Contextualización general del estado del arte

En la transformación digital que estamos siendo participes al tener a nuestra disposición tecnologías como transacciones en línea, la seguridad financiera se ha convertido en la principal preocupación para las instituciones como para sus usuarios. Varios autores abordan el problema de la detección de fraude a través de técnicas de análisis de datos avanzados y modelado predictivo. Por ejemplo, Dal Pozzolo, Caelen, Le Borgne, Waterschot y Bontempi (2014) mostraron que los patrones de fraude se pueden determinar en los sistemas tradicionales mediante el uso de algoritmos de aprendizaje automatizados como bosques y redes neuronales (Dal Polo, 2014), 2014 no tiene piedad. Bhattacharya, JHA, Tharkunnel y Westland (2011) señalaron que una combinación de métodos estadísticos y tecnologías de inteligencia artificial es altamente efectiva para combatir el problema del desequilibrio entre las clases de registros de datos financieros para mejorar las tasas de detección y reducir los falsos positivos (Bhattachayyyya et al., 2011).

En tecnología e informática se entiende como el área de conocimiento denominada Apredizaje Automatico o machine learning; muy cercano a lo denominado como inteligencia artificial (IA); la finalidad de esta técnica es que las computadoras puedan aprender ya que se considera un agente que aumente la experiencia; el aprendizaje automático ha encontrado gran utilidad, sobre todo en el análisis de investigaciones y procesos que generan grandes cantidades de datos; para este artículo se hace una revisión documental sobre el estado del arte mediante los principales métodos de machine learning, en base a publicaciones y artículos que no tienen más de dos años de antigüedad, para así poder conocer conceptos, identificar y entender el funcionamiento de los diversos tipos de técnicas de Machine Learning con vistas a la detección de fraudes financieros.

De esta forma, el estado del arte muestra una línea clara hacia la implementación de sistemas híbridos y de aprendizaje continuo. No obstante, surgen interrogantes sobre la validez externa de estos modelos y la necesidad de profundizar en el análisis cualitativo de los errores del sistema. En este sentido, el presente proyecto se propone no sólo mejorar la detección de fraudes a partir de un algoritmo Random Forest (Breiman, 2001), sino también integrar un proceso de validación cruzada riguroso y una visualización detallada de los resultados que incluya el “antes y después” del aprendizaje de la máquina.

1.2. Proceso investigativo metodológico

1.2.1. Enfoque de la Investigación

El presente estudio utiliza un enfoque cualitativo para la recolección y análisis de datos. Este enfoque no sigue métodos específicos u organizados, por lo que son utilizados en los procesos sociales. También es una vía de investigación que no necesariamente implica mediciones numéricas, se toman entrevistas, encuestas, las opiniones de los investigadores y descripciones. En este enfoque cualitativo, la persona que investiga realiza primero el estudio del contexto con el que trabajará, luego observa el fenómeno de estudio y después prosigue con teorías según de lo que observa. Sus métodos son basados en inductivos (fenomenología, la etnografía) o interpretativos (teoría critica, feminismo, constructores personales, psicología). Según el ritmo con el que se desenvuelve la investigación, no se trabaja con la prueba de hipótesis por lo que estas son desarrolladas y otra característica de este enfoque es que es humanista (Course hero, sf).

El diseño de muestra el cual ha sido adaptado para este proyecto es la muestra no probabilística, se lo seleccionó debido a que la elección de los elementos de la muestra no está determinada por la probabilidad, sino por causas relacionadas a las características de la investigación o los objetivos del investigador (Hernández-Sampieri et al. 2018).

Este enfoque metodológico cualitativo enfocado en entrevistas estructuradas abre la posibilidad a un análisis profundo de las experiencias y formas de ver de experimentados en seguridad financiera, punto que es valioso para poder comprender que tan efectivo resulta el sistema de detección de fraudes.

La elección de una muestra no probabilística en este proyecto se encuentra justificada debido a la escasez de profesionales especializados en esta área en el medio en el que se pretende su implementación, debido a esto se estará obligado a selección de profesionales específicos en la materia, garantizando que la información obtenida sea altamente relevante para el estudio, lo que hará posible la validación y constante mejora de la propuesta en el contexto de Big Data y Ciencias de Datos.

Esta entrevista "cualitativa", es una conversación fluida donde uno de los participantes reflexiona y revive su vida, ante la escucha atenta y cuasi invisible del entrevistador. Se enfoca aquí como un recurso insustituible porque logra la descripción del mundo desde la perspectiva histórica de quien la ha vivido directamente, es especial, los sectores menos privilegiados de la sociedad que han sido olvidados por la historia oficial. En este tipo de entrevistas, el investigador

debe poseer al menos, cinco cualidades básicas: identificación con su trabajo, honestidad, confianza, naturalidad y curiosidad. Finalmente, se describen los procedimientos para realizarla (Fernández Carballo, R. (2013)).

1.2.2. Población y Muestra

Para la determinación de población y muestra del proyecto, y dada a la escasez de personas conocedoras del tema en el sector financiero, que posean larga experiencia en el sector financiero y que además estén en la lista de contactos del desarrollador del proyecto, la población y muestra se centró en cuatro profesionales del sector financiero que trabajan en una misma institución. Al armar este grupo compacto con el mismo conocimiento de fondo que se mueva en un mismo entorno tecnológico, nos aseguramos de que la información recopilada responda a las características especiales del medio ambiente y proporcione una visión integral y cualitativa del proyecto a realizarse.

1.2.3. Instrumentos

Debido a que este estudio es parte de un paradigma cualitativo, el instrumento principal utilizado fueron las entrevistas. En este modelo, las preguntas se definieron previamente, por lo que todas las entrevistas podrían ser un aspecto sistemático y relevante de la propuesta. Para este propósito, se desarrolló una guía de entrevista semiestructurada, como se muestra en el Anexo 1. Esta guía proporciona muchas preguntas específicas para extraer más información de cada experto. La estructura de la entrevista facilita la comparación de respuestas e identificación de patrones de opiniones que contribuyen a la verificación cualitativa de la propuesta.

1.3. Análisis de resultados

Para los resultados se realizó un análisis en base a entrevistas realizadas a cuatro expertos en ramas de la seguridad financiera y ciencias de datos, mismos que se los puede encontrar en el Anexo 1 ("Guía de Entrevista Semiestructurada – Expertos en Seguridad Financiera y Ciencia de Datos"). Cada una de las preguntas principales se analizó para identificar diferencias relacionadas con las tendencias generales y las diferencias relacionadas en las percepciones de los encuestados.

Pregunta 1: Esta pregunta se enfocó en la percepción que se tiene sobre qué tan eficaz resulta la utilización del Algoritmo Random Forest, mismo que se está utilizando en el proyecto

Pregunta 2: Evaluación del Proceso de Limpieza y Preprocesamiento de Datos

Pregunta 3: Utilidad de la Integración de Herramientas de Visualización (PowerBI).

Tabla 1

Cuadro resumen respuestas de encuestas

Pregunta	Respuestas	Análisis
¿Cuál es su percepción de la efectividad del algoritmo Random Forest aleatorio para reconocer el fraude en las transacciones financieras?	El bosque aleatorio es extremadamente efectivo en la detección temprana de fraude, ya que le permite identificar patrones desconectados en grandes registros de datos.	La mayoría de los expertos enfatizan la efectividad de los bosques aleatorios y la robustez y la capacidad de usar datos desproporcionados. En opinión, se observa que la convergencia respalda la selección de este algoritmo en el proyecto para la detección de fraude
	El algoritmo tiene una notable robustez, le es posible adaptarse de forma adecuada a una gran variabilidad de transacciones financieras que se encuentran disponibles y son un común entre todas las instituciones financieras	
	La estructura de múltiples árboles que es la característica de Random Forest es la principal característica que le permite reducir significativamente los errores de predicción durante su utilización y ejecución	
	La eficiencia y la capacidad de aprendizaje continuo del modelo son fundamentales para detectar fraudes	
¿Qué valoración otorga al proceso de limpieza y preprocesamiento de datos utilizado en el proyecto?	La doble eliminación y la estandarización son extremadamente importantes para garantizar la calidad de los datos	Es un común de los entrevistados que en base a su experiencia y estudios dan un peso valioso a los procesos de limpieza de datos y preprocesamiento en el éxito o fracaso del sistema. Sus respuestas reflejan una alta calificación de estos procedimientos, reforzando la importancia de las bases de datos uniformes e inconsistentes
	El preprocesamiento estricto es la base para el funcionamiento eficiente de los modelos de bosques aleatorios	
	una preparación adecuada de los datos es esencial para eliminar ruidos y asegurar la relevancia de la información	
	La metodología aplicada en la limpieza aportó claridad y solidez a la entrada de datos	
¿Cómo valora la utilidad de PowerBI para la visualización y análisis de los resultados del modelo?	PowerBI facilita la interpretación de datos y permite un seguimiento continuo del rendimiento del modelo	Todos los participantes están de acuerdo en que usar PowerBI es una estrategia invaluable para visualizar los resultados. Esto no solo le permite monitorear la salida del modelo, sino que también facilita la comunicación de resultados y apoya los procesos de decisión a nivel de gestión.
	La interacción visual es fundamental para comprender la evolución del sistema en tiempo real	
	la personalización del dashboard en PowerBI resulta muy útil para extraer insights relevantes	
	la claridad en la presentación de resultados permite realizar ajustes estratégicos cuando es necesario	

En términos generales, el análisis cualitativo de las entrevistas pone de manifiesto que los expertos en este ámbito de seguridad financiera consideran de una manera positiva y consensuada tanto la eficacia del algoritmo Random Forest y la metodología utilizada en el preprocesado de los datos. Junto a ello, la vinculación de PowerBI se debe considerar un componente estructural a la hora de interpretar y analizar los resultados. Estos resultados, pues, respaldan la propuesta del proyecto y de esta manera se pone de manifiesto la conveniencia de una aproximación cualitativa para recoger información en un ámbito como el del presente estudio de forma contextualizada y detallada.

CAPÍTULO II: PROPUESTA

2.1. Fundamentos teóricos aplicados

Para poder justificar el algoritmo a ser utilizado en el proyecto, a continuación, se muestra un cuadro comparativo en el que se muestran los principales algoritmos utilizados hoy en día, junto con un análisis de cada uno de ellos y así poder llegar a la selección correcta.

Tabla 2

Cuadro Comparativo de Algoritmos

Algoritmo	Ventajas	Desventajas
Random Forest	- **Robustez:** Reduce el sobreajuste mediante la agregación de múltiples árboles (Breiman, 2001). - **Manejo de datos ruidosos y desequilibrados:** Funciona eficazmente en escenarios con clases desbalanceadas (Bhattacharyya et al., 2011). - **Facilidad de ajuste:** Requiere menos tuning de hiperparámetros en comparación con modelos complejos. - **Paralelismo:** Los árboles se pueden entrenar de forma independiente, lo que facilita el escalado.	- **Consumo de recursos:** Puede requerir gran capacidad de procesamiento y memoria cuando se trabaja con volúmenes muy altos de datos. - **Interpretabilidad:** La interpretación de los resultados es menos directa en comparación con modelos lineales.
Support Vector Machine (SVM)	- **Eficiente en alta dimensionalidad:** Funciona bien cuando el número de características es elevado (Cortes & Vapnik, 1995). - **Separación clara:** La utilización de distintos núcleos para la clasificación permite una buena separación de clases.	- **Escalabilidad:** No es la opción más eficiente para datasets muy grandes, ya que implica un alto consumo computacional. - **Ajuste complejo:** La elección del tipo de núcleo y de parámetros requiere un ajuste cuidadoso y puede ser complicado.
K-Nearest Neighbors (KNN)	- **Simplicidad:** Es fácil de entender e implementar. - **Sin entrenamiento explícito:** Se basa en medidas de distancia para clasificar, sin requerir una fase de entrenamiento extensa.	- **Lento en predicción:** Con grandes volúmenes de datos, la predicción puede volverse ineficiente. - **Sensibilidad a escalas:** Altamente afectado por características irrelevantes y por la escala de las variables.
Logistic Regression	- **Rápida y sencilla:** Su implementación es directa, con un alto grado de interpretabilidad. - **Eficiente en problemas lineales:** Funciona bien cuando las relaciones entre las variables son lineales.	- **Capacidad limitada:** requiere un gran tamaño de muestra y datos independientes y distribuidos de forma idéntica. - **Sensibilidad a datos desbalanceados:** puede ser sensible a los valores atípicos y al ruido, por lo que debe detectarlos y eliminarlos o tratarlos.
Modelos de Deep Learning	- **Potencial en complejidad:** Son capaces de capturar patrones complejos en datos de gran volumen (LeCun, Bengio, & Hinton, 2015). - **Adaptabilidad:** Se pueden ajustar para trabajar con diversos tipos de datos y problemas.	- **Requerimientos elevados:** Necesitan grandes volúmenes de datos y recursos computacionales significativos. - **Interpretabilidad:** Los modelos son a menudo considerados como "cajas negras" debido a la dificultad de interpretar internamente sus decisiones.

Después del análisis, el algoritmo seleccionado del bosque aleatorio es el algoritmo principal que reconoce el fraude de transacciones financieras por una variedad de razones técnicas y prácticas:

1. **Robustez y precisión:** combina varias decisiones para tomar madera a través de técnicas de bolsas para reducir significativamente el riesgo de voladizo y aumentar la estabilidad de la clasificación (Breiman, 2001; Bhattacharyya et al., 2011). Esta robustez es esencialmente importante para detectar patrones sutiles en entornos con datos fuertes y desequilibrados que son característicos del sistema financiero.

2. **Gestión inevitable del ruido:** el fraude es una parte muy pequeña de la transacción total, por lo que es importante tener un algoritmo que pueda manejar este desequilibrio sin usar técnicas de reenvío extensas. Random Forest demuestra su efectividad en estos escenarios y también ofrece alta precisión en condiciones de datos desequilibradas (Dal Pozzolo et al., 2014).

3. **Implementación y verificación simples:** la adaptación simple del hiperparámeno y la posibilidad de implementar técnicas de validación cruzada (validación cruzada de K-compatibilidad) nos permitirá evaluar y certificar el rendimiento del modelo (PHUA, Lee, Smith y Gayler, 2010).

4. **Escalabilidad y paralelismo:** Con la opción de capacitar independientemente entre sí, cada árbol escala fácilmente el modelo cuando se trabaja con grandes cantidades de datos, lo que lo hace adecuado para aplicaciones en entornos de Big Dad con eficiencia cuando el procesamiento es un factor crítico.

5. **Integración con herramientas de visualización:** la solidez inherente del modelo y la extracción simple de las métricas de rendimiento permite una integración efectiva con herramientas de visualización como PowerBI. Esto facilita la interpretación de los resultados y el monitoreo continuo del rendimiento, y demuestra claramente la efectividad del aprendizaje del modelo en la detección de fraude.

La selección de Random Forest está basada en su potencial para detectar complejas relaciones en el contexto de conjuntos de datos desbalanceados, su robustez, su capacidad de validación y su capacidad de integración con el trabajo en un sistema de análisis interactivo. Estas ventajas hacen de Random Forest la mejor elección para la implementación del sistema de detección de fraudes que se propone realizar en el marco de este proyecto.

La propuesta se apoya sobre diversas áreas del conocimiento que van desde la Big Data hasta las avanzadas técnicas de machine learning. Entre los conceptos principales se destacan:

- **Big Data:** Término que se refiere a conjuntos masivos de datos con una estructura extensa, variada y compleja, cuyas dificultades para almacenarlos, analizarlos y visualizarlos permiten obtener resultados o procesos posteriores.
- **Ciencia de Datos:** Es considerada un enfoque novedoso y promisorio empleado en la obtención y análisis de información en múltiples disciplinas científicas.
- **Random Forest:** Método de aprendizaje conjunto para clasificación, regresión y otras tareas. Funcionan mediante la creación de múltiples árboles de decisión durante el entrenamiento.
- **Validación Cruzada:** Técnica estadística que permite estimar la capacidad predictiva del modelo al dividir los datos en varios pliegues y promediar los resultados obtenidos.

Los parámetros a utilizarse en el proyecto durante sus distintas fases de programación se detallan en el siguiente cuadro explicativo justificativo

Tabla 3

Detalle de variables Random Forest

Parámetro	Valor Configurado	Descripción	Justificación del Valor
n_estimators	50	Este parámetro sirve para determinar el número de árboles en el bosque. Mientras más árboles existan en el modelo, generalmente mejoran la precisión, pero aumentan el tiempo de cómputo.	Se ha configurado un valor de 50 ya que logra un equilibrio entre precisión y eficiencia computacional, esto es usado para manejar grandes volúmenes de datos.
max_depth	10	Este parámetro sirve para determinar la profundidad máxima de cada árbol. Limitarla ayuda a evitar el sobreajuste (overfitting).	Se ha configurado un valor de 10 lo que permite generalizar bien los datos sin perder precisión al clasificar las transacciones como fraudulentas.
random_state	42	Este parámetro es una semilla que se utiliza para generar números aleatorios. El parámetro garantiza que los resultados del	Se ha configurado un valor de 42, que es un valor estándar ampliamente usado para reproducibilidad

		modelo sean reproducibles.	y consistencia en experimentos.
criterion	Predeterminado: gini	Este parámetro es una función que se usa para medir la calidad de las divisiones (impureza de Gini).	Este parámetro no se ajustó, pero gini es una métrica eficiente para optimizar divisiones en problemas de clasificación binaria como detección de fraudes.
bootstrap	Predeterminado: True	Este parámetro indica si se utilizan muestras bootstrap (muestras con reemplazo) al construir cada árbol.	Este parámetro se configuró en,True lo cual asegura que cada árbol sea construido a partir de datos diferentes, esta configuración va mejorando la diversidad entre árboles del bosque.
min_samples_split	Predeterminado: 2	Este parámetro muestra el número mínimo de muestras necesarias para dividir un nodo.	El haber configuradp este parámetro en 2 permite que las divisiones se realicen con pocos datos, maximizando la capacidad de ajuste en conjuntos pequeños.
min_samples_leaf	Predeterminado: 1	Este parámetro muestra el número mínimo de muestras requeridas en un nodo hoja.	El haber configurado este valor en 1 asegura que el modelo pueda capturar patrones finos en datos con desequilibrio, como el análisis de fraudes financieros.
max_features	Predeterminado: auto (raíz cuadrada de características)	Este parámetro muestra el úmero máximo de características consideradas para buscar la mejor división en cada nodo.	Este valor auto optimiza el balance entre sesgo y varianza, reduciendo la complejidad de cada árbol sin afectar mucho la precisión general.

2.2. Descripción de la propuesta

A manera de propuesta para este proyecto se plantea el desarrollo de un sistema integral para la detección de fraude en transacciones financieras. Un mayor detalle que se encuentra incluido en el ANEXO 2, muestra el proceso completo de detección de fraude en sus fases completas, a manera de resumen, a continuación, se desglosan los componentes del producto:

2.2.a. Estructura General

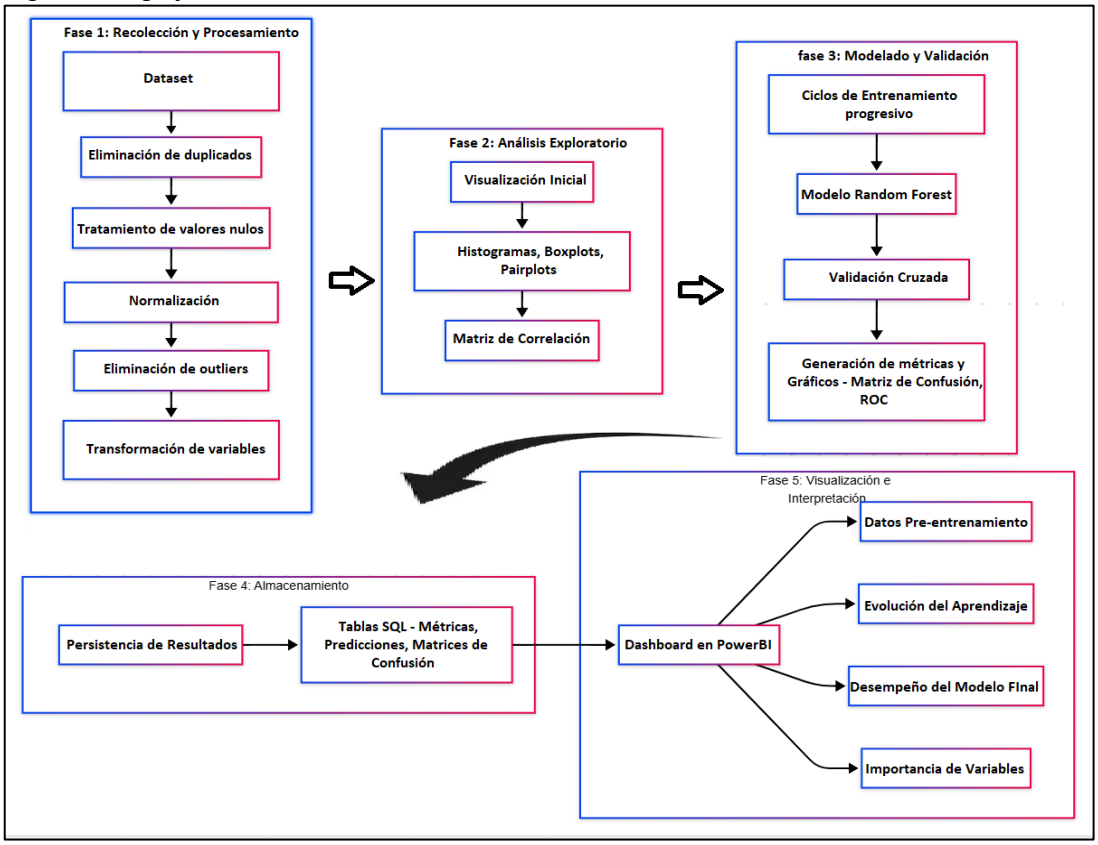
El sistema se compone de las siguientes capas:

- **Capa de Recolección y Preprocesamiento:** En esta capa se realiza el proceso de importación del dataset y se realiza un proceso de limpieza y normalización dejándolo listo para en lo posterior realizar el análisis exploratorio correspondiente.
- **Capa de Modelado:** Esta es la parte donde se pone a trabajar al modelo elegido, implementándolo como primera fase y en lo posterior efectuando el entrenamiento respectivo para al final realizar el proceso de validación cruzada.
- **Capa de Almacenamiento:** Esta es la capa que recolecta la información procesada, recopilando datos de los resultados y métricas en una base de datos SQLite para su posterior consulta.
- **Capa de Visualización:** Esta es la parte que se encuentra de cara al usuario final, consiste en la implementación de dashboards interactivos desarrollados en PowerBI que permiten analizar gráficamente los resultados del modelo.

A continuación, se presenta un organizador gráfico que resume la estructura general:

Figura 1

Organizador gráfico



2.2.b. Explicación del Aporte

En el proyecto en sus distintas fases, aparte de cumplir con su función específica, realiza un aporte determinado que se lo detalla a continuación por cada fase:

- **Recolección y Preprocesamiento:** Aporta al proyecto y a su usuario final, garantizando la calidad de la data e integridad completa del dataset, asegurando así que el modelo cumpla con su objetivo esencial y se entrene con datos homogéneos.
- **Modelado y Validación:** el aporte realizado en esta fase, tras la implementación del algoritmo Random Forest en conjunto con la aplicación de validación cruzada, proporciona la robustez y confiabilidad necesaria para una correcta detección de fraudes.
- **Almacenamiento:** La idea de realizar una estructuración de la información en una base de datos SQLite, facilita enormemente a futuras consultas, así como también a la integración con sistemas de visualización y reportería que se deseen implementar a futuro.
- **Visualización:** Quizás la parte más importante de todas las fases, la implementación del dashboard en PowerBI ofrece una potente herramienta interactiva para la interpretación de resultados, permitiendo visualizar el "antes y después" del entrenamiento del modelo brindando la información necesaria al personal encargado para realizar su análisis.

2.2.c. Estrategias y/o Técnicas

Durante el desarrollo del proyecto, se han utilizado las siguientes estrategias:

- **Estrategia de Preprocesamiento:** El preprocesamiento de datos es un paso fundamental para analizarlos, permite transformarlos sin procesar en un formato comprensible y utilizable para el análisis, garantiza que los datos estén preparados para las etapas posteriores de exploración, modelado e interpretación. Incluye también la limpieza de datos, el manejo de nulos y la normalización de datos.
- **Técnica de Modelado:** Implementación de Random Forest junto con validación cruzada (k=5) para evaluar el comportamiento del modelo en distintos escenarios.
- **Técnica de Integración:** Realizando la implementación y uso de una base de datos SQL (SQLite) como método de almacenamiento y, posteriormente fuente de presentación dinámica y comparativa, siendo de gran utilidad para el usuario final.

2.3. Validación de la propuesta

La validación de la propuesta se llevó a cabo a través de la aplicación de criterios de los especialistas de la disciplina de las ciencias de datos. Se realizaron entrevistas semiestructuradas sobre los expertos, se explicaron las maneras de funcionamiento del sistema y se recogieron las retroalimentaciones que se presentan en el ANEXO 3.

Los criterios de validación fueron, entre otros, la aplicabilidad del modelo ante la variabilidad de los datos, la viabilidad de integración del sistema con plataformas de análisis y la capacidad en la detección temprana de fraudes. A los especialistas les parecía que la combinación de un modelo de Random Forest junto con un análisis interactivo de resultados es una solución integral y novedosa.

Para la elección de especialistas se ha considerado un perfil acorde a los siguientes criterios: formación académica relacionada con el tema investigativo, experiencia académica y/o laboral orientada a la gestión administrativa y motivación para participar. La siguiente tabla presenta información detallada de los actores seleccionados para la validación del modelo.

Tabla 4

Descripción de perfil de validadores

Nombres y Apellidos	Años de experiencia	Titulación Académica	Cargo
Rolando Ochoa	18	Tecnólogo	Coordinador de base de datos y data analítica
Jorge Enríquez	9	Magister en Inteligencia de Negocios y Analítica de Datos Masivos	Administrador de Data analítica

A continuación, los criterios de los especialistas y sus conclusiones y recomendaciones

Tabla 5

Validación de especialistas

ESPECIALISTA	CALIF.	OBSERVACIONES	RECOMENDACIONES
Rolando Ochoa	30/35	Resultaría de gran utilidad el tener una solución como esta ya que brindaría una herramienta clave para el giro del negocio alertando a las áreas de primera línea sobre posibles transacciones fraudulentas y poniendo principal atención en ellas	Se recomienda realizar una propuesta para que la institución la analice y realizar dicha implementación de forma inmediata
Jorge Enríquez	32/35	El análisis que se pudiera conseguir con este proyecto empataría correctamente con el giro que la institución está presentando al tornarse en 100% digital, el análisis bridaría una herramienta que se tornaría crucial para el día a día en el trabajo del área, girando en torno a esta información y adaptar los subsistemas de TI a ella	La implementación se la debería realizar en la nube privada de la institución para de esta forma proteger de manera efectiva los datos que se pretenden analizar

2.4. Matriz de articulación de la propuesta

A continuación, se presenta la matriz de articulación que sintetiza la relación entre los sustentos teóricos, metodológicos, estratégicos-técnicos y tecnológicos empleados, y los resultados obtenidos.

Tabla 6

Matriz de articulación

EJES O PARTES PRINCIPALES	SUSTENTO TEÓRICO	SUSTENTO METODOLÓGICO	ESTRATEGIAS / TÉCNICAS	DESCRIPCIÓN DE RESULTADOS	INSTRUMENTOS APLICADOS
Recolección y Preprocesamiento	Según Chen, Chiang y Storey (2012), la recolección y el procesamiento de datos son fundamentales en el manejo de Big Data, ya que permiten transformar grandes volúmenes de datos en información valiosa mediante la normalización y eliminación de sesgos. Esto asegura que los datos sean coherentes y estén listos para análisis posteriores.	Hernández, Fernández y Baptista (2014) indican que la recolección y el procesamiento de datos pueden abordarse de forma sistemática a través de métodos cualitativos (como la revisión documental y la observación), garantizando la calidad de la información para estudios exploratorios.	Uso de Pandas para la manipulación de datos y StandardScaler para la normalización.	Datos homogéneos y sin duplicados, lo que facilita el entrenamiento y mejora la precisión del modelo.	Scripts en Python, Jupyter Notebook
Modelado y Validación	Según Breiman (2001), la aplicación de algoritmos de ensamblaje, como Random Forest, mejora la precisión y reduce el sobreajuste al combinar múltiples árboles de decisión. Esta técnica es especialmente útil en escenarios donde los datos presentan un alto grado de variabilidad y desequilibrio.	Kuhn y Johnson (2013) destacan que el modelado y la validación se pueden llevar a cabo mediante técnicas estructuradas, como la validación cruzada (k-fold), complementadas con métodos cualitativos para evaluar la interpretación de los modelos a partir de entrevistas y revisiones detalladas.	Implementación de Random Forest y validación cruzada (k=5) utilizando Scikit-learn.	Modelo robusto y preciso en la detección de fraudes, validado mediante técnicas de cross-validation y entrevistas estructuradas.	Scikit-learn, Python

Almacenamiento de Resultados	De acuerdo con Elmasri y Navathe (2015), el almacenamiento estructurado mediante sistemas de bases de datos relacionales es esencial para gestionar información proveniente de grandes volúmenes de datos y facilitar su posterior consulta y análisis. Esto es vital para proyectos que requieren integración de resultados en tiempo real.	Hernández et al. (2014) señalan que el almacenamiento de datos se puede abordar utilizando soluciones SQL, que permiten una integración eficiente y una organización clara de los resultados, facilitando análisis adicionales posteriores.	Integración con SQLite a través de SQLAlchemy para una gestión estructurada de los resultados.	Los resultados se almacenan de forma estructurada en una base de datos, lo que permite un acceso rápido y consultas interactivas.	SQLite, SQLAlchemy
Visualización e Interpretación	Según Few (2009), la visualización de datos es fundamental para transformar información compleja en insights comprensibles y accionables, permitiendo la identificación de patrones relevantes y facilitando la toma de decisiones basada en datos claros y visualmente representados.	Hernández et al. (2014) enfatizan que las herramientas de visualización, como los dashboards, son eficaces para analizar y comunicar resultados de investigaciones cualitativas, permitiendo interpretar los datos y validar las conclusiones a través de representaciones gráficas interactivas.	Uso de PowerBI para el desarrollo de dashboards interactivos que integran y muestran las métricas y comparaciones “antes y después” del modelo.	Se genera un dashboard interactivo que permite visualizar la evolución del modelo y analizar el impacto de la detección de fraudes de forma clara y dinámica.	PowerBI Desktop

CONCLUSIONES

La investigación permitió profundizar en los fundamentos teóricos de diferentes algoritmos de aprendizaje, resaltando la importancia del algoritmo Random Forest para analizar indicadores importantes en el mundo financiero, permitió también verificar el comportamiento de las distintas variables que intervienen en el proceso. Dicha etapa fue la que permitió asentar el conocimiento del comportamiento de datos transaccionales, demostrando que una buena interpretación teórica afirma el modelo metodológico elegido.

El diagnóstico realizado mostró con claridad la importancia del problema, destacándose el marcado desbalance entre las transacciones legítimas y las fraudulentas. Mediante el análisis de datos, los patrones y las anomalías fueron muchos los que evidenciaron la importancia de adoptar sistemas robustos en la detección de fraudes evidenciándose de esta forma la necesidad de adoptar técnicas avanzadas para interpretar los datos transaccionales y detectar de manera automática las conductas sospechosas en un entorno complejo.

La aplicación del modelo Random Forest seguido de un importante proceso de pre-procesamiento y de validación cruzada, dieron lugar a un análisis correcto y detallado de las transacciones. El modelo demostró ser muy capaz de clasificar las transacciones en legítimas o fraudulentas poniendo en relevancia la importancia que tiene un aprendizaje iterativo y el ajuste progresivo a través de ciclos. Esto no sólo permitía una mayor precisión y estabilidad del sistema sino que además tuvo la capacidad de identificar los atributos más relevantes para la detección del fraude.

El diagnóstico y el análisis cualitativo de expertos corroboró la viabilidad y la aplicabilidad de la propuesta. Los expertos resaltaron sobre todo el proceso de pre-procesamiento, la aplicación del modelo Random Forest, su integración con herramientas de visualización como PowerBI. Estas consideraciones en conjunto ratifican que el sistema propuesto es viable y eficiente para la detección de fraudes en el contexto de un entorno financiero.

RECOMENDACIONES

Se aconseja llevar a cabo un mayor nivel de profundización de la revisión bibliográfica actualizada sobre algoritmos de machine learning y de estrategias de análisis de datos. Por ejemplo, se puede considerar la incorporación de estudios de casos actuales y comparativos tanto procedentes del mundo académico como del mundo profesional, ya que se informática a enriquecer el marco teórico permitido una mayor adecuación de la selección de indicadores y de la interpretación de la evolución histórica de las transacciones.

Se aconseja la puesta en marcha de un sistema de seguimiento en tiempo real, que actualice periódicamente el diagnóstico y se reutilicen orígenes de datos adicionales y de tipo multidimensional. Esta estrategia permitiría captar mejor la evolución y las nuevas tendencias en los patrones de fraude y responder entre otras cosas a los cambios en el comportamiento transaccional.

Se recomienda estudiar combinaciones de algoritmos y construir estrategias híbridas y enlaces con otros tipos de machine learning (incluidos los de deep learning) que permiten dar un mayor recorrido y una mayor amplitud a las capacidades del Random Forest. Asimismo, la optimización de los hiperparámetros y la puesta en marcha de procesos de actualización dinámica del modelo permitirán una mejor adaptación ante la evolución de los patrones fraudulentos.

Finalmente, se recomienda ampliar las rondas de evaluación mediante nuevas entrevistas y grupos de discusión con expertos, de modo que se puedan volver a validar los resultados y obtener sugerencias para futuras mejoras. La colaboración interdisciplinaria y la retroalimentación constante con profesionales de la actividad permitirá alimentar la propuesta y asegurarse que, efectivamente, dar respuesta a las necesidades emergentes del sector financiero.

BIBLIOGRAFÍA

Bhattacharyya, S., Jha, S., Tharakunnel, K., & Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems, 50*(3), 602–613. https://doi.org/10.1016/j.dss.2010.08.005

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly, 36*(4), 1165–1188. https://doi.org/10.2307/41703503

Credit Card Fraud Detection Dataset. (n.d.). *Kaggle*. Retrieved Month Day, Year, from https://www.kaggle.com/mlg-ulb/creditcardfraud *(Reemplaza "Month Day, Year" por la fecha de consulta real)*

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2015). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications, 41*(10), 4915–4928. https://doi.org/10.1016/j.eswa.2013.12.015

Cortés, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*(3), 273–297. https://doi.org/10.1007/BF00994018

Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). SAGE Publications.

Elmasri, R., & Navathe, S. B. (2015). *Fundamentals of Database Systems* (7th ed.). Pearson.

Few, S. (2009). *Now you see it: Simple visualization techniques for quantitative analysis*. Analytics Press.

Hernández, R., Fernández, C., & Baptista, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444. https://doi.org/10.1038/nature14539

Phua, C., Lee, V., Smith, K., & Gayler, R. (2010). A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*.

Course Hero. (s.f.). *La investigación científica surge de la necesidad del hombre de dar solución.* Course Hero. https://www.coursehero.com/es/file/p5pa6p0g/

ANEXO 1

INSTRUMENTO DE VALIDACIÓN

Universidad Israel

ESPOG | Escuela de Posgrados

UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

GUÍA DE ENTREVISTA SEMIESTRUCTURADA – EXPERTOS EN SEGURIDAD FINANCIERA Y CIENCIA DE DATOS

Esta guía fue diseñada para entrevistar a especialistas en seguridad financiera y ciencia de datos, con el objetivo de validar cualitativamente la propuesta del sistema de detección de fraude. A continuación, se listan las preguntas y respuestas obtenidas de cuatro expertos que trabajan en instituciones financieras.

Preguntas de Entrevista

1. **Pregunta 1:** ¿Cuál es su percepción sobre la efectividad del algoritmo Random Forest para la detección de fraudes en transacciones financieras?
2. **Pregunta 2:** ¿Qué opina de la metodología aplicada en el proceso de limpieza y preprocesamiento de datos en este proyecto?
3. **Pregunta 3:** ¿Considera que la validación cruzada aplicada (k=5) proporciona una estimación confiable del rendimiento del modelo?
4. **Pregunta 4:** ¿Qué tan útil le resulta la integración de la herramienta PowerBI para la visualización de los resultados del modelo?
5. **Pregunta 5:** ¿Cuáles son sus recomendaciones para mejorar aún más la propuesta de detección de fraude presentada?

Protección de datos: Los resultados de esta encuesta serán analizados con absoluta reserva y su tratamiento será de manera estadística y exclusivamente para los fines investigativos.

Aplicación a 4 Expertos

Entrevistado 1 – Ing. Armando Arce Jefe de Seguridad de la Información – Banco ProCredit

- **P1:** "El uso de Random Forest es adecuado, ya que ha mostrado alta efectividad en pruebas internas."
- **P2:** "La limpieza y normalización aplicadas son suficientes para garantizar datos de calidad, lo que redunda en un mejor desempeño del modelo."
- **P3:** "La estrategia de validación es correcta; 5 pliegues ofrecen una visión equilibrada del rendimiento."
- **P4:** "La integración con PowerBI expone de forma clara y concisa la evolución del modelo y facilita ajustes posteriores."
- **P5:** "Sería interesante probar con otros algoritmos y combinarlos con Random Forest para mejorar la detección de anomalías."

Entrevistado 2 – Mg. Maria Fernanda Ramos Analista de Riesgos – Banco ProCredit

- **P1:** "Considero que Random Forest es robusto para detectar fraudes, aunque siempre se debe evaluar en diversos escenarios."
- **P2:** "El preprocesamiento es crucial, y en este caso se ha ejecutado de manera satisfactoria."
- **P3:** "La validación cruzada realmente refuerza la confianza en el modelo."
- **P4:** "PowerBI permite una visualización que simplifica la interpretación incluso para usuarios no técnicos."
- **P5:** "Recomiendo la integración de datos en tiempo real para mejorar la capacidad de respuesta."

Entrevistado 3 – Mauricio Ochoa (Coordinador de Base de Datos y Data Analítica – Banco ProCredit)

- **P1:** "Random Forest ha demostrado ser muy efectivo en escenarios de detección de fraude, aportando estabilidad al modelo."
- **P2:** "La metodología de limpieza es acertada, eliminando ruido y duplicidades que podrían alterar los análisis."
- **P3:** "Estoy de acuerdo en que el método de validación cruzada es adecuado para asegurar la robustez del modelo."
- **P4:** "La herramienta PowerBI expone los resultados de forma tal que facilita la toma de decisiones inmediatas."

- **P5:** "Se podría complementar la solución actual con modelos híbridos para cubrir mejor los casos atípicos."

Entrevistado 4 – Mg. Jorge Enriquez (Administrador de Data analítica – Banco ProCredit)

- **P1:** "El algoritmo seleccionado es apropiado y se alinea con las mejores prácticas en análisis de fraudes."
- **P2:** "La estandarización de los datos es un paso crítico y se ha implementado correctamente."
- **P3:** "La validación cruzada con 5 pliegues es suficiente para conferir robustez al análisis."
- **P4:** "PowerBI facilita significativamente la interpretación de las métricas de evaluación."

P5: "Se sugiere la exploración de técnicas adicionales, especialmente en escenarios de gran volumen de datos."

ANEXO 2

PROPUESTA

IMPLEMENTACION DEL PROCESO DE DETECCION DE FRAUDE

1. Importación de Librerías

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split, cross_val_score, learning_curve
from sklearn.ensemble import RandomForestClassifier
from sklearn.preprocessing import LabelEncoder
from sklearn.metrics import classification_report, roc_curve, auc
from sqlalchemy import create_engine
import os
```

Se importan las librerías necesarias para la manipulación de datos (pandas, numpy), visualización (matplotlib, seaborn), aprendizaje automático (sklearn) y conexión a base de datos (sqlalchemy).

---

2. Configuraciones Iniciales

```
RANDOM_STATE = 42
CHUNK_SIZE = 50000  # Tamaño del bloque para leer los archivos
DB_CONNECTION_STRING = 'mssql+pyodbc://sa:Test1@QUINB000032/AnalisisFraude?driver=ODBC+Driver+17+for+SQL+Server'
```

Parámetros utilizados:

- RANDOM_STATE: Fijado en 42 para garantizar reproducibilidad.
- CHUNK_SIZE: Tamaño de los bloques al leer archivos grandes.
- DB_CONNECTION_STRING: Cadena de conexión a la base de datos SQL Server.

---

3. Conexión a la Base de Datos

```
engine = create_engine(DB_CONNECTION_STRING)
```

Se crea un motor de conexión a la base de datos usando sqlalchemy.

---

4. Limpieza de Tablas Existentes

```
def clean_database(engine):
    with engine.connect() as connection:
        tables_query = """SELECT TABLE_NAME FROM INFORMATION_SCHEMA.TABLES WHERE TABLE_TYPE = 'BASE TABLE'"""
        existing_tables = connection.execute(tables_query).fetchall()
        for table in existing_tables:
            connection.execute(f"DROP TABLE IF EXISTS {table[0]}")
    print("Base de datos limpiada.")
```

clean_database(engine)

Esta función elimina todas las tablas existentes en la base de datos para evitar conflictos con datos antiguos.

---

5. Carga de Datos de Forma Incremental

```
def load_large_datasets_incrementally(directory, files):
    combined_data = pd.DataFrame()
    for file in files:
        file_path = os.path.join(directory, file)
        chunk_iter = pd.read_csv(file_path, chunksize=CHUNK_SIZE, low_memory=False, dtype={
            'MONTO_ACREDITADO': 'float64',
            'MONTO_AUTORIZADO': 'float64',
            'GRUPO_TRANSFERENCIA_NOMBRE': 'string',
            'EJECUTORA_NOMBRE': 'string',
            'RUBRO_NOMBRE': 'string',
            'FUENTE_NOMBRE': 'string',
            'RECURSO_NOMBRE': 'string'
        })
        for chunk in chunk_iter:
            combined_data = pd.concat([combined_data, chunk], ignore_index=True)
    return combined_data
```

Se carga un conjunto de archivos de transferencias en bloques (chunks) para manejar grandes volúmenes de datos.

---

6. Generación de Datos Fraudulentos

```
def generate_fraudulent_data(data, n_specific=2500, n_random=2500):
    specific_fraud = data.sample(n=n_specific, random_state=RANDOM_STATE).copy()
    specific_fraud['MONTO_ACREDITADO'] *= 10

    random_fraud = data.sample(n=n_random, random_state=RANDOM_STATE).copy()
    random_fraud['MONTO_ACREDITADO'] = np.random.uniform(1e6, 5e6, n_random)

    return pd.concat([specific_fraud, random_fraud], ignore_index=True)
```

Se generan transacciones fraudulentas modificando el MONTO_ACREDITADO en dos formas:

1. Multiplicándolo por 10.
2. Asignando montos aleatorios elevados.

---

7. Preprocesamiento de Datos

```
X = data_sampled[['GRUPO_TRANSFERENCIA_NOMBRE', 'EJECUTORA_NOMBRE', 'RUBRO_NOMBRE', 'FUENTE_NOMBRE', 'RECURSO_NOMBRE']]
```

y = (data_sampled['MONTO_AUTORIZADO'] != data_sampled['MONTO_ACREDITADO']).astype(int)

Se define X (variables independientes) y y (variable dependiente que indica si una transacción es fraudulenta).

---

8. Configuración de Random Forest

rf_model = RandomForestClassifier(n_estimators=50, max_depth=10, random_state=RANDOM_STATE)

Parámetros de Random Forest

Parámetro	Valor	Descripción
n_estimators	50	Número de árboles en el bosque. Un valor moderado equilibra rendimiento y tiempo de entrenamiento.
max_depth	10	Profundidad máxima de los árboles. Un valor bajo evita el sobreajuste.
random_state	42	Asegura reproducibilidad en los resultados.

---

9. Entrenamiento y Evaluación del Modelo

```
rf_model.fit(X_train, y_train)
cv_scores = cross_val_score(rf_model, X_train, y_train, cv=5)
y_pred = rf_model.predict(X_test)
print(classification_report(y_test, y_pred))
```

Se entrena el modelo y se valida con cross_val_score. Se imprimen las métricas de clasificación.

---

10. Visualización de Resultados

Curva ROC

```
fpr, tpr, _ = roc_curve(y_test, rf_model.predict_proba(X_test)[:, 1])
roc_auc = auc(fpr, tpr)
plt.plot(fpr, tpr, label=f'Curva ROC (Área = {roc_auc:.2f})')
plt.legend()
plt.show()
```

Importancia de Variables

```
importances = rf_model.feature_importances_
plt.bar(range(X.shape[1]), importances, align='center')
plt.xticks(range(X.shape[1]), X.columns, rotation=90)
plt.show()
```

---

11. Almacenamiento de Resultados en la Base de Datos

fraudulent_transactions.to_sql('TransaccionesFraudulentas', con=engine, if_exists='replace', index=False)

Se guardan las transacciones detectadas como fraudulentas en la base de datos.

ANEXO 3

VALIDACION DE ESPECIALISTAS

Universidad Israel
ESPOG | Escuela de Posgrados

UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

INSTRUMENTO PARA VALIDACIÓN DE LA PROPUESTA

Estimado colega:

Se solicita su valiosa cooperación para evaluar la calidad del siguiente contenido digital **"Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos"**. Sus criterios son de suma importancia para la realización de este trabajo, por lo que se le pide que brinde su cooperación contestando las preguntas que se realizan a continuación.

Datos informativos

Validado por: Rolando Mauricio Ochoa Albuja
Título obtenido: Tecnólogo en Sistemas
C.I.:1712332301
E-mail: mauricio.ochoa@bancoprocredit.com.ec
Institución de Trabajo: Banco Procredit
Cargo: Coordinador de Base de Datos y Data Analítica
Años de experiencia en el área: 18

Universidad Israel

ESPOG | Escuela de Posgrados

Instructivo:

- Responda cada criterio con la máxima sinceridad del caso.
- Revisar, observar y analizar la propuesta de la plataforma virtual, blog o sitio web.
- Coloque una X en cada indicador, tomando en cuenta que Muy adecuado equivale a 5, Bastante Adecuado equivale a 4, Adecuado equivale a 3, Poco Adecuado equivale a 2 e Inadecuado equivale a 1.

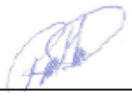
Tema: " "

Indicadores	Muy adecuado	Bastante Adecuado	Adecuado	Poco adecuado	Inadecuado
Pertinencia	X				
Aplicabilidad	X				
Factibilidad		X			
Novedad			X		
Fundamentación pedagógica		X			
Fundamentación tecnológica		X			
Indicaciones para su uso	X				
TOTAL	30				

Observaciones: Resultaría de gran utilidad el tener una solución como esta ya que brindaría una herramienta clave para el giro del negocio alertando a las áreas de primera línea sobre posibles transacciones fraudulentas y poniendo principal atención en ellas

Recomendaciones: Se recomienda realizar una propuesta para que la institución la analice y realizar dicha implementación de forma inmediata

Lugar, fecha de validación: Quito, 07 marzo 2025



Firma del especialista
MAURICIO OCHOA

Universidad Israel
ESPOG | Escuela de Posgrados

UNIVERSIDAD TECNOLÓGICA ISRAEL

ESCUELA DE POSGRADOS "ESPOG"

MAESTRÍA EN BIG DATA Y CIENCIA DE DATOS

INSTRUMENTO PARA VALIDACIÓN DE LA PROPUESTA

Estimado colega:

Se solicita su valiosa cooperación para evaluar la calidad del siguiente contenido digital **"Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos"**. Sus criterios son de suma importancia para la realización de este trabajo, por lo que se le pide que brinde su cooperación contestando las preguntas que se realizan a continuación.

Datos informativos

Validado por: Jorge Luis Enriquez Rivadeneira
Título obtenido: Magister en Inteligencia de Negocios y Analítica de Datos Masivos
C.I.: 1718049503
E-mail: Jorge.enriquez@bancoprocredit.com.ec
Institución de Trabajo: Banco Procredit
Cargo: Administrador de Data analítica
Años de experiencia en el área: 9

Universidad Israel

ESPOG | Escuela de Posgrados

Instructivo:

- Responda cada criterio con la máxima sinceridad del caso.
- Revisar, observar y analizar la propuesta de la plataforma virtual, blog o sitio web.
- Coloque una X en cada indicador, tomando en cuenta que Muy adecuado equivale a 5, Bastante Adecuado equivale a 4, Adecuado equivale a 3, Poco Adecuado equivale a 2 e Inadecuado equivale a 1.

Tema: "Detección de Fraude en Transacciones Financieras utilizando Modelos de Aprendizaje Automático con Datos Públicos".

Indicadores	Muy adecuado	Bastante Adecuado	Adecuado	Poco adecuado	Inadecuado
Pertinencia	X				
Aplicabilidad		X			
Factibilidad		X			
Novedad	X				
Fundamentación pedagógica		X			
Fundamentación tecnológica	X				
Indicaciones para su uso	X				
TOTAL	20	12			

Observaciones: El análisis que se pudiera conseguir con este proyecto empataría correctamente con el giro que la institución está presentando al tornarse en 100% digital, el análisis bridaría una herramienta que se tornaría crucial para el día a día en el trabajo del área, girando en torno a esta información y adaptar los subsistemas de TI a ella

Recomendaciones: La implementación se la debería realizar en la nube privada de la institución para de esta forma proteger de manera efectiva los datos que se pretenden analizar

Lugar, fecha de validación: Quito, 07 marzo 2025

Firma del especialista
JORGE ENRIQUEZ