

Informe Final:

“Detección de Corrupción, Fraude y Delitos Financieros mediante técnicas de Machine Learning”

Pasantía Part-Time

Sección 1

Sede Santiago

Doble Titulación:

Ing. Civil Informática

Ing. Civil Industrial

Autor:

Alberto Vergara Cerda

(albvergara@alumnos.uai.cl)

Fecha:

Lunes 27 de Noviembre 2023

Índice

Resumen ejecutivo	2
Executive Summary	3
Introducción	4
Contexto Histórico	4
Contexto Actual de la Empresa	5
Identificación de la Oportunidad y el Dolor Existente	6
Objetivos	7
Objetivos Específicos	7
Estado del Arte	8
Casos Reales	9
Entidades Reguladoras	12
Propuestas de Solución	14
Solución 1	14
Solución 2	15
Solución 3	15
Solución Escogida	16
Evaluación Económica	17
Esquema de Negocio	18
Análisis de Sensibilidad	19
Metodologías	21
Medidas de Desempeño	23
Desarrollo del proyecto	25
Matriz de Riesgos	26
Mitigación de Riesgos	28
Resultados	29
Discusión	33
Conclusión	35
Referencias y Anexos	36
Diagrama de Ishikawa	36
Carta Gantt	36
Algoritmos	38
Flujo Trabajo	40
Flujo Económico	41
Link Repositorio	41
Anexos	42
Bibliografía	47

Resumen ejecutivo

El proyecto realizado en Ceptinel durante el año 2023 se centró en el desarrollo de un modelo de machine learning para la detección de fraudes en transacciones. Esto fue necesario debido a la baja tasa de detección que tiene Ceptinel y a sus limitados mecanismos de detección.

En un mercado competitivo, con jugadores clave como IBM, SAS y FICO, y emergentes como Featurespace y Kount, Ceptinel necesita demostrar un progreso significativo. Se propusieron dos estrategias de precios: un modelo de suscripción basado en niveles y un esquema de pago por transacción, proyectando una rentabilidad anual de \$96.000, asegurando la viabilidad del proyecto incluso en escenarios menos optimistas.

La efectividad de Ceptinel en la publicidad en redes sociales y su participación en conferencias han sido cruciales para adquirir y retener clientes. El análisis de sensibilidad se centró en las ganancias, considerando dos escenarios de variabilidad en la adquisición de clientes. A pesar de los desafíos, el proyecto cumplió la mayoría de los objetivos específicos, excepto en la reducción de la tasa de falsos negativos. Sin embargo, se espera que esta mejore con el reentrenamiento del modelo.

Se identificaron riesgos clave como la calidad de los datos, interpretación de resultados, actualización y adaptación, sesgo y manejo de datos sensibles, con estrategias de mitigación establecidas para cada uno.

Los resultados mostraron una alta precisión en la identificación de transacciones no fraudulentas y se identificaron áreas de mejora en la detección de fraudes. La integración del modelo en Ceptinel permite predecir el nivel de riesgo de transacciones y elaborar reportes personalizados, mejorando significativamente la capacidad de la empresa para manejar transacciones sospechosas y posicionándose competitivamente en el mercado.

En conclusión, el proyecto en Ceptinel marcó un hito significativo en la detección de fraude, con un modelo de machine learning que mejoró la capacidad de la empresa para identificar y gestionar transacciones sospechosas. Este esfuerzo no solo representa un avance tecnológico sino también una valiosa experiencia de aprendizaje y crecimiento profesional para el pasante, contribuyendo significativamente al campo de la seguridad financiera y al cumplimiento normativo.

Executive Summary

The project undertaken at Ceptinel during the year 2023 focused on developing a machine learning model for fraud detection in transactions. This was necessitated by Ceptinel's low detection rate and limited detection mechanisms. In a competitive market, with key players like IBM, SAS, and FICO, and emerging ones such as Featurespace and Kount, Ceptinel needed to demonstrate significant progress. Two pricing strategies were proposed: a tier-based subscription model and a pay-per-transaction scheme, projecting an annual profitability of \$96,000, ensuring the project's viability even in less optimistic scenarios.

Ceptinel's effectiveness in social media advertising and participation in conferences have been crucial in acquiring and retaining customers. The sensitivity analysis focused on earnings, considering two scenarios of variability in customer acquisition. Despite challenges, the project met most specific objectives, except in reducing the false negative rate. However, this is expected to improve with the retraining of the model.

Key risks identified included data quality, result interpretation, updating and adaptation, bias, and handling of sensitive data, with mitigation strategies established for each.

The results showed high accuracy in identifying non-fraudulent transactions and areas for improvement in fraud detection were identified. The integration of the model into Ceptinel allows for predicting the risk level of transactions and generating customized reports, significantly enhancing the company's ability to handle suspicious transactions and positioning it competitively in the market.

In conclusion, the project at Ceptinel marked a significant milestone in fraud detection, with a machine learning model that improved the company's ability to identify and manage suspicious transactions. This effort represents not only a technological advancement but also a valuable learning experience and professional growth for the intern, contributing significantly to the field of financial security and regulatory compliance.

Introducción

En el actual panorama financiero, donde la integridad y transparencia son fundamentales para garantizar la estabilidad económica, las empresas se enfrentan constantemente al desafío de prevenir y detectar prácticas fraudulentas, corrupción y lavado de activos. Estas actividades ilícitas no solo amenazan la reputación y el buen funcionamiento de las organizaciones, sino que también socavan la confianza de los inversores y del público en general.

Este informe, pudo ser posible gracias a la experiencia obtenida durante un semestre, en el cual se desarrolló una pasantía en una empresa cuyo objetivo principal es identificar y combatir las malas prácticas financieras, mediante el monitoreo de transacciones. Esta compañía desempeña un papel crucial al salvaguardar la integridad del sistema financiero, identificando y previniendo el fraude, la corrupción y el lavado de activos. Dicha pasantía se desarrolló en el área TI de la empresa, específicamente en la célula de datos, la cual se encarga principalmente de recopilar, gestionar y analizar los datos de los clientes de Ceptinel. Además de garantizar la seguridad y confidencialidad de los datos transaccionales.

Contexto Histórico

A través de la historia, específicamente dentro del sector financiero, han ocurrido numerosos escándalos que sacudieron los cimientos de la economía global. Desde el colapso de importantes instituciones financieras hasta casos de corrupción masiva, es por esto que existe la necesidad de contar con mecanismos efectivos de monitoreo y prevención se ha vuelto imperativa.

En respuesta a estos desafíos, las autoridades regulatorias han intensificado sus esfuerzos para fortalecer las medidas de control y supervisión. Surgieron nuevas regulaciones y estándares internacionales que exigen a las empresas implementar sistemas sólidos de detección y prevención de prácticas financieras ilícitas.

Contexto Actual de la Empresa

La empresa de monitoreo de transacciones Ceptinel, fue fundada en 2019 con el objetivo de apoyar a diversas organizaciones a combatir casos de corrupción, fraude y lavado de activos. Hoy Ceptinel opera en Chile, Perú y Panamá, con la misión de abrir sus redes internacionales para captar clientes en el resto de Latinoamérica y de Estados Unidos, esta se encuentra en un momento crucial de su desarrollo. Enfrenta la tarea de adaptarse a un entorno financiero en constante evolución, donde las amenazas son cada vez más sofisticadas y difíciles de detectar. Los clientes de la empresa son mayormente entidades financieras como bancos, aseguradoras, cajas de compensación o administradores de fondos de pensiones. Dentro de la lista actual de clientes se encuentra el Banco BCI, Plan Vital, Mercantil, Canal Security, Prima y MF Tech.

La empresa, consciente de los retos actuales, ha invertido considerablemente en tecnología de última generación y en el desarrollo de modelos de análisis avanzados para identificar patrones de comportamiento sospechosos. Sin embargo, a pesar de estos esfuerzos, se ha identificado una brecha significativa en la tasa de detección del fraude, lo que sugiere que, aún con todos los avances tecnológicos implementados, se están escapando transacciones sospechosas del sistema.

Además de esta brecha, su crecimiento se ha visto ralentizado, ya que aún no están preparados para procesar la gran cantidad de datos que poseen los clientes más grandes del rubro, los servicios aún están montados en un servidor monolítico, lo que dificulta dimensionar sus métricas de rendimiento y consumo de recursos individuales de forma clara. Este problema es bastante crítico considerando la misión de Ceptinel y el compromiso con la industria.

Finalmente, en los últimos meses quedó en evidencia un feedback negativo por parte de algunos clientes antiguos de la empresa, lo que trajo como consecuencia la pérdida de un cliente importante, esto es algo que llama la atención y confirma que existen dudas respecto a la calidad del servicio y su eficacia por parte de los clientes.

Identificación de la Oportunidad y el Dolor Existente

Como oportunidad detectada, Ceptinel recibe grandes cantidades de datos, estos datos provienen de diversas fuentes y de distintos rubros. De este gran volumen de datos se puede extraer información valiosa para la mejora de los procesos de Ceptinel y asegurar la continuidad operacional mediante la optimización de los mecanismos de detección de fraudes.

Por otro lado, se identificó una brecha que consiste en la baja tasa de detección de transacciones fraudulentas actual, que no supera el 0,01%, esta baja tasa de detección del fraude representa una oportunidad clara de mejora para la empresa. Aumentar la efectividad de sus mecanismos de monitoreo y detección permitirá fortalecer su posición en el mercado y reforzar su papel como referente en la lucha contra las malas prácticas financieras, todo esto considerando que dentro de la industria, el estándar de las mejores empresas es de alrededor de un 0,1% a un 1% de las transacciones etiquetadas como sospechosas.

Al mismo tiempo, es importante reconocer que esta brecha pone en peligro la integridad de la empresa y su capacidad para salvaguardar los intereses de sus clientes. Cada transacción sospechosa que pasa desapercibida representa un riesgo potencial tanto para la organización como para el sistema financiero en general.

En este informe, se exploraron las posibles causas de esta baja tasa de detección del fraude y se proponen recomendaciones concretas para mejorar los mecanismos de monitoreo y detección de transacciones sospechosas. Con ello, se busca fortalecer la posición de la empresa como un actor clave en la prevención de prácticas financieras ilícitas y contribuir a la construcción de un entorno financiero más seguro y confiable.

Actualmente Ceptinel tiene una tasa de detección de aproximadamente 1 en 10.000 (0,01%) de las transacciones que procesa, con un 20% de Falsos Positivos y un 25% de Falsos Negativos. Quedando atrás en relación con algunos competidores directos e indirectos del rubro que manejan tasas de detección que superan el 0,1% y llegan hasta un 1% de las transacciones, es decir entre 10 y 100 transacciones detectadas cada 10.000 transacciones procesadas, todo esto sin superar un 10% de Falsos Positivos.

Objetivos

En base a este contexto se plantea un objetivo SMART para el proyecto el cual consiste en:

“Aumentar la tasa de detección de transacciones sospechosas del sistema de monitoreo de Ceptinel hasta un 1%, considerando un plazo de 8 meses”

Objetivos Específicos

Siguiendo este objetivo general se plantean 5 objetivos específicos.

- 1.** Consolidar procesos ETL y que estos demoren menos de 1 minuto por iteración.
- 2.** Disminuir la Tasa de Falsos Positivos y Falsos Negativos en un 10%.
- 3.** Aumentar la precisión del sistema de monitoreo hasta alcanzar un 90% de transacciones sospechosas identificadas correctamente.
- 4.** Incrementar el volumen de datos procesados en un 25% en términos de capacidad por unidad de tiempo.
- 5.** Aumentar el nivel de satisfacción del usuario en un 10%.

Estado del Arte

En los últimos años, el procesamiento de grandes volúmenes de datos se ha convertido en un componente clave para prevenir el fraude financiero. Las empresas están utilizando tecnologías avanzadas y técnicas analíticas sofisticadas para detectar patrones, anomalías y comportamientos sospechosos en grandes conjuntos de datos financieros. Algunos enfoques destacados en el estado del arte incluyen:

1- Aprendizaje automático y minería de datos: El uso de algoritmos de aprendizaje automático y técnicas de minería de datos permite identificar patrones y anomalías en los datos financieros. Estos enfoques utilizan modelos predictivos para detectar transacciones fraudulentas y realizar análisis de comportamiento para identificar actividades inusuales.

2- Análisis de red y detección de conexiones: Mediante el análisis de las conexiones entre diferentes entidades y transacciones financieras, se pueden identificar redes de fraude complejas. Los algoritmos de análisis de red ayudan a descubrir relaciones ocultas y patrones de comportamiento sospechoso.

3- Procesamiento en tiempo real: Para combatir el fraude financiero, es fundamental contar con sistemas de procesamiento de datos en tiempo real que puedan analizar y detectar actividades fraudulentas en tiempo casi real. Esto permite una respuesta inmediata y una mitigación temprana del riesgo.

4- Integración de fuentes de datos heterogéneas: El procesamiento de grandes volúmenes de datos requiere la integración de fuentes de datos heterogéneas, como transacciones financieras, registros de actividad, datos demográficos y más. La capacidad de unificar y analizar estos datos provenientes de diversas fuentes es fundamental para obtener una imagen completa y precisa del riesgo de fraude.

Casos Reales

JPMorgan Chase & Co.

JPMorgan Chase & Co., una de las instituciones financieras más grandes del mundo, ha utilizado el procesamiento de grandes volúmenes de datos para mejorar la detección y prevención del fraude financiero.

En 2013, JPMorgan Chase & Co. sufrió un importante ataque cibernético en el que se comprometieron datos de aproximadamente 76 millones de hogares y 7 millones de pequeñas empresas. Este incidente puso de manifiesto la necesidad de fortalecer las medidas de seguridad y mejorar el procesamiento de grandes volúmenes de datos para prevenir futuros ataques y actividades fraudulentas.

Tras el ataque, JPMorgan Chase & Co. realizó inversiones significativas en tecnologías de procesamiento de datos, como el uso de algoritmos avanzados de aprendizaje automático y análisis de datos a gran escala. Estas tecnologías permitieron a la empresa procesar grandes volúmenes de datos financieros en tiempo real y detectar patrones y comportamientos anómalos que indican posibles actividades fraudulentas.

JPMorgan Chase & Co. también mejoró sus capacidades de análisis de datos mediante la integración de múltiples fuentes de información, como transacciones financieras, datos de clientes y patrones de comportamiento. Esto permitió un enfoque más completo y preciso en la detección de fraude, al identificar anomalías en los datos y establecer correlaciones entre diferentes transacciones y cuentas.

Además, JPMorgan Chase & Co. implementó medidas de seguridad adicionales, como el fortalecimiento de la autenticación y la vigilancia continua de las actividades en sus sistemas. Estas medidas ayudaron a prevenir y mitigar futuros ataques y proteger la integridad de los datos financieros de la empresa y sus clientes.

El caso de JPMorgan Chase & Co. demuestra la importancia del procesamiento de grandes volúmenes de datos en la detección y prevención del fraude financiero. Mediante la aplicación de

tecnologías avanzadas y un enfoque integral en el análisis de datos, las instituciones financieras pueden fortalecer su capacidad para identificar y abordar eficazmente actividades fraudulentas, protegiendo así los intereses de sus clientes y su propia reputación.

Amazon

Amazon es conocida por su enfoque en el manejo de grandes volúmenes de datos para impulsar su modelo de negocio. Utiliza diversas tecnologías y prácticas para gestionar y analizar eficientemente grandes cantidades de datos. Algunos aspectos clave de su estado del arte en este campo incluyen:

1- Arquitectura de datos escalable: Amazon ha desarrollado su propia arquitectura de datos escalable y altamente disponible, basada en servicios en la nube como Amazon S3 (Simple Storage Service) y Amazon Redshift. Estos servicios permiten almacenar y procesar grandes volúmenes de datos de manera eficiente y confiable.

2- Sistemas de procesamiento distribuido: Amazon emplea sistemas de procesamiento distribuido, como Apache Hadoop y Apache Spark, para realizar análisis de datos a gran escala. Estas herramientas permiten el procesamiento paralelo de datos en clústeres de servidores, lo que acelera los tiempos de respuesta y permite el análisis de grandes conjuntos de datos.

3- Aprendizaje automático y análisis predictivo: Amazon utiliza técnicas de aprendizaje automático (machine learning) y análisis predictivo para extraer información valiosa de sus datos y mejorar la personalización de la experiencia del cliente. Al aplicar algoritmos avanzados, como la recomendación de productos basada en el historial de compras, Amazon puede ofrecer a sus clientes recomendaciones altamente relevantes.

Paypal

PayPal es una de las compañías más reconocidas y exitosas en el ámbito de los pagos en línea. Desde su fundación en 1998, ha revolucionado la forma en que las personas realizan transacciones comerciales en internet, brindando un servicio rápido, seguro y conveniente. Una de las áreas clave en las que PayPal ha invertido significativamente es la detección de fraudes, ya que es fundamental garantizar la seguridad de las transacciones y proteger a sus usuarios.

PayPal ha desarrollado sistemas avanzados de detección de fraudes que combinan tecnología de vanguardia y análisis de datos para identificar y prevenir actividades fraudulentas. Su enfoque se basa en el uso de algoritmos de machine learning, técnicas de inteligencia artificial y análisis en tiempo real para analizar patrones de transacciones y comportamiento del usuario.

Una de las fortalezas de PayPal en la detección de fraudes es su vasta cantidad de datos transaccionales y de comportamiento acumulados a lo largo de los años. Esto les permite construir modelos de aprendizaje automático altamente sofisticados que pueden identificar patrones y anomalías en las transacciones. Los algoritmos de machine learning utilizados por PayPal pueden reconocer señales sutiles de actividad fraudulenta y aprender de nuevas formas de fraude a medida que surgen.

Para mejorar aún más la precisión de sus sistemas de detección de fraudes, PayPal utiliza una combinación de técnicas de machine learning supervisado y no supervisado. El aprendizaje supervisado se basa en datos etiquetados previamente, lo que permite al algoritmo identificar patrones conocidos de transacciones fraudulentas. El aprendizaje no supervisado, por otro lado, se enfoca en descubrir patrones no etiquetados y anomalías que pueden indicar actividad sospechosa.

La detección de fraudes en PayPal no se limita solo a los patrones de transacciones, sino que también considera otros factores como la ubicación geográfica, el dispositivo utilizado para realizar la transacción y el historial de comportamiento del usuario. El análisis de estos datos adicionales ayuda a reducir aún más los falsos positivos y mejorar la precisión de la detección.

Además de los algoritmos de detección automatizados, PayPal cuenta con un equipo dedicado de expertos en seguridad y prevención de fraudes. Este equipo se encarga de investigar y analizar casos sospechosos, así como de adaptar y actualizar constantemente los modelos de detección para mantenerse al tanto de las últimas tendencias y técnicas utilizadas por los estafadores.

La prioridad de PayPal es brindar una experiencia segura y confiable a sus usuarios. Para lograr esto, han implementado medidas adicionales de seguridad, como la verificación en dos pasos y el

monitoreo en tiempo real de las transacciones. Estas medidas ayudan a proteger las cuentas de los usuarios contra accesos no autorizados y transacciones fraudulentas.

Es importante destacar que PayPal tiene una política de tolerancia cero hacia el fraude. En caso de detectar actividad sospechosa, toman medidas rápidas para proteger a sus usuarios y colaborar con las autoridades correspondientes en la investigación y el enjuiciamiento de los delincuentes.

En resumen, PayPal ha invertido significativamente en sistemas de detección de fraudes basados en tecnología de vanguardia y análisis automático que le permite seguir siendo una empresa líder en el rubro.

Entidades Reguladoras

Finalmente, es necesario entender el papel de las entidades reguladoras del ámbito financiero, como la UAF y la CMF.

Unidad de Análisis Financiero (UAF)

La UAF es una entidad autónoma del estado encargada de prevenir el lavado de activos y el financiamiento del terrorismo. Su función principal es recibir, analizar y remitir información a las autoridades judiciales pertinentes cuando detecta operaciones sospechosas que podrían estar relacionadas con estos delitos. Las instituciones financieras tienen la obligación de reportar transacciones inusuales a la UAF, proporcionando un mecanismo clave para detectar y prevenir actividades ilícitas en el sector financiero.



Comisión para el Mercado Financiero (CMF)

La CMF es una entidad pública autónoma que tiene como misión supervisar a los diferentes actores y entidades del mercado financiero en aras de asegurar el correcto funcionamiento, desarrollo y estabilidad del mercado financiero, resguardando los intereses del público. Esta supervisión implica garantizar que las entidades cumplen con la normativa vigente y actúan en beneficio de sus clientes y del sistema financiero en su conjunto. La CMF también tiene competencias en áreas relacionadas con la prevención del lavado de activos y financiamiento del terrorismo.

La interacción con entidades como la UAF y la CMF es crucial para cualquier empresa o sistema que gestione grandes volúmenes de transacciones financieras. Asegurarse de que las transacciones sean legítimas y que no estén vinculadas a actividades ilícitas no solo es esencial desde una perspectiva ética y legal, sino que también puede proteger a la empresa de posibles sanciones y daños a su reputación.



Propuestas de Solución

En relación a lo investigado se proponen 3 soluciones que resuelven el problema en su totalidad, cada una sigue una línea de las referencias del estado del arte. Enriquecimiento con datos de varias fuentes, Aprendizaje automático para la detección de patrones y Personalización según cliente y rubro.

Solución 1

“Aumentar la efectividad de los modelos de detección existentes mediante el enriquecimiento de los datos.”

Pros:

- Mayor precisión en la detección de fraudes al utilizar datos adicionales, como información contextual y características de comportamiento.
- Posibilidad de descubrir patrones y tendencias más sutiles que podrían indicar actividad fraudulenta.
- Mejora continua a medida que se incorporan nuevos datos y se actualizan los algoritmos.

Contras:

- Requiere un esfuerzo significativo para recopilar y procesar los datos adicionales necesarios.
- Puede haber desafíos en la calidad y disponibilidad de los datos, lo que puede afectar la efectividad de los modelos.
- Mayor complejidad en el mantenimiento y actualización de los modelos a medida que se agregan nuevos tipos de datos.

Solución 2

“Implementar un modelo de aprendizaje automático combinando modelos supervisados con no supervisados.”

Pros:

- Aprovechamiento de las fortalezas de ambos enfoques: los modelos supervisados utilizan datos etiquetados para identificar patrones conocidos, mientras que los no supervisados pueden descubrir anomalías y patrones desconocidos.
- Mayor capacidad para detectar nuevas formas de fraude y adaptarse a cambios en los patrones de actividad fraudulenta.
- Posibilidad de mejora continua a medida que se actualizan los modelos y se incorporan nuevas técnicas de aprendizaje automático.

Contras:

- Mayor complejidad en el desarrollo y entrenamiento de los modelos combinados.
- Posible aumento de los falsos positivos debido a la detección de anomalías legítimas como transacciones sospechosas.
- Requiere un análisis y validación rigurosos para garantizar que los modelos produzcan resultados precisos y confiables.

Solución 3

“Ofrecer la integración y personalización del sistema de monitoreo para cada cliente.”

Pros:

- Mayor satisfacción del cliente al adaptar el sistema a las necesidades y requisitos específicos de cada institución financiera.
- Posibilidad de ofrecer un servicio más completo y centrado en el cliente.
- Mayor efectividad en la detección de fraudes al tener en cuenta las reglas y políticas internas de cada cliente.

Contras:

- Requiere una personalización individualizada para cada cliente, lo que puede ser laborioso y requerir recursos adicionales.
- Necesidad de un fuerte enfoque en la comunicación y colaboración con los clientes para comprender sus necesidades y garantizar una implementación exitosa.
- Posible complejidad técnica al integrar y mantener la compatibilidad del sistema de monitoreo con los sistemas internos de las instituciones financieras.

Solución Escogida

Los criterios que se utilizan para determinar cuál es la solución óptima son Costo, Complejidad, Escalabilidad, Efectividad y Tiempo de Implementación cada uno comparte el mismo peso en la decisión final.

La metodología usada fue la siguiente: Para cada criterio se realiza un Top 3 donde el 1 es el mejor y el 3 el peor para cada una de las variables, al final se hace un promedio de sus calificaciones y se determina cual es la mejor solución en base al que está más cercano a 1.

	Costo	Complejidad	Escalabilidad	Efectividad	Tiempo de Implementación	Posiciones
Enriquecer Datos	2	1	2	2	2	1.8
Machine Learning	1	2	1	3	1	1.6
Personalizar Motor de Reglas	3	3	3	1	3	2.6

Fig 1: Tabla para ponderar posibles soluciones.

En este caso la solución escogida fue el Machine Learning como método para aumentar la tasa de detección de transacciones fraudulentas. Esta solución destaca principalmente en el costo, en su escalabilidad, y en el corto tiempo de implementación necesario para llevarla a cabo.

Evaluación Económica

El objetivo de esta evaluación es medir el impacto económico que puede tener el proyecto para la empresa, esto se hizo considerando un plazo de 2 años desde su paso a Producción. Para esto, se consideraron los costos relevantes en el desarrollo, despliegue y monitoreo del sistema de detección de fraude transaccional.

	2023	2024	2025
Costo Desarrollador	\$ 1.000	\$ 3.500	\$ -
Costo Infraestructura Servidor	\$ 200	\$ 1.200	\$ 1.200
Costo Almacenamiento	\$ 65	\$ 420	\$ 420
Consultoría Senior	\$ 50	\$ 400	\$ -
Costo Implementación	\$ 250	\$ 4.100	\$ 1.200
Flujo de Costos	\$ 1.565	\$ 9.620	\$ 2.820

VAN de COSTOS	\$-12.878
Tasa de descuento Anual	12%
Tasa de descuento Mensual	0,95%

Fig 2: Resultados Evaluación de Costos.

Dado que solo se consideran los costos relevantes, se calcula un VAN de costos con flujos mensuales, por lo tanto, para la evaluación del VAN de costos se usa una tasa efectiva mensual. Como referencia se toma una tasa de 12% anual, que es considerada en promedio para proyectos de riesgo medio o bajo. Para el cálculo de esta tasa, se considera que la empresa ya la calculó por medio de algún método como WACC.

Así mismo, dado que solo se han considerado costos relevantes, el cálculo de alguna TIR o tasa que mide la flexibilidad a los cambios de riesgo y que afecte a la tasa de descuento (Y que haga que el VAN se haga cero) no tiene sentido, ya que por ahora no hay cambios en los flujos positivos o negativos, lo mismo pasa para el payback. Es por eso que se requiere de la definición de un plan de negocios que convierta a este proyecto en un servicio o módulo adicional a lo ofrecido actualmente por Ceptinel y permita obtener rentabilidad a mediano y largo plazo.

Esquema de Negocio

En el mercado de detección de fraude mediante machine learning, hay varios jugadores clave que ofrecen soluciones variadas. Empresas como IBM, SAS y FICO son líderes reconocidos en este campo, ofreciendo soluciones integrales que combinan análisis avanzado, inteligencia artificial y experiencia en datos. Estas empresas han establecido una fuerte presencia en el mercado y gozan de la confianza de clientes en sectores como banca, comercio electrónico y telecomunicaciones.

Sin embargo, también existen empresas emergentes y de tecnología avanzada como Featurespace, Kount y Feedzai, que se están haciendo un nombre en el sector por su enfoque innovador y adaptabilidad a los requisitos específicos de los clientes. Estas empresas están ganando terreno gracias a su capacidad para integrar nuevas tecnologías y ofrecer soluciones personalizadas a diferentes tipos de fraude.

El estado actual del mercado muestra una tendencia creciente hacia soluciones más adaptativas y personalizadas, con una demanda cada vez mayor de herramientas capaces de aprender y evolucionar continuamente para enfrentar las sofisticadas tácticas de fraude. Es debido a esto que se hicieron 2 propuestas de estrategias de precios.

Propuesta 1: Modelo de Suscripción Basado en Niveles

Básico: \$500/mes - Incluye detección de fraude estándar con actualizaciones trimestrales.

Avanzado: \$700/mes - Incluye detección de fraude avanzada, actualizaciones mensuales y soporte técnico prioritario.

Premium: \$1000/mes - Incluye todas las características anteriores, además de análisis personalizados y acceso a nuevas características antes de su lanzamiento oficial.

Propuesta 2: Pago por Transacción

Pequeñas Empresas: \$0,01 por transacción - Ideal para empresas con bajo volumen de transacciones.

Medianas Empresas: \$0,008 por transacción - Para empresas con un volumen moderado de transacciones.

Grandes Empresas: Acuerdo personalizado - Para empresas con un alto volumen de transacciones, se ofrece un precio personalizado con descuentos basados en volumen.

Para calcular la rentabilidad esperada en un año se tomó la propuesta 1 bajo el supuesto de que Ceptinel tenga 12 clientes durante todo 2024. Resultando lo siguiente:

- Básico: 5 clientes x \$500/mes x 12 meses = \$30.000
- Avanzado: 5 clientes x \$700/mes x 12 meses = \$42.000
- Premium: 2 clientes x \$1000/mes x 12 meses = \$24.000

Total anual: \$96.000

Considerando esta rentabilidad anual bajo la propuesta 1 el proyecto pasaría a ser rentable antes de que termine el primer año desde su paso a producción, cabe destacar que cualquier aumento de costos debido a la adquisición de nuevos clientes será siempre superado por la rentabilidad que un nuevo suscriptor entrega a la empresa. Lo que invita a buscar la escalabilidad y mejora continua del modelo entrenado.

Es importante destacar que Ceptinel emplea eficazmente canales de publicidad en redes sociales y participa constantemente en conferencias. Estas actividades no solo ayudan a adquirir nuevos clientes, sino también a fidelizar a los existentes, creando una sólida base para el crecimiento sostenido del negocio.

Análisis de Sensibilidad

Considerando que los costos no van a variar demasiado, y que las variaciones siempre pueden ser amortizadas mediante la correcta administración de los recursos Cloud, se decidió que se analizaría la sensibilidad del proyecto en base a las Ganancias, debido a la incertidumbre que puede haber sobre el precio del mercado, la adopción del producto o la retención de los clientes.

Para este análisis se consideran 2 escenarios, el primero es un contexto en donde Ceptinel sólo logra vender este servicio a 6 clientes, el segundo escenario es aún más pesimista, se supone que Ceptinel no logra que sus clientes contraten el servicio durante el primer semestre pero durante el segundo semestre 4 clientes se suscriben, dos al plan básico, uno al plan avanzado, y el último se suscribe al servicio premium.

Escenario 1: Total anual: \$55.200

- Básico: 3 clientes x \$500/mes x 12 meses = \$18.000
- Avanzado: 3 clientes x \$700/mes x 12 meses = \$25.200
- Premium: 1 clientes x \$1000/mes x 12 meses = \$12.000

Escenario 2: Total anual: \$16.200

- Básico: 2 clientes x \$500/mes x 6 meses = \$6.000
- Avanzado: 1 clientes x \$700/mes x 6 meses = \$4.200
- Premium: 1 clientes x \$1000/mes x 6 meses = \$6.000

Finalmente, incluso en el peor escenario se cubren los costos totales de todo el proyecto solamente con el primer año desde su paso a producción. Esto significa que aún si las condiciones no son óptimas el proyecto pasa a ser rentable desde el segundo año en adelante.

	2023	2024	2025
Ingresos ML	\$ -	\$ 16.200	\$ 55.200
Costo Desarrollador	\$ -1.000	\$ -3.500	\$ -
Costo Infraestructura Servidor	\$ -200	\$ -1.200	\$ -1.200
Costo Almacenamiento	\$ -65	\$ -420	\$ -420
Consultoría Senior	\$ -50	\$ -400	\$ -
Costo Implementación	\$ -250	\$ -4.100	\$ -1.200
Flujo de Caja	\$ -1.565	\$ 6.580	\$ 52.380

VAN	\$41.131,26
-----	-------------

Fig 3: Resultados Finales.

Metodologías

Cada objetivo tiene una metodología de validación propia.

1. **Validación por pares:** Este primer objetivo es la base para seguir avanzando con el proyecto, es por eso que se debe pasar por una validación de la célula de datos de Ceptinel, generando un estándar en la ingesta de datos.
2. **Análisis FODA y Six-Sigma:** El análisis FODA permite extraer información valiosa sobre las oportunidades y amenazas que significa una tasa alta de FP o FN, por otra parte la cobertura estadística que otorga el trabajar con un estándar de calidad como Six-Sigma abre paso a un proceso sano y de gran calidad.
3. **Costo-Beneficio y Validación por pares:** Se debe hacer un análisis profundo para determinar hasta qué punto es conveniente aumentar la precisión del sistema de monitoreo sin generar más costos que beneficios para la empresa. Esta cifra o punto crítico debe ser validada por la célula de datos y la célula de desarrollo.
4. **SEMMA:** Este proceso ampliamente utilizado en la minería de datos atraviesa todas las etapas que se necesitan para que el proceso de ingesta de datos sea eficiente y a prueba de fallas, etapas como el muestreo, exploración, modificación, modelado y evaluación son esenciales en esta fase.
5. **Encuestas de Satisfacción:** El feedback que entregan los clientes debe ser tomado en cuenta siempre en este tipo de empresas, no sólo entender qué tan satisfecho está el cliente, sino también conocer métodos para aumentar ese nivel de satisfacción, las encuestas son el mejor método para esta misión.

A continuación se muestra un diagrama simplificado de los procesos de Ceptinel, lo encerrado en el recuadro negro representa las tareas de la célula de datos, es en esta secuencia en la cual se trabajará directamente en este proyecto.

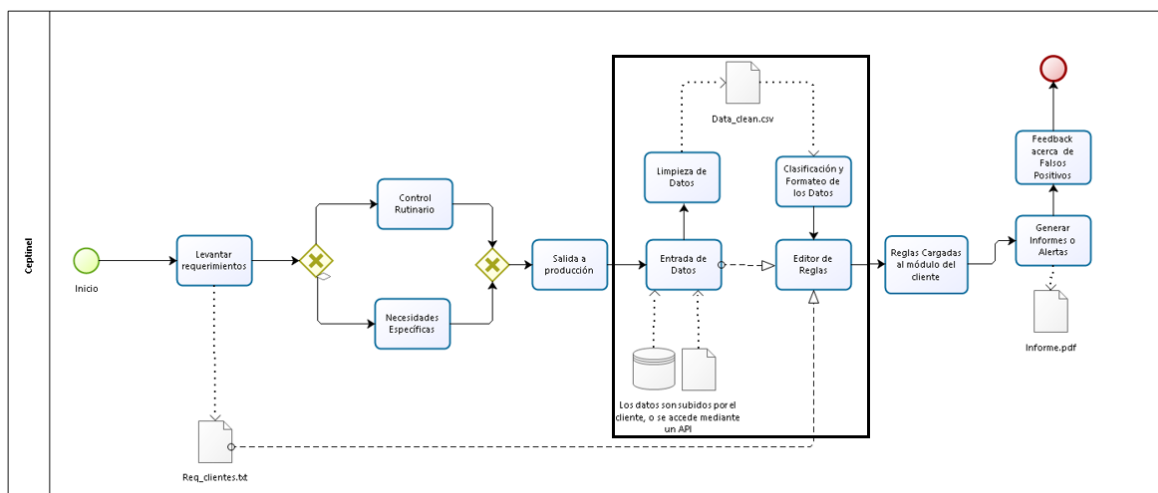


Fig 4: Diagrama de Procesos Ceptinel.

Como metodología de trabajo interna de la empresa, se desarrolló el proyecto mediante trabajo ágil de Scrum, usando herramientas de seguimiento y despliegue continuo como Jira. Este método consiste en reuniones diarias y semanales de avance, denominadas Daily o Weekly respectivamente. El objetivo de estas reuniones era definir las tareas a realizar durante el día, declarar avance de días anteriores y también tener espacio para levantar dudas, consultas o problemáticas que surjan durante el desarrollo. Para el proyecto se utilizaron los 6 estados de tareas que ofrece Jira.

- Blocked: Tareas Bloqueadas.
- To Do: Tareas por hacer o pendientes.
- In Progress: Tareas en progreso, que toman más de un día.
- Daily: Tareas que se deben completar durante el día.
- In Review: Tareas finalizadas que por algún motivo requieren de una validación por parte del supervisor del proyecto u otra persona relevante para el mismo.
- Done: Tareas finalizadas que ya pasaron la revisión y están listas.

Medidas de Desempeño

Las medidas de desempeño están alineadas con los objetivos específicos.

1. **Tiempo de ejecución del proceso ETL:** El proceso consiste en 3 etapas: la extracción, la transformación y la carga de los datos. Todas las etapas fueron medidas en minutos.

$$TT_{ETL} = T_E + T_T + T_L$$

2. **Tasa de falsos positivos:** Medida en porcentaje, representa la proporción de transacciones clasificadas incorrectamente como fraudulentas o lavado de activos por el modelo de datos.

$$T_{FP} = \frac{\text{Número de Falsos Positivos}}{\text{Falsos Positivos} + \text{Verdaderos Negativos}}$$

3. **Tasa de falsos negativos:** Medida en porcentaje, representa la proporción de transacciones fraudulentas o lavado de activos que no fueron detectadas por el modelo de datos.

$$T_{FN} = \frac{\text{Número de Falsos Negativos}}{\text{Verdaderos Positivos} + \text{Falsos Negativos}}$$

4. **Tasa de transacciones fraudulentas detectadas correctamente:** Medido en porcentaje. (%)

$$T_{Precisión} = \frac{\text{Número de Transacciones identificadas correctamente}}{\text{Total Transacciones identificadas}} \times 100$$

5. **Volumen de datos procesados:** Medido en GB o TB, representa la cantidad de datos que el modelo de datos es capaz de procesar en un determinado período de tiempo. También se puede medir la cantidad de registros procesados en una unidad de tiempo.

$$V_{\frac{\text{Procesados}}{\text{Hora}}} = \frac{V_{\text{Total}} - V_{\text{Por Procesar}}}{\text{Hora}}$$

6. **Nivel de satisfacción del usuario:** Medido en escala de 1 a 7.

$$N_{\text{Satisfacción}} = \frac{\sum \text{Resultados Encuesta}}{\text{Número de Encuestados}}$$

Desarrollo del proyecto

El plan de trabajo considerado para implementar esta solución se basó en la metodología CRISP-DM, la cual es la más usada actualmente en los proyectos de Machine Learning debido a su gran versatilidad. El plan se presenta a continuación:

Fase 1: Comprensión del negocio y definición del objetivo.

- Reunión con los stakeholders
- Definición objetivo SMART

Fase 2: Comprensión de los datos.

- Recopilación y análisis de datos
- Evaluación de los datos

Fase 3: Preparación de los datos.

- Limpieza y transformación
- Selección de características
- Consolidación del proceso ETL

Fase 4: Modelado.

- Selección de algoritmos
- Entrenamiento modelo no supervisado
- Entrenamiento modelo supervisado

Fase 5: Evaluación.

- Evaluación de los modelos seleccionados
- Ajuste y optimización del modelo

Fase 6: Implementación.

- Integración del modelo en el sistema de monitoreo
- Pruebas y validación

Fase 7: Monitoreo y mantenimiento.

- Monitoreo continuo del rendimiento
- Actualización y mejora

Matriz de Riesgos

Se identificaron 5 riesgos durante la etapa de implementación, estos fueron ubicados en la matriz de riesgo según su severidad y probabilidad de ocurrencia. Para luego evaluar y definir sus planes de mitigación respectivos.

- A. Baja calidad de los datos:** El éxito de un modelo de aprendizaje automático depende en gran medida de la calidad de los datos utilizados para entrenar y validar el modelo. Si los datos de entrada contienen errores, sesgos o información incompleta, esto puede afectar la precisión y confiabilidad de los resultados. Es crucial garantizar la calidad y la integridad de los datos utilizados en el proceso de implementación.
- B. Interpretación de resultados:** Los modelos de aprendizaje automático, especialmente los que involucran técnicas no supervisadas, pueden generar resultados difíciles de interpretar. La complejidad del modelo puede dificultar la comprensión de cómo se toman las decisiones y qué características o variables son las más relevantes para la detección del fraude. Es esencial contar con mecanismos de validación y explicación de los resultados para garantizar su interpretación correcta.
- C. Actualización y adaptación:** El entorno financiero está en constante evolución, y los métodos utilizados para cometer fraude también evolucionan rápidamente. Por lo tanto, es necesario mantener los modelos actualizados y adaptados a las nuevas tendencias y patrones de fraude. Esto implica la necesidad de un proceso de retroalimentación constante y una capacidad de actualización y ajuste continuo del modelo implementado.

- D. **Sesgo:** Los modelos de aprendizaje automático pueden verse afectados por sesgos inherentes a los datos utilizados para entrenarlos. Si los datos históricos contienen sesgos o discriminación, el modelo puede replicar y amplificar esos sesgos, lo que puede resultar en decisiones discriminatorias o injustas. Es fundamental evaluar y mitigar el sesgo en el modelo, asegurando la equidad y la imparcialidad en la detección del fraude.
- E. **Manejo de datos sensibles:** Al trabajar con empresas del rubro financiero se da el caso de que no quieren que sus datos se ocupen para otra cosa que no sea el monitoreo, por lo que no todos los clientes están dispuestos a que sus datos se utilicen para entrenar un modelo de aprendizaje automático, aun sabiendo que esto puede representar una ventaja o un mejor servicio a futuro.

		Severidad				
Probabilidad		1	2	3	4	5
1					A	
2				E		
3		D	B			
4						
5		C				

Fig 5: Matriz de Riesgos.

Mitigación de Riesgos

Podemos ver en la matriz los 5 riesgos identificados durante esta etapa de implementación:

- A. **Calidad de los datos:** Se consolida el proceso ETL en el cual se define los pasos para la limpieza y preprocesamiento de los datos, se utilizan técnicas de muestreo estratificado, combinando sobremuestreo de la clase minoritaria y submuestreo de la clase mayoritaria para lidiar con el desbalance de las clases Fraude y no Fraude.
- B. **Interpretación Resultados:** Aplicar técnicas de visualización de los datos para tener gráficos más claros y entendibles, una vez se recibe una salida del modelo esta es discutida por la célula de datos, para no caer en errores de interpretación. Para esta tarea se emplearon gráficos de barras, histogramas, mapas de calor y diagramas de dispersión, además se emplearon boxplot para detectar outlier fácilmente.
- C. **Actualización y Adaptación:** Fijar a un grupo de 2 o más personas que se encarguen de mantener y adaptar el modelo según sea necesario, fijar plazos para reentrenamiento, establecer un protocolo para evaluar en tiempo real el desempeño del modelo. Esto corresponde a la última etapa de la planificación Monitoreo y Mantenimiento.
- D. **Sesgo:** Se realiza un análisis exploratorio profundo de los datos para eliminar cualquier sesgo previamente existente, además se utilizan técnicas de submuestreo y sobremuestreo para balancear las clases. Además se seleccionaron algoritmos que son conocidos dentro del rubro por ser insensibles al sesgo, junto con que son más transparentes e interpretables. Por otra parte, se revisan las fuentes de información de forma crítica para entender sus limitaciones, posibles sesgos y la representatividad de las muestras. Finalmente se realizan pruebas rigurosas de validación cruzada K-Fold, para asegurar que el modelo generaliza bien y no perpetúa sesgos.
- E. **Manejo de Datos Sensibles:** Se obtuvo el consentimiento explícito de los clientes cuyos datos fueron usados para el entrenamiento del modelo, explicando al cliente de qué forma se van a utilizar los datos y con qué medidas de seguridad se van a trabajar las variables o columnas que contengan información sensible. Se opta por enmascarar datos personales o derechamente no trabajar con ellos en el entrenamiento.

Resultados

En términos de resultados, se tienen aquellos que salen directamente del modelo entrenado y que se representan mediante herramientas como la matriz de confusión y el área bajo la curva de Precisión-Recall



Fig 6: Matriz de Confusión RF.

Esta matriz se divide en cuatro partes:

- **Verdaderos Positivos (VP):** Se obtuvieron 740 casos que el modelo ha predicho correctamente como fraude.
- **Verdaderos Negativos (VN):** Acá se tienen 449,212 casos que el modelo ha identificado correctamente como no fraude.
- **Falsos Positivos (FP):** Hay 70 casos que el modelo incorrectamente marcó como fraude (estos son en realidad no fraude).
- **Falsos Negativos (FN):** En esta parte, 402 casos de fraude no fueron detectados por el modelo.

En términos simples, se busca que los números de Verdaderos Positivos y Verdaderos Negativos sean lo más altos posibles, mientras que los Falsos Positivos y Falsos Negativos sean lo más bajos posibles. En este caso, el modelo es muy bueno para identificar no fraudes, pero necesita mejorar en la detección de fraudes, ya que hay una cantidad considerable de fraudes que no detectó

(Falsos Negativos). Esto se debe mejorar mediante un reentrenamiento con nuevos datos etiquetados.

Por otro lado, se tiene la curva de Precisión-Recall, que es un gráfico que muestra la relación entre dos métricas importantes:

- **Precisión:** De todas las transacciones que el modelo predijo como fraude. ¿Cuántas fueron realmente fraude?
- **Recall (Sensibilidad):** De todas las transacciones que son fraude. ¿Cuántas pudo encontrar el modelo?

La línea azul muestra cómo cambia la precisión del modelo a medida que intenta mejorar su recall. El área bajo la curva (AUC) nos da una idea general de la calidad del modelo; cuanto más cercano es el valor a 1, mejor es el modelo. Un AUC de 0.83 es considerado bueno, indicando que el modelo tiene una buena capacidad para distinguir entre transacciones fraudulentas y no fraudulentas.

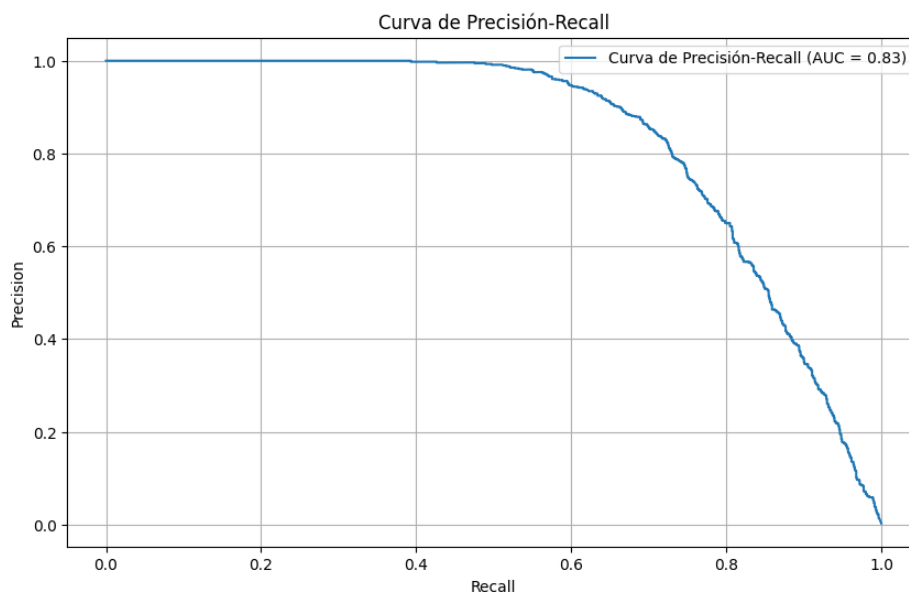
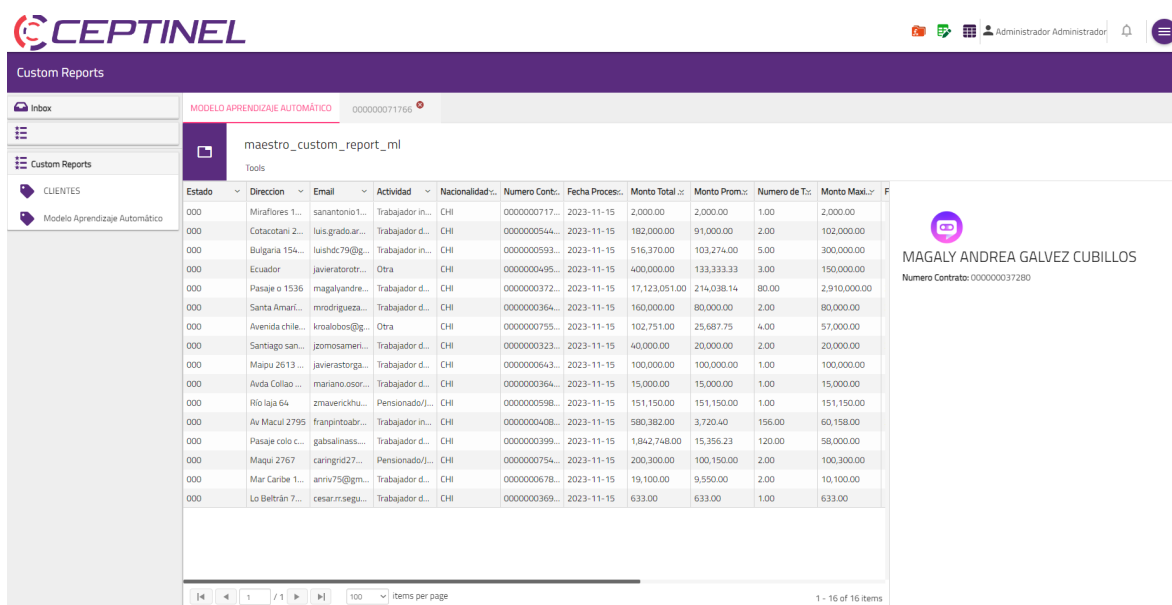


Fig 7: Curva de Precisión-Recall.

En resumen, mientras que la matriz de confusión otorga números exactos de las predicciones del modelo, la curva de precisión-recall entrega una vista más detallada de su rendimiento a lo largo de diferentes umbrales, lo que es especialmente útil en situaciones donde el equilibrio entre precisión y recall es crítico, como en la detección de fraude.

Una vez se entrenó el modelo este fue empaquetado para su posterior consumo en la plataforma Ceptinel, este modelo se implementó de manera que fuera capaz de predecir el nivel de riesgo de transacciones y elaborar reportes personalizados por cliente indicando su factor de riesgo. La vista también incluye su historial transaccional y da la posibilidad de revisar los resultados del modelo a nivel individual, ya sea por cliente o por transacción.



The screenshot shows the Ceptinel Custom Reports interface. The top navigation bar includes the Ceptinel logo, user information (Administrador), and a search icon. The main content area is titled 'Custom Reports' and displays a table of transaction data for a specific client. The table has columns for Estado, Dirección, Email, Actividad, Nacionalidad, Numero Contr., Fecha Proces., Monto Total, Monto Prom., Numero de T., and Monto Maxi. The data is filtered by 'MODELO APRENDIZAJE AUTOMÁTICO' and 'maestro_custom_report_ml'. The table shows 16 items, with the first 15 visible. The data includes transaction details for various clients, such as 'Miraflores 1...', 'Cotacani 2...', 'Bulgaria 154...', 'Ecuador', 'Pasaje o 1536', 'Santa Amar...', 'Avenida Chile...', 'Santiago san...', 'Maipú 2613...', 'Aída Collao...', 'Río Iaja 64', 'Av Macul 2795', 'Pasaje colo c...', 'Maqui 2767', 'Mar Caribe 1...', and 'Lo Beltrán 7...'. The table also includes a summary row for 'MAGALY ANDREA GALVEZ CUBILLOS' with a 'Numero Contrato: 000000037280'.

Estado	Dirección	Email	Actividad	Nacionalidad	Numero Contr.	Fecha Proces.	Monto Total	Monto Prom.	Numero de T.	Monto Maxi
000	Miraflores 1...	sanantonio1...	Trabajador in...	CHI	0000000717...	2023-11-15	2,000.00	2,000.00	1.00	2,000.00
000	Cotacani 2...	luisgrado.ar...	Trabajador d...	CHI	0000000544...	2023-11-15	182,000.00	91,000.00	2.00	102,000.00
000	Bulgaria 154...	luisd79@...	Trabajador in...	CHI	0000000593...	2023-11-15	516,370.00	103,274.00	5.00	300,000.00
000	Ecuador	javieratorotr...	Otra	CHI	0000000495...	2023-11-15	400,000.00	133,333.33	3.00	150,000.00
000	Pasaje o 1536	magalyandre...	Trabajador d...	CHI	0000000372...	2023-11-15	17,123,051.00	214,038.14	80.00	2,910,000.00
000	Santa Amar...	mrodriguez...	Trabajador d...	CHI	0000000364...	2023-11-15	160,000.00	80,000.00	2.00	80,000.00
000	Avenida Chile...	kroalobos@...	Otra	CHI	0000000755...	2023-11-15	102,751.00	25,687.75	4.00	57,000.00
000	Santiago san...	jzomosen...	Trabajador d...	CHI	0000000323...	2023-11-15	40,000.00	20,000.00	2.00	20,000.00
000	Maipú 2613 ...	javierastorga...	Trabajador d...	CHI	0000000643...	2023-11-15	100,000.00	100,000.00	1.00	100,000.00
000	Aída Collao ...	mariano.osor...	Trabajador d...	CHI	0000000364...	2023-11-15	15,000.00	15,000.00	1.00	15,000.00
000	Río Iaja 64	zmaverickhu...	Pensionado/...	CHI	0000000598...	2023-11-15	151,150.00	151,150.00	1.00	151,150.00
000	Av Macul 2795	francintoabr...	Trabajador in...	CHI	0000000408...	2023-11-15	580,382.00	3,720.40	156.00	60,158.00
000	Pasaje colo c...	gabrielinas...	Trabajador d...	CHI	0000000399...	2023-11-15	1,842,748.00	15,356.23	120.00	58,000.00
000	Maqui 2767	caringsid27...	Pensionado/...	CHI	0000000754...	2023-11-15	200,300.00	100,150.00	2.00	100,300.00
000	Mar Caribe 1...	amiv75@gm...	Trabajador d...	CHI	0000000678...	2023-11-15	19,100.00	9,550.00	2.00	10,100.00
000	Lo Beltrán 7...	cesar.rcsaga...	Trabajador d...	CHI	0000000369...	2023-11-15	633.00	633.00	1.00	633.00

Fig 8: Mockup integración a Ceptinel.

La figura 9 muestra la salida final del modelo y su implementación en Ceptinel, mostrando como ejemplo la vista de un cliente, en donde se pueden ver claramente sus datos personales, datos transaccionales, nivel de riesgo. Además como agregado se enseñan los factores de importancia que el modelo asigna a las columnas incluidas en el dataset.

También se implementó dentro de la salida final del modelo la posibilidad de validar los resultados por parte del usuario final, esta colaboración y comunicación fluida con el usuario es esencial para las futuras mejoras y adaptación de las herramientas de ML que ofrece Ceptinel.

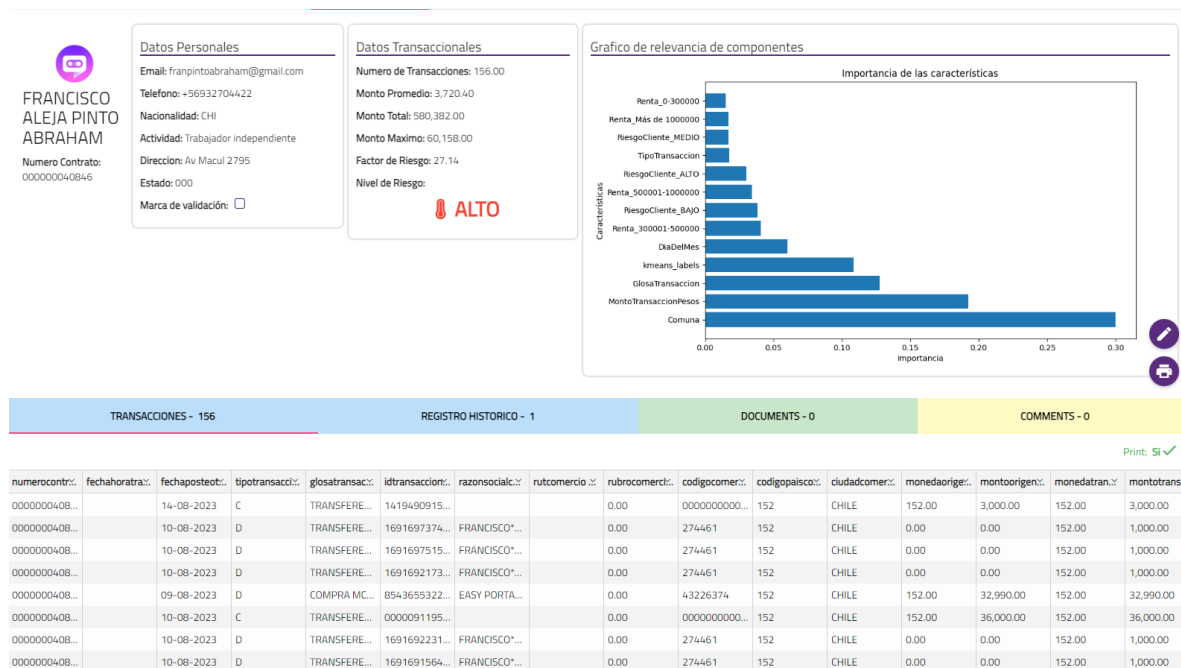


Fig 9: Mockup integración a Ceptinel.

Como resultado del proyecto se analizó si se cumplieron los objetivos tanto específicos como el objetivo general, esto se realizó considerando las métricas o medidas de desempeño definidas previamente en el informe.

Métrica	Valor Inicio	Valor Final	Estado
Duración proceso ETL	3[min]	1[min]	
Tasa de Falsos Positivos	20%	0,02%	
Tasa de Falsos Negativos	25%	35%	
Precisión	?	99,9%	
Volumen de Datos Procesados	1[Millón/H]	1,5[Millón/H]	
Nivel de Satisfacción Clientes	5	6	
Tasa de Detección	0,01%	0,2%	

Discusión

En general los resultados cumplieron las expectativas, se lograron los 5 objetivos específicos planteados al inicio, exceptuando disminuir la tasa de FN, esto habla del gran desempeño del modelo final. Existen razones suficientes para afirmar que esta tasa va a mejorar cuando se realice el primer reentrenamiento del modelo con datos productivos. Esta primera instancia de reentrenamiento está prevista para el primer trimestre de 2024.

Si bien el valor de precisión del modelo es muy alto no se debe confiar a ciegas, primero porque en Ceptinel anterior a la implementación del proyecto no existía una manera de obtener esta métrica de forma confiable y precisa. Además, debido a la naturaleza desbalanceada de los datos, esta métrica pasa a ser un poco irrelevante y pierde valor, de forma contraria a las métricas de FP y FN que aportan gran valor a la toma de decisiones en cuanto a métodos de entrenamiento y adaptación.

Por otra parte, el objetivo general no fue alcanzado, es vital reconocer que el objetivo propuesto al inicio de este proyecto resultaba muy ambicioso sin reflejar realmente cual era la realidad a nivel nacional, donde según estudios la tasa de fraude es menor que en Europa y Norteamérica, a pesar de esto se logró aumentar hasta 20 veces la capacidad de detección de transacciones sospechosas. Alcanzando valores que acercan a Ceptinel a sus competidores internacionales, y lo ubican dentro de los grandes representantes del cumplimiento normativo dentro de Latinoamérica.

Los problemas que ocurrieron durante el desarrollo no fueron relevantes, por lo que no afectaron los resultados, pero es importante ahondar en más detalle acerca de cómo la cultura percibe los términos de aprendizaje automático o machine learning, muchas veces el cliente exhibe un nivel de conocimiento limitado en la materia en cuestión y manifiesta expectativas que no se ajustan de manera realista a las circunstancias o condiciones asociadas, las cuales dificultan que estos entiendan y acepten lo que realmente Ceptinel les está ofreciendo.

Para finalizar, como componente ético del proyecto, es importante destacar que todo el desarrollo y el trabajo realizado durante la pasantía tuvo como finalidad aportar y mantener un legado de seguridad financiera siguiendo las buenas prácticas, junto con los consensos internacionales que llevan a nuestra sociedad a convivir en armonía evitando conductas de fraude y corrupción que tanto han afectado a la sociedad moderna.

Conclusión

Aunque el objetivo general propuesto al inicio del proyecto no fue alcanzado, debido a que existió una gran ambición. La implementación de esta solución tecnológica avanzada mejoró notablemente la capacidad de la empresa para identificar y gestionar transacciones sospechosas, aumentando hasta 20 veces su capacidad de detección, posicionando a Ceptinel como un actor competitivo en el mercado.

Durante el desarrollo del proyecto, se enfrentaron diversos desafíos, desde la conceptualización inicial hasta la implementación práctica del modelo. Estas dificultades brindaron oportunidades valiosas para aprender sobre la importancia de ser adaptable e innovador en un entorno tecnológico en constante cambio. La comprensión profunda de las necesidades del cliente y del mercado se revelaron como un factor crucial para guiar la estrategia de desarrollo y comercialización del proyecto.

Asimismo, este proyecto representó una excelente oportunidad para el crecimiento profesional y personal del pasante. Se adquirieron habilidades avanzadas en el ámbito de machine learning y análisis de datos, así como una mayor comprensión de las estrategias de negocio y marketing aplicadas a soluciones tecnológicas. Trabajar en un entorno desafiante, aplicando metodologías ágiles de desarrollo fortaleció la capacidad de adaptación y fomentó la búsqueda de soluciones creativas ante problemas complejos.

Finalmente, se pudo concluir que el sector financiero desempeña un papel crucial en el desarrollo de nuestras vidas, por lo que mantener este sector libre de malas prácticas que perjudican a todos es vital.

Referencias y Anexos

Diagrama de Ishikawa

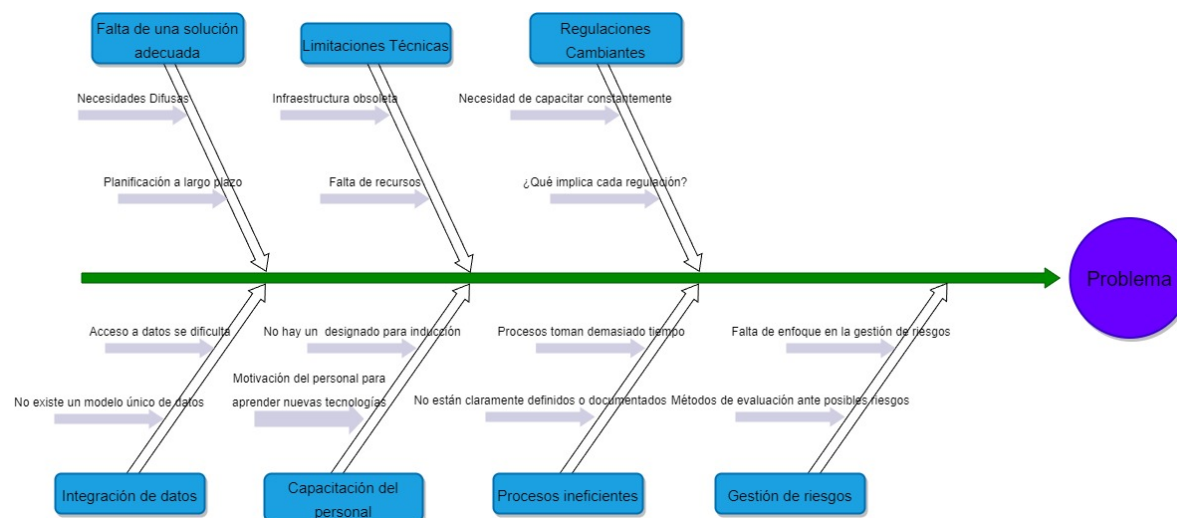


Fig 10: Diagrama Causas Posibles.

Carta Gantt

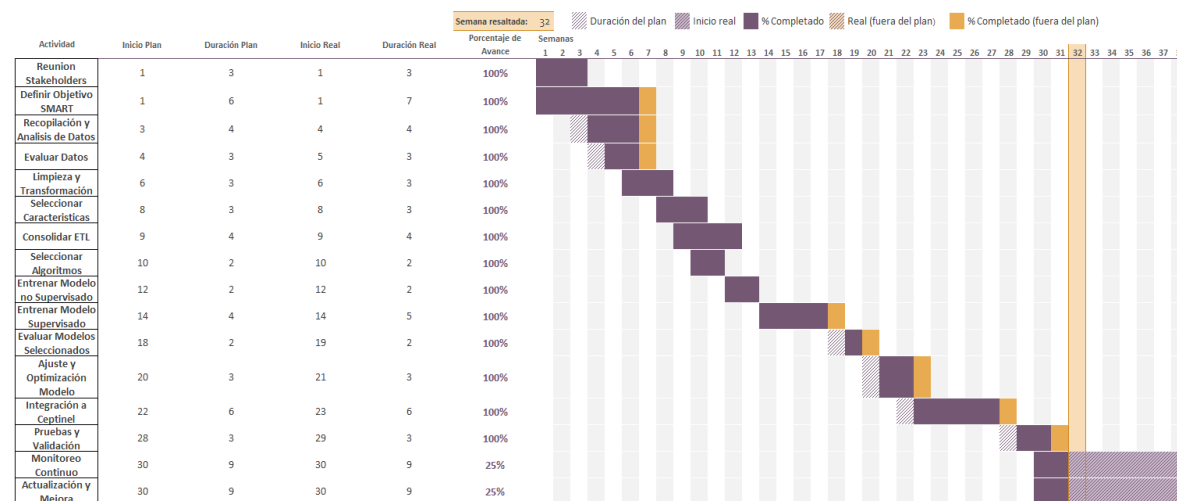
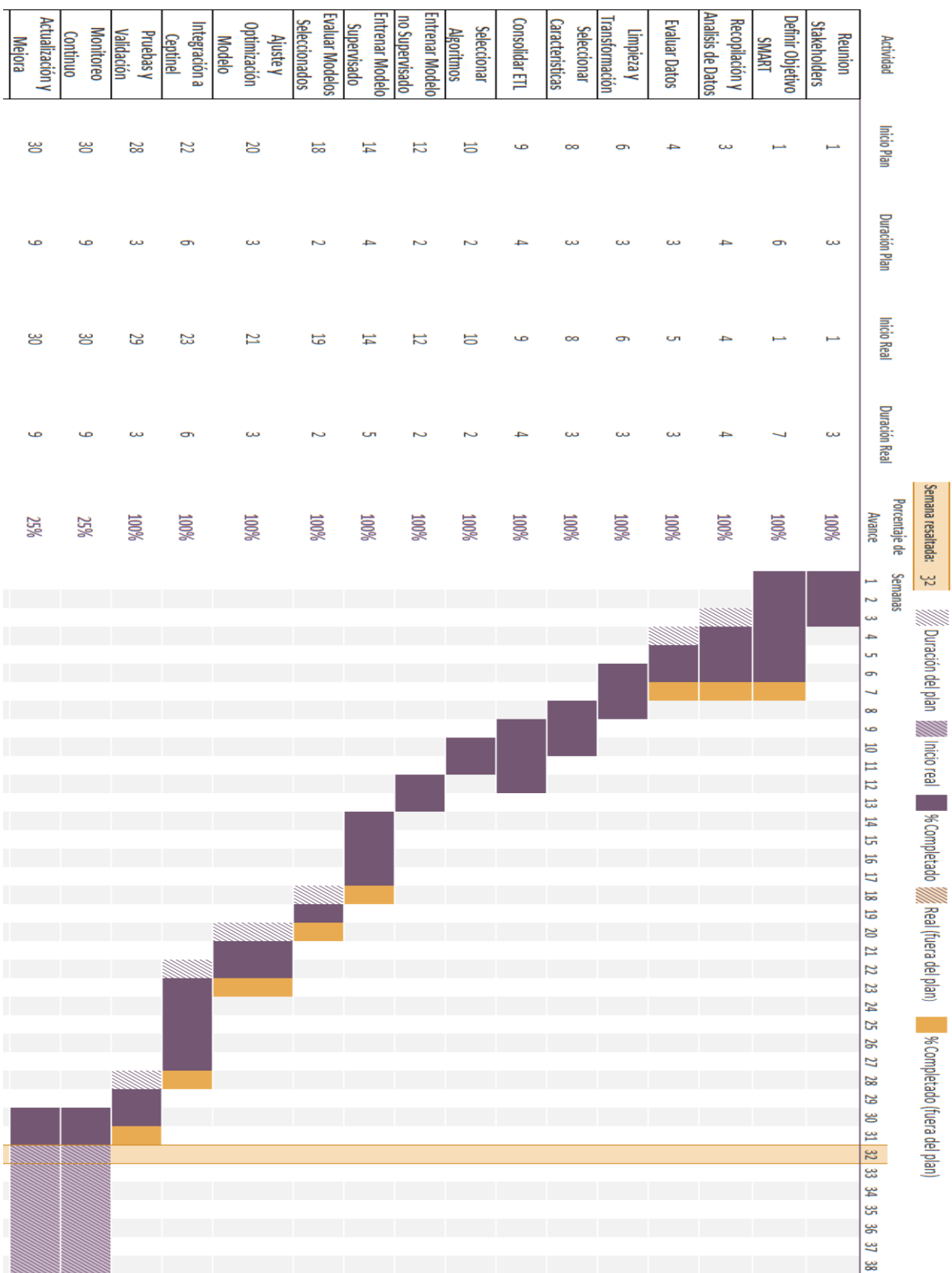


Fig 11: Planificación.



Algoritmos

K-means (No Supervisado)

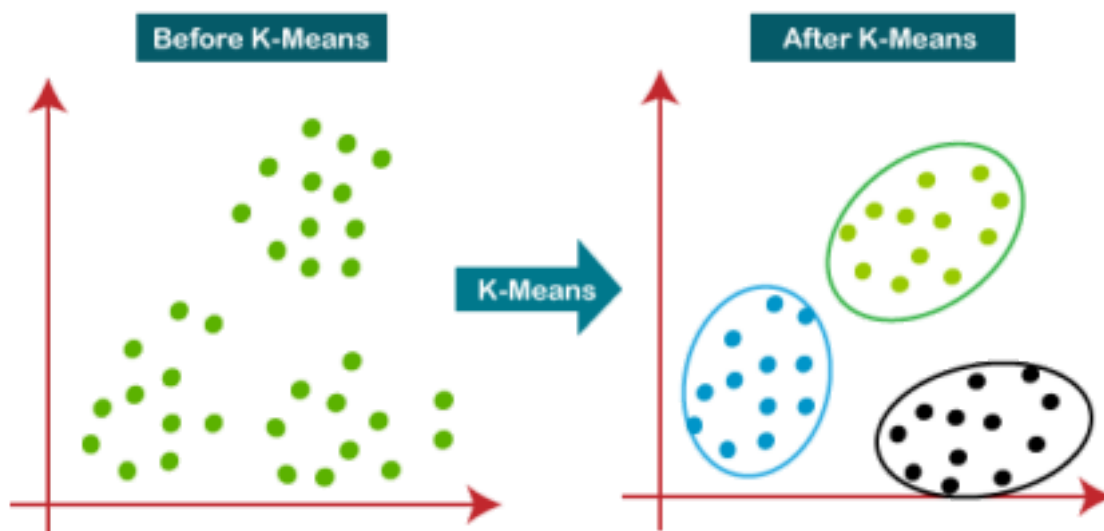


Fig 12: Ejemplo de uso K-means.

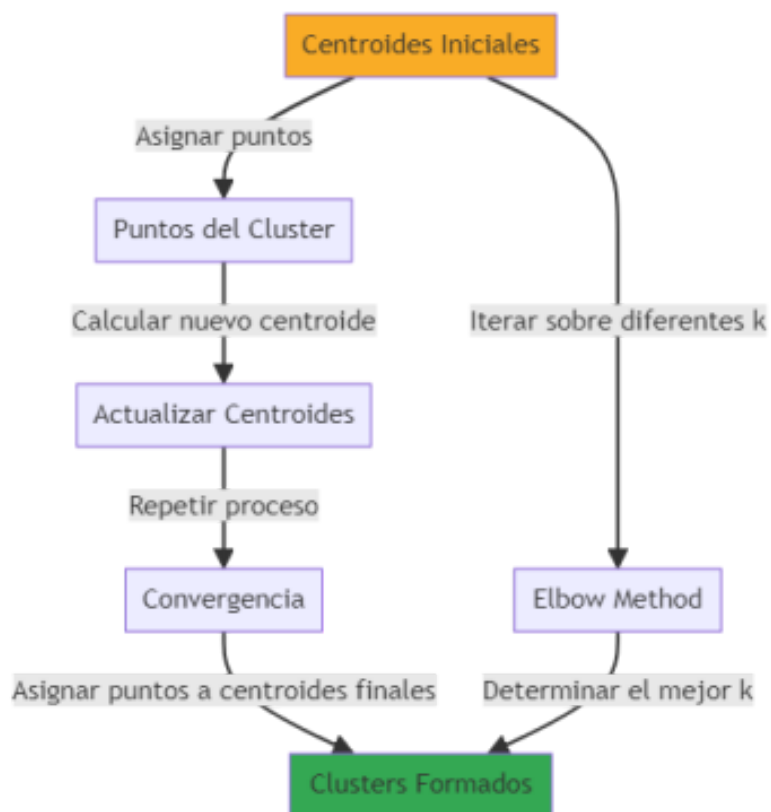


Fig 13: Flujo de uso K-means.

Random Forest (Supervisado)



Fig 14: Ejemplo de uso RF.

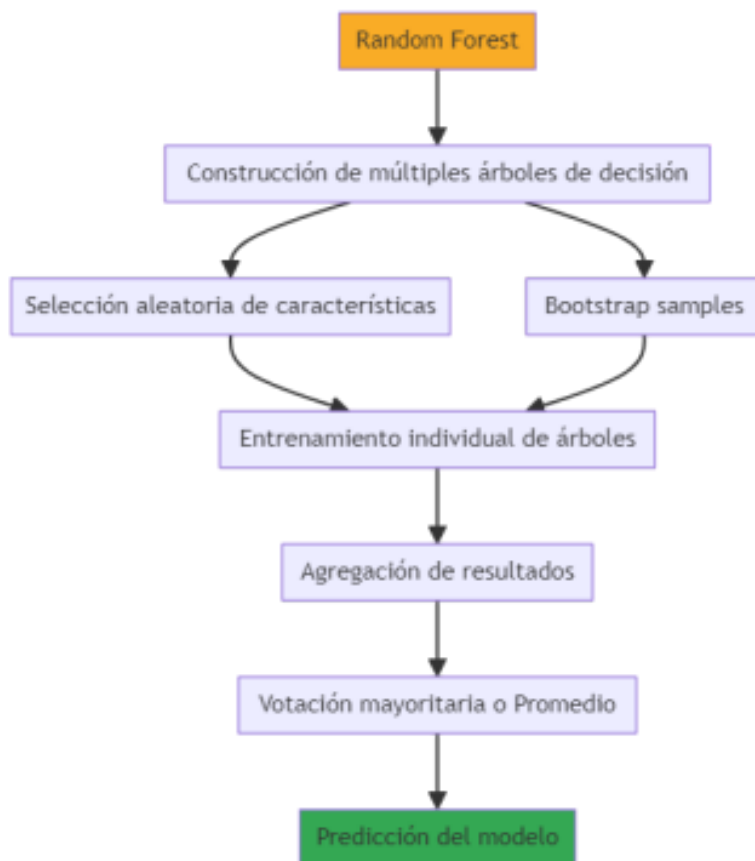


Fig 15: Flujo de uso RF.

Flujo Trabajo

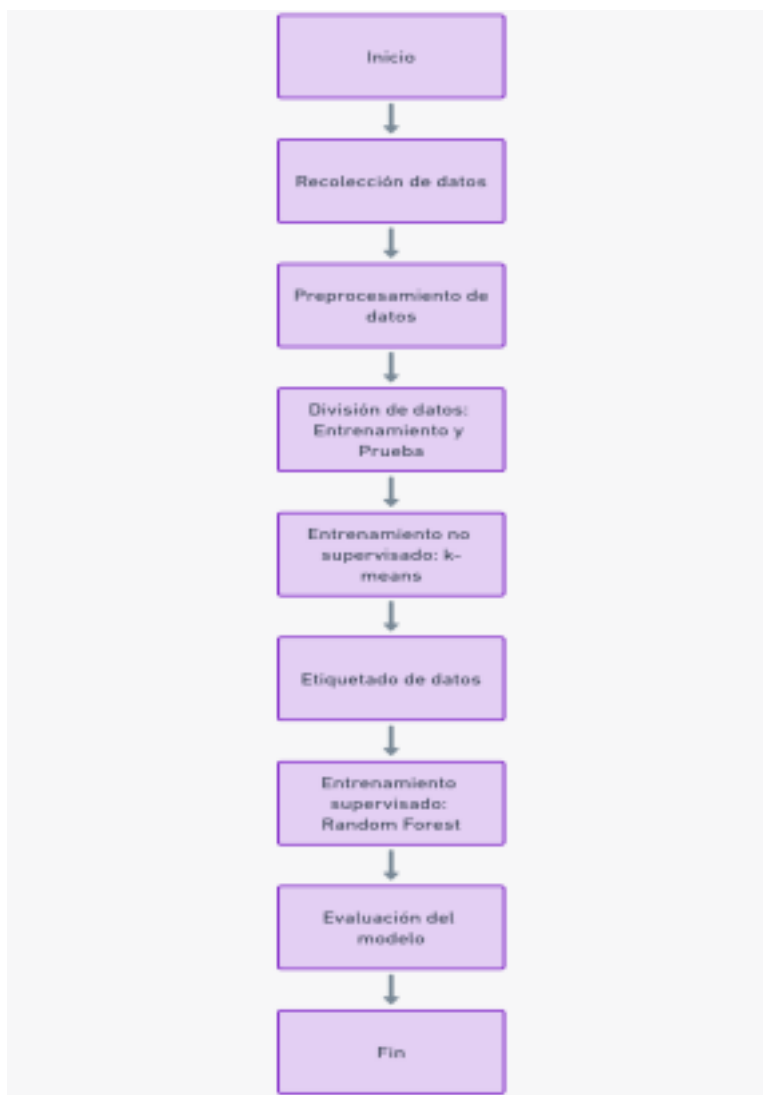


Fig 16: Flujo Trabajo.

Flujo Económico

	2023	2024	2025
Costo almacenamiento	\$ 65	\$ 420	\$ 420
Consultoría Senior	\$ 50	\$ 400	\$ -
Costo implementación	\$ 250	\$ -	\$ 1.200
Flujo de Costos	\$ 365	\$ 820	\$ 1.620

Tabla de costos fijos relevantes	
Ítem	Valor Mensual
Costo Desarrollador	\$ 500
Costo Infraestructura Servidor	\$ 100

Fig 17: Costos Fijos Y Variables.

	nov-23	dic-23	ene-24	feb-24	mar-24	abr-24	may-24	jun-24	jul-24	ago-24	sept-24	oct-24	nov-24	dic-24	ene-25	feb-25	mar-25	abr-25	may-25	jun-25	jul-25	ago-25	sept-25	oct-25	nov-25	dic-25
Costo Desarrollador	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500	\$ 500
Costo Infraestructura Servidor	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100
Costo almacenamiento	\$ 30	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35	\$ 35
Consultoría Senior	\$ 50	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -	\$ -
Costo implementación	\$ 100	\$ 150	\$ 150	\$ 250	\$ 400	\$ 700	\$ 700	\$ 700	\$ 700	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100	\$ 100
Flujo de Costos	\$ 730	\$ 835	\$ 885	\$ 985	\$ 1.135	\$ 1.435	\$ 1.335	\$ 1.335	\$ 1.335	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235	\$ 235

Fig 18: Flujo Económico.

Link Repositorio

<https://github.com/AVergara97/Proyecto-Pasantia>

Anexos

Anexo 1: Causas posibles

1. Falta de una solución adecuada: La empresa actualmente no tiene un modelo de datos escalable diseñado para manejar grandes volúmenes de datos y detectar patrones sospechosos de fraude financiero y lavado de activos. Esto ha llevado a problemas de capacidad de almacenamiento y procesamiento, lo que ha impedido el monitoreo y la detección efectiva de actividades sospechosas.
2. Limitaciones técnicas: La infraestructura técnica existente de la empresa puede no ser lo suficientemente robusta para manejar grandes volúmenes de datos. Las herramientas y sistemas actuales pueden no ser escalables y no cumplir con los requisitos de la empresa para detectar patrones sospechosos de fraude y lavado de activos.
3. Falta de recursos especializados: La empresa puede carecer de personal técnico con experiencia y conocimientos especializados en el diseño e implementación de soluciones escalables para el monitoreo de fraude financiero y lavado de activos. La falta de personal adecuado puede restringir la capacidad de la empresa para implementar una solución efectiva.
4. Regulaciones cambiantes: Las regulaciones gubernamentales y de la industria están en constante evolución y se actualizan regularmente. La falta de una solución escalable puede impedir el cumplimiento de las regulaciones actuales, lo que puede resultar en multas y sanciones financieras para la empresa.
5. Falta de integración de datos: La empresa puede estar utilizando varias fuentes de datos que no están integradas entre sí, lo que dificulta el monitoreo y la detección de patrones sospechosos de fraude y lavado de activos.
6. Falta de herramientas de análisis adecuadas: La empresa puede tener herramientas de análisis de datos limitadas o desactualizadas que no pueden manejar grandes volúmenes de datos o detectar patrones complejos de manera efectiva.
7. Falta de capacitación del personal: El personal de la empresa puede no estar completamente capacitado para utilizar las herramientas y sistemas existentes o no estar al tanto de las mejores prácticas en la detección de fraude y lavado de activos.

8. Procesos ineficientes: La empresa puede tener procesos eficientes para la detección y verificación de actividades sospechosas, lo que puede retrasar la identificación de posibles casos de fraude o lavado de activos.
9. Fallas en la gestión de riesgos: La empresa puede no estar evaluando adecuadamente los riesgos asociados con actividades potencialmente fraudulentas o de lavado de activos, lo que puede llevar a la falta de monitoreo y detección de estos casos.

Causas Menores:

1. Falta de una solución adecuada:
 - a. Falta de comprensión de las necesidades de la empresa en términos de capacidad de almacenamiento y procesamiento.
 - b. Falta de recursos para adquirir o desarrollar una solución adecuada.
 - c. Falta de planificación a largo plazo para la implementación de una solución escalable.
2. Limitaciones técnicas:
 - a. Infraestructura técnica obsoleta o inadecuada.
 - b. Herramientas y sistemas actuales que no cumplen con los requisitos de la empresa.
 - c. Falta de recursos para actualizar o reemplazar la infraestructura técnica.
3. Falta de recursos especializados:
 - a. Falta de personal técnico con experiencia en la implementación de soluciones escalables.
 - b. Dificultad para encontrar y atraer talentos especializados.
 - c. Limitaciones presupuestarias que impiden la contratación de personal especializado.
4. Regulaciones cambiantes:
 - a. Falta de recursos para mantenerse actualizado sobre las regulaciones cambiantes.
 - b. Falta de comprensión de las implicaciones de las regulaciones actuales para el monitoreo y la detección de fraudes financieros y lavado de activos.
 - c. Falta de recursos para adaptarse a los cambios regulatorios.
5. Falta de integración de datos:

- a. Falta de comprensión de cómo integrar diferentes fuentes de datos para un monitoreo y detección efectiva de fraudes y lavado de activos.
 - b. Dificultad para acceder a fuentes de datos relevantes.
 - c. Limitaciones técnicas que impiden la integración de datos.
6. Falta de herramientas de análisis adecuadas:
- a. Falta de comprensión de las necesidades de la empresa en términos de herramientas de análisis.
 - b. Herramientas de análisis desactualizadas o limitadas.
 - c. Falta de recursos para adquirir o desarrollar herramientas de análisis adecuadas.
7. Falta de capacitación del personal:
- a. Falta de recursos para proporcionar capacitación adecuada.
 - b. Falta de motivación del personal para aprender nuevas habilidades y conocimientos.
 - c. Falta de una cultura de aprendizaje y desarrollo en la empresa.
8. Procesos ineficientes:
- a. Procesos de detección y verificación de casos de fraude y lavado de activos que son demasiado complejos o que toman demasiado tiempo.
 - b. Procesos que no están claramente definidos y documentados.
 - c. Falta de herramientas y sistemas automatizados para acelerar el proceso de detección y verificación.
9. Fallas en la gestión de riesgos:
- a. Falta de un enfoque estructurado para la gestión de riesgos en la empresa
 - b. Falta de recursos para llevar a cabo evaluaciones de riesgos adecuadas.
 - c. Falta de compromiso de la alta dirección en la gestión de riesgos.

Anexo 2: Algoritmos

K-means (Aprendizaje no Supervisado)

El algoritmo K-means es una técnica de aprendizaje no supervisado que se utiliza para clasificar datos en diferentes grupos (o clusters). La idea es dividir un conjunto de datos en k grupos basados en las características de los datos. K-means identifica k centros de clusters y asigna cada punto de datos al centro más cercano, basándose en la distancia euclidiana.

Imagine que tiene un conjunto de transacciones financieras representadas en un gráfico bidimensional como puntos. K-means funcionaría de la siguiente manera:

Inicialización: Se seleccionan aleatoriamente k puntos como los centros de los clusters. Asignación: Cada transacción (punto) se asigna al centro de cluster más cercano. Actualización: Se recalcula el centro de cada cluster tomando el promedio de todos los puntos asignados a ese cluster.

Repetición: Los pasos 2 y 3 se repiten hasta que la asignación de puntos a clusters ya no cambia significativamente.

Random Forest (Aprendizaje Supervisado)

Random Forest es un algoritmo de aprendizaje supervisado que utiliza múltiples árboles de decisión para realizar predicciones. Cada árbol se entrena con una muestra aleatoria de los datos y realiza su propia predicción. El resultado final del Random Forest es la combinación (por ejemplo, la moda en clasificación) de las predicciones de todos los árboles individuales.

En el contexto de la detección de fraude, un Random Forest podría entrenarse con datos históricos etiquetados como 'fraudulentos' o 'no fraudulentos'. Cada árbol de decisión en el bosque consideraría diferentes atributos y sus interacciones, como el monto de la transacción, la frecuencia, el origen, etc., para identificar si una nueva transacción parece ser fraudulenta.

Anexo 3: Flujo de Trabajo

Aquí hay una breve explicación de cada paso del flujo de trabajo:

1. Inicio: Se marca el comienzo del proceso de detección de fraude.
2. Recolección de datos: Este es el primer paso operativo donde se recopilan los datos necesarios para el análisis. Estos pueden incluir transacciones financieras, registros de clientes y otros datos relevantes.
3. Preprocesamiento de datos: Los datos crudos se limpian y se preparan para el análisis. Esto puede incluir la eliminación de valores atípicos, la normalización de variables y el manejo de valores faltantes.
4. División de datos: Entrenamiento y Prueba: Los datos preprocesados se dividen en dos conjuntos. Uno para entrenar el modelo (entrenamiento) y otro para evaluar su rendimiento (prueba).
5. Entrenamiento no supervisado: k-means: Se utiliza el algoritmo K-means para explorar los datos y encontrar patrones naturales o grupos sin etiquetas previas. Esto puede ayudar a identificar comportamientos inusuales que podrían indicar fraude.
6. Etiquetado de datos: Después del entrenamiento no supervisado, los datos son etiquetados, lo que significa que se asigna una categoría a cada grupo identificado por K-means, basado en alguna lógica o criterio específico, como puede ser la intervención de un experto.
7. Entrenamiento supervisado: Random Forest: Ahora que se tienen datos etiquetados, se entrena un modelo de Random Forest. Este modelo aprende de los datos etiquetados para poder hacer predicciones sobre nuevos datos.
8. Evaluación del modelo: Una vez entrenado el modelo, se evalúa su rendimiento utilizando el conjunto de prueba y varias métricas, como la matriz de confusión y la curva de precisión-recall.
9. Fin: El proceso termina, idealmente con un modelo capaz de detectar actividades fraudulentas de manera eficiente.

El flujo de trabajo muestra un enfoque combinado de aprendizaje no supervisado y supervisado, aprovechando las fortalezas de ambos para mejorar la precisión en la detección de fraude.

Bibliografía

- Raghavan, P., & El Gayar, N. (2019, December). Fraud detection using machine learning and deep learning. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 334-339). IEEE.
- Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oblé, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. *Information sciences*, 557, 317-331.
- Shirgave, S., Awati, C., More, R., & Patil, S. (2019). A review on credit card fraud detection using machine learning. *International Journal of Scientific & technology research*, 8(10), 1217-1220.
- Álvarez-Jareño, J. A., Badal-Valero, E., & Pavía, J. M. (2017). Using machine learning for financial fraud detection in the accounts of companies investigated for money laundering. *Economics Department, Universitat Jaume I, Castellón (Spain)*, (2017/07).
- JPMorgan Chase & Co. (2023). Estrategias de prevención de fraude financiero. Sitio web de JPMorgan Chase & Co. [<https://www.jpmorganchase.com>]
- Amazon. (2023). Tecnologías de procesamiento de datos en Amazon. Sitio web de Amazon. [<https://www.amazon.com>]
- PayPal. (2023). Sistemas avanzados de detección de fraudes en PayPal. Sitio web de PayPal. [<https://www.paypal.com>]
- Unidad de Análisis Financiero. (2023). Prevención del lavado de activos y financiamiento del terrorismo. Sitio web de la UAF. [<https://www.uaf.gov.cl>]
- Comisión para el Mercado Financiero. (2023). Supervisión del mercado financiero en Chile. Sitio web de la CMF. [<https://www.cmfchile.cl>]