Universidad Nacional Federico Villarreal

VRIN | VICERRECTORADO DE INVESTIGACIÓN

FACULTAD DE INGENIERÍA ELECTRÓNICA E INFORMÁTICA

PROCESO DE INGENIERÍA DE DATOS APLICADO A UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA EFECTIVIDAD DE UNA ALERTA DE LAVADO DE ACTIVOS EN EL ÁREA DE CUMPLIMIENTO DE UNA ENTIDAD BANCARIA

Línea de investigación:
Sistema de Información y Optimización

Trabajo de Suficiencia Profesional para optar el Título Profesional de Ingeniero Informático

Autora

Aguilar Gutiérrez, Ashly Loren

Asesor

Rodriguez Rodriguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado

Solis Fonseca, Justo Pastor

Flores Masias, Edward José

Peña Carrillo, César Serapio

Lima – Perú

2026

PROCESO DE INGENIERÍA DE DATOS APLICADO A UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA EFECTIVIDAD DE UNA ALERTA DE LAVADO DE ACTIVOS EN EL ÁREA DE CUMPLIMIENTO DE UNA ENTIDAD BANCARIA

INFORME DE ORIGINALIDAD

6%	5%	1%	1%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	repositorio.autonoma.edu.pe Fuente de Internet	1%
2	alicia.concytec.gob.pe Fuente de Internet	<1%
3	Submitted to Universidad de Lima Trabajo del estudiante	<1%
4	www.lilab.io Fuente de Internet	<1%
5	revistas.pucp.edu.pe Fuente de Internet	<1%
6	www.researchgate.net Fuente de Internet	<1%
7	Submitted to Universidad de Guayaquil Trabajo del estudiante	<1%
8	repositorio.unc.edu.pe Fuente de Internet	<1%
9	Submitted to Unidades Tecnológicas de Santander Trabajo del estudiante	<1%
10	repositorio.upt.edu.pe Fuente de Internet	<1%
11	es.scribd.com Fuente de Internet	<1%

FACULTAD DE INGENIERÍA ELECTRÓNICA E INFORMÁTICA

PROCESO DE INGENIERÍA DE DATOS APLICADO A UN MODELO DE APRENDIZAJE AUTOMÁTICO PARA MEJORAR LA EFECTIVIDAD DE UNA ALERTA DE LAVADO DE ACTIVOS EN EL ÁREA DE CUMPLIMIENTO DE UNA ENTIDAD BANCARIA

Línea de investigación
Sistema de Información y Optimización

Modalidad de Suficiencia Profesional para optar el Título Profesional de Ingeniero Informático

Autora
Aguilar Gutiérrez, Ashly Loren

Asesor
Rodriguez Rodriguez, Ciro

ORCID: 0000-0003-2112-1349

Jurado
Solis Fonseca, Justo Pastor
Flores Masias, Edward José
Peña Carrillo, César Serapio

Lima – Perú
2026

Dedicatoria

A mis padres, por haberme inculcado buenos valores y acompañado en mi crecimiento profesional.

A mis hermanos, por ser una referencia para seguir adelante a pesar de los obstáculos que se puede presentar en la vida.

A mis sobrinos Ariadna y Jhosep por su motivación y aliento.

A mi novio quien siempre me apoya y me motiva en las metas que me propongo con amor y paciencia.

Agradecimientos

Agradezco profundamente a Dios por ser mi guía constante. A mi familia y a mi novio, por su amor incondicional, apoyo y motivación para seguir adelante. A los docentes de la FIEI por compartir sus conocimientos y contribuir a mi desarrollo profesional. A mis líderes del Tech Team de Cumplimiento, por su guía, confianza y por inspirarme a seguir creciendo en el ámbito tecnológico. Y finalmente, a mis amigos por su valiosa amistad y compartir conmigo sus experiencias en el apasionante mundo de los datos.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

RESUMEN

El presente informe se enfoca en el objetivo de implementar un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático, con el propósito de mejorar la efectividad de la generación de la alerta en una entidad bancaria. De esta manera, se busca perfeccionar los reportes preventivos ante la SBS y robustecer la lucha contra delitos de lavado de activos en el país. El proyecto se gestionó bajo el marco de trabajo ágil Scrum, respetando los estándares técnicos de la entidad bancaria para proyectos tecnológicos. La base de datos construida comprende 321,850 personas naturales y 41,011 personas jurídicas, la cual fue analizada mediante el algoritmo de aprendizaje automático Isolation Forest. Los resultados evidencian una mejora sustancial en la capacidad de detección, logrando que la efectividad de la alerta incremente de un 20% a un 45%, reduciendo significativamente los falsos positivos, y optimizando la carga operativa del equipo de investigaciones. El proyecto logró disponibilizar los datos dentro del flujo de extracción, y automatizar la generación de la alerta, cumpliendo con los objetivos planteados. Asimismo, este proyecto contó con el desarrollo de un equipo interno de la entidad bancaria, generando un ahorro económico al prescindir de proveedores externos y garantizando la sostenibilidad del mantenimiento futuro del modelo.

Palabras clave: Aprendizaje automático, efectividad, ingeniería de datos, lavado de activos, entidad bancaria.

ABSTRACT

This report focuses on the objective of implementing a data engineering process applied to a machine learning model, with the aim of improving the effectiveness of alert generation at a banking institution. In this way, the goa lis to enhance preventive reporting to the SBS and strengthen the fight against money laundering in the country. The Project was managed using the Scrum agile framework, in accordance with the bank's technical standards for technology projects. The database comprises 321,850 individuals and 41, 011 legal entities, which was analyzed using the Isolation Forest machine learning algorithm. The results show a substancial improvement in detection capabilities, increasing the effectiveness of alerts from 20% to 45%, significantly reducing false positives, and optimizing the operational workload of the investigative team. The Project successfully made the data available within the data extraction workflow and automated the generation of alerts, thereby meeting the established objectives. Furthermore, the Project was carried out b yan in-house team at the bank, resulting in cost savings by eliminating the need for external vendors and ensring the sustainability of future model maintenance.

Keywords: Machine learning, effectiveness, data engineering, money laundering, banking institution

I. INTRODUCCIÓN

En el ámbito financiero, el lavado de activos representa uno de los mayores desafíos que enfrentan las instituciones bancarias, particularmente en lo que respecta a transferencias internacionales de montos pequeños y recurrentes. Estas acciones corresponden a la técnica de ‘smurfing’ o fraccionamiento, que se distingue por la fragmentación intencionada de montos significativos en múltiples transacciones diminutas, concebidas para eludir los protocolos de supervisión y los límites de reporte impuestos por las regulaciones. (Financial Crimes Enforcement Network [FinCEN], 2005). Los sistemas analíticos ineficaces para identificar delitos financieros exponen a las entidades a riesgos reputacionales y operativos críticos. Según los estándares internacionales, estas deficiencias suelen materializarse en sanciones onerosas, la pérdida de financiamiento mayorista y un agotamiento de recursos humanos destinados a la resolución de fallas de cumplimiento (Basel Committee on Banking Supervision [BCBS], 2020).

Ante el incremento exponencial de los delitos financieros en los últimos años, el desarrollo de modelos predictivos se ha transformado en un factor determinante para garantizar la seguridad de las transacciones en las organizaciones bancarias. La literatura subraya que la efectividad en la identificación de operaciones sospechosas no depende solo del uso de inteligencia artificial, sino de la importancia de abordar aspectos específicos del modelado, resaltando que una fase de preparación y estructuración rigurosa de la información es lo que permite potenciar el rendimiento de los algoritmos. (Soria et al.,2024).

El informe tiene como objetivo mejorar la efectividad desde el momento en que se genera la alerta en el área de cumplimiento de una entidad bancaria. Para ello, se llevó a cabo un análisis exhaustivo del proceso actual y se elaboró una solución basada en un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático. Esta solución también destaca la necesidad

de adoptar un enfoque holístico a la hora de abordar el lavado de activos, contribuyendo a la defensa del sistema financiero y, por lo tanto, a la consolidación de la estabilidad económica y social del país.

1.1. Trayectoria del autor

Bachiller en Ingeniería Informática con más de cinco años desempeñándome como analista de datos, ingeniera de procesos y de datos. En la actualidad, estoy laborando como Data Engineer en el equipo de Data Value del Área de Cumplimiento del BCP. He participado en proyectos de BI, proyectos de estrategia de negocio con machine learning y proyectos de innovación con IA. Fui responsable de la exploración y extracción de datos, automatización de pipelines de datos, generación de reportes de visualización y participación en la identificación, definición y desarrollo de patrones de riesgos y oportunidades de mejora en eficiencia. Profesional apasionado por la interpretación y el análisis de datos para brindar conocimientos valiosos y soluciones efectivas.

1.1.1. Formación académica

Bachiller en Ingeniería Informática de la Universidad Nacional Federico Villarreal (marzo 2015-julio 2021).

1.1.2. Estudios complementarios

- Ceps UNI: Sql Server- Data Base Administrator.
- Big Data Academy: Programa de Especializaciòn en Big Data.
- DMC Perú:
  -Especialización en Python for Analytics
  -Especialización en Power BI y Visualización de datos.
  - Curso Oracle for BI

- Datapath:

  -Programa de Data Engineer

  - Especialización en Azure

1.1.3 Voluntariado

Institución de Ingenieros Eléctricos y Electrónicos de la Universidad Nacional Federico Villarreal:

- Miembro activo de la RE IEE UNFV (2016-2020)
- Presidenta de la Directiva General de la RE IEEE UNFV (2019)

1.1.4 Experiencia laboral

Data Engineer Data Value

Banco de Crédito del Perú (BCP)

Enero 2025 – Actual

Participación en proyectos estratégicos orientados al diseño e implementación de soluciones de datos alineadas con los objetivos del negocio. Lidero proyectos de recolección, almacenamiento, transformación y optimización de procesos de información, promoviendo la adopción de tecnologías en la nube. Se utilizan herramientas como Databricks (SQL y PySpark), Azure Blob Storage, SAS Enterprise y Oracle SQL, bajo la metodología Scrum y gestión de requerimientos con Jira.

Data Engineer Innovación

Banco de Crédito del Perú (BCP)

Enero 2024 – diciembre 2024

Participación en proyectos piloto centrados en la mejora del manejo de datos transaccionales, perfilamiento de clientes, control de calidad de reportes y análisis de datos web

(Copilot Web).En este rol, se colabora con equipos ágiles aplicando metodología Scrum, y equipos técnicos para garantizar la calidad y utilidad de los datos en soluciones innovadoras, esto incluye: Levantamiento de información, análisis de modelos, diseño de variables, ingesta y transformación de datos, automatización de pipelines de datos, y colaboración interdepartamental.

Data Engineer Analytics

Banco de Crédito del Perú (BCP)

Julio 2023 – diciembre 2023

Participación en proyectos clave para la implementación de escenarios de alerta relacionados con transacciones internacionales y venta responsable, fortaleciendo los procesos de cumplimiento y prevención en el programa PLAFT. Las responsabilidades técnicas incluyen: Análisis de modelos, definición de variables, diseño y desarrollo del ETL, modelamiento con Python, automatización de pipelines de datos, manejo de repositorios y control de versiones con Bitbucket y gestión de proyectos con metodología ágil Scrum.

Analista Senior BI

Banco de Crédito del Perú (BCP)

Abril 2022 – junio 2023

Participación en proyectos clave enfocados en la optimización y migración de herramientas de análisis, así como en la implementación de indicadores de calidad y completitud de datos, fortaleciendo la toma de decisiones basada en datos. Las responsabilidades técnicas incluyen: Mantenimiento y desarrollo de consultas en SQL, automatización de flujos de datos, migración de dashboards, implementación de indicadores de calidad y completitud y optimización de procesos.

Analista Junior BI

Banco de Crédito del Perú (BCP)

Mayo 2021 – marzo 2022

Contribución en proyectos de automatización, optimización de indicadores y generación de reportes adhoc, apoyando procesos de análisis de datos y monitoreo de riesgos. Las responsabilidades técnicas incluyen: Desarrollo en SQL, automatización de flujos de datos, automatización de procesos automatizados adhoc y optimización de indicadores.

Asistente BI

Banco de Crédito del Perú (BCP)

Abril 2020 – abril 2021

Contribución en el diseño, análisis y elaboración de informes de indicadores de riesgo, además de la implementación y mantenimiento de procesos automatizados para la gestión de datos. Las responsabilidades técnicas incluyen: Elaboración de informes, consultas SQL, automatización de procesos e implementación de indicadores.

Practicante BI

Banco de Crédito del Perú (BCP)

Octubre 2019 – marzo 2020

Contribución en la gestión de datos y mantenimiento de herramientas de visualización para facilitar el análisis y la toma de decisiones. Las responsabilidades técnicas incluyen: Identificación, extracción y limpieza de datos, consultas SQL y mantenimiento de dashboards.

1.2 Descripción de la empresa

El Banco de Crédito del Perú es una institución del sistema financiero peruano y proveedor líder de servicios financieros en el país. Además, representan el principal activo del grupo financiero Credicorp, el holding financiero más importante del Perú. Su objetivo social es

desarrollar todas aquellas actividades y operaciones que la legislación permite a las empresas bancarias (CIIU Nª 6519)

La historia de la institución se remonta a finales del siglo XIX. El Banco de Crédito del Perú se constituyó como sociedad anónima el 3 de abril de 1889 con el nombre de Banco Italiano. Seis días después se iniciaron sus operaciones. Sin embargo, el 21 de enero de 1942 adoptaron la razón social actual. La escritura pública se custodia en el Archivo General de la Nación, asentado a fojas 87 del protocolo de instrumentos públicos del notario Carlos Sotomayor y bajo el número 126. (GRI 2-1).

El Banco de Crédito del Perú ofrece una completa variedad de productos y servicios de banca corporativa, empresarial y personal a través de su red de distribución a nivel nacional. Al cierre de 2023, cuentan con la red de sucursales más grande entre los bancos peruanos, con 308 oficinas bancarias en todo el Perú, además de 2,439 cajeros automáticos, 10,683 Agentes BCP, una sucursal en Panamá y una agencia en Miami.

1.2.1 Propósito

Ser aliados de nuestros colaboradores, clientes y país para que transformen sus planes en realidad.

1.2.2 Aspiración

- Ser la empresa peruana que brinda la mejor experiencia a los clientes simple, cercana y oportuna.
- Ser la comunidad laboral de preferencias en el Perú, que inspira, potencia y dinamiza a los mejores profesionales
- Ser referentes regionales en gestión empresarial potenciando nuestro liderazgo histórico y transformador de la industria financiera en el Perú.

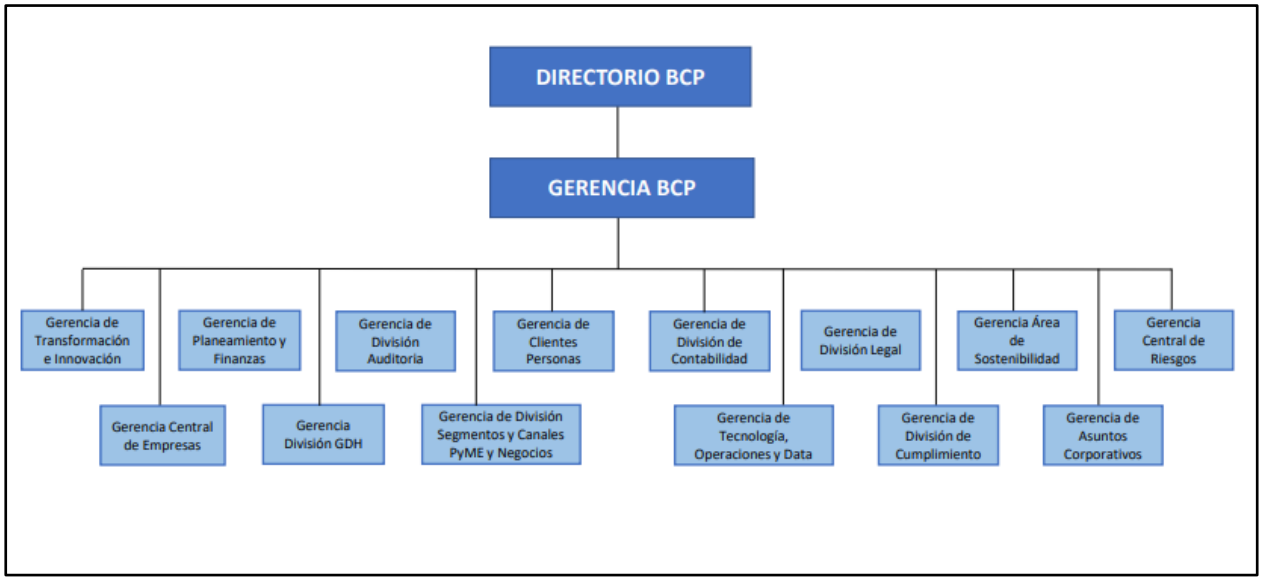
1.2.3 Principios

- Cliente céntrico
- Potenciamos nuestra mejor versión
- Sumamos para multiplicar
- Mínimo damos lo máximo
- Emprendemos y aprendemos
- Seguros y derechos.

1.3 Organigrama de la empresa

Figura 1

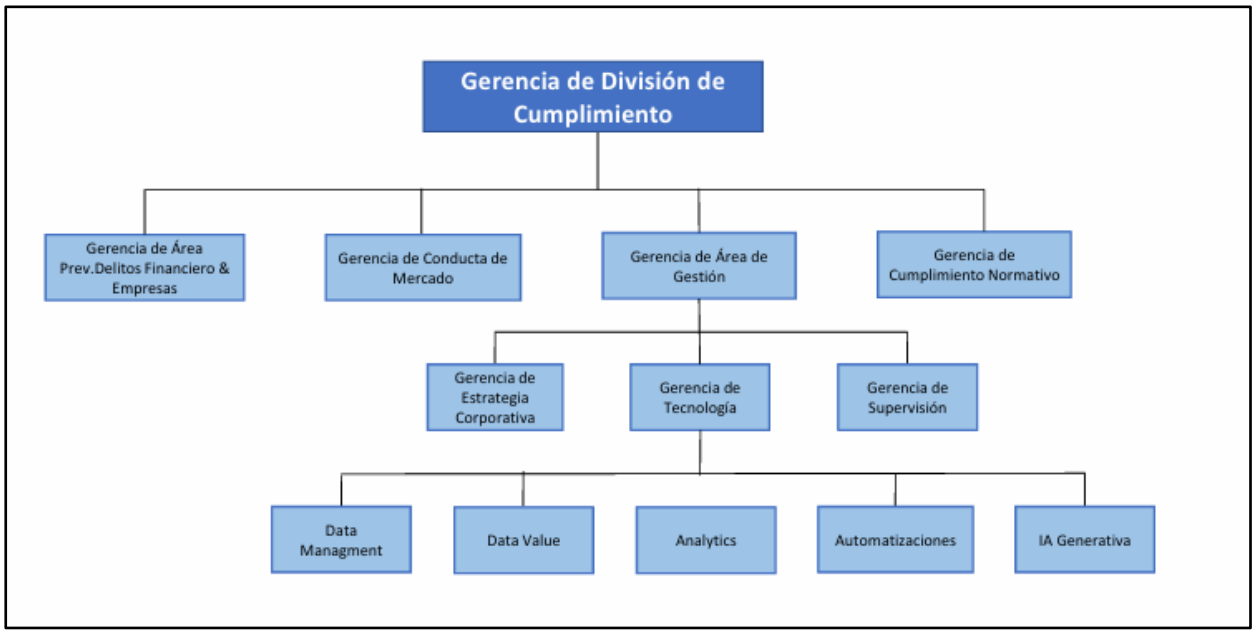
Organigrama de la Empresa Directorio



Nota. La figura muestra un organigrama funcional de tipo jerárquico y vertical, en el cual la estructura organizacional se encuentra distribuida por áreas especializadas que dependen directamente de la Gerencia General, la cual a su vez responde al Directorio. Este organigrama facilita la comprensión de la relación entre las áreas estratégicas y operativas de la entidad. La gerencia en la que se enfoca el presente informe es la Gerencia de División de Cumplimiento.

Figura 2

Organigrama del Área de Cumplimiento



Nota. La Gerencia de División de Cumplimiento tiene una modelo funcional jerárquico y vertical, en el cual las áreas se organizan por especialización y dependen directamente de la gerencia de división. El organigrama permite visualizar de forma clara la jerarquía, distribución de funciones y articulación entre áreas técnicas y de cumplimiento, siendo relevante la Gerencia d Tecnología para el desarrollo del presente informe.

1.4 Áreas y funciones desempeñadas

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Data Engineer Data Value

Enero 2025 – Actualidad

- Analizo las necesidades estratégicas de información para diseñar soluciones de datos alineados a objetivos de negocio.
- Lidero proyectos de recolección, almacenamiento, transformación y optimización de procesos de información, promoviendo la adopción de tecnologías cloud.

- Diseño e implemento nuevas soluciones end-to-end en entornos Oracle y Azure, utilizando herramientas como Databricks (SQL y PySpark), Azure Blob Storage, SAS Enterprise y Oracle SQL.
- Atención y desarrollo de requerimientos de información solicitados por usuarios de equipos clave, aplicando metodología ágil Scrum y herramienta Jira.
- Proyecto-Automatización del control de calidad en Archivo Central:
- Implementé solución de data utilizando Databrick SQL, Blob Storage y SAS, generando un ahorro de 80 hh anuales y mejorando la satisfacción del equipo de Transparencia Fiscal.
- Proyecto- Optimización del proceso de identificación de PJ reportables
- Implementé una solución de data, utilizando Pyspark, SQL, Oracle, Blob Storage y SAS, logrando un ahorro de 20 hh anuales y mejorando la eficiencia operativa del equipo de Transparencia Fiscal.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Data Engineer Innovación

Enero 2024 – diciembre 2024

- Colaboré con el levantamiento de requerimientos con usuarios clave para identificar necesidades de información.
- Participé en el análisis de procesos operativos (estado As-Is), e identificación de oportunidades de mejora para diseñar flujos To-Be eficientes.
- Lideré el diseño del flujo de datos para la ingesta, modelado y exportación de información utilizando Azure Blob Storage, Databricks SQL, Oracle-SQL, Batch Scripts, Power Automate, AI Builder y Copilot Web en los pilotos:
  - Nuevo Paquete Transaccional de Clientes: Diseñé una solución de datos que permitió reducir

en 40% el tiempo de análisis del equipo de investigaciones PLAFT.

- Control de calidad de Informes de Cierre de alertas PLAFT: Implementé un modelo con precisión superior al 90% de efectividad, optimizando la validación de perfiles de cliente y sustento documental.

- Resumen de Noticias Web: Automaticé la lectura y resumen de noticias mediante inteligencia artificial, logrando un ahorro de 790 hh anuales para el equipo de investigaciones PLAFT.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Data Engineer Analytics

Julio 2023 – diciembre 2023

- Participé en el análisis, diseño y construcción del modelo conceptual de datos para proyectos de analytics.
- Desarrollé procesos ETL en SQL y generación de logs automatizados utilizando Batch Scripts.
- Diseñé, construí y orquesté un pipeline de datos end-to-end que habilitó la puesta en producción de un modelo de machine learning, utilizando SQL, Python y Batch Scripts.
- Utilicé Git y Bitbucket para el desarrollo, test y puesta en producción de los proyectos de analytics.
- Utilicé Jira para el registro de tareas y desarrollo de proyectos aplicando la metodología ágil Scrum.
- Proyecto-Implementación de modelos ML para alertas PLAFT:
- Lideré el diseño del flujo de datos, construcción del pipeline y la implementación que permitió poner en producción dos escenarios clave de alerta sobre Transacciones de pagos con el exterior y Venta responsible.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Analista Senior BI

Abril 2022 – junio 2023

- Diseñé y desarrollé pipelines de datos automatizados para la ingesta, procesamiento y exportación de información utilizando SQL, Python, Batch Scripts y visualiazación en Power BI.
- Migré tableros clave de Qliksense a Power BI, mejorando la experiencia del usuario y centralizando la gestion de dashboards.
- Proyecto-Construí un pipeline de datos para integrar fuentes, automatizar el cálculo de Kpis y entregar resultados al cliente final utilizando SQL y Batch Scripts, logrando implementar 4 indicadores de completitud y calidad en el proceso de Onboarding de clientes digitales y presenciales.
- Proyecto-Optimicé proceso de integración y validación de datos mediante la mejora de Scripts SQL, lo que permitió incrementar la calidad y completitud de datos en un 50% en el proceso de Debida Diligencia de Clientes.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Analista Junior BI

Mayo 2021 – marzo 2022

- Desarrollé y mantuve dashboards operativos utilizando Oracle-PLSQL.
- Construí pipeline de ingesta para integrar fuentes internas y externas mediante SQL y Batch Scripts.

- Proyecto-Optimicé fuentes de ingesta, lógicas de extracción, y carga de datos utilizando SQL, logrando mantener el 70% de indicadores implementados dentro de los límites de control, y reduciendo la carga operativa del monitoreo de riesgos PLAFT.
- Proyecto-Diseñé, y orquesté un proceso ETL utilizando SQL y Batch Scripts, logrando implementar el 50% de indicadores de riesgo de lavado PLAFT, y generando nuevos insights para la toma de decisiones estratégicas.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Asistente BI

Abril 2020 – abril 2021

- Construí scripts SQL.
- Di soporte a los procesos automatizados de macros en Excel.
- Participé en el proyecto BI – Implementación de Indicadores de riesgo PLAFT: Análisis, extracción de datos, creación de lógicas personalizadas para los indicadores.

Área: Banco de Crédito del Perú (BCP) – División de Cumplimiento y Ética

Cargo: Practicante BI

Octubre 2019 – marzo 2020

- Ejecuté consultas SQL.
- Di soporte al mantenimiento y actualización de dashboards en Power BI
- Diseñé y construí un módulo de mantenimiento para un Dashboard de indicadores Web.

II. DESCRIPCIÓN DE UNA ACTIVIDAD ESPECÍFICA

2.1 Planteamiento del problema

2.1.1 Determinación del problema

En el ámbito internacional, las técnicas de lavado de dinero han evolucionado para adoptar formas cada vez más sofisticadas como la modalidad de “smurfing” o fraccionamiento de operaciones. Según un estudio señala que el smurfing está relacionado con la inyección de dinero ilícito en sistemas financieros mediante múltiples transacciones pequeñas en un corto periodo de tiempo, y que los métodos tradicionales de detección presentan elevadas tasas de falsos positivos. (Shadrooh, 2024). Además, se estima que el lavado de dinero representa entre el 2% y el 5% del Producto Bruto Interno (PBI) mundial, lo que significa que una tasa de detección imperfecta tiene consecuencias en el riesgo operativo, reputacional y regulatorio. (Go Beyond Biz, 2024).

En el contexto peruano, la modalidad de “smurfing” se caracteriza por introducir capitales ilícitos en el sistema financiero y luego dispersarlos mediante redes de cuentas, empresas y operadores, lo cual es una práctica recurrente identificada por las autoridades nacionales. (Jacopo, 2022). Esta práctica se inserta en un problema más amplio de incremento de fondos de origen ilícito que ingresan al sistema financiero peruano, fenómeno que ha sido documentado por la Superintendencia de Banca, Seguros y AFP(SBS) y la Unidad de Inteligencia Financiera (UIF-Perú) a través del creciente volumen de reportes de operaciones sospechosas (ROS) y de montos potencialmente lavados. (Ruffner, 2014).

El problema identificado en esta investigación radica sobre la baja efectividad en la generación de la alerta denominada “Estructuración de pagos del exterior” en el equipo de Analytics del área de Cumplimiento de una entidad bancaria. Este problema impacta negativamente en el sistema de monitoreo transaccional, al reflejarse en un elevado volumen de

alertas irrelevantes, lo cual dificulta la identificación oportuna de operaciones de riesgo real y puede permitir que pasen desapercibidos casos de lavado de activos bajo la modalidad de smurfing.

Las causas del problema se dividen en dos ejes críticos. Por un lado, la baja efectividad surge de parámetros desactualizados en la disponibilidad de datos ante el crecieminto de transacciones digitales, impulsado por la adopción masiva de canales como Yape y Plin POS. Esta desincronización técnica amplía la brecha entre la data transaccional y el modelo de alerta. Asimismo, la persistencia de flujos manuales y no optimizados genera un entorno propenso al retraso, donde el 80% del esfuerzo del equipo de investigaciones se desperdicia en casos sin valor real. Finalmente, la situación se agrava por una ingeniería de datos ineficaz que no logra generar variables predictivas robustas, dejando al aprendizaje automático sin los recursos necesarios para detectar patrones complejos de lavado de activos.

Las consecuencias del problema son considerables. Como principal punto, la baja efectividad en las alertas impide que se realicen los reportes correspondientes ante el regulador. Esta omisión expone a sanciones económicas que podrían alcanzar los US$ 2,000,000.00, comprometiendo seriamente la reputación la entidad financiera. Por otro lado, el crecimiento de falsos positivos obliga al equipo de investigaciones a analizar casos que carecen de relevancia ilícita, lo cual disminuye la eficiencia operativa del equipo lo que representa 266.7 horas hombre al mes.

Ante este panorama, el presente informe de trabajo propone implementar un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático para mejorar la efectividad de la generación de la alerta denominada “Estructuración de pagos del exterior”. Con la propuesta, se busca no solo reducir drásticamente la tasa de falsos positivos, sino también mejorar la detección de actividades de lavado de activos. De este modo, se contribuye a un ecosistema de cumplimiento

más robusto, optimizando los recursos de la entidad y minimizando los riesgos de sanciones por parte del regulador.

2.1.2 Problema general

¿Cómo un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático mejora la efectividad de la alerta de lavado de activos en el área de Cumplimiento de una entidad bancaria?

2.1.3 Problemas específicos

- ¿Cómo un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático impacta en la disponibilidad de los datos en la generación de la alerta de lavado de activos?
- ¿Cómo un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático impacta en la optimización del flujo en la generación de la alerta de lavado de activos?

2.1.4 Objetivo principal

Implementar un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático para mejorar la efectividad de la alerta de lavado de activos en el área de Cumplimiento de una entidad bancaria.

2.1.5 Objetivos secundarios

- Impactar en la disponibilidad de los datos con la implementación de un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático en la generación de la alerta de lavado de activos.
- Impactar en la optimización del flujo con la implementación de un proceso de ingeniería de datos aplicado a un modelo de aprendizaje automático en la generación de la alerta de lavado de activos.

2.1.6 Justificación

La justificación teórica de este trabajo no radica en el uso de algoritmos, sino en la conexión que existe entre un proceso de ingeniería de datos robusto y la precisión de los resultados obtenidos. A menudo se comete el error de interpretar el procesamiento de datos como algo aislado; sin embargo, en este informe se plantea como el eje fundamental para que el modelo de aprendizaje automático sea efectivo y confiable. Con ello, se busca enriquecer el conocimiento académico sobre prevención de delitos financieros, ofreciendo una base teórica que sirva de guía para futuras optimizaciones de procesos analíticos en la entidad bancaria.

La justificación práctica radica en la reducción del volumen de falsos positivos, siendo el equipo de investigaciones los principales beneficiados. Al mejorar la efectividad en la generación de la alerta y descartar los falsos positivos, el equipo podrá optimizar su tiempo centrándose en casos de riesgo real, y así tener una respuesta más rápida ante operaciones sospechosas. Esto no solo alivia la carga operativa, sino que permite una toma de decisiones mucho más ágil y fundamentada frente a las exigencias regulatorias vigentes.

Según la Superintendencia de Banca, Seguros y AFP (SBS, 2024), entre 2010 y 2023 se registraron reportes de operaciones sospechosas por un monto que superó los US$ 1 802 millones, reflejando la persistencia de deficiencias en los mecanismos de detección y prevención dentro del sistema financiero peruano. Fortalecer estos mecanismos, permitirá a la entidad anticiparse a minimizar el riesgo de multas y sanciones por operaciones no detectadas, fortaleciendo su capacidad de prevención de delitos financieros. Finalmente, la solución contribuye al conocimiento del cumplimiento normativo y la mitigación del riesgo de lavado de activos.

2.1.7 Alcances y limitaciones

El presente estudio se desarrolló en el área de Cumplimiento de una entidad bancaria, centrándose exclusivamente en el proceso de generación de la alerta denominada “Estructuración de pagos del exterior (EPIC024)”. El análisis, diseño e implementación de la propuesta de solución se llevó a cabo tomando como base la experiencia profesional del autor como ingeniero de datos durante el año 2023.

En cuanto a las limitaciones, el estudio estuvo condicionado por el tiempo asignado para su ejecución, el cual correspondió a un trimestre, conforme a la planificación establecida por la entidad. Asimismo, se consideró como limitante los recursos de infraestructura computacional disponibles, los cuales estuvieron sujetos a los lineamientos, políticas internas y normas de seguridad tecnológica de la organización. Estas condiciones influyeron directamente en el desarrollo y en el alcance del proyecto.

2.2 Marco teórico

2.2.1 Antecedentes bibliográficos

Valencia (2024) presentó su tesis “Diseño de Metodología Integral de Ingeniería de Datos Empresariales para Optimización de Procesos y Toma de Decisiones Estratégicas”. Su objetivo consiste en mejorar los procesos de preparación, limpieza, calidad y gobernanza de datos en entornos empresariales. El diseño se desarrolló bajo una metodología integral de ingeniería de datos, aplicado a un caso empresarial del sector bancario, y verificando el rendimiento de cada fase. Como resultado se obtuvo que la metodología propuesta mejora la calidad de la información como la efectividad de los procesos para la toma de decisiones.

Luna (2024) presentó su tesis “Detección de fraude transaccional mediante modelos de aprendizaje automático: Una aplicación a una entidad financiera chilena”. Su objetivo se centró en realizar una comparación de la variedad de modelos de aprendizaje automático que puedan detectar fraudes en transacciones financieras y determinar el mejor. Su investigación fue de tipo teórica, aplicado a una data de prueba de 48.438 registros, data de validación de 58.126 registros y data de entrenamiento con 135.626 registros, utilizando como instrumento de medida la aplicación en los modelos de Regresión Logística, Redes Neuronales Artificiales, Máquinas de Vectores de Soporte, AdaBoost, CatBoost, Bosques Aleatorios y Naive Bayes, y XGBoost. Como resultado se concluyó que el modelo CatBoost logró el mejor rendimiento, el cuál combatirá el fraude y protegerá los activos de los clientes de la entidad bancaria.

Rayo (2020) presentó su tesis “Prototipo de detección de fraudes con tarjetas de crédito basado en inteligencia artificial aplicado a un banco peruano”. Su objetivo se centró en implementación de una solución que detecte el fraude en tiempo real con inteligencia artirficial y machine learning, sin depender de reglas preconfiguradas. Su investigación fue de tipo experimiental, aplicado a una data que comprende 3,960 transacciones tipo fraude y 56,047 tipo genuinas, distribuidas entre 80% como data de entrenamiento y 20% data de validación, utilizando como instrumento la aplicación del algoritmo Random Forest. Como resultado se obtuvo una precisión del 48.10% con un falso positivo de 4.35, el cual progresivamente mejorará el valor con el aumento del volumen transaccional y la ejecución del modelo.

Barboza (2023) presentó su tesis “Efectividad de un modelo de clasificación basado en Deep Learning en la detección de Covid-19”. Su objetivo consiste en determinar qué tan efectivo y eficiente es el modelo propuesto para la detección del virus. El diseño de la investigación es de tipo experimental, aplicado a una muestra de 642 imágenes de radiografías de COVID 19,

recolectados a través de 4 datasets que permiten entrenar, validar y evaluar el modelo propuesto. Como resultado se obtuvo que el modelo propuesto llega a una efectividad del 98%, y cuenta con una eficiencia de 18.1M parámetros y 3.2 GFLOPs, lo cual es fiable, efectivo y de bajo costo para la detección del COVID-19.

Alvarez et al. (2024) presentaron su tesis “Niveles de efectividad del modelo de redes neuronales en el análisis del comportamiento y predicción del PBI de Perú: 1980-2021”. Su objetivo está relacionado con la estimación del coeficiente de determinación para evaluar la efectividad del modelo de redes neuronales para el periodo de estudio. La investigación es de tipo descriptivo-aplicada, y la muestra correspondiente a la población de los datos trimestrales del Producto Bruto Interno del Perú se ha obtenido mediante el método de muestreo no probabilístico por conveniencia. Como resultado se obtuvo que el modelo alcanzó un 98.79% de la variabilidad observada en el PBI durante el periodo que se estudia, es decir que el modelo se ajusta peculiarmente bien a los datos observados del PBI de Perú.

Fuentes y Navarro (2025) presentaron su tesis “Implementación de un modelo basado en algoritmos de Machine Learning para el análisis predictivo de diagnósticos clínicos en REDLAB Perú S.A.C”. Su objetivo general se centró en optimizar la precisión y la rapidez en los diagnósticos clínicos. Realizó una investigación de tipo Aplicada-Explicativa, con diseño Pre-Experimental, empleando algoritmos supervisados como regresión logística y árboles de decisión, aplicado a una muestra de 26 pruebas médicas de laboratorio, y utilizando como instrumento de recolección las fichas de registro y cuestionarios. Como resultado se obtuvo una precisión entre el 98% y el 99%, reduciendo considerablemente los tiempos de análisis y mejorando la calidad de los diagnósticos clínicos. Además, la implementación de las herramientas tecnológicas permitió

optimizar los flujos de trabajo, y en consecuencia transformar los procesos de diagnóstico en el sector clínico-laboratorial.

Suxo (2024) presentó su tesis “Implementación de una solución de machine learning para la asignación de créditos financieros en una entidad bancaria”. Su objetivo fue proponer un modelo que permita estimar los ingresos potenciales, determinar tasas de interés y montos apropiados a los clientes top según la estimación. Realizó una investigación con la metodología TDSP y las etapas de desarrollo de modelos analíticos y probabilísticos de la entidad bancaria, aplicando la técnica de machine learning a una muestra de 23 000 clientes de negocios de la entidad bancaria. Como resultado se logró reducir el tiempo de atención y evaluación de los clientes en un 90%, lo cual elevó la satisfacción, también se logró elevar las colocaciones en un 65% y además de reducir la provisión en un 35%.

Inga et al. (2023) presentó su tesis “Implementación de técnicas de Machine Learning para la segmentación de clientes en una empresa del sector farmacéutico”. Su objetivo se centró en la investigación e implementación de técnicas de Machine Learning para una empresa del sector farmacéutico. Su investigación fue de tipo No experimental con diseño transversal, aplicado a una muestra de 30 mil transacciones comerciales relacionado con las compras de productos farmacéuticos, utilizando como instrumento de medida las técnicas de aprendizaje no supervisado, tales como K-Means y Clúster Jeráquico. Como resultado se concluyó que el modelo K-Means es el más adecuado para adaptarse a las necesidades de la empresa, ya que las agrupaciones se ajustaban más a la realidad del negocio.

2.2.2 Bases teóricas

2.2.2.1. Efectividad. Existen diferentes definiciones según:

Barboza (2023) retoma la definición de efectividad de Galia (2019), la cual integra los conceptos de eficacia y eficiencia. En sus palabras:

> Si consideramos que la eficacia se relaciona con el logro de metas específicas y la eficiencia se relaciona con lograr esas metas con el uso más eficaz de recursos limitados, la efectividad abarca ambos conceptos. En otras palabras, ser efectivo implica alcanzar las metas de manera eficaz y eficiente al mismo tiempo, y buscar la optimización de los recursos disponibles. (p. 26)

Al respecto, Oza (2018, como se citó en Luna, 2024) sostiene que una plataforma robusta debe identificar el fraude con alta precisión, evitando que los usuarios maliciosos lleven a cabo sus actos fraudulentos, y también no generar obstáculos que afecten a los clientes. El autor advierte que el incremento desmedido de falsos positivos puede impactar en la experiencia del cliente, generando una fricción operativa que, en última instancia, motiva a los clientes a abandonar la institución en busca de otras alternativas de servicios en el mercado.

Se considera que el marco conceptual de Oza (2018, como se citó en Luna, 2024) ofrece la perspectiva más robusta sobre la efectividad. Su valor académico radica en que no se limita exclusivamente en la precisión estadística sobre transacciones fraudulentas, sino que integra la eficiencia en el sistema operativo. Esta definición proporciona el equilibrio técnico-operativo requeridos para garantizar la detección oportuna de operaciones sospechosas y una experiencia satisfactoria para el usuario legítimo, lo cual constituye un criterio fundamental para el desarrollo de esta tesis.

2.2.2.2. Medición efectividad. La medición de la efectividad tiene diferentes interpretaciones según:

Triviño (2021) destaca la importancia de considerar la interacción organizacional en entornos digitales al momento de medir la efectividad:

> Se establece que la medición de la efectividad debe proporcionar un parámetro claro que aporte visión al logro de objetivos organizacionales, por esta razón, se debe tener en cuenta la naturaleza bidireccional de cualquier interacción entre una organización y un usuario, ya que la mayoría de los esfuerzos se enfocan en un ambiente digital que es altamente interactivo. Para esto, se debe asegurar que la medición de la efectividad refleje la creación de valor para los actores del sistema en términos de salidas y procesos, y también la satisfacción de los participantes en la interacción (p. 34).

Por su parte, Álvarez et al. (2024) explican cómo se mide la efectividad en modelos de redes neuronales aplicados a la predicción económica:

> Para medir la efectividad de un modelo de redes neuronales que se utiliza para predecir el Producto Bruto Interno (PBI), se puede usar diferentes métricas, pero una forma bastante habitual de hacerlo es usando el 'r cuadrado', con el cual se puede calcular el coeficiente de determinación (R^2) como una medida de cuánto se parecen los valores predichos a los valores observados. Se puede seguir este enfoque basado en la regresión lineal para calcular R^2. (p. 19).

Finalmente, Rayo (2020) propone una forma concreta y aplicada para medir la efectividad en sistemas de detección de fraude: "Uno de los indicadores de medición del proyecto, es la identificación del falso positivo, que permite identificar qué tan efectiva es la solución que permitirá identificar el fraude. Mediante dicho indicador, se podrá medir la efectividad del sistema para detectar el fraude en tiempo real" (p. 55).

A partir del análisis de las distintas definiciones, se concluye que la propuesta de Rayo (2020) se ajusta con mayor precisión al enfoque de esta investigación, ya que introduce una métrica concreta —los falsos positivos— que permite evaluar objetivamente la efectividad en contextos

críticos como la detección de fraude en tiempo real, combinando precisión técnica y aplicabilidad operativa.

2.2.2.3. Aprendizaje automático. Se define desde enfoques teóricos y prácticos:

Desde la óptica de Murphy (2012, como se citó en Luna,2024) sostiene que esta disciplina comprende una metodología orientada al reconocimiento automatizado de regularidades en la información, empleando estos hallazgos para generar inferencias precisas sobre eventos aún no observados. Partiendo de esta base, la potencia de los algoritmos radica en su autonomía evolutiva, prescindiendo de una intervención humana constante frente a nuevos volúmenes de datos.

Bajo una perspectiva integradora, Adepoju et al. (2029, como se citó en Rayo,2020) describen esta disciplina como una variante de la inteligencia computacional orientada al entrenamiento de sistemas para la detección de estructuras lógicas en almacenamientos masivos de la información. El valor más resaltante de esta disciplina es su facultad de autoperfeccionamiento, permitiendo que las computadoras mejoren sus propios modelos de manera orgánica, sin la necesidad de una supervisión manual recurrente para la interpretación de nuevas regularidades dentro de los datos.

Al respecto, Fontalvo et al. (2024) sostienen que esta disciplina se divide en dos metodologías predominantes según el tratamiento de los insumos: la supervisada y la no supervisada. En el primer esquema, los algoritmos trabajan con datos previamente clasificados y un objetivo de predicción claro, mientras que, en el enfoque no supervisado, los modelos exploran depósitos conjuntos de información sin etiquetar, orientándose a descubrir patrones internos por sí mismos, sin necesidad de un control previo.

Luego de contrastar las posturas teóricas, se considera que el marco conceptual de Fontalvo et al. (2024) proporciona el sustento más sólido, ya que su teoría reside en su capacidad para

establecer comunicación directa entre el aprendizaje automático y los datos, empleando el algoritmo adecuado según el caso para detección preventiva, permitiendo una interpretación precisa de la metodología aplicada para el diseño de la alerta.

2.2.2.4. Etapas de implementación de modelos ML. Para asegurar que los modelos de aprendizaje automático aporten un valor real en el entorno financiero, se necesita adoptar un marco de trabajo sistemático. Al respecto, Chapman et al. (2000, como se citó en Abdala, 2022) proponen la metodología CRISP-DM, la cual estructura el ciclo de vida del proyecto en seis dimensiones interconectadas que garantizan la viabilidad técnica y operativa. La etapa inicial del proyecto inicia con la comprensión del negocio, etapa donde se traducen las necesidades corporativas en una definición técnica de minería de datos. Posteriormente, la etapa del entendimiento de los datos, donde se valida la calidad y las limitaciones de las fuentes. La etapa de preparación es indispensable, ya que abarca la limpieza y el modelado estructural de los datos antes de ser procesados por los algoritmos. Finalmente, tras la modelación y evaluación, donde se calibran los parámetros, el ciclo culmina con el despliegue en la puesta en producción, estableciendo protocolos de seguimiento y mantenimiento.

Asimismo, Álvarez (2020) plantea que el despliegue técnico inicie con la ingeniería de características, etapa donde se analiza el histórico transaccional para consolidar variables explicativas y etiquetar la clasificación. Luego, el conjunto de datos se divide en dos grupos: uno se usa para entrenar el algoritmo y el otro para verificar su rendimiento. Con los datos de entrenamiento, se construye el modelo inicial. Posteriormente, se evalúan sus resultados comparando las predicciones obtenidas con los valores reales del grupo de prueba. Finalmente, se culmina con la calibración de parámetros definitiva del modelo para alcanzar los niveles de exactitud requeridos antes de la integración con el negocio. (Alvarez, 2020)

2.2.2.5. Ingeniería de datos. La ingeniería de datos se rige como un componente fundamental estructural que garantiza la viabilidad del aprendizaje automático. Al respecto, Mejía y Huaman (2024) definen esta disciplina como un flujo integrador de actividades que abarca desde la captura de registros hasta la entrega final de los datos como insumo analítico. Su enfoque destaca que el valor de la información no reside únicamente en su volumen, sino en la planificación del ciclo de vida del dato, garantizando mantener la integridad técnica y su operatividad en modelos automatizados. Por lo tanto, la ingeniería de datos deja de verse como un proceso aislado y pasa a ser el eje central del despliegue de modelos predictivos.

Desde una perspectiva orientada a la mitigación de riesgos financieros, Rodríguez (2022) define la ingeniería de datos como la gestión integral de la infraestructura técnica de la información, abarcando etapas críticas de limpieza, transformación y estandarización desde su captura inicial. En su propuesta, se establece una conexión entre el tratamiento de la información y la potencia predictiva de los esquemas de aprendizaje automático. Esta evidencia sostiene que, para los rubros de banca, la ingeniería de datos transciende la operatividad técnica para convertirse en el pilar estratégico de que los sistemas de detección alcancen los umbrales de precisión exigidos por el marco regulatorio de la institución.

Desde una perspectiva de madurez analítica, Santamaría (2023) define la ingeniería de datos como una arquitectura de flujos automatizados y escalables diseñados para garantizar alta disponibilidad de los datos de calidad en entornos productivos. Este enfoque, remplaza la visión tradicional de preparación de datos estática hacia un ciclo de vida de monitoreo y control de cambios, asegurando la sostenibilidad del modelo frente a la evolución constante de los datos. Dicha postura resulta adecuada para este estudio al integrar la automatización de datos como motor de la estabilidad operativa a lo largo de todo el sistema predictivo.

2.2.2.6. Rol de la ingeniería de datos en modelos de ML. Dentro del marco preventivo bancario, Rodríguez (2022) posiciona a la ingeniería de datos como el soporte estructural que potencia al aprendizaje automático. Su investigación plantea que la integridad y la disponibilidad oportuna de los datos, logrados mediante la madurez técnica y el conocimiento del entorno institucional, son requisitos indispensables para que los algoritmos logren detectar patrones atípicos. De este modo, se establece que la sinergia entre el funcionamiento tecnológico y el contexto operativo de la organización es el único camino para garantizar la viabilidad de un modelo frente a riesgos financieros.

Asimismo, Santamaría (2023) posiciona a la ingeniería de datos como la columna vertebral que sostiene la operatividad y confiabilidad del aprendizaje automático, en especial en la integración con prácticas de MLOps. Su función estratégica radica en el diseño de arquitecturas de flujos robustos y automatizados que garanticen un viaje ininterrumpido de la información hacia los sistemas de producción. Este enfoque comprende un ciclo del seguimiento constante sobre los cambios en los datos, actuando como el principal sustento técnico frente al deterioro de los algoritmos. Por lo tanto, implementar estos flujos permite que los modelos operen con una escalabilidad óptima, favoreciendo que el sistema evolucione de manera estable, posicionando a la ingeniería de datos como el pilar fundamental de la infraestructura analítica.

2.2.2.7. Etapas de la ingeniería de datos en modelos ML. Siguiendo los pasos que marca la ingeniería de datos, Aronés (2021) señala que el punto inicial clave es consolidar la información y asegurar que los datos sean de calidad y de utilidad para el proyecto. En su propuesta, detalla las herramientas de limpieza, transformación y estandarización de los datos que se utilizan, antes que la información pueda ser procesada por el algoritmo. Un paso vital es la división planificada de los datos en grupos: un grupo para que el modelo aprenda y otro para ser testeado, de esta manera

asegurar que los resultados finales sean reales y justos. Po último, se considera la creación de nuevas variables que puedan explicar mejor el problema, logrando que el algoritmo sea más certero en sus predicciones.

Por su parte, Fuentes y Navarro (2024) diseñaron un sistema inteligente para predecir resultados en el sector salud. Su propuesta consiste en obtener la información de sus propios sistemas, dar tratamiento y crear flujos automatizados para que el proceso se repita constantemente. Luego, con la data ya trabajada, pasar por diferentes algoritmos matemáticos como la regresión logística y los árboles de decisión, para evaluar con cual se acierta mejor y con cual se obtiene resultados más precisos y exactos. Esta forma de trabajo no solo facilita que el modelo aprenda, sino que el proceso se mucho más ágil y que la información se trate de manera ordenada en todo momento.

2.2.3 Definición de términos básicos

2.2.3.1. Alerta. Notificación automatizada generada por la ejecución del monitoreo de un modelo de machine learning, que se dispara cuando detecta comportamientos inusuales en el modelo de producción. (Bodor et al.,2023).

2.2.3.2. Anomalía. Conjunto de datos que se aparta significativamente del comportamiento habitual del sistema, donde se indica la presencia de eventos inusuales ó atípicos. (Aylin, 2025)

2.2.3.3. Apache spark. Es un sistema diseñado para procesar datos masivos con rapidez, dividiendo las tareas entre distintos equipos y aprovechar la memoria para obtener resultados eficientes en diferentes tipos de análisis. (Calvo,2018)

2.2.3.4. Calidad de datos. Es el grado en que un conjunto de datos cumple con los requisitos establecidos y es adecuado para el uso previsto en procesos de análisis, toma de decisiones o sistemas operativos. Es decir, la calidad no es una característica absoluta sino la

capacidad de los datos de ser “aptos para el uso” específico al que están destinados, lo cual implica que estos sean completos, consistentes, precisos y libres de errores que puedan afectar la interpretación o los resultados de un análisis. (Hassenstein & Vanella, 2022)

2.2.3.5. Dataset. Representa una agrupación organizada de información que ha sido recopilada para ser utilizad en operaciones de procesamiento, modelado o entrenamiento automatizado. (Databricks,2025)

2.2.3.6. Datawarehouse. Es un sistema que unifica datos de diferentes fuentes de manera estructurada, ofreciendo información histórica y confiable, para apoyar en el análisis, y toma de decisiones en procesos administrativos y estratégicos de una organización.

2.2.3.7. Datos. Se define como hechos, valores o registros que representan aspectos observables de la realidad y que, en su estado bruto, constituyen unidades de información que pueden ser analizadas y procesadas para generar conocimiento y apoyar la toma de decisiones. (Olson., 2021)

2.2.3.8. Estandarización. Se define como el proceso mediante el cual se transforman datos provenientes de diversas fuentes para que adopten un formato y una estructura uniformes y consistentes, de modo que puedan integrarse, compararse y utilizarse de forma eficiente en análisis y operaciones de ingeniería de datos. (Pure Storage, 2023)

2.2.3.9. ETL. Es un proceso de integración de datos que consiste en extraer los datos de diferentes fuentes, transformar los datos para que sean consistentes, y cargarlos a un repositorio para que sea explotado por otro proceso. (Encalada, 2025)

2.2.3.10. Lavado de activos. Proceso mediante el cual una persona o entidad introduce en el sistema económico y financiero bienes o recursos que provienen de actividades

delictivas, con propósito de ocultar su origen ilícito y darles apariencia de legalidad. (SBS, 2025)

2.2.3.11. Lenguaje SQL. Es un lenguaje diseñado para interactuar con la base de datos, donde se puede realizar tareas como buscar, modificar o eliminar datos, incluso construir operaciones complejas. (International Business Machines Corporation [IBM], s.f)

2.2.3.12. Modelo. Un modelo es el resultado de entrenar un algoritmo matemático con datos de entrada, los cuales procesando sus parámetros internos pueden generar predicciones confiables sobre nuevos casos, y tomar las mejores decisiones. (Mendoza et al., 2024).

2.2.3.13. On-premise. Se refiere a la implementación de sistemas tecnológicos dentro de la propia infraestructura empresarial, donde la entidad asume la responsabilidad completa sobre su instalación, gestión operativa y seguridad de los datos. (Vargas, 2023)

2.2.3.14. Oracle. Es una herramienta usada para manejar grandes cantidades de datos, que permite guardar, consular, y proteger la información de manera rápida y segura. (Oracle, s.f)

2.2.3.15. Pipeline. Un pipeline de datos es un proceso que permite conectar distintos pasos del procesamiento de información, desde su origen hasta su destino final, asegurando que la información viaje de forma continua y controlada hacia repositorios o modelos analíticos. (IBM, 2023)

2.2.3.16. Plsql. Es un lenguaje de procesamiento que permite combinar sentencias SQL para la manipulación y almacenamiento de los datos, que se ejecutan directamente en la base de datos. (Oracle, s.f)

2.2.3.17. Python. Es un lenguaje de programación que se caracteriza por su facilidad de uso, ya que su sintaxis es clara y comprensible, además, es multiplataforma, lo que permite ejecutar código en diversos entornos. (EAE Barcelona,2024)

2.2.3.18. Ros. Documento que elaboran las empresas para reportar a la UIF-Perú sobre eventos de lavado de activos o del financiamiento del terrorismo. Luego, este documento es revisado y posteriormente tramitado al Ministerio Público en caso exista actividades de lavado de activos y/o financiamiento del terrorismo. (SBS,2024)

2.2.3.19. Script. Reúne múltiples operaciones de extracción, transformación y carga, integrándolos en un solo archivo, y que permite ser ejecutado de manera automatizada. (TiDB Team,2024)

2.2.3.20. Scrum. Metodología ágil que permite desarrollar proyectos de manera iterativa y colaborativa, enfocándose en entregar valor continuo a través de ciclos cortos de trabajo, y obtener un mejor resultado en el proyecto. (Proyectos Ágiles, s.f)

2.2.3.21. Variable. Es un atributo cuantificable que describe cada caso del conjunto de datos y que alimenta al modelo. Las variables pueden ser numéricas, categóricas o binarias, y son esenciales para que el modelo pueda aprender y generar predicciones. (Domino Data Lab, 2025).

2.3 Propuesta de solución

2.3.1 Descripción de la propuesta

La presente propuesta de solución surge a partir del análisis de la problemática identificada en el área de Cumplimiento de una entidad bancaria donde se evidenció baja efectividad de la alerta denominada “Estructuración de pagos del exterior”.

La propuesta consiste incorporar un proceso de ingeniería de datos en un modelo de aprendizaje automático, lo que permitirá reducir falsos positivos y mejorar la detección de operaciones ilícitas, contribuyendo al cumplimiento normativo y alineándose con los objetivos estratégicos de la entidad bancaria.

Asimismo, esta propuesta se sustenta en base a la experiencia adquirida durante el desempeño profesional en la institución, donde se identificó los aspectos críticos del proceso y se propuso una alternativa de solución acorde al contexto y los recursos de la entidad. Se ha considerado el uso de tecnología On-premise y metodología ágil Scrum, que se adaptan a la infraestructura existente y al perfil del equipo involucrado en el proyecto.

En ese sentido, esta propuesta no solo busca resolver el problema identificado, sino también establecer las bases para la mejora continua con la posibilidad de ser replicada en otros procesos o áreas con características similares.

2.3.2 Desarrollo de la propuesta

En este capítulo se presenta el desarrollo de la propuesta planteada, la cual se basa en la experiencia profesional dentro del área de Cumplimiento de la entidad bancaria, así como el conocimiento adquirido durante la experiencia profesional.

El desarrollo contempla 7 etapas bajo la metodología ágil tipo SCRUM en Jira estructurados en 4 Sprints (2 semanas cada sprint), y considerando el entorno tecnológico On-premise de la organización.

Tabla 1

Desglose de Trabajo

SPRINT	ITEM	SEMANA
Sprint 1	Modelo conceptual y ETL	Semana 1 y 2
Sprint 2	EDA y Modelamiento (parte 1)	Semana 3 y 4
Sprint 3	Modelamiento (parte 2) y Test	Semana 5 y 6

Sprint 4	Producción y Documentación	Semana 7 y 8

2.3.2.1 Etapa 1 - modelo conceptual. En esta etapa se realizan las definiciones como objetivo, tipo de clientes, alcance, universo, productos, canales, meses de extracción, factores de riesgo, así como operaciones que se pueden filtrar. Los usos tecnológicos en esta etapa fueron el Teams y Power Point.

A. Revisión de antecedentes. Se reúne el equipo data science con el PO para analizar los antecedentes del proyecto vigente.

-La alerta denominada *"Estructuración pagos del exterior"* se originó por un caso remitido por Visa, donde se detectó un comportamiento poco habitual de una tarjeta de débito que recibió múltiples transacciones diarias por cifras redondas del exterior.

-Presenta alto número de volumen de alertas (400 aprox. mensual) con 20% de efectividad, generando alto número de falsos positivo (320 aprox.).

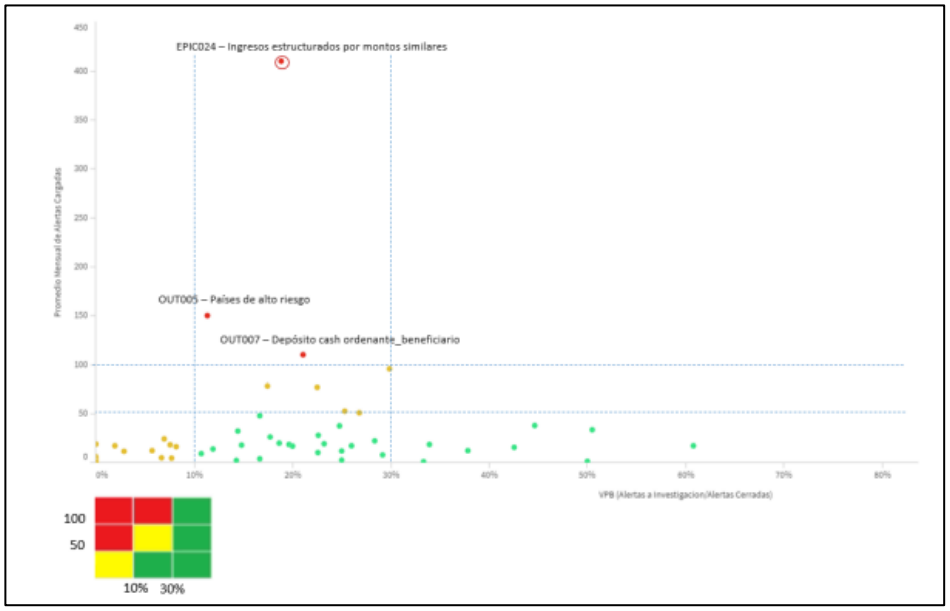
-Tiempo total mensual invertido en alertas falsas es de 266.7 horas, lo cual significa que esa cantidad de horas se pierde en alertas que no debieron generarse. En jornada laboral de 8 horas, se están consumiendo más de 33 días laborales al mes.

-Disponibilidad de los datos insuficiente debido al incremento sustancial en el volumen de transacciones de tipo Visa, generado por la incorporación del YAPE-PLIN POS, lo cual impactó en la capacidad de procesamiento y gestión de la información.

-Ausencia de un flujo estructurado automatizado.

Figura 3

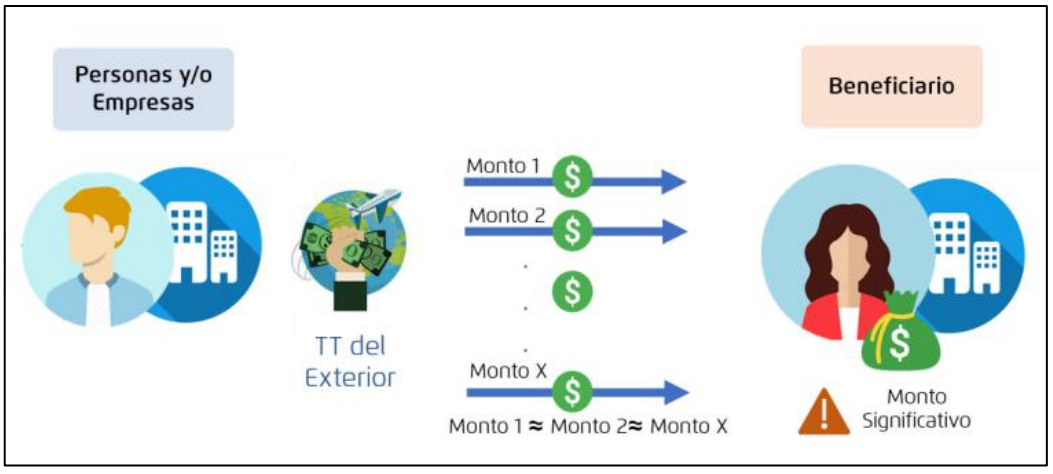
Antecedentes



B. ***Modelo conceptual.*** El equipo Data Science construye el flujo de trabajo del modelo y la segmentación de riesgo.

Figura 4

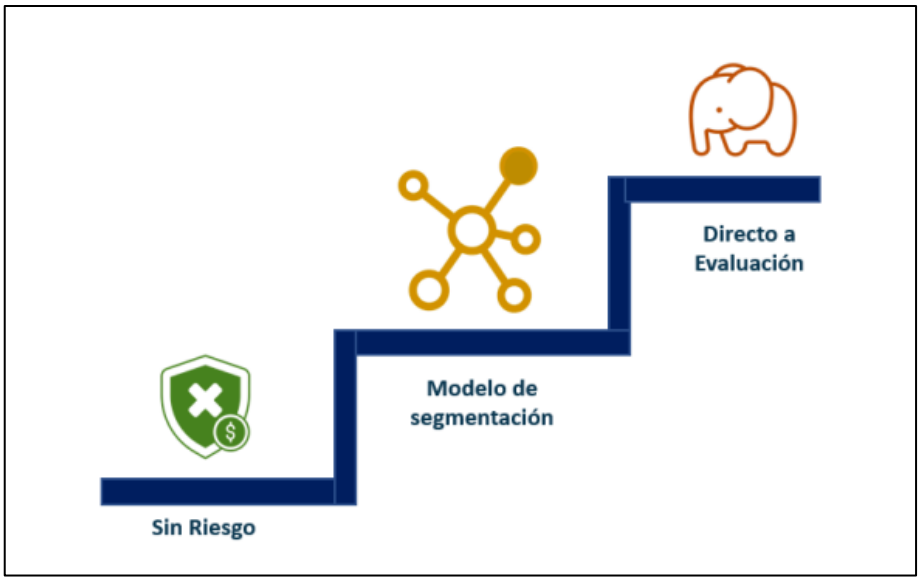
Propuesta de mejora



El modelo se abordó considerando un esquema de escalera para la segmentación de riesgo, con el propósito de diferenciar a los clientes con grandes diferencias entre sus ingresos recibidos, cuyos cortes fueron definidos por el negocio:

Figura 5

Segmentación Riesgo



PPNN:

-Escalón 1(sin riesgo): Clientes cuyo acumulado de ingreso en el mes es menor a US$ 150.

-Escalón 2(segmentación): Clientes cuyo acumulado de ingreso en el mes es mayor igual a US$ 150 Y menor igual a US$ 1MM.

-Escalón 3(Directo a evaluación): Clientes cuyo acumulado de ingreso en el mes es mayor a US$ 1MM

PPJJ:

-Escalón 1(sin riesgo): Clientes cuyo acumulado de ingreso en el mes es menor a US$1,100

-Escalón 2(segmentación): Clientes cuyo acumulado de ingreso en el mes es mayor igual a US$ 1,100 y menor igual a US$ 30 MM

-Escalón 3(Directo a evaluación): Clientes cuyo acumulado de ingreso en el mes mayor a US$30MM

C. Validación de definiciones. El equipo data science y data enginner se reúnen para realizar validaciones lógicas para la data de entrenamiento.

Universo: Transacciones de transferencias del exterior: Base Remittance y GGTT, específicamente las transacciones de: Transferencias exterior remesa y Exterior ventanilla.

Periodo de modelamiento:

-Periodo de análisis: abril 2023-setiembre 2023

-Periodo de test: octubre 2023

Variables:

Tabla 2

Listado Variables

VARIABLES	DESCRIPCIÓN	CATEGORÍA	TIPO	Uso para el modelo
CODCLAVECIC	ID del cliente		ID	No
PERIODO	Mes de análisis		Cuantitativa	No
TIPPER	Tipo de persona (PPNN ó PPJJ)		Cuantitativa	No
MTO_INGRESOS_MES	Monto acumulado dolarizado de ingresos del cliente en el mes		Cuantitativa	No
FLG_PERFIL_INGRESOS_3DS	Flag si sale de su perfil transaccional de ingresos en el	1:Sí / 0:No	Cualitativa	Si

	mes respecto a los 6 meses previos		
MTO_CONOTROSPROXIMOS	Monto acumulado del mayor monto repetido junto a valores +- 10% del monto máximo.	Cuantitativa	Si
CTD_CONOTROSPROXIMOS	Cantidad acumulada del mayor monto repetido junto a valores+-10% del monto máximo	Cuantitativa	Si

Consideraciones:

-No se incluyeron transacciones menores a $100.

-Los modelos fueron separados por PPNN y PPJJ.

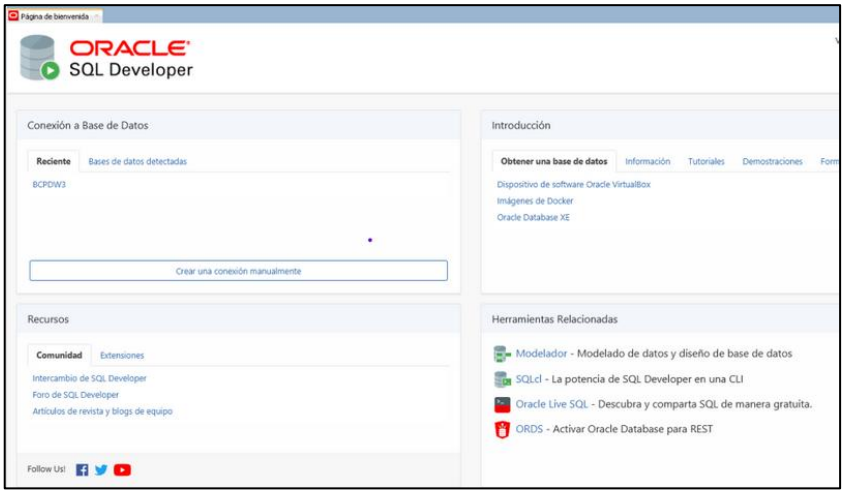
-Se utilizó una lista blanca de empresas para los clientes PPJJ.

2.3.2.2. Etapa 2 - ETL. En esta etapa el data engineer construye el ETL para disponibilizar la data de entrenamiento utilizando PL/SQL en Oracle. Este flujo permitió centralizar, transformar y disponer la información requerida para el modelado.

A. Fuentes de información. Las fuentes utilizadas corresponden a tablas transaccionales almacenadas en el Data Warehouse de la entidad Financiera.

Figura 6

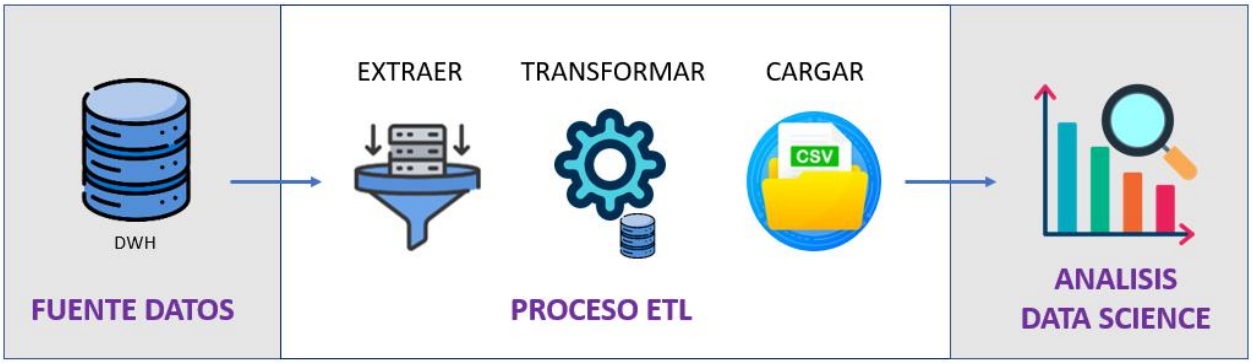
Herramienta de Trabajo



B. Flujo de datos. El ETL se estructuró en tres fases principales:

Figura 7

Flujo de datos



Extracción (Universo):

En esta fase se implementó una lógica en SQL para extraer las tablas transaccionales de los clientes correspondientes a un periodo de 7 meses previo al mes solicitado (202304-202310). Posteriormente, la información fue centralizada en una tabla única que se sirvió como universo de análisis.

Transformación (Variables):

A partir de la tabla del universo, se procesó el movimiento transaccional proveniente de los canales Remittance y Giros-Transferencias (transacciones históricas del periodo 202210 al 202310). En esta fase se aplicaron procesos de limpieza, estandarización y enriquecimiento de los datos, culminando con la construcción de las variables necesarias para el tablón final consolidado.

Carga (Output):

Finalmente, se generó el dataset de entrenamiento en formato .csv, el cual quedó disponible para las etapas posteriores de modelado y validación.

2.3.2.3. Etapa 3 – EDA. El data science toma como input la data de entrenamiento compartida por el data engineer y utiliza Python para realizar el análisis estadístico descriptivo.

A. PPNN. Se obtuvo la siguiente información con respecto a las variables:

-Número de observaciones: 321,850:

-Número de variables: 3(Numéricos:2 – Categóricos:1)

-Valores perdidos: 0%

Variables cuantitativas:

Las variables mto_conotrosproximos, ctd_conotrosproximos tiene una distribución asimétrica hacia la derecha.

Variables cualitativas:

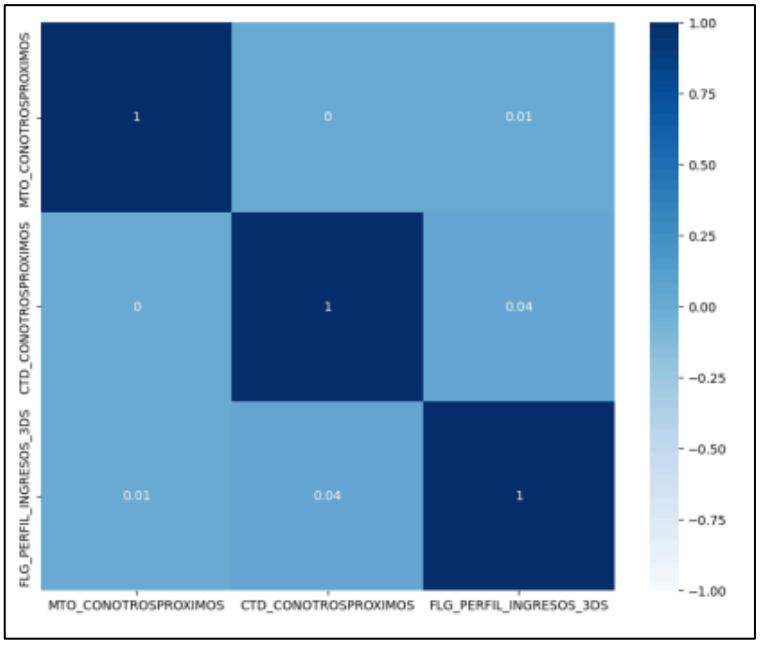
El 4.9% de los clientes salen de su perfil de ingresos basado en las fuentes del universo que se está trabajando.

Matriz de correlaciones

Se evaluó la correlación Phik de todas las variables, concluyendo que ninguna variable esta correlacionada

Figura 8

Gráfico de correlación PN

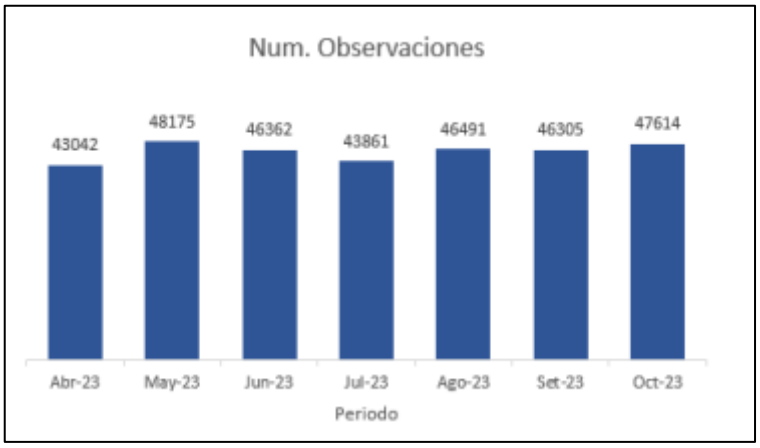


Análisis de Estabilidad

Se analizó la estabilidad desde el periodo Abril-23 hasta Octubre-23 y se obtuvo lo siguiente:

Figura 9

Gráfico de Estabilidad PN



Se observa una cantidad similar de clientes en todos los periodos, por lo cual podemos afirmar que existe estabilidad en los periodos de análisis.

Outliers

Se tiene percentiles de las siguientes variables:

Tabla 3

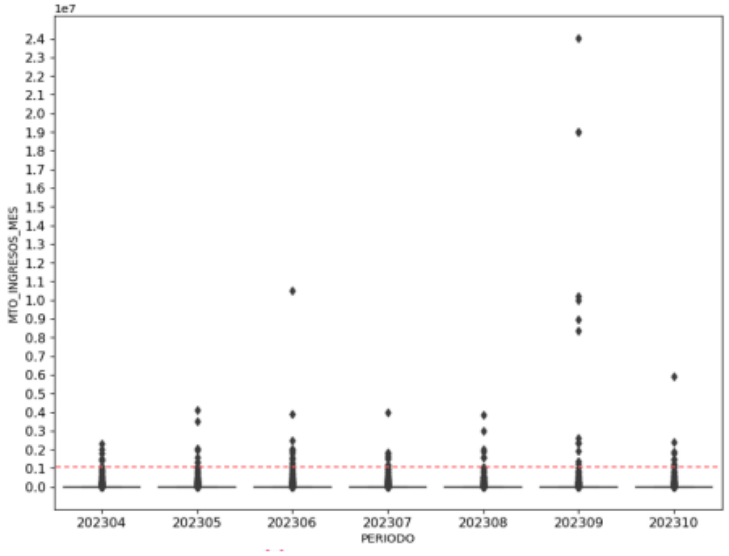
Percentiles variables montos y cantidades PN

Variables	P50	P95	P99
MTO_INGRESOS_MES	500	9050.232	45000
CTD_INGRESOS_MES	1	2	4
MTO_CONOTROSPROXIMOS	497	8311.7965	40000
CTD_CONOTROSPROXIMOS	1	2	2

En relación a esta tabla anterior, negocio optó por observar los outliers de la variable MTO_INGRESOS_MES por periodo para definir los escalones definidos en el modelo conceptual.

Figura 10

Gráfico de outliers PN



Según el gráfico se observa que los outliers a partir de US$1MM, el cual fue aprobado por negocio para establecer a los clientes que van directo a evaluación.

B. PPJJ. Se obtuvo la siguiente información con respecto a las variables:

-Número de observaciones: 41,011

-Número de variables: 3(Numéricos:2 – Categóricos:1)

-Valores perdidos: 0%

Variables Cuantitativas

Las variables mto_conotrosproximos, ctd_conotrosproximos tiene una distribución asimétrica hacia la derecha.

Variables Cualitativas

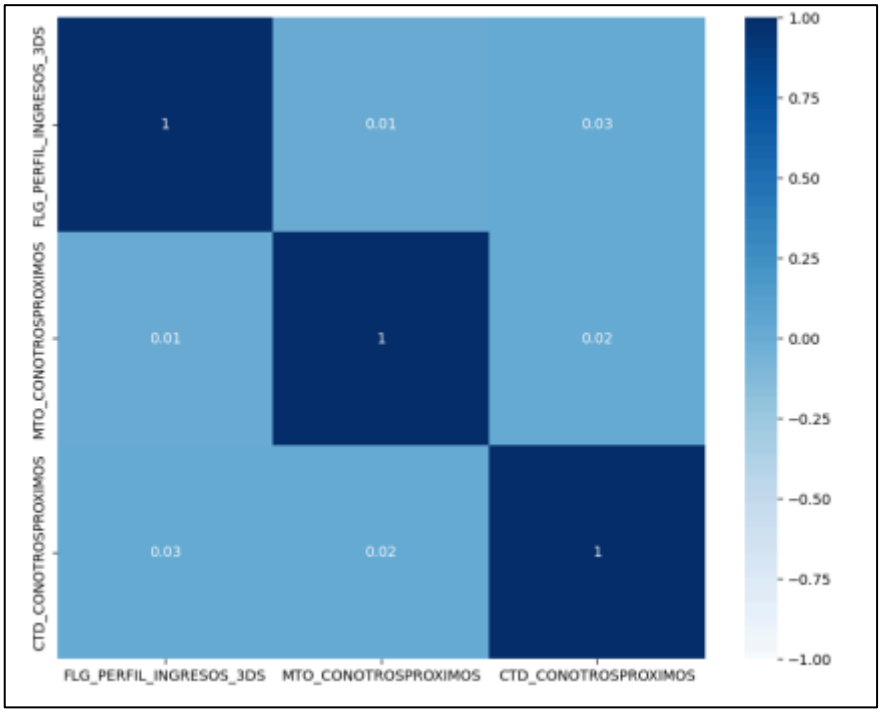
El 6.7% de los datos salen de su perfil de ingresos basado en las fuentes del universo que se está trabajando.

Matriz de correlaciones

Se evaluó la correlación Phik de todas las variables, concluyendo que ninguna variable esta correlacionada

Figura 11

Gráfico de correlación PJ

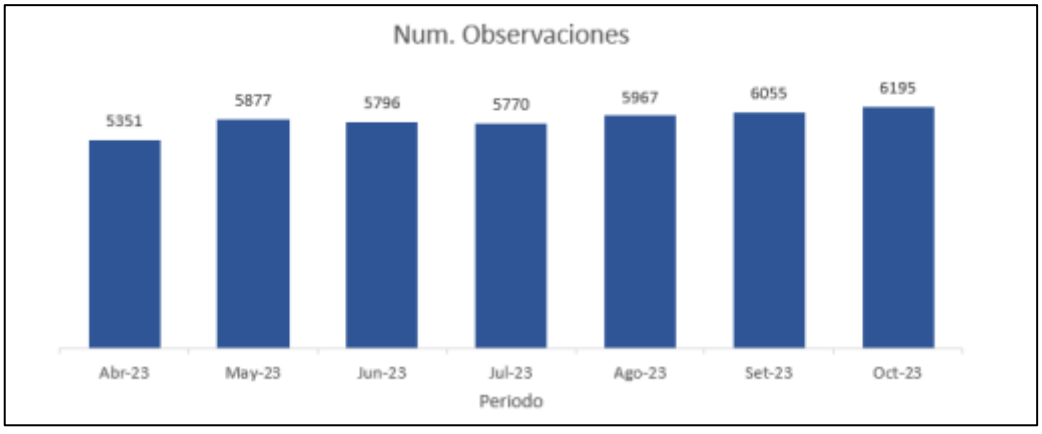


Análisis de Estabilidad

Se analizó la estabilidad poblacional desde el periodo Abril-24 hasta Octubre-24, y se obtuvo lo siguiente:

Figura 12

Gráfico de Estabilidad PJ



Del gráfico anterior se observa una cantidad similar de clientes en todos los periodos, por lo cual podemos afirmar que existe estabilidad en los periodos de análisis.

Outliers

Se tiene percentiles de las siguientes variables:

Tabla 4

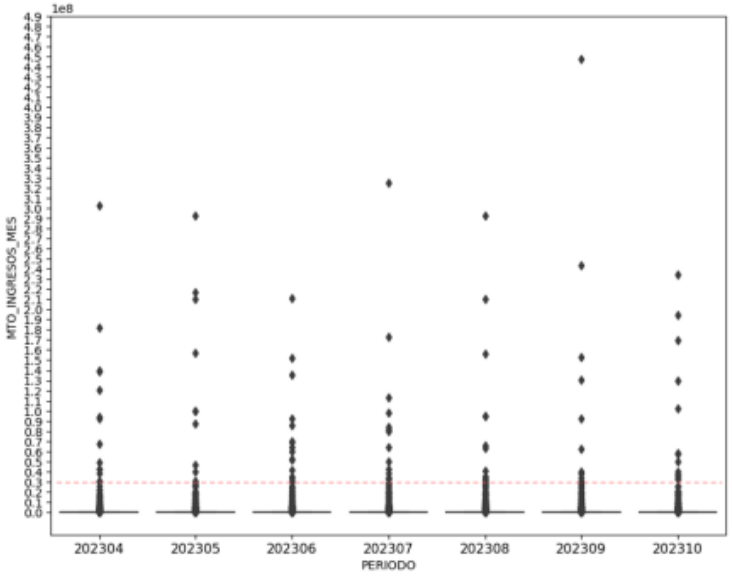
Percentiles variables montos y cantidades PJ

Variables	P50	P95	P99
MTO_INGRESOS_MES	16920	1000000	6583254.8
CTD_INGRESOS_MES	1	10	24
MTO_CONOTROSPROXIMOS	1300	506288.19	3747233.2
CTD_CONOTROSPROXIMOS	1	2	4

Para las PPJJ también se observó los outliers de la variable MTO_INGRESOS_MES por periodo para definir los escalones en el modelo conceptual:

Figura 13

Gráfico de outliers PJ



Según el gráfico se observa que los outliers a partir de US$30MM, el cual fue aprobado por negocio para establecer a los clientes que van directo a evaluación.

2.3.2.4. Etapa 4 – modelamiento. EL data science se encarga de la construcción de dos modelos por lo menos para que compitan entre sí por cada tipo de persona, y elegir el modelo final para el escenario.

A. PPNN.

Comparativa de modelos

El siguiente cuadro muestra el conjunto de técnicas con las que se experimentó para este escenario, donde se muestra el mejor score de silueta obtenido en cada modelo.

Tabla 5

Comparación modelos PN

TÉCNICA	SILUETA
K-means	94.7%
Isolation Forest	96.9%

Para el cálculo de la silueta se extrajo con una muestra del 30% por Periodo del total de casos, debido a la gran cantidad de casos.

Se optó por elegir el Isolation Forest como modelo final, debido a que tiene una mayor silueta y con ello se obtuvo a los clientes diferenciados con mayor riesgo respecto a los demás.

Variables del modelo final

- MTO_CONOTROSPROXIMOS
- CTD_CONOTROSPROXIMOS
- FLG_PERFIL_INGRESOS_3DS

B. PPJJ.

Comparativa de modelos

El siguiente cuadro muestra el conjunto de técnicas con las que se experimentó para este escenario, donde se muestra el mejor score de silueta obtenido en cada modelo:

Tabla 6

Comparación modelos PJ

TÉCNICA	SILUETA
K-means	88.4%
Isolation Forest	97.5%

Se eligió Isolation Forest como modelo final, debido a que tiene una mayor silueta y con ello se obtuvo a los clientes diferenciados con mayor riesgo respecto a los demás.

Variables del modelo final

- MTO_CONOTROSPROXIMOS
- CTD_CONOTROSPROXIMOS
- FLG_PERFIL_INGRESOS_3DS

2.3.2.5. Etapa 5 – test.

A. Muestra de resultados. El data science aplica el modelo final elegido para la data de entrenamiento, la cual se obtiene el resultado de alertas que genera el modelo por mes.

Figura 14

Resultados Modelo para PN

Periodo	Escalón 2 Segmentación	Escalón 3 Directo a Evaluación
Abr-23	13	8
May-23	14	10
Jun-23	20	13
Jul-23	21	9
Ago-23	15	9
Set-23	19	14
Promedio	**17**	**10**

Figura 15

Resultados Modelo para PJ

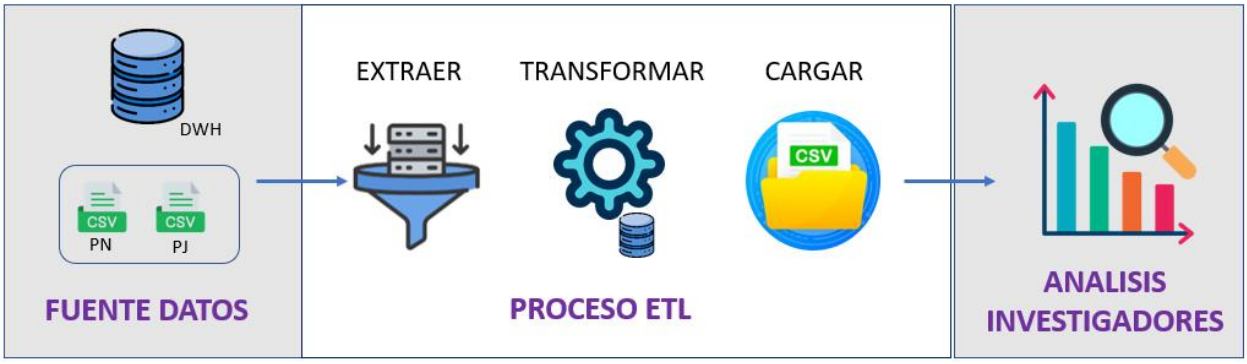
Periodo	Escalón 2 Segmentación	Escalón 3 Directo a Evaluación
Abr-23	8	13
May-23	8	10
Jun-23	10	15
Jul-23	11	14
Ago-23	11	13
Set-23	14	11
Promedio	**10**	**13**

B. Generación paquetes de información El data science remite a la data engineer los resultados de las alertas generadas para el mes de testeo Octubre-2023, correspondientes a un total de 51 alertas. De este resultado, el analista de investigaciones seleccionó 31 casos para su análisis (16 de PPJJ y 15 de PPNN). A partir de esta selección, el Data Engineer se encargó de generar los paquetes adhoc.

Para la implementación se desarrolló un ETL estructurado en tres fases:

Figura 16

Flujo datos – Paquetes Información



Extracción:

Se particionó el archivo .csv consolidado en dos subcojuntos (PN y PJ). Posteriormente, se importaron los 31 casos en dos tablas sandbox del Data Warehouse en Oracle.

Transformación (packages_adhoc):

Una vez cargada la información, se desarrollaron lógicas en SQL que permitan construir las variables requeridas para los paquetes de información.

Carga (packages):

Finalmente, se desarrolló una lógica iterativa en SQL para generar el paquete de información en archivos .csv por cliente, luego exportarlos y almacenados en una ruta definida para su posterior análisis.

Finalmente, se compartió los archivos de paquetes de información con el data science.

Validación Analista:

El data science se reúne con el analista de investigación para presentarle los paquetes de información y estiman un tiempo de validación**,** el analista deberá confirmar si las 31 alertas compartidas son efectivamente un posible ROS.

Resultado Analista

El analista se reúne con el data science, y le indica los resultados del testeo:

Resultado A Investigación (Posible ROS): 14 alertas

-Clientes sin actividad comercial o cuya actividad no está claramente definida.

-Volumen de transferencias fuera del perfil del cliente

-No se puede identificar la relación con los ordenantes de las transferencias

Resultado NAI (No Amerita Investigación): 17 alertas

-Clientes de Banca Privada con operaciones relacionadas a inversiones (dentro del perfil)

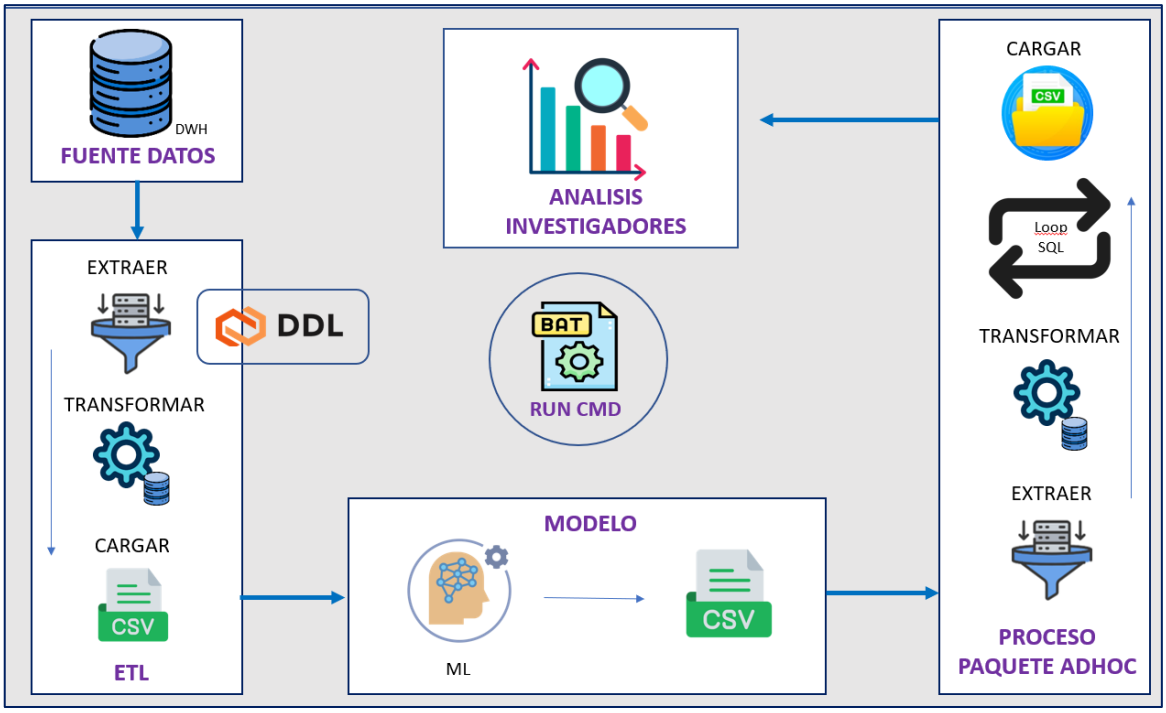
-Clientes de Banca Corporativa, con volumen significativo de exportaciones en Inforcorp y operaciones con empresas del mismo grupo en el extranjero.

Concluyendo que 14 alertas pasan a Investigación de 31 alertas evaluadas, el modelo logra un 45% de efectividad. Asimismo, al proyectar el comportamiento del modelo en un escenario mensual de 400 alertas generadas, y considerando la misma tasa de efectividad (45%), se estimaría la generación de 180 alertas efectivas y 220 alertas falsas, lo cual representa una reducción de 101 alertas falsas mensuales. Dicha disminución implica un ahorro de 83,33 horas de trabajo al mes, equivalente a 10,4 días laborales, lo que contribuye significativamente a la optimización del tiempo operativo del equipo de investigación y a una gestión más eficiente de los recursos.

2.3.2.6. Etapa 6 – producción. La puesta en producción del flujo ETL automatizado tuvo como objetivo garantizar la disponibilidad oportuna y confiable de los datos requeridos para la ejecución del modelo y la generación de los paquetes de información de manera automatizada.

Figura 17

Flujo puesto en producción



A. ***Scripts DDL.*** Se desarrollan los scripts de Data Definition Language(DDL) necesarios para la gestión de la estructura de base de datos empleados en el proceso ETL.

-Creación de Tablas: Script destinado a la generación de las tablas que almacenarán la información procesada durante la ejecución del flujo ETL

-Eliminación de Tablas: Script encargado del dropping de las tablas, permitiendo liberar espacio y asegurar la correcta regeneración de los objetos en cada nueva ejecución mensual del proceso.

B. ***Scripts ETL.*** En este proceso de preparación de datos, se desarrollaron tres scripts con parametrización de variables en función del periodo de generación de la alerta. Estos componentes permiten estandarizar la extracción, transformación y disposición de la información para el uso en el modelo.

Universo:

Script encargado de la extracción de información de los clientes que han realizado operaciones del exterior por el canal Giros y Transferencias durante los últimos siete meses del periodo parametrizado.

Variables:

Script orientado a la construcción de las variables de entrada requeridas para el modelo, consolidando los resultados en una tabla final que servirá como insumo principal.

Output:

Script diseñado para generar un spool que define la ruta de salida y genera un archivo .csv con la data raw destinada al proceso de modelado.

En conjunto estos tres scripts conforman el núcleo del flujo ETL, garantizando la calidad, consistencia y trazabilidad de la información necesaria para el desarrollo del modelo.

C. Script del modelo. Se construyó un archivo inference automatizado, el cual utiliza los inputs provenientes del modelo previamente entrenado por el data science. En este caso, se implementaron dos archivos diferenciados: uno para clientes de Persona Natural (PN) y otro para clientes de Persona Jurídica (PJ)

El archivo inference comprende las siguientes etapas:

-Load Data: Carga de la data raw proveniente del output del ETL.

-Type Validation: Validación de los tipos de datos de las variables incluidas en el data raw, garantizando consistencia y calidad de información.

-Feature Engineering: Transformación de las variables y construcción de un archivo preprocesado que servirá como insumo para el siguiente paso.

-Rules: Aplicación de filtros y reglas de negocio, así como la ejecución del modelo entrenado por el data science.

-Output: Generación y almacenamiento de los resultados de la predicción en un archivo estructurado de tipo csv.

-Run: Ejecución secuencial de cada uno de los pasos descritos, lo que asegura la correcta operación del flujo de inferencia.

Finalmente, se construyó un archivo loader encargado de cargar los resultados de la predicción desde archivo csv hacia una tabla sandbox en Oracle, con el objetivo de disponer la información para la siguiente etapa.

D. ***Script del paquete adhoc.*** Se construyeron 8 scripts en total: cuatro destinados para Persona Natural (PN) y cuatro a clientes de Persona Jurídica (PJ). Estos scripts permiten estructurar, y automatizar la generación de los paquetes de información de cada cliente según el periodo de generación de la alerta:

Paquetes Adhoc:

Script encargado de la extracción de información transaccional, incorporando variables relacionadas con las características de los clientes que serán objeto de alerta.

Deploy:

Script encargado de integrar las variables vinculadas con las características propias de la alerta, y consolidar en una tabla para su posterior consumo en los procesos de análisis.

Packages:

Se construyó la lógica para consolidar el listado de clientes a quienes se les generará el respectivo paquete de información.

Packages_trx:

Finalmente, se diseñó un script que genera el paquete de información individual por cliente en formato csv, y cual es exportado a una ruta definida y queda disponible para su análisis posterior.

E. ***Steps.*** Se construye un archivo .cmd encargado de estructurar la secuencia de ejecución de los scripts, asegurando que cada uno de ellos se ejecute en el orden correcto y sin intervención manual. Este componente permite automatizar el procesamiento de la información, optimizar los tiempos de ejecución y reducir la probabilidad de errores operativos.

F. ***Readme.*** Se elaboró un archivo Readme con el propósito de documentar los pre-requisitos y las configuraciones necesarias para la correcta ejecución del proceso. El archivo contiene el detalle de la estructura de carpetas, los archivos, y las instrucciones de configuración, garantizando así la adecuada replicabilidad y mantenimiento del flujo desarrollado.

2.3.2.7. Etapa 7: documentación. Se elaboran los documentos de cierre de la implementación

-Ficha Técnica

Documento que detalla los responsables del proyecto, fuentes de información utilizadas, y un resumen de las principales lógicas implementadas

-Manual metodológico

Contiene información conceptual y estructurada de las etapas de la implementación, con el fin de facilitar la comprensión y replicabilidad del proceso.

-Anexos

Contiene evidencias sobre el comentario de la alerta, el correo de conformidad emitido por Product Owner, así como el repositorio de archivos de ejecución y de la data de entrenamiento empleada.

2.3.3 Factibilidad técnica – operativa

2.3.3.1. Factibilidad técnica. La factibilidad Técnica se evaluó la disponibilidad y adecuación de los recursos tecnológicos existentes con los que se cuenta en el área de Cumplimiento de la entidad bancaria para la implementación del proyecto.

La institución financiera cuenta con los siguientes:

Hardware:

Para el personal de desarrollo del Proyecto 3 Laptops

Software:

-Oracle Sql Developer

-Variables de entorno SQL

-Visual Studio Code

-Python

-Excel

2.3.3.2. Factibilidad operativa. En la factibilidad Operativa considera la capacidad del equipo de trabajo para llevar a cabo el desarrollo de la implementación del proyecto según los perfiles adecuados.

Equipo de Trabajo:

-Product Owner

-Data Science

-Data Engineer

Beneficiarios

-Equipo Investigaciones

-Área Cumplimiento de la Entidad Bancaria

2.3.4 Cuadro de inversión

En esta sección se presentan los costos estimados respecto al desarrollo e implementación del proyecto. El análisis contempla dos categorías principales: costo del personal y costo tecnológico, con la finalidad de determinar el nivel de inversión requerido.

2.3.4.1. Costo personal. Se calculó considerando la dedicación del personal interno involucrado en el proyecto.

Tabla 7

Resumen Costos Personal

Cargo	**Costo mensual**	**Dedicación (hxmes)**	**Cantidad Tiempo Meses**	**Cantidad Personas**	**Importe Total**
Product Owner	S/.8,000.00	160	2	1	S/.16,000.00
Data Science	S/.5,000.00	160	2	1	S/.10,000.00
Data Engineer	S/.3,000.00	160	2	1	S/.6,000.00
				TOTAL	S/.32,000.00

2.3.4.2. Costo tecnológico. En cuanto a los recursos tecnológicos, la entidad bancaria mantiene estos costos como parte del equipamiento y la infraestructura ya disponible, sin requerir

nuevas licencias adicionales, lo cual no forma parte del centro de costo. Se presenta una estimación referencial de los costos asociados.

Tabla 8

Resumen Costos Tecnológicos.

Nombre	Costo Estimado	Importe Total
Equipo- Laptop	Instalación disponible	S/.0.00
Servicio-Licencia Oracle y Soporte Anual	Instalación disponible	S/.0.00
Visual Studio	Instalación Gratuita	S/.0.00
Python	Instalación Gratuita	S/.0.00
Microsoft Excel	Instalación de Office	S/.0.00
TOTAL		S/.0.00

2.3.4.3. Costo total del proyecto. Finalmente, el costo global del proyecto se presenta en el siguiente cuadro resumen:

Tabla 9

Resumen Costo Total Proyecto

COSTO DESCRIPCIÓN	**UNITARIO**	**SUB TOTAL**
Costo Personal	S/ 32,000.00	S/ 32,000.00
Costo Tecnológico	S/.0.00	S/.0.00
TOTAL		S/ 32,000.00

2.4 Análisis de resultados

2.4.1 Análisis costo – beneficio

El análisis del costo beneficio que se usó en este proyecto fue el ahorro que se obtuvo al utilizar recursos de personal internos del área de cumplimiento y no contratar consultores o proveedores externos especializados, generando un ahorro directo en costo de personal.

Tabla 10

Resumen Salario Personal

Cargo	Salario estimado	Salario estimado con Proveedor (+ 30%)	Importe Total
Product Owner	S/.7,000.00	S/.9,100.00	S/.18,200.00
Data Science	S/.10,000.00	S/.13,000.00	S/.26,000.00
Data Engineer	S/.8,000.00	S/.10,400.00	S/.20,800.00
		TOTAL	S/65,000.00

Tabla 11

Comparación de recurso Interno y Externo

N	Recursos	Tiempo Meses	Recursos	Costo
1	Costo Personal-Interno	2	3	S/.32,000.00
2	Costo Personal -Externo	1	3	S/65,000.00

Como se observa en la Tabla 13, la propuesta del personal externo utilizando la misma cantidad de recursos, en un tiempo estimado de 1 mes conlleva a S/65,000.00, sin embargo, utilizando la propuesta del personal interno nos costaría S/ 32,000.00 en un periodo de 2 meses. El área de cumplimiento tomó la decisión de llevar el proyecto con el equipo interno, y se logró un ahorro de S/. 33,000.00.

Utilizaremos el cálculo del ROI, para obtener el beneficio económico del proyecto.

Figura 18

Fórmula ROI

$$ROI = \frac{Utilidad\ o\ beneficio\ neto\ de\ la\ actividad}{Inversiones\ realizadas\ o\ costos}$$

ROI = (S/.65,000.00 – S/32,000.00) / S/32,000.00 = 1,03

ROI= 1,03 X 100 = 103.1%

Según lo calculado, el proyecto con equipo interno generó un retorno de 1.03 veces la inversión con un porcentaje equivalente a un 103.1% de retorno, lo que significa que por cada sol invertido se obtuvo una horro de 1.03 soles adicionales. La relación de costo-beneficio calculado con el ROI es de 1.03, como el valor es mayor que uno, podemos afirmar que el proyecto fue rentable.

III. APORTES MÁS DESTACABLES A LA EMPRESA / INSTITUCIÓN

Los aportes más destacables del proyecto a la empresa se refleja diferentes aspectos:

El principal aporte fue la contribución al fortalecimiento del sistema de prevención del lavado de activos mediante la reducción en la omisión de clientes con potenciales operaciones sospechosas. Este logro permitió disminuir el riesgo de sanciones por parte de la Superintendencia de Banca, Seguros y AFP(SBS) derivadas de la omisión o detección tardía de dichas operaciones. Al mismo tiempo, fortalecemos la reputación institucional de la entidad bancaria, y a su vez, contribuimos con el país en la lucha contra el lavado de activos.

En el ámbito operativo, la reducción de alertas irrelevantes contribuyó al ahorro de tiempo para los investigadores, permitiéndoles centrar sus esfuerzos en alertas con fundamento sólido hacia posibles casos ROS. Esta mejora se traduce en una capacidad de respuesta mucho más rápida, asegurando que los recursos de la entidad se utilicen donde realmente aporten valor.

En el plano económico, el proyecto fue desarrollado por un equipo propio interno lo que contribuyó a la reducción de costos totales sin necesidad de contratar proveedores externos. Este aporte sentó las bases para una gestión de recursos más eficiente y autónoma en futuras soluciones de cumplimiento.

Finalmente, el proyecto contribuyó a una nueva metodología de trabajo, un marco de buenas prácticas y conocimientos técnicos en la generación de alertas de lavado de activos. Este resultado garantizó un precedente para futuras implementaciones de procesos de ingeniería de datos dentro el área de cumplimiento.

IV. CONCLUSIONES

- La implementación del proceso de ingeniería de datos no solo potenció el modelo de aprendizaje automático de la alerta, sino que impactó en la disponibilidad de los datos, asegurando que la información se encuentre lista en cada fase, y se optimizó el flujo de procesamiento, facilitando incluso los reprocesos futuros. Esta mejora se traduce en un resultado tangible de la efectividad de la generación de la alerta, incrementando de un 20% a 45%, demostrando el valor de una implementación en la detección de riesgos.
- La reducción de alertas irrelevantes impactó en la eficiencia operativa de los investigadores, reduciendo su tiempo de análisis de casos de alertas, lo que representa un ahorro estimado de 83.33 horas hombres mensuales. Además, se desarrollaron paquetes de información que consolidan las alertas generadas, lo que permite a los investigadores realizar un análisis más detallado y fortalecer su capacidad analítica en la evaluación de casos.
- El costo beneficio de la implementación generó un valor de 1.03, y dado que el valor es mayor a 1, se afirma que el proyecto fue rentable. En otras palabras, por cada sol invertido la entidad bancaria obtuvo un ahorro adicional de 1.03 soles, demostrando la viabilidad económica de la propuesta.
- Finalmente, la implementación con un equipo interno de la entidad bancaria refleja la capacidad técnica y el dominio en el desarrollo de este tipo de iniciativas, lo cual permite enfrentar nuevos desafíos tecnológicos y brindar soporte de mantenimiento en la implementación de futuros modelos de detección de alertas.

V. RECOMENDACIONES

- Se recomienda que el equipo de desarrollo identifique y mapee previamente las tablas o fuentes de datos requeridas, así como verificar los permisos y el tiempo estimado para su obtención, debido a que existen políticas internas para gestionar accesos a las tablas de la base de datos. De esta manera, permitirá optimizar la planificación del proyecto, mejorar la gestión del tiempo y determinar con anticipación la viabilidad del acceso a la información necesaria.
- Se recomienda que el equipo de desarrollo implemente técnicas de anonimización sobre los datos de alta criticidad (DAC) para que la información pueda ser compartida de forma segura y confiable. De esta manera, respaldamos la información de los clientes, cumpliendo con las normas internas y los protocolos de seguridad de datos en la entidad bancaria.
- Se recomienda que el equipo de validación de modelos mantenga un seguimiento continuo sobre los responsables de las fuentes que generan la información almacenada en las tablas de datos, con el fin de identificar oportunamente actualizaciones en la información, accesos o responsables. Asimismo, anticipamos desviaciones en los resultados del modelo, y aseguramos la efectividad de la alerta de lavado de activos.
- Se recomienda que el equipo de desarrollo elabore una documentación técnica detallada que integre los requisitos y el desglose de cada fase del proyecto. Esta documentación, no solo sirve como una guía metodológica para la implementación de alertas de lavado de activos, sino que optimiza tiempos de respuesta ante ajustes preventivos, asegurando la continuidad operativa y facilitando futuras auditorías.

VI. REFERENCIAS

Abdala, G. (2022). *Procesos de machine learning para la prevención de fraude en Mercado Libre* [Tesis de maestría, Universidad Torcuato Di Tella]. Repositorio Institucional UTDT. https://repositorio.utdt.edu/server/api/core/bitstreams/772963ea-c454-458c-b977-25b9b7541647/content

Adepoju, O., Wosowei, J., Lawte, S., & Hemaint, J. (18 y 20 de octubre de 2019). Comparative evaluation of credit card fraud detection using machine learning techniques. [Conferencia]. *2019 Global Conference for Advancement in Technology (GCAT)*, Bangalore, India. https://es.scribd.com/document/570994590/COMPARATIVE-EVALUATION-OF-CREDIT-CARD-FRAUD-DETECTION

Alvarez, F. (2020) *Machine Learning en la detección de fraudes de comercio electrónico aplicado a los servicios bancarios*. [Tesis de Maestría, Universidad de Palermo]. Repositorio Institucional UP. https://dspace.palermo.edu/dspace/handle/10226/2240?show=full

Alvarez, J., Enriquez, C., & Tineo, R. (2024). *Niveles de efectividad del modelo de redes neuronales en el análisis del comportamiento y predicción del PBI del Perú (1980–2021)* [Tesis de licenciatura, Universidad Nacional Hermilio Valdizán]. Repositorio Institucional UNHEVAL. https://repositorio.unheval.edu.pe/backend/api/core/bitstreams/e2a1e7ac-9c59-461b-8de4-0311fa7829c6/content

Aronés, E. (2021). *Predicción del rendimiento académico basado en machine learning* [Tesis de licenciatura, Universidad Nacional de San Cristóbal de Huamanga]. Repositorio Institucional UNSCH.

https://repositorio.unsch.edu.pe/server/api/core/bitstreams/21894746-aefb-4699-a75f-2f25277106dc/content

Aylin, T. (28 de febrero de 2025). Drift, anomaly, and novelty in machine learning. *Baeldung*. https://www.baeldung.com/cs/ml-drift-anomaly-novelty

Barboza, J. L. (2023). *Efectividad de un modelo de clasificación basado en deep learning en la detección de COVID-19* [Tesis de licenciatura, Universidad Nacional José María Arguedas]. Repositorio Institucional UNAJMA. https://repositorio.unajma.edu.pe/bitstream/handle/20.500.14168/822/Jos%C3%A9_Luis_Tesis_Bachiller_2023.pdf?sequence=1&isAllowed=y

Basel Committee on Banking Supervision (BCBS). (2020**).** *Sound management of risks related to money laundering and financing of terrorism.* Bank for International Settlements. https://www.bis.org/bcbs/publ/d353.pdf

Bodor, A., Hnida, M., & Daoudi, N. (2023). Machine learning models monitoring in MLOps context: Metrics and tools. *Revista Internacional de Tecnologías Móviles Interactivas, 17*(23), 125-139. https://doi.org/10.3991/ijim.v17i23.43479

Breunig, M., Kriegel, H., Ng, R., & Sander, J. (2000). LOF: Identifying density-based local outliers. *ACM SIGMOD Record*, 29(2), 93–104. https://doi.org/10.1145/335191.335388

Calvo, D. (20 de junio de 2018). Apache Spark: *Diego Calvo*. https://www.diegocalvo.es/spark/

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). *CRISP-DM 1.0 Step-by-step data mining guide*. https://www.semanticscholar.org/paper/CRISP-DM-1.0%3A-Step-by-step-data-mining-guide-Chapman/54bad20bbc7938991bf34f86dde0babfbd2d5a72

Cortés, J. (2023). *Detección de anomalías transaccionales aplicando técnicas de machine learning con grafos* [Tesis de maestría, Universidad del Rosario]. Repositorio Institucional UR. https://repository.urosario.edu.co/server/api/core/bitstreams/148e3ef8-0f0c-4563-9e9e-71480ef7b41e/content

Databricks. (2025*). What is a dataset?* https://www.databricks.com/glossary/what-is-dataset

Dávila-Morán, R., Castillo-Sáenz, R., Vargas-Murillo, A., Velarde-Dávila, L.,García-Huamantumba, E.,García-Huamantumba,C., Pasquel-Cajas, R., & Guanillo-Paredes, C. (2023). Aplicación de modelos de aprendizaje automático en la detección de fraudes en transacciones financieras. *Data and Metadata*, 2, 109. https://doi.org/10.56294/dm2023109

Domino Data Lab. (2025). *What is a Feature in Machine Learning and Data Science?*. https://domino.ai/data-science-dictionary/feature

Encalada, D. (2025). *Diseño de un marco de trabajo para la implementación de procesos ETL* [Tesis de licenciatura, Universidad Politécnica Salesiana]. Repositorio Institucional UPS. https://dspace.ups.edu.ec/bitstream/123456789/29947/1/UPS-GT006112.pdf

Escuela de Administración de Empresas [EAE] Barcelona. (09 de abril de 2024). *Python: ¿Qué es? Y como se aplica en Big Data.* https://www.eaebarcelona.com/es/blog/python-que-es

Financial Crimes Enforcement Network [FinCEN]. (2005). *Suspicious Activity Reporting (Structuring).* U.S. Department of the Treasury. https://www.fincen.gov/resources/statutes-regulations/administrative-rulings/suspicious-activity-reporting-structuring

Fontalvo, J., Guarín, E., & Díaz, M. (2024). *La analítica de datos y el comportamiento del fraude bancario.* [Trabajo de grado, Universidad del Norte] Repositorio Institucional UN.

https://manglar.uninorte.edu.co/bitstream/handle/10584/11934/La%20Anal%C3%ADtica%20de%20Datos%20y%20el%20Comportamiento%20del%20Fraude%20Bancario.pdf?sequence=1&isAllowed=y

Fuentes, A. y Navarro, C. (2024). *Implementación de un modelo basado en algoritmos de machine learning para el análisis predictivo de diagnósticos clínicos en REDLAB Perú S.A.C*. [Tesis de licenciatura, Universidad Autónoma del Perú]. Repositorio Institucional UA. https://repositorio.autonoma.edu.pe/item/f6f074ef-2eff-4c6f-b165-4833fa74d297

Go Beyond Biz. (2025). *Money laundering statistics 2025 (US & Worldwide)*. https://www.go-beyond.biz/data-statistics/money-laundering-statistics

Hassenstein, M.J & Vanella,P (2022). Data Quality – Concepts and Problems. *Encyclopedia, 2(1), 498-510. https://www.mdpi.com/2673-8392/2/1/32*

Inga, F., Miranda, K., Quispe, D., Reyna, J., & Turriate, S. (2023). *Implementación de técnicas de machine learning para la segmentación de clientes en una empresa del sector farmacéutico* [Trabajo de suficiencia profesional, Universidad ESAN]. Repositorio Institucional ESAN. https://repositorio.esan.edu.pe/server/api/core/bitstreams/b6a9e4f3-5698-4e2f-ad51-05e90cd93477/content

International Business Machines Corporation [IBM] (s.f.). *¿Qué es SQL?* https://www.ibm.com/mx-es/think/topics/structured-query-language

International Business Machines Corporation [IBM]. (s.f.). *Data pipeline examples: ETL, data science, e-commerce + more.* https://www.ibm.com/mx-es/think/topics/data-pipeline

Jacopo, C. (2022). El nexo entre la corrupción y el lavado de dinero: deconstruyendo la red Toledo-Odebrecht en Perú. *Tendencias del Crimen Organizado,* 27, 342-363. https://doi.org/10.1007/s12117-021-09439-6

Liu, F., Ting, K., & Zhou, Z. (2008). Isolation forest. In 2008 Eighth IEEE International *Conference on Data Mining.IEEE*. https://ieeexplore.ieee.org/document/4781136

Llanos, L. (2021). *Propuesta de implementación de un sistema de alertas tempranas para minimizar fraudes en cuentas pasivas en una entidad financiera* [Tesis de licenciatura, Universidad San Ignacio de Loyola]. Repositorio Institucional USIL. https://repositorio.usil.edu.pe/server/api/core/bitstreams/f4c5c6a2-fa23-445e-8462-d26b31e2e0b4/content

Luna, C. (2024). *Detección de fraude transaccional mediante modelos de aprendizaje automático: Una aplicación a una entidad financiera chilena* [Tesis de licenciatura, Universidad de Concepción]. Repositorio Institucional UDEC. https://repositorio.udec.cl/server/api/core/bitstreams/92e3968a-f91a-4a83-8ce9-f6be134aee7a/content

Mejía, I. y Huamán, M. (2024). *Solución Big Data para la integración de datos de contacto de clientes para una empresa financiera* [Tesis de maestría, Universidad Peruana de Ciencias Aplicadas]. Repositorio Institucional UPC. https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/673775/Mejia_CI.pdf?sequence=1&isAllowed=y

Mendoza-Álava, J., Macías-Bermeo, L., Morales-Carrillo, J., & Cedeño-Valarezo, L. (2024). Modelos de aprendizaje automático: Aplicación y eficiencia. Revista Cientifica De Informática *ENCRIPTAR* ,7(14), 87-114. https://doi.org/10.56124/encriptar.v7i14.005

Murphy, K. P. (2012). *Machine learning: A probabilistic perspective. MIT Press.* https://mitpress.mit.edu/9780262018029/machine-learning/

Olson, K. (2021). What Are Data?. *Qualitative Health Research*, 31(9), 1567-1569. https://journals.sagepub.com/doi/10.1177/10497323211015960

Oracle. (s.f.). *Oracle Database*. https://www.oracle.com/pe/database/

Oracle. (s.f.). *PL/SQL Oracle.* https://www.oracle.com/pe/database/technologies/appdev/plsql.html

Oza, A. (2018). *Fraud detection using machine learning* [Proyecto final, CS229, Stanford University]. https://www.semanticscholar.org/paper/Fraud-Detection-using-Machine-Learning-Oza-aditya/9f2c08d9efaa53cfabdd0ec47afa8015c7ff5bb9#citing-papers

Peña, G. (2023). *Estudio de técnicas de aprendizaje no supervisado y detección de anomalías: Aplicación en perfiles de potencia consumida en trenes urbanos* [Tesis de maestría, Escuela Técnica Superior de Ingeniería, Universidad de Sevilla]. Repositorio Institucional US. https://biblus.us.es/bibing/proyectos/abreproy/72537/fichero/TFM-2537+Pe%C3%B1a+Arg%C3%BCeso.pdf

Proyectos Ágiles. (s.f.). *¿Qué es Scrum?* https://proyectosagiles.org/que-es-scrum/

Puerto, G. (12 de septiembre de 2019). *Eficacia, eficiencia y efectividad: qué las diferencia*. https://galiapuerto.es/eficacia-eficiencia-y-efectividad-que-las-diferencia/

Pure Storage. (2023). *¿Qué es la estandarización de datos?* https://www.purestorage.com/es/knowledge/what-is-data-standardization.html

Rayo, C. (2020). *Prototipo de detección de fraudes con tarjetas de crédito basado en inteligencia artificial aplicado a un banco peruano* [Tesis de licenciatura, Universidad de Lima]. Repositorio Institcional UL. Uhttps://repositorio.ulima.edu.pe/bitstream/handle/20.500.12724/15294/Trabajo.pdf?sequence=1&isAllowed=y

Rodríguez, J. (2022). *Proyecto de implementación de un modelo machine learning para la evaluación de riesgo de operaciones sospechosas a los clientes de una entidad bancaria* [Trabajo de suficiencia profesional, Universidad Nacional Mayor de San Marcos]. Repositorio Institucional UNMS. https://cybertesis.unmsm.edu.pe/backend/api/core/bitstreams/700743f0-3f54-48d9-8279-6eeca503c1f1/content

Ruffner, J. (2014). Control, Prevención y Represión ante el lavado de activos en el Perú. *Quipukamayoc, 18(35), 209-220*. https://doi.org/10.15381/quipu.v18i35.3812

Santamaría, M. (2023). *Machine learning operations (MLOps): Contexto actual y tendencias futuras* [Tesis de maestría, Universidad Internacional de La Rioja]. Repositorio Institucional UIR. https://investigacion.unirioja.es/documentos/655c98b6da93c5320dbe7690/f/655c98b6da93c5320dbe768f.pdf

Shadrooh, S., Norvag, K. (2024). Smotef money laundering detection using temporal order and flow analysis. *Applied Intellingence*, 54(15), 7461–7478. https://doi.org/10.1007/s10489-024-05545-4

Soria, J., Segura, L. y Loayza, R. (17 al 19 de julio de 2024). Machine Learning Models for Money Laundering Detection in Financial Institutions: A Systematic Literature Review [conferencia]. *22nd LACCEI International Multi-Conference for Engineering, Education, and Technology*, San José, Costa Rica. https://laccei.org/LACCEI2024-CostaRica/papers/Contribution_1682_final_a.pdf

Superintendencia de Banca, Seguros y Administradoras Privadas de Fondos de Pensiones[SBS]. (2024). *Guía de calidad de reportes de operaciones sospechosas (ROS).*

https://www.howdengroup.com/sites/peru.howdenprod.com/files/2024-10/manual-gremial-para-la-prevencion-del-laft.pdf

Superintendencia de Banca, Seguros y Administradoras Privadas de Fondos de Pensiones[SBS]. (2025). *Nociones básicas del Sistema Nacional contra LAFT: Lavado de Activos.* *https://www.sbs.gob.pe/prevencion-de-lavado-activos/Nociones-basicas-del-Sistema-Nacional-contra-LAFT?utm_source=chatgpt.com*

Suxo, G. (2024). *Implementación de una solución de machine learning para la asignación de créditos financieros en una entidad financiera* [Trabajo de suficiencia profesional, Universidad Nacional Mayor de San Marcos]. Repositorio Institucional UNMS. https://cybertesis.unmsm.edu.pe/backend/api/core/bitstreams/4086399a-c6b1-4681-9c15-810a43f183eb/content

Tarazona, P., Mateous, S., Becerra, L., & Pérez, J. (2024). *Métodos de machine learning para detección de fraude en transacciones financieras en tiempo real* [Trabajo académico de posgrado, Universidad EAN]. Repositorio Institucional EAN. https://repository.universidadean.edu.co/server/api/core/bitstreams/f055e77b-109e-4047-a1ba-9e7cec92c6bc/content

TiDB Team. (2024). *What is script in SQL?* https://www.pingcap.com/article/what-is-script-in-sql-simple-explanation/

Triviño, M. (2021). *Modelo de machine learning para la evaluación de efectividad de una estrategia de marketing digital* [Tesis de maestría, Universidad Católica de Colombia]. Repositorio Institucional UCC. https://repository.ucatolica.edu.co/server/api/core/bitstreams/cc8f614c-67fe-4f90-a448-edb5f4fda723/content

Valencia, J. (2024). *Diseño de Metodología Integral de Ingeniería de Datos Empresariales para Optimización de Procesos y Toma de Decisiones Estratégicas* [Tesis de grado, Universidad Evaluación de Impacto Ambiental]. Repositorio Institucional EIA. https://repository.eia.edu.co/server/api/core/bitstreams/a0e57657-8b17-48e2-84a0-d14c1bc8d1e6/content

Vargas, A. (2023). *Consideraciones de implementación de sistemas y servicios de TI en modalidad on-premise o en nube de computación en el Ecuador* [Tesis de licenciatura, Escuela Politécnica Nacional]. Repositorio Institucional EPN. https://bibdigital.epn.edu.ec/bitstream/15000/24415/1/CD%2013339.pdf

Wilson, J., Rodrigues, J. (2020). Classification accuracy and efficiency of writing screening using automated essay scoring. *Journal of School Psychology*, 82, 123–140. https://doi.org/10.1016/j.jsp.2020.08.008