



Universidad Nacional Mayor de San Marcos

Universidad del Perú. Decana de América

Facultad de Ingeniería de Sistemas e Informática

Escuela Profesional de Ingeniería de Sistemas

**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

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Josef Renato RODRÍGUEZ MALLMA

ASESOR

John Ledgard TRUJILLO TREJO

Lima, Perú

2022



Reconocimiento - No Comercial - Compartir Igual - Sin restricciones adicionales

<https://creativecommons.org/licenses/by-nc-sa/4.0/>

Usted puede distribuir, remezclar, retocar, y crear a partir del documento original de modo no comercial, siempre y cuando se dé crédito al autor del documento y se licencien las nuevas creaciones bajo las mismas condiciones. No se permite aplicar términos legales o medidas tecnológicas que restrinjan legalmente a otros a hacer cualquier cosa que permita esta licencia.

Referencia bibliográfica

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 de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios

Datos de autor	
Nombres y apellidos	Josef Renato Rodriguez Mallma
Tipo de documento de identidad	DNI
Número de documento de identidad	47111929
URL de ORCID	https://orcid.org/0000-0002-8819-1985
Datos de asesor	
Nombres y apellidos	John Ledgard Trujillo Trejo
Tipo de documento de identidad	DNI
Número de documento de identidad	06187585
URL de ORCID	https://orcid.org/0000-0002-0563-4809
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Javier Cabrera Diaz
Tipo de documento	DNI
Número de documento de identidad	08692591
Miembro del jurado 1	
Nombres y apellidos	Rosa Menéndez Mueras
Tipo de documento	DNI
Número de documento de identidad	10246770
Datos de investigación	
Línea de investigación	No Aplica
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Lima Distrito: Cercado de Lima

	<p>Jr. Carlos Amezaga No. 375</p> <p>Universidad Nacional Mayor de San Marcos</p> <p>Latitud: -12.0564232</p> <p>Longitud: -77.0843327</p>
Año o rango de años en que se realizó la investigación	2021
URL de disciplinas OCDE	<p>2.02.04 -- Ingeniería de sistemas y comunicaciones</p> <p>https://purl.org/pe-repo/ocde/ford#2.02.04</p>



UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA
Escuela Profesional de Ingeniería de Sistemas

**Acta Virtual de Sustentación
del Trabajo de Suficiencia Profesional**

Siendo las 20:00 horas del día 04 de enero del año 2022, se reunieron virtualmente los docentes designados como Miembros del Jurado del Trabajo de Suficiencia Profesional, presidido por el Mg. Cabrera Díaz Javier (Presidente), Mg. Menéndez Mueras Rosa (Miembro) y el Lic. Trujillo Trejo John Ledgard (Miembro Asesor), usando la plataforma Meet (<https://meet.google.com/gfv-qdyi-szt>), para la sustentación virtual del Trabajo de Suficiencia Profesional intitulado: **“PROYECTO DE IMPLEMENTACIÓN DE UN MODELO MACHINE LEARNING PARA LA EVALUACIÓN DE RIESGO DE OPERACIONES SOSPECHOSAS A LOS CLIENTES DE UNA ENTIDAD BANCARIA”**, por el Bachiller **Rodríguez Mallma Josef Renato**; para obtener el Título Profesional de Ingeniero de Sistemas.

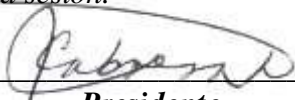
Acto seguido de la exposición del Trabajo de Suficiencia Profesional, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los miembros del Jurado.

El Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el Bachiller obtuvo la nota de **17 DIECISIETE**.

A continuación el Presidente de Jurado el Mg. Cabrera Díaz Javier, declara al Bachiller **Ingeniero de Sistemas**.

Siendo las 21 horas, se levantó la sesión.


Presidente
Mg. Cabrera Díaz Javier


Miembro
Mg. Menéndez Mueras Rosa


Miembro Asesor
Lic. Trujillo Trejo John Ledgard

Dedicatoria

Al Excelentísimo Dr. Luis Morales Galarreta docente de la facultad de medicina de la UNMSM a quien gracias a su ayuda pude culminar el grado de bachiller.

AGRADECIMIENTOS

A mi familia.

Quienes estuvieron pendientes de mi cuando más lo
requería.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERÍA DE SISTEMAS E INFORMÁTICA ESCUELA
PROFESIONAL DE INGENIERÍA DE SISTEMAS

Proyecto de Implementación de un Modelo Machine Learning para la
Evaluación de Riesgo de Operaciones Sospechosas a
los Clientes de una Entidad Bancaria.

Autor: Rodriguez Mallma Josef R.

Asesor: Trujillo Trejo John Ledgard

Título: Trabajo de Suficiencia Profesional para optar el Título Profesional
de Ingeniero de Sistemas

Fecha: noviembre 2021

RESUMEN

El presente trabajo de experiencia profesional describe la implementación de un modelo Machine Learning realizada para determinar del nivel de riesgo de operaciones sospechosas a los clientes pertenecientes a una entidad bancaria, para lo cual se siguió la metodología, procesos y normas establecidas por el Banco, junto a las buenas prácticas y diseño de modelos.

El esfuerzo realizado tiene como fin lograr tener un modelo que sea útil para la institución permitiéndole identificar clientes de alto riesgo,

En el informe se detallan los pasos que se siguieron, así como las conclusiones y recomendaciones a los que se llegó en base a la experiencia profesional.

Palabras claves: Modelo, Lavado de Activos, Financiamiento al Terrorismo, Machine Learning, Banco.

MAJOR NATIONAL UNIVERSITY OF SAN MARCOS

FACULTY OF SYSTEMS ENGINEERING

PROFESSIONAL SCHOOL OF SYSTEMS

ENGINEERING

Project Implementation of a Machine Learning Model for Risk Assessment
Of Suspicious Operations to The Clients of a Banking Entity.

Author: Rodriguez Mallma Josef Renato

Adviser: Trujillo Trejo John Ledgard

Title: Professional Sufficiency Work for opt for the Professional Title of
Systems Engineer

Date: November 2021

ABSTRACT

This work of professional experience describes the implementation of a Machine Learning model carried out to determine the level of risk of money laundering or terrorist financing to clients belonging to a banking entity

For which used the established methodology, processes, and standards, along with good practices in Machine Learning model design.

The effort made is to achieve a model that is useful for the institution, allowing it to identify high-risk clients,

The report details the steps that were followed, as well as the conclusions and recommendations that were reached based on professional experience.

Keywords: Model, Money Laundering, Terrorism Financing, Machine Learning, Bank

Contenido

ÍNDICE DE TABLAS	x
ÍNDICE DE FIGURAS	xi
CAPÍTULO I TRAYECTORIA PROFESIONAL.....	3
1.1. PRESENTACION PROFESIONAL	3
1.2. EXPERIENCIA PROFESIONAL.....	3
1.3. FORMACION ACADEMICA.....	5
1.4. CERTIFICACIONES	5
1.5. IDIOMAS	6
CAPÍTULO II CONTEXTO EN EL QUE SE DESARROLLA LA EXPERIENCIA	7
2.1 EMPRESA - ACTIVIDAD QUE REALIZA	7
2.2 MISION	8
PROPOSITO.....	8
2.3 VISION	8
ASPIRACION.....	8
2.4 ORGANIZACION DE LA EMPRESA	9
2.4.1 ORGANIZACION DE LA DIVISION DE CUMPLIMIENTO	10
2.5 AREA, CARGO Y FUNCIONES DESEMPEÑADAS	11
FUNCIONES DESEMPEÑADAS:.....	11
FUNCIONES DESEMPEÑADAS:.....	11
FUNCIONES DESEMPEÑADAS:.....	12
CAPÍTULO III CONTEXTO EN EL QUE SE DESARROLLA LA EXPERIENCIA	13
3.1 SITUACION PROBLEMATICA.....	13
3.1.1 DEFINICION DEL PROBLEMA.....	15
3.2 SOLUCION.....	15
3.2.1 OBJETIVOS	15
OBJETIVO GENERAL	15
OBJETIVOS ESPECIFICOS	15
3.2.2 ALCANCE	15
3.2.3 ETAPAS Y METODOLOGIA.....	16
3.2.4 FUNDAMENTOS UTILIZADOS	26

3.2.4.1	SQL.....	26
3.2.4.2	SPARK.....	26
3.2.4.3	PYTHON.....	27
3.2.4.4	PLSQL DEVELOPER	28
3.2.4.5	CLOUDERA.....	28
3.2.4.6	LAPTOP.....	28
3.2.5	IMPLEMENTACION DE LAS AREAS, PROCESOS, SISTEMAS Y SUS BUENAS PRACTICAS	29
3.3	EVALUACION	32
3.3.1	EVALUACION ECONOMICA.....	32
3.3.2	BENEFICIOS OBTENIDOS	33
CAPÍTULO IV REFLEXIÓN CRÍTICA DE LA EXPERIENCIA		34
CAPÍTULO V CONCLUSIONES Y RECOMENDACIONES		36
4.1	CONCLUSIONES.....	36
4.2	RECOMENDACIONES	37
4.3	FUENTES DE INFORMACIÓN	39
4.4	GLOSARIO.....	39
ANEXOS		40

ÍNDICE DE TABLAS

Tabla 1: Experiencia Profesional del Autor.....	3
Tabla 2: Formación Académica del Autor.....	5
Tabla 3: Certificaciones del Autor.....	5
Tabla 4: Idiomas del Autor.....	6
Tabla 5: Costo estimado del proyecto	32
Tabla 6: Estimado beneficio del País al evitar ROS	33

INDICE DE FIGURAS

Figura 1: Mapa Credicorp.....	8
Figura 2: Organigrama BCP	9
Figura 3: Organigrama de la Divisi�n de Cumplimiento	10
Figura 4: Evoluci�n n�mero de ROS	13
Figura 5: Dinero congelado por tipo de Delito ROS	14
Figura 6: Fases de la creaci�n de un modelo ML.....	16
Figura 7: Aprendizaje Supervisado.....	17
Figura 8: Aprendizaje no Supervisado.....	18
Figura 9: Ejemplo Dataset.....	20
Figura 10: Separaci�n del Dataset.....	21
Figura 11: Comparaci�n de Modelos.....	23

INTRODUCCIÓN

El presente informe de experiencia profesional describe el proceso de implementación de un modelo Machine Learning realizado para la detección de personas que puedan estar realizando operaciones sospechosas en un Banco en Perú. Con ello se buscaba lograr un modelo de buen performance, permitiendo que pueda integrarse a los objetivos del negocio, diferentes análisis y procesos en la unidad para así poder realizar las evaluaciones de forma más rápida, así como también tener un modelo con base matemática y estadística que cumpla los estándares de la SBS.

El proyecto es representativo en la trayectoria profesional del autor del presente informe, debido a que le permitió abarcar la mayor parte de las etapas de un proyecto y aplicar los conocimientos, herramientas y técnicas de implementación de un Modelo Machine Learning.

El presente informe está organizado de la siguiente manera:

En el CAPÍTULO I se detalla cronológicamente la trayectoria profesional del autor, los cargos, funciones, actividades, así como también la especialización realizada en diferentes ámbitos.

En el CAPÍTULO II se describe la historia de la empresa donde se realizó el desarrollo del proyecto en mención, su estructura orgánica, la visión, la misión, y los servicios que brinda. También se resaltan las funciones realizadas por el autor dentro de la empresa.

En el CAPÍTULO III se detalla el trabajo realizado, el cual se refiere al proyecto de implementación de un Modelo Machine Learning para la evaluación de clientes que puedan cometer actos de operaciones sospechosas, se describe la metodología, los procesos y normas, asimismo también se especifican los fundamentos utilizados y se resaltan los puntos más importantes que se presentaron durante la realización de este trabajo.

En el CAPITULO IV se menciona al aporte del autor del informe, la experiencia obtenida, los conocimientos que demandó así como el desarrollo profesional que obtuvo con la realización del presente trabajo. También se expone de forma crítica la introspección de la práctica laboral.

En el CAPITULO V se menciona los resultados conclusivos y los consejos y sugerencias.

CAPÍTULO I TRAYECTORIA PROFESIONAL

1.1. PRESENTACION PROFESIONAL

Profesional de la carrera de Ingeniería de Sistemas e Informática de la Universidad Nacional Mayor de San Marcos. Actualmente me desempeño con el rol de Subgerente Adjunto de Analytics de Cumplimiento en el Banco de Crédito del Perú, creando y supervisando modelos de Machine Learning.

He desempeñado distintos cargos enfocados en implementación de modelos Machine Learning con distintos algoritmos. Tengo conocimientos en desarrollo de Software, Ciencia de Datos, Inteligencia Artificial y Big Data.

1.2. EXPERIENCIA PROFESIONAL

Tabla 1: Experiencia Profesional del Autor

Junio 2021 - Actual	Empresa: Banco de Crédito del Perú - BCP Cargo: Sub Gerente Adjunto de Analytics de Cumplimiento Desempeño: Creación, mantenimiento, seguimiento de los modelos Machine Learning, así como innovación en Ciencia de Datos.
Agosto 2019 - Junio 2021	Empresa: Banco de Crédito del Perú - BCP Cargo: Data Scientist Desempeño: Creación y mantenimiento de modelos Machine Learning para detectar la probabilidad de pago del cliente.

Septiembre 2017 - julio 2019	<p>Empresa: Banco de Crédito del Perú - BCP</p> <p>Cargo: Analista de Gestión de la Información</p> <p>Desempeño:</p> <p>Gestionar la base de datos de todos los reclamos de la entidad, así como brindar ideas para evitar el aumento de estos, creación de modelos Machine Learning para predecir el volumen de reclamos mensuales.</p>
Mayo 2017 - septiembre 2017	<p>Empresa: Indra</p> <p>Cargo: Junior System Engineer</p> <p>Desempeño:</p> <p>Analista de la Base de datos del cliente telefónica manteniendo la disponibilidad del servicio para los clientes.</p>
Junio 2016 - diciembre 2016	<p>Empresa: Osiptel</p> <p>Cargo: Técnico en Cableado y Conectividad</p> <p>Desempeño:</p> <ul style="list-style-type: none"> - Creación y mantenimiento del sistema de transferencia de información de la red de datos.

Fuente: Elaboración Propia

1.3. FORMACION ACADEMICA

Tabla 2: Formación Académica del Autor

2010 - 2016	Grado Académico de Bachiller en Ingeniería de Sistemas Escuela Académico Profesional de Ingeniería de Sistemas - Facultad de Ingeniería de Sistemas e Informática - Universidad Nacional Mayor de San Marcos.
-------------	--

Fuente: Elaboración Propia

1.4. CERTIFICACIONES

Tabla 3: Certificaciones del Autor

2020	Data Scientist in Python. Microsoft
2019	Introduction Python for Data Scientist.
2019	Maths for Machine Learning AWS

Fuente: Elaboración Propia

1.5. IDIOMAS

Tabla 4: Idiomas del Autor

Nativo	Español
Intermedio/Avanzado	Inglés

Fuente: Elaboración Propia

CAPÍTULO II CONTEXTO EN EL QUE SE DESARROLLA LA EXPERIENCIA

2.1 EMPRESA - ACTIVIDAD QUE REALIZA

El banco de Crédito del Perú es una institución del sistema financiero peruano y el proveedor más grande y líder de servicios financieros en el país, representando el principal activo del grupo financiero Credicorp.

A través de sus divisiones de Banca Corporativa y Banca Empresa provee servicios a clientes corporativos y empresas medianas, mientras que desde Banca Minorista atiende a pequeñas empresas y clientes individuales con una amplia gama de productos con alto valor agregado. Las actividades de su sector se rigen por la Ley 267025, norma que tiene como objetivo promover al funcionamiento de un sistema financiero y un sistema de seguros competitivos, sólidos y confiables, que contribuyan al desarrollo nacional.

El Banco de Crédito del Perú (BCP) se constituye como sociedad anónima, con el nombre de Banco Italiano el 3 de abril de 1889. 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. Inicia sus operaciones el día 9 de abril de 1889 y el 21 de enero de 1942 cambia de razón social a Banco de Crédito del Perú.

Credicorp

Holding de servicios financieros líder en el Perú con presencia en Bolivia, Chile, Colombia y Panamá. Cuenta con un portafolio diverso de servicios organizados en cuatro líneas de negocio: Banca Universal, a través del Banco de Crédito del Perú - BCP y Banco de Crédito de Bolivia; Microfinanzas, a través de Mibanco y Mibanco Colombia; Seguros y Fondos de Pensiones, a través de Grupo Pacífico y Prima AFP; y Banca de Inversión y Gestión de Patrimonios, a través de Credicorp Capital, Gestión de Patrimonios del BCP y Atlantic Security Bank. Asimismo, a través de Krealo, el brazo innovador de la corporación, se crean, invierten y gestionan Fintech en la región.



Figura 1: Mapa Credicorp

Fuente Web Grupo Credicorp

2.2 MISIÓN

PROPOSITO

Transformar Planes en Realidad.

Estar siempre contigo, alentando y transformando tus sueños y planes en realidad y con el Perú, construyendo su historia de desarrollo y superación.

2.3 VISIÓN

ASPIRACION

- ¿ Ser la empresa peruana que brinda la mejor experiencia a los clientes. Simple, cercana y oportuna.
- ¿ Ser la comunidad laboral de preferencia 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.

2.4 ORGANIZACION DE LA EMPRESA

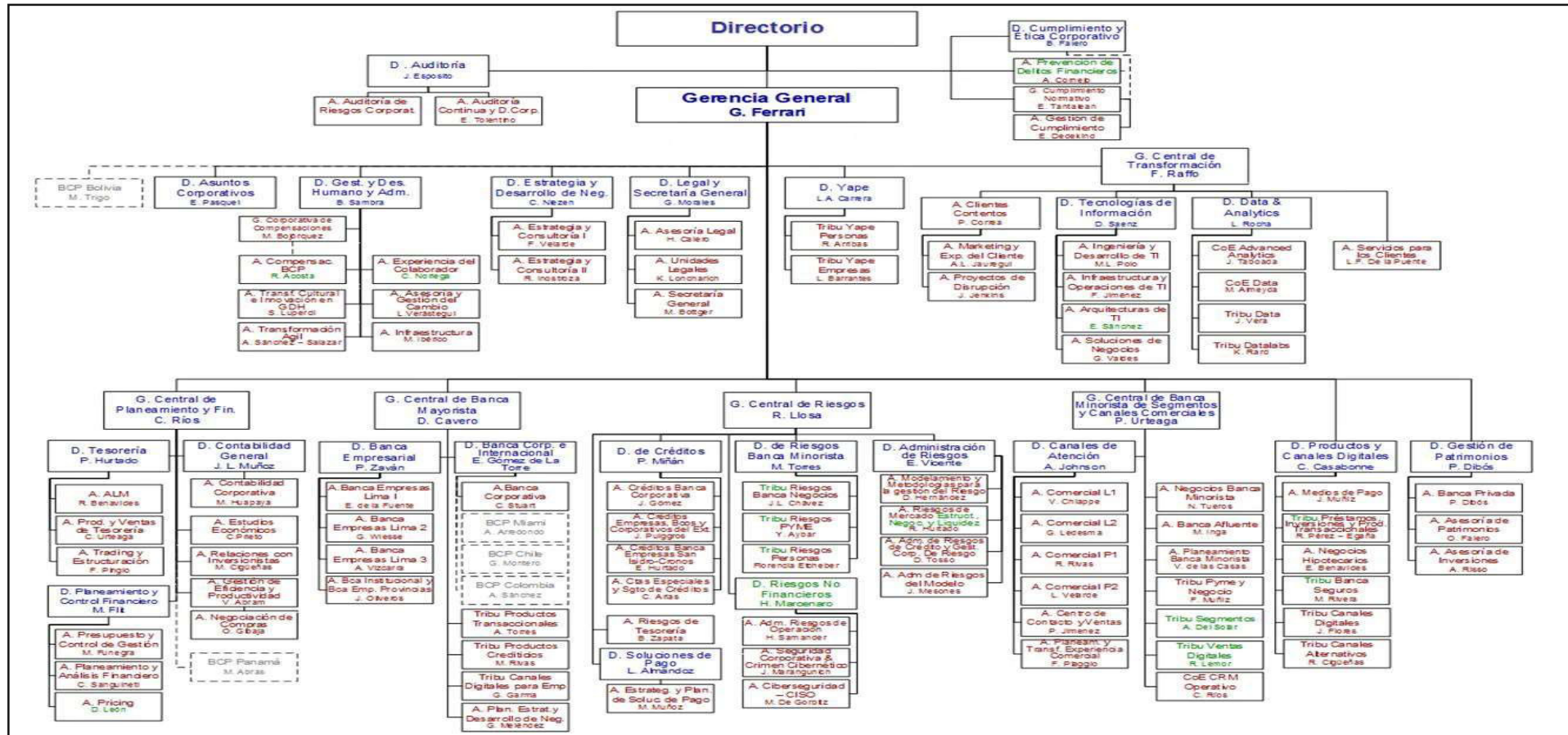


Figura 2: Organigramma BCP

Fuente: BCP - 2019

2.4.1 ORGANIZACIÓN DE LA DIVISIÓN DE CUMPLIMIENTO

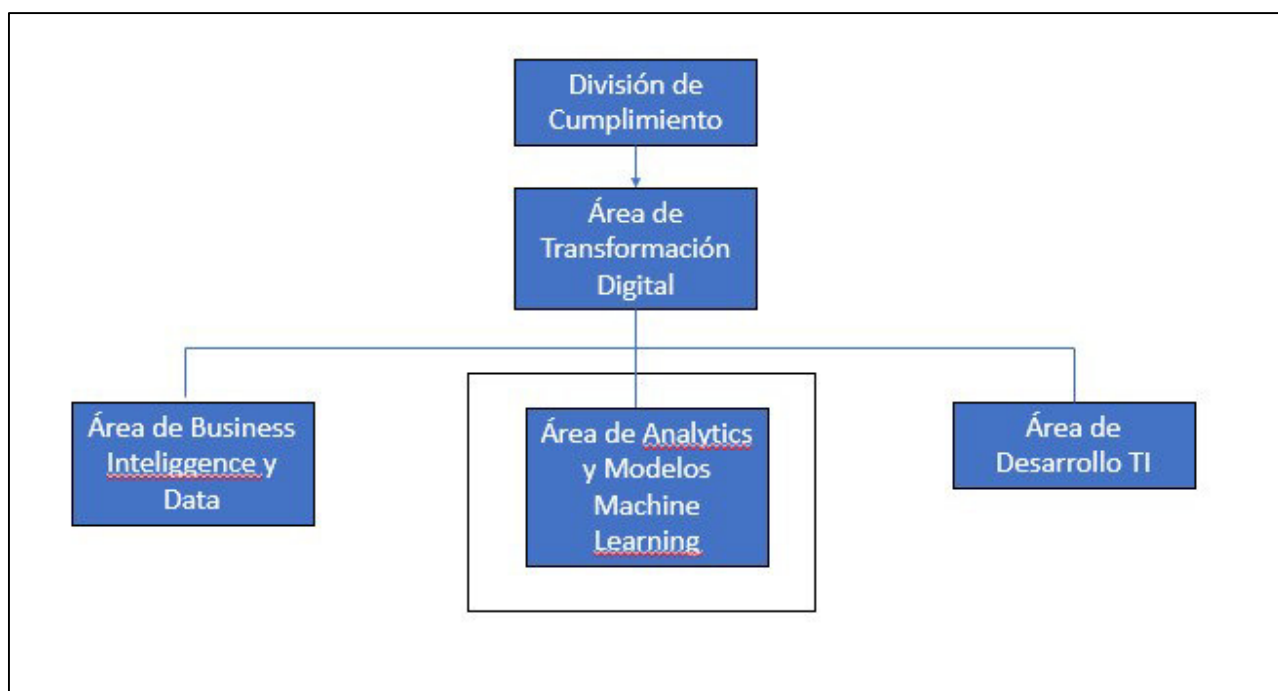


Figura 3: Organigrama de la División de Cumplimiento
Fuente elaboración propia,

2.5 AREA, CARGO Y FUNCIONES DESEMPEÑADAS

AREA: División de Cumplimiento

CARGO: Subgerente Adjunto de Analytics de Cumplimiento

FUNCIONES DESEMPEÑADAS:

- ¿ Creación de Modelos Machine Learning para detectar clientes con Operaciones Sospechosas.
- ¿ Promover la innovación con nuevas técnicas estadísticas y algoritmos avanzados.
- ¿ Reta el status quo a Nivel de Squad.
- ¿ Promover la cultura de mejora y aprendizaje continuo
- ¿ Realizar constante retroalimentación con todos los miembros del equipo
- ¿ Realizar seguimiento a los modelos creados.
- ¿ Realizar presentaciones de alto impacto para nuestros stakeholders.

AREA: Soluciones de Pago

CARGO: Data Scientist

FUNCIONES DESEMPEÑADAS:

- ¿ Definir los objetivos y alcance de proyectos y requerimientos.
- ¿ Creación de Modelos de Machine Learning para la cartera morosa.
- ¿ Seguimiento a los modelos viendo su discriminación y performance de los modelos de riesgo Crediticio.
- ¿ Consolidar el plan del modelo y asegurar el compromiso de todos los interesados.
- ¿ Identificar y gestionar riesgos y resolver problemas oportunamente.
- ¿ Brindar soluciones analíticas a los equipos que lo requieran.
- ¿ Manejar grandes volúmenes de datos.

AREA: Servicio para los Clientes

CARGO: Analista de Gestión de Información.

FUNCIONES DESEMPEÑADAS:

- ¿ Mantener la Base de datos de reclamos, para que pueda ser utilizada por el Área.
- ¿ Elaborar y presentar indicadores de gestión de reclamos.
- ¿ Creación de Modelos Machine Learning para detectar el número de reclamos mensuales.
- ¿ Resolver consultas que se presentan acerca de la data de reclamos.
- ¿ Apoyo constante en temas de data a los miembros del equipo.

CAPÍTULO III CONTEXTO EN EL QUE SE DESARROLLA LA EXPERIENCIA

3.1 SITUACION PROBLEMATICA

En el Perú según fuentes de la SBS el número de Riesgo de Operaciones Sospechosas ha ido en aumento, esto genera daños a la economía nacional e incluso al bienestar del País, es el motivo por el cual la banca peruana está obligada según norma de la SBS a informar de clientes con posible Riesgo a un equipo especial de Investigación Financiera para que puedan verificar el ROS y poder aplicar las sanciones correctivas.

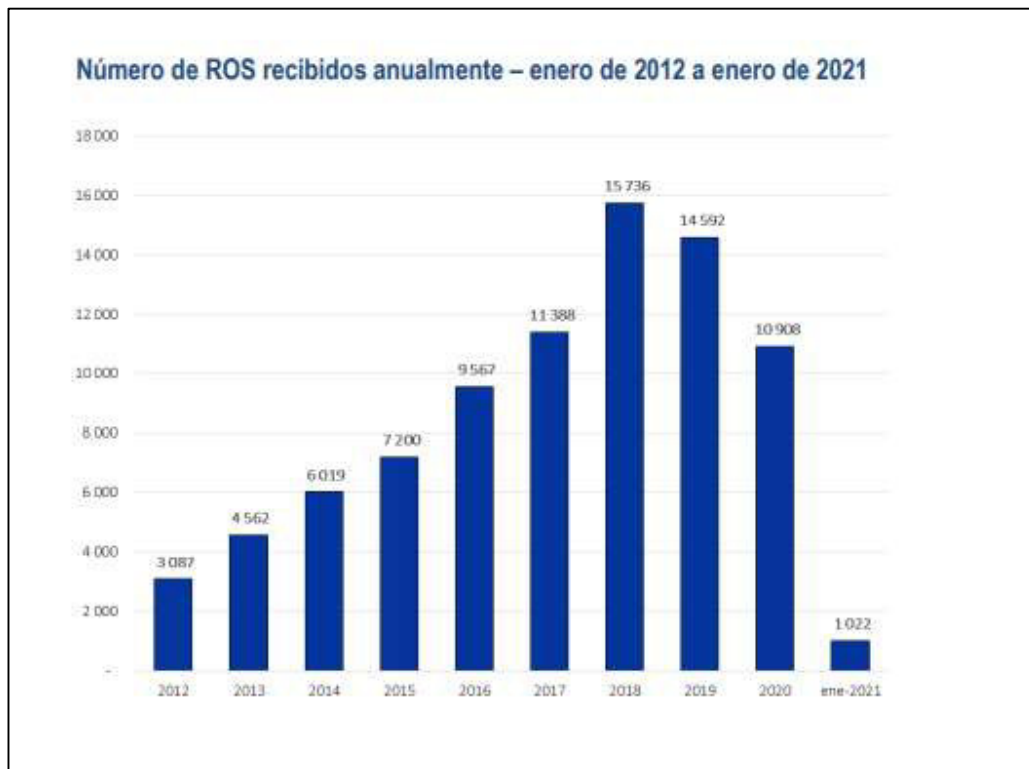


Figura 4: Evolución número de ROS
Fuente SBS,

Congelamiento de Fondos u otros Activos Convalidados Judicialmente por Posible Delito Precedente por año en que se iniciaron los casos - abril de 2012 a enero de 2021												
Delito Precedente	2012	2013	2014	2015	2016	2017	2018	2019	2020	ene-2021	Total general	Monto total en US\$
Corrupción	-	1	1	-	1	11	8	6	-	-	28	33 064 683
TID	2	1	2	2	3	1	2	-	2	-	15	4 534 088
Estafa	-	-	-	-	1	3	-	4	2	-	10	4 393 561
Crimen Organizado	1	3	-	-	-	-	-	2	-	-	6	3 136 679
Fraude Administración Personas Jurídicas	-	-	-	-	-	-	-	1	-	-	1	1 463 979
Defraudación Tributaria	-	-	-	-	-	-	1	2	-	-	3	921 428
No Se Precisa	-	-	-	2	-	1	-	-	-	-	3	549 447
Tráfico de personas	-	-	-	-	-	-	-	-	1	-	1	133 241
Phishing	-	-	-	-	-	-	-	-	1	-	1	100 977
Asociación Ilícita Para Delinquir	-	-	-	2	1	-	-	-	-	-	3	99 113
Enriquecimiento Ilícito	-	-	-	-	-	1	-	-	-	-	1	64 716
BEC y fraude financiero	-	-	-	-	-	-	-	-	1	-	1	61 551
Minería Illegal	-	-	-	2	-	-	-	-	-	-	2	39 945
Fraude	-	-	-	-	-	-	-	-	1	-	1	28 737
Extorsión	-	-	-	-	-	-	1	-	-	-	1	26 621
Terrorismo y TID	-	1	-	-	-	-	-	-	-	-	1	18 838
Peculado	-	-	-	-	-	-	-	2	-	-	2	15 444
Total general	3	6	3	8	6	17	12	17	8	0	80	48 653 047

Figura 5: Dinero congelado por tipo de Delito ROS
Fuente: SBS

El banco, dentro de la división de Cumplimiento, cuenta con procesos que permiten la identificación de ROS, a lo largo del tiempo se han creado reglas que permiten una buena identificación para cumplir con la norma de la SBS.

El banco, dentro de la división de Cumplimiento también posee un equipo de Científicos de Datos los cuales, analizando la problemática de encontrar ROS con ayuda de herramientas innovadoras, lenguajes de programación, análisis de datos y manejo de grandes volúmenes de estos se opta por implementar un modelo machine Learning que pueda identificar clientes ROS de una manera más rápida y con bases estadísticas, matemáticas y sistémicas.

Para este importante proceso Core de la división, el Área de Analytics cuenta con conocimientos implementando modelos Machine Learning o Aprendizaje Automático.

3.1.1 DEFINICION DEL PROBLEMA

El área de analytics de la división de cumplimiento tiene en sus miembros científicos de datos que pueden crear una solución con algoritmos Machine Learning para detectar Clientes con Riesgo de Operación Sospechosa.

Debido a la criticidad de este proceso se tiene que realizar un análisis completo desde la extracción de la data, limpieza de la data, análisis estadísticos de la data input, entrenamiento del modelo, escoger el modelo, revisar el performance del modelo, revisión de resultados y finalmente la puesta en producción del modelo Machine Learning.

3.2 SOLUCION

3.2.1 OBJETIVOS

OBJETIVO GENERAL

Lograr implementar un modelo que permita identificar Clientes con riesgo de Operaciones Sospechosas con ayuda de algoritmos de Machine Learning.

OBJETIVOS ESPECIFICOS

- ¿ Exploración, tratamiento y análisis estadístico de los datos input para el modelo.
- ¿ Análisis del algoritmo Machine Learning óptimo, revisión de resultados.
- ¿ Implementación del Modelo en un ambiente productivo.

3.2.2 ALCANCE

ALCANCE FUNCIONAL

El desarrollo de este proyecto es para cubrir una necesidad core del negocio, se requiere identificar clientes con Riesgo de Operación Sospechosa, para lo cual mediante técnicas de Machine Learning, análisis de datos y entendimiento del negocio se busca suplir un modelo anterior basado en juicio de experto por un modelo con algoritmos avanzados ya probados en la industria.

ALCANCE ORGANIZACIONAL

A nivel de organización, el Área de Cumplimiento es la que se beneficia con esta implementación. El proyecto además tiene un alcance a nivel de la empresa debido a que gracias a este se cuida la reputación de la entidad frente a otras empresas y la SBS que exige cumplir con lineamientos que debe seguir la entidad informando sobre clientes con Riesgo de Operación Sospechosa.

3.2.3 ETAPAS Y METODOLOGIA

Si bien el hacer modelos Machine Learning aún no es muy común en empresas peruanas ya se tiene cierta metodología dentro de la Entidad y también dentro del equipo de Data Scientist al cual pertenezco, esta metodología se ha creado en base a la experiencia y los buenos resultados que se han venido dando en los modelos anteriores.

En general se usan las siguientes 5 fases.

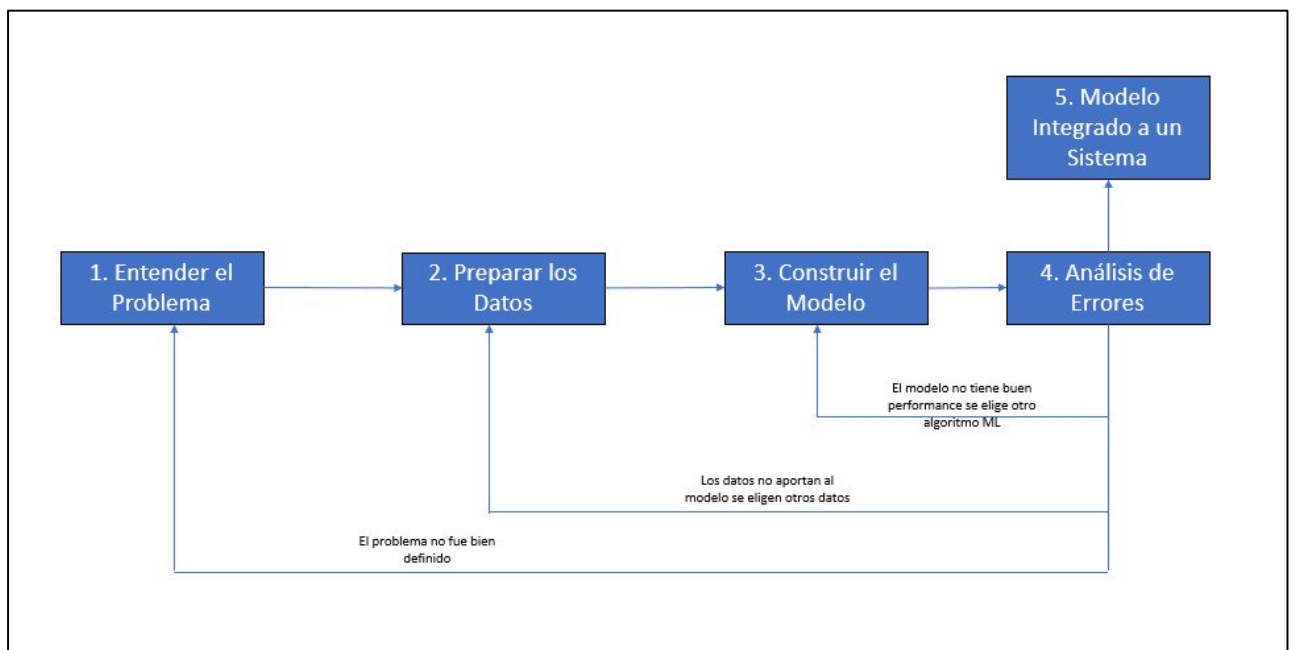


Figura 6: Fases de la creación de un modelo ML
Fuente Elaboración Propia.

Fase 1 Entender el Problema:

Cómo se menciona en el punto 3.1.1 de este informe se definió el problema el cual consiste en Implementar un modelo Machine Learning para la entidad que permita identificar clientes con Riesgo de Operación Sospechosa.

Cabe resaltar que el entendimiento del problema va de la mano con el negocio y con los expertos que ya vienen haciendo este proceso con buenos resultados pero que con técnicas de Machine Learning podrá aumentar su efectividad.

Los algoritmos de Aprendizaje Automático o Machine Learning son de varios tipos entre los más comunes tenemos:

- ¿ Aprendizaje Supervisado: En este tipo de algoritmos se requiere de datos previamente etiquetados, esto quiere decir que por cada conjunto de datos se conoce el objetivo, mediante esto el algoritmo podrá aprender una función que le sea posible predecir el objetivo para un conjunto de datos nuevo.



Figura 7: Aprendizaje Supervisado
Fuente Elaboración Propia.

- ¿ Aprendizaje no Supervisado: En este tipo de algoritmos funciona mediante datos no etiquetados explícitamente, sino que el algoritmo intenta encontrar algún tipo de relación o estructura en el conjunto de datos.

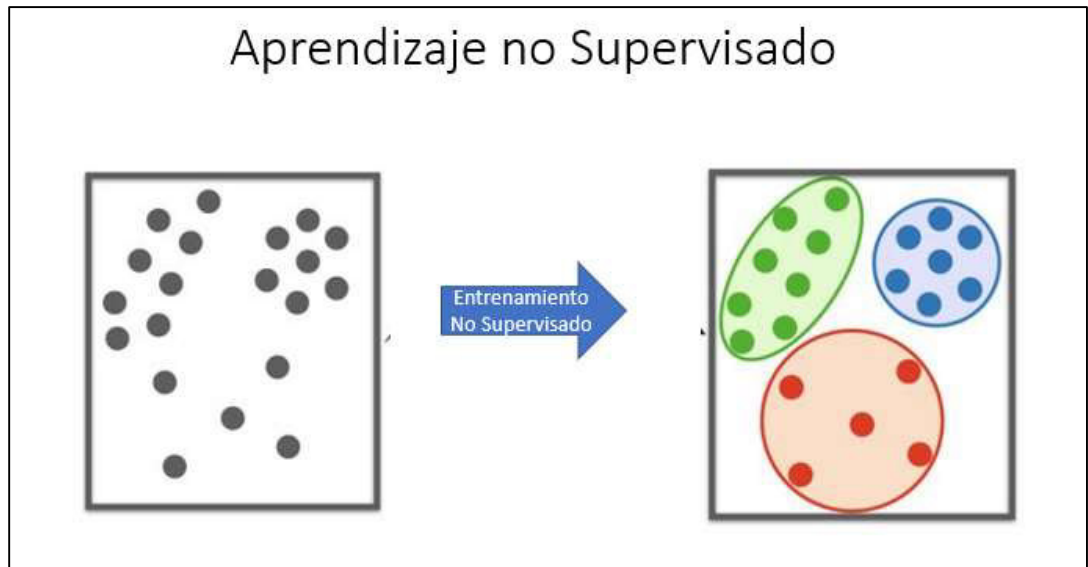


Figura 8: Aprendizaje no Supervisado
Fuente Elaboración Propia.

Por la naturaleza de nuestro problema el cual busca predecir quienes realizan ROS se eligió el modelo Supervisado.

Fase 2 Preparar los datos:

La fase de preparación de datos es una de las que requiere mayor esfuerzo y cuidado debido a que el resultado del modelo dependerá en gran medida de que datos vamos a elegir y la forma como van a entrar al modelo.

Al momento de Extraer la data se consideran diferentes fuentes, en el caso de la entidad se hizo uso del Datawarehouse y del Datalake, para su extracción se hace uso del lenguaje SQL en el caso se tengan los datos estructurados y para datos no estructurados como el BIG DATA se usó SPARK que también usa sentencias SQL.

Extraer los datos solo es la parte inicial, luego estos tienen que ser limpiados, existen diferentes formas para preparar el set de datos que ingresar al modelo

Se tienen criterios de nulidad por ejemplo si una variable como la edad presenta muchos nulos esta puede ser descartada o se rellenan los valores con técnicas estadísticas como la mediana, la moda o el promedio.

En este paso también se hacen transformaciones a los datos agrupándolos o creando más variables a partir de las ya existentes esto es conocido como darle formato a los datos o manipulación de datos.

Se realizaron las siguientes consideraciones para limpiar la data y tener el set de datos final.

- ¿ Análisis de % de nulos, se excluyen aquellos que tienen más del 80% de nulos.
- ¿ Cuando la desviación estándar es igual a 0 (son variables que solo tienen un valor sin considerar los nulos)
- ¿ Se hace un análisis de correlación la cual consiste en encontrar algún tipo de relación entre las variables ya sea positiva o negativa, usamos la correlación de Spearman para datos de tipo continuo y la correlación de Pearson para datos cuantitativos.

Un set de datos antes de entrar al modelo debería quedar de la siguiente manera:

[7]:

	CODMES	CODCLAVECIC	PROM_U6M_MOVI	PROM_U6M_TRX	FACTOR_CANAL	FACTOR_CLI	FACTOR_ZONA_F	FACTOR_PROD_F	SCOREFINAL	SCORE_ZG_AGRUPADO	CLI_AGP	PROD_AGP	CANAL_AGP	ZONA_AGP	MTD_AGP
0	202106	12405247	0.0	0	100.0	46.00	10.0	100.0	28.0	10	50	100	100	10	0
1	202106	8721557	0.0	0	100.0	23.50	10.0	100.0	35.5	50	10	100	100	10	0
2	202106	14099085	0.0	0	100.0	52.75	10.0	100.0	NaN	10	50	100	100	10	0
3	202106	2800984	0.0	0	100.0	46.00	10.0	100.0	NaN	10	50	100	100	10	0
4	202106	3182200	0.0	0	100.0	23.50	10.0	100.0	28.0	10	10	100	100	10	0

Figura 9: Ejemplo Dataset
Fuente Elaboración Propia.

Si bien para la extracción de datos se hizo uso de SQL y SPARK para su manipulación y análisis estadístico se utilizó Python.

Para terminar esta fase se quedó con 13,000 clientes, una ventana de tiempo de 1 año y un set de datos de 195 variables, entre las cuales destacan variables transaccionales, demográficas, pagos a entidades como SUNAT entre otras.

Fase 3 Construir el Modelo:

Cómo se menciona en la fase anterior ya tenemos nuestro dataset listo para hacer uso de un algoritmo de machine learning, previo a esto se tiene que partir la data en 2 universos y crear un set de datos de validación:

- ¿ TRAIN: Contiene 70% del total del set de datos y sirve para entrenar al modelo.
- ¿ TEST: Contiene el 30% del total del set de datos y sirve para testear o medir el desempeño y performance del modelo.
- ¿ VALIDATION: Contiene datos que no pertenecen al dataset del modelo pero que nos van a servir para hacer una validación del performance y desempeño del modelo, comúnmente también se le conoce como datos fuera de tiempo.

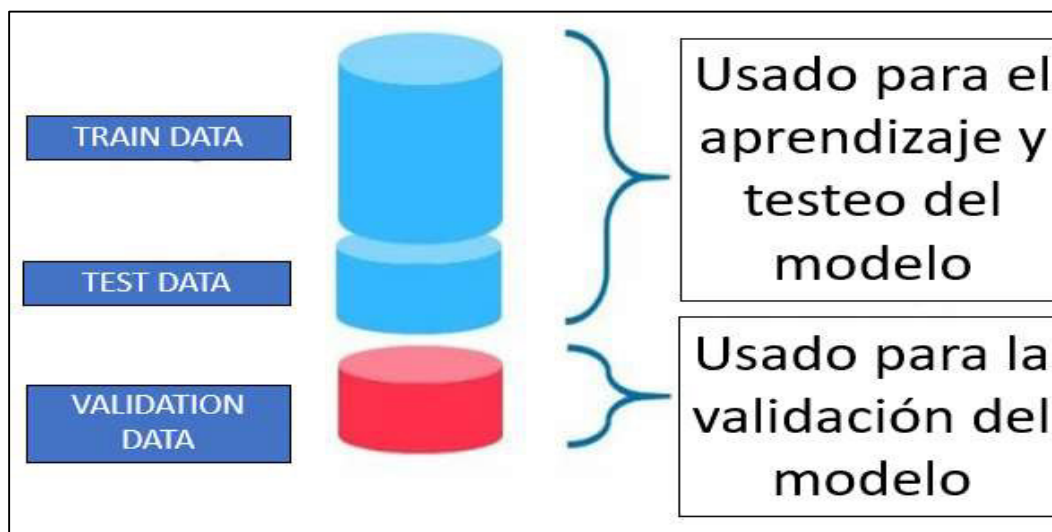


Figura 10: Separación del Dataset
Fuente Elaboración Propia.

En esta fase se eligió el algoritmo para llegar al objetivo planteado, se probaron diferentes técnicas entre ellas tenemos:

- ¿ Random Forest o Bosques Aleatorios: Algoritmo basado en árboles de decisión en donde cada árbol depende de los valores de un vector aleatorio.
- ¿ XGBOOST: Algoritmo supervisado muy utilizado y con buenos resultados debido al uso del paralelismo, muy útil para grandes cantidades de datos como el nuestro.
- ¿ LGBM: Algoritmo mucho más optimizado que el XGBOOST debido a que es más ligero y se adapta de manera perfecta a nuestro dataset como se podrá observar más adelante.

Para ver si un modelo es mejor que el otro para nuestro set de datos utilizamos el AUC, que es una métrica que nos permite identificar que tan bien funciona nuestro modelo, por ejemplo, si queremos distinguir un modelo que predice verdadero y falso una mediada de 0.7 de AUC significa que hay 70% de probabilidad de que el modelo pueda acertar.

La medida optima del AUC es cuando es cercana a 1 y la mínima es 0.5 entre ese rango se puede observar si el modelo es excelente, bueno, regular o malo.

Bajo esta métrica comparamos nuestros modelos y se obtuvieron los siguientes resultados:

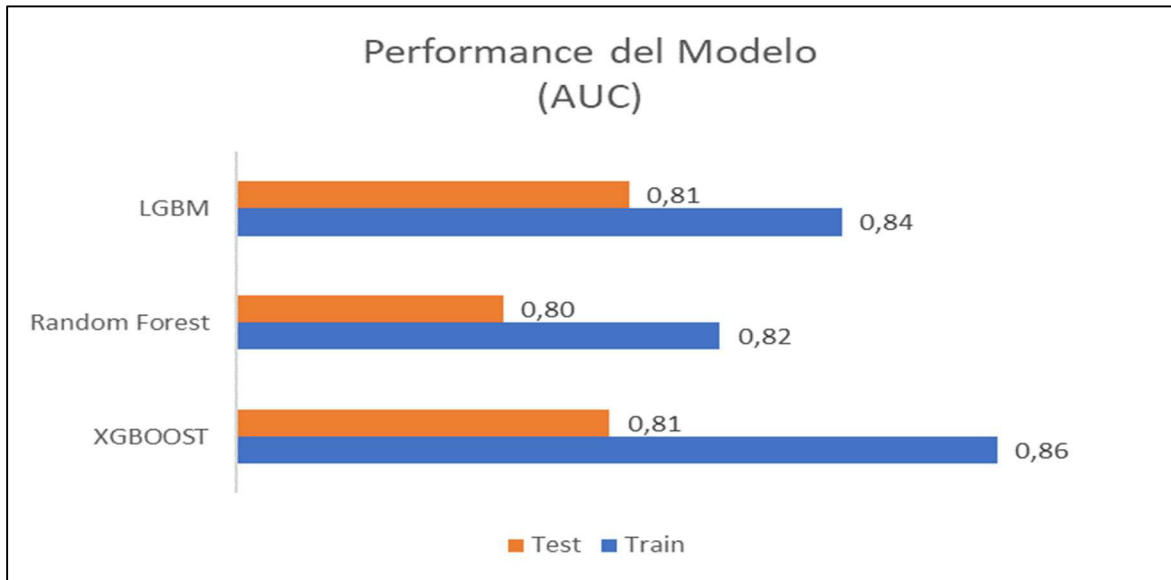


Figura 11: Comparación de Modelos
Fuente Elaboración Propia.

Como se puede observar los modelos XGBOOST y LGBM tienen métricas similares, podríamos optar por cualquiera de ellos sin embargo también se observa la diferencia de AUC en el periodo TRAIN y TEST donde el modelo LGBM presenta menor diferencia esto quiere decir que se comporta de manera parecida es por ello y por el costo computacional bajo del algoritmo que optamos por el modelo LGBM.

Fase 4 Análisis de Errores:

En esta Fase es necesario a parte del AUC de los periodos TRAIN y TEST es necesario ver el periodo VALIDATION debido a que es aquí donde pueden surgir algunos errores.

Para nuestro caso la medida AUC para el periodo VALIDATION fue de 0.847 que es muy parecido al periodo TRAIN y es bastante alto por lo que podemos concluir que el modelo funciona bien.

En casos donde el AUC tiende a bajar en el periodo de validación se tiene que revisar si el problema fue definido de manera óptima con los expertos del negocio y si ese no es el problema revisar la data, incluir o quitar variables que puedan aumentar el AUC y en caso este tampoco sea el problema buscar un algoritmo Machine Learning que se adecue a nuestro dataset y seguir probando hasta encontrar un AUC decente.

En esta fase también se comprueba con los expertos del negocio si la salida de nuestro modelo tiene sentido para ellos, se les proporciona un grupo de clientes para que ellos lo evalúen con su conocimiento y nos puedan corroborar que efectivamente nuestro modelo asigna una probabilidad de Riesgo de Operación Sospechosa real.

Se genera una reunión con el gerente de división y los equipos que harán uso del modelo y se presentan los resultados quedando con el conforme del pase a producción,

Fase 5 Modelo Integrado al Sistema:

En esta fase ya se tiene el modelo aprobado y listo para su puesta en producción, se conversa con el equipo de Data Engineers, se les brinda el Script y el archivo del modelo para que ellos puedan correrlo en su servidor.

Para esta fase se requiere un entendimiento entre los científicos de Datos y los Data Engineers, se trabaja de la mano para que los resultados del modelo no se vean alterados en la puesta en producción.

Los entregables para el equipo de Data son:

- ¿ Script en Python del tratamiento de la data.
- ¿ Archivo del modelo en formato `.pkl`
- ¿ Script en SQL para la extracción de datos.
- ¿ Documentación y Manual de fuentes de datos.

Cuando el Modelo ya esté puesto en producción se tiene que hacer un seguimiento

para ver si el modelo aún mantiene un AUC a lo largo del tiempo, la corrida del modelo es de manera mensual y se pueden ver los resultados también.

Si el modelo en algún momento presenta bajo AUC se entra a revisar la data para verificar que esté pasando y poder ver que variable esté afectando el AUC, en casos extremos se tiene que reentrenar el modelo con data más actual para que el modelo aprenda los nuevos comportamientos del cliente.

3.2.4 FUNDAMENTOS UTILIZADOS

3.2.4.1 SQL

Por sus siglas en inglés 'Structure Query Language_' es un lenguaje de acceso y manipulación de bases de datos que nos permite en nuestro caso extraer y darle forma a la data.

Para un Científico de Datos es primordial el tener este conocimiento, debido a que en las organizaciones se cuenta con una base de datos generalmente estructurada.

Como se menciona en la Fase 2 Preparar los datos, este lenguaje es de vital importancia para extraer datos y manipularlos antes de ingresarlos al modelo.

SQL hoy en día se ha vuelto un lenguaje muy solicitado no solo para profesionales que se dedican al desarrollo de software sino también al análisis de datos e incluso para otras profesiones, la data se ha vuelto un bien extremadamente importante para las empresas es por ellos que se requiere de un lenguaje que pueda administrar los datos, es donde se ve el poder de SQL.

3.2.4.2 SPARK

Apache Spark se encarga del procesamiento de datos distribuidos, es un framework de programación muy rápido y multipropósito, posee flexibilidad e integración con módulos como HIVE o IMPALA para ambientes DATALAKE o mar de datos que la entidad posee.

SPARK permite entrar al mundo del BIG DATA, en nuestros procesos lo utilizamos porque tenemos en nuestro mar de datos millones de registros que SQL no podr a administrarlos debido a su limitante de performance, una consulta podr a demorar demasiado tiempo frente a una consulta hecha sobre SPARK con el m dulo HIVE o IMPALA.

Se pueden utilizar algunas sentencias SQL dentro del m dulo HIVE o IMPALA, pero son mucho m s r pidos debido a que trabajan sobre datos con formato HDFS que permite paralelizaci n y por ende hacen much simo m s r pido las consultas.

En nuestro Proyecto aproximadamente la mitad de las variables fueron extra das y moldeadas usando SPARK.

3.2.4.3 PYTHON

Python es uno de los lenguajes de programaci n mas usados hoy en d a debido a que puede ser usado tanto para desarrollo de software, an lisis de datos, Inteligencia Artificial, automatizaci n de procesos y otros.

Python es el lenguaje que se eligi  para hacer el an lisis de datos debido a que posee librer as espec ficas para el an lisis de datos y los modelos Machine Learning.

Numpy es una librer a que te permite hacer operaciones matem ticas entre matrices y es muy  til para hacer las operaciones que la data requiere.

Pandas es una librer a que nos permite manipular dataframes o set de datos, combinado con el an lisis matem tico y estad stico.

LGBM es una librer a de Python con modelo Machine Learning que permite hacer uso de este algoritmo e incluso cambiar par metros para balancear la data, ponerle hiper par metros y ajustar el algoritmo, esto es conocido como tunear el modelo para conseguir mejores resultados y se ajuste a la forma de nuestro dataset.

3.2.4.4 PLSQL DEVELOPER

Es un Sistema Gestor de Base de Datos que nos permite extraer datos del DATAWAREHOUSE de la entidad con sentencias SQL.

Como cualquier SGBD permite insertar, actualizar, crear y eliminar registros de tablas o datos estructurados.

3.2.4.5 CLOUDERA

Herramienta basada en Apache Hadoop para explotar Data Lakes, como se menciona la entidad posee grandes cantidades de datos para ello necesitamos hacer uso del BIG DATA, manipular estos millones de registros es una tarea en la que CLOUDERA nos fue de mucha ayuda con los módulos HIVE e IMPALA de SPARK.

3.2.4.6 LAPTOP

Se hizo uso de una laptop workstation con 16gb de RAM marca HP, con procesador INTEL XEON, para manejar grandes cantidades de data y poder entrenar un modelo se requiere equipos de gran potencia.

3.2.5 IMPLEMENTACION DE LAS AREAS, PROCESOS, SISTEMAS Y SUS BUENAS PRACTICAS

Ahora se van a describir cada una de las fases del proyecto y los entregables que evidencian las buenas prácticas usadas a lo largo del Proyecto, se hará énfasis en las partes donde el autor tuvo mayor involucramiento.

Los artefactos se encuentran en la parte de Anexos, algunos de ellos tuvieron que ser simulados debido a la cláusula de seguridad y no divulgación de información que tiene la Entidad bancaria.

Fase 1 Entender el Problema:

En esta etapa se tuvo 4 reuniones mediante el aplicativo Teams para plasmar lo que los stakeholders requieren.

Se definieron los tiempos de trabajo siguiendo la metodología SCRUM en 6 Sprints cada uno de ellos de 2 semanas de duración en total 12 semanas.

Al final de esta fase de reuniones se tiene como entregable:

- ¿ Diseño conceptual de la solución.

Fase 2 Preparar los datos:

Según la metodología propia que seguimos se hizo uso de:

- ¿ PLSQL
- ¿ CLOUDERA
- ¿ IMPALA
- ¿ SPARK
- ¿ PYTHON

Cada uno de estos softwares tienen vital importancia en el tratamiento de datos tanto estructurados como no estructurados.

El autor tuvo participaci3n con cada uno de los softwares mencionados y asegur3ndose de la calidad de la data mediante an3lisis de estabilidad de la data, validaci3n de fuentes y todo lo que se fuese necesario para tener un set de datos 3ptimo para el algoritmo de Machine Learning.

Como entregable de esta etapa se tiene el set de datos en PLSQL-ORACLE y PYTHON para seguir con el proceso de modelaci3n y los Scripts en PYTHON y SQL para el equipo encargado de la puesta en producci3n el modelo.

Fase 3 Construir el Modelo:

En esta fase ya teniendo el set de datos cada miembro del equipo que desarrolla la solución se encargó de probar distintos algoritmos de machine Learning todo esto utilizando PYTHON.

El autor probó el algoritmo de LGBM el cual finalmente fue elegido por su bajo consumo en recursos y tener buenos indicadores y además que cumplía con el objetivo.

El entregable para esta etapa es el un archivo en formato `.sav_` el cual contiene el modelo ya entrenado con las métricas y objetivos alcanzados.

Fase 4 Análisis de Errores:

En esta Fase según la metodología propia se ven las métricas sobre una data de Validación el cual indicaba buenos resultados.

El autor de este informe se encargó de preparar los datos de validación y correr el modelo sobre esta nueva data.

Como entregable de esta fase se tiene las métricas del modelo sobre la nueva data y el conforme del equipo de científicos de datos.

Fase 5 Modelo Integrado al Sistema:

En esta fase se realizó el seguimiento a la puesta en producción de la mano con el equipo de Data Engineers.

El autor de este informe tuvo que validar que la salida del modelo productivo sea 100% igual que el modelo de construcción para dar el conforme y aceptar que el modelo productivo cumple con lo esperado.

3.3 EVALUACIÓN

3.3.1 EVALUACIÓN ECONÓMICA

El costo del proyecto lo dividimos en gastos de personal, y gastos de tecnología

Tabla 5: Costo estimado del proyecto

Descripción	Cantidad	Costo unitario	Costo total	Costo total
Gastos del Personal				
Product Owner	1	10000	3	S/30,000.00
Team members	3	6000	3	S/18,000.00
Gastos tecnológicos estimados de depreciación				
Laptops	4	500	3	S/1,500.00
Todos los Software	4	400	3	S/1,200.00
Total				S/50,700.00

3.3.2 BENEFICIOS OBTENIDOS

El beneficio obtenido lo vamos a calcular como cuanto de lavado de dinero podemos evitar al encontrar clientes con Riesgo de Operación sospechosa, la SBS tiene un estimado de ROS por año de entidades bancarias además de la participación de la entidad en el Perú así que podemos ver cuanto dinero pudimos evitar que sea lavado y afecte al País.

El beneficio que la entidad valora también es la reputación bancaria y la reputación del holding.

Tabla 6: Estimado beneficio del País al evitar ROS

59% [[9	1! [hw
wh{ τ Π I 0%XT ↑ ot oXEt ↑ ixΠT ¼ Vx	えんをいれ
Ξot ixΠT ¼ Vx	えん
. τ ΠT ¼ Vx	えん

CAPÍTULO IV

REFLEXIÓN CRÍTICA DE LA EXPERIENCIA

- ¿ La participación del autor del presente informe y en el proyecto de implementación del modelo Machine Learning fue como Subgerente Adjunto de Analytics, específicamente como Científico de Datos, si bien el autor cuenta con 3 años de experiencia creando modelo. Cabe mencionar que el proyecto tuvo bastante complejidad y aprendizaje continuo al ser un escenario nuevo.
- ¿ El equipo compuesto de trabajo 3 Científicos de datos y un Product Owner tuvieron que realizar el trabajo en el contexto de trabajo remoto lo cual provocó el alto compromiso de cada miembro y la comunicación eficaz a pesar de la distancia.
- ¿ El equipo de Científicos de Datos tiene diferentes habilidades, algunos más en el tema de data, otros en el tema de modelos y puesta en producción, sin embargo, se logró usar cada habilidad para finalizar el proyecto de manera correcta.
- ¿ La etapa de extracción de datos fue la que llevo mayor esfuerzo al equipo debido a la gran cantidad de data y la búsqueda de variables que fueran importantes para el modelo y para el negocio.
- ¿ El uso de SCRUM ayudó bastante al equipo en definir metas en cada Sprint, si bien en algunos casos hubo retrasos estos se subsanaban en el siguiente Sprint adelantando y cumpliendo con lo comprometido.
- ¿ No existe hoy en día una metodología de trabajo que se adapte exactamente para equipo que crean modelos de machine learning, sin embargo, el uso de SCRUM ayudó bastante con la agilidad y cumplir metas.
- ¿ Durante la fase de elección del algoritmo de machine learning se revisaron papers para poder entender el funcionamiento de cada algoritmo y como mejorarlos.
- ¿ En la etapa de corrección de errores hubo opiniones diferentes sobre métricas e indicadores los cuales fueron resueltos con ayuda del Product Owner.

- ¿ En la fase de puesta en producción se tuvo que trabajar de la mano con el equipo de Ingenieros de Datos, esto provoco un trabajo bastante cercano y el resultado fue el esperado.
- ¿ El potencial que tiene los algoritmos de Machine Learning en las empresas es muy grande, estamos en la época donde la data vale muchísimo y aprovecharla aún más, estos algoritmos se pueden utilizar en diferentes negocios no solo en Banca.
- ¿ El potencial de la data y del BIG DATA en las empresas está siendo muy aprovechado, aquellas empresas que no se unan a la ola seguramente quedarán rezagadas.
- ¿ Se culmino de manera exitosa con todas las fases del proyecto y de manera satisfactoria, hoy en día el modelo corre una vez cada mes e incluso se han recibido premiaciones por parte de la división y reconocimiento a nivel Banco.

CAPÍTULO V CONCLUSIONES Y RECOMENDACIONES

4.1 CONCLUSIONES

- ¿ La etapa más importante fue la fase que tiene que ver con los datos, extraerlos y analizarlos, con el conocimiento en SQL, SPARK y PYTHON se pudo hacer frente a esta tarea de manera satisfactoria.
- ¿ El dataset final con todo el tratamiento de datos antes de ingresar al modelo sirvió no solo para el modelo sino también para sacar conclusiones más analíticas y relación entre variables que puedan ayudar en otros procesos del negocio.
- ¿ Se logró hacer todos los análisis matemáticos y estadísticos a los datos a pesar de tener que el proceso pudiese ser repetitivo.
- ¿ La etapa de elección del modelo se logró gracias a que cada miembro del equipo probó diferentes técnicas de acuerdo con su conocimiento, El modelo LGBM demostró su potencia con nuestros datos y con el negocio.
- ¿ Las métricas del modelo son consideradas como muy buenas y esto se demostró cuando comparamos el equipo de analistas versus el modelo.
- ¿ Se logró la puesta en producción con el equipo de Ingenieros de Datos debido al feedback continuo con los Científicos de Datos, ambos trabajaron de la mano por tener el entregable preciso que sea de uso para el negocio.
- ¿ Hoy en día se hace un proceso de seguimiento al modelo, evidenciando que este sigue funcionando y tiene las mismas métricas de cuando fue creado.
- ¿ El modelo sigue siendo observado de manera mensual por si en algún momento pierde poder de discriminación a esta fase post modelación se le conocen como etapa de seguimiento del modelo.

4.2 RECOMENDACIONES

- ¿ Hoy en día no existe una metodología al nivel de SCRUM o KANBAN para equipos de Científicos de datos, Machine Learning o Inteligencia Artificial, es recomendable para tener proyectos exitosos es tener una adaptación propia de una de estas metodologías de trabajo para tener proyectos exitosos.
- ¿ Todas las fases que tienen que ver con la data ya sea extracción, análisis y otros deben ser tomadas como la etapa en donde se le debe poner todo el esfuerzo posible, ya que de este dependerá en su mayoría el éxito del proyecto.
- ¿ Actualmente existen algoritmos mucho más avanzados como las redes neuronales es posible que incluso den mejores métricas, probar estos modelos debe ser los siguientes pasos.
- ¿ Dentro de los algoritmos de Machine learning existe hoy en día mucha innovación, estar al corriente con la tendencia mundial es vital para poder aplicarlo en las empresas.
- ¿ Los conocimientos de un Científico de datos van desde estadística, matemática, programación y hoy en día también de tecnología pues la puesta en producción se hace sobre contenedores DOCKER.
- ¿ Actualmente se está tratando de migrar todo este proceso a la nube con servicios como AWS, AZURE, GCP, para los Científicos de datos es indispensable tener también estos conocimientos.
- ¿ Un tema importante para tomar en cuenta es la ley de protección de datos, esto puede ser una limitante para crear modelos robustos así que se puede codificar estos datos para poder usarlos.
- ¿ Un modelo Machine Learning con el tiempo puede perder discriminación básicamente porque el comportamiento de los seres humanos cambia con el tiempo o se ven afectados por eventos como el COVID para poder vigilar si el modelo sigue funcionando de manera correcta es necesario hacer un seguimiento a la data y al propio modelo.

- ¿ Una competencia muy importante para científicos de datos es el poder comunicar su solución frente a personas que no tienen mucho conocimiento técnico o sobre modelos o Machine learning.
- ¿ La definición del problema es muy importante es por eso por lo que escuchar y entender al negocio es de suma importancia para tener un buen proyecto.

4.3 FUENTES DE INFORMACIÓN

- ¿ Modelo XGBOOST Recuperado de
<https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf>
- ¿ Modelo Random Forest (2001) Recuperado de
<https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>
- ¿ Modelo LightGBM Recuperado de
<https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>
- ¿ BCP (2019). Nosotros. Recuperado de
<https://www.viabcp.com/nosotros>
- ¿ BCP (2019). Memoria anual 2019. Lima, Perú. Recuperado de
<https://www.viabcp.com/buscador?buscar=Memoria%20Anual%202018>
- ¿ SBS (2020). Información Estadística de UIF Recuperado de
https://www.sbs.gob.pe/Portals/5/jer/ESTADISTICAS-OPERATIVAS/Bol_abril_2020.pdf

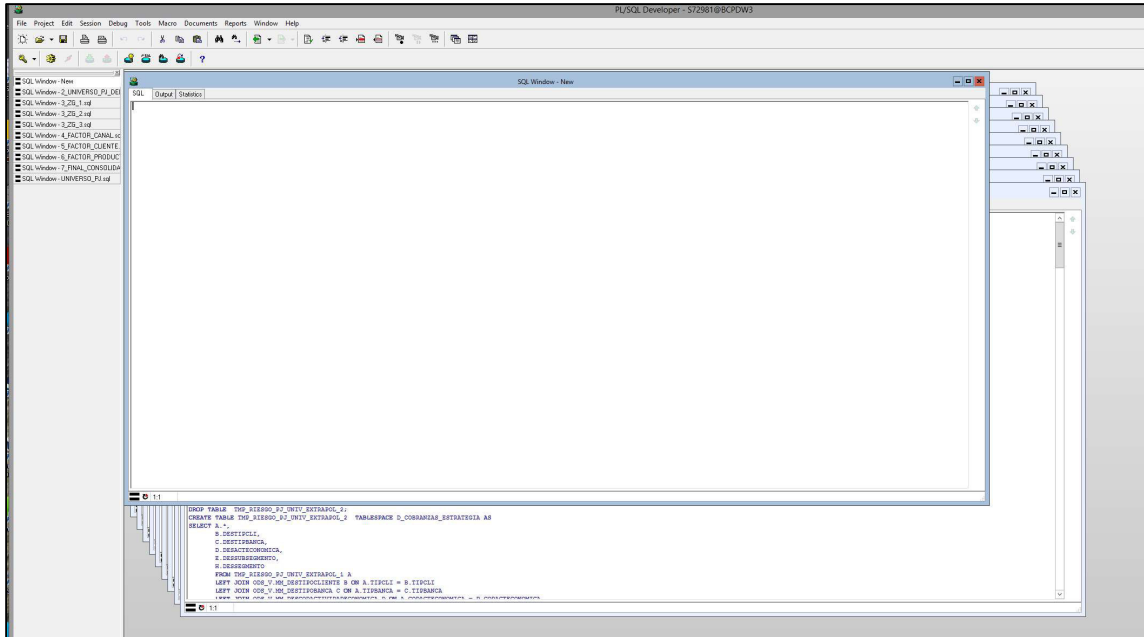
4.4 GLOSARIO

- ¿ SBS: Superintendencia de Banca y Seguros y AFP, organismo encargado de regular y supervisar del sistema financiero.
- ¿ ROS: Riesgo de Operación Sospechosa en el sistema financiero.
- ¿ Machine Learning: Campo de la inteligencia artificial y que se centra en diseñar y desarrollar algoritmos para emular la inteligencia humana aprendiendo de los datos.
- ¿ Big Data: Término que se usa para referirse a la manipulación de gran cantidad de datos.
- ¿ SQL: Lenguaje de programación para manipular datos estructurado.

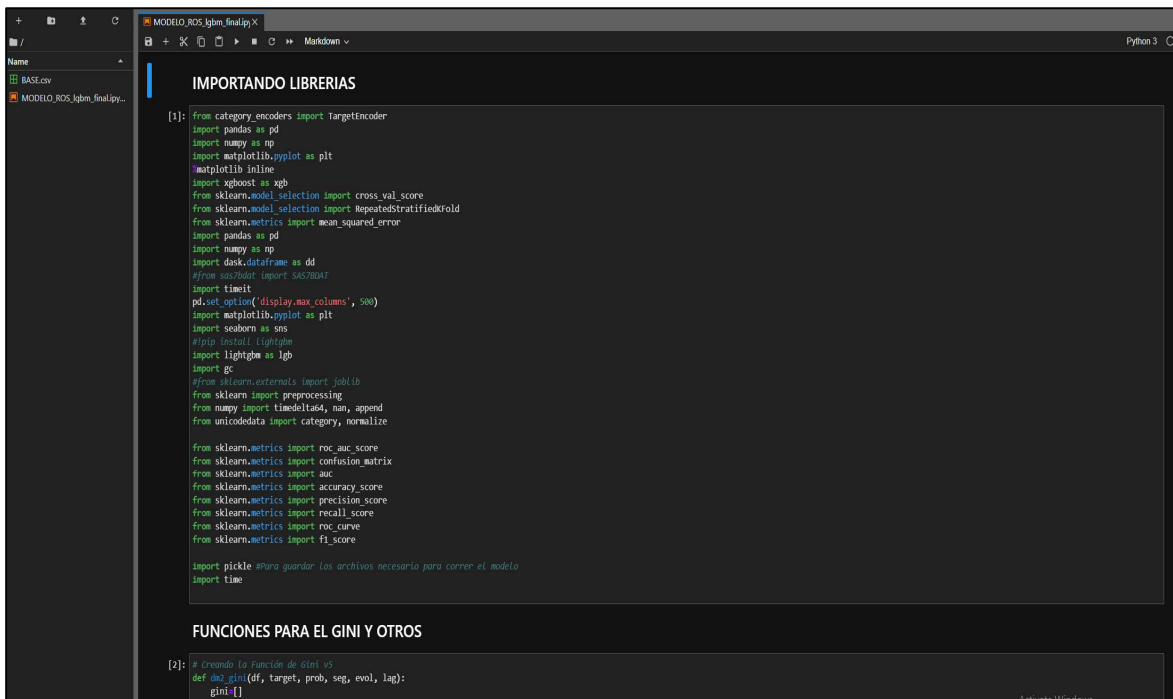
ANEXOS

Implementación en Fases del Modelo de Machine Learning

- Entregable: Script SQL

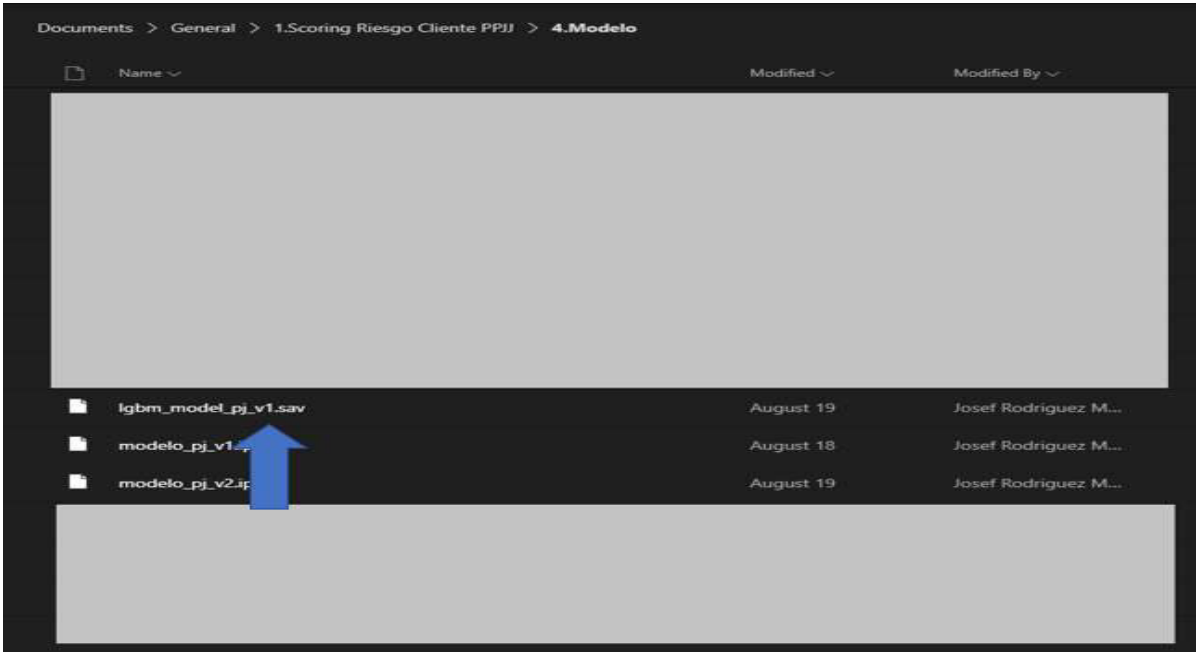


2 Entregable: Script Python



ANEXO 3A. FASE 3 PREPARAR LOS DATOS

¿ Entregable: Modelo entrenado en formato .sav



ANEXO 4A. FASE 4 ANÁLISIS DE ERRORES

¿ Entregable: AUC en el periodo de Validación.

