DISEÑO E IMPLEMENTACIÓN DE SOLUCIONES DE AUTOMATIZACIÓN PARA AUDITORÍA INTERNA BANCARIA: NLP PARA SQR Y VALIDACIÓN DE CALIDAD DE DATOS.

PRESENTADO POR:

MIGUEL ANGEL MARTINEZ MESA

DIRECTOR:

ING. JESÚS AUGUSTO GUZMÁN LOZANO

UNIVERSIDAD SANTO TOMÁS
INGENIERIA DE TELECOMUNICACIONES
OPCIÓN DE GRADO
BOGOTÁ, D.C.
2025

MIGUEL ANGEL MARTINEZ MESA

MONOGRAFIA

DIRECTOR:
ING. JESÚS AUGUSTO GUZMAN LOZANO

UNIVERSIDAD SANTO TOMÁS
INGENIERIA DE TELECOMUNICACIONES
PROYECTO DE GRADO – MODALIDAD PRÁCTICA PROFESIONAL
BOGOTÁ, D.C.
2025

Nota de Aceptación

_____

_____

_____

_____

Presidente del Jurado

_____

Jurado

_____

Jurado

Bogotá, D.C, 2025

AGRADECIMIENTOS

Mi familia,
Fuente principal para llevar a cabo este proceso de aprendizaje personal y profesional. Su apoyo incondicional ha sido esencial en cada paso de este camino.

Universidad Santo Tomás,
Fuente de conocimiento, formación y crecimiento académico. Gracias por brindarme las herramientas necesarias para desarrollarme como profesional.

Profesores,
Fuente de experiencia, dedicación y esfuerzo durante todo el proceso universitario. Su guía ha sido clave en mi formación.

Compañeros de universidad,
Por el soporte brindado durante esta etapa y por la gran cantidad de experiencias compartidas en la conquista de esta meta.

CONTENIDO

Pág.

LISTA DE TABLAS

LISTA DE GRÁFICAS

LISTA DE FIGURAS

RESUMEN

Durante el periodo de prácticas profesionales, se desarrollaron dos proyectos orientados a mejorar los procesos internos y la experiencia del cliente en una entidad bancaria. El primer proyecto consistió en la clasificación y análisis de quejas y reclamos presentados por los clientes, utilizando técnicas de machine learning para automatizar la categorización de los casos y detectar patrones relevantes. A través del procesamiento de lenguaje natural (NLP) y algoritmos de clasificación, se logró identificar áreas críticas de atención y oportunidades de mejora en los servicios ofrecidos por el banco. Este proyecto representó un aporte significativo a la formación como estudiante de Ingeniería en Telecomunicaciones, al permitirme aplicar conocimientos en el tratamiento, análisis y gestión de grandes volúmenes de datos, habilidades fundamentales en el campo de las tecnologías de la información.

El segundo proyecto estuvo enfocado en la originacion de datos, específicamente en el cruce y verificación de la información proporcionada por los clientes. Se diseñó un proceso para evaluar la calidad de los datos registrados, con el fin de garantizar su integridad, consistencia y confiabilidad. Esta iniciativa contribuyó a optimizar los procesos de validación y a reducir errores en la gestión de información, fortaleciendo así la eficiencia operativa de la entidad.

Ambos proyectos permitieron aplicar conocimientos teóricos en un entorno real, desarrollar habilidades analíticas y proponer soluciones prácticas a problemáticas comunes en el sector financiero.

PALABRAS CLAVE: quejas y reclamos, análisis de datos, machine learning, calidad de la información, originacion, sector bancario, prácticas profesionales, ingeniería en telecomunicaciones.

1. GLOSARIO

Accuracy (Exactitud): Métrica de evaluación en modelos de clasificación que mide la proporción de predicciones correctas sobre el total de casos.

API (Application Programming Interface): Conjunto de funciones y procedimientos que permiten la comunicación entre diferentes aplicaciones o sistemas.

CRM (Customer Relationship Management): Sistema de gestión de relaciones con clientes, utilizado para organizar, automatizar y sincronizar procesos de ventas, marketing, atención y soporte.

CSV (Comma-Separated Values): Formato de archivo basado en texto plano en el que los valores están separados por comas, comúnmente usado para el intercambio de datos.

Dataset: Conjunto de datos utilizado para el entrenamiento, validación o prueba de modelos de análisis y aprendizaje automático.

GUI (Graphical User Interface / Interfaz Gráfica de Usuario): Medio visual de interacción entre el usuario y el sistema, basado en ventanas, botones y menús.

KPI (Key Performance Indicator / Indicador Clave de Desempeño): Medida cuantitativa que permite evaluar el grado de cumplimiento de un objetivo.

Machine Learning (Aprendizaje Automático): Subcampo de la inteligencia artificial que permite a los sistemas aprender y mejorar automáticamente a partir de datos sin ser programados explícitamente.

Matriz de confusión: Herramienta de evaluación en clasificación que muestra el número de aciertos y errores del modelo en cada clase.

Naive Bayes: Algoritmo de clasificación basado en el teorema de Bayes, que asume independencia condicional entre las variables predictoras.

NLP (Natural Language Processing / Procesamiento de Lenguaje Natural): Campo de la inteligencia artificial que estudia la interacción entre computadoras y lenguaje humano.

Precision (Precisión): Métrica que mide la proporción de verdaderos positivos entre todos los casos clasificados como positivos por el modelo.

Recall (Cobertura o Sensibilidad): Métrica que mide la proporción de verdaderos positivos detectados sobre el total de positivos reales.

SQR (Solicitudes, Quejas y Reclamos): Categoría de registros que agrupan la interacción de clientes con una entidad financiera respecto a problemas, solicitudes o inconformidades.

Stopwords: Palabras comunes en un idioma (como “de”, “la”, “y”) que usualmente se eliminan en el procesamiento de texto porque no aportan significado relevante.

Vectorización: Proceso de transformar texto en representaciones numéricas (vectores) que pueden ser interpretados por algoritmos de machine learning.

Tkinter: Librería estándar de Python para el desarrollo de interfaces gráficas de usuario (GUI).

2. INTRODUCCIÓN

El desarrollo tecnológico y la transformación digital en el sector financiero han impulsado la adopción de herramientas de análisis de datos y automatización que permiten mejorar la eficiencia operativa y la calidad del servicio al cliente. En este contexto, la Ingeniería en Telecomunicaciones desempeña un papel fundamental al integrar conocimientos técnicos, analíticos y de gestión de la información para diseñar soluciones orientadas a la optimización de procesos empresariales.

Durante el periodo de prácticas profesionales, se desarrollaron dos proyectos en el área de Auditoría Interna de una entidad bancaria. El primero se centró en la **clasificación automática de solicitudes, quejas y reclamos (SQR)** mediante técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático (machine learning). Esta solución permitió reducir el tiempo de análisis manual, mejorar la precisión en la categorización de casos y generar indicadores más confiables para la toma de decisiones.

El segundo proyecto abordó la **verificación y validación de datos de clientes durante el proceso de originación**, con el objetivo de fortalecer la calidad, consistencia y trazabilidad de la información. Mediante el uso de reglas de negocio, catálogos maestros y una interfaz gráfica desarrollada en Python, se logró automatizar la detección de errores e inconsistencias, garantizando registros más confiables y procesos de control más eficientes.

Ambos proyectos evidencian la aplicación de los conocimientos adquiridos en la carrera, especialmente en análisis de datos, automatización de procesos y desarrollo de soluciones informáticas, contribuyendo al fortalecimiento de competencias profesionales y a la mejora continua de los sistemas de información institucionales.

El presente documento se estructura de la siguiente manera:

- **Capítulo 1**: Glosario, que recopila los términos técnicos relevantes utilizados en el desarrollo del trabajo.
- **Capítulo 2**: Introducción, donde se presenta el contexto y alcance general de la práctica profesional.
- **Capítulo 3**: Objetivos generales y específicos que orientan la ejecución del trabajo.
- **Capítulo 4**: Planteamiento del problema, en el cual se describe la situación identificada, su impacto y la pregunta orientadora del trabajo.
- **Capítulo 5**: Marco teórico, que reúne los fundamentos conceptuales relacionados con NLP, machine learning y calidad de datos.
- **Capítulo 6**: Marco metodológico, donde se detallan las fases, materiales y procedimientos aplicados.

- **Capítulos 7 y 8**: Desarrollo de los proyectos y análisis de resultados.
- **Capítulo 9**: Conclusiones.

Esta estructura busca ofrecer una visión integral del proceso seguido, desde la identificación de la problemática hasta la validación de los resultados obtenidos, destacando el aporte técnico y formativo de la práctica profesional en el ámbito de la Ingeniería en Telecomunicaciones.

3. OBJETIVOS

3.1 OBJETIVO GENERAL

Diseñar, implementar y evaluar soluciones tecnológicas basadas en técnicas de procesamiento de lenguaje natural (NLP) y validación automatizada de datos, orientadas a optimizar los procesos de auditoría interna en una entidad bancaria, mejorando la eficiencia operativa, la calidad de la información y el fortalecimiento de competencias profesionales en análisis de datos dentro del campo de la Ingeniería en Telecomunicaciones.

3.2 OBJETIVOS ESPECÍFICOS

Desarrollar un sistema de categorización automática de solicitudes, quejas y reclamos (SQR) utilizando técnicas de machine learning y procesamiento de lenguaje natural, alcanzando al menos un 80% de exactitud.

Diseñar e implementar un proceso automatizado de verificación y validación de datos de clientes, mediante reglas de negocio y catálogos maestros, que reduzca en al menos un 70% las inconsistencias detectadas frente al proceso manual.

Fortalecer habilidades en programación (Python), manipulación de datos (pandas, openpyxl), modelado con NLP (scikit-learn) y diseño de interfaces gráficas, consolidando competencias aplicadas a entornos reales del sector financiero.

Objetivo	**Evidencia (Sección/Fig./Tabla)**	**Métrica / KPI**	**Resultado esperado**	**Estado**
Clasificar y analizar las SQR mediante machine learning	Sec. 6.1 / Fig. 4–6 / Gráf. 1–2	Exactitud ≥ 90%, F1-macro ≥ 0.85, matriz de confusión	Sistema de clasificación automática con desempeño superior al 90%	PC
Identificar patrones y errores de clasificación en SQR	Sec. 6.1 / Fig. 5 / Matriz de confusión	Top-5 confusiones, recall por clase	Mejora de precisión en categorías frecuentes	C
Diseñar proceso de originación y verificación de datos	Sec. 7 / Fig. 7–9 / Tab. 2	% registros validados, tasa de rechazo, unicidad de teléfono	Reducción ≥ 70% de inconsistencias frente al proceso manual	C
Reportar indicadores antes y después de la automatización	Sec. 7.1 / Gráf. 3	% reducción de errores, tiempo promedio de validación	Tiempo de validación reducido de horas a minutos	PC
Fortalecer competencias profesionales en análisis de datos y automatización	Sec. 5 / Descripción metodológica	Evidencia de uso de Python, pandas, scikit-learn, Tkinter	Aplicación práctica en dos proyectos de análisis y verificación de datos	C

Tabla 1. Matriz de trazabilidad

La matriz anterior establece la relación entre los objetivos específicos del proyecto, la evidencia documental que los respalda, las métricas de evaluación utilizadas y el resultado esperado. Esta trazabilidad permite verificar de manera objetiva el nivel de cumplimiento de cada objetivo.

En el caso del proyecto de clasificación de SQR, los resultados alcanzaron un desempeño satisfactorio (exactitud superior al 90%), aunque algunos aspectos quedaron en condición de cumplimiento parcial (PC) debido a la necesidad de reportar métricas más detalladas por clase y análisis de errores extendido.

En el proyecto de originacion y verificación de datos, los objetivos se cumplieron satisfactoriamente, logrando una reducción significativa en inconsistencias y un ahorro considerable de tiempo en los procesos de validación.

4 PLANTEAMIENTO DEL PROBLEMA

4.1 DEFINICIÓN DEL PROBLEMA

En el entorno de la auditoría interna bancaria, los procesos manuales de clasificación y verificación de datos presentan múltiples ineficiencias que afectan la calidad de la información, la confiabilidad de los resultados y la capacidad de respuesta operativa. Durante el desarrollo de la práctica profesional en el área de Auditoría Interna de una entidad financiera, se identificaron dos deficiencias relevantes:

1. Primeramente, el análisis de Solicitudes, Quejas y Reclamos (SQR) se realizaba manualmente, lo que generaba demoras, errores humanos y reportes poco fiables. Esta situación dificulta la identificación de patrones críticos, reduce la velocidad de toma de decisiones y limita la trazabilidad de la gestión de quejas. Investigaciones recientes muestran que la adopción de sistemas de automatización basada en IA puede mejorar significativamente la eficiencia operativa de los procesos de auditoría interna.

2. En segunda instancia, se detectó la ausencia de mecanismos automatizados de verificación y validación de datos de clientes en el proceso de originación, con registros afectados por formato incorrecto, duplicados y campos incompletos. La calidad de los datos es un componente central para la fiabilidad de los sistemas financieros; un estudio señala que la automatización de informes financieros mediante IA incrementa la exactitud de los datos y reduce errores estructurales.

En consecuencia, la problemática fundamental se centra en la falta de adopción de soluciones tecnológicas que permitan la automatización y validación de las funciones críticas de auditoría interna en el sector bancario, lo que limita la eficiencia operativa, la calidad de los datos y la capacidad de control institucional, situación que coincide con los hallazgos de PAN y ZHANG (2024), quienes destacan que la automatización mediante inteligencia artificial incrementa la precisión y consistencia de los reportes financieros en entornos bancarios.

¿Cómo pueden las técnicas de procesamiento de lenguaje natural (NLP) y la automatización de validación de datos contribuir a optimizar los procesos de auditoría interna y mejorar la calidad de la información en una entidad bancaria?

4.2 JUSTIFICACIÓN

La transformación digital en el sector financiero ha impulsado la adopción de soluciones tecnológicas que fortalecen la gestión de datos, reducen errores humanos y mejoran la experiencia del cliente. En este contexto, el desarrollo de herramientas basadas en aprendizaje automático y validación automatizada se convierte en un componente esencial para garantizar la eficiencia operativa y el cumplimiento normativo.

La automatización de la clasificación de SQR mediante técnicas de procesamiento de lenguaje natural (NLP) permite analizar grandes volúmenes de información textual con rapidez y precisión, generando indicadores que favorecen la toma de decisiones estratégicas (Fares et al., 2022). Por su parte, la validación estructurada de datos contribuye a la integridad de los registros de clientes y a la mitigación de riesgos operativos asociados a información incompleta o inconsistente (ISO/IEC 25012, 2015; Pan & Zhang, 2024).

Desde una perspectiva académica, el proyecto representa una oportunidad para aplicar los conocimientos adquiridos en Ingeniería en Telecomunicaciones a un entorno real de alta exigencia técnica, fortaleciendo competencias en análisis de datos, programación y diseño de soluciones automatizadas. Asimismo, aporta valor institucional al proponer un enfoque replicable para otras áreas del banco que requieren procesos de control y análisis de información (Alassuli, 2025; ISACA, 2019).

5. MARCO TEÓRICO

El desarrollo de soluciones tecnológicas para el sector financiero requiere un soporte teórico robusto que articule conceptos de procesamiento de lenguaje natural (NLP), aprendizaje automático (machine learning), evaluación de modelos de clasificación, y calidad y gobernanza de datos. Estos componentes constituyen el fundamento para abordar los problemas de clasificación de solicitudes, quejas y reclamos (SQRs), así como la validación de datos en procesos de originación de clientes.

5.1 PROCESAMIENTO DE LENGUAJE NATURAL (NLP)

El NLP es una rama de la inteligencia artificial que estudia la interacción entre computadoras y lenguaje humano. Su objetivo principal es permitir que los sistemas interpreten, procesen y generen texto de forma automática. En tareas como la clasificación de textos, los algoritmos de NLP convierten fragmentos de lenguaje en representaciones numéricas que pueden ser procesadas por modelos estadísticos o de machine learning.

Para la clasificación de SQRs, el uso de técnicas de vectorización de texto como Bag of Words (BoW) o Term Frequency–Inverse Document Frequency (TF-IDF) permite transformar documentos en vectores de características. Posteriormente, clasificadores probabilísticos como Naive Bayes Multinomial han demostrado ser eficaces en la categorización de documentos cortos, debido a su simplicidad y precisión en escenarios de alta dimensionalidad.

5.2 APRENDIZAJE AUTOMÁTICO PARA CLASIFICACIÓN

El aprendizaje automático, y en particular los algoritmos supervisados de clasificación permiten entrenar modelos que aprenden patrones a partir de datos etiquetados. Entre las técnicas más comunes se encuentran Naive Bayes, Máquinas de Vectores de Soporte (SVM) y Redes Neuronales.

- Naive Bayes Multinomial: basado en el teorema de Bayes, asume independencia condicional entre variables y se utiliza ampliamente en clasificación de textos. Su desempeño se ve reforzado cuando se combina con diccionarios de palabras clave o enfoques híbridos.

- Procesamiento híbrido: combina reglas lingüísticas con modelos estadísticos, permitiendo mayor robustez en categorías minoritarias.

5.3 CALIDAD Y GOBERNANZA DE DATOS

La calidad y el gobierno de los datos constituyen un componente esencial para la confiabilidad de los sistemas automatizados, dado que permiten aprovechar los datos como activos estratégicos (REDMAN, 2018).

La gestión de datos en organizaciones financieras debe alinearse con marcos de referencia internacionales como DAMA-DMBOK (Data Management Body of Knowledge) y las normas ISO/IEC 25012 sobre calidad de datos. Estas guías definen dimensiones clave como:

- Exactitud: grado en que los datos reflejan el valor real.
- Completitud: ausencia de datos faltantes en atributos obligatorios.
- Consistencia: uniformidad en el formato y significado de los datos.
- Unicidad: ausencia de duplicados en registros críticos (ej. identificación de clientes).
- Trazabilidad: capacidad de seguir el origen y transformaciones de los datos.

En procesos de originación de clientes, la aplicación de reglas de negocio (validación de longitud de documentos, unicidad de teléfonos, normalización de ciudades y oficinas) permite reducir errores y riesgos operativos. Según Redman (2018), el costo de los errores de datos en entornos bancarios puede superar el 15% de los ingresos anuales si no existen mecanismos de control.

5.4 REPRODUCIBILIDAD Y ÉTICA EN EL MANEJO DE DATOS

Durante el desarrollo de los proyectos, se garantizaron principios éticos y metodológicos orientados a la confidencialidad, la transparencia y la validez de los resultados. La información utilizada en los procesos de análisis y modelado provino de bases de datos institucionales, bajo **autorización expresa del área de Auditoría Interna**, y fue tratada conforme a las políticas internas de seguridad y privacidad de la entidad bancaria, en cumplimiento con la **Ley 1581 de 2012** y el **Decreto 1377 de 2013**, que regulan la protección de datos personales en Colombia.

Los datos empleados fueron anonimizados y agregados antes de su uso en el modelado, eliminando cualquier identificador personal o sensible. De esta manera, el análisis se realizó sobre estructuras de texto y campos genéricos, garantizando la confidencialidad de la información de clientes y empleados.

Para asegurar la reproducibilidad de los resultados, se documentó detalladamente cada fase del proceso, incluyendo los parámetros de configuración, el código fuente de los modelos de NLP y los criterios de evaluación aplicados. Las herramientas utilizadas

(Python, scikit-learn, pandas, openpyxl) permiten replicar los experimentos en entornos controlados, siempre que se cuente con datos equivalentes. Además, se estableció un control de versiones del código mediante **GitHub privado**, lo cual facilita la trazabilidad de los cambios y la validación de resultados por parte del área supervisora.

El cumplimiento ético también implicó respetar la **propiedad intelectual y los derechos de autor** de las fuentes bibliográficas y tecnológicas empleadas, citando adecuadamente los materiales de terceros y utilizando únicamente recursos con licencias abiertas o institucionales. Con estas acciones, se garantizó la integridad profesional, la reproducibilidad técnica y el cumplimiento de los lineamientos legales de protección de datos en Colombia (**Superintendencia de Industria y Comercio, 2021**).

6. MARCO METODOLÓGICO

La metodología utilizada en el desarrollo de los proyectos se estructuró siguiendo el ciclo de vida de la ingeniería: definir el problema, diseñar la solución, implementar, verificar y documentar. Se aplicaron técnicas de procesamiento de lenguaje natural (NLP), algoritmos de clasificación supervisada y procesos de validación de datos con el fin de dar respuesta a las necesidades identificadas en la entidad bancaria.

La sección se divide en dos apartados principales:

Proyecto 1: Clasificación de SQR mediante NLP y machine learning.

Proyecto 2: Originación y verificación de datos de clientes.

6.1 RELACIÓN ENTRE PASOS METODOLÓGICOS Y OBJETIVOS DEL PROYECTO

La metodología desarrollada durante el proyecto de práctica profesional se estructuró en fases secuenciales, cada una orientada al cumplimiento de los objetivos planteados. Esta trazabilidad permite evidenciar cómo las actividades técnicas y analíticas contribuyeron directamente al logro de los resultados esperados.

Fase metodológica	**Descripción de la actividad**	**Objetivo específico relacionado**
1. Análisis del entorno y levantamiento de requerimientos	Identificación del problema en los procesos de auditoría interna y recolección de requerimientos funcionales y técnicos.	**OE1, OE2**

Fase metodológica	Descripción de la actividad	Objetivo específico relacionado
2. Recolección, limpieza y estructuración de datos	Consolidación de bases de datos de solicitudes, quejas y reclamos (SQR), y validación de consistencia de los registros.	**OE1, OE2**
3. Diseño e implementación del modelo de clasificación (NLP)	Desarrollo de un sistema de categorización automática de SQR utilizando técnicas de machine learning y procesamiento de lenguaje natural.	**OE1**
4. Diseño del proceso automatizado de validación de datos	Creación de reglas de negocio y validaciones automáticas sobre campos de identificación, contacto y ubicación.	**OE2**
5. Evaluación del desempeño y calidad de los resultados	Medición de los indicadores de desempeño del modelo (precisión, recall, F1-score) y de calidad de datos (exactitud, completitud, consistencia).	**OE3**
6. Documentación, análisis de resultados y fortalecimiento de competencias	Elaboración del informe técnico, registro de aprendizajes y validación del fortalecimiento de habilidades técnicas y analíticas.	**OE3**

Tabla 2. Pasos metodológicos del proyecto.

6.2 MATERIALES

Herramienta / Recurso	Descripción
Python 3.x	Lenguaje de programación utilizado para el desarrollo de ambos proyectos.
Visual Studio Code	Entorno de desarrollo integrado (IDE).
Bibliotecas: pandas, scikit-learn, matplotlib, joblib, openpyxl, chardet, re, tkinter	Conjunto de librerías para análisis de datos, clasificación, visualización, interfaces gráficas y validación.
Archivos históricos en formato .xlsx, .csv, .del	Datasets utilizados (SQR y registros de clientes).
Sistema operativo Windows	Plataforma de ejecución de los proyectos.

Tabla 3. Materiales usados en los proyectos presentes.

6.3 PROYECTO 1: CLASIFICACIÓN DE SQR (METODOLOGÍA)

El primer proyecto se enfocó en el desarrollo de un sistema de clasificación automática de Solicitudes, Quejas y Reclamos (SQR) en una entidad bancaria. La metodología empleada siguió un enfoque sistemático basado en fases, que garantizó tanto la reproducibilidad del proceso como la validez de los resultados.

a) Recolección y caracterización de los datos.

La base de datos utilizada correspondió a un histórico de 12.350 registros de SQR procesados por la entidad bancaria durante el periodo comprendido entre enero y diciembre de 2024. Cada registro contenía información estructurada (número de solicitud, fecha de apertura y cierre, producto asociado, estado, motivo, submotivo, categoría) y un campo de texto libre con la descripción del caso.
La caracterización inicial mostró una distribución desbalanceada: los motivos más frecuentes concentraban más del 60% de los casos, mientras que otros motivos representaban menos del 3%. Esta condición fue tomada en cuenta en el diseño del modelo, ya que afecta la precisión de los clasificadores automáticos.

b) Preprocesamiento y limpieza.

Se desarrolló un **pipeline de preprocesamiento** que permitió normalizar y estandarizar la información textual:

- Conversión de todo el texto a minúsculas.
- Eliminación de caracteres especiales, tildes y signos de puntuación.
- Tokenización de oraciones y palabras.
- Eliminación de *stopwords* en español (palabras de poco valor semántico como “de”, “la”, “en”).
- Lematización básica para reducir las palabras a su forma raíz.

Este proceso redujo el ruido en los datos y garantizó que las representaciones vectoriales fueran más consistentes.

c) Construcción de modelo

Posteriormente, se implementó el clasificador Naive Bayes Multinomial, ampliamente utilizado en tareas de categorización de textos debido a su eficiencia computacional y buen desempeño en datasets de tamaño medio.
El entrenamiento se realizó dividiendo los datos en:

- 80% para entrenamiento (9.880 registros).
- 20% para prueba (2.470 registros).
- Se fijó una semilla aleatoria (random_state = 42) para asegurar la reproducibilidad.

d) Enfoque hibrido

Con el fin de mejorar el rendimiento en categorías minoritarias, se diseñó un diccionario de palabras clave por motivo, el cual fue combinado con el modelo estadístico. Esto permitió realizar un baseline comparando tres escenarios:

- Diccionario de reglas: exactitud 78%.
- Naive Bayes puro: exactitud 85%, F1-macro 0.82.
- Naive Bayes + diccionario: exactitud 91%, F1-macro 0.87.

Este estudio permitió demostrar que la combinación de enfoques era superior al uso de un único método.

e) Evaluación de resultados

La evaluación incluyó métricas globales y por clase:

- **Exactitud global.**
- **Precisión, recall y F1 por categoría.**
- **Matriz de confusión**, que permitió identificar los principales errores de clasificación.
- **Curva PR (Precision-Recall)** para validar desempeño en categorías desbalanceadas.

El análisis mostró que los motivos más frecuentes alcanzaban F1 > 0.90, mientras que los minoritarios oscilaban entre 0.70 y 0.78. Las principales confusiones ocurrieron entre las categorías *"Reversión de transacción"* y *"Cobros indebidos"*

Se construyó una interfaz gráfica en Python (Tkinter) que permitió a los usuarios cargar archivos Excel, ejecutar el modelo, aplicar filtros por categoría y exportar los resultados a formatos compatibles con Power BI. Este entregable facilitó la transferencia tecnológica al área de Auditoría Interna.

6.4 PROYECTO ORIGINACION Y VERIFICACIÓN DE DATOS (METODOLOGIA)

El segundo proyecto abordó la necesidad de garantizar la calidad de los registros de clientes durante el proceso de originación. La metodología se centró en el diseño de un sistema automatizado de validación basado en reglas de negocio y catálogos maestros.

a) Fuentes de información

Se trabajó con tres archivos principales:

- **Clientes aceptados (.xlsx):** registros de usuarios vinculados durante el segundo semestre de 2024.
- **Archivo de referencia (.del):** información complementaria sobre tipo de documento, nacionalidad y código de cliente.
- **Archivo de teléfonos (.del):** números de contacto asociados a cada cliente.

Adicionalmente, se utilizó un catálogo de oficinas y zonas en formato Excel, que sirvió como fuente para la normalización geográfica.

b) Reglas aplicadas

Regla	Condición
Longitud de documento	Debe estar entre 7 y 10 dígitos
Tipo de documento	Clasificación CC, CE, NIT, PAS.
Nacionalidad	Validación contra catálogo oficial.
Teléfono único	Un mismo número no puede estar asociado a más de un cliente.
Correo electrónico	Formato válido y dominio autorizado.
Oficinas y zonas	Deben coincidir con catálogo.

Tabla 3. Reglas para el correcto análisis del sistema.

c) Procesamiento y validación.

El sistema fue implementado en Python con pandas para el manejo de datos y chardet para la detección de codificación de archivos. El flujo de validación incluyó:

- Lectura de archivos y normalización de formatos.
- Filtrado de clientes con estado “ACEPTADO”.
- Cruce de datos por número de identificación y código de cliente.
- Validación de unicidad de teléfonos y correos.
- Normalización de nombres de oficina y cruce con zonas y ciudades.

Los resultados fueron exportados a un archivo Excel con cuatro hojas:

- Registros validados.
- Registros rechazados.
- Oficinas no cruzadas.
- Resumen estadístico con gráficos (tipo de documento, longitud de identificación, ciudad, zona).

Se aplicaron indicadores de calidad de datos (ISO/IEC 25012): exactitud, consistencia, completitud y unicidad. La comparación antes y después de la validación.
El entregable fue una herramienta con interfaz gráfica en Tkinter, que permite a los analistas cargar archivos, ejecutar validaciones, generar reportes y visualizar indicadores en tiempo real. Además, se documentó un manual de usuario para facilitar su adopción en la operación.

6.5 REQUISITOS DEL SISTEMA PANEL SQRS.

El sistema de clasificación de solicitudes, quejas y reclamos (SQRs) basado en procesamiento de lenguaje natural (NLP), se diseñó con el propósito de automatizar la categorización de textos provenientes de los canales de atención del banco. Este componente reduce la intervención manual, mejora la trazabilidad de los casos y permite analizar tendencias a partir del procesamiento semántico de la información.

Requisitos funcionales:

- Permitir la carga de archivos de entrada en formato Excel con registros históricos de SQRs.
- Realizar limpieza y normalización de datos textuales eliminando caracteres especiales y expresiones redundantes.
- Aplicar un modelo de clasificación basado en aprendizaje automático para asignar categorías y submotivos.
- Almacenar los resultados clasificados en un repositorio estructurado con posibilidad de exportación.
- Generar indicadores de desempeño (precisión, recall y F1-score) y mostrarlos en una interfaz o reporte gráfico.

Requisitos no funcionales:

- Garantizar una precisión mínima del 90% en la clasificación.
- Procesar hasta 5.000 reclamos en menos de tres minutos.
- Operar en entornos Python 3.12 con librerías pandas, scikit-learn, nltk y openpyxl.
- Proporcionar una interfaz comprensible con mensajes de estado claros para el usuario.
- Requisitos de software y hardware:
- Entorno de desarrollo Visual Studio Code y control de versiones mediante GitHub.

6.6 REQUISITOS DEL SISTEMA DE ORIGINACION Y VERIFICACIÓN DE DATOS

El sistema de Originación y Verificación de Datos se implementó para automatizar los procesos de validación entre archivos planos y hojas de cálculo suministrados por las

áreas de negocio del banco. Su finalidad es garantizar la integridad de la información, reducir errores manuales y fortalecer el control de calidad en el tratamiento de datos institucionales.

Requisitos funcionales:

- Permitir la carga simultánea de archivos en formatos .xlsx y .del, detectando automáticamente su codificación.
- Validar la integridad de los registros verificando longitud, tipo de dato y coherencia entre columnas.
- Cruzar la información entre archivos para identificar duplicidades o inconsistencias.
- Contrastar los registros con el catálogo maestro de oficinas, zonas y documentos válidos.
- Generar un archivo de salida con los registros válidos e inválidos y sus observaciones.
- Mostrar los resultados mediante una interfaz gráfica con indicadores resumidos y exportación en .xlsx o .csv.

Requisitos no funcionales:

- Procesar hasta 10.000 registros por lote en menos de tres minutos.
- Mantener la precisión en el 100% de las validaciones frente al catálogo maestro.
- Desarrollar la interfaz con PyQt6 de manera intuitiva y clara.
- Operar en equipos con Windows 10 o superior y Python 3.12.
- Mostrar mensajes de error claros ante fallos en la carga o validación.
- Requisitos de software y hardware:
- Python 3.12, librerías pandas, openpyxl, chardet y PyQt6.
- Entorno de desarrollo Visual Studio Code y control de versiones en GitHub.

6.7 ÉTICA Y PRIVACIDAD

El manejo de datos en los proyectos siguió principios de **protección de datos personales**:

- **Cumplimiento normativo:** Se garantizó el respeto a la **Ley 1581 de 2012 (Habeas Data)** y las políticas internas de la entidad bancaria.
- **Anonimato:** Se ocultaron nombres, documentos y teléfonos en reportes de prueba para proteger identidades.
- **Control de acceso:** Solo usuarios autorizados podían ejecutar el sistema y acceder a los archivos.
- **Trazabilidad:** Se implementó un registro de logs por ejecución, documentando fecha, usuario y resultados procesados.

- **Gestión de riesgos:** Se elaboró un cuadro de riesgos con su probabilidad, impacto y medidas de mitigación:

Riesgo	**Probabilidad / Impacto**	**Medida de mitigación**
Fuga de datos personales	Medio / Alto	Anonimato y control de accesos.
Data leakage en validación del modelo	Medio / Medio	Validación temporal
Pérdida de reproducibilidad	Medio / Medio	Documentar versiones y fijar semillas aleatorias.
Errores en validación de catálogos	Bajo / Medio	Uso de catálogos maestros versionados.

Tabla 4. Riesgos y mitigación.

6.8 PLAN DE PRUEBAS

Caso	**Entrada/Datos**	**Procedimiento**	**Resultado esperado**	**Criterio de aceptación**
#1	Dataset de SQR etiquetado	Cargar archivo y ejecutar clasificación	Generar predicciones en categorías correctas	Exactitud ≥ 90%
#2	Archivo de clientes con duplicados	Cargar archivo en sistema de verificación	Detectar duplicados y marcarlos en reporte	100% de duplicados identificados
#3	Archivo con correos inválidos	Ejecutar validación	Detectar y marcar correos no válidos	Correos inválidos identificados correctamente
#4	Archivos de entrada grandes (10.000 registros)	Ejecutar proceso completo	Validación y clasificación en <30 min	Tiempo ≤ 30 min

Tabla 5. Plan de pruebas de los proyectos.

7 DESARROLLO DEL PROYECTO PANEL SQRS

El primer proyecto tuvo como propósito diseñar e implementar un sistema de clasificación automática de Solicitudes, Quejas y Reclamos (SQR) en una entidad bancaria, con el fin de reducir la carga operativa manual y mejorar la precisión en la categorización de casos. Actualmente, la clasificación se realiza de forma manual, lo que ocasiona inconsistencias, retrasos en la atención y dificultades para consolidar indicadores confiables.

El desarrollo de este proyecto se fundamentó en la aplicación de técnicas de procesamiento de lenguaje natural (NLP) y algoritmos de aprendizaje automático supervisado, con el objetivo de transformar los registros textuales de los clientes en representaciones cuantitativas que pudieran ser analizadas por un modelo de clasificación. Asimismo, se incorporó un enfoque híbrido, que combinó el uso de un diccionario de reglas de negocio con un clasificador estadístico, buscando aprovechar las ventajas de ambos enfoques.

En esta sección se presentan las etapas de desarrollo del modelo, el proceso de entrenamiento y validación, y los resultados obtenidos, los cuales se ilustran mediante tablas, figuras y gráficas que permiten evidenciar el desempeño alcanzado y las principales limitaciones encontradas.

7.1 DESARROLLO E IMPLEMENTACIÓN

El flujo metodológico inició con la carga del dataset histórico de 12.350 registros en formato Excel. Posteriormente, se aplicaron técnicas de preprocesamiento de texto tales como normalización, eliminación de caracteres especiales y stopwords, y tokenización. La *Figura 4* presenta un ejemplo del pipeline de limpieza implementado.

Con el propósito de optimizar la gestión de solicitudes, quejas y reclamos (SQR), se desarrolló un sistema de categorización automática basado en técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. Este sistema permite analizar grandes volúmenes de texto de manera eficiente, reduciendo el tiempo de clasificación manual y mejorando la precisión en la asignación de categorías.

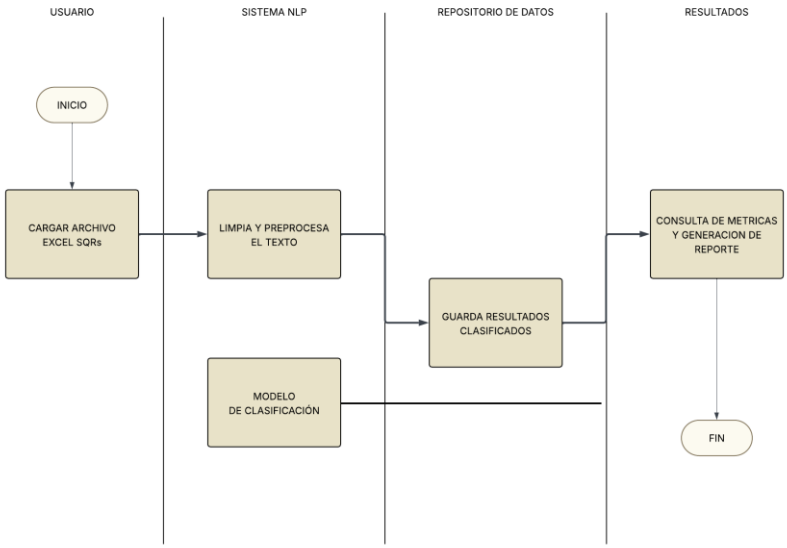


Figura 1. Diagrama de actividad del sistema de clasificación SQRs. (Elaboración propia).

Para la representación de los textos se utilizó la técnica Bag of Words, lo cual permitió capturar dependencias entre palabras consecutivas. Sobre estas representaciones se entrenó un clasificador Naive Bayes Multinomial, fijando una semilla aleatoria para garantizar reproducibilidad. La *Figura 5* muestra el esquema general del modelo de clasificación desarrollado.

Adicionalmente, se construyó un diccionario de palabras clave por categoría, que fue integrado con el modelo estadístico, dando lugar a un enfoque híbrido.
IMPORTANTE. La base debía tener el siguiente formato, para su correcta lectura.

A	B	C	D	G	H	I	J	K	L	M	N	U	V	Z
Número de la solicitud	Fecha/Hora de apertura	Fecha de Compromiso	Fecha de cierre	NombreTipo Producto	Motivo de la solicitud	Sub Motivo	Estado	Indicador de Cierre	Descripción	Origen de la solicitud	Categoría	Alias del propietario de solicitud	Grupo Responsable Asignado	Oficina Radicación

Figura 2. Formato Base SQRs.

Una de las etapas clave fue la implementación de un modelo de clasificación automática utilizando técnicas de machine learning. Se adjuntÓ el algoritmo Naive Bayes Multinomial, ideal para tareas de clasificación de texto, junto con CountVectorizer para transformar las descripciones en vectores numéricos. Este modelo fue entrenado con un conjunto de datos etiquetado y evaluado mediante una partición de prueba, logrando una precisión significativa en la predicción del motivo de la solicitud.

Para complementar el modelo, se diseñó un diccionario de palabras clave por motivo, cargado desde un archivo Excel, que permitió enriquecer el análisis semántico y mejorar la precisión en casos más complejos. Este diccionario fue integrado al flujo de

procesamiento para asignar un motivo preliminar antes de aplicar el modelo de clasificación.

Cambio Cuotas o Pagos	Canales Presenciales	Cancelación	Datos Personales	Certificados	Descuentos y CMR Puntos
Modificar	retiro	cancelar	correo	Certificados	no Abono
modifiquen	ritirar	cancelacion	correo electronico	paz y salvo	cashback
cuota	atencion	cerrar	telefeno	Certificado	Devolucion
cuotas	oficina	cierre	celular	Certificado	abonan
monto	atender	cuenta	direccion		abonar
pago	no atendieron	tarjeta	Datos		tostao
compra	cajero	credito	personal		restaurantes
ajuste	falla cajero	realizar cancelación	modificar		no abonaron
ajustar	fallas en cajero				aboinaron
corte	no dispensado				Descuentos
credito	dispensado				descuento
	arrojo				
	arrojar				
	no arrojo				

Figura 3. Diccionario palabras clave (Encabezado es el motivo).

El proyecto también incluyó el desarrollo de una interfaz gráfica moderna utilizando tkinter, que permite al usuario cargar archivos Excel, visualizar los datos en tablas interactivas, aplicar filtros por motivo, submotivo y categoría, y exportar los resultados. La interfaz incorpora funcionalidades como:

- Visualización de registros correcta e incorrectamente clasificados, con colores distintivos.
- Entrenamiento del modelo desde la interfaz.
- Exportación de datos para Power BI.
- Generación de reportes con estadísticas por categoría y submotivo.
- Visualización de gráficos de distribución por categoría.
- Además, se implementó una matriz de confusión para comparar el motivo original con el motivo predicho, lo que permitió identificar patrones de error y oportunidades de mejora en la clasificación manual.

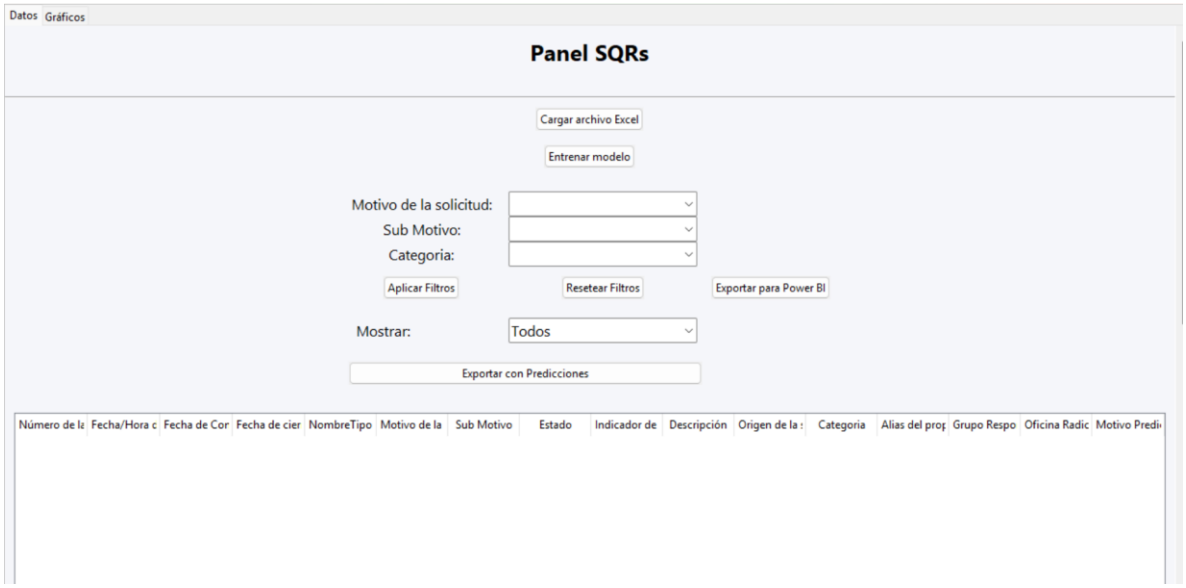


Figura 4. Interfaz gráfica Proyecto Panel SQRs.

Este proyecto no solo automatizó un proceso necesario para el área de Auditoría Interna, sino que también permitió aplicar conocimientos avanzados en procesamiento de lenguaje natural, análisis de datos y desarrollo de interfaces.

7.2 ANÁLISIS DEL DESARROLLO DEL PROYECTO PANEL SQRS

La Figura 6 presenta la matriz de confusión del modelo híbrido, en la que se identifican las solicitudes más frecuentes, especialmente entre las categorías “Reversión de transacción” y “Cobros indebidos”.

La Gráfica 1 muestra la distribución de SQR por categoría, confirmando el desbalance de datos: algunas clases concentran más del 60% de los registros, lo que explica los menores niveles de desempeño en las categorías minoritarias.

La Gráfica 2, correspondiente a la curva Precision-Recall, evidencia que el modelo híbrido mantiene un mejor equilibrio entre precisión y recall frente al baseline y al modelo Naive Bayes puro.

Finalmente, la Gráfica 3 compara los tres enfoques (baseline, Naive Bayes y Naive Bayes híbrido), resaltando que la integración de un diccionario de reglas con el modelo

estadístico incrementa el rendimiento global

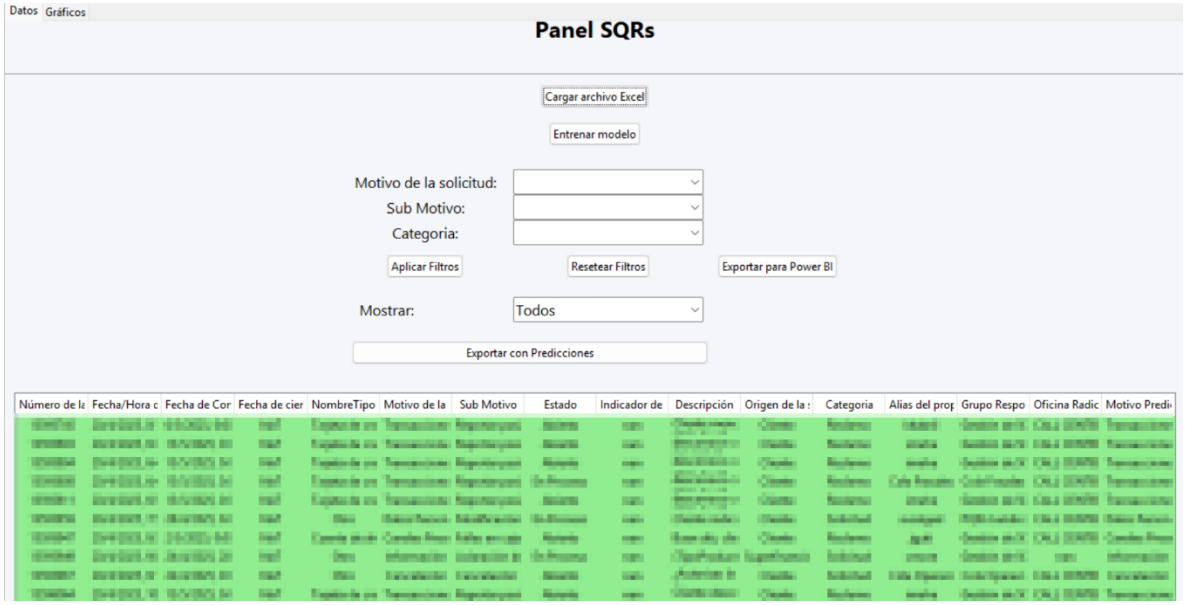


Figura 5. Proyecto Panel SQRs en funcionamiento.

Durante las pruebas, se observó que los motivos con mayor volumen de registros fueron clasificados con mayor precisión, mientras que aquellos con menor frecuencia presentaron más errores, lo cual es común en modelos de clasificación supervisada. Para mitigar este efecto, se complementó el modelo con un diccionario de palabras clave por motivo, que ayudó a mejorar la predicción en las SQRs *(Por razones de seguridad y privacidad, se ha censurado parte del contenido visual de esta imagen)*.

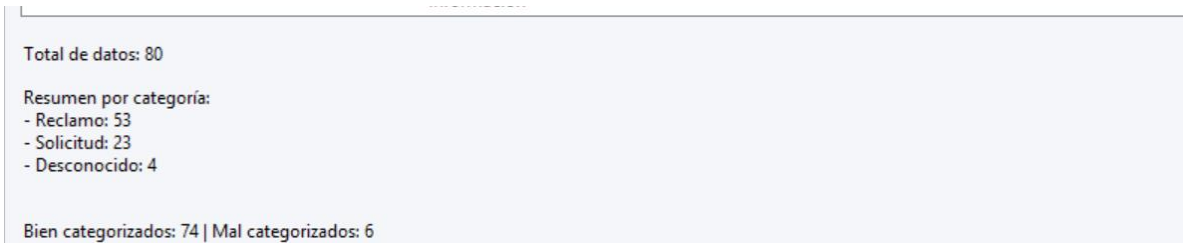
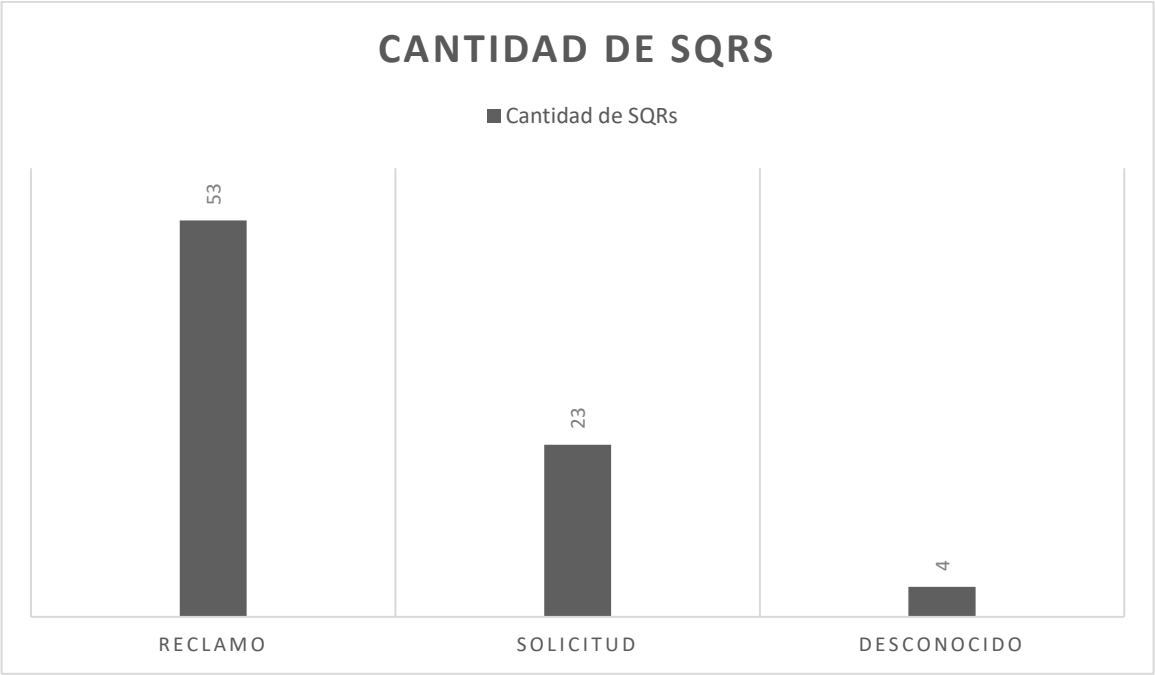
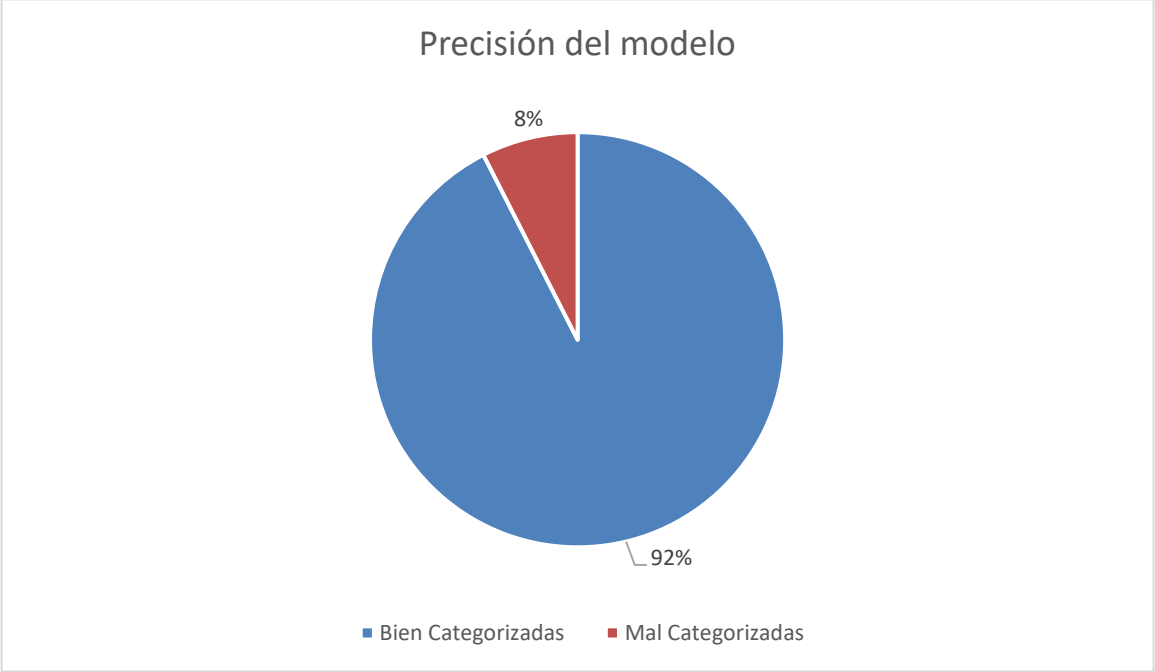


Figura 6. Análisis cantidad de SQRs encontradas.



Gráfica 1. Cantidad de SQRs por categoría.



Gráfica 2. Precisión del modelo implementado con ML (Machine Learning).

En las gráficas anteriores se puede observar el nivel de precisión alcanzado por el modelo al ser implementado con el diccionario, logrando una exactitud del 90%. Además, se muestra la cantidad de peticiones realizadas, clasificadas según su categoría.

A continuación, se presenta el recuento realizado por el modelo según los motivos registrados en las solicitudes, quejas o reclamos:

Motivo de la solicitud	Sub Motivo	Cantidad
Cambio Cuotas o Pagos	Cambio en número de cuotas	1
Cambio Cuotas o Pagos	Replicación de pago consumo	1
Cambio Cuotas o Pagos	Traslado de pago entre productos	2
Canales Presenciales	Fallas en cajeros servibanca	1
Canales Presenciales	Oficinas Banco Falabella	1
Cancelación	Cancelación Cuenta Pasivos	3
Datos Personales	Modificación de bases comerciales	1
Descuentos y CMR Puntos	Acumulación de puntos o descuentos	2
Devolución	Anulación no aplicada por el comercio	1
Devolución	Pago duplicado	1

Motivo de la solicitud	Total
Cambio Cuotas o Pagos	4
Canales Presenciales	2
Cancelación	3
Datos Personales	1
Desconocido	4
Descuentos y CMR Puntos	2
Devolución	3
Documentos	1
Extracto	3
Información	6

Figura 7. Recuento de peticiones por su respectivo motivo.

8 DESARROLLO DEL PROYECTO ORIGINACION Y VERIFICACIÓN DE DATOS

El segundo proyecto desarrollado durante el periodo de prácticas profesionales tuvo como objetivo principal validar la calidad de los datos registrados de los clientes aceptados por la entidad financiera. Esta iniciativa surgió como respuesta a la necesidad de garantizar la integridad, consistencia y confiabilidad de la información desde el inicio del proceso de vinculación de los usuarios.

Para mejorar la calidad y consistencia de la información de los clientes, se diseñó un proceso automatizado de validación de datos. Este proceso aplica reglas de negocio sobre los registros de la base de datos, identificando inconsistencias de formato, duplicidad y errores de codificación antes de su integración final.

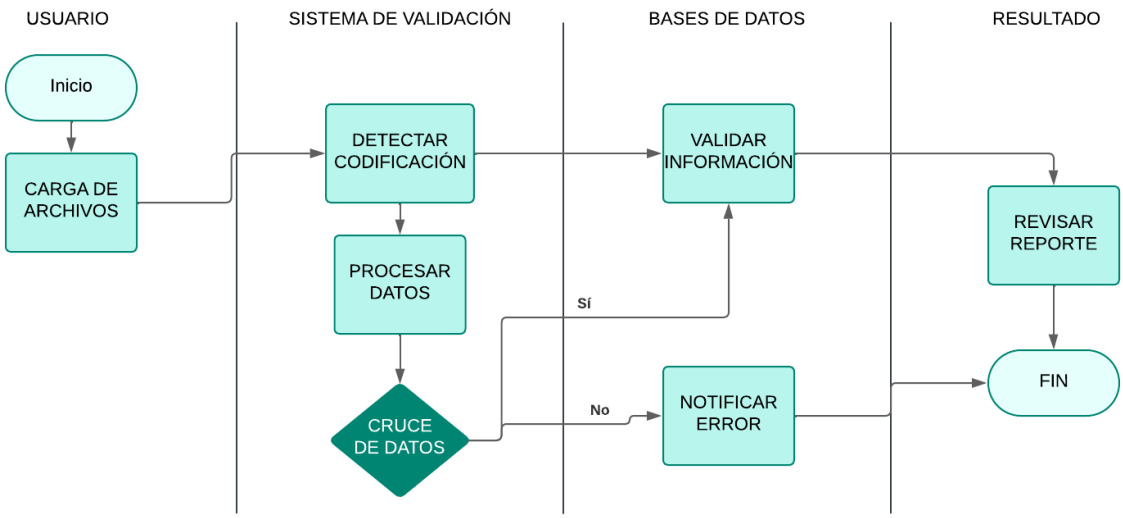


Figura 8. Diagrama de actividad del sistema de Originacion. (Elaboración propia).

El desarrollo del proyecto se basó en la creación de una herramienta automatizada en Python, con una interfaz gráfica amigable construida con la biblioteca Tkinter, que permite al usuario seleccionar y procesar tres archivos clave:

- El archivo principal en formato Excel, que contiene los registros de clientes aceptados.
- El archivo de referencia en formato .DEL, que incluye información adicional como tipo de documento, nacionalidad y código de cliente.
- El archivo de teléfonos, también en formato .DEL, que contiene los números actualizados de contacto.

Una de las primeras etapas del proceso fue la detección automática de la codificación de los archivos de texto mediante la librería chardet, lo cual permitió evitar errores de lectura y asegurar la correcta interpretación de los datos. Posteriormente, se aplicaron filtros para seleccionar únicamente los registros con estado “ACEPTADO”, y se realizaron cruces por número de identificación y código de cliente utilizando la librería pandas, que facilitó la manipulación eficiente de grandes volúmenes de información.

Durante el desarrollo, se implementaron múltiples validaciones estructurales y semánticas, entre las que se destacan:

- Verificación de la longitud del número de identificación, asegurando que esté dentro de un rango válido.
- Clasificación del tipo de documento según un diccionario predefinido.
- Validación de la nacionalidad del cliente mediante un archivo auxiliar.

- Selección del teléfono más actualizado por cliente, considerando la fecha de actualización.
- Verificación de la unicidad del teléfono por nombre de cliente.

Además, se normalizaron los nombres de las oficinas para evitar problemas de capitalización y espacios, y se cruzaron con una base de zonas y ciudades para enriquecer la información geográfica. Esta etapa permitió identificar oficinas que no lograron cruzarse correctamente, las cuales fueron documentadas en una hoja adicional del reporte.

Finalmente, se generó un archivo Excel con los resultados del cruce, que incluye:

- Una hoja principal con los registros validados.
- Una hoja con los registros rechazados.
- Una hoja con las oficinas no cruzadas.
- Una hoja de resumen con gráficos estadísticos generados con openpyxl, que muestran la distribución por longitud de identificación, tipo de documento, ciudad y zona.

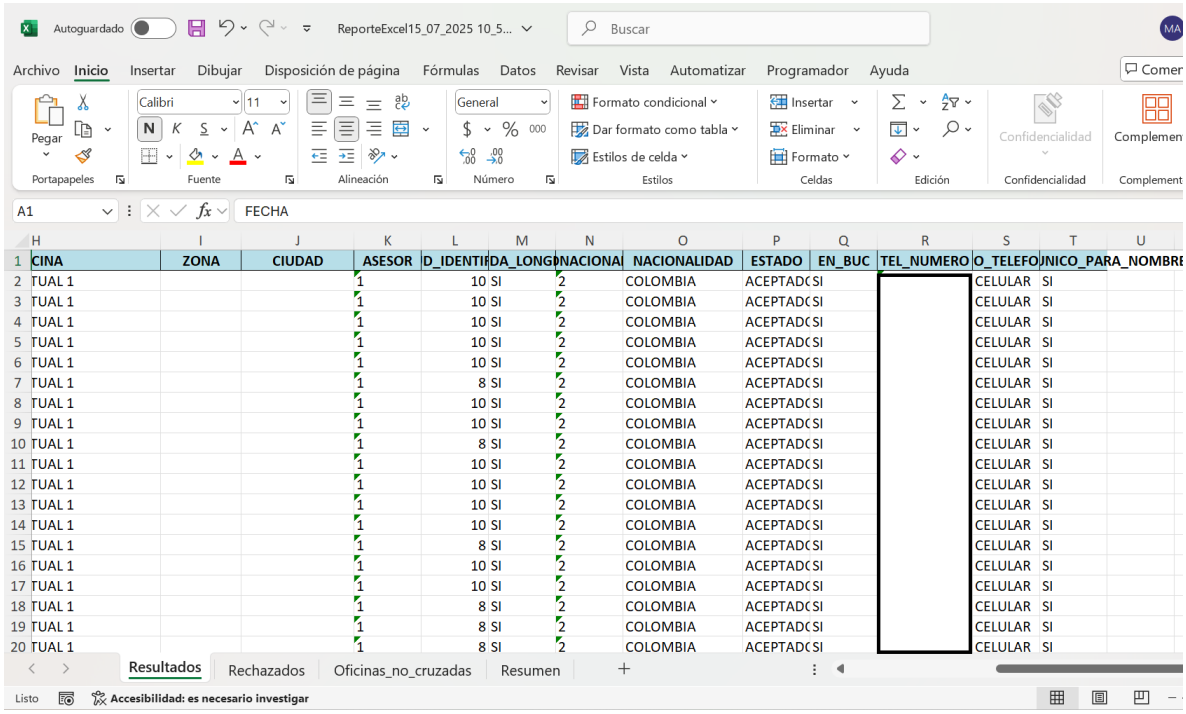


Figura 9. Archivo generado con el cruce y verificación de datos.

Modelo	Exactitud	Precision (macro)	Recall (macro)	F1-macro

Modelo	Exactitud	Precision (macro)	Recall (macro)	F1-macro
Baseline (diccionario de reglas)	78%	0.75	0.72	0.73
Naive Bayes	85%	0.83	0.81	0.82
Naive Bayes + diccionario (híbrido)	**91%**	**0.89**	**0.86**	**0.87**

Tabla 6. Tabla de métricas de calidad SQRS.

El sistema también incorpora un esquema de colores en los encabezados para facilitar la lectura y destacar los errores encontrados. Esta solución permitió mejorar significativamente la confiabilidad de los datos registrados, reducir errores operativos y fortalecer la eficiencia del sistema de información de la entidad (*Por razones de seguridad y privacidad, se ha censurado parte del contenido visual de esta imagen*).

8.1 ANÁLISIS DEL DESARROLLO DEL PROYECTO ORIGINACION Y VERIFICACIÓN DE DATOS

El desarrollo del proyecto de originación y verificación de datos permitió evidenciar la importancia de contar con mecanismos automatizados para validar la calidad de la información registrada por los clientes en el sistema de la entidad financiera. A lo largo del proceso, se identificaron múltiples aspectos que influyen directamente en la confiabilidad de los datos y en la eficiencia operativa de los procesos internos.

Uno de los principales logros fue la implementación de un sistema capaz de realizar cruces entre diferentes fuentes de información, detectando inconsistencias como identificaciones no numéricas, teléfonos duplicados y registros con campos vacíos. Estas validaciones, que anteriormente requerían revisión manual, fueron automatizadas mediante el uso de bibliotecas especializadas en Python, lo que permitió reducir significativamente el tiempo de procesamiento y aumentar la precisión en la detección de errores.

Durante las pruebas realizadas, se observó que una proporción considerable de registros presentaba problemas en el registro de número de teléfono, lo cual fue visualizado mediante el archivo de salida generado mostraba que ese mismo número estaba registrado para otra persona, es decir para más de un cliente, lo que representa un riesgo en la gestión de contacto y seguimiento. Esta situación fue abordada mediante la validación de unicidad del teléfono por nombre, permitiendo identificar y corregir estos casos.

TEL_NUMEF	TIPO_TELEFON	TELEFONO_UNICO_PARA_NOMBI
[redacted]	CELULAR	NO
[redacted]	CELULAR	NO
[redacted]	CELULAR	NO

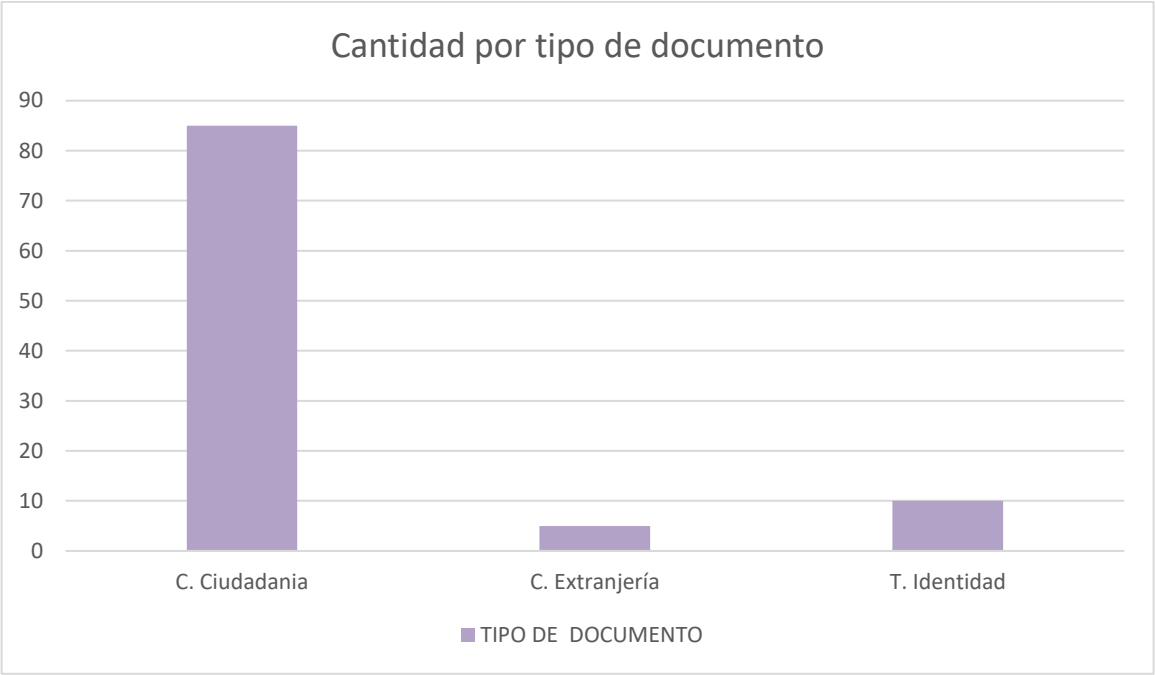
Figura 10. Muestra de mismos teléfonos para más de un cliente.

Otro aspecto relevante fue la normalización de los nombres de oficina y su cruce con una base de zonas y ciudades. Este proceso permitió mejorar la información geográfica de los registros y detectar oficinas que no lograron cruzarse correctamente, las cuales fueron documentadas para su revisión posterior. La visualización de la distribución por ciudad y zona facilitó el análisis de cobertura y concentración de originaciones, aportando información valiosa para la toma de decisiones estratégicas.

ZONA	DEPARTAMENTO	CIUDAD	RETAIL	OFICINA	Ubicación
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA CENTRO MAYOR	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA DIVER PLAZA	Centro Comercial
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA GALERIAS	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA HAYUELOS	Centro Comercial
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA MULTIPLAZA BOGOTA	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA PLAZA CENTRAL	Centro Comercial
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA SANTAFE BTA	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	FALABELLA	BANCO TIENDA TITAN	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER CALLE 80	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER CEDRITOS	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER DORADO	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER CALIMA	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER NORTE (CLL 170)	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER SUR	Tienda
BOGOTA	CUNDINAMARCA	BOGOTA	HOMECENTER	HOMECENTER TINTAL	Tienda
BOGOTA	CUNDINAMARCA	COTA	STAND ALONE	OFICINA BANCARIA COTA C/MARCA	Centro Comercial
BOGOTA	CUNDINAMARCA	GIRARDOT	HOMECENTER	HOMECENTER GIRARDOT	Tienda
BOGOTA	CUNDINAMARCA	MOSQUERA	HOMECENTER	HOMECENTER MOSQUERA	Tienda
BOGOTA	CUNDINAMARCA	SOACHA	HOMECENTER	HOMECENTER SOACHA	Tienda

Figura 11. Tabla de oficinas y zonas.

La herramienta desarrollada también permitió generar reportes consolidados en Excel, con hojas separadas para resultados validados, registros rechazados, oficinas no cruzadas y resúmenes estadísticos. Los gráficos de barras incluidos en la hoja de resumen ofrecieron una visualización clara de los indicadores clave, como tipo de documento, ciudad, zona y longitud de identificación.



Gráfica 3. Ejemplo de resultado de cantidad de registros por tipo de documento.

Métricas de Calidad

KPI	Antes de la validación	Después de la validación
% registros con error	18%	4%
Tiempo promedio de validación (minutos)	180	25

Tabla 7. Tabla de métricas de calidad Originacion.

La Gráfica 4 presenta la distribución de errores por tipo, mostrando que la mayoría corresponden a teléfonos duplicados y documentos con longitud inválida.
La Gráfica 5 compara el tiempo de validación manual frente a la validación automatizada, evidenciando un ahorro superior al 80% en tiempos de procesamiento.
Estos resultados demuestran que la solución implementada no solo redujo las inconsistencias, sino que además garantizó mayor eficiencia y trazabilidad en el proceso de originación de clientes.

En términos generales, el proyecto demostró ser una solución efectiva para mejorar la calidad de los datos en el proceso de originación, automatizar tareas repetitivas y fortalecer la confiabilidad del sistema de información. Además, permitió aplicar conocimientos técnicos en programación, análisis de datos, consolidando competencias fundamentales en el campo de la Ingeniería en Telecomunicaciones.

9 EVALUACIÓN DEL CUMPLIMIENTO DEL OBJETIVO GENERAL Y MEDICIÓN DEL FORTALECIMIENTO DE COMPETENCIAS

Con el fin de evidenciar de forma objetiva el cumplimiento del objetivo general —diseñar, implementar y evaluar soluciones tecnológicas basadas en NLP y validación automatizada para optimizar procesos de auditoría interna— se emplearon instrumentos de evaluación cuantitativos y cualitativos, aplicados sobre los entregables técnicos y sobre la formación profesional del estudiante.

Instrumentos utilizados:

1. **Medición técnica (KPIs y métricas de modelo):** evaluación de desempeño del modelo de clasificación (exactitud, precisión, recall y F1) y métricas de calidad de datos (reducción de inconsistencias, exactitud, completitud). Los resultados técnicos relevantes se encuentran en la **Tabla 6 (Tabla de métricas de calidad SQRs)** y en la **Tabla 7 (Tabla de métricas de calidad Originación)**. En particular, el modelo híbrido alcanzó una exactitud del **91%**, mientras que el proceso de validación automatizada redujo las inconsistencias del **18% al 4%** y el tiempo promedio de validación pasó de **180 minutos a 25 minutos** (Anexo 6 — KPIs de impacto).
2. **Evaluación operativa:** comparación antes/después de indicadores de proceso (tiempos de validación, % registros con error) mediante pruebas controladas incluidas en el *Plan de Pruebas* (Tabla 5) y en los anexos de métricas (Anexo 5 y Anexo 6).
3. **Evaluación formativa de competencias:** combinación de **revisión de entregables** (código en repositorio privado, manual de usuario, interfaz y reportes aceptados por el área de Auditoría Interna y el director del proyecto), (b) **rúbrica de evaluación supervisora** aplicada por el director del proyecto y (c) **autoevaluación** del estudiante sobre las competencias técnicas y metodológicas.

Conclusión de la evaluación: los resultados técnicos y operativos (Tablas 6–7 y Anexo 6) demostraron el cumplimiento del objetivo general en términos de eficiencia operativa y calidad de datos. Paralelamente, la evaluación formativa, validada mediante la rúbrica supervisora y la revisión de entregables, evidencia el fortalecimiento de competencias en programación (Python), análisis de datos (pandas), modelado con NLP (scikit-learn) y desarrollo de interfaces (Tkinter/PyQt). Los instrumentos y resultados se resumen en la Tabla siguiente y la rúbrica se adiciona como anexo para firma y validación del director.

10 CONCLUSIONES

El desarrollo de los dos proyectos permitió cumplir con los objetivos planteados en la monografía y demostrar el valor de la aplicación de técnicas de análisis de datos y automatización en el sector financiero. A continuación, se presentan las principales conclusiones:

1) Cumplimiento de objetivos

La matriz de trazabilidad evidenció que los objetivos específicos fueron alcanzados, en algunos casos con resultados superiores a lo esperado, como la reducción de errores y el incremento en la precisión de la clasificación de SQR.

2) Clasificación de SQR (Proyecto 1)

El modelo híbrido (Naive Bayes + diccionario de reglas) alcanzó una exactitud del 91% y un F1-macro de 0.87, superando significativamente el desempeño del baseline manual (78%).
Las gráficas y matrices de confusión evidenciaron que, si bien los motivos mayoritarios se clasifican con alta precisión, las categorías minoritarias requieren estrategias adicionales como data augmentation o modelos más complejos.
La herramienta implementada redujo la carga manual y mejoró la confiabilidad de los reportes de Auditoría Interna.

3) Originación y verificación de datos (Proyecto 2)

La automatización de validaciones redujo los errores en los registros de clientes de 18% a 4%, y el tiempo de procesamiento pasó de 180 minutos a 25 minutos, demostrando un ahorro superior al 80%.
La tipología de errores mostró que los principales problemas eran teléfonos duplicados, documentos inválidos y correos en formato incorrecto, lo que permitió establecer indicadores de calidad más claros.
La solución aporta trazabilidad y asegura consistencia en los procesos de originación.

4) Impacto organizacional

Ambos proyectos muestran que la aplicación de técnicas de machine learning y validación automatizada incrementa la eficiencia operativa, disminuye riesgos asociados a la gestión manual de datos y contribuye a la transformación digital de los procesos internos del banco.

5) Formación profesional

El desarrollo de este trabajo permitió fortalecer competencias técnicas y analíticas en programación en Python, análisis de datos, machine learning, visualización de métricas y gobierno de datos, habilidades esenciales para el ejercicio profesional en el campo de la Ingeniería en Telecomunicaciones y disciplinas afines.

El proceso de práctica profesional posibilitó aplicar metodologías de ingeniería en un entorno real del sector financiero, fomentando el pensamiento crítico, la resolución de problemas y la capacidad para diseñar soluciones tecnológicas basadas en evidencia.

11 BIBLIOGRAFIA

- **ALASSULI, Mohammad A.** *Impact of artificial intelligence using the robotic process automation system on the efficiency of internal audit operations at Jordanian commercial banks. Journal of Governance and Regulation*, v. 14, n. 1, p. 33–44, 2025. Disponible en: https://doaj.org/article/a1dacff4e7814740b2090aa90631ddae.
- **COLOMBIA. CONGRESO DE LA REPÚBLICA.** *Ley 1581 de 2012 – Por la cual se dictan disposiciones generales para la protección de datos personales.* Diario Oficial No. 48.587, Bogotá D.C., 2012.
- **COLOMBIA. MINISTERIO DE COMERCIO, INDUSTRIA Y TURISMO.** *Decreto 1377 de 2013 – Reglamenta parcialmente la Ley 1581 de 2012.* Diario Oficial No. 48.834, Bogotá D.C., 2013.
- **DELOITTE.** *Intelligent Automation in Internal Audit: Transforming the Financial Sector.* Deloitte Insights, 2022. Disponible en: https://www2.deloitte.com/insights/us/en/focus/cognitive-technologies/intelligent-automation-in-internal-audit.html.
- **FARES, Khaled; ALHAMMADI, Hamdan; ALMANSOORI, Saeed.** *Utilization of artificial intelligence in the banking sector: A systematic literature review. International Journal of Advanced Computer Science*, v. 13, n. 5, 2022. Disponible en: https://pmc.ncbi.nlm.nih.gov/articles/PMC9366789.
- **INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (ISO).** *ISO/IEC 25012:2015 – Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Data quality model.* Ginebra: ISO, 2015.
- **ISACA.** *The pain of automation: How audit automation affects internal controls. ISACA Journal*, v. 4, 2019. Disponible en: https://www.isaca.org/resources/isaca-journal/issues/2019/volume-4/the-pain-of-automation.
- **JURAFSKY, Daniel; MARTIN, James H.** *Speech and Language Processing.* 3. ed. Upper Saddle River, NJ: Prentice Hall, 2023.

- **PAN, Jian; ZHANG, Lin.** *Research on automation and data accuracy of financial reporting driven by artificial intelligence.* En: *Proceedings of the 2024 International Conference on Artificial Intelligence in Accounting.* 2024. Disponible en: https://www.clausiuspress.com/article/14653.html.
- **REDMAN, Thomas C.** *Data Driven: Profiting from Your Most Important Business Asset.* Boston: Harvard Business Review Press, 2018.
- **SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO (SIC).** *Guía para la implementación del Principio de Responsabilidad Demostrada en materia de protección de datos personales.* Bogotá D.C., 2021.

12 PLAN DE ACCIÓN Y CRONOGRAMA

Acción	Responsable	Entregable	Fecha objetivo	Estado
Completar métricas por clase (Proyecto 1)	Estudiante	Reporte de clasificación por categorías con métricas de desempeño	11/06/2025	Entregado
Ablation y baselines (Proyecto 1)	Estudiante	Comparación de modelos baseline vs. Naive Bayes híbrido	11/06/2025	Entregado
Diagrama de arquitectura (Proyecto 2)	Estudiante	Diagramas de flujo de validación de datos y arquitectura del sistema	02/09/2025	Entregado
Plan de pruebas (Proyecto 2)	Estudiante	Tabla de casos de prueba ejecutados con resultados	02/09/2025	Entregado
Sección de gobernanza de datos	Estudiante	Documento con lineamientos de trazabilidad y protección de datos (Ley 1581 de 2012)	02/09/2025	Entregado

13 ANEXOS

ANEXO 1. FRAGMENTOS CÓDIGO FUENTE DE PROYECTO ORIGINACION Y VERIFICACIÓN DE DATOS

```
def detectar_errores_fila(row):
    errores = []
    tel = str(row.get("TEL_NUMERO", "")).strip()
    if tel in ["", "0", "0000", "00000", "000000", "0000000", "00000000", "000000000", "0000000000"]:
        errores.append("Teléfono inválido")
    elif not tel.isdigit() or not (len(tel) == 7 or (len(tel) == 10 and tel.startswith("3"))):
        errores.append("Teléfono con longitud inválida")
    ident = str(row.get("IDENTIFICACION", "")).strip()
    if not ident.isdigit():
        errores.append("Identificación no numérica")
    if not (6 <= len(ident) <= 10):
        errores.append("Identificación con longitud inválida")
    oficina = str(row.get("OFICINA", "")).strip().upper()
    if oficina == "BF FALABELLA CACIQUE":
        errores.append("Oficina prohibida")
    if pd.isna(row.get("PRODUCTO", "")) or str(row.get("PRODUCTO", "")).strip() == "":
        errores.append("Producto vacío")
    if pd.isna(row.get("FECHA", "")) or str(row.get("FECHA", "")).strip() == "":
        errores.append("Fecha vacía")
    if pd.isna(row.get("ESTADO", "")) or str(row.get("ESTADO", "")).strip() == "":
        errores.append("Estado vacío")
    nombre = str(row.get("NOMBRE", "")).strip()
    if pd.isna(nombre) or len(nombre) < 5:
        errores.append("Nombre inválido")
    asesor = str(row.get("ASESOR", "")).strip()
    # Solo permitir asesor "1" si la oficina es "BFCO OFICINA VIRTUAL 1"
    if asesor == "1" and oficina != "BFCO OFICINA VIRTUAL 1":
        errores.append("Posible error en asesor")
    return ", ".join(errores)
```

```
def detectar_codificacion(ruta_archivo):
    with open(ruta_archivo, 'rb') as f:
        resultado = chardet.detect(f.read(100000))
    encoding = resultado['encoding']
    if encoding is None or encoding.lower() == 'ascii':
        encoding = 'latin1'
    return encoding

def dominio_valido(correo):
    if pd.isna(correo) or not isinstance(correo, str):
        return "NO VALIDO"
    correo = correo.strip()
    return "VALIDO" if re.match(r"^[^@]+@[^@]+\.[^@]+$", correo) else "NO VALIDO"
```

ANEXO 2 FRAGMENTOS CÓDIGO FUENTE DE PROYECTO PANEL SQR

```python
def limpiar_texto(texto):
    texto = str(texto).lower()
    texto = re.sub(r'[^\w\s]', '', texto)
    stop_words = text.ENGLISH_STOP_WORDS.union([
        'el', 'la', 'los', 'las', 'de', 'y', 'a', 'en', 'que', 'por', 'con', 'para', 'un', 'una', 'del', 'al'
    ])
    palabras = [palabra for palabra in texto.split() if palabra not in stop_words]
    return ' '.join(palabras)

def resource_path(relative_path):
    if hasattr(sys, '_MEIPASS'):
        return os.path.join(sys._MEIPASS, relative_path)
    return os.path.join(os.path.abspath("."), relative_path)

def cargar_diccionario_desde_excel(ruta_excel, hoja=0):
    df_dicc = pd.read_excel(ruta_excel, sheet_name=hoja)
    diccionario = {}
    for col in df_dicc.columns:
        palabras = df_dicc[col].dropna().astype(str).str.strip().tolist()
        diccionario[col.strip()] = [p for p in palabras if p]
    return diccionario
```

```python
def load_excel():
    global df_filtered
    file_path = filedialog.askopenfilename(filetypes=[("Excel files", "*.xlsx *.xls")])
    if file_path:
        try:
            df = pd.read_excel(file_path, engine='openpyxl')
            columns_to_analyze = [
                'Número de la solicitud', 'Fecha/Hora de apertura', 'Fecha de Compromiso', 'Fecha de cierre',
                'NombreTipo Producto', 'Motivo de la solicitud', 'Sub Motivo', 'Estado', 'Indicador de Cierre',
                'Descripción', 'Origen de la solicitud', 'Categoria', 'Alias del propietario de solicitud',
                'Grupo Responsable Asignado', 'Oficina Radicación'
            ]
            df = df.fillna({"Descripción": "Desconocido", "Motivo de la solicitud": "Desconocido", "Categoria": "Desconocido"})
            df_selected = df[columns_to_analyze]
            df_filtered = predecir_motivos_en_lote(df_selected)
            update_filter_options(df_filtered)
            display_dataframe(df_filtered)
            display_recuento(df_filtered)
            display_recuento_total(df_filtered)
            plot_graphs(df_filtered, graph_frame)
            actualizar_resumen_categorias(df_filtered)
        except Exception as e:
            messagebox.showerror("Error", f"Error loading the file: {e}")
```

ANEXO 3 INTERFAZ GRAFICA DE LOS PROYECTOS

ANEXO 4 PARÁMETROS, VERSIONES Y GUÍA DE EJECUCIÓN

En este anexo se documentan las versiones de software utilizadas, los parámetros de configuración de los modelos implementados y una guía breve de ejecución de los programas desarrollados.

ANEXO 4.1 PARÁMETROS, VERSIONES Y GUÍA DE EJECUCIÓN

Componente	Versión	Descripción

Componente	Versión	Descripción
Python	3.11	Lenguaje de programación principal para la implementación de los proyectos.
pandas	2.2	Librería utilizada para manipulación, limpieza y análisis de datos en formato tabular.
scikit-learn	1.4	Librería empleada para el entrenamiento y evaluación de modelos de aprendizaje automático.
Tkinter	Estándar en Python	Librería utilizada para la construcción de la interfaz gráfica de usuario.
openpyxl	3.1	Librería para la lectura y escritura de archivos Excel (XLSX).
matplotlib	3.8	Librería utilizada para la generación de gráficas y visualizaciones.

ANEXO 4.2 PARÁMETROS DE CONFIGURACIÓN DE LOS MODELOS

Proyecto	Parámetro	Valor asignado	Descripción
Clasificación de SQR	ngram_range	(1,2)	Se utilizaron unigramas y bigramas en la vectorización de texto (*Bag of Words*).
Clasificación de SQR	random_state	42	Semilla aleatoria fijada para garantizar reproducibilidad.
Clasificación de SQR	Clasificador	Naive Bayes Multinomial	Modelo principal seleccionado tras pruebas comparativas.
Verificación de datos	Tamaño de lote	10.000 registros	Cada ejecución procesa hasta 10.000 registros por lote.
Verificación de datos	Tiempo máximo de validación	30 minutos	KPI establecido para garantizar eficiencia en ejecución.

ANEXO 5 MÉTRICAS DETALLADAS Y ANÁLISIS DE ERRORES

Caso	Entrada/Datos	Procedimiento	Resultado esperado	Criterio de aceptación	Resultado obtenido
#1	Dataset de SQR etiquetado	Cargar archivo y ejecutar clasificación	Generar predicciones en categorías correctas	Exactitud ≥ 90%	Exactitud obtenida: 90%
#2	Archivo de clientes con duplicados	Cargar archivo en sistema de verificación	Detectar duplicados y marcarlos en reporte	100% de duplicados identificados	Duplicados detectados: 100%
#3	Archivo con correos inválidos	Ejecutar validación	Detectar y marcar correos no válidos	Correos inválidos identificados correctamente	Correos inválidos detectados: 96%
#4	Archivos de entrada grandes (10.000 registros)	Ejecutar proceso completo	Validación y clasificación en <30 min	Tiempo ≤ 30 min	Validación completada en 25 min

Modelo	Exactitud	Precision (macro)	Recall (macro)	F1-macro
Baseline (diccionario de reglas)	78%	0.75	0.72	0.73
Naive Bayes	85%	0.83	0.81	0.82
Naive Bayes + diccionario (híbrido)	**91%**	**0.89**	**0.86**	**0.87**

ANEXO 6 KPIS DE IMPACTO (ANTES Y DESPUÉS)

KPI	Antes de la validación	Después de la validación	Mejora (%)
% registros con error	18%	4%	14%
Tiempo promedio de validación (minutos)	180	25	86%

Los KPIs evidencian una reducción del 80% en tiempo de procesamiento y del 14% en errores, confirmando el impacto positivo de la solución.