Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL

Juan David Restrepo Cifuentes

Guiancarlo Javier Velasco Gómez

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

Director

Jurado

Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencias de Datos.

HERNAN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali 26 junio 2023

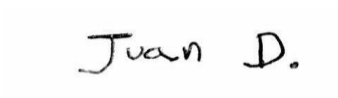
Santiago de Cali, 28 de mayo del 2023

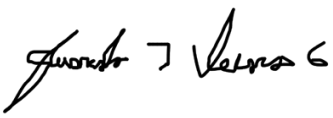
Ingeniero:
Juan Carlos Martínez Arias.
Directora Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado " Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL", el cual será realizado por el (la) estudiante Juan David Restrepo Cifuentes con código: 8973464 y el estudiante Guiancarlo Javier Velasco Gómez con código (s) 8972750 pertenecientes a la Maestría en Ciencia de Datos, bajo la dirección del profesor Mario Julián Mora Cardona.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

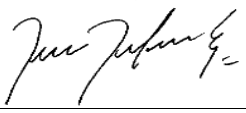
Atentamente,





Juan David Restrepo Cifuentes
C.C. 115225722 de Medellín

Guiancarlo Javier Velasco Gómez
C.C. 10.293.980 de Popayán



Mario Julián Mora Cardona
CC. 94.400.220 de Cali
**Director**

**Documentación anexa:**
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

**Maestría en** Ciencias de Datos
**Facultad de Ingeniería y Ciencias**
**FICHA RESUMEN**
**TRABAJO DE GRADO DE MAESTRIA**

**Título: Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL**

1. **ÉNFASIS:** NLP (Procesamiento de lenguaje natural)
2. **TIPO DE PROYECTO:** Aplicado – innovación
3. **ÁREA DE TRABAJO:** Sector Financiero
4. **ESTUDIANTE(S):** Juan David Restrepo Cifuentes; Guiancarlo Javier Velasco Gómez
5. **CORREO ELECTRÓNICO:** jdrestrepc@javerianacali.edu.co; guiancarlo07@javerianacali.edu.co
6. **DIRECCIÓN Y TELEFONO:** Calle 80 # 45-18 B/ Campo Valdez #1, **Ce**l:3182133312; Calle 7 # 34-26 B/San José, **CEL**: 3113269857 respectivamente.
7. **DIRECTOR:** Mario Julián Mora cardona
8. **VINCULACIÓN DEL DIRECTOR:** Profesor catedra
9. **CORREO ELECTRÓNICO DEL DIRECTOR:** mariomora@javerianacali.edu.co
10. **CO-DIRECTOR (Si aplica):** N/A
11. **GRUPO O EMPRESA QUE LO AVALA (Si aplica):** N/A
12. **OTROS GRUPOS O EMPRESAS:** N/A
13. **PALABRAS CLAVE:** Gestión de incidentes, inteligencia artificial, lenguaje de texto, base de datos, auditoria basada en datos, lenguaje natural.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Industria, innovación e Infraestructura
15. **FECHA DE INICIO:** 15 MAYO 2022
16. **RESUMEN:**


El área de auditoría interna de una entidad financiera se ha venido enfrentando a diversos desafíos a raíz de la falta de eficacia para llevar a cabo la clasificación de incidentes en los procesos sujetos a auditoría. Estas fallas han obstaculizado la toma de decisiones basadas en datos y ha llevado al incumplimiento de los acuerdos de nivel de servicio (ANS), resultando en el cierre de incidencias sin una solución adecuada.

En el marco del trabajo de grado, se propuso desarrollar un prototipo para el análisis y clasificación de incidentes en una entidad financiera utilizando el Procesamiento de Lenguaje Natural (PLN). Para abordar esta problemática, se decidió crear una aplicación de Machine Learning que pudiera clasificar los incidentes de la mesa de servicio de acuerdo con su prioridad.

Para lograr este objetivo, fue necesario llevar a cabo un proceso de limpieza de las descripciones de los incidentes, eliminando palabras irrelevantes que no aportaban al contexto y al significado de cada incidente. A continuación, se adaptaron y vectorizaron los datos textuales para que fueran fáciles de procesar por los modelos de clasificación.

Posteriormente, se evaluaron las métricas de diferentes modelos y se seleccionaron los mejores, optimizando sus hiperparámetros y probando su capacidad de predicción utilizando registros de incidentes diferentes a los utilizados en el entrenamiento.

Como resultado, se presentaron a la entidad financiera dos modelos con TF-IDF que habían sido optimizados y mostraban una precisión superior al 80%. Sin embargo, al probar los modelos con registros distintos a los utilizados en el entrenamiento, se observaron diferencias en la clasificación de hasta el 19%.

Es importante destacar que esta discrepancia no implica que el modelo esté equivocado en la clasificación, sino que invita al personal del banco a validar los incidentes en los cuales difiere de la prioridad asignada manualmente por los colaboradores de la entidad financiera.


# Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL

*Juan David Restrepo Cifuentes*

*Código* 8973464

*Guiancarlo Javier Velasco Gómez*

*Código* 8972750

*Proyecto de grado aplicado para optar al título de*
*Magister en Ciencia de Datos*

Director:
Mario Julián Mora Cardona

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 1 DE 2023

Contenido

# ÍNDICE DE FIGURAS

# ÍNDICE DE TABLAS

# INTRODUCCIÓN

En el área de auditoría interna, la falta de comprensión de herramientas tecnológicas hacia el tratamiento de datos mediante el PLN (Procesamiento de Lenguaje Natural), así como la incorrecta categorización y descripción de incidentes por parte de personal humano, y el desconocimiento de las técnicas de PLN aplicables al registro de incidentes, han generado una falta de certeza en la clasificación de estos para cada uno de los procesos que son objeto de auditoría en esta entidad.

Este proyecto se trabajó con conjuntos de datos que contienen registros de eventos reportados por los colaboradores de la entidad. Para análisis de información, se utilizaron técnicas de preprocesamiento y se seleccionaron técnicas de clasificación, seguidas de una evaluación del desempeño de estas.

Tras la aplicación de las diferentes técnicas propuestas en cada fase del proyecto, se logró establecer una categorización adecuada de los eventos según su nivel de criticidad (alto, medio y bajo). Esto permitió al área de auditoría interna emitir recomendaciones a las áreas de operación correspondientes, para que puedan brindar una respuesta acorde al tipo de evento que se presente dentro de la organización.

De esta manera, se logró desarrollar un prototipo para el análisis y categorización de incidentes en una entidad financiera, utilizando técnicas de PLN.

# 1. DEFINICIÓN DEL PROBLEMA

## 1.1. PLANTEAMIENTO DEL PROBLEMA

La inteligencia artificial ha visto incrementadas sus posibilidades en los últimos años permitiendo desarrollo de numerosas aplicaciones en el sector financiero, el empleo de las herramientas de inteligencia artificial resulta en trascendentales beneficios, dado que permiten la posibilidad de automatizar procesos e incremento de capacidades analíticas comparado con técnicas tradicionales.

En la actualidad las entidades financieras han visto la necesidad de pensar en estrategias de transformación digital debido a la incorporación de nuevos competidores en el sector bancario; éstas vienen adoptando nueva tecnología para el mejoramiento de sus procesos internos, permitiendo generación de valor frente a sus clientes.

Según The Institute of Internal Auditors [1] indica que la AI se aplica en los entes públicos, privados, industriales y gubernamentales, y específicamente en auditoría interna involucra mejoras en eficiencia de los procesos, rebaja de errores mediante automatización, métodos capaces de realizar procesos de inferencias y gestión de conocimiento de acuerdo a actividades realizadas previamente.

Pedrosa, Laureano, & Costa [2] resalta que el empleo de la tecnología de información en auditoria cobra relevancia en organismos y entidades que regulan la profesión, el uso de estas herramientas garantizan eficiencia y efectividad en las actividades de auditoría.

Las incidencias recibidas por la mesa de servicio, la cual tiene como función principal brindar soporte y responder de manera eficiente las solicitudes de servicio técnico que reportan los colaboradores, se comportan como datos de entrada o patrones de eventos,

esta recepción y análisis puede convertirse en un proceso complejo y difícil de manejar por parte de las entidades financieras, impidiendo el análisis e interpretación de los datos [3]. Para ello, se hace necesario el empleo de algoritmos, técnicas y mecanismos de verificación de minería de datos para su análisis e interpretación.

En el área de auditoría interna debido al desconocimiento de herramientas tecnológicas para el tratamiento de los datos en el lenguaje natural, la falta de categorización y descripción acertada de incidentes, la ausencia de capacitación en formación especializada en temas relacionados en ciencias de datos, dificultad para realizar búsquedas de incidentes en la descripción realizada por el creador de la novedad y el desconocimiento de técnicas de lenguaje natural aplicable a registro de incidentes, ha provocado que se presente falta de certeza en la clasificación de los incidentes para cada uno de los procesos objeto de auditar en esta entidad, lo que ha impedido toma de decisiones basadas en datos, incumplimiento de acuerdos de nivel de servicios (ANS), cierre de incidencias sin respuesta acertada, poca credibilidad de los datos de consulta del registro de incidentes, además de posibles sanciones de entes regulatorios por no aplicar los controles de manera efectiva y se pueden presentar materialización de los incidentes, convirtiéndose en riesgo operativo para el Banco.

## 1.2. FORMULACIÓN DEL PROBLEMA

Se dio respuesta a los siguientes interrogantes: ¿Qué características se consideraron relevantes para el desarrollo de un prototipo funcional de aplicación para gestionar datos? ¿Qué herramientas tecnológicas se ajustaron para el desarrollo del prototipo? ¿Cuáles fueron las técnicas de inteligencia artificial apropiadas para el desarrollo del prototipo?

¿Cuáles fueron los mecanismos necesarios para clasificar automáticamente los incidentes que se presentan en entidad financiera? ¿Cuál fue la herramienta más adecuada que permitió visualizar los resultados del prototipo funcional?

# 2. OBJETIVOS DEL PROYECTO

## 2.1. OBJETIVO GENERAL

Desarrollar un prototipo funcional de aplicación para categorizar eficientemente los datos del registro de incidencias en una entidad financiera.

## 2.2. OBJETIVOS ESPECÍFICOS

- Determinar las técnicas del procesamiento del lenguaje natural adecuadas para el desarrollo del prototipo.
- Clasificar automáticamente incidentes correspondientes a la categorización de las alertas en la entidad bancaria.
- Seleccionar herramientas tecnológicas adecuadas para el desarrollo del prototipo.

# 3. MARCO TEÓRICO Y ANTECEDENTES

## 3.1. MARCO TEÓRICO

El proyecto de grado se centra en el desarrollo de prototipo para el análisis y clasificación de incidentes en una entidad financiera utilizando PLN. Para comprender el alcance y la importancia de este trabajo, se abordaron varios temas principales.

En primer lugar, se realizó una definición exhaustiva de PLN, explorando sus

fundamentos teóricos y conceptuales. Se analizaron las técnicas y algoritmos más utilizados en PLN, como el análisis léxico, la desambiguación, la generación de lenguaje natural. Además, se examinaron las aplicaciones prácticas del PLN en el ámbito financiero y cómo puede beneficiar a las entidades en la gestión de incidentes.

En cuanto a la gestión de incidentes, se exploraron diversas metodologías, marcos de referencia y buenas prácticas relacionadas con este campo específico. Se analizaron enfoques eficientes para la detección, análisis y resolución de incidentes en entidades financieras, considerando aspectos como la priorización, la escalación y la comunicación efectiva.

### 3.1.1. Definición de Procesamiento de Lenguaje Natural

El objetivo de PLN es capacitar a las computadoras para comprender y generar lenguaje humano de manera efectiva. El PLN se basa en una serie de mecanismos y algoritmos que permiten a las computadoras interpretar y procesar texto y habla en un lenguaje de programación definido, lo que les permite comunicarse con los seres humanos en su propio lenguaje de manera natural y fluida [7].

#### Usos actuales de PLN

El PLN se emplea en variedad de aplicaciones en la vida cotidiana, desde asistentes virtuales hasta análisis de sentimientos en redes sociales. Algunos de los usos actuales más comunes del NPL incluyen: asistentes virtuales, análisis de sentimientos, traducción automática, chatbots y análisis de texto [24].

**Evolución de PLN**

Los primeros enfoques del PLN en la década de 1950 se centraron en la sintaxis de lenguaje natural y creación de gramáticas formales para analizar el lenguaje [19]. A mediados de la década de 1980, surgieron nuevos enfoques del PLN que se centraron en la estadística y el aprendizaje automático. En los últimos años, el PLN se ha centrado en la comprensión del lenguaje natural en su contexto. Modelos basados en redes neuronales [20] han demostrado ser altamente efectivos para comprender el lenguaje natural en su contexto y generar texto que suena más humano.

**Avances del PLN**

En los últimos años, el PLN ha experimentado avances significativos que han revolucionado la forma en que las computadoras comprenden y generan lenguaje natural humano. Estos avances han llevado a mejoras sustanciales en la precisión y eficacia de las aplicaciones basadas en NLP. Algunos de los desarrollos más destacados en el campo del NLP son los siguientes: Modelos de lenguaje basados en Transformers, Pre-entrenamiento de modelos de lenguaje, Incorporación de conocimiento externo y Generación de lenguaje natural interactivo, para mayor detalle ver **Anexo D**.

### 3.1.2. Marco de referencia de buenas prácticas COBIT 2019 aplicado a gestión de incidentes.

COBIT 2019 es un marco de referencia empleado en la industria para el gobierno y la gestión de la información y la tecnología (IT). Creado por ISACA (Asociación de Auditoría y Control de Sistemas de Información) y proporciona un enfoque integral para instituir y

conservar sistema de gobierno efectivo de TI en las organizaciones [4].

En COBIT 2019, los objetivos se agrupan en cinco dominios principales, cada uno con su conjunto de objetivos: Evaluar, Dirigir y Monitorizar (EDM), Alinear, Planificar e Implementar (APO), Entregar, Dar Servicio y Soporte (DSS) y Monitorizar, Evaluar y Valorar (MEA), Construir, Adquirir e Implementar (BAI). Dentro de estos dominios se encuentra el objetivo DSSO2 (Gestionar peticiones y los incidentes de servicios) de particular interés para el desarrollo de este proyecto de grado [4].

Este objetivo se encuentra dentro del dominio DSS y se relaciona con la gestión eficiente y efectiva de las peticiones e incidentes relacionados con servicios de Tecnología de la información. Al enfocarte en este objetivo, podrá explorar y desarrollar técnicas y herramientas de PLN para analizar y clasificar los incidentes en una entidad financiera.

### 3.1.3. Guía de buenas prácticas ITIL aplicado a gestión de incidentes.

ITIL V4 comprende última versión del marco de referencia de ITIL (Information Technology Infrastructure Library), ITIL V4 eleva el perfil de la gestión del servicio al reconocer su importancia estratégica en las organizaciones e industrias [5].

En el contexto del proyecto de grado, el elemento de interés en la cadena de valor del servicio es una de las prácticas de gestión de servicios. La práctica es gestión de incidentes y tiene como objetivo minimizar el impacto negativo de los incidentes en los servicios de TI, restaurando el funcionamiento normal del servicio. La gestión de incidentes se centra en responder de manera efectiva a los incidentes que afectan la disponibilidad o el rendimiento de los servicios de Tecnología e información.

Gestión de incidentes desempeña un papel fundamental en las empresas, ya que su implementación adecuada permita garantizar que el flujo de trabajo y el servicio no se interrumpan por periodos prolongados e inaceptables. Esto es especialmente importante en entidades financieras, donde la continuidad de los servicios y la respuesta rápida a los incidentes son cruciales para mantener la confianza de los clientes y cumplir con los requisitos regulatorios [5].

### 3.1.4. Herramientas para el procesamiento del lenguaje natural

Existen numerosas aplicaciones, herramientas, librerías y programas que están diseñados específicamente para facilitar el PLN. Estas herramientas son utilizadas por investigadores, desarrolladores y profesionales del PLN para realizar tareas, como análisis de texto, clasificación de documentos, extracción de información, traducción automática, generación de lenguaje natural, entre otras [9].

la ambigüedad del lenguaje natural (LN) es una de las principales dificultades que se encuentran en el (PLN). El lenguaje natural es inherentemente ambiguo, lo que significa que una misma palabra o expresión puede tener múltiples significados o interpretaciones dependiendo del contexto [9].

La tabla registrada en el **Anexo A** resume el número de estas herramientas y librerías y sus respectivos enlaces al sitio web oficial, donde se pueden observar los lenguajes de programación que pueden usar para su implementación.

Con base en la estructuración del proyecto, se realizó con un enfoque de lenguaje de código abierto y libre, donde la manipulación, limpieza, aplicación y análisis de los datos

sea sin problemas de licencia, para ello se tomó como base librerías y paquetes del lenguaje Python, en el Anexo B se relacionan las diferentes alternativas de programas.

### 3.1.5. Técnicas para el pre-procesamiento y la normalización del texto

El preprocesamiento y normalización del texto son aspectos fundamentales en el campo del PLN. Estos procesos implican la preparación y adecuación de los datos de texto antes de utilizarlos en actividades de análisis y modelado. En el desarrollo de este proyecto, se ha dado especial atención a estas etapas para garantizar la calidad de los datos y facilitar su análisis posterior.

En este trabajo se han aplicado diversas técnicas de preprocesamiento y normalización del texto, como la remoción de acentos, *stemming*, remoción de caracteres especiales, *tokenización* y el manejo de *stopwords*. Para obtener más detalles sobre cada técnica, puedes consultar el **ANEXO C**.

### 3.1.6. EDA: análisis exploratorio de datos

El análisis exploratorio de datos (EDA siglas en inglés) es una técnica utilizada en ciencia de datos para examinar y comprender los datos antes de realizar análisis más detallados o modelado. El objetivo principal del EDA es descubrir patrones, identificar relaciones, detectar valores atípicos y obtener una visión general de los datos [25].

En el desarrollo de este proyecto, se utilizaron varias bibliotecas clave de Python que permitieron realizar un análisis eficiente de los datos. Entre ellas se encuentran Pandas, NumPy y Matplotlib.

Pandas: es una biblioteca poderosa y flexible que brinda herramientas para manipular y

analizar datos de manera eficiente. Permite trabajar con estructuras de datos como DataFrames, lo cual facilita la limpieza, transformación y exploración de los datos. Con Pandas, se realizaron operaciones como la selección y filtrado de datos, cálculos estadísticos, y fusionar conjuntos de datos para un análisis más completo.

Por otro lado, NumPy se enfoca principalmente en la manipulación de matrices y cálculos numéricos. Esta biblioteca es fundamental para el análisis de datos, ya que proporciona funciones y métodos eficientes para realizar operaciones matemáticas y estadísticas en nuestros datos. Con NumPy, se realizaron cálculos numéricos complejos, trabajar con matrices multidimensionales y realizar manipulaciones y transformaciones de datos de manera eficiente.

Además, se utilizó Matplotlib para la visualización de datos. Matplotlib es una biblioteca ampliamente utilizada y poderosa que permite crear una variedad de gráficos y visualizaciones para explorar y comprender mejor los datos. Con Matplotlib, se pudo crear gráficos de líneas, gráficos de barras, histogramas, diagramas de dispersión y muchas otras visualizaciones que ayudaron a identificar patrones, tendencias y relaciones en los datos [25].

### 3.1.7. Modelos supervisados

Los modelos supervisados son un tipo de técnica de aprendizaje automático en la cual se entrena un algoritmo utilizando conjunto de datos etiquetados. Estos datos etiquetados consisten en pares de entrada y salida esperada, donde la salida esperada también se conoce como etiqueta o variable objetivo.

El objetivo de un modelo supervisado es aprender una función o relación entre las características de entrada y las salidas esperadas para poder hacer predicciones precisas sobre nuevas entradas sin etiquetar [26].

### 3.1.8. Modelos no supervisados

Los modelos no supervisados son aquellos en los que el algoritmo se encarga de explorar y descubrir patrones o estructuras ocultas en los datos sin tener información previa o etiquetas que indiquen la respuesta correcta.

En los modelos no supervisados, el objetivo principal suele ser la agrupación (clustering) o la reducción de dimensionalidad. En la agrupación, el algoritmo busca identificar grupos o clústeres de datos similares entre sí, sin saber de antemano a qué grupo pertenece cada dato. Por otro lado, la reducción de dimensionalidad se refiere a la técnica de representar los datos en un espacio de menor extensión, manteniendo la mayor cantidad de información relevante. Estos modelos se utilizan principalmente para la segmentación de datos, la reducción de dimensionalidad y la detección de anomalías. Existen varias técnicas y algoritmos utilizados en modelos no supervisados, entre los cuales se incluyen: Clustering (Agrupamiento), Anomaly detection (Detección de anomalías), Reducción de dimensionalidad, Reglas de asociación [26].

### 3.1.9. Modelos de representación

#### 3.1.9.1 Bolsa de palabras PLN

Bolsa de Palabras, es una representación simplificada de un texto o documento en el campo del PLN. En este enfoque, se ignora el orden de las palabras y se considera la

frecuencia de aparición de las palabras individuales en el texto. La Bolsa de Palabras se construye siguiendo los siguientes pasos: Tokenización, Construcción del vocabulario, Creación de vectores de características [26].

#### 3.1.9.2 TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) es una técnica empleada en PLN y la recuperación de información para evaluar la importancia relativa de una palabra en un documento dentro de un corpus. La idea principal de TF-IDF consiste en determinar qué tan relevante es una palabra en un documento específico en comparación con su aparición en otros documentos del corpus. Esto se logra mediante dos componentes principales: la frecuencia de termino (TF) encargada de medir la frecuencia de una palabra especifica en un documento, cuantas más veces aparezca una palabra en el documento, mayor será su valor de frecuencia de termino, el otro componente principal corresponde a la frecuencia inversa de documento (IDF) encargada de medir la rareza de una palabra en todo el corpus de documento. Se calcula como el logaritmo inverso de la fracción de documentos en el corpus que contienen la palabra específica. En otras palabras, cuantos menos documentos contengan la palabra, mayor será su valor de IDF.
La fórmula general para calcular TF-IDF es: **TF-IDF = TF * IDF**
El resultado final del cálculo TF-IDF es un valor numérico que representa la importancia relativa de una palabra en un documento. Cuanto mayor sea el valor de TF-IDF, más relevante se considera la palabra en ese documento específico.
La técnica de TF-IDF se utiliza ampliamente en tareas como la recuperación de información, la clasificación de documentos, la recomendación de contenido y el análisis

de texto en general. Ayuda a identificar las palabras clave y destacar las características más importantes de un documento dentro de un corpus más amplio [6].

### 3.1.10 .Métricas de evaluación de modelos

Las métricas utilizadas para evaluar modelos supervisados pueden variar dependiendo de la tarea específica y tipo de modelo. A continuación, se presentan algunas métricas comunes utilizadas en los modelos supervisados [26].

Métricas para modelos supervisados:

- ✓ **Exactitud (Accuracy):** Es la proporción de los ejemplos clasificados correctamente sobre total de ejemplos en el conjunto de prueba. Es una medida general de la precisión del modelo [26].
- ✓ **Precisión (precision):** Representa la proporción de ejemplos positivos clasificados de manera correcta sobre la totalidad de ejemplos clasificados como positivos. Se enfoca en la precisión de las predicciones positivas [26].
- ✓ **Recall o Sensibilidad (Recall or Sensitivity):** Es la proporción de ejemplos positivos clasificados sobre total de ejemplos positivos en el conjunto de prueba. Se enfoca en la capacidad del modelo para identificar correctamente ejemplos positivos [26].
- ✓ **F1-score:** Es una medida que combina la precisión y el recall en una única métrica. Representa la media armónica entre ambas medidas y proporciona un equilibrio entre ellas [26].
- ✓ **Matriz de confusión:** Tabla que muestra la cantidad de ejemplos clasificados

correcta e incorrectamente para cada clase. Puede utilizarse para calcular diversas métricas, como la exactitud, la precisión y el recall [26].

- ✓ **R_Cuadrado:**

  La métrica R cuadrado es una medida utilizada para evaluar la calidad de un modelo de regresión y representa la proporción de la varianza de la variable dependiente que es explicada por el modelo. Un valor de R cuadrado cercano a 1 indica un buen ajuste del modelo, mientras que un valor cercano a 0 indica que el modelo no puede explicar la variabilidad de la variable dependiente [26].

### 3.1.11 Optimización de hiperparametros

La optimización de hiperparámetros es un proceso crucial en el aprendizaje automático (machine learning) que implica encontrar la combinación óptima de valores para los hiperparámetros de un modelo. Estos son variables que no se aprenden directamente del conjunto de datos, pero que influyen en el rendimiento y comportamiento del modelo.

Para el desarrollo de este proyecto se empleó la biblioteca scikit-learn de Python que proporciona una clase llamada Pipeline que permite construir y administrar pipelines en el contexto del aprendizaje automático. Esta clase permite encadenar varios estimadores (como transformadores y modelos) en una secuencia definida, donde cada paso se configura con sus propios hiperparámetros.

También fue empleado el módulo GridSearchCV de la biblioteca scikit-learn encargado de proporcionar una herramienta útil para realizar la búsqueda exhaustiva de hiperparámetros en una cuadrícula de valores predefinidos [6].

### 3.1.12 Metodología Crisp-DM (Cross-Industry Standard Process for Data Mining)

Enfoque sistemático y estructurado utilizado por los científicos de datos para realizar proyectos de minería de datos. Esta metodología consta de seis fases principales, que son la comprensión del negocio, comprensión de los datos, preparación de los datos, modelización, evaluación y la implementación [6].

- ✓ Primera fase, comprensión del negocio, se debe identificar el problema a resolver y se define el objetivo del proyecto.
- ✓ En la segunda fase, comprensión de los datos, se realiza una exploración inicial de los datos para identificar su calidad, relevancia y limitaciones.
- ✓ Tercera fase, preparación de los datos, se debe realizar el procesamiento de datos para limpiarlos, integrarlos y transformarlos para su posterior uso en la modelización.
- ✓ Cuarta fase, el modelamiento, implica selección de técnicas de modelado adecuadas y la construcción de modelos predictivos a partir de datos procesados.
- ✓ En la quinta fase, la evaluación, se evalúan los modelos construidos para determinar su precisión y fiabilidad.
- ✓ Por último, en la fase de implementación, se implementan los modelos seleccionados en un entorno operativo y se monitorean para asegurarse de que continúen cumpliendo su objetivo.

La metodología CRISP-DM es una herramienta valiosa para los científicos de datos, ya que permite planificar y gestionar proyectos de minería de datos de manera

estructurada y rigurosa [6].

### 3.1.13 Modelos de clasificación

Los modelos de clasificación son algoritmos utilizados en el aprendizaje automático supervisado para asignar instancias o muestras a diferentes categorías o clases predefinidas. Estos modelos se construyen a partir de un conjunto de datos de entrenamiento que contiene instancias etiquetadas con sus respectivas clases.

El objetivo de un modelo de clasificación es aprender patrones y relaciones en los datos de entrenamiento para poder clasificar correctamente nuevas instancias no etiquetadas en una de las clases existentes. Existen diversos tipos de modelos de clasificación, en el Anexo G se puede validar la descripción de los modelos empleados en el desarrollo de este trabajo de grado.

## 3.2. ANTECEDENTES

Los auditores de los bancos pueden desempeñar un papel importante para contribuir a la estabilidad financiera cuando realizan auditorías bancarias de calidad que fomentan la confianza del mercado en los estados financieros. El Comité de Supervisión Bancaria de Basilea [8] publica información sobre las auditorías de los bancos para mejorar calidad de las auditorías y aumentar la eficacia de la supervisión prudencial, que contribuyen a la estabilidad financiera.

Las herramientas de interfaz podrían utilizarse para compartir automáticamente la información en tiempo real con las herramientas de IA del auditor, que a su vez podría analizar, probar y señalar las anomalías o problemas que requieren la atención del

auditor. De este modo, la interacción humana se centraría en las transacciones de alto riesgo y no en las consultas rutinarias [14].

En este caso, las herramientas de IA podrían identificar las transacciones inusuales y, al mismo tiempo, proporcionar información sobre las consideraciones pertinentes que el auditor podría tener en cuenta, incluidas las normas aplicables (normas de contabilidad, de divulgación, de auditoría o reglamentarias), situaciones históricas similares o resultados de fuentes disponibles públicamente (incluidas situaciones similares de grupos de pares del sector) [14].

En el área de auditoría interna del Banco, el enfoque principal del grupo de auditores se centra en las auditorías continuas y basadas en datos. Por esta razón, la implementación de técnicas de inteligencia artificial en el desarrollo de actividades es de gran importancia. En este sentido, la tesis mencionada previamente es relevante ya que emplea PLN para el análisis y clasificación de incidentes en la entidad financiera. Los modelos creados en esta investigación tienen una gran relación con el tema de interés en este trabajo de maestría, por lo que considerar esta tesis puede resultar valioso para el desarrollo de las actividades propias de esta investigación.

### 3.2.1 Auditoria basada en datos

Las tecnologías avanzadas proporcionan una gran cantidad de información a un auditor que le permite emitir un juicio. Pero el auditor seguirá siendo quien emita ese juicio. La tecnología es un facilitador y no tiene rival cuando se trata de identificar correlaciones entre conjuntos de datos o variables.

Sin embargo, se necesita la visión y la experiencia humanas para comprender en última

instancia el contexto subyacente al resultado, así como la causalidad del resultado en relación con las entradas proporcionadas. Los resultados de la IA son, en el mejor de los casos, predicciones probabilísticas basadas en inferencias en la correlación de datos y no deben tomarse como verdades (es decir, las predicciones no son necesariamente la respuesta "correcta ") [15].

El auditor debe emplear su criterio profesional para evaluar los resultados de la IA en combinación con otras pruebas. Las herramientas de IA pueden ofrecer otro nivel de conocimiento, pero no son la única respuesta. Un auditor confirma la información y determina si se trata de una anomalía y, lo que es más importante, determina lo que implica o cómo concluir sobre lo apropiado del tratamiento de información.

En el área de auditoría interna del Banco se vienen creando iniciativas en torno al desarrollo de modelos de analítica de datos, permitiendo la toma de decisiones basadas en datos y poder de esta manera emitir recomendaciones a las diferentes Gerencias del Banco en relación a los hallazgos encontrados durante el desarrollo de las auditorias definidas en plan anual de Auditoria, luego el tema tratado en esta tesis es de mucha trascendencia para el desarrollo del trabajo de grado, dado que se ajusta a la temática que se viene tratando al interior del área de auditoría interna del Banco.

### 3.2.2 Transición de auditorías basadas en riesgos a datos

El uso de herramientas de IA también puede plantear cuestiones en torno al uso del muestreo. Por ejemplo, con respecto a los procedimientos de auditoría sustantiva, puede ser más eficiente utilizar el muestreo de auditoría si el auditor no puede diseñar un procedimiento con parámetros suficientemente precisos para un conjunto de datos y

evitar la necesidad de gestionar el número resultante de valores atípicos, ya que puede aumentar (a veces hasta los miles) cuando se analiza el 100% de la población [16].

En la actualidad, los informes de auditoría suelen publicarse tras el cierre del periodo de auditoría (normalmente entre 25 y 120 días después del cierre del periodo) [17]; sin embargo, los usuarios exigen cada vez más información puntual. Muchas partes de la auditoría ya pueden automatizarse o ejecutarse en paralelo y, por lo tanto, dar lugar a una finalización más rápida de los pasos individuales de la auditoría (a pesar de las limitaciones de la IA identificadas anteriormente).

La auditoría continua (por ejemplo, mensual, trimestral u otro plazo pertinente) o la información en tiempo real (por ejemplo, a medida que se producen las transacciones) pueden llegar a ser habituales.

La velocidad a la que el auditor puede emitir su opinión tras el cierre de un periodo está intrínsecamente limitada por la velocidad de información del cliente. El uso de herramientas de IA por parte del auditor puede dar lugar a que la herramienta de IA del auditor identifique continuamente las transacciones significativas a medida que se contabilizan (suponiendo que los controles estén diseñados y funcionen eficazmente) y realice automáticamente procedimientos para validar dichas transacciones (por ejemplo, relacionándolas con los detalles de las transacciones bancarias, evaluándolas con respecto a las condiciones contractuales), de modo que el auditor sólo tenga que evaluar si se requieren procedimientos adicionales en las afirmaciones que no pudieron comprobarse en tiempo real [18].

De momento en el área de auditoría interna del Banco se viene trabajando en auditorias de cumplimiento para dar respuesta a los organismos de control, además se trabajan

auditorias basadas en riesgo mediante el apoyo de marcos de referencia y guías de buenas prácticas, en la actualidad se está enfocando en auditorias basadas en datos, de ahí la importancia de lo resaltado en la tesis de grado comentada anteriormente.

### 3.2.3 Modelo de procesamiento de lenguaje natural

A diario se comparte en redes sociales y portales indeterminado número de noticias entre familiares y amigos, desconociendo el impacto que puede generar al momento de la toma de decisiones frente a un evento en particular, por ejemplo: en los últimos acontecimientos sucedidos en Colombia (2021) relacionados con las protestas sociales y elecciones políticas fue notorio como la comunidad en general ha podido llegar a sentir miedo y un sin sabor ante las publicaciones de noticias que se generaban, afectando el ambiente económico, político y social de la región, luego este trabajo de grado buscó determinar el impacto que producen las noticias.

El objetivo de ese trabajo de grado consiste en implementar un modelo de aprendizaje automático para predecir fidelidad posible, la viralidad de artículos en línea. Sus autores se enfocaron en la metodología CRISP-DM.

Para validar este trabajo de grado emplearon encuestas que permitió comprobar los objetivos planteados. Obtuvieron como mejor modelo un núcleo de arquitectura basado en modelo pre-entrenado, denominado BERT, que permitía predecir, para pareja de títulos de noticias, si el primer título era más viral que el segundo [6].

La tesis mencionada guarda estrecha relación con el tema abordado en nuestro proyecto de grado, ya que se enfoca en el PLN en artículos. Su metodología es similar a la adoptada en el presente trabajo de maestría, lo que ha sido de gran utilidad al considerar

el enfoque tratado y la similitud de algunos de los objetivos planteados.

# 4. METODOLOGÍA

El presente trabajo de grado se desarrolló bajo la modalidad de proyecto teórico-práctico y se trabajó de la siguiente manera:

- Validación de marcos de referencia y buenas prácticas que permitieron definir las actividades necesarias para la clasificación de los incidentes.
- Realización de extracción de (características) destinadas a ser informativas y no redundantes.
- Abstracción de información para determinar las técnicas de PLN adecuadas para el desarrollo de un prototipo.
- Definición de las herramientas tecnológicas adecuadas para el desarrollo de un prototipo.

Para la recopilación de la información, se emplearon entrevistas e investigación documental, de igual manera se utilizaron herramientas y técnicas necesarias para el desarrollo de prototipo.

## 4.1 ESQUEMA DE TRABAJO

Se destino tiempo promedio de 12 horas semanales por estudiante. Se realizaron reuniones semanales con el director para validar progreso del proyecto y recibir retroalimentación del trabajo de grado, además de reuniones con el profesor asignado para la materia proyecto aplicado III.

## 4.2 FASES DE DESARROLLO DEL PROYECTO

En el desarrollo del proyecto de grado se empleó metodología CRISP-DM, que fue considerada debido a que tiene un enfoque sistemático y estructurado para realizar proyectos de ciencias de datos. La metodología consta de las siguientes fases: entendimiento del negocio, entendimiento de los datos, preparación de los datos, modelado, evaluación y el despliegue.

Cabe resaltar que para el desarrollo de este trabajo no fue considerada la etapa de despliegue, dado que el alcance de la propuesta está definido solo hasta el desarrollo de prototipo.

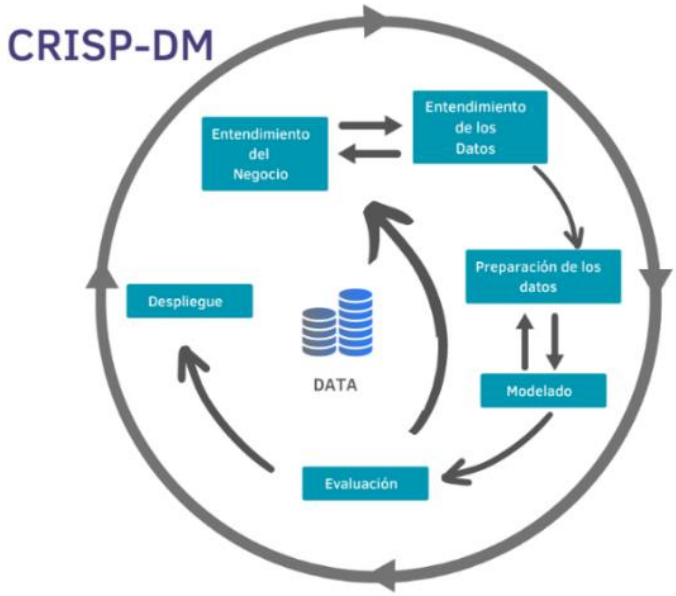



***Figura 4.2.1*** *Fases del Modelo Crisp-DM*

*Fuente: Elaboración propia*

En la tabla 1 se relaciona las fases de CRISP-DM con sus correspondientes actividades y entregables acordes al desarrollo del proyecto de grado.

***Tabla 1*** *Descripción de las actividades desarrolladas (CRISP-DM).*

| Fases CRIPD-DM | Actividades | Entregables |
|---|---|---|
| Entendimiento del negocio | _ Definición del problema<br>_ Objetivos del proyecto.<br>_ Alcance<br>_ Justificación<br>_ Marco teórico de referencia y antecedentes<br>_ Antecedentes<br>_ Trabajos relacionados<br>_ Metodología<br>_ Recursos a emplear<br>_ Cronograma | _ Anteproyecto de grado. |
| Entendimiento de los datos | _ Recolección de los datos (registro de incidentes reportados en mesa de servicio de entidad financiera).<br>_ Descripción de Las variables.<br>_ Explorar los datos. | _ Análisis exploratorio de los datos. |
| Preparación de los datos | _ Remover acentos<br>_ Colocar en minúsculas<br>_ Eliminar líneas sobrantes<br>_ Stemming<br>_ Insertar espacios entre caracteres especiales para aislarlos<br>_ Remover caracteres especiales y dígitos<br>_ Remover espacios en blanco extras<br>_ Tokenización<br>_ Eliminación de Stopwords | _ Código con la preparación y limpieza de datos. |
| Modelado | _ Dividir datos en datos de prueba y datos de entrenamiento.<br>_ Configurar modelos de representación: bolsa de palabras y modelo de TF-IDF<br>_ Aplicar modelos de clasificación con cada uno de los modelos de representación.<br>_ Cálculo de métricas de modelos<br>_ Selección de mejores modelos<br>_ Optimización de hiperparametros | _ Marco teórico del anteproyecto.<br>_ Código del modelado en Python. |
| Evaluación | _ Evaluar predicción de los modelos optimizados con los datos de prueba. | _ Documento con revisión de los resultados obtenidos. |

Fuente: Elaboración propia

En la Tabla 2 se mapea cada fase del modelo CRISP-DM de acuerdo con los objetivos específicos:

***Tabla 2*** *Fases de CRISP-DM mapeado con los objetivos específicos del proyecto.*

| Fase CRISP-DM | Objetivos específicos relacionados |
|---|---|
| Entendimiento del negocio | Objetivo 1 |
| Entendimiento de los datos | Objetivo 1 |
| Preparación de los datos | Objetivo 2 |
| Modelado | Objetivo 2 |
| Evaluación | Objetivo 3 |

Fuente: Elaboración propia

# 5. PRESENTACIÓN DE LA PROPUESTA

## 5.1 ENTENDIMIENTO DE LOS DATOS

**Objetivo específico 1**: Determinar las técnicas del procesamiento del lenguaje natural adecuadas para el desarrollo del prototipo.

### 5.1.1 Recolección de los datos (registro de incidentes reportados en mesa de servicio de entidad financiera).

La entidad financiera facilitó una muestra en formato Excel de los tickets que se escalaron durante enero de 2022. Estos tuvieron a cabo un proceso previo de “anonimización” con el objetivo de mantener la confidencialidad de ciertos puntos claves. Para ese mes se escalaron 14.286 tickets, esta muestra resultó bastante significativa ya que permitió ejecutar sin problemas las fases posteriores.

Con relación a los tickets, estos son escalados a la mesa de servicios por parte de las distintas gerencias de la entidad a través de los distintos medios habilitados para tal fin, los más importantes para ese mes fueron el portal de reporte de tickets con 10.273 y los

flujos de trabajo con 4005.

### 5.1.2 Descripción de Las variables.

Cada una de las 14.286 solicitudes o tickets del conjunto de datos trabajado cuenta con una serie de características diferentes, a saber:

**1. ID:** un identificador único para la solicitud o ticket.
**2. Fecha de Creación:** la fecha de creación de la solicitud o ticket.
**3. Título:** título de la solicitud o ticket.
**4. Descripción:** texto en el cual se explica la problemática que se le está presentando al colaborador de la entidad.
**5. Prioridad:** clasificación de prioridad que se le da al ticket, puede ser: Bajo, Baja - Media, Medio, Alto y Urgente.
**6. Origen:** aplicativo desde el cual llega la solicitud.
**7. Creado Por:** identificador único del colaborador que crea la solicitud.
**8. Oficina:** punto de venta, establecimiento o unidad empresarial de la entidad financiera.
**9. Regional:** departamento donde se encuentra la oficina registrada en la solicitud.
**10. Gerencia:** área de la entidad financiera donde trabaja el colaborador que realiza la solicitud.
**11. Clasificación:** clasificación que los colaboradores de la mesa de servicio le asignan a las solicitudes con características similares.
**12. Grupo de Soporte:** la solicitud se les escala a los distintos niveles de la mesa de servicios para encontrar la solución a la problemática.
**13. Asignado A:** identificador único del colaborador que atiende la solicitud.
**14. Estado:** al tratarse de solicitudes del mes de enero de 2022 ya todas se encuentran cerradas.
**15. Detalle Respuesta:** texto en el cual se explica la solución a la problemática de la solicitud.
**16. Fecha Solución:** la fecha en que se implementa la solución al problema.
**17. Tiempo Solución Minutos**: tiempo transcurrido en minutos para implementar la solución al problema.
**18. Fecha Cierre:** la fecha en que se cierra la solicitud o ticket.
**19. Tipo Ticket:** todas están clasificadas como solicitudes.
**20. Cumplió SLA:** se menciona si la solicitud esta vencida o no con respecto al tiempo de respuesta acordada por la mesa de servicio para las solicitudes.

Con estas variables se realizó un análisis exploratorio de datos que tenía como objetivo analizar su comportamiento y dar un valor añadido al trabajo al presentar datos valiosos que le permitieran a la entidad bancaria tomar futuras decisiones relacionadas a los hallazgos de la exploración. Algunos de los datos más valiosos están relacionados en el **ANEXO E.**

### 5.1.3 Explorar los datos.

Como se mencionó anteriormente, se realizó un análisis exploratorio de datos para las distintas variables del conjunto de datos de tickets como valor agregado para la entidad financiera. Sin embargo, dando alcance a los objetivos definidos en el trabajo se continuó el proceso con dos variables: las descripciones de las distintas solicitudes y la prioridad asignada a los tickets. El análisis de estas variables también se puede consultar en el **ANEXO E.**

De la variable “Prioridad” se identificó que había algunas categorías que contaban con muy pocos registros como, la de Baja – Media, que solo contaba con 8 registros y por lo tanto representaba el 0.1% del total de tickets. Otra prioridad que fue apareció con baja frecuencia fue la de Urgente que solo contaba con 75 apariciones lo que representaba el 0.5% del total de tickets. Al final para simplificar los pasos posteriores se tomó la decisión de unificar los tickets clasificados como Urgente con los de Alto y los tickets clasificados como Baja – Media con los de clasificación Medio.

Las razones principales para soportar estas modificaciones fueron las siguientes:

**1.** Las categorías con pocos registros tienen un bajo significado, esto hubiese llevado a que los modelos implementados en los pasos posteriores tuvieran demasiados problemas para clasificar correctamente estas categorías.

**2.** Recordemos que la muestra utilizada en el proyecto es de enero de 2022, en los siguientes meses la entidad financiera simplifico su forma de clasificar las prioridades de las solicitudes limitando el alcance a bajo, medio y alto.

Con relación a la variable que almacenaba las descripciones de los distintos tickets de la muestra se analizó el top 100 de palabras que más se repetían. En este punto identificaron las palabras que, en nuestro idioma sirven de conectores entre palabras, como las que más se repiten. Esto se debía solucionar en la limpieza de los datos ya que contar con estas palabras no aportaban valor al proceso de clasificación de solicitudes y además podía opacar palabras que se repiten menos pero que son claves para entender porque un ticket se clasifica como Alto y no como Medio o Bajo.

## 5.2 PREPARACIÓN DE LOS DATOS

**Objetivo Específico 2:** Clasificar automáticamente incidentes correspondientes a la categorización de las alertas en la entidad bancaria.

Una vez finalizada toda la fase de entendimiento de los datos se continuó con la preparación de datos, en esta fase se construyó una función que tomaba el corpus que para el caso específico es la variable donde se almacenan las descripciones de las 14.286 solicitudes y se aplicaron distintos pasos que buscaban volverlas descripciones más comprensibles para los modelos. A continuación, se hace una breve descripción del

paso a paso de la función de preprocesamiento y normalización del corpus.

### 5.2.1 Remover acentos

Este paso buscaba identificar en el corpus los caracteres o letras acentuadas. Posteriormente se usaba la librería “unicodedata” para transformarlos en caracteres ASCII o caracteres ingleses lo que ayudaba a normalizar las palabras del corpus.

### 5.2.2 Colocar en minúsculas

Para facilitar el manejo del texto lo más conveniente era convertir todos los caracteres en minúsculas.

### 5.2.3 Eliminar líneas sobrantes

Utilizando expresiones regulares se realizó un proceso de eliminación de las líneas sobrantes entre cada una de las descripciones para de esta manera facilitar el posterior manejo de los datos sin que se diera algún error que se fuera difícil de identificar.

### 5.2.4 Stemming

Para entender el proceso de separación de palabras primero se debe entender que en español, los afijos son unidades como los prefijos y sufijos. Estos se unen a la raíz de una palabra para cambiar su significado o crear nuevas palabras. Este proceso es conocido como inflexión. El paso de stemming nos ayuda a normalizar y estandarizar las palabras según su raíz independientemente de sus inflexiones.

Para este proceso se utilizó uno de los stemmers más empleados, el Porter, este algoritmo fue desarrollado por Martin Porter y está disponible en la librería NLTK. Es de

amplio uso porque soporta los idiomas de inglés y español.

### 5.2.5 Insertar espacios entre caracteres especiales para aislarlos

En este paso se llevó a cabo un proceso de aislamiento de los caracteres especiales y dígitos almacenados en las descripciones por medio de expresiones regulares. Este paso permitiría su posterior eliminación.

### 5.2.6 Remover caracteres especiales y dígitos

Por medio de las expresiones regulares se eliminaron los caracteres especiales y los dígitos que se encontraban en las descripciones de los distintos tickets. Este paso era de vital importancia ya que en las descripciones existían algunos datos personales de los colaboradores como la cédula o el número telefónico, así que se logró eliminar toda esta información privada junto con el resto de los caracteres no alfanuméricos que no aportaban a la clasificación y que por el contrario le sumaban ruido al texto.

### 5.2.7 Remover espacios en blanco extras

Este paso se realizó por medio de expresiones regulares y buscaba eliminar espacios en blanco que no eran necesarios y que podían generar algún error de compilación del código de los modelos de clasificación.

### 5.2.8 Tokenizar

Tokenizar es un paso de vital importancia en la normalización del corpus, esta tiene como objetivo dividir o segmentar las frases en las palabras que la componen. La librería de NLTK ofrece varias opciones para este proceso, por lo que se decidió implementar el

tokenizador TokTok que es uno de los últimos introducidos por NLTK y que asume que la entrada (corpus) tiene una frase por línea.

### 5.2.9 Eliminación de Stopwords

Las stopwords o palabras vacías son aquellas que tienen poco o ningún significado y suelen eliminarse del texto al procesarlo para que no generen ruido y se puedan conservar aquellas palabras que tienen el máximo significado y contexto para aportar a la clasificación.

La eliminación de stopwords se realizó en simultaneo con el proceso de Tokenización ya que este permitía ir haciendo una identificación de su frecuencia de aparición basado en tokens singulares para proceder a filtrarlas del corpus. A continuación, se puede observar la lista de stopwords en español disponible en la Librería de NLTK.

```
print(stopword_list)
```

```
['de', 'la', 'que', 'el', 'en', 'y', 'a', 'los', 'del', 'se', 'las', 'por', 'un', 'para', 'con', 'no', 'una', 'su', 'al', 'lo', 'como', '
más', 'pero', 'sus', 'le', 'ya', 'o', 'este', 'sí', 'porque', 'esta', 'entre', 'cuando', 'muy', 'sin', 'sobre', 'también', 'me', 'hasta',
'hay', 'donde', 'quien', 'desde', 'todo', 'nos', 'durante', 'todos', 'uno', 'les', 'ni', 'contra', 'otros', 'ese', 'eso', 'ante', 'ellos'
, 'e', 'esto', 'mí', 'antes', 'algunos', 'qué', 'unos', 'yo', 'otro', 'otras', 'otra', 'él', 'tanto', 'esa', 'estos', 'mucho', 'quienes',
'nada', 'muchos', 'cual', 'poco', 'ella', 'estar', 'estas', 'algunas', 'algo', 'nosotros', 'mi', 'mis', 'tú', 'te', 'ti', 'tu', 'tus', 'e
llas', 'nosotras', 'vosotros', 'vosotras', 'os', 'mío', 'mía', 'míos', 'mías', 'tuyo', 'tuya', 'tuyos', 'tuyas', 'suyo', 'suya', 'suyos',
'suyas', 'nuestro', 'nuestra', 'nuestros', 'nuestras', 'vuestro', 'vuestra', 'vuestros', 'vuestras', 'esos', 'esas', 'estoy', 'estás', 'e
stá', 'estamos', 'estáis', 'están', 'esté', 'estés', 'estemos', 'estéis', 'estén', 'estaré', 'estarás', 'estará', 'estaremos', 'estaréis'
, 'estarán', 'estaría', 'estarías', 'estaríamos', 'estaríais', 'estarían', 'estaba', 'estabas', 'estábamos', 'estabais', 'estaban', 'estu
ve', 'estuviste', 'estuvo', 'estuvimos', 'estuvisteis', 'estuvieron', 'estuviera', 'estuvieras', 'estuviéramos', 'estuvierais', 'estuvier
an', 'estuviese', 'estuvieses', 'estuviésemos', 'estuvieseis', 'estuviesen', 'estando', 'estado', 'estada', 'estados', 'estadas', 'estad'
, 'he', 'has', 'ha', 'hemos', 'habéis', 'han', 'haya', 'hayas', 'hayamos', 'hayáis', 'hayan', 'habré', 'habrás', 'habrá', 'habremos', 'ha
bréis', 'habrán', 'habría', 'habrías', 'habríamos', 'habríais', 'habrían', 'había', 'habías', 'habíamos', 'habíais', 'habían', 'hube', 'h
ubiste', 'hubo', 'hubimos', 'hubisteis', 'hubieron', 'hubiera', 'hubieras', 'hubiéramos', 'hubierais', 'hubieran', 'hubiese', 'hubieses',
'hubiésemos', 'hubieseis', 'hubiesen', 'habiendo', 'habido', 'habida', 'habidos', 'habidas', 'soy', 'eres', 'es', 'somos', 'sois', 'son',
'sea', 'seas', 'seamos', 'seáis', 'sean', 'seré', 'serás', 'será', 'seremos', 'seréis', 'serán', 'sería', 'serías', 'seríamos', 'seríais'
, 'serían', 'era', 'eras', 'éramos', 'erais', 'eran', 'fui', 'fuiste', 'fue', 'fuimos', 'fuisteis', 'fueron', 'fuera', 'fueras', 'fuéramo
s', 'fuerais', 'fueran', 'fuese', 'fueses', 'fuésemos', 'fueseis', 'fuesen', 'sintiendo', 'sentido', 'sentida', 'sentidos', 'sentidas', '
siente', 'sentid', 'tengo', 'tienes', 'tiene', 'tenemos', 'tenéis', 'tienen', 'tenga', 'tengas', 'tengamos', 'tengáis', 'tengan', 'tendré
', 'tendrás', 'tendrá', 'tendremos', 'tendréis', 'tendrán', 'tendría', 'tendrías', 'tendríamos', 'tendríais', 'tendrían', 'tenía', 'tenía
s', 'teníamos', 'teníais', 'tenían', 'tuve', 'tuviste', 'tuvo', 'tuvimos', 'tuvisteis', 'tuvieron', 'tuviera', 'tuvieras', 'tuviéramos',
'tuvierais', 'tuvieran', 'tuviese', 'tuvieses', 'tuviésemos', 'tuvieseis', 'tuviesen', 'teniendo', 'tenido', 'tenida', 'tenidos', 'tenida
s', 'tened']
```

***Figura 5.2.1*** *Stopwords disponibles en español para la librería NLTK*

*Fuente: Elaboración propia*

## 5.3 MODELADO

**Objetivo Específico 2:** Clasificar automáticamente incidentes correspondientes a la

categorización de las alertas en la entidad bancaria

Una vez se finaliza la etapa de preparación de los datos se sigue con el modelado del prototipo de clasificación de solicitudes de la entidad financiera, ver **Anexo H**. A continuación, se describen los pasos realizados en esta etapa:

### 5.3.1 Dividir el conjunto de datos en datos de prueba y datos de entrenamiento.

Para este paso ya se contaba con un corpus de descripciones que había sido normalizado y que ya no tenía mucho de los ruidos del idioma que podían afectar el rendimiento de los modelos a implementar. Aquí es importante aclarar que estos procesos de limpieza y preparación son iterativos hasta encontrar un corpus adecuado para el problema que se enfrentaba.

Para este punto se utilizó la librería sklearn para dividir el corpus y etiquetas con la intención de obtener las que se usarían para entrenar el modelo y las que servirían para evaluarlo al final del proceso. A continuación, se describen las variables que son definidas en este paso:

- ✓ **train_corpus** = se almacenan 9571 descripciones de las 14.286 que se tienen en el corpus inicial, esto con el objetivo de que sean utilizadas para entrenar los modelos de clasificación.
- ✓ **test_corpus** = se almacenan 4715 descripciones de las 14.286 que se tienen en el corpus inicial, esto con el objetivo de que sean utilizadas como nuevos registros que permitan evaluar la capacidad de predicción de los modelos de clasificación entrenados.

- ✓ **train_label_names** = se almacenan 9571 prioridades de las 14.286 que se tienen en el corpus inicial, esto con el objetivo de apoyar el proceso de entrenamiento de los modelos.
- ✓ **test_label_names** = se almacenan 4715 prioridades de las 14.286 que se tienen en el corpus inicial, esto con el objetivo de analizar en cuales descripciones de prueba o testeo el modelo clasifica el ticket con una prioridad distinta a la asignada por el personal de la entidad financiera.

```
from sklearn.model_selection import train_test_split

train_corpus, test_corpus, train_label_names, test_label_names =\
                                train_test_split(np.array(data_clean['Descripcion_limpia']),
                                                 np.array(data_clean['Category']), test_size=0.33, random_state=42)

train_corpus.shape, test_corpus.shape
```

```
((9571,), (4715,))
```

***Figura 5.3.1*** *División del corpus y las etiquetas*

*Fuente: Elaboración propia*

### 5.3.2 Configurar modelos de representación: bolsa de palabras y modelo de TF-IDF

Contar con el corpus ya tokenizado y dividido en los grupos de muestra y prueba es un paso indispensable para que pueda usarse en aplicaciones de machine learning. Sin embargo, aún faltaba un detalle importante para facilitarle la tarea a los modelos y es que el corpus seguía siendo un conjunto de texto y los modelos operan con valores numéricos organizados en vectores, por esto se debía realizar la vectorización del corpus de entrenamiento. Para este paso utilizamos 2 modelos de representación: la bolsa de palabras y el TF-IDF.

A nivel teórico lo que realiza los modelos de representación es construir una tabla en la que cada fila es un documento del corpus, en este caso una descripción por fila y cada columna es un token del vocabulario de palabras que se tiene en el corpus.

| | bolsa | cancha | comer | compramese | decisoo | ensuciar | fútbol | ganar | jugar | juro | noche | palato | pásticoo | suadero | taco | vida |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | | | | | | | | | | | | | | | | |
| Document 2 | | | | | | | | | | | | | | | | |
| Document 3 | | | | | | | | | | | | | | | | |

***Figura 5.3.2*** *Modelo de Representación*

*Fuente: Elaboración propia*

Para rellenar esta tabla se cuenta con la opción de la bolsa de palabras que lo que hace es contar la cantidad de veces que aparece un token en cada una de las descripciones de los tickets.

| | bolsa | cancha | comer | compramese | decisoo | ensuciar | fútbol | ganar | jugar | juro | noche | palato | pásticoo | suadero | taco | vida |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| Document 2 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| Document 3 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |

***Figura 5.3.3*** *Modelo de Bag of Words*

*Fuente: Elaboración propia*

El modelo TF-IDF se diferencia del anterior en que este toma en cuenta la influencia de todo el corpus en cada una de las descripciones.

Es por esto que para tener en cuenta el hecho de que algunas palabras son generalmente más populares que otras, el valor de cada token aumenta proporcionalmente al número de veces que aparece una palabra en el documento, pero se compensa con la frecuencia de la palabra en la colección de documentos.

| | bolsa | cancha | comer | compramee | decicios | ensuciar | fútbol | ganar | jugar | juro | noche | plato | plástico | sueadero | taco | vida |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Document 1 | 0.00 | 0.00 | 0.33 | 0.00 | 0.25 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | v | 0.75 | 0.20 | 0.33 |
| Document 2 | 0.43 | 0.00 | 0.00 | 0.00 | 0.32 | 0.43 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.43 | 0.43 | 0.32 | 0.25 | 0.00 |
| Document 3 | 0.00 | 0.40 | 0.00 | 0.40 | 0.00 | 0.00 | 0.40 | 0.40 | 0.40 | 0.00 | 0.40 | 0.00 | 0.00 | 0.00 | 0.23 | 0.00 |

***Figura 5.3.4*** *Representación Modelo TF-IDF*

*Fuente: Elaboración propia*

### 5.3.3 Aplicar modelos de clasificación con cada uno de los modelos de representación.

Una vez listos ambos modelos de representación pasamos a aplicar las distintas técnicas que permitieran realizar clasificación. Este proceso se hizo tanto con las representaciones de bolsa de palabras como con TF-IDF. Los siguientes son los modelos de clasificación aplicados: Naive Bayes, Regresión logística, K-vecinos, Árbol de decisión, Máquina de soporte vectorial, Random forest, Boosting o Bagging, Red neuronal.

### 5.3.4 Cálculo de métricas de modelos

Para cada modelo de clasificación se calcularon las siguientes métricas: Accuracy, precisión, Recall, F1 Score, Matriz de confusión y para los modelos de regresión se utilizó

la siguiente métrica: $R^2$ (R cuadrado). Con estas se indagaron cuáles eran los modelos más precisos con la intención de continuar con la optimización de sus hiperparametros.

### 5.3.5 Selección de mejores modelos

***Tabla 3*** *Métricas para los modelos de clasificación con bolsa de palabras*

| MODELO BOLSA DE PALABRAS | | | | | |
|---|---|---|---|---|---|
| **Técnicas de clasificación de Aprendizaje Automático** | **Métricas** | | | | |
| | **Métricas de clasificación** | | | | **Métricas de Regresión** |
| | **Accuracy** | **Precisión** | **Recall** | **F1-score** | **R2** |
| **Regresión logística** | | | | | 0,8075 |
| **Naive bayes** | 0,7890 | 0,7926 | 0,7890 | 0,7858 | |
| **K-vecinos** | 0,6526 | 0,6579 | 0,6526 | 0,6548 | |
| **Árboles de decisión** | 0,7559 | 0,7545 | 0,7559 | 0,7549 | |
| **Máquinas de soporte vectorial** | 0,7913 | 0,7905 | 0,7913 | 0,7907 | |
| **Random forest** | 0,7758 | 0,7756 | 0,7758 | 0,7736 | |
| **Boosting o Bagging** | 0,6477 | 0,6800 | 0,6477 | 0,6278 | |
| **Redes neuronales** | 0,7631 | 0,7629 | 0,7631 | 0,7628 | |

Fuente: Elaboración propia

Para la bolsa de palabras se observa que la regresión logística fue el que arrojó los mejores resultados.

Para el modelo TF-IDF la regresión logística y la máquina de soporte vectorial eran las que arrojaban las mejores métricas.

***Tabla 4*** *Métricas para los modelos de clasificación con TF-IDF*

| MODELO TF-IDF | | | | | |
|---|---|---|---|---|---|
| **Técnicas de clasificación de Aprendizaje Automático** | **Métricas** | | | | |
| | **Métricas de clasificación** | | | | **Métricas de Regresión** |
| | **Accuracy** | **Precisión** | **Recall** | **F1-score** | **R2** |
| **Regresión logística** | | | | | 0,8059 |
| **Naive bayes** | 0,7589 | 0,7889 | 0,7589 | 0,7358 | |

| | | | | | |
|---|---|---|---|---|---|
| **K-vecinos** | 0,5832 | 0,7285 | 0,7285 | 0,5404 | |
| **Árboles de decisión** | 0,7419 | 0,7414 | 0,7419 | 0,7414 | |
| **Máquinas de soporte vectorial** | 0,8127 | 0,8124 | 0,8127 | 0,8115 | |
| **Random forest** | 0,7644 | 0,7634 | 0,7644 | 0,7615 | |
| **Boosting o Bagging** | 0,6704 | 0,6704 | 0,6704 | 0,6482 | |
| **Redes neuronales** | 0,7665 | 0,7665 | 0,7665 | 0,7665 | |

Fuente: Elaboración propia

Finalmente, se llevó a cabo una validación de las variaciones en las métricas de los distintos modelos cuando se intercambiada del modelo de representación de bolsa de palabras al de TF-IDF. Algo que se observó como interesante fue que en la mayoría de las métricas los modelos desmejoraban la clasificación cuando se intercambiaban las representaciones de bolsa de palabras a TF-IDF. En algunos en donde no se observó este comportamiento y por el contrario mejoraron con el TF-IDF fue en la máquina de soporte vectorial y en las redes neuronales.

***Tabla 5*** *Variaciones en las métricas de los modelos de clasificación cuando se pasa a bolsa de palabras a TF-IDF*

| **VARIACIONES** | | | | | |
|---|---|---|---|---|---|
| **Técnicas de clasificación de Aprendizaje Automático** | **Métricas** | | | | |
| | **Métricas de clasificación** | | | | **Métricas de Regresión** |
| | **Accuracy** | **Precisión** | **Recall** | **F1-score** | **R2** |
| **Regresión logística** | | | | | -0,0016 |
| **Naive bayes** | -0,0301 | -0,0037 | -0,0301 | -0,0500 | |
| **K-vecinos** | -0,0694 | 0,0706 | 0,0759 | -0,1144 | |
| **Árboles de decisión** | -0,0140 | -0,0131 | -0,0140 | -0,0135 | |
| **Máquinas de soporte vectorial** | 0,0214 | 0,0219 | 0,0214 | 0,0208 | |
| **Random forest** | -0,0114 | -0,0122 | -0,0114 | -0,0121 | |
| **Boosting o Baggin** | 0,0227 | -0,0096 | 0,0227 | 0,0204 | |
| **Redes neuronales** | 0,0034 | 0,0036 | 0,0034 | 0,0037 | |

Fuente: Elaboración propia

Una vez se analizaron los datos de las 3 tablas se tomó la decisión de continuar el proceso con la regresión logística y la máquina de soporte vectorial de TF-IDF. Estas decisiones se tomaron porque eran los modelos con las métricas más altas y porque en la regresión logística con bolsa de palabras y con TF-IDF la diferencia era casi imperceptible a nivel de sus métricas.

### 5.3.6 Optimización de hiperparametros

Una vez seleccionados los dos modelos se siguió con un proceso de optimización de hiperparametros; para esto se apoyó en los módulos Pipeline y GridSearchCV de sklearn para comenzar a buscar los parámetros que optimizaban las métricas y la capacidad de clasificación tanto de la regresión logística como de la máquina de soporte vectorial con TF-IDF.

Con la implementación de este proceso al final se logró conseguir una leve mejoría en las métricas de ambos modelos. Dejándolos a ambos casi que con la misma exactitud.

```
# evaluar el modelo mejor sintonizado en el conjunto de datos de prueba

best_lr_test_score = gs_lr.score(test_corpus, test_label_names)
print('Test Accuracy :', best_lr_test_score)
```

```
Test Accuracy : 0.8203605514316012
```

***Figura 5.3.5*** *Métrica de Accuracy de la regresión logística cuando se realiza optimización de sus hiperparametros*

*Fuente: Elaboración propia*

```
# evaluar el modelo mejor sintonizado en el conjunto de datos de prueba

best_svm_test_score = gs_svm.score(test_corpus, test_label_names)
print('Test Accuracy :', best_svm_test_score)
```

```
Test Accuracy : 0.824390243902439
```

***Figura 5.3.6*** *Métrica de la máquina de soporte vectorial cuando se realiza la optimización de sus hiperparametros*

*Fuente: Elaboración propia*

## 5.4 EVALUACIÓN

**Objetivo específico 3:** Seleccionar herramientas tecnológicas adecuadas para el desarrollo del prototipo.

### 5.4.1 Evaluar predicción de los modelos optimizados con los datos de prueba.

Una vez finalizada la fase de modelado, se contaba con dos modelos optimizados y con métricas por encima del 80% que podían ser evaluados con los datos de prueba.

En la siguiente figura se observa una parte de las 4715 descripciones de prueba con su respectiva etiqueta “category” que fue asignada manualmente por el personal de la entidad financiera, posteriormente en la columna Predicted Name se observa la clasificación que le asigna el modelo de regresión logística optimizado con TF-IDF a cada descripción del corpus de prueba.

| | Descripcion_limpia | Category | Predicted Name | Predicted Confidence |
|---|---|---|---|---|
| 4450 | buen dia envia solicitud correccion pago operacion cliente fabiola bermudez casierra c c error recaudo numero credito adjunto carta solicitud soport pago gracias | Medio | Medio | 0.985052 |
| 7444 | bueno dia favor rechazar delegar insatncia permit tomar dar continuidad proceso | Alto | Alto | 0.982927 |
| 7050 | bueno dia carpeta siniestro encuntra reclamacion seguro fp lui iriart salgado cc | Bajo | Bajo | 0.994699 |
| 10313 | buena tardes solicto colaboracion aplicar pago credito client diego marin agudelo cc cobro comis juridica acordada abogado adjunta pago correo | Medio | Medio | 0.910490 |
| 12922 | buen dia adjunto documento client monica mafla mora credito condonacion conceptos | Bajo | Bajo | 0.962882 |
| 8764 | solicito favor habilitar coordinadora servicio neiva altico gina lorena salgado roja usuario gsalgado partir hoy enero debido retoma cargo despu reemplazar vacacion c r c region huila deshabilitar... | Medio | Medio | 0.939832 |
| 6575 | buen dia solicito colaboracion actualizacion correo usuario afburbano | Medio | Alto | 0.801614 |
| 5403 | buen dia adjunta report disfrut vacaciones | Medio | Medio | 0.978253 |
| 476 | desbloqueo clave datacredito | Bajo | Bajo | 0.974783 |
| 10950 | buen dia solicito cambio perfil cajero princip oswal fernando rozo c c orozo cajera princip laura daniela ceron c c ldceron incapacitada dia gracia | Medio | Medio | 0.994858 |
| 1139 | bueno dias adjunto report client rosmeri montenegro parodi cc fin validar categoria puesto aparec recurrent ant pandemia elit plus anterior sujetro nuevo cambio sarc agradezco colaboracion afectac... | Medio | Medio | 0.969271 |
| 12063 | cordial saludo favor autorizar poliza plan hipertens tiroid rosa lobon gil cc quedo atenta autorizacion gracia | Alto | Alto | 0.936694 |
| 4245 | solicita revisar cartera digit analista credito permit conexion usuario registrado | Bajo | Bajo | 0.980520 |
| 884 | buen dia adjunta report cambio cargo | Medio | Medio | 0.986997 |
| 11515 | bueno dias favor enviar provis gracia | Medio | Medio | 0.957410 |
| 10558 | buen dia solicito habilitar luz arelli valencia usuario lvalencia coordinadora servicios martha luci jurado usuario mjurado cajera princip juliana jaramillo usuario jjaramil asesora servicio pda s... | Medio | Medio | 0.974485 |
| 99 | buen dia favor consutar lista restrictiva titular ampliacion fabian andr gomez identificado cc mucha gracias quedo atenta | Medio | Medio | 0.989715 |
| 3467 | buena tardes solicito apoyo habilitar coordinador servicio popayan ciudad jardin usuario apgonzale angela patricia gonzalez santand deshabilitar usuario yramirezt yuri katerin ramirez tobar estara... | Medio | Medio | 0.950571 |
| 9214 | buena tarde solicita favor deshabilitar clave cajero princip estara vacacion asignar clave bakcup boveda asesor servicios adjunto envio formato solicitud | Medio | Medio | 0.928050 |
| 9419 | bueno dia envio certificado libertad viabilidad credito gracia | Alto | Alto | 0.808107 |
| 14010 | bueno dias permito solicitar habilitar cajero princip lui miguel mayorga asesora servicio magda milena martinez perfil asesora servicio cajera auxiliar mucha gracia | Medio | Medio | 0.981531 |
| 6325 | celular conecta vpn celular analista credito carlo antonio casanova c c usuario ccasanova celular | Bajo | Bajo | 0.994571 |
| 5226 | buen dia solicito colaboracion autorizacion cuarto sistema part prosegur validar disponibilidad ampliacion sistema seguridad techo oficina libertador robo presntado agencia quedo atenta | Medio | Medio | 0.975414 |
| 1984 | solicita prueba panico | Alto | Alto | 0.994615 |
| 290 | bueno dias solicito colaboracion validacion usuario rcastaneda acceso adminfo genera do errores horario permitido error integracion mucha gracias adjunto pantallazo | Bajo | Bajo | 0.705554 |

***Figura 5.4.1*** *Predicción de la prioridad en los datos de testeo con el modelo de regresión logística.*

*Fuente: Elaboración propia*

En la figura 5.8 se puede observar que el 19.4% de los tickets de prueba fueron clasificados con una categoría diferente por parte del modelo a comparación de lo asignado manualmente por el analista. Estos tickets deberían ser analizados para concluir si se trata de un error del modelo o de un error humano del analista que clasifico manualmente los tickets en mención.

```
Comparación
False     883
True     3832
dtype: int64
```

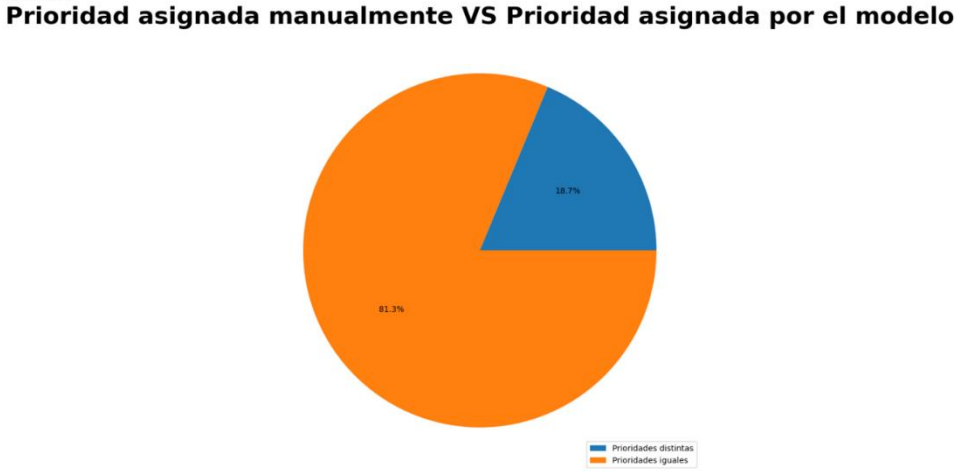



***Figura 5.4.2*** *Prioridad asignada manualmente vs prioridad asignada por modelo.*

*Fuente: Elaboración propia*

A continuación, se comparte el mismo ejercicio, pero con el modelo de máquina de soporte vectorial optimizado con TF-IDF:

| | Descripcion_limpia | Category | Predicted Name | Predicted Confidence |
|---|---|---|---|---|
| 4450 | buen dia envia solicitud correccion pago operacion cliente fabiola bermudez casierra c c error recaudo numero credito adjunto carta solicitud soport pago gracias | Medio | Medio | 1.119387 |
| 7444 | bueno dia favor rechazar delegar insatncia permit tomar dar continuidad proceso | Alto | Alto | 1.203841 |
| 7050 | bueno dia carpeta siniestro encuntra reclamacion seguro fp lui iriart salgado cc | Bajo | Bajo | 1.255771 |
| 10313 | buena tardes solicto colaboracion aplicar pago credito client diego marin agudelo cc cobro comis juridica acordada abogado adjunta pago correo | Medio | Medio | 0.678883 |
| 12922 | buen dia adjunto documento client monica mafla mora credito condonacion conceptos | Bajo | Bajo | 0.922564 |
| 8764 | solicito favor habilitar coordinadora servicio neiva altico gina lorena salgado roja usuario gsalgado partir hoy enero debido retoma cargo despu reemplazar vacacion c r c region huila deshabilitar... | Medio | Medio | 0.737344 |
| 6575 | buen dia solicito colaboracion actualizacion correo usuario afburbano | Medio | Alto | 0.258155 |
| 5403 | buen dia adjunta report disfrut vacaciones | Medio | Medio | 0.957791 |
| 476 | desbloqueo clave datacredito | Bajo | Bajo | 1.069401 |
| 10950 | buen dia solicito cambio perfil cajero princip oswal fernando rozo c c orozo cajera princip laura daniela ceron c c ldceron incapacitada dia gracia | Medio | Medio | 1.388007 |
| 1139 | bueno dias adjunto report client rosmeri montenegro parodi cc fin validar categoria puesto aparec recurrent ant pandemia elit plus anterior sujetro nuevo cambio sarc agradezco colaboracion afectac... | Medio | Medio | 1.024909 |
| 12063 | cordial saludo favor autorizar poliza plan hipertens tiroid rosa lobon gil cc quedo atenta autorizacion gracia | Alto | Alto | 0.918321 |
| 4245 | solicita revisar cartera digit analista credito permit conexion usuario registrado | Bajo | Bajo | 1.004977 |
| 884 | buen dia adjunta report cambio cargo | Medio | Medio | 0.970941 |
| 11515 | bueno dias favor enviar provis gracia | Medio | Medio | 0.981435 |
| 10558 | buen dia solicito habilitar luz arelli valencia usuario lvalencia coordinadora servicios martha luci jurado usuario mjurado cajera princip juliana jaramillo usuario jjaramil asesora servicio pda s... | Medio | Medio | 0.926440 |
| 99 | buen dia favor consutar lista restrictiva titular ampliacion fabian andr gomez identificado cc mucha gracias quedo atenta | Medio | Medio | 1.258077 |
| 3467 | buena tardes solicito apoyo habilitar coordinador servicio popayan ciudad jardin usuario apgonzale angela patricia gonzalez santand deshabilitar usuario yramirezt yuri katerin ramirez tobar estara... | Medio | Medio | 0.828089 |
| 9214 | buena tarde solicita favor deshabilitar clave cajero princip estara vacacion asignar clave bakcup boveda asesor servicios adjunto envio formato solicitud | Medio | Medio | 0.624532 |
| 9419 | bueno dia envio certificado libertad viabilidad credito gracia | Alto | Alto | 0.505366 |
| 14010 | bueno dias permito solicitar habilitar cajero princip lui miguel mayorga asesora servicio magda milena martinez perfil asesora servicio cajera auxiliar mucha gracia | Medio | Medio | 1.050605 |
| 6325 | celular conecta vpn celular analista credito carlo antonio casanova c c usuario ccasanova celular | Bajo | Bajo | 1.463462 |
| 5226 | buen dia solicito colaboracion autorizacion cuarto sistema part prosegur validar disponibilidad ampliacion sistema seguridad techo oficina libertador robo presntado agencia quedo atenta | Medio | Medio | 1.131851 |
| 1984 | solicita prueba panico | Alto | Alto | 1.548528 |
| 290 | bueno dias solicito colaboracion validacion usuario rcastaneda acceso adminfo genera do errores horario permitido error integracion mucha gracias adjunto pantallazo | Bajo | Bajo | 0.331075 |

***Figura 5.4.3*** *Predicción de la prioridad en los datos de testeo con el modelo de máquina de soporte vectorial*

*Fuente: Elaboración propia*

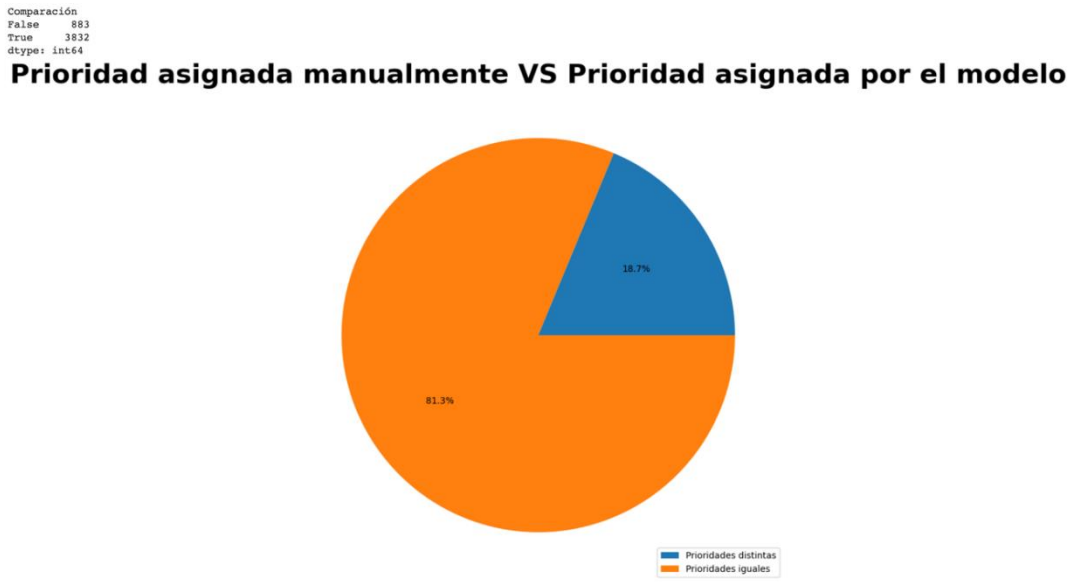



***Figura 5.4.4*** *Prioridad asignada manualmente vs prioridad asignada por modelo.*

*Fuente: Elaboración propia*

En la Figura 5.10 se logran evidenciar casi los mismos resultados que con la regresión logística. El 18.7% de los tickets de testeo fueron clasificados de forma distinta por el modelo a comparación de la clasificación dada manualmente por el analista.

Finalmente, se aclara porque los valores de la columna Predicted Confidence dan distinto cuando se aplican la regresión logística y la máquina de soporte vectorial. Esto se debe a un tema interpretativo de los modelos.

Para calcular la confianza prevista en la predicción se utilizaron las respectivas funciones disponibles en la documentación de cada modelo. Estas fueron predict_proba para la regresión logística y decision_function para la máquina de soporte vectorial. La diferencia entre predict_proba y decision_function depende del tipo de problema de aprendizaje automático que estábamos abordando.

En general, tanto predict_proba como decision_function son métodos utilizados en

modelos de clasificación para obtener información sobre la confianza o certeza de las predicciones del modelo, pero lo hacen de diferentes maneras.

predict_proba es un método que devuelve la probabilidad estimada de que una muestra pertenezca a una de las clases posibles. En otras palabras, devuelve un arreglo de probabilidades donde cada elemento representa la probabilidad de pertenencia a una clase específica. Este método se utiliza comúnmente en clasificadores probabilísticos como la Regresión Logística y los clasificadores basados en árboles, como Random Forests o Gradient Boosting.

Por otro lado, decision_function es un método que devuelve un valor numérico que representa la "distancia" o "confianza" de una muestra con respecto al hiperplano de decisión del clasificador. La interpretación exacta de este valor depende del algoritmo específico utilizado, pero en general, cuanto mayor sea el valor devuelto por decision_function, más segura será la clasificación positiva, y cuanto menor sea, más segura será la clasificación negativa. Este método se utiliza comúnmente en clasificadores basados en máquinas de vectores de soporte (SVM) y también en otros clasificadores lineales.

En resumen, mientras que predict_proba proporciona probabilidades de pertenencia a cada clase, decision_function proporciona una medida numérica de la confianza o certeza de la clasificación del modelo. La elección entre uno u otro depende de las necesidades específicas del problema y del algoritmo de clasificación utilizado.

# 6. VALIDACIÓN PROPUESTA PROTOTIPO PARA ANÁLISIS Y CLASIFICACIÓN DE INCIDENTES EN UNA ENTIDAD FINANCIERA UTILIZANDO NPL

La evaluación del modelo se centró en validar el prototipo desarrollado para el análisis y clasificación de incidentes en una entidad financiera utilizando Procesamiento del Lenguaje Natural (PLN). Con el fin de respaldar esta validación, se optó por utilizar la modalidad de encuesta a expertos, considerando tres factores principales: pertinencia, coherencia y aplicación.

Con base en estos factores, se elaboraron una serie de preguntas relevantes que se enumeran en la siguiente tabla. Estas preguntas serían evaluadas utilizando los criterios que se detallan a continuación [27]:

- No se cumple (menor a 3)
- Se cumple insatisfactoriamente (3.0 a 3.5)
- Se cumple aceptablemente (3.6 a 4.0)
- Se cumple en alto grado (4.1 a 4.5)
- Se cumple plenamente (4.6 a 5)

***Tabla 6*** *Diseño de encuestas de validación*

| **Diseño de encuesta de validación** | |
|---|---|
| **Factor** | **Preguntas** |
| **Pertinencia** | ¿Considera que mediante esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL se está respondiendo a las necesidades de las partes interesadas, como el Auditor Interno, los Auditores de Riesgos, los Auditores Senior y los Auditores? |

| | |
|---|---|
| | ¿Considera que esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es relevante y tiene aplicabilidad en el ámbito práctico? |
| | ¿La propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es adecuada para abordar el problema planteado? |
| **Coherencia** | ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es coherente con la formulación del problema planteado? |
| | ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL está interconectada y se complementa de manera lógica y coherente con todas las partes del proyecto? |
| | ¿Considera que la formulación del problema de investigación, los objetivos planteados, la metodología propuesta, los resultados esperados y las conclusiones se encuentran alineados de manera coherente y se sustentan mutuamente en el contexto del proyecto propuesto? |
| | ¿Considera que el proyecto de grado presenta una estructura clara y ordenada, estableciendo una relación directa entre el problema de investigación planteado y los objetivos propuestos? Además, ¿opina que los métodos y técnicas seleccionados son apropiados para abordar la problemática planteada? |
| **Aplicación** | ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL podría ser aplicada en otras áreas de la misma entidad? |
| | ¿Cree que esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL puede ser aplicada en otras actividades dentro del campo de la auditoría interna? |
| | ¿Considera usted que el prototipo planteado tiene el potencial de ser aplicado en el futuro para generar propuestas que faciliten la toma de decisiones fundamentadas en datos, aprovechando el crecimiento en el uso y la disponibilidad de información? |

Fuente: Elaboración propia

Se realizó esta encuesta a un total de 11 colaboradores de Entidad Financiera del área de auditoría Interna, quienes desempeñan un papel directo en el proceso y cuentan con una amplia experiencia en sus cargos, lo que los convierte en expertos en la materia.

Una vez definidas las preguntas y los criterios de evaluación, se procedió a crear la encuesta utilizando la plataforma virtual "Office 365 Forms". Esta plataforma nos permitió generar una encuesta en línea fácil de responder y compartir con los encuestados. Los resultados obtenidos podrán ser visualizados en el **ANEXO F**.

## 7. RESULTADOS OBTENIDOS

Partiendo del diseño y la construcción de la encuesta a los colaboradores del área de auditoría interna de entidad financiera, se resaltan los siguientes resultados por cada pregunta evaluada en la figura 7.1.

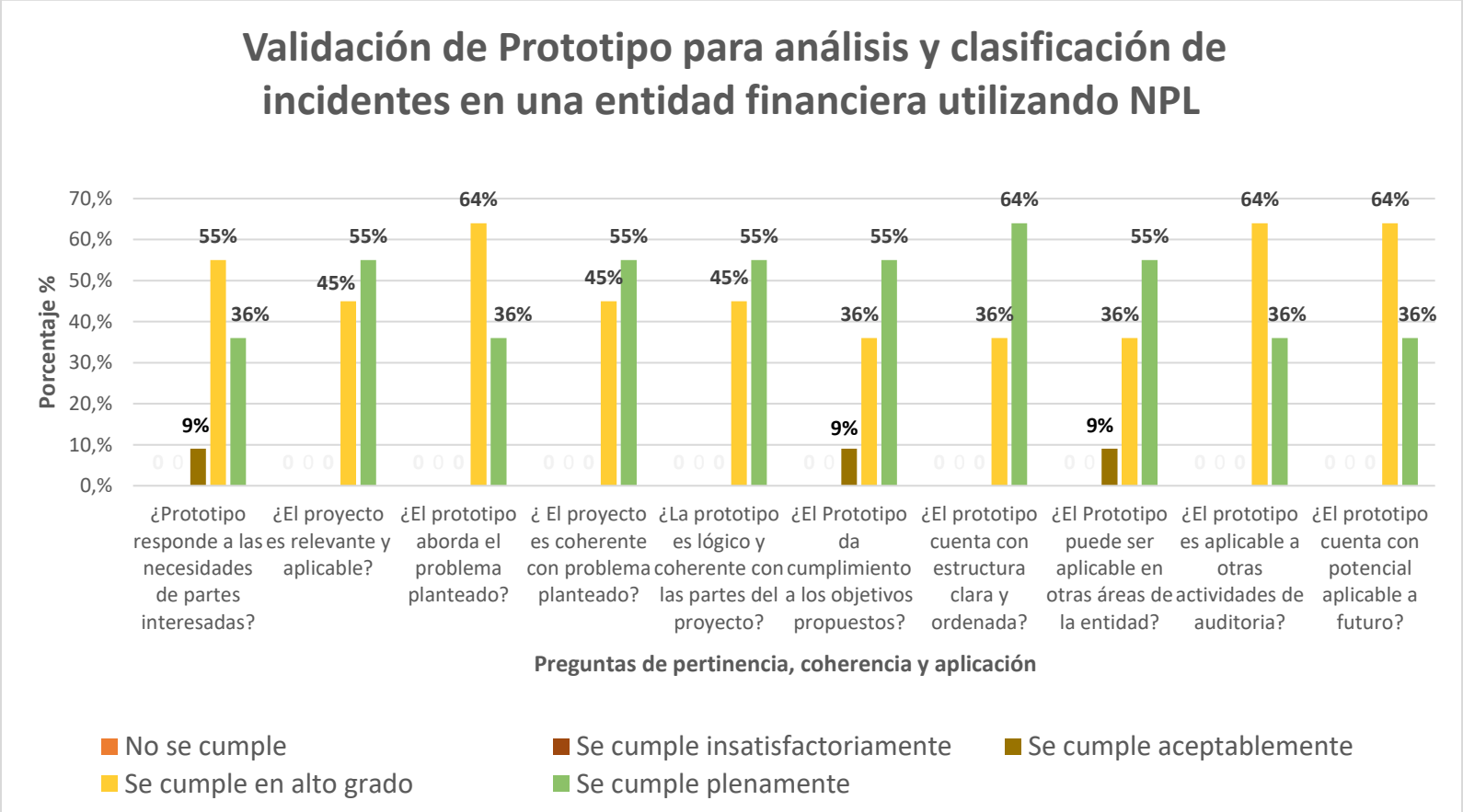



***Figura 5.4.1*** *Resultados de la validación de la propuesta*

*Fuente: Elaboración propia*

En reunión concertada con el grupo de colaboradores del área de auditoría interna se socializo la propuesta de prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL, para ello se consideraron criterios evaluativos específicos y sensibles, como son pertinencia, coherencia y aplicación. A continuación, se detalla cada uno de los criterios asociados a la validación de la propuesta:

**Pertinencia**

La propuesta de prototipo para el análisis y clasificación de incidentes en una entidad financiera utilizando PLN tiene como objetivo satisfacer la necesidad de clasificar con precisión los incidentes para cada uno de los procesos que son objeto de auditoría en esta entidad financiera. Según los resultados de la encuesta, el 55% de los encuestados afirman que este prototipo responde a las necesidades de las partes interesadas en alto grado, el 36% considera que cumple completamente con estas necesidades y solo el 9% indica que se cumple aceptablemente.

De manera similar, en cuanto a la relevancia y aplicabilidad del proyecto, el 55% de los entrevistados indican que se cumple completamente, el 45% afirma que se cumple en alto grado. Además, en respuesta a la pregunta de si el prototipo aborda el problema planteado, el 64% de los encuestados indica que se cumple en alto grado y el 36% considera que se cumple completamente.

En resumen, según la opinión de los expertos, la propuesta tiene una alta pertinencia, ya que, en promedio, las respuestas indican un cumplimiento en alto grado, con una calificación promedio de 4.1 a 4.5 en las preguntas evaluadas.

De esta manera podemos afirmar de acuerdo con el juicio de los expertos que la solución propuesta es adecuada y tiene un grado de adecuación satisfactorio para resolver el

problema planteado y cumplir con los objetivos establecidos.

**Coherencia**

Teniendo en cuenta que el objetivo principal de la coherencia del proyecto es garantizar una relación lógica y consistente entre todas las partes y componentes de este, se realizaron una serie de preguntas a los encuestados para evaluar el nivel de coherencia con respecto a la propuesta presentada.

En respuesta a la pregunta sobre si el proyecto es coherente con el problema planteado, el 55% de los encuestados indicaron que se cumple plenamente, el 45% afirmó que se cumple en alto grado. Por otro lado, con relación a si el prototipo es lógico y coherente con las partes del proyecto, el 55% de los entrevistados indicaron que se cumple plenamente, y el 45% consideró que cumple en alto grado.

En cuanto a la pregunta sobre si el prototipo cumple con los objetivos propuestos, el 55% de los encuestados afirmó que se cumple plenamente, el 36% indicó que se cumple en alto grado, y solo el 9% consideró que se cumple aceptablemente. Con relación a la claridad y organización de la estructura del prototipo, el 64% de los entrevistados indicó que cumple plenamente y el 36% consideró que cumple en alto grado.

Según las respuestas proporcionadas por los expertos, el análisis consolidado indica que el proyecto cumple plenamente con una calificación entre 4.6 y 5 en términos de coherencia. Esto confirma que todas las partes y componentes del proyecto están alineados y trabajan de manera lógica y consistente para lograr los objetivos establecidos. Al lograr esta coherencia, se contribuye a mejorar la eficacia, eficiencia y éxito general del proyecto.

Esta evaluación positiva de la coherencia refleja que las diferentes acciones y resultados

del proyecto se encuentran en armonía y se complementan entre sí. Esto asegura que el proyecto avance en la dirección correcta y que no existan contradicciones ni brechas que puedan afectar su implementación exitosa a futuro.

**Aplicación**

Se ha validado la propuesta de prototipo para el análisis y clasificación de incidentes en una entidad financiera utilizando NP en términos de su aplicabilidad y los beneficios que conlleva. Al realizar preguntas relacionadas con la aplicabilidad del prototipo en otras áreas de la entidad, el 55% de los encuestados respondió que cumple plenamente, el 36% en alto grado y el 9% lo consideró aceptable. En cuanto a su aplicabilidad en otras actividades de auditoría, el 64% de los entrevistados indicó que cumple en alto grado y el 36% lo consideró plenamente aplicable. Con respecto al potencial de aplicación futura del prototipo, el 64% de los encuestados indicó que cumple en alto grado y el 36% lo consideró plenamente aplicable.

Posteriormente, y con base en la validación de la aplicación del prototipo a través del juicio de expertos, que está relacionado con su implementación práctica y efectiva en el contexto real, se logró determinar que, en términos de su nivel de aplicación, obtuvo un promedio de cumplimiento en alto grado, con una calificación entre 4.1 y 4.5. Esto indica que es posible llevar a cabo las actividades y acciones planificadas en el proyecto de manera eficiente y efectiva, utilizando los recursos disponibles de manera adecuada.

# 8. IMPACTOS DEL PROYECTO

El proyecto tiene diversos impactos que vale la pena destacar. En primer lugar, se espera que la implementación del prototipo de análisis y clasificación de incidentes en la entidad financiera mejore significativamente la eficiencia y efectividad en la gestión de incidentes. Esto permitirá una detección temprana y una respuesta más rápida y precisa ante cualquier incidente que pueda surgir, lo que a su vez reducirá los tiempos de resolución y minimizará el impacto negativo en las operaciones.

Además, se espera que el prototipo proporcione una mayor capacidad de análisis y generación de informes, lo que facilitará la toma de decisiones basada en datos sólidos. Esto permitirá a la entidad financiera obtener una visión más clara de los incidentes ocurridos, identificar tendencias y patrones, y tomar medidas proactivas para prevenir futuros incidentes.

Otro impacto importante del proyecto es su potencial de aplicación en otras áreas de la entidad financiera y en actividades de auditoría. Esto permitirá ampliar los beneficios y la eficacia del prototipo en diferentes contextos y contribuirá a mejorar la seguridad y la gestión de riesgos en la organización en su conjunto.

En términos más amplios, el proyecto también podría tener un impacto positivo en la reputación y confianza de la entidad financiera, al demostrar su compromiso con la adopción de tecnologías innovadoras y mejores prácticas en la gestión de incidentes.

La gestión más eficiente de los incidentes puede ayudar a minimizar los costos asociados con interrupciones operativas, pérdida de datos o impactos negativos en la reputación. Al detectar y resolver los incidentes de manera oportuna, la entidad financiera puede evitar

costosos tiempos de inactividad o pérdidas financieras significativas.

El prototipo de análisis de incidentes puede contribuir a fortalecer la seguridad en la entidad financiera. Al identificar y clasificar rápidamente los incidentes, se pueden implementar medidas de seguridad adecuadas para prevenir futuros problemas y proteger la información sensible de la entidad y sus clientes.

Otro de los impactos a resaltar, sería las mejoras en la toma de decisiones estratégicas, esto debido a que la disponibilidad de datos y análisis más sólidos derivados del prototipo puede proporcionar a la alta dirección de la entidad financiera una visión más clara de los riesgos y desafíos asociados con los incidentes. Esto permite tomar decisiones estratégicas más informadas y desarrollar medidas de mitigación más efectivas.

## CONCLUSIONES

- ✓ Al final de todo el proceso se presentaron 2 modelos como prototipo funcional de aplicación para categorizar eficientemente los tickets de incidentes en una entidad financiera. Ambos modelos cuentan con métricas muy similares y su evaluación con los datos de testeo nos permitió evidenciar la alta capacidad que tienen para lograr la clasificación de los incidentes de acuerdo con su prioridad. Nuestra recomendación sobre cual usar radica sobre todo en el tema interpretativo ya que la regresión logística al manejar probabilidades se hace más fácil de comprender para todos los públicos que la máquina de soporte vectorial.
- ✓ Se determinaron las técnicas de procesamiento de lenguaje natural más idóneas para preparar los datos textuales para ser usados en una aplicación de machine

learning que buscaba categorizar por su prioridad los incidentes registrados en una entidad financiera.

- ✓ La implementación del prototipo tiene el potencial de generar impactos significativos en la entidad financiera. Estos incluyen mejoras en la gestión de incidentes, capacidad de análisis y generación de informes, satisfacción de los colaboradores, reducción de costos y fortalecimiento de la seguridad.
- ✓ Se utilizaron librerías como matplotlib como herramienta de visualización de la capacidad de clasificación de nuestro prototipo funcional. Esta herramienta permite identificar los puntos donde los modelos finales difieren en comparación con las clasificaciones asignadas por el analista de la entidad financiera, lo cual es un gran aporte para la futura revisión y mejora continua.
- ✓ De acuerdo con el juicio de expertos y considerando las variables de validación como pertinencia, coherencia y aplicación se puede indicar que el prototipo de análisis y clasificación de incidentes en la entidad financiera en promedio cumple en alto grado. Esto indica que existe una relación lógica y consistente entre las diferentes partes y componentes del proyecto. Además, las acciones planificadas pueden llevarse a cabo de manera eficiente y efectiva utilizando los recursos disponibles de manera adecuada.

# TRABAJOS FUTUROS

- ✓ Mejora y refinamiento del prototipo para el análisis y clasificación de incidentes en una entidad financiera aplicando mayor robustez a la etapa de preparación de los datos por medio de otras librerías que permitan realizar optimizaciones en la información mejorando de esta manera el rendimiento y precisión de la clasificación.
- ✓ Aplicabilidad del prototipo para análisis y clasificación de incidentes en sectores industriales utilizando NPL.
- ✓ Integración del prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL con sistemas existentes para lograr una mayor automatización y una mejor interoperabilidad, optimizando así la eficiencia en la gestión de incidentes.
- ✓ Evaluación de la efectividad del prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL a largo plazo para analizar su rendimiento, identificar áreas de mejora continua y asegurar que sigue siendo relevante y beneficioso para la entidad financiera.

# BIBLIOGRÁFIA


[1] IIA, “Global Perspectives and Insights Artificial Intelligence-Considerations for the Profession of Internal Auditing Special Edition,” pp. 1–9, 2017, [Online]. Available: www.theiia.org/gpi.

[2] I. Pedrosa, R. M. S. Laureano, and C. J. Costa, “Motivações dos auditores para o uso das Tecnologias de Informação na sua profissão: Aplicação aos Revisores Oficiais de Contas,” RISTI - Rev. Iber. Sist. e Tecnol. Inf., no. 15, pp. 101–118, 2015, doi: 10.17013/risti.15.101-118.

[3] D. M. Garces-Eslava, “Metodo de procesamiento de lenguaje natural y tecnicas de mineria de datos aplicadas a la clasificacion de incidentes informaticos,” Interfases, no. 12, pp. 11–29, 2019, doi: 10.26439/interfases2019.n012.4635.

[4] ISACA, COBIT 2019 MARCO DE REFERENCIA. Schaumburg, 2019. [Online]. Available: www.isaca.org

[5] Z. M. C. QUIÑONES, “Implementacion del servicio de gestión de incidencias aplicando ITIL V3, caso de estudio: financiera efectiva,” p. 114, 2016.

[6] D. Sarkar, Text Analytics with Python - A Practitioner’s Guide to Natural Language Processing. 2019.

[7] J. Ordóñez et al., “Modelo de procesamiento de lenguaje natural para detectar la tasa de éxito de un artículo sobre otro,” 2021, [Online]. Available: https://repository.icesi.edu.co/biblioteca_digital/handle/10906/89008

[8] B. Committee and Bank for International Settlements, “Basel Committee on Banking Supervision: The relationship between banking supervisors and banks’ external auditors,” no. January 2002, 2013.

[9] J. C. Díaz Pérez, “Pragmalingüística del disfemismo y la descortesía. Los actos de habla hostiles en los medios de comunicación virtual,” p. 517, 2012.

[10] montylingua :: a free, commonsense-enriched natural language understander. (s.f.). Welcome to alumni.media.mit.edu. http://alumni.media.mit.edu/~hugo/montylingua/

[11] Ali, A.-R. (2017b, 3 de mayo). Presentando el Natural Language Toolkit (NLTK). Code Envato Tuts+. https://code.tutsplus.com/es/tutorials/introducing-the-natural-language-toolkit-nltk--cms-28620

[12] M. Ángel and M. Ramírez, “Inteligencia artificial aplicada a la ley de protección de datos,” 2020.

[13] Copyright 2020 by Tutorials Point. (2020). TutorialsPoint Gensim.

[14] O. E. Material, “International Standard on Auditing 265 Communicating Deficiencies in Internal Control To Those Charged With Governance,” Control, vol. 610, no. December, pp. 7–11, 2009.

[15] R. Yzquierdo Herrera, “Miner{í}a de proceso como herramienta para la auditoria.,” Ciencias la Inf., vol. 44, no. 2, pp. 25–32, 2013, [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true%7B&%7Ddb=a9h%7B&%7DAN=100723751%7B&%7Dsite=ehost-live

[16] A.- Barral Rivada, R. Bautista Mesa, and H. -Molina Sanchez, “Aplicación de Normas Internacionales de Auditoría,” p. 229, 2013, [Online]. Available: http://www.ctcp.gov.co/puerta/athena/_files/docs/1472852232-7796.pdf

[17] "Cierre de auditorías de estados financieros 2021. Aspectos importantes". Herramientas y Cursos Online para Auditores | Auditool. https://www.auditool.org/blog/auditoria-externa/que-aspectos-se-deben-tener-en-cuenta-al-cierre-de-una-auditoria-de-informacion-financiera (accedido el 20 de mayo de 2023).

[18] Castello Ricardo J. (n.d.). Auditoría en entornos informáticos.

[19] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4171-4186).

[20] Zhang, Y., Yang, L., Wu, H., & Huang, D. (2021). Incorporating external knowledge to improve natural language processing. IEEE Transactions on Neural Networks and Learning Systems, 32(2), 480-492.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[22] Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning (pp. 160-167).

[23] Zipf, G. K. (1949). Human behavior and the principle of least effort: An introduction to human ecology. Addison-Wesley.

[24] 8 ejemplos comunes del procesamiento del lenguaje natural y su impacto en la comunicación". Tableau. https://www.tableau.com/es-es/learn/articles/natural-language-processing-examples (accedido el 10 de mayo de 2023).

[25] G. D. Buzai and C. A. Baxendale, “A e d e,” no. 1977, pp. 1–11, 2009.

[26] K. Wosińska et al., Correlations of neutral and charged particles in 40Ar- 58Ni reaction at 77 MeV/u, vol. 32, no. 1. 2007. doi: 10.1140/epja/i2006-10279-1.

[27] C. E. N. Pr, C. Institucionales, E. N. Un, P. Tecnol, and G. D. E. L. Sena, “Cambios en prácticas institucionales en un programa tecnológico del sena 1,” pp. 1–98, 2015

# ANEXOS

## Anexo A. Tipos de herramientas y lenguajes de programación empleado para NPL.

| Lenguaje de programación | Nombre de la herramienta | Link de la herramienta | Licencia |
|---|---|---|---|
| **.NET framework** | Antelope framework | https://www.3ds.com/products-services/netvibes/products/proxem/ | Free |
| | NooJ (Basado en INTEX) | http://www.nooj4nlp.net | Free |
| | Rosette | https://www.basistech.com | Free |
| **C y C#** | Antelope framework | https://www.3ds.com/products-services/netvibes/products/proxem/ | Free |
| | Ellogon | https://www.ellogon.org | Free |
| **C++** | Apertium | https://wiki.apertium.org/wiki/Main_Page | Free |
| | DELPH-IN | https://github.com/delph-in/docs/wiki/ | Free |
| | Distinguo | https://ultralingua.com/en/semantic-search.htm | Free |
| | Ellogon | https://www.ellogon.org | Free |
| | FreeLing | https://nlp.lsi.upc.edu/freeling/ | Free |

| | | | |
|---|---|---|---|
| | Rasp | http://www.informatics.susx.ac.uk/research/groups/nlp/rasp/index.html | Free |
| | Rosette | https://www.basistech.com | Free |
| | UIMA | https://uima.apache.org/index.html | Free |
| | UniteX | https://unitexgramlab.org | Free |
| | VisualText | http://www.textanalysis.com | Free |
| | WebLab-project | http://weblab-project.org | Free |
| **JAVA** | Apertium | https://wiki.apertium.org/wiki/Main_Page | No free |
| | ClearTk | https://code.google.com/archive/p/cleartk/ | No free |
| | Factorie | https://code.google.com/archive/p/factorie/ | No free |
| | Graph Expression | https://code.google.com/p/graph-expression/ | No free |
| | IceNLP | https://sourceforge.net/projects/icenlp/ | Free |
| | Learning Based Java | https://cogcomp.seas.upenn.edu/page/software_view/11 | No free |
| | LingPipe | http://www.alias-i.com/lingpipe/index.html | Free |
| | LinguaStream | http://www.linguastream.org | Free |
| | Mallet | https://mimno.github.io/Mallet/index | Free |
| | MII nlp toolkit | https://mii.ucla.edu/nlp/ | Free |

| | | | |
|---|---|---|---|
| | MontyLingua | http://alumni.media.mit.edu/~hugo/montylingua/ | Free |
| | OpenNLP | https://opennlp.apache.org/index.html | Free |
| | Palladian | https://palladian.ai | Free |
| | The Dragon Toolkit | http://dragon.ischool.drexel.edu | Free |
| | UIMA | https://uima.apache.org/index.html | Free |
| **JavasScript** | Natural | https://github.com/NaturalNode/natural | Free |
| **LISP** | DEPLH-IN | https://github.com/delph-in/docs/wiki/ | Free |
| **Node JS** | Natural | https://github.com/NaturalNode/natural | Free |
| **Perl** | Treex | https://ufal.mff.cuni.cz/treex | No free |
| **Python** | MontyLingua | http://alumni.media.mit.edu/~hugo/montylingua/ | Free |
| | Natural Languaje Toolkit (NLTK) | https://www.nltk.org/Home | Free |
| | Silpa Indic Language Processing Toolkit | https://smc.org.in/silpa | Free |

| | | | |
|---|---|---|---|
| | Pytorch | https://pytorch.org | Free |
| | SpaCy | https://spacy.io | Free |
| | Gensim | https://radimrehurek.com/gensim/ | Free |
| **Scala** | ScalaNLP | http://www.scalanlp.org | Free |

**Anexo B. Descripción de herramientas para el PLN**

| Nombre del Programa | Descripción del Programa |
|---|---|
| MontyLingua | MontyLingua [10] es un programa gratuito, enriquecido por el sentido común, para entender el lenguaje natural del inglés de principio a fin. Introduzca un texto en inglés sin procesar en MontyLingua y el resultado será una interpretación semántica de ese texto. Es perfecto para la recuperación y extracción de información, el procesamiento de solicitudes y la respuesta a preguntas. A partir de frases en inglés, extrae tuplas de sujeto/verbo/objeto, extrae adjetivos, frases sustantivas y verbales, y extrae nombres de personas, lugares, eventos, fechas y horas, y otra forma de información semántica. |

| | |
|---|---|
| Natural Language Toolkik (NLTK) | NLTK es una biblioteca de código abierto que ofrece una amplia gama de funcionalidades para el procesamiento del lenguaje natural, incluyendo tareas como tokenización, etiquetado de partes del discurso, análisis sintáctico, desambiguación léxica, análisis de sentimientos, entre otros. También cuenta con una amplia colección de corpus y recursos lingüísticos que se pueden utilizar para entrenar modelos y realizar investigaciones en el campo del procesamiento del lenguaje natural [11].<br>_ Librería NLTK (Stop Words): Las palabras vacías son palabras de uso común, estas son generalmente ignoradas por el motor de búsqueda, estas palabras se eliminan con el fin de ahorrar espacio en las bases de datos y mejorar tiempos de procesamiento, el paquete NLTK es empleado para eliminar palabras vacías sobre el texto en Python, este contiene palabras vacías en muchos idiomas. El paquete de stop words se emplea para eliminar palabras vacías del texto en Python. |

| | |
|---|---|
| | **Lematización**<br>La lematización es una técnica importante en el procesamiento del lenguaje natural (NLP) que se utiliza para reducir las palabras a su forma base o raíz, también conocida como lemma. La lematización se utiliza para normalizar las palabras y reducir la variabilidad en los textos [23].<br>La lematización se utiliza para reducir la variabilidad en las palabras y para facilitar la búsqueda de información en los textos.<br>Una de las ventajas de la lematización es que puede mejorar la precisión del análisis de texto al reducir la variabilidad de las palabras. Además, la lematización puede ser útil en la identificación de palabras relacionadas y en la detección de relaciones semánticas entre las palabras [23].<br>**Tokenización**<br>La tokenización es el proceso de dividir un texto en palabras individuales. El NLTK proporciona varias opciones para la tokenización, como la tokenización de palabras, la tokenización de |

| | |
|---|---|
| | oraciones y la tokenización de tweets. La tokenización es un paso fundamental en el procesamiento del lenguaje natural, ya que permite trabajar con las palabras de forma individual [22]. |
| SpaCy | SpaCy es una biblioteca de Python que puede usar para crear aplicaciones de procesamiento de lenguaje natural (NLP). SpaCy ofrece modelos pre-entrenados [12] en varios idiomas. Esto, junto con su sintaxis clara, es perfecto para principiantes en el campo de la PNL.<br><br>Además, puede entrenar el modelo en un campo específico creando un nuevo modelo o volviendo a entrenar el modelo que proporcionó con sus propios datos. spaCy está diseñado para su uso en entornos de producción y, a diferencia de otras bibliotecas como TensorFlow, donde puede experimentar con diferentes arquitecturas de redes neuronales e implementar los últimos modelos de desarrollo, es completo y requiere procesamiento de lenguaje natural. Proporciona un marco para crear aplicaciones. Además, |

| | |
|---|---|
| | spaCy es eficiente en la CPU a pesar de usar un modelo de red neuronal. La velocidad y precisión del modelo spaCy es una de las mejores del mercado [12]. |
| Gensim | Gensim [13] está diseñado para procesar texto digital no estructurado ("texto sin formato") utilizando algoritmos de aprendizaje automático no supervisados. Los algoritmos de Gensim como Word2Vec, FastText, Latent Semantic Indexes (LSI, LSA, LsiModel) y Latent Dirichlet Allocation (LDA, LdaModel) capturan la estructura semántica de un documento al encontrar patrones estadísticos de co-ocurrencia en el corpus. automáticamente. Examine el documento de capacitación. Estos algoritmos no son monitoreados. Es decir, no se requiere intervención humana. Todo lo que necesita es un corpus de documentos de texto sin formato. |
| scikit-learn | Scikit-learn es una biblioteca de aprendizaje automático (machine learning) de código abierto para el lenguaje de programación Python. |

| | |
|---|---|
| | También conocida como sklearn, es una de las bibliotecas más utilizadas y populares en el campo del aprendizaje automático debido a su amplia gama de algoritmos y herramientas que facilitan el desarrollo de modelos predictivos y análisis de datos [6]. |

## Anexo C. Técnicas para el pre-procesamiento y la normalización del texto empleadas en el proyecto.

| TÉCNICAS | DESCRIPCIÓN |
|---|---|
| **Remoción de acentos.** | Consiste en eliminar los acentos gráficos de las palabras para evitar posibles discrepancias en el análisis.<br>La biblioteca unicodedata es una biblioteca estándar de Python que proporciona funciones para trabajar con caracteres Unicode. Esta biblioteca permite realizar diversas operaciones relacionadas con la manipulación y normalización de caracteres Unicode [6]. |
| **Stemming** | Es un proceso de reducción de las palabras a su forma base o raíz, lo que ayuda a unificar términos similares y simplificar el análisis.<br>PorterStemmer es un algoritmo de stemming desarrollado por Martin Porter en 1979. El stemming es el proceso de reducir una palabra a su forma raíz o base, eliminando sufijos y prefijos, con el objetivo de tratar palabras similares como variantes del mismo término [6]. |
| **Remoción de caracteres especiales** | Implica eliminar símbolos o caracteres que no aportan información relevante al texto [6]. |

| | |
|---|---|
| **Tokenización** | Se encarga de dividir el texto en unidades más pequeñas llamadas tokens, como palabras o frases, para facilitar el análisis y el modelado posterior.<br>Toktok es una biblioteca de Python que proporciona una forma sencilla y eficiente de realizar la tokenización de texto<br>La biblioteca toktok se basa en reglas heurísticas para separar el texto en tokens. Estas reglas están diseñadas para funcionar bien en varios idiomas y para manejar correctamente situaciones comunes, como puntuación, abreviaturas y números [6]. |
| **Stopwords** | Son palabras comunes en un lenguaje determinado y suelen tener poco valor semántico. Estas palabras se eliminan del texto para reducir el ruido y enfocarse en las palabras más relevantes.<br>En el contexto de NLTK, las stopwords son palabras comunes que se consideran menos relevantes para el análisis de texto debido a su alta frecuencia de aparición en un lenguaje determinado. Estas palabras a menudo carecen de significado contextual y no aportan mucha información para tareas como clasificación de texto, recuperación de información o análisis de sentimientos [6]. |

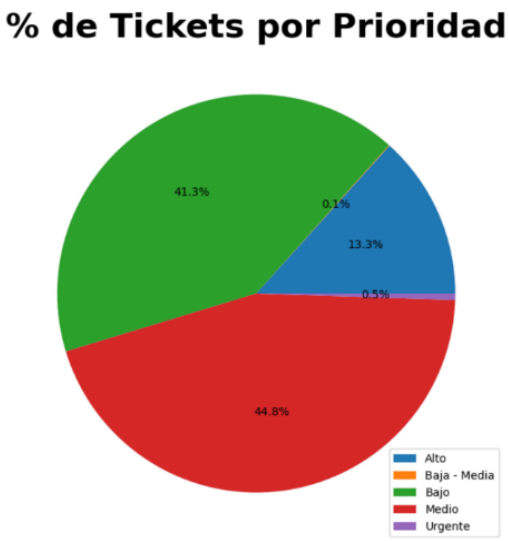
## Anexo D. Avances del NPL

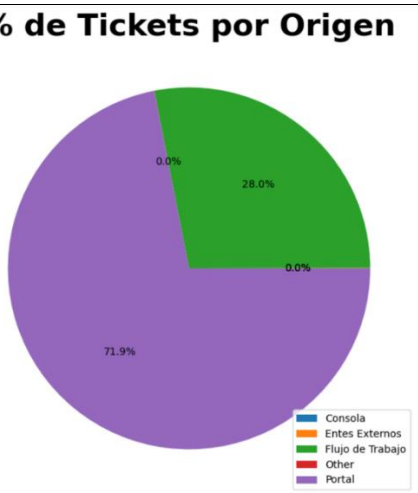
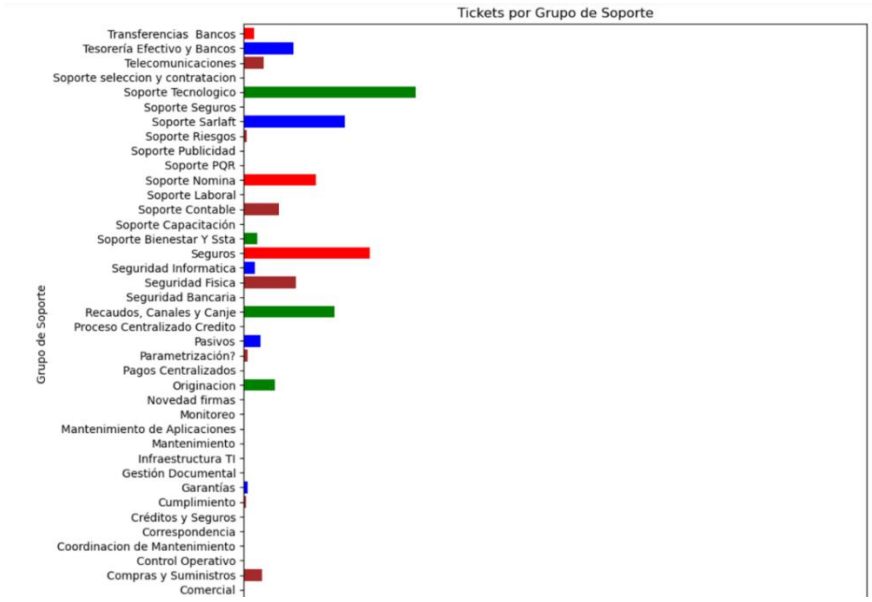
| Avances | Descripción |
|---|---|
| **Modelos de lenguaje basados en Transformer** | Los modelos de lenguaje basados en Transformer, como GPT-3 de OpenAI, han demostrado ser altamente efectivos en la generación de texto coherente y natural en una variedad de contextos, incluyendo la creación de contenido, la traducción automática y la conversación con chatbots. Estos modelos se basan en una arquitectura de redes neuronales llamada |

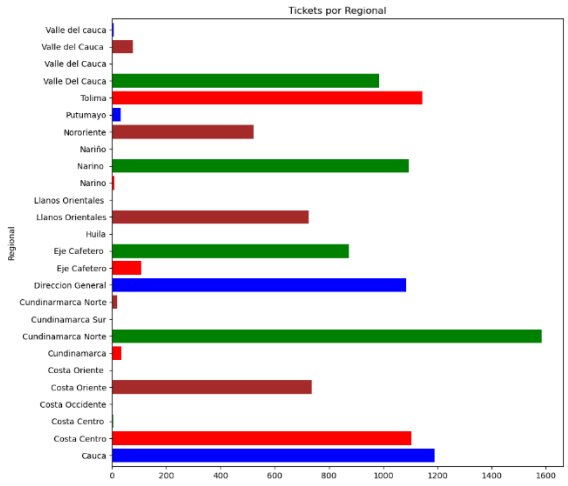
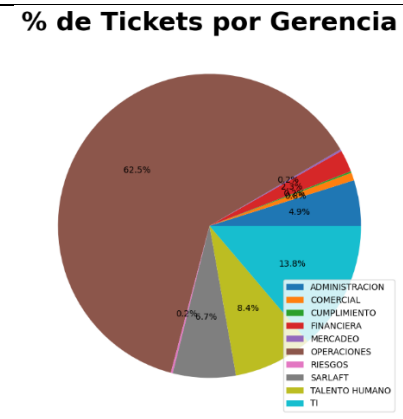
| | |
|---|---|
| | Transformer, que permite que el modelo tenga en cuenta el contexto del texto en su análisis [20]. |
| **Pre-entrenamiento de modelos de lenguaje** | El pre-entrenamiento de modelos de lenguaje, como BERT de Google, ha demostrado ser una técnica efectiva para mejorar la capacidad de los modelos de NLP para entender el lenguaje natural. En lugar de entrenar un modelo para una tarea específica, el pre-entrenamiento se enfoca en entrenar el modelo en una variedad de tareas de procesamiento del lenguaje natural, lo que le permite adquirir una comprensión más profunda del lenguaje natural. |
| **Incorporación de conocimiento externo** | Los modelos de NLP que incorporan conocimiento externo, como ontologías y bases de conocimiento, han demostrado ser efectivos en tareas como el análisis de texto y la extracción de información. Estos |

| | |
|---|---|
| | modelos pueden utilizar el conocimiento externo para mejorar su comprensión del lenguaje natural y hacer inferencias más precisas. |
| **Generación de lenguaje natural interactivo.** | Los sistemas de generación de lenguaje natural interactivo, como el modelo de conversación Meena de Google, han demostrado ser efectivos en la conversación humana y en la creación de chatbots más naturales y efectivos. Estos sistemas utilizan técnicas de generación de texto y aprendizaje por refuerzo para mejorar su capacidad de conversar con los usuarios [21]. |

## Anexo E. Análisis exploratorio de datos

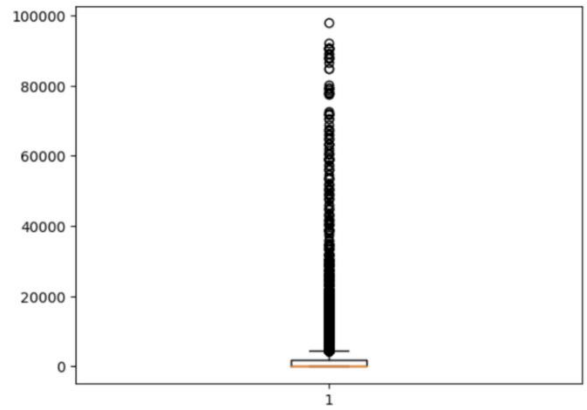
| | |
|---|---|
|  <br>Figura. % de Tickets por prioridad | Del total de tickets de enero de 2023 se puede observar que el 41.3% tienen una prioridad baja y el 44.8% es media, aunque es bajo el porcentaje de tickets que son de categoría alta se debe tener en cuenta que esta clasificación se realiza manualmente y un ticket mal clasificado se puede traducir en un perjuicio con riesgo de materialización que puede traer consecuencias de índole económico para la entidad financiera. |
| | Se puede observar que más del 70% de los tickets son escalados por medio |

| | |
|---|---|
|  <br>Figura. % de Tickets por Origen | del portal definido por la organización para este tipo de solicitudes lo cual evidencia una importante apropiación de las herramientas definidas por la entidad para la atención de incidentes. |
|  <br>Figura. Tickets por grupo de soporte | Los incidentes que son reportados por los colaboradores son escalados a los distintos grupos de soporte para su revisión y solución, en esta grafica se puede visualizar que la mayoría de tickets son de origen tecnológico por lo que se escalan a este grupo de soporte. |

| | |
|---|---|
|  <br>Figura. Tickets por Regional | En esta grafica se pueden analizar varios elementos importantes, el primero tiene que ver con la falta de estandarización de los datos ya que hay regiones que aparecen varias veces por algún carácter especial, esta parte se debe revisar para que arroje el consolidado de toda la región. La región que más escala incidente es la de Cundinamarca. |
|  <br>Figura. Tickets por Gerencia | La mayoría de los tickets vienen por parte de la gerencia de operaciones, esto suele ser muy común en varias de las organizaciones, ya que la mayoría de los incidentes se pueden dar en el día a |

<table>
<tr><td></td><td>día de las operaciones.</td></tr>
<tr><td>Tiempo Solución Minutos<br>In [14]:<br>

```
plt.boxplot((data['Tiempo Solución Minutos']))
plt.show()
data['Tiempo Solución Minutos'].describe()
```




Out[14]:

```
count    14286.000000
mean      2668.984530
std       6965.021269
min          1.000000
25%         24.000000
50%        159.000000
75%       1782.750000
max      97858.000000
Name: Tiempo Solución Minutos, dtype: float64
```

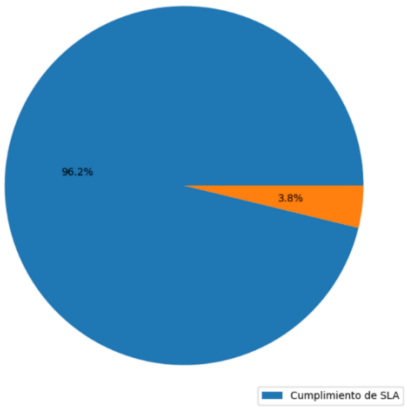
Figura. Tiempo solución en minutos</td><td>En esta grafica se puede observar que los tiempos de repuesta de los tickets son en promedio 2668 minutos, lo que se traduce en casi 2 días.<br>Es importante aclarar que estos cálculos han arrojado datos atípicos con tickets que han demorado hasta los 97858 lo que se traduce en 67 días.</td></tr>
<tr><td>% de Tickets por SLA<br> 

Figura. % de Tickets por SLA</td><td>En términos generales se ha evidenciado un buen cumplimiento en los acuerdos de servicio que tiene la mesa de atención de incidentes con un 96.2%.</td></tr>
</table>

## Anexo F. Resultados de encuestas de evaluación del modelo.

## PERTINENCIA PROTOTIPO PARA ANÁLISIS Y CLASIFICACIÓN DE INCIDENTES MEDIANTE PLN.

5. ¿Considera que mediante esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL se está respondiendo a las necesidades de las partes interesadas, como el Auditor Interno, los Auditores de Riesgos, los Auditores Senior y los Auditores?

Más detalles  Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 1 |
| Se cumple en alto grado | 6 |
| Se cumple plenamente | 4 |

6. ¿Considera que esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es relevante y tiene aplicabilidad en el ámbito práctico?

Más detalles  Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 5 |
| Se cumple plenamente | 6 |

7. ¿La propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es adecuada para abordar el problema planteado?

Más detalles  Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 7 |
| Se cumple plenamente | 4 |

# COHERENCIA PROTOTIPO PARA ANÁLISIS Y CLASIFICACIÓN DE INCIDENTES MEDIANTE PLN.

8. ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL es coherente con la formulación del problema planteado?

Más detalles  Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 5 |
| Se cumple plenamente | 6 |

9. ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL está interconectada y se complementa de manera lógica y coherente con todas las partes del proyecto?

Más detalles  Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 5 |
| Se cumple plenamente | 6 |

10. ¿Considera que la formulación del problema de investigación, los objetivos planteados, la metodología propuesta, los resultados esperados y las conclusiones se encuentran alineados de manera coherente y se sustentan mutuamente en el contexto del proyecto propuesto?

Más detalles

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 1 |
| Se cumple en alto grado | 4 |
| Se cumple plenamente | 6 |

11. ¿Considera que el proyecto de grado presenta una estructura clara y ordenada, estableciendo una relación directa entre el problema de investigación planteado y los objetivos propuestos? Además, ¿opina que los métodos y técnicas seleccionados son apropiados para abordar la problemática planteada?

Más detalles

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 4 |
| Se cumple plenamente | 7 |

13. ¿Considera que la propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL podría ser aplicada en otras áreas de la misma entidad?

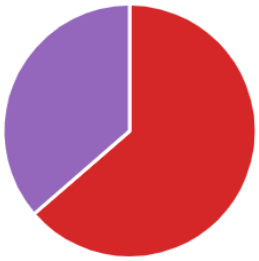
Más detalles

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 1 |
| Se cumple en alto grado | 4 |
| Se cumple plenamente | 6 |

14. ¿Cree que esta propuesta de Prototipo para análisis y clasificación de incidentes en una entidad financiera utilizando NPL puede ser aplicada en otras actividades dentro del campo de la auditoría interna?

Más detalles Información

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 7 |
| Se cumple plenamente | 4 |



16. ¿Considera usted que el prototipo planteado tiene el potencial de ser aplicado en el futuro para generar propuestas que faciliten la toma de decisiones fundamentadas en datos, aprovechando el crecimiento en el uso y la disponibilidad de información?

Más detalles

| | |
|---|---|
| No se cumple | 0 |
| Se cumple insatisfactoriamente | 0 |
| Se cumple aceptablemente | 0 |
| Se cumple en alto grado | 7 |
| Se cumple plenamente | 4 |

## Anexo G. Descripción de modelos de clasificación.

| Modelos de Clasificación | Descripción |
|---|---|
| **Naive Bayes** | Naive Bayes es un algoritmo de clasificación supervisada basado en el teorema de Bayes. Se utiliza comúnmente en problemas de clasificación de texto y minería de datos. El enfoque "ingenuo" (naive) de este algoritmo radica en la suposición de independencia condicional entre las características (variables predictoras) dadas las clases (categorías a predecir). Aunque esta suposición es simplificadora y rara vez se cumple en la práctica [6]. |
| **Regresión logística.** | La regresión logística es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación. A diferencia de la regresión lineal, que se utiliza para problemas de regresión, la regresión logística se emplea cuando la variable de salida (o variable dependiente) es categórica, es decir, se busca predecir la pertenencia a una o varias categorías [6]. |

| | |
|---|---|
| **K-vecinos.** | Es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y regresión. El algoritmo se basa en el principio de que las instancias similares tienden a compartir características y pertenecer a la misma clase.<br>En el caso de la clasificación, dada una instancia de prueba, el algoritmo k-NN busca en el conjunto de datos de entrenamiento las k instancias más cercanas a ella en términos de distancia. La distancia puede ser calculada utilizando diferentes medidas, como la distancia euclidiana o la distancia de Manhattan. Luego, toma la clase más común entre los k vecinos más cercanos y asigna esa clase a la instancia de prueba [6]. |
| **Árbol de decisión.** | Un árbol de decisión es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y regresión. El objetivo principal de un árbol de decisión es crear un modelo predictivo que aprenda reglas de decisión a partir de los datos de entrenamiento.<br>El árbol de decisión se construye en forma de estructura de árbol, donde cada nodo interno representa una característica o atributo, y cada rama representa una posible decisión o resultado basado en el valor de esa característica. Las hojas del árbol representan las clases o valores de salida [6]. |

| | |
|---|---|
| **Máquina de soporte vectorial.** | Es un algoritmo de aprendizaje supervisado utilizado tanto para problemas de clasificación como para problemas de regresión.<br>En el caso de la clasificación, SVM busca encontrar un hiperplano óptimo que permita separar las clases de manera eficiente en el espacio de características. El hiperplano es seleccionado de tal manera que maximiza la distancia entre los puntos de datos más cercanos de diferentes clases, conocidos como vectores de soporte. Estos vectores de soporte son los puntos más representativos y determinantes para definir la ubicación del hiperplano [6]. |
| **Random forest** | Random Forest (bosque aleatorio) es un algoritmo de aprendizaje supervisado utilizado para problemas de clasificación y regresión. Se basa en la idea de combinar múltiples árboles de decisión para obtener predicciones más precisas y estables.<br>El algoritmo Random Forest crea un conjunto de árboles de decisión, donde cada árbol se entrena con una muestra aleatoria de los datos de entrenamiento y utilizando un subconjunto aleatorio de las características disponibles. Esto se conoce como muestreo bootstrap y muestreo de características.<br>Durante la etapa de entrenamiento, cada árbol se construye utilizando una combinación de muestreo aleatorio y selección |

| | |
|---|---|
| | aleatoria de características. En cada nodo del árbol, se selecciona la mejor característica para dividir los datos en función de la ganancia de información o el índice de Gini.<br>Una vez que se ha construido el conjunto de árboles, las predicciones se realizan tomando la mayoría de votos en el caso de la clasificación, o el promedio en el caso de la regresión, de los resultados de todos los árboles [6]. |
| **Boosting o Bagging.** | Boosting y Bagging son dos técnicas de ensamblado utilizadas en el aprendizaje automático para mejorar la precisión y el rendimiento de los modelos de predicción.<br>Bagging (Bootstrap Aggregating) es una técnica que implica la construcción de múltiples modelos de forma independiente utilizando muestras de datos bootstrap, que son muestras aleatorias con reemplazo del conjunto de entrenamiento original. Cada modelo se entrena en una muestra diferente y produce una predicción individual. Luego, las predicciones individuales se |

| | |
|---|---|
| | promedian (en el caso de regresión) o se realiza una votación (en el caso de clasificación) para obtener la predicción final [6]. |
| **Red neuronal.** | Una red neuronal, también conocida como red neuronal artificial o artificial neural network (ANN) en inglés, es un modelo computacional inspirado en el funcionamiento del cerebro humano. Se utiliza en el campo del aprendizaje automático y se basa en la interconexión de nodos o neuronas artificiales.<br>En una red neuronal, las neuronas artificiales están organizadas en capas, generalmente en una estructura de capas de entrada, capas ocultas y una capa de salida. Cada neurona está asociada con una función de activación que determina su comportamiento.<br>El proceso de entrenamiento de una red neuronal implica la propagación hacia adelante (forward propagation) de los datos de entrada a través de la red, donde se calculan y ajustan los pesos de las conexiones entre las neuronas. Esto se hace para |

| | |
|---|---|
| | minimizar una función de pérdida o error, que mide la diferencia entre las salidas predichas por la red y las salidas reales esperadas. La optimización se realiza utilizando algoritmos de descenso del gradiente, como el retropropagación del error (backpropagation), que ajustan los pesos de las conexiones en función del gradiente de la función de pérdida [6]. |

## Anexo H. NOTEBOOK Trabajo de Grado.

# DATOS

Examinaremos un conjunto de datos de solicitudes a una mesa de servicio en una entidad financiera. Cada solicitud o ticket cuenta con una serie de características diferentes, a saber:

1. **ID:** un identificador unico para la solicitud o ticket.
2. **Fecha de Creación:** la fecha de creación de la solicitud o ticket.
3. **Título:** título de la solicitud o ticket.
4. **Descripción:** texto en el cual se explica la problematica que se le esta presentando al colaborador de la entidad.
5. **Prioridad:** clasificación de prioridad que se le da al ticket, puede ser: Bajo, Baja - Media, Medio, Alto y Urgente.
6. **Origen:** aplicativo desde el cual llega la solicitud.
7. **Creado Por:** identificador unico del colaborador que crea la solicitud.
8. **Oficina:** punto de venta, establecimiento o unidad empresarial de la entidad financiera.
9. **Regional:** departamento donde se encuentra la oficina registrada en la solicitud.
10. **Gerencia:** área de la entidad financiera donde trabaja el colaborador que realiza la solicitud.
11. **Clasificación:** clasificación manual que los colaboradores de la mesa de servicio le asignan a las solicitudes con caracteristicas similares.
12. **Grupo de Soporte:** la solicitud se le escala a los distintos niveles de la mesa de servicios para encontrar la solución a la problematica.
13. **Asignado A:** identificador unico del colaborador que atiende la solicitud.
14. **Estado:** al tratarse de solicitudes del mes de Enero de 2022 ya todas se encuentran cerradas.
15. **Detalle Respuesta:** texto en el cual se explica la solución a la problematica de la solicitud.
16. **Fecha Solución:** la fecha en que se implementa la solución al problema.
17. **Tiempo Solución Minutos:** tiempo transcurrido en minutos para implementar la solución al problema.
18. **Fecha Cierre:** la fecha en que se cierra la solicitud o ticket.
19. **Tipo Ticket:** todas estan clasificadas como solicitudes.

20. **Cumplio SLA:** se menciona si la solicitud esta vencida o no con respecto al tiempo de respuesta acordada por la mesa de servicio para las solicitudes.

Nos interesa principalmente el campo "Descripción", ya que contiene datos de texto en lenguaje natural y puede ser nuestro corpus. La variable respuesta será la de "Prioridad".

In [1]:
```
import nltk # importa el kit de herramientas de lenguaje natural
nltk.download('stopwords')

import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np
import re
import nltk
import matplotlib.pyplot as plt

pd.options.display.max_colwidth = 200
%matplotlib inline
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     /Users/juanda/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [2]:
```
# CARGAR EL CONJUNTO DE DATOS Y VER LOS DETALLES

df = pd.read_excel('Listado Tickets Diario-Enero_2022_Prueba_V2.xlsx')

df.head(1)
```

| Prioridad | Origen | Creado Por | Oficina | Regional | Gerencia | Clasificación | Grupo de Soporte | Asignado A | Estado | Detalle Respuesta | Sc |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Medio | Flujo de Trabajo | 1 | Fenalcontac | Valle del Cauca | OPERACIONES | OPERACIONES\CIS\RECORDAR USUARIO APP DIA NO HABIL | Auxilares CIS | 1 | Cerrado | Cordial saludo:\nMe permito informar que el usuario con el que se registró el cliente en la App de BMM es derianylf15, el usuario debe proceder a restablecer la contraseña ingresando el usuario, s... | 07 |

In [3]:
```
df.loc[df['Prioridad'] == 'Urgente', 'Prioridad'] = 'Alto'
df.loc[df['Prioridad'] == 'Baja - Media', 'Prioridad'] = 'Medio'
```

In [4]:
```
corpus = df['Descripción']
category = df['Prioridad']
```

In [5]:
```
df["Prioridad"].value_counts()
```

Out[5]:
```
Medio    6403
Bajo     5904
Alto     1979
Name: Prioridad, dtype: int64
```

# Analizando vocabulario del corpus antes del preprocesamiento

In [6]:
```
from collections import Counter

texto_descripciones = ' '.join(corpus)
palabras_tokenizadas = nltk.word_tokenize(texto_descripciones)
word_freq = Counter(palabras_tokenizadas)
word_freq.most_common(100)
```

Out[6]:
```
[('de', 25677),
 (',', 19575),
 ('.', 11122),
 ('la', 9847),
 ('el', 7713),
 ('para', 6385),
 ('por', 6233),
 ('se', 5955),
 ('a', 5911),
 ('su', 5484),
 ('que', 4972),
 ('en', 4889),
 ('y', 4453),
 ('del', 4164),
 ('usuario', 3697),
 ('con', 3623),
 ('solicito', 3511),
 ('Buen', 2960),
 ('Buenos', 2902),
 ('no', 2894),
 ('colaboracion', 2879),
 ('Gracias', 2822),
 ('dia', 2746),
 ('cliente', 2614),
 ('gracias', 2572),
 ('favor', 2498),
 ('CC', 2399),
 ('DE', 2329),
```

```
('dias', 2237),
('al', 2221),
('Buenas', 2209),
('ya', 2172),
('tardes', 2122),
('solicita', 2000),
('cc', 1844),
(':', 1821),
('adjunto', 1647),
('%', 1558),
('me', 1461),
('colaboración', 1405),
('día', 1323),
('los', 1261),
('credito', 1253),
('como', 1211),
('solicitud', 1169),
('servicios', 1164),
('habilitar', 1136),
('$', 1108),
('realizar', 1080),
('cuenta', 1020),
('pago', 992),
('Se', 983),
('atenta', 911),
('días', 908),
('amable', 902),
('le', 863),
('adjunta', 850),
('valor', 846),
('contraseña', 843),
('C.C', 836),
('quien', 821),
('saludo', 801),
('principal', 793),
('Buena', 786),
('un', 778),
('LA', 765),
('oficina', 761),
```

```
('cambio', 746),
('equipo', 745),
('tarde', 744),
('analista', 739),
('permite', 738),
('Solicito', 730),
('perfil', 722),
('documentos', 713),
('Cordial', 704),
(')', 699),
('(', 696),
('agencia', 688),
('esta', 684),
('nombre', 677),
('traslado', 656),
('debido', 653),
('las', 639),
('USUARIO', 628),
('envio', 628),
('EL', 624),
('es', 623),
('funcionaria', 615),
('quedo', 603),
('Muchas', 590),
('ingreso', 585),
('SE', 564),
('validar', 563),
('IP', 559),
('vacaciones', 551),
('cual', 551),
('encuentra', 526),
('c.c', 526),
('-', 525)]
```

# PREPROCESAMIENTO Y NORMALIZACIÓN

## Remover acentos

In [7]:
```
import unicodedata
def remove_accented_chars(text):
    text = unicodedata.normalize('NFKD', text).encode('ascii', 'ignore').decode('utf-8', 'ignore')
    return text
```

## Stemming

In [8]:
```
def simple_stemmer(text):
 ps = nltk.porter.PorterStemmer()
 text = ' '.join([ps.stem(word) for word in text.split()])
 return text
```

## Remover caracteres especiales

In [9]:
```
def remove_special_characters(text, remove_digits=False):
    pattern = r'[^a-zA-z0-9\s]' if not remove_digits else r'[^a-zA-z\s]'
    text = re.sub(pattern, '', text)
    return text
```

## Tokenización y eliminar stopwords

In [10]:
```
from nltk.tokenize.toktok import ToktokTokenizer

tokenizer = ToktokTokenizer()

stopword_list = nltk.corpus.stopwords.words('spanish')

def remove_stopwords(text, is_lower_case=False, stopwords=stopword_list):
    tokens = tokenizer.tokenize(text)
    tokens = [token.strip() for token in tokens]
    if is_lower_case:
        filtered_tokens = [token for token in tokens if token not in stopwords]
    else:
        filtered_tokens = [token for token in tokens if token.lower() not in stopwords]
    filtered_text = ' '.join(filtered_tokens)
    return filtered_text
```

In [11]:
```
print(stopword_list)
```

```
['de', 'la', 'que', 'el', 'en', 'y', 'a', 'los', 'del', 'se', 'las', 'por', 'un', 'para', 'con', 'no', '
una', 'su', 'al', 'lo', 'como', 'más', 'pero', 'sus', 'le', 'ya', 'o', 'este', 'sí', 'porque', 'esta', '
entre', 'cuando', 'muy', 'sin', 'sobre', 'también', 'me', 'hasta', 'hay', 'donde', 'quien', 'desde', 'to
do', 'nos', 'durante', 'todos', 'uno', 'les', 'ni', 'contra', 'otros', 'ese', 'eso', 'ante', 'ellos', 'e
', 'esto', 'mí', 'antes', 'algunos', 'qué', 'unos', 'yo', 'otro', 'otras', 'otra', 'él', 'tanto', 'esa',
'estos', 'mucho', 'quienes', 'nada', 'muchos', 'cual', 'poco', 'ella', 'estar', 'estas', 'algunas', 'alg
o', 'nosotros', 'mi', 'mis', 'tú', 'te', 'ti', 'tu', 'tus', 'ellas', 'nosotras', 'vosotros', 'vosotras',
'os', 'mío', 'mía', 'míos', 'mías', 'tuyo', 'tuya', 'tuyos', 'tuyas', 'suyo', 'suya', 'suyos', 'suyas',
'nuestro', 'nuestra', 'nuestros', 'nuestras', 'vuestro', 'vuestra', 'vuestros', 'vuestras', 'esos', 'esa
s', 'estoy', 'estás', 'está', 'estamos', 'estáis', 'están', 'esté', 'estés', 'estemos', 'estéis', 'estén
', 'estaré', 'estarás', 'estará', 'estaremos', 'estaréis', 'estarán', 'estaría', 'estarías', 'estaríamos
', 'estaríais', 'estarían', 'estaba', 'estabas', 'estábamos', 'estabais', 'estaban', 'estuve', 'estuvist
e', 'estuvo', 'estuvimos', 'estuvisteis', 'estuvieron', 'estuviera', 'estuvieras', 'estuviéramos', 'estu
vierais', 'estuvieran', 'estuviese', 'estuvieses', 'estuviésemos', 'estuvieseis', 'estuviesen', 'estando
', 'estado', 'estada', 'estados', 'estadas', 'estad', 'he', 'has', 'ha', 'hemos', 'habéis', 'han', 'haya
', 'hayas', 'hayamos', 'hayáis', 'hayan', 'habré', 'habrás', 'habrá', 'habremos', 'habréis', 'habrán', '
habría', 'habrías', 'habríamos', 'habríais', 'habrían', 'había', 'habías', 'habíamos', 'habíais', 'había
n', 'hube', 'hubiste', 'hubo', 'hubimos', 'hubisteis', 'hubieron', 'hubiera', 'hubieras', 'hubiéramos',
'hubierais', 'hubieran', 'hubiese', 'hubieses', 'hubiésemos', 'hubieseis', 'hubiesen', 'habiendo', 'habi
do', 'habida', 'habidos', 'habidas', 'soy', 'eres', 'es', 'somos', 'sois', 'son', 'sea', 'seas', 'seamos
', 'seáis', 'sean', 'seré', 'serás', 'será', 'seremos', 'seréis', 'serán', 'sería', 'serías', 'seríamos'
, 'seríais', 'serían', 'era', 'eras', 'éramos', 'erais', 'eran', 'fui', 'fuiste', 'fue', 'fuimos', 'fuis
teis', 'fueron', 'fuera', 'fueras', 'fuéramos', 'fuerais', 'fueran', 'fuese', 'fueses', 'fuésemos', 'fue
seis', 'fuesen', 'sintiendo', 'sentido', 'sentida', 'sentidos', 'sentidas', 'siente', 'sentid', 'tengo',
'tienes', 'tiene', 'tenemos', 'tenéis', 'tienen', 'tenga', 'tengas', 'tengamos', 'tengáis', 'tengan', 't
endré', 'tendrás', 'tendrá', 'tendremos', 'tendréis', 'tendrán', 'tendría', 'tendrías', 'tendríamos', 't
endríais', 'tendrían', 'tenía', 'tenías', 'teníamos', 'teníais', 'tenían', 'tuve', 'tuviste', 'tuvo', 't
uvimos', 'tuvisteis', 'tuvieron', 'tuviera', 'tuvieras', 'tuviéramos', 'tuvierais', 'tuvieran', 'tuviese
', 'tuvieses', 'tuviésemos', 'tuvieseis', 'tuviesen', 'teniendo', 'tenido', 'tenida', 'tenidos', 'tenida
s', 'tened']
```

# Preprocesamiento del texto

In [12]:
```
def normalize_corpus(corpus,
                     accented_char_removal=True, text_lower_case=True,
                     stemming_text=True, special_char_removal=True,
                     stopword_removal=True, remove_digits=True):

    normalized_corpus = []
    # normalize each document in the corpus
    for doc in corpus:
        # remove accented characters
      if accented_char_removal:
        doc = remove_accented_chars(doc)
        # lowercase the text
      if text_lower_case:
        doc = doc.lower()
        # remove extra newlines
      doc = re.sub(r'[\r|\n|\r\n]+', ' ', str(doc))
        # simple_stemmer
      if stemming_text:
        doc = simple_stemmer(doc)
        # remove special characters and\or digits
      if special_char_removal:
            # insert spaces between special characters to isolate them
        special_char_pattern = re.compile(r'([{.(-)!}])')
        doc = special_char_pattern.sub(" \\1 ", str (doc))
        doc = remove_special_characters(doc, remove_digits=remove_digits)
        # remove extra whitespace
      doc = re.sub(' +', ' ', str (doc))
        # remove stopwords
      if stopword_removal:
        doc = remove_stopwords(doc, is_lower_case=text_lower_case)

      normalized_corpus.append(doc)

    return normalized_corpus
```

In [13]:
```
normalize_corpus = normalize_corpus(corpus)
```

In [14]:
```
print(normalize_corpus[0:25])
```

```
['olvido usuario contrasena', 'olvido usuario contrasena', 'bueno dias solicito colaboracion habilitar u
suario restablec contrasena usuario iquiroga cedula reintegro vacacion', 'cordial saludo manera atenta p
ermito informar permit ver menu bantot analista credito olga lucia cano hernandez usuario ocano gracia',
'buen dia amablement solicita colaboracion habilitar usuario aromerog astrid romero gaspar cc reintegro
vacacion', 'buen dia solicitamo activen usuario hverag hernando vera garcia encontraba periodo vacacion
retornoel dia hoy', 'buen dia amablement solicita colaboracion habilitar usuario apalvar adriana milena
alvarez parga reintegro vacacion', 'favor revisar usuario rcaicedol ningun boton habilitado bantotal caj
era princip asesora perfil autorizado dobl perfil aunqu cajera principal ahora despliega ningun boton',
'solicito colaboracion habilitar cajera princip nuevament abvillalba ana beatriz villalba sala', 'buen d
ia amablement solicita colaboracion habilitar usuario svallejo sandra milena vallejo garcia reintegro va
cacion', 'buen dia solicito habilitar usuario bantot poder dar inicio sistema lidia lori lopez cuesta c
c coordinadora servicio carolina correa villa c c cajera principal mucha gracia', 'solicitud realizar si
guient cambio perfil habilitar cajro princip fchiquillo fladi chiquillo castro cedula cajera princip ret
orna vacacion devolverl asesor psuarez paula suarez cruz cedu asesora servicio agradecemo valiosa colabo
racion agradecemo valiosa colaboracion realizar cambio tener retroceso gracia colaboracion', 'buen dia m
anera atenta solicito amabl colaboracion habilitacion usuario bantot funcionaria diana isabel guzman seg
ura cedula usuario diguzman regreso licencia mucha gracia', 'buena tarde solicita habilitar perfil cajer
a princip yurinei garcia cc ypgarcia restablec contrasena reintegra vacacion cambiar perfil funcionaria
ynovoa asesor cc', 'buen dia favor realizar traslado asesor mfrancor miguel franco sede pereira cuba dos
quebradas mucha gracia', 'amablement solicita trasladar usuario lcolorado luz adriana colorado giraldo a
gencia armenia centro agencia sur cargo cajera princip', 'buen dia solciito amabl coolaboracion asesor s
ervicio andr felip gonzalez segura cc usuario afgonzalez desplega menu bantotal agradzeco amabl colabora
cion', 'buen dia adjunto solicitud traslado adriana soto lsoto sede pereira cuba pereira centro gracia',
'buen dia solicito colaboraccion habilitar usuario restablec contrasena jcabril juan carlo abril garzon
cc reintegra vavacion dia hoy usuario ycastroc yenni alexandra castro correacc disfrutara vacacion parti
r dia hoy', 'bueno dias solicito amabl colaboracion crear usuario aprendiz sena paola alexandra garc mun
oz cc realiza ingreso dia hoy gracia', 'bueno dias solicito colaboracion habilitar jose armando torr vil
lega cc usuario javillegas cajero princip inhabilitar funcionaria sandra leonor bernal moreno dia hoy en
ero gracia', 'buen dia adjunto solicitud traslado nhora roja tangarif nrojast sede pereira cuba sede dos
quebradas gracia', 'buen dia cordial saludo solicito amablement usuario window analista credito eyber ga
briel gomez ruano identificado cc gracia', 'bueno dias solicitar comedidament habilitar jessica cort usu
ario jcort coordinadora servicio lina wong usuario lwong cajera princip', 'bueno dia solicito amabl cola
boracion deja ingresar bantotal digito varia vece usuario contrasena permit abrir lleva vario dia asi de
moro aprox media hora poder ingreasar ip equipo mucha gracia quedo atenta']
```

# Corpus limpio

In [15]:
```
data_clean = pd.DataFrame(normalize_corpus)
```

In [16]:
```
data_clean.to_csv('clean_data.csv', index=False)
```

In [17]:
```
data_clean = pd.read_csv('clean_data.csv')
```

In [18]:
```
data_clean = data_clean.rename(columns={'0':'Descripcion_limpia'})
data_clean.head(5)
```

Out[18]:

| | Descripcion_limpia |
|---|---|
| **0** | olvido usuario contrasena |
| **1** | olvido usuario contrasena |
| **2** | bueno dias solicito colaboracion habilitar usuario restablec contrasena usuario iquiroga cedula reintegro vacacion |
| **3** | cordial saludo manera atenta permito informar permit ver menu bantot analista credito olga lucia cano hernandez usuario ocano gracia |
| **4** | buen dia amablement solicita colaboracion habilitar usuario aromerog astrid romero gaspar cc reintegro vacacion |

# Analizando vocabulario del corpus después del preprocesamiento

In [19]:
```
from collections import Counter

texto_descripciones_limpias = ' '.join(normalize_corpus)
palabras_tokenizadas = nltk.word_tokenize(texto_descripciones_limpias)
word_freq = Counter(palabras_tokenizadas)
word_freq.most_common(100)
```

Out[19]:
```
[('dia', 5549),
 ('gracia', 4773),
```

```
('cc', 4689),
('usuario', 4655),
('colaboracion', 4503),
('solicito', 4436),
('bueno', 3421),
('buena', 3365),
('buen', 3293),
('client', 3178),
('c', 3072),
('favor', 2820),
('dias', 2380),
('adjunto', 2238),
('solicita', 2232),
('credito', 2134),
('tardes', 1737),
('solicitud', 1614),
('servicio', 1559),
('cuenta', 1397),
('habilitar', 1365),
('gracias', 1362),
('pago', 1192),
('realizar', 1155),
('mucha', 1147),
('ip', 1081),
('agencia', 1072),
('quedo', 1023),
('atenta', 1021),
('analista', 1018),
('valor', 970),
('contrasena', 968),
('cambio', 954),
('equipo', 952),
('tard', 951),
('amabl', 946),
('oficina', 939),
('cordial', 919),
('documento', 914),
('adjunta', 914),
('perfil', 899),
```

```
('princip', 897),
('envio', 893),
('saludo', 892),
('agradezco', 846),
('pda', 799),
('traslado', 788),
('permit', 787),
('reclamacion', 748),
('nombr', 739),
('debido', 697),
('ingreso', 696),
('funcionaria', 685),
('tarde', 656),
('operacion', 650),
('cajera', 649),
('cedula', 647),
('cajero', 638),
('asesor', 623),
('validar', 615),
('seguro', 615),
('cartera', 563),
('funcionario', 562),
('encuentra', 554),
('correo', 550),
('error', 539),
('ingresar', 538),
('asesora', 536),
('maria', 528),
('titular', 522),
('coordinador', 514),
('autorizacion', 501),
('enero', 489),
('amablement', 479),
('hoy', 478),
('aparec', 478),
('bantot', 472),
('coordinadora', 455),
('carpeta', 454),
('senor', 452),
```

```
('permito', 442),
('impresora', 439),
('lista', 436),
('mil', 429),
('permiso', 422),
('desd', 414),
('vacacion', 413),
('tien', 410),
('cdt', 398),
('atencion', 394),
('cargo', 389),
('nuevo', 388),
('restablec', 379),
('soport', 362),
('report', 361),
('rodriguez', 359),
('formato', 359),
('requier', 352),
('clave', 351),
('senora', 339)]
```

## Ajustes para aplicar modelos

In [20]:
```
data_clean = data_clean.assign(Category=df['Prioridad'])
```

In [21]:
```
data_clean
```

Out[21]:

| | Descripcion_limpia | Category |
|---|---|---|
| **0** | olvido usuario contrasena | Medio |
| **1** | olvido usuario contrasena | Medio |
| **2** | bueno dias solicito colaboracion habilitar usuario restablec contrasena usuario iquiroga cedula reintegro vacacion | Medio |
| **3** | cordial saludo manera atenta permito informar permit ver menu bantot analista credito olga lucia cano hernandez usuario ocano gracia | Medio |
| **4** | buen dia amablement solicita colaboracion habilitar usuario aromerog astrid romero gaspar cc reintegro vacacion | Medio |
| **...** | ... | ... |
| **14281** | buena tardes solicitamo colaboracion habilitar nuevament curso actualizacion mejor ruta construir banco lider cumunidad ano jlballest jose lui ballestero quintero cc | Medio |
| **14282** | bueno tardes favor habilitar usuario nbarrios ninibeth barrio barrio cc coordinador servicios modificar usuario lbertel leidi bertel bello cc coordinador servicio cajero princip modificar usuario ... | Medio |
| **14283** | buena tard solicita inhabilitar cajero princip sebastian viloria castellano dia lune empezara licencia paternidad habilitar reemplazo tempor cajera princip maryra marquez rodriguez cc usuario mmar... | Medio |
| **14284** | olvido usurio | Medio |
| **14285** | cliet claudia milena causado garcia cc comunico usted nuevament dado hice linea nacion senorita lina corrales debido realic pago total credito codigo medio boton pse valor total valor anterior cor... | Medio |

14286 rows × 2 columns

In [22]:
```
data_clean = data_clean.fillna(' ')
```

In [23]:
```
from sklearn.model_selection import train_test_split

train_corpus, test_corpus, train_label_names, test_label_names =\
                                train_test_split(np.array(data_clean['Descripcion_limpia']),
                                np.array(data_clean['Category']), test_size=0.33, random_state=42)

train_corpus.shape, test_corpus.shape
```

Out[23]:
```
((9571,), (4715,))
```

In [24]:
```
from collections import Counter

trd = dict(Counter(train_label_names))
tsd = dict(Counter(test_label_names))

(pd.DataFrame([[key, trd[key], tsd[key]] for key in trd],
              columns=['Category', 'Train Count', 'Test Count'])
.sort_values(by=['Train Count', 'Test Count'],
              ascending=False))
```

Out[24]:

| | Category | Train Count | Test Count |
|---|---|---|---|
| **1** | Medio | 4317 | 2086 |
| **0** | Bajo | 3919 | 1985 |
| **2** | Alto | 1335 | 644 |

In [25]:
```
from sklearn import metrics

def get_metrics(true_labels, predicted_labels):

    print('Accuracy:', np.round(
                        metrics.accuracy_score(true_labels,
                                               predicted_labels),
                        4))
    print('Precision:', np.round(
                        metrics.precision_score(true_labels,
                                               predicted_labels,
                                               average='weighted'),
                        4))
    print('Recall:', np.round(
                        metrics.recall_score(true_labels,
                                               predicted_labels,
                                               average='weighted'),
                        4))
    print('F1 Score:', np.round(
                        metrics.f1_score(true_labels,
                                               predicted_labels,
                                               average='weighted'),
                        4))
```

# MODELO BOLSA DE PALABRAS

In [26]:
```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import cross_val_score

# build BOW features on train articles
cv = CountVectorizer(binary=False, min_df=0.0, max_df=1.0)
cv_train_features = cv.fit_transform(train_corpus)
```

In [27]:
```
# transform test articles into features
cv_test_features = cv.transform(test_corpus)
```

In [28]:
```
print('BOW model:> Train features shape:', cv_train_features.shape,
      ' Test features shape:', cv_test_features.shape)
```

```
BOW model:> Train features shape: (9571, 12674)  Test features shape: (4715, 12674)
```

## Naive Bayes

In [29]:
```
from sklearn.naive_bayes import MultinomialNB

mnb = MultinomialNB(alpha=1)
mnb.fit(cv_train_features, train_label_names)
```

Out[29]:
```
MultinomialNB(alpha=1)
```

In [30]:
```
mnb_predictions = mnb.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=mnb_predictions)
```

```
Accuracy: 0.789
Precision: 0.7926
Recall: 0.789
F1 Score: 0.7858
```

In [31]:
```
from sklearn.metrics import confusion_matrix

matriz = confusion_matrix(test_label_names, mnb_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[31]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 351 | 81 | 212 |
| **Alto** | 19 | 1597 | 369 |
| **Medio** | 61 | 253 | 1772 |

## Regresión logística

In [32]:
```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(penalty='l2', max_iter=100, C=1, random_state=42)
lr.fit(cv_train_features, train_label_names)
```

Out[32]:
```
LogisticRegression(C=1, random_state=42)
```

In [33]:
```
#R2

lr_predictions = lr.predict(cv_test_features)
metrics.precision_score(test_label_names, lr_predictions, average='weighted')
```

Out[33]:
```
0.807578060062458
```

## K-vecinos

In [34]:
```
from sklearn.neighbors import KNeighborsClassifier

n_neighbors = 3
k_neighbors = KNeighborsClassifier(n_neighbors)
k_neighbors.fit(cv_train_features, train_label_names)
```

Out[34]:
```
KNeighborsClassifier(n_neighbors=3)
```

In [35]:
```
k_neighbors_predictions = k_neighbors.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=k_neighbors_predictions)
```

```
Accuracy: 0.6526
Precision: 0.6579
Recall: 0.6526
F1 Score: 0.6548
```

In [36]:
```
matriz = confusion_matrix(test_label_names, k_neighbors_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[36]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 322 | 152 | 170 |
| **Alto** | 184 | 1322 | 479 |
| **Medio** | 234 | 419 | 1433 |

## Arboles de decisión

In [37]:
```
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier()
clf.fit(cv_train_features,train_label_names)
```

Out[37]:
```
DecisionTreeClassifier()
```

In [38]:
```
clf_predictions = clf.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=clf_predictions)
```

```
Accuracy: 0.7618
Precision: 0.7603
Recall: 0.7618
F1 Score: 0.7607
```

In [39]:
```
matriz = confusion_matrix(test_label_names, clf_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[39]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 380 | 119 | 145 |
| **Alto** | 68 | 1562 | 355 |
| **Medio** | 131 | 305 | 1650 |

## Máquinas de soporte vectorial

In [40]:
```
from sklearn.svm import LinearSVC

svm = LinearSVC(penalty='l2', C=1, random_state=42)
svm.fit(cv_train_features, train_label_names)
```

Out[40]:
```
LinearSVC(C=1, random_state=42)
```

In [41]:
```
svm_predictions = svm.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=svm_predictions)
```

```
Accuracy: 0.7913
Precision: 0.7905
Recall: 0.7913
F1 Score: 0.7907
```

In [42]:
```
matriz = confusion_matrix(test_label_names, svm_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[42]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 414 | 83 | 147 |
| **Alto** | 48 | 1624 | 313 |
| **Medio** | 134 | 259 | 1693 |

## Random forest

In [43]:
```
from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier(n_estimators=10, random_state=42)
rfc.fit(cv_train_features, train_label_names)
```

Out[43]:
```
RandomForestClassifier(n_estimators=10, random_state=42)
```

In [44]:
```
rfc_predictions = rfc.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=rfc_predictions)
```

```
Accuracy: 0.7758
Precision: 0.7756
Recall: 0.7758
F1 Score: 0.7736
```

In [45]:
```
matriz = confusion_matrix(test_label_names, rfc_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[45]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 368 | 104 | 172 |
| **Alto** | 27 | 1594 | 364 |
| **Medio** | 92 | 298 | 1696 |

## Boosting o Bagging

In [46]:
```
from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(n_estimators=10, random_state=42)
gbc.fit(cv_train_features, train_label_names)
```

Out[46]:
```
GradientBoostingClassifier(n_estimators=10, random_state=42)
```

In [47]:
```
gbc_predictions = gbc.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=gbc_predictions)
```

```
Accuracy: 0.6477
Precision: 0.68
Recall: 0.6477
F1 Score: 0.6278
```

In [48]:
```
matriz = confusion_matrix(test_label_names, gbc_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[48]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 125 | 291 | 228 |
| **Alto** | 3 | 1568 | 414 |
| **Medio** | 11 | 714 | 1361 |

## Redes neuronales

In [49]:
```
from sklearn.neural_network import MLPClassifier

model = MLPClassifier()
model.fit(cv_train_features, train_label_names)
```

Out[49]:
```
MLPClassifier()
```

In [50]:
```
model_predictions = model.predict(cv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=model_predictions)
```

```
Accuracy: 0.7656
Precision: 0.764
Recall: 0.7656
F1 Score: 0.7643
```

In [51]:
```
matriz = confusion_matrix(test_label_names, model_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[51]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 384 | 90 | 170 |
| **Alto** | 56 | 1604 | 325 |
| **Medio** | 122 | 342 | 1622 |

# MODELO TF - IDF

In [52]:
```
from sklearn.feature_extraction.text import TfidfVectorizer

# build BOW features on train articles
tv = TfidfVectorizer(use_idf=True, min_df=0.0, max_df=1.0)
tv_train_features = tv.fit_transform(train_corpus)
```

In [53]:
```
# transform test articles into features
tv_test_features = tv.transform(test_corpus)
```

In [54]:
```
print('TFIDF model:> Train features shape:', tv_train_features.shape,
      ' Test features shape:', tv_test_features.shape)
```

```
TFIDF model:> Train features shape: (9571, 12674)  Test features shape: (4715, 12674)
```

## Naive Bayes

In [55]:
```
from sklearn.naive_bayes import MultinomialNB

mnb = MultinomialNB(alpha=1)
mnb.fit(tv_train_features, train_label_names)
```

Out[55]:
```
MultinomialNB(alpha=1)
```

In [56]:
```
mnb_predictions = mnb.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=mnb_predictions)
```

```
Accuracy: 0.7589
Precision: 0.7889
Recall: 0.7589
F1 Score: 0.7358
```

In [57]:
```
from sklearn.metrics import confusion_matrix

matriz = confusion_matrix(test_label_names, mnb_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[57]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 137 | 120 | 387 |
| **Alto** | 1 | 1556 | 428 |
| **Medio** | 6 | 195 | 1885 |

## Regresión logística

In [58]:
```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression(penalty='l2', max_iter=100, C=1, random_state=42)
lr.fit(tv_train_features, train_label_names)
```

Out[58]:
```
LogisticRegression(C=1, random_state=42)
```

In [59]:
```
#R2

lr_predictions = lr.predict(tv_test_features)
metrics.precision_score(test_label_names, lr_predictions, average='weighted')
```

Out[59]:
```
0.8059326602768081
```

## K-vecinos

In [60]:
```
from sklearn.neighbors import KNeighborsClassifier

n_neighbors = 3
k_neighbors = KNeighborsClassifier(n_neighbors)
k_neighbors.fit(tv_train_features, train_label_names)
```

Out[60]:
```
KNeighborsClassifier(n_neighbors=3)
```

In [61]:
```
k_neighbors_predictions = k_neighbors.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=k_neighbors_predictions)
```

```
Accuracy: 0.5832
Precision: 0.7285
Recall: 0.5832
F1 Score: 0.5404
```

In [62]:
```
matriz = confusion_matrix(test_label_names, k_neighbors_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[62]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 155 | 473 | 16 |
| **Alto** | 10 | 1934 | 41 |
| **Medio** | 33 | 1392 | 661 |

## Arboles de decisión

In [63]:
```
from sklearn.tree import DecisionTreeClassifier

clf = DecisionTreeClassifier()
clf.fit(tv_train_features,train_label_names)
```

Out[63]:
```
DecisionTreeClassifier()
```

In [64]:
```
clf_predictions = clf.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=clf_predictions)
```

```
Accuracy: 0.7417
Precision: 0.7417
Recall: 0.7417
F1 Score: 0.7413
```

In [65]:
```
matriz = confusion_matrix(test_label_names, clf_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[65]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 387 | 99 | 158 |
| **Alto** | 91 | 1486 | 408 |
| **Medio** | 142 | 320 | 1624 |

## Máquinas de soporte vectorial

In [66]:
```
from sklearn.svm import LinearSVC

svm = LinearSVC(penalty='l2', C=1, random_state=42)
svm.fit(tv_train_features, train_label_names)
```

Out[66]:
```
LinearSVC(C=1, random_state=42)
```

In [67]:
```
svm_predictions = svm.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=svm_predictions)
```

```
Accuracy: 0.8127
Precision: 0.8124
Recall: 0.8127
F1 Score: 0.8115
```

In [68]:
```
matriz = confusion_matrix(test_label_names, svm_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[68]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 412 | 73 | 159 |
| **Alto** | 30 | 1656 | 299 |
| **Medio** | 105 | 217 | 1764 |

## Random forest

In [69]:
```
from sklearn.ensemble import RandomForestClassifier

rfc = RandomForestClassifier(n_estimators=10, random_state=42)
rfc.fit(tv_train_features, train_label_names)
```

Out[69]:
```
RandomForestClassifier(n_estimators=10, random_state=42)
```

In [70]:
```
rfc_predictions = rfc.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=rfc_predictions)
```

```
Accuracy: 0.7644
Precision: 0.7634
Recall: 0.7644
F1 Score: 0.7615
```

In [71]:
```
matriz = confusion_matrix(test_label_names, rfc_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[71]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 347 | 111 | 186 |
| **Alto** | 30 | 1590 | 365 |
| **Medio** | 97 | 322 | 1667 |

## Boosting o Bagging

In [72]:
```
from sklearn.ensemble import GradientBoostingClassifier

gbc = GradientBoostingClassifier(n_estimators=10, random_state=42)
gbc.fit(tv_train_features, train_label_names)
```

Out[72]:
```
GradientBoostingClassifier(n_estimators=10, random_state=42)
```

In [73]:
```
gbc_predictions = gbc.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=gbc_predictions)
```

```
Accuracy: 0.6704
Precision: 0.7027
Recall: 0.6704
F1 Score: 0.6482
```

In [74]:
```
matriz = confusion_matrix(test_label_names, gbc_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[74]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 119 | 292 | 233 |
| **Alto** | 2 | 1625 | 358 |
| **Medio** | 9 | 660 | 1417 |

## Redes neuronales

In [75]:
```
from sklearn.neural_network import MLPClassifier

model = MLPClassifier()
model.fit(tv_train_features, train_label_names)
```

Out[75]:
```
MLPClassifier()
```

In [76]:
```
model_predictions = model.predict(tv_test_features)
unique_classes = list(set(test_label_names))
get_metrics(true_labels=test_label_names, predicted_labels=model_predictions)
```

```
Accuracy: 0.7635
Precision: 0.763
Recall: 0.7635
F1 Score: 0.7627
```

In [77]:
```
matriz = confusion_matrix(test_label_names, model_predictions)

pd.DataFrame(matriz, index = unique_classes, columns = unique_classes)
```

Out[77]:

| | Bajo | Alto | Medio |
|---|---|---|---|
| **Bajo** | 386 | 78 | 180 |
| **Alto** | 61 | 1557 | 367 |
| **Medio** | 136 | 293 | 1657 |

NOTA: Los mejores algoritmos son la máquina de soporte vectorial y la regresión logística, en este caso vamos a continuar con ambos para realizar el proceso de optimización de hiperparámetros, al final vamos a comparar cuál de los 2 realiza de forma más adecuada la clasificación de los tickets por su prioridad.

# Ajuste del modelo de regresión logística con TF-IDF

In [78]:
```
from sklearn.pipeline import Pipeline
from sklearn.model_selection import GridSearchCV

lr_pipeline = Pipeline([('tfidf', TfidfVectorizer()),
                        ('lr', LogisticRegression(penalty='l2', max_iter=100, random_state=42))
                       ])

param_grid = {'tfidf__ngram_range': [(1, 1), (1, 2)],
              'lr__C': [1, 5, 10]
}

gs_lr = GridSearchCV(lr_pipeline, param_grid, cv=5, verbose=2)
gs_lr = gs_lr.fit(train_corpus, train_label_names)
```

```
Fitting 5 folds for each of 6 candidates, totalling 30 fits
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 1); total time=   0.6s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 1); total time=   0.6s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 2); total time=   2.1s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 2); total time=   2.0s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 2); total time=   2.1s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 2); total time=   2.3s
[CV] END .................lr__C=1, tfidf__ngram_range=(1, 2); total time=   2.2s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 1); total time=   0.8s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 1); total time=   0.8s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 1); total time=   0.8s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 2); total time=   2.2s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 2); total time=   2.2s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 2); total time=   2.1s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 2); total time=   2.4s
[CV] END .................lr__C=5, tfidf__ngram_range=(1, 2); total time=   2.3s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 1); total time=   0.7s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 2); total time=   2.2s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 2); total time=   2.5s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 2); total time=   2.3s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 2); total time=   2.2s
[CV] END ................lr__C=10, tfidf__ngram_range=(1, 2); total time=   2.3s
```

In [79]:
```
gs_lr.best_estimator_
```

Out[79]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer(ngram_range=(1, 2))),
                ('lr', LogisticRegression(C=10, random_state=42))])
```

In [80]:
```
# evaluar el modelo mejor sintonizado en el conjunto de datos de prueba

best_lr_test_score = gs_lr.score(test_corpus, test_label_names)
print('Test Accuracy :', best_lr_test_score)
```

```
Test Accuracy : 0.8203605514316012
```

In [81]:
```
# Extraer los números de fila del documento de prueba

train_idx, test_idx = train_test_split(
    np.array(range(len(data_clean['Descripcion_limpia']))), test_size=0.33, random_state=42)
test_idx
```

Out[81]:
```
array([4450, 7444, 7050, ..., 7163, 4320, 6258])
```

In [82]:
```
predict_probas = gs_lr.predict_proba(test_corpus).max(axis=1)
test_df_1 = data_clean.iloc[test_idx]
test_df_1['Predicted Name'] = lr_predictions
test_df_1['Predicted Confidence'] = predict_probas
test_df_1.head(60)
```

Out[82]:

| | Descripcion_limpia | Category | Predicted Name | Predicted Confidence |
|---|---|---|---|---|
| **4450** | buen dia envia solicitud correccion pago operacion cliente fabiola bermudez casierra c c error recaudo numero credito adjunto carta solicitud soport pago gracias | Medio | Medio | 0.985052 |
| **7444** | bueno dia favor rechazar delegar insatncia permit tomar dar continuidad proceso | Alto | Alto | 0.982927 |
| **7050** | bueno dia carpeta siniestro encuntra reclamacion seguro fp lui iriart salgado cc | Bajo | Bajo | 0.994699 |
| **10313** | buena tardes solicto colaboracion aplicar pago credito client diego marin agudelo cc cobro comis juridica acordada abogado adjunta pago correo | Medio | Medio | 0.910490 |
| **12922** | buen dia adjunto documento client monica mafla mora credito condonacion conceptos | Bajo | Bajo | 0.962882 |
| **8764** | solicito favor habilitar coordinadora servicio neiva altico gina lorena salgado roja usuario gsalgado partir hoy enero debido retoma cargo despu reemplazar vacacion c r c region huila deshabilitar... | Medio | Medio | 0.939832 |

| | | | | |
|---|---|---|---|---|
| **6575** | buen dia solicito colaboracion actualizacion correo usuario afburbano | Medio | Alto | 0.801614 |
| **5403** | buen dia adjunta report disfrut vacaciones | Medio | Medio | 0.978253 |
| **476** | desbloqueo clave datacredito | Bajo | Bajo | 0.974783 |
| **10950** | buen dia solicito cambio perfil cajero princip oswal fernando rozo c c orozo cajera princip laura daniela ceron c c ldceron incapacitada dia gracia | Medio | Medio | 0.994858 |
| **1139** | bueno dias adjunto report client rosmeri montenegro parodi cc fin validar categoria puesto aparec recurrent ant pandemia elit plus anterior sujetro nuevo cambio sarc agradezco colaboracion afectac... | Medio | Medio | 0.969271 |
| **12063** | cordial saludo favor autorizar poliza plan hipertens tiroid rosa lobon gil cc quedo atenta autorizacion gracia | Alto | Alto | 0.936694 |
| **4245** | solicita revisar cartera digit analista credito permit conexion usuario registrado | Bajo | Bajo | 0.980520 |
| **884** | buen dia adjunta report cambio cargo | Medio | Medio | 0.986997 |
| **11515** | bueno dias favor enviar provis gracia | Medio | Medio | 0.957410 |
| **10558** | buen dia solicito habilitar luz arelli valencia usuario lvalencia coordinadora servicios martha luci jurado usuario mjurado cajera princip juliana jaramillo usuario jjaramil asesora servicio pda s... | Medio | Medio | 0.974485 |
| **99** | buen dia favor consutar lista restrictiva titular ampliacion fabian andr gomez identificado cc mucha gracias quedo atenta | Medio | Medio | 0.989715 |
| **3467** | buena tardes solicito apoyo habilitar coordinador servicio popayan ciudad jardin usuario apgonzale angela patricia gonzalez santand deshabilitar usuario yramirezt yuri katerin ramirez tobar estara... | Medio | Medio | 0.950571 |
| **9214** | buena tarde solicita favor deshabilitar clave cajero princip estara vacacion asignar clave bakcup boveda asesor servicios adjunto envio formato solicitud | Medio | Medio | 0.928050 |
| **9419** | bueno dia envio certificado libertad viabilidad credito gracia | Alto | Alto | 0.808107 |
| **14010** | bueno dias permito solicitar habilitar cajero princip lui miguel mayorga asesora servicio magda milena martinez perfil asesora servicio cajera auxiliar mucha gracia | Medio | Medio | 0.981531 |
| **6325** | celular conecta vpn celular analista credito carlo antonio casanova c c usuario ccasanova celular | Bajo | Bajo | 0.994571 |
| | buen dia solicito colaboracion autorizacion cuarto sistema part prosegur validar | | | |

| | | | | |
|---|---|---|---|---|
| **5226** | disponibilidad ampliacion sistema seguridad techo oficina libertador robo presntado agencia quedo atenta | Medio | Medio | 0.975414 |
| **1984** | solicita prueba panico | Alto | Alto | 0.994615 |
| **290** | bueno dias solicito colaboracion validacion usuario rcastaneda acceso adminfo genera do errores horario permitido error integracion mucha gracias adjunto pantallazo | Bajo | Bajo | 0.705554 |
| **6684** | buena tard adjunto solicitud devolucion familia protegida karina medina torr cc precancelacion credito | Bajo | Bajo | 0.984562 |
| **7632** | bueno dia alexis favor restablec credencial acceso sftp talento humano | Bajo | Bajo | 0.909551 |
| **2039** | buen dia solicito amablement cuenta operacion senor rocha gomez manuel mucha gracias | Medio | Medio | 0.528537 |
| **13200** | adjunto solicitud ajust pago credito sarria chasoy yulli tatiana operacion client cancelo credito mora | Medio | Medio | 0.976084 |
| **9097** | solicito favor asignar perfil ppardo auxopereca ambient preproduccion sucurs | Medio | Medio | 0.974105 |
| **9052** | envian soport reclamacion devolucion prima seguro familia protegida nombr sra lucero lopez cc sede palmira saman | Bajo | Bajo | 0.986877 |
| **3187** | buen dia favor reestablec contrasena funcionario giovanni alexand chamorro c c nro | Medio | Alto | 0.575353 |
| **5623** | bueno dias favor pido amabl colaboracion configurar impresora ip equipo ip area caja cajero auxiliar | Medio | Bajo | 0.769566 |
| **6920** | buen dia solicito colaboracion revisar equipo ip permit descargar archivos salen blanco | Bajo | Bajo | 0.918319 |
| **2519** | buen dia solicita verificacon revis telefono linea judith eliana gallego aria cc dond apliacion wasap pidiendo actualizacion vencio enero deja ingresar pide contrasena gracia | Bajo | Bajo | 0.945838 |
| **3728** | solicita deshabilitar cajero princip usuario gpaleman cc guisella paola aleman aria habilitar asesora servicios habilitar cajero princip usuario afpenam cc andr felip pena mendoza motivo reintegro | Alto | Medio | 0.996351 |
| **3341** | bueno diassolicito amabl colaboracion trasladar saldo relacionado archivo adjunto deudabeneficio mucha gracia | Bajo | Bajo | 0.980242 |
| **533** | buen dia adjunto envio formato traslado correspondient funcionario hernan dario ruano pasa region narino ipial mistares | Bajo | Bajo | 0.673486 |
| **7745** | buena tardes favor validar dac client lui antonio taquez mues cc operacion cta cuota venc fecha debitado gracia | Medio | Medio | 0.784734 |

| | | | | |
|---|---|---|---|---|
| **862** | buena tard favor amabl colaboracion configuracion scaner permit scanear desd bantot gracia | Medio | Medio | 0.589931 |
| **8784** | buen dia solicito favor habilitar usuario fjblanco fernando blanco blanco cc cajero princip tien perfil cajero gracia | Medio | Medio | 0.996178 |
| **10870** | buena tardes informo carpeta constitucion garantiaminuta oficina encuentran documento elaboracion minuta senora sorleni yueth uriel dodino cc quedo atenta | Bajo | Bajo | 0.959643 |
| **13571** | bueno dias adjunto fm realizar eliminacion adicion firma oficina pda gualmatan momento solo contamo sola firma gracia colaboracion | Medio | Bajo | 0.866326 |
| **3684** | requier reclasificar saldo encuentran rubro saldo cuenta ahorro cancelada saldo cuenta ahorro saldadas saldo pertenecen traslado tesoro nacion realizado rubro cancelada debe existir saldo acuerdo ... | Bajo | Bajo | 0.964830 |
| **8268** | cordial saludo favor colaboran permit ingresar sara radicar incapacidad favor colaboran gracias ip | Medio | Bajo | 0.571670 |
| **12001** | buena tarde tan amabl pued colaborar habilitando opcion leer proceso estan dispuesto intranet permit leer contenido tal mucha gracia yonathan david ch calamba cel asesor servicio pda balboa | Bajo | Bajo | 0.447875 |
| **10608** | cordial saludo solicito favor trasladar jagarciap kennedi bosa brasil gracia | Medio | Medio | 0.955282 |
| **11887** | bueno dias adjunto solicitud chequ gerencia pago impuesto declaracion reteica autorretencion mpio espin diciembr gracias | Medio | Medio | 0.737590 |
| **10836** | lista restrictiva cliente leidi viviana nastar coral adjunta documentacion consulta lista restrictiva cliente leidi viviana nastar coral cc cliente conyug codeudor tipo producto credinegocios tipo... | Alto | Medio | 0.984340 |
| **13371** | siniestro jhon blanco gamez seguro familia protegida adjunta documento gestion gracia | Bajo | Bajo | 0.998853 |
| **2760** | amablement solicita tapet bienvenida agencia armenia centro debido actual presenta deterioro adjunto foto quedo atenta | Bajo | Bajo | 0.709557 |
| **2471** | buen dia envio report pantalla dinamica agencia cartago aparec cobtenido tien mensaj pantalla presion ctrl alt supr gracias | Bajo | Bajo | 0.949788 |
| **10886** | buena tardes solicito aprovisionamiento efectivo dosciento cincuenta millon peso moneda corriente | Medio | Medio | 0.990938 |
| | bueno dia solicita colaboracion habilitar usuario fbanol flor camila banol largo cc analista | | | |

| | | | | |
|---|---|---|---|---|
| **7313** | credito ingresa incapacidad correspondient pda riosucio | Medio | Medio | 0.601745 |
| **9875** | buena tard solicita autorizacion ingreso cuarto sistema retiro antena cisco | Medio | Medio | 0.941011 |
| **3601** | buena tardes pido colaboracion verificar conincidencia sarlaft senor diego fernando rodriguez cc client titular credito renovacion segmento agropecuario gracia | Medio | Medio | 0.993422 |
| **3780** | buen dia solicito colaboracion autorizar apertura cdt nombr maria rubiela lozano c c plazo dia dinero efectivo provien cancelacion cdt pasado diciembr ano banco mundo mujer igual manera anexa decl... | Alto | Alto | 0.970120 |
| **14206** | buena tardes adjunto solicitud compensacion vacacion nombr analista credito pedro pablo romero moral cc | Bajo | Bajo | 0.783516 |
| **2619** | buen dia solicito colaboracion anular cxc analista credito wilmar fernando nino cta valor error quedo ingresado numero credito errado adjunto soport | Medio | Medio | 0.922723 |
| **13316** | solicito amabl colaboracion realizar envio tarjeta debito agencia piendamo debido contamo suficient operacion diaria oficina | Bajo | Bajo | 0.848876 |

# Ajuste del modelo de Máquinas de soporte vectorial con TF-IDF

In [83]:
```
svm_pipeline = Pipeline([('tfidf', TfidfVectorizer()),
 ('svm', LinearSVC(random_state=42))
 ])

param_grid = {'tfidf__ngram_range': [(1, 1), (1, 2)],
 'svm__C': [0.01, 0.1, 1, 5]
}

gs_svm = GridSearchCV(svm_pipeline, param_grid, cv=5, verbose=2)
gs_svm = gs_svm.fit(train_corpus, train_label_names)
```

```
Fitting 5 folds for each of 8 candidates, totalling 40 fits
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 1); total time=   0.1s
```

```
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END .............svm__C=0.01, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 1); total time=   0.1s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 2); total time=   0.3s
[CV] END ..............svm__C=0.1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 1); total time=   0.2s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 1); total time=   0.2s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 1); total time=   0.2s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 1); total time=   0.2s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 1); total time=   0.2s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=1, tfidf__ngram_range=(1, 2); total time=   0.4s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 1); total time=   0.3s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 1); total time=   0.3s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 1); total time=   0.3s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 1); total time=   0.3s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 1); total time=   0.3s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 2); total time=   0.6s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 2); total time=   0.6s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 2); total time=   0.6s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 2); total time=   0.6s
[CV] END ................svm__C=5, tfidf__ngram_range=(1, 2); total time=   0.7s
```

In [84]:
```
gs_svm.best_estimator_
```

Out[84]:
```
Pipeline(steps=[('tfidf', TfidfVectorizer(ngram_range=(1, 2))),
                ('svm', LinearSVC(C=1, random_state=42))])
```

In [85]:
```
# evaluar el modelo mejor sintonizado en el conjunto de datos de prueba

best_svm_test_score = gs_svm.score(test_corpus, test_label_names)
print('Test Accuracy :', best_svm_test_score)
```

```
Test Accuracy : 0.824390243902439
```

In [86]:
```
predict_probas_2 = gs_svm.decision_function(test_corpus).max(axis=1)
test_df_2 = data_clean.iloc[test_idx]
test_df_2['Predicted Name'] = svm_predictions
test_df_2['Predicted Confidence'] = predict_probas_2
test_df_2.head(60)
```

Out[86]:

| | Descripcion_limpia | Category | Predicted Name | Predicted Confidence |
|---|---|---|---|---|
| **4450** | buen dia envia solicitud correccion pago operacion cliente fabiola bermudez casierra c c error recaudo numero credito adjunto carta solicitud soport pago gracias | Medio | Medio | 1.119387 |
| **7444** | bueno dia favor rechazar delegar insatncia permit tomar dar continuidad proceso | Alto | Alto | 1.203841 |
| **7050** | bueno dia carpeta siniestro encuntra reclamacion seguro fp lui iriart salgado cc | Bajo | Bajo | 1.255771 |
| **10313** | buena tardes solicto colaboracion aplicar pago credito client diego marin agudelo cc cobro comis juridica acordada abogado adjunta pago correo | Medio | Medio | 0.678883 |
| **12922** | buen dia adjunto documento client monica mafla mora credito condonacion conceptos | Bajo | Bajo | 0.922564 |
| **8764** | solicito favor habilitar coordinadora servicio neiva altico gina lorena salgado roja usuario gsalgado partir hoy enero debido retoma cargo despu reemplazar vacacion c r c region huila deshabilitar... | Medio | Medio | 0.737344 |
| **6575** | buen dia solicito colaboracion actualizacion correo usuario afburbano | Medio | Alto | 0.258155 |
| **5403** | buen dia adjunta report disfrut vacaciones | Medio | Medio | 0.957791 |
| **476** | desbloqueo clave datacredito | Bajo | Bajo | 1.069401 |

| | | | | |
|---|---|---|---|---|
| **10950** | buen dia solicito cambio perfil cajero princip oswal fernando rozo c c orozo cajera princip laura daniela ceron c c ldceron incapacitada dia gracia | Medio | Medio | 1.388007 |
| **1139** | bueno dias adjunto report client rosmeri montenegro parodi cc fin validar categoria puesto aparec recurrent ant pandemia elit plus anterior sujetro nuevo cambio sarc agradezco colaboracion afectac... | Medio | Medio | 1.024909 |
| **12063** | cordial saludo favor autorizar poliza plan hipertens tiroid rosa lobon gil cc quedo atenta autorizacion gracia | Alto | Alto | 0.918321 |
| **4245** | solicita revisar cartera digit analista credito permit conexion usuario registrado | Bajo | Bajo | 1.004977 |
| **884** | buen dia adjunta report cambio cargo | Medio | Medio | 0.970941 |
| **11515** | bueno dias favor enviar provis gracia | Medio | Medio | 0.981435 |
| **10558** | buen dia solicito habilitar luz arelli valencia usuario lvalencia coordinadora servicios martha luci jurado usuario mjurado cajera princip juliana jaramillo usuario jjaramil asesora servicio pda s... | Medio | Medio | 0.926440 |
| **99** | buen dia favor consutar lista restrictiva titular ampliacion fabian andr gomez identificado cc mucha gracias quedo atenta | Medio | Medio | 1.258077 |
| **3467** | buena tardes solicito apoyo habilitar coordinador servicio popayan ciudad jardin usuario apgonzale angela patricia gonzalez santand deshabilitar usuario yramirezt yuri katerin ramirez tobar estara... | Medio | Medio | 0.828089 |
| **9214** | buena tarde solicita favor deshabilitar clave cajero princip estara vacacion asignar clave bakcup boveda asesor servicios adjunto envio formato solicitud | Medio | Medio | 0.624532 |
| **9419** | bueno dia envio certificado libertad viabilidad credito gracia | Alto | Alto | 0.505366 |
| **14010** | bueno dias permito solicitar habilitar cajero princip lui miguel mayorga asesora servicio magda milena martinez perfil asesora servicio cajera auxiliar mucha gracia | Medio | Medio | 1.050605 |
| **6325** | celular conecta vpn celular analista credito carlo antonio casanova c c usuario ccasanova celular | Bajo | Bajo | 1.463462 |
| **5226** | buen dia solicito colaboracion autorizacion cuarto sistema part prosegur validar disponibilidad ampliacion sistema seguridad techo oficina libertador robo presntado agencia quedo atenta | Medio | Medio | 1.131851 |
| **1984** | solicita prueba panico | Alto | Alto | 1.548528 |

| | | | | |
|---|---|---|---|---|
| **290** | bueno dias solicito colaboracion validacion usuario rcastaneda acceso adminfo genera do errores horario permitido error integracion mucha gracias adjunto pantallazo | Bajo | Bajo | 0.331075 |
| **6684** | buena tard adjunto solicitud devolucion familia protegida karina medina torr cc precancelacion credito | Bajo | Bajo | 1.063379 |
| **7632** | bueno dia alexis favor restablec credencial acceso sftp talento humano | Bajo | Bajo | 0.678403 |
| **2039** | buen dia solicito amablement cuenta operacion senor rocha gomez manuel mucha gracias | Medio | Bajo | -0.041600 |
| **13200** | adjunto solicitud ajust pago credito sarria chasoy yulli tatiana operacion client cancelo credito mora | Medio | Medio | 1.086626 |
| **9097** | solicito favor asignar perfil ppardo auxopereca ambient preproduccion sucurs | Medio | Medio | 0.967523 |
| **9052** | envian soport reclamacion devolucion prima seguro familia protegida nombr sra lucero lopez cc sede palmira saman | Bajo | Bajo | 1.384665 |
| **3187** | buen dia favor reestablec contrasena funcionario giovanni alexand chamorro c c nro | Medio | Alto | 0.056969 |
| **5623** | bueno dias favor pido amabl colaboracion configurar impresora ip equipo ip area caja cajero auxiliar | Medio | Medio | 0.321440 |
| **6920** | buen dia solicito colaboracion revisar equipo ip permit descargar archivos salen blanco | Bajo | Bajo | 0.798591 |
| **2519** | buen dia solicita verificacon revis telefono linea judith eliana gallego aria cc dond apliacion wasap pidiendo actualizacion vencio enero deja ingresar pide contrasena gracia | Bajo | Bajo | 0.936408 |
| **3728** | solicita deshabilitar cajero princip usuario gpaleman cc guisella paola aleman aria habilitar asesora servicios habilitar cajero princip usuario afpenam cc andr felip pena mendoza motivo reintegro | Alto | Medio | 1.502490 |
| **3341** | bueno diassolicito amabl colaboracion trasladar saldo relacionado archivo adjunto deudabeneficio mucha gracia | Bajo | Bajo | 0.916232 |
| **533** | buen dia adjunto envio formato traslado correspondient funcionario hernan dario ruano pasa region narino ipial mistares | Bajo | Bajo | 0.054764 |
| **7745** | buena tardes favor validar dac client lui antonio taquez mues cc operacion cta cuota venc fecha debitado gracia | Medio | Medio | 0.300456 |
| **862** | buena tard favor amabl colaboracion configuracion scaner permit scanear desd bantot gracia | Medio | Medio | -0.008865 |
| | buen dia solicito favor habilitar usuario fjblanco fernando blanco blanco cc cajero princip | | | |

| | | | | |
|---|---|---|---|---|
| **8784** | tien perfil cajero gracia | Medio | Medio | 1.462431 |
| **10870** | buena tardes informo carpeta constitucion garantiaminuta oficina encuentran documento elaboracion minuta senora sorleni yueth uriel dodino cc quedo atenta | Bajo | Bajo | 0.873424 |
| **13571** | bueno dias adjunto fm realizar eliminacion adicion firma oficina pda gualmatan momento solo contamo sola firma gracia colaboracion | Medio | Bajo | 0.543087 |
| **3684** | requier reclasificar saldo encuentran rubro saldo cuenta ahorro cancelada saldo cuenta ahorro saldadas saldo pertenecen traslado tesoro nacion realizado rubro cancelada debe existir saldo acuerdo ... | Bajo | Bajo | 0.865475 |
| **8268** | cordial saludo favor colaboran permit ingresar sara radicar incapacidad favor colaboran gracias ip | Medio | Bajo | 0.019234 |
| **12001** | buena tarde tan amabl pued colaborar habilitando opcion leer proceso estan dispuesto intranet permit leer contenido tal mucha gracia yonathan david ch calamba cel asesor servicio pda balboa | Bajo | Bajo | -0.220535 |
| **10608** | cordial saludo solicito favor trasladar jagarciap kennedi bosa brasil gracia | Medio | Medio | 0.831628 |
| **11887** | bueno dias adjunto solicitud chequ gerencia pago impuesto declaracion reteica autorretencion mpio espin diciembr gracias | Medio | Medio | 0.310814 |
| **10836** | lista restrictiva cliente leidi viviana nastar coral adjunta documentacion consulta lista restrictiva cliente leidi viviana nastar coral cc cliente conyug codeudor tipo producto credinegocios tipo... | Alto | Medio | 0.977978 |
| **13371** | siniestro jhon blanco gamez seguro familia protegida adjunta documento gestion gracia | Bajo | Bajo | 1.743151 |
| **2760** | amablement solicita tapet bienvenida agencia armenia centro debido actual presenta deterioro adjunto foto quedo atenta | Bajo | Bajo | 0.225196 |
| **2471** | buen dia envio report pantalla dinamica agencia cartago aparec cobtenido tien mensaj pantalla presion ctrl alt supr gracias | Bajo | Bajo | 0.836377 |
| **10886** | buena tardes solicito aprovisionamiento efectivo dosciento cincuenta millon peso moneda corriente | Medio | Medio | 1.246465 |
| **7313** | bueno dia solicita colaboracion habilitar usuario fbanol flor camila banol largo cc analista credito ingresa incapacidad correspondient pda riosucio | Medio | Bajo | 0.071171 |
| **9875** | buena tard solicita autorizacion ingreso cuarto sistema retiro antena cisco | Medio | Medio | 0.796530 |

| | | | | |
|---|---|---|---|---|
| **3601** | buena tardes pido colaboracion verificar conincidencia sarlaft senor diego fernando rodriguez cc client titular credito renovacion segmento agropecuario gracia | Medio | Medio | 1.323234 |
| **3780** | buen dia solicito colaboracion autorizar apertura cdt nombr maria rubiela lozano c c plazo dia dinero efectivo provien cancelacion cdt pasado diciembr ano banco mundo mujer igual manera anexa decl... | Alto | Alto | 1.066835 |
| **14206** | buena tardes adjunto solicitud compensacion vacacion nombr analista credito pedro pablo romero moral cc | Bajo | Bajo | 0.204229 |
| **2619** | buen dia solicito colaboracion anular cxc analista credito wilmar fernando nino cta valor error quedo ingresado numero credito errado adjunto soport | Medio | Medio | 0.645735 |
| **13316** | solicito amabl colaboracion realizar envio tarjeta debito agencia piendamo debido contamo suficient operacion diaria oficina | Bajo | Bajo | 0.523454 |

# Análisis de escenarios

## Escenario 1: regresión logística con TF-IDF

In [87]:
```
grafica_1 = pd.DataFrame(test_df_1["Category"]==test_df_1["Predicted Name"])
```

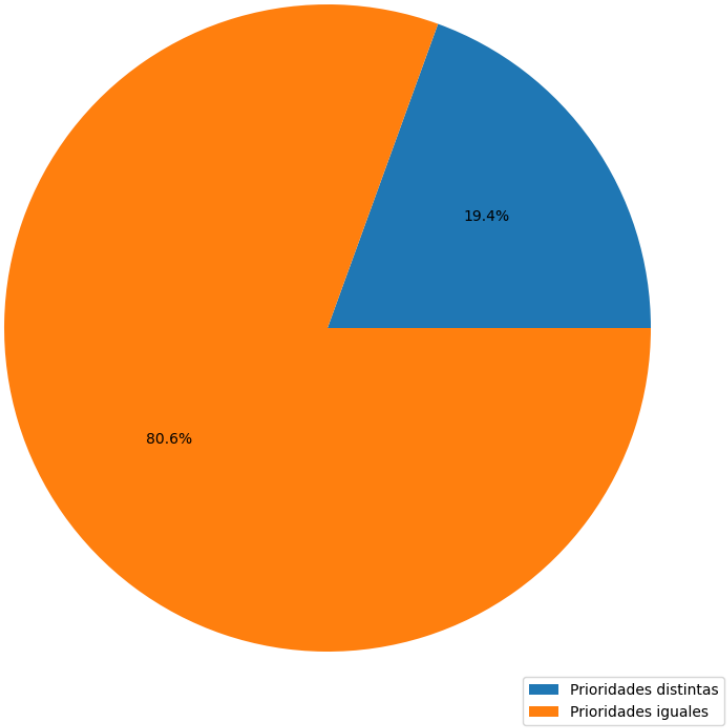
In [88]:
```
grafica_1 = grafica_1.rename(columns={0:'Comparación'})
```

In [89]:
```
print(grafica_1.groupby('Comparación').size())

fig, ax = plt.subplots(figsize = (10, 10))
ax.pie(grafica_1.groupby(grafica_1['Comparación']).size(), autopct='%1.1f%%')
ax.set_ylabel('')
labels=["Prioridades distintas", "Prioridades iguales"]
ax.legend(labels, loc = 'lower right')
ax.set_title('Prioridad asignada manualmente VS Prioridad asignada por el modelo',
             fontdict = {'fontsize':30,'fontweight':'bold','color':'black'})
plt.show()
```

```
Comparación
False     917
True     3798
dtype: int64
```

**Prioridad asignada manualmente VS Prioridad asignada por el modelo**




# Escenario 2: Máquinas de soporte vectorial con TF-IDF

In [90]:
```
grafica_2 = pd.DataFrame(test_df_2["Category"]==test_df_2["Predicted Name"])
```

In [91]:
```
grafica_2 = grafica_2.rename(columns={0:'Comparación'})
```

In [92]:
```
print(grafica_2.groupby('Comparación').size())

fig, ax = plt.subplots(figsize = (10, 10))
ax.pie(grafica_2.groupby(grafica_2['Comparación']).size(), autopct='%1.1f%%')
ax.set_ylabel('')
labels=["Prioridades distintas", "Prioridades iguales"]
ax.legend(labels, loc = 'lower right')
ax.set_title('Prioridad asignada manualmente VS Prioridad asignada por el modelo',
             fontdict = {'fontsize':30,'fontweight':'bold','color':'black'})
plt.show()
```

```
Comparación
False     883
True     3832
dtype: int64
```

# Prioridad asignada manualmente VS Prioridad asignada por el modelo