UNIVERSIDAD CENTRAL	**FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS** **MAESTRÍA EN ANALÍTICA DE DATOS** **REGISTRO PROPUESTA DE TRABAJO DE GRADO** **2025-I**

REGISTRO DE TRABAJO DE GRADO

FECHA	02	05	2025

DATOS DEL ESTUDIANTE (S)

NOMBRES: Michael David							APELLIDOS: Gualteros Garcia
TIPO IDENTIFICACIÓN:	T.I.		C.C.	X	C.E.		NÚMERO: 1023980438
CORREO INSTITUCIONAL: mgualterosg@ucentral.edu.co							TELÉFONO: 3223750389

NOMBRES: Raúl Andrés							APELLIDOS: Gamba Hastamorir
TIPO IDENTIFICACIÓN:	T.I.		C.C.	x	C.E.		NÚMERO: 1023003554
CORREO INSTITUCIONAL: rgambah@ucentral.edu.co							TELÉFONO: 3197660529

NOMBRES: Diana Carolina							APELLIDOS: Gómez Boada
TIPO IDENTIFICACIÓN:	T.I.		C.C.	X	C.E.		NÚMERO: 1030583926
CORREO INSTITUCIONAL: dgomezb10@ucentral.edu.co							TELÉFONO: 3212539359

MODALIDAD DE TRABAJO DE GRADO (Seleccione una opción)

	II. Modalidad de profundización: □
	a. Trabajo monográfico

Línea de profundización
Modelo de machine learning y procesamiento de lenguaje natural.

UNIVERSIDAD CENTRAL	**FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS** **MAESTRÍA EN ANALÍTICA DE DATOS** **REGISTRO PROPUESTA DE TRABAJO DE GRADO** **2025-I**

AVAL DEL DOCENTE DIRECTOR

NOMBRES: luis andres campos maldonado	DEPARTAMENTO:
CORREO INSTITUCIONAL: lcamposm@ucentral.edu.co	TELÉFONO-EXT. : +

COMPONENTES

1. TÍTULO DEL TRABAJO DE GRADO

Segmentación Automatizada de Incidentes en una Mesa de Servicio para un Fondo de Pensiones mediante Machine Learning, con el diseño de una interfaz interactiva para el usuario

2. INTRODUCCIÓN Y JUSTIFICACIÓN (máximo 1500 palabras)

Los fondos de pensiones en Colombia fueron creados en 1945 mediante la Ley 1600, que dio origen a la Caja Nacional de Jubilaciones y Pensiones de los Empleados Públicos (la Caja de Jubilaciones), con el objetivo de proporcionar pensiones a los empleados del Estado (Zúñiga, s.f.). Posteriormente, en 1993, se implementó la Ley 100, que reformó el sistema de seguridad social, estableciendo el Sistema General de Pensiones, Salud y Riesgos Profesionales. Esta ley aplicó a todos los habitantes del territorio nacional, con algunas excepciones, como los miembros de las Fuerzas Militares y de la Policía Nacional, personal civil del Ministerio de Defensa vinculado con anterioridad a la ley, trabajadores de la Empresa Colombiana de Petróleos (Ecopetrol) y los maestros públicos afiliados al Fondo Nacional del Magisterio (Zúñiga, s.f.).

La Ley 100 introdujo dos regímenes: el Régimen de Prima Media (RPM) y el Régimen de Ahorro Individual con Solidaridad (RAIS). Posteriormente, con la entrada en vigencia de la reforma pensional, se creó la Ley 797 de 2003, con el objetivo de hacer el sistema pensional más sostenible financieramente para el Régimen de Prima Media (RPM). Estas entidades tienen como propósito

principal asegurar que los trabajadores cuenten con una pensión o garantizarles una fuente de ingresos tanto al alcanzar la edad de jubilación como en situaciones excepcionales, como la invalidez o la supervivencia.

En un mundo donde las actualizaciones tecnológicas son cada vez más intensas y el avance de las inteligencias artificiales es cada vez más impactante, en Colombia se han comenzado a implementar estas herramientas con el propósito de optimizar procesos y permitir el crecimiento de las empresas colombianas. El país ha venido creciendo exponencialmente con el desarrollo y aplicación de estas tecnologías. Algunas entidades educativas como, la Universidad Central, Universidad de los Andes, Universidad Jorge Tadeo Lozano y Universidad El Bosque, entre otras, optaron por ofrecer carreras que permiten el aprendizaje en este medio, como es la Maestría en Análisis de Datos, que permite generar profesionales más competentes para brindar un uso adecuado de estas herramientas y hacer uso de lo aprendido en diferentes campos empresariales del país. Esto permite a las empresas que se encuentran interesadas en esta tecnología desarrollar capacidades para contar con información a tiempo y mejorar sus decisiones.

En virtud a la automatización de procesos y la mejora en los tiempos de respuesta, muchas empresas han mostrado interés en dar uso a estas tecnologías. Una de ellas es Colfondos, cuyo propósito es brindar respuestas más rápidas a sus clientes, ya que sus operaciones diarias generan diferentes solicitudes que esperan ser atendidas y resueltas en el menor tiempo posible. Para ello, se propone implementar un asistente virtual que facilite a los operadores de Colfondos brindar las respuestas de manera más ágil.

Este proyecto se centrará en establecer un modelo de procesamiento de lenguaje natural (PLN), el cual permitirá automatizar la interpretación, análisis y procesamiento del texto de las solicitudes radicadas por los usuarios funcionales.

Actualmente, el fondo de pensiones y cesantías Colfondos S.A. enfrenta un alto volumen de solicitudes relacionadas con la atención al cliente. Para gestionar dicho volumen, cuentan con una mesa de servicio encargada de resolver las incidencias que no pueden ser solucionadas en un análisis previo. Esto ocurre

porque, en algunos casos, los escalamientos requieren una solución técnica o los responsables del área usuaria carecen del conocimiento necesario para resolverlos. La mesa de servicio centraliza las incidencias escaladas por los funcionarios internos, que están relacionadas con fallas identificadas en las aplicaciones y productos ofrecidos por el fondo de pensiones y cesantías.

Estas incidencias permiten identificar problemas técnicos, solicitudes recurrentes y requerimientos de análisis de información. Sin embargo, debido al elevado número de solicitudes registradas mensualmente, surge la necesidad de atender casos reiterativos, lo que genera reprocesos y un aumento en la carga operativa para dar respuesta a los casos. La importancia de plantear mejoras en esta parte del proceso radica en que, al revisar las cifras de escalamientos, se encontró que, en 2024, Colfondos recibió 12,741 quejas, lo que representa el 17.8% del total a nivel gremial, ubicándose como el segundo fondo con más quejas escaladas. Esto hace necesario un argumento para proponer mejoras en este proceso de atención a incidencias, con el fin de reducir las cifras de quejas mediante metodologías que impulsen respuestas oportunas en los escalamientos de primer nivel de los clientes.

Con la implementación del PLN, se busca reducir los tiempos de atención y, al mismo tiempo, realizar una segmentación automática de los incidentes, lo que permitirá una respuesta automática de primer nivel. De esta manera, el usuario funcional podrá auto gestionarse a través de una respuesta predefinida, optimizando los recursos dedicados a la gestión de solicitudes y permitiendo que se enfoquen en casos de mayor complejidad y valor estratégico para la organización.

La pregunta de investigación que busca resolver este trabajo es: **¿Cómo se puede segmentar de manera efectiva los casos recurrentes en la mesa de servicio de Colfondos, con el fin de ofrecer una respuesta automatizada de primer nivel?** Esta interrogante surge de la necesidad de optimizar la gestión de incidentes en la organización, dado que el volumen de solicitudes de soporte ha crecido significativamente, afectando los tiempos de respuesta y aumentando la carga operativa del personal encargado. Actualmente, muchos de estos casos corresponden a problemas recurrentes que podrían resolverse de manera automatizada si se cuenta con un modelo capaz de clasificar y responder solicitudes de primer nivel de manera efectiva.

Este estudio emplea técnicas de procesamiento de lenguaje natural y aprendizaje automático para transformar los datos históricos de la mesa de servicio en información estructurada y útil. Se usarán modelos de embeddings (Raschka, 2019) para representar el contenido de los casos, permitiendo identificar similitudes entre solicitudes mediante la reducción de dimensionalidad y la captura de contexto. Posteriormente, se aplicarán técnicas de clustering, con el objetivo de agrupar incidentes similares y generar categorías estandarizadas. Este proceso permitirá estructurar los datos históricos y establecer patrones que faciliten la automatización de respuestas.

La implementación de este procedimiento no solo incrementará la eficiencia de la operación en la mesa de servicio al disminuir la intervención manual, sino que también favorecerá una mejor distribución de los recursos dentro de Colfondos. Varios estudios han evidenciado que el uso de NLP y aprendizaje automático en sistemas de solución de casos puede disminuir significativamente el tiempo de respuesta y mejorar la solución de incidentes a través de modelos de clasificación automática (Qamili et al., 2018; Venegas Villarreal, Villar García & Mendoza De Los Santos, 2022). Además, incorporar estas tecnologías en la mesa de servicio de Colfondos facilitará una transición hacia un modelo más escalable y flexible, en el que las soluciones puedan mejorar gradualmente a medida que se aprenden nuevos casos.

Finalmente, este estudio no solo tendrá un impacto en la optimización operativa, sino que también sentará las bases para el desarrollo de una interfaz interactiva, capaz de proporcionar respuestas automáticas basadas en el conocimiento adquirido. Esta interfaz no solo reducirá la carga del equipo de soporte, sino que mejorará la experiencia de los usuarios internos al proporcionarles respuestas precisas y rápidas. La combinación de NLP, clustering y embeddings permitirá que el sistema evolucione de manera continua, asegurando que las respuestas ofrecidas sean cada vez más efectivas y alineadas con las necesidades reales de la organización.

3. OBJETIVO GENERAL

Desarrollar modelos de Machine Learning para la segmentación de casos en la mesa de servicio interna de Colfondos, con el fin de automatizar y a través de una interfaz interactiva brindar respuestas de primer nivel a casos recurrentes para reducir la carga operativa del equipo de soporte.

4. OBJETIVOS ESPECÍFICOS

- Realizar la recolección, limpieza y análisis exploratorio de la base de datos de casos generados en la mesa de servicio, con el fin de identificar patrones y tendencias relevantes que permitan optimizar su gestión.

- Implementar técnicas de clustering para agrupar casos recurrentes según patrones comunes, asegurando una segmentación adecuada y evaluando su desempeño mediante métricas como el índice de silueta y el coeficiente de Dunn.
- Desplegar los resultados obtenidos a través de una interfaz interactiva, que facilite al usuario escribir un caso y le sugiera soluciones de primer nivel basadas en los grupos definidos.

5. ANTECEDENTES Y MARCO TEÓRICO (máximo 3000 palabras)

El crecimiento acelerado de las Tecnologías de la Información ha transformado la manera en que las empresas gestionan la atención al cliente y la resolución de incidencias. La gestión de servicios ha evolucionado para garantizar la continuidad operativa y mejorar la eficiencia de los procesos de soporte (Zuev et al., 2018). En este contexto, la gestión de incidentes se ha convertido en una práctica clave, ya que permite registrar, clasificar y dar respuestas de primer nivel a distintos problemas.

Las mesas de servicio o service desks en la gestión de casos juegan un papel importante en el registro de incidencias, sin embargo, su dependencia histórica con la intervención manual en el procesamiento

de tickets o casos de soporte, ha generado retos significativos como, demoras en la atención de errores, clasificación y asignación ineficiente de recursos. En este punto, la implementación de la inteligencia artificial y el aprendizaje automatizado pueden optimizar el proceso de atención a incidentes, siendo más exactos con la respuesta al usuario, disminuyendo tiempos de respuesta y minimizando reprocesos

Dado que el número de solicitudes en las empresas sigue aumentando debido a los esfuerzos de digitalización, los errores en la clasificación y asignación de casos pueden incrementar significativamente los costos operativos y los tiempos de resolución. Esto, a su vez, afecta negativamente la satisfacción del cliente y perjudica la experiencia del usuario final (Fuchs et al., 2022).

El avance de la inteligencia artificial (IA) y el procesamiento de lenguaje natural (PLN) ha permitido la implementación de soluciones más eficientes para la gestión de incidentes, y esta tendencia se refleja en el creciente número de publicaciones sobre el tema.

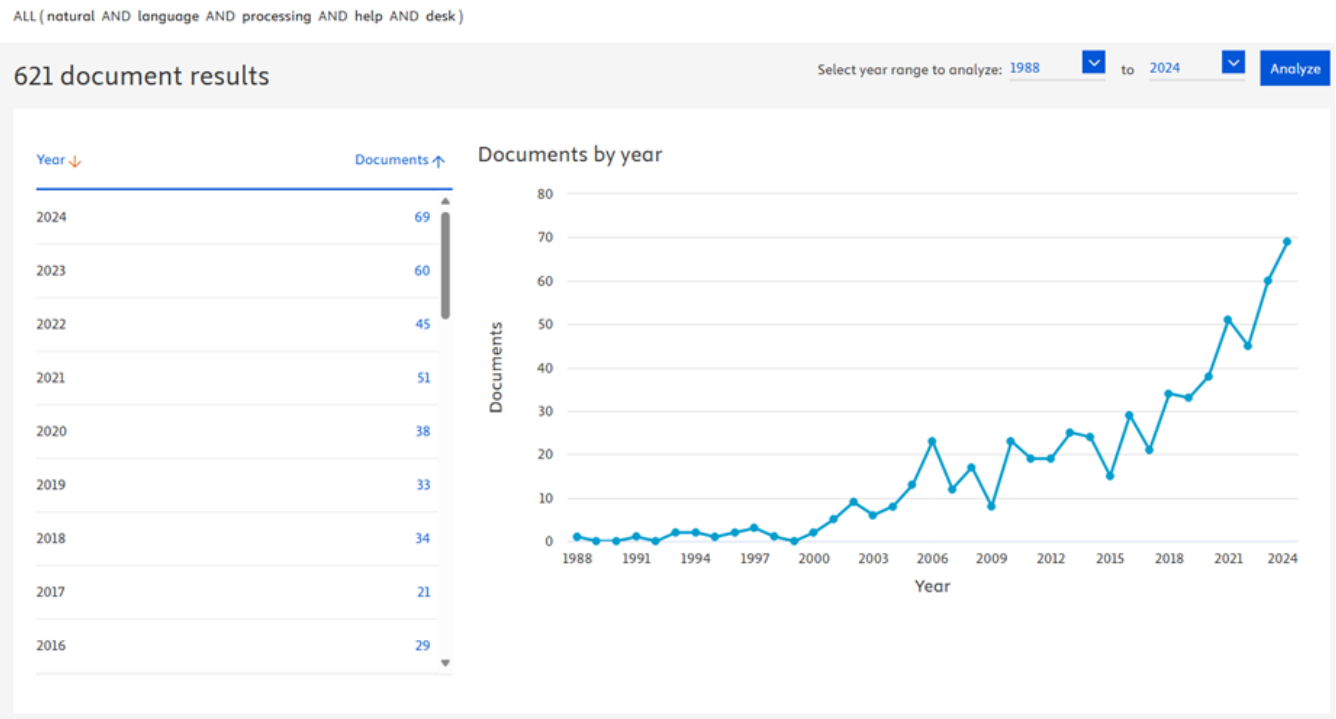

Imagen 1. Documentos publicados anualmente relacionados con lenguaje natural y sistemas de mesa de ayuda, según datos obtenidos de Scopus. Elaboración propia.

En la actualidad, se utilizan modelos basados en aprendizaje automático para perfeccionar la categorización automática de casos, disminuir el tiempo de respuesta y mejorar la distribución de recursos en las mesas de servicio (Vital et al., 2024).

Dentro de los enfoques más empleados para mejorar la eficiencia en la gestión de incidentes se encuentran:

- Modelos de clasificación supervisada: Algoritmos como random forest y redes neuronales profundas han sido utilizados para la categorización de incidentes.
- Embeddings y representación semántica: Técnicas como embeddings basados en transformers han mejorado la comprensión contextual de los incidentes reportados.
- Modelos avanzados de PLN: Arquitecturas como BERT, T5 y GPT han demostrado gran capacidad para analizar y generar respuestas a partir de texto libre, facilitando la automatización de procesos en las mesas de servicio.

Estos avances han permitido reducir la carga operativa del equipo de asistencia de primer nivel, mejorar la categorización de incidentes y aumentar la exactitud en la asignación de casos a los grupos de solución pertinentes. En muchos casos, el uso de machine learning no sólo apoya, sino que puede incluso reemplazar algunas tareas realizadas por los operadores de primer nivel, mejorando la eficacia del proceso (Venegas Villarreal, Villar García & Mendoza De Los Santos, 2022).

Diversos estudios han aplicado modelos de machine learning para mejorar la gestión de incidentes en mesas de servicio. Un ejemplo es el trabajo de Qamili et al. (2018), quienes utilizaron aprendizaje automático para proponer un marco inteligente que optimiza la gestión de sistemas de tickets. Este marco aborda tres desafíos principales: la detección de spam, la asignación automática de tickets y el análisis de sentimientos. El estudio empleó un conjunto de datos de 18,917 registros provenientes de un sistema de tickets de una empresa de desarrollo de software. El proceso incluyó la aplicación de técnicas de limpieza de texto para eliminar signos de puntuación y palabras irrelevantes, seguido de la

representación de los datos mediante un enfoque de "bag-of-words", donde cada palabra individual se convirtió en una característica. Los modelos entrenados, como SVM, Random Forest y SGD, mostraron un desempeño superior en términos de precisión y consistencia, mientras que los Decision Trees presentaron menores niveles de precisión. Este enfoque permitió mejorar la eficiencia en la asignación de tickets a los departamentos correspondientes y minimizar los falsos positivos en la clasificación de spam.

A nivel local, un ejemplo de implementación exitosa lo encontramos en la investigación de Ramírez Devia (2021), quien desarrolló un modelo de clasificación y priorización de gestión de PQRS en Colsubsidio, basado en procesamiento de lenguaje natural y aprendizaje automático. En su estudio, se utilizó Naïve Bayes, árboles de decisión y máquinas de vectores de soporte (SVM) para vectorizar los textos de las PQRS y clasificarlos de manera eficiente. El modelo alcanzó un 94.5% de equilibrio constante entre las peticiones de los clientes, demostrando que este tipo de sistemas puede optimizar la gestión de solicitudes en organizaciones como Colsubsidio, con la posibilidad de seguir entrenando el modelo para mejorar la precisión y el recall.

Otro ejemplo claro se presenta en el repositorio de la Universidad de Antioquia, que centró su investigación en validar la información manejada a nivel interno por la empresa Bancolombia. La compañía utilizó Microsoft Teams para conectar a sus colaboradores independientemente de su ubicación geográfica, con el fin de optimizar la gestión de comunicaciones internas. Dado el aumento en el uso de esta herramienta, se propuso un proyecto para validar, mediante modelos de machine learning, si los mensajes intercambiados tenían un propósito laboral o no. Para ello, se empleó la técnica GloVe, que utiliza vectores de 500 dimensiones para facilitar la clasificación de los mensajes en tres categorías: Negativo, Neutral y Positivo.

Dentro del análisis que se realiza en este repositorio se puede entender que los diferentes métodos de vectorización de palabras (KNN, SVM, RF y XGBoost), ajustaban a la clasificación de palabras y que en general 12 de las palabras ajustaban a una conversación de manera laboral y cuando una

conversación contiene más de 40 palabras es porque no se trataba de algo laboral. El porcentaje de conversaciones no laborales equivale aproximadamente a un 25% de las conversaciones obtenidas.

Para las conclusiones generales de este proyecto se pudo indicar que las metodologías más efectivas fueron las XG Boost y RF sobresalieron con una efectividad del 96% y de las demás técnicas utilizadas, pudieron brindar respuestas claras para lo que se estaba solicitando. finalmente son metodologías aplicables para resolver problemáticas relacionadas con el lenguaje natural.

6. FUENTE DE LOS DATOS (Cómo se obtendrán los datos)

Los datos utilizados en este estudio provienen de Colfondos, específicamente de su mesa de servicio interna. La información corresponde a un histórico de 45,000 casos registrados entre agosto de 2023 a febrero de 2025, los cuales contienen descripciones de problemas y resolución en texto libre. El acceso a estos datos fue otorgado como parte de una solicitud formal, en la que se especificó que la información sería utilizada exclusivamente con fines académicos para el desarrollo de este trabajo de grado.

Para garantizar el cumplimiento de las políticas internas de seguridad y privacidad de Colfondos, la empresa realizó un proceso de enmascaramiento antes de la entrega de los datos, asegurando que no contengan información sensible o identificable; el tratamiento de los datos incluirá pre procesamiento y limpieza, aplicando técnicas de procesamiento de lenguaje natural para estructurar la información y facilitar su análisis. Se garantizará el cumplimiento de normativas de privacidad y uso ético de los datos a lo largo de todo el proyecto.

7. APLICACIÓN Y/O APORTE ESPECÍFICO AL CAMPO

Este trabajo se enmarca principalmente en el campo de la analítica de datos, poniendo especial atención en el procesamiento de lenguaje natural y el aprendizaje automático. La implementación de estas técnicas permite convertir datos textuales sin estructura en información valiosa para la toma de decisiones, lo cual es crucial en la mejora de procesos en ambientes corporativos.

Bajo la perspectiva de la analítica de datos, este análisis ayuda a segmentar y organizar grandes cantidades de datos históricos de la mesa de servicio a través de técnicas de embeddings y clustering. Al agrupar incidentes parecidos y generar grupos estandarizados, el sistema puede identificar de una mejor forma patrones de recurrencia, lo que permite la automatización de respuestas a consultas frecuentes o casos similares.

Además de su impacto en la analítica de datos, este trabajo también realiza aportes significativos al campo de la ingeniería de software, específicamente en el desarrollo de asistentes virtuales inteligentes o chatbots para la gestión de soporte técnico. La implementación de un modelo de segmentación basado en NLP dentro de la mesa de servicio de Colfondos representa un avance hacia la automatización de procesos en las áreas de tecnología y servicio al cliente, lo que optimiza la asignación de recursos y mejora la experiencia del usuario.

Así mismo, desde una perspectiva organizacional, este proyecto tiene implicaciones en el ámbito de la gestión empresarial y la optimización de procesos operativos, al reducir la carga de trabajo manual en la mesa de servicio y mejorar los tiempos de respuesta a incidentes recurrentes.

8. METODOLOGÍA O ACTIVIDADES ESPECÍFICAS

La metodología que será usada en el proyecto será la CRISP-DM (Cross Industry Standard Process for Data Mining) ya que es uno de los métodos más usados en los proyectos de analítica de datos por su propuesta de trabajo estructurado en fases y etapas que permiten llevar un orden en la ejecución y brinda la posibilidad de identificar oportunidades de mejora en el ciclo del proyecto.

Esta metodología consta de 6 etapas que abarcan desde el entendimiento de las necesidades del negocio y de los datos que se usarán en el proyecto, hasta el despliegue de la solución que cumple con la necesidad planteada.

"CRISP-DM ayuda a las organizaciones a comprender el proceso de minería de datos y proporcionan una hoja de ruta a seguir mientras se planifica y lleva a cabo un proyecto de minería de datos" (The CRISP-DM model: The New Blueprint for Data Mining, 2000).

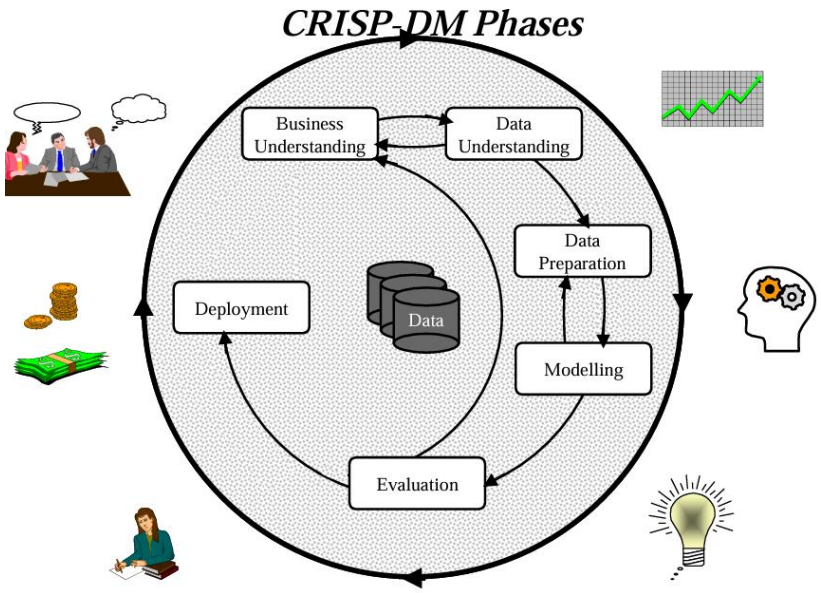

Imagen 2. Esquema de las fases del modelo CRISP-DM tomado de Chapman, P. (1999). The CRISP-DM User Guide.

El diseño del modelo CRISP-DM muestra de manera evidente su carácter cíclico y organizado, resaltando cómo las etapas interrelacionadas facilitan una implementación organizada y adaptable de proyectos de minería de datos. Este ciclo permite el intercambio constante entre fases como la comprensión del negocio, el entendimiento de los datos, su preparación, el modelado, la evaluación y la implementación. Además, su diseño visual resalta la relevancia de la iteración, dado que en numerosas situaciones se requiere retornar a etapas anteriores para efectuar modificaciones y mejoras, garantizando de esta manera la calidad y la capacidad de adaptación de las soluciones sugeridas. Esta perspectiva asegura que cada etapa del proyecto se encuentre en sintonía con los objetivos iniciales y con las demandas empresariales.

A continuación, se describe cada una de las fases del ciclo de vida en la metodología CRISP-DM:

- **Fase 1: Comprensión del Negocio**

Esta fase es la parte inicial y fundamental para comprender el comportamiento actual del negocio y de la misma forma poder identificar la necesidad que motiva la ejecución del proyecto planteado en este anteproyecto. A partir de lo anterior, se identificó que una de las problemáticas radica en los largo tiempos de respuesta que se generan en la atención de casos de la mesa de servicio y el uso de intervención humano requerida para la gestión de solicitudes, lo cual incrementa la carga operativa, genera reprocesos y afecta negativamente la satisfacción del usuario interno.

Actividades:

- ✓ Reuniones con la persona que conoce los datos para entender los tipos de incidentes y sus resoluciones.
- ✓ Identificar y segmentar las incidencias más frecuentes presentadas.
- ✓ Definición de criterios de éxito: precisión mínima del modelo, tasa de cobertura de respuestas automatizadas, reducción de tiempos de respuesta.

- ✓ Análisis del flujo de proceso de la atención de incidentes actual, para de esta forma proyectar la incorporación de la interfaz que se encarga de dar una respuesta de primer nivel

- **Fase 2: Comprensión de los Datos**

En la fase se analizan los objetivos y el contexto desde la perspectiva de la organización. donde se obtiene un conjunto de datos del fondo de pensiones el cual contiene el tipo de solicitud escalada por el usuario interno, donde se evidencia si es un requerimiento o un incidente. Dentro de la información, también se ve un ID específico para cada uno de los casos, el cual nos permite trabajar de manera más específica y ordenada permitiendo generar una trazabilidad de cada caso escalado dentro de la mesa de servicio. para posteriormente segmentar los datos por el asunto que conlleva cada uno de los casos teniendo en cuenta diferentes variables.

La información obtenida dentro de la base de datos pasó por un proceso de enmascaramiento con el propósito de proteger la información confidencial del fondo de pensiones y se está manejando con fines académicos.

Actividades:

- ✓ Análisis exploratorio de los datos (tipos de incidentes, frecuencia, columnas disponibles).
- ✓ Identificación de problemas de calidad de datos: datos faltantes, entradas duplicadas, variabilidad en las descripciones.
- ✓ Identificar si la base cuenta con algún sesgo.
- ✓ Evaluación inicial de la cantidad y representatividad de los casos.

- **Fase 3: Preparación de los Datos**

En esta fase construimos el conjunto de datos final que será utilizado para el modelado. Aunque en la fase anterior ya se haya realizado una comprensión inicial de los datos, en esta etapa se realiza un trabajo más específico para limpiar, transformar, seleccionar y estructurar los datos de manera adecuada. El

objetivo es garantizar que la calidad y formato de los datos permitan extraer patrones significativos y obtener resultados confiables.

En el caso de nuestro proyecto, la preparación de los datos será fundamental para lograr modelos precisos y robustos. Basándonos en antecedentes como el estudio de Qamili et al. (2018), donde se aplicaron técnicas de limpieza de texto y representación mediante bag-of-words, y en la investigación de Ramírez Devia (2021) en PQRS de Colsubsidio utilizando técnicas de vectorización y clasificación de textos, adoptaremos estrategias similares. Realizaremos una limpieza profunda de los textos, normalización, eliminación de ruido (stopwords, puntuaciones), y posteriormente aplicaremos técnicas modernas de representación semántica, como la generación de embeddings a partir de modelos de lenguaje, para mejorar la capacidad de clasificación.

Este procedimiento busca maximizar la calidad de las entradas al modelo de machine learning, asegurando un mejor desempeño en la segmentación de incidentes y priorización de casos.

Actividades:

1. Revisión y limpieza inicial de datos:

- ✓ Identificación de datos duplicados o inconsistentes.
- ✓ Eliminación de registros incompletos o irrelevantes.
- ✓ Conversión de todo el texto a minúsculas.
- ✓ Eliminación de signos de puntuación, números no relevantes y caracteres especiales.
- ✓ Eliminación de palabras vacías (stopwords) que no aportan significado.

- **Fase 4: Modelado**

En esta fase se hará uso y entrenamiento de modelos de agrupación y segmentación de incidentes basados en sus embeddings vectorizados.

Actividades detalladas:

- Clustering:
  - ✓ Aplicación de algoritmos de agrupación como **K-Means** o **DBSCAN** para descubrir grupos naturales de incidentes similares.
  - ✓ Determinación del número óptimo de clusters utilizando métricas como *Silhouette Score* o *Elbow Method*.
- Diseño de la interfaz interactiva:
  - ✓ Configuración de flujos de respuesta de la interfaz interactiva basados en los clusters generados.
  - ✓ Integración básica para simular cómo se le daría respuesta a un usuario interno.
- **Fase 5: Evaluación**

Se evalúa la calidad del modelo en función de los objetivos del negocio definidos en la fase 1.

Actividades:

- ✓ Evaluación cuantitativa de la precisión y cobertura del modelo en el conjunto de prueba.

- ✓ Validación cualitativa mediante pruebas con incidentes reales recientes.
- ✓ Revisión con el área de soporte técnico de los resultados obtenidos para ajustar criterios si es necesario.
- ✓ Documentación de limitaciones y recomendaciones para fases futuras de mejora.

- **Fase 6: Implementación**

Aunque en el proyecto académico no se realizará una implementación productiva, se diseñará un plan de despliegue teórico.

Actividades:

- ✓ Propuesta de integración del modelo en un asistente virtual o sistema de ticketing de Colfondos.
- ✓ Diseño de flujos básicos de respuesta automatizada.
- ✓ Recomendaciones para reentrenamiento periódico del modelo a medida que se generen nuevos datos.

9. RECURSOS

Recursos de personal:

Para la elaboración del proyecto se cuenta con el siguiente recurso humano:

- ❖ Diana Carolina Gómez Boada- Ingeniera Industrial
- ❖ Raúl Andrés Gamba Hastamorir - Contador Publico
- ❖ Michael David Gualteros Garcia - Ingeniero Industrial
- ❖ Luis Andrés Campos Maldonado - Tutor de Maestría en Analítica de Datos.

Recursos tecnológicos:

- Equipos de cómputo personales con capacidad adecuada para procesamiento de datos, entrenamiento de modelos de machine learning y manejo de grandes volúmenes de texto.
- Plataformas de almacenamiento en la nube, como Google Drive, para la gestión y respaldo de datasets y modelos y Mongo Atlas para almacenamiento de información.
- Acceso a bibliotecas de procesamiento de lenguaje natural y machine learning, como Hugging Face Transformers, Scikit-learn y TensorFlow.
- Se utilizará visual studio code para programar en python, miniconda como administrador de paquetes y se controlarán las versiones de los avances a través del repositorio de GitHub al que pertenecen todos los miembros del equipo de trabajo.

10. CRONOGRAMA DE ACTIVIDADES

	Semanas de ejecución de cada actividad															
ACTIVIDADES A REALIZAR	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Actividad 1 = Identificación del proyecto	x	x														
Actividad 2 = Búsqueda de los datos		x	x													
Actividad 3 = Presentación de los datos			x	x												
Actividad 4 = Definición de metodología y objetivos					x	x										
Actividad 5 = Complemento del anteproyecto						x	x									
Actividad 6 = Primer avance del anteproyecto								x	x							
Actividad 7 = Generación de repositorios y espacios de trabajo en la nube									x	x						
Actividad 8 = Correcciones y retroalimentación del documento												x				

11. PRESUPUESTO (En caso de modalidad Investigación) **Y FUENTES DE FINANCIACIÓN** (En caso de modalidad Profundización)

Este proyecto se desarrolla bajo la modalidad de profundización y no cuenta con financiación externa. Todos los recursos utilizados son aportados de manera individual por los integrantes del equipo, quienes asumen los costos asociados al desarrollo académico del proyecto.

Aunque no se contempla una financiación específica, se realiza un presupuesto estimado de los recursos necesarios para un proyecto de esta naturaleza:

Categoría	Subcategoría	Descripción detallada	Costo estimado (COP)
Recursos Computacionales	Equipos	Uso de tres computadores personales de alto rendimiento	$ 6.000.000,00
Infraestructura en la Nube	Almacenamiento en MongoDB Atlas	Servicio de almacenamiento de base de datos para casos históricos de la mesa de servicio.	$ 81.000,00
Licencias de Software	Acceso a modelos de Hugging Face	Licencias premium para el uso extendido de modelos de procesamiento de lenguaje natural en el proyecto.	$ 75.600,00
Servicios Básicos	Internet	Conectividad necesaria para acceso a plataformas, servidores y recursos en la nube.	$ 300.000,00
	Energía Eléctrica	Consumo de energía asociado al uso de computador personal durante el desarrollo del proyecto.	$ 300.000,00
Otros gastos	Fondo de contingencia	Costos imprevistos	$ 300.000,00
TOTAL PRESUPUESTO			**$ 7.056.600,00**

Observaciones:

- Por tratarse de un proyecto de carácter académico, no se incluye el valor correspondiente a mano de obra en el presupuesto actual. Sin embargo, para fines de estimaciones en posibles implementaciones profesionales futuras, se calcula que el costo de mano de obra para el desarrollo de este proyecto sería aproximadamente de $58,000,000 COP, considerando la dedicación en

horas/hombre necesarias para el análisis de datos, procesamiento de lenguaje natural, desarrollo del modelo y su implementación en un plazo estimado de 8 meses.

- Se hace uso de tecnologías gratuitas o de libre acceso para minimizar los costos.
- El acceso a los datos fue otorgado por Colfondos, bajo un acuerdo de uso exclusivo para fines académicos y posterior a un proceso de enmascaramiento de datos.
- El procesamiento de datos y entrenamiento de modelos se realiza principalmente en equipos propios de los integrantes del proyecto.
- Para almacenamiento y acceso a bases de datos, se utiliza un servicio en la nube (MongoDB Atlas).
- Se utiliza la suscripción a Hugging Face para acceder a modelos pre entrenados especializados.

12. RESULTADOS ESPERADOS

Al desarrollar este proyecto, se espera obtener los siguientes resultados:

Segmentación efectiva de casos recurrentes: A través del uso de técnicas de procesamiento de lenguaje natural y modelos de machine learning, se agruparán los casos históricos en categorías estandarizadas lo que permitirá estructurar la información y facilitar su análisis.

Generación de embeddings para comprensión semántica: Se desarrollará una representación vectorial de los casos históricos mediante embeddings, lo que permitirá capturar relaciones y similitudes entre solicitudes.

Interfaz interactiva para respuestas automatizadas: Se espera diseñar una interfaz que pueda sugerir respuestas de primer nivel a los casos recurrentes, reduciendo la carga operativa de la mesa de servicio y mejorando los tiempos de respuesta, este asistente solo responderá a casos del contexto abordado en el presente trabajo.

13. BIBLIOGRAFÍA

Debe incluir las fuentes referencias y las fuentes que, aunque no referencie, haya leído. No olvide que deben ser puestas de acuerdo con la manera que se indica en las normas APA.

- Chapman, P. (1999). The CRISP-DM User Guide. The CRISP-DM User Guide.
- Fuchs, S., Drieschner, C., & Wittges, H. (2022). Improving support ticket systems using machine learning: A literature review. Proceedings of the Annual Hawaii International Conference on System Sciences, 2022-January. https://doi.org/10.24251/hicss.2022.238
- Qamili, R., Shabani, S., & Schneider, J. (2018). An intelligent framework for issue ticketing system based on machine learning. Proceedings - IEEE International Enterprise Distributed Object Computing Workshop, EDOCW, 2018-October. https://doi.org/10.1109/EDOCW.2018.00022
- Ramírez Devia, C. L. (1129). Priorización inteligente de PQRS en Colsubsidio: Un estudio de caso en análisis de texto y aprendizaje automático. N.p.
- Shearer, C., Watson, H. J., Grecich, D. G., Moss, L., Adelman, S., Hammer, K., & Herdlein, S. A. (2000). The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing.
- Venegas Villarreal, A., Villar García, E., & Mendoza De Los Santos, A. C. (2022). Machine learning para automatizar los sistemas de tickets de soporte: Una revisión literaria. Campus, 27(34), 209–218. https://doi.org/10.24265/campus.2022.v27n34.04
- Vital Jr, A., Silva, F. N., Oliveira Jr, O. N., & Amancio, D. R. (2024). Predicting citation impact of research papers using GPT and other text embeddings. ArXiv Preprint ArXiv:2407.19942.
- Zuev, D., Kalistratov, A., & Zuev, A. (2018). Machine learning in IT service management. Procedia Computer Science, 145. https://doi.org/10.1016/j.procs.2018.11.063

14. FIRMAS

FIRMA DEL ESTUDIANTE:	**FIRMA DEL DOCENTE DIRECTOR O TUTOR**

15. DATOS DE TRÁMITE COMITÉ DE INVESTIGACIÓN DEL PROGRAMA (Espacio para diligenciar por el Comité del Programa)

No. CONSECUTIVO	
No. ACTA	
FECHA	