PLAN DE PROYECTO

Presentado por: GRUPO 1

CINDY JOHANNA ZAPATA ROMERO

HECTOR GEOVANY BELLO SANTAMARÍA

MARIO GUERRA GUALY

LEANDRO REYES JORDÁN

Docente:

CARLOS ISAAC ZAINEA MAYA

UNIDAD DE ESTUDIO:

GERENCIA DE PROYECTOS PARA CIENCIA DE DATOS

MAESTRÍA - GRUPO 1 - M1V - VIRTUAL - 2025

FACULTAD DE INGENIERÍA

BOGOTÁ, 12 DE FEBRERO DE 2025

UNIVERSIDAD EAN

Tabla de Contenido

INTRODUCCIÓN. 4
OBJETIVOS 5
Objetivo General 5
Objetivos del Proyecto (SMART): 5
ALCANCE DEL PROYECTO 6
Descripción del Alcance 6
Entregables Principales 6
Límites y Restricciones 7
Supuestos 7
ESTRUCTURA DEL PROYECTO 8
Roles y Responsabilidades 8
Cronograma 9
Tareas y Actividades (Hitos y Responsables) 10
Recursos Necesarios 11
Recursos Humanos 11
Recursos Tecnológicos 11
Recursos Financieros 11
GESTION DE RIESGOS 12
Identificación de Riesgos 12
Análisis y Plan de Mitigación 12
Monitoreo y Control de Riesgos 13
PRESUPUESTO 13

Costos 13

PLAN DE COMUNICACIÓN 14

Objetivos de la Comunicación 14

Canales de Comunicación 15

Tipos de Comunicación y Frecuencia 15

Normas de Comunicación 16

CIERRE DEL PROYECTO 16

Actividades de Cierre 16

Entregables Finales 17

Evaluación del Proyecto 17

INTRODUCCIÓN.

En la era digital, el acceso a grandes volúmenes de información ha generado la necesidad de desarrollar métodos eficientes para analizar y extraer conocimiento útil de los datos. En este contexto, el Procesamiento de Lenguaje Natural (PLN) se ha convertido en una herramienta clave para comprender y organizar información textual de manera automatizada.

El presente proyecto se enfoca en el análisis de datos y modelado predictivo mediante técnicas de Procesamiento de Lenguaje Natural (PLN). Se busca identificar patrones en un corpus de noticias y generar clusters de información relevante usando TF-IDF y Análisis de Componentes Principales (PCA).

Para la gestión y desarrollo del proyecto, se utilizará un enfoque híbrido que combina CRISP-DM (Cross Industry Standard Process for Data Mining), una metodología estructurada para proyectos de minería de datos, con Scrum, un marco ágil que facilitará la organización del equipo y la entrega iterativa de resultados. De esta manera, se garantizará un flujo de trabajo flexible y eficiente, permitiendo adaptar el análisis y modelado de datos a los requerimientos específicos del proyecto.

El resultado esperado es la generación de un modelo que facilite la clasificación y análisis de noticias, permitiendo detectar tendencias y relaciones entre los datos.

OBJETIVOS

Objetivo General

Desarrollar un modelo automatizado para la clasificación y agrupamiento de noticias mediante técnicas de Procesamiento de Lenguaje Natural (PLN), utilizando TF-IDF para la representación de datos, PCA para la reducción de dimensionalidad y K-Means como algoritmo de clustering, con el fin de optimizar la organización de la información textual y mejorar la eficiencia en el análisis de grandes volúmenes de noticias.

Objetivos del Proyecto (SMART):

- Automatizar la clasificación de noticias en distintas categorías temáticas utilizando técnicas de Procesamiento de Lenguaje Natural (PLN) para mejorar la eficiencia en la organización de información textual.

- Implementar y evaluar un modelo de clustering en una base de datos de noticias, utilizando TF-IDF para la representación de datos, PCA para reducción de dimensionalidad y K-Means como algoritmo de agrupamiento, con el objetivo de optimizar la clasificación de noticias en clusters con alta cohesión y baja dispersión.

- Desarrollar una solución escalable mediante la integración de un pipeline automatizado de preprocesamiento y clustering de noticias con K-Means, asegurando procesamiento eficiente y bajo costo computacional.

- Completar la implementación del modelo y su evaluación final, asegurando entregables parciales en cada fase del proyecto y la documentación técnica correspondiente.

ALCANCE DEL PROYECTO

Descripción del Alcance

Este proyecto se enfoca en la aplicación de técnicas de Procesamiento de Lenguaje Natural (PLN) para el análisis de un corpus de noticias, con el objetivo de identificar patrones y generar clusters de información relevante. Se implementarán métodos como TF-IDF (Term Frequency-Inverse Document Frequency) y Análisis de Componentes Principales (PCA) para extraer conocimiento útil a partir del texto.

El desarrollo del proyecto se basará en la metodología CRISP-DM para estructurar el proceso de minería de datos y en Scrum para la gestión ágil del equipo, garantizando entregas iterativas y mejoras continuas.

Entregables Principales

Los productos finales del proyecto incluirán:

- Un corpus de noticias preprocesado y estructurado.
- Un modelo de agrupamiento de noticias basado en TF-IDF y PCA.
- Visualizaciones y análisis de los clusters generados.
- Un informe detallado con hallazgos, interpretación de resultados y recomendaciones.
- Presentación final con los resultados del proyecto.

Límites y Restricciones

- Se trabajará únicamente con un conjunto de datos de noticias en idioma español.
- El corpus de noticias será obtenido de fuentes públicas disponibles en línea.
- No se realizará análisis en tiempo real; el estudio será retrospectivo sobre un conjunto de datos estático.
- Se utilizarán herramientas y entornos de desarrollo de código abierto, como Python, Pandas, Scikit-learn y NLTK.

Supuestos

- Se espera que el corpus de noticias contenga suficientes datos para realizar un análisis significativo.
- Se asumirán ciertos niveles de ruido en los datos que deberán ser tratados mediante técnicas de preprocesamiento.
- Se contará con la colaboración de todos los integrantes del grupo para el cumplimiento de los objetivos.

✓ Herramientas analíticas para ofrecer mejores recomendaciones a sus clientes.

✓ Acceso a datos precisos y predicciones confiables.

✓ Flexibilidad para personalizar soluciones según las necesidades de cada cliente.

ESTRUCTURA DEL PROYECTO

Roles y Responsabilidades

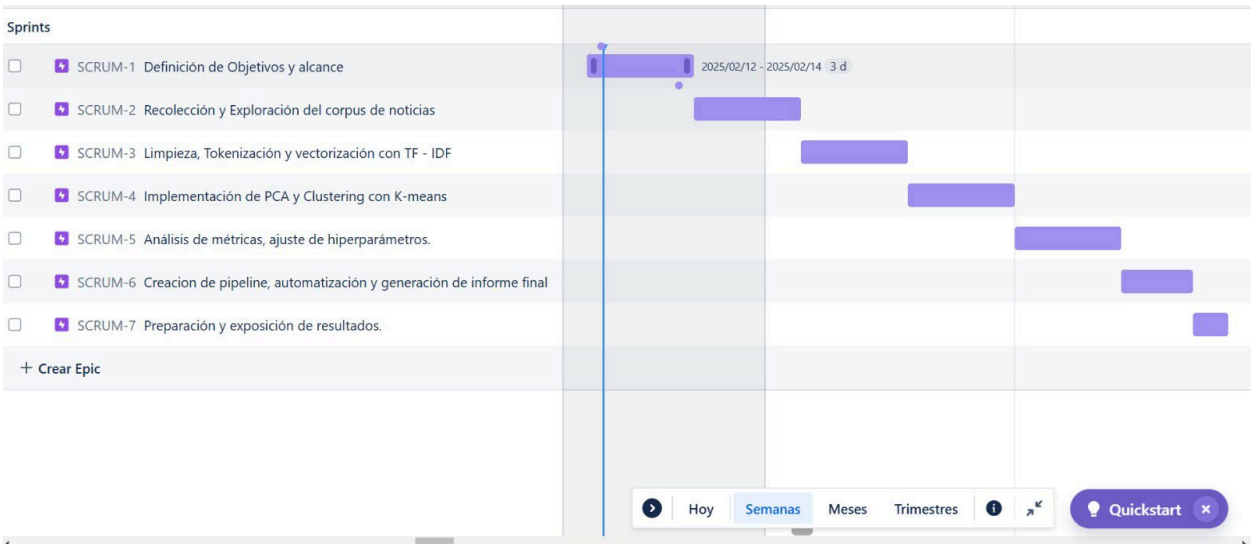
Considerando que este proyecto se gestionará bajo la metodología híbrida CRISP-DM + SCRUM, los roles se han definido de la siguiente manera conforme a las definiciones indicadas en el documento “Roles”:

Rol	Responsabilidades en Scrum	Responsabilidades en CRISP-DM
(Scrum Master)	- Facilita la metodología Scrum. - Coordina reuniones y resuelve bloqueos. - Asegura la colaboración y mejora continua.	- Facilita la organización de las fases de CRISP-DM dentro de los sprints. - Apoya en la planificación y gestión de historias de usuario. - Elimina impedimentos técnicos o de comunicación en el equipo.
(Product Owner)	- Representa las necesidades del negocio. - Prioriza las historias de usuario en el backlog. - Se comunica con stakeholders para definir requerimientos.	- Define los objetivos del proyecto basados en el negocio. - Valida si los modelos y análisis cumplen con las expectativas. - Da feedback en Sprint Reviews sobre los resultados obtenidos.
(Developer - Científico de Datos / Ingeniero de ML)	- Desarrolla modelos y experimentos. - Implementa soluciones técnicas. - Trabaja en iteraciones ágiles con entregables parciales.	- Lidera la parte de modelado y evaluación de modelos. - Prueba y ajusta diferentes algoritmos. - Implementa los modelos en producción si es necesario.
(Developer - Ingeniera de Datos / Analista de Datos)	- Gestiona los datos y pipelines. - Apoya en el análisis exploratorio. - Implementa transformaciones y optimizaciones.	- Lidera la preparación y transformación de datos. - Desarrolla scripts para limpieza de datos y extracción de información. - Apoya en la interpretación de datos y visualización.

Cronograma

El proyecto seguirá un enfoque basado en CRISP-DM y Scrum, organizando las actividades en sprints iterativos.

Fase	Actividad	Fecha de Inicio	Fecha de Fin	Responsables
1. Comprensión del Negocio	Definir objetivos y alcance del proyecto	12-Feb	14-Feb	Todo el grupo
2. Comprensión de los Datos	Recolección y exploración del corpus de noticias	15-Feb	17-Feb	Leandro, Cindy
3. Preparación de los Datos	Limpieza, tokenización y vectorización con TF-IDF	18-Feb	20-Feb	Mario, Héctor
4. Modelado	Implementación de PCA y clustering con K-Means	21-Feb	23-Feb	Mario, Héctor
5. Evaluación	Análisis de métricas, ajuste de hiperparámetros	24-Feb	26-Feb	Todo el grupo
6. Implementación y Documentación	Creación de pipeline automatizado y generación de informe final	27-Feb	28-Feb	Cindy, Héctor
7. Presentación Final	Preparación y exposición de resultados	1-Mar	1-Mar	Todo el grupo



Tareas y Actividades (Hitos y Responsables)

Hito 1: Definición del Proyecto

- Entregable: Documento de alcance y objetivos.
- Responsables: Todo el equipo.

Hito 2: Procesamiento y Preparación de Datos

- Entregable: Dataset limpio y transformado con TF-IDF.
- Responsables: Mario, Héctor.

Hito 3: Desarrollo del Modelo de Clustering

- Entregable: Implementación del modelo con PCA y K-Means.
- Responsables: Mario, Hector

Hito 4: Evaluación del Modelo

- Entregable: Reporte de métricas de desempeño y ajuste de hiperparámetros.
- Responsables: Todo el equipo.

Hito 5: Documentación y Presentación

- Entregable: Informe final y exposición del proyecto.
- Responsables: Cindy, Leandro.

Recursos Necesarios

Recursos Humanos

- **Scrum Master** : Leandro Reyes Jordán
- **Product Owner :** Hector Geovany Bello Santamaría
- **Developer - Científico de Datos / Ingeniero de ML:** Mario Guerra Gualy
- **Developer - Ingeniera de Datos / Analista de Datos :** Cindy Johanna Zapata Romero

Recursos Tecnológicos

Lenguajes y librerías: Python (Pandas, Scikit-learn, NLTK, Matplotlib).

Plataforma de desarrollo: Google Colab / Jupyter Notebook.

Herramientas de gestión: Trello, Jira, Microsoft Project, Excel.

Entorno de documentación: Google Docs, OneDrive

Recursos Financieros

Se utilizarán herramientas y datasets de acceso gratuito.

No se requiere inversión económica adicional más allá del uso de equipos personales.

GESTION DE RIESGOS

Identificación de Riesgos

ID	Riesgo	Descripción	Impacto	Probabilidad	Categoría
R1	Retraso en la recolección de datos	Dificultad para acceder a un corpus de noticias adecuado o demoras en su procesamiento.	Alto	Medio	Técnico
R2	Calidad de los datos	Posibles problemas con datos incompletos, sesgados o con ruido que afecten el modelo.	Alto	Alto	Técnico
R3	Desempeño del modelo	El clustering podría no generar grupos significativos o interpretables.	Alto	Medio	Técnico
R4	Problemas técnicos o computacionales	Limitaciones en el procesamiento de datos debido a recursos computacionales limitados.	Medio	Medio	Tecnológico
R5	Falta de coordinación en el equipo	Problemas de comunicación o desorganización en la ejecución de tareas.	Medio	Medio	Organizacional
R6	Retrasos en la documentación y presentación	Falta de tiempo para completar la documentación y preparar la exposición.	Alto	Medio	Organizacional

Análisis y Plan de Mitigación

ID	Riesgo	Estrategia de Mitigación	Responsable
R1	Retraso en la recolección de datos	Definir fuentes alternativas de datos y realizar pruebas iniciales con datasets pequeños antes de la recolección completa.	Cindy, Leandro
R2	Calidad de los datos	Aplicar técnicas de limpieza y preprocesamiento como eliminación de ruido y manejo de valores nulos.	Mario, Héctor

R3	Desempeño del modelo	Ajustar hiperparámetros y probar diferentes combinaciones de algoritmos si K-Means no funciona bien.	Leandro, Mario
R4	Problemas técnicos o computacionales	Utilizar plataformas como Google Colab para aprovechar recursos en la nube.	Todo el grupo
R5	Falta de coordinación en el equipo	Mantener reuniones semanales y usar herramientas como Trello para el seguimiento de tareas.	Scrum Master (Leandro)
R6	Retrasos en la documentación y presentación	Asignar tareas específicas para la documentación desde el inicio y realizar revisiones continuas.	Héctor, Cindy

Monitoreo y Control de Riesgos

Para garantizar la correcta gestión de riesgos, se implementarán las siguientes acciones:

- ✓ Reuniones semanales para evaluar el avance y posibles bloqueos.
- ✓ Seguimiento en Trello para asegurar que las tareas se cumplen en los tiempos establecidos.
- ✓ Revisión continua del modelo para realizar ajustes en caso de bajo desempeño.
- ✓ Plan de contingencia en caso de problemas técnicos, como cambiar de entorno de ejecución.

PRESUPUESTO

Costos

Categoría	Descripción	Costo Estimado (COP)	Observaciones
Recursos Humanos	Trabajo del equipo (no remunerado, académico)	$0	Trabajo realizado por los integrantes del grupo.

Software y Herramientas	Python, Pandas, Scikit-learn, NLTK, Jupyter Notebook, Google Colab	$0	Todo el software es de código abierto.
Plataformas de Computación	Google Colab Pro (opcional)	$50000 - $100000	Solo si se requiere mayor capacidad computacional.
Almacenamiento de Datos	Google Drive / GitHub	$0	Uso de cuentas personales gratuitas.
Internet y Electricidad	Conexión para trabajo remoto	$40000-$60000	Estimado de uso adicional en casa.
Materiales y Documentación	Impresión de informes (opcional)	$20000-$40000	Solo si se requiere versión impresa.
Presentación Final	Diseño de diapositivas, conexión a internet para exposición	$0	Se utilizarán herramientas gratuitas como Google Slides o PowerPoint.

El presupuesto aproximado es de $110000 - $200000 COP

PLAN DE COMUNICACIÓN

Dado que el equipo trabaja de manera remota, se implementarán herramientas digitales para la planificación, seguimiento y entrega de resultados.

Objetivos de la Comunicación

- ✓ Asegurar la coordinación efectiva entre los miembros del equipo.
- ✓ Garantizar el seguimiento de tareas y cumplimiento de plazos.
- ✓ Facilitar la documentación y el acceso a la información del proyecto.
- ✓ Mantener una comunicación clara con el profesor y posibles stakeholders.

Canales de Comunicación

Canal	Uso Principal	Frecuencia	Responsable
WhatsApp	Comunicación rápida, dudas y recordatorios	Diario	Todo el equipo
Google Meet / Zoom/Teams	Reuniones semanales de avance y revisión	2 veces por semana	Scrum Master (Leandro)
Trello / Jira	Seguimiento de tareas y asignaciones	Continuo	Todo el equipo
Google Drive / OneDrive	Almacenamiento de documentos, datasets y código	Continuo	Todo el equipo
Google Docs	Edición colaborativa de documentación	Continuo	Héctor, Cindy
GitHub	Control de versiones del código y experimentos	Continuo	Mario, Leandro
Correo Electrónico	Envío de entregables y comunicación con el profesor	Según necesidad	Product Owner (Héctor)

Tipos de Comunicación y Frecuencia

Tipo de Comunicación	Propósito	Frecuencia	Participantes
Reuniones de Seguimiento	Evaluar avances, resolver bloqueos, asignar tareas	2 veces por semana	Todo el equipo
Actualizaciones en Trello	Marcar tareas completadas y asignar nuevas	Continuo	Todo el equipo
Sincronización rápida (WhatsApp/Telegram)	Solucionar dudas puntuales y coordinar actividades	Diario	Todo el equipo
Revisión de Código en GitHub	Asegurar calidad del código y evitar conflictos	Según necesidad	Mario, Leandro
Entrega de Documentos en Google Drive	Almacenar y compartir avances de informes y presentación	Continuo	Héctor, Cindy
Ensayo de Presentación Final	Revisar exposición y mejorar claridad	29 de febrero	Todo el equipo

Normas de Comunicación

- Mantener un lenguaje claro y preciso en todos los canales.
- Responder en un máximo de 24 horas a mensajes y asignaciones.
- Respetar los horarios de reunión y llegar preparados con avances.
- Usar Trello y GitHub para documentar el progreso y evitar pérdida de información.
- Guardar todas las versiones de documentos en Google Drive/OneDrive para fácil acceso.

CIERRE DEL PROYECTO

Actividades de Cierre

Actividad	Descripción	Responsable	Fecha Límite
Validación de resultados	Evaluar si los objetivos del proyecto se cumplieron de acuerdo con los criterios establecidos.	Todo el equipo	27-Feb
Revisión de documentación	Completar y revisar el informe final del proyecto.	Héctor, Cindy	28-Feb
Entrega de Código	Subir versión final del código en GitHub con comentarios y documentación.	Mario, Leandro	28-Feb
Ensayo de Presentación	Realizar al menos un ensayo de la exposición final.	Todo el equipo	29-Feb
Presentación Final	Exposición ante el profesor y stakeholders.	Todo el equipo	1-Mar

Retroalimentación y Reflexión	Discusión sobre aprendizajes, mejoras y desafíos enfrentados.	Todo el equipo	1-Mar
Cierre Administrativo	Confirmar la entrega de todos los entregables y archivar documentos en Drive.	Scrum Master (Leandro)	1-Mar

Entregables Finales

Entregable	Formato	Ubicación
Informe Final	Documento PDF / Word	OneDrive
Código del Proyecto	Repositorio GitHub	GitHub / OneDrive
Presentación	Diapositivas (Google Slides o PowerPoint)	OneDrive
Dataset y Documentación	CSV / JSON y PDF explicativo	OneDrive
Resumen Ejecutivo	1-2 páginas con hallazgos clave	OneDrive

Evaluación del Proyecto

El equipo realizará un análisis de los siguientes aspectos:

- Cumplimiento de Objetivos: ¿Se lograron los objetivos SMART definidos?
- Calidad del Modelo: ¿El clustering fue efectivo? ¿Se pueden interpretar los resultados?
- Cumplimiento de Plazos: ¿Se respetaron las fechas del cronograma?
- Gestión del Equipo: ¿La comunicación y la coordinación fueron adecuadas?
- Lecciones Aprendidas: ¿Qué se puede mejorar para futuros proyectos?