Nicolás Rincón - 202021963
Santiago Jaimes – 201912921
Nicolás Casas – 202212190

Proyecto 1 Etapa 1

1. Introducción

El Fondo de Poblaciones de las Naciones Unidas (UNFPA) en colaboración con entidades públicas y la participación ciudadana desea evaluar la información dada por los ciudadanos con los Objetivos de Desarrollo Sostenible (ODS). Para este propósito han recopilado la información textual con las opiniones de los ciudadanos y la categorización que se les dio.Dado el objetivo de los actores, así como la información proporcionada se plantea construir un modelo de aprendizaje automático que sea capaz de reconocer las opiniones de los ciudadanos y categorizarlas en uno de los Objetivos de Desarrollo Sostenible (3,4,5).

Elemento	Descripción
Oportunidad/problema Negocio	El UNFPA busca una solución automatizada para relacionar de forma precisa y eficiente las opiniones de los ciudadanos con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Este proceso, actualmente realizado manualmente, es costoso en términos de tiempo y recursos humanos.
Objetivos y criterios de éxito desde el punto de vista del negocio	El éxito se medirá por la precisión, recall, F1-score y capacidad de adaptación del modelo en la clasificación de opiniones ciudadanas relacionadas con los ODS. El objetivo final es crear un modelo replicable y automatizable que permita ahorrar recursos y mejorar la toma de decisiones basada en datos.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El Fondo de Poblaciones de las Naciones Unidas (UNFPA) es la entidad principal beneficiada, junto con otras instituciones públicas de Colombia que colaboran en la implementación de políticas relacionadas con los ODS. Los analistas de datos y los responsables de políticas públicas se beneficiarán del análisis automatizado.
Impacto que puede tener en Colombia este proyecto.	El proyecto tiene el potencial de mejorar la toma de decisiones y la implementación de políticas públicas en Colombia relacionadas con la salud (ODS 3), la educación (ODS 4) y la igualdad de género (ODS 5). La automatización de la relación entre opiniones ciudadanas y ODS puede generar una respuesta más rápida y eficaz a los problemas sociales.
Enfoque analítico. Descripción de la categoría de análisis (descriptivo, predictivo, etc.) , tipo y tarea de aprendizaje e incluya las técnicas y algoritmos que propone utilizar.	El enfoque principal es **predictivo**, ya que el objetivo es predecir qué ODS está relacionado con un texto dado. El tipo de tarea es **clasificación multiclase**, y se han empleado tres algoritmos principales: **Naive Bayes**, **Regresión Logística Multinomial**, y **Random Forest**. Estos algoritmos fueron seleccionados por su capacidad de manejar tareas de clasificación de texto y se evaluaron usando métricas como accuracy, precision, recall y F1-score para determinar el rendimiento del modelo.

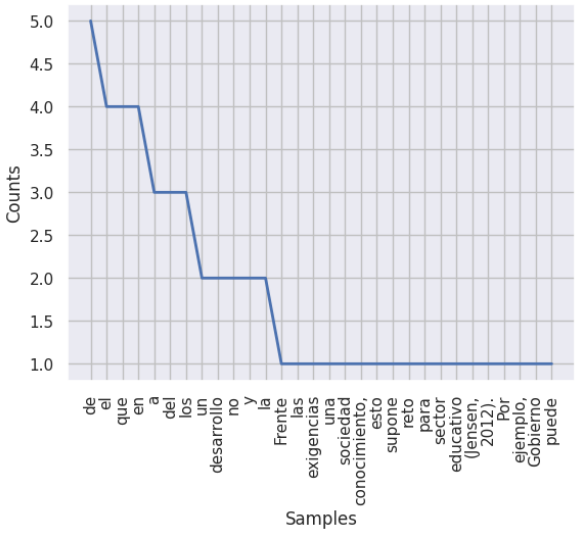
2. Entendimiento y preparación de los datos

2.1 Entendimiento de los datos

Al observar la información proporcionada se aprecia que hay 4050 registros cada uno con su columna de Texto español que contiene la opinión de los ciudadanos y otra columna con la clasificación asignada del ODS (3,4,5). Las opiniones son de tipo no estructurado puesto son texto y no cuentan con un valor numérico por lo que no se pueden ordenar nominal u ordinalmente. Por otro lado, se tiene que cada opinión tiene en promedio 700 caracteres y 41 palabras únicas. Además, la clasificación más popular fue 5 con 1451 opiniones y que el idioma identificado en las opiniones fue 99.6% español, 0.1% inglés y 0.1% francés.

Idioma	count
es	0.996789
en	0.001729
fr	0.001482

Para visualizar cuáles son las palabras más frecuentes en todo el corpus y de ahí en adelante ejemplos de palabras que se encuentran con menor frecuencias se usó una librería de procesamiento de lenguaje natural (nltk.probability).



2.2 Verificación de Calidad

La verificación de calidad de los datos reveló que el conjunto es completo, sin registros vacíos, y que todos los valores son únicos, sin duplicados. No se encontraron valores atípicos, aunque se detectaron algunos caracteres mal codificados, como "Ã¡", que se corregirán durante la limpieza de los datos. Además, los valores son consistentes, con las opiniones en formato `Object` y las categorías en formato `Int`, lo que asegura una estructura adecuada para el análisis posterior.

2.3 Limpieza de los datos

Este proceso de limpieza se compone de varias etapas:

- Corrección y eliminación de ruido.
- Elminación de stopwords
- Eliminación de registros que se les haya detectado un idioma distino al español.

La **corrección de los datos** se centró en solucionar errores de codificación en UTF-8, específicamente en textos con caracteres especiales en español. Se creó un mapa de caracteres mal codificados y una función (`correct_encoding_errors`) para reemplazarlos por su representación adecuada, como cambiar "Ã¡" por "á". Además, se implementó la función `remove_non_ascii` para eliminar caracteres no pertenecientes al conjunto ASCII, asegurando que solo se utilicen caracteres estándar en la codificación final.

En cuanto a la **eliminación de ruido**, se ajustaron los textos mediante tres pasos: conversión a minúsculas, eliminación de signos de puntuación y transformación de números a palabras, con un ajuste adicional para convertir correctamente porcentajes. También se **eliminaron las **stopwords**** (artículos, preposiciones y conjunciones) al considerarlas irrelevantes para el análisis.

Finalmente, se **retiraron los registros en inglés y francés**, ya que más del 99% del corpus estaba en español. La eliminación de estos registros, que representaban menos del 1% del total, no afectó significativamente las predicciones.

2.4 Preparación de los datos

Tokenización

Primero se tokenizarón las palabras, es decir, se separan individualmente sus palabras. Cada palabra se convierte en un token al momento de realizar el parsing de cada uno de los registros del DataFrame.

Normalización

Luego de haber tokenizado las palabras se pasa a la normalización, donde se realiza un Stemming y una Lematización. Estos dos procesos se encargan de eliminar los prefijos y sufijos de las palabras, así como de generar la raíz léxica de las palabras.
De esta manera, reducen la cantidad de palabras del corpus al quitar todas las variantes posibles.

Vectorización

La vectorización se realizó sobre las palabras procesadas tras el **stemming** y la **lematización,** uniendo los términos en una representación final. Se utilizó el método **TF-IDF** para equilibrar la frecuencia y relevancia de las palabras en el corpus, reduciendo la importancia de las palabras más comunes y dando mayor peso a las más específicas. Como hiperparámetro, se seleccionó un máximo de **10,000 características** para evitar problemas de sobreajuste o subajuste y optimizar el tiempo de procesamiento. Esto asegura una representación eficiente sin sacrificar el rendimiento del modelo.

3. Modelo y evaluación

En esta sección se presenta la aplicación de tres algoritmos diferentes para la tarea de **clasificación multiclase,** con el fin de predecir automáticamente la categoría de los textos relacionados con los Objetivos de Desarrollo Sostenible 3,4 y 5. Se

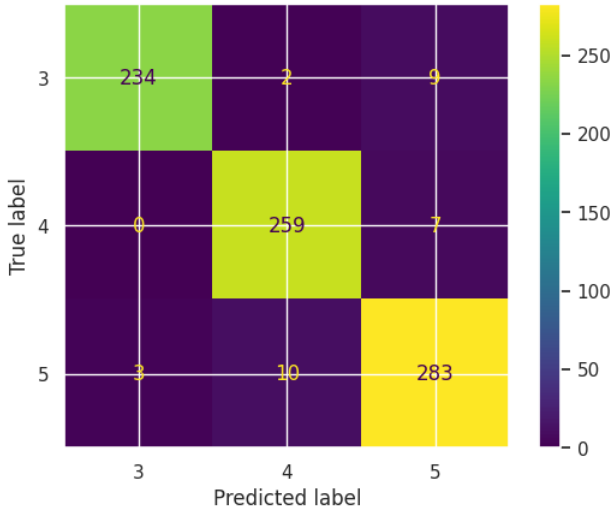
evaluaron y compararon los modelos utilizando métricas estándar de rendimiento, como la precisión, el recall, el F1-score, para determinar el mejor enfoque.

- **Algoritmo 1: Naive Bayes Multinomial (responsable - Nicolás Casas)**

El **Naive Bayes Multinomial** es un modelo probabilístico basado en el teorema de Bayes, que es eficaz para tareas de clasificación de texto, particularmente cuando los datos están representando por características discretas como la frecuencia de palabras, lo cual se ajusta bien al uso de la técnica de vectorización TF-IDF. En este modelo, se asume que las características (en este caso, las palabras) son independientes entre sí, lo que simplifica el cálculo de probabilidades y lo hace adecuado para conjuntos de datos grandes.

Entrenamiento del modelo
El modelo fue entrenado con los textos procesados y vectorizados mediante TF-IDF, lo que permitió transformar los textos en una representación numérica basada en la importancia de las palabras en el corpus. El modelo se probó con un conjunto de prueba divido aleatoriamente con una proporción de 80/20 para entrenamiento y prueba.

Resultados de las métricas de evaluación




```
>> Las métricas del modelo generado son las siguientes:

Accuracy: 0.9615861214374225
Precision: 0.9619339176063125
Recall: 0.9615861214374225
F1 Score: 0.9616423556967129
```

- **Algoritmo 2: Regresión Logística Multinomial (responsable - Santiago Jaimes)**

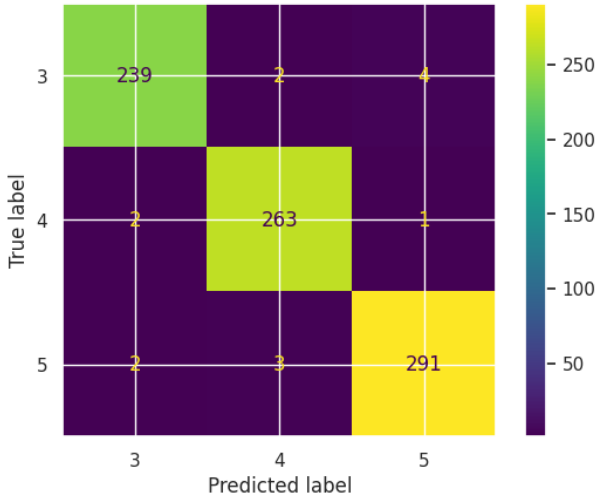
La **Regresión Logística Multinomial** es una extensión de la regresión logística tradicional que permite realizar tareas de **clasificación multiclase**. Este modelo se basa en la función **softmax**, la cual calcula las probabilidades de cada clase y selecciona aquella con la mayor probabilidad para realizar la predicción. Es un modelo muy adecuado para tareas de clasificación de texto, ya que puede capturar relaciones complejas entre las características, especialmente cuando los datos se han procesado usando técnicas como **TF-IDF**.

Entrenamiento del modelo

El modelo se entrenó con datos previamente procesados y vectorizados usando **TF-IDF** para convertir los textos en vectores numéricos. La tarea de clasificación se basa en predecir cuál de las tres clases de ODS (3, 4, 5) es la más adecuada para cada texto dado. El conjunto de datos se dividió en una proporción del 80% para entrenamiento y el 20% restante para prueba, asegurando que las predicciones fueran evaluadas sobre un conjunto no visto.

Se utilizó el solver **"lbfgs"**, que es una técnica de optimización utilizada comúnmente en regresión logística para manejar problemas multiclase. Esto garantiza que el modelo converja eficientemente hacia una solución óptima.

Resultados de las métricas de evaluación




```
>> Las métricas del modelo generado son las siguientes:

Accuracy: 0.9826517967781908
Precision: 0.9826572386891403
Recall: 0.9826517967781908
F1 Score: 0.9826449683644728
```

- **Algoritmo 3: Random Forest (responsable - Nicolás Rincón)**

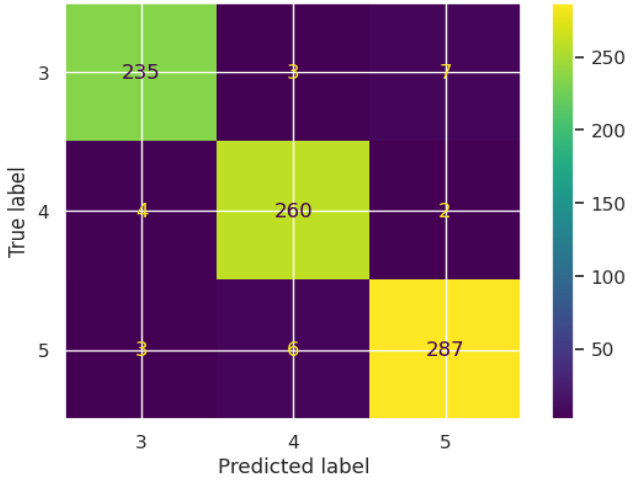
El algoritmo de **Random Forest** es un método de aprendizaje supervisado de ensamble que crea un conjunto de múltiples árboles de decisión para mejorar el rendimiento de un modelo y reducir su riesgo de sobreajuste. En lugar de entrenar un solo árbol, entrena varios árboles de decisión independientes utilizando diferentes subconjuntos de datos mediante el muestreo con reemplazo (técnica llamada bootstrap). Cada árbol toma decisiones basadas en la división de los datos según un criterio como la impureza de Gini o la ganancia de información (entropy).

Para una predicción final, Random Forest utiliza la votación mayoritaria de las predicciones individuales de cada árbol. Al promediar o tomar la moda de los resultados de todos los árboles, el modelo tiende a ser más robusto y menos propenso a sobreajustarse que un único árbol de decisión.

Entrenamiento del modelo

El modelo de Random Forest fue entrenado utilizando el conjunto de datos vectorizado mediante **TF-IDF**, lo que permitió obtener representaciones numéricas de los textos basadas en la frecuencia de los términos. El conjunto de datos se dividió en un 80% para entrenamiento y un 20% para prueba, asegurando que el modelo pudiera evaluarse de manera confiable.

Resultados de las métricas de evaluación

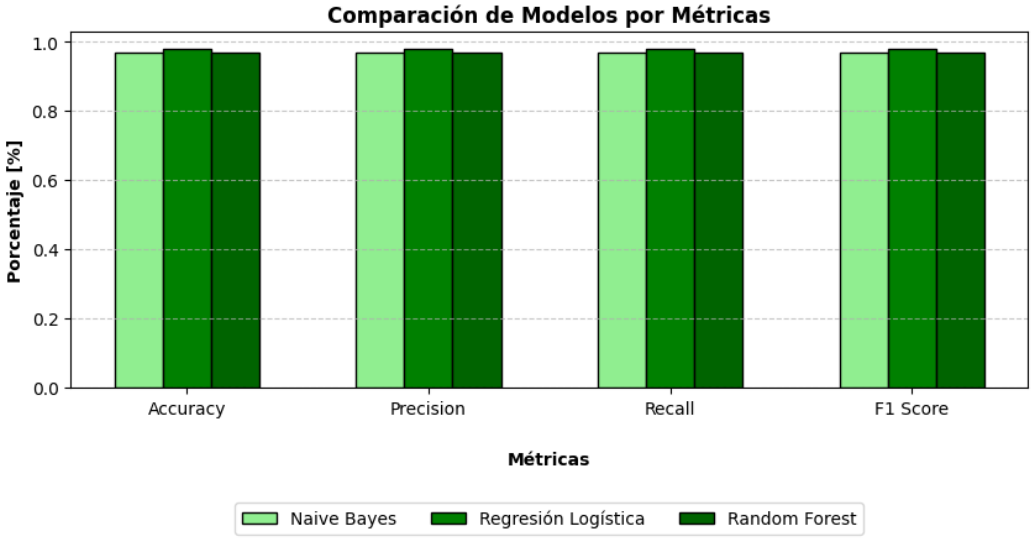



```
>> Las métricas del modelo generado son las siguientes:

Accuracy: 0.9690210656753407
Precision: 0.9690379118806788
Recall: 0.9690210656753407
F1 Score: 0.9690082957601766
```

4. Resultados

a)



Seleccionamos el modelo de Regresión Logística Multinomial, que obtuvo las métricas más altas en todas las áreas, destacándose con un F1 Score de 0.9826 (el cual fue nuestro criterio de decisión). Al priorizar el F1 Score (0.9826), nos

aseguramos de seleccionar un modelo que balanceara bien tanto la Precision como el Recall. Esto es fundamental cuando buscamos un buen rendimiento tanto en la correcta identificación de los textos de cada ODS como en evitar perder información relevante.

Dado que los datos de entrenamiento estaban balanceados, la métrica de Accuracy (Exactitud) sigue siendo un buen indicador de rendimiento general. El Accuracy del modelo de Regresión Logística Multinomial fue de 0.9826, lo que significa que el 98.26% de las predicciones realizadas fueron correctas en términos de clasificar los textos en las categorías correctas de ODS 3, 4 y 5. Esta métrica es útil para darnos una visión global de qué tan bien el modelo clasifica, pero no toma en cuenta cómo se comporta en cada clase individual.

Además, la Precisión (0.9826) nos indica el porcentaje de predicciones correctas entre todas las instancias que el modelo etiquetó como pertenecientes a una categoría en particular. En este caso, casi el 98.26% de los textos que fueron clasificados como ODS 3, 4 o 5 eran efectivamente de esa categoría. Esto es importante en escenarios donde los falsos positivos son costosos, pero en nuestro caso queríamos también asegurar que no dejáramos de identificar textos relevantes.

Por esta razón, el Recall (0.9826) también fue clave, ya que mide la capacidad del modelo para identificar correctamente las instancias positivas dentro de cada categoría. En nuestro caso, esto significa que el modelo fue capaz de recuperar el 98.26% de los textos que efectivamente pertenecían a las categorías ODS 3, 4 o 5, evitando que se pierdan textos relevantes.

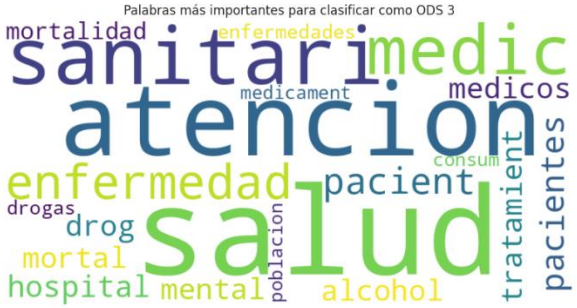
b)Al analizar los resultados obtenidos para el proyecto con el Fondo de Poblaciones de las Naciones Unidas (UNFPA), cuyo objetivo es relacionar automáticamente opiniones ciudadanas con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5, se observa un excelente desempeño del modelo implementado. En particular, el modelo de clasificación, basado en regresión logística, muestra un F1-score del 98.26%, lo que indica un equilibrio casi perfecto entre precisión y exhaustividad en la identificación correcta de las opiniones asociadas a estos ODS.

El análisis tanto cuantitativo como cualitativo respalda la efectividad del modelo. Las palabras que el modelo prioriza como relevantes e irrelevantes para la clasificación de las opiniones hacia los ODS 3 (Salud y Bienestar), ODS 4 (Educación de Calidad) y ODS 5 (Igualdad de Género) presentan una lógica consistente con los temas que cubren estos objetivos. Los coeficientes de regresión asignados a estas palabras refuerzan la lógica detrás del comportamiento del modelo, lo que asegura que las clasificaciones son tanto precisas como comprensibles.

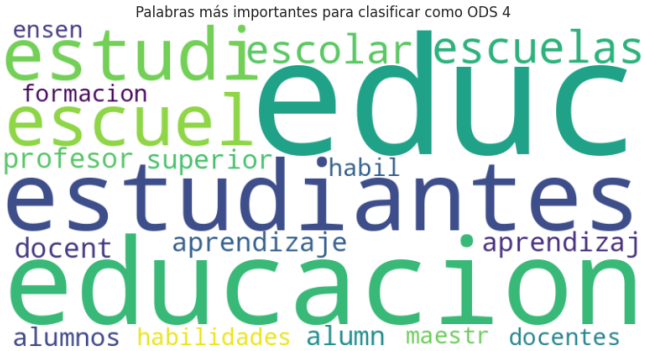
Si bien el desempeño actual del modelo es excepcional, con un error promedio muy bajo, se puede sugerir continuar monitoreando y ajustando el modelo a medida que se recolecten más datos o que las opiniones ciudadanas abarquen una mayor diversidad de temas. No obstante, consideramos que la versión actual ya es una herramienta muy efectiva que permitirá al UNFPA automatizar el análisis de opiniones y obtener valiosos insights para su labor en el cumplimiento de los ODS.

<u>Estrategias para la empresa:</u>

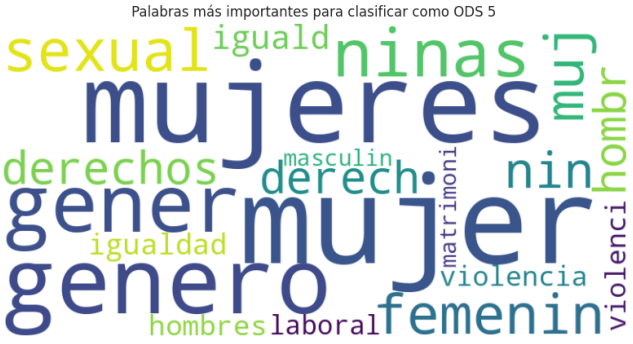
A partir de la identificación de las palabras más importantes para la categorización de las opiniones en cada ODS, se proponen las siguientes estrategias a implementar para favorecer las opiniones positivas de las personas respecto al cumplimiento de los objetivos.



- Enfoque en mejorar la atención hospitalaria
- Enfoque en asegurar medidas sanitarias
- Enfoque en tratamiento de enfermedades
- Poner al paciente en primer lugar



- Enfoque en el mejoramiento de la educación
- Desarrollo de formación de profesores
- Mejoramiento de la calidad de las escuelas
- Asegurar condiciones de aprendizaje para los estudiantes



- Protección de los derechos de las mujeres y niñas
- Enfoque en la igualdad laboral de mujeres y hombres
- Políticas en contra de la violencia hacia la mujer
- Reforzar las medidas en contra de la discriminación sexual

c) Los datos de prueba etiquetados se encuentran en los siguientes enlaces: https://github.com/nrincon2302/BI-Proyecto1-G27/blob/main/source/predicciones.csv

https://github.com/nrincon2302/BI-Proyecto1-G27/blob/main/source/predicciones.xlsx

5. Mapa de actores relacionado con el producto de datos creado.

Rol dentro del Ministerio	Tipo de actor	Beneficio	Riesgo
Ciudadanos	Beneficiado	Sus opiniones serán analizadas automáticamente para identificar problemáticas relevantes relacionadas con los ODS 3, 4 y 5.	Si el modelo no tiene un buen desempeño, las opiniones pueden clasificarse incorrectamente, afectando la priorización de problemas.
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Usuario-cliente	Obtiene un mecanismo eficiente para analizar grandes volúmenes de datos de opiniones ciudadanas, lo que agiliza la toma de decisiones para los ODS.	Si el modelo no funciona como se espera, las decisiones basadas en datos podrían no ser precisas.
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Financiador	Proporciona los recursos financieros necesarios para el desarrollo y ejecución del proyecto, asegurando la continuidad del mismo.	Si el proyecto no tiene éxito, los recursos invertidos podrían no generar el impacto deseado, perdiéndose oportunidades de otras inversiones.
Entidades públicas colaboradoras	Proveedor	Aseguran que el modelo cumpla con los estándares de calidad y privacidad de datos, lo que mejora la confianza en el análisis de opiniones.	El manejo incorrecto de los datos puede comprometer la privacidad y generar desconfianza en la ciudadanía.

6. Trabajo en Equipo

Actividad	Santiago Jaimes	Nicolas Rincón	Nicolas Casas
Planeamiento de reuniones	X		
Gestión del repositorio		X	
Documentación del negocio			X
Documentación del cuaderno		X	
Investigación inicial de técnicas de análisis de vectorización	X		
Ejecución de limpieza de datos	X		
Ejecución de entendimiento de datos			X
Modelado del algoritmo NB multinomial			X
Modelado del algoritmo Regresión logística multinomial	X		

Modelado del algoritmo Random Forest		X	
Análisis de resultados		X	
Documentación de actores			X
Aporte total	33.34%	33.33%	33.33%

Rol desempeñado	Santiago Jaimes	Nicolas Rincón	Nicolas Casas
Líder de proyecto	X		
Líder de negocio			X
Líder de datos		X	
Líder de analítica	X		

Reunión	Fecha	Descripción
Reunión de lanzamiento y planeación	01/09/2024	Para el lanzamiento del proyecto de inteligencia de negocios se establecen objetivos, alcance, roles, requisitos del cliente y plan de acción. Se asignan roles, se delinean hitos clave, se identifican riesgos y se establecen estrategias de mitigación.
Reunión de ideación	02/09/2024	Se generan ideas creativas y soluciones innovadoras para el análisis de los textos. Se fomenta la participación abierta y la colaboración entre los asistentes. Se establecen los siguientes pasos para desarrollar y evaluar las ideas.
Reunión de seguimiento 1	04/09/2024	Se hace el seguimiento del progreso del proyecto, revisar el estado actual y discutir los hitos alcanzados desde el lanzamiento. Se identifican posibles desafíos y se establecen medidas correctivas si es necesario. Se asignan responsabilidades para las próximas etapas y se fija la fecha para la próxima reunión de seguimiento.
Reunión de seguimiento 2	06/09/2024	Se continua el seguimiento del progreso del proyecto, el avance desde la última reunión y evaluamos el cumplimiento de los hitos establecidos. Se discuten los problemas emergentes y se implementan medidas correctivas.
Reunión de finalización	08/09/2024	Se evalúa la finalización exitosa del proyecto y revisamos los logros alcanzados en relación con los objetivos iniciales. Se analizan lecciones aprendidas y se destacan los puntos fuertes y áreas de mejora. Se discuten los próximos pasos, incluyendo la entrega del resultado.

7. Enlace del Video

El enlace del video en el Padlet es el siguiente: https://youtu.be/QtHfDEEWhAM

8. Referencias

[1]Awan, A.A. & Naviani, A. (2023). Naive Bayes Classification Tutorial using Scikit-learn. DataCamp. Disponible en: https://www.datacamp.com/tutorial/naive-bayes-scikit-learn

[2]Geeks for Geeks. (2024). Logistic Regression in Machine Learning. Disponible en: https://www.geeksforgeeks.org/understanding-logistic-regression/

[3]Shafi, A. (2023). Random Forest Classification with Scikit-Learn. DataCamp. Disponible en: https://www.datacamp.com/tutorial/random-forests-classifier-python