INTELIGENCIA DE NEGOCIOS

ISIS - 3301

PROYECTO 1 – ETAPA 1

POR:

DAVID ALEJANDRO FUQUEN

CÓDIGO: 202021113

VALERIA CARO RAMIREZ

CÓDIGO: 202111040

MARIANA RUIZ GIRALDO

CÓDIGO: 202011140

UNIVERSIDAD DE LOS ANDES

FACULTAD DE INGENIERÍA

SEPTIEMBRE DE 2024

Tabla de Contenido

1. Entendimiento del negocio y enfoque analítico

Tabla 1. Análisis del negocio

Oportunidad/problema Negocio	El problema que el Fondo de Poblaciones de las Naciones Unidas (UNFPA) quiere resolver es la identificación de problemas y evaluación de soluciones relacionadas con los objetivos de Desarrollo Sostenible (ODS) basadas en las opiniones y necesidades que expresan los ciudadanos. Esto requiere el análisis de grandes cantidades de información textual, acompañado del concepto de expertos. Dado este problema se presenta una oportunidad para implementar un modelo analítico de clasificación de opiniones ciudadanas respecto a los ODS 3 (Salud y bienestar), 4 (Educación de Calidad) y 5 (Igualdad de género) (United Nations, 2023). El propósito es permitir un análisis más eficiente y económico de las opiniones ciudadanas para encontrar insights que permitan la toma de

	decisiones y el diseño de políticas públicas.
Objetivos y criterios de éxito desde el punto de vista del negocio	Se definen los siguientes objetivos para el proyecto: - Implementar un modelo de aprendizaje que sea capaz de clasificar automáticamente las opiniones ciudadanas en relación con los ODS 3, 4 y 5 - Reducir la dependencia del análisis manual de expertos y optimizar la asignación de recursos para el análisis de datos textuales. Se espera que la solución planteada cumpla con los siguientes criterios: - El modelo debe alcanzar una precisión mínima del 80% en la clasificación por ODS de las opiniones ciudadanas - Reducción del tiempo necesario para clasificar la información textual.
Organización y rol de ella que se beneficia con la oportunidad definida	La organización que se ve beneficiada por este proyecto es El Fondo de Poblaciones de las Naciones Unidas (UNFPA) y entidades públicas de Colombia, tales como los ministerios de salud y educación, que trabajan en el diseño de políticas para cumplir con los ODS. Los roles específicos que se verán beneficiados son: - Analistas de Datos y Científicos de Datos quienes pueden reducir su carga de trabajo y pueden enfocarse en tareas más estratégicas y no tan manuales - Políticos y Planificadores de políticas públicas quienes van a obtener insights para facilitar la toma de decisiones de forma más rápida, lo que les permitirá definir políticas efectivas y acordes al momento que vive la ciudadanía

	- Expertos en desarrollo sostenible quienes verán su trabajo facilitado, pues podrán identificar las áreas prioritarias de intervención.
Impacto que puede tener en Colombia este proyecto	Este proyecto puede tener un impacto significativo en Colombia, ya que permitirá el mejoramiento de las políticas públicas encaminadas al cumplimento de los ODS. La optimización de recursos y el alineamiento de las estrategias gubernamentales con las prioridades de los ciudadanos representan un impacto importante para Colombia, que puede resultar en una mejor calidad de vida, mayor equidad y participación ciudadana en la toma de decisiones.
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 análisis por realizar es principalmente de carácter **predictivo,** ya que el modelo debe predecir a cuál de los ODS 3, 4 o 5 corresponde una opinión ciudadana basada en su contenido textual. El tipo de aprendizaje será **supervisado** donde se utilizará el conjunto de datos etiquetados previamente para entrenar el modelo de clasificación. Como ya ha sido mencionado, este modelo debe resolver una tarea de **clasificación,** donde las opiniones sean categorizadas en una de las tres categorías de ODS. Los algoritmos propuestos para solucionar este requerimiento son: El Procesamiento del lenguaje natural para convertir los datos a un formato idóneo para la aplicación de algoritmos de clasificación. Se utilizarán 3 algoritmos de clasificación: Regresión logística, K-vecinos más cercanos y Naive Bayes Multimodal. Además, se utilizarán métricas precisión, recall y f1-score para evaluar el modelo.

2. Entendimiento y preparación de los datos

El entendimiento de los datos preliminar se realiza con la librería `ydata-profiling` cuyo reporte nos permitió evidenciar que los datos tienen una dimensión total de 4049 registros. Además, se pudo identificar la existencia de 2 columnas: “sdg” (categórica) y “Textos_espanol” (textual)
Para el caso de este proyecto, la variable objetivo es “sdg” que corresponde a el número de Objetivo de Desarrollo Sostenible al que pertenece cada registro en la columna de “Textos_espanol”. Los datos disponibles hacen referencia a los ODS de Salud y bienestar (3), Educación de Calidad (4) e Igualdad de género (5).
En cuanto a la **completitud**, no se encontraron valores ausentes en ningún registro, por lo que no fue necesario realizar transformaciones adicionales. Respecto a la **unicidad**, se confirmó que no existían duplicados, asegurando que cada registro es único.

En términos de **validez**, se verificó que todos los valores de la columna "Textos_espanol" son datos de tipo texto y que los valores de "sdg" corresponden exclusivamente a los ODS {3, 4, 5}. Por último, se validó la **consistencia estructural**, ya que los datos de texto están almacenados como cadenas de caracteres y los datos categóricos como datos numéricos. También se pudo evidenciar **consistencia semántica** porque los datos que representan el mismo concepto tienen la misma representación.

Tras completar la exploración de los datos con `ydata-profiling` y visualizar las palabras más frecuentes utilizando un WordCloud, se determinó que la única transformación necesaria es la conversión de los textos a vectores numéricos. Esto es crucial para aplicar los algoritmos de clasificación: Regresión Logística, K-Vecinos más Cercanos y Naive Bayes, implementados con scikit-learn. Para la transformación textual, se aplicaron los siguientes pasos: Conversión de los textos a minúsculas, Eliminación de caracteres especiales, Corrección de codificación (encoding), Eliminación de *stopwords*, Tokenización, Aplicación de *stemming* para obtener las raíces de las palabras. Por último, se vectoriza por medio del método TF-IDF y se crea un pipeline de preprocesamiento de textos. Es importante notar que este proceso fue realizado por los 3 integrantes en conjunto.

3. Modelado y evaluación

Aplicación de mínimo tres algoritmos diferentes para la tarea de aprendizaje automático seleccionada. En esta parte deben describir los algoritmos utilizados para la construcción de los modelos y presentar los resultados de las métricas de evaluación para justificar la selección del modelo. Además, deben incluir los nombres de los estudiantes que trabajaron en cada modelo analítico. Al igual que en los laboratorios, en la nota del proyecto hay un porcentaje específico para el aporte de cada miembro del grupo.

Para la tarea de aprendizaje automático se utilizaron tres algoritmos de clasificación: Logistic Regression, K-Nearest Neighbors y Multinomial Naive Bayes. Cada uno de estos fue realizado por uno de los integrantes del equipo: respectivamente, David Fuquen, Valeria Caro y Mariana Ruiz.

En primer lugar, la Regresión Logística Multinomial es un modelo lineal para clasificación que nace a partir de la Regresión Logística. Como su nombre lo indica, es una adaptación de la Regresión Logística hecha para poder manejar variables objetivo con más de dos categorías. Como se explica en la documentación de Scikit Learn (Scikit Learn, S.F.), el objetivo de la Regresión Logística es modelar la probabilidad de que un dato de entrada pertenezca a una de las K clases de la variable objetivo. Tras estimar la probabilidad de que la observación haga parte de cada clase, la clase con mayor probabilidad es elegida como la predicción. Después, el modelo realiza un método de máxima verosimilitud para realizar un ajuste entre las probabilidades predichas y las reales. En el caso de nuestro algoritmo, se maneja también regularización L2 para penalizar coeficientes muy grandes. Para encontrar el valor óptimo del hiperparámetro “C” (fuerza inversa de la regularización) realizamos una K-Fold Cross Validation, llegando a un C=5 que proporcionaba los mejores resultados, indicando que la regularización debía ser más débil que el default de C=1.

En segundo lugar, utilizamos un modelo de K-Nearest Neighbor. En este, como se explica en la documentación de Scikit Learn (Scikit Learn, S.F.), el dato de entrada es comparado con sus K-vecinos más cercanos, siéndole asignado la clase más repetida en estos vecinos (mayoría de votación). Para definir la métrica con la cual se calcula la distancia entre los datos (vecinos), la cantidad de vecinos a revisar y el método de ponderación de vecinos, se utilizó un K-Fold Cross Validation. Según este, los hiperparámetros que permiten el mejor modelo son: 7 vecinos, ponderación según la distancia y una métrica de tipo Minkowski.

Finalmente, utilizamos un modelo Naive Bayes multinomial, que es una adaptación del modelo Naive Bayes especialmente diseñado para clasificación de textos con múltiples clases. Como se explica en la documentación de Scikit Learn (Scikit Learn, S.F.), el modelo se basa en el teorema de Bayes, donde se asume que todos los features son independientes entre sí, dada la clase. Utilizando este teorema, el algoritmo predice la clase de un nuevo dato calculando la probabilidad posterior de cada clase y eligiendo la que tenga mayor probabilidad.

4. Resultados

En primer lugar, se analizan los resultados arrojados por el modelo de Regresión logística. Este modelo muestra un excelente desempeño con un F1-score del 97.9%, lo que indica que clasifica correctamente la mayoría de los registros. Las métricas de precisión (0.9793), recall (0.9794) y exactitud (0.9790) también son consistentemente

altas, lo que demuestra equilibrio adecuado entre la capacidad del modelo para identificar correctamente las clases (precisión) y su capacidad para no omitir verdaderos positivos (recall). Al analizar cada clase por separado, es posible notar que el ODS 3 tiene un mejor desempeño, sin embargo, el modelo también se ajusta bien para las demás categorías. En general, el modelo tiene un excelente ajuste y es confiable para clasificar textos relacionados con los ODS 3, 4 y 5 en este conjunto de datos.

En segundo lugar, al analizar el modelo de K-Vecinos más cercanos se puede concluir que también se ajusta bien a los datos, mostrando un F1-score del 96.3%, lo que refleja una clasificación correcta en la mayoría de los casos. Las métricas de precisión (0.9647), recall (0.9626) y exactitud (0.9629) son consistentes y sugieren que el modelo logra un buen equilibrio entre la identificación de verdaderos positivos y la reducción de falsos positivos. A nivel de clase, el ODS 3 tiene la mayor precisión (1.00), pero su recall es ligeramente menor (0.94), lo que indica que el modelo es excelente para predecir correctamente los casos de esta clase, pero omite algunos ejemplos. En general, el modelo ofrece un rendimiento sólido y es confiable, aunque no tan preciso como la Regresión Logística, especialmente en la clase ODS 3, donde tiene un ligero margen de mejora en la identificación de todos los casos.

Por último, el modelo de Naive Bayes Multidimensional se ajusta muy bien a los datos, con un F1-score del 96.7% y métricas consistentes de precisión (0.9677), recall (0.9664) y exactitud (0.9667). A nivel de clase, el ODS 3 presenta una alta precisión (99%), aunque con un recall ligeramente inferior (96%), lo que indica que el modelo es muy preciso, pero omite algunos ejemplos. Las clases ODS 4 y ODS 5 tienen un rendimiento balanceado con F1-scores cercanos a 0.97. En general, el modelo ofrece un desempeño sólido y consistente, siendo muy confiable para clasificar textos relacionados con los ODS 3, 4 y 5, aunque podría mejorar ligeramente en la captura de todos los casos del ODS 3.

El mejor modelo, basándonos en las métricas proporcionadas, es el de Regresión Logística, ya que ofrece el mayor F1-score y un excelente equilibrio entre precisión, recall y exactitud. Además, su rendimiento es superior o comparable al de los otros modelos en todas las clases, especialmente en la clase ODS 3, donde mantiene un F1 de 0.99, el más alto de todos los modelos evaluados.

En cuanto a si los modelos cumplen con los objetivos de negocio, que incluye la correcta clasificación de textos para ayudar a la toma de decisiones sobre los Objetivos de Desarrollo Sostenible (ODS), la Regresión Logística sería la opción más adecuada. Su alto desempeño en todas las métricas asegura que las predicciones sean precisas y consistentes, minimizando tanto los errores de clasificación como las omisiones, lo que es crucial para aplicaciones que involucren decisiones basadas en los ODS.

Ya construidos los modelos, se presenta el análisis de las palabras más comunes presentes en los textos asociados a cada categoría:

Ilustración 1. Palabras comunes en ODS 3, 4 y 5 respectivamente

Como es posible notar en la Ilustración 1, las palabras encontradas en los textos asociados a cada ODS están íntimamente relacionadas a el tema de este. Por ejemplo, para el ODS 3 (Salud y Bienestar) las palabras más frecuentes son “salud”, “servicio” y “paciente”. Estas relaciones entre palabras y ODS, permitieron la construcción de los modelos ya mencionados, y la elección de la regresión logística como mejor modelo. Con este modelo se realizaron las predicciones sobre nuevos datos de la misma forma que El Fondo de Poblaciones de las Naciones Unidas lo podrá realizar en un futuro.

Además, el Fondo de Poblaciones de las Naciones Unidas (UNFPA) puede utilizar los resultados del modelo de Regresión Logística para identificar problemas prioritarios asociados a los ODS 3, 4 y 5 y ajustar políticas públicas de forma ágil, priorizando las necesidades más urgentes. Este enfoque permitirá segmentar a la población según sus preocupaciones, diseñar intervenciones específicas, y monitorear el impacto de las políticas a lo largo del tiempo. Además, el reentrenamiento continuo del modelo garantizará que las políticas respondan a nuevas tendencias y necesidades emergentes. Al comunicar de manera clara los resultados y su impacto en las decisiones, se fomentará la participación ciudadana, lo que fortalecerá la transparencia y confianza en las instituciones públicas, optimizando así los recursos destinados a salud, educación e igualdad de género.

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

Tabla 2. Actores relacionados al producto creado

Rol dentro de la organización	Tipo de actor	Beneficio	Riesgo
Dirección de políticas públicas	Usuario-cliente	Acceso a información oportuna y basada en datos para diseñar políticas más efectivas y focalizadas.	Si el modelo clasifica incorrectamente, se pueden formular políticas basadas en datos erróneos.
Entidades gubernamentales (ministerios)	Financiador	Toma de decisiones más eficiente y basada en evidencia para la asignación de recursos públicos.	Inversión de recursos en un proyecto que, si falla, podría generar una mala asignación de fondos.
Equipo de desarrollo tecnológico	Proveedor	Garantiza la implementación técnica y la adaptación del modelo en plataformas web o móviles.	Mal manejo o protección inadecuada de los datos ciudadanos, lo que podría conllevar problemas legales.
Ciudadanos (habitantes locales)	Beneficiado	Se beneficiarán de políticas más alineadas a sus necesidades reales, mejorando su calidad de vida.	Si el modelo clasifica incorrectamente, se podrían ignorar las verdaderas necesidades de ciertos grupos.
ONGs y organizaciones sociales	Colaborador	Facilitan la recolección de datos ciudadanos, y obtienen datos que pueden usar para abogar por cambios.	Si los resultados no son precisos, podrían basar su activismo en información poco confiable.
Analistas de datos y científicos	Usuario interno	Herramienta que reduce el tiempo y esfuerzo en el análisis de grandes volúmenes de texto.	Dependencia excesiva en el modelo, lo que podría limitar la supervisión manual y análisis contextual.

6. Trabajo en equipo

A continuación, se define cada uno de los roles del equipo:

- **Líder de proyecto:** Mariana Ruiz Giraldo, se encargó de pactar las reuniones semanales, de definir cada una de las tareas y revisar el trabajo de cada uno de los integrantes con el fin de mantener coherencia en el proyecto.
- **Líder de negocio:** David Alejandro Fuquen, fue responsable de velar por resolver el problema o la oportunidad identificada y alinearse con la estrategia del negocio.

- **Líder de datos:** Valeria Caro, encargada de gestionar los datos que se van a usar en el proyecto y la asignación de las tareas sobre los datos.
- **Líder de analítica:** David Alejandro Fuquen, se encarga de gestionar las tareas de analítica del grupo. Se encarga de verificar que los modelos escogidos cumplieran con los estándares del análisis para obtener los mejores resultados.

Cada integrante realizó un algoritmo,

- Logistic Regression: David Alejandro Fuquen
- K-Nearest Neighbors: Valeria Caro Ramírez
- Multinomial Naive Bayes: Mariana Ruiz Giraldo

Adicionalmente, se realizaron aproximadamente 5 reuniones virtuales/presenciales en la semana para comprender los avances del proyecto y revisar oportunidades de mejora. En las primeras dos reuniones nos enfocamos en entender el problema del negocio y los datos brindados. Después en la reunión 3 y 4 nos enfocamos en la limpieza de los datos (entre todos) y la implementación de cada uno de los algoritmos propuestos. Por último, la última sesión se utilizó para definir el documento y la preparación del video.

En cuanto a los 100 puntos que hay que repartir entre los integrantes, se decidido repartir los puntos equitativamente entre los integrantes, en este caso a cada integrante le corresponden 33.33 puntos para completar los 100 puntos. Esta decisión, se tomó debido a que cada uno de los aportes de los integrantes fue fundamental para la realización del proyecto.

7. Bibliografía

United Nations. (2023). *Objetivos de desarrollo sostenible*. Obtenido de https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/#

SciKit Learn (S.F.). *Linear Models 1.1.11 Logistic Regression* Retribuido el 07/09/24 de https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression

SciKit Learn (S.F.). *Naive Bayes 1.9.2 Multinomial Naive Bayes* Retribuido el 07/09/24 de https://scikit-learn.org/stable/modules/naive_bayes.html#multinomial-naive-bayes

SciKit Learn (S.F.). *Nearest Neighbors 1.6.2 Nearest Neighbors Classification* Retribuido el 07/09/24 de https://scikit-learn.org/stable/modules/neighbors.html#nearest-neighbors-classification