UNIVERSIDAD DE LOS ANDES
FACULTAD DE INGENIERIA
INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

Universidad de los Andes
Facultad de Ingeniería

PROYECTO ENTREGA GRUPAL

Presentado a
Haydemar Nuñez

Presentado por
Tomas Acosta Bernal - 202011237
Diego Andres Rubio - 201816492
David Alvarez Polo - 201911318

BOGOTÁ DC - COLOMBIA
14 de Octubre 2023

Índice general

Capítulo 1

Primera Parte

1.1. Entendimiento del negocio y enfoque analítico

1.1.1. Oportunidad / Problema negocio

La comunicación efectiva y la retroalimentación del cliente son esenciales para cualquier empresa o industria, ya que permiten comprender la experiencia real del cliente y su perspectiva sobre los productos o servicios ofrecidos. En la actualidad, hay muchas formas de obtener grandes volúmenes de datos relacionados con la opinión de los clientes sobre una empresa. Sin embargo, procesar esta retroalimentación puede ser un desafío cuando se trata de un gran número de comentarios, ya que leer y analizar cada uno de ellos llevaría mucho tiempo. En este contexto, la analítica de texto se presenta como una oportunidad significativa para las empresas, ya que permite automatizar el procesamiento de grandes cantidades de comentarios y extraer información valiosa.

Un objetivo de desarrollo sostenible ODS está diseñado para acabar con la pobreza, el hambre, la discriminación contra mujeres y niñas, entre otras problemáticas. La creatividad, el conocimiento, la tecnología y los recursos financieros de toda la sociedad son necesarios para alcanzar los ODS en todos los contextos. En este caso de estudio específico, la analítica de texto se utilizará para determinar el tipo de ODS (objetivo de desarrollo sostenible) al cual se relaciona cada texto. Esta herramienta permite a la empresa obtener información de entrada valiosa que puede interpretarse rápidamente, lo que a su vez facilita la toma de decisiones más informadas y acertadas en cuanto a los habitantes locales y sus problemáticas de su entorno particular. Los ODS tratados en este proyecto serán el 3, 4 y 5.

ODS 3: Salud y Bienestar

Objetivo: Garantizar una vida saludable y promover el bienestar para todas las edades.

Impacto en Colombia: El logro del ODS 3 en Colombia implicaría mejoras significativas en el acceso a servicios de salud de calidad, la reducción de la mortalidad infantil y materna, y la lucha contra enfermedades como el VIH/SIDA, la malaria y la tuberculosis. Esto tendría un impacto directo en la salud de la población colombiana y en la reducción de las desigualdades en el acceso a la atención médica.

ODS 4: Educación de Calidad

Objetivo: Garantizar una educación inclusiva, equitativa y de calidad, y promover oportunidades de aprendizaje durante toda la vida para todos.

Impacto en Colombia: El cumplimiento del ODS 4 en Colombia significaría una mejora significativa en la calidad y el acceso a la educación en todos los niveles. Esto tendría un impacto positivo en la formación de la fuerza laboral, reduciría la brecha de desigualdad educativa y contribuiría al desarrollo sostenible y al crecimiento económico del país.

ODS 5: Igualdad de Género

Objetivo: Lograr la igualdad de género y empoderar a todas las mujeres y niñas.

Impacto en Colombia: La realización del ODS 5 en Colombia sería fundamental para la promoción de la igualdad de género y el empoderamiento de las mujeres y las niñas. Esto tendría un impacto positivo en la reducción de la violencia de género, el acceso a oportunidades económicas, la participación en la toma de decisiones y la construcción de una sociedad más equitativa en Colombia.

1.1.2. Enfoque analitico

En el contexto de esta problemática, donde se busca implementar una clasificación sobre un vector "bag of words"(bolsa de palabras), se han seleccionado tres algoritmos de clasificación particulares: Regresión Logística Multivariada, Naive Bayes y Random Forest. A continuación, se explica por qué estos algoritmos son óptimos para este problema:

Regresión Logística Multivariada

- **Optimalidad para el Problema:** La regresión logística multivariada es una elección adecuada cuando se trata de problemas de clasificación. Su capacidad para modelar relaciones entre características y asignar probabilidades a múltiples clases es beneficiosa en este contexto.
- **Flexibilidad:** Permite la clasificación de datos en múltiples categorías, lo que se alinea con la naturaleza de las categorías de este problema (por ejemplo, categorías 3, 4 y 5).
- **Interpretabilidad:** La regresión logística proporciona coeficientes que indican la importancia relativa de cada palabra en la clasificación, lo que puede ayudar a entender qué características influyen en la toma de decisiones.

Naive Bayes

- **Optimalidad para el Problema:** Naive Bayes es particularmente útil en problemas de clasificación de texto, como el presente. Su suposición de independencia condicional entre características (en este caso, palabras) puede ser adecuada en muchas situaciones de procesamiento de lenguaje natural.
- **Eficiencia:** Naive Bayes es computacionalmente eficiente y suele funcionar bien con conjuntos de datos de alta dimensionalidad, como los vectores "bag of words".
- **Capacidad para Modelar Probabilidades:** El enfoque probabilístico de Naive Bayes es útil para la clasificación y proporciona una medida de confianza en las predicciones.

Random Forest

- **Optimalidad para el Problema:** Random Forest es una elección sólida para problemas de clasificación, especialmente cuando se busca un alto rendimiento predictivo. Su capacidad para manejar múltiples características y la construcción de múltiples árboles de decisión lo hace adecuado para tareas de clasificación de texto.
- **Resistencia al Sobreajuste:** Random Forest es conocido por su capacidad para mitigar el sobreajuste, lo cual es crucial en problemas con datos ruidosos o conjuntos de datos pequeños.
- **Importancia de Características:** El algoritmo puede proporcionar información sobre la importancia de cada palabra en la clasificación, lo que puede ser útil para interpretar y entender los resultados.

1.1.3. Organización y rol dentro de ella que se beneficia con la oportunidad definida

Es evidente que este enfoque tiene el potencial de ser beneficioso para el Fondo de Poblaciones de las Naciones Unidas (UNFPA), ya que la retroalimentación de los habitantes locales puede ser una fuente valiosa de información. Esto es crucial porque permite a la UNFPA comprender tanto sus puntos fuertes como las áreas en las que pueden mejorar. Es importante destacar que cuando se necesita analizar una gran cantidad de datos, las herramientas de

procesamiento automático de texto ofrecen ventajas significativas debido a su capacidad de escalamiento.

En el contexto específico de los objetivos de desarrollo sostenible 3, 4 y 5, la implementación exitosa de estos ODS en Colombia tendría un impacto significativo en la salud, la educación y la igualdad de género en el país, lo que a su vez contribuiría al desarrollo sostenible, la reducción de la desigualdad y el bienestar de la población colombiana.

En realidad, los posibles beneficiarios son numerosos y variados, ya que esta tecnología brinda la oportunidad de obtener información valiosa para cualquiera interesado en comprender mejor las necesidades de la sociedad. Además, es importante destacar que esta implementación también sería ventajosa para los clientes, ya que la UNFPA podría conocer mejor a la sociedad y ofrecerles soluciones que se ajusten más a sus necesidades.

1.2. Entendimiento y preparación de los datos

1.2.1. Entendimiento Datos

En el proceso de entendimiento de los datos, se realizó un análisis detallado del archivo proporcionado, el cual presentaba dos columnas. La primera columna incluía textos en español, mientras que la segunda columna correspondía a un valor denominado 'sdg'. En total, el DataFrame contenía 3000 registros.

Se llevaron a cabo las siguientes observaciones clave:

- **Conteo de Palabras Frecuentes:** Se identificó que las palabras 'de', 'la', 'y', 'en' y 'los' eran las más frecuentes en los textos analizados. Estas palabras son comunes en el idioma español y, en la mayoría de los casos, no aportan información distintiva al análisis, por lo que se consideraron candidatas para su eliminación o inclusión en la lista de stopwords.
- **Balance de Categorías en la Columna 'sdg':** Se notó que los datos en la columna 'sdg' estaban equilibrados, ya que cada una de las tres categorías (3, 4 y 5) tenía exactamente 1000 registros. Este equilibrio en la distribución de categorías es valioso, ya que asegura que el modelo no esté sesgado hacia ninguna categoría específica.

Estas observaciones iniciales proporcionan una base importante para la comprensión de los datos y orientan las futuras decisiones en el proceso de análisis y modelado de datos en el contexto de la ciencia de datos. Además, la identificación de palabras frecuentes puede ser útil para refinar aún más el preprocesamiento de los textos, mientras que el equilibrio en las categorías 'sdg' es un aspecto relevante a considerar en tareas de clasificación o modelado predictivo.

Por otro lado, la inclusión de un análisis de la longitud de palabras en los textos es una observación valiosa para el entendimiento de los datos. En este caso, se determinó que la longitud de las palabras variaba entre un mínimo de 9 caracteres y un máximo de 83 caracteres. Esta información proporciona una visión más completa de la estructura y la variabilidad de los textos en el conjunto de datos.

Este conocimiento sobre la longitud de las palabras puede ser útil en varias fases del proceso de análisis de datos, como la tokenización, el preprocesamiento de texto y la selección de características en tareas de procesamiento de lenguaje natural. Además, puede ayudar a identificar posibles problemas en los datos, como palabras demasiado cortas o largas que podrían requerir un tratamiento especial en futuros pasos del análisis.

1.2.2. Preparacion Datos

En la fase inicial de preparación de datos, se implementó un proceso de limpieza exhaustiva con el objetivo de garantizar la calidad y uniformidad de los datos. Este proceso se dividió en las siguientes etapas:

- **Eliminación de Caracteres No ASCII:** Se inició el proceso mediante la eliminación de caracteres no ASCII de la lista de palabras recibida. Este paso resultó fundamental para asegurar la integridad y consistencia de los datos.

- **Conversión a Minúsculas:** A continuación, se uniformizaron los datos al convertir todas las palabras a minúsculas. Esto facilitó la estandarización y el manejo eficiente de los mismos.
- **Eliminación de Signos de Puntuación:** Dado que los signos de puntuación no aportan información relevante al modelo, se procedió a eliminarlos. Esto permitió enfocarse en el contenido esencial de los textos.
- **Remoción de Stopwords:** Se utilizó una biblioteca de palabras comunes en el idioma español para eliminar stopwords, o palabras de paro, que no contribuyen significativamente al análisis.
- **Corrección de Caracteres Extraños:** Se identificaron y corrigieron palabras que contenían caracteres extraños, que a menudo se originan por errores en la importación de datos desde fuentes externas.
- **Reemplazo de Números:** Se llevó a cabo la sustitución de palabras que contenían dígitos numéricos por palabras, empleando una herramienta de procesamiento específica.

Estas operaciones de preprocesamiento se encapsularon en una función para asegurar su reutilización y aplicabilidad sistemática a los datos. A continuación, se describen las etapas posteriores:

- **Tokenización:** Se procedió a la tokenización, que implica la segmentación de frases u oraciones en palabras individuales. Sin embargo, se realizó una corrección previa de contracciones para preservar la integridad semántica.
- **Normalización de Datos:** En esta etapa, se llevó a cabo la normalización de palabras, incluyendo la eliminación de prefijos y sufijos, junto con una lematización, con el propósito de estandarizar las palabras y reducir la variabilidad.
- **Vectorizacion de Datos:** A continuación, se procedió con la vectorización de palabras utilizando la técnica conocida como "bag of words"(bolsa de palabras). Esta elección se basó en la preferencia por crear un modelo en el que se asigne una clasificación binaria, indicando si una palabra está presente o no en el documento. La técnica de bag of words.[es] una forma eficaz de representar documentos de texto como vectores numéricos, donde cada dimensión del vector corresponde a una palabra única en el conjunto de datos. En este enfoque, se registra la presencia o ausencia de cada palabra en el documento, sin considerar el orden de las palabras o su frecuencia.

Estas operaciones de preparación de datos son fundamentales en el proceso de análisis y modelado de datos en proyectos de ciencia de datos y procesamiento de lenguaje natural.

1.3. Modelado y evaluación

En el contexto del proceso de modelado, se implementó un pipeline que integraba las funciones necesarias para llevar a cabo las transformaciones esenciales de los datos. Este pipeline permitió gestionar eficazmente el flujo de datos y aplicar las preparaciones requeridas de manera sistemática.

En un paso crucial, se procedió a dividir el conjunto de datos de entrenamiento en dos subconjuntos distintos: uno destinado al entrenamiento (denominado "train") y otro al conjunto de prueba ("test"). Esta división es fundamental para evaluar y validar el rendimiento de los modelos de aprendizaje automático antes de pasarle los datos reales para la prueba.

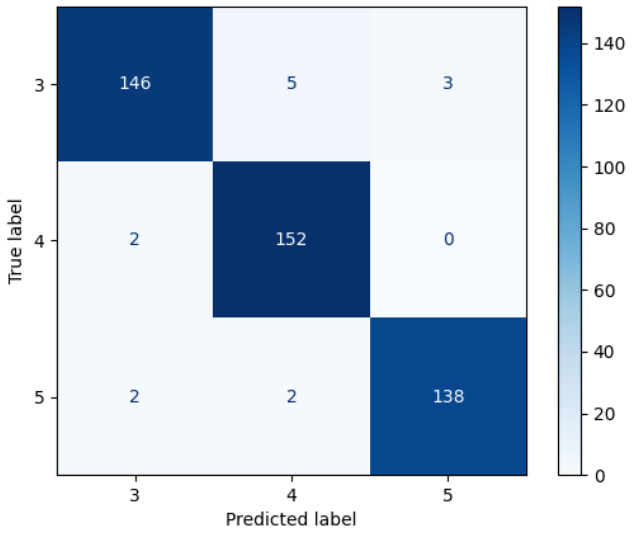
El proceso de evaluación de modelos se realizó siguiendo un enfoque escalonado. En primer lugar, se evaluaron varios algoritmos sobre el pipeline, los cuales se mencionaron previamente. Esta evaluación inicial permitió observar cómo se desempeñaban estos algoritmos utilizando los datos de entrenamiento, lo que proporcionó información valiosa sobre su capacidad para aprender de los datos.

Posteriormente, se aplicaron los modelos entrenados a los datos de prueba. Esta fase de evaluación en datos de prueba es crítica, ya que permite determinar cómo se comportan los modelos en datos que no han sido vistos durante el proceso de entrenamiento. Esta evaluación es esencial para medir la capacidad de generalización de los modelos y determinar su utilidad.

1.3.1. Regresion Logistica (Tomás Acosta)

La Regresión Logística Multivariada emplea el sigmoide para modelar la relación entre las variables independientes y la probabilidad de pertenecer a una categoría particular. Esta función garantiza que las probabilidades se mantengan en el rango de 0 a 1. Debido a esto sigue las siguientes caracteristicas:

- **Coeficientes de Características:** El modelo asigna coeficientes a cada característica o variable independiente. Estos coeficientes indican la importancia relativa de cada característica en la clasificación y pueden interpretarse para comprender cómo cada característica influye en la decisión.
- **Interpretabilidad:** La Regresión Logística Multivariada es conocida por su interpretabilidad. Los coeficientes de características proporcionan información sobre cómo y cuánto influye cada característica en la probabilidad de pertenecer a una categoría.
- **Suposiciones y Limitaciones:** Este modelo asume una relación lineal entre las características y la probabilidad logarítmica de la pertenencia a una categoría. Además, asume independencia condicional entre las características, lo que a veces puede ser una suposición simplista en la vida real.



Grafica 1. Matriz Confusion Regresion Logistica Multivariada

Como se puede apreciar en la imagen, de un conjunto de 154 datos destinados a la clasificación 3, se identificaron 8 de ellos que fueron incorrectamente asignados. En el caso de la clasificación 4, que constaba de 154 datos, se detectaron 2 instancias que fueron erróneamente ubicadas. Finalmente, en la clasificación 5, compuesta por 142 datos, se evidenciaron 4 predicciones incorrectas. Estos hallazgos proporcionan información valiosa sobre el rendimiento y la precisión del modelo en la tarea de clasificación en las diferentes categorías.

1.3.2. Naive Bayes (Diego Rubio)

Para Naive Bayes, se obtuvo una precisión del 0.971, posicionandose como el mejor de los 3 algoritmos, eso significa que el modelo de Naive Bayes fue capaz de realizar predicciones correctas para aproximadamente el 97.1 % de los casos en el conjunto de datos de prueba. En otras palabras, el modelo fue el mas acertado en la clasificación de las instancias.

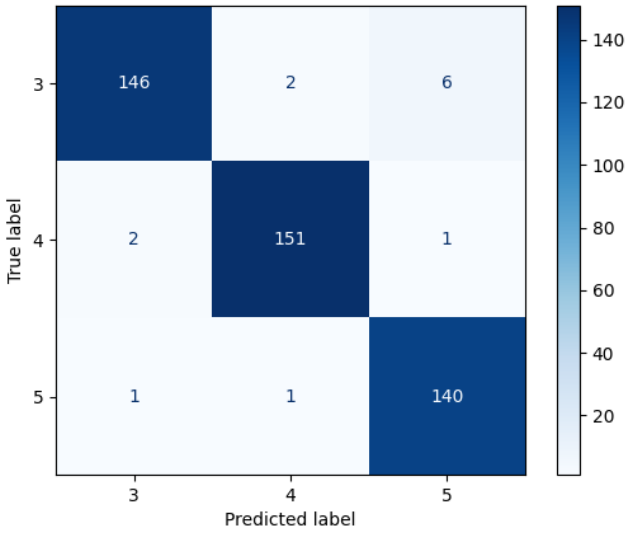
La razón por la cual el modelo de Naive Bayes ha obtenido la mayor precisión podría deberse a varias razones:

- **Suposición de Independencia Condicional:** Naive Bayes hace suposiciones sobre la independencia condicional de las características, lo que puede ser adecuado en ciertos conjuntos de datos, especialmente en

problemas de clasificación de texto. Esta suposición puede ayudar al modelo a capturar patrones sutiles en los datos.

- **Eficiencia Computacional:** Naive Bayes es un algoritmo computacionalmente eficiente y rápido en comparación con algunos otros algoritmos, como Random Forest, lo que puede ser ventajoso en términos de tiempo de entrenamiento y predicción.
- **Adecuación para Datos de Texto:** Naive Bayes es conocido por su buen rendimiento en tareas de procesamiento de lenguaje natural, como clasificación de texto, análisis de sentimientos y detección de spam. Si estás trabajando con datos de texto, Naive Bayes puede ser una elección sólida.
- **Menos Sensible a Overfitting:** En algunos casos, Naive Bayes puede ser menos sensible al sobreajuste (overfitting) en comparación con modelos más complejos, como Random Forest. Esto puede ser beneficioso cuando se dispone de un conjunto de datos limitado.

En este caso, tambien se realizo la matriz de confusion:

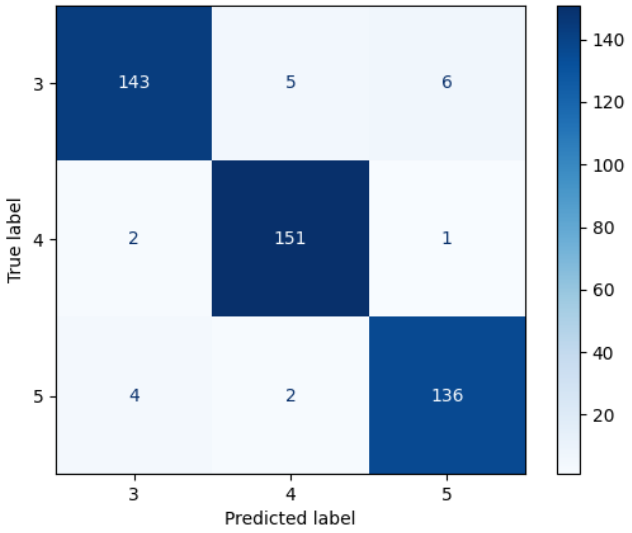


Grafica 2. Matriz Confusion Naive Bayes

Como se puede identificar en la imagen, de un conjunto de 154 datos destinados a la clasificación 3, se identificaron 8 de ellos que fueron incorrectamente asignados. En el caso de la clasificación 4, que constaba de 154 datos, se detectaron 3 instancias que fueron erróneamente ubicadas. Finalmente, en la clasificación 5, compuesta por 142 datos, se evidenciaron 2 predicciones incorrectas. Estos hallazgos proporcionan información valiosa sobre el rendimiento y la precisión del modelo Naive Bayes en la tarea de clasificación en las diferentes categorías.

1.3.3. Random Forest (David Polo)

Para el caso de Random forest, Se eligio de hiperparametro n_estimators = 200, ya que se observo que despues de este valor el modelo empezaba a sobreajustarse y ya no estaba aumentando mas el valor accuracy del modelo. Por lo tanto, se obutvo un accuracy de 95,7 %. Un valor bastante alto al igual que los otros 2 algoritmos.



Grafica 3. Matriz Confusion Random Forest

Como se puede observar en la imagen, de un conjunto de 154 datos destinados a la clasificación 3, se identificaron 11 de ellos que fueron incorrectamente asignados. En el caso de la clasificación 4, que constaba de 154 datos, se detectaron 3 instancias que fueron erróneamente ubicadas. Finalmente, en la clasificación 5, compuesta por 142 datos, se evidenciaron 6 predicciones incorrectas. Estos hallazgos proporcionan información valiosa sobre el rendimiento y la precisión del modelo en la tarea de clasificación en las diferentes categorías.

Es relevante destacar que el modelo Random Forest mostró el peor rendimiento entre los tres algoritmos evaluados en este análisis cuantitativo. Esto puede atribuirse a varios factores, como la configuración de los hiperparámetros y la complejidad inherente del algoritmo. Dado que el Random Forest es un modelo más complejo y potencialmente propenso al sobreajuste, es esencial realizar ajustes adecuados de hiperparámetros y considerar la naturaleza de los datos para mejorar su rendimiento.

1.4. Resultados

A continuación se presentan los resultados obtenidos con los tres modelos, su exactitud para cada uno y las matrices de confusión generadas con cada algoritmo las cuales ya fueron mostradas en *Gráfica 1*, *Gráfica 2* y *Gráfica 3*.

Para el algoritmo de Regresión Logística se obtuvo una exactitud de 0.968. En el modelo que utiliza el algoritmo Naive Bayes se obtuvo una exactitud de 0.971. Por último, para el modelo que utilizó el algoritmo de Random Forest se obtuvo una exactitud de 0.957. Siendo así el modelo con mejor exactitud es el de Naive Bayes.

Dado que el objetivo por parte de la UNFPA y las demás entidades públicas radica en el procesamiento de grandes volúmenes de texto para extraer conocimiento (herramientas de participación ciudadana) que respalde la toma de decisiones, se aconseja la elección de un algoritmo con un rendimiento sólido en términos de precisión y recall. De los tres algoritmos que hemos implementado, se ha observado que Naive Bayes sobresale en términos de precisión, recall y puntuación F1 tanto para las categorías positivas como negativas, en comparación con la Regresión Logística y el Random Forest. Además, Naive Bayes es un modelo de fácil interpretación, lo que podría resultar valioso para la organización en el proceso de extracción de conocimiento.

Por lo tanto, se recomienda la implementación del modelo de Naive Bayes como el algoritmo de elección para la clasificación de objetivos de desarrollo sostenible en el análisis de comentarios por parte de la sociedad. Un escenario en el que este algoritmo podría beneficiar a la UNFPA es en la clasificación automatizada de opiniones de ciudadanos sobre sus problemas sociales, económicos, entre otros. La empresa podría utilizar Naive Bayes para etiquetar automáticamente estas opiniones entre las categorías de ODS 3, 4 o 5, lo que permitiría identificar

rápidamente las tendencias de opinión sobre sus problemáticas de algún entorno en particular. Esto, a su vez, beneficiaría a la empresa en:

1. Análisis de opiniones que representan la voz de los habitantes locales: La compañía podría detectar con prontitud el tipo de ODS al cual se relaciona la problemática que comenta algún habitante por medio de las herramientas de participación ciudadana y ajustar sus estrategias de solución de tales situaciones.

2. Ahorro de tiempo y recursos: La automatización de la clasificación de opiniones reduciría la necesidad de la clasificación manual de comentarios, lo que agilizaría la respuesta a las problemáticas de los diferentes entornos y sus habitantes locales, permitiendo un uso más eficiente de los recursos estatales.

En resumen, la elección de Naive Bayes como algoritmo de clasificación para analizar las opiniones sobre los ciudadanos podría proporcionar una ventaja significativa a la UNFPA en términos de comprensión de las problemáticas del público y optimización de recursos del estado para el alcance de tales objetivos de desarrollo sostenible.

1.5. Mapa de actores

A continuación se presenta el mapa de los actores identificados a lo largo del proyecto.

Actor	Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Cliente	Proporcionar retroalimentación sobre los productos o servicios	Usuario-cliente	Mejora de los productos o servicios basada en la retroalimentación del cliente	Insatisfacción del cliente debido a la mala calidad del producto o servicio
Analista de datos	Procesar y analizar la retroalimentación del cliente	Empleado	Mejora de la eficiencia operativa y toma de decisiones informada	Malinterpretación de los datos del cliente
Gobierno local	Implementar y supervisar los ODS en Colombia	Entidad gubernamental	Mejora en la salud, educación y equidad de género en Colombia	Incumplimiento de los ODS

Proveedor de servicios de TI	Proporcionar la infraestructura tecnológica necesaria para el análisis de datos	Proveedor externo	Mejora de la eficiencia operativa y toma de decisiones informada	Interrupción del servicio debido a problemas técnicos
Reguladores	Supervisar el cumplimiento de las normativas y leyes relevantes	Organismo regulador	Cumplimiento de las normativas y leyes, lo que mejora la reputación de la empresa	Sanciones y multas debido al incumplimiento de las normativas y leyes
Gerente de Proyecto	Garantizar que el proyecto se realice de manera eficiente y logre sus objetivos	Empleado de la empresa	Asegurarse que se cumplan los objetivos del proyecto.	El proyecto podría retrasarse o exceder su presupuesto

Mapa de actores

1.6. Trabajo en equipo

Tomás Acosta

Roles: Líder de proyecto, líder de datos
Tareas realizadas: Entendimiento del negocio y enfoque analítico, entendimiento de los datos, preparación de los datos, creación de modelo
Tiempo: 20 horas
Algoritmo trabajado: Regresión Logística Multivariada
Retos: Preparar datos para mantenerlo en lenguaje natural, aprender herramientas y conceptos nuevos para la preparación de los datos y obtención de palabras más útiles
Soluciones: Definición de funciones dentro de pipeline para remover caracteres que no sean ASCII, pasar todas la palabras a minúsculas, remover signos de puntuación, convertir caracteres especiales, implementar código para eliminar prefijos y sufijos, reemplazar números por palabras, realizar tokenización y lemmatización de los datos (normalización de los datos mediante el uso de lems y stems)
Puntaje: 33.3

Diego Rubio

Rol: Líder de negocio
Tareas realizadas: Entendimiento del negocio y enfoque analítico, entendimiento de los datos, preparación de los datos, documento/informe
Tiempo: 20 horas
Algoritmo trabajado: Naive Bayes
Retos: Exploración de temáticas correspondientes para cada ODS, diseño de mapa de actores dentro del negocio, selección de algoritmo apropiado, evaluación y validación del mismo
Soluciones: Usando Naive Bayes se encontró el modelo con mayor exactitud, respuesta al caso propuesto por la UNFPA
Puntaje: 33.3

David Polo

Rol: Líder de analítica
Tareas realizadas: Entendimiento de los datos, análisis del mejor modelo, documento/informe, wiki
Tiempo: 20 horas
Algoritmo trabajado: Random Forest
Retos: Selección de algoritmo apropiado, evaluación y validación del mismo
Soluciones: Usando Random Forest se detecta si existe algún sobreajuste por medio de métricas de rendimiento y ajuste de hiperparámetros
Puntaje: 33.3

1.7. Referencias

1. Scikit-learn developers. (2021). Scikit-learn: Machine Learning in Python. Recuperado de: `https://scikit-learn.org/stable/index.html`

2. SciELO. (2014). Clasificación automática de textos usando redes de palabras. Recuperado de: `https://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342014000300001`.