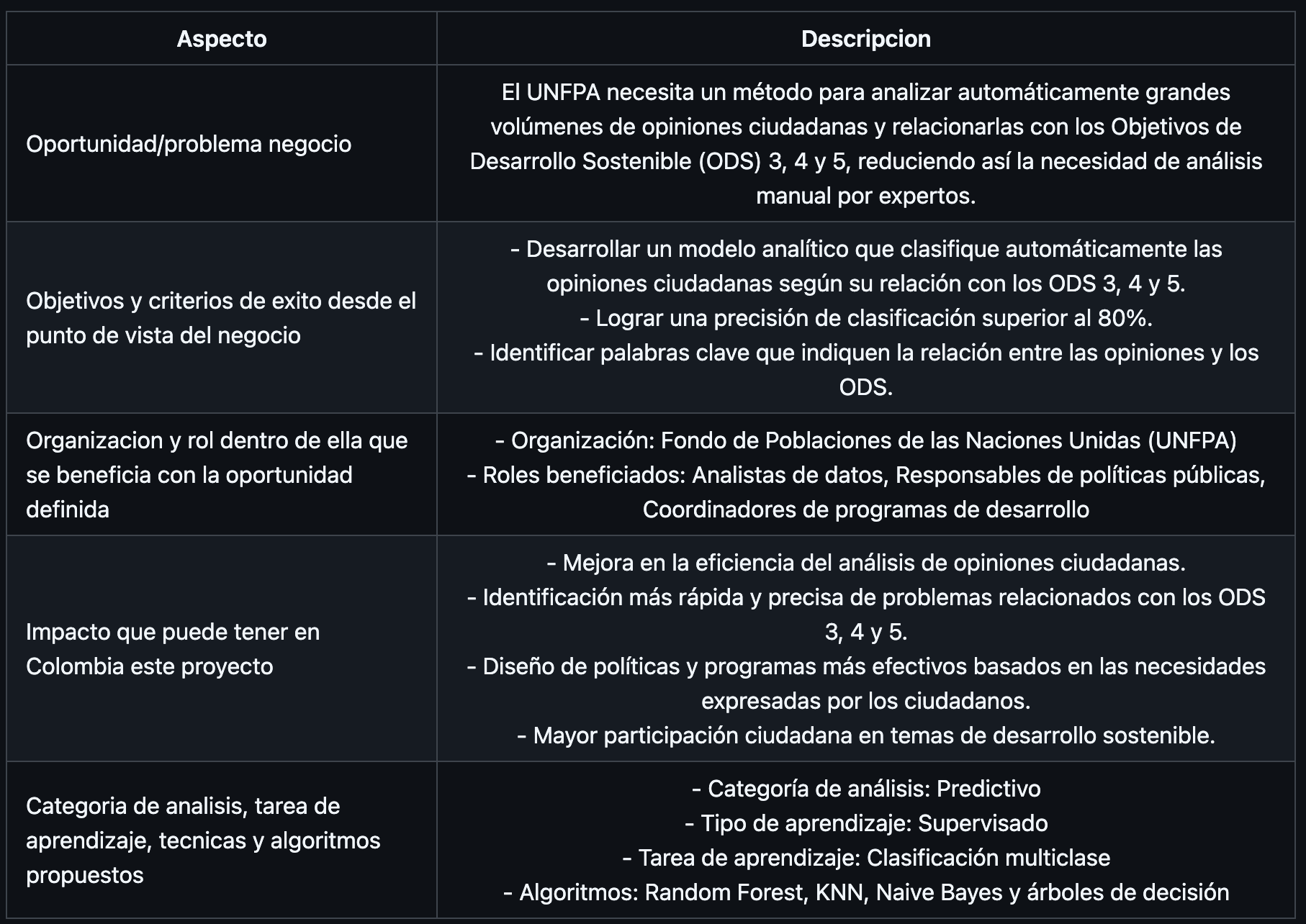
Documento: Proyecto 1 – Etapa 1

*Daniel Clavijo*

*Juan David Briceño*

*Carlos Medina*

**Fase 1 – Entendimiento de los datos**



**Fase 2 – Entendimiento y preparación de los datos**

En esta sección, nos enfocamos en el análisis y preparación de los datos para su uso en los modelos de aprendizaje automático. Los pasos clave incluyeron:

1. Perfilamiento de Datos:

Análisis de la estructura y contenido de los datos.

Identificación de columnas relevantes para la clasificación de ODS.

1. Análisis de Calidad:

Detección y manejo de valores nulos o inconsistentes.

Evaluación de la distribución de las clases (ODS 3, 4, y 5).

1. Preprocesamiento de Texto:

Limpieza de texto: eliminación de caracteres especiales y normalización.

Tokenización y eliminación de stopwords.

Aplicación de técnicas de stemming y lematización.

1. Transformaciones:

Vectorización de texto utilizando TF-IDF (Term Frequency-Inverse Document Frequency).

Normalización de características para mejorar el rendimiento de los modelos.

1. División de Datos:

Separación en conjuntos de entrenamiento y prueba para una evaluación robusta.

**Fase 3 - Modelado y evaluación**

En esta sección, implementamos y evaluamos cuatro algoritmos diferentes para la clasificación automática de textos en los ODS 3, 4 y 5:

1. Random Forest:

Implementación utilizando RandomForestClassifier de scikit-learn.

Ajuste de hiperparámetros como número de árboles y profundidad máxima.

Análisis de importancia de características para entender los factores clave en la clasificación.

1. K-Nearest Neighbors (KNN):

Uso de KNeighborsClassifier con ajuste del número de vecinos.

Evaluación de diferentes métricas de distancia para optimizar el rendimiento.

1. Naive Bayes:

Implementación de MultinomialNB, especialmente adecuado para clasificación de texto.

Ajuste de parámetros como alpha para suavizado.

Evaluación de la eficacia del modelo en el manejo de características de alta dimensionalidad.

1. Árbol de Decisión:

Uso de DecisionTreeClassifier para crear un modelo interpretable.

Experimentación con diferentes criterios de división y profundidades máximas.

Evaluación de Modelos:

Utilizamos métricas como precisión, recall, y F1-score para cada modelo.

Implementamos validación cruzada para una evaluación más robusta y para evitar el sobreajuste.

Analizamos las matrices de confusión para entender los patrones de error en la clasificación.

Selección del Modelo:

Comparamos exhaustivamente los resultados de los cuatro modelos.

El modelo Naive Bayes fue seleccionado como el mejor debido a su alto F1-score, eficiencia computacional y buen manejo de la naturaleza dispersa de los datos de texto.

Aunque Random Forest y KNN mostraron un buen rendimiento, Naive Bayes ofreció un mejor equilibrio entre precisión y velocidad.

El Árbol de Decisión, aunque más interpretable, no alcanzó el nivel de rendimiento de los otros modelos en este caso específico.

En esta sección se demostró la aplicación práctica de diferentes algoritmos de aprendizaje automático y el proceso riguroso de selección del modelo más adecuado para nuestra tarea de clasificación de ODS. La comparación entre estos cuatro enfoques diferentes proporcionó insights valiosos sobre las fortalezas y debilidades de cada técnica en el contexto de la clasificación de texto para los ODS.

**Fase 4 – Resultados**

1. **Descripción de los resultados obtenidos**

En este proyecto, hemos implementado y evaluado cuatro modelos de aprendizaje automático para clasificar textos relacionados con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. A continuación, presentamos un análisis detallado de los resultados obtenidos y su relevancia para los objetivos de la organización.

**Resumen de Rendimiento de los Modelos**

Naive Bayes:

Accuracy: 0.98

F1-score ponderado: 0.98

Random Forest:

Accuracy: 0.97

F1-score ponderado: 0.97

K-Nearest Neighbors (KNN):

Accuracy: 0.95

F1-score ponderado: 0.95

Árbol de Decisión:

Accuracy: 0.83

F1-score ponderado: 0.83

Análisis de las Métricas de Calidad

1. Precisión (Precision)

Naive Bayes y Random Forest muestran una precisión excepcional (0.97-1.00) para todas las clases.

KNN mantiene una precisión alta (0.94-0.96) en todas las clases.

El Árbol de Decisión muestra variabilidad, con alta precisión para ODS 3 y 5, pero baja para ODS 4.

1. Exhaustividad (Recall)

Naive Bayes y Random Forest tienen un recall consistentemente alto (0.97-0.98) para todas las clases.

KNN muestra un recall muy bueno (0.94-0.96) en todas las clases.

El Árbol de Decisión tiene un recall bajo para ODS 3, pero alto para ODS 4 y 5.

1. F1-Score

Naive Bayes y Random Forest logran F1-scores excepcionales (0.97-0.98) para todas las clases.

KNN mantiene F1-scores muy buenos (0.95) en todas las clases.

El Árbol de Decisión muestra variabilidad, con F1-scores de 0.71 a 0.94 dependiendo de la clase.

**Implicaciones para los Objetivos del Negocio**

1. Clasificación Precisa de Opiniones

Los modelos Naive Bayes y Random Forest, con accuracies del 98% y 97% respectivamente, permiten una clasificación altamente precisa de las opiniones ciudadanas en relación a los ODS 3, 4 y 5. Esto facilita el análisis y categorización de grandes volúmenes de datos textuales.

1. Reducción de Recursos Manuales

La alta precisión de los modelos significa que se puede reducir significativamente la necesidad de revisión manual, liberando recursos humanos para tareas más estratégicas.

1. Identificación Equilibrada de ODS

Los altos valores de recall en Naive Bayes y Random Forest indican que los modelos son capaces de identificar correctamente la mayoría de las opiniones relacionadas con cada ODS, minimizando el riesgo de pasar por alto información importante.

1. Confiabilidad en la Toma de Decisiones

Los F1-scores consistentemente altos (0.97-0.98) proporcionan una base sólida para la toma de decisiones basada en estos análisis.

1. Flexibilidad en la Elección del Modelo

La disponibilidad de múltiples modelos con buen rendimiento ofrece flexibilidad para elegir el más adecuado según las necesidades específicas.

1. Áreas de Mejora Identificadas

El rendimiento inferior del Árbol de Decisión, especialmente en la clasificación del ODS 3, señala áreas donde se podría profundizar el análisis para mejorar la precisión.

Los resultados obtenidos demuestran la viabilidad y eficacia de utilizar técnicas de aprendizaje automático para clasificar automáticamente opiniones ciudadanas en relación con los ODS 3, 4 y 5.

1. **Análisis de Palabras Clave, Estrategias Propuestas y Justificación**

**Análisis de Palabras Identificadas para Relacionar Opiniones con ODS**

* ODS 3 (Salud y Bienestar)

Palabras clave: "salud", "atención", "servicio", "paciente", "mental", "enfermedad", "médico", "tratamiento", "sanitario", "mortalidad".

Enfoque: Sistema de salud, atención médica, salud mental, calidad de los servicios sanitarios.

* ODS 4 (Educación de Calidad)

Palabras clave: "escuela", "educación", "estudiante", "docente", "aprendizaje", "alumno", "evaluación", "profesor", "programa", "habilidad".

Enfoque: Sistema educativo, calidad de la enseñanza, evaluación del aprendizaje, desarrollo de habilidades.

* ODS 5 (Igualdad de Género)

Palabras clave: "mujer", "género", "hombre", "igualdad", "derecho", "trabajo", "política", "violencia", "participación", "brecha".

Enfoque: Igualdad de derechos, participación laboral y política, violencia de género, brechas de género.

**Análisis de Clasificaciones Erróneas**

* Superposición temática entre ODS, especialmente entre educación (ODS 4) e igualdad de género (ODS 5).
* Complejidad en textos que abordan múltiples ODS simultáneamente.
* Desafíos en la clasificación de textos que mencionan aspectos transversales como políticas públicas o desarrollo.

**Estrategias Propuestas para la Organización**

* Implementar un sistema de salud holístico que integre la atención física y mental.
* Modernizar el sistema educativo, enfocándose en la calidad de la enseñanza y el desarrollo de habilidades.
* Desarrollar políticas que promuevan la equidad laboral y política, así como programas para prevenir la violencia de género.
* Adoptar un enfoque interseccional en el diseño de políticas públicas que aborden salud, educación e igualdad de género de manera integrada.
* Implementar programas de formación continua para profesionales en estos campos.
* Establecer un sistema de monitoreo y evaluación multidimensional que permita seguir el progreso en los tres ODS de manera integrada.



**Justificación de la Utilidad de esta Información**

La información obtenida a través del análisis de texto y la clasificación automática de opiniones ciudadanas es de vital importancia para el UNFPA en su misión de identificar problemas y evaluar soluciones relacionadas con los ODS 3, 4 y 5. Esta herramienta analítica permite una precisión sin precedentes en la focalización de esfuerzos, permitiendo a la organización dirigir sus recursos hacia aspectos específicos dentro de cada ODS que son de mayor preocupación para los ciudadanos. La comprensión de las complejidades y superposiciones entre los ODS, revelada a través del análisis de clasificaciones erróneas, facilita el diseño de estrategias más efectivas, abordando múltiples objetivos simultáneamente. Esto no solo optimiza los recursos, sino que también mejora el impacto global de las intervenciones. La adaptación de las comunicaciones basada en las palabras clave identificadas permite al UNFPA mejorar su diálogo con diferentes audiencias, lo que es crucial para la participación ciudadana efectiva. Además, la identificación de brechas y áreas menos abordadas en las opiniones ciudadanas señala oportunidades para investigación y acción futura. La capacidad de procesar y analizar grandes volúmenes de texto de manera automática reduce significativamente la necesidad de análisis manual por expertos, liberando recursos valiosos para otras tareas críticas. Finalmente, esta herramienta proporciona una base sólida para la mejora continua de políticas, permitiendo al UNFPA ajustar y refinar sus estrategias basándose en feedback ciudadano en tiempo real, lo que es esencial para alcanzar los ambiciosos objetivos de los ODS para 2030.

1. Excel en el repositorio de Github
2. Video: <https://youtu.be/BHpQNMoM_kc>

**Fase 5 - Mapa de actores relacionados con el producto de datos creado**



**Fase 6 – Trabajo en equipo**

El trabajo en equipo para este proyecto 1 aunque se repartio todos los miembros participaron de manera equitativa en todos los puntos que conforman el proyecto. Sin embargo, cabe resaltar las tareas en las cuales cada persona se enfocó más.

El primer reto que nos encontramos como grupo fue el total entendimiento tanto del contexto como de la construcción del modelo, en cuanto al preprocesamiento de texto el mayor encargado fue Juan David Briceño, el primer reto encontrado en esta punto fue el tratar los datos para que tomara tanto el idioma español como los caracteres que aparecian como lo era "A°", para resolverlo se trabajo con la libreria ftfy. Otro reto que nos encontramos durante la resolución del proyecto fue la correcta implementacion del algoritmo Naive Bayes. La siguiente tarea del entrenamiento del modelo se llevó a cabo por parte de Daniel Clavijo.

En cuanto a la implementaición de los algoritmos utilizados se repartieron de la siguiente manera:

Random Forest -> Daniel Clavijo.

KNN -> Juan David Briceño.

Naive Bayes -> En conjunto Juan David y Daniel.

Árboles de decisión -> Carlos Medina.

Para las dos tareas de presentación de gráficas de análisis y resultados se llevaron a cabo por Juan David y Daniel respectivamente.

Finalmente en las secciones 4, 5 y 6 del proyecto se trabajaron conjuntamente entre los todos los miembros del equipo y el video por parte de Carlos.

La distribución de puntos se tomara 33.33 para Daniel, 33.33 para Juan David y 33.33 para Carlos Medina

Lider de proyecto: El encargado de la gestión del proyecto fué Juan David Briceño.

Lider de negocio: El responsable de velar por resolver el problema o la oportunidad identificada y estar alineado con la estrategia del negocio para el cual se plantea el proyecto fué Daniel Clavijo.

Lider de datos: El encargado de gestionar los datos a usar en el proyecto y sus asignaciones fué Juan David Briceño y Carlos Medina.

Lider de analítica: El encargado de gestionar las tareas de analítica del grupo fué Daniel Clavijo y Carlos Medina.

Las reuniones que se llevaron a cabo durante la primer etapa del desarrollo del proyecto 1 en todos sus aspectos: Reunión de lanzamiento y planeación, reunión de ideación, reuniones de seguimiento, reunión de finalización, fueron siempre acordadas por los miembros del grupo y cubrieron todos los aspectos que corresponden a cada uno de los componentes anteriormente mencionados.