Nicolás Ruiz Pérez – 202123608

Nicolás Coca Peña – 202111176

Inteligencia de Negocios

Proyecto 1

Etapa 1: Construcción de modelos de analítica de textos

Sección 1: Entendimiento del negocio y enfoque analítico.

Oportunidad/problema Negocio	Necesidad de implementar un sistema eficaz de identificación y evaluación de problemas sociales a través de la información proporcionada por los ciudadanos.
Objetivos y criterios de éxito desde el punto de vista del negocio.	Objetivos: • Automatizar la clasificación de opiniones ciudadanas. • Mejorar la capacidad de respuesta. • Optimizar el uso de recursos. • Fortalecer la alineación de la UNFPA con los ODS. Criterios de éxito: • Precisión y consistencia en la clasificación. • Reducción del tiempo de análisis. • Capacidad de reentrenamiento y mejora continua. • Adopción del modelo por parte de la UNFPA. • Impacto significativo en la toma de decisiones.
Organización y rol dentro de ella que se beneficia con la oportunidad definida.	Organización: La UNFPA y entidades públicas. Rol dentro de la organización que se beneficia: Personas encargadas de la toma de decisiones con respecto a las políticas que se implementan en torno a los ODS y los ejecutores de estos proyectos.
Impacto que puede tener en Colombia este proyecto.	El impacto sería muy importante, ya que este proyecto permitiría mejorar la capacidad de respuesta a problemas

	sociales, beneficiando así a millones de personas que se benefician de estas políticas. Esto gracias a una optimización en los recursos y un enfoque en aquellos problemas que requieren una atención más urgente.
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.	Categoría de análisis: Predictivo. Tipo de aprendizaje: Supervisado. Tarea de aprendizaje: Clasificación.

Sección 2. Entendimiento y preparación de los datos.

Entendimiento:

Se obtuvieron las siguientes propiedades para los textos de entrenamiento:

```
Cantidad de filas: 4049

Cantidad promedio de palabras por fila:  110

Porcentaje de cada label:
sdg
5    35.8
4    33.4
3    30.7
Name: count, dtype: float64
```

```
Palabras más comúnes:
  - "mujeres", conteo: 2449
  - "atención", conteo: 1172
  - "educación", conteo: 1134
  - "salud", conteo: 1131
  - "países", conteo: 1124

Palabras más comúnes para el label 3 (Salud y Bienestar):
  - "salud", conteo: 1029
  - "atención", conteo: 1029
  - "servicios", conteo: 545
  - "países", conteo: 392
  - "pacientes", conteo: 300

Palabras más comúnes para el label 4 (Educación de Calidad):
  - "educación", conteo: 908
  - "estudiantes", conteo: 773
  - "escuelas", conteo: 594
  - "aprendizaje", conteo: 391
  - "evaluación", conteo: 329

Palabras más comúnes para el label 5 (Igualdad de género):
  - "mujeres", conteo: 2316
  - "género", conteo: 1055
  - "hombres", conteo: 515
  - "igualdad", conteo: 508
  - "trabajo", conteo: 414
```

Preparación:

Se implementó un procesamiento de lenguaje natural para preparar los datos con el objetivo de limpiarlos, normalizarlos y transformarlos antes de ser utilizados en el modelo de ML. El proceso comienza recuperando caracteres mal codificados (codificados con la letra "Ã") y corrigiéndolos. Luego, convierte el texto a minúsculas y reemplaza ciertos patrones específicos definidos en el diccionario `conversiones` (por ejemplo, cambia el carácter “á” por el carácter “a”). A continuación, tokeniza el texto (divide en palabras) y elimina espacios dentro de las palabras. Después, remueve las palabras vacías (stopwords) y cualquier palabra que contenga el carácter "ã". Los números son reemplazados por la palabra "num". Posteriormente, se aplica un stemming para reducir las

palabras a sus raíces. Finalmente, las palabras procesadas se concatenan de nuevo en una cadena usando la función de Python join.

Lo que permite que un texto como este: “Por ejemplo, el nÃºmero de consultas externas”

Sea transformado a este: “ejempl numer consult extern”

Sección 3. Modelado y evaluación.

Se probaron tres modelos de clasificación con tres modelos de clasificación distintos:

- Random Forest: Es un algoritmo basado en la construcción de múltiples árboles de decisión. Cada árbol se entrena con un subconjunto de los datos, y la predicción final es la mayoría de las predicciones de todos los árboles.
- Multinomial Naive Bayes (NB): Es un clasificador probabilístico basado en el teorema de Bayes que asume una distribución multinomial entre los atributos.
- K-Nearest Neighbors (KNN): Este algoritmo clasifica un punto basándose en los "k" puntos más. Se calcula la distancia entre el punto nuevo y los puntos de entrenamiento, y la etiqueta más común entre los "k" vecinos cercanos se asigna al punto.

Random Forest (Realizado por Nicolás Ruiz):

Se hizo una elección entre las siguientes opciones de hiper parámetros:

```
param_grid = {
    'n_estimators': [100, 200],
    'max_depth': [None, 20, 30],
    'min_samples_split': [5, 10],
    'min_samples_leaf': [2, 4],
    'bootstrap': [True, False]
}
```

De los cuales se obtuvo que los mejores hiper parámetros son:

```
Mejores hiperparámetros: {'n_estimators': 100, 'min_samples_split': 10, 'min_samples_leaf': 1, 'max_depth': None, 'bootstrap': False}
0.97
```

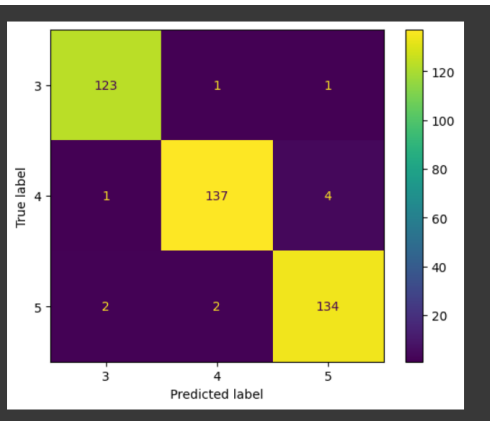
Las métricas obtenidas para este modelo fueron las siguientes y la siguiente matriz de confusión:

```
Accuracy score: 0.9728
Precision score: 0.9729
Recall score: 0.9728
F1 score: 0.9728

              precision    recall  f1-score   support

           3       0.98      0.98      0.98       125
           4       0.98      0.96      0.97       142
           5       0.96      0.97      0.97       138

    accuracy                           0.97       405
   macro avg       0.97      0.97      0.97       405
weighted avg       0.97      0.97      0.97       405
```



Multinomial Naive Bayes (Realizado por Nicolás Coca):

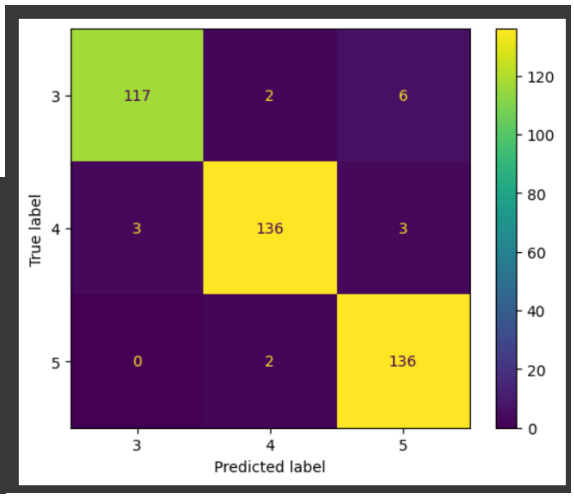
Métricas obtenidas:

```
Accuracy score: 0.9605
Precision score: 0.9611
Recall score: 0.9605
F1 score: 0.9605

              precision    recall  f1-score   support

           3       0.97      0.94      0.96       125
           4       0.97      0.96      0.96       142
           5       0.94      0.99      0.96       138

    accuracy                           0.96       405
   macro avg       0.96      0.96      0.96       405
weighted avg       0.96      0.96      0.96       405
```



K-Nearest Neighbors (Realizado por Nicolás Coca):

Se realizó una prueba de validación cruzada para seleccionar el mejor valor del hiperparámetro n_neighbors en el modelo K-Nearest Neighbors (KNN). La prueba consistió en una búsqueda en rejilla (GridSearchCV) con los valores de n_neighbors definidos como [3, 5, 7, 9], utilizando validación cruzada con 5 particiones (cv=5). Esta metodología nos permite encontrar el valor óptimo de n_neighbors que maximiza el rendimiento del modelo, minimizando el riesgo de overfitting o underfitting. Como resultado, el valor óptimo encontrado fue 9.

```python
from sklearn.model_selection import GridSearchCV
param_grid = {'n_neighbors': [3, 5, 7, 9]}
grid_search = GridSearchCV(KNeighborsClassifier(), param_grid, cv=5)
grid_search.fit(X_train_tfidf, y_train)
grid_search.best_params_['n_neighbors']
```

```
9
```

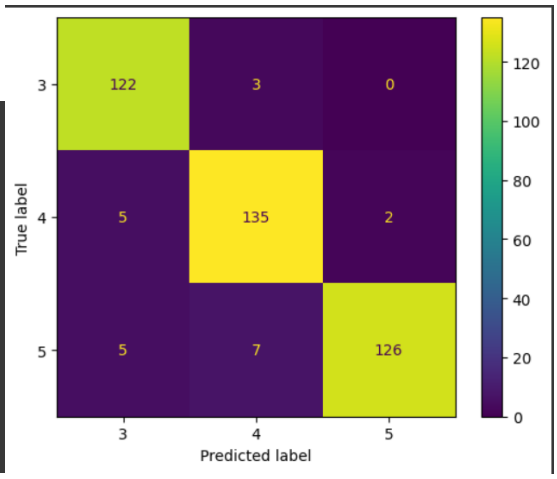
Métricas obtenidas:

```
Accuracy score: 0.9457
Precision score: 0.9471
Recall score: 0.9457
F1 score: 0.9457

              precision    recall  f1-score   support

           3       0.92      0.98      0.95       125
           4       0.93      0.95      0.94       142
           5       0.98      0.91      0.95       138

    accuracy                           0.95       405
   macro avg       0.95      0.95      0.95       405
weighted avg       0.95      0.95      0.95       405
```



Sección 4. Resultados:

a)

Se pueden los recoger las métricas de los modelos en la siguiente tabla:

	Model	Accuracy	Precision	Recall	F1
0	Random Forest	0.9728	0.9729	0.9728	0.9728
1	Multinomial NB	0.9605	0.9611	0.9605	0.9605
2	KNN	0.9457	0.9471	0.9457	0.9457

El valor de accuracy es el porcentaje de textos clasificados correctamente por el modelo.

El valor de precision representa el porcentaje en que un texto de cierta categoría se clasificó correctamente en esta categoría. Esto se hace para cada categoría y luego se promedió.

El recall es el porcentaje de textos de cada categoría que se logró identificar en la categoría correcta, se hace para cada categoría y luego se promedia.

El F1 es el promedio armónico de precisión y recall.

En el análisis de los modelos de machine learning presentados, el modelo de Random Forest destaca como el más efectivo en términos de precisión, recall y F1-score, todas con el mayor valor. Estas métricas indican que el modelo tiene una excelente capacidad (comparado con los otros dos modelos planteados) para clasificar correctamente los textos mientras minimiza los falsos positivos y falsos negativos.

Dado que el Random Forest supera en todas las métricas al resto de modelos, se escogió este para clasificar en uno de los tres Objetivos de Desarrollo Sostenible (3,4,5) los testimonios de los ciudadanos brindados por el UNFPA.

El hecho de escoger este modelo permite que se minimice la cantidad de errores de clasificación de los Objetivos de Desarrollo Sostenible, lo que permitirá que la UNFPA pueda dirigir mejor sus esfuerzos y crear las políticas correctas, encaminadas en los ods que más preocupan a los ciudadanos.

b)

Se utilizó el atributo feature_importances_ del modelo escogido para obtener las importancias de los atributos usados, los resultados son los siguientes (las palabras son raíces, no la palabra como tal. Por lo que es posible que algunas palabras no existan):

1. mujer (0.06): Palabras como mujer, mujeres.
2. gener (0.04): Palabras como Género, Géneros.
3. escuel (0.03): Palabras como escuelas, escuela, escuelero.
4. salud (0.03): Palabras como Salud, Saludable.
5. educ (0.03): Palabras como educación, educable.
6. atencion (0.02): Principalmente la palabra atención.
7. hombr (0.02): Palabras como hombre, hombres.
8. medic (0.02): Palabras como médico, médicos, médicas, medicina.
9. estudi (0.02): Palabras como estudio, estudiar, estudios.
10. aprendizaj (0.02): Principalmente aprendizaje.

Dadas estas palabras, se proponen las siguientes estrategias para cada ods:

Ods 3:

- Desarrollar programas para mejorar el acceso a atención médica, promoción de la salud preventiva y el suministro de medicamentos.

Las palabras relacionadas con la salud, como "salud", "medic" y "atencion", destacan una preocupación significativa en los testimonios analizados. Esto sugiere que la comunidad ve la salud y el acceso a servicios médicos como una prioridad. Invertir en campañas de salud preventiva, mejorar la infraestructura de atención médica y garantizar el acceso a medicamentos y servicios adecuados alinearía los esfuerzos de la organización con esta demanda, contribuyendo directamente al cumplimiento de ods 3.

Ods 4:

- Implementar programas que mejoren el acceso a la educación y refuercen la calidad de la enseñanza en todas las etapas del aprendizaje.

La importancia de las palabras "educ", "escuel" y "aprendizaj" muestran una clara preocupación por la educación, tanto en términos de infraestructura ("escuel") como de calidad educativa y aprendizaje. Esto indica que la comunidad prioriza la mejora en el acceso y la calidad educativa. Al enfocar recursos en la mejora de las instalaciones educativas y el desarrollo de programas de aprendizaje efectivos, la organización puede impactar significativamente en est ods, asegurando que más personas tengan acceso a una educación inclusiva y de calidad.

Ods 5:

- Crear programas de empoderamiento femenino y campañas para sensibilizar sobre la igualdad de género, incluyendo a hombres en la discusión.

La alta relevancia de las palabras "mujer" y "gener" indica una preocupación central por la igualdad de género y los derechos de las mujeres en los testimonios. La mención de "hombr" sugiere que también es importante incluir a los hombres en las discusiones sobre equidad. La organización podría desarrollar iniciativas que promuevan la inclusión y la igualdad de oportunidades para mujeres y hombres, combatan la violencia de género y fomenten la sensibilización y educación sobre el tema. Estos programas apoyarían directamente el ods 5, trabajando hacia un entorno más equitativo y justo.

c. El archivo se encuentra en el repositorio.

d. El video se encuentra en el padlet.

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

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Equipo de Proyectos de Sostenibilidad	Cliente	El modelo permite priorizar las iniciativas en función de las necesidades y opiniones de los ciudadanos, lo que puede mejorar la asignación de recursos para ODS específicos.	Si el modelo clasifica incorrectamente, puede enfocar recursos en prioridades erróneas, descuidando áreas clave de los ODS.

Departamento de Análisis de Datos	Financiador	Utilización de datos para mejorar la toma de decisiones basada en información, optimizando las políticas relacionadas con salud, educación y equidad de género.	Si el modelo no es preciso, los recursos invertidos en su desarrollo podrían verse malgastados y generar dudas sobre la viabilidad de proyectos futuros.
Proveedores de Infraestructura TI	Proveedor	Asegura que el sistema utilizado para correr el modelo cumple con los estándares de seguridad y manejo eficiente de datos.	Una mala gestión de la infraestructura puede llevar a filtraciones de datos confidenciales, comprometiendo la privacidad de los ciudadanos.
Ciudadanos afectados por los ODS 3, 4 y 5	Beneficiados	Reciben una atención más específica y dirigida a sus necesidades, lo que puede mejorar su acceso a servicios de salud, educación y equidad.	Si el modelo clasifica mal a las personas, se puede generar desatención a quienes realmente lo necesitan o sobrecargar servicios innecesarios.
Oficinas de Cooperación Internacional y Alianzas Estratégicas	Aliado/Financiador	Pueden utilizar los resultados del modelo para identificar áreas de cooperación y financiación con mayor impacto en los ODS.	Si el modelo no refleja correctamente las necesidades, pueden financiar proyectos que no alineen correctamente con las necesidades reales.
Ministerios de Salud,	Implementadores	El modelo les ayuda a diseñar	Si el modelo tiene sesgos o errores,

Educación y Género		políticas públicas y estrategias más informadas basadas en datos reales, optimizando las intervenciones.	las políticas implementadas podrían estar basadas en información incorrecta, afectando su efectividad.

Sección 6. Trabajo en equipo:

Líder de proyecto y negocio: Nicolás Coca

Tareas:

- Entendimiento del negocio y enfoque analítico: 1.5 horas
- Creación e implementación del modelo que usa el algoritmo de KNN: 40 minutos.
- Creación e implementación del modelo que usa el algoritmo de Naive Bayes: 40 minutos.
- Creación mapa de actores: 1 hora.
- Elaboración del script del vídeo. 1.5 horas

Líder de datos y analítica: Nicolás Ruíz

Tareas:

- Entendimiento y procesamiento de datos de entrenamiento y prueba: 3 horas
- Creación del modelo de ML de Random Forest: 15 minutos.
- Cálculo de métricas de desempeño de todos los modelos creados y elección del mejor modelo: 1 hora.
- Obtención de los atributos más importantes para el mejor modelo: 1 hora
- Procesamiento de los datos de prueba y creación del archivo con los textos clasificados: 45 minutos.
- Creación del repositorio de GitHub y subida de archivos: 20 minutos
- Elaboraciones dispositivas del video: 30 minutos

Repartición de puntos:

Nicolás Ruiz: 50

Nicolás Coca: 50

Puntos para mejorar:

- Realizar más reuniones de planificación previo a la entrega.
- Utilizar más los canales de comunicación entre integrantes del grupo.
- Definir plazos para cada subtarea y cumplirlos para evitar la acumulación de trabajo.
- Utilizar más los canales de comunicación del curso como slack.