Inteligencia de Negocios

Entrega 1 del Proyecto

Grupo 28 - Sección 4

Logotipo, Icono

Descripción generada automáticamente

Santiago Gómez

Juan Andrés Carrasquilla

David Ernesto Zamora

Contenido

[1 Entendimiento del Negocio y Enfoque Analítico 2](#_Toc176617110)

[2 Entendimiento y Preparación de los Datos 4](#_Toc176617111)

[3 Modelado 7](#_Toc176617112)

[4 Resultados 9](#_Toc176617113)

[5 Mapa de Actores 9](#_Toc176617114)

[6 Trabajo en Equipo 10](#_Toc176617115)

# 1 Entendimiento del Negocio y Enfoque Analítico

Oportunidad/Problema de Negocio: UNFPA es el fondo de poblaciones de las naciones unidas y en este momento buscan realizar un proceso conjunto que ayude a identificar problemas y evaluar soluciones. Específicamente el proyecto a desarrollar necesita desarrollar un modelo que logre relacionar la información dada por diferentes ciudadanos con los ODSs 3 4 y 5 automáticamente.

Después de explorar los datos y ver lo que el UNFPA quiere, proponemos el siguiente enfoque analítico para resolver el problema

Objetivos:

* Desarrollar un modelo analítico que automatice la relación entre las opiniones de los ciudadanos y los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Hay que tener en cuenta que la automatización para que sea posible el uso de ese modelo por gente externa a nuestra organización (quienes vamos a desarrollar el modelo) hace falta la parte 2 de la entrega.
* Optimización del análisis de texto. Al desarrollar este proyecto, UNFPA puede lograr disminuir los tiempos y los costos relacionados con el análisis de texto. Eso es porque gran parte del trabajo lo hará el modelo desarrollado, disminuyendo la cantidad de horas de personal dedicado a eso.
* Aumento de la participación ciudadana.

Criterios de Éxito:

* Se debe obtener una reducción en el tiempo de clasificación de opiniones para los trabajadores como mínimo del 50% con relación al análisis manual. (Tener en cuenta que se hace hipotéticamente y no podemos medir realmente cuanto ahorrarían).
* El valor de precisión para cada clase para las clasificaciones generadas por el modelo sobre los datos de prueba debe ser mayores a 0.7.
* El valor del *recall* para cada clase sobre el conjunto de prueba debe ser mayor a 0.8. Esto significa que el modelo debe clasificar correctamente al menos el 80% de las opiniones reales que pertenecen a cada clase.
* El f1 score para cada clase debe ser al menos 0.75, para encontrar un balance entre precisión y *recall* en el modelo.
* Al facilitar el acceso a todos a esta herramienta se puede mejorar la participación ciudadana, obteniendo un 20% más de opiniones que en periodos anteriores o en estimados de números de opiniones sin esta herramienta.

Organización y sus roles internos que se benefician: La organización que se beneficiaría con este proyecto es la misma que lo está proponiendo, UNFPA.

Dentro del UNFPA se beneficiarían los roles:

* Analista de políticas públicas: Estas personas utilizarán los datos que salgan clasificados por el modelo para poder tomar decisiones referentes a políticas públicas para mejorar el desarrollo sostenible.
* Los investigadores podrán crear reportes con los resultados obtenidos por el modelo. De esta forma, esos reportes podrán tener un impacto social, y se podrán encontrar más fácilmente los problemas relacionados a cada ODS.

Impacto en Colombia:

* Políticas públicas: Al tener una herramienta que facilite el proceso de clasificación, se podrán responder de forma más ágil y eficiente con respecto a los problemas relacionados con esos 3 ODS.
* Impacto social: Al aumentar la participación ciudadana, es probable que más personas den su opinión y por tanto se sientan representados cuando se tomen decisiones y se adopten nuevas políticas. De esa forma, personas que antes no hubieran podido dar su opinión y encontrar una solución a su problema, ahora si tengan la posibilidad de expresarse y encontrar soluciones.

Enfoque Analítico:

* Tipo de aprendizaje: Supervisado.
* Tarea de aprendizaje: Clasificación multiclase.
* Técnicas y algoritmos a utilizar:
  + Preprocesamiento: tokenización, lematización, eliminación de stopwords, Bag of Words.
  + Para el modelado: Árbol de decisión, KNN, Random Forest y Redes Neuronales.
* Por último, se va a usar un análisis descriptivo y predictivo. El primero se usará porque se necesita entender las opiniones ciudadanas relacionadas con cada uno de los ODS y el segundo porque se construye un modelo de clasificación a partir de machine learning que permita relacionar nuevas opiniones con ODS 3, 4 y 5.

# 2 Entendimiento y Preparación de los Datos

Entendimiento

La columna SDG en el conjunto de datos es una variable categórica que se representa con números y que funciona como etiqueta. Esta será la columna que se pretende predecir a partir de opiniones. Dicho de otra manera, el modelo usara opiniones en texto para clasificarlas en las 3 clases que SDG representa.

Validez

Al ver la descripción de los datos, vemos que coincide con nuestros supuestos, que decían que cada valor de SDG corresponde a una categoría de ODS. Esto porque los valores se mueven entre 3 y 5 y no hay valores fuera de este rango. Eso significa que esta columna cumple con el atributo de calidad validez. Con respecto a la columna Textos\_espanol revisamos caracteres no válidos que pudieran estar en las opiniones, palabras no validas en español o palabras repetidas como “a a a” o “y y y”. Se encontró que había muchos problemas de interpretación de letras con tilde, por lo que se tuvieron que corregir estos errores para hacer una correcta interpretación de los datos. Con respecto a las palabras repetidas, se encontraron pocos casos y se corrigieron borrando las palabras duplicadas de cada opinión. También se revisó que no hubiera demasiado cortas, para esto se revisó la distribución de la cantidad de palabras de las opiniones y no se encontraron casos con opiniones muy cortas que no dieran información.

Gráfico, Histograma

Descripción generada automáticamente

*Fig 1. Distribución de la longitud de las opiniones.*

Imagen de la pantalla de un video juego

Descripción generada automáticamente con confianza media

*Fig 2. Descripción de la columna sdg del conjunto de datos.*

Completitud

Además, vemos que el número de valores en la columna de sdg es igual al número de filas en train.shape, lo que significa que no hay valores nulos en la etiqueta. Por esto podemos decir que la columna cumple con la completitud de los datos.

Consistencia

No hay problemas de consistencia, cada categoría se representa con un número. No hay problemas que se muestre, por ejemplo, una categoría con 3 y otra con 'tres'.

**Interfaz de usuario gráfica, Aplicación

Descripción generada automáticamente**

*Fig 3. Cuenta de registros por categoría.*

Unicidad

Ninguna opinión debería estar dos veces en el conjunto de datos, en este caso no se identificó ninguna tupla duplicada.

Texto

Descripción generada automáticamente

*Fig 4. Código relacionado a Unicidad de tuplas.*

Preparación de datos

Una vez realizado el entendimiento podemos determinar que los principales problemas a corregir son:

* La consistencia: puede pasar que palabras diferentes tengan un mismo significado. Para esto debemos lematizar.
* Validez: vimos que puede haber formas de escribir palabras que no son válidas, o puede haber palabras repetidas consecutivamente. Para esto intentaremos corregir los errores más comunes. Por ejemplo, las vocales con tildes la mala codificación las vuelve caracteres extraños.

Seguido de esto, se hizo la preparación para poder pasar los datos a los modelos. Específicamente se hizo tokenizado, lematizado, eliminación de stopwords y luego TF-IDF y N-grams. Estos dos últimos se harán por separado, lógicamente no se puede hacer los dos al tiempo, entonces en una variable dejaremos TD-IDF y en otra N-grams.

Tener tanto N-grams como TF-IDF nos permitió probar los modelos con ambas formas de representación para identificar cuál de las dos nos daba mejores resultados, pero eso se discute luego.

Revisión de balance de clases

Se revisó el balance de clases ya que podría afectar al modelo, sin embargo, encontramos que en este caso no es un problema porque las clases se encuentran bien balanceadas, por lo que no vimos necesario un *upsampling.*

Texto

Descripción generada automáticamente con confianza baja

*Fig 5. Balanceo de clases.*

# 3 Modelado

Se realizaron 4 modelos con distintos algoritmos. Se usaron: árbol de decisión, random forest, KNN y redes neuronales. Al final el objetivo es escoger el mejor modelo de los 4 para entregar al cliente.

Las métricas que se mostraran consisten en accuracy (porcentaje de predicciones correctas), precisión (de los que fueron clasificados como positivos, ¿cuántos realmente lo eran?), recall (cuantos de los que se predijeron como positivos realmente eran y f1 que es una media entre precisión y recall. Los macros son los promedios de estas métricas para cada clase.

Árbol de Decisión

Como hiper parámetros iniciales para este algoritmo usamos:

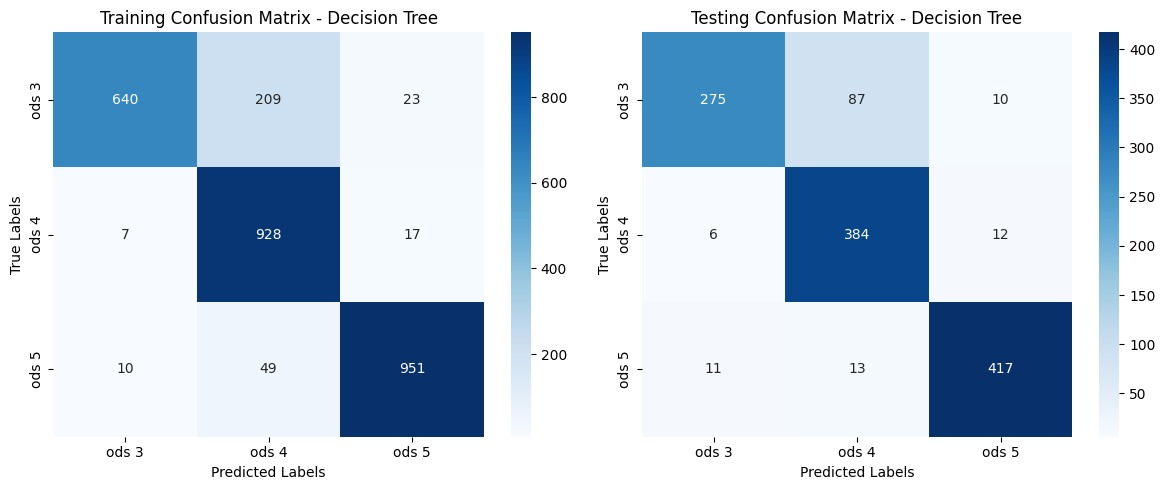
* Criterion: Gini
* Max\_depth: 5
* Random\_state: 42

Imagen de la pantalla de un celular con letras

Descripción generada automáticamente con confianza mediaCalendario

Descripción generada automáticamente

*Fig 6. Métricas de calidad DT.*

**

*Fig 7. Matriz de decisión DT.*

Después de realizó GridSearch para buscar los mejores hiper parámetros y ver si hay mejoras. Los hiper parámetros encontrados fueron:

* Criterion: Gini
* Max\_depth: None
* Min\_samples\_leaf: 1
* Min\_samples\_split: 5

Calendario

Descripción generada automáticamente

*Fig 8. Métricas de calidad DT con grid search.*

Como se ve en la figura 7, las métricas de calidad obtenidas al usar los hiper parámetros obtenidos con grid search fueron mejores. Macro avg f1 de 97% vs 88%.

Los principales determinantes para los ODS son los siguientes ODS 3: Menciona pocas veces mujer, nombra bastante salud, poco educación, poco niña, poco la palabra reproductiva Si no menciona mucho salud, los documentos que nombren bastante la palabra paciente o sanitario pueden coincidir con este ODS.

ODS 4: Documentos que mencionan poco la palabra mujer, poco la palabra salud, poco género, poco paciente, y poco sanitario.

ODS 5: Documentos que mencionen mucho mujer, poco enfermedad, poco mortalidad, poco consumo, poco embarazado. Si un documento no menciona mucho mujer y no menciona salud pero si género, poco la palabra equidad y estudiante entonces puede pertenecer al ODS 5.

Random Forest

Como hiper parámetros iniciales para este algoritmo usamos:

* N\_estimators: 100
* Random\_state: 42

Calendario

Descripción generada automáticamente

*Fig 9. Métricas de calidad RF.*

Después de realizó grid search para buscar los mejores hiper parámetros y ver si hay mejoras. Los hiper parámetros encontrados fueron:

* N\_estimators: 300
* min\_samples\_split: 2
* min\_samples\_leaf: 1
* max\_features: sqrt
* max\_depth: None

Calendario

Descripción generada automáticamente

*Fig 10. Métricas de calidad RF con grid search.*

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

*Fig 11. Matriz de confusión RF.*

Como se ve en la figura 9, las métricas de calidad obtenidas al usar los hiper parámetros obtenidos con grid search fueron levemente mejores. Macro avg f1 de 97% vs 96%.

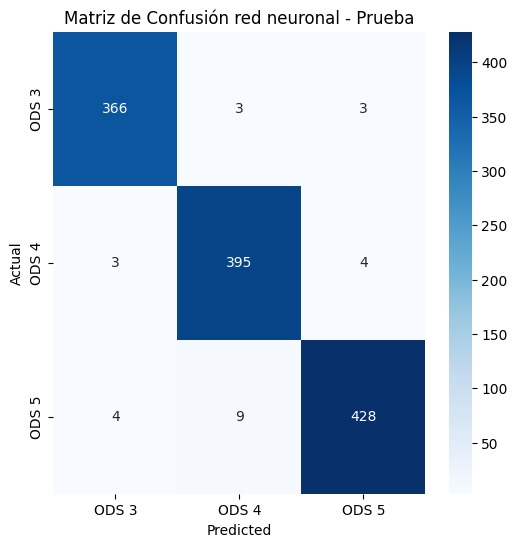
Red Neuronal Estándar

En la red neuronal estándar estamos usando los hiper parámetros de 3 capas de neuronas, la primera de 128, segunda de 64 y finalmente de 3. Cada una usa RELU como función de activación y la última capa tiene softmax. Esto se hace para permitir la clasificación. Esencialmente cada neurona hace una regresión lineal con los parámetros que llegan de la capa anterior y luego aplica una función no lineal. La última capa debe usar softmax porque asi, cada regresión de cada neurona de la última capa se transforma en un numero de 0 a 1, representando la probabilidad de que el dato pertenezca a una determinada clase. Son 3 neuronas en la última capa porque cada una tiene como output la probabilidad de que el dato corresponda a cada una de las 3 clases (ODS 3, 4 y 5). Además, como función de perdida se usó spase categorical crossentropy, que calcula el error entre las clases predichas por el modelo y las etiquetas reales. Finalmente, señalamos que usamos dropout entre cada capa con el 50% de probabilidad de que un input no llegue la siguiente neurona. Esto se hace para evitar el sobreajuste.

**Calendario

Descripción generada automáticamente**

*Fig 12. Métricas de calidad Red Neuronal Estándar.*



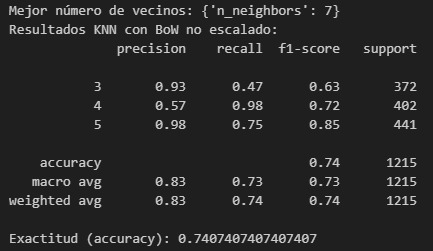
*Fig 13. Matriz de Confusión red neuronal.*

Como se puede ver en la figura 10, las métricas de calidad obtenidas son bastante buenas, sin embargo, se tiene la desventaja de que el modelo tarda más en entrenarse y ejecutarse. Se usaron 10 épocas de entrenamiento.

KNN

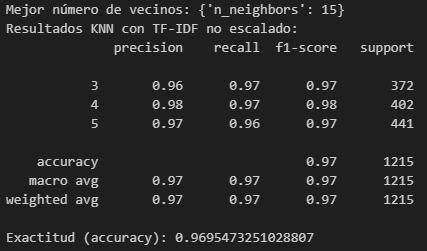
Para este algoritmo se hicieron varias iteraciones. Se probó tanto con TF-IDF como con BoW, escalado y no escalado, cada una de esas combinaciones con y sin grid search. A continuación, se evidencian las métricas de calidad para la iteración realizada con BoW no escalado y sin grid search. Como hiper parámetros iniciales para este algoritmo usamos:

* n\_vecinos: 7

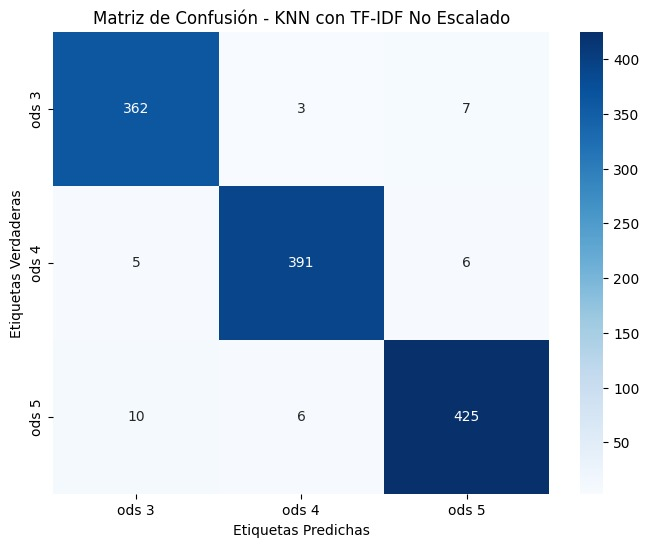


*Fig 14. Métricas de calidad KNN.*

Después de realizó grid search con para buscar los mejores hiper parámetros y ver si hay mejoras, lo que nos arrojó que el número de vecinos óptimos son 15. Es importante anotar que en este caso se usó TF-IDF.



*Fig 15. Métricas de calidad KNN con grid search.*

**

*Fig 16. Matriz de Confusión KNN.*

# 4 Resultados

Como se mostró en las secciones anteriores, en general los modelos tuvieron un muy buen desempeño pues todos tienen métricas de calidad bastante favorables, por ejemplo, todos presentan un F1-score y una precisión de mas de 0.95. Aún asi, la red neuronal estándar fue la que tuvo un mejor desempeño de las 4. Teniendo esto en cuenta se decidió que ese sería el modelo a utilizar en la solución para el cliente ya que es el modelo más confiable.

Por ahora se podría decir que se están cumpliendo con los objetivos definidos, ya que el F1-Score del modelo supera con buen margen el propuesto al inicio de este documento. En ese sentido podemos decir que el despliegue y uso a futuro que se le puede dar al modelo tiene el potencial de aportar mucho en el proceso de clasificación de opiniones y de esa forma ahorrar tiempo al UNFPA. Además, una solución como esta puede empoderar a las personas para poder dar opiniones sobre su situación actual y mejorar la respuesta de los entes gubernamentales de cara a los distintos tipos de problemas que puede enfrentar tanto el país como el mundo entero.

# 5 Mapa de Actores

**Tabla

Descripción generada automáticamente**

*Fig 17. Mapa de Actores*

# 6 Trabajo en Equipo

Como equipo dividimos el trabajo de acuerdo a los espacios libres de cada integrante, se fue trabajando en orden cada uno de los pasos del proyecto. Se realizaron dos reuniones principales, una de planeamiento y otra de seguimiento, el resto del tiempo se mantuvo comunicación activa a través del grupo de WhatsApp del proyecto para ir viendo resultados y avances de las actividades que cada uno de los integrantes iba desarrollando. Para la primera reunión todos teníamos el contexto del proyecto bastante claro y David empezó con el trabajo de Entendimiento del negocio y enfoque analítico. Así mismo, David desarrolló el Entendimiento y la preparación de los datos, teniendo en cuenta el feedback y las opiniones que se daban por el grupo para estos pasos por parte de los demás integrantes. En la reunión de planeamiento también se declaró que cada integrante del proyecto trabajaría en un modelo para la clasificación de textos. En este caso, Andrés trabajó en el algoritmo de Árbol de decisión, Santiago en el de KNN y David desarrolló dos modelos, uno con el algoritmo de Random Forest y otro con Redes neuronales estándar. En la reunión de seguimiento se plantearon las actividades finales como el desarrollo del documento y el video y otras cosas como el mapa de actores. En este caso, el documento quedó a cargo de Andrés y Santiago, el video a cargo de Andrés y el mapa de actores a cargo de Santiago. Así mismo, el encargado de generar el CSV final fue David. Los detalles de los algoritmos se estuvieron modificando hasta último momento por el feedback recibido para cada uno de los modelos por los demás integrantes del grupo. El tiempo distribuido a lo largo de las actividades fue relativamente parecido, aunque cabe destacar que David invirtió más tiempo en la primera fase, realizando el entendimiento y preparación de los datos.

En cuanto a los roles que asumimos como equipo, todos asumimos el rol de desarrolladores de modelos, pues todos trabajamos en alguno de los modelos para la clasificación de textos desarrollados por el equipo. Así mismo, todos nos aseguramos de tomar el rol de redacción, tanto David al principio del desarrollo del proyecto, cada uno por su cuenta en la explicación de los modelos y Andrés y Santiago con la creación del documento. Del mismo modo, todos tomamos el rol de gestión de las tareas y de verificar que todo estuviera en orden y completo para poder realizar la entrega.

En cuanto a los retos del proyecto fueron varios, sin embargo, algunos de los más importantes fueron el escoger qué modelo era el mejor y más representativo del trabajo que el grupo realizó, esto se solucionó con el análisis de las gráficas de las matrices de confusión, el análisis de las métricas de los modelos (precisión, recall, f1-score) y la votación del equipo. Otro de los problemas importantes fue la unificación de los archivos porque cada uno trabajó por aparte en sus modelos, sin embargo, esto se solucionó copiando y pegando todo en un solo archivo. También, hubo un problema en la realización del video, pues Andrés tuvo problemas con su PC.

En base a todo lo anterior, se plantea una distribución equitativa de los puntos para los integrantes del proyecto (33,33). También, se plantean distintos puntos de mejora para la siguiente entrega del proyecto, por ejemplo, el empezar la realización del mismo con más tiempo, con el objetivo de tener la presión de la entrega tan encima, así como realizar más reuniones para poder trabajar en conjunto con más facilidad.