PROYECTO ETAPA 1

Construcción de modelos de analítica de textos

ISIS3301

Juan Camilo Beltrán Garnica
Leidy Johana Lozano Florez
Juan Andrés Reyes

Índice

Sección 1. Entendimiento del negocio y enfoque analítico.

Como parte de la metodología ASUM-DM para procesos de analítica descriptiva, es necesario empezar con un proceso de entendimiento del negocio y de ahí, partir a un enfoque analítico. Esta información es presentada a continuación.

Entendimiento del negocio	
Oportunidad/problema Negocio	El Fondo de Poblaciones de las Naciones Unidas (UNFPA), en colaboración con entidades públicas y utilizando diversas herramientas de participación ciudadana, busca identificar problemas y evaluar soluciones vigentes, alineándose con los Objetivos de Desarrollo Sostenible (ODS) de las Naciones Unidas (ONU). No obstante, la clasificación de la información recopilada presenta dificultades, lo que complica determinar a qué ODS específicos corresponde cada entrada.
Objetivos y criterios de éxito desde el punto de vista del negocio.	Para la UNFPA, el objetivo principal es facilitar la tarea de clasificación de las entradas de texto que reciben conforme a los ODS. De esta forma, pueden ocupar menos tiempo en tareas logisticas y más tiempo evaluando soluciones.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La organización beneficiada es la UNFPA, ya que ella hará uso del producto de clasificación. Como equipo, teorizamos que el rol beneficiado por esta oportunidad es el ingeniero de datos. Consideramos que el proyecto podría facilitar una de sus tareas, que es el recibimiento y clasificación de la información. Con esta información ya etiquetada, el ingenierio podría, por ejemplo, centrar su atención en diseñar sistemas de análisis que permitan a la UNFPA obtener información valiosa de estos comentarios. (Coursera, 2023)
Impacto que puede tener en Colombia este proyecto	Para conocer el impacto que tiene el proyecto en colombia, es necesario resaltar que la UNFPA desea emplear el producto para discernir los ODS 3, 4 y 5, los cuales, son definidos por la ONU (2022) de la siguiente manera: • **3 – Salud y Bienestar**. Alcanzar la cobertura universal de salud y garantizar que todas las personas tengan acceso a medicamentos y vacunas seguros y a precios accesibles. • **4 – Educación de Calidad**. Educación primaria y secundaria gratuita, dando pie a una educación superior asequible que elimine las disparidades de género y riqueza. • **5 – Igualdad de género**. Eliminación del matrimonio infantil y las ausencias de protección juridica para las mujeres. Alcanzar la paridad en puestos de poder, liderazgo y representación. Con esta información, se evidencia por qué la UNFPA tiene enfoque particular en estos tres ODS. La UNFPA en Colombia desarrolló un enfoque especializado alrededor del bienestar sexual y reproductivo de las mujeres, jovenes y adolescentes. Garantizando así que todas reciban acceso a una vida productiva, saludable e informada, habilitando una dinámica poblacional libre de situaciones no deseadas o derechos reproductivos violados o ignorados, donde todas las jovenes puedan alcanzar su potencial (UNFPA, s.f.). De esta manera, es posible afirmar que el proyecto estaría generando un impacto positivo en relación a la salud sexual y reproductiva de las

	mujeres colombianas. Este impacto se vería manifestado en la forma como el proyecto apoya la labor de recolección y clasificación de información de la UNFPA. Al tener información clasificada y actualizada de los ODS 3,4 y 5, la UNFPA podría monitorear con facilidad los avances de estos ODS a nivel mundial, y tratar de replicar o aplicarlos en la sociedad Colombiana, contribuyendio así a su misión.

Enfoque analítico	El equipo consideró el siguiente enfoque analítico:
	• **Tipo de análisis – Predictivo**. Se va a determinar la variable objetivo, el SDG (ODS), con base a la variable texto.
	• **Tipo de aprendizaje – Supervisado.** Los datos de entrenamiento se enucentran completamente etiquetados.
	• **Tarea de aprendizaje – Clasificación.** Se desea clasificar nuevas entradas dado un entrenamiento previo.
	• **Algoritmos seleccionados** – Cada miembro del equipo seleccionó un algoritmo de clasificación distinto, con el propósito de evaluar cuál de ellos presenta mejores resultados. Los algoritmos seleccionados fueron:
	  ○ Árbol de decisión
	  ○ K-Nearest Neighbor (KNN)
	  ○ Bayes ingenuo

Sección 2. Entendimiento y preparación de los datos

Tras hacer una revisión al archivo de entrenamiento, se comprobó que cada fila del archivo hace referencia a un comentario o fragmento de texto hecho con relación a los ODS 3, 4 o 5, y su respectiva clasificación. En total, se encontraron un total de 4049 registros.

Análisis de calidad

- Unicidad – No se encuentran registros duplicados.
- Completitud – Todos los registros están completos.
- Consistencia – No hay columnas que puedan afectar la consistencia.
- Validez – Se eliminaron algunos registros que estaban en inglés y en francés.

Limpieza de los datos y tokenización

En primer lugar, se convirtió cada uno de los textos a Tokens, haciendo uso de la libería NLTK, tras haber tokenizado todo, se eliminaron registros nulos. Luego, se procedió a limpiar las entradas para convertirlas a texto plano, para lograr esto, se realizaron las siguientes tareas:

- Remoción de caracteres no ASCII – Esto incluye Emojis, caracteres especiales, etc.
- Conversión a minúsculas – Para evitar confusiones posteriores, se pasaron todas las palabras a minúscula.
- Remoción de *stopwords* – Se emplea la librería de NLTK para remover *stopwords.*
- Remoción de puntuación – Para evitar que sean contados como caracteres.

Normalización

En la normalización de los datos se realiza la eliminación de prefijos y sufijos, usando el Stemmer de NLTK Snowball. No fue posible aplicar una lematización.

Vectorización

Para la fase final de la preparación, se vectorizó la información obtenida de fases anteriores. Para realizar esta tarea, se empleó el enfoque *Bolsa de Palabras,* facilitado por la libería SKLearn.

	000	001	003	004	005	006	007	008	009	01	...	zogl	zoles	zoll	zomb	zon	zonmw	zuck	zukowski	zupanc	zusatzentgelt
0	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

5 rows × 11424 columns

Figura 1 - Resultado de la vectorización

Sección 3. Modelado y evaluación

Algoritmo 1 – Arbol de decisión [Juan Camilo Beltrán]

Se modeló un algoritmo de árbol de decisión, este es una estructura en forma de árbol donde cada nodo interno representa una "pregunta" o "condición" sobre una característica específica del conjunto de datos. Cada rama del nodo representa el resultado de la condición, y cada nodo hoja (es decir, un nodo terminal) representa una clase o una decisión final, en este caso, representa la categoría final a la que pertenece un comentario.

Preparación para un Árbol de decisión

A diferencia de otros algoritmos probados, un árbol de decisión no requiere una preparación especial de los datos además de la hecha en el preprocesamiento, ya que puede recibir datos numéricos continuos o discretos y Strings categóricos, por lo que en este caso no necesitamos realizar ninguna transformación.

Selección de hiperparámetros para Arboles de decisión.

El algoritmo de Arboles de decisión emplea dos hiperparámetros:

- Criterio – Refiere al criterio de división que se usara para tomar la decisión en cada nodo, la librería usada soporta 3 criterios, Entropy, Gini y Logloss.
- Profundidad Máxima – Refiere a la profundidad máxima a la que puede llegar cada rama del árbol.

La búsqueda de hiperparametros fue automatizada con la función *best_estimator,* la cual retornó los mejores hiperparametros Criterio = Gini y Profundidad Máxima = 8.

Resultados de las métricas de evaluación.

Para la evaluación, se dividió el conjunto de prueba en 80% de los datos para el entrenamiento, y el 20% para la evaluación. Se obtuvieron los siguientes resultados:

	precision	recall	f1-score	support
3	0.62	0.73	0.67	255
4	0.64	0.57	0.60	275
5	0.62	0.58	0.60	259
accuracy			0.63	789
macro avg	0.63	0.63	0.62	789
weighted avg	0.63	0.63	0.62	789

Figura 2 - Métricas de resultado para AD

Algoritmo 2 – K-Nearest Neighbor (KNN) [Leidy Lozano]

Preparación para KNN

El algoritmo de KNN es intolerante a variables nulas o variables no numéricas, por esta razón, se eliminaron 108 registros con valores vacíos. Adicionalmente, se estandarizó la información, ya que KNN es sensible a la variación de los datos. Esto se logró usando el *StandardScaler* de SKLearN.

Selección de hiperparámetros para KNN

KNN emplea dos hiperparámetros:

- K – Refiere al número de "vecinos" que van a ser tomados en cuenta.
- P – Refiere a la ecuación que va a calcular la distancia entre la entrada nueva y las existentes.

Es posible probar distintas combinaciones de K y de P para encontrar el mejor resultado; sin embargo, esta tarea puede ser automatizada usando la función de *best_estimator.*
Para el caso del análisis de las ODS, se obtuvo un valor de K = 9, y un P = 1, que refiere a la ecuación euclidiana de la distancia entre dos puntos.

Resultados de las métricas de evaluación.

Para la evaluación, se dividió el conjunto de prueba en 80% de los datos para el entrenamiento, y el 20% para la evaluación. Se obtuvieron las siguientes métricas.

	precision	recall	f1-score	support
3	0.91	0.85	0.88	253
4	0.80	0.84	0.82	263
5	0.82	0.83	0.82	252
accuracy			0.84	768
macro avg	0.84	0.84	0.84	768
weighted avg	0.84	0.84	0.84	768

Figura 3 - Métricas de evaluación KNN

Algoritmo 3 – Bayes Ingenuo [Andrés Reyes]

Bayes Ingenuo es un algoritmo de agrupamiento que se basa en el teorema de Bayes, que ayuda a encontrar la probabilidad de un evento dado que se cumple alguno otro. Este algoritmo es llamado ingenuo porque hace algunas suposiciones acerca del problema. Se asume que todos los predictores en el modelo son condicionalmente independientes y que todos los *features* del modelo contribuyen de igual manera al resultado.

En la librería de *sklearn* hay 3 tipos de algoritmos Ingenuos de Bayes: Bernoulli, Multinomial y Gaussiano. Bernoulli sirve para *features* discretos y binarios o *booleanos.* Multinomial es similar, pero trata con datos discretos sin importar si son binarios o no. Por último, la variante Gaussiana se usa

para *features* con valores continuos. De estos 3, el mejor para realizar análisis de texto para posteriormente agrupar es el Multinomial. Un ejemplo de su uso es predecir la calificación de una película según las palabras en sus reseñas.

Preparación para Bayes Ingenuo

Un cambio que se debe realizar con los datos tras el proceso de vectorización es el de remover los datos nulos o no numéricos que se encuentren en el *dataframe*. Tras quitarlos, ya se está listo para realizar el análisis con Bayes Ingenuo.

Resultados de las métricas de evaluación

En esta evaluación, se usaron 75% de datos para entrenamiento y el 25% restante para las pruebas del modelo.

	precision	recall	f1-score	support
3	0.84	0.81	0.83	315
4	0.80	0.79	0.79	324
5	0.75	0.79	0.77	321
accuracy			0.80	960
macro avg	0.80	0.80	0.80	960
weighted avg	0.80	0.80	0.80	960

Figura 4 - Métricas de evaluación NB

Sección 4. Resultados.

Tras observar los tres resultados consolidados de los modelos, se concluyó que el mejor algoritmo para la tarea de clasificación asignada era el algoritmo KNN, seguido de *Naive Bayes.* Los resultados se comparan de forma más clara en la siguiente gráfica.

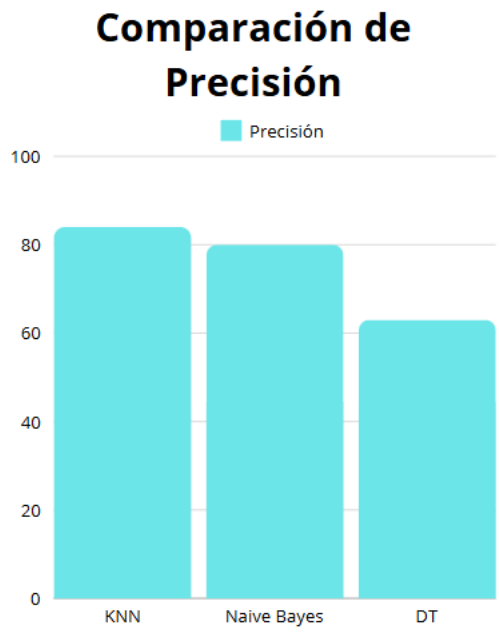

Figura 5 - Comparación de precisión

No obstante; no es posible determinar la calidad de un modelo únicamente por su precisión, veamos cómo se desempeña cada modelo individualmente en la clasificación de cada uno de los ODS.

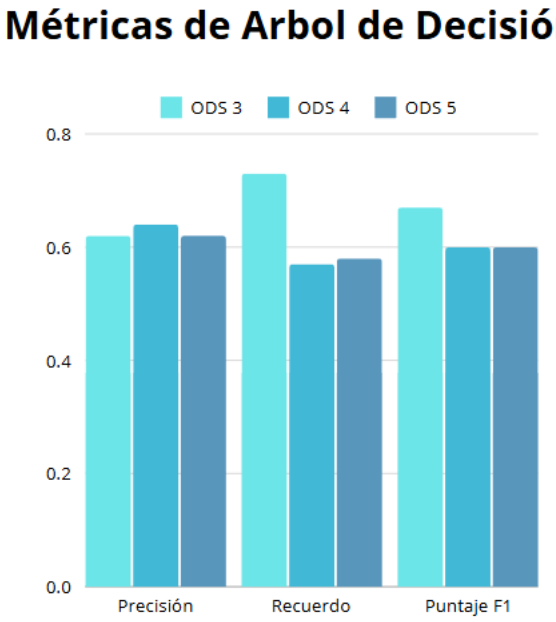

Figura 6 - Gráfica de barras AD

Figura 7 - Gráfica de barras KNN

De esta forma, es posible comprobar con mayor facilidad como KNN obtiene resultados superiores. Para observar los resultados de etiquetado del algoritmo, dirigirse al repositorio de Github..

Algunas estrategias que la UNFPA podría aplicar son.

1. **Monitoreo y Evaluación**: Analizar informes, proyectos y comunicaciones internas y externas. Esto ayudará a evaluar el impacto y alineación de las actividades con los ODS 3 (Salud y Bienestar), 4 (Educación de Calidad) y 5 (Igualdad de Género).
2. **Informes y Transparencia**: Generar informes periódicos que muestren cómo las actividades de la organización contribuyen a los ODS. Esto no solo mejora la transparencia, sino que también puede atraer a donantes y socios interesados en estos objetivos.
3. **Optimización de Recursos**: Identificar áreas que necesitan más atención o recursos. Por ejemplo, si el modelo muestra que hay menos actividades relacionadas con el ODS 5, la organización puede enfocar más recursos en iniciativas de igualdad de género.
4. **Comunicación y Sensibilización**: Utilizar los resultados del modelo para crear campañas de sensibilización y comunicación. Esto puede incluir publicaciones en redes sociales,

boletines informativos y presentaciones que destaquen el compromiso de la organización con los ODS.

5. **Colaboración y Alianzas**: Compartir los hallazgos con otras organizaciones y agencias que trabajan en los mismos ODS. Esto puede fomentar colaboraciones y alianzas estratégicas para abordar desafíos comunes.
6. **Capacitación y Desarrollo**: Capacitar al personal en el uso del modelo y en la importancia de los ODS. Esto asegurará que todos en la organización comprendan cómo sus roles contribuyen a los objetivos globales.
7. **Innovación y Mejora Continua**: Utilizar el modelo para identificar tendencias y oportunidades de innovación. Esto puede incluir el desarrollo de nuevos proyectos o la mejora de los existentes para alinearse mejor con los ODS.

Figura 8- Gráfica de barras NB

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

Se identificaron los siguientes actores dentro de la UNFPA.

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Coordinador/a de respuesta humanitaria.	Usuario-Cliente	La clasificación eficiente de nueva información apoyará su labor de gestión y supervisión de proyectos relacionados con la respuesta humanitaria.	Si el modelo no tiene buen desempeño, puede que se pierda información relevante a algún proyecto.
Consultoría para el apoyo de políticas a favor de la educación sexual/reproductiva.	Usuario-cliente	La información ayudaría a generar decisiones y propuestas informadas con respecto a las nuevas políticas de educación sexual.	Si el modelo falla, el consultor podría ver su tarea de levantamiento de información retrasada al encontrar datos no relevantes al ODS 3 o 5.
Profesional especializado en la	Usuario-cliente	Información actualizada sobre el	Si el modelo falla, el especialista podría

prevención combinada de VIH.		ODS3 ayudaría al especialista a promover estrategias de comunicación y demanda de derechos más actualizadas y asertivas.	encontrar ruido o incluir información no relevante en su proceso de investigación.
Profesional de enlace para proyecto de parteras local.	Usuario-Cliente	Tener información al día le ayudará a validar la eficacia y la relevancia de los avances que se estén realizando en su proyecto.	Si el modelo falla, puede que el profesional no tenga forma de corroborar información relevante a su proyecto.

La información sobre los puestos laborales de la UNFPA fue extraída, en parte, del siguiente enlace.

Sección 6. Trabajo en equipo

Consolidación del trabajo en equipo						
Nombre integrante y rol	Tareas Asignadas	Tiempo dedicado	Algoritmo Trabajado	Retos encontrados	Solución Propuesta	Puntos Asignados
Juan Camilo Beltrán – Líder de datos	-Realizar la preparación de los datos. -Realizar el pipeline de resultado. -Implementar su algoritmo. -Apoyar diligenciamiento del documento.	9-11 horas	-Arboles de decisión.	-Errores en la preparación de datos -Uso de librerías externas para tareas como Stemming o detectar el lenguaje. -No existe una librería de lematización en español, por lo que realizar este paso fue imposible.	-Para los errores en la preparación de datos, se usaron diferentes técnicas de vectorización de los comentarios hasta obtener los resultados deseados. -Para detectar el idioma se usó la librería langdetect y para el stemming se usó una librería de Stemming en español.	34
Andrés Reyes – Líder de negocio	-Asistir en la preparación de los datos. -Realizar la presentación. - Implementar su algoritmo. -Apoyar diligenciamiento del documento.	9 horas	Bayes Ingenuo – Multinomial (NB-multinomial)	-Se eligió usar un algoritmo basado en Bayes, pero se encontraron varias implementaciones -Se hallaron 3 variantes del algoritmo Bayes Ingenuo -Algunos errores con el manejo de	-Buscar cuál algoritmo es más usado en tareas similares a la de esta clasificación. Ahí se encontró el algoritmo Bayes Ingenuo (Naive Bayes)	33

	-Revisión de ortografía en el documento			archivo .xlsx a .csv	-Indagar acerca de la funcionalidad de cada variante de NB. Se encontró que la variante multinomial es la que se usa para clasificación de textos. -Rectificar que se estuviera cambiando de tipo con *encoding* utf-8	
Leidy Lozano – Líder del grupo	-Asistir en el entendimiento de los datos. - Diligenciar el documento. - Crear los entregables y el repositorio. - Grabar el video. -Implementar su algoritmo. -Realizar la entrega.	10 Horas	KNN	-El archivo de resultado de la fase de entendimiento contaba con valores nulos extraños que no eran detectados por Pandas. - Aprendizaje del manejo de SKLearn.	-Se usaron métodos de pandas para encontrar los valores nulos y se eliminaron los registros nulos manualmente desde Excel. - Se apoyó el proceso con prácticas de clase y videos.	33

Referencias

What are naïve Bayes classifiers? | IBM. (n.d.). Obtenido de IBM: https://www.ibm.com/topics/naive-bayes

sklearn.naive_bayes. (n.d.). Scikit-learn. Obtenido de SciKit Learn https://scikit-learn.org/stable/api/sklearn.naive_bayes.html

Fondo de Poblaciones de las Naciones Unidas. (s.f.). *UNFPA en Colombia*. Obtenido de UNFPA Colombia: https://colombia.unfpa.org/es/unfpa-en-colombia

Coursera. (15 de Junio de 2023). *¿Qué es un ingeniero de datos? Una guía para esta carrera tan demandada*. Obtenido de Coursera : https://www.coursera.org/mx/articles/what-does-a-data-engineer-do-and-how-do-i-become-one

Organización de Las Naciones Unidas. (22 de Mayo de 2022). *Onjetivos de desarrollo Sostenible*. Obtenido de United Nations: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/