Inteligencia de Negocios – ISIS 3301
Departamento de Ingeniería de Sistemas y Computación

Entrega 1: Construcción de modelos analíticos para clasificación de textos

Grupo 37

Antonin Bouillaud – 202325830

Ernesto José Duarte Mantilla – 202014279

Lina María Gómez Mesa – 201923531

Universidad de Los Andes

Octubre 15 2023

Contenido

Introducción

En el presente trabajo se llevará a cabo una propuesta de clasificación textual de acuerdo con los objetivos 3, 4 y 5 del Desarrollo Sostenible (ODS) como colaboración entre la Universidad de los Andes y la UNFPA. Este texto tiene como objetivo, en primera instancia, explicar la elaboración un análisis automatizado de opiniones que representan la voz de los habitantes locales sobre problemáticas de su entorno particular mediante el uso de un modelo de clasificación de textos supervisado. Para ello, se llevará a cabo un entendimiento preliminar de los datos recopilados, teniendo en cuenta características generales de los datos como idioma, unicidad, duplicidad y completitud en relación con los objetivos y criterios de éxito desde el punto de vista del negocio. A continuación, se realizará la limpieza de los datos teniendo en cuenta el análisis exploratorio inicial para luego presentar los modelos seleccionados (Random Forest, KNN, Multinomial Naive Bayes y Ridge Classifier). Posteriormente, se presentarán los resultados cuantitativos obtenidos y las recomendaciones preliminares para la organización. Finalmente, se describirá el mejor modelo y conclusiones. Todos estos pasos tienen como finalidad determinar oportunidades de éxito para el negocio con el fin de automatizar el proceso de clasificación de textos de acuerdo con los objetivos de desarrollo sostenible seleccionados.

1. Entendimiento del Negocio Y Enfoque Analítico

Los objetivos de desarrollo sostenible fueron adoptados por las Naciones Unidas en el 2015 como un llamado universal para erradicar la pobreza, proteger al planeta y que para en el 2030 se disfrute de mayor prosperidad a nivel mundial [4]. En total son 17 metas que se centran en un área particular para alcanzar tres dimensiones del desarrollo: ámbito económico, social y ambiental. En particular, los objetivos 3,4 y 5 se centran en [5]:

- Salud Y Bienestar (3): Garantizar una vida sana y promover el bienestar de todos a todas las edades. Intenta combatir las desigualdades en acceso a asistencia sanitaria.
- Educación de Calidad (4): Garantizar una educación inclusiva y equitativa de calidad y promover oportunidades de aprendizaje permanentes para todos. Intenta fomentar la tolerancia entre personas y contribuir a sociedades más pacíficas.
- Igualdad de Género (5): Lograr igualdad de género y empoderar a todas las mujeres y las niñas. Intenta ir en contra los prejuicios y las asociaciones implícitas a menudo invisible para la igualdad de oportunidades.

En Colombia, estos objetivos son de suma importancia dado que ayudan a la erradicación de la pobreza, reducción de la mortalidad infantil y mejoramiento de la cobertura en educación [6]. Además, sirven como indicadores de medición del progreso del país hacia el desarrollo y como herramienta para la integración y la coherencia entre las agendas y las políticas públicas. El objetivo de salud (3) se centra especialmente en reducir la mortalidad materna a 32 por cada 100 mil nacidos vivos, acabar con las muertes prevenibles de menores de 5 años, luchar contra enfermedades transmisibles (malaria, VIH, enfermedades tropicales) entre otras. La educación se centra en aumentar la cobertura de educación superior pasando de 49,4% en 2015 a 80,0% en 2030, educación básica y media gratuita, aumentar el número de personas con habilidades relevantes para el éxito financiero, entre otros. Finalmente, el objetivo de igualdad de género se centra en que las mujeres ocuparán el 50% los cargos decisorios dentro del Estado Colombiano, acceso universal a los derechos y salud reproductiva y hacer cumplir la legislación que promueve la igualdad de género.

A. Definición de los objetivos y criterios de éxito desde el punto de vista del negocio

En primer lugar, es importante recalcar que la organización a la que va enfocada el proyecto es una subunidad de las Naciones Unidas que busca alcanzar el cumplimiento de los ODS: la UNFPA. Esto lo logra gracias a la participación ciudadana quienes son quienes brindan los datos a través de distintas propuestas como ECHO que buscan escuchar la voz de las personas y traducirla al lenguaje de los ODS [8]. Por tanto, el objetivo principal de este proyecto es clasificar automáticamente el texto en las categorías de los ODS 3, 4 y 5 para identificar problemas y evaluar soluciones en la información textual recopilada.

Algunos otros objetivos de negocio son:

- Mejorar la eficiencia en el procesamiento de información: Debido a que es un proceso automatizado la herramienta clasifica de manera más consistente y de forma rápida, las opiniones, necesidades y perspectivas de las personas en torno a sus territorios frente a los ODS. Lo que reduce gastos operativos en la organización en expertos que clasifiquen las opiniones ciudadanas.
- Aumentar la satisfacción de la ciudadanía: Tener una herramienta automatizada reduce el tiempo de espera para que las voces de los ciudadanos sean escuchadas y transformadas en propuestas de políticas públicas que sean aplicadas como ordenamiento territorial.

Los criterios de éxito para la clasificación automática de textos de manera supervisada son:

- **Precisión:** La precisión brinda la proporción de textos clasificados correctamente, por lo que permite saber del número de textos bien predichos, qué porcentaje es verdaderamente positivo de acuerdo con los *labels*. Una alta precisión minimiza los falsos positivos. Para cada clase C_i esta se calculará de la forma:

$$PPV(C_i) = \frac{TP(C_i)}{TP(C_i) + FP(C_i)}$$

donde TP son los verdaderos positivos y FP los falsos positivos.

- ***Recall*:** El *recall* brinda la proporción de anotaciones bien identificadas, lo cual nos permitirá saber sobre el total de positivos identificados en la matriz, qué porcentaje se clasifica correctamente para cada clase. Un alto *recall* minimiza los falsos negativos. Para cada clase esta se calculará de la forma:

$$TPR(C_i) = \frac{TP(C_i)}{TP(C_i) + FN(C_i)}$$

donde TP son los verdaderos positivos y FN los falsos negativos.

- **F-Medida (F_1):** La medida F1 es un promedio armónico de la precisión y la cobertura que privilegia un buen balance entre ambas medidas. Para cada clase C_i se calcula:

$$F_1(C_i) = 2\,\frac{TPR(C_i) \cdot PPV(C_i)}{TPR(C_i) + PPV(C_i)}$$

- **Accuracy:** La proporción de predicciones correctamente clasificadas sobre el total de predicciones.

$$\text{Accuracy} = \frac{TP + \text{TN}}{TP + FP + TN + FN}$$

B. Descripción del enfoque analítico

Uno de los enfoques analíticos que puede ayudar a alcanzar los objetivos del negocio es implementar un sistema de clasificación de textos que permita clasificar los textos de los habitantes locales de acuerdo con los objetivos de desarrollo sostenible 3,4 y 5 de forma automática. Esto se realizará utilizando técnicas de procesamiento de lenguaje natural y aprendizaje automático para entrenar el modelo con un conjunto de datos etiquetados. Una vez implementado el modelo, se utilizarán las métricas de evaluación detalladas en la anterior sección para medir el rendimiento del modelo y ajustarlo en caso de ser necesario.

C. Objetivos del Negocio

Tabla 1. Principales Objetivos del Negocio

Oportunidad/problema Negocio	La UNFPA busca clasificar automáticamente el texto en las categorías de los ODS 3, 4 y 5 para identificar problemas y evaluar soluciones en la información textual recopilada, reduciendo así el esfuerzo y el uso de recursos en la organización.
Enfoque analítico (Descripción del requerimiento desde el punto de vista de aprendizaje automático) e incluya las técnicas y algoritmos que propone utilizar	Implementar un modelo de clasificación de textos utilizando técnicas de procesamiento de lenguaje natural y aprendizaje supervisado para entrenar y evaluar el modelo con un conjunto de datos de textos etiquetados con los ODS 3, 4 y 5.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	La UNFPA es la organización que se beneficia del proyecto ya que es la encargada de “identificar problemas y evaluar soluciones actuales, relacionando la información con los diferentes ODS”. Asimismo, dentro de la organización, los roles que se benefician de este proyecto son los (usuarios) que son participantes internos de la organización como los directores ejecutivos y coordinadores nacionales y asesores de asuntos políticos e intergubernamentales quienes son los encargados de plantear las políticas públicas y planes de mitigación para llegar a los ODS en el 2030 [1,2,3]. Los ciudadanos (habitantes locales) también se beneficiarían de manera indirecta ya que se identificarían de manera más rápida y eficaz las problemáticas de su entorno.
Contacto con experto externo al proyecto	Juan José Suárez jj.suarezm1@uniandes.edu.co Canal Seleccionado: Correo y reunión presencial

2. Entendimiento y preparación de los datos

A. Perfilamiento de los Datos

En primer lugar, se observa que con los datos suministrados de los ODS 3,4,5 el *DataFrame* contiene 3000 filas y 2 columnas: *Textos_espanol* y *sdg*. Existen en total 3 *labels* en los que se puede clasificar un texto: 3, 4, 5; dónde cada número es el objetivo del desarrollo sostenible con el que se encuentra asociado el texto. Siendo estos salud y bienestar, educación de calidad e igualdad de género, respectivamente. A su vez, se tiene que para el *Dataset* entregado el 33% de las *reviews* están asociadas con el ODS 3, el 33% con el ODS 4 y el otro 33% con el ODS 5. Es decir, el *dataset* se encuentra balanceado. Por otro lado, se observa que los textos se encuentran en su mayoría, un 99,8%, en español, y el 0,2% restante corresponde a inglés y francés. Por último, se observa que no existen valores duplicados ni nulos en todo el *dataset*. Finalmente, vale la pena resaltar que se decidió dividir el *dataset* 80% train y 20% *test* con el fin de poder evaluar los modelos seleccionados.

B. Preparación de los Datos

Inicialmente se realizaron tres tipos de modificaciones a los datos de entrada: 1) limpieza de los datos, 2) tokenización y 3) normalización. Dentro de la limpieza de los datos se consideró que una posible modificación era eliminar las columnas que no se encontraban en español; es decir, inglés y francés. No obstante, luego de comparar resultados obtenidos entre los modelos con y sin columnas no se evidenció una diferencia significativa entre las métricas (f_1, precisión y recall) por lo que no se realizó.

En la segunda etapa, se eliminaron los símbolos de puntuación o comillas para evitar que este tipo de caracteres interfirieran en la interpretación de palabras por los algoritmos. Luego, se eliminaron los caracteres que no fueran tipo ASCII, se pasaron todos los caracteres a minúscula y se pasaron todos los dígitos o números en su representación de texto en español. Posteriormente, se eliminaron las palabras tipo *stopwords* que no aportan significado al problema de clasificación (a, el, lo... etc).

Finalmente, en la tercera etapa en la normalización de los datos, se realiza la eliminación de prefijos y sufijos, además de realizar una lematización. Esto da como resultado los datos limpios listos para ser convertidos a valores numéricos por uno de los siguientes dos métodos que se describen a continuación:

- **Count Vectorizer:** Es una técnica basada en bolsa de palabras en la que se representa el texto como un conjunto de palabras. En este modelo no importa la información relativa a la posición de los tokens ni su contexto, solo su frecuencia. Esta técnica fue seleccionada dado que puede ayudar a verificar la presencia o ausencia de una palabra en el texto y puede ayudar a identificar las palabras que más aportan o más caracterizan a cada uno de los ODS.
- **TF-IDF:** Esta es una técnica de procesamiento de lenguaje natural en la que se calcula la importancia de cada uno de los términos en un texto analizando la frecuencia con la que aparecen en el texto. Este considera tanto la frecuencia de una palabra en un documento como la frecuencia de la misma palabra en todo el texto, lo que ayuda a reducir la importancia de palabras comunes. Se eligió esta técnica ya que permite que el modelo sea sensible a las palabras que son más importantes para cada clase.

3. Modelado Y Evaluación

A. Modelo Random Forest - Antonin Bouillaud

Un paso crítico en la clasificación de texto es la representación de los documentos en forma de vectores. Aquí, se exploran dos enfoques diferentes: CountVectorizer y TfidfVectorizer. CountVectorizer crea vectores de palabras basados en la frecuencia de las palabras en los documentos. En contraste, TfidfVectorizer combina la frecuencia de las palabras en los documentos con la importancia relativa de las palabras en el corpus. Ambos enfoques tienen sus ventajas y se evalúan en el contexto de la tarea de clasificación de texto. Después se aplica el algoritmo *Random Forest* a las matrices generadas por CountVectorizer y TfidfVectorizer para entrenar modelos de clasificación de texto.

Justificación: Random Forest es un algoritmo de aprendizaje automático que construye múltiples árboles de decisión a partir de muestras aleatorias de los datos de entrenamiento. Luego, combina las predicciones de estos árboles mediante votación en clasificación o promediación en regresión. Este enfoque, conocido como Bagging, reduce la varianza y el sobreajuste, mejorando la generalización. Random Forest también utiliza la selección aleatoria de características en cada división de un árbol, lo que agrega aleatoriedad y diversidad al conjunto de árboles. Esto lo convierte en una herramienta efectiva para tareas de clasificación y regresión, especialmente en conjuntos de datos con alta dimensionalidad y muchas características como es el caso de este *dataset*.

Sin embargo, la elección de hiperparámetros óptimos es esencial para el rendimiento del modelo. Por lo tanto, se utiliza la búsqueda en cuadrícula (GridSearchCV) para encontrar la mejor combinación de hiperparámetros, como el número de árboles en el bosque y la profundidad máxima de los árboles. Esta optimización se realiza con el objetivo de maximizar la exactitud (accuracy) del modelo.

Los modelos entrenados se evalúan con un conjunto de datos de prueba. Se calcula el informe de clasificación, que incluye métricas clave como la precisión, el recall y el puntaje F1. Dado que las clases tienen aproximadamente la misma proporción en los datos, la exactitud (accuracy) es una métrica adecuada para evaluar cuántas predicciones correctas hace el modelo en general. Esto es importante para tener una visión general de su rendimiento en la clasificación de todas las clases sin sesgos hacia ninguna de ellas. A continuación, se muestran las métricas obtenidas en la tabla 2.

Tabla 2. Métricas Asociadas a RF (Average Macro y Weighted)

	Precisión (↑)	***Recall* (↑)**	***Accuracy* (↑)**	F_1 **(↑)**
CountVectorizer	0.97	0.97	0.97	0.97
TF-IDF	0.966	0.963	0.97	0.966

B. Modelo KNN – Ernesto Duarte

El algoritmo k-Nearest Neighbors (KNN) es un método de aprendizaje supervisado que se utiliza para la clasificación y regresión. En el contexto de la clasificación de textos, KNN opera considerando un texto como un punto en un espacio multidimensional. Cuando se necesita clasificar un nuevo texto, KNN identifica los ‘k’ textos en el conjunto de entrenamiento que son más cercanos al texto en cuestión, según alguna métrica de distancia (como la distancia euclidiana). La clasificación del nuevo texto se determina por la mayoría, basada en las categorías de estos “k” textos más cercanos. Esencialmente, se predice que el texto pertenece a la categoría que es más común entre sus vecinos más cercanos.

Justificación: La elección de este algoritmo radica en su simplicidad y su capacidad para manejar eficazmente grandes conjuntos de datos. Además, como no asume ninguna distribución subyacente para los datos, es especialmente útil cuando la relación de las características es compleja o difícil de comprender mediante modelos paramétricos.

i. Reducción de dimensionalidad

Como se mostró anteriormente, los datos fueron codificados bajo dos métodos: Count Vectorizer y TF - IDF Vectorizer. Estos métodos convierten el texto plano como parte del proceso de NLP en valores numéricos para luego ser procesados por cada uno de los algoritmos. Estas codificaciones arrojan como resultados matrices dispersas y debido a la naturaleza del algoritmo KNN que se basa en distancia numéricas, es más conveniente tener una matriz densa, por lo que se procedió a realizar un procedimiento de reducción de dimensionalidad con LSA (*Latent Semantic Analysis*).

Inicialmente se redujo la dimensionalidad a 100 componentes, para posteriormente, por medio del método del codo, verificar el número de componentes ideal para cada codificación. Los resultados obtenidos son los siguientes:

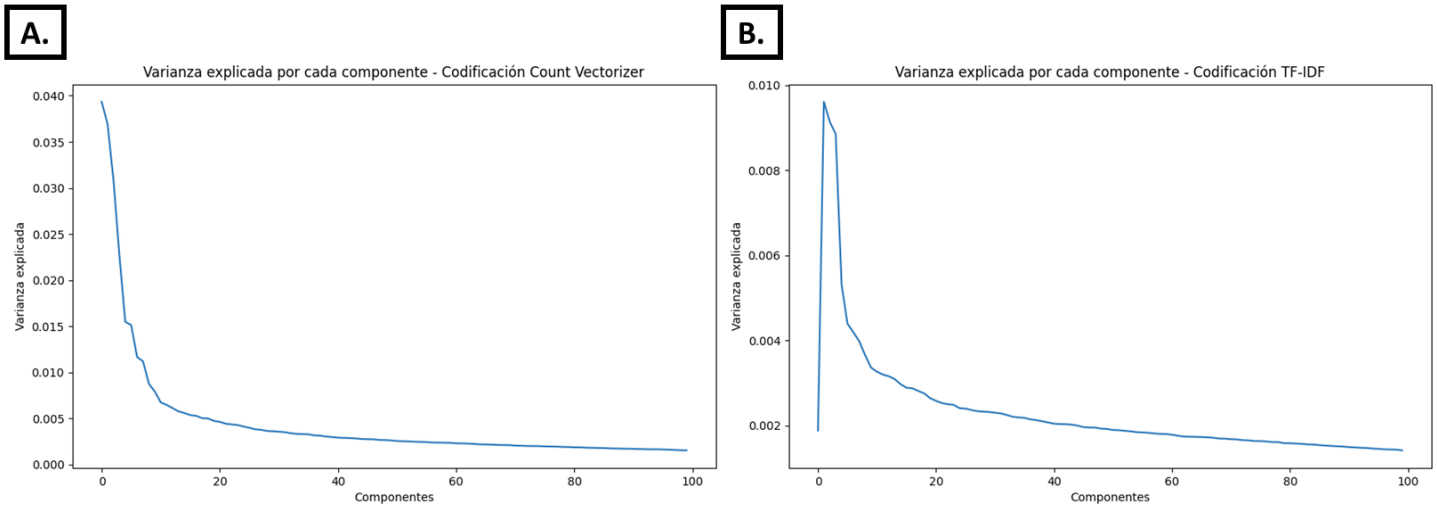

Figura 1. Varianza Explicada vs Número de componentes para el modelo KNN para A) CountVectorizer y B) TD-IDF

Como se aprecia en la gráfica 1A, el número óptimo de componentes para esta codificación sería 20. A partir de este valor, la varianza explicada no muestra un incremento considerable. Por otro lado, en la gráfica 1B es evidente que alrededor de 10 es el número de componentes óptimo para esta codificación. Después de este punto, el aumento en la varianza explicada es marginal. Por lo tanto, se usará 12 como número de componentes. Luego, de este proceso de reducción de dimensionalidad para cada tipo de codificación, se procede a realizar el algoritmo KNN. A continuación, en la tabla 3 se presentan los resultados cuantitativos tanto con CountVectorizer y TF-IDF.

Tabla 3. Métricas Asociadas al KNN (Average Macro y Weighted)

	Precisión (↑)	***Recall* (↑)**	***Accuracy* (↑)**	F_1 **(↑)**
CountVectorizer	0.923	0.916	0.916	0.917
TF-IDF	**0.965***	**0.965***	**0.965***	**0.965***

Para validar estos resultados, se empleó la función cross_val_score() de scikit-learn. Los resultados arrojaron una *accuracy* de 91% con una variación de +/- 0.02 para Count Vectorizer y de 96% con una

fluctuación de +/- 0.02 para TF-IDF. Es decir, los datos presentan una consistencia razonable en las diferentes particiones evaluadas.

C. Modelo Multimodal Naive Bayes – Ernesto Duarte

El algoritmo Multinomial Naive Bayes (MNB) es un método de clasificación probabilística basado en el teorema de Bayes, adapatado específicamente para características discretas, como la frecuencia de palabras en textos. En el contexto de la clasificación de textos, MNB estima la probabilidad de una palabra dada su categoría y, utilizando el teorema de Bayes, calcula la probabilidad posterior de una categoría dada un texto. Se asume que cada característica (por ejemplo, una palabra) es independiente de las demás, de ahí el término "*naive*" o "ingenuo". A pesar de esta simplificación, MNB ha demostrado ser particularmente efectivo en tareas de clasificación de textos, especialmente cuando se trata de grandes conjuntos de datos.

Justificación: La razón para optar por el algoritmo *Multinomial Naive Bayes* en clasificación de textos, radica en su eficiencia y robustez, especialmente cuando los datos son dispersos, como suele ser el caso de representaciones de texto como bolsas de palabras o matrices de términos-documento. Además, es capaz de manejar un gran número de características y se adapta rápidamente a nuevos datos, lo que lo hace adecuado para entornos dinámicos.

Para la ejecución de este algoritmo de clasificación, se entrenó de maneras separadas con los datos codificados con Count Vectorizer y TD – IDF. A continuación, se presentan los resultados para cada uno de los casos mencionados en la Tabla 4.

Tabla 4. Métricas Asociadas a Naive Bayes Multinomial (Average Macro y Weighted)

	Precisión (↑)	***Recall* (↑)**	***Accuracy* (↑)**	F_1 **(↑)**
CountVectorizer	**0.968***	**0.968***	**0.968***	**0.968***
TF-IDF	0.951	0.951	0.951	0.951

Es importante resaltar, que los primeros resultados que se muestran para cada modelo por codificación (Accuracy, Precision, Recall, y F1) son métricas generales para el modelo. Están calculados con un promedio macro, lo que significa que cada clase es tratada con igual importancia, sin tener en cuenta el desbalance entre clases.

D. Ridge Classifier – Lina Gómez

Por último, se decidió utilizar un modelo de clasificación lineal como *Ridge Classifier*. Este se caracteriza por ser un modelo lineal discriminativo que penaliza los coeficientes del modelo para evitar el sobreajuste a los datos. Por lo tanto, agrega un término de penalización a la función de costo que usualmente es la suma de los coeficientes al cuadrado de los *features.* Una penalización mayor resulta en una mayor regularización y en valores de coeficientes más pequeños lo cual funciona cuando hay pocos datos como es este caso. Este clasificador tiene un enfoque one-vs- all para problemas de multi-clasificación.

Justificación: Se eligió este clasificador, en particular, ya que como se discutió evita el sobreajuste, penaliza los *features* irrelevantes asignándoles un menor coeficiente y es un método inherentemente interpretable; por lo que, le puede dar indicios al negocio de porque el modelo está clasificando las cosas de cierta manera. A continuación, en la tabla 5 se presentan los resultados cuantitativos tanto con CountVectorizer y TF-IDF.

Tabla 5. Métricas Asociadas al Ridge Classifier (Average Macro y Weighted)

	Precisión (↑)	***Recall* (↑)**	***Accuracy* (↑)**	F_1 **(↑)**
CountVectorizer	0.94	0.94	0.94	0.94
TF-IDF	**0.979 ***	**0.979***	**0.979 ***	**0.979 ***

El valor promedio de la puntuación de la validación cruzada para f_1 fue de 0.9514 y 0.9665 respectivamente para CountVectorizer y TF-IDF.

4. Resultados Y Recomendaciones

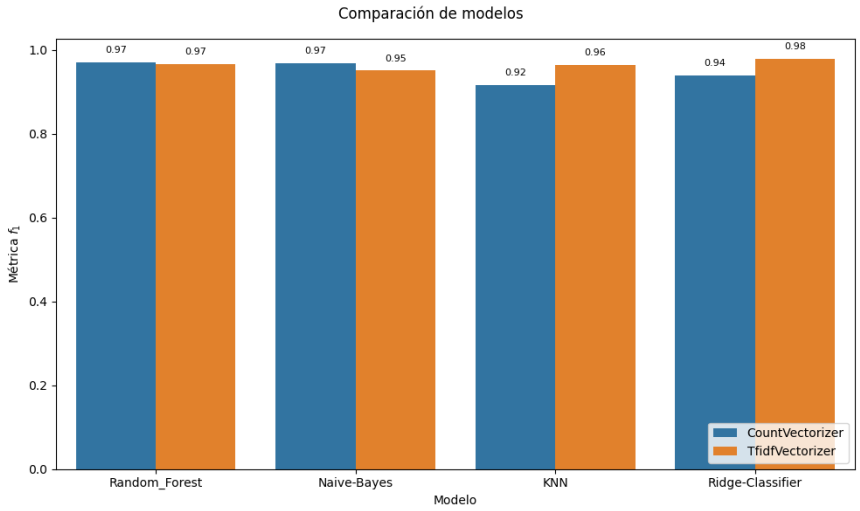

Figura 2. Métrica f_1 obtenida para todos los modelos

En la figura 2 se logra observar que para la mayoría de los modelos se obtuvieron mejores métricas con TF-IDF. Esto se puede deber a que para clasificar los ODS y distinguir entre el 3 (salud), 4 (educación) y 5 (igualdad de género) es importante la relevancia de cada una de las palabras que distingue a cada clase. En particular, TD-IDF justamente le asigna un peso a cada palabra en el texto en función de su frecuencia y relevancia en todos los documentos en general; por lo que, este método resalta palabras que son trascendentales para cada clase lo cual pudo llevar a que tuviera buen f-medida en todos los modelos. Asimismo, si hay una palabra que se repite en todos los documentos, le brinda menos relevancia en comparación con aquellas que aparecen menos. Ahora bien, Naive Bayes Multinomial es el único modelo que obtuvo mejores resultados con CountVectorizer 0.97 > 0.95. Esto se pudo deber a que Count Vectorizer es una representación más simple de los datos de texto lo que brinda relaciones menos complejas para que Naive Bayes lo generalice.

En general, se observó en los resultados que los modelos presentan un rendimiento similar en término de accuracy y f-medida. No obstante, el modelo que obtuvo mejores métricas por centésimas de unidad fue el clasificador Ridge. 0.979 > 0.97. Esto se puede deber a la regularización de norma l_2 que utiliza para penalizar el sobreajuste a los datos. Asimismo, dado que es más simple que otros modelos y es computacionalmente eficiente y con alta interpretabilidad puede ser una buena decisión para el negocio. Por lo que, utilizar el clasificador Ridge con TF-IDF es una buena opción para resolver este problema para UNFPA. Esto se debe a:

- Es un modelo computacionalmente eficiente fácil de ajustar a nuevos datos y entrenar constantemente dada la posibilidad de que nuevos habitantes locales presenten nuevas propuestas u opiniones.
- Los datos de clasificación de texto suelen tener muchas dimensiones y cada documento está representado por una gran cantidad de características. Los clasificadores Ridge manejan datos multidimensionales [9].
- Son altamente interpretables. El negocio puede verificar cuáles *features* está utilizando más el modelo para clasificar los datos. Sabiendo, a qué palabras le está poniendo más atención el modelo, se puede volver a reentrenar.

Es esencial tener en cuenta que el modelo se construye utilizando datos pasados y una pequeña base de datos, lo que significa que su rendimiento puede verse afectado por cambios en el comportamiento de los usuarios y cantidad de datos. Por lo tanto, es fundamental realizar un seguimiento constante y actualizar el modelo de manera regular para garantizar que siga siendo válido y preciso.

5. Mapa de actores relacionado con el producto

A continuación, en la tabla 6 se muestra el mapa de actores que se pueden ver beneficiados con el modelo.

Tabla 6. Principales Objetivos del Negocio

Rol dentro de la Empresa	Tipo de Actor	Beneficio	Riesgo
Asesor y Coordinador de Programas Locales del País	Usuario-cliente	Este actor trabaja para garantizar la cohesión en la ejecución de los distintos programas financiados. Puede ayudarle a llegar a políticas públicas de una manera más rápida y justificar cuantitativamente procesos políticos e intergubernamentales.	Si el modelo tiene métricas bajas puede estar comunicando de manera errónea a demás stakeholders qué políticas son indicadas para implementar o que clases tienen más textos asociados y se deberían enfocar.
Director ejecutivo	Usuario-cliente	Dado que es quien comunica resultados generales, el clasificador puede ayudarle en sus roles de: estrategia al Identificar y contribuir a definir las prioridades estratégicas y operativas globales del UNFPA.	Dado que es quien comunica resultados generales, puede generar desinformación si las métricas son bajas.
Habitante local	Proveedor de los datos / Beneficiador	Apoya con su visión e identifican cuáles son los principales problemas del país para que luego se formulen buenas	Manejo incorrecto de los datos que lleve a la violación de la privacidad de los datos y planteamiento de políticas que no

		políticas que los beneficien.	benefician a la comunidad.
Gobierno central	Financiador	Recibe consejería de las políticas en las que se debería estar entrando Colombia.	En caso de que el modelo no funcione es dinero mal invertido y pudo dejarse de hacer un proyecto con mayor impacto y viabilidad.

Es importante recalcar que este modelo de analítica también le puede servir a empresas del sector privado que deseen estar alineados con los ODS. Dado que el input del modelo son textos, se puede utilizar el modelo para clasificar los textos de la empresa de acuerdo con los ODS y saber en qué ODS se está centrando más la empresa. Por tanto, este modelo no se reduce únicamente a la UNFPA. Hay empresas colombianas como CORONA, Isa, Colombina que se guían por los ODS a quienes también les puede servir.

6. Trabajo en equipo

Se realizó una reunión inicial de lanzamiento y planeación en la que se definieron los roles y forma de trabajo. En esta reunión se definieron los roles y cómo se iba a realizar la primera parte de la entrega del proyecto. Se llegó al acuerdo de que Antonin era el encargado del rol de líder de datos, Ernesto el líder de negocio y Lina la líder de analítica y proyecto. Adicionalmente, cada integrante estaba encargado de implementar uno de los algoritmos de aprendizaje automático. Posteriormente, se hicieron dos reuniones subsecuentes: una para elegir los algoritmos y otra con el experto de analítica para recibir retroalimentación. A continuación, en la Tabla 7 se muestra la repartición de puntos asignados.

Tabla 7. Repartición de Puntuación entre Integrantes

Puntaje Obtenido	**Antonin Bouillaud**	**Ernesto Duarte**	**Lina Gómez**
	0.33	0.33	0.33

En total, cada miembro le dedicó por su parte aproximadamente 8-12 horas al proyecto sin contar las reuniones grupales. Uno de los problemas con los que se enfrentó el grupo fue el tiempo de corrida para buscar los mejores parámetros por lo que tocó hacerlo con antelación para cumplir con la fecha de entrega.

7. Retroalimentación de Estudiante de Estadística

El estudiante de estadística nos contactó el martes y acordamos tener una reunión presencial el día viernes 14 de octubre. En esta reunión se habló sobre las posibles modificaciones que se le podía hacer a la entrega y se discutió el trabajo realizado. Juan nos comentó que en general estaba bien solo había problemas de redacción que nos ayudó a corregir. También nos comentó que en cuanto a cuestiones de estadística presenta unas dudas con respecto a unas definiciones que va a discutir con el equipo docente y más adelante nos comenta los cambios respectivos que se deben hacer.

8. Referencias

[1] “Political and Intergovernmental Affairs Adviser, WCARO, Dakar, Senegal, P5,” United Nations Population Fund, https://www.unfpa.org/jobs/political-and-intergovernmental-affairs-adviser-wcaro-dakar-senegal-p5.

[2] “Deputy executive director (management) (assistant secretary-general), Office of the Executive director, New York,” United Nations Population Fund, https://www.unfpa.org/jobs/deputy-executive-director-management-assistant-secretary-general-office-executive-director-new.

[3] “Multi Country Programme Coordinator, Suva, Fiji, PSRO, P4, FTA,” United Nations Population Fund, https://www.unfpa.org/jobs/multi-country-programme-coordinator-suva-fiji-psro-p4-fta.

[4] https://www.undp.org/es/sustainable-development-goals

[5] https://ods.mma.gob.cl/que-son-los-ods/

[6] https://colaboracion.dnp.gov.co/CDT/Sinergia/Documentos/UNFPA_14092016.pdf

[7] https://ods.dnp.gov.co/es/objetivos/igualdad-de-genero

[8] https://colombia.unfpa.org/es/news/%C2%BFcomo-promover-la-participacion-ciudadana-traves-de-la-innovacion

[9] https://www.mdpi.com/2076-3417/9/4/743#:~:text=Dimensionality%20Reduction,an%20average%2Dsized%20document%20collection.