INTELIGENCIA DE NEGOCIOS - ISIS 3304

"Proyecto 1 Analítica de textos"

Profesora: Haydemar Núñez

Nicolas Rozo Fajardo - n.rozo
Sara Benavides Mora - s.benavidesm
Laura Calderón Molina - l.calderonm

Índice

Entendimiento del negocio y enfoque analítico

Antes de iniciar con el análisis de los datos y la construcción del modelo, resulta esencial entender cuales son los objetivos y naturaleza del negocio y, de la misma forma, cuál es el enfoque analítico que se debería llevar a cabo para satisfacer sus necesidades. La siguiente tabla resume tanto el entendimiento como las necesidades analíticas identificadas.

Oportunidad/problema Negocio	El problema principal que enfrenta el Fondo de Poblaciones de las Naciones Unidas (UNFPA) en este contexto es la alta demanda de recursos para analizar la información textual proporcionada por los ciudadanos, particularmente en la relación de estas opiniones con los Objetivos de Desarrollo Sostenible (ODS) 3 (salud y bienestar), 4 (educación de calidad) y 5 (igualdad de género). La oportunidad radica en automatizar este proceso, lo que permitiría identificar rápidamente las problemáticas locales y aplicar soluciones de manera más eficiente. Además, al automatizar la relación entre las opiniones y los ODS, se puede agilizar la toma de decisiones y optimizar el uso de los recursos, tanto humanos como tecnológicos.
Objetivos y criterios de éxito desde el punto de vista del negocio	A partir del contexto brindado, se puede afirmar que el objetivo del negocio es contar con una herramienta análistica predictiva que pueda ser reentrenada con datos clasificados por expertos y les permita reducir el tiempo y los recursos necesarios para desarrollar de forma satisfactoria el proceso de clasificación de textos en sus categorías de ODS correspondientes. A partir de esto, también se puede afirmar que los criterios de éxito del negocio serían contar con una herramienta análitica que se puede reentrenar pero, sobre todo, que clasifique correctamente los textos brindados en un tiempo menor y con un menor uso de recursos.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El UNFPA y las entidades públicas involucradas en la implementación de los ODS serían los principales beneficiarios del proyecto. El rol que se beneficia directamente incluye a los analistas de datos y los encargados de la toma de decisiones dentro del UNFPA, quienes podrán utilizar el modelo para identificar áreas clave que necesitan intervención en temas de salud, educación e igualdad de género al mismo tiempo que reducen el tiempo necesario para

	llevar a cabo su labor.
Impacto que puede tener en Colombia este proyecto.	En Colombia, este proyecto podría tener un impacto significativo al mejorar la capacidad de los gobiernos locales y nacionales para abordar problemáticas sociales alineadas con los ODS. Al permitir una toma de decisiones más rápida y basada en datos, las intervenciones en temas como la salud pública, la educación y la igualdad de género serían más efectivas. Esto contribuiría a mejorar el bienestar de la población y a cumplir con los compromisos del país en relación con los ODS.
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.	**Descripción de la categoría de análisis:** La categoría de análisis que se propone es predictiva, ya que el objetivo es clasificar automáticamente las opiniones ciudadanas en función de su relación con los ODS 3, 4 y 5. **Tipo de aprendizaje:** Utilizaremos aprendizaje supervisado, donde entrenaremos varios modelos con datos previamente etiquetados para aprender a identificar las características que vinculan una opinión a uno de los tres ODS. **Tarea de aprendizaje:** La tarea principal es la clasificación de textos, donde los modelos deben clasificar cada opinión en una de las clasificaciones ODS (de 3 a 5). **Modelos:** **Random Forest:** Es un modelo de aprendizaje supervisado que utiliza una combinación de múltiples árboles de decisión para hacer predicciones más precisas y reducir el sobreajuste. Funciona creando múltiples árboles a partir de diferentes subconjuntos de datos y promediando los resultados para hacer una predicción final. **K-Nearest Neighbors (KNN):** Este algoritmo busca clasificar cada nueva opinión basándose en las opiniones más cercanas en el espacio de características. Es simple pero efectivo para casos donde las relaciones entre datos similares son importantes. **Árboles de decisión:** Este modelo complementará los anteriores al crear un árbol donde las ramas representan decisiones basadas en características

	clave de las opiniones. Los árboles de decisión son interpretables y pueden ser útiles para visualizar las características que más influencian la clasificación de las opiniones. Cabe resaltar que, además de los modelos, se propone usar el proceso de vectorización Tf-Idf debido a su robustez y capacidad de representar correctamente la importancia de un término dentro de un corpus de documentos.

Entendimiento y preparación de los datos

Luego de entender el objetivo del negocio y proponer una metodología analítica, resulta necesario analizar los datos para determinar características importantes acerca de su naturaleza y, posteriormente, tomar decisiones que favorezcan la construcción de los modelos y el rendimiento de los mismos. Aunque en el Notebook se presenta una descripción procesa y detallada de este proceso (Sección 1, 2 y 3), a continuación se presenta un resumen de los resultados obtenidos y acciones llevadas a cabo.

- **Entendimiento de los datos:** El conjunto de datos entregado por el negocio se compone de dos columnas, la columna *Textos_espanol* presenta las opiniones a analizar en formato de lenguaje natural y la columna *sdg* indica la categoría ODS correspondiente a la opinión. Aunque no se presentan valores faltantes en ninguna de las dos columnas, se identificaron problemas asociados a la codificación incorrecta de caracteres, la presencia de textos en idiomas distintos al Español y el alto contenido de enlaces o rutas hacía sitios web que no aportan valor real en el modelo. Para ver una descripción detallada del proceso de entendimiento de datos, favor revisar la sección 1 del Notebook.

- **Limpieza de los datos:** Luego de entender la naturaleza de los datos y tomar las decisiones necesarias para aumentar su calidad y, por ende, la de los modelos, se llevó a cabo la limpieza de los datos. Dado a que no era necesario imputar valores ni llevar a cabo normalizaciones, se decidió limpiar la totalidad del conjunto de datos cargados al mismo tiempo. Para esto, se implementó una función por cada decisión tomada y, al final, se unificaron en una única función de limpieza que fue aplicada sobre los datos. Para ver una descripción detallada del proceso de limpieza de datos, favor revisar la sección 2 del Notebook.

- **Preparación de los datos:** Finalmente, luego de llevar a cabo la limpieza de los datos, se realizó la preparación final para su uso en el entrenamiento de los modelos predictivos. Aunque algunos de los algoritmos usados pueden trabajar con palabras en formato de String, la librería Scikit Learn no lo permite, por lo cual resulta necesario codificar todos los textos presentes en los datos a vectores numéricos. Para llevar a cabo esta tarea se utilizó el método de vectorización Tf-Idf debido a su robustez, inclusión del contexto, y capacidad de representar la importancia de los términos dentro de un corpus de documentos. Para ver una descripción detallada del proceso de preparación de datos, favor revisar la sección 3 del Notebook.

Modelado y evaluación

Con el fin de obtener los mejores resultados para el negocio, se construyeron tres modelos de clasificación usando algoritmos distintos y, luego de su evaluación, se escogió el que presentaba un mejor desempeño general. A continuación, se presenta la información relevante de cada modelo.

- **Random Forest:** Es un algoritmo utilizado para tareas de aprendizaje supervisado que, a nivel general, producen un árbol de decisión final combinando múltiples árboles de decisión creados con distintos parámetros. Suelen utilizarse ampliamente en el mundo del machine learning debido a que, gracias a la mezcla de varios árboles de decisión, brindan un desempeño superior a otros algoritmos y además ayudan a reducir el sobreajuste del modelo a los datos. Finalmente, este algoritmo se puede usar para tareas tanto de regresión como de clasificación [1].

  Ahora bien, en lo que respecta al proyecto, se eligió este algoritmo debido a su amplio uso en tareas de clasificación, su particular buen desempeño en tareas de clasificación de lenguaje natural y la posibilidad que brinda de tener en cuenta una gran variedad de árboles de clasificación al mismo tiempo y elegir su combinación más óptima. Cabe aclarar que, a pesar de su gran cantidad de ventajas, es un modelo que demanda recursos computacionales, por lo cual su entrenamiento y selección puede ser un proceso costoso y demorado.

- **K Nearest Neighbors (KNN):** Es un algoritmo utilizado para tareas de aprendizaje supervisado que, a nivel general, clasifica los nuevos datos individuales basándose en la tendencia de clasificación de sus datos cercanos. Más formalmente, es un algoritmo que utiliza la proximidad para hacer clasificaciones o predicciones de grupo sobre un dato individual. Aunque se puede usar tanto para tareas de regresión como de clasificación,

se suele usar ampliamente en tareas de clasificación, donde destaca por su desempeño y simplicidad [2].

En lo que respecta al proyecto, se eligió usar este algoritmos ya que, a pesar de tener una base teórica sencilla, es conocido por obtener buenas resultados en tareas de clasificación de textos y, al mismo tiempo, por que permite el entrenamiento y selección de modelos sin la necesidad de contar con una gran potencia computacional.

- **Árbol de Decisión:** Es un algoritmo utilizado para tareas de aprendizaje supervisado que utiliza una estructura de árbol (muy común en el área de la estadística) para representar una serie de decisiones y consecuencias. Su funcionamiento se basa en dividir repetidamente un conjunto de datos en conjuntos más pequeños basándose en ciertas características y, haciendo uso de criterios como la impureza o entropía, encontrar la mejor relación posible entre las características y las clases a predecir. Es importante aclarar que los nodos hoja del árbol representan las predicciones de categoría y los nodos intermedios las decisiones que se toman [3].

  En lo que respecta al proyecto, se escogió este algoritmo ya que es ampliamente utilizado en tareas de clasificación y, al mismo tiempo, debido a su naturaleza no requiere de una gran potencia computacional para su entrenamiento. A su vez, se quiso incluir el modelo ya que, como es la base para el algoritmo Random Forest, resulta importante entender su funcionamiento y poder comparar los resultados generados.

La creación de los modelos se puede observar en las secciones 4, 5 y 6 del Notebook.

Resultados

En el desarrollo del proyecto de analítica de textos, se implementaron tres algoritmos de aprendizaje automático: Random Forest, K Nearest Neighbors (KNN) y Árbol de Decisión. Estos modelos fueron evaluados para determinar su eficacia en la clasificación de textos, con el objetivo de identificar cuál es el que mejor se adapta a las necesidades del proyecto y nos permite generar insights de valor al negocio.

El modelo de Random Forest demostró el mejor desempeño general con una precisión del 96% para la clase 3, del 97% para la clase 4 y del 96% para la clase 5. También logró un F1-score ponderado de 0.97, lo que indica una sólida capacidad de generalización. El modelo K Nearest Neighbors (KNN) presentó una ligera disminución en el rendimiento en comparación con el anterior modelo pues tuvo una precisión del 95% para las clases 3 y 4, y del 96% para la clase 5. Además, obtuvo

un F1-score ponderado de 0.95, mostrando su capacidad de generalización pero manteniéndose por debajo del anterior modelo por 0.02 y un poco más de errores de clasificación. Por último, el modelo de árbol de decisión sigue siendo confiable pero presentó una capacidad de generalización menor que la de los dos modelos anteriores. Obtuvo precisiones de 93% para la clase 3 y de 94% para las clases 4 y 5, y un F1-score promedio de 0.94 mostrando una reducción en términos de precisión y sensibilidad. Así que, según los resultados obtenidos se recomienda el uso del algoritmo Random Forest para el proyecto porque presenta un sólido desempeño en la generalización y su capacidad para manejar de forma eficaz la clasificación de textos, este modelo es confiable y será de utilidad para el negocio al ser una herramienta analítica que sirve para el soporte de la toma de decisiones.

En el análisis de las palabras más importantes por clase, se identificaron términos clave que reflejan el enfoque temático de cada clase. Para la clase 3, relacionada con Salud y Bienestar, las palabras de mayor peso incluyen “salud”, “atención” y “mental”, subrayando la relevancia de aspectos de la salud en la clasificación. En la clase 4, que abarca la educación, términos como “educación”, “estudiantes” y “escuelas” destacan la importancia del entorno académico. En la clase 5, enfocada en igualdad de género, las palabras “mujeres”, “género” e “igualdad” son las dominantes y reflejan el énfasis en cuestiones de género y derechos. Es necesario resaltar que, además de que estas palabras brindan información relevante para clasificar los textos, también son de alta relevancia para el negocio, ya que sirven como base para la toma de decisiones y acciones en la sociedad. Por ejemplo, para el ODS 5 se podría ahondar en políticas de igualdad de género en el trabajo y de reducción de violencias basadas en género, ya que estas son las palabras que dominan en la vectorización.

Poder clasificar textos para identificar las opiniones de los habitantes locales es crucial para el Fondo de Poblaciones de las Naciones Unidas (UNFPA) en su esfuerzo por alinear las preocupaciones ciudadanas con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Al analizar y clasificar automáticamente las opiniones, el UNFPA puede detectar y entender mejor los problemas de salud, educación e igualdad de género que afectan a las comunidades. Esto no solo optimiza el uso de recursos al reducir la necesidad de hacer un análisis manual, sino que también permite una respuesta más rápida y efectiva a las problemáticas identificadas.

En conclusión, el uso del modelo de Random Forest para la analítica de textos para clasificar las opiniones de la población en alguno de los ODS de tipo 3, 4 y 5 permitirá que el negocio tenga una herramienta valiosa para optimizar el análisis de las problemáticas de las comunidades. Esto se da gracias a que este modelo destaca por su confiabilidad, precisión y a su sólido desempeño en la generalización y su capacidad para manejar de manera eficaz la clasificación de textos.

Mapa de actores

Al crear un producto de analítica de datos para una organización o negocio, diversos roles o actores se pueden ver beneficiados por el producto pero, a su vez, pueden generarse riesgos que deben ser considerados. A continuación, se presenta el mapa de actores en el cual se especifican estas características.

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Usuario - Cliente	Facilita la evaluación y seguimiento de políticas públicas relacionadas con los ODS, permitiendo una toma de decisiones informada y basada en datos, lo que mejora la eficacia y eficiencia de sus proyectos y actividades.	Si el modelo de clasificación no funciona correctamente o presenta errores y sesgos considerables, podría llevar a interpretaciones erróneas de la información recopilada y, por ende, podría desembocar en tomas de decisiones inadecuadas.
Entidades públicas y gubernamentales	Financiador	Obtiene información de forma rápida y precisa sobre cómo sus políticas impactan en la sociedad con relación a los ODS 3,4 y 5, lo cual puede ayudar a una toma de decisiones más efectiva y a mejorar procesos de gobierno.	En caso de que el modelo no presente los resultados esperados y no pueda ser incorporado en la práctica, estarían perdiendo el dinero invertido. A su vez, el modelo podría ignorar matices importantes o presentar sesgos significativos, lo cual podría llevar a decisiones incorrectas si se

			depende mucho del mismo.
Universidad de los Andes (entidad colaboradora)	Proveedor	Contribuye al desarrollo de tecnología y soluciones que tienen un impacto significativo en la mejora de las políticas públicas y la calidad de vida de las comunidades locales, a la vez que permite poner a prueba los conocimientos de los estudiantes.	Al proveer el modelo, la universidad es responsable de los problemas de calidad del mismo. Por ejemplo, si se toman decisiones públicas inadecuadas debido a que el modelo tuvo un entrenamiento sesgado, la universidad sería responsable por esta mala configuración y podría tener implicaciones éticas.
Habitantes locales	Beneficiado	Sus opiniones y preocupaciones se ven reflejadas en las políticas y acciones de desarrollo sostenible, lo que potencialmente podría mejorar su calidad de vida. A su vez, se benefician ya que se aceleran los procesos de decisión y análisis asociados a los ODS:	Vulneración de la privacidad y seguridad de los datos personales en caso de que la información sea manejada de forma inadecuada. A su vez, los habitantes podrían verse afectados si el modelo se encuentra mal entrenado y genera resultados que no se encuentran alineados con la realidad.

Trabajo en equipo

A continuación, se presenta toda la información asociada a la distribución del trabajo realizada por el equipo, la contribución individual, el rol dentro de esta etapa y el puntaje asignado a cada integrante.

Nombre	Rol	Algoritmo	Puntaje Otorgado
Nicolas Rozo Fajardo	Líder de Proyecto y Líder de Datos	Random Forest	33.33
Laura Calderón Molina	Líder de Negocio	KNN (K Nearest Neighbors)	33.33
Sara Benavides Mora	Líder de Analítica	Árbol de Decisión	33.33

Posteriormente, se presenta la lista de tareas asignadas, los responsables de su realización y el tiempo invertido en completar la tarea. Cabe resaltar que la totalidad de las tareas asignadas fueron completadas.

Tarea	Responsables	Tiempo Invertido
Entendimiento y limpieza de los datos	Nícolas Rozo Fajardo	4 horas
Vectorización de los datos	Nicolas Rozo Fajardo	0.5 horas
Construcción de los modelos de clasificación	Nicolas Rozo Fajardo, Laura Calderón Molina, Sara Benavides Mora	2 horas
Evaluación de los resultados individuales de cada modelo	Nicolas Rozo Fajardo, Laura Calderón Molina, Sara Benavides Mora	1 hora
Construcción y finalización del documento	Nicolas Rozo Fajardo, Laura Calderón Molina, Sara Benavides Mora	3 horas
Creación del script del video, material audiovisual, y grabación	Laura Calderón Molina, Sara Benavides Mora	3 horas

De la misma forma, se presentan los retos enfrentados durante la solución del proyecto y las decisiones tomadas para poder superarlos.

Reto	Solución
Corrección de codificación errónea de los caracteres	Inicialmente se optó por usar las librerías disponibles para la corrección de la codificación usada, sin embargo, se obtuvieron errores durante este proceso. La solución definitiva fue consultar sobre los caracteres mal codificados y crear un diccionario de reemplazos el cual se aplicó posteriormente al corpus de textos.
Eliminación de enlaces a páginas web o rutas de archivos	Durante la etapa de entendimiento de los datos se identificó que existían una gran cantidad de enlaces a páginas web o rutas de archivos que debían ser eliminadas. Aunque se intentó usar una expresión regular, al final se optó por solucionar el problema usando comparación directa de cadenas de texto.
Tiempo de entrenamiento excesivo para el algoritmo Random Forest	Buscando obtener el mejor modelo se propuso llevar a cabo un proceso de cross validación con 6 hiper parámetros, sin embargo, el espacio de búsqueda era considerablemente grande y el proceso demoraba más de 15 minutos. Para solucionar este problema se optó por reducir el espacio de búsqueda, dejando únicamente los hiperparametros más relevantes y ajustando sus valores al contexto del problema.

A modo de conclusión, se puede afirmar que el trabajo llevado durante esta etapa fue satisfactorio ya que, además de obtener excelentes resultados en lo que respecta a la construcción de los modelos, se cumplió con todas las tareas propuestas y no se presentaron grandes inconvenientes a la hora de desarrollar las distintas tareas como grupo. Para las próximas entregas, recomendamos seguir trabajando bajo la misma dinámica y mantener la calidad tanto del trabajo técnico como del trabajo en grupo.

Bibliografía

[1] *¿Qué es el random forest?* (s.f.). IBM. https://www.ibm.com/mx-es/topics/random-forest

[2] *¿Qué es KNN?* (s.f.). IBM. https://www.ibm.com/mx-es/topics/knn#:~:text=algoritmo%20de%20k%20vecinos%20m%C3%A1s,un%20punto%20de%20datos%20individual.

[3] *¿Qué es un árbol de decisión?* (s.f.). IBM. https://www.ibm.com/es-es/topics/decision-trees