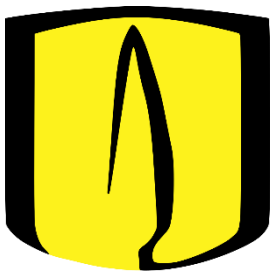
Proyecto 1

Fondo de Poblaciones de las Naciones Unidas (UNFPA)



Integrantes (Grupo 15)

Sara Sofia Cárdenas Rodríguez - 202214907

Daniel Felipe Diaz Moreno - 202210773

Juan Sebastián Urrea López - 201914710

Universidad de Los Andes

Departamento de Ingeniería de Sistemas y Computación

Inteligencia de Negocios - ISIS 3301

Bogotá D.C., Colombia

2024

Tabla de contenido

1. Entendimiento del negocio y enfoque analítico ……. 3
2. Entendimiento y preparación de los datos …..…….. 4
3. Modelado y evaluación …….. 4
4. Resultados ……..…….. 5
4.1. Descripción de los resultados obtenidos ……...…….. 5
4.2. Análisis de las palabras identificadas ……..…..….. 5
4.3. Datos de prueba compartidos …..……..……...…. 6
4.4. Video del proyecto ……...…. 6
5. Mapa de actores relacionado con el producto de datos creado ……. 7
6. Trabajo en equipo …….. 10
6.1. Roles y las tareas realizadas por cada integrante …….……... 10
6.2. Reuniones de grupo ……..……..…… 11
7. Referencias …….....……. 12

1. Entendimiento del negocio y enfoque analítico

Oportunidad / Problema de negocio	El Fondo de Poblaciones de las Naciones Unidas (UNFPA) necesita analizar grandes volúmenes de datos textuales (opiniones ciudadanas) para identificar problemas y evaluar soluciones en relación con los ODS 3 (Salud y Bienestar), ODS 4 (Educación de Calidad), y ODS 5 (Igualdad de Género). El proceso actual es manual y consume muchos recursos.
Objetivos y criterios de éxito desde el punto de vista del negocio	Clasificar automáticamente las opiniones ciudadanas en categorías de los ODS (3, 4, 5) con alta precisión y eficiencia. Reducir el tiempo de análisis y generar resultados accionables que ayuden en la toma de decisiones políticas y estratégicas. El éxito se mide por la precisión, recall y f1-score del modelo.
Organización y rol dentro de ella que se beneficia con la oportunidad definida	El UNFPA y los entes gubernamentales responsables de los ODS en Colombia. Dentro de estas organizaciones, los analistas de políticas públicas, los líderes de proyectos sociales y los gestores de recursos son los principales beneficiarios, ya que pueden usar la información para priorizar recursos y acciones.
Impacto que puede tener en Colombia este proyecto	Mejora de la eficiencia en la identificación de problemas sociales clave a través del análisis automatizado de opiniones. Esto podría optimizar la implementación de políticas públicas y programas que mejoren los indicadores de bienestar relacionados con la salud, la educación y la igualdad de género.
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	Predictivo, enfocado en clasificación de texto. Este proyecto utiliza técnicas de aprendizaje supervisado para clasificar opiniones ciudadanas en las categorías relevantes de los ODS 3, 4 y 5. Los algoritmos propuestos son K Nearest Neighbors (KNN), Árboles de Decisión y Gradient Boosting. El procesamiento del texto incluye técnicas de NLP para preparar los datos antes de la clasificación, y se emplea SMOTE para el balanceo de clases en el entrenamiento, asegurando que el

	modelo sea efectivo incluso con datos desbalanceados. Estos algoritmos han sido seleccionados por su capacidad para manejar datos de alta dimensionalidad y su efectividad en tareas de clasificación con múltiples clases.

2. Entendimiento y preparación de los datos

Se cargaron los datos y se inició su proceso de entendimiento. En este, se pudo hacer un análisis descriptivo de una columna con texto y de un atributo categórico que indicaba la clasificación del comentario respecto a los ODS en cuestión. No se evidenciaron problemas de calidad de datos asociados a completitud, unicidad, validez ni consistencia. Posteriormente, en la preparación de datos se realizó One Hot Encoding. El procesamiento de texto utilizó técnicas de stemming, lematización y nm-gramas. También se dividió el conjunto de entrenamiento en dos para tener un conjunto de validación, se aplicó el balanceo de clases SMOTE y se estandarizaron los atributos resultantes.

3. Modelado y evaluación

Se utilizaron tres algoritmos diferentes de clasificación (Árbol de decisión, KNN y Gradient Boosting). Se presentaron las métricas de evaluación respectivas junto a la matriz de confusión con el fin de determinar el mejor modelo para esta labor dados los datos.

4. Resultados

4.1. Descripción de los resultados obtenidos

En el análisis de los tres algoritmos de clasificación implementados (Árbol de Decisión, K-Vecinos más Cercanos, y Gradient Boosting), se observa una clara diferencia en su desempeño tanto en el conjunto de entrenamiento como en el de validación.

El modelo de Árbol de Decisión mostró un rendimiento sólido en el conjunto de entrenamiento, con una exactitud del 93% y un f1-score promedio de 0.93, lo que indica un buen ajuste al conjunto de datos. En el conjunto de validación, la exactitud fue ligeramente menor, alcanzando un 89%, con un f1-score de 0.89. Aunque el modelo generaliza bien, se observó una pequeña caída en el rendimiento para la clase ODS 5 (f1-score de 0.88), lo que sugiere que el modelo podría mejorar en la clasificación de esta categoría.

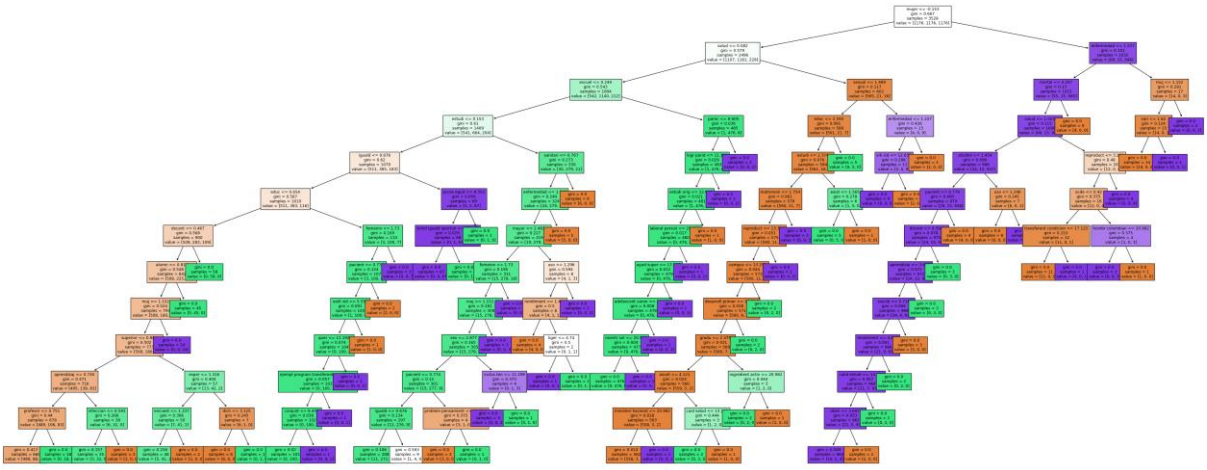
Por su parte, en el conjunto de entrenamiento, el KNN alcanzó una exactitud del 100%, lo que indica un claro sobreajuste, ya que el modelo memorizó los datos de entrenamiento. Sin embargo, en el conjunto de validación, el rendimiento cayó drásticamente a un 53% de exactitud, con un f1-score promedio de 0.51. El modelo mostró un rendimiento inconsistente entre las clases, con un f1-score de solo 0.35 para la clase ODS 5, lo que refleja su incapacidad para generalizar.

Por último, el modelo de Gradient Boosting demostró un rendimiento excelente en el conjunto de entrenamiento, con una exactitud del 97% y un f1-score de 0.97, lo que refleja su capacidad para capturar las relaciones subyacentes en los datos. En el conjunto de validación, el modelo mantuvo un rendimiento alto, con una exactitud del 93% y un f1-score de 0.93. Las clases ODS 3, 4 y 5 presentaron resultados consistentes, siendo el f1-score para la clase ODS 5 de 0.93, lo que indica una buena generalización del modelo. A pesar de sus resultados prometedores, Gradient Boosting presenta la dificultad de un tiempo de entrenamiento excesivo. Por ejemplo, realizar un Grid Search de 2x2 puede tardar más de 5 horas de entrenamiento, por lo que se plantea un tradeoff adicional al seleccionar el modelo.

De los tres modelos, Gradient Boosting es el modelo más adecuado para cumplir con los objetivos del negocio del UNFPA. Su alto rendimiento en generalización y su capacidad para clasificar con precisión las opiniones ciudadanas lo convierten en una herramienta eficaz para automatizar el análisis de datos textuales, permitiendo a la organización tomar decisiones informadas en torno a la implementación de los ODS, optimizando recursos y tiempo de análisis.

4.2. Análisis de las palabras identificadas

Incluir el análisis de las palabras identificadas para relacionar las opiniones con los ODS y posibles estrategias que la organización debe plantear utilizando los resultados obtenidos en los modelos analíticos y una justificación de por qué esa información es útil para ellos.



En el árbol de decisión, las palabras que tienen mayor impacto general son aquellas que aparecen cerca de la raíz. La variable "mujer" es la más influyente ya que se encuentra en el nodo raíz, dividiendo el conjunto de datos de manera significativa.

Otras palabras como "salud", "educ", "escolar" y "enfermedad" también tienen un impacto relevante, ya que aparecen en los primeros niveles del árbol y afectan una gran cantidad de muestras.

Para el ODS 3, las palabras más influyentes son "dcomet", "alumn", "mtij", que aparecen en las ramas que llevan a esta categoría, proporcionando una buena separación de los datos. Además, palabras como "esper", "encuest" y "profesor" también tienen un rol clave en la clasificación dentro de esta categoría, especialmente en los nodos hoja.

En el ODS 4, destacan "salud", "sexigual" y "alcohol" como las más relevantes para llevar a esta clasificación, mientras que en en el ODS 5, las palabras "enfermedad", "mortal" y "tranferent" son las más influyentes. Estas palabras específicas de cada categoría juegan un papel crucial en la clasificación final de los datos en las hojas del árbol.

A partir de estos resultados, la organización podría plantear estrategias específicas enfocadas en mejorar los indicadores relacionados con los ODS más mencionados. Por ejemplo, si "salud" aparece como una palabra clave de gran impacto, podrían desarrollar iniciativas para mejorar el acceso a servicios de salud o implementar programas educativos que promuevan la prevención de enfermedades. De manera similar, el énfasis en palabras como "educ" o "escolar" sugiere la necesidad de programas que fortalezcan la educación y fomenten la igualdad de oportunidades.

Esta información es útil para la organización porque permite priorizar acciones basadas en los intereses y necesidades expresadas en los datos. Al comprender qué palabras están asociadas con mayores preocupaciones o influencia, pueden diseñar campañas, políticas y proyectos que maximicen el impacto social, económico y ambiental, contribuyendo directamente al cumplimiento de los ODS y mejorando la percepción pública de la organización.

4.3. Datos de prueba compartidos

Esta es la carpeta de los archivos csv obtenidos

Archivos BI

Estos son los datos de prueba obtenidos por las predicciones

Predicciones.csv

Estos son los datos obtenidos luego de preparar el conjunto de entrenamiento

Proyecto 1 - X y Y smote.csv

Proyecto 1 - X smote.csv

Proyecto 1 - Y smote.csv

4.4. Video del proyecto

Este es el link del video

https://youtu.be/qsHRt4P-29M

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

Este análisis se basa en la información suministrada por la UNFPA (2021) y el PNUD (2016) en sus sitios web

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Usuario - Cliente	Apoya la automatización del análisis de datos, disminuyendo la carga sobre los expertos. De esta manera, puede existir una mejora en la toma de decisiones, aumento de la escalabilidad del proyecto y un otorgamiento de resultados relevantes a través de una interfaz	El sesgo en los datos es riesgoso, ya que el modelo podría clasificar de forma errónea y esto llevaría a decisiones equivocadas. También puede existir una mala interpretación de los comentarios o presentarse una mala inversión de las donaciones en el proyecto dados los costos de mantenimiento y reentrenamiento
Programa de las Naciones Unidas para el Desarrollo (PNUD)	Proveedor de información de los ODS	Aumenta la visibilidad de los ODS y la colaboración interinstitucional dentro de la ONU. De igual forma, el producto podría permitirle una mejor medición de su impacto si tiene acceso a sus resultados	Puede existir un conflicto entre actores, un desacuerdo en la clasificación de comentarios y una mala inversión de los recursos necesarios para la integración de estas instituciones
Organización de las Naciones Unidas (ONU)	Financiador indirecto	Apoya a la implementación de los ODS, alineándose con las metas globales de la organización. También puede darse	Podría incurrirse en resultados de clasificación incorrectos, costos adicionales por el proyecto y un impacto

		un mejor uso del capital humano y económico. Si el proyecto tiene éxito, esto favorece su escalabilidad y replicabilidad	en la reputación de la organización si se toman malas decisiones con los datos obtenidos
Gobierno Nacional	Financiador directo	Mejora la toma de decisiones relacionadas con la integración con la ONU y la implementación de políticas públicas. También se puede fomentar la participación ciudadana y enfocar el monitoreo del progreso de los ODS si cuenta con los resultados del proyecto	Puede realizar aportes a la organización que resulten en una mala inversión. De igual forma, puede tomar decisiones con base en datos imprecisos y así fracasar en cumplir las expectativas de los ciudadanos
Administraciones departamentales	Financiador indirecto	Apoya la toma de decisiones informadas, mejora la asignación de recursos y posibilita la adaptación de las políticas locales si cuenta con acceso a los resultados del proyecto	Los resultados podrían no estar conectados con la realidad local, entonces esto podría llevar a malas decisiones administrativas
Instituciones nacionales, territoriales y locales (Sistema Nacional de Juventud en Colombia, Profamilia, DANE)	Proveedor de opiniones e información estadística	Mejora el aprovechamiento de la información recolectada y favorece la colaboración interinstitucional de cara al desarrollo sostenible. Esto ayuda a tomar mejores decisiones para prestar servicios o tomar la retroalimentación de los ciudadanos.	Podría darse una mala interpretación de los datos o problemas de coordinación. Igualmente, como se tiene información con cierto grado de sensibilidad, un manejo inadecuado de los datos pone en riesgo la privacidad de las personas y la confianza en estas instituciones
Organizaciones de sociedad civil o no	Beneficiario y/o Proveedor de	Aumenta su incidencia en políticas públicas y participación en procesos estratégicos.	Puede darse una sobrecarga operativa que no sea compensada por los

gubernamentales (ONGs)	opiniones e información estadística	De igual forma, las decisiones tomadas pueden favorecer al cumplimiento de la misión de estas organizaciones	beneficios del proyecto. De igual forma, hay riesgos de privacidad y manejo de datos
Empresas donantes del sector privado colombiano	Financiador directo	Mejora la reputación corporativa, contribuye al desarrollo sostenible y podría darle acceso a información estratégica	Este proyecto por definición no tendría un retorno de la inversión en términos financieros, pero aun así podría ser una mala inversión si no tiene el impacto esperado
Ciudadanos colombianos y habitantes de Colombia	Beneficiario y/o Proveedor de opiniones	Mejora su participación en la toma de decisiones y el acceso a información relevante si los resultados son abiertos. Igualmente, se pueden identificar problemas locales y mejorar la calidad de las políticas públicas con los comentarios	Existe una posible falta de impacto de la información suministrada y de desigualdad en la representación. También hay riesgos de seguridad y privacidad de los datos, al igual que en la interpretación incorrecta de sus opiniones

6. Trabajo en equipo

6.1. Roles y las tareas realizadas por cada integrante

Integrante 1: Juan Sebastián Urrea López

Rol(es): Líder de negocio

Contribución: 33.3 / 100

Horas dedicadas: 10

Algoritmo implementado: Gradient Boosting

Tareas:

- Asegurar el cumplimiento del problema identificado
- Verificar que el producto se pueda comunicar con el negocio
- Implementar el algoritmo de Gradient Boosting escogiendo el mejor modelo
- Documentar el entendimiento del negocio y el enfoque analítico a utilizar
- Definir el Describir los resultados obtenidos
- Analizar y entregar las palabras identificadas por el modelo

- Plantear posibles estrategias a la organización para el uso de los datos
- Elaborar la wiki con la líder del proyecto

Integrante 2: Sara Sofía Cárdenas Rodríguez

Rol(es): Líder de proyecto y Líder de analítica

Contribución: 33.3 / 100

Horas dedicadas: 10

Algoritmo implementado: K vecinos más cercanos (KNN)

Tareas:

- Definir fechas de reuniones
- Asignar equitativamente los entregables
- Subir la entrega final
- Implementar el algoritmo de KNN escogiendo el mejor modelo
- Elaborar el mapa de actores
- Entregar los datos de prueba compartidos
- Realizar el video
- Verificar el cumplimiento de estándares en los entregables
- Elaborar la wiki con el líder de negocio

Integrante 3: Daniel Felipe Diaz Moreno

Rol(es): Líder de datos

Contribución: 33.3 / 100

Horas dedicadas: 10

Algoritmo implementado: Árbol de decisión

Tareas:

- Asignar tareas con respecto a la elaboración del notebook
- Realizar entendimiento de datos
- Realizar procesamiento de datos
- Dejar disponibles los datos para ser trabajados por el equipo
- Implementar el algoritmo de árboles de decisión escogiendo el mejor modelo
- Documentar el trabajo en equipo realizado

- Documentar las reuniones realizadas
- Garantizar consistencia del notebook
- Almacenar el trabajo realizado en un repositorio

6.2. Reuniones de grupo

Reunión de lanzamiento y planeación

Fecha: 23 de agosto de 2024. **Integrantes:** Todos

Resumen: Durante esta reunión se definieron los roles y las tareas asignadas a cada miembro del equipo. Se establecieron las primeras acciones necesarias para el desarrollo del proyecto y se acordó la fecha de la próxima reunión. El líder de datos se comprometió a presentar la información relevante en la siguiente sesión.

Reunión de ideación

Fecha: 27 de agosto de 2024. **Integrantes:** Todos

Resumen: El líder de datos presentó los resultados acordados. Se discutió el modelo de negocio, identificando a los beneficiarios y cómo se verían favorecidos por la solución propuesta. El líder de negocio se encargó de documentar esta información en los entregables y de compartirla con el equipo.

Reuniones de seguimiento

Fecha: 30 de agosto de 2024. **Integrantes:** Todos

Resumen: Esta reunión se realizó de manera escrita a través de WhatsApp, dirigida por la líder del proyecto. Se revisaron los avances individuales de cada miembro en sus respectivas tareas.

Fecha: 4 de septiembre de 2024. **Integrantes:** Todos

Resumen: También realizada de forma escrita mediante WhatsApp y dirigida por la líder del proyecto, en esta reunión se discutió cómo organizar y entregar los productos finales al cliente.

Reunión de finalización

Fecha: 7 de septiembre de 2024. **Integrantes:** Todos

Resumen: Se revisó el cumplimiento de las tareas asignadas y se reflexionó sobre la efectividad en la distribución del trabajo y la comunicación. No se identificaron mayores inconvenientes, por lo que no se propusieron acciones adicionales.

REFERENCIAS

Fondo de Poblaciones de las Naciones Unidas (UNFPA). (26 de noviembre del 2021). UNFPA en Colombia - Asegurando derechos y opciones para todas las personas. https://colombia.unfpa.org/es/unfpa-en-colombia

Programa de las Naciones Unidas para el Desarrollo (PNUD). (9 de septiembre del 2016). ¿Qué son los Objetivos de Desarrollo Sostenible? https://www.undp.org/es/sustainable-development-goals