Inteligencia de Negocios

Proyecto: Parte 1

Carlos Julio Pinto Rodríguez – 201616667
Juan Camilo González – 201911030
Laura Daniela Arias Flórez – 202020621

Tabla de Contenido

Entendimiento del negocio y enfoque analítico

La Agenda 2030 para el Desarrollo Sostenible de la ONU, adoptada en 2015, establece 17 Objetivos de Desarrollo Sostenible (ODS) con 169 metas específicas. Estos ODS abordan cuestiones cruciales como la erradicación de la pobreza, acceso a la salud y educación, igualdad de género, y sostenibilidad ambiental. El proyecto asignado se centra en el trabajo colaborativo con el Fondo de Poblaciones de las Naciones Unidas (UNFPA) y la Universidad de los Andes para implementar estrategias de clasificación de textos. El objetivo principal es desarrollar un modelo de clasificación con técnicas de aprendizaje automático para relacionar automáticamente un texto con los ODS. Este modelo, junto con la aplicación asociada, facilitará la interpretación y análisis de información recopilada a través de diferentes fuentes durante los procesos de planificación participativa para el desarrollo a nivel territorial.

Los objetivos del proyecto 1 de inteligencia de negocio es que la implementación de un modelo de clasificación automática sea exitosa y que asocie textos con los ODS correspondientes. El éxito se medirá por la precisión y eficacia del modelo en la clasificación, así como su utilidad práctica para UNFPA en el análisis automatizado de opiniones. La aplicación que se desarrollará a lo largo del proyecto también debe ser intuitiva y facilitar la interacción con los resultados del modelo. Por otro lado, la aplicación desarrollada determinará en comprender el impacto de abordar estos ODS específicos en el contexto colombiano, destacando cómo la aplicación de esta tecnología puede contribuir a la evaluación de políticas públicas y el desarrollo sostenible en el país.

Para avanzar en la segunda etapa del proyecto, se estableció contacto con las estudiantes del curso de estadística Letitia Sofia Russi Bello (ls.russi) y Maria Gabriela Vidal Salcedo (m.vidals) y se llevó a cabo una corta reunión el pasado viernes 13 de octubre. Aunque no se ha discutido directamente un plan sobre los pasos a seguir, se acordó programar prontamente otro espacio de encuentro. Este encuentro permitirá discutir los resultados de la primera etapa y establecer una estrategia conjunta para avanzar en la implementación del modelo y la aplicación en la siguiente fase del proyecto.

Entendimiento y perfilamiento de los datos.

Todo lo mencionado en el siguiente apartado se puede evidenciar en el archivo de Jupyter Notebook (Proyecto1.ipynb) que acompaña a este documento.

En la fase de entendimiento y preparación de los datos, se comenzó por importar las bibliotecas y paquetes necesarios para el proyecto. Luego, se procedió a cargar el conjunto de datos desde un archivo Excel que contiene los textos a clasificar. Con el objetivo de obtener una comprensión más profunda de los datos, se aplicó una herramienta de perfilamiento de datos mediante la librería "pandas_profiling", lo cual permitió obtener estadísticas descriptivas, visualizaciones y explorar las características clave de los textos. Además, se llevaron a cabo tareas fundamentales de preprocesamiento de texto, como la eliminación de caracteres no ASCII, la supresión de signos de puntuación, la conversión de números en palabras y la eliminación de palabras de parada (stop words) en español. También se aplicaron técnicas de lematización y se realizaron ajustes para eliminar prefijos y sufijos especiales en las palabras. Estas transformaciones aseguran que los datos estén limpios y listos para ser utilizados en los modelos de clasificación. Por último, se dividió el conjunto de datos en subconjuntos de entrenamiento y prueba, lo que permitirá evaluar el rendimiento de los modelos en una etapa posterior del proyecto.

Modelado y evaluación

En la fase de modelado y evaluación, se aplicaron varios algoritmos de clasificación para abordar la tarea de categorizar los textos en las categorías específicas de interés. Estos algoritmos se eligieron con el propósito de explorar diferentes enfoques y evaluar cuál de ellos ofrece el mejor rendimiento en términos de precisión de clasificación. Entre los algoritmos utilizados se incluyen árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM) y k vecinos más cercanos (KNN). Para cada algoritmo, se llevaron a cabo varios pasos, incluyendo la transformación de los textos en representaciones numéricas mediante técnicas como CountVectorizer y TfidfVectorizer. Además, se realizó una búsqueda de hiperparámetros para afinar los modelos utilizando la técnica de GridSearchCV. Esto permitió identificar la combinación óptima de parámetros que maximiza el rendimiento del modelo.

Después de entrenar cada modelo, se evaluaron en términos de su capacidad para clasificar correctamente los textos. Se calcularon métricas de calidad como la precisión, la recuperación y la puntuación F1. Además, se generaron matrices de confusión para visualizar la cantidad de clasificaciones correctas y erróneas. Con esta evaluación exhaustiva, se pudo determinar cuál de los modelos presentaba el mejor rendimiento en función de las métricas de calidad, y se seleccionó el modelo más adecuado para su posterior uso en la clasificación de nuevos textos. Este enfoque de modelado y evaluación permitió tomar decisiones informadas sobre la elección del modelo que se ajusta mejor a los requisitos del proyecto.

Resultados

En la fase de resultados, se llevó a cabo una exhaustiva evaluación de los modelos de clasificación de textos implementados. Tras aplicar distintos algoritmos, entre ellos árboles de decisión, bosques aleatorios, máquinas de soporte vectorial (SVM) y k

vecinos más cercanos (KNN), se obtuvieron valiosas métricas de calidad que proporcionan una visión detallada del rendimiento de estos modelos. El algoritmo que arrojó los mejores resultados fue la SVM con kernel RBF, utilizando la representación TF-IDF de los textos. Este modelo alcanzó una puntuación F1 de 0.98, lo que indica una alta precisión y capacidad de recuperación en la clasificación de textos en las categorías deseadas. Estas métricas respaldan la elección de este modelo como la solución más adecuada para la tarea de clasificación de textos.

Los resultados también se exportaron en un nuevo archivo, lo que permitirá a la organización utilizar estos resultados para la clasificación de nuevos textos de manera eficaz y precisa. La información generada a partir de estos modelos puede ser de gran utilidad para la organización en la toma de decisiones informadas y en la optimización de sus procesos relacionados con la categorización de textos. La precisión y el rendimiento del modelo elegido brindan una herramienta valiosa que puede mejorar la eficiencia y la efectividad de la organización en la gestión y clasificación de textos en el contexto de sus objetivos comerciales.

Mapa de actores

En la siguiente tabla se presentan los actores relacionados:

Rol dentro de la empresa	**Tipo de actor**	**Beneficio**	**Riesgo**
Fondo de Poblaciones de las Naciones Unidas (UNFPA)	Usuario-cliente	El modelo simplificará el trabajo de seguimiento y evaluación de políticas públicas relacionadas con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Esto permitirá a la organización tomar decisiones más fundamentadas y eficaces.	Si el modelo no tiene un buen desempeño, la información clasificada podría ser incorrecta, lo que podría llevar a decisiones erróneas y una pérdida de tiempo y recursos en la implementación de políticas inadecuadas.
Entidades gubernamentales	Usuario-Cliente	Obtendrán datos más precisos y análisis relacionados con las políticas públicas que están implementando para los ODS 3, 4 y 5. Esto les ayudará a tomar decisiones más	Los resultados inexactos del modelo podrían llevar a decisiones subóptimas en la implementación de políticas, lo que afectaría negativamente a la comunidad.

		informadas y a mejorar la calidad de sus programas.	
Comunidades locales	Beneficiario	El enfoque basado en datos mejorará la calidad de vida en sus comunidades al abordar problemas de salud, educación y género de manera más efectiva.	Si el modelo no funciona correctamente, las políticas pueden no abordar adecuadamente las necesidades reales de la comunidad, lo que podría resultar en un impacto insatisfactorio.
Académicos	Usuario	Acceso a datos clasificados y análisis que pueden respaldar investigaciones académicas relacionadas con los ODS 3, 4 y 5, lo que facilita la generación de conocimiento en estas áreas.	Afectación sicológica que lo lleva a bajar el nivel académico dado que fue alertado o acompañado de forma incorrecta.
Desarrolladores y científicos de datos (Nosotros)	Proveedor	Contribuirán al desarrollo y mejora continua del modelo de clasificación de textos, lo que podría resultar en beneficios financieros y de reputación. Su experiencia en la creación y ajuste del modelo es fundamental para su éxito.	Si el modelo no funciona correctamente, podrían enfrentar críticas por su contribución al proyecto y riesgos de pérdida de oportunidades de negocio.
Grupo de estadística	Colaborador	Su experiencia en estadísticas contribuirá a la validación y mejora del modelo de clasificación de textos. Ayudarán a garantizar que los métodos	Si el modelo no se valida adecuadamente desde una perspectiva estadística, existe el riesgo de obtener resultados

		estadísticos utilizados sean sólidos y confiables.	incorrectos o sesgados.

Trabajo en equipo

Roles para el proyecto:

- **Líder de proyecto**: Juan Camilo González
  Está a cargo de la gestión del proyecto. Define las fechas de reuniones, pre-entregables del grupo y verifica las asignaciones de tareas para que la carga sea equitativa. Se encarga de subir la entrega del grupo. Si no hay consenso sobre algunas decisiones, tiene la última palabra.
- **Líder de negocio**: Laura Daniela Arias
  Es responsable de velar por resolver el problema o la oportunidad identificada y esta alineado con la estrategia del negocio para el cual se plantea el proyecto.
- **Líder de datos**: Carlos Julio Pinto
  Se encarga de gestionar los datos que se van a usar en el proyecto y de las asignaciones de tareas sobre datos. Debe dejarlos disponibles para todo el grupo.
- **Líder de analítica**: Juan Camilo González
  Se encarga de gestionar las tareas de analítica del grupo. Se encarga de verificar que los entregables cumplen con los estándares de análisis y que se tiene el "mejor modelo" según las restricciones existentes.

Se realizaron las siguientes reuniones de grupo:

- **Reunión de lanzamiento y planeación:** Se realizo esta reunión el día lunes 9 de octubre de 2023. La reunión, con duración de 40 minutos, consistió en definir los roles y la forma de trabajo del grupo. Se acuerdan las fechas de reuniones y pre-entregables del grupo, además de los métodos de comunicación para el desarrollo del proyecto.
- **Reunión de ideación:** Se realizo esta reunión el día miércoles 11 de octubre de 2023. La reunión duró 20 minutos. Se discutieron y se definieron las organizaciones, empresas o instituciones que se beneficiarán de la solución analítica desarrollada por el grupo.
- **Reunión con el grupo de estadística:** Se realizo esta reunión el día viernes 13 de octubre de 2023. La reunión, con duración de 30 minutos, consistió en presentar el problema al grupo de estadística y definir como apoyarían este proyecto.
- **Reuniones de seguimiento:** Se realizaron diferentes informes de seguimiento del proyecto mediante mensajes de texto. En estos mensajes se informaba de cada una de las tareas que realizaba cada uno de los integrantes del grupo. Sirven para mantener un registro constante del avance del proyecto y para asegurarse de que las tareas se completen de manera eficiente.
- **Reunión de finalización:** Se realizo esta reunión el día domingo 15 de octubre de 2023. La reunión, con duración de 2 horas, consistió en una verificación del trabajo realizado por cada uno de los estudiantes en el proyecto, además se discutieron los posibles puntos de mejora para la siguiente etapa. En la reunión, se consolidó todo el trabajo realizado. Una vez que se llegó a un

consenso sobre la completitud y la satisfacción con los resultados, se realizó la entrega del proyecto.