Etapa 1 Proyecto Inteligencia de Negocios

Integrantes

Daniel Pedroza 202123283
Miguel Gomez 202122562
Pablo Martinez 202122937

Entendimiento del Negocio y Enfoque Analítico

Problema/Oportunidad del Negocio:
El Fondo de Poblaciones de las Naciones Unidas (UNFPA) enfrenta el reto de identificar y abordar los problemas relacionados con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5. Estos ODS están centrados en la salud, educación, y equidad de género, y el UNFPA busca evaluar las opiniones de los ciudadanos sobre estos temas para guiar el desarrollo de políticas públicas efectivas.

La oportunidad radica en utilizar técnicas avanzadas de analítica de textos para extraer información relevante de las opiniones de los ciudadanos, facilitando la toma de decisiones basada en datos. Este enfoque permitirá a las entidades relacionadas con los ODS evaluar problemas clave a nivel local y formular soluciones más precisas y efectivas.

Con esto dicho, se realizan técnicas de analítica de textos para extraer información relevante de las opiniones de los ciudadanos, facilitando la toma de decisiones basada en datos. Este enfoque permitirá a las entidades relacionadas con los ODS evaluar problemas clave a nivel local y formular soluciones más precisas y efectivas.

Objetivos y Criterios de Éxito:
El principal objetivo de esta etapa del proyecto es desarrollar un modelo analítico que permita:

- Identificar automáticamente las opiniones de los ciudadanos que se relacionan con los ODS 3, 4 y 5.
- Evaluar y entender las problemáticas locales relacionadas con la salud, educación y equidad de género.
- Generar insights accionables para el UNFPA y otras organizaciones involucradas, con el fin de crear estrategias más efectivas y centradas en la mejora de los ODS.

Los criterios de éxito incluyen:

- La precisión del modelo para clasificar las opiniones correctamente en relación con los ODS seleccionados.
- La utilidad de los resultados obtenidos para las organizaciones beneficiarias, proporcionando datos claros y relevantes a la hora de tomar decisiones.
- La capacidad del modelo para mejorar la formulación de políticas basadas en los datos obtenidos de las opiniones de los ciudadanos.

Organización y rol que se beneficia:
La organización beneficiaria de este proyecto es el Fondo de Poblaciones de las Naciones Unidas (UNFPA). Los beneficiarios dentro del UNFPA incluyen:

- Analistas de políticas, que pueden utilizar los resultados para desarrollar y adoptar políticas públicas

- Directores de proyectos, van a guiar las estrategias a corto y largo plazo basándose en los insights generados por el modelo.

Aparte de esto el modelo le dará información valiosa a otras organizaciones terceras y entidades colaboradoras, ayudando a identificar áreas críticas de los diversos procesos, guiando en acciones concretas en pro de los ODS.

Impacto en Colombia
El impacto de este proyecto en Colombia será significativo, ya que permitirá a las organizaciones interpretar y utilizar las opiniones de los ciudadanos de manera más eficiente para guiar el desarrollo de políticas públicas. Al relacionar automáticamente las opiniones con los ODS 3, 4 y 5, el modelo ayudará a:

- Identificando áreas críticas en salud, educación y equidad de género. Problemáticas que requieren atención y acción inmediata.
- Promover un desarrollo sostenible al asegurar que las políticas públicas sean tomadas desde un punto de información para focalizar problemas claves identificados por la población.
- Aumentar la efectividad de las intervenciones públicas, esto asegura que los recursos se asignen en áreas que requieren de mayor atención y necesidades, por lo que los resultados tendrán un impacto real en el bienestar de la población.

Enfoque Analítico:
El enfoque que se emplea para abordar el problema es de analítica predictiva, específicamente centrado en la clasificación de texto. La tarea principal consiste en clasificar automáticamente las opiniones de los ciudadanos y relacionarlas con los Objetivos de Desarrollo Sostenible (ODS) 3, 4 y 5, que abarcan temas como la salud, la educación y la equidad de género.

Categoria de Analisis:
Como se mencionó previamente la categoría de análisis es predictiva, esto se basa en el hecho de que el objetivo principal es predecir la categoría (ODS) a la que pertenece una opinión o reseña de un contenido textual. Al ser predictivo se identificarán patrones y características clave en las opiniones públicas para mejorar la toma de decisiones y políticas públicas.

Tarea de Aprendizaje:
La tarea de aprendizaje es de clasificación supervisada, esto viene del hecho que los modelos aprenden a partir de datos etiquetados, siendo las opiniones categorizadas en ODS 3,4, y 5. Esto se utiliza para predecir la categoría de nuevas opiniones no etiquetadas.

Técnicas y Algoritmos Propuestos:
Las técnicas y algoritmos seleccionados son:

- Support Vector Machine(SVM): Este es un algoritmo de clasificación que se utiliza para maximizar el margen entre diferentes clases. En nuestro caso es útil ya que maneja bien las clases separables linealmente como aquellas que no lo son.

- Regresión Logística: Este algoritmo se utiliza para modelar la relación entre las características de las opiniones y la probabilidad de pertenecer a una categoría específica de ODS.

Entendimiento y preparación de datos

Se nos entregó un archivo Excel con dos columnas. La primera columna, llamada "textos_espanol", contiene comentarios en español que las personas han realizado sobre los Objetivos de Desarrollo Sostenible (ODS), específicamente los ODS 3, 4 y 5. La segunda columna, denominada "sdg", contiene valores numéricos entre 3 y 5, que indican a cuál objetivo corresponde cada comentario.

Los objetivos trabajados son:

- Buena Salud y Bienestar (3)
- Educación de Calidad (4)
- Igualdad de Género (5)

Al revisar los datos en la columna "textos_espanol", detectamos errores significativos en palabras que contenían caracteres diacríticos, como tildes y la letra "ñ". Estos errores resultan en la aparición de caracteres extraños, por ejemplo, "comparación" en lugar de "comparación". A continuación, realizamos una búsqueda de valores nulos en los datos, la cual arrojó un 0% de nulos, lo que implica que no había datos faltantes ni vacíos que requirieron manejo adicional.

Posteriormente, verificamos el tipo de datos en ambas columnas. La columna "textos_espanol" fue identificada con el tipo object, mientras que la columna "sdg" fue correctamente detectada como int.

Preparación de Datos

El proceso de limpieza de datos comenzó reemplazando el símbolo "%" por la palabra "porciento". Esta decisión se tomó debido a la presencia de números en los comentarios que incluían porcentajes, como en el caso específico de "15.3%". Si solo eliminamos el símbolo de porcentaje, se perdería el contexto del número, lo cual afectaría el análisis.

A continuación, se corrigieron los errores de codificación relacionados con los caracteres diacríticos. Esto se logró reemplazando las palabras que contenían estos caracteres extraños por sus equivalentes sin acentos ni tildes (por ejemplo, cambiando "á" por "a", y "ñ" por "n"). Luego, se eliminaron los caracteres ASCII especiales, se convirtieron todas las palabras a minúsculas y se eliminó la puntuación.

Después, se manejaron los números escritos en los textos, convirtiéndolos en su forma textual. Posteriormente, se eliminaron las stopwords (palabras vacías) con el fin de reducir el tamaño de los datos y evitar palabras genéricas que no aportan valor al análisis.

Tokenización, lematización y stemming

El siguiente paso fue realizar la tokenización de las palabras, generando una nueva columna con los tokens resultantes. A continuación, se llevó a cabo la normalización de los datos mediante lematización y stemming, lo que permitió reducir las palabras a su forma base. Este proceso mejora la consistencia de los datos y facilita su análisis.Finalmente, se creó una nueva columna con las palabras modificadas tras los procesos de limpieza, tokenización y normalización.

Vectorización

Para preparar los datos para el modelo, se utilizó la técnica de vectorización TF-IDF (Term Frequency-Inverse Document Frequency). Esta técnica mide la importancia de una palabra en un documento en relación con el corpus completo. Ayuda a identificar las palabras más relevantes asignando un peso a aquellas que son más importantes, lo que permite una mejor representación y análisis por parte del modelo.

Modelado y Evaluación

Como se menciono en la seccion de enfoque analitico se trabajaron tres algoritmos:

- Support Vector Machine
- Regresion Logistica
- Random Forest

Support Vector Machine (SVM)

El algoritmo SVM funciona para clasificar datos en múltiples clases, incluso cuando estas no son linealmente separables. En nuestro caso, las opiniones relacionadas con los ODS son complejas y pueden no seguir patrones lineales claros. SVM es ideal para esta situaciones, ya que permite encontrar un hiperplano óptimo que separa las clases de manera efectiva, maximizando el margen entre los diferentes grupos.

Además, le permite transformar los datos en un espacio de mayor dimensión donde las clases sean más fáciles de separar. Este aspecto es útil para la clasificación de texto, donde las palabras y frases pueden tener interacciones complejas. Por estas razones, SVM fue una elección natural para este problema.

Funciona, al buscar la mejor línea o plano que separe las clases de datos de la manera más clara posible, maximizando la distancia entre los puntos más cercanos de cada clase. Si los datos no se pueden separar linealmente, SVM transforma el espacio para hacerlo posible, usando un kernel

Regresión Logística

La Regresión Logística es un algoritmo de clasificación que es eficiente y fácil de interpretar. Su capacidad para modelar probabilidades y asignar observaciones a clases con base en estas probabilidades lo hace una opción adecuada para problemas de clasificación multiclase, como el que enfrentamos con los ODS.

Este algoritmo es especialmente útil cuando las relaciones entre las características (en este caso, las palabras en los textos) y las clases son principalmente lineales. Aunque puede no ser tan flexible como SVM para manejar relaciones no lineales, su simplicidad y eficiencia lo convierten en una excelente solución para cualquier problema de clasificación.

Funciona, ya que asigna una probabilidad a cada clase basándose en las características de los datos. Utiliza una fórmula matemática que transforma la información en probabilidades y luego clasifica la observación en una clase en función de estas probabilidades.

Random Forest

Random Forest es un algoritmo de ensamblaje que combina múltiples árboles de decisión, lo que le permite capturar relaciones no lineales en los datos de manera robusta. Se eligió este modelo debido a su capacidad para manejar conjuntos de datos con gran cantidad de características (como es el caso con los datos textuales que hemos vectorizado) y su baja susceptibilidad al sobreajuste.

Otra ventaja clave de Random Forest es que proporciona interpretabilidad mediante la importancia de las características. En nuestro caso, esto permite identificar qué palabras o términos son más relevantes para la clasificación de las opiniones, lo que aporta valor adicional al análisis. Por su versatilidad y capacidad para manejar tanto datos lineales como no lineales, Random Forest fue elegido como uno de los modelos principales para este ejercicio.

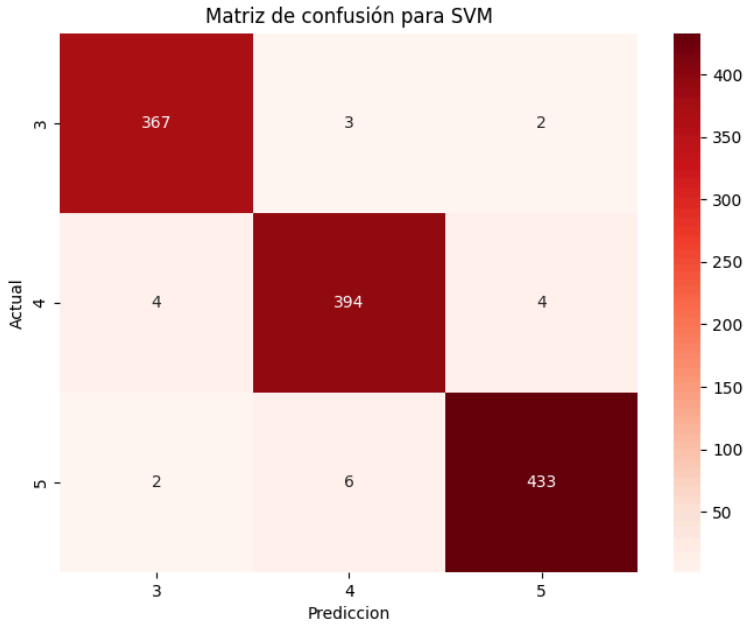
La forma en la que funciona es que construye muchos árboles de decisión pequeños y toma la decisión final por votación mayoritaria. Al promediar las predicciones de múltiples árboles, mejora la precisión y reduce el riesgo de errores debido a ruido en los datos.

Resultados

Descripción de los resultados obtenidos

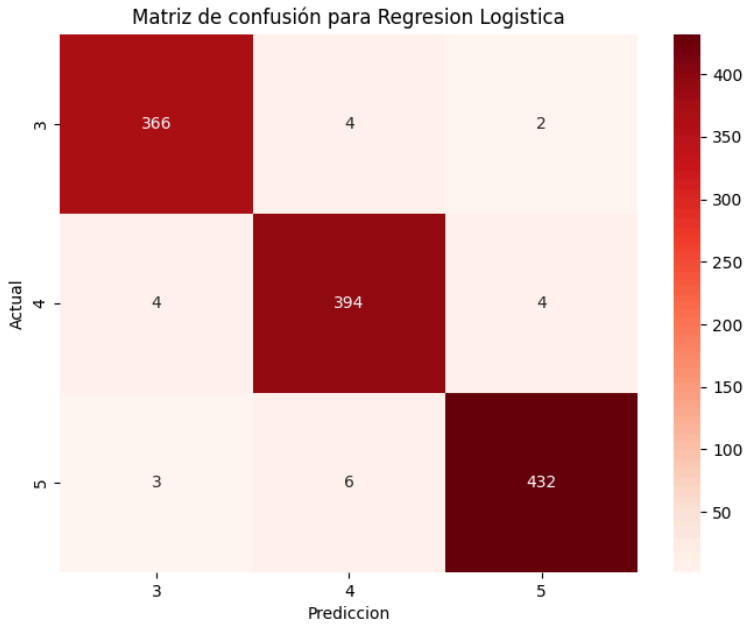
Los resultados de los modelos analíticos, mostraron métricas de precisión, recall y F1-Score con una consistencia alta, esto indica que se logra una clasificación precisa de las opiniones relacionadas con los Objetivos de Desarrollo Sostenible 3,4 y 5.
Los valores de precisión y recall en los tres modelos se acercan al 98% en la mayoría de las clases. Esto, indica que los modelos son altamente efectivos para identificar las opiniones correctas en cada categoría.

Support Vector Machine (SVM): Este mostró una precisión del 98% en las clases 3 y 4, y del 99% en la clase 5. El recall en todas las clases fue igualmente alto, esto asegura que la mayoría de las opiniones en los datos fueron clasificadas correctamente.



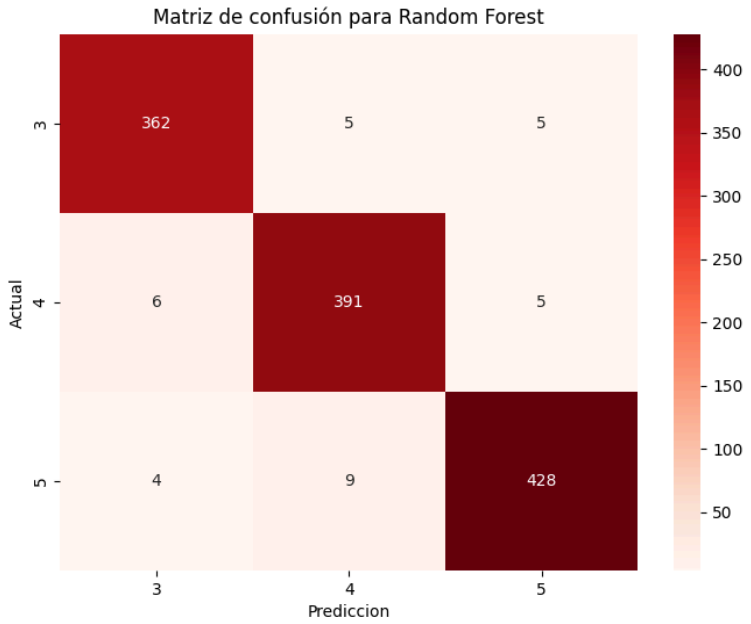
¿Qué nos dice la matriz de confusión?
La matriz de confusión para el modelo Support Vector Machine (SVM) nos muestra que, de los 372 ejemplos que pertenecen a la clase 3 (Buena Salud y Bienestar), 367 fueron correctamente clasificados, mientras que 3 fueron incorrectamente clasificados como clase 4 y 2 como clase 5. En el caso de la clase 4 (Educación de Calidad), de 402 ejemplos, 394 fueron clasificados correctamente, con 4 clasificaciones erróneas como clase 3 y 4 como clase 5. Finalmente, para la clase 5 (Igualdad de Género), de 441 ejemplos, 433 fueron correctamente clasificados, mientras que 6 fueron clasificados como clase 4 y 2 como clase 3.

Regresión Logística: Este alcanzó resultados similares a los que demostró SVM, teniendo una precisión general del 98%, esto implica una muy buena opción a la hora de realizar la tarea de clasificación multiclase. El algoritmo se basa en el concepto de márgenes y vectores de soporte, este modelo genera la relación de la clase específica, permitiendo una mayor interpretabilidad sobre las preocupaciones cambiantes de los ciudadanos en relación con los ODS.



¿Qué nos dice la matriz de confusión?
La matriz de confusión para el modelo de Regresión Logística nos muestra que para la clase 3 (Buena Salud y Bienestar), de 372 ejemplos, 366 fueron correctamente clasificados, mientras que 4 fueron incorrectamente clasificados como clase 4 y 2 como clase 5. En la clase 4 (Educación de Calidad), de 402 ejemplos, 394 fueron clasificados correctamente, con 4 errores al ser clasificados como clase 3 y 4 como clase 5. Finalmente, para la clase 5 (Igualdad de Género), de 441 ejemplos, 432 fueron clasificados correctamente, mientras que 6 fueron clasificados erróneamente como clase 4 y 3 como clase 3.

Random Forest: Con una precisión del 97%, se vio un rendimiento ligeramente inferior a las otros, pero continúa mostrando resultados significativos por lo que se puede considerar un modelo fiable. Este permite medir la importancia de las características lo que significa que el modelo indica cuales son las palabras o características más relevantes a la hora de clasificar las opiniones.



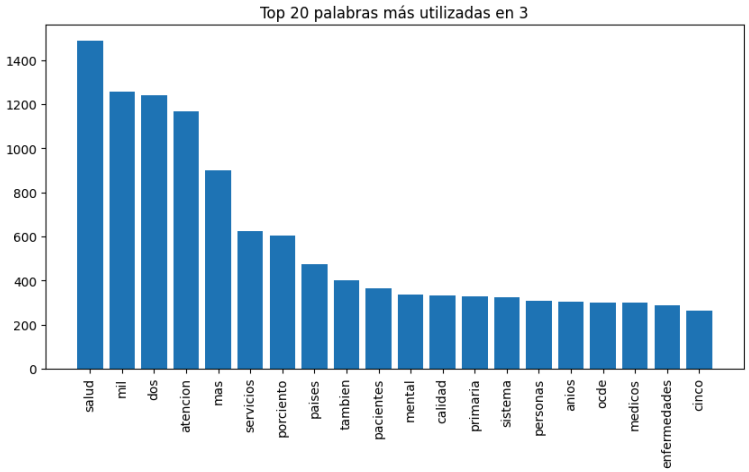
Qué nos dice la matriz de confusión?
La matriz de confusión para el modelo de Regresión Logística nos muestra que para la clase 3 (Buena Salud y Bienestar), de 372 ejemplos, 362 fueron correctamente clasificados, mientras que 5 fueron clasificados incorrectamente como clase 4 y 5 como clase 5. Para la clase 4 (Educación de Calidad), de 402 ejemplos, 391 fueron clasificados correctamente, con 6 errores como clase 3 y 5 como clase 5. Finalmente, para la clase 5 (Igualdad de Género), de 441 ejemplos, 428 fueron correctamente clasificados, mientras que 9 fueron clasificados como clase 4 y 4 como clase 3.

Análisis de las palabras identificadas y estrategias para la organización

A partir de los gráficos que muestran las 20 palabras más utilizadas en los comentarios relacionados con los ODS 3 (Buena Salud y Bienestar), ODS 4 (Educación de Calidad) y ODS 5 (Igualdad de Género), se pueden extraer insights valiosos que orienten a la organización en el diseño de estrategias focalizadas y efectivas para abordar las preocupaciones de los ciudadanos.

ODS 3 (Buena Salud y Bienestar)

En el gráfico del ODS 3, destacan términos como "salud", "atención", "servicios", "pacientes" y "mental", lo que indica que las preocupaciones principales giran en torno a la calidad de los servicios de atención sanitaria, el acceso a la salud mental y la atención a los pacientes. Palabras como "países" y "sistema" también sugieren comparaciones con los servicios de salud en otros lugares y cómo se percibe la calidad del sistema de salud en general.



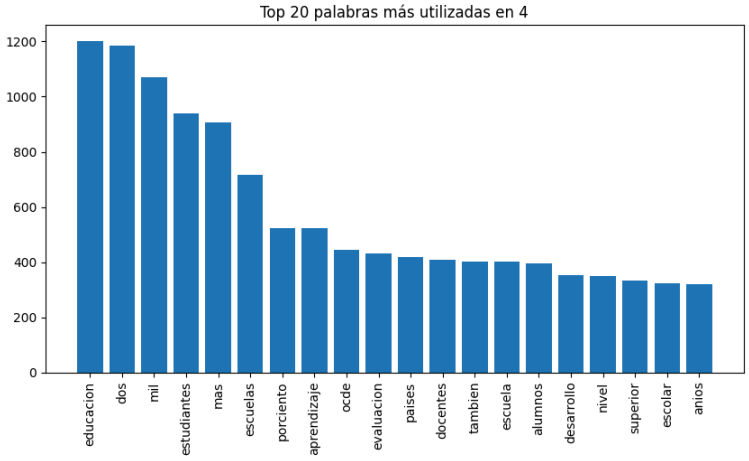
Estrategias para la Organización:

- Mejorar los servicios de atención sanitaria: El foco en "atención" y "servicios" indica la necesidad de mejorar la eficiencia y calidad en la prestación de servicios médicos, incluyendo la atención a la salud mental, que es otra preocupación frecuente.

- Fortalecer los sistemas de salud pública: La mención del "sistema" y "países" sugiere la necesidad de crear políticas que fortalezcan el sistema de salud local en comparación con estándares internacionales.

ODS 4 (Educación de Calidad)

Las palabras más recurrentes en el gráfico del ODS 4 incluye "educación", "estudiantes", "escuelas" y "docentes", lo que indica que las preocupaciones giran en torno a la infraestructura educativa, el acceso a una educación de calidad y la capacitación de los docentes. Términos como "evaluación" y "OCDE" reflejan un interés en mejorar la evaluación educativa en comparación con estándares globales.



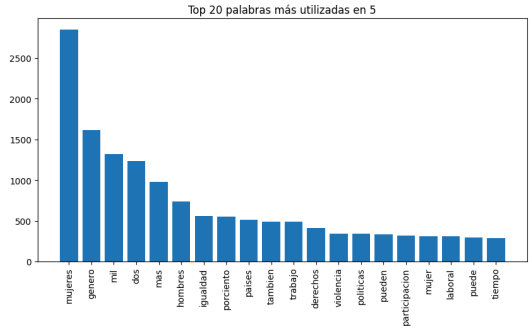
Estrategias para la Organización:

- Mejorar la infraestructura educativa y el acceso a escuelas para asegurar que más estudiantes puedan beneficiarse de un entorno adecuado para el aprendizaje.

- Fortalecer la capacitación de docentes y mejorar los métodos de evaluación educativa, asegurando que el sistema se alinee con las mejores prácticas internacionales.

ODS 5 (Igualdad de Género)

En el gráfico del ODS 5, términos como "mujeres", "género", "igualdad", "hombres", "violencia" y "trabajo" destacan como las principales preocupaciones. Esto refleja que la equidad de género, la lucha contra la violencia de género y la igualdad laboral son temas prioritarios para los ciudadanos.



Estrategias para la Organización:

- Promover políticas contra la violencia de género y fortalecer las leyes que protegen a las mujeres, dado que "violencia" es una de las preocupaciones clave.

- Fomentar la igualdad laboral entre hombres y mujeres mediante la creación de programas que promuevan la igualdad de oportunidades en el trabajo y en otras esferas sociales.

Mapa de actores relacionado con el producto de datos creado

Rol dentro de la organización	Tipo de actor	Beneficio	Riesgo
Ministerio de Salud	Usuario-cliente	Utiliza el modelo para identificar preocupaciones de los ciudadanos sobre la	Si el modelo no clasifica correctamente, se podrían ignorar temas críticos o

		calidad de los servicios de salud.	centrarse en áreas menos relevantes.
Ministerio de Educación	Usuario-cliente	El modelo ayuda a identificar problemas clave en la infraestructura educativa y en la calidad de la enseñanza.	En caso de errores en la clasificación, los recursos podrían no enfocarse en las áreas más importantes.
ONGs enfocadas en Igualdad de Género	Usuario-cliente	El modelo permite priorizar programas basados en las preocupaciones sobre igualdad de género y violencia.	Si el modelo tiene un mal desempeño, las inversiones pueden no lograr los resultados esperados o ser mal dirigidas.
Ciudadanos	Beneficiado	Sus opiniones son escuchadas y consideradas en la creación de políticas más efectivas.	Si el modelo falla, se podrían ignorar sus preocupaciones reales, afectando la confianza en las instituciones.
Organizaciones Internacionales (ONU, OCDE)	Financiador	Utilizan los resultados del modelo para generar informes globales y guiar políticas a nivel internacional.	Si los datos locales no son precisos, las políticas internacionales podrían basarse en información incorrecta.

Trabajo en equipo

En el desarrollo de este proyecto, los roles fueron distribuidos de la siguiente manera:

Líder de Proyecto: Daniel Pedroza

Daniel fue responsable de la gestión general del proyecto. Se encargó de definir las fechas de reuniones y los pre-entregables del grupo, asegurando que las asignaciones fueran equitativas entre los miembros. Además, verificó el progreso del equipo y fue responsable de subir la entrega final. A lo largo del proyecto, Daniel dedicó aproximadamente 9 horas a estas tareas. Los retos principales que enfrentó fueron la coordinación de los tiempos de trabajo de todos los miembros y la resolución de desacuerdos en cuanto a la asignación de recursos y tiempos. Para resolverlos, Daniel mantuvo comunicación constante por el servidor de discord creado y el grupo de whatsapp, asegurando que cada integrante cumpliera su parte. Aparte de esto ayudo con una gran parte del desarollo del documento y genero diversas adiciones al jupyter notebook a la hora de generar los modelos.

Líder de Analítica: Miguel Gómez

Miguel asumió la responsabilidad de gestionar las tareas relacionadas con el análisis de datos. Miguel dedicó unas 15 horas al análisis de datos y la verificación de los modelos. Los principales desafíos enfrentados por Miguel fueron lograr un modelo que equilibrara precisión y simplicidad. Para abordar este problema, trabajó en iteraciones del modelo y ajustó los algoritmos seleccionados para optimizar el rendimiento. Fue el desarollador principal del codigo y los modelos realizados en el jupyter notebook, por lo que evidentemente fue una parte crucial a la hora de desarollar el modelo propuesto, cumpliendo con los objetivos del negocio en el proceso.

Líder de Negocios y Datos: Pablo Martínez

Pablo fue el encargado de garantizar que el enfoque del proyecto estuviera alineado con los objetivos del negocio. Estuvo a cargo de comunicar el valor del producto analítico al negocio. Pablo dedicó 6 horas a esta tarea, enfocado en identificar oportunidades de mejora desde una perspectiva de negocio y en ajustar la solución analítica para que responda efectivamente al problema planteado. El reto principal de Pablo fue garantizar que las soluciones técnicas fueran comprensibles para las partes interesadas no técnicas. Solucionó este problema mediante la creación de informes de resultados claros y concisos. Pablo fue una parte crucial en el desarollo del documento, buscando priorizar la claridad a la hora de exponer los algoritmos implementados y los beneficios que le generan al negocio, priorizando el cumplimiento de los objetivos del negocio, buscando enfatizar en soluciones que ayudan a los diversos actores del negocio.

Repartición de Puntos

Basado en la contribución de cada miembro del equipo, se decidió repartir los 100 puntos de la siguiente manera:
Daniel Pedroza (Líder de Proyecto): 30 puntos
Miguel Gómez (Líder de Analítica): 40 puntos
Pablo Martínez (Líder de Negocios y Datos): 30 puntos
Reuniones

- **Reunión de planeamiento y preparación:** Definición de roles, objetivos y estrategia general del proyecto.
- **Reunión de revisión sobre lo discutido en el planeamiento:** Evaluación de avances respecto a los acuerdos del plan inicial.
- **Reunión de distribución de tareas:** Asignación específica de responsabilidades entre los miembros.
- **Reunión de revisión de entrega:** Verificación final del trabajo antes de la entrega del proyecto.

Reflexiones y Mejoras

En el desarrollo del proyecto, un área de mejora identificada es la organización de las reuniones presenciales y virtuales. Aunque el equipo se mantuvo en contacto a través de medios de comunicación digitales como mensajería instantánea y correos electrónicos, estos no siempre fueron suficientes para resolver problemas complejos o asegurar una alineación

completa entre los miembros. Se recomienda aumentar la frecuencia de las reuniones presenciales o virtuales en tiempo real, ya que estas permiten una mayor interacción, discusión más profunda de los desafíos y una toma de decisiones más rápida y efectiva. Establecer una combinación equilibrada entre reuniones virtuales y presenciales puede mejorar significativamente la coordinación y reducir la posibilidad de malentendidos en la ejecución del proyecto.