Proyecto 1

Inteligencia de Negocios

Etapa 1

Grupo 30

German Alberto Rojas Cetina 202013415

María Paula Almeciga Moreno 202023369

Juan Diego Sarmiento Sánchez 202121484

Contenido

Sección 1. Entendimiento del negocio y enfoque analítico.

Oportunidad/problema Negocio.

El proyecto se encuentra enfocado para la empresa de El Fondo de Poblaciones de las Naciones Unidas (UNFPA). Esta empresa es la agencia de la ONU para la salud sexual y reproductiva. Esta empresa Apoya a Colombia en la producción y utilización de datos socio-demográficos que sirven de base para la formulación de políticas de desarrollo sostenible y programas de reducción de la pobreza y construcción de paz, considerando aspectos humanitarios. [1]

En concreto el proyecto se enfocó en los Objetivos de Desarrollo Sostenible (ODS), los cuales consisten en objetivos globales los cuales tienen como objetivo un llamamiento universal para poner fin a la pobreza, proteger el planeta y garantizar que para el 2030 todas las personas disfruten de paz y prosperidad. [2] Existen en total 17 ODS cada uno tratando sobre un tema diferente, los cuales podemos ver a continuación:

Ilustración 1 17 ODS

Dentro del proyecto se concentrará sobre los ODS de salud y bienestar (3), Educación de calidad (4), Igualdad de género (5). Se realizan encuestas y fórums para recopilar información sobre la opinión y experiencia de las personas sobre el tema.

Objetivos y criterios de éxito desde el punto de vista del negocio.

El objetivo de la primera etapa del proyecto es a partir de las encuestas realizadas y de los comentarios para entrenar a un modelo de inteligencia artificial para estimar que tipo es la consulta respecto al ODS. Puede procesar el texto natural introducido por los usuarios y

estimar que tipo es el comentario recibido. En específico que se desea poder identificar si pertenece al grupo 3,4 o 5 de los ODS el comentario brindado por la persona. Esto a partir de comentarios previos ya etiquetados.

Para obtener el éxito dentro del proyecto se espera poder obtener un modelo que pueda realizar esta estimación a del lenguaje natural en español. Teniendo en cuenta los errores y casos que se puedan presentar a la hora de presentar el texto, como frases mal escritas o errores en la codificación de texto, etc… El modelo debe ser adecuado y preciso para esta estimación. En concreto, el modelo debe tener una relevancia dentro de la estimación con un valor aceptable de error.

Organización y rol dentro de ella que se beneficia con la oportunidad definida.

La principal organización beneficiada por el proyecto, es el UNFPA1. Gracias a que maneja el control de los ODS y ayuda a poder cumplir con objetivos deseados en los propios ODS. Dentro de la propia organización se pueden ver varios roles beneficiados por este proyecto, entre estos podemos ver varios roles tomados de [3]:

- Asistentes de seguridad local: El rol debe brindar apoyo con la implementación de políticas de seguridad y procedimientos. A partir del modelo entrenado se pueden manejar los comentarios de las personas de manera más rápida. Además, pueden tener de forma más sencilla una perspectiva amplia y completa de los ODS más frecuentes y sus características generales. Permitiendo tomar decisiones a partir de esto.
- Personal general de la UNPFA: El rol tiene trabajo de cumplir con todas las políticas de seguridad, procedimientos, directrices y directivas del UNSMS y del UNFPA. El manejo de los ODS puede facilitarle el trabajo al no tener que revisar manualmente tantos comentarios y poder tener información adicional brindada por el propio modelo.
- Representantes del UNFPA: El rol es responsable y rinde cuentas ante el Secretario General a través del Director Ejecutivo e informar a su director regional para la seguridad del UNFPA. El modelo facilita la realización de los reportes para identificar los ODS y dar datos sobre estos. Para poder realizar estrategias y políticas.

Adicionalmente, cualquier organización que desee utilizar los ODS puede beneficiarse del modelo. Al poder identificar los ODS más urgentes y necesitados por las personas y poder tener estrategias y políticas mejores para el desarrollo sostenible.

Impacto que puede tener en Colombia este proyecto.

Gracias que la UNPFA está en asociación con Colombia se pueden ver el impacto directo dentro de esta. Gracias que a partir de los datos obtenidos con el modelo se puede ver las necesidades y las falencias que las personas están pasando en estos momentos. Además, poder ver las más críticas al afectar a una mayor cantidad de personas en el país. Lo cual le daría información a la UNPFA para que esta pueda promover y ayudar con diferentes estrategias y políticas a implementar dentro del país.

En general esta información también se puede utilizar por el propio gobierno colombiano y demás organizaciones interesadas en el desarrollo sostenible colombianos. Donde se pueden tener una mejor perspectiva respecto a los ODS 3, 4, y 5. Sobre todo tener un enfoque adecuado para la salud, la igualdad de género y la educación de calidad.

Incluso a partir del modelo se pueden procesar una mayor cantidad de comentarios y poder identificar que estrategias están funcionando y si se están cumpliendo los objetivos planteados de estas mismas.

Enfoque analítico.

El enfoque analítico nos podrá guiar para la realización del proyecto, donde cada fase nos brinda información de que se debe realizar y como se debe realizar. Podemos ver el enfoque analítico en 4 fases:

1. Tipo de analítica:
   Predictiva: el objetivo del proyecto es ver el ODS que pertenece un texto. A partir del texto poder estimar y predecir a que ODS pertenece. Se identificarán patrones y palabras claves que ayudarán y nos brindarán estimaciones para obtener un ODS adecuado.
2. Tipo de aprendizaje:
   Supervisado: Para la realización del proyecto se posee una gran variedad de datos previamente tomados cada uno con el valor que le corresponde del ODS. Esto implica que se tienen ya etiquetas a los datos que se utilizarán para poder estimar la variable objetivo que son los ODS.
3. Tareas de aprendizaje:
   Clasificación: Gracias a que se tiene un tipo de aprendizaje supervisado, además de que la variable objetivo es discreta, categórica se puede usar esta tarea de aprendizaje. La clasificación es la ideal para poder estimar las categorías del ODS de datos que no posean etiquetas.
4. Algoritmo:
   a. Suport Vector Machine (SVM) es un algoritmo de clasificación que se basa en funciones lineales (generalmente) que dividen la categoría de los datos. Ideal para el caso gracias al requerir una separación clara de las clases.
   b. Random Forest es un algoritmo que combina el uso de múltiples árboles. Se puede utilizar en la clasificación de un texto a partir de los ODS, capaz de manejar múltiples características de texto mejorando el rendimiento en la clasificación.
   c. K-Nearest Neighbors (KNN) es un algoritmo de clasificación que funciona buscando los k datos más cercanos a un nuevo dato y eligiendo la etiqueta

más común entre ellos. Es útil para clasificar textos porque puede adaptarse a la variabilidad en los datos y no requiere un modelo complejo. Sin embargo, el número de vecinos k es importante, teniendo en cuenta que un k muy bajo puede hacer que el modelo sea muy sensible al ruido, mientras que un k muy alto puede hacer que el modelo sea menos preciso.

Sección 2. Entendimiento y preparación de los datos

Para iniciar el proceso de entendimiento y preparación de los datos, se leyó el archivo utilizando la biblioteca pandas y se asignó a una variable llamada "df". A continuación, se identificó el idioma de los textos con el fin de filtrar aquellos que aportaban valor al entrenamiento y evaluación de los modelos. Se detectó la presencia de textos en inglés y francés, los cuales fueron eliminados al no ser significativos en relación con la cantidad total de información del conjunto de datos.

Después, se verificó la ausencia de valores nulos y duplicados en el conjunto. Posteriormente, se realizó una limpieza exhaustiva de los textos, que incluyó la corrección de palabras mal codificadas, la eliminación de caracteres especiales y signos de puntuación, la tokenización de palabras y la normalización del texto. Esta normalización abarcó también la obtención de las raíces y lemas de cada palabra.

Finalmente, la columna "words" se unificó en una sola cadena de texto, y el nuevo DataFrame resultante se guardó en un archivo CSV. Todos estos pasos fueron esenciales para preparar adecuadamente los datos de cara a su análisis y modelado.

Sección 3. Modelado y evaluación.

SVM - Juan Diego Sarmiento Sánchez

Para empezar, Una máquina de vectores de soporte (de sus siglas en inglés, support vector machine SVM) es un modelo de aprendizaje automático capaz de realizar clasificación lineal o no lineal, regresión e incluso detección de novedades.[4]

El modelo se basa en el uso de las funciones para brindar márgenes en la separación y clasificación de los datos. Estos márgenes se pueden definir de diferentes maneras, ya sea que estos son estrictos o con un error asociado y aceptando valores mal clasificados. [4]

Cuando estos márgenes no son suficientes a causa de la naturaleza de los datos se pueden utilizar herramientas que ayudan al algoritmo a escalar en espacios n-dimensionales. Para este propósito se suele utilizar el kernel. Este consiste en una transformación de los datos de entrenamiento a partir de vector formado de estos mismos y un producto punto que me escala los datos a las dimensiones deseadas. Existen varios kernel, ya sean polinomiales o radiales, cada uno de estos con varios hiperparámetros asociados. [4]

Dentro del desarrollo se utilizará la búsqueda en grilla para realizar la validación cruzada a partir de una partición de 5, para obtener los mejores hiperparámetros asociados al modelo. Para luego utilizar la matriz de confusión y las estadísticas de “presision”, “recall” y el “f1 score”.

Random Forest - Germán Rojas

La elección de utilizar el algoritmo Random Forest se basó en su capacidad para manejar múltiples características de texto, lo cual es fundamental para la clasificación de textos en relación con los Objetivos de Desarrollo Sostenible (ODS). El modelo fue entrenado empleando la representación TF-IDF de los textos.

El proceso se desarrolló en varias etapas: primero, se dividieron los datos en conjuntos de entrenamiento y prueba; luego, se realizó la vectorización del texto utilizando TF-IDF; posteriormente, se entrenó el modelo y se visualizó la importancia de las características. Finalmente, se hicieron predicciones sobre ambos conjuntos, de entrenamiento y prueba.

Para evaluar el rendimiento del modelo, se utilizarán la matriz de confusión y métricas clave como la precisión, el recall y el F1-score.

KNN - María Paula Almeciga Moreno

K-Nearest Neighbors (KNN) es un algoritmo de clasificación supervisado no paramétrico que clasifica los datos en función de la proximidad a los puntos de datos cercanos en el espacio de características. A diferencia de otros algoritmos, KNN no requiere una fase de entrenamiento explícita, ya que simplemente almacena el conjunto de datos de entrenamiento. Cuando se necesita realizar una predicción, el algoritmo calcula la distancia entre el punto de consulta y todos los puntos de entrenamiento, utilizando métricas como la distancia euclidiana.

El valor de k es fundamental en KNN, pues define el número de vecinos más cercanos que se consideran para hacer una predicción. En clasificación, el punto de datos se asigna a la clase que aparece con mayor frecuencia entre estos k vecinos. Para problemas de regresión, KNN toma el promedio de los valores de los vecinos más cercanos para hacer la predicción. Elegir el valor adecuado de k es crucial, ya que valores bajos pueden llevar a un modelo sensible al ruido, mientras que valores altos pueden suavizar en exceso la clasificación.

El algoritmo KNN se ajusta bien a datos con pocas dimensiones y se utiliza en una variedad de aplicaciones, como sistemas de recomendación, detección de intrusos, y reconocimiento de patrones. Sin embargo, no escala bien con grandes volúmenes de datos y puede verse afectado por la "maldición de la dimensionalidad", donde la adición de más características puede degradar el rendimiento del modelo. Para mejorar su rendimiento, se puede utilizar la validación cruzada para seleccionar el valor óptimo de k y reducir el riesgo de sobreajuste. [5]

Sección 4. Resultados.

SVM - Juan Diego Sarmiento Sánchez

Los mejores parámetros:

-C (Penalización): 10

-Kernel usado: RBF

- Gamma: Scale

Al ver las matrices de confusión:

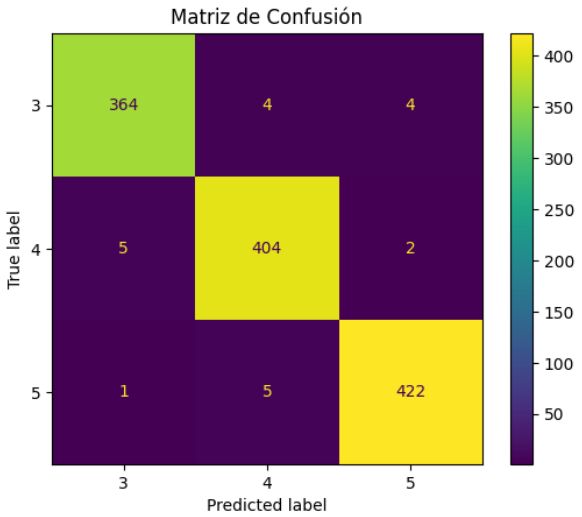

Ilustración 2 Matriz Confusión sin normalizar

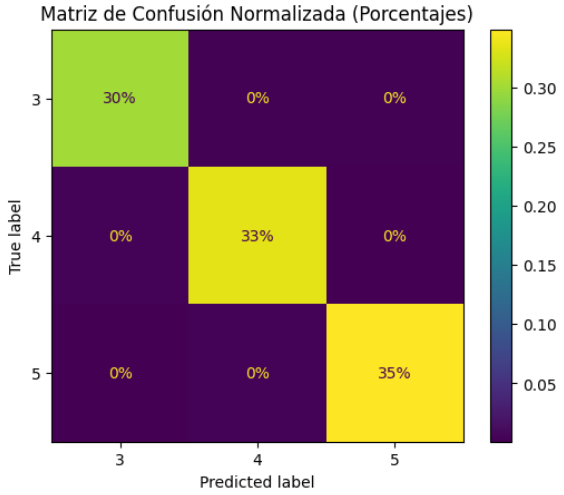

Ilustración 3 Matriz Confusión sin Normalizar

Se ve que el modelo es preciso al realizar las predicciones de los ODS. En general se tiene los valores precedidos adecuados para cada ODS, con porcentajes mayores a 30%. Si bien en la normalización los errores muestran 0% este valor en realidad no es exactamente 0, gracias que se tienen valores erróneos en la matriz sin normalizar. La matriz en general tiene un gran balanceo dentro de los datos evitando así sesgos y preferencia por alguna categoría.

Y las métricas:

Tabla 1 Resultados SVM

	precision	recall	f1-score	support
3	0.98	0.98	0.98	372
4	0.98	0.98	0.98	398
5	0.99	0.98	0.99	441
accuracy			0.98	1211
macro avg	0.98	0.98	0.98	1211
weighted av	0.98	0.98	0.98	1211

Random Forest - Germán Rojas

Reporte de clasificación de muestra

	precision	recall	f1-score	support
3	0.97	0.97	0.97	248
4	0.97	0.99	0.98	271
5	0.98	0.97	0.97	289
accuracy			0.97	808
macro avg	0.97	0.97	0.97	808
weighted avg	0.97	0.97	0.97	808

En el conjunto de entrenamiento, el modelo alcanzó una precisión del 100%, lo que indica que clasificó correctamente todos los casos en dicho conjunto. Sin embargo, en el conjunto de prueba, las métricas de precisión, recall y F1-score promediaron 0.98. Aunque este valor es ligeramente inferior al obtenido en el conjunto de entrenamiento, sigue reflejando un desempeño sólido.

Estas métricas sugieren que el modelo es muy efectivo en la identificación de textos relacionados con los ODS en el conjunto de prueba. Además, se llevó a cabo un proceso de validación cruzada (cross-validation) con 5 iteraciones, donde el modelo alcanzó un puntaje promedio de 0.969, lo que confirma su excelente desempeño en términos de precisión general para la predicción de clases.

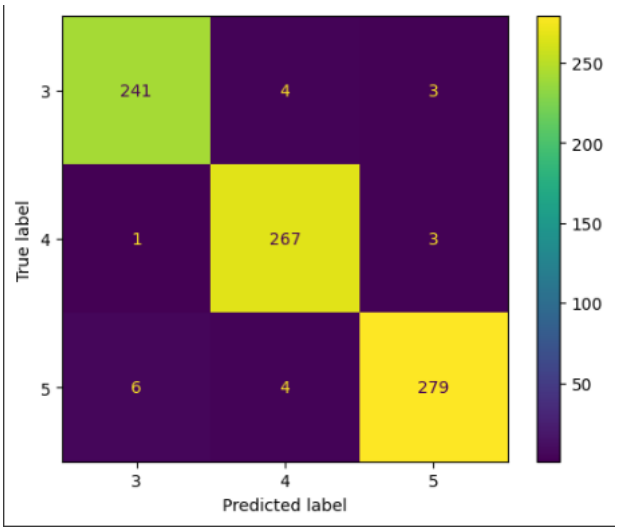

Ilustración 4 Matriz Confusión Random Forest

KNN - María Paula Almeciga Moreno

El reporte de clasificación muestra:

Clase	Precision	Recall	F1-score
ODS 3	0.93	0.97	0.95
ODS 4	0.94	0.94	0.94
ODS 5	0.95	0.92	0.94
Promedio (macro)	**0.94**	**0.94**	**0.94**

Se encuentra que el modelo alcanzó una precisión del 94%, pudiendo clasificar bien los textos con una alta precisión general. Se encuentra que el modelo tiende a confundir más entre ODS 4 y 5. Esto implica que aunque el modelo tiene un buen rendimiento, puede que no sea la mejor opción cuando la separación precisa entre estas fases es crítica.

La matriz de confusión igualmente muestra un buen desempeño general con un alto número de clasificaciones correctas en todas las clases. Sin embargo, aún así comete algunos errores.

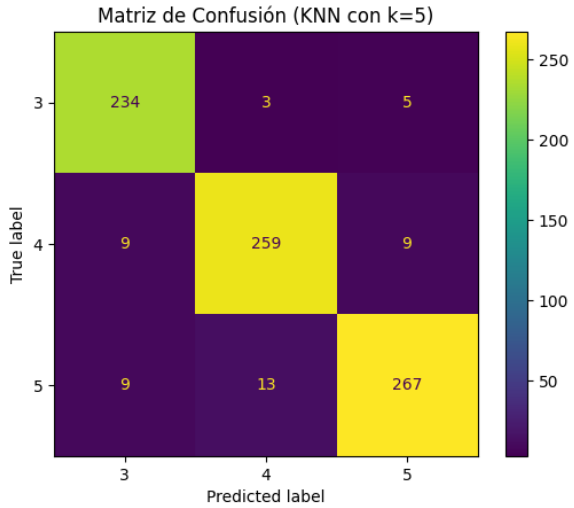

Ilustración 5 Matriz Confusión KNN

Mejor modelo

KNN mostró un rendimiento sólido pero un poco inferior en comparación con los otros dos. Por otro lado, SVM y Random Forest mostraron métricas similares. Sin embargo, aunque SVM muestra métricas ligeramente mejores en el reporte de clasificación en términos de precisión y recall por clase, Random Forest supera a SVM en términos de generalización, estabilidad y robustez del modelo. La capacidad de **Random Forest** para manejar grandes volúmenes de datos, su resistencia al sobreajuste, y el sólido desempeño en la validación cruzada lo hacen una opción fuerte en este contexto.

Análisis de palabras identificadas

El análisis de las opiniones recogidas ha permitido identificar palabras clave asociadas a distintos ODS, proporcionando una visión clara de las principales preocupaciones y áreas de interés de los participantes. A partir de las palabras más frecuentes en los textos analizados, es posible sugerir estrategias específicas para que la organización alinee sus acciones con estos objetivos, enfocándose en mejorar la calidad de los servicios de salud, fortalecer el sistema educativo, y promover la igualdad de género en diversas esferas.

El análisis de las opiniones vinculadas con el ODS 3 muestra que las palabras más frecuentes incluyen “salud”, “servicio”, “médico” y “enfermedad”. Estos términos resaltan una preocupación recurrente por el acceso a servicios de salud y a calidad de atención médica. También aparecen palabras relacionadas con el “sistema” de salud y la “capacidad” de respuesta ante enfermedades, lo que sugiere que los usuarios valoran la eficiencia de los sistemas sanitarios y su capacidad para proporcionar atención integral. Este análisis indica que la organización podría enfocarse en mejorar el acceso y la calidad de los servicios médicos, optimizar los recursos en los sistemas de salud y trabajar en la prevención de enfermedades para cumplir con los objetivos de salud establecidos.

Para el ODS 4, se observan términos clave como “estudio”, “educación”, “profesor”, “escuela” y “aprendizaje”, lo que indica que las opiniones destacan la importancia de la calidad educativa y el rol de los docentes en los niveles escolares. También se mencionan conceptos como “nivel” y “escuela”, lo que refleja la necesidad de asegurar que el aprendizaje ocurra de manera efectiva en todos los niveles del sistema educativo. La organización podría utilizar esta información para desarrollar estrategias que mejores la capacitación docente, fortalecer las instituciones educativas y promover políticas que aseguren una educación inclusiva y de calidad para todos.

Las palabras más frecuentes asociadas con el ODS 5 incluyen “mujer”, “género”, “hombre”, “igualdad” y “participación”. Estos términos reflejan la preocupación constante por la equidad de género, el papel de la mujer en la sociedad y la necesidad de fomentar su participación en diversas esferas, incluida la política y el trabajo. Se identifican también palabras como “trabajo” y “política”, lo que indica que las discusiones sobre igualdad de género se centran en gran medida en la necesidad de crear condiciones equitativas en el ámbito laboral y en la participación de las mujeres en la toma de decisiones. La organización podría enfocarse en promover políticas que fortalezcan la igualdad de género, fomentar la participación de mujeres en espacios de liderazgo y eliminar las barreras para su inclusión en el mercado laboral.

NOTA:

El mapa de actores relacionados con el producto de datos creado y el trabajo en equipo puede encontrarse en la Wiki del repositorio del equipo de trabajo.

Referencias.

[1]. UNFPA en Colombia. (s. f.). UNFPA-Colombia. https://colombia.unfpa.org/es/unfpa-en-colombia

[2]. *Objetivos de desarrollo sostenible*. (s. f.). UNDP. https://www.undp.org/es/sustainable-development-goals

[3]. UNFPA. (2013). Roles and Responsibilities of actors within UNFPA. En *UNFPA*. https://www.unfpa.org/sites/default/files/admin-resource/OSC_Annex%20Security%20Accountability_1.pdf

[4]. Géron, A. (2022). *Hands-on machine learning with Scikit-Learn, Keras and TensorFlow : concepts, tools, and techniques to build intelligent systems* (Third edition.). O'Reilly Media, Inc.

[5]. ¿Qué es KNN? | IBM. (s.f.). IBM - United States. https://www.ibm.com/mx-es/topics/knn