Proyecto 1 – Etapa 1: Analítica de Textos

Universidad de los Andes Departamento de Ingeniería de Sistemas y Computación
ISIS 3301 Inteligencia de Negocios
202320

Andrés Francisco Borda 201729184
Juan Pablo Lora Hernández 202012524
Gabriela Vargas Rojas 202013830

Contenido

1.1 Entendimiento del negocio y enfoque analítico 2
1.2 Entendimiento y Preparación de los datos 5
1.3 Modelado y evaluación 6
1.4 Resultados 8
1.5 Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido 9
1.6 Trabajo en equipo 11
1.7 Referencias 12

1.1 Entendimiento del negocio y enfoque analítico

Oportunidad/Problema Negocio	En primer lugar, vamos a establecer el contexto del proyecto definiendo qué es un ODS. Los Objetivos de Desarrollo Sostenible (ODS), son un conjunto de 17 objetivos interconectados establecidos por las Naciones Unidas en septiembre de 2015 como parte de la Agenda 2030 para el Desarrollo Sostenible. Estos objetivos fueron diseñados para abordar una amplia gama de desafíos globales, incluyendo la pobreza, el hambre, la salud, la educación, la igualdad de género, el agua limpia, la energía asequible, el trabajo decente, la paz y la justicia, entre otros. Para el caso que nos compete, vamos a trabajar con los ODS 6, 7 y 16, siendo estos respectivamente: **Objetivo de Desarrollo Sostenible 6 (ODS 6): Agua Limpia y Saneamiento:** Garantizar la disponibilidad y la gestión sostenible del agua y el saneamiento para todos. **Objetivo de Desarrollo Sostenible 7 (ODS 7): Energía Asequible y No Contaminante:** Garantizar el acceso a una energía asequible, fiable, sostenible y moderna para todos. **Objetivo de Desarrollo Sostenible 16 (ODS 16): Paz, Justicia e Instituciones Sólidas:** Promover sociedades justas, pacíficas e inclusivas para el desarrollo sostenible, brindar acceso a la justicia para todos y construir instituciones eficaces, responsables y transparentes a todos los niveles.

	Teniendo esto en cuenta, para continuar la agenda del desarrollo sostenible, es necesario apoyar a la ONU con el avance de los ODS que se han venido planteando. Con este fin, se busca ayudar en la clasificación de la información proporcionada por el Fondo de Poblaciones de las Naciones Unidas (UNFPA) para poder hacer un buen uso de las herramientas de participación ciudadana, identificar problemas y evaluar correctamente las soluciones actuales que se contrastan para satisfacer los objetivos planteados. De esta manera, la oportunidad principal para este negocio será poder hacer un análisis automatizado de las voces de los habitantes locales sobre las problemáticas de su entorno particular y así llevarlas de la mano al incluirlas en la aplicación de los objetivos de desarrollo sostenible. Finalmente, aterrizando la situación problema a nuestro contexto como país (Colombia), el abordar estos objetivos de desarrollo sostenible permitiría una mejora sustancial para nuestra población. Colombia tiene una gran cantidad de recursos naturales, tanto agua como posibles fuentes de energía sostenible que no se han aprovechado de esta manera, por lo que reconducir este camino sería uno de los pilares en el cumplimiento de la agenda para 2030. Así mismo, dos de las grandes problemáticas que nos afecta diariamente son la corrupción y los grupos al margen de la ley que imposibilitan la paz y un mejor conducto de nuestros recursos para avanzar. De esta manera, el objetivo de desarrollo sostenible 16 se acopla perfectamente en la solución de estos problemas y por lo tanto impactaría muy positivamente el país. El simple hecho de reducir la corrupción y poder formar una sociedad más pacífica permitiría un gran adelanto en Colombia.
Enfoque analítico	Teniendo en cuenta el contexto presentado por el negocio y lo que se busca lograr con el proyecto, se propone un enfoque analítico predictivo de aprendizaje supervisado, específicamente usando la tarea de clasificación sobre los datos recopilados, buscando relacionar los testimonios de las personas con el respectivo ODS. Se considera apropiado utilizar tres diferentes técnicas, RandomForest, LogisticRegression y MultinomialNB, esto permitirá formar distintos modelos y llegar al más optimo

	respecto a la comparación de las métricas de calidad de estos.
Organización y rol que se beneficia	Principalmente, la organización beneficiada en este caso sería la UNFPA, ya que el modelo de clasificación les permitirá un gran ahorro de tiempo y una mejora sustancial para apoyar los procesos de los diferentes ODS antes mencionados. Su rol beneficiado en esta situación sería fundamental en varios aspectos: **Eficiencia y Escalabilidad:** El UNFPA se beneficiaría al poder recopilar y analizar una gran cantidad de información de manera automatizada y eficiente **Toma de Decisiones Informada**: Al automatizar el proceso de recopilación y análisis de datos, el UNFPA podría tomar decisiones más informadas sobre políticas y programas destinados a promover el desarrollo sostenible. **Mayor Participación Ciudadana**: La automatización promoverá la participación de la ciudadanía en los procesos de planificación y desarrollo a nivel territorial, dado que facilitaría la recopilación de sus opiniones y perspectivas en las decisiones y políticas. Monitoreo en Tiempo Real: El UNFPA podría monitorear de manera más efectiva el progreso hacia los ODS en tiempo real al contar con datos actualizados y análisis en curso. **Mayor Transparencia**: La automatización del proceso de recopilación y análisis de datos podría aumentar la transparencia en el trabajo del UNFPA al proporcionar una base de datos sólida y objetiva que respalde sus decisiones y políticas, lo que a su vez podría generar una mayor confianza en la organización.
Contacto con experto externo al proyecto	Nuestro contacto experto para el proyecto es Alexandra Pulido Alvarado. Iniciamos el contacto a través de su correo electrónico: a.pulidoa@uniandes.edu.co Gracias a esto, intercambiamos números de teléfono y acordamos una reunión para explicarle en detalle el proyecto que se va a realizar e iniciar el trabajo conjunto. Por lo tanto, los canales a utilizar serán WhatsApp y zoom para el caso de la reunión virtual. Fecha de la reunión: 13 de octubre Se planea otra reunión para la realización y consecución de los resultados necesarios para la etapa 2.

1.2 Entendimiento y Preparación de los datos

Perfilamiento: Al explorar los datos encontramos que nos dieron registros de 3000 testimonios de diferentes personas, relacionadas a alguno de los objetivos de desarrollo sostenible (6,7 o 16). Cada registro cuenta con 2 columnas, teniendo una llamada “texto_espanol” que es lo dicho por la persona en una cadena de texto y otra columna de nombre “sdg” donde está la referencia en un valor numérico respecto al objetivo que se asocia. De esta manera, nuestra columna principal a tratar es la primera, la cual nos da información para relacionar con algunos de los objetivos desarrollos sostenibles, siendo esto lo que queremos lograr con el modelo.

Análisis de calidad de datos:

- **Completitud:** Para el análisis de completitud se observó que porcentaje de valores nulos tiene cada una de las columnas de los datos. Se encontró que las 2 columnas están completas, teniendo un 0% de incompletitud. Esto quiere decir que los 3000 registros están completos
- **Unicidad:** Para el análisis de unicidad primero se buscó si había registros repetidos y se encontró que no existe ninguno que este duplicado. Adicionalmente, es importante mencionar que este aspecto es importante para nuestra primero columna de texto, ya que para el caso de la numérica los valores solo pueden ser 6,7 o 16 por lo que se repiten en los diferentes registros.
- **Consistencia:** Se encontró que la columna “Textos_espanol” cuenta con 4 registros que tienen una mezcla de idiomas en la cadena de texto entre inglés y español, por lo tanto, rompen en este esquema de consistencia.
- **Validez:** Para el análisis de esta característica, se tomaron todos los registros como válidos, debido a que no poseemos un diccionario de datos para comparar y al estar hablando de cadenas de texto, es muy complicado llegar a evaluarla. Para la columna numérica, si se revisó que todos los valores estuvieran en el rango de 6,7 o 16 acorde a los objetivos que se seleccionaron para el proyecto.

Preparación de datos

Para la implementación de los modelos que nos ayudaran a la tarea propuesta es necesario corregir los problemas encontrados anteriormente, adicionalmente se deberá transformar los datos de manera que queden de tal manera que puedan ser utilizados por los modelos. Debido a que se encontraron textos que no están en español se eliminaron estos 4 registros. Con esto los datos están libres de problemas que impidan realizarlos modelos.

Después de esto se realizó la preparación de los datos que consta de la limpieza, tokenización y lemmatización de los mensajes. En términos de la limpieza se remplazaron los números por su representación en texto, se eliminaron los caracteres que no están en el alfabeto latino, y se eliminaron los espacios y la puntuación. Después de esto se realizo la lematización que reduce todas las

palabras a su raíz y se genera un token por cada una de estas raíces. Como resultado se obtuvo por cada uno de los mensajes la lista de tokens.

Para finalizar la preparación se realizo la division y vectorización de los datos. Los datos fueron divididos en la proporción 80-20 para entrenamiento y test. Para finalizar, se realizo la vectorización de los datos utilizando TFid para obtener los vectores de cada mensaje que se utilizaran para la construcción del modelo.

1.3 Modelado y evaluación

Para la realización y el alcance de los objetivos planteados para este proyecto, se desarrollaron tres diferentes modelos en base a tres diferentes algoritmos de clasificación apropiados para el contexto y los datos que se prepararon para esto, a continuación, explicaremos cada uno de ellos:

- **Algoritmo MultinomialNB (Juan Pablo Lora):** El algoritmo MultinomialNB (Naive Bayes Multinomial) es un clasificador de aprendizaje automático que se utiliza comúnmente en problemas de clasificación de texto. Se basa en el teorema de Bayes y asume independencia entre las características. Este algoritmo se utiliza para estimar la probabilidad de que un dato pertenezca a una de varias clases posibles, en función de la frecuencia de ocurrencia de características (como palabras en un documento). Es especialmente útil en aplicaciones de procesamiento de lenguaje natural, donde las características son términos o palabras y se cuenta cuántas veces aparecen en un documento de texto. De esta manera, obtuvimos estas métricas en nuestro modelo:

	precision	recall	f1-score	support
6	0.98	0.96	0.97	213
7	0.97	0.97	0.97	200
16	0.98	0.99	0.99	187
accuracy			0.98	600
macro avg	0.98	0.98	0.98	600
weighted avg	0.98	0.98	0.98	600

  Esto nos indica un muy buen modelo ya que esta muy cercano a obtener el 100% de las predicciones en todos los casos, siendo muy eficiente y acertado independiente del ODS en cuestión. Con un promedio de 98% de acierto y un recall mínimo del 96% indicándonos que identifica correctamente ese porcentaje de los pertenecientes a ese objetivo de desarrollo sostenible.
- **Algoritmo Logistic Regression (Andrés Borda):** La regresión logística es un método estadístico utilizado para predecir una variable categórica binaria (como sí/no, éxito/fracaso) en función de una o más variables predictoras. Utiliza la función logística para modelar la relación entre las variables predictoras y la probabilidad de pertenecer a una categoría específica. Para la realización de este algoritmo se deben determinar 3 hiperpárametros, el tipo de regularización (penalty), la fuerza de la regularización (C) y el

algoritmo de optimización que se utilizara (solver). Con el uso de un parameter grid se encontró que los hiperparametros óptimos son penalty: Ridge, C: 10000000000, solver: newton-cg. Con estos parámetros se entreno el modelo y se procedió a probar lo tanto en el set de entrenamiento como el de prueba. Se encontró que el accuracy del modelo en el set de entrenamiento es 0.985 y en el set de prueba también 0.986. La matriz de confusión y el reporte de clasificación encontrado es el siguiente:

```
[[208   3   2]
 [  3 197   0]
 [  0   0 187]]
```

	precision	recall	f1-score	support
6	0.99	0.98	0.98	213
7	0.98	0.98	0.98	200
16	0.99	1.00	0.99	187
accuracy			0.99	600
macro avg	0.99	0.99	0.99	600
weighted avg	0.99	0.99	0.99	600

Como podemos ver el modelo es bueno para la tarea propuesta con un accuracy muy alto mostrando que el modelo es muy bueno en predecir correctamente a que categoría se le debe asignar cada entrada.

- **Algoritmo Random Forest (Gabriela Vargas):** Random Forest es un algoritmo usado para tareas de clasificación y regresión. Este algoritmo es basado en árboles de decisión el cual logra combinar la salida de varios árboles de decisión para poder lograr a un resultado. El funcionamiento de este algoritmo se basa en tres hiperpárametros que son el tamaño del nodo, la cantidad de árboles y la cantidad de características muestreadas. Para su implementación, primero, importar las respectivas bibliotecas (en este caso RandomForestClassifier), luego se realizó la creación del modelo en donde el random state se especifica en 0 para afianzar la reproducibilidad. Una vez con esto, se entrenan el modelo, para así poder comenzar a realizar las particiones para la validación cruzada, en donde se usó 5 particiones teniendo en cuenta la cantidad de datos que se está manejando. Por otro lado, se realiza la definición de los hiperpárametros para realizar una búsqueda sistemática de los mejores hiperpárametros para este modelo y con este resultado, lograr arrojar el informe de clasificación con sus respectivos resultados. Por último, pudimos obtener los siguientes datos:

```
{'criterion': 'entropy', 'max_features': 100, 'min_samples_split': 2}

Classification Report
```

	precision	recall	f1-score	support
6	0.98	0.94	0.96	213
7	0.96	0.97	0.97	200
16	0.98	0.99	0.99	187
accuracy			0.97	600
macro avg	0.97	0.97	0.97	600
weighted avg	0.97	0.97	0.97	600

Por el lado de la precisión, recall y f1-score, los resultados arrojados fue que el promedio de estas es de 97%, lo cual estas métricas nos pueden indicar que el modelo logra hacer buenas predicciones y puede ser predecir datos

no etiquetados de manera precisa. Dado estos resultados, obtuvimos un accuracy_score de 0,97, lo cual nos puede indicar que este modelo nos permite relacionar de manera automática un texto según los ODS de manera casi precisa.

1.4 Resultados

Se determinó que se utilizará el modelo de regresión logística dado que su accuracy es más alto en comparación a los demás. Con el objetivo de la organización de predecir el ODS de documentos no etiquetados, se utilizó el modelo para clasificar las entradas no etiquetadas presentadas. Los resultados de este se presentan en el archivo “Resultado_predicciones.xlsx”.

En términos más cualitativos, gracias a la alta eficacia de este modelo, se podrá automatizar el proceso de la nueva información de participación ciudadana, logrando predecir rápidamente a que objetivo de desarrollo sostenible se está haciendo referencia y así mismo agilizar todo el proceso que se lleva a cabo para tener en cuenta estas voces dentro de las metas que se plantean y logrando de esta forma también poder poner este tiempo ganado en esfuerzos mucho más centrados.

Este modelo puede ser una base importante para que realmente las personas que participan en estas actividades vean de una manera mucho más transparente y efectiva cómo es que se están apoyando las causas de las cuales se hace eco su voz. Si se logra realmente mostrar los resultados a este grupo de gente, será posible entonces que muchos otros se unan a colaborar y el proyecto de los ODS sea cada vez más cercano y no se vea tan lejano como algunas veces puede parecer.

En conclusión, considerando que el modelo de regresión logística ha demostrado la mayor precisión en la predicción de qué textos se relacionan con los ODS, su implementación puede proporcionar beneficios significativos. Esta precisión mejorada permitirá una identificación más precisa de contenido relacionado con los ODS, lo que facilitará promocionar y seguir efectivamente estos objetivos. Por todo esto se recomienda al cliente utilizar este modelo para sus futuras tareas de predicción dado que su utilización representa una herramienta valiosa para impulsar un impacto más positivo y específico.

1.5 Mapa de actores relacionado con un producto de datos creado con el modelo analítico construido

Organización/Apartado	Rol dentro de la empresa	Tipo de Actor	Beneficio	Riesgo
Equipo de Investigación de la Universidad de los Andes	Investigadores	Usuarios Internos	Los investigadores de la Universidad de los Andes utilizan el producto de datos para realizar investigaciones académicas relacionadas con los ODS en Colombia, lo que contribuye al avance del conocimiento en el campo del desarrollo sostenible.	Si los datos no son precisos o completos, las investigaciones académicas basadas en ellos podrían carecer de fiabilidad.
Estudiantes de la Universidad de los Andes	Participantes	Usuarios Internos	Los estudiantes pueden aprender sobre la aplicación práctica de los ODS en Colombia a través de proyectos y análisis basados en el producto de datos, lo que enriquece su	Si la participación de los estudiantes en la recopilación o análisis de datos no se gestiona adecuadamente, podría haber desafíos logísticos o éticos.

			educación y conciencia sobre el desarrollo sostenible.	
UNFPA Colombia	Socio Estratégico	Beneficiarios/Colaboradores	UNFPA Colombia se beneficia al contar con un producto de datos que le permite evaluar la percepción de las comunidades locales en relación con los ODS en Colombia. Esto facilita la toma de decisiones informadas y la planificación de programas más efectivos.	Si los datos recopilados y analizados no reflejan con precisión la realidad de las comunidades, esto podría llevar a decisiones ineficaces o inadecuadas.
Comunidades y Grupos Locales en Colombia	Beneficiarios/Interesados	Beneficiarios	Las comunidades locales se benefician al tener sus voces y necesidades representadas en el proceso de planificación de proyectos y programas	Si las comunidades no confían en el proceso de recopilación de datos o no comprenden su importancia, podrían no

			relacionados con los ODS en Colombia.	participar de manera efectiva.

1.6 Trabajo en equipo

Estudiantes	Rol	Horas trabajadas	Tareas hechas	Desafíos	Puntaje
Andrés Borda	Líder de proyecto	14	Algoritmo – Documento - Vídeo	Investigación para la preparación de datos	33,33
Juan Pablo Lora	Líder de negocio	14	Algoritmo Documento – Vídeo	Investigación de los algoritmos y mejores modelos	33,33
Gabriela Vargas	Líder de analítica	14	Algoritmo Documento- Vídeo	Investigación del manejo de las librerías necesarias	33,33

Luego de realizar una reunión, logramos decidir que el puntaje de cada uno es de 33,33 (para así lograr una suma de 100) debido a que el trabajo realizado fue hecho y dividido de manera equitativa con el objetivo de ser más eficientes al realizar este proyecto. Con respecto a las horas, se definió que fue de 14 horas totales debido a que cada uno trabajó dos horas diario por 7 días. En cuanto a los puntos a mejorar para la siguiente entrega, consideramos que el punto más importante puede ser empezar con más tiempo el proyecto para poder lograr mejores resultados ya que se ha logrado tener un buen trabajo en grupo y una buena comunicación, lo cual creemos que debe seguir prevaleciendo para el siguiente proyecto. Cabe aclarar que también se realizó una reunión inicial para poder compartir ese planteamiento y definir cómo se iba a trabajar, especialmente en la comunicación para lograr el trabajo, ya en la reunión final se concretaron los últimos detalles de cada uno de los entregables.

1.7 Referencias

Naciones Unidas. “Objetivos De Desarrollo Sostenible”. Recuperado de: https://www.un.org/sustainabledevelopment/es/objetivos-de-desarrollo-sostenible/

Naciones Unidas. “Objetivo 6: Garantizar la disponibilidad de agua y su gestión sostenible y el saneamiento para todos”. Recuperado de: https://www.un.org/sustainabledevelopment/es/water-and-sanitation/

Naciones Unidas. “Objetivo 7: Garantizar el acceso a una energía asequible, segura, sostenible y moderna”. Recuperado de: https://www.un.org/sustainabledevelopment/es/energy/

Naciones Unidas. “Objetivo 16: Promover sociedades justas, pacíficas e inclusivas”. Recuperado de: https://www.un.org/sustainabledevelopment/es/peace-justice/

Fondo de población de las Naciones Unidas (UNFPA). (enero de 2018). “Quiénes Somos”. Recuperado de: https://www.unfpa.org/es/acerca-del-unf