Proyecto 1 – Construcción de modelos de analítica de Textos

INTELIGENCIA DE NEGOCIOS

Juan Andrés Eslava Tovar

Santiago Osorio Osorio

Alejandro Segura

Departamento de Ingeniería de Ssistemas y Computación

Universidad de los Andes

7 de septiembre de 2024

1. Entendimiento del negocio y enfoque analítico

1.1. Oportunidad/Problema de negocio

En Colombia, los Objetivos de Desarrollo Sostenible (ODS) representan una prioridad tanto para el gobierno como para las organizaciones sociales y económicas. Sin embargo, el seguimiento, medición y evaluación de los avances en los indicadores relacionados con los ODS 3 (Salud y bienestar), ODS 4 (Educación de calidad), y ODS 5 (Igualdad de género) presentan grandes desafíos debido a la dispersión de datos y la complejidad de las fuentes de información. Existe una necesidad de mejorar la precisión en la predicción de avances y posibles áreas de mejora en estos indicadores, basados en datos históricos y otras fuentes relevantes. Este proyecto busca abordar esta problemática mediante la implementación de un modelo de clasificación que ayude a predecir el cumplimiento de estos objetivos de manera eficiente y a identificar áreas que requieren más atención.

1.2. Objetivos y Criterios de Éxito desde el Punto de Vista del Negocio

El objetivo principal de este proyecto es implementar 1 modelo de análisis predictivo basado en técnicas de aprendizaje automático para predecir la tendencia de los indicadores relacionados con los ODS 3, 4 y 5 en Colombia. Los criterios de éxito incluyen:

1. Mejora en la precisión predictiva: Al implementar el modelo, se espera aumentar la capacidad de predecir correctamente los avances y retrocesos en los indicadores clave.
2. Identificación de áreas críticas: El modelo debe ser capaz de identificar áreas donde los avances son limitados, lo que permitiría a las entidades gubernamentales y ONGs focalizar sus esfuerzos y recursos.
3. Capacidad de toma de decisiones: Se busca que el modelo entregue insights claros que sirvan para la toma de decisiones tanto en el sector público como en las organizaciones que trabajan en torno a los ODS.
4. Reducción de la dispersión de datos: Estandarizar y unificar la información de diversas fuentes para facilitar el análisis y la toma de decisiones.

1.3. Organización y Rol Dentro de la Organización que se Beneficia con la Oportunidad Definida

Las principales organizaciones que se beneficiarían de este proyecto son:

1. **Gobierno Nacional y Local:** Principalmente el Departamento Nacional de Planeación (DNP) y los Ministerios de Salud, Educación y la Consejería Presidencial para la Equidad de la Mujer, quienes gestionan los programas relacionados con los ODS 3, 4 y 5. La información proporcionada por el modelo ayudará a priorizar políticas y recursos en las áreas de mayor necesidad.
2. **Organizaciones No Gubernamentales (ONGs):** Aquellas que trabajan en áreas de salud, educación y equidad de género podrán usar el modelo para planificar sus intervenciones de manera más efectiva.

1.4. Impacto que Puede Tener en Colombia Este Proyecto

Este proyecto tiene un impacto significativo en el país, ya que la correcta evaluación y predicción del avance en los ODS puede permitir un uso más eficiente de los recursos y un enfoque más claro en las áreas que requieren mayor atención. El éxito de este proyecto contribuiría directamente al mejoramiento de la calidad de vida de los colombianos al proporcionar información relevante para mejorar las políticas públicas en salud, educación y equidad de género. Además, un sistema predictivo robusto puede servir como modelo para otros países de la región que enfrentan desafíos similares en el seguimiento de los ODS.

1.5. Enfoque Analítico

El enfoque analítico del proyecto es predictivo, ya que se busca anticipar el comportamiento de los indicadores asociados con los ODS 3, 4 y 5 en el futuro basados en datos históricos. El tipo de aprendizaje que se implementará es el resultado de comparar 3 modelos: Regresión logístoca, Naive Bayes y árbol de decisión. Luego de compararlos, se escogerá el que mejor rendimiento tenga. Este tipo de análisis permitirá estimar probabilidades de cumplimiento de los ODS y clasificar el estado de cada indicador en diferentes categorías (avances, retrocesos o estancamiento).

El proceso de análisis incluye las siguientes etapas:

1. Preparación de datos: Limpiar y estructurar los datos de los indicadores de los ODS 3, 4 y 5 provenientes de diversas fuentes, eliminando valores atípicos y manejando datos faltantes.

2. Selección de algoritmos: Se escogerá el mejor modelo de los 3 mencionados anteriormente (Regresión logístoca, Naive Bayes y árbol de decisión), y con este se trabajará para los siguientes pasos.

3. Entrenamiento y validación del modelo: Se entrena el modelo con datos históricos y se valida utilizando técnicas como la búsqueda para optimizar su desempeño.

4. Evaluación del desempeño: Se evalúa la precisión del modelo mediante métricas como la matriz de confusión, precisión, recall y F1-score.

Al implementar este enfoque, se espera entregar un sistema capaz de generar reportes predictivos de los indicadores ODS, mejorar la toma de decisiones y optimizar las políticas y programas sociales en Colombia.

2. Entendimiento y preparación de los datos

2.1. Entendimiento de los datos

El proyecto comenzó con la exploración inicial de los datos proporcionados, compuestos por 4049 registros y 2 columnas: Textos_espanol (que contiene los textos a analizar) y sdg (que representa la categoría de Objetivos de Desarrollo Sostenible - ODS). En esta fase, verificamos la integridad estructural de los datos, asegurándonos de que los valores fueran válidos y que no hubiera inconsistencias significativas.

Durante el análisis preliminar, se identificaron las categorías de ODS a trabajar (3, 4, y 5), observándose una distribución de clases desbalanceada, con más ejemplos de las clases 4 y 5 en comparación con la clase 3. Este desbalance es relevante porque puede influir en el rendimiento de los modelos de clasificación.

Posteriormente, se llevó a cabo una visualización exploratoria, generando gráficos para revisar la frecuencia de palabras y combinaciones de palabras (bigramas y trigramas), lo que ayudó a identificar patrones temáticos relevantes para cada ODS. Además, se identificaron palabras irrelevantes, como URLs, que fueron eliminadas en la fase de limpieza.

2.2. Preparación de los datos

2.2.1. Eliminación de Registros no en Español

Dado que el análisis debía centrarse en textos en español, se aplicó un filtro de idioma utilizando langdetect, eliminando registros en inglés y francés. Tras este proceso, se mantuvieron 4036 registros en español.

2.2.2. Preprocesamiento y Limpieza

En la fase de preprocesamiento, los textos fueron normalizados eliminando puntuación, caracteres especiales, y URLs. Además, se eliminaron las stopwords para reducir el ruido en los datos y se aplicó lemmatización para reducir las palabras a su forma base. Esto permitió que el modelo generalizara mejor durante la etapa de entrenamiento, asegurando que solo las palabras más relevantes se utilizaran en el análisis.

2.2.3. Vectorización (TF-IDF)

Para transformar los textos en una representación numérica comprensible para los modelos de aprendizaje automático, se utilizó la técnica TF-IDF (Term Frequency-Inverse Document Frequency). Esto generó una matriz de características con 4036 documentos y 14,652 características, representando cada documento como un vector numérico. TF-IDF permitió asignar mayor peso a las palabras más relevantes y reducir la importancia de las palabras más comunes.

2.2.4. Exportación de Datos Procesados

Finalmente, el conjunto de datos limpio y procesado fue exportado en formato CSV, asegurando que estaba listo para ser utilizado en la siguiente fase del modelado. Este conjunto de datos preparado proporcionó una base sólida para el posterior análisis predictivo utilizando el mejor de 3 algoritmos.

3. Modelado y evaluación

Para este proyecto, se implementaron y compararon tres modelos de aprendizaje supervisado con el fin de predecir correctamente la clasificación de los textos en las categorías de los Objetivos de Desarrollo Sostenible (ODS 3, 4 y 5). Los modelos evaluados fueron:

3.1. Regresión Logística

La Regresión Logística es un modelo de clasificación lineal que estima la probabilidad de que una instancia pertenezca a una clase utilizando una función sigmoide. Este modelo fue el que mejor rendimiento presentó en comparación con los demás.

3.1.1. Implementación

El modelo de Regresión Logística se ajustó a los datos de entrenamiento y fue optimizado utilizando validación cruzada. No se observó sobreajuste en este modelo, lo que indica que generaliza bien tanto en el conjunto de entrenamiento como en el de prueba.

3.1.2. Evaluación

- Métricas de rendimiento: La Regresión Logística obtuvo el mejor rendimiento global en términos de precisión, recall y f1-score. Las métricas mostraron una excelente capacidad de predicción, con un equilibrio adecuado entre las clases.
- Interpretación: Dado que la Regresión Logística modela las probabilidades de pertenencia a las clases, fue particularmente efectiva en la clasificación de los textos de los ODS, lo que la hizo el modelo más adecuado para este problema.

3.2. Naive Bayes

El Naive Bayes es un clasificador probabilístico que asume independencia entre las características. Aunque este modelo es conocido por ser eficiente en tareas de clasificación de texto, en este caso presentó el rendimiento más bajo de los tres modelos comparados.

3.2.1. Implementación

Se utilizó el clasificador Naive Bayes Gaussiano. A pesar de la simplicidad y velocidad del modelo, los resultados no fueron tan precisos como en los otros dos modelos debido a la suposición de independencia entre las características, lo que puede no ser del todo adecuado para el conjunto de datos analizado.

3.2.2. Evaluación

- Métricas de rendimiento: El modelo Naive Bayes tuvo un rendimiento más bajo, con métricas de precisión y f1-score inferiores en comparación con la Regresión Logística y el Árbol de Decisión. En particular, mostró dificultades para separar correctamente las clases, lo que se reflejó en una matriz de confusión con mayores errores en las predicciones.
- Interpretación: Aunque es un modelo útil en muchas aplicaciones de clasificación de texto, la estructura y relaciones entre las palabras en este conjunto de datos parecieron requerir un enfoque más complejo.

3.3. Árbol de Decisión

El Árbol de Decisión fue el segundo mejor modelo en términos de rendimiento. Este modelo crea un árbol de decisiones basándose en la división de las características para maximizar la separación entre las clases.

3.3.1. Implementación

El Árbol de Decisión se entrenó utilizando el criterio de Gini para medir la pureza de las divisiones en los nodos. Aunque el modelo es fácil de interpretar, es susceptible al sobreajuste, lo que fue mitigado ajustando parámetros como la profundidad máxima del árbol.

3.3.2. Evaluación

- Métricas de rendimiento: El Árbol de Decisión se desempeñó bien, logrando un balance entre precisión y recall. Sin embargo, presentó un rendimiento ligeramente inferior al de la Regresión Logística, aunque superior al de Naive Bayes.
- Interpretación: Si bien el Árbol de Decisión proporcionó un buen rendimiento, su tendencia a sobreajustar en los datos de entrenamiento se gestionó limitando la profundidad del árbol. Aun así, el modelo mostró ser eficaz en la clasificación, pero no logró superar la precisión de la Regresión Logística.

4. Resultados

4.1. Rendimiento de resultados

El modelo de Regresión Logística ha demostrado ser el más preciso de los tres modelos comparados, obteniendo un rendimiento sólido en la clasificación de los textos asociados a los ODS 3, 4 y 5. Los resultados en el conjunto de prueba se detallan a continuación:

- ODS 3 (Salud y bienestar):
  Precisión: 0.73
  Recall: 0.73
  F1-Score: 0.73
- ODS 4 (Educación de calidad):
  Precisión: 0.74
  Recall: 0.79
  F1-Score: 0.76
- ODS 5 (Igualdad de género):
  Precisión: 0.78
  Recall: 0.74
  F1-Score: 0.76
- Promedio General:
  Precisión Macro: 0.75
  Recall Macro: 0.75
  F1-Score Macro: 0.75
  Precisión Ponderada: 0.75
  Recall Ponderado: 0.75
  F1-Score Ponderado: 0.75

4.2. Análisis de palabras clave por clase

- ODS 3 (Salud y bienestar):

    - Palabras clave con coeficientes positivos: "maternal", "sensibilizado", "dignidad".
    - Palabras clave con coeficientes negativos: "ciclista", "divergent".
- ODS 4 (Educación de calidad):
    - Palabras clave con coeficientes positivos: "importante", "adquirido", "tolerada".
    - Palabras clave con coeficientes negativos: "oncológica", "sacrificiir".
- ODS 5 (Igualdad de género):
    - Palabras clave con coeficientes positivos: "difundido", "antibacteriano", "letalidad".
    - Palabras clave con coeficientes negativos: "afectar", "recibirar".

4.3. Contribución a los Objetivos del Negocio

- Mejora en la Precisión Predictiva: El modelo de regresión logística ha mostrado una precisión general del 75%, lo que permite identificar mejor los avances y áreas críticas en los ODS 3, 4 y 5.
- Identificación de Áreas Críticas: Los coeficientes de las palabras clave proporcionan información valiosa para las entidades gubernamentales y ONGs, permitiéndoles focalizar sus esfuerzos en las áreas más relevantes para cada ODS.
- Capacidad de Toma de Decisiones: Con base en los resultados del modelo, los responsables de políticas pueden ajustar sus estrategias para maximizar el impacto en los indicadores de los ODS.
- Reducción de la Dispersión de Datos: El modelo ayuda a consolidar información proveniente de diferentes fuentes, estandarizando los datos y facilitando la toma de decisiones informadas.

4.4. Conclusión

El modelo de Regresión Logística no solo clasifica de manera precisa los textos sobre los ODS, sino que también proporciona insights útiles para las políticas públicas. Su capacidad predictiva ayuda a identificar áreas críticas y a mejorar la eficiencia en la toma de decisiones, contribuyendo al cumplimiento de los ODS en Colombia.

5. Mapa de actores

Rol dentro de la empresa	Tipo de actor	Beneficio	Riesgo
Departamento de Planeación	Usuario-cliente	Mejor toma de decisiones basada en datos, alineada con los ODS 3, 4 y 5	Reduce el uso del juicio humano ya que depende excesivamente del modelo.
Gerencia de Proyectos	Financiador	Optimización de recursos y enfoque en proyectos que generen mayor impacto	Si el modelo no funciona correctamente, los recursos podrían malgastarse en áreas no prioritarias

Departamento de Tecnologías de la Información	Proveedor	Garantía de que el modelo se integra con otros sistemas y cumple con los estándares de seguridad de datos	Riesgo de filtración de datos si no se implementan medidas de seguridad adecuadas
Ciudadanos	Beneficiados	Participación ciudadana efectiva en la toma de decisiones mediante la alineación con los ODS	Desconfianza en el sistema si los resultados no son percibidos como justos o precisos

6. Trabajo en equipo

6.1. Asignación de roles y algoritmos

NOMBRE	ROL	ALGORITMO
Santiago Osorio Osorio	Lider de proyecto	Regresión llogistica
Juan Andrés Eslava	Líder de analítica	Naive Bayes
Alejandro Segura	Líder de datos	Árbol de decisión

6.2. Distribución de tareas

Santiago:

Entendimiento y preparación de datos, Algoritmo de regresión logística, resultados.

Juan:

Algoritmo de Naive Bayes, resultados, entendimiento del negocio, crear documento.

Alejandro:

Algoritmo de árbol de decisión, mapa de actores, crear presentación.

6.3. Retos y soluciones

RETOS	SOLUCIONES
Entendimiento de datos	Revisión detallada para entender como es la mejor forma de preparar los datos para que funciones en cualquier algoritmo que vayamos a desarrollar
Escoger algoritmos	Intentamos buscar 3 algoritmos que se pudieran diferenciar de la mejor manera y que tuivieran el mejor rendimiento posible

6.4. Reuniones

- Reunión de lanzamiento:

  24/08/2024, se entendió el proyecto y se dieron tareas iniciales.

- Reunión de seguimiento de entendimiento:

29/08/2024, se consolidó el entendimiento y preparación, se asignaron los algoritmos.

- Reunión de seguimiento algoritmos:

  2/09/2024, se vieron los resultados de los algoritmos, se definió cuál usar para la siguiente etapa.

- Reunión de resultados:

  5/09/2024, se socializaron los resultados.

- Reunión final:

  7/09/2024, se consolidó la entrega y se creó el video.

6.5. Reparto de puntos

Alejandro Segura: 33.3%

Santiago Osorio: 33.3%

Juan Eslava: 33.3%