Proyecto 1 Etapa 1

Inteligencia de negocios

Universidad de los Andes

Samuel Freire
Felipe Garcia
Lucciano Franco

Entendimiento del negocio y enfoque analitico

Oportunidad/ Problema de negocio

- El proyecto busca automatizar la identificación y clasificación de opiniones o textos en función de los Objetivos de Desarrollo Sostenible (ODS).
- Además de los objetivos empresariales, el proyecto tiene un componente social significativo al permitir escuchar y analizar las opiniones de las personas sobre problemáticas.
- Los ODS de la ONU abordan cuestiones como bienestar, educación de calidad e igualdad de género, entre otros.
- El impacto del proyecto se traduce en la posibilidad de actuar y solucionar problemáticas de manera eficaz, especialmente en el contexto colombiano, al dar voz a las personas afectadas.
- La métrica clave para evaluar el éxito del proyecto radica en su capacidad para clasificar textos en los ODS de manera precisa, lo cual representa el punto crítico para determinar su efectividad.

Enfoque analitico

- Se ha elegido un enfoque analítico basado en aprendizaje supervisado, debido a la disponibilidad de etiquetas que definen los objetivos de desarrollo sostenible asociados a cada texto.
- La tarea de aprendizaje se centra en la clasificación, lo que facilitará la división de los datos en los tres objetivos de desarrollo sostenible asignados.
- Se utilizarán tres algoritmos: árbol de decisión (específicamente, el algoritmo de random forest) por su facilidad de implementación y capacidad de explicación; regresión logística por su capacidad de clasificación versátil y reconocimiento de patrones no lineales; y SVM, un método basado en probabilidad que ofrece un enfoque diferente para la clasificación.
- La propuesta es útil para la empresa ya que proporcionará un modelo con alto porcentaje de precisión en la clasificación de opiniones en los objetivos deseados.
- Además, ofrecerá una visualización clara y comprensible de los resultados propuestos.

Organización y rol dentro de ella

- La ONU y la UNFPA son las principales beneficiarias del proyecto, al proporcionar la información y ser la entidad contratante. También se espera un impacto positivo en las personas relacionadas con las problemáticas estudiadas.
- La UNFPA se beneficia directamente al ser la entidad contratante. Además, las víctimas de las problemáticas y los líderes de propuestas pueden obtener beneficios a partir de los resultados del proyecto.

Entendimiento de datos y Limpieza de datos

Entendimiento de los datos

- El análisis inicial se centra en un conjunto de datos con pocas columnas, incluyendo el ID organizador, las opiniones de las personas sobre los objetivos de desarrollo y la referencia al objetivo específico.
- A pesar de ser una columna numérica, es importante recordar que representa categorías de objetivos de desarrollo. Se analizaron 3000 registros con una media de 4, una desviación de 0.81, un valor mínimo de 3 (objetivo de desarrollo 3) y un máximo de 5 (otro objetivo de desarrollo).

Entendimiento de datos y Limpieza de datos

Calidad de los datos

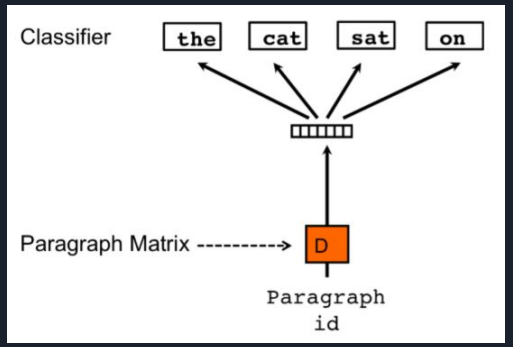
- Completitud de datos: No se observan valores nulos en el conjunto de datos, lo que indica una alta completitud y no requiere tomar decisiones sobre la inclusión o exclusión de registros.
- Unicidad: No se encuentran valores duplicados en la información, lo que demuestra una buena variedad de datos y no es necesario eliminar duplicados.
- Consistencia: Todos los valores se encuentran dentro de los rangos predefinidos, lo que indica una consistencia adecuada en los datos.
- Validez: Dado que los datos cumplen con todos los formatos y no existe una relación significativa entre columnas, no es necesario realizar un análisis adicional de validez.

Entendimiento de datos y Limpieza de datos

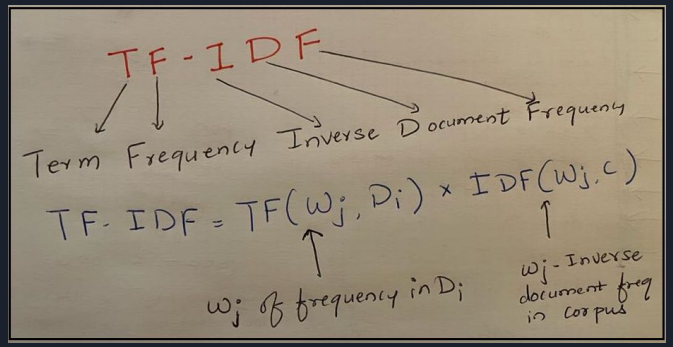
Limpieza de datos

- Transformación ASCII: Eliminar símbolos o caracteres que no se ajusten al formato ASCII o similares, para preparar el texto para su análisis.
- Normalización a minúsculas: Homogeneizar el texto convirtiendo todas las letras a minúsculas para evitar problemas de comparación entre palabras en mayúsculas y minúsculas.
- Eliminación de signos de puntuación: Remover puntos, comas y punto y comas del texto para mantener el contenido limpio y enfocado en el análisis.
- Sustitución de números: Cambiar los números por palabras para evitar que afecten el análisis y facilitar la comprensión del contenido.
- Eliminación de stopwords: Excluir palabras comunes pero poco informativas del análisis, para evitar densidad y mejorar la relevancia del texto.
- Tokenización y preparación para los algoritmos: Dividir los datos en tokens y realizar diversas transformaciones para prepararlos para su procesamiento en los algoritmos seleccionados.

Algoritmos de vectorización



Doc2Vec



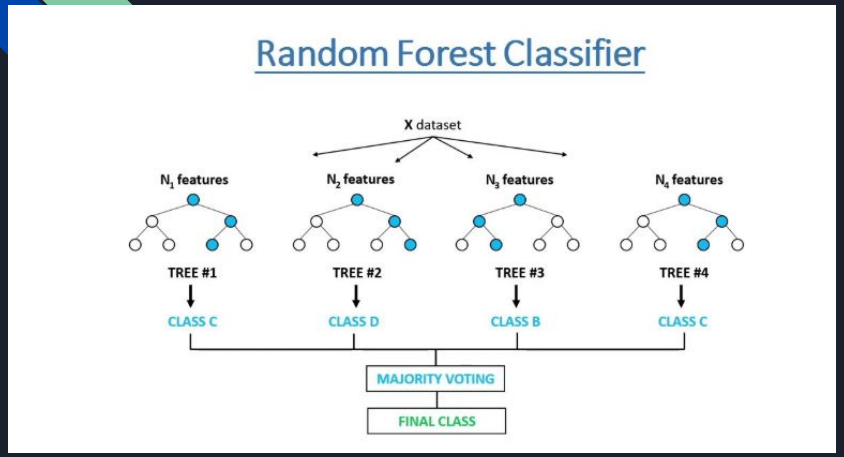
TF-IDF

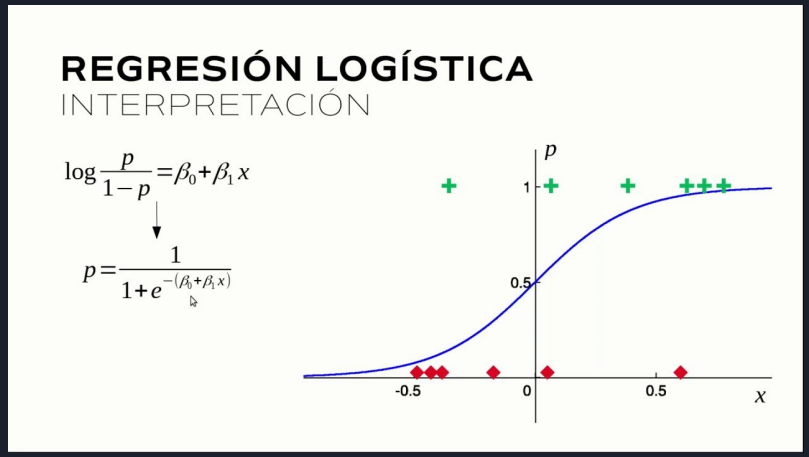
	esta	pelicula	es	malisima	buenisima
Esta pelicula es buenisima	1	1	1	0	1
Esta pelicula es malísima	1	1	1	1	0
malísima	0	0	0	1	0

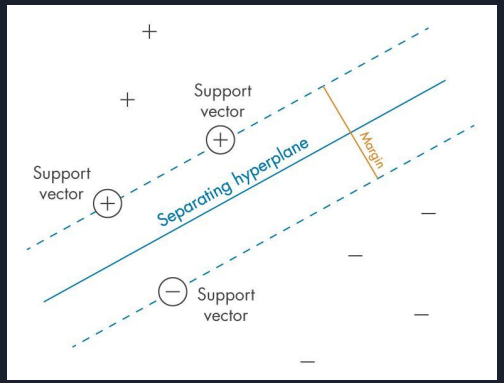
Bow

GloVe

Algoritmos de Clasificación







SVM(Support Vector Machine)

Métricas

1. Precision (Precisión): La precisión mide la proporción de predicciones positivas correctas (verdaderos positivos) con respecto al total de predicciones positivas realizadas. Es útil para evaluar cuán confiables son las predicciones positivas del modelo.

2. Recall (Recuperación): El recall evalúa la proporción de casos positivos reales que el modelo ha identificado correctamente como positivos (verdaderos positivos). Se utiliza para medir la capacidad del modelo para detectar todos los casos positivos en el conjunto de datos.

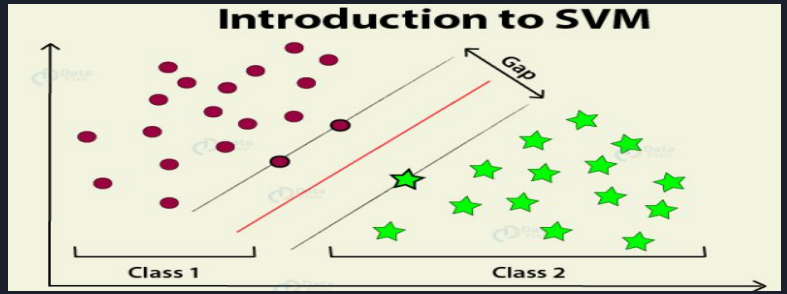
3. F1 Score: El puntaje F1 es una métrica que combina la precisión y el recall en un solo valor, proporcionando un equilibrio entre ambas métricas. Es útil cuando se necesita encontrar un compromiso entre la identificación precisa de positivos y la cobertura de todos los positivos en el conjunto de datos.

TF-IDF

Modelo Seleccionado

Modelo TF-IDF con SVM

- Precisión: 97.55%
- Recall: 97.55%
- F1: 97.55%



Resultados, evaluación y oportunidad

Se proponen algunas estrategias clave que la organización puede implementar a partir de nuestro proyecto.

1. Alta precisión y calidad de predicciones
2. Ampliación a diferentes departamentos
3. Equipo interdisciplinario
4. Umbral de confianza y monitoreo continuo
5. Experimentación continua