Proyecto 1: AT de reseñas de peliculas

Grupo 26

Índice

01 Introducción

02 Entendimiento de los datos

03 Limpieza de datos

04 Modelos

05 resultados

Introducción

El proyecto es de analítica de sentimientos usando reseñas de películas, donde se quiere clasificar si la reseña es positiva o negativa, de manera que con el modelo en construcción se pueda facilitar la clasificación de reseñas.

Este modelo podría llegar a tener un beneficio en empresas como lo puede ser cineColombia, Netflix y plataformas de streaming de películas . Pues la ventaja está en que rápidamente se podría llegar a una idea de cómo se ha recibido las películas entre el público para tomar decisiones de negocio como quitarla de taquilla, o incluso en ciertas platafomas de streaming se podría hablar de recomendarla.

Entendimiento de los datos

Datos:

- 5000 registros
- 2 datos categóricos
- 1 dato numérico

Completitud:

- Todos los datos están completos.

Unicidad:

- 2 Datos repetidos en la columna de reseñas.

Nulos:

- No hay nulos

Validez:

- Son validos

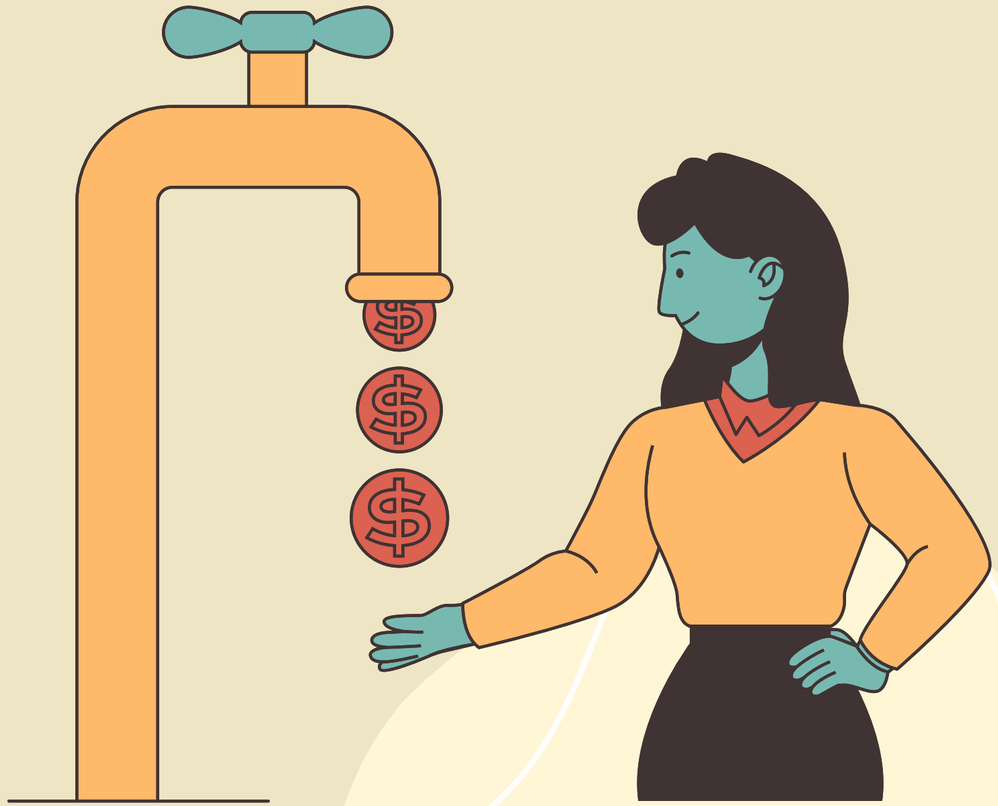
Limpieza y tratamiento de datos:

LIMPIEZA DE DATOS:

1. Se eliminó los duplicados.

PREPARACIÓN DE LOS DATOS:

1. Se eliminó todos los caracteres No ASCII
2. Todo carácter pasó a estar en minúscula.
3. Se eliminó toda puntuación. (./,/;/(/))
4. Se remplazó todo numero a palabras con num2words.
5. Se cambió las palabras vacías/ palabras comunes (stopwords) con la librería nltk a espacios.
6. Todo mensaje clasificado como negativo pasó a ser 1 y 0 positivo.
7. Con countVectorize se construyó una matriz que mirara cuantas si aparece una palabra en especifico.



Técnicas y algoritmos:

Tecnica: Clasificación

- regresión logistica
- Arboles de decisión
- SVM (Maquinas de vectores de soporte)

Regresión lógistica

El modelo de regresión logistica para análisis de textos es una técnica de aprendizaje supervisado que se utiliza para predecir la probabilidad de que un texto pertenezca a una o más categorias predefinida

```
Matriz de confusión:
 [[594 143]
 [ 95 668]]
```

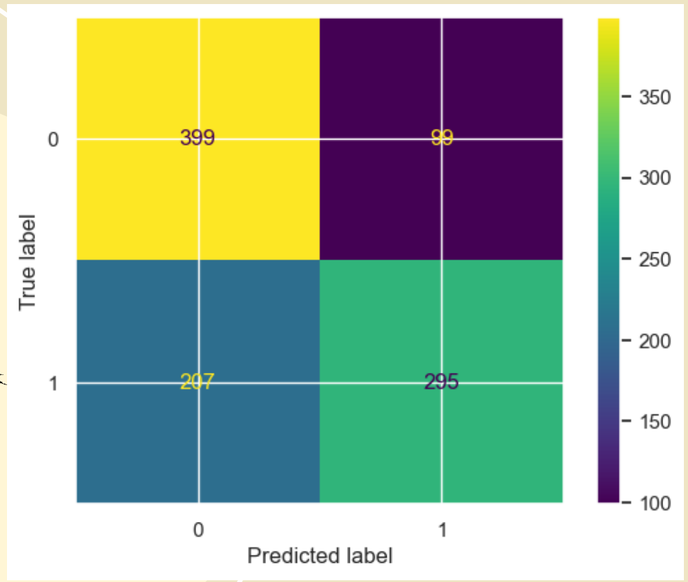
```
Informe de clasificación:
              precision    recall  f1-score   support

           0       0.86      0.81      0.83       737
           1       0.82      0.88      0.85       763

    accuracy                           0.84      1500
   macro avg       0.84      0.84      0.84      1500
weighted avg       0.84      0.84      0.84      1500
```

Árbol de decisión

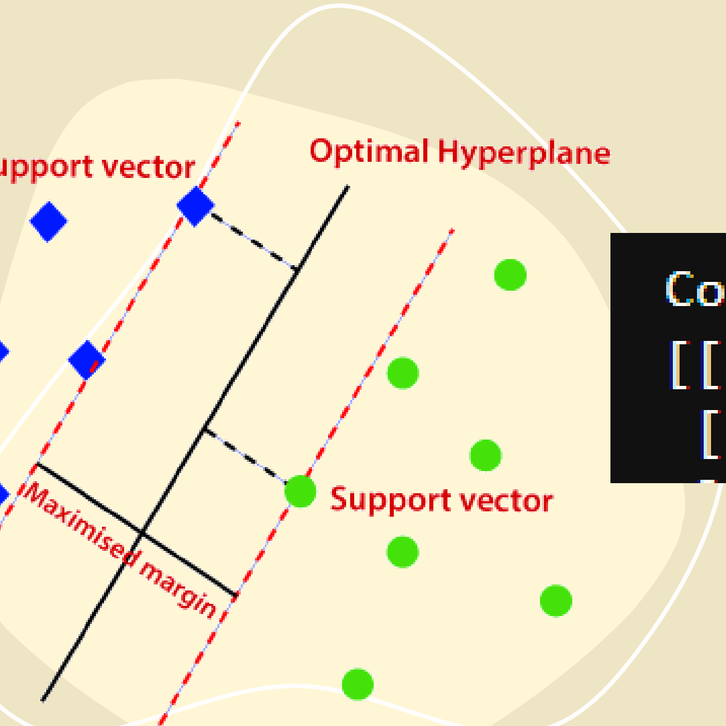
Los arboles de decisión es un algoritmo de clasificación en un aprendizaje supervisado Los nodos hojas representan los resultados posibles dentro del conjunto de datos, donde hay un numero de reseñas y un numero de entropia. Cabe resaltar que entre menor sea dicho número mejor son la clasificación de las reseñas.



	precision	recall	f1-score	support
0	0.66	0.80	0.72	498
1	0.75	0.59	0.66	502
accuracy			0.69	1000
macro avg	0.70	0.69	0.69	1000
weighted avg	0.70	0.69	0.69	1000

SVM (Maquinas de vectores de soporte)

El algoritmo de máquinas de vectores de soporte es un algoritmo de aprendizaje supervisado, el cual se usó para la clasificación binaria y poder separar de la mejor forma posible dos clases diferentes de puntos de datos.




```
Confusion Matrix:
[[388  97]
 [ 81 434]]
```

```
Classification Report:
              precision    recall  f1-score   support

           0       0.83      0.80      0.81       485
           1       0.82      0.84      0.83       515

    accuracy                           0.82      1000
   macro avg       0.82      0.82      0.82      1000
weighted avg       0.82      0.82      0.82      1000
```

Conclusiones

Técnicas y algoritmos:

Tecnica: Clasificación

- regresión logistica
- Arboles de decisión
- SVM (Maquinas de vectores de soporte)

Técnicas y algoritmos:

Tecnica: Clasificación

- regresión logistica

¡Gracias!