



PRESENTACIÓN **PROYECTO 1**

Daniel Esteban Aguilera Figueroa – 202010592
Laura Vanessa Martínez Prieto – 202012624
Cristian Armando Sánchez Ocampo – 202022112

UNIVERSIDAD DE LOS ANDES

INTRODUCCIÓN

Este proyecto se enfoca en fortalecer las competencias necesarias para una aplicación efectiva del proceso de descubrimiento de conocimiento a partir de textos, utilizando una metodología propia del mundo de la analítica. Una de las opciones de problemáticas a abordar en este proyecto es el análisis de sentimientos de películas en español, se tienen comentarios de películas que deben ser clasificados en las categorías de positivo o negativo.

ENTENDIMIENTO DE DATOS

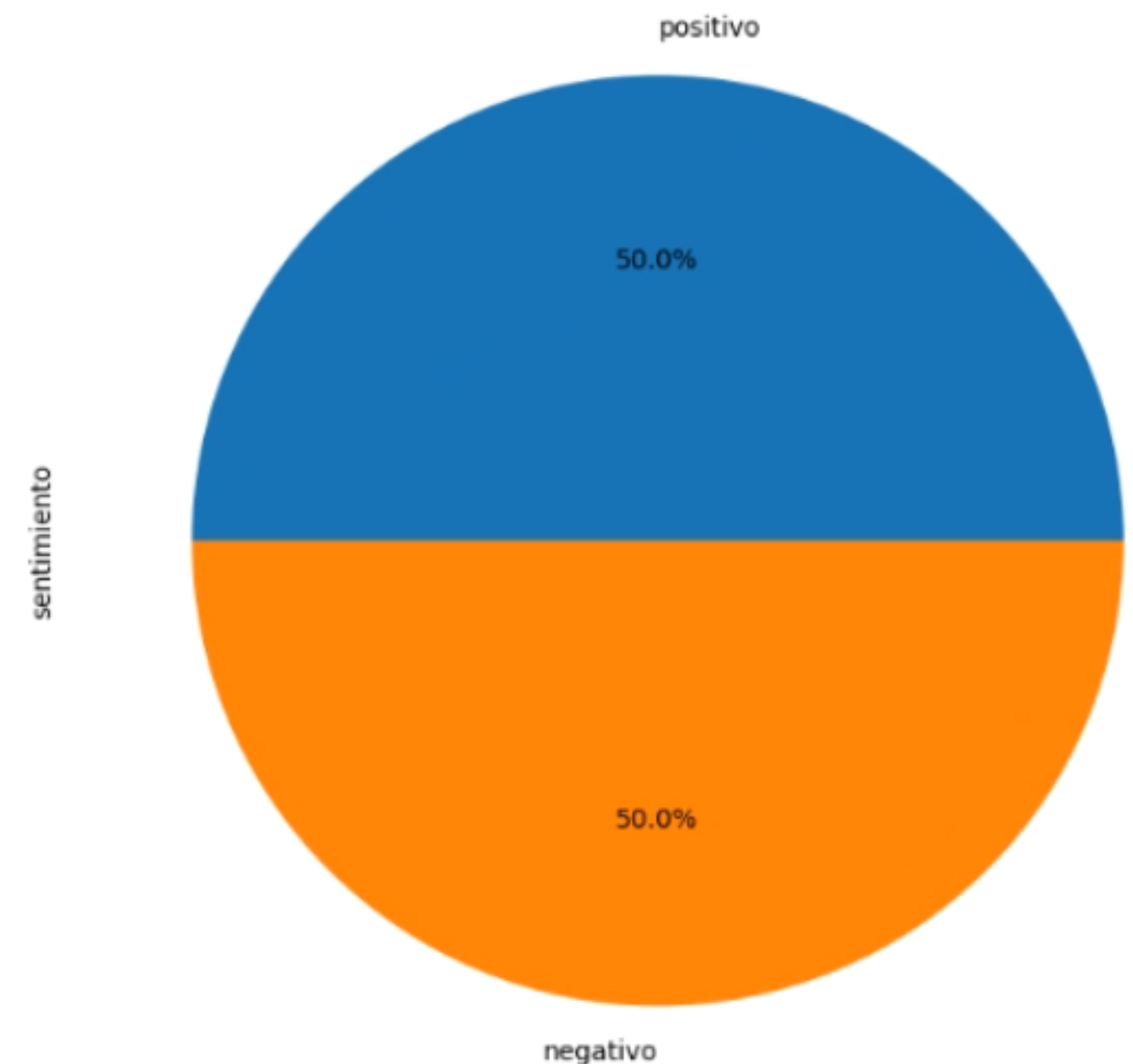
Se analizan los datos en términos completitud, unicidad, consistencia y validez.

Para cumplir con los elementos ya mencionados se verifica:

- La existencia de nulos.
- La duplicidad de los datos: Teniendo como resultado un mismo comentario tanto en sentimiento negativo como positivo.

La validez de los datos no se puede verificar ya que al ser una comentario es libre por el usuario.



Al momento de revisar la variable objetivo vemos que es balanceada con 2499 comentarios clasificados con cada una.





TRATAMIENTO DE DATOS

Para esta etapa se decidió seguir con los siguientes pasos

- Se transforman todas las palabras a minúsculas.
 - Se eliminan todos los signos de puntuación y caracteres especiales.
 - Se transforman los caracteres con tilde a su forma base.
 - Se eliminan las palabras que no tienen una representación relevante para el modelo (Stopwords).
 - Se reemplazan los caracteres numéricos que posiblemente tienen una representación textual.
 - Tokenización, Lematización y Stemming a todas las palabras para su representación base y fructífera para el modelo
 - Se transforma la variable objetivo a una representación numérica (1 – Positivo, 0 – Negativo)
- 
- 

TRATAMIENTO DE DATOS

Para esta etapa se decidio seguir con los sigientes pasos

Como resultado de lo anterior se puede observar el siguiente dataframe:

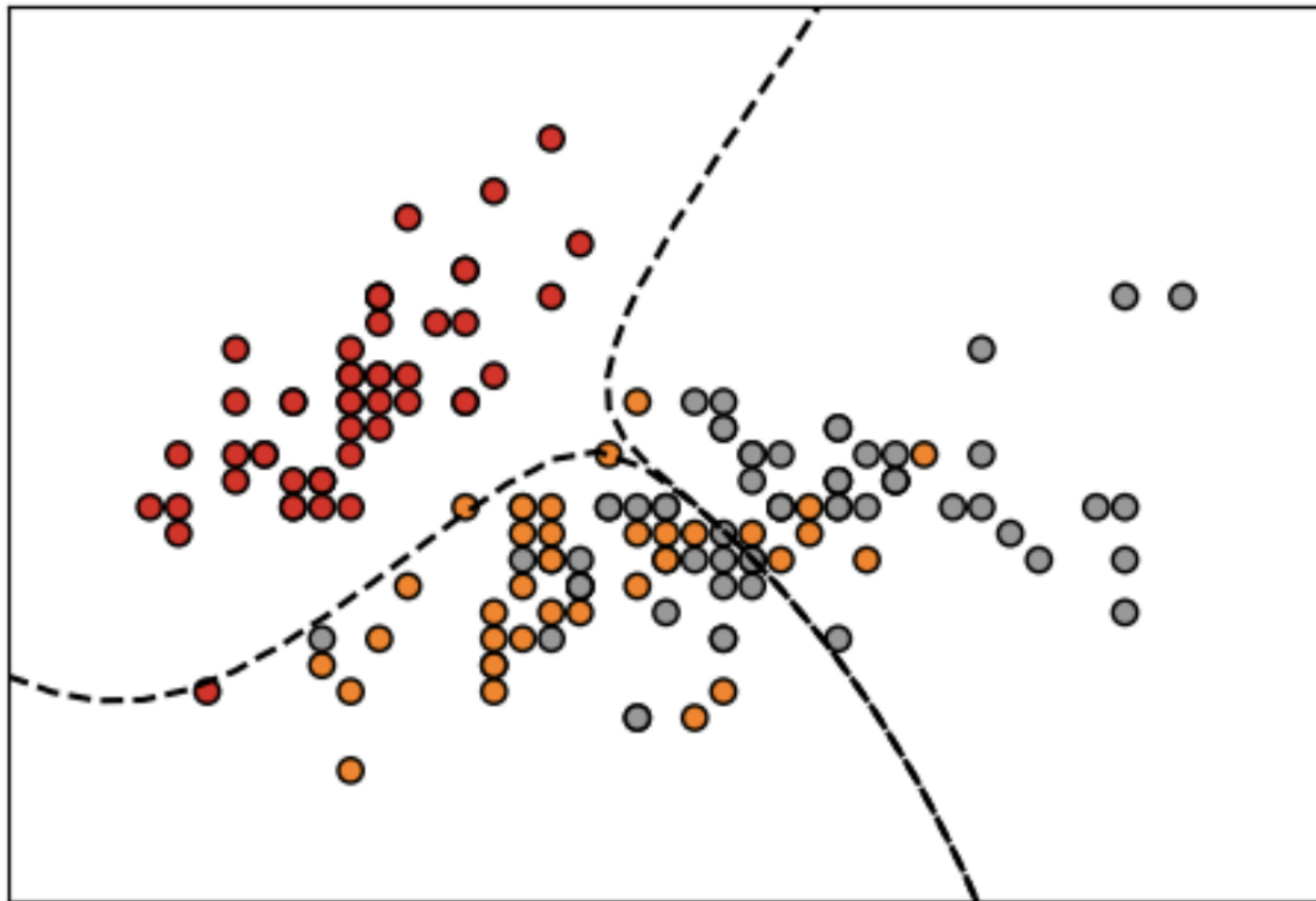
	review_es	sentimiento	processed_text
0	Si está buscando una película de guerra típica...	1	si busc pelicul guerr tipic asi not aficion gu...
1	Supongo que algunos directores de películas de...	1	supon director pelicul luj sent busc abrig gra...
2	Es difícil contarle más sobre esta película si...	1	dificil contar el mas pelicul estropearlal dis...
3	La película comienza muy lentamente, con el es...	1	pelicul comenz lent estil vid wallac napalm as...
4	Esta película es verdadera acción en su máxima...	1	pelicul verdader accion maxim expresion mejor ...

NAIVE BAYES

El modelo de Naive Bayes es un modelo de clasificación probabilístico, de aprendizaje automático supervisado, que se basa en el teorema de Bayes y en la suposición ingenua (naive) de independencia condicional entre las características de los datos.

Este modelo es ampliamente utilizado en el análisis de sentimientos en redes sociales, comentarios de productos en línea y reseñas de películas, entre otros.

Es un enfoque simple pero efectivo para la clasificación de sentimientos en grandes conjuntos de datos.



Resultados

```
Exactitud: 0.82
Recall: 0.7575360419397117
Precisión: 0.8639760837070254
Puntuación F1: 0.8072625698324023
```

	precision	recall	f1-score	support
0	0.78	0.88	0.82	737
1	0.86	0.76	0.81	763
accuracy			0.82	1500
macro avg	0.82	0.82	0.82	1500
weighted avg	0.82	0.82	0.82	1500

NAIVE BAYES

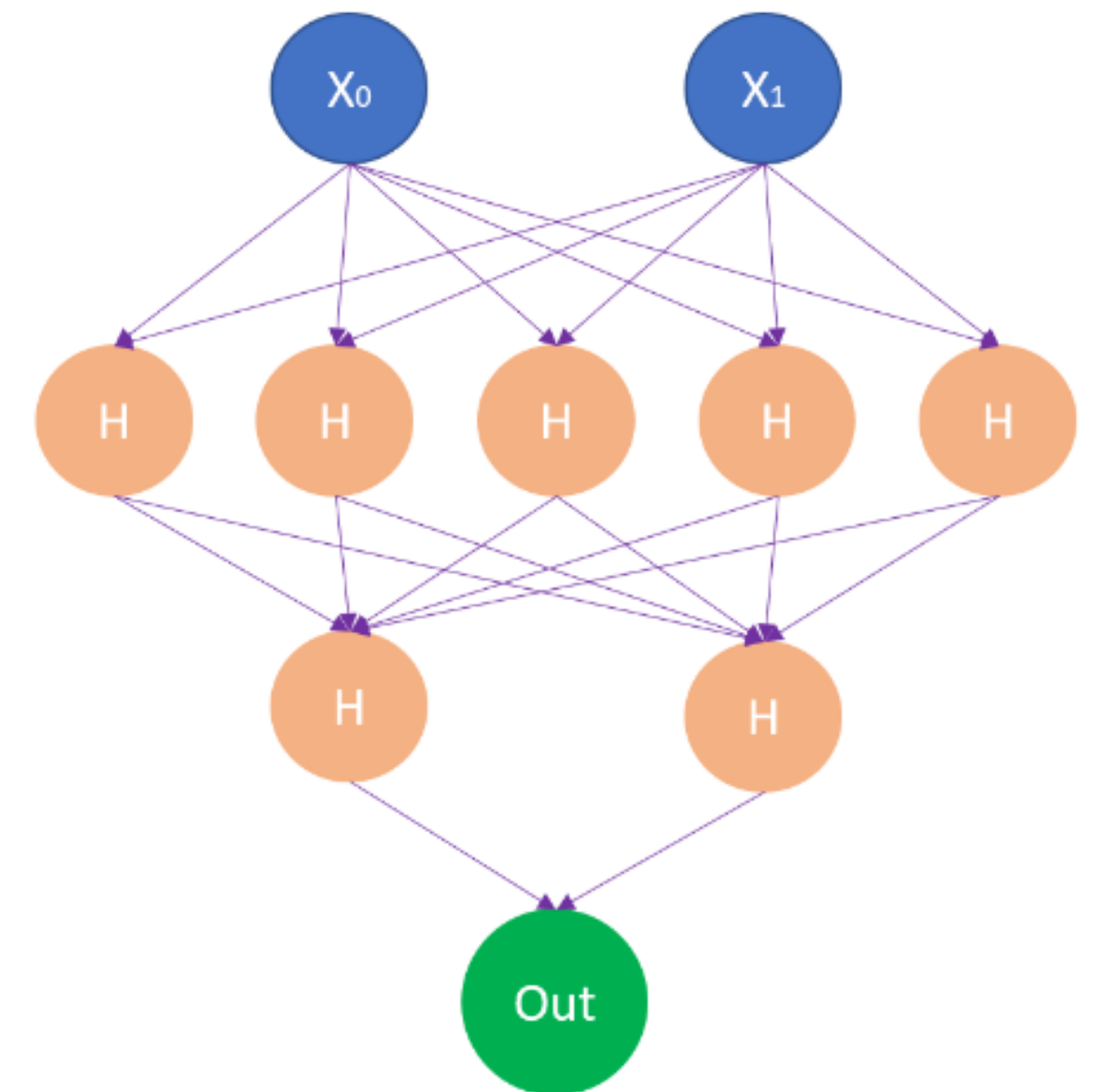
Se vectorizan los datos con ayuda de CountVectorizer. Asimismo, se entrenó y se predijo el modelo con ayuda de MultinomialNB() de la librería Sklearn

El modelo puede funcionar bien para clasificar comentarios de películas como positivos o negativos, la precisión del modelo puede verse afectada por factores como la ambigüedad en el lenguaje y el sarcasmo.

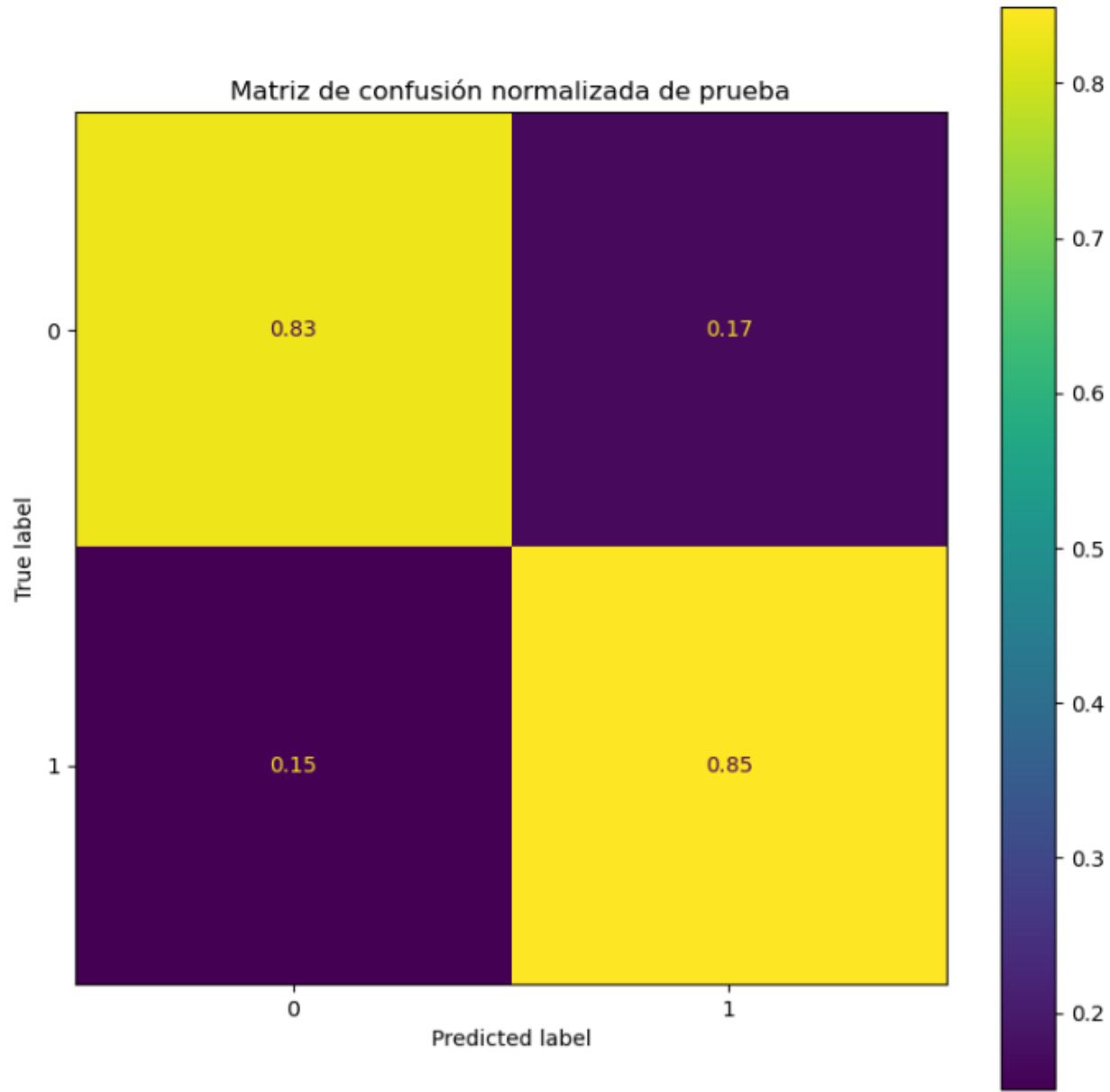
Este algoritmo en específico fue usado para realizar la tarea de clasificación con ayuda de una red neuronal que funciona por medio de iteraciones e hiperparametros funcionales

En este caso, nuestras palabras son los nodos azules y su clasificación iterativa son los nodos naranjas. Finalmente, el nodo verde es el resultado

NEURAL NETWORK



NEURAL NETWORK



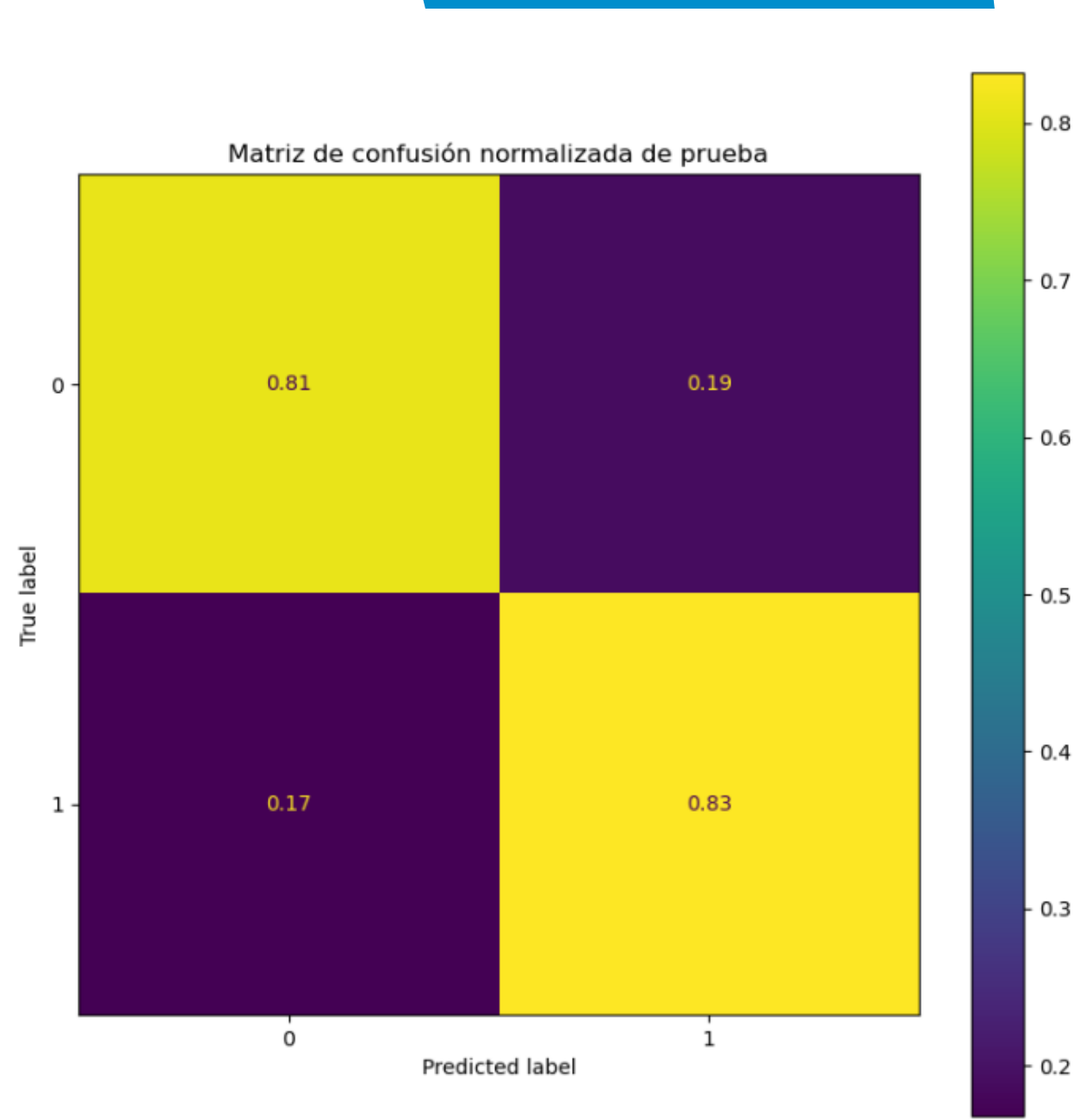
	Accuracy	F1	Precision	Recall
Entrenamiento	0.981990	0.981990	0.982010	0.981990
Prueba	0.840667	0.840644	0.840683	0.840667

En esta prueba se usó **CountVectorizer** como vectorizador para los datos.

NEURAL NETWORK

Al igual que en el caso anterior, realizamos el proceso de clasificación pero usando el vectorizador TF IDF

	Accuracy	F1	Precision	Recall
Entrenamiento	0.964265	0.964265	0.964266	0.964265
Prueba	0.822000	0.821968	0.822024	0.822000



Como puede observarse, el vectorizador CountVectorizer genera un mejor resultado.

RANDOM FOREST

Este modelo es la combinación de varios arboles de decisión. Se usa para aumentar la precisión del modelo, ya que cada árbol tiene acceso a una proporción de los datos de entrenamiento y al unirse se compensan los errores de uno con otro.

Para este modelo es necesario vectorizar los datos. Se usa como método de vectorización TF-IDF que mide la frecuencia de las palabras y le asigna una importancia.

Con los datos ya vectorizados hacemos las primera pruebas del modelo obteniendo los siguientes resultados:

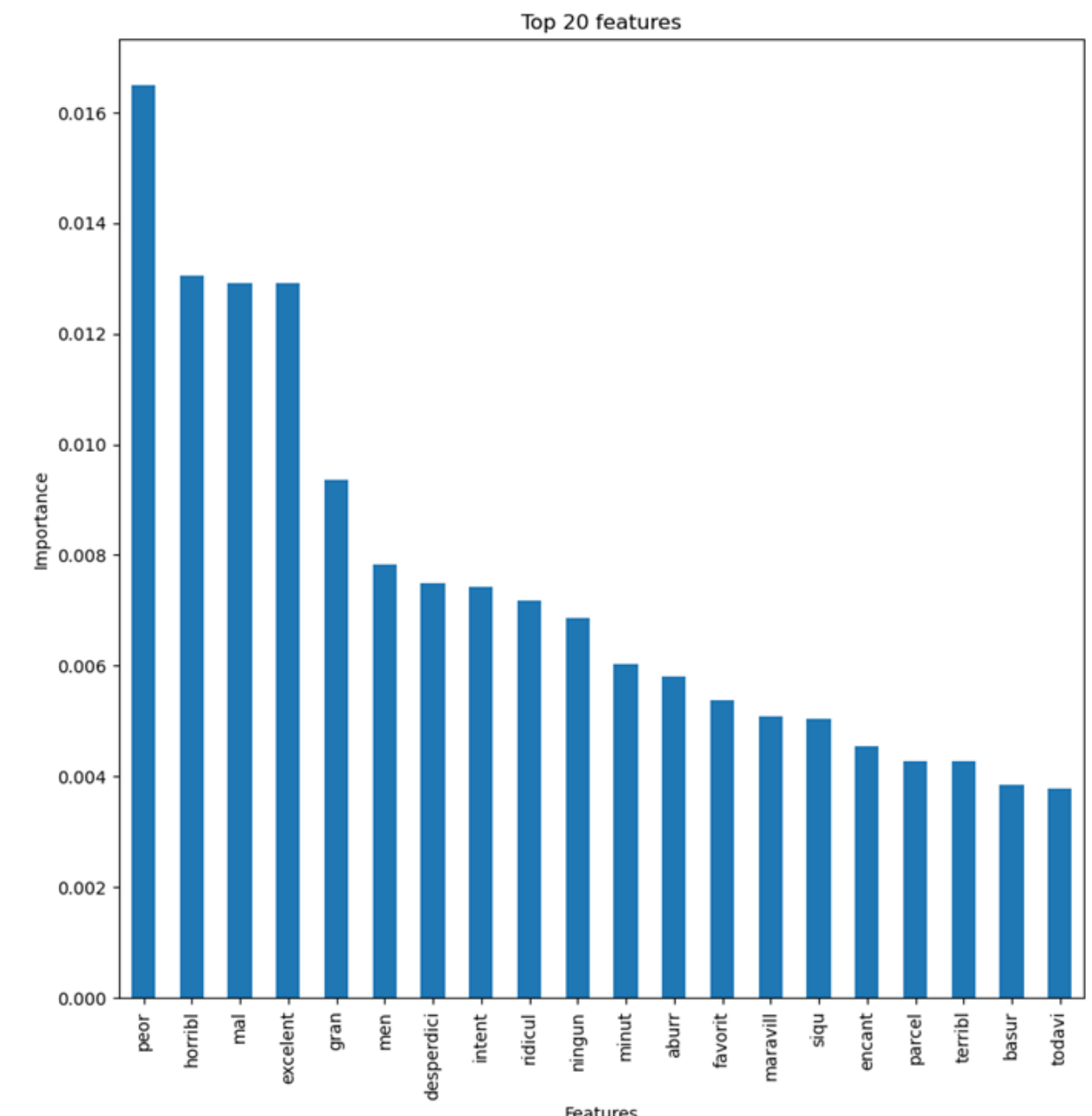
```
[0.77      0.786      0.79      0.766      0.79759519]  
0.7819190380761524
```

RANDOM FOREST

Para mejorar el modelo se hizo una búsqueda de hiperparametros que permitieran aumentar el rendimiento del modelo.

```
MAX DEPTH: 20 / # OF EST: 100 -- A: 0.806 / P: 0.845 / R: 0.757 / F1: 0.799
```

Por ultimo, se puede observar las palabras más importantes para el modelo, que en su mayoría son adjetivos positivos y negativos lo que tiene sentido en el problema propuesto.





CONCLUSIONES

Este proyecto describe la evaluación de tres modelos de análisis de sentimiento (RandomForest, Naive Bayes y red neuronal) para clasificar opiniones en positivas o negativas. Los resultados muestran que el modelo de red neuronal tuvo un rendimiento más alto con un 84.1%, seguido por Naive Bayes y RandomForest. Se concluye que el modelo de red neuronal es el más efectivo en la tarea de clasificación de opiniones en este caso específico, pero siempre hay margen de mejora y se pueden explorar diferentes técnicas y enfoques para mejorar el desempeño.



CONCLUSIONES

Por otra parte, se destaca que los tres modelos pueden ser útiles en diferentes contextos, y se sugiere que la organización podría utilizar el modelo de red neuronal para clasificar opiniones y el modelo RandomForest para identificar las palabras clave y características importantes.