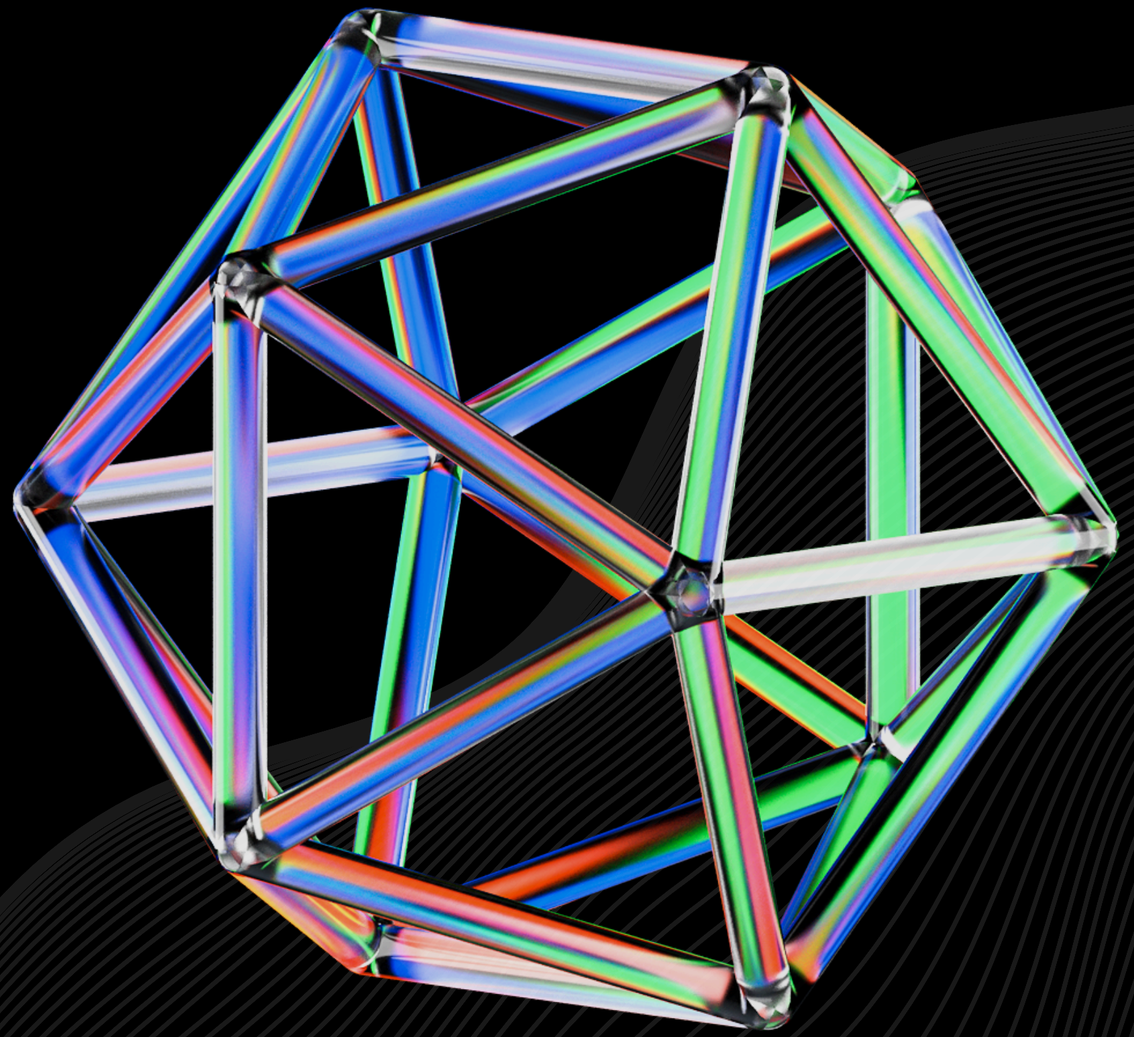
PROYECTO #1

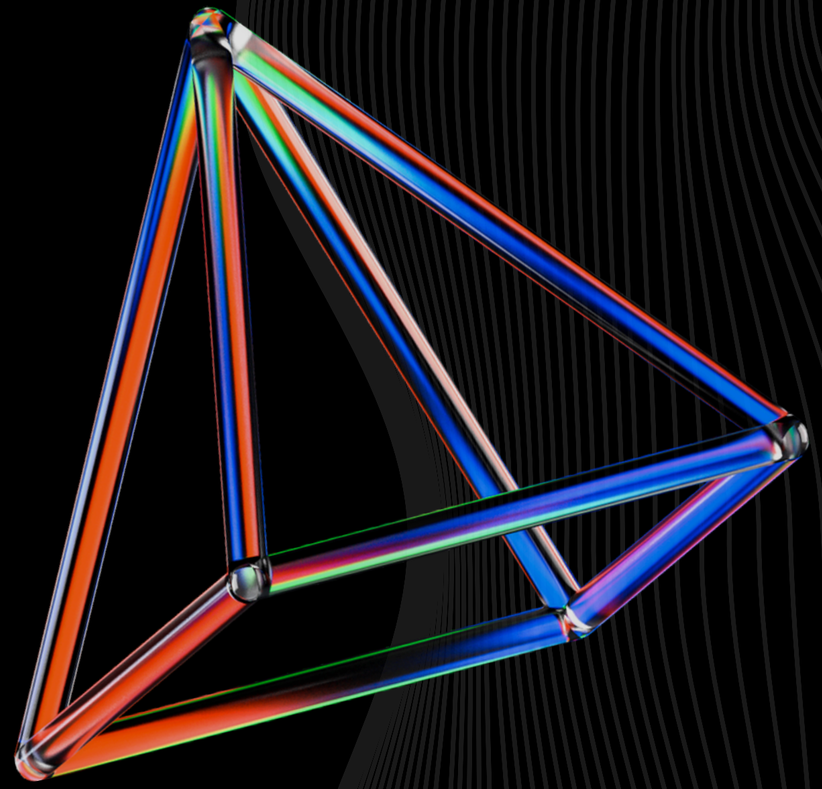
ANALÍTICA DE TEXTOS



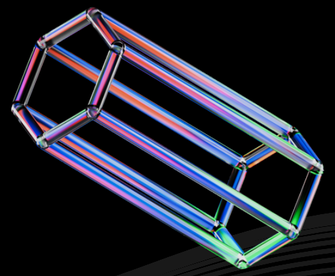
JAIME TORRES - 202014866
MARIA LUISA RODRIGUEZ - 202121549
JUAN GUILLERMO GUTIERREZ - 202122659

Clasificación de Texto

La primera etapa tiene como objetivo analizar las opiniones de los habitantes locales para identificar problemáticas de su entorno relacionadas con los ODS 3, 4 y 5, de forma que se pueda construir un modelo analítico que relacione automaticamente ciertas opiniones con ciertas palabras relevantes a su contexto.

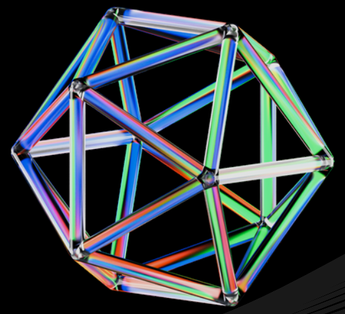


Pre-Procesamiento de Datos



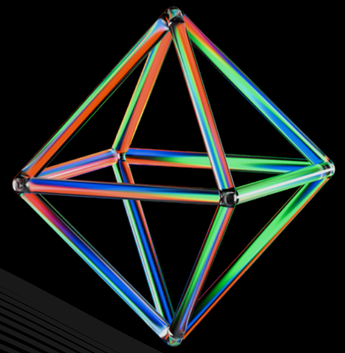
Limpieza

Los datos del archivo de prueba deben limpiarse para que sólo la información relevante al modelo sea analizado por este. Esto incluye que todo el texto este en minúsculas, que se elimine la puntuación y que se omitan las ‘stopwords’



Tokenización

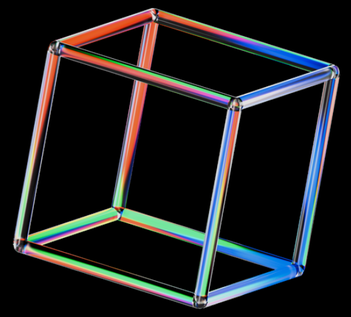
Posteriormente, las palabras se dividen de forma que los modelos puedan analizarlas de manera individual facilmente. Estas se almacenan en una columna adicional para no alterar los datos originales del archivo.



Normalización

Finalmente, se utiliza la técnica de 'stemming', la cual permite reducir las palabras a su raíz. Esto permite que varias palabras que se utilizan para referirse al mismo concepto puedan ser entendidas por el modelo como un solo token.

K-Nearest Neighbor

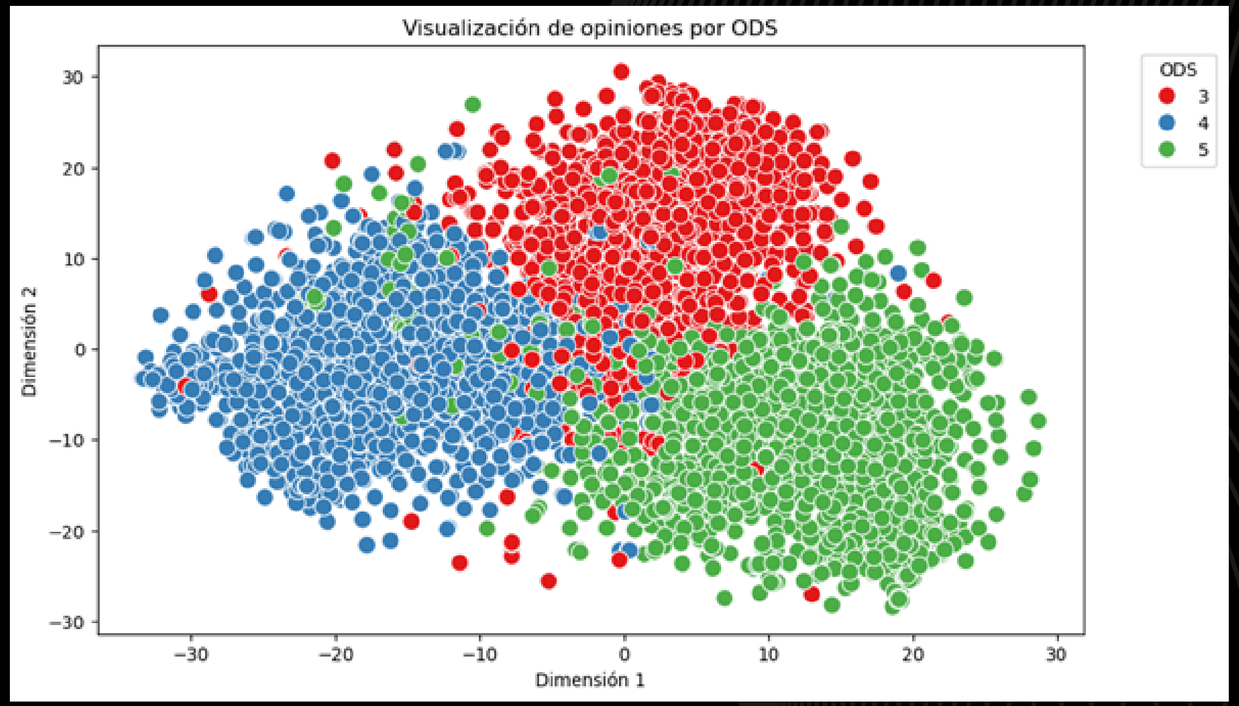


Este algoritmo nos permite, a partir de comparar los datos del problema con otros datos cercanos, decidir una clasificación para este

-Algunos valores dan bastante lejanos a su distribucion esperada

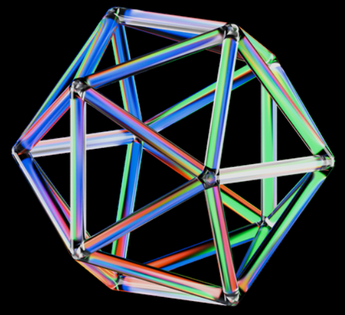
-en general agrupa bastantes datos

-k = 15



	precision	recall	f1-score	support
3	0.96	0.94	0.95	269
4	0.90	0.97	0.93	266
5	0.97	0.92	0.94	275
accuracy			0.94	810
macro avg	0.94	0.94	0.94	810
weighted avg	0.94	0.94	0.94	810

Random Forest

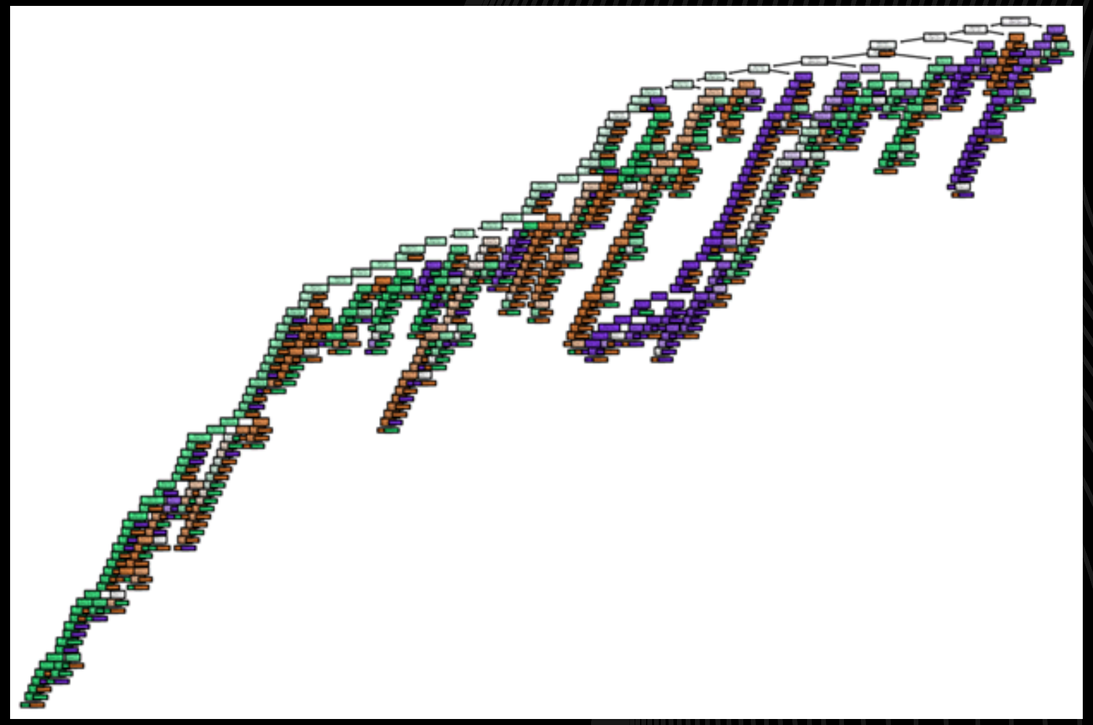


Genera un grupo de arboles de decisión a partir de los datos encontrados

-Arboles bastante balanceados

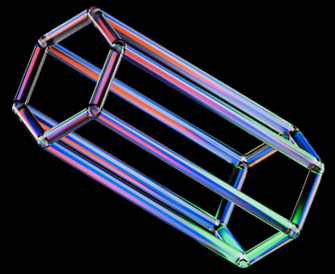
-Depende de la complejidad del problema

-No controlamos por profundidad. por lo que podrían en teoría llegar a una profundidad de n datos



	precision	recall	f1-score	support
3	0.98	0.97	0.98	269
4	0.97	0.96	0.97	266
5	0.95	0.97	0.96	275
accuracy			0.97	810
macro avg	0.97	0.97	0.97	810
weighted avg	0.97	0.97	0.97	810

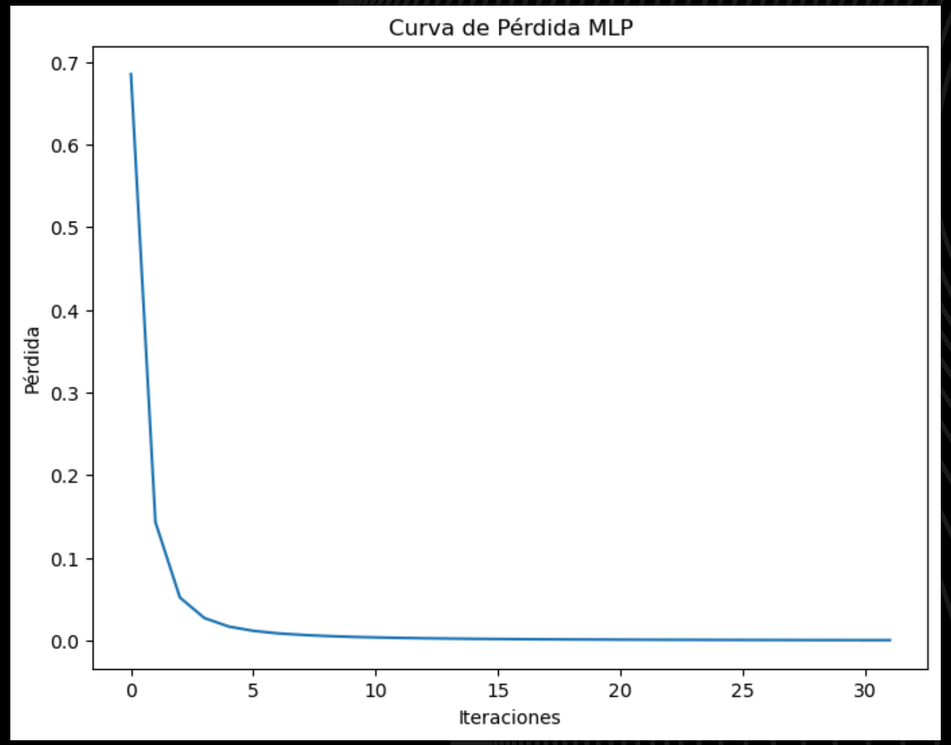
Multi-Layer Perceptron



Sistema de redes neuronales que encuentra patrones dentro del conjunto de datos presentado

-fue el mejor modelo encontrado

-Tiene una curva de perdida bastante pronunciada y suave (implica que esta aprendiendo bien)



	precision	recall	f1-score	support
3	0.99	0.99	0.99	269
4	0.97	0.98	0.98	266
5	0.97	0.97	0.97	275
accuracy			0.98	810
macro avg	0.98	0.98	0.98	810
weighted avg	0.98	0.98	0.98	810

Comparación de Resultados

El modelo MLP Classifier obtuvo el mejor desempeño de los tres modelos, teniendo en cuenta todos sus valores de f1-score. Esto podría explicarse por la capacidad de MLP de entender relaciones más complejas que los demás modelos, al igual que su habilidad de manejar la gran cantidad de tokens obtenidos de los datos de prueba.

accuracy			0.94	810
macro avg	0.94	0.94	0.94	810
weighted avg	0.94	0.94	0.94	810

accuracy			0.97	810
macro avg	0.97	0.97	0.97	810
weighted avg	0.97	0.97	0.97	810

accuracy			0.98	810
macro avg	0.98	0.98	0.98	810
weighted avg	0.98	0.98	0.98	810