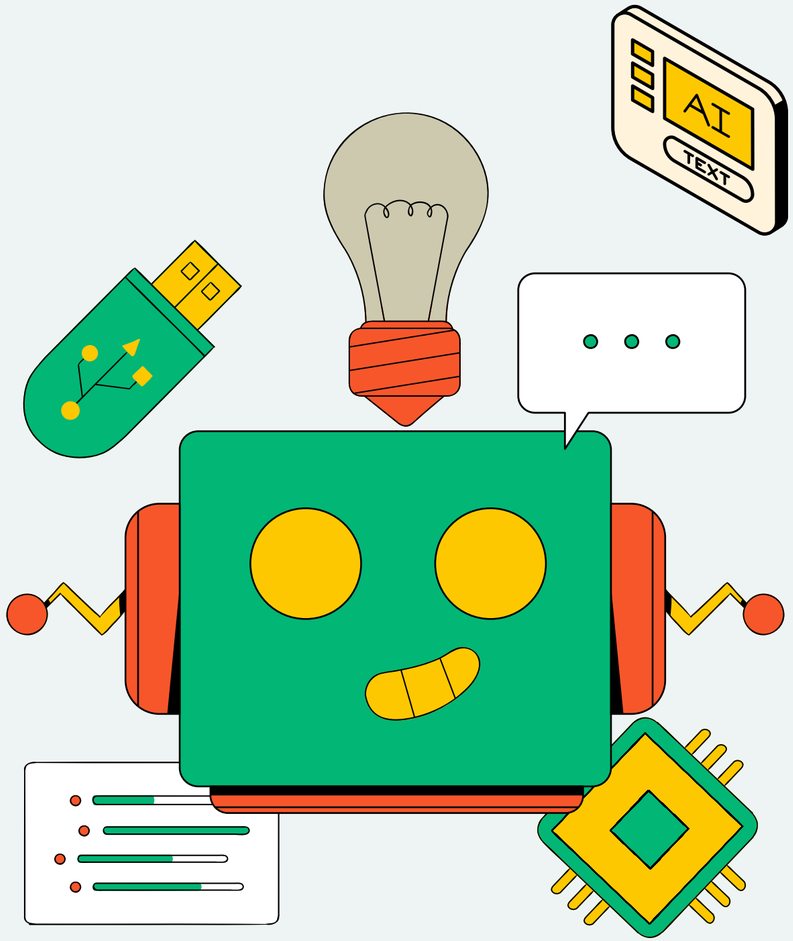
THYNK UNLIMITED

WE LEARN FOR THE FUTURE



PROYECTO ETAPA1 ANALÍTICA DE TEXTOS

PRESENTATION

PRESENTED BY:

NICOLÁS RINCÓN

SANTIAGO JAIMES

NICOLÁS CASAS

ETAPA1

BENEFICIOS Y REQUERIMIENTOS FUNCIONALES

Como organización sin animo de lucro junto a entidades del sector público deseamos a partir de la recopilación de opiniones de los ciudadanos, clasificarlas entre los Objetivos Desarrollo Sostenible (3,4,5), de esta forma, se ahorran recursos en saber las necesidades de las personas.

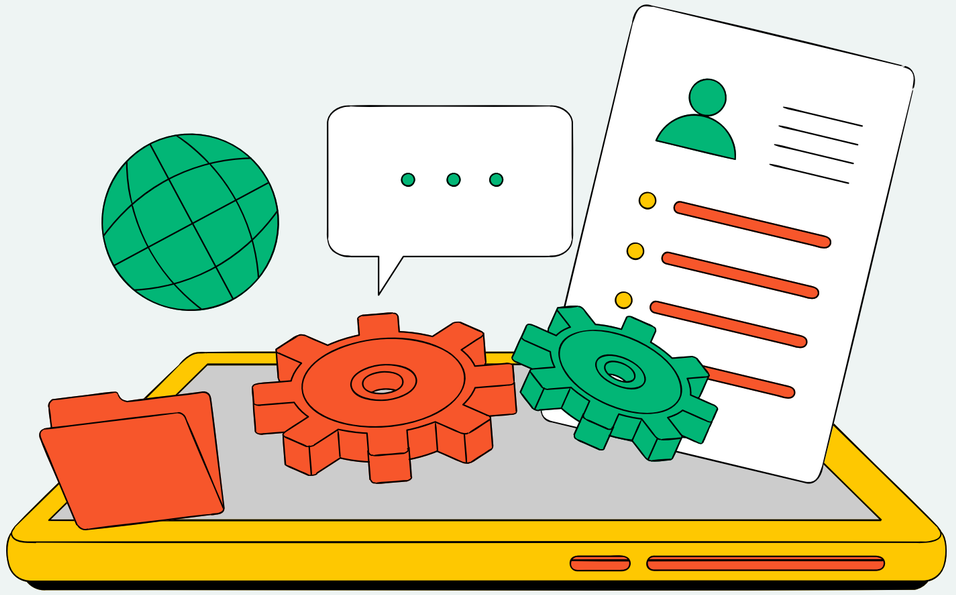
- Crear modelos que permiatan clasificar las opiniones de los ODS
- Identificar las palabras clave para caracterizar los ODS.

ENTENDIMIENTO DE DATOS

CALIDAD DE DATOS

Número de Registros: 4049

Número de Categorías: 3



1. **Unicidad:** Correcto
2. **Completitud:** Correcto
3. **Consistencia:** Correcto
4. **Validez**: Correcto

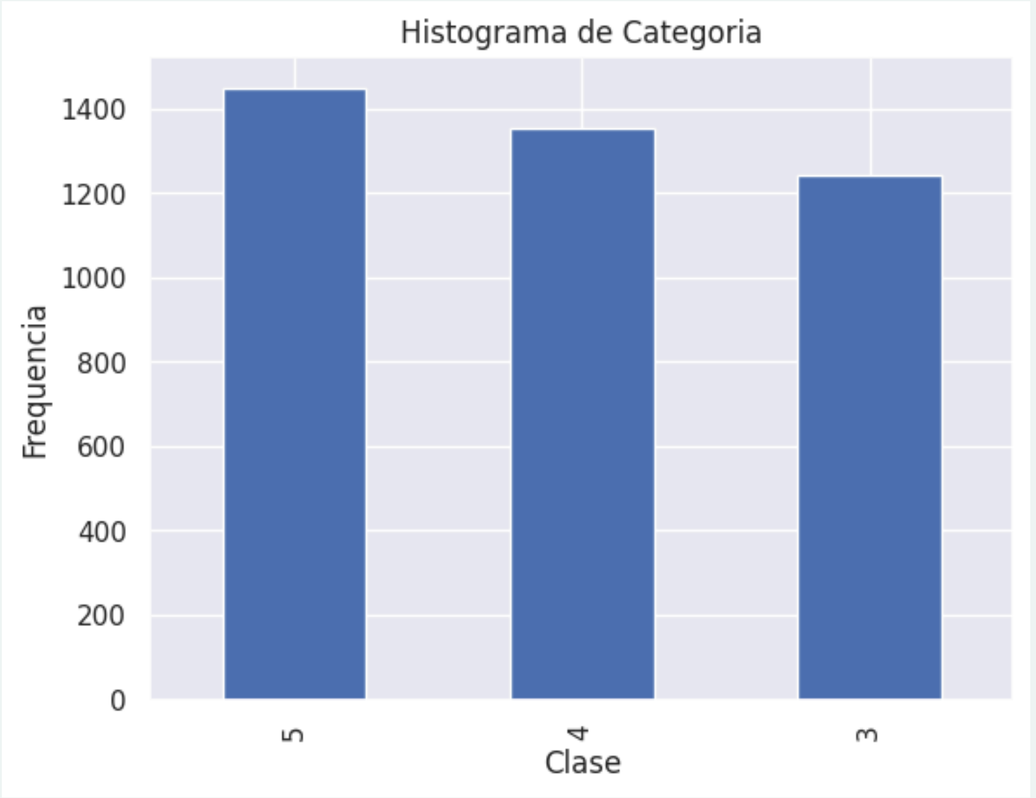
EXPLORACIÓN DE LOS DATOS

- Idiomas detectados:

99.6% Español, 0.2% Inglés, 0.1% Francés.

Menor # palabras: 699

Mayor # palabras: 1513



PREPARACIÓN DE LOS DATOS: DECISIONES TOMADAS

1. **Usar solo Minúsculas:** Disminuye conjunto de las palabras
2. **Eliminar todos lo relaciona a números:** No aporta información
3. **Eliminar Puntuación:** No aporta información
4. **Codificación de latin1 y decodificación de UTF-8:** Para asegurarnos que el texto no tuviera errores por la codificación que tiene la información proporcionada.
5. **Eliminar stopwords:** No aportan información
6. **Lematizar Palabras:** Extraer los lemas de las palabras
7. **Stemming:** Reducir la palabra a su raiz.

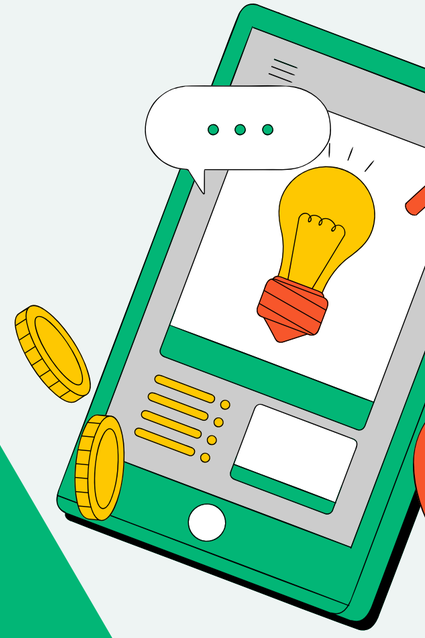
REPRESENTACIÓN: BOLSAS DE PALABRAS TF-IDF

Frecuencia Absoluta

Equilibrio de Frecuencia

Equilibrio de Relevancia

Alta dimensionalidad



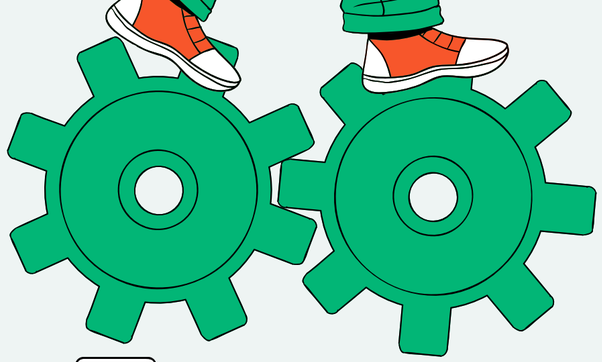
CREACIÓN MODELOS DE CLASIFICACIÓN

- Se desarrollaron 3 modelos con el propósito de encontrar el más preciso para clasificar las opiniones de los ciudadanos.

En este sentido el modelo se hizo a través de:

- Naive Bayes.
- Regresión Logística Multinomial
- Random Forest

MODELO 1: NAIVE BAYES



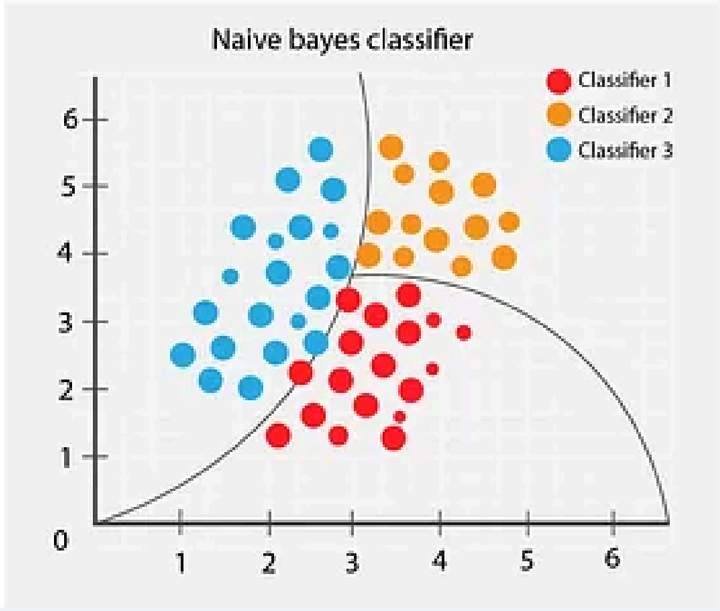
thatware.co

In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naive) independence assumptions between the features.

$$P(A|B) = \frac{P(B|A)\ P(A)}{P(B)}$$

using Bayesian probability terminology, the above equation can be written as

$$\text{Posterior} = \frac{\text{prior x likelihood}}{\text{evidence}}$$

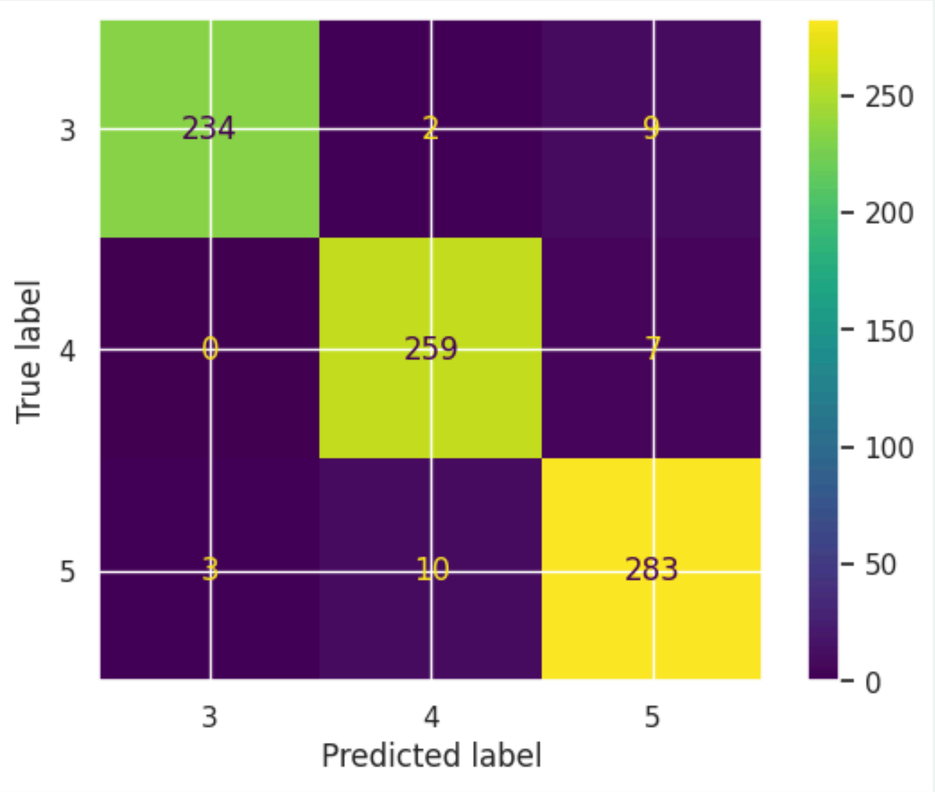


Tomado de: https://thatware.co/wp-content/uploads/2020/04/naive-bayes.png

RESULTADOS NAIVE BAYES

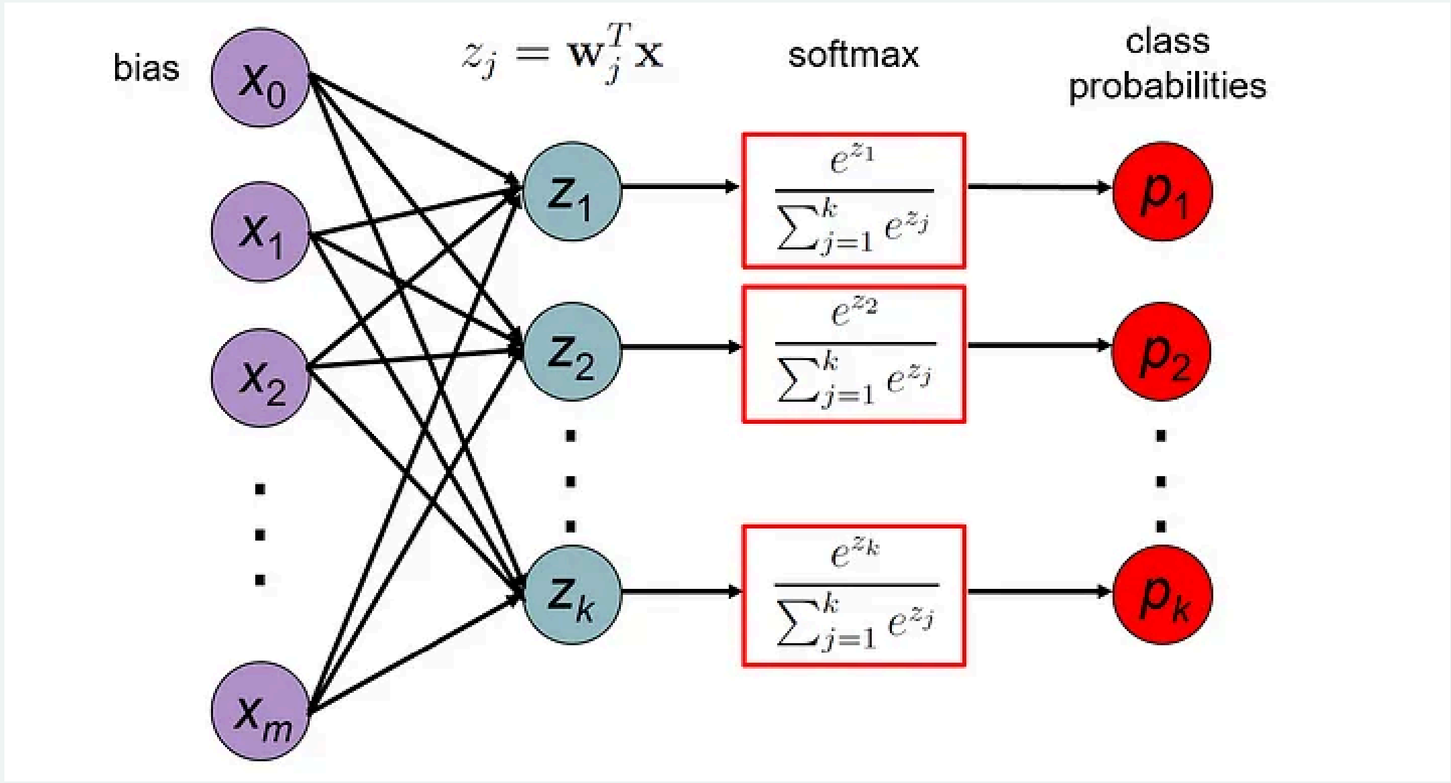
Métricas del Modelo

- **Accuracy: 96.1%**
- **Precision: 96.2%**
- **Recall: 96.1%**
- **F1 Score: 96.1%**



Se obtiene que el acurracy fue de 96.1% y se refiere a la exactitud en que tantos datos fueron catalogados correctamente en promedio. También la precisión fue de 96.2% y se refiere a los que fueron categorizados correctamente de una categoría, el Recall cuantos fueron categorizados correctamente del todo de datos de una categoría real. Y el F1 score media el precision y recall.

MODELO 2: REGRESIÓN LOGÍSTICA MULTINOMINAL

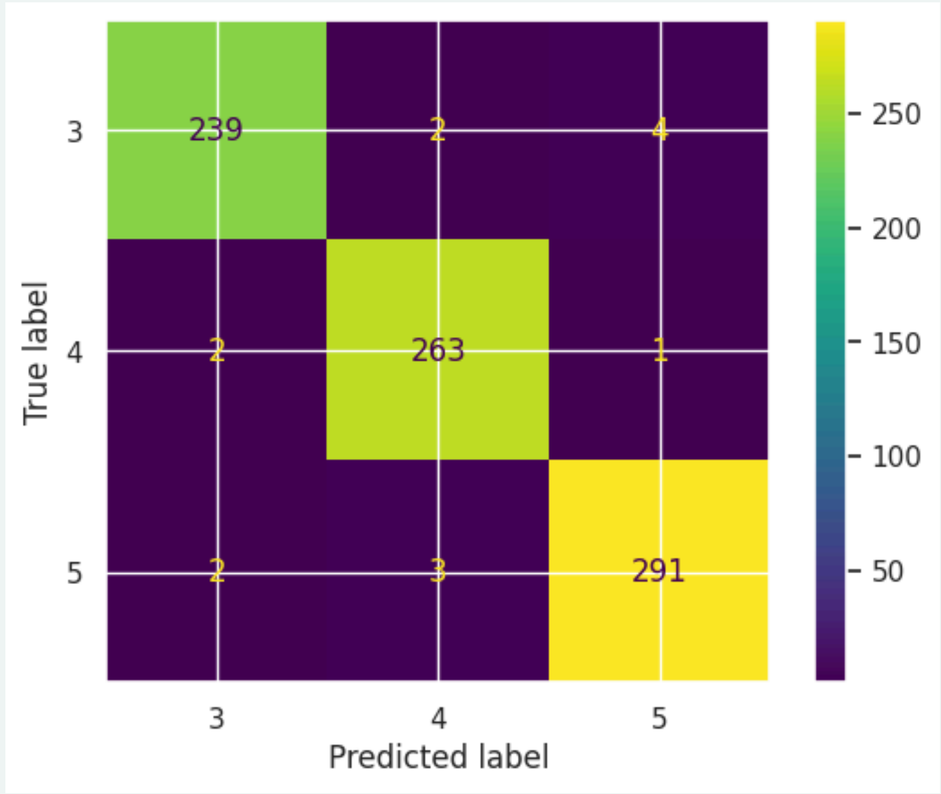


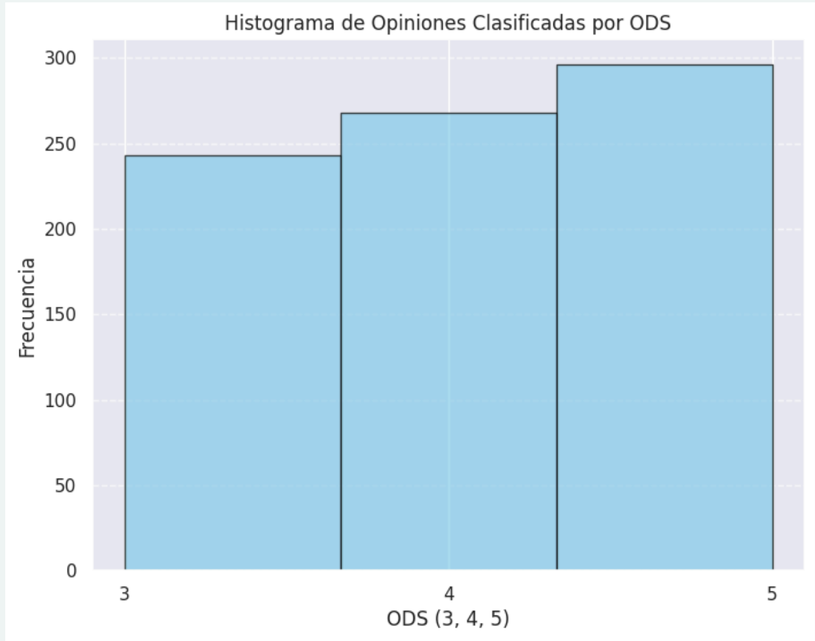
Tomado de: https://towardsdatascience.com/deep-dive-into-softmax-regression-62deea103cb8

RESULTADOS REGRESIÓN LOGÍSTICA MULTINOMIAL

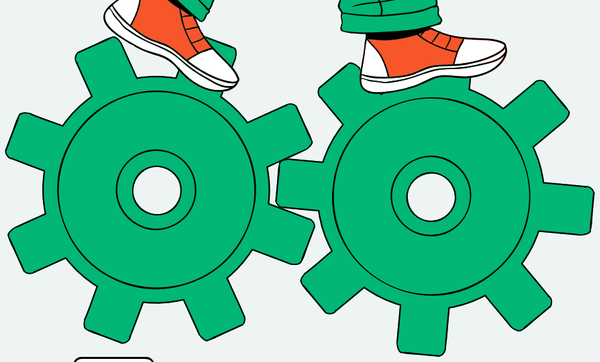
Métricas del Modelo

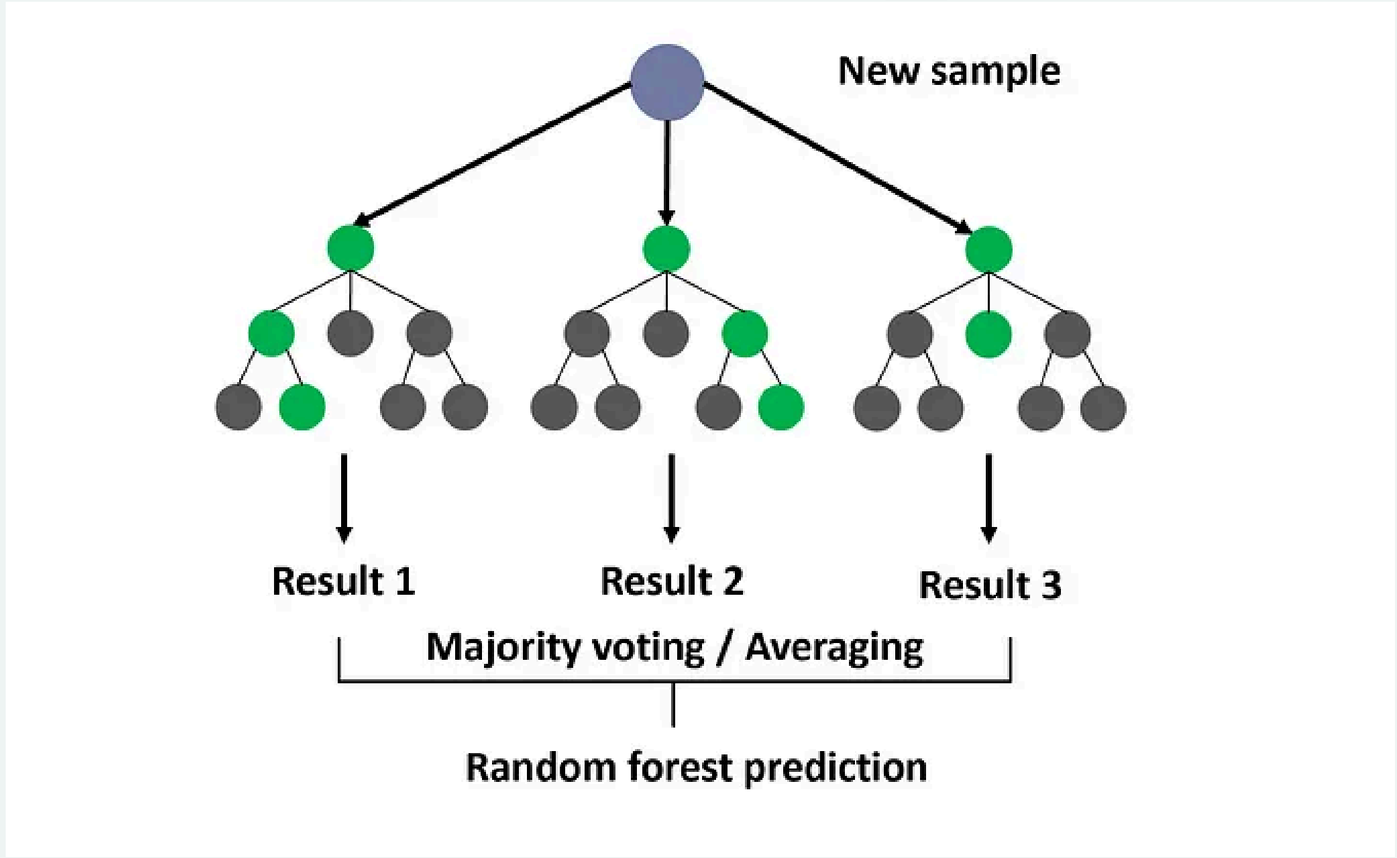
- **Accuracy: 98.2%**
- **Precision: 98.2%**
- **Recall: 98.2%**
- **F1 Score: 98.2%**





MODELO 3: RANDOM FORESTS



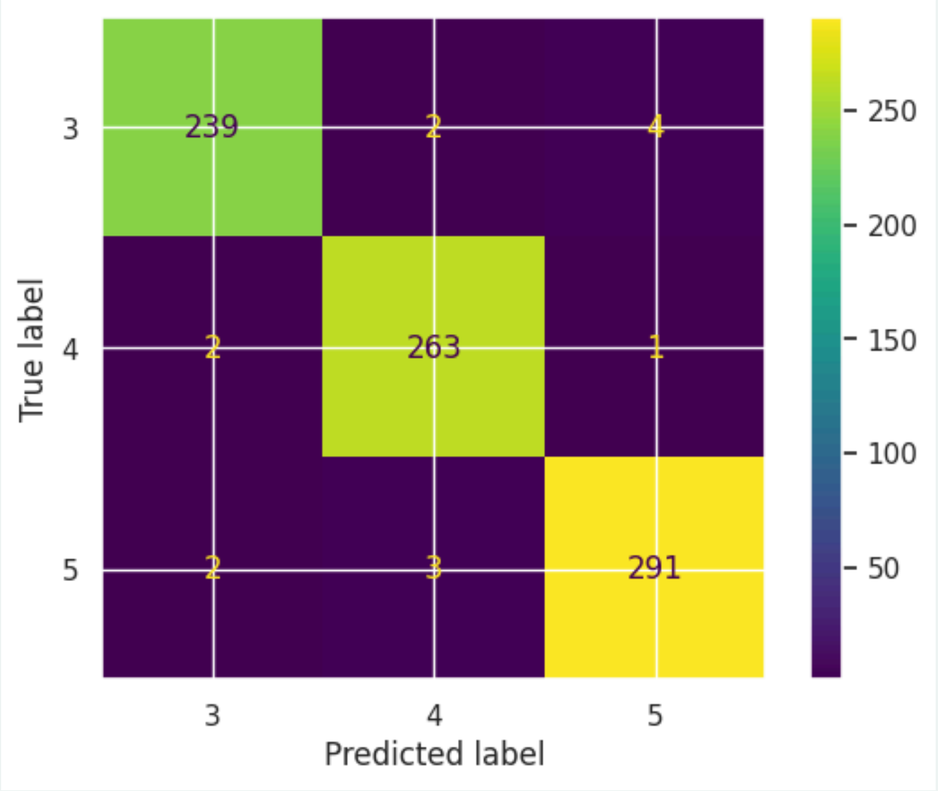


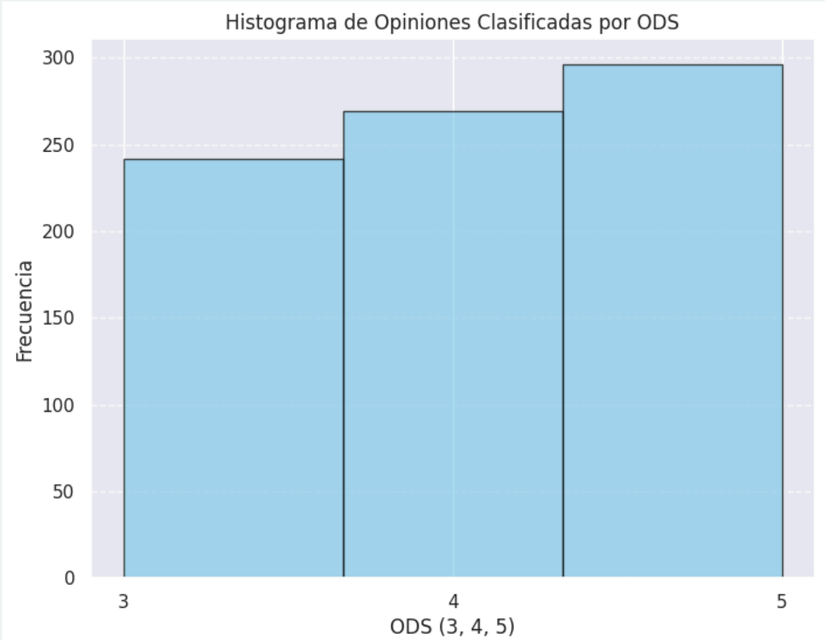
Tomado de: https://medium.com/@roiyeho/random-forests-98892261dc49

RESULTADOS RANDOM FOREST

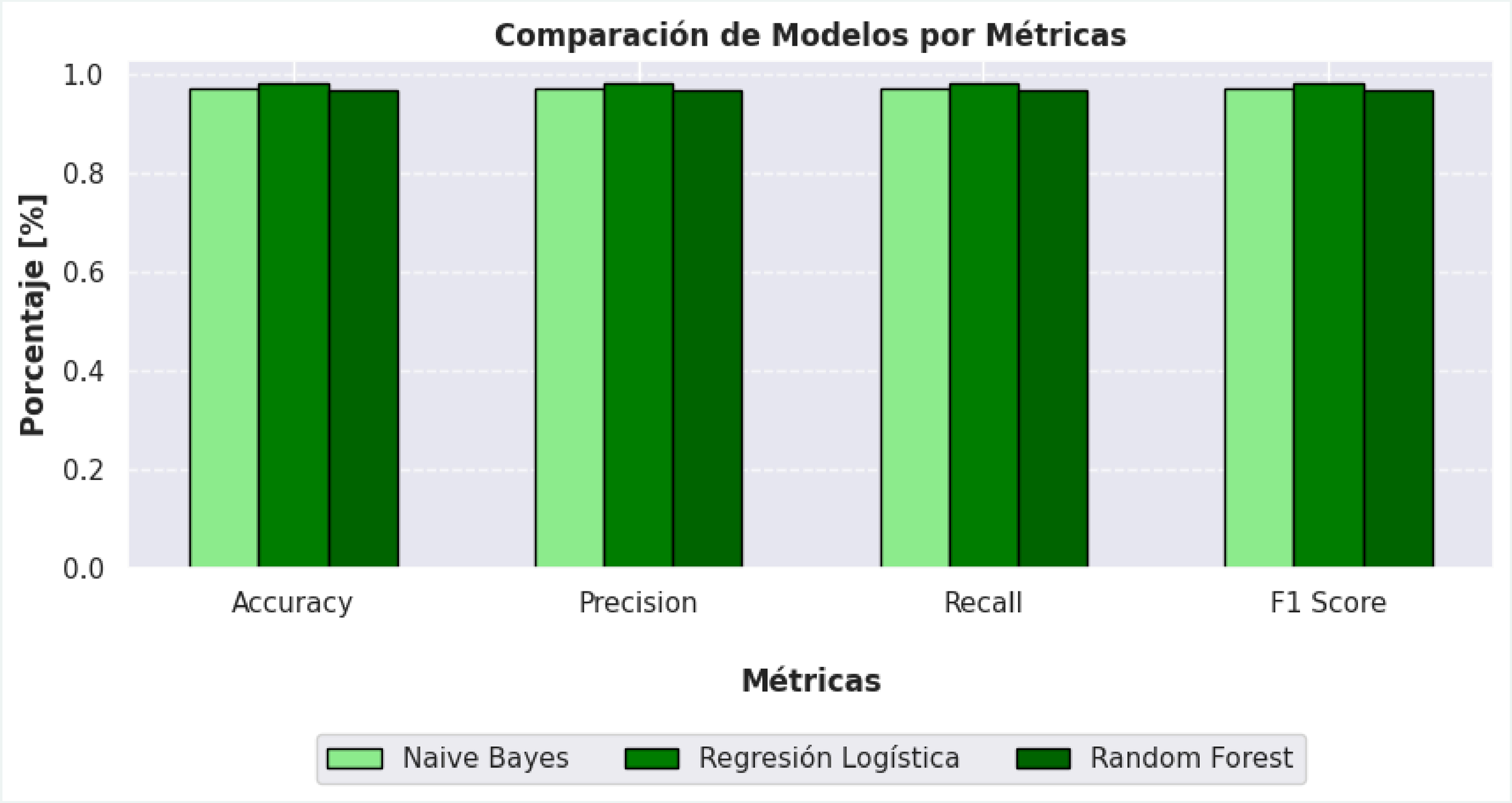
Métricas del Modelo

- **Accuracy: 96.9%**
- **Precision: 96.9%**
- **Recall: 96.9%**
- **F1 Score: 96.9%**

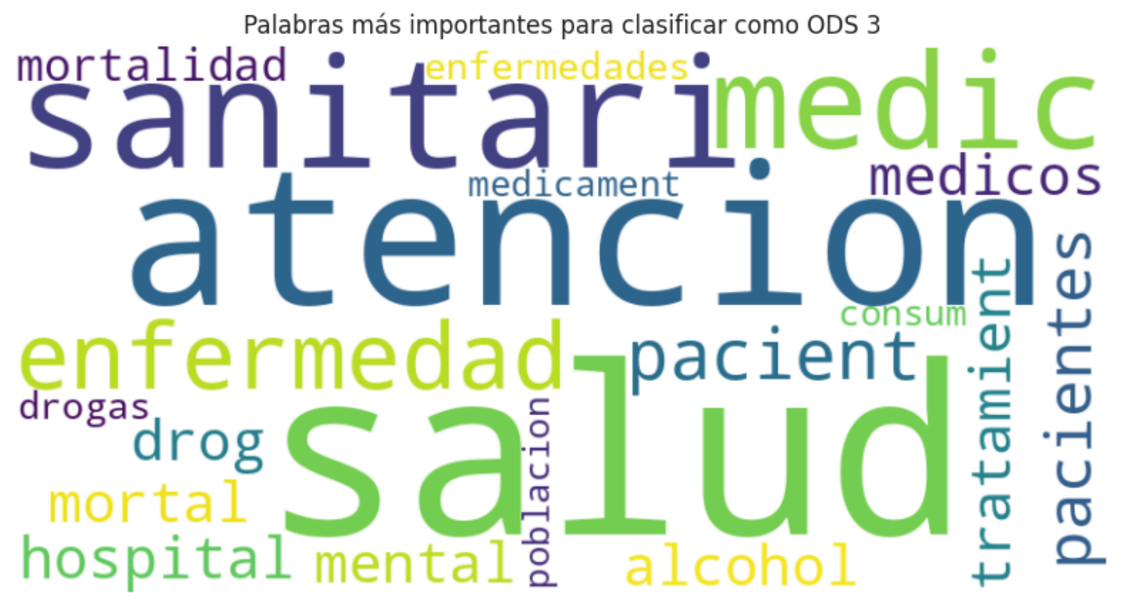


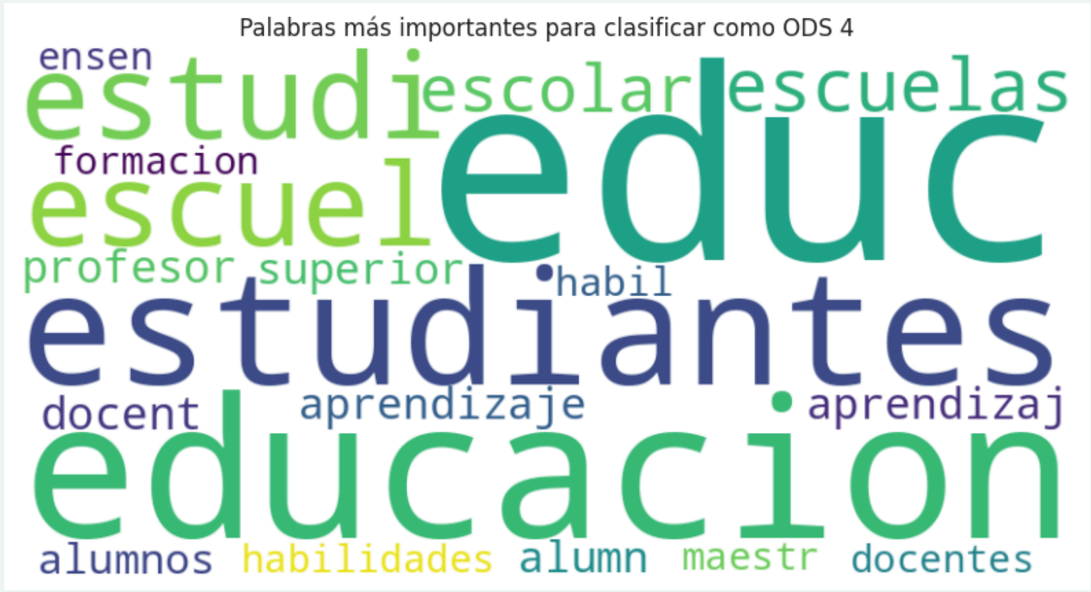


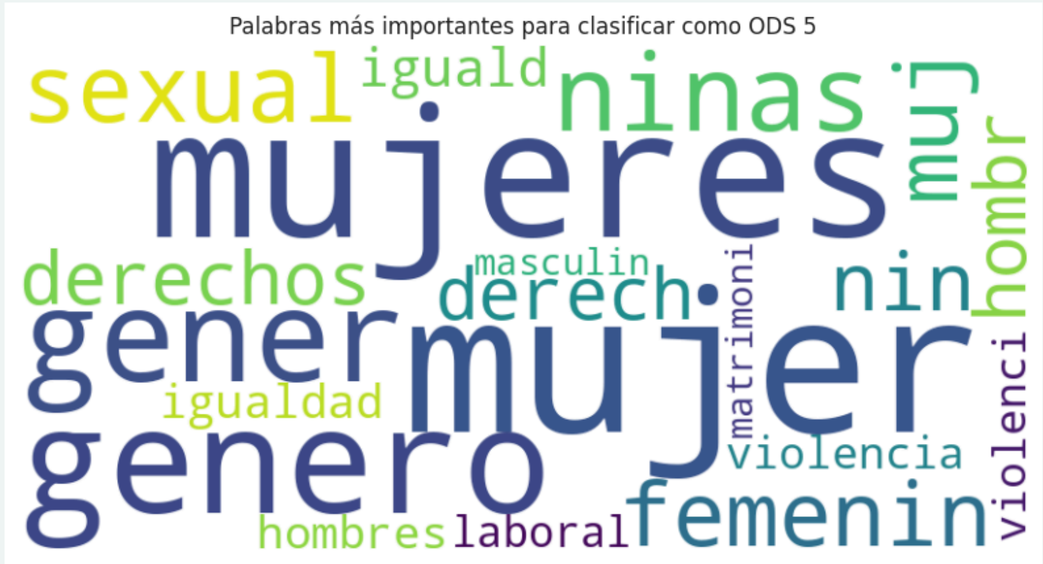
Comparación de Resultados de cada modelo



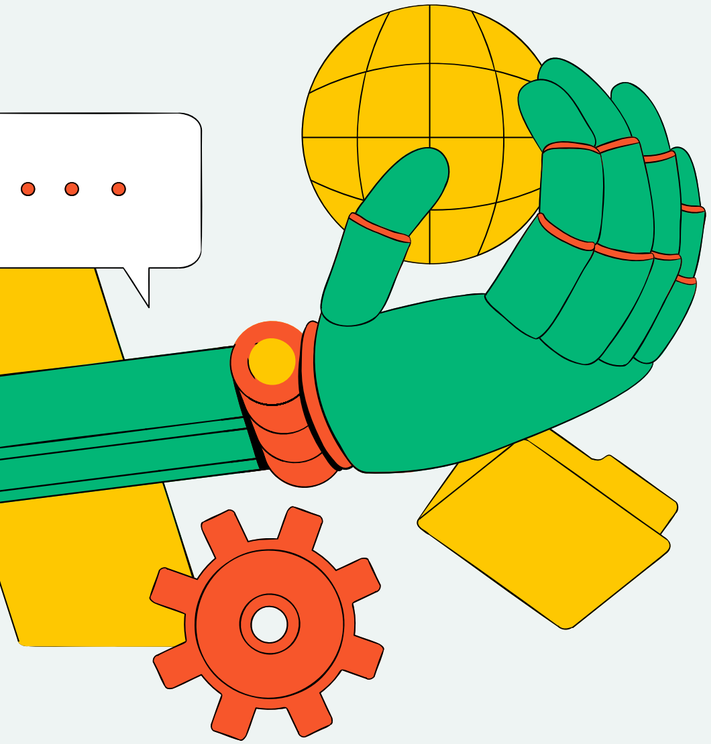
Mejor Modelo: Análisis de Resultados en relacion a los objetivo de negocio.Regresión Logística Multinomial







MEJOR MODELO: CONCLUSIONES



1. Casi el 98% de las opiniones escritas por los ciudadanos se lograron clasificar correctamente, lo cual está en un nivel bastante aceptable e indica que el modelo es altamente confiable.
2. El modelo es de gran utilidad para poder identificar las características que más impactan en la vida cotidiana de cada ciudadano.
3. A partir del objeto de negocio otorgado por la UNFPA y reconociendo su participación con otras entidades públicas y ciudadanos, se recomendaría que el cliente (UNFPA) utilize el mejor modelo construido para generar predicciones de las opiniones de los ciudadanos en el futuro y cumplir su metas al 2030.