

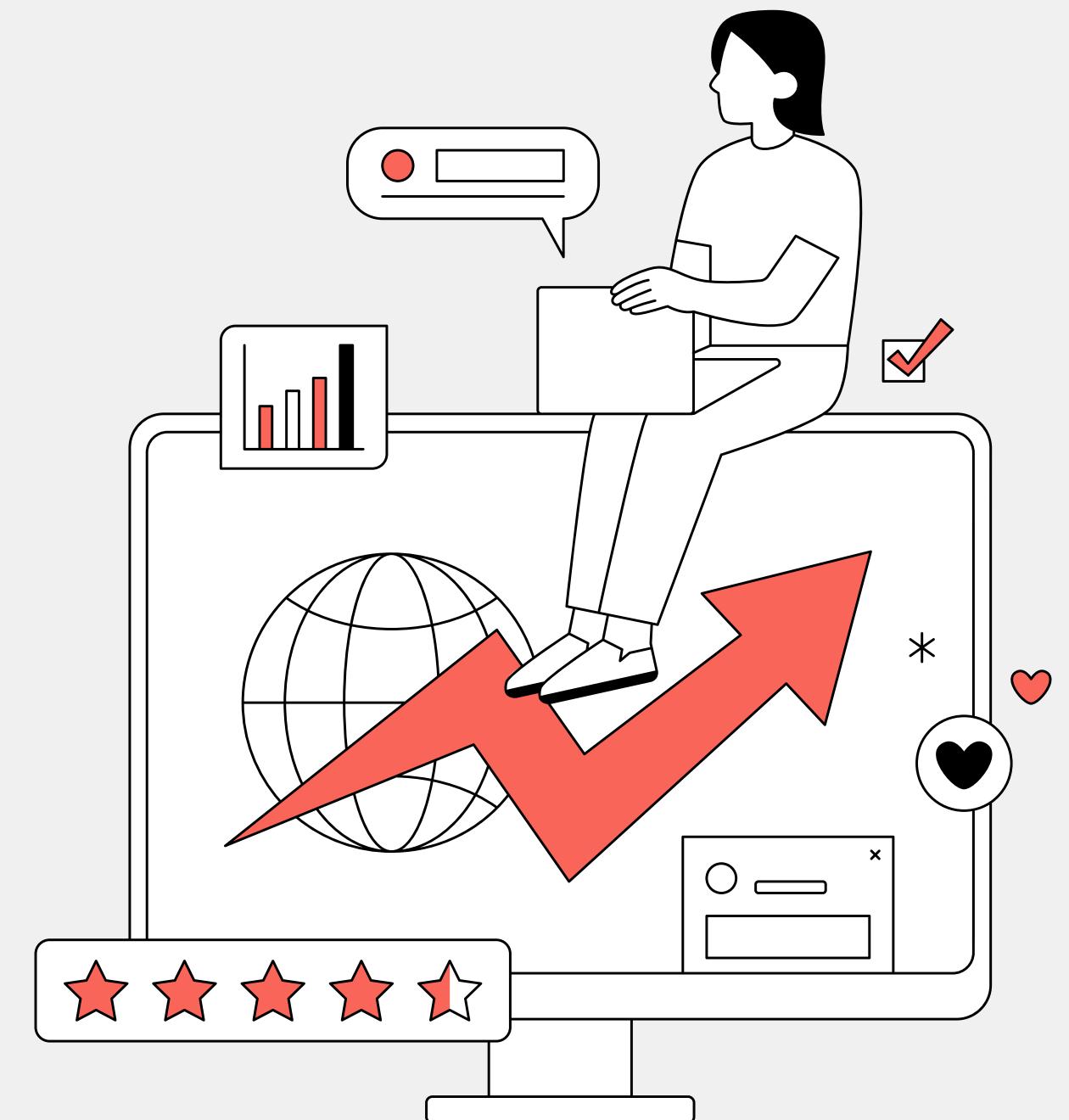
Hecho por:
Juan Manuel Ramírez Tamayo
Santiago Celis Rengifo
Luis Alfredo Borbón Holguín

PROYECTO 1

ETAPA 1

Inteligencia de Negocios

GRUPO 22



Agenda

Contexto del
negocio

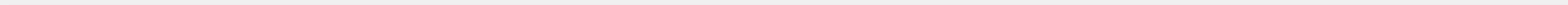
Entendimiento de
los datos

Limpieza de los
datos

Implementación
de modelos

Resultados

Conclusión





Contexto del negocio

INTRODUCCIÓN AL PROBLEMA

- La distribución de noticias falsas afecta procesos democráticos y de seguridad a lo largo del mundo.
- Las fake news alteran votaciones, generan desinformación y disminuyen la confianza en las instituciones.
- Este proyecto busca desarrollar un modelo capaz de identificar noticias falsas analizando su contenido.
- **Beneficiarios finales:** Empresas de comunicación, redes sociales y verificadores de datos.

¿Qué queremos lograr?

- Clasificar noticias como verdaderas (1) o falsas (0) usando un modelo de aprendizaje automático.
 - Utilizar datos textuales (descripciones - cuerpo de las noticias) para entrenar algoritmos predictivos y priorizar confiabilidad.
 - Comparar y seleccionar el mejor modelo en términos de precisión y recall.
- 

Contexto del negocio



Entendimiento de los datos

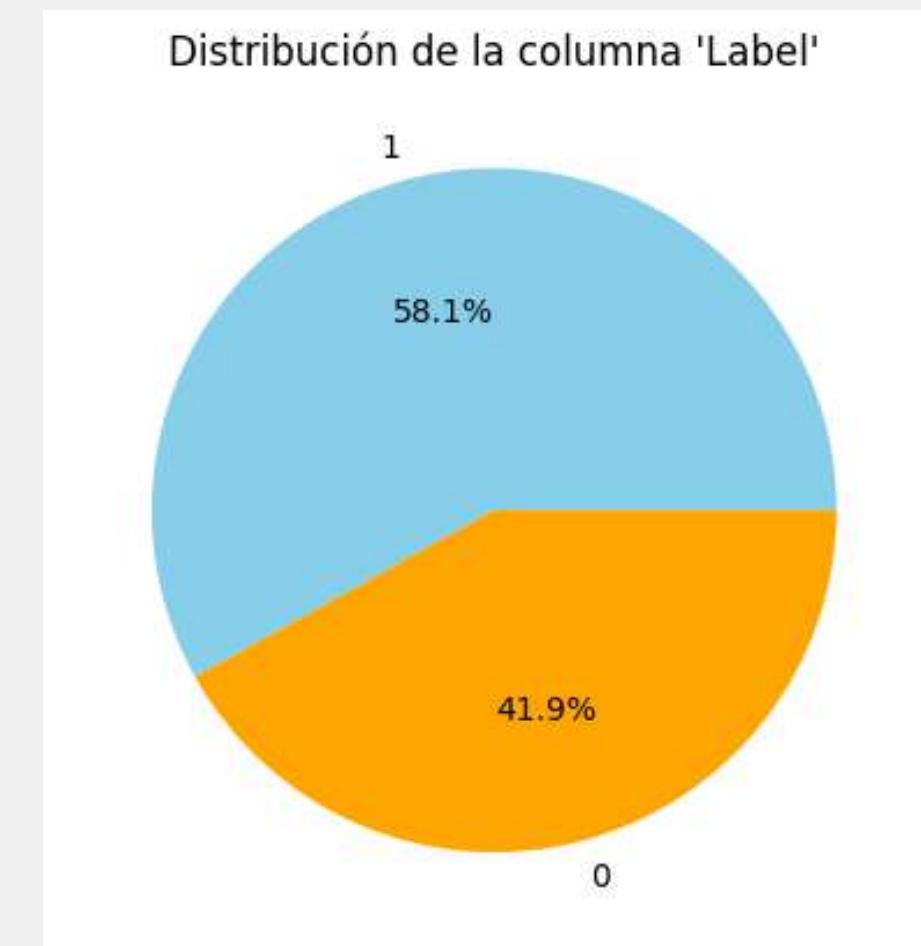
Descripción del conjunto de datos

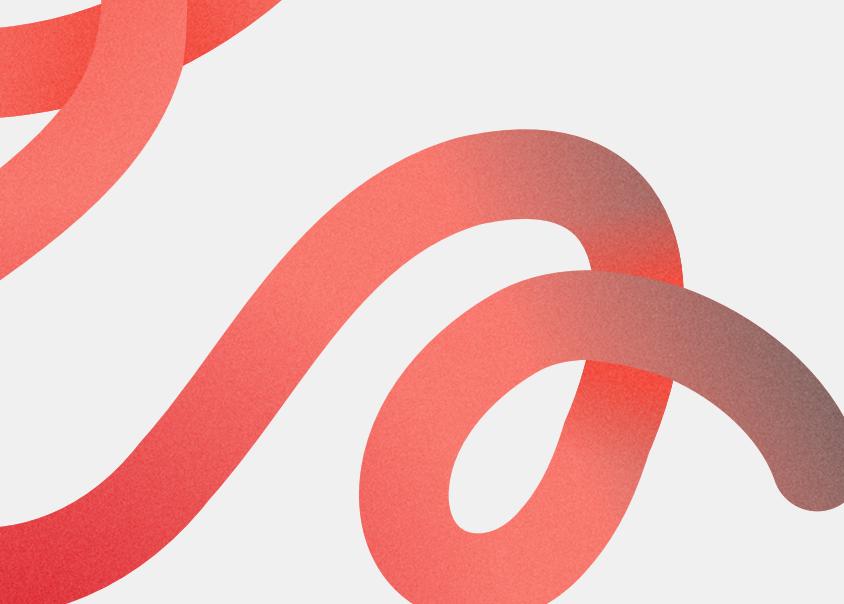
Columnas clave:

- **ID:** Identificación única de la noticia.
- **Label:** 1 (verdadera) o 0 (falsa).
- **Título:** Encabezado de la noticia.
- **Descripción:** Contenido principal de la noticia (texto).
- **Fecha:** Fecha de publicación.

Distribución de las etiquetas:

- 58% noticias verdaderas.
- 42% noticias falsas.





Limpieza de datos

Proceso de Transformación

- **Eliminación de duplicados:** Se eliminaron registros basados en la columna 'Descripción', reduciendo los registros de 60215 a 49638.
- **Limpieza textual:**
 - Eliminación de caracteres especiales.
 - Paso a minúsculas.
 - Eliminación de stopwords (palabras vacías).
 - Lematización para reducir palabras a su forma base.
- **Resultado:** Un dataset limpio, único y consistente para su análisis.

Transformaciones Posteriores

- Textos categóricos (descripciones) transformados en vectores numéricos utilizando TF-IDF.
 - TF-IDF: Asigna mayor peso a palabras relevantes descartando las comunes.
 - Vector final: Cada descripción convertida a una matriz numérica procesable por los algoritmos.
- 



Implementación de modelos

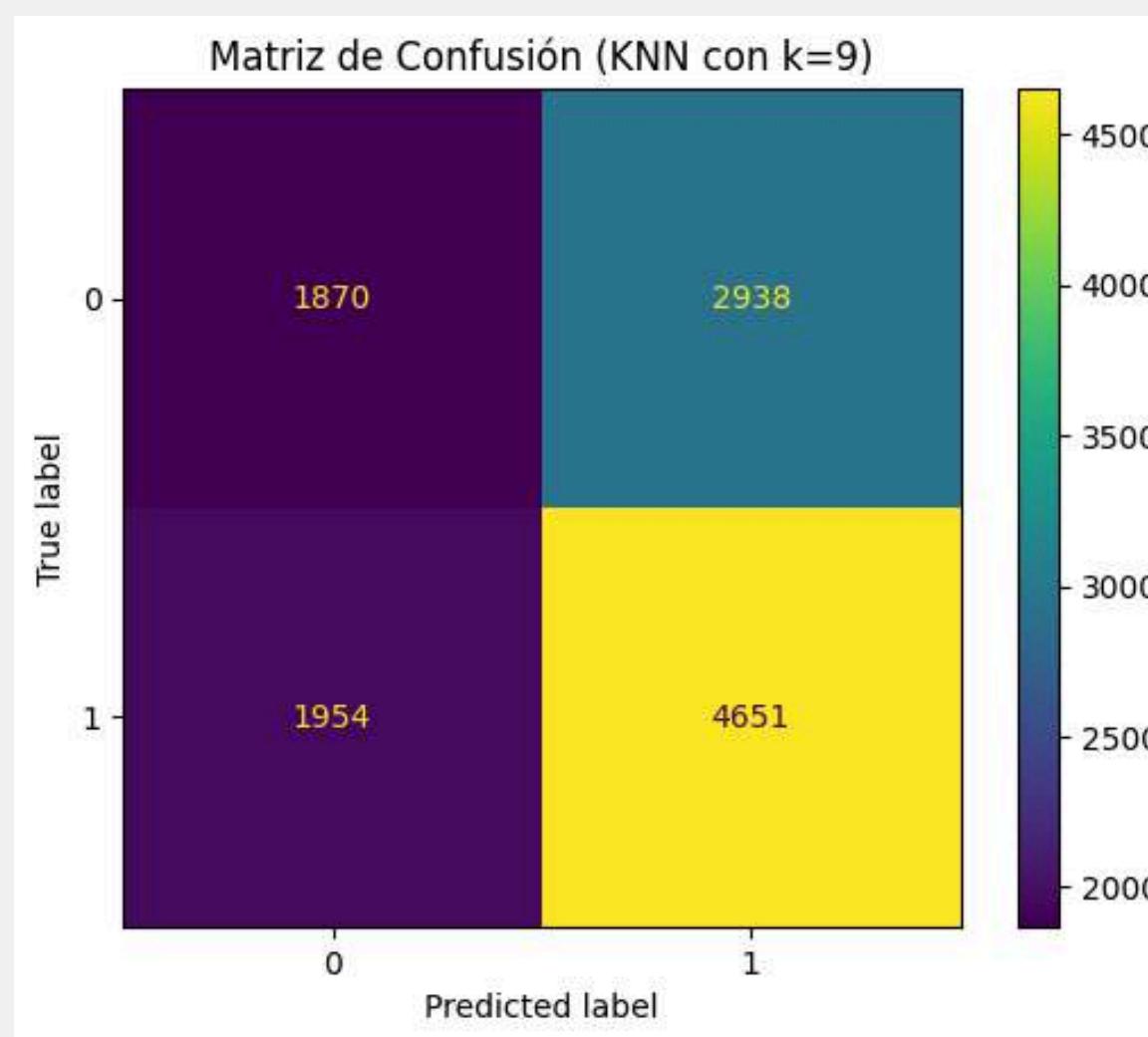
Pasos

- **1. Preparación de Datos:**
 - Eliminación de duplicados, corrección de errores de codificación y procesamiento textual.
 - Tokenización y eliminación de stopwords
 - Conversión de textos en representaciones numéricas mediante TF-IDF.
 - **2. División de Datos:**
 - Uso de train_test_split para crear conjuntos de entrenamiento y prueba.
 - **3. Implementación de Modelos:**
 - K-Nearest Neighbors (KNN)
 - Árboles de Decisión:
 - Naive Bayes
 - **4.. Evaluación y Optimización:**
 - Métricas: Exactitud, precisión, recall y F1-score.
 - Visualización: Matrices de confusión y curvas Precision-Recall.
 - Ajuste de Hiperparámetros: Validación cruzada y GridSearch para optimizar el rendimiento.
- 

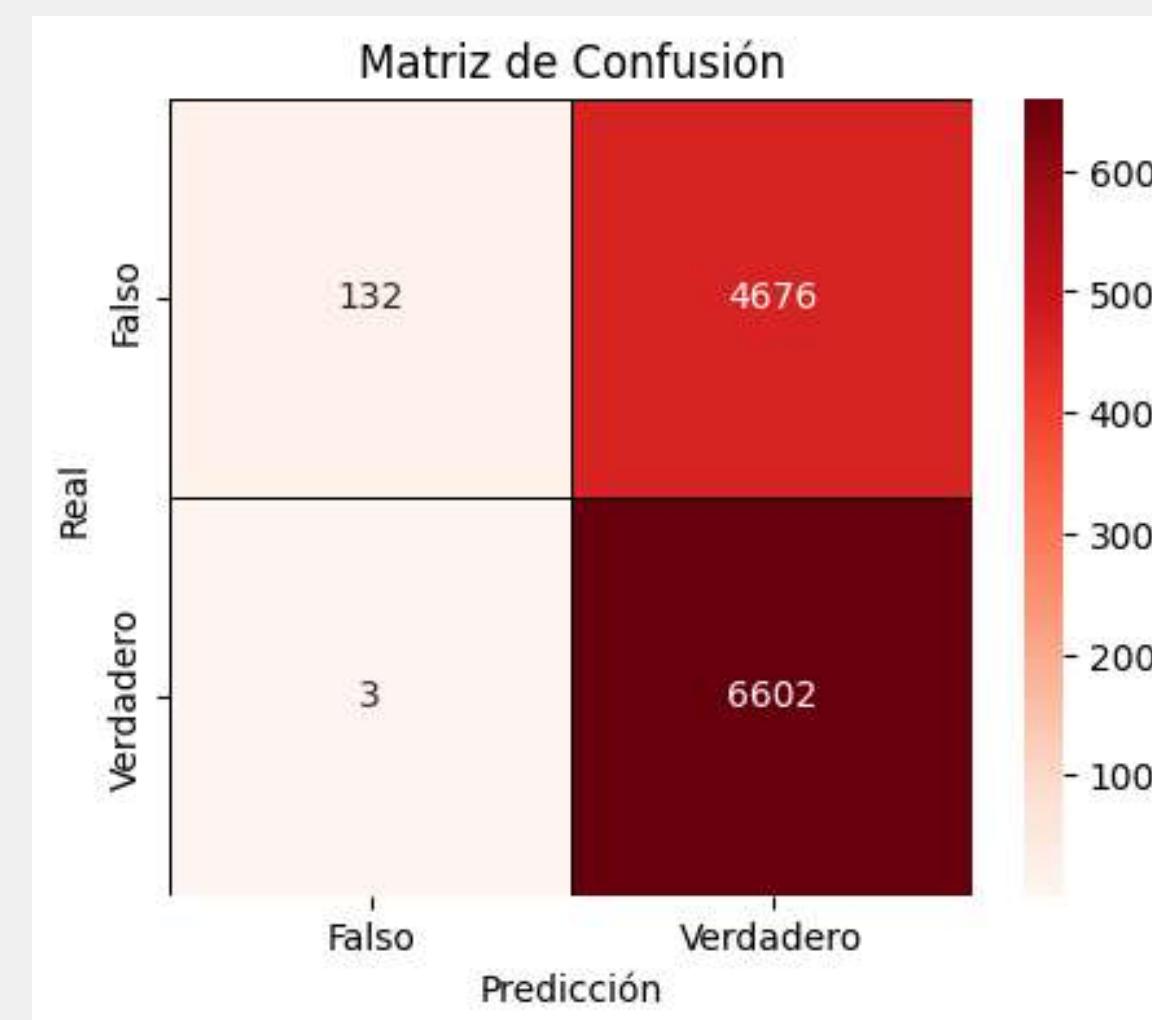
Resultados obtenidos

Mediante estas gráficas usando los tres algoritmos se apreciaron los siguientes resultados:

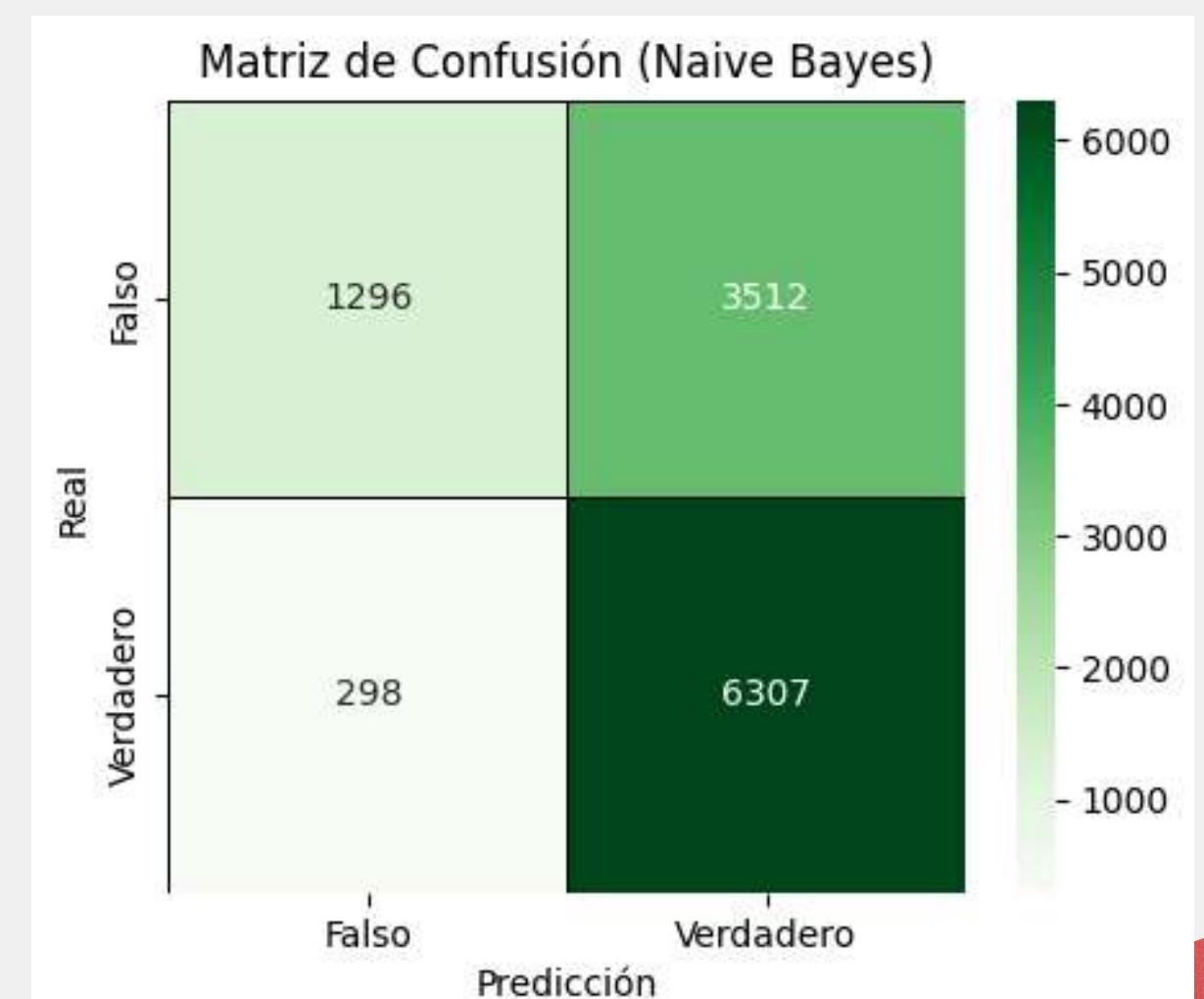
Algoritmo de KNN



Algoritmo de Árboles de decisión



Algoritmo de Naive bayes



Conclusión

- La selección del modelo óptimo dependerá de los objetivos del proyecto, considerando la importancia de minimizar ciertos tipos de error y la interpretabilidad del modelo.
- Sin embargo, con base a los resultados el mejor modelo es el de Árboles de decisión por su recall del casi 100%.
 - Esto debido a que queremos que cada noticia falsa sea identificada.



Gracias

GRUPO 22

