

DE CLICS A CIENCIA: PREDICTING ONLINE NEWS POPULARITY WITH MACHINE LEARNING

Equipo 10

Curso: Aprendizaje de Máquina (1INFo2)

Fecha: 17/06/2025

EL DESAFÍO: ¿QUÉ HACE POPULAR A UNA NOTICIA?

EL PROBLEMA

- Sobrecarga de contenido digital.
- ¿Cómo predecir qué artículos clasificador preciso destacarán?
- Misión: Predecir si un artículo superará los **1400** 'shares'.

NUESTROS OBJETIVOS

1. Desarrollar un modelo clasificador preciso.
2. Evaluar múltiples algoritmos avanzados.
3. Superar el benchmark académico ($AUC > 0.73$).
4. Identificar factores clave de popularidad.



NUESTRO ENFOQUE: UNA METODOLOGÍA ESTRUCTURADA

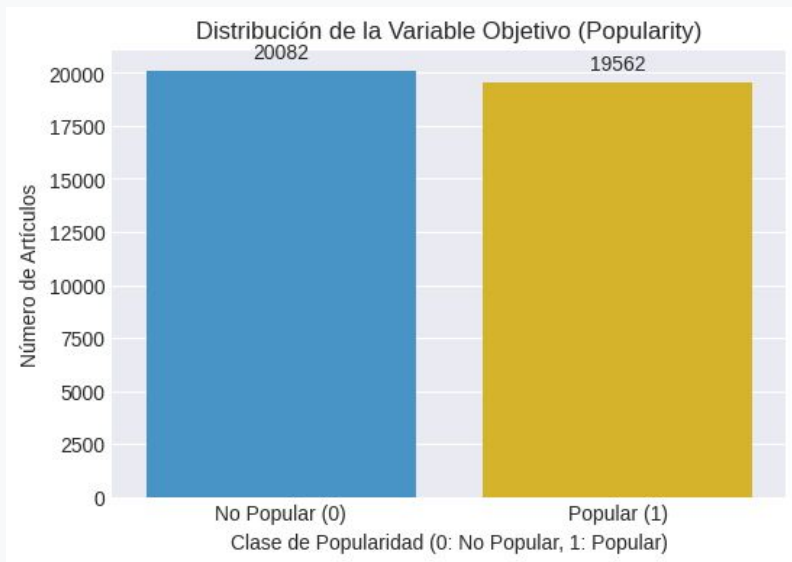
Adoptamos una metodología inspirada en CRISP-DM para asegurar un desarrollo riguroso, desde los datos hasta los insights.



EXPLORANDO LOS DATOS: PRIMEROS HALLAZGOS

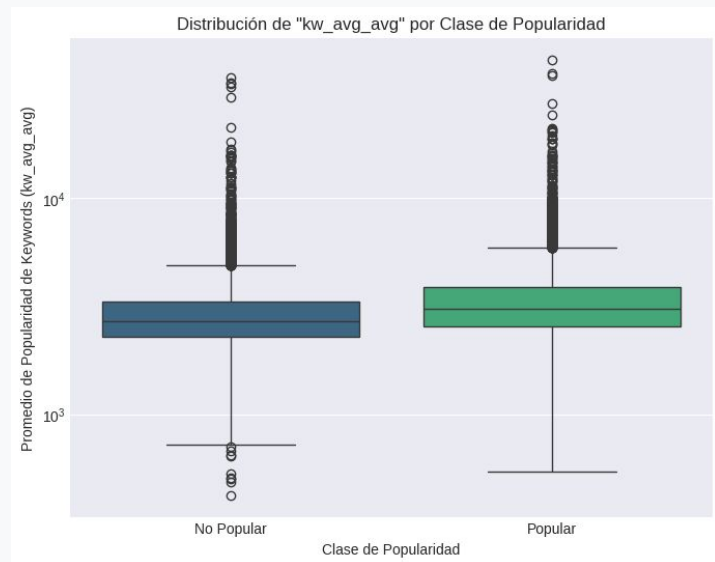
CLAVE

Balance de Clases



💡 **Insight:** El dataset está razonablemente balanceado (49% vs 51%).

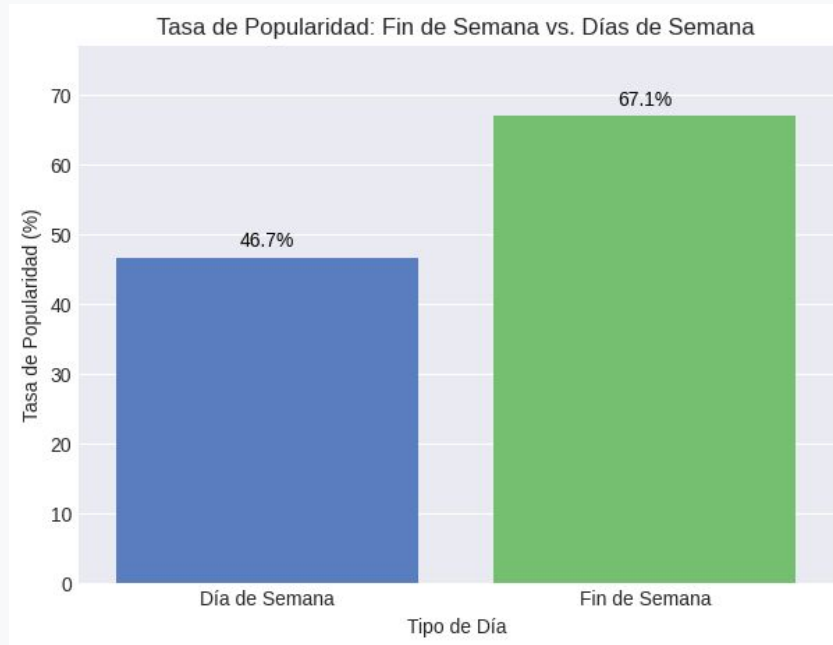
Impacto de Keywords




💡 **Insight:** Las noticias populares tienden a usar keywords que ya son, en promedio, más populares.

EXPLORANDO LOS DATOS: PRIMEROS HALLAZGOS CLAVE

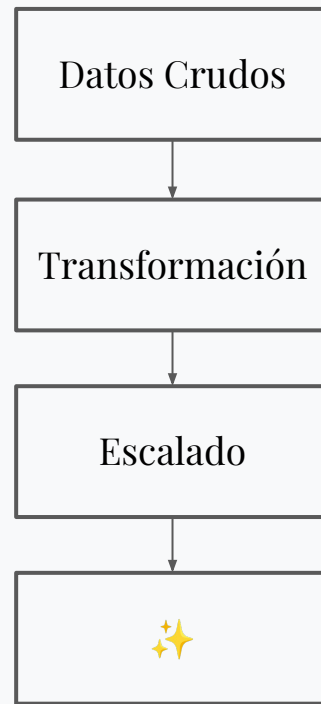
Influencia de la Temporalidad



 **Insight:** El tiempo es un factor crucial. Los artículos publicados durante el fin de semana tienen una mayor probabilidad de volverse populares.

DE DATOS CRUDOS A FEATURES LISTAS

- 1. Definición del Objetivo:**
 - Transformamos `shares` en un objetivo binario (`popularity`)
- 2. Transformación de Características:**
 - Aplicamos $\log(1+x)$ a variables asimétricas para estabilizarlas.
- 3. Estandarización:**
 - Escalamos todas las 58 características con `StandardScaler`.





NUESTRA CARTERA DE MODELOS: DE LA BASE AL SOTA



Regresión Logística (GPU): Nuestra línea base lineal, rápida e interpretable.



Random Forest (GPU): Un ensamble robusto para capturar interacciones no lineales.



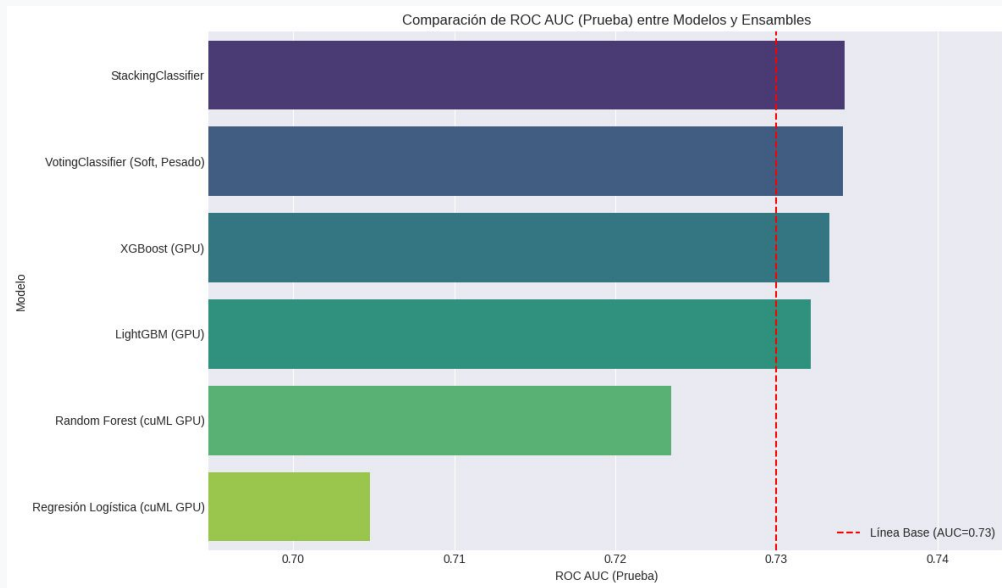
XGBoost & LightGBM (GPU): Modelos de Gradient Boosting de vanguardia, conocidos por su rendimiento superior.



Ensamblados de 2º Nivel (Stacking): Para combinar las fortalezas de todos los modelos y maximizar el rendimiento.



RESULTADOS: ¡MISIÓN CUMPLIDA! SUPERAMOS EL BENCHMARK



Modelo	AUC
Reg. Log.	0.7048
Rand. For.	0.7235
LGBM	0.7321
XGBoost	0.7333
Stacking	0.7342



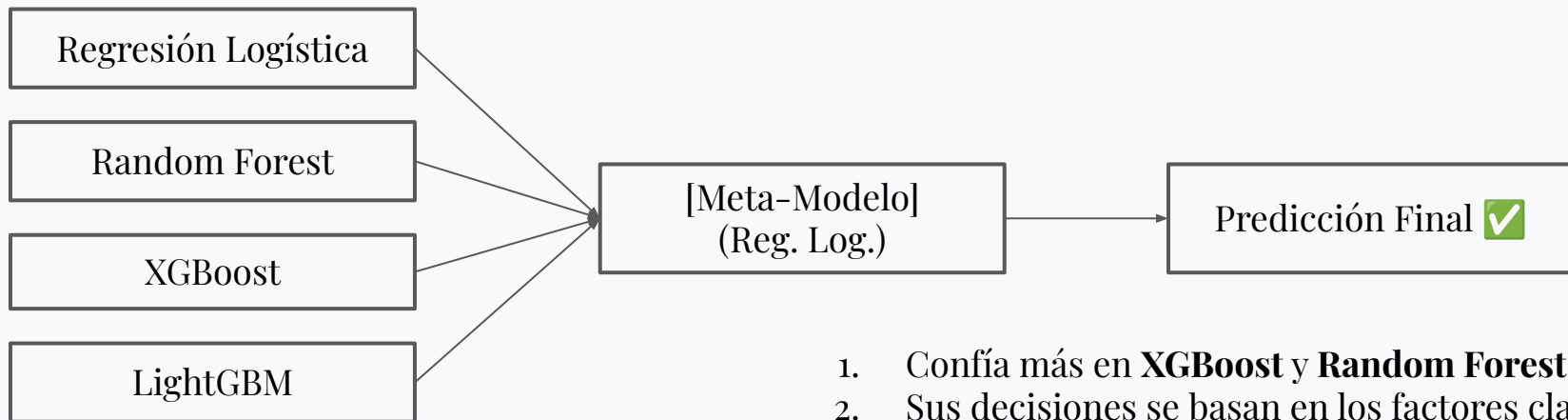
Insight:

- XGBoost, LightGBM y Ensamblas superaron la línea base de 0.73.
- Nuestro mejor modelo, el StackingClassifier, alcanzó 0.7342.



ANÁLISIS DEL CAMPEÓN: ¿CÓMO FUNCIONA EL STACKING?

El `StackingClassifier` aprende a combinar las predicciones de forma inteligente, actuando como un "gerente de expertos".



1. Confía más en **XGBoost** y **Random Forest**.
2. Sus decisiones se basan en los factores clave que estos modelos valoran: **Canal**, **Temporalidad** y **Keywords**.



DE INSIGHTS A IMPACTO: ESTRATEGIAS ACCIONABLES

1. **Amplificar el Contenido de Entretenimiento** 🧠📺
 - Asignar más recursos a la creación y promoción en el canal `Entertainment`, nuestro predictor más fuerte.
2. **Optimizar el Calendario para el Fin de Semana** 📅
 - Publicar o re-promocionar contenido de alto potencial los sábados y domingos para maximizar el engagement.
3. **Usar Inteligencia de Contenido y SEO** 🧠
 - Crear un sistema para priorizar el uso de keywords con un historial de alta popularidad (`kw_avg_avg`).



CONCLUSIONES DEL PROYECTO

1. **Objetivo Cumplido y Superado:**

- Desarrollamos un modelo (Stacking) que supera la línea base académica, alcanzando un **ROC AUC de 0.7342**.

2. **Factores Clave Descubiertos:**

- El **contexto de una noticia (su canal y temporalidad)** es más predictivo que las métricas simples de su contenido.

3. **Impacto Tecnológico Clave:**

- La **aceleración por GPU fue fundamental** para la viabilidad del proyecto, permitiendo la optimización de modelos avanzados en tiempos manejables.

GRACIAS
¿Preguntas? ?



Python

Pandas

NumPy

SKLearn

cuML

XGBoost

LightGBM