Predicción de la Popularidad de Noticias Online Mediante Ensambles de Modelos Acelerados por GPU

Ricardo Meléndez, Juan Galindo, Joseph Gutiérrez, Juan Coronado
Facultad de Ciencias e Ingeniería
Pontificia Universidad Católica del Perú
Lima, Perú
{20200485, 20181410, 20191826, 20191672}@pucp.edu.pe

***Resumen*—** La predicción de la viralidad del contenido digital es un desafío crítico para los medios de comunicación. Este estudio aborda el problema mediante el desarrollo de un pipeline de aprendizaje automático para clasificar la popularidad de noticias online, utilizando el dataset "Online News Popularity" de Mashable. Se evaluaron y optimizaron cuatro modelos base (Regresión Logística, Random Forest, XGBoost, LightGBM) y dos estrategias de ensamble de segundo nivel (Voting y Stacking), empleando aceleración por GPU (a través de RAPIDS cuML y las capacidades nativas de los modelos de boosting) para garantizar la viabilidad computacional. El modelo final, un StackingClassifier, alcanzó un ROC AUC de 0.7342 en el conjunto de prueba, superando el benchmark académico de referencia (~0.73 AUC). El análisis del modelo revela que el canal temático, la temporalidad (publicación en fin de semana) y la popularidad inherente de las palabras clave son los predictores más influyentes. Este trabajo demuestra la efectividad de los ensambles avanzados y la importancia de la computación acelerada para resolver problemas de clasificación en el mundo real.

Palabras Clave — *Aprendizaje de Máquina, Clasificación, Predicción de Popularidad, XGBoost, LightGBM, Stacking, Aceleración por GPU.*

I. INTRODUCCIÓN

En el ecosistema mediático digital contemporáneo, la capacidad de discernir qué contenido captará la atención de la audiencia es un diferenciador estratégico crucial. Este estudio aborda el problema de predecir la popularidad de noticias online, formalizado como una tarea de clasificación binaria. Utilizando el dataset "Online News Popularity" de UCI, el objetivo principal de este trabajo es desarrollar y evaluar un pipeline de modelado predictivo de extremo a extremo, con la meta específica de superar la línea base de rendimiento establecida en la literatura académica e identificar los factores más determinantes de la popularidad. Este documento presenta el estado del arte, el diseño experimental, los resultados, una discusión de los hallazgos y las conclusiones.

II. ESTADO DEL ARTE

La contextualización de este trabajo se basa principalmente en el estudio seminal de Fernandes et al., que introdujo el dataset y estableció el benchmark de rendimiento principal. Utilizando un modelo Random Forest, reportaron un Área Bajo la Curva ROC (ROC AUC) de aproximadamente 0.73, definiendo la popularidad con un umbral de 1400 'shares'. Trabajos posteriores, como el de Obiedat et al., reforzaron este benchmark, logrando

resultados similares. Adicionalmente, la literatura en el campo sugiere que para superar significativamente este umbral, es probable que se requieran técnicas avanzadas de ingeniería de características, como métricas de legibilidad o embeddings de texto basados en Deep Learning. Nuestro enfoque investiga si la aplicación de algoritmos de ensamble más potentes y una optimización exhaustiva, utilizando únicamente el conjunto de características original, pueden mejorar el benchmark establecido.

III. DISEÑO DEL EXPERIMENTO

A. Descripción del Conjunto de Datos

Se utilizó el dataset "Online News Popularity", que contiene 39,644 artículos de Mashable, cada uno descrito por 58 características predictivas. Estas características se agrupan en categorías como atributos de contenido (ej., número de tokens), estructurales (ej., número de imágenes), de canal (ej., data_channel_is_entertainment), de keywords, temporales, de tópicos LDA y de sentimiento. Se verificó que el dataset no contiene valores nulos. La variable objetivo binaria, popularity, fue construida a partir de la variable original shares, etiquetando un artículo como "popular" (clase 1) si shares > 1400. El análisis exploratorio inicial confirmó que las clases resultantes están razonablemente balanceadas (49.3% populares), y que varias características numéricas presentan una alta asimetría positiva, justificando el uso de transformaciones.

La Fig. 1 y la Fig. 2 ilustran dos de los hallazgos más significativos del análisis exploratorio. Como se observa en la Fig. 1, los artículos que alcanzan la popularidad presentan una distribución de kw_avg_avg (popularidad promedio de las keywords) con una mediana considerablemente más alta que los artículos no populares, sugiriendo que la relevancia previa del tema es un fuerte predictor. Adicionalmente, la Fig. 2 muestra que la tasa de popularidad es notablemente mayor para los artículos publicados durante el fin de semana en comparación con los días de semana, destacando la importancia del contexto temporal.

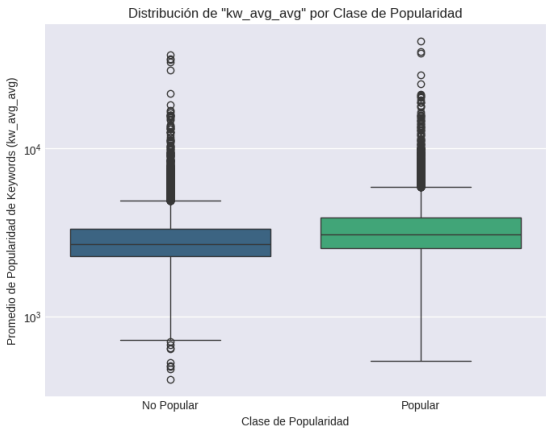

Fig. 1. Distribución de la popularidad promedio de las keywords (kw_avg_avg) por clase de popularidad, en escala logarítmica.

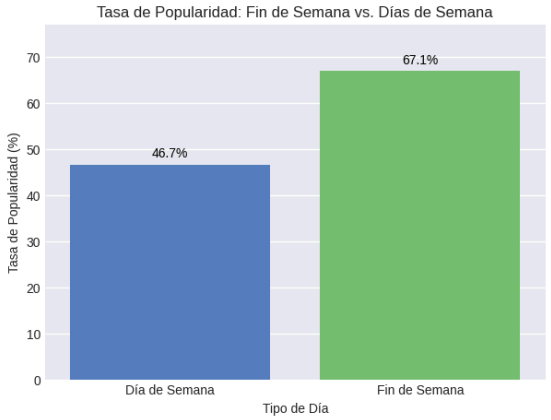

Fig. 2. Tasa de popularidad (%) de los artículos publicados en días de semana frente a los publicados en fin de semana.

B. Metodología

El proceso metodológico comprendió los siguientes pasos:

- **Preparación de Datos:** El dataset fue dividido en conjuntos de entrenamiento (80%) y prueba (20%) mediante una división estratificada para preservar la proporción de clases, utilizando una

semilla aleatoria (random_state=42) para garantizar la reproducibilidad.

- **Preprocesamiento:** Se definieron dos pipelines de preprocesamiento con ColumnTransformer: uno que aplica transformación logarítmica (np.log1p) a 14 características asimétricas y luego StandardScaler a todas (para Regresión Logística), y otro que aplica únicamente StandardScaler a todas las 58 características (para modelos basados en árboles).
- **Modelado:** Se evaluaron cuatro modelos base (Regresión Logística, Random Forest, XGBoost, LightGBM) y dos ensambles de segundo nivel (VotingClassifier y StackingClassifier).
- **Optimización:** Se empleó GridSearchCV con validación cruzada de 5 folds, utilizando roc_auc como la métrica de scoring.
- **Evaluación:** Las métricas finales incluyeron ROC AUC, Exactitud, Precisión, Recall, F1-Score y la Matriz de Confusión.
- **Entorno Computacional:** Los experimentos se ejecutaron en Google Colab con un acelerador GPU, utilizando Scikit-learn, RAPIDS cuML, XGBoost y LightGBM.

IV. EXPERIMENTACIÓN Y RESULTADOS

Los resultados de los cuatro modelos base optimizados y los dos ensambles de segundo nivel, evaluados en el conjunto de prueba, se resumen en la Tabla I y la Fig. 3. Los modelos de gradient boosting (XGBoost y LightGBM) y los ensambles de segundo nivel superaron consistentemente al modelo lineal y al Random Forest, además de exceder la línea base de referencia de 0.73 AUC. El VotingClassifier, que promedió las predicciones ponderadas de los modelos base, alcanzó un notable AUC de 0.7341. Sin embargo, el ensamble StackingClassifier, que entrenó un meta-estimador de Regresión Logística sobre las predicciones de los modelos base, logró el rendimiento más alto, aunque por un margen marginal.

TABLA I
COMPARACIÓN DE RENDIMIENTO DE MODELOS Y ENSAMBLES

```
--- Tabla Comparativa Final de Modelos y Ensambles ---
Tabla Comparativa de Rendimiento de Modelos y Ensambles
```

	Modelo	ROC AUC (CV)	ROC AUC (Prueba)	Accuracy (Prueba)	Tiempo Opt/Entr. (s)
0	Regresión Logística (cuML GPU)	0.7068	0.7047	0.6522	15.38
1	Random Forest (cuML GPU)	0.7272	0.7235	0.6643	172.11
2	XGBoost (GPU)	0.7336	0.7333	0.6721	408.28
3	LightGBM (GPU)	0.7325	0.7321	0.6728	78.76
4	VotingClassifier (Soft, Pesado)	nan	0.7341	0.6734	3.87
5	StackingClassifier	nan	0.7342	0.6732	39.28

```
<ipython-input-17-1177387153>:98: FutureWarning:
```

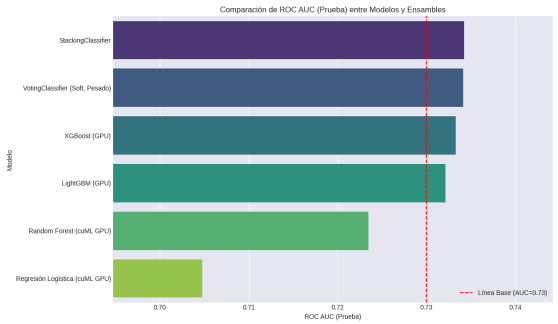

Fig. 3. Comparación de ROC AUC en el conjunto de prueba para todos los modelos y ensambles evaluados. La línea de base de referencia (0.73) se muestra como una línea discontinua.

Basado en su rendimiento superior en la métrica principal, el StackingClassifier fue seleccionado como el modelo final del proyecto, con un **ROC AUC de 0.7342** y una **Exactitud del 67.32%**.

V. DISCUSIÓN

El rendimiento superior del StackingClassifier se atribuye a su capacidad para aprender una combinación óptima de las predicciones de sus modelos base. El análisis de los coeficientes de su meta-estimador de Regresión Logística reveló una alta dependencia de las predicciones de XGBoost (coef. 1.99) y Random Forest (coef. 1.53), sugiriendo que el ensamble valora tanto el poder

predictivo del mejor modelo individual (XGBoost) como la diversidad aportada por el Random Forest. A su vez, el análisis de importancia de características del componente XGBoost indicó que los predictores más influyentes son el **contexto de la noticia**, específicamente el canal temático (data_channel_is_entertainment) y la temporalidad (is_weekend), por encima de las características intrínsecas del contenido como su longitud o número de imágenes.

El análisis de errores cualitativo del StackingClassifier mostró que sus fallos ocurren en los límites de la capacidad predictiva del dataset, donde las señales son ambiguas o son invalidadas por factores externos.

Los **Falsos Positivos** (artículos predichos como populares que no lo fueron) tienden a ser instancias que presentan múltiples características fuertemente asociadas con el éxito. Por ejemplo, una instancia clasificada erróneamente pertenecía al canal tech (un predictor importante), tenía un kw_avg_avg alto y un self_reference_avg_sharess extremadamente elevado. La decisión del modelo fue lógicamente consistente con los patrones aprendidos; su fracaso real para ganar tracción probablemente se debió a factores latentes no capturados en los datos, como una baja calidad de redacción o la competencia con un evento noticioso de mayor impacto ese día.

Inversamente, los **Falsos Negativos** (artículos que sí fueron populares a pesar de ser predichos como no populares) representan "éxitos inesperados" que carecían de los predictores típicos. Un ejemplo notable fue un artículo que, a pesar de ser del popular canal entertainment, no se publicó en fin de semana y tenía valores de auto-referencia modestos. Su éxito imprevisto pudo haber sido impulsado por factores externos, como la promoción por parte de una figura influyente en redes sociales o la conexión con una comunidad de nicho muy activa, dinámicas que el modelo no puede "ver". Estos casos subrayan la naturaleza estocástica de la viralidad.

Limitaciones del Estudio

Es importante reconocer las limitaciones inherentes a este estudio, que a su vez abren oportunidades para futuras investigaciones. Primero, el análisis se basa exclusivamente en las 58 características tabulares proporcionadas en el dataset, sin incorporar el contenido textual completo de los artículos. Esto impide el uso de técnicas de NLP avanzadas que podrían capturar matices semánticos o estilísticos. Segundo, la definición de popularidad se basa en un umbral binario fijo (1400 shares), lo cual es una simplificación de lo que en realidad es un espectro continuo de viralidad. Finalmente, el dataset representa una instantánea temporal de noticias publicadas entre 2013 y 2015; los patrones de consumo de medios y los factores que impulsan la popularidad en las plataformas digitales pueden haber evolucionado desde entonces.

Desde una perspectiva de negocio, estos hallazgos implican que las estrategias editoriales deberían enfocarse en:

- Potenciar la creación y promoción de contenido en los canales de mayor impacto, como el de entretenimiento.
- Optimizar el calendario de publicaciones para capitalizar el mayor engagement del fin de semana.
- Utilizar sistemas de inteligencia de contenido para priorizar keywords con un historial probado de alta popularidad.

VI. CONCLUSIONES Y TRABAJOS FUTUROS

Este estudio ha demostrado con éxito la aplicación de un pipeline de aprendizaje automático avanzado

para predecir la popularidad de noticias online. Las conclusiones principales son:

- Se desarrolló un modelo StackingClassifier que, con un **ROC AUC de 0.7342**, superó la línea base académica establecida.
- Se identificó que el **contexto de una noticia (su canal y temporalidad)** es un predictor más fuerte de su popularidad que las métricas simples de su contenido.
- La **aceleración por GPU fue un habilitador tecnológico fundamental** para la optimización exhaustiva de modelos complejos en tiempos viables.

Las líneas de trabajo futuro más prometedoras incluyen la **ingeniería de características avanzadas** mediante el procesamiento de lenguaje natural del contenido textual completo, la **optimización del meta-estimador** del ensamble Stacking, y la aplicación de **técnicas de interpretabilidad avanzadas** como SHAP para un análisis más granular a nivel de predicción individual.

REFERENCIAS

[1] K. Fernandes, P. Vinagre, and P. Cortez, "A Proactive Intelligent Decision Support System for Predicting the Popularity of Online News," in *Portuguese Conference on Artificial Intelligence*, 2015, pp. 535–546.

[2] R. Obiedat et al., "Predicting the Popularity of Online News Using Classification Methods with Feature Filtering Techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 583–591, 2020.

[3] P. Singh et al., "Online news popularity prediction before publication: effect of readability, emotion, psycholinguistics features," *Multimedia Tools and Applications*, vol. 81, pp. 34293–34320, 2022.

[4] U. Saeed et al., "An Automated System to Predict Popular Cybersecurity News Using Document Embeddings," *IEEE Access*, vol. 9, pp. 126685–126698, 2021.