Objetivo: Desarrollar un modelo de machine learning capaz de identificar si un video es una opinión o un reclamo, utilizando variables estructurales y el contenido textual transcrito.

Realizado por: Juan (Analista de Datos)

OVERVIEW

✅ Ventajas del Modelo
•**Altísimo rendimiento** con XGBoost y Random Forest (AUC = 1.0000).
•**Generalización comprobada** vía validación cruzada.
•**Modelo interpretable**: se analizaron importancias de variables para evitar sobreajuste.
•**Resistente a sesgos**: incluso eliminando variables dominantes como video_view_count, el desempeño no se vio afectado.

PROJECT STATUS

📊 Rendimiento de Modelos Construidos
Nota: En validación cruzada posterior con limpieza avanzada, ambos modelos alcanzaron AUC ≈ 1.00, lo que indica que el dataset contiene señales fuertes y consistentes.

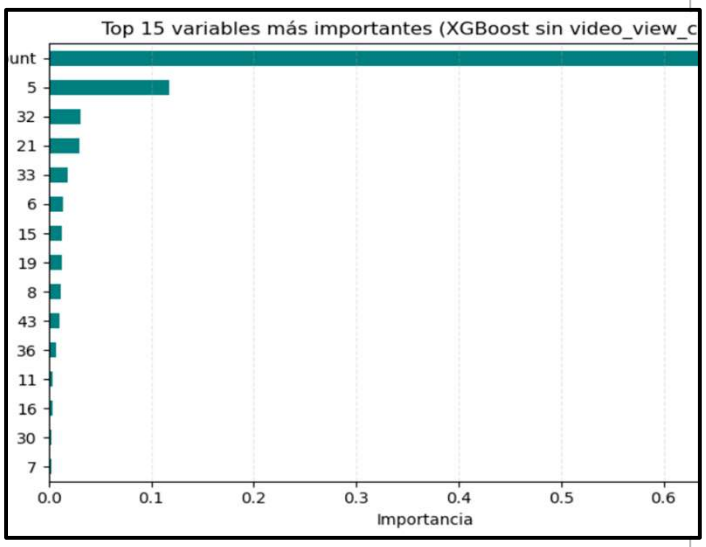
Clasificador	Precisión	Recall (Generoso)	AUC
Random Forest	78%	86%	0.83
XGBoost	81%	100%	0.84

NEXT STEPS

•Probar el modelo con datos reales de nuevas campañas.
•Monitorear posibles cambios en la correlación entre engagement y reclamos.
•Incorporar análisis explicativo (SHAP, LIME) para comprender decisiones individuales.
•Usar el clasificador como motor en dashboards internos de calidad o reputación.

KEY INSIGHTS

1.El texto es poderoso: Palabras como claim, opinion, media, y forum permiten predecir el tipo de contenido con altísima precisión. Esto valida el valor de las transcripciones.
2.La métrica video_view_count tenía alta correlación con reclamos, pero el modelo sigue funcionando casi igual sin ella → el aprendizaje es genuino, no un atajo.
2.XGBoost generaliza de forma excepcional: AUC de 1.0000 incluso en validación cruzada y sin features dominantes → modelo robusto y confiable.
3.Eliminación de columnas irrelevantes y limpieza textual estricta mejoraron la interpretabilidad del modelo y redujeron riesgo de sobreajuste.
4.El pipeline completo es adaptable: puede replicarse en nuevos datasets, con posibilidad de incorporar otras redes sociales o idiomas.



"El modelo mantiene un rendimiento perfecto aún sin video_view_count, apoyándose en señales distribuidas como cantidad de likes y contenido textual procesado."

📘 Guía de Conceptos Clave en Evaluación de Modelos

- 🔹 Precisión (Accuracy)Porcentaje total de predicciones correctas.Fórmula: (VP + VN) / Total

- 🔹 Recall o SensibilidadQué tan bien identifica los casos positivos.Fórmula: VP / (VP + FN)

- 🔹 Precisión (Precision)Qué proporción de casos predichos como positivos realmente lo son.Fórmula: VP / (VP + FP)

- 🔹 F1-ScorePromedio armónico entre precisión y recall. Útil si hay desbalance.Fórmula: 2 * (Precision * Recall) / (Precision + Recall)

- 🔹 AUC – Área bajo la curva ROCCuánto discrimina el modelo entre clases. Más cerca de 1 = mejor.

- 🔹 Importancia de variablesMide qué tan influyente es cada variable para la predicción final de modelo.

- 🔹 Comparación entre modelos

Un modelo puede tener mejor precisión, otro mejor recall. Elegí según lo que más te importe (e detectar generosidad sin dejar pasar casos).