1. Introducción

Frente al desafío de desarrollar un modelo para predecir la cantidad de *likes* de una publicación, el problema fue abordado desde dos enfoques complementarios, con objetivos analíticos distintos.

En un primer escenario, la predicción de *likes* se considera un fin en sí mismo. Es decir, la idea es que, a partir de ciertas características del post —como el tipo de publicación, su categoría, el día y horario de publicación o el alcance de la página— se pueda estimar cuántos *likes* podría obtener antes de ser publicado.

Bajo este enfoque, se excluyeron deliberadamente variables que introducen un claro problema de *data leakage*, como *comments*, *shares*, *interactions* y aquellas relacionadas con el comportamiento histórico de la audiencia (variables del tipo *lifetime*). Si bien estas variables son altamente predictivas dentro del dataset, no estarían disponibles en un entorno productivo al momento de realizar la predicción, lo que volvería al modelo poco realista desde una perspectiva operacional.

En un segundo escenario, el modelado de *likes* se utiliza como una herramienta para comprender qué factores influyen con mayor fuerza en este KPI y, en consecuencia, orientar decisiones estratégicas. Por ejemplo, un influencer que define su *fee* en función de los *likes* obtenidos en publicaciones patrocinadas podría beneficiarse de entender cómo variables como *comments* y *shares* contribuyen al desempeño del post. En este contexto, resulta razonable incorporar estas variables al modelo, no con fines estrictamente predictivos, sino para obtener *insights* accionables sobre qué tácticas priorizar.

---

2. Preprocesamiento de datos

El preprocesamiento se diseñó con el objetivo de garantizar consistencia entre variables, minimizar sesgos y asegurar que los datos fueran adecuados para el entrenamiento de modelos de *machine learning*.

Las variables numéricas fueron normalizadas mediante *RobustScaler*, con el fin de reducir el impacto de valores extremos y evitar que las diferencias de escala influyan en el entrenamiento del modelo. Si bien los algoritmos seleccionados son relativamente robustos a la escala, esta normalización aporta mayor estabilidad y comparabilidad entre *features*.

Para las variables categóricas se utilizó *OneHotEncoder* en aquellos casos de baja cardinalidad, priorizando la interpretabilidad del modelo. En particular, se codificaron variables como el día de la semana, la categoría y el tipo de publicación. La variable de hora fue

agrupada en intervalos de seis horas, con el objetivo de capturar patrones temporales relevantes sin introducir una granularidad excesiva.

En cuanto al manejo de valores faltantes, se eliminó un único registro con valor nulo en la variable objetivo (*likes*), dado que no aporta información útil en un contexto de aprendizaje supervisado. Además, se imputaron cuatro valores faltantes en la variable *share* utilizando información disponible de *interactions*, *likes* y *comments*. Finalmente, se imputó un valor faltante en la variable *paid* asignando la clase dominante, decisión respaldada por un análisis bivariado previo.

2.3 Selección de variables

Para el primer escenario se incluyeron únicamente variables que no estuvieran sujetas a *data leakage*. En este sentido, se excluyeron todas las variables asociadas al comportamiento histórico de la audiencia (*Lifetime*), así como aquellas directamente relacionadas con la interacción posterior a la publicación, como *comments*, *shares* y *Total Interactions*.

Para el segundo escenario, se decidió excluir la variable *Total Interactions*, dado que constituye una combinación lineal de *likes*, *comments* y *shares*, lo que podría introducir redundancia y problemas de multicolinealidad en el modelo.

2.4 Modelado

Se evaluaron distintos enfoques de modelado, incluyendo regresión lineal regularizada mediante Ridge, Random Forest, KNN y XGBoost. Para cada modelo se realizó una optimización bayesiana de hiperparámetros, utilizando validación cruzada de cinco *folds* y el coeficiente de determinación (R^2) como métrica objetivo.

Una vez seleccionados los mejores hiperparámetros, los modelos fueron entrenados sobre el conjunto de entrenamiento y evaluados sobre el conjunto de test, utilizando una partición 80/20. Adicionalmente, se decidió excluir las publicaciones virales (con más de 1000 *likes*) del entrenamiento, ya que estos casos extremos deterioraban el desempeño del modelo debido a su alta impredecibilidad y baja representatividad dentro del dataset.

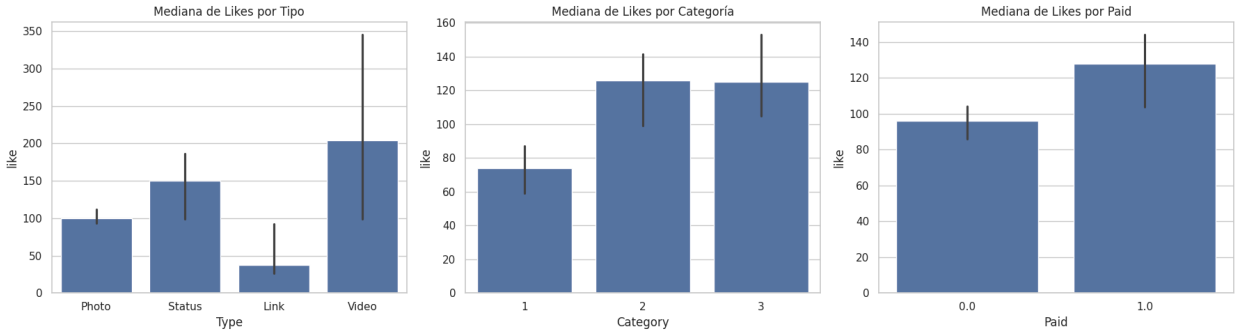
Finalmente, se optó por utilizar Random Forest como modelo final para ambos enfoques, dado que fue el algoritmo que mostró mejor desempeño de manera consistente a lo largo de las distintas evaluaciones.

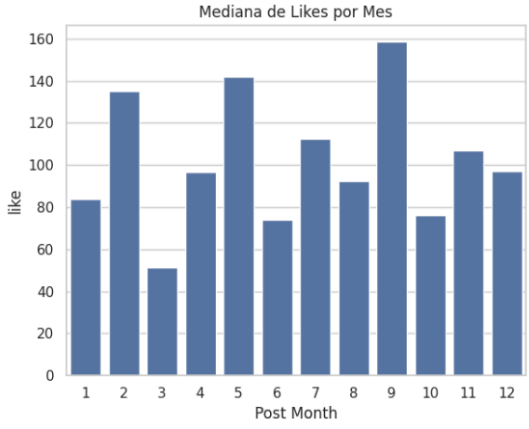
---

3. Conclusiones

El desarrollo de modelos bajo el primer escenario resulta altamente desafiante y no alcanza un desempeño que permita considerarlo un modelo predictivo eficaz. Esto se explica principalmente porque el dataset no contiene información ex ante lo suficientemente rica como para anticipar con precisión el éxito de una publicación.

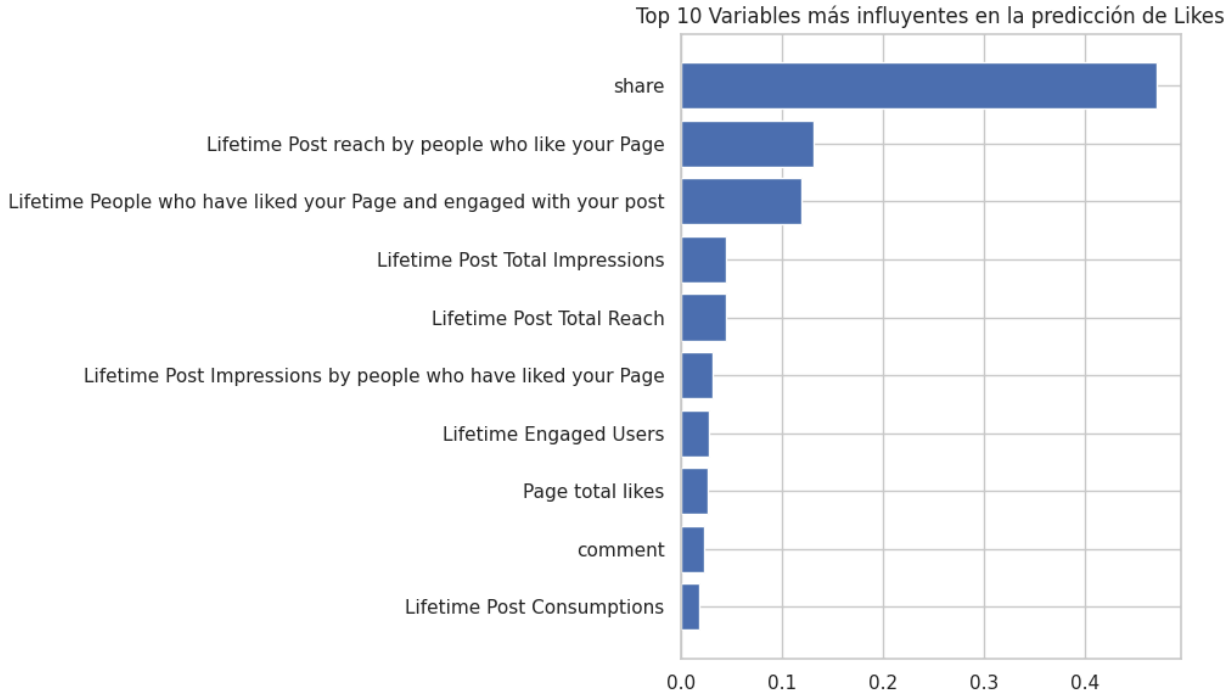
No obstante, el análisis descriptivo permite extraer algunos *insights* relevantes. En particular, se observa que las publicaciones de tipo *link* y aquellas pertenecientes a la categoría 1 tienden a obtener una menor cantidad de *likes*. Asimismo, se identifican patrones temporales que sugieren priorizar publicaciones en los meses de mayo y septiembre, y despriorizar febrero y octubre.





En contraste, para el segundo escenario se obtienen resultados predictivos considerablemente más sólidos. Al modelar todos los tipos de publicación, el modelo alcanza un R^2 cercano a 0.60. Sin embargo, al restringir el análisis únicamente a publicaciones de tipo *foto*, que constituyen la clase dominante, el desempeño mejora de manera significativa, alcanzando un R^2 cercano a 0.80, junto con una reducción sustancial del error absoluto.

Desde una perspectiva práctica, el análisis del segundo escenario indica que la variable con mayor poder predictivo es *share*. Esto sugiere que, para mejorar el desempeño en términos de *likes*, un influencer debería priorizar estrategias orientadas a fomentar la compartición de contenido, por encima de aumentar comentarios o depender exclusivamente del alcance potencial de su plataforma.



En conjunto, este trabajo muestra las limitaciones y oportunidades de modelar el engagement en redes sociales a partir de información disponible. Mientras que la predicción ex ante de *likes* presenta fuertes restricciones, el uso de modelos como herramienta analítica permite obtener conclusiones útiles para la toma de decisiones.