Reporte Entrega 1: EDA
Proyecto de Machine Learning con CRISP-DM

Kristian Restrepo Osorio - Sebastian Restrepo Ortiz

October 2, 2025

1 Introducción y planteamiento del problema

El mercado de aplicaciones móviles crece de forma acelerada, y la valoración de las apps en plataformas como Google Play Store se ha convertido en un indicador importante de éxito y visibilidad. Sin embargo, los factores que influyen en que una aplicación obtenga una calificación alta no son triviales, ya que intervienen variables como el número de descargas, reseñas de usuarios, categoría de la app, precio, entre otros.

El objetivo de este proyecto es **predecir la calificación de una aplicación** a partir de sus características disponibles en el dataset de Google Play Store y las derivadas/seleccionadas a partir de feature engineering [Lava18, 2018]. Este trabajo sigue la metodología CRISP-DM, comenzando en esta entrega con un análisis exploratorio de datos (EDA).

2 Trabajos relacionados

Diversos estudios han explorado la predicción de ratings y el éxito de aplicaciones móviles, aportando distintas perspectivas metodológicas:

- Saleem et al. [2024] proponen un enfoque basado en *ensemble learning* para predecir el éxito de aplicaciones en la Google Play Store. Usando atributos como rating, número de descargas, reseñas, tamaño y precio, además de aplicar técnicas de procesamiento de lenguaje natural (NLP) sobre las descripciones y reseñas de las apps, evaluaron modelos como Decision Tree, Random Forest, KNN, Gradient Boosting y LightGBM. Los resultados muestran que los métodos de ensamble, en particular el *hard voting ensemble*, alcanzan una precisión cercana al 97%, superando a los modelos individuales y mostrando mayor robustez frente a ruido y variabilidad en los datos.

- Mahmood et al. [2020] investigan la influencia de factores del dataset de Google Play (installs, reviews, tamaño de la app, tipo de app, categoría, content rating y variables derivadas como número de palabras o caracteres en el título) sobre el rating promedio. Utilizan Random Forest, regresión lineal y SVR para calcular la importancia de cada variable. Sus resultados muestran que **el número de reseñas, el tamaño de la app, el género y la longitud del título (en caracteres y palabras)** son los factores

con mayor impacto en la valoración de una aplicación, mientras que aspectos como incluir la palabra "free" o números en el título tienen poca relevancia. Estos hallazgos refuerzan la elección de dichas características en nuestro proyecto, pues demuestran su relevancia real en la predicción del rating.

- Aleem and Noor [2024] estudian la predicción del éxito de aplicaciones Android antes de su publicación en Google Play Store. Para ello, utilizan un conjunto de características iniciales disponibles en la etapa previa al lanzamiento, como categoría, tamaño de la app, tipo de contenido y metadatos asociados. Evalúan distintos modelos de aprendizaje automático, incluyendo regresión logística, Random Forest y máquinas de soporte vectorial (SVM), con el fin de anticipar tanto la valoración como el número de instalaciones de las aplicaciones. Sus resultados muestran que, aun sin información de reseñas o descargas, es posible obtener predicciones consistentes sobre el desempeño futuro de una app. Este trabajo aporta evidencia de que las características tempranas pueden ser suficientes para construir modelos baseline útiles y complementa directamente nuestro enfoque.

En conjunto, estos trabajos muestran que, aunque la predicción del rating es un problema con alta variabilidad y ruido, que no permite definir reglas estrictas para determinarlo, existen patrones generales detectables mediante machine learning.

3 Aprendizajes relevantes del análisis exploratorio

El análisis exploratorio de datos (EDA) reveló hallazgos fundamentales que guían las decisiones metodológicas del proyecto. Los principales aprendizajes se organizan en las siguientes categorías:

3.1 Calidad y estructura de los datos

- **Duplicados significativos**: Se identificaron 483 filas duplicadas (4.46% del dataset), lo que evidencia la necesidad de limpieza previa para evitar sesgos en el entrenamiento.
- **Valores faltantes estratégicos**: Las variables críticas como `Rating` (13.6% faltantes) y `Size` (15.6% faltantes) requieren estrategias diferenciadas de imputación por categorías, mientras que la variable objetivo (`Rating`) no debe imputarse para evitar sesgo en la predicción.
- **Valores inválidos**: La presencia de un `Rating = 19` confirma errores de captura que deben eliminarse, no corregirse, para mantener la integridad del rango [1,5].

3.2 Distribuciones y transformaciones necesarias

- **Asimetría extrema**: Variables como `Reviews` (media: 444,111 vs mediana: 2,094) e `Installs Numeric` muestran distribuciones con colas largas que requieren transformaciones logarítmicas ($\log_{1+p}$) para estabilizarlas.

- **Desbalance categórico**: El 93.1% de las apps son gratuitas (`Type = Free`) y el 79.2% tienen clasificación `Everyone`, lo que refleja la realidad del mercado pero plantea desafíos para el modelado.

- **Outliers legítimos**: El 32.2% de registros fueron marcados como outliers por el método IQR, pero representan apps exitosas con descargas masivas (no errores), por lo que se preservan tras transformaciones.

3.3 Relaciones entre variables y señal predictiva

- **Correlaciones débiles con el target**: La variable objetivo `Rating` muestra correlaciones prácticamente nulas con las demás variables numéricas (todas < 0.1), lo que sugiere relaciones no lineales y la necesidad de feature engineering avanzado.

- **Relación instalaciones-reseñas**: La correlación más fuerte se da entre `Installs Numeric` y `Reviews` (Spearman: 0.97), confirmando que mayor popularidad genera más interacciones, patrón útil para variables derivadas.

- **Impacto del precio**: Las correlaciones negativas entre `Price` y variables de popularidad (-0.17 con `Reviews`, -0.24 con `Installs`) validan la hipótesis de que las apps pagas tienen menor alcance.

3.4 Feature engineering efectivo

- **Variables derivadas exitosas**: Se crearon 24 nuevas variables, incluyendo `review_rate` (engagement), `popularity_score` (métrica compuesta), y `days_since_update` (recencia), que capturan patrones más complejos que las variables originales.

- **Binning estratégico**: La categorización de variables continuas (`Price_bin`, `Size_bin`, `Installs_bin`) facilita la interpretabilidad y manejo de distribuciones asimétricas.

- **Flags de calidad**: Los indicadores de valores faltantes (`size_missing`, `price_missing`) preservan información sobre patrones de ausencia que pueden ser predictivos.

3.5 Implicaciones para el modelado

- **Dataset final robusto**: Tras la limpieza parcial, se obtuvieron 8,892 registros limpios para una futura selección de variables significativas con 38 de las mismas disponibles.

- **Necesidad de modelos no lineales**: Las correlaciones débiles y distribuciones complejas sugieren que algoritmos como Random Forest o Gradient Boosting pueden capturar mejor los patrones que modelos lineales simples.

- **Validación por categorías**: El fuerte desbalance hacia categorías como `FAMILY` (19%) y `GAME` (12%) requiere estratificación en la división train/test para mantener representatividad.

Nota: Todos estos hallazgos se detallan exhaustivamente en el notebook "01_eda_baseline" incluyendo visualizaciones, estadísticas descriptivas completas, análisis de correlaciones multivariados, y la implementación paso a paso de cada transformación aplicada como avance parcial de la preparación de los datos. El notebook también incluye la metodología CRISP-DM aplicada, la justificación teórica de cada decisión de preprocesamiento, y código reproducible para la generación del dataset limpio parcial, con probables futuras mejoras. Además, tambien se destacan los data cards generados de los datasets para ver información más a detalle.

References

Muhammad Aleem and Noorhuzaimi Mohd Noor. Predicting android app success before google play store launch using machine learning. *Library Progress International*, 44(1): 75–88, 2024. URL `https://bpasjournals.com/library-science/index.php/journal/article/view/641`.

Lava18. Google play store apps dataset. `https://www.kaggle.com/datasets/lava18/google-play-store-apps`, 2018.

Ahsan Mahmood et al. Identifying the influence of various factor of apps on google play apps ratings. *Journal of Data, Information and Management*, 2020. doi: 10.1007/s42488-019-00015-w. URL `https://www.researchgate.net/publication/337547293_Identifying_the_influence_of_various_factor_of_apps_on_google_play_apps_ratings`.

Aqsa Saleem, Muhammad Aleem, Nayem Uddin Prince, Md Mehedi Hassan Melon, Salman Mohammad Abdullah, Shah Md. Wasif Faisal, and Mohd Abdullah Al Mamun. Predicting mobile app success using a robust hard voting ensemble learning approach. *Letters in High Energy Physics*, 2024, 2024. URL `https://www.lettersinhighenergyphysics.com/index.php/LHEP/article/view/865`. Regular Issue, published 2024-02-04.