Análisis de Datos de Google PlayStore

Mónica Gisselle Auler Simonis, Camila Montserrat Alderete González, Máxima Soledad Ayala, Luis Fernando Caballero Ramoa

Data Science - Universidad Nacional de Asunción

Abstract

El proyecto consistió en utilizar un dataset con diferentes características de las Aplicaciones Móviles (Apps) disponibles para smartphones Android, realizar una limpieza, análisis descriptivo de los datos y aplicación de modelos de Machine Learning para clasificación de las aplicaciones con respecto a la cantidad de descargas de las mismas.

Introducción y justificación

Hoy en día, con la importante adopción de los teléfonos inteligentes en prácticamente todas las esferas de la sociedad, y las diferentes oportunidades de negocio que posibilita, vemos que el desarrollo de aplicaciones móviles (Apps) va más allá de simplemente elegir un lenguaje de programación o framework y empezar a implementarlas; sino que desde su misma concepción, como producto que busca ofrecer una solución a alguna situación. Es por esto que debemos aprovechar las herramientas disponibles a modo de clasificar las aplicaciones con respecto a una cantidad aceptable de descargas.

Objetivos

- Desarrollar habilidades para obtener, tratar e interpretar grandes volúmenes de datos estructurados.
- Realizar un análisis descriptivo integral de los datos de las aplicaciones disponibles en la Google Play Store de dispositivos Android.
- Determinar e implementar modelos de clasificación de machine learning.

Metodología

Primeramente se realizó una limpieza de los datos del correspondiente dataset. Luego se procedió al análisis descriptivo del conjunto de datos, donde revisamos algunas propiedades generales del dataset, sintetizando características, revelando patrones, y realizando un análisis estadístico descriptivo con gráficos relevantes. En base al análisis descriptivo se tomaron como variables independientes las columnas Category, Content Rating, Free, In App Purchases, Ad Supported y Minimum Android. Como variable dependiente se asignaron valores binarios que indican si una app superó las mil descargas,lo cual a nuestro criterio indica si una app es exitosa. Luego se procedió a particionar los datos en 80% para entrenamiento del modelo y 20% para test. Se realizó la implementación de modelos de machine learning para clasificar las aplicaciones. Los modelos implementados fueron los algoritmos de Decision Tree, Random Forest, Gradient Boosting, Naive Bayes y Voting Classifier. Voting Classifier fue entrenado en base a Random Forest, Decision Tree y Gradient Boosting. Una vez entrenados los modelos se procedió a validar los modelos, empleando métricas como accuracy, precision, recall y matriz de confusión.

Resultados

Modelo	Accuracy (%)	Precisión (%)	Recall (%)
Decision Tree	65.9239	64.1756	58.8078
Random Forest	65.9416	64.1500	58.9781
Gradient Boosting	64.4739	61.6641	60.3479
Gaussian Naive Bayes	56.5890	69.1386	10.3005
Voting Classifier	65.9458	64.1792	58.9039

Figure 1:Resultados según metricas de evaluacion de los modelos

Análisis Descriptivo

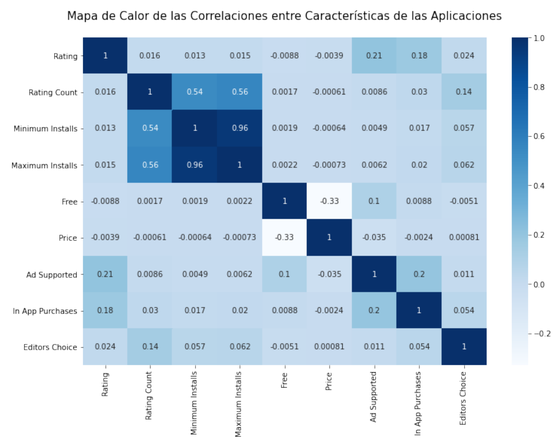

Figure 2:Mapa de calor

Análisis Predictivo

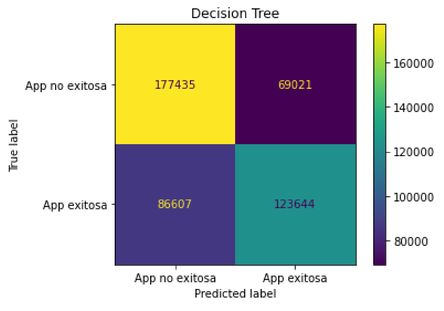

Figure 3:Decision Tree

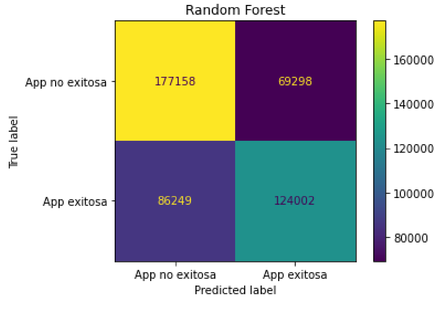

Figure 4:Random Forest

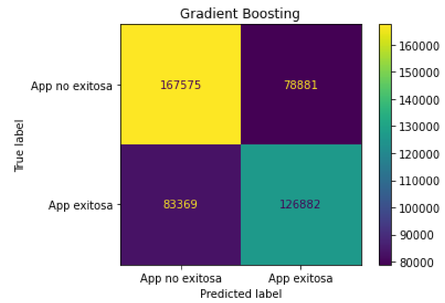

Figure 5:Gradient Boosting

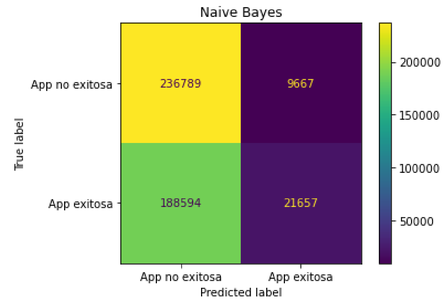

Figure 6:Naive Bayes

Conclusión

Finalmente, observamos que cada modelo implementado asignó una cierta valoración a una característica particular del dataset para la clasificación de las Apps. Al lograr entrenar los modelos, se obtuvieron los valores de las métricas de evaluación que dieron un resultado general por un rango del 60 % aproximadamente, por lo que el modelo con el mejor resultado de evaluación fue el de Random Forest.

Referencias

- Web: https://n9.cl/cy647
- Web: https://n9.cl/okuc9
- Web: https://n9.cl/6npmr
- Web: https://n9.cl/91rxe