GOOGLE PLAY STORE RATING PREDICTION UTILIZING RANDOM FOREST AND K-NEAREST NEIGHBORS REGRESSION WITH FEATURE SENSITIVITY ANALYSIS AND K-MEANS CLUSTERING FEATURE ANALYSIS

Composed by: Zhafira Elham Fawnia

ABSTRACT

Within this age of digitalization, the competition to predict successful attributes of an application is getting more and more versatile. The Google Play Store is amongst the top sources for consumers to obtain mobile applications and games specifically. Therefore, it raises the question of whether it is possible to predict the rating of games based on their market features, what kind of clusters exist within the data and what could contribute to their success.

Continuous variable prediction and clustering are under the umbrella term of predictive analytics. A regression method is commonly used for continuous variable prediction. Hence, a machine learning model often used for both prediction and classification problems that fits both discrete and continuous variables is Random Forest; as it builds a series of decision trees with a decided variation value, with a combination of bagging and random forest selections. To compare the results of the Random Forest regression, a K-Nearest Neighbors regression model is applied to the same problem. The results show Random Forest (95.19%) performs better than K-Nearest Neighbors (94.3%) when predicting continuous variables for mobile games rating.

Clustering analysis of each feature is performed to understand the importances, and success factors of the applications within the Google Play Store. The clustering approach is using the unsupervised K-Means Clustering; an algorithm that clusters data into multiple clusters to see its likelihood of one another. The results show that there are three distinct clusters within the Google Play Store dataset.

Keyword: Regression, Random Forest, K-Nearest Neighbors, Clustering, K-Means, Continuous Variable Prediction, Market Feature Analysis