Big Data Course Project
Proposal

Team #18

Team members

Name	Sec	BN
Marim Naser	2	23
Abeer Hussein	2	1
Mariem Muhammed	2	22
Ayman Mohamed	1	19

Problem description

- If a company wants to develop a new app, What's the best way to develop it to keep it highly profitable and highly rated? (**classification**).
- In addition to **predicting** the best price for this app -if it's paid- and predicting the number of installations for this app based on its given features.
- Lastly, if this company wants to hire new mobile app developers, we can help it to know those whose apps have the highest ratings and number of installations (**clustering**).

Dataset Source

This dataset was scraped via a python script running on a cloud. (we didn't scrape it, rather, we downloaded it).

- https://www.kaggle.com/datasets/gauthamp10/google-playstore-apps

Proposed Pipeline

1. **Data Preprocessing and cleansing**
2. **Data Exploration** (Involves visualization to extract knowledge from the data):
    I. **Descriptive analysis**: using Map Reduce.
    II. **Diagnostic analysis**: Using Pearson and Spearman's correlation.
    III. **Clustering** to gain insights about data: Using K-means, K-Medoids or ISODATA.
3. **Model training** and validation For Prediction and Classification: Using SVM, LR or Decision Trees, plus K-Fold.

Proposed Framework

- ***Spark***