Big Data Course Project Proposal

Team members

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

Idea

Developing a new profitable android app

Due to the vast usage of android phones we all use everyday, there emerged some patterns of android apps that make users hate the app instantly and decide to uninstall it, or like it and perhaps get addicted to it.
you can notice it yourself, you install a new app, an ad appears right in front of your touch screen, you become frustrated, then the app asks for micro-transactions, you will instantly hate it then uninstall.

Problem

Choosing the best price that maximizes number of installations plus high ratings

you are a manager in a software company and want to develop a profitable app, you want to pick the maximum price that guarantees very high ratings and large number of installations,
it will be a combination of choosing the idea of the app, its category, its target audience, the devs who will implement, presence and patterns of ads, etc.

Business Problems:
1-company wants to choose the most profitable and highly rated app category (category in our case carries some features from the app idea). classification
2- company wants to predict the best price for this app -if it' is paid- given its features (category, developer, ads, etc.). prediction
3-company wants to predict the number of installations for this app based on its given features (category, developer, ads, size, estimated price, etc.). prediction
4- company wants to hire new mobile app developers, and wants 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 Plan

Pipeline:

1- Data Preprocessing and cleansing

- might merge our dataset with another one, joining is done on app_id.
- remove useless columns, detect outlier ratings, interpolate missing data.

2- Data Exploration

Involves visualization to extract knowledge from the data:

I-Descriptive analysis:

- average number of installs per free/paid app (map red)
- number of installs for each category (map red)
- average and max price for each category (map red)
- average app size (in mb) for each category (map red)
- number of developed apps per developer (map red)

II- Diagnostic analysis:

- correlation between user rating and app price. (Pearson's correlation, it is better for intervals)
- Correlation between user rating and app price.
- correlation between user installations and app category and size. (spearman's correlation)

III- Clustering:

- cluster apps which gets high ratings, with medium to high price and look at (and maybe store) the developer's id and profile -which is in our case, number of developments-
- Cluster apps by ads and ratings
    - expectation: apps with more ads have medium to low ratings
- Cluster apps which have high ratings with low to medium price.
- We may use k-mean or k-medoid for clustering based on the results of data preprocessing .

3- Model training and validation:

- For prediction and classification problems we stated above we may use the following models: SVM, LR, Decision Trees.
- we may use k-fold cross validation to make the best use of our data.

need to write names of models for classification and prediction + i think the clustering part above needs to be here? and write name of the clustering technique