Project Proposal

**Data set**

Google Play Store Apps Open Data

* Google PlayStore Android App Data. (2.3 Million+ App Data)  
  Backup repo: <https://github.com/gauthamp10/Google-Playstore-Dataset>
* Reference: <https://www.kaggle.com/gauthamp10/google-playstore-apps>
* This Data set contains 2.3 million data of apps on Google Play Store. Several information will be implemented.

**Data Dictionary**

|  |  |
| --- | --- |
| App Name | The name of the app |
| App ID | Unique ID of the app on Google Play Store |
| Category | The category of the app on Google Play Store |
| Rating | Float, a total of 5. |
| Rating Count | The number of users who rated this app |
| Installs | The number of installations, only show big number with + |
| Minimum Install | The minimum of the installation in a period |
| Maximum Install | The maximum of the installation in a period |
| Free | Boolean, the app is free to download or not |
| price | The price of the app for downloading |
| currency | The currency of the app in its area |
| size | The Size of the app |
| Minimum Android | The lowest require Android version |
| Developer ID | Developer’s unique ID on Google Play Store |
| Developer Website | Developer’s personal website |
| Developer Email | Developer’s personal email |
| released | Released data of the app |
| Last updated | The date when the app was updated last time |
| Content rating | The rating of the content |
| Privacy Policy | Own policy of the app |
| Ad supported | Boolean, if the ads was supported in the app |
| In app purchase | Boolean, if there is in app purchase |
| Editor’s Choice | Unique choice of the editors |
| Scraped Time | Timestamp of the scraped time of the app. |

**Research**

There are four research goals to chase.

Part 1 is figure out that if there are linear relationship between the number of Installations and the rating by simple linear regression. In addition, multi-linear relationship is expected between the number of Installations and rating, rating count, price, and the size.

Part 2. Applying big data analysis classification and models for the name of the app to find out what kind of apps are installed the most and evaluate the models by confusion matrix.

This project will use all pyspark code to do the analysis. Data will be classified in RDD.

Top 20,000 app sorted by the number of installation will be used for classification and learning models.

**Learning model**

The learning models would be used in this project are: linear model, logistic regression, SVM

**Objective**

The objective of this project is successfully applied classifications and learning models to specific columns by pyspark code and evaluate them.

**Expectation**

1. There is strong simple linear relationship between the number of installation and the rating of the app.
2. Find out the multi-linear regression parameters between the number of installation and rating, rating count, price, and size. Expected to see the most important elements and the least important element.
3. Applied logistic regression to the name of apps to find out the most popular type of apps.
4. Evaluate the logistic regression model and get the accuracy over 70%, F1 score over 0.7 as well.
5. Applied SVM to the name of apps and evaluate the model with expected 0.7 accuracy and 0.7 F1 score.

**Model evaluation**

The method to evaluate the logistic regression and SVM is that random select 10000 apps from the data set and save as the test file. Building functions to find out the confusion matrix of the result. Apply F1 = 2 \* TP / (2 \* TP + FN + FP) to find out the F1 score and compute the accuracy. These two value will be used to evaluate the model and expect to have more than 0.7.