**CAPSTONE PROJECT**

**Project name :- Android google play store applications analysis**

Name :-Akashdeep Singh

Student ID: - 0771949

Name of Supervisor:- Dr. Savita Seharawat

Date of Submission:- 31st January

Logo, company name

Description automatically generated with medium confidence

Table of Contents

[Abstract 3](#_Toc452457337)

[Keywords 3](#_Toc452457338)

[Research Questions 3](#_Toc452457338)

[Tools 3](#_Toc452457338)

[Github Source 3](#_Toc452457338)

[Introduction: 3](#_Toc452457338)

[Literature Review 3](#_Toc452457338)

[Methodology: 3](#_Toc452457338)

[Data Details 3](#_Toc452457338)

[Table 1 3](#_Toc452457338)

[Heading 3 3](#_Toc452457339)

[Sample Chart 3](#_Toc452457340)

[Sample Table 3](#_Toc452457341)

# Abstract

Our project name is Android Google Play, moreover known as the Google Play Store, and in which I used androiddataset.csv dataset in which 1048575 records and 24 attributes. With the help of android google play store we will download or purchase millions of apps, diversions, and other media onto your Android gadget. You'll be able discover programs for a wide cluster of interface. You'll be able download apps or recreations by exploring to the app or amusement page inside the Play Store and tapping Introduce. Numerous apps will be free, a few will have in-app promotions, a few will fetched cash, whereas others may offer in-app buys, or a combination of any of these things. For those who are interested, there's the Google Play Pass, which allows you to download hundreds of apps and recreations for complimentary, without advertisements or in-app buys.so in this data we have some important attributes and value that help us to understand all functions of this dataset. Its a stage. It’s Google’s stage for advertising different advanced substance to its customers. Opposite to what a few individuals may think, the Google Play Store isn't fair an app store, not at all. You'll discover all sorts of substance accessible here. The Google Play Store is domestic to music, motion pictures, books, and recreations

**KEYWORDS:** Data cleaning, data visualization, Logistic regression model, KNeighbors, Random Forest Classifier, Naïve bayes model , Sampling , K-Fold Cross Validation , Stratified K-Fold , Confusion Matrix .

**RESEARCH QUESTIONS:**

**Question 1)-** find how many applications are free ,ad supported ,in purchases using pie chart?

ANSWER: 99.4 % applications are free of cost for the customers , also same percentage is noticed in applications that supported ads that is 99.4% and 0.6% applications are paid and also not supported ads.

**Question – 2) Which category of content rating show highest numbers that shows rating corresponding to application.**

**ANSWER: From all categories of content rating I find that people those are 18+ , they give more rating to the applications.**

**Question 3)** find accuracy of this project(dataset) by using different methods and also compare each other model to check which one is best as compare to the others?

ANSWER: I used four models to find accuracy of this project That is Logistic regression , KNN model ,Naïve bayes model and random forest classifier model .

Logistic regression accuracy is 59.33

KNN model accuracy is 54.91

Naïve bayes model accuracy is 59.09

Random forest classifier accuracy is 56.88

**So accuracy that we find with the help of logistic regression is best for our model that is 59.33**

**TOOLS:**

We will use python software for performing operations and to visualize our dataset.

**GITHUB INFO:**

**https://github.com/Akashdeepsingh1949**

**LITERATURE REVIEW:**

From the research I found that **Mr. Lavanya** define dataset as a name of google play store apps (**in 2019)** in which he was define lots of functions on the given dataset.in which he explains While numerous public datasets (on Kaggle and so forth) give Apple App Store information, there are very few partner datasets accessible for Google Play Store applications anyplace on the web. On burrowing further, I discovered that iTunes App Store page sends a well listed reference section like construction to consider basic and simple web scratching. Then again, Google Play Store utilizes complex advanced procedures (like unique page load) utilizing jQuery making scratching seriously testing. Along with this google play store apps data cleaning dataset by **Saba Siddiqi** in which they explain about data cleaning **(in 2018).**they were explaining about questions and answer about given dataset that is also related to google play store apps .**In 2019 ,n1kshan** publish dataset related to playstore in which he define seaborn functions and also explain how show information via graphs , The google play store is one of the largest and most popular Android app stores. It has an enormous amount of data that can be used to make an model. We have used a raw data set of Google Play Store from the Kaggle website. This data set contains 13 different features that can be used for predicting whether an app will be successful or not using different features. This data set is scraped from the Google Play Store. This journal talks about different classifier models that we used for prediction purposes and finding which one gives the highest accuracy. This journal also gives detailed information on feature extraction and the complete Data visualization done on this data set. [1] Kaggle.com. (2018). Google Play Store Apps.[online]https://www.kaggle.com/lava18/google-play-store-apps [Accessed 3 Mar. 2020].[2] “Mining and Analysis of Apps in Google Play,” Pro-ceding’s of the 9th International Conference on Web Information Systems and Technologies, 2013.[3] Google play store: number of apps2018(2018). [online] https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/ [Accessed 3 Mar. 2020].[4] Amit Chile, Dr. P. R. Gundalow.(2019). Anal-lysis of Google Play Store Application.[online]http://ijraset.com/ﬁleserve.php?FID=24134 [Accessed3 Mar. 2020][5] T. Doneux, M. Banger, Induction of decision trees from partially classiﬁed data, in: Proceedings of the 2000 IEEE International Conference on Systems, Man and Cybernetics (SMC’00), IEEE, Nashville, TN,2000, pp. 2923–2928.[6] Harman, M., Jia, Y., and Zhang, Y. (2012). App store mining and analysis: Mrs. for app stores. In 2012 9thIEEE Working Conference on Mining Software Repos-ivories (MSR),pages 108–111.[7] R. P. Rajeswari, K. Juliet, and Aradhana, “Text Classification for Student Data Set using Naive Bayes Classiﬁer and KNN Classiﬁer,” Int. J. Com-put. Trends Technol., vol. 43, no. 1, pp. 8–12, 2017.https://doi.org/10.14445/22312803/ijctt-v43p103[8] Jong, J. (2011). Predicting rating with sentiment anal-lysis. [online] http://cs229.stanford. Edu/proj2011/Jong-PredictingRatingwithSentimentAnalysis.pdf.[9] [2015].Grover, S. 3 apps that failed (and what they teach us about app marketing). [online]https://blog.placeit.net/apps-fail-teach-us-app-marketing

## Methodology:

Firstly, our dataset is uncleaned like there are some missing values . First, we will clean our dataset to perform father operations. In methodology part we will use regression model to find out relationship between dependent and independent variables.

**DATA PROCESSING**

**EXPLORATORY DATA ANALYSIS(EDA)**

**EXPERIMENTAL DESIGN**

**MODELING IMPLEMENTATION AND VALUES**

**CONCLUSION**

**DETAIL DATA :-DATA DICTIONARY HERE :-**

|  |  |  |  |
| --- | --- | --- | --- |
| **Attributes name** | **description** | **datatype** | **Count**  (Label name=count value) |
| App name | Define name of each application like call of duty.pub-g etc | object | Flashlight = 23  Calculator = 17  Gallery = 11 |
| App id | It define ID of the application that is unique for every application | object | com. ishakwe.gakondo = 1  petrosidagresik.sidacare =1  com.appsislamik.quranne =1 |
| category | It shows categories of application that stored in this dataset. | object | Education = 14751  Music & Audio = 9801  Tools = 9012  Business = 8963 |
| Rating | It shows rating of each application like 4.4, 7,9 etc. | Float64 | 0.0 = 65845  5.0 = 6146  4.2 = 5469  4.4 = 5375 |
| Rating count | It defines how many times user give ratings to each application. | Float64 | 0.0 = 65845  5.0 = 3991  6.0 = 3328  7.0 = 2940 |
| installs | How many times people install each application | object | 100+ = 27559  1,000+ = 24737  10+ = 18695  10,000+ = 16048 |
| Minimum installs | Minimum no of install apps | Float64 | 1.000000e+02 = 27559  1.000000e+03 = 24737  1.000000e+01 = 18695  1.000000e+04 = 16048 |
| Maximum installs | Shows no of apps that maximum install. | Float64 | 5.0 = 1062  2.0 = 1040  3.0 = 1037  4.0 = 1023 |
| Free | It shows application is free or not. | category | True = 140745  False = 2863 |
| Price | It define price of each application. | Float64 | 0.000000 = 140755  0.990000 = 786  1.990000 = 388  2.990000 = 253 |
| Currency | It shows currency type that used to buy application. | category | USD = 142998  XXX = 592  INR = 2  BRL = 1 |
| Size | Shows size of application. | category | Varies with device = 4538  11M = 3826  12M = 3433  13M = 2972 |
| Minimum android(version) | Shows version of application like 4.4 , 5,5. | object | 4.1 and up = 37546  5.0 and up = 24591  4.4 and up = 24183  4.0.3 and up = 11227 |
| Developer id | Define developer id that published their app in play store. | object | TRAINERIZE = 339  Subsplash Inc = 338  ChowNow = 310  Phorest = 176 |
| Developer websites | It shows website of each developer that makes apps. | object | <http://www.subsplash.com/> = 454  [http://www.chownow.com](http://www.chownow.com/) = 286  <http://www.sharefaith.com/category/church-websites.html> = 119  [https://foodsoul.pro](https://foodsoul.pro/) = 113 |
| Developer email | Define email id of developer. | object | [support@classplus.co](mailto:support@classplus.co) = 639  [appsupport@subsplash.com](mailto:appsupport@subsplash.com) = 458  [help@trainerize.com](mailto:help@trainerize.com) = 315  [eng-android@chownow.com](mailto:eng-android@chownow.com) = 310 |
| Released | Define release date of application. | category | 15-Jun-20 = 145  18-Feb-20 = 138  9-Jun-20 = 138  5-Jun-20 = 138 |
| Last updated | Define date when application is updated | object | 11-Jun-21 = 666  9-Jun-21 = 661  10-Jun-21 = 644  14-Jun-21 = 627 |
| Content rating | Define age group that gives rating for applications. | object | Everyone = 125710  Teen = 12158  Mature 17+ = 3657  Everyone 10+ = 2065 |
| Privacy policy | Shows instruction of privacy policy to the user. | object | <http://www.subsplash.com/legal/privacy> 447  <http://www.trainerize.com/privacy.aspx> 340  <https://unity3d.com/legal/privacy-policy> 338  <http://www.chownow.com/privacy-policy> 304 |
| Ad supported | Shows app is ad supported or not | object | False = 72010  True = 71598 |
| In app purchases | Define application in purchases or not. | object | False = 131555  True = 12053 |
| Editor’s choice | It shows application under editor choice or not. | object | False = 143563  True = 45 |
| Scraped Time | It shows scraped time of application. | object | 6/15/2021 21:59 = 1107  6/15/2021 22:33 = 1095  6/15/2021 22:17 = 1095  6/15/2021 21:42 = 1095 |

## In special data dictionary, we check the exclusive degrees of categories of an attribute and also total wide variety of counts in each. We also test the data type of each attribute and convert them into appropriate kind to get preferred outcomes. So, we convert object attributes into category and float into integer as a way to get accurate result.

## 

## Numerical Attributes:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column1** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **Rating** | 1038265 | 2 | 2 | 0 | 0 | 3 | 4 | 5 |
| **Rating\_Count** | 1038265 | 2937 | 255967 | 0 | 0 | 6 | 42 | 1.39E+08 |
| **Minimum\_Installs** | 1048529 | 176200 | 13259642 | 0 | 50 | 500 | 5000 | 5E+09 |
| **Maximum\_Installs** | 1048575 | 304457 | 20466334 | 0 | 84 | 695 | 7361 | 9.77E+09 |
| **Price** | 1048575 | 0 | 3 | 0 | 0 | 0 | 0 | 400 |

**Summary of numeric attributes:-**contain summary of dataset it is showing the numeric attributes of count, mean or average, is calculated by adding up the scores and dividing the total by the number of scores, standard deviation is a measure of dispersion of data values from the mean.

## Detailed Data Dictionary:-

In this step, I took one variable at a time and do the modifications in each attributes I exchange the datatype similar to my dataset. For numeric attribute, I created five range of precis and discover the count, mean. Standard deviation of every characteristic. After that I checked the quantity of outliers in every numeric attribute. Next, I checked levels, counts and categories of columns and characteristic along side imbalance inside the dataset.

**Missing data:-**

Missing values are the null values in a dataset. When there's no cost stored for any variable in an observation then that condition is dealt with as null/lacking or N/A fee. If there's an availability of null values in any dataset, then our result will in no way be accurate. We can see that our dataset carries sufficient amount of NA values and to perform analysis we need to take away these missing values from dataset with the aid of the use of suitable standards. If we aren't doing away with those null values, then we get inaccurate output.

**Criteria for cleaning the missing values:**

Next step is to analyze if there is any pattern for the missing values then I remove it for further analysis. I drop those attributes with machine algorithm techniques those have sufficient number of NA values. After that I replacing NA values in each attribute. Criteria for cleaning the missing values:As we can see ‘Developer\_website’, ‘Privacy\_policy’, ‘Released’, ‘Rating’, ‘Rating\_Count’ contains more than 10000 null values so if we randomly drop N/A values without any argument, it will remove the whole row wherever missing value exist, then almost half of our data will be deleted. That is why for that we dropped those attributes which contains highest number (10000 or more) of missing values in our dataset and less than 10000 missing values, we dropped the whole row where NA value exists.

**Exploratory Data Analysis:-**

Chart, box and whisker chart

Description automatically generated

**FIGURE 1 . Maximum\_installs**

Above boxplot suggests positively skewed distribution as it suggests that median is greater toward the lowest of the field and lower values towards lower bound. It constitute maximum of the values took place between zero to a thousand. It also carries huge quantity of outliers at the upper certain.

Chart, scatter chart

Description automatically generated

**FIGURE 2. Minimum\_installs**

The above plot is positive skewed because median is more in the direction of lower bound. Its maximum of the values passed off among 0 to a thousand and additionally consists of few outliers in the direction of upper bound.

**Data Visualization:-**

**Counts vs currency:**

Chart, bar chart

Description automatically generated

**FIGURE 3 . Count vs currency**

This bar graph represents the information between currency and counts. We can see that from this graph that the counts of USD foreign money has highest as compared to others. It is nearly 700000.

**Show highest Content rating category :-**

Chart, bar chart

Description automatically generated

**FIGURE 4. Count vs Content\_Rating**

We can noticed that people who are 18+ that human beings give highest score to the content material of all software as evaluate to the alternative classes( adults most effective , every body , mature , 17+ , Teen , Unrated) like extra then as a 60,00,00 counts number.

**Show how many applications are supported purchased feature**.:-

Chart, pie chart

Description automatically generated

**FIGURE 5. Is App Supported or Not?**

In above pie chart the blue color represent the percentage of apps that are purchased this is 99.4% and the yellow coloration represents the apps which are not bought which might be 0.6%.

**Show how many applications are free of cost to the customer:-**

Chart, pie chart

Description automatically generated

**FIGURE 6. Is App Free or Not?**

The above pie chart is showing the percentage of apps which can be free or not loose. There are 99.4% apps which can be free of price. And the relaxation aren't free.

## Modeling Implementation: -

**Modelling**

In the modelling part there are four methods Logistic Regression, KNN, Naive Bayes, random forest that I have applied on my dataset.

**Logistic Regression**: It is classification machine learning algorithm that is used to predict the probability of categorical dependent variable.

**KNN**: The KNN stands for “K-Nearest Neighbour”. It is a supervised machine learning algorithm. The algorithm can be used to solve both classification and regression problem statements. The number of nearest neighbours to a new unknown variable that has to be predicted or classified is denoted by the symbol 'K'.

**Naive Bayes**:  It is classification algorithm that is suitable for binary and multiclass classification.

It performs well in cases of categorical input variables compared to numerical variables.

**Random Forest CLASSIFIER**: It is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It is also a meta estimator that fits a number of decision tree classifier on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.

|  |  |  |
| --- | --- | --- |
|  | MODEL | TRAIN\_TEST\_SPLIT |
| 0 | Logistic Regression | 59.224857 |
| 1 | KNN | 53.829545 |
| 2 | Naïve Bayes | 58.923835 |
| 3 | Random Forest | 57.429588 |

**Cross Validation:-** In cross validation, the training set is randomly split into k(usually between 5 to 10) subsets known as folds. Where k-1 folds are used to train the model and the fold is used to test the model.

Stratified K Fold:

This skip-validation item returns stratified folds and is a version of K-Fold. The folds are made via the usage of keeping the proportion of samples in every elegance. I divided the facts into five stratified folds. The 4 folds are then used to healthful the model, and the 5th fold is used to test it. Repeat till each fold has been used as a check set. Then add together all of the effects and calculate the common. That might be the version's metric of achievement.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | MODEL | TRAIN\_TEST\_SPLIT | KFOLDS\_5 | STRATIFIEDKFOLD\_5 |
| 0 | Logistic Regression | 59.224857 | 52.973865 | 52.880016 |
| 1 | KNN | 53.829545 | 51.017988 | 56.029452 |
| 2 | Naïve Bayes | 58.923835 | 53.882115 | 54.376866 |
| 3 | Random Forest | 57.429588 | 53.504490 | 56.071677 |

## Random Train-Test-Split

This approach combines the k-fold-cross-validation method with ordinary train-test-splits. I create random divides of the facts inside the training-check set, similar to the circulate-validation method, after which repeat the method of splitting and attempting out the algorithm many times. I divided the information into five Repeated Random Test-Train Splits

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | MODEL | TRAIN\_TEST\_SPLIT | KFOLDS\_5 | STRATIFIEDKFOLD\_5 | RRTESTTRAINSPLITS\_5 |
| 0 | Logistic Regression | 59.22 | 52.97 | 52.88 | 53.69 |
| 1 | KNN | 53.83 | 51.02 | 56.03 | 58.80 |
| 2 | Naïve Bayes | 58.92 | 53.88 | 54.38 | 58.54 |
| 3 | Random Forest | 57.43 | 53.50 | 56.07 | 58.12 |

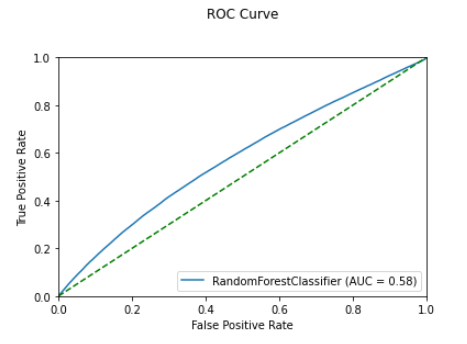
## Confusion Matrix Corresponding to Random Forest Classifier Algorithm:- Well, it's far a overall performance measurement for machine learning trouble where output may be two or extra lessons. It is a table with 4 exceptional mixtures of predicted and actual values. And it a tabular summary of the quantity of correct and wrong predictions made by way of a classifier.Chart, treemap chart Description automatically generated

**FIGURE 7. Actual Values Vs Predicted Values**

## Confusion Matrix Corresponding to Random Forest

When the target variable's actual value is 0 and the predicted value is also 0 and 38.60% of observation fall into first quadrant. 18.20% of observations fall into the second quadrant; whilst the target variable's actual value is 0 and predicted value is 1. 24.37% of observations fall into the third quadrant; and while the target variable's actual value is 1 and predicted value 0, 18.83% of observations fall into the fourth quadrant. The actual and predicted values are each 1.

**ROC CURVE (RECEIVER OPERATING CHARACTERISTIC) :-**Roc curves are widely used in binary classification to study the output of a classifier



**FIGURE 8. True Positive Rate Vs False Positive Rate**

FROM ABOVE GRAPH, AREA UNDER THE CURVE SCORE IS 0.58. IT SHOW THAT IT IS AN OPTIMAL CLASSIFIER.

**Conclusion: -** In conclusion I can say that I carry out all feasible operations and algorithms to find a few basic things from our project and after locating da accuracy of this undertaking by using 4 models , I discover properly result which means its smooth to use in the future. Secondly By using pie charts and bar graphs I defined visualization of some few attributes that assist us to see direct consequences like minimum and most values from the special categories.

**References:-**

["Google Play Store APK"](https://www.apkmirror.com/apk/google-inc/google-play-store/google-play-store-29-0-14-release/google-play-store-29-0-14-21-0-pr-423376132-android-apk-download/).

Jump up to “Info. Google Help. Retrieved November 13, 2020.

["Number of Google Play Store apps 2017 | Statistic"](https://www.statista.com/statistics/266210/number-of-available-applications-in-the-google-play-store/). Statista. Retrieved January 3, 2018.

["Number of Android applications on the Google Play store"](https://www.appbrain.com/stats/number-of-android-apps). appbrain.com. Retrieved August 12, 2020.

Jump up to “Number. AppBrain. February 9, 2017. Archived from [the original](https://www.appbrain.com/stats/number-of-android-apps) on February 10, 2017. Retrieved February 24, 2017.

 Pasternak’s, Jacob (February 28, 2017). ["Google now lets apps display sale prices in the Play Store"](https://www.theverge.com/2017/2/28/14760940/google-play-store-app-sales). [The Verge](https://en.wikipedia.org/wiki/The_Verge). [Vox Media](https://en.wikipedia.org/wiki/Vox_Media). Retrieved March 1, 2017.

 Rossignol, Derrick (February 28, 2017). ["Google now lets developers offer sales on Android apps"](https://www.engadget.com/2017/02/28/google-now-lets-developers-offer-sales-on-android-apps/). [Engadget](https://en.wikipedia.org/wiki/Engadget). [AOL](https://en.wikipedia.org/wiki/AOL). Retrieved March 1, 2017.

["How to use a Google Play gift card"](https://www.androidcentral.com/how-use-google-play-gift-card). Android Central. February 5, 2019. Retrieved June 30, 2020.

O'Brien, Terrence (May 2, 2012). ["Google Play adds carrier billing for music, movies and books"](https://www.engadget.com/2012/05/02/google-play-adds-carrier-billing-for-music-movies-and-books/). [Engadget](https://en.wikipedia.org/wiki/Engadget). [AOL](https://en.wikipedia.org/wiki/AOL). Retrieved February 23, 2017.

["Two Years Later, Google Play Support Docs Acknowledge 48-Hour Refund Window"](https://www.androidpolice.com/2016/06/02/two-years-later-google-play-support-docs-acknowledge-48-hour-refund-window/). Android Police. June 2, 2016. Retrieved June 30, 2020.