Android Authenticity Prediction

Capstone Project

Supervised Machine Learning

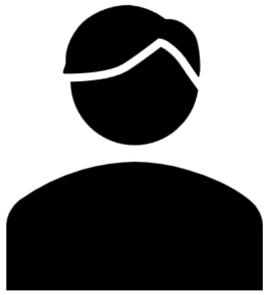
Hello & welcome to presentation

Android Authenticity Prediction

Welcome to the Android Authenticity Prediction ML classification . This is designed to analyze and evaluate the authenticity of Android devices, providing you with a reliable prediction of whether app on the device is malware or benign

LETS GET STARTED

Team Representation



Vinayak Marathe

Riya Patel

Project Timeline

Android Authenticity Prediction

Step 1	Step 2	Step 3	Step 4	Step 5
				
Introduction, Understanding Data	Data Wrangling & Data Visualization	Hypothesis Testing & Data Preprocessing	Feature Manipulation & Model Implementation	Future Work & Conclusion

1

Introduction

Information about Android Authentication Prediction

Introduction

The Android Authenticity Prediction ML model is based on supervised machine learning techniques. It is trained using a large dataset of Android applications, both malicious and benign. The model is then tested on a separate dataset to evaluate its performance. The performance of the model is measured using metrics such as precision, recall, and accuracy.

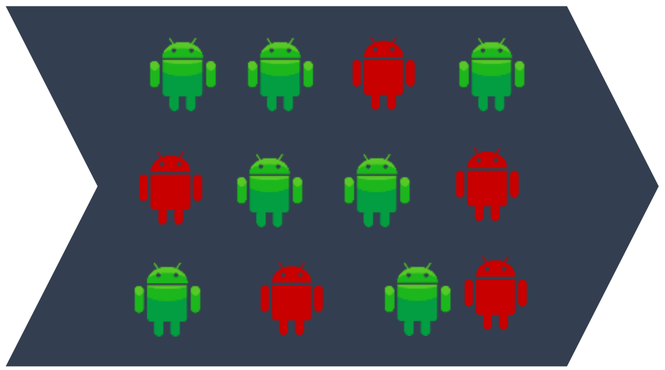
One of the key advantages of the Android Authenticity Prediction ML model is that it is highly accurate in detecting malicious applications. The model can accurately detect up to **91%** of malicious applications with a low false positive rate. This makes it an effective tool for protecting Android users from malicious applications.

Another advantage of the model is that it is scalable and can be easily integrated into existing mobile security solutions. The model can be deployed on mobile devices or on servers to detect malicious applications in real-time. This makes it an ideal solution for mobile security providers, app stores, and other organizations that need to protect their users from malicious applications.

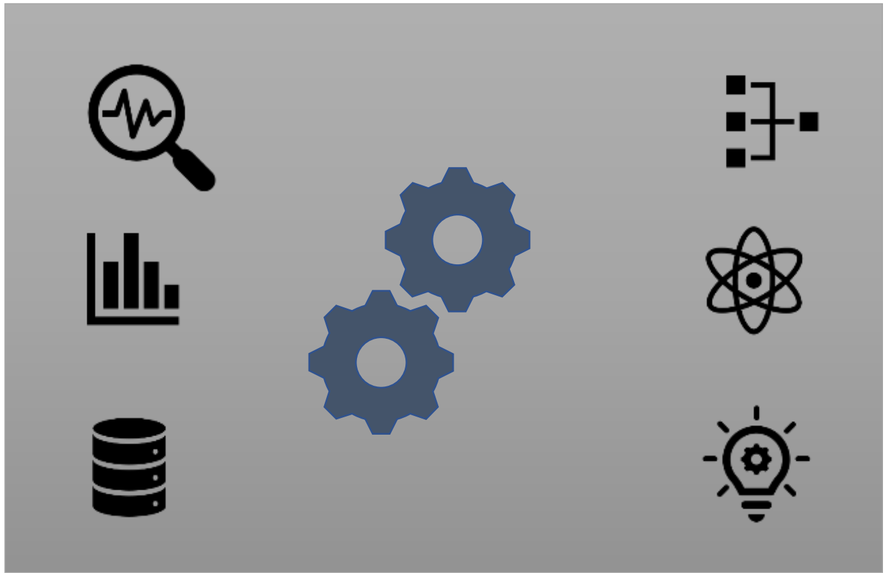
However, there are also some challenges that need to be addressed when using the Android Authenticity Prediction ML model. One of the challenges is the need for a large dataset of Android applications for training and testing the model. This requires a significant amount of resources and expertise to collect and analyze the data.

Project Architecture

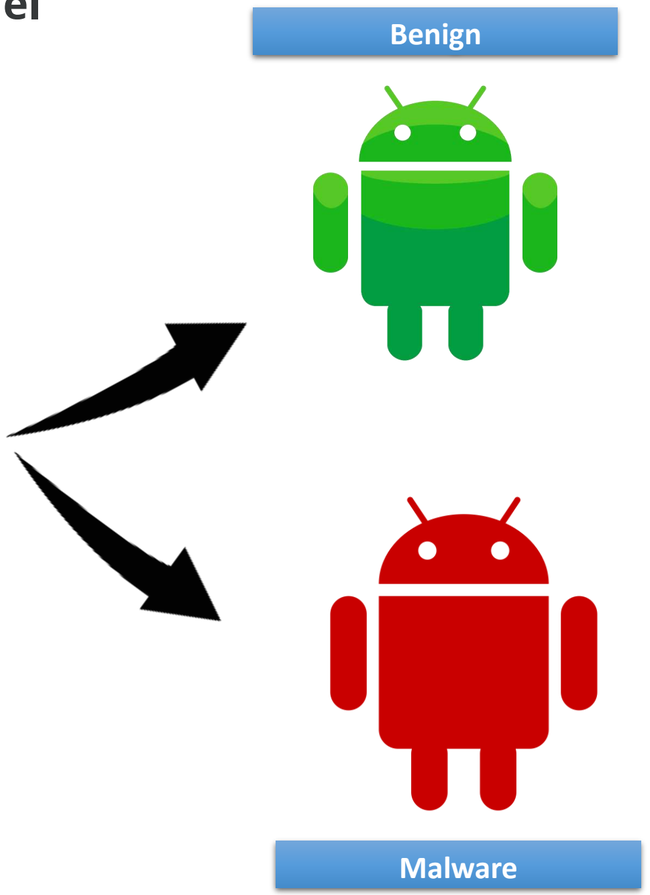
Android Authenticity Prediction Classification Model



Data



Machine Learning Model



2

Problem Statement

Our business problem overview

Problem Statement

The problem of Android Authenticity Prediction is to develop a machine learning model that can accurately predict whether an Android application (app) is authentic or not. With the increase in the number of mobile apps, the risk of downloading malicious apps has also increased. Malicious apps can steal sensitive user data, perform unwanted actions, and damage the user's device. Therefore, it is essential to develop a model that can accurately predict the authenticity of Android apps and help users make informed decisions about which apps to download and install on their devices. The challenge is to identify the relevant features that can distinguish between authentic and malicious apps and to train a classification model that can generalize well to new, unseen apps. Additionally, the model should be able to handle the large and dynamic nature of the mobile app ecosystem, where new apps are constantly being developed and released.

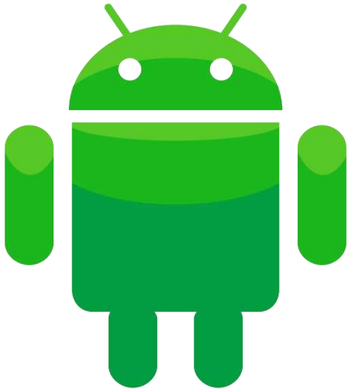
3

Understanding Data

What are features and label in our data

Default Permissions

- Access DRM content. (S)
- Access email provider data (S)
- Access all system downloads (S)
- Access download manager. (S)
- Advanced download manager functions. (S)
- Audio file access (S)
- Install DRM content. (S)
- Modify google service configuration (S)
- Modify google settings (S)
- Move application resources (S)
- Read google settings (S)
- Send download notifications. (S)
- Voice search shortcuts (S)
- Access surfaceflinger (S)
- Access checkin properties (S)
- Access the cache filesystem (S)
- Access to passwords for google accounts (S)
- Act as an account authenticator (S)
- Bind to a wallpaper (S)
- Bind to an input method (S)
- Change screen orientation (S)
- Coarse (network-based) location (S)
- Control location update notifications (S)
- Control system backup and restore (S)
- Delete applications (S)
- Delete other applications caches (S)
- Delete other applications data (S)
- Directly call any phone numbers (S)
- Directly install applications (S)
- Disable or modify status bar (S)
- Discover known accounts (S)
- Display unauthorized windows (S)
- Enable or disable application components (S)
- Force application to close (S)
- Force device reboot (S)
- Full internet access (S)
- Interact with a device admin (S)
- Manage application tokens (S)
- Mock location sources for testing (S)
- Modify battery statistics (S)
- Modify secure system settings (S)
- Modify the google services map (S)
- Modify/delete USB storage contents modify/delete SD card contents (S)
- Monitor and control all application launching (S)
- Partial shutdown (S)
- Permanently disable device (S)
- Permission to install a location provider (S)
- Power device on or off (S)
- Press keys and control buttons (S)
- Prevent app switches (s)
- Read frame buffer (S)
- Read instant messages (S)
- Read phone state and identity (S)
- Record what you type and actions
- You take (S)
- Reset system to factory S)
- Run in factory test mode (S)
- Set time (S)
- Set wallpaper size hints (S)
- Start IM service (S)
- Update component usage statistics (S)
- Write contact data (S)
- Write instant messages (S)



Development Tools Permissions

- Enable application debugging (D)
- Limit number of running processes (D)
- Make all background applications close (D)
- Send linux signals to applications (D)

Hardware controls Permissions

- Test hardware (S)
- Control flashlight (S)
- Control vibrator (S)

- Record audio (D)
- Take pictures and videos (D)
- Change your audio settings (D)

Network Communication Permissions

- Receive data from internet (S)
- View wi-fi state (S)
- View network state (S)
- Broadcast data messages to applications. (S)
- Download files without notification (S)

- Control near field communication (D)
- Create bluetooth connections (D)
- Full internet access (D)
- Make/receive internet calls (D)

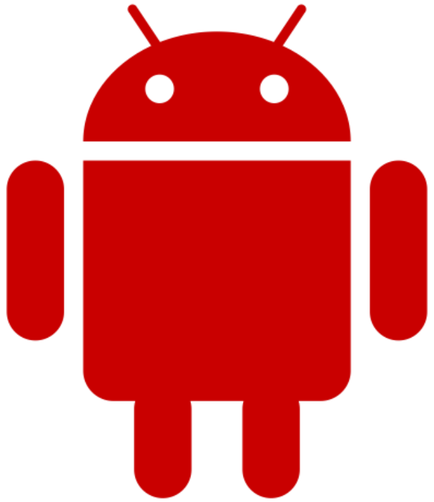
Phone Calls Permissions

- Modify phone state (S)

- Read phone state and identity (D)
- Intercept outgoing calls (D)

Service that cost you Money Permissions

- Directly call phone numbers (D)



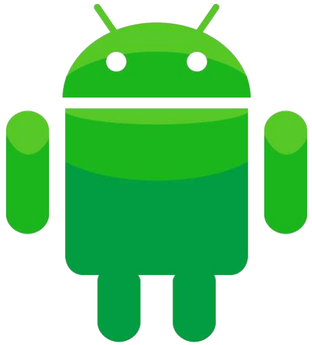
- Send SMS messages (D)

Modify/delete USB storage contents modify/delete SD card contents (D)

Storage Permissions

System Tools Permissions

- Automatically start at boot (S)
- Change background data usage setting (S)
- Expand/collapse status bar (S)
- Force stop other applications (S)
- Kill background processes (S)
- Measure application storage space (S)
- Read subscribed feeds (S)
- Read sync settings (S)
- Read sync statistics (S)
- Read/write to resources owned by diag (S)
- Send package removed broadcast (S)
- Send sticky broadcast (S)
- Set preferred applications (S)
- Set wallpaper (S)
- Set wallpaper size hints (S)



- Write access point name settings (D)
- Write subscribed feeds (D)
- Write sync settings (D)
- Bluetooth administration (D)
- Change wi-fi state (D)
- Allow wi-fi multicast reception (D)

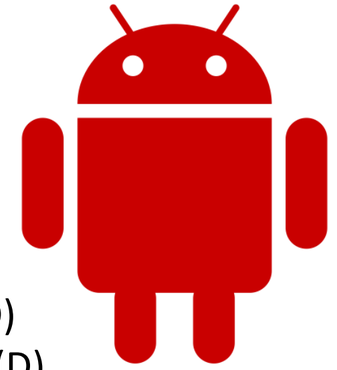
- Change network connectivity (D)
- Change your UI settings (D)
- Delete all application cache data (D)
- Disable keylock (D)
- Display system-level alerts (D)
- Format external storage (D)
- Modify global animation speed (D)
- Make application always run (D)
- Modify global system settings (D)
- Mount and unmount filesystems (D)
- Prevent device from sleeping (D)

- Reorder running applications (D)
- Retrieve running applications (D)
- Set time zone (D)

Your Accounts Permissions

- Act as the accountmanagerservice (S)
- Discover known accounts (S)
- Access all google services (S)
- Read google service configuration (S)
- View configured accounts (S)

- Blogger (D)
- Google app engine (D)
- Google docs (D)
- Google finance (D)
- Google maps (D)
- Google spreadsheets (D)
- Google voice (D)
- Google mail (D)
- Picasa web albums (D)
- Youtube (D)
- Youtube usernames (D)
- Access other google services (D)
- Act as an account authenticator (D)
- Contacts data in google acjcounts (D)
- Manage the accounts list (D)
- Use the authentication credentials of an account (D)



Your Location Permissions

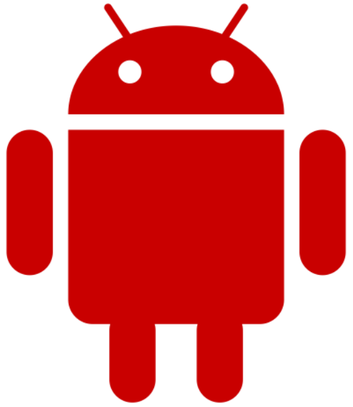
- Access extra location provider commands (S)

- Coarse (network-based) location (D)
- Fine (GPS) location (D)
- Mock location sources for testing (D)

Your Messages Permissions

- Send gmail (S)
- Send sms-received broadcast (S)
- Send wap-push-received broadcast (S)

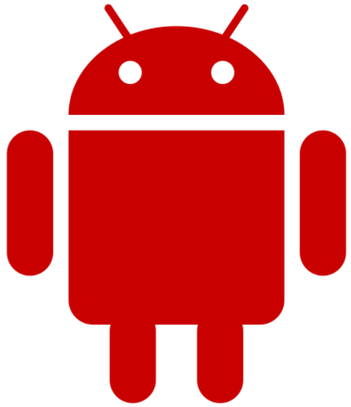
- Receive SMS (D)
- Receive WAP (D)
- Write instant messages (D)
- Read email attachments (D)
- Edit SMS or MMS (D)
- Modify gmail (D)
- Read gmail (D)
- Read gmail attachment previews (D)
- Read SMS or MMS (D)
- Read instant messages (D)
- Receive MMS (D)



Your Personal Information Permissions

- Retrieve system internal state (S)
- Set alarm in alarm clock (S)
- Choose widgets (S)
- Write to user defined dictionary (S)

- Add or modify calendar events and send email to guests (D)
- Read browser s history and bookmarks (D)
- Read calendar events (D)
- Read contact data (D)
- Read sensitive log data (D)
- Read user defined dictionary (D)
- Write browsers history and bookmarks (D)
- Write contact data (D)

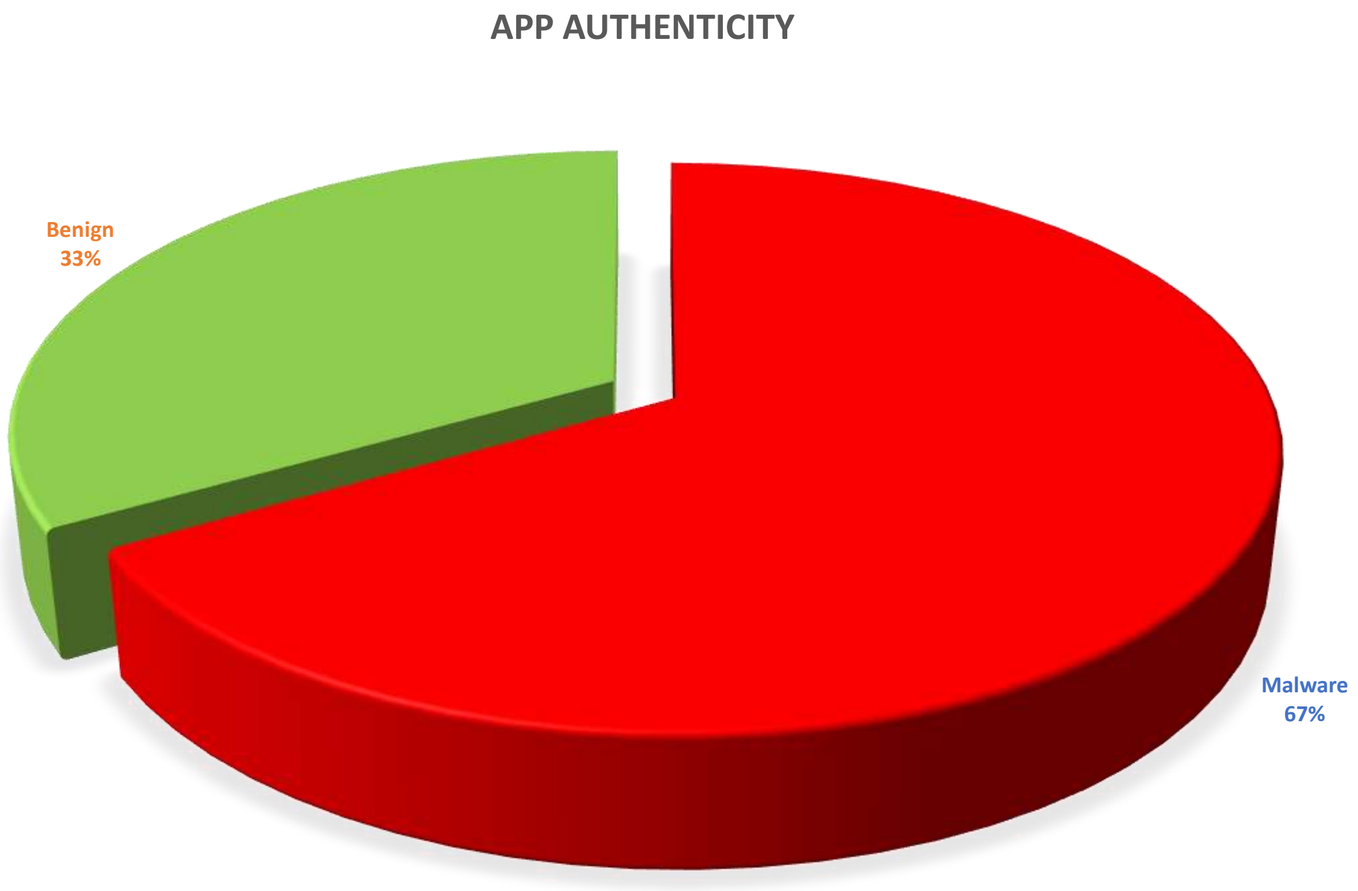


4

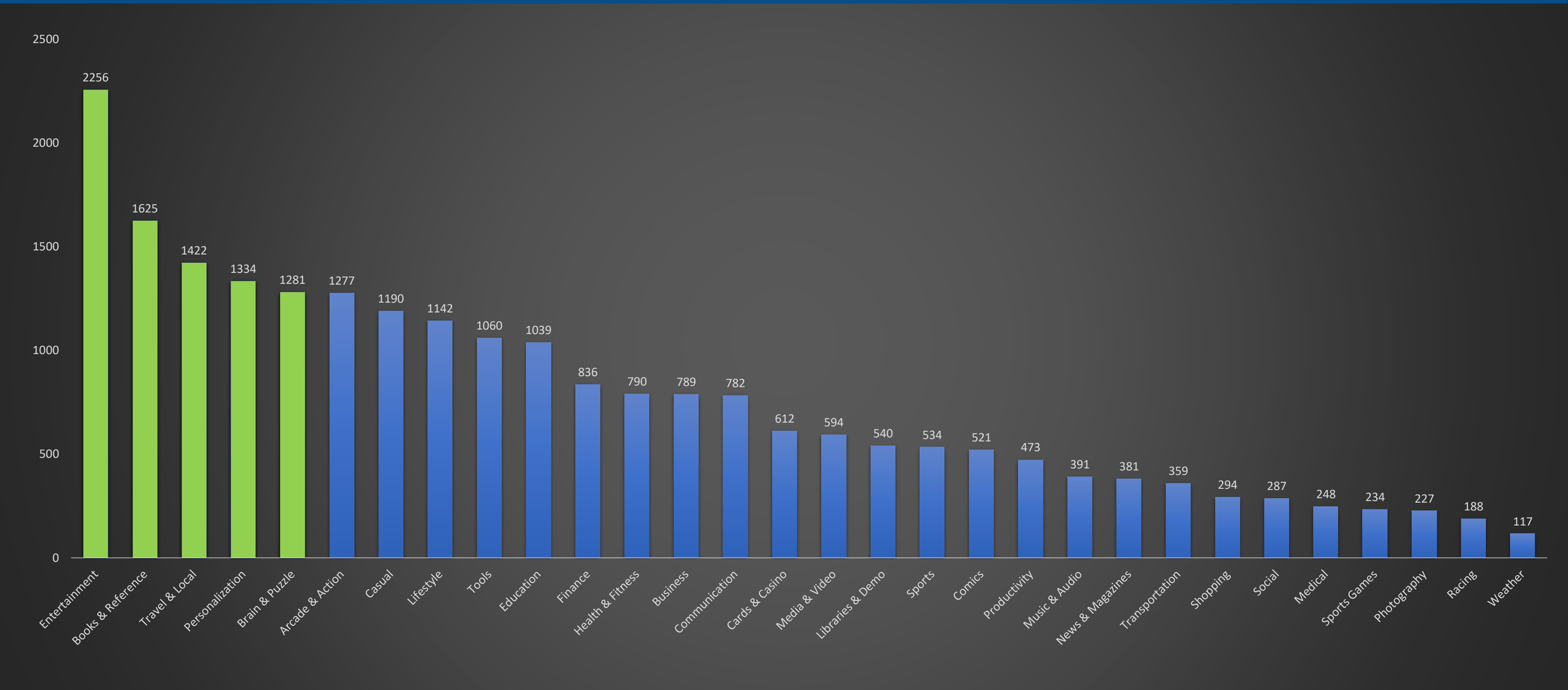
Data Wrangling

Finding meaningful insights

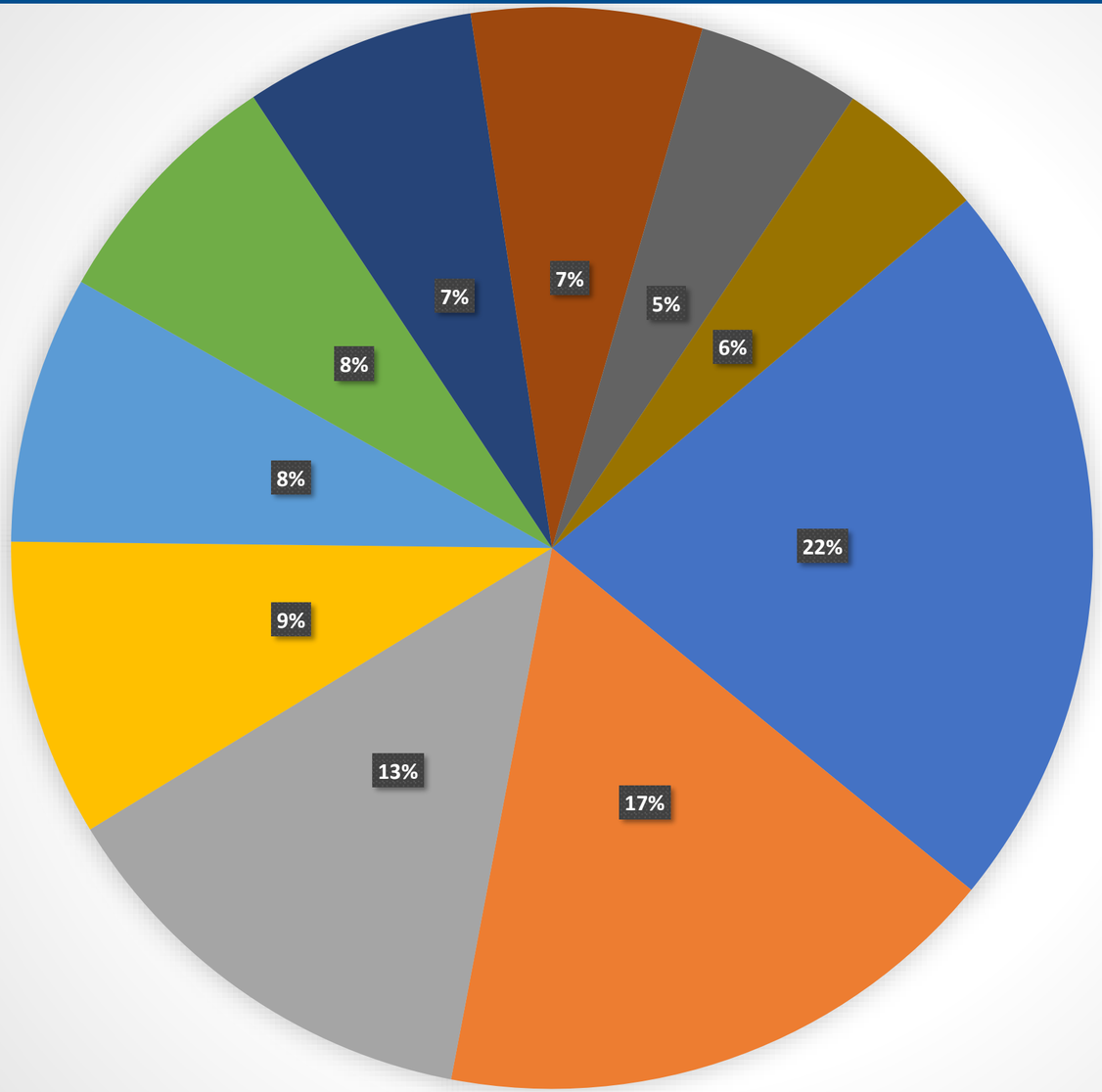
Count of Malware & Benign App



Number of apps by category

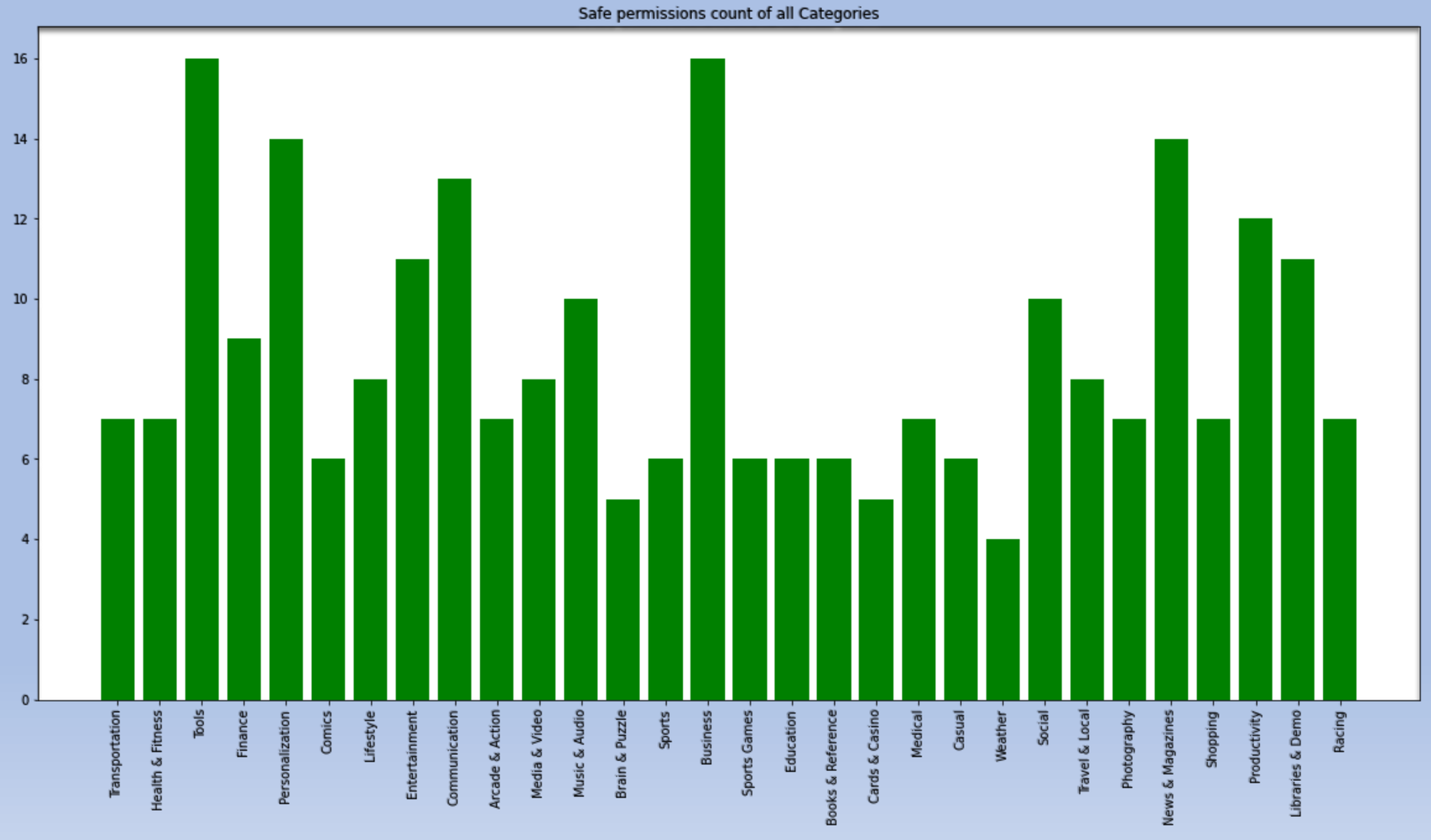


Top apps by category

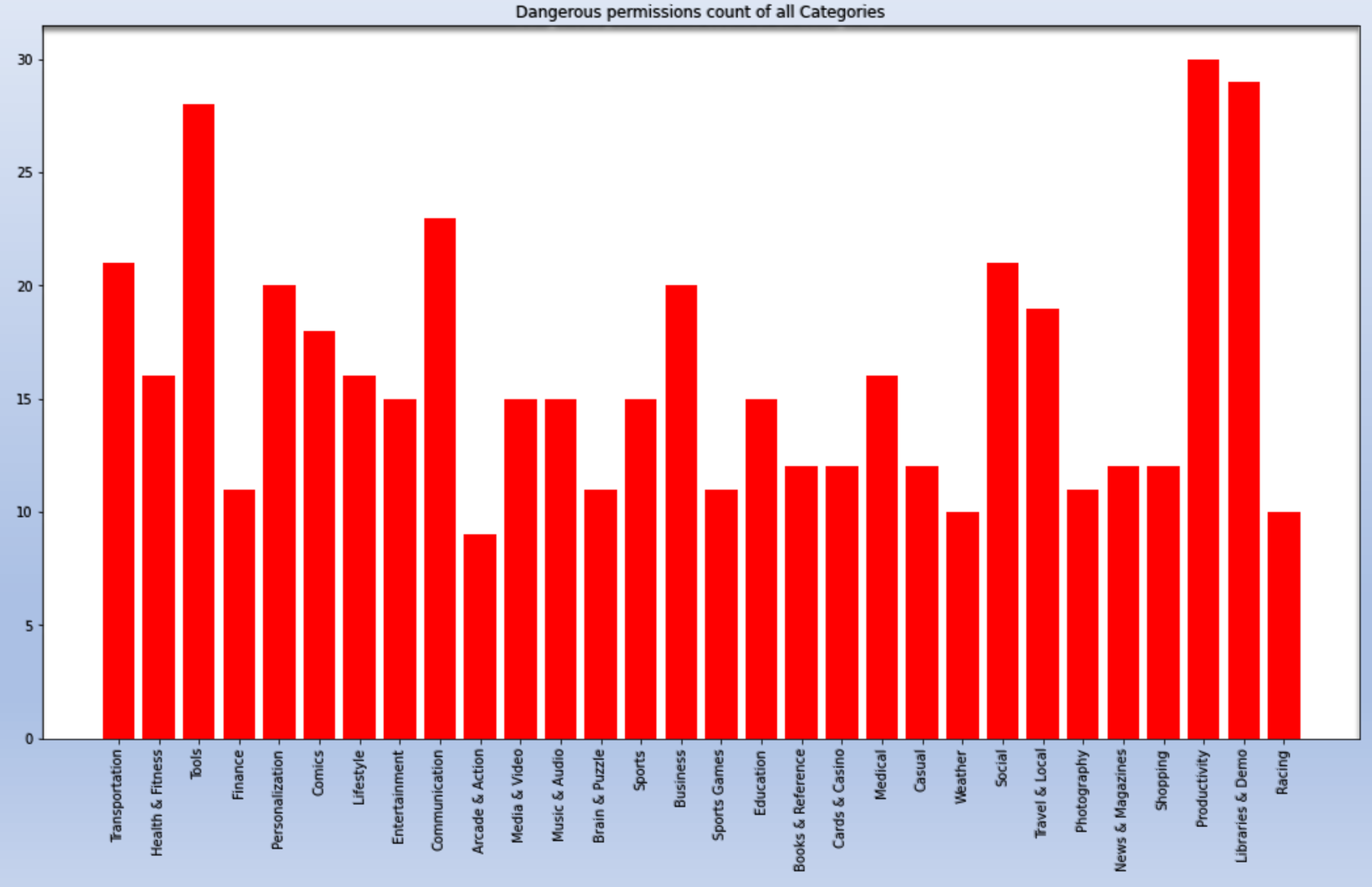


Communication has highest percentage Rating followed by Racing and Media & Video

Safe Permission Count of all category



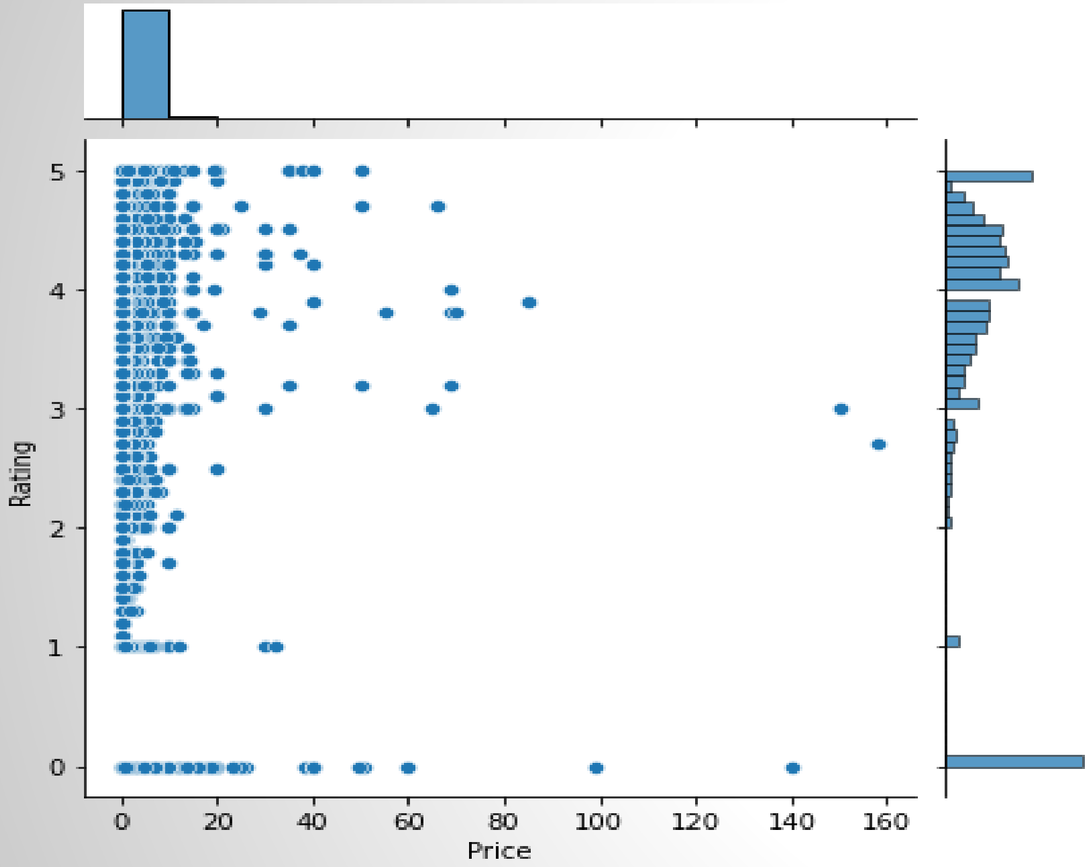
Dangerous Permission Count of all category



Correlation

	Rating	Number of ratings	Price	Dangerous permissions count	Safe permissions count
Rating	1.00	0.06	-0.15	0.08	0.12
Number of ratings	0.06	1.00	-0.02	0.11	0.12
Price	-0.15	-0.02	1.00	-0.00	-0.04
Dangerous permissions count	0.08	0.11	-0.00	1.00	0.70
Safe permissions count	0.12	0.12	-0.04	0.70	1.00

Joint plot



It is showing the relation between ratings and price of the app. This graph can analyze us that does the price of an app affect its rating?

We find that the majority of top rated apps (rating over 4) range has the vast majority of apps price th emselves under $10.

Insights

- There are 174 unique permission
- There are 12 main categories of permission Count namely Default, Development tools, Hardware controls, Network communication, Phone calls, Services that cost you money, Storage, System tools, Your accounts, Your location, Your messages & Your personal information
- There are 7175 duplicate values in our app name column.
- there are many repetitive apps with different ratings and number of ratings
- we keeping the first instance of the app with maximum number of ratings
- As number of rating keeps on increasing as app getting older. Rating may drop but Number of ratings never decreases
- There are 2827 apps in the Entertainment category which is the highest number among all the category.
- Travel & Local has the second highest i.e. 2154 number of available applications.
- Followed by Entertainment and Travel, Books & Reference and Arcade & Action has 1959 apps are there in the dataset.
- Weather has the least number of application available.
- Applications with 0 rating has the highest number with 3303. while application with 5 star, 4 start and 4.2 start has probably equal number of applications
- There are very less number of application with 1 or 2 star.

Insights

- Free apps with Number of ratings greater than 5000 and Rating greater than 4
- Super stickman Golf tops the list. followed by Rage Reader and Titanium backup
- here are 20000 malwares and 9999 benign apps in the dataset so the data is imbalanced and we have to deal with it.
- We have found top 10 features that determine whether the app is malware or not
- We have found the top 5 category with top 5 apps in each category where rating is 5 star.
- We have also found some apps with 4 star for each of the different category.
- Mostly every category has application with 4star and 5 star.
- We can see that Communication has an average of 230 numbers of overall ratings for a particular app in that particular category followed by Racing and Media & Video with 179 and 130 number of overall ratings.
- Next we have found the data frame of top 10 apps with average rating of 4.2 star named Weather, Transportation, Tools, and Productivity.

5

Hypothesis Testing

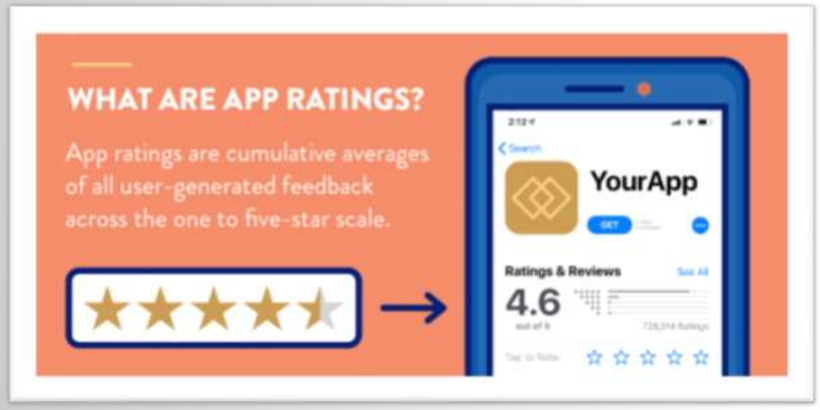
Three different hypothesis on our project

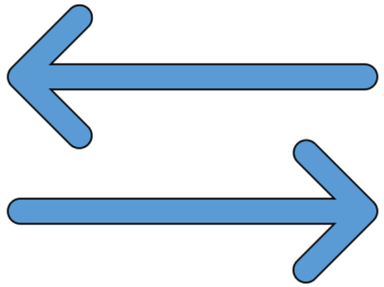
Hypothesis Testing

Hypothetical Statement - 1 : The app's rating is a significant predictor of its authenticity

Null Hypothesis: The app's rating is not significantly related to its authenticity.

Alternative hypothesis: The app's rating is significantly related to its authenticity.





Result

The t-statistic represents the difference between the means of the two groups (authentic and fake app ratings) relative to the variability in the data. In this case, the negative t-statistic value indicates that the mean rating of the fake apps is lower than the mean rating of the authentic apps.

The p-value is the probability of observing the difference in mean ratings between the authentic and fake apps, or a more extreme difference, assuming that the null hypothesis (that there is no significant relationship between app rating and authenticity) is true. A p-value of 0.0 means that the probability of observing the difference in mean ratings or more extreme differences is extremely low, virtually zero. Therefore, we reject the null hypothesis and conclude that there is a statistically significant relationship between app rating and authenticity.

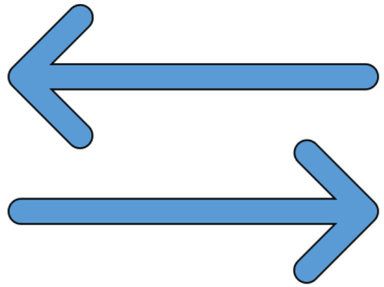
In summary, the output suggests that the app's rating is a significant predictor of its authenticity, and the mean rating of authentic apps is significantly higher than that of fake apps.

Hypothesis Testing

Hypothetical Statement - 2 : The app's category is a significant predictor of its authenticity.

Null Hypothesis: The app's category is not significantly related to its authenticity.

Alternative hypothesis: The app's category is significantly related to its authenticity.



Result

The chi-square statistic measures the difference between the observed and expected frequencies in each cell of the contingency table. In this case, the large chi-square statistic indicates a significant difference between the observed and expected frequencies, suggesting that there is a relationship between app category and authenticity.

The degrees of freedom represent the number of independent observations in the contingency table. In this case, the degrees of freedom is 29, which is the number of categories minus one, multiplied by the number of authenticity levels minus one.

The p-value is the probability of observing the relationship between app category and authenticity, or a more extreme relationship, assuming that there is no relationship between the two variables (i.e., the null hypothesis). A p-value of 0.0 indicates that the probability of observing the relationship or a more extreme relationship by chance is essentially zero. Therefore, we reject the null hypothesis and conclude that there is a statistically significant relationship between app category and authenticity.

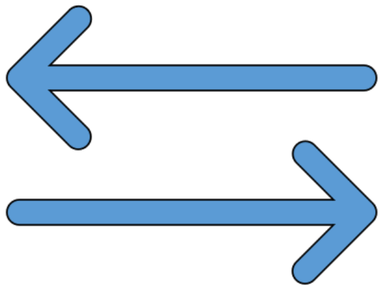
In summary, the output suggests that the app's category is a significant predictor of its authenticity, and there is a significant relationship between app category and authenticity.

Hypothesis Testing

Hypothetical Statement – 3 : The number of reviews is a significant predictor of an app's authenticity.

Null Hypothesis: The number of reviews is not significantly related to an app's authenticity.

Alternative hypothesis: The number of reviews is significantly related to an app's authenticity.



Result

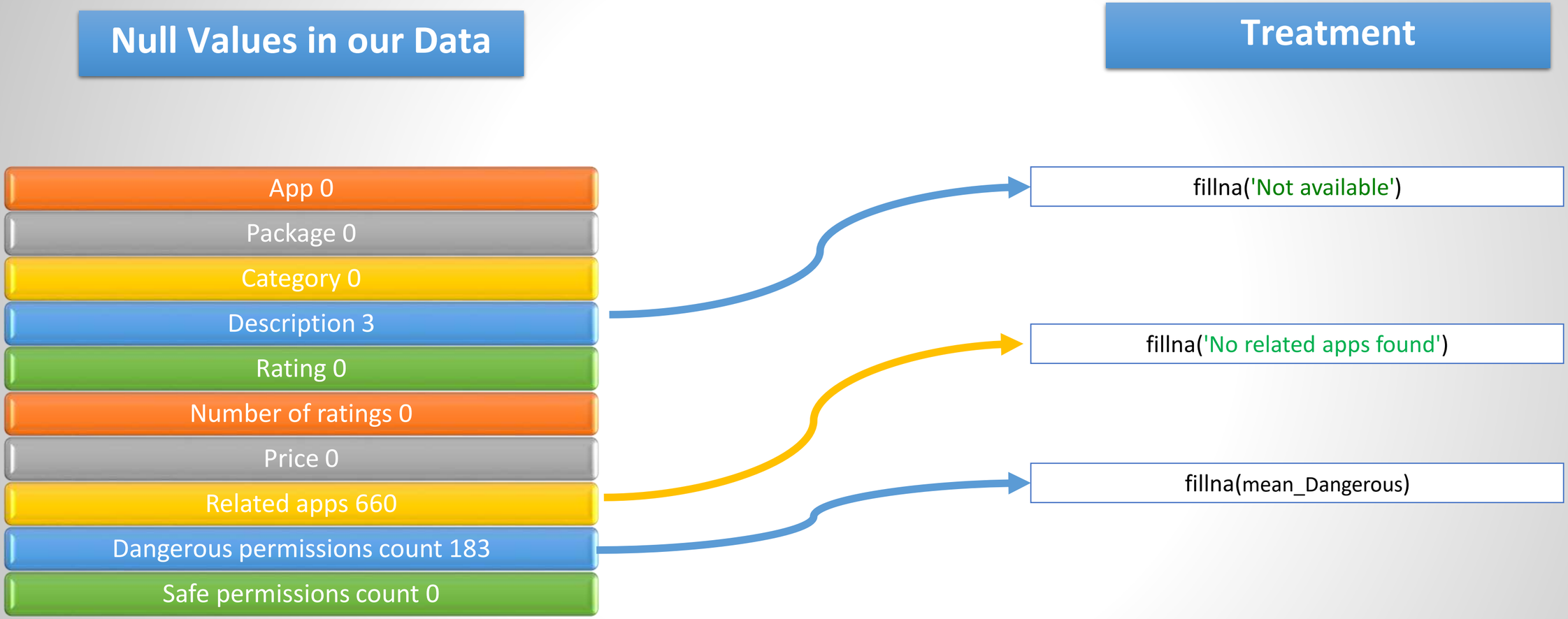
The result of the chi-square test in the code provided indicates that the p-value is 0.0, which is less than the typical significance level of 0.05. This means that we reject the null hypothesis that the number of ratings is not significantly related to the Class of the app and conclude that there is a significant relationship between the number of ratings and the Class of the app.
The chi-square statistic is a measure of how different the observed values are from the expected values under the null hypothesis. In this case, the observed values are the contingency table of the number of ratings and Class, and the expected values are calculated assuming the null hypothesis is true. The degrees of freedom tell us how many values can vary freely in the chi-square distribution. The large chi-square statistic and degrees of freedom indicate that there are many categories with large differences between the observed and expected values, supporting the conclusion of a significant relationship between the number of ratings and the Class of the app.

6

Feature Engineering

Feature Manipulation , Handling Imbalanced Data & Feature selection

Null values treatment



Categorical Encoding

In our data there is 30 different categories

Here we are using **one hot encoding** is a technique used to represent categorical variables as numerical values in a machine learning model. As the category column in our dataset is an important feature for evaluation, we have to convert it into numeric for model implementation so we have used this technique. The advantages of using one hot encoding includes:

It allows the use of categorical variables in models that require numerical input.

It can improve model performance by providing more information to the model about the categorical variable.

It can help to avoid the problem of ordinality, which can occur when a categorical variable has a natural ordering (e.g. “small”, “medium”, “large”).

Textual Data Pre-processing

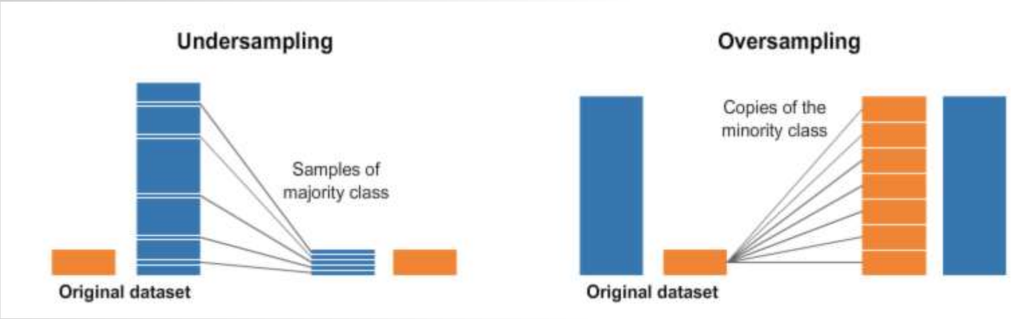
In our data , App column which contains App name has many non English text that we have to remove

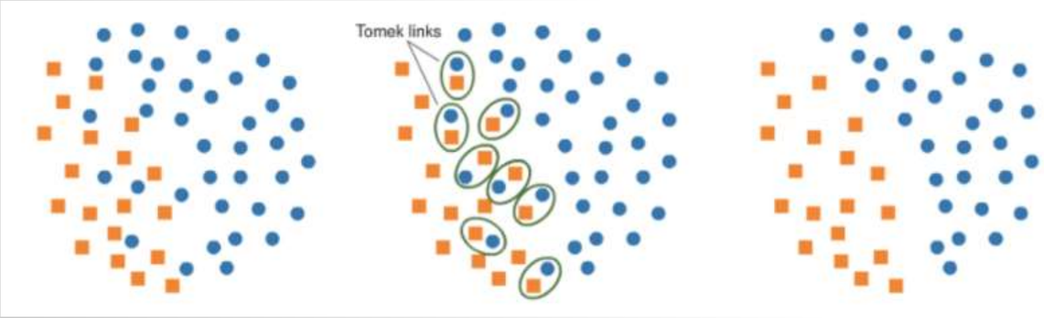
Here we are using **Re (Regular Expression) Function** to remove non English , stop words, etc. from the app name column
For example:
App name : ¼ê¸° Toast (Tour a.. is converted to empty string

We found out there is many non English app names
And 52 of them is converted into Empty string., we drop that rows from our Data Frame

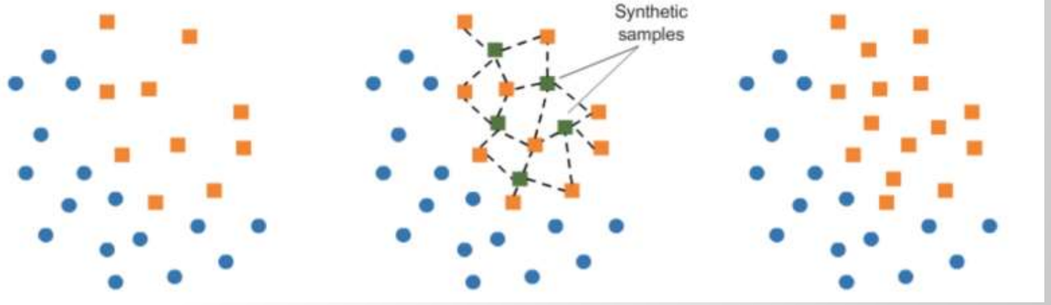
Handling Data Imbalanced

Our target variable ‘Class’ is imbalanced. we used following techniques





Tomek Link



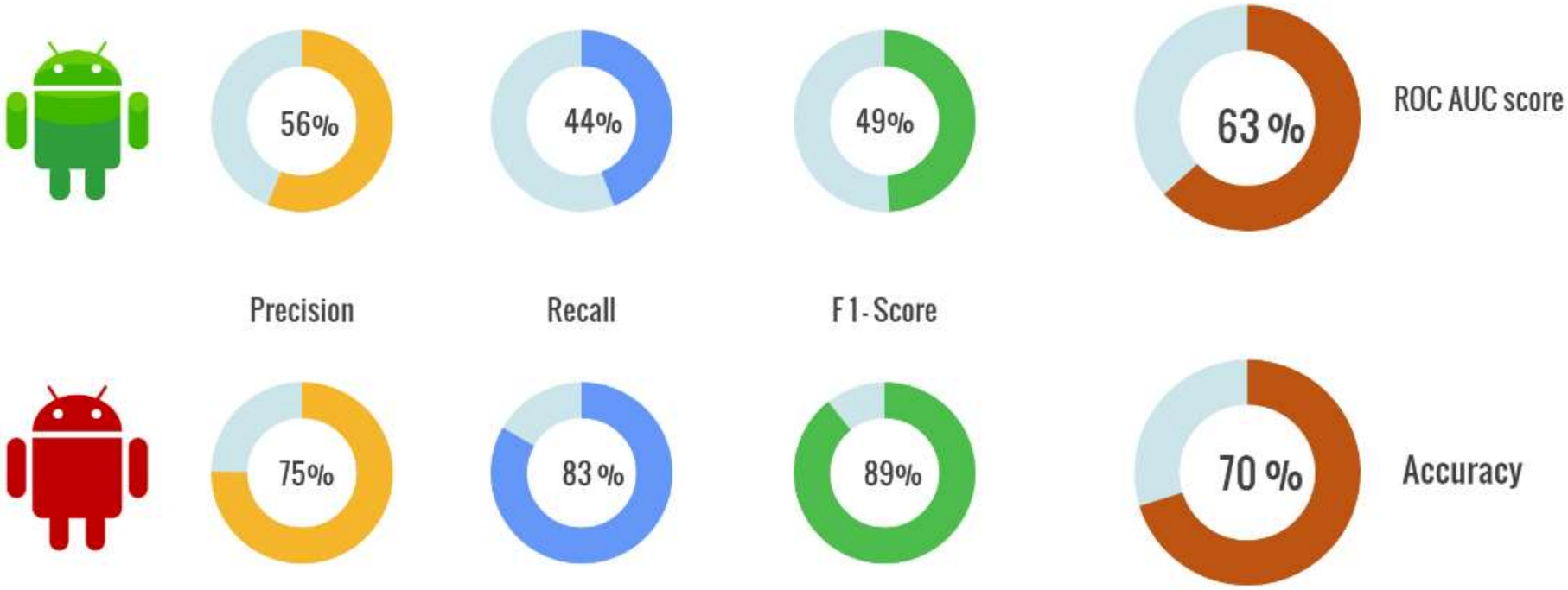
SMOTE

7

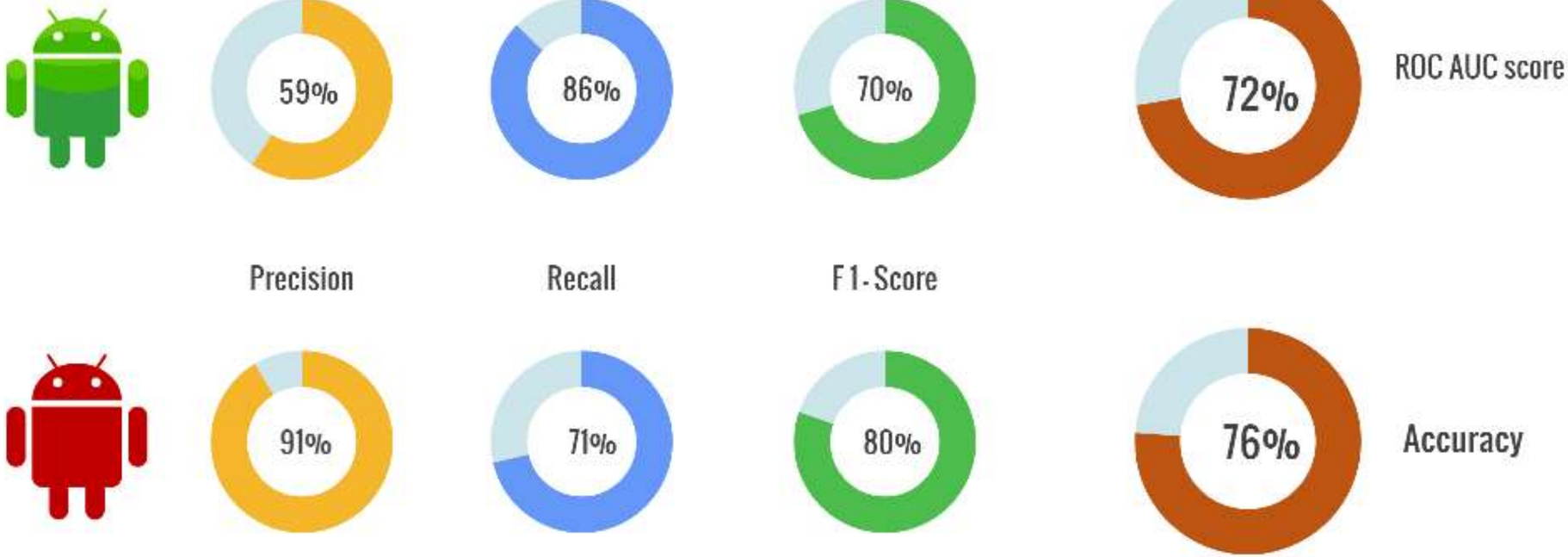
Model Impementation

Implement various ML models

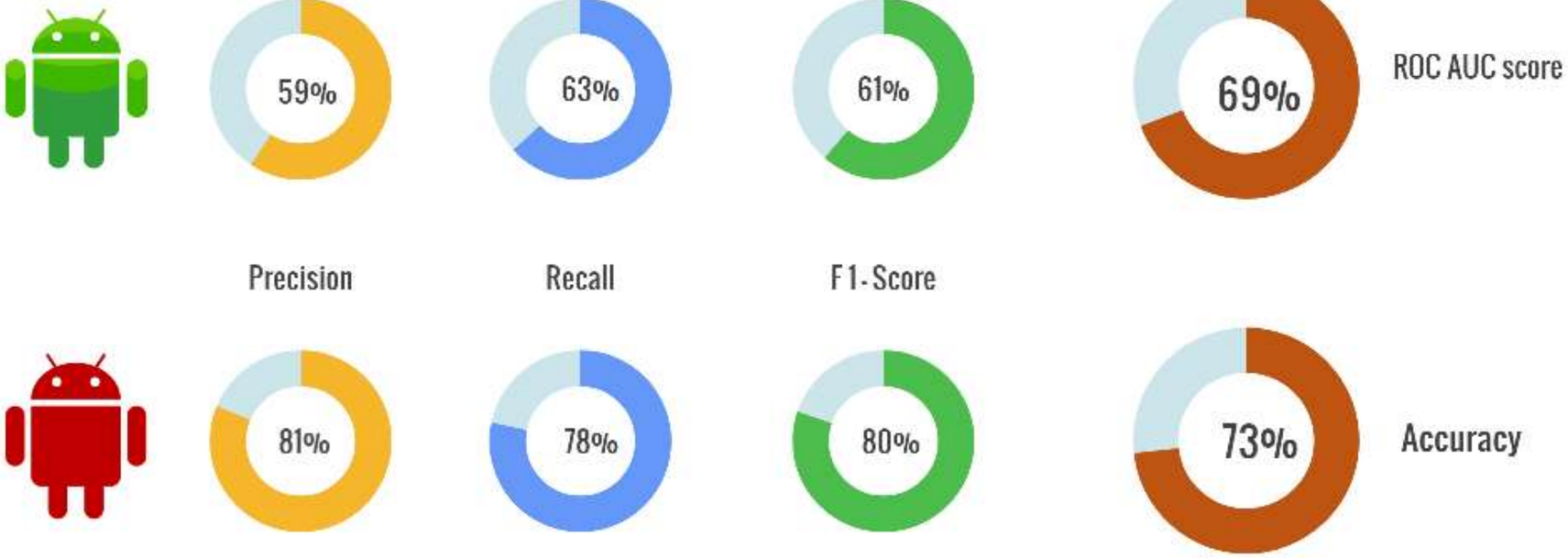
Model 1: Logistic Regression



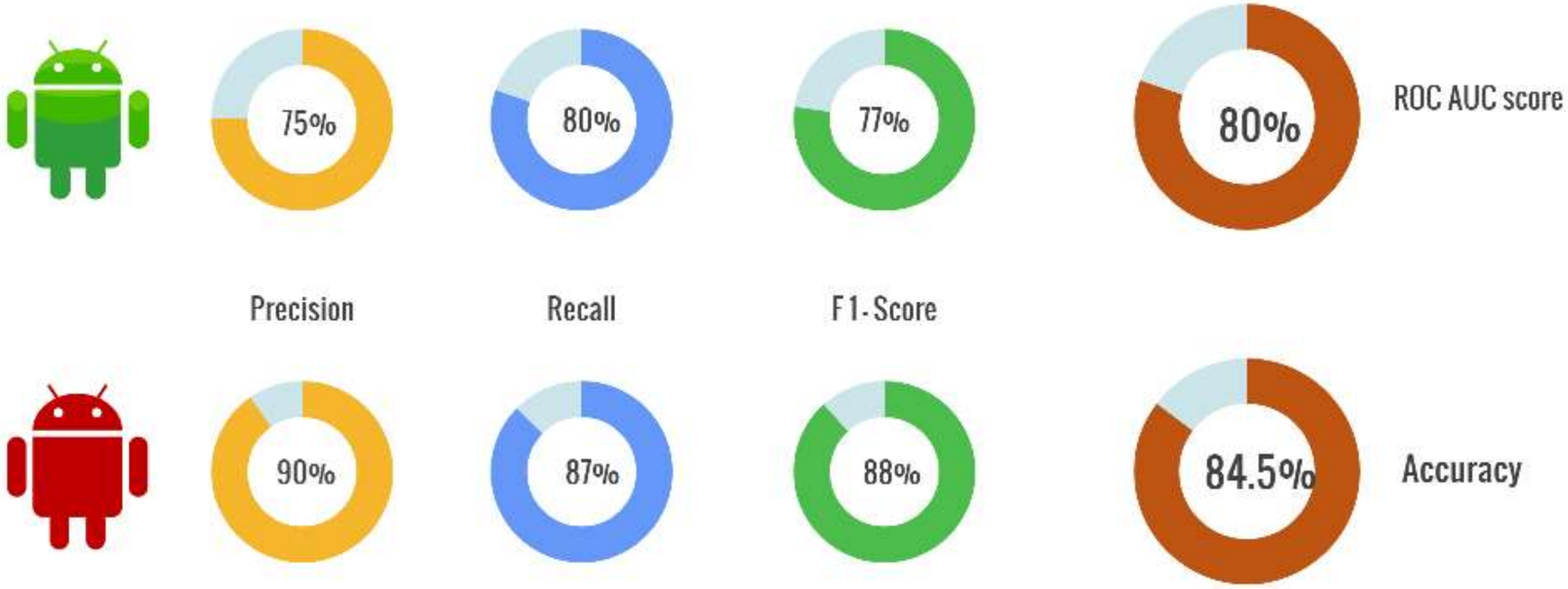
Model 2: Decision Tree Classifier



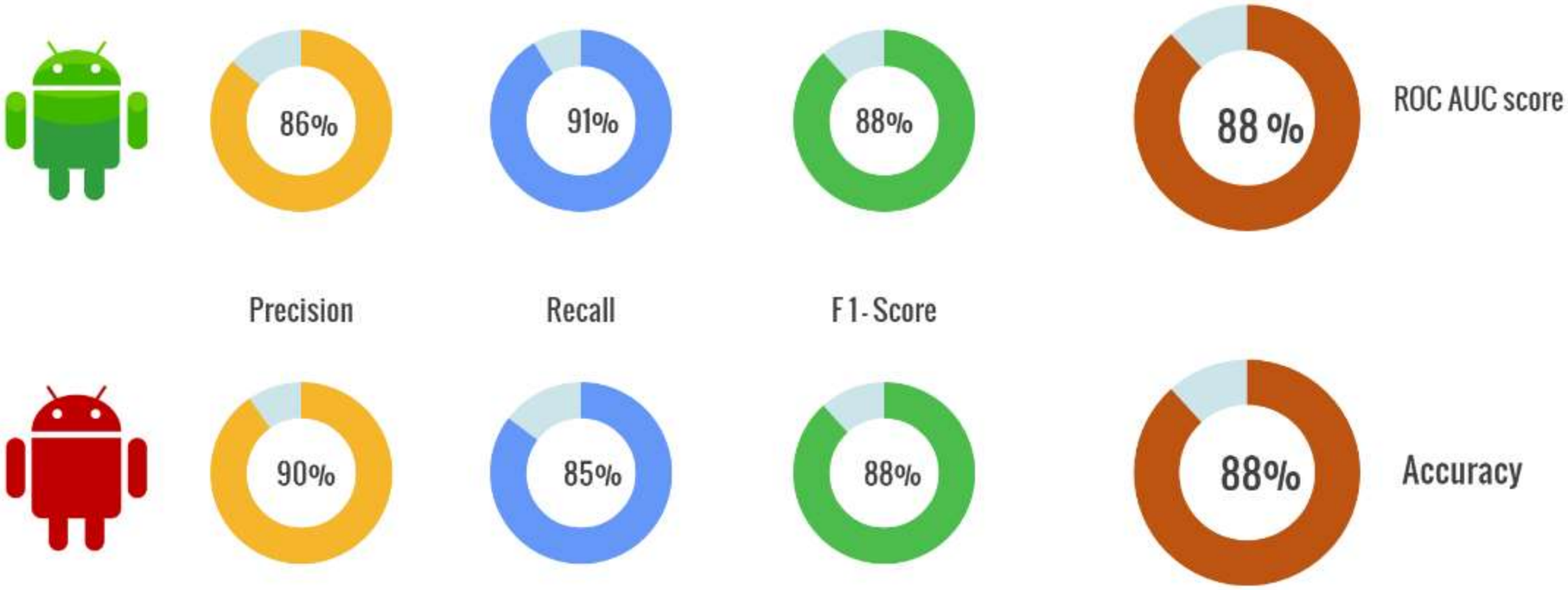
Model 3: KNN Classifier



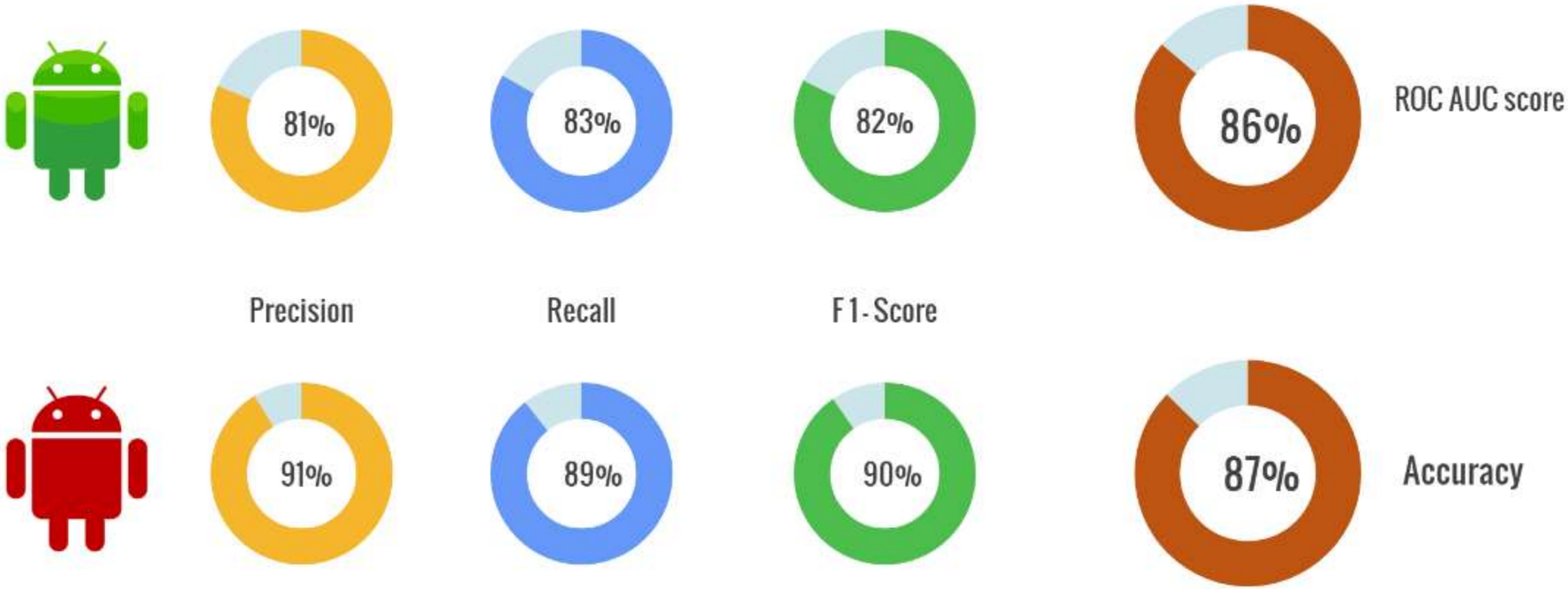
Model 4: Gradient Boosting Classifier



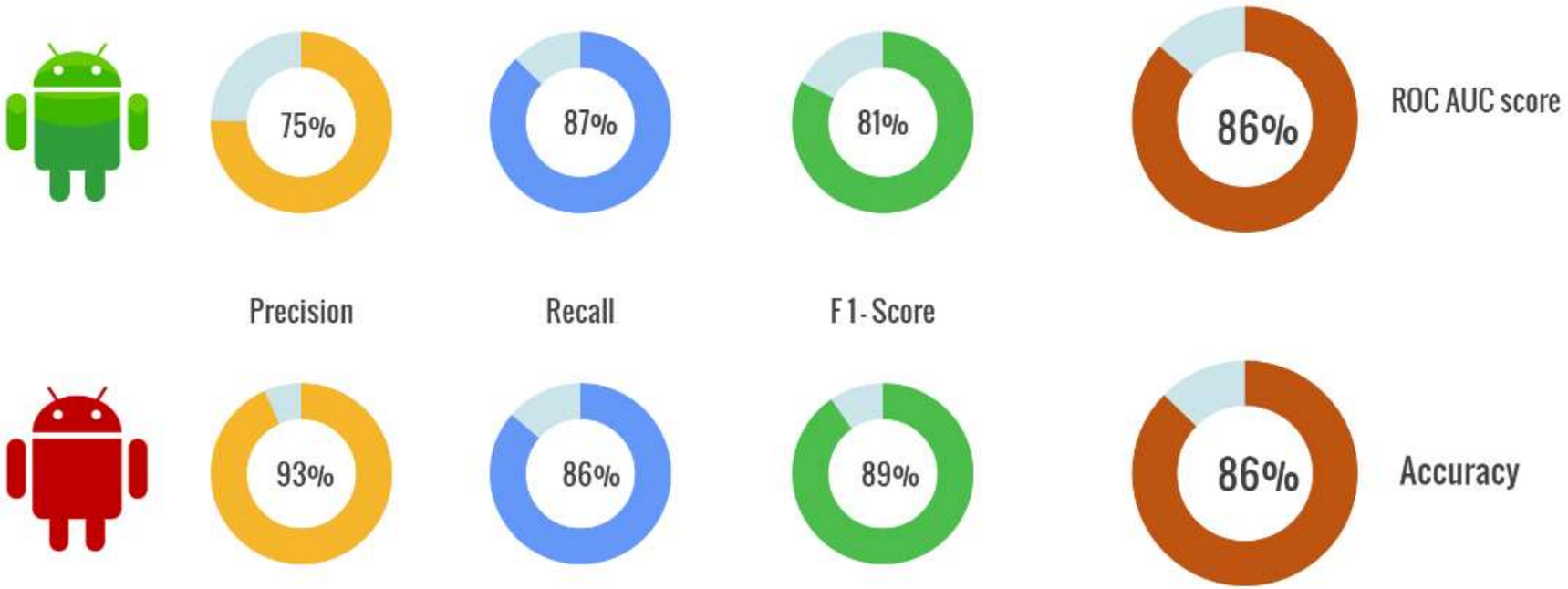
Model 5.1: Random Forest using Smote



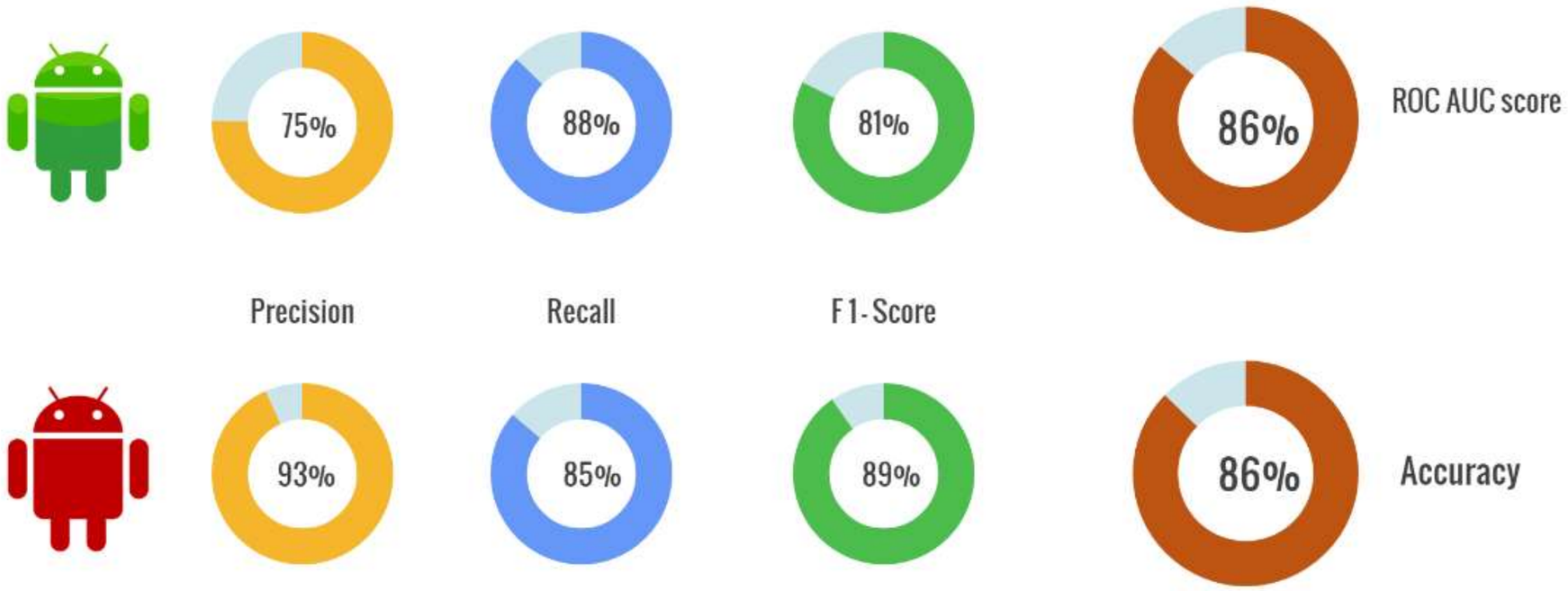
Model 5.2: Random Forest using Tomek Link



Model 6.1: XG boost using Smote

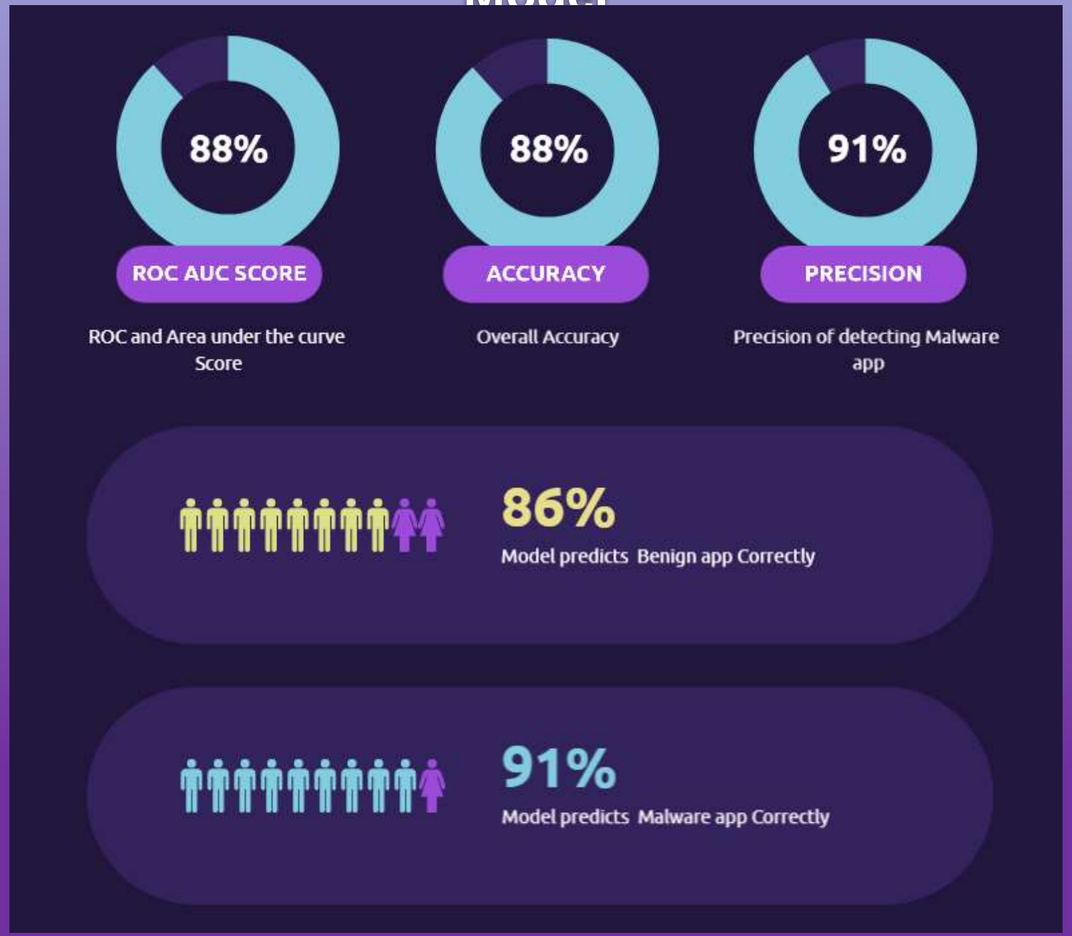


Model 6.2: XG boost using Tomek link



BEST MODEL

After apply various ML model, Random Forest Classifier using SMOTE comes out the best Model



Model Explainability

We use Feature Importance for model Explanability

index	Feature	Importance
1	Number of ratings	0.236409
0	Rating	0.124835
2	Price	0.108716
3	Dangerous permissions count	0.043639
206	Category_Travel & Local	0.038866
204	Category_Tools	0.027912
187	Category_Entertainment	0.023499
4	Safe permissions count	0.022575
202	Category_Sports	0.014302
184	Category_Comics	0.013401
190	Category_Libraries & Demo	0.012147
150	Your location : fine (GPS) location (D)	0.011999
180	Category_Brain & Puzzle	0.011675
91	Storage : modify/delete USB storage contents m...	0.011200
205	Category_Transportation	0.010857
198	Category_Productivity	0.010799
85	Network communication : view network state (S)	0.010625
195	Category_News & Magazines	0.010616
88	Phone calls : read phone state and identity (D)	0.010601
178	Category_Arcade & Action	0.010058

In Android Authenticity Prediction classification project, the feature importance analysis has identified that the **number of ratings** is the most important feature followed by **rating** and **price**.

These three features provide valuable information for predicting the authenticity of an Android app, and their importance highlights the importance of user ratings, popularity, and pricing in assessing the trustworthiness of an app.

8

Future Work

Future work

Future Work

Future directions for improving android authenticity prediction

1. **Using advanced machine learning techniques:**
Advanced machine learning techniques, perceptron, Data Mining can be applied to android authenticity prediction. These techniques can help capture complex patterns in the data that may be difficult to detect using traditional modelling approaches.

2. **Developing AI bases Real-time Approach**-AI-
based approaches of real time Deep learning models, can be used to identify and examine new malicious applications. These methods can be used to recognize malicious apps by analyzing the app's short history, browsing activity, and code structure. Further research sh ould also be conducted to minimize online fraud and malware attacks by developing robust AI-based solutions.

3. **Cloud Computing for Huge Data**-
Cloud computing can be used to store a large volume of Android data and detect malicious activities. So that we can get accurate pred iction of authenticity of the applications.

9

Conclusion

Final conclusion

Conclusion

Final Conclusion

The permissions that an application requests are an important factor to consider when determining its authenticity. Users should be wary of applications that request dangerous permissions, such as device admin, SMS and call, camera and microphone, and location permissions. Conversely, applications that request safe permissions such as internet access, network state and WiFi state, vibration, and wake lock permissions can generally be considered safe as long as they a re from a trusted source

In this project, our goal was to use classification techniques to solve the problem of detecting malware applications on Android phones. We experimented wit h different supervised classification techniques and identified the best technique for each approach. The results obtained from this project show that artificial intelligence can be used effectively to detect and prevent malicious Android app downloads. This authenticator can save Android users from the potential dan gers of downloading and installing malicious Android apps. By incorporating this AI-based authenticator into existing anti-malware safeguards, users can rest assured that their devices are protected from a wide variety of malicious threats. In addition, Random Forest and Gradient Boosting Classifier provide the great results in the approaches.

Android Authenticity Prediction ML model is a promising tool for detecting malicious Android applications. While it has some limitations, it has the potential t o be used in a variety of settings to improve mobile security. As the mobile threat landscape continues to evolve, it is important to continue to develop and im prove models like this to stay ahead of emerging threats.

Android Authenticity Prediction ML model is an effective tool for detecting malicious Android applications. It is highly accurate and scalable, making it an ideal solution for mobile security providers, app stores, and other organizations. However, there are also some challenges that need to be addressed when using th e model. We need to continue to develop and improve the model to keep up with the evolving threat landscape.

THANKS!

Any questions?

You can find me at:

- https://github.com/riyapatelrp