*Machine Learning Classifier for Mobile Malware Detection*

Dissertation Chapter # 4: Analysis & Result

Student Name: Harjinder

Table of Contents

[Chapter # 4 Results & Analysis 4](#_Toc95831930)

[Support Vector Machine Classifier, Random Forest Classifier, Decision Tree Classifier 4](#_Toc95831931)

[Programing Task 5](#_Toc95831932)

[Importing python libraries 5](#_Toc95831933)

[Loading dataset 5](#_Toc95831934)

[Data Cleaning 6](#_Toc95831935)

[Removing Null Values and Fixing data types for column 7](#_Toc95831936)

[Convert text columns to integer data type 7](#_Toc95831937)

[Stripping white spaces in columns 8](#_Toc95831938)

[Top 10 permission rating Bar graph 9](#_Toc95831939)

[Data Visualization 9](#_Toc95831940)

[Bar Plot 10](#_Toc95831941)

[Distribution plot of Rating of Malicious App 10](#_Toc95831942)

[Distribution of rating of Benign App 11](#_Toc95831943)

[Malicious Distribution of Malware App 11](#_Toc95831944)

[Android Version Malicious Ratio 12](#_Toc95831945)

[Distribution days of publish Histogram 12](#_Toc95831946)

[Machine learning python libraries 13](#_Toc95831947)

[Removing unnecessary columns from the dataset 14](#_Toc95831948)

[StandardScaler feature 14](#_Toc95831949)

[Training and testing features Model development 15](#_Toc95831950)

[Non Normalized Confusion Matrix 15](#_Toc95831951)

[Non Normalized Confusion Matrix 16](#_Toc95831952)

[Model Tuning 19](#_Toc95831953)

[Non Normalized Confusion Matrix Plot 20](#_Toc95831954)

[Outcome 21](#_Toc95831955)

[Confusion matrix without normalization 21](#_Toc95831956)

[Optimal classifier configuration 22](#_Toc95831957)

[Decision Tree classifier result 22](#_Toc95831958)

[Artificial Neural Network with SVM 24](#_Toc95831959)

[Importing APK libraries 25](#_Toc95831960)

[Importing System, OS Statistical 26](#_Toc95831961)

[Print Predicted validated Data value 26](#_Toc95831962)

[Support vector machine with Grid search to tune model 27](#_Toc95831963)

[EPOCH Call Kernel 27](#_Toc95831964)

[Classification Report 28](#_Toc95831965)

[Prediction 28](#_Toc95831966)

[Android Permission Data 29](#_Toc95831967)

[Multilayer Perceptron (Artificial neural network) 29](#_Toc95831968)

[Sequential RELU Metrics 30](#_Toc95831969)

[Android API APK 31](#_Toc95831970)

[Android malware prediction accuracy 31](#_Toc95831971)

[Android malware prediction accuracy 32](#_Toc95831972)

[Analysis 32](#_Toc95831973)

[Genetics Algorithm 33](#_Toc95831974)

[Importing Python Libraries in Notebook 34](#_Toc95831975)

[Importing dataset 34](#_Toc95831976)

[SVC Mean Squared ERROR 35](#_Toc95831977)

[Genetic Selector 35](#_Toc95831978)

[Score Generation 36](#_Toc95831979)

[Mean square ERROR after feature selection 36](#_Toc95831980)

[Import Pickle library 37](#_Toc95831981)

[Trainings & Testing dataset 37](#_Toc95831982)

[ANN Sequential 37](#_Toc95831983)

[EPOCH 38](#_Toc95831984)

[Andro Guard 38](#_Toc95831985)

[SVC 39](#_Toc95831986)

[Classification Report 39](#_Toc95831987)

[Results & Analysis 39](#_Toc95831988)

[References 44](#_Toc95831989)

# Chapter # 4 Results & Analysis

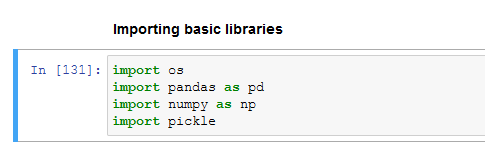
## Support Vector Machine Classifier, Random Forest Classifier, Decision Tree Classifier

Highly efficient random forest based malware detection (Cheng, L., 2018) framework for android phone, in these days android phone is rapidly the growing industry of telephonic network, various malware attacked by the android phone in certain period of time to capture the sensitive data, mobile phone posing to increasing the threats including various security threats. Google’s android is the most popular smart phone platform in the world of permission mechanism, this research propose machine learning classification technique, feature group selection including permission rate mined to characterize each android application and the most essential component of random forest classifier actively works to detect the app either it contain with malicious or not. The proposed performance method is evaluated on the actual data set using tenfold cross validation and the experiment result presented that which present the high accuracy rate formulated on 89.91% rate performance accuracy analysis of malicious android app dataset. (Hofmann, M., 2017) android is most popular operating system with share of 86% of mobile phone user, its open source operating system to support various third party application to executes within the android phone device. Google provides security features that built in under the android operating system releases but the issues is still exist by obtaining malware from the outsources by redirect the user to third party link to open malware applications by using the adware and other malware threats exists openly in android device, however the static and dynamic analysis of machine learning analysis provide guard against such analysis to secure the android application by scanning the machine learning feature selection to best prediction and remove malware from the app. The study conduct to scan the 216 malicious app to construct feature vector of training and testing by selection various feature selection, the system call behavior detected the 216 malicious app and 278 normal app to develop the feature vector for training the classifier. Machine learning data classification techniques has been used with various method to predict and detect the malware apps by scanning code of APK. Including decision tree, random forest, gradient boost, support vector KNN, artificial neural network including deep learning approach has been applied on dataset. There are three feature ranking technique has been identified by selecting the suitable feature selection set of 337 attributes of dataset system calls, the techniques ranks the various feature selection method based on chi square static and correlation analysis by determining weights of feature, the feature selection method founded on high & low rank to predict the performance analysis of android application behavior which consist of malicious code and finally the experiment result present with various machine learning method support vector machine classifier through selecting feature through the correlation analysis of including accuracy of 97% and achieve the recall performance of 98% by identifying the malicious apps, the contribution of system call identifying set of performance by predicting the malicious apps of android phone.

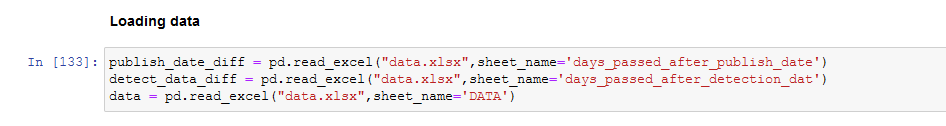
## Programing Task

Python libraries to validate the basic operation

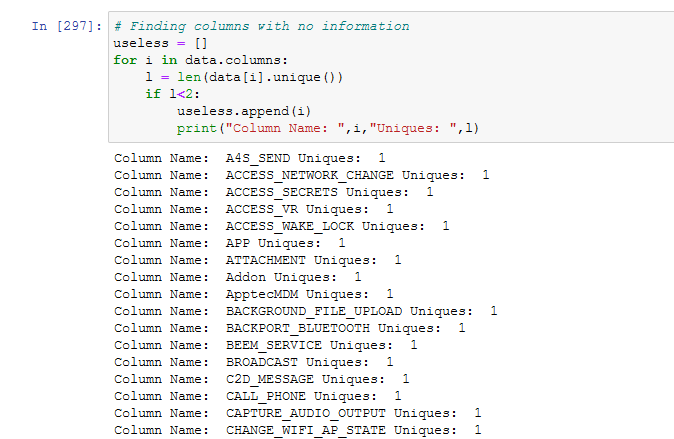
### Importing python libraries



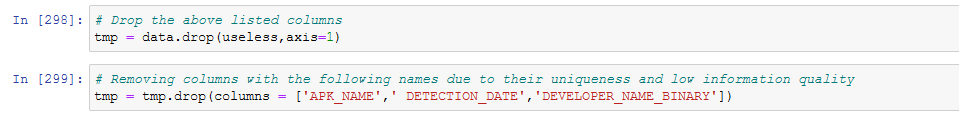
### Loading dataset



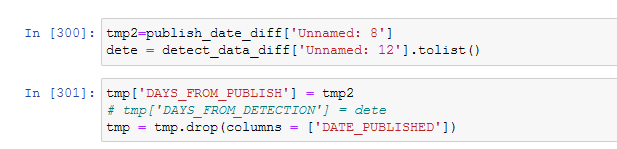
### Data Cleaning



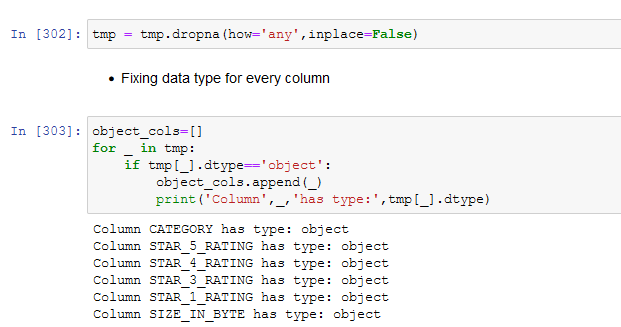
Drop column with following names due to their uniqueness and low information quality



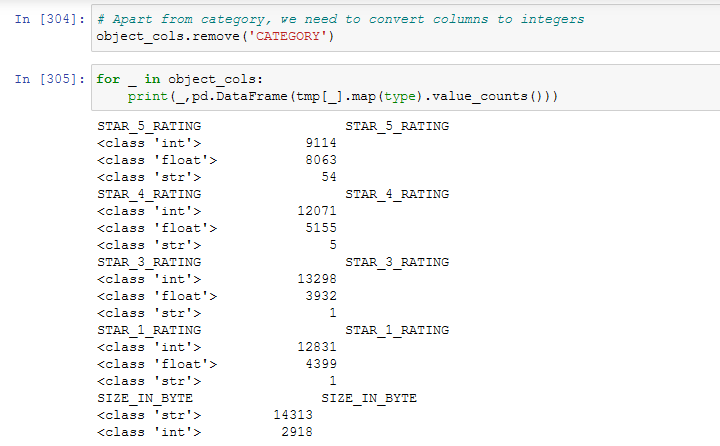
Calculating the days Elapsed since uploading on play store and days since first detection result has been made.



### Removing Null Values and Fixing data types for column



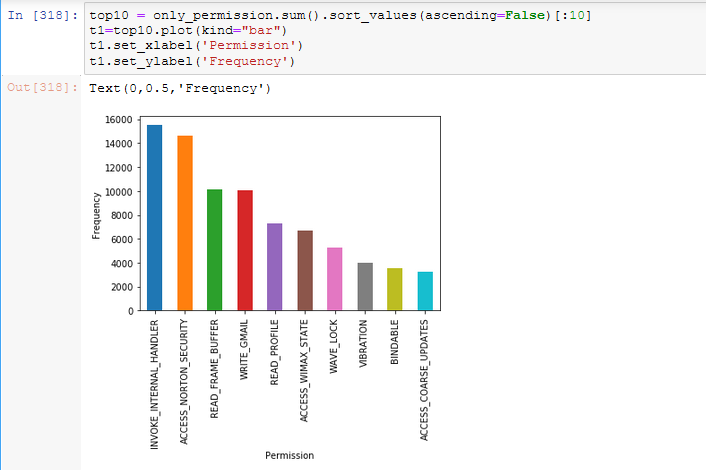
### Convert text columns to integer data type



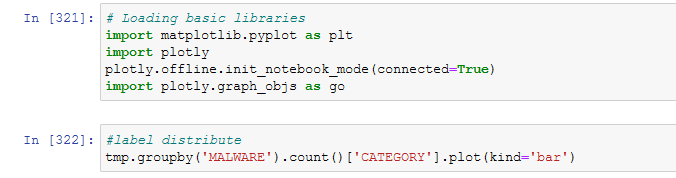
### Stripping white spaces in columns



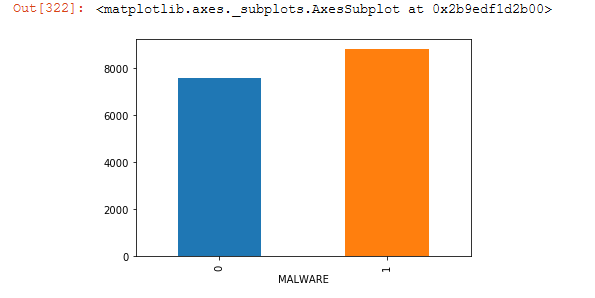
### Top 10 permission rating Bar graph



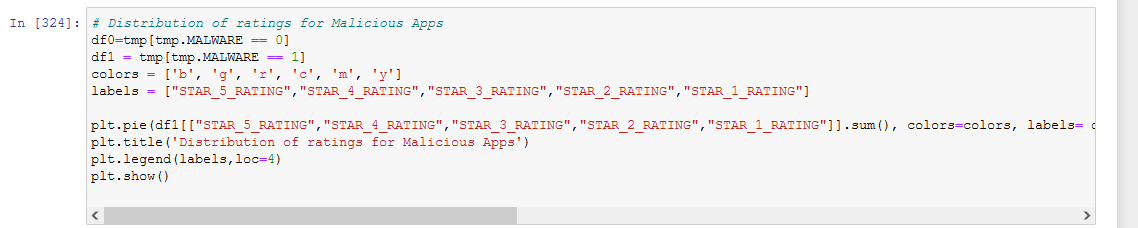
### Data Visualization

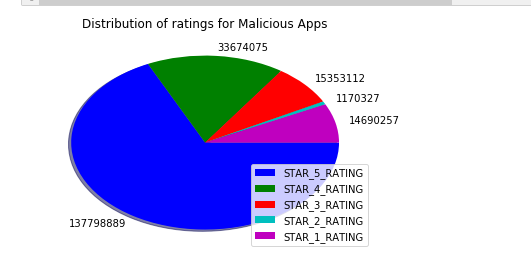


### Bar Plot

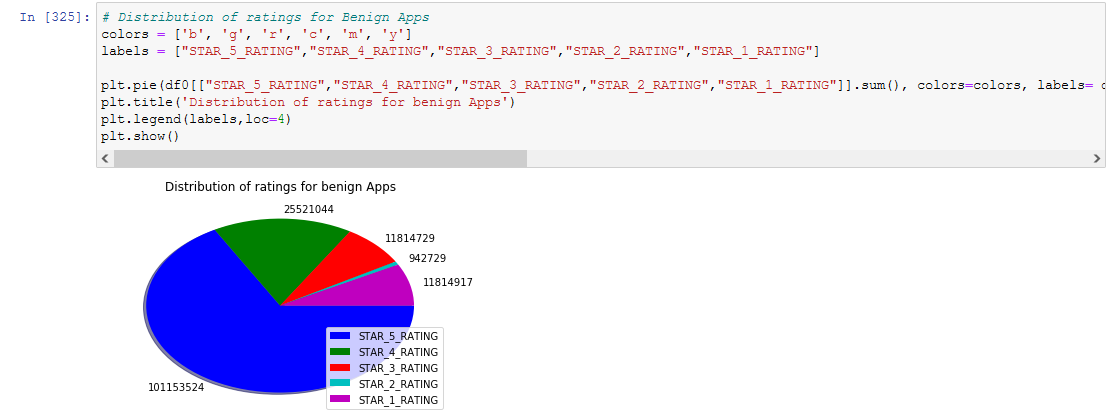


### Distribution plot of Rating of Malicious App

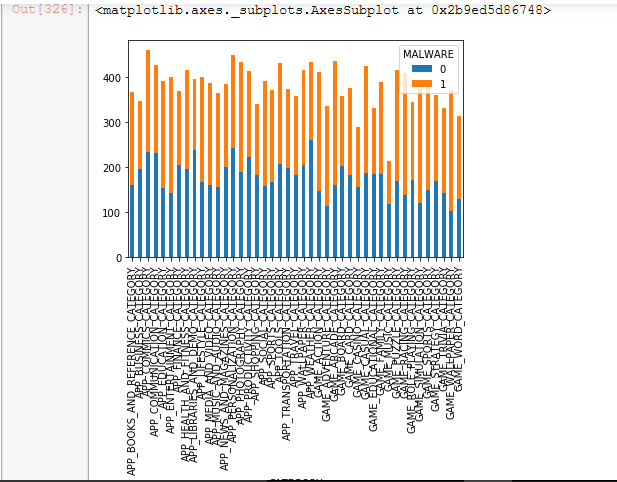




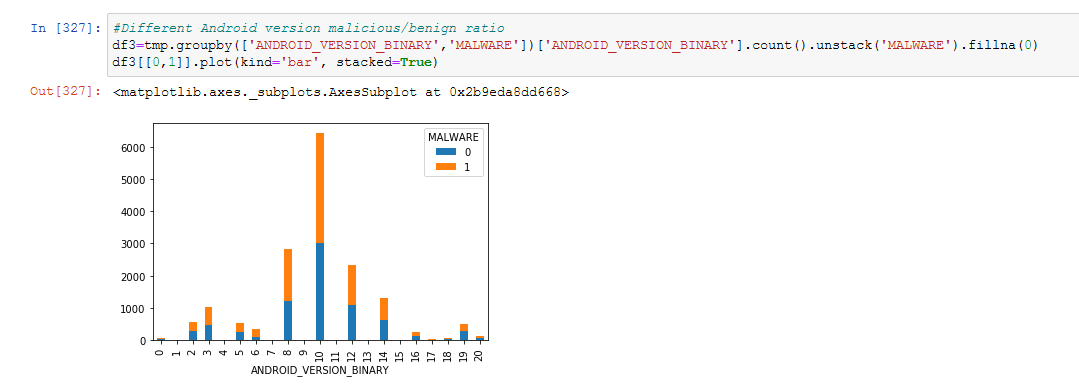
### Distribution of rating of Benign App



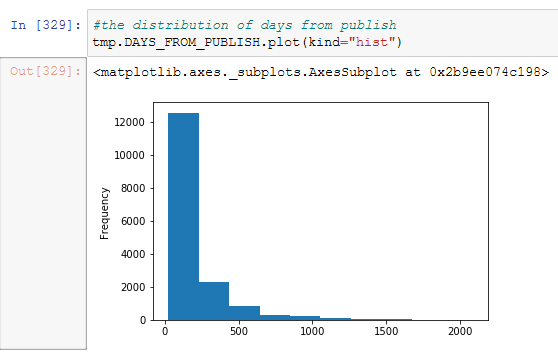
### Malicious Distribution of Malware App



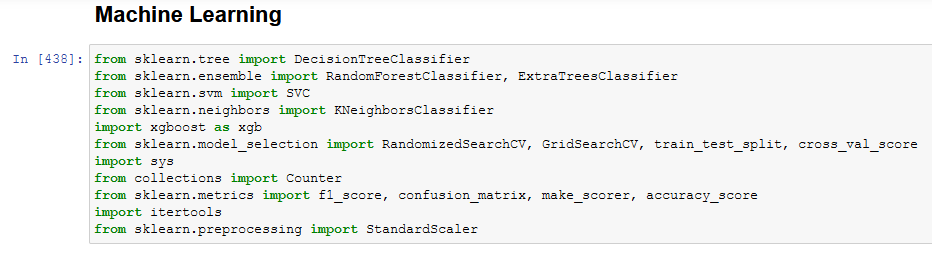
### Android Version Malicious Ratio

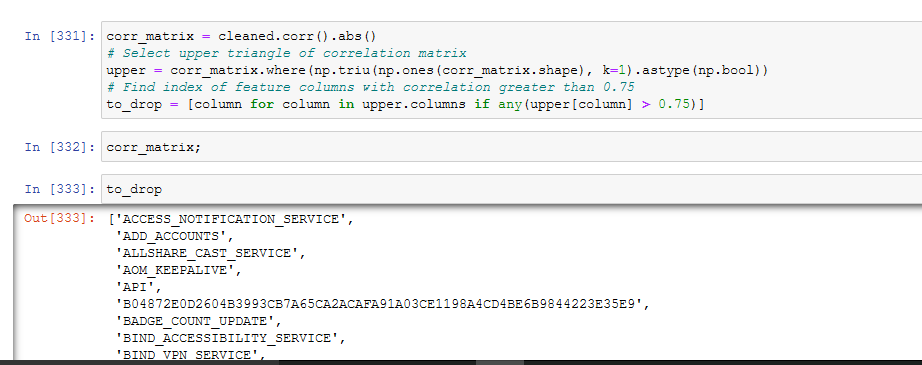


### Distribution days of publish Histogram

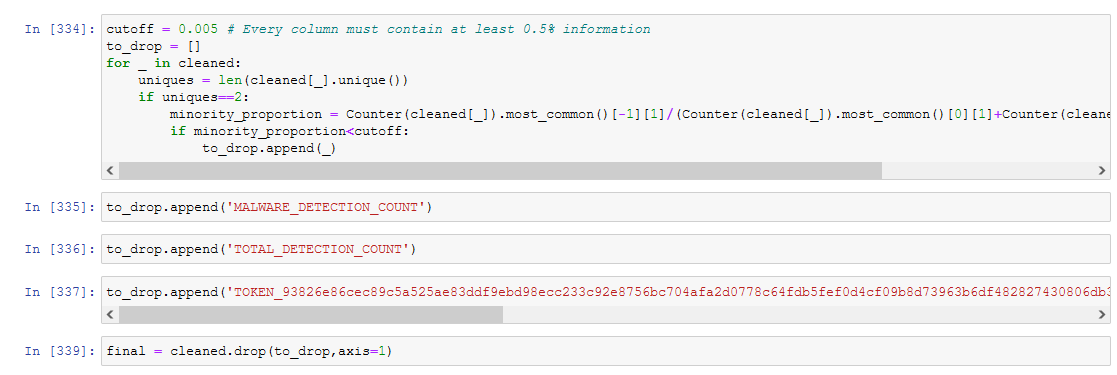


### Machine learning python libraries

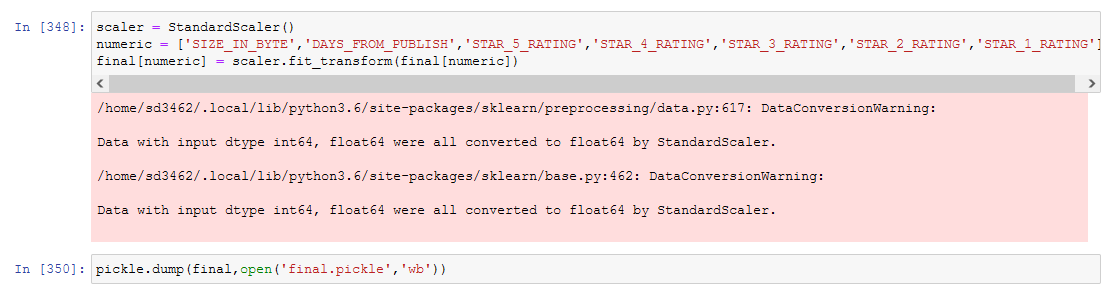
Removing highly correlated features decrease redundancy of features



### Removing unnecessary columns from the dataset

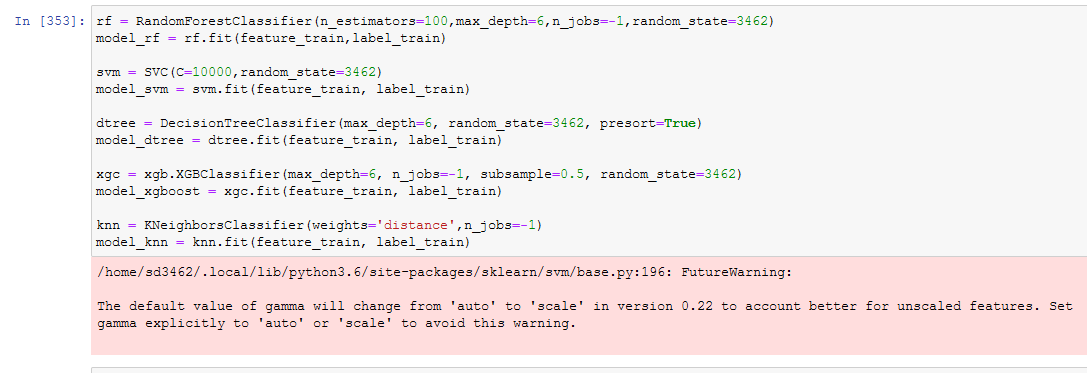


### StandardScaler feature

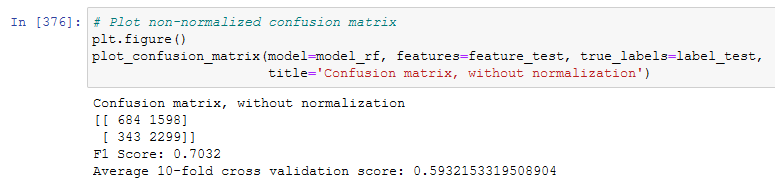


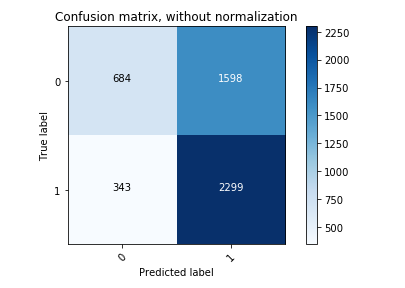
### Training and testing features Model development



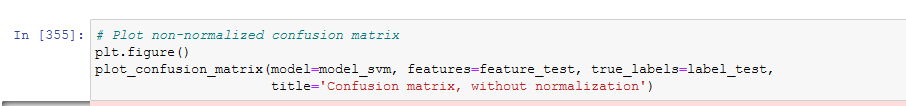


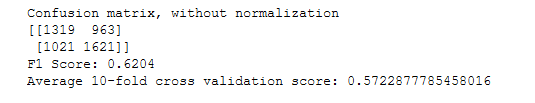
### Non Normalized Confusion Matrix

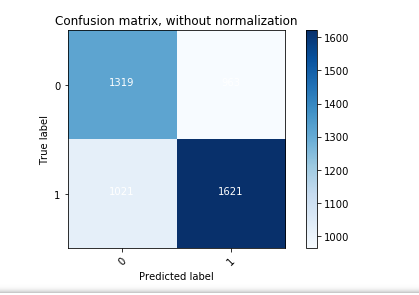


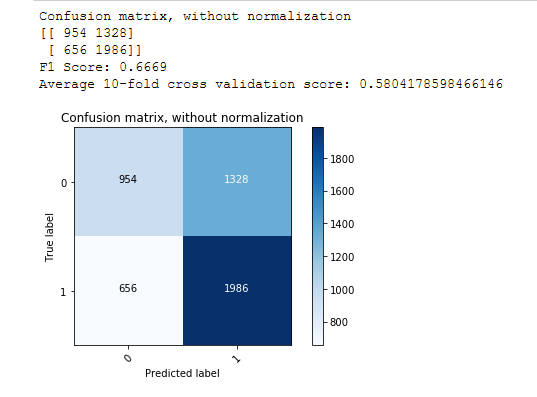


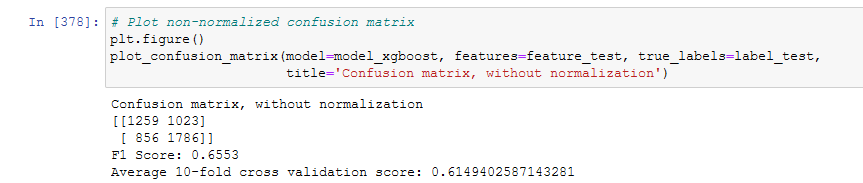
### Non Normalized Confusion Matrix

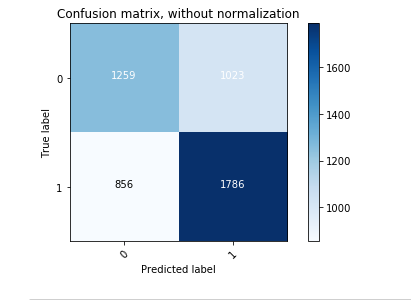


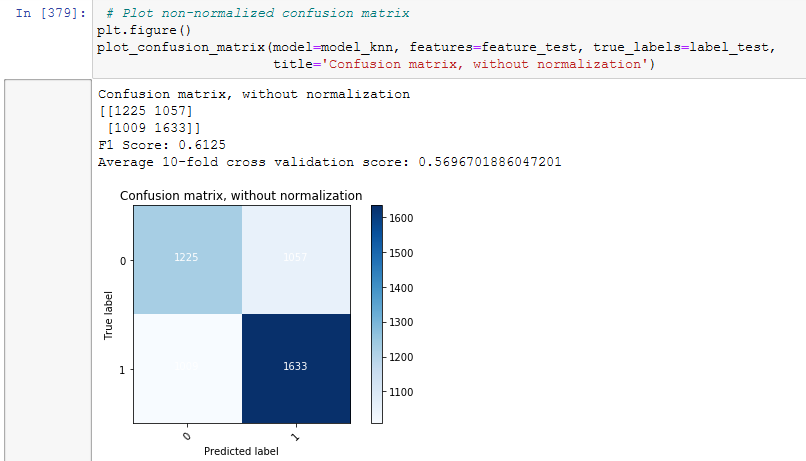






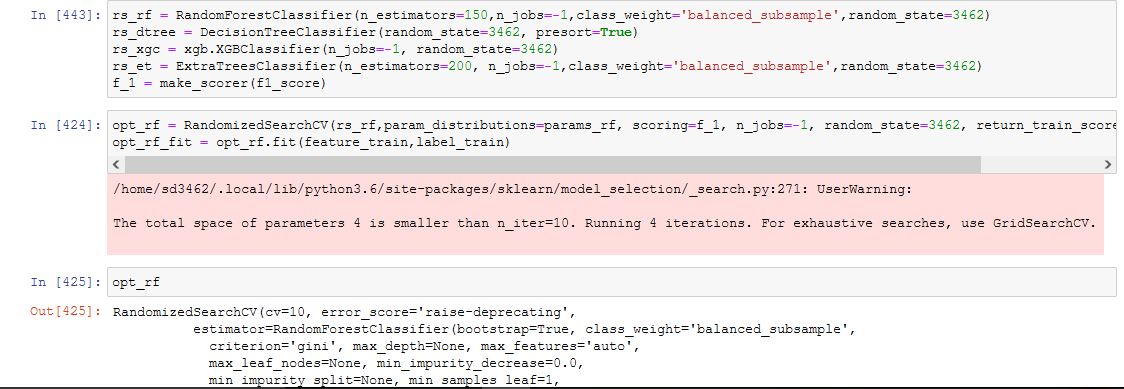
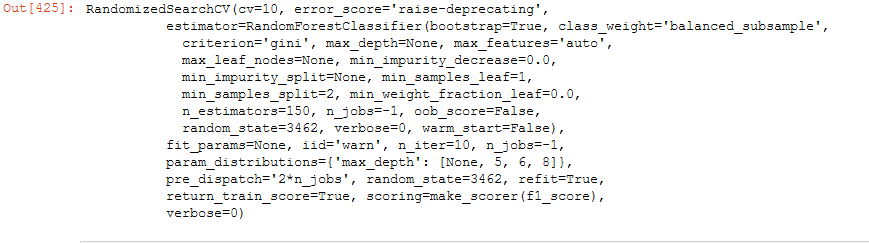




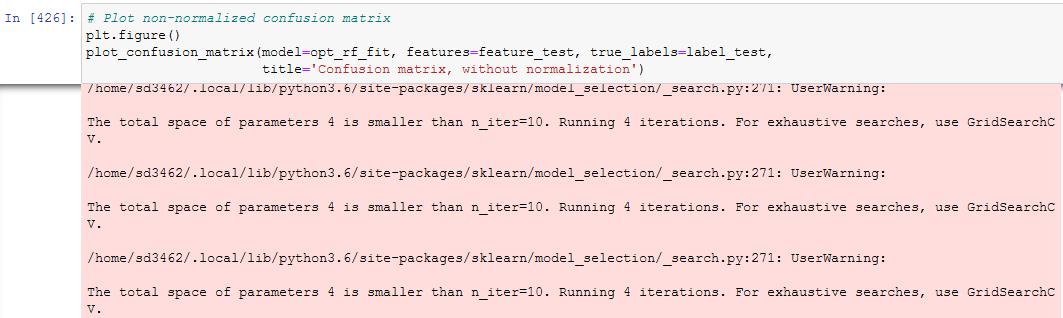


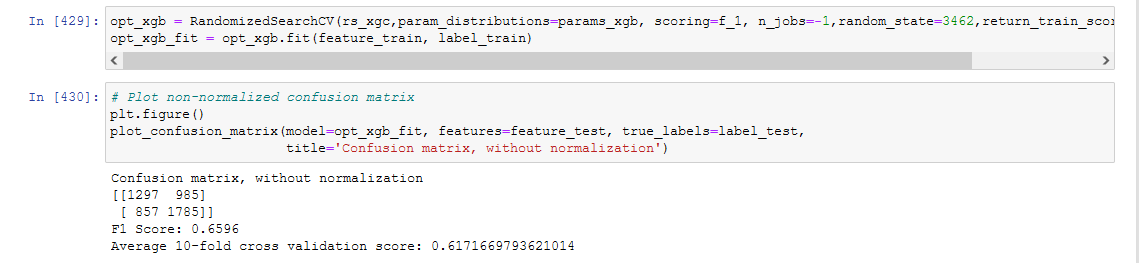
### Model Tuning

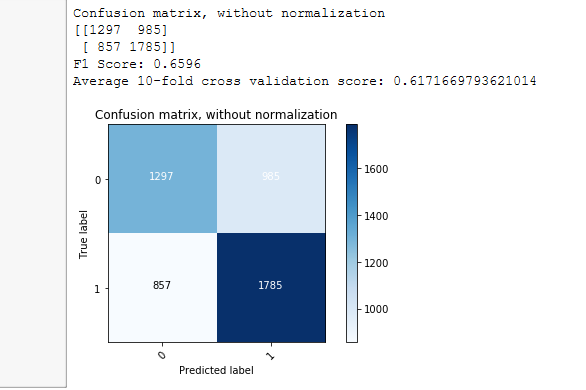


### Non Normalized Confusion Matrix Plot

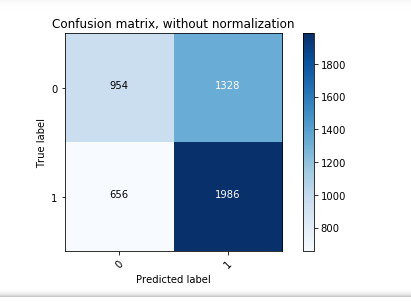






### Outcome

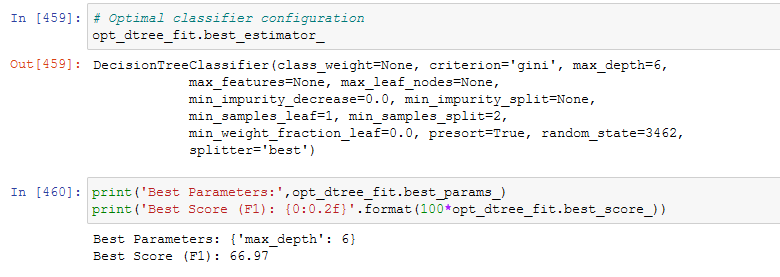
### Confusion matrix without normalization



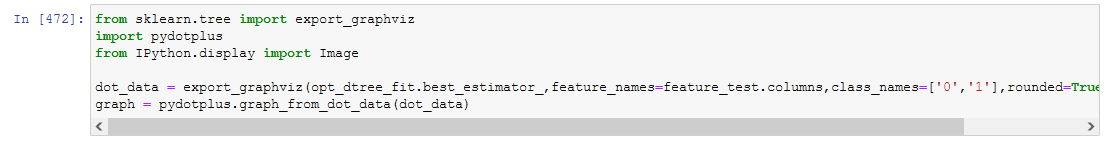
### 

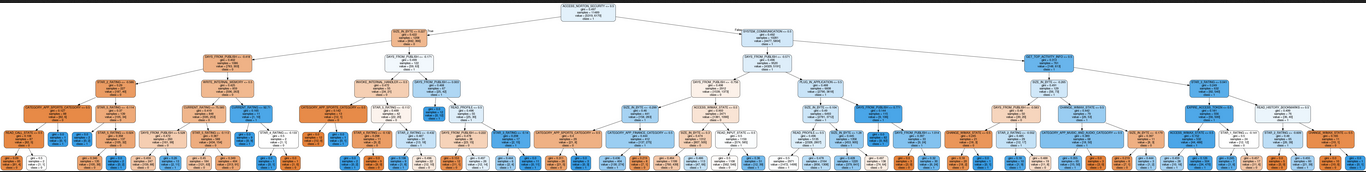
The experiential assessment numerous machine learning replicas aided us accomplish that the decision tree performs the best on the data. The evaluation result of metric score is presented with F1 score which takes into account the precision and the recall of every classifier and while it is noted that the detect true positive which is important to reduce false positive and false negatives. The false positive result might be cause inconvenience to the user while false negatives would be unsafe to the user safety and privacy, (Ali, W., 2019) hybrid intelligent android malware detection method using support vector machine classifier which proposed on genetic algorithm, the optimization problem in support vector machine is solving using the genetic algorithm technique with practical swarm optimization referred to DROID-HESVMPSO approach to achieve the best detection result obtaining by using the machine learning techniques including hybrid malware detection techniques approach. When the feature set was constructed the classified model were trained & validated feature ranking & selection technique has been identified on performance and after application comparison.

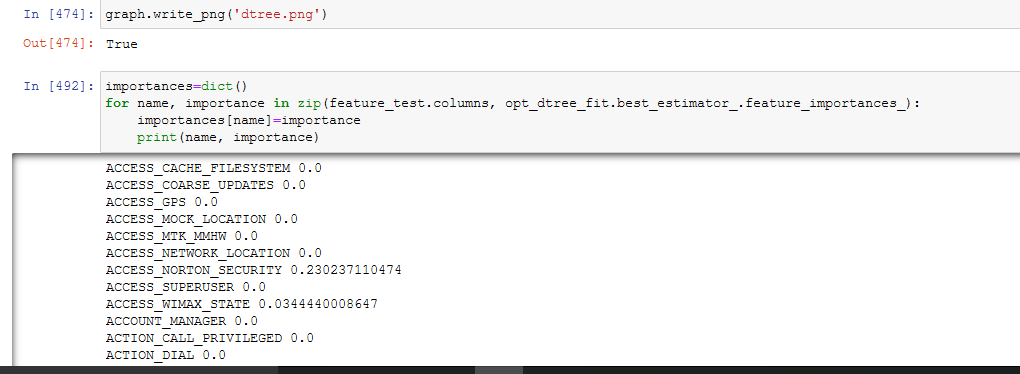
### Optimal classifier configuration

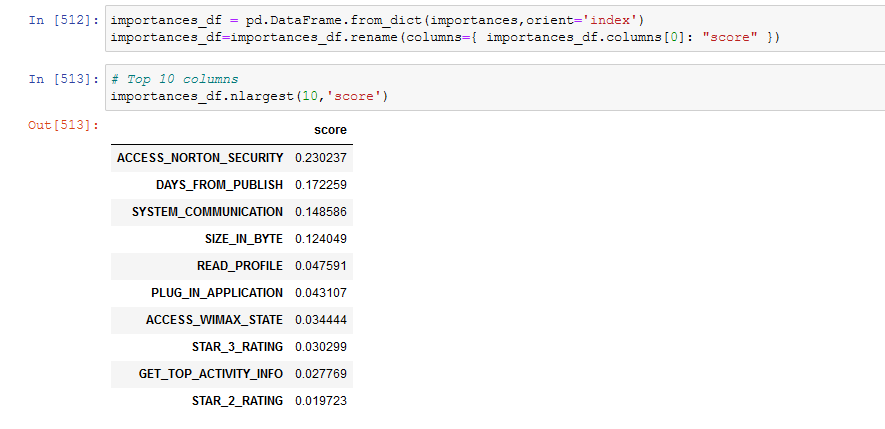


### Decision Tree classifier result



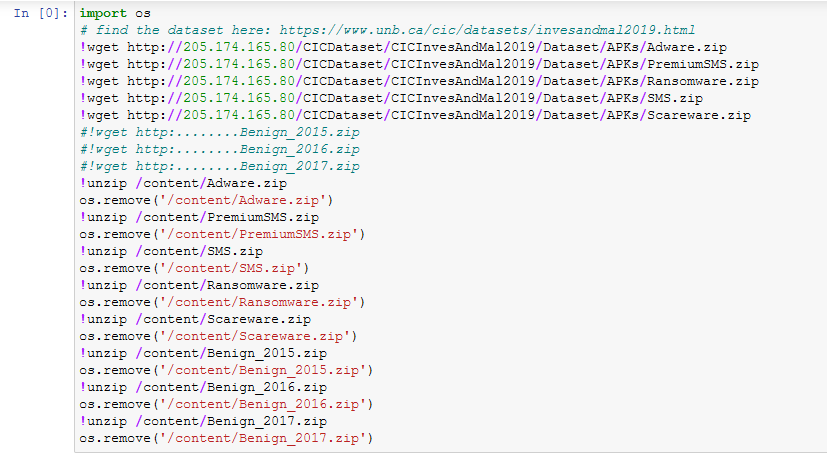






### Artificial Neural Network with SVM

#### Dataset



#### Importing Libraries

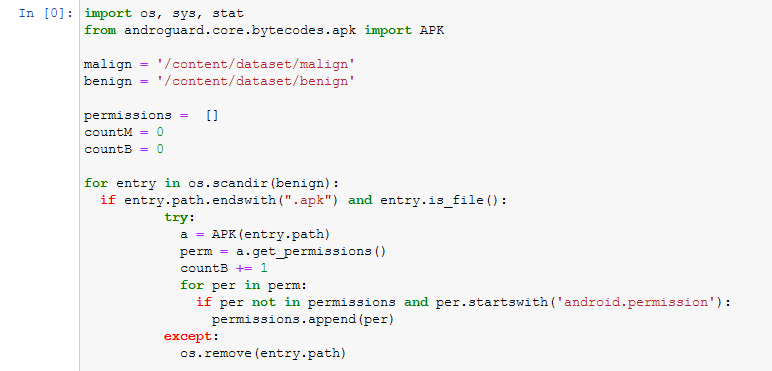




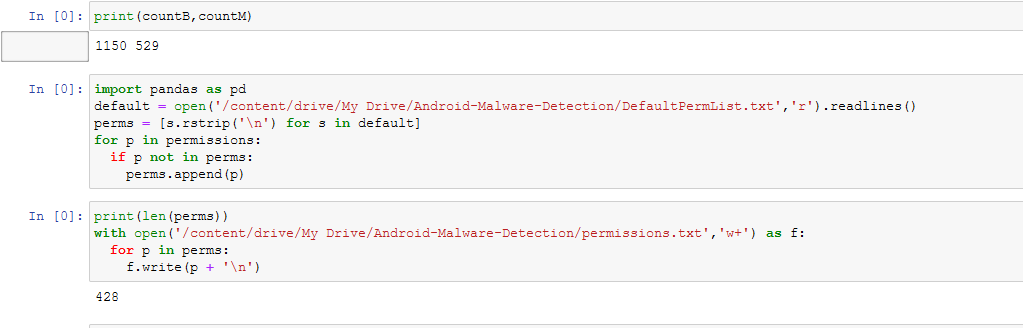
### Importing APK libraries



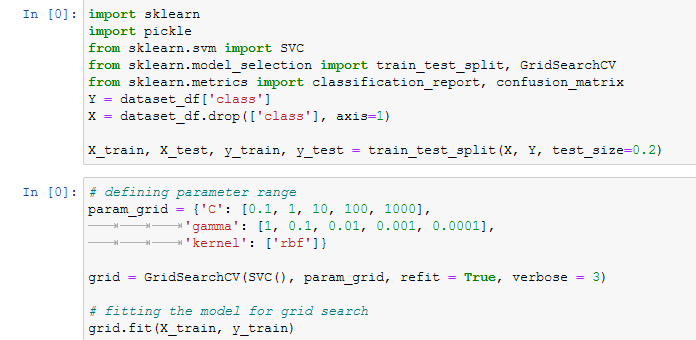
### Importing System, OS Statistical



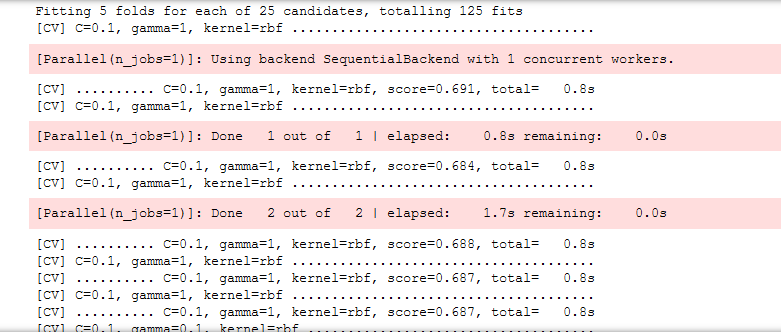
### Print Predicted validated Data value



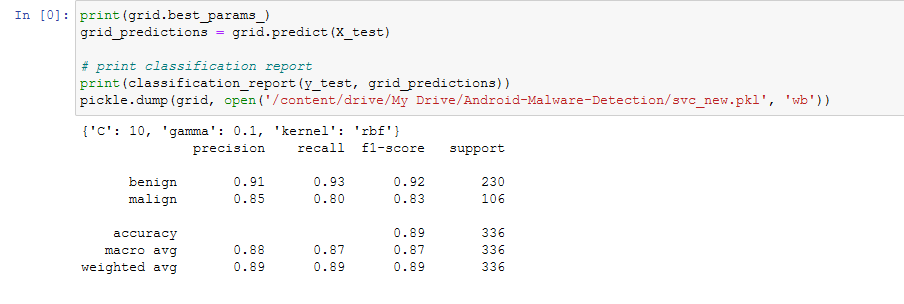
### Support vector machine with Grid search to tune model



### EPOCH Call Kernel



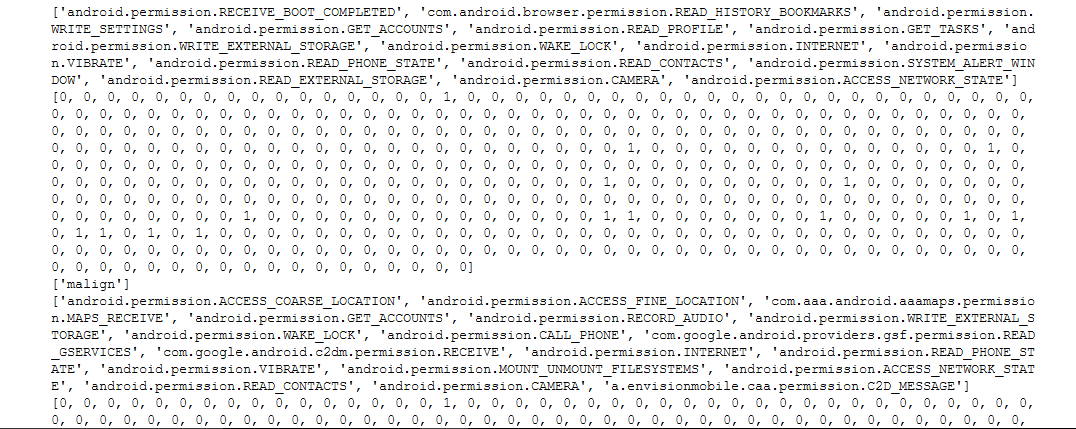
### Classification Report



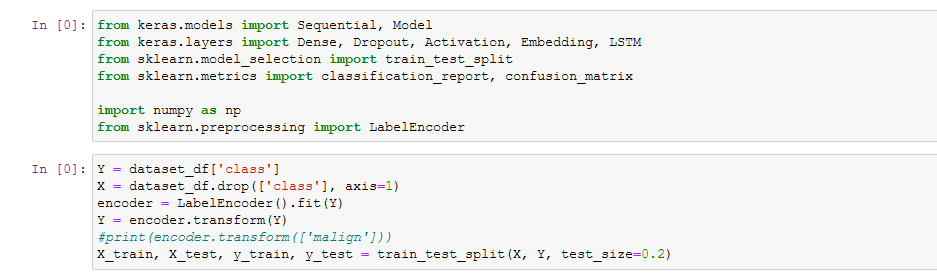
### Prediction



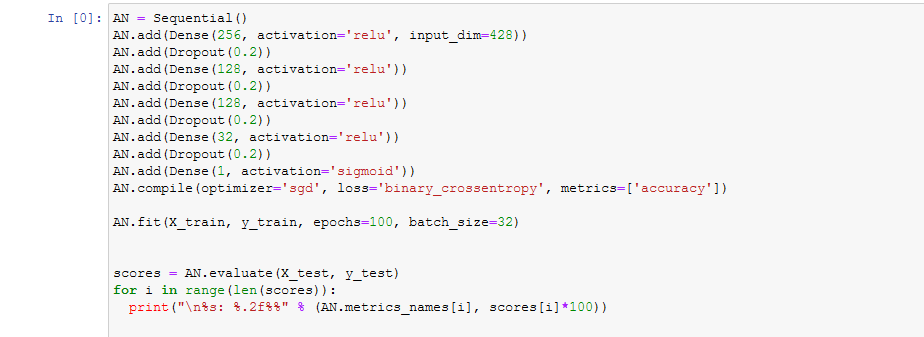
### Android Permission Data

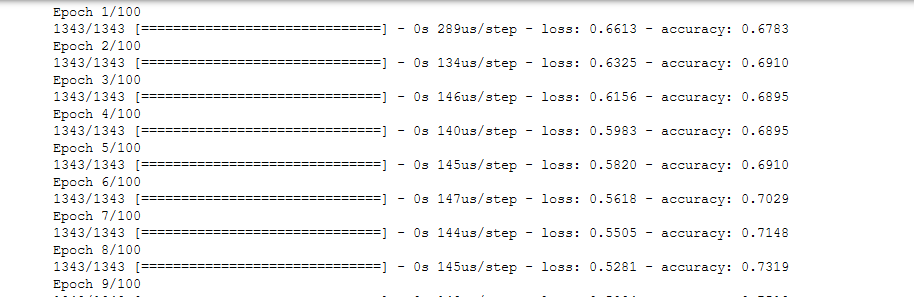


### Multilayer Perceptron (Artificial neural network)



### Sequential RELU Metrics



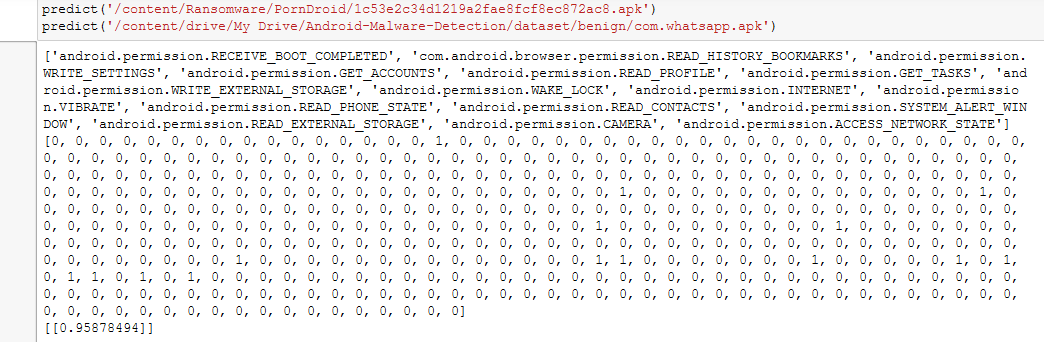


During the data preprocessing the steps involved by scanning the feature selection to identifying the attribute in zero variance which was removed. The feature corresponds to the system call which were never invoked by set of samples to proceed the with 337 system calls and 43 attributes. The classification report constructed formulated on training dataset based on input and output pairs to analyze the algorithms which learn the mapping functions, the gradient boosting tree by training and testing the random data and growing the weighted version of dataset by obtaining the correlated regions which formulated on different values. SVM technique based on statistical theory the linear SVM techniques identifying the hyperplanes with maximum margins of two separate value of classes.

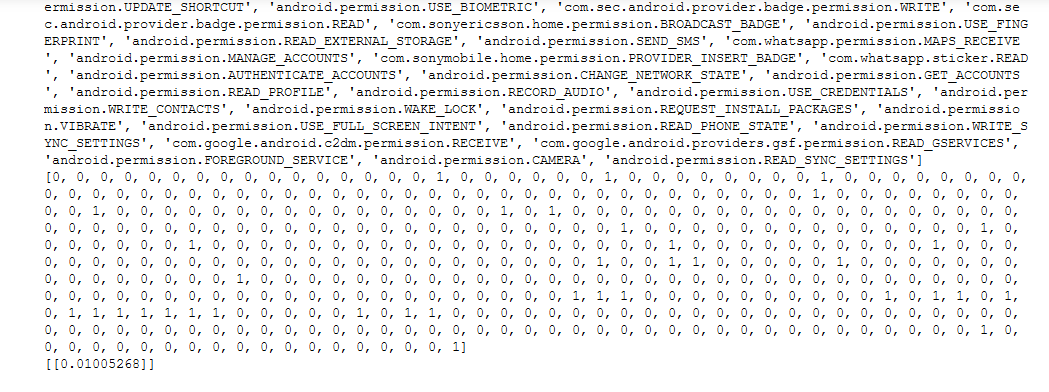
### Android API APK



### Android malware prediction accuracy



### Android malware prediction accuracy



## Analysis

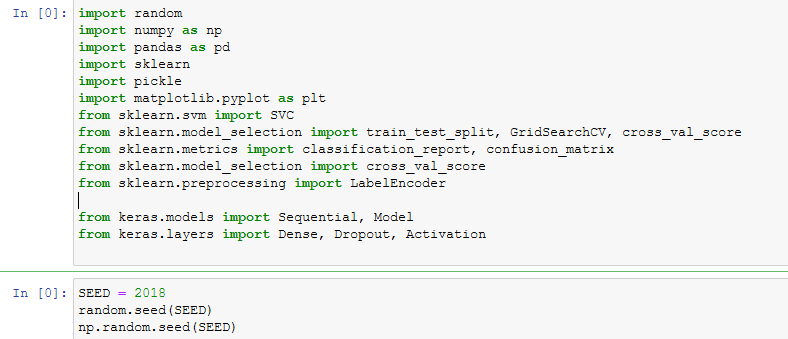
Mobile malware detection using the machine learning classifier (Lukman, S., 2017) in regards to latest technological platform, smart phone and other mobile devices facing issues due to malware, it caused problematic situation over the internet web. Android mobile platform using the various Google operating system API perform various mobile malware detection & prevention including various computing platform. Dataset consist of both benign and malicious applications were installed on android device to analyze the various pattern of behavioral network rules. This research generate the system metrics feature vector from each app by executing the controlled environment of machine learning operation in order to achieve the accuracy of malicious malware detection from the security dataset of android phone. Various machine learning algorithm has been proposed and used in python notebook in order to achieve the great efficiency of machine learning detection accuracy Decision tree, KNN, SVM, Multilayer perceptron neural network with naïve bays classifier random forest classify the app as benign or malware, individual algorithm evaluate on different criteria of prediction and detection of machine learning accuracy, the resultant outcomes of research evaluate that the random forest and support vector machine classifier predict the best accuracy outcomes including effective techniques of malware detection. (Sung, A.H., 2018) evaluated machine learning models for android malware detection, popular android mobile billions of active users, developer deploy malware for various purposes, Google play store providing the security features to Google app by user ID based security tool in which the hacker identification easy to prone while capturing the sensitive data from the android phone. The system permission applications needs various security features from different user identification based configurations, machine learning static classification method has been implemented to test the android malware dataset, android operating system scans various system application while installing the new app, this research present the 5560 malware samples including android permission APK dataset with 5560 Benign samples of DERBIN dataset to validate and predict the accuracy performance including F1 score and precision call, with 94% accuracy achieved (J. Lin, X., 2015) effective behavior based android malware detection system, rapid growing of android application become challenging issues with research to avoid malware from the android phone. The behavioral analysis predict the android malware detection by using SVM and decision tree classification technique the promising approach predict the accuracy and resilience to malware alternatives the android API with Google interface to scan the mobile applications to predict the sensitive application behavior. Results of this research based on machine learning technique which would evaluate the different ML algorithm such as Support vector machine, random forest naive bays, decision tree and genetic algorithm classifier technique, experiment conducted on 1136 real world samples composed of various malware samples including malware and benign application, the evaluation results present that the behavior system of android application used to predict the malware detection composed of application behavior method. (Verma, H.K., 2020) static feature selection based method including effective machine learning method effective large scale detection method dataset extracted from online API tool and unzip in python notebook and predict the security features of APK application to validate the outcomes with precision recall and F1 score prediction of machine learning principle component analysis used the feature selection approach by training and testing the dataset and building the random forest and support vector and decision tree classifier to validate the result by achieving the dataset trained and tested approach and getting the random forest classifier accuracy with 96.5% classification accuracy including classification report of confusion matrix.

## Genetics Algorithm

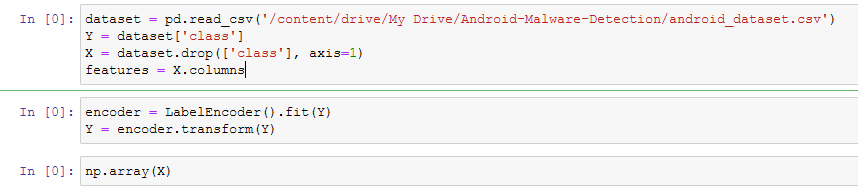
Using python notebook by importing various machine learning libraries and perform and malware detection by using the Github and Kaggle repository dataset. (Doğru, I.A., 2019.) Permission based android malware detection system using the feature selection of Genetic algorithm since the android is Linux based open source Google operating system, android developer mostly target android devices to capture the sensitive data, machine learning method rapidly used in android malware analysis by various genetic based techniques by combining the support vector machine classifier to other random forest and decision tree classifier and get the highest prediction accuracy of 98% with 16 selected permission using the dataset sample of 1740 consisting of 1119 malware and 621 benign samples.

## Programing Task

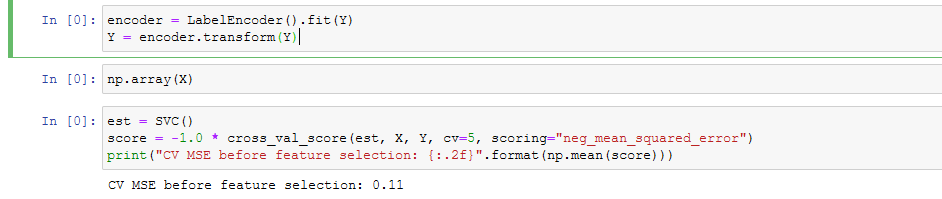
### Importing Python Libraries in Notebook



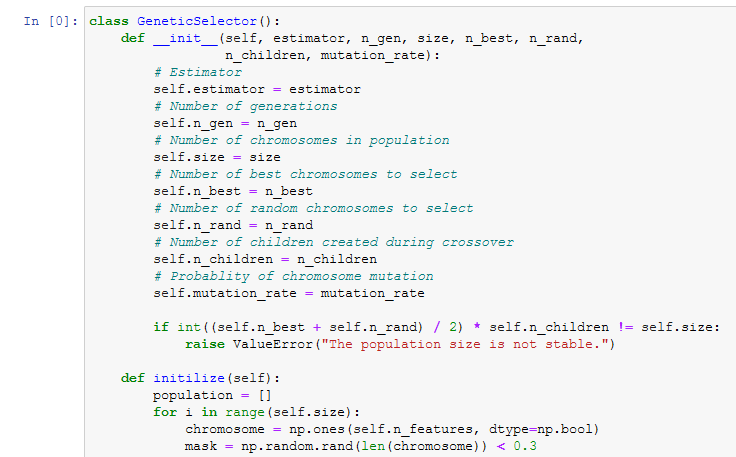
### Importing dataset



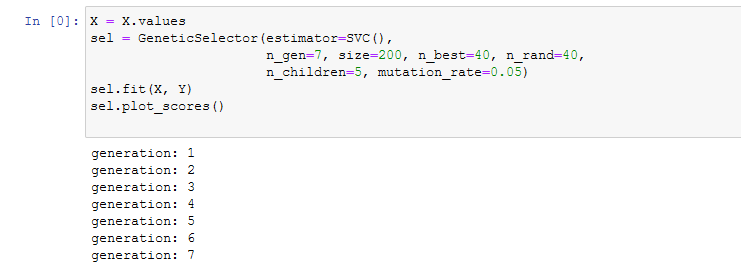
### SVC Mean Squared ERROR

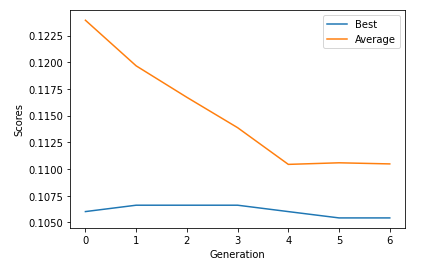


### Genetic Selector

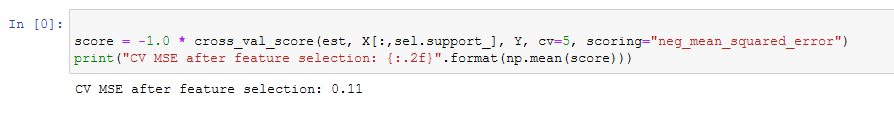


### Score Generation

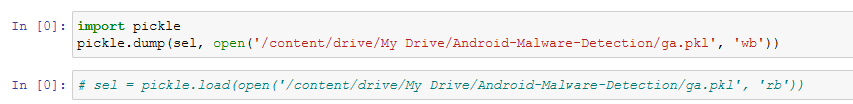




### Mean square ERROR after feature selection



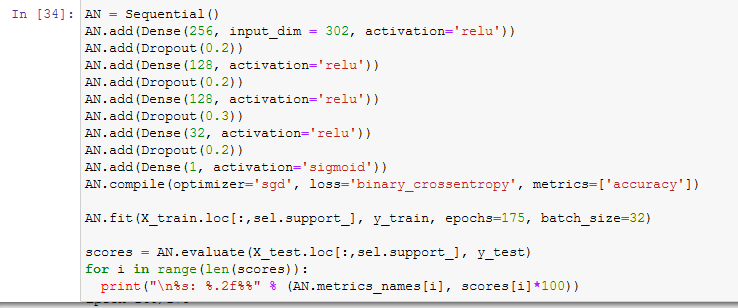
### Import Pickle library



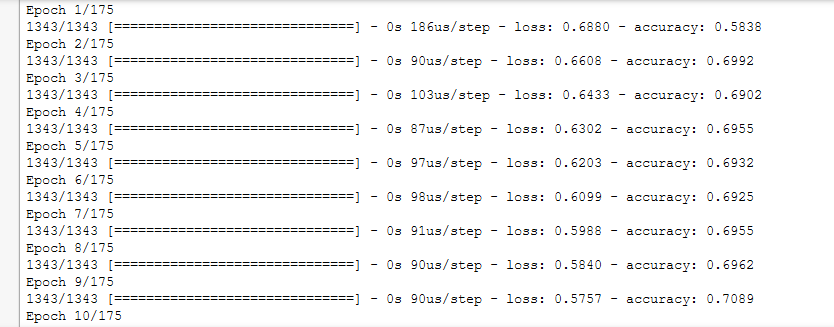
### Trainings & Testing dataset



### ANN Sequential



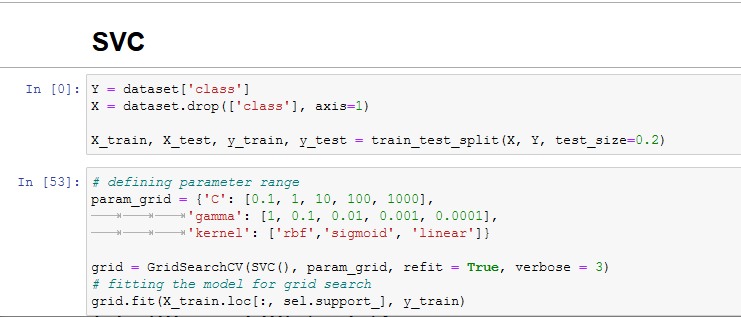
### EPOCH



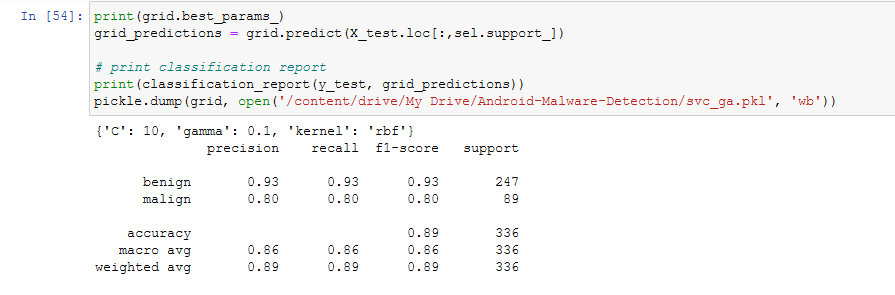
### Andro Guard



### SVC

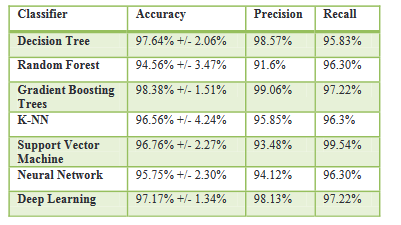


### Classification Report



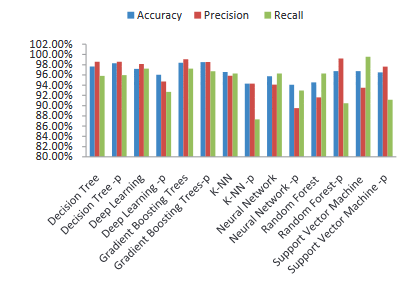
## Results & Analysis

Decision tree is classification algorithm which classify the various object of attribute to validate the different usages of malware detection and prevention. The decision tree classifier internally classify the node with edge to edge connected nodes based on the dataset attribute which validated the internal nodes which classify the internal label (Darmawan, M.F., 2021) android malware family detection using the classification method of machine learning algorithm various malicious software application of android identified & predicted formulated on the dataset value the resultant feature selection approach used to identify the different attribute with false positive and false negative value the ability to identify the malware from the android device which prevent on different classification techniques. The machine learning and deep learning classifier support vector machine, KNN, Naïve bays decision tree, random forest tree classifier which formulated on 25 malware families is converted into fixed gray scale image and the gray scale image transformed on 100x100 size which based on single format into 100000 columns value the phase test identify the pattern of predicted attribute value with 92% accuracy of random forest classifier and 88% accuracy of support vector machine classifier & 81% accuracy of decision tree classifier which revealed the best accuracy & prediction method. (D. Roshni, A., 2022) dynamic analysis based mobile malware classification using the supervised machine learning method, during the pandemic period, mobile hacking and mobile malware attacked happened to capture the essential information from cell phone and target the person and track the location, supervised machine learning models investigate the machine learning algorithm which formulated on classify the malware attacks, the systematic method of supervised machine learning approach detect the malware data points and classify them into binary classification to detect the malware of benign is main component of malware attack which needs to identify theme from the dataset values. The objective of this research is to investigate the mobile malware attack to classify them by using the machine learning approach and identify the pattern with high efficacy rate, it is important to note that the performance measure like precision call and F1 score has been measured and classify with great prediction accuracy rate including mean absolute error and root mean square error are experimented to conduct the benchmark analysis of malware detection and prevention to discover the latest roll back analysis of machine learning modeling, it is suggested that the latest android 11, and 12 version based on good security features with great machine learning classifier method. Machine learning method such as decision tree, support vector genetic algorithm random forest classifier Ada Boost, multilayer perceptron are experimented and discussed in this research.



*Table 1: performance of classification after feature selection using the information gain (*RENJITH, P., 2021)

The evaluation of machine learning (Obuzor, Z., 2022) algorithm tested on IOT system to predict the malware attack with OPCODE prediction. The popularity of advanced machine learning approach including various applications to scan & predict the various identification and detection approach to identify and secure the malware attack from the system. The OPCODE method is used to classify the malware from IOT system and the OPCODE was generated from dissembled malware program. The supervised machine learning method used to obtain the random forest classification technique including decision tree classification approach naïve bays support vector have been implemented in research and the experiment result represented the classification report of including 512 IOT benign and malware files showed the random forest classifier which performed the best accuracy & prevention techniques including better accuracy of 100% (Mišić, V.B., 2022) android malware detection using feature ranking of permission the use of android permission including effective differentiation method to identify the benign and malware app the extraction of android permission eliminate the zero impact of which apply feature ranking technique including chi-square test and fisher exact test to rank the additional filter to test them. The comparative result has been tested and implemented by obtaining the classification report and summary report of machine learning which essentially tested and implemented based on the small set of relevant permission. The decision tree, support vector machine and random forest classifier used to detect the malware app the analysis indicate the approach with better accuracy and great F1 score and reported the other features which proposed on different attributed values. Beside this the random forest classifier approach which combination of fisher exact test, and achieve the 99% accuracy in 92% accuracy in F score which corresponds to positive rate of dataset question to formulate the research problem, the improved results of this research with 99% accuracy of prediction including great precision call system approach which validated the android malware app infected with malware or not with false positive rate has been tested which considered on various malware detection & prevention method. (CHANDRIKA, A.T.B 2021) android malware detecting method by using of extra trees classification approach which discriminant feature selection, and the feature selected on training and testing the machine learning classification decision tree, random forest tree, genetic algorithm including support vector machine learning without using hyper parameter tuning and the identification capability used with different learning method of feature selection & proposed with various parameter tuning approach. The experiment result present that the validation of extra tree classifier method designed to proposed the tree classification method the optimization method of feature subset techniques helping in reduction of feature reduction dimension to less than the half of original dataset. The classification report accuracy about to 70% prediction accuracy which maintained the hyper parameter tuning while working on much reduced feature dimensions there having positive impact on computational complexity of prediction & detection of learning classification techniques approach has been tested carefully (RENJITH, P., 2021) performance metric system for malicious URL data using revised random forest algorithm phishers used ridicules emails phishing programing takes to phishing cost which developed to capture the data from mobile phone, the common use of phishing attack served billions dollar every year, beside this phishers used satirized emails phishing programming to take the budgeting outcome to proposed the record of individual data samples including various information, phishing attack capture the device user name and password and sends link to redirect them to on particular link which investigate different highlight of machine learning calculations. The utilized technique of phishing web locales depends on the surface of page local properties method which capture the location of the user by recognizing the particular location and the proposed model has been tested and identified including five distinct calculation by machine learning calculation.



*Figure 1: comparison performance after feature selection ((RENJITH, P., 2021))*

The effect of sampling (Kumar, M.A., 2022) in the machine learning malware analysis proposed on testing and training the dataset to select the feature on target variable and train the machine learning model and getting the outcome of prediction on app permission dataset of android phone the malware depends on the user permission given by the user and user tend to grant permission unknowingly on such dangerous application to cause the problem over device machine learning random forest classifier , genetic algorithm gradient boost, decision tree and support vector machine classifier produce the great outcome to detect and prevent the malware from the android phone, it is suggested that the latest android phone security updated with avoid of malware attack due to secure application permission access based on the user grant. The precision call and F score present with great accuracy report to identify the malware from android phone by using advance machine learning method.

# References

Al Ali, M., Svetinovic, D., Aung, Z. and Lukman, S., 2017, December. Malware detection in android mobile platform using machine learning algorithms. In *2017 International Conference on Infocom Technologies and Unmanned Systems (Trends and Future Directions)(ICTUS)* (pp. 763-768). IEEE.

Rana, M.S., Gudla, C. and Sung, A.H., 2018, December. Evaluating machine learning models for Android malware detection: A comparison study. In *Proceedings of the 2018 VII International Conference on Network, Communication and Computing* (pp. 17-21).

Zou, S., Zhang, J. and Lin, X., 2015. An effective behavior‐based Android malware detection system. *Security and Communication Networks*, *8*(12), pp.2079-2089.

Sangal, A. and Verma, H.K., 2020, September. A static feature selection-based android malware detection using machine learning techniques. In *2020 International conference on smart electronics and communication (ICOSEC)* (pp. 48-51). IEEE.

Yildiz, O. and Doğru, I.A., 2019. Permission-based android malware detection system using feature selection with genetic algorithm. *International Journal of Software Engineering and Knowledge Engineering*, *29*(02), pp.245-262.

Zhu, H.J., Jiang, T.H., Ma, B., You, Z.H., Shi, W.L. and Cheng, L., 2018. HEMD: a highly efficient random forest-based malware detection framework for Android. *Neural Computing and Applications*, *30*(11), pp.3353-3361.

Singh, L. and Hofmann, M., 2017, December. Dynamic behavior analysis of android applications for malware detection. In *2017 International Conference on Intelligent Communication and Computational Techniques (ICCT)* (pp. 1-7). IEEE.

Ali, W., 2019. Hybrid intelligent android malware detection using evolving support vector machine based on genetic algorithm and particle swarm optimization. *IJCSNS*, *19*(9), p.15.

Osman, M.Z., Abidin, A.F.Z., Romli, R.N. and Darmawan, M.F., 2021, August. Pixel-based Feature for Android Malware Family Classification using Machine Learning Algorithms. In *2021 International Conference on Software Engineering & Computer Systems and 4th International Conference on Computational Science and Information Management (ICSECS-ICOCSIM)* (pp. 552-555). IEEE.

Ganapathi, P., Shanmugapriya, D. and Roshni, A., 2022. Dynamic Analysis Based Mobile Malware Classification Using Supervised Machine Learning Methods. *Dynamic Analysis Based Mobile Malware Classification Using Supervised Machine Learning Methods*, pp.1-168.

Anidu, A. and Obuzor, Z., 2022. Evaluation of machine learning algorithms on Internet of Things (IoT) malware opcodes. In *Handbook of Big Data Analytics and Forensics* (pp. 177-191). Springer, Cham.

Saleem, M.S., Mišić, J. and Mišić, V.B., 2022. Android Malware Detection using Feature Ranking of Permissions. *arXiv preprint arXiv:2201.08468*.

CHANDRIKA, A.T.B., VIJAYANAND, R. and RAO, D.P.S., ANDROID MALWARE DETECTION USING EXTRA TREES CLASSIFIER BASED FEATURE SELECTION AND MACHINE LEARNING. 2021

RAMESH, K., BENNET, M.A., VEERAPPAN, J. and RENJITH, P., 2021, April. Performance Metric System for Malicious URL Data using Revised Random Forest Algorithm. In *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)* (pp. 1188-1191). IEEE.

Thimmaiah, K.S., Raj, L.S., Bolimera, P. and Kumar, M.A., 2022. The Effect of Sampling in the Machine Learning-Based Malware Analysis. In *Proceedings of International Conference on Intelligent Cyber-Physical Systems* (pp. 143-156). Springer, Singapore.