Methodology

* Random Forest, Support Vector Machine and PCA - **3 classifiers** used
* **Ensemble pipeline** – We are using GridSearchCV for Hyperparameter tuning in our SVM classifier. Best parameter to be chosen are extracted by using Grid Search in (gamma, C) pattern, radial basis function (RBF), Polynomial and Linear Kernel functions being used. Best params are fed into the SVM classifier in 5 folds for each of 126 candidates, total in 630 fits. This is following the ensemble pipeline techniques.
* **Other models considered** – No third model considered.
* **Hyper-parameter tuning** –
  + Done for GridSearchCV of optimum combination of parameters in SVM Classifier [C, Gamma, Kernel].
  + Dataset for training and test set sizes for model execution while doing random forest classification.
  + Number of components/features tweaked while doing PCA.

Overview

* **Objective** – Identify if an application is authentic application or a Malware.
* **Methodology** –
  + Perform EDA on the data, Data Wrangling
  + Perform Feature Engineering
  + Extract Top 10 Features,
  + Machine Learning model implementation
  + Perform model Comparison
  + Share conclusions and results

Dataset

* **How many features?** - 330 independent and one dependent features.
* **Size of the dataset?** – 398 rows and 331 columns
* **Multiple files?** – Single Dataset
* **What kind of data – numerical or character?** – numerical
* **Balanced or imbalanced – what is the distribution?** – Both classes 1 and 0 for target variable type have same ratio, hence we have a balanced dataset.
* **Distribution of Training set, validation set, testing set?** – Training set with 70%, Testing with 30%, No validation set
* **Missing data?** - No missing values in the dataset
* **Preprocessing challenges?** – Duplication data checked to reduce the overfitting in the model, else would be the result in bias addition.

Feature Engineering Techniques

FE1 + ML1 is Feature Selection using Gini Importance + Random Forest Classifier

1. Features removed – Feature importance have been calculated by using Gini importance measure which gives us top 10 features, thereby removing 320 features excluding the target variable.

2. Feature creation – Zero feature added

3. Feature ranking – made based on Gini Importance (10 Features ranked)

4. Class imbalance treatment - Both classes 1 and 0 for target variable type have same ratio, hence we have a balanced dataset.

5. Given our test dataset the Random Forest Classifier model correctly predicted that 30 out of 37 applications were malware.

FE1 + ML2 is Feature Selection using Gini Importance + Support Vector Machine

1. Features removed – Zero feature removed.

2. Feature creation – Zero feature added.

3. Feature ranking – Feature importance extracted based on recall score.

4. Class imbalance treatment - Both classes 1 and 0 for target variable type have same ratio, hence we have a balanced dataset.

5. Given our test dataset the Support Vector Classifier model correctly predicted that 31 out of 37 applications were malware.

FE2 + ML1 is PCA + Random Forest Classifier

1. Features removed – 330 features removed down to 30 features. We initially started with 80 features because PCA algorithm didn’t accept 330 features and we had to initialize with 80 features(components).

2. Feature creation – Zero feature added.

3. Feature ranking – made based on Explained Variance Ratio.

4. Class imbalance treatment - Both classes 1 and 0 for target variable type have same ratio, hence we have a balanced dataset.

5. Given our test dataset the Random Forest Classifier model correctly predicted that 32 out of 37 applications were malware.

FE2 + ML2 is PCA + Support Vector Machine

1. Features removed – 330 features removed down to 30 features. We initially started with 80 features because PCA algorithm didn’t accept 330 features and we had to initialize with 80 features(components).

2. Feature creation – Zero feature added.

3. Feature ranking – made based on Explained Variance Ratio.

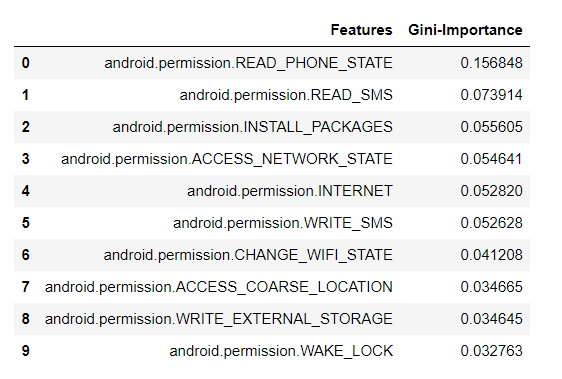
4. Class imbalance treatment - Both classes 1 and 0 for target variable type have same ratio, hence we have a balanced dataset.

5. Given our test dataset the Support Vector Machine model correctly predicted that 30 out of 37 applications were malware.

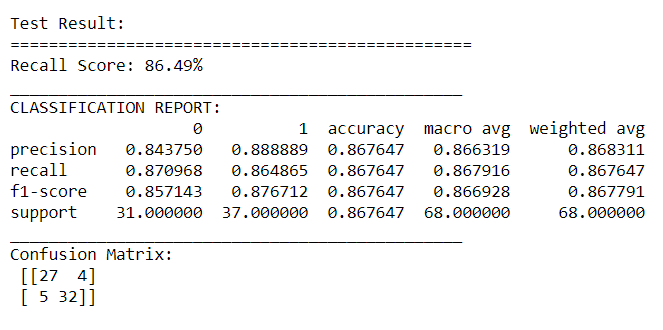
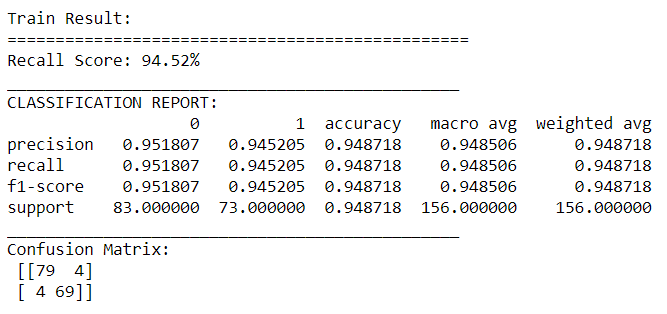
Results

* Table for the evaluation metric for each ML technique used

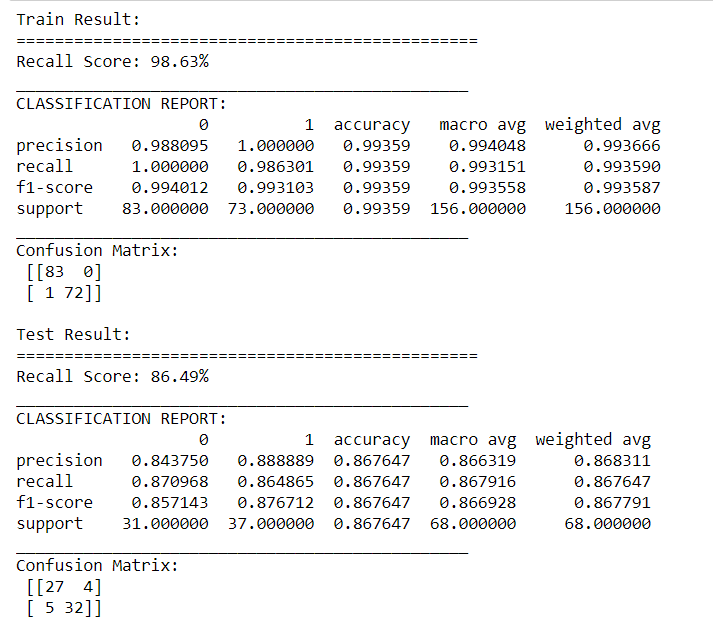
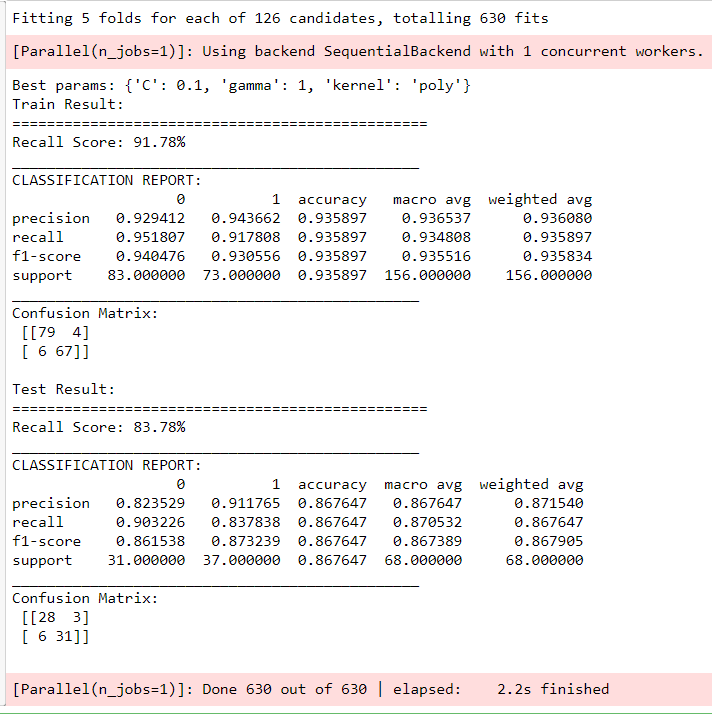
List of Top 10 features by Gini-Importance Scores:



Classification Report of the random forest model

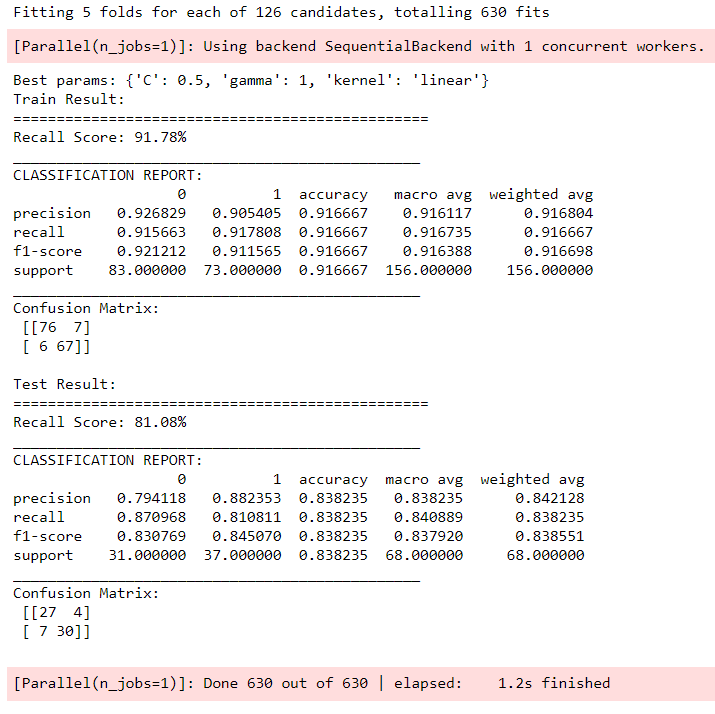


Classification Report of the Support Vector Classifier Classification Report of the Updated Random Forest Model



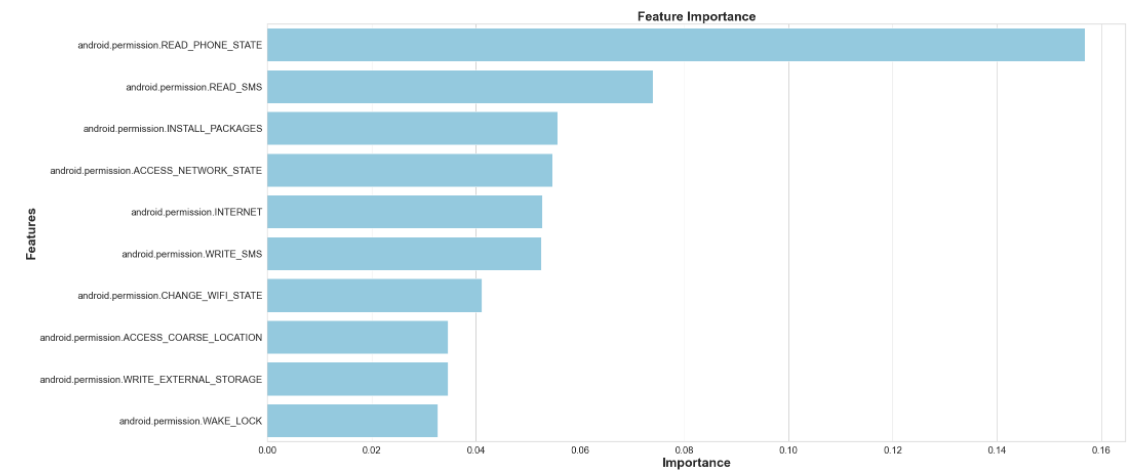
Results

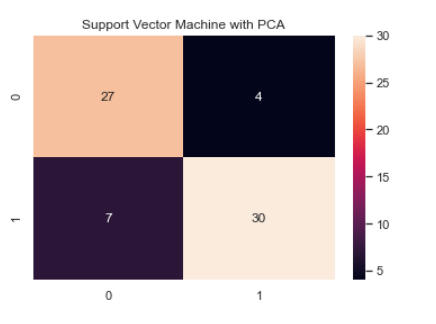
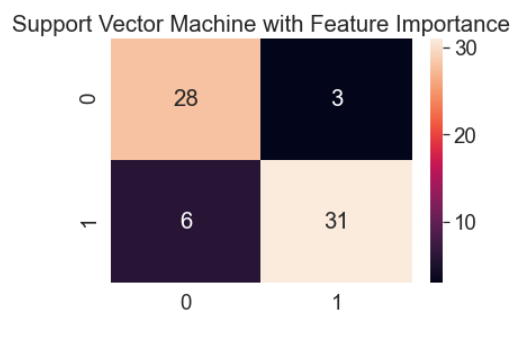
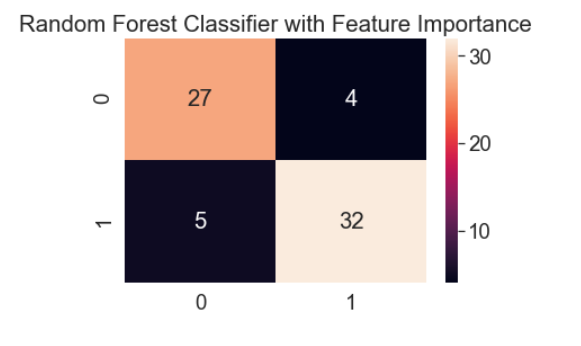
GridSearchCV for HyerParameter tuning of optimal combination of Parameters of SVC



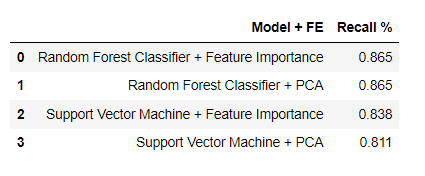
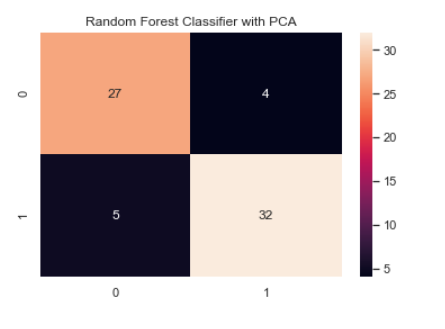
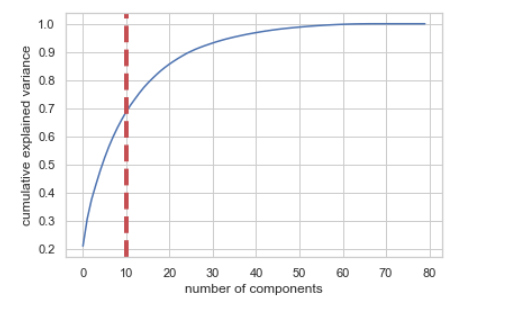
Results

* Plot of the curves –

The feature importance have been calculated below by using Gini Importance measure, which gives us the top 10 features by importance.



**This graph shows that after more than 30 components, we don’t gain very much explained variance**



* **Conclusion**-

From the above Dataframe, we can come to the conclusion, that the Random Forest Classifier with PCA as Feature Engineering technique, helped perform well on the Test data in Accurately Classifying 86.5% of the Applications as Malware