Data-Mining CS-F415
Bonus-Assignment

Pre-Processing Data

The data was scaled using MinMax scaler for KNN because it uses distance metric,for all other models the data was not scaled .The data were combined into one dataframe for training using number 0 and 1 to identify benign and malignant class.The attribute 'FileName' was dropped before training because it was a unique identifier.The data did not contain any null values or any illegal values.
RandomForest was used to perform feature selection by finding the least important features and the resultant data was trained on RandomForest.
PCA with 5 components was used to reduce the features to 5 and on which RandomForest was trained.

Discussion About Data:

After preprocessing there were 6769 records of which 4060 records belonged to malicious class and 2709 belonged to benign class.
The data was split into training and validation sets to find the optimal parameters for different models.

Comparing Models-Used:

- **Logistic Regression:**

The data was split into testing and training data and the classifier was trained on training data and tested on testing data.the accuracy on both testing and training data was plotted against log10(C) in "random-forest_accuracy.jpg" to determine an optimal value of C as 0.01.

f1-score= 0.98
classification accuracy=0.97769

- **Support Vector Machines:**

The training split was trained on SVM with radial basis function as kernel was tested on testing data.the accuracy on both testing and training data was plotted against log10(C)
in "svmt_accuracy.jpg" to determine an optimal value of C as 0.1.

f1-score= 0.74
classification accuracy=0.64603

- **K-Nearest Neighbors :**

The training split was trained on KNN was tested on testing data.the accuracy on both testing and training data was plotted against number of neighbors parameter in “knn_accuracy.jpg” to determine an optimal value of n_neighbors as 3.

f1-score= 0.99
classification accuracy=0.988033

The same model was trained on scaled features which produced the following results

f1-score= 0.99
classification accuracy=0.98640

- **Decision Tree:**

The training split was trained on DecsionTree was tested on testing data.the accuracy on both testing and training data was plotted against maximum depth parameter in “decision-tree_accuracy.jpg” to determine an optimal value of max_depth as 12.

f1-score= 0.99
classification accuracy=0.98566

- **RandomForest:**

The training split was trained on Random Forest with 50 trees was tested on testing data.the accuracy on both testing and training data was plotted against maximum depth parameter in “random-forest_accuracy.jpg” to determine an optimal value of max_depth as 20.

f1-score= 1
classification accuracy=0.99881

Least Important features were removed using the above model reducing the number of features to 587.Random Forest with 200 trees were trained on the resultant dataset.The test-training accuracy was plotted against max-depth in “random-forest-accuracy_vs_depth_after_dropping features.jpg” to determine an optimal value of max_depth as 20.

f1-score= 1
classification accuracy=0.998965

PCA with 5 components was applied to the data with 587 features used in above model.Random Forest with 300 trees and max_depth of 20 were trained on the resultant dataset.

f1-score= 1
classification accuracy=0.998965

Results And Conclusion:

On comparison SVM performed poorly on the data as compared to other algorithms.All other model gave good result with Random Forest models scoring the highest in all metrics used.

RandomForest with 200 trees and max_depth of 20 trained with 1220 features dropped gave the highest accuracy and f1 score with the following result.

precision=1
recall=0.997
f1-score= 0.998
classification accuracy=0.998965

And Random Forest with 300 trees and max_depth of 20 trained on the above data with 1220 features dropped after applying PCA with 5 principal components gave same result.

precision=1
recall=0.997
f1-score= 0.998
classification accuracy=0.998965

Some of the most Important opcodes are 472,7,447,494,568 as per Random Forest's feature importance.

The final model proposed:
Classifier:Random Forest Classifier
Number of estimators=300
Maximum depth of trees=20
The KKD diagram of the model is plotted in "ModelKDD.jpg"

Submitted by:
Rusafid Mirza Pottengad.
2015B4A70572G