Observations and Training Information

Team Name : Hawk Eyes

1. **Files Provided:**
    a. **Static Analysis**
        - **String.txt** file contains all the strings present in the executable under consideration.
        - **Structure_Info.txt** has the structure of the executable.
    b. **Dynamic analysis**
        - **<Hash_Value>.json** file has information about the dynamic analysis of the executable. In dynamic analysis the files are run in sandbox environments and their behaviour is monitored.

2. Observations from the files provided:

- For static analysis on a String.txt file we focused on encoding the string information in vectorized form and then using it as our training data.
- For static analysis on **Structure_Info.txt** file we considered major observations in the files ,i.e., Difference in the virtual size and raw data size of the different section, and presence of unnecessary shell scripts in the files.
- From the **<Hash_Value>.json** file, we can observe that an antivirus is run during dynamic analysis, so using some clever string search on .json file we can figure out if the executable is malicious or not.

3. Data Preparation, Features creation and Training:

For Static analysis on **String.txt** we first filtered the files to extract useful strings and than used string encoding to convert the information in vectorized form.

For Static analysis on **Structure_Info.txt** we first filtered the files to extract two features - the count of sections having abruptly large differences in the virtual and rawdata size. And the count of shell scripts in the file .encoding to convert the information in vectorized form.

We did an 80-20(train-test) split of the data. We then used the Random Forest Classifier Algorithm to train our model.
Our results on the test dataset for combined models are as follows:

Total number of files in test dataset: 1999
Correctly Predicted: 1998
Accuracy: 0.99949974987

Confusion Matrix:

	Predicted 0	Predicted 1
Actual 0	1017	0
Actual 1	1	981

Precision: 1
Recall: 0.99898167006
F1 Score: 0.99949057564