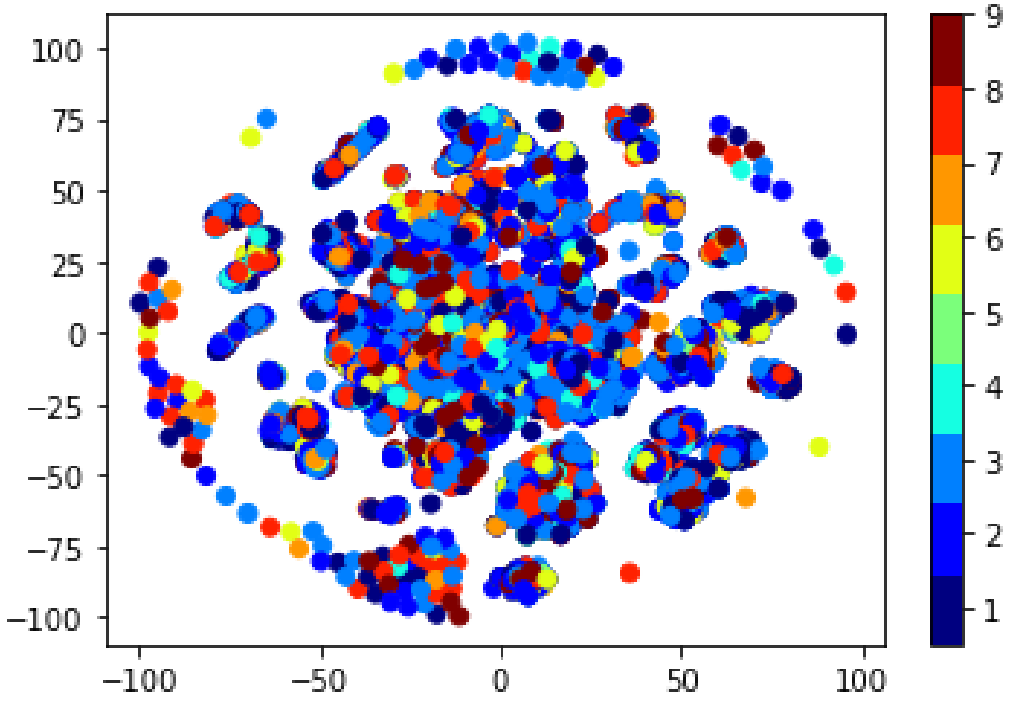
Malware Assignment

Steps performed:

1. **Feature extraction:**
    - **Bytesfile**
        1. Unigram feature extraction (BOW)
        2. Bigram feature extraction

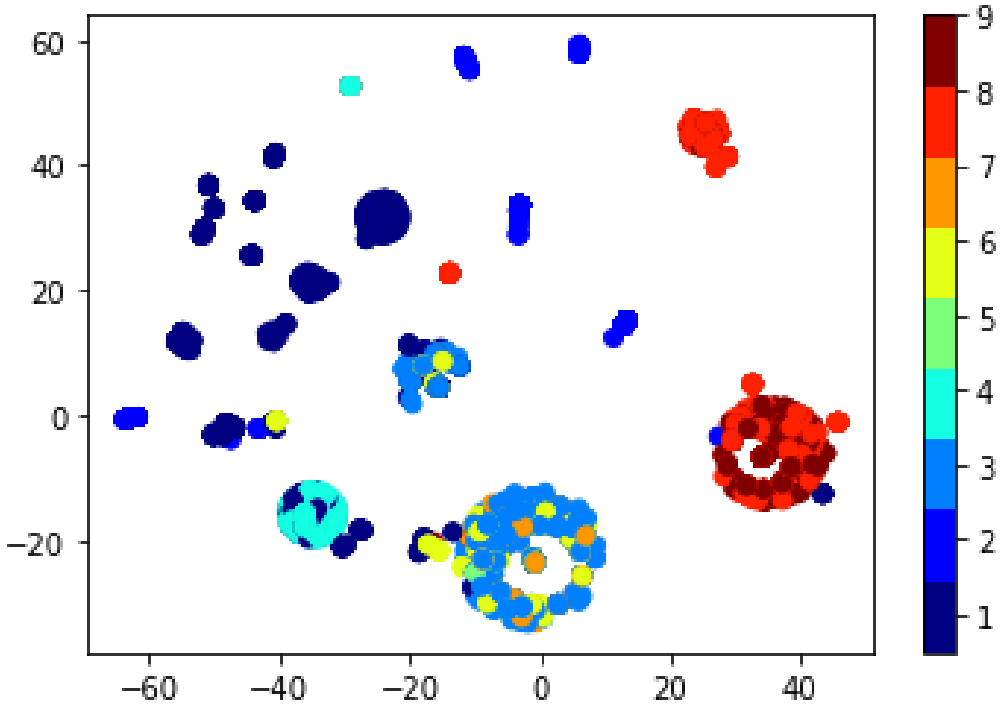
           Since the bigram feature are 32896 features so I used svd decomposition to decrease the number of feature and selected only top 1000 features
        3. Size of file as a feature
    - **Asmfile**
        1. Asm feature like .data, .exb features
        2. ASM image pixel features
        3. Size of file as feature.

2. TSNE Visulization:
    - Bigram features:



The bigram features are forming small clusters. Although the features are not well clustered, they can be useful for classification.

- ASM image features:



The top 800 image pixel based features are very well clustered in the above plot indicatine them to be useful for classification.

3. Train test split:

train:val = 75:25 (stratified)

3. Feature Selection :

Even after applying svd on bigram still there are total 2110 features (1000 bigram after svd + 257 unigram features from bytesfile + 1 bytefilesize + 51 asmfile feature + 800 asm-image pixel feature + 1 asm filesize)

So I have use random forest model for feature selection on the basis of feature importance. And after validation on cross-validation data it was found that selecting top 200 features were giving the best result.

4 Modelling :
After using a variety of model It was found that LIGHTGBM worked better.

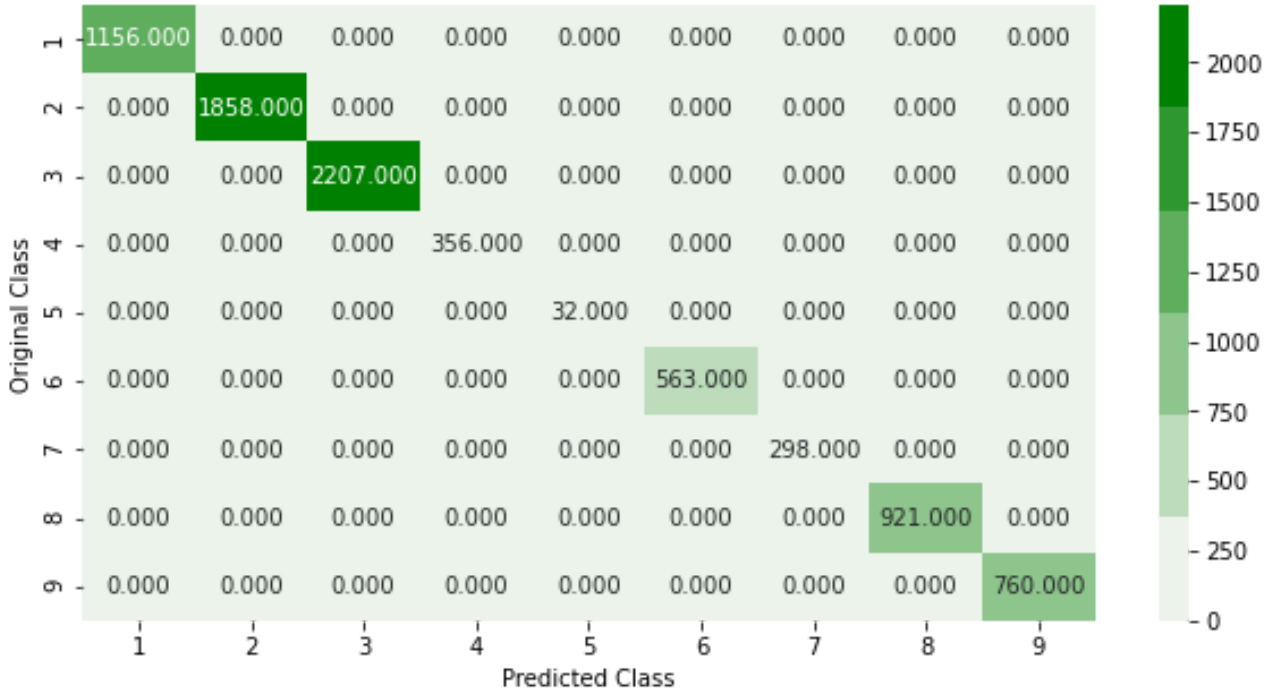
So later I used again lightgbm for feature selection on the basis of feature importance in place of random forest (top 300 features selected)t and use the selected feature for classification.

5. Result:

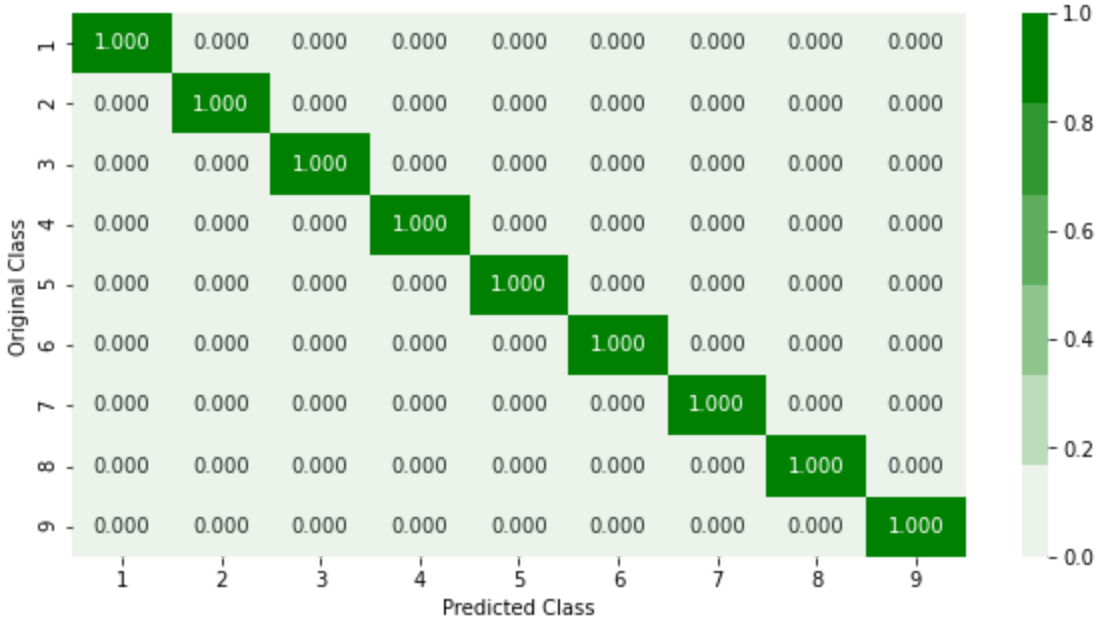
Training data:

Logloss = 0.00017021813719157107

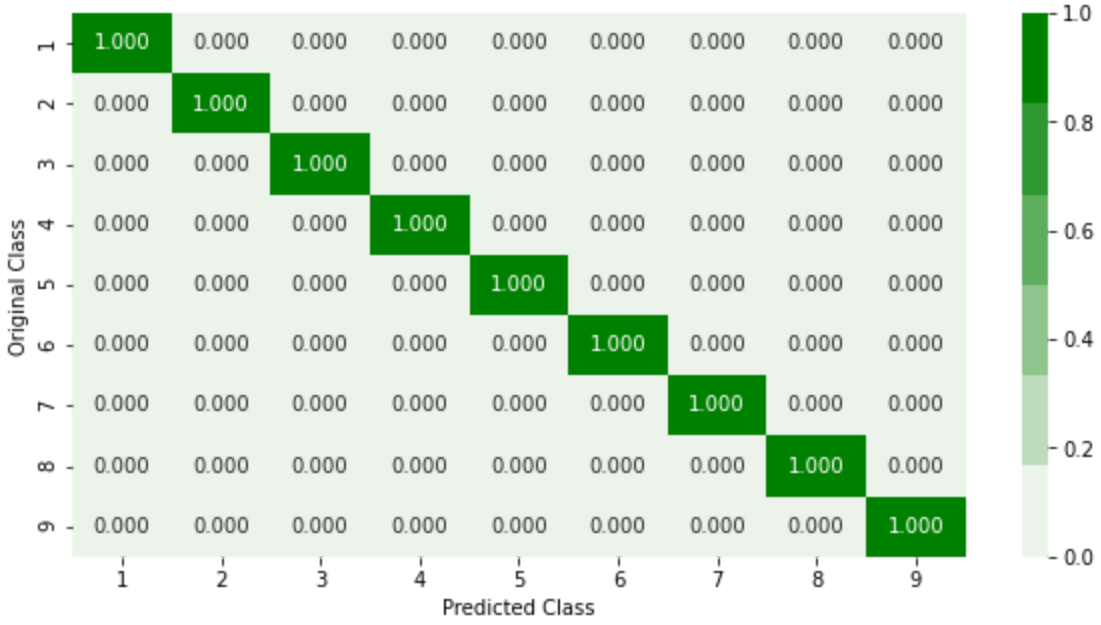
Confusion matrix:



Precision matrix:

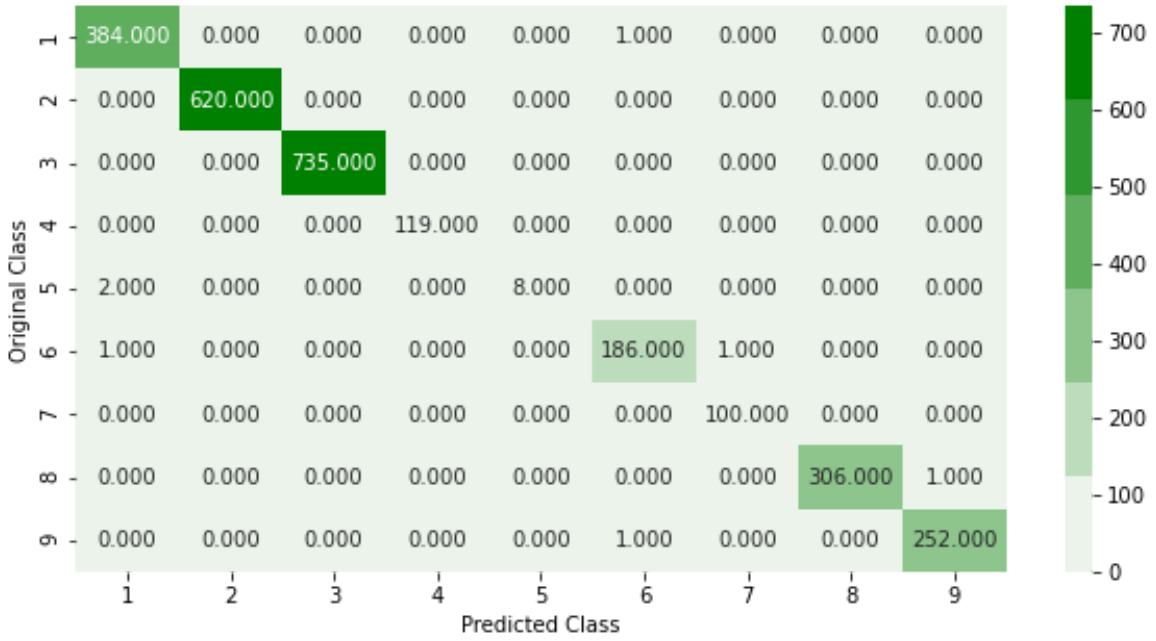


Recall metrix:

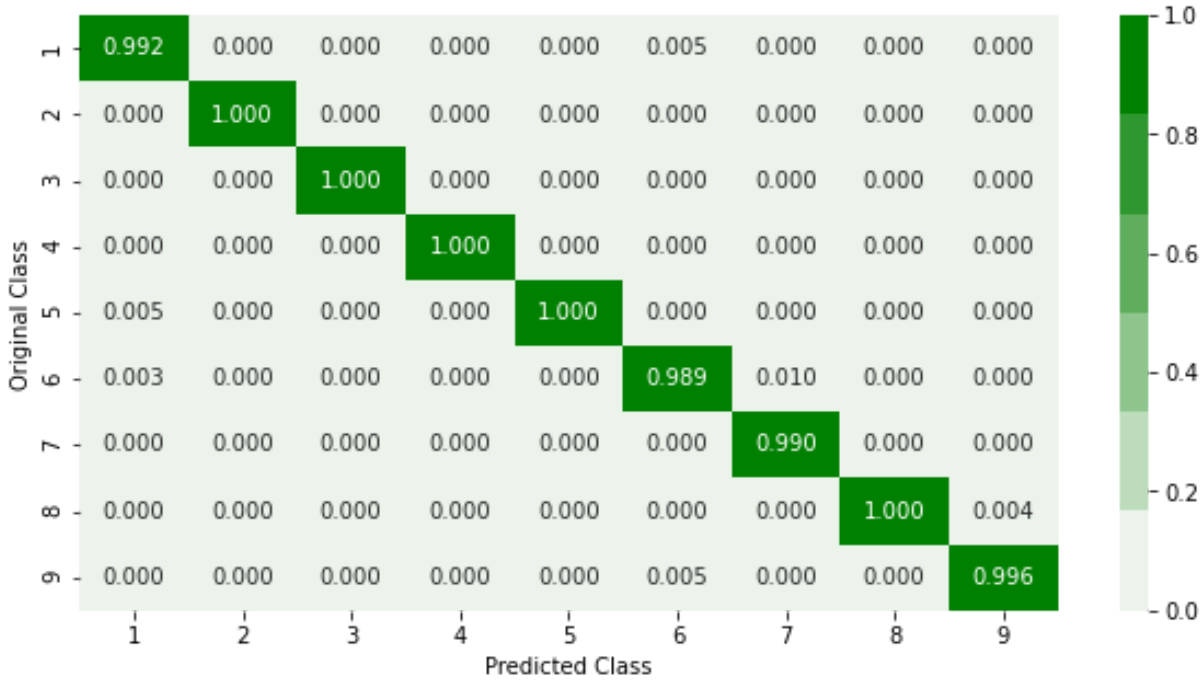


Validation data:
Logloss = 0.00967886438881175

Confusion Matrix:



Precision Matrix



Recall matrix: