Attack Classification using Naïve Bayes Algorithm

Step 1:

Downloading the dataset and checking class distribution.

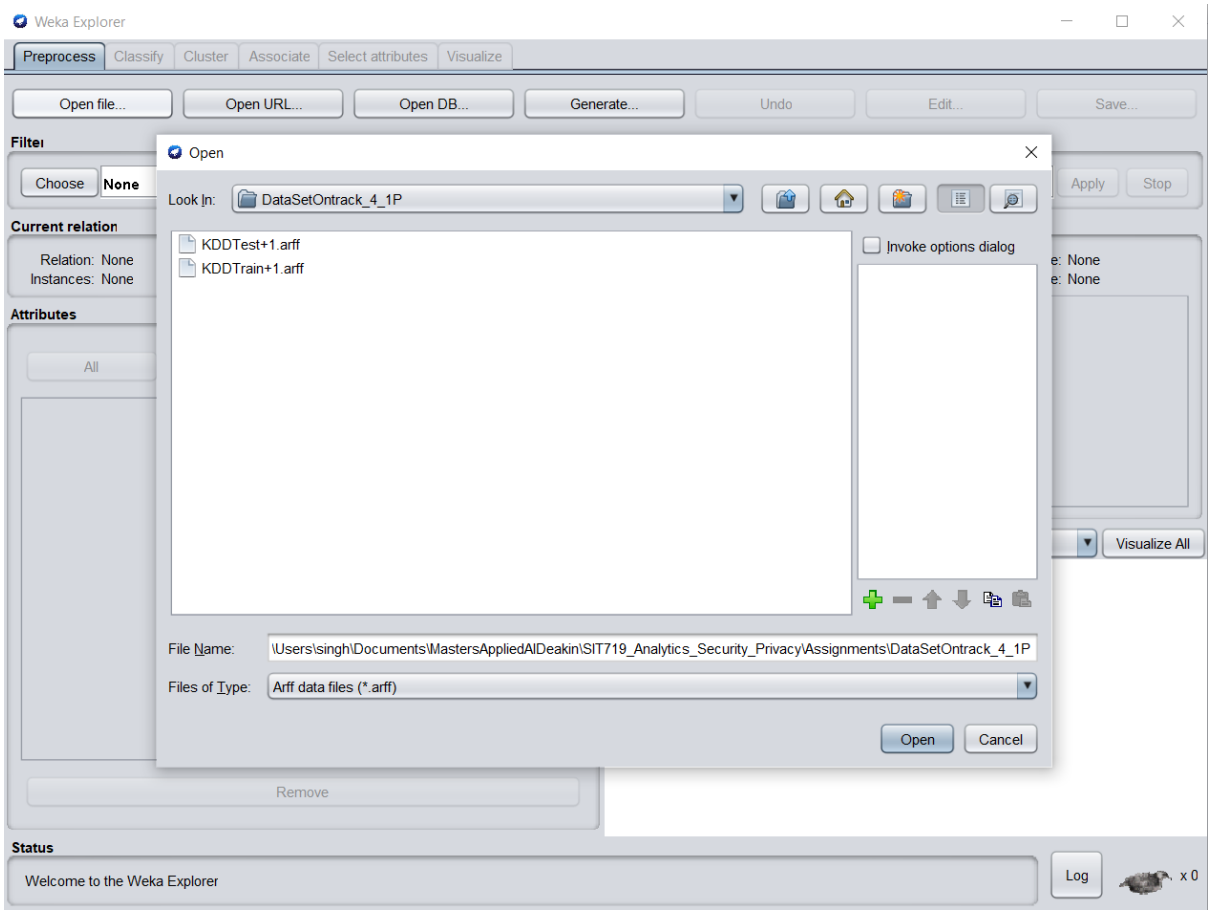

Fig: Downloaded data, train and test.

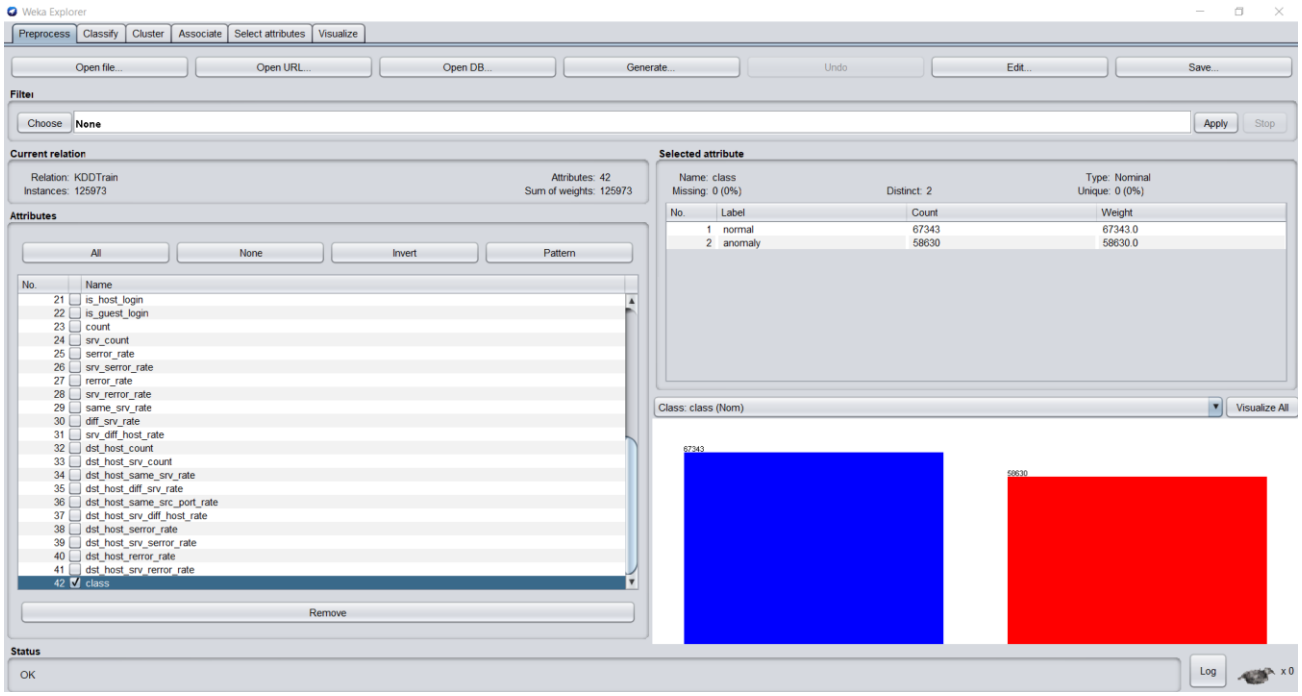

Fig: Data distribution

Step 2:

Applying Naïve Bayes Classifier

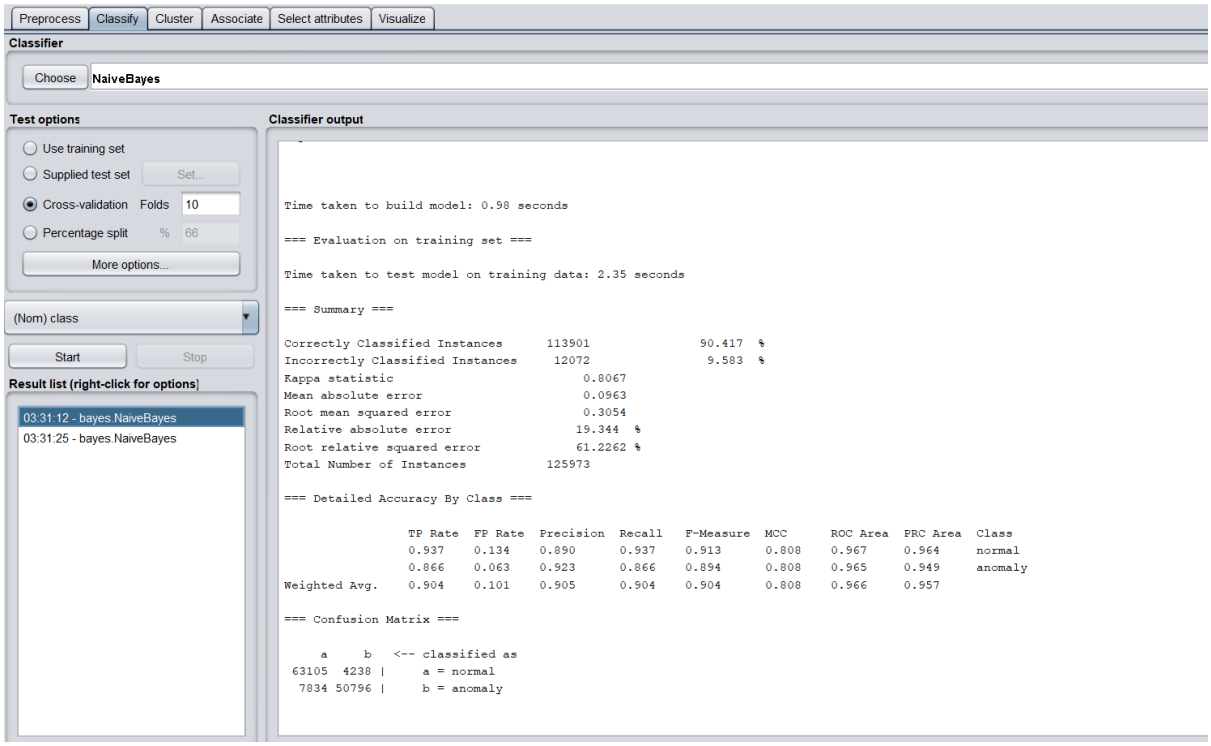

Fig: Classification summary after applying Naïve Bayes classifier.

Step 3:

Performing 10-fold cross validation

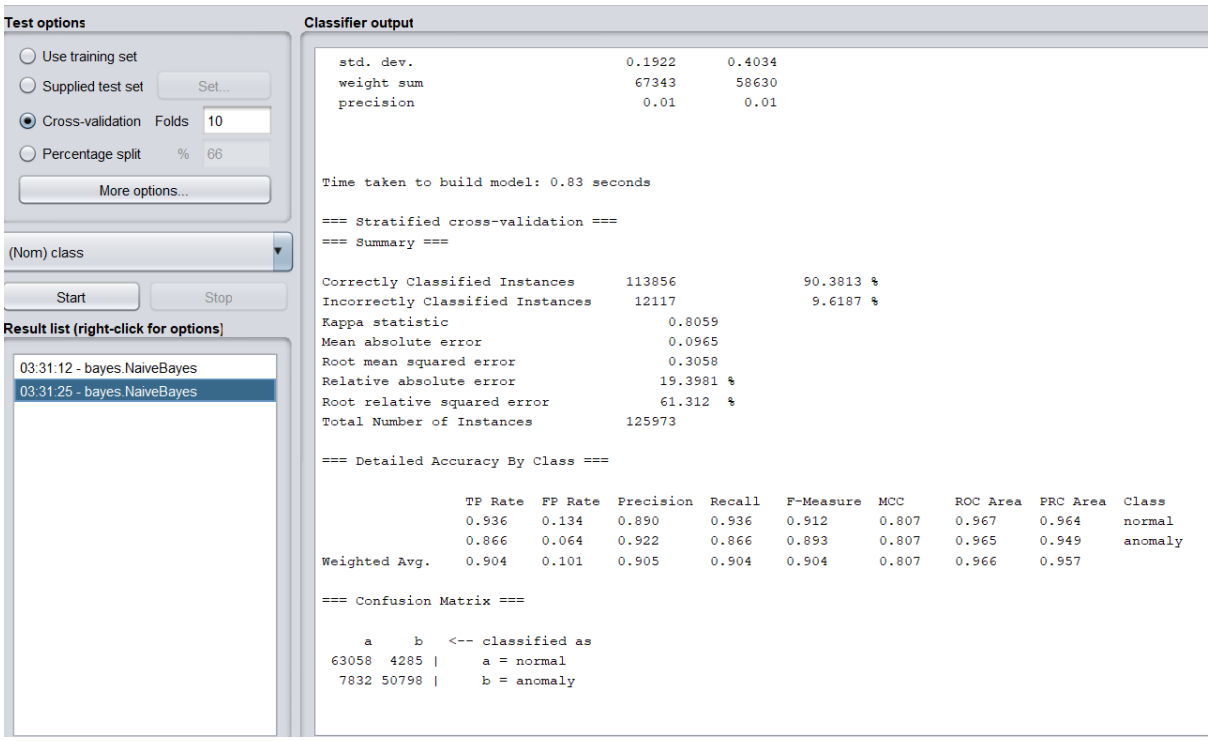

Fig: Summary of 10-fold cross validation

Step 4:

Upload test data and checking classification result

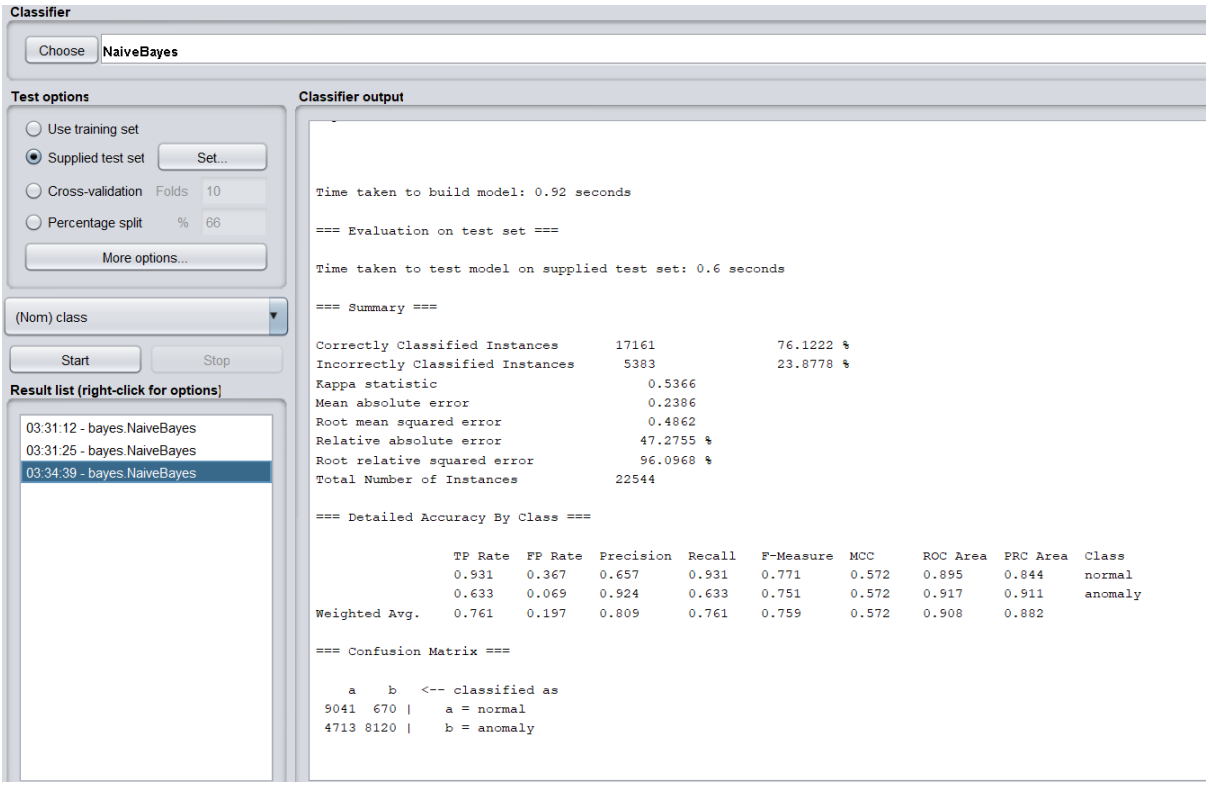

Fig: Summary on uploaded test data

Step 5:

Compare results between 10-fold cross validation and test dataset.

Ground Truth\Classification	**Normal - pred**	**Anomaly - pred**
Normal - gt	63058	4285
Anomaly - gt	7832	50798

Table 1: 10-fold Cross validation result

Ground Truth\Classification	**Normal - pred**	**Anomaly - pred**
Normal - gt	9041	670
Anomaly - gt	4713	8120

Table 2: Test set result

Here, we can see the difference between cross validation and test set results.

Table 1 shows that during cross validation a total of 113856 samples were correctly classified (63058 + 50798) and 12117 were incorrectly classified (4285 + 7832).

Table 2 shows that during classification on test data a total of 17161 samples were correctly classified (9041+ 8120) and 5383 were incorrectly classified (4285 + 670).

Here, correct classification is constituted of two elements, True Positives and True Negatives, similarly, misclassification constitutes of two elements, False Positives and False Negatives.

True Positives: When sample is normal and classified as normal.

True Negatives: When sample is anomaly and classified as anomaly.

False Positives: When sample is anomaly and classified as normal.

False Negatives: When sample is normal and classified as anomaly.

(*This is when normal is considered as positive and anomaly as negative, if we interchange the label assigned to these classes then the meaning will change accordingly).*

Based on these values we have the following metrics:

	10 fold cross validation	Test data
Accuracy	90.38 %	76.12 %
Precision (Weighted avg)	90.5 %	80.9 %
Recall (Weighted avg)	90.4 %	76.1 %

We can see that there is a performance drop in test data as compared to 10-fold cross validation results. This indicates that the model does not generalize well on unseen data and is possibly overfitted.