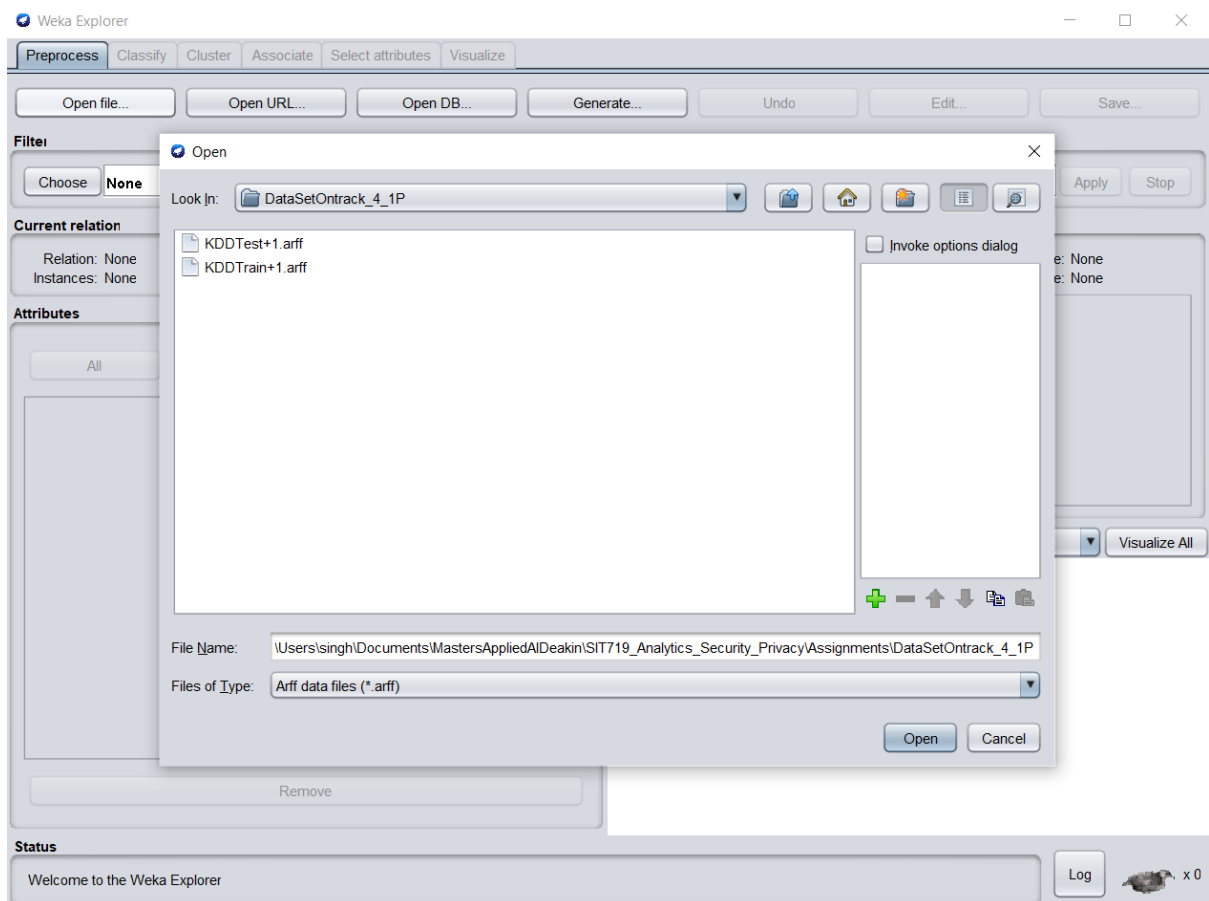


# Attack Classification using Naïve Bayes Algorithm

## Step 1:

Downloading the dataset and checking class distribution.



*Fig: Downloaded data, train and test.*

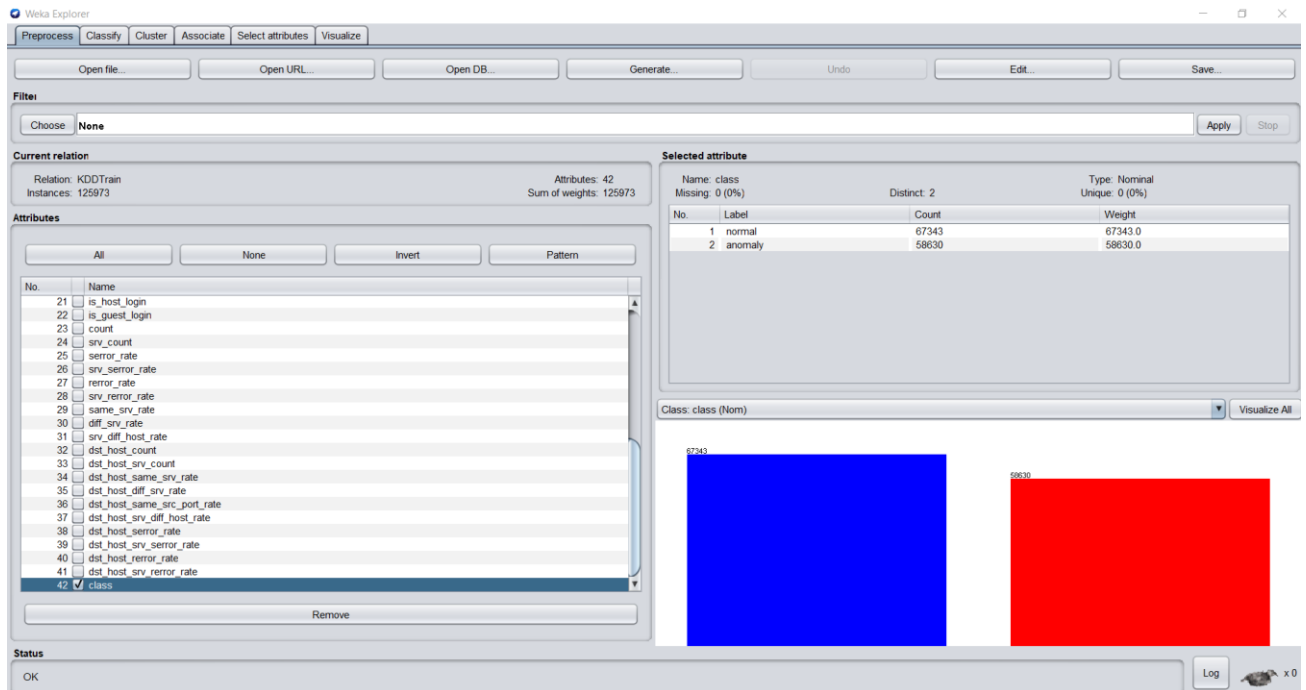


Fig: Data distribution

## Step 2:

### Applying Naïve Bayes Classifier

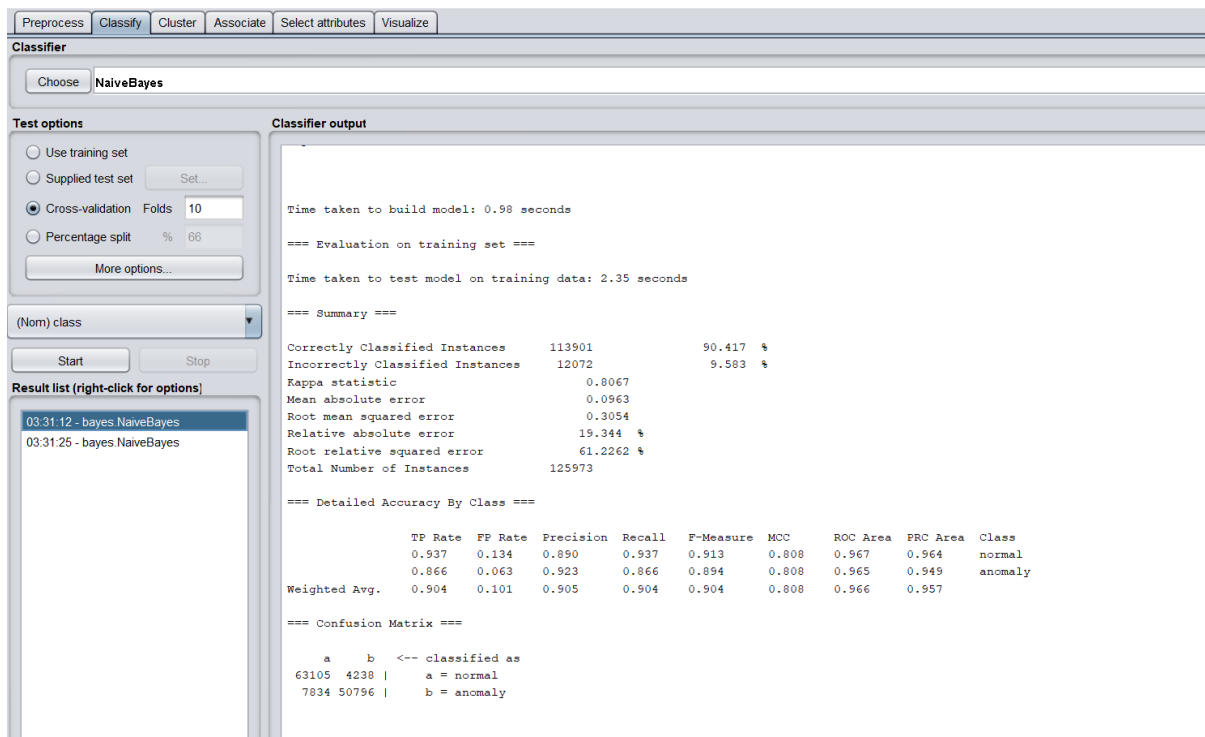


Fig: Classification summary after applying Naïve Bayes classifier.

### Step 3:

#### Performing 10-fold cross validation

☐ Use training set

☐ Supplied test set

☒ Cross-validation Folds

☐ Percentage split %

(Nom) class

Result list (right-click for options)

03:31:12 - bayes.NaiveBayes

03:31:25 - bayes.NaiveBayes

Classifier output

std. dev. 0.1922 0.4034

weight sum 67343 58630

precision 0.01 0.01

Time taken to build model: 0.83 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 113856 90.3813 %

Incorrectly Classified Instances 12117 9.6187 %

Kappa statistic 0.8059

Mean absolute error 0.0965

Root mean squared error 0.3058

Relative absolute error 19.3981 %

Root relative squared error 61.312 %

Total Number of Instances 125973

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.936	0.134	0.890	0.936	0.912	0.807	0.967	0.964	normal
	0.866	0.064	0.922	0.866	0.893	0.807	0.965	0.949	anomaly
Weighted Avg.	0.904	0.101	0.905	0.904	0.904	0.807	0.966	0.957	

=== Confusion Matrix ===

a b <-- classified as

63058 4285 | a = normal

7832 50798 | b = anomaly

Fig: Summary of 10-fold cross validation

#### Step 4:

Upload test data and checking classification result

The screenshot shows the 'Classifier' window with 'NaiveBayes' selected. The 'Test options' panel on the left has 'Supplied test set' selected. The 'Classifier output' panel on the right displays the following text:

```
Time taken to build model: 0.92 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0.6 seconds

=== Summary ===

Correctly Classified Instances      17161      76.1222 %
Incorrectly Classified Instances    5383      23.8778 %
Kappa statistic                    0.5366
Mean absolute error                 0.2386
Root mean squared error             0.4862
Relative absolute error             47.2755 %
Root relative squared error         96.0968 %
Total Number of Instances          22544

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.931   0.367   0.657     0.931   0.771     0.572   0.895    0.844    normal
          0.633   0.069   0.924     0.633   0.751     0.572   0.917    0.911    anomaly
Weighted Avg.   0.761   0.197   0.809     0.761   0.759     0.572   0.908    0.882

=== Confusion Matrix ===

  a    b  <-- classified as
9041  670 |  a = normal
4713 8120 |  b = anomaly
```

The 'Result list' on the left shows three entries for '03:34:39 - bayes NaiveBayes'.

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.