Task 4.1
Attach Classification using Naïve Bayes Algorithm
SIT719

Sanket Thakur

Summary of both the data sets derived by using WEKA

Test Data set

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       18200               80.731  %
Incorrectly Classified Instances      4344               19.269  %
Kappa statistic                          0.623
Mean absolute error                      0.1924
Root mean squared error                  0.4371
Relative absolute error                 39.2297 %
Root relative squared error             88.2712 %
Total Number of Instances            22544

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.950    0.301    0.705      0.950    0.809      0.652    0.958     0.949     normal
                 0.699    0.050    0.949      0.699    0.805      0.652    0.949     0.944     anomaly
Weighted Avg.    0.807    0.158    0.844      0.807    0.807      0.652    0.953     0.946

=== Confusion Matrix ===

    a    b   <-- classified as
 9225  486 |    a = normal
 3858 8975 |    b = anomaly
```

Training Data Set

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      113858               90.3829 %
Incorrectly Classified Instances     12115                9.6171 %
Kappa statistic                          0.806
Mean absolute error                      0.0965
Root mean squared error                  0.3058
Relative absolute error                 19.3947 %
Root relative squared error             61.3067 %
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
 63060  4283 |     a = normal
  7832 50798 |     b = anomaly
```

Confusion Matrix Comparison :
A confusion matrix is a way to summarise the efficiency of a classification algorithm

	Event	No-event
Event	True positive	False positive
No-event	False negative	True negative

Test data	Train Data
=== Confusion Matrix === a b <-- classified as 9225 486 | a = normal 3858 8975 | b = anomaly	=== Confusion Matrix === a b <-- classified as 63060 4283 | a = normal 7832 50798 | b = anomaly

From the matrices we learn that

- The classifier made total of 22544 and 125973 predictions in the test data and train data sets respectively
- The number of true positive and True negative in both the sets is high
- False positives in both the sets are very low, that means the type 1 errors in the data sets are low
- False positives in both the sets are low as well, that means the type 2 errors in the sets are low
- The accuracy of the classifier in test data is 81% and train data is 90%
- The misclassification or the error rate of the classifier for the test data is 19% and train data is 10%
- The true positive rate or sensitivity for the test data is 95% and for the train data is 94%
- The False positive rate for test data is 5% and for the train data is 6%
- The true negative rate or specificity for the test data is 70% and for the train data is 87%