Question 1

Q1 dataset. Load the diabetes.arff into Weka. The dataset already in Weka data folder.

1a) After loading the dataset, select the **Classify** tab, and then click on NaiveBayesSimple. Use 10-fold cross-validation to estimate the error of the Naive Bayes classifier.

Answer:

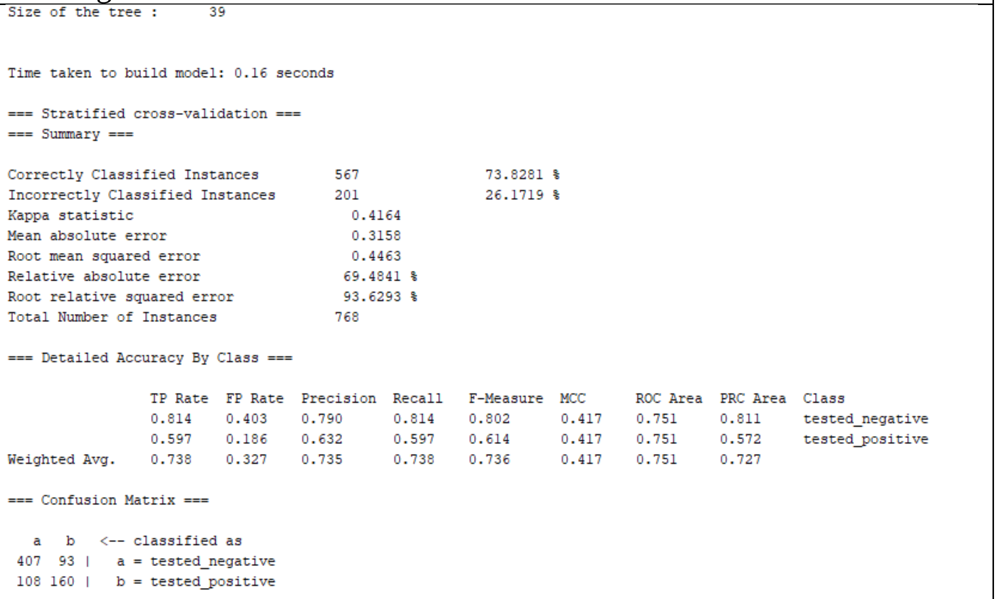
Mean absolute error: 0.2841

Root mean squared error: 0.4168

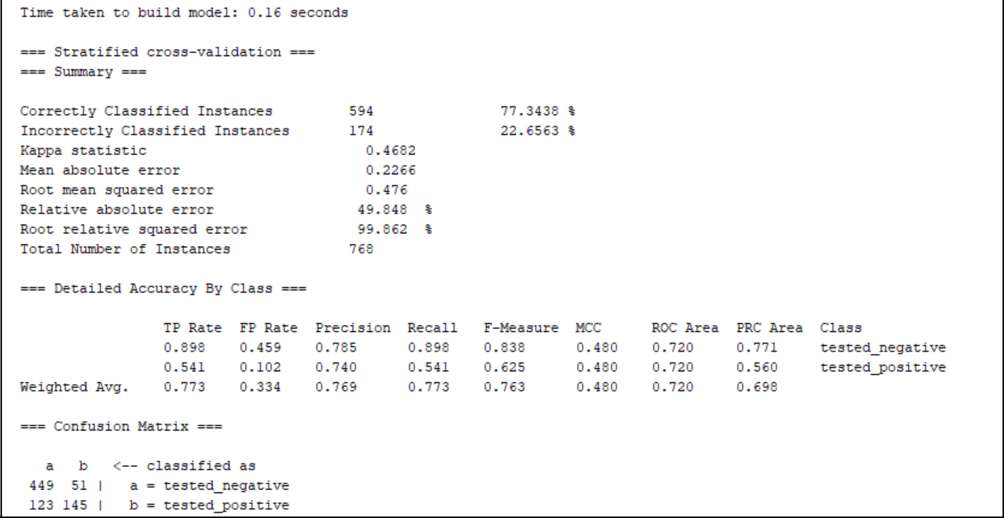
Relative absolute error: 62.5028%

Root relative squared error: 87.4349%

1b) Apply two other classifiers that you can used in WEKA. For this comparison use the default parameters for all the classifiers. Put the result in the table and give the conclusion.

Classifiers	Readings
Decision Tree (J48)	

```
Size of the tree :      39


Time taken to build model: 0.16 seconds

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

Correctly Classified Instances         567               73.8281 %
Incorrectly Classified Instances       201               26.1719 %
Kappa statistic                          0.4164
Mean absolute error                      0.3158
Root mean squared error                  0.4463
Relative absolute error                 69.4841 %
Root relative squared error             93.6293 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.814    0.403    0.790      0.814    0.802      0.417    0.751     0.811     tested_negative
                 0.597    0.186    0.632      0.597    0.614      0.417    0.751     0.572     tested_positive
Weighted Avg.    0.738    0.327    0.735      0.738    0.736      0.417    0.751     0.727

=== Confusion Matrix ===

   a   b   <-- classified as
 407  93 |   a = tested_negative
 108 160 |   b = tested_positive
```

Support Vector Machine	

```
Time taken to build model: 0.16 seconds

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

Correctly Classified Instances         594               77.3438 %
Incorrectly Classified Instances       174               22.6563 %
Kappa statistic                          0.4682
Mean absolute error                      0.2266
Root mean squared error                  0.476
Relative absolute error                 49.848  %
Root relative squared error             99.862  %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.898    0.459    0.785      0.898    0.838      0.480    0.720     0.771     tested_negative
                 0.541    0.102    0.740      0.541    0.625      0.480    0.720     0.560     tested_positive
Weighted Avg.    0.773    0.334    0.769      0.773    0.763      0.480    0.720     0.698

=== Confusion Matrix ===

   a   b   <-- classified as
 449  51 |   a = tested_negative
 123 145 |   b = tested_positive
```

1c) Perform feature selection on WEKA. Used filter method. The filter is called “Attribute Selection” that you have to choose the “Attribute Evaluator” and “Search Method” . Apply the filter and save the filtered dataset with another name (use the Save button).

1d) Load the newly filtered dataset and run the Naive Bayes classifier again. How does it compare to the other classification models in 1b). Is there any changes on the result with and without feature selection on this data.

Answer:

a) The newly filtered dataset has an average of precision and recall which are 0.759 and 0.763 respectively, while for Decision Tree it has an average of precision and recall of 0.735 and 0.738 respectively. For SVM it has an average of precision and recall of 0.769 and 0.773. It can be concluded that, Naïve Bayes has a slightly better readings than Decision Tree but SVM is more accurate than both classifier.

b) The feature selected dataset compared to the previous non feature selected dataset has a no changes.

Question 2

Q2 dataset. Load the weather dataset (weather.arff) into Weka. The dataset already in Weka data folder.

2a). Select the **Cluster** panel and choose SimpleKMeans as clustering method. Use the default number of clusters. Observe the clustering result in the output window. Identify the meaning of Cluster centroids in the cluster result. How would you interpret the clusters produced by this experiment?

2b). Visualise the clustering results by right-clicking the result set on the left "Result list" panel and selecting "Visualize cluster assignments". Screenshot the image and paste here.

2c) Change the number of clusters to 3 (click on SimpleKMeans), and analyse the output. Which clustering is better? What is the “within cluster sum of squared errors”? Find out what the **seed parameter** is about. Why is it important for k-means clustering? Repeat the experiment with different seed values and compare the results.