Big Data

Task 1: What are the attributes can be used to diagnose diabetes and what are the classes? How many samples are there from each class? Produce a screenshot of the bar-chart of the values for the age attribute.

Attributes:

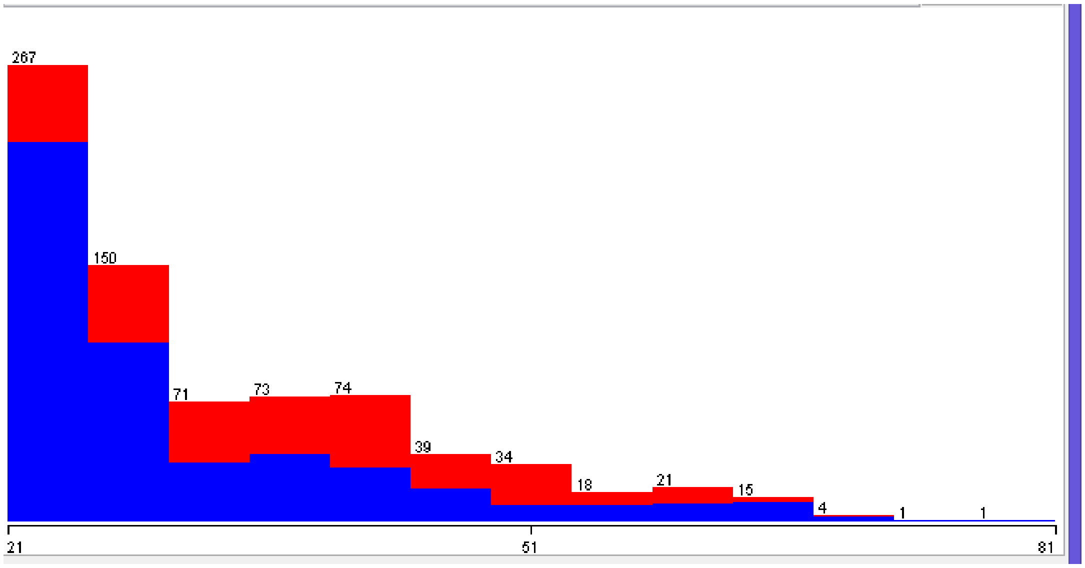
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

[1: National Institute of Diabetes and Digestive and Kidney Diseases]

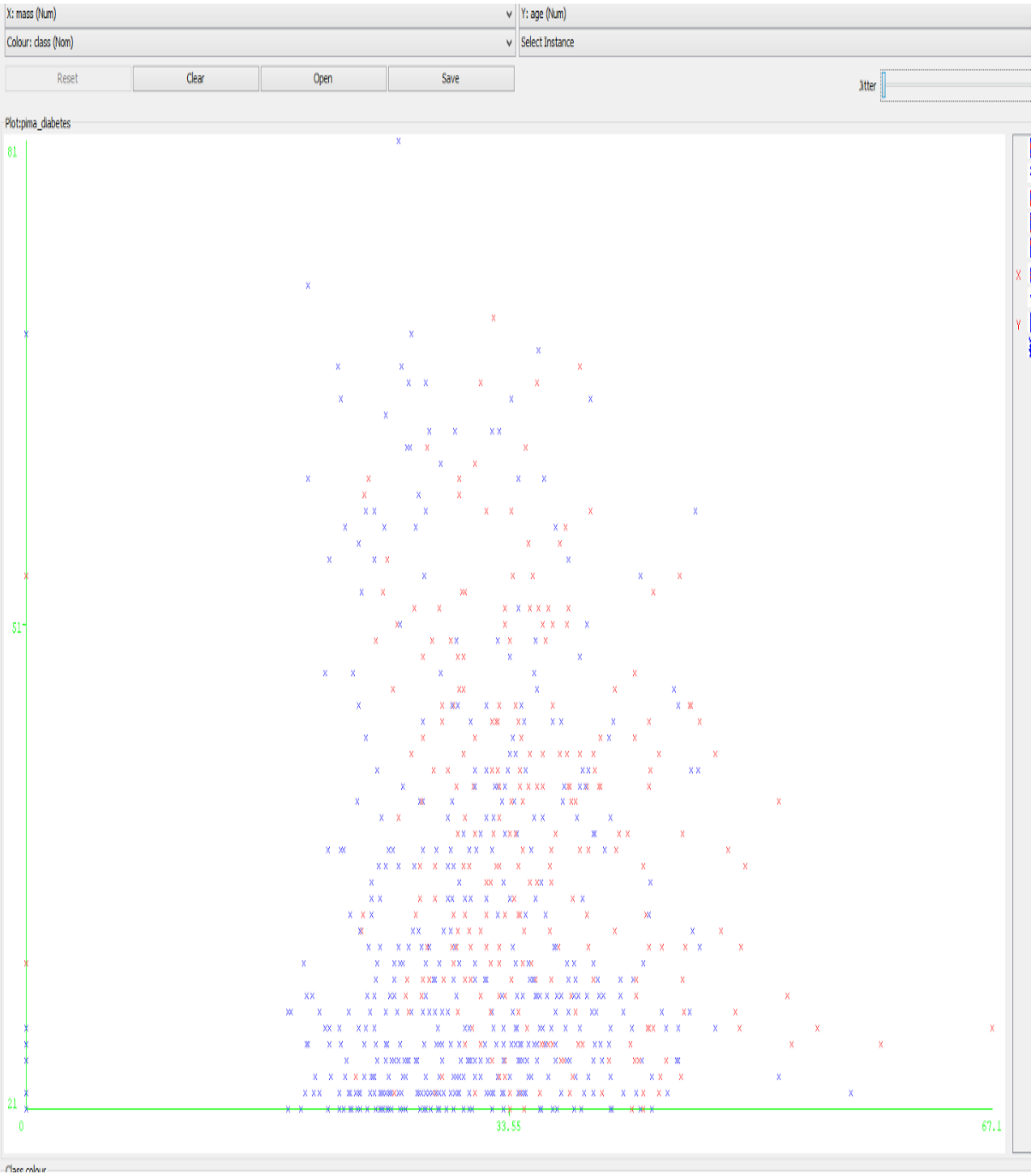
Samples

There are 768 samples for each class.

Age bar-chart



Scatter Graph of age(y) against mass(x)



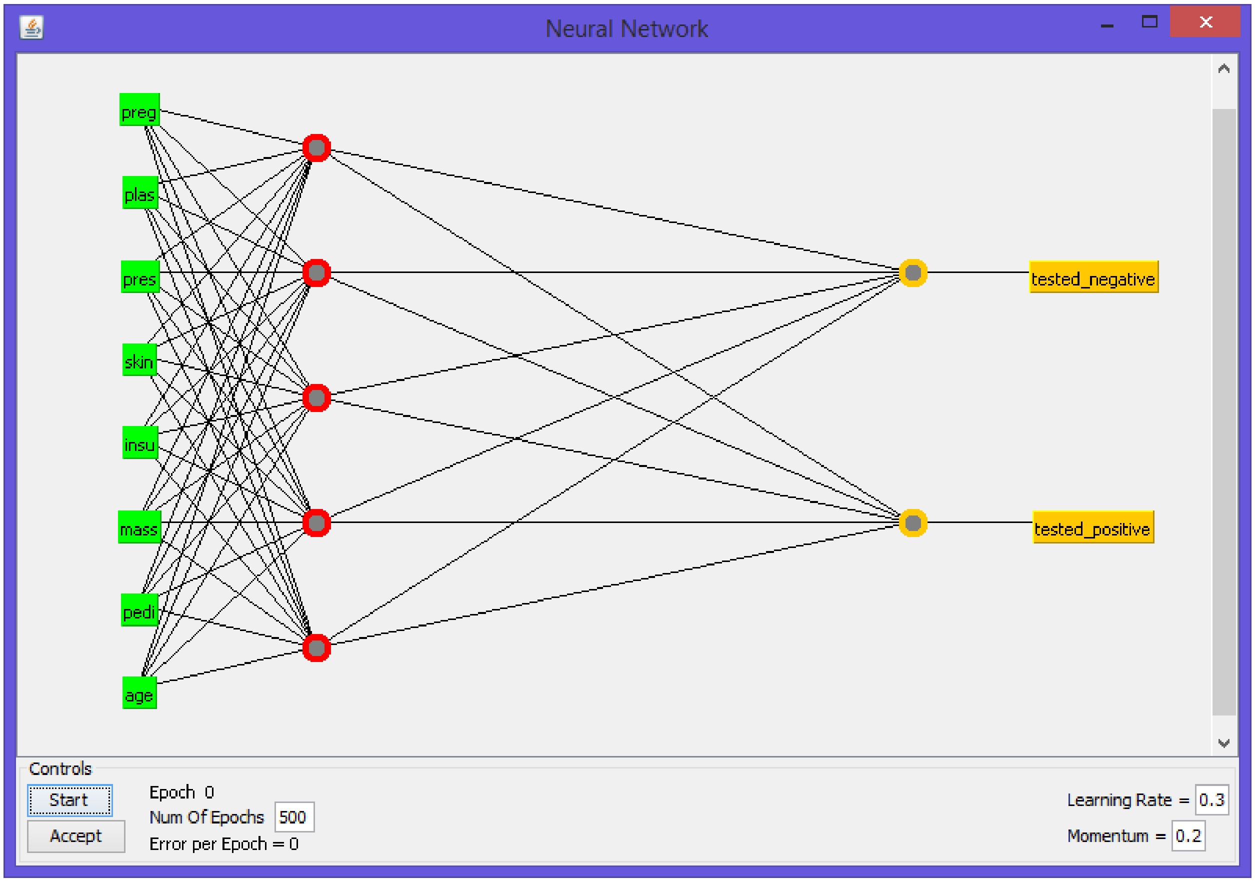
X: mass (Num)

Y: age (Num)

Colour: class (Nom)

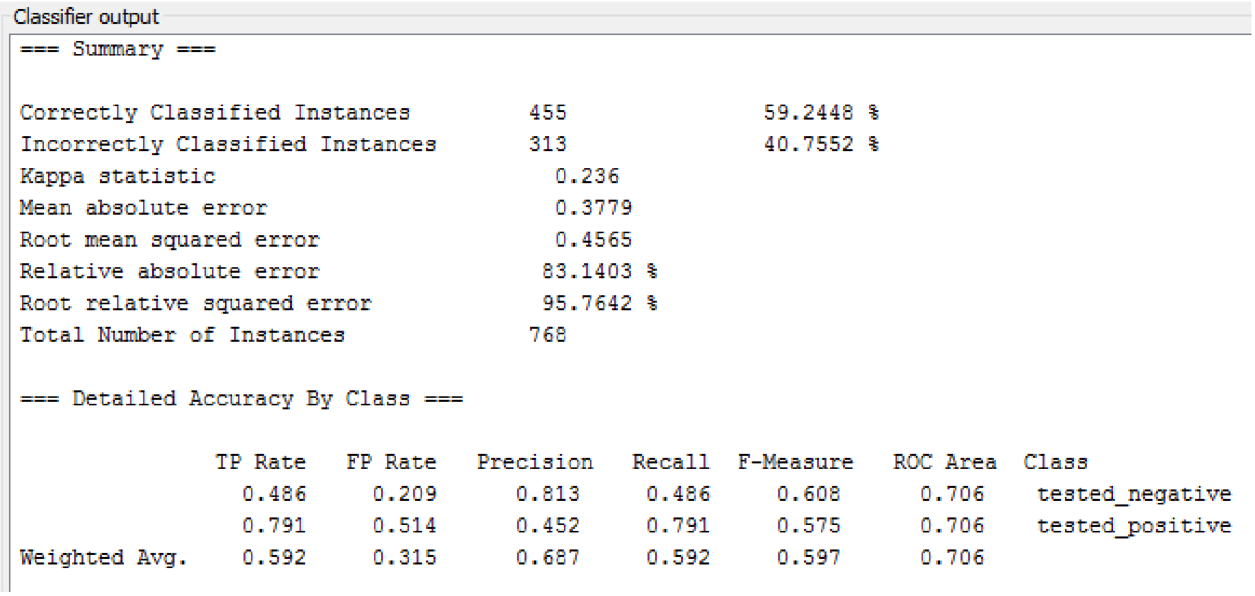
Select Instance

Activity 3: Neural Networks



Epoch 0
Num Of Epochs 500
Error per Epoch = 0

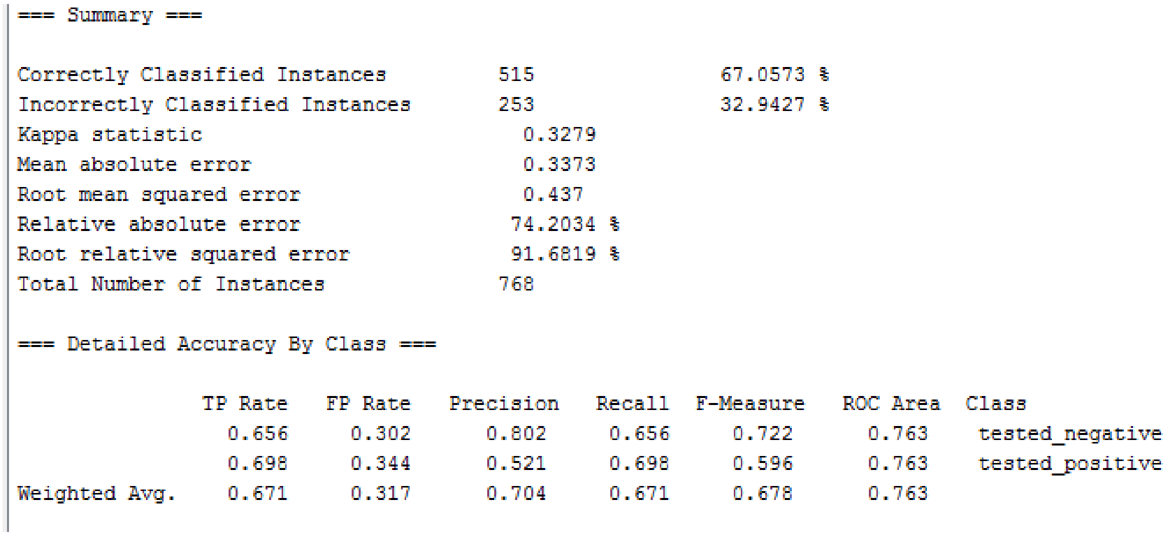
Learning Rate = 0.3
Momentum = 0.2



```
=== Summary ===

Correctly Classified Instances         455               59.2448 %
Incorrectly Classified Instances       313               40.7552 %
Kappa statistic                          0.236
Mean absolute error                      0.3779
Root mean squared error                  0.4565
Relative absolute error                 83.1403 %
Root relative squared error             95.7642 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.486     0.209      0.813     0.486     0.608      0.706    tested_negative
                 0.791     0.514      0.452     0.791     0.575      0.706    tested_positive
Weighted Avg.    0.592     0.315      0.687     0.592     0.597      0.706
```

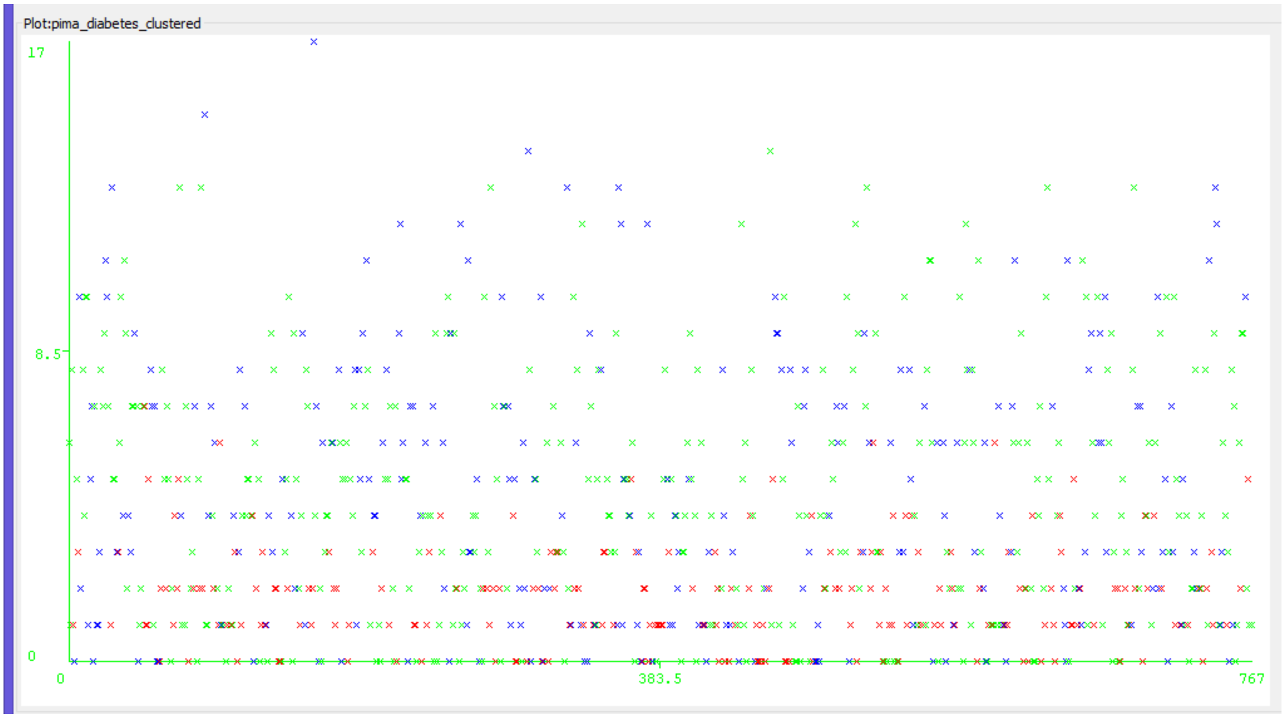
Performance of Network

Hidden: 8
Training time: 200



```
=== Summary ===

Correctly Classified Instances         515               67.0573 %
Incorrectly Classified Instances       253               32.9427 %
Kappa statistic                          0.3279
Mean absolute error                      0.3373
Root mean squared error                  0.437
Relative absolute error                 74.2034 %
Root relative squared error             91.6819 %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.656     0.302      0.802     0.656     0.722      0.763    tested_negative
                 0.698     0.344      0.521     0.698     0.596      0.763    tested_positive
Weighted Avg.    0.671     0.317      0.704     0.671     0.678      0.763
```

Clustering



References

[1.http://storm.cis.fordham.edu/~gweiss/data-mining/weka-data/diabetes.arff]