Discussion Board 1

Daniel Prusinski

**Non-discretized Regular Diabetes Data Set**

Scheme:weka.classifiers.bayes.NaiveBayes

Instances: 768

Attributes: 9

Correctly Classified Instances 586 76.3021 %

Incorrectly Classified Instances 182 23.6979 %

Root mean squared error 0.4168

a b <-- classified as

422 78 | a = tested\_negative

104 164 | b = tested\_positive

F-Measure

0.823 tested\_negative

0.643 tested\_positive

0.76 Weighted Avg.

This is the base measurement that the other experiments will be scored against. Apart from the F-measure, I will be also analyzing the false negative given that this is medical data. In addition, the root mean squared error is also important in that it gauges the error from an overall perspective.

**Discretized Regular Diabetes Data Set**

Scheme:weka.classifiers.bayes.NaiveBayes

Instances: 768

Attributes: 9

Correctly Classified Instances 598 77.8646 %

Incorrectly Classified Instances 170 22.1354 %

Root mean squared error 0.3905

a b <-- classified as

415 85 | a = tested\_negative

85 183 | b = tested\_positive

F-Measure

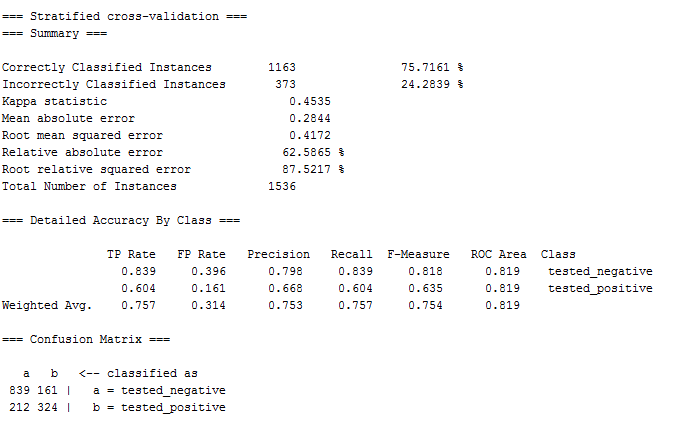
0.83 tested\_negative

0.683 tested\_positive

0.779 Weighted Avg.

The F-measure is better and the Root Mean Square error is smaller, which is a positive sign that the discretization is a positive step. Although, the False Negative increased by 7, which is a 9 percent increase that I am not willing to sacrifice. I would stick with the original data instead of switching to this data.

**Non-discretized Doubled Diabetes Data Set**



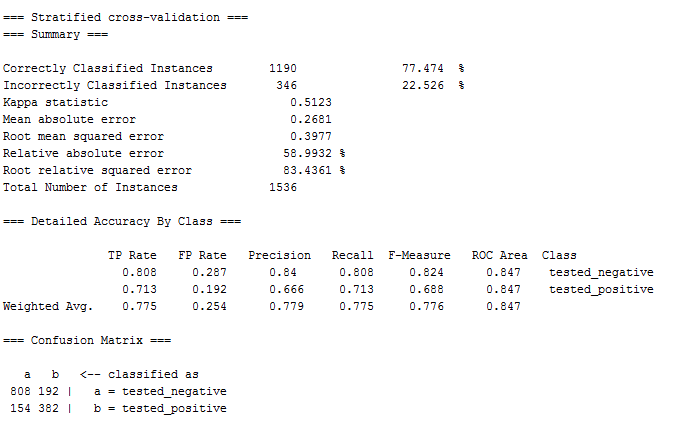
The past F measure was, so this output is consistent but slightly less. The false positives went up more than twice, but it is still within five of double.

0.823 tested\_negative

0.643 tested\_positive

0.76 Weighted Avg.

**Discretized Doubled Diabetes Data Set**



Before doubling the discretized data:

a b <-- classified as

415 85 | a = tested\_negative

85 183 | b = tested\_positive

F-Measure

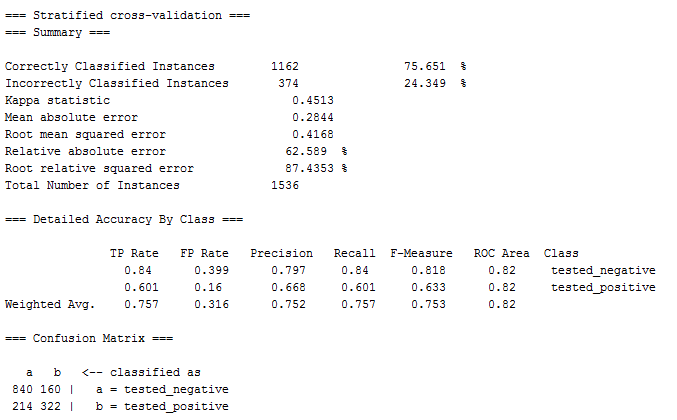
0.83 tested\_negative

0.683 tested\_positive

0.779 Weighted Avg.

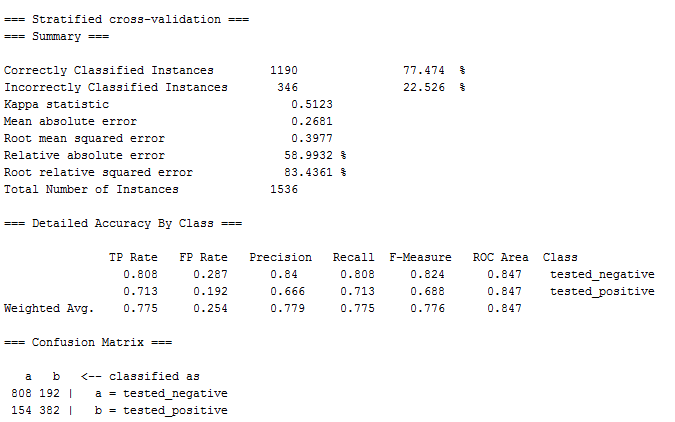
Doubling the data has for the most part maintained its consistency, with the exception of the false negatives. Again, given that this is medical diagnostic data I would not favor this method because 12.5% of the diagnostics are giving the patient an “all-clear” when in fact they have diabetes.

**Non-discretized Doubled Diabetes Data Set with Noise**



Surprisingly, with noise there is one less false negative and the f-measure only slightly decreased. This is an amazing feat given that there is noise in the data.

**Discretized Doubled Diabetes Data Set with Noise**

ggdfdgf

With the noise, the algorithm performs the same as the dataset without the noise. Overall it has been neat to see the consistency of F-measurements. I am still not sold on this discretization because of the increase in the false negatives and diabetes is a disease that is time sensitive to contraction. I would not use discretization over the continuous data.