Diabetes analysis
Weka & Machine Learning

Joel Dos Santos Iraha

s545242@nwmissouri.edu / (203)-300-9996

NORTHWEST MISSOURI STATE UNIVERSITY

Introduction

The Pima Indian population have been heavily studied since 1965 on account of high rate of diabetes. This dataset contains measurements for 768 female subjects, all aged 21 and above. Can we figure out if someone will likely have diabetes just by taking a few of these measurements?

Materials/Mehtods

- Data: Labeled dataset with nominal features. Use of training set as testing option.
- Algorithm: Supervised learning algorithm using Random Tree classifier to visualize classification accuracy.

Results

```
=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         768               100      %
Incorrectly Classified Instances         0                 0      %
Kappa statistic                          1
Mean absolute error                      0
Root mean squared error                  0
Relative absolute error                  0      %
Root relative squared error              0      %
Total Number of Instances              768

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     tested_negative
                 1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000     tested_positive
Weighted Avg.    1.000    0.000    1.000      1.000    1.000      1.000    1.000     1.000

=== Confusion Matrix ===

   a   b   <-- classified as
 500   0 |   a = tested_negative
   0 268 |   b = tested_positive
```

The Random Tree classifier achieved perfect accuracy (100%) on the dataset, with no misclassifications among 768 instances. The model demonstrated ideal precision, recall, and F-measure for both tested_negative and tested_positive classes, resulting in a perfect performance for the predictive model. The confusion matrix confirms accurate predictions, with 500 instances of tested_negative and 268 instances of tested_positive correctly classified.

Conclusion

Based on the evaluation metrics and the structure of the decision tree, the model appears to be highly effective in predicting diabetes within the given dataset. The tree structure provides a clear decision path based on various input features such as pregnancy occurrences, glucose levels (plas), blood pressure (pres), skin thickness (skin), insulin levels (insu), body mass index (mass), pedigree function value (pedi), and age.

Additional Resources

This is a public training dataset offered by Weka.

https://storm.cis.fordham.edu/~gweiss/data-mining/datasets.html

Acknowledgements

Project made for Intro to data science and mining at Northwest Missouri State University

Further Information

https://joeldossantospersonal.github.io/DosSantosJ.github.io/