CENG414-Introduction to Data Mining

Programming HW1

Task 3-Data Visualization

Zeynepsu Kesim

2292340

1. **What does linearly-separable mean?**

   Linearly-separable means that if you have 2 classes, then there is a point, line, plane or hyperplane that divides the input features so that all points of one class are in one half of space and all the points of the other class are in the other half of space without conflicting.

2. **By only using price and maintenance-cost attributes, is it linearly separable?**

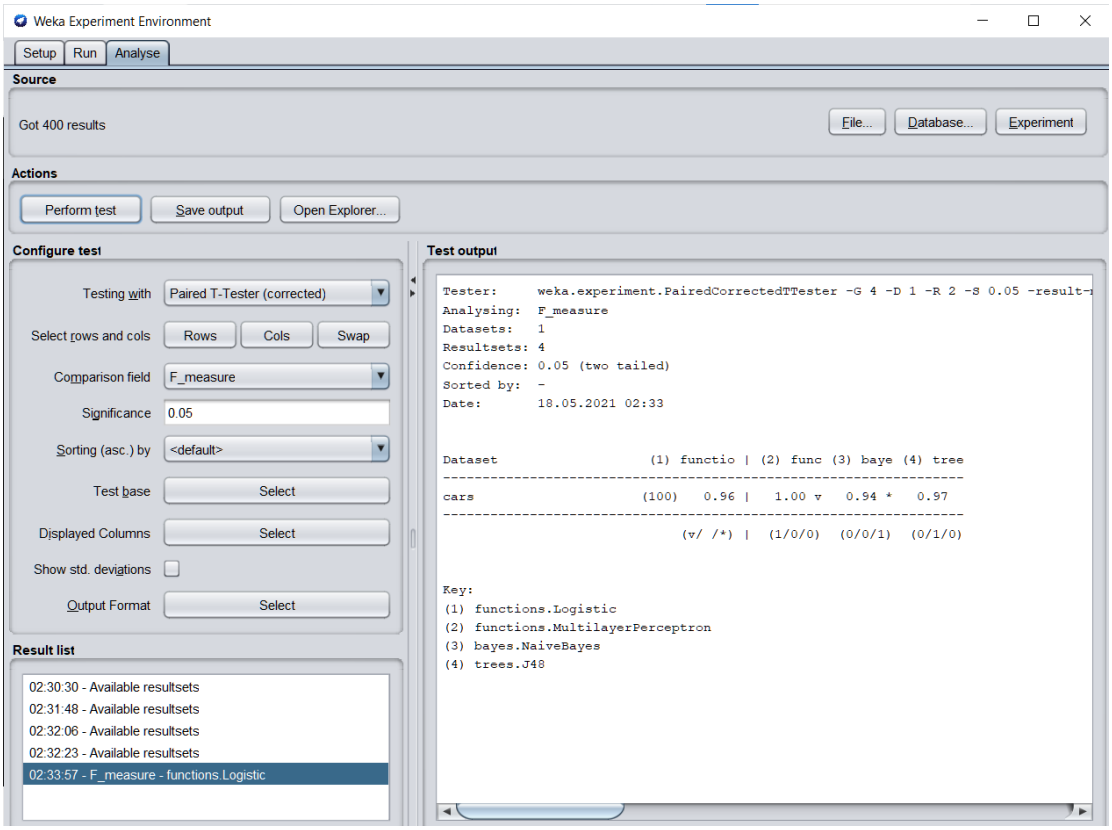
   No, it is not. I removed all the attributes from the preprocessing window except price, maintenance-cost and class. Then I ran two classifiers namely logistic and SMO on training set and the accuracy levels were around 65% which are quite low. And I visualized the plot and saw that the classes are conflicting. After that, I use the simplekmeans clustering technique and EM with using 4 clusters to see if I can obtain pure clusters, but I couldn't.

3. **Is this dataset linearly-separable with all of its attributes? How did you decide?**

   No, it is not. I tried two different classifiers namely logistic and SMO. I evaluated on the training set and see if any misclassifications occur. If no misclassification occur we can say that the dataset is linearly-separable. However, there were misclassifications for both of the classifiers. Moreover, I used cluster techniques and observed whether the clusters are pure. I use the simplekmeans clustering technique and EM with using 4 clusters to see if I can obtain pure clusters, but I couldn't. Same classes occurred in more than one clusters.

4. **Which classifier would you use to classify this dataset? Explain briefly.**

   Naive bayes classifier is mostly better than other models when the training set is small, but here we have a big dataset. Therefore, using naïve bayes is not the best choice.



When we perform a test with 0.05 significance level, we can see that the best one is the second function which is the MLP.

When accuracies are included, the best one is MLP too.

When tested on cross validation 10 folded:

MLP Accuracy → 99.4216 %

Logistic → 93.4066 %

J48 → 91.9028 %

Naïve Bayes → 86.3505 %

Thus, I would use MLP for this dataset.