Knowledge Mining and Big Data Part 1

Pedram Ghazi

Student Number: 267640

1) A) The original sample is randomly divided into 10 equal sized subsamples. From these subsamples, a single subsample is kept as the validation data for testing the model, and the remaining 9 subsamples are used for training data. The cross-validation process is repeated 9 more times and in these iterations every fold is used exactly once as the validation data.

   B) Cross-validation is a model validation technique for evaluating how the results of a statistical analysis can generalize to an independent data set. Its main usage is in settings where the goal is prediction, and we want to estimate how accurately a predictive model will perform in practice. The goal of cross validation is to define a dataset to test the model in the training phase to reduce errors caused by problems like overfitting.

2) A)

Correctly Classified Instances	500	65.1042 %
Incorrectly Classified Instances	268	34.8958 %

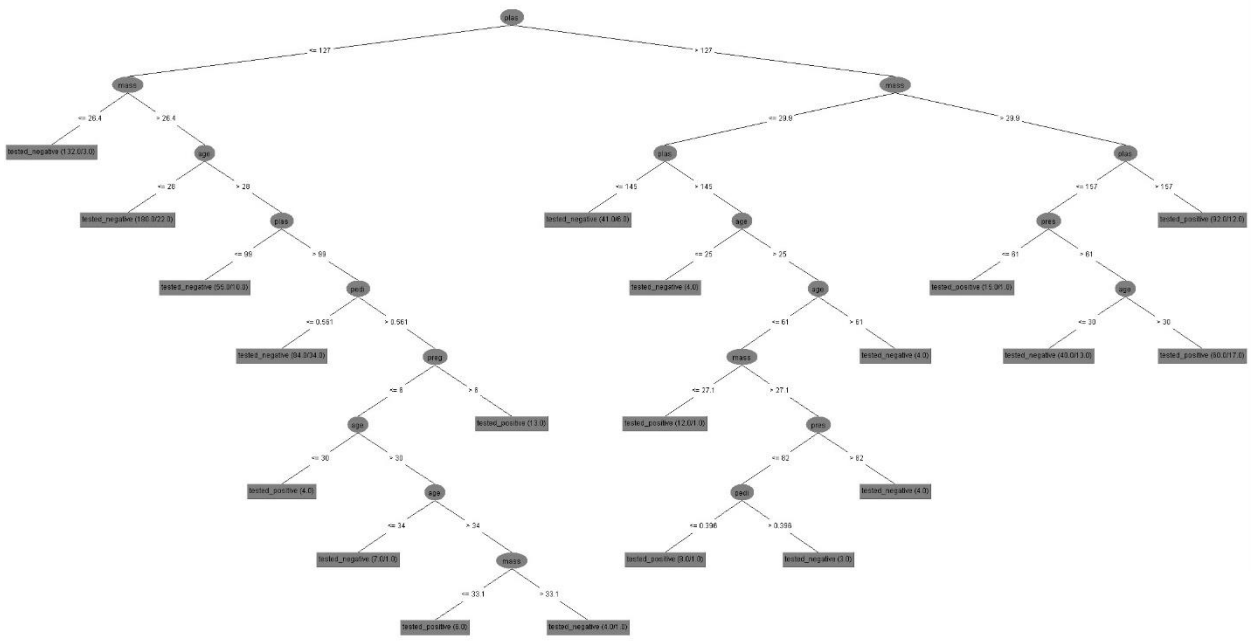
   B)

Correctly Classified Instances	500	65.1042 %
Incorrectly Classified Instances	268	34.8958 %

3) A)

Correctly Classified Instances	646	84.1146 %
Incorrectly Classified Instances	122	15.8854 %

B)



C) tested-positive (60.0/17.0)

D) Correctly Classified Instances 567 73.8281 %
Incorrectly Classified Instances 201 26.1719 %

4) A) Bagging is a machine learning ensemble meta-learning technique which improves the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting. This method achieves this goal by training many classifiers on different partitions of the training data and using the majority vote on the results of all those classifiers to define the final answer for a test pattern.

B) Correctly Classified Instances 729 94.9219 %
Incorrectly Classified Instances 39 5.0781 %

C) Correctly Classified Instances 573 74.6094 %
Incorrectly Classified Instances 195 25.3906 %

5) A) As it was shown by the results for these two classifiers, J48's performance was significantly better as expected since ZeroR is the simplest classification method which relies on the target and ignores all predictors.

B) The results show that bagging J48 classifier performs better(about 10 percent) than normal J48 classifier only when we do not use cross validation. When we utilize cross validation, bagging J48 classifier perform worse. Bagging increases the accuracy by avoiding to overfit.