1. Have a look at both datasets and give a brief explanation on both datasets and what do they describe?

Answer: The cars sales data covers information like car's capacity of taking persons or how much cost needed to maintain the cars. With this data, we can generate a decision tree to determine which attributes make more contribution to evaluate the price for customers to buy a car.

The health sector data set contains patient's information who have thyroid disease, for example whether the patient is pregnant, has tumor or not etc. With this data, we can generate a decision tree to determine some medical results for example which attributes likely to tells that the patient has a problem with his or her hormone.

2. To learn the effect of decreasing training set on the data, we use the first dataset (cars) with the classification algorithm J48 and record the precision results in three different settings:

Answer:

Use 50% of the dataset for training and the other 50% for testing

Accuracy without ID = 17.243%

Accuracy with ID =~ 99%

Use 25% of the dataset for training and the other 75% for testing

Accuracy without ID = 17.593%

Accuracy with ID =~ 99%

Use 10% of the dataset for training and the other 90% for testing

Accuracy without ID = 22.444%

Accuracy with ID =~ 96.656%

3. Use the second dataset (thyroid) with the same algorithm and record the precision results in three different settings:

Answer:

Use 50% of the dataset for training and the other 50% for testing

Accuracy = 96.643%

Use 25% of the dataset for training and the other 75% for testing

Accuracy = 96.857%

Use 10% of the dataset for training and the other 90% for testing

Accuracy = 95.595%

4. What are the effects of the different sizes of training and testing datasets on the learned models and accuracy?

Answer: Normally, the accuracy drops with smaller training set. Moreover, it is not going to reflect that much in the results as the result is come from evenly distributed elements in the dataset.

5. Compare the different precision results between the two datasets and explain the change of accuracy.

Answer: The two datasets give two different types of information and based on their information the accuracy differs. Using the thyroid patient dataset, we can get good accuracy based on the relationship between the datasets. On the other hand, car's dataset and their distribution and correlation doesn't help in the accuracy of their model.

6. For the first dataset, use the safety attribute as a class, how accurate is the model with the same settings (50-50%, 25-75%. 10-90%).

Answer:

With 50-50% : 24.074%

With 25-75% : 27.006%

With 10-90% : 30.418%

7. Why is the model accuracy with the safety attribute different from the one with the price attribute?

Answer: The accuracy of the model depends on the correlation between safety and other qualities. Forecasting, the safety based on the other attributes is somewhat better than price.