Question 6

(i) Dummy Dataset 1:

The tree for the dummy dataset 1 is very simple with only size of 3.

By observing the dataset, we can find that the 5th feature itself discriminates the dataset perfect. Our decision tree noticed this because it asks the question that has the maximal gain.

In the testing set, the rule persists: the label of the data is still completely determined by the 5th feature. Since our decision tree grasped this rule, it got a 1.0 classification rate.

(ii) Dummy Dataset 2:

Its tree size is 11 and the classification rate is 0.65.

The cause for its low classification rate may be that there is not enough training data. Note that our decision tree only uses feature 0, 2, 4, 5, 6 to classify. Maybe there other features have influences to the class of the data which our training set does not represent.

(iii) Car:

Tree size: 408

Average classification rate: 0.9485

The reason why our decision tree performs good on the Car dataset may be that:

(a) The training sample contains 1728 instances and there is only $3^3 \cdot 4^3 = 1728$ possible combination of attributes. The training sample is large.

(b) In our experience, the different features in the Car dataset influence to the class of the data independently. Our decision tree does not need to capture relation between different attributes to perform well.

(iv) Connect4

Tree size: 41521

Classification rate: 0.762

The reason why the size of the tree size for Connect4 is larger than that of Car Dataset is that the Connect4 Dataset has much more attributes and training samples.

Even though there is 67557 instances, it is still relatively small when considering the number of possible setting of the Connect4 game, which is 3^{42}. Another fact that may have effect on the classification rate is that the value of different attributes in Connect4 game have relation with each other. Our decision tree may not capture that very well.

Question 7

(i) A similar dataset may contains information about certain type of products that people bought on a website. For a product like smartphone, it may contain their price, size, brand, and processor type. The label for the dataset could be the product related to smartphones that they probably need in the future. A decision tree based on this dataset can help a website recommend products for their customers.

(ii) We have learned about using MINIMAX algorithm to implement a program that play board game. We can build a more efficient playing bot for Connect4 by combining the MINIMAX algorithm and our decision tree. In the original MINIMAX algorithm, we explore all the situation until reaching the end of the game. With the help of our decision tree, we can set a depth limit and use the result of the game predicted by our decision tree once we reach that limit.