Question 6

DummyDataSet 1
Tree Size : 3
Average Classification Rate: 1.0

DummyDataSet 2
Tree Size: 11
Average Classification Rate: .65

Connect4 Data
Tree Size: 41521
Average Classification Rate: 0.80

Car Data
Tree Size: 408
Average Classification Rate: 0.97

Analysis
The two DummyDataSets have the same amount of instances that are being tested; however they have a very different average classification rates. At first I was confused as to why this would happen because of the fact that the tree is essentially testing the same data. However, one of the biggest difference in the two models is the fact that the tree size for the second dataset is almost 4 times as large as the tree size for the first dummy data set. This shows that a larger tree does not necessarily result in a more accurate classification rate. This phenomenon is called overfitting, and between these two datasets we can see the second model overfitted the data and therefore did not do very well in testing. The testing dataset also differs in the different possibilities that the outputs can take. There are only two possible classes for the first dataset while there are 2^10 different outputs for the second dataset.

For Connect4 dataset, the tree has a size of around 41521 and an average classification rate of around 80%. The dataset for connect 4 is very large and it has a large amount of possible classes. It is also very hard for a decision tree to predict the outputs for this dataset since each instance could be part of any number of classes. Also in connect 4 each datapoint is as important as others since each place you put the token is equally important. Since all attributes of an instance are important the information gain from each branch of the tree is very low and this will negatively affect the performance of the model.

In comparison, the car dataset has much less instances of data but still performs better. This could be due to the fact that it posses less attributes per instance as well as the fact that not every attribute for each instance is as important. This means that this decision tree posses more overall information gain in the model. This will, therefore, result in a much higher accuracy than the other real dataset.

Question 7

Connect4

For connect4, it is better to use the K-Nearest Neighbors algorithm to classify this dataset since in KNN all attributes are treated as equally in the classification. This is perfect for Connect 4 since in connect 4 all attributes are independent.

Cars

This model is perfect to combine with a website to help predict user behavior to see what kind of car people prefer when searching for one.

Question 8 (Extra Credit)

Balloon Dataset

Examples: 20
Tree Size: 9
Average Classification Rate: 0.67

The balloon dataset does not have a lot of examples, however it still has a decently large average classification rate. One of the more surprising aspects of the model is that the tree size is still somewhat large for the amount of examples in the data. The attributes of this dataset are not independent, because there seems to be a strong relationship between the size and act of the balloon and whether or not the balloon is inflated or not. This is why I believe that the lower size of the dataset does not affect the accuracy score that much.