Nick Liccini
CS 3600 – Hrolenok
Project 4 Analysis

Files in this submission:
DecisionTree.py
DataInterface.py
runExtraCredit.py
heart-disease-data.txt
analysis.pdf

Results: <dataset>: (classification rate, tree size)
Dummy Set 1: (1.0, 3) for 20 examples
Dummy Set 2: (0.65, 11) for 20 examples
Cars Dataset: (0.95, 408) for 1728 examples
Connect4 Dataset: (0.78, 2089) for 4000 examples

Question 6:
For the dummy datasets, the decision tree performed relatively well. Dummy Set 1 had a classification rate of 1.0 with a tree size of 3; Dummy Set 2 had 0.65 and 11 respectively. Dummy Set 1 had a flawless classification rate with a very small tree size, most likely due to the diversity of the training examples, allowing the algorithm to see a variety of possible data thus making it more accurate in classifying test examples. Dummy Set 2 had a relatively good classification rate and a small tree size; this was probably due to a lower degree of diversity in the training examples (that is, the examples were more similar to each other, or had the same attribute values among different examples). Since the dummy datasets used binary data, it is more likely to have repeated attribute values.

For the real datasets, the decision tree performed differently, but still relatively well. The Cars dataset has a classification rate of 0.95 and a tree size of 1728; the Connect4 dataset had 0.78 and 2089, respectively (the Connect4 dataset was tested using only 4000 examples because the testing methods reached a recursion limit when there was no limit to the number of examples). The Cars dataset had a very good classification rate and a relatively small tree size (compared to the number of examples and attributes), most likely due to the high number of examples in the training set. This would allow the algorithm to see a lot of different outcomes and thus be prepared for new test data, also since the attributes had non-binary values, they were more helpful in building the decision tree. The Connect4 dataset also had a good classification rate and a relatively small tree size (compared to the number of examples and attributes); this was probably due to having more attributes with a smaller number of attribute values. Since there are many different combinations of attribute values, it is harder for the algorithm to know every outcome and these combinations produce a bigger tree.

Question 7:
For the Car dataset, this classification algorithm could be applied to a used car dealership's website. When a new car comes in the lot, the dealership can fill in the attributes of that car and the decision tree can determine whether that car is valuable (in this case, is its condition acceptable). If the tree produced an outcome of 'vgood' with a 0.95 classification rate, it's very likely that the car is worth selling!

For the Connect4 dataset, this classification algorithm could be applied to an intelligent agent to improve the performance, or win rate, of a Connect4 playing bot. For example, assume an agent uses reinforcement learning (Q-learning) to become better at playing Connect4. If this agent incorporated a decision tree using this dataset, the agent would be able to make more informed, even high reward, exploratory moves instead of choosing a random action given the state of the game!

Question 8:
Heart Disease Dataset: (0.47, 3908) for 303 examples

The Heart Disease dataset was taken from the UCI Machine Learning Repository using the following link: https://archive.ics.uci.edu/ml/datasets/Heart+Disease with the source data coming from the 'processed.cleveland.data'. The outcome classifications are integers ranging from 0 (absence of heart disease) to 4 (presence of heart disease). The dataset included 303 examples, with 14 attributes (some discrete, some continuous).

The Heart Disease dataset had a classification rate of 0.47 and a tree size of 3908. This classification rate is quite low, and the tree size is quite large for such a dataset. This is most likely due to the high number of continuous attributes and the low number of training examples. With so many different possible combinations of data, a very high number of examples would be needed to produce a higher classification rate.

For the Heart Disease dataset, this classification algorithm could be applied to a hospital system that needed to determine what course of action to take for a patient who potentially has heart disease. Perhaps a man is experiencing chest pain, so he goes in to the hospital and the staff insert his attributes into the decision tree. If the decision tree was trained properly and had a high classification rate, the outcome could be used to determine how immediately the man needs medical attention.