Anmol Chhabria
902931430
CS 3600
Project 4a

Analysis: Document has analysis for Q6, Q7, Q8

Question 6
Performance

- **Dummy Dataset 1**
  - Tree Size: 3
  - Classification Rate: 1.0
  - Examples (dataset instances): 20
- **Dummy Dataset 2**
  - Tree Size: 11
  - Classification Rate: 0.65
  - Examples (dataset instances): 20
- **Connect4 Dataset**
  - Tree Size: 41521
  - Classification Rate: 0.762850
  - Examples (dataset instances): 67557
- **Cars Dataset**
  - Tree Size: 408
  - Classification Rate: 0.944000
  - Examples (dataset instances): 1728

Explanation
The classification rate for the dummy dataset 1 is 100%, whereas the classification rate for dummy dataset 2 is 65%. Both have the same number of instances, 20, however, their tree sizes are different. The reason that dummy dataset 2 has a much lower classification rate could be because of the size of the tree. A large tree can result in overfitting, which reduces the ability of the model to generalize the data and results in lower classification rates. Thus, dummy dataset 2 with a tree size of 11 results in 65% classification rate, while dummy dataset 1 with a tree size of 3 results in a 100% classification rate.

The other reason for the better performance of dummyData1 could be because for this dataset when DummyData1 has fifth attribute equal to 0, the class is 1, and vice versa. This clearly allows to build a reliable classifier. On the other hand, for dummy dataset 2, this is not the case and there are 1024 (2^10) possible combinations. It is not possible to understand how all the combinations affect the class label because there are only 20 instances in the data.

Tree Dummy Data1 is as follows:
5 = 0
|---1
5 = 1
|---0

Testing dummy dataset 2. Number of examples 20.
Tree is as follows:
2 = 0
|---0 = 0
|---|---0
|---0 = 1
|---|---4 = 0
|---|---|---1
|---|---4 = 1
|---|---|---0
2 = 1
|---5 = 0
|---|---6 = 0
|---|---|---0
|---|---6 = 1
|---|---|---1
|---5 = 1
|---|---1

The Connect4 dataset has a tree size of 41521 and a classification rate of 76%. The decision tree algorithm does not perform as well on the Connect4 dataset because all the attributes are equally important, there isn't a specific location on the Connect4 board that can predict a win or loss. Decision trees generally perform better on datasets in which specific attributes result in more information gain, or are significantly more important, than others. Moreover, Connect4 has 43 attributes, which each have 3 combinations. It is not possible for the classifier to undersant all ~80,000 of these combinations in 41,000 instances.

The cars dataset has less instances than the connect4 dataset, but it still results in a better classification rate of 94%, compared to 76%. The decision tree performs well on the car dataset because some attributes are more important than others. The cars dataset has 6 attributes, and depending on the overall data gathered, some attributes will result in more information gain, as opposed to the connect4 dataset attributes. With the cars dataset, we also notice that there are 1728 combinations possible of all the attributes, and the dataset has exactly 1728 instances. Therefore, the classifier is better able to label the test data with more overall data.

Question 7
Applications

Cars
Websites can analyze the user preferences and compile this data to give user preferences. The cars dataset can be used to analyze buyer behaviors and be able to predict car types that will be suitable to a specific buyer.

Connect4

The algorithm that this can be used with is KNN. Connect4 dataset has attributes that are independent from each other. In Connect4 having a token in any particular location does not imply a win. Therefore, no attribute is more important than another one. Since KNN is an algorithm that is preferred when the attributes are independent from each other, Connect4 data would perform well with it.

Question 8

Acute Inflammations Dataset (Extra Credit)

- Tree Size: 18
- Classification Rate: 0.983333
- Examples: 120

Despite the low number of examples, the classification rate is very high for the dataset. Revising the dataset, it can be observed that when one of the attributes ('urine pushing') is always 'yes', then the class value is always 'yes'. Moreover, the possible values for the other 5 attributes other than temperature_of_patient are only 'yes' and 'no'. Therefore, all the possible combinations of those attributes are 32, and the dataset is 120 instances, so with all the instances every combination is covered multiple times.

urine_pushing = yes
|---micturition_pains = yes
|---|---yes
|---micturition_pains = no
|---|---burning = yes
|---|---|---temperature_of_patient = 39
|---|---|---|---no
|---|---|---temperature_of_patient = 38
|---|---|---|---no
|---|---|---temperature_of_patient = 42
|---|---|---|---lumbar_pain = yes
|---|---|---|---|---occurrence_nausea = yes
|---|---|---|---|---|---no
|---|---|---|---|---occurrence_nausea = no
|---|---|---|---|---|---no
|---|---|---|---lumbar_pain = no
|---|---|---|---|---no
|---|---|---temperature_of_patient = 37
|---|---|---|---no
|---|---|---temperature_of_patient = 36
|---|---|---|---no
|---|---|---temperature_of_patient = 40
|---|---|---|---no
|---|---|---temperature_of_patient = 41
|---|---|---|---no

|---|---burning = no
|---|---|---yes
urine_pushing = no
|---no

Acute Inflammations Data Set: The data was created by a medical expert as a data set to test the expert system, which will perform the presumptive diagnosis of two diseases of the urinary system. This can be used to predict if someone has an inflammation of the urinary bladder. After asking a patient all the questions corresponding to the attributes, even a program can output an accurate class label after it has been trained well.