**CS 460G Homework 1 Report**

**Program Usage Instructions:**

1. Place all homework CSV files in the same directory as the script decision.py.
2. You will be tasked to enter a filename; this can be any CSV file included (e.g. synthetic-1.csv)
3. When tasked to “enter value for prediction”…
   1. For synthetic data, you may enter any combination of these:
      1. rainy
      2. sunny
      3. 50s
      4. 60s
   2. For game data, you may enter the combination of any existent value in any column
   3. Type stop and hit enter to stop collecting guess data.
4. The program will make a prediction based on the created tree and print it out to you. It will also print the training set error.
5. The program will then ask you if you want to print the tree out. Type yes or hit enter.

**Data Storage & Tree**

* The data is read in as a CSV reader and is cast to a list. The list ends up being a list of lists.
* My tree is stored as nested dictionaries. It is traversed recursively when finding the result of a prediction.

**Synthetic Data Report:**

* For the synthetic data discretization, I chose the number of bins to be 2. It is possible to change this value anywhere from 1 through 5. I found the best way to deal with the data was to initially discretize before doing anything else, and I did this using the weather labels “rainy”, “sunny”, “overcast”, “snowy”, and “clear” for the first column of data. For the second column, I based these values off of temperature: “50s”, “60s”, “70s”, “80s”, and “90s”. I thought of the last data set as “do/do not play tennis”: 0 is do not, 1 is do. This was all to help me visualize the tree in my head.
* The discretization method used was equidistant bins. I found the max and min values of each column and created bins dynamically based on those values (e.g. 0 being min and 10 being max and bin count being 5, bins would categorize like this: 0-2, 2-4, 4-6, 6-8, 8-10). Once again, the number of bins can be changed from 1 to 5 and everything still works just as well.
* Training set errors:
  + Synthetic 1: 0.0
  + Synthetic 2: 0.115
  + Synthetic 3: 0.165
  + Synthetic 4: 0.385

**Video Game Data Report:**

* No data was discretized in the video game sales CSV.
* Training set error: 0.903