THE GEORGE WASHINGTON UNIVERSITY
WASHINGTON, DC

Introduction to Data Mining
DATS 6103
Date: *April-14-2020*

Exam 2

- **Show ALL Work, Neatly and in Order.**
- **No** credit for Answers Without Work.
- Open Books, Open Notes.
- Use Your Own Paper. Submit one PDF file to BB.
- Write Only on the Front of Each Sheet.

Question 1: 2.5 pts

Assume you are living in an remote island. The only food which you can eat are mushrooms. There are various type of them are existed. By looking at different features of mushrooms you will determined that it edible or not. This comes information comes with your pervious experience. You recorded the features and your experience. The following table is your record.

Example	NotHeavy	Smelly	Spotted	Smooth	Edible
A	1	0	0	0	1
B	1	0	1	0	1
C	0	1	0	1	1
D	0	0	0	1	1
E	1	1	1	0	1
F	1	0	1	1	1
G	1	0	0	1	1
H	0	1	0	0	1

i.What is the entropy edible?

ii. Which feature should you choose as the root of a decision tree? Hint: You can find out this by looking at each feature and draw the nodes.

iii. What is the information gain of the attribute you chose in the previous question?

Question 2 - Python Decision Tree: 10 pts

The file weather.csv is a csv file that contains weather data. We collected this data from a weather station in Washington DC. Station has sensors that are capable of recording weather measurements such as air temperature, air pressure, and relative humidity. We collect this dat over several years. We have sufficient data samples in this set.

Answer ALL questions in Q2.py. Please comment your result in the same py file if needed. Answer the questions carefully.

Question 3 - Python Support Vector Machine: 10 pts

We use the same dataset in question 2.

Answer ALL questions in Q3.py. Please comment your result in the same py file if needed. Answer the questions carefully

Question 4: 2.5 pts

Compare your best results in Q2 and Q3 answer the following questions.
i. Can we definitively say one model will perform better on the other model? If yes why, If no Why?

ii. Explain the difference between this two models? You can use math or draw diagrams? Be specific

iii. Which metric is the best way to explain the results? Please explain why?