IDS 472, Spring 2021

Homework 1

Due date: Tuesday, February 2 (11:59pm)

Please provide answers to the questions below. For full points, please show the intermediary steps you took to arrive at your answers. You should submit an electronic pdf or word file in blackboard.

1. A survey was sent to UIC Business students which asked them the following questions:

   - Do you typically spend more than an hour per day on *TikTok*? (Y or N)
   - Do you typically spend more than an hour per day on *Instagram*? (Y or N)
   - Do you typically spend more than an hour per day on *Facebook*? (Y or N)
   - Are you an upperclassman (i.e., have you spent three or more years in college)?

   The following table summarizes the responses to the survey. For each entry, "Number of Instances" represents the number of respondents having the corresponding values for the attributes *TikTok*, *Instagram*, and *Facebook*.

TikTok	Instagram	Facebook	Number of Instances	
			Upperclassman	Not Upperclassman
Y	Y	Y	5	0
N	Y	Y	0	20
Y	N	Y	20	0
N	N	Y	0	5
Y	Y	N	0	0
N	Y	N	25	0
Y	N	N	0	0
N	N	N	0	25

   (a) Find the support and confidence for the rule

   if *Instagram* = N then class = *Not Upperclassman*.

   (b) Find the support and confidence for the rule

   if *TikTok* = N and *Facebook* = N then class = *Upperclassman*.

   (c) Using the 1-rule method discussed in class, find the relevant sets of classification rules for the target variable by testing each of the input attributes *TikTok*, *Instagram*, and *Facebook*. Which of these three sets of rules has the lowest misclassification rate?

2. In a few sentences, explain what overfitting is and how we can usually handle overfitting in a classification model?

3. Rebecca works at *Croissants4Ever*, a internet startup that ships freshly-baked croissants to customers in various geographic regions in Illinois and surrounding states. Motivated by the surge of internet sales due to the pandemic, the startup is looking to expand into new geographic regions. Rebecca is tasked with building a model which can identify geographic regions in which *Croissants4Ever* will be profitable.

   Rebecca compiled a training dataset (see table below) comprised of eleven geographic regions that are currently served by the company.

avg. income	pop. density	# competitors	profitable
low	rural	2	yes
low	suburban	0	no
low	suburban	0	yes
low	urban	0	no
med	suburban	2	no
med	suburban	0	yes
med	urban	2	no
med	urban	1	yes
high	suburban	0	yes
high	rural	2	no
high	urban	1	yes

   (a) Considering "profitable" as the target variable, which of the attributes would you select as the root in a decision tree that is constructed using the information gain impurity measure?

   (b) Use the Gini index impurity measure and construct the full decision tree for this data set.

   (c) Consider the following set of points as your test data.

avg. income	pop. density	# competitors	profitable
low	suburban	2	yes
med	suburban	0	yes
med	suburban	2	yes
med	suburban	1	yes
low	rural	1	no
med	suburban	0	no
high	suburban	0	yes
med	urban	1	yes

   What is the accuracy of your decision tree built in part (b) on the test data?