ELE 364: Assignment #1

1. (10 pts) Consider the alphabetical sequence:

   $\{M, A, C, H, I, N, E, L, E, A, R, N, I, N, G, A, L, G, O, R, I, T, H, M, S\}$.

   (a) Compute the corresponding entropy (in bits).

   (b) If this set is split into two sets of vowels and consonants, what would the information gain be?

   (c) Suppose you are to play Scrabble with these letters, would you prefer the set to have a high entropy or low entropy? Justify your answer.

2. (10 pts) The following table shows a dataset containing details of five participants in a heart disease study. The descriptive features are: (i) EXERCISE: how regularly do they exercise, (ii) SMOKER: do they smoke, (iii) OBESE: are they overweight, and (iv) FAMILY: is there a family history of disease. The target feature is Risk that describes their risk of heart disease.

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

   (a) Build a random forest predictive model for heart disease based on the three bootstrap samples given below. Use Gini index based information gain for feature selection.

   (b) Assuming your random forest model uses majority voting, what prediction will it return for the following query: EXERCISE = weekly, SMOKER = false, OBESE = true, FAMILY = yes.

Bootstrap Sample A				Bootstrap Sample B				Bootstrap Sample C			
ID	EXER.	FAM.	RISK	ID	SMOKER	OBESE	RISK	ID	OBESE	FAM.	RISK
1	daily	yes	low	1	false	false	low	1	false	yes	low
2	weekly	yes	high	2	true	false	high	1	false	yes	low
2	weekly	yes	high	2	true	false	high	2	false	yes	high
5	rarely	no	high	4	true	true	high	4	true	yes	high
5	rarely	no	high	5	true	true	high	5	true	no	high

3. (10 pts) Suppose a dataset of size n has been given to us. We induce a decision tree with it. It has a total error of ϵ in the set of predictions it makes for the instances in the dataset. A weighted dataset is then created for boosting. Prove that the sum of all the weights in this weighted dataset is 1.0.

4. (10 pts) Compute the probability of a model ensemble, which uses simple majority voting, making a correct prediction in the following scenario: the ensemble contains seven independent models, all of which have an error rate of 0.1.

5. (20 pts) **Coding project**

   In this project, you will train classifiers based on decision trees to identify types of glass. The dataset consists of 214 data instances of glass with the following nine numerical descriptive features per instance: refractive index, sodium, magnesium, aluminum, silicon, potassium, calcium, barium, and iron.

   The target class of each glass instance is one of six glass types: float processed building window, non-float processed building window, float processed vehicle window, container, tableware, and headlamp.

   The data will be divided into a training set and a validation set. You will start with training the default decision tree classifier in the Scikit-learn library and then train a decision tree classifier that is pre-pruned based on depth. You will also train an ensemble of decision tree classifiers using bagging and boosting.

   Once you have built these classifiers, you will evaluate them to find the one with the best performance.

   Please refer to the Jupyter notebook for more details.

   GitHub repository for ECE364 coding projects: https://github.com/JHA-Lab/ece364_2022