ECE 364: Assignment #1

1. (10 pts) Consider the following dataset:

ID	AGE	EDUCATION	OCCUPATION	INCOME
1	28	MS	teacher	25-50
2	30	BS	professional	75-100
3	42	PhD	professional	100-150
4	28	BS	farmer	75-100
5	32	BS	teacher	25-50
6	64	PhD	professional	100-150
7	50	PhD	professional	75-100
8	37	MS	farmer	25-50
9	42	BS	teacher	75-100
10	70	MS	professional	100-150

   (a) Calculate the entropy and Gini index of this dataset using the income target feature.

   (b) Calculate the information gain for the education feature using entropy.

   (c) Calculate the information gain ratio for the education feature using entropy.

   (d) Calculate information gain for the education feature using the Gini index.

2. (10 pts) We would like to construct a decision tree for a dataset consisting of n data instances and m descriptive features.

   (a) Assume that there exist descriptive features i and j such that for each data instance in the training dataset, these features have identical values (i.e., the two descriptive feature columns, say $\mathbf{d}[i]$ and $\mathbf{d}[j]$, are identical). Assume that we break ties between them by using $\mathbf{d}[i]$ (that is, if both lead to the same remainder entropy, we would use $\mathbf{d}[i]$). Can removing descriptive feature $\mathbf{d}[j]$ from our training dataset change the decision tree we learn for this dataset? Explain briefly.

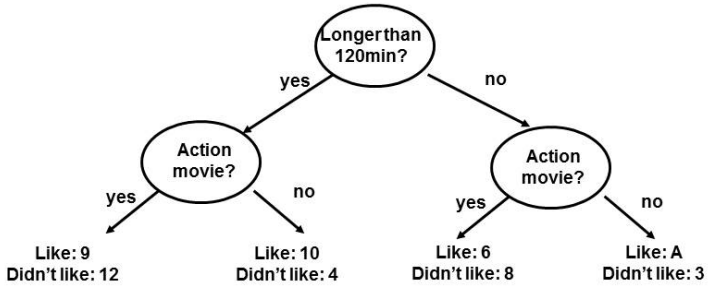
   (b) Assume we have two equal data instances, $\mathbf{d}_i$ and $\mathbf{d}_j$, in our training dataset (that is, all descriptive feature values of $\mathbf{d}_i$ and $\mathbf{d}_j$ including the labels are exactly the same). Can removing $\mathbf{d}_j$ from the training dataset change the decision tree we learn for this dataset? Explain briefly.

   For the next set of questions consider a dataset with continuous descriptive features. For such features, we can use threshold values to determine the best partition for a set of data instances. Assume we are at the root node and we have n data instances, all with different (continuous) values for descriptive feature $\mathbf{d}[k]$.

   (c) Assume we would like to use binary partitions at each node in the decision tree. For such splits, we need to choose a value a and partition the dataset by including all data instances with $\mathbf{d}[k] < a$ to the first and those with $\mathbf{d}[k] \geq a$ to the second partition. For any value of a we consider, we would like to have at least one data instance assigned to each of the two partitions. How many values of a do we need to consider in the worst case?

(d) Assume we would like to use three-way partitions at each node in the decision tree. For such partitions, we need to choose values a and b such that $a < b$ and partition the data into three sets: $\mathbf{d}[k] < a$, $a \leq \mathbf{d}[k] < b$, and $\mathbf{d}[k] \geq b$. Again, we require that for any value of a and b that we consider, at least one data instance should be assigned to each of the three partitions. How many $\{a, b\}$ pairs do we need to consider in the worst case?

3. (10 pts) The following figure presents the top two levels of a decision tree learned to predict the attractiveness of a movie. It has descriptive feature Length at the root node that denotes the length of the movie and descriptive feature Action at the second level. The target feature takes two values: Like, Didn't Like. What should be the value of A if the decision tree was learned using entropy-based information gain. You can either say 'At most X' or 'At least X' or 'Equal to X' where you should replace X with a number based on your calculation. Explain your answer?



4. (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

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

(a) Build a random forest predictive model for heart disease based on the three bootstrap samples given above. 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.

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

For this project, you will train classifiers based on decision trees to determine whether a patient has heart disease.

The dataset consists of the following set of descriptive features:

(a) numeric: age, resting blood pressure, serum cholesterol, max. heart rate achieved, level of exercise-induced ST segment depression, number of major vessels colored by flouroscopy.

(b) categorical: sex, chest pain level, whether the fasting blood sugar > 120 (mg/dl), resting ECG type, whether the patient suffers from exercise-induced angina, slope type of the peak exercise ST segment, type of thalassemia.

We will only use the numerical descriptive features due to implementation constraints.

The target feature is a binary variable, where 1 indicates the presence of heart disease.

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.

See the Jupyter notebook for more details.

GitHub repository for ECE364 coding projects: https://github.com/JHA-Lab/ece364_2025