Exercise sheet: Decision trees and ensemble methods

1. The table below lists a sample of data from a census There are four descriptive features and one target feature in this dataset: AGE, EDUCATION, MARITAL STATUS and OCCUPATION. The target feature is the ANNUAL INCOME.

ID	AGE	EDUCATION	MARITAL STATUS	OCCUPATION	ANNUAL INCOME
1	39	bachelors	never married	transport	25K-50K
2	50	bachelors	married	professional	25K-50K
3	18	high school	never married	agriculture	≤25K
4	28	bachelors	married	professional	25K-50K
5	37	high school	married	agriculture	25K-50K
6	24	high school	never married	armed forces	≤ 25K
7	52	high school	divorced	transport	25K-50K
8	40	doctorate	married	professional	≥ 50K

   (a) Calculate **information gain** (based on entropy) for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

   (b) Calculate **information gain** using the **Gini index** for the EDUCATION, MARITAL STATUS, and OCCUPATION features.

   (c) When building a decision tree, the easiest way to handle a continuous feature is to define a threshold around which splits will be made. What would be the optimal threshold to split the continuous AGE feature (use information gain based on entropy as the feature selection measure)?

2. The following table lists a dataset of the scores students achieved on an exam described in terms of whether the student studied for the exam (STUDIED) and the energy level of the lecturer when grading the student's exam (ENERGY). Which of the two descriptive features should we use as the testing criterion at the root node of a decision tree to predict students' scores?

ID	STUDIED	ENERGY	SCORE
1	yes	tired	65
2	no	alert	20
3	yes	alert	90
4	yes	tired	70
5	no	tired	40
6	yes	alert	85
7	no	tired	35

3. The following table lists a dataset containing the details of five participants in a heart disease study, and a target feature RISK which describes their risk of heart disease. Each patient is described in terms of four descriptive features: EXERCISE (how regularly do they exercise?), SMOKER (do they smoke?), OBESE (are they overweight?) FAMILY (did any of their parents or siblings suffer from 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) As part of the study researchers have decided to create a predictive model to screen participants based on their risk of heart disease. You have been asked to implement this screening model using a **random forest**. The three tables below list three bootstrap samples that have been generated from the above dataset. Using these bootstrap samples create the decision trees that will be in the random forest model (use entropy based information gain as the feature selection criterion).

ID	EXERCISE	FAMILY	RISK
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high

Boostrap Sample A

ID	SMOKER	OBESE	RISK
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high

Boostrap Sample B

ID	OBESE	FAMILY	RISK
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

Boostrap Sample C

(b) Assuming the random forest model you have created uses majority voting, what prediction will it return for the following query:

EXERCISE=rarely, SMOKER=false, OBESE=true, FAMILY=yes.