Instructions

- This homework assignment is worth 102 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: November 4, 2022 by 11:59 pm.**

Exercise 1

(5 points) If a decision tree is under-fitting the training dataset, is it a good idea to try scaling the input features?

Exercise 2

(5 points) If a decision tree is over-fitting the training dataset, is it a good idea to try decreasing `max_depth`?

Exercise 3

(4 points) Why would you use a random forest instead of a decision tree?

(a) For a lower training error.

(b) To reduce the variance of the model.

(c) For a model that it is easier for human to interpret.

(d) (a) and (b)

(e) (a) and (c)

(f) (b) and (c)

(g) All of the above

(h) None of the above

Exercise 4

(4 points) Which of the following is/are **TRUE** about bagging trees?

(a) In bagging trees, the trees are grown independent of each other.

(b) In bagging trees, the trees are grown in sequence.

(c) Bagging is a method for improving the performance by aggregating the results of weak learners.

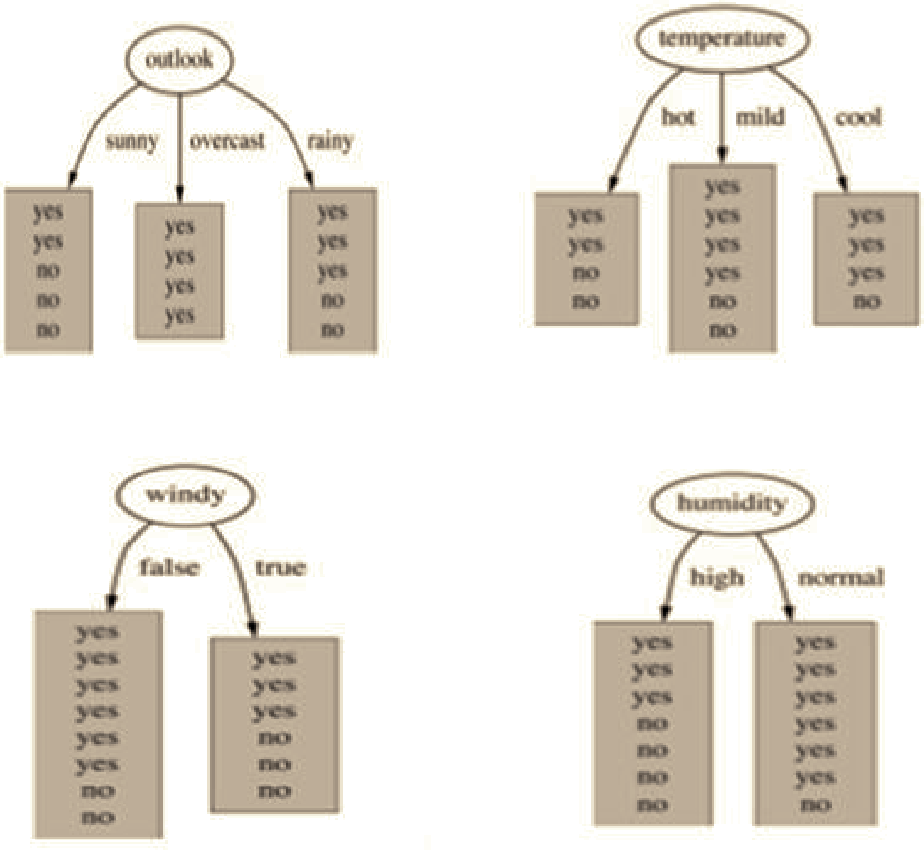
(d) (a) and (c)

(e) (b) and (c)

(f) None of the above.

Exercise 5

(12 points) Suppose you are building random forest model, which split a node on the attribute, that has highest information gain (using the Gini index). In the below image, which attribute which has the highest information gain? Show all your calculations.



Exercise 6

Consider the `framingham.csv` data file. The dataset is publically available on the Kaggle website, and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD). The dataset provides the patients? information. It includes over 4,000 records and 15 attributes. Each attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

- Demographic:
  - Sex: male or female (Nominal)
  - Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
- Behavioral
  - Current Smoker: whether or not the patient is a current smoker (Nominal)
  - Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)
- Medical (history)
  - BP Meds: whether or not the patient was on blood pressure medication (Nominal)
  - Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)
  - Prevalent Hyp: whether or not the patient was hypertensive (Nominal)
  - Diabetes: whether or not the patient had diabetes (Nominal)
- Medical (current)
  - Tot Chol: total cholesterol level (Continuous)
  - Sys BP: systolic blood pressure (Continuous)
  - Dia BP: diastolic blood pressure (Continuous)
  - BMI: Body Mass Index (Continuous)
  - Heart Rate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
  - Glucose: glucose level (Continuous)
- Predict variable (desired target)
  - 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

In Python, answer the following:

(a) (4 points) Load the data file to you S3 bucket. Using the pandas library, read the csv data file and create a data-frame called `heart`.

(b) (3 points) Remove observations with missing values.

(c) (30 points) Using all the available variables as the predictor variables, and `TenYearCHD` as the target variable, do the following:

  (i) Split the data into train (80%) and test (20%) (taking into account the proportion of 0s and 1s).

  (ii) Using the train data-frame, build a random forest classifier (using 500 trees).

  (iii) Extra the feature importance of each of the variables.

  Repeat (i)-(iii) 100 times. Compute the average importance of each of the variables across the 100 splits. After that, select the top 5 variables (the ones with top 5 average importance) as the predictor variables.

(d) (35 points) Using the top 5 variables from part (c) as the predictor variables and `TenYearCHD` as the target variable, do the following:

  (i) Split the data into train (80%) and test (20%) (taking into account the proportion of 0s and 1s).

  (ii) Using the train data-frame, build a random forest classifier (using 500 trees and maximum depth tree equal to 3). Using this model, predict the likelihood of risk of coronary disease of the patients in the test data-frame. Using 10% as cutoff value, report the recall.

  (iii) Using the train data-frame, build a random forest classifier (using 500 trees and maximum depth tree equal to 5). Using this model, predict the likelihood of risk of coronary disease of the patients in the test data-frame. Using 10% as cutoff value, report the recall.

  (iv) Using the train data-frame, build a random forest classifier (using 500 trees and maximum depth tree equal to 7). Using this model, predict the likelihood of risk of coronary disease of the patients in the test data-frame. Using 10% as cutoff value, report the recall.

  Repeat (i)-(iii) 100 times. Compute the average recall of each of the models across the 100 iterations. What model would use to predict `TenYearCHD`? Be specific.