Instructions

- This homework assignment is worth 78 points.
- Please submit a **.ipynb** file to Blackboard.
- One submission per team.
- **Please strive for clarity and organization.**
- **Due Date: November 10, 2023 by 11:59 pm.**

Exercise 1

(4 points) Which of the following statements is **true**?

(a) logistic regression is a parametric model while random forest is a non-parametric model.

(b) logistic regression is a non-parametric model while random forest is a parametric model.

(c) logistic regression and random forest are both parametric models.

(d) logistic regression and random forest are both non-parametric models.

(e) All of the above.

(f) None of the above.

Exercise 2

(4 points) Consider the below table that shows the predicted churn probabilities of five decision trees for a given customer of cable company. What is the bagged predicted churn probability using the five decision trees?

Number of Tree	Probability of Churn
1	0.78
2	0.81
3	0.76
4	0.82
5	0.80

Exercise 3

(4 points) What is the difference between random forest and boosted trees? Be specific.

Exercise 4

(6 points) Consider the following decision tree, show how new observation in the table would be classified by filling the last column in the table. If an observation can not be classified, enter **unknown** in the last column.

```
                        Color
                          |
           --------------------------------
      Blue |          Red |         Green |
           |              |               |
         Width          No            Height
     -------------                  -------------
Thin |       Fat |            Short |      Tall |
     |           |                  |           |
    No          Yes                No          Yes
```

Observation	Color	Height	Width	Class
1	Red	Short	Thin	
2	Blue	Tall	Fat	
3	Green	Short	Fat	
4	Green	Tall	Thin	
5	Blue	Short	Thin	

Exercise 5

Consider the `Default.csv` datafile presented in class. In this data set, we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit balance. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and called `credit_default`.

(b) (4 points) Using the `where` function from the `numpy` library, create a new variables in the `credit_default` data-frame called `default_numb` that takes the value of 1 when `default = Yes` and 0 when `default = No`.

(c) (5 points) Using the `sklearn.ensemble`, build a random forest classifier model model in which `default_numb` is the target variable, `balance` and `income` are input variables.

(d) (4 points) Using the model from part (c), predict the likelihood of default of a customer with `balance` $= \$2,000$ and `income` $= \$70,000$.

(e) (4 points) Using the `where` function from the `numpy` library, create a new variables in the `credit_default` data-frame called `student_numb` that takes the value of 1 when `student = Yes` and 0 when `student = No`.

(f) (5 points) Using the `sklearn.ensemble`, build a random forest classifier model in which `default_numb` is the target variable, `balance`, `income` and `student_numb` are input variables.

(g) (5 points) Using the model from part (e), predict the likelihood of default of a customer with `balance` $= \$2,500$ and `income` $= \$20,000$ and `student_numb = 1`.

Exercise 6

Consider the `Churn_Data.csv` datafile. This datafile contains information related to a multinational bank. Assume that you work at a multinational bank that is aiming to increase its market share in Europe. Recently, it has been noticed that the number of customers using the banking services has declined, and the bank is worried that existing customers have stopped using them as their main bank. This bank hired you as a consultant, you are tasked with finding out the reasons behind customer churn and to predict customer churn. The marketing team, in particular, is interested in your findings and want to better understand existing customer behavior and possibly predict customer churn. Your results will help the marketing team to use their budget wisely to target potential churners. **In Python**, answer the following:

(a) (3 points) Using the `pandas` library, read the csv file and create a data-frame called `customer_churn`.

(b) (3 points) Using the `numpy` library, create a variable called `Churn_numb` that takes the value of 1 when `Churn = Yes` and 0 when `Churn = No`.

(c) (5 points) Using the `RandomForestClassifier` function from the `sklearn.ensemble` library, build a random forest classifier, in which `Churn_numb` is the target variable, `Age`, `EstimatedSalary`, `CreditScore`, `Balance`, and `NumOfProducts` are the input variables. Use `n_estimators = 500` and `max_depth = 3`.

(d) (5 points) Using the model from part (c), predict the likelihood of churn of a customer with `Age = 50`, `EstimatedSalary = $100,000`, `CreditScore = 600`, `Balance = $100,000`, and `NumOfProducts = 2`.

(e) (5 points) Using the `GradientBoostingClassifier` function from the `sklearn.ensemble` library, build a boosted tree classifier, in which `Churn_numb` is the target variable, `Age`, `EstimatedSalary`, `CreditScore`, `Balance`, and `NumOfProducts` are the input variables. Use `n_estimators = 500` and `max_depth = 3`.

(f) (5 points) Using the model from part (e), predict the likelihood of churn of a customer with the characteristics presented in part (d)

(g) (4 points). Using the results from part (d) and (f), answer the following: assuming the marketing department of this multinational uses 40% as a reference to flag out customers. That is, a customer with likelihood of churn of 40% or greater would be flagged out as a customer who is likely to churn. Would the customer with the characteristics presented in part (c) be flagged out? Explain.