Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur

Part-IV
Machine Learning with Pandas

Instructions:

- Solve the problems using "Jupyter Notebook" linked to "Anaconda Navigator".
- For each problem, create a "Markdown" window, mention the problem you have asked to solve followed by "Code" window containing your program and output of the program.
- As comments as much as you can.
- Upload the .ipynb file. Give the name of your file as **A8-<RollNo>.ipynb.** You have to upload only one file in this lab.

Time: 2 hours Full Marks: 100

Problem 1:

(a) Create a emp.csv file with the following data.

ID	Name		Gender	Age	Score	City
1	Alice		Female		9.8	NY
2	Bob	Male	27	6.5	MN	
3	Tom		Male	24		PA
4	Jerry			26	8.4	PN
5	Kim	Female	7.9		NY	
6	Honey			25		MM
7	Jisa	Female	21	9.0	PA	

(b) Create a data frame from this emp.csv file.
(c) Replace 'Female' as 0 and 'Male' as 1
(d) Filling all the missing values in the column 'Score' with the average value in that column
(e) Fill all the missing value in the column 'Gender' as 2
(f) Change the name of the column 'Gender' as 'Sex'
(g) Arrange the data group by 'City'
(h) Add a column, say 'YOA' which is initially empty
(i) Fill the values in each row of the column 'YOA' with data in the range [2022, 2023, 2024] at random.
(j) Modify all the values in each row of the column 'Score' by adding 0.5 to each
(k) Modify a specific entry, say Tom's age by 28
(l) Save the data frame as emp2.csv

[36 points]

Problem 2:

Consider the "Breast Cancer data" from the link HERE.

(a) Read the .csv file as a Pandas dataframe.
(b) Convert all categorical attribute values to numerical values
(c) Remove the column ID from the dataframe.
(d) Split the dataframe into training and test sets using random sampling and in the ration 2:1.
(e) Normalize the training and test data using "Standard Scaling"
(f) Build the classification models using SVM, LR, RF, DT, XGBoost, Gradient Boosting algorithm.

(g) Print the Confusion Matrix for each model
(h) Store Accuracy, Precision, Recall and F1 Score for all the models in the form of a table.
(i) Draw a group bar chart showing the comparison of all the models with respect to each performance estimation.

[64 points]