About the Exam !

- Examples are provided in the slides that follow
- Some of the questions test the skills acquired and ability to interpret results
- Some questions test the ability to handle open ended problems, define the problem concretely, and comprehensively solve it and present the results.
- The exam will be **open all resources** including your own notes, uploaded presentations, internet resources - including LLMs.
- It is expected that LLMs will be as used as tools for code generation, and not for solving the problem!
- It is expected that you will guide the LLMs to create code under your control and be responsible for the code that is generated, completely understand it, and stand by it, and defend it when required. Inability to explain code, completely, will result in the loss of all credits associated with that question.

- NOTE: The samples provided, and the marks allocated, should not be taken as a **pattern** that will be followed in the forthcoming end semester examination. There are no fixed question patterns that will be followed. Each examination is unique.

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Question – 1 [10 marks]

Datasets **D1.csv** and **D2.csv** are relevant to this question.

a. Perform EDA on these datasets and generate the relevant plots. Explain why you have chosen these plots and report your initial observations. **[2 marks]**

b. For each dataset, perform **regression analysis** and **report** the following metrics: **[2 marks]**

   i. Coefficients (slope and intercept)

   ii. MSE

   iii. R^2 score

   iv. Plot of data points along with the regression line.

c. Compare the slope, intercept and R^2 values related to D1 and D2. What do you observe? Explain why this is possible. What does this tell you about the **quality of fit** and **data variability**? **[2 marks]**

d. Suppose dataset D2 had a very low R^2 score. Can we say that the regression model is incorrect or poorly fitted? Justify your answer. **[2 marks]**

e. What changes can you possibly make to D2 to obtain better models? Would transforming the data help? **[2 marks]**

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Question – 2 [10 marks]

Dataset **health_data.csv** is relevant to this question, and the column **is_diabetic** is the dependent variable of this dataset.

a. This dataset has problem(s). Identify and report. **[2 marks]**
b. Rectify the problem(s) you have identified. List the steps taken and justify your approach. **[2 marks]**
c. Split the final dataset into train and test sets (80-20 split) using random sampling. **[1 mark]**
d. Train a classifier of your choice to predict the dependent variable. **[1 mark]**
e. Using the classifier you have created, calculate and report at least four metrics based on the test set. Comment briefly on the model performance based on these metrics. **[2 marks]**
f. How could feature engineering or domain knowledge improve preprocessing or model performance in this problem? **[2 marks]**

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Question – 3 [20 marks]

The dataset **fraud.csv** is relevant to this question.

a. Report the imbalance ratio of the dataset. **[1 mark]**
b. Visualize the dataset in 2D and comment about it. **[3 marks]**
c. Train a baseline logistic regression or random forest model without any class rebalancing. Report and interpret: **[2 marks]**
    i. Accuracy
    ii. Precision, Recall, F1-score (especially for the minority class)
d. To handle the imbalance, apply all of the following techniques (stating you reasons, choose your own rebalancing ratio): **[8 marks]**
    i. Random under-sampling of majority class.
    ii. Random over-sampling of minority class.
    iii. SMOTE.
    iv. Tomek links.
e. Visualize the dataset before and after your rebalancing technique. **[2 marks]**
f. Train a model on your rebalanced datasets for each technique. Evaluate this model on the original dataset and compare its performance to the baseline model that you trained in 3(c), above. What do you think, did your rebalancing step help train a better model? Justify your answer. **[4 marks]**

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.

Question – 4 [30 marks]

Read the following statements carefully and work out your solutions following Data Science best practices. The assessment scheme will be as follows:

Problem understanding, definition and solution approach: completeness and correctness	20 marks
Observations, analysis, results, and conclusions: completeness and correctness	10 marks

You have applied for the post of a Data Scientist (DS) at FitCo Pvt. Ltd., a company with a pan-India presence, and engaged in the manufacture, marketing, sales, and support of fitness equipment. With Indians focussing on fitness goals, the company has been growing steadily since the pandemic. As per the advertisement for the DS post, it now wants to introduce and emphasize **data orientation** in all its activities. The company's Marketing and Sales Head is tasked with the responsibility of building the team of Data Scientists and Analysts.

On the day of the interview the Head addresses the candidates as follows:

> "We are giving you one of the files – **fitco.csv** – from our data archives. See what best you can do with it! It contains the total sales value (Sales), and unit Sales price (SellingPrice) related to the sales of some of our products. It also tells us if any product return happened related to a particular sale. **Tell us how we can use this data for some useful purposes**". He continued, "BTW, there is some relevant information for you. In general, we know that people in class-A cities have more expendable money, and we enjoy selling to them!
>
> Finally, he said, "And one more point, I will need good explanation of whatever you do, with justifications! You will have to convince me that whatever you create, it is the best and reliable!"
>
> "Wish you the very best, all of you"

Do your best to satisfy the executive, and get the job. It will help if you create logical, complete, yet precise and well-organized material. Good luck!

This file is meant for personal use by mepravintpatil@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.