Programming for Machine Learning and Data Science

Semester-end Examination
May 10, 2025: 10:00 – 13:00
Marks: 50, Weightage: 50%

INSTRUCTIONS:

- This is a **paper-less** examination. Your solutions should be uploaded to Olympus, by 13:15, as explained later.
- Solutions should be created using coding + documentation environment like Google Colab / Jupyter Notebook / etc. Every question should be answered in a separate Notebook (Note: all sub-parts of a question should be in the same Notebook!)
- You are recommended to use the Notebook itself for also documenting your explanations / observations / analysis / conclusions.
    - However, if you so desire, you can submit a separate document (PDF) containing your analysis and report.
- **Note**: In case you use online tools such as Google Colab, be sure to download and submit '.ipynb' file(s) and **NOT** links to the online Notebook. Likewise, if you create any (report) document online, submit its PDF.
- File names of the Notebook / PDF files SHOULD have your full name.
- The Notebooks should flawlessly execute with your data file(s) kept in the same folder.
- Upload the Notebooks, PDF of your report (if separately created) to Olympus, to the submission point: **Final term exam – 1st Semester**
    - Submissions will be accepted only until 13:15 Hrs.

- . You can access your notes, slides, Internet based tools (**including code generation tools**) during the test. However, the following restrictions will strictly apply:
    - **AUTOMATIC GENERATION OF OVERVIEWS / SUMMARIES / REPORTS / ANALYSIS / CONCLUSIONS WILL ATTRACT PENALTIES and MAY LEAD TO REJECTION OF THE SUBMISSION AND AWARD OF ZERO MARKS. All these should be your own!**
    - Copying, collaboration, use of ANY online collaboration tool / other means of collaboration is **NOT ALLOWED** at any time.

- **In case of difficulties in uploading to Olympus, use one of the following links to upload your solution:**

  https://docs.google.com/forms/d/e/1FAIpQLScS3MhWy2xa6ADAX7LS-GxCGm4hdlp0MqIzNaNshM9s54sazg/viewform?usp=dialog
  https://forms.gle/7FoamyW6Mr5N8xfb7

Under no circumstances will submissions be accepted after 13:15 Hrs.

In case of any queries related to these instructions, contact an invigilator

***

Note - 1:

- Budget approximately **3 minutes for every mark** allotted to a question (E.g. budget 30 minutes for 10 marks question)
- Every question should be answered in a separate Notebook (and all sub-parts of a question should be in the same Notebook!)
- In the Notebooks clearly mention the question number, such as 1(a), 2(b), etc. before starting the solution.

Note – 2:

- Four problems have been given below, along with the total marks allocated to them.
- You should choose a combination of questions, such that the maximum marks total to 50. For example: [Q1, Q2, Q4] or [Q3, Q4]. You can also answer all questions, if time permits, and the best combination will be considered!

==========

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]**

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]**

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]**

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!

---

oooOOOooo