Programming for Machine Learning and Data Science

Semester-end Examination – Part 1
Nov 29, 2025: 10:00 – 12:15
Total Marks: 35

INSTRUCTIONS

The semester examination has two parts:

1. Part-1, 35 marks: 10:00 am – 12:15 pm (this question paper)
2. Part-2, 20 marks: 12:15 pm – 1:00 pm (question paper will be distributed at 12:15 pm)

Instructions for Part-1

- This part is a **paper-less** examination. Your solution files (see below) should be **ZIP**ped and uploaded to Olympus, by 12:15 pm
- Solutions should be created using coding + documentation environment like Google Colab / Jupyter Notebook / etc.
- 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 the following format "**FULL NAME_EOID**".
- 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 **"Endsem-Part-1"**
- You can access your notes, slides, Internet based tools (**including code generation tools**) during the test. However, the following conditions and restrictions will strictly apply:
    - **You will be deemed to be the creator of the code. This requires you to completely own and understand the generated code and, if required, explain it later.**
    - **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.

Under no circumstances will submissions be accepted after 12:15 Hrs.

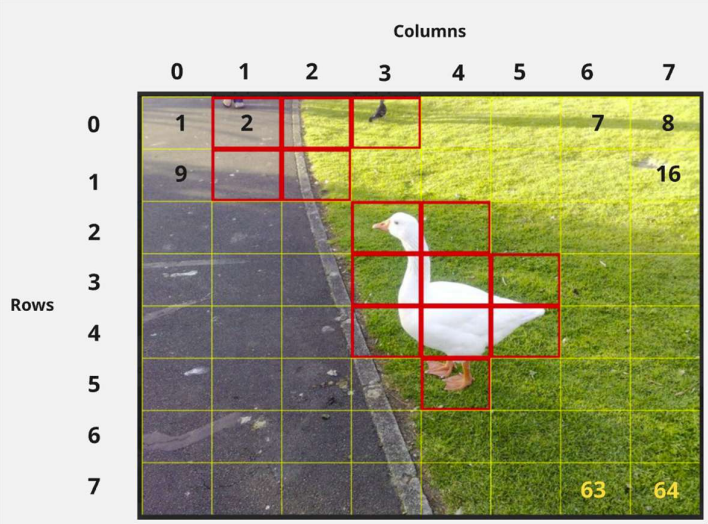
In case of any queries related to these instructions, contact an invigilator

***

Dataset: All the questions are based on the dataset **image-data-ePGD-endsem.csv**.

An overview of this dataset follows:

- **Each row in the dataset represents a cell of an image**. Images are divided into 64 cells and into 8 columns and 8 row numbered as shown below using a sample image (the red cells have no specific significance in the context of this dataset; consider them to be yellow!).



- Each record (row) in the dataset contains the following:
  - Name of the image file (eg. ***IMG_20250705_121136381~2.jpg***)
  - The cell number (1 to 64) of the image
  - The cell_row index (0-8)
  - The cell_col index (0-8)
  - Label (0-1) where:
    - '1' indicates that wild-life is contained in the cell
    - '0' indicates no wild-life is contained in the cell
  - Values of the feature vector created by applying image processing methods to the cell (eg. colour histogram, histogram values of oriented gradients (HOG), etc.

==========

Questions follow on the next page.

Note:

- Budget approximately **3 minutes for every mark** allotted to a question (e.g. budget 30 minutes for 10 marks' question)
- In the Notebook clearly mention the question number, such as 1(a), 2(b), etc. before starting the solution.

Question – 1 [15 marks]

a. Perform and document all the EDA steps executed on the dataset by you.

*Irrespective of **when** the EDA steps are implemented (initial EDA, done to understand the dataset, or subsequent EDA, as you progress through the questions), all the EDA steps should be documented and explained at one place. Following is the suggested format for EDA documentation.*

- EDA step 1:
- Reason for this step:
- Observation and analysis:
- Actions to be taken (if any):
- Remarks

- EDA step 2:
- Reason for this step:
- …
- …

Question – 2 [20 marks]

a. Is it correct to say that Box Plots and Histograms are closely related to each other? Justify your answer with the help of concrete evidence based on the provided dataset. [**3 marks**]
b. Using a relevant feature from the dataset, illustrate how data skew can be reduced using Log Transformation and Box-Cox Transformation. [**3 marks**]
c. Are there any features that indicate almost normal distribution? Identify such features and provide convincing evidence for arriving at this conclusion. [**3 marks**]
d. If you carry out clustering on the dataset using t-SNE coordinates, find out the most optimum number of clusters to be created. Create the required plot(s) to justify your answer [**5 marks**]
e. There is a hypothesis that wild-life photographers always take pictures by placing the animals somewhere near the centre of the photograph. Based on the data given to you, and by using either ***DecisionTreeClassifier*** or ***KNeighborsClassifier*** classification algorithm, can you prove / disprove this hypothesis?
    i. Describe the steps you will take to solve this problem [**3 marks**]
    ii. Implement the solution, analyse the results, state your conclusions [**3 marks**]

---

oooOOOooo