DS203-2023-Sem2: Exercise – 7 (Project)

- **This Exercise (Project) carries 50 marks and 50% weightage**
- **It should preferably be done in teams of at least 2 participants, but not exceeding 4 participants**
    - **Compulsorily register your group at: https://tinyurl.com/2024-03-E7-Groups**
        - **Only ONE member of the group should fill the form!**
    - **Team of '1' is strongly discouraged, but not barred. Solo attempt should be only the last alternative.**
- **Read the instructions and the evaluation criteria carefully.**
- **Note and adhere to the submission requirements and deadlines carefully.**
- **Submissions due by: April 14, 2024, 11:55pm**

Group Registration QR Code

Submission Guidelines

You **should** submit a **SINGLE ZIP FILE** containing the following:

1. **A carefully prepared presentation** that conveys, succinctly, the essence of all your work, analysis, and results. Refer to the presentation guidelines outlined later in this document.
2. **All the Python code files created by you** (code should execute with the data files kept the same directory)
3. **All the data files created by you**. These could be intermediate input data files containing feature information after feature engineering, or output files resulting from analysis or model executions. DO NOT submit the original image files uploaded to Moodle as part of the exercise definition.

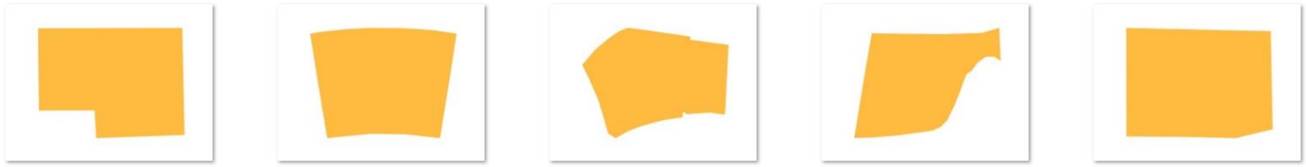
1) If linearity holds then SVM, MLR
2) If linearity x holds KNN, decision tree

Evaluation Criteria

Evaluation Criteria	Marks
• Are all the questions posed by the problem statements effectively addressed with solutions? • Are the solutions relevant, and correctly applied, and backed up with proper logic / reasons? • Has there been any creative thinking and innovation while solving the problems? • Have any possibilities, other than those asked for, been implemented, or listed in the presentation?	15
• Are the major steps of data analysis diligently followed and correctly applied and documented (wherever required …)? ○ EDA, correct data transformation, structural analysis, correct feature engineering, correct model building and model validation, final analysis.	15
• Quality of results: are they backed with appropriate metrics, comparisons, analysis, explanations, and justifications?	10
• Quality of presentation: ○ Completeness and preciseness of the final slide deck: design and readability of the slides. Are **all the above questions** covered in the presentation?	10
• **Does the presentation contain raw and hyped-up outputs generated using LLM tools like ChatGPT / Gemini etc.**	**-25**
• Viva will be conducted, if deemed necessary, to ascertain originality of work, and to ascertain contribution by team members. • The project will be evaluated **ONLY** by reviewing the presentation. Source code / Jupyter Notebook will **NOT** be referred to, to understand your work; they will be used to *verify* the claims you have made in the presentation. Therefore, if you forget to mention some part of your work / analysis in your presentation, it would be deemed that you have not done it! • If you do not submit your source and data files, your project becomes unverifiable and the submission will not be given any credit.	

Problem Description

An architecture company has recently completed an initiative to digitalize their building layout designs. As part of this exercise, they have converted the plan view of buildings architected by them into bitmaps of size 640 x 480 pixels, as shown by the sample images below.



A repository of 1183 such views (ie. buildings) has been created in the first phase. These images are available in the file **E7-images.zip** uploaded to Moodle.

They are now ready to launch initiatives to use this data to improve their productivity, and facilitate robust designs. They are also very keen to mine the information potentially hidden in these views, and establish new and innovative processes to respond rapidly to customer requirements and demands. They are convinced that methods of Data Science in general, and Machine Learning in particular, can be effectively used to achieve these goals. Following are some of the requirements / questions they have posed:

1. *We have a hunch that our designs can be grouped into families, based on their shapes. Knowledge of these families will not only improve our insights into our own designs, but also help in standardizing them by creating design templates. We have been informed that multiple approaches can be used to do this, and we would like to see the results of at least two approaches.* < for grouping
2. *Further, we would like to classify the complexity of layouts as **Low Complexity**, **Medium Complexity**, **High Complex**. Based on a formal analysis of the layouts we would like to create the criteria to decide their complexity. Can you carry out an appropriate analysis, establish the criteria, and classify the layouts?*
3. *We are very keen to speed up our layout design process by retrieving relevant prior layouts based on a set of gross parameters. For example, the architect usually knows the dimensions (length, width) of the tight-fitting box (see Appendix 1), the layout area, and the permissible layout complexity. When the architect specifies these parameters, the **design family / families** likely to provide the closest designs should be predicted. Layouts from such families can be retrieved for further detailing. This will save a lot of time and effort, and ensure that designs are based on past, successful designs!*
4. *We are open to other suggestions, about what else can be done using the image data and information mined from it. Please suggest more possibilities and ideas!*

Solution guidelines

- Doing justice to this project involves much work, and perhaps, learning some new skills, related to image analysis and creating and making presentations. Do what it takes to provide solutions to the above requests / problems – and effectively communicate them!
- Re-visit all major aspects of Data Science that you have learned so far, and check if they can be / need to be used to solve the posed problems. EDA PE, Regression, Classification,
- It will really help to work in a team, and divide work amongst the members – to do a good job.
- It is important to have an overall **solution design** in place to ensure that:
  - All the customer's requests and requirements are taken care of
  - The appropriate (additional) data, required to implement the solutions, get generated.

EDA → Feature eng. → PCA t-SNE → Logistic regression / decision tree KNN / SVM / Naive bias →

+ Converting image to pixels 640x480

Image Processing is imp.

Presentation guidelines:

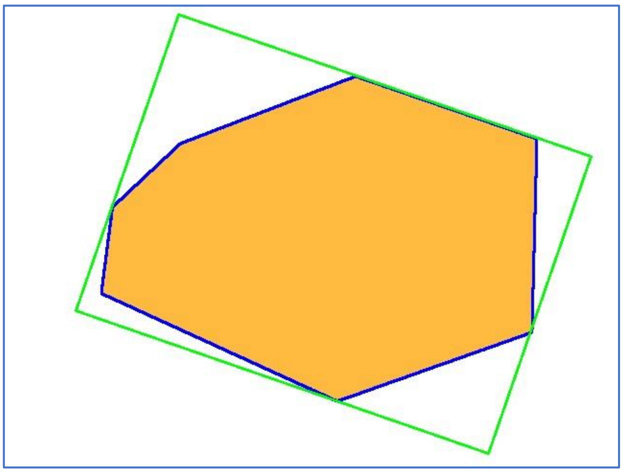
1. As mentioned, succinctly communicating your work and results is a very important part of the Data Science process. One of your important submissions is the presentation – that summarizes your approach, work, results, achievements, learnings, and possibilities. Budget adequate time for this activity and design your presentation well; last minute work will invariably be shoddy.
2. **Mention the names and roll numbers of all group members on the title slide. No credit will be given to members who are not mentioned on the title slide.**
3. Provide an executive overview (1-2 slides) at the start of the presentation.
4. DO NOT use verbose paragraphs, or storytelling, to explain your steps, observations, results, and recommendations. All these should be presented precisely and point-wise. *Review the uploaded sample presentations to get some idea about effective presentations.*
5. Summarize your observations and results using charts / Tables and adequately explain them and explain and draw conclusions from them. Merely including charts and Tables is not enough. Slides should be well designed. Use as many slides as required to completely convey your work.
6. If you do not include something in your presentation, it will be deemed that you have not done it. **Source code / Notebooks will NOT be reviewed to *understand* your work**.
7. Towards the end of the presentation, include slides that clearly answer all the questions posed in the ***Evaluation Criteria*** table. In addition to focussing the evaluator's attention, this will also ensure that you have covered all the expected points in the presentation.
8. Finally, include a slide or two outlining your learnings from this project, and your experiences and hurdles while doing the project.

Note that the presentation is just one of the submission requirements. **Re-visit the Submission Guidelines described at the beginning of this document.**

APPENDIX – 1: Tight Fitting Bounding Box

A tight-fitting bounding box around a contour is the one with minimum area. Such a box is usually not aligned with the horizontal and vertical axes of the image containing the contour – as shown in the image below.

- Outer blue box: Total extents of the image
- Inner dark blue polygon: Contour / layout under consideration
- Green box: The tight-fitting box covering the contour under consideration



oooOOOooo