German University in Cairo
Faculty of Media Engineering and Technology
Assoc. Prof. Mohamed Hamed

Machine Learning CSEN-1022

Project Description WS24

General Information:

Teams:

- Each team can have 6 members or less
- ALL team members must contribute
- You are required to submit your team members names and IDs and your chosen dataset in this form before Thursday 14th of December 2024.

Grade:

- This project accounts for 35% of your total grade of the course.
- For more information, check the **Deliverables** section.

Submission:

The project needs to be submitted as a zip file containing the deliverables as stated in the **Deliverables** section

Deadline:

Your submission deadline is at **11:59 pm On Thursday 12th December 2024**, any late submissions will not be considered.

Evaluation:

- After the deadline passes, each team will be asked to present the presentation submitted as per the **Deliverables** section.
- The presentation will have a total allocated time of 20 minutes, 12 minutes for presenting and 8 minutes for discussion.
- During the presentation, you will be judged based on your presentation, we will also be cross-referencing your notebook mentioned in the **Deliverables** section against your presentation while your team presents.

Deliverables:

Presentation:

- The presentation must follow the outline stated in the **Presentation Outline** section.
- If the presentation isn't a PowerPoint presentation (.pptx) please add a readme file to your submission file with instructions on how to preview the contents of the presentation.

Notebook:

- The notebook must adhere to the guidelines stated in the **Notebook Guidelines** section.
- If there are any required libraries (modules) that need to be installed for the notebook to work, aka, if it isn't a built-in library, then please either state the library clearly in a readme file or embed the installation within the notebook.

German University in Cairo
Faculty of Media Engineering and Technology
Assoc. Prof. Mohamed Hamed

Presentation Outline:

Background:

- Context and motivation for the project.
- Description of the problem to be tackled and why it matters.

Dataset:

- Description of the data being used, including size, features, descriptive statistics and visualizations.
- Choose one dataset from 11 from the spreadsheet uploaded on the CMS.

Objective/Goal:

The problem statement and what the project aims to achieve.

Method: Outline of the methodology including:

Data Exploration and visualization:

→ Doing some descriptive analysis for your selected dataset as well as plotting different plots taken in the course

Data cleaning and processing:

➔ Appropriate steps to prepare data and get it ready for the rest of the pipeline.

- ◆ Handling missing values
- ◆ Deal with Outliers
- ◆ Smoothing and deal with duplicates
- ◆ Normalization & Splitting data
- ◆ etc.

Feature extraction and selection:

➔ Identifying and selecting significant variables for the model and stating clearly the approach.
➔ Applying Dimensionality Reduction, and justifying why you used it.
➔ PCA analysis is a must do especially for the classification tasks.

Model selection:

➔ Choosing the appropriate algorithms based on the problem.
➔ Please not that you are required to:

a. Applying unsupervised learning (Clustering) to your data.
b. Applying either Regression or Classification to your data (according to the chosen dataset), apply at least 3 models for either of them.

c. Compare between the outcomes of different regression/classification algorithms for the same task.

Model training:

➔ Mentioning your data split and hyperparameters.

Model evaluation:

➔ please apply all evaluation methods (taught in the lectures) to assess the performance of the model against test data and across different hyperparameters, if possible.

Results:

- The outcomes of the model training and evaluation, the final model performance along with the below for each algorithm.

  Clustering:

  ➔ Visualization of the clusters generated with comparison to ground truth.
  ➔ Comparison between the algorithms used.

  Classification:

  ➔ Visualization of the decision boundary along with the original data points.

  Regression:

  ➔ Visualization of the regression boundary along with the original data points.

-Discuss the evaluation results of your models

Discussion and Future Insight:

Summarizing the findings and their implications, discuss how the results can be improved from all phases of the pipeline.

Notebook Guidelines:

- **Narrative Structure:**

  The notebook should follow the same outline and structure of the presentation.

- **Documentation:**

  The notebook should be well documented through markdown cells and code comments.

- **Robustness:**

  The notebook should not result in an error or faulty output if run in order.

- **Reproducibility:**

The output should be consistent across multiple runs, i.e. add a constant seed value (in case we had to run your notebook).

Special Notes to consider:

1. As a rule of thumb, we will **NOT** be running any notebooks. Therefore, the output of each code cell should be present in the submitted notebook.
2. Any insights, observations and/or conclusions should be backed up by code cells that prove how these insights, observations and/or conclusions were reached.
3. The submitted work needs to be authentically the team's work. Therefore, if plagiarism is detected (either from online resources or from other teams), the entire team will be penalized (could be to the extent of losing the entire percentage of the project).
4. Each team member MUST present a part of the work. ALL members should be prepared to be asked in all sections of the presentation.

Sources:

- Please note that if the data source selected is distributed across multiple files, you are required to join it into a single csv file before ***Data cleaning and processing***.
- Make sure to follow the columns Classification Target and Regression Target when doing classification or regression respectively.
- There are 11 datasets that you can choose from, the links are in a csv file uploaded on the CMS.