Programming in Python

Final Term Project - Fall 24-25

Project Description

The goal of the final term project is to implement different machine learning techniques in a dataset using Python. You should carefully read this document to complete the project successfully.

- The project must be completed in a three-person group. The groups will remain the same as the midterm groups. There are some students who completed the midterm project alone or in a four-person group. Those who did so, please form a three-person group to do the final project.
- Select a unique dataset to do this project that means your selected dataset must not match with the dataset of any other group in your class. Details about dataset selection will be discussed below.
- A spreadsheet has been posted in MS Teams containing the group information (Groups and Projects Information - Final Term Fall 24-25.xlsx). Please check the information and correct it if you find any inconsistency. Insert your selected dataset information along with the source URL in that file in the designated column.

Project Deliverables

- Submit the implemented python program ("python_final_project_group_XX.ipynb" file) and the report ("python_final_project_group_XX.pdf" file) in MS Teams. Replace "XX" in all filenames with your group number such as for group-1, the report file name will be "python_final_project_group_01.pdf". Additionally, please submit your dataset file. Create a ZIP file ("python_final_project_group_XX.zip") using these three files and submit your ZIP file only.
- During the VIVA session, you will bring this implemented program and we may ask you to execute the program. The program file must not contain any comments.
- See the instruction section below for the report details.

Before You Start

- Make sure your group information is complete in the spreadsheet (Check the given spreadsheet in MS Teams - Groups and Projects Information - Final Term Fall 24-25.xlsx).
- Make sure you have selected a unique dataset by reviewing the already selected datasets (Check the given spreadsheet in MS Teams).

General Instructions

- **The submission deadline for all deliverables is January 18, 2025 (you must submit the program and report by 11:59 PM). There will be a penalty for the late submission.**
- **Comments are not allowed in the implemented program.**
- At the beginning of the report (after the cover page), write a short description of your selected dataset.
- For each implemented task, write a paragraph in the report. In the paragraph, describe how the related task is implemented in your project along with the relevant code segment. While writing a description, only write the content (do not write unnecessary content) that is sufficient to understand the solution and its associated code segment. No need to write anything in the report for the pre-tasks.

Project Requirements

- **Pre-Task 1:** Form your group and make sure the information is updated in the spreadsheet by **28-12-2024**.
- **Pre-Task 2:** Select a labelled dataset that contains both numerical and categorical data. Provide the dataset information in the spreadsheet by **28-12-2024**. The dataset must have at least 4 features and 200 entries. You are not allowed to select the datasets: Iris, US Arrest and Titanic.
- **Task 1:** Read/Load the dataset file in your program. Use Pandas library to complete this task.
- **Task 2:** Apply appropriate data cleaning techniques to the dataset. In this step, replace bad data using proper methods and do not delete any record except duplicate records. Use Pandas library to complete this task.
- **Task 3:** Draw graphs to analyze the frequency distributions of the features. Use Matplotlib library to complete this task. Draw all the plots in a single figure so that all plots can be seen in one diagram (use subplot() function).
- **Task 4:** Perform scaling to the features of the dataset. Remember that you will need to apply data conversion before performing scaling whenever necessary.
- **Task 5:** Split your data into two parts: Training dataset and Testing dataset. You must use the function train_test_split() to complete this task and use value 3241 as the value of the `random_state` parameter of this function.
- **Task 6:** Apply Support Vector Machine (SVM) Classifier to the dataset. Build (train) your prediction model in this step.
- **Taks 7:** Calculate the confusion matrix for your model. Interpret it in detail in the report.
- **Task 8:** Calculate the train and test accuracy of your model and compare them.
- **Each task must be completed in a separate program cell in your python notebook file. That means a program cell for task-1, another program cell for task-2 and so on. No need to do anything for pre-tasks. I will provide a python notebook file named "python_final_project_group_XX.ipynb" in the portal and you will use that for your solution.**
- **Use Scikit-Learn library to complete a task wherever applicable. If a feature is available in Scikit-Learn, then do not use any other library such as PyTorch to implement that feature.**
- **Scikit-Learn supports various implementations of SVM method. You must implement the SVC version of SVM for this project.**
- **Do not use any library that needs to be installed other than matplotlib, numpy, pandas and scikit-learn.**