School of Computing and Information Sciences
Willamette University

DATA-599-02: Python for Data Science (Summer 2023)

Final Project

Instructor: Fred Agbo

Week 11 - 13:	**Task 1 – 3 Due date:**	**Task 4 – 5 Due date:**
17th Jul – 7th Aug 2023	24th July 2023 at 12 noon	7th August 2023

Assessment overview

This project is expected to allow you to demonstrate your knowledge and understanding of the application of machine learning techniques using Python. The tasks are to be completed in groups, lasting for three weeks. However, there are deliverables at the end of week one. The tasks are divided into five compulsory components with their allotted scores and an extra task to attract an extra score (not compulsory). Each group is expected to follow the instructions carefully and provide solutions and deliverables for each component on the due dates.

Task 1:

Project Proposal (7%): The group is required to submit a project proposal outlining the topic and problem they will address, research questions that motivate their choice, a description of datasets/sources to use, and the machine learning techniques they plan to use. The datasets and problem domain **must** be different from the ones presented in previous homework or mid-term review. The proposal must be brief and should not exceed one page.

Dataset Requirements: Each group must obtain a reasonable dataset for their chosen topic excluding the datasets used in the lectures and past homework. The dataset can be obtained from any public data repository and should be referenced. The dataset must meet the following requirements:

- Must have a column containing labels representing the classes identified in the dataset and should contain not less than 4 features.
- Must contain no less than 3 classes.
- Must have at least 100 samples.
- Your dataset can be sourced from any of the open data sources - (NYC Open Data, Kaggle Datasets, UCI Repository, e.t.c)

Task 2:

Data Pre-processing and Visualization (7%): Once the group has identified a Use Case (problem domain) and obtained the dataset, it is time to get familiar with the data. You are expected to apply data pre-processing techniques to process data using the Jupyter Notebook. The data pre-processing activities could include identifying missing values and fixing them, treating any outliers, scaling/normalization, and other kinds of processing that may be required to make the data ready for implementing machine learning models. You must explain all the steps taken in data pre-processing and their outcomes using Markdown in the Jupyter Notebook. Pay attention to details and describe any surprises from the process.

Task 3:

Effort Diary (2%)
Each group is expected to provide an effort diary that contains the name of each group member, the task (he/she/they) completed within the group at the point of submission, and the overall percentage of that task. Include the effort diary section in the Jupyter Notebook to be submitted.

Task 1 to 3 deliverables:

As a group, submit the following to WIS by 24th July 2023 at 12: noon

- One-page project proposal
- Dataset(s) obtained to address the identified problem
- Jupyter Notebook containing the data preprocessing steps.

****

Task 4:

Implementation & Evaluation (10%): Each group is expected to proceed to implement the proposed machine learning solution in Python programming language using the dataset obtained. You should demonstrate a clear understanding of the chosen techniques and apply them effectively to the chosen problem by answering the research questions proposed.

In addition to the machine learning models/techniques proposed earlier, you should implement machine learning solutions for the following tasks (if not included already in the original proposal).

Task 4.1. Implement a random forest classifier using sci-kit-learn. Create a subsection in your report describing the chosen parameters of the classifier and training method, and how they were set. Also, report the accuracy of the resulting method.

Task 4.2. Modify your solution in Task 4.1 to use a Support Vector Machine (SVC) classifier using sci-kit-learn with the appropriate parameters including *kernel = 'linear'*. Report the accuracy of the resulting method. Compare the results of the SVM model with the random forest model implemented in task 4.1. You may also justify the other parameters used or tuned in both models and how that affects the results.

Task 4.3. Use a clustering method of your choice to cluster the dataset. Describe any insights gained from the clusters. Also, create a subsection in your report where you could mention your methods, resulting accuracy, and any other metrics used in the evaluation of clustering results.

Advanced tasks for extra grades (5%):

For higher grades, you may consider adding advanced ML functionality to extend some of the basic ML solutions implemented. You should remember that the marks for advanced functionalities are considered only if **all** the basic solutions described above are implemented. There is no restricted list of extra functionalities, but some examples are as follows:

- Applying other ML techniques to improve the performance of your models and discuss the process in your report.
- Apply features engineering process, cross-validation of models, dimensionality reduction, etc, and explain the impact of these activities on your model accuracy.
- Describe how these advanced functionalities affect the accuracy of your models compared to applying the basic implementation process described above.

Task 5:

Report presentation (10%):

- Design a web (HTML) page where you will present a detailed project report
- The webpage report should follow the structure below and should contain text, tables, figures (if possible, codes), and references.

  Project title
  Brief introduction
  Problem statement
  Research questions
  Methodology
  Results
  Discussions and Implications
  Conclusion
  References
  Appendix: List of contributors (group members), link to their web/LinkedIn/other professional or social platform

- Pitch the project in the classroom (10 minutes per group)
- Submit the Jupyter Notebook containing the machine learning implementation and a link to the group webpage report.

All the best!