Course: Data Science **Instructor: M. Saif ul Islam**

Project Description Document

Introduction:

This project is an opportunity for you to apply your data science skills to real-world problems. The goal is to apply the knowledge learned from this course to solve a data-related problem. In this project, you have to perform all the data science steps including data collection, wrangling, exploration and visualization, modeling, evaluation, and presentation. This project constitutes **15%** of your course weight-age. Detailed instructions are given in this document. You are advised to thoroughly study this document before starting the project.

Project Requirements:

1. Data Collection:

Choose a topic and data set that is relevant to the course content and of interest to you. You may scrap the data from a website or use publically available data from different websites discussed during the class sessions. The dataset must be chosen given the following restrictions.

Nature	Minimum Features	Minimum Instances
Numeric/Categorical	10	5000
Images/Audio	(100x100)	1000
Video	-	100
Textual	-	5000

2. Data Wrangling and Transformation:

Clean and preprocess the data as needed to prepare it for analysis. Wrangling includes but not limited to removing duplicates, filling the missing data, data type conversion, aggregating, integration, categorical encoding, label encoding, scaling and normalization.

3. Exploration and Visualization:

Conduct exploratory data analysis to gain insights into the data and identify relevant trends and patterns. It provides a preliminary examination of the data and allows for an overview of its properties, characteristics, and patterns. It is an important phase because it can help to determine the most appropriate modeling techniques based on the distribution and characteristics of the data. You have to present your insights using different charts and graphs studied during the course.

4. Machine Learning Modeling:

Develop a predictive model to make predictions about the target variable based on the other variables in the data set. You can apply regression, classification, or clustering according to your problem. Apply at least 3 different algorithms to your problem and then save the best model for further processing.

5. Evaluation:

Evaluate the performance of your model and refine it as needed. Evaluation metrics depend on the nature of your problem domain. Follow the below given chart to evaluate your model.

Regression	R^2 Score, MAE, MSE, RMSE
Classification	Accuracy, Loss, Confusion Matrix, Precision, Recall, F1 Score, ROC/AUC
Clustering	Extrinsic Measures: Mutual Information, Rand Index, Adjusted Rand Index Intrinsic Measures: Silhouette Score Davies-Bouldin Index

6. Presentation:

Communicate your findings and conclusions through a well-structured presentation. Schedule of the presentation will be shared later.

7. Application or Research Paper:

Create an application and integrate it with your machine learning model. Integrating a machine learning model with a GUI-based application involves choosing a suitable GUI library, building the interface, loading the model, connecting the model to the GUI, testing the integration, and deploying the application. You may choose any python based GUI for front end or develop a web based application using JS, Django, Flask etc.

Alternatively, you may write a research paper explaining the problem, background, literature review, methodology, implementation and results section in detail. Download the template from the link given below and customize it accordingly.

https://www.ieee.org/content/dam/ieee-org/ieee/web/org/conferences/Conference-template-A4.doc

Deliverables:

- Dataset
- A brief explanation of the data set and any challenges you faced in working with it. (.pdf file)
- A Jupyter notebook (Use the provided notebook without any alterations.)
- Machine learning model (After training and evaluation)
- Research paper/Report (According to the template shared) and Source code of your application.

Grading Criteria:

- Data preprocessing and cleaning (Assignment 2)
- Data analysis and visualization (Assignment 3)
- Machine learning modeling (5 marks)
- Model evaluation and improvement (2 marks)
- Presentation and documentation (3 mark)
- Application or Research paper (5 marks)

Tips for Success:

- Start early and plan your time effectively to ensure that you have enough time to complete the project.
- Choose a data set that is well-suited to your interests and skill level.
- Work collaboratively with classmates and seek help from your instructor as needed.
- Be creative and innovative in your approach to the problem.
- Focus on quality over quantity, and prioritize clarity and effective communication in your report or presentation.
- Stick to the instruction provided in this document.