National University

Of Computer & Emerging Sciences Chiniot- Faisalabad Campus

CS 4048 – Data Science (Fall 2025)

Project I

Topics Covered:	**Submission Deadline:** *Sunday – November 30, 2025 by 23.59 sharp*
▪ EDA, Model training, Regression, K-fold CV, Bootstrapping	**Group size:** 1 to 2 persons
Submission Guidelines:	
▪ Submit all your deliverables (see on the last page) as a zip file on the Google Classroom.	

Dataset

The dataset contains students' assessment scores including assignments, quizzes, mid-terms, and final exam scores. The data has been anonymized to hide identities of the students and course(s). The data is shared in six excel sheets (1 to 6), where each sheet may contain a different number of assessments, total scores and weightages.

Problem

We are interested in the following three research questions.

- RQ1: How accurately can we predict student marks in Midterm I?
- RQ2: How accurately can we predict student marks in Midterm II?
- RQ3: How accurately can we predict student final examination marks?

Your task is to perform EDA (include visualizations for data understanding), necessary preprocessing, and model training using the regression models covered in the course (i.e., simple, multiple, polynomial). However, keep the domain knowledge in consideration during model training (e.g., using mid-term II marks to predict mid I mark is an incorrect approach). After preprocessing, you must combine records of all sheets to build a complete dataset.

- You must build at least two different models for each of the RQs above.
- Perform bootstrapping (500 samples, use train data only) and report 95% confidence

interval of the MAE and interpret your findings.

- Evaluate the models using MAE, RMSE, and R^2, and interpret your models.
- Also compare your results with the dummy/baseline models (explore the *dummy regressor* in Scikit-learn).
- Create a comparison table showing all the results of (at least) three models for each RQ (i.e., two variants of the regression and a dummy).
- Show the train and test accuracies of your "best model" to show model overfitting/underfitting.

CAVEAT: beware of the "data leakage" (during preprocessing)!

Deliverables

1. Preprocessed dataset (used for model fitting)
2. Jupyter notebook (you must have a good understanding of all the code written exceptions are there for data visualizations)
3. Build an interactive dashboard on Streamlit (https://streamlit.io/) or Gradio (https://www.gradio.app/) to present your results.
4. A diagram representing your workflow/pipeline of all the steps performed (including cleaning/preprocessing, transformation, data split, etc.). Include this in the notebook and dashboard.

References:

Articles on data leakage in ML:

- https://machinelearningmastery.com/data-leakage-machine-learning/
- https://www.analyticsvidhya.com/blog/2021/07/data-leakage-and-its-effect-on-the-performance-of-an-ml-model/