**CS 49200 Machine Learning**

**Homework 3: Data Analysis**

**Due on February 13, 2025, 11:59 PM**

The goal of this homework assignment is to help you get familiar with pandas, numpy, matplotlib, and sklearn Python libraries for data analysis and processing. Please refer to the class example in file *“end\_to\_end\_project.ipynb”* inside the zipped file *“ML\_examples.zip”* that was posted in Brightspace.

In Brightspace, under *“Content/Homework/HW 3”* module, you can find an Excel file *“extra\_credits\_dataset.xlsx”* that contains students’ report *“0.1 Extra Credit”* bonus points and their final scores, which the instructor collected from Fall 2015 to Spring 2023. The *“Extra”* column records students’ report *“0.1 Extra Credit”* bonus points, whereas the *“Total”* column contains students’ final scores.

Please analyze the dataset using Jupyter Notebook at Google Colab and answer the following questions:

1. How many students reported their *“0.1 Extra Credit”* bonus points? 503
2. What are the mean, the minimum, and the maximum of students’ reported *“0.1 Extra Credit”* bonus points? Mean, Min, Max of Extra Credit: 0.7976540755467197 0.1 5.0
3. Among students who reported their extra credits, what is the percentage of students who reported no more than 1 point? Percentage of students reporting ≤1 extra credit: 73.95626242544732
4. What is the correlation coefficient between students’ extra credits and their final score? Is the correlation positive or negative? Correlation Coefficient: 0.18706068470027773 (Positive)
5. Please plot a scatter figure to show a student’s extra credits and their final score, where the x-axis represents a student’s extra credits and the y-axis represents their final score.

A graph of blue dots

AI-generated content may be incorrect.

1. For this task and the remaining tasks, only consider students who reported *“0.1 Extra Credit”* bonus points. That is, remove the data points with missing data from the dataset. We want to use the student’s extra credits to predict their final score using a linear regression model and all data samples. What is the linear equation?

Linear Regression Equation: y = 2.7867x + 84.5402

1. What is the root mean square error (RMSE) of the prediction of the linear regression model when the model is applied to all data samples?

RMSE: 10.711492451359439

1. Please plot a figure to show a student’s extra credits and their final score in scatter points (i.e., the same as Task 5) and the linear regression model prediction line from Task 6, where the x-axis represents a student’s extra credits and the y-axis represents their final score.

A graph with blue dots and red line

AI-generated content may be incorrect.