
Design and Analysis of Algorithms

# SOFE3770U Final Project: Part 1: Data Cleaning & Linear Regression

| | |
|---|---|
| **Course Name** | Design and Analysis of Algorithms |
| **Course Code** | SOFE 3770U |
| **Course CRN** | 43513 |
| **Date** | Oct 24, 2025 |
| **Group Number** | Final Project 20 |
| **Group Member Names** | Vinujen Dilogen - 100870390<br>Rukshan Baskaran - 100864410<br>Khushi Patel - 100940709<br>Prabhnoor Saini - 100946515 |

## Objective

The goal of this project is to predict the overall State of Health (SOH) of a battery pack using a linear regression model, and then use that prediction to inform a chatbot system that can communicate the battery's health condition and answer user queries. The goal of this part is to develop a linear regression model to predict the State of Health of battery cells based on voltage readings. This model will aim to estimate the SOH and classify the battery as Healthy or Unhealthy, depending on a user-defined threshold (default 0.6)

## Dataset Description

The dataset used in this experiment is a PulseBat Dataset, which contains experimental measurements collected from a 21-cell lithium-ion battery pack. It consists of 670 samples and 30 columns, combining metadata, charge-discharge characteristics, voltage readings, and the targeted output variable, State of Health (SOH).

## Model and Evaluation Metrics

A linear regression model was trained using an 80-20 train-test split. Three evaluation metrics were used:

**$R^2$**: Measures how much variance in SOH is explained by the model

**MSE (Mean Squared Error)**: Penalizes larger prediction errors more heavily

**MAE (Mean Absolute Error)**: Average of the absolute errors, interpretable in the same units as SOH.

## Classification

A classification rule was applied for the given SOH, if the SOH was 0.8:

- **Healthy:** SOH >= 0.8, **Unhealthy:** SOH < 0.8

## Actual vs Predicted SOH Scatter Plot

This scatter plot compares the actual SOH value from the dataset against the predicted SOH values produced by the Linear Regression model. Each point in the plot represents one test sample from the dataset, x-axis showing the actual SOH value, while the y-axis shows the predicted SOH value. Points close to the red dashed diagonal line, where y=x, represent accurate predictions where the model output closely matches the real SOH. The color coding is based on the threshold value. Green points are predicted as healthy with SOH being greater than or equal to the threshold, while red points are predicted as unhealthy where the SOH is less than the threshold. This plot helps visualize how well the model predicts battery health, with points lying close to the diagonal indicating good correlation between predicted and actual SOH.

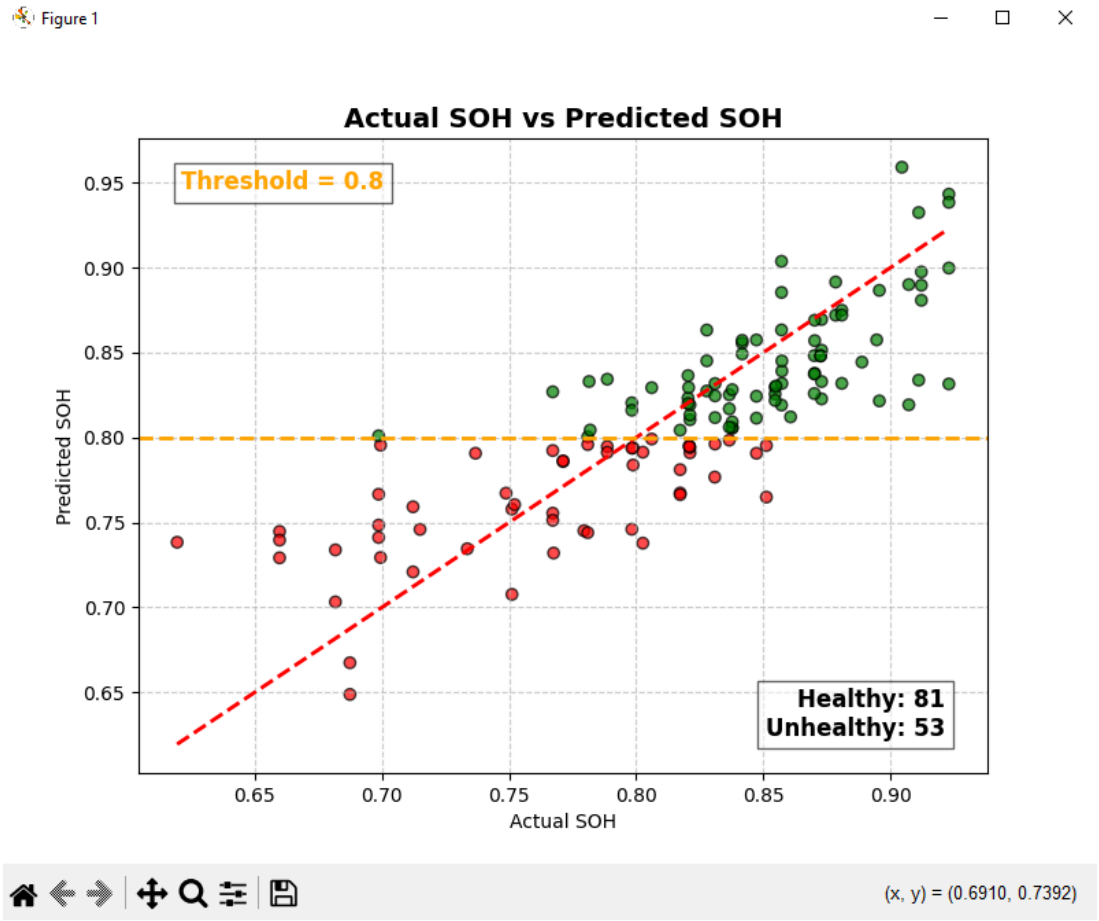


**Figure 1: Actual vs Predicted SOH Scatter Plot**

## Residuals (Error) Distribution Histogram

This histogram of the residuals represents the difference between the actual and predicted State of Health (SOH) values.

**Residual = Actual SOH - Predicted SOH**

The x-axis displays the range of residual values, showing how far each prediction deviates from the true SOH. The y-axis represents the frequency, the number of test samples that fall within each residual. The black dashed vertical line at zero marks the point of perfect prediction, where actual and predicted SOH values are equal. A smooth density curve (KDE_ is drawn on top of the bars to show the overall distribution trends. The annotation box (|Residual| > 0.05) counts the number of samples with an absolute error greater than 0.05, identifying them as potential outliers or less accurate predictions.

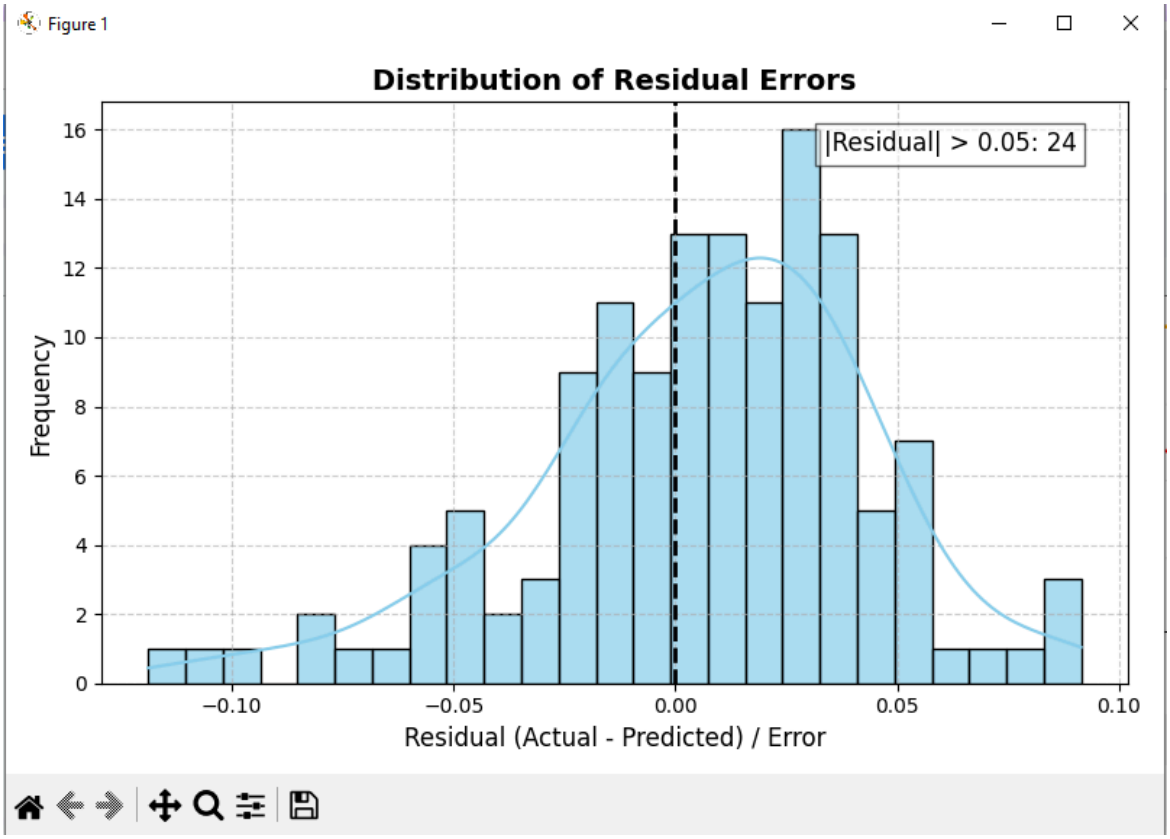


**Figure 2: Distribution of Residual Errors**

## Data Preprocessing and Linear Regression Model Performance on Unsorted, Ascending, and Descending Voltage Data

The output shows that the dataset was successfully loaded, cleaned, and analyzed. Only the voltage readings (U1-U21) and the target variable (SOH) were selected from the original file. While unnecessary columns were removed. No missing values were found, confirming that the dataset is complete. To test whether the order of voltages affects prediction accuracy, three dataset versions were created: unsorted, ascending, and descending. Each was used to train a Linear Regression model, and their performance was evaluated using $R^2$, MSE, and MAE metrics. All three versions produced similar accuracy, showing that sorting has minimal effect on results.

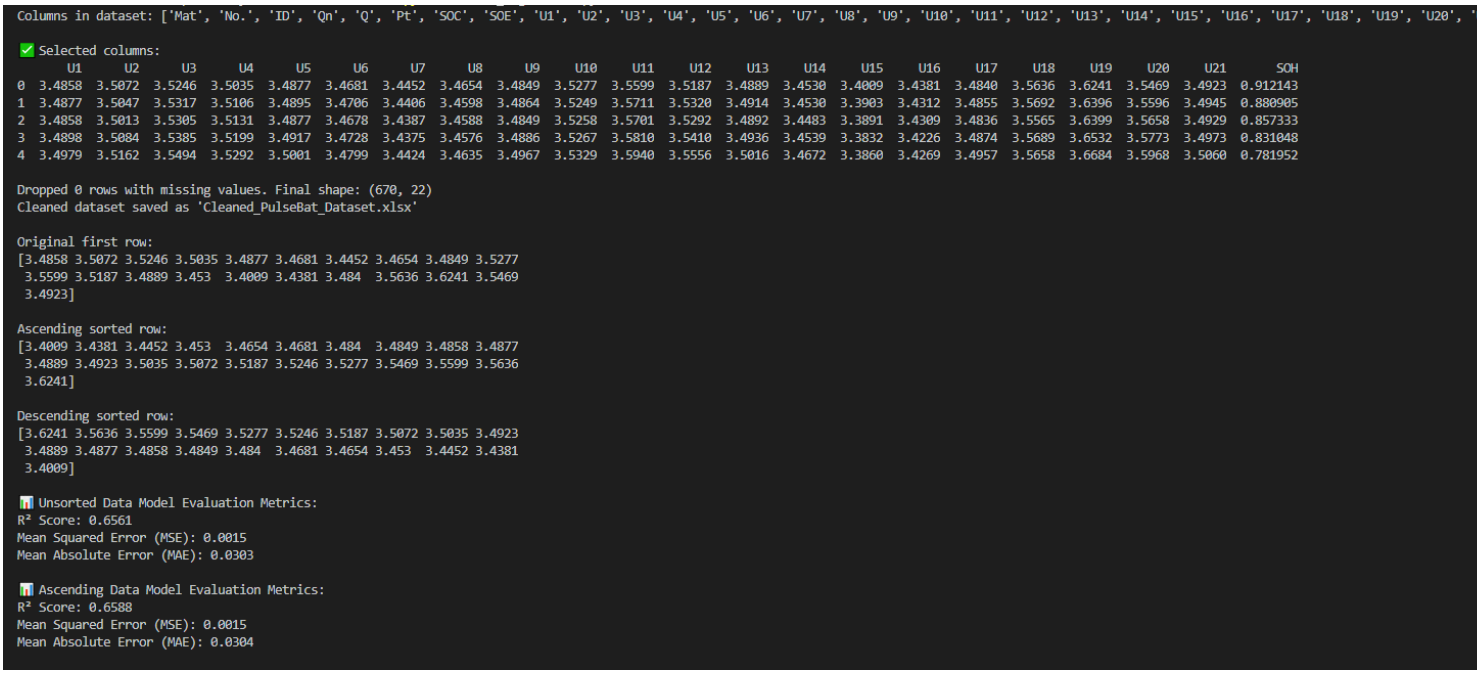


**Figure 3: Data Preprocessing and Linear Regression Model Performance**

## Descending Data Model Evaluation Metrics

This figure displays the performance metrics and sample classifications for the linear regression model trained on descending-sorted battery voltage data. The first value present is the $R^2$, which is the coefficient of determination. It shows how much of the variation in the actual SOH can be explained by the regression model. In this model, 65.88% of the variation can be explained, which is a moderate correlation and not perfect, but reasonably strong for battery data. MSE is the average of squared differences between predicted and actual SOH. The smaller the MSE, the better. The value shown, 0.0015, represents that most predictions are close to the actual value. Lastly, the MAE is the average absolute difference between predicted and actual SOH. It shows that on average, there is a 3% margin of error, which is extremely good. Moving on, the total execution time shows how long our Python script took to run, which includes data preprocessing, model training, evaluation, and plot saving. Finally, with the SOH threshold for classification entered as 0.8, with SOH >= 0.8 -> Healthy and SOH < 0.8 -> Unhealthy, the sample predictions with the classification table show how well your predictions matched with reality. You can see that most predictions match with only a few that are slightly off the near 0.8 threshold boundary.

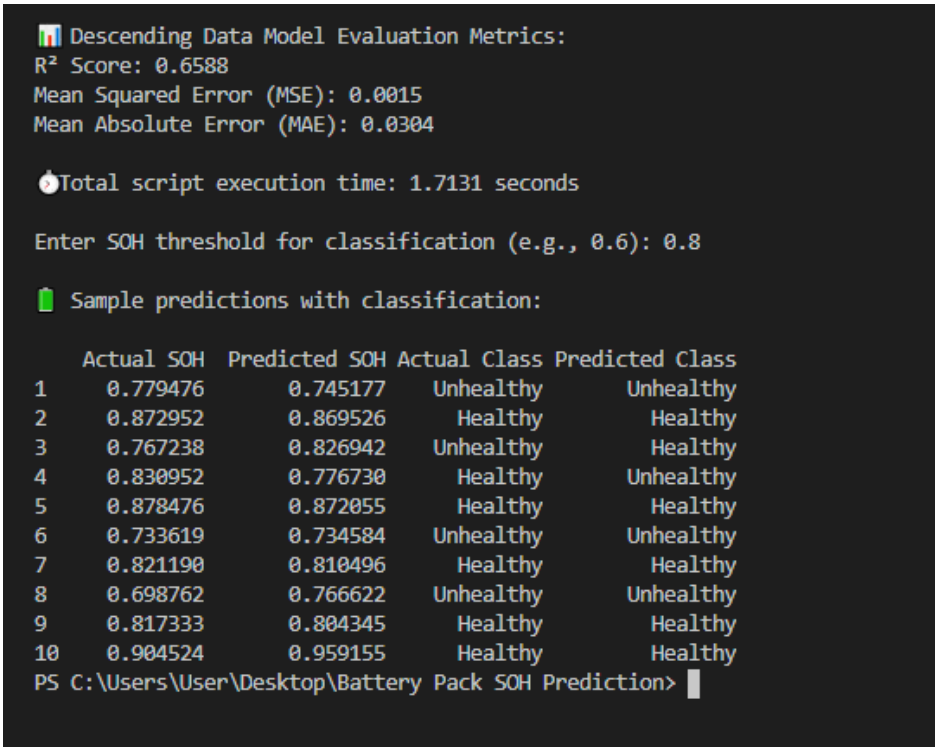


**Figure 4: Descending Data Model Evaluation Metrics in Terminal**