Project Process:

- Read and load data: There are 1000 rows and 6 columns
- Describe data: no. of rows, mean, standard deviation, etc.
- Check for missing values: no missing values
- Sample data visualization and corresponding inference

Project Process:

- Convert categorical data to integer: To prevent errors when training the model
- Define Performance Index as the target variable
- Define the rest as input features
- Data splitting: 20% of the data used for testing the model, 80% used for training

Project Process:

- Define model: Linear regression as lr
- Model training: Train linear regression model on data from x and y train
- Make prediction on training and testing set of x

Project Process:

- Performance evaluation: Normally, classification from sklearn is used to calculate accuracy, precision and F1 score, but because of an error message involving multiclass and continuous data, I decided to use mean squared error and r2 score metrics to calculate.
- LR Mean(Train): average value predicted by the LR model on the training data
- LR Mean(Test): average value predicted by the LR model on the testing data
- LR R2(Train): The R-squared (R^2) value on Train shows how well the model fits the training data. R^2 value of 0.9887 means that about 98.87% of the variance in the training data is explained by the model meaning it is an excellent fit.

Project Process:

- LR R2(Test): The R^2 value on Test shows how well the model generalizes to unseen data. The value of 0.9890 means that about 98.90% of the variance in the test data is explained by the model which shows the strong predictive power on the test set.
- Inference: The model performs very well, both on the training and test sets, with high R^2 values close to 1 showing a strong correlation between predicted and actual values.