Continuous Evaluation: 70%, Viva: 30%

Assignment 1:

i. Download House Prices Data Set from https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data. Analyze the features of the dataset.
Upload the dataset in the "ML_DRIVE/Assign_1" folder, if executing through COLAB. Access the dataset from there.

ii. Read the dataset in the Pandas data frame. Estimate the missing values with any technique of your choice. Divide the dataset into two sets using stratified k-fold cross validation technique entitled to train and test set respectively.

iii. Use the linear regression method to estimate the slope and intercept for predicting "SalePrice" based on "LotArea".

iv. Use the multiple regression method to estimate the value of the weights/coefficients for predicting "SalePrice" based on the following features:
  a. Model 1: LotFrontage, LotArea
  b. Model 2: LotFrontage, LotArea, OverallQual, OverallCond
  c. Model 3: LotFrontage, LotArea, OverallQual, OverallCond, 1stFlrSF, GrLivArea

v. Calculate and compare the Mean Squared Error, R2 score for each of the model using the training set and test set.

vi. Use the multiple regression method to estimate the value of the weights/coefficients for predicting "SalePrice" based on the following set of mixed (numerical and categorical) features:
  a. Model 4: LotArea, Street
  b. Model 5: LotArea, OverallCond, Street, Neighborhood
  c. Model 6: LotArea, OverallCond, Street, 1stFlrSF, Neighborhood, Year

vii. Compare the feature "LotArea" weights/coefficients for all the six trained models and plot a graph using the Matplotlib library.

viii. Use the polynomial regression of degree 2 and 3, to estimate the value of the weights/coefficients for predicting "SalePrice" based on "LotArea". Print the graph on the training and test set (Bonus).

Submit a report with the result.