IS670- Assignment 3 (25 pts)

Predicting Wine Quality

Objective: In this assignment, you will predict the **quality** of wines (target variable) based on various chemical attributes. The goal is to perform regression analysis using OLS, regression trees, SVR, and MLP models on a dataset, and then compare these models.

Dataset: You can use the "Wine Quality" dataset, which is available on the UCI Machine Learning Repository. This dataset contains information about various chemical properties of wines and their quality ratings (I have uploaded the dataset and the description file on Canvas).

Instructions:

1- **Data Preparation (5 points)**
- Load the Wine Quality dataset into Google Colab.
- Perform descriptive analytics (i.e., statistics, visualizations) to understand the dataset, the variables and the relationships among them.
- Handle missing values, highly correlated predictors (at 90%+ level), outliers, and encode categorical variables, if necessary.
- Split the dataset into training and testing sets.

2- **Simple Linear Regression (2 points)**
- Select a single chemical attribute (e.g., alcohol content) to predict wine quality using OLS.
- Fit a linear regression model, compute the R-squared, and make predictions.
- Visualize the results (any relevant visualization will work).

3- **Multiple Linear Regression (3 points)**
- Choose multiple chemical attributes as predictor variables for OLS regression.
- Interpret the coefficients and their significance and R-squared value.
- Describe the significance of top 3 most important variables and how they change with the target variable

4- **Regression Trees (3 points)**
- Build a regression tree model with a desired setting for max_depth.
- Visualize the tree and make predictions.
- Experiment with different depths and variables to create a more accurate tree.

5- **Support Vector Regression (SVR) (3 points)**
- Run SVR models with different hyperparameters.
- Evaluate the SVR models and interpret their performance.

6- **Multi-layer Perceptron (MLP) (3 points)**
- Build MLP models with different architectures.
- Evaluate the MLP models and interpret their performance.

7- Model Comparison (5 points)

- Create a table comparing MAE and RMSE for all models you have created so far and discuss the performance and trade-offs of each model. I strongly suggest creating a single multi-class bar chart that encompasses the MAE and RMSE for all models.
- Consider which model is the most suitable for predicting wine quality and explain why.

8- HTML Report (1 points)

- Create an HTML report in Google Colab that documents the entire analysis.
- Include explanations, code, visualizations, and conclusions in the report.

9- Extra credits (5 points):

Follow the steps above on another dataset of your choice to do a numeric prediction task such as predicting stock prices, car values, or real estate/house prices. The data can be from UCI repository, Kaggle, Github, or any other publicly available source. However, it should contain at least 10 columns and 500 rows. You should submit your dataset, so I can check your implementation.

10- Extra credits (5 points):

It's interesting to note that the wine dataset can also be analyzed using classification methods by categorizing wines based on their quality. Can you explain the advantages and disadvantages of employing the classification approach? Next, apply the classification algorithms you've learned so far to identify the best performing model (in your report include only the top three ones)? How do the results compare to those obtained from the regression models? Finally, based on your findings, which approach do you prefer, and why?