IDATG2208 - Mandatory Assignment - 1

Deadline for submission - 26 Sep, 2025

Dataset

Download the **Wine Quality Dataset** from the UCI repository (also available from Kaggle `https://www.kaggle.com/datasets/yasserh/wine-quality-dataset`). This dataset focuses on red variants of the Portuguese "Vinho Verde" wine. It contains measurements of various chemical properties found in the wine and their influence on its overall quality. The data can be used for either classification or regression tasks. Note that the quality ratings are ordered but not evenly distributed—there are significantly more average-quality wines than those rated as excellent or poor. Your objective is to build a model that predicts wine quality based on the provided features.

- Features: physicochemical properties (e.g., alcohol, pH, citric acid, density).
- Target variable: `quality` (integer score from 0-10).

The answers should be submitted through a pdf document on Blackboard. The answers should be supplemented with the figures and code snippets. Alternatively, solutions can be provided on Github (link should be provide on Blackboard).

Each question is mandatory and needs to be answered. A minimum of 70 points is needed for the solution to qualify as valid submission for this mandatory assignment.

1 Exercise-1 [60 Points]

1.1 Data Exploration [4 Points]

Q1.1.1 Load the dataset into a DataFrame and display the first 5 rows. Print the dataset information and summary statistics.

Q1.1.2 Which features show the highest variation based on summary statistics?

1.2 Correlation Analysis [12 Points]

Q1.2.1 Compute the correlation matrix of all features.

Q1.2.2 Plot a heatmap of the correlation matrix.

Q1.2.3 Which variable has the strongest positive correlation with `quality`? Which variable has the strongest negative correlation with `quality`?

Q1.2.4 Between alcohol and pH, which do you expect to better predict wine quality? Justify your answer.

1.3 Linear Regression [12 Points]

Q1.3.1 Fit a simple linear regression model using gradient descent to predict `quality` using only `chlorides`.

Q1.3.2 Fit a simple linear regression model predicting `quality` using only `alcohol`.

Q1.3.3 Report the regression coefficient and intercept and compare both the models.

Q1.3.4 Plot the regression line against the data points. Does the regression line fit the data well for chlorides or alcohol? Why or why not?

1.4 Train-Test Split [16 Points]

Split the dataset into training (80%) and test (20%) sets in 5 different folds. Train the simple linear regression model (using gradient descent) for each split on the train-test data in each fold. Evaluate the model on the test set in each fold using:

- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R^2 score

Q1.4.1 How well does alcohol alone predict wine quality in each split?

Q1.4.2 How well does chloride alone predict wine quality in each split?

Q1.4.3 Do you think the model underfits? Why?

Q1.4.4 Provide the mean and variance from the 5 different folds and comment on the variation in performance across all 5 folds when using alcohol versus chloride.

1.5 Multiple Linear Regression [16 Points]

Q1.5.1 Train a multiple linear regression model using all features to predict `quality` using the same splits as used in previous question. Evaluate the model on the test set using MSE, RMSE, and R^2.

Q1.5.2 Compare the results of simple vs multiple regression in terms of MSE, RMSE, and R^2.

Q1.5.3 Provide comparison plots for multiple versus simple linear regression solved in previous exercise. At-least one of the following plots among (i) Cost vs Iteration, (optimization) (ii) Parameter Convergence (coefficients) (iii) Predicted vs Actual (performance) and (iv) Residuals Plot (assumptions check) should be provided

Q1.5.4 Which model performs better and why?

2 Exercise-2 [40 Points]

Q2.1 Which features are most suitable/influential in predicting wine quality? (Tip - You can consider feature importance ranking.)

Q2.2 The models you trained so far assume a linear relationship between features and target.

a) **Polynomial regression:** Extend the feature space to include quadratic or interaction terms. Does this improve performance?

b) **Regularization:** Train models using Ridge and Lasso regression. How do these methods affect the coefficients and model generalization?

c) **Model comparison:** Compare your linear regression results to a non-linear model (e.g., Decision Tree or Random Forest). Which performs better, and why?