Machine Learning - Coursework

Assignment 01

Course Name: Machine Learning(24-25)

Course Number: COMP4139

Group Number: C4_18

Group Members:

- **Juntian Xiao**
- **Luqi Xin**
- **Guangzheng Dong**
- **Yuhong Yuan**
- **Qifeng He**

Github Link: https://github.com/CloudSkytec/ML-CW1

1. Introduction

- **Classification Dataset** : Wine Quality - UCI Machine Learning Repository
  - Two datasets are included, related to red and white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests.
- **Regression Dataset**: Abalone - UCI Machine Learning Repository
  - Predict the age of abalone from physical measurements

2. Method Description

- **Models**
  - *Linear regression*
  - *Logistic regression*
  - *Support vector machines (SVM)*
  - *Decision trees*
  - *Multilayer perceptron neural network*
- **Evaluate Methods**
  - *K-fold cross-validation*
  - *Mean squared error (MSE) for Regression*

2.1 Analyzing Data

Wine Quality

There are 11 features and 1 target value in this dataset, including: - Fixed Acidity - Volatile Acidity - Citric Acid - Residual Sugar - Chlorides - Free Sulfur Dioxide - Total Sulfur Dioxide - Density - pH - Sulphates - Alcohol The target value is: - Quality .

- Looking at the correlation matrix, it is evident that Fixed Acidity and pH are inversely correlated, as expected due to the relationship between acidity and pH levels.
- When summing up Free Sulfur Dioxide and Bound Sulfur Dioxide (derived from Total Sulfur Dioxide minus Free Sulfur Dioxide), it corresponds closely to Total Sulfur Dioxide.
- Some obvious outliers exist, such as extremely high values in Residual Sugar and Chlorides, which might indicate wines with significantly different processing or contamination.

Abalone

There are 8 features and 1 target value in this dataset, including 'Sex', 'Length', 'Diameter', 'Height', 'Whole_weight', 'Shucked_weight', 'Viscera_weight', 'Shell_weight', and target value 'Rings'

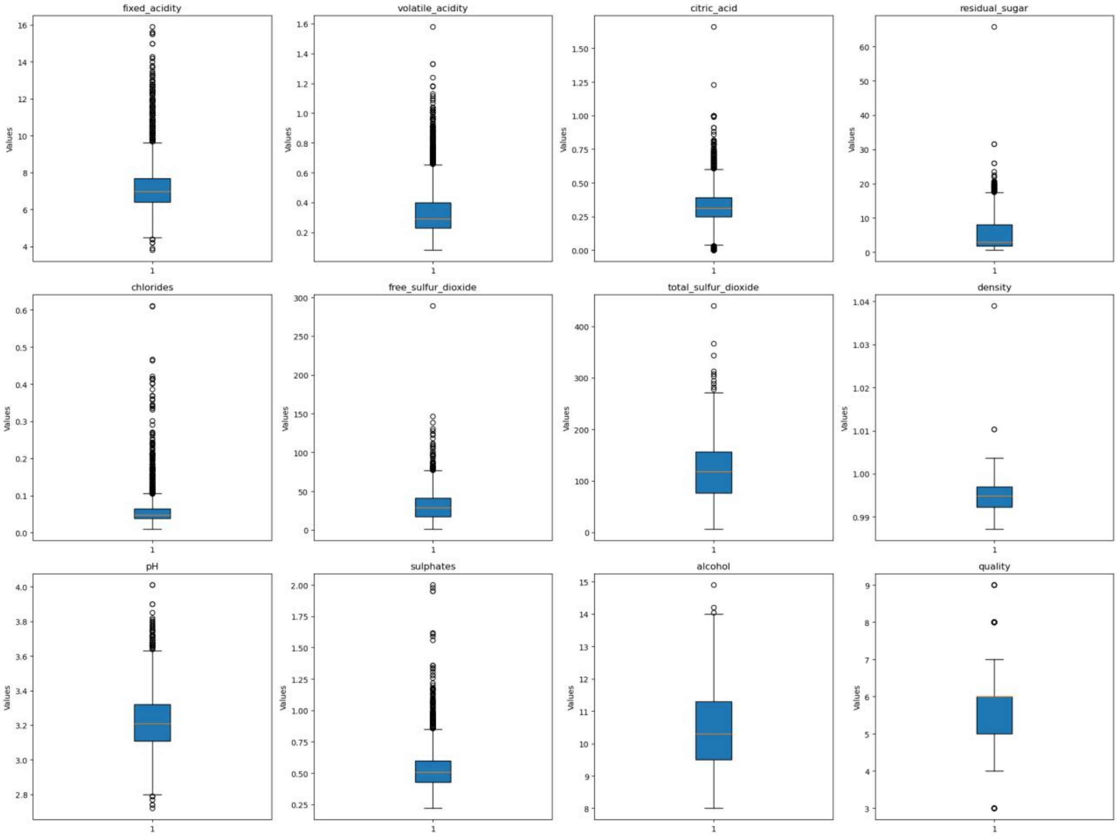
- Looking at graph it is evident that Length and Diameter are highly correlated.
- when summing up 'Shucked_weight', 'Viscera_weight', 'Shell_weight' it is quite close to Whole_weight.
- Some obvious outliers exist like 0 value and huge value in Height .

2.2 Data Preprocessing

Wine Quality

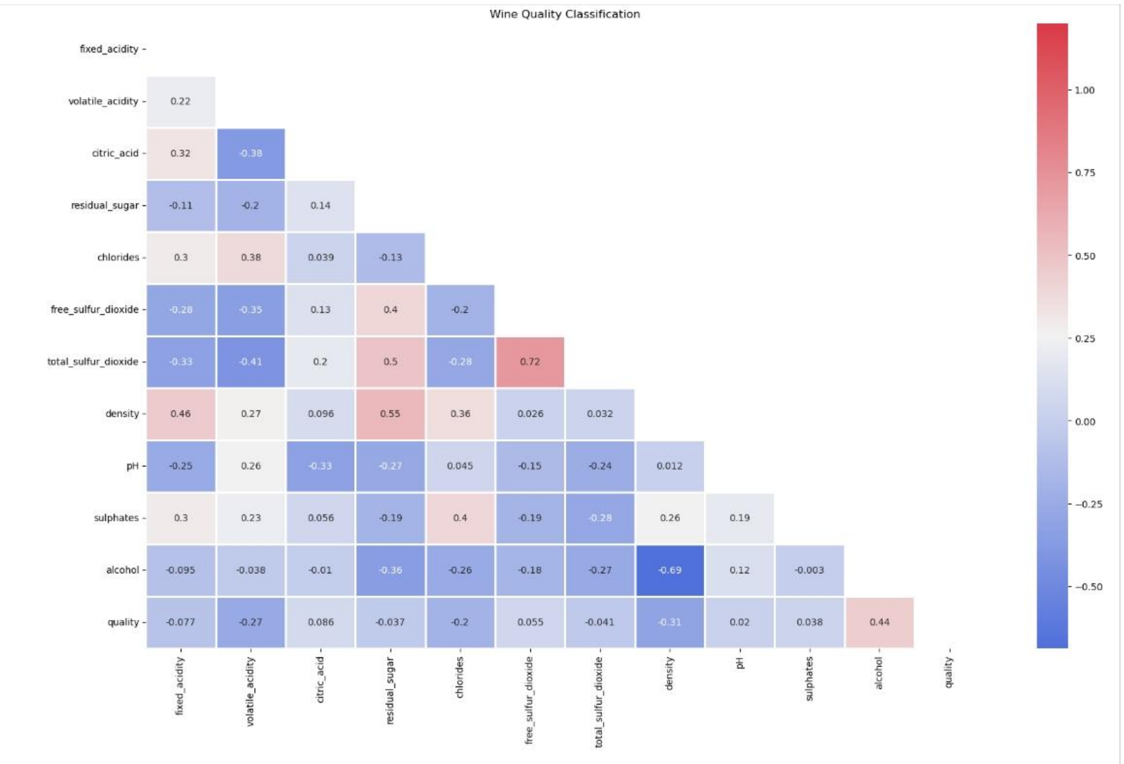
Outliers

The Interquartile Range (IQR) method was applied to define the bounds for detecting outliers. Certain features or extreme values in both the feature and quality are defined as meaningful and are retained based on the actual situation. For example, very low citric acid may indicate low-acid white wines or sweet wines, while high residual sugar could indicate sweet wines like Sauternes



Feature Engineering

According to the heatmap of correlation, high-correlated features free sulfur dioxide and total sulfur dioxide were selected and combined into free_sulfur_ratio to minimize the correlation.

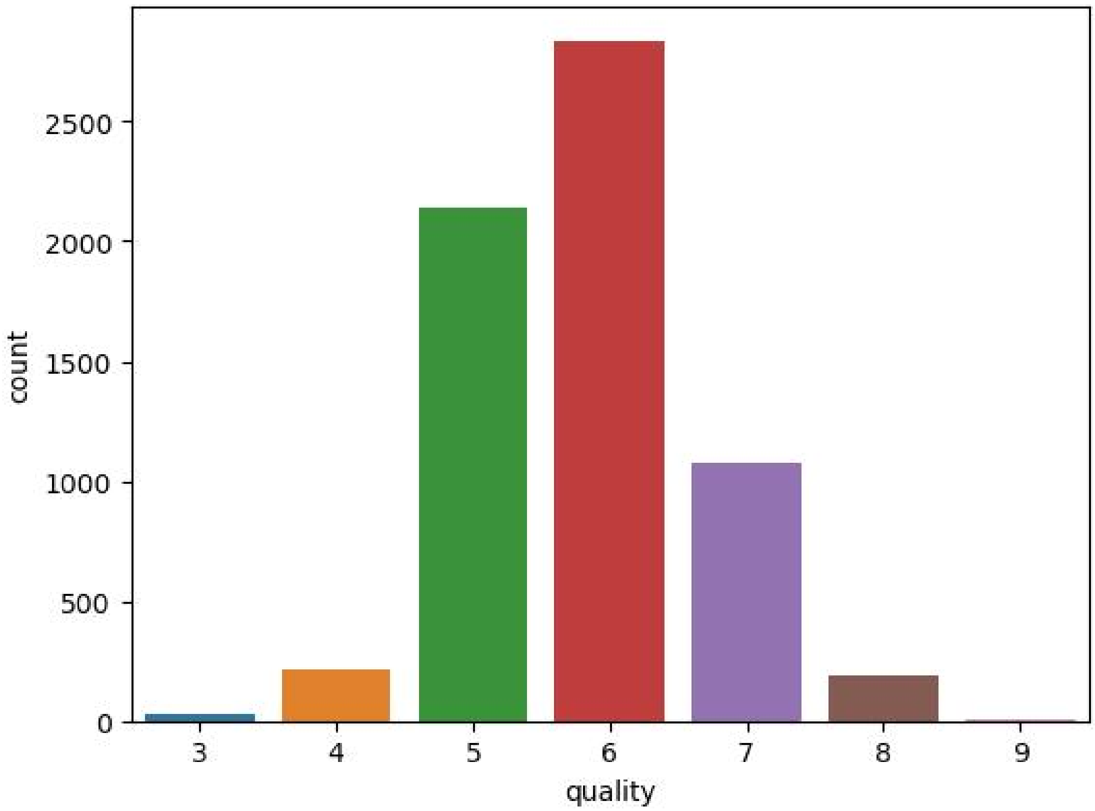


Scaling

Given that features like fixed acidity, volatile acidity, residual sugar, and alcohol have different ranges, Standard Normalization is applied to prevent features with varying scales from dominating the model.

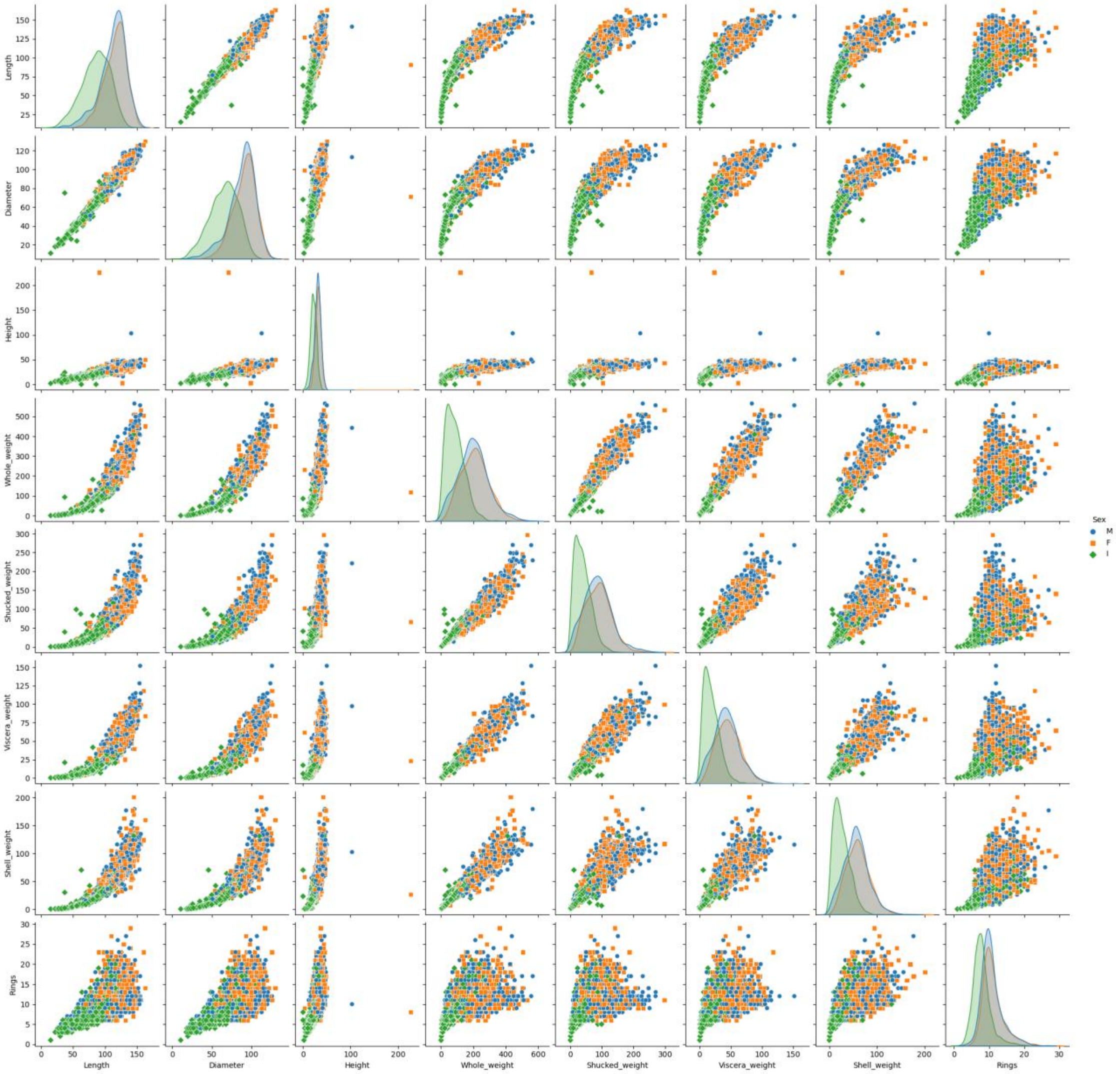
Oversampling

The count plot shows that the target variable quality is imbalanced, so SMOTE is applied to increase the samples for classes 3 and 9.

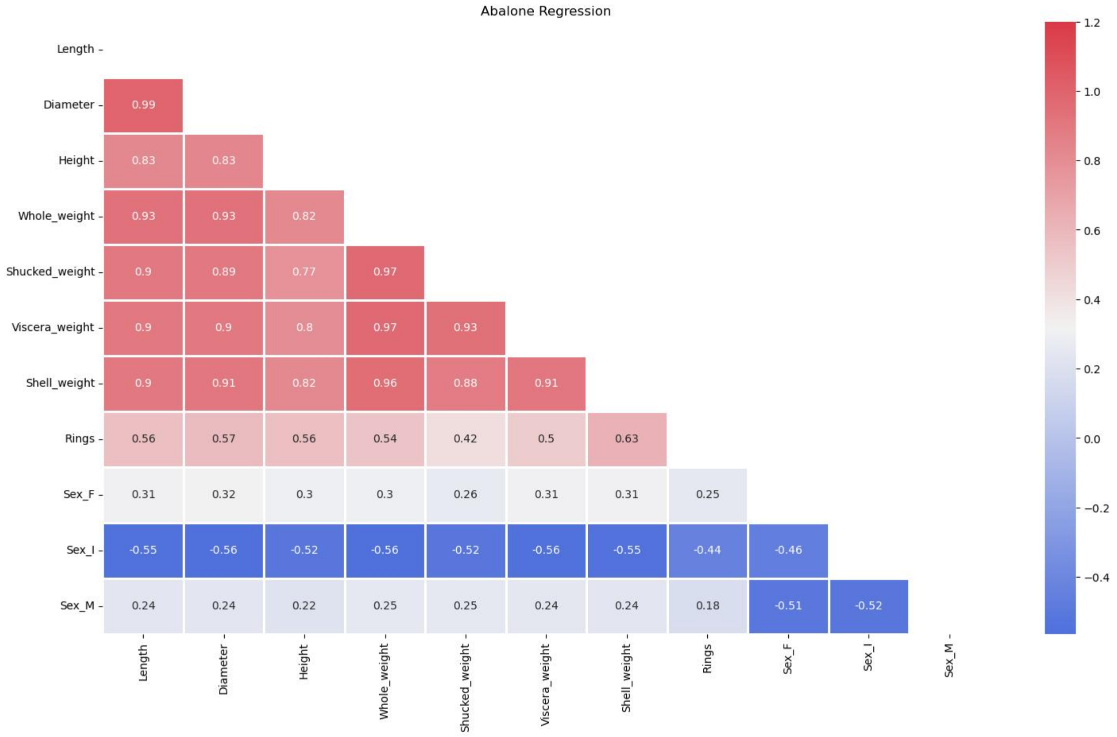


Abalone

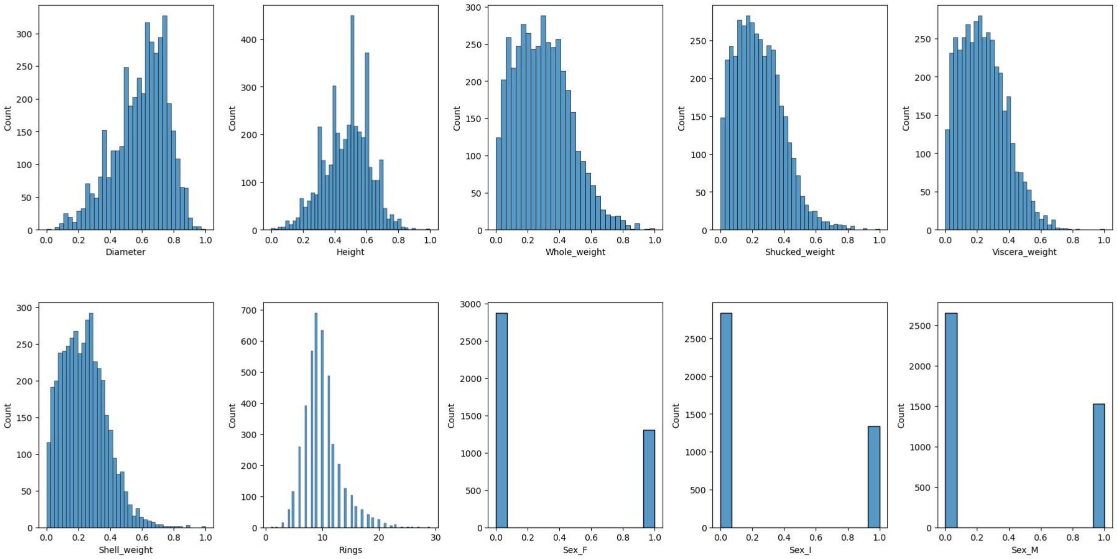
- Based on pairwise scatter plots of each feature, the correlations between features can be initially observed.



- Feature 'Sex' is a string type, in order to fit it into regression model, it should be applied One-Hot-Encoding.
- The graph indicates that 'Length' and 'Diameter' are highly correlated, since decided to remove 'Length'.



- IQR Method was applied to other features that also have outliers.
- Then using min-max-scaler to scale this dataset, a relatively balanced dataset was eventually obtained.



3. Parameters settings

3.1 Wine Quality

3.1.1 Decision Trees Classifier

parameters: **min_samples_split**=3, **random_state**=42

- **min_sample_split:** By increasing it, the total number of splits is reduced, which limits the number of parameters in the model and helps mitigate overfitting.

3.1.2 Logistic regression

parameters: C=1, **solver**='saga', **max_iter**=500, **penalty**='l1'

- **solver**: For multiclass problems ‘saga’ and ‘lbfgs’ can handle multinomial loss; By testing, we eventually chose 'saga'.

- **max_iter**: the model can not converge when using default value '100'. By increasing it the model successfully converged.

3.1.3 Support vector machines (SVM)

parameters: **kernel**='rbf', **gamma**='scale', **degree**=3, C=10, **random_state**=0

- **C**: By increasing C from 1 to 10, we improved model’s accuracy, which indicates that the SVM model is less tolerant of misclassifications and is better able to fit the training data closely.

3.1.4 Multilayer perceptron neural network

parameters: **hidden_layer_sizes**=(100,64,32), **max_iter**=300, **alpha**=0.1, **learning_rate**=0.005

- **Hidden Layer Size**: Add two hidden layers with decreasing numbers to improve accuracy and reducing the risk of overfitting.

- **alpha**: Alpha was initially set to 0.01, but due to persistent overfitting, it was increased to 0.1, which successfully reduced the variance to 4.7%

3.2 Abalone

3.2.1 Decision trees regression

parameters: **max_depth**=5, **criterion**='squared_error'

- **max_depth**: The maximum depth of the tree. Increasing this value may lead to overfitting, and smaller values may lead to underfitting by adjusting the maximum depth of the tree from 3-8 to determine 5 as the optimal parameter.
- **criterion**: Default option to minimize mean squared error

3.2.2 Linear regression

parameters: **fit_intercept**=True

- **fit_intercept**: Since the data is not centralized, setting fit_intercept=True allows for a better fit to the data, reducing model bias.

3.2.3 Support vector machines (SVM)

parameters: **kernel**='rbf', **gamma**='scale', **C**=100, **epsilon**=0.5

- **epsilon**: By increasing epsilon value, we increased tolerance for error, making the model more robust and reducing the risk of overfitting to small errors.

3.2.4 Multilayer perceptron neural network

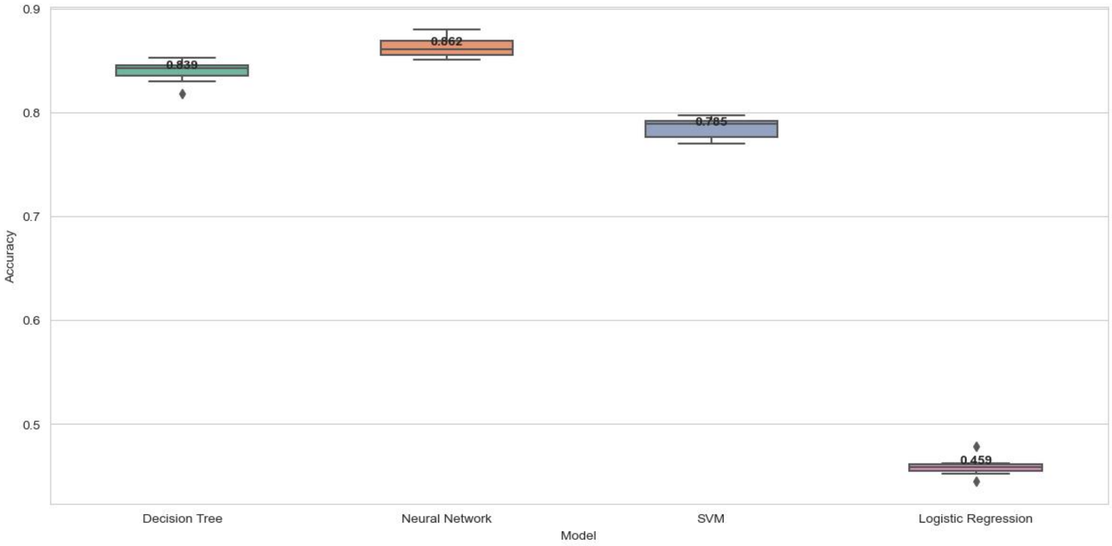
parameters: **hidden_layer_sizes**=(64, 32, 16, 8), **max_iter**=500, **learning_rate**=0.0015, **batch_size**=80

- **Hidden Layer Size**: Starting with 64 neurons captures broad features, and decreasing the number in deeper layers helps refine them to balance model capacity
  and reducing the risk of overfitting.
- **Learning Rate**: Increasing learning rate helps accelerate convergence. However, when it increased to 0.002, the model experienced premature fluctuations, indicating overfitting. Setting the learning rate to 0.0015 proved to be the most effective.

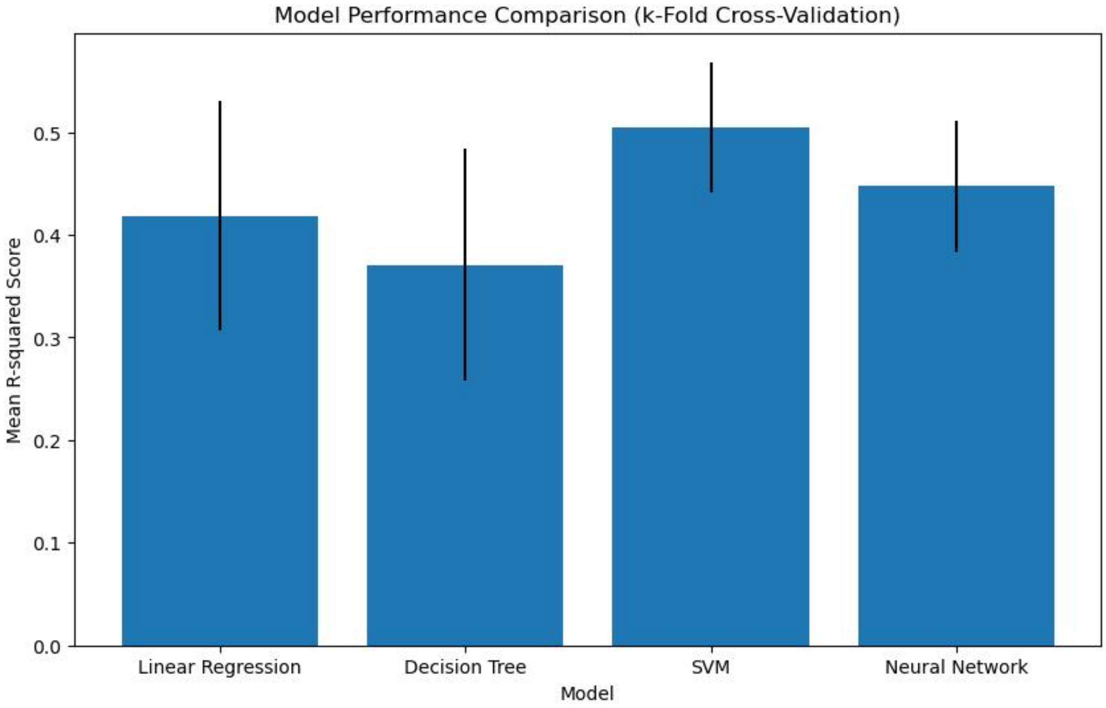
- **Batch Size**: A batch size of 80 can provide a stable estimate of the gradient and avoid overfitting.
- **Max Iterations**: The epochs of 500 can not only effectively reduce the model's MSE but also maintain stability without significant fluctuations.

4. Assessment

Wine Quality



Abalone



5. Conclusion

Different types of datasets have corresponding models that are suitable for them

The features of dataset ***Abalone*** shows a clear linear relationship with the target values. While the features of ***Wine Quality*** did not show a clear linear relationship, by analyzing the correlation of each feature with quality, it is evident that customers tend to choose specific ranges of parameters such as 'alcohol', 'residual sugar', 'total SO2', and so on. It turns out that the combination of specific range of parameters is crucial. Therefore, the performance of **logistic classifier model** on this dataset is unsatisfactory.

Data preprocessing matters

Outliers can significantly affect models, especially some models could be sensitive to them. Appropriate preprocessing can make the distribution of feature more balanced, thereby making the model easier to converge. What's more, different data processing methods should be applied to different models to achieve the best fit.