

Wine Quality Prediction Machine Learning Summary

1. Data Import and Preprocessing:

- The dataset is loaded and checked for missing values, which are then imputed with the mean of their respective columns.
- Redundant features (highly correlated) are identified and removed from the dataset.

2. Exploratory Data Analysis (EDA):

- Histograms are plotted to visualize the distribution of continuous features, giving insights into their spread and central tendency.
- A count plot is drawn to display the number of data points for each quality of wine, providing an understanding of the dataset's class distribution.
- A heatmap is generated to identify highly correlated features, aiding in feature selection or engineering.

3. Model Development:

- The dataset is prepared for training by segregating features and the target variable.
- The target variable is transformed into a binary format suitable for classification.
- The dataset is split into training and test sets using an 80:20 ratio.
- Data normalization is performed using MinMaxScaler to ensure stable and efficient training.
- Three classification models (Logistic Regression, XGBoost, SVC) are trained on the training data.

4. Model Evaluation:

- The trained models are evaluated using the validation data.
- The accuracy of each model on both training and validation data is printed, helping to assess model performance.
- A confusion matrix is plotted for the best performing model (Logistic Regression), providing insights into the model's predictive performance on the validation data.
- A classification report is printed for the best performing model, providing detailed metrics such as precision, recall, and F1-score for each class, aiding in understanding the model's overall performance.

5. Summary:

- In summary, the code provides a comprehensive pipeline for wine quality prediction, including data preprocessing, exploratory data analysis, model training, evaluation, and performance analysis. It helps in understanding the dataset, selecting suitable features, training and comparing different machine learning models, and evaluating their performance for the task at hand.