Iran Izadyariaghmirani

Practical Machine Learning **Homework 3: ML Algorithm Selection**

I. Introduction

This investigation explores the efficacy of various machine learning models to be applied on the "Winequality-red" dataset. The objective of the ML task is predicting the quality of the red wine based on its physicochemical properties. The goal is to determine the most accurate model using an AutoML framework, which simplifies the selection process through automation. By analyzing the intrinsic characteristics of wine, the objective is to identify a model that can reliably predict quality.

II. Dataset description

The dataset utilized in this exercise is the "winequality-red" dataset, which includes data on various features of wines, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The dataset comprises a total of 1,599 rows and eleven features. The target variable, "quality" rates the wine on a scale from zero to ten (0-10). There are no missing values in the dataset. Data splitting is performed to separate the dataset into training and testing sets, with 80% of the data allocated for training and the remaining 20% for testing.

III. Methodology

Utilizing Python, we process the data with Pandas and NumPy and visualize results using Matplotlib library. The Scikit-learn library supports our machine learning algorithms. We preprocess the data with a logarithmic transformation to reduce skewness and standardization for scale normalization, performed by a "**ColumnTransformer"**.

We evaluate three classifiers: *Random Forest, Support Vector Machine (SVM)*, and *Logistic Regression*, each selected for their distinct learning strategies and potential for high performance. The hyperparameters of these models are listed in Table 1.

The evaluation of model performance relies on a combination of 5-fold cross-validation on the training sets and accuracy measurement on the test datasets. The cross-validation process allows the model to mitigate overfitting by providing a more generalized performance indicator across multiple training subsets, while test accuracy offers a straightforward assessment of how well each model predicts unseen data.

IV. Results

The results of performing the developed models on the training and tests datasets are summarized in Table 1, and the accuracy scores have also been depicted in a bar chart (see Figure 1). The mean cross-validation (CV) accuracy scores demonstrate that the **Random Forest model** had a clear lead in both mean CV and test accuracy scores, indicating its superior performance in predicting red wine quality within this dataset.

Moreover, for each model, the observed difference between the CV and test accuracy score was negligible, which implies none of the models contained significant overfitting. In other words, all models, including the Random Forest, SVM, and Logistic Regression, were well-fitted and capable of generalizing their predictions to unseen data. The performance of the models demonstrate a good balance between learning from the training data and maintaining flexibility to apply this knowledge to new, unseen data effectively.

Table 1: The hyperparameters of the classifiers and the results of evaluation tests

Classifier	Hyperparameters	Mean CV accuracy	Test accuracy
Random Forest	N_estimators=100, criterion='gini', max_features='auto', random_state=42	0.6779	0.6469
Support Vector Machine	C=1.0, kernel='rbf', gamma='scale' random_state=42	0.5943	0.5750
Logistic Regression	C=1.0, solver='lbfgs', penalty='l2', max_iter=1000, random_state=42	0.6044	0.5719

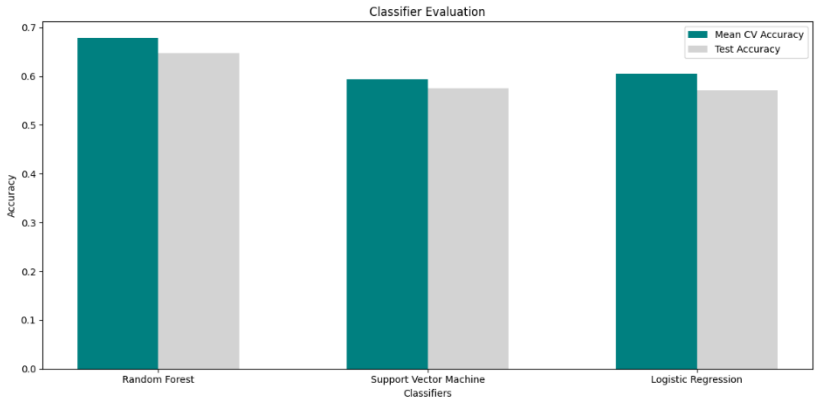

Figure 1: Comparative performance of the machine learning models on the "Red Wine Quality" Prediction

V. Conclusion

In this study, we evaluated several machine learning models to determine which could best predict red wine quality. The results indicated that the Random Forest model demonstrated the highest average accuracy in both cross-validation and test datasets. This suggested a superior ability of the Random Forest model to generalize beyond the training data, making it the most effective model among those tested for predicting the quality of red wine. The similarity between the mean CV and test accuracies for all models implied that there was no significant overfitting occurring with any of the models. The evaluation based on cross validation and test accuracy scores provides a framework to determine the most accurate classifier model and simplifies the selection process.

VI. References

1. https://www.openml.org/search?type=data&sort=runs&id=1003&status=active

2. “Winequality-red” dataset - UCI Machine Learning Repository

3. Wine Quality Prediction Using Machine Learning. https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/

4. Dahal, K. R., et al. "Prediction of wine quality using machine learning algorithms." Open Journal of Statistics 11.2 (2021): 278-289.