COSC 4557/5557 Practical Machine Learning Spring 2024
ML Algorithm Selection
Submitted by: Iqbal khatoon

Introduction

The quality of wine is influenced by its intrinsic characteristics, which can be quantitatively assessed through various physicochemical properties. In an effort to automate the process of model selection and enhance predictive accuracy, this study employs an AutoML framework. AutoML not only simplifies the modeling process but also systematically evaluates multiple machine learning algorithms to determine the most effective model for predicting wine quality. The focus is on leveraging the AutoML's capability to handle preprocessing and model selection efficiently thereby identifying a model that can reliably predict the quality of red wine. This approach aims to minimize human intervention and bias in the model selection process, ensuring that the best-performing model is chosen based on objective performance metrics.

Wine Quality Dataset

ML algorithm selection exercise employs the "winequality-red" dataset, containing information on different attributes of red wines. These attributes encompass fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. The dataset consists of eleven features. The target variable, denoted as "quality," assesses the wine's quality on a scale ranging from zero to ten (0-10). Based on our data analysis on the provided data set, we can ascertain that the dataset contains a total of 1599 entries, each with non-null values across all features and the label. This implies that there are no missing values present in the dataset, which is a positive aspect for our analysis. Furthermore, upon inspecting the data types assigned to each column, we observe that all features have been appropriately assigned the 'float64' data type, indicating numerical values. Similarly, the label class 'quality' comprises integer values exclusively, aligning with its assigned 'integer' data type.

Preprocessing Steps

The preprocessing involved two main steps:

- Log Transformation: Applied to reduce skewness in feature distributions.
- Standard Scaling: Used to normalize features, thus ensuring that they contribute equally to the analysis.

Model Training

Four models were trained on the "winequality-red" dataset:

- Linear Regression
- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)

Each model was trained on both raw and preprocessed data. The data was split into training (80%) and testing (20%) sets, with model performance evaluated on the testing set.

Evaluation Metrics

- Linear Regression: R^2 score
- Classification Models (Logistic Regression, Random Forest, SVM): Accuracy

Results

The performance of each model on the raw and preprocessed data is summarized below in Figure 1 and Figure 2.

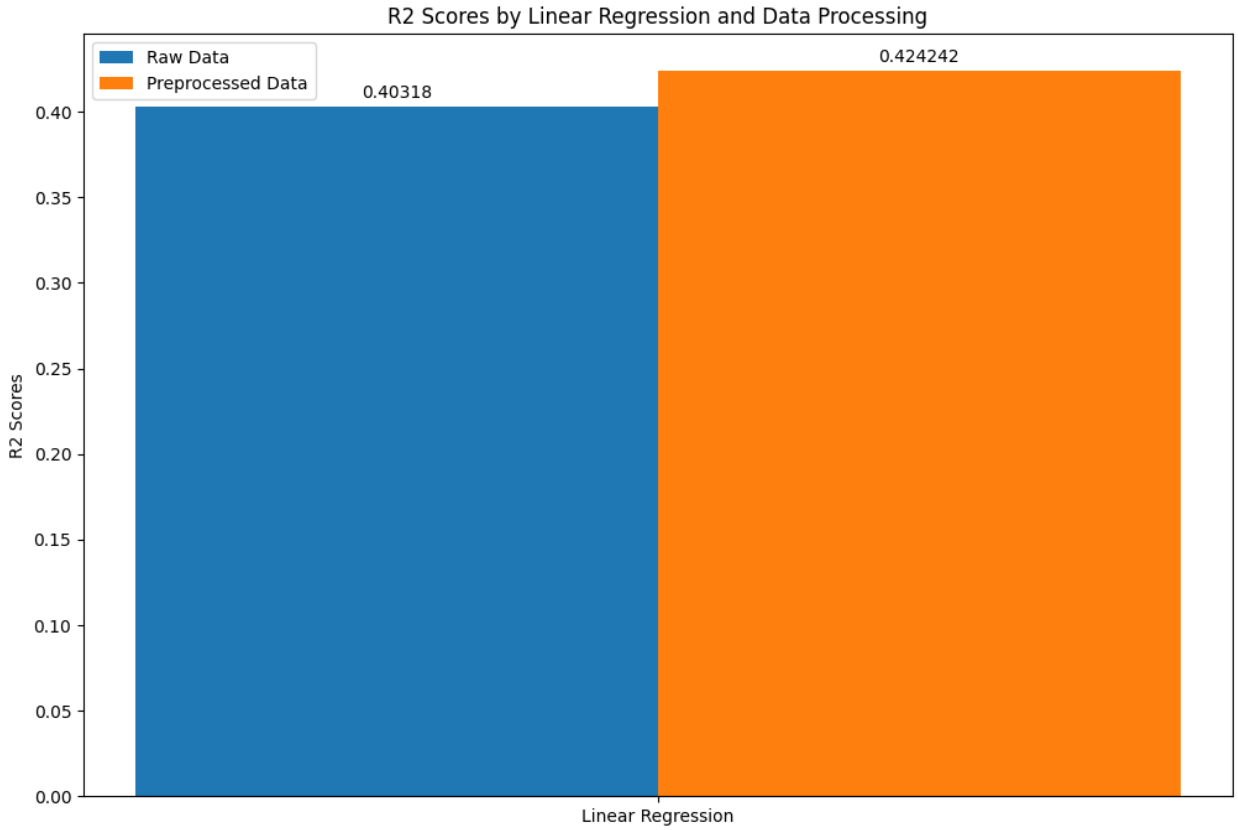

Figure 1: Performance of Linear Regression model on the "Red Wine Quality" Prediction

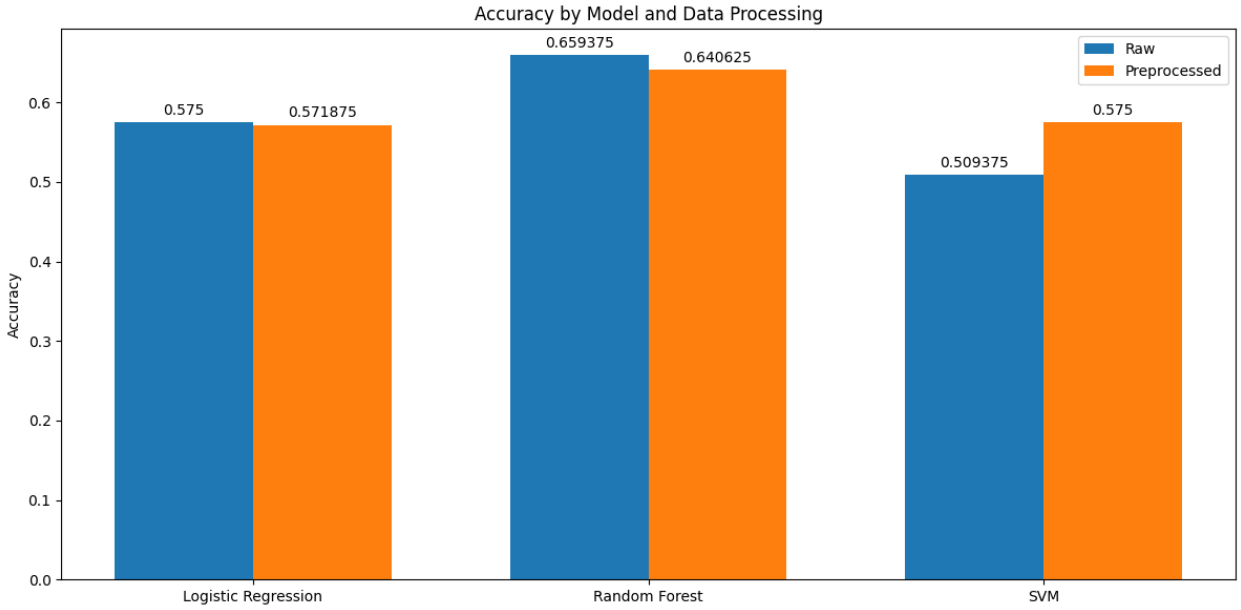

Figure 2: Performance comparison of the machine learning models on the "Red Wine Quality" Prediction

Analysis

The results indicate that preprocessing generally improves the performance of linear regression and SVM. The Random Forest model shows a slight decrease in performance with preprocessing, which might be due to the model's inherent capability to handle raw, untransformed data effectively. Logistic Regression's performance remained largely unaffected by the preprocessing.

Conclusion

In this study, we systematically evaluate several machine learning models for predicting the quality of red wine based on its physicochemical characteristics. The analysis encompassed both raw and preprocessed data to discern which models most accurately predict wine quality. Among the tested models, the Random Forest classifier demonstrated the highest accuracy on raw data, suggesting its robust capability to generalize well beyond the training dataset. This superiority in performance makes Random Forest the most effective model for this task among those evaluated.

The preprocessing steps improved the performance of the SVM and Linear Regression models, indicating that certain models benefit more from data normalization and transformation. Notably, the consistent accuracy rates between the raw and preprocessed datasets for most models suggest minimal overfitting, underscoring the effectiveness of the AutoML framework in selecting well-tuned models. The comprehensive evaluation using both raw and preprocessed data provides a robust framework for identifying the most accurate model, thereby streamlining the selection process in predictive modeling of wine quality.

References

[1] Wine Quality - UCI Machine Learning Repository
[2] **https://scikit-learn.org/stable/**