ML Algorithm Selection

Mohammad Irfan Uddin

Introduction:

The objective of this report is to explore an Automated Machine Learning (AutoML) approach for the selection of the most suitable machine learning model for a given task. In this endeavor, we emphasize the methodological aspect of model selection, prioritizing the process itself over the predictive performance outcomes. This report serves as an instructional guide on how to leverage AutoML techniques for identifying the optimal machine learning algorithm for a given problem.

Dataset Description:

The dataset under consideration, known as the Wine Quality Dataset, is composed of 11 features and contains 1599 samples. The target variable is wine quality, which is a categorical attribute. One notable aspect of this dataset is the absence of missing values, simplifying the preprocessing phase. From dataset, we establish a clear understanding of the dataset's characteristics and set the stage for subsequent model selection.

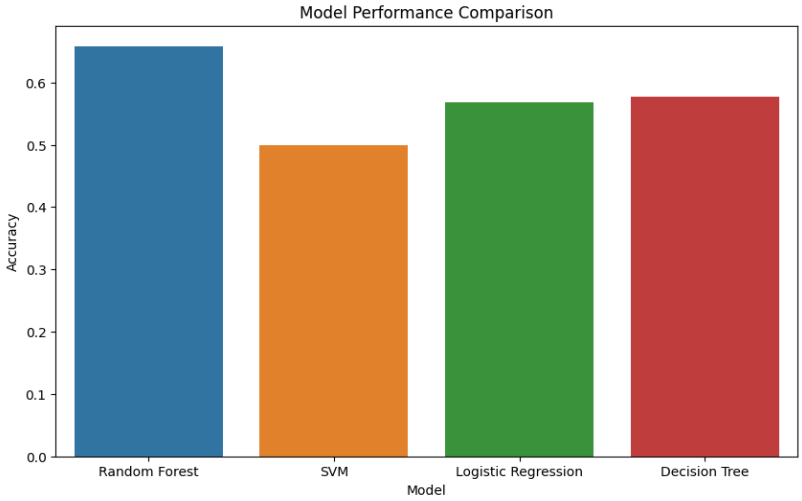
Experimental Setup:

The practical implementation of the AutoML approach begins with the selection of programming languages and libraries. In this case, Python is the chosen language, and we utilize essential libraries such as Pandas for data manipulation, Scikit-learn for machine learning, SciPy for statistical analysis, and Warnings for managing alerts. The dataset is loaded and subsequently divided into training and testing sets through an 80-20 split.

Four machine learning algorithms are considered for evaluation: Random Forest, Support Vector Machine (SVM), Logistic Regression, and Decision Tree. To assess their performance, we employ a 5-fold cross-validation strategy, with the primary evaluation metric being accuracy. Furthermore, statistical tests are used to compare the performance of these models to select the most suitable algorithms for further analysis.

Results:

The results of the model evaluation are summarized as follows:



P-value for Random Forest vs SVM: 0.0002961522700452354

P-value for Random Forest vs Logistic Regression: 0.0027856183135524255

P-value for Random Forest vs Decision Tree: 0.003304394074522585

P-value for SVM vs Logistic Regression: 0.00042009485006459595

P-value for SVM vs Decision Tree: 0.00027634985948454534

P-value for Logistic Regression vs Decision Tree: 0.03277166790018237

I have used pairwise t-test to compare the performance of models. A nested loop is used to compare each pair of models. The paired t-test is applied to the cross-validated accuracy scores of two models. The resulting p-value indicates whether there is a statistically significant difference between the performances of the two models. With significance level 0.05, Random Forest, SVM, Decision Tree and Logistic regression are selected Algorithms.

Specifically, Random Forest demonstrates statistically significant superiority over SVM, Logistic Regression, and Decision Tree. SVM also outperforms Logistic Regression and Decision Tree, while Logistic Regression performs better than Decision Tree. Each machine learning algorithm is evaluated using cross-validation, and their accuracy scores are calculated. The best-performing model, among the selected algorithms, is Random Forest Regression. It achieves the highest accuracy score.