Machine Learning Algorithm Selection

Narges Helmisiasifariman

University of Wyoming

Introduction

In this paper, a set of machine learning algorithms has been chosen to predict the red wine quality based on its attributes. To evaluate the performance of ML models, a compression of all performance metrics is used for scaled and unscaled data. For this purpose, random forest, supper vector machine, linear/logistic regression, and gradient boosting are analyzed. In addition, model accuracy for scaled/unscaled data is discussed.

Dataset description

In this study, a dataset "winequality-red" is selected that contains twelve various attributes of selected wine such as acidity measures, alcohol, sulfur dioxide content, quality measures, etc. (see Table 1). The dataset contains 1599 entries, and each column has no missing data. While most columns store decimal numbers (float 64), quality columns have integer numbers (int64). Figure 1 describes red wine attributes, stating that the majority of red wine quality is rated between 5 and 6. In this study, the wine quality column is set as the target variable, and other red wine attributes are considered feature variables.

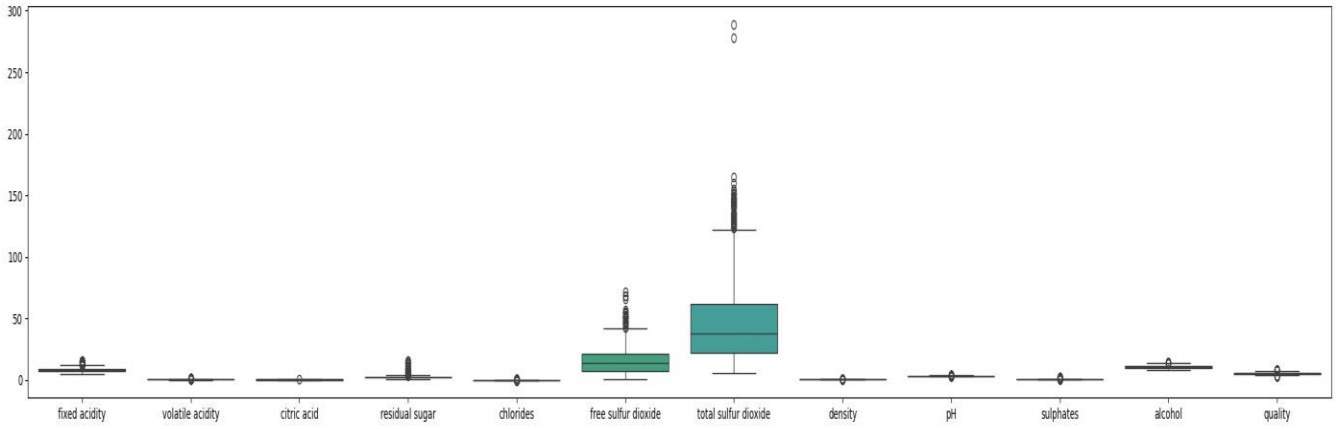

Figure 1 Attributes of selected wine and its rating quality

Table ١ Dataset description

Column	Attribute	Datatype
0	Fixed acidity	float64
1	Volatile acidity	float64
2	Citric acid	float64
3	Residual sugar	float64
4	Chlorides	float64
5	Free sulfur dioxide	float64
6	Total sulfur dioxide	float64
7	Density	float64
8	PH	float64
9	Sulphates	float64
10	Alcohol	float64
11	Quality	int64

Experimental setup

In this section, programming languages and libraries, data processing approaches, and machine learning algorithm selection are explained in detail.

Programming languages and libraries

In this work, python as a programming language is used for data analysis and machine learning. Also, the following libraries are used[1]:

- Pandas: Python data manipulation and analysis.
- NumPy supplies a library of high-level numerical computing that operates on arrays and matrices.
- Matplotlib and Seaborn: used for data visualization to analyze and achieve results
- Scikit-learn: it is a free and open-source machine-learning library for the Python programming language.

Data Processing Approach

The sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators. In general, many learning algorithms benefit from the standardization of the data set.

1- **Standardization**

Standardization of datasets is a common requirement for many machine learning algorithms implemented in scikit-learn that remove mean and scale based on variance.

2- **Splitting data**

The data was split into training (80%) and testing (20%) data.

Model selection

A small number of machine learning algorithms are chosen for the prediction of the quality of the selected wine[2].

- Linear/logistic regression
- random forest
- support vector machine
- gradient boosting
- decision tree

Evaluation: Performance Metrics

some evaluation metrics are used, such as mean square error (MSE), root mean square error (RMSE), R-squared (R^2), and accuracy in evaluating the performance of each model in predicting the quality of selected wine accurately.

MSE is calculated as the average of squares differences between predicted and actual values.

RMSE is calculated by the square root of the mean square error and provides a measure of the average magnitude of the prediction error.

$\mathbf{R^2}$ Provides how well the predictions approximate the actual data.

Accuracy is a way to evaluate the performance of classification.

Results

The various machine learning models, such as linear/logistic regression, SVM, random forest, decision tree, and gradient boosting, are trained with scaled and unscaled data. Some metrics are used, such as mean square error (MSE), square root of mean square error (RMSE), R-square (R^2) and accuracy to evaluate the various models.

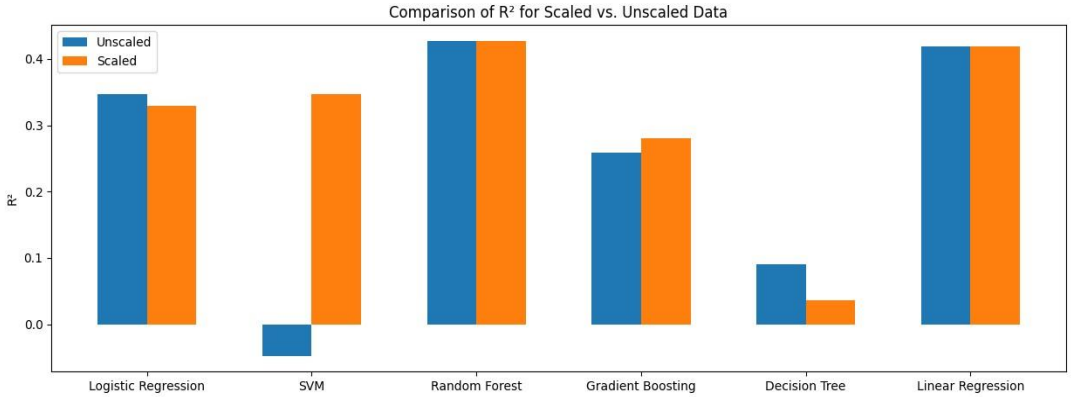

Figure 2 Comparison of R-square of various models for scaled and unscaled data

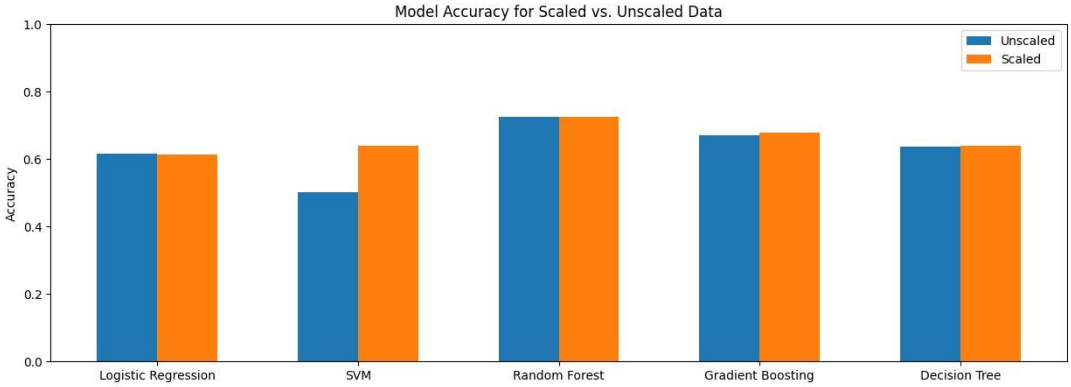

Figure 3 Comparison of accuracy of various models for scaled and unscaled data

Figure 2 displays a comparison of R-square for different ML algorithms on both scaled and unscaled data. As shown, scaling of data does impact the R2 of logistic regression, SVM, gradient boosting and decision tree, while random forest and linear regression do not. According to Figure 2-3, display random forest model has more accuracy compared to other models. As a result, the random forest could display a good performance model among evaluated ML models based on the accuracy of the analysis.

Reference

1. https://scikit-learn.org.
2. Mahesh, B., *Machine learning algorithms-a review*. International Journal of Science and Research (IJSR).[Internet], 2020. **9**(1): p. 381-386.