Practical Machine Learning: ML Algorithm Selection

Milana M. Wolff

May 06, 2024

1 Introduction

In this assignment, we approach the problem of selecting the best-performing machine learning model for a wine classification prediction problem. Wines are rated based on expert assessments; using a number of physicochemical quantitative measures associated with wines, such as density and acidity, we attempt to replicate expert opinions and predict the score assigned to a wine.
In order to solve the problem of determining the best-performing machine learning algorithm to predict expert scoring of red Vinho Verde wines, we test a variety of well-known classification models, starting with default hyperparameter configurations, and evaluate across metrics commonly used for classification problems, such as balanced accuracy and confusion matrices.

2 Dataset Description

The dataset used for this assignment contains physicochemical quantitative input features and sensory quantitative output features (i.e., an expert wine score) for the red variant of the Portuguese "Vinho Verde" wine [1]. The dataset includes 1599 observations and eleven input features, including fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. According to the UC Irvine Machine Learning Repository website, "the classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones)", with a total of 1319 observations rated as 5 or 6 and a mere 28 observations rated with the highest and lowest scores (3 and 8) [2]. This robust dataset includes no missing values to be imputed. We use the eleven listed input features to predict the wine quality measurement.

Classifier	Balanced Accuracy
Ridge	0.2299 ± 0.0271
AdaBoost	0.2649 ± 0.0769
Bagging	0.2582 ± 0.0475
Random Forest	0.2932 ± 0.0394
Logistic Regression	0.2370 ± 0.0243
Perceptron	0.2144 ± 0.0573
Stochastic Gradient Descent	0.2531 ± 0.0554
Multi-Layer Perceptron	0.2453 ± 0.0280

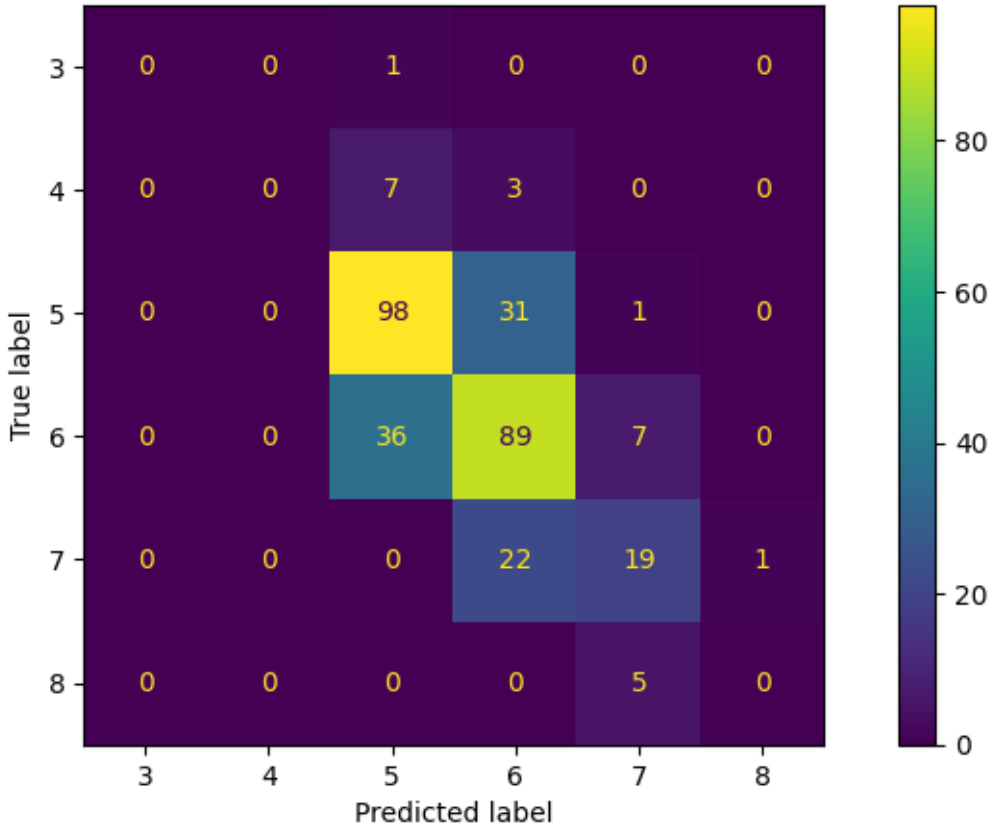
3 Experimental Setup

We use Python 3.10.12 in a Jupyter/interactive Python notebook running on Google Colaboratory. After importing the `scikit-learn` library, we use out-of-the-box implementations of the following classification algorithms with default hyperparameter settings: Ridge classifier, AdaBoost classifier, bagging classifier, random forest classifier, logistic regression classifier, perceptron classifier, stochastic gradient descent classifier, and a multi-layer perceptron classifier.
We load the wine data using the `pandas` library and use all eleven features for classification without additional pre-processing steps. For each machine learning algorithm, using default hyperparameter settings, we perform 10-fold cross-validation with balanced accuracy as the scoring metric. We use the highest test score for each algorithm to determine the best estimator (model with best parameters) for that algorithm. We calculate the average performance as the mean of the test scores obtained in the 10 cross-validation folds, and report the standard deviation of these test scores as well. We report the algorithm with the highest average performance and provide the confusion matrix for the best model of that algorithm type on an 80/20 train/test split of the data.

4 Results

We observed the highest performance using the balanced accuracy metric with the random forest classifier, with an average balanced accuracy score of 0.2932. Results for each model type are tabulated below for comparison.

The confusion matrix below shows the performance of the best random forest model on test data on a random 80/20 train/test split of the dataset.



Code is uploaded to GitHub at the following link: https://github.com/COSC5557/ml-algorithm-selection-mwolff2021-1

References

[1] In: (). URL: `http://www.vinhoverde.pt/en/`.

[2] Paulo Cortez, A. Cerdeira, F. Almeida, et al. "Wine Quality". In: (2009). DOI: https://doi.org/10.24432/C56S3T.