Practical Machine Learning: ML Algorithm Selection

Milana M. Wolff

December 19, 2023

1 Introduction

In this assignment, we approach the problem of selecting the best-performing model 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.

3 Experimental Setup

We use Python 3.10.12 in a Jupyter/interactive Python notebook. After importing the scikit-learn library, we use out-of-the-box implementations of the following classification algorithms: Ridge classifier, AdaBoost classifier", bagging Classifier, random forest classifier, logistic regression classifier, perceptron classifier, and SGD 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, we use a nested resampling approach, choosing the best model parameters over a ten-fold cross-validation using `cross_validate`, and averaging the performances of five of these best models over a randomly selected 80/20 train-test split by storing the test performance of the best model for each of the five. We use balanced accuracy as the primary metric for comparison between algorithms during evaluation.

4 Results

We observed the highest performance using the balanced accuracy metric with the random forest classifier method, at 0.3218 (performance of best parameters averaged across five trials) and an overall accuracy of 0.65 (a randomly selected trial). See below for a comparison of the confusion matrix and other performance estimators for this approach to a less performant method. Note that the metrics reported are from a random trial of five and the confusion matrix graph represents the best trial of five.

```
RandomForest Classifier Performance:
Accuracy: 0.65
0.3225718725718726
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        10
           5       0.72      0.74      0.73       130
           6       0.62      0.70      0.65       132
           7       0.58      0.50      0.54        42
           8       0.00      0.00      0.00         5

    accuracy                           0.65       320
   macro avg       0.32      0.32      0.32       320
weighted avg       0.62      0.65      0.64       320

[[ 0  0  1  0  0  0]
 [ 0  0  7  3  0  0]
 [ 0  0 96 34  0  0]
 [ 0  0 30 92 10  0]
 [ 0  0  0 20 21  1]
 [ 0  0  0  0  5  0]]
```

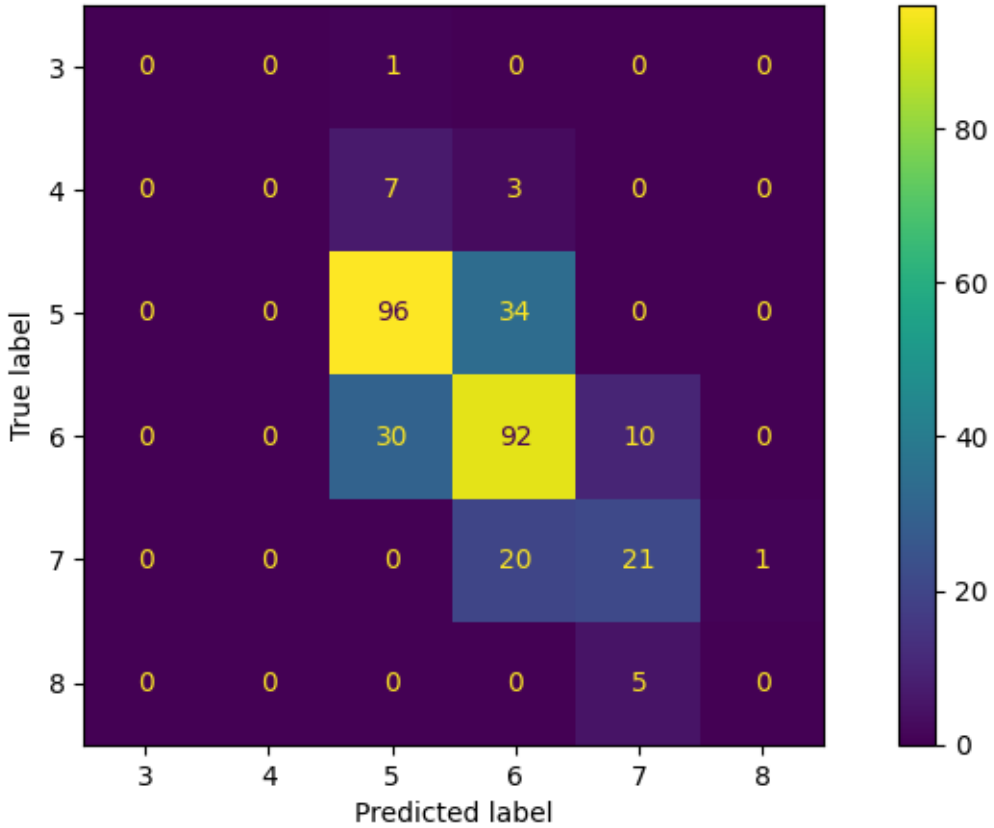
```
SGD Classifier Performance:
Accuracy: 0.43
0.21944166944166943
              precision    recall  f1-score   support

           3       0.00      0.00      0.00         1
           4       0.00      0.00      0.00        10
           5       0.60      0.18      0.28       130
           6       0.42      0.73      0.53       132
           7       0.34      0.40      0.37        42
           8       0.00      0.00      0.00         5

    accuracy                           0.43       320
   macro avg       0.23      0.22      0.20       320
weighted avg       0.46      0.43      0.38       320

[[  0   0   0   1   0   0]
 [  0   0   4   6   0   0]
 [  0   0  24 101   5   0]
 [  0   1  12  96  23   0]
 [  0   0   0  25  17   0]
 [  0   0   0   0   5   0]]
```

	Accuracy	Balanced Accuracy
Ridge	0.57	0.2318
AdaBoost	0.53	0.2309
Bagging	0.58	0.3021
Random Forest	0.65	0.3226
Logistic Regression	0.55	0.2253
Perceptron	0.42	0.2524



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

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.