Charles Myers

Introduction

For this assignment I will be testing different parameters and auto-searches to determine which combination is the best for a given task. I will be using the White Wine dataset.

Dataset Description

The white wine dataset has 11 features and 1 target variable. The features are different aspects of wine such as its acidity, sugar, sulfur dioxide, and alcohol. The target variable is the quality of the wine. I chose this dataset for two reasons. The first is that I enjoy white wine and the second is because it has a decent amount of data inside of it.

Experimental Setup

I will be using scikit-learn machine learning libraries, JupyterLab as an IDE, and other libraries such as seaborn and panda for data and graphs. The means of comparing the different algorithms will be the coefficient of determination, and best_score_ function sklearn has.I chose the coefficient of determination when running the algorithm without any parameters being changed at all.

The different Machine Learning Algorithms I will be using are random forest, and support vector machine. I chose these mostly because they had a decent number of parameters that could be changed. The two sklearn auto-searches I used are GridSearchCV and HalvingGridSearchCV. I chose GridSearchCV because it seemed to be the most basic of exhaustive searches, and I chose HalvingGridSearchCV because of its experimental and I thought it sounded interesting. I will be using several different parameters for SVM and RF.

The SVM parameters are:

The kernel between linear, poly, rbf, and sigmoid

The degree with 3,6,9

Gamma to be either scale or auto

C to be either 1.0 or 1.2

The cache_size to be 1000

The RF parameters are

The number of n_estimators between 10 and 100 increasing by 10 each time.

The max_features to be None or sqrt

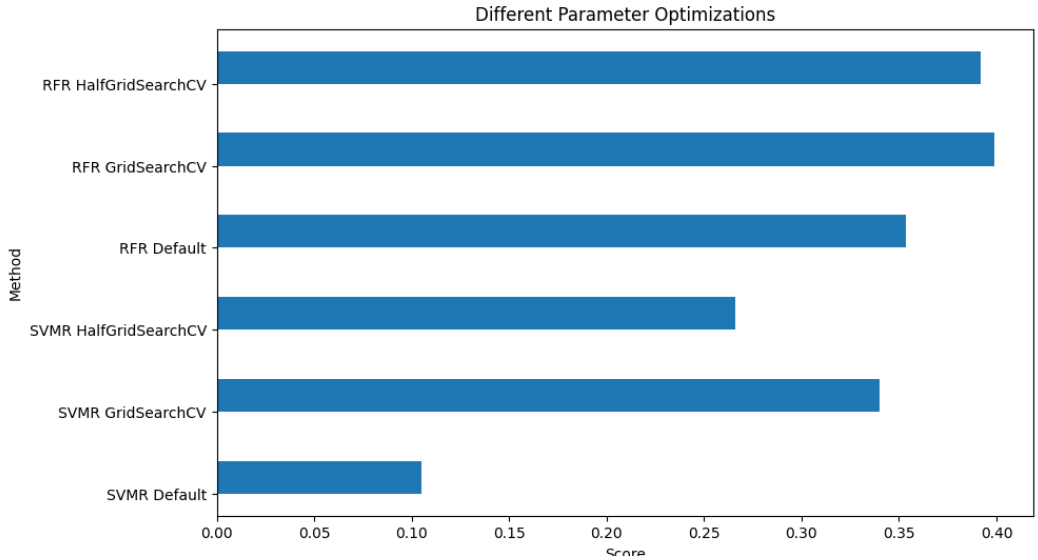
The max_depth to be None, 2 or 4

The main_samples_split to be 2

The bootstrap between True or False.

Results

RF Default
Coefficient of determination of RF: 0.35
RF GridSearchCV
Best Score: 0.3987673633042612
Best Parameters: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 80}
RF HalfGridSearchCV
Best Score: 0.39193583325657594
Best Parameters: {'bootstrap': False, 'max_depth': None, 'max_features': 'sqrt', 'min_samples_split': 2, 'n_estimators': 100}
SVM Default
Coefficient of determination of SVM: 0.10
RF GridSearchCV
Best Score: 0.34004873357840165
Best Parameters: {'svr__C': 0.8, 'svr__cache_size': 1000, 'svr__degree': 3, 'svr__gamma': 'scale', 'svr__kernel': 'rbf'}
RF HalfGridSearchCV
Best Score: 0.26576443841011305
Best Parameters: {'svr__C': 1.2, 'svr__cache_size': 1000, 'svr__degree': 3, 'svr__gamma': 'auto', 'svr__kernel': 'rbf'}



Looking at the results it seems that the random forest using the GridSearchCV gives us the best score out of the bunch. It also seems that all the Support Vector Machines perform worse than any of the random forest. The reason why is probably because the svr data is being standardized using a pipe and MinMaxScaler. The reason why the data is being preprocessed is because SVM would run painfully slow otherwise.