ST3189 Machine learning (2023-24)

Coursework

Student Number: 220656407

Table of contents:

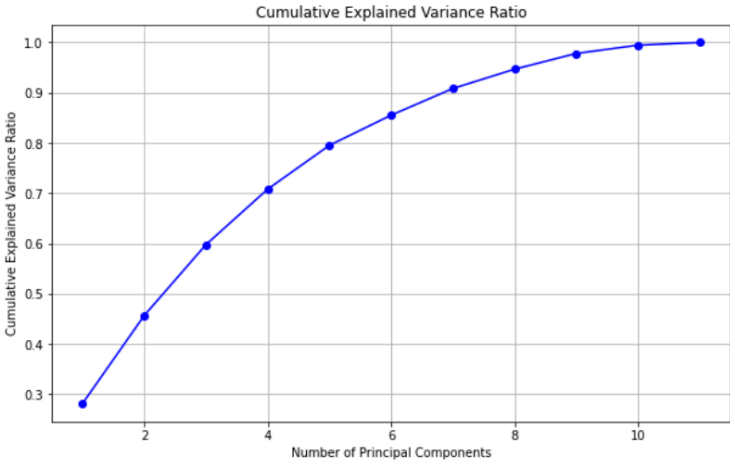
Main description of dataset and tasks structure decisions:

I took only one dataset through which I conducted all the tasks. Classification problem is done with the target variable - “quality”, representing quality of wine with integer values from 3 to 8. For Regression I chose a continuous variable as a target one from the same dataset - “alcohol”, presenting the strength of wine. According to the Unsupervised learning task - it is applied twice for different datasets that I have changed from the initial one for different target variables. So, regression and classification tasks are done independently from each other both having unsupervised learning methods application.

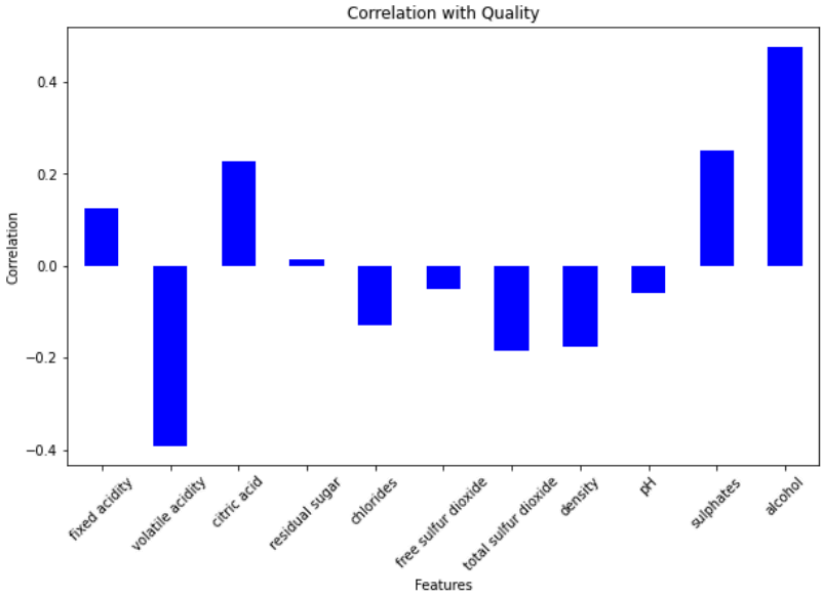
The chosen dataset was “Wine Quality” having 11 features it is possible to apply many methods using it, the dataset is signed as belonging to the business area. According to the business area it might be useful for wineries to get some of the informative predictions for further work in terms of wine quality. Probably the wine quality is determined by the market, so, it might be helpful to be able to make predictions for wineries by themself before the mass production of wine, thus, having less expenses as in the most of the real life cases quality is determined in comparison by the sommelier, having more subjective views. While the percent of alcohol prediction in this case is important too to make the process of wine production more automatic, as from all of the other examples using the machine learning methods people would be able to predict the strength of wine based on the initial components behavior.

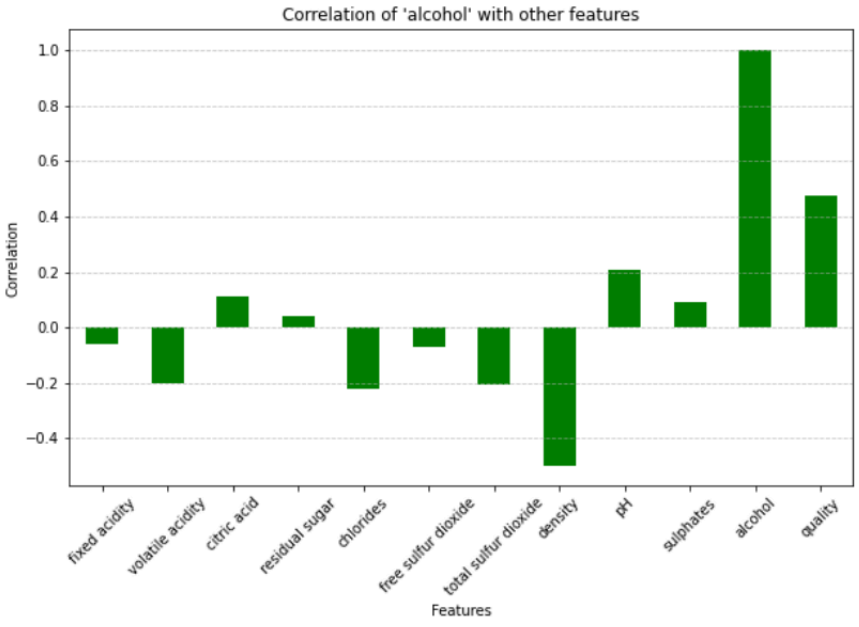
This dataset is valuable for its simplicity for understanding together with data which is easy to use as it does not have any of the missing values which is important for more precise prediction. Also, the chosen dataset initially consists of 1599 instances, it is not a high number in terms of machine learning, so it is more convenient for code with lower time taken for compilation.

To prepare the data for predictions in most cases it is vital to make the **preliminary analysis** and some changes. In this part described changes and analysis do not belong to any of the given tasks, however, the actions might not be skipped. Firstly, it is vital to look at Principal Component Analysis performance and find the best parameters. It is helpful to visualize the Cumulative Explained Variance Ratio with respect to the number of components in PCA, by which we could see which number we should take in an algorithm where 0.85-0.95 of variance should be explained.

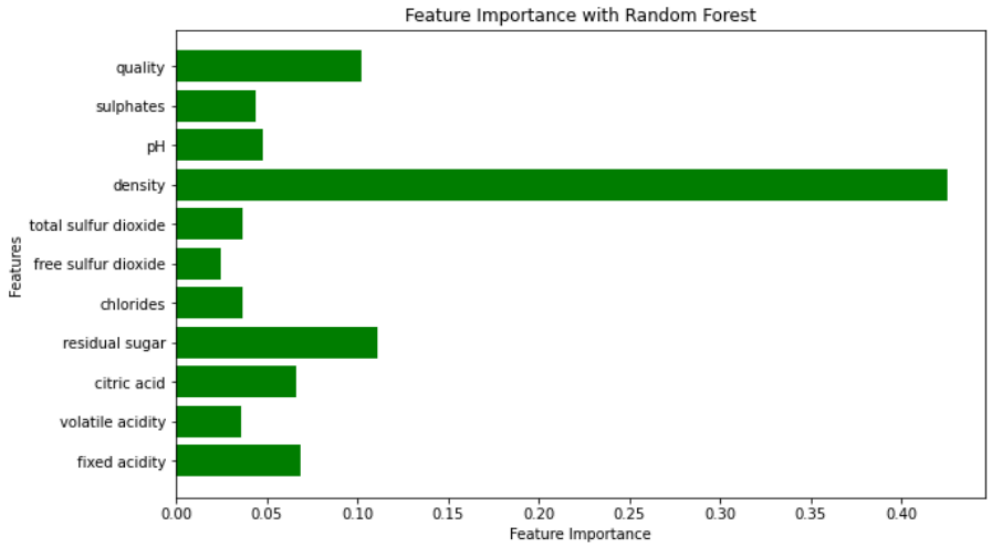


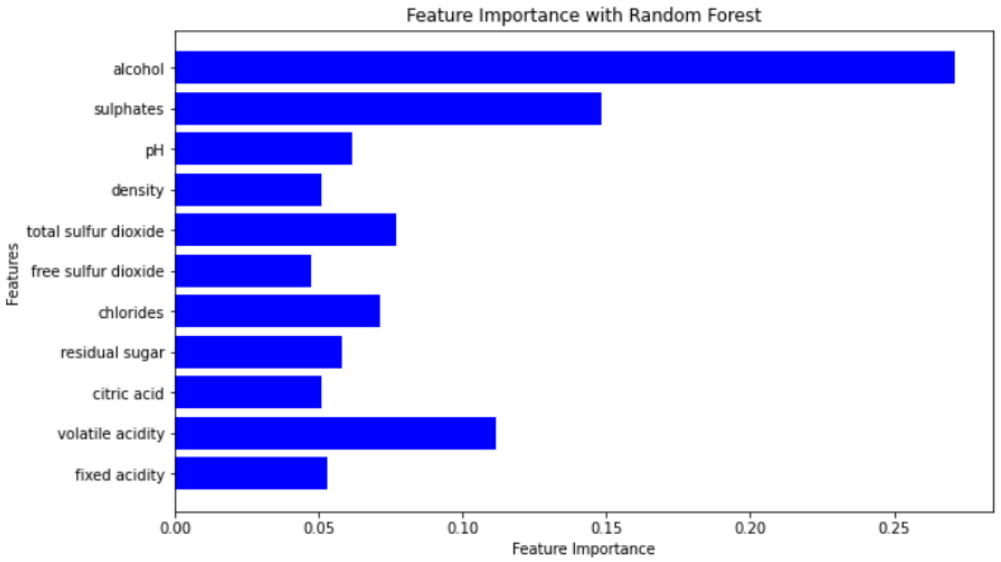
Secondly, to find out about the data it is useful to look at the correlation between the target variable and other variables together for better understanding of features that the target depends on. Here presented correlation for “quality” and “alcohol”:





After seeing visualized methods it is useful to apply the relevant algorithms, as feature importance to find ones having highest impact on the target variable and addition of outliers filtering with help of interquartile range (IQR) which leaves only values that are stated between 25% and 75% of all data values. All of these methods are applied in terms of getting more precise predictions. These methods were used for both of the tasks and conducted before unsupervised learning methods.

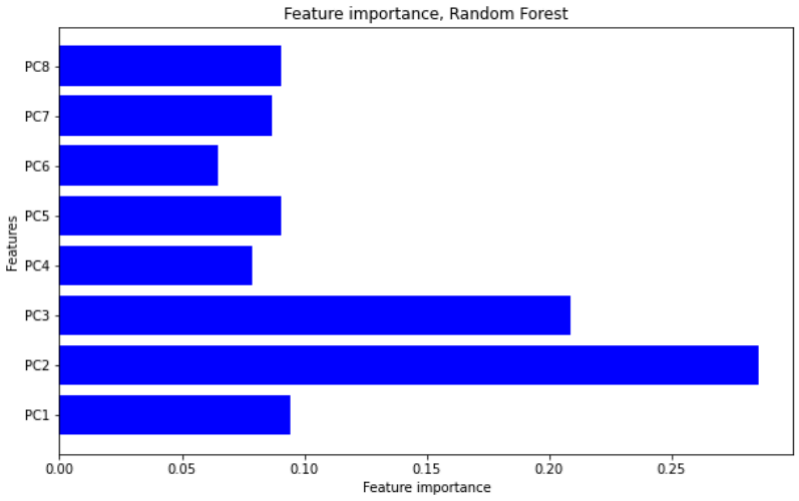


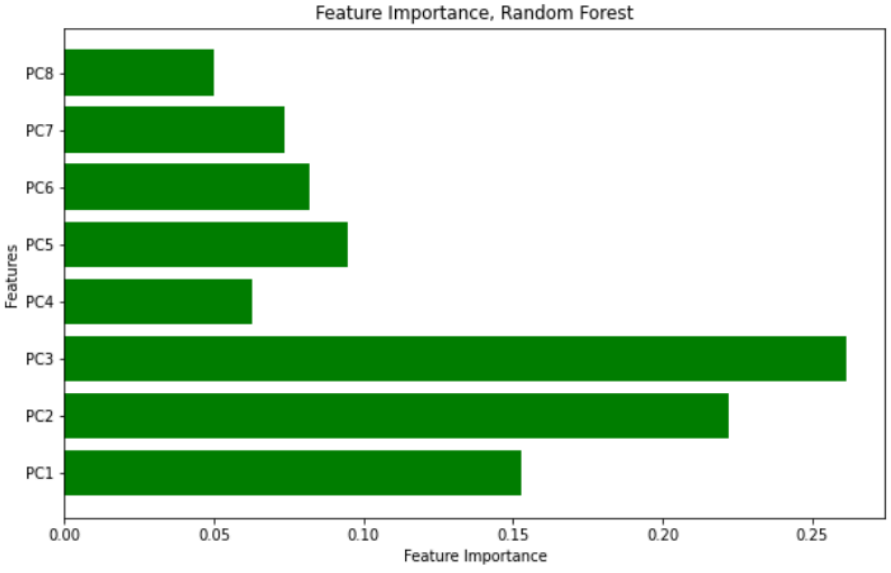


Unsupervised learning:

Mostly this method is about lowering the dimensionality of the data, here the chosen algorithm is Principal Component Analysis. Firstly, it allows us to find the new main components that mostly influence the target variable in which data is mostly volatile, creating them as new axes. Then, the algorithm automatically reduces dimensionality by applying it to the new axes, helping to reduce the number of

features while keeping the highest amount of information. One of the major principles PCA uses - is data centering by subtraction of mean value of each feature from itself. Applying to my data, PCA is useful for a couple of reasons: easier interpretation of results; improvement of data visualization; improvement of the machine learning models performance. As PCA takes the main components it makes it easier to focus on the most valuable features for prediction. As the dimensionality is reduced with PCA, it becomes easier to visualize the data on the graph. PCA also affects the noise in data, in most cases making the models work faster and with more precise predictions. Also, the standard scaler method was applied for normalization and standardization of values in the dataset. After the PCA standard scaler techniques conduction the results in changes of the data were further used in models as new datasets.





On the pictures above the feature importance visualization is presented after PCA conducting, green bars are for “alcohol” data, blue - “quality” data.

Regression:

In the chosen dataset the target variable for regression techniques is “alcohol” with the research question of predicting its levels (how strong the wine is) based on the other features. The work with models that would be presented further is done after the previously mentioned methods. The model evaluation here is done with Mean Squared Error (MSE) and R^2 score metrics, for main rules, R^2 metric has a range of values between 0 and 1. To analyze the work of the model it is important to understand that it is better if the MSE metric score is the least (0.1 better than 0.9) and the opposite for R^2 where best results are the highest ones. MSE shows the average squared difference between the real values and the predicted ones, while R^2 shows the proportion of the variance in the target variable that is explained by other features in the model. All the models use various parameters and have different working principles, the simplest model is Linear Regression which assumes that there is a linear relationship between the features and target variable, together with the Ridge model. The other models that were used are: DecisionTreeRegressor, RandomForestRegressor, GradientBoostingRegressor, KNeighborsRegressor and Support Vector Regression(SVR). These models have the following results displayed on figure 2.1

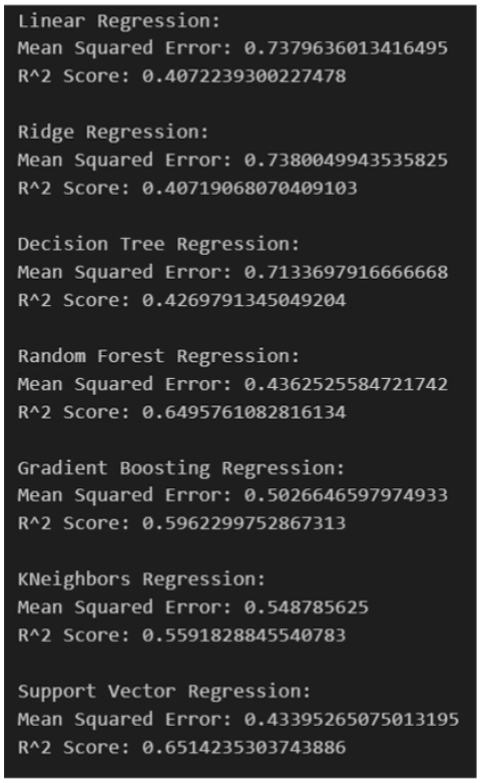


Figure 2.1

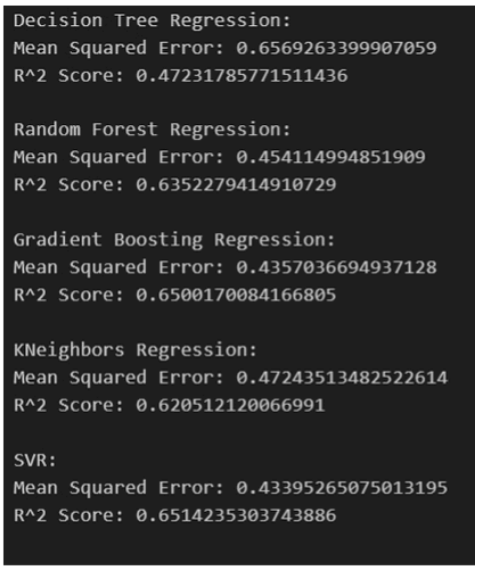
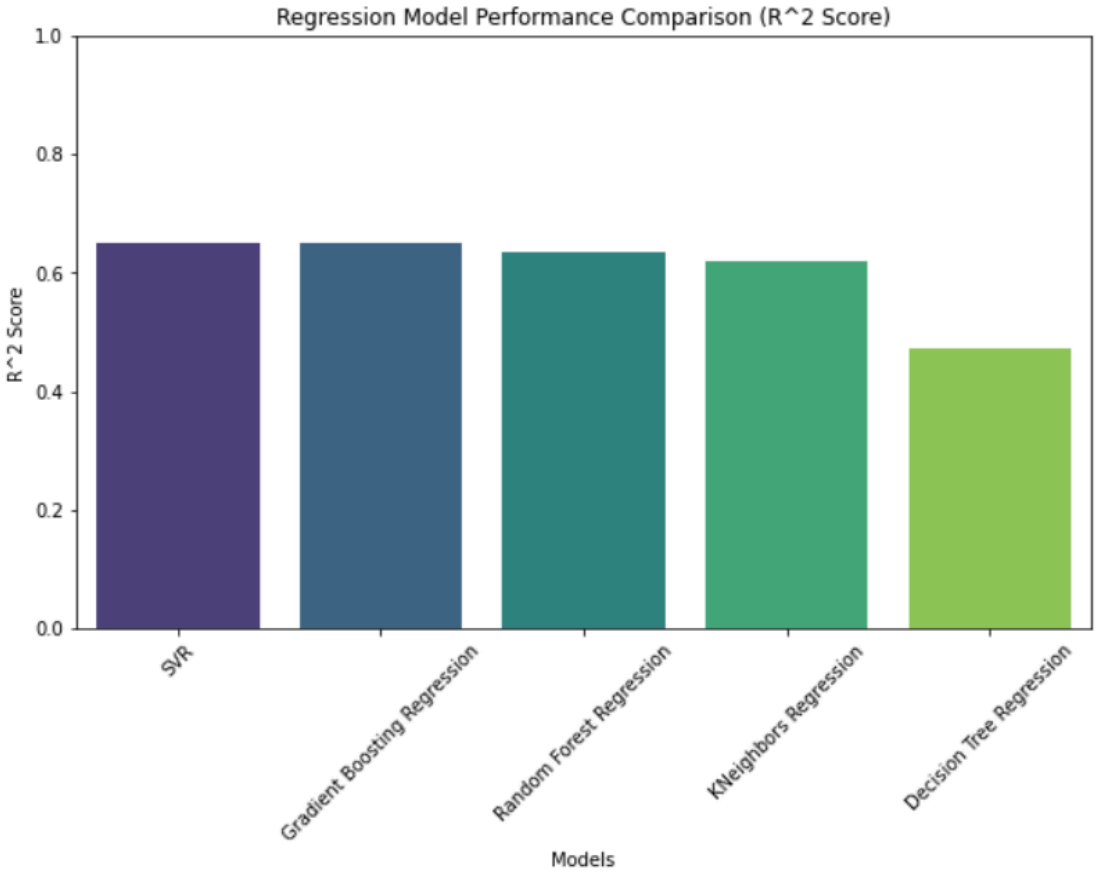


Figure 2.2

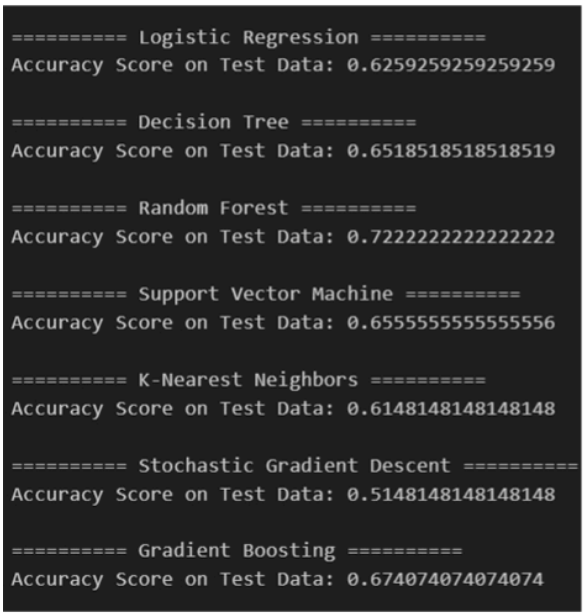
After the performance of models with no parameters tuned, I used the algorithm “GridSearchCV” for tuning of the hyperparameters which is supposed to find better

settings for the model looking through various combinations of hyperparameters and got these scores displayed on figure 2.2.
Not all the models were tuned due to its possible settings, the Linear Regression model might be barely tuned, keeping nearly the same result, so as the Ridge model, thus, there was no need to try improving their performance.
According to the appearing scores it could be seen that the Support Vector Regression (SVR) model has the best performance where 65.1% of the variance of target variable are explained by the model, and the Gradient Boosting Regression (GBR) with 65% explained. Thus, here the most suitable model should be the SVR together with GBR. On the other hand, Decision tree regressor became the worst performing model with only 47% of the explained variance of the target variable.
Mostly, the Grid Search algorithm performed well in this case, on the other hand, as it could be seen, in some cases, as with the Random Forest Classifier model, the score got slightly lower, which means that initial parameters were the best for it.
Although many models would show different results, here for predicting the percent of alcohol in wine by the other features, the Support Vector Regression model gave the best result.



Classification:

For the classification problem I use the same dataset with different target variable - “quality”. From the data preprocessing and analysis it could be seen that this variable is a categorical one, which may take only 6 available values all of them are whole numbers [3, 4, 5, 6, 7, 8] where 8 is the best quality and 3 is the worst wine. The main research question is to predict the wine quality based on the other features. All the models that would be seen further are done after unsupervised learning and data preprocessing methods. The model performance is evaluated here with an accuracy score metric which is calculated as percent of correctly classified values divided by the whole number of values. Accuracy score is calculated in percentages where the best score is 1 which is equivalent to 100%. The models I have tried for this task are using various parameters and have different abilities and difficulty of working. The models of my choice are presented here with the following accuracy scores displayed on figure 3.1.



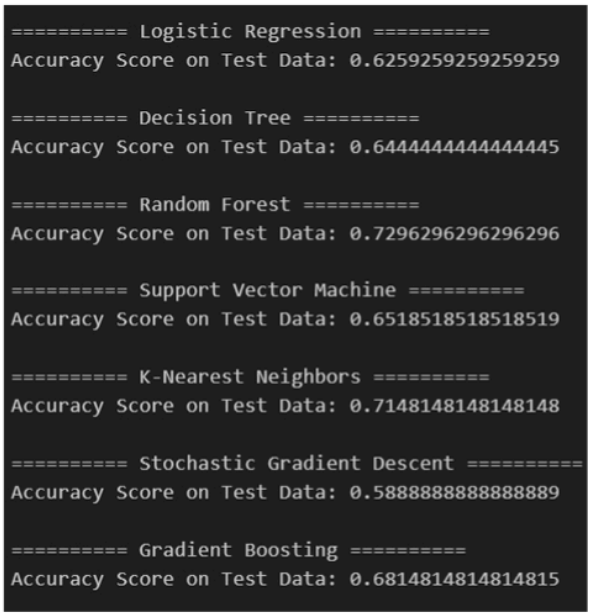
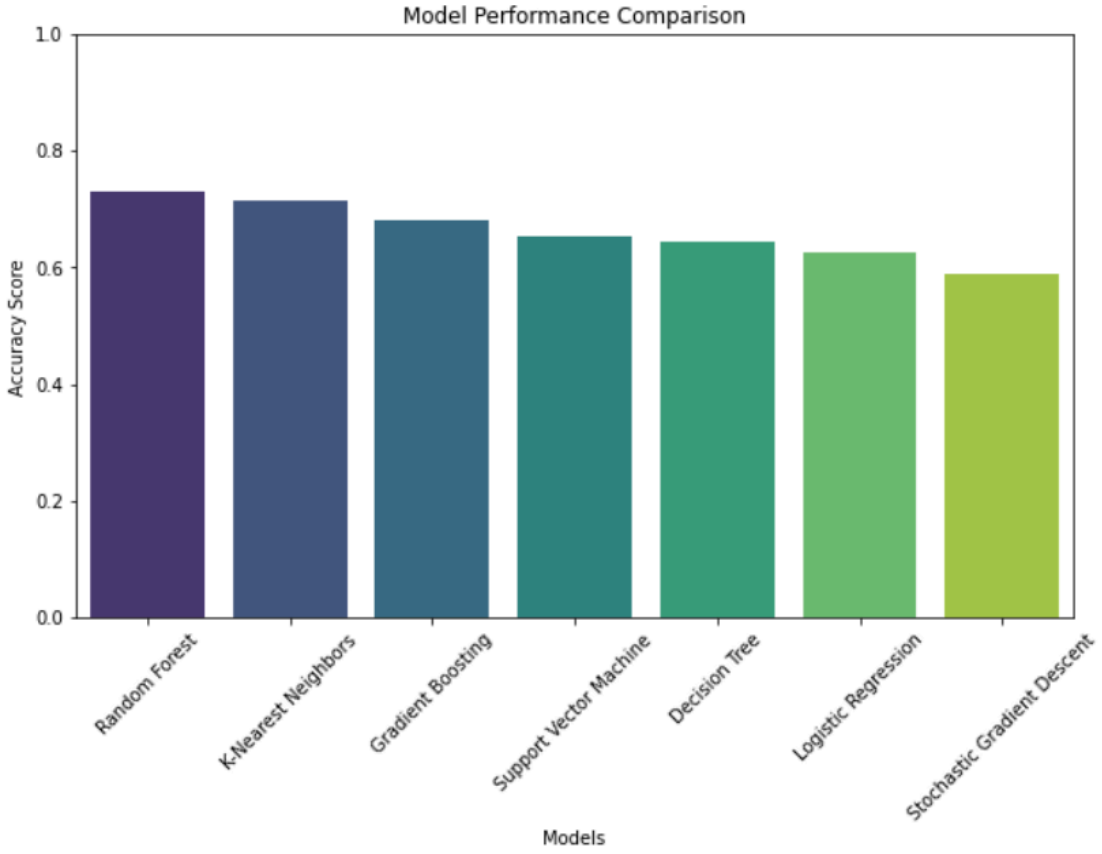


Figure 3.1 Figure 3.2

However, I decided to use the “GridSearchCV” to tune the hyperparameters of the models which should allow for better performance, and got the following results displayed on figure 3.2.
So, here the models with best performance are the “Random Forest Classifier”(RFC) model with 72.9% of the correctly predicted values, together with the K-Nearest Neighbors which has 71.4% of the correctly predicted values. While the “Stochastic Gradient Descent” (SGD) model had the worst performance with 58%. After tuning the parameters , some of the models’ scores left the same, it might be caused by the fact that initial parameters were best for it. So, I could say that the SGD and Logistic Regression models might not be the best choice for this data. Thus, in this case of

predicting the quality of wine based on the other feature, the Random Forest Classifier model gives the best prediction.



Summary:

At all, through the analysis and models implementation, the research questions were fully understood by me and I got the answers to them as models I have made give predictions of wine quality and wine strength for classification and regression tasks, respectively. Furthermore, in some of the specified studies in the business area some of this answer might be really helpful.

References:

According to the research area, to understand the assigned questions better I have looked through this articles:

1. State of the world wine - which explains the way wine is produced and which factors in real life affect its quality, also the article presents the data of the wine grown by the countries which might be useful knowledge for further studies.

2. Also, the following source was useful for understanding of the way wine is classified and the way its quality determined in the real world examples

According to the code and methods applications I used the sources:

1. Linear models (Logistic Regression, Ridge Regression, Lasso)
2. Classification models : Decision Tree Classifier, Random Forest Classifier, K-Nearest Neighbours Classifier, Stochastic Gradient Descent Classifier, Gradient Boosting Classifier, Support Vector Classifier.
3. Regression models : Decision Tree Regressor, K-Nearest Neighbors Regressor, Gradient Boosting Regressor.
4. Methods and models for analysis and unsupervised learning : PCA, Standard Scaler, Grid Search CV.