Supervised Learning
Project 1

T-409-TSAM,Computer communication, 2022-1

Reykjavik University - School of Computer Science, Menntavegi 1, IS-101 Reykjavík, Iceland

Ágúst Þór Þrastarsson
agustt20@ru.is
Ragnar Smári Ómarsson
ragnaro20@ru.is

26. September 2022

1 Introduction

Wine is a popular alcoholic drink that is typically made from fermented grapes. Wine has been produced for thousands of years.[4] The dataset used in this project is based on wine quality[3], it comes with two different data files, one with information about red wine and the other one with information about white wine. In this report we will only be using the white wine dataset. We will use classification to predict white wine quality based on chemicals measured in physicochemical test. We will use three classes, poor, good and excellent which are based on wine quality score between 0 and 10 that are in the dataset. The report will go into detail with machine learning models, and how they are predicting the quality of the wine based on some input.

2 Process

Typically when looking at a data set the first step is to do an exploratory analysis, which we explore the data, data types, mean of variables, etc. Then proceed on data pre-processing, where we process or change the data if we see a need to. When that is done, looking at what models we want to use, it is typically trained a few different models and looking at the accuracy for each one and then choose the one with the highest accuracy.

2.1 Exploratory Analysis

In Machine Learning it is important to explore the data set a little bit before going straight into data pre-processing as we want to know if we actually need to do some pre-processing on the data, why we would do it and prevent data misunderstanding.
First we will look at the variables or attributes of the data set.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4898 entries, 0 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         4898 non-null   float64
 1   volatile acidity      4898 non-null   float64
 2   citric acid           4898 non-null   float64
 3   residual sugar        4898 non-null   float64
 4   chlorides             4898 non-null   float64
 5   free sulfur dioxide   4898 non-null   float64
 6   total sulfur dioxide  4898 non-null   float64
 7   density               4898 non-null   float64
 8   pH                    4898 non-null   float64
 9   sulphates             4898 non-null   float64
 10  alcohol               4898 non-null   float64
 11  quality               4898 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 459.3 KB
```

From the image we can see that we have a total of 12 attributes and 4898 rows of data. Note that the type of each variable is also seen in the picture as well as if there are any null values in the data set. Which is very important to notice as having a null attribute can have bad influence on the predictions.
To understand the data further we looked duplicates and if there were any missing values in some rows of the data. It is important to look at those as they will have direct influence on the data. Although duplicates in a data set can sometimes be very useful, in this data set we don't want any duplicates as when training the model it is important to have distinctive attribute values because it brings more consistency within the data.
After deleting all duplicates from the data set we can see that the total number of entries are now at 3961 instead of 4898.

```
Int64Index: 3961 entries, 0 to 4897
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         3961 non-null   float64
 1   volatile acidity      3961 non-null   float64
 2   citric acid           3961 non-null   float64
 3   residual sugar        3961 non-null   float64
 4   chlorides             3961 non-null   float64
 5   free sulfur dioxide   3961 non-null   float64
 6   total sulfur dioxide  3961 non-null   float64
...
 10  alcohol               3961 non-null   float64
 11  quality               3961 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 402.3 KB
```

2.1.1 Outliers

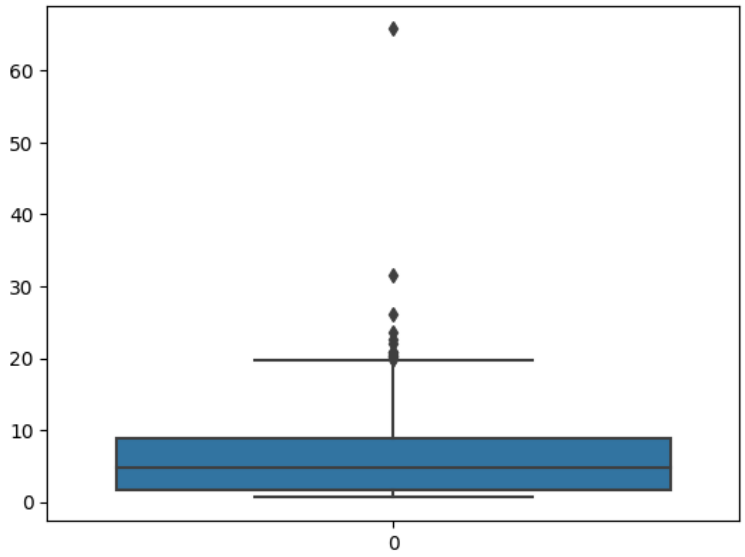
It is important to look at outliers in the data set before continuing as they have direct influence on the data set. Outliers are values that differs from normality and can cause inconsistency in the results from training the data with algorithms and analytical systems, as so they always need some attention. To detect the outliers within data it is good to create a box-plot which can visually represent what values are causing the outlier. To see which attribute has some kind of worth looking at, as we don't want to look at each and every attribute if we don't have to, is to get the sense of where most the points within each attribute is.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol
count	3620	3620	3620	3620	3620	3620	3620	3620	3620	3620	3620
mean	6.83	0.27	0.33	5.90	0.04	34.38	136.23	0.99	3.20	0.49	10.62
std	0.81	0.09	0.10	4.70	0.01	15.48	41.63	0.00	0.14	0.11	1.21
min	4.4	0.08	0	0.6	0.012	2	21	0.99	2.79	0.22	8.4
25%	6.3	0.21	0.27	1.6	0.035	23	106	0.99	3.1	0.41	9.6
50%	6.8	0.26	0.32	4.8	0.042	33	132	0.99	3.19	0.47	10.5
75%	7.3	0.32	0.38	8.825	0.049	45	165	1.00	3.29	0.55	11.4
max	9.4	0.59	0.7	20.4	0.115	86	260	1.00	3.65	0.83	14.2

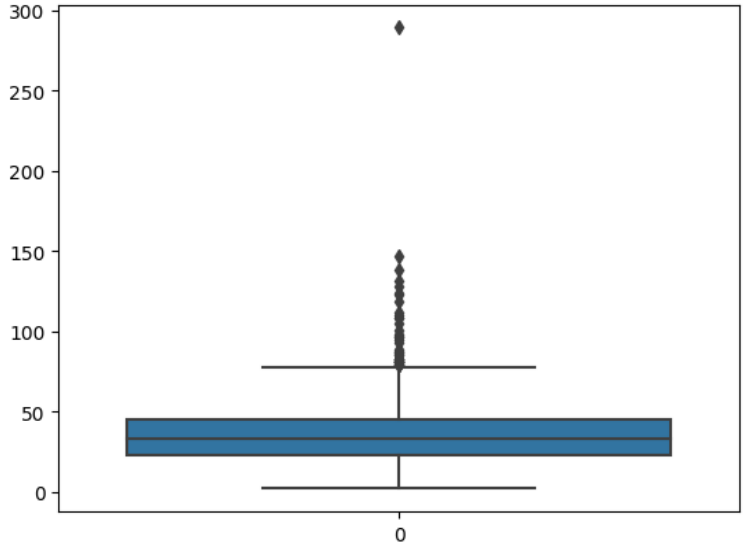
From this picture we can see what min, 25%, 50%, 75% and max values are within each

attribute. An attribute has a high probability of having and outlier is when the difference between the values 75% and max is high. Attributes that this rule applies are as following:

- Residual sugar



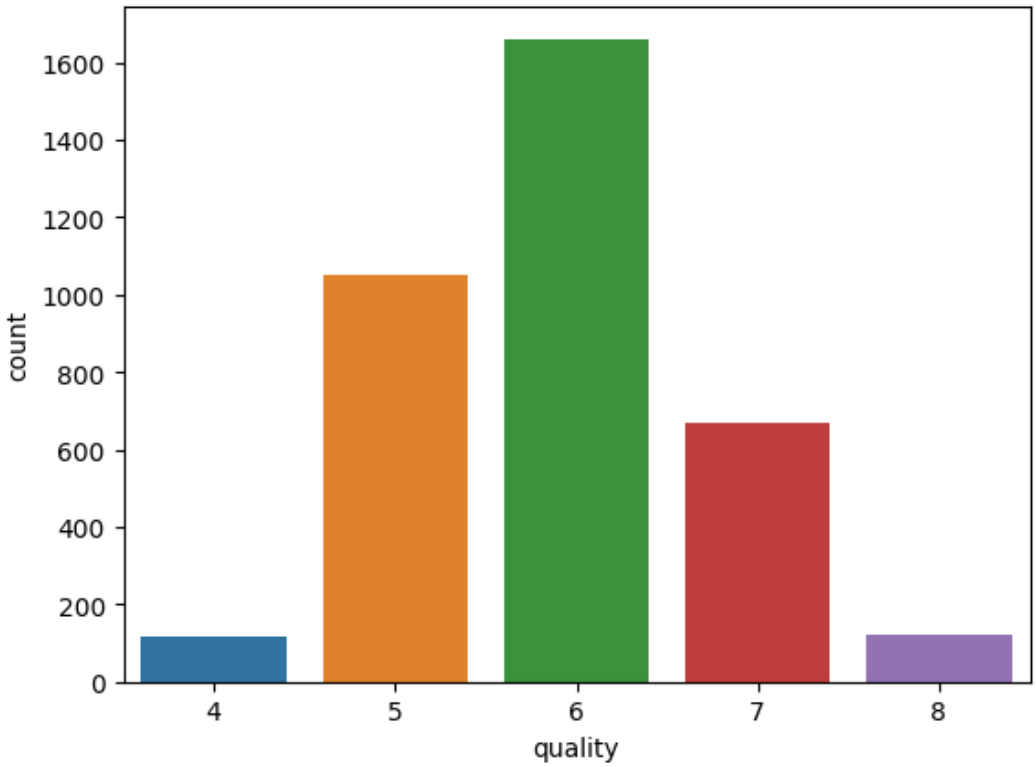
- Free sulfur dioxide



As so we removed the outliers within the data set with the help of the z-score. [5]

2.1.2 Class Labeling

In the dataset all of the attributes are of a numerical value (float) but one, quality, which has an integer value ranging from 1 - 10. Where the numbers represent the quality of the wine. In the data set there is no wine with the quality value of 10 but we are still going to include that as that might happen. Next we want to see the quality distribution within the data set to get a good sense of the variance.



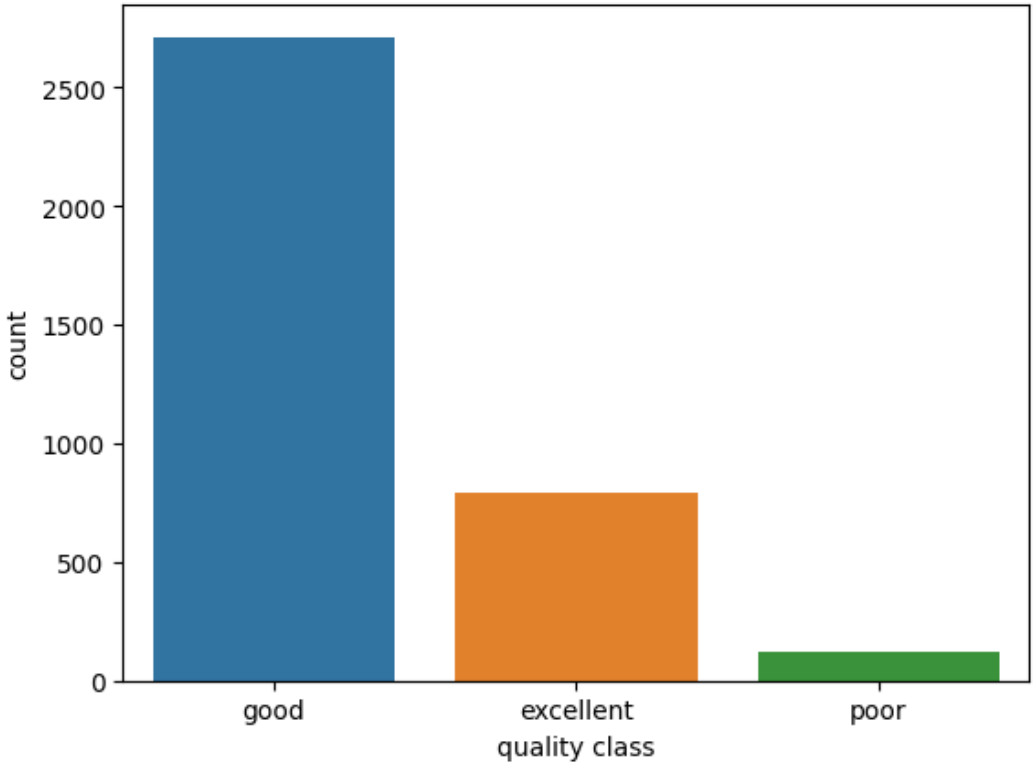
In the picture above we can see that most wines are within quality of 6 and 5.

Since we are using classification we are going minimize the classification by creating a new attribute called quality class. The quality class attribute is going to represent the ranking of wine quality in 3 categories:

- 0 - 4 = 'Poor'
- 5 - 6 = 'Good'
- 7 - 10 = 'Excellent'

2.1.3 Class Imbalance

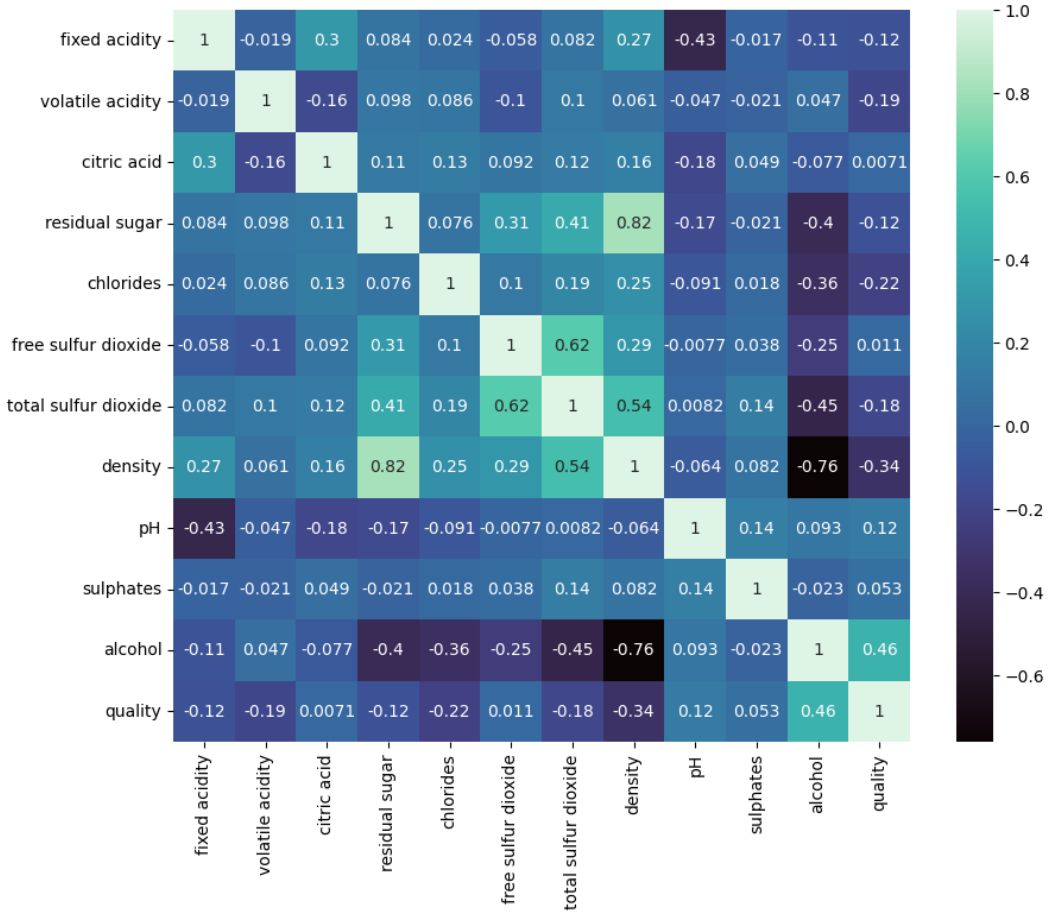
To get check if we have a class imbalance after classifying the quality to 3 classes, we create a plot to visualize that better.



After creating a new attribute and classifying the old attribute 'quality' we can see that there is a major class imbalance within the data set.

2.1.4 Correlation

To see which feature is impacting the attribute quality the most, we have to see the relationships between them. The best way to do that is with correlation. Correlation explains how one or more variables are related to each other. It gives us an idea about the degree of the relationship of the two variables.



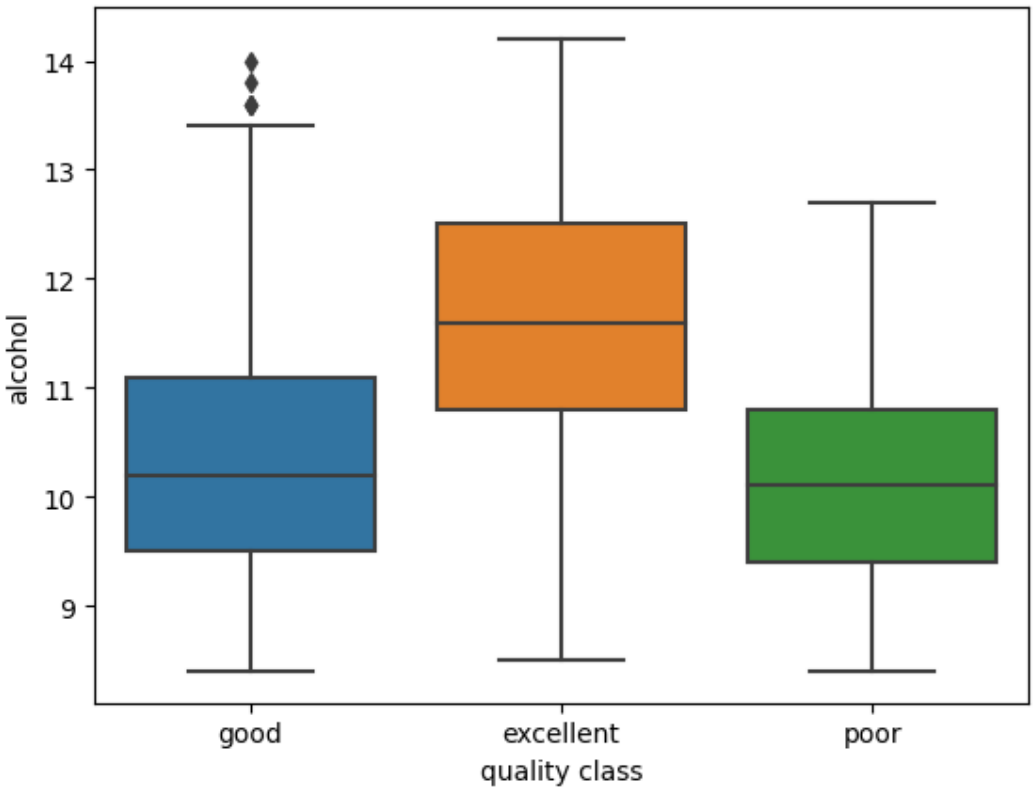
From this picture we want to especially look at the attribute quality and see which attribute correlates most with that, no matter if it is negative correlation or positive correlation.

Alcohol has the most correlation with quality The attributes which have almost no correlation with the quality attribute:

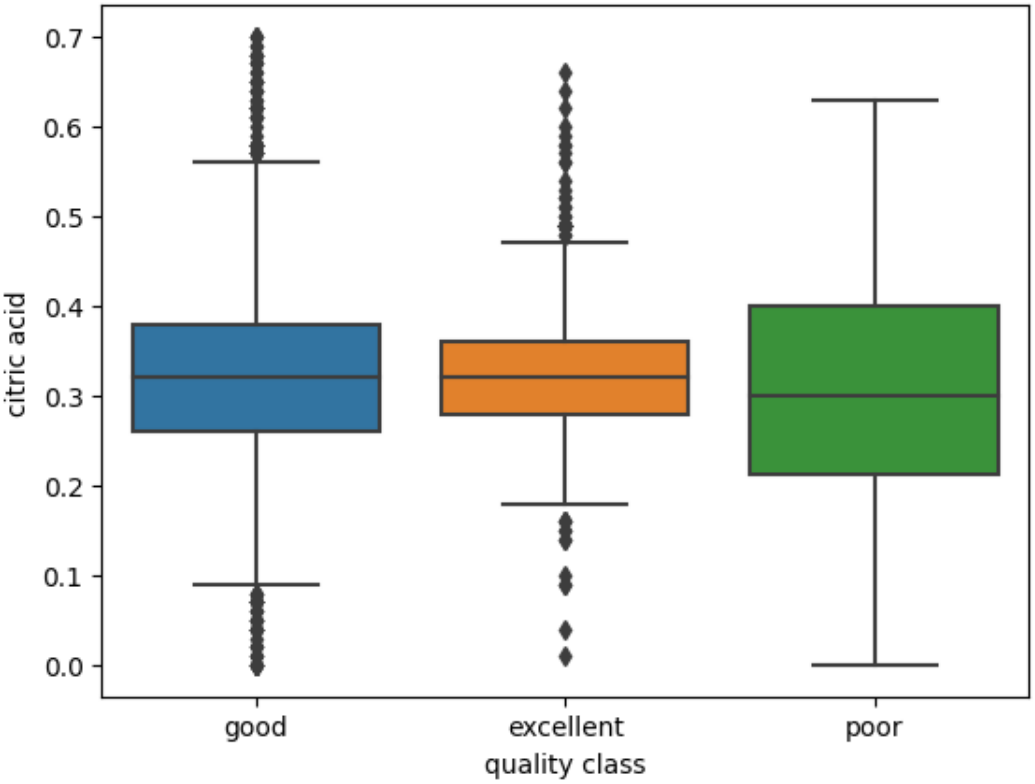
- Citric Acid
- Fixed Acidity
- Sulphates
- Free sulfur dioxide

Because of this analysis we can look more in depth of the correlation between the new attribute we just created, class level, and the attributes which have almost no correlation with the old quality attribute.

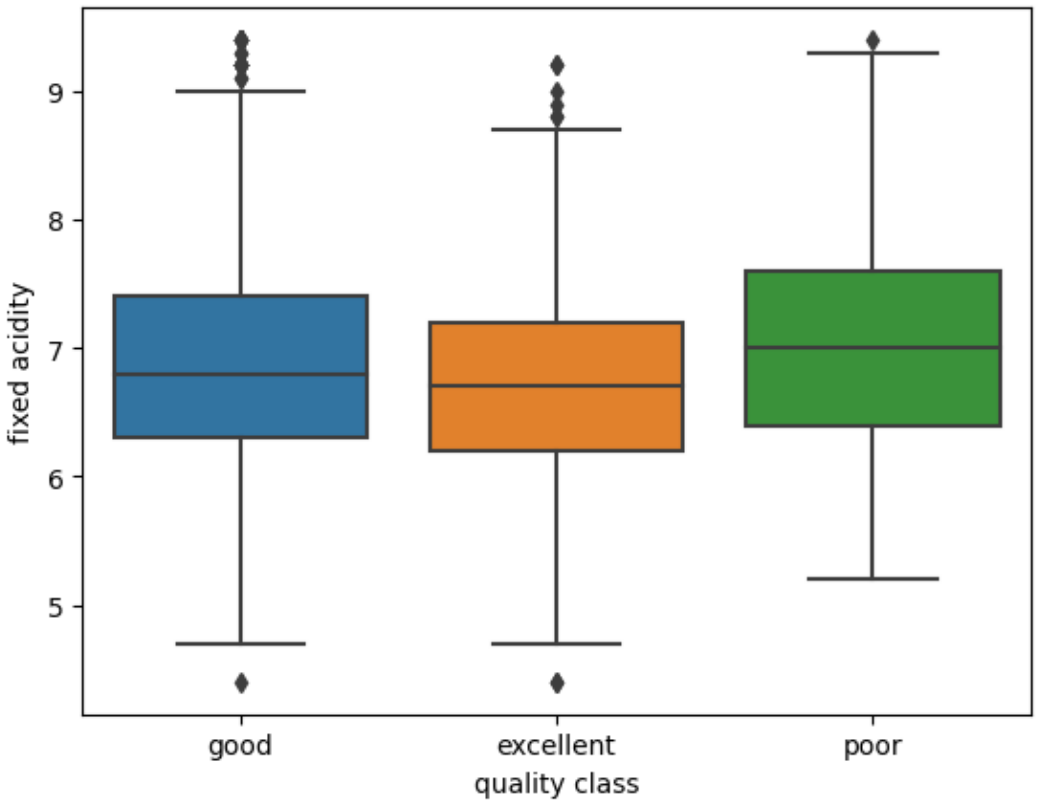
After seeing the correlation plot and creating the new attribute 'quality class' we can now also finally drop the column 'quality' so it wont affect our training and tests.
Correlation between quality class and alcohol.



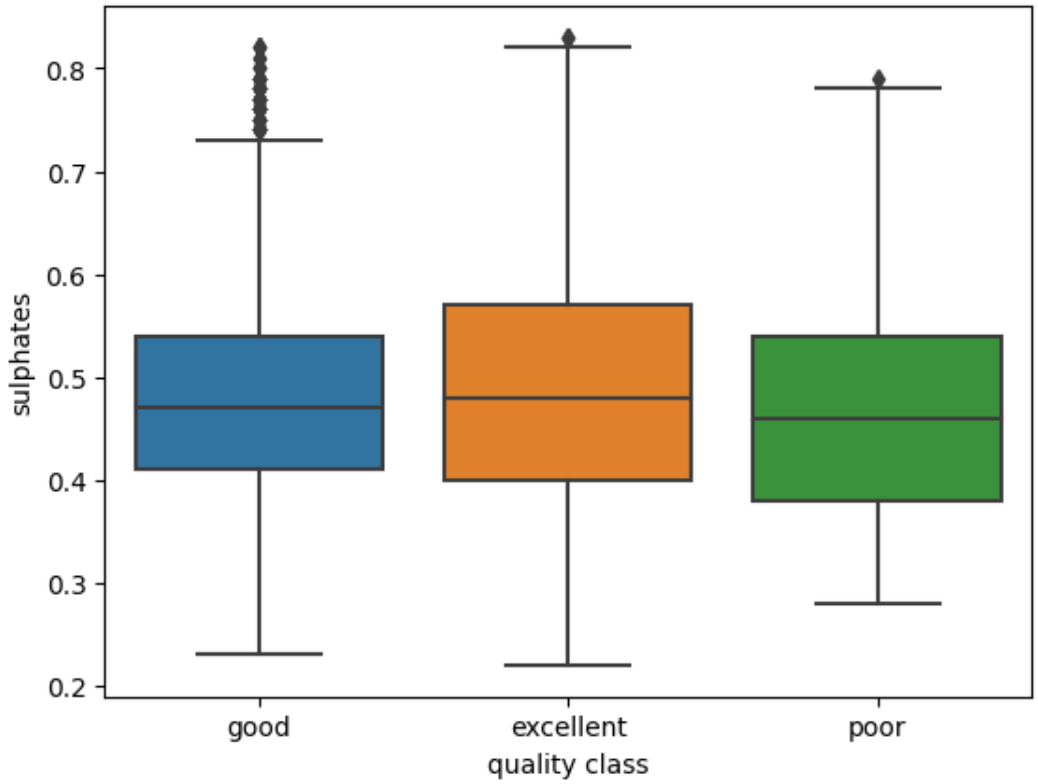
Correlation between quality class and Citric Acid



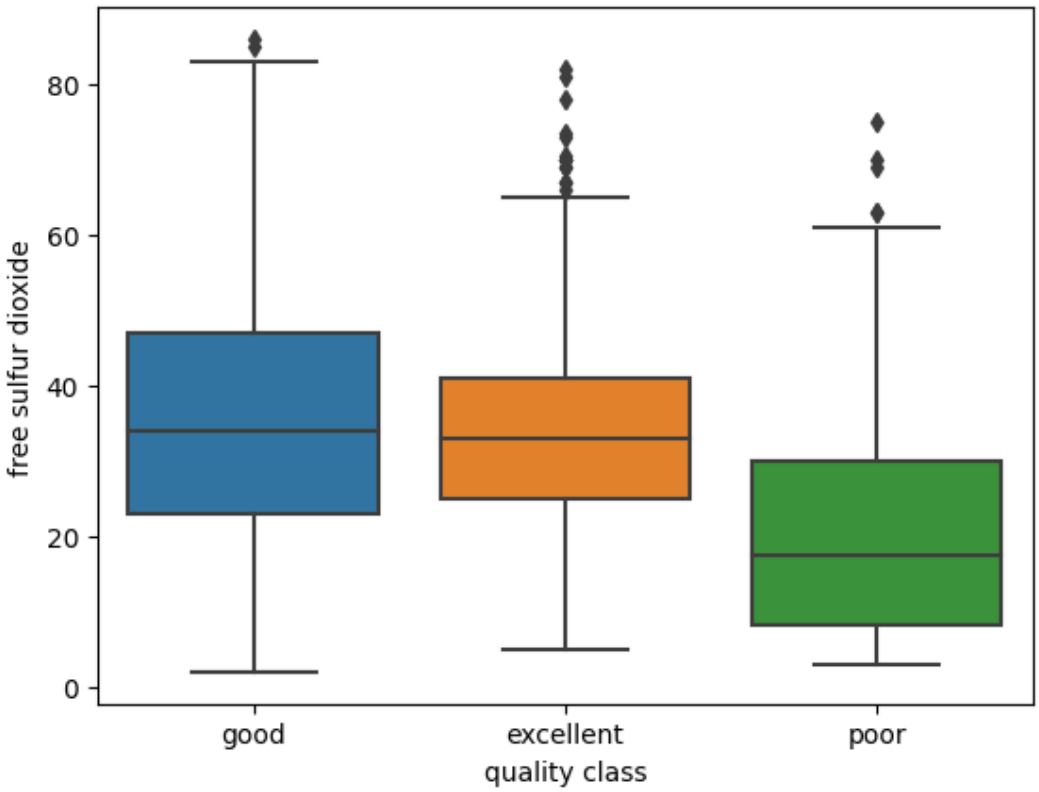
Correlation between quality class and Fixed Acidity



Correlation between quality class and Sulphates



Correlation between quality class and Free sulphur dioxide



2.2 Training models

Next we are going to train different classification algorithms and choose the one with the best results. The best results are measured with the accuracy.
The classification algorithms we are going to test:

- K-nearest neighbors (Knn)
- Decision Tree
- Random Forest

2.2.1 Model selection

When deciding on model selection for this data set we have to take into consideration that we are having a classification problem. Which means that we can not have any regression models as regression algorithms find correlations between dependent and independent variables such as continuous variables. Classification is more about dividing the data into classes and then categorizing it and the algorithm is then dividing the data into these categories. This applies to our data as quality attribute is not a linear attribute like stock prices, etc.

Since we picked the 3 popular classification algorithms to train and test our data off.

2.2.2 K-Nearest Neighbors Classification

Starting with the Knn. To classify a record with Knn it computes the distance to other training records, identifies the k nearest neighbors and then uses class labels of the nearest neighbors to determine the class label by taking the majority vote.

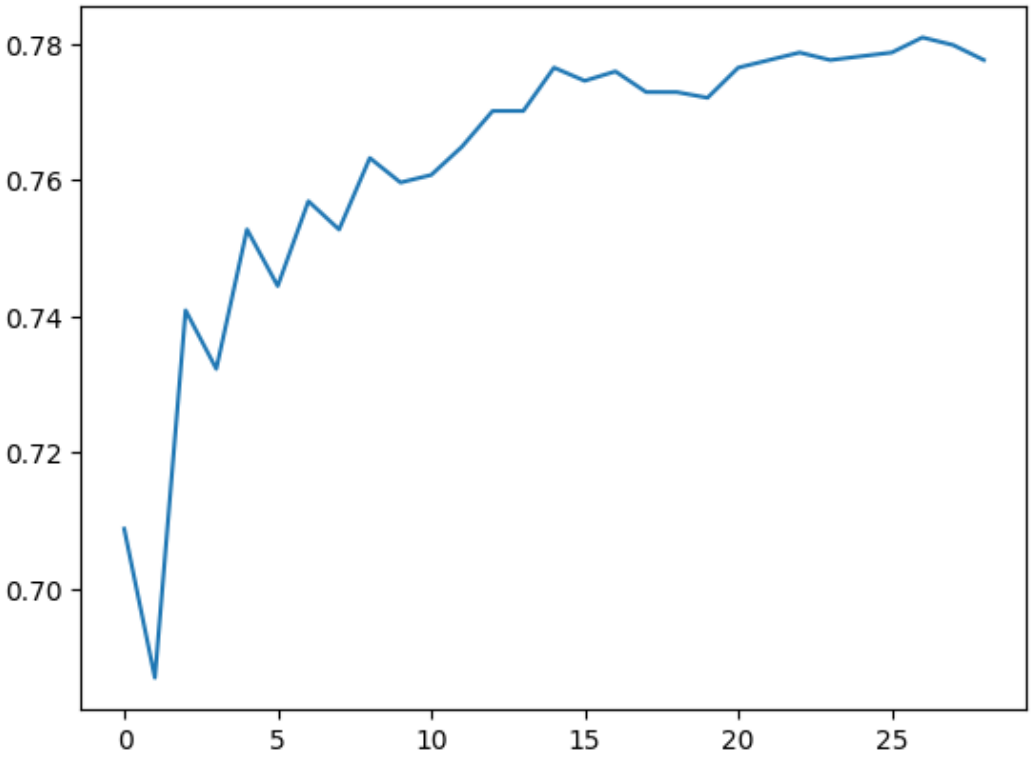
Most problems with choosing the value of k for the algortihm is that if k is to small it is sensitive to noise points and if it is too large it may include points from the other classes e.g. points that are irrelevant.

Good thing to mention is that the reason we scaled the data before going into testing the models is that when testing the Knn and looking at the numbers of which each point represents in the exploratory chapter is that some values are way bigger than others. Scaling is a good thing to prevent distance measures from being dominated by one of the attributes in the data.

This algorithm requires three things:

- The set of labeled records
- Distance metric to compute distance between records
- The value of k, the number of nearest neighbors to retrieve.

To see the best value of k for the algorithm we defined a range for the k to be tested on, 1 to 30, and ran the algorithm through each number and did a cross validation score, with kfold as 15, on each one to measure the accuracy with each k.



Here we can see that the best k value for accuracy is when k = 26

2.2.3 Decision Tree

Decision trees can perform both classification and regression tasks. Decision trees organize the data into a tree-like structure by asking a question. Each time the algorithm asks a question a node is added to the tree. That raises the question on deciding the best split or question. The algorithm tries to divide the data set as small as possible while minimizing the loss function. The loss function evaluates the split based on the amount of data in each class before deciding on a split and after. One of the best ways to evaluate a split is to measure the gini impurity before and after each split. The gini impurity is a measure of variance across the classes. [1]

2.2.4 Random Forest

Random Forest is an algorithm that consists of some number of individual decision trees that operate as an ensemble. Each tree in the random forest algorithm calculates an accuracy and

decides on the best one. [2]

3 Results

Results from testing the data for each of the models:

K-nearest-neighbour results:

	precision	recall	f1-score	support
excellent	0.65	0.35	0.45	157
good	0.81	0.94	0.87	545
poor	1.00	0.00	0.00	22
accuracy			0.79	724
macro avg	0.82	0.43	0.44	724
weighted avg	0.78	0.79	0.75	724

Cross validation score: 0.7787181509550426

Decision Tree results:

	precision	recall	f1-score	support
excellent	0.46	0.46	0.46	157
good	0.82	0.83	0.83	545
poor	0.31	0.18	0.23	22
accuracy			0.73	724
macro avg	0.53	0.49	0.51	724
weighted avg	0.73	0.73	0.73	724

Cross Validation with accuracy: 0.6869986168741353

Random Forest results:

	precision	recall	f1-score	support
excellent	0.67	0.33	0.44	169
good	0.79	0.95	0.86	533
poor	0.50	0.05	0.08	22
accuracy			0.78	724
macro avg	0.65	0.44	0.46	724
weighted avg	0.75	0.78	0.74	724

Cross Validation with accuracy: 0.7748659739606555

As we can see the K-Nearest Neighbors Classification has the highest cross validation, but that does not mean that it´s the most ideal model, it never predicted a poor wine because the recall for the poor wine is 0, which means we need to improve the model. Random forest had the second highest cross validation, but it also had 0 recall for poor wine. Decision tree had the lowest cross validation of the three models we trained.

As we can see the K-Nearest Neighbors Classification has the highest cross validation, but that does not mean that it´s the most ideal model, it never predicted a poor wine because

the recall for the poor wine is 0, which means that the model never took the poor quality into consideration and might not even have had one in the training set to begin with. Which means that to get a better prediction with the K-nearest-neighbour algorithm we needed more data to work with. Even though the Knn never took poor quality into consideration the accuracy prediction with cross validation method is not a good method because it uses linear prediction so we have to look at precision, recall and f1-score instead. Since the knn had 0 in both recall and f1-score we can not take that model into consideration. As for the others we can see that the weighted average numbers: recall, f1-score, precision, are overall higher in random forest algorithm, we can say that we would use that algorithm to predict with as it has 75% average prediction accuracy while decision tree has 73% average prediction accuracy.

4 Conclusion

In conclusion these methods had not a good accuracy of prediction because the data set was too small. If there would be more data and to improve the accuracy further would use the help of neural networks to improve the accuracy.

References

[1] Carolina Bento. Decision tree classifier explained in real-life: picking a vacation destination. https://towardsdatascience.com/decision-tree-classifier-explained-in-real-life-picking-a-vacation-destination-6226b1

[2] ony Yiu. Understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.

[3] F. Almeida T. Matos P. Cortez, A. Cerdeira and J. Reis. Modeling wine preferences by data mining from physicochemical properties.

[4] Madeline Puckette. What is wine exactly? https://winefolly.com/deep-dive/what-is-wine/.

[5] Fares Sayah. Outlier detection using pdf and z-score. https://www.kaggle.com/code/faressayah/outlier-detection-using-pdf-and-z-score.