Regression Analysis Final Project

Drink Quality Classification

Farzad Roiintan: 98100153

Ahmadreza Tavana: 98104852

Professor: Dr. Mirsadeghi

1. Preview

First of all, we examine the dataset and search on the columns and rows to knowing the dataset better. The dataset which we use in this project is collected from Kaggle which is a website that contains many good datasets for analysis in machine learning and data science. Our data set is about some drink and their ingredients and we want to do some classification to classify the quality of the drinks by methods that we learned in the course. In the next part we do some visualizing in the dataset to understand that better and discuss about the questions that we can answer with the dataset and also make some changes in data to make it better to use for our models.

2. Data visualizing and Data Cleaning

At first we introduce the properties of the dataset. We put the first rows of the data in below to figure out the features.

index	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8	6
4	7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

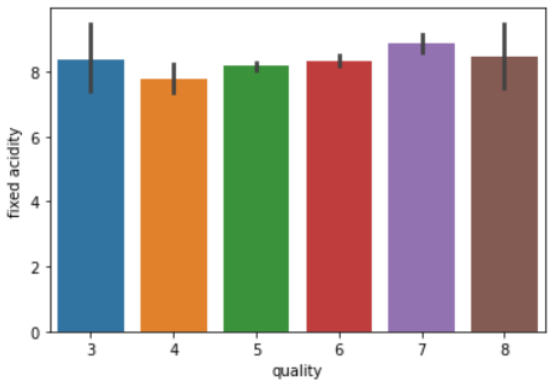
As we see the data has 12 columns which is number of the drink (index), fixed acidity, volatile acidity, citric sugar, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, pH, sulphates, alcohol, and the main parameter with is our response parameter that we classify the drinks by that is quality. The number of drinks in the dataset is 1600 so n (number of data is 1600) and p (number of features is 11 for prediction). In the dataset we have some duplicated rows which we remove them to have better validation in the models and have better conclusions on the data. The duplicated rows are as below. We removed them all for our models.

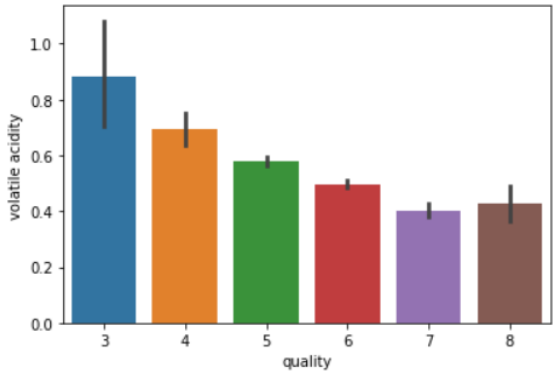
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
4	7.4	0.700	0.00	1.90	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
11	7.5	0.500	0.36	6.10	0.071	17.0	102.0	0.99780	3.35	0.80	10.5	5
27	7.9	0.430	0.21	1.60	0.106	10.0	37.0	0.99660	3.17	0.91	9.5	5
40	7.3	0.450	0.36	5.90	0.074	12.0	87.0	0.99780	3.33	0.83	10.5	5
65	7.2	0.725	0.05	4.65	0.086	4.0	11.0	0.99620	3.41	0.39	10.9	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1563	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1564	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1567	7.2	0.695	0.13	2.00	0.076	12.0	20.0	0.99546	3.29	0.54	10.1	5
1581	6.2	0.560	0.09	1.70	0.053	24.0	32.0	0.99402	3.54	0.60	11.3	5
1596	6.3	0.510	0.13	2.30	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6

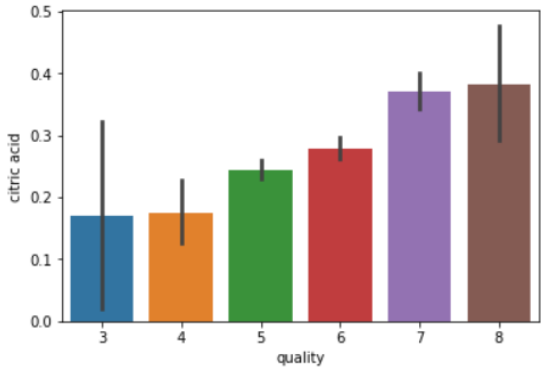
240 rows × 12 columns

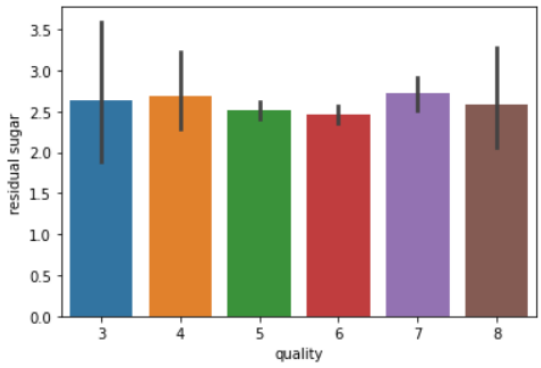
The pure data without duplicated rows has 1360 rows so in our models we work with 1360 drinks.

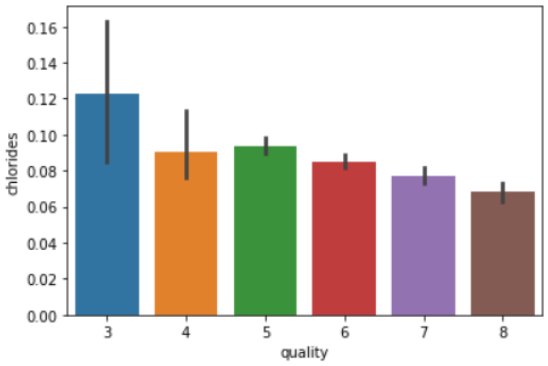
Now we show bar plot of data in below to show the connections between each parameter with the quality of drinks. Notice that the quality is a number between 3 to 8 shows how the drink is good and desirable.

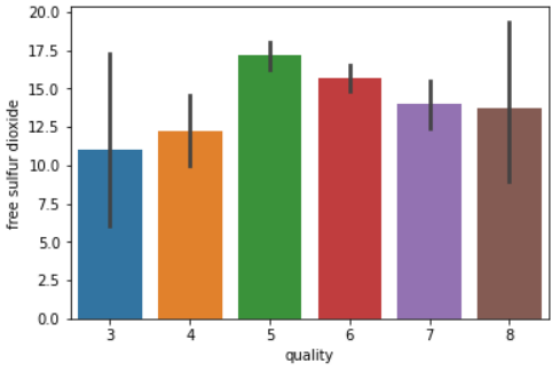


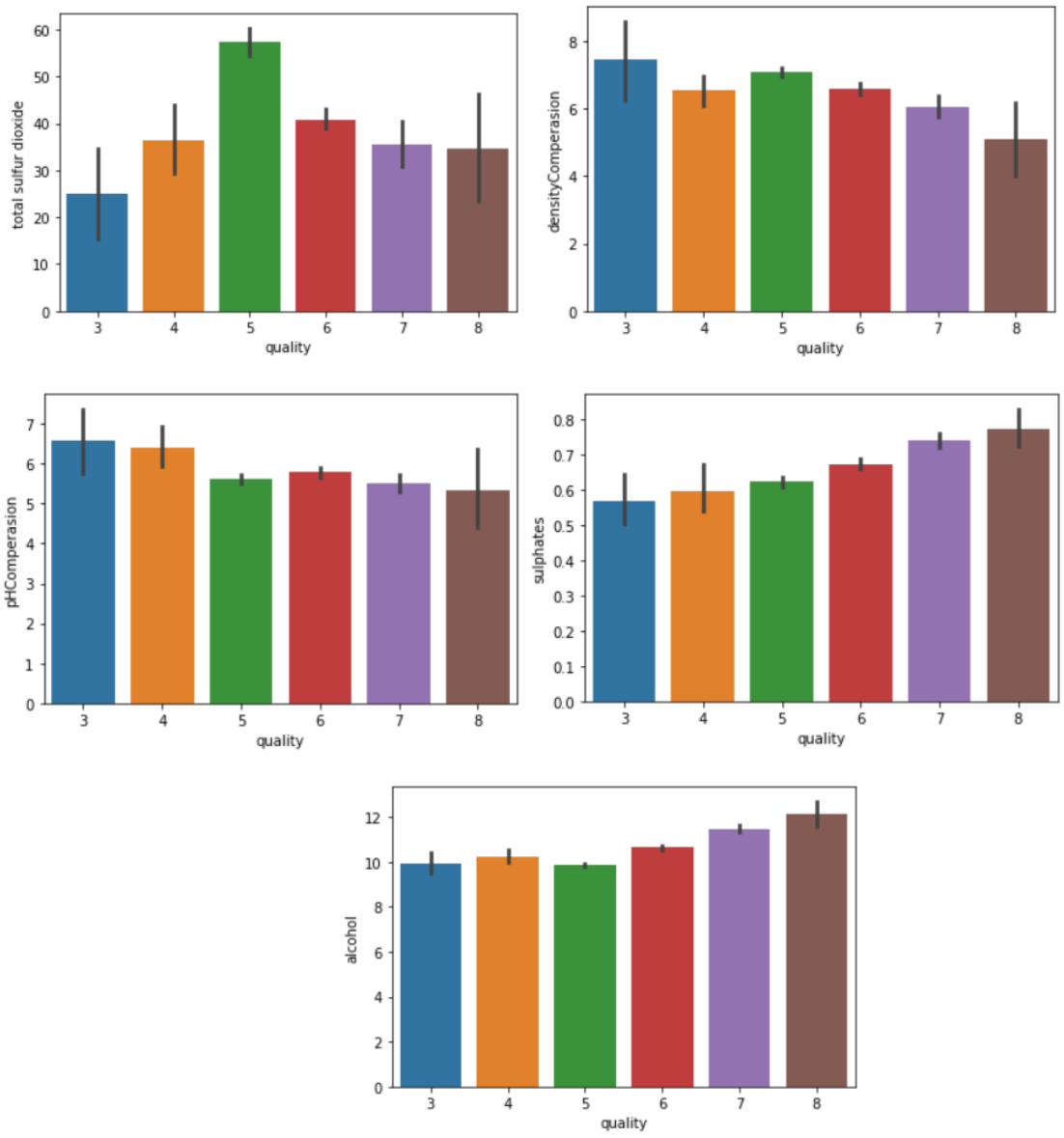












Note that pH, and density of drinks are almost equal so we normalize them to use them in the model for better comparison between these parameters for our prediction.

Now base on some connections that we can see in the bar plots, we ask some question in the next parts and fit a model to see whether there would be some connections between parameters or not and learn that model and test it with cross validation.

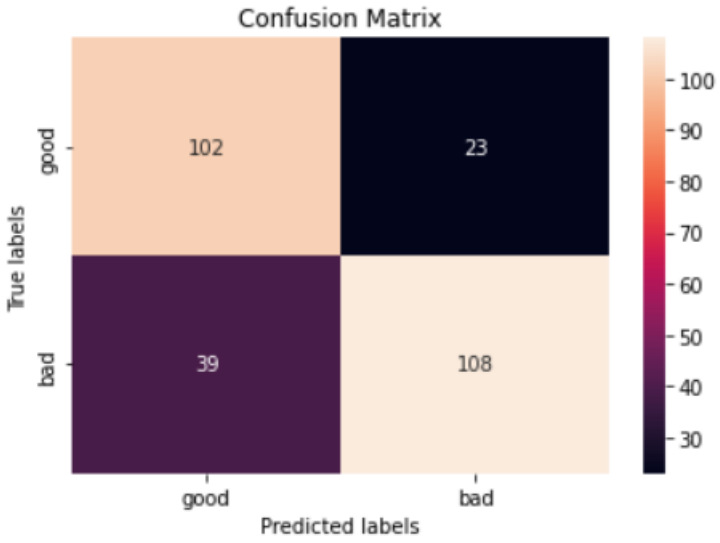
In the next parts we would name the models that we fit and show the answers that we get from the questions and the models.

Then we calculate VIF of variables to figure out collinearity between variables so we can consider these collinearities in the models which we fit. The VIF result is as below:

	feature	VIF
0	fixed acidity	182.128549
1	volatile acidity	16.439427
2	citric acid	8.866208
3	residual sugar	7.141456
4	chlorides	5.961549
5	free sulfur dioxide	6.469125
6	total sulfur dioxide	6.133908
7	sulphates	23.312328
8	alcohol	192.541000
9	densityComperasion	81.566470
10	pHComperasion	46.300526

3. Logistic Regression

For this part we separate the drink quality to 2 groups for logistic regression to classify drinks base on the qualities. So we label drinks with quality of 2 to 5.5 as a bad drink and 5.5 to 8 as a good drink and do logistic regression. For producing the cross validation, we keep 0.2 of the data set as cross validation data and do the logistic regression on that. Now let's see the results of the regression in below. The confusion matrix of the regression is as below:



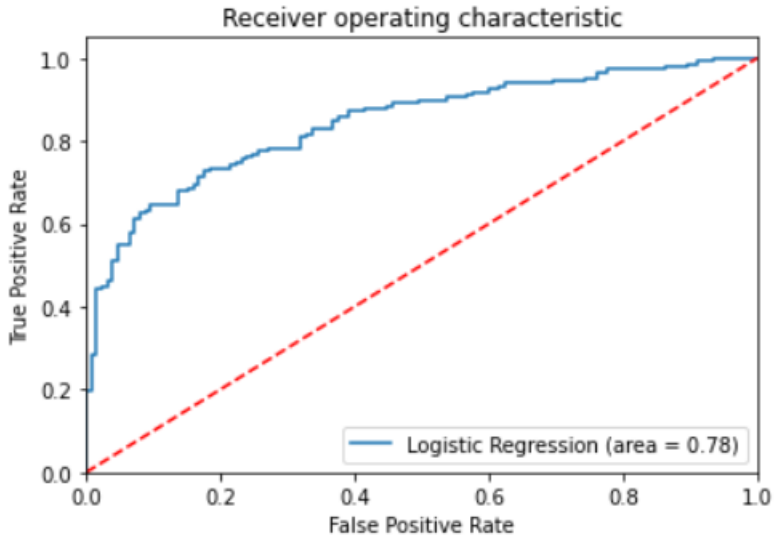
The coefficient of each parameter from the logistic regression is as below:

	features	coef
8	alcohol	0.978997
6	total sulfur dioxide	0.554364
1	volatile acidity	0.546115
7	sulphates	0.467124
2	citric acid	0.230819
5	free sulfur dioxide	0.199027
4	chlorides	0.163906
0	fixed acidity	0.120693
10	pHComperasion	0.072380
3	residual sugar	0.022965
9	densityComperasion	0.015421

And the final accuracy for classification is as below:

```
Accuracy of logistic regression classifier on test set: 0.77
```

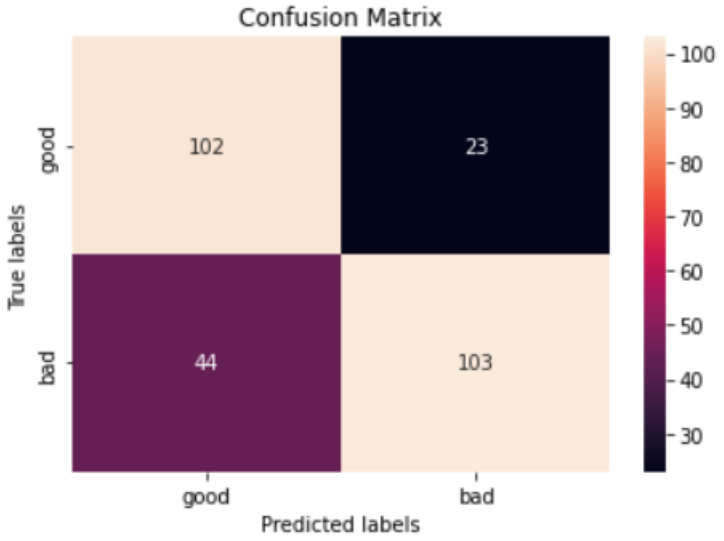
The ROC curve of the logistic regression is as below:



4. LDA (Linear discriminant analysis)

For this model, we use the last training and cross validation dataset to compare the result of LDA with logistic regression. We put the results of this model in following and then compare it with logistic regression.

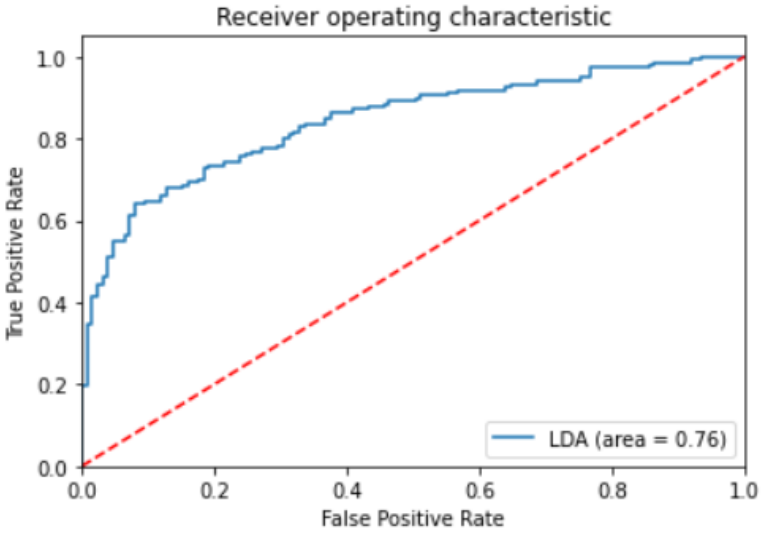
The confusion matrix of LDA is as below and as we see, it's almost as the same as logistic regression but the accuracy and sensitivity of logistic regression was better but the results doesn't have major difference:



The accuracy of the LDA model is:

```
Accuracy of LDA classifier on test set: 0.75
```

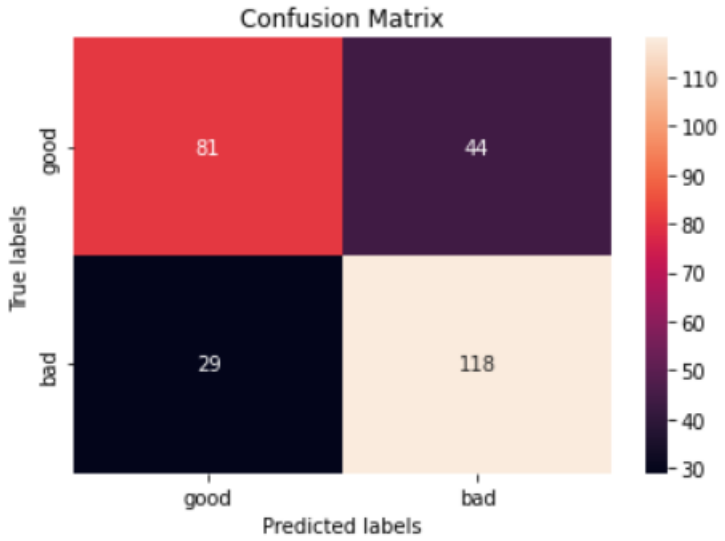
The ROC curve of LDA model is as below:



5. QDA (Quadratic Discriminant Analysis)

For this model, we use the last training and cross validation dataset to compare the result of QDA with logistic regression and LDA. We put the results of this model in following and then compare it with logistic regression and LDA.

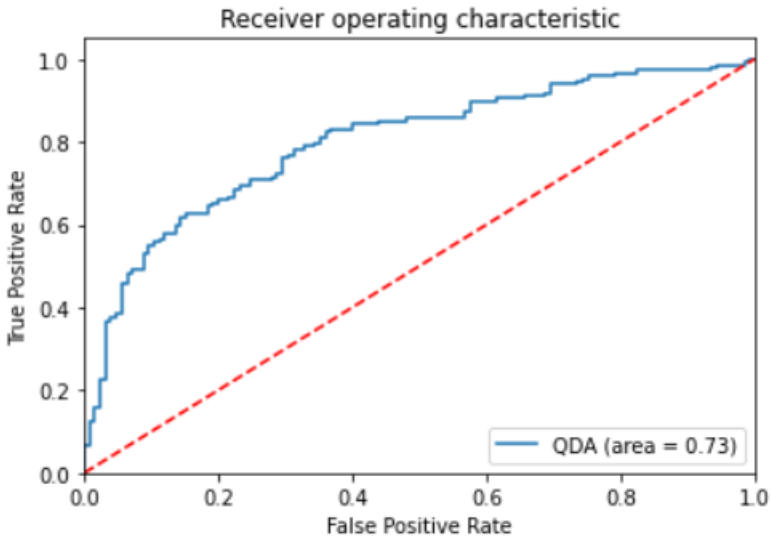
The confusion matrix of QDA is as below and as we see, it's almost as the same as logistic regression and LDA but the accuracy and sensitivity of logistic regression was better than both LDA and QDA but the results doesn't have major difference:



The accuracy of the QDA model is:

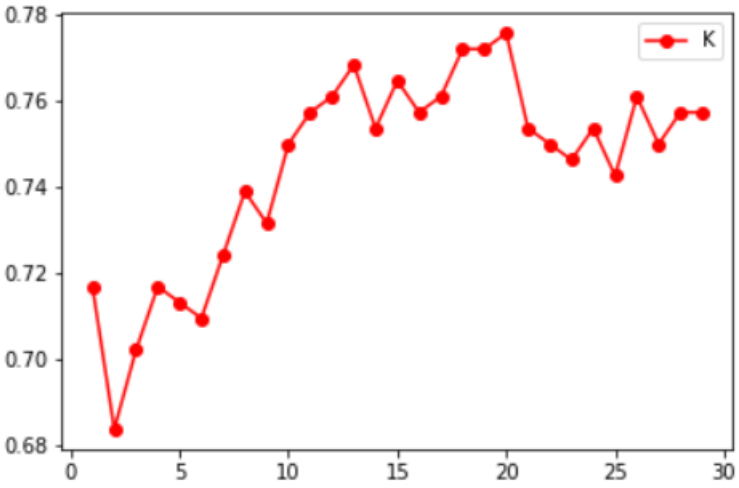
```
Accuracy of QDA classifier on test set: 0.73
```

The ROC curve of LDA model is as below:



5. KNN (k-nearest neighbors)

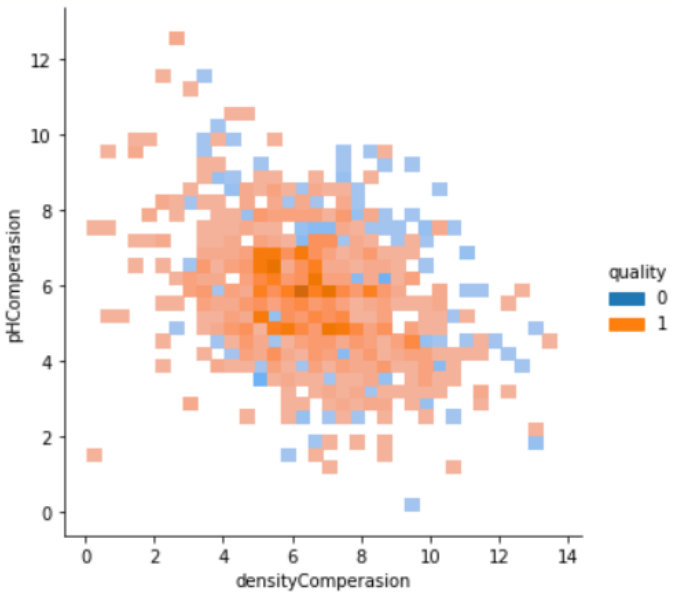
For KNN we used the data set as the previous parts and try to find the best K with best accuracy. The result is as below and K is about 20 and the accuracy is about 78%:



As we can see, with many fall and downs in accuracy, the best K for this data set and the way of producing validation dataset, is K = 20. The accuracy is a little bit more than logistic regression in this method.

6. Feature Combination

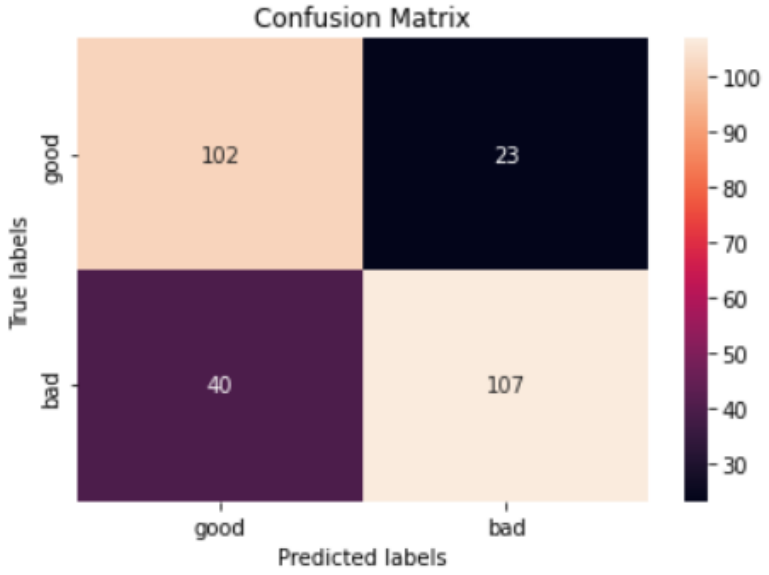
With more attention to relations between features, we can understand that pHCompersion (the parameter that ourselves made as a new parameter) and densityCompersion have low coefficients so we can check the relation between these 2 parameters and quality of the drink in a plot to figure out if there is a relation:



it seems we can draw a circle with x = 7 and y = 6 center, in circle higher chance of quality = 1 so we define a new variable, like $z = (x-7)^2 + (y-6)^2$, linear relation between z and x, z and y can show themselves in coefficients of x and y, so we just need to describe 2 new variable like x^2 and y^2.

volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	sulphates	alcohol	quality	densityComperasion	pHComperasion	extfeature1	extfeature2
0.70	0.00	1.9	0.076	11.0	34.0	0.56	9.4	0	7.8	7.7	60.84	59.29
0.88	0.00	2.6	0.098	25.0	67.0	0.68	9.8	0	6.8	4.6	46.24	21.16
0.76	0.04	2.3	0.092	15.0	54.0	0.65	9.8	0	7.0	5.2	49.00	27.04
0.28	0.56	1.9	0.075	17.0	60.0	0.58	9.8	1	8.0	4.2	64.00	17.64
0.66	0.00	1.8	0.075	13.0	40.0	0.56	9.4	0	7.8	7.7	60.84	59.29

Now we do a logistic regression on the new data with new features. The confusion matrix of logistic regression is as below:



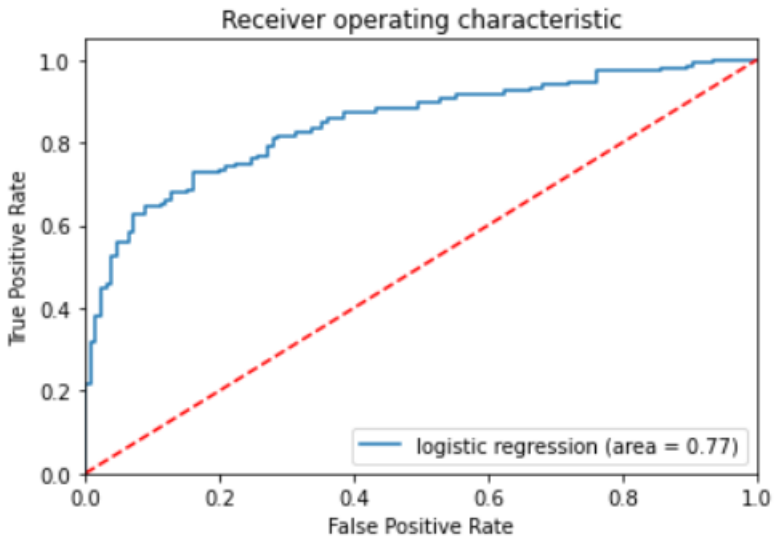
The accuracy of the model is:

```
Accuracy of logistic regression classifier with new featcures on test set: 0.77
```

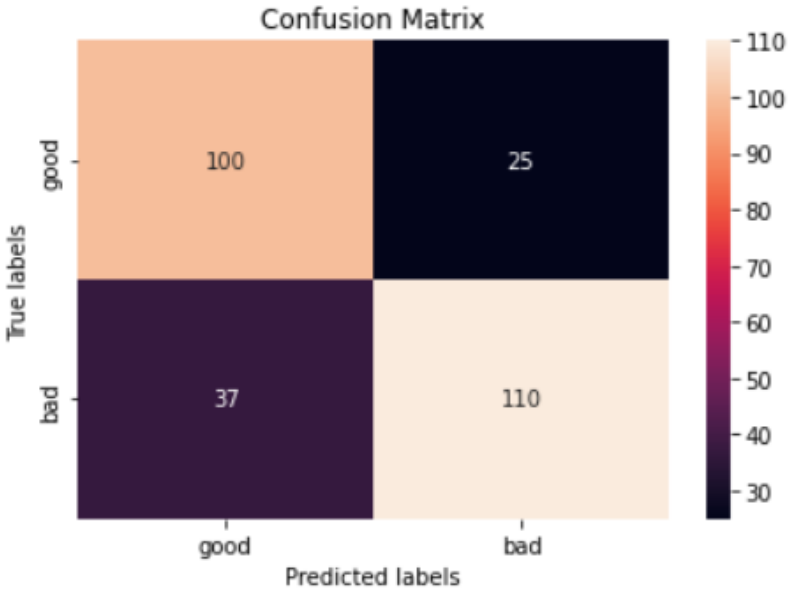
The coefficients of parameters are as below:

	features	coef
8	alcohol	0.968561
6	total sulfur dioxide	0.557041
1	volatile acidity	0.536250
7	sulphates	0.478142
12	extfeature2	0.262656
11	extfeature1	0.261802
9	densityComperasion	0.248294
2	citric acid	0.221128
5	free sulfur dioxide	0.196050
10	pHComperasion	0.181043
4	chlorides	0.161906
0	fixed acidity	0.111339
3	residual sugar	0.004965

And the ROC curve of that is as below:



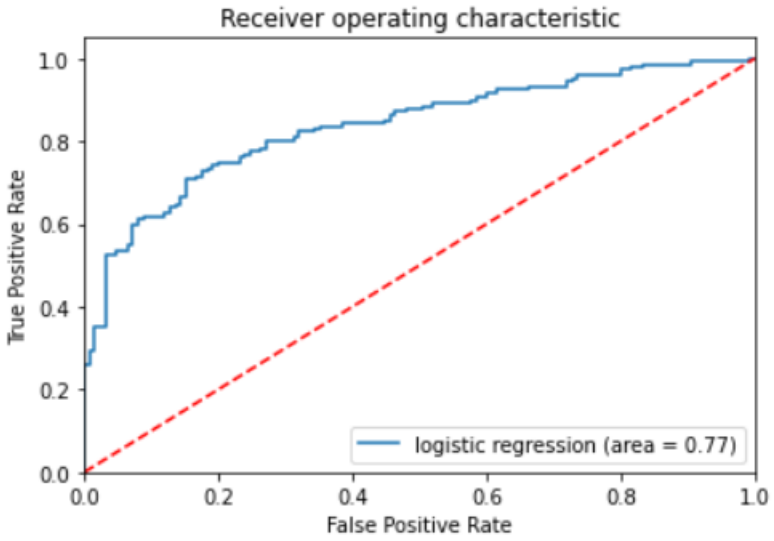
Now as the accuracy of the model increased we can add some other features, then we can choose the best p of them. We started to adding X^2 of the features and then do another logistic regression and see the results. The confusion matrix of the new model is as below:



The coefficients of this logistic regression are as below:

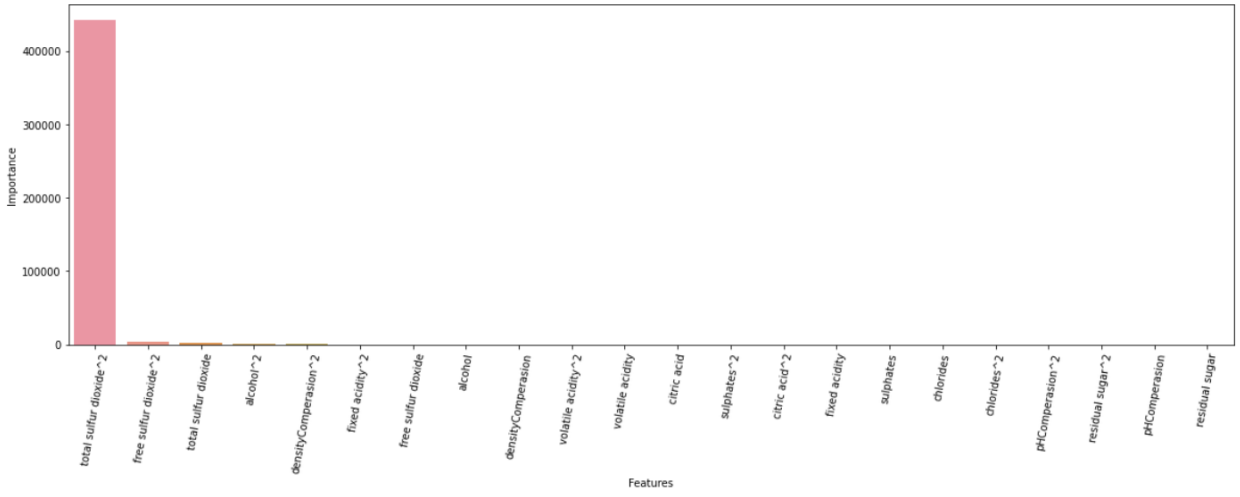
	features	coef
8	alcohol	0.968561
6	total sulfur dioxide	0.557041
1	volatile acidity	0.536250
7	sulphates	0.478142
12	extfeature2	0.262656
11	extfeature1	0.261802
9	densityComperasion	0.248294
2	citric acid	0.221128
5	free sulfur dioxide	0.196050
10	pHComperasion	0.181043
4	chlorides	0.161906
0	fixed acidity	0.111339
3	residual sugar	0.004965

The ROC curve of that is as below:

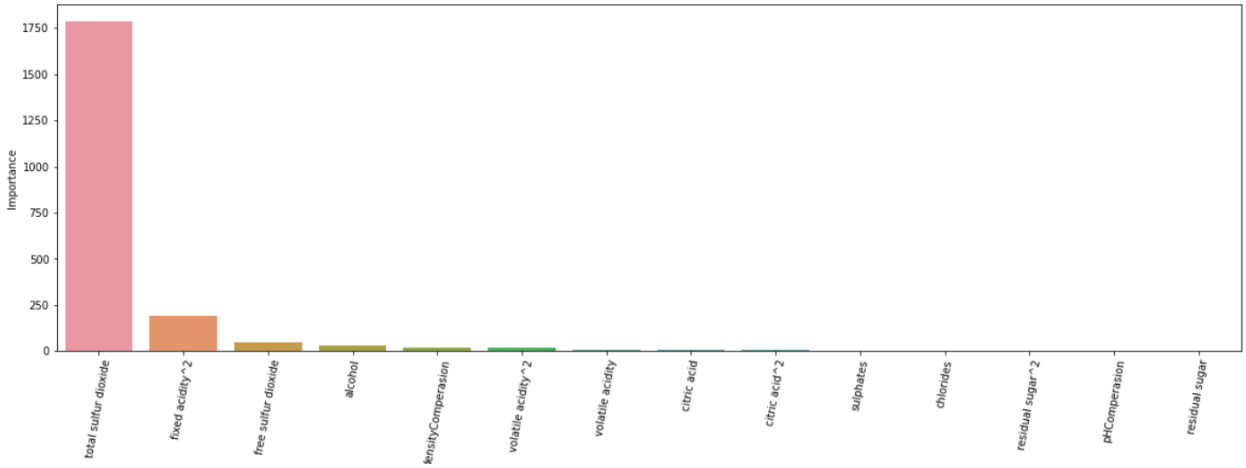


7. Feature Selection

Now with some edition that we made in the data set and producing some new features, we now want to select the ones which are the most effective ones to clarify quality of drinks. As we add some parameters we have 22 parameters overall and we do the feature selection on these 22 features. The result is as below:



As we can see in above the importance of the squared of total sulfuric dioxide is the most important parameter for quality of drinks. The differences between other parameters is not as clear as enough so we remove the squared of the total sulfuric to compare other features.



With removing that feature we can see the importance other factors better and as we can see total sulfur dioxide, squared of fixed acidity, free sulfur dioxide, alcohol, densityComperasian are other important factors by sort. For next parts we use this feature reduction to fit our models.

The next method for feature combination is a polynomial method to combine parameters to make new features for fitting models. The result of this combination is a dataset with 78 parameters that almost have many connections between parameters. The result of this polynomial combination is as below:

	features	coef
68	sulphates^2	1.045940
69	sulphates alcohol	0.979055
15	fixed acidity residual sugar	0.907656
60	free sulfur dioxide alcohol	0.839195
43	residual sugar chlorides	0.804127
..	...	...
76	densityComperasion pHComperasion	0.007060
42	residual sugar^2	0.003405
39	citric acid alcohol	0.002508
46	residual sugar sulphates	0.002207
0	1	0.000000

Another method that we use for feature selection is sequential feature selector and the result of this feature selection is as below:

```
array(['volatile acidity sulphates',
       'volatile acidity densityComperasion', 'chlorides alcohol',
       'sulphates^2', 'sulphates alcohol'], dtype=object)
```

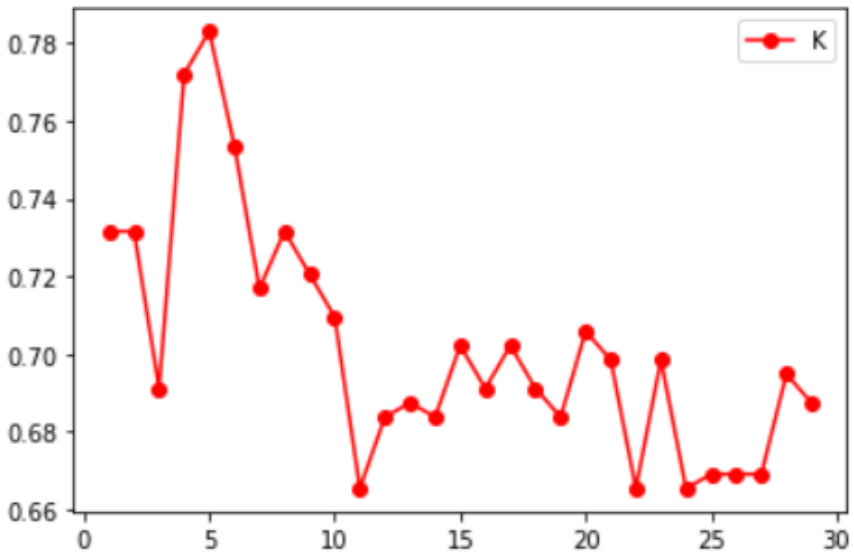
For making sure that we didn't select some useless features, we calculate VIF data that collected from the previous sequential feature selection. The result of this calculation is as below:

	feature	VIF
0	volatile acidity sulphates	25.401374
1	volatile acidity densityComperasion	12.856046
2	chlorides alcohol	5.780141
3	sulphates^2	9.946207
4	sulphates alcohol	15.325321

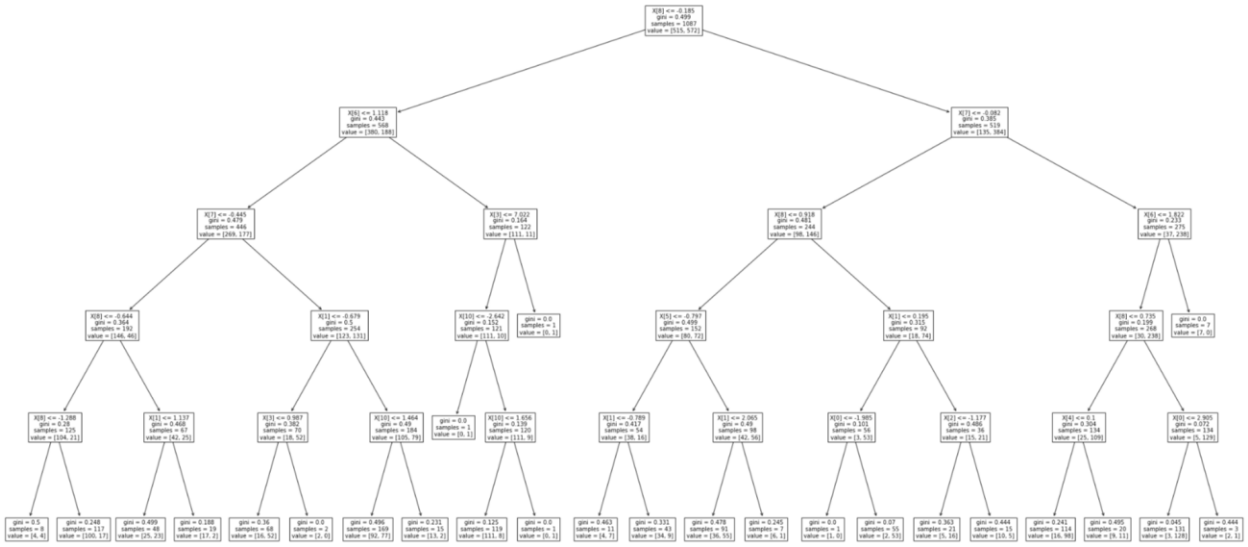
As we can see VIF of volatile acidity sulphates and sulphate alcohol is large so if we use this data set for next parts it's better to remove these parameters.

8. Decision Tree

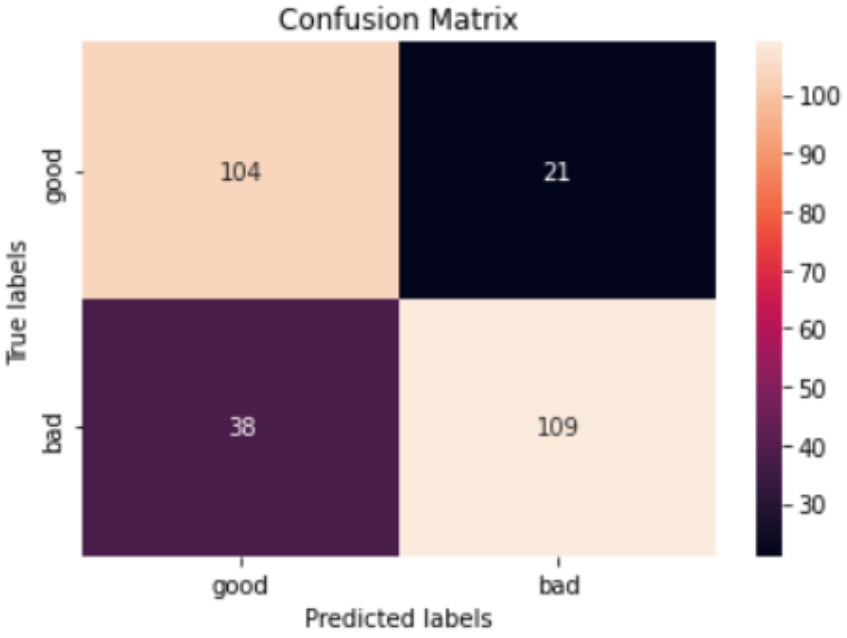
In this part we use decision tree for classifying drinks. In this method at first we try to find best depth with the most accuracy by using cross validation. The result of this is method is as below:



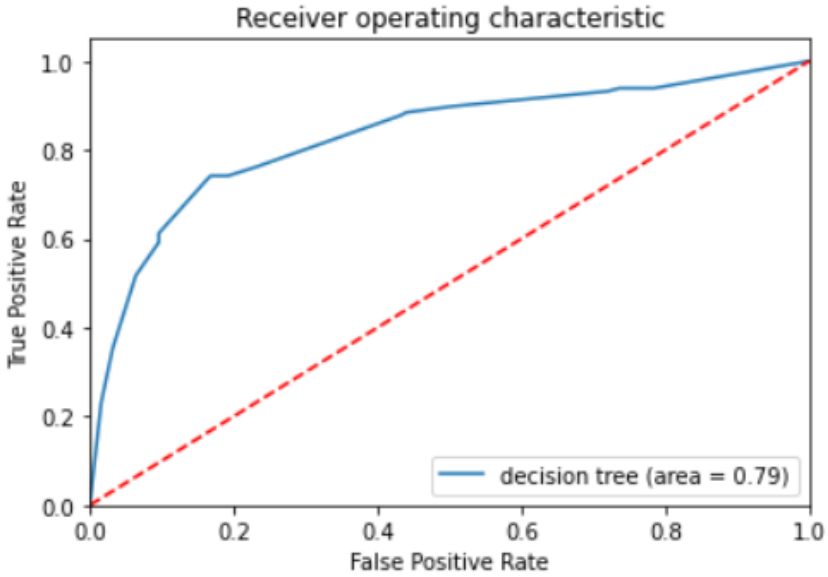
As we can see the best accuracy is in length of 5 and after that we came to realize that we may have overfitting. So we choose the length of 5 and try to fit the model with this length and see the tree itself, confusion matrix and ROC curve of that. The tree that we claimed is as blew:



The confusion matrix of this model is as below:



Now the ROC curve of tree that we fit with length of 5 is as below:



As we can see the accuracy of this model is a bit better that logistic regression, LDA, QDA and KNN. The reason is that with a small length of like 5 we prevent overfitting very well and the model does as optimal as it can and this simplicity in modeling that comes from small number of classes cause this increasing accuracy.

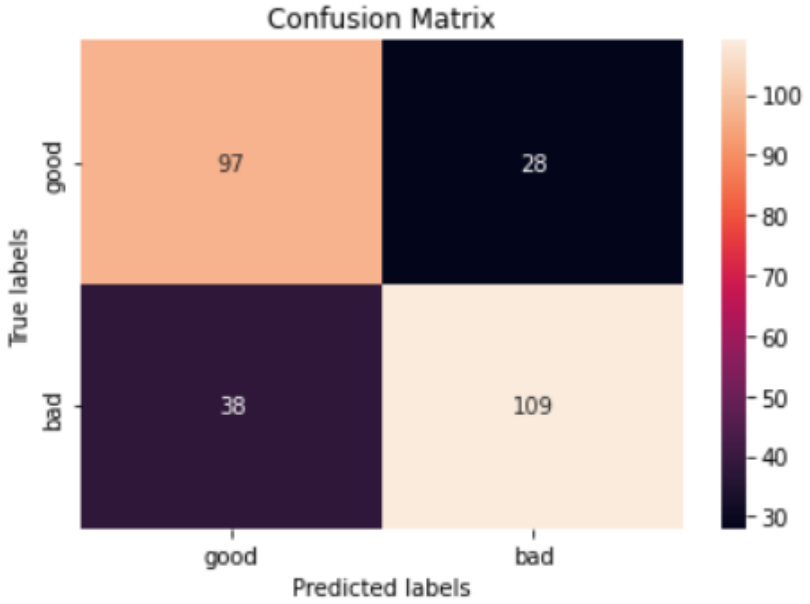
9. Random Forest

Random forest is a method that it chooses in order of square root of number of original data features and train a tree for that features. We train a random forest with length of 3 for drinks to classify and the result is as below:

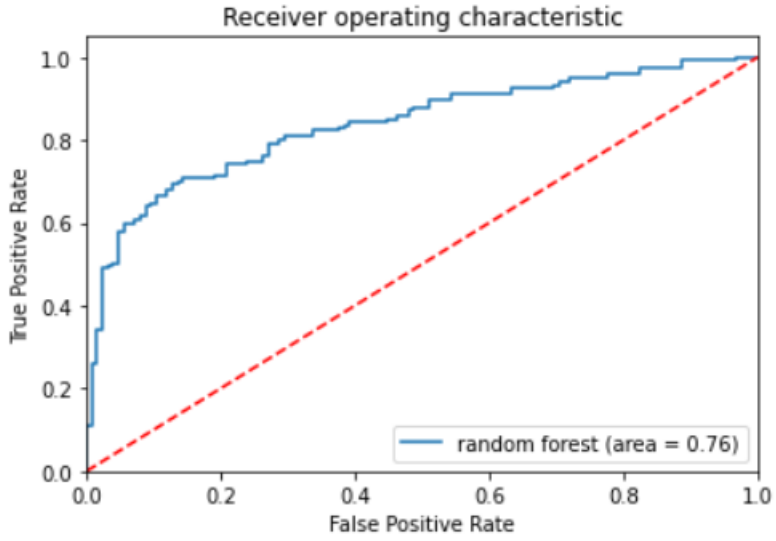
	precision	recall	f1-score	support
0	0.72	0.78	0.75	125
1	0.80	0.74	0.77	147
accuracy			0.76	272
macro avg	0.76	0.76	0.76	272
weighted avg	0.76	0.76	0.76	272

As we can see the accuracy of random forest with length of 3 is a bit lower than other models but the difference is not major.

The confusion matrix of this model is as below and as we can see the sensitivity decreased and false true rates increased so this can't be a good model for this dataset.



The ROC curve of this random forest model is as below:



As I explain in above the accuracy decreased so we can't count this model as an optimal model for our drink dataset.

10. SVC (Support Vector Classifier)

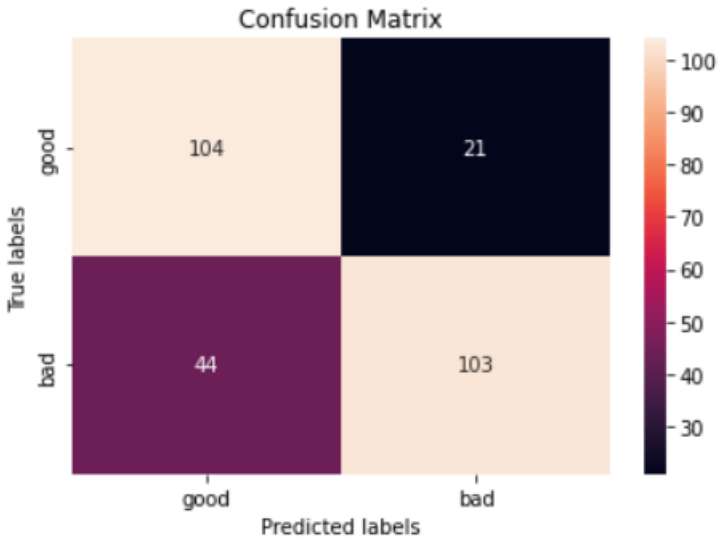
SVC of support vector classifier is another classifier that we learn in the course and we fit it with different kernels on the dataset and the results of these fits are as below:

10.1. SVC with “liner” kernel

The accuracy for this model is as below:

	precision	recall	f1-score	support
0	0.70	0.83	0.76	125
1	0.83	0.70	0.76	147
accuracy			0.76	272
macro avg	0.77	0.77	0.76	272
weighted avg	0.77	0.76	0.76	272

The confusion matrix of SVC with linear kernel is as below:



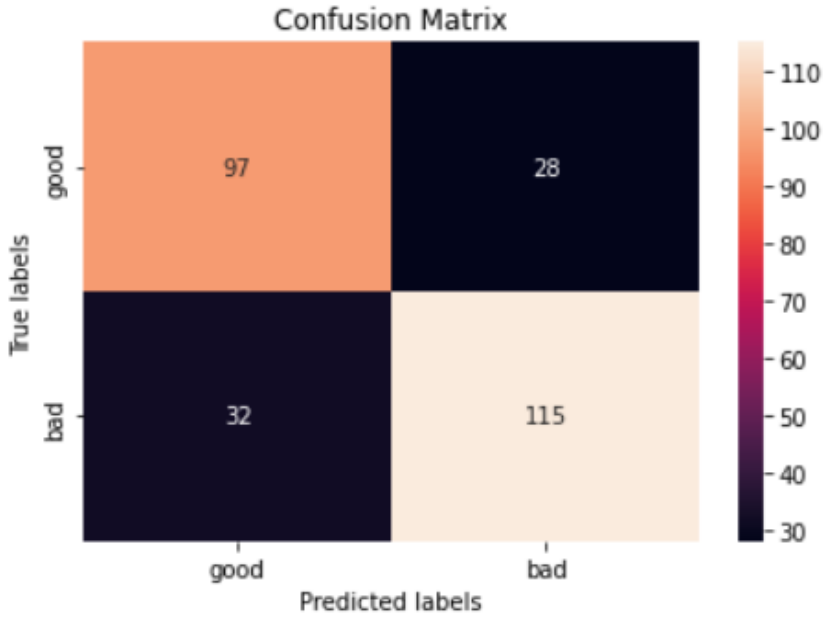
As we can see we have an improve in increasing true positives and true negatives and the accuracy was better than random forest, LDA and QDA.

10.2. SVC with “polynomial” kernel

The accuracy for this model is as below:

	precision	recall	f1-score	support
0	0.75	0.78	0.76	125
1	0.80	0.78	0.79	147
accuracy			0.78	272
macro avg	0.78	0.78	0.78	272
weighted avg	0.78	0.78	0.78	272

The confusion matrix of SVC with polynomial kernel is as below:



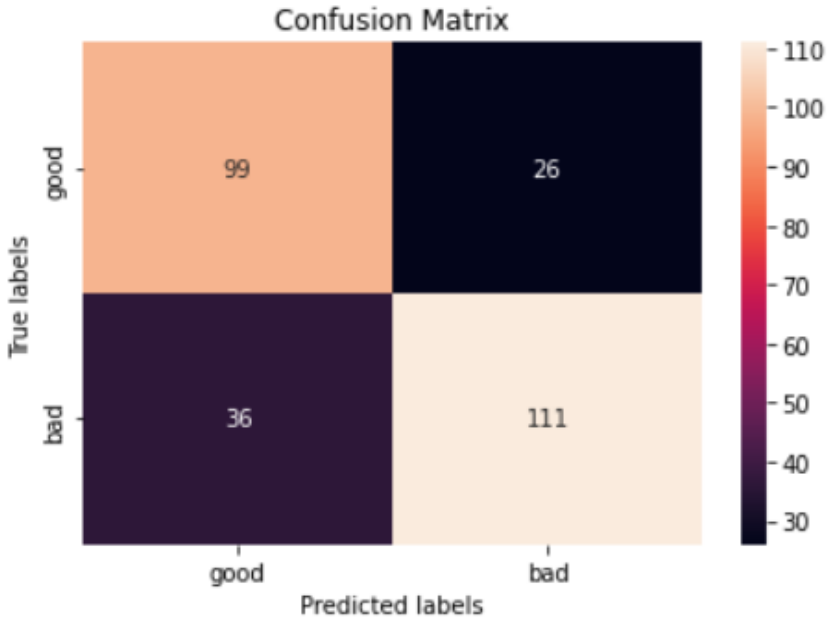
As we can see the accuracy and true positives got less so we can conclude that the model gets worse and the last kernel was better for the model.

10.3. SVC with “RBF” kernel

The accuracy for this model is as below:

	precision	recall	f1-score	support
0	0.73	0.79	0.76	125
1	0.81	0.76	0.78	147
accuracy			0.77	272
macro avg	0.77	0.77	0.77	272
weighted avg	0.77	0.77	0.77	272

The confusion matrix of SVC with RBF kernel is as below:



As we can see the true positives of this model got larger and the accuracy of the total model increased to and we can conclude that the features of the model are distributed in a circular matter that the SVC could separate them better.

11. Conclusion

In this project we tried to find best model for classifying the list of drinks that we had in the dataset. We tried 9 models with changing dataset and using cross validation in some parts. The result was that logistic regression, KNN and decision tree had best accuracies and they could predict the test model properly. The accuracy was about 80 percent and that was for that the number of features were a bit large for classification and the dataset itself was not prepared well to be classified in the best way. Classifying these drinks in the real life is so important generally because it directly works with health and having the high accuracy models is essential for such usage.

12. References

1. https://www.analyticsvidhya.com/blog/2021/04/beginners-guide-to-decision-tree-classification-using-python/

2. https://scikit-learn.org/stable/modules/classes.html#module-sklearn.ensemble

3. https://data36.com/coding-a-decision-tree-in-python-classification-tree-gini-impurity/

4. https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC

5. https://machinelearningmastery.com/random-forest-ensemble-in-python/

6. https://vitalflux.com/random-forest-classifier-python-code-example/

7. https://scikit-learn.org/stable/auto_examples/tree/plot_iris_dtc.html#sphx-glr-auto-examples-tree-plot-iris-dtc-py

8. https://stackoverflow.com/questions/59447378/sklearn-plot-tree-plot-is-too-small

9. https://stackabuse.com/implementing-lda-in-python-with-scikit-learn/

10. https://towardsdatascience.com/quadratic-discriminant-analysis-ae55d8a8148a

11. https://scikit-learn.org/0.16/modules/generated/sklearn.qda.QDA.html