ECE595IDMProject

Chengjun Guo

December 2022

1 Introduction

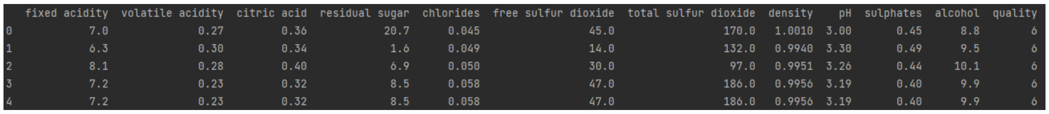
Wine quality dataset is a dataset of red wines and white wines from the Vinho Verde region of Portugal by P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.[Cor+09]. They propose a data mining approach to predict human taste preference which is quantified as wine quality from 0 to 10. The input variables of the dataset is based on physico chemical tests including: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates and alcohol. The model designed is useful to predict the wine tasting evaluations and improve wine production. Furthermore, it can also help wine industry model consumer tastes.

2 Formulation

Since the wine quality is rating in integers, it is a multiclass classification problem from 0 to 10. The input includes fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates and alcohol. All of them are float numbers. The output is the wine quality which can be considered as discreted labels. The statistic information of them can be found in the next section.

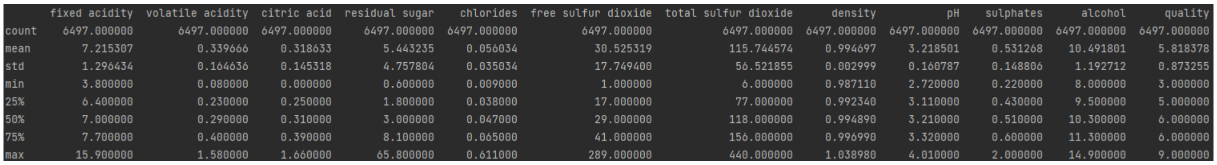
3 Datasets

This dataset is documented by UCI Machine learning Repository: https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/. For the implementation, I concatenated white and red wines together. As we can see in 1, this is the first five lines of the white wine dataset. We can find the statistical information of the input data in 2. In this dataset, some property such as chlorides has mean as low as 0.056 while property such as total sulfur dioxide has mean as high as 115.745. We can notice that there are multiple attributes named with acid. Thus, the correlation between the attributes would be necessary. It can be found in 3.



```
   fixed acidity  volatile acidity  citric acid  residual sugar  chlorides  free sulfur dioxide  total sulfur dioxide  density    pH  sulphates  alcohol  quality
0            7.0              0.27         0.36            20.7      0.045                 45.0                 170.0   1.0010  3.00       0.45      8.8        6
1            6.3              0.30         0.34             1.6      0.049                 14.0                 132.0   0.9940  3.30       0.49      9.5        6
2            8.1              0.28         0.40             6.9      0.050                 30.0                  97.0   0.9951  3.26       0.44     10.1        6
3            7.2              0.23         0.32             8.5      0.058                 47.0                 186.0   0.9956  3.19       0.40      9.9        6
4            7.2              0.23         0.32             8.5      0.058                 47.0                 186.0   0.9956  3.19       0.40      9.9        6
```

Figure 1: first 5 lines of the dataset



```
       fixed acidity  volatile acidity  citric acid  residual sugar    chlorides  free sulfur dioxide  total sulfur dioxide      density           pH    sulphates      alcohol      quality
count    6497.000000       6497.000000  6497.000000     6497.000000  6497.000000          6497.000000           6497.000000  6497.000000  6497.000000  6497.000000  6497.000000  6497.000000
mean        7.215307          0.339666     0.318633        5.443235     0.056034            30.525319            115.744574     0.994697     3.218501     0.531268    10.491801     5.818378
std         1.296434          0.164636     0.145318        4.757804     0.035034            17.749400             56.521855     0.002999     0.160787     0.148806     1.192712     0.873255
min         3.800000          0.080000     0.000000        0.600000     0.009000             1.000000              6.000000     0.987110     2.720000     0.220000     8.000000     3.000000
25%         6.400000          0.230000     0.250000        1.800000     0.038000            17.000000             77.000000     0.992340     3.110000     0.430000     9.500000     5.000000
50%         7.000000          0.290000     0.310000        3.000000     0.047000            29.000000            118.000000     0.994890     3.210000     0.510000    10.300000     6.000000
75%         7.700000          0.400000     0.390000        8.100000     0.065000            41.000000            156.000000     0.996990     3.320000     0.600000    11.300000     6.000000
max        15.900000          1.580000     1.660000       65.800000     0.611000           289.000000            440.000000     1.038980     4.010000     2.000000    14.900000     9.000000
```

Figure 2: Statistic information

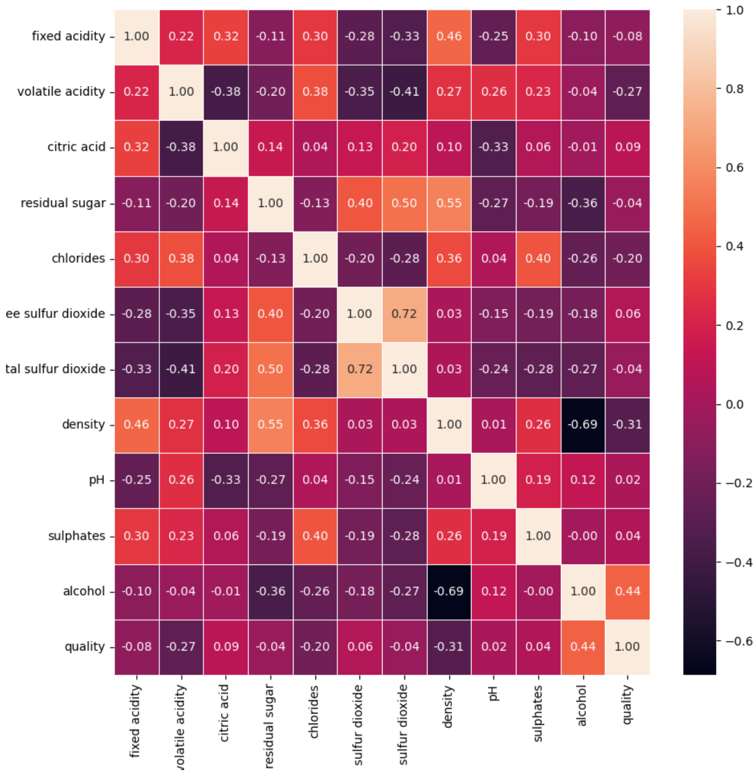


Figure 3: Correlation between the attributes

Algorithm	accuracy
Naive bayes Gaussian	0.440
K nearest Neighbour	0.522
XGB classifier	0.548
SVM rbf	0.555
Adaboost classifier	0.577

Table 1: Cross Validation Score

We are preprocessing the data with the predefined function fit_transform from sklearn library. It is normalizing with formula:

$$\frac{x-u}{s}$$

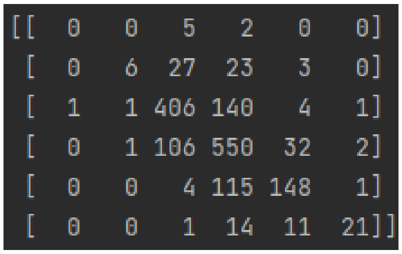
where x is each attribute and u is the mean of the attribute and s is the standard deviation of the attribute.

4 Algorithm

In this project, I implemented several classic models for the multiclass classification problem including: support vector machine, XGBoost(gradient boosted tree), Naive Bayes Gaussian, K Nearest Neighbour, Decision tree classifier, Adaboost classifier and Logistic Regression. The original paper introduced three models: support vector machine, neural network and the multiple regression. In their comparison, SVM achieves highest score. In my implementation, Adaboost achieves higher score than SVC.

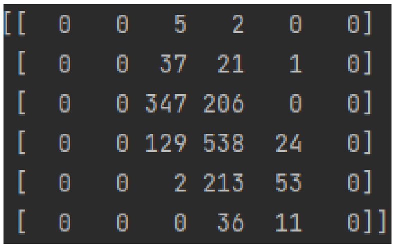
5 Experiments

The score of the cross validation is shown in tabular 1. The SVC achieves best performance with rbf kernel. Among all the models, Adaboost classifier achieves best performance.



```
[[  0   0   5   2   0   0]
 [  0   6  27  23   3   0]
 [  1   1 406 140   4   1]
 [  0   1 106 550  32   2]
 [  0   0   4 115 148   1]
 [  0   0   1  14  11  21]]
```

Figure 4: Adaboost classifier confusion matrix



```
[[  0   0   5   2   0   0]
 [  0   0  37  21   1   0]
 [  0   0 347 206   0   0]
 [  0   0 129 538  24   0]
 [  0   0   2 213  53   0]
 [  0   0   0  36  11   0]]
```

Figure 5: SVM confusion matrix

As we can see in the confusion matrix of SVM5, The model have low accuracy for the outer quality wines. This phenomenon is also noted in paper[Cor+09]. Compared to SVM, Adaboost4 fit the distribution better. It has a relatively high detecting rate for the outer wine quality.

6 Conclusion

With the development of the modeling algorithms, a better algorithm Adaboost is shown to better fit the true modeling of the wines. Also, XGBoost also shows a competitive performance compared to an old technique SVM.

7 Code

It can be found at: https://github.com/gcj13/ECE595IDMproject

References

[Cor+09] Paulo Cortez et al. "Modeling wine preferences by data mining from physicochemical properties". In: *Decision Support Systems* 47.4 (2009). Smart Business Networks: Concepts and Empirical Evidence, pp. 547–553. ISSN: 0167-9236. DOI: https://doi.org/10.1016/j.dss.2009.05.016. URL: https://www.sciencedirect.com/science/article/pii/S0167923609001377.