**Project15**

**Topic:** Wine Quality

**Group members:**

CHEN Tianlu 22098248d

GUO JIe Nan 22099604d

QU Jiaying 22101681d

WU Xinyi 22097723d

**Project Report：**

It is well known that many factors, such as fixed acidity, chlorides, free sulfur dioxide, etc., influence wine quality. Our group project aims to develop a classification model to classify the goodness of wine by these variables.

This time we had a total of 11 variables, and our assumption is to identify two groups of variables from the 11 groups that have the most influence on the wine quality. We took a total of two methods: a scatter plot to observe the separation as a traditional method and a decision tree method used as a complementary method.

There were 1599 data for each variable. First, we divided the data into train data and test data in a ratio of seven to three. Secondly, there are 55 cases based on choosing any two groups of 11 variables. So our group made 55 scatter plots and calculated all the accuracy. The graph is shown below.

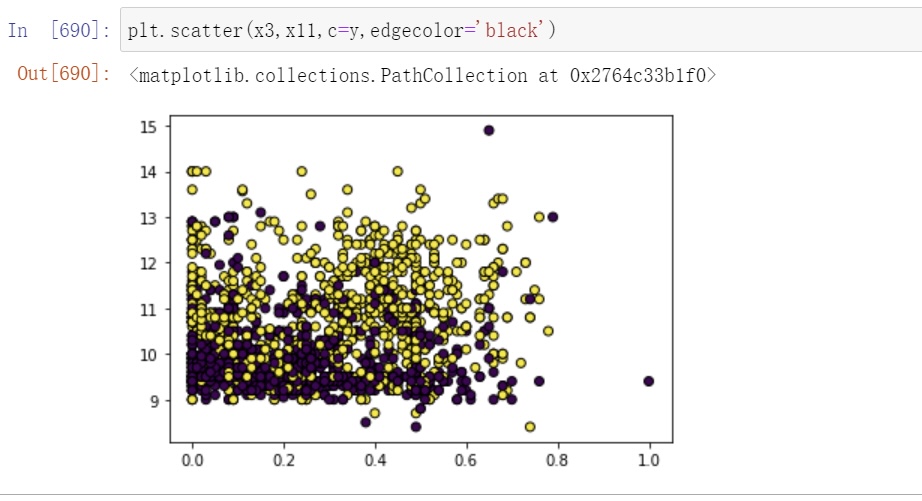
|  |  |
| --- | --- |
| variables | Accuracy |
| X1&X2 | 0.52 |
| X1&X3 | 0.52 |
| X1&X4 | 0.46 |
| X1&X5 | 0.50 |
| X1&X6 | 0.46 |
| X1&X7 | 0.54 |
| X1&X8 | 0.46 |
| X1&X9 | 0.51 |
| X1&X10 | 0.54 |
| X1&X11 | 0.54 |
| X2&X3 | 0.46 |
| X2&X4 | 0.46 |
| X2&X5 | 0.46 |
| X2&X6 | 0.56 |
| X2&X7 | 0.54 |
| X2&X8 | 0.47 |
| X2&X9 | 0.46 |
| X2&X10 | 0.61 |
| X2&X11 | 0.54 |
| X3&X4 | 0.54 |
| X3&X5 | 0.49 |
| X3&X6 | 0.49 |
| X3&X7 | 0.54 |
| X3&X8 | 0.50 |
| X3&X9 | 0.54 |
| X3&X10 | 0.60 |
| X3&X11 | 0.65 |
| X4&X5 | 0.52 |
| X4&X6 | 0.54 |
| X4&X7 | 0.54 |
| X4&X8 | 0.47 |
| X4&X9 | 0.54 |
| X4&X10 | 0.55 |
| X4&X11 | 0.53 |
| X5&X6 | 0.53 |
| X5&X7 | 0.54 |
| X5&X8 | 0.46 |
| X5&X9 | 0.46 |
| X5&X10 | 0.59 |
| X5&X11 | 0.48 |
| X6&X7 | 0.56 |
| X6&X8 | 0.46 |
| X6&X9 | 0.54 |
| X6&X10 | 0.54 |
| X6&X11 | 0.62 |
| X7&X8 | 0.54 |
| X7&X9 | 0.54 |
| X7&X10 | 0.54 |
| X7&X11 | 0.60 |
| X8&X9 | 0.46 |
| X8&X10 | 0.60 |
| X8&X11 | 0.54 |
| X9&X10 | 0.56 |
| X9&X11 | 0.61 |
| X10&X11 | 0.54 |

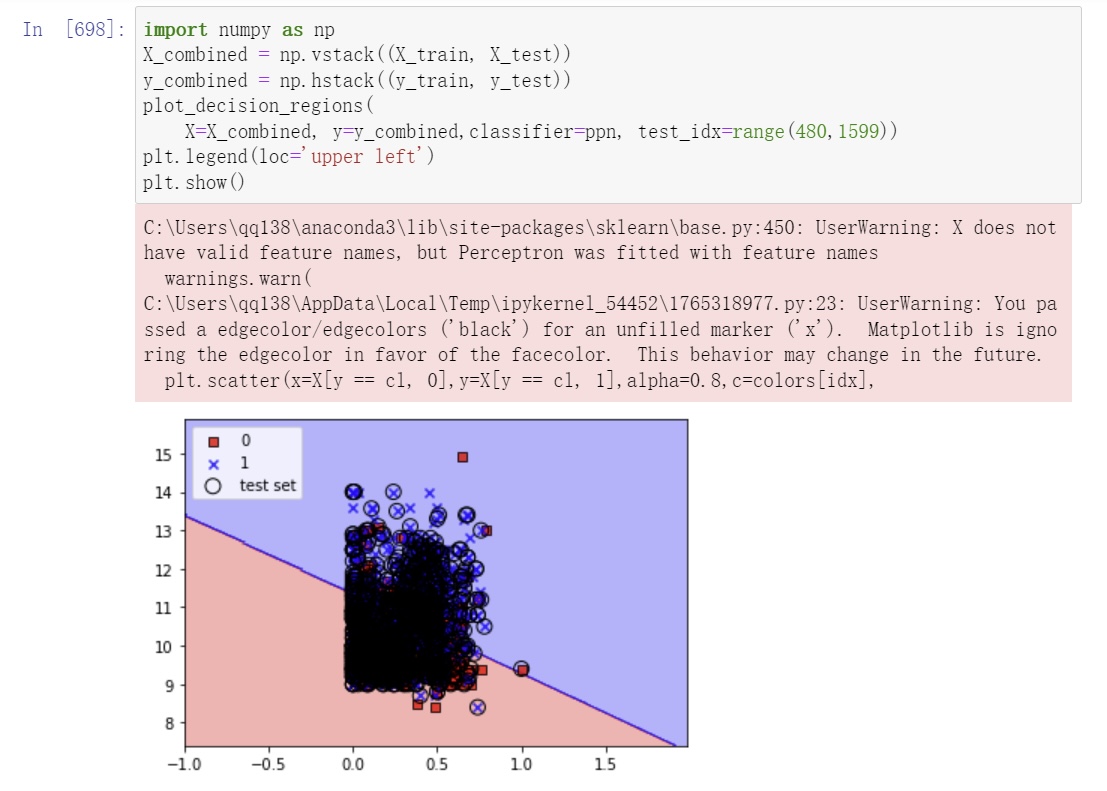
The two groups with the highest accuracy are the 3X and the 11X.

To exclude exceptional cases, we also selected multiple groups of data for testing, and the results are shown below.

|  |  |
| --- | --- |
| variables | Accuracy |
| X3&X10&X11 | 0.54 |
| X2&X3&X10&X11 | 0.55 |

It is easy to see that the accuracy of multiple data sets is low.

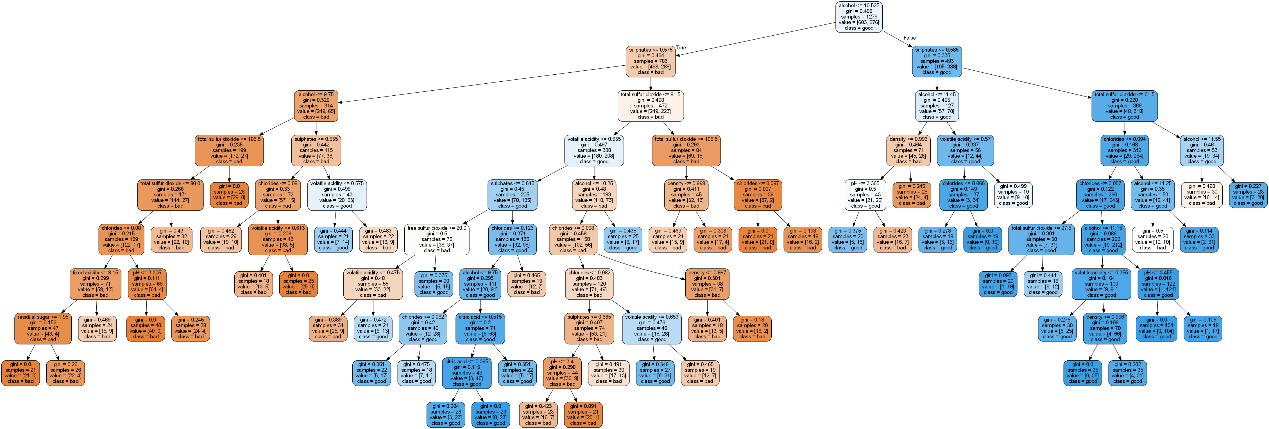




Finally, we did the logarithmic function to verify our conjecture. Two groups of variables had the most significant impact on wine quality, and we considered these two data sets, X3 and X11.

The model we use is a scatter plot classification model. We first draw a scatter plot using points of different colors as labels and observe the separation. Then the best quality solution is selected by evaluating the accuracy. Since our highest accuracy is limited to 0.65, our solution has some shortcomings. When we learn more, we will improve our scheme.

In conclusion, from the scatter plot approach, the two factors affecting wine quality are citric acid and alcohol.

According to Hajjej (2022), in artificial intelligence, **decision trees** are a classification model known for their visual aid in making**decisions.** Therefore, in addition, we used the decision tree to think about the problem from another perspective. For the decision tree, we measured the accuracy as the Gini index.

The decision tree gives us a more accurate and intuitive view of the data and is easier to understand.  We will be able to see which algorithms are the most efficient and appropriate to use when delving into this data and arrive at an informed **decision (Hajjej,2022).** Based on what the decision tree shows, we can follow the guidelines and classify the wines as good or bad.

**Reference**

Hajjej (2022). A Comparison of Decision Tree Algorithms in the Assessment of Biomedical Data. *BioMed Research International*., 1–9. https://www.hindawi.com/journals/bmri/2022/9449497/