7820 Final Project Report

Selected Topic	Topic 2: Visual Analytics on Red Wine Quality	
Group Membership	**Student ID**	**Name**
Member 1	**22407987**	**Zhang Chi**
Member 2	**22427481**	**Qiu Chen**

Visual Analytics on Red Wine Quality

I. Introduction

Wine has been a popular beverage of mankind for thousands of years. Our natural love for this drink stems from its wonderful taste, nutritional properties, and its psychoactive (intoxicating) effects. So, in this industry, the quality is really important part for both consumers and manufactors. In the past, evaluating the quality could happened after production, and it is a time-consuming process that requires experts to give an assessment, making the process very expensive. Furthermore, ‘there are a thousand Hamlets in a thousand people's eyes’, so as to say that the price of red wine depends on the rather abstract concept of tasters' appreciation of the wine, and opinions among them can be highly variable, which means this mode is not a fair but challenging task. However, In modern times, the wine industry has grown exponentially of late with the rise of social drinking. Industry players are using product quality certification to promote their products and also help them earn more market share.

Therefore, based on the background above, we believe the wine market will gain traction if human tasting qualities can be linked to the wine's chemical properties, thereby making the certification, quality assessment and assurance process more controllable. But not all parameters are suitable for the evaluation. So we still need to identify important parameters

such as acidity, pH, sugar content which can be evaluated using existing technologies. So this project aims to identify which characteristics are the best indicators of wine quality and gain insight into each factor that affects the quality of wine in our model while giving appropriate predictors.

The report is divided into 6 sections, including this one. And in Section 2, our group will related work from other researchers, and we will formulate our research questions. In the section 3, we describe the methodologies and their advantages for sovling the task. Section 4 we will do data-preprocessing and describe the dataset. In Section 5, we show the the results and discussion of the whole work. In Section 6, we discuss the conclusions and future work.

II. Related Work

Lee et al., (2015) has proposed a method about decision tree-based to predict the wine quality and compare approachs using some ML algorithms such as support vector machine, multi-layer perceptron, and BayesNet. As a result, they found the method is better compared to other stated methods.

Er, and Atasoy, (2016) has proposed the method for classifing the quality of the red wine and white wine by using three machine learning algorithms such as k-nearest-neighborhood, random forest, and support vector machine. They finally concluded that they have achieved the best result using the random forest algorithm by using principal component analysis to select the feature.

Véronique Gomes et al. (2021) has predicted the wine quality based on comparing four different machine learning methods (including deep learning), assessing their generalization capacity for different vintages and varieties not included in the training process. Ridge regression, partial least squares, neural networks and convolutional neural networks were the methods considered to conduct this comparison. The results show that the estimated models can successfully predict the sugar content from hyperspectral data, with the convolutional neural network outperforming the other methods.

Piyush Bhardwaj et al. (2022) have used prediction of wine quality using Seven machine learning algorithms for the prediction, they used Adaptive Boosting (AdaBoost) classifier, XGB, Random Forest (RF) classifier and so on. According to the results, AdaBoost predicted wine quality with higher accuracy(100%) during without feature selection, with feature selection (XGB) and with essential variables. In the presence of essential variables, the Random Forest (RF)

classifier performance was increased. Overall, performance of all classifiers (except KNN) improved when model trained and tested using essential variables. The usefulness of data generation algorithms and importance of feature selection is the key feature in this study.

Shengnan Di et al. (2023) identify that these early methods ignore the inner relationship between the physical and chemical properties of the wine, and then they has proposed a combined methods to conduct the Pearson correlation analysis, PCA analysis, and Shapiro-Wilk test on those properties and incorporates 1D-CNN architecture to capture the correlations among neighboring features. The result indicated that 1DCNN has a significant improvement where the mean value of Precision, Accuracy, Recall and F1 Score have increased at least 4%, 2.7%, 2.5%, and 4.1%, respectively. Besides, the DNN-based models show a better overall performance on red wine quality prediction than traditional machine learning models such as kNN, SVM, LR, and RF.

Based on the research, we also concluded six questions that we want to solve in this project:

1. Which variable(s) is (are) the most influential one(s) to the quality of the red/white wine?
2. Are there any correlations between different variables?
3. Are the influential variables of the red wine the same as those of the white wine?
4. Are these given variables enough for making prediction/evaluation on the quality of the red/white wine? If yes, state your reason; if not, what extra attributes/features of the wine do we need to evaluate its quality?
5. Which methodology(s) is(are) the most proper to make predicition/evaluation on the quality of the red/white wine?
6. Whether different variables of the data set used for prediction will have an impact on the prediction results?

III. Methodology

1. Logistic Regression

The logistic model (or logit model) is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables.

Advantages：

Simplicity of implementation and wide application to industrial problems.

Very low computational effort for classification, high speed and low storage resources.

Convenient observation of sample probability scores.

2. KNN

The k-nearest neighbors algorithm (k-NN) is a non-parametric supervised learning method, it's used for classification and regression.

Advantages:

Theoretical maturity and simplicity of thought, which can be used for both classification and regression.

Can be used for nonlinear classification.

Training time complexity of O(n).

no assumptions on the data, high accuracy and insensitivity to outlier.

3. Random Forest

It is an integrated learning method based on Bagging, which can handle classification and regression very well.

Advantages:

Decision trees can be generated by different hosts with high efficiency.

Random forests inherit a bit of CART.

Combined in the form of bagging to avoid overfitting.

IV. Data Description and Preprocessing

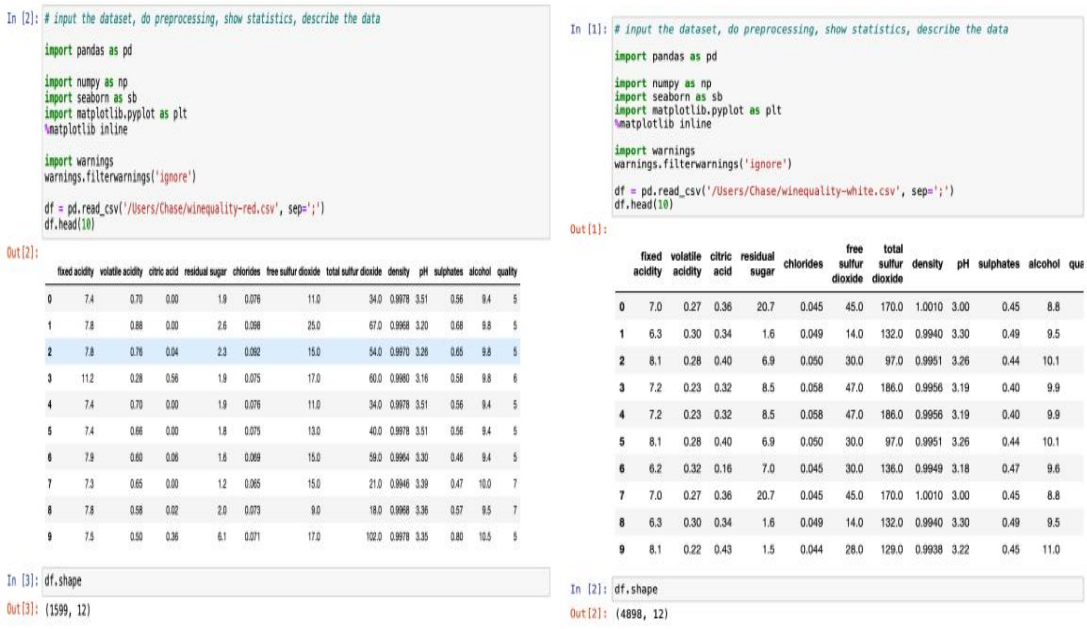


Figure 1-1 & Figure 1-2

The Figure 1-1 and Figure 1-2 shows that there are 11 independent attributes from two dataset:

1. **Alcohol:** the amount of alcohol in wine.
2. **Volatile acidity:** acetic acid content which leading to an unpleasant vinegar taste.
3. **Sulphates:** a wine additive that contributes to SO2 levels and acts as an antimicrobial and antioxidant.
4. **Citric Acid:** acts as a preservative to increase acidity.
5. **Total Sulfur Dioxide:** is the amount of SO2.
6. **Density:** sweeter wines have a higher density.
7. **Chlorides:** the amount of salt.
8. **Fixed acidity:** are non-volatile acids that do not evaporate easily.
9. **pH:** the level of acidity.
10. **Free Sulfur Dioxide:** it prevents microbial growth and the oxidation of wine.
11. **Residual sugar:** it is the amount of sugar remaining after fermentation stops.

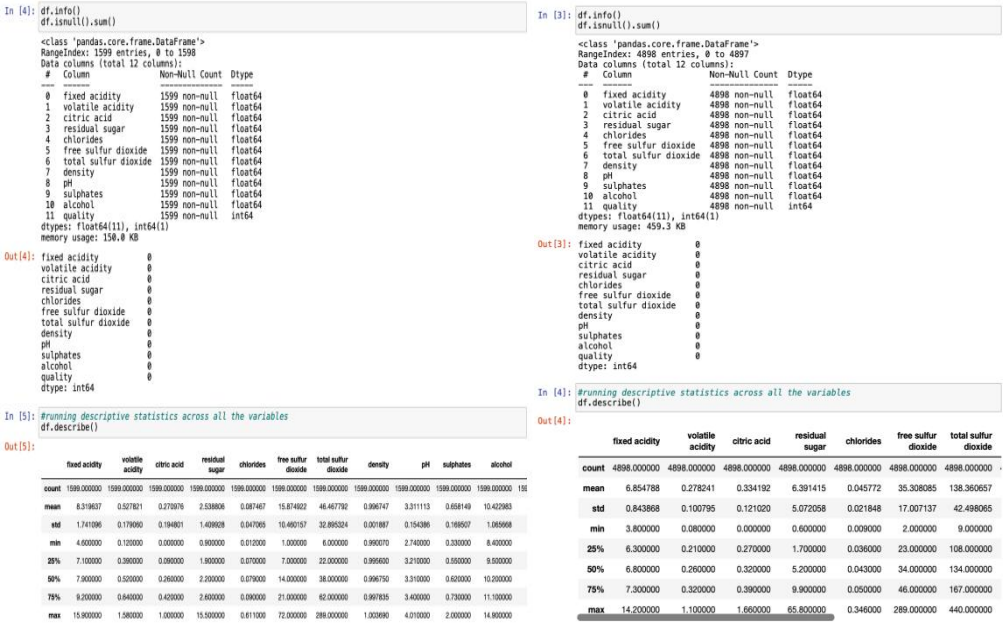


Figure 2-1 & Figure 2-2

From the Figure 2-1 and Figure 2-2 show that there are 1599 and 4898 records in red and white wine dataset, and it does not have any null cell in each attributes, so that we do not need to address missing value. Besides, those 11 independent attributes type are float64, and the only dependent attribute, quality, it type is int64. The bottom part of Figure 2-1 and Figure

2-2 display the Statistics value, we can find some classical and useful value such as count, mean, std, min, max and so on.

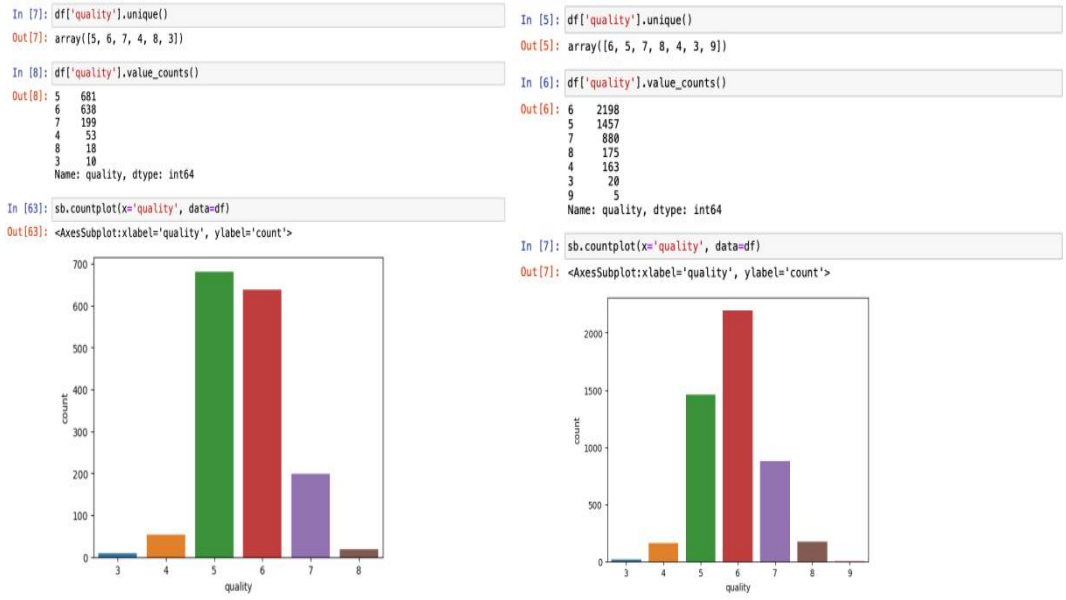


Figure 3-1 & Figure 3-2

As for the ‘quality’, we can know that its range from 3 to 8, and the number of level 5 and level 6 take the most two share, which has 681 and 638, respectively. There is only 18 data results belong to the best level 8.

Although the quality ranges from 0 to 10. However, for this dataset, it is in the range of 3 to 8. The middle classes have higher counts. Therefore the entire model will be biased toward these three classes. Since the data are imbalanced through the classes, we may need to perform class-balancing after splitting the data.

From the Figure 3-2, we can find that the white wine dataset has a similar situation with the red wine one. But totally the quality of white wine is better than that of red. Besides, There are 5 samples belong to level 9, which it does not show in the red wine dataset.

Data Preprocessing:

First, we use boxplot of the attributes to check the outliers, the result showed in Figure 4.

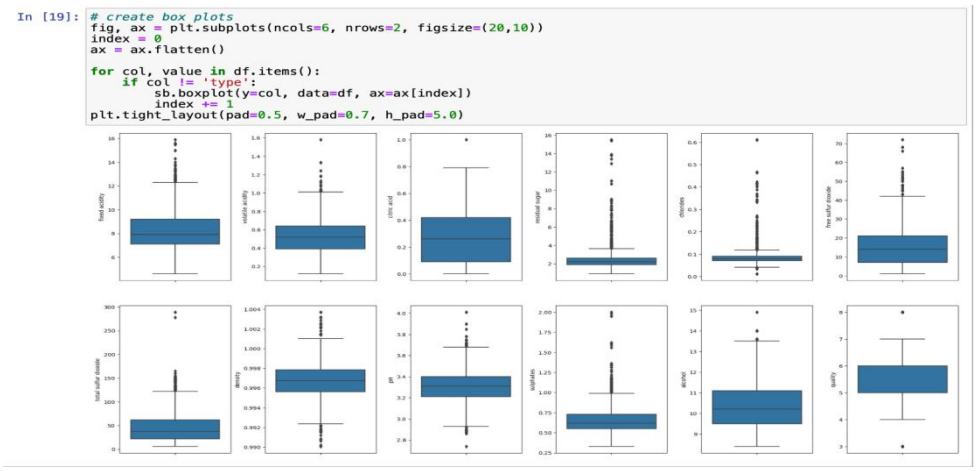


Figure 4

After visualization, outliers can be find out from these box plots. Although eliminating these outliers will improve the accuracy of the model, we will ignore this outlier since it won't affect the outcome of the project.

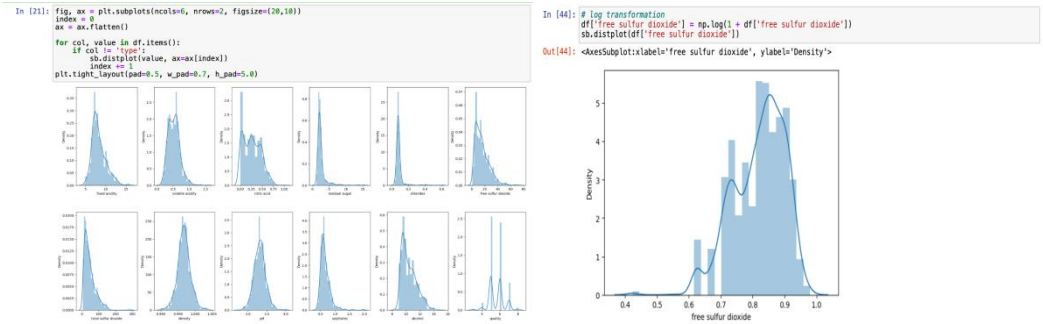


Figure 5-1 & Figure 5-2

From the Figure 5-1, It can be seen that most red wines' pH levels are always between 3 – 4 and chlorides, the amount of salt is most prevalent at level 0.1. However, the column 'Free sulfur dioxide' is slightly right-skewed. So we can normalize it using log transformation, and other skewed plot can also do in the same way(Figure 5-2). After transformation, it can be observed taht a normal distribution in a form of a bell curve. And we can use the same method in the white wine dataset.

V. Experimental Results

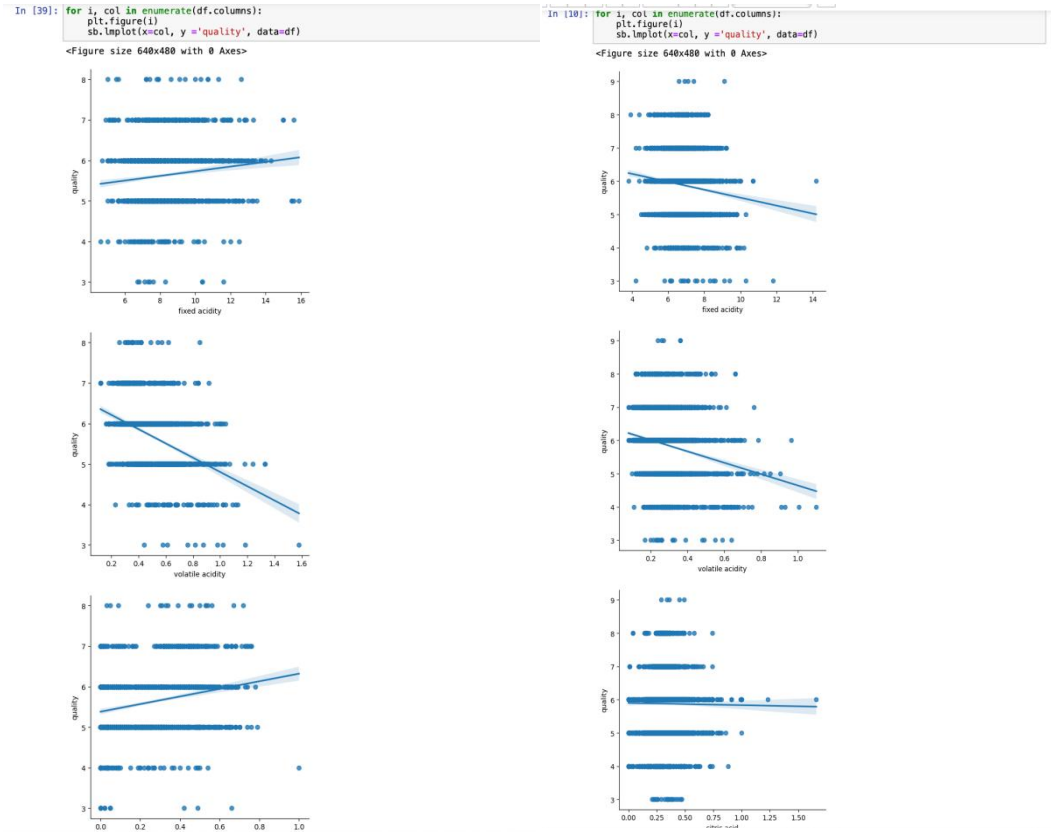


Figure 7-1 & Figure 7-2

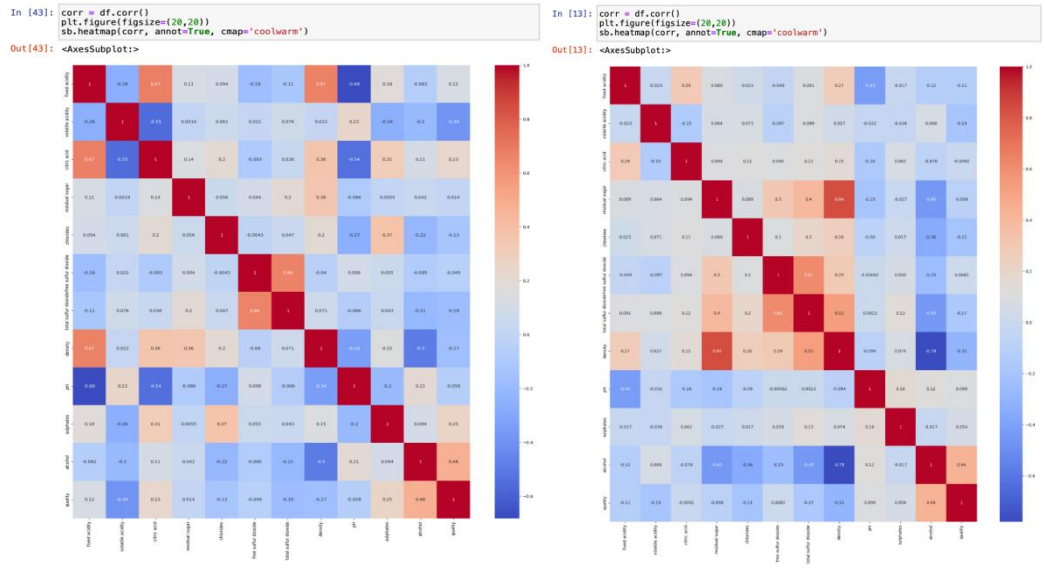


Figure 8-1 & Figure 8-2

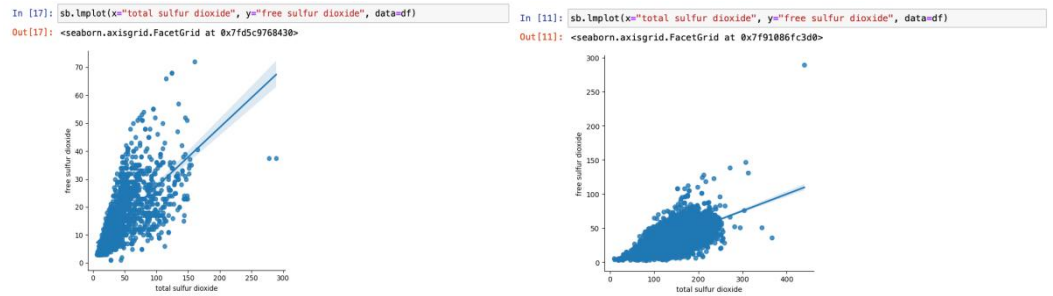


Figure 9-1 & Figure 9-2

```
In [41]: plt.subplots(figsize=(20,15))
         ax = plt.axes()
         ax.set_title("Correlation Differences between Red and White Wines")
         corr = diff_corr
         sns.heatmap(corr, annot=True,
                     xticklabels=corr.columns.values,
                     yticklabels=corr.columns.values,
                    cmap="coolwarm")
Out[41]: <AxesSubplot:title={'center':'Correlation Differences between Red and White Wines'}>
```

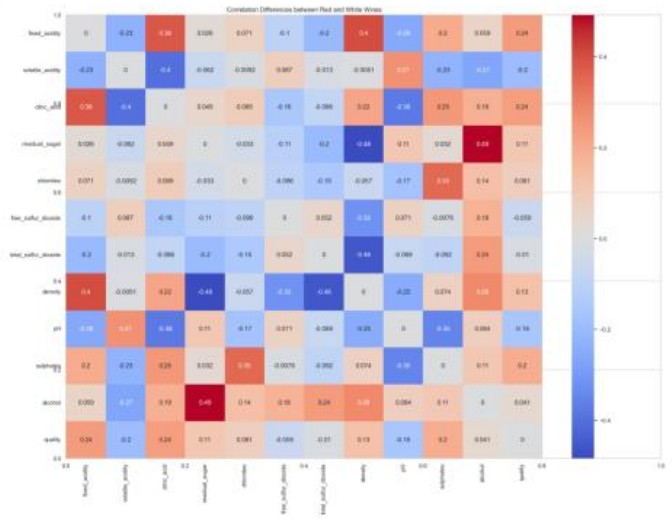


Figure 10

```
!pip install -U scikit-learn
```

Figure 11

```
from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.20)
```

Figure 12

```
from sklearn.preprocessing import StandardScaler

sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)
```

Figure 13

```
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier
```

Figure 14

```
from sklearn.model_selection import GridSearchCV
```

Figure 15

Logistic Regression

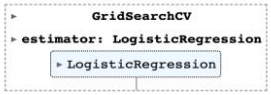
```
from sklearn.model_selection import GridSearchCV
```

```
method = 'Logistic Regression'

lr = LogisticRegression()

lr_params = {"C": [0.5, 0.75, 1, 1.5, 1.5, 2]}

lr_cv_model = GridSearchCV(lr, lr_params, cv = 10)
lr_cv_model.fit(x_train, y_train)
```




```
print("optimal C value: " + str(lr_cv_model.best_params_["C"]))
```

```
optimal C value: 1.5
```

```
lr_model = LogisticRegression(C = 1.5)
lr_model.fit(x_train, y_train)
```


```
from sklearn.metrics import accuracy_score
```

```
y_pred = lr_model.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)

result_list = store_result(result_list, method, accuracy)
```

Figure 16

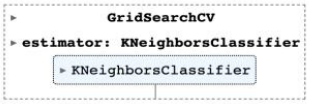
KNN

```
method = 'KNN'

knn = KNeighborsClassifier()
```

```
knn_params = {"n_neighbors": np.arange(2, 50),
              "weights": ["uniform", "distance"],
              "p": [1, 2]}

knn_cv_model = GridSearchCV(knn, knn_params, cv = 10)
knn_cv_model.fit(x_train, y_train)
```

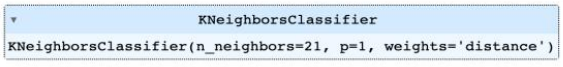



```
print("best K value: " + str(knn_cv_model.best_params_["n_neighbors"]),
      "\nbest weights: " + knn_cv_model.best_params_["weights"],
      "\nbest value of p: " + str(knn_cv_model.best_params_["p"]))
```

```
best K value: 21
best weights: distance
best value of p: 1
```

```
knn_model = KNeighborsClassifier(n_neighbors = knn_cv_model.best_params_["n_neighbors"],
                                 weights = knn_cv_model.best_params_["weights"],
                                 p = knn_cv_model.best_params_["p"],
                                 )

knn_model.fit(x_train, y_train)
```




```
y_pred = knn_model.predict(x_test)
```

Figure 17

Random Forest

```
method = 'Random Forest'

rf = RandomForestClassifier()
```

```
rf_params = {
              "n_estimators": [100, 150, 200],
              "max_depth": [2, 3, 4, 5],
              "min_samples_split": [2, 3, 4, 5]
            }

rf_cv_model = GridSearchCV(rf, rf_params, cv = 10, n_jobs = -1)
rf_cv_model.fit(x_train, y_train)
```

GridSearchCV
estimator: RandomForestClassifier
RandomForestClassifier

```
print("\nbest n_estimators: " + str(rf_cv_model.best_params_["n_estimators"]),
      "\nbest max_depth: " + str(rf_cv_model.best_params_["max_depth"]),
      "\nbest min_samples_split: " + str(rf_cv_model.best_params_["min_samples_split"]))
```

```
best n_estimators: 100
best max_depth: 5
best min_samples_split: 2
```

```
rf = RandomForestClassifier(
                                max_depth = rf_cv_model.best_params_["max_depth"],
                                n_estimators = rf_cv_model.best_params_["n_estimators"],
                                min_samples_split = rf_cv_model.best_params_["min_samples_split"])
rf_model = rf.fit(x_train, y_train)
```

```
y_pred = rf_model.predict(x_test)
accuracy = accuracy_score(y_test, y_pred)
```

Figure 18

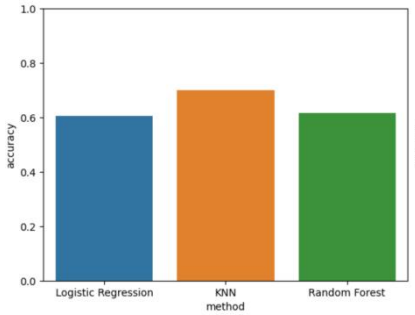


Figure 19

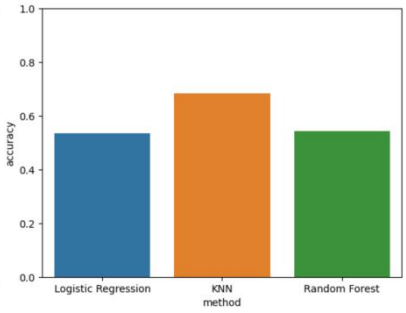


Figure 20

	method	accuracy
0	Logistic Regression	0.546875
1	KNN	0.684375
2	Random Forest	0.606250

Figure 21

	method	accuracy
0	Logistic Regression	0.596875
1	KNN	0.684375
2	Random Forest	0.615625

Figure 22

1. Which variable(s) is (are) the most influential one(s) to the quality of the red/white wine?

From the Figure 7-1, we can roughly identify the relationship between quality and other 11 variables. To be specific:

In the red wine dataset, ‘fixed acidity’, ‘citric acid’, ‘sulphates’, ‘alcohol’, ‘residual sugar’, all these 5 variables show a positive correlation with 'quality', and the remaining variables all have a negative correlation with ‘quality’. Besides, we can generate a correlation matrix(Figure 8-1), which is a table showing correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value is in the range of -1 to 1. If two variables have a

high correlation, we can neglect one variable from those two.

Therefore, we can know that ‘alcohol’ is the most positive influential variable to the quality of the red wine and the ‘volatile acidity’ is the one has the most negative influential variable.

Similarly, we can draw another conclusion about the white wine from the Figure 8-2 that ‘alcohol’ is the most positive influential variable to the quality of the red wine and the ‘density’ is the one has the most negative influential variable.

2. Are there any correlations between different variables?

From the Figure 8-1, we identify that ‘fixed acidity’ has a highly positive correlation with ‘citric acid’(0.67) and ‘density’(0.67). And ‘free sulfur dioxide’ also has a high positive correlation with the ‘total sulfur dioxid’(0.66). On the other hand, ‘fixed acidity’ has a high negative correlation with ‘pH’(-0.68). And there also have 3 high negative set: {volatile acidity, citric acid(-0.55)},{pH, citric acid(-0.54)},{alcohol, density(-0.50)}.

Similarly, from the Figure 8-2, we identify that ‘residual sugar’ has a very highly positive correlation with ‘density’(0.84). On the other hand, ‘density’ also has a high negative correlation with ‘alcohol’(-0.78).

3. Are the influential variables of the red wine the same as those of the white wine?

We can draw a conclusion that the positive influential variables of the red wine are as same as those of the white wine - (alcohol). However, the most negative influential variables are totally different, which ‘volatile acidity’ for red wine and ‘density’ for white wine.

To be specific, we use Pearson's Correlation to do the correlation diffenence of two dataset. From the Figure 10, There are some obvious differences that certain variables interact depending on the variety of wine. The darker the square, the larger the difference that interaction is between Red and White wines. For example, the correlation between ‘alcohol’ and ‘residual sugar’ content is much higher for Red wines than it is for white wines(showed darked with the number ‘0.49’ in the Figure 10). And closer inspection indicates that the correlation between ‘residual sugar’ and ‘alcohol’ is positive for Red wines (weak positive, 0.042), but it is much more strongly negative for White wines (-0.45).

4. Are these given variables enough for making prediction/evaluation on the quality of the red/white wine? If yes, state your reason; if not, what extra attributes/features of the wine do we need to evaluate its quality?

In fact, in the above prediction/evaluation of red wine quality by variables such as acidity, density, pH, alcohol, etc., we only analyzed the quality through the essential information on the substance of red wine, i.e., the analysis based on the objective factors of red wine itself. The evaluation system is widely applicable and reasonable only for the general evaluation of red wine. However, from a more complex and comprehensive perspective, when thinking about the analysis, we believe that, in addition to the analysis of the objective factors of the wine itself, it is also necessary to include the objective and subjective factors other than the wine itself, including the production process, production environment (such as the tools of production), color, taste, etc. of the type of wine. In the analysis of quality issues (i.e. wine quality prediction/analysis) combining subjective and objective factors, increasing the complexity and comprehensiveness of the analysis dimension, the results obtained, to a certain extent, will also improve its accuracy and usability.

5. Which methodology(s) is(are) the most proper to make predicition/evaluation on the quality of the red/white wine?

In this project, we used different algorithms to predict and evaluate the quality of white/red wine, and compared and analyzed the accuracy of the results produced under the different methods using a unified approach.

We imported scikit-learn as a tool library for performing operations such as prediction on datasets. Scikit-learn is a library in Python that provides many unsupervised and supervised learning algorithms, it's a simple and efficient tools for predictive data analysis. (Figure 11)

In order to make prediction and evaluation, we need to split the dataset first. In this project, we use 80-20 split of training and Hold-Out Data(splitting dataset into training(80%) and testing(20%)). Then, we should make a standardization to avoid that two different scales data may end up affecting our model differently.(Figure 12&13)

In our project, we use LR, KNN and RF to predict the red/white wine quality and find out which is the best method getting the highest accuracy score among them. (Figure 14)

We use GridSearchCV to helps us in finding the optimal values of the parameters for our model. It will end up getting the best accuracy scores from these 3 methods.(Figure 15&16&17&18)

Finally, we obtained the computational results(accuracy) obtained using the different methods. By observing the results, it is clear that KNN is the most accurate method among the

three methods, followed by RF and LR. Figure 19 and Figure 20 show the predicted results for red and white wine respectively (their results are similar).

6. Whether different variables of the data set used for prediction will have an impact on the prediction results?

In the first of the six problems we listed in this project, we concluded that there are some variables that have a positive or negative correlation with the quality of red wine. In this problem, we set the data of positively correlated variables as one set and the data of negatively correlated variables as another set, as different data sets used to predict the quality of red/white wine, and the corresponding results obtained were compared to draw the results of the discussion about this problem.

Taking red wine as an example, we can learn from the data shown in Figure 21 and Figure 22 that the results of the two sets of data are basically similar. Obviously we can conclude that using different variables to predict the quality of red wine does not have an impact on its prediction results.

VI. Future Work and Discussion

In morden society, quality forecasting of products is very important. There have been many studies on the application of artificial intelligence in the field of wine quality identification.

In this project, we have successively asked, discussed, and tried to solve the six problems presented. These include: finding the variables that most affect the quality of red/white wine - 'alcohol' (positively correlated with quality) and 'volatile acidity'. 'density' (negative correlation with quality, corresponding to red wine and white wine, respectively). Correlation problems between different variables were found, such as: the positive correlation between 'aitric' and 'density' with the positive correlation between 'fixed acidity' and negative correlation between 'pH', etc. The Pearson's Correlation was used to prove that the variables affecting red and white wine are independent of each other. More possible factors affecting the quality of red/white wine are proposed. Different machine learning algorithms were used to train and compare the prediction of quality for the same dataset (in this project, i.e., red wine/white wine), and the most appropriate algorithm to use in this project, KNN, was derived. Experimental control groups were divided by the effect of positive and negative correlations between variables and red wine/white wine quality, and it was demonstrated that in this project, different groups of variables had different effects on

There is essentially no effect on red/white wine quality prediction.

However, exploring the quality of wine is complex due to the sophisticated correlations within its attributes. It is necessary for a model to consider the global relationship between those features and their interactions.

In the future, it may have some better algorithms can be developed which involves the combination of best features of all other data mining techniques so that we may explore more interaction between the features. And from the perspective of improving the accuracy, it is clear that the algorithm or the data must be adjusted. We rsuggested other researchers can use potential relationships between wine quality, or applying the boosting algorithm on the more accurate method.

VII. Author Contributions

Zhang Chi and Qiu Chen contributed equally. Both authors conceived the project and participated in drafting the report. Zhang Chi finished the Introduction, Related Work, Data Description and Data Preprocessing, and the first three questions of the Experimental Results part, Qiu Chen Methodology, Future Work and Discussion, last three questions of the Experimental Results part. Both authors read and approved the final report. Besides, both authors finished Author Contributions and References together.

VIII. References

[1] Zhang, H., Wang, Z., He, J., & Tong, J. (2021). Construction of Wine Quality Prediction Model based on Machine Learning Algorithm. ACM International Conference Proceeding Series, 53–58. https://doi.org/10.1145/3480433.3480443

[2] Bednárová, A., Kranvogl, R., Brodnjak-vončina, D., & Jug, T. (2014). Prediction of wine sensorial quality by routinely measured chemical properties. Nova Biotechnologica et Chimica, 13(2), 182–196. https://doi.org/10.1515/nbec-2015-0008

[3] Sáiz-Abajo, M. J., González-Sáiz, J. M., & Pizarro, C. (2006). Prediction of organic acids and other quality parameters of wine vinegar by near-infrared spectroscopy. A feasibility study. Food Chemistry, 99(3), 615–621. https://doi.org/10.1016/j.foodchem.2005.08.006

[4] S Di, & Yang, Y. (2023). Prediction of Red Wine Quality Using One-dimensional

Convolutional Neural Networks. arXiv.org.

[5] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Using Data Mining for Wine Quality Assessment. Discovery Science, 5808, 66–79. https://doi.org/10.1007/978-3-642-04747-3_8

[6] Machine Learning; Study Findings from University of Piemonte Orientale Broaden Understanding of Machine Learning (Authenticity assessment and protection of high-quality Nebbiolo-based Italian wines through machine learning) (p. 467–). (2018). NewsRx.

[7] Diako, C. (2016). Influence of wine components on the chemical and sensory quality of wines. ProQuest Dissertations & Theses.

[8] Yogesh Gupta.(2018). Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science, Volume 125, 305-312. https://doi.org/10.1016/j.procs.2017.12.041.

[9] Koranga, M., Pandey, R., Joshi, M., & Kumar, M. (2021). Analysis of white wine using machine learning algorithms. Materials Today: Proceedings, 46, 11087-11093.

[10] Chiu, T. H. Y., Wu, C. W., & Chen, C. H. (2021). A Hybrid Wine Classification Model for Quality Prediction. In Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part IV (pp. 430-438). Springer International Publishing.

[11] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104095.