KerianTeam_TC3313_MiniProject

by liangttan 1

Submission date: 19-Jan-2022 02:46AM (UTC-0500)
Submission ID: 1744028930
File name: KerianTeam_TC3313_Mini_Project.docx (1.01M)
Word count: 5763
Character count: 31063

Wine Quality Prediction

Ng Hao Lin[1], Ng Xing Ning[2], Tan Jing Xuan[3]
Faculty of Information Science & Technology
The National University of Malaysia (UKM)
Bangi, Selangor, Malaysia
Email: a175838@siswa.ukm.edu.my[1], a176493@siswa.ukm.edu.my[2], a175711@siswa.ukm.edu.my[3]

***Abstract*—This Wine Quality is an important feature for every winemakers to focus on for a profitable market sales. It can literally be figure out with investigation on its concentrations of its chemical substances. To learn an ideal idea for the wine production, this research has been carried out to study the relationship between the chemical properties of experimental wine with its quality with Machine Learning algorithms. The dataset used for this research is Wine Quality Dataset which is obtained from the UCL Machine Learning Repository, a popular machine learning databases community. This dataset describes an analysis of 13 features found in different samples of experimental wines. At the beginning, we perform data visualization and figuring the correlations between each feature in the dataset. Then, Data Pre-processing task was conducted to handle the missing values and categorical columns. Outliers are detected and removed as well. After that, we also splitting our pre-processed dataset into training and testing sets in ratio of 70:30. This split datasets were then saved and exported in CSV file for backup purpose. This process is followed by data normalization before we perform our modeling. Among various machine learning algorithms, we select 3 machine learning models to predict this dataset, which are Random Forest, Support Vector Machine(SVM), and Logistic Regression. For a better analytical result, we summarize our results with performance measurement matrices in our reports. As a conclusion of this research, we achieved our result which is Random Forest as the best prediction algorithms among these three different machine learning models.**

***Keywords*—Classification, Random Forest, Support Vector Machine, Logistic Regression, Machine Learning**

I. INTRODUCTION

The quality of drinks generally becomes one of the most important factors for the consumer to consider when selecting and purchasing beverages. This is especially true when it comes to wine, a sort of drink which has long been considered as luxury beverages. Wine is a kind of alcoholic drink fermented from fresh fruits especially grapes. It has a long history which discovered thousand years ago and its popular in European countries such as England. Since wine is brewed from fresh fruits, it also consists of antioxidants which can protects our cells from free radicals and polyphenols which manages our health. These make wine the healthiest alcoholic drink, which surprisingly brings benefits to our body with moderate consumption [1] [2].

This project aims to investigates about the application of various machine learning models on Wine Quality Dataset. Wine Quality Dataset is a dataset that described about an analysis of different chemical features contains in numerous samples of wine, as well as their quality. It was published on UCL Machine Learning Repository and owned by Forina, M. et al, PARVUS, from the Institute of Pharmaceutical and Food Analysis and Technologies in Brigata Salerno, Italy [3]. This dataset shows the results of the experimental wines which are grown in same region of Italy but derived from three different cultivars. There are 13 chemical properties shown in this dataset, which are wine types, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol, and quality.

Commonly, wine with good quality always comes with good sales. No mans like to spending money on drinking terrible wines which spoiled their moods. This is the of the key reasons that wine companies always putting their best efforts to do research and discover the best recipes to produces good wine which come with high profit margin. Hence, it is quite necessary for the winemakers to conduct analytical research like machine learning or deep learning in order to fitting the demands of the market. A great wine recipes will definitely bring the wineries large profits and popularity, and also help them to plan great marketing strategy as well.

Machine learning is a study of Artificial Intelligence (AI) which allows the system to have the ability to learn experience from dataset and improve their predicting outcomes without explicitly coding. The techniques of machine learning are very important nowadays as it supports human to perform tasks in a variety of fields and industries. In order to analysis this wine dataset, three different machine learning models will be developed to perform the analysis task. The machine learning models are Random Forest, Support Vector Machine (SVM) and Logistic Regression. Among these three models, we will compare their performance by classified which models provide us with a better analytical result.

Through this research, we will perform various tasks around the Wine Quality Dataset. First and foremost, we will import the dataset into our notebook. Secondly, we will perform data visualization task to visualize and have a better understanding on our dataset. We will also figure out the correlations between the features and perform data pre-processing by handle the missing values and categorical columns in our dataset. The outliers in this dataset will also be figured out and will be removed if there is any. Next, we have split our data into two datasets: training and testing, which then are saved and exported as two separate csv file. It will be followed by normalization of data. Finally, we will put those three models to the test and determine their performance. For a better analytical result, we apply performance measurement matrices to our reports which

are precision, recall, f-1 score and accuracy. All these phases will be explained with more details in the related sections.

II. LITERATURE REVIEW

This literature review will show about relevant literature to our problems and also explaining their related works. Since the wine quality is such an intriguing and important issues which the consumers are concerned about, there are quite many researchers conducts research on the best wine recipes and which are the key features that affect it. One of the most commonly well-known wine quality experiment is this Wine Quality Dataset that we have talked about in the introduction.

First and foremost, we will have a look to the wine quality prediction research by Rohan Dilip Kothawade. The purpose of this research is to investigates about the Wine Quality Dataset with classification algorithms. They are 3 types of machine learning models been applied to it, which are Support Vector Machine (SVM), Naïve Bayes and Artificial Neural Network (ANN). This result ends out with ANN as the better result in classification problems for this dataset [4].

Another research that we have gone through about is conducted by Mayur Badole in 2021. This result only performs a machine learning models which is Random Forest Classifier. Random Forest Classifier is a straight forward and simple to use algorithms when it comes to classification task. However, if we only have one machine learning model, we cannot confirm that how well our outcome is [5]. So, it's not really recommended to carry out the research with solely one model.

The last research which we have looked into is the research carried out by K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire. This research carried out with the Red Wine Quality Dataset and 4 different machine learning model. Red Wine Quality Dataset is similar to our original dataset, but it only studies about the red wines. The 4 datasets which has performed are Ridge Regression, Support Vector Machine, Grading Boosting Regressor and Artificial Neural Network (ANN). It results come out with Gradient Booster as a better choice among other machine learning models [6].

III. RESEARCH METHODOLOGY

A. Data Preparation

The wine quality dataset used in this project is obtained from the UCL Machine Learning Repository, which contains an extensive collection of datasets that the machine learning community has widely used. The dataset that we use in our project combines two datasets, which are red and white Vinho Verde wine samples, respectively. According to the dataset's authors, only physicochemical variables and the sensory result are available due to privacy problems and logistic issues. Therefore, there is no detailed information about wine grapes varieties, wine's cost, wine selling price, and wine brands in the datasets. The dataset aims to model wine quality based on physicochemical tests [7]. The wine dataset is useful in building a valuable model for wine manufacture and consumers as it can be used by oenologists in wine evaluations, potentially increasing decision-making speed and quality. Apart from that, measuring the impact of different analytical data on wine's quality can enhance the production process and help in target marketing [7].

B. Data Pre-processing

Data pre-processing is the process of transforming raw data into data that is suitable for a machine learning model. Data in real world is frequently uncleaned and corrupted with noise, incomplete information, and missing value. Therefore, data pre-processing is required to ensure data quality as the quality of data directly affects the ability and performance of learning model. The dataset used in this project is structured data as it is organized in tabular format. It contains 6497 rows with 13 features. Among 13 features, 12 features are independent variables and a dependent variable. The first feature is the type of wine which included 4898 white wine samples and 1599 red wine samples. Each sample of both types of wine consists of 12 physiochemical variables: fixed acidity (g/dm^3), volatile acidity (g/dm^3), citric acid (g/dm^3), residual sugar (g/dm^3), chlorides (g/dm^3), free sulfur dioxide (mg/dm^3), total sulfur dioxide (mg/dm^3), density (g/cm^3), pH, sulphates (g/dm^3) and alcohol (vol%). The attributes and their data type are shown in the TABLE I below.

TABLE I. RAW DATA ATTRIBUTE

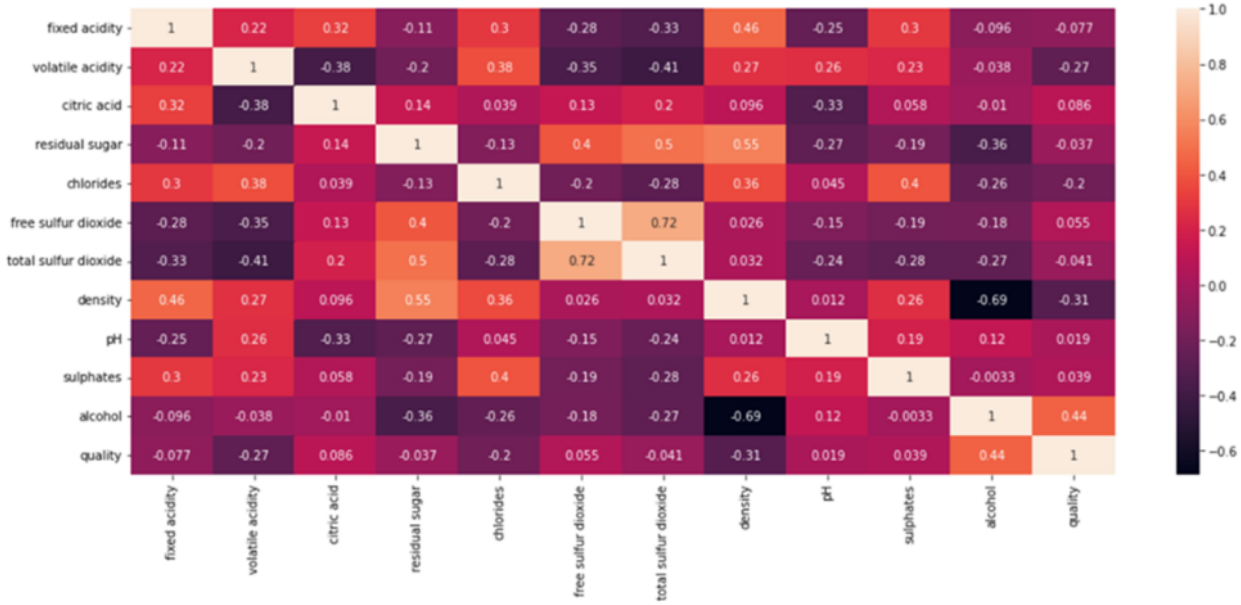
Attribute	**Data Type**
type	Nominal
fixed acidity	Ratio
volatile acidity	Ratio
citric acidity	Ratio
residual sugar	Ratio
chlorides	Ratio
free sulfur dioxide	Ratio
total sulfur dioxide	Ratio
density	Ratio
pH	Interval
sulphates	Ratio
alcohol	Ratio
quality	Ordinal

To enhance the computational performance and avoid overfitting problems, only the relevant features are selected for the model training process. The Pearson correlation is used in the feature selection process. All the independent variables should be uncorrelated with each other. For this reason, the variables which have correlation value greater than 0.7 should be kept only one and drop the rest. The correlation coefficient of features with each other is shown in figure 1 below. From the diagram, the feature "total sulfur dioxide" and "free sulfur dioxide" are highly correlated with each other with the coefficient of 0.72. Hence "total sulfur dioxide" is dropped and kept only "free sulfur dioxide" since its correlation with the target variable "quality" is higher than that of "total sulfur dioxide." (TABLE II).

TABLE II. HIGH CORRELATED VARIABLES

	Total sulfur dioxide	Free sulfur dioxide	Quality
Total sulfur dioxide	1	0.72	-0.041
Free sulfur dioxide	0.72	1	0.055
Quality	-0.041	0.055	1

Fig. 1. Correlation Heatmap



The next step is to handle missing values and categorical columns. The rows containing missing values are dropped since the number of missing values is considered less, only 0.5% (34 out of 6463) of the whole data. The categorical data need to be transformed into numerical form as the machine learning algorithms cannot operate on categorical data directly. Therefore, in the "type" column, the values "white" and "red" are converted to numerical form which is 1 and 0 respectively. This project aims to predict whether the wine quality is good or bad. Hence, to perform binary classification task, the target attribute "quality" is transformed into binary data by separating samples into two classes: good quality (label with 1) and bad quality (label with 0). The wine samples with quality greater than or equal to 7 are considered good quality, while those with quality lower than 7 are bad.

There are outliers in attribute "residual sugar" and "free sulfur oxide" which can be obviously observed from the statistical report (TABLE III). The maximum value of attribute "residual sugar" is 65.80, which is way too large compared to the 25th percentile (1.80), 75th percentile (8.10), and value of mean (5.44). For the attribute "free sulfur dioxide" as well, there is an apparent huge gap of range between mean and percentile with the maximum value. (Figure 2). Hence, the rows with outliers are removed.

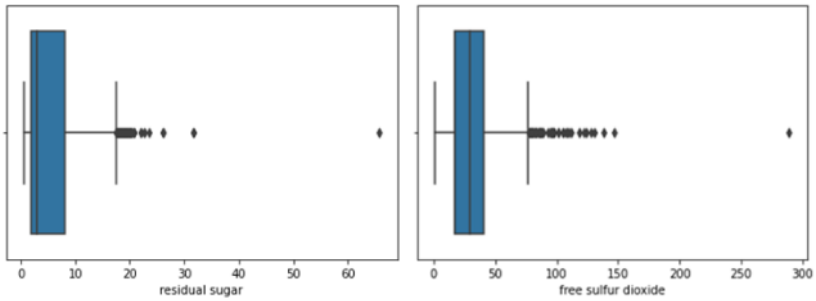
After that, min-max normalization is used for feature scaling. Feature scaling is essential in the training model process, especially for machine learning algorithms that use gradient descent as an optimization technique. For instance, logistic regression that we use in the following training step. The features on similar scale can efficiently improve the speed of the optimization process. Besides, the range of features can affect the performance of distance-based algorithms (such as SVM). The different scales may cause the model to be biased towards one feature [8]. Since each data range is very wide, the independent variables are transformed into the range between 0 and 1 by using the min-max scaling technique.

TABLE III. STATISTICAL DATA FOR NUMERICAL ATTRIBUTE (BEFORE PRE-PROCESSING)

Variable	Mean	Standard deviation	Minimum	Maximum	Median
Fixed acidity	7.2178	1.2979	3.8000	15.9000	7.0000
Volatile acidity	0.3396	0.1646	0.0800	1.5800	0.2900

Citric acidity	0.3188	0.1453	0.0000	1.6600	0.3100
Residual Sugar	5.4440	4.7569	0.6000	65.8000	3.2000
Chlorides	0.0561	0.0351	0.0090	0.6110	0.0470
Free sulfur dioxide	30.5169	17.7588	1.0000	289.0000	29.0000
density	0.9947	0.0030	0.9871	1.0390	0.9949
pH	3.2183	0.1607	2.7200	4.0100	3.2100
sulphates	0.5312	0.1489	0.2200	2.0000	0.5100
alcohol	10.4928	1.1931	8.0000	14.9000	10.3000

Fig. 2. Boxplot of Attribute Residual Sugar and Free Sulfur Dioxide



C. Descriptive Analysis

TABLE V show the attributes and the statistical analyses in the final dataset after pre-processing. The attributes that we used in training process included fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, density, pH, sulphates, alcohol, type white and best quality. The description and type of each attribute are shown in the TABLE IV below.

TABLE IV. ATTRIBUTES IN FINAL DATASET [7]

Attribute	Data Type	Data Description
Fixed acidity	Ratio	The amount of tartaric acids in wine sample. (g/dm^3)
Volatile acidity	Ratio	The amount of acetic acid in wine sample. (g/dm^3)
Citric acidity	Ratio	The amount of citric acidity in wine sample. (g/dm^3)
Residual Sugar	Ratio	The amount of sugar remaining from natural grape sugars after the alcoholic fermentation process. (g/dm^3)
Chlorides	Ratio	The amount of sodium chloride (salt) in the wine sample. (g/dm^3)
Free sulfur dioxide	Ratio	The amount of free form of SO2 (SO2 that has not yet reacted) in wine sample. (mg/dm^3),
density	Ratio	The density of wine sample (Depend on the percent alcohol and sugar content) (g/cm3)
pH	Interval	The pH value of wine samples. (from 0 to 14)
sulphates	Ratio	The amount of potassium sulphate in wine sample (g/dm3)
alcohol	Ratio	The percent of alcohol content of the wine. (vol%)
Type white	Ordinal	The type of wine. (White wine 1/red wine 0)
Best quality	Ordinal	Output variable. (1/0)

TABLE V. STATISTICAL DATA FOR NUMERICAL ATTRIBUTE (AFTER PRE-PROCESSING)

Variable	Mean	Standard deviation	Minimum	Maximum	Median
Fixed acidity	7.0970	1.0503	3.9000	11.1000	6.9000
Volatile acidity	0.3264	0.1459	0.0800	0.8300	0.2900

Citric acidity	0.3130	0.1332	0.0000	0.7400	0.3100
Residual Sugar	5.5051	4.6491	0.6000	19.500	3.2000
Chlorides	0.0516	0.0210	0.0090	0.1610	0.0460
Free sulfur dioxide	30.7158	16.3816	1.0000	83.000	29.0000
density	0.9945	0.0029	0.9871	1.0026	0.9947
pH	3.2182	0.1544	2.7400	3.7000	3.2100
sulphates	0.5192	0.1272	0.2200	0.9700	0.5000
alcohol	10.5098	1.1965	8.0000	14.050	10.4000

The final dataset contains the records of 5979 wine samples. The statistical analysis for the type of wine and its quality is reported in the TABLE VI. The majority of wine samples are white wine (78.69%), there are only 21.3% of red wine samples. Among 4705 white wine samples, 1041 (21.25%) of them have a range of quality greater than 7, which is considered good quality. The rest of 72.75% is bad. For red wine, only 170 samples are considered good quality among 1274 samples, which is only 13.34%.

TABLE VI. THE NUMBER OF QUALITY BY WINE TYPE GROUPS

Attribute	Date	Count	% quality > 7
Type white	1 (White wine)	4705 (78.69%)	1041 (21.25%)
	0 (Red wine)	1274 (21.3%)	170 (13.34%)

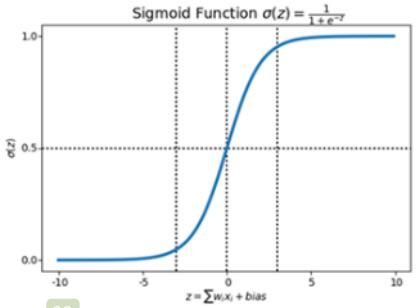
D. Modeling the Data

The project goal is to classify the quality of the wine, whether the wine sample is good or bad, based on physiochemical tests through machine learning models. Classification algorithms used in the research are given below.

- Logistic Regression

Logistic regression is a supervised learning algorithm mainly used to predict the problems with two class values. The probability of a class occurrence is calculated based on the values of a set of provided independent variables. Hence, the prediction of the logistic regression function is always between 0 and 1. The core of logistic regression algorithm is logistic function (or also known as Sigmoid function) (Figure 3), which is an S-shaped curve that can map every real-valued input into outcome between 0 and 1. To turn probability outcome into categorical form, we need to decide the threshold probability. The threshold probability is set to 0.5, which means the quality of wine is considered good if the probability is greater than 0.5, and vice versa.

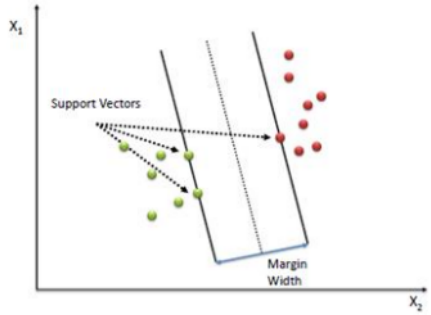
Fig. 3. Logistic Regression Graph



- Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm model widely used to solve binary classification problems. SVM algorithm classifies two classes by separating them using a decision boundary, called hyperplane (Figure 4). In a dataset, there could be many possible hyperplanes. The objective of the SVM algorithm is to find the best hyperplane with the maximum distance between data points of both classes so that the generalization error during the classification process can be greatly reduced.

Fig. 4. Support Vector Machine Graph

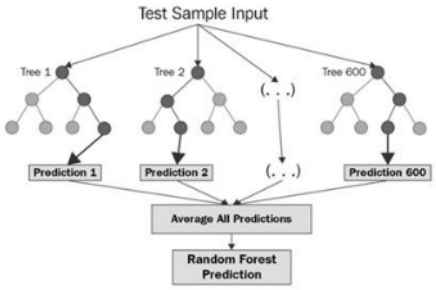


- Random Forest

Random forest algorithm is an administered algorithm for the grouping. This technique utilizes a blend of tree indicators; each depends on a random vector. This arbitrary vector has

indistinguishable and a similar circulation for all trees in the forest. It was portrayed by Breiman in 2001[11]. Random forest helps predict the important variables in classification and regression problems in a simple way.

Fig. 5. Random Forest Graph



All 12 attributes are used in the model building process. The wine dataset is split into 70% as training set and 30% as testing set. For each model, the training set was used in hyperparameters tuning process to select the optimal combination of hyperparameters that can produce the high-performance model. Hyperparameter is the parameters that define the model architecture, whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning [9]. While hyperparameters tuning is the process of searching for the ideal model architecture for a learning algorithm. A good choice of hyperparameters can make a highly performance model or on the contrary it can lead to an unending cycle of continuous training and optimization [10].

This project uses the Grid Search method for hyperparameters tuning activity. Gird search is an optimization algorithm that builds and evaluates a model for every possible combination of the provided hyperparameter values through cross-validation. In the process of hyperparameter tuning, we set the range of possible values for the selected hyperparameters for each model. Accuracy is set as the evaluative criteria for every model. Lastly, 10-folds cross-validation is defined for each model for evaluation. The values of hyperparameter set for each model are as shown in TABLE VII below. The combination of hyperparameters that has the best scored in accuracy are used in final prediction model. In the following model evaluation, the performance of models is evaluated by comparing the value of accuracy, precision, recall and F measure.

TABLE VII. HYPERPARAMETER TUNING RESULT

Model	**Parameter**
Logistic Regression	*penalty: l1, l2* *solver: liblinear, lbfgs*
Random Forest	*criterion: gini, entropy* *max_feature: 2, 4, 6, 8, 10* *n_estimators: 30, 60, 80, 100*
Support Vector Machine	*c: 1, 5, 10, 100, 1000* *kernel: linear, rbf* *gamma: 0.1, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4*

E. Communicating and Visualizing the Results

Visualization is a better and easiest way for a programmer to understand and digest the data. It is beneficial before and after doing the processing data and it also helps to know how the data can be used in the particular machine learning model.

Fig. 6. Histogram of numeric form features

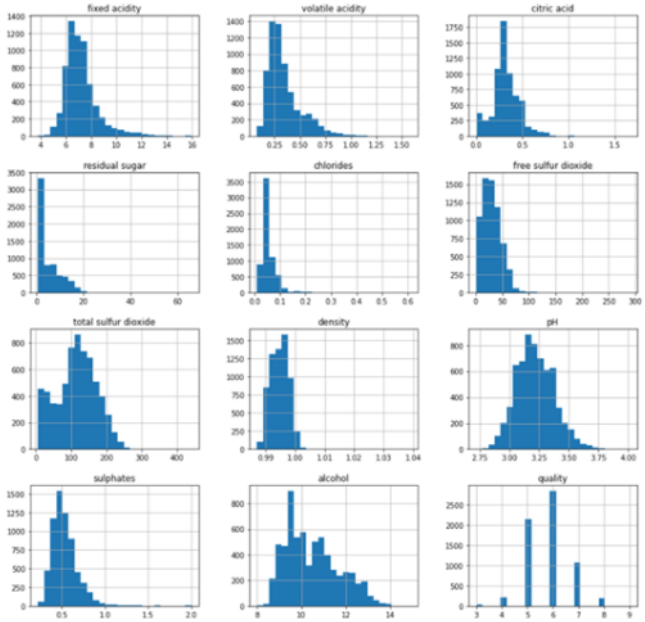


Figure 6 shows the histogram of the independent and dependent features. In our project, there are 13 types of features. We tried to plot out all of it into a histogram form to look at the distribution of the values. However, there are only 12 features that can be drawn out. Thus, the first thing we got from visualization was the 'type' is not in numeric form and it is an object datatype. It may be a categorical feature, and we had to handle it later since the computer could not understand object data. From figure 6, the values of overall features look like not so much problem except the quality feature. The histogram of quality feature shows that the range of data was quite messy. Most of the data were in the range of 5-7 and the data in quality 3,4,8 and 9 were quite small. Quality is the dependent feature in our project, and it was important. We cannot ignore it and we should do something about it. In conclusion, from figure 6, we knew at least two things that we should handle: handle type feature (object form) and handle dependent feature, quality.

Fig. 7. Correlation Heatmap

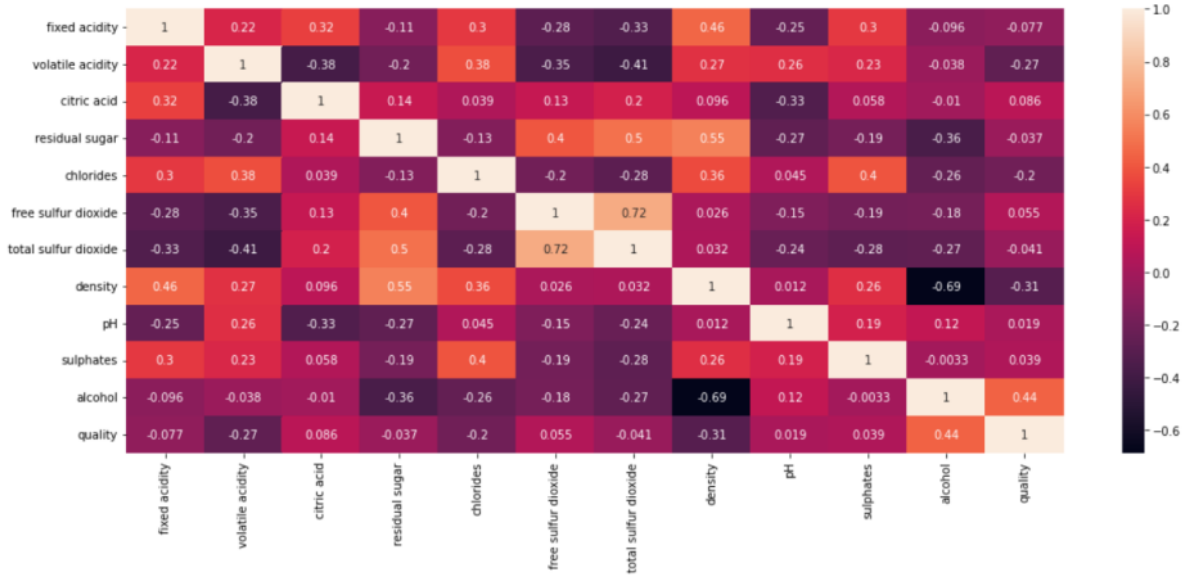


Figure 7 shows the heatmap that figures out the correlation between each feature. From the heatmap, the first information we got was we noticed that alcohol is the most strongly correlated feature with quality (0.44). It means alcohol is the most important feature that will affect the prediction of wine quality. If we have to do any processing on alcohol features, we should think carefully. Besides, if the correlation between each feature is higher than 0.7, it may give the same meaning. For example, the correlation value between quality and quality is 1 and we knew that both are the same things. From above, we noticed the correlation value between free sulfur dioxide and total sulfur dioxide is 0.72. It means both features are strongly correlated, and they might be given the same meaning for our research. Thus, we had to choose to drop one of them. The correlation values for other feature looks not so much problem. In conclusion, from figure 7, the two important messages we got are alcohol is the important feature in predicting wine quality and we have to drop free sulfur dioxide or total sulfur dioxide to prevent duplicate meanings.

Fig. 8. Histogram of Best Quality Feature

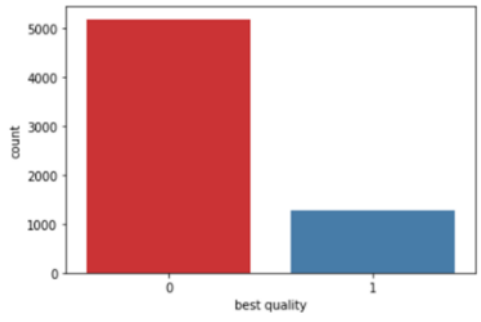


Figure 8 shows the histogram of the best quality feature. We plot this graph after we handled the quality feature. We rearranged the quality into binary form. This graph was plotted to know whether there is any mistake in our handling. We noticed only two values in the graph, which are 1 and 0. Thus, it should be no problem for our processing.

Fig. 9. Boxplot of Residual Sugar Feature

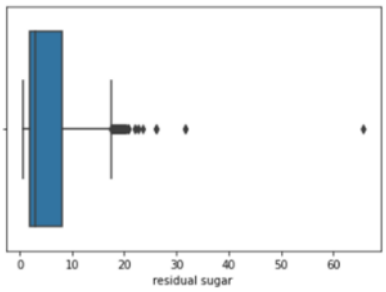


Fig. 10. Boxplot of Free Sulfur Dioxide Feature

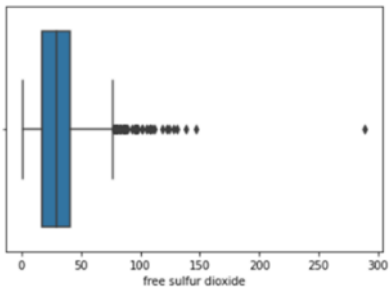


Figure 9 and figure 10 show the boxplot of residual sugar and free sulfur dioxide features. We plotted these two graphs because we wanted to make sure whether there are any outliers in our data since the statistical info shows both features look got

outliers. From figures 9 and 10, we can conclude our data has outliers. For example, the max range of residual sugar feature was among 20, but there is one data that more than 60. So, we have to handle the outliers' problem.

IV. RESULTS AND DISCUSSION

TABLE VIII. TABLE BEST PARAMETER AND BEST SCORE

Model	best_score_	best_params_
Logistic Regression	0.817442	{'penalty': 'l2', 'solver': 'lbfgs'}
Random Forest	0.871440	{'criterion': 'entropy', 'max_features': 2, 'n_estimators': 80}
SVM	0.834642	{'C': 100, 'gamma': 1.4, 'kernel': 'rbf'}

TABLE VIII shows the results after using GridSearchCV. GridSearchCV is mainly used for parameter searching. The result will come out with the best parameter for a model to predict. best_score_ is the average r2 scores on left-out test folds for the best parameter combination. For example, the best_score_ of Logistic Regression is 0.817. In Logistic Regression using GridSearchCV, we used 10-fold cross-validation to split the data into training and testing folds 10 times. The model will be fitted on training values and scored on testing values. The average score will come out when the 10-testing score is averaged. The process will be repeated for the whole parameter combinations. After that, the best average score and parameter will be recommended. It means 0.817 is the best average score throughout the entire parameter combinations. After having the best parameter, the model will train on full data. Random Forest and SVM used the same method with Logistic Regression and came out with the best_score_ 0.871 and 0.834. The score is relatively high and accurate for every model, which means our data doesn't have many problems.

TABLE IX. TABLE RESULT FROM CLASSIFICATION REPORT

Model	Best Quality	Precision	Recall	F1-Score	Accuracy
Logistic Regression	0	0.84	0.95	0.89	0.82
	1	0.61	0.29	0.39	
Random Forest	0	0.89	0.96	0.93	0.88
	1	0.79	0.56	0.66	
SVM	0	0.87	0.94	0.90	0.84
	1	0.64	0.44	0.53	

TABLE IX shows the precision, recall, f1-score and accuracy for three selected models. From TABLE IX, we noticed that the highest accuracy of our models is 0.88 (Random Forest), followed by 0.84 (SVM) and 0.82 (Logistic Regression). The accuracy was relatively high and meant our models could predict 80% and above correctly. Based on three models, we can conclude the best model in our project is Random Forest since it can be accurately predicted by 88%. Random Forest is a supervised machine learning algorithm that mainly used for classification and regression problem. Random Forest will create decision trees on the various sample and comes out with the majority vote results for classification or average for regression case. It will perform better results for classification problems [12]. Random Forest can perform better than other models because Random Forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature when splitting nodes, it searches for the best features among the random group of features. Thus, this makes the result in a wide variety, which often leads to better models [13].

As mentioned before, 1 in best quality represents the wine quality value in the range seven and above, and the rest will be 0. For the best model Random Forest, the precision for best quality 0 is 0.87. Precision means the percentage of correct predictions. 0.87 of precision means 87% correctly predicted. Besides, the recall of the Random Forest model for best quality 0 is 0.94. Recall means what percent of positive cases we get. 0.94 of recall means that 94% of the data was caught as best quality 0. F1-score for Random Forest on best quality 0 is 0.90. F1-score will represent the percentage of the correctness of positive predictions. 0.90 of f1-score means 90% of the data correctly predicted as best quality 0. However, results of best quality 1 are quite lower than best quality 0. This problem occurred because values of wine quality in the range of 7 and above are quite less. Logistic Regression and SVM also meet the same problems with logistic regression. Their precision, recall and f1-score in best quality 0 (quality range below seven) are quite high compared with best quality 1 (quality range seven and above). This problem may be solved by collecting more of the range of data. Since the overall results for our model are pretty good since the accuracy is 80% and above, we can conclude our project can be considered successful.

V. CONCLUSION

In this project, the first thing we did was import the necessary libraries and modules that were needed, such as pandas (used for analysing the data in the data frame), seaborn (used for data visualizing) and etc. We also deleted the useless warning message and imported the data into the data frame. In the visualization part, we draw some histograms to read the data more clearly. We also look at the heatmap for knowing the

correlation value between each feature. For the data preparation part, we had handled missing values, handled categorical columns, made some fundamental changes on the dependent feature, handled outliers and did normalization for the independent data. The data was split into 30% for testing purposes and 70% for training purposes. Three models were used in our report to classify the quality of the wine, such as Random Forest, Logistic Regression and SVM. We found that Random Forest performed better result than others. Thus, in our project, we can conclude Random Forest is the best model for predicting the wine quality by the features given.

The advantage of using a machine learning algorithm in this project for prediction is that it can analyse a massive quantity of data. Although it has thousands of data in our project, it can still provide a faster and more accurate result. However, apart from the advantage, our project has some limitations. The first limitation is the range of the dependent variable is quite messy. For example, most of the quality of data lies between 5,6 and 7. There are 2820 data in quality 7, 2128 data in quality 5 and 1074 data in 7. However, there are only 214 in quality 4, 192 in quality 8, 30 in quality 3 and 5 in quality 9. These kinds of values will make our prediction more challenging to identify the good or bad quality of wine and may come out with a less accurate result. Since quality is the main focus of the analysis, we have to balance it to improve our predictive model. We chose to rearrange it into the best quality such as values lies in quality 7 and above will be 1 and others will be 0. Throughout this whole process, the data may lose a lot of meaning.

Another limitation in this project is that only 13 features (including dependent and independent features) were used to make a prediction. It may be narrow down the accuracy of the predicting quality of a wine. The features used in the project were only type, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and quality. The important feature that will affect the quality of wine should be added, such as year of harvest, brew time, viticulture and ways of vinification [14]. In the future, we can collect more data on the various kinds of wine quality and collect more important features that will affect the prediction of wine quality. The limitations can be reduced to the minimum throughout the processes and can get a better accurate result. Besides, the project's analysis may also help the wine businesses predict the wine quality based on specific attributes and manufacture good quality wine.

ACKNOWLEDGMENT

First of all, we wish to thank our Principle of Data Science lecturer, Dr. Nor Samsiah Binti Sani. Thank you so much to Dr for the ability of teaching and discussion. From the knowledge learned, we successfully completed our project on time. Besides, we are also very grateful that we can meet very cooperative teammates. In every meeting, everyone actively discusses opinions and strives to complete this project. We had done the coding together and separated the job for writing the report. The below shows the biography of each member with the contributions:

Group member 1: Ng Hao Lin (A175838)

Ng Hao Lin is a student of computer science, data science track. In this project, he played the role of a team leader. The parts that he focused on in this project for the report were result and discussion, conclusion, acknowledgments, and report combining.

Group member 2: Ng Xing Ning (A176493)

Ng Xing Ning is a student of computer science, data science track. In this project, she was an active member. The part that she focused on in this project for the report was research methodology.

Group member 3: Tan Jing Xuan (A175711)

Tan Jing Xuan is a student of computer science, data science track. In this project, he was a member that has much inspiration. The parts that he focused on in this project for the report were abstract, introduction and literature review.

REFERENCES

[1] Massimo Lucarini, M. Durazzo, Ginevra Lombardi-Boccia, E. B. Souto, Antonello Santini (2021) Wine Polyphenols and Health: Quantitative Research Literature Analysis. https://doi.org/10.3390/app11114762

[2] Hua Li,Hua Wang,Huanmei Li,Steve Goodman,Paul van der Lee,Zhimin Xu,Alessio Fortunato,Ping Yang (2018) The worlds of wine: Old, new and ancient. https://doi.org/10.1016/j.wep.2018.10.002

[3] Wine Data Set. Stefan Aeberhard. "UCI Repository of machine learning databases." https://archive.ics.uci.edu/ml/datasets/wine

[4] Rohan Dilip Kothawadem (2021) Wine Quality Prediction Model Using Machine Learning Techniques. https://www.diva-portal.org/smash/get/diva2:1574730/FULLTEXT01.pdf

[5] Mayur Badole (2021) Wine Quality Prediction Using Machine Learning. https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/

[6] Dahal, K. , Dahal, J. , Banjade, H. and Gaire, S. (2021) Prediction of Wine Quality Using Machine Learning Algorithms. Open Journal of Statistics, 11, 278-289. https://doi.org/10.4236/ojs.2021.112015

[7] Paulo Cortez, University of Minho, Guimarães, Portugal, http://www3.dsi.uminho.pt/pcortez A. Cerdeira, F. Almeida, T. Matos and J. Reis, Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal, 2009. https://archive.ics.uci.edu/ml/datasets/Wine+Quality

[8] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Wine Quality Dataset. 2009. http://www3.dsi.uminho.pt/pcortez/wine/

[9] Kizito Nyuytiymbiy. (2020, Dec 30). Parameters and Hyperparameters in Machine Learning and Deep Learning. Retrieved from towards datascience: https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac

[10] Niwratti Kasture. (2020, Nov 16). Why Hyper parameter tuning is important for your model? Retrieved from Analytics Vidhya: https://medium.com/analytics-vidhya/why-hyper-parameter-tuning-is-important-for-your-model-1ff4c8f145d3#:~:text=Hyper%20parameter%20tuning%20(optimization)%20is,of%20continuous%20training%20and%20optimization

[11] W. L. Martinez, A. R. Martinez, "Supervised Learning" in Computational Statistics Handbook with MATLAB, 2nd ed., Boca Raton, FL, USA: Chapman & Hall/CRC, 2007, pp. 363-431.

[12] R, S. E. (2021, June 17). Understanding Random Forest. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/#:~:text=Random%20forest%20is%20a%20Supervised,average%20in%20case%20of%20regression.&text=It%20performs%20better%20results%20for%20classification%20problems.

[13] Donges, N. (2021, July 22). A Complete Guide to the Random Forest Algorithm. Retrieved from builtin BETA: https://builtin.com/data-science/random-forest-algorithm

[14] JAMES, S. (2020). FAQs - Wine Quality. Retrieved from BERRY BRO & RUDD: https://www.bbr.com/wine-knowledge/faq-quality

PEERS EVALUATION

TABLE X. PEERS EVALUATION

Performance	**Ng Hao Lin (A175838)**	**Ng Xing Ning (A176493)**	**Tan Jing Xuan (A175711)**
Shows strong initiative	1	1	1
Works well with others in group-based projects	1	1	1
Takes instructions and follows leaders well/	1	1	1
Gives instructions and discusses well	1	1	1
Stays focused on tasks at hand	1	1	1
Knows how to prioritise tasks	1	1	1
Has good communication with team members	1	1	1
Is dependable	1	1	1
Gets assignments in on time	1	1	1
Responsive to discussions online and offline	1	1	1
Work is of high quality	1	1	1

KerianTeam_TC3313_MiniProject

ORIGINALITY REPORT

15%	11%	8%	9%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	www.scirp.org Internet Source	2%
2	Submitted to Universiti Kebangsaan Malaysia Student Paper	2%
3	Sunny Kumar, Kanika Agrawal, Nelshan Mandan. "Red Wine Quality Prediction Using Machine Learning Techniques", 2020 International Conference on Computer Communication and Informatics (ICCCI), 2020 Publication	1%
4	www.diva-portal.org Internet Source	1%
5	espace.library.uq.edu.au Internet Source	1%
6	Yogesh Gupta. "Selection of important features and predicting wine quality using machine learning techniques", Procedia Computer Science, 2018 Publication	1%
7	ijsrcseit.com	

	Internet Source	1 %
8	Submitted to University of Sydney Student Paper	1 %
9	stackoverflow.com Internet Source	1 %
10	Submitted to Taylor's Education Group Student Paper	<1 %
11	zombiedoc.com Internet Source	<1 %
12	Submitted to Tilburg University Student Paper	<1 %
13	akdemir.user.cos.ucf.edu Internet Source	<1 %
14	Submitted to RMIT University Student Paper	<1 %
15	medium.com Internet Source	<1 %
16	www.compstat2014.org Internet Source	<1 %
17	Pradeepta Mishra. "Practical Explainable AI Using Python", Springer Science and Business Media LLC, 2022 Publication	<1 %

18	dataaspirant.com Internet Source	<1 %
19	"Intelligent Internet of Things", Springer Science and Business Media LLC, 2020 Publication	<1 %
20	euvs-vintage-cocktail-books.cld.bz Internet Source	<1 %
21	Submitted to University College London Student Paper	<1 %
22	www.igi-global.com Internet Source	<1 %
23	www.coursehero.com Internet Source	<1 %
24	Submitted to CSU, San Jose State University Student Paper	<1 %
25	Submitted to Nottingham Trent University Student Paper	<1 %
26	"International Conference on Innovative Computing and Communications", Springer Science and Business Media LLC, 2022 Publication	<1 %
27	Submitted to Auckland University of Technology Student Paper	<1 %
	www.ijert.org	

28	Internet Source	<1%
29	Submitted to University of Iowa Student Paper	<1%
30	towardsdatascience.com Internet Source	<1%
31	www.kaggle.com Internet Source	<1%
32	Submitted to Hong Kong Baptist University Student Paper	<1%
33	Ripon Patgiri, Udit Varshney, Tanya Akutota, Rakesh Kunde. "An Investigation on Intrusion Detection System Using Machine Learning", 2018 IEEE Symposium Series on Computational Intelligence (SSCI), 2018 Publication	<1%
34	amslaurea.unibo.it Internet Source	<1%
35	www.science.gov Internet Source	<1%
36	upcommons.upc.edu Internet Source	<1%
37	www.slideshare.net Internet Source	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		