I.J. Intelligent Systems and Applications, **2022, 5, 35-46**
Published Online on October 8, 2022 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2022.05.04

Detailed Study of Wine Dataset and its Optimization

Parneeta Dhaliwal
Department of Computer Science and Technology, Manav Rachna University, Faridabad Sector 43, Haryana 121001, India
E-mail: parneeta07@gmail.com

Suyash Sharma
Department of Computer Science and Technology, Manav Rachna University, Faridabad Sector 43, Haryana 121001, India
E-mail: suyashsharma9211@gmail.com

Lakshay Chauhan
Department of Computer Science and Technology, Manav Rachna University, Faridabad Sector 43, Haryana 121001, India
E-mail: lakshaychauhan100@gmail.com

Received: 07 March 2022; Revised: 16 June 2022; Accepted: 12 August 2022; Published: 8 October 2022

Abstract: The consumption of wine these days is becoming more common in social gatherings and to monitor the health of individuals it's very important to maintain the quality of the wine. For the assessment of wine quality many methods have been proposed. We have described a technique to pre-process the “Vinho Verde” wine dataset. The dataset consists of red and white wine samples. The wine dataset size has been reduced from a total of 13 attributes to 9 attributes without any loss of performance. This has been validated through various classification techniques like Random Forest Classifier, Decision tree Classifiers, K-Nearest Neighbor Classifier and Artificial Neural Network Classifier. These classifiers have been compared based on two performance metrics of accuracy and RMSE values. Among the three classifiers Random Forest tends to outperform the other two classifiers in various measures for predicting the quality of the wine.

Index Terms: Machine Learning, Optimisation, Data Analytics, Wine dataset.

1. Introduction

In the current era of modernization and digitization, the amount of data generated is unbelievably high. Data has always been the source of information, when processed in numerous ways and the information so generated can be used for making future predictions [1]. With the Internet of Things (IoT) taking over the world, the future of automation is going to increase. As in 2012, every day 2.5 exabytes of data had been created [2]. The amount of data processed on a day-to-day basis has also increased resulting in reduced efficiency of software’s and automation due to lack of proper data management techniques. Thus, data management tools are required for processing large amounts of data and extracting the relevant information efficiently [3].

Artificial intelligence (AI) and machine learning (ML) has shown rapid growth in recent years in the context of data analysis. The new optimized computing techniques typically allow the applications to function efficiently in the real world [4]. Machine learning [5], a branch of Artificial intelligence, deals with training a machine according to a particular dataset, to use its learning for future decision making. The trained system can imitate the way human beings learn and analyse, gradually increasing their predictive accuracy.

In the current age of the fourth revolution, the use of machine learning tools has spread across various industries, providing accurate decision making and lower processing time [6,7]. There are various applications around the globe that use data-driven decision-making such as facial recognition, e-commerce, business intelligence etc. Facial recognition means identifying the facial features of a particular person. It helps in identifying a criminal in an offense using CCTV cameras [8]. Another application is E-commerce product recommendation where the system recommends a product to the customer based on his earlier shopping experience [9].

Data analytics [1] can either be predictive, descriptive, diagnosis or prescriptive. Descriptive analysis determines the current system state and provides current and previous data in the form of graphical or statistical output [1,10]. To find out “Why something is happening” [11] or “Why did it happen”, we need Diagnostic analysis for going deep in data to

search for valuable insights [1,10]. Predictive analysis is invoked when a certain prediction is needed. It can be defined as “what will happen?” [11]. Actions recommended on the basis of diagnosis or a prediction [1,10] refers to Prescriptive analysis. It says “What action should we take” [11]. It often recommends how to fix or optimize something or how to achieve objectives like customer satisfaction, profits, and cost saving.

Organizations across industries use prescriptive analytics for a range of use cases spanning strategic planning, operational and tactical activities. Various OTT platforms use predictive modelling techniques based on machine learning to determine what users will love to watch and suggest their choice. It also helps in making customized suggestions for a business model.[5]

Machine learning techniques have been further categorized as Supervised machine learning, Unsupervised machine learning, semi supervised machine learning and Reinforcement machine learning. Supervised algorithms such as Classification and regression use labelled datasets to train the algorithms that classify data and predict outcome accurately. Unsupervised Algorithms use unlabelled data to discover patterns that help in clustering or association problems and have Data-Driven Approach. Semi-supervised Models are built using a small amount of labelled data and large amounts of unlabelled data. In Reinforcement learning, agents are trained on a reward and punishment mechanism, they are rewarded for each correct prediction and move and punished for wrong moves.

In our paper, we have performed comparative analysis of various classification algorithms using Wine dataset based on various performance metrics such as accuracy and RMSE. Our objective was to optimize the size of the dataset without affecting the performance metrics. We have used a dataset which consists of two types of wine that is red and white wine of Portuguese "Vinho Verde", for predicting the wine quality. We have removed various attributes on the basis of correlation, Ranking etc.

Our paper provides a step-by-step procedure of data pre-processing and we have also applied ANN based model for wine quality assessment and comparative analysis which in comparison with other related work is more helpful for a beginner level programmer as we have defined a procedure to help them understand pre-processing in a better way whereas the current related paperwork does not completely specify data pre-processing techniques.

In the next Section, we are giving a brief description about the Wine dataset. Section 3 shares the related work. In Section 4, we have identified the various research objectives that we tend to answer through our work. Section 5 gives the conclusion and the future work.

2. Dataset Description

Our paper focuses on performing supervised machine learning on the Portuguese "Vinho Verde" wine dataset [Cortez et al., 2009]. The dataset can be viewed as a classification or regression task. It contains 6,497 instances with 13 attributes each. The Wine quality dataset uses the 12 different attributes of wine to predict wine of a certain “quality “on a scale of (3-9) in ascending order. In this dataset there is no data about grape types, wine brand or selling price of the wine due to logistic issues. There are only one sensory feature and 11 psycho chemical features present. The various attributes in the Wine dataset are float type except the quality attribute that represents an integer data.

We have done data pre-processing wherein the shape of the dataset came out as (6497, 13), where 6497 are the number of rows and 13 are the number of columns. The description of our dataset in terms of count, mean, standard deviation, minimum value,1st quartile, median value, 3rd Quartile and maximum value can be seen in Fig.1. It was observed that fixed acidity has a minimum value of 3.80 and 75% value of fixed acidity is 7.70 and the maximum value came out to be 15.90 which was very high as compared to other values ranging from 3.80 to 7.70. Further, residual sugar has a minimum value of 0.60 and 3rd Quartile value as 8.100 and the maximum value was 26.05. The minimum and Q3 values of free sulfur dioxide are 1.0 and 41.0, respectively and the maximum value being 289.0, which is relatively higher as compared to the other observed values. It has been observed that the minimum value of total sulfur dioxide is 6.0 and the Q3 value is 156.0, the maximum value being 440.0. Hence, there seems to be some outliers in the dataset. The detailed description about the Wine dataset is as shown in Table 1.

Table 1. Description of Wine dataset

	color	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pH	sulphates	alcohol	quality
count	6497	6487	6489	6494	6495	6495	6497	6497	6497	6488	6493	6497	6497
mean	1.246114	7.216579	0.339691	0.318722	5.444326	0.056042	30.525319	115.744574	0.994697	3.218395	0.531215	10.491801	5.818378
std	0.430779	1.29675	0.164649	0.145265	4.758125	0.035036	17.7494	56.521855	0.002999	0.160748	0.148814	1.192712	0.873255
min	1	3.8	0.08	0	0.6	0.009	1	6	0.98711	2.72	0.22	8	3
25%	1	6.4	0.23	0.25	1.8	0.038	17	77	0.99234	3.11	0.43	9.5	5
50%	1	7	0.29	0.31	3	0.047	29	118	0.99489	3.21	0.51	10.3	6
75%	1	7.7	0.4	0.39	8.1	0.065	41	156	0.99699	3.32	0.6	11.3	6
max	2	15.9	1.58	1.66	65.8	0.611	289	440	1.03898	4.01	2	14.9	9

To get these attributes within range, we applied outlier detection techniques to help remove the outliers and noisy data which then helped the attributes to come within reasonable range as their values would show a large difference between minimum value and 3rd Quartile value present in the dataset. We have not considered remaining attributes as

they were within range. Further we have plotted a graph of wine quality vs count through which it is clearly visible that that wine quality which is equal to 6 resides in more numbers in the dataset than the rest.

Next, we discovered that there are some null values in the dataset in the form of noisy data which were then filled by mode values of respective attributes. Further in the program we tend to scale the values of all 12 attributes to help make the classification and prediction of quality faster and easier, so we applied normalization technique which helped to scale the values from 0 to 1 for the columns color, fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. Only the attribute quality was not scaled from 0 to 1 as it had the wine quality ranging from 3 to 9 which is to be considered as constant.

3. Related Work

The wine quality [3] was predicted from physicochemical properties data which was provided by UC Irvine machine learning repository. The dataset contains 1599 instances for red wine and 4898 instances of white wine respectively, Including 11 features of physicochemical data. The best accuracy attained was 99.5229% from Random Forest classifier. Overall, three classifiers were used to classify both red and white quality wine including random forest, k-nearest-neighbor, and support vector machine. It was also observed that by performing principal component analysis in feature selection the accuracy rate of random forest algorithm increased.

In this paper [12] the author has explained the use of product quality certification by industries to boost their product. This is a time-consuming task if done by human beings, so the help of machine learning is taken into consideration. Also, it states that better prediction can be achieved if selected attributes are used for classification. The dependency of target variable on independent variable is determined and value of targeted value is predicted. For prediction of dependent variable linear regression, neural network and support vector machine are used.

The author [5] presents a comprehensive view on various machine learning techniques such as supervised, unsupervised, reinforcement and semi-supervised learning. Various machine learning algorithms from above techniques have been used such as Random Forest, Decision Tree, and K- Nearest-Neighbour. In the paper [13], the importance of quality of wine for consumption to preserve human health was discussed. Red wine dataset is used as it is observed that consumption of red wine has increased. The results are compared among training set and testing set and the best out of the three techniques (Random Forest, Support vector machine, Naive bayes) according to the training set results are predicted. All the best features from other techniques are merged together for better accuracy and efficiency. An online ensemble approach DDWM (Diversified dynamic weighted majority) [14] to classify new data instances has been discussed. An expert in either of the ensembles is updated or removed asper its classification accuracy and a new expert is added based on the final global prediction of the algorithm and the global prediction of the ensemble for any data instance.

The author [15] used a wine dataset and a data mining approach to predict human wine taste preferences. Three regression techniques such as multiple\linear regression, neural network and Support Vector Machine were used for analysis. It was observed that SVM (Support Vector Machine) got the best results, outperforming the various regression and classification models. Comparison of various machine learning models [16] such as support vector machine, rigid regression and other machine learning models have been done. After comparing their R, MSE and MAPE values, it was found that the gradient boosting regressor proved to be the best model with values of 0.6057,0.3741 and 0.0873 respectively.

Wine Quality was predicted [17] using a dataset of red wine and white wine using machine learning algorithms like naive bayes and decision tree algorithms. The comparative analysis led to the conclusion that classification can help in improving the quality of wine during production. The proposed decision-tree based method [18] proved to be better than other machine learning models such as LibSVM (Support vector machine), Bayes Net, MultIPerceptron (Multi-layer perceptron). The various performance metrics used were Precision, Recall, F-measure, and Accuracy and found that the proposed method had the best performance with the values of 60.10, 60.70, 60.30 ,60.66 respectively. A clear roadmap [19] was provided about the importance of feature selection for predicting the quality of wine. The author also concluded that SA based feature sets performed better than GA based feature sets (Genetic algorithm sets). The best classifier that worked best for the feature selection technique was SVM (Support vector machine) compared to other various machine learning classifiers with an accuracy range from 95.23% to 98.81%.

In Random Forest Classifier many decision trees operate as a group. Each individual tree in random forest gives its prediction and the class with most no of votes becomes our model's prediction [20]. This classifier works on parallel ensembling [5] and uses the best decision tree classifier from multiple ones, for achieving the best accuracy and minimizes over-fitting [21]. It is compatible with classification and regression problems [5]and can be implemented for linear and non-linear datasets.

4. Research Objectives

In this section, we have listed various research objectives that have been identified and answered.

Q1. What would be the minimum size of the Wine dataset that needs to be maintained, without affecting its predictive

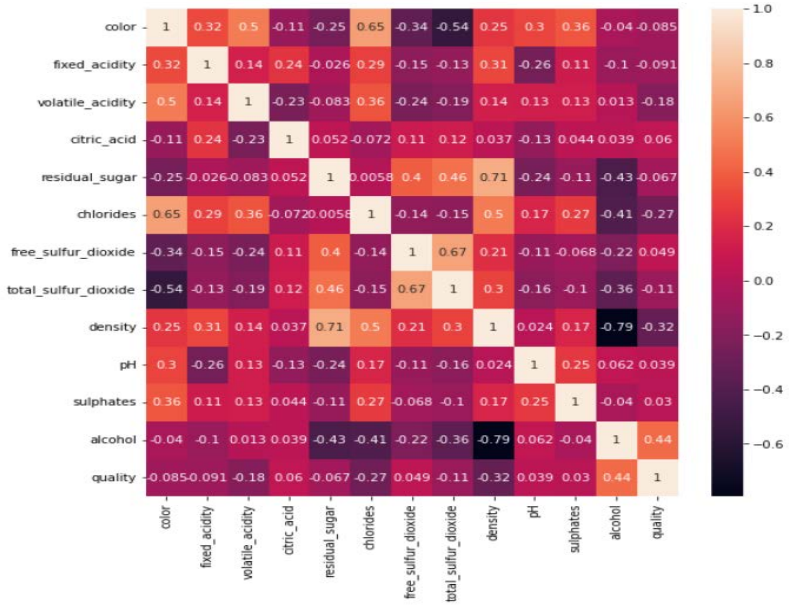
performance?

Q2. Measure the average value of the various attributes that clearly classify the wine as a good quality wine.

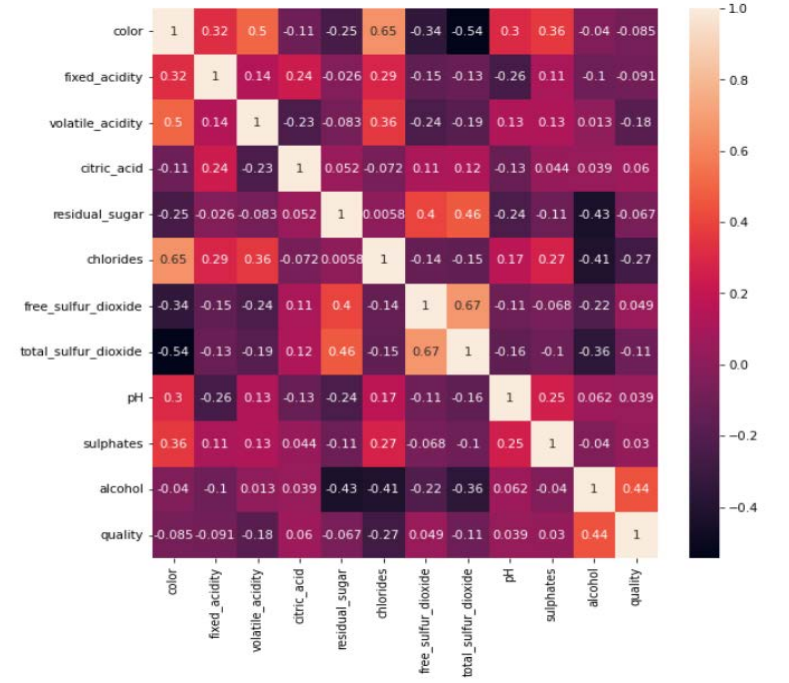
Q3. Measure the Concentration of various attributes in good red and white wine.

Q4. Which is the best classifier for predicting the Wine quality?

To answer the first research objective, we have generated a heat map which gives us the Pearson correlation coefficient value among the various attributes as shown in Fig. 1 (a). As per the literature, the value of Pearson's coefficient determines the type and level of correlation between the various attributes as seen in Table 2. It has been found in the heat map that the coefficient value between the attributes- density and alcohol is -0.79 and density and residual sugar is 0.71. These values of density-alcohol and density- residual sugar to be more than 0.7, so the density attribute was dropped from the dataset. After dropping the density attribute the coefficient value of each attribute was observed to be below 0.7 as seen in Fig. 1(b).



(a)



(b)

Fig.1. (a) Heat Map for Wine Dataset before removing Density attribute. b). Heat map for Wine dataset after removing the density attribute

Table 2. Pearson's Coefficient value and its interpretation

Size of Correlation	Interpretation
.90 to 1.00 (-.90 to -1.00)	Very high positive (negative) correlation
.70 to .90 (-0.70 to -.90)	High positive (negative) correlation
.50 to .70 (-.50 to -.70)	Moderate positive (negative) correlation
.30 to .50 (-.30 to -.50)	Low positive (negative) correlation
.00 to .30 (.00 to -.30)	negligible correlation

After performing bivariate analysis, we observed that there is some strong relationship between the attributes (numeric) and the target variable quality as shown in Fig. 2. We found that some attributes showcase a very strong relationship with the target variable. The bivariate analysis showed a sharp decline and incline variation of numeric variables. However, it was observed that the attribute "total_sulphur_dioxide" had almost a consistent value with the change in quality. Hence, total_sulphur_dioxide can be removed from the dataset.

We have calculated the p-value for all the attributes considering wine quality as target variable. All the attributes i.e. color, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH and sulphates had p-value lower than 0.05 as seen in Table 3, proving them to be highly significant. However, citric acid had a p-value of 0.3478 and fixed acidity had a p-value of 0.3434 that was more than 0.05. Hence, citric acid and fixed acidity proved to be highly insignificant to the model and could be removed from the dataset.

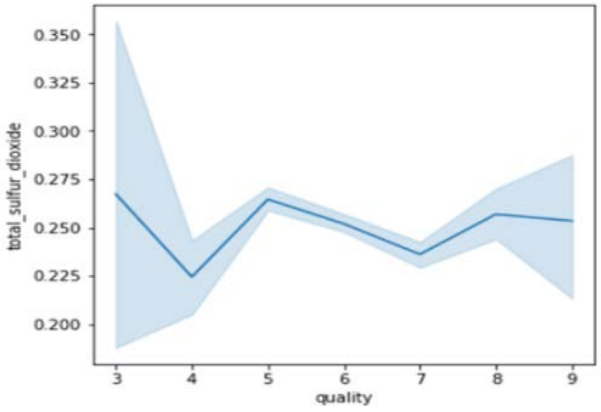


Fig.2. Bivariate analysis of total sulfur dioxide

Table 3. P-value of various attributes

Attributes	pvalue
color	0
fixed_acidity	0.3434
volatile_acidity	0
citric_acid	0.3478
residual_sugar	0
chlorides	0
free_sulfur_dioxide	0
pH	0.0448
sulphates	0
alcohol	0

Finally, we conclude that the most optimum size of the Wine dataset without affecting the actual performance of the model is 9 attributes as compared to the original dataset with 12 attributes in total, without comprising the performance of the system.

Further validating our results, we rank the various attributes in our dataset by performing feature importance using Extra Trees Classifier [22,23] as shown in Fig. 3. Feature importance [23] calculates the scores of each feature present in the dataset. It has been observed that higher the score more relevant is the feature towards the target variable. As seen in Fig. 3, the score of color is the least as compared to the other features. However, the color feature is not removed from the dataset as it distinguishes the type of wine, either to be red wine or white wine.

We have tried to optimize the size of the dataset by analysing the performance of Random Forest Classifier (RFC), k-NN and Decision Tree classifiers in terms of accuracy [24] and RMSE performance metrics, with variations in the number of attributes. As seen in Table 3, fixed acidity has the score of 0.098, so first we check the performance of various

classifiers by removing the low scored fixed acidity feature as it has very low relevance in predicting the quality of the wine. As shown in Table 4, RMSE [25] value increased from 0.38 to 0.40 for the Random Forest classifier,0.49 to 0.54 of k-NN, RMSE value of Decision Tree classifier remained unchanged and for Seq_ANN (Sequential model Artificial Neural Network) the RMSE value increased from 0.42 to 0.45 which is acceptable.

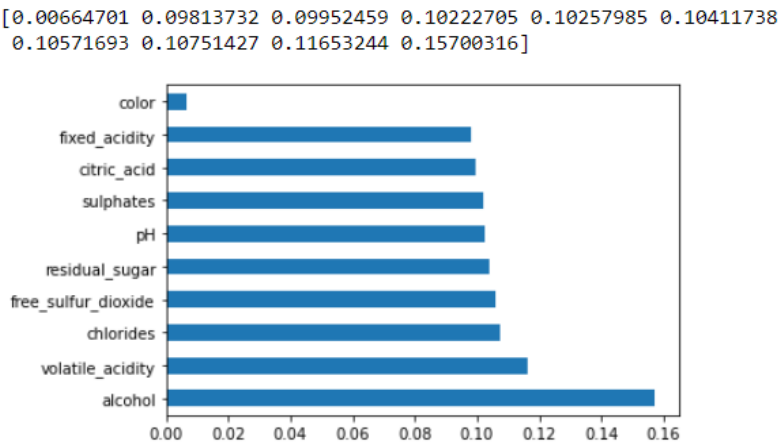


Fig.3. Graphical Representation of scores of each attribute with respect to quality (Target variable).X-axis represents the score

Apart from the three classifier (Random Forest classifier, K-Nearest Neighbor classifier and Decision Tree Classifier) we have also applied Artificial neural network Sequential model on our dataset for a better comparative analysis. We have used 3 layers of neural network, The first dense layer had 84 units of neurons and activation function 'relu', Second dense layer had 14 units of neurons and activation function 'relu', and the third layer was output layer with 7 units of neurons with activation function 'sigmoid'. The third output layer had 7 units of neurons because the 'quality' attribute was divided into 7 categories (ranging from 3 to 9). The input layer parameter was changed subsequently as the number of input features changed according to feature selection. For our Sequential model we have used 'adam' optimizer and loss function 'categorical_crossentropy', further we will move towards feature selection and comparative analysis of different classifiers and Sequential model (ANN) [26].

The removal of fixed acidity resulted in an accuracy change of RFC from 90.08% to 89.79%; and k-NN accuracy slightly changed from 86.69% to 86.66%, however accuracy of DT remained unchanged, which is 83.54%, Next for Seq_ANN (Sequential model Artificial Neural Network) we saw a slight dip in accuracy from 81.00% to 77.19% which is acceptable as seen in Table 5. A decrease of 0.29% in accuracy was observed for Random Forest Classifier and for Seq_ANN decrease of 3.81% was observed whereas negligible change was observed in KNN classifier of 0.03%. (9 attributes).

Further, the next least relevant attribute was citric acid with a score of 0.099. The removal of citric acid resulted in a decrease in accuracy of K-NN, DT, RFC and Seq_ANN as seen in Table 5. The accuracy of DT decreased more drastically by 1.44% (from 83.54% to 82.10%). As shown in Table 4, RMSE values for KNN and DT remained the same however RFC observed a very minimal decrease of 0.01 (from 0.40 to 0.39) whereas for Seq_ANN the accuracy drastically decreased by 1.73% (from 77.19 to 75.56%) also the RMSE value showcased a similar trend of drastic decrease from 0.45 to 0.54. Hence, it was better to maintain this attribute in the Wine dataset. So, the conclusion came that the features like total sulfur dioxide, fixed acidity and density were less relevant for the target variable(quality). These 3 features didn't impact accuracy of the classifiers to a very large extent, without a huge increase in their RMSE values.

Table 4. RMSE Values after removing attributes.

S.no	Classifiers	RMSE(All,12)	RMSE(10)Removed-Density, Total Sulfur Dioxide	RMSE(9)Removed-Fixed Acidity	RMSE(8)Removed-Citric acid
1	RFC	0.38	0.37	0.40	0.39
2	KNN	0.49	0.49	0.54	0.54
3	DT	0.60	0.58	0.60	0.60
4	Seq_ANN	0.42	0.50	0.45	0.54

Table 5. Accuracy of the classifiers

S.no	Classifiers	Accuracy(All,12)	Accuracy(10) Removed- Density, Total Sulfur Dioxide	Accuracy(9) Removed-Fixed Acidity	Accuracy(8)Removed-Citric acid
1	RFC	90.08	90.14	89.79	89.91
2	KNN	86.69	87.31	86.66	86.30
3	DT	83.54	83.61	83.54	82.10
4	Seq_ANN	81.00	79.52	77.19	75.46

We have taken 9 attributes in consideration because the accuracy of Random Forest classifier decreased by 0.17%, Decision Tree classifier decreased by 1.44% and for k-NN classifier depleted by 0.39% and Seq_ANN depleted by 1.73% after removal of the next relevant attribute i.e., citric acid. Whereas for the selected 9 attributes compared with all the 12 attributes the accuracy of Decision tree classifier remained the same, for k-NN classifier depleted by 0.03%, for Random Forest Classifier it depleted by 0.29% and for Seq_ANN it depleted by 3.81%. Hence, we concluded that KNN, DT and RFC perform better with 9 attributes, and further removal of more attributes results in decrease in the accuracy of all three classifiers and Seq_ANN (Sequential model Artificial Neural network).

Next, we answer the second research objective here. The quality of wine above 6 is considered as good quality wine. We eliminated all the tuples having wine quality value less than 7 and calculated the descriptive statistics of all the attributes. The mean value, standard deviation, minimum and maximum value for the selected 9 best attributes is shown in Table 6.

Table 6. The descriptive statistics for the selected 9 attributes

	color	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	pH	sulphates	alcohol	quality
count	1036	1036	1036	1036	1036	1036	1036	1036	1036	1036
mean	1.078185	0.269502	0.32806	4.852944	0.039545	31.799228	3.218983	0.509546	11.463401	7.162162
std	0.268593	0.087828	0.067865	3.920995	0.011958	12.830961	0.151463	0.135243	1.199058	0.381653
min	1	0.08	0.15	0.8	0.012	3	2.84	0.22	8.5	7
25%	1	0.2	0.28	1.8	0.03175	23	3.11	0.4	10.8	7
50%	1	0.26	0.32	3.2	0.037	31	3.22	0.485	11.5	7
75%	1	0.33	0.36	6.8625	0.045	40	3.33	0.59	12.4	7
max	2	0.52	0.5	14.8	0.077	67	3.66	0.88	14.2	9

Next, we answer the third research objective hereto get the measure of the concentration of the various attributes in red and white wine, we generate 3D visualization of a Violin Plot [27] as seen in Fig. 4. A violin plot describes the kernel density estimation [27,28] of each attribute with respect to the quality of red and white wine. A violin plot is a combination of a box plot and kernel density plot [28], which shows peaks in the data. It is used to visualize the distribution of numerical data. In fig. 4, we observe that the concentration of white wine is more for quality 3 to 6 as compared to red wine. Moreover, the frequency of red wine of quality 9 is zero that is why in each violin plot where the quality is 9 there is no red wine concentration. The concentration of each attribute in white wine and red wine is determined by using the kernel density estimation. Further a comparison of both (red and white wine) concentrations is represented in the form of a violin plot for wine qualities (ranging from 3 to 9). These graphs clearly depict the concentration of selected 9 attributes with respect to wine quality, that are required to make a wine quality good. Further, wine quality which is greater than 6 is considered to be good quality wine and wine quality between 3-6 is considered as low quality or poor-quality wine.

As seen in Fig. 4. (b), concentration of pH in red wine is more as compared to white wine, between wine quality 3 to 6. Moreover, the concentration of red wine was more than that of white wine. Similarly, in the wine quality between 7 to 9(considered as good quality wine) the red wine still dominated. For Wine quality 9 red wine does not exist in the violin plot because of the absence of red wine in wine quality number 9 as shown in Fig. 4(a). As shown in Fig. 4(c), concentration of free sulfur Dioxide in low quality red wine (3 to 6) is less as compared to low quality white wine (3 to 6). For Quality between 7-9 (considered as good quality wine) the concentration in white wine dominates red wine. The concentration of residual sugar is much higher in both low-quality white wine (3 to 6) and good quality white wine (7 to 9) as compared to red wine as shown in Fig. 4(d). As seen in Fig. 4(e), the concentration of Sulphates in low quality red wine (3 to 6 wine quality) is more than the concentration low quality white wine but as we move further towards good quality wine (7 to 9 quality) the concentration in red wine increases and the white wine remains the same.

The concentration of chlorides in quality 3 to 6 (considered as low quality wine) white wine is less than red wine (quality 3-6) and as we see when the quality improves(quality 7-9) concentration of red wine is dominated in good quality red wine(7 to 8) as shown in Fig. 4 (f).The concentration of Volatile acidity in low quality red wine (quality 3 to 6) is much higher than the volatile acidity in low quality white wine(3 to 6) and as the quality improves that is in quality 7 to 9 the concentration of volatile acidity is still dominated by red wine as seen in Fig. 4(g). A seen in Fig. 4(h), the concentration of alcohol in low quality white wine (quality 3 to 6) is more than low quality red wine (quality 3 to 6) but as quality improves (quality 7 to 9) the concentration of alcohol increases in both red and white wine and alcohol concentration becomes higher in good quality red wine (quality 7-9). As seen in Fig. 4(h), the concentration of alcohol in low quality white wine (quality 3 to 6) is more than low quality red wine (quality 3 to 6) but as quality improves (quality 7 to 9) the concentration of alcohol increases in both red and white wine and alcohol concentration becomes higher in good quality red wine (quality 7-9). The concentration of citric acid as seen in Fig.4(i) for wine quality between 3 to 6 (considered as low-quality wine) is observed higher in white wine as compared to red wine but as we move further towards good quality wine (quality 7-9) we can see that concentration of red wine is dominated over white wine.

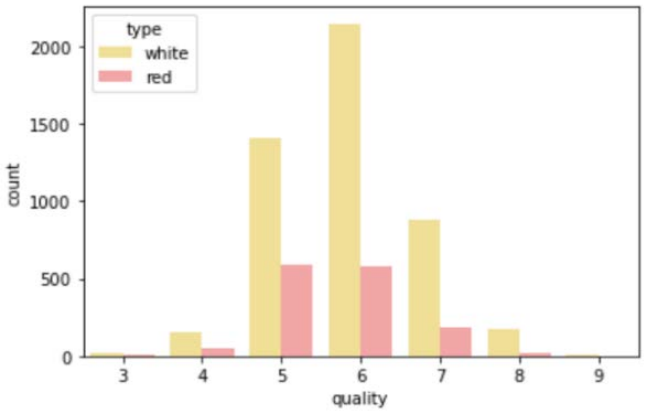


Fig.4. (a) The following figure depicts the number of each quality wine present in the dataset, yellow bar depicts white wine and red bar depicts red wine. It is clearly evident that in Quality 9 red wine is not present in our dataset

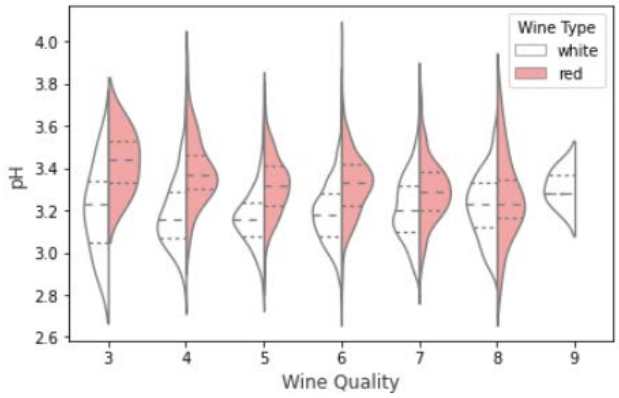


Fig.4. (b) Concentration of pH

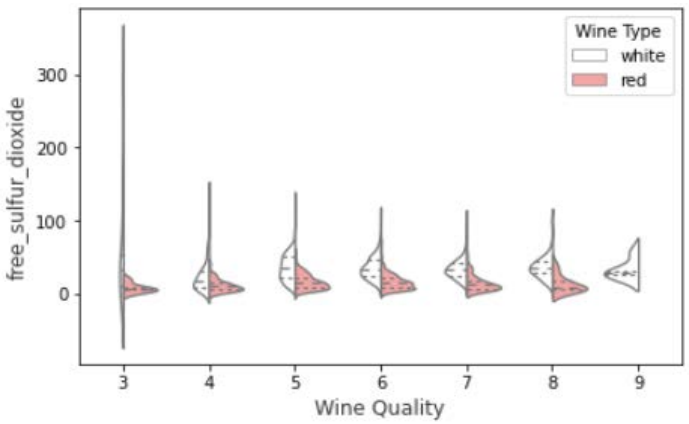


Fig.4. (c) Concentration of free sulfur dioxide

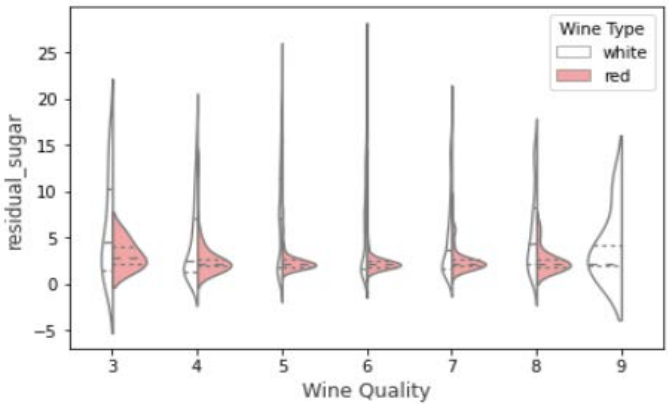


Fig.4. (d) Concentration of residual sugar

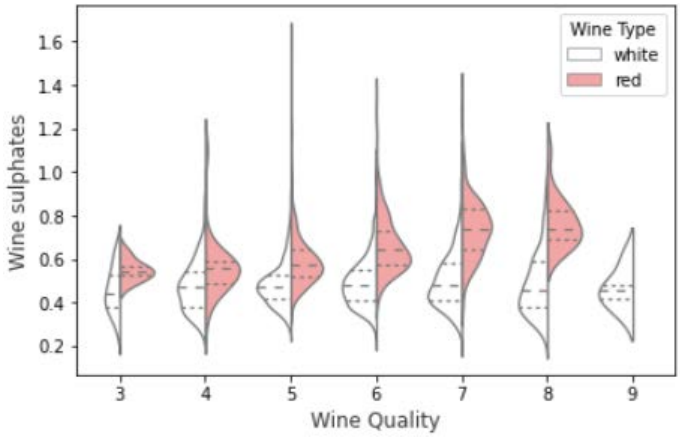


Fig.4. (e) Concentration of sulphates

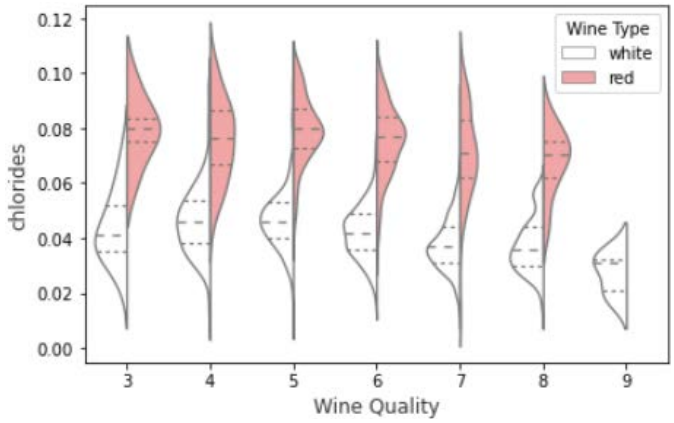


Fig.4. (f) Concentration of chlorides

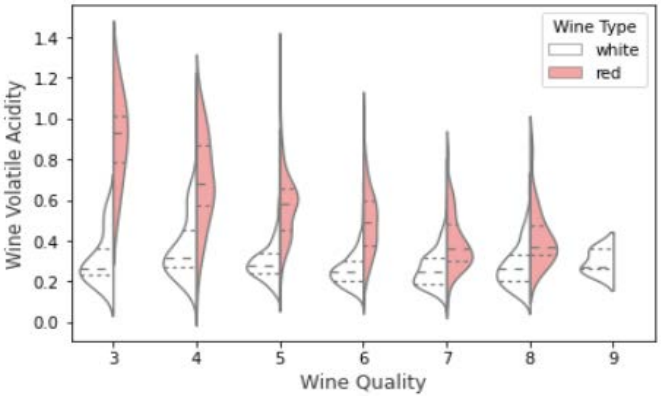


Fig.4. (g) Concentration of volatile acidity

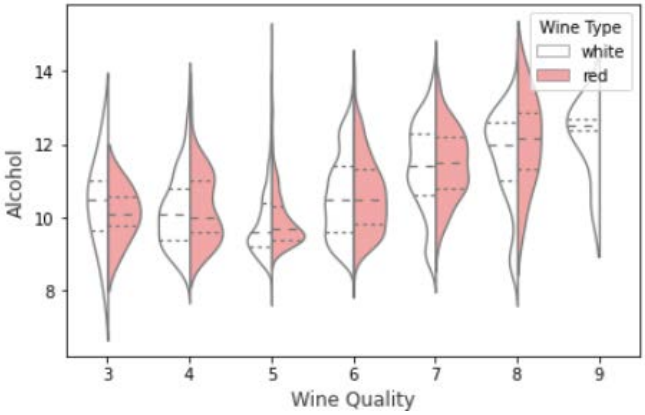


Fig.4. (h) Concentration of alcohol

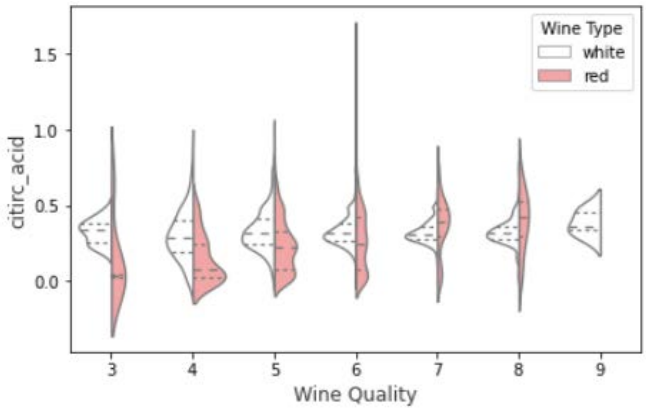


Fig.4. (i) Concentration of citric acid.

Table 7. Descriptive analysis of figure 4 representing the concentration of attributes in wine quality ranging from 3 to 9. High quality white wine ranges from 7 to 9 because the white wine falls in quality 9 and good quality red wine ranges from 7 to 8.

Attributes	Concentration in low quality red wine (3 to 6)	Concentration in low quality white wine (3 to 6)	Concentration in high quality red wine (7 to 8)	Concentration in high quality white wine (7 to 9)
pH	high	low	similar	similar
free_sulfur_dioxide	low	high	low	high
residual_sugar	low	high	low	high
sulphates	high	low	high	low
chlorides	high	low	high	low
volatile_acidity	high	low	high	low
alcohol	low	high	high	low
citric_acid	low	high	high	low

Hence, we conclude that the concentration of Ph, Sulphates, Chlorides, Volatile acidity, Alcohol and Citric Acid is more in good quality red wine whereas the concentration of Free sulfur Dioxide and Residual Sugar is more in good quality white wine as compared to red wine

Now answering the last research objective. According to the results depicted in Table 8, we can clearly see that “RFC” (Random Forest Classifier) shows the best performance with the highest accuracy i.e., 89.79% (for selected 9 attributes) as compared to the other three classifiers i.e., Decision tree classifier, k-nearest neighbor classifier and Sequential Model (ANN classifier). RFC shows a lower RMSE value than the other three classifiers, 0.40. Decision tree classifier has an accuracy of 83.54% and RMSE value of 0.60 whereas Sequential Model has accuracy of 77.19% and RMSE value of 0.45. Similarly, k-nearest neighbor classifier has an accuracy of 86.66% and RMSE value 0.54 which was showing lower performance as compared to Random Forest classifier.

Table 8. Accuracy performance for various classifiers

S.no	Classifiers	Accuracy for selected 9 attributes	RMSE for selected 9 attributes
1	RFC	89.79	0.40
2	Knn	86.66	0.54
3	DT	83.54	0.60

5. Conclusion

We have used a dataset which consists of data of two types of wine that is red and white wine of Portuguese "Vinho Verde", for predicting the wine quality. Pre-processing on the dataset was done using heatmap for observing Pearson coefficient value of each attribute with others. It was observed that “density” had a coefficient value of more than 0.7, with attributes “residual sugar” and “alcohol” which is considered as insignificant, so we removed the attribute “density”. Next, we applied Bivariate analysis after which we found that there exists a strong bond between our dataset numeric variable(attributes) and target variable(quality), Hence we eliminated one feature that had a minimum deviation 0.05% with our target variable(quality) the selected attribute was “total sulfur dioxide”. Further, using feature selection and extra trees classifier was used for ranking our attributes from least to most significant according to their ranking scores. The first least significant variable was “Fixed acidity”. The dataset size was reduced and optimized the performance of our model. For the selected final 9 attributes along-with quality being the predicted attribute, the accuracy of KNN classifier is 86.66% and RMSE value is 0.54, whereas the accuracy of Decision Tree Classifier is 83.54% and RMSE value is 0.60, Sequential model (ANN) has accuracy of 77.19% and RMSE value of 0.45. The best classifier turned out to be Random

Forest Classifier with accuracy of 89.76% and RMSE value of 0.40.

Looking towards the future, our methodology and techniques can be used to shorten wine datasets of similar type. Moreover, we can use our method for pre-processing any dataset using the combination of our technique by implementing bivariate analysis, correlation, and ranking.

References

[1] F. Balali, J. Nouri, A. Nasiri, and T. Zhao, “Data Analytics,” in Data Intensive Industrial Asset Management, Cham: Springer International Publishing, 2020, pp. 105–113.

[2] “Big Data Analytics,” IBM. [Online]. Available: https://www.ibm.com/analytics/big-data-analytics. [Accessed: 24-Apr-2022].

[3] Y. Er and A. Atasoy, "The Classification of White Wine and Red Wine According to Their Physicochemical Qualities", International Journal of Intelligent Systems and Applications in Engineering, vol. 4, no. Special Issue-1, pp. 23-26, Dec. 2016, doi:10.18201/ijisae.265954

[4] I.H. Sarker, M. H. Furhad, and R. Nowrozy, “AI-driven cybersecurity: An overview, security intelligence modeling and research directions,” SN Computer Science, vol. 2, no. 3,2021

[5] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[6] H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, “Cybersecurity data science: an overview from machine learning perspective,” J. Big Data, vol.7, no. 1, 2020.

[7] Marchand and P. Marx, Automated product recommendations with preference-based explanations, J. Retail., vol. 96, no. 3, pp. 328–343, 2020.

[8] V. Singh, S. Singh, and P. Gupta, “Real-time anomaly recognition through CCTV using neural networks,” Procedia Comput. Sci., vol. 173, pp. 254–263, 2020.

[9] Sarker, M. Hoque, M. Uddin and T. Alsanoosy, "Mobile Data Science and Intelligent Apps: Concepts, AI-Based Modeling and Research Directions", Mobile Networks and Applications, vol. 26, no. 1, pp. 285-303, 2020, doi: 10.1007/s11036-020-01650-z

[10] A.Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015.

[11] D. Tamburini, “Describe, diagnose, and predict with IoT Analytics,” Microsoft.com. [Online]. Available: https://azure.microsoft.com/en-in/blog/answering-whats-happening-whys-happening-and-what-will-happen-with-iot-analytics/. [Accessed: 24-Feb-2022]

[12] Y. Gupta, Selection of important features and predicting wine quality using machine learning techniques, Procedia Comput. Sci., vol. 125, pp. 305–312, 2018.

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

[14] P. Sidhu and M. Bhatia, "A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority", International Journal of Machine Learning and Cybernetics, vol. 9, no. 1, pp. 37-61, (2015) [Online]. Available: https://link.springer.com/article/10.1007/s13042-015-0333-x

[15] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, “Modeling wine preferences by data mining from physicochemical properties,” In Decision Support Systems, Elsevier, 47 (4):547-553. ISSN: 0167-9236.

[16] K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire, “Prediction of wine quality using machine learning algorithms,” Open J. Stat., vol. 11, no. 02, pp. 278–289, 2021.

[17] P. Appalasamy, A. Mustapha, N. Rizal, F. Johari and A. Mansor, "Classification-based Data Mining Approach for Quality Control in Wine Production", Journal of Applied Sciences, vol. 12, no. 6, pp. 598-601, 2012. Available: 10.3923/jas.2012.598.601.

[18] S. Lee, J. Park and K. Kang, "Assessing wine quality using a decision tree," 2015 IEEE International Symposium on Systems Engineering (ISSE), 2015, pp. 176-178, doi: 10.1109/SysEng.2015.7302752.

[19] S. Aich, A. A. Al-Absi, K. Lee Hui and M. Sain, "Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques," 2019 21st International Conference on Advanced Communication Technology (ICACT), 2019, pp. 1122-1127, Doi: 10.23919/ICACT.2019.8702017.

[20] L. Breiman, Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[21] X. Ying, "An Overview of Overfitting and its Solutions", Journal of Physics: Conference Series, vol. 1168, p. 022022, 2019. Available: 10.1088/1742-6596/1168/2/022022.

[22] B. Baranidharan, A. Pal and P. Muruganandam, "Cardio-Vascular Disease Prediction based on Ensemble Technique Enhanced using Extra Tree Classifier for Feature Selection", International Journal of Recent Technology and Engineering, vol. 8, no. 3, pp. 3236-3242, 2019.doi:10.35940/ijrte.C5404.098319.

[23] P. Geurts, D. Ernst, and L. Wehenkel, “Extremely randomized trees,” Mach. Learn., vol.63, no. 1, pp. 3–42, 2006.

[24] S. Walker, W. Khan, K. Katic, W. Maassen and W. Zeiler, "Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings", Energy and Buildings, vol. 209, p. 109705, 2020. Available: 10.1016/j.enbuild.2019.109705 [Accessed 24 Feb 2022].

[25] T. Chai and R. R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?, Geosci, Geosci. Model Dev. Discuss, vol. 7, pp. 1525–1534, 2014.

[26] Wikipedia contributors, “Kernel density estimation,” Wikipedia, The Free Encyclopedia,11 Apr. 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Kernel_density_estimation&oldid=1082123335.

[27] Ahmed Iqbal, Shabib Aftab, "A Feed-Forward and Pattern Recognition ANN Model for Network Intrusion Detection", International Journal of Computer Network and Information Security, Vol.11, No.4, pp.19-25, 2019.

[28] J. Hintze and R. Nelson, "Violin Plots: A Box Plot-Density Trace Synergism", The American Statistician, vol. 52, no. 2, pp. 181-184, 1998 [Online]. Available: https://www.jstor.org/stable/2685478

Authors' Profiles

Parneeta Dhaliwal was born in 1980.She is presently working as an Associate Professor in CST Department, Manav Rachna University, Faridabad. She has over 16 years of academics & research experience. She is currently serving as Head, Research Cluster of Computing Lab at Manav Rachna University. She has worked as full time Teaching & Research Fellow for four years at Netaji Subhas Institute of Technology, University of Delhi. She has been awarded Ph.D. for thesis in the area of "Machine Learning" from Netaji Subhas Institute of Technology, University of Delhi in 2017.She has published many research papers in international level journals with good indexing and high impact factor.

She has presented very high-quality research papers in National & International Conferences and also attended various workshops & FDPs. She is currently mentoring Ph. D students working in the area of Blockchain Technology and Digital Forensics. She has mentored many consultancy projects in latest technical areas for the industry. She has mentored many student teams in their projects at the State Level and National level Competitions. She is the University Coordinator for MRU Code Chef Campus Chapter. The chapter works towards improving the culture of competitive programming in the Campus and organizes various events, workshops, and contests from time to time.

She is a life member of Computer Society of India. She is currently serving as a reviewer with many reputed international journals and conferences. She has also contributed as Session chairs and as a Resource Person in reputed international conferences and FDPs, respectively.

Suyash Sharma was born in Kanpur, India on 13th April 1999.He is currently Pursuing B.Tech. In Computer Science from Manav Rachna University, Haryana 2018-2022.

He is a mentor in Research cluster computing in Manav Rachna University for Machine Learning.

Lakshay Chauhan was born in Delhi, India on 25th July 2000.He is currently Pursuing B.Tech. In Computer Science from Manav Rachna University, Haryana 2018-2022.

He is a mentor in Research cluster computing in Manav Rachna University for Machine Learning.

How to cite this paper: Parneeta Dhaliwal, Suyash Sharma, Lakshay Chauhan, "Detailed Study of Wine Dataset and its Optimization", International Journal of Intelligent Systems and Applications(IJISA), Vol.14, No.5, pp.35-46, 2022. DOI:10.5815/ijisa.2022.05.04