Wine Quality Prediction by Using Backward Elimination Based on XGBoosting Algorithm

Umer Zukaib[1]†, Mir Hassan[2]†, Tariq Khan[3]†, and Shoaib Ali[4]†,

[1]Department of Computer Science, Comsats University Islamabad, Abbottabad Campus, Pakistan
[2]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania
[3]Department of Information Engineering, University of Politecnico Delle Marche, Ancona, Italy
[4]Department of Computer Science, Virtual University, Pakistan

Abstract

Different industries mostly rely on quality certification for promoting their products or brands. Although getting quality certification, specifically by human experts is a tough job to do. But the field of machine learning play a vital role in every aspect of life, if we talk about quality certification, machine learning is having a lot of applications concerning, assigning and assessing quality certifications to different products on a macro level. Like other brands, wine is also having different brands. In order to ensure the quality of wine, machine learning plays an important role. In this research, we use two datasets that are publicly available on the “UC Irvine machine learning repository”, for predicting the wine quality. Datasets that we have opted for our experimental research study were comprised of white wine and red wine datasets, there are 1599 records for red wine and 4898 records for white wine datasets. The research study was twofold. First, we have used a technique called backward elimination in order to find out the dependency of the dependent variable on the independent variable and predict the dependent variable, the technique is useful for predicting which independent variable has maximum probability for improving the wine quality. Second, we used a robust machine learning algorithm known as “XGBoost” for efficient prediction of wine quality. We evaluate our model on the basis of error measures, root mean square error, mean absolute error, R2 error and mean square error. We have compared the results generated by “XGBoost” with the other state-of-the-art machine learning techniques, experimental results have showed, “XGBoost” outperform as compared to other state of the art machine learning techniques.

Keywords:
Deep learning, Data mining, Machine learning

1 Introduction

Wine is a product that is liked in different communities, but the quality of the wine is the most important factor. Most of the times consumers acquire quality certification in order to ensure the quality of their product [1]. In the modern era mostly consumers increase the market value of their product on behalf of the quality certification [2]. In this regard, wine quality also needs some sort of quality certification, and mostly the testing of the product will be done after the completion of manufacturing [3]. This is a tedious and time-consuming process and also acquire a lot of resources. As it needs a human expert to assess the product quality and although it is an expensive process. Quality certification by human experts is a challenging task as every person is having their own way of judgment [4].

Now a day’s most of the industries have adopted the latest technology and from time to time they are applying it in different areas, these methods are much more efficient to enhance and smooth the overall process for quality assessment and certification [5]. Machine learning has made this a bit easier to perform some sort of quality prediction and quality assessment regarding different product certification [6]. Likewise, the human assessment for predicting the quality of wine was replaced by some sort of modern machine learning techniques for the purpose of wine quality prediction [7]. Some machine learning techniques like support vector machine, random forest, k means clustering, neural network, simulated annealing, genetic algorithm, and gradient boosting were mostly used for predicting the quality of wine [8]. These techniques minimize the inference of human experts and automate quality assurance by categorizing the product [8].

Machine learning methods are helpful for assessing the quality of wine, these assessments are also useful for

Manuscript received February 5, 2024
Manuscript revised February 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.2.4

quality certification and assuring the quality of wine in different competitive markets [9]. There are different characteristics like PH value, density, alcohol, and some other acids. Generally, there are two types of test for quantifying the wine, the first is physiochemical, and the second is a sensory test. In the physiochemical test, there is no involvement of a human expert, and the test is performed in the lab. But in a sensory test, a human expert is needed. It is difficult to adopt physicochemical and sensory tests for wine quality assessment due to their complex analysis [3]. Different researchers have worked on wine quality assessment using machine learning, but still, there is a need for some improvements.

In this paper, we used two datasets that are publicly available on the "UC Irvine machine learning repository", for predicting the wine quality. Datasets that we have opted for our experimental research study were comprised of white wine and red wine datasets, there are 1599 records for red wine and 4898 records for white wine datasets. The research study was in twofold. First, we have used a technique called backward elimination in order to find out the dependency of the dependent variable on the independent variable and predict the dependent variable, the technique is useful for predicting which independent variable has maximum probability for improving the wine quality. Second, we used a robust machine learning algorithm known as "xgboost" for efficient prediction of wine quality. We evaluate our model on the basis of error measures, root mean square error, mean absolute error, R2 error and mean square error. We have compared the results generated by "XGBoost regressor" with the other state of the art machine learning techniques that are, support vector machine, random forest, k means clustering, simulated annealing, genetic algorithm, and XGBoot regressor for predicting the quality of wine, experimental results have showed, "XGBoost regressor" outperform as compared to other state of the art machine learning techniques.

Rest of the paper is organized as, section 2 describes the related work, section 2 describes the proposed methods, section 3 describes the results, section 4 describes the overall discussion about results and section 5 describes the conclusion and future work.

2 Related work

Many researchers have worked on wine quality prediction by using classical machine learning techniques. Atasoy and Er have used K-nearest neighbor, Random Forest and Support-Vector machine for classifying the quality of wine. For feature section they have used principal component analysis and achieved better accuracy on random forest model [10]. Chen et al. have proposed a human based survey for grading the wine, besides this they have also used hierarchical clustering technique and association rule mining for processing the reviews and predicting the wine quality and achieved 85 % accuracy results [11]. Applalasmy et al. have used physiochemical test-data for predicting the quality of wine. They have concluded, during the production process, the classification approach is helpful for grading the wine quality [12]. Baltran et al. have used Aroma chromatography for classifying wine quality, for feature selection they have opted wavelet transformation, principal component analysis for reducing the dimensionality, they have concluded, support-vector-machine with wavelet transform have achieved better results in term of accuracy as compared to other classifiers [13].

Thakkar et al. have used different machine learning techniques along with analytical hierarchical process for ranking the features, they have achieved 70 % accuracy on SVM and 66 % accuracy on random forest [14]. Reddy et al. have used red wine dataset for initial survey study, and also used centric clustering approach for product recommendation, used Gaussian distribution for assigning weights and finalized the quality of wine on the basis of user-preference group [15]. S. Kumar and N. Mandan have worked on red wine quality prediction, for their research study they have opted three machine learning techniques, naïve Bayes, SVM and random forest, they have separately trained and then tested these techniques and record their outcomes, SVM have scored best accuracy of 67.25 % as compared to naïve Bayes and random forest [16]. B. Shaw and Ak. Suman have proposed a comparative study for predicting the wine quality, for that purpose they applied different machine learning models and lastly compared their results, the used models were SVM, multi-layer perceptron and random forest, but SVM have achieved better results as compared to rest of the techniques [7].

Sun et al. have used neural network for predicting the six globally wine origins, they feed neural network with 15 input-variables. They have collected experimentation dataset from Germany and achieved 100 % prediction results [17]. Vlassides et al. have used neural-network for classifying the Californian wine dataset. They have classified the wine quality on the basis of grapes maturity. They have used 36 samples for experimentation and achieved 6 % error [18]. Moreno et al. have used probabilistic-neural-network for classifying the two red wine-dataset with 54 samples [19]. Yu et al. have used experimental data comprised of 147 bottle that contains rice wine and used spectral measurements technique for predicting three categories from rice wine [20]. Beltran et al. have used Chilean wine for experimental study and used SVM and neural-network for classifying the data, and for feature selection they have opted linear-discriminate-analysis [13].

Y. Gupta have worked on wine quality prediction, used machine learning techniques in two folds, first used linear regression in order to find out the dependency between features, secondly used most significant features for predicting the dependent variables, the techniques they have opted were SVM, neural network and linear regression, their results have demonstrated that, significant features are useful for efficient prediction of wine quality [6]. S. Llic and S. Pitulic have used data-mining techniques for studying the wine quality, on the basis of physiochemical-properties of wine made prediction of wine quality, for that purpose different machine-learning techniques they have used, made comparisons between different methods but random forest give satisfactory results as compared to other machine-learning methods [21].

S. Aich and A Absi have worked on wine quality-prediction using machine-learning techniques, they have adopted different techniques for feature selection, the first one was recursive-feature-elimination and second was principal-component-analysis, and used decision tree for making prediction, they have also generated results by using random-forest, as compared to features selection by using decision tree, random forest gave efficient results on same feature selection-techniques [22]. Besides machine learning techniques, several researchers have adopted physiochemical test in order to predict wine-quality. Ashen Filter have conducted research study for wine quality, he mentioned that, wine quality directly related to the quality of grapes that were used during production, and based on grapes quality be predicted the price of wine, he proposed a equation for estimating the wine price, the major factors that were involved for deciding the price were climate change, grapes quality and expert opinion [23].

Riberio et al. used data mining techniques for predicting wine quality, the used machine learning techniques were, ANN, decision tree, linear regression, by using these techniques they have predicted the organoleptic-parameters, on data-mining tools they got efficient results in terms of accuracy [24]. Lee et al, conducted experimental research study in order to predict wine quality, their work was in two fold, first was to use decision tree and secondly used WEKA tool for generating results on some of the machine learning models SVM, Multi-layer perceptron and Bayes net, it was found that decision tree achieved better results compared to other techniques [25]. Yeo et al used multi-task based learning technique and Gaussian-regression process for prediction of wines quality, for their research study wine's historical price data was used, it was concluded that machine learning have a great potential for wines prediction [26].

3 Materials and Methods

3.1 Dataset Description

We have selected wine dataset, there are two subtypes of wine dataset, and the first one is white wine and the second is red wine dataset. The number of samples in white wine dataset are 1599 in numbers and number of samples in red wine dataset are 4898 in numbers. And collectively there are 12 physiochemical variables. That are fixed acidity, citric acid, volatile acidity, residual sugar, free sulfur dioxide, chlorides, density, pH, alcohol, total sulfur dioxide, sulphates, and quality rating. Out of these 12 variables 11 are independent variables and 1 is dependent variable, all others are independent except quality rating. We can say, on the basis of independent variables we are going to predict dependent variable by using machine learning techniques. PH values indicate the how much a wine is acidic or basic, range from 0 to 14.

The ultimate goal is to predict the quality rating of wines that are assigned by a quality expert like acidity or alcohol composition. Due to logistic and privacy issues we just have information about physiochemical and sensory variables. We does not have any information about types of grapes, wine brank or selling prices.

3.2 Machine Learning Models

3.2.1 SVM

Support vector machine is the well-known technique used for classifying and regressing the data. SVM is efficient to construct a hyper-plane in a high dimensional space. And helpful for solving classification or regression related problems. If there is a larger distance between the nearest data points we can say that the hyper-plane has achieved the better results. In simple words we can say that if there is a larger distance between data points the generalization error of the classifier might be minimum.

The closest data points passing through the plane is known as support vector. And the distance between optimal hyper plane and hyper plane is called margin. Simply said, margin is the area which does not contain any data point. In some cases the data points exists in margin area, but generally there are minor cases in which data points really exists in margin area. The main goal is choose an optimal hyper plane, the hyper plane in which there is a highest distance from closest data-points. The margin will be smaller if there is minimum distance between data points, it is consider to be the best generalization practice for training the data. But, when some sort of unseen record or data will come the model fails to generalize well as we have already discussed. The main goal is to maximize the margin such that the classifier will be able to generalize the seen record in efficient manner [27].

3.2.2 Random Forest

Random forest is a well-known machine learning technique that is used to handle regression and classification problem in a quite well manner. It is ensemble based technique that actually combine many classifiers in order to provide best optimal solution to the complex problem. Random forest consist of different numbers of decision trees and generate a forest. The algorithm of random forest is trained by the bootstrap aggregation or bagging. These are called meta-algorithms that are used for improving the accuracy of machine-learning algorithm. The model generate their results on the basis of prediction made by decision tree. It predicts the resultant values by averaging the output results from different trees. Increasing the number of trees may result in increasing the precision of predicted outcomes.

Some of the main features of random forest are, it is quite efficient and accurate technique as compared to decision tree. It can handle missing values in an efficient way. Without the hyper parameter tuning the model can made reasonable predictions. The issues related to over-fitting can easily be solved by decision trees.

Actually the decision trees are the basic building block of random forest, it is a decision supporting technique that makes a tree like-structure. It consists of three components, the first one is decision node, second is leaf node and third is root node. The algorithm divides the input data into branches that further divide into other branches. The sequence remain continuous until a new leaf node is generated. The nodes indicate the attribute for predicting the output. Decision node generate a linkage to the leaves, the leaf node represents the final outcome [28].

3.2.3 K-means clustering

K-means clustering is the un-supervised based machine learning technique, un-supervised algorithm makes inference from the input data without knowing any label information. The aim of K-means is a bit simple, grouping the similar input data pints and discovering the underlying information. To achieve the objective, K-mean generate fixed number of clusters from the input data. A cluster is generated by the collection of data points on the basis of certain similarities. A number K is defined, that refers to the centroid in the input data, and centroid indicates the center of cluster. In K-mean clustering there are K numbers of centroids that indicates the center of each cluster, generally, the best practice is to keep minimum number of centroids. In K-mean algorithm, the term “mean” refers to calculating the mean of input data points in order to find the centroid.

Initially in the K-mean algorithm, the first group of centroid is randomly selected, which is consider to be the starting point for each cluster and later on iteration is performed for optimizing the position of centroid. It halts optimizing and creating new clusters when the centroids have been stabilized, means when there is no change in the values it means that the cluster is successful, secondly when the defined number of iterations have made then it halts [29].

3.2.4 Simulated Annealing

Annealing is the process in metallurgy, where a metal is slowly cooled down in order to attain a state of low-energy where it is consider to be more robust. Simulated annealing is an optimization algorithm that is described in terms of thermo-dynamics. Simulated annealing is the process, in which the temperature is slowly reduced that is started from a random high temperature and it becomes pure greedy descent when it approaches to the zero temperature. The algorithm maintains a current assignment of variables, on each step it randomly picks a value from a random variable. If the assignment of that value to the selected variable is an improvement and it does not increase the conflict, the algorithm of simulated annealing accept that assignment. Otherwise it tries to accept the assignment in different probabilities, it depends on the temperature and worseness of the current assignment, and if the change is unaccepted the current assignment might be unchanged [30].

3.2.5 Genetic Algorithm

Genetic Algorithm is the optimization based algorithm that is used to solve many complex problems that takes a long for their solution. There are many real-world applications of Genetic Algorithm like data centers, code breaking, circuit designs, artificial creativity and image processing. There are some of the major terms used in Genetic Algorithm.

First one is population, it is the subset of all possible solutions that refers to the solution of the problem.

Second one is chromosome, it is also one of the solution in the population.

Next is Gene, gene is the component of chromosome.

Allele, this is the specific value given to the gene for a specific-chromosome.

Fitness function, this is a function that contains a specific value for input in order to generate an improved output.

Genetic operators, in Genetic Algorithm, some of the best individual mate and reproduce such an offspring that are quite better than their parents. The purpose of using the Genetic operators is, to change the genetic composition for their next generation.

Genetic Algorithm is a heuristic, search based optimization algorithm that is subset of evolutionary algorithm. It use the natural selection and the concept of genetics for providing the solution to the problem. Algorithm is based on structure of genes and the behavior of chromosomes, where every chromosome provides a possible solution, fitness function is responsible for providing the characteristics to all the individuals with in the population. Let's see how Genetic Algorithm work.

3.2.5.1 Initialization

The algorithm starts from the initial population. The initial populations contains all the possible solutions to the problem. The common method used for initialization is using the random-binary string.

3.2.5.2 Fitness assignment

Fitness function is helpful for assigning fitness score to all the individuals in the population. On the basis of fitness the individual have chosen for reproduction. Higher the fitness score, higher chances for reproduction.

3.2.5.3 Selection

Individuals are selected for producing the offspring, individuals pass their genes to the next generation. The main purpose of selection phase is to ensure the maximum chances of generating best optimal solution to the problem.

3.2.5.4 Reproduction

This phase is the actual creation of new population. There are two main operators, cross over and mutation. In the cross-over the genetic information of parents were swapped in order to reproduce offspring. But in mutation a new genetic information is added to reproduce offspring. In mutation some of the bits of chromosomes were flipped and it is much helpful to solve local minima and also enhancing the diversification.

3.2.5.5 Replacement

This is also known as generational replacement, in which the old population is being replaced by new child population. The newly generated population contains high fitness score as compared to older population.

3.2.5.6 Termination

After the replacement, the last step is termination. Algorithm terminates after achieving the certain threshold that might be in the form of fitness score.

3.2.6 Gradient boosting

Gradient boosting is consider to be the robust model used for the purpose of classification and regression problems. The basic idea behind the boosting is to modify a weak learner in order to achieve a better learnable model. A weak learner is the one, whose performance is less than a random chance. So we can say boosting is the efficient model a weak hypothesis into better hypothesis.

The term gradient boosting means, adding a weak learner like gradient decent in order to minimize the loss of the model. Let's talk about how a gradient boosting model works. There are three basic steps involved. First step is the optimization of loss function, second is, need a weak learner for making the prediction. Third is, there should be an additive model that will add weak learner in order to minimize the loss function. There three steps are the most important elements for gradient boosting. Opting the loss function may depend on the type of problem. Different loss functions may use like squared error for regression problem and logarithmic for classification purpose. In gradient boosting, decision trees are used for purpose of weak learner. Regression trees are used that generate the output result in the form of real values. Usually trees are constructed in the greedy manner, on the basis of purity score such as minimizing the loss or Gini. Usually trees are added just one time and in the model the existing trees remains unchanged. During the addition of tree, gradient descent is used for minimizing the loss [31].

3.3 Feature selection

We have red wine and white wine datasets and first of all we select the features from these datasets, for the purpose of feature selection we are using backward elimination with gradient boosting regression.

3.3.1 Backward elimination using Gradient boosting Regression

We need to find an optimal metrics of features that contains independent variables that are statistically significant for dependent variable (wine quality in our case). Optimal metrics only contains the independent variables that have high impact on the quality of wine. We are using backward elimination that contains all the independent variables at first and we remove all the independent variables that are not statistically significant.

3.3.1.1 Steps for backward elimination

The first step is to select a significance level (SL) to stay in the model (we need to select a significance level, so that if the P-value of independent variable is below the significant level then the independent variable will stay in the model, but on other hand if the P-value of independent variable is above the significance level then it will not stay in the model and we will remove it) we are choosing (SL = 0.05).

The second step is to fit the model with all the possible predictors (in our case, all the independent variables used for predicting the wine quality).

Third step is, consider the predictor with highest P-value. If P > SL then we will remove that independent variable whose P-value is greater than SL.

Fourth step is to remove the predictor (independent variable with value greater P-value than SL).

Fifth and the last step is fit the model without this variable.

Those independent variables that are having P-value < 0.05 are consider to be the most powerful predictors for predicting the wine quality. Finally, if we follow the backward elimination, the optimal team of independent variables that can predict the quality of wine with highest statistical significant.

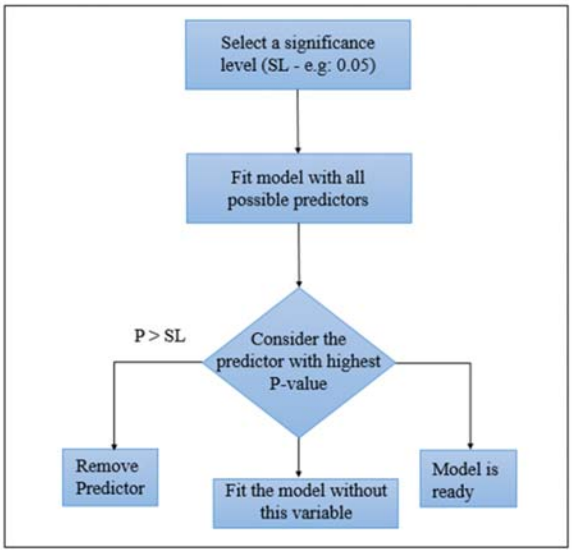

Select a significance level (SL - e.g: 0.05)
Fit model with all possible predictors
Consider the predictor with highest P-value
P > SL
Remove Predictor
Fit the model without this variable
Model is ready

Fig 1: Steps for backward elimination

3.3.1.2 Backward elimination for red wine dataset

Table 1: P-values for different independent variables before applying backward elimination

Indep Variables	Coeff	Std error	t	P > |t|
Fixed Acidity	0.0042	0.016	0.255	0.799
Volatile Acidity	-1.0997	0.120	-9.157	0.000
Citric Acid	-0.1841	0.147	-1.251	0.211
Residual sugar	0.0071	0.012	0.587	0.557
Chlorides	-1.9114	0.418	-4.575	0.000
Free sulfur-dioxide	0.0045	0.002	2.102	0.036
Total sulfur-dioxide	-0.0033	0.001	-4.565	0.000
Density	4.5291	0.625	7.243	0.000
PH	-0.5229	0.160	-3.268	0.001
Sulphates	0.8871	0.111	8.006	0.000
Alcohol	0.2970	0.17	17.217	0.000

In the Table 1. We calculated P-value for all the independent variables, in order to find the best optimal metrics (independent variables that are statistically significant for predicting wine quality). After analyzing the P-value the independent variables that have P-value greater than 0.05 are going to remove one by one. We need to find optimal metrics that only contains P-value less than significant level (SL = 0.05) means P < 0.05, after analyzing the table we came to know that first of all we need to remove Fixed Acidity, because it contains the P-value == 0.799 that is greater than 0.05, we follow the steps discussed in the fig 1. After removing Fixed Acidity we again calculate the P-value for all the remaining independent variables. After analyzing the new P-values for all predictors we came to know that the independent variable named (residual sugar) contains P-value == 0.547 that is greater than 0.05 (SL) we remove residual sugar and again follow the same procedure then we find the citric acid contains P-value == 0.209 that need to be removed (p > 0.05). Again we follow the same steps for backward elimination and calculate P-values for rest of the predictors and this time we don't get any predictor that have P-value greater than 0.05 (SL). Hence we have achieved the optimal metrics.

Table 2: P-values for different independent variables after applying backward elimination

Indep Variables	Coeff	Std error	t	P > |t|
Volatile Acidity	-1.0997	0.120	-9.157	0.000
Chlorides	-1.9114	0.418	-4.575	0.000
Free sulfur-dioxide	0.0045	0.002	2.102	0.017
Total sulfur-dioxide	-0.0033	0.001	-4.565	0.029
Density	4.5291	0.625	7.243	0.032
PH	-0.5229	0.160	-3.268	0.000
Sulphates	0.8871	0.111	8.006	0.012
Alcohol	0.2970	0.17	17.217	0.008

We find the best optimal metrics that consist of those independent variables which are statistically significant for predicting the wine quality. We used backward elimination in order to find the optimal team of independent variables that have highest statistical significance for predicting the wine quality. After analyzing the Table 2, we have concluded, there are total eight independent variables that are having P-value < 0.05 and consider to be the team of best optimal variables for predicting the quality of red wine. The independent variables are volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, density, PH, Sulphates and alcohol. The P-values of some independent variables that equals to 0.000, is not totally zero but it is a minor or smaller value but not zero.

3.3.1.3 Backward elimination for white wine dataset

In order to find the best optimal metrics of independent variables that are statistically significance for predicting the quality of white wine. We just focus on optimal metrics that only contains independent variables that have high impact on wine quality (white wine). We follow the same procedure of backward elimination as we have followed for red wine quality prediction. We find all the independent variables that are statistically significant for wine quality prediction and remove all other independent variables that are not statistically significant.

Table 3: P-values for different independent variables before applying backward elimination

Indep Variables	Coeff	Std error	t	P > |t|
Fixed Acidity	-0.0506	0.015	-3.356	0.001
Volatile Acidity	-1.9585	0.114	-17.196	0.000
Citric Acid	-0.0293	0.096	-0.305	0.760
Residual sugar	0.0250	0.003	9.642	0.000
Chlorides	-0.9426	0.543	-1.736	0.083
Free sulfer-dioxide	0.0048	0.001	5.710	0.000
Total sulfer-dioxide	-0.0009	0.000	-2.352	0.019
Density	2.0420	0.353	5.780	0.000
PH	0.1684	0.084	2.014	0.044
Sulphates	0.4165	0.097	4.279	0.000
Alcohol	0. 3656	0.011	32.880	0.000

The Table 3, shows all the P-values for different independent variables before applying the backward elimination, we set the significant level = 0.05, for backward elimination we just consider those independent variables whose P-value is less than significant level (P < 0.05). We remove all the independent variables that are having p-value greater than 0.05 (P > 0.05). First of all we remove the citric acid that is having P-value = 0.76 (p > 0.05), then we follow the same procedure for backward elimination that is depicted in fig 1. Again we calculate P-values for all remaining variables and found chlorides with highest P-value and remove it. Again we calculate P-values and this time we do not find any other independent variable that is having P-value < 0.05.

Table 4: P-values for different independent variables after applying backward elimination

Indep. Variables	Coeff	Std error	t	P > |t|

Fixed Acidity	-0.0506	0.015	-3.356	0.001
Volatile Acidity	-1.9585	0.114	-17.196	0.012
Residual sugar	0.0250	0.003	9.642	0.000
Free sulfer-dioxide	0.0048	0.001	5.710	0.000
Total sulfer-dioxide	-0.0009	0.000	-2.352	0.014
Density	2.0420	0.353	5.780	0.021
PH	0.1684	0.084	2.014	0.027
Sulphates	0.4165	0.097	4.279	0.037
Alcohol	0. 3656	0.011	32.880	0.000

After analyzing the Table 4, we conclude that for predicting the quality of white wine, we follow the backward elimination in order to find the team of best optimal independent variables that have highest statistical significance for predicting wine quality. The strongest impact is actually composed of nine independent variables that happens to be fixed acidity, volatile acidity, residual sugars, free sulfur dioxide, total sulfur dioxide, density, PH, Sulphates and alcohol.

4 Results and Discussion

We are using various machine learning models in order to generate the results for wine quality prediction. We have used backward elimination for feature selection and also used two separate techniques principal component analysis (PCA) and linear discriminant analysis (LDA) for feature extraction. We have generated the results by using different machine learning techniques and also used PCA and LDA to generate the results on the same techniques. And finally make comparisons of the results generated by PCA and the results generated by LDA techniques.

Table 5: Machine learning results generated on White wine dataset

Technique	**PCA**				**LDA**			
	RMSE	**MSE**	**MAE**	**R2**	**RMSE**	**MSE**	**MAE**	**R2**
SVM	0.955	0.91	0.44	0.69	0.871	0.76	0.58	0.19
Random Forest	0.993	0.99	0.61	0.42	0.823	0.61	0.46	0.31
K-means Clustering	1.041	1.08	0.72	0.33	0.936	0.88	0.61	0.21
Genetic Algorithm	0.934	0.63	0.51	0.73	0.821	0.68	0.46	0.63
Simulate Annealing	1.409	1.61	0.98	1.08	1.335	1.02	0.69	0.21
XGBoost	**0.6422**	**0.41**	**0.34**	**0.28**	**0.721**	**0.56**	**0.39**	**0.14**

Table 5 have showed results generated by different machine learning models on the white wine dataset with respect to PCA and LDA. The machine learning models that we have used, SVM, Random Forest, Neural Network, K-means Clustering, Genetic Algorithm, Simulate Annealing and XGBoosting. We have also recorded the results in term of two feature extraction techniques that are PCA and LDA, finally we have compared the results in term of mean square error, mean absolute error, root mean square error and R square error. On white wine dataset XGBoosting model performed well with least errors as compared to rest of the techniques. The comparisons of PCA and LDA with respect to XGBoosting model demonstrated the LDA technique showed better results as compared to PCA. On white wine dataset, in term of PCA, XGBoosting score 0.642 % RMSE, 0.41 % MSE, 0.34 % MAE and 0.28 % R2 error. On the other size same XGBoosting model in terms of LDA score 0.721 % RMSE, 0.56 % MSE, 0.39 % MAE and 0.14 % R2 error.

Table 6: Machine learning results generated on red wine dataset

Technique	**PCA**				**LDA**			
	RMSE	**MSE**	**MAE**	**R2**	**RMSE**	**MSE**	**MAE**	**R2**
SVM	0.796	0.63	0.51	0.14	0.682	0.51	0.58	0.19
Random Forest	0.815	0.67	0.50	0.16	0.663	0.54	0.61	0.23
K-means Clustering	0.842	0.71	0.57	0.24	0.728	0.53	0.65	0.07
Genetic Algorithm	0.982	0.63	0.69	0.41	0.856	0.49	0.58	0.32
Simulate	1.192	1.20	0.80	1.01	0.970	0.51	0.55	0.24

Anneal ing								
XGBo ost	**0.38 6**	**0.3 2**	**0.1 1**	**0. 10**	**0.57 0**	**0.4 5**	**0.1 9**	**0. 17**

Table 6 have showed results generated by different machine learning models on the red wine dataset with respect to PCA and LDA. The machine learning models that we have used, SVM, Random Forest, Neural Network, K-means Clustering, Genetic Algorithm, Simulate Annealing and XGBoosting. We have also recorded the results in term of two feature extraction techniques that are PCA and LDA, finally we have compared the results in term of mean square error, mean absolute error, root mean square error and R square error. On white wine dataset XGBoosting model performed well with least errors as compared to rest of the techniques. The comparisons of PCA and LDA with respect to XGBoosting model demonstrated the LDA technique showed better results as compared to PCA. On red wine dataset, in term of PCA, XGBoosting score 0.386 % RMSE, 0.32 % MSE, 0.11 % MAE and 0.10 % R2 error. On the other size same XGBoosting model in terms of LDA score 0.570 % RMSE, 0.45 % MSE, 0.19 % MAE and 0.17 % R2 error.

4.1 Comparisons of results generated by different techniques on white wine dataset

On white wine dataset we have applied different machine learning models, SVM, Random Forest, Neural Network, K-means Clustering, Genetic Algorithm, Simulate Annealing and XGBoosting. And we have also visualized their results in tabular form and also in graphical form. The comparisons of different models were made on the basis of errors measure, the model which score least errors were considered to be the better model for predicting white wine quality. From the range of different machine learning models XGBoosting model score least errors in term of RMSE, MSE, MAE and R2 error. The overall comparisons of their results were clearly depicted in Fig 2.

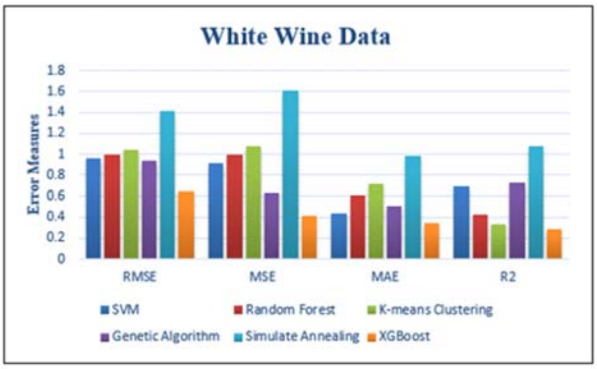


Fig 2: Errors measure on white wine dataset

Fig 2. Demonstrated results generated by using different machine learning techniques with respect to their error measures. The XGBoost model score least errors as compared to rest of the models.

4.2 Comparisons of results generated by different techniques on red wine dataset

On red wine dataset we have also applied machine learning models like, SVM, Random Forest, Neural Network, K-means Clustering, Genetic Algorithm, Simulate Annealing and XGBoosting. And we have visualized their results in tabular form and also going for graphical representation. The comparisons of different models were made on the basis of errors measure, the model which score least errors were considered to be the better model for predicting white wine quality. From the range of different machine learning models XGBoosting model score least errors in term of RMSE, MSE, MAE and R2 error. The overall comparisons of their results were clearly depicted in Fig 3.

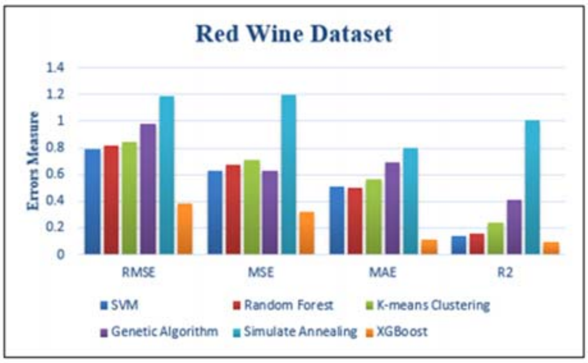


Fig 3: Errors measure on red wine dataset

Fig 3. Demonstrated results generated by using different machine learning techniques with respect to their error

measures. The XGBoost model score least errors as compared to rest of the models.

4.3 Overall comparisons of result with respect to PCA and LDA for both red and white wine datasets

We have also compared the results generated by different machine learning techniques on red and white wine datasets with respect to PCA and LDA techniques for feature extraction. The tabular representation have already mentioned in Table 5 and Table 6. Now we are depicting the graphical representation of different machine learning models with respect to PCA and LDA.

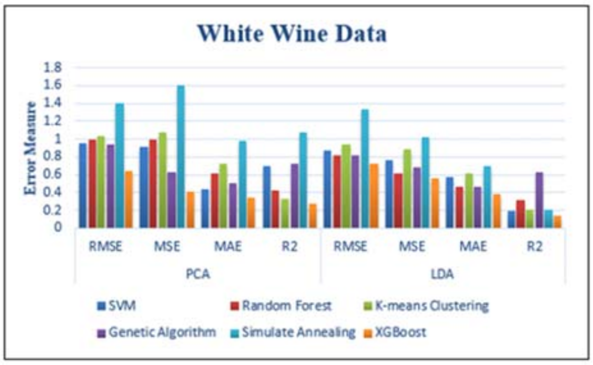


Fig 4: Comparisons of PCA and LDA on white wine dataset

Fig 4. Showed the results generated by different machine learning models on white wine dataset with respect to PCA and LDA (feature extraction techniques). The comparisons were made on the basis of error measures, RMSE, MSE, MAE and R2 error. It is clearly depicted the XGBoost score least errors with respect to LDA technique.

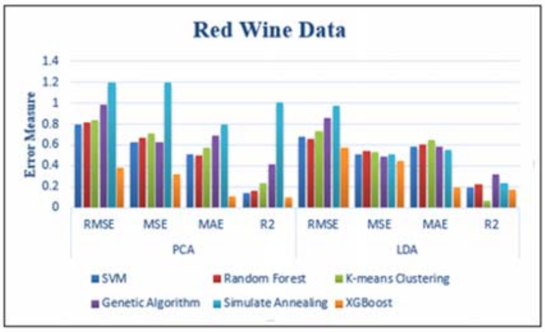


Fig 5: Comparisons of PCA and LDA on white wine dataset

Fig 5. Showed the results generated by different machine learning models on red wine dataset with respect to PCA and LDA (feature extraction techniques). The comparisons were made on the basis of error measures, RMSE, MSE, MAE and R2 error. It is clearly depicted the XGBoost score least errors with respect to LDA technique.

4.4 Discussion

For predicting the wine quality we have used red and white wine datasets. For feature selection a famous technique called "backward elimination" was used. Backward elimination is useful for selection of most significant features, means the independent variables those have significant impact on dependent variables were opted by using backward elimination. After selecting the most significant features from red and white wines dataset, we have used different machine learning techniques for the prediction of wine quality, the machine leaning techniques were support vector machine, random forest, k means clustering, neural network, simulated annealing, genetic algorithm and XGBoost regressor. We have used two different techniques for feature extraction that were PCA and LDA. The outcome of different techniques were evaluated on the basis of RMSE, MSE, MAE and R2 error. We have separately used PCA and LDA for generating results on all the mentioned machine learning techniques. Lastly we have compared the results in two folds, first compare the overall results generated by different techniques on the basis of least errors, however XGBoost regressor have scored least error and consider to be the best technique for predicting the wine quality. Secondly we have also compared the results generated by both of the feature extraction techniques (PCA and LDA) for different machine learning models. The comparison of LDA and PCA with respect to all the machine learning techniques (specifically XGBoost regressor) have showed, PCA have achieved least errors and considered to be the best feature extraction for predicting the wine quality.

5 Conclusion

Feature selection is the most crucial step in machine learning and data mining. Because if most significant features were selected, it will reasonably improve the prediction result of machine learning models. From our research study it is concluded that backward elimination is the best technique for feature selection, After successful selection of most significant features, next step is to adopt feature extraction technique, we have opted two separate techniques and finally compared their results, from our research study it is concluded that PCA (feature extraction) is consider to be the best feature extraction technique for efficient prediction of wine quality on both the datasets (red and white wine). Although the final evaluation were made on the basis of least errors, XGBoost regressor score least errors on both of the feature extraction (PCA and LDA) with respect to other machine learning techniques but when we made comparison between PCA and LDA with respect to XGBoost regressor, PCA our performs and score least errors (RMSE, MSE, MAE and R2 error) and consider to be the best technique for predicting the wine quality.

Bibliography

1. Botonaki, A., et al., *The role of food quality certification on consumers' food choices.* British Food Journal, 2006.
2. Corduas, M., L. Cinquanta, and C. Ievoli, *The importance of wine attributes for purchase decisions: A study of Italian consumers' perception.* Food Quality and Preference, 2013. **28**(2): p. 407-418.
3. Cortez, P., et al., *Modeling wine preferences by data mining from physicochemical properties.* Decision support systems, 2009. **47**(4): p. 547-553.
4. Veale, R. and P. Quester, *Consumer sensory evaluations of wine quality: The respective influence of price and country of origin.* Journal of wine economics, 2008. **3**(1): p. 10-29.
5. Kupis, J., et al., *Assessing the usability of the automated self-administered dietary assessment tool (ASA24) among low-income adults.* Nutrients, 2019. **11**(1): p. 132.
6. Gupta, Y., *Selection of important features and predicting wine quality using machine learning techniques.* Procedia Computer Science, 2018. **125**: p. 305-312.
7. Shaw, B., A.K. Suman, and B. Chakraborty, *Wine Quality Analysis Using Machine Learning*, in *Emerging Technology in Modelling and Graphics*. 2020, Springer. p. 239-247.
8. Gupta, U., et al., *Wine quality analysis using machine learning algorithms*, in *Micro-Electronics and Telecommunication Engineering*. 2020, Springer. p. 11-18.
9. Tingwei, Z. *Red wine quality prediction through active learning*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.
10. Er, Y. 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, 2016: p. 23-26.
11. Chen, B., et al. *Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel*. in *2014 IEEE International Conference on Data Mining Workshop*. 2014. IEEE.
12. Appalasamy, P., et al., *Classification-based data mining approach for quality control in wine production.* Journal of Applied Sciences, 2012. **12**(6): p. 598-601.
13. Beltrán, N.H., et al., *Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer.* IEEE Transactions on Instrumentation and Measurement, 2008. **57**(11): p. 2421-2436.
14. Thakkar, K., et al., *AHP and Machine Learning Techniques for Wine Recommendation.* International Journal of Computer Science and Information Technologies, 2016. **7**(5): p. 2349-2352.
15. Reddy, Y.S. and P. Govindarajulu, *An Efficient User Centric Clustering Approach for Product Recommendation Based on Majority Voting: A Case Study on Wine Data Set.* IJCSNS, 2017. **17**(10): p. 103.
16. Kumar, S., K. Agrawal, and N. Mandan. *Red Wine Quality Prediction Using Machine Learning Techniques*. in *2020 International Conference on Computer Communication and Informatics (ICCCI).* 2020. IEEE.
17. Sun, L.-X., K. Danzer, and G. Thiel, *Classification of wine samples by means of artificial neural networks and discrimination analytical methods.* Fresenius' journal of analytical chemistry, 1997. **359**(2): p. 143-149.
18. Vlassides, S., J.G. Ferrier, and D.E. Block, *Using historical data for bioprocess optimization: modeling wine characteristics using artificial neural networks and archived process information.* Biotechnology and Bioengineering, 2001. **73**(1): p. 55-68.
19. Moreno, I.M., et al., *Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks.* Talanta, 2007. **72**(1): p. 263-268.
20. Yu, H., et al., *Prediction of enological parameters and discrimination of rice wine age using least-squares support vector machines and near infrared spectroscopy.* Journal of agricultural and food chemistry, 2008. **56**(2): p. 307-313.
21. Radosavljević, D., S. Ilić, and S. Pitulić, *A DATA MINING APPROACH TO WINE QUALITY PREDICTION.*
22. Aich, S., et al. *A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques*. in *2018 20th International conference on advanced communication technology (ICACT)*. 2018. IEEE.
23. Ashenfelter, O., *Predicting the quality and prices of Bordeaux wine.* Journal of Wine Economics, 2010. **5**(1): p. 40-52.
24. Ribeiro, J., et al. *Wine vinification prediction using data mining tools*. in *ECC'09 Proceedings of the 3rd international conference on European computing conference. Computing and Computational Intelligence. WSEAS*. 2009.
25. Lee, S., J. Park, and K. Kang. *Assessing wine quality using a decision tree*. in *2015 IEEE International*

Symposium on Systems Engineering (ISSE). 2015. IEEE.

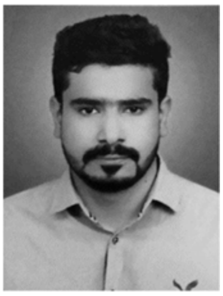
26. Yeo, M., T. Fletcher, and J. Shawe-Taylor, *Machine learning in fine wine price prediction.* Journal of Wine Economics, 2015. **10**(2): p. 151-172.
27. Noble, W.S., *What is a support vector machine?* Nature biotechnology, 2006. **24**(12): p. 1565-1567.
28. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* Journal of chemical information and computer sciences, 2003. **43**(6): p. 1947-1958.
29. Kanungo, T., et al., *An efficient k-means clustering algorithm: Analysis and implementation.* IEEE transactions on pattern analysis and machine intelligence, 2002. **24**(7): p. 881-892.
30. Van Laarhoven, P.J. and E.H. Aarts, *Simulated annealing*, in *Simulated annealing: Theory and applications*. 1987, Springer. p. 7-15.
31. Wang, S.-C., *Genetic algorithm*, in *Interdisciplinary computing in java programming*. 2003, Springer. p. 101-116.

Umer Zukaib is serving as a lecturer at Comsats University, Islamabad, Pakistan. He has done his master’s degree from Comsats University Islamabad, majoring in Computer Science. His research interest includes Machine Learning, Deep Learning, Data Mining, and Computer Vision.

Tariq Khan is a Ph.D. Researcher at the University of Politecnico Delle Marche, Ancona, Italy. He completed his master’s degree from Beijing University of Post and Telecommunication, Beijing, China. As well as he acquired a bachelor's from Mehran University of Engineering and Technology Jamshoro, Sindh, Pakistan.

His area of interest is Process Mining, Text Analytics, Semantic Analytics.

Mir Hassan is pursuing his Ph.D. degree in Computer Engineering at Vilnius University, Vilnius, Lithuania. He also served as a post-doctoral researcher at the University of Glasgow, Glasgow. The UK. He did his Master of Engineering in Software Engineering from Wuhan University, Wuhan, China. His research interest areas are Blockchain Technology, Machine Learning, and the Internet of Things.



Shoaib Ali is pursuing a master’s degree in computer science at the Virtual University of Pakistan, He has also acquired a master’s degree in Commerce from the University of Sindh and an Advanced Diploma in Software Engineering from Aptech Computer Education. His research interest includes Machine Learning, Data Science, and IoT.