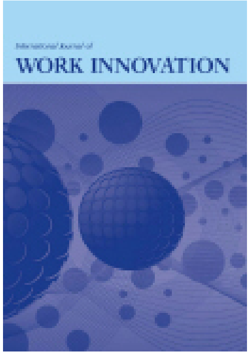


International Journal of Work Innovation

ISSN online: 2043-9040 - ISSN print: 2043-9032
https://www.inderscience.com/ijwi

Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques

Swarna Prabha Jena, Bijay Kumar Paikaray, Jitendra Pramanik, Rishi Thapa, Abhaya Kumar Samal

DOI: 10.1504/IJWI.2023.10055314

Article History:

Received:	04 January 2023
Last revised:	29 January 2023
Accepted:	30 January 2023
Published online:	20 April 2023

Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques

Swarna Prabha Jena

Department of ECE,
Centurion University of Technology and Management,
Odisha, India
Email: prabha.jena@gmail.com

Bijay Kumar Paikaray*

School of Information and Communication Technology,
Medhavi Skills University,
Sikkim, India
Email: bijaypaikaray87@gmail.com
*Corresponding author

Jitendra Pramanik

Department of Mining,
National Institute of Technology,
Rourkela Odisha, India
Email: bidun35@gmail.com

Rishi Thapa

School of Management,
Amity University,
Madhya Pradesh, India
Email: rishi.m.thapa@gmail.com

Abhaya Kumar Samal

Department of CSE,
Trident Academy of Technology,
Bhubaneswar, Odisha, India
Email: abhaya@tat.ac.in

Abstract: Proving the quality of a food product is challenging for any country. Every country recommends using products whose quality has been assured. A similar thing applies to the wine industry. To promote their products, wine industries acquire quality certifications through expert assessments. It is an expensive and time-consuming process. This paper explores the usage of machine algorithms like principle component analysis (PCA), linear discriminant analysis (LDA), random forest (RF), Gaussian naive Bayes (GNB), decision trees (DT), K-nearest neighbour (KNN), logistic regression

(LR), and gradient boost (GB) for classifying the wine data into three main categories. The experimental work provides a comparative study of the accuracy of all classifiers is discussed in detail.

Keywords: classification; machine learning; decision tree; principle component analysis; PCA; linear discriminant analysis; LDA; Gaussian naive Bayes; GNB; K-nearest neighbour; KNN.

Reference to this paper should be made as follows: Jena, S.P., Paikaray, B.K., Pramanik, J., Thapa, R. and Samal, A.K. (2023) 'Classifications on wine informatics using PCA, LDA, and supervised machine learning techniques', *Int. J. Work Innovation*, Vol. 4, No. 1, pp.58–73.

Biographical notes: Swarna Prabha Jena is currently working as an Assistant Professor in the Department of ECE, Centurion University of Technology and Management, Odisha. She has received BTech and MTech degree from Biju Patnaik University of Technology. Currently, she is pursuing her PhD in the Field of IoT in Precision Agriculture at Centurion University of Technology and Management, Odisha. Her area of interest is the embedded system, internet of things and deep learning algorithms.

Bijay Kumar Paikaray is currently working as an Assistant Professor in the Department of Information and Communication Technology, Medhavi Skills University, Sikkim, India. He has completed his MTech and PhD from Centurion University of Technology and Management, India. His areas of research include high-performance computing, information security, ML and IoT.

Jitendra Pramanik is currently pursuing his PhD in the Department of Mining Engineering, National Institute of Technology, Rourkela, Odisha, India. His areas of research include internet of things (IoT)/wireless sensor network (WSN), machine learning (ML), soft computing, image processing, and video processing.

Rishi Thapa is currently working as an Assistant Registrar in Medhavi Skills University, Sikkim, India. He has completed his MBA, MCom and pursuing his PhD from Amity University Gwalior, India. His areas of research include strategic management, digital marketing, marketing analytics, lead content and CTA.

Abhaya Kumar Samal is working as a Professor and Dean (Research and Consultancy) in the Department of Computer Science Engineering, Trident Academy of Technology (TAT), Bhubaneswar, Odisha India. Her areas of research interest include fault-tolerant scheduling, internet of things (IoT)/wireless sensor network (WSN), ML/deep learning, and soft computing.

1 Introduction

Around 7,000 BCE in China, the most primitive evidence of wine preparation was found, made up of vinified rice/fruit/honey. With the growth of society and humankind, there has been a rise in the variety and quality of wine. It is the most popularly consumed beverage globally, having both commercials and social importance. Hence quality checking of wine has become a significant aspect for both consumers and manufacturers.

Due to the rise in wine consumption, the wine industry is likewise seeking ways to produce high-quality wine at reasonable prices (Pérez-Álvarez et al., 2019). The chemicals used for the production of wine are almost the same for a variety of wines. However, the quantity/concentration of chemicals used for a particular wine varies. Since the quality of wine cannot be compromised, the wine industry's main focus is to identify or classify wines via different intelligent methods. The most conventional way of judging the quality of wine is through sensory evaluation using trained experts. However, the accuracy of this method cannot be trusted as the judgement can be easily affected by environment, personal or physical conditions.

Other methods of classifying wine include chemical analysis-based approaches like chromatography, mass spectrometry, and gas chromatography-mass spectrometry (Lona-Ramirez et al., 2016; Papageorgiou et al., 2018; Nicolli et al., 2018). Component by component approach is considered highly authentic but has disadvantages like high costs, online measurement, and low capacity for on-site (Jing et al., 2014). Innovative methods like multitask learning (MTL) have also been implemented in wine quality analysis. Numerous studies have shown that learning several related activities simultaneously is more effective than learning them separately (Caruana, 1997; Chang and Yang, 2014). In this paper, we provide a robust framework for classifying wine data using principle component analysis (PCA) and a variety of ML classifiers, including SVC, K-nearest neighbour (KNN), decision trees (DT), Gaussian naïve Bayes (GNB), random forest (RF), logistic regression (LR), and gradient boost (GB). The dataset used in all of the studies in this work was taken from the UCI library and used under identical experimental circumstances.

The work done is introduced in Section 2 presents the related work, Section 3 dataset, PCA, machine learning classifiers, and evaluation metrics. The various experimental results are provided along with an interpretation in Section 4. In Section 5, the study concludes with a discussion on future work.

2 Related works

In the past, efforts have been made to implement different machine-learning algorithms, and feature selection approaches on wine datasets. PCA is one of the oldest and most popular multivariate statistical techniques used to reduce the dimensionality of a dataset with many interrelated variables while retaining as much variation as possible (Wold et al., 1987).

Reddy and Govindarajulu (2017) implemented an approach to the red wine data set of centric clustering for the survey. By allocating relative voting to features and assigning weights to the features through the Based on the user's preferences, they forecasted the wine quality using the Gaussian distribution technique. Beltrán et al. (2008) reported using aroma chromatograms for wine classification. PCA was implemented to reduce the dimension of the dataset along wavelet transform for feature extraction. The support vector classifier with wavelet transform was proposed to perform better than other classifiers.

Appalasamy et al. (2012) reported the use of the physiochemical test to enhance wine quality during production. Panahi et al. (2011) reported a hybrid method involving different classifiers like Bayesian, KNN, and Parzen window along with PCA and linear discriminant analysis (LDA). It claimed to have a fast and efficient classification with low computational. Aich et al (2019) proposed using different ML techniques to anticipate the quality of the wine. By using feature selection methods like genetic algorithm (GA) and simulated annealing (SA) based feature selection, the accuracy was examined. To forecast the wine quality, Shruthi (2019) used a variety of data mining classification techniques, including naive Bayes, simple logistic, KStar, JRip, and J48. The classifiers divided the wine into three groups, and the accuracy of the classifiers was compared. Patra et al. (2022) reported the application of ensemble learning to prognosticate sugar levels in the human body. The applied model was assessed through different classification metrics. Dogra et al. (2021) studied the potency of machine learning algorithms in predicting mine fire scenarios.

3 Methodology

With Subsections 3.1, 3.2, and 3.3 explaining the dataset, ML algorithms, and hardware and metrics used to evaluate the framework, respectively, this section concentrates on the materials and methods utilised for this study in the literature.

3.1 Dataset

The ML techniques are implemented on the publicly available wine dataset present in the UCI library. The 178 rows and 13 dependent variables that make up the wine dataset provide information on the various chemical components of a single wine. Three distinct clusters (1, 2, and 3) of consumers who liked https://archive.ics.uci.edu/ml/datasets/Wine) particular wines make up the dependent variable.

3.2 Machine learning algorithms

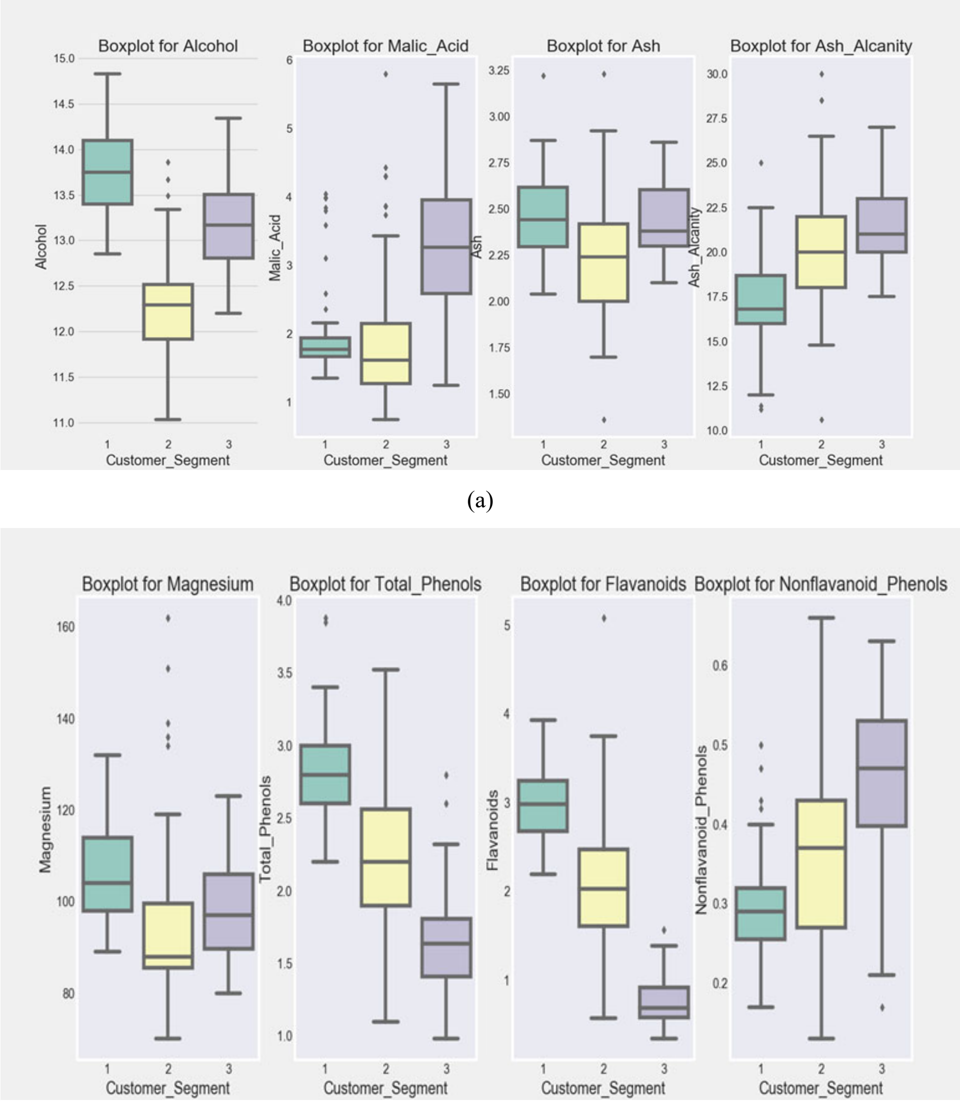
The ML algorithm's performance is enhanced by increasing the attribute's dimension. Unfortunately, sometimes the classifier accuracy tends to reduce when the sample size stays the same while the attribute's dimension grows. This is referred to as the curse of dimensionality.

In this section, we have discussed the most popular methods, like PCA, LDA, and RFE, used for selecting the dominant attributes from a dataset along with different ML algorithms such as KNN, LR, GNB, RF, SVC, DT, and GB are also implemented.

3.2.1 Principle component analysis

One of the most widely used techniques for multivariate analysis involves the principal-component analysis of the spectral decomposition of a correlation coefficient or covariance matrix. It is the most preferred technique for dimensionality reduction and feature extraction. The direction of principle components of data is taken eigenvector, and their corresponding eigenvalues give their statistical significance. When PCA is implemented for dimension reduction a fraction of the eigenvector is kelp (Panahi et al., 2011).

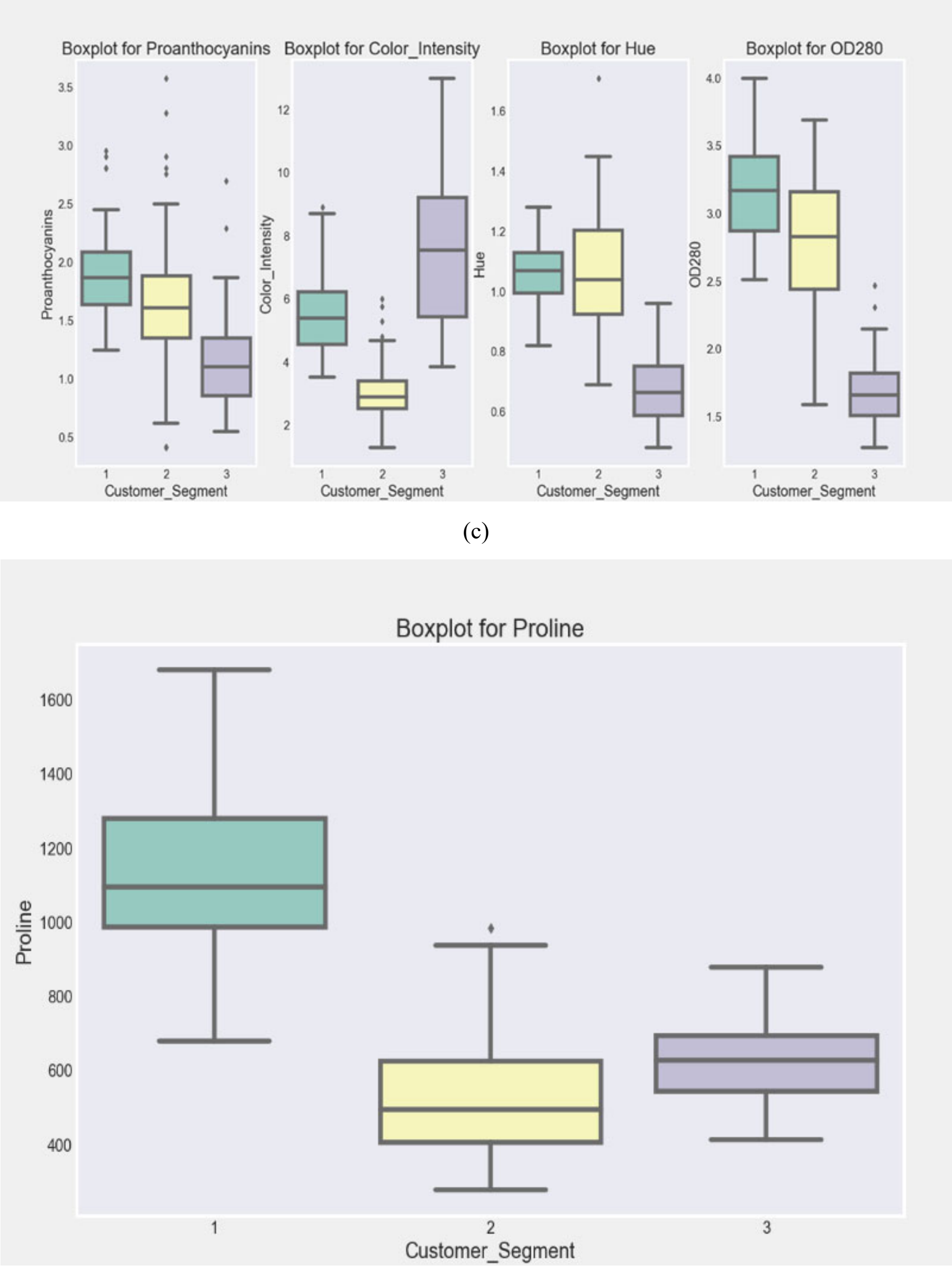
Figure 1 The wine dataset's population distribution of each attribute, (a) population distribution of alcohol, malic acid, ash, ash alcanity, (b) population distribution of magnesium, phenols, flavonoids, and non-flavonoids, (c) population distribution of proanthocyanins, colour intensity, hue, OD280, (d) population distribution of praline (see online version for colours)



(a)

(b)

Figure 1 The wine dataset's population distribution of each attribute, (a) population distribution of alcohol, malic acid, ash, ash alcanity, (b) population distribution of magnesium, phenols, flavonoids, and non-flavonoids, (c) population distribution of proanthocyanins, colour intensity, hue, OD280, (d) population distribution of praline (continued) (see online version for colours)



(c)

(d)

$X \in Rn$ with N number of samples the mean is

$$\underline{X} = \frac{1}{N}\sum\nolimits_{i=1}^{N} Xi \cdot \text{ where } \underline{X} \in Rn$$

And the covariance matrix is

$$C_{n\times n} = \sum_{i=1}^{N}(Xi - \underline{X})(Xi - \underline{X})^{T}$$

The eigenvectors are computed $C_{n\times n}$ as PDP – 1, where $P \in Rn$ is the matrix of eigenvectors and Dn is the diagonal matrix with eigenvalues. By sorting the eigenvectors in descending order, we can choose eigenvectors having variance greater than the threshold.

3.2.2 K-nearest neighbours

It is another supervised ML algorithm used for classifications and regression problems. However, it is mostly employed in the sector to categorise predictive issues. KNN is a non-parametric and lazy learning algorithm since it makes no assumptions about the underlying data point and stores the training set at the time of classification rather than learning from it immediately. It maintains all accessible data samples and classifies with the new data. Then new data appears, by using KNN algorithm, the data points can be easily classified into a good suite category (Singh, 2019; Sanal Kumar and Bhavani, 2019).

3.2.2.1 KNN algorithm

In the case of the KNN algorithm, on basis of similarity in features new data points are being predicted. This implies that the degree to which the new data point resembles the points in the training set will determine the value that will be assigned to it.

The workflow of the algorithm will be as follows:

Step 1 Dataset loading.

Step 2 Initialising the K value.

Step 3 Data follow for each point in given below steps:

- Step 3.1: measure the separation between each row of training data and the query data.
- Step 3.2: add the distance and index of the example to an ordered collection.
- Step 3.3: based on the distance and indices value, now sort them in ascending order.
- Step 3.4: select the top rows K from the array of the sorted collection.
- Step 3.5: give the data point class based on the most prevalent class of these rows.

Step 4 End.

3.2.3 Gaussian naïve Bayes

With the 'naive' assumption that each featured pair is independent given the value of the class variable, this classification model is based on Bayes' theorem.

For example, consider the value of a class variable 'y' and a dependent set of features 'xi', then Bayes' theorem represented relationship among the following variables.

$$P\left(x_1, ..., x_n\right) = \frac{P(y)P\left(x_1, ..., x_n \mid y\right)}{P\left(x_1, ..., x_n\right)}$$

Using the naïve independence assumption, the above equation is represented by:

$$P\left(x_1, ..., x_n\right) = \frac{P(y)\prod_{i=1}^{n} P\left(x_i \middle| y\right)}{P\left(x_1, ..., x_n\right)}$$

Since $P(x_1, \ldots, x_n)$ is constant for given the input data, the above equation can be represented as:

$$P\left(x_1, ..., x_n\right) \propto P(y)\prod_{i=1}^{n} P\left(x_i \middle| y\right)$$

Now to implement naïve Bayes classification model for the above equation, With respect to each potential value of the target set (y), the probability of the feature input datasets '$x_1, \ldots, x_n$' is computed, and the probability with the highest value is selected as the output, which is represented as follows:

$$y = \arg\max x_y P(y)P(y)\prod_{i=1}^{n} P\left(x_i \middle| y\right)$$

The class and conditional probability are $P(y)$ and $P(x_i| y)$, respectively.

The naive Bayes classification model can be used for multi-class prediction, spam filtering, text classification, real-time prediction, sentiment analysis, and recommendation system (Sanal Kumar and Bhavani, 2019).

3.2.4 Random forest

In ML, both classification and regression problems can be solved using RF. It is a supervised learning technique, it is created on the ensemble learning concept, in which multiple classifiers can be combined to solve the complex problem as well as for the best model of the performance (Sanal Kumar and Bhavani, 2019).

It is a classifier model which has a number of DT that are created based on various subsets of given data samples and to improve the prognostic accuracy, the prediction for each tree is used to select the best possible solution by means of voting.

The working process of the RF algorithm can be explained with the help of the below steps.

Step-1 Randomly select K data points from the practise set.

Step-2 Build the DT using the subsets of the selected data points.

Step-3 Specifying N, the number of DT to be constructed.

Step-4 Repeat steps 1 through 2.

Step-5 For the new data sample find the forecasts of each decision tree, then assign that new data samples to the category with the highest number of votes.

The RF model is mostly used in four sectors i.e., banking, medicine, land use, and marketing.

3.2.5 Gradient boost

A simple parameterised function (base learner) is incrementally fitted using least squares to the current 'pseudo'-residuals at each iteration to create additive regression models (Hao, 2021).

The gradient of the loss function, which is being minimised with respect to the values of the model at each training data point, is what makes up the pseudo-residuals. It is shown that the approximation accuracy and execution speed of gradient boosting can be considerably improved by including randomisation in the process. More specifically, from the complete training data set, a subsample of the training data is randomly chosen (without replacement) at the beginning of each cycle. Then, instead of computing the model update for the current iteration using the complete sample, the base learner is fitted to this randomly selected subsample. This randomisation method also strengthens resistance to overcapacity of the base learner.

As an illustration, take into account a system with a random 'output' or 'response' variable y and a collection of random 'input' or 'explanatory' variables $x = (x_1, \ldots, x_n)$.

Given a 'training' sample $\{y_i, x_1\}_1^N$ of known $(y; x)$ values, the goal is to *nd a function.

x to y is the maps of the $F^*(x)$, such that all (y, x) values over the joint distribution of the expected value of some specified loss function $\Psi(y, F(x))$ is minimised.

$$F^*(x) = \arg\min_{F(x)} E_{y,x} \Psi(y, F(x)):$$

$F^*(x)$ boosting approximates by an 'additive' extension of the form

$$F(x) = \sum_{m=0}^{M} \beta_m h(x; a_m),$$

3.2.6 Logistic regression

It is a method of categorisation that utilises the idea of probability. Using the logistic sigmoid function, it transforms its output into a probability value that may be divided into two or more discrete categories. Contrary to what its name might imply, it is not an algorithm for problems involving regression in which a continuous result needs to be predicted. Instead, the method of binary categorisation is chosen (problems with two class values). It will result in a discrete binary value, or, to put it another way, either 0 or something else (Sanal Kumar and Bhavani, 2019).

The underlying logistic function, commonly known as the sigmoid function, is used to estimate probabilities when examining the interaction between the independent variables and the dependent variable (our label, the thing we want to forecast) (our features). The sigmoid function, an S-shaped curve, may change any real-valued number into a value between 0 and 1, but never precisely at those values. The threshold classifier will then change these values between 0 and 1 to either 0 or 1.

The sigmoid function's job is to convert these probabilities into binary values so that a prediction can be made.

$$S(z) = \frac{1}{1+e^{-z}}$$

where $S(z)$ is the output between 0 and 1, z = input to the function, e = base of natural log.

Because the procedure divides the input data into two 'regions' by a linear boundary, one for each class, one of the disadvantages of the algorithm is that we cannot tackle nonlinear problems with LR. Your data must therefore be able to be separated linearly.

3.2.7 Decision tree

The tree algorithm is a different supervised learning method that can be used for classification and regression problems. First, it categorises data objects into a limited number of pre-established classifications. Then, starting from the tree's root node, which branches into other nodes of many potential outcomes, we attempt to predict the class label. Each node in the tree represents a test case for a different property, and the edges descending from each node reflect a variety of potential answers to the test case. This process is repeated for each subtree that has its root at the new node (Sanal Kumar and Bhavani, 2019; Awasthi and Kumar, 2022).

A decision tree is made up of two primary methods: one is building the tree, and the other one is classification.

1. The edges descending from each node represent potential answers to the test case, and as the tree is built, each node serves as a test case for a distinct attribute. This iterative process will be repeated for each subtree that is anchored at the new node.
2. Classification. The induced tree is used as the basis for classifying new items. Therefore, to categorise something, we begin at the root, assess the relevant test attribute, and select the branch corresponding to the test result. Up until a leaf is found, this process is repeated. The newly discovered item is subsequently assigned to the leaf's class.

3.3 Hardware and evaluation metrics

On a PC running the Windows 10 operating system, the trials were carried out with the following hardware configuration: which was used to implement the machine learning algorithms: processor: Intel R CoreTM i5 CPU@2:20 GHz, installed memory (RAM): 12 GB.

Figure 2 Performance plot of all classifiers without PCA (see online version for colours)

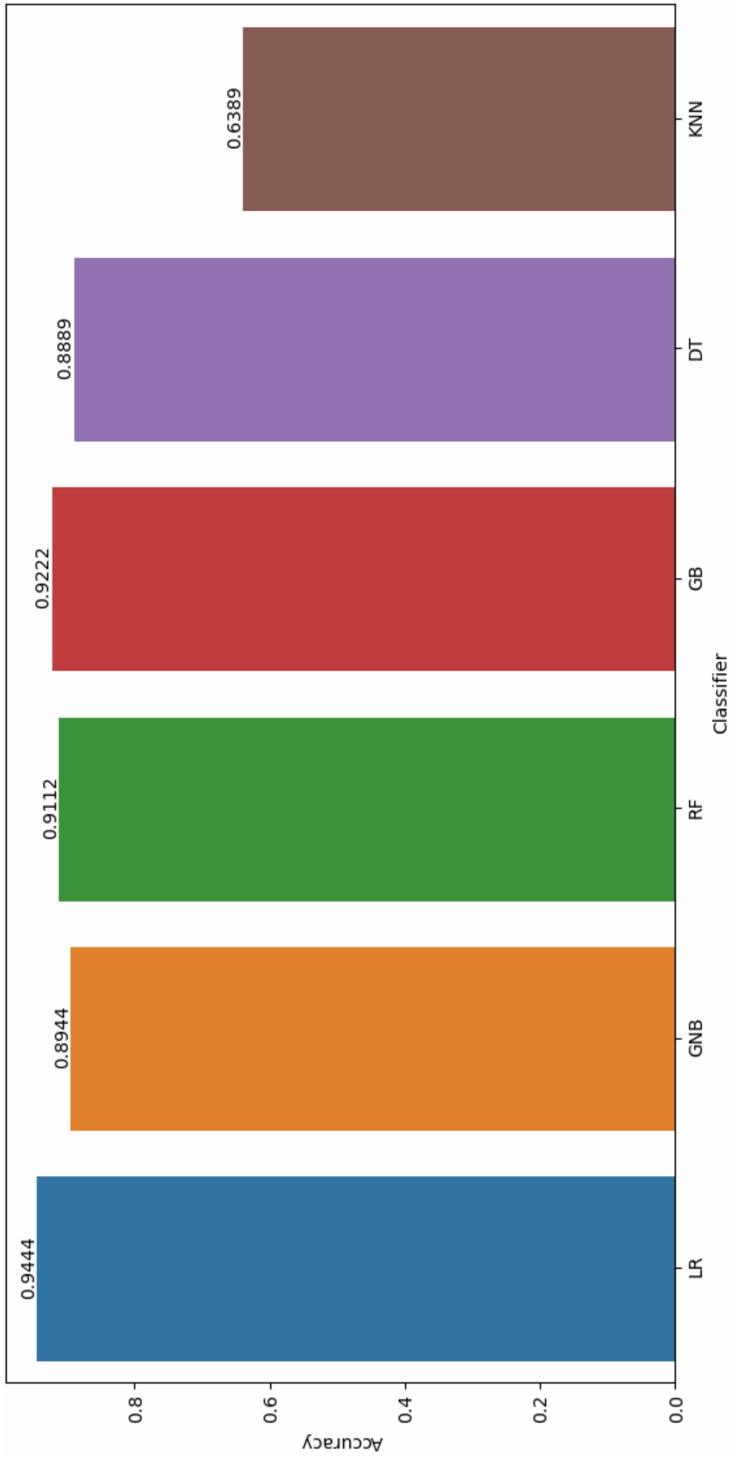


Figure 3 Performance plot of all classifiers with PCA (see online version for colours)

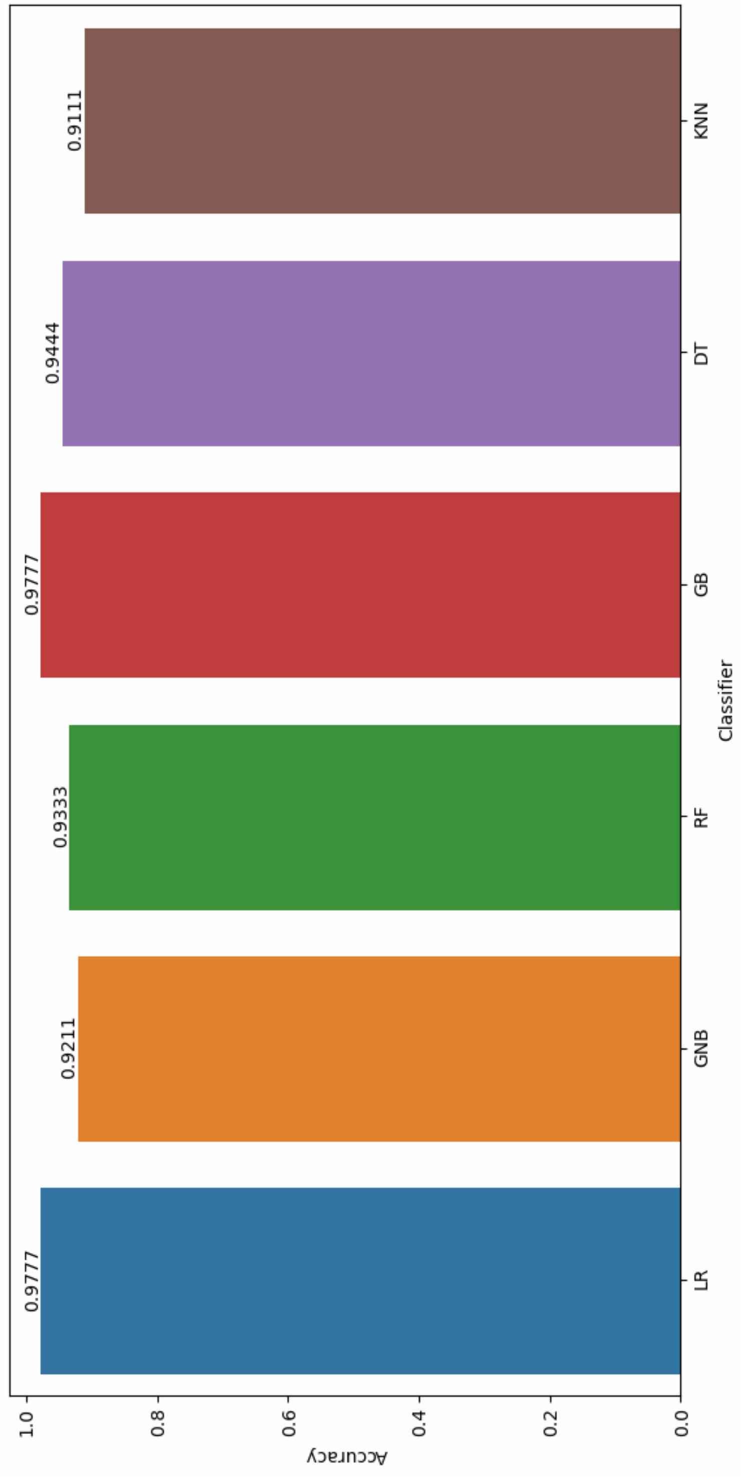
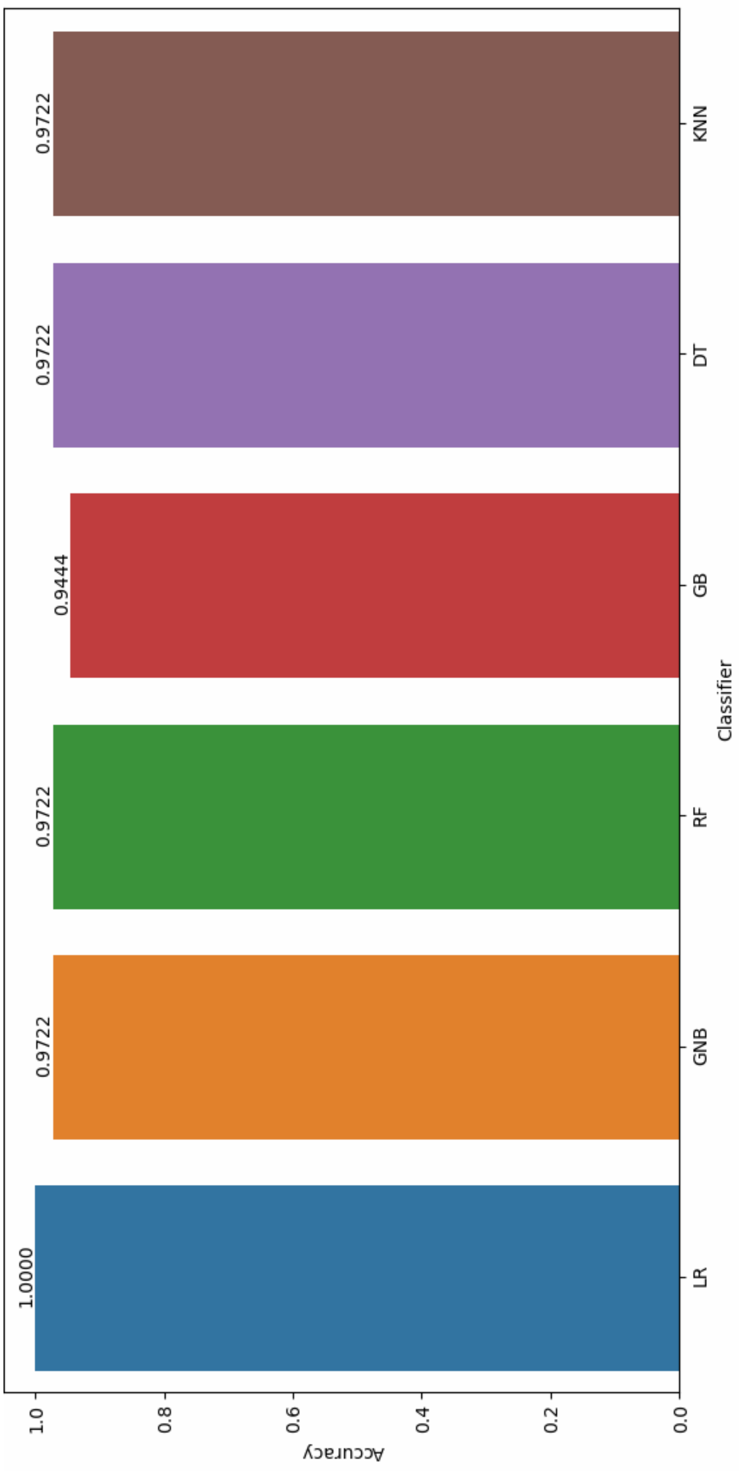


Figure 4 Performance plot of all classifiers with LDA (see online version for colours)



4 Experimental results and comparison

Out of 13 variables in the dataset, 12 are considered independent variables, and the customer segment is regarded as a dependent variable. At first, the accuracy of ML classifiers is tested on the raw dataset without implementing any feature extraction or elimination technique like PCA and a comparison has been drawn based on it. The comparative performance plot is shown in Figure 2. The highest accuracy is shown by GB and LR, whereas KNN and SVC show the lowest performance.

To improve the classifier performance, other algorithms can be implemented to figure out the dominant features present in the dataset. Firstly, PCA is implemented to find out the dominant features present in the dataset. Then, it optimises the dataset whenever it has more than one of the attributes. In our procedure, we have selected two no-principle components that have been taken for further process. Finally, the output of PCA is fed to the group classifier, and performance is shown in Figure 3; as we can see, the accuracy of all the classifiers has attended a consistent level.

Table 1 Assessment of performance of classifiers with PCA

Performance score of different classifiers with PCA				
Classification algorithms	*Quality metric parameters*			
	Score	*Precision*	*Recall*	*F1-score*
Logistic regression	0.977	0.970	0.977	0.978
Gaussian naïve Bayes	0.921	0.918	0.918	0.917
Support vector classifier	0.901	0.901	0.901	0.901
Random forest	0.933	0.932	0.933	0.933
Gradient boost	0.977	0.970	0.977	0.978
Decision tree	0.944	0.932	0.932	0.944
K-nearest neighbour	0.911	0.921	0.921	0.921

Table 2 Assessment of performance of classifiers with LDA

Performance score of different classifiers with LDA				
Classification algorithms	*Quality metric parameters*			
	Score	*Precision*	*Recall*	*F1-score*
Logistic regression	1	1	1	1
Gaussian naïve Bayes	0.972	0.972	0.972	0.972
Support vector classifier	1	1	1	1
Random forest	1	1	1	1
Gradient boost	0.944	0.944	0.944	0.944
Decision tree	0.972	0.972	0.972	0.972
K-nearest neighbour	0.972	0.972	0.972	0.972

We used LDA on the input dataset in the second and third processes before feeding it to the classifier models. LDA was implemented it eliminate the redundant information present in the dataset. The output of LDA was given as input to the classifier model to predict the target value.

5 Conclusions

The use of machine learning techniques to forecast wine quality is explored in this research. The old method was time-and money-consuming to implement, but the feature selection strategy provided an explicit knowledge of the value of the features for quality prediction. Furthermore, the experiment demonstrates that, as opposed to considering all features, only essential aspects can be considered when predicting the value of the dependent variable. It should also be noted that the methodology created in this work is broad and might be applied to various applications, such as the classification of coffee, olive oil, explosives, etc.

References

Aich, S., Al-Absi, A.A., Lee Hui, K. and Sain, M. (2019) '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)*, Pyeong Chang Kwangwoon_Do, Korea (South), pp.1122–1127, DOI: 10.23919/ICACT.2019.8702017.

Appalasamy, P., Mustapha, A., Rizal, N.D., Johari, F. and Mansor, A.F. (2012) 'Classification-based data mining approach for quality control in wine production', *Journal of Applied Sciences*, Vol. 12, No. 6, pp.598–601.

Awasthi, K.K. and Kumar, M. (2022) 'Survey of supervised machine learning techniques in wireless sensor network', in: Dhawan, A., Mishra, R.A., Arya, K.V. and Zamarreño, C.R. (Eds.): *Advances in VLSI, Communication, and Signal Processing. Lecture Notes in Electrical Engineering*, Vol. 911, Springer, Singapore, https://doi.org/10.1007/978-981-19-2631-0_18.

Beltran, N.H., Duarte-Mermound, M.A., Vicencio, V.A.S., Salah, S.A. and Bustos, M.A. (2008) 'Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer', *IEEE Trans. Instrum. Measurement*, Vol. 57, No. 11, pp.2421–2436.

Caruana, R. (1997) 'Multitask learning', *Machine Learn.*, Vol. 28, No. 1, pp.41–75.

Chang, X. and Yang, Y. (2014) 'Semisupervised feature analysis by mining correlations among multiple tasks', *IEEE Trans. Neural Netw. Learn. Syst.*, July, Vol. 28, No. 10, pp.2294–2305.

Dogra, S.K., Jayanthu, S., Samal, A.K., Pramanik, J. and Pani, S.K. (2021) 'Machine learning approach to implement mine fire predicting for underground coal mines', *2021 2nd Global Conference for Advancement in Technology (GCAT)*, pp.1–4, DOI: 10.1109/GCAT52182.2021.9587499.

Hao, J. (2021) 'Supervised machine learning', in: von Davier, A.A., Mislevy, R.J. and Hao, J. (Eds.): *Computational Psychometrics: New Methodologies for a New Generation of Digital Learning and Assessment. Methodology of Educational Measurement and Assessment*, Springer, Cham, https://doi.org/10.1007/978-3-030-74394-9_9.

Jing, Y., Meng, Q., Qi, P., Zeng, M., Li, W. and Ma, S. (2014) 'Electronic nose with a new feature reduction method and a multi-linear classifier for Chinese liquor classification', *Rev. Sci. Instrum.*, Art. No. 055004, Vol. 85, No. 5, p.055004.

Lona-Ramirez, J., Gonzalez-Alatorre, G., Rico-Ramírez, V., Perez-Perez, M.C.I. and Castrejón-González, E.O. (2016) 'Gas chromatography/mass spec-trometry for the determination of nitrosamines in red wine', *Food Chem.*, April, Vol. 196, No. 1, pp.1131–1136.

Nicolli, K.P., Biasoto, A.C., Souza-Silva, É.A., Guerra, C.C., dos Santos, H.P., Welke, J.E. and Zini, C.A. (2018) 'Sensory, olfactometry and comprehensive two-dimensional gas chromatography analyses as appropriate tools to characterize the effects of vine management on wine aroma', *Food Chem.*, March, Vol. 243, No. 1, pp.103–117.

Panahi, N., Shayesteh, M.G., Mihandoost, S. and Zali Varghahan, B. (2011) 'Recognition of different datasets using PCA, LDA, and various classifiers', *2011 5th International Conference on Application of Information and Communication Technologies (AICT)*, Baku, pp.1–5, DOI: 10.1109/ICAICT.2011.6110912.

Papageorgiou, M., Lambropoulou, D., Morrison, C., Namieśnik, J. and Plotka-Wasylka, J. (2018) 'Direct solid phase microextraction combined with gaschromatography – mass spectrometry for the determination of biogenicamines in wine', *Talanta*, June, Vol. 183, No. 1, pp.276–282.

Patra, N., Pramanik, J., Samal, A.K. and Pani, S.K. (2022) 'Machine learning application in primitive diabetes prediction – a case of ensemble learning', in Mallick, P.K., Bhoi, A.K., Barsocchi, P. and de Albuquerque, V.H.C. (Eds.): *Cognitive Informatics and Soft Computing. Lecture Notes in Networks and Systems*, Vol. 375, Springer, Singapore, https://doi.org/10.1007/978-981-16-8763-1_64.

Pérez-Álvarez, E.P., Garcia, R., Barrulas, P., Dias, C., Cabrita, M.J., Garde-Cerdán, T. (2019) 'Classification of wines according to several factors by ICP-MS multi-element analysis', *Food Chem.* January, Vol. 1, pp.270–273, 280, DOI: 10.1016/j.foodchem.2018.07.087, Epub: 2018 Jul 20. PMID: 30174046.

Reddy, Y.S. and Govindarajulu, P. (2017) 'An efficient user centric clustering approach for product recommendation based on majority voting: a case study on wine data set', *IJCSNS*, Vol. 17, No. 10, p.103.

Sanal Kumar, K.P. and Bhavani, R. (2019) 'Human activity recognition in egocentric video using PNN, SVM, kNN and SVM+kNN classifiers', *Cluster Comput.*, Vol. 22, No. Suppl 5, pp.10577–10586, https://doi.org/10.1007/s10586-017-1131.

Shruthi, P. (2019) 'Wine quality prediction using data mining', *2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE)*, Bangalore, India, pp.23–26, DOI: 10.1109/ICATIECE45860.2019.9063846.

Singh, P. (2019) 'Supervised machine learning', in *Learn PySpark*, Apress, Berkeley, CA, https://doi.org/10.1007/978-1-4842-4961-1_6.

Wold, S., Esbensen, K. and Geladi, P. (1987) 'Principal component analysis', *Chemometrics Intell. Lab. Syst.*, Vol. 2, Nos. 1–3, pp.37–52.