An investigation of wine quality testing using machine learning techniques

Sathishkumar Mani[1], Reshmy Avanavalappil Krishnankutty[2], Sabaria Swaminathan[3], Prasannavenkatesan Theerthagiri[1]
[1]Department of Computer Science and Engineering, GITAM School of Technology, GITAM University, Bengaluru, India
[2]Department of Computational Intelligence, SRM Institute of Science and Technology, Kattankulathur, India
[3]Department of Computer Applications, B.S. Abdur Rahman Crescent Institute of science and technology, Vandalur, India

Article Info

Article history:

Received Jul 17, 2020
Revised Oct 31, 2022
Accepted Nov 30, 2022

Keywords:

Linear regression
Machine learning
Quality
Random forest
Wine testing

ABSTRACT

Quality is the most determining factor for any product. Optimal care and best measures are to be taken in assessing the quality of any product. This work deals with determining the quality of wine using intelligence-based learning techniques. In order to estimate the quality of wine, several experiments are performed on wine datasets. The main purpose of our work is to study and discover an efficient machine learning (ML) model that could determine the quality of wine given some Physico-chemical features. This study establishes that selecting important features to evaluate rather than all of them can lead to improved forecasts. According to the results, this approach may provide people who are not wine experts a greater opportunity to choose a fine wine.

Corresponding Author:

Prasannavenkatesan Theerthagiri
Department of Computer Science and Engineering, GITAM School of Technology, GITAM University
Bengaluru, India
Email: prasannait91@gmail.com

1. INTRODUCTION

An alcoholic beverage derived from fermented grapes is called wine. By consuming the sugar in the grapes and converting it to ethanol, carbon dioxide, and heat, the yeast is utilised to prepare wine. Red wine, rose wine, white wine, and other wine types are produced using various grape varietals and yeast strains. White wine is a type of wine that undergoes skin-free fermentation [1]. While preparing the wine, its quality must be tested to verify the acidity and pH level. Currently, the food industry incorporates advanced technology for their developments.

In such a way, machine learning (ML) models was applied for the classification of wine quality by the food industry for their research quality management. As the demand for wine is drastically increasing these days, the urge to better estimate its quality has become the need of the hour. Wine quality prediction is a part of wine informatics where we predict the quality of the wine. Wine quality and its certification can be done based on its Physico-chemical properties or by sensory tests. Physico-chemical properties include pH, dissolved salts, sodium levels, the acidity of liquid and conductivity. Since human specialists are mostly used in sensory testing and taste is the sense that humans understand the least, this is a challenging task.

This work involves determining the quality of the wine using artificial intelligence-based learning methods. The process of wine quality prediction using ML involves constructing a trained model based on the available input data. The model should predict the value accurately, thus minimizing the error and increasing the overall performance. Many of the existing algorithms like logistic regression [2], support vector machine (SVM) [3], decision tree [4], Adaboost regression [5], [6] are suffering from performance problems. To ensure better performance, the obtained results are evaluated using evaluation metrics, namely accuracy, recall, precession, F1 score [7], [8]. A good evaluation metric helps in better estimating the performance of the model. The evaluation metrics should be chosen such that they suit the current problem.

The task of any quality prediction comes under supervised learning as the target variable is already known. Generally, for the continuous data, regression techniques are used. The regression models intended to find a function F: X->Y, given X and Y. Independent input variables (X) are mapped to dependent output variables using this approach (Y). From labelled training data made up of a collection of training instances, the regressor function is derived. Finally, the regressor function F should correctly predict the output value for unseen input tuples [9], [10].

This work estimates the quality of the wine as ranging from 0 to 10, based on different Physico-chemical features. The quality estimation of wine would help both the sellers and the buyers of the wine understand the quality of wine and the levels of different ingredients contained in the wine. Here we use different types of regression algorithms and identify the one which performs the best by using evaluation metrics. The learning models are optimized to choose the best-performed model for obtaining better results.

The remainder of the article is structured. The essential research studies on the practises of ML are detailed in this section. The suggested approach, tools, and algorithms of this study are covered in section 2. The performance outcome of the prediction is compared and analysed with various metric performances in section 3. Enhancements to the features round up section 4.

This section summarizes the usage of ML techniques for wine quality estimation. The investigation of the relationship between wine qualities and quality was done by Chen *et al.* [1]. To perform this study, they took region-specific wine samples consisting of 1200 different wine reviews and constructed a dataset. The dataset contains the attributes regarding the flavors used to prepare the wine. The authors used variables taken from wine reviews to estimate the quality of the wines using an association rule-based classification system. Support and confidence metrics were used for each association rule generated from the dataset. However, this paper only concerns wines belonging to a particular region. Authors had concluded wine reviews might help as a basis for Prediction of quality by drawing association rules from the reviews.

Fan *et al.* [9] has performed the statistical analysis and developed a mathematical model application that describes whether the physical-chemical indicators can alone predict the quality correctly or not. To conduct this study, firstly, they took evaluation results of 2 sommeliers. They conducted three different statistical tests to conclude the result. The first test is fitting analysis, and it deals with calculating significant differences between the evaluation results of both sommeliers. The second test is the variance analysis; here, the results from this are more dependable. This test uses different statistical measures like total difference and group difference. The final test is the Q cluster analysis. This test finds the effect of physical-chemical components on the quality of the wine. Further, they analyzed the dataset to find whether physical-chemical components can be used to predict wine quality. The model concludes that when physical-chemical components are alone used to predict the quality of wine, they may not give accurate results.

Hu *et al.* [11] has constructed a model to predict the quality of wine by balancing the imbalanced data. The authors have developed the model using the balanced percentage of data. The remaining percentage of imbalanced data was applied with the synthetic minority over-sampling technique (SMOTE), which increases the number of cases to make the data balanced. They have used different classifiers, namely, adaptive boosting, decision tree, random forest, and concluded that the random forest techniques produce desired results when applied to imbalanced data. For concluding the best classifier, they have calculated different performance metrics, namely sensitivity, specificity, accuracy, and error rates before and after applying SMOTE. This technique had reduced the over-fitting problem, and SMOTE along with random forest gave the best results. Aich *et al.* [12] also used the classification, linear, non - linear classifiers, and probabilistic algorithms on a balanced dataset. Some feature selection techniques such as genetic algorithms based and simulated annealing-based feature selection to evaluate the performance of the prediction. The authors had concluded that the SVM classifier gives the best results for red and white wine datasets.

Andonie *et al.* [13] had focused on cost minimization by finding the best features that correctly predict the quality of the wine. So that it can collect the required best features to predict the quality, thus reducing the cost. In doing this, they have used a separate wine dataset of their own and data quality predict by some wine experts. In their study, they have ranked the features that best predict the quality. The ranking is age, co-pigmentation, region, total SO2, isobutanol, color intensity, volatile acidity, polymeric anthocyanins, color anthocyanins, total phenols, Brettanomyces, free SO2, pH, acetic acid bacteria, lactic acid bacteria, ethanol, titratable acidity, active amyl alcohol. Also, the author has analyzed the cost estimation and accuracy for different sets of features.

Nebotl *et al.* [14] had used fuzzy techniques, namely fuzzy inductive reasoning (FIR) and other genetic fuzzy systems, to predict wine quality. FIR is based on fuzzy logic and ML. Among all the techniques, they found that FIR gives the best results as it takes less computational time for obtaining the system model and performing Prediction. They concluded that the results obtained by fuzzy techniques are really appropriate for different aspects of the wine industry. Also, based on their observations, they suggested that fuzzy logic is not always accurate.

Cortez *et al.* [4] had developed prediction models using neural networks, linear regression, and support vector machines (SVMs). Each one of them used large data sets for the predictions. The authors have concluded that SVM gives better results when compared to regressors and neural networks. Also, for better accurate results, rather than all the variables or features, they selected few variables to predict the target variable. The best variables are selected based on the impact of the target variable [3].

Yesim [5] had presented a method to predict wine quality based on physicochemical properties using classification techniques. For the experiments, the authors took two large data sets containing white wine and red wine; and classified them using the random forest algorithm. They used the K nearest neighbor and support vector machines (SVMs) for classifying the data. Based on the observations authors suggested that the random forests give the best classification. The outcomes are categorized in percentage by applying cross-validation mode or split percentage mode. The authors infer from the results of the main component analysis that the quality classification decreases in white wine and increases in red wine in both cross-validation and split-rate mode.

2. METHOD

Quality is the most determining factor for any product. Hence, optimal care and best measures are to be taken in assessing the quality of the product. In order to estimate the quality of wine, a system should be developed for predicting the quality of a wine based on the acidity levels and ingredients of the wine. This work aims to determine an efficient ML model that could estimate the quality of wine given some Physico-chemical features. The best quality of wine can be manufactured and made available for consumers.

In the present scenario, the existing system requires chemical titrations to be performed to determine the quality of the wine. One of the other ways of finding the quality of wine is by human expertise. The major drawback of this method is the taste of the wine will change significantly by the time it will be consumed [14]. ML techniques are used in this work for solving the problem of wine quality testing. Since features like acidity and volatility determine the wine's quality, a regression model can be developed from these features. In this way, the quality of the wine could be predicted, aiming at producing an efficient model.

The main functionality of the work is to return the quality value of wine when a set of input physio-chemical properties are provided. To attain this, a dataset containing a large number of wine samples is considered. In order to ensure that the data is clean and ready for use, data exploration and data preprocessing are performed. Through data exploration, we can get a clear insight into data. Data exploration can be better understood by using visualization techniques. Data preprocessing helps find and replace the missing values, detect outliers and duplicate values. If they are of low significance or misleading, the values can be removed from the dataset, increasing the prediction standard.

The model construction requires algorithms that should be trained on our dataset and evaluate. This is done by using evaluation metrics called accuracy, precisions, recall, and F1 score. Once the model construction is completed, it finalizes a well-performed model based on the evaluation metrics for further optimization. Optimization helps increase the model's performance by finding better hyper-parameters like n-estimators, max-features, max-depths, and criteria to be provided for the model. The obtained hyper-parameters are passed to the finalized model, and the model is trained using them. The final trained model is used for estimating the quality of the unseen tuples in the future.

2.1. Experimental setup

The tools such as Anaconda, Jupyter Notebook, and python are adopted for the proposed work on wine quality estimation. The ML libraries required for this work are sklearn which contains various regression algorithms, including random forest model, SVM, Adaboost regressor, and decision tree. The tool pandas are used for loading the dataset and data manipulation. The numpy provides support for mathematical and scientific operations on data. Finally, matplotlib which are used for visualizing the data.

Anaconda is an open-source distribution for performing data science and ML on any platform. Anaconda helps quickly download and manage libraries required for all the ML tasks for the wine quality estimation [5]. Python code may be written and executed using the free and open-source online application Jupyter Notebook. It also includes data visualization, data cleaning, and other ML related to the wine quality estimation tests [15]. The current work was simulated using the 8 GB RAM, Intel core i5 processor, 120 GB of Hard disk space, and Windows 10 operating system.

2.2. Data source

The dataset for the wine quality estimation testing is taken from University of California, Irvine (UCI) Machine Learning Repository. The dataset consists of some physicochemical characteristics of red wine from the Vinho Verde region of Portugal. Only the physicochemical variables and the output variable are supplied due to logistical and privacy concerns and do not contain any information about wine brand and selling price. The dataset has 1359 rows and 12 columns [3]. The 12 variables and their description are,

- Fixed acidity: acids associated with wine are fixed or non-volatile.
- Volatile acidity: the volume of volatile acids.
- Citric acid: It is found in minute amounts and gives wines freshness and taste.
- Residual sugar: the remaining sugar after fermentation has stopped.
- Chlorides: the number of salts in the wine.
- Free Sulfur dioxide: the wine's oxidation and microbiological development are stopped by the free form of SO2.
- Total sulfur dioxide: the amount of free and combined forms of SO2.
- Density: depends on the percentage of sugar content and alcohol.
- pH: describes how acidic or basic the wine is on a scale of 0-14. Acids fall under the pH range of 0-7. Wines have a pH between 3 to 4.
- Sulphates: an additive to wine that contributes to SO2 levels.
- Alcohol: the alcohol percentage in the wine.
- Quality: It is the output variable ranging from 0-10.

2.3. Algorithms and techniques for quality estimation

The algorithms used in this work are logistic regressor, decision tree regressor, SVM, Adaboost regressor, and random forest regressor. A statistical technique for analysing a dataset in which one or more independent variables produce a result is called logistic regression. It is used for binary scoring. By fitting data to a logistic or sigmoid function, it predicts the likelihood of an event occurring [15].

For a more precise and reliable forecast, random forest constructs many decision trees and combines them. In a random forest, the process for splitting a node only takes into account a random subset of the characteristics. Using random criteria for each feature rather than looking for the best feasible thresholds might even increase the randomness of trees (like a normal decision tree does). In general, a forest appears more robust the more trees there are in it. Similar to this, the random forest classifier produces results with high accuracy the more trees there are in the forest [16].

A supervised ML approach called the SVM may be applied to classification and regression problems. However, categorization issues are where it's most frequently employed. In this approach, each data point is represented as a point in an n-dimensional space (where n is the number of features), with each feature's value having a specific coordinate value. Then, classification is carried out by identifying the hyper-plane that effectively distinguishes the two classes [17]. This work selects the hyper-plane that better segregate the two classes. Then it selects the hyper-plane with higher margin and robustness. Then, selecting the hyper-plane that correctly classifies the classes before optimizing the margin SVM has the advantage of ignoring outliers and choosing the hyper-plane with maximum margin.

Models using decision trees as regressors have a tree-like structure. It divides a dataset into ever-tinier sections while also developing an associated decision tree progressively. A tree containing decision nodes and leaf nodes will be used to display the ultimate outcome. Additionally simple to understand is this regression algorithm. Hence, this regressor can be chosen for this work [4].

Boosting is a method of converting weak learners into strong learners. Adaboost regressor is also a boosting algorithm. Here, the output of a weak regressor is sent to the next level regressor by adding high weights to the values that it failed to predict. The next level regressor concentrates more on these high-weighted values and tries to predict them. This process continues until the number of iterations that the user had given or no more regression is possible. It is efficient, simple, and easy to program, and it is also resistant to over-fitting. So this regressor is also chosen for the proposed work [3].

2.4. Performance evaluation metrics

The predictions of the proposed results are analyzed using metrics like precision, recall, and F1 score. The accuracy score metrics are used to evaluate the performance of applied models. Accuracy is used to calculate the performance of a model. It is used to know the accuracy of the learning models was over other existing models. This metric is used for the balanced type of dataset. The current adopted dataset in this work is having 719 ones and 640 zeros (nearly equal in number). Such that the dataset is balanced and hence is used metrics (1) gives the accuracy score calculation.

$$\text{Accuracy score } = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Where TP is true positive, TN is a true negative, FP is false positive, FN is a false negative [18], [19]. There are other metrics called Precision, Recall, F1 score [20], [21].

The proportion of accurately predicted positive observations to all expected positive observations is known as precision. [20], [22]. It is given in (2). Recall is the proportion of accurately anticipated positive observations to all of the actual class observations that we have observed [20], [22]. The equation for the calculation of recall is given in (3). F1 score is the weighted average of Precision and recall [20], [22] as given in (4).

$$\text{Precision } = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall } = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 score } = 2 \times \frac{(\text{Recall} \times \text{Precision})}{(\text{Recall} + \text{Precision})} \quad (4)$$

2.5. Data preprocessing

Data preprocessing is a demonstrated technique for resolving any issues with the dataset. The preprocessing checks for any missing values, perform normalization on numerical data to reduce the range of values of variables if necessary and convert numerical to categorical data. If all the attributes in the dataset are numerical, there is no need for any encoding. This work also checks for duplicates on the dataset and takes action accordingly [22]. Figure 1 gives the various steps involved in data preprocessing.

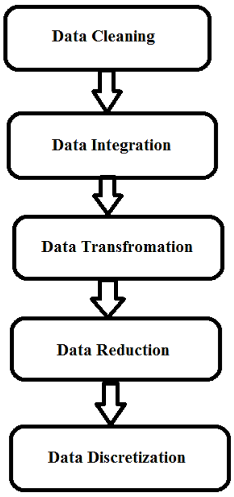

Figure 1. Data preprocessing

Data Cleaning: The data may contain a large number of irrelevant and missing pieces. Data cleansing is done to handle this portion. It entails dealing with erroneous data and noisy data. Data that is noisy cannot be interpreted by computers. It may be produced as a result of poor data gathering, incorrect data input [23]. Data integration: A data preparation method called data integration gives customers a uniform picture of this data by combining data from many sources. Data transformation: This step changes the data in appropriate forms suitable for the mining process by attribute selection, normalization and discretization [24].

Data Reduction: Data reduction is done to reduce the dataset size by considering only those data features which are relevant to the task. While working with huge amounts of data and many features, data reduction technique helps increase storage efficiency, reducing data storage and analysis costs. It includes dimensionality reduction, data cube aggregation, and numerosity reduction.

Data discretization is the process of grouping values into buckets or ranges, which reduces the number of potential values for the data and makes it discrete. The buckets themselves are handled as ordered and discrete values. Both numeric and string columns can be discretized [25].

3. RESULTS AND DISCUSSION

In Figure 2, the bar graph represents accuracy values obtained by various models. The horizontal plane represents models, and the vertical plane represents accuracy values. The ML methods are evoluated to find the accuracy of different models. The accuracy of models is followed as for logistic regression it is 0.621, for SVM classifier it is 0.625, for Adaboost classifier it is 0.0672, for the decision tree it is 0.665. Finally, for the random forest Classifier, it is 0.716. Actually, in the Figure 2, there are two random forest models called optimized and un-optimized. But to increase the model's performance, we used to optimize the random forest model by using hyper-parameters like max-features, n-estimators, max-depth, and criterion. By that, we got accuracy from 0.709 to 0.716.

The random forest model has good accuracy above 70% because random forest models are based on decision trees. We used hyper-parameters like n-estimators, max-depth, max-features, and criterion. Basically, the dataset is divided into random subgroups. Only a random set of features are taken into account at each node of the decision tree to determine the appropriate split since the decision tree model matches each subset. Voting predictions from all decision trees calculate the final Prediction. So, the random forest model can be used by the wine industry to check the consistency of the wine before it can be published on the market.

These existing results are taken from the references [5], [6], [12], [14]. Here along with accuracy, we have evaluated other metrics even though the dataset is balanced. Table 1 summarizes the accuracy performance results of various ML algorithms. In Table 1, the precision provides how relevant the positive detections are; the weighted Harmonic mean of accuracy and recall is used to get the F1 score. The recall is calculated as the number of accurate results divided by the number of results that should have been returned.

Figure 3 illustrates performance metrics such as precision, accuracy, F1 score and recall. We may deduce from Figure 3 that the random forest method performs best in terms of accuracy, precision, and recall. Thus, it can be concluded that the food industry may use the random forest method to assess the quality of wine.

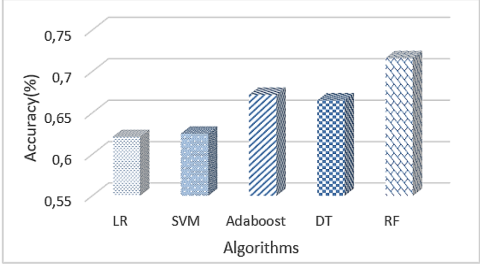

Figure 2. Accuracy scores for ML algorithms

Table 1. Performance analysis

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic regressor	0.621	0.653	0.648	0.650
SVM classifier	0.625	0.623	0.750	0.680
Adaboost classifier	0.672	0.692	0.716	0.704
Decision tree	0.665	0.696	0.682	0.689
Random forest	0.716	0.726	0.770	0.747

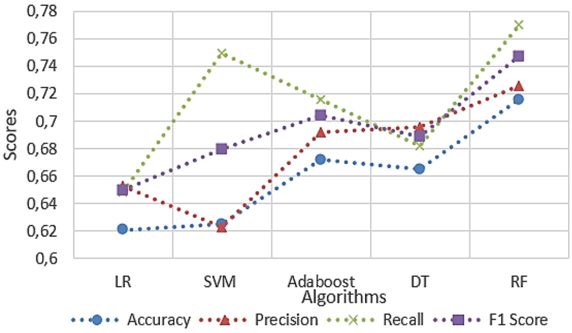

Figure 3. Performance metrics

4. CONCLUSION

This work estimates the quality of wine using various ML techniques. The results of each technique give different accuracy rates. Among those techniques, the random forest gives an accuracy above 70% to state the random forest gives the highest accuracy rate. Businesses are investing in new technologies to improve their production and distribution procedures as a result of the increase in wine consumption. The important phase of quality certification relies on human wine tasting at the moment. The experiments of this work show random forest ML techniques can predict a more accurate quality of the wine. Future forecasts of wine quality may need the use of various ML techniques and a huge dataset that can be used for tests.

REFERENCES

[1] B. Chen, V. Velchev, B. Nicholson, J. Garrison, M. Iwamura, and R. Battisto, “Wineinformatics: Uncork Napa’s cabernet Sauvignon by association rule based classification,” *Proceedings - 2015 IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015*, pp. 565–569, 2016, doi: 10.1109/ICMLA.2015.44.

[2] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, “Wineinformatics: Applying data mining on wine sensory reviews processed by the computational wine wheel,” *IEEE International Conference on Data Mining Workshops, ICDMW*, vol. 2015-January, no. January, pp. 142–149, 2015, doi: 10.1109/ICDMW.2014.149.

[3] Y. Gupta, “Selection of important features and predicting wine quality using machine learning techniques,” *Procedia Computer Science*, vol. 125, pp. 305–312, 2018, doi: 10.1016/j.procs.2017.12.041.

[4] P. Cortez, J. Teixeira, A. Cerdeira, F. Almeida, T. Matos, and J. Reis, “Using data mining for wine quality assessment,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5808 LNAI, pp. 66–79, 2009, doi: 10.1007/978-3-642-04747-3_8.

[5] Y. Er, “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, 2016, doi: 10.18201/ijisae.265954.

[6] J. Palmer, “Multi-Target Classification and Regression in Wineinformatics,” *University of Central Arkansas*, 2018.

[7] K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire, “Prediction of wine quality using machine learning algorithms,” *Open Journal of Statistics*, vol. 11, no. 02, pp. 278–289, 2021, doi: 10.4236/ojs.2021.112015.

[8] P. Theerthagiri, “Stress emotion recognition with discrepancy reduction using transfer learning,” *Multimedia Tools and Applications*, 2022, doi: 10.1007/s11042-022-13593-6.

[9] F. Fan, J. Li, G. Gao, and C. Ma, “Mathematical model application based on statistics in the evaluation analysis of grape wine quality,” *2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2015*, pp. 107–110, 2016, doi: 10.1109/ICCWAMTIP.2015.7493956.

[10] P. A. Flach and M. Kull, “Precision-recall-gain curves: PR analysis done right,” *Advances in Neural Information Processing Systems*, vol. 2015-Janua, pp. 838–846, 2015.

[11] G. Hu, T. Xi, F. Mohammed, and H. Miao, “Classification of wine quality with imbalanced data,” *Proceedings of the IEEE International Conference on Industrial Technology*, vol. 2016-May, pp. 1712–1717, 2016, doi: 10.1109/ICIT.2016.7475021.

[12] 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,” *International Conference on Advanced Communication Technology, ICACT*, vol. 2019-Febru, pp. 1122–1127, 2019, doi: 10.23919/ICACT.2019.8702017.

[13] R. Andonie, A. M. Johansen, A. L. Mumma, H. C. Pinkart, and S. Vajda, “Cost efficient prediction of Cabernet Sauvignon wine quality,” *2016 IEEE Symposium Series on Computational Intelligence, SSCI 2016*, 2017, doi: 10.1109/SSCI.2016.7849995.

[14] À. Nebot, F. Mugica, and A. Escobet, “Modeling wine preferences from physicochemical properties using fuzzy techniques,” *SIMULTECH 2015 - 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications, Proceedings*, pp. 501–507, 2015, doi: 10.5220/0005551905010507.

[15] B. Chen, C. Rhodes, A. Yu, and V. Velchev, “The computational wine wheel 2.0 and the trimax triclustering in wineinformatics,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9728, pp. 223–238, 2016, doi: 10.1007/978-3-319-41561-1_17.

[16] B. Flanagan, N. Wariishi, T. Suzuki, and S. Hirokawa, “Predicting and visualizing wine characteristics through analysis of tasting notes from viewpoints,” *Communications in Computer and Information Science*, vol. 528, pp. 613–619, 2015, doi: 10.1007/978-3-319-21380-4_104.

[17] P. Theerthagiri and A. U. Ruby, “RFFS: Recursive random forest feature selection based ensemble algorithm for chronic kidney disease prediction,” *Expert Systems*, 2022, doi: 10.1111/exsy.13048.

[18] A. Sinha and A. Kumar, “Wine quality and taste classification using machine learning model,” *International Journal of Innovative Research in Applied Sciences and Engineering*, vol. 4, no. 4, pp. 715–721, 2020, doi: 10.29027/ijirase.v4.i4.2020.715-721.

[19] C. Goutte and E. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *Lecture Notes in Computer Science*, vol. 3408, Springer Berlin Heidelberg, 2005, pp. 345–359.

[20] P. Theerthagiri, C. Gopala Krishnan, and A. H. Nishan, “Prognostic analysis of hyponatremia for diseased patients using multilayer perceptron classification technique,” *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 7, no. 26, 2021, doi: 10.4108/eai.17-3-2021.169032.

[21] P. Theerthagiri and J. Vidya, “Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques,” *Expert Systems*, 2022, doi: 10.1111/exsy.13064.

[22] S. Lee, J. Park, and K. Kang, “Assessing wine quality using a decision tree,” *1st IEEE International Symposium on Systems Engineering, ISSE 2015 - Proceedings*, pp. 176–178, 2015, doi: 10.1109/SysEng.2015.7302752.

[23] A. Frank and A. Asuncion, “UCI machine learning repository,” 2010, [Online]. Available: http://archive.ics.uci.edu/ml.

[24] B. Chen, H. Le, C. Rhodes, and D. Che, “Understanding the wine judges and evaluating the consistency through white-box classification algorithms,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9728, pp. 239–252, 2016, doi: 10.1007/978-3-319-41561-1_18.

[25] Mahima, U. Gupta, Y. Patidar, A. Agarwal, and K. P. Singh, “Wine quality analysis using machine learning algorithms,” in *Micro-Electronics and Telecommunication Engineering, Lecture Notes in Networks and Systems, vol. 106*, D. K. Sharma, V. E. Balas, L. H. Son, R. Sharma, and K. Cengiz, Eds. Springer, Singapore, 2020, pp. 11–18, doi: 10.1007/978-981-15-2329-8_2.

BIOGRAPHIES OF AUTHORS

Sathishkumar Mani has obtained his B.E degree in Computer Science and Engineering from Bharathiar University, Coimbatore, India and M.Tech degree in Information Technology from Punjabi University, Patiala, India. He earned his Ph.D in Computer Science and Engineering from Saveetha University, Chennai, India. He has over 25 years of experience in multiple domains like teaching, research and software development. His research area is network security, Machine Learning and IOT. He can be contacted at email: sathishkumarmani17@gmail.com.

Reshmy Avanavalappil Krishnankutty is working as an Assistant Professor in the Department of Computational Intelligence, SRM Institute of Science and Technology, Chennai, TamilNadu, India. She has more than 18 years of experience in teaching and research. She has completed her doctorate (Ph.D. in Information and Communication Engineering) from Anna University in the field of Big Data. Her current research focus is on Data Science, Machine Learning. She received Professor T.R. Natesan Endowment Award (Gold Medal-Instituted by Operational Research Society of India). She can be contacted at email: reshmyak@gmail.com.

Sabaria Swaminathan working as an Assistant Professor in B.S. Abdur Rahman Crescent Institute of Science and Technology. She completed my M.E. Computer Science and Engineering in Srinivasan Engineering College. She is having 8 years of teaching experience. Her research interests are Machine learning and Artificial Intelligence. She can be contacted at email: ssabaria89@gmail.com.

Prasannavenkatesan Theerthagiri is working as the Assistant Professor in the Department of Computer Science and Engineering, GITAM Deemed to be University, Bengaluru, India. He was awarded PhD (Full-Time) degree in the year 2021 on the work of wireless communication with Machine Learning from Anna University, Chennai, India. He was awarded the Mobility grant award by the Republic of Slovenia in the year 2017-2018. He has published his research works in 12 SCI indexed journals, 16 SCOPUS indexed journals. His research interests are Data Science, AI, IoT, MANET. He can be contacted at email: prasannait91@gmail.com.