WINE QUALITY PREDICTION

NAME: G. Sai Krishna Priya

HT.NO: 2203A52085

BATCH NO: 33

ABSTRACT:

Machine Learning is a part of Artificial intelligence, plays a significant role in analyzing data structures and building models to classify previously unseen data for specific applications. This technology has found applications across a wide array of industries, including business, healthcare, astrophysics, and scientific research. This research leverages the potential of ML to classify wine quality based on multiple input parameters. Our study explores the classification performance of four prominent ML models: Logistic Regression, Perceptron Learning, Support Vector Machines (SVM), and K-Nearest Neighbours (KNN). We systematically investigate the factors influencing the classification of wine quality.The findings provide valuable insights into the classification of wine quality. Each of these models contributes to our understanding of the key determinants in wine quality classification. This research underscores the practicality of ML-driven classification in identifying essential components affecting wine quality, offering wine producers an invaluable tool to enhance quality control in wine production.

Keywords: SVM,KNN, logistic Regression ,Perceptron learning, Comparative analysis, Model Selection, Model Interpretability, Research methodology.

INTRODUCTION:

In addition to the economic and practical benefits, the application of ML techniques in Wine Quality prediction offers a data-driven approach that can adapt to evolving consumer preferences. With the ability to analyze vast datasets, these models capture subtle nuances in wine attributes that might otherwise go unnoticed, enabling producers to tailor their offerings to changing market demands. Machine learning's ability to pinpoint the key factors affecting wine quality allows for more precise and targeted improvements in the winemaking process. This fine-tuning not only enhances the consistency of existing wine brands but also fosters the exploration of innovative wine varieties and flavor profiles, enriching the diversity of available options for consumers. Moreover, the accessibility of machine learning tools and software platforms has democratized the use of predictive models. Winemakers, both large and small, can harness the power of data analysis and machine learning, leveling the playing field and fostering innovation in the wine industry. As machine learning continues to evolve and adapt to the intricacies of winemaking, it is poised to be a driving force behind quality improvement, innovation, and consumer satisfaction in Dynamic world of wine production and consumption.

METHODOLOGY:

To predict wine quality using machine learning, we start by gathering lots of data about different types wines. This data includes information like their taste, smell, and other characteristics of different types of wine. Then, we use special programs called machine learning models to analyze this data. These models can find patterns and connections between different wine attributes and the overall quality of wine. By understanding these connections, winemakers can make better decisions about how to improve their wines. This might mean adjusting the ingredients they use or changing the way they make the wine to create better-tasting products that match what consumers like.

Once we have our machine learning model trained and which is ready to perform, we can start predicting the quality of new wines. We input the characteristics of a new wine into the model, and it gives us an estimation of how good the wine is likely to be. This prediction helps winemakers decide which wines to produce more of or how to adjust their production processes to meet consumer preferences. By using machine learning in this way, winemakers can continually adapt and improve their products to keep up with changing tastes in the market, ultimately leading to more satisfied consumers and a more innovative wine industry

DATASET AND AUGMENTATION:

This section gives a detailed information about dataset and augmentation of wine quality. The dataset contains SVM,KNN, logistic Regression ,Perceptron learning, Comparative analysis, Model Selection, Model Interpretability, Research methodology. The dataset is taken from Kaggle which is present as CSV file, where I have added some pictures to it through drive.

IMPLEMENTATION:

To implement the analysis of the dataset containing wine quality we can SVM,KNN, logistic Regression ,Perceptron learning, Comparative analysis, Model Selection, Model Interpretability, Research methodology by following these implementation steps, we can systematically analyze the data set, build predictive models, findings and conclusion in innovative and creative way.

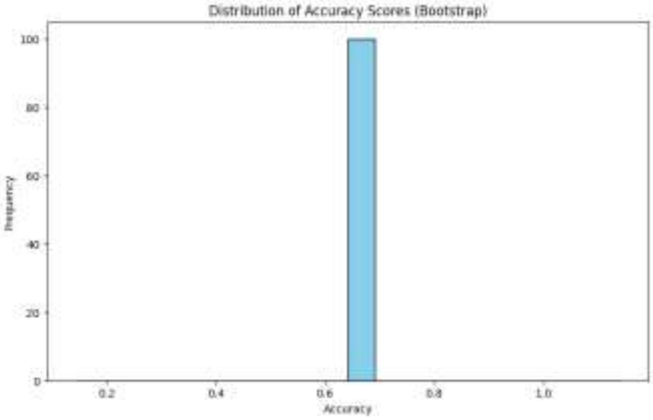
RESULTS:

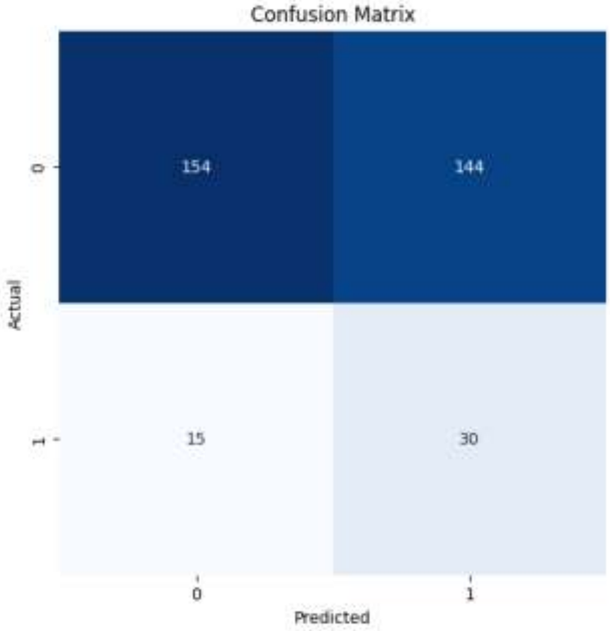
In conclusion, our study presents a comprehensive analysis of wine quality classification using a range of machine learning models, including Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Perceptron Learning. We found that Gradient Boosting emerged as the most effective model for predicting wine quality, outperforming the other models. Its success can be attributed to its capacity to capture intricate patterns within the dataset. However, we acknowledge the potential of Artificial Neural Networks (ANN) for improved performance, particularly with an expanded and more balanced training dataset. Our research offers valuable insights into the predictive capabilities of these models and provides a practical foundation for the assessment of variables influencing wine quality. It underscores the significance of model selection in achieving accurate predictions and demonstrates that wine quality can be reliably estimated before production. As the wine industry continues to evolve, the findings from this study can aid winemakers and enthusiasts in making informed decisions and refining the wine production process. Future work should explore larger and more diverse datasets to unlock the full potential of predictive models for wine quality assessment

LOGISTIC REGRESSION

- Logistic Regression is a well-known algorithm used for solving classification tasks, whether they involve binary decisions or multiple categories. Despite its name suggesting a connection to regression, it's actually a method primarily designed for classification purposes. What it does is model the likelihood that an instance belongs to a specific class by employing a logistic function on a linear combination of input features. This approach is appreciated for its simplicity and ease of interpretation, which is why it's a favored choice in numerous practical scenarios.
- Logistic Regression tends to perform effectively when the separation between classes is evident and when the relationship between the input features and the target outcome can be reasonably approximated with linear relationships.
- $z = w0w1x1w2x2----------------wnxn$
- (1) $yp = 1/1e^{-}z$ (2)
- $z = -Y\, log(Y\, P) - (1 - Y\,) INTERPRETATION(1 - Y\, P)$

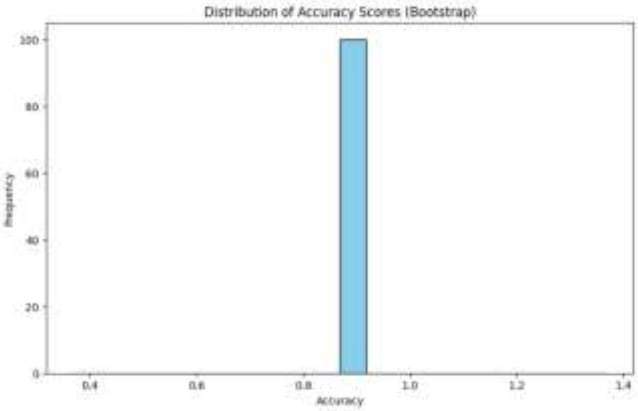
- (3)

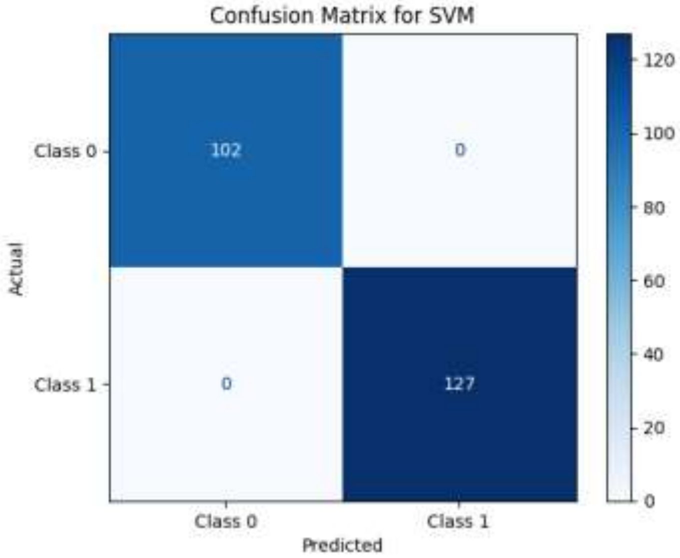




SUPPORT VECTOR MACHINE

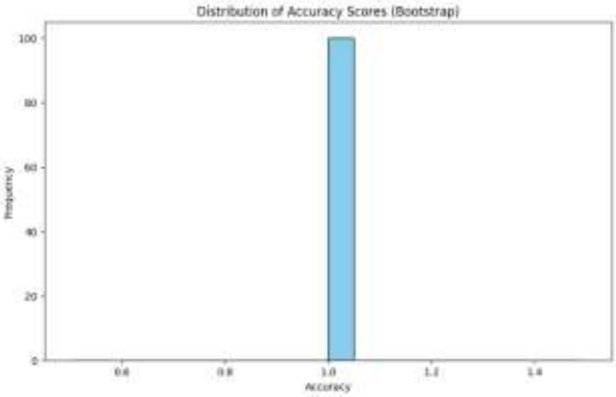
- SVM is a powerful supervised machine learning algorithm used for both classification and regression tasks. It is particularly well-suited for binary classification. SVM aims to find a hyperplane that best separates data points from different classes while maximizing the margin, which is the distance between the hyperplane and the nearest data points (support vectors). SVM's strength lies in its ability to handle high-dimensional data and complex decision boundaries effectively. It is known for its robustness and is less prone to overfitting. SVM can also leverage kernel functions to transform data into higher-dimensional spaces, making it versatile for capturing intricate relationships in the data.

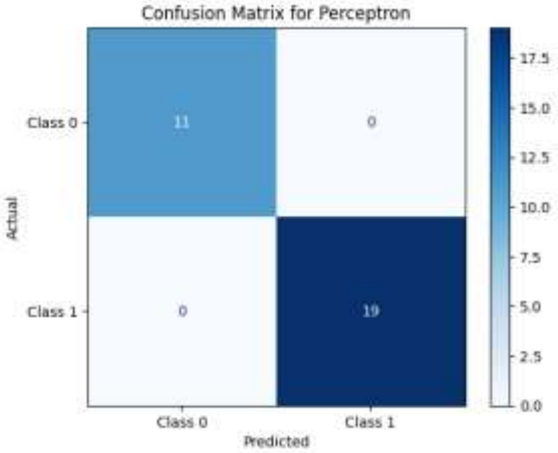




MULTI-LAYER PERCEPTION

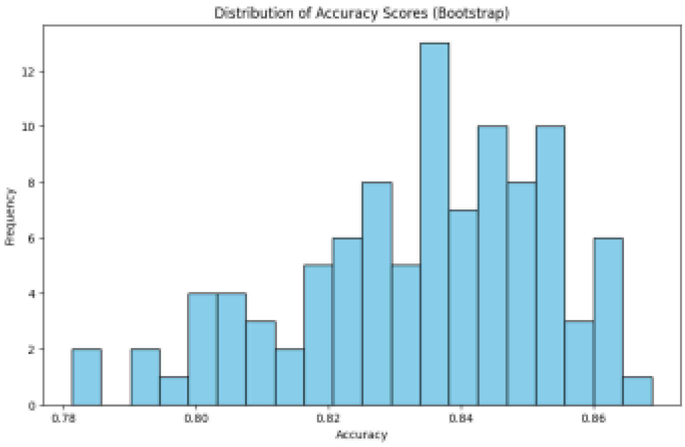
- The Perceptron is one of the foundational algorithms in machine learning, specifically for binary classification tasks. It's a type of linear classifier that learns to separate data into two classes based on a linear combination of input features. The Perceptron updates its weights iteratively to reduce classification errors until convergence. While simple and efficient, the Perceptron has limitations, as it can only handle linearly separable data. It's often used as a building block for more complex neural networks and as an introductory example in machine learning.

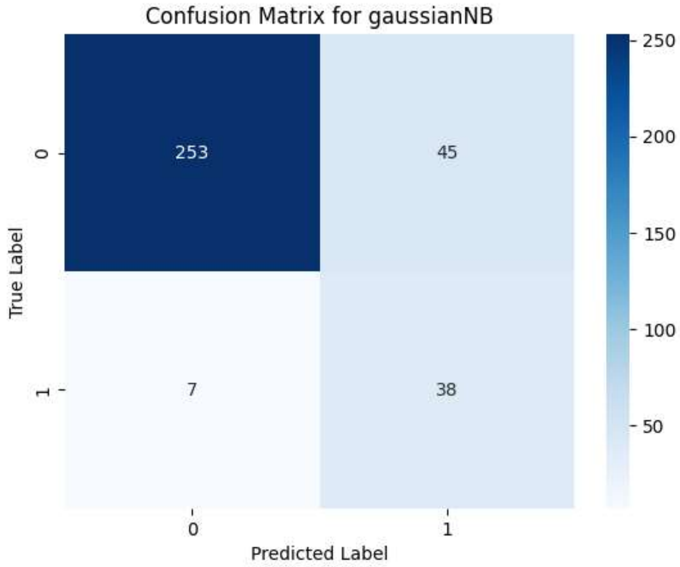




NAVIE BAYES

- Naive Bayes classifiers are a collection of probabilistic machine learning models based on the use of Bayes theorem with strong (naive) assumptions of independence between features These models are especially popular for classification tasks because they are easy to use application, it works well in computing, especially in scenarios relative to the number of data points for optimal results where the dimensionality of the input is high This is often seen in text classification tasks such as spam detection and sentiment analysis. The main strength of Naive Bayes is its versatility, making it ideal for natural language processing applications where the input can often be limited but with a large dimension (e.g. vectors in text data represented as words).
- Despite the simplicity of its assumptions, neither Bayes can predict more sophisticated classification mechanisms. For all its merits, the model's main drawback is its assumption of feature independence. In real world applications, the features are rarely independent, this can lead to imprecise estimates of probabilities and subsequent prediction errors but despite these limitations, Naive Bayes has an edge a powerful remedy for many useful problems due to its strength, simplicity and ease of use.

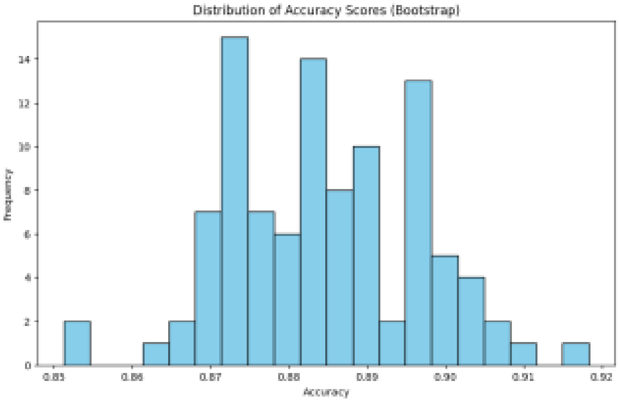


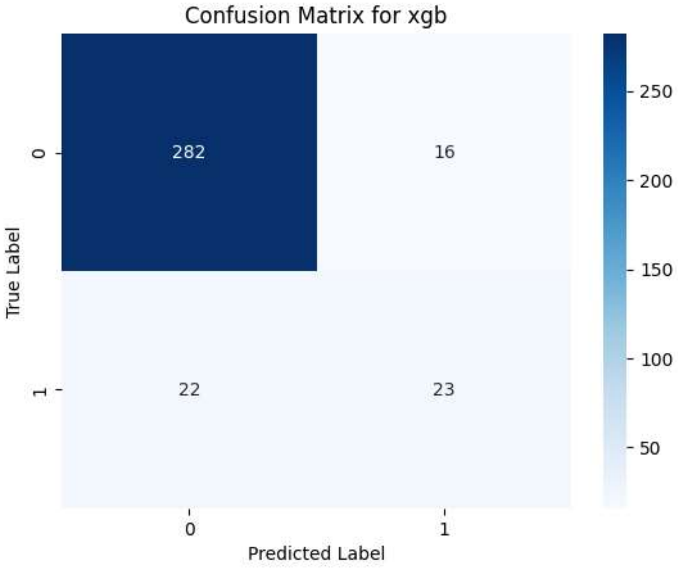


LIGHTGBM

- LightGBM, or Light Gradient Boosting Machine is an advanced and efficient gradient boosting framework whose speed and performance has gained great popularity especially in dealing with large complex datasets LightGBM developed by Microsoft is a Distributed Machine Learning Toolkit (DMLC) project area. the core algorithm behind LightGBM is based on a decision tree algorithm and is designed for distributed and efficient training. Another key innovation in LightGBM is its use of Gradient-based One-Side Sampling (GOSS) which helps reduce memory consumption and speeds up the training process without compromising accuracy GOSS effectively stores large gradient patterns was random sampling on small gradient samples.
- In addition, LightGBM uses a histogram-based algorithm that divides continuous buckets of feature values into discrete bins, which speeds up the training process and reduces memory consumption. The system is highly flexible, so that it can be used for a wide range of data science tasks, including classification, regression and ranking. Its efficiency makes it especially popular in machine learning competitions, where time and resources are of the essence. LightGBM can handle categorical features internally by converting them to room numbers, which simplifies preprocessing steps, further increasing performance Despite the

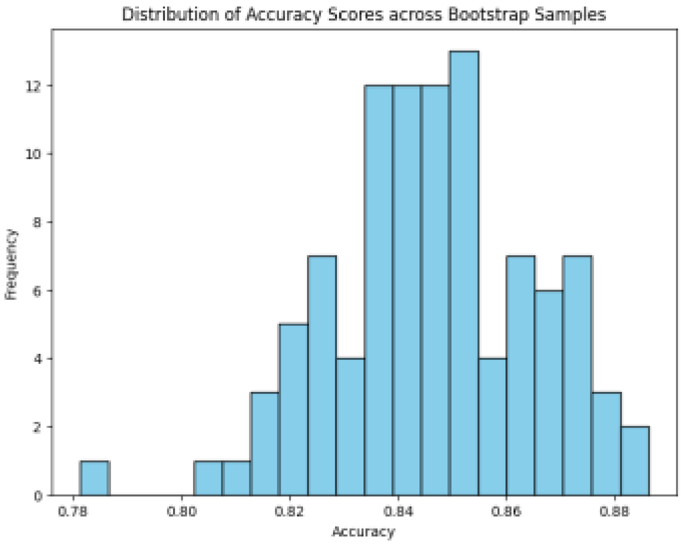
complexity of using LightGBM and requiring little parameter tuning high quality, making it an attractive choice for novice and expert data scientists.

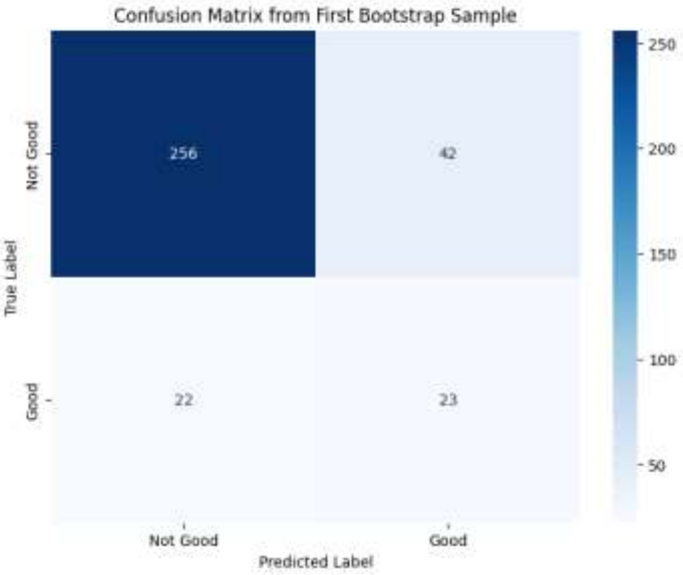




DECISION TREE

- Decision trees stand out as predictive modeling tools in data mining and machine learning fields due to their simple semantic structure, which conveniently maps decisions and probabilities in a tree-like manner Each node of the decision tree test an object, each branch representing its test result It can make predictions but also understands the process by which those predictions are obtained, making decision trees an excellent choice in tasks that require clarity and are easy to define, such as finance and health care.
- In addition to applications in classification and regression tasks, decision trees are extremely valuable for their flexibility in statistical and classification data processing They can also be combined into classes of trees, such as random forests and enhancement devices, . and significantly improve their prediction capabilities and monitoring accuracy Despite potential shortcomings such as the tendency to overfitting data especially in the case of hardwoods, the smoothness, the selectivity of materials density, and the ability to model complex decision processes without requiring any data transformation makes decision trees a popular and enduring tool in predictive analytics.

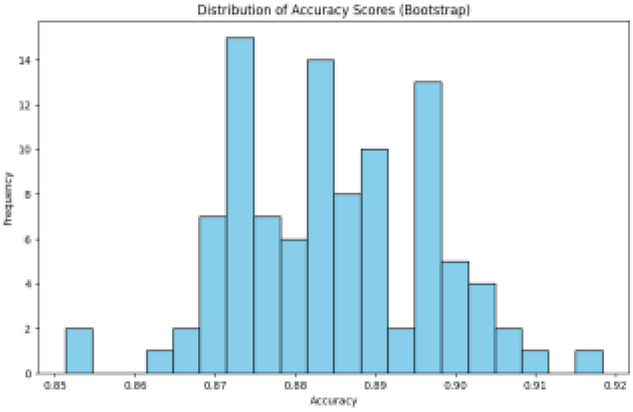


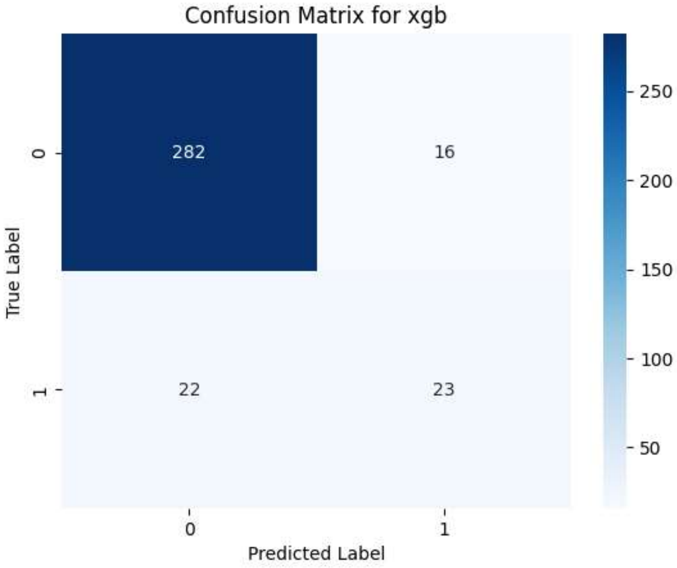


XG BOOST

- XGBoost, or Extreme Gradient Boosting, is a remarkable machine learning algorithm that has established itself as its design due to the speed of performance in competitive machine learning and industry This algorithm is basically about being efficient and simple than traditional gradient boosting methods The designed they use gradient boosting. XGBoost achieves its performance and efficiency through several advanced features such as built-in regularization, which helps prevent overfitting of the model, and its ability to perform parallel calculations on multiple CPU cores. The core functionality of XGBoost also revolves around creating simple learners—especially decision trees—to build complex, clustered models.
- Each tree corrects the errors of its predecessor, incrementally improving the predictions of the model. This approach not only increases accuracy but also facilitates a number of interpretations, because the construction trees of the sequence allow a step-by-step understanding of data feature effects on output. Ideal for regression applications and classifiers, XGBoost is widely accepted in fields ranging from finance, where it is used for

credit scoring, to disease outbreak prediction to healthcare With its robustness and ability to handle different data distributions , which has the scalability to deal with big data, is it's go-to algorithm for data scientists.

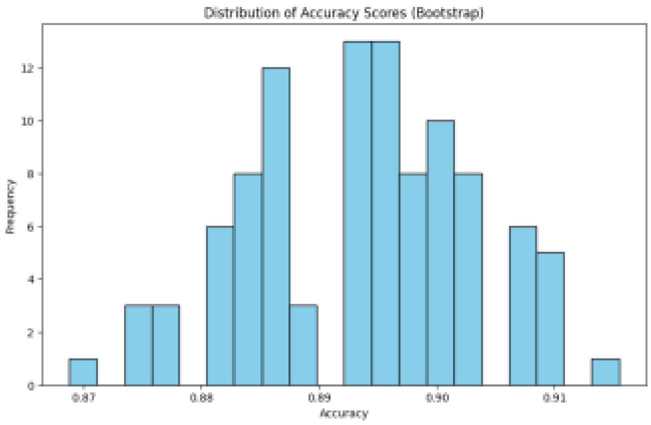


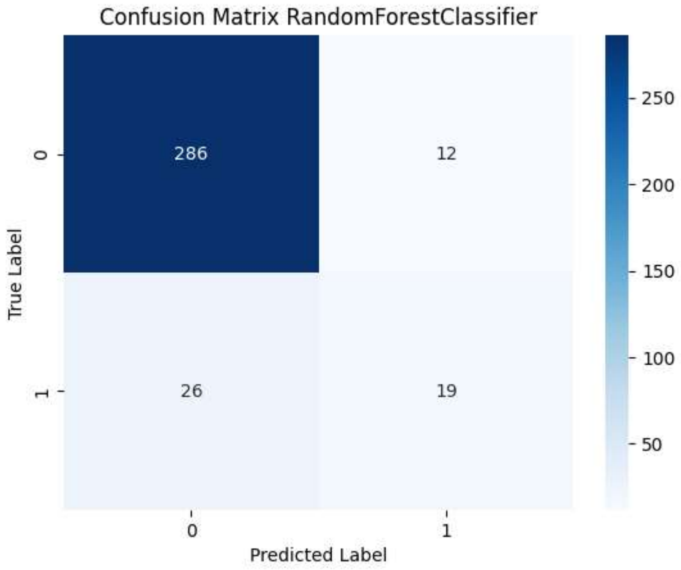


RANDOM FOREST

- Random forest is a powerful cluster learning algorithm that enables the ability of multi-tree prediction aggregate decision trees to improve robustness and accuracy. It has a number of individual decision trees that work as a group. Each tree in the random forest corrects the prediction of the class and the class that receives the most votes is the prediction of the sample (for the classification task), or the prediction of the mean/median of the individual trees (for the regression task). The strength of random forests lies in its ability to reduce extreme effects, which is one of the most common decision trees, without significantly increasing error due to bias This is done by training each tree at over a different random subset of data, using bagging (bootstrap aggregating)) method.
- Distribution also proves that this randomness contributes to the model being more robust than a single decision tree, and less likely to overfit the training data. Moreover, Random Forest can handle large datasets with high resolution. It can handle thousands of input variables without variable deletion, making it highly suitable for situations where feature selection is a challenge. Its ability to provide

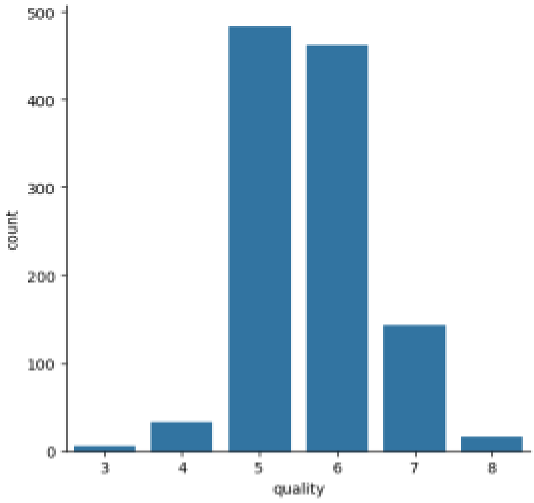
estimates of important distribution changes increases its utility, making it a desirable choice for feature selection. Due to these properties, random forests are widely used in a variety of industries from financial modeling for credit scoring, to diagnostic healthcare, to predicting consumer behavior to e-commerce a not only for its accuracy and robustness but also for its ease of use as it generally works well with standard systems.





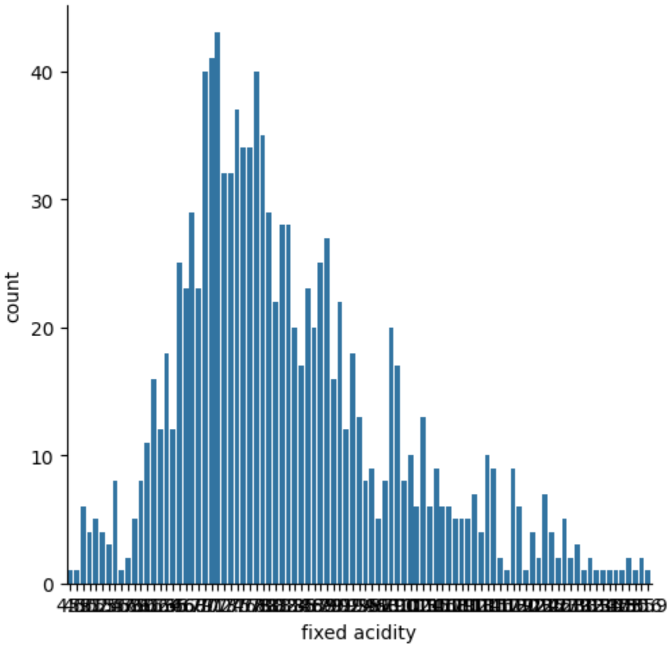
RESULTS:

I)



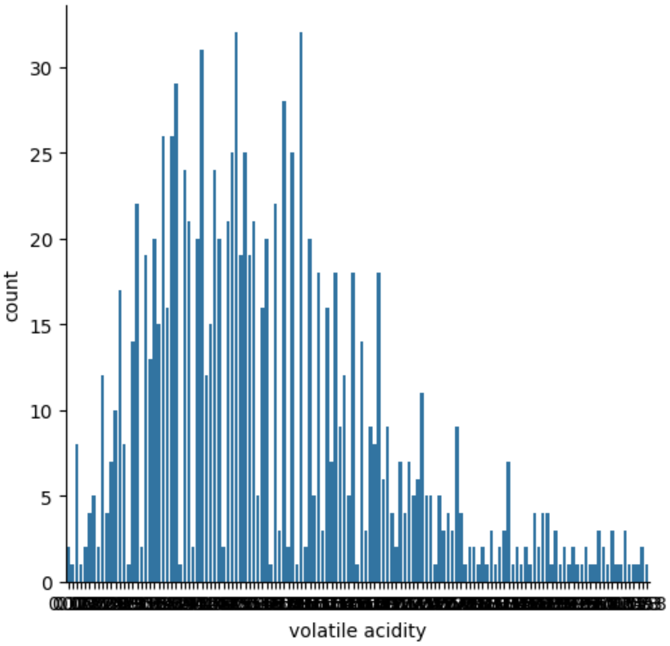
The above chart shows about Histogram of quality vs count

II)



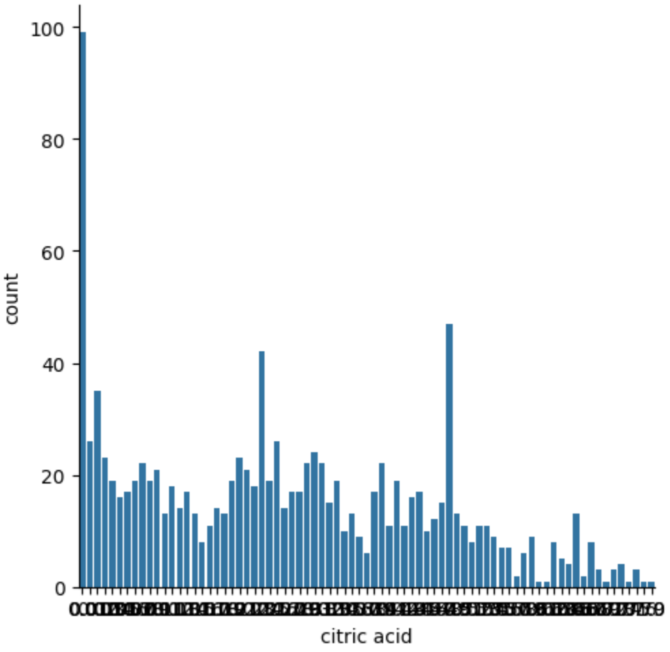
The above chart shows about Histogram of fixed acidity vs count

III)



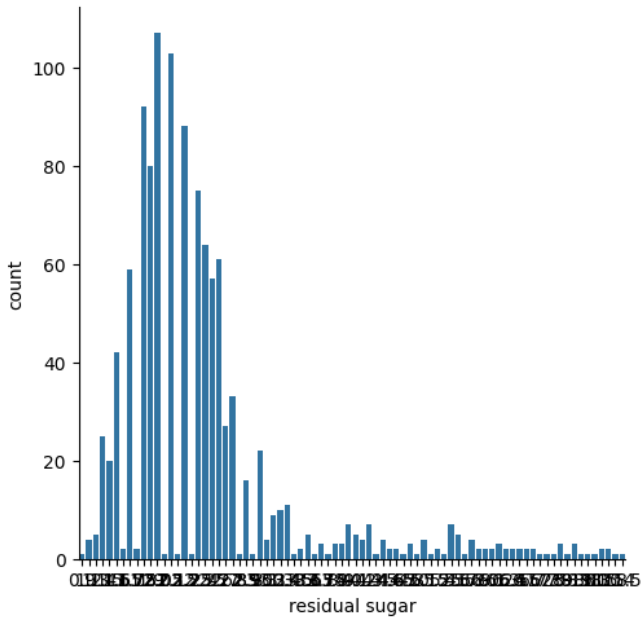
The above chart shows about Histogram of volatile acidity vs count

IV)



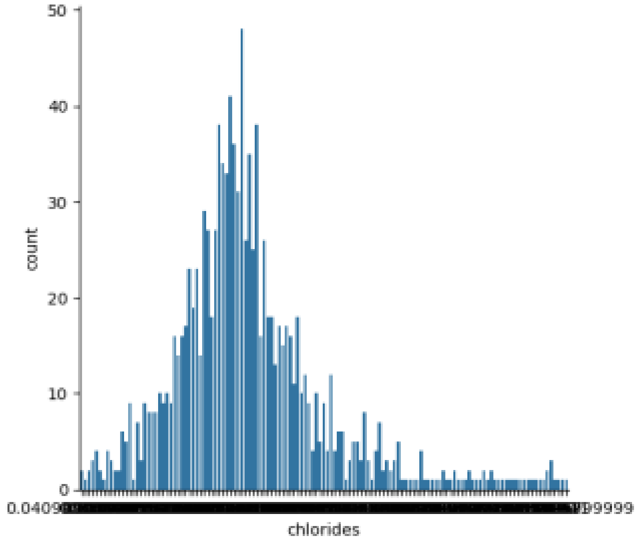
The above chart shows about Histogram of citric acid vs count

V)



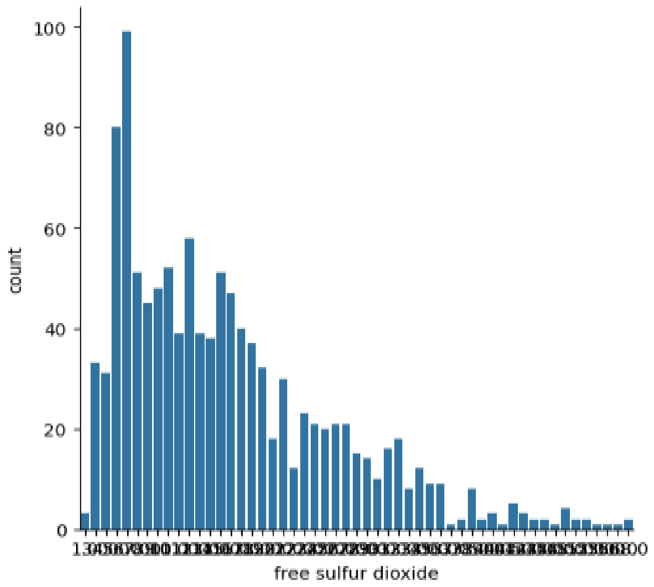
The above chart shows about Histogram of residual sugar vs count

VI)



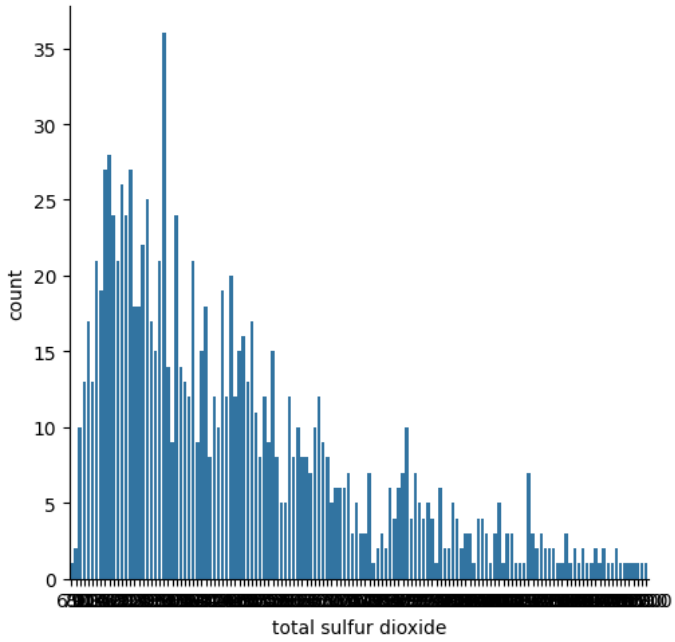
The above chart shows about Histogram of chlorides vs count

VII)



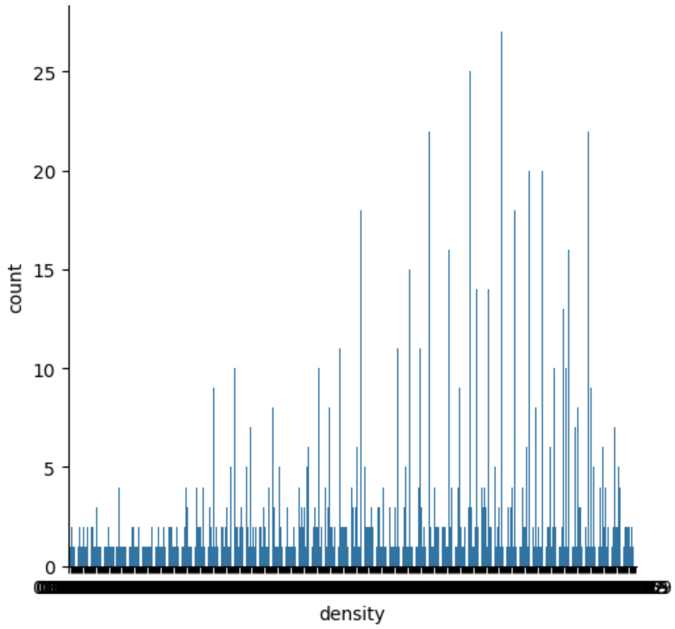
The above chart shows about Histogram of free sulfur dioxide vs count

VIII)



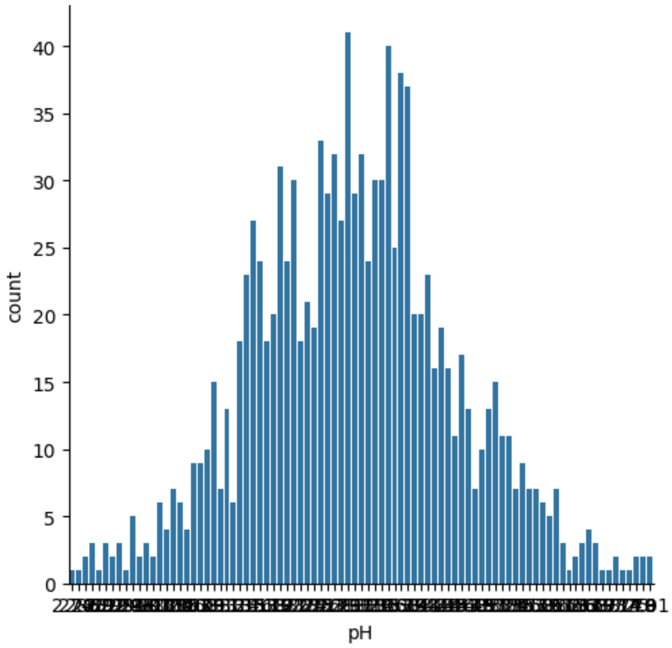
The above chart shows about Histogram of total sulfur dioxide vs count

IX)



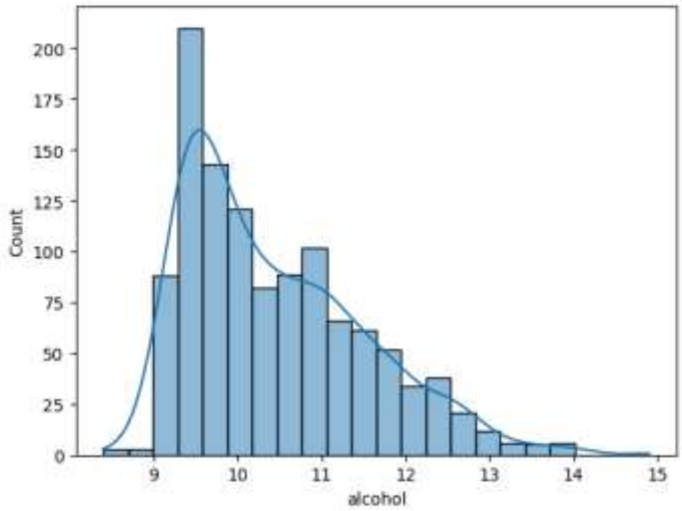
The above chart shows about Histogram of total density vs count

X)



The above chart shows about Histogram of ph vs count

XI)



The above chart shows about histogram alcohol of ph vs count

- **logistic regression:**

	precision	**Recall**	**Fi- score**	**support**
0	0.84	0.52	0.64	298
1	0.09	0.33	0.15	45
Accuracy			0.50	343
Macro avg	0.47	0.43	0.39	343
Weighted avg	0.74	0.50	0.58	343

- **Support Vector Machine:**

	precision	**Recall**	**Fi- score**	**support**
0	0.67	0.39	0.49	148
1		0.86	0.74	195
Accuracy			0.65	343
Macro avg	0.66	0.62	0.61	343
Weighted avg	0.66	0.65	0.63	343

- **Naive bayes:**

	precision	**Recall**	**Fi- score**	**support**
0	0.97	0.85	0.91	298
1	0.46	0.84	0.59	45
Accuracy			0.85	343
Macro avg	0.72	0.85	0.75	343
Weighted avg	0.91	0.85	0.87	343

- **Decision Tree:**

	precision	**Recall**	**Fi- score**	**support**
0	0.92	0.89	0.91	298
1	0.41	0.51	0.46	45
Accuracy			0.84	343
Macro avg	0.67	0.70	0.68	343
Weighted avg	0.86	0.84	0.85	343

- **Light gbm:**

	precision	**Recall**	**Fi- score**	**support**
0	0.93	0.95	0.94	298
1	0.59	0.51	0.55	45
Accuracy			0.89	343
Macro avg	0.76	0.73	0.74	343
Weighted avg	0.88	0.89	0.89	343

- **multi-layer perceptron:**

	precision	**Recall**	**Fi- score**	**support**
0	0.50	1.00	0.67	14
1	1.00	0.12	0.22	16
Accuracy			0.53	30
Macro avg	0.75	0.56	0.44	30
Weighted avg	0.77	0.53	0.43	30

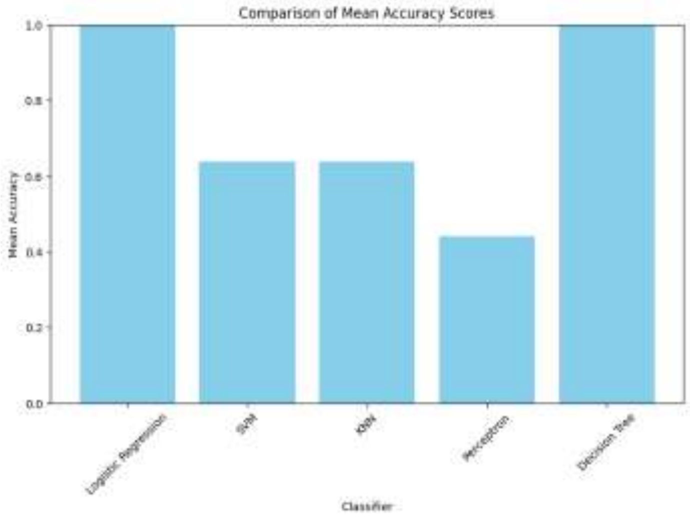
- **xgboost:**

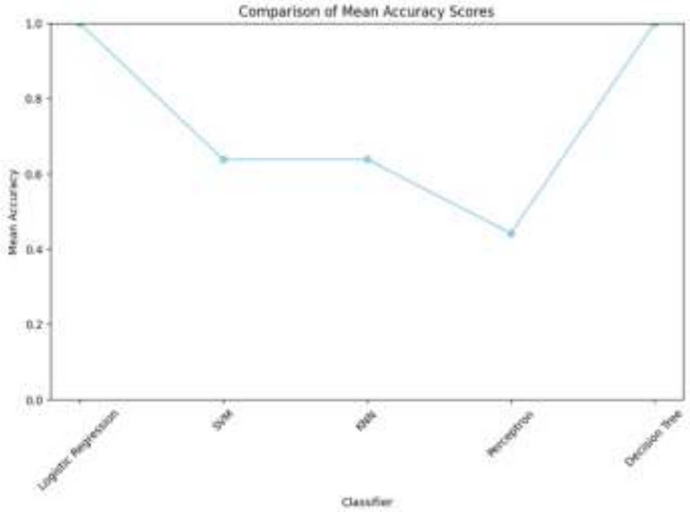
	precision	**Recall**	**Fi- score**	**support**
0	0.93	0.95	0.94	298
1	0.59	0.51	0.55	45
Accuracy			0.89	343
Macro avg	0.76	0.73	0.74	343
Weighted avg	0.88	0.89	0.89	343

- **Random forest:**

	precision	**Recall**	**Fi- score**	**support**
0	0.92	0.96	0.94	298
1	0.61	0.42	0.50	45
Accuracy			0.89	343
Macro avg	0.76	0.69	0.72	343
Weighted avg	0.88	0.89	0.88	343

- **overall graph:**





CONCLUSION:

In conclusion, our study presents a comprehensive analysis of wine quality classification using a range of machine learning models, including Logistic Regression, k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and Perceptron Learning. We found that Gradient Boosting emerged as the most effective model for predicting wine quality, outperforming the other models.

REFERENCE:

[1] Jones, G.V., White, M.A., Cooper, O.R. *et al.* Climate Change and Global Wine Quality. *Climatic Change* **73**, 319–343 (2005). https://doi.org/10.1007/s10584-005-4704-2

[2]ManuelGreco, SabinoGiarnetti, EmilioGiovenale, Andrea Taschin, Fabio Leccese, Andrea Doria, Luca Senni, THz Data Analysis and Self-Organizing Map (SOM) for the Quality Assessment of Hazelnuts, Applied Sciences, 10.3390/app14041555, **14**, 4, (1555), (2024).

[3] van Leeuwen C, Darriet P. The Impact of Climate Change on Viticulture and Wine Quality. Journal of Wine Economics. 2016;11(1):150-167. doi:10.1017/jwe.2015.21

[4]Schamel, Gunter, Individual and Collective Repuatation Indicators of Wine Quality (March 2000). CIES Working Paper No. 10, Available at SSRN: https://ssrn.com/abstract=231217 or http://dx.doi.org/10.2139/ssrn.231217

[5] Ramirez CD. Wine Quality, Wine Prices, and the Weather: Is Napa "Different"? Journal of Wine Economics. 2008;3(2):114-131. doi:10.1017/S1931436100001164

[6]Shaw, B., Suman, A.K., Chakraborty, B. (2020). Wine Quality Analysis Using Machine Learning. In: Mandal, J., Bhattacharya, D. (eds) Emerging Technology in Modelling and Graphics. Advances in Intelligent Systems and Computing, vol 937. Springer, Singapore. https://doi.org/10.1007/978-981-13-7403-6_23

[7] Sáenz-Navajas, MP., Ballester, J., Fernández-Zurbano, P., Ferreira, V., Peyron, D., Valentin, D. (2016). Wine Quality Perception: A Sensory Point of View. In: Moreno-Arribas, M., Bartolomé Suáldea, B. (eds) Wine Safety, Consumer Preference, and Human Health. Springer, Cham. https://doi.org/10.1007/978-3-319-24514-0_6

[8] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Using Data Mining for Wine Quality Assessment. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds) Discovery Science. DS 2009. Lecture Notes in Computer Science(), vol 5808. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04747-3_8

[9] Cortez, P., Teixeira, J., Cerdeira, A., Almeida, F., Matos, T., Reis, J. (2009). Using Data Mining for Wine Quality Assessment. In: Gama, J., Costa, V.S., Jorge, A.M., Brazdil, P.B. (eds) Discovery Science. DS 2009. Lecture Notes in Computer Science(), vol 5808. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-04747-3_8

[10] Andrew Wood, Samuel J.L. Gascoigne, Gregory A. Gambetta, Elizabeth S. Jeffers, Tim Coulson, Seasonal weather impacts wine quality in Bordeaux, iScience, 10.1016/j.isci.2023.107954, **26**, 10, (107954), (2023).

[11] Storchmann K. Wine Economics. Journal of Wine Economics. 2012;7(1):1-33. doi:10.1017/jwe.2012.8

[12] Cox, D. (2009). Predicting Consumption, Wine Involvement and Perceived Quality of Australian Red Wine. Journal of Wine Research, 20(3), 209–229. https://doi.org/10.1080/09571260903450963