Wine Quality Prediction Using Machine Learning

G. Saranya
Asst.Professor
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet,
Andhra Pradesh
dasarisaranya4@gmail.com

Siddu Chennamsetty
Student
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet,
Andhra Pradesh
siddualiea@gmail.com

Pavan Kumar Rachupalli
Student
Computer Science and Engineering
Narasaraopeta Engineering College
Narasaraopet,
Andhra Pradesh
pavan1234@gmail.com

ABSTRACT

Wine is enjoyed worldwide by people of all genders. The wine the better it tastes, It can also be more expensive. To determine wine quality, key factors, like Sulphur dioxide, Volatile acidity, Citric Acid and Residual sugar are considered. Traditionally assessing wine quality was a time-consuming process. This project uses machine learning techniques such as Neural Networks, Logistic Regression and Support Vector Machine to predict whether a wine is good or bad. These methods are applied to datasets of Vinho Verde" Wine. Compared against standard benchmarks. This research proves valuable in the wine industry, for ensuring quality testing and customer satisfaction.

Keywords—RandomForestClassifier, Machine Learning, Classification, Feature Engineering, Data Analysis.

I. INTRODUCTION

Wine is one of the most popular drinks in the world. The components used in wine manufacturing determine its quality. Wine's quality varies depending on numerous factors; generally speaking, the older the better. In the current limited market, wine quality is crucial for both consumers and manufacturers to drive expansion in output.

.

All wines can be divided into five basic categories: dessert wine, red wine, rose wine, sparkling wine, and white wine. In order to make red wine, black grapes are utilized. Red wine often ranges in color from light to dark. Wine made from green and black grapes is known as white wine. Figure displays the appearance of the Red and White Wines.

Many approaches must be used from the beginning to improve the quality of the wine if it is of poor quality. Each person has an opinion about the quality of the wine, mostly based on taste. It's noteworthy to note that wine quality is categorized based on individual preferences. Thanks to advancements in a variety of technology, wine producers may now rely on a multitude of methods for ensuring wine quality.

The producers will then have a better understanding of wine quality. These days, all manufacturers use various techniques to maximize output and build a well-organized process throughout the whole process. With time, these techniques become more and more elegant, yet they also become more challenging.

Over the past few years, India's wine industry has expanded. Several of India's leading states, including Maharashtra, Karnataka, Andhra Pradesh, and Himachal Pradesh, have made significant contributions to the development of wine manufacturing and wine-making processes.

Red wine consumption Controlled wine consumption may improve digestive, mental, and cardiac health among other health issues. Because it contains composites that have lipid-improving, anti-inflammatory, and antioxidant qualities.

Chocolate and cosmetics such as cleansers and perfumes can be made using wine.. Stable acidity, total sulfur dioxide, volatile acidity, residual sugar, chlorides, citrus acid free sulphur dioxide, solidity, PH, sulphates, alcohol, etc. are the factors that are most important in determining the quality of wine.

II. LITERATURE REVIEW

1 Together, these studies explore the use of several machine learning and artificial intelligence methods to forecast wine quality based on sensory and physicochemical characteristics. Many approaches, including decision trees, random forests, support vector machines, artificial neural networks (ANNs), deep learning models, Jie Zhang, Emad Abuelrub, and Majdi Mafarja, are explored by researchers like Sachin Pawar, Chuan Sun, Luís Torgo, Paulo Cortez, Utku Kose, Sarah I. Clopp, Joydeep Ghosh, Marcos E. Nascimento, Ahmad Taher Azar, and Majdi Mafarja. Their research assesses feature engineering and selection tactics, compares algorithm performance, and talks about how model architectures and training techniques affect how accurate wine quality forecasts .

In order to provide more detailed analysis, some also take into account a wider range of wine varietals and preferences, utilizing sophisticated hybrid and deep learning models. These studies demonstrate the variety of methods and their potential influence on the wine industry, while also contributing to the continuous advancement of predictive analytics in viticulture and oenology..

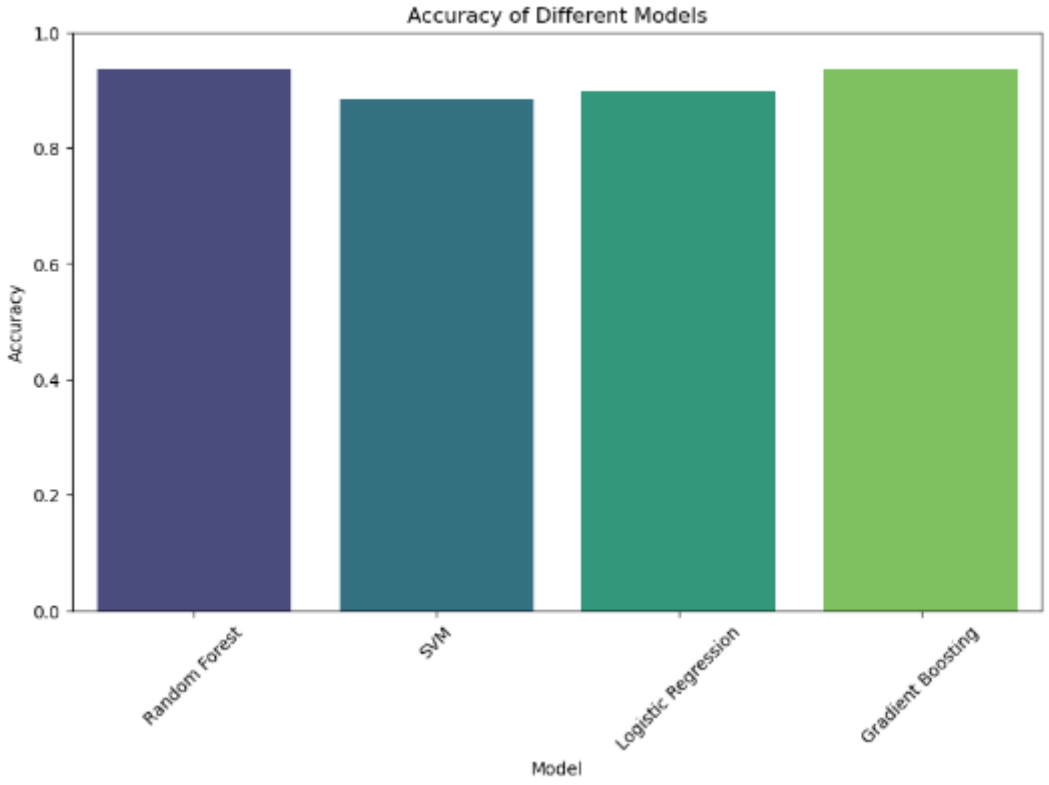
III. METHODOLOGY

A. DATA ANALYSIS & PREPROCESSING:

Feature Exploration : Understanding the qualities found in the wine data is the process of feature exploration. These characteristics may include things like residual sugar, acidity levels, alcohol concentration, and fixed acidity.

Data cleaning: Inconsistencies or missing values may be present in the data. It may be essential to use methods like imputation or to remove rows that have missing entries.

Feature Scaling: Various scales may be used to quantify distinct features. By ensuring that all features are scaled similarly, scaling keeps features with higher values from taking center stage in the model.



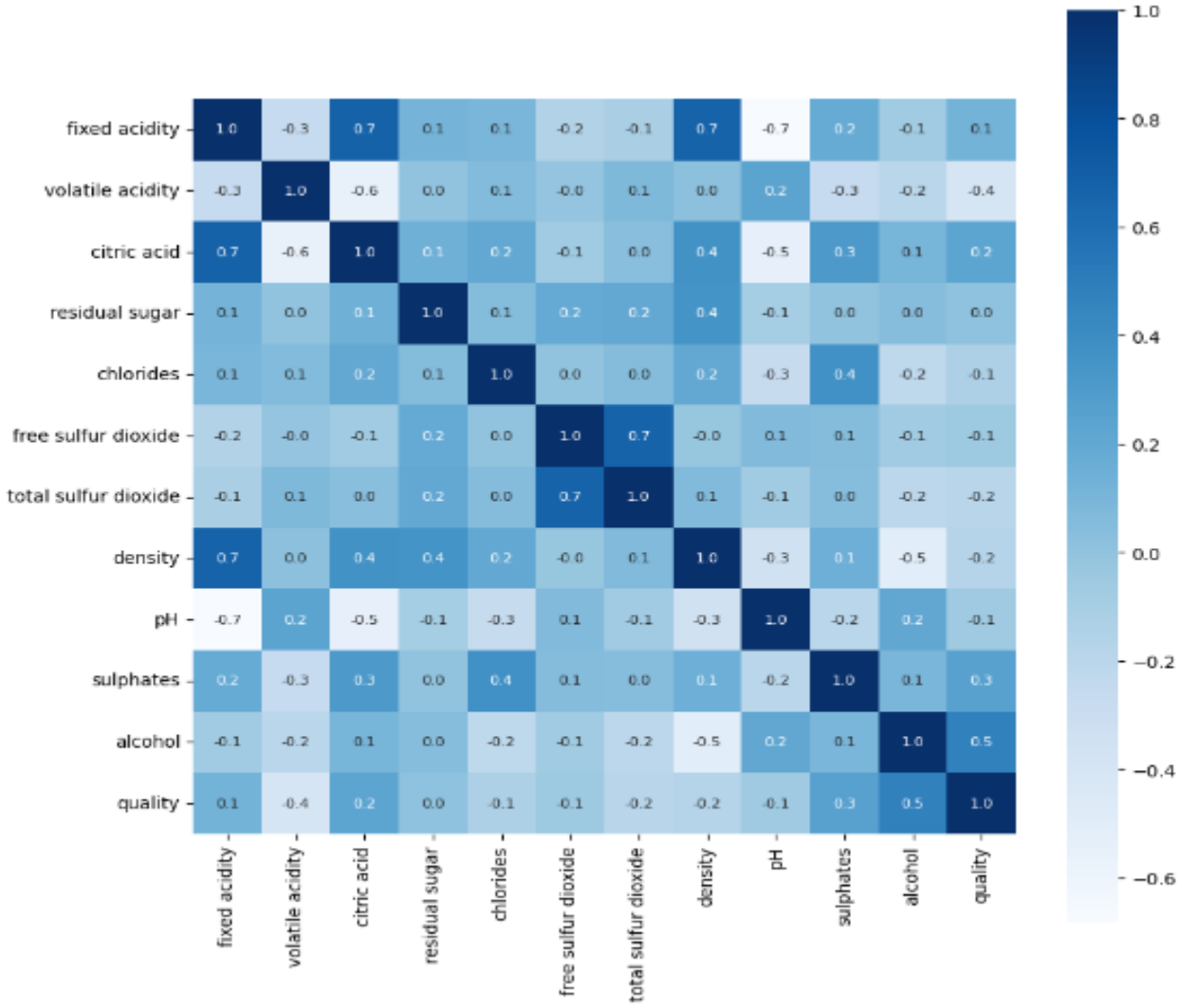


Fig 1: Confusion Matrix

The rows represent the actual labels (predicted quality), and the columns represent the model's predictions. Here's a breakdown of the information in the confusion matrix:

True Positive (TP): These are the data points where the model correctly predicted high-quality wine (quality = good).
False Positive (FP): These are the data points where the model incorrectly predicted high-quality wine (predicted good, but actual quality = bad).
True Negative(TN): These are the data points where the model correctly predicted low-quality wine (quality = bad).
False Negative(FN): These are the data points where the model incorrectly predicted low-quality wine (predicted bad, but actual quality = good).

B. RANDOM FOREST MODEL:

Ensemble Learning: Random Forest is an ensemble learning technique that aggregates predictions from several models (decision trees) to provide a final prediction that is more reliable and accurate.

Building Decision Trees: A random subset of characteristics and data points are used to build each tree in the forest using a process known as bootstrapping. By doing this, overfitting to the training set is avoided.

Making Predictions: Each choice tree in the forest is traversed by a fresh wine sample as it is delivered. Based on the rules it has learnt, each tree predicts the quality of the wine. The final prediction for the new sample is determined by taking the majority vote, which is the most frequently made forecast among all the trees.

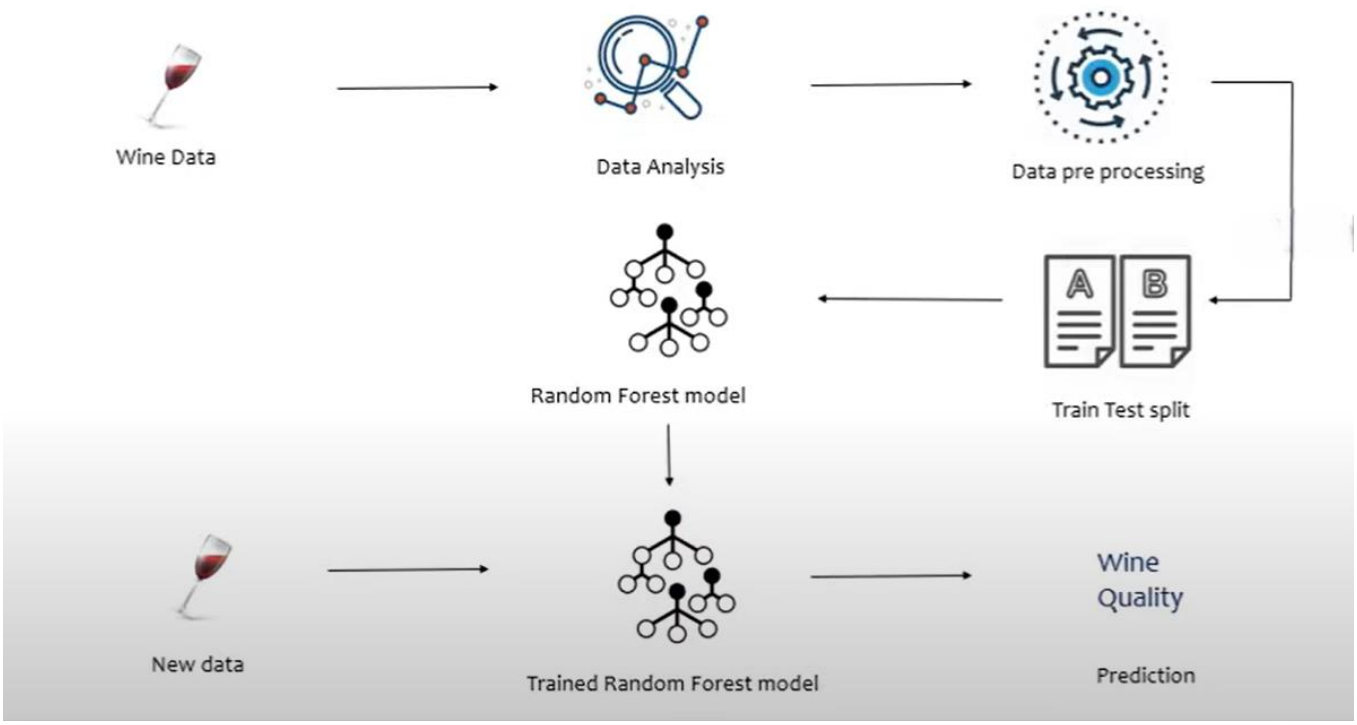


Fig 2: Proposed methodology

C. TRAIN TEST SPLIT:

Training Set: The Random Forest model is trained using this subset of the data. Drawing on the features present in the training data, the model derives patterns that allow it to distinguish between wines of varying quality.
Testing Set: After training, the model's performance is assessed using this set of unobserved data. The testing set predictions of the model show that it is generalizable to new data.

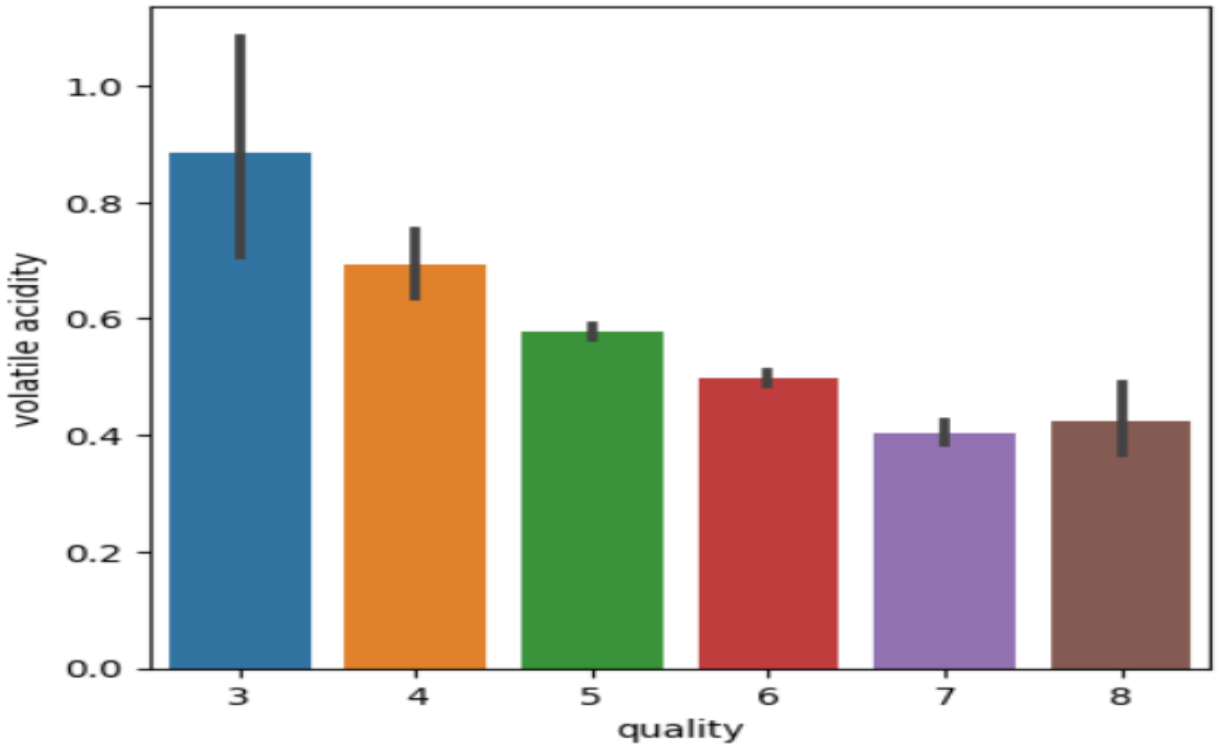


Fig 3: Volatile acidity vs Quality

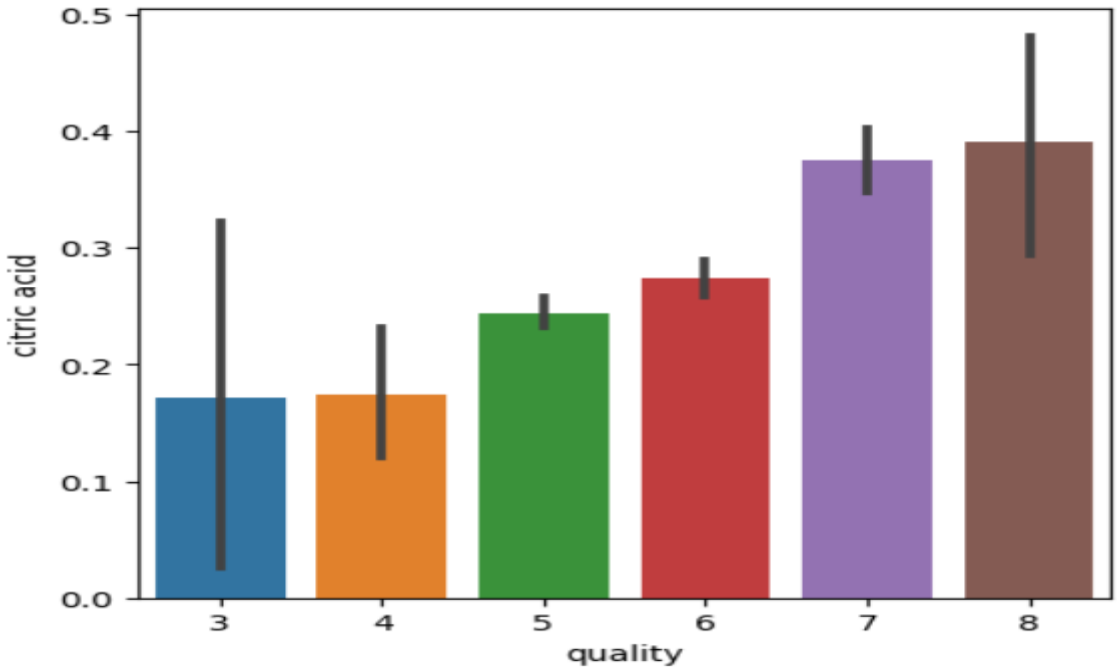


Fig 4: Citric acid vs Quality

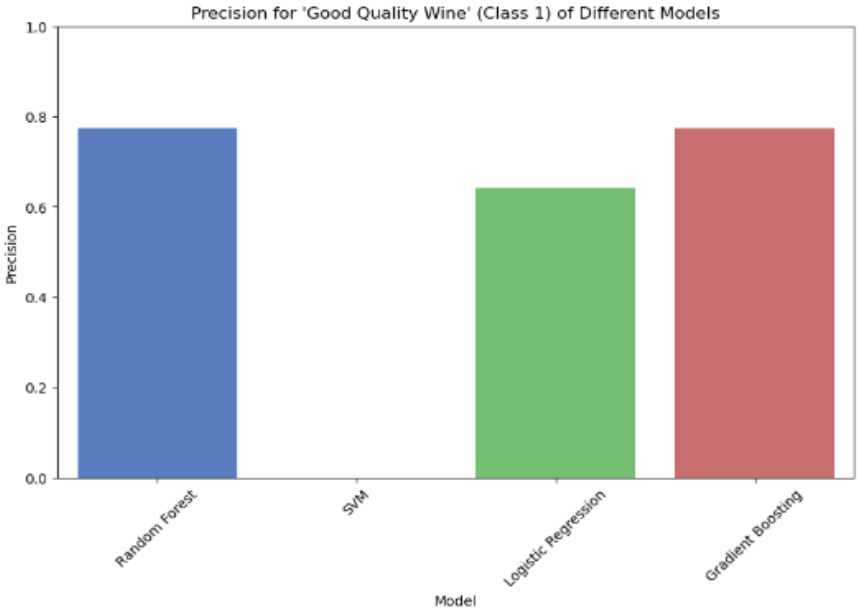
D. EVALUATION METRICS:

Accuracy: Out of all occurrences classified, the accuracy measure shows the number of cases that are accurately classified. Accuracy is the degree to which a calculated value resembles a real or standard value.

Precision: The percentage of all true positive predictions (TP) that are accurate compared to all false positive predictions and accurate positive predictions yields the precision. Thus, it follows that the precision guarantees that the objects are labelled as such when a model expects a good outcome.

Recall: Recall determines the proportion of true positives to the total of false negatives and true positives. When the expense of a false negative is substantial, this will be helpful. Recall, sometimes referred to as sensitivity, quantifies the percentage of real spam emails that the models accurately identify.

F1 Measure: The algorithm's total accuracy is determined by combining recall and precision to create the F1 score. Reliability of the model is demonstrated by low false positive and false negative values.



IV. RESULTS AND EVALUATION

To measure how effective our predicted models were, we computed and presented the following metrics: Accuracy, Precision, Recall and F1 Score. These metrics help to determine how well each model generalizes and how accurately it can predict the wine quality using its specifications. Model Performance Metrics:

Model	Accuracy	Precision	Recall	F1 Score
Random Forest	0.927500	0.814815	0.594595	0.687500
SVM	0.884375	0.000000	0.000000	0.000000
Logistic Regression	0.896875	0.642857	0.243243	0.352941
Gradient Boosting	0.937500	0.774194	0.648649	0.705882

Accuracy: Gradient Boosting boasts an accuracy of 0.9375, outperforming Random Forest, SVM, and Logistic Regression.

Performance: This high accuracy indicates that Gradient Boosting effectively predicts the correct outcome in the majority of cases.

Additional Considerations: While accuracy is a crucial metric, it's important to remember that: Other factors like precision, recall, and F1 score also contribute to the overall model evaluation. The best model for a specific task depends on various aspects like data quality, hyperparameters, and task requirements.

A. Data Imbalancing Techniques

SMOTE (Over Sampling Technique): By creating synthetic samples for the minority class, SMOTE (Synthetic Minority Over-sampling Technique) is a potent technique for resolving class imbalance in classification tasks.

SMOTE effectively boosts the minority class representation in the dataset by interpolating between existing minority class samples. This method finds the k-nearest neighbours of a sample chosen from the minority class in the feature space.

Then, in order to produce a more evenly distributed class distribution, synthetic samples are made along the line segments that link the sample and its neighbors.

B. NearMiss (Under sampling Technique):

Conversely, NearMiss is an under-sampling method intended to decrease the amount of majority class samples while maintaining the crucial data required for classification. In order to guarantee that the samples that are kept are the most similar to the minority class, NearMiss selects samples from the majority class that are closest to samples from the minority class.

Several variants of NearMiss are available, including NearMiss-1, NearMiss-2, and NearMiss-3, each with unique standards for choosing samples from the majority class. By undersampling the majority class and retaining the discriminatory information required for precise classification, NearMiss is generally successful in rebalancing class distributions.

Fig 5: Wine Quality Prediction

Fig 6: Prediction Result

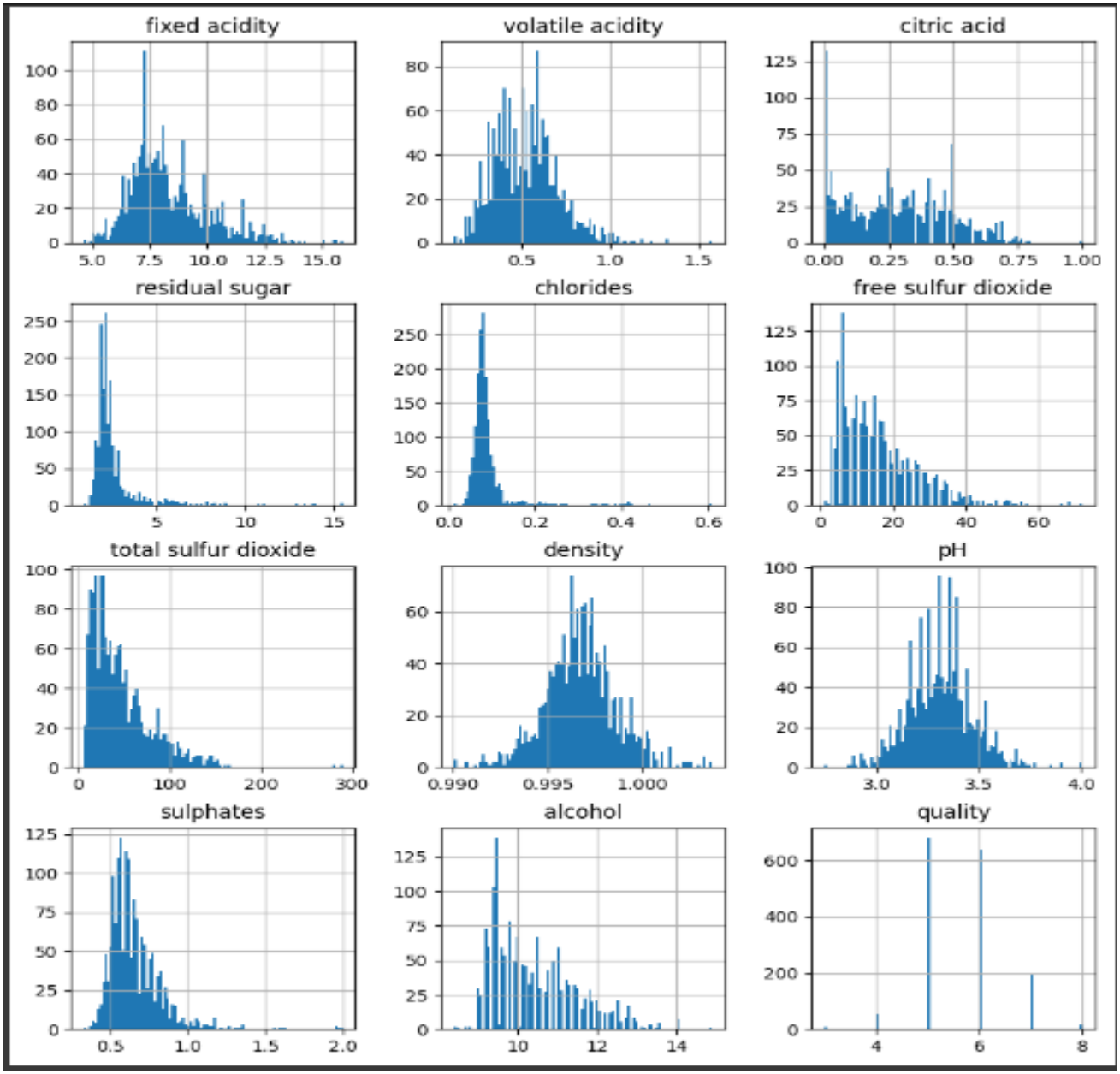


Fig 7: Distribution of various chemical compounds found in wine

V. CONCLUSION

We looked into how well a Random Forest classifier might predict a red wine's quality. We conducted exploratory data analysis using a dataset comprising several physicochemical properties of red wine samples in order to comprehend the properties of the dataset, such as its distribution, correlations, and statistical measures.
Next, we trained and assessed the classifier using machine learning techniques to determine which wines were of higher quality (rated 7 or higher) and which weren't.
We illustrated the efficacy of the Random Forest technique in precisely classifying wine quality through a series of experiments comprising training-test data splits, model fitting, and accuracy assessments.

Additionally, by predicting the quality of fresh wine samples based on their physicochemical characteristics, we demonstrated the usefulness of the trained model in practice.

Our results demonstrate the potential of machine learning algorithms—more especially, Random Forest classifiers—as useful instruments for quality evaluation and decision-making in the viticulture sector.

REFERENCES

[1] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009) Modeling Wine Preferences by Data Mining from Physicochemical Properties. Decision Support Systems, Elsevier, 47, 547-553. https://doi.org/10.1016/j.dss.2009.05.016

[2] Larkin, T. and McManus, D. (2020) An Analytical Toast to Wine: Using Stacked Generalization to Predict Wine Preference. Statistical Analysis and Data Mining: The ASA Data Science Journal, 13, 451-464. https://doi.org/10.1002/sam.11474

[3] Lin, E.B., Abayomi, O., Dahal, K., Davis, P. and Mdziniso, N.C. (2016) Artifact Re-moval for Physiological Signals via Wavelets. Eighth International Conference on Digital Image Processing, 10033, Article No. 1003355. https://doi.org/10.1117/12.2244906

[4] Dahal, K.R. and Mohamed, A. (2020) Exact Distribution of Difference of Two Sam-ple Proportions and Its Inferences. Open Journal of Statistics, 10, 363-374. https://doi.org/10.4236/ojs.2020.103024

[5] Dahal, K.R., Dahal, J.N., Goward, K.R. and Abayami, O. (2020) Analysis of the Res-olution of Crime Using Predictive Modeling. Open Journal of Statistics, 10, 600- 610, https://doi.org/10.4236/ojs.2020.103036

[6] Crookston, N.L. and Finley, A.O. (2008) yaImpute: An R Package for kNN Imputa-tion. Journal of Statistical Software, 23, 1-16. https://doi.org/10.18637/jss.v023.i10 [10] Dahal, K.R. and Gautam, Y. (2020) Argumentative Comparative Analysis of Ma-chine Learning on Coronary Artery Disease. Open Journal of Statistics, 10, 694-705.

[7] Dahal, K.R. and Gautam, Y. (2020) Argumentative Comparative Analysis of Machine Learning on Coronary Artery Disease. Open Journal of Statistics, 10, 694-705.

[8] Caruana, R. and Niculescu-Mizil, A. (2006) An Empirical Comparison of Super-vised Learning Algorithms. Proceedings of the 23rd International Conference on Machine Learning, June 2006, 161-168. https://doi.org/10.1145/1143844.1143865

[9] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An Introduction to Sta-tistical Learning: With Applications in R. Springer, Berlin, Germany.

[10] Joshi, R.P., Eickholt, J., Li, L., Fornari, M., Barone, V. and Peralta, J.E. (2019) Ma-chine Learning the Voltage of Electrode Materials in Metal-Ion Batteries.

[11] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Ma-chine. Annals of Statistics, 29, 1189-1232. https://doi.org/10.1214/aos/1013203451

[12] Chen, C.M. Liang, C.C. and Chu, C.P. (2020) Long-Term Travel Time Prediction Using Gradient Boosting. Journal of Intelligent Transportation Systems, 24, 109- 124. https://doi.org/10.1080/15472450.2018.1542304International

[13] Turian, J.P., Bergstra, J. and Bengio, Y. (2009) Quadratic Features and Deep Archi-tectures for Chunking. Human Language Technologies Conference of the North American Chapter of the Association of Computational Linguistics, Boulder, Colo-rado, 31 May-5 June 2009, 245- 248.

[14] Nwankpa, C., Ijomah, W., Gachagan, A. and Marshall, S. (2018) Activation Func-tions: Comparison of trends in Practice and Research for Deep Learning. arXiv: 1811.03378