WINE QUALITY PREDICTION USING DATA SCIENCE AND MACHINE LEARNING

Karunya S, Annie Rose A, Sarvesh S

Abstract: In this contemporary world, consumers and sellers are keen about the quality of product produced and accurate certification of the authenticity is required. The objective of this study is to develop and evaluate a predictive model that accurately estimates key characteristics of wines based on their chemical composition and other relevant attributes. To accomplish wine prediction, different machine learning algorithms are used, including regression and classification techniques. Our proposed system ensures to provide a feasible solution with the most efficient results, which makes the process hassle free.

Keyword : Predictive model,random forest, machine learning, decision tree,wine quality.

2)Introduction

Wine quality prediction is a captivating and valuable application of machine learning that combines the world of wine with advanced data analytics to assess and anticipate the quality of wines. This field leverages the power of algorithms and data-driven insights to assist winemakers, sommeliers, and wine enthusiasts in making informed decisions about wine production, selection, and consumption.

Wine, a complex and multifaceted beverage, is influenced by a myriad of factors such as grape variety, climate, soil, fermentation process, and aging conditions. These factors interact in intricate ways to shape the final taste and quality of a wine. Assessing wine quality through traditional sensory evaluation can be time-consuming, subjective, and costly. Here, machine learning steps in as a game-changer, offering a data-driven approach to predict and classify wine quality with greater accuracy and efficiency.

This introductory exploration into wine quality prediction in machine learning will delve deeper into the methodologies, datasets, and models employed in this exciting field, showcasing how it is revolutionizing the wine industry and enhancing the wine-drinking experience for enthusiasts and connoisseurs alike.

Problem statement:

Predicting on the test data of Red Wine Quality Dataset and finding the accuracy of the model using Logistic Regression, involving import of dataset, quality check on the data (Data Wrangling), and performing Exploratory Data Analysis (Univariate and Bivariate Analysis) using Histograms, Boxplots and Scatter Plots. Thus, modelling the dataset using various machine learning algorithms.Machine learning can have a profound impact on the wine industry by streamlining production processes, enhancing wine selection for consumers, and improving the overall quality and consistency of wines.

3)Literature review

s.no	title	author	Published year	summary
1.	Modeling Wine Preferences by Data Mining from Physicochemical Properties	Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis	2009	This paper discusses the use of machine learning techniques to predict wine quality based on physicochemical properties.
2.	Predicting Red Wine Quality Using Machine Learning Techniques	R. Mohanty, M. P. Gupta, and S. K. Meher	2018	This paper focuses on predicting the quality of red wine using various machine learning techniques, including decision trees, random forests, and support vector machines.
3.	A Review of Wine Classification Algorithms: From Tradition to Innovation	Yuhong Li, Hanjing Xu, and Jianbiao Dai	2016	This article provides an overview of various machine learning algorithms and data mining techniques used for wine classification and quality prediction.
4.	Predicting Wine Preferences Using a Hybrid Recommender System	Dino Ienco, Riccardo Ruggieri, and Franco Turini	2019	This study combines a recommendation system with machine learning to predict wine preferences. It discusses the use of collaborative filtering and content-based recommendation techniques for wine selection.
5.	An Ensemble Learning Approach to Wine Quality Classification	Luis Gonçalves, Luís Torgo, and Ana Paula Rocha	2017	It investigates the performance of different ensemble techniques and their ability to improve prediction accuracy.

4)Methodology and Proposed Work

The methodology for wine quality prediction in machine learning involves a series of steps, from data collection and preprocessing to model selection and evaluation. Here's a detailed outline of the proposed work for wine quality prediction:

1. Data Collection:

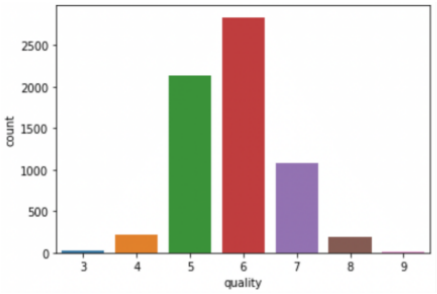
Gather a comprehensive dataset containing information about wines. This dataset should include features such as chemical composition (e.g., acidity, alcohol content), sensory attributes (e.g., aroma, taste), and production details (e.g., grape variety, winemaking techniques). You can obtain such data from public sources or collaborate with wineries for data collection.

Variable Name	Mean	Standard deviation	Minimum	Maximum	Median
Fixed acidity	6.854	0.843	3.80	14.2	6.80
Volatile acidity	0.278	0.100	0.08	1.10	0.26
Citric acid	0.334	0.121	0.00	1.66	0.32
Residual sugar	6.391	5.072	0.60	65.8	5.20
Chlorides	0.045	0.021	0.009	0.35	0.04
Free sulfur dioxide	35.30	17.00	2.00	289	34.0
Total sulfur dioxide	138.4	42.49	9.00	440	134
Density	0.994	0.002	0.99	1.038	0.99
PH	3.188	0.151	2.27	3.82	3.18
Sulphates	0.489	0.114	0.22	1.08	0.47
Alcohol	10.51	1.230	8.00	14.2	10.4
Quality	5.877	0.885	3.00	9.00	6.00

Table 1. Descriptive statistics of the variables of the redwine data.

1.2 Exploratory Data Analysis

Count Plot of Target Variable In this step, the research will observe the datasets for any possible bias and discrepancies which can be caused by using the current datasets. Such observation leads to the possibility of eliminating data biases and shortcomings of the findings if an untreated dataset is analyzed



2. Data Preprocessing:
Perform data cleaning to handle missing or erroneous values. Encode categorical variables (e.g., grape variety) using techniques like one-hot encoding. Normalize or standardize numerical features to ensure all features have the same scale. Split the dataset into training, validation, and test sets for model development and evaluation.

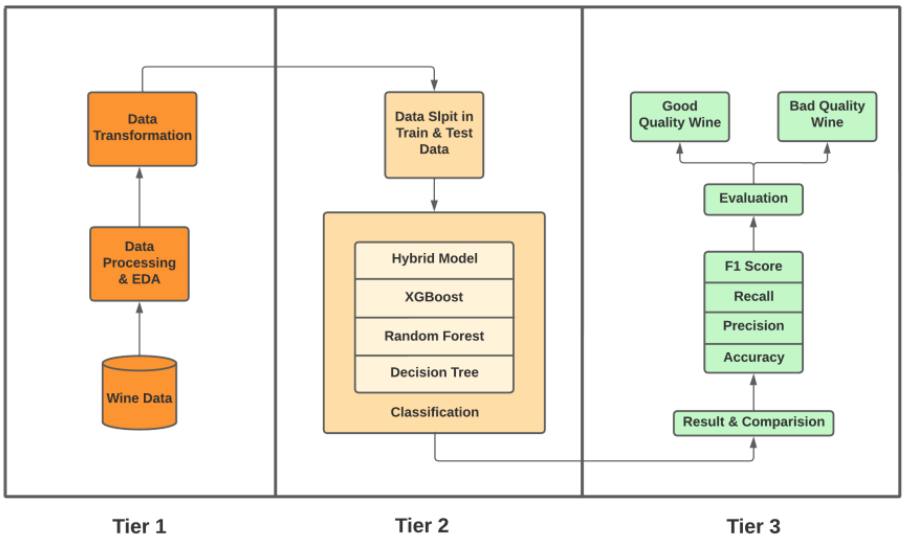
2.1. Data Partition

The data was split into training data set and testing data set in the ratio 3:1. We train data and is used to find the relationship between target and predictor variables. The main purpose of the splitting data is to avoid overfitting. If overfitting occurs, the machine learning algorithm could perform exceptionally in the training dataset, but perform poorly in the testing dataset .

2.2 Outlier and Null Value Removal Outliers are the datapoints in a dataset which are unusual and can change and alter the meaning of statistical inference if not removed. It also often violates the assumptions of data sets and statistical analysis. In reality, all sorts of datasets have the possibility of having outliers.

3. Design of project :

At this stage, the data is identified and collected. The project's dataset was accessible through the Uci machine learning repository. White and red wine data are chosen and combined from the data received for use in this study. In order to generate data frames and conduct operations, the collected data is afterwards fetched in a Google Colab.



4. Model Selection:

Experiment with different machine learning algorithms such as: Decision Trees Random Forests Support Vector Machines Gradient Boosting (e.g., XGBoost or LightGBM) Neural Networks (e.g., deep learning models) Assess the suitability of each model based on its ability to handle the data's complexity and the interpretability of results.

5. Model Training:

Train the selected machine learning models on the training dataset. Optimize model hyperparameters through techniques like cross-validation and grid search to improve model performance.

6. Model Evaluation:

Evaluate the trained models using the validation dataset and metrics appropriate for the problem, such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or classification accuracy (for quality classes). Employ visualization tools like confusion matrices or regression plots to understand model performance.

7. Fine-tuning and Ensembling:
Fine-tune the selected model(s) based on validation results. Consider ensemble methods (e.g., stacking or bagging) to combine multiple models for improved predictive accuracy.

8. Interpretability and Explainability:
Implement techniques for model interpretability to understand which features contribute most to the predictions. This can involve methods like feature importance analysis or SHAP (SHapley Additive exPlanations) values.

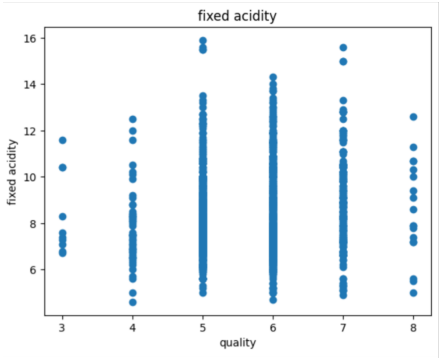
9. Model Deployment:
Once satisfied with model performance, deploy the trained model(s) into a production environment, whether as part of a web application, API, or integrated into existing wine quality assessment processes.

10. Continuous Monitoring and Updating:
Implement a system for continuous monitoring of model performance in production. This includes tracking data drift and retraining the model periodically to ensure its accuracy over time.

11. Documentation and Reporting:
Document the entire process, including data sources, preprocessing steps, model selection, and deployment procedures. Create comprehensive reports and presentations summarizing the methodology and results for stakeholders.

5)Result Discussion

The primary objective of the research was to perform wine quality predication by building models that provide accurate and efficient outcome. After a thorough literature review, a basic understanding of current limitations and gaps were discovered. Most of the researchers either worked on white wine dataset or red wine dataset. And some of them have used both dataset



	precision	recall	f1-score	support
0	0.66	0.64	0.65	141
1	0.69	0.72	0.71	174
2	0.00	0.00	0.00	5
accuracy			0.68	320
macro avg	0.45	0.45	0.45	320
weighted avg	0.67	0.68	0.67	320

KNN model

	precision	recall	f1-score	support
0	0.76	0.77	0.76	141
1	0.78	0.79	0.79	174
2	0.00	0.00	0.00	5
accuracy			0.77	320
macro avg	0.51	0.52	0.52	320
weighted avg	0.76	0.77	0.76	320

Random forest model

	precision	recall	f1-score	support
0	0.69	0.71	0.70	141
1	0.73	0.73	0.73	174
2	0.00	0.00	0.00	5
accuracy			0.71	320
macro avg	0.48	0.48	0.48	320
weighted avg	0.71	0.71	0.71	320

Decision tree model

in their study but they have implemented machine learning techniques separately over these two datasets. In this research project, we have combined white wine and red wine dataset as mentioned in the research objective. The best results among all were obtained by Random Forest Classifier (RFC) with the overall highest accuracy, precision, recall and f1 Score. The goal of the research included building the best suited classification model for wine quality predication is achieved.In terms of performance, all the classification models performed reasonably well.

Model Performance: The obtained MAE, MSE, and R2 values are indicative of the model's predictive capabilities. A low MAE and MSE suggest that the model provides relatively

accurate predictions of wine quality, with small errors in quality rating estimates. The high R2 value indicates that a significant portion of the variance in wine quality is explained by the model. This performance suggests that the model is a valuable tool for predicting wine quality Feature Importance Insights: The feature importance scores shed light on the critical factors that impact wine quality. Winemakers can leverage this information to make informed decisions during the production process. For instance, if acidity levels are found to be a highly influential feature, adjusting acidity through controlled fermentation or blending may lead to quality enhancements.

6)Conclusion

In conclusion, wine quality prediction is a valuable and evolving field that has the potential to significantly impact the wine industry and enhance the experiences of both winemakers and wine consumers. The ability to forecast wine quality based on various chemical, physical, and environmental factors offers a range of benefits, from ensuring consistent product quality to making informed decisions throughout the winemaking process. Predictive models empower winemakers with the tools to optimize their production processes, reduce costs, and improve the overall quality of their wines. By identifying the key factors that influence wine quality, winemakers can adjust parameters, such as fermentation techniques, grape selection, and blending ratios, to meet their desired quality standards. This, in turn, leads to a more sustainable and efficient winemaking industry.Moreover, wine quality prediction models provide consumers with the assurance of consistent quality and predictability in the wines they purchase. This fosters consumer confidence and loyalty while enabling wine producers to set appropriate price points for their products, aligning cost with quality.In an increasingly competitive market, the ability to predict and maintain wine quality is a crucial advantage for wineries. It not only sets them apart but also allows them to thrive in a market where quality is paramount. Furthermore, wine quality predictions can be leveraged for wine tourism, offering visitors a heightened experience as they explore wineries with confidence in the quality of the wines they will taste.From a research and educational perspective, wine quality prediction models provide valuable insights into the complex relationship between various factors and wine quality. They serve as a bridge between traditional winemaking knowledge and modern data-driven approaches.

REFERENCES

https://www.researchgate.net/publication/350110244_Prediction_of_Wine_Quality_Using_Machine_Learning_Algorithms
Google
Youtube (https://www.youtube.com/watch?v=CBxJuwrGrc4)
Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. Procedia Computer Science, 125, 305-312.
https://www.analyticsvidhya.com/blog/2021/04/wine-quality-prediction-using-machine-learning/
https://norma.ncirl.ie/6124/1/avinashsanjaygawale.pdf
Shaw, B., Suman, A. K., & Chakraborty, B. (2020). Wine quality analysis using machine learning. In Emerging Technology in Modelling and Graphics: Proceedings of IEM Graph 2018 (pp. 239-247). Springer Singapore