

Red Wine Quality Prediction

Group Members: Sanyukta Joshi	16O1O12OO19
Manas Marathe	16O1O12OO27
Zenith Mehta	16O1O12OO28

Guide: Dr. Shruti Javkar

Motivation

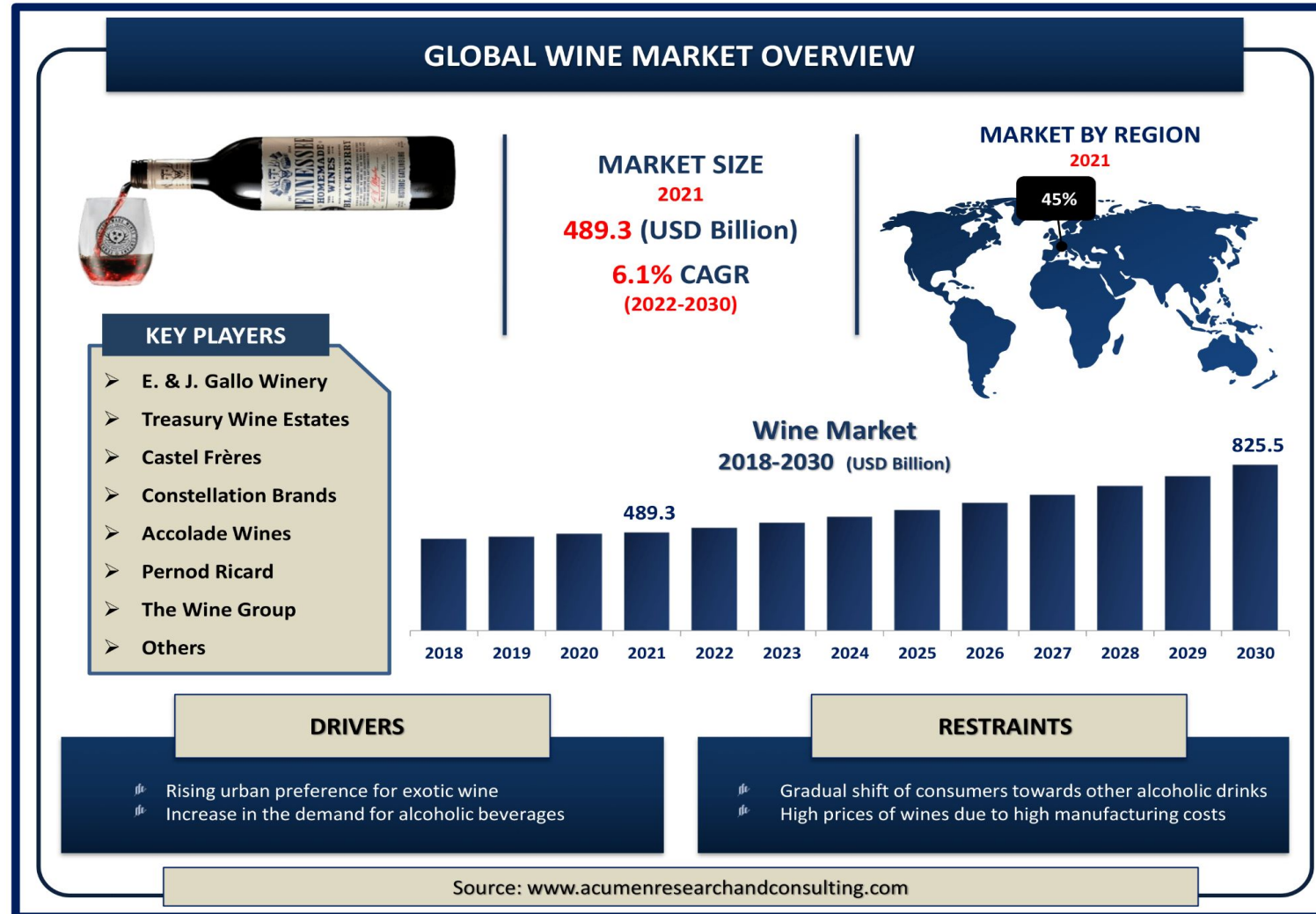
National Scale



Global Scale

So why not invest where risk is very low and returns are high due to its ever rising demands.

Yield of 6.1% CAGR



Problem Statement

To leverage comprehensive dataset encompassing various chemical parameters, sensory descriptors, and quality ratings. By employing machine learning techniques, we seek to develop predictive models capable of providing valuable insights to winemakers, guiding decisions related to production processes and quality assurance

Scope

- Comparative studies to evaluate the performance of machine learning algorithms like Random Forest, Support Vector Machine, and ensemble methods specifically for wine quality prediction, focusing on metrics such as accuracy, precision, and computational efficiency.
- Extend datasets by incorporating additional attributes and samples, considering factors such as grape variety, geographical location, and environmental conditions, to improve the robustness and generalization capabilities of predictive models.
- Used by industry partners to implement developed models into practical tools for winemakers, enabling real-time assessment and decision support in vineyard management and wine production processes.

Literature Survey

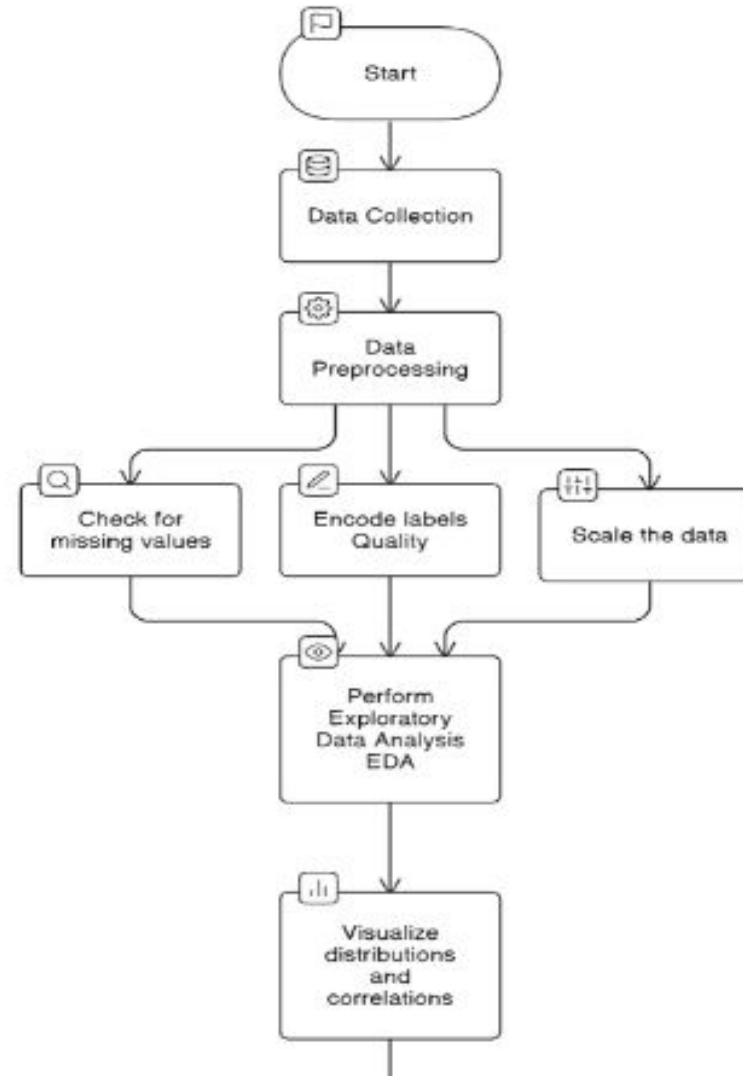
Sr.No	Title	Objective	Model/Algorithm	Dataset	Conclusion
1.	Quality of Red Wine: Analysis and Comparative Study of Machine Learning Models [1]	To develop and optimize a predictive model using machine learning algorithms, focusing on decision tree classification, for accurately classifying the quality of red wine based on chemical parameters.	Naive Bayes, Logistic Regression, Support Vector Machine and Random Forest Classifier	Consists of 12 attributes and was divided into two classes based on wine quality (>6 considered good).	Random Forest Classifier demonstrated superior performance with, outperforming other classifiers.
2.	Red Wine Quality Prediction Using Machine Learning Techniques [2]	Aims to predict the quality of red wine using various attributes, employing various techniques. By comparing results between training and testing sets, the most effective technique will be determined,	Random Forest, Support Vector Machine, and Naïve Bayes.	The dataset contains 1599 instances with 12 variables for red wine data. Qualities are in the range 3-8, where '3' predicts poor quality of red wine and '8' predicts excellent quality of red wine.	SVM showing higher accuracy compared to Random Forest and Naïve Bayes.

3.	Enhancing red wine quality prediction through Machine Learning approaches with Hyperparameters optimization technique [3]	Compare classification and regression methods for predicting red wine quality, evaluating the performance of various classifiers and regressors with hyperparameter tuning and imbalance data handling techniques	Logistic regression, Gradient boosting, Extra Tree Classifier, AdaBoost classifier, Decision Tree, Support Vector	Red wine samples with their chemical compositions and qualitative evaluations have been included in the UCI Machine-Learning database Wine Grade Collection.	Employing machine learning algorithms like Random Forest, AdaBoost, and Gradient Boosting, among others, demonstrated high accuracy in predicting red wine quality. Importance of EDA.
4.	Utilization of Random Forest Classifier (RFC) To Predict the Quality of Beverages [4]	Aims to predict beverage quality by analyzing its attributes using the Random Forest algorithm, with performance measures compared between training and testing sets to determine overall quality prediction accuracy.	Random Forest Classifier	Data obtained from the UCI ML repository. The dataset comprises 1599 observations and 8 variables pertaining to beverage data	Using Random forest algorithm, enabled precise prediction of wine quality, achieving an impressive accuracy rate.
5.	Smart Agriculture and Digital Transformation on Case of Intelligent System for Wine Quality Prediction	Propose an intelligent system for wine quality prediction using decision tree-based machine learning methodology, leveraging emerging technologies	Decision Tree Based ML Methodology	UCI Machine learning Repository (2009.) Wine Quality Data Set	Potential for digital transformation in smart agriculture by automating wine quality assessment. The effectiveness of classification and regression trees has improved a great deal to help in prediction

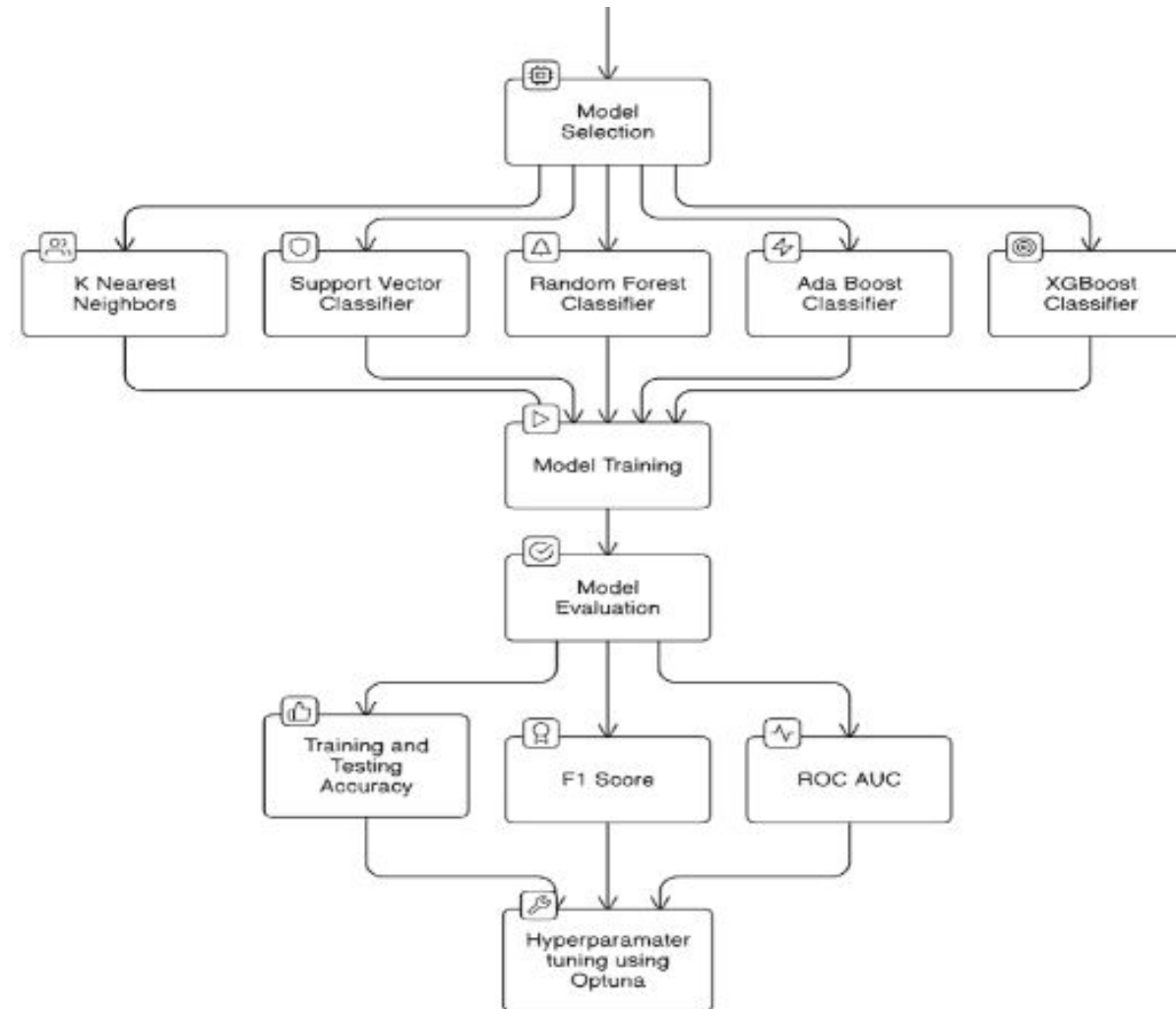
Objectives

- Investigate a variety of predictive models for assessing the quality of red wine, encompassing classification algorithms.
- Conduct thorough performance evaluations utilizing established metrics.
- Employ advanced optimization techniques like hyperparameter tuning with Optuna to refine model performance iteratively, aiming for optimal configurations that enhance predictive accuracy and generalization capabilities.

Implementation Flow (change name later)



Implementation Flow (change name later)



Implementation

Dataset










Title: Red Wine Quality

# fixed acidity	# volatile aci...	# citric acid	# residual su...	# chlorides	# free sulfur ...	# total sulfur...	# density	# pH	# sulphates	# alcohol
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
7.8	0.88	0.0	2.6	0.098	25.0	67.0	0.9968	3.2	0.68	9.8
7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.997	3.26	0.65	9.8
11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.998	3.16	0.58	9.8
7.4	0.7	0.0	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4
7.4	0.66	0.0	1.8	0.075	13.0	40.0	0.9978	3.51	0.56	9.4
7.9	0.6	0.06	1.6	0.069	15.0	59.0	0.9964	3.3	0.46	9.4

Dataset - Why this one specific ?

36 Results

Relevance ▾

	Wine Quality Dataset Dataset · 2y ago · by M Yasser H Wine Quality Prediction - Classification Prediction	606 45,567 downloads
	Red Wine Dataset Dataset · 7y ago · by piyushgoyal443 This is a subset of wine quality dataset which contains only red wine samples	93 12,799 downloads
	Red Wine Quality Dataset · 6y ago · by UCI Machine Learning Context The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine	2720 225,828 downloads
	Spanish Wine Quality Dataset Dataset · 2y ago · by fedesoriano Spanish Wine Quality Dataset . Retrieved Date Retrieved from	114 6,144 downloads
	Wine Quality Dataset · 6y ago · by Raj Parmar The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine .	251 37,932 downloads
	White Wine Quality Dataset · 3y ago · by Piyush Agnihotri repository, to get both the dataset i.e. red and white vinho verde wine samples, from the north of	51 10,182 downloads
	Wine Quality Dataset · 2y ago · by Brenda N Wine quality for red and white wine	40 1,083 downloads
	Wine Quality Dataset · 6y ago · by Daniel S. Panizzo Number of Instances: red wine - 1599; white wine - 4898. 6.	41 5,752 downloads
	Wine Quality Classification	48



241155 views

Dataset - Parameters Insights

Term	Need	Value Range
Fixed Acidity	Provides backbone of flavor, acidity balance	4 - 15 g/L
Volatile Acidity	Indicates spoilage	< 0.8 g/L
Citric Acid	Adds freshness, balances flavors	Up to 0.5 g/L
Residual Sugar	Determines sweetness level	0 - 220 g/L
Chlorides	Contributes to taste	< 0.1 g/L
Free Sulfur Dioxide	Preserves wine	10 - 40 mg/L
Total Sulfur Dioxide	Preserves wine	Up to 150 mg/L
Density	Indicates body and texture	0.98 - 1.05 g/mL
pH	Affects taste and stability	3.0 - 4.0
Sulphates	Acts as preservative	0.3 - 2.0 g/L
Alcohol	Influences body and richness	9% - 16% by volume

Quality	Overall evaluation of the wine	Typically rated on a scale of 3 to 8 or 0 to 10
---------	--------------------------------	---

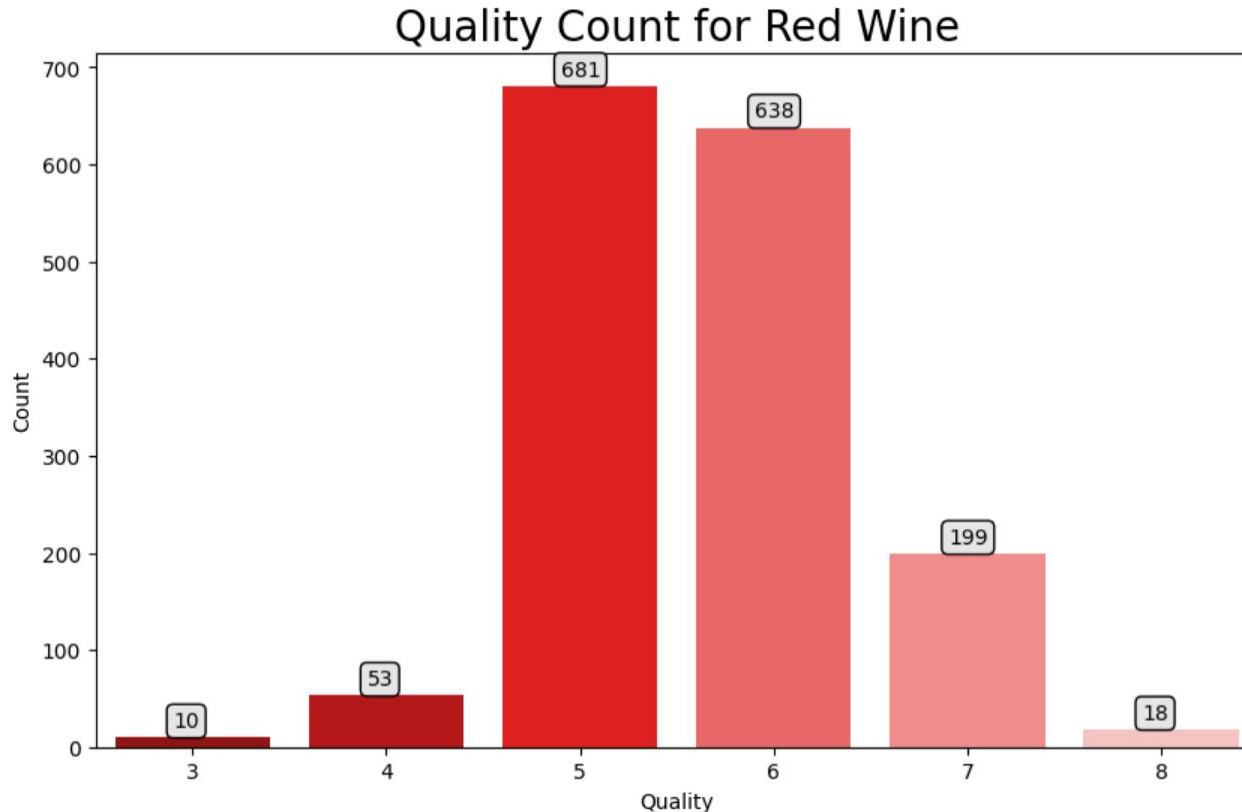
Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables.

The conducted Exploratory Data Analysis (EDA) on the red wine quality dataset encompasses three primary visualizations:

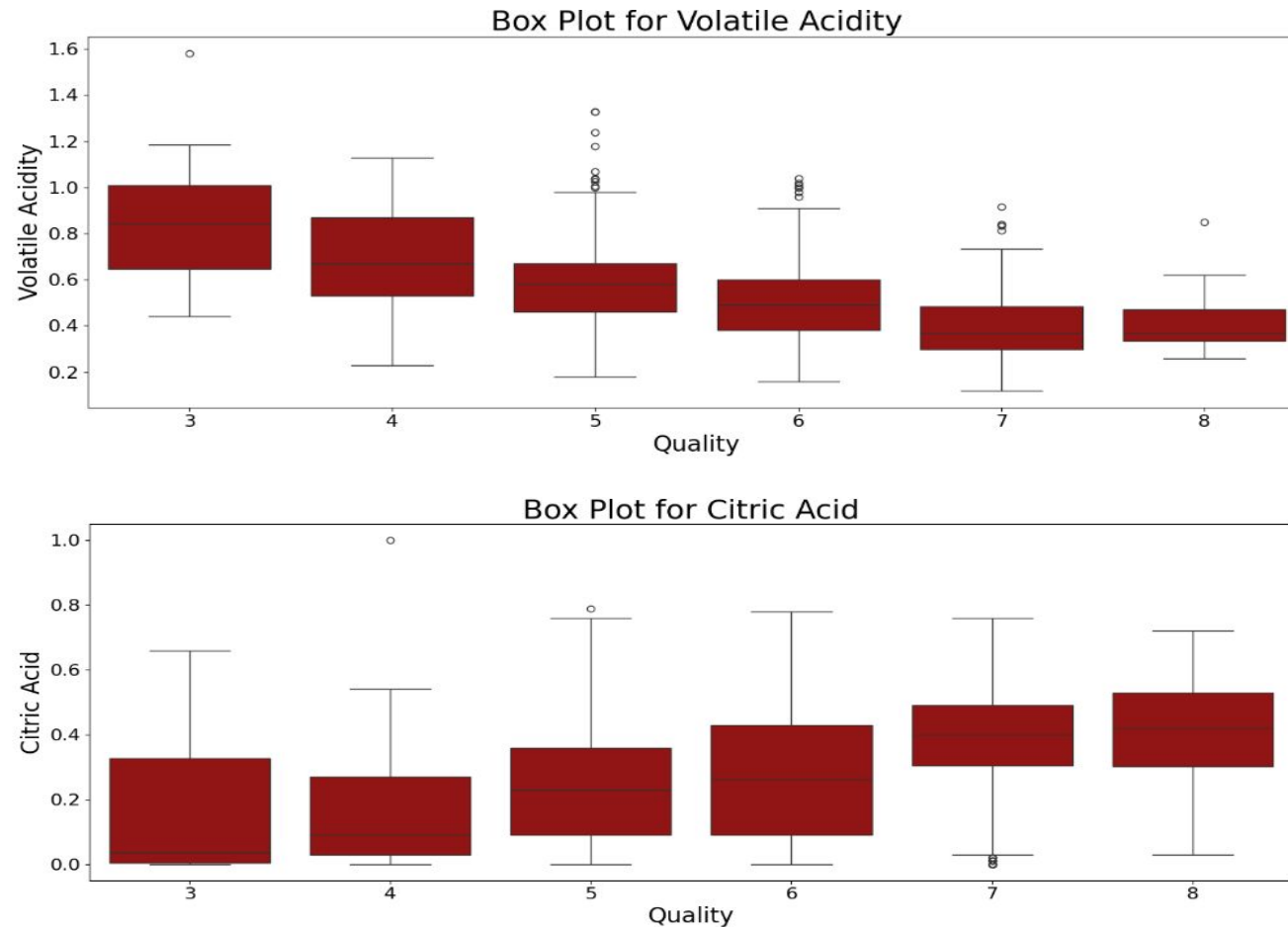
- **A countplot** provides an overview of wine quality distribution, indicating that wines with a quality rating of 5 are most prevalent, while those with a rating of 3 are notably scarce
- Series of **boxplots** depict the distributions of various wine features, such as fixed acidity, volatile acidity, and pH, offering insights into their central tendencies and variability.
- **Correlation matrix** elucidates the interrelationships between these features, revealing potential dependencies and guiding feature selection for subsequent modeling efforts.

EDA: Countplot



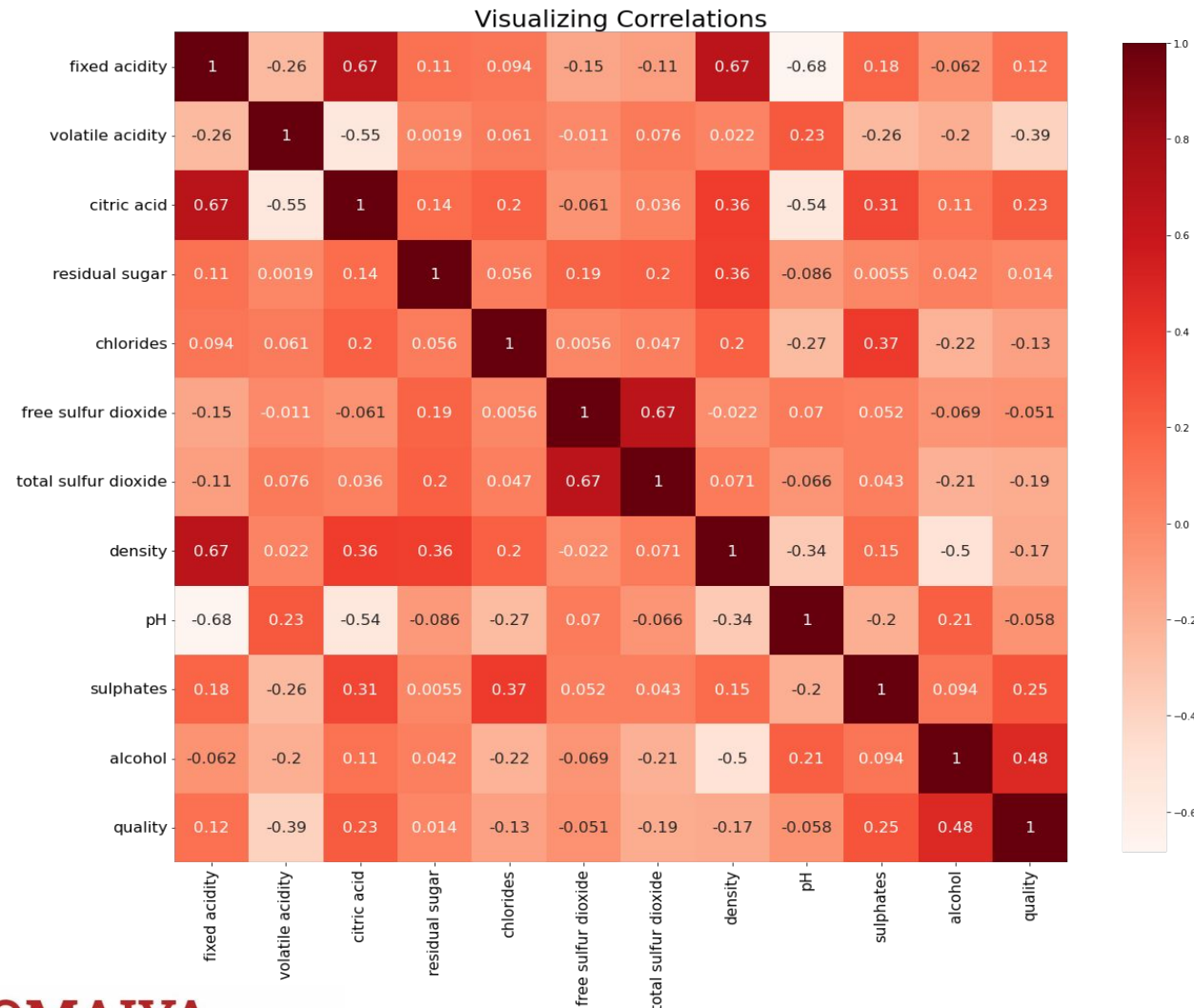
- It is observed that wines with a **quality rating of 5** exhibit the **highest frequency**, while those with a **quality rating of 3** are notably scarce, representing the **lowest count**.
- This discrepancy in frequency across quality ratings underscores the presence of a **class imbalance** within the dataset.

EDA: Box Plot



- Boxplots display the **distribution of data**. It gives a good indication of how the values in the data are spread out for each parameter of the dataset.
- Volatile Acid:** There is an inverse relationship between wine quality and volatile acidity, where higher quality wines exhibit lower median values of volatile acidity.
- Citric Acid:** Contrary to volatile acidity, an increase in wine quality corresponds to higher median values of citric acid, indicating a positive correlation between quality and citric acid content.

EDA: Correlation Matrix

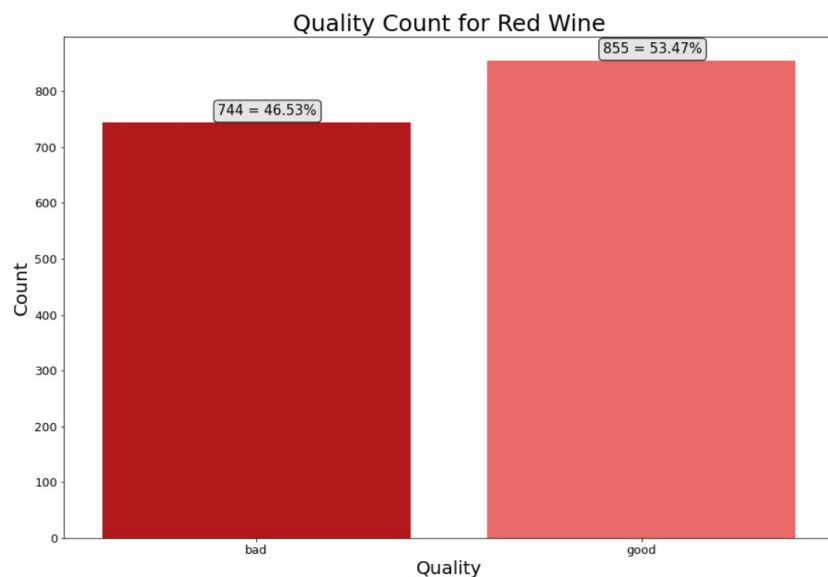


- Correlation analysis is a fundamental technique used to examine the relationships between variables within a dataset.
- Shades of red are employed to represent correlation values, with **lighter shades** indicating **higher values** and **darker shades** indicating **lower values**.

Data Preprocessing

Solving Class Imbalance

- Class imbalance occurs when one class in a classification problem has **significantly fewer samples than the others**, potentially leading to biased model performance.
- We redefined the wine quality ratings into two categories, '**bad**' and '**good**', based on a predefined



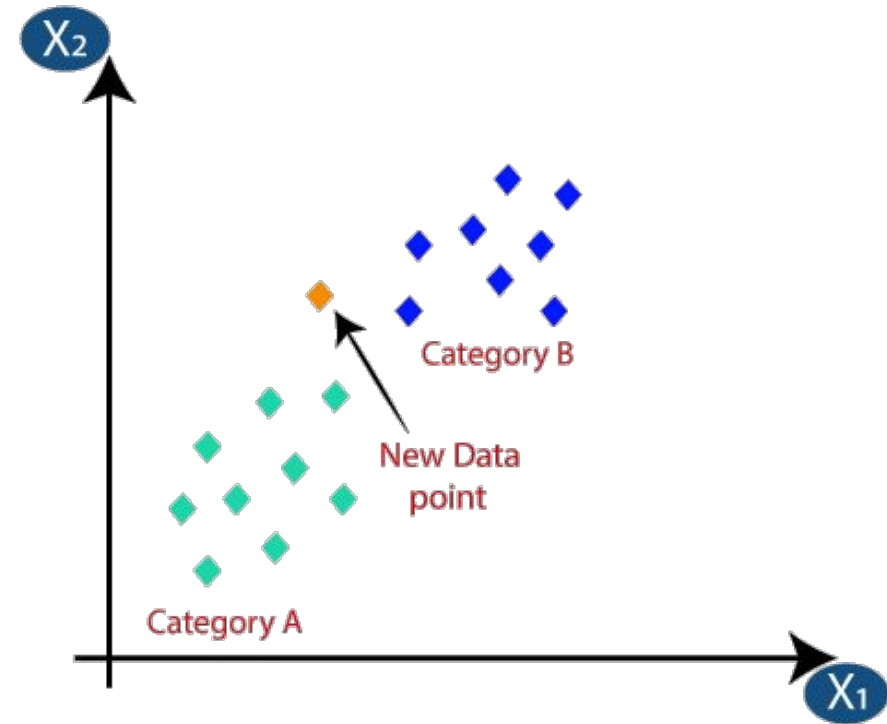
Scaling

- Scaling is a crucial preprocessing step in machine learning. It ensures that **all features contribute equally** to the model training process by standardizing their distributions.

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH
0	-0.528360	0.961877	-1.391472	-0.453218	-0.243707	-0.466193	-0.379133	0.558274	1.288643
1	-0.298547	1.967442	-1.391472	0.043416	0.223875	0.872638	0.624363	0.028261	-0.719933
2	-0.298547	1.297065	-1.186070	-0.169427	0.096353	-0.083669	0.229047	0.134264	-0.331177
3	1.654856	-1.384443	1.484154	-0.453218	-0.264960	0.107592	0.411500	0.664277	-0.979104
4	-0.528360	0.961877	-1.391472	-0.453218	-0.243707	-0.466193	-0.379133	0.558274	1.288643
...

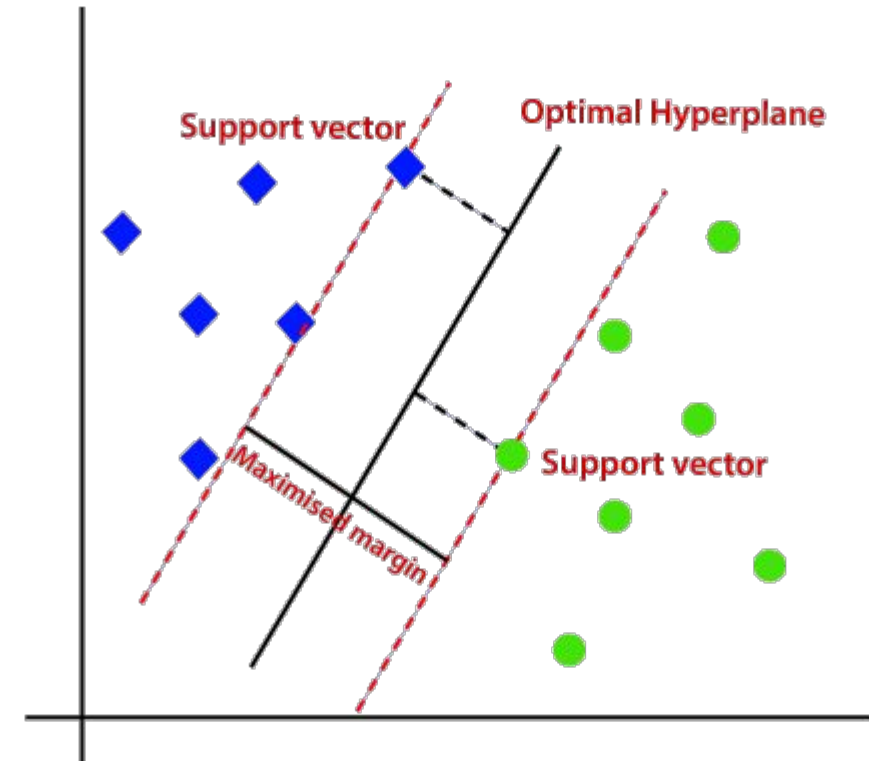
Training Models: K-Nearest

- K-Nearest Neighbors (KNN) is based on the principle that objects or instances with similar characteristics exist close to each other in the feature space.
- When a new data point is encountered, KNN identifies its closest neighbors from the training set based on a chosen distance metric.
- Once the nearest neighbors are identified, KNN classifies the new data point by a majority voting scheme, where the class label of the majority of its nearest neighbors determines its predicted class.



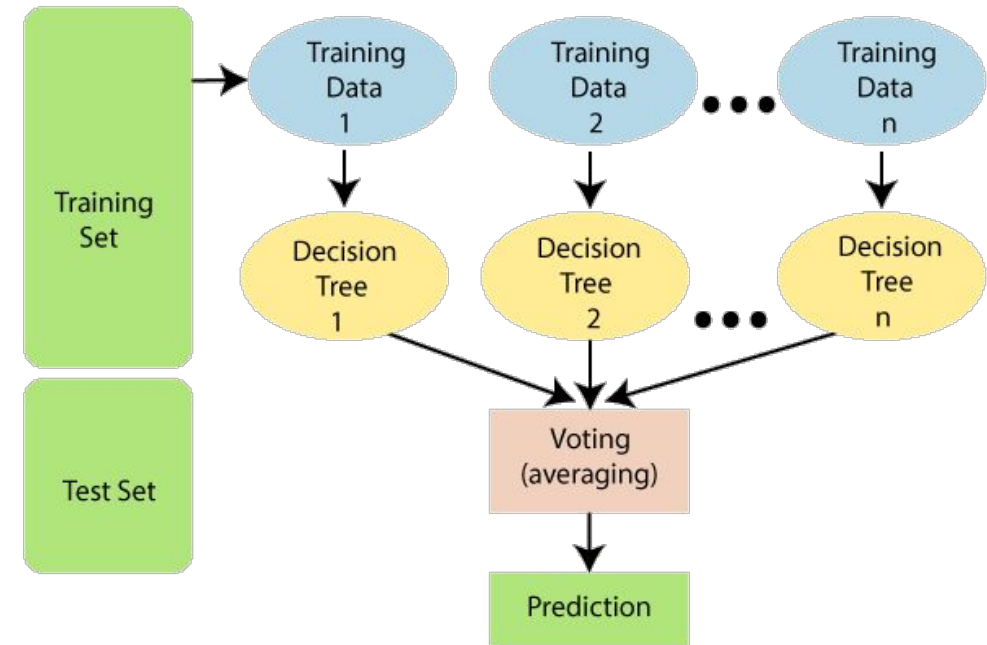
Training Models: Support Vector Classifier

- Support Vector Classification (SVC) finds the best line (or plane) to separate different groups in data. It identifies the most important data points, called support vectors, to create the boundary with the largest gap between classes.
- VC aims to create a wide margin between classes, making it less sensitive to noisy data and outliers. By focusing on the most relevant data points, it tends to generalize well to new, unseen data. However, proper parameter tuning is crucial for optimal performance.



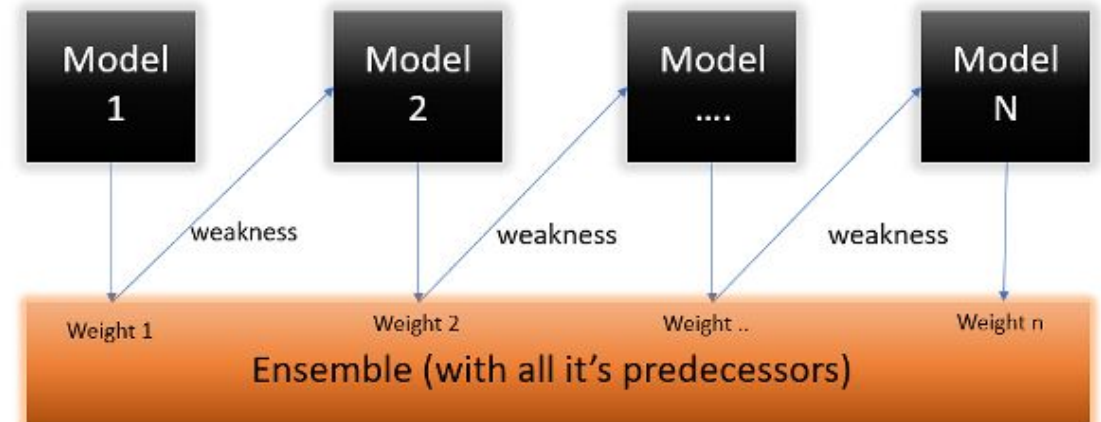
Training Models: Random Forest Classifier

- The Random Forest Classifier employs the ensemble learning technique by combining multiple decision trees to improve prediction accuracy and robustness. Each decision tree is trained independently on a random subset of the data and features.
- Random Forest is highly versatile and can handle various types of data, including numerical and categorical features. It automatically handles missing values and can be applied to both classification and regression tasks.
- Random Forest tends to be less prone to overfitting compared to individual decision trees, making it suitable for a wide range of real-world applications.
-



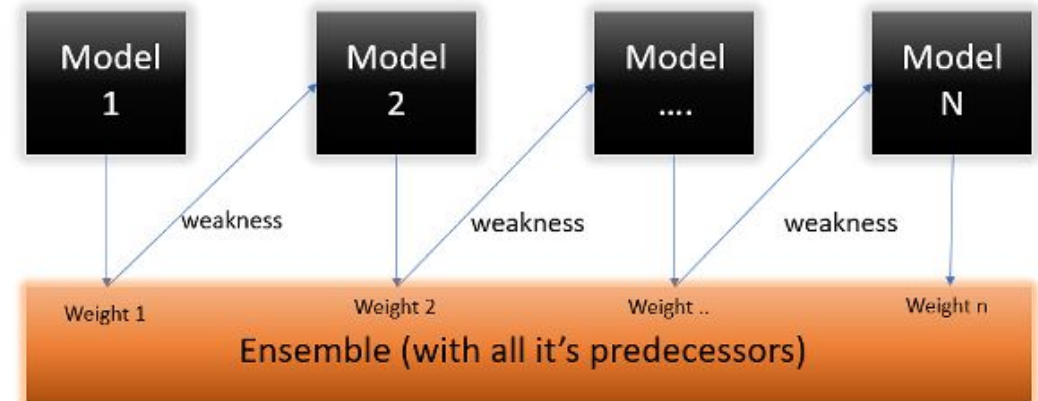
Training Models: AdaBoost

- ADABOOST (Adaptive Boosting) is an ensemble learning method that combines multiple weak learners, typically decision trees, to create a strong classifier. It sequentially trains a series of weak learners, adjusting their weights based on the performance of previous models.
- it focuses on improving the performance of misclassified data points from the previous weak learners. It assigns higher weights to misclassified instances, allowing subsequent weak learners to focus more on these challenging cases, gradually improving overall accuracy.
- ADABOOST is versatile and can be applied to various types of classification problems.



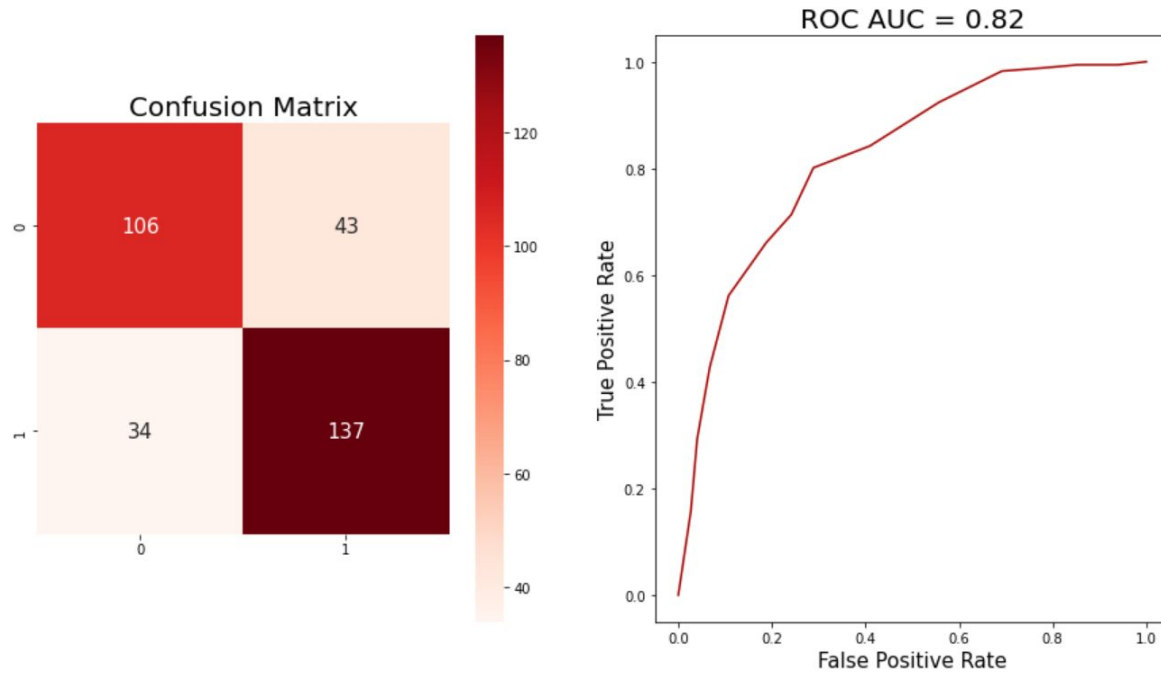
Training Models: XGBoost

- It builds an ensemble of weak learners, typically decision trees, in a sequential manner, where each subsequent model learns from the mistakes of its predecessors.
- XGBoost is designed for speed and efficiency, making it well-suited for large datasets.
- XGBoost often achieves state-of-the-art results in various machine learning competitions and real-world applications. It excels in handling complex datasets with high-dimensional features and provides built-in support for handling missing values and categorical variables.

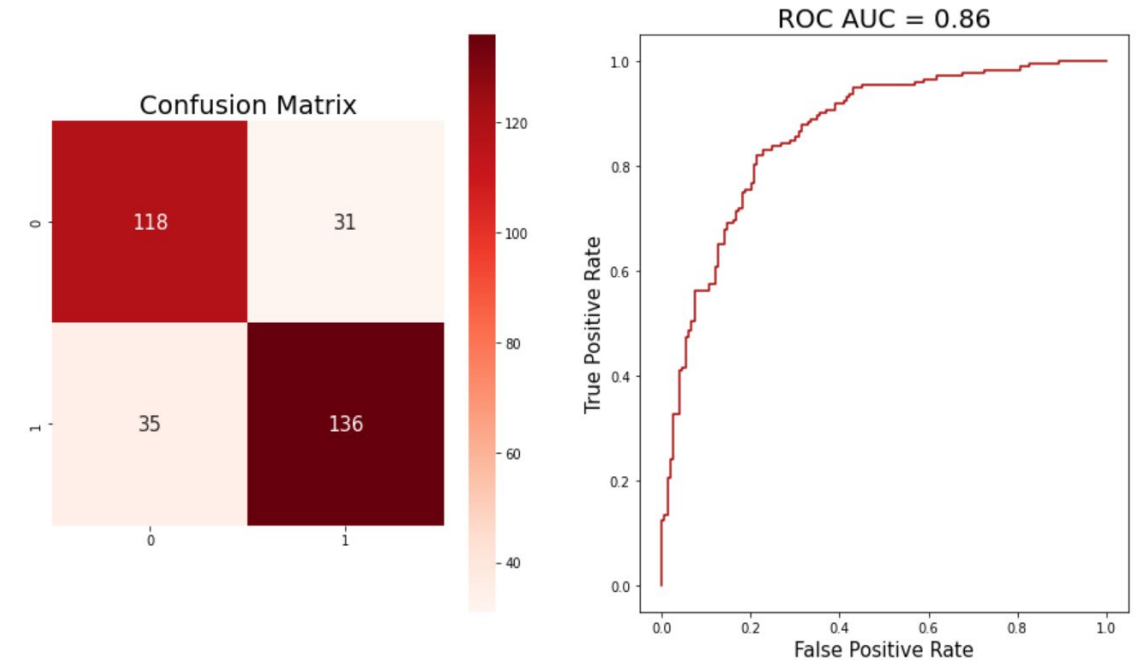


Results

Graphs for K Nearest Neighbors

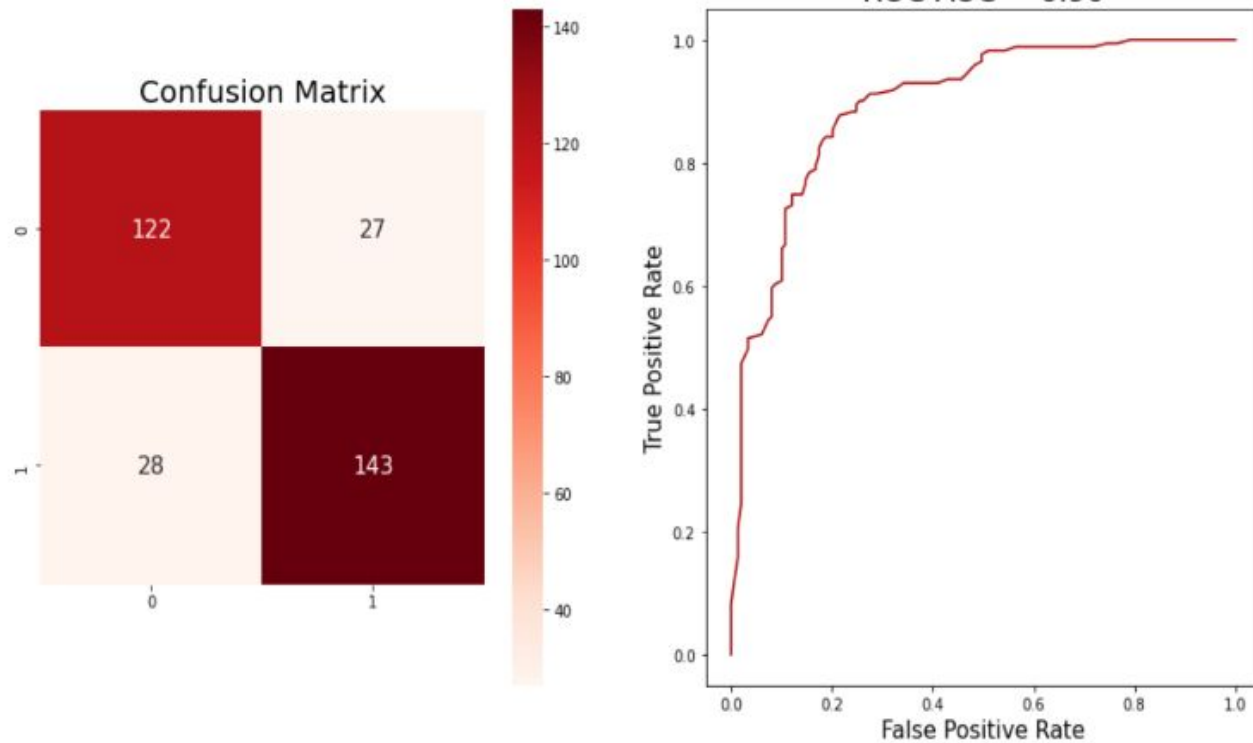


Graphs for Support Vector Classifier

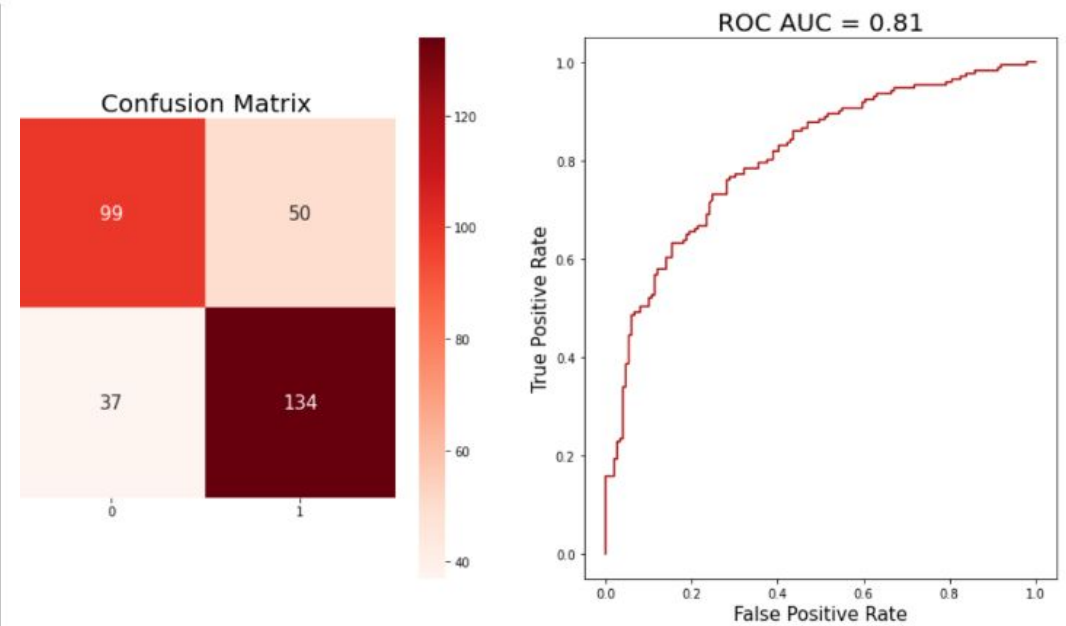


Results

Random Forest:

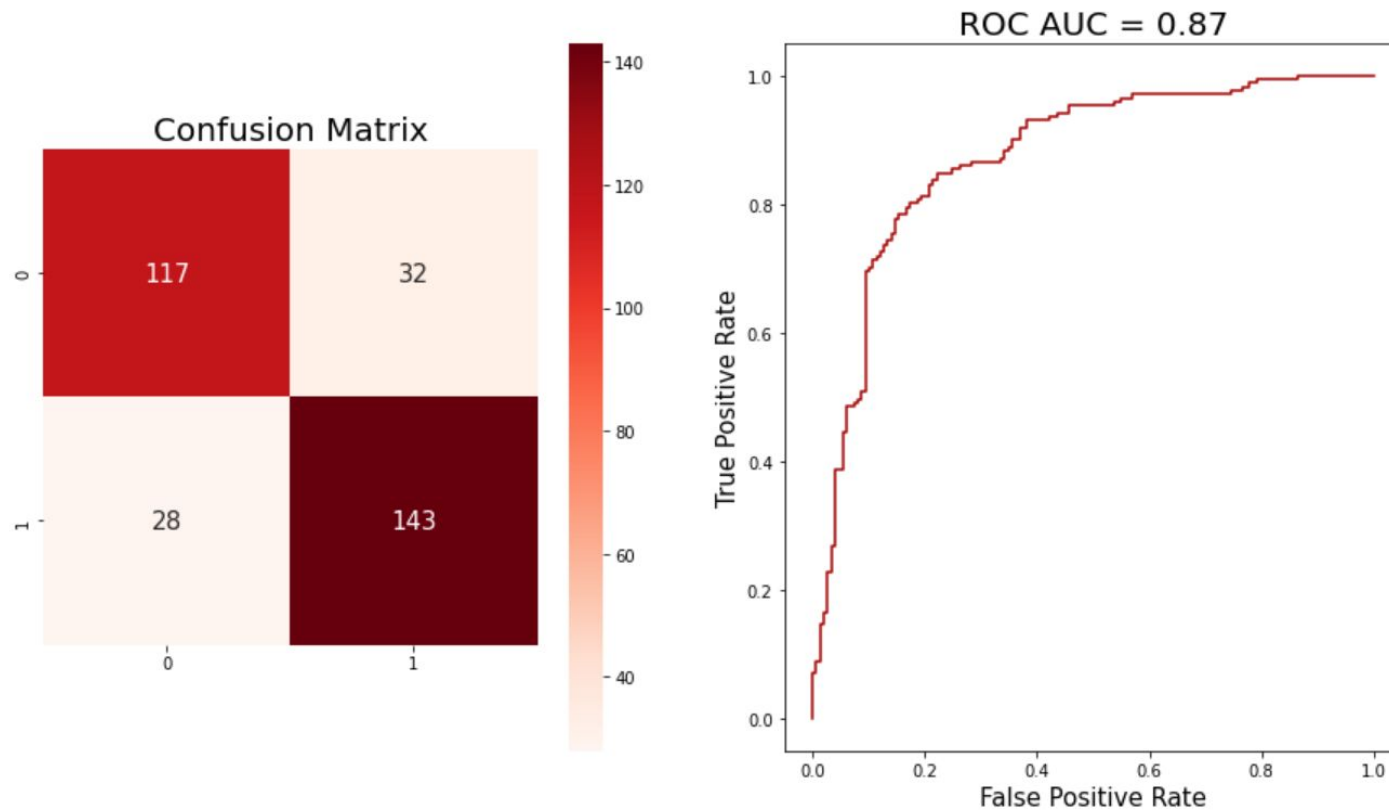


Graphs for Ada Boost Classifier



Results

Graphs for XGBoost Classifier



Hyperparameter Tuning using Optuna

Hyperparameter of XGBoost Classifier:

- **learning_rate:** The step size shrinkage used to prevent overfitting. Range: [0,1].
- **n_estimators:** Number of boosting rounds or trees to build.
- **max_depth:** Maximum depth of a tree. Controls the complexity of the trees.
- **min_child_weight:** Minimum sum of instance weight (hessian) needed in a child. Controls regularization.
- **subsample:** Subsample ratio of the training instances. Range: (0,1].
- **colsample_bytree:** Subsample ratio of columns when constructing each tree.
- **gamma:** Minimum loss reduction required to make a further partition on a leaf node of the tree. Controls tree complexity.
- **lambda (reg_lambda):** L2 regularization term on weights. It's an alternative regularization term.
- **alpha (reg_alpha):** L1 regularization term on weights. It's another alternative regularization term.
- **objective:** The learning objective, e.g., 'binary:logistic' for binary classification.
- **eval_metric:** The evaluation metric to be used for model performance evaluation during training.

Result:

Number of finished trials: 300

Best Parameters: {'lambda': 0.00478927224660259, 'alpha': 0.004140276642810999, 'colsample_bytree': 0.5, 'subsample': 0.6, 'learning_rate': 0.014, 'n_estimators': 900, 'max_depth': 13, 'min_child_weight': 1}

Improvement in XGBClassifier Accuracy: 3.125%

Conclusion

- The Random Forest Classifier and the XGBoost Classifier with hyperparameter tuning emerged as the top-performing algorithms for wine quality prediction.
- These algorithms showcased strong predictive capabilities and demonstrated effectiveness in handling the complexities of the wine quality dataset.
- Support Vector Classifier also exhibited competitive performance, underscoring its suitability for classification tasks in high-dimensional spaces.
- While KNN and AdaBoost demonstrated respectable performance, they may benefit from further optimization or exploration of alternative algorithms.

References

- [1] S. Kumari, A. Misra, A. Wahi and P. S. Rathore, "Quality of Red Wine: Analysis and Comparative Study of Machine Learning Models," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023
- [2] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6, doi: 10.1109/ICCCI48352.2020.9104095. keywords: {processes;data extraction;Naïve Bayes;SVM;Random Forest;quality},
- [3] M. S. Amzad Basha, K. Desai, S. Christina, M. M. Sucharitha and A. Maheshwari, "Enhancing red wine quality prediction through Machine Learning approaches with Hyperparameters optimization technique," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023
- [4] M. V. Gupta and S. K, "Utilization of Random Forest Classifier (RFC) To Predict the Quality of Beverages," 2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI), Greater Noida, India, 2023
- [5] D. Oreški, I. Pihir and K. Cajzek, "Smart Agriculture and Digital Transformation on Case of Intelligent System for Wine Quality Prediction," 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2021
- <https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners>
- <https://xgboost.readthedocs.io/en/stable/>
- <https://www.ibm.com/topics/random-forest>
- <https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/>
- <https://optuna.org/>

Thank You!