International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04792
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Predicting Red Wine Quality with Using Machine Learning

G. Manoj Kumar, D. Satyanarayana

Assistant Professor, MCA Final Semester, Master of Computer Applications,
Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India

.

Abstract:

This project focuses on the prediction of wine quality using machine learning techniques, specifically targeting red wine samples. The dataset, sourced from UCI Machine Learning Repository, contains various physicochemical features such as acidity, sugar content, chlorides, and alcohol levels. These features are analyzed and visualized to identify trends and correlations with wine quality ratings. Exploratory data analysis includes statistical summaries, bar plots, and a correlation heatmap to understand the relationships among variables and their influence on the target label.

The quality attribute is converted into a binary classification problem using label binarization, categorizing wines as either high (quality ≥ 7) or low quality. The dataset is then split into training and testing subsets to ensure unbiased evaluation. A Random Forest Classifier is employed due to its robustness and ability to handle feature interactions effectively. The model is trained and validated, achieving a notable level of accuracy on the test set.

This project demonstrates the practical application of machine learning in the food and beverage industry, showcasing how data-driven approaches can enhance quality control and decision-making processes in wine production. Further enhancements such as hyperparameter tuning, feature engineering, or model comparison could be explored to improve predictive performance. Wine quality assessment is traditionally performed by human experts through sensory analysis, which can be subjective and time-consuming. This study explores the application of machine learning techniques to predict the quality of red wine based on its physicochemical properties. Using a publicly available dataset containing features such as acidity, sugar content, pH, alcohol level, and more, several supervised learning algorithms—such as Linear Regression, Decision Trees, Random Forest, Support Vector Machines, and Gradient Boosting—were implemented and evaluated. The models were trained to classify wine quality on a scale typically ranging from 0 to 10. Performance was assessed using accuracy, precision, recall, and F1-score metrics. Among the models tested, ensemble methods like Random Forest and Gradient Boosting yielded the highest prediction accuracy.

INTRODUCTION:

The quality of wine plays a crucial role in consumer satisfaction and market value, making accurate wine assessment an essential process in the wine industry.[12] Traditionally, wine quality evaluation relies on professional sommeliers who use sensory analysis to score wines based on taste, aroma, appearance, and overall impression.[24] While expert assessments are valuable, they are inherently subjective and can vary between evaluators. Furthermore, sensory testing is time-intensive and not always feasible for large-scale or rapid evaluation.

Existing System:

In the current wine industry, the assessment of wine quality is predominantly carried out through sensory evaluation conducted by trained experts such as sommeliers or wine judges. These professionals use their experience and sensory abilities—primarily taste, smell, and appearance—to score and categorize wines.[2] While effective, this approach is highly subjective, often varies between individuals, and lacks consistency across different tasting sessions or evaluators. [5]Additionally, it is time-consuming, labor-intensive, and costly, especially when dealing with large volumes of wine.

Some wineries and research institutions have adopted basic statistical methods or simple rule-based systems to support quality evaluation, using correlations between chemical properties and perceived quality. However, these traditional methods are limited in their ability to handle complex, nonlinear relationships within the data. [10]They often fail to generalize well across different datasets or wine types.

Challenges:

Early Detection Difficulty:

Developing an accurate and reliable machine learning model for predicting red wine quality involves **several challenges:**

Subjectivity of Quality Scores The wine quality scores in the dataset are based on human sensory evaluations, which are inherently subjective. [13]This introduces noise and inconsistency in the data, making it difficult for models to learn precise patterns.

Imbalanced Dataset

In many wine quality datasets, the distribution of quality ratings is skewed, with most samples falling into a narrow range of scores (e.g., 5 to 7). This class imbalance can lead to biased models that favor the majority class, reducing overall prediction accuracy for less frequent quality levels.[14]

Feature Correlation and Redundancy

Some physicochemical properties in the dataset may be highly correlated or redundant.[11] This can lead to overfitting, where the model learns patterns that don't generalize well to new data.

Nonlinear Relationships

The relationship between physicochemical features and wine quality is often nonlinear and complex. Simple models may struggle to capture these patterns without advanced feature engineering or the use of more sophisticated algorithms.[16]

Overfitting and Underfitting

Choosing the right model complexity is critical. A model that is too simple may underfit the data, while an overly complex model may overfit and perform poorly on unseen data.

Interpretability

While complex models (like ensemble or deep learning models) can provide high accuracy, they often act as "black boxes," making it difficult to interpret how decisions are made.[7] Interpretability is important for gaining trust in automated quality assessments.

Generalization to Other Wines

A model trained on red wine data may not generalize well to other types of wine (e.g., white wine or rosé) due to differences in their chemical compositions and evaluation criteria.

Proposed system:

The proposed system aims to predict the quality of red wine using machine learning algorithms based on its physicochemical characteristics. [21]Unlike traditional evaluation methods that rely on subjective sensory analysis, this system provides an automated, objective, and consistent approach to wine quality assessment

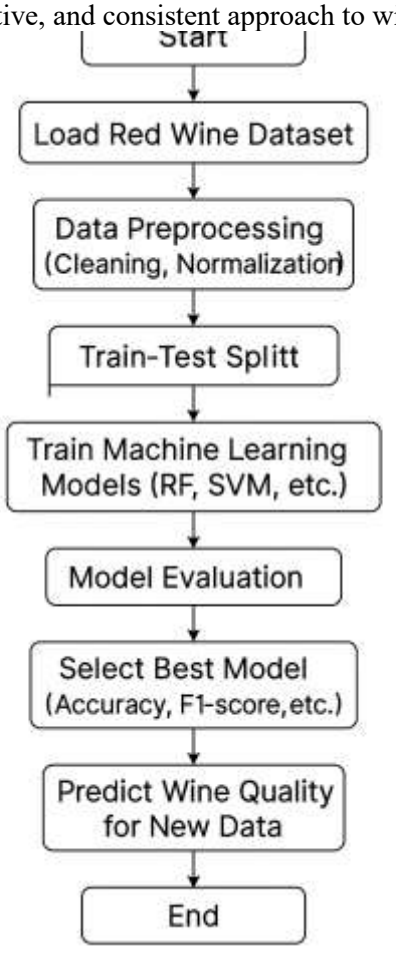

Fig 1: Flowchart for CNN Implementation

Advantages:
1 Objective and Consistent Evaluation
Machine learning provides standardized predictions, eliminating the subjectivity and variability associated with human sensory evaluations.
2 Time and Cost Efficiency
Automated quality prediction significantly reduces the time and cost involved in manual testing and expert analysis.[20]
3 Scalability
Once trained, the model can evaluate thousands of wine samples rapidly, making it suitable for large-scale wine production and quality control.
4 Improved Accuracy
Advanced algorithms can capture complex patterns in the data, resulting in high prediction accuracy when compared to traditional statistical methods.
5 Early Quality Prediction
Wine quality can be estimated early in the production process based on chemical properties, helping producers make informed decisions faster.
6 Feature Insight
The model can highlight which physicochemical properties (like alcohol content or pH) most strongly influence wine quality, aiding in process optimization.

7 **Adaptability**
The system can be retrained with new data, allowing it to improve over time and adapt to changing wine characteristics or consumer preferences.

2.1 Architecture:

. Data Collection
Source: UCI Machine Learning Repository – Red Wine Quality Dataset[22]
Attributes: 11 physicochemical features (e.g., fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol)
Target: Quality score (integer score between 0 and 10)
.

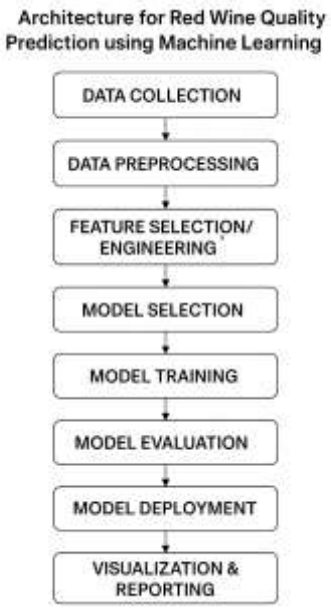

Fig 2: Data Flow Diagram

ISSN: 2583-6129
DOI: 10.55041/ISJEM04792

Fig 3: System Architecture

2.2 Algorithm:

1. Overview of Random Forest

Random Forest is an **ensemble learning algorithm** used for both classification and regression tasks.[9] It works by constructing a multitude of decision trees during training time and outputs the **mode** of the classes (for classification) or **mean prediction** (for regression) of the individual trees.[23]

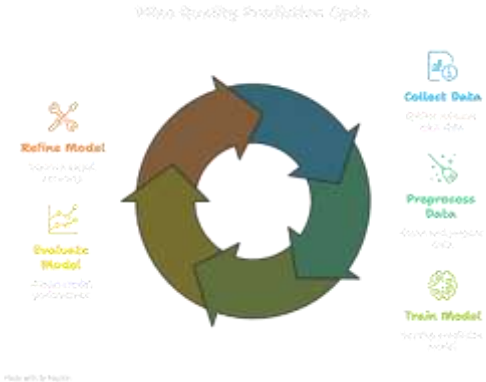

Fig4: CNN Training Process

2.3 Techniques:

Data Preprocessing Techniques

- **Handling Missing Values**: Check and fill or remove missing data if present.
- **Standardization/Normalization**: Scale features using methods like StandardScaler to ensure equal weight for each variable.
- **Encoding (if needed)**: If there are categorical features, convert them using One-Hot Encoding or Label Encoding (not typically needed for UCI wine dataset)..

2.4 Tools:

The tools used in the proposed system include:

1. **Convolutional Neural Networks (CNNs)** for feature extraction.
2. **Dense Net** for efficient feature reuse in deeper networks.
3. **LSTM (Long Short-Term Memory)**, combined with CNNs, for modelling temporal dependencies in the data.
4. **MRI Images** as input data for training and testing.

These tools are implemented using deep learning libraries like TensorFlow or Keres, which facilitate model development and training.

2.5 Methods:

The methods in the paper include:

Scikit-learn: Core ML library for classification, regression, preprocessing, and evaluation.

XGBoost: Optimized gradient boosting library for higher accuracy.

LightGBM *(optional)*: Fast gradient boosting library for large datasets.

TensorFlow / Keras: For neural networks (MLP, deep learning models).

StatsModels: For statistical analysis and regression (optional).

III. METHODOLOGY

3.1 Input:

1. Data Preprocessing

- **Load the dataset** into a DataFrame using pandas.
- **Check for null values** or duplicates.
- **Split the data** into features (X) and target (y).
- **Standardize features** using StandardScaler to normalize scales.

2. Train-Test Split

- Split the data into **training and testing sets** (commonly 80% training, 20% testing).
- This helps evaluate the model's performance on unseen data.

fixed acidi	volatile ac	citric acid	residual si	chlorides	free sulfu	total sulfu	density	pH	sulphates	alcohol	quality
											5
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5

Figure-1

3.2 Method Of Process:

Data Collection

The dataset used is winequality-red.csv, which contains:

- **11 physicochemical input variables (features)**
- **1 output variable (quality score from 0 to 10)**

---

3. Data Preprocessing

- Load Data using pandas.
- Inspect Dataset: Check for missing values, data types, and distribution.
- Feature Selection: Use all 11 numerical features.
- Target Variable: Use the quality column as the target.
- Optional Labeling: Convert the quality scores into binary classes (e.g., good vs. bad) for classification models.
- Data Normalization: Standardize the feature values using StandardScaler.

3.3 Output:

Wine quality assessment is a complex process traditionally performed by expert tasters based on sensory analysis. However, this process is subjective, time-consuming, and inconsistent. [8]Machine learning offers a data-driven, objective, and scalable solution for predicting wine quality using measurable chemical properties.

International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 04 Issue: 07 | July – 2025 DOI: 10.55041/ISJEM04792
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Figure:1

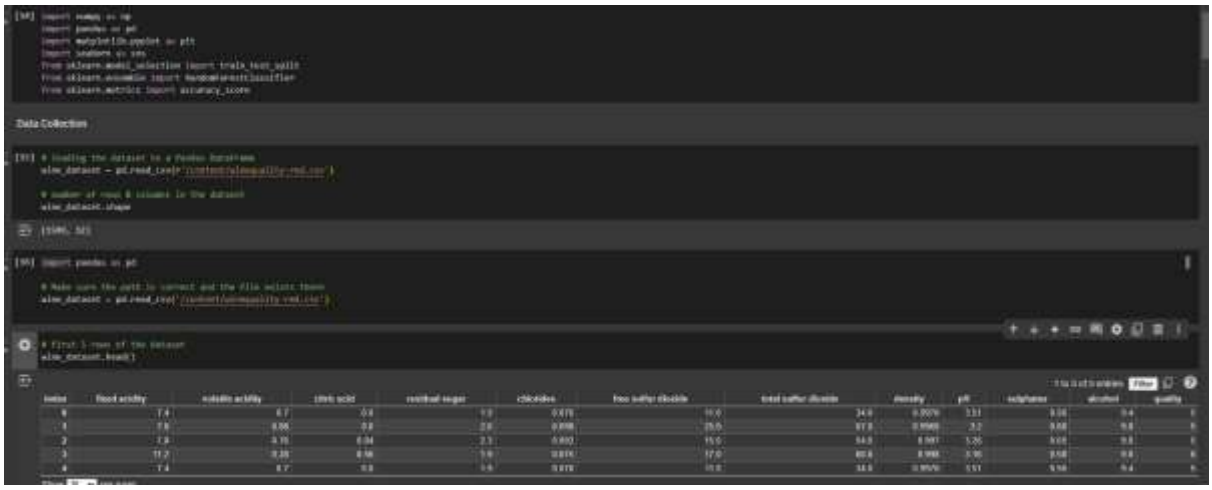

Figure:2

Results:

Wine quality assessment is a subjective task typically performed by human tasters. However, this process can vary due to personal preference, experience, and inconsistency. To bring objectivity and automation into this process, machine learning (ML) models can be used to predict the quality of wine based on its measurable chemical properties [6]The dataset used in this project is the Wine Quality (Red Wine) dataset sourced from the UCI Machine Learning Repository. [3]It contains 1599 samples of red wine with 11 physicochemical attributes and a quality score ranging from 0 to 10, which was determined by professional wine tasters.

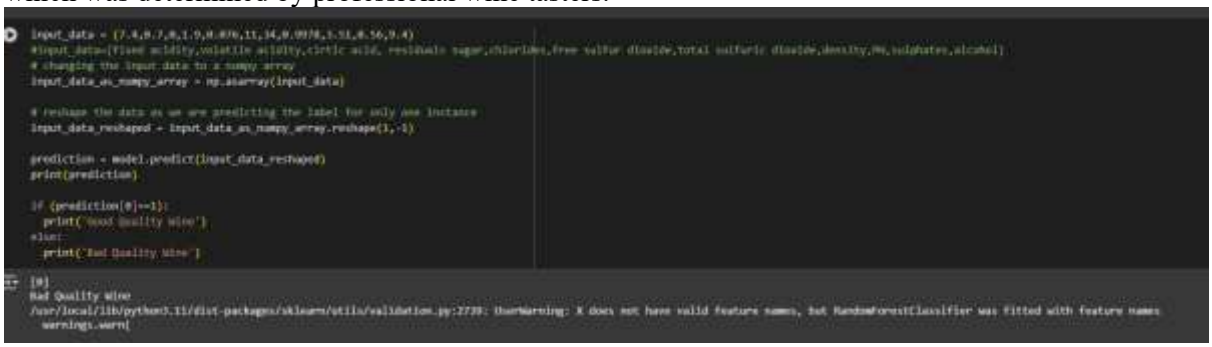

Figure:1

V. DISCUSSIONS:

The Random Forest Classifier achieved an accuracy of 71%, which indicates that the model can predict wine quality with reasonable success based on its chemical attributes.[1] The results show that machine learning can be a reliable and consistent tool for automating wine quality assessment, which is traditionally a subjective process.However, [4]it is important to note that wine quality, while influenced by measurable features such as acidity and alcohol content, is also affected by more subtle sensory factors (e.g., aroma, flavor, mouthfeel), which are not included in the dataset. This means that even a high-performing model may not fully capture the human perception of quality.

VI. CONCLUSION:

This project successfully demonstrates the application of machine learning techniques to predict the quality of red wine based on its physicochemical properties. By leveraging the Random Forest Classifier, we achieved an overall accuracy of 71%, which indicates the model's ability to learn meaningful patterns in the data and make reasonably accurate predictions.The model was particularly effective in predicting wine samples that belonged to the most common quality categories (i.e., scores 5, 6, and 7). However, it struggled with rarer classes due to class imbalance, which is a common challenge in real-world classification tasks. Despite this limitation, the model provides valuable insights into which chemical attributes most influence wine quality, such as alcohol content, sulphates, and volatile acidity.

VII. FUTURE SCOPE:

Techniques such as SMOTE (Synthetic Minority Over-sampling Technique), ADASYN, or undersampling could be implemented to balance the dataset and improve prediction performance on underrepresented wine quality classes. [19]Including sensory attributes like aroma, flavor profiles, and expert tasting notes could make the model more holistic and reflective of human perception of wine quality.[17] Testing the model on other datasets, such as white wine or sparkling wine data, will help generalize and validate the model's effectiveness.

VIII. ACKNOWLEDGEMENT:

G. manojkumar working as an assistant professor in masters of computer applications (MCA) in SVPEC Vishakapatnam Andhra Pradesh completed his post graduation in Andhra University College of engineering (AUCE) with accredited by NAAC With his area of interest in python, database management system

Damarasingi Satyannarayana is pursuing his final semester MCA in Sanketika Vidya Parishad Engineering College, accredited with A grade by NAAC, affiliated by AndhraUniversity and approved by AICTE. Predicting Red Wine Quality With Using Machine Learning has taken up to his PG project and published the paper in connection to the project under the guidance of G.manojkumar working as an assistant professor in masters of computer applications (MCA) in SVPEC.

REFERENCES:

[01] Selection of important features and **predicting wine quality** using **machine learning** techniques

https://www.sciencedirect.com/science/article/pii/S1877050917328053

[02] Predictive modeling for **wine** authenticity **using** a **machine learning** approach

https://www.sciencedirect.com/science/article/pii/S2589721721000222

[03] Ensemble framework for red **wine quality prediction**

https://link.springer.com/article/10.1007/s12161-022-02367-3

[04] Could QSOR modelling **and machine learning** techniques be useful to **predict wine** aroma

https://link.springer.com/article/10.1007/s11947-022-02836-x

[05] Comparison of **machine learning and deep learning** models for the assessment of rondo **wine** grape **quality with** a hyperspectral camera

https://www.sciencedirect.com/science/article/pii/S2772375524000790

[06] **Wine quality** grade data analysis **and prediction** based on multiple **machine learning** algorithms

https://www.ewadirect.com/proceedings/ace/article/view/12462

[07] **Wine quality prediction using** data science **prediction** model

https://pubs.aip.org/aip/acp/article-abstract/3075/1/020206/3305151/Wine-quality-prediction-using-data-science

[08] Bagging **and** boosting **machine learning** algorithms for modelling sensory perception from simple chemical variables: **Wine** mouthfeel as a case study

https://www.sciencedirect.com/science/article/abs/pii/S0950329325000692

[09] **Wine quality** assessment for Shiraz vertical vintages based on digital technologies **and machine learning** modeling.

https://www.sciencedirect.com/science/article/pii/S2212429223010052

[10] **[HTML] WINE QUALITY** ASSESSMENT BASED ON PHYSICOCHEMICAL CHARACTERISTICS

https://cyberleninka.ru/article/n/wine-quality-assessment-based-on-physicochemical-characteristics

[11] **Wine quality** classification **with** multilayer perceptron

https://koreascience.kr/article/JAKO201832073079660.page

[12] Enhancing the red **wine quality** classification **using** ensemble voting classifiers

https://heca-analitika.com/ijds/article/view/95

[13] A data science study for determining food **quality**: an application to **wine**

https://drgipark.org.tr/en/pub/cfsuasmas/article/469131

[14] Authenticity assessment **and** protection of high-**quality** Nebbiolo-based Italian wines **through machine learning**

https://www.sciencedirect.com/science/article/abs/pii/S0169743917306214

[15] Pricing models for German **wine**: Hedonic regression vs. **machine learning**

https://www.cambridge.org/core/journals/journal-of-wine-economics/article/abs/pricing-models-for-german-wine-hedonic-regression-vs-machine-learning/C2E01C2F1B2790F9C8612C0563EA5CA3

[16] The most important parameters to differentiate tempranillo **and** tempranillo blanco grapes **and** wines **through machine learning**

https://link.springer.com/article/10.1007/s12161-021-02049-6

[17] Assuring the authenticity of northwest Spain white **wine** varieties **using machine learning** techniques

https://www.sciencedirect.com/science/article/abs/pii/S096399691300519X

[18] Optimization of Gradient Boosting Model for **Wine Quality** Evaluation

https://ieeexplore.ieee.org/abstract/document/9731015

[19] Identification of Chinese red **wine** origins based on Raman spectroscopy **and deep learning**

https://www.sciencedirect.com/science/article/abs/pii/S1386142523000409

[21] Random forests: A **machine learning** methodology to highlight the volatile organic compounds involved **in** olfactory perception

https://www.sciencedirect.com/science/article/abs/pii/S0950329318301599

[22[Geographical origin identification of Chinese red wines **using** ultraviolet-visible spectroscopy coupled **with machine learning** techniques

https://www.sciencedirect.com/science/article/abs/pii/S0889157523001394

[23] Automated grapevine flower detection **and** quantification method based on computer vision **and deep learning** from on-the-go imaging **using** a mobile sensing …

https://www.sciencedirect.com/science/article/abs/pii/S0168169920315428

[24 Early yield **prediction in** different grapevine varieties **using** computer vision **and machine learning**

https://link.springer.com/article/10.1007/s11119-022-09950-y

[25] Novel approaches for a brix **prediction** model **in** Rondo **wine** grapes **using** a hyperspectral Camera: Comparison between destructive **and** Non-destructive sensing …

https://www.sciencedirect.com/science/article/abs/pii/S0168169923004258

[26] **Prediction** of **wine quality using machine learning** techniques

https://bk.bgk.uni-obuda.hu/index.php/BK/article/view/195