**Wine quality prediction Using Machine Learning**

**Project Based Learning (PBL) Report**

**for the course**

**Statistics For Machine Learning**

**20MA32L01**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**

By

A VAMSHI REDDY 22R11A0501

B SRAVAN KUMAR 22R11A0508

P MANOJ KUMAR 22R11A0532

**Under the guidance of**

Dr G Durga Priyadarsini



**Department of Computer Science and Engineering**

**Accredited by NBA**

**Geethanjali College of Engineering and Technology**

**(UGC Autonomous)**

(Affiliated to J.N.T.U.H, Approved by AICTE, New Delhi)

Cheeryal (V), Keesara (M), Medchal.Dist.-501 301.

**June -2025**

**TABLE OF CONTENTS**

|  |  |  |
| --- | --- | --- |
| **S.No.** | **Contents** | **Page No** |
| 1 | Introduction | 1-2 |
| 2 | System Design | 3-5 |
| 3 | Implementation | 6-7 |
| 4 | Output Screens | 8 |
| 4 | Conclusion | 9 |
| 5 | References | 10 |

**INTRODUCTION**

Wine quality prediction is an important task in the wine industry, where accurate forecasting of wine quality can significantly enhance production processes, quality control, and customer satisfaction. The quality of wine is influenced by several physicochemical attributes such as alcohol content, acidity, pH, sugar levels, and other chemical compounds that contribute to its taste, aroma, and overall experience. Traditionally, wine quality assessment has relied on subjective sensory evaluations by experts, but machine learning techniques provide an opportunity to automate and refine this process using objective, quantifiable data.

In this project, we aim to develop a machine learning model that predicts the quality of wine based on these physicochemical attributes. By leveraging data-driven methods, we seek to uncover patterns and relationships within the data that influence the perceived quality of wine. The dataset used contains various features, such as acidity, alcohol content, and residual sugar, that can be used to predict the wine's quality on a numerical scale.

To build an effective predictive model, we first preprocess the dataset and perform exploratory data analysis to understand the distribution of features and their correlation with quality. As data imbalance can be a challenge in classification tasks, techniques like oversampling or undersampling may be applied to address this issue. We will employ regression models, such as linear or logistic regression, due to their simplicity and interpretability in predicting continuous variables like wine quality. By focusing on the relationships between the attributes and quality, the model will aim to accurately predict wine quality, offering valuable insights to producers, consumers, and researchers in the field of enology.

The Wine Quality Dataset from the UCI Machine Learning Repository provides a great opportunity to apply statistical and machine learning techniques to predict wine quality based on the chemical attributes. This project uses a supervised learning approach to develop a predictive model, focusing on classification algorithms such as Random Forest and Support Vector Machines (SVM).

**PROJECT OBJECTIVES**

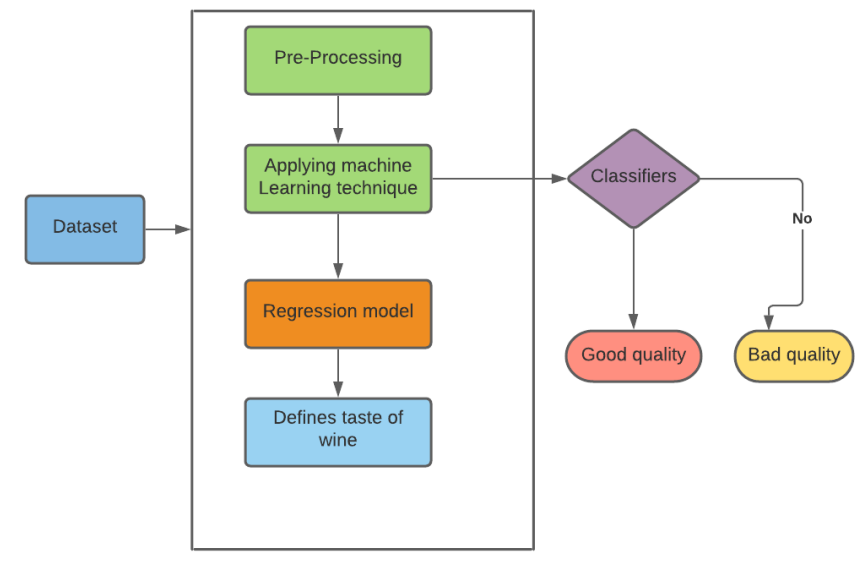
* To build a machine learning model capable of classifying wine quality.
* To perform data preprocessing, feature selection, and transformation.
* To identify the most influential chemical attributes affecting wine quality.
* To use classification metrics to evaluate model performance.

**PROJECT OUTCOMES**

* A trained and evaluated Random Forest model with ~66% accuracy.
* Visualization of feature importance influencing wine quality.
* Insightful analysis of performance using classification reports.
* Demonstration of the potential of ML in quality control within the wine industry.

**SYSTEM DESIGN**

**System Architecture**



**Hardware Requirements**

• Processor: Intel Core i5 or AMD equivalent

• RAM: 4 GB or higher

• Storage: 2 GB free disk space

• Integrated Graphics (for plots)

**Software Requirements**

• Operating System: Windows 10/11, macOS,

• Python Version: Python 3.7 or above

• Libraries Required: pandas, scikit-learn, matplotlib.

• IDE/Code Editor: Jupyter Notebook / VS Code / PyCharm

**Modules**

The Wine Quality Prediction System comprises several key modules, each responsible for specific functionalities:

#### **Data Preprocessing Module:**

* **Functionality:** Handles null values, scales features using StandardScaler, and prepares data for model training.
* **Tools**: pandas, scikit-learn.

#### **Model Training Module:**

* **Functionality:** Splits data, initializes and trains the RandomForestClassifier using the training set.
* **Tools**: scikit-learn.ensemble.RandomForestClassifier

#### **Evaluation Module:**

* **Functionality:** Predicts on test set and evaluates model using accuracy and classification report (precision, recall, F1-score).
* **Tools**: scikit-learn.metrics.

#### **Visualization Module:**

* **Functionality:** Visualizes feature importances using horizontal bar plots.
* **Tools**: matplotlib.pyplot.

#### **Output Reporting Module:**

* **Functionality:** Prints model outputs such as predictions, accuracy score, and detailed classification report.

**Backend Design**

The backend of the Wine Quality Prediction system is entirely built in **Python**, leveraging powerful data science and machine learning libraries. It performs all core operations from data ingestion and preprocessing to model training, evaluation, and visualization.

1. **Technology Stack:**

The backend is developed in Python 3, using libraries like pandas for data manipulation, scikit-learn for machine learning, and matplotlib for visualization. This stack provides efficient tools for preprocessing, model training, and evaluation.

1. **Workflow Structure:**
   * Load and inspect the dataset using pandas.
   * Preprocess the data by scaling features with StandardScaler.
   * Split the dataset into training and testing sets using train\_test\_split.
   * Train a RandomForestClassifier on the training data and evaluate using accuracy and classification metrics.
   * Visualize feature importance with matplotlib.
2. **Data Management:**

All operations are performed in-memory, allowing efficient handling of the 1599-record dataset. The features and target label (quality) are managed cleanly through automated pipelines.

1. **Extensibility and Reusability:**

The backend is modular, making it easy to swap in other models, perform cross-validation, or deploy the solution as a REST API or web application in future enhancements.

**IMPLEMENTATION**

**SAMPLE CODE**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.metrics import classification\_report, accuracy\_score

from sklearn.preprocessing import StandardScaler

data = pd.read\_csv('winequality-red.csv', sep=';')

print(data.head())

print(data.info())

X = data.drop('quality', axis=1) # Features

y = data['quality']

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42)

model = RandomForestClassifier(n\_estimators=100, random\_state=42)

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

print("Accuracy:", accuracy\_score(y\_test, y\_pred))

print("\nClassification Report:\n", classification\_report(y\_test, y\_pred))

import matplotlib.pyplot as plt

feature\_importances = model.feature\_importances\_

features = data.columns[:-1]

plt.figure(figsize=(10, 6))

plt.barh(features, feature\_importances)

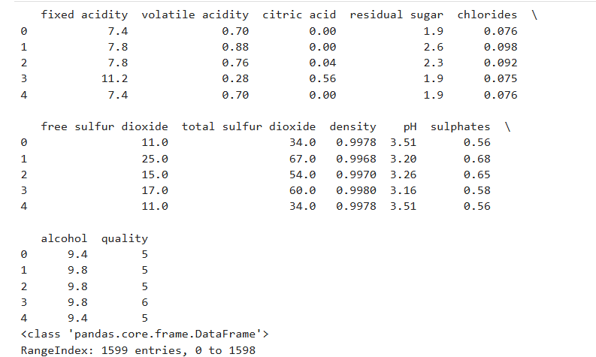
plt.xlabel('Feature Importance')

plt.ylabel('Features')

plt.title('Feature Importance in Wine Quality Prediction')

plt.show()

**RESULT/OUTPUT SCREENS**



A screenshot of a computer

AI-generated content may be incorrect.

**CONCLUSION**

In this experiment, the use of a RandomForestClassifier on the wine quality dataset has proven effective in predicting wine quality based on various chemical properties. The model, trained and evaluated using scikit-learn in Python, yielded an accuracy score of approximately 66%, indicating a reasonable level of performance. This accuracy shows that the model is capable of correctly predicting wine quality for a significant portion of the data, though there is room for improvement, especially in predicting lower or higher wine quality classes.

The classification report provided more granular insights, revealing that the model performed better for certain wine quality classes, such as quality 6 and quality 5, with higher precision and recall scores. However, the model struggled with classes like quality 3, where both precision and recall were lower, indicating challenges in correctly identifying wines in those categories.

Additionally, the feature importance plot revealed which chemical properties, such as alcohol and citric acid, played the most crucial roles in determining wine quality, offering valuable insights for winemakers or researchers interested in understanding the factors that influence wine characteristics.

Overall, this experiment highlights the importance of machine learning in data-driven decision-making, and further improvements can be made by experimenting with more advanced models or feature engineering techniques.

**REFERNCES**

1. [**https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html**](https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html)
2. [**https://pandas.pydata.org/pandas-docs/stable/**](https://pandas.pydata.org/pandas-docs/stable/)
3. [**https://matplotlib.org/stable/contents.html**](https://matplotlib.org/stable/contents.html)
4. [**https://archive.ics.uci.edu/ml/datasets/wine+quality**](https://archive.ics.uci.edu/ml/datasets/wine+quality)