

INDIAN INSTITUTE OF INFORMATION
TECHNOLOGY, ALLAHABAD

INTRODUCTION TO MACHINE LEARNING

Wine Quality Prediction

Author:

Abhishek ANAND

Supervisor:

Dr. Muneendra OJHA



December 2, 2023

Chapter 1

Literature Review

1.1 Prediction of Wine Quality Using Machine Learning Algorithms K. R. Dahal^{1*}, J. N. Dahal², H. Banjade³, S. Gaire⁴

The research paper discusses the application of machine learning (ML) algorithms to predict wine quality based on various parameters.

The authors highlight the importance of understanding the structure of data and fitting it into models using ML, which has been widely used in various sectors such as businesses, medicine, astrophysics, and scientific problems. In this specific study, the authors compare the performance of four ML models: Ridge Regression (RR), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and multi-layer Artificial Neural Network (ANN) to predict wine quality.

The dataset used in the study is obtained from the UCL Machine Learning Repository and focuses on red wine, containing physicochemical properties and sensory scores. The authors discuss data preprocessing steps, including feature scaling and data partitioning into training and testing sets. Various statistical analyses, such as Pearson correlation coefficients, are performed to understand the relationships between different variables and wine quality.

The results indicate that the GBR model surpasses other models in terms of Mean Squared Error (MSE), correlation coefficient (R), and Mean Absolute Percentage Error (MAPE). The authors emphasize the importance of identifying the key parameters that control wine quality, which can be valuable for wine manufacturers in optimizing production processes.

The article concludes by discussing the significance of ML techniques in predicting wine quality and how it can serve as an alternative to traditional methods, potentially saving time and resources in the wine production industry.

1.2 Prediction of Wine Quality: Comparing Machine Learning Models in R Programming Olatunde Akanbi

The paper, "Prediction of Wine Quality: Comparing Machine Learning Models in R Programming," by Olatunde David Akanbi, Taiwo Mercy Faloni, and Sunday Olaniyi, explores the use of machine learning in predicting wine quality. Using the R programming language, the study analyzes physicochemical features of red variants of Portuguese "Vinho Verde" wine. Collaborative efforts across diverse fields, including data science, computer science, chemical, and material engineering, are emphasized.

The dataset comprises physicochemical tests, with input variables like fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The output variable represents wine quality. Challenges in traditional quality assessment methods and the advantages of machine learning are discussed.

The literature review covers previous works in wine quality prediction, highlighting applications of probabilistic neural networks, spectral measurements, and diverse machine learning models. The authors stress the underutilization of R in such studies, despite its accessibility and effectiveness.

Data preprocessing involves cleaning, removal of unnecessary variables, and addressing missing values. Visualization explores variable relationships and patterns. Machine learning algorithms, including linear regression, neural network, naive Bayes, LDA, CART, kNN, SVM, and RF, are applied using R packages such as neuralnet, naivebayes, ggplot2, and caret.

The study identifies random forest as the most accurate model, with alcohol being a key influencer in predicting wine quality. The conclusion underscores the importance of alcohol content, fixed acidity, citric acid, and sulphates. Future research suggestions include addressing unbalanced datasets and further exploring R's capabilities in machine learning.

1.3 Why predict wine quality

The objective of creating a wine quality prediction model is to leverage machine learning and data analysis techniques to accurately forecast or assess the quality of wines. This involves developing a predictive model that can analyze various input features, typically related to the physicochemical composition of the wine, and associate them with a qualitative measure of wine quality. The ultimate goal is to provide winemakers, vineyards, or consumers with a reliable tool to assess and potentially improve wine quality based on measurable characteristics.

1. **Quality Assessment:** Develop a model that can objectively evaluate and quantify the quality of wines. This is particularly useful for winemakers who want an automated and data-driven approach to assess their products..

2. **Decision Support:** Provide a tool that can assist winemakers in making informed decisions during the winemaking process. For example, adjusting certain parameters based on the model's predictions to enhance the overall quality of the final product.
3. **Machine Learning and Artificial Intelligence:** Leveraging machine learning algorithms and artificial intelligence to analyze complex datasets, including electronic health records, imaging data, and genomic information, for more accurate prediction models.
4. **Process Optimization:** Identify key physicochemical factors that significantly influence wine quality. This information can guide winemakers in optimizing the winemaking process to achieve desired quality attributes.
5. **Consumer Guidance:** Offer consumers insights into the potential quality of a wine based on its composition. This can aid consumers in making informed choices when purchasing wines.
6. **Research and Development:** Research and Development: Contribute to the broader understanding of the relationship between physicochemical characteristics and wine quality. This can guide further research and development efforts in the field of winemaking.
7. **Efficiency Improvement:** Streamline quality control processes by replacing or complementing traditional sensory evaluations with a more efficient and potentially more accurate predictive model.

In summary, the objective is to create a robust and reliable model that enhances the ability to assess, understand, and potentially improve wine quality, benefiting both producers and consumers in the wine industry.

Chapter 2

Problem Statement

I plan to predict wine quality using machine learning algorithms. In this report i will explore the methodology, analyze various models, and conclude with future possibilities.

Chapter 3

Abstract

Wine quality prediction is a machine learning task where the goal is to develop a model that can accurately predict the quality of wines based on various features. Typically, datasets used for this task include attributes such as alcohol content, acidity, residual sugar, and more. A popular approach is to use regression models, like Random Forest or Gradient Boosting, to predict the wine quality score on a scale. The model is trained on a dataset containing information about different wines and their corresponding quality ratings. The predictive model can then be applied to new, unseen data to estimate the quality of wines. Evaluating the model's performance involves metrics such as Mean Squared Error or R-squared to assess how well it generalizes to new wine samples. This predictive modeling approach can assist in quality control and optimization of wine production processes.

Chapter 4

Proposed Methodology

Proposed Methodology

Data Import:

Approach

Import the necessary libraries (NumPy, Matplotlib, Pandas, Seaborn, scikit-learn, and XGBoost).

Read the dataset using `pd.read_csv('storeitinaPandasDataFrame(wine)')`.

Data Visualization:

Visualize relationships between features and the target variable using various plots (box plots, violin plots, pair plots).

Exploratory Data Analysis (EDA):

Print basic information about the dataset using `head()`, `shape`, and `describe()` functions.

Check for missing values using `isna().sum()` and visualize them.

Compute the correlation matrix using `corr()` and display it using a heatmap.

Explore the distribution of the target variable ('quality') and other features using count plots, KDE plots, and histograms.

Feature Engineering:

Create a new binary column 'goodquality' based on a threshold (quality >= 7).

Data Visualization:

Visualize relationships between features and the target variable using various plots (box plots, violin plots, pair plots).

Decision Tree, Gaussian Naive Bayes, Random Forest, XGBoost).

Model Training and Evaluation:

Train each model on the training set (X_{train}, Y_{train}). Make predictions on the test set (X_{test}) and evaluate.

Chapter 5

Machine Learning algorithms that I plan to utilize.

5.1 Logistic regression

Logistic Regression is a statistical method used for binary classification, predicting the probability of an observation belonging to one of two classes. It employs the logistic function to model the relationship between independent variables and the log-odds of the event, ensuring predictions fall within the range of 0 to 1.

5.2 KNN

KNN is a simple and intuitive algorithm for classification and regression. It classifies a data point based on the majority class of its k nearest neighbors in the feature space.

5.3 Support Vector Classifier (SVC)

SVC is a machine learning algorithm for classification and regression tasks. It works by finding the hyperplane that best separates classes in a high-dimensional space, maximizing the margin between them.

5.4 Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' theorem. It assumes that features are normally distributed and calculates the probability of a data point belonging to a particular class.

5.5 Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and merges their predictions. It improves accuracy and control overfitting by combining the strength of individual trees.

5.6 XGBoost

XGBoost (Extreme Gradient Boosting) is a powerful gradient boosting algorithm designed for speed and performance. It sequentially builds a series of weak learners and combines them to create a robust predictive model, often used for classification and regression tasks.

Chapter 6

Exploratory Data Analysis

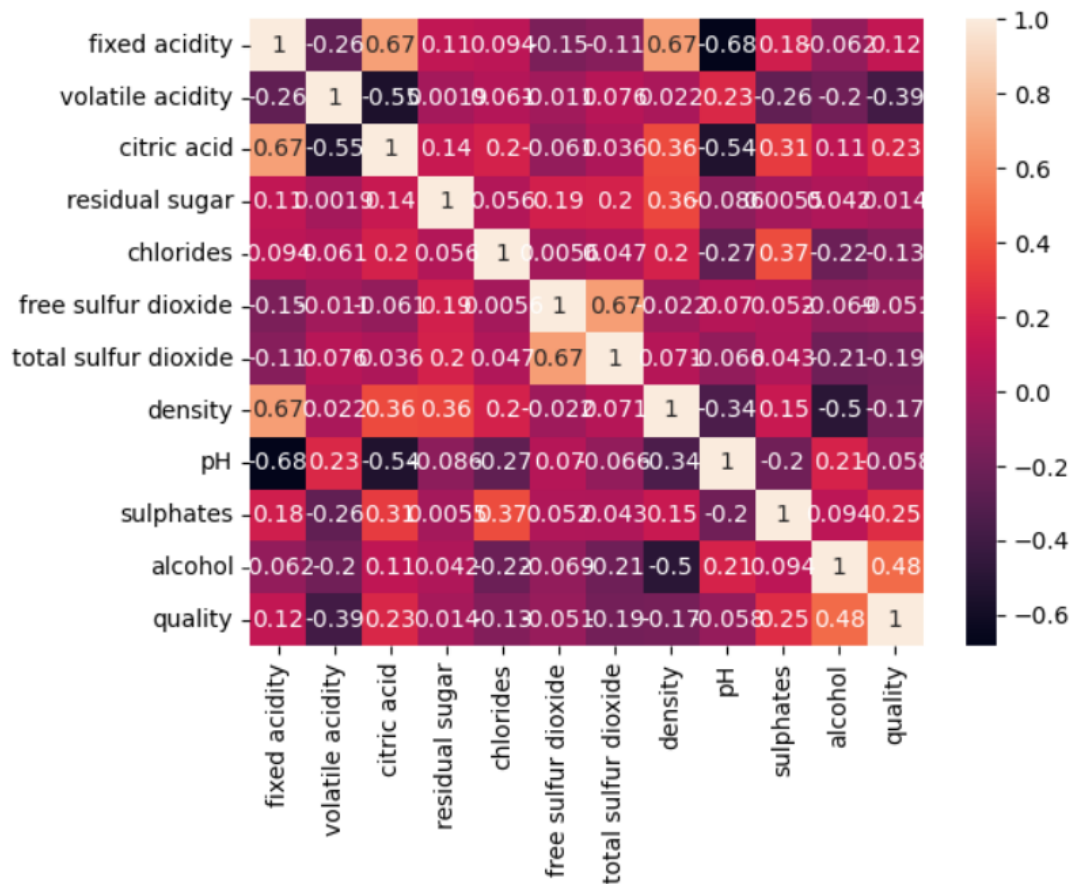


FIGURE 6.1: correlation matrix

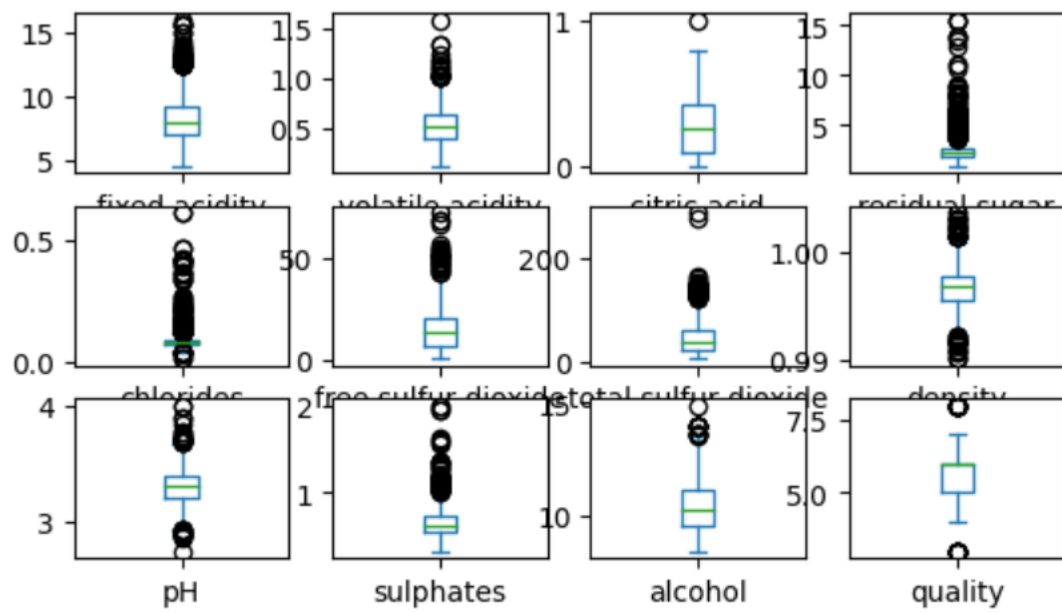


FIGURE 6.2: box plot

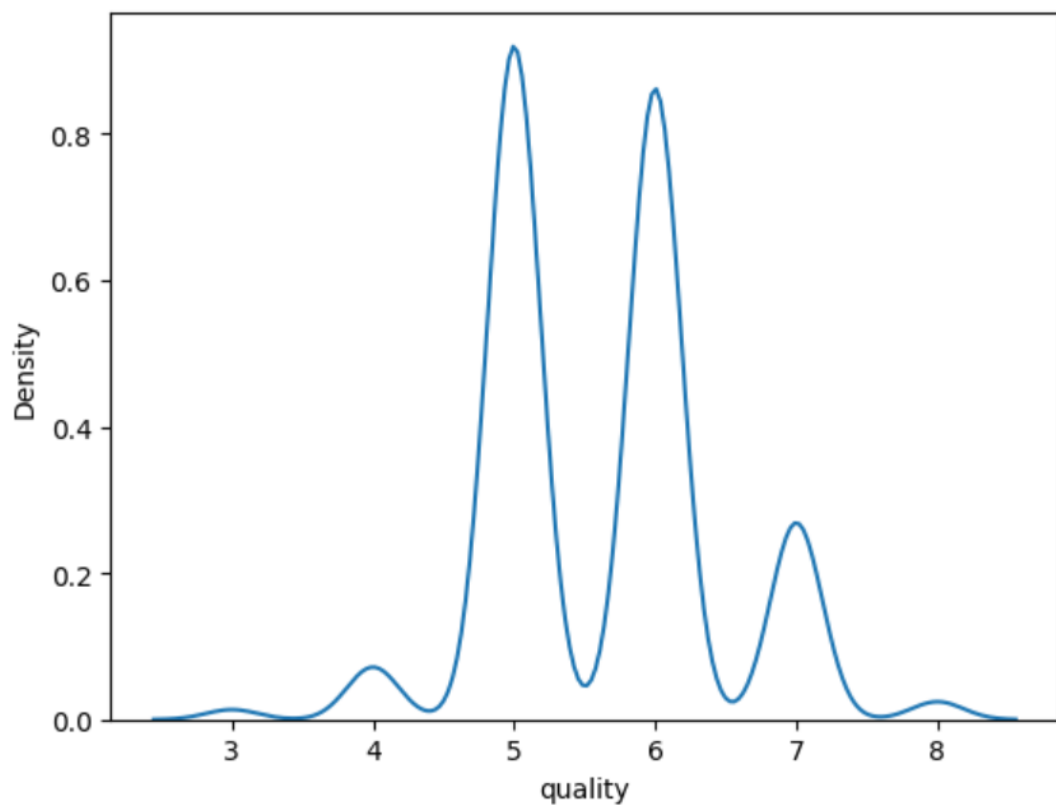


FIGURE 6.3: kdeplot

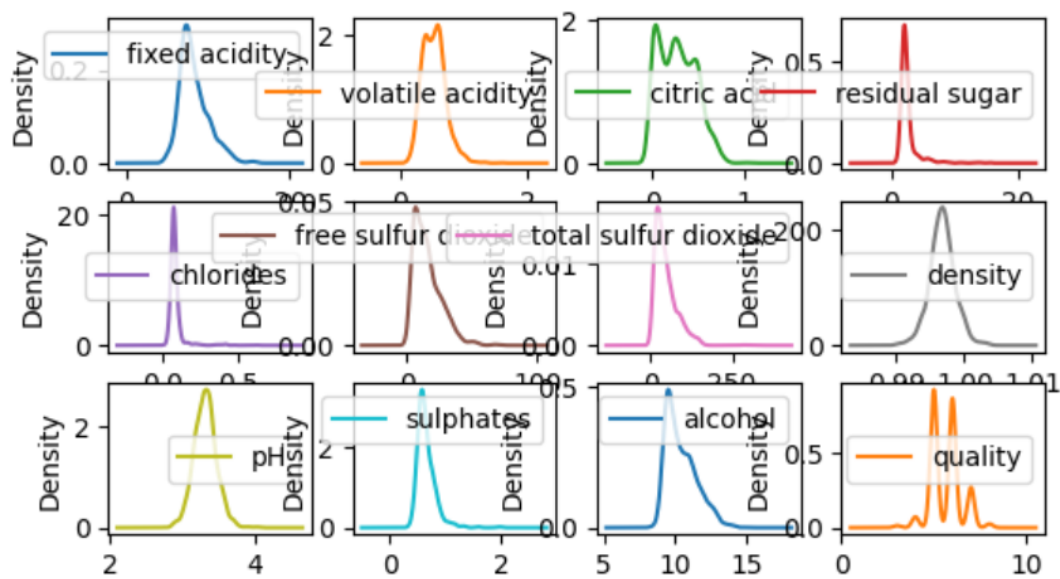


FIGURE 6.4: density vs other features

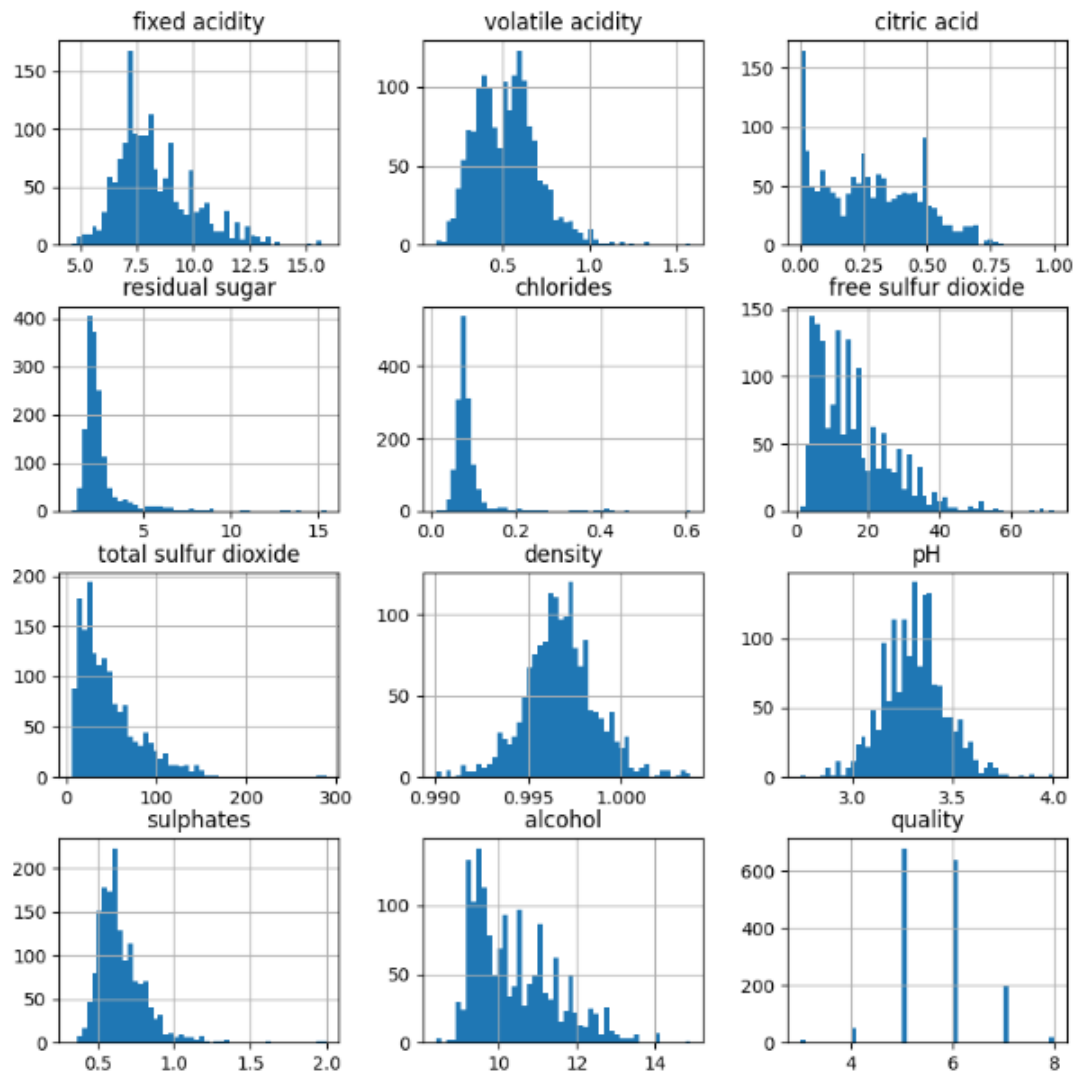
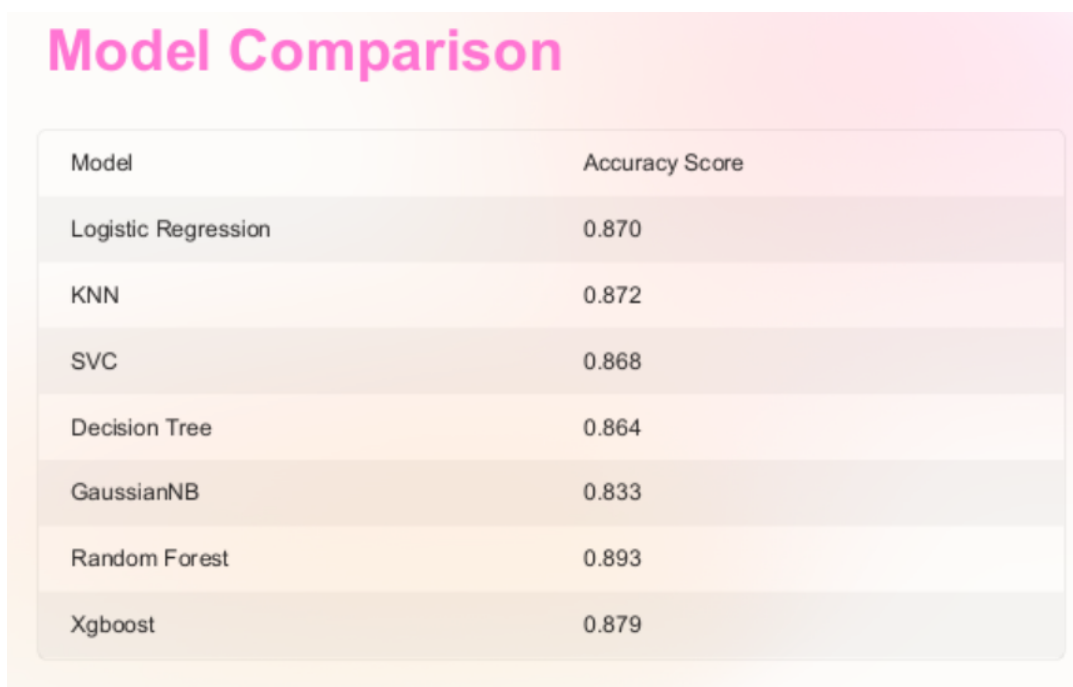


FIGURE 6.5: histogram

Chapter 7

Comparison of Performance with other Models

A table titled "Model Comparison" with two columns: "Model" and "Accuracy Score". It lists eight machine learning models and their corresponding accuracy scores. The table is presented within a light pink rectangular frame.

Model	Accuracy Score
Logistic Regression	0.870
KNN	0.872
SVC	0.868
Decision Tree	0.864
GaussianNB	0.833
Random Forest	0.893
Xgboost	0.879

FIGURE 7.1: Comparison of Performance with other Models

Chapter 8

Comparison of Performance with other Models after hyperparameter tuning

Model	Accuracy Score
Logistic Regression	0.883
KNN	0.852(only one whose accuracy decreases)
SVC	0.893
Decision Tree	0.879
GaussianNB	cannot apply hyperparameter tuning
Random Forest	0.902
Xgboost	0.891

FIGURE 8.1: Comparison of Performance with other Models after hyperparameter tuning

Chapter 9

Conclusion and Future Work

9.1 Conclusion

we can clearly see that random forest gives highest accuracy accuracy both before and after hyperparameter tuning.

9.2 Future Work

1. Assess the applicability of deep learning models
2. Explore the possibility of incorporating additional relevant data sources that could enhance performance.
3. use ensemble techniques, such as stacking, to combine the strengths of multiple models.

Chapter 10

References

1. Kaggle Datasets:

Kaggle Datasets : winequality-red.csv

2. Review Papers:

1. of Wine Quality Using Machine Learning Algorithms K. R. Dahal1*, J. N. Dahal2, H. Banjade3, S. Gaire4
2. of Wine Quality: Comparing Machine Learning Models in R Programming Olatunde Akanbi