Red Wine Quality Prediction Using Machine Learning

A Project report submitted in the partial fulfilment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE AND ENGINEERING

Submitted by

Ch. Sai Sri Ram	**(19471A0573)**
J. Avinash	**(19471A0579)**
K. Raghu Ram Sri Rishik	**(19471A0585)**
V. Narendra Reddy	**(19471A05C9)**

Under the esteemed guidance of

A. Thanuja **M.Tech.,(Ph.D.)**

Assistant Professor

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

NARASARAOPETA ENGINEERING COLLEGE: NARASARAOPET
(AUTONOMOUS)

Accredited by NAAC with A+ Grade and NBA under Cycle -1
NIRF rank in the band of 251-320 and an ISO 9001:2015 Certified
Approved by AICTE, New Delhi, Permanently Affiliated to JNTUK, Kakinada
KOTAPPAKONDA ROAD, YALAMANDA VILLAGE, NARASARAOPET-522601
2022-2023

NARASARAOPET ENGINEERING COLLEGE: NARASARAOPET
(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

This is to certify that the project entitled "RED WINE QUALITY PREDICTION USING MACHINE LEARNING" is a bonafide Work done by **"Ch. Sai Sri Ram (19471A0573), J. Avinash (19471A0579), K. Raghu Ram Sri Rishik (19471A0585), V. Narendra Reddy (19471A05C9)"** in partial fulfilment of the requirements for the award of the degree of **BACHELOR OF TECHNOLOGY** in the Department of **COMPUTER SCIENCE AND ENGINEERING** during the academic year 2022- 2023.

PROJECT GUIDE
A. Thanuja **M.Tech.,(Ph.D.)**
Assistant Professor

PROJECT CO-ORDINATOR
M.Sireesha **M.Tech., Ph.D.**
Associate Professor

HEAD OF THE DEPARTMENT
Dr. S. N.TirumalaRao **M.Tech,Ph.D.**

EXTERNAL EXAMINER

ACKNOWLEDGEMENT

We wish to express our thanks to carious personalities who are responsible for the completion of the project. We are extremely thankful to our beloved chairperson sir **M.V.Koteswara Rao**, **B.sc** who took keen interest on us in every effort throughout this course. We owe out gratitude to our principal **Dr. Sreenivasa Kumar**, **M.Tech, Ph.D(UK),MISTE,FIE(1)** for his kind attention and valuable guidance throughout the course.

We express our deep felt gratitude to **Dr.S.N.Tirumala Rao**, M.Tech, Ph.D. H.O.D,CSE department and our guide **A. Thanuja** **M.Tech, (Ph.D)** of CSE department whose valuable guidance and unstinting encouragement enable us to accomplish our project successfully in time.

We extend our sincere thanks to **Ms.M.Sireesha** **M.Tech.,Ph.D.** Coordinator of the project for extending her encouragement. Their profound knowledge and willingness have been a constant source of inspiration for us throughout this project work.

We extend our sincere thanks to all other teaching and non-teaching staff of department for their cooperation and encouragement during our B.Tech degree. we have no words to acknowledge the warm affection, constant inspiration and encouragement that we receive from our parents.

We affectionately acknowledge the encouragement received from our friends and those who involved in giving valuable suggestions and clarifying out doubts, which had really helped us in successfully completing our project.

By
Ch. Sai Sri Ram (19471A0573)
J. Avinash (19471A0579)
K. Raghu Ram Sri Rishik (19471A0585)
V. Narendra Reddy (19471A05C9)

ABSTRACT

To analyze the quality of red wine before its consumption to preserve human health. A number of variables affects the quality of prediction. These considerations of different factors contribute to the prediction of red wine quality. Using the machine learning algorithms, this study provides a computational intelligence approach for predicting the red wine quality. The proposed research approach uses Support Vector Machine, Naive Bayes Algorithm and Random Forest Classifier. We had used a red wine quality dataset for the prediction purpose, and machine learning methods are used to predict the quality by models comparing their accuracies.

NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

INSTITUTE VISION AND MISSION

INSTITUTION VISION

To emerge as a Centre of excellence in technical education with a blend of effective student centric teaching learning practices as well as research for the transformation of lives and community.

INSTITUTION MISSION

M1: Provide the best class infra-structure to explore the field of engineering and research.

M2: Build a passionate and a determined team of faculty with student centric teaching, imbibing experiential, innovative skills.

M3: Imbibe lifelong learning skills, entrepreneurial skills and ethical values in students for addressing societal problems.

NARASARAOPETA
NEC ENGINEERING COLLEGE
(AUTONOMOUS)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

VISION OF THE DEPARTMENT

To become a centre of excellence in nurturing the quality Computer Science & Engineering professionals embedded with software knowledge, aptitude for research and ethical valuesto cater to the needs of industry and society.

MISSION OF THE DEPARTMENT

The department of Computer Science and Engineering is committed to

M1: Mould the students to become Software Professionals, Researchers and Entrepreneursby providing advanced laboratories.

M2: Impart high quality professional training to get expertize in modern software tools and technologies to cater to the real time requirements of the industry.

M3: Inculcate team work and lifelong learning among students with a sense of societal and ethical responsibilities.

NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Specific Outcomes (PSO’s)

PSO1: Apply mathematical and scientific skills in numerous areas of Computer Science and Engineering to design and develop software-based systems.

PSO2: Acquaint module knowledge on emerging trends of the modern era in Computer Science and Engineering

PSO3: Promote novel applications that meet the needs of entrepreneur, environmental and social issues.

NEC NARASARAOPETA ENGINEERING COLLEGE
(AUTONOMOUS)

Program Educational Objectives (PEO’s)

The graduates of the programme are able to:

PEO1: Apply the knowledge of Mathematics, Science and Engineering fundamentalsto identify and solve Computer Science and Engineering problems.

PEO2: Use various software tools and technologies to solve problems related toacademia, industry and society.

PEO3: Work with ethical and moral values in the multi-disciplinary teams and can communicate effectively among team members with continuous learning.

PEO4: Pursue higher studies and develop their career in software industry.

Program Outcomes

1. **Engineering knowledge:** Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.

2. **Problem analysis:** Identify, formulate, research literature, and analyse complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. **Design/development of solutions:** Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. **Conduct investigations of complex problems:** Use research-based knowledge and research methods including design of experiments, analysis and interpretation of data, andsynthesis of the information to provide valid conclusions.

5. **Modern tool usage:** Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modelling to complex engineering activities with an understanding of the limitations.

6. **The engineer and society:** Apply reasoning informed by the contextual knowledgeto assess societal, health, safety, legal and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. **Environment and sustainability:** Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. **Ethics:** Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. **Individual and team work**: Function effectively as an individual, and as a member or leader in diverse teams, and in multidisciplinary settings.

10. **Communication**: Communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.

11. **Project management and finance**: Demonstrate knowledge and understanding of the engineering and management principles and apply these to one's own work, as a member and leader in a team, to manage projects and in multidisciplinary environments.

12. **Life-long learning**: Recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.

Project Course Outcomes (CO'S):

CO425.1: Analyse the System of Examinations and identify the problem.

CO425.2: Identify and classify the requirements.

CO425.3: Review the Related Literature**. CO425.4:** Design and Modularize the project.

CO425.5: Construct, Integrate, Test and Implement the Project.

CO425.6: Prepare the project Documentation and present the Report using appropriate method.

Course Outcomes – Program Outcomes mapping

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1		✓											✓		
C425.2	✓		✓		✓								✓		
C425.3				✓		✓	✓	✓					✓		
C425.4			✓			✓	✓	✓					✓	✓	
C425.5					✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
C425.6									✓	✓	✓		✓	✓	

Course Outcomes – Program Outcome correlation

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PO12	PSO1	PSO2	PSO3
C425.1	2	3											2		
C425.2			2		3								2		
C425.3				2		2	3	3					2		
C425.4			2			1	1	2					3	2	
C425.5					3	3	3	2	3	2	2	1	3	2	1
C425.6									3	2	1		2	3	

Note: The values in the above table represent the level of correlation between CO's and PO's:

1. Low level
2. Medium level
3. High level

Project mapping with various courses of Curriculum with Attained PO's:

Name of the course from Which principles are applied in this project	Description of the device	Attained PO
C3.2.4, C3.2.5	Gathering the requirements and defining the problem, plan to develop a smart bottle for health care using sensors.	PO1, PO3
CC4.2.5	Each and every requirement is critically analyzed, the process model is identified and divided into five modules	PO2, PO3
CC4.2.5	Logical design is done by using the unified modelling language which involves individual team work	PO3, PO5, PO9
CC4.2.5	Each and every module is tested,integrated, and evaluated in our project	PO1, PO5
CC4.2.5	Documentation is done by all our four members in the form of a group	PO10
CC4.2.5	Each and every phase of the work ingroup is presented periodically	PO10, PO11
CC4.2.5	Implementation is done and the project will be handled by the hospital management and in future updates in our project can be done based on air bubbles occurring in liquid in saline.	PO4, PO7
CC4.2.8 CC4.2.	The physical design includes hardware components like sensors, gsm module, software and Arduino.	PO5, PO6

INDEX

S. No. CONTENTS PAGENO

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1 INTRODUCTION

Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, the presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.

ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on red wine data. The model can then accurately predict quality by using the necessary elements as inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict the quality of red wine. Classification is the best choice available to fulfill our needs. We use classification models in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for quality prediction is used.

Preprocessing of the dataset done first. Then we trained models with training data and finally evaluated these models based on testing data. In this article, we used several models of classification, for example, Support Vector Machine, Naïve Bayes Algorithm and Random Forest Classifier. It is found that the random forest classifier provides the highest accuracy of 80.24. The inclusion of a novel method of quality prediction is the main goal of this work.

1.2 EXISTING SYSTEM

Nowadays people try to lead a luxurious life. They tend to use the things either for show off or for their daily basis. These days the consumption of red wine is very common to all. So it became important to analyze the quality of red wine before its consumption to preserve human health.

Disadvantages

- Doesn't generate accurate and efficient results.
- Computation time is very high.
- Lacking of accuracy may result in lack of efficient further prediction.

1.3 PROPOSED SYSTEM

By using this system we can reduce its consumption, enhance quality. It is easier to predict the quality and it will also help to reduce the consumption.

Advantages:

- Generates accurate and efficient results.
- Computation time is greatly reduced.
- Reduces manual work.
- Efficient further prediction.

1.4 SYSTEM REQUIREMENTS

1.4.1 HARDWARE REQUIREMENTS

- Processor : Intel Core i5
- Cache Memory : 4MB
- Hard Disk : 30GB or more
- RAM : 1GB or more

1.4.2 SOFTWARE REQUIREMENTS

- Operating System : Window 10
- Coding Language : Python
- Python Distribution : Anaconda, Flask
- Browser : Any Latest Browser Like Chrome

2. LITERATURE

2.1 MACHINE LEARNING

Today, various customers appreciate wine to an ever increasing extent. Wine industry is looking into new advances for both wine making and offering structures in order to back up this development [1]. Physicochemical and tactile tests are utilized for assessing wine confirmation [2]. The segregation of wines isn't a simple procedure inferable from the intricacy and heterogeneity of its headspace. The arrangement of wines is significant in light of the fact that of various reasons. These reasons are financial estimation of wine items, to secure and guarantee the nature of wines, to preclude corruption of wines, and to control refreshment preparing [3].

Data mining innovations have been applied to plan wine quality. The point of machine learning techniques like various applications is to make models from information to anticipate wine quality. In 1991, a "Wine" informational index which contains 178 occurrences with estimations of 13 distinctive synthetic constituents, such as, alcohol, magnesium was given into UCI store to order three cultivars from Italy [4]. For new information mining classifiers this data has been significantly utilized as a benchmark since it is exceptionally simple to separate.

For wine characterization as indicated by geological area; Principal Component Analysis (PCA) was done and announced [5]. The information they utilized in their examination incorporates 33 Greek wines with physicochemical factors. Another work of wine grouping relied upon the physicochemical data. This data associated with wine smell chromatograms as estimated with a Fast GC Analyser [6]. In the last investigation, three portrayal methods, for example, Naïve Bayes, Random Forest and Support Vector Machines (SVM) are contrasted agreeing and their exhibition in a two-organized architecture. Some have proposed a couple of uses of data mining frameworks to wine quality appraisal. Cortez et al. [1] proposed a taste desire framework.

2.2 SOME MACHINE LEARNING METHODS

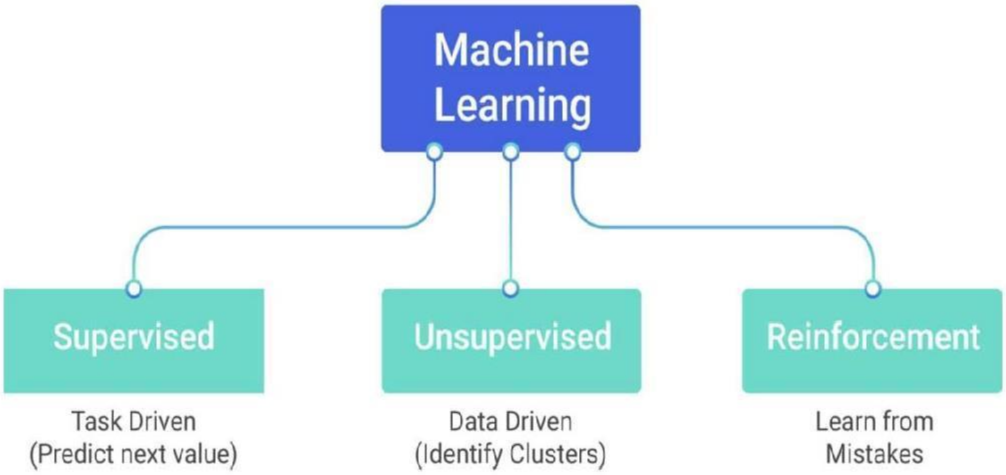


Figure: 2.2. Types of Machine Learning

Machine learning algorithms are often categorized as supervised and unsupervised.

- **Supervised machine learning algorithms:**

Supervised machine learning algorithms can apply what has been learned in the past to new data using labeled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

- **Unsupervised machine learning algorithms:**

Unsupervised machine learning algorithms are used when the information used to train is neither classified nor labeled. Unsupervised learning studies how systems can infer a function

to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.

- **Reinforcement machine learning algorithms:**

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best. This is known as the reinforcement signal.

2.3 APPLICATIONS OF MACHINE LEARNING

1. Virtual Personal Assistants
2. Predictions while Commuting
3. Videos Surveillance
4. Social Media Services
5. Email Spam and Malware Filtering
6. Online Customer Support
7. Search Engine Result Refining
8. Product Recommendations
9. Online Fraud Detection

3.SYSTEM ANALYSIS

3.1 IMPORTANCE OF MACHINE LEARNING IN PYTHON

The importance of machine learning in wine quality is increasing because of its ability to process huge datasets efficiently beyond the range of human capability, and then dependably convert analysis of that data into clinical insights that assist in planning and providing care, which ultimately leads to better outcomes, reduces the consumption. Using these types of advanced analytics, we can provide better information at the point of consumption.

3.2 IMPLEMENTATION OF MACHINE LEARNING USING PYTHON

Python is a popular programming language. It was created in 1991 by Guido van Rossum. It is used for:
1. web development (server-side), 2. software development, 3. mathematics,
4. system scripting.

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than security updates, is still quite popular.

It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that used in Machine Learning are:

1. Numpy 2. Scipy 3. Scikit-learn 4. Pandas 5. Matplotlib

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

SciPy is a very popular library among Machine Learning enthusiasts as it contains different modules for optimization, linear algebra, integration and statistics. There is a difference between the SciPy library and the SciPy stack. The SciPy is one of the core packages that make up the SciPy stack. SciPy is also very useful for image manipulation.

Skikit-learn is one of the most popular Machine Learning libraries for classical Machine Learning algorithms. It is built on top of two basic Python libraries, NumPy and SciPy. Scikit-learn supports most of the supervised and unsupervised learning algorithms. Scikit learn can also be used for data-mining and data-analysis, which makes it a great tool who is starting out with Machine Learning.

Pandas is a popular Python library for data analysis. It is not directly related to Machine Learning. As we know that the dataset must be prepared before training. In this case, Pandas comes handy as it was developed specifically for data extraction and preparation. It provides high-level data structures and wide variety tools for data analysis. It provides many inbuilt methods for groping, combining and filtering data.

Matpoltlib is a very popular Python library for data visualization. Like Pandas, it is not directly related to Machine Learning. It particularly comes in handy when a programmer wants to visualize the patterns in the data. It is a 2D plotting library used for creating 2D graphs and plots. A module named pyplot makes it easy for programmers for plotting as it provides features to control line styles, font properties, formatting axes, etc. It provides various kinds of graphs and plots for data visualization, histogram, error charts, bar chats, etc.

3.3 SCOPE OF THE PROJECT

The scope of this system is to maintain patient details in datasets, train the model using the large quantity of data present in datasets and predict whether presence or absence of disease on new data during testing.

3.4 DATA SET ANALYSIS

We collected the data set from the Kaggle. Data Set consists of 12 variables. They are:

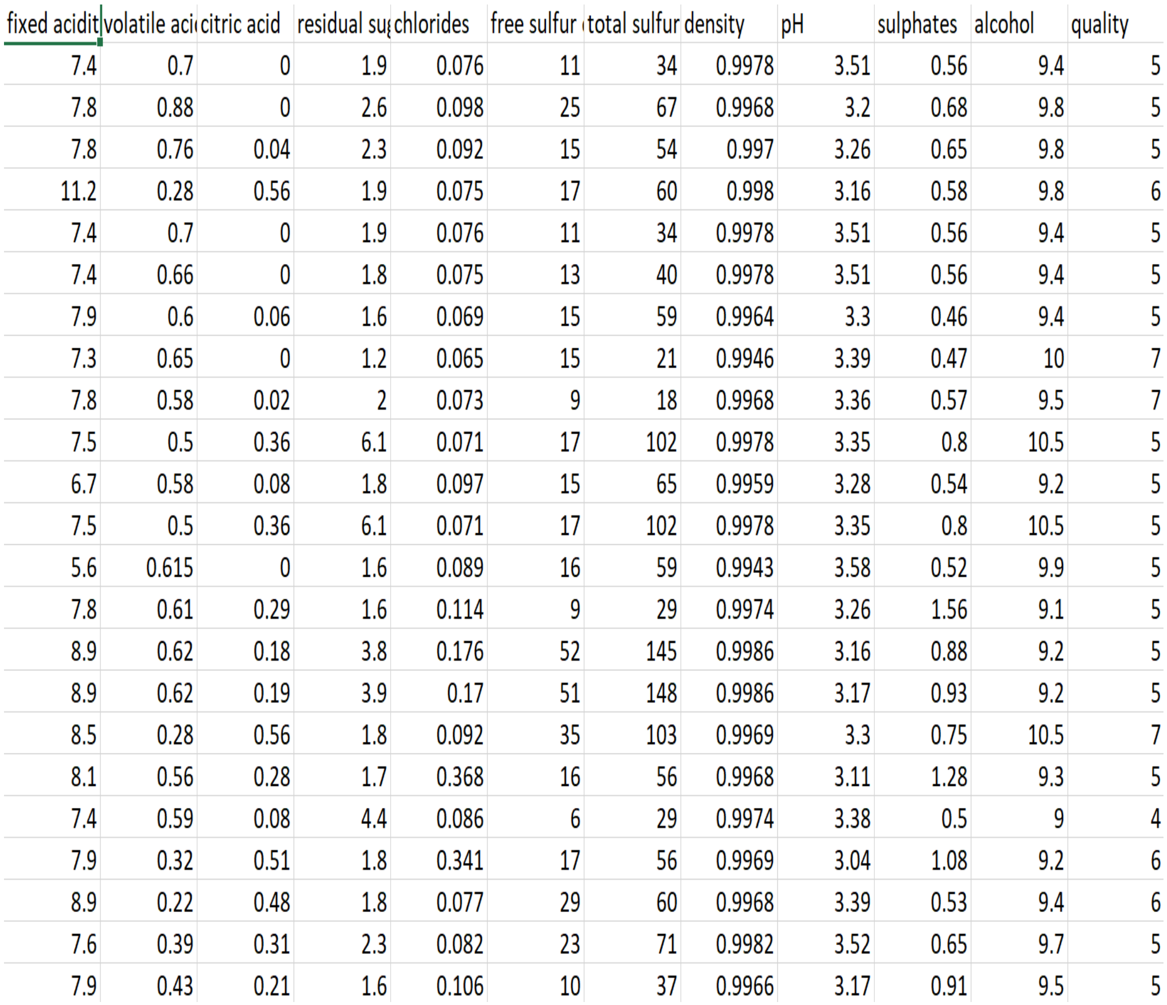
- **Fixed Acidity:** These are non-volatile acids that do not evaporate readily.
- **Volatile Acidity:** These are high acetic acid in wine which leads to an unpleasant vinegar taste.
- **Citric Acid:** It acts as a preservative to increase acidity (small quantities add freshness and flavor to wines).
- **Residual Sugar:** It is the amount of sugar remaining after fermentation stops.
- **Chlorides:** It is the amount of salt in the wine.
- **Free Sulfur Dioxide:** It prevents microbial growth and the oxidation of wine.
- **Total Sulfur Dioxide:** It is the amount of free + bound forms of SO2.
- **Density:** The sweeter wines have a higher density.
- **pH:** The level of acidity.
- **Sulphates:** It is a wine additive that contributes to SO2 levels and acts as an antimicrobial

and antioxidant.

- **Alcohol:** The amount of alcohol in wine.
- **Quality:** It describes the quality whether it is good or bad.

Data Set Link:

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009



fixed acidit	volatile aci	citric acid	residual su	chlorides	free sulfur	total sulfur	density	pH	sulphates	alcohol	quality
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	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

Figure: 3.4. Data Set

3.5 METHADODLOGY

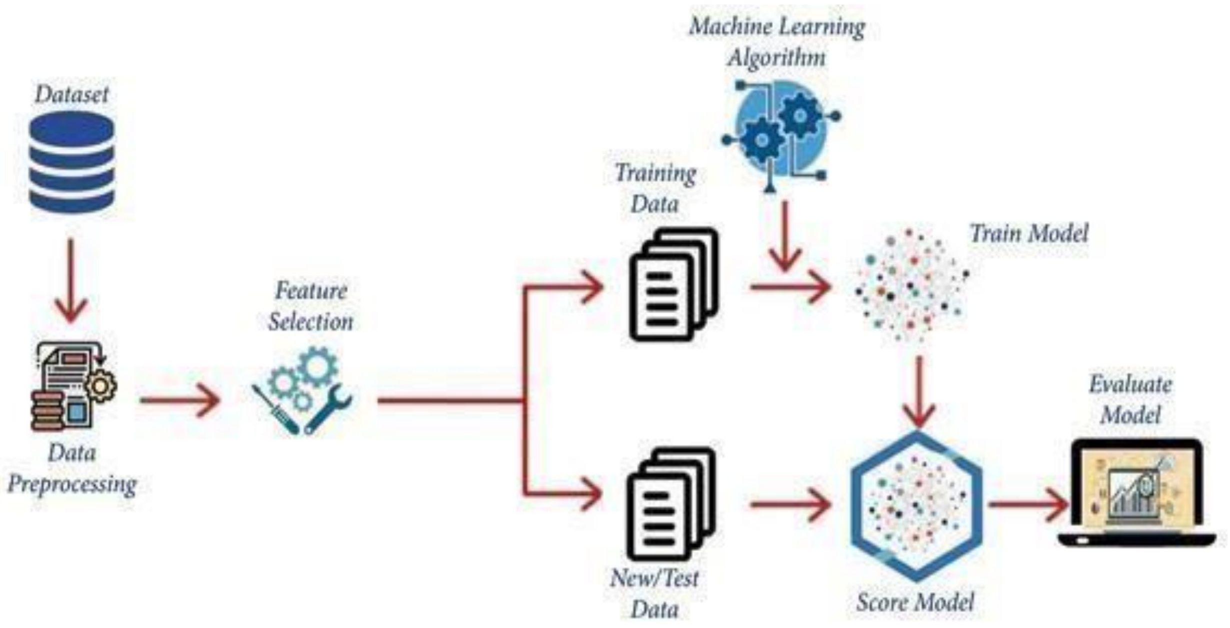


Figure: 3.5. Methodology

3.6 DATA PREPROCESSING

Before feeding data to an algorithm, we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing, the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.

fixed acidity	Sample's fixed acidity value in float
volatile acidity	Sample's volatile acidity value in float
citric acid	Sample's citric acid value in float
residual sugar	Sample's residual sugar value in float
chlorides	Sample's chloride value in float
free sulfur dioxide	Sample's free sulfur dioxide value in float
total sulfur dioxide	Sample's total sulfur dioxide value in float
density	Sample's density value in float
pH	Sample's pH value in float
sulphates	Sample's sulphates value in float
alcohol	Sample's alcohol value in float
quality	Good or Bad

Table: 3.6. categorical variables after translated into numeric or binary values

Data Pre-processing:

Pre-processing refers to the transformations applied to our data before feeding it to the algorithm. Data Pre-processing is a technique that is used to convert the raw data into a clean data set. In other words, whenever the data is gathered from different sources it is collected in raw format which is not feasible for the analysis.

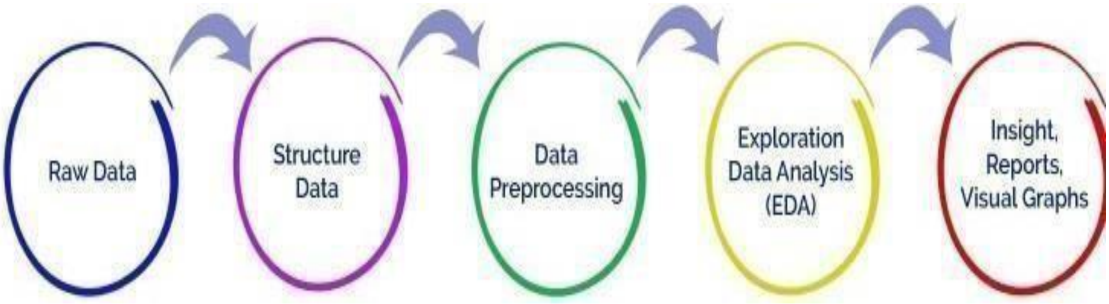

Raw Data
Structure Data
Data Preprocessing
Exploration Data Analysis (EDA)
Insight, Reports, Visual Graphs

Figure: 3.6. Data Preprocessing

Need of Data Preprocessing:

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

3.7 CORRELATION COEFFICIENT METHOD

We can find dependency between two attributes p and q using Correlation coefficient method using the formula.

$rp, q= \sum(pi-p)(qi-q)/n\sigma p\sigma q$

$=\sum(pi\ qi)-np\ q/\ n\sigma p\sigma q$

n is the total number of patterns, pi and qi are respective values of p and q attributes in patterns i, p and q are respective mean values of p and q attributes, σp , σq are respective

standard deviations values of p and q attributes. Generally, -1≤ rp,q ≤ +1. If rp,q < 0, then p and q are negatively correlated. If rp,q =0, then p and q are independent attributes and there is no correlation between them. If rp,q > 0, then p and q are positively correlated. We can drop the attributes that are having correlation coefficient value as 0 as it indicates that the variables are independent with respect to the prediction attribute. Fig:3.8.2 is the correlation heat map. After applying correlation the attributes are PR interval , QRS duration , QT interval , QTc interval, P wave , T wave , QRS wave and problem . The attribute Vent_rate got dropped.

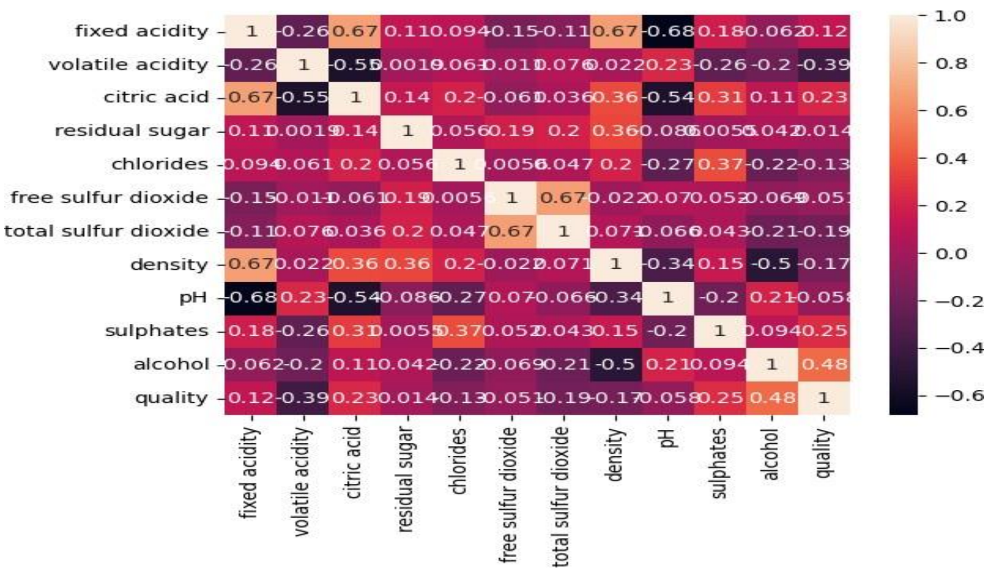


Figure: 3.7. Correlation

3.8 CROSS VALIDATION:

Cross-validation is a technique in which we train our model using the subset of the data- set and then evaluate using the complementary subset of the data-set. The three steps involved in cross-validation are as follows:

- Reserve some portion of sample data-set.
- Using the rest data-set train the model.
- Test the model using the reserve portion of the data-set.0

3.9 CLASSIFICATION

It is a process of categorizing data into given classes. Its primary goal is to identify the class of our new data.

Machine learning algorithms for classification

Research on data mining has led to the formulation of several data mining algorithms. These algorithms can be directly used on a dataset for creating some models or to draw vital conclusions and inferences from that dataset. Some popular data mining algorithms are Random Forest, Gaussian Naïve Bayes, Support vector machine etc.

Gaussian Naive Bayes :

It is a simple technique for constructing classifiers. It is a probabilistic classifier based on Bayes' theorem. All Naive Bayes classifiers assume that the value of any particular feature is independent of the value of any other feature, given the class variable. Bayes theorem is given as follows: $P(C|X) = P(X|C) * P(C)/P(X)$, where X is the data tuple and C is the class such that P(X) is constant for all classes. Though it assume an unrealistic condition that attribute values are conditionally independent, it performs surprisingly well on large datasets where this condition is assumed and holds.

Support vector machine :

Support vector machine is a linear model for classification and regression problems. It is a supervised machine learning algorithm. It can solve linear and non-linear problems and work well many practical problems. The idea of support vector machine is simple: The algorithm creates a line or hyperplane which separates the data into classes. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification, implicitly mapping their inputs into high dimensional feature spaces.

Random Forest Classifier :

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.

3.10 TESTING DATA

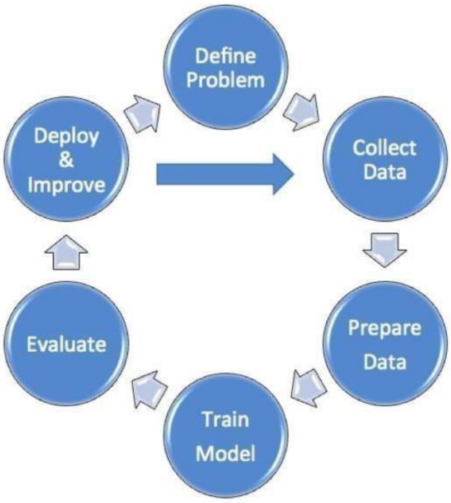


Figure: 3.10. Prediction model for red wine quality

Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

3.11 IMPLEMENTATION OF CODE

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.model_selection import KFold
from sklearn.svm import SVC
from sklearn.naive_bayes import GaussianNB
from sklearn.ensemble import RandomForestClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix,accuracy_score
from sklearn.metrics import r2_score,mean_squared_error,mean_absolute_error
import warnings
warnings.filterwarnings('ignore')

data = pd.read_csv('winequality-red.csv')
data.head()

data.shape

data.columns

data.describe()

data.info()

sns.countplot(x='quality',data=data)

corr = data.corr()
sns.heatmap(corr,annot=True)
```

```
data.hist(figsize=(10,10),bins=50)
plt.show()

data.plot(kind="box",subplots=True,sharex=False,sharey=False,figsize=(20,10),color
='deeppink')

def outliers(df,ft):
        Q1=df[ft].quantile(0.25)
        Q3=df[ft].quantile(0.75)
        IQR=Q3-Q1
        lower_bound=Q1-1.5*IQR
        upper_bound=Q3+1.5*IQR
        ls=df.index[(df[ft]<lower_bound)|(df[ft]>upper_bound)]
        return ls

 index_list=[]
 for feature in ['fixed acidity','volatile acidity','citric acid','residual sugar','chlorides','free
 sulfur dioxide','total sulfur dioxide','density','pH','sulphates','alcohol','quality']:
 index_list.extend(outliers(data,feature))

 for i in index_list:
        print(i,end=" ")

 def remove(df,ls):
        ls=sorted(set(ls))
        df=df.drop(ls)
        return df

 data_cleaned=remove(data,index_list)

data_cleaned.shape

data_cleaned.plot(kind="box",subplots=True,sharex=False,sharey=False,figsize=(20,10)
,color='deeppink')
```

```
data_cleaned = data_cleaned.loc[data_cleaned['fixed acidity']<11,:]
data_cleaned = data_cleaned.loc[data_cleaned['volatile acidity']<0.95,:]
data_cleaned = data_cleaned.loc[data_cleaned['residual sugar']<3.2,:]
data_cleaned = data_cleaned.loc[data_cleaned['chlorides']<0.10,:]
data_cleaned = data_cleaned.loc[data_cleaned['free sulfur dioxide']<35,:]
data_cleaned = data_cleaned.loc[data_cleaned['total sulfur dioxide']<90,:]
data_cleaned = data_cleaned.loc[data_cleaned['density']<1.000,:]
data_cleaned = data_cleaned.loc[data_cleaned['pH']<3.65,:]
data_cleaned = data_cleaned.loc[data_cleaned['sulphates']<0.9,:]
data_cleaned = data_cleaned.loc[data_cleaned['alcohol']<13,:]


data_cleaned = data_cleaned.loc[data_cleaned['sulphates']>0.4,:]
data_cleaned = data_cleaned.loc[data_cleaned['chlorides']>0.05,:]

data_cleaned.plot(kind="box",subplots=True,sharex=False,sharey=False,figsize=(20,10),color
='deeppink')

data_cleaned['quality'].value_counts()

data_cleaned['quality'] = data_cleaned['quality'].map({3 : 'bad', 4 :'bad', 5: 'bad',
                                6: 'good', 7: 'good', 8: 'good'})

data_cleaned['quality'].value_counts()

le = LabelEncoder()
data_cleaned['quality'] = le.fit_transform(data_cleaned['quality'])
data_cleaned['quality'].value_counts

data_cleaned['quality'].value_counts()

x = data_cleaned.iloc[:,:11]
y = data_cleaned.iloc[:,11]
print(x.shape)
print(y.shape)
```

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size = 0.3, random_state = 42)
print(x_train.shape)
print(y_train.shape)
print(x_test.shape)
print(y_test.shape)

modelsvm = SVC()
modelsvm.fit(x_train, y_train)
y_predsvmtrain=modelsvm.predict(x_train)
y_predsvmtest=modelsvm.predict(x_test)

ConfusionMatrix=confusion_matrix(y_test,y_predsvmtest)
print(ConfusionMatrix)

print('MAE= ',metrics.mean_absolute_error(y_test,y_predsvmtest))
print('MSE= ',metrics.mean_squared_error(y_test,y_predsvmtest))
print(f"r2 score: {r2_score(y_test,y_predsvmtest)}")
print('Adjusted R2 value= ',1 - (1 - (modelsvm.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,y_predsvmtrain)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_predsvmtest)))

modelnb=GaussianNB()
modelnb.fit(x_train,y_train)
y_prednbtrain=modelnb.predict(x_train)
y_prednbtest=modelnb.predict(x_test)

ConfusionMatrix=confusion_matrix(y_test,y_prednbtest)
print(ConfusionMatrix)

print('MAE= ',metrics.mean_absolute_error(y_test,y_prednbtest))
print('MSE= ',metrics.mean_squared_error(y_test,y_prednbtest))
print(f"r2 score: {r2_score(y_test,y_prednbtest)}")
print('Adjusted R2 value= ',1 - (1 - (modelnb.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,y_prednbtrain)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_prednbtest)))

modelrfc=RandomForestClassifier(n_estimators = 100)
modelrfc.fit(x_train, y_train)
y_predrfctrain=modelrfc.predict(x_train)
y_predrfctest=modelrfc.predict(x_test)

ConfusionMatrix=confusion_matrix(y_test,y_predrfctest)
print(ConfusionMatrix)

print('MAE= ',metrics.mean_absolute_error(y_test,y_predrfctest))
print('MSE= ',metrics.mean_squared_error(y_test,y_predrfctest))
print(f"r2 score: {r2_score(y_test,y_predrfctest)}")
print('Adjusted R2 value= ',1 - (1 - (modelrfc.score(x_test,y_test))) * ((756 - 1)/(756-10-1)))
print('RMSE (train)= ',np.sqrt(mean_squared_error(y_train,y_predrfctrain)))
print('RMSE (test)= ',np.sqrt(mean_squared_error(y_test,y_predrfctest)))

kf=KFold(n_splits=12)
kf

for train_index,test_index in kf.split(['fixed acidity','volatile acidity','citric acid','residual
sugar','chlorides','free sulfur dioxide','total sulfur
dioxide','density','pH','sulphates','alcohol','quality']):
    print(train_index,test_index)
def get_score(model,x_train,x_test,y_train,y_test):
    model.fit(x_train,y_train)
    return model.score(x_test,y_test)

get_score(SVC(),x_train,x_test,y_train,y_test)


get_score(GaussianNB(),x_train,x_test,y_train,y_test)


get_score(RandomForestClassifier(n_estimators =100),x_train,x_test,y_train,y_test)
```

```
import pickle
filename='model.pkl'
pickle.dump(modelrfc,open(filename,'wb'))
load_model=pickle.load(open(filename,'rb'))
```

app.py

```
import numpy as np
from flask import Flask, request, render_template
import pickle

app = Flask(_name_, template_folder='template')

model = pickle.load(open("model.pkl", "rb"))

@app.route('/')
def home():
    return render_template('about.html')

@app.route('/formsg')
def formsg():
    return render_template('formsg.html')

@app.route('/predict', methods= ["POST", "GET"])
def predict():
    float_features = [float(x) for x in request.form.values()]
    features = [np.array(float_features)]
    result = model.predict(features)
    return render_template('submit.html', result = result)

if__name__== '_main_':
    app.run(debug=True)
```

about.html

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Wine Quality Test</title>
    <link rel="stylesheet" href="{{ url_for('static',filename='css/style.css') }}">
</head>
<body>
<div class="heading">
        <h1>Wine Quality Test</h1>
    </div>
    <section class="about-us">
        <img src ="{{url_for('static', filename='wine image.jpg')}}">
        <div class="content">
            <h2>About Us</h2>
            <p>It is a wine quality prediction system.</p>
            <p>This website helps you to predict the wine quality whether it is good or bad based
on the composition values given by the user.</p>
            <p>To check the quality, click on "Start".</p>
            <button class="read-more-btn" onclick="window.location.href='{{ url_for('formsg')
}}';">Start</button>
        </div>
    </section>
</body>
</html>
```

Style.css

```
*{
            margin:0px;
            padding:0px;
            box-sizing: border-box;
            font-family: segoe ui;
        }
        body{
            background: linear-gradient(90deg, #d5eff4, #ffc7a6);
        }
        .heading{
            text-align: center;
            margin-top: 25px;
```

```
}
.heading h1{
   font-size: 45px;
   color: #36455c;
   margin-bottom: 0px;
}
.about-us{
   display: flex;
   align-items: center;
   width: 85%;
   margin: auto;
}
.about-us img{
   flex: 0 50%;
   max-width: 50%;
   height: auto;
   padding-top: 50px;
}
.content{
   padding: 35px;
   padding-top: 60px;
}
.content  h2{
   color: #36455c;
   font-size: 37px;
   margin: 15px 0px;
}
.content p{
   color: #666;
   font-size: 18px;
   line-height: 1.5;
   margin: 15px 0px;
}
.read-more-btn{
   background: #1c00b5;
   width: 100px;
   border: none;
   outline: none;
   color: #fff;
   height: 35px;
   border-radius: 30px;
   margin-top: 6px;
   cursor: pointer;
   box-shadow: 0px 5px 15px 0px rgba(28,0,181,0.3);
}
```

Formsg.html :

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="UTF-8">
    <title>Forms</title>
  <link rel="preconnect" href="https://fonts.googleapis.com">
<link rel="preconnect" href="https://fonts.gstatic.com" crossorigin>
<link
href="https://fonts.googleapis.com/css2?family=Poppins:wght@400;600;800&display=swap"
rel="stylesheet">
  <link rel="stylesheet" href="{{ url_for('static',filename='css/formstyle.css') }}">
</head>
<body>
<div class="container">
  <h1>Wine Quality Test</h1>
  <p>Predict the quality of your wine sample</p>
  <div class="contact-box">
    <div class="contact-left">
      <h3>Send your composition values</h3>
      <form action='/predict' method="POST">
        <div class="input-row">
```

```
          <div class="input-group">
            <label>fixed acidity</label>
            <input type="number" name="fixed_acidity" placeholder="Enter Fixed Acidity value"
max="16" min="4" step="0.01" required="required">
          </div>
          <div class="input-group">
            <label>volatile acidity</label>
            <input type="number" name="volatile_acidity" placeholder="Enter Volatile Acidity
value" max="1" min="0.3" step="0.001" required="required">
          </div>
        </div>
        <div class="input-row">
          <div class="input-group">
            <label>citric acid</label>
            <input type="number" name="citric_acid" placeholder="Enter Citric Acid value"
max="1" min="0" step="0.01" required="required">
          </div>
          <div class="input-group">
            <label>residual sugar</label>
            <input type="number" name="residual_sugar" placeholder="Enter Residual  Sugar
value" max="7" min="1" step="0.1" required="required">
          </div>
        </div>
        <div class="input-row">
          <div class="input-group">
            <label>chlorides</label>
            <input type="number" name="chlorides" placeholder="Enter Chlorides value"
max="1.000" min="0.000" step="0.001" required="required">
          </div>
          <div class="input-group">
            <label>free sulfur dioxide</label>
            <input type="number" name="free_sulfur_dioxide" placeholder="Enter Free Sulfur
Dioxide value" max="72" min="1" step="0.1" required="required">
          </div>
```

```
    </div>
    <div class="input-row">
      <div class="input-group">
        <label>total sulfur dioxide</label>
        <input type="number" name="total_sulfur_dioxide" placeholder="Enter Total Sulfur Dioxide value" max="289" min="6" step="0.1" required="required">
      </div>
      <div class="input-group">
        <label>density</label>
        <input type="number" name="density" placeholder="Enter Density value" max="1.00000" min="0.99000" step="0.00001" required="required">
      </div>
    </div>
    <div class="input-row">
      <div class="input-group">
        <label>pH</label>
        <input type="number" name="pH" placeholder="Enter pH value" max="4.01" min="2.7" step="0.01" required="required">
      </div>
      <div class="input-group">
        <label>sulphates</label>
        <input type="number" name="sulphates" placeholder="Enter Sulphates value" max="2" min="0.33" step="0.01" required="required">
      </div>
    </div>
    <div class="input-row">
      <div class="input-group">
        <label>alcohol</label>
        <input type="number" name="alcohol" placeholder="Enter Alcohol value" max="14.9" min="8.4" step="0.1" required="required">
      </div>
    </div>
    <button type="submit">Submit</button>
  </form>
```

```
    </div>
  </div>
</div>
</body>
</html>
```

Formstyle.css :

```
*{
    margin: 0px;
    padding: 0px;
    }
    body{
    background: linear-gradient(135deg, #71b7e6, #9b59b6);
    font-size: 14px;
    font-family: 'Poppins', sans-serif;
    }
    .container{
    width: 59%;
    margin: 30px auto;
    }
    .contact-box{
    background: #fff;
    display: flex;
    }
    .contact-left{
    flex-basis: 84%;
    padding: 9px 55px;
    }
    h1{
    margin-bottom: 2px;
    }
    .container p{
    margin-bottom: 5px;
    }
```

```
p{
font-size: 18px;
}
.input-row{
display: flex;
justify-content: space-between;
margin-bottom: 10px;
}
.input-row .input-group{
flex-basis: 45%;
}
input{
width: 85%;
border: none;
border-bottom: 1px solid #ccc;
outline: none;
padding-bottom: 0px;
}
label{
margin-bottom: 6px;
display: block;
font-size: 17px;
color: #1c00b5;
}
button{
background: #1c00b5;
width: 100px;
border: none;
outline: none;
color: #fff;
height: 35px;
border-radius: 30px;
margin-top: 10px;
margin-bottom: 15px;
```

```
    cursor: pointer;
    box-shadow: 0px 5px 15px 0px rgba(28,0,181,0.3);
    }
    .contact-left h3{
    color: #1c00b5;
    font-weight: 600;
    padding-top: 15px;
    margin-bottom: 20px;
    }
```

Submit.html :

```
<!DOCTYPE html>
<html>
<head>
    <link rel="preconnect" href="https://fonts.googleapis.com">
<link rel="preconnect" href="https://fonts.gstatic.com" crossorigin>
<link
href="https://fonts.googleapis.com/css2?family=Poppins:wght@400;600;800&display=swap"
rel="stylesheet">
<link rel="stylesheet" href="{{ url_for('static',filename='css/submission.css') }}">
</head>
<body>
<h1> Prediction of your sample</h1>
{% if result == 1 %}
<img src="{{url_for('static', filename='good.png')}}" class="center">
{% else %}
<img src="{{url_for('static', filename='bad.png')}}" class="center">
{% endif %}
</body>
</html>
```

Submission.css :

```
body{
background: linear-gradient(90deg, #d0ffae, #34ebe9);
}
h1{
text-align: center;
font-size: 33px;
font-family: 'Poppins', sans-serif;
margin-top: 60px;
color: black;
}
.center {
  display: block;
  margin-left: auto;
  margin-right: auto;
  width: 28%;
  padding-top: 35px;
}
```

3.12 RESULT ANALYSIS

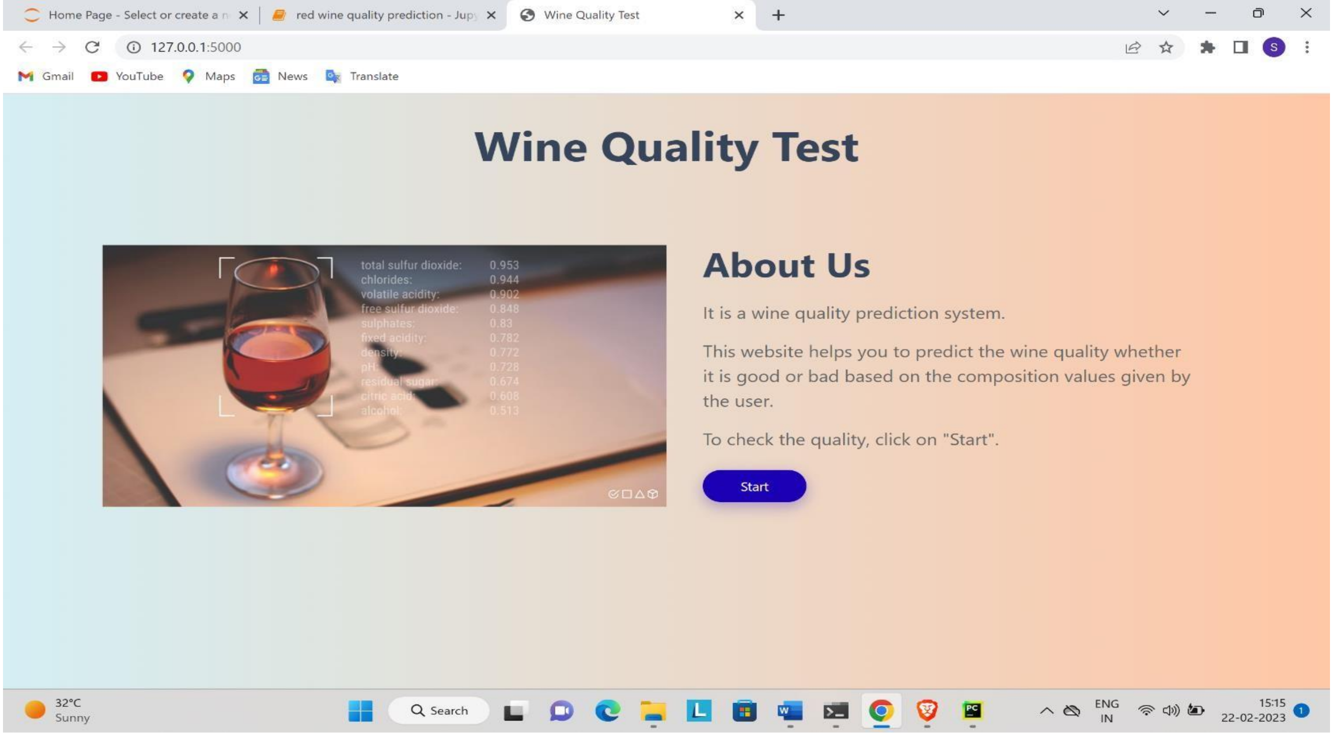
The performance of the several regressors were analysed in order to pick the most effective fundamental algorithm and establish the correctness and reliability of the model's deployment. The efficiency of the regressor can be explained using the estimation coefficient.

MODEL	ACCURACY
Support Vector Machine	56.8%
Naïve Bayes Algorithm	71.7%
Random Forest Classifier	80.2%

Table 3.12. Comparison of models

Comparing the models used in Table 2 for predicting the quality revealed that the random forest is larger. With an accuracy of 80.2, the Random Forest Classifier model was determined to be the best model.

4. OUTPUT SCREENS



Wine Quality Test

total sulfur dioxide:	0.953
chlorides:	0.944
volatile acidity:	0.902
free sulfur dioxide:	0.848
sulphates:	0.83
fixed acidity:	0.782
density:	0.772
pH:	0.728
residual sugar:	0.674
citric acid:	0.608
alcohol:	0.513

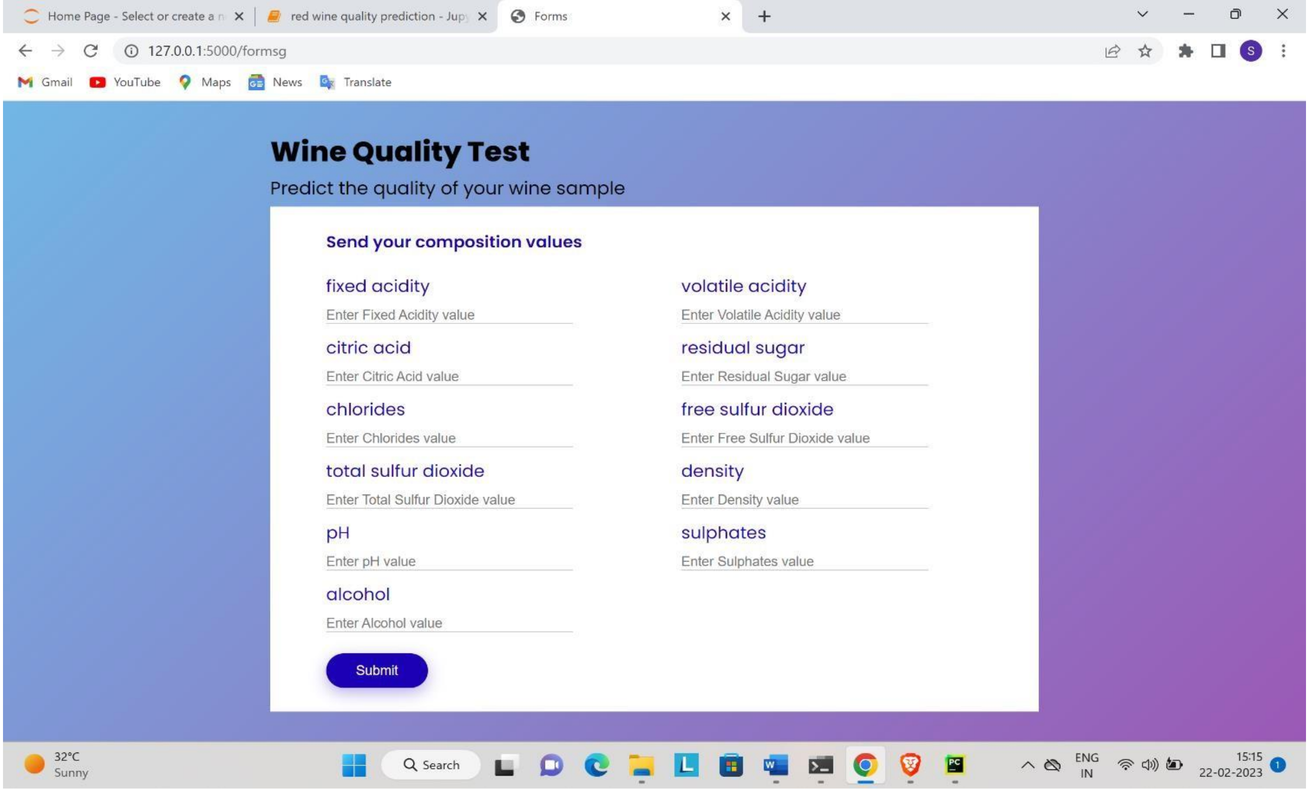
About Us

It is a wine quality prediction system.

This website helps you to predict the wine quality whether it is good or bad based on the composition values given by the user.

To check the quality, click on "Start".

Figure: 4.1. Home Screen



Wine Quality Test

Predict the quality of your wine sample

Send your composition values

fixed acidity
Enter Fixed Acidity value

volatile acidity
Enter Volatile Acidity value

citric acid
Enter Citric Acid value

residual sugar
Enter Residual Sugar value

chlorides
Enter Chlorides value

free sulfur dioxide
Enter Free Sulfur Dioxide value

total sulfur dioxide
Enter Total Sulfur Dioxide value

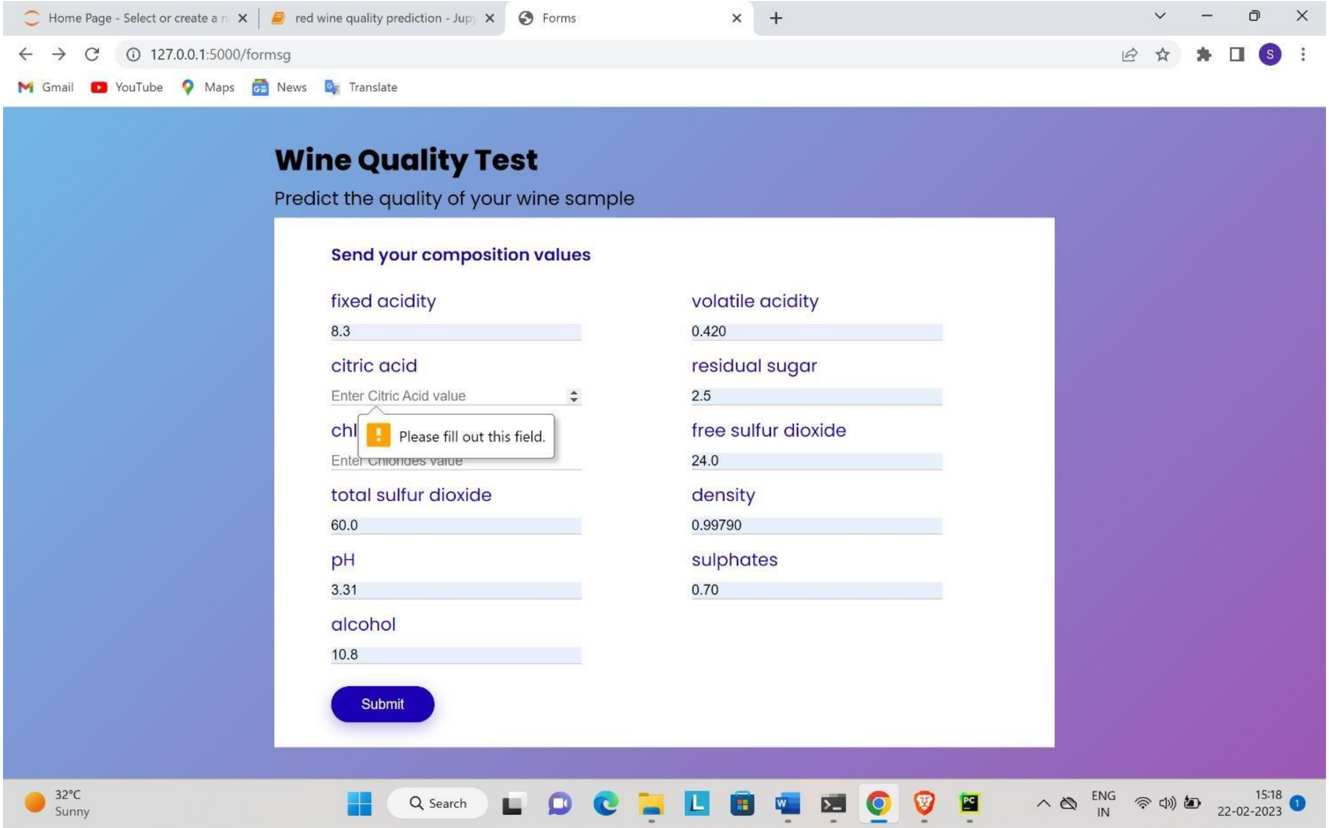
density
Enter Density value

pH
Enter pH value

sulphates
Enter Sulphates value

alcohol
Enter Alcohol value

Figure 4.2 : Prediction Form



Wine Quality Test

Predict the quality of your wine sample

Send your composition values

fixed acidity: 8.3

volatile acidity: 0.420

citric acid: Enter Citric Acid value

residual sugar: 2.5

chl[orides]: Enter Chlorides value

Please fill out this field.

free sulfur dioxide: 24.0

total sulfur dioxide: 60.0

density: 0.99790

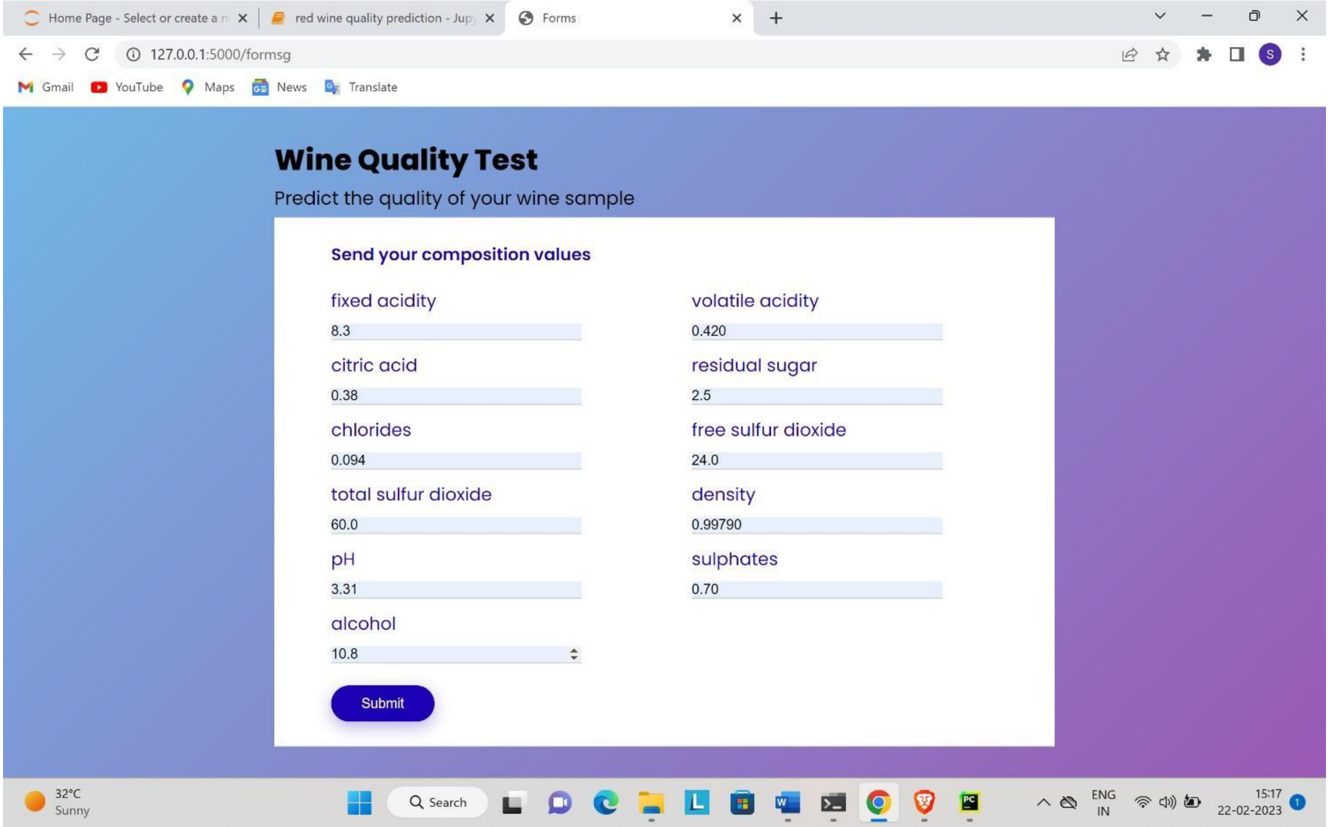
pH: 3.31

sulphates: 0.70

alcohol: 10.8

Submit

Figure 4.3 : Missing Field Validation



Wine Quality Test

Predict the quality of your wine sample

Send your composition values

fixed acidity: 8.3

volatile acidity: 0.420

citric acid: 0.38

residual sugar: 2.5

chlorides: 0.094

free sulfur dioxide: 24.0

total sulfur dioxide: 60.0

density: 0.99790

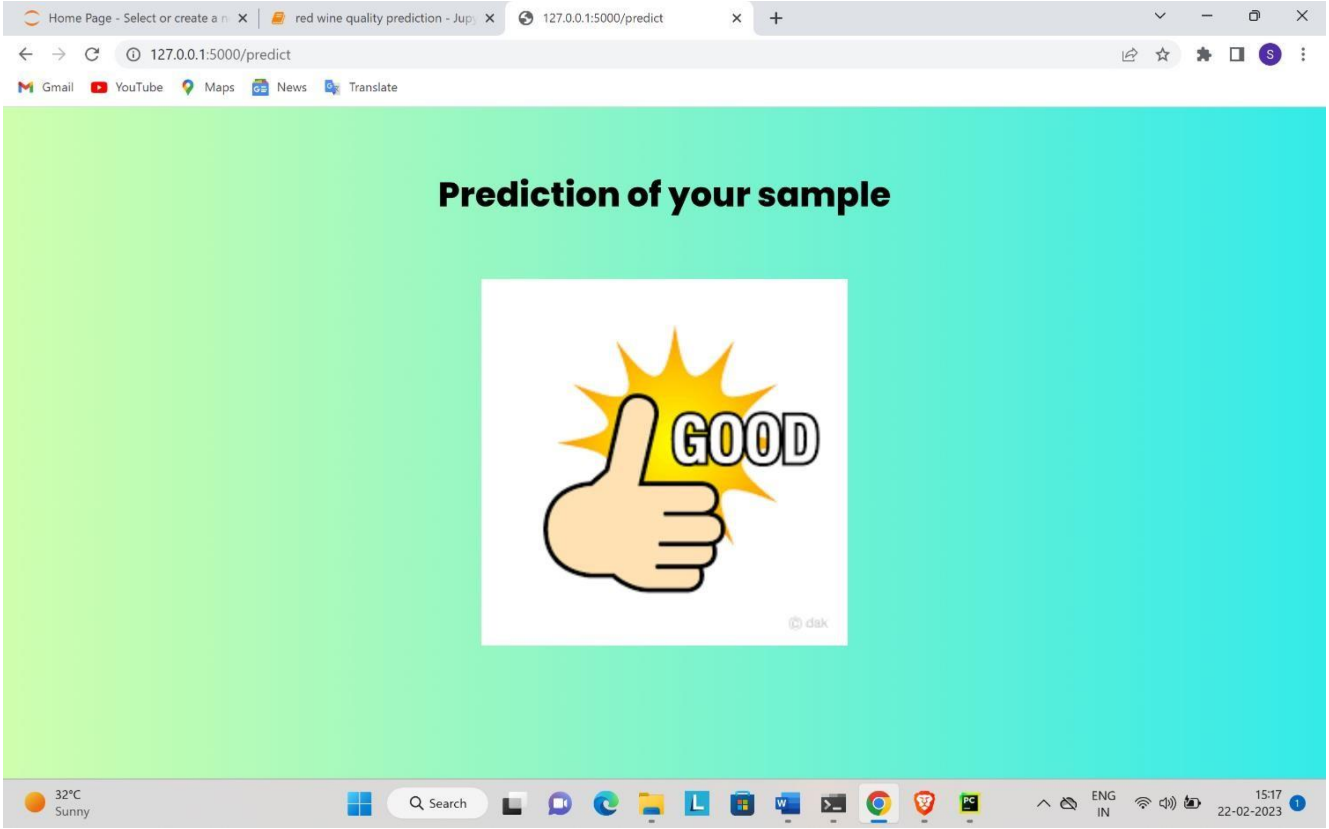
pH: 3.31

sulphates: 0.70

alcohol: 10.8

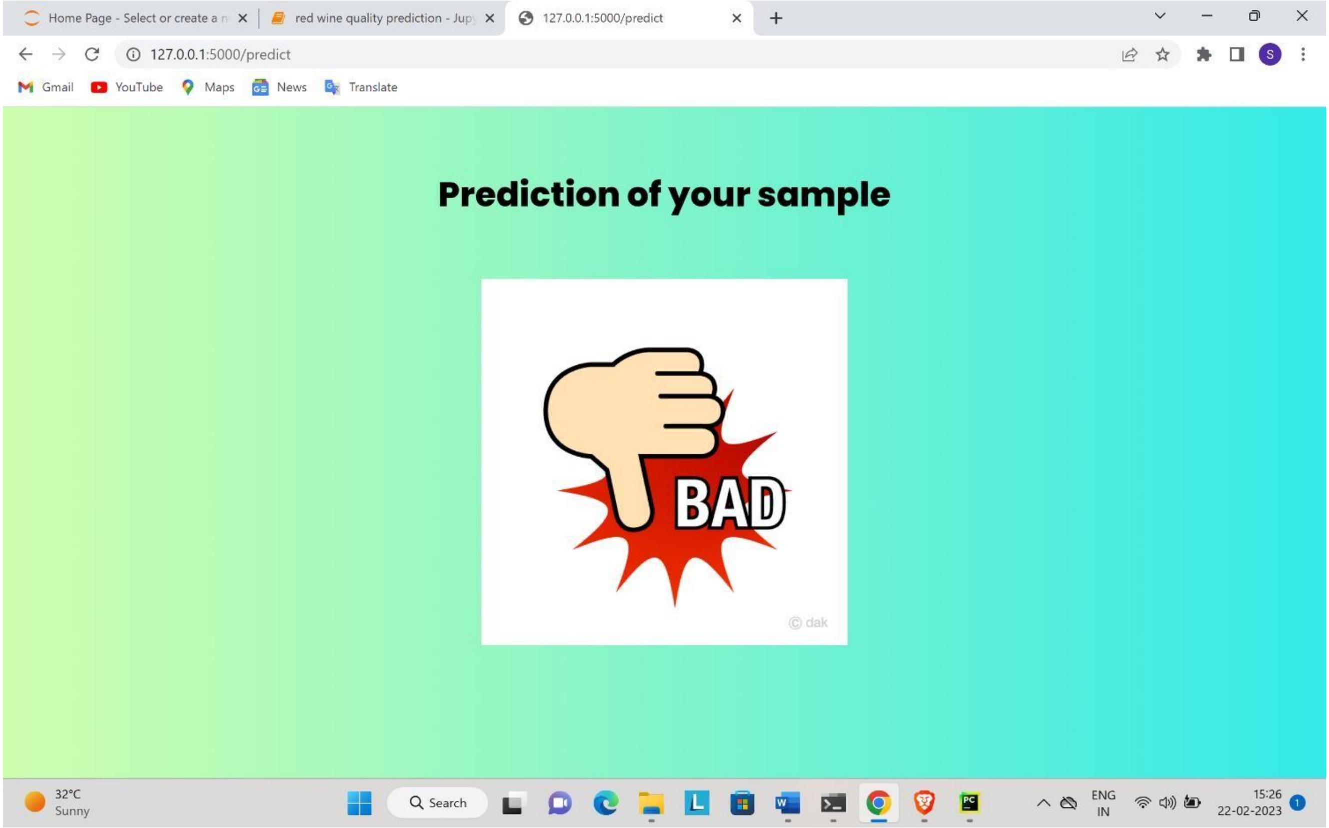
Submit

Figure 4.4 : Inserting Values in the form



Prediction of your sample

Figure 4.5 : Output Screen Good Quality



Prediction of your sample

Figure 4.6 : Output Screen Bad Quality

5. CONCLUSION AND FUTURE SCOPE

5.1 Conclusion

The goal of using machine learning methods was the study's goal to predict whether a red wine will be good or awful. The analysis revealed a considerable improvement in the performance of the models, and we found that, Compared to the support vector machine and naive bayes methods, the random forest classifier has a greater accuracy. We selected a random forest classifier model because our goal was to forecast the quality of red wine.

5.2 Future Scope

Future research, however, can focus on exploring a number of additional deep learning applications. The technique that can converge the changes and do multiple frame work can be improved using improved approaches from machine learning and other fields.

6. BIBLIOGRAPHY

[1] P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, “Modelling wine preferences by data mining from physicochemical properties,” In Decision Support Systems, Elsevier, 47 (4): 547-553. ISSN: 0167-9236.

[2] S. Ebeler, “Linking Flavour Chemistry to Sensory Analysis of Wine,” in Flavor Chemistry, Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409-422.

[3] A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[4] S. Kallithraka, IS. Arvanitoyannis, P. Kefalas, A. El-Zajouli, E. Soufleros, and E. Psarra, “Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin,” Food Chemistry, 73(4): 501-514, 2001.

[5] N. H. Beltran, M. A. Duarte- MErmound, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, “Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer,” Instrum. Measurement, IEEE Trans., 57: 2421-2436, 2008.

[6] S. Shanmuganathan, P. Sallis, and A. Narayanan, “Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality,” IEEE International Conference on Computational Intelligence Communication Systems and Networks, pp. 82-89, July 2010.

[7] B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, “Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel,” IEEE International Conference on Data Mining Workshop, pp. 142-149, Dec. 2014.

[8] K. Agrawal and H. Mohan, "Cardiotocography Analysis for Fetal State Classification Using Machine Learning Algorithms," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019, pp. 1-6.

[9] K. Agrawal and H. Mohan, "Text Analysis: Techniques, Applications and Challenges," presented in 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019.

[10] J. Han, M. Kamber, and J. Pei, “Classification: Advanced Methods,” in Data Mining Concepts and Techniques, 3rd ed., Waltham, MA, USA: Morgan Kaufmann, 2012, pp. 393-443.

Red Wine Quality Prediction Using Machine Learning

Ch. Sai Sri Ram[1], J. Avinash[2], K. Raghu Ram Sri Rishik[3], V. Narendra Reddy[4], A. Thanuja[5]

[1,2,3,4]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India
[5]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India
saisriramchunduri@gmail.com[1], jagarlamudirakesh2@gmail.com[2], rishik12457@gmail.com[3], vennan19@gmail.com[4], a.thanuja18@gmail.com[5]

1. ABSTRACT- The goal of this work was to create a model to forecast red wine quality based on its physicochemical characteristics. Several factors influence the accuracy of prediction while analysing the quality of red wine. This paper offers a computational intelligence method using machine learning techniques. In this instance, Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine were used. With this information and these machine learning techniques, we can forecast the quality of a sample of red wine.

2. KEYWORDS: Red wine, Naive Bayes algorithm, Support vector machine, quality prediction, and Random forest classifier.

3. INTRODUCTION

Machine learning, a fast growing field of artificial intelligence, allows computers to automatically learn from experience and improve over time without explicit programming. Machine learning allows computers to examine massive volumes of data, spot patterns and trends, and then use that information to predict the future or make decisions. There are many useful uses for machine learning speech recognition, and personalised recommendations. Machine learning is anticipated to have a significant impact on many businesses and facets of daily life as it develops.

Red wine quality prediction using machine learning seeks to increase the precision and effectiveness of processing.

Machine learning models can be trained to recognise patterns and forecast the likelihood that a claim will be approved or refused by utilising historical data and prediction algorithms.

Machine learning, a cutting-edge field of research, enables computers to learn on their own using past data.

In order to build mathematical models and generate predictions based on previously collected data or information, machine learning employs a range of methodologies.

The purpose of utilising machine learning to forecast wine quality is to increase the precision and effectiveness of processing. Some machine learning software packages that can be used to create this system.

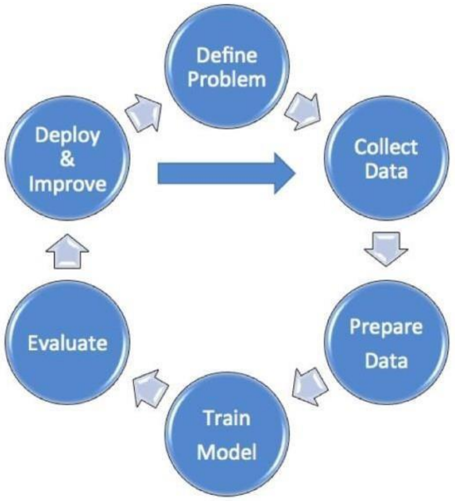


Fig.1 5 steps involved in Model

4. LITERATURE SURVEY

Existing literature was examined in order to gain the necessary knowledge on numerous ideas linked to the current use of our model. Through those, some of the most significant conclusions were drawn.

The research on estimating the quality of red wines utilising a variety of methodologies, including chemical analysis, machine learning, and sensory evaluation. It would go over each method's benefits and drawbacks and point out any gaps in the existing body of knowledge.

A description of red wine's chemical makeup and the elements that affect its quality. It would go over how different substances, including acids, sugars, and phenolics, affect the flavour, fragrance, and colour of red wine.

The numerous machine learning techniques that have been applied to forecast depends on the quality of red wine its physicochemical characteristics. It would emphasise the most important performance measures used to assess each algorithm's performance and go over its advantages and disadvantages.

The techniques used to choose the features that are most useful for forecasting the quality of red wine and to adjust the model's parameters for the greatest performance. It would also go over the difficulties and restrictions of these techniques and make recommendations for the future.

In this study, we attempt to forecast the quality of a sample of red wine using data from the dataset.

The model is trained using the machine learning algorithms Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine, where we have achieved accuracy up to 80% using RFC, 72% with Naive Bayes, and nearly 57% with Support Vector Machine.

5. MATERIALS AND METHODOLOGY

Our model is suggested based on the following characteristics.

5.1 Dataset Analysis: We downloaded the datasets from the Kaggle website on the internet, and a dataset of red wine quality is a must if we are to make any predictions.

The dataset includes 12 columns: citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol, quality, fixed acidity, and volatile acidity.

Column	Description
fixed acidity	Sample's fixed acidity
volatile acidity	Sample's volatile acidity
citric acid	Sample's citric acid
residual sugar	Sample's residual sugar
chlorides	Sample's chloride
free sulfur dioxide	Sample's free sulfur dioxide
total sulfur dioxide	Sample's total sulfur dioxide
density	Sample's density
pH	Sample's pH
sulphates	Sample's sulphates
alcohol	Sample's alcohol
quality	Sample's quality

5.1. Data set

To better comprehend the features, we created a graphical representation of the dataset.

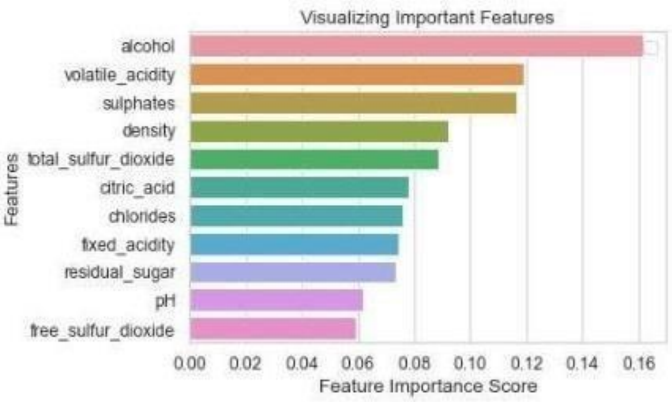


Fig 5.2. Features Visualizations

The dataset above shows several quality values within the specified range.

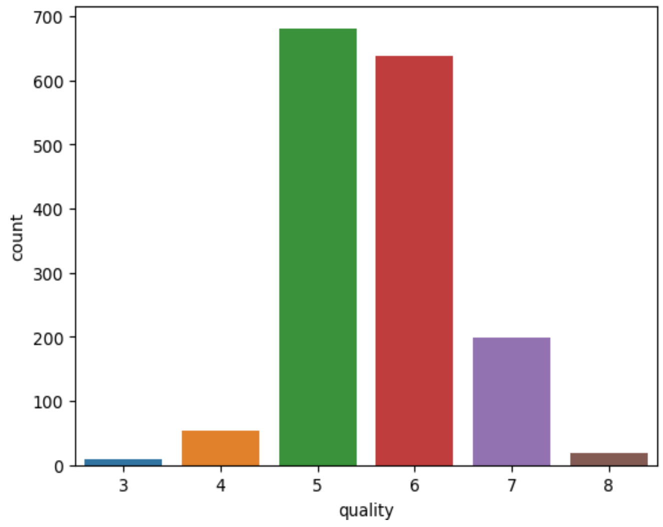


Fig 5.3. Quality

The quality values across the ranges are depicted in the above diagram.

Making adjustments to our data before submitting it to the algorithm is known as pre-processing. a procedure for turning raw data into uncleaned data. In comparable, whenever data are gathered from various sources, they are collected in raw format, making analysis impossible. Data must be in a specific format because of a particular machine learning algorithm. For instance, null values must be handled from the original raw data set in order to apply the Random Forest method because the Random Forest algorithm does not allow null values. A data set should be organised so that many algorithms for machine learning can be utilised to achieve the best outcomes.

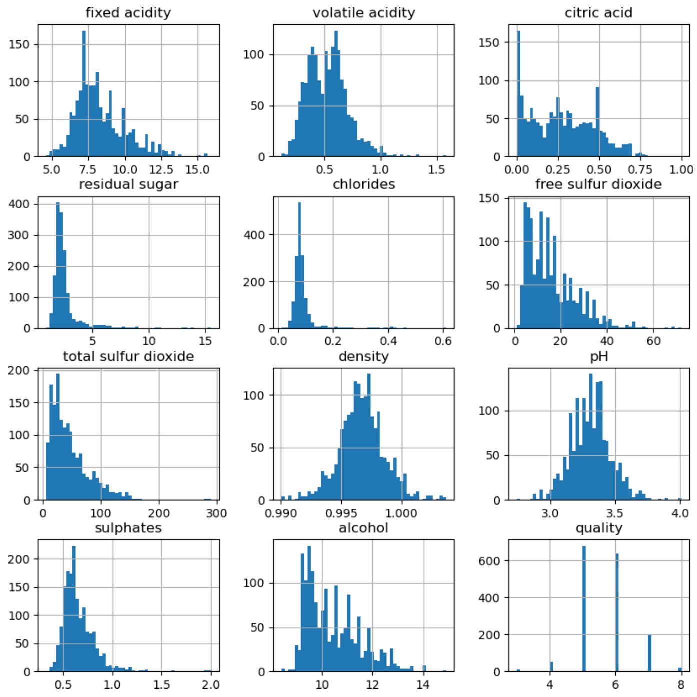


Fig 5.4. Histogram Graph

Eventually, it is discovered that the data contains some outliers. For a clear understanding of the outliers for each column, we employed a boxplot. After utilising percentile to remove the irrelevant data, a cleaned dataset was produced.

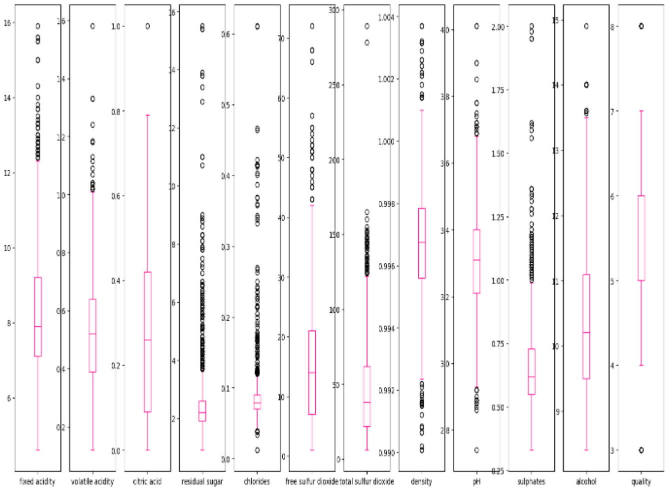


Fig. 5.5 Outliers Detection

We separated the quality feature values from the cleaned datasets into excellent and bad. To convert the excellent and bad values into binary values of 0 and 1, we used LabelEncoder(). Good is found to be transformed to the binary value 1 and bad to the binary value 0.

We divided the dataset for conducting the models into 70% for training and 30% for testing, and we removed the quality attribute because it is thought of as a goal variable.

Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine are the tools we're employing for this. For training, a random forest classifier model is employed.

We are measuring both the accuracy of the model and the projected result using the sklearn accuracy score.

The Support Vector Machine and Naive Bayes Algorithm were also used in the same way.

These are the outcomes:

56.8% of support vector machines.

Algorithm of Naive Bayes, 71.7%

80.2% for Random Forest Classifier

Due to this accuracies, we discovered that Random Forest Classifier delivers the best among other two models . Therefore, we believed that this model was the most accurate for assessing the red wine sample's quality.

6. CONCLUSION AND FUTURE SCOPE

The goal of using machine learning methods was the study's goal to predict whether a red wine willbe good or awful. The analysis revealed a considerable improvement in the performance of the models, and we found that, Compared to the support vector machine and naive bayes methods,the random forest classifier has a greater accuracy. We selected a random forest classifier model because our goal was to forecast the quality of redwine.

Future research, however, can focus on exploring a number of additional deep learning applications.The technique that can converge the changes and do multiple frame work can be improved using improved approaches from machine learning and other fields.

7. REFERENCES

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, &José Reis. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

[2] Edelmann, Andrea , et al. "Rapid Method for the Discrimination of Red Wine Cultivars Based on Mid- Infrared Spectroscopy of Phenolic Wine Extracts." Journal of Agricultural & Food Chemistry49.3(2001):1139-1145.

[3] Zhang Shiling, Xu Ruimin. Formation and prevention of volatile acidin wine. New Rural Technology, 2008 (06): 81-82.

[4] Dahal, K., Dahal, J., Banjade, H., Gaire, S., 2021. Prediction of Wine Quality Using Machine Learning Algorithms. Open J. Stat. 11, 278–289.

[5] Rish, I., 2001. An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell 3.

[6] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) “Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks”. Talanta 72 263–268.

BB-2

ORIGINALITY REPORT

7%	3%	5%	2%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	"Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems", Springer Science and Business Media LLC, 2022 Publication	2%
2	S. Siva Sunayna, S. N. Thirumala Rao, M. Sireesha. "Chapter 25 Performance Evaluation of Machine Learning Algorithms to Predict Breast Cancer", Springer Science and Business Media LLC, 2022 Publication	2%
3	Submitted to University of Winchester Student Paper	2%
4	www.diva-portal.org Internet Source	1%
5	docs.google.com Internet Source	1%