International Journal of Progressive Sciences and Technologies (IJPSAT)
ISSN: 2509-0119.

Vol. 44 No. 2 May 2024, pp. 372-403

A Supervised Machine Learning Classification Framework for Beverage Quality Prediction

[1,]Jules MUHAYIMANA,[2]Dr Leopord Hakizimana

[1]Candidate in MIT, Graduate school, University of Kigali
[2]Lecturer, Graduate School, University of Kigali
* E-mail of the corresponding author: muhayejules@gmail.com

Abstract – Since the production of food and beverages is energy-intensive, the quality of food and beverage is important for the consumers as well as the food and beverage industry, the economic, political and social condition are posing challenge to Food and beverage small and medium and large industries assessment is an evaluation method used to measure the strengths and weaknesses of a food and beverage system to make improvements. With the start-up business success help of a machine learning model and several features of beverages, this thesis would focus on important features that affect the quality of beverage production and have a model to predict a beverage quality. This review would also compare and discuss each technique and provide suggestions based on the current technology. This review would deliberate technology integration and the involvement of deep learning to enable several types of current technologies and the results demonstrate the model's ability to accurately predict beverage quality based on chemical composition. Furthermore, the developed model allows for the identification of critical chemical parameters influencing beverage quality. Manufacturers can use this information to make targeted adjustments in the formulation and production process, leading to enhanced product quality and consistency.

Keywords – Machine Learning, Classification, Beverage and Prediction.

I. INTRODUCTION

In this section introduction of the study, the background of the study, the problem statement, the objectives of the study, the scope and limitation of the research, and the significance of the study are all outlined.

1.1. Introduction

Over the four decades we've been in business, we've seen beverage change and evolve with consumer demands and beverage trends. The beverage industries include all industries involved in processing raw beverage materials, similarly to those who package and distribute them. This includes fresh, prepared beverages as well as packaged beverages (Coles, 2011).

Machine learning (ML) refers to an application of AI that provides the system which is able to automatically learn and enhance through practical knowledge instead of relying on direct programming. This is too simple because a large amount of data today is available which makes it easier to machines to be trained rather than programmed. It is deemed a significant technological breakthrough capable of scrutinizing vast volumes of data. ML is rapidly transforming the world by changing all segments including healthcare services, transport, food, education, and different assembly line and many more (Singh, 2020). Machine learning is a swiftly expanding domain with limitless potential applications. Over the upcoming years, we anticipate witnessing machine learning revolutionize numerous sectors, such as manufacturing, retail, and healthcare. In manufacturing, machine learning can be used for quality control, automation and customization (Ambadipudi, 2023)

The food and beverages manufacturing and processing create significant hazards related to the risk of fire and exposure to toxic gases. Therefore, the development of food and beverage assessments is essential and should be a priority to ensure food and beverage safety and public health. Numerous methodological processes and technologies have been created for the evaluation of food and beverages, encompassing imaging, odor, taste, electromagnetic sensing, and various other approaches. (Alabi, 2020).

According to Georgia Pratt, the beverage industry is sub-divided into two segments, those are the production and the distribution of the finished products. The first group, production, includes the processing of creation of soft drinks, alcoholic beverages and other modified beverages. Every item intended for human consumption, excluding pharmaceuticals, goes through this industry. Production also covers the local production, dairy and alcoholic beverages. The manufacturing sector does not include beverages and fresh produce directly cultivated through farming, as these are categorized under agriculture (Georgia Pratt, 2022).

According to the Rwanda Standards Board (RSB) notice, it plans to remove substandard processed food products and beverages from the Rwandan market. The primary regulatory authority overseeing various food products and beverages in Rwanda, the Rwanda Standards Board, has instructed manufacturers and processors of all types of food items to register their products with its quality assurance unit. Failure to register locally manufactured food and beverage stuff would attract penalties (FoodStuff, 2016).

1.2. Problem statement

In the beverages industry high-quality product is essential; in contrast, most of the beverage companies fail to guarantee the quality of products for consumers. There is sequence of assessments that can be put in place to ensure beverage like monitoring the quality of raw material, production processes, sensory analysis, and packaging examinations. Examining the nutritional content and ingredient quality of beverages is crucial for ensuring the safety and quality of the products (Aadil, 2019).

Currently, safety of beverages and food has been a universal health issues of global concern, especially in food and beverage hygiene and their quality (Tan, 2023). In the global market for food products, quality has emerged as a crucial distinguishing factor for competition. To have the best quality of end product, Quality is increasingly being overseen along the whole food chain from the raw material's suppliers to the consumption. In this study should come up with a model or framework that would predict the beverage quality.

1.3. Research Objectives

1.3.1. General objective

The general objective of this thesis is to develop a supervised machine learning classification framework for beverage quality prediction.

1.3.2. Specific objectives

To ensure that the general objective of the study is reached, the following objectives are formulated:

1. To determine different existing machine learning algorithms to help in generation of beverage quality prediction model
2. To create an AI system that can detect potential safety hazards and hygiene issues in beverage preparation and handling processes.
3. To implement an AI system that can analyze data from beverage production to predict and maintain the quality.

II. LITERATURE REVIEW

Introduction

This chapter presents a review of theoretical framework concepts and fundamental definitions based and used in this research. Its main objective is to provide useful information and an overview of theories and concepts that lead to the development of an Artificial Intelligent based system "A Framework for Beverage Quality Prediction Using Assembling of Machine Learning Classifiers". Besides that, also emphasizing the software and languages that are used. It included definitions of some keywords, gaps, weaknesses, immediate relevance and also an overview of this research topic.

A clear conceptual framework was developed and presented in the last section of this chapter. A conceptual framework reflects the relations among concepts or variables that the researcher analyzed to achieve the stated objectives of the study.

2.1. Conceptual Review

2.1.1. Beverage

Beverages means alcoholic drinks and non-alcoholic drinks and other potable liquids intended for human consumption, including beer, wine, soft drinks, fruit juices, milk, liquid dietary supplements and packaged or bottled water but excluding products that constitute Pharmaceuticals (Kalpana, 2019).

2.1.2. Prediction

Machine learning prediction, or prediction in machine learning, refers to the output of an algorithm that has been trained on a historical dataset. The algorithm proceeds to produce likely values for unknown variables in each record of the new data. In machine learning, the goal of prediction is to create an anticipated dataset that correlates with the initial data. Generally speaking, to anticipate is to foresee with accuracy using knowledge, math, or astute deduction from data or experience.

2.1.3. Machine learning

Machine learning is an application of artificial intelligence that uses statistical techniques to enable computers to learn and make decisions without being explicitly programmed (Ed Burns, 2021). It is based on the concept that computers can learn from data, recognize patterns, and make decisions with minimal human intervention. It is a subset of Artificial Intelligence. It is the study of giving robots the capacity to learn and create their own programs in order to make them behave and make decisions more like humans. This is done with minimum human intervention, i.e., no explicit programming. The process of it uses to learn is automated and improved refering on the experiences of the machines throughout the process.

2.1.4. Quality

Product Quality refers to how well an item meets regulatory requirements and standards. It gives consumers peace of mind that products are safe for use. Having such a system is proof of a brand's commitment to tangible quality and is a key element in converting satisfied customers into loyal followers.

2.1.5. A classifier in machine learning

Classifiers are algorithms in machine learning that are used to categories data into distinct groups or classes. They are essentially mathematical models that use statistical analysis and optimization to identify patterns in the data. These patterns help to determine the class or category that each instance belongs to.

2.2. Machine Learning methods

2.2.1. Supervised machine learning

Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is characterized by the way it trains algorithms to accurately classify data or predict outcomes using labeled datasets. The model refines its weights as input data is provided, ensuring proper fitting, and this adjustment is a crucial step within the cross-validation process (Gillis, 2023).

2.2.1.1. Types of Supervised machine learning

Classification: Utilizing an algorithm, it precisely categorizes test data into distinct categories, identifying specific entities in the dataset and striving to make informed conclusions about how those entities should be labeled or defined.

Regression: Is employed to comprehend the connection between dependent and independent variables, often applied to make predictions, such as estimating sales revenue for a particular business.

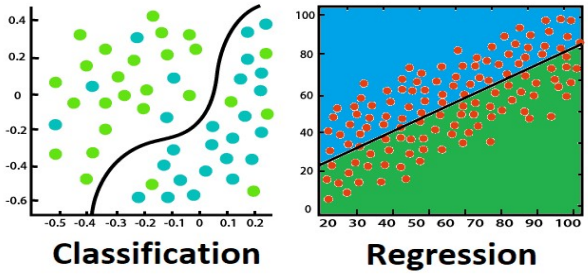


Figure 1: Supervised machine learning

Source: (datacamp, 2019)

2.2.2. Unsupervised Learning

Unsupervised learning is a kind of machine learning (ML) approach that finds patterns in data sets that are neither labeled nor classed by using artificial intelligence (AI) algorithms. unsupervised methods learn concise representations of the data input, that can be used for data exploration or to analyze or generate new data (Pykes, 2023).

2.2.3. Semi-Supervised Learning

Semi-supervised learning is supervised learning with a small number of labeled instances and a large number of unlabeled examples in the training data. In contrast to supervised learning, the purpose of a semi-supervised learning model is to make good use of all available data rather than just the labeled data. Unsupervised and supervised learning techniques are used in semi-supervised learning.

2.2.4. Reinforcement learning

Reinforcement Learning (RL) is the study of decision-making, focusing on acquiring the most effective behavior in an environment to achieve the highest possible reward. In RL, the data is accumulated from machine learning systems that use a trial-and-error method (Bhatt, 2018). Data is not part of the input that we would find in supervised or unsupervised machine learning.

2.3. Machine learning algorithms(classifiers)

2.3.1. Naive Bayes

A collection of supervised learning algorithms known as "naive Bayes methods" are based on the application of Bayes' theorem under the "naive" assumption that every pair of features would have conditional independence given the value of the class variable.

The Naïve Bayes algorithm is comprised of two words Naïve and Bayes, which can be described as:

Naïve: It is termed "Naïve" because it presupposes that the presence of a specific feature is unrelated to the presence of other features.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

2.3.2. Support Vector Machine

A collection of supervised learning algorithms known as "naive Bayes methods" are based on the application of Bayes' theorem under the "naive" assumption that every pair of features would have conditional independence given the value of the class variable. This hyperplane is known as the decision boundary, separating the classes of data points (e.g., oranges vs. apples) on either side of the plane (Gandhi, 2018). Support Vector Machines have been widely used in various applications, including image classification, text classification, handwriting recognition, and bioinformatics, among others. Their effectiveness, especially in high-dimensional spaces, and the ability to handle non-linear relationships through kernel tricks make SVMs a popular choice in machine learning

2.3.3. Random forest

Random forest stands as a versatile supervised machine learning algorithm employed for classification and regression tasks. The term "forest" denotes an assembly of independent decision trees that are subsequently amalgamated to diminish variance and enhance the precision of data predictions (Mbaabu, 2020).

The model also overcomes the over fitting problem usually faced by decision trees as it consist of multiple trees from which a random choice is made. Overfitting happens when the algorithm models the training data too well i.e., it learns the detail and noise in the training data as concepts of the model hence impacting the model's ability to generalize and model new data Invalid source specified. Random forests can also handle null values, which in our case is not a problem as we treated all missing values as shown in the previous chapters. Our data set consists of multiple categorical values which the model is well known for handling properly.

i. Modeling pseudo code:

1. Randomly choose a number of variables 'a' from total variables 'b'.

   Where a << b.

2. From the chosen variables, calculate the node "d" using the best separation point.
3. Separate the node into daughter nodes using the best separation point.
4. Repeat 1 to 3 stages until only one node is reached.
5. Build forest model by recapping stages 1 to 4 for "x" number of times to generate "x" number of trees.

ii. Prediction pseudo code:

1. Utilize the test variables and the rules of each randomly generated decision tree to predict the outcome and store the result.
2. Compute the votes for every predicted target.
3. Take the high voted predicted target as the final likelihood or prediction from the random forest model.

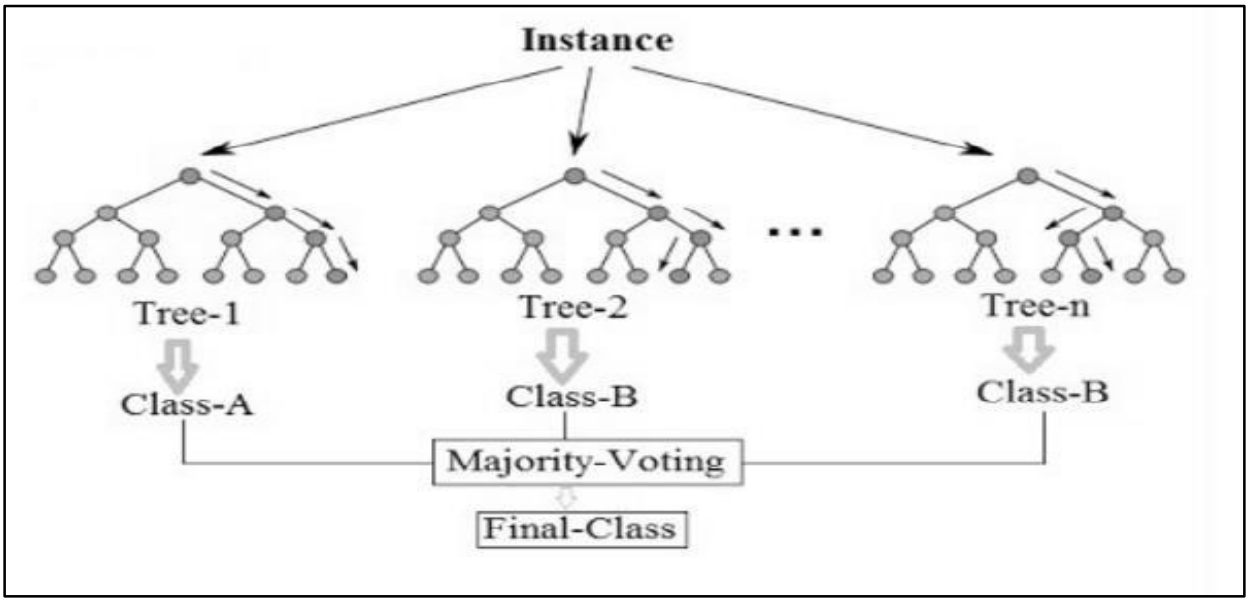


Figure 2: Random Forest

2.3.4. K-Nearest Neighbours

K nearest neighbors is a straightforward method that maintains all existing examples and categorizes new ones using a similarity metric (e.g., distance functions). According to this technique, comparable data points can be located close to one another. Therefore, it aims to compute the distance between data points, typically using Euclidean distance, and subsequently assigns a category based on the most prevalent category or the average. Data scientists appreciate this technique because of its simplicity and quick computation time, however as the test dataset gets larger, the processing time increases, making it less useful for classification tasks. KNN is typically used for recommendation engines and image recognition.

2.3.5. Detailed Algorithm design for beverage quality prediction

Step1: Import Libraries: Import the necessary libraries, including NumPy, Pandas, scikit-learn, and specific modules for SVM.

Step2: Load the Dataset: Load the Beverage Quality dataset using scikit-learn's load_beverage() function. This dataset is a sample dataset that comes with scikit-learn, and it contains features related to wine, milk and water composition and a target variable indicating beverage class.

Step3: Data Preprocessing: Separate the features (X) and the target variable (y). In this dataset, the target variable represents the beverage class.

Step4: Model Selection: on this step suitable algorithm for regression is chosen. SVMs can be used for both classification and regression tasks.

Step5: Split the Data: Split the dataset into training and testing sets using train_test_split().

Step6: Model Training: Split the dataset into training and testing sets. Train the SVR model on the training data. The model learns to map the input features to the target variable.

Step7: Prediction: Use the trained SVR model to predict the beverage quality for new, unseen data.

Step8: Model Interpretation: SVMs provide coefficients associated with each feature, indicating their contribution to the prediction.

Step9: Evaluate the Model: Assess the model's performance using accuracy, classification report, and confusion matrix.

2.4. Empirical Review

Many studies have been conducted to understand the characteristics and complexity of beverage. Additionally, many researchers seek to understand which characteristics to be considered for the best quality of both alcoholic and none alcoholic beverages. Research has also sought to determine whether consumers could differentiate the beverage quality, in addition to investigating how to find other factors that affect their quality.

According to Ghulam Muhammad in book named quality control in beverage production stated that beverages are very important sector of the food industry based on all types of liquid foods including alcoholic (beers, wines, and spir-its) and nonalcoholic drinks (water, soft or cola drinks, fruit juices and smoothies, tea, coffee, dairy beverages, and carbonated and noncar-bonated drinks). The quality of any beverage manufacturing system is related to its quality management system’s effectiveness, that is, raw ingredients quality, processing layout, equipment’s quality, and satisfaction of consumers. The main objective of quality management is to develop knowledge and understanding, identify suitable methods for determination, evaluation, and monitoring of the product quality (according to the specifications of international quality standards), and prepare documentation and records.

Instances of assessments in the analysis of beverage and food quality, as well as the detection of beverage adulteration, involve examining both alcoholic and non-alcoholic drinks through various technologies. The emphasis in imaging technologies, particularly hyperspectral imaging, lies in examining the color, shape, and texture of substances. Odour and taste sensing technology (Quartz Crystal Microbalance (BAW), Metal Oxide Semiconductor (MOS-based electronic nose), and Electrochemical biosensors are focused on the specific components of an aroma or solution and analyses their chemical composition by contact with its headspace and immersed in sample respectively, whereas electromagnetic sensing technology measures the electromagnetic wave transmission coefficient using the frequency, polarization, and angle of incidence of the electromagnetic wave, as well as the object's permittivity and conductivity. (Mustafa et al. 2019; Wei et al. 2018).

Table 1: Comparison between electronic senses, sensory analysis and conventional laboratory instruments features.

Feature	Imaging technology	Odor sensing technology	Taste sensing technology	Electromagnetic sensing technology
Rapidness	Yes	Yes	Yes	Yes
Low-cost analysis	Yes	Yes	Yes	Yes
Use of chemicals	No	No	No	No
Objectiveness	Yes	Yes	Yes	Yes
Non-destructive measurements	Yes	Yes	No	Yes
Sample pre-treatment	No	No	Yes	No
Simplicity	Yes	Yes	Yes	Yes
Single operator	Yes	Yes	Yes	Yes
Permanent storage of data	Yes	Yes	Yes	Yes

Numerous variables which could result in variation, thereby affecting homogeneity of beverage product are there in the industry. The use of different grades of raw materials, like fruits and water, also result into variations in tastes, flavours and overall quality. In addition, it is possible to have risks of microbial contamination at various stages, problems such as storing under unconventional conditions as well as packing concerns. Such implementation of quality assurance is also complex because it involves incurring additional cost for regulatory compliance and adjustment due to dynamics of consumers' needs. Investment in modern technology including strict quality controls as well as continuous process improvements are required for the industry in order to be able to successfully solve those highly multi-dimensional problems that include real time monitoring and traceability along the supply chain.

2.5. Research Gap Analysis

Beverage quality testing, while essential, faces several gaps that can impact the accuracy and comprehensiveness of assessments. One notable gap is the potential for human error in the testing process. Manual testing procedures are inherently susceptible to variations introduced by individuals, leading to inconsistencies in results. Automation and the implementation of standardized testing protocols can help mitigate this gap, ensuring more reliable and reproducible outcomes.

In the beverage industry, commitment to quality is the only rule to swear by since the end-users directly consume the products delivered. Also, there are multiple supply chains involved in the process. Overlooking quality at any single stage could be of huge harm to the consumer's health as well as the brand's reputation (Chellappa, R. K., & Saraf, N, 2010). Beverage industries are facing various challenges in quality assurance, one of these challenges is:

- Testing procedures can be performed with outdated systems and lack coordination, which negatively impacts raw materials that need to be expedited to production and bottleneck dock or quarantine areas.

2.6. Conceptual Framework

A conceptual framework illustrates the expected relationship between variables. It establishes the pertinent goals for your research procedure and outlines how these goals align to derive cohesive conclusions.

In this research, a conceptual framework has been developed to aid the researcher in developing an understanding of the effects of AI-based hybrid deep learning models on beverage industries.

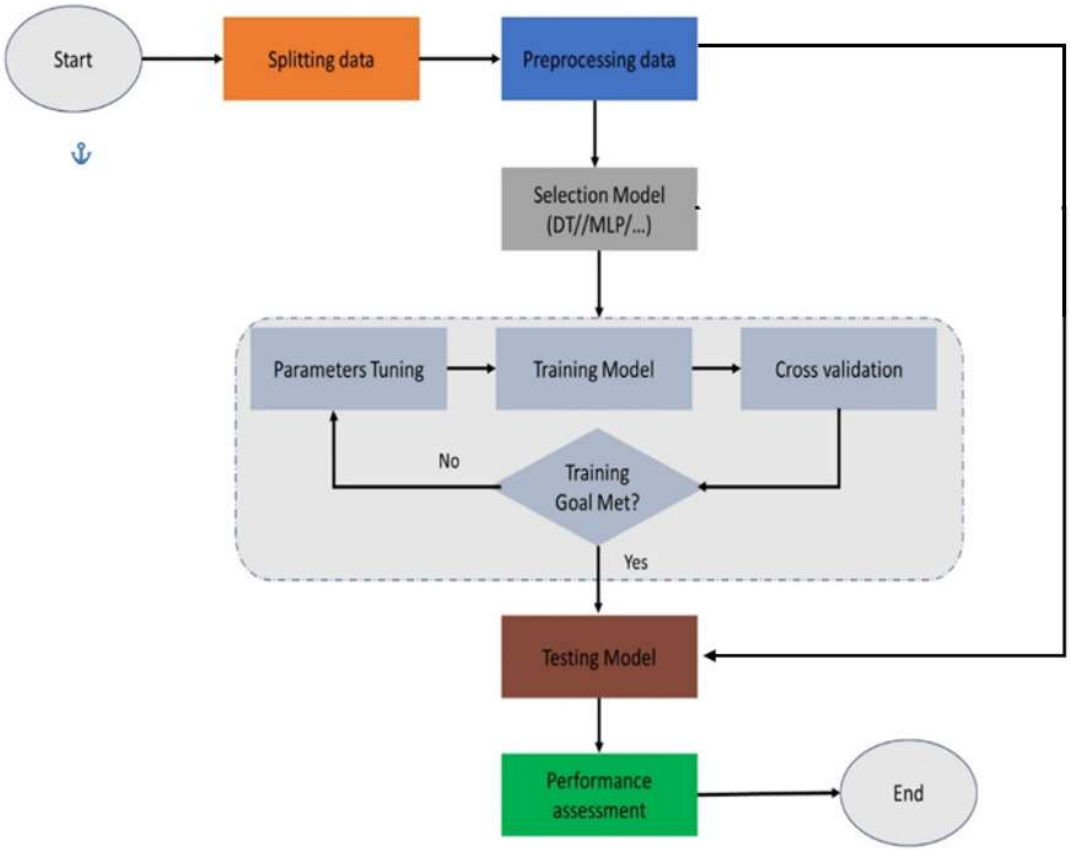


Figure 3: Data Processing Map(Conceptual Framework)

III. METHODOLOGY

Introduction

Research methodology simply refers to the practical "how" of a research study. More specifically, it involves how a researcher methodically plans a study to guarantee results that are both valid and reliable, effectively addressing the research aims, objectives, and questions.

3.1. Research Design

Refers to the framework of research methodologies and techniques selected by a researcher to carry out a study is known as research design. The design allows researchers to refine the research methods appropriate for the subject matter and establish their studies for success (Sileyew, 2019).

3.2. Study population

A population is the entire set of individuals within a group, be it a nation or a collection of people sharing a certain trait. In statistics, a population refers to the group of individuals from which a statistical sample is selected for a study. Thus, any selection of individuals sharing the same feature in common can be said to be a population.

Typically, a research population constitutes a substantial assembly of individuals or objects that forms the primary focus of a scientific inquiry (MOMOH, 2023). It is for the advantage of the population that researches are carried out. Population study is a cross-disciplinary area of scientific exploration employing diverse statistical methods and models to examine, identify, tackle, and forecast population issues and trends. This analysis is conducted using data gathered through various methods, including population census, registration processes, sampling, and other data sources.

3.3. Sampling Frame

The sample design plans encompassed details concerning sampling frames and their scope, offering explanations of national sample designs that incorporated sampling stages, probabilities of selection, sampling units, and sample descriptions. The plans for sample selection provided comprehensive details regarding the procedures for selecting samples at every level of the sampling process.

3.3.1. Sample size determination

A sample is a smaller set of data that a researcher chooses or selects from a larger population using a pre-defined selection method. These elements are known as sample points, sampling units, or observations. Kumekpor (2002) also states that a sample of a population consists of that proportion of the number of units selected for investigation. Jankowiez (2002) further states that, sampling is the deliberate choice of a number of people who are to provide the data from which conclusions about these people can be drawn. Based on the structure of the study population the size of the sample of this study was calculated by use of the formula of Yamane, where:

$$n = \frac{N}{1 + N(e)^2}$$

Here: n = Sample
N = Population
e = level of precision (error)

3.3.2. Sampling Technique

Berg (2009) explained "Sampling" as the way of selecting a finite number of items from a population of interest. Probability sampling was used in the research. The reasons for choosing this sampling technique were its convenience and economy. A simple random sampling method was used for the probability sampling, which is the most basic sampling method assumed in statistical calculations and methods. To collect a simple random sample, each unit in the target population is assigned a number. A set of random numbers is then generated and units with those numbers are included in the sample.

3.4. Data collection methods and Tools

Data collection methods are techniques and procedures used to gather information for research purposes. These techniques can be either quantitative or qualitative in nature, and they can range from straightforward self-reported surveys to intricate experiments.

3.4.1. Quantitative research

The research is based on quantitative research because it is based on real dataset and prediction. This research approach involves quantifying variables through a numerical system, analyzing these measurements with various statistical models, and presenting the relationships and associations observed among the variables studied (Bhandari, 2023).

3.4.2. Secondary data source

Refers to the data that a researcher has not gathered or created themselves. Secondary data can include some of the most comprehensive and unique investigations, as well as some of the biggest and most meticulous data sets. (Ajayi, 2017). Secondary research can be qualitative or quantitative in nature. Frequently, it utilizes information obtained from published peer-reviewed articles, meta-analyses, or databases and datasets from government or private sectors. It means that the information is already available, and someone analyses it. The secondary data includes magazines, newspapers, books, journals, etc. It may be either published data or unpublished data.

3.4.3. Publicly available databases

Publicly available databases in machine learning are datasets accessible to the public for research and model development. These datasets, often spanning diverse domains like image recognition, natural language processing, and healthcare, serve as valuable resources for training and evaluating machine learning models. With open accessibility and standardized formats, such as MNIST for digit recognition or ImageNet for image classification, these databases enable researchers and practitioners to benchmark algorithms, foster collaboration, and advance the field's collective knowledge. The availability of ground truth labels facilitates supervised learning tasks, making these datasets crucial for exploring and addressing a wide range of machine learning challenges (Altexsoft, 2019).

3.5. Data Processing

This in research means the collection and translation of a data set into valuable, usable information. In this procedure, a researcher, data engineer, or data scientist transforms raw data into a more comprehensible format, such as a graph, report, or chart, either manually or with the assistance of an automated tool.

3.5.1. Editing

Editing refers to the process of identification of errors and possible misstatement in the questionnaires and interviews. During editing process, raw data are checked for mistakes arising from either the compiler of the questionnaires and interview schedules or the respondents. During this process, unclear responses are clarified. For the purpose of this study as any survey studies, editing of data collected aims at improving the quality of information from respondents and eliminates the useless responses for easy analysis and interpretation.

3.5.2. Tabulation

Following the editing process, which verifies the accuracy and appropriate categorization of information on the schedule, the data are organized into tables and may undergo various forms of statistical analysis. For the purpose of this study, frequency distribution tables and percentages for tabulating the gathered information, edited and coded data. Tables were built based on themes developed from questions in a respective order for easy interpretation.

3.5.3. Coding

Coding involves composing instructions for computers and other hardware. Subsequently, the computer interprets these instructions, referred to as "programs," and performs the tasks you have specified. During this study, this process was used to sum up data different categories of responses into an easily interpretable way. For the purpose of this study, coding questions numerical values were used.

3.7. Data Analysis

Data analysis in machine learning is modifying, altering, and visualizing data in order to derive significant conclusions from the findings. People, companies, and even governments frequently make decisions based on these observations.

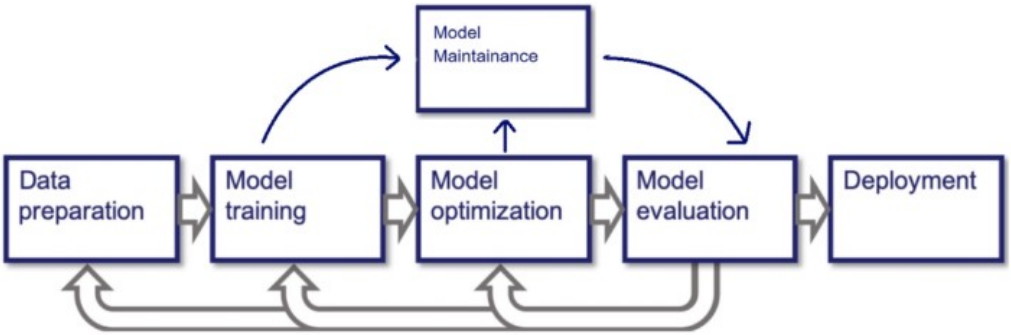


Figure 4: Data Analysis

ource: (analyticssteps, 2020)

3.8. Hardware and Software Requirements

3.8.1. RAM

Although 32 GB of RAM would be ideal because training any algorithm would take a lot of time, 16 GB is the minimal requirement. Less than 16 GB can make multitasking challenging.

3.8.2.CPU

Because it is more potent and provides High Performance, processors above Intel Corei7 7th Generation are encouraged.

3.8.3.GPU

This is the most crucial factor since neural networks, a computationally expensive subfield of Deep Learning, are necessary for it to function. Matrix calculations are heavily required when working on images or videos. These matrices can be processed in parallel thanks to GPUs. The procedure can take days or months without a GPU. But with it, you can complete the same activity on your Best Laptop for Machine Learning in a matter of hours.

3.8.4. Storage

Given that the datasets become steadily bigger every day, a minimum of 1TB HDD is required. A minimum of 256 GB is suggested if you have an SSD-equipped PC. However, if you don't have enough storage, you can use cloud storage options. Even machines with powerful GPUs are available there.

3.8.5. Operating System

The majority of people use Linux, but you can also work on Windows and MacOS if you install a virtual Linux environment on those platforms.

3.9. Agile Development Methodology

The meaning of Agile is swift or versatile."Agile process model" refers to a software development approach based on iterative development. Agile approaches divide tasks into smaller iterations, avoiding extensive long-term planning. The project scope and requirements are established at the commencement of the development process. Plans outlining the number of iterations, their duration, and scope are distinctly defined in advance.

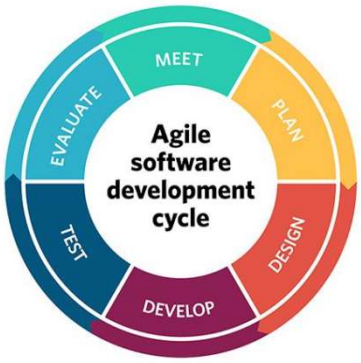


Figure 5: Agile SDLC

Source: (medium, 2015)

3.10. Tools to be used

3.10.1. Python

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python brings an exceptional amount of power and versatility to machine learning environments. The language's straightforward syntax makes it easier to validate data and expedites the procedures of scraping, processing, refining, cleaning, organizing, and analyzing; as a result, working with other programmers becomes less difficult. Additionally, Python provides a rich ecosystem of libraries that eliminate a large portion of the tedious routine function writing activities, freeing developers to concentrate on writing code and lowering the likelihood of errors when programming (RedMonk, 2021).

3.10.2. Anaconda

An open-source distribution of Python and R for data research called Anaconda seeks to make package management and deployment easier. Conda, the package management system used by Anaconda, is in charge of managing package versions. Conda examines the current environment before initiating an installation to prevent conflict with other frameworks and packages (Tutorialspoint, 2023).

3.10.3. Spyder

An open-source, cross-platform IDE is Spyder. Python is the only language used to create the Python Spyder IDE. It was created particularly for scientists, data analysts, and engineers and was designed by scientists. It also goes by the name Scientific Python Development IDE and contains a long list of outstanding capabilities, some of which are covered here.

3.10.4. Jupyter Notebook

Jupyter notebooks find application in various data science activities, including exploratory data analysis (EDA), data cleaning and transformation, data visualization, statistical modeling, machine learning, and deep learning. Notebooks take the console-based approach to interactive computing in a new direction, offering a web-based application that encompasses the entire computational process, including code development, documentation, execution, and communication of results (Nelli, 2015). Jupyter notebook are especially for "showing the work" that your data team has completed using a combination of code, markdown, links, and photos, Jupyter notebooks are quite helpful. They are simple to use, and you may run them cell by cell to learn more about the functions of the code.

3.10.5. NumPy(Numerical Python)

A powerful library for numerical computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays. NumPy can be used to perform a wide variety of mathematical operations on arrays. It introduces robust data structures to Python, ensuring efficient calculations involving arrays and matrices, and it provides an extensive library of high-level mathematical functions designed to operate on these arrays and matrices.

3.10.6. Pandas

Pandas is a powerful and widely used open-source data manipulation and analysis library for Python. It provides data structures for efficiently storing and manipulating large datasets and tools for working with structured data. Constructed atop another package called Numpy, it leverages support for multi-dimensional arrays. As a widely used data manipulation tool, Pandas seamlessly integrates with numerous other data science modules within the Python ecosystem. It is commonly included in all Python distributions, ranging from those bundled with your operating system to commercial vendor distributions like ActiveState's ActivePython. (Activestate, 2020).

IV. 4. Presentation, Analysis and Interpretation of Findings

Introduction

The process of system analysis, design, development, implementation, and testing is covered in this chapter. The results of the data collection are used to inform the system design. During the data collecting phase, all user needs that were gathered from feedback are used to design a system that meets the needs of the users. Following system implementation, targeted users would be tested to verify that the system is appropriate for them and to ascertain whether the goals have been met.

4.2. System Implementation

4.2.1. Wine Quality

As the most widely consumed beverage in the world, wine is valued highly in society. Wine quality is always important to consumers, but it's especially important for producers to increase revenue in the current competitive market.

4.2.1.2. Data Source and Description

The wine quality dataset, which is accessible to the public, was acquired from the UCL Machine Learning Repository. This repository encompasses an extensive array of datasets that have been extensively utilized by the machine learning community. The wine dataset contains 11 physiochemical properties: fixed acidity (g[tartaric acid]/dm^3),volatile acidity (g[acetic acid]/dm^3), total sulfur dioxide (mg/dm^3), chlorides (g[sodium chloride]/dm^3), pH level, free sulfur dioxide (mg/dm^3), density (g/cm^3), residual sugar (g/dm^3), citric acid (g/dm^3), sulphates (g[potassium sulphate]/dm^3), and alcohol (vol%).

Alongside these properties, a sensory score was acquired from several different blind taste testers which graded each wine sample with a score ranging from zero (poor) to 10 (excellent). The median was recorded and serves as the response variable. The dataset includes 4898 randomly selected wine manufacturing samples' records. Many statistical analyses were carried out in order to comprehend the dataset's nature. The degree of correlation between two distinct variables is indicated by the Pearson correlation coefficient (r). If "r" is near to 1, the relationship between two variables is deemed highly positive; if "r" is near to -1, it is deemed highly negative. We determined the Pearson correlation coefficient between each variable and the wine quality (i.e., target property) in our dataset before feeding the data into the ML models.

According to our analysis, the quantity of alcohol has the highest correlation coefficient (0.435) with the target property, while the lowest (−0.009) is with citric acid. In the statistical analysis, the variables that have a significantly lower correlation coefficient (almost zero) with the target property may be deemed irrelevant. These variables introduced noise into the dataset and misled the training process, which can have a significant impact on the predicted property during ML model training. As a result, there are subpar models and lower prediction accuracy.

There are various strategies for reducing noise. Eliminating the superfluous, redundant, and insignificant predictors is one of the most well-liked and frequently applied denoising techniques. A statistician prioritizes the method because it is easy to use and straightforward. ML algorithms are outlier-sensitive. It may taint and misdirect the instruction. This could lead to subpar models that produce results that are ultimately less accurate. Therefore, it is standard practice to examine outliers when preprocessing the data. A boxplot is a standard method for showing the data's distribution.

It is frequently employed to determine the distribution's shape and the potential presence of outliers. boxes for every feature. All the variables, with the exception of alcohol, are either skewed or may contain outliers, according to these boxplots. We might remove the extreme values from the data set in order to eliminate outliers. However, removing data is always a drastic measure,

so it should only be done in the direst circumstances when we are positive that the outliers are caused by measurement errors. We are unable to confirm these extreme values as measurement errors at this time, so we are unable to drop them.

Table 2: Descriptive statistics of the variables of the wine data.

Variable Name	Mean	Standard deviation	Minimum	Maximum	Median
Fixed acidity	6.854	0.843	3.80	14.2	6.80
Volatile acidity	0.278	0.100	0.08	1.10	0.26
Citric acid	0.334	0.121	0.00	1.66	0.32
Residual sugar	6.391	5.072	0.60	65.8	5.20
Chlorides	0.045	0.021	0.009	0.35	0.04
Free sulfur dioxide	35.30	17.00	2.00	289	34.0
Total sulfur dioxide	138.4	42.49	9.00	440	134
Density	0.994	0.002	0.99	1.038	0.99
PH	3.188	0.151	2.27	3.82	3.18
Sulphates	0.489	0.114	0.22	1.08	0.47
Alcohol	10.51	1.230	8.00	14.2	10.4
Quality	5.877	0.885	3.00	9.00	6.00

Table 3: The value of the pearson correlation coefficient (r) of the predictors with respect to the target variable: quality

Predictor	r	Predictor	r	Predictor	r
Alcohol	0.435	Citric acid	−0.009	Volatile acidity	−0.194
pH	0.099	Residual sugar	−0.097	Chlorides	−0.209
Sulphates	0.053	Fixed acidity	−0.113	Density	−0.307
Free sulfur dioxide	0.008	Total sulfur dioxide	−0.174		

4.2.1.3. Feature Scaling

The variables are widely dispersed, as Table 1 illustrates. For example, the total sulfur dioxide values are much higher than the chlorides. When training machine learning models, a single variable with an extremely high value may exert a dominant influence over other quantities. For example, if one does not standardize the nonuniform data when performing K-nearest neighbor KNN or SVM, the performance of the KNN or SVM model would be dominated by the datapoints with high distance. Thus, before training any machine learning model, feature scaling is a crucial step that needs to be attended to. Numerous techniques exist for feature scaling. Standardization and normalization are the most widely used and well-liked methods in the machine learning community. There isn't enough theoretical data to conclude which approach is the most effective. Standardization has been applied to the dataset in order to scale its features. The following formulas were used to determine the standardization:

$$z = \frac{\mathrm{x} - \mathrm{mean}}{\mathrm{std}}$$

where z, x, mean, and std are standardized input, input, mean and standard deviation of the feature, respectively.

4.2.1.4. Data partition

A 3:1 split of the data was made into training and testing sets. To determine the relationship between the target and predictor variables, we train the data. Preventing overfitting is the major goal of data splitting. The machine learning algorithm may perform exceptionally well in the training dataset but poorly in the testing dataset if overfitting takes place.

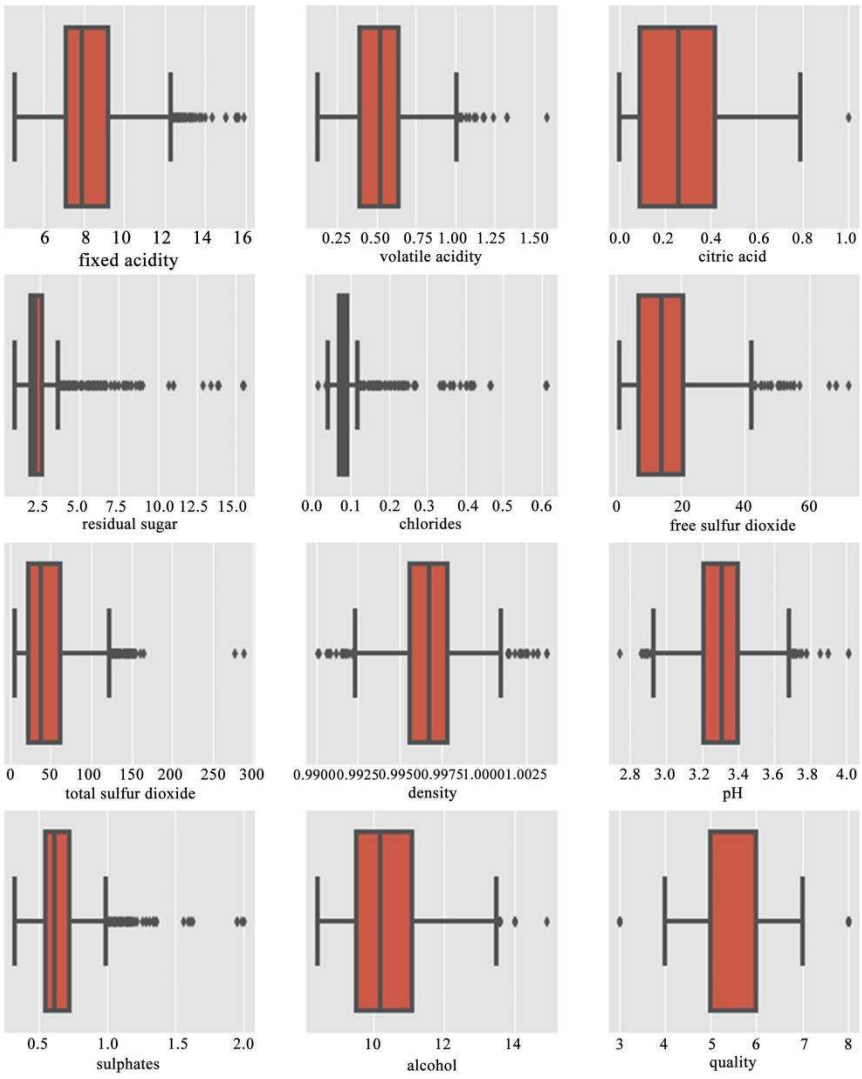


Figure 6: Box plot of the variables of the wine data.

4.2.1.4. Machine Learning Algorithms

For the learning process, a variety of machine learning algorithms are available, including neural networks, support vector machines, kernel methods, logistic regression, linear regression, and many more. Every technique has advantages and disadvantages. In this work, we predict wine quality using the following supervised learning algorithms.

4.2.1.5. Support Vector Machine

Support Vector Machine (SVM) is one of the most popular and powerful machine learning algorithms which was introduced in the early 90s. When used for regression, SVM is also known as Support Vector Regressor (SVR). SVR is a kernel-based

regression technique which maps nonlinearly separable data in real space to higher dimension space using kernel function. It is equipped with various kernels such as linear, sigmoid, radial, and polynomial. In this work, we have used radial basis kernel (RBF) because it outperformed other kernels based SVR in wine dataset. The performance of the SVR is controlled by two important tuning parameters (cost: regularization parameter and gamma: kernel coefficient for RBF). The tuning parameter cost control the bias and variance trade-off. The small value of the tuning parameters cost underfits the data, whereas the large value overfit.

4.2.1.6. Gradient Boosting Regressor

Gradient Boosting Regression (GBR) is one of the leading ensemble algorithms used for both classification and regression problems. Which builds an ensemble of weak learners in sequence with each tree and together make an accurate predictor. Decision tree is one of the most popular choices of such ensemble models. Each new tree added to the ensemble model (combination of all the previous tree) minimize the loss function associated with the ensemble model. The loss function depends on the type of the task performed and can be chosen by the user. For GBR, the standard choice is the squared loss. A key factor of this model is that adding sequentially trees that minimize the loss function, the overall prediction error decreases. By tuning many hyperparameters such as the learning rate, the number of trees, maximum depth we can control the gradient boosting performance which helps to make model fast and less complex. Detailed explanation of the GBR algorithm can be found in Friedman et al.

4.2.1.7. Artificial Neural Network (ANNs)

ANNs are a very primitive generalization of biological neurons. They are composed of layers of computational units called neurons, with a connection between different layers through the adjustable weights. The major constituents of ANNs are weights, bias, and the activation function. An excellent choice of the activation function results in the proper accuracy of an ANN model. The most widely used activation functions are Logistic (known as Sigmoid) Rectified linear unit, and the SoftPlus. Passage of information along a predetermined path between the neurons is the fundamental idea behind the construction of ANNs. Its architecture is very flexible, and various network parameters (such as weights, bias, number of nodes, and number of hidden layers) can be tuned to improve the performance of the network. One can add up the information from multiple sources to the neurons and apply a non-linear transformation at each node, which helps the network to learn the complexity present in the data. With the application of linear and non-linear transformation in the input data, ANNs transform those initial representations up to a specific outcome. Depending on the learning task, the outcome of the network could be either classification or regression. The schematic diagram for the ANN used in this work is presented in Figure 2.

4.2.1.7. Results and Discussion

With the goal of assessing the performance of the different ML algorithms, we have used three most popular machine learning algorithms, namely: Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and Artificial Neural Network (ANN) to predict the wine quality in the wine data. This allows us the freedom to select the most suitable ML algorithm to predict the wine quality with the given variables.

All of the model's performance (on the training and test data) explained in the previous sections are evaluated by using Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and correlation coefficient (R) defined as

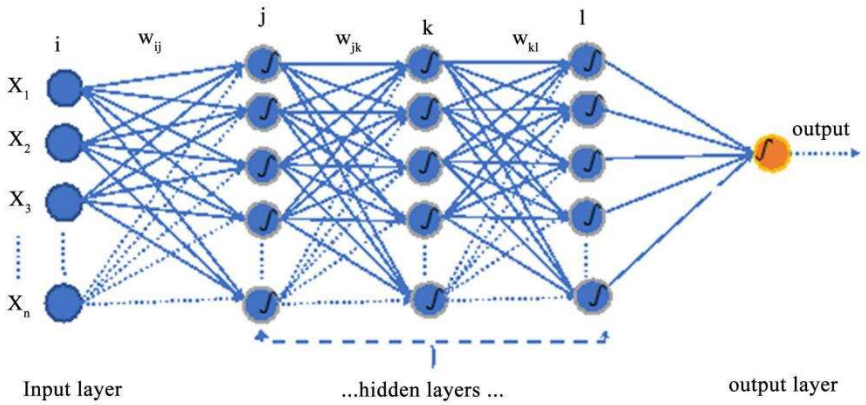


Figure 7: The schematic of an ANN, with three hidden layers and one output layer with Relu activation function at each node

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)\,1^2$$

$$MAPE = \frac{1}{n}\sum_{n=1}^{\infty}\left|\frac{y_{i-\hat{y}_i}}{y_i}\right|$$

$$R = \frac{\sum_{i=1}^{n}\left(y_{i-\hat{y}_i}\right)(\hat{y}_{i-}\hat{y})}{\sqrt{\sum_{i=1}^{n}\left(y_{i-\hat{y}_i}\right)^2\left(\hat{y}_{i-\hat{y}_i}\right)}}$$

where the target predicted values (yi) and the target observed values ($\hat{y}_i$) are, respectively. Moreover, n represents the total number of observations.

The optimal tuning parameter λ for the RR model is selected using the 10-fold cross-validation. By varying λ with increments of 0.01 from 0.00001 to 100, we employed the grid search technique to determine the optimal parameter. The variation in MSE with λ is depicted in Figure 3. We found that 45.25 is the ideal value of λ since it minimizes the MSE. This tuning parameter λ is used to fit an RR model, and Table 3 shows the model's performance.

In addition to the RR, the performance of the kernel-based machine learning algorithm SVM is compared. Similar to RR, the 10-fold cross-validation is used to obtain the tuning parameters cost and gamma. We used the grid search technique to obtain these tuning parameters by varying each between 0.01 to 10. For each possible combination of cost and gamma, MSE is computed. A heatmap of

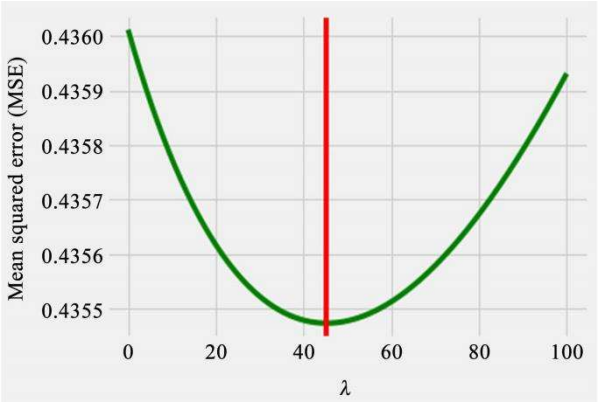


Figure 8: Graph of tuning parameter λ versus the Mean squared error (MSE) of the RR model.

Table 4: Model performance metrics obtained using training and test datasets.

Models		Training data set			Test data set	
	R	MSE	MAPE	R	MSE	MAPE
RR	0.6029	0.4281	0.0934	0.5897	0.3869	0.0888
SVM	0.7797	0.267	0.1426	0.5971	0.3862	0.1355
GBR	0.7255	0.3286	0.0826	0.6057	0.3741	0.0873
ANN	0.66	0.37	0.14	0.58	0.4	0.12

these tuning parameters versus MSE is plotted in Figure 4. The optimal values the parameters computed using 10-fold cross-validation are cost = 0.95 and gamma = 0.13. An SVM model with these tuning parameters is fitted, and its performance is presented in Table 3.

Gradient boosting was also used to predict the wine quality. It has hyperparameters to control the growth of Decision Trees (e.g., max_depth, learning rate), as well as hyperparameters to control the ensemble training, such as the number of trees (n_estimators). In the tuning of the model parameters, we test the learning rate from low (0.01) to high (0.2) as well as a number of trees in the range 1 to 200. The results show that setting the learning rate to (0.05) has better predictive outcomes. Figure 5 shows the change in validation error with the number of iterations. We use an early stopping process that performs model optimization by monitoring the model's performance on a separate test data set and stopping the training procedure once the performance on the test data stops improving beyond a certain number of iterations. We found better predictive outcomes at n_estimators = 40, which is indicated by a red star in Figure 5. The GBR model based on this tuning parameter is fitted, and its performance is presented in Table 3.

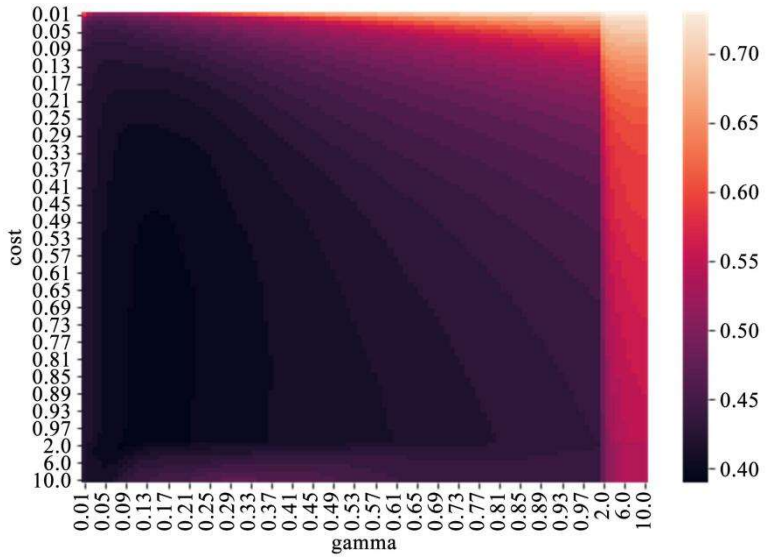


Figure 9: Heatmap showing tuning parameters cost and gamma with colors bars displaying mean squared error.

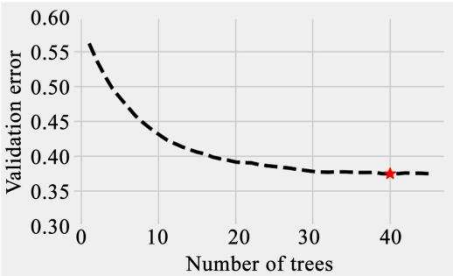


Figure 10: Variation of validation error with number of trees (i.e. n_estimators).

As ANN performs very well in compared to other mathematical models in most of the dataset, we test its performance to predict the wine quality in wine dataset. Before using the model in the test dataset, we train the model by tuning various network parameters such as the number of layers and the number of nodes in each layer. For the sake of comparison, we used gradient descent (GD) and Adam as an optimization algorithm to update the network weights. The comparison shows that the Adam optimizer outperforms the prediction of wine quality than Gradient descent, so we use Adam as an optimization algorithm and the optimized network that can make the best prediction is obtained. The detailed architecture and the working of ANNs can be found elsewhere. In this work, we use ANN with one input layer, three hidden layers (each with 15 neurons) and one output layer.

For the training and test process, we choose a 60-20-20 train-validation-test split. Before passing to the network for the training purpose, the data were normalized by using the method described earlier in Equation. The model was trained on the training set and validated on the validation set to make sure there is no overfitting or underfitting during the training process. By tuning various hyperparameters such as learning rate, batch size, and number of epochs an optimized ANN model is obtained. Once, the model is optimized, it is tested on the test dataset. and its performance is evaluated by using MSE, R and MAPE. The performance comparison between various mathematical models and the ANN used in this work is presented in Table 3.

As presented in Table 3, GBR model shows the best performance (highest R as well as least MSE and MAPE) among the four models we used to predict the wine quality. The performance of ANN is very close to other models, but it is unable to surpass the accuracy obtained for GBR. It might happen because of the small number of datasets, we used to train the ANN, or the dataset is too simple, and the model is complex to learn enough the data. In addition, importance features from GBR that determines the wine quality is presented in Figure 6. When we plot the feature importance of all features for our GBR model we see that the most important feature to control the wine quality is turn out to be an alcohol. Which perfectly make sense because it is not only about the feelings after drinking in fact it effects the teste, texture and structure of the wine itself. The second

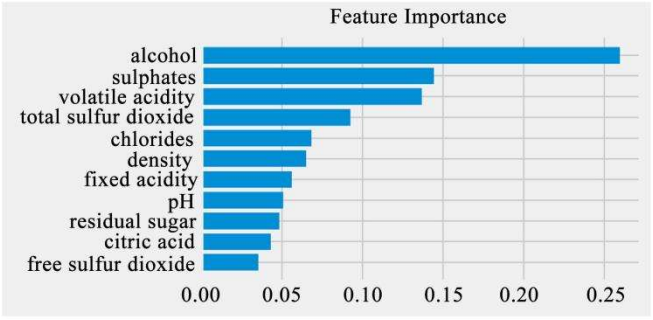


Figure 11: Feature importance for the wine quality for our best model GBR.

most important feature is the sulphates, which is by definition somewhat correlated with the first feature. From plot what we also observed is the least important feature is the free sulfur dioxide. Which is a measure of the amount of SO_2 (Sulfur Dioxide) which is used throughout all stages of the winemaking process to prevent oxidation and microbial growth.

4.2.2. Water Quality

Analysis of water quality is a complicated topic because of the various factors that affect it. The different uses that water is put to are intrinsically related to this idea. Standards must vary according to needs. Numerous studies are being conducted on the prediction of water quality. A number of physical and chemical factors that are directly related to the intended use of the water are typically used to determine the quality of the water. Next, each variable's acceptable and unacceptable values need to be determined. Water deemed suitable for a given application is that which satisfies the specified requirements.

Dataset

The water_potability.csv file contains water quality metrics for 3276 different water bodies.

1. pH value: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. Solids (Total dissolved solids - TDS): Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity: Pure water is not a good conductor of electric current Rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the number of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceed 400 µS/cm.

7. Organic carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

Table 5: Water Quality Dataset

	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramin	Sulfate	Conductiv	Organic_ca	Trihalome	Turbidity	Potability
2		204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3	3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
4	8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
5	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
6	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
7	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
8	10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
9	8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
10		118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11	11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
12	7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
13	7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
14	7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
15		150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
16	7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
17	6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
18	7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
19	9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
20	8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0
21	7.37105	214.4966	25630.32	4.432669	335.7544	469.9146	12.50916	62.79728	2.560299	0
22		227.435	22305.57	10.33392		554.8201	16.33169	45.38282	4.133423	0
23	6.660212	168.2837	30944.36	5.858769	310.9309	523.6713	17.88424	77.04232	3.749701	0
24		215.9779	17107.22	5.60706	326.944	436.2562	14.18906	59.85548	5.459251	0
25	3.902476	196.9032	21167.5	6.996312		444.4789	16.60903	90.18168	4.528523	0
26	5.400302	140.7391	17266.59	10.05685	328.3582	472.8741	11.25638	56.93191	4.824786	0
27	6.514415	198.7674	21218.7	8.670937	323.5963	413.2905	14.9	79.84784	5.200885	0
28	3.445062	207.9263	33424.77	8.782147	384.007	441.7859	13.8059	30.2846	4.184397	0
29		145.7682	13224.94	7.906445	304.002	298.9907	12.72952	49.53685	4.004871	0

water_potability

Classification

To estimate river water quality class, two data mining methods were used: Decision Tree (DT) and K- Nearest Neighbor (KNN). These methods are both parametric and nonparametric classifiers, and their goal is to develop a function that maps input variables to output variables from a training dataset. Because the function's form is unknown, different algorithms make different assumptions about the function's form and how training data is learned to produce the output. The parametric learning classifier makes more confident assumptions about the data. If the assumptions for any data set are true, these classifiers would make rectification judgments. However, if the assumptions are incorrect, the same classifier performs poorly. In order to learn classification tasks, these classifiers do not rely on the quantity of the sample data set; rather, their working principles are their assumptions. This classifier is susceptible to prediction mistakes such as bias, in addition to its parametric character. When the model makes multiple assumptions, the Decision Tree yields substantial bias. Nonparametric classifiers, unlike parametric learning classifiers, do not make any assumptions about the form of the mapping function, and by not making any assumptions, they are having more accuracy. These classifiers can create any function from the training data set. The DT and KNN classifiers are included in this category. Learning techniques are used in DT, whereas the similarity principle is used in KNN. To put it another way, DT Small data sets with complete domain expertise, on the other hand, are equally advantageous for these classifiers. Instead of learning from data, the KNN classifier finds a group of k items in the training set that are the most similar to the test object. Unlike other classifiers, DT does not rely on domain expertise. To make classification decisions, it simply calculates the distance between two characteristics. Because each algorithm's mode of operation differs, a comparison of all of them is necessary to determine which one is better at approximating the underlying function for the same training and testing water quality datasets.

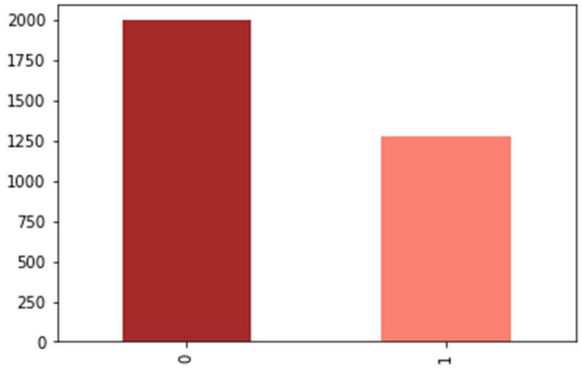


Figure 12: Potability Counts of Dataset

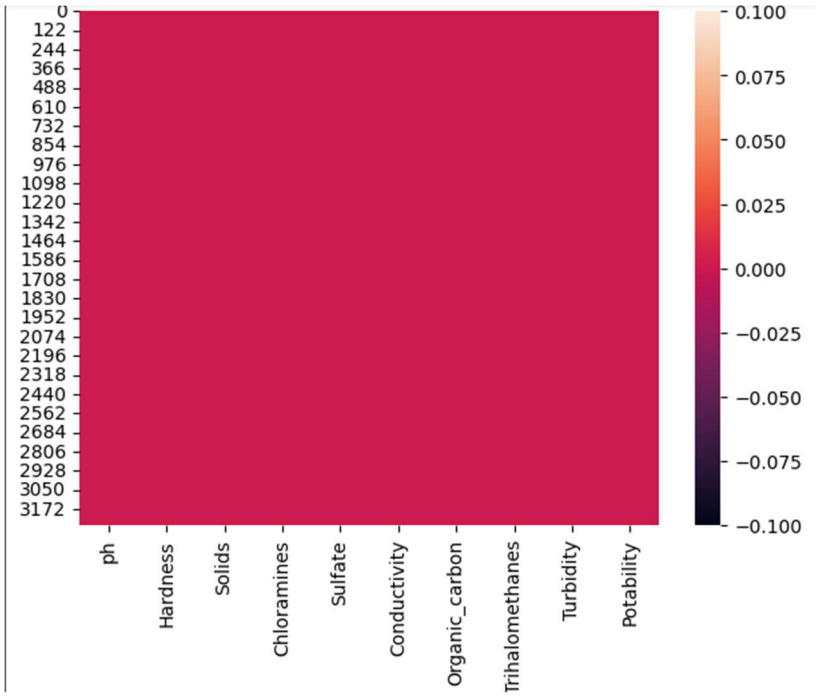


Figure 13: Heat Map for Water Quality

In figure 6, heat map was used as They are useful for exploring the patterns and trends within large datasets, particularly in correlation matrices or confusion matrices. Furthermore, using heat maps makes it easier to identify patterns visually and quickly discern areas of interest or concern within the data.

Performance Measures Results

True Positives (TP) are when the model predicts the positive class properly.

True Negatives (TN) is one of the components of a confusion matrix designed to demonstrate how classification algorithms work. Positive outcomes that the model predicted incorrectly are known as False Positives (FP). False Negatives (FN) are negative outcomes that the model predicts negative class. Accuracy is the most basic and intuitive performance metric, consisting of the ratio of successfully predicted observations to total observations.

Accuracy = TP+TN/(TP+FP+FN+TN)

Table 6: Comparison of algorithms

SN.	Algorithm Type	Accuracy score	Precision	Recall	f1-Score
1	Decision Tree	58.5	0.42	0.38	0.40
2	K-Nearest Neighbour	61.7	0.43	0.12	0.18

4.2.3. Milk Quality

4.2.3.1. Milk quality dataset

The Milk Quality dataset used in this study was taken from the open source Kaggle data storage area. The dataset data used were obtained from manual observations. Milk samples have 7 attributes: pH, Temperature, Taste Odor, Oil, Turbidity and Color. Generally, the quality of milk is determined by looking at these characteristics. The target is to classify the milk as Low (Poor), Medium (Medium) and High (Good). Taste, Smell, Oil and Turbidity properties take the value 1 or 0. Temperature, pH and Color properties have true color values.

Table 2 shows the values of 15 randomly selected milk samples (Kaggle,2023)

Table 7: Milk Quality dataset and classification (Kaggle, 2023)

pH	Temperature	Taste	Odor	Fat	Turbidity	Colour	Grade
6.6	38	1	0	1	0	255	high
6.8	45	1	1	1	1	245	high
6.8	36	0	1	1	0	253	high
6.6	45	0	1	1	1	250	high
6.8	45	1	1	1	1	245	high
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
6.8	43	1	0	1	0	250	medium
7.4	65	0	0	0	0	255	low
3	40	1	0	0	0	255	low
9	43	1	1	1	1	248	low
3	40	1	1	1	1	255	low
8.6	55	0	1	1	1	255	low

4.2.3.2. Results and Discussions

In the study, it was seen that the pH value and temperature value of the milk are an important factor in determining the milk quality. It has been observed that the pH value of high and medium quality milk is between 6-7, and the temperature values are at most 45 degrees. In low quality milk, it was observed that these two conditions did not provide at the same time. The distribution of the results obtained is shown on figure 5.

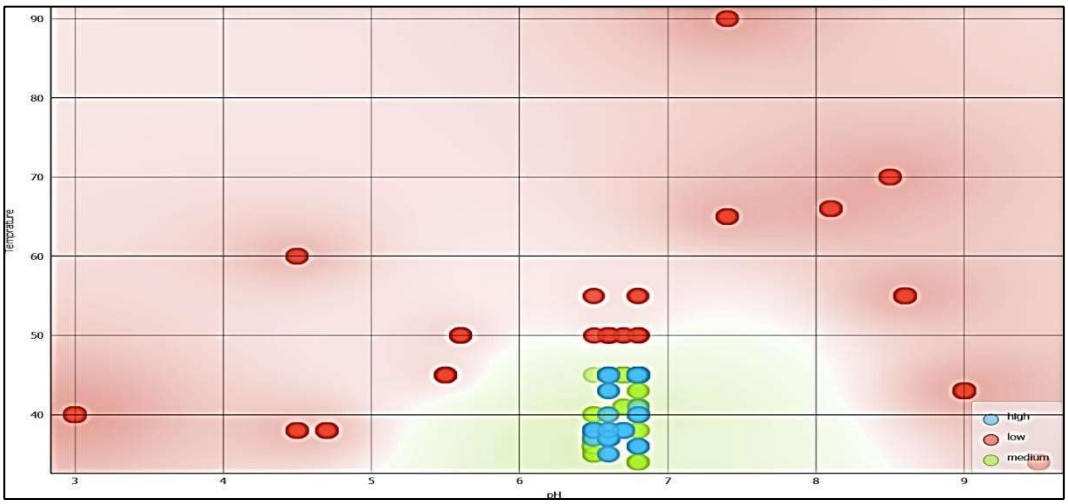


Figure 14: Classification distribution of temperature and pH values

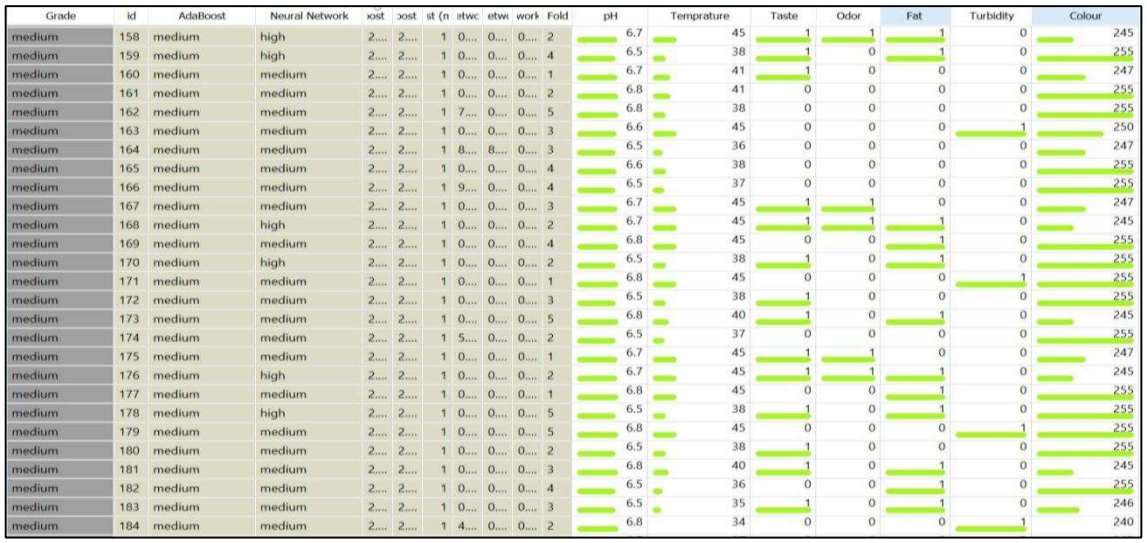
The "high" classification results and attribute data of Milk Quality data with AdaBoost and Neural Network algorithm using Orange Data Mining Tool are shown on Figure 6. The “high” actual values are shown in the Grade field, and the estimated classification prediction results in the AdaBoost and Neural Network field.

With the Neural Network algorithm, in the example with Id number 153, the quality result that should have been "high" was incorrectly estimated as "medium".

Grade	id	AdaBoost	Neural Network	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
high	143	high	high	6.6	45	0	1	1	1	250
high	144	high	high	6.8	45	0	1	1	1	255
high	145	high	high	6.6	45	0	1	1	1	250
high	146	high	high	6.8	45	1	1	1	1	245
high	147	high	high	6.8	36	0	1	1	0	253
high	148	high	high	6.6	37	1	0	1	0	255
high	149	high	high	6.5	38	1	1	1	1	255
high	15	high	high	6.8	36	0	1	1	0	253
high	150	high	high	6.8	40	1	1	1	1	255
high	151	high	high	6.8	45	1	1	1	0	245
high	152	high	high	6.8	40	1	1	1	1	255
high	153	high	medium	6.8	45	1	1	1	0	245
high	154	high	high	6.6	37	1	1	1	1	255
high	155	high	high	6.7	38	1	0	1	0	255
high	156	high	high	6.8	43	1	0	1	0	250
high	16	high	high	6.6	45	0	1	1	1	250
high	17	high	high	6.8	45	1	1	1	1	245
high	18	high	high	6.8	36	0	1	1	0	253
high	19	high	high	6.6	37	1	0	1	0	255
high	2	high	high	6.8	45	1	1	1	1	245
high	20	high	high	6.5	38	1	1	1	1	255
high	21	high	high	6.8	40	1	1	1	1	255
high	22	high	high	6.8	45	1	1	1	0	245
high	23	high	high	6.6	37	1	1	1	1	255
high	24	high	high	6.6	35	0	1	1	1	255
high	25	high	high	6.8	45	0	1	1	1	255
high	26	high	high	6.5	38	1	1	1	1	255

Figure 15:“high” classification results and attribute data obtained with AdaBoost and Neural Network algorithms

The results of “medium” classification and attribute data of Milk Quality data with AdaBoost and Neural Network algorithm using Orange Data Mining Tool are shown on figure 7. The “medium” real values are shown in the Grade field, and the predicted classification prediction results in the AdaBoost and Neural Network field. In the samples with Id numbers 158,159,168,176 and 178 with the Neural Network algorithm, the quality result that should have been "medium" was incorrectly estimated as "high".

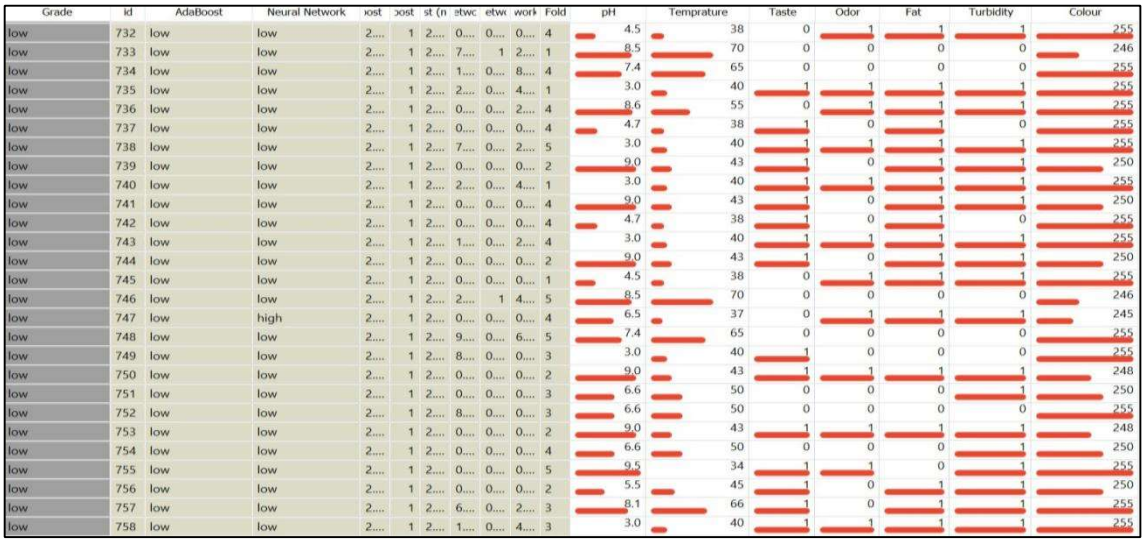


Grade	id	AdaBoost	Neural Network	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
medium	158	medium	high	6.7	45	1	1	1	0	245
medium	159	medium	high	6.5	38	1	0	1	0	255
medium	160	medium	medium	6.7	41	1	0	0	0	247
medium	161	medium	medium	6.8	41	0	0	0	0	255
medium	162	medium	medium	6.8	38	0	0	0	0	255
medium	163	medium	medium	6.6	45	0	0	0	1	250
medium	164	medium	medium	6.5	36	0	0	0	0	247
medium	165	medium	medium	6.6	38	0	0	0	0	255
medium	166	medium	medium	6.5	37	0	0	0	0	255
medium	167	medium	medium	6.7	45	1	1	0	0	247
medium	168	medium	high	6.7	45	1	1	1	0	245
medium	169	medium	medium	6.8	45	0	0	1	0	255
medium	170	medium	high	6.5	38	1	0	1	0	255
medium	171	medium	medium	6.8	45	0	0	0	1	255
medium	172	medium	medium	6.5	38	1	0	0	0	255
medium	173	medium	medium	6.8	40	1	0	1	0	245
medium	174	medium	medium	6.5	37	0	0	0	0	255
medium	175	medium	medium	6.7	45	1	1	0	0	247
medium	176	medium	high	6.7	45	1	1	1	0	245
medium	177	medium	medium	6.8	45	0	0	1	0	255
medium	178	medium	high	6.5	38	1	0	1	0	255
medium	179	medium	medium	6.8	45	0	0	0	1	255
medium	180	medium	medium	6.5	38	1	0	0	0	255
medium	181	medium	medium	6.8	40	1	0	1	0	245
medium	182	medium	medium	6.5	36	0	0	1	0	255
medium	183	medium	medium	6.5	35	1	0	1	0	246
medium	184	medium	medium	6.8	34	0	0	0	1	240

Figure 16: “medium” classification results and attribute data obtained with AdaBoost and Neural Network algorithms

The "low" classification results and attribute data of Milk Quality data with AdaBoost and Neural Network algorithm using Orange Data Mining Tool are shown on Figure 8.

The "low" actual values are shown in the Grade field, and the predicted classification estimation results are shown in the AdaBoost and Neural Network field. In the example with the Id number 747 with the Neural Network algorithm, the quality result that should have been "low" was incorrectly predicted as "high".



Grade	id	AdaBoost	Neural Network	ost	post	st (n	etwc	etw	work	Fold	pH	Temprature	Taste	Odor	Fat	Turbidity	Colour
low	732	low	low	2....	1	2....	0....	0....	0....	4	4.5	38	0	1	1	1	255
low	733	low	low	2....	1	2....	7....	1	2....	1	8.5	70	0	0	0	0	246
low	734	low	low	2....	1	2....	1....	0....	8....	4	7.4	65	0	0	0	0	255
low	735	low	low	2....	1	2....	2....	0....	4....	1	3.0	40	1	1	1	1	255
low	736	low	low	2....	1	2....	0....	0....	2....	4	8.6	55	0	1	1	1	255
low	737	low	low	2....	1	2....	0....	0....	0....	4	4.7	38	1	0	1	0	255
low	738	low	low	2....	1	2....	7....	0....	2....	5	3.0	40	1	1	1	1	255
low	739	low	low	2....	1	2....	0....	0....	0....	2	9.0	43	1	0	1	1	250
low	740	low	low	2....	1	2....	2....	0....	4....	1	3.0	40	1	1	1	1	255
low	741	low	low	2....	1	2....	0....	0....	0....	4	9.0	43	1	0	1	1	250
low	742	low	low	2....	1	2....	0....	0....	0....	4	4.7	38	1	0	1	0	255
low	743	low	low	2....	1	2....	1....	0....	2....	4	3.0	40	1	1	1	1	255
low	744	low	low	2....	1	2....	0....	0....	0....	2	9.0	43	1	0	1	1	250
low	745	low	low	2....	1	2....	0....	0....	0....	1	4.5	38	0	1	1	1	255
low	746	low	low	2....	1	2....	2....	1	4....	5	8.5	70	0	0	0	0	246
low	747	low	high	2....	1	2....	0....	0....	0....	4	6.5	37	0	1	1	1	245
low	748	low	low	2....	1	2....	9....	0....	6....	5	7.4	65	0	0	0	0	255
low	749	low	low	2....	1	2....	8....	0....	0....	3	3.0	40	1	0	0	0	255
low	750	low	low	2....	1	2....	0....	0....	0....	2	9.0	43	1	1	1	1	248
low	751	low	low	2....	1	2....	0....	0....	0....	3	6.6	50	0	0	0	1	250
low	752	low	low	2....	1	2....	8....	0....	0....	3	6.6	50	0	0	0	0	255
low	753	low	low	2....	1	2....	0....	0....	0....	2	9.0	43	1	1	1	1	248
low	754	low	low	2....	1	2....	0....	0....	0....	4	6.6	50	0	0	0	1	250
low	755	low	low	2....	1	2....	0....	0....	0....	5	9.5	34	1	1	0	1	255
low	756	low	low	2....	1	2....	0....	0....	0....	2	5.5	45	1	0	1	1	250
low	757	low	low	2....	1	2....	6....	0....	2....	3	8.1	66	1	0	1	1	255
low	758	low	low	2....	1	2....	1....	0....	4....	3	3.0	40	1	1	1	1	255

Figure 17: “low” classification results and attribute data obtained with AdaBoost Neural Network algorithm

The frequencies of classification distributions of AdaBoost and Neural Network algorithms are shown on figure 9. Blue color indicates "high", Green color indicates "medium" and red color indicates "low" classification. When the results are compared, it is seen that there is only a very small amount of error in the "medium" classification in the AdaBoost algorithm. However, it is seen that there are incorrect results in all classification predictions with the Neural Network algorithm.

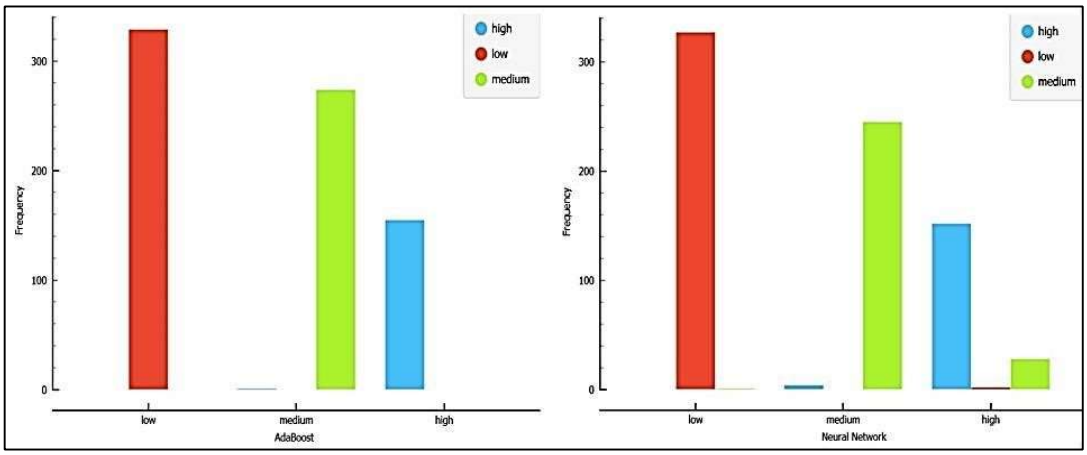


Figure 18:. Classification distributions and frequencies of machine learning algorithms (a) AdaBoost algorithm classification prediction distribution (b) Neural Network algorithm classification prediction distribution

The stability graphs of the algorithms according to the classification estimation results are shown in Figure 10. It is seen that the Adaboost algorithm becomes stable in a very short time. However, the Neural Network algorithm passed steady state slower.

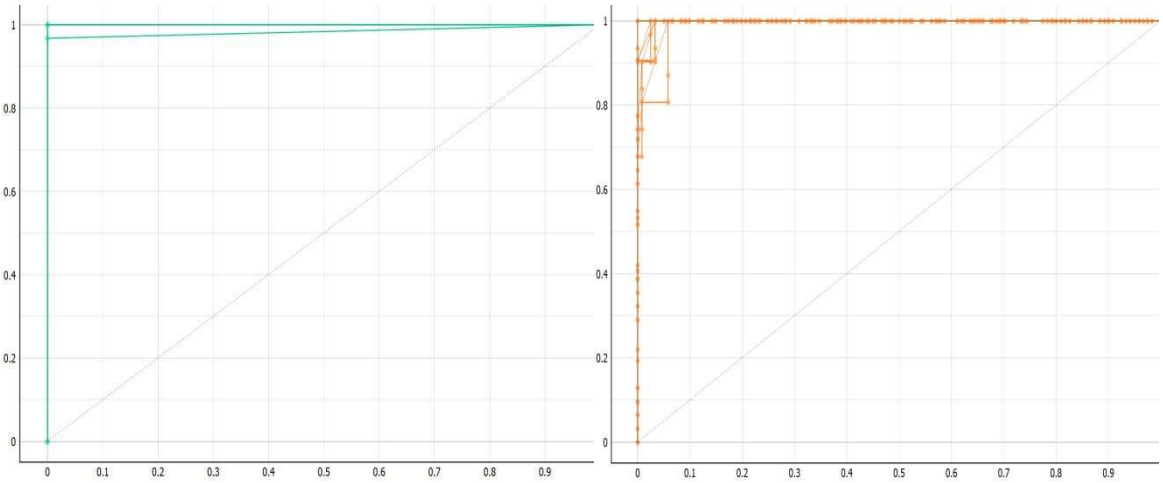


Figure 19: Classification ROC curves of machine learning algorithms (a) ROC curve of AdaBoost Algorithm (b) ROC curve of Neural Network Algorithm

The complexity matrix value graphs of the algorithms according to the classification estimation results are shown on Figure11. With the Adaboost algorithm, only one data belonging to the "high" class was incorrectly predicted as the "medium" class. However, with the Neural Network algorithm, 4 data belonging to the “high” class were incorrectly predicted as “medium”, 28 data belonging to the “medium” class were incorrectly predicted as “high” and 2 data belonging to the “low” class were also “high” was incorrectly predicted.

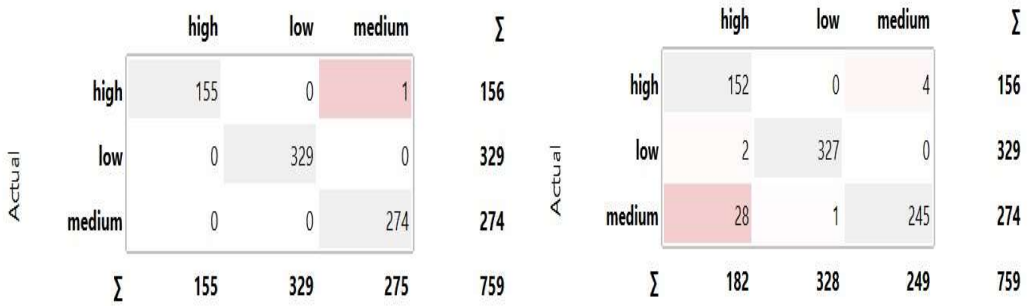


Figure 20: Machine learning algorithms, classification Complexity Matrix values

The metric values of the results obtained with both algorithms are shown on table 2. Area Under Curve (AUC), Classification accuracy (CA), F1 score, Precision and Recall parameters were used as metric values. With the AdaBoost algorithm, 99.9% accuracy estimation was achieved in all metric values. The highest success rate with the Neural Network algorithm was obtained with the AUC metric value, but the CA parameter was used as the success metric in this study. As a CA parameter, 99.9% success rate was obtained with the AdaBoost algorithm and 95.4% with the Neural Network algorithm.

Table 8: Metric success rates of algorithms

Model Name	AUC	CA	F1	Precision	Recall
AdaBoost	0.999	0.999	0.999	0.999	0.999
Neural Network	0.997	0.954	0.955	0.959	0.954

4.3. System Screenshots

This section displays the various graphical user interfaces of the machine learning model used to forecast the quality of beverages. Those interfaces were created using the Python streamlit library. As shown in below graphs:

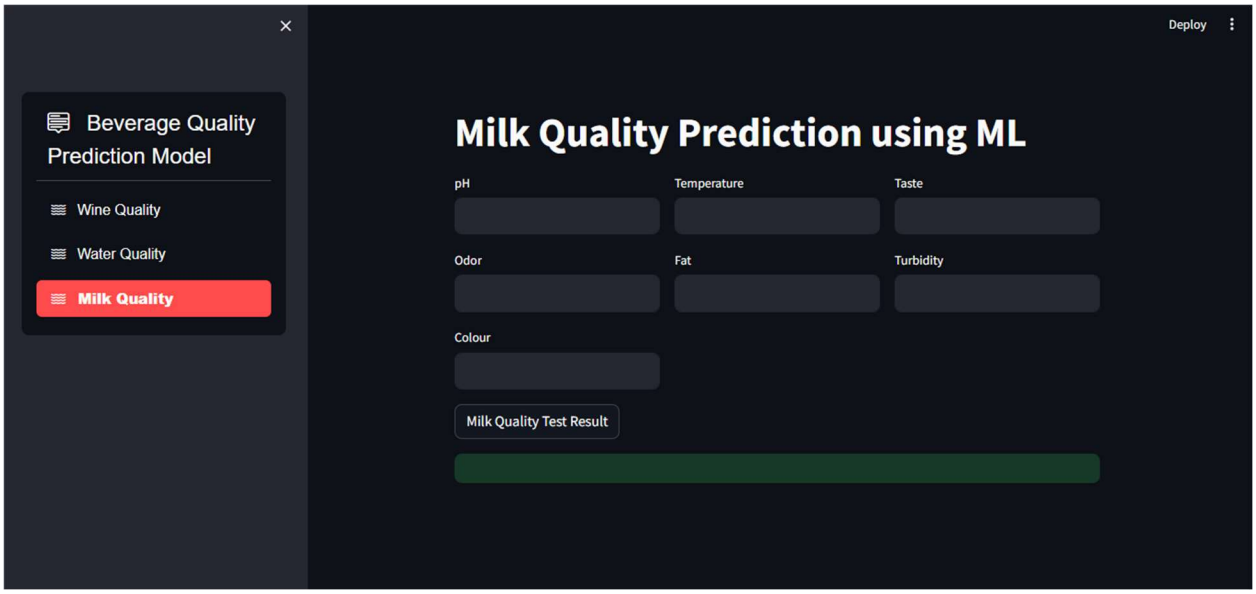


Figure 21: Milk Quality Prediction Page

The milk prediction interface is in Figure 14. The user must fill out all the fields on that page, including pH, temperature, taste, and so on, then click on the milk quality test result to see if the milk is medium, high, or low quality.

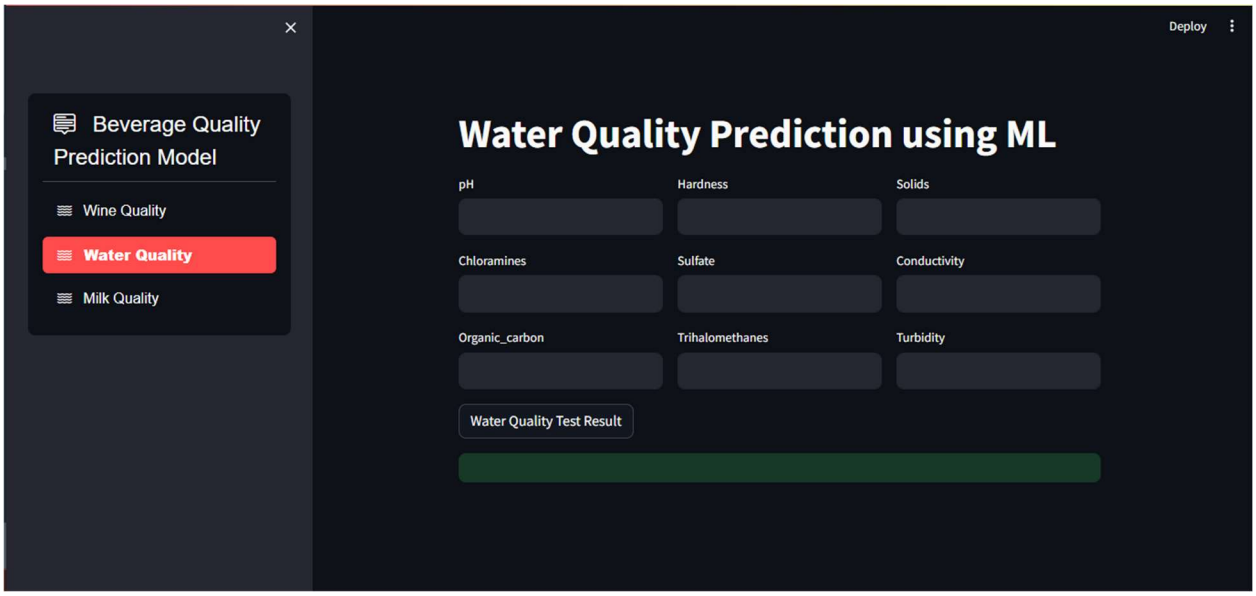


Figure 22: Water Quality Prediction Page

Figure 15 displays the water prediction interface. To determine whether the water is safe to drink, the user must fill out all the fields on the page, including those for ph, hardness, solids, sulfates, turbidity, and so on. Then, they must click the button for the water quality test result.

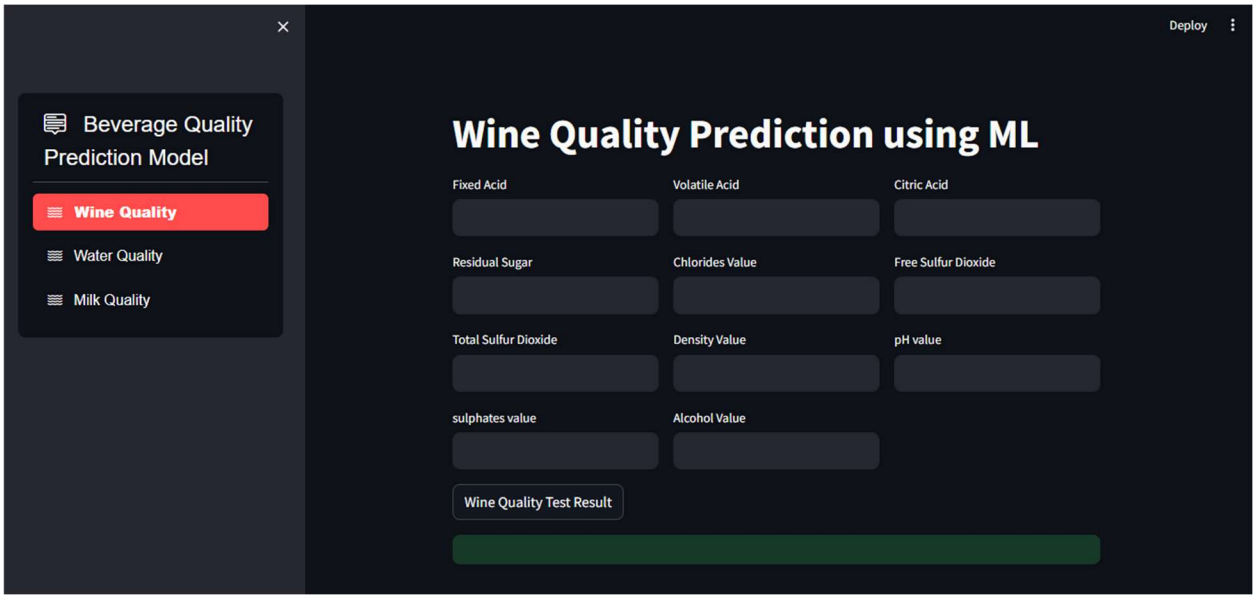


Figure 23: Wine Quality Prediction Page

Figure 16 displays the interface for predicting wine quality. To find out if the wine is safe to drink or not, the user must fill out all the fields on the page, including those for residual sugar, alcohol, citric acid, fixed acid, and volatile acid as well as other fields on the page.

V. 5. CONCLUSION

Introduction

This is the last section of research; the section is divided into two parts. The first part is entitled conclusions which include the contribution that the study has made to knowledge. The second sections deal with the recommendations.

5.1. Conclusion

The intention of this study was to come up with model which is able to predict the quality of beverage as its general objective and it was achieved systematically through addressing these research questions; (1)to determine different existing machine learning algorithms to help in generation of beverage quality prediction model and this was addressed in chapter 2.3, (2)to implement an AI system that can analyze data from beverage production to predict and maintain the quality and this was addressed as shown in system analysis and interpretation(Chapter 4).

In conclusion, the development and implementation of a beverage quality prediction model represent a significant advancement in ensuring consistency and excellence in the production of beverages across various industries. By leveraging machine learning algorithms, we have the potential to gain valuable insights into the intricate relationships between diverse factors influencing beverage quality, ranging from ingredient variations to processing conditions. The success of such a model hinge on the careful curation of high-quality and comprehensive datasets, encompassing the complexities of real-world production scenarios. The journey toward an effective beverage quality prediction model involves continuous refinement, validation, and collaboration between data scientists and domain experts to create a solution that aligns with industry standards.

5.2. Recommendations

5.2.1. To society

The adoption of beverage quality prediction models presents a transformative opportunity for society, offering benefits that extend beyond the industrial realm. The first and most important reason for quality control in beverage manufacturing is the protection of public health. There is a need for increased awareness and education regarding the implications of these models on product quality and safety.

5.2.2. To Industry

The integration of beverage quality prediction models represents a game-changing opportunity for the industry to elevate its standards and operational efficiency. Thus, beverage testing throughout the manufacturing process helps ensure product quality and consumer safety. By using this model companies can significantly reduce their production costs by conducting effective inspections and controls during production and operations in the beverage industry. Real-time monitoring and integration of predictive models into existing quality control processes can enable proactive decision-making and adjustments, minimizing quality variations. Additionally, fostering a culture of continuous learning and adaptation is crucial. Industry stakeholders should stay abreast of advancements in machine learning techniques and regularly update their models to remain at the forefront of quality assurance.

5.2.3. To Researcher

For researchers involved in the development and enhancement of beverage quality prediction models, several avenues can contribute to the success and impact of their work. Researchers should actively engage with beverage manufacturers and quality control experts to ensure that the models are not only scientifically robust but also practically applicable to real-world production settings. Furthermore, the exploration of advanced machine learning techniques, such as deep learning and ensemble methods, can enhance the predictive capabilities of models. Continuous validation and benchmarking against industry standards would be essential to ensure the real-world applicability of research findings. By addressing these recommendations, researchers can contribute significantly to the development of robust and applicable beverage quality prediction models, advancing both scientific knowledge and industry practices.

5.3. Future Work

In the future development of a beverage quality prediction model, a key area of focus should be the integration of advanced sensor technologies and real-time data monitoring. Furthermore, to enhance the model's adaptability across different beverage types and production processes, future work should involve the development of a modular and customizable architecture. Moreover, the implementation of self-learning mechanisms that continuously update the model based on new data and user feedback can ensure its longevity and relevance in dynamic beverage production environments.

REFERRENCES

[1]. Activestate. (2020, October 9). *What Is Pandas in Python? Everything You Need to Know*. Retrieved from Activestate: https://www.activestate.com/resources/quick-reads/what-is-pandas-in-python-everything-you-need-to-know/

[2]. Alabi, O. A. (2020). Production usage, and potential public health effects of aluminum cookware. *Annals of Science and Technology*, 20-30.

[3]. Altexsoft. (2019, December 18). *Best Public Datasets for Machine Learning and Data Science: Sources and Advice on the Choice*. Retrieved from Altexsoft: https://www.altexsoft.com/blog/best-public-machine-learning-datasets/

[4]. Ambadipudi, R. (2023, February 27). *How Machine Learning Would Transform Your Industry*. Retrieved from Forbes: https://www.forbes.com/sites/forbestechcouncil/2023/02/27/how-machine-learning-will-transform-your-industry/?sh=1fac9f4e1a3b

[5]. Bhandari, P. (2023, June 22). *What Is Quantitative Research? | Definition, Uses & Methods*. Retrieved from Scribbr: https://www.scribbr.com/methodology/quantitative-research/

[6]. Bhatt, S. (2018, March 19). *Reinforcement Learning 101*. Retrieved from towardsdatascience: https://towardsdatascience.com/reinforcement-learning-101-e24b50e1d292

[7]. Chellappa, R. K., & Saraf, N. (2010). Alliances, rivalry, and firm performance in enterprise systems software markets. *Information Systems Research*, 849-871.

[8]. Coles, R. &. (2011). Food and beverage packaging technology.

[9]. Ed Burns. (2021, march). *machine learning*. Retrieved from Techtarget: https://www.techtarget.com/searchenterpriseai/definition/machine-learning-ML#:~:text=Machine%20learning%20(ML)%20is%20a,to%20predict%20new%20output%20values.

[10]. FoodStuff. (2016). *Rwanda tightens rules on food safety*. Retrieved from FoodStuff-Africa: https://foodstuff-africa.com/rwanda-tightens-rules-food-safety/

[11]. Gandhi, R. (2018, Jun 7). *Support Vector Machine — Introduction to Machine Learning Algorithms*. Retrieved from Towardsdatascience: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[12]. Georgia Pratt. (2022, November 18). *Industry Overview: Food and Beverage* . Retrieved from crowcon: https://www.crowcon.com/blog/industry-overview-food-and-beverage/

[13]. Gillis, A. s. (2023, July 2). *supervised learning*. Retrieved from techtarget: https://www.techtarget.com/searchenterpriseai/definition/supervised-learning

[14]. Kalpana, V. N. (2019). In Preservatives and preservation approaches in beverages. In *Preservatives in beverages* (pp. 1-30). Academic Press.

[15]. Mbaabu, O. (2020, December 11). *Introduction to Random Forest in Machine Learning*. Retrieved from section: https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

[16]. Mcleod, S. (2023, May 15). *Questionnaire: Definition, Examples, Design And Types*. Retrieved from simplypsychology: https://www.simplypsychology.org/questionnaires.html

[17]. MOMOH, O. (2023, April 29). *Population Definition in Statistics and How to Measure It*. Retrieved from Investopedia: https://www.investopedia.com/terms/p/population.asp#:~:text=Investopedia%20%2F%20Matthew%20Collins-,What%20Is%20Population%3F,is%20drawn%20for%20a%20study.

[18]. Nelli, F. (2015). Data analysis and science using PANDAs, Matplotlib and the Python Programming Language. *Python data analytics*.

[19]. Pykes, K. (2023, March 4). *Introduction to Unsupervised Learning*. Retrieved from datacamp: https://www.datacamp.com/blog/introduction-to-unsupervised-learning

[20]. Sileyew, K. J. (2019). *Research design and methodology.* Cyberspace.

[21]. Singh, V. (2020, February 28). *How Machine Learning Is Changing the World*. Retrieved from Datasciencecentra: https://www.datasciencecentral.com/how-machine-learning-is-changing-the-world/#:~:text=Machine%20learning%20is%20changing%20the%20world%20by%20transforming%20all%20segments,shopping%2C%20food%20ordering%2C%20etc.

[22]. Tan, W. K. (2023). Recent technology for food and beverage quality assessment. *Food Science and Technology*, 1681-1694.

[23]. Tutorialspoint. (2023, January 2). *What are the differences between Python and an Anaconda*. Retrieved from Tutorialspoint: https://www.tutorialspoint.com/what-are-the-differences-between-python-and-an-anaconda

[24]. Butler, T., & Fitzgerald, B. (2019). "Unpacking the systems development process: an empirical application of the CSF concept in a research context. *The Journal of Strategic Information Systems*, 351–371.

[25]. Chris, P. (2021). What is a Machine Learning Model? *The Journal of Machine Learning Research*, 10-14.

[26]. Dan, R. (2020). An Overview of The Anaconda Distribution. *Towards Data Science*, 57-65.

[27]. jake, f. (2022). What Is Artificial Intelligence (AI). *Investopedia*, 45-47.

[28]. Jay, P. (2021). Agile Development Applied to Machine Learning Projects. *Facilitating the Spread of Knowledge and Innovation in Proffesional Software Development*, 94-97.

[29]. Kate, B., & Valerie, S. (2020). Agile Software Development. *TechTarget*, 49.

[30]. Kauffman, R. J., & Wang, B. (2017). The success and failure of dotcoms: A multimethod survival analysis. *proceedings of the 6th INFORMS Conference on Information Systems and Technology (CIST)*, 3–4.

[31]. Pádraig, C., Matthieu, C., & Sarah, J. (2019). Supervised Learning. In C. Matthieu, & C. Pádraig, *Machine Learning Techniques for Multimedia* (pp. 21-49). Berlin: Springer Science & Business Media.

[32]. Pavlo, S. (2021). Machine Learning in Banking: Top Use Cases. *SDK*, 61-67.