Utilizing Machine Learning Techniques for Excellent Coffee Prediction

MSc Research Project
Data Analytics

Methmi Kaveesha Melewwe Thantrige
Student ID: x22156879

School of Computing
National College of Ireland

Supervisor: Prof. Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Methmi Kaveesha Melewwe Thantrige
Student ID:	22156879
Programme:	Data Analytics
Year:	2023 - 2024
Module:	MSc Research Project
Supervisor:	Prof. Jorge Basilio
Submission Due Date:	31st January 2024
Project Title:	Utilizing Machine Learning Techniques for Excellent Coffee Prediction
Word Count:	6525
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Methmi Kaveesha Melewwe Thantrige
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	✓
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	✓
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	✓

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Utilizing Machine Learning Techniques for Excellent Coffee Prediction

Methmi Kaveesha Melewwe Thantrige
22156879

Abstract

As universal coffee utilization rises, the coffee industry remains highly valued on sustaining and rising coffee quality. To this is added the requirement to understand the essential components affecting the quality of coffee accurately. By bridging the gap between traditional cupping techniques for coffee quality detection, this study utilizes modern machine learning approaches to explore the sensory as well as non-sensory attributes shaping coffee quality. Three machine learning models namely, Random Forest, Gradient Boosting Machine, and Support Vector Regressor, along with a Hybrid Model which combines all three algorithms, are used for predictive analysis and the models are evaluated using evaluation metrics namely, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and coefficient of determination (R Squared). The outcomes deliver valuable perceptions into the features critical for coffee quality evaluation. Results express that among all four predictive models, Support Vector Regression performance is best in forecasting coffee quality by settling good generalizations on unseen or trained data.

Keywords: Machine Learning Models, Quality Prediction, Non-sensory Attributes, Support Vector Regression, Gradient Boosting Machine, Random Forest, Hybrid Model.

1 Introduction

1.1 Background

The coffee industry globally stands as one of the keystones of the world economy, rolling into daily routines, social meetings, as well as international market. As of 2023, coffee has conquered its spot among the most imported agricultural products internationally, sustaining the livelihoods of millions (Vijayan et al. 2022). Nonetheless, the industry's sustainability and succession spin around the level of quality of the coffee it plants and sells. Coffee, for buyers, goes beyond being a regular beverage; it's an experience, with extraordinary quality being the main motivator of this trip through the senses.

Quality in coffee is a multi-aspect thought, manipulated by numerous factors involving the coffee's variety, its country of origin, as well as processing techniques to the complex sensory factors such as aroma, flavor, aftertaste, bitterness, body, and balance. Holding on to the complexities of these components and their extreme effect on the product holds a major concern for all involved in the coffee supply chain, from farmers and makers to suppliers and buyers. This perception provides stakeholders with the right tools to boost their procedures, adapt novelty, and hold a competitive edge in a marketplace where consumers are progressively trained and observing their food preferences.

1.2 Significance of the Study

Despite the apparent significance of coffee quality, the procedures for its assessment continue to be varied and occasionally subjective. Customary cupping sessions, while valued, may lack the complexity required to understand the sophisticated chemistry of factors affecting quality. Furthermore, evolving trends in sustainability, changing climate situations, and evolving processing methods require a more sophisticated methodology for coffee quality analysis.

This study is motivated by the requirement to bridge existing gaps in the understanding of coffee quality and to contribute to the current discussion within the coffee industry. By utilizing advanced data analytics and machine learning algorithms such as Random Forest, Gradient Boosting Machine (GBM), Support Vector Regression (SVR) as well as Hybrid approach, this research aims to crack the complex relationships between several features and the ultimate sensory feel of coffee users. Such perceptions hold the potential to reform quality control procedures, foster the development of new and modern coffee varieties, and encourage participants throughout the coffee value chain.

1.3 Problem Statement

The coffee industry is faced with a critical challenge in precisely assessing and improving coffee quality due to the nature of traditional cupping procedures and the dynamic landscape of agriculture practices. As consumer preferences grow and sustainability becomes vital, there is a critical need for an advanced, data-driven approach to broadly analyze the non-sensory characteristics of coffee. Existing approaches often fall short in acquiring the complicated chemistry of aspects such as processing method, variety, and country of origin. This research project aims to bridge this gap by utilizing advanced data analytics and machine learning techniques to present actionable insights for farmers(manufacturers), processors, and investors. Through a complete analysis, the research seeks to enhance quality control processes, perfect production methods, and meet the changing requirements of consumers, contributing to the overall progression and sustainability of the coffee industry.

1.4 Research Goal

The major goal of this study is to utilize advanced data analytics and machine learning techniques for a broad analysis of coffee quality, surpassing the limitations of traditional cupping procedures. By analytically evaluating the non-sensory characteristics of coffee, involving factors such as variety, processing methods, and geographical location, the research intends to develop a subtle understanding of the complex factors influencing coffee quality. This research aspires to contribute to the optimization of coffee quality evaluation and support the growing needs of consumers and the coffee industry at a great scale.

1.5 Research Objectives

1. To Implement Predictive Models: Develop effective predictive models that forecast coffee quality based on these fundamental features, providing a model for coffee investors to evaluate and boost coffee quality in their strategies.

2. To Recognize Significant Quality Factors: Analyze historical coffee quality data to identify the fundamental factors that constantly contribute to high-quality coffee among diverse parameters like country of origin, processing methods, and coffee varieties.

1.6 Research Questions

1. How can advanced machine learning and data analytics techniques enhance the accuracy and efficacy of coffee quality evaluation?

1.7 Paper Structure

The subsequent chapters will dive into the Literature Review which discusses the related studies that have been carried out by academic researchers in Chapter 2 and Chapter 3 discusses the methodology which presents the entire analysis process carried out in the research. Chapter 4 is dedicated to presenting the Design Specification followed by Implementation the Chapter 5 and Evaluation in Chapter 6. Finally, in chapter 7 it focuses on establishing the conclusion to the research project and potential future work.

1.8 Summary

The background of this study dives into the complex world of coffee quality estimation, investigating the combination of traditional cupping methods and innovative machine learning techniques. It sets the stage by emphasizing the progressing landscape of the coffee industry and the requirement for more sophisticated quality assessment processes. The significance of the study lies in its ability to transform how coffee is perceived and assessed, offering an understanding of the composite relationship of components shaping its quality. The problem statement highlights the existing gaps in current approaches, motivating the hunt for a broad understanding of the sensory as well as non-sensory characteristics influencing coffee quality. The research goal is to boost coffee quality evaluation through a data-driven approach utilizing machine learning techniques. The research objectives highlight the key dimensions to be surveyed, involving the impact of varieties, processing methods, and geological origins. The subsequent research questions guide the research. Lastly, the paper structure presents a brief overview of how the paper is organized, mapping out the sequential flow of material in the subsequent segments.

2 Literature Review

This chapter explores an exploration of prior studies regarding coffee and its quality attributes. Several research works have studied various elements of coffee and its fundamental characteristics, aiming to recognize how each component influences coffee quality, mostly within the fields of agriculture as well as food science. Nonetheless, there is an evident deficiency in applying data analytics techniques, especially machine learning and deep learning methods, for the comprehensive analysis of these quality features. However, these analytical techniques have mostly been applied for coffee bean recognition and classification through image recognition. Nevertheless, the insights gained from these studies serve as a valuable means for this study.

This sector is ordered into three sub-sectors:

1. The first section analyses and interprets comparisons between studies that examine the facts of coffee quality influences.
2. The subsequent section focuses on how genetic and farming components affect the overall quality of coffee.
3. The final part studies the application of machine learning and deep learning techniques in the classification of coffee beans.

Within each of these three portions, specific paragraphs separate and evaluate a few related studies, explaining their conclusions and contributions to the wider understanding of coffee quality characteristics.

2.1 Complexity of the Factors on Coffee Quality

Hinsene et al., Paulo et al., and Giselle et al. have all focused their investigations on different facets of the quality of coffee. In their research, (Garuma et al. 2014) examines the impact of various coffee production systems in Ethiopia on the prevalence of bean defects and the overall quality of the coffee. The researchers classify coffee production into four distinct systems: forest, semi-forest, garden, and cultivation. Within each system, they assess the incidence of peaberry, which is a notable aberration in beans. According to their findings, the rate of bean defects is influenced by the production system chosen, with cultivation producing a higher percentage of peaberry anomalies.

Paulo and colleagues (Toledo et al. 2016) provide an exhaustive examination of a wide range of variables linked to the quality of coffee beans and the frequently uncertain attributes they possess. The variables that their research investigates include species of coffee, origin, storage, roasting, processing methods, and defects in the beans. In addition, a comprehensive examination of the current body of literature is undertaken to decipher the complex relationship between the quality of coffee seeds and the volatile components that contribute to aroma. The research conducted by the authors sheds light on discrepancies in the aroma compositions of Robusta and Arabica coffee, the effects of after-harvest processing on coffee aroma, and the consequences of the storage environment on volatile substances. Furthermore, both studies investigate the impact that coffee cultivars and genera have on its quality. The initial investigation posits that the establishment and dissemination of peaberry within the monoculture plantation system could be ascribed to particular genotypes of coffee shrubs. On the contrary, the second study investigates how variations in the concentration of specific aroma compounds among Arabica and Robusta coffee result in noticeable distinctions in flavor.

In summary, both investigations substantially augment our understanding of the quality of coffee from diverse perspectives. In contrast, Hinsene et al. investigate the effects of various production systems on defects in coffee beans, whereas Paulo et al. clarifies the complex array of elements that impact the aromatic characteristics of coffee. As a whole, these studies highlight the complexity of the factors that influence the quality of coffee.

2.2 Genetic Deviation and Farming Strategies for Enhancing Coffee Quality

An investigation was conducted by (Malau 2018) to examine the genetic and phenotypic variations in the sensorial attributes of Arabica coffee among twenty-eight distinct genotypes. The research results indicate that the majority of characteristics, such as aroma, flavor, residue, bitterness, body, consistency, equilibrium, sweetness, and overall evaluation, demonstrate minimal genetic and phenotypic diversity. The limited heritability and genetic progress in these characteristics pose difficulties when it comes to the selection of coffee varieties according to their sensory qualities. Prior investigations have proposed increased variability, which suggests that crossbreeding may offer a means to achieve improvement. However, this research emphasizes the significance of genotypic diversity as a fundamental aspect in the development of coffee profiles that are more varied and exceptional.

On the other hand, (Ferreira et al. 2019) present a comprehensive analysis that establishes the quality of coffee as a critical factor in the pursuit of improving worldwide coffee production. Their primary focus is on the complex interaction that occurs among genetics, environmental elements, and after-harvest circumstances. The research emphasizes substantial genetic diversity in the composition of beans and sensory characteristics, both among and among distinct coffee species. This highlights the successful implementation of quality enhancement techniques by employing interspecies and crossbreeding approaches, specifically in the case of Arabica and Robusta. Arabica prioritizes the maintenance of high-quality standards while concurrently improving disease resistance and harvest optimization. The primary objective of Robusta is to enhance quality by employing within-a-species and interspecies crosses, to achieve particular bean sizes and molecular combinations that elevate the overall quality of the cup. To improve reproductive efficiency, the study additionally presents genomic tools, including expressed sequence tags (ESTs) and genetic maps, which exhibit the potential to facilitate marker-guided choice and gene conversion.

In brief, both research studies concur regarding the critical significance of coffee quality as perceived by consumers, as well as the indispensable function that genetic diversity plays in the emergence of unique coffee attributes. The research team, led by Malau, investigates the variation in sensory attributes among distinct Arabica genotypes. In contrast, Ferreira et al. take a more comprehensive approach by analyzing the genetic elements that regulate the differentiation in coffee quality between Robusta and Arabica. Both research studies emphasize the enormous possibilities of hybridization and the application of genetic tools in the pursuit of improving agricultural productivity and coffee quality.

2.3 Usage of Deep Learning and Artificial Intelligence on Coffee Classification

The challenges faced in coffee production, specifically in the domain of coffee bean categorization, are examined in the following three studies: (Micaraseth et al. 2022) present an automated coffee bean inspection system built upon a conveyor-mounted camera and Raspberry Pi 4. For image recognition, the authors employ various deep learning approaches, which include Enhanced, ResNet-50, and AlexNet. ResNet-50 achieves the highest accuracy among the models evaluated. Among these, ResNet-50 attains the most elevated accuracy rate of 93.33%. By effectively differentiating defective coffee beans from normal ones, this streamlined system significantly decreases the need for human labor and manufacturing time, ultimately increasing the worth of the coffee beans.

(Huang et al. 2019) conducted research where they introduced an automated coffee bean harvesting system. This system makes use of image processing, data augmentation, and a convolutional neural network (CNN). The CNN model is trained using grayscale photos. The identification system, which relies on a camera, has an outstanding overall accuracy rate of 93.34% in identifying coffee beans. Additionally, it has a very low false positive rate of 0.1007. This technology's ability to automate selection processes results in significant time and manpower cost savings, providing substantial advantages to the specialty coffee market.

In 2022, (Febriana et al. 2022) provided the USK-Coffee dataset, which consists of 8,000 photos of green coffee beans categorized into four distinct groups. The research attains overall accuracy in classification around 81.31% and 81.12% using deep learning models like MobileNetV2 and ResNet-18, respectively. While the accuracy rates in these trials are somewhat lower than in earlier ones, the use of deep-learning algorithms on the dataset lays a strong groundwork for future study in coffee bean sorting.

To summarise, these researches jointly emphasize the ability of deep learning and artificial intelligence to improve the effectiveness as well as the accuracy of coffee bean categorization. Advancements such as real-time production line apps, webcam-based recognition systems, and the generation of huge datasets help to make coffee manufacturing processes more efficient, reduce jobs that need a lot of manpower, and eventually improve the overall quality of coffee beans.

The coffee business faces major challenges arising from the effects of climate change insects, and illnesses, which have a substantial influence on both output and quality. The research undertaken by (Aunsa-Ard & Kerdcharoen 2022) investigates the suitability of an electronic nose (e-nose) in discerning the fragrances of coffee originating from different locations in northern Thailand. The e-nose technology effectively distinguishes coffee smells impacted by parameters such as altitude, processing techniques, and roasting temperatures by using eight metal oxide semiconductor gas sensors and using PCA for pattern recognition. This research emphasizes the significance of meticulous sample handling and ideal e-nose settings for exact measurements, and it showcases the e-nose's potential in quality checks and tracking applications in the coffee sector.

Furthermore, (Vijayan et al. 2022) specifically concentrates on the output of coffee in South India. To evaluate the data from 1990 to 2018, Keerthi used the ARIMA time series model. The analysis predicts a small rise in Arabica output and a minor decline in Robusta production starting in 2021. The ARIMA model with parameters (1, 0, 2) is considered appropriate for analyzing Arabica data, while the ARIMA model with parameters (1, 1, 1) is used for studying Robusta production. This study provides significant observations on arising patterns in coffee products in the local region.

Combined, these researches provide valuable insights and advancements to the coffee business. Wandee and their team highlight the capabilities of e-nose technology in identifying and improving the quality of aromas. Meanwhile, Keerthi and their team make important forecasts about the future developments in coffee production in South India. Overall, these papers were evaluated to comprehend the use of machine learning methods in this domain.

2.4 Research Gap

The review of the literature examines the current research on coffee quality, including several aspects such as aromatic identification, genetic adaptability, and the use of machine learning for coffee bean classifying. Previous research has examined certain factors that affect the quality of coffee. However, there is a lack of utilization of modern data analytics approaches, such as machine learning or deep learning approaches, to comprehensively assess and improve coffee production and quality.

The research focus is on examining the efficacy of sophisticated data analytics methods in the coffee industry as a whole. The research aims to identify crucial characteristics and successful strategies that might enhance the quality of coffee across various regions, processing methods, and varieties of coffee, utilizing historical data on coffee quality. This study aims to bridge the gap between conventional methodology and cutting-edge data-driven approaches, therefore introducing a fresh aspect to the analysis and improvement of coffee quality.

This research aims to provide a comprehensive knowledge of the many aspects that impact coffee quality by using modern analytics approaches. The goal is to identify the most important factors and their connections, providing insight into the intricate characteristics that impact coffee attributes. Moreover, the impact of this study extends beyond the realm of academia, since its findings and insights may be directly applied to the coffee business. The research seeks to make a contribution by using the capabilities of data analytics.

3 Research Methodology

This section covers the methodology followed to implement the predictive model. The research method used in this study is KDD (Knowledge Discovery in Databases) (Marbán et al. 2009) which is a comprehensive technique that links machine learning, data mining, statistics, and more to bring out meaningful insights from large datasets. The method contains data preparation and cleaning, data preprocessing, data mining, and afterward visualizing the outcomes. Certain procedures such as clustering, classification, association rule mining, and discovery of outliers are used at several stages to demonstrate hidden patterns and relations to support accurate decision-making, and advancing various operations. Figure 1 1 illustrates the flow of the methodology.

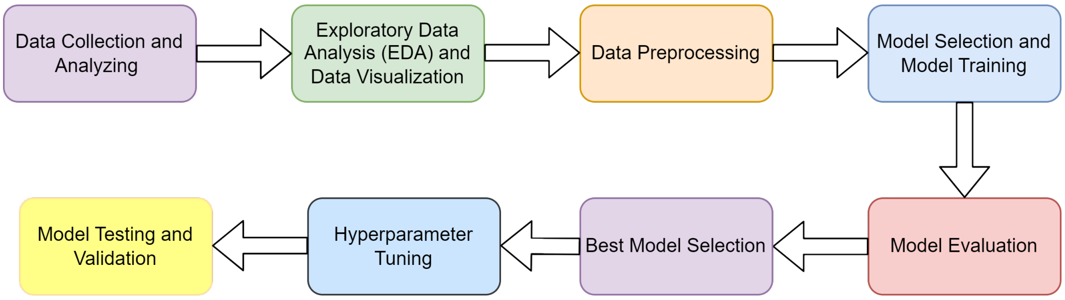
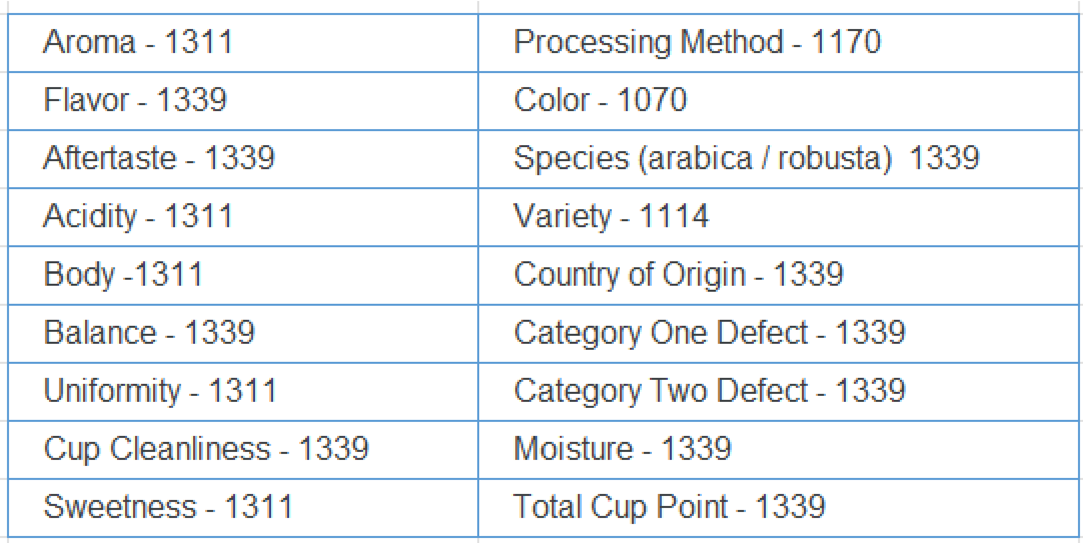

Data Collection and Analyzing
Exploratory Data Analysis (EDA) and Data Visualization
Data Preprocessing
Model Selection and Model Training
Model Evaluation
Best Model Selection
Hyperparameter Tuning
Model Testing and Validation

Figure 1: Methodology Flow.

3.1 Data Collection and Preparation

This is an essential step since the quality and amount of data collected will have a direct impact on the predictive model's accuracy (Mariscal et al. 2010). The data collected is then compiled and classed as Training Data. After arranging training data, the subsequent step is data preparation, which requires importing the data into a suitable location and getting it organized for use. Additionally, the data is separated into two sets. Most of the data set will be used to train the predictive model using the first section, and the second will be used to evaluate the trained model's performance.

The dataset was obtained from the Kaggle platform which is an open-source website [1]. The dataset was in two different CSV files one file having data regarding Arabica coffee and the other containing Robusta coffee data. For this study purpose, both files were merged and the final dataset was renamed as 'Coffee'.u The coffee dataset consists of 1340 rows and 44 columns. However, for this research, it will use only 18 attributes which is shown in Figure 2 with each count of attributes.



Aroma - 1311	Processing Method - 1170
Flavor - 1339	Color - 1070
Aftertaste - 1339	Species (arabica / robusta) 1339
Acidity - 1311	Variety - 1114
Body -1311	Country of Origin - 1339
Balance - 1339	Category One Defect - 1339
Uniformity - 1311	Category Two Defect - 1339
Cup Cleanliness - 1339	Moisture - 1339
Sweetness - 1311	Total Cup Point - 1339

Figure 2: Coffee Attributes.

3.2 Exploratory Data Analysis

Visualizing data to recognize and eliminate unnecessary features is known as exploratory data analysis, or EDA (Behrens 1997). Both visual and mathematical approaches can be used for exploratory data analysis. For a data analyst to initiate analyzing the data, this is an essential tool. It assists analysts in making sense of the dataset by offering insights. Which machine learning algorithms to use in an analysis is a decision that could be made by an analyst using EDA. It also assists analysts in choosing the best technique to apply while studying data. It enables figuring out how unique dataset features connect. For data analysis study, EDA may hence be considered one of the most crucial methods.

[1] CQI Dataset: `https://www.kaggle.com/datasets/volpatto/coffee-quality-database-from-cqi`

3.2.1 Coffee Quality

The majority of coffees in the dataset have a total cup point score between 75 and 80, with a peak at 82.5. This suggests that most coffees are of good quality, but there are some that are of exceptional quality (above 85 points) and some that are of lower quality (below 70 points). By analyzing the below distribution shown in Figure 3, it is evident that the dataset is biased toward higher-quality coffees.

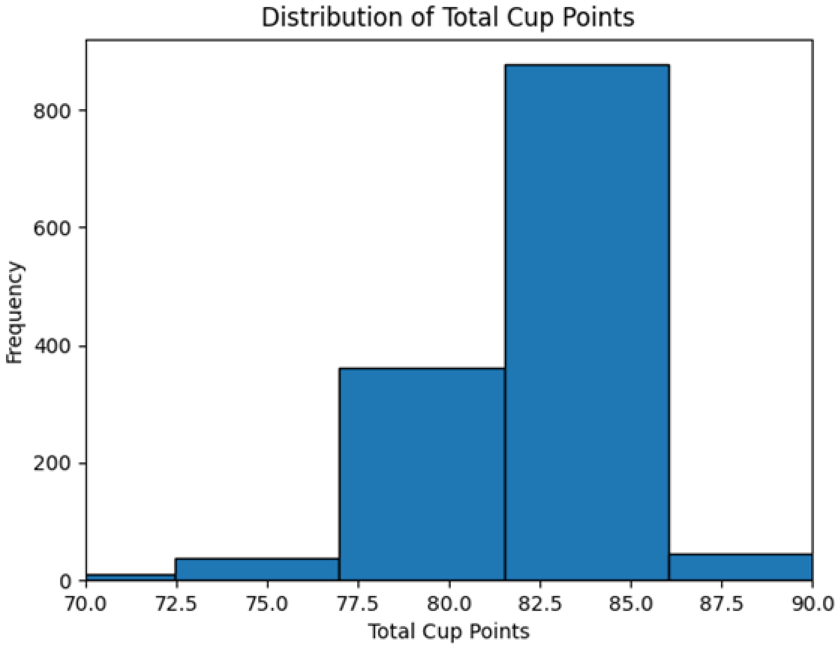


Figure 3: Coffee Quality Distribution

3.2.2 Processing Method

First, the methods that are available in the 'processing.Method' variable were analyzed, and the following methods were found,

Washed / Wet, Natural / Dry, Semi-washed / Semi-pulped, Other, Pulped natural / honey.

According to Figure 4, the distribution of processing method majority of coffee belongs to the washed/wet method however when analyzing the impact on quality natural/dry seems to be the better method.

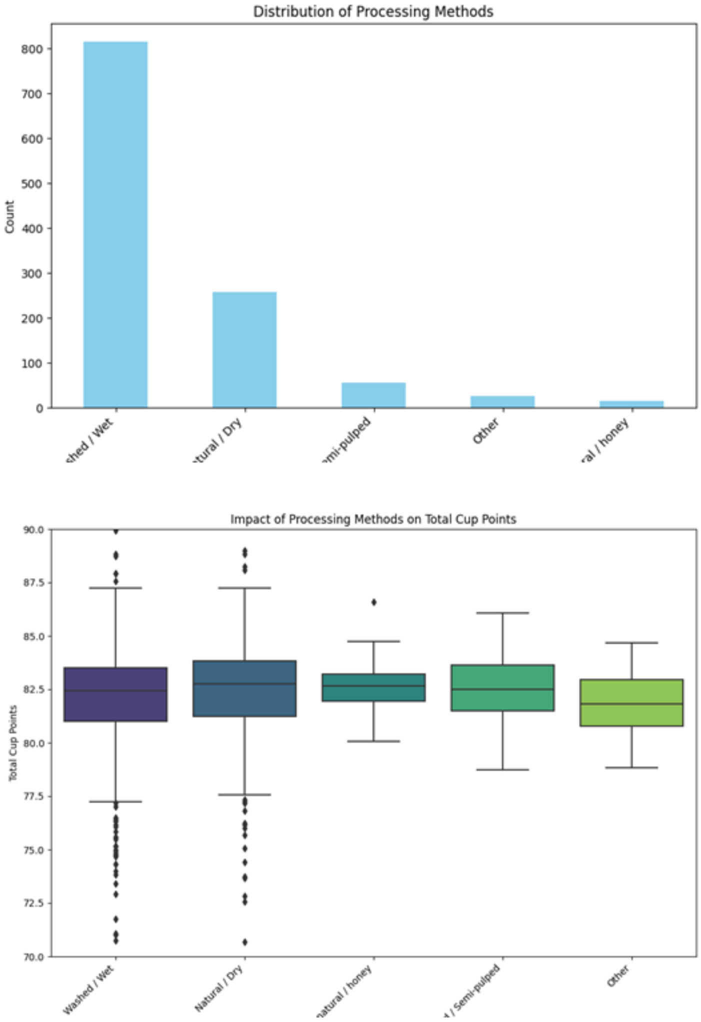


Figure 4: Processing Method Frequency and Impact on Quality

3.2.3 Variety

The varieties are sorted according to their occurrence. All the varieties which had a frequency smaller than 5 and the unknown varieties were put into the new category called ‘Other’.

The most common variety appeared to be Caturra, followed by Bourbon and Typica as shown in Figure 5. The least common varieties are Arusha, Camara, and SL35. When considering the impact on the quality of coffee SL34, SL26, and Bourbon seem to be having higher mean.

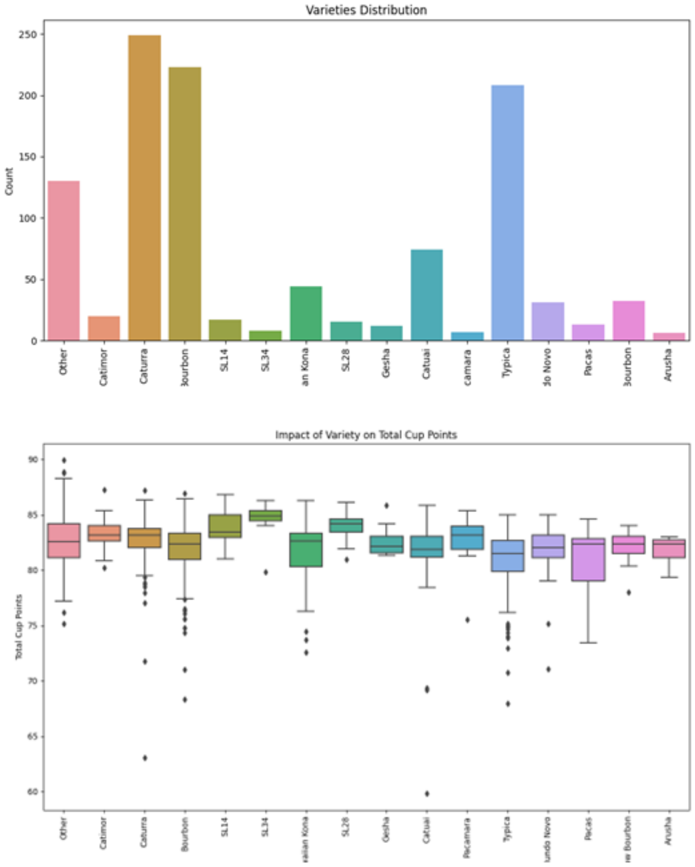


Figure 5: Variety Count and Impact on Quality

3.3 Data Preprocessing

3.3.1 Handling Outliers and Missing values

Abnormal data points in a dataset are known as outliers, and if they remain in the dataset, they can significantly affect the analysis of statistical inferences (Mishra et al. 2020). Furthermore, it often goes against the assumptions of statistical analysis and data sets. Thus, dealing with outliers by removing those are essential step in data analysis.

Afterward, missing values were handled to ensure the cleanliness of the dataset. Missing values for numerical features were managed by filling it with the mean value of the column whereas, for categorical variables, missing values were managed by the most frequent value. Once the process was done cleaned data was obtained to utilize for further analysis process.

3.3.2 Feature Engineering

Feature engineering is an essential step in the data preprocessing pipeline that contains transforming and creating new attributes to boost the performance of machine learning models (Dong & Liu 2018). It seeks to provide the models with more important or meaningful and relevant information, letting them better find patterns and connections within the data. In this study to improve the predictive model's performance One Hot Encoding was performed on categorical variables namely 'Processing Method', 'Variety', and 'Country of Origin'.

3.4 Model Building and Training

The idea of data mining depends strongly on data mining models. These are the virtual structures that are used to classify data so that predictive analysis be done properly (Mariscal et al. 2010). Although they appear to have a shape related to data tables, data mining models are profoundly distinct from data tables. Tables are meant to denote real data sets; nonetheless, data mining procedures are used to interpret data and get meaningful insights. In this research project, four machine learning models were implemented namely, Random Forest, Gradient Boosting Machine(GBM), Support Vector Regression(SVR), and lastly hybrid model which is a combination of all three models. These models are discussed in detail under Chapter 5 Implementation.

3.5 Interpretation of Data and Evaluation

The predictive performance of each model was measured on both the training and testing sets. Once the prediction is done each model's performance is evaluated using evaluation metrics namely, Mean Squared Error(MSE), Mean Average Error(MAE), and coefficient of determination (R squared). Comparative analysis of evaluation metrics delivered insights into the efficiency or effectiveness of each model in predicting the quality of coffee in other words 'Total Cup Points'. More regarding results and comparisons are discussed in Chapter 6 Evaluation.

4 Design Specification

All the techniques that have been used in this study and the predictive model usage are discussed in this chapter. The research project is implemented in Python language using PyCharm integrated development environment (IDE). The project design is shown in Figure 6

4.1 Phase 1

In the initial stage, the data is discovered and gathered from the open-source machine learning website called Kaggle. Arabica and Robusta Coffee data are selected and merged for use in this study. Afterward, collected data is fetched into the Pycharm IDE for further analysis process. Data pre-processing is performed on the dataset, which includes processes such as removing outliers and missing values, and feature engineering using one hot encoding. Cleaned data was necessary for the machine learning algorithms as the study's purpose was to predict coffee quality. Hence, cleaned data was produced.

4.2 Phase 2

The cleaned and preprocessed data are then split into two separate sets training dataset which contains 80 percent of cleaned data and the rest into the test dataset. Afterward, the regression task is carried out using four different machine learning algorithms namely, Random Forest, Gradient Boosting Machine(GBM), Support Vector Regression(SVR), and Hybrid Model which is a combined model of all three models mentioned. The hybrid model was created to compare the results properly and present the best predictive model for coffee quality prediction.

4.3 Phase 3

In this phase, produced results are obtained and then model performance is evaluated using evaluation metrics namely, Root Mean Squared Error(RMSE), Mean Average Error(MAE), and R squared. The actual values against predicted values are plotted and visualized using a scatter plot to understand the model performances better.

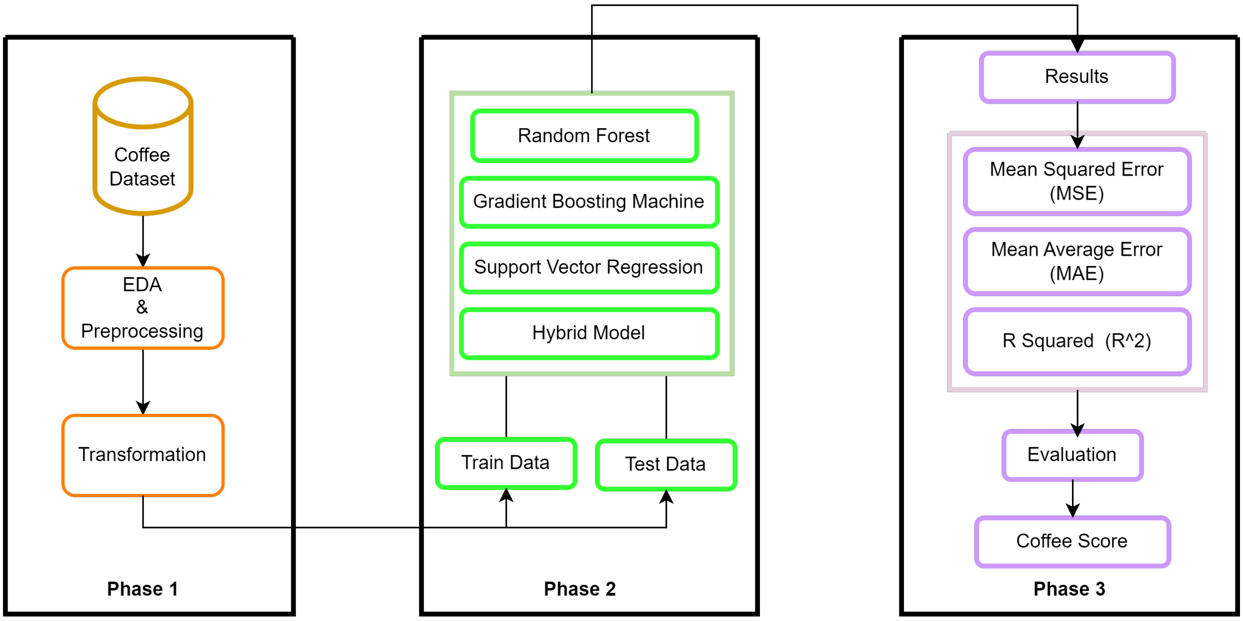


Figure 6: Methodology Flow.

5 Implementation

This chapter digs into implementing the machine learning models selected for the study and their performances. All the models are trained using the same set of training data and evaluated using the evaluation metrics mentioned below. Finally, the model that shows the best performance is selected and hyperparameter tuning is performed to obtain more accurate predictions then the best model is used to test and validate data using the test dataset.

5.1 Implementation of Machine Learning Algorithms

A machine learning model forms a mathematical methodology, or a set of algorithms constructed to recognize patterns and correlations within data, allowing the generation of predictions or estimations for new, hidden data. These models can be found in common applications such as image and audio recognition, Natural Language Processing (NLP), as well as predictive analytics.

For this study, three machine learning algorithms were selected and a hybrid model was built to find out which model performs well in predicting the quality of coffee. Those algorithms are Random Forest, Gradient Boost Machine(GBM), and Support Vector Regressor(SVR).

5.1.1 Random Forest

Random Forest is a combination learning method in machine learning that structures a variety of decision trees during the training process and outputs the mode of the classes (classification) or the mean prediction (regression) of the distinct trees (Schonlau & Zou 2020). It excels in accuracy and strength by mitigating overfitting and acquiring complicated relationships within the data through the sequence of various decision trees. Random Forest is broadly utilized for tasks such as classification and regression, offering a strong and flexible tool in predictive modeling.

Random Forest model is implemented to predict quality ('Total Cup Points') in the coffee dataset. It starts by selecting relevant attributes, containing sensory attributes and non-sensory attributes such as processing methods, Variety, and country of origin. The dataset is then separated into training and testing sets. The Random Forest Regressor is initiated with 100 trees and trained on the training dataset. Afterward, predictions are generated for both the training and test sets, and evaluation metrics namely Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared are computed.

5.1.2 Gradient Boost Machine(GBM)

Gradient Boosting Machine (GBM) is a common machine learning technique that develops a predictive model in the form of an ensemble of weak learners, usually decision trees, in a serial manner (Konstantinov & Utkin 2021). GBM minimalizes errors by serially fitting new models to the residual errors of the current ensemble, thus iteratively advancing the overall analytical performance. It is known for its extraordinary predictive accurateness, flexibility to diverse datasets, and ability to handle both regression and classification tasks efficiently. GBM has become a broadly employed method in various fields due to its strength and capability to capture complicated connections within the data.

Gradient Boosting Regressor model is initiated, with the creation of 100 decision trees within the ensemble, allowing for a possibly more sophisticated model, although at the cost of enhanced computational demands. The inclusion of a random state as 42 is vital for reproducibility, as it settles the random seed. This makes sure that, even though the inherent randomness in the model's training process, the results remain stable across sequences, raising reliability in the analysis.

5.1.3 Support Vector Regressor(SVR)

Support Vector Regression (SVR) is an algorithm used for regression tasks. Gained from the basis of Support Vector Machines (SVM), SVR targets to predict constant outcomes by discovering a hyperplane that best fits the data while minimalizing prediction errors (Awad et al. 2015). Unlike traditional regression techniques, SVR is efficient in high-dimensional spaces and is remarkably useful when dealing with non-linear relationships in the dataset. It depends on a set of support vectors, which are the data points that provide the most substantial to defining the hyperplane. SVR has proven to be a strong and flexible tool for regression applications, showing efficiency in various realms by acquiring complex relationships in the data.

The SVR model begins with a linear kernel, indicating a linear relationship between the attributes and the target variable. The model is trained on the training data, and the predictions are made both on the training as well as test sets. Evaluation metrics, containing Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared, are estimated to evaluate the model's performance on both sets.

5.1.4 Hybrid Model

Given that all three models provide unique properties and values, as well as pros and consequences of their own, a hybrid model was selected. To observe if it outdoes any or all the predictive models, this research planned to mix all of the models into a single one. Due to their extensive use and often greater effectiveness, hybrid machine learning models have observed a significant rise in use in the current (Bies et al. 2006).

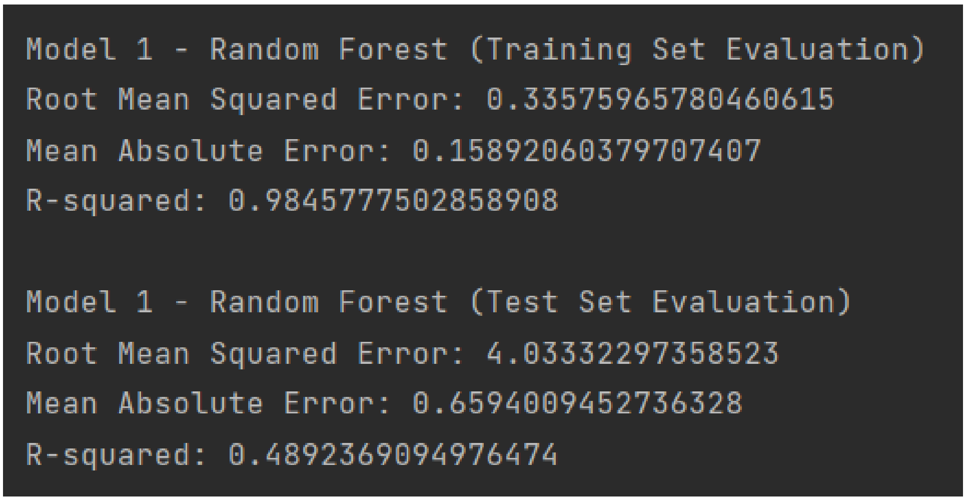
There are numerous ways to build a hybrid model and in this study, a hybrid model is built by linking predictions from three individual models: Random Forest, Gradient Boosting, and Support Vector Regressor (SVR). For both the training and test datasets, predictions from all models are averaged to produce hybrid predictions. This collective approach powers up the strengths of various models to potentially enhance overall predictive performance. RMSE, MAE, and R-squared, are measured to evaluate the hybrid model's effectiveness. This hybridization seeks to catch the complementary features of the data obtained by each model, potentially ensuing in a stronger and correct prediction of the 'Total Cup Points' in the coffee dataset

6 Evaluation

This chapter demonstrates and discusses the results gained from predictive models in predicting the quality of coffee. As described in the previous chapters gathered coffee data is preprocessed and split into two parts and put through machine learning models as training and testing datasets. The results are then evaluated using Root Mean Squared Error(RMSE), mean Average Error(MAE), and R Squared, and then the predicted values are plotted against actual values to illustrate the model prediction ability.

6.1 Experiment 1: Random Forest

According to the Figure 8 in the evaluation of the Random Forest model, the training set metrics disclose a creditable performance, with a low Root Mean Squared Error (RMSE) of 0.336 and a Mean Absolute Error (MAE) of 0.159, demonstrating accurate predictions with negligible deviation from actual values. The great R-squared (R^2) value of 0.985 implies that the model properly captures the essential patterns in the training data, describing 98.5% of the variance. Nevertheless, the transition to the test set introduces challenges, as exposed in a higher RMSE of 4.033, telling a larger deviation of roughly 4.033 points on average. The MAE on the test set is likewise higher at 0.659, suggestive of a somewhat increased prediction error. The R-squared value of 0.489 on the test set is remarkably lower, emphasizing a weakened ability of the model to generalize to new data, obtaining only 48.9% of the variance indicating potential over-fitting. Figure8 represents the scatter plot of the performance of the Random Forest model on training and testing data.



```
Model 1 - Random Forest (Training Set Evaluation)
Root Mean Squared Error: 0.33575965780460615
Mean Absolute Error: 0.15892060379707407
R-squared: 0.9845777502858908

Model 1 - Random Forest (Test Set Evaluation)
Root Mean Squared Error: 4.03332297358523
Mean Absolute Error: 0.6594009452736328
R-squared: 0.4892369094976474
```

Figure 7: Random Forest Evaluation

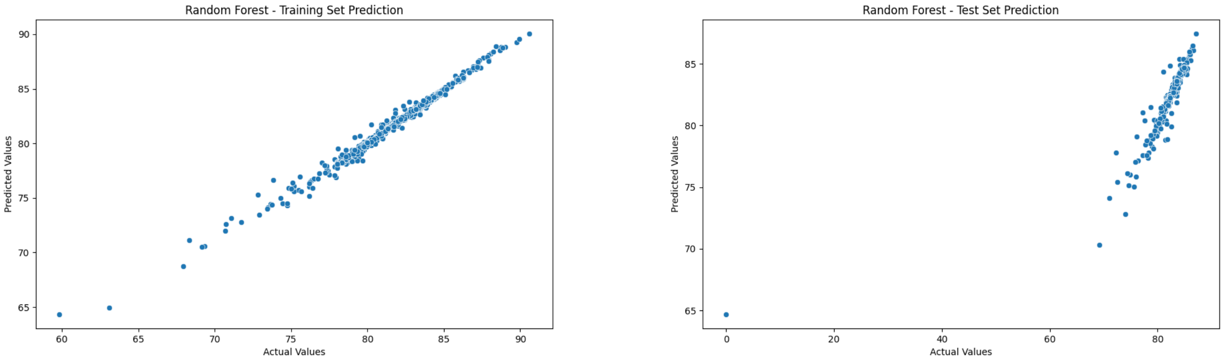
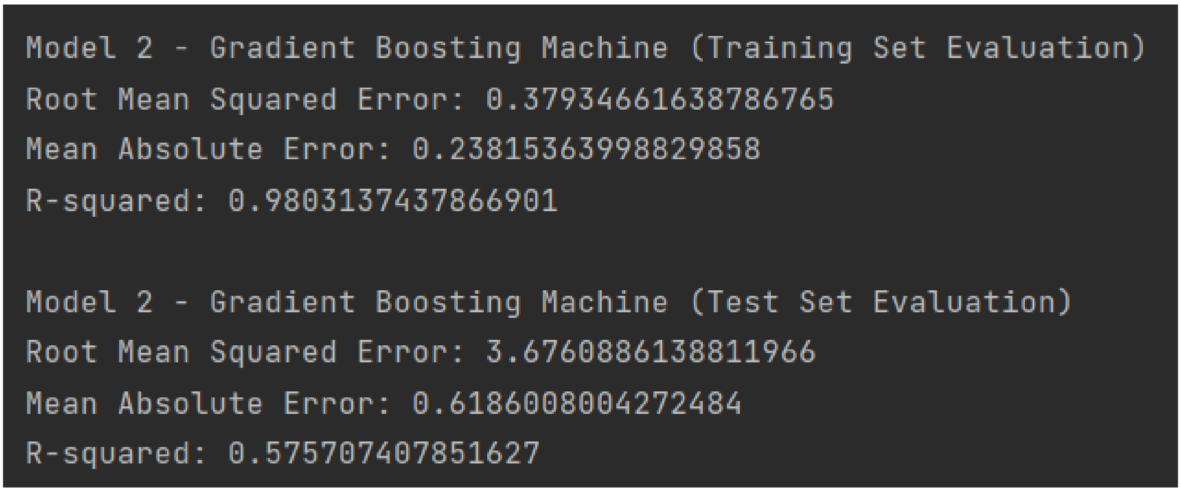


Figure 8: Random Forest Performance Plot

6.2 Experiment 2: Gradient Boosting Machine

As can be seen in Figure 9 In the evaluation of the Gradient Boosting Machine (GBM) model, the training set metrics show strong predictive performance. The low Root Mean Squared Error (RMSE) of 0.379 and Mean Absolute Error (MAE) of 0.238 denote accurate predictions with negligible differences from actual values. The high R-squared (R^2) value of 0.980 emphasizes the model's fitness to explain 98% of the variance in the training data, indicating a robust catch of basic patterns. Nevertheless, on the test set, while the GBM model retains a competitive performance, there is an evident rise in RMSE to 3.676 and MAE to 0.619. The R-squared value of 0.576 implies that the model describes 57.6% of the variance in the test data. Even though this signifies a small decrease from the training set performance, it still signifies a good capability to generalize. Overall, the stability in performance among training and testing sets implies good generalization, emphasizing the reliability of the GBM model. Figure10 represents the scatter plot of the performance of the GBM model on training and testing data.



```
Model 2 - Gradient Boosting Machine (Training Set Evaluation)
Root Mean Squared Error: 0.37934661638786765
Mean Absolute Error: 0.23815363998829858
R-squared: 0.9803137437866901

Model 2 - Gradient Boosting Machine (Test Set Evaluation)
Root Mean Squared Error: 3.6760886138811966
Mean Absolute Error: 0.6186008004272484
R-squared: 0.575707407851627
```

Figure 9: Gradient Boosting Machine Evaluation

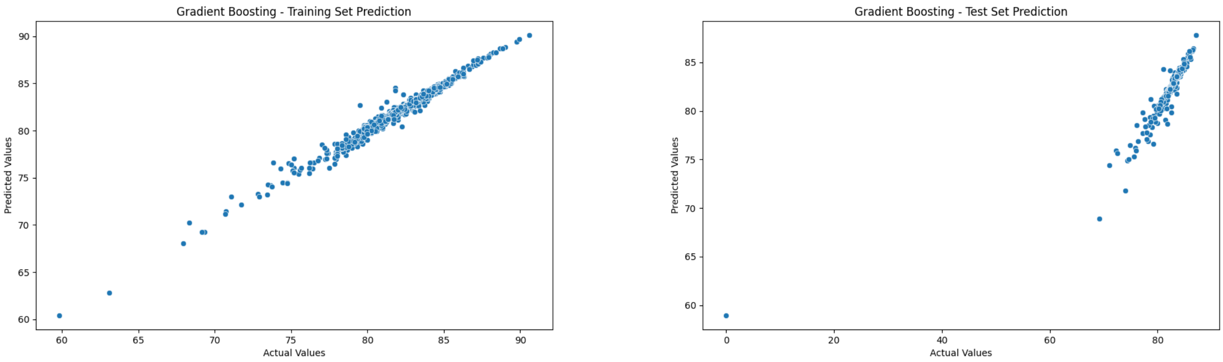
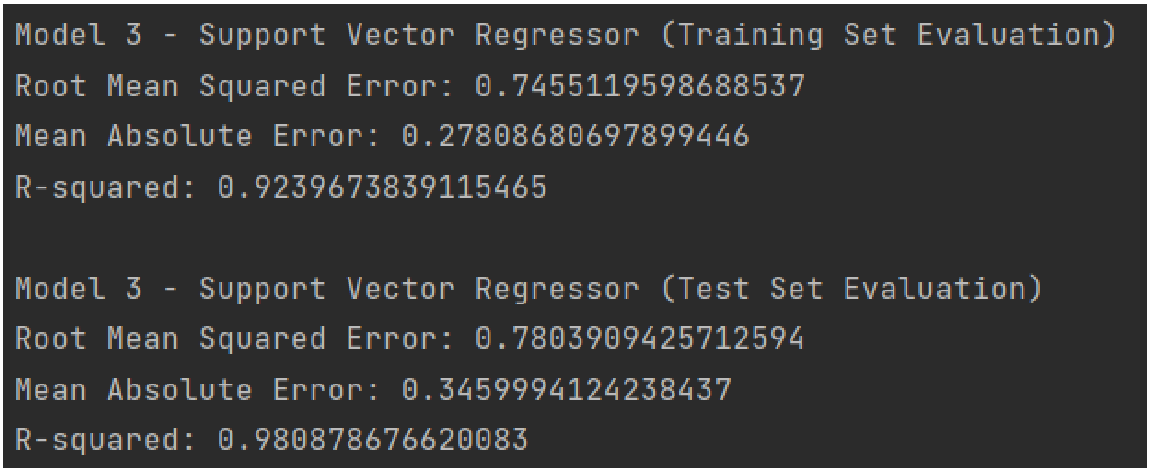


Figure 10: Gradient Boosting Machine Performance Plot

6.3 Experiment 3: Support Vector Regression

In the evaluation of the Support Vector Regressor (SVR) model, as shown in Figure 11, both the training and test set metrics imply strong predictive abilities. The low Root Mean Squared Error (RMSE) of 0.746 and 0.780 for the training and test sets, correspondingly, indicates that, on average, the model's predictions turn aside by roughly 0.746 and 0.780 points from the actual values. The Mean Absolute Error (MAE) values of 0.278 (training) and 0.346 (test) further verify the model's accuracy. The high R-squared (R^2) value of 0.924 on the training set and 0.981 on the test set signifies that the SVR model explains about 92.4% and 98.1% of the variance in the particular datasets. These results emphasize the SVR model's robustness in acquiring fundamental patterns in both the training and test data, indicating its efficiency in predicting coffee quality. The SVR model proves a strong capability for predicting coffee quality, with stability between training set correctness and generalization to new data. Figure12 represents the scatter plot of the performance of the SVR model on training and testing data.



```
Model 3 - Support Vector Regressor (Training Set Evaluation)
Root Mean Squared Error: 0.7455119598688537
Mean Absolute Error: 0.27808680697899446
R-squared: 0.9239673839115465

Model 3 - Support Vector Regressor (Test Set Evaluation)
Root Mean Squared Error: 0.7803909425712594
Mean Absolute Error: 0.3459994124238437
R-squared: 0.980878676620083
```

Figure 11: Support Vector Regression Evaluation

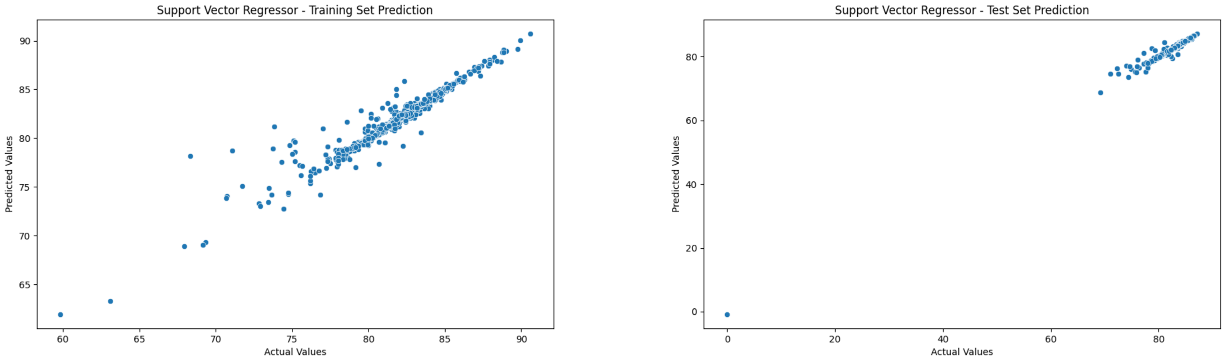
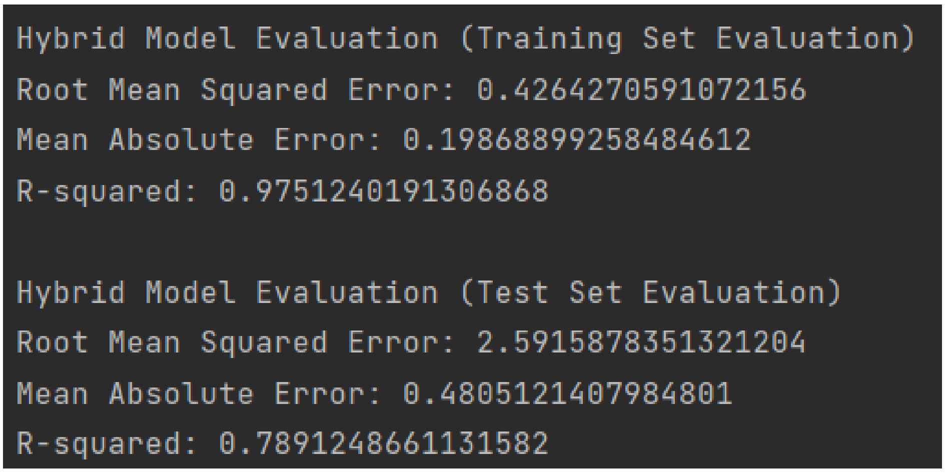


Figure 12: Support Vector Regression Performance Plot

6.4 Experiment 4: Hybrid Model

According to the Figure 13 Hybrid Model, both the training as well as test sets denote ensuring results in predicting coffee quality. On the training set, the Root Mean Squared Error (RMSE) of 0.426, Mean Absolute Error (MAE) of 0.199, and R-squared (R^2) of 0.975 indicate effective predictive performance. These values indicate that, on average, the model's predictions divert by approximately 0.426 points from the actual values, with an average absolute error of 0.199 points. The R-squared value of 0.975 implies that the model explains about 97.5% of the variance in the training data, showing its capability to capture primary patterns efficiently.

On the test set, the model determines good generalization with an RMSE of 2.592, MAE of 0.481, and R-squared of 0.789. While the test set metrics are somewhat higher than those on the training set, the RMSE of 2.592 implies that, on average, predictions divert by approximately 2.592 points from the real values. The R-squared value of 0.789 signifies that the model describes about 78.9% of the variance in the test data, highlighting its ability to present reliable predictions on new, unseen data. Overall, these results imply that the Hybrid Model, merging Random Forest, Gradient Boosting, and Support Vector Regressor models, produces a robust and precise prediction of coffee quality. Figure14 represents the scatter plot of the performance of the Hybrid model on training and testing data.



```
Hybrid Model Evaluation (Training Set Evaluation)
Root Mean Squared Error: 0.4264270591072156
Mean Absolute Error: 0.19868899258484612
R-squared: 0.9751240191306868

Hybrid Model Evaluation (Test Set Evaluation)
Root Mean Squared Error: 2.5915878351321204
Mean Absolute Error: 0.4805121407984801
R-squared: 0.7891248661131582
```

Figure 13: Hybrid Model Evaluation

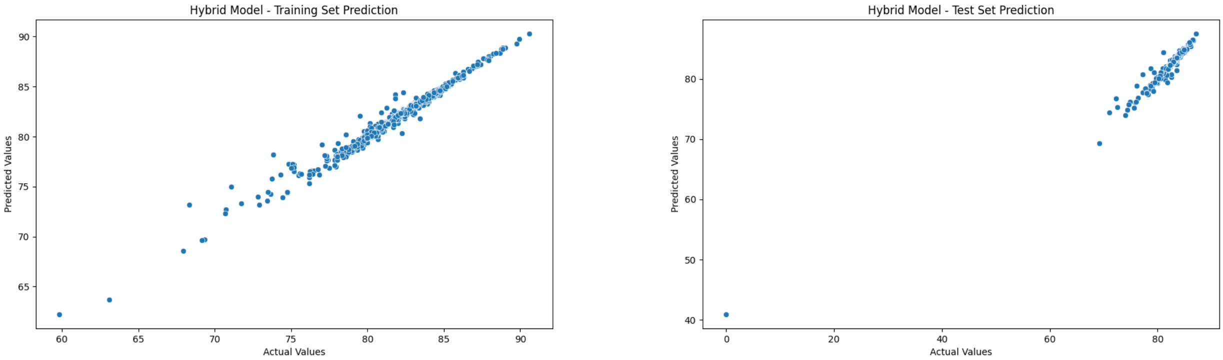


Figure 14: Hybrid Model Performance Plot

6.5 Discussion

6.5.1 Results Comparison

As shown in Table 1. a comparison of all four models, it is evident that Support Vector Regression outperforms Random Forest, GBM, and even Hybrid model by being consistent on both training and testing datasets. Obtaining 0.7455 and 0.7803 in RMSE for training and Testing data respectively shows that SVR succeeded in generalizing to the test or unseen data. Likewise, 0.2780 and training and testing data confirm its capability to predict coffee quality scores accurately. The hybrid model shows the second-best performance in predictions by acquiring 2.5915(RMSE), 0.4805(MAE), and 0.7891(R Squared) whereas GBM and Random Forest gained 3rd and 4th position in predicting coffee quality value accurately.

6.5.2 Results Discussion

The results acquired from the models provide valuable insights into the components affecting overall coffee quality and how advanced machine learning techniques can improve

Table 1: Evaluation of Machine Learning Models

Model Name	**RMSE**		**MAE**		**R Squared**	
	Training Set	**Testing Set**	**Training Set**	**Testing Set**	**Training Set**	**Testing Set**
Random Forest	0.3357	4.0333	0.1589	0.6594	0.9845	0.4892
GBM	0.3793	3.6760	0.2381	0.6186	0.9803	0.5757
SVR	0.7455	0.7803	0.2780	0.3459	0.9239	0.9808
Hybrid Model	0.4264	2.5915	0.1986	0.4805	0.9751	0.7891

accuracy in coffee quality assessment. This provides the solution to the research question mentioned in the Chapter 1 Introduction and achieves the primary objective of the research which is utilizing cutting-edge machine learning techniques to predict the coffee quality scores by using sensory as well as non-sensory attributes in the dataset. When exploring the literature review thoroughly, it can be found the lack of employing advanced machine learning techniques in assessing coffee quality. Addressing this gap in past studies three machine learning models and a hybrid model are employed to evaluate and obtain accurate coffee quality predictions.

The models employed, containing Random Forest, GBM, SVR, and the Hybrid Model, represent enhanced machine learning techniques. These models are capable of capturing complicated relationships between non-sensory features and coffee quality. The excellent execution of these models on the training set indicates their capability to learn from the data and make precise predictions. Furthermore, hyperparameter tuning, as performed with SVR's best parameters, shows the fine-tuning feasible with these techniques to achieve the finest performance.

The discussed models provide accurate predictions of coffee quality and the analysis of coffee quality factors sheds light on the interaction between sensory features as well as non-sensory features and coffee quality. The advanced models present the potential of machine learning in understanding and boosting complex processes like coffee production.

7 Conclusion and Future Work

In conclusion, this study investigated into the sophisticated domain of coffee quality assessment, utilizing advanced machine learning techniques to discover the factors affecting overall quality. The Random Forest, Gradient Boosting Machine, and Support Vector Regression models, along with a Hybrid Model, presented valuable insights into the sensory features and non-sensory features essential for predicting coffee quality. The study demonstrated the efficiency of data-driven approaches in improving accuracy and identifying complex associations within the coffee production process.

Looking ahead, future work could study more sophisticated hybrid approaches or neural network architectures for even more accurate predictions. Moreover, expanding the dataset to add in different coffee varieties and growing conditions could produce a broader understanding of quality factors. Exploring explainable Artificial Intelligence methodologies would further improve the interpretability of models, assisting in better communication of results to stakeholders. Additionally, integrating real-time data from IoT appliances in coffee farms and managing units could enable dynamical quality observation. This holistic methodology could contribute to the current evolution of the coffee industry, guaranteeing quality and sustainability at every stage of the supply chain.

Acknowledgement

My extreme appreciation to the National College of Ireland, and the MSc in Data Analytics department for making it possible for me to achieve this research project. More importantly, my mentor and supervisor Prof. Jorge Basilio, for his guidance to succeed in this research. His continued assistance allowed me to fulfill and deliver this thesis. I highly appreciate his time and his insights.

References

Aunsa-Ard, W. & Kerdcharoen, T. (2022), 'Electronic nose for analysis of coffee beans obtained from different altitudes and origin', *2022 14th International Conference on Knowledge and Smart Technology (KST)* .

Awad, M., Khanna, R., Awad, M. & Khanna, R. (2015), 'Support vector regression', *Efficient learning machines: Theories, concepts, and applications for engineers and system designers* pp. 67–80.

Behrens, J. T. (1997), 'Principles and procedures of exploratory data analysis.', *Psychological methods* **2**(2), 131.

Bies, R. R., Muldoon, M. F., Pollock, B. G., Manuck, S., Smith, G. & Sale, M. E. (2006), 'A genetic algorithm-based, hybrid machine learning approach to model selection', *Journal of pharmacokinetics and pharmacodynamics* **33**(2), 195.

Dong, G. & Liu, H. (2018), *Feature engineering for machine learning and data analytics*, CRC press.

Febriana, A., Muchtar, K., Dawood, R. & Lin, C.-Y. (2022), 'Usk-coffee dataset: A multi-class green arabica coffee bean dataset for deep learning', *2022 IEEE International Conference on Cybernetics and Computational Intelligence (CyberneticsCom)* .

Ferreira, T., Shuler, J., Guimarães, R. & Farah, A. (2019), 'Chapter 1. introduction to coffee plant and genetics', *Coffee* p. 1–25.

Garuma, H., Berecha, G. & Abedeta, C. (2014), 'Influence of coffee production systems on the occurrence of coffee beans abnormality: Implication on coffee quality', *Asian Journal of Plant Sciences* **14**(1), 40–44.

Huang, N.-F., Chou, D.-L. & Lee, C.-A. (2019), 'Real-time classification of green coffee beans by using a convolutional neural network', *2019 3rd International Conference on Imaging, Signal Processing and Communication (ICISPC)* .

Konstantinov, A. V. & Utkin, L. V. (2021), 'Interpretable machine learning with an ensemble of gradient boosting machines', *Knowledge-Based Systems* **222**, 106993.

Malau, S. (2018), 'Variability of organoleptic quality of arabica coffee', *Anadolu Journal of Agricultural Sciences* p. 241–245.

Marbán, Ó., Mariscal, G. & Segovia, J. (2009), A data mining & knowledge discovery process model, *in* 'Data mining and knowledge discovery in real life applications', IntechOpen.

Mariscal, G., Marban, O. & Fernandez, C. (2010), 'A survey of data mining and knowledge discovery process models and methodologies', *The Knowledge Engineering Review* **25**(2), 137–166.

Micaraseth, T., Pornpipatsakul, K., Chancharoen, R. & Phanomchoeng, G. (2022), 'Coffee bean inspection machine with deep learning classification', *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)* .

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F. & Rutledge, D. N. (2020), 'New data preprocessing trends based on ensemble of multiple preprocessing techniques', *TrAC Trends in Analytical Chemistry* **132**, 116045.

Schonlau, M. & Zou, R. Y. (2020), 'The random forest algorithm for statistical learning', *The Stata Journal* **20**(1), 3–29.

Toledo, P. R., Pezza, L., Pezza, H. R. & Toci, A. T. (2016), 'Relationship between the different aspects related to coffee quality and their volatile compounds', *Comprehensive Reviews in Food Science and Food Safety* **15**(4), 705–719.

Vijayan, K., Sebastian, L., J, V. & S, R. (2022), 'Predictive modelling for coffee production using r programming', *2022 3rd International Conference on Communication, Computing and Industry 4.0 (C2I4)* .