Apple Perfection: Assessing Apple Quality with Waterwheel Plant Algorithm for Feature Selection and Logistic Regression for Classification

Abdelhameed Ibrahim[1, *], Ehsan Khodadadi[2], Ehsaneh Khodadadi[2], P.K. Dutta[3], Nadjem Bailek[4,5], Abdelaziz A. Abdelhamid[6,7]

[1]Computer Engineering and Control Systems Department, Faculty of Engineering, Mansoura University, Mansoura 35516, Egypt.
[2]Department of Chemistry and Biochemistry, University of Arkansas, Fayetteville, AR 72701, USA.
[3]School of Engineering and Technology, Amity University Kolkata, India.
[4]Energies and Materials Research Laboratory, Faculty of Sciences and Technology, University of Tamanghasset, Tamanrasset, 10034, Algeria.
[5]Sustainable Development and Computer Science Laboratory, Faculty of Sciences and Technology, Ahmed Draia University of Adrar, Adrar, Algeria.
[6]Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, Cairo 11566, Egypt
[7]Department of Computer Science, College of Computing and Information Technology, Shaqra

* Corresponding author: afai79@mans.edu.eg

Emails: afai79@mans.edu.eg; Ehsank@uark.edu; ekhodada@uark.edu; pkdutta@kol.amity.edu; bailek.nadjem@univ-adrar.edu.dz; abdelaziz@su.edu.sa

Abstract

This study concentrates on the evaluation of apple quality, which is a vital part of the agricultural industry. The quality of apples is examined through several factors, such as the cultivation techniques, the harvesting methods, and the post-harvest procedures. The dataset, titled "Apple Perfection," contains important characteristics such as the size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and overall quality of the apple. To make the apple quality prediction more accurate, we used different feature selection algorithms, mainly the binary Waterwheel Plant Algorithm (bWWPA), which, in fact, had the lowest average error of 0.52153, and several of the types of classification models, especially Logistic Regression, which had the highest accuracy of 0.88625. The attribute selection process found the most important attributes, which cut down the dimensionality, and hence, the model performance became better. The results of the study show that the combination of bWWPA for feature selection and logistic regression for classification can predict apple quality with high accuracy. This way of dealing with the problem gives us information that is useful for the improvement of the cultivation techniques and the post-harvest handling to the extent that we will be able to have the best quality apples. The findings of this research have a great impact on the farming industry, meaning a strong way to evaluate the quality of apples.

Keywords: Apple Quality; Feature Selection; Waterwheel Plant Algorithm; Logistic Regression; Machine Learning; Agricultural Data Analysis.

MSC: 68N01; 68N30; 78M50

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

1. Introduction

Apple quality is not just a concern, but a crucial aspect in the agricultural industry. It significantly impacts marketing, customer satisfaction, and the economic value of a country. The quality of an apple is determined by a multitude of parameters, including size, weight, sweetness, crunchiness, juiciness, ripeness, acidity, and overall appearance. High-quality apples are not only a source of consumer satisfaction but also a key driver of profitability for stockholders. In a competitive market for fresh produce, consistently delivering superior apples can enhance a brand's reputation, foster long-term customer loyalty, and ensure compliance with market standard regulations [1-3].

Numerous factors influence apple quality, and cultivation practices are among them. The correct fertilization of apple trees ensures that they get the needed nutrients to grow healthily and produce the best fruits you can get. The nutrients have different influences on the factors of apple growth; for example, nitrogen helps vegetative growth, while potassium increases fruit quality and disease resistance. Irrigation is to be controlled really so that there will be enough water, but it will not be too much, which will cause the water to be logged, and thus, it will affect the root health and the fruit development. Effective irrigation systems, like drip irrigation, which ensure water is directly delivered to the roots, thus reducing waste and optimizing water usage, are the ones that are used more and more. Pest control is a key factor in protecting apples from pests and diseases that can affect their quality. Integrated pest management (IPM) is a major composition of the biological, cultural, and chemical methods of handling the pest sustainably. The choice of apple species is also essential since the different varieties have certain needs and characteristics that need to be taken into account in order to get the best quality. For instance, some cultivars may be more immune to certain pests and diseases, or they may have specific flavor profiles that are popular in some markets [4-5].

Moreover, the methods of gathering data are also the main factors that decide the apple's quality. The picking time is the key factor; the apples picked too early will not be as sweet as the ones harvested during the season, and the ones picked too late will be spoiled. The level of maturity of the fruit at the time of its harvest is a significant factor in determining its storage life and final eating quality. The way to deal with the handling procedures during harvesting must be gentle to avoid bruising and other physical damage. Objects such as hand-picking are usually the preferred ones for their accuracy and gentleness, but they are time-consuming and tiring. Mechanical harvesting is a more efficient way, but it has to be managed properly to avoid damage. At the same time, the workers in the harvesting of the crops can be trained in the proper harvesting methods, and the equipment can be given to them to help reduce post-harvest losses [6-8].

The post-harvest processes, such as sorting, grading, and storage methods, are also, as they are, as important as the approaches to apple quality during transit and distribution. Sorting and grading are the procedures that help in the grouping of apples of the same quality, which, in turn, assures the market standards and consumer expectations. Grading systems usually rate the size, color, and lack of defects of apples to sort them into different quality grades. The storage of apples under the right conditions, according to the temperature and humidity level, is necessary for the preservation of the freshness of the apples and the extension of their shelf life. Refrigeration of apples is a way to decrease the metabolic processes of the fruits and thus the level of spoilage and maintain the quality. Controlled atmosphere storage, which adjusts the level of oxygen and carbon dioxide, can prolong storage life. These processes guarantee that apples reach consumers in the most perfect condition, free from any damage and deterioration. Moreover, the packaging technology has progressed so that the apples will be safer to transport. The packaging materials that have cushioning and are able to keep the humidity level are the ones that can help preserve the quality of apples [9-11].

Since these factors are very complex and important, we can say that getting data-driven methods to measure and predict apple quality is a must. Usually, the conventional ways of quality evaluation are grounded on subjective judgments, which are not always reliable and are ineffective. Through the use

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

of the latest algorithms and machine learning models, we can get more accurate and unbiased evaluations of apple quality. Through this method, the reliability of quality evaluation is improved, and it also gives knowledge on the optimization of agriculture and post-harvest stages. Now, data-driven techniques can process big datasets to find patterns and correlations that are not seen through human inspection. To machining algorithms can discover the relationship between soil nutrients, weather conditions and apple quality. Thus, these algorithms will give us the formation that will be helpful in improving practices [12-14].

This study aims to examine and assess the efficiency of several feature selection algorithms and classification models in determining the quality of apples. The main part of our study is the bWPPA for feature selection and Logistic Regression for classification. The bWWPA algorithm is based on the water wheel concept and seeks to find out the features that are the most important for determining the quality of the apple. This algorithm copies the way water flows from the waterwheel to use a certain feature in the data set to find the most important features. Logistic Regression, a name that signifies its simple and powerful nature in binary classification tasks, is used for the creation of predictive models on the basis of the chosen features. Logistic Regression models the connection between the chosen features and the binary result (high vs. low-quality apples), which is the basis of the model interpretation and the prediction of the result. Therefore, the model is reliable, and the results are interpretable. Through the examination of these methods, we hope to distinguish the most significant features and thus get a high predictive accuracy, which, in the end, will contribute to the app development industry. The results of this investigation could be of great help to the apple producers in choosing the best methods for the cultivation, harvesting and post-harvest processes to produce consistently high-quality apples. Thus, the producers would be more competitive in the market. Moreover, the study's findings can be used for other crops. Thus, the data-driven quality assessment methods in agriculture can be made more widely known.

2. Literature Review

India is the second-largest fruit producer in the world after China. The main problem in India's fruit industry is the wastage of 30–35% of the harvested fruits, which is mainly caused by the lack of skilled labour. Besides, the subjective nature of human perception results in the imprecise identification, classification, and grading of fruits. Hence, the automation of the fruit industry is of vital importance. The combination of machine learning and the latest image processing concepts is a good way to design systems that are able to differentiate fruits by type, variety, maturity, and integrity. As [15] says, a thorough review of the research articles from 2010 to 2019 shows the different methods of identifying, classifying, and grading fruits. Besides, this paper also analyses the present achievements and limitations and suggests the directions for future research. The quality of fresh apples is a major issue for both consumers and manufacturers, and the classification based on the ripening stage is a major factor in the quality. [16] is trying to find a non-destructive way of classifying the ripening state of Fuji apples using the hyperspectral information in the visible and near-infrared (Vis/NIR) regions. The study examined spectra from 172 apple samples of four ripening stages, from 450 to 1000 nm. A convolutional neural network (CNN) model was used for classification, and its performance was compared with the artificial neural networks (ANN), support vector machines (SVM) and k-nearest neighbours (KNN). The CNN model obtained a correct classification rate (CCR) of 96.5%, outperforming ANN (89.5%), SVM (95.93%), and KNN (91.68%). These results show that there is a great scope for the development of a device for the fast and accurate quality estimation of apples.

Automatic pruning is a very important but, at the same time, a very labour-intensive and costly process in specialty crop production. In winter, professional pruners cut certain main branches of dormant trees according to preestablished rules. The main goal of automatic pruning applications is to cut down the need for manual labour and the associated costs by automating the pruning decisions. Thus, the integration of intelligent robotic pruners will be easier. In [17], a strong 3D reconstruction scheme based on colour information and time-of-flight depth data from the Kinect2 sensor was designed for the accurate modelling of the trunk and primary branches of dormant apple trees. The addition of colour information was intended for accurate 3D reconstruction, even with poor-quality depth data. The suggested method proved a performance accuracy of 93.94% for branch identification and 71% for branch identification. 13% and 89.26% for the diameter estimation of primary branches within the error margins of 3 mm and 5 mm, respectively. Besides, the designed algorithm demonstrated a considerable reduction in time complexity when compared to the base approach. The digitalization of data has resulted in a data explosion in the data-driven industries. Man-to-machine (M2M) digital data handling has greatly increased this information wave; thus, digital agriculture management applications have significantly improved. These advancements have helped farmers and consumers by introducing technological solutions into rural areas. [18] deals with the possible use of ICT technologies in traditional agriculture and the difficulties in their implementation. The research explains the functions of robotics, IoT devices, machine learning, artificial intelligence, and sensors in agriculture. Besides, drones are being used for crop observation and yield optimization management. The review contains a study of the world-wide and the latest IoT-based farming systems and platforms. The detailed evaluation ends with a conversation about the current and future AI trends in agriculture and the highlighted research challenges.

Automated machine learning (AutoML) is a great progress in artificial intelligence, which is designed to provide high-performance end-to-end machine learning pipelines with the least amount of user input. Although AutoML has been successfully used in computer vision tasks, no studies have applied it to hyperspectral imaging for the classification of fertilization levels. For this purpose, [19] studies the implementation of AutoML for the classification of fertilization levels using hyperspectral and CIELAB colour space datasets. A comparative analysis between PyCaret, an open-source AutoML framework, and traditional machine learning using the PLS-DA algorithm showed that PyCaret had the highest accuracy (1.00) with the hyperspectral dataset, while PLS-DA had 0.91. The CIELAB dataset was not as good, with accuracies of 0.72. These findings show that AutoML has great potential to improve hyperspectral imaging applications in agriculture, especially for fertilization tasks. The spectrochemical estimation of pH and titratable acidity (TA) in Fuji apples at different ripening stages is the subject of [20]. A novel near-infrared (NIR) spectral analysis method that uses hybrid machine learning methods combining ANN and metaheuristic algorithms is suggested. Spectral data from 120 samples of three ripening stages were used to predict the acidity properties. Besides, the four most effective wavelengths were chosen by a combination of ANN and cultural algorithms. The models showed a correlation coefficient, R, of 0.926 for pH and 0.925 for TA using spectral bands and 0.924 and 0.920, respectively, for the second approach. Even though the models had a hard time with extreme pH and TA values because of clusters after regression, both the methods reached high classification accuracy (100% for pH and 99.2% for TA) for low/high acidity levels.

[21] studies the factors that affect the users' intention to adopt wearable payment systems, which are the perceived aesthetics, technology readiness, mobile usefulness, and ease of use. The study employs a dual-stage analysis and deep learning on 307 responses. The results show that all relationships were confirmed except the connection between the ease of use of mobile and behavioural intention. These findings give payment companies and smart wearable device manufacturers the opportunity to devise effective marketing strategies. The combined theoretical model of the Mobile Technology Acceptance Model, Fashion Theory, and Technology Readiness Theory gives a better insight into wearable payment acceptance among consumers. Through the application of computer vision in the fruit processing industry, tasks like quality classification and gradation are automated. [22] concentrates on the identification of rotten or fresh apples by detecting the peel defects via deep learning-based semantic segmentation. The research uses UNet and its improved version, En-UNet, and the latter attain the training and validation accuracies of 97.46% and 97.54%, respectively, which is a lot less than 95%.36% for UNet. The most accurate mean IoU score under a threshold of the best 0.95 for En-UNet was 0.866, compared to 0.66 for UNet. These findings prove that En-UNet is more suitable for real-time segmentation, detection, and categorization of apples.

In [23], multivariate techniques like machine learning are applied to estimate the biochemical traits from the spectral and colour characteristics of foodstuffs and agricultural commodities. The research analysed dried apples of various cultivars through different drying methods and evaluated their spectral, chromatic, and biochemical characteristics using five machine-learning algorithms. The highest success estimation was for the total phenolic content ($R \geq 0.\ 85$). Multilayer Perceptron, Support Vector Regression, and Gaussian Processes were the best algorithms for estimating the biochemical compositions of dried apples. Moreover, [24] investigates the politics of classification within machine learning systems, stressing that automated image interpretation is a social and political process. The study investigates the role of images in computer vision systems, the process of introducing images into these systems, and the building of taxonomies that determine the interpretations of the system. Besides it also talks about the consequences of labelling, for instance, the way AI systems categorize humans by race, gender, emotions, ability, sexuality, and personality. The study demands data archaeology to critically analyse the politics and values that are included in AI systems and their social consequences.

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

3. Proposed Methodology

3.1 Dataset

The "Apple Perfection" dataset, which is also called "Your Source for Everything Apple Quality," is a comprehensive and carefully curated collection of data points that describe the wide variety of attributes that are necessary for evaluating Apple quality [25]. The dataset constitutes the whole picture of the factors that determine apple quality. Hence, it is a treasure for detailed analysis and predictive modeling. Thus, the dataset that contains both physical and sensory characteristics of apples is able to provide a strong basis for the examination of how the different attributes contribute to the overall quality of the apples.

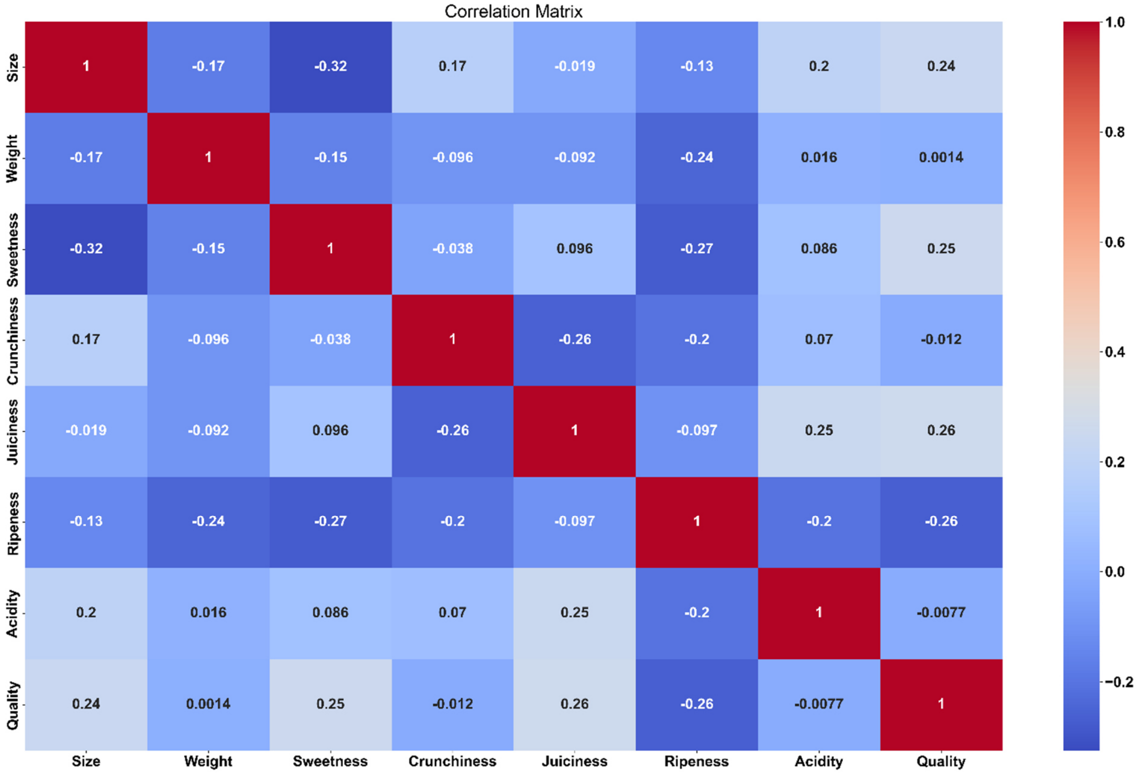
Explanation of Key Features:

- A_id (Apple ID): Each apple in the dataset carries its unique tag number. This feature is the main feature of the system for tracking and differentiating apples in the dataset, thus ensuring that every data point in the system can be analyzed correctly and accurately. With our Apple ID, you will be able to simplify data management and also research longitudinal studies and comparisons in detail.
- Size: The measurements of an apple are generally the diameter and the length. Size is a significant indicator that always coincides with a customer's tasting or market rules. Bigger apples are evident in some markets for the visual effect they have and the value they show, but the opposite is true for the smaller apples, which are best for snacks or children's lunches. The appropriate measurement of the size is done using calibrated tools or automated image analysis systems.
- Weight: The weight of the apple is usually measured in grams. The weight of the apple is the key factor in its density and toughness, which influences the way it is handled and transported. Heavier apples are usually considered juicier and more substantial, and this can influence customers' choices. Weight is the measurement of mass, which is done on precision scales so that the result can be achieved accurately. Besides, it also enlightens about the apple's water content and nutrient composition.
- Sweetness: The concept of sugar content in the apple is measured by the percentage of sugar in the apple's juice, which is indicated by the Brix units. Sweetness is the main cause of the consumer's happiness, and the sweetness of different apple cultivars is very different from that of other apples. Usually, high sweet levels are the ones liked by consumers, especially when it comes to fresh eating. Sweetness is measured by using refractometers or chemical analysis of sugar content.
- Crunchiness: The evaluation of the apple's texture, namely, how crispy and firm it is when you bite into it. A nice, crunchy texture is a plus that makes the eating better and tells that the product is fresh. This attribute can be assessed either by way of mechanical testing, which judges the amount of pressure to be used to bite through the apple, or the opinion of the trained experts. Crunchiness is a characteristic of the apple's flesh that is related to the structural integrity of the cell walls, and it is affected by the composition of the cell wall and the moisture content.
- Juiciness: This function finds the moisture level of the apple, which is a good indicator of how much juice the apple has. Juiciness is one of the essential attributes that affect tastiness and is the most important criterion for many buyers. The acme of juiciness in a fruit indicates its freshness and succulence. Juiciness is calculated by taking out and counting the juice content of the apple or by means of sensory evaluation. This character is shaped by the apple's water content, cellular structure, and the harmony of sugars and acids.
- Ripeness: A sign of the apple's maturity at the moment of the examination. The suitability of the apple plays a role in both the flavor profile and the shelf life of the fruit. The right ripeness of the apples is a must for the perfect taste and texture. Ripeness can be measured by visual inspection, firmness testing, and expressing the starch content. The iodine-starch test or chlorophyll fluorescence are techniques that are employed to find out the condition of ripeness. Ripeness is the most essential factor for picking the apples at the perfect time for flavor and storage capacity.
- Acidity: The tartness or sourness of an apple can be determined by its pH level or specific acid content. The acidity grips the sweetness and contributes to the apple's flavor complexity. Varied degrees of acidity are aimed at different consumer tastes and the purpose of use (e.g., a lemonade could be made to be either very sour or sweeter). fresh consumption vs. cooking). Titration methods and pH meters are the techniques that are used to measure acidity. This is a significant factor that can influence the taste of the apple and, therefore, dictate its use in various cooking techniques.
- Quality: An overall grade that is given to the apple and also shows the total evaluation of all the different features of the apple. This characteristic is generally developed by a combination of physical inspection and consumer feedback, and it is the variable of the target in the predictive modeling of this study. The quality grade includes all the individual characteristics, thus creating a single judgment that is the total

evaluation of an apple's worthiness. Quality evaluations are carried out with the help of grading norms, which take into account features like appearance, texture, taste, and defects.

The "Apple Perfection" dataset is created to show the whole picture of apple quality. Thus, it allows for the detailed analysis of how the various attributes interact and, therefore, contribute to the general view. This dataset gives a strong base for the application of data-driven techniques to predict and improve apple quality. Through the use of this dataset, we will be able to find the major patterns and relationships that can inform the development of better cultivation, harvesting and post-harvest practices; thus, the information will be both comprehensive and applicable to the agricultural industry. This way, it not only helps improve Apple's production processes but also fulfills consumers' expectations and market standards. Thus, the profitability and the sustainability of the apple producers are in the end.

By both helping the students in their research and serving as a tool for the professionals in the industry, the "Apple Perfection" data set is an invaluable resource. The data obtained through the sources can be used by the growers, agronomists, and quality control specialists to ascertain the right choices regarding cultivar selection, orchard management, and post-harvest handling. The data-centric strategy, on the other hand, provides the opportunity for continuous improvement and flexibility in adapting to the unpredictability of markets. As a result, Apple producers are competitive and able to respond to changes in consumer tastes. Using the most modern algorithms and machine learning models, this research aims to raise the quality assurance and improvement bar for apples, which will be the future producers and consumers.



	Size	Weight	Sweetness	Crunchiness	Juiciness	Ripeness	Acidity	Quality
Size	1	-0.17	-0.32	0.17	-0.019	-0.13	0.2	0.24
Weight	-0.17	1	-0.15	-0.096	-0.092	-0.24	0.016	0.0014
Sweetness	-0.32	-0.15	1	-0.038	0.096	-0.27	0.086	0.25
Crunchiness	0.17	-0.096	-0.038	1	-0.26	-0.2	0.07	-0.012
Juiciness	-0.019	-0.092	0.096	-0.26	1	-0.097	0.25	0.26
Ripeness	-0.13	-0.24	-0.27	-0.2	-0.097	1	-0.2	-0.26
Acidity	0.2	0.016	0.086	0.07	0.25	-0.2	1	-0.0077
Quality	0.24	0.0014	0.25	-0.012	0.26	-0.26	-0.0077	1

Figure 1: Correlation Matrix for the Dataset

Figure 1 presents the correlation matrix for the 'Apple Perfection' dataset. This matrix reveals the interconnections between the features. It contains coefficients that signify the strength and direction of the linear relationships between each feature pair. Correlations nearing 1 or -1 indicate robust linear relationships, while those close to 0 suggest weak or no linear relationships. This visual representation allows us to comprehend the interplay of features and identify multicollinearity, which could impact the accuracy of predictive models.

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

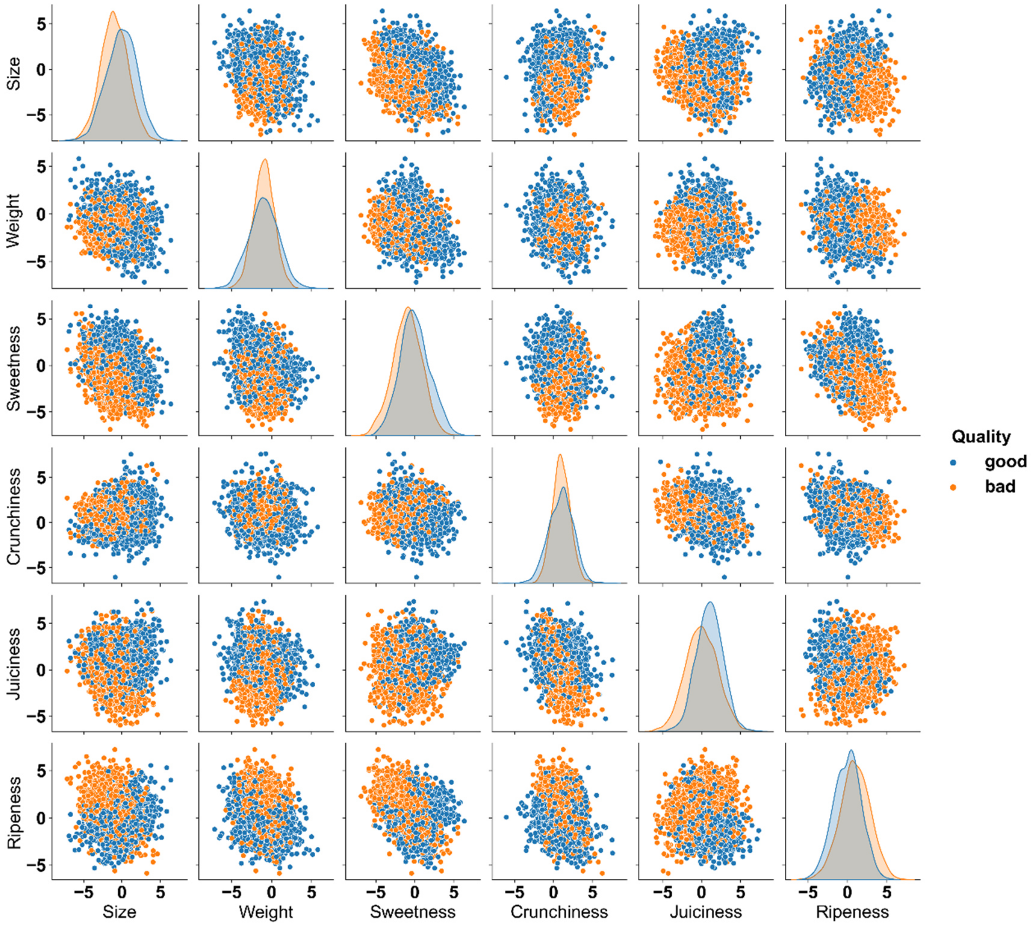


Figure 2: Pair Plot for the Dataset

Figure 2 encompasses the pair plot for the 'Apple Perfection' dataset, which is a powerful visualization that demonstrates correlations between any two variables. Every pair plot, a scatter plot of two features, manifests patterns, trends, and possible outliers graphically. The diagonal scatter plots show the distribution of individual features. This is a crucial visualization tool in structuring the data, with features selected and modeled using feature correlation detection and distribution analysis.

The preprocessing step is an essential stage for good quality and integrity of the dataset that will be used for analysis or modeling purposes. This phase includes several key steps:

1. Handling Missing Values:

The fact that data is usually missing in real-world datasets and hence it should be taken care of, there are biases in analysis. Through our study, the missing values in the Sleep Health and Lifestyle Dataset are found, and the impute methods are used to replace them efficiently. These techniques are the mean or median imputation, where missing values are substituted with the mean or median of the corresponding feature and also the other techniques like the k-nearest neighbors' imputation, which is based on the generalization of the data points to secure the missing values correctly.

2. Encoding Categorical Variables:

The categorical variables, such as gender and occupation, are non-numeric and cannot be used in computational analysis. They should be transformed suitably. These techniques are known as encoding methods, and we employ them to encode categories such as one-hot encoding or label encoding. One-hot encoding transforms the categorical variables into binary columns, with each category indicated by a binary indicator, and label encoding gives a numerical label to each category. Through the encoding of categorical variables, we are able to make sure that the algorithms will be able to properly understand and use these variables when they are in the modeling stage.

3. Feature Scaling:

The elements of the dataset usually have different scales, which can, in turn, influence the durability of the machine learning algorithms. We solve this problem by using feature scaling techniques; feature scaling techniques are the processes of standardizing or normalizing the features into a comparable range. Standardization recalibrates features to have a mean of zero and a standard deviation of one, while normalization scales them to a predefined range between zero and one. This, hence, amplifies the features and the convergence properties of the algorithms, thereby making the model training more efficient.

Through the careful cleaning and preprocessing of the data, we pave the way for the Sleep Health and Lifestyle Dataset to be ready for further analysis and modeling. Through the treatment of missing values, application of the encoding of the categorical variables, and scaling of features, we facilitate machine learning algorithms to obtain insight from the data and make an accurate prediction of sleep disorders based on the comprehensive sleep and lifestyle metrics included in the dataset.

3.2 Feature Selection

Feature selection is a significant phase in the data preprocessing process, which consists of finding and picking the most important features from a dataset to use as input for the model training. The main aim of feature selection is to improve the machine learning models by getting rid of the superfluous or irrelevant features that are not going to be useful, thus lowering the dimensionality of the data [26-28]. This process is essential for several reasons:

1. Improved Model Performance: Through the concentration on the vital features, the models can gain the maximum accuracy and, thus, the highest predictive performance. Superfluous or noisy features can hinder the model's performance and thus result in the wrong predictions. Feature selection makes the input data be taken in a streamlined way; for this, the most informative variables will be used for training to make the model to be able to generalize well on new, unseen data.

2. Reduced Overfitting: The models with a lot of features have the tendency to overfit; they work well on the training data, but they do not perform well on the unseen test data. The model is learning not only the underlying patterns in the training data but also the noise and outliers, thus, the overfitting. Feature selection is the process of removing the redundant features that do not have any impact on the model's predictive power, which, as a result, will not be overfitted and will be able to focus on the real underlying patterns.

3. Enhanced Interpretability: Models with fewer features are simple and, thus, easy to understand and interpret. This is especially significant in healthcare and finance, where the factors that are responsible for the predictions are taken into consideration to be understood. The model becomes easier to understand, and the decision-making process becomes more transparent with fewer features, which is why the stakeholders will have more trust in it.
4. Efficient Computation: The cut of features decreases the computational cost of the training of models. This is particularly advantageous when you are faced with large datasets and complex models because it allows you to train the model faster and with fewer resources. Computation is a fundamental factor in the practical fields where time and hardware resources are scarce.

In this research, we applied different binary optimization algorithms for feature selection to find the most important features that affect apple quality. These algorithms are built to operate in binary areas, which makes them good for feature selection problems. Through the use of such two-step programs, we intend to improve the quality and performance of our predictive models.

To choose the most important features from the "Apple Perfection" dataset, we applied several binary optimization algorithms. These algorithms are based on natural and physical processes and have been modified to work in binary search spaces for better feature selection.

- bWWPA: Binary Waterwheel Plant Algorithm: This algorithm is based on the waterwheel plant's water flow mechanism, which was used here to optimize the feature selection. The efficiency of the

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

waterwheel plant in water flow utilization is used to select the most important features from the dataset iteratively. Therefore, this approach maintains exploration and exploitation processes to find the best subset of the features.

- bMVO: Binary Multiverse Optimization: This algorithm, akin to the multiverse theory in cosmology, embarks on a journey through different 'universes' or solution spaces in search of the best features. It creates a parallel existence of multiple solutions, refining the task of finding the best feature subset. The process of identifying simultaneous available solutions is a powerful tool that significantly enhances the probability of discovering the most valuable features.

- bGWO: The Grey Wolf Optimizer: This optimization algorithm is rooted in the hunting behavior and social hierarchy of grey wolves. It dynamically selects features, mirroring the hierarchical structure and collaborative hunting tactics of these animals. The algorithm ensures a uniform search through its alpha, beta, and delta nodes, providing a fascinating glimpse into the world of grey wolves.

- bWOA: the Binary Whale Optimization Algorithm: Based on the bubble-net feeding behavior of humpback whales, the algorithm exploits the same strategy in the search for the best features in binary spaces. The synchronous and spiral movements of whales are mastered to increase diversity. This way of surrounding the prey is analogous to the precise run of the algorithm that comes down to the subset selection.

- bPSO: the binary Particle Swarm Optimization through social bird flocking and fish schooling study uses a population of particles that transform their positions according to their personal and collective experiences. The collaborative intelligence and communication within particles enhance the feature selection process. This algorithm maintains the balance of exploration and exploitation by altering the particle speeds and locations.

- bBA: binary Bat Algorithm: This algorithm, inspired by bats' echolocation behavior, uses bats' ability to detect objects and optimize feature selection. The algorithm's adaptive and dynamic features allow more relevant features to be found. The feature set is based on the assumption of echolocation and concentrates on the most promising areas.

- bGA: Binary Genetic Algorithm: Based on the foundations of natural selection and genetics, this algorithm involves crossover, mutation, and selection to find the best features. The algorithm's ability to grow and change over generations increases its efficiency in feature selection. GA's genetic algorithm ensures that in each iteration, the most efficient genes are preserved.

These algorithms were used to detect the most significant features that determine the quality of apples. Through the selection of the most significant features, we make sure that the subsequent modeling stage is based on the most informative data. Thus, the accuracy and reliability of the predictive models are improved. This thorough feature selection process is the key to the creation of reliable models that can predict apple quality based on the selected attributes. The application of these cutting-edge optimization algorithms allows us to attain a high degree of accuracy and efficiency in feature selection, thus leading to the achievement of our predictive modeling goals.

3.3 Machine Learning

This study made use of different classification models to classify the apple quality according to the selected features. Every model has its own strengths and features. Thus, it is applicable to various types of data and problem situations [29-31]. Here is an explanation of the classification models used:

- Logistic Regression: Logistic Regression is a linear model that is used for binary classification tasks. It illustrates the probability of a certain input being a particular class by using a logistic function to transform the linear output into a probability. This probability aids in the making of decisions about class membership. Logistic Regression is famous for its simplicity and interpretability. Thus, it is the most common choice for binary classification problems. It presumes a linear connection between the independent variables and the log odds of the dependent variable.

- Decision Tree: A Decision Tree is a non-linear model that divides the data into subsets according to the value of the input features. Every node in the tree is a decision point, and the branches are the possible outcomes. The leaves of the tree are the last classification results. Decision Trees are simple to understand

and can represent complicated decision boundaries. They are very intuitive and can deal with both numerical and categorical data. Nevertheless, Decision Trees can be vulnerable to overfitting if not well pruned. Thus, they might do well on training data but badly on unseen data.

- Random Forest: Random Forest is a technique in which several decision trees are created during training time, and the output is the mode of the classes that are used for classification tasks. It is an ensemble of the predictions of many trees which are probabilistically consistent. By averaging the results, Random Forest lowers the possibility of overfitting the model and, thus, increases its generalization ability. This method deploys an ensemble of varied models to acquire a more accurate and consistent forecast. Random Forest takes into account a high number of input features and is relatively immune to outliers compared to a single decision tree.

- Support Vector Machine (SVM): SVM is a powerful classification method that identifies the hyperplane, which splits the data into different classes absolutely. It is based on the principle of optimization of the margin between the classes, which makes it energetic for high-dimensional spaces. SVMs have exclusive applications in classification problems where the number of dimensions is more than the number of observations. They can contend with the non-linear decision boundaries through the use of kernel functions such as the polynomial or the Radial Basis Function (RBF) kernels.

- K-Nearest Neighbors (KNN): KNN is a simple and popular instance-based learning algorithm that classifies a point based on the majority class among its K closest neighbors in the feature space. It is a non-parametric method, which means it does not assume anything about the data distribution. KNN is easy to apply and understand, but it can be computationally intensive for large datasets since it requires computing the distance between the query point and all the points in the dataset. The decision of K number and distance metric (d). g. (Euclidean, Manhattan) greatly influences the performance of the model.

- Naive Bayes: Naive Bayes is a probabilistic classifier that is based on the Bayes rule, and it is assumed that the features are independent. Surprisingly, the model performs better than you expect, especially for text classification tasks and other ones with categorical input features. It is effective and suitable for big data. Naive Bayes models are easy to learn and can handle missing data and features that are irrelevant.

- Neural Network (MLP): A Neural Network, particularly a Multilayer Perceptron (MLP), is an example of an architecture that mimics the brain's neural networks. It has an input layer, one or more hidden layers, and an output layer. Each layer consists of neurons that deal with the input data by using weights, biases, and activation functions. MLPs can establish complex relationships and patterns in the data, that is why they are used for a lot of classification tasks. They can model non-linear decision boundaries better and typically perform well on large and complex datasets.

- Gradient Boosting: Gradient Boosting is a sequential method which is composed of one model that amends the mistake of the preceding model. It integrates the forecasts of several weak learners (mostly decision trees) to get a strong learner. Gradient Boosting is well-known for a high accuracy and the ability to handle various data kinds. However, it is prone to be overfitting in which case it should be adjusted. Methods like regularization or early stopping are commonly used to reduce the problem of overfitting. The Gradient Boosting models have both regression and classification capabilities and they are also good with imbalanced datasets.

- SVM (RBF Kernel): This type of SVM utilizes an RBF kernel to transform the input features into a higher-dimensional space and then a linear separator can be identified. The RBF kernel is most effective in non-linear classification problems; therefore, SVM creates a complex decision boundary. It enhances model's capability to deal with non-linear separable data. The RBF kernel works out the similarity between the data points. So, SVM can look for the optimal hyperplane in the transformed feature space.

- AdaBoost: AdaBoost (Adaptive Boosting) is an ensemble method that merges multiple weak classifiers to create a strong classifier. It works by giving more weight to misclassified instances in each iteration, thus making the model concentrate on the difficult cases. AdaBoost boosts the model's performance in each iteration, correcting the errors. Thus, it becomes robust enough to overfit and capable of getting high accuracy.

To evaluate the performance of the classification models, we used several key metrics:

- Accuracy: Accuracy is the ratio of the correctly classified instances to the total instances. It is a common parameter but can be deceiving for imbalanced datasets, where the number of instances in different classes is not the same. Accuracy is the ratio of the true positive and true negative predictions to the total number of predictions. Accuracy gives a quick overview of the model performance, but it does not take into account the distribution of classes.

- Sensitivity (True Positive Rate, TPR): Specificity, which is the same as recall, is the proportion of real positives that the model has correctly identified. It is the most important factor when we consider cases where the lack of positive cases has the most serious consequences. Sensitivity is the number of true positives divided by the sum of true positives and false negatives. High sensitivity means that the model can detect positive cases, which is very important for applications like disease detection.

- Specificity (True Negative Rate, TNR): The specificity is the percentage of the negatives that the model correctly identifies. It plays a crucial role in reducing false positives. The specificity expresses the number of true negatives divided by the sum of true negatives and false positives. High sensitivity shows that the model effectively identifies the negative case. On this note, the chances of false alarms are lessened.

- P-Value PPV (Positive Predictive Value): PPV, which is also called precision, is the ratio of correct positive predictions to the total number of positive predictions. It proves useful when the cost of false positives is high. PPV is the proportion of true positives in the sum of true positives and false positives. A high PPV indicates that if the model predicts a positive case, then it is likely to be true; this is very important in tasks like spam detection.

- P-value NPV (Negative Predictive Value): NPV is the percentage of negative predictions that turn out to be true. It assumes a high importance in scenarios where the cost of false negatives is high. NPV is the ratio of true negatives to the sum of true negatives and false negatives. High NPV indicates that the prediction model shows negative cases most of the time, and this is very important in screening tests where the negative cases should not be missed.

- F-Score: The F-Score is the arithmetic mean of precision and recall. It provides a particular way of determining the accuracy and completeness of a system. Subsequently, it is prudent for model evaluation if both false positives and false negatives have a significant impact. The F-Score is a great tool for imbalanced datasets because it provides a more balanced assessment of model performance.

- These evaluation measures give us an overall image of the model's performance to check its efficacy in predicting apple quality. Through the examination of these metrics, we will find out the strong and weak points of each model and choose the most effective ones for our predictive tasks in this way. Through this comprehensive study, the models that are selected are not just accurate but also consistent and stable; they can make exact predictions in real-life situations.

4 Results

In this part, here we illustrate our research findings, which are mainly derived from feature selection and their performances, which were based on different classification models. We provide a comprehensive consideration of the efficiency of the binary optimization algorithm in selecting important features and a detailed examination of the models using several performance indicators. The purpose of the feature selection process was to find the most critical features from the "Apple Perfection" dataset so that the model would use only the most salient variables for the training. To achieve this objective, we used various binary optimization algorithms. Table 1 lists the performance of each algorithm as average error, average select size, average fitness, best fitness, worst fitness, and standard deviation fitness.

Table 1: Feature Selection Results

	bWWPA	bGWO	bPSO	bBA	bWAO	bMVO	bGA
Average error	0.52153	0.53873	0.57253	0.58213	0.57233	0.54923	0.55233
Average Select size	0.47433	0.67433	0.67433	0.81373	0.83773	0.77083	0.61673
Average Fitness	0.58473	0.60093	0.59933	0.62223	0.60713	0.62903	0.61233
Best Fitness	0.48653	0.52123	0.57963	0.51193	0.57123	0.55423	0.51563
Worst Fitness	0.58503	0.58813	0.64733	0.61353	0.64733	0.67223	0.63073

Standard deviation Fitness	0.40703	0.41173	0.41113	0.42103	0.41333	0.46183	0.41333

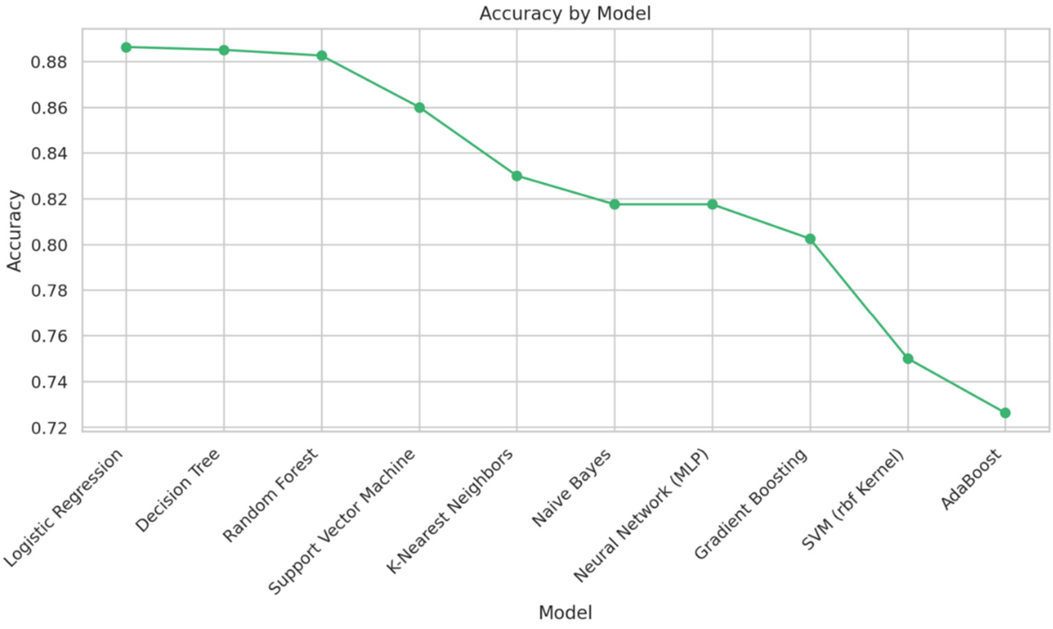
From the table, it is evident that the binary Waterwheel Plant Algorithm (bWWPA) stands out with the lowest average error of 0.52153. This remarkable performance underscores the algorithm's ability to select the most significant features, a crucial aspect in predicting the quality of apples. The bWWPA's success in identifying the main features that influence apple quality enhances the accuracy and reliability of the feature subset for subsequent modeling, a key finding of our study.

Following the feature selection, we assessed the performance of several classification models. The results are given in Table 2, which summarizes the accuracy, sensitivity (True Positive Rate, TPR), specificity (True Negative Rate, TNR), positive predictive value (PPV), negative predictive value (NPV), and F-Score for each model.

Table 2: Classification Results

Models	**Accuracy**	**Sensitivity (TRP)**	**Specificity (TNP)**	**Pvalue PPV**	**Pvalue NPV**	**FScore**
Logistic Regression	0.88625	0.886364	0.886139	0.884131	0.888337	0.885246
Decision Tree	0.885	0.868687	0.90099	0.895833	0.875	0.882051
Random Forest	0.8825	0.863636	0.90099	0.895288	0.870813	0.879177
Support Vector Machine	0.86	0.835859	0.883663	0.875661	0.845972	0.855297
K-Nearest Neighbors	0.83	0.833333	0.826733	0.825	0.835	0.829146
Naive Bayes	0.8175	0.782828	0.851485	0.837838	0.8	0.809399
Neural Network (MLP)	0.8175	0.782828	0.851485	0.837838	0.8	0.809399
Gradient Boosting	0.8025	0.79798	0.806931	0.80203	0.802956	0.8
SVM (rbf Kernel)	0.75	0.744949	0.75495	0.748731	0.751232	0.746835
AdaBoost	0.72625	0.69697	0.75495	0.736	0.717647	0.715953

From the table, Logistic Regression turned out to be the best model, with the highest accuracy of 0.88625. This model also demonstrated high sensitivity 0.88636, specificity 0.88614, PPV 0.88413, NPV 0.88834, and F-Score 0.88525, which shows that it is very good at predicting the quality of the apple. Decision Tree and Random Forest models also did a great job, with accuracies of 0.885 and 0.8825, respectively. The bWWPA was the most effective feature selection algorithm. It had the lowest average error and showed its ability to select the most important features. Logistic Regression, the best-performing classification model, achieved high accuracy and gave a balanced performance in all the evaluation metrics.



Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

Figure 3: Accuracy by Model

Figure 3 illustrates the accuracy of the classification models used, a key performance metric that measures the ratio of correctly classified instances to the total number of instances. The bar chart, a powerful tool, facilitates a direct comparison of model performance, clearly indicating the models that were most successful in predicting apple quality. This chart is instrumental in identifying the most reliable models and understanding their key strengths in terms of accuracy.

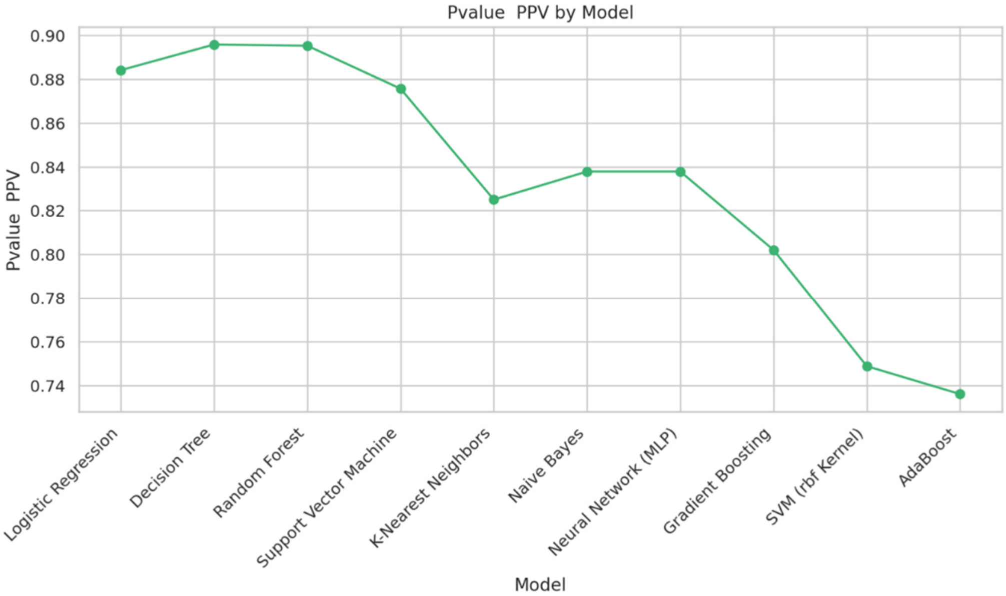


Figure 4: Pvalue PPV by Model

Figure 4 shows the PPV for all the classifiers. This is also known as precision. PPV signifies the fraction of positive predictions that accurately describe reality, therefore, the model's capability to avoid false positives. The bar graph allows for the straight comparison of PPV across models, allowing one to see the relative effectiveness of different models in producing correct, positive forecasts. It is a key indicator of how well each model performs, especially where mistakes are not acceptable. The power of joint feature selection of bWWPA and classifier of Logistic Regression in our research is beyond doubt. This strategic mix of the two techniques highlighted their strengths and finally culminated in better performance predictions. This aided in capturing the essential elements to build a more robust and accurate model, which highlighted the reliability of our approach.

In sum, this study's outcome established the effectiveness of predicting apple quality using feature selection algorithms and robust classification models in conjunction. The information used during the feature selection process is vital to improving agricultural practices and meeting consumers' expectations.

5. Conclusion

This study evaluated apple quality using advanced feature selection and classification techniques. The binary Waterwheel Plant Algorithm (bWWPA) proved to be the best feature selection method, identifying the most important features such as sweetness, crunchiness, juiciness, ripeness, acidity, size, and weight. Logistic Regression emerged as the most effective classification model, achieving the highest accuracy of 0.88625 and demonstrating balanced performance across other metrics. The combination of bWWPA and Logistic Regression provided a robust and reliable method for predicting apple quality. These findings are highly significant for the agricultural industry, impacting areas such as quality control, cultivation practices, consumer satisfaction, and data-driven decision-making. Future research should focus on using larger and more diverse datasets, integrating additional features, exploring advanced modeling techniques, and developing real-time quality monitoring systems. However, the study's limitations include the limited scope of features, the potential for overfitting, reliance on a single dataset, and the need for more consideration for environmental variability. Addressing these limitations in future research will further enhance Apple's quality assessment and management.

Funding: “This research received no external funding”

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Ashraf, A. R., Somogyi-Végh, A., Merczel, S., Gyimesi, N., & Fittler, A. (2024). Leveraging code-free deep learning for pill recognition in clinical settings: A multicenter, real-world study of performance across multiple platforms. Artificial Intelligence in Medicine, 150, 102844. https://doi.org/10.1016/j.artmed.2024.102844

[2] Bai, S., Shi, S., Han, C., Yang, M., Gupta, B. B., & Arya, V. (2024). Prioritizing user requirements for digital products using explainable artificial intelligence: A data-driven analysis on video conferencing apps. Future Generation Computer Systems, 158, 167–182. https://doi.org/10.1016/j.future.2024.04.037

[3] Doudesis, D., Lee, K. K., Yang, J., Wereski, R., Shah, A. S. V., Tsanas, A., Anand, A., Pickering, J. W., Than, M. P., Mills, N. L., Strachan, F. E., Tuck, C., Shah, A. S., Anand, A., Chapman, A. R., Ferry, A. V., Lee, K. K., Doudesis, D., Bularga, A., … Duncan, C. (2022). Validation of the myocardial-ischaemic-injury-index machine learning algorithm to guide the diagnosis of myocardial infarction in a heterogenous population: A prespecified exploratory analysis. The Lancet Digital Health, 4(5), e300–e308. https://doi.org/10.1016/S2589-7500(22)00025-5

[4] Feng, Y., Mei, D., & Zhao, H. (2023). Auction-based deep learning-driven smart agricultural supply chain mechanism. Applied Soft Computing, 149, 111009. https://doi.org/10.1016/j.asoc.2023.111009

[5] Giannakos, M., Voulgari, I., Papavlasopoulou, S., Papamitsiou, Z., & Yannakakis, G. (2020). Games for Artificial Intelligence and Machine Learning Education: Review and Perspectives. In M. Giannakos (Ed.), Non-Formal and Informal Science Learning in the ICT Era (pp. 117–133). Springer. https://doi.org/10.1007/978-981-15-6747-6_7

[6] Iqbal, M. J., Javed, Z., Sadia, H., Qureshi, I. A., Irshad, A., Ahmed, R., Malik, K., Raza, S., Abbas, A., Pezzani, R., & Sharifi-Rad, J. (2021). Clinical applications of artificial intelligence and machine learning in cancer diagnosis: Looking into the future. Cancer Cell International, 21(1), 270. https://doi.org/10.1186/s12935-021-01981-1

[7] Jackulin, C., & Murugavalli, S. (2022). A comprehensive review on detection of plant disease using machine learning and deep learning approaches. Measurement: Sensors, 24, 100441. https://doi.org/10.1016/j.measen.2022.100441

[8] Jones, O. T., Matin, R. N., Schaar, M. van der, Bhayankaram, K. P., Ranmuthu, C. K. I., Islam, M. S., Behiyat, D., Boscott, R., Calanzani, N., Emery, J., Williams, H. C., & Walter, F. M. (2022). Artificial intelligence and machine learning algorithms for early detection of skin cancer in community and primary care settings: A systematic review. The Lancet Digital Health, 4(6), e466–e476. https://doi.org/10.1016/S2589-7500(22)00023-1

[9] Manlhiot, C., van den Eynde, J., Kutty, S., & Ross, H. J. (2022). A Primer on the Present State and Future Prospects for Machine Learning and Artificial Intelligence Applications in Cardiology. Canadian Journal of Cardiology, 38(2), 169–184. https://doi.org/10.1016/j.cjca.2021.11.009

[10] Nayyar, A., Gadhavi, L., & Zaman, N. (2021). Chapter 2 - Machine learning in healthcare: Review, opportunities and challenges. In K. K. Singh, M. Elhoseny, A. Singh, & A. A. Elngar (Eds.), Machine Learning and the Internet of Medical Things in Healthcare (pp. 23–45). Academic Press. https://doi.org/10.1016/B978-0-12-821229-5.00011-2

[11] Prabha, K. (2021). Disease sniffing robots to apps fixing plant diseases: Applications of artificial intelligence in plant pathology—a mini review. Indian Phytopathology, 74(1), 13–20. https://doi.org/10.1007/s42360-020-00290-3

[12] Sahoo, A., Rathi, A., Bashishth, S., Roy, S., & Pradhan, C. (2023). Predictive Farmland Optimization and Crop Monitoring Using Artificial Intelligence Techniques. In M. A. Ahad, G. Casalino, & B. Bhushan (Eds.), Enabling Technologies for Effective Planning and Management in Sustainable Smart Cities (pp. 79–121). Springer International Publishing. https://doi.org/10.1007/978-3-031-22922-0_4

[13] Wieme, J., Mollazade, K., Malounas, I., Zude-Sasse, M., Zhao, M., Gowen, A., Argyropoulos, D., Fountas, S., & Van Beek, J. (2022). Application of hyperspectral imaging systems and artificial intelligence for quality assessment of fruit, vegetables and mushrooms: A review. Biosystems Engineering, 222, 156–176. https://doi.org/10.1016/j.biosystemseng.2022.07.013

[14] Yao, J., Tran, S. N., Sawyer, S., & Garg, S. (2023). Machine learning for leaf disease classification: Data, techniques and applications. Artificial Intelligence Review, 56(3), 3571–3616. https://doi.org/10.1007/s10462-023-10610-4

[15] Ayoub Shaikh, T., Rasool, T., & Rasheed Lone, F. (2022). Towards leveraging the role of machine learning and artificial intelligence in precision agriculture and smart farming. Computers and Electronics in Agriculture, 198, 107119. https://doi.org/10.1016/j.compag.2022.107119

Doi : https://doi.org/10.21608/jaiep.2024.355003
Received: 22 March 2024; Revised: 19 April 2024; Accepted: April 29, 2024

[16] Behera, S. K., Rath, A. K., Mahapatra, A., & Sethy, P. K. (2020). Identification, classification & grading of fruits using machine learning & computer intelligence: A review. Journal of Ambient Intelligence and Humanized Computing. https://doi.org/10.1007/s12652-020-01865-8

[17] Benmouna, B., García-Mateos, G., Sabzi, S., Fernandez-Beltran, R., Parras-Burgos, D., & Molina-Martínez, J. M. (2022). Convolutional Neural Networks for Estimating the Ripening State of Fuji Apples Using Visible and Near-Infrared Spectroscopy. Food and Bioprocess Technology, 15(10), 2226–2236. https://doi.org/10.1007/s11947-022-02880-7

[18] Çetin, N., & Sağlam, C. (2022). Rapid detection of total phenolics, antioxidant activity and ascorbic acid of dried apples by chemometric algorithms. Food Bioscience, 47, 101670. https://doi.org/10.1016/j.fbio.2022.101670

[19] Crawford, K., & Paglen, T. (2021). Excavating AI: The politics of images in machine learning training sets. AI & SOCIETY, 36(4), 1105–1116. https://doi.org/10.1007/s00146-021-01162-8

[20] Elfiky, N. (2023). Application of Artificial Intelligence in the Food Industry: AI-Based Automatic Pruning of Dormant Apple Trees. In A. E. Hassanien & M. Soliman (Eds.), Artificial Intelligence: A Real Opportunity in the Food Industry (pp. 1–15). Springer International Publishing. https://doi.org/10.1007/978-3-031-13702-0_1

[21] Lee, V.-H., Hew, J.-J., Leong, L.-Y., Tan, G. W.-H., & Ooi, K.-B. (2020). Wearable payment: A deep learning-based dual-stage SEM-ANN analysis. Expert Systems with Applications, 157, 113477. https://doi.org/10.1016/j.eswa.2020.113477

[22] Malounas, I., Lentzou, D., Xanthopoulos, G., & Fountas, S. (2024). Testing the suitability of automated machine learning, hyperspectral imaging and CIELAB color space for proximal in situ fertilization level classification. Smart Agricultural Technology, 8, 100437. https://doi.org/10.1016/j.atech.2024.100437

[23] Pourdarbani, R., Sabzi, S., Rohban, M. H., García-Mateos, G., Paliwal, J., & Molina-Martínez, J. M. (2022). Using metaheuristic algorithms to improve the estimation of acidity in Fuji apples using NIR spectroscopy. Ain Shams Engineering Journal, 13(6), 101776. https://doi.org/10.1016/j.asej.2022.101776

[24] Roy, K., Chaudhuri, S. S., & Pramanik, S. (2021). Deep learning based real-time Industrial framework for rotten and fresh fruit detection using semantic segmentation. Microsystem Technologies, 27(9), 3365–3375. https://doi.org/10.1007/s00542-020-05123-x

[25] Apple Perfection: (n.d.). Retrieved May 15, 2024, from https://www.kaggle.com/datasets/zeesolver/apple-quality

[26] Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., & Lang, M. (2020). Benchmark for filter methods for feature selection in high-dimensional classification data. Computational Statistics & Data Analysis, 143, 106839. https://doi.org/10.1016/j.csda.2019.106839

[27] Chauhan, N. K., & Singh, K. (2022). Performance Assessment of Machine Learning Classifiers Using Selective Feature Approaches for Cervical Cancer Detection. Wireless Personal Communications, 124(3), 2335–2366. https://doi.org/10.1007/s11277-022-09467-7

[28] Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., & Saboor, A. (2020). Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. IEEE Access, 8, 107562–107582. https://doi.org/10.1109/ACCESS.2020.3001149

[29] Pande, S., Khamparia, A., & Gupta, D. (2023). Feature selection and comparison of classification algorithms for wireless sensor networks. Journal of Ambient Intelligence and Humanized Computing, 14(3), 1977–1989. https://doi.org/10.1007/s12652-021-03411-6

[30] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2. https://doi.org/10.3389/fbinf.2022.927312

[31] Raj, R. J. S., Shobana, S. J., Pustokhina, I. V., Pustokhin, D. A., Gupta, D., & Shankar, K. (2020). Optimal Feature Selection-Based Medical Image Classification Using Deep Learning Model in Internet of Medical Things. IEEE Access, 8, 58006–58017. https://doi.org/10.1109/ACCESS.2020.2981337