



Comparison of Performance of Some Classification Methods to Evaluate the Quality of Vegetables from its Morphology

Joy Deb ¹ * , Dibyojyoti Bhattacharjee ¹ 

¹ Department of Statistics, Assam University

Abstract

One important aspect of Data Science is its ability to classify subjects into non-overlapping groups based on one or several input variables. Several methods and algorithms are available in the literature for classifying subjects based on the values of multiple observed variables. Such classification tools are Naive Bayesian Classifiers, Logistic Regression, Discriminant Analysis, k-nearest neighborhood etc. This paper attempts to recognize if the morphological variables, identified either through literature review or from expert opinion, can be utilized to understand the quality of vegetables. Consequently, the current researchers obtained primary data about the morphology of the vegetables through experimentation. The outcome variable is the quality of the vegetables classified as eatable or not-eatable because of worm attack. Several classification methods are then compared for the classification exercise by building the model based on the training sample and testing the performance of the models in the holdout sample. Methods of classification performance statistics like sensitivity, specificity, precision etc. are used for their comparison. The study finds that Naive Bayes and Logistic Regression models perform better for this classification exercise. For example, only eggplant (brinjal) is considered for the study.

Keywords: Machine learning; classification; morphology of Vegetables; data science.

1. Introduction

Vegetables plays a significant role in individuals' daily diet. Vegetables offer essential nutrients, contain minimum fat and carbohydrates, but are rich in vitamins, minerals, and dietary fibre. Vegetables are excellent minerals providers, particularly calcium, iron, vitamins A and C [1]. In the financial year 2022-23 India produced an estimated 200 million metric tons of vegetables. These vegetables include, among others, potatoes, onions, eggplants and cabbage.

The eggplant (*Solanum melongena* L.) is called Brinjal in South Asia, particularly in India, Pakistan and Bangladesh. The name "eggplant" was given by the USA and Canada because some varieties are shaped like eggs [3]. Eggplant is a versatile crop suitable for cultivation in different agro-climatic regions, with annual cultivation. It is India's second most popular vegetable, after potatoes, tomato, and onion [3]. India has various eggplants with differing tastes, shapes, colours and sizes. Cultivated varieties of eggplants exhibit a wide range of sizes, including small to large, and various shapes, from oblong to round, and oval to club-shaped. They come in diverse colors, including green, white, yellow, and a spectrum of purple hues, spanning from nearly black to striped patterns and gradients [4].

Eggplants are a great provider of many vitamins and nutrients, including protein, fat, Carbohydrates, sugar, vitamin A, vitamin B complex, etc. Additionally, contains the full complement of vitamins, minerals, nutritious fiber, antioxidants, and phytochemicals with scavenging properties [5] [6]. According to the production volume of vegetables in India FY, 2008-2022, on average, 12.98 million metric tons of brinjal are produced in India [2].

Ensuring optimal nutrition requires selecting high-quality vegetables from the market. Unfortunately, pest infestations pose a challenge to consistently obtaining such vegetables. Insects contaminate the vegetables, degrading their quality and making them unsuitable for human consumption. A significant risk to brinjal crops is the brinjal fruit and shoot borer, "Leucinodes orbonalis Guenee," which is responsible for suspectable damages and losses in production. As per AVRDC's research on eggplant entomology, *Leucinodes orbonalis* stands out as the most prominent and severely impacting pest across several Asian nations, including but not limited to India, Pakistan, Sri Lanka, Nepal, Bangladesh, Thailand, and the Philippines [7]. *Leucinodes orbonalis* has four distinct growth stages such as egg, larva, pupa, and adult. Usually, eggs are typically laid one by one on the undersides of young leaves, green stems, flower buds, or the calyces of fruit. Afterwards, the larvae infiltrate flower buds and also gain entry in the same way. During the reproductive period, they invade susceptible fruits through the calyx [8].

*Corresponding author

E-mail address: jdeb48389@gmail.com



Figure 1. *The picture of infected eggplant.*

Brinjal is a popular vegetable that is frequently consumed. However, the quality of brinjal compromised due to insect infestations. This problem in brinjal is a multifaceted issue having profound implications for consumers health and financial losses. Consumers often lack the expertise in accurately identifying such infested products, potentially exposing themselves to health risks and economic losses. Furthermore, infestation may not be easily detectable at the time of purchasing but become evident only during the initial stage of preparation for consumption. This situation poses a significant challenge for consumers in identifying infested brinjal eventually leading to potential financial losses. Addressing this issue is essential to ensure that consumers can make informed decisions about the quality of the vegetables they purchase, promoting food safety and reducing economic losses.

In the field of agriculture, several research studies have utilised classification techniques. For instance, Ajaz and Hussain researched seed classification based on morphological features [9]. At the same time, Malyadri, M.S., and J. focused on developing a classification model to identify suitable crops for specific soil and weather conditions, along with the corresponding fertilizer recommendations [10]. Padoa and Maravillas employed the Naïve Bayesian method for plant classification [11], while Bishnoi et al. utilized classification tools to identify cotton genotypes [12]. Gawande and Dhande reviewed the application of classification techniques in grading fruits based on quality before packaging [13].

Additionally, Lauguico et al. performed lettuce life stage classification using texture attributes [14]. Iroliam et al. employed different machine learning techniques to predict okra's shelf life using various parameters. They found that SVM, Naïve Bayes, and decision tree algorithms effectively predicted okra's shelf life [15]. Davis et al. predicted added sugar content in packaged foods using machine learning, with k-NN showing similar capability to explain variation in added sugar compared to existing approaches [16]. Despite brinjal (eggplant) being India's second most popular vegetable, previous research in agriculture utilising machine learning has yet to address infested brinjals' identification adequately. This research gap motivates our study, where we aim to identify infested eggplants based on their morphological traits using machine learning techniques. Through an assessment of different machine learning techniques, our objective is to identify the most efficient approach for identifying pest infestations in eggplants.

2. Variable Selection and Different Classification Tool

This section presents the variable selection process and the different classification tools employed in this study.

2.1. Variable selection

Here we investigated six morphological characteristics related to fruit and shoot borer resistance. These characteristics were fruit hardness, deformity, surface spots, density (Weight/Volume), stem colour distance, and fruit colour distance. Some of these characteristics were identified based on a literature review. For instance, Krishnaiah & Vijay reported that lower susceptibility to borer injury in specific cultivars might be attributed to the hardness of the fruit skin [17]. Garewal & Singh observed that long-fruited varieties were more susceptible to borer infestations [18]. Additionally, Hazra et al. found a correlation between eggplant weight and susceptibility to fruit infestation by the pest [19]. Furthermore, some morphological characteristics were identified through consultations with farmers and experts in the field.

To evaluate fruit hardness, we manually assessed the fruit's texture using tactile perception. Fruit deformities and surface spots were identified through careful visual examination, analysing their visual attributes.

We used an electric weighing machine to measure the fruit's weight to determine the fruit density. Concurrently, the volume of the fruit was measured using the water displacement method, enabling us to calculate the density as weight per volume accurately.

We used the "Color Identification" application to measure the RGB values of individual stems to assess stem colour distance. We obtained the stem colour distance by subtracting its RGB value from its

corresponding standard RGB value. The standard RGB value for the stem was calculated by averaging the RGB values obtained from various stems. Similarly, the fruit colour distance was estimated using the same approach. The Euclidean distance formula was employed to calculate the distance measurements.

Let x and y be two colour surfaces, the Euclidean distance of RGB value of and is given by,

$$d(x, y) = \sqrt{(r_x - r_y)^2 + (g_x - g_y)^2 + (b_x - b_y)^2} \quad (1)$$

Where r_x, g_x and b_x represents the RGB value as its individual red, green, and blue components of x -surface respectively while r_y, g_y and b_y represent the red, green and blue of RGB value of y -surface respectively.

2.2. Different Classification Tools

2.2.1 k-Nearest Neighbour (k-NN) Classifier

The k-Nearest Neighbour (k-NN) algorithm is a nonparametric and supervised classification method. It was originally developed by Fix & Hodges [20] and subsequently expanded upon by Cover [21]. It works by categorizing data points based on their proximity to nearby neighbours. To improve validity, multiple neighbours are considered, hence the name k-Nearest Neighbour (k-NN) classifier. In this approach, the class of a new observation is determined by analysing the k nearest neighbours in the dataset.

This algorithm classified as Lazy Learning, as it only stores and memorizes the training data without performing significant computation during the time of training. As a results, it doesn't generalize the training dataset. Consequently, during the testing phase the entire training dataset is needed [22]. The Nearest Neighbourhood classifier is used to classify a new, unlabeled observation based on the positions of its neighbouring data points.

2.2.2 Naïve Bayesian Classifier

The Naïve Bayesian classifier stands out as one of the most frequently employed machine learning techniques for classification tasks. It operates on probabilistic principles rooted in the Bayes theorem and offers flexibility in handling an arbitrary number of independent variables, whether they are continuous or categorical [23]. The Naïve Bayes classifier assumes that all the predictor values are independent. In other words, the presence of one particular feature in a class does not affect the presence of another one. This concept is known as class conditional independence [24]. A significant advantage of this classifier is its ability to perform well with small training datasets and it can also handle incomplete data, including instances with missing values.

2.2.3 Logistic Regression

Logistic Regression (LR) is a widely used supervised classification technique that derive from the field of Statistics. Its main goal is to discover the appropriate model to explain how a categorical dependent variable is related to one or more independent variables.

2.2.4 Linear Discriminant Analysis (LDA)

Another effective classification technique is Linear Discriminant Analysis. It was developed by Ronald A. Fisher in the year 1936. LDA is commonly used for dimension reduction techniques, i.e., separating two or more classes. According to Manage et al. [25], LDA creates the axes that effectively distinguish or separate different classes in the data. The LDA algorithm divides and uses linear boundaries to categorize data based on the predictor variables.

2.2.5 Decision Tree

A decision tree (DT) was built by Hunt et al. [26]. It is an essential nonparametric supervised machine learning algorithm. A tree structure with a root node, branches, internal nodes, and leaf nodes characterizes it. It is often drawn from left to right, starting at the root and moving downward. The root node is the node from which the tree begins. Leaf nodes are the nodes at the endpoints of chains. Internal nodes are not leaf nodes and may stretch across two or more branches. While branches show a range of values, nodes reflect a specific attribute. Different feature selection methods exist for nodes, and by those methods, many decision tree construction algorithms exist. CART is the most widely used algorithm for creating decision trees. Depending on whether the dependent variable is categorical or continuous, the nonparametric decision tree algorithm CART generates classification or regression trees [27].

2.2.6 Random Forest

Leo Breiman developed the Random Forest (RF) algorithm in 2001 [28]. It is a supervised ensemble machine learning algorithm. The fundamental idea behind the ensemble technique is that it's like having a team of different experts who aren't super accurate on their own, but when they all work together, they become really good at making predictions. In random forests, decision trees serve as the fundamental classifier. A random

subsample of data from the available sample generates each decision tree. In each decision split, features are chosen at random in Random Forest. The association between the trees and the prediction accuracy decreases as the features are randomly chosen. The random forest is appropriate for high-dimensional data sets because it can handle categorical, continuous, and missing values.

The Random Forest algorithm is executed through the following steps:

- If the training group consists of N cases selected with replacement, then N cases are randomly taken from the original data to serve as the training group for the tree's growth.
- For an M -variable input, the variable m is selected in such a way that $m \ll M$ for each node. m variables are then randomly selected from M , and used to split the node. The value of m remains constant throughout the forest's widening.
- Each tree in the Random Forest is permit to grow its full potential without any pruning or trimming [29].

2.3 Performance Evaluation Method

In this study, we evaluate the performance of the classification models by assessing their accuracy and error rate. These metrics are commonly used to measure the effectiveness of classifiers and are derived from the confusion matrix. The confusion matrix comprehensively summarises the model's predictions and the actual class labels. The accuracy is determined by dividing the total number of correct predictions by the total number of observations. By examining these measures and analysing the confusion matrix, we gain insights into the performance of the classification models in our study.

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + fn + tn} \tag{2}$$

Where, tp = true positive, tn = true negative, fp = false positive and fn = false negative. When there is a significant imbalance in the number of observations between classes, accuracy and error rate may not be reliable performance indicators [30]. Hence, additional metrics are employed to evaluate the classification model's performance comprehensively. Sensitivity and G-mean are positive measures, ranging from 0 to 1, with higher values indicating better performance. Specificity, precision, and F1 score are also positive measures, with values ranging from 0 to 1, where higher values indicate better performance. These metrics, derived from the confusion matrix of each respective classification model, provide a more comprehensive assessment of the model's performance beyond accuracy.

$$\text{Sensitivity} = \frac{tp}{tp + fn} \tag{3}$$

$$\text{Specificity} = \frac{tn}{tn + fp} \tag{4}$$

$$\text{Precision} = \frac{tp}{tp + fp} \tag{5}$$

$$\text{F1 score} = \frac{2 \times \text{sens} \times \text{prec}}{\text{sens} + \text{prec}} \tag{6}$$

$$\text{G-mean} = \sqrt{\text{sens} \times \text{prec}} \tag{7}$$

3. Findings and Deliberations

This segment presents the findings and deliberation of the data analysis phase, where the dataset was partitioned into training and holdout samples in an 80:20 ratio. To ensure the robustness and reliability of the findings, the analysis was conducted through ten iterations with replacement. This approach accounts for the variations in model performance due to specific data instances and allows for a comprehensive evaluation.

Table 1 presents a comprehensive overview of the results obtained from the various methods utilized, encompassing metrics such as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). Additionally, it includes the average accuracy percentage computed over ten iterations. These findings offer insights into how well each method can detect and categorize the target variable, offering valuable information about their performance and appropriateness for the task at hand.

Table 1. Confusion matrix results for 10 iterations of sample size 80:20

Method	Correct Prediction		Wrong Prediction		Average Accuracy per cent
	TP	TN	FP	FN	
k-Nearest Neighbour	10,12,9,9,11, 8,7,7,10,12	4,9,9,11,12 13,11,6,10,7	1,5,3,3,1 2,5,2,3,1	6,2,5,5,2 6,2,2,6,10	72.4
Naïve Bayes	8,4,8,13,13,8, 10,7,10,10	14,7,6,8,7,6,7, 13,12,13	3,4,2,6,2,8,4,7,0,2	0,5,0,2,1, 3,4,2,3,3	75.30
Logistic Regression	12,7,6,11,9,8, 16,10,5,13	8,15,13,11,12, 11,17,8,9,14	2,4,5,0,0,2,5,2,3,4	3,2,2,3,3, 2,2,4,2,0	80.81
Linear Discriminant Analysis	7,5,9,9,5,9, 9,9,10,9	9,12,10,5,5,8, 10,11,13,15	2,3,9,4,2,2,4,2,4,2	4,5,0,4,3, 5,3,6,1,1	72.52
CRAT	10,11,8,12,7, 6,7,5,3,8	18,12,6,12,9, 10,10,15,11,9	6,4,3,1,4,1,2,5,4,3	1,1,1,5,3, 3,2,0,3,6	76.25
Random Forest	4,6,10,10,11, 8,8,9,8,9	11,9,8,11,9, 13,8,14,6,10	4,7,2,4,4,7,3,2,4,4	0,2,6,6,2, 1,1,3,4,5	72.14

A critical aspect of implementing the *k*-NN classifier is choosing the optimal value for *k*. In this study, the square root of the total of observations in the train sample determined *k* = 11. However, it is worth noting that the *k*-NN classifier may introduce bias due to using the Euclidean distance, which tends to favour variables with larger values. A subsequent step to address this potential bias involved normalizing the data using the min-max normalization function. The *k*-NN model was then constructed using the training data, and the testing data was utilised for cross-validation purposes. For the *k*-NN model, due to multiple iterations, the accuracy shows fluctuations, making it challenging to pinpoint precise results. However, when averaging the outcomes over all samples, the model achieved an accuracy rate of 72.4 per cent.

The Naïve Bayes model also exhibited varying accuracy across ten iterations. On average, the model demonstrated an overall accuracy of 75.3 per cent, indicating its competitive performance.

In contrast, the Logistic Regression model outperformed the others with an average accuracy of 80.80 per cent, showcasing its effectiveness in classifying the data.

The Linear Discriminant Analysis (LDA) model achieved an average accuracy of 72.58 per cent across ten iterations, showing its capability to perform reasonably well on the given task.

For the decision tree constructed using the CRAT algorithm, the model attained an average accuracy of 76.25 per cent after ten iterations, making it a viable choice for classification tasks.

In the case of the Random Forest model, a crucial aspect was determining the appropriate number of trees. In this study, 1000 decision trees were utilized to achieve better performance and accuracy. On average, across ten iterations, the Random Forest model achieved an accuracy of 72.14 per cent, demonstrating its suitability for the task.

Overall, the Logistic Regression model showed the highest average accuracy, followed closely by the decision tree constructed using the CRAT algorithm. These findings provide valuable insights into the effectiveness of different machine learning models and can be helpful to guide the selection of the most fruitful approach for similar classification tasks.

In an effort to explore how these models perform with smaller training samples, we conducted an additional test using a 60:40 samples split, as presented in **Table 2**. It is evident from **Table 2**, that the *k*- Nearest Neighbour (*k*-NN) model exhibited variability in accuracy across ten iterations, ultimately yielding an average accuracy of 74.68 per cent. Notably, Naïve Bayes consistently delivered a competitive performance, demonstrating an average accuracy of 75.03 per cent, reaffirming its reliability in the context of pest infestation identification. Likewise, the Logistic Regression model remain effective, with an average accuracy of 74.67 per cent, albeit slightly lower than the previous results, underscoring its robust classification capabilities.

Linear Discriminant Analysis (LDA), consistently achieved an average accuracy of 74.73 per cent, reflecting its reliability in delivering consistent and reasonable performance. Moreover, both the CRAT and Random Forest models showcased their suitability for the task, achieving average accuracies of 73.88 per cent and 72.72 per cent respectively. The findings from the 60:40 sample split indicate that the performance patterns of these models remain consistent with the 80:20 results.

Table 2. Confusion matrix results for 10 iterations of sample size 60:40

Method	Correct Prediction		Wrong Prediction		Average Accuracy per cent
	TP	TN	FP	FN	
k-Nearest Neighbour	28,19,19,11,14, 15,15,17,21,16	18,23,22,17,18, 15,16,21,21,26	4,7,2,9,9, 1,2,5,4,8	13,6,9,5,4, 12,7,4,8,6	74.68
Naïve Bayes	13,15,19,17,19, 19,20,24,19,18	26,22,20,23,18, 22,24,21,16,18	9,4,4,4,6, 4,7,11,6,8	6,6,8,4,5, 15,4,4,10,7	75.03
Logistic Regression	17,15,19,17,16, 14,20,12,14,18	25,22,25,16,18, 25,23,22,22,17	8,8,8,7,10, 5,3,7,9,6	4,5,3,5,6, 6,10,7,4,6	74.67
Linear Discriminant Analysis	18,20,17,7,18, 17,19,14,15,18	18,21,17,24,17, 16,24,26,22,21	7,10,8,3,10, 8,4,5,10,6	4,2,4,7,2, 5,11,11,3,5	74.73
CRAT	15,10,19,9,14, 12,16,17,8,10	15,18,17,21,18, 23,23,25,16,23	6,10,5,8,7, 7,8,8,6,4	5,10,2,2,6, 4,2,4,5,6	73.88
Random Forest	13,19,17,13,14, 18,18,14,18,19	21,21,19,20,16, 19,23,19,25,21	11,5,7,7,14, 3,6,10,4,2	4,8,8,7,6, 7,8,3,7,10	72.72

3.7 Comparison of different classification models

Figure 2. illustrates the results of the various classification models analysed in this study. According to the findings, the logistic regression model exhibited the highest accuracy rate of 80.80 per cent, indicating success.

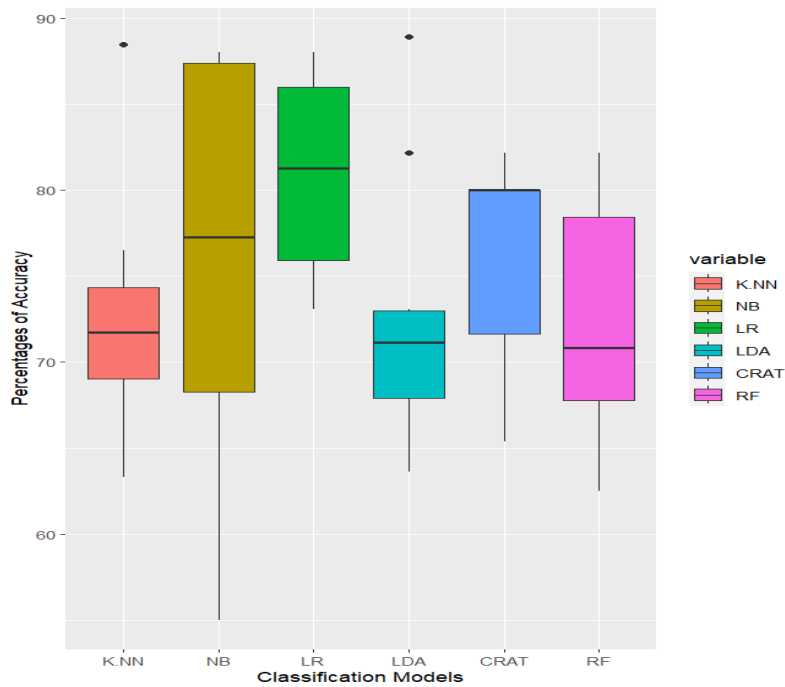


Figure 2. Boxplot of the accuracy for the different classification model

When considering the F1 score and G-mean metrics, the Naive Bayes and Logistic Regression models consistently rank at the forefront. These models consistently exhibit strong performance in both F1 score and G-mean, showcasing their effectiveness in accurately classifying the data.

Based on the results, the Naive Bayes and Logistic Regression models consistently outperform the other models in detecting infected eggplants. These two models exhibit superior performance across multiple evaluation metrics, indicating their effectiveness in accurately identifying infested samples.

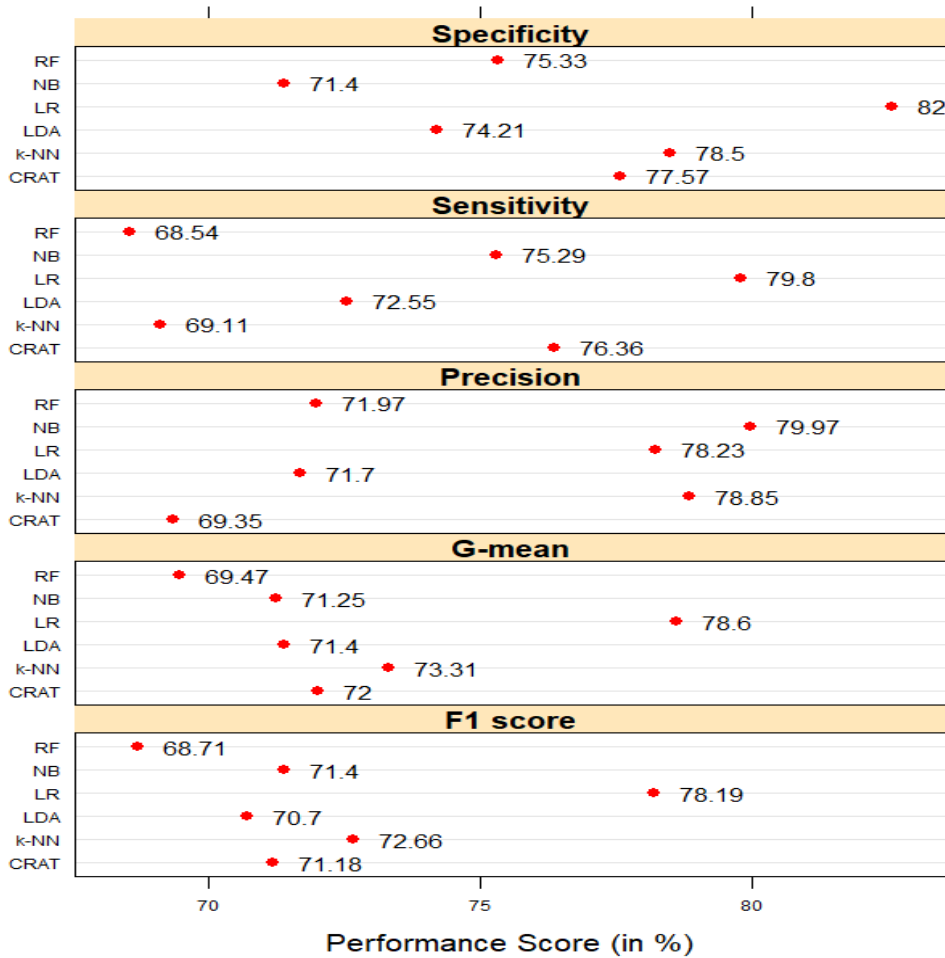


Figure 3. Sensitivity, Specificity, Precision, F1 score and G-mean for different classification models classification models.

4. Conclusion

The goal of this research is to assess the effectiveness different supervised classification models, namely *k*-Nearest Neighbour (*k*-NN), Naïve Bayes (NB), Logistic Regression (LR), Linear Discriminant Analysis (LDA), CRAT of decision tree, and Random Forest. The aim is to evaluate the efficiency of these models in classification.

The study considers six morphological characteristics of the eggplants, including fruit hardness, deformity, surface spots, density (Weight/Volume), stem colour distance, and fruit colour distance. The database is categorized into two parts: training data for build-up models and testing data for cross-validation. The accuracy of the models is assessed using a confusion matrix that recorded accurate and inaccurate classifications for all classes. The analysis revealed that Naive Bayes, logistic regression, and CRAT consistently demonstrated favorable performance, particularly when considering the average accuracy over ten iterations.

The *k*-NN model achieves an average accuracy of 72.4 per cent in classifying infested eggplants. Comparatively, the NB model exhibits a relatively higher accuracy of 75.3 per cent. The LR model demonstrates the highest accuracy among all classification models, with a rate of 80.8 per cent. Following LR, the accuracy scores for LDA, CRAT, and RF were 72.52 per cent, 76.25 per cent, and 72.14 per cent, respectively. When considering the classification of infested eggplants, the Logistic Regression model emerges as the most accurate among all other models, with CRAT and Naive Bayes occupying the second and third positions, respectively. However, Random Forest has poor accuracy compared to classification models in this dataset.

To facilitate a comprehensive comparison, we computed additional performance measures, including sensitivity, specificity, accuracy, F1 score, and G-mean. Our analysis revealed that logistic regression achieved the highest performance value among the evaluated metrics. Following closely, Naive Bayes and CRAT decision trees ranked second and third, respectively, for the classification of infected eggplants. This assessment allowed for a more detailed evaluation of the models' effectiveness in accurately classifying the target samples.

Based on the findings, it can be concluded that the logistic regression, Naive Bayes, and CRAT decision tree models demonstrate sufficient capability to categorize the infected eggplants effectively. These models exhibit promising performance across various evaluation metrics, suggesting their suitability for accurate classification in this context.

In agriculture, there is a significant potential for future research utilizing machine learning techniques. Currently, our focus is primarily on one specific vegetable, namely eggplant. However, the study can encompass a broader range of vegetables. Also, utilizing machine learning to assess the quality attributes of vegetables, such as ripeness, freshness, and nutritional content can be studied. This information can aid in quality control and consumer satisfaction.

Furthermore, using regression-based machine learning tools, one can evaluate the nitrogen content in vegetables, as elevated levels of nitrates have been associated with endogenous nitrosation. This process has been implicated in developing thyroid conditions, several types of human cancers, neural tube defects, and diabetes. Assessing the nitrogen content levels in eggplant or in any other horticultural crop can disclose their potential effect on health.

References

- [1] Ulger TG, Songur AN, Cirak O and Cakiroglu FP, "Role of Vegetables in Human Nutrition and Disease Prevention", in *Vegetables-Importance of Quality Vegetables to Human Health*, Intechopen, 2018, pp. 7-32; doi: 10.5772/intechopen.77038.
- [2] Gowda LR, "Genetically Modified Aubergine(Also Called Brinjal or Solanum melongena)" in *Genetically Modified Organisms in Food*, (2016), 27-37; doi: 10.1016/B978012802259700004-X.
- [3] S. Herbst, "The New Food Lover's Companion: Comprehensive Definitions of Nearly 6,000 Food, Drink, and Culinary Terms. Barron's Cooking Guide," *Hauppauge, NY : Barron's Educational Series. ISBN 0764112589*.
- [4] Y. Noda , T. Kaneyuki, K. Igarashi and A. Mori, "Antioxidant activity of nasunin, an anthocyanin in eggplant peels," *Toxicology*, pp. 119-123, 2000.
- [5] B. Whitaker and J. Stommel, "Distribution of Hydroxycinnamic Acid Conjugates in Fruit of Commercial Eggplant (*Solanum melongena* L.) Cultivars," *Journal of Agricultural Food Chemistry*, vol. 51, pp. 3448-3454, 2003.
- [6] A. Minhas, "Production volume of vegetables India FY 2008-2022," 22 03 2023. [Online]. Available: <http://www.statista.com>.
- [7] AVRDC Eggplant entomology, "Control of eggplant fruit and shoot borer. Progress Report," Asian Vegetable Research and Development Center,(AVRDC), Shanhua,Taiwan, 1994.
- [8] E. A. Netam M ., "Screening of shoot and Fruit Borer(*Leucinodes orbonalis* Guenee) for Resistance in Brinjal (*Solanum melongena* L.) Germplasm Lines," *International Journal of current Microbiology and Applied Sciences* 7.2, pp. 3700-3706, 2018.
- [9] R. H. Ajaz and L. Hussain, "Seed Classification using Machine Learning Techniques," *Journal of Multidisciplinary Engineering Science and Technology(JMEST)*, pp. 1098-1102, 2015.
- [10] Manoj Kumar D P, Malyadri N, Srikanth MS, Ananda Babu J, "A Machine Learning model for Crop and Fertilizer recommendation", *Natural Volatiles & Essential Oils*, 8(5) (2021) 10531-10539.
- [11] Padoa F.R.F. & Maravillas E. A. "Using Naive Bayesian method for plant Leaf classification based on shape and texture features" *2015 International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment Management(HNICEM)*, 2015.
- [12] S. Bishnoi, N. A. Ansari, M. Khan, S. Heddami and A. Malik, "Classification of Cotton Genotypes with Mixed Continuous and Categorical Variables: Application of Machine Learning Models," *Sustainability*, pp. 14,13685, 2022.
- [13] A. P. & D. S. S. Gawande, "Implementation of fruit grading system by image processing and data classifier-a review," *International Journal of Engineering Research and General Science*, vol. 2, no. 6, pp. 411-413, 2014.
- [14] S. C. Lauguico, R. I. S. Cocepcion, J. D. Alejandrino, R. R. Tobias and E. P. Dadios, "Lettuce life stage classification from texture attributes using machine learning estimators and feature selection processes," *International Journal of Advances in Intelligent Informatics*, pp. 173-184, 2020.
- [15] I. ., Iorliam, " Application of Machine Learning Techniques for Okra Shelf Life Prediction," *Journal of Data Analysis and Information Processing*, pp. 136-150, 2021.

- [16] Davies T, Yu Louie JC, Ndanuko R, Barbieri S, Perez-Concha O, H Y Wu J, "A Machine Learning Approach to Predict the Added-Sugar Content of Packaged Foods", *The Journal of Nutrition* 152(1), (2022) 343-349.
- [17] K. Krishnaiah and O. Vijay, "Evaluation of brinjal varieties for resistance to shoot and fruit borer," *Indian J Hort*, pp. 84-86, 1975.
- [18] R. Garewal and D. Singh, "Fruit characters of brinjal in relation to infestation by *Leucinodes orbonalis* Guen," *Indian J Ent*, vol. 57, pp. 336-343, 1995.
- [19] P. Hazra, R. Dutta and T. Maity, "Morphological and Biochemical characters associated with field tolerance of brinjal (*Solanum melongena* L.) to shoot and fruit borer and their implication in breed for tolerance," *Indian Journal Genet*, pp. 255-256, 2004.
- [20] E. Fix and J. L. Hodges, "Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties," *USAF School of Aviation Medicine, Randolph Field, Texas*, 1951.
- [21] T. M. Cover and P. E. Hart, " "Nearest neighbor pattern classification", " *IEEE Transactions on Information Theory*, p. 13 (1): 21–27, 1967.
- [22] N. a. N. D. a. T. P. Ali, "Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets," *SN Applied Sciences*, 2019.
- [23] H. Zhang, "The Optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence* .
- [24] P. A. a. L. N. Flac, "Naive Bayesian Classification of Structured Data," *Machine Learning, Boson: Kluwer Academic Publisher*, pp. 1-37, 2004.
- [25] A. W. Manage, " classification of all rounders in limited over cricket - a machine learning approach," *Journal of Sports analytics*, pp. 6(4),295-306, 2020.
- [26] E. Hunt, J. Marin and P. Stone, "Experiment in induction academic press," *N.Y.*, p. 247, 1966.
- [27] L. Breiman, J. Friedman, R. Olshen and C. Stone , "Classification and Regression Trees," *Chapman Hall/ CRC Press: New York, NY, USA*, 1984.
- [28] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5-32, 2001.
- [29] J. Ali, R. Khan and N. Ahmad, "Random Forests and Decision Trees," *IJCSI International Journal of Computer Science Issues*, vol. 9, no. 5, pp. 272-278, 2012.
- [30] M. Kubat, R. Holte and S. Matwin, "Machine learning for the detection of oil spoils in satellite radar images.," *Mach. Learn*, vol. 30, pp. 195-215, 1998.