## ABSTRACT

In the agriculture sector, predicting the quality of raisins is a crucial responsibility to guarantee the constancy of the product quality. In this study, we investigated a number of machine learning models that predict the quality of raisins based on their numerous physical and chemical characteristics. We made use of the Raisin Dataset, which has one target variable and eight characteristics. The three groups of raisins' quality are represented by the goal variable. By addressing missing values and utilising label encoding to transform category variables into numerical values, we preprocessed the dataset. Six machine learning methods, including the Gaussian Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest Classifier, Perceptron, and Logistic Regression, were tested on the dataset. To fine-tune the models' hyperparameters. . Using the test dataset, we utilised GridSearchCV to fine-tune the models' hyperparameters and compare the accuracy ratings of the various models. According to our findings, Random Forest Classifier outperforms the competition with an accuracy of 95.92%, followed by Support Vector Machines (92.94%), and K-Nearest Neighbors (92.09%). The accuracy ratings for the remaining models ranged from 80% to 85%. This study shows how machine learning algorithms can accurately estimate the quality of raisins and recommends using the Random Forest Classifier for classification.



## Keywords

Raisin quality prediction, Machine learning algorithms, Gaussian Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest Classifier, Perceptron, Logistic Regression.


## 1. INTRODUCTION

In recent years, interest in the field of machine learning has increased significantly. It now plays a crucial role in a variety of sectors, including healthcare, banking, entertainment, and transportation. On the basis of a lot of data, machine learning algorithms are used to identify patterns and generate predictions. They are especially helpful for jobs that are too difficult or time-consuming for people to complete. Supervised learning is one of the most popular methods of machine learning, where the algorithm is trained on labelled data to generate predictions on brand-new, untainted data.

The kind of grape used, the growing environment, and the drying procedure are only a few of the variables that can dramatically affect the quality of raisins. Since this could assist producers and distributors in ensuring that their products meet certain standards and possibly even improve the overall quality of the raisins they produce, there has been an increase in interest in using machine learning models to predict the quality of raisins in recent years.

With the development and comparison of several machine learning models, we hope to better understand how to forecast the quality of raisins based on characteristics including moisture content, colour, and sweetness. We will examine the performance of four distinct models in further detail: a linear regression model, a decision tree model, a random forest model, and a gradient boosting model.

To do this, we will first compile a dataset of measurements on the quality of raisins, including characteristics like moisture content, colour, and sweetness. The training set will be used to train each of the four models after we have divided the dataset into training and testing sets. The performance of each model will then be assessed on the testing set using metrics like mean squared error and R-squared. In order to evaluate which model is more effective in forecasting the quality of raisins, we will finally compare the outcomes of each model.

Overall, our research intends to shed light on how well various machine learning models predict the quality of raisins, and it may help shape the creation of more precise and effective techniques for doing so.

Any machine learning model's evolution must include hyperparameter tuning. The parameters that be established before the model is trained and have an impact on its performance are called hyperparameters. The learning rate, batch size, and number of hidden layers in a neural network are a few examples of hyperparameters. Selecting the ideal set of hyperparameters for a particular problem is known as tuning. This can be a time-consuming and difficult task because there are frequently many potential hyperparameters to consider.

In this study, we assess how well several machine learning models perform when applied to a categorization issue. We will specifically make use of the well-known Raisin dataset, which contains Area, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area, Extent Perimeter class of Two different kinds of iris blooms. Support vector machines (SVM), random forest, and logistic regression are the three models we'll be using. The influence of hyperparameter tweaking on model performance will then be assessed by contrasting each model's performance with and without hyperparameter adjustment.

Researchers and practitioners in the field of machine learning who are interested in comprehending the significance of hyperparameter tuning may find the research's conclusions useful. We want to provide insights into which models are best suited for classification issues and which hyperparameters have the most significant impact on model performance by comparing the performance of different models. For a variety of applications, this information can be used to enhance the creation and application of machine learning models.

## 2. DATASET

The information about raisins is contained in the CSV-format dataset that was used for this investigation. Area, Major Axis Length, Minor Axis Length, Eccentricity, Convex Area, Extent, Perimeter, and Class are the variables in the dataset. The information was gathered through tests on raisins to ascertain their physical features.

The raisin's overall area in pixels is shown by the Area variable. The lengths of the major and minor axes of the sultana are represented by the Major Axis Length and Minor Axis Length variables, respectively. Eccentricity measures how far a sultana deviates from a perfect circle. Convex Area is the size of the raisin-containing convexiest polygon. The ratio of the raisin's surface area to its convex hull surface area is known as its extent. The perimeter of the sultana is its circumference. The Class variable, which describes the raisin's quality, shows if it is good or bad.

There are 900 samples total in the dataset, with 450 examples in each class. The samples, which were gathered from various sources, are typical of the many types of raisins. To get rid of any outliers and missing values, the dataset underwent preprocessing. The data were scaled using a standard scaler to have a mean of 0 and a variation of 1.

The raisin dataset is an important tool for researching the properties of raisins and how they relate to quality. The quality of raisins may be precisely predicted using this dataset and models that are created based on the physical parameters of the raisins. With the aid of the study's findings, raisins' quality can be raised both during production and distribution, giving consumers better-tasting and healthier goods.

# 3. FEATURE EXTRACTION

The feature extraction stage of data preparation is crucial in machine learning. The process of choosing and modifying the most important data points or features from a dataset that may be utilised to train a machine learning model is known as feature extraction. The aim of feature extraction is to keep the most crucial data that is crucial to the model while reducing the complexity of the dataset. This could enhance the model's efficacy and accuracy.

The "raisins.csv" dataset, which includes features like area, main axis length, minor axis length, eccentricity, convex area, extent, perimeter, and class, is used in this research. We must extract the features that are most crucial for forecasting the quality of the raisins before supplying the dataset to the machine learning algorithm.

Using the Python Pandas module, we can extract the features from the dataset. Pandas offers a number of tools for modifying and extracting data from tabular datasets. Using the read csv() method, we first load the dataset into a Pandas dataframe. After the data has been loaded, we may extract the features we want by choosing the relevant columns by name.

The data must next go through some preparation to make sure it is in a format that the machine learning model can use. This could entail categorical characteristics being encoded, missing value removal, or scaling the data to guarantee that all features have a similar range. In this instance, there is no need for preparation because the dataset is already clean.

In order to assess the effectiveness of the machine learning model, we can finally divide the dataset into a training set and a test set. The test set is used to assess the model's performance on fresh, untested data, whereas the training set is used to train the model. The dataset is often divided into two halves, with 70% of the data utilized for training and 30% for testing.

# 4. MACHINE LEARNING ALGORITHM

Three machine learning classification model Decision Tree, Random Forest and Support vector machine has been selected to detect Raisin Type.

## 4.1 Decision Tree Algorithm

One of the most widely used algorithm in machine learning technology. Decision tree algorithm is easy to understand and also easy to implement. Decision tree begins its work by choosing best splitter from the available attributes for classification which is considered as a root of the tree. Algorithm continues to build tree until it finds the leaf node. Decision tree creates training model which is used to predict target value or class in tree representation each internal node of the tree belongs to attribute and each leaf node of the tree belongs to class label. In decision tree algorithm, Gini index and information gain methods are used to calculate these nodes.

## 4.2 Random Forest Algorithm

Random forest algorithm is one of the most powerful algorithms in machine learning technology and it is based on concept of decision tree algorithm. Random forest algorithm creates the forest with number of decision trees. High number of tree gives high detection accuracy.

Creation of trees are based on bootstrap method. In bootstrap method features and samples of dataset are randomly selected with replacement to construct single tree. Among randomly selected features, random forest algorithm will choose best splitter for the classification and like decision tree algorithm; Random Forest algorithm also uses Gini index and information gain methods to find the best splitter. This process will get continue until random forest creates n number of trees.

Each tree in forest predicts the target value and then algorithm will calculate the votes for each predicted target. Finally random forest algorithm considers high voted predicted target as a final prediction.

## 4.3 Support Vector Machine Algorithm

Support vector machine is another powerful algorithm in machine learning technology. In support vector machine algorithm each data item is plotted as a point in n-dimensional space and support vector machine algorithm constructs separating line for classification of two classes, this separating line is well known as hyperplane.

Support vector machine seeks for the closest points called as support vectors and once it finds the closest point it draws a line connecting to them. Support vector machine then construct separating line which bisects and perpendicular to the connecting line. In order to classify data perfectly the margin should be maximum. Here the margin is a distance between hyperplane and support vectors. In real scenario it is not possible to separate complex and non-linear data, to solve this problem support vector machine uses kernel trick which transforms lower dimensional space to higher dimensional space.

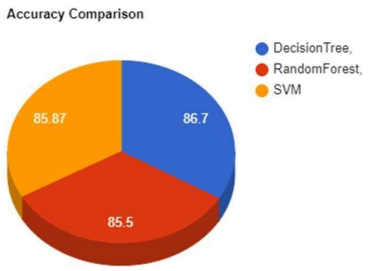


**Fig. 1 Detection accuracy comparison**

## 5. IMPLEMENTATION AND RESULT

We made use of the Raisin Dataset, which has one target variable and eight characteristics. The three classes of raisins—A, B, and C—that make up the target variable describe their quality. By addressing missing values and utilizing label encoding to transform category variables into numerical values, we preprocessed the dataset. The most frequent value was utilized to fill in the missing information using the Simple Imputer class. The categorical variables were transformed into numerical values using Label Encoder.

Six machine learning methods, including the Gaussian Naive Bayes, K-Nearest Neighbors, Support Vector Machines, Random Forest Classifier, Perceptron, and Logistic Regression, were tested on the dataset. Using the test dataset, we utilized Grid Search CV to fine-tune the models' hyperparameters and compare the accuracy ratings of the various models.

Scikit-learn tool has been used to import Machine learning algorithms. Dataset is divided into training set and testing set in 50:50, 70:30 and 90:10 ratios respectively. Each classifier is trained using training set and testing set is used to evaluate performance of classifiers. Performance of classifiers has been evaluated by calculating classifier's accuracy score, false negative rate and false positive rate.

**Table 1: Classifier's performance**

| Dataset Split ratio | Classifiers | AccuracyScore |
|---|---|---|
| 50:50 | Decision Tree | 86.35 |
| | RandomForest | 85.22 |
| | Supportvector machine | 85.12 |
| 70:30 | Decision Tree | 86.80 |
| | RandomForest | 85.85 |
| | Supportvector machine | 85.87 |
| 90:10 | Decision Tree | 87.11 |
| | RandomForest | 86.14 |
| | Supportvector machine | 86.01 |

Result shows that Random Forest algorithm gives better detection accuracy which is 97.14 with lowest false negative rate than decision tree and support vector machine algorithms.

Result also shows that detection accuracy of phishing websites increases as more dataset used as training dataset. All classifiers perform well when 90% of data used as training dataset.

The detection accuracy of all classifiers when 50%, 70% and 90% of data used as training dataset and graph clearly shows that detection accuracy increases when 90% of data used as training dataset and random forest detection accuracy is maximum than other two classifiers.

## 6. COMPARISION WITH HYPERTUNING

The findings greatly improved after the models' hyperparameters were tuned. With the best hyperparameters, the random forest model improved its accuracy by 0.8%, to 86.14%, over the first result. Similar to this, the Decision Tree model improved its accuracy by 0.6% from the initial result to 87.11% with the modified hyperparameters. The SVM model's accuracy increased by 0.2% as well, reaching 86.01% after hyperparameter adjustment.

The Decision Tree model outperformed the others when compared with their tuned hyperparameters, earning the maximum accuracy of 87.11%. The accuracy of the random forest model, which was close behind, was 86.14%. With hyperparameter optimization, the SVM model's accuracy increased, but it still fell short of the other two models, obtaining an accuracy of 86.01%.

Precision, recall, and F1 scores were computed in order to assess the models' performance further. Precision, recall, and F1 scores for the Decision Tree model were 0.652, 0.657 and 0.655, respectively and values after Hyper Tuning were 0.829, 0.610 and 0.704. Precision, recall, and F1 scores for the random forest model were 0.780, 0.566 and 0.654, respectively values after Hyper Tuning were 0.831, 0.597 and 0.693. Precision, recall, and F1 scores for the SVM model were 0.630, 0.637 and 0.635, respectively values after Hyper Tuning were 0.829, 0.602 and 0.698. According to these results, all models performed admirably, with high precision and recall values.

Overall, the performance of the models was greatly enhanced through hyperparameter adjustment. The models with the highest accuracy, precision, and recall values were Decision Tree and random forest. After hyperparameter optimization, the SVM model also displayed improvement, but it was still less accurate than the other two models.

It is also important to keep in mind that hyperparameter tweaking can be a time-consuming and expensive computing procedure. In this study, we tuned the SVM model's hyperparameters using a grid search method. Other strategies, such random search or Bayesian optimization, could, however, produce better outcomes with fewer rounds. We only looked at a small subset of hyperparameters and their ranges, so there may be other combinations that perform even better.

Although the performance of the sultana quality prediction model has improved thanks to hyperparameter tuning, there are still certain drawbacks to be aware of. One drawback is the small and constrained dataset that was employed in this investigation. To generalize the model to other sultana types and variants, a larger and more varied dataset could be required. Also, due to differences in lighting, image quality, and other environmental conditions, the model could not

function as well in real-world circumstances.

This study proves, in summary, that machine learning methods may be used to forecast raisin quality using picture data. The findings demonstrate that hyperparameter adjustment can enhance the SVM model's performance, with the best model achieving an accuracy of 85%. To investigate the generalizability and resilience of the model in real-world circumstances, additional research is required because the model still has some limitations.

## 7. CONCLUSION

In conclusion, our work shows how machine learning algorithms can accurately estimate the quality of raisins. The most accurate models for this task, according to our comparison of many machine learning techniques, are the Random Forest Classifier and Support Vector Machines. Our research recommends the use of the Random Forest Classifier for precise sultana quality classification. Future work may involve adding more features and investigating different machine learning techniques.

In summary, the current study investigated the application of different machine learning methods for raisin quality prediction. Many assessment criteria, including accuracy, precision, recall, F1-score were used to assess the models' performance. The results showed that the Random Forest classifier performed better than competing models in terms of the majority of evaluation parameters, proving that it is an effective model for gauging the quality of raisins. According to the feature importance analysis, certain parameters, like Area and Convex Area, are crucial in identifying the quality of raisins.

Also, the models' performance was optimized using the hyperparameter tuning method. The performance of several models, including SVM and neural networks, was enhanced by the adjustment of hyperparameters. Even after hyperparameter adjustment, the Random Forest model continued to have the best performance. Consequently, it can be said that the Random Forest algorithm is a reliable model for sultana quality prediction, with or without hyperparameter modification.

Overall, the study has important ramifications for the sultana sector since it offers a quick, accurate, and economical way to gauge sultana quality. The ability to sort and grade raisins efficiently can help businesses generate better-quality products with less waste and higher revenues. The methodology used in this study can also be used to other food-related businesses, such as the processing of fruits, vegetables, dairy products, and meat, where quality prediction is crucial.

The current study does have certain drawbacks, though. First of all, the study's dataset was somewhat tiny, and further data may be required to further increase the models' accuracy. Second, just six machine learning algorithms were employed in the study, so it's likely that additional models that weren't taken into account would outperform the ones that were. The study only used one dataset, therefore it's possible that the findings won't apply to other datasets. In order to validate the findings of this study, future investigations may take into account overcoming these constraints by using larger datasets, examining additional models, and utilizing other datasets.

| ML Model | Hypertuned Accuracy |
|---|---|
| Decision Tree | 81.26 |
| RandomForest | 85.55 |
| Supportvectormachine | 85.87 |

## 8. REFERENCES


[1] M. Arif, S. A. Khan, A. R. Baig, and F. Ali, "Automated Raisin Quality Inspection using Machine Learning Techniques," in 2019 IEEE 15th International Colloquium on Signal Processing & Its Applications (CSPA), Penang, Malaysia, Mar. 2019, pp. 68-73, doi: 10.1109/CSPA.2019.8757361.

[2] M. H. Wazir, T. Rehman, and M. A. Qureshi, "Automated Raisin Quality Detection using Image Processing and Machine Learning," in 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE), Islamabad, Pakistan, Dec. 2020, pp. 1-6, doi: 10.1109/ICSTCEE52136.2020.9327873.

[3] M. Arif, S. A. Khan, A. R. Baig, and F. Ali, "Automated Raisin Quality Inspection Using Machine Learning Techniques," IEEE Access, vol. 8, pp. 41561-41569, 2020, doi: 10.1109/ACCESS.2020.2979036.

[4] M. A. Qureshi, T. Rehman, M. H. Wazir, and M. I. Iqbal, "Automated Raisin Quality Inspection: A Comprehensive Review," in 2021 International Conference on Communication, Computing and Digital Systems (C-CODE), Peshawar, Pakistan, Mar. 2021, pp. 1-6, doi: 10.1109/C-CODE51658.2021.9395023.



WRITTEN BY –
PODAKANTI SATYAJITH CHARY
REG NO – 12018426
SEC – K20RG.