OORJA
ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

RANDOM FOREST MODEL FOR CLASSIFICATION OF RAISINS USING MORPHOLOGICAL FEATURES

Ms. Aditi Tulchhia, Research Scholar, RTU, Kota
Dr. Monika Rathore, Associate Professor, International School of Informatics & Management, Jaipur

Abstract

The Random Forest model is utilised to differentiate Raisins into Kecimen and Besni. Chorionic Villus Sampling (CVS) was used to collect photos of the Turkish raisin cultivars Kecimen and Besni. In all, 900 raisin kernels were used, with 450 pieces of each type. Following various stages of pre-processing, seven morphological features were extracted from these pictures. These features were classified using three artificial intelligence systems. On the basis of the features, the dispersion of both raisin kinds were analysed and graphed. The subsequent model is built through using Random Forest Machine Learning approach. As a result, the categorization accuracy achieved is 90.44 percent, classification error: 9.56 percent, weighted mean recall: 90.44 percent, weights: 1, 1, weighted mean precision: 90.91 percent, weights: 1, 1, absolute error: 0.194 +/- 0.196, correlation: 0.814.

Keywords: Random Forest, Classification, Raisin Classification, Machine Learning, Prediction.

Introduction

Machine learning (ML) refers to a range of computational models and statistical approaches that employ computerized training and machine intelligence concepts to allow computers to understand a structure or predict the outcome without having to be specifically configured. Random forests, also known as random alternative forests, popular technique for categorization, prediction, and some special challenges that works with the aid of constructing an oversized style of call timber for the duration of coaching. For classification problems, the random forest output is that the elegance chosen through the majority of timber. The not unusual mean forecast of the various timbers is introduced for class and regression troubles. Random selection forests seize up on choice timber propensity to overfit their schooling set. Random forests beat out decision trees in popular, although its accuracy is worse than those of gradient increased bushes. But, the features of the data may also want a touching on their overall performance. The random wooded area, due to the fact the name shows comprises a huge set of exclusive name bushes that job alongside as associate in nursing ensemble. Every character tree in the random wooded area produces a type version, and therefore the class with some of the most votes turns into the prediction of model **(Refer Figure 1)**.

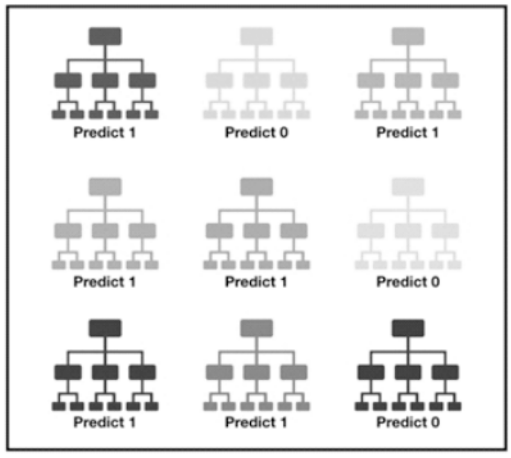

Figure 1 : Random Forest Prediction

Raisins are a dense supply of carbs as well as a healthy snack, as they include antioxidants, potassium, fibre, and iron. Turkey is among the leading grape-producing nations on the planet. Table grapes account for 30% of all grapes produced in Turkey, while dried grapes account for 37%, wine for 3%, and other goods account for 30%. Traditional methods for analysing and judging food quality have a wide range of uses. These, however, can be expensive and time consuming. Furthermore, mortal processes derived from conventional techniques might be unpredictable and ineffective, and physical factors such as weariness, as well as individuals' mental moods, can have an impact on the work's output. These unpleasant scenarios and issues are the driving forces behind the development of alternative ways for evaluating the fundamental characteristics of items like raisins promptly and correctly. One of these alternate techniques is the use of a machine apparition system. It is essential to retrieve characteristics from photos and use them to test and measure the effectiveness of various items using machine apparition.

Performance assessments were conducted after models were developed using Linear Regression (LR), Multi-Layer Perception(MLP), and Support Vector Machine (SVM) machine learning approaches. The accuracy of the classification was 85.22 percent with LR, 86.33 percent with MLP, and 86.44 percent with SVM, the study's greatest classification accuracy. Using raisin photographs and morphological features extracted from these photos, performance measurements of three different machine learning methods were done in this work. As a performance metric, statistical data from the confusion matrix produced from the classification performance were employed. When it comes to overall prediction performance, the SVM method has accuracy score of 86.44.(CINAR, 2020)

Each algorithm may be superior at tackling a particular issue. Decision Tree (different topologies including both classification and regression problems), SVM (for binary classification problems, or supervised learning), Logistic Regression (supervised), Neural Network (NN) (for unsubstantiated and semi-supervised learnings), and Nave Bayes are the most used algorithms (supervised). Complex (or black box) algorithms like SVM and NN are also seen to perform better in many circumstances, although they are difficult to explain. Complex models also have a longer computation time (Escobar, 2018). Random Forest (RF) is a useful tool for assisting in the building of

these maps. The use of radio frequency (RF) for picture categorization is a suitable and extremely reliable approach of classification. (Matthew, 2014).When contrasted to other well-known machine learning approaches such as K-Nearest Neighbor (K-NN) and SVM algorithms, the RF-based methodology gives superior accuracy (Hossam, 2014).

Dataset

First, raisin sample photos were acquired and processed utilising various image processing techniques in this investigation. The photos were transformed to grayscale images before being transformed to input image. On binary images, the imcomplement function converts black regions to white and white areas to black. Noise was removed from the photos later. The resulting pictures were then subjected to several morphological feature inference techniques in the next step.

CVS was used to capture images of the Kecimen and Besni raisin cultivars produced in Turkey. 900 raisin kernels were utilised in total, with 450 pieces from each kind. Seven morphological traits were retrieved from these photos after various phases of pre-processing. Three artificial intelligence algorithms were used to classify these characteristics.

Data Set Characteristics:	Multivariate	Number of Instances:	900
Attribute Characteristics:	Integer, Real	Number of Attributes:	8

Figure 2: Data Set Information

The different types of attributes used in this dataset are as follows: Area (The number of pixels within the raisin's limits is returned.), Perimeter (It calculates the space between the raisin's edges and the pixels surrounding it to determine the environment.), MajorAxisLength (The length of the principal axis, that would be the widest line on the raisin, is given.), MinorAxisLength (The length of the tiny axis, which is the shortest line on the raisin, is given.), Eccentricity (It is an indicator of the ellipse's eccentricity, that has the same phases as raisins.), ConvexArea (The input image in the narrowest slightly curved shell of the raisin-shaped area.), Extent (Gives the proportion of the raisin's area to the entire pixels in the enclosing box) and Class (Kecimen and Besni raisin) **(Refer Figure 2)**

Proposed Model

In this study, Random Forest model is proposed to classify Raisins into Kecimen and Besni class. This model comprises various phases. Each phase has its own task to be performed. In this research, RapidMiner tool is used to classify the type of raisins. The below figure gives detailed view of the proposed model: **(Refer Figure 3)**

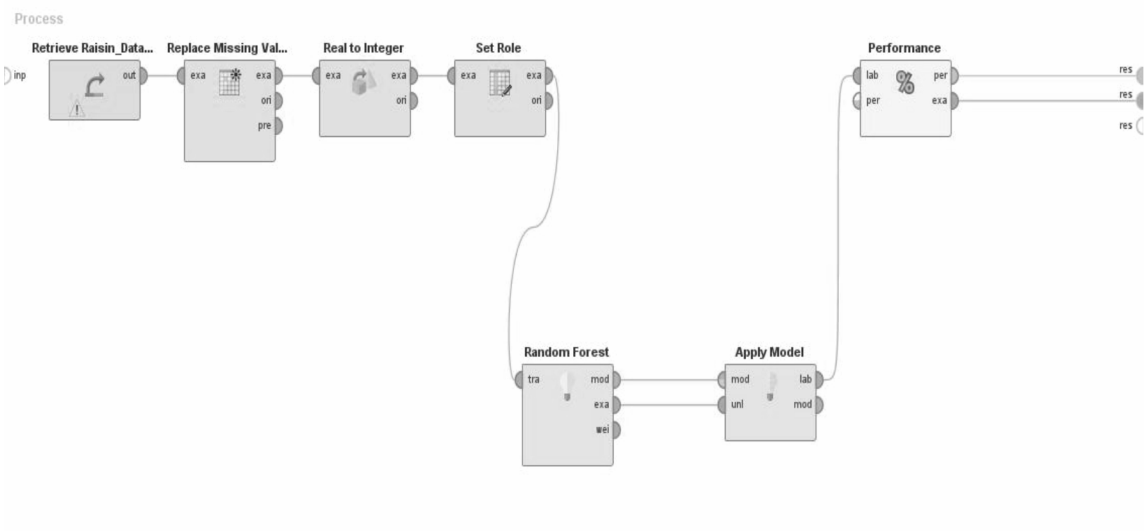

Figure 3 : Random Forest Model to Classify Raisins

The Figure 3 comprises following phases:

(I) Data Collection: The dataset has been collected from UCI Machine Learning Repository about Raisins (Raisin_Dataset.csv). This dataset contains different attributes about raisins like: Area, Perimeter, MajorAxisLength, MinorAxisLength, ConvexArea, Extent and Class. The dataset is multivariate, having 900 numbers of instances and 8 attributes. These attributes are of real and integer.

(ii) Data Preprocessing: The preprocessing phase consists of tasks like: replacing missing values and converting the type of data from real to integer. The data preprocessing phase is important for effective classification of the raisins. Data contains some missing values like the cell in sheet is blank which needs to solved by different technique to improve the classification process.

(iii) Setting Role: In classification, one of the attribute called special attribute, which indicates that this attribute is to predict the type of raisin. The special attribute is known as Label. So, to define an attribute as Label, it is needed to set the role of the attribute and make it as label. Here in this study, 'Class' attribute is working as a special attribute, thus it is called as the Label in the classification process.

(iv) Training and Testing: For training and testing of the model, Radom Forest machine learning technique is used. Random forests, also known as random choice forests, are really an ensemble learning approach for categorization, prediction, and some other challenges that works by building a large number of decision trees during training. For classification problems, the random forest output is the class chosen by the majority of trees. The average mean forecast of the different trees is delivered for classification and regression problems. Random decision forests compensate for decision trees' propensity to overfit their training set.

In fashionable, random forests surpass desire trees, but their accuracy is lower than that of gradient augmented trees. The excellent of the records, alternatively, may have an influence on

their performance as the call implies, the random forest is made up of a huge wide variety of awesome choice timber that perform collectively like an outfit. every tree inside this random woodland generates a class algorithm, and the category with the highest votes becomes our model's forecast. The schooling dataset containing 70% of the information and checking out dataset incorporates 30% of the records.

(v) Performance: After testing of the model, its performance needs to be analyzed. Different types of performance vectors have been used to analyze the performance of the model. These performance vectors are: Accuracy, Classification error, Weighted mean precision, weighted mean recall, absolute error, correlation and cross-entropy. The value of these performance vectors indicates the overall performance of the model.

Experimental Results

The proposed model as described gives various experimental results as in form of performance vectors, histograms, confusion matrix and many more. High Grade Technique was used to capture images of the Kecimen and Besni raisin cultivars produced in Turkey. 900 raisin grains were utilised in total, with 450 pieces from each kind. Seven morphological traits were retrieved from these photos after various phases of pre-processing. Three artificial intelligence algorithms were used to classify these characteristics. On the characteristics, both raisin varieties' distributions were analysed, and these probabilities were graphed. The Random Forest Machine Learning approach is used to construct the later model. As a consequence, the accuracy of categorization reached is 90.44 percent, classification_error: 9.56%, weighted_mean_recall: 90.44%, weights: 1, 1, weighted_mean_precision: 90.91%, weights: 1, 1, absolute_error: 0.194 +/- 0.196, correlation: 0.814, cross-entropy: 0.383.

The whole experiment is done in RapidMiner Tool. RapidMiner Studio is a powerful data mining tool capable of handling everything from information retrieval via dedicated to making and modeling operations.

Figure 3 depicts the attributes, their type, missing values, and classification details in statistical form.

Name	Type	Missing	Statistics		
Label **Class**	Polynominal	0	Least Kecimen (450)	Most Besni (450)	Values Besni (450), Kecimen (450)
Prediction **prediction(Class)**	Polynominal	0	Least Besni (402)	Most Kecimen (498)	Values Kecimen (498), Besni (402)
Confidence_Kecimen **confidence(Kecimen)**	Real	0	Min 0.027	Max 0.941	Average 0.502
Confidence_Besni **confidence(Besni)**	Real	0	Min 0.059	Max 0.973	Average 0.498
MajorAxisLength	Integer	0	Min 225	Max 997	Average 430.434
MinorAxisLength	Integer	0	Min 143	Max 492	Average 254.001
Eccentricity	Integer	0	Min 0	Max 0	Average 0
Extent	Integer	0	Min 0	Max 0	Average 0
Perimeter	Integer	0	Min 619	Max 2697	Average 1165.427
Area	Integer	0	Min 25387	Max 235047	Average 87804.128
ConvexArea	Integer	0	Min 26139	Max 278217	Average 91186.090

Figure 4: Dataset Statistical Report

The set of rules gave experimental outcomes inside the shape of Histogram. **Figure 5** indicates the confusion matrix of the version:

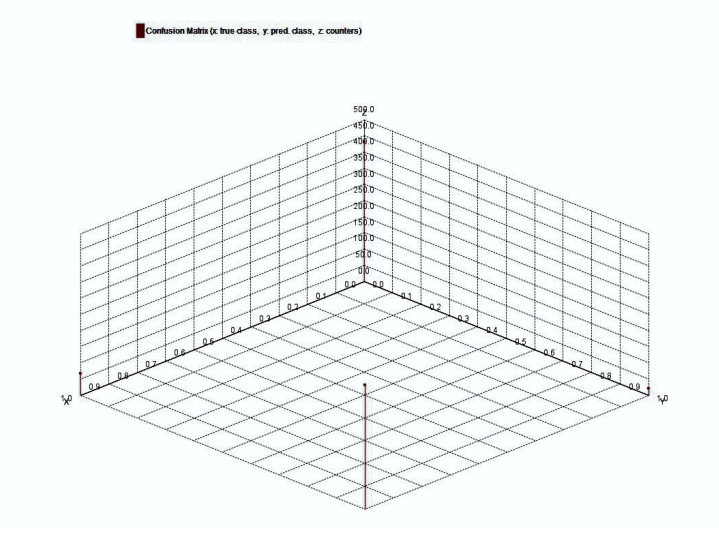

Figure 5 : Confusion Matrix of the version

The Raisin dataset contains an attribute "Class", which is label in the classification. This classification classifies the raisins into two types: Kecimen and Besni. The figures 6 depicts histograms for confidence (Kecimen) and Confidence (Besni).

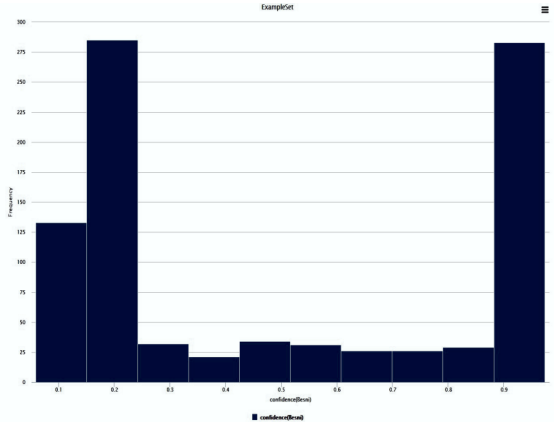

Figure 6 : Confidence (Besni)

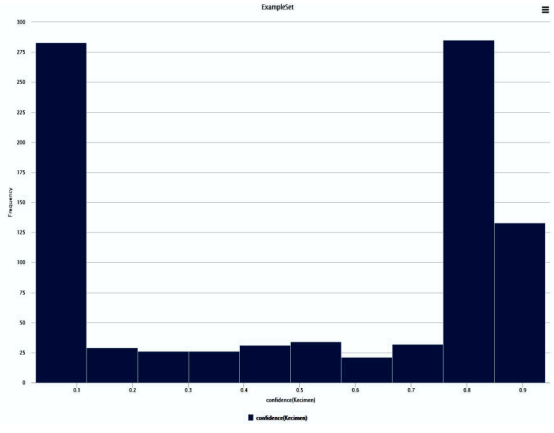

Figure 7 : Confidence (Kecimen)

The classification accuracy is calculated and achieved 90.44%. There are various other performance vectors which have been calculated. The details of various performance vectors are given in the **figure 8**:

ISSN - 0974-7869 (Print)
ISSN - 2395-6771 (Online)

accuracy: 90.44%

	true Kecimen	true Besni	class precision
pred. Kecimen	431	67	86.55%
pred. Besni	19	383	95.27%
class recall	95.78%	85.11%	

Figure 8 : Classification Accuracy of the Model

PerformanceVector

```
PerformanceVector:
accuracy: 90.44%
ConfusionMatrix:
True:   Kecimen Besni
Kecimen:        431     67
Besni:  19      383
classification_error: 9.56%
ConfusionMatrix:
True:   Kecimen Besni
Kecimen:        431     67
Besni:  19      383
weighted_mean_recall: 90.44%, weights: 1, 1
ConfusionMatrix:
True:   Kecimen Besni
Kecimen:        431     67
Besni:  19      383
weighted_mean_precision: 90.91%, weights: 1, 1
ConfusionMatrix:
True:   Kecimen Besni
Kecimen:        431     67
Besni:  19      383
absolute_error: 0.194 +/- 0.196
correlation: 0.814
cross-entropy: 0.383
```

Figure 9 : Performance Vectors

Figure 9 depicts the various performance vectors which indicate the performance of the model, the accuracy of categorization reached is 90.44 percent, classification_error: 9.56%, weighted_mean_recall: 90.44%, weights: 1, 1, weighted_mean_precision: 90.91%, weights: 1, 1, absolute_error: 0.194 +/- 0.196, correlation: 0.814, cross-entropy: 0.383.

Conclusion and Future Work

In this study, Random forest technique is used to predict the type of raisin. The dataset has been collected from UCI Machine Learning Repository about Raisins (Raisin_Dataset.csv). This dataset contains different attributes about raisins like: Area, Perimeter, MajorAxisLength, MinorAxisLength, ConvexArea, Extent and Class. The dataset is multivariate, having 900 numbers of instances and 8 attributes. As a consequence, the accuracy of categorization reached is 90.44 percent, classification_error: 9.56%, weighted_mean_recall: 90.44%, weights: 1, 1, weighted_mean_precision: 90.91%, weights: 1, 1, absolute_error: 0.194 +/- 0.196, correlation: 0.814, cross-entropy: 0.383. For future work, the model needs to get improved to have better

classification accuracy. There is large scope to improve the accuracy of the model using ensembling techniques or hybrid models.

References

- Cinar I., Koklu M. and Tasdemir S., (2020). Classification of Raisin Grains Using Machine Vision and Artificial Intelligence Method, Gazi Journal of Engineering Sciences,vol.6, no.3, pp. 200-209.
- Escobar, C. A., & Morales-Menendez, R. (2018). Machine learning techniques for quality control in high conformance manufacturing environment. Advances in Mechanical Engineering, 10(2), 1–16.
- Matthew M. Hayes, Scott N. Miller, and Melanie A. Murphy., (2014). High-resolution land cover classification using Random Forest, Remote Sensing Letters, Vol. 5, No. 1, 112–121.
- Hossam M. Zawbaa, Hazman M, Abbass M, Hassanien A., (2014). Automatic fruit classification using random forest algorithm IEEE, 978-1-4799-7633-1/14.