

**Project Report**

**on**

**Raisin Classification by Using Machine Learning Classifiers**



**By**

**Aditya V. Shinde**

**MIS- 112015003**

**Under the Guidance**

**Dr. Nagendra Kushwaha**

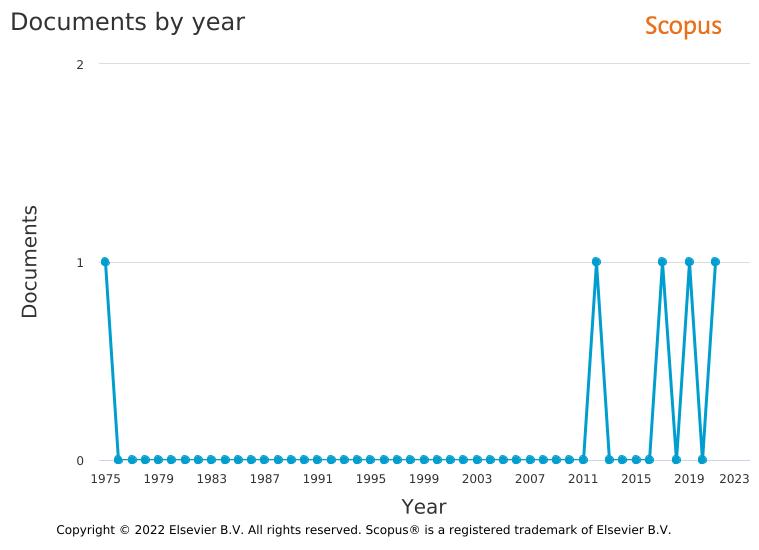
**Contents**

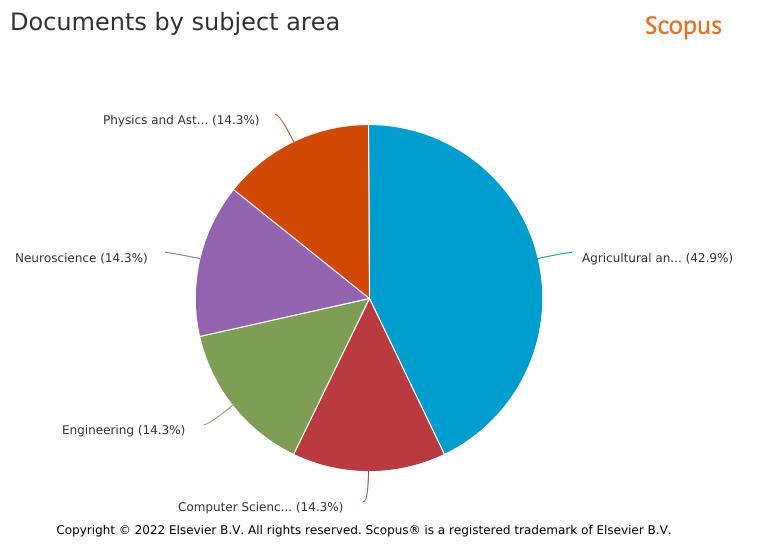
|  |  |
| --- | --- |
| **Sr. No** | **Chapter** |
| 1 | Introduction |
| 2 | Problem Statement |
| 3 | Methodology |
| 4 | Dataset |
| 5 | WEKA |
| 6 | Experimental results |
| 7 | Conclusion |
| 8 | Resferences |

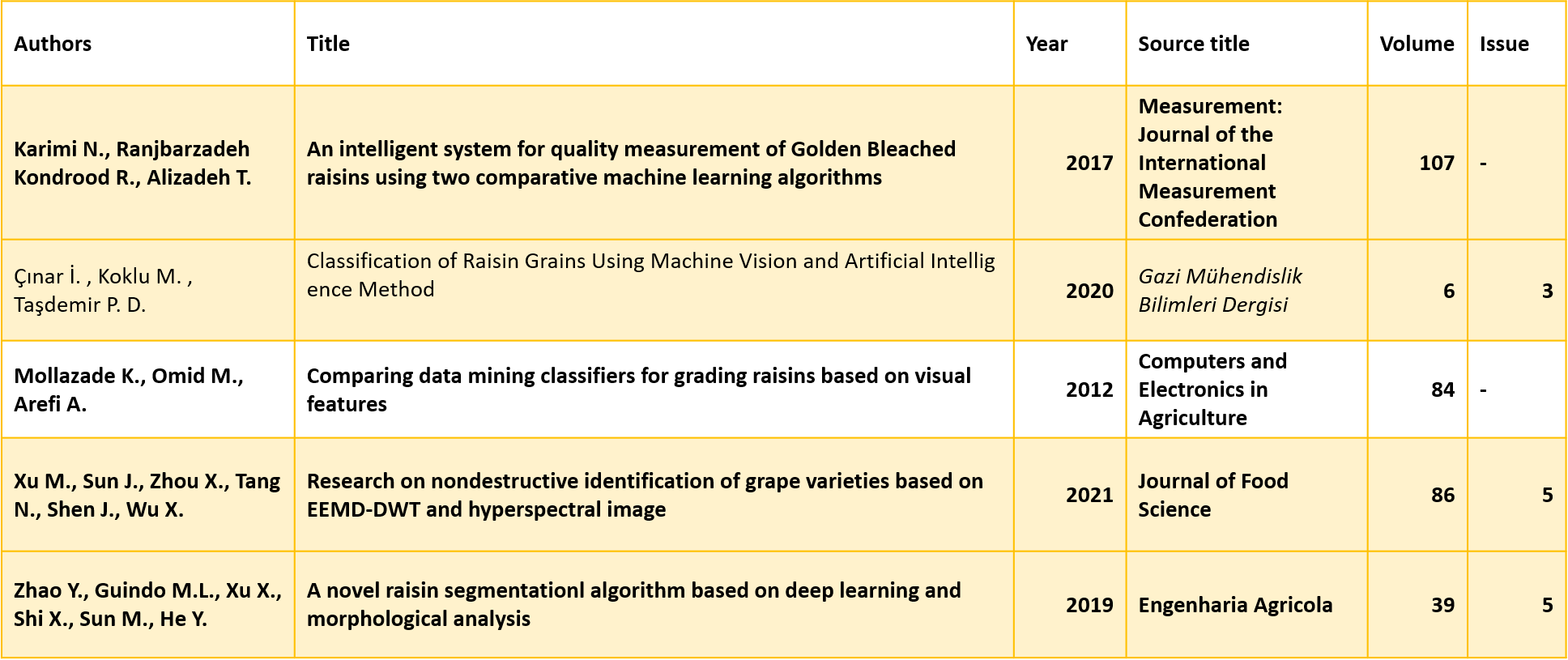
**Chapter 1 - Introduction**

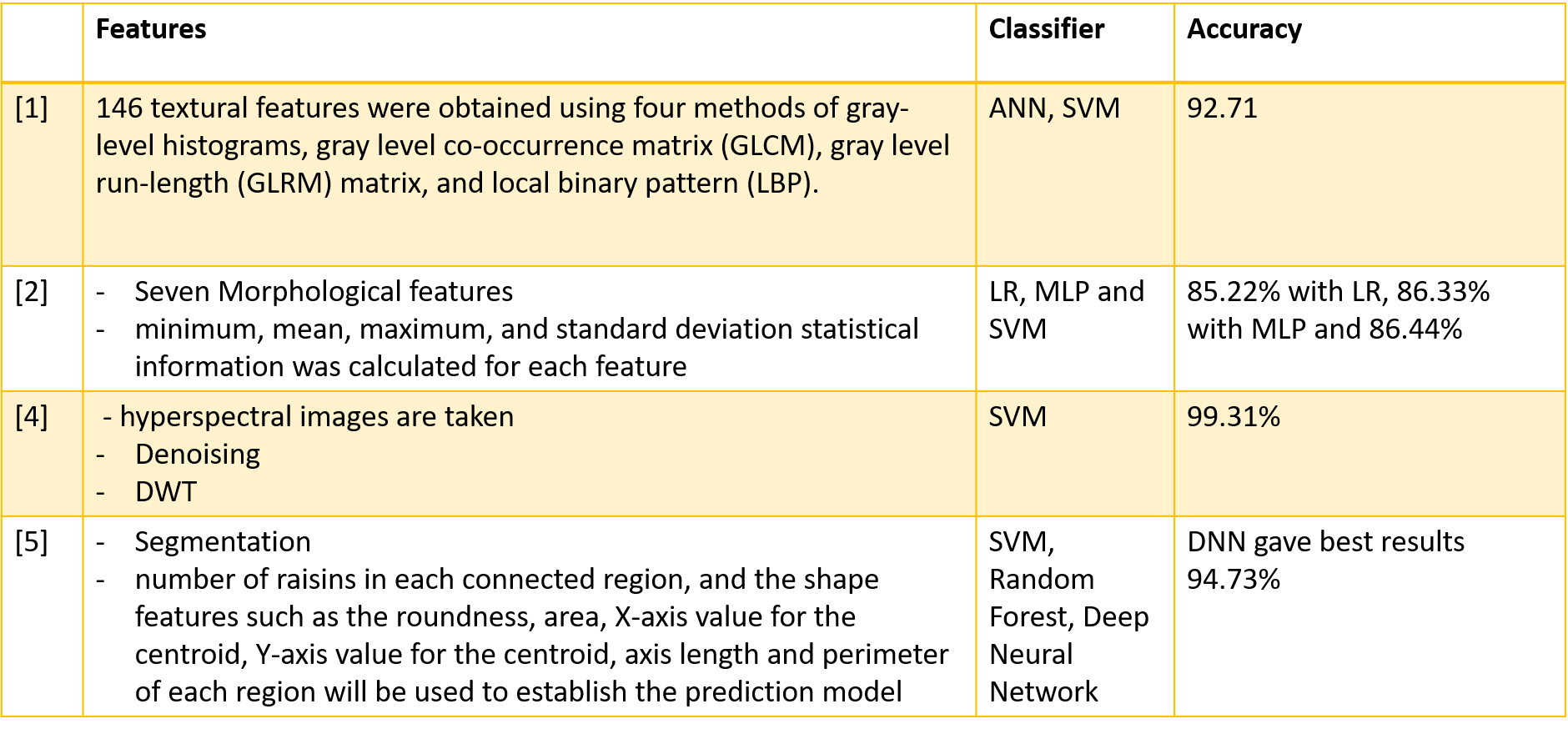
* ***Raisins* are a concentrated source of carbohydrates and a nutritious snack, containing antioxidants, potassium, fiber and iron**
* ***Manual Methods* to assess the quality of raising are time-consuming and expensive and non-efficient**
* **Machine *learning* plays important role in the quick and precise evaluation of raisin quality.**
* ***This project* focuses on machine learning for raisin classification**

**Literature Review: Search Query ‘Raisin+Machine Learning’**









**Chapter 2 – Problem Statement**

**Problem Statement**- Raisin Classification by Using Machine Learning Classifiers

**Domain Area-** Machine Learning

**Objective-** To improve accuracy of the Raisins classification by using different machine learning algorithms

**Chapter 3 – Methodology**

1. **Technology used: Machine Learning algorithms**
2. Decision Tree Classification Algorithm

* Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
* In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
* The decisions or the test are performed on the basis of features of the given dataset.
* ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
* It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
* In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
* A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
* Below diagram explains the general structure of a decision tree:



## **Why use Decision Trees?**

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

* Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand
* The logic behind the decision tree can be easily understood because it shows a tree-like structure.

## **Decision Tree Terminologies**

 **Root Node:** Root node is from where the decision tree starts

. It represents the entire dataset, which further gets divided into two or more homogeneous sets.

 **Leaf Node:** Leaf nodes are the final output node

, and the tree cannot be segregated further after getting a leaf node.

 **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.

 **Branch/Sub Tree:** A tree formed by splitting the tree.

 **Pruning:** Pruning is the process of removing the unwanted branches from the tree.

 **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

**How does the Decision Tree algorithm Work?**

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.

For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

* **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
* **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
* **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
* **Step-4:** Generate the decision tree node, which contains the best attribute.
* **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

# 2.Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of **ensemble learning,** which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.

As the name suggests, ***"Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset."***

 Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.

**The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.**

The below diagram explains the working of the Random Forest algorithm:

1.3M

CNN's New Streaming Service to Shut Down by End of April



## **Assumptions for Random Forest**

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

* There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
* The predictions from each tree must have very low correlations

## **Why use Random Forest?**

Below are some points that explain why we should use the Random Forest algorithm:

* It takes less training time as compared to other algorithms.
* It predicts output with high accuracy, even for the large dataset it runs efficiently.
* It can also maintain accuracy when a large proportion of data is missing.

## **How does Random Forest algorithm work?**

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

The Working process can be explained in the below steps and diagram:

**Step-1:** Select random K data points from the training set.

**Step-2:** Build the decision trees associated with the selected data points (Subsets).

**Step-3:** Choose the number N for decision trees that you want to build.

**Step-4:** Repeat Step 1 & 2.

**Step-5:** For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

**2.Classification via regression**

In weka this algorithm is implemented.

Class for doing classification using regression methods. Class is binarized and one regression model is built for each class value. For more information, see, for example

## **What Does Classification Mean?**

Classification is the process of identifying and and grouping objects or ideas into predetermined categories. In [data management](https://www.techopedia.com/definition/5422/data-management), classification enables the separation and sorting of data according to set requirements for various business or personal objectives.

In [machine learning](https://www.techopedia.com/definition/8181/machine-learning-ml) (ML), classification is used in [predictive modeling](https://www.techopedia.com/definition/14004/predictive-modeling) to assign input data with a class label. For example, an email security program tasked with identifying spam might use natural language processing ([NLP](https://www.techopedia.com/definition/653/natural-language-processing-nlp)) to classify emails as being "spam" or "not spam."

Data classification is a diverse process that involves various methods and criteria for sorting data within a database or repository. This is generally done through a database or business intelligence software that provides the ability to scan, identify and separate data. Some examples and applications of data classification include:

* Separating customer data based on gender
* Identifying and keeping frequently used data in disk/memory cache
* Data sorting based on content/file type, size and time of data
* Sorting for security reasons by classifying data into restricted, public or private data types

# What is Regression?

Regression analysis is a statistical method that helps us to analyze and understand the relationship between two or more variables of interest. The process that is adapted to perform regression analysis helps to understand which factors are important, which factors can be ignored, and how they are influencing each other.

For the regression analysis is be a successful method, we understand the following terms:

* **Dependent Variable:**This is the variable that we are trying to understand or forecast.
* **Independent Variable:** These are factors that influence the analysis or target variable and provide us with information regarding the relationship of the variables with the target variable.

1. **Proposed Solution**





**Figure 2. The computer vision system used to acquisition images**



Figure 2 -Sample image of raisin varieties used in the study ((a) Besni, (b) Kecimen)

Image Pre-processing



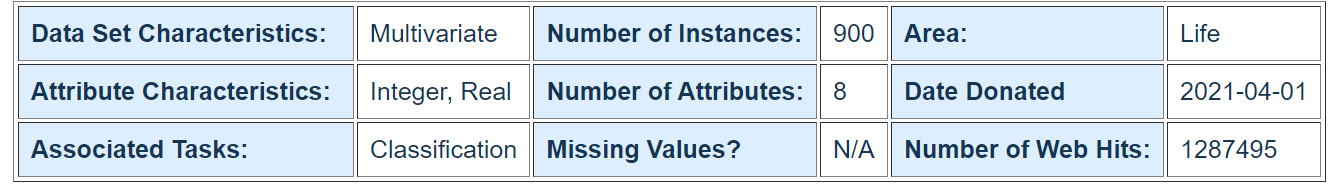
Figure 4. The preprocessing steps performed on images ((a) Real image, (b) Grayscale image, (c) Binary image, (d) Imcomplement image)

Feature Extraction

1. Area
2. Perimeter
3. MajorAxisLength
4. MinorAxisLength
5. Eccentricity
6. ConvexArea
7. Extent

**Chapter 4 – Dataset**

**Abstract**: Images of the Kecimen and Besni raisin varieties were obtained with CVS. A total of 900 raisins were used, including 450 from both varieties, and 7 morphological features were extracted.

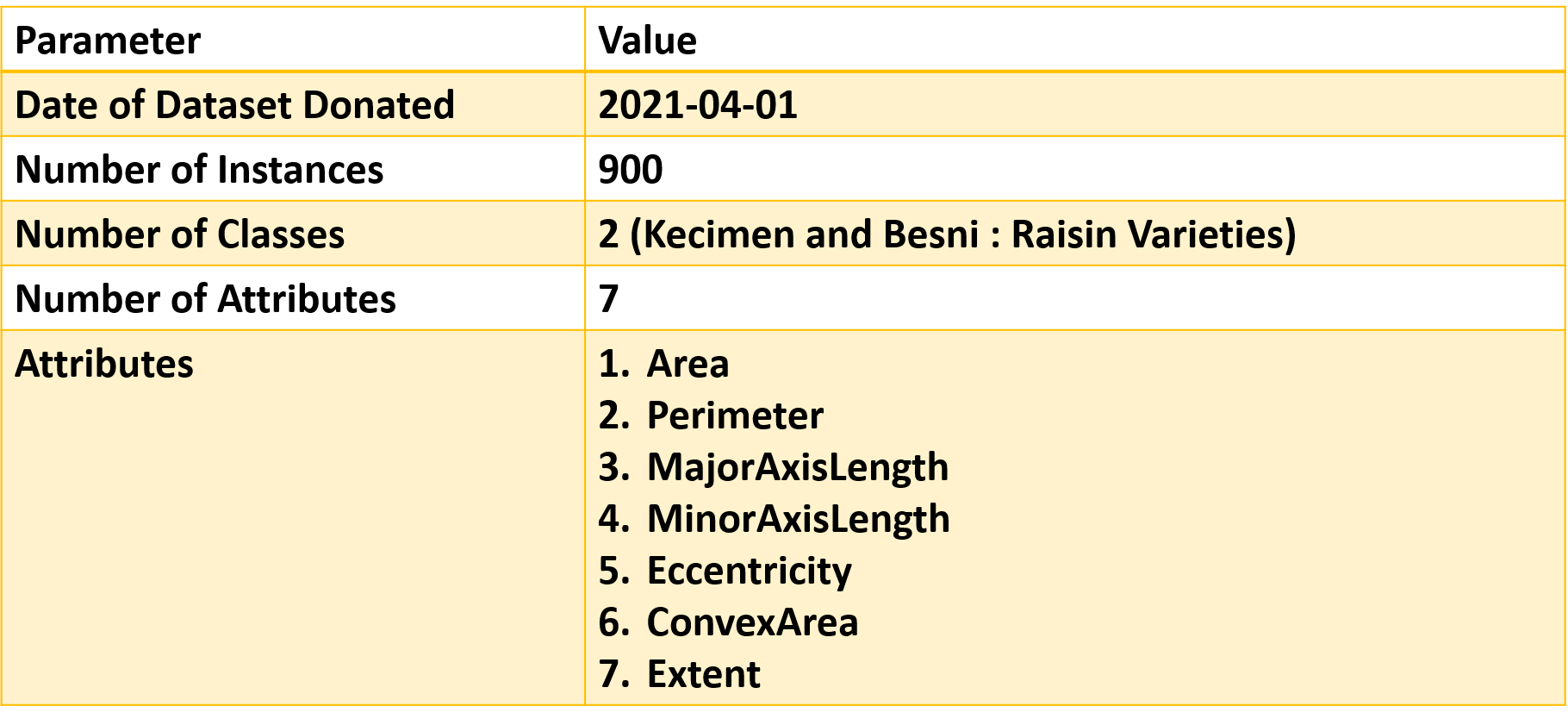


**Data Set Information:**

Images of Kecimen and Besni raisin varieties grown in Turkey were obtained with CVS. A total of 900 raisin grains were used, including 450 pieces from both varieties. These images were subjected to various stages of pre-processing and 7 morphological features were extracted. These features have been classified using three different artificial intelligence techniques.

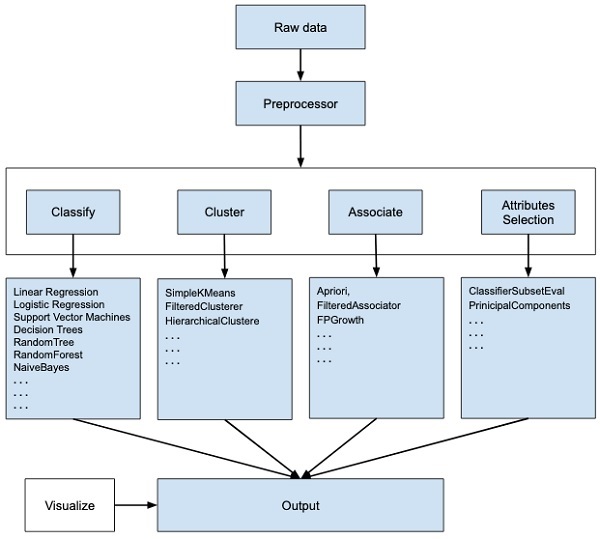
**Attribute Information:**

1.) Area: Gives the number of pixels within the boundaries of the raisin.  
2.) Perimeter: It measures the environment by calculating the distance between the boundaries of the raisin and the pixels around it.  
3.) MajorAxisLength: Gives the length of the main axis, which is the longest line that can be drawn on the raisin.  
4.) MinorAxisLength: Gives the length of the small axis, which is the shortest line that can be drawn on the raisin.  
5.) Eccentricity: It gives a measure of the eccentricity of the ellipse, which has the same moments as raisins.  
6.) ConvexArea: Gives the number of pixels of the smallest convex shell of the region formed by the raisin.  
7.) Extent: Gives the ratio of the region formed by the raisin to the total pixels in the bounding box.  
8.) Class: Kecimen and Besni raisin.



**Chapter 5 – WEKA (Waikato Environment for Knowledge Analysis) - Machine Learning Tool**

WEKA - an open source software provides tools for data preprocessing, implementation of several Machine Learning algorithms, and visualization tools so that you can develop machine learning techniques and apply them to real-world data mining problems. What WEKA offers is summarized in the following diagram −



First, you will start with the raw data collected from the field. This data may contain several null values and irrelevant fields. You use the data preprocessing tools provided in WEKA to cleanse the data.

Then, you would save the preprocessed data in your local storage for applying ML algorithms.

Next, depending on the kind of ML model that you are trying to develop you would select one of the options such as **Classify, Cluster**, or **Associate**. The **Attributes Selection** allows the automatic selection of features to create a reduced dataset.

Note that under each category, WEKA provides the implementation of several algorithms. You would select an algorithm of your choice, set the desired parameters and run it on the dataset.

Then, WEKA would give you the statistical output of the model processing. It provides you a visualization tool to inspect the data.

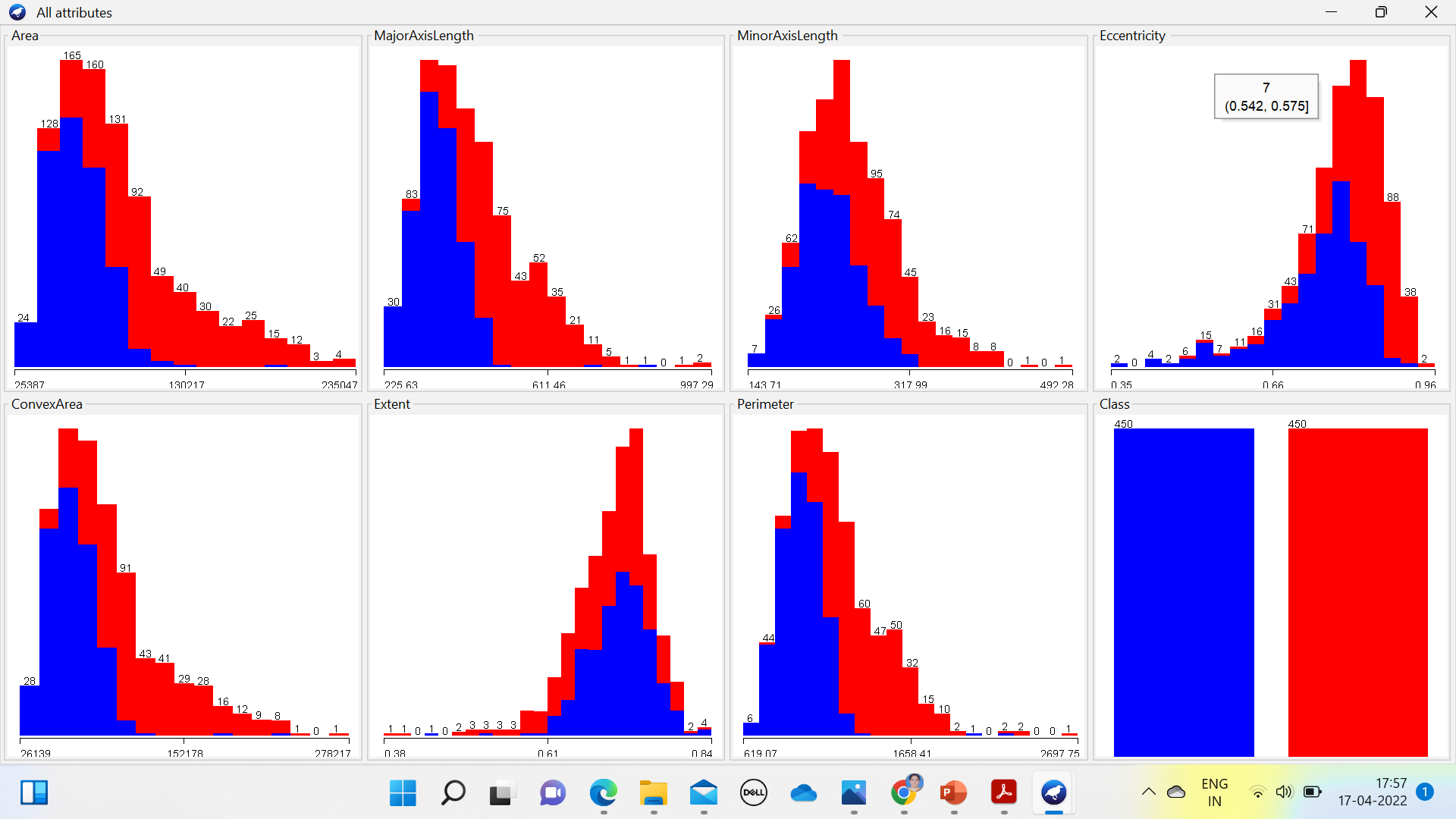
The various models can be applied on the same dataset. You can then compare the outputs of different models and select the best that meets your purpose.

Thus, the use of WEKA results in a quicker development of machine learning models on the whole.

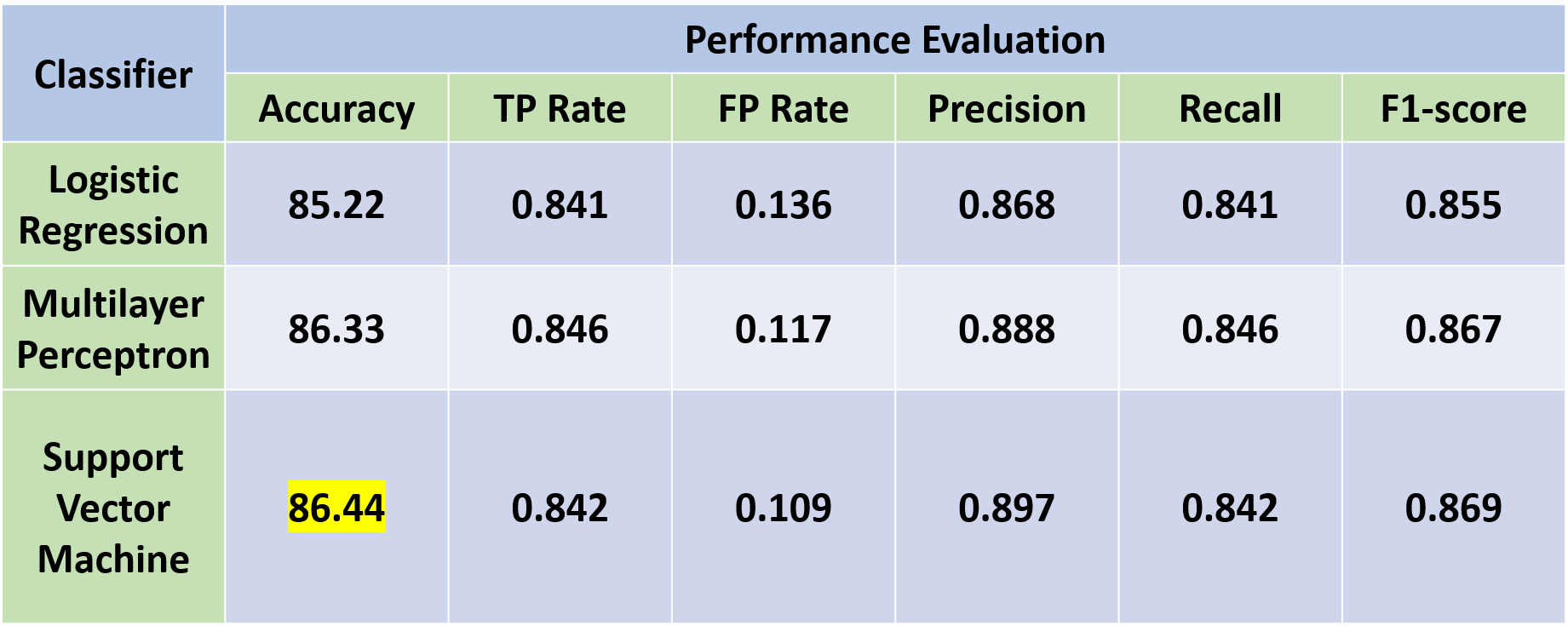
Now that we have seen what WEKA is and what it does, in the next chapter let us learn how to install WEKA on your local computer.

**Chapter 6 – Experimental Results**

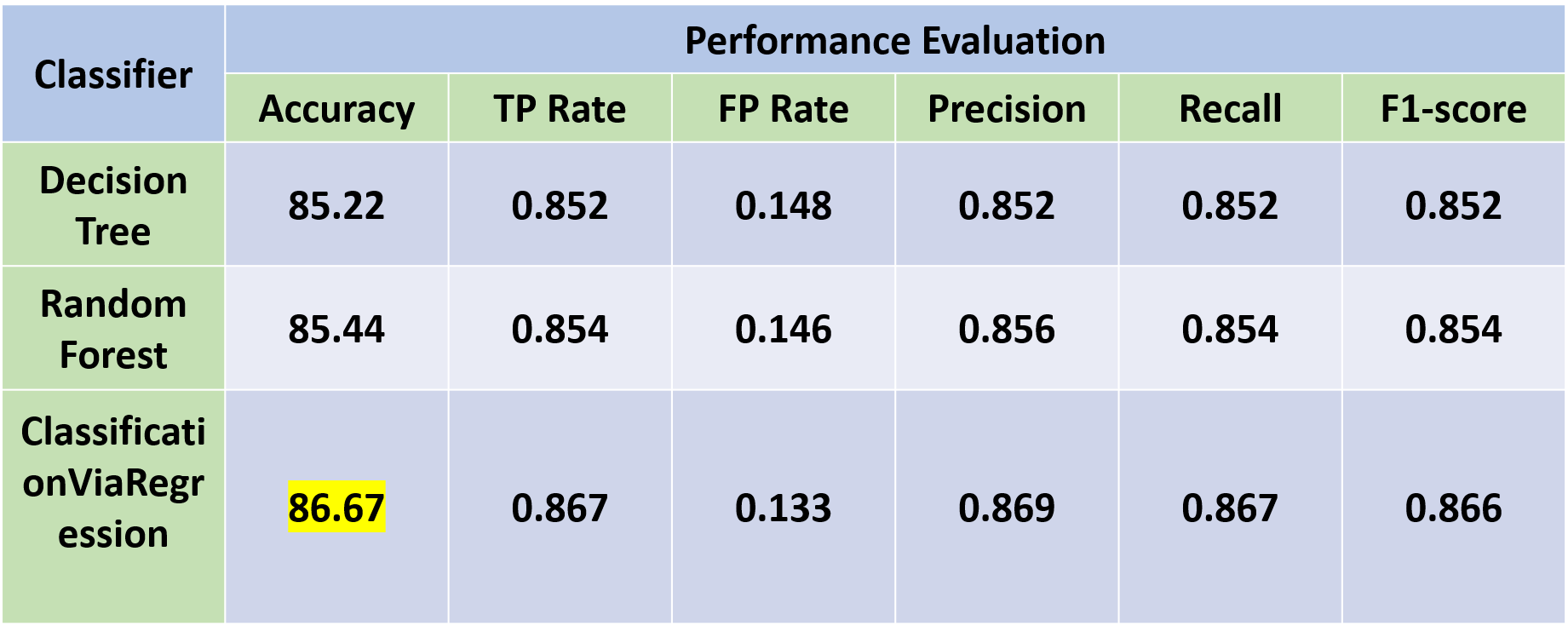
**Attributes Information**



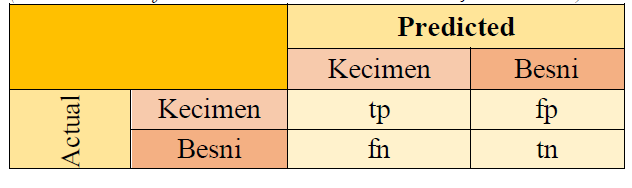
**Performance Evaluation in given in [2]**

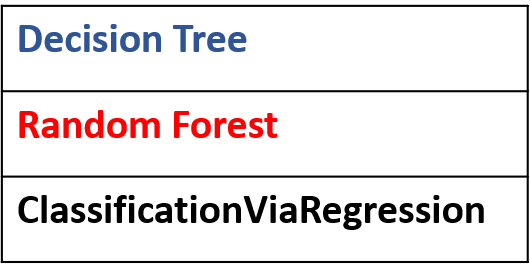


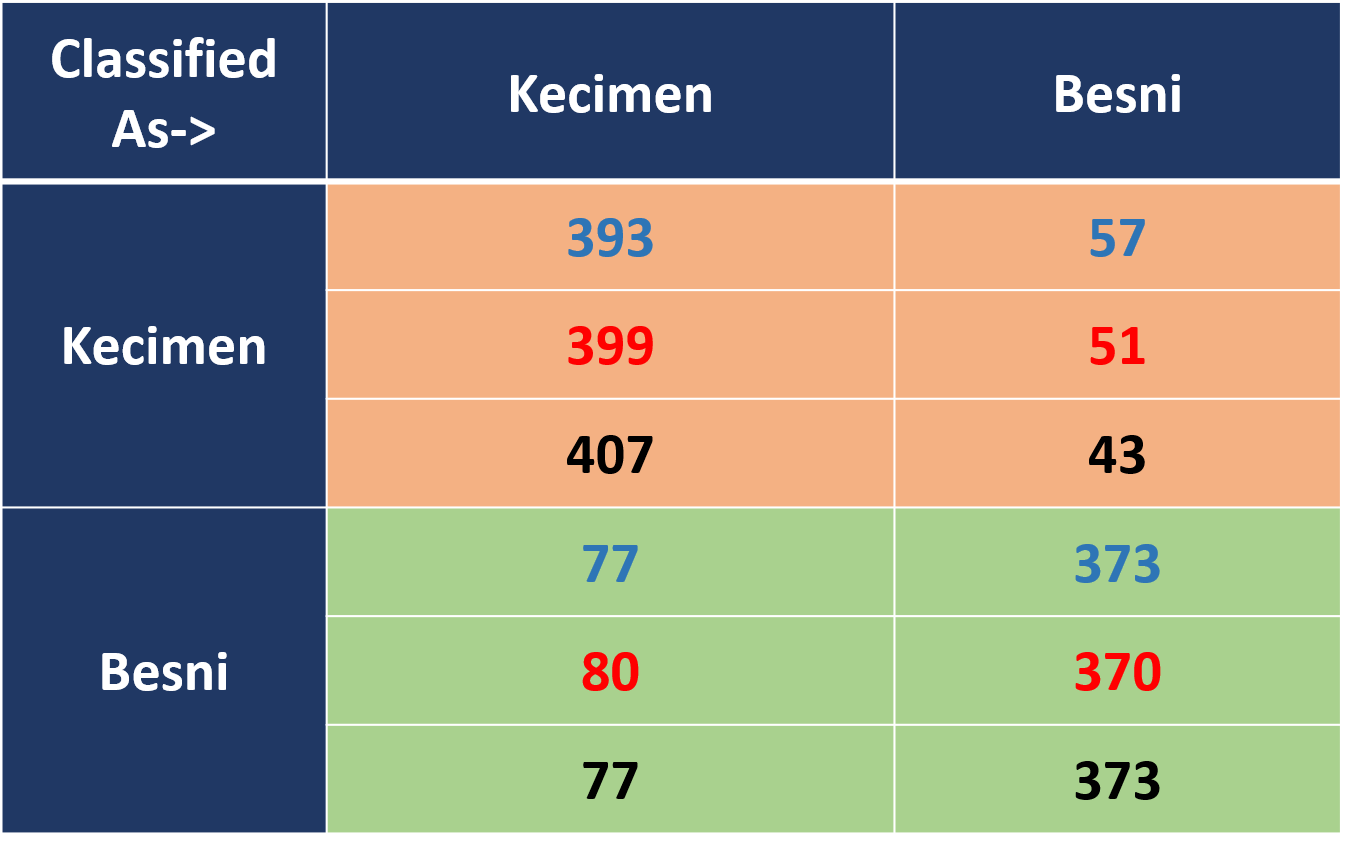
**Performance Evaluation**



**Confusion Matrix-**







**Chapter 7 – Conclusion**

* Machine Learning is an efficient method for raisin classification.
* In this project different classifiers like Decision Tree, Random forest and ClassificationViaRegression are applied.
* Among these, classificationViaRegression has given the highest accuracy of 86.67%

**Chapter 8 – References**

1. avab Karimi, Ramin Ranjbarzadeh Kondrood, Tohid Alizadeh, An intelligent system for quality measurement of Golden Bleached raisins using two comparative machine learning algorithms, Measurement, Volume 107, 2017, Pages 68-76, ISSN 0263-2241, https://doi.org/10.1016/j.measurement.2017.05.009.
2. Frank, E., Wang, Y., Inglis, S., Holmes, G. and Witten, I.H. (1998) Using Model Trees for Classification. Machine Learning, 32, 63-76.  
   <http://dx.doi.org/10.1023/A:1007421302149>.
3. *Kaveh Mollazade, Mahmoud Omid, and Arman Arefi. 2012. Comparing data mining classifiers for grading raisins based on visual features. Comput. Electron. Agric. 84 (June, 2012), 124–131. DOI:https://doi.org/10.1016/j.compag.2012.03.004*
4. Xu M, Sun J, Zhou X, Tang N, Shen J, Wu X. Research on nondestructive identification of grape varieties based on EEMD-DWT and hyperspectral image. J Food Sci. 2021 May;86(5):2011-2023. doi: 10.1111/1750-3841.15715. Epub 2021 Apr 22. PMID: 33885160.
5. Zhao, Yun et al. A NOVEL RAISIN SEGMENTATION ALGORITHM BASED ON DEEP LEARNING AND MORPHOLOGICAL ANALYSIS. Engenharia Agrícola [online]. 2019, v. 39, n. 5 [Accessed 17 April 2022] , pp. 639-648. Available from: <https://doi.org/10.1590/1809-4430-Eng.Agric.v39n5p639-648/2019>. Epub 04 Nov 2019. ISSN 1809-4430. <https://doi.org/10.1590/1809-4430-Eng.Agric.v39n5p639-648/2019>.
6. Dayaset- <https://archive.ics.uci.edu/ml/machine-learning-databases/00617/>
7. Weka- https://www.cs.waikato.ac.nz/ml/weka/