CPS 844 Assignment 1

02/28/2022
Section 1
Ryan Soliven
500840952

Dataset Information

The dataset used was taken from the UCI Machine Learning Repository and is the Raisin Dataset[1]. The abstract taken from the site says “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.” The data set is multivariate and its attributes are integer and real. The associated tasks are classification with 900 number of instances and 8 number of attributes.

Attribute Information

The attribute information taken from data set is as follows:

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.

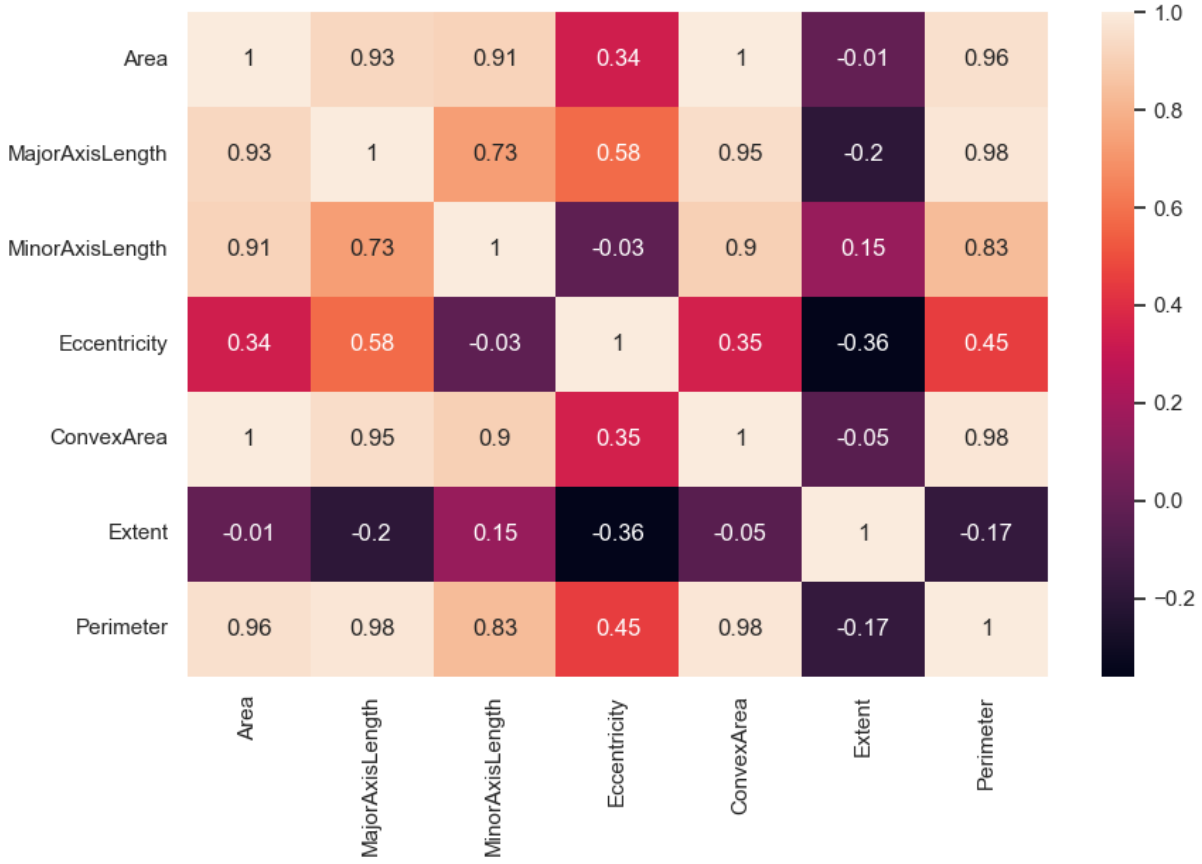
[1] Source link of dataset: https://archive.ics.uci.edu/ml/datasets/Raisin+Dataset

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.

Important Attributes for Classification Models



- **Image 1**

To find the attributes that are most important for our classifiers, we generate a correlation matrix and visualize the heatmap as we can see in **Image 1**. We see that most of the attributes appear to have a strong positive correlation with each other.

When looking at the ‘Area’ attribute, it has strong positive correlations with the ‘MajorAxisLength’, ‘MinorAxisLength’, and ‘Perimeter’ attributes.

When looking at the ‘MajorAxisLength’ attribute, it has strong positive correlations with the ‘Area’, ‘MinorAxisLength’, ‘Convex’, and ‘Perimeter’ attributes.

When looking at the ‘MinorAxisLength’ attribute, it has strong positive correlations with the ‘Area’, ‘MajorAxisLength’, ‘Convex’, and ‘Perimeter’ attributes.

When looking at the ‘Eccentricity’ attribute, it has positive correlations with the ‘Area’, ‘MajorAxisLength’, ‘Convex’, and ‘Perimeter’ attributes.

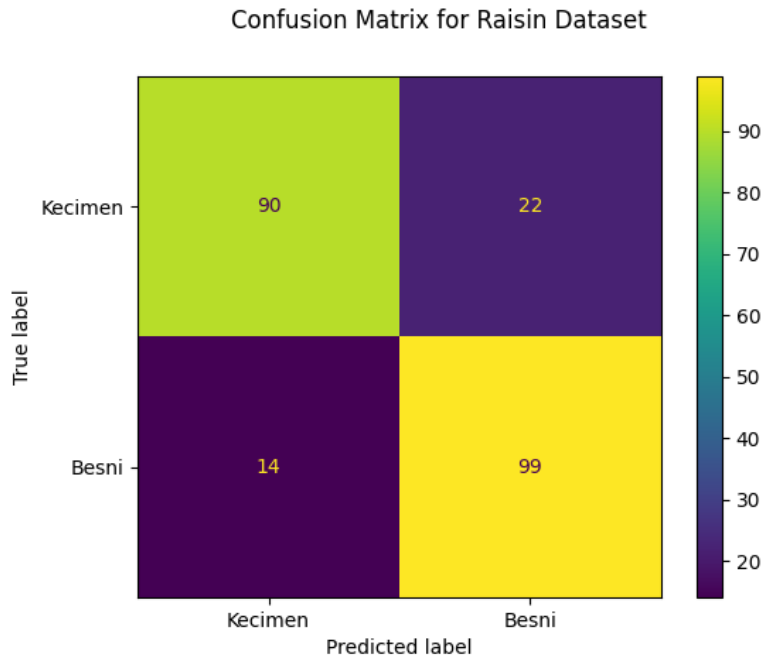
When looking at the ‘Perimeter’ attribute, it has positive correlations with all attributes except for the ‘Extent’ attribute.

The only attribute that does not have positive correlations with other attributes is the ‘Extent’ attribute. We see that it has negative correlations with all attributes.

Goals

The goal is to use five different classifications techniques that can find a model to map each attribute to either the Kecimen class, or the Besni class, so that we can use this model on new records to assign them to one of the classes as accurately as possible. For each classifier, we first standardize the data, to get the best accuracy score. We then split the data, to train the classifier, and later to test the classifier. We will use functions that provide a classification report, the accuracy, and a confusion matrix to compare the results of each classification technique.

K-Nearest Neighbors Classifier

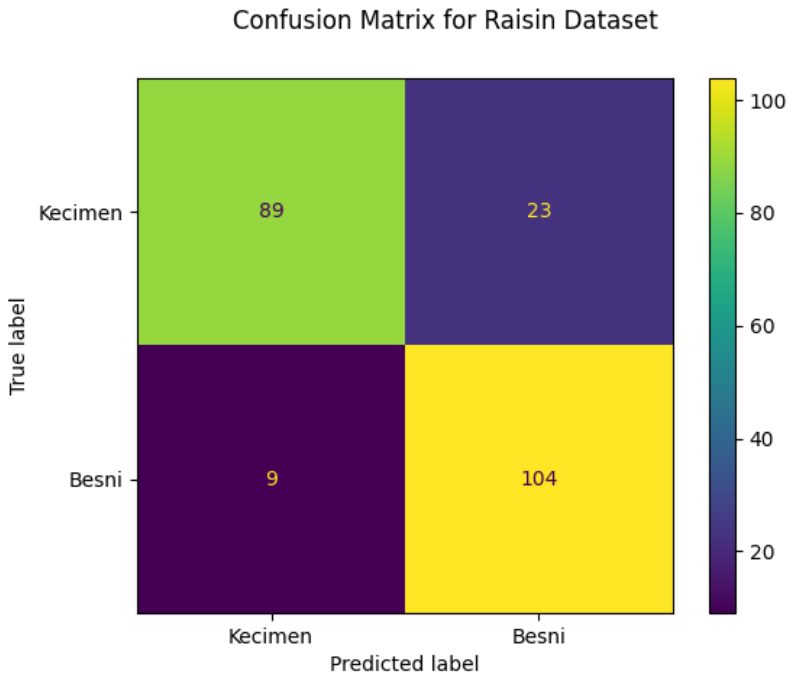


- **Image 2**

Results

After training the classifier with the training attributes and class, and applying the K-Nearest Neighbors classifier to the test data, we compute the accuracy of the classifier with the test data to be 84%. To get the best accuracy, the data was first scaled instead of normalized. When normalized and tested, the accuracy was lower. When we look at **Image 2**, we find that 90 raisins were correctly classified as Kecimen, and 99 were correctly classified as Besni. We also see that 22 Kecimen raisins were incorrectly classified as Besni and 14 Besni were incorrectly classified as Kecimen.

Artificial Neural Networks Classifier

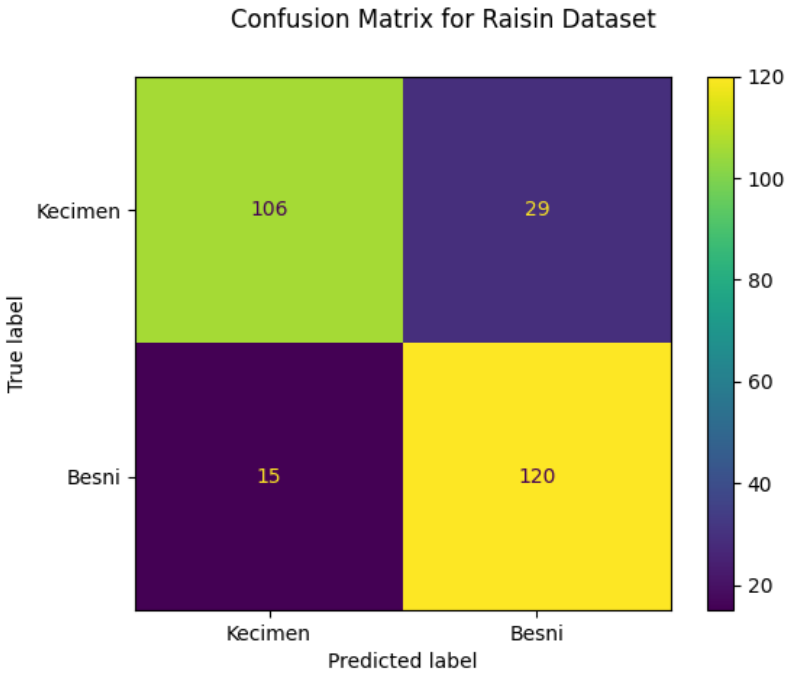


- **Image 3**

Results

After training the classifier with the training attributes and class, and applying the Artificial Neural Networks classifier to the test data, we compute the accuracy of the classifier with the test data to be 85%. To get the best accuracy, the data was first scaled instead of normalized. When normalized and tested, the accuracy was lower. When we look at **Image 3**, we find that 89 raisins were correctly classified as Kecimen, and 104 were correctly classified as Besni. We also see that 23 Kecimen raisins were incorrectly classified as Besni and 9 Besni were incorrectly classified as Kecimen.

Decision Tree Classifier

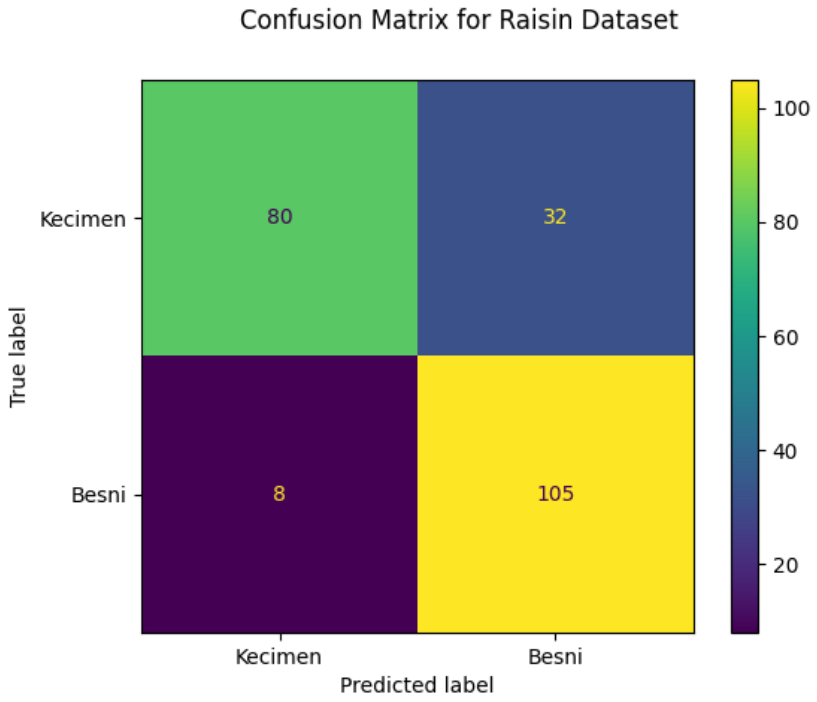


- **Image 4**

Results

After training the classifier with the training attributes and class, and applying the Decision Tree classifier to the test data, we compute the accuracy of the classifier with the test data to be 84%. To get the best accuracy, the data was first scaled instead of normalized. When normalized and tested, the accuracy was lower. When we look at **Image 4**, we find that 106 raisins were correctly classified as Kecimen, and 120 were correctly classified as Besni. We also see that 29 Kecimen raisins were incorrectly classified as Besni and 15 Besni were incorrectly classified as Kecimen.

Gaussian Naive Bayes Classifier

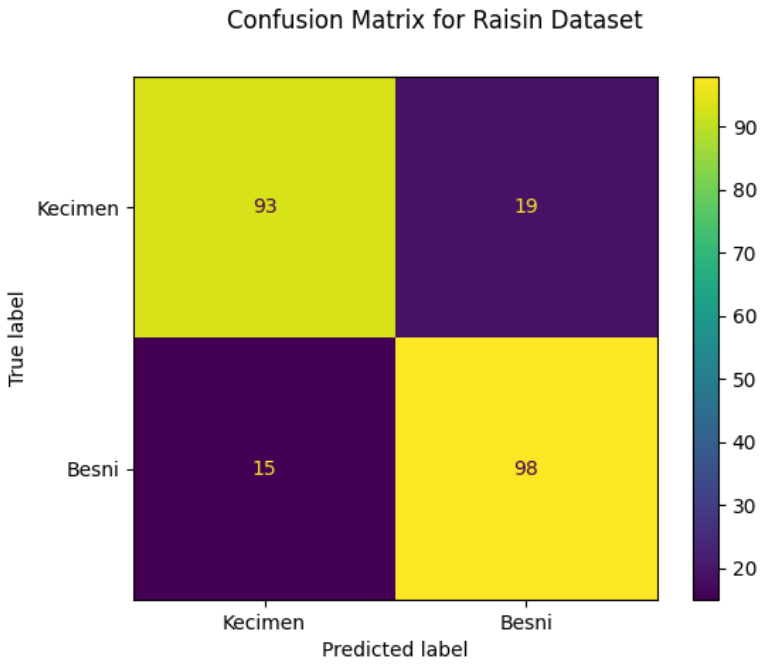


- **Image 5**

Results

After training the classifier with the training attributes and class, and applying the Decision Tree classifier to the test data, we compute the accuracy of the classifier with the test data to be 82%. To get the best accuracy, the data was first scaled instead of normalized. When normalized and tested, the accuracy was lower. When we look at **Image 5**, we find that 80 raisins were correctly classified as Kecimen, and 105 were correctly classified as Besni. We also see that 32 Kecimen raisins were incorrectly classified as Besni and 8 Besni were incorrectly classified as Kecimen.

Logistic Regression Classifier



- **Image 6**

Results

After training the classifier with the training attributes and class, and applying the Decision Tree classifier to the test data, we compute the accuracy of the classifier with the test data to be 85%. To get the best accuracy, the data was first scaled instead of scaled. When we look at **Image 6**, we find that 93 raisins were correctly classified as Kecimen, and 98 were correctly classified as Besni. We also see that 19 Kecimen raisins were incorrectly classified as Besni and 15 Besni were incorrectly classified as Kecimen.

Conclusion

When comparing all five classification techniques, we find that they all share a similar accuracy score. The most accurate classifiers were the Artificial Neural Networks classifier, and the Logistic Regression classifier, sharing the accuracy of 85%. The K-Nearest Neighbors classifier and Decision Tree classifier had an accuracy of 84%. The worst performing one was the Gaussian Naive Bayes Classifier with an accuracy of 82%.