Pavlo Nikolaiev

Practical 3

Title name is - Dry Bean Dataset

Link for my database - <https://archive.ics.uci.edu/ml/datasets/Dry+Bean+Dataset>

Authors :

Murat KOKLU

Faculty of Technology,

Selcuk University,

TURKEY.

ORCID : 0000-0002-2737-2360

[mkoklu@selcuk.edu.tr](mailto:mkoklu@selcuk.edu.tr)

Ilker Ali OZKAN

Faculty of Technology,

Selcuk University,

TURKEY.

ORCID : 0000-0002-5715-1040

[ilkerozkan@selcuk.edu.tr](mailto:ilkerozkan@selcuk.edu.tr)

a way how the dataset was collected:

Images of 2019 grains of 7 different registered dry beans were taken with a high-resolution camera. A total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains.

the number of data objects in the dataset – 2019

a representation of the features (attributes) of the dataset together with their roles in the Orange tool;

Attribute Information:

1.) Area (A): The area of a bean zone and the number of pixels within its boundaries.

2.) Perimeter (P): Bean circumference is defined as the length of its border.

3.) Major axis length (L): The distance between the ends of the longest line that can be drawn from a bean.

4.) Minor axis length (l): The longest line that can be drawn from the bean while standing perpendicular to the main axis.

5.) Aspect ratio (K): Defines the relationship between L and l.

6.) Eccentricity (Ec): Eccentricity of the ellipse having the same moments as the region.

7.) Convex area (C): Number of pixels in the smallest convex polygon that can contain the area of a bean seed.

8.) Equivalent diameter (Ed): The diameter of a circle having the same area as a bean seed area.

9.) Extent (Ex): The ratio of the pixels in the bounding box to the bean area.

10.)Solidity (S): Also known as convexity. The ratio of the pixels in the convex shell to those found in beans.

11.)Roundness (R): Calculated with the following formula: (4piA)/(P^2)

12.)Compactness (CO): Measures the roundness of an object: Ed/L

13.)Class (Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sira)

The number of classes in the dataset, the meaning of each class and the way of representing classes (explanation of the labels assigned to classes); if the data set provides several possible data classifications, then the report must identify which classification is considered in the assignment; o the number of data objects belonging to each class;

In this database we have 7 classes - Seker, Barbunya, Bombay, Cali, Dermosan, Horoz and Sir.The name of classes it is the sorts of dry beams.

the number of data objects belonging to each class;

Seker – 87;

Barbunya – 149;

Bombay – 522;

Cali – 514;

Horoz – 289;

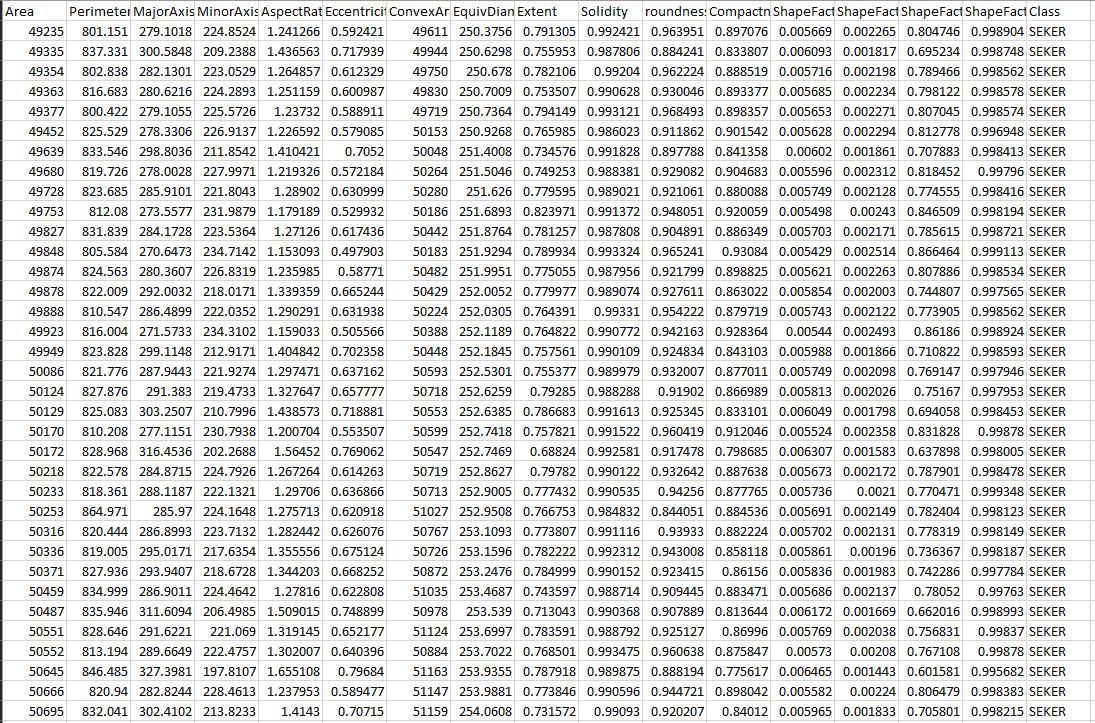
Sir – 38;

Dermosan – 418;

the number and meaning of features in the dataset, as well as their value types and ranges (this information should be presented in a table consisting of the feature representation, its meaning, value type and range of values available in the dataset);

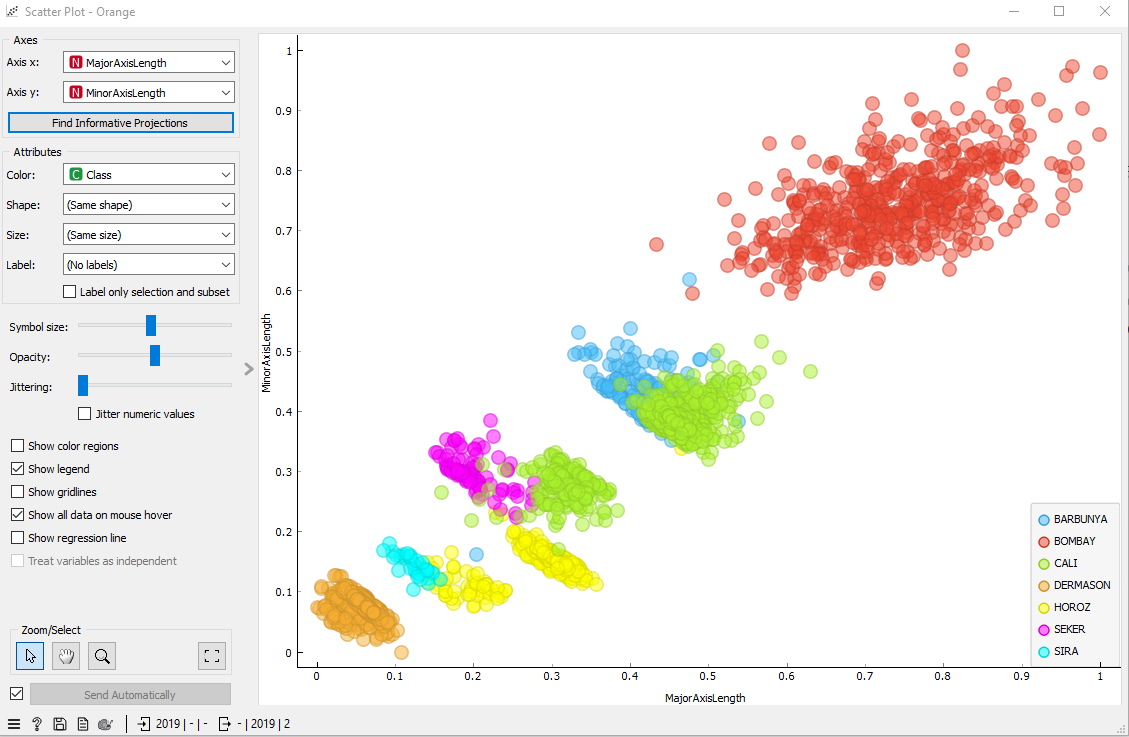
|  |  |  |  |
| --- | --- | --- | --- |
| features | Min | Max | Type |
| Area | 21558 | 254616 | Int |
| Perimeter | 531.318 | 1985.37 | float |
| MajorAxisLength | 187.1686 | 738.8602 | float |
| MinorAxisLength | 122.5127 | 460.1985 | float |
| AspectRation | 1.148116 | 2.387395 | float |
| Eccentricity | 0.491297 | 0.908048 | float |
| ConvexArea | 21808 | 263261 | float |
| EquivDiameter | 165.6759 | 569.3744 | float |
| Extent | 0.581852 | 0.85842 | float |
| Solidity | 0.944568 | 0.994678 | float |
| roundness | 0.607867 | 0.990685 | float |
| Compactness | 0.646562 | 0.932794 | float |
| ShapeFactor1 | 0.002778 | 0.010451 | float |
| ShapeFactor2 | 0.000564 | 0.00347 | float |
| ShapeFactor3 | 0.418042 | 0.870104 | float |
| ShapeFactor4 | 0.957325 | 0.999505 | float |

a snippet of the structure of your datafile in which the columns of your datafile and class labels are shown together with some data objects;

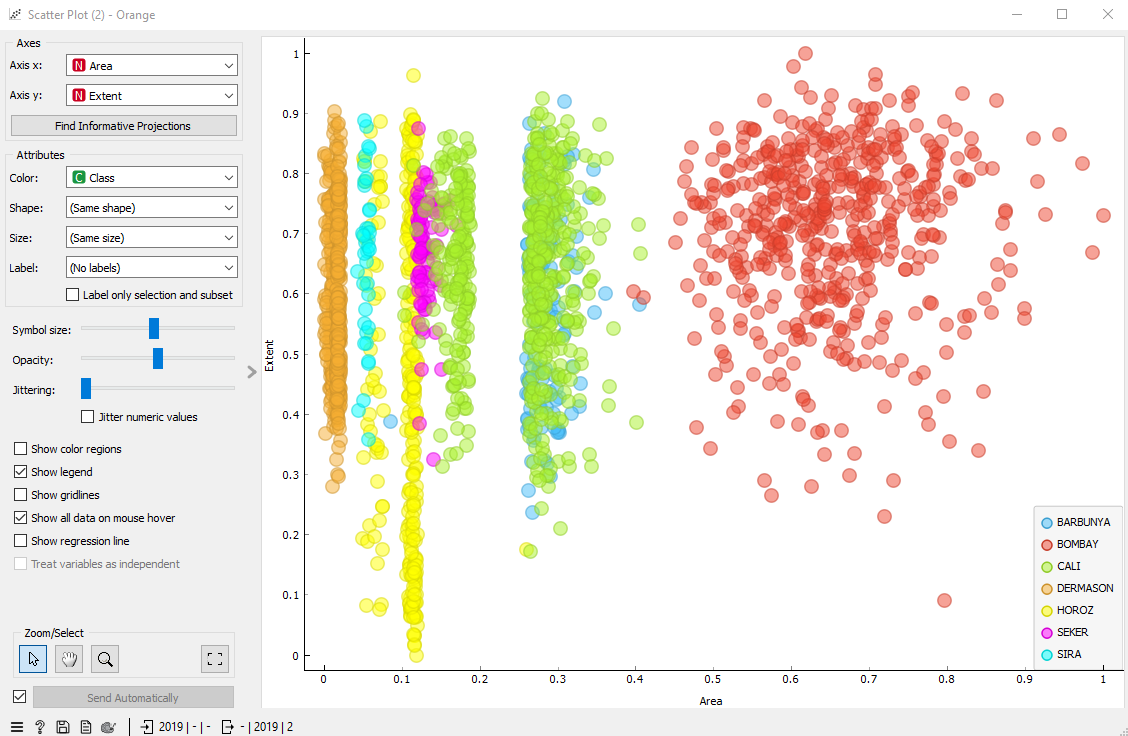


Representing your training dataset visually and statistically:

You must create at least two 2- or 3-dimensional scatter plots illustrating the separability of classes in your dataset based on different features; the student should avoid using the data object ID or a class label as a variable in the scatterplot

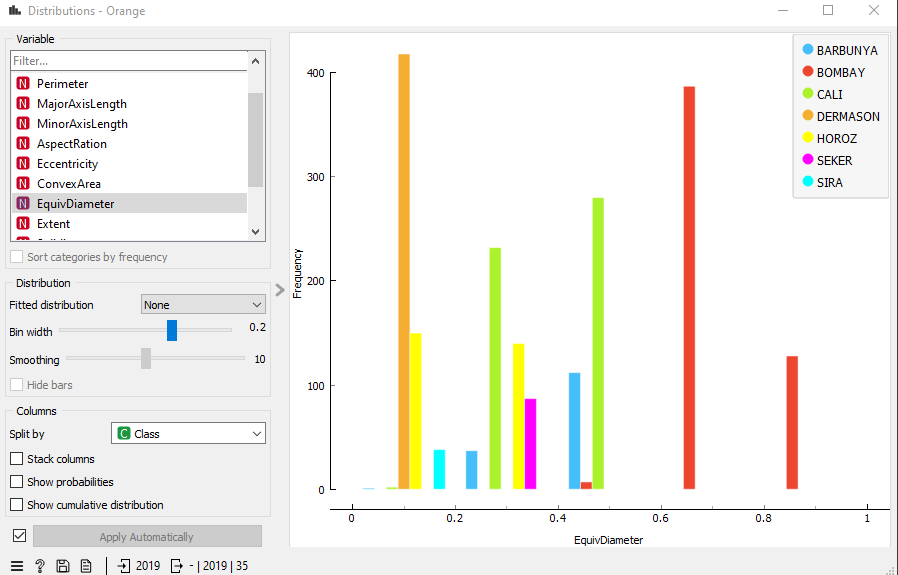
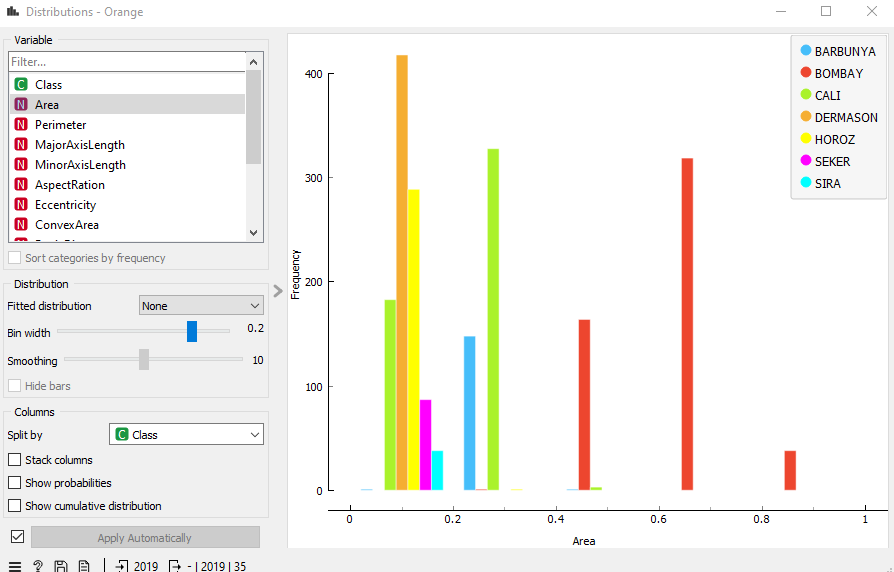


By graphing the major axis length (L) against intensity (I), we gain insight into how bean intensity varies with respect to their major axis length. Such a plot may reveal any patterns or links between these two variables.

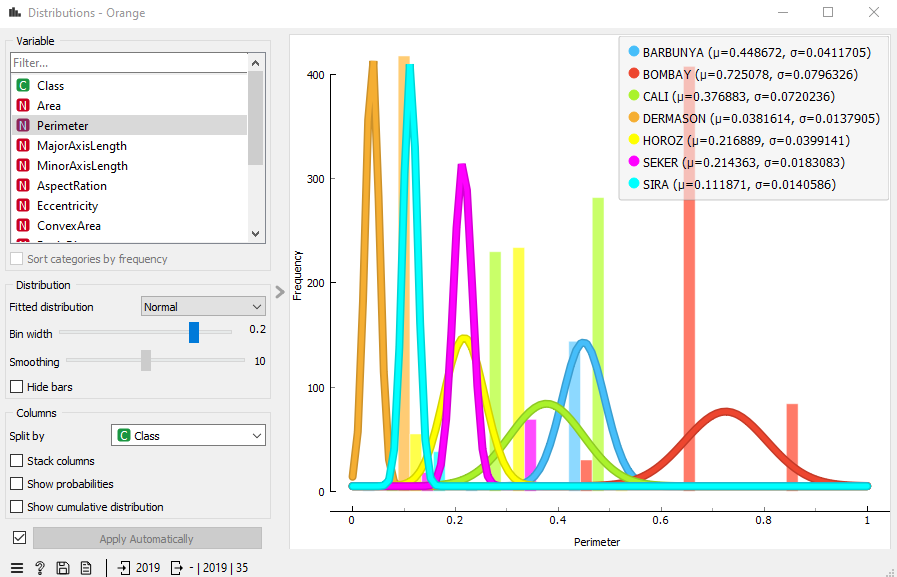
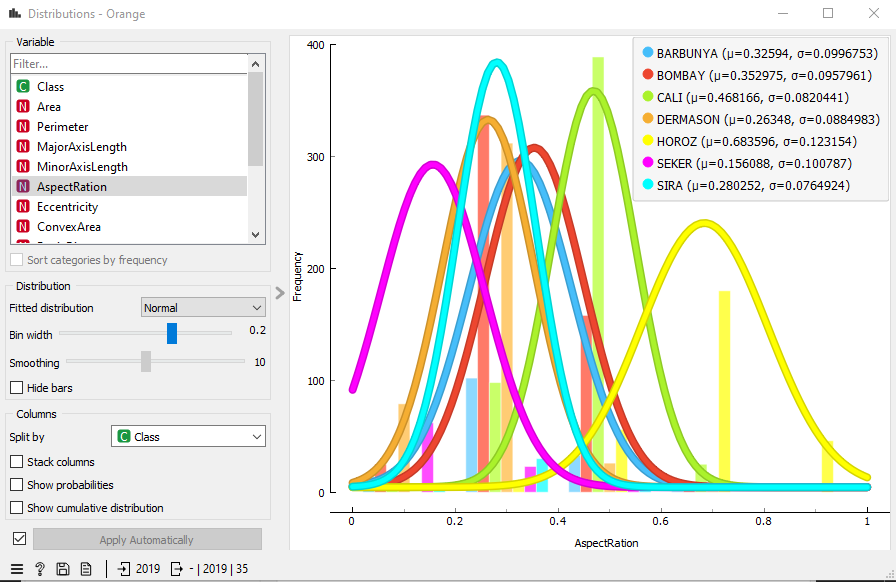


If we depict Area (A) on the x-axis and Extent (Ex) on the y-axis, a plot can demonstrate how changes in bean size impact their extent. By analyzing pixel ratios within a given bounding box, this visualization may reveal relationships or patterns between these two measured variables.

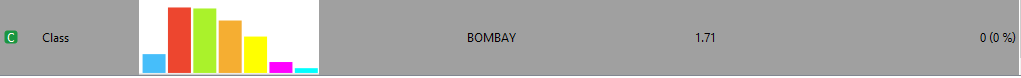
you must create at least 2 histograms showing the separation of classes for the features of interest;



you must show 2 distributions for the features of interest;



you must calculate statistics on your data (at least the central tendency and the dispersion of the feature values).



1- conclusions coming from the analysis of scatter plots, histograms and distributions (from Step 5 in Part I) about the separability of your classes (remember to include your graphs in the report). Try to answer the following questions:

Are classes in your dataset balanced, or is one class (several classes) prevailing? It is determined by how many data objects belong to each class: In my classes, the entered data from the video camera and they cannot have the same values, in my data, most of the data is in the BOMBAI class, namely 522 values.

2- Does the visual representation of the data allow the structure of the data to be seen? It is a question of whether data objects belonging to different classes are separable.

The diagram shows that the product data and if you make more comparisons, then some varieties will stand out more than others.

3- Yes, the beans differ only visually and there is nothing more.

No ,identification signs are far apart.

Conclusions coming from the analysis of statistical calculations (central tendency and dispersion).

Based on my data, I realized that my data is asymmetric and also that there is a lot of variability and spread in feature values.

Part 2

Perform at least 3 experiments with Hierarchical clustering, freely moving the cut-off line and analysing how the number and content of clusters change;

In this algorithm we will define the

A screenshot of a computer

Description automatically generated with medium confidence

In the first slide we see the cut-off line placing in the end of diagram and do not cut anything and like shown in the screen we have all the results consist of.

A screenshot of a computer

Description automatically generated

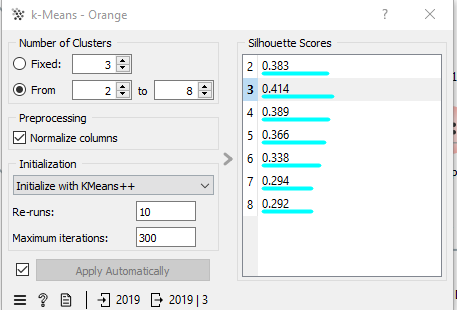
In the second screen shown that cut-off line placing in the 73.8% saving the data which values is bigger then the max in line that’s values will not accepted. And we see the different Class BARBUNYA,CALI,SEKIR and SIRA did not changed because the values was to small.

A screenshot of a computer

Description automatically generated with medium confidence

And the same situation with the second screen ,here the cut-of line saving the 30,1% out of the biggest value. And here we see only 1 class left because this class has the smallest values on dataset.

For the K-means algorithm, calculate the Silhouette Score for at least 5 k values, and analyze the performance of the algorithm.



In this algorithm we have 7 clusters 2 to 8 . The Clusters has range -1 to 1 and here we see only positive numbers that shown the clusters work correctly .In the slide shown – That clusters is stable because the different is little bit more than 0.1.The most powerful cluster it is the number 3 then 4,2,5,6,7 and 8.

Part 3