Project 3 Dry Beans Clustering and Classification

Dataset

About data:

Seven different types of dry beans were used in this research, taking into account the features such as form, shape, type, and structure by the market situation. A computer vision system was developed to distinguish seven different registered varieties of dry beans with similar features in order to obtain uniform seed classification. For the classification model, images of 13,611 grains of 7 different registered dry beans were taken with a high-resolution camera. Bean images obtained by computer vision system were subjected to segmentation and feature extraction stages, and a total of 16 features; 12 dimensions and 4 shape forms, were obtained from the grains.

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.) ShapeFactor1 (SF1)

14.) ShapeFactor2 (SF2)

15.) ShapeFactor3 (SF3)

16.) ShapeFactor4 (SF4)

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

Procedure

1. Import Data
2. Check dataset size
3. Find and treat missing values (If any)
4. Check column types and describe which columns are numerical or categorical
5. Perform Univariate analysis
    1. Calculate mean, median, std dev, and quartiles of numerical data
    2. Plot histogram of ‘Class.’
    3. Check the distribution of numerical variables and comment on it
6. Perform Bivariate analysis

    1. Plot box plot of ‘AREA’ segregated by different classes (Hint take ‘Class’ on the y-axis)
    2. Plot box plot of ‘PERIMETER’ segregated by different classes
    3. Calculate Pearson correlation, and plot the heatmap
7. Drop columns having high correlation and any unnecessary columns
8. Split into train and test set
9. Scale the variables
10. Use KNN to predict “Class.”
11. Use the elbow method to select the best K in KNN.
12. Drop the “Class” column, and cluster the data using K means.
13. Use the Elbow method to find the best K for K means
14. Take K = 7, and do K means clustering
15. Plot a histogram of different Kmeans clusters segregated by “Class” (Hint take “Class” as hue)
16. Comment on your findings
17. PCA
    0. Decrease the dimensions to two using PCA
    1. Use K means clustering with K found earlier
    2. Plot clusters

Compulsory

1. Use Multi-class logistic regression to predict “Class” and compare the performance with KNN.
2. Use PCA to reduce the dimensions to two and perform classification to predict the “Class,” and compare performance with previous models.