DS5230 Final: Phase 1

Michael Massone and Nelson Farrell

Spring 2024

Dry Bean Dataset

Size: (13611 x 17) – 13611 rows, 17 columns

This dataset is composed of data derived from 13,611 images of 7 species of beans. 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 generated.

KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507.
DOI: https://doi.org/10.1016/j.compag.2020.105507
Link: https://archive.ics.uci.edu/dataset/602/dry+bean+dataset

Target Class

TARGET: 7 SPECIES OF DRIED BEANS

- Barbunya
- Bombay
- Cali
- Dermosan
- Horoz
- Seker
- Sira

NUMERICAL ENCODING

- Barbunya: 0
- Bombay: 1
- Cali: 2
- Dermosan: 3
- Horoz: 4
- Seker: 5
- Sira: 6

Features

Total: 16

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.

Features

(continued)

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)

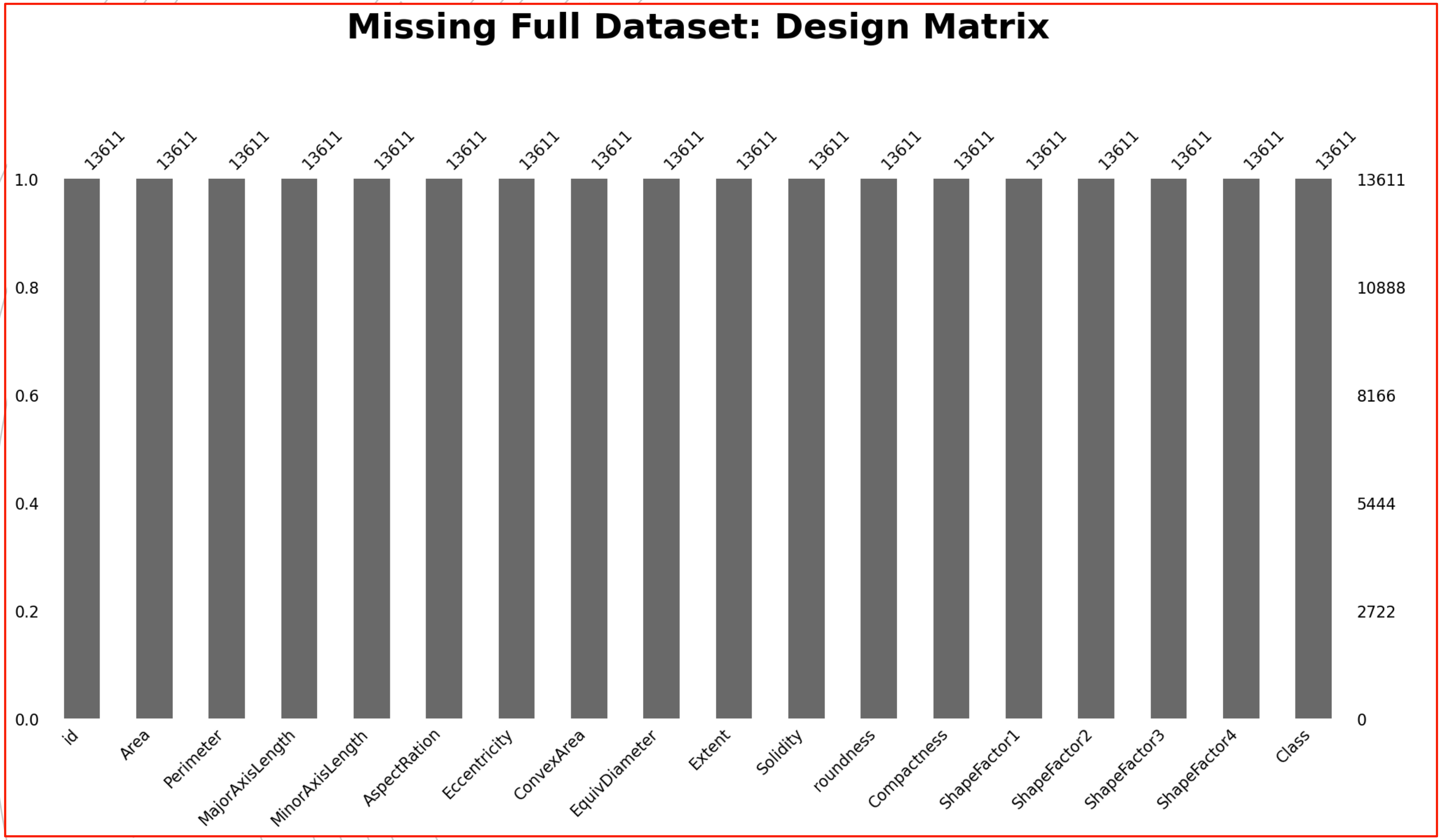
16) ShapeFactor4 (SF4)

Features

Characteristics

All features in dataset are measurements based on image processing of the original bean images. Each feature is a ratio measure, based on a length measurement or some function of various lengths. The ShapeFactors are also calculated from measured dimensions, although which dimensions are not specified.

Features	Data Type	Measurement Scale
Area	float	Ratio
Perimeter	float	Ratio
Major Axis Length	float	Ratio
Minor Axis Length	float	Ratio
Aspect Ratio	float	Ratio
Eccentricity	float	Ratio
Convex Area	float	Ratio
Equivalent Diameter	float	Ratio
Extent	float	Ratio
Solidity	float	Ratio
Roundness	float	Ratio
Compactness	float	Ratio
ShapeFactor1	float	Ratio
ShapeFactor2	float	Ratio
ShapeFactor3	float	Ratio
ShapeFactor4	float	Ratio



Missingness

- Target Missingness: 0
- Feature Missingness: 0

Acknowledgments

- KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans Using Computer Vision and Machine Learning Techniques.” Computers and Electronics in Agriculture, 174, 105507. DOI: https://doi.org/10.1016/j.compag.2020.105507
- Dr. Steve Morin – Class slides and labs.
- UC Irvine Machine Learning Repository