Dry Bean Classification

Abinash Sonowal 210101004 Anoop Singh 210101016 Kundan Meena 210101060 Ravi Lahare 210101086

Abstract

This project revolves around the development of a classification model aimed at accurately identifying various types of dry beans based on their attributes. Leveraging a dataset comprising features extracted from images of dry beans, we employed machine learning techniques to construct a predictive model. The dataset, which can be accessed through a provided link, contains 13611 instances across 17 attributes, including 12-dimensional and 4 shape features along with the target class for bean classification. Seven distinct dry bean classes are represented in the dataset. Our methodology involved thorough data cleaning to ensure integrity, univariate analysis to understand feature distributions and outliers, and bivariate analysis to explore relationships between features and classes. Notable findings include distinct characteristics of the Bombay class and the identification of outlier instances predominantly in the Bombay class. Through comprehensive analysis and visualization, this project contributes to quality control and agricultural research, potentially enhancing bean classification accuracy and efficiency.

1. Introduction

Our project focuses on developing a classification model for identifying different types of dry beans based on their attributes. We'll be utilizing a dataset containing various features extracted from images of dry beans. Through machine learning techniques, we aim to build a predictive model for accurate bean classification. The dataset can be accessed via this link. This project holds significance for quality control and agricultural research.

2. Methods

2.1. Data Description

To gain preliminary insights into our dataset, we conducted a comprehensive Pandas profiling analysis. This rigorous examination provided us with a detailed overview and statistical summary of the dataset's characteristics.Panda Profiling Output.

The data set contains 13611 rows and 17 columns. These columns includes 12-dimensional and 4 shape features attributes(Area, Perimeter, MajorAxisLength, MinorAxisLength, AspectRation, Eccentricity, ConvexArea, EquivDiameter, Extent, Solidity, roundness, Compactness, ShapeFactor1, ShapeFactor2, ShapeFactor3, ShapeFactor4) and last attribute is target class for bean(that is our classification output). We have seven specific dry bean classes: Seker, Barbunya, Bombay, Cali, Horoz, Sira, and Dermason. In the dataset we have count value 2027, 1322, 522, 1630, 1928, 2636, 3546 for Seker bean, Barbunya bean, Bombay bean, Cali bean, Horoz bean, Sira bean, and Dermason bean respectively.

2.2. Data Cleaning

During the data cleaning process, we conducted checks for null values in attributes and identified duplicate rows. Fortunately, we did not encounter any null value attributes within the dataset. However, our examination revealed the presence of 68 duplicate rows, which were subsequently removed to ensure data integrity and consistency.

2.3. Univariate Analysis

During the univariate analysis phase, we conducted a comprehensive examination of each feature of the bean dataset in relation to their respective classes. We analyzed and visualized the conditional distribution(Figure 4) of each feature of the bean dataset with respect to the class labels.

Upon inspection, a notable observation emerged: the Bombay class exhibited distinct characteristics markedly different from the other classes, particularly in terms of grain size, which appeared substantially larger. This observation suggested the potential for effective separation of Bombay class instances from other classes with a high degree of certainty.

Further analysis via box plots for each feature revealed the presence of outlier points in the 'area' and 'convex area' attributes.

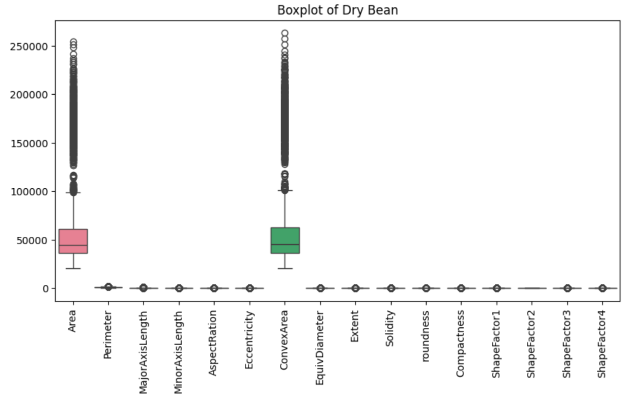

Figure 1. Box Plot.

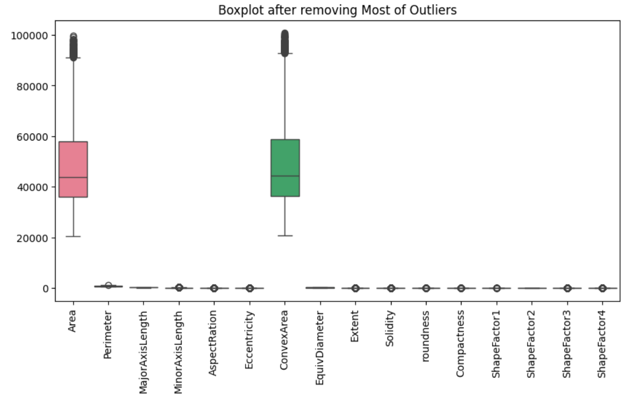

Figure 2. Box Plot After removal.

Consequently, outlier detection procedures were employed, resulting in the identification of 522, 22, and 5 outlier points for the Bombay, Cali, and Barbunya classes, respectively. These outliers were subsequently removed from the dataset to ensure data integrity and robustness of subsequent analyses.

A critical observation arose from this process: all instances falling under the Bombay class were identified as outliers. Consequently, the Bombay class was excluded from further consideration in subsequent analyses.

2.4. Bivariate Analysis

We accounted for relationships between features that appeared to be independent of each other. This analysis provided additional insights into the structure of the data and contributed to our understanding of the underlying patterns

We generated joint scatter plots(Figure 5) of two features and distinguished each class by assigning different colors. Upon visual inspection, notable clusters corresponding to different classes were observed. Furthermore, it was evident that the interpretation of the data was facilitated by the discernible presence of complex patterns.

A key observation was the emergence of two-dimensional regions wherein representatives of almost exclusively one class were contained. Additionally, it was noted that many scatterplots exhibited a certain degree of "isomorphism" - clusters for different classes displayed similar relative locations, potentially up to symmetry. For instance, the Sira class was positioned between other classes, suggesting a distinct pattern in the dataset.

2.5. Citations and References

Source code on gihub -
https://github.com/abinashrasonowal/CS361$_{m}lproject$.

Data set -
https://www.kaggle.com/datasets/nimapourmoradi/drybean-dataset-classification/data.

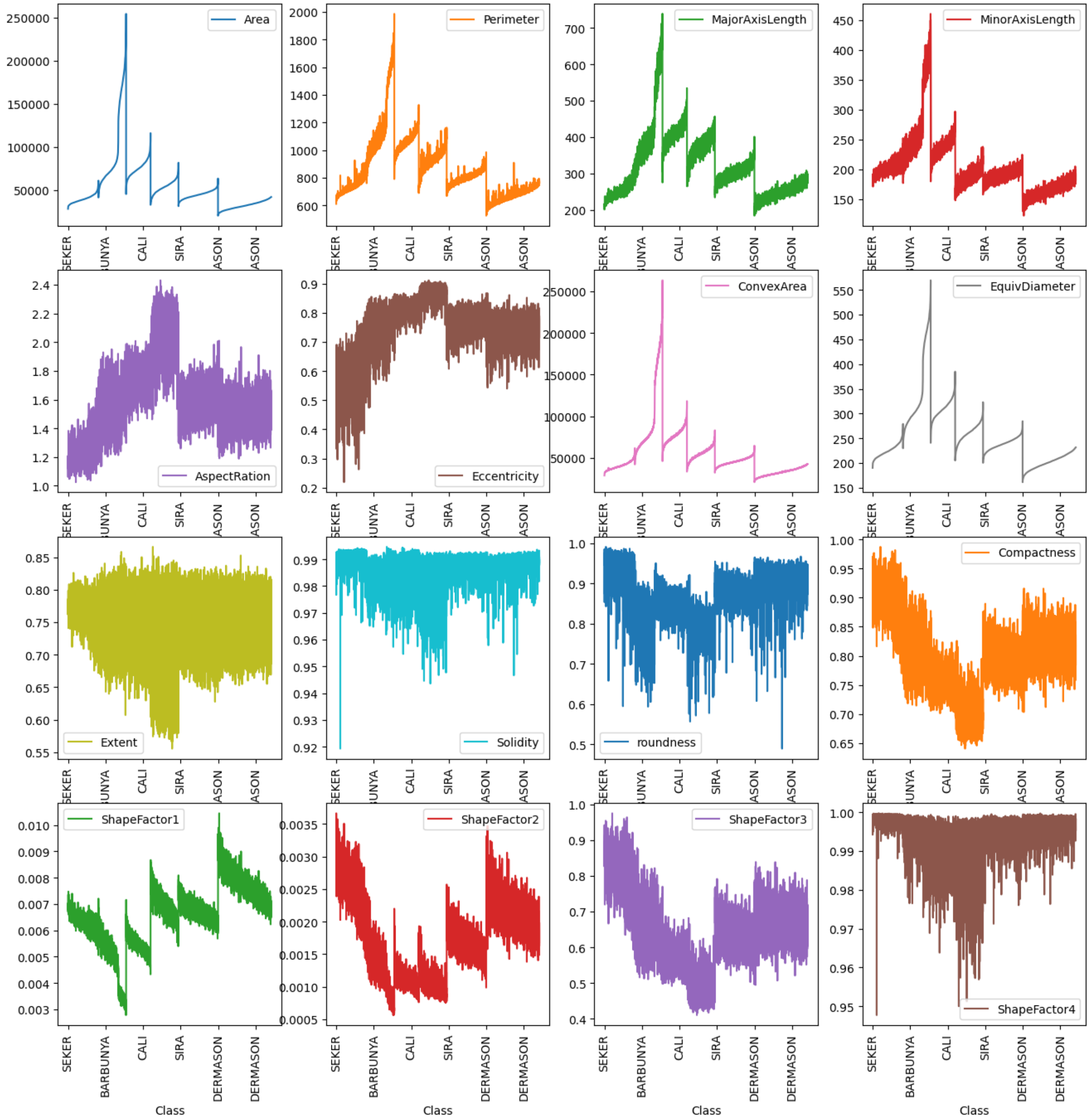

Figure 3. Distribution of Features across Class

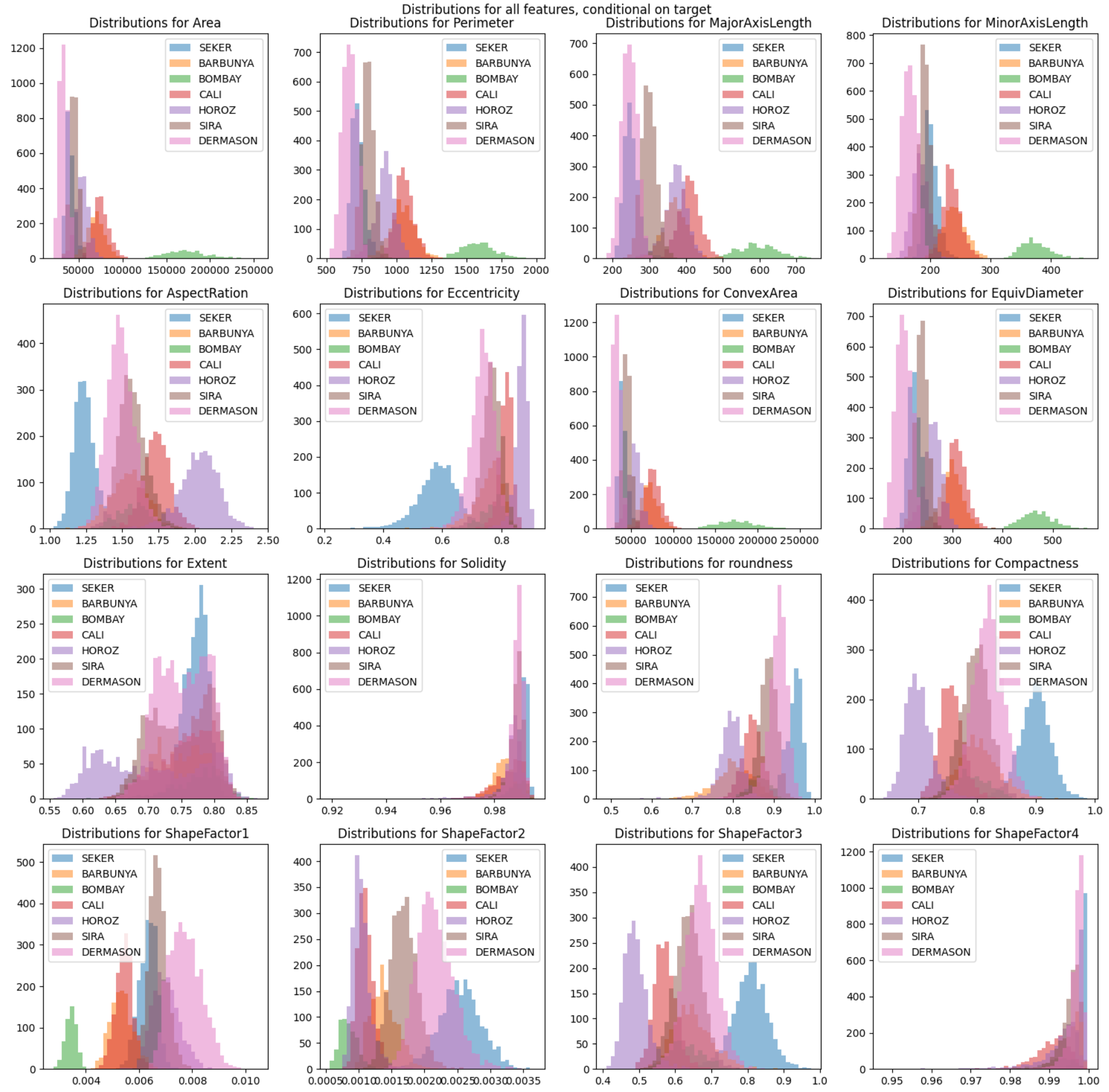

Figure 4. Conditional Distribution on Class

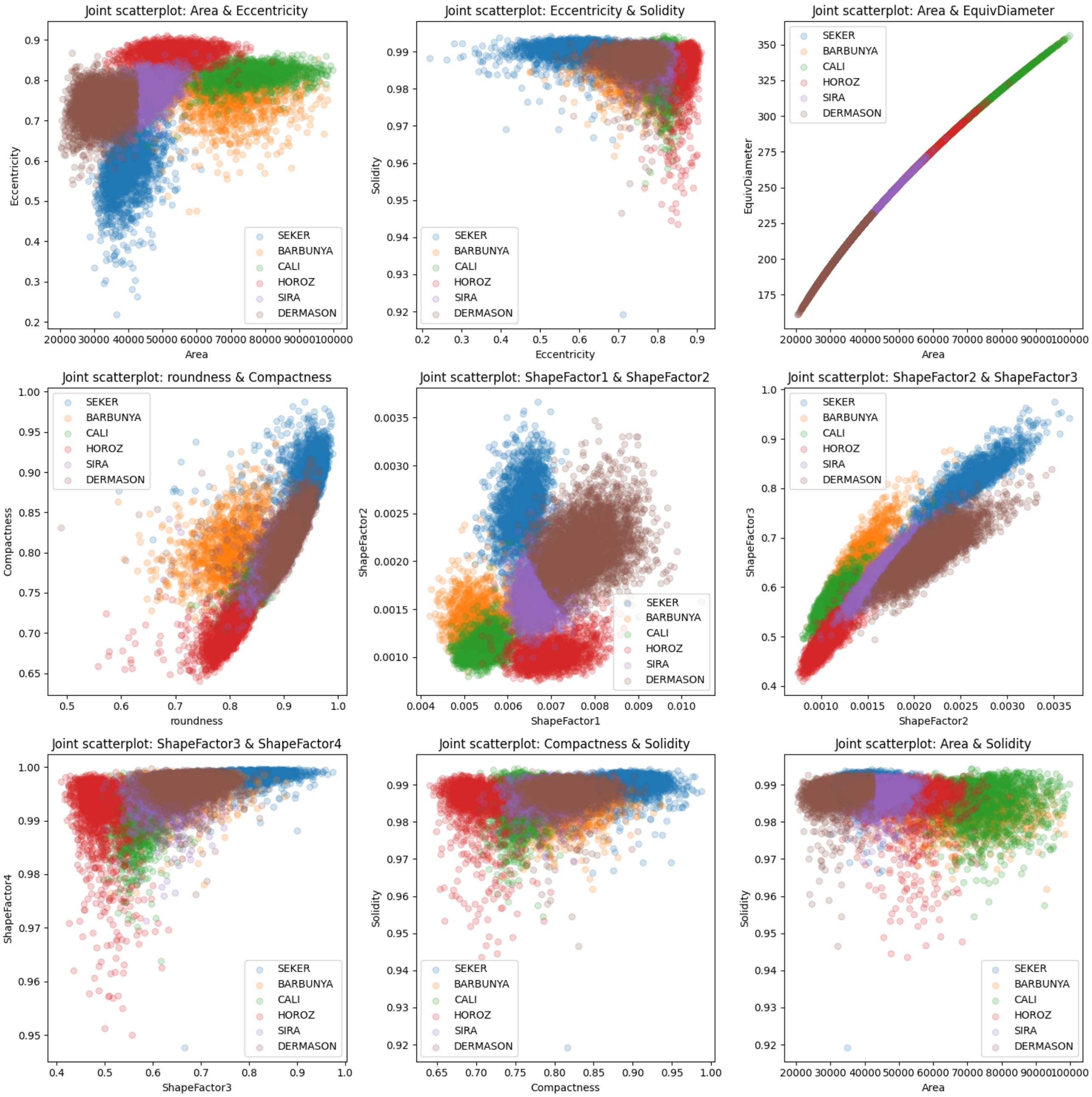

Figure 5. Scatter Plot of Independent Features