Technical University of Denmark

Project 1 - Abalone

Authors

Mathias Munck Nikolajsen - s204271

Jakub Bouzan - s251155

Antonio Javier Cayetano Matamoros - s172861

March 6, 2025

Author/Section	Dataset	Attributes	Visualization	PCA	Discussion and conclusion	Abstract
s251155	20%	20%	60%	20%	20%	50%
s204271	60%	60%	20%	20%	20%	20%
s172861	20%	20%	20%	60%	60%	30%

Table 1: Table of contributions

1 Abstract

This report presents an analysis of the Abalone dataset, focusing on understanding its structure through feature extraction and visualization. The primary objective is to apply the data analysis techniques learned in the course "02450 - Introduction to Machine Learning and Data Mining" to gain insights into the dataset before further modeling. The dataset contains physical measurements of abalones, and the goal is to predict their age, which is a crucial factor in fisheries management and ecological research. The problem centers around determining the most relevant features for predicting age based on measurable attributes such as length, diameter, height, and weight. Since age is not directly recorded, it is inferred by counting shell rings, making it a challenging regression and classification task in report 2. The analysis also explores potential correlations between physical attributes and sex classification. the exploration of the dataset,where various visualization techniques, such as histograms, boxplots, and correlation matrices, were applied to examine attribute distributions, detect outliers, and identify relationships among variables. Additionally, Principal Component Analysis (PCA) was conducted.

.

Contents

1 Abstract 1

2 Description of the dataset 3

3 Explanation of attributes 4

4 Data visualization 5

4.1 The histogram 5

4.2 Boxplot diagram 6

4.3 Correlation between attributes 8

5 Principal Component Analysis 9

6 Discussion and Conclusion 11

2 Description of the dataset

The primary objective of the Abalone dataset is to predict the age of abalones (marine snails) based on physical measurements, specifically using attributes such as sex, length, diameter, height, whole weight, shucked weight, viscera weight, and shell weight. These physical characteristics are critical for understanding the growth and development of abalones, which can aid in fisheries management and ecological studies. Traditionally, the age is estimated by counting the rings on the shell after dissection—a labor-intensive and time-consuming process. The ultimate goal is to build a model that can estimate age from easily obtainable physical characteristics.

The data set is obtained from: https://archive.ics.uci.edu/dataset/1/abalone.

Data were collected through a study in which abalones were harvested and a series of physical measurements were recorded. The data set includes 4177 samples, each with eight attributes related to the physical characteristics of the abalone, as well as the class label that indicates its age. The age of the abalone is not directly measured, but inferred from the number of rings on its shell, which can be seen through dissection.

The data set provides a valuable resource for modeling the relationship between the physical characteristics of abalones and their age. The age of an abalone is determined by slicing its shell through the cone, staining it, and counting the rings under a microscope, a tedious and time-consuming process. Instead, more easily obtainable measurements can be used and, by analyzing these characteristics, we aim to predict the age of abalones.

Given the range of characteristics (from physical measurements to weight properties), the challenge is to determine which characteristics are the most important in predicting age and how well these characteristics correlate with each other.

For regression, we expect to predict the age of abalones (Rings + 1.5 years) based on attributes such as Length, Diameter, Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight, as these physical measurements are directly related to growth. For classification, we can categorize abalones into age groups for example,(Young: 1-8 rings, Adult: 9-10 rings, Old: 11+ rings) using the same attributes. Additionally, we can classify the sex of abalones (Male, Female, Infant) based on physical measurements, including Length, Diameter, Height, and weight attributes, to explore potential correlations between physical characteristics and sex.

3 Explanation of attributes

Missing values from the original dataset were removed to ensure data completeness. Furthermore, all measurement values from the continuous attributes were scaled by dividing by 200, allowing them to be approximately within the range [-1, 1] for better consistency and compatibility with machine learning models.

Table 2 shows type and description of the attributes from the dataset as they were. We however one-hot encoded the 'Sex' attribute, to make sure that one 'sex' would be weighed higher with a value of 2, compared to 1 and 0. This is also how the data is visualized in figure 1 and 2.

Attribute	**Type**	**Description**
Sex, *sex*	Discrete, Nominal	Sex of the abalone (M = Male , F = Female, I = Infant)
Length (L), *len*	Continuous, Ratio	Length of the abalone (mm)
Diameter (D), *dia*	Continuous, Ratio	Diameter of the abalone (mm)
Height (H), *height*	Continuous, Ratio	Height of the abalone (mm)
Whole Weight (WW), *ww*	Continuous, Ratio	Whole weight of the abalone (grams)
Shucked Weight (SW), *shw*	Continuous, Ratio	Weight of the abalone's meat (grams)
Viscera Weight (VW), *vw*	Continuous, Ratio	Weight of the abalone's gut (grams)
Shell Weight (SW), *sw*	Continuous, Ratio	Weight of the abalone's shell (grams)
Class label		
Age of Abalone, *age*	Discrete, Ratio	Age of the abalone (in years), calculated as the number of rings on the shell

Table 2: Description of Attributes/Class

Summary statistics from the dataset is shown in table 3 and visualized in figure 2

Attribute	Min	Max	Mean	SD	25%	Median	75%
Length	0.075	0.815	0.524	0.120	0.450	0.545	0.615
Diameter	0.055	0.650	0.408	0.099	0.350	0.425	0.480
Height	0.000	1.130	0.140	0.042	0.115	0.140	0.165
Whole Weight	0.002	2.826	0.829	0.490	0.442	0.799	1.153
Shucked Weight	0.001	1.488	0.359	0.222	0.186	0.336	0.502
Viscera Weight	0.001	0.760	0.181	0.110	0.094	0.171	0.253
Shell Weight	0.002	1.005	0.239	0.139	0.130	0.234	0.329

Table 3: Statistics for Numeric Attributes

4 Data visualization

4.1 The histogram

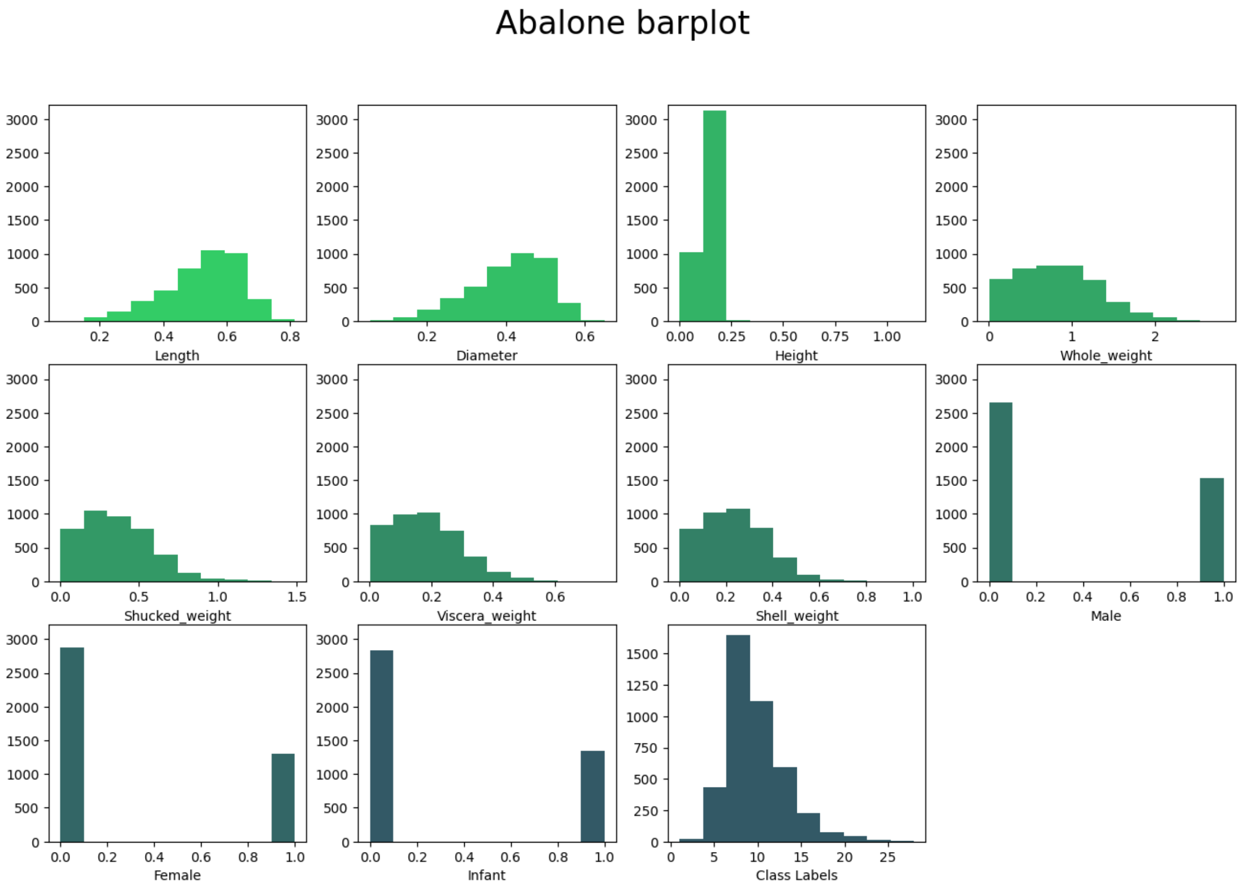


Figure 1: Barplot (histogram) of the attributes

The histogram in figure 1 provides an overview of the distribution of various attributes of Abalone's dataset represent, respectively. The distribution of the attributes in figure 1 depending on their type. The sex attributes are categorical and encoded as Female = 1, Otherwise 0, Male = 1, Otherwise 0, and Infant = 1, Otherwise 0. The distribution of sex categories is imbalanced due to the category chosen, as the only possible values are 1 or 0. Related attributes such as Length, Diameter, and Height are approximately normally distributed, which means that most abalones fall within a medium range, while very small or very large ones are less frequent. However, the Height attribute contains some outliers. The weight-related attributes (Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight) tend to be right-skewed, indicating that most abalones are relatively light, while a few significantly heavier ones create a long tail in the distribution. The Rings attribute, which serves as the target variable for predicting age, follows an approximately normal distribution but is slightly skewed to the right, as younger abalones are more common than older ones. Since the age of the abalones is calculated as Rings + 1.5 years, the data set mainly contains young and middle-aged abalones, with fewer older ones. In general, while some attributes exhibit normal distributions, others exhibit skewness or imbalances.

4.2 Boxplot diagram

The boxplot diagram 2 provides a visual representation of the distribution of attributes in the Abalone dataset, helping to identify he center, range, and outliers in the data. Each box represents the interquartile range (IQR), which contains the middle 50% of the data, with the median marked inside. The whiskers extend from the box to represent the range of data, excluding outliers, while any points beyond **1.5 times the IQR** are displayed as black circles, indicating outliers.
From the table 3 and figure 2, the median of the attribute's Height is 0.140, while its max value is 1.130. The IQR is calculated as:

$$IQR = q_{75} - q_{25}$$

which for Height is: 0.165 - 0.115 = 0.050. The upper bound is given by:

$$q_{75} + 1.5 \times IQR$$

resulting in:

$$0.165 + (1.5 \times 0.050) = 0.240$$

Since 1.130 is much higher than 0.240, this confirms that Height has extreme outliers, as seen in the boxplot2.. Similarly, Weight-related attributes such as Whole, Shucked, Viscera, and Shell weights exhibit numerous black circles, indicating that their maximum values exceed 1.5 × IQR, confirming them as outliers. For example, the maximum Whole Weight is 2.826, while the median is 0.799, highlighting a large range and reinforcing the presence of outliers. The lower bound using:

$$q_{25} - 1.5 \times IQR$$

, which applies to Length, Diameter, and Height, for detecting outliers.

The outliers in the dataset appear primarily in Height, Whole Weight, Shucked Weight, Viscera Weight, and Shell Weight, whereas extreme values, which are even farther from the IQR threshold, are especially evident in Height. The boxplot 2 and statistics table 3 confirm that multiple attributes exhibit significant outliers and extreme values, particularly in Height and Weight-related features, making the dataset skewed. On the other hand, Length and Diameter have a more even spread with fewer outliers, suggesting a more normal distribution.

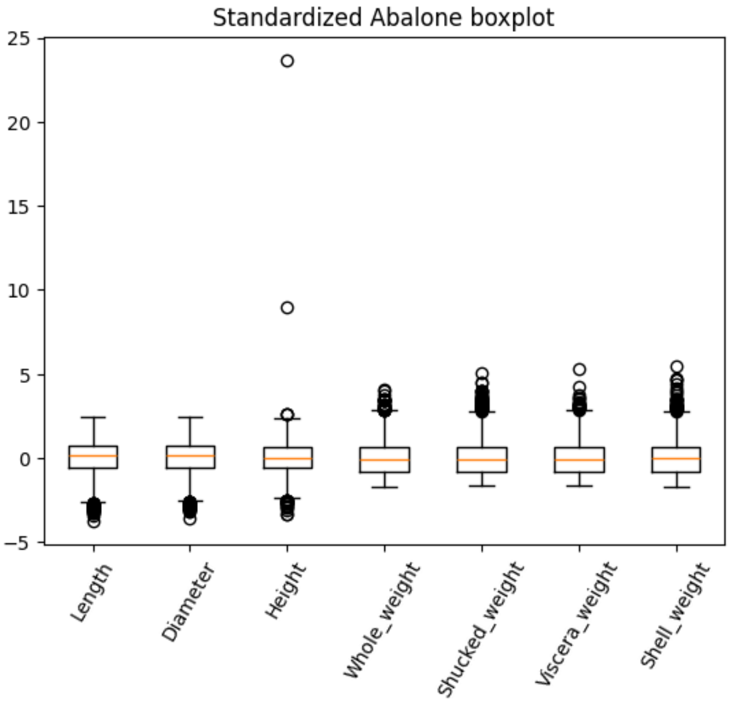


Figure 2: Boxplot of the attributes

4.3 Correlation between attributes

In figure 3 the correlation between attributes can be observed. The class label is represented on a color map using the clearer red for young abalones and the darker red for older ones. As is expected, the correlation of the attributes against themselves result in a diagonal line. Others are directly proportional for most samples, such as the length and diameter of the abalones. Meanwhile, others such as the length and weight follow an exponential correlation. Another aspect which deserves to be highlighted is that the 2 outliers samples are present in most of the graphs, but they are particularly visible in the height graphs.

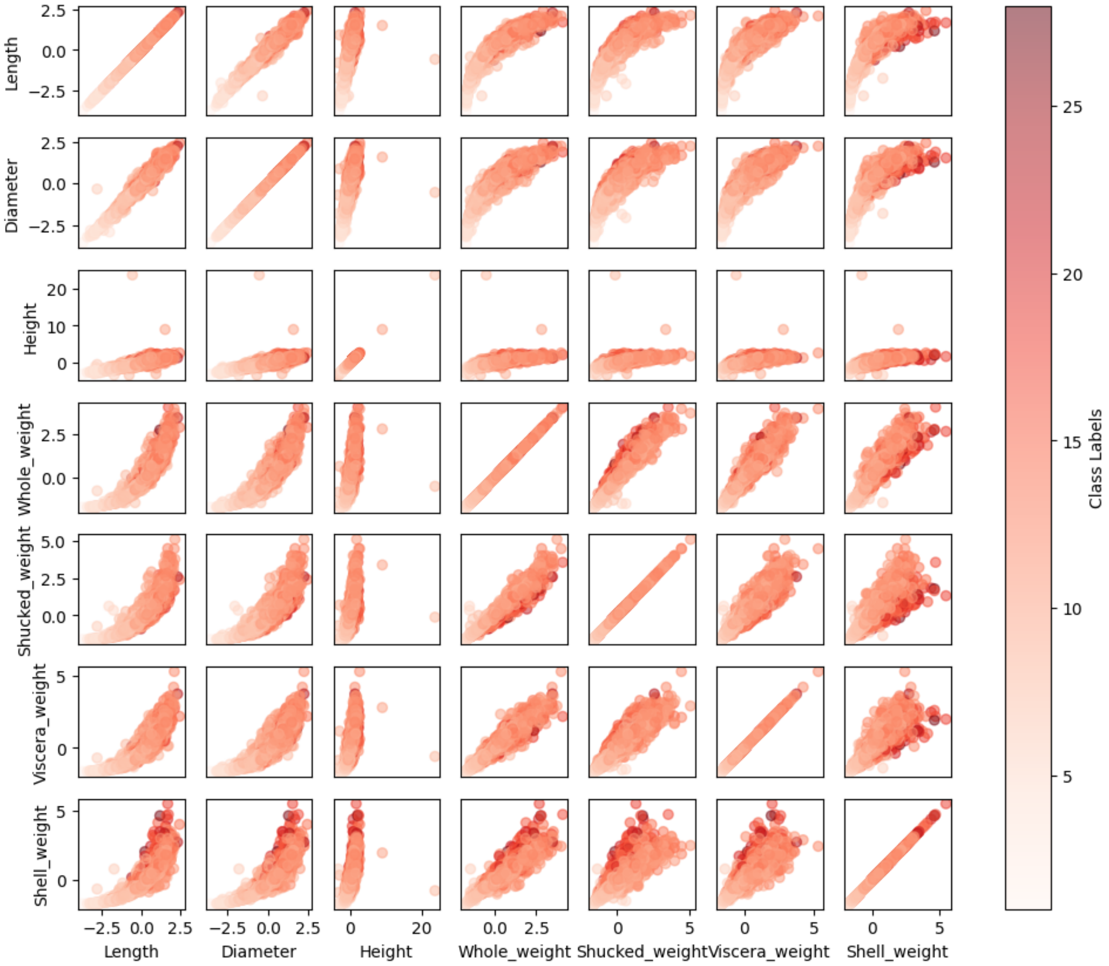


Figure 3: Correlation between attributes

5 Principal Component Analysis

In order to extract further information from the data, a Principal Component Analysis (PCA) was performed. This analysis will be useful for creating a model to classify new data points using these components in a later stage.

In figure 4 the variance associated with each component and the cumulative variance can be observed. To reach the desired threshold of cumulative variance 90 %, the first three components would be enough.

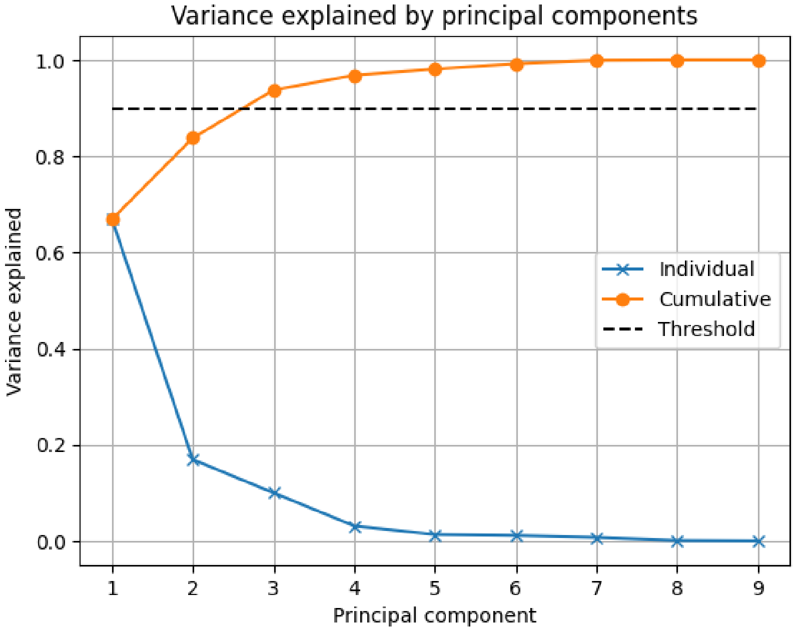
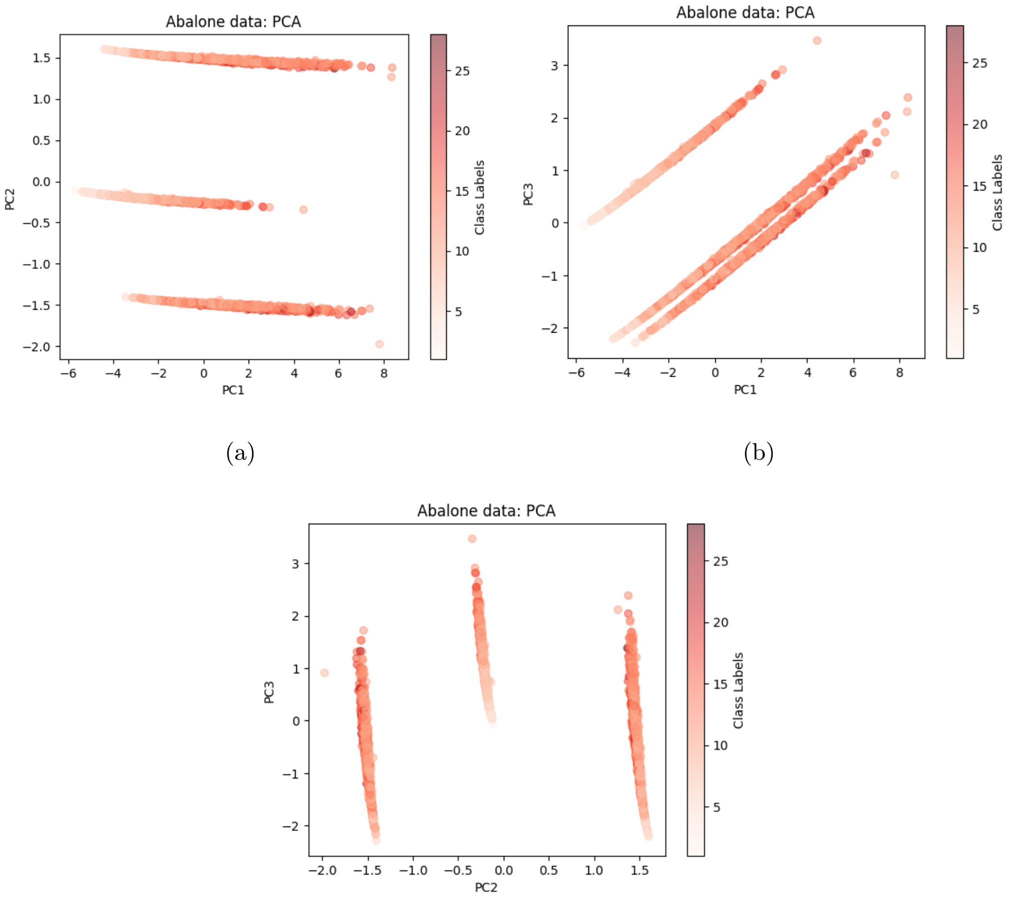


Figure 4: Variance explained by principal components

To further analyze the main 3 principal components(PC), figure 5 shows the three components against each other. PC2 is almost constant against PC1 5(a). PC3 and PC1 are directly proportional as shown in figure 5(b). And because of this, PC2 is also almost constant over PC3, as can be observed in figure 5(c). In each graph of the figure, 3 different populations can be observed. These are directly related to sex attributes. It can be seen that in PC1 and PC3 the class label (age) increases as the component increases but for PC2 which remains quite constant.



(a)

(b)

(c)

Figure 5: Main principal component 1, 2 and 3 against each other.

6 Discussion and Conclusion

In this report, the Abalone dataset has been analyzed to use it later to predict the age based on physical measures such as length, diameter, or weight. For that, the data has been preprocessed to remove missing values, and the attributes have been scaled for consistency.

To ease the analysis, the different data attributes have been represented in a histogram 1 and boxplot 2 to explore the shape of the distribution over all available samples. Some bar plots have also been computed to visualize the statistical properties of the data.

The results show that some attributes, particularly weight-related features, exhibit skewed distributions and significant outliers, which may require transformations in future modeling. The correlation analysis indicates strong relationships between certain attributes, supporting their relevance for predictive modeling.

Furthermore, the correlation 3 between different attributes has been explored, showing a strong correlation between them and age, which will aid in age prediction once a model is based on the data.

In addition, a principal component analysis 4, 5 was performed to reduce dimensionality showing that the first three principal components account for more than a 90 % of variability. Also, the analysis of the principal components against each other revealed distinct clusters related to the sex of abalones with age increasing along certain components.

In the next phase of the project, we will apply regression models to predict the age of abalones based on physical attributes and experiment with classification models to categorize them into age groups. Additionally, we will investigate whether sex classification can be effectively performed based on available measurements. The findings from this initial analysis provide a comprehensive understanding of the dataset, laying the groundwork for these predictive modeling tasks.