DTU ML Project 1

Jai Murhekar s237013
Nicolai Pavliuc s240366
Alice Coimbra s237108

February 2024

	Section 1	Section 2	Section 3	Section 4	Section 5
Jai (s237013)	10%	10%	80%	20%	20%
Nicolai (s240366)	80%	10%	10%	30%	50%
Alice (s237108)	10%	80%	10%	50%	30%

Table 1: Contributions

1 Description of the Dataset

The selected dataset is called Abalone. It contains diverse physical measurements of abalones - marine mollusks belonging to the family Haliotidae.

The overall problem of interest is predicting the age of an abalone by analyzing its physical measurements. This method was introduced since the 'classical' approach of determining the age is both difficult and time-consuming. It involves cutting the shell through the cone, staining it, and counting the number of rings through a microscope.

The dataset is based on real-life measurements and the authors of the dataset did some data manipulation. Firstly, they removed all measurements (rows) that had missing values (the resulting dataset has 0 missing values). Secondly, a linear transformation has been done to the continuous attributes, all values have been divided by 200. A description of the attributes is shown in Table 2. There are a total of 4176 records in the dataset.

The main classification problem is predicting the number of rings. Since the original dataset contains 29 discrete values (1-29) for the rings attribute, for convenience it is possible to introduce a new attribute 'Rings_group' that groups them into 4 intervals as shown in Table 3.

Another classification problem can be predicting the sex (nominal attribute) based on physical measurements of the abalone.

An example of regression problem is, just like in the first classification problem, predicting the age of an abalone. The rings attribute would have to be

Table 2: Attributes of Abalone Dataset

Name	Description	Type 1	Type 2
Sex	M, F, and I (infant)	Discr.	Nominal
Length	Longest shell measurement	Cont.	Ratio
Diameter	Perpendicular to length	Cont.	Ratio
Height	With meat in shell	Cont.	Ratio
Whole_weight	Whole abalone	Cont.	Ratio
Shucked_weight	Weight of meat	Cont.	Ratio
Viscera_weight	Gut weight	Cont.	Ratio
Shell_weight	After being dried	Cont.	Ratio
Rings	+1.5 = Age	Discr.	Ratio

Table 3: Possible values and intervals for 'Rings_group' attribute. Intervals were picked based on 25th, 50th, and 75th percentile for Rings attribute

Value	0	1	2	3
Interval	[0,8]	(8,9]	(9,11]	(11,29]

converted into a continuous type. Physical attributes can be used as input attributes for this regression.

Another example of regression problem is determining the diameter of abalone based on its length. These 2 attributes have high correlation as shown later in the report.

Before proceeding to this data modeling, some additional data preparation needs to be done. The sex attribute, for example, needs to be converted into numeric values using one-of-K encoding. In addition to that, it is necessary to deal with outliers and standardize data.

2 A detailed explanation of the attributes of the data

Table 4 and Table 5 show the summary statistics of all the attributes, except the attribute sex. Since this attribute is nominal this type of analysis is not adequate. A total of 4176 observations were made. From these summary tables we can see that the range of values for most of the continuous attributes are in the same scale. An exception to this is the rings attribute, whose values are around 10 times bigger than the rest of the attributes. For this reason, it might be relevant to standardize all the values before performing more in depth analysis, such as PCA, so that the results can be comparable.

In order to further analyse the attributes, plots were made to examine the distribution of values for each of the attributes. These plots can be seen in Figure 2.1. For all the attributes except the attribute "sex", the data is displayed in histograms, whereas the "sex" attribute is represented as a bar chart. From this

Table 4: Summary Statistics of attributes of Abalone Dataset (1/2)

	length	diameter	height	whole weight	shucked weight
mean	0.524	0.408	0.140	0.828	0.359
std	0.120	0.099	0.042	0.490	0.222
min	0.075	0.055	0.000	0.002	0.001
25%	0.450	0.350	0.115	0.442	0.186
50%	0.545	0.425	0.140	0.800	0.336
75%	0.615	0.480	0.165	1.153	0.502
max	0.815	0.650	1.130	2.826	1.488

Table 5: Summary Statistics of attributes of Abalone Dataset (2/2)

	viscera weight	shell weight	rings
mean	0.180	0.239	9.933
std	0.110	0.139	3.224
min	0.001	0.002	1.000
25%	0.093	0.130	8.000
50%	0.171	0.234	9.000
75%	0.253	0.329	11.000
max	0.760	1.005	29.000

graph it can be seen that the distribution across the three categories (Female, Infant, Male) is quite balanced, with there being slightly more observations for the Male category. As for the continuous attributes, both the attribute length and diameter seem to follow a normal distribution. In the case of the weight attributes (whole weight, shucked weight, viscera weight and shell weight), all four distributions appear to be skewed to the left, meaning there are few observations with high values for these attributes. As for the height, the values are not so disperse, concentrating around the 0.1 to 0.2 range. This is in accordance with the statistical analysis previously performed, where we can see that the standard deviation for the height attribute is relatively low.

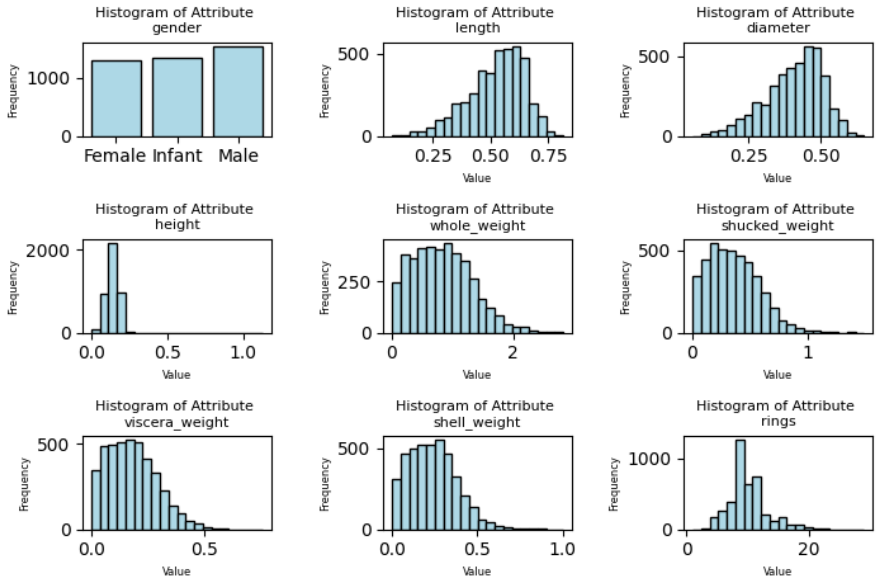


Figure 2.1: Distribution of values for each attribute

3 Data Visualization and Principal Component Analysis (s237013)

We aim to understand our dataset by visualizing it effectively and applying Principal Component Analysis (PCA). Initially, we check for outliers by plotting scatter plots of attribute pairs using the Seaborn library. For example, in Fig 3.1 (Diameter vs Height), we spot two outliers. Nevertheless, since the number of outliers are very few, we can ignore them in our analysis. Next, we assess the distribution of our variables by plotting histograms for each column. We see that the variables generally follow a normal distribution, although there is some skew (example 3.2 which is the histogram for the Diameter feature). Additionally, we observe strong correlations among variables, as depicted in Fig 3.3 (heatmap of the correlation matrix for normalized data). Since the data is more or less normally distributed and largely free from outliers, predicting the age/number of rings appears feasible.

After visualizing the data, we proceed with Principal Component Analysis (PCA). Before that, we standardize the feature variables to ensure that they have zero mean and unit variance. This is done by subtracting the mean from the data and dividing by the standard deviation. Then, we compute the singular value decomposition of the normalized data using the scipy.linalg library. We also analyze the amount of variance explained by the principal components (Fig 3.4). Finally, we visually represent the variance and principal components to

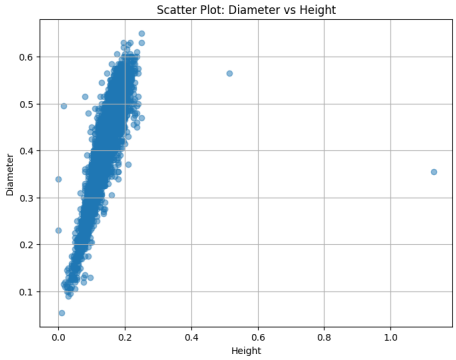


Figure 3.1: Diameter vs Height

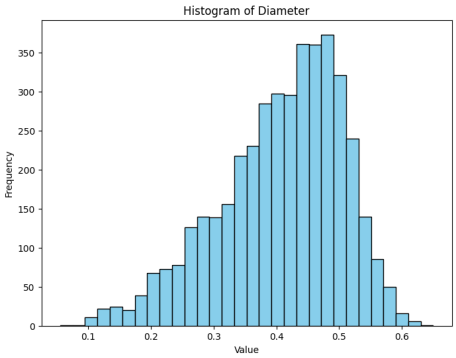


Figure 3.2: Histogram of Diameter

gain insights into the underlying patterns of the data.

3.1 Interpretation of the PCA

From Fig 3.4, we observe that the first principal component explains nearly 90% (specifically 91%) of the dataset's variance. In contrast, the second principal component only contributes about 4% of the total variance. This suggests that a single principal component suffices to capture the majority of the data's variability, which is not surprising considering the high correlation among features. From this, we see that PCA is highly effective in reducing the dimensionality of the data. Fig 3.5 gives the contribution of each feature to the principal component. Each feature makes a nearly equal contribution (in the range 0.35-0.4), as depicted in the plot.

Finally, Figures 3.6 and 3.7 show the projection of the data using one and two Principal Components respectively. From Fig 3.6, we see that the data does not easily separate into classes. However it is clear from both Fig 3.6 and Fig 3.7 that older abalones (with more number of rings) have higher values of the PC1 projection.

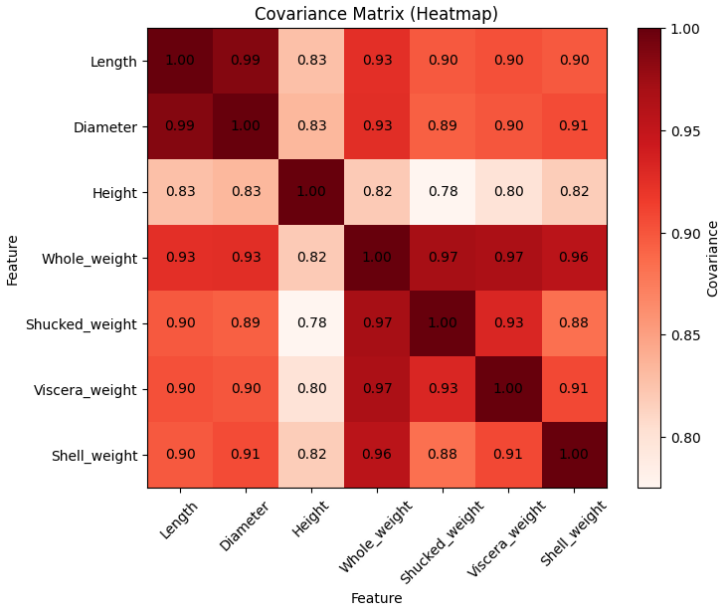


Figure 3.3: Covariance Matrix (Heatmap)

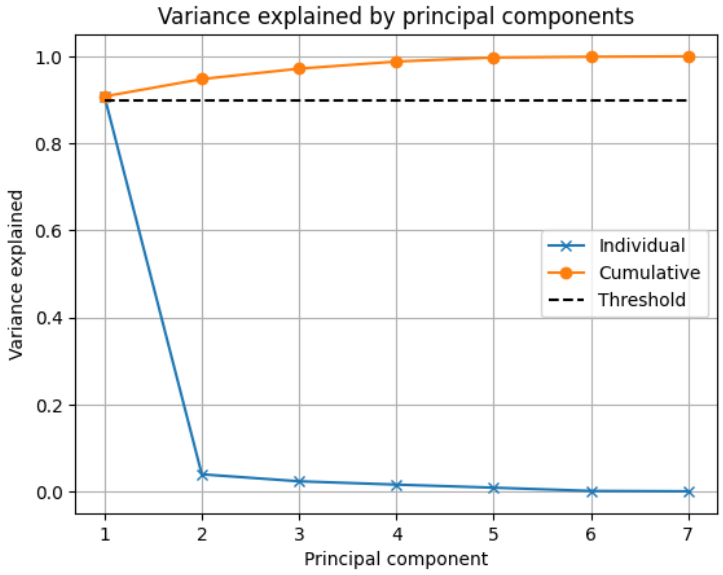


Figure 3.4: Variance explained by principal components

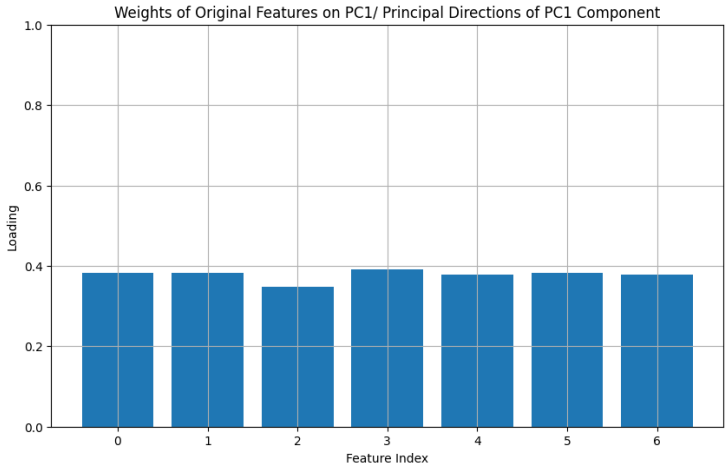


Figure 3.5: Weights of Original Features on PC1

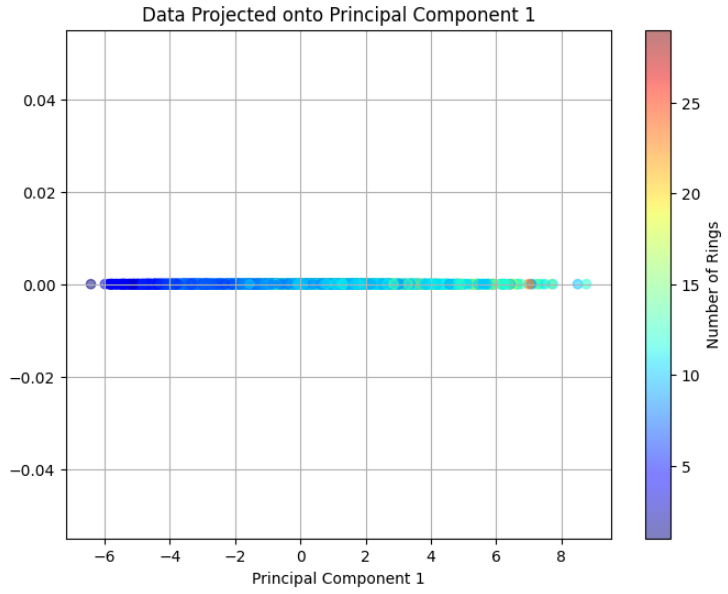


Figure 3.6: Projection on PC1

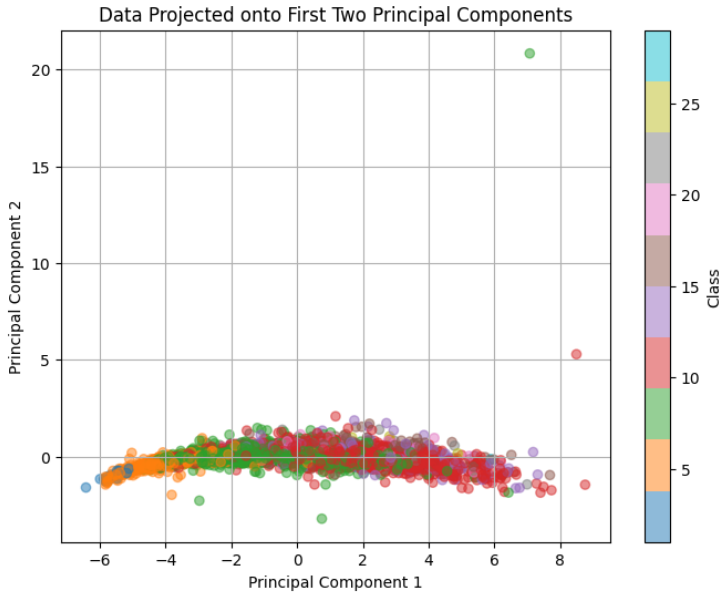


Figure 3.7: Projection on PC1 and PC2

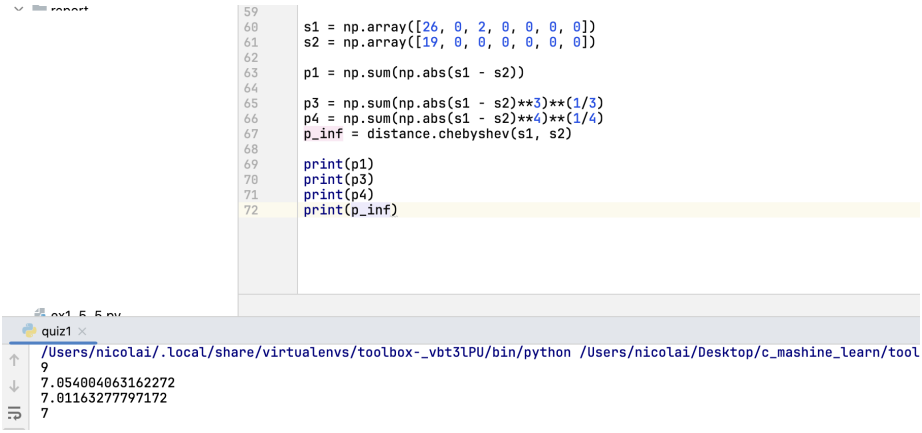
4 A discussion explaining what you have learned about the data

We have learned that the attributes are highly correlated, which is a benefit for making a good model. We saw that the first principal component already explains more than 90% of the dataset's variance. However, even though not a lot of variance is lost when projecting into either the first or the first two principal components, a visual classification is quite hard to perform, from which we might conclude that PCA is most likely not the best tool to perform classification.

Overall, the choice of the dataset was a successful one. Apart from the presence of some outliers and lack of normalization, it is of a high quality. Attributes are highly understandable and relevant. The classification and regression problems to be tackled later should be very straightforward.

5 Exercises

Question 2. Spring 2019 question 2:
The correct answer is option A ($d_{p=\infty}(x_{14}, x_{18}) = 7.0$) as seen below from the code




```
59
60  s1 = np.array([26, 0, 2, 0, 0, 0, 0])
61  s2 = np.array([19, 0, 0, 0, 0, 0, 0])
62
63  p1 = np.sum(np.abs(s1 - s2))
64
65  p3 = np.sum(np.abs(s1 - s2)**3)**(1/3)
66  p4 = np.sum(np.abs(s1 - s2)**4)**(1/4)
67  p_inf = distance.chebyshev(s1, s2)
68
69  print(p1)
70  print(p3)
71  print(p4)
72  print(p_inf)
```

Question 4. Spring 2019 question 4:
The correct answer is option D. The reasoning for this as follows :

$$v_2 = \begin{bmatrix} -0.5 \\ 0.23 \\ 0.23 \\ 0.09 \\ 0.08 \end{bmatrix} \qquad \begin{bmatrix} - \\ + \\ + \\ + \\ + \end{bmatrix} \rightarrow \begin{bmatrix} + \\ + \\ + \\ \sim 0 \\ + \end{bmatrix}$$

Given that Time of day has a negative coefficient of PCA2 and that the value for this attribute is low (negative), this will contribute with a positive value for the projection. On the other hand, Broken Truck, Accident victims and defects all have positive coefficients of PCA2, while the values are also positive (high values). Thus, these attributes will also contribute with a positive value for the projection. The remaining attribute, Immobilized bus, has a coefficient that is close to zero, so it's contribution for the projection will be minimal. Therefore, the overall projection onto PCA2 will most likely be a positive result.

Question 5. Spring 2019 question 14:
The correct answer is option A. The calculations are shown below :
$f_{11} = 1+1 = 2, M = 20,000($ words in vocabulary$), f_{11}+f_{10}+f_{01} = M - f_{00} =$ 13 Hence,

$$\hat{j}(s_1, s_2) = \frac{2}{13} = 0.015386$$

Question 6. Spring 2019 question 27: The correct answer is option B.

The calculation is shown below

$$P(A|C) = \sum_{i=1}^{n} P(AB_i|C)$$

Thus,

$$\begin{aligned} P(\hat{x_2} = 0|y = z) &= P(\hat{x_2} = 0, \hat{x_7} = 0|y = z) + P(\hat{x_2} = 0, \hat{x_7} = 1|y = z) \\ &= 0.8 + 0.04 = 0.84 \end{aligned}$$

6 References

Data set - https://archive.ics.uci.edu/dataset/1/abalone