Danmarks Tekniske Universitet

02450 Introduction to Machine Learning and Data Mining Project 1

Authors
Aleksandar Lukic - (AL) - s194066
Marco Andreis - (MA) - s243116
Martina Bellini - (MB) - s243118

Lecturer

Georgios Arvanitidis

Section No. & Title	AL	MA	MB
1 Description of data set	30%	40%	30%
2 Detailed explanation of data set	20%	50%	40%
3 PCA analysis	20%	30%	50%
4 Discussion	60%	20%	20%
5 Exam problem solutions	33%	33%	33%

Contribution % to report sections per group member.

October 3rd, 2024 at 17:00

Summary

1 Description of data set

1.1 Problem of interest

The dataset contains physical measurements of the shell and meat of specimens of *Haliotis rubra*, a species of marine snails commonly known as *blacklip abalone*. The original data was collected in a 1994 technical report of the Sea Fisheries Division of the Marine Research Laboratories of Taroona. It aimed to assess the effect of the growing fishing practices on stunting the abalone population of the north coast of Tasmania and islands of the Bass Strait [1].
Following the original publication, the gathered data was utilised by Samuel George Waugh in his PhD thesis (1995) at the University of Tasmania. In his work, the dataset was used as a real-life example for the benchmarking of Cascade-Correlation feed-forward artificial neural networks in the task of classifying the abalones' age [2].

After his thesis, S. Waugh donated the dataset to the UCI Machine Learning Repository [3], the source from which it was obtained for this project.

1.2 Summary of study analysis and results

The 1994 study does not have a machine learning approach, but the researchers did thoroughly investigate the relation between an abalone's physical features and their age, sexual maturation and fecundity.

The data were collected in 1988 in five different locations, recording physical measurements of the shell and meat. Additionally, the ages of the abalones were determined using a method that counts the growth rings on the shell, based on a relationship of one ring per year (to which are added 1.5 years to account for initial development). Sexual maturation and fecundity were obtained through inspection of the gonad and its cross-sectional area.

Initially, the researchers studied the correlation between the abalone's morphometrics to quantify the variability of length-weight-shape relations across sites. Then, they focused on how these findings related to the abalones' ages, onset of sexual maturity and fecundity. The results were finally used to investigate the yield-per-recruit and egg-per-recruit to relate how an abalone characteristics could affect its probability of being captured, thus providing valuable insights regarding depletion of fishing stocks and the growth stunting of the abalone population.

Afterwards, in his thesis, Waugh introduced and examined different extensions to the Cascade-Correlation algorithm, such as patience and alternative training methods, for building neural networks with increased performance in classification tasks. In this context, the abalone dataset was used as a real-life example, as the process to determine the age by counting the rings of a specimen is tedious and time-consuming. Consequently, neural networks were promising as a tool capable of predicting it from physical measurements, that can be more easily collected.

The data were processed to make them suitable as input for a neural network through scaling and selection of the attributes. As the age attribute assumes discrete values (from 1 to 29 rings), the problem was approached as a 29-class classification task. Additionally, models were also trained in two variations of this approach, either by conserving all data points, but reducing complexity by grouping the examples into three more general classes (abalone with 1-8, 9, or 10 and more than 10 rings), or by focusing instead on classifying only the classes containing the largest amount of abalones (8-9-10-11 rings).

Results were not satisfactory in terms of performance. Outside of determining the importance of data scaling, the prediction accuracy was acceptable only in the grouped variation of the task. This was mainly attributed to the significant classes overlapping due to the different initial purpose of the dataset, which made it not particularly suitable for classification.

1.3 Expectations and predictions

The dataset contains seven continuous variables, being different physical measurements of the abalones, which are all suitable to be analysed with regression, with the aim of finding if there are any linear relations occurring between them. In the original work on the dataset, they did in fact investigate this aspect, recognising from the scatter plots that simple linear regression was not fitting for all attribute combinations. In these cases, they fitted the data to the log-transformation of the power function:

$$y = a \cdot x^b$$
$$\log(y) = \log(a) + b \cdot \log(x)$$

We intend to do the same and then compare the results with those obtained by the authors.

In the classification task, two potential attributes can be used as targets: sex and the number of rings. However, we anticipate that sex would be challenging to analyse, as the three sex classes (male, female, and infant) differ only slightly in terms of each other's attribute.

Focusing on the number of rings appears to be a more promising approach. Considering the results of Waugh, we will mainly focus on the classification of abalones in the three age classes.

2 Detailed explanation of data set

2.1 Description of attributes

The dataset consists of 4177 samples, each with 9 attributes. The only non-numerical variable is *Sex*, which can assume three values, being female (F), male (M) or infant/immature (I).

Name	Data type	Measurement	Description
Sex	nominal	M, F, I	the sex of the abalone
Length	continuous / ratio	mm	longest shell measurement
Diameter	continuous / ratio	mm	perpendicular to length
Height	continuous / ratio	mm	with meat in shell
Whole weight	continuous / ratio	grams	whole abalone
Shucked weight	continuous / ratio	grams	weight of meat
Viscera weight	continuous / ratio	grams	gut weight (after bleeding)
Shell weight	continuous / ratio	grams	weight of the dehydrated shell
Rings	discrete / ratio	integer	number of rings on the shell

Table 1: *Summary of the attributes*

The number of rings can be directly use to calculate the age of the abalone, as described in the original paper.

$$\text{abalone's age} = \text{no. of rings} + 1.5$$

2.2 Data issues

The available data was provided by Waugh, therefore the dataset was already processed to be suitable as input to a neural network. The processing conducted included the removal of all entries having at least one missing attribute value, reducing the number of examples from 8233 to 4177. Also, the information about the abalone collection site and fecundity are not provided, as they were not utilised in the PhD thesis.
Another consequence is that all continuous attributes have been scaled by dividing by 200. This was purely done because neural networks perform better when trained on numbers in small ranges. However, this scaling does not serve the aim of this project, as it would impair the graphs of our analysis by making the values more difficult to interpret. Thus, we decided to revers it.
Scaling will be applied subsequently when needed for specific analysis. Some data were provided with an unexplained number of significant figures, probably due to the handling of the data across different databases. During the cleaning of the dataset, we standardised these values at the first decimal digit.

2.3 Data exploration

Understanding the basic distribution of the variables is essential for deciding how the future analysis should be carried out.

From Table 2, we can see that standard deviations tend to be really significant. This might be due to the fact that the dataset contains data about abalones gathered from different locations. The median and upper extrema for *Height* and all the weight measurements, show the presence of outliers with significantly higher values, compared to the rest of population. The means being only slightly bigger than the medians further suggests, that these outliers are limited in numbers. Additionally, *Height* has a minimum value equal to 0.0 - possibly

an error in data collection that was not detected by Waugh. On account of the above, the entry was removed from the dataset.

	Length	Diameter	Height	Whole weight	Shucked weight	Viscera weight	Shell weight	Rings
mean	104.8	81.6	27.9	165.7	71.9	36.1	47.8	9.9
std	24.0	19.8	8.4	98.1	44.4	21.9	27.8	3.2
min	15.0	11.0	0.0	0.4	0.2	0.1	0.3	1.0
median	109.0	85.0	28.0	159.9	67.2	34.2	46.8	9.0
max	163.0	130.0	226.0	565.1	297.6	152.0	201.0	29.0

Table 2: *Summary statistics of the dataset*

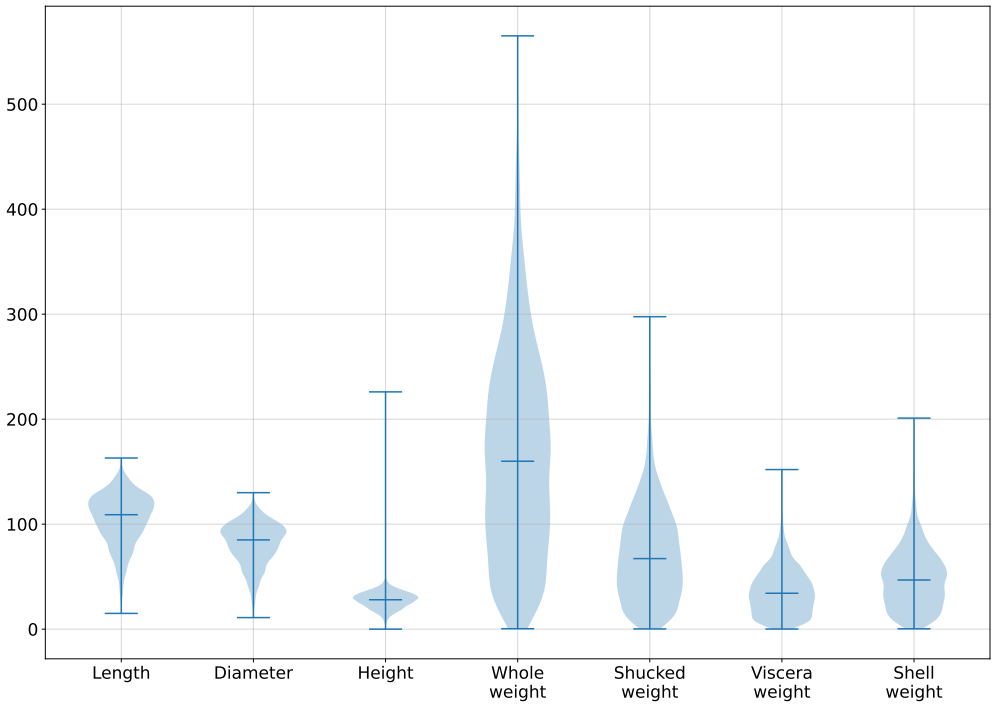

Figure 1: *Distribution of physical measurements: median and extrema displayed*

A similar conclusion can be drawn by observing Figure 1, which provides a clear visualisation of the data distributions. The values of the shell height, and all the weight measures, lie within a wide range compared to their standard deviations. This is particularly evident for the attribute *Height*, where the points are clustered around the mean, supporting the hypothesis of outliers. Additionally, all the attributes appear to be normally distributed.

For deciding which analysis to perform on the dataset, we chose to stratify the physical measurements on *Sex*.

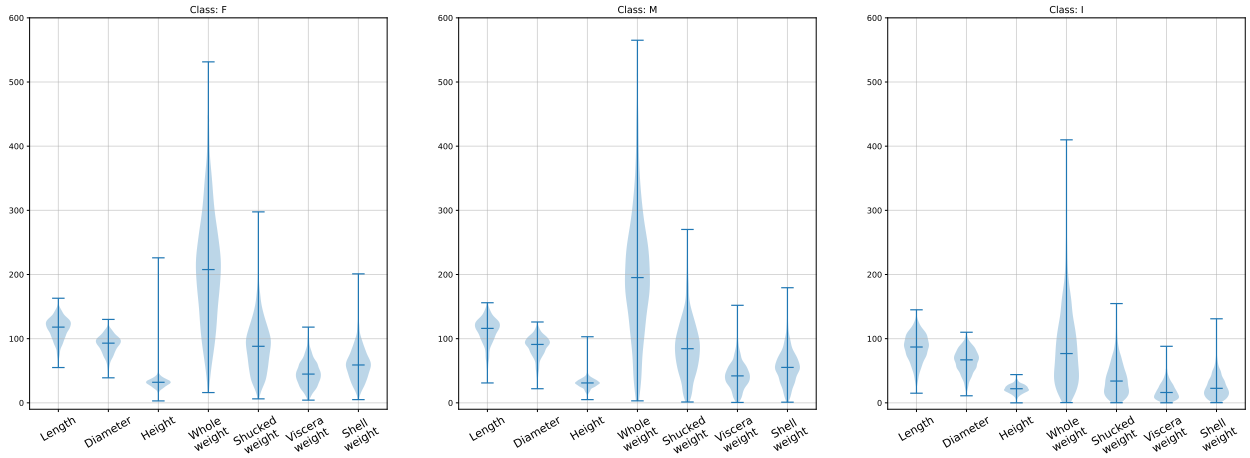

Figure 2: *Distribution of physical measurements stratified on sex.*

From Figure 2, it is immediately clear how minimal the differences between males (M) and females (F) are. This highlights how a classification on *Sex* would likely not yield good results. For this reason, we tried with a different stratification - based on the number of rings. Abalones were split in three groups: those with 1-8, 9-10 and more than 10 rings, as done by Waugh. The stratification on rings can be seen in Figure 3.

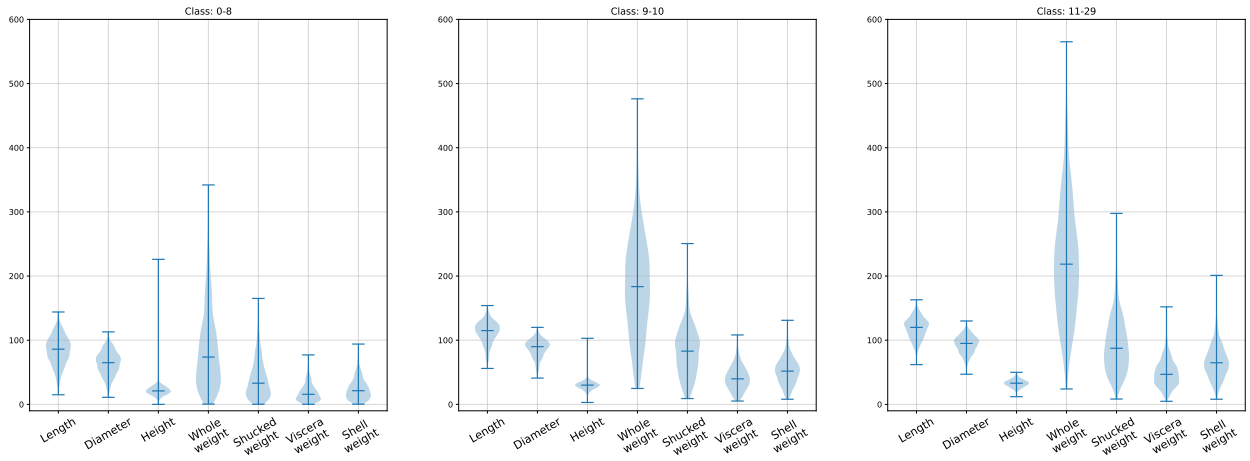

Figure 3: *Distribution of physical measurements stratified on number of rings.*

The second stratification didn't provide a significant improvement, as expected from the unsatisfactory outcome of the Waugh classification. Therefore, we moved on and analysed the correlation between the single attributes, as shown in Figure 4.
The correlation between *Length* and *Diameter* observed in the matrix plot has been confirmed by the calculation, obtaining a Pearson correlation value of 0.99. Another interesting relationship discovered was between *Whole weight* and *Length*. Just like in the original 1994 research, they were found to be in an exponential relationship.

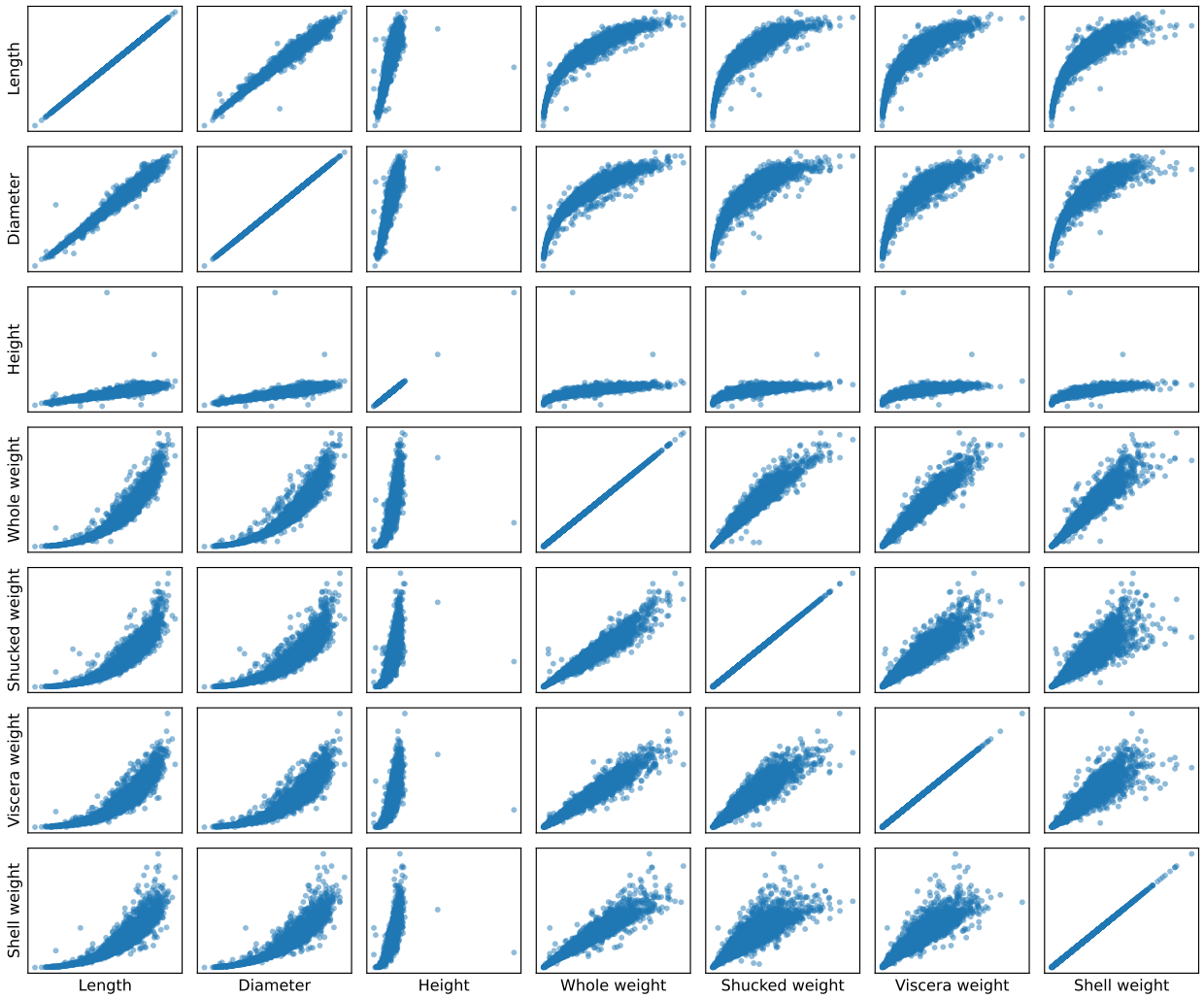

Figure 4: *Matrix representation of 1:1 relations between attributes*

Figure 5 shows the two attributes plotted on a logarithmic scale. After applying the logarithm of the two attributes, the Pearson correlation coefficient was calculated to be 0.98. We believe it is not interesting to carry on further analysis on the correlation between the different weight attributes since they are dependent on each other.

Lastly, by examining the row of the *Height* attribute in Figure 4, we can clearly see two repetitive outlier data points. These support the hypothesis of outliers in the data, which corresponds to the samples removed earlier.

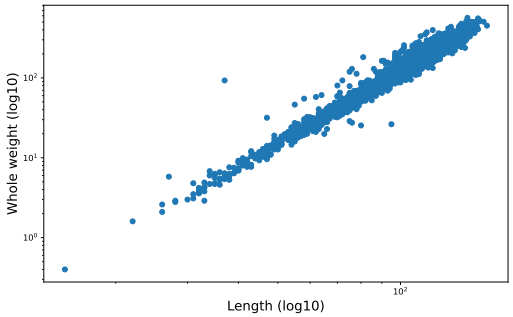

Figure 5: *Linear relationship between abalones' Length and Whole weight*

3 PCA analysis

For a further exploration of the data, a PCA analysis was performed. Observing the varying ranges of the attributes (especially *Whole weight*), the data was standardised.
Firstly, we analysed the amount of variance explained by each principal component. More than 90% of the variance is explained by the first principal component (PC1), as presented in Figure 6.

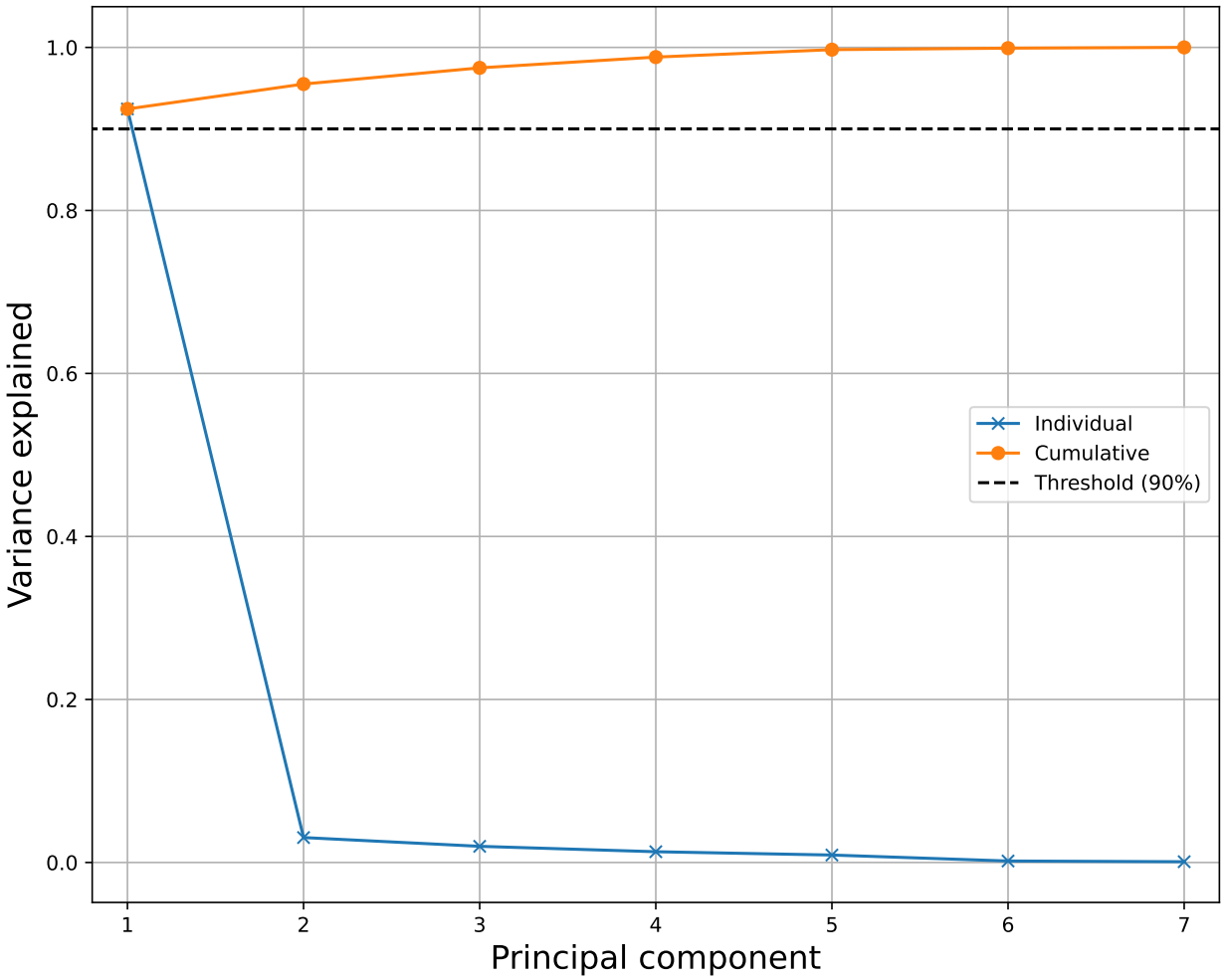

Figure 6: *Variance explained by each principal component*

Thus, we evaluated the contribution of the individual attributes to PC1. From this, we observed that all of the attributes provide approximately the same contribution to PC1. This supports that it is possible to explain almost the totality of the variance from a single component. Furthermore, this entails that all of the physical measurements are linearly dependent.
This is evident from calculating the contribution to PC1 of the attributes:

$$\text{PC1 contributes} = [0.380, 0.381, 0.367, 0.387, 0.375, 0.378, 0.376]^T$$

Considering these results, the data were projected only on PC1, emphasising the three age groups based on number of rings, as previously described in Figure 3. Again, it becomes evident, that it is not possible to separate the three groups. This is especially true for the groups representing a number of rings over 9, which are completely superimposed.

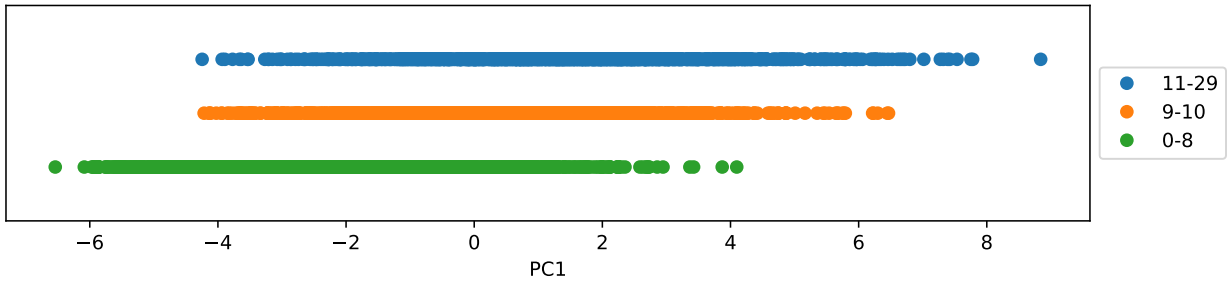

Figure 7: *Projection of data on PC1 (artificial vertical axis to avoid overlapping)*

4 Discussion

This project has revealed some interesting insights about the dataset. However, it has also highlighted some challenges. The data shows significant variability, especially in the *weight* attribute, and three datapoints where removed due to their *height*. This might be due to differences in collection locations, as suggested in the original research paper.

The machine learning aim of this project, was to create a classification rule for abalones, using PCA to find explained variances and correlation analysis. To visually showcase our findings, we used violin plots and a correlation matrix. From out visualisations, we were able to graphically explain flaws in the data, as well as the statistical findings.

Correlations between physical attributes were strong, particularly between *Length* and *Diameter* (PCC of 0.99) and between *Whole weight* and *Length* (0.98 on a logarithmic scale). However, the weight attributes were dependent on each other, making further correlation analysis unnecessary.

The PCA analysis showed that over 90% of the variance was captured by the first principal component, with all attributes contributing practically equally, indicating linear dependence. However, the projection of data onto PC1 did not separate the three age groups, especially for abalones with more than 9 rings, as these two groups completely overlapped. Thus, we were not able to define a classification based on sex nor age.

5 Exam problem solutions

5.1 Question 1: OPTION A

x1 (Time of day) is nominal since each observation corresponds to a 30-minute interval, specified by a numerical identifier, e.g. 1 = 07:00 - 07:30, 2 = 07:30 - 08:00, etc.. x2 (Broken Trucks) and x7 (Running over) are both quantifiable ratios, as we can have 0 broken trucks or as many as is unfortunate to occur. y (Congestion level) is a scale from low to high, which is based on the predicted level of congestion.

5.2 Question 2: OPTION A

The exercise can be solved with the help of python, where we find the absolute values of each vector and finding the distance between them as such:

```
import numpy as np
x_14 = np.array([26, 0, 2, 0, 0, 0, 0])
x_18 = np.array([19, 0, 0, 0, 0, 0, 0])
p_norm_dist = np.linalg.norm(x_14-x_18, ord = np.inf)
```

The p-norm distance for the two vectors when $p = infinity$: $d_{p=\infty}(x_{14}, x_{18}) = 7.0$

5.3 Question 3: OPTION A

We can apply the formula for explained variance: Variance Explained $= \frac{\sum_{i=1}^{n} \sigma_i^2}{\sum_{i=1}^{M} \sigma_i^2}$. From the decomposition, we can see each attributes standard deviation in the diagonal. These can then be parsed into the formula. For option A, $i = [1..5]$ while $j = [1..4]$:

$$\frac{\sum_{j=1}^{n} \sigma_j^2}{\sum_{i=1}^{M} \sigma_i^2} = \frac{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2}{13.9^2 + 12.47^2 + 11.48^2 + 10.03^2 + 9.45^2} = 0.867 \Rightarrow 0.8 < 0.867 \tag{1}$$

5.4 Question 4: OPTION D

Based on the provided PCA, we want see how different constellations of standardized values affect the sign of their projection onto a given principal component. The second principle component is taken from V, $PC2 = [v_{21}, v_{22}, v_{23}, v_{24}, v_{25}]^T = [-0.5, 0.23, 0.23, 0.09, 0.8]^T$. Suppose then, that we apply high values for standardized values of x_2, x_3 and x_5, i.e. ones, and low values for x_1, i.e. one. The projection onto PC2 thus can be estimated as:

$$(1) \cdot (-0.5) + 1 \cdot (0.23) + 1 \cdot (0.23) + 0 \cdot (0.09) + 1 \cdot (0.8) = 0.76 \Rightarrow 0 < 0.76 \tag{2}$$

5.5 Question 5: OPTION A

Given a vocabulary size of $M = 20000$, and since that $M = f_{11} + f_{10} + f_{01} + f_{00}$, we can use the formula for *Jaccard Similarity*:

$$J(x, y) = \frac{f_{11}}{f_{11} + f_{10} + f_{01}} = \frac{f_{11}}{M - f_{00}} \tag{3}$$

Here $f_{11} = 2$ containing {"*the*", "*words*"}, and $f_{00} = 20000 - 13 = 19987$, containing all other words from M except the unique words between s_1 and s_2, i.e. "*the*" and "*words*" are only counted once. We thus get:

$$J(x, y) = \frac{2}{20000 - 19987} = \frac{2}{13} = 0.153846 \tag{4}$$

5.6 Question 6: OPTION B

From Table 1, we know that light congestion is given by $y = 2$, and we are checking this against the binarized $\hat{x}_2 = 0$. Thus, we are only interested in the first two rows of Table 2, where $p(0, 0|2) = 0.81$ and $p(0, 1|2) = 0.03$. We then get The probability to be:

$$p(\hat{x}_2 = 0|y = 2) = p(\hat{x}_2 = 0, \hat{x}_7 = 0|y = 2) + p(\hat{x}_2 = 0, \hat{x}_7 = 1|y = 2) = 0.81 + 0.03 = 0.84 \tag{5}$$

References

[1] Warwick J. Nash et al. "The Population Biology of Abalone (Haliotis species) in Tasmania. I. Blacklip Abalone (H. rubra) from the North Coast and the Islands of Bass Strait". In: (1994). ISSN: 1034-3288.

[2] Samuel George Waugh. "Extensions to the Cascade-Correlation architecture and benchmarking of feed-forward supervised artificial neural networks". PhD thesis. University of Tasmania, 1995.

[3] Warwick J. Nash et al. *Abalone*. UCI Machine Learning Repository. 1994. DOI: https://doi.org/10.24432/C55C7W.