Statistical analysis of the characteristics of abalones

Marcelo Colares da Silva
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brasil
colaresmarcelo2018@gmail.com

***Abstract*—In the current scenario of expansion of data analysis techniques, applications are emerging in the most varied areas of knowledge. This work presents a statistic of the physical characteristics of the abalones (edible mollusc), aiming to perform univariate, bivariate and multivariate analyzes with the intuition of studying how these variables are distributed and checking for the existence of possible correlations between them. The tools used were histograms for the individual study of each variable, correlation matrix to analyze the variables in pairs and PCA (principal component analysis) as a way to verify which components are most important**

***Index Terms*—statistical learning, abalones, PCA.**

I. Introduction

Currently, with the existence of several data analysis techniques, it is possible to evaluate some phenomena and check which factors are most inactivating for their occurrence. The present work has as its object of study the abalones. The number of species recognized worldwide varies between 30 and 130.The objective is to study how their body characteristics such as diameter, length, among others, along with their age, in order to verify how they are distributed and what is the relationship between them.

II. Methods

The data set has 9 predictors with 4177 observations. Variables include type, length, diameter, height, total weight, shell weight, viscera weight, shell weight and number of rings.

TABLE I
DATA CHARACTERISTICS.

Nome	Tipo de dados	Unidade	Observações
Type	categorical		4177
LongestShell	continuous	mm	4177
Diameter	continuous	mm	4177
Height	continuous	mm	4177
WholeWeight	continuous	gramas	4177
ShuckedWeight	continuous	gramas	4177
VisceraWeight	contínuo	grams	4177
ShellWeight	continuous	grams	4177
Rings	continuous		4177

A. Pre-processing

The Yeo-Johnson transformation to stabilize the variance, and bring the data closer to the normal distribution, in addition to improving the correlation between variables. Where y_i corresponds to the vector of data to be transformed. The transformation is given by:

$$y_i^{\lambda} = \begin{cases} \left((y_i+1)^{\lambda}-1\right)/\lambda & if\ \lambda \neq 0, y \geq 0 \\ log\left(y_i+1\right) & if\ \lambda = 0, y \geq 0 \\ -\left[(-yi+1)^{(2-\lambda)}-1\right]/(2-\lambda) & if\ \lambda \neq 2, y < 0 \\ -log\left(-y_i+1\right) & if\ \lambda = 2, y < 0 \end{cases} \tag{1}$$

B. Statistical measures

1) Mean: In statistics, mean (μ) is defined as the value that shows the concentration of the data in a distribution, as the equilibrium point of the frequencies in a histogram. Average is also interpreted as a significant value in a list of numbers. Let x be a vector with N samples, its average will be given by:

$$\mu = \frac{1}{N}\sum_{i=0}^{N} x_i \tag{2}$$

2) Standard deviation: The standard deviation indicates how the samples are distributed around the mean, while a small standard deviation indicates that they are condensed close to the mean. In a nutshell, the lower the standard deviation, the more homogeneous the sample is.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=0}^{N}(x_i-\mu)^2} \tag{3}$$

3) Skewness: Oblique measures the asymmetry of the tails of the distribution. Asymmetrical distributions that have a "heavier" tail than the other have slanting. Symmetric distributions have zero obliquity.

$$\gamma = \frac{m_3(\mu)}{\sigma^3} \tag{4}$$

Where:

$$m_3(\mu) - \text{Third central moment}$$
$$\sigma - \text{Standard deviation}$$

- Case $\gamma > 0$, then the distribution has a higher number of values above the average. Asymmetry is said to be negative
- Case $\gamma < 0$, then the distribution has a greater amount of below average. Asymmetry is said to be positive

- If $\gamma = 0$, then the distribution is approximately symmetric.

C. Histograms

The histogram, also known as frequency distribution, is the graphical representation in columns or bars of a previously tabulated data set and divided into uniform or non-uniform classes. The base of each rectangle represents a class.

D. Boxplot

The boxplot consists of the first (lower limit) and third quartile (upper limit) and the median (value that separates the larger and the lower half of the data set). The lower and upper stems extend, respectively, from the lower quartile up to the minimum value not less than the lower limit and from the upper quartile to the highest value not greater than the upper limit. The limits are calculated as follows, points outside these limits are considered outliers.

$$Lim_{inf} = Q_1 - 1.5IQR \tag{5}$$

$$Lim_{sup} = Q_3 + 1.5IQR \tag{6}$$

Where IQR is the difference between the quartiles $(Q_3 - Q_1)$ that represents a measure of the variability of the data.

$$IQR = Q_3 - Q_1 \tag{7}$$

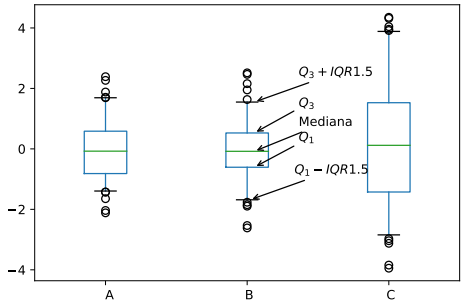

Fig. 1. Example of bloxplot

E. Principal component analysis

PCA is a technique that makes it possible to reduce the dimensions of a data matrix $\boldsymbol{X}$ of size $n \times p$. However, in a way, the characteristics of the original data are preserved. The first component is a normalized linear combination of the original variables.

$$Z_1 = \underbrace{\phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p}_{\text{PC1}} \tag{8}$$

The components generated have a greater variance. The coefficients have the following limitation $sum_j^p\ phi_{ij}^2 = 1$. Following is the graph of the explained variance accumulated as a function of the number of main components, that is, the sum of the individual explained variances for each component.

III. Results

Figure 1 presents the histograms of each numerical variable in an unconditional way, that is, considering all classes. It appears that the distributions do not present such clear symmetries, such asymmetry (γ) may indicate that there is a greater amount of observations above or below the average.

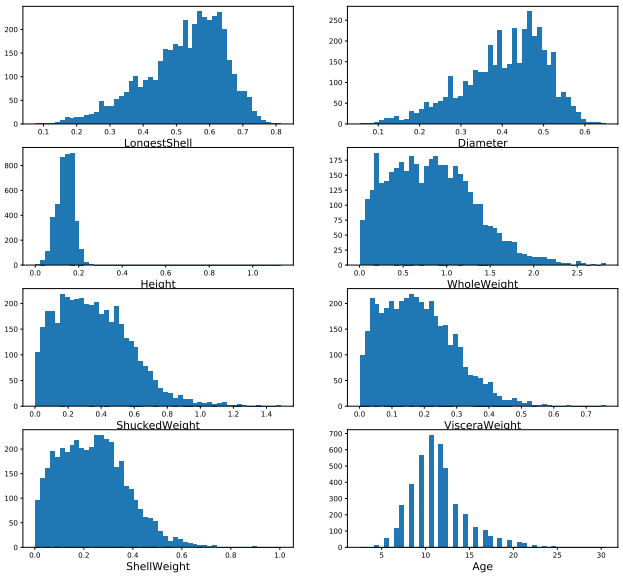

Fig. 2. Unconditional histograms of variables.

Considering the division of data based on the classes defined by the variable Type, Male, female and Child. Below are the respective distributions, the blue color represents the male class, the red one the female class and the green one the child class, as indicated in the legend:

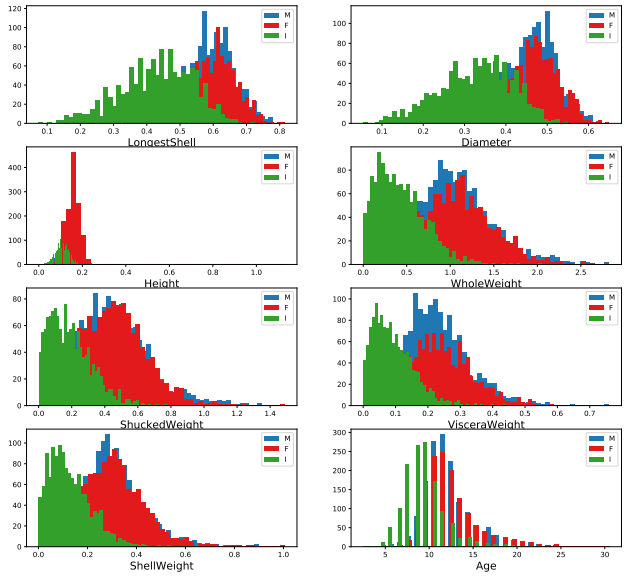

Fig. 3. Conditional histograms of variables

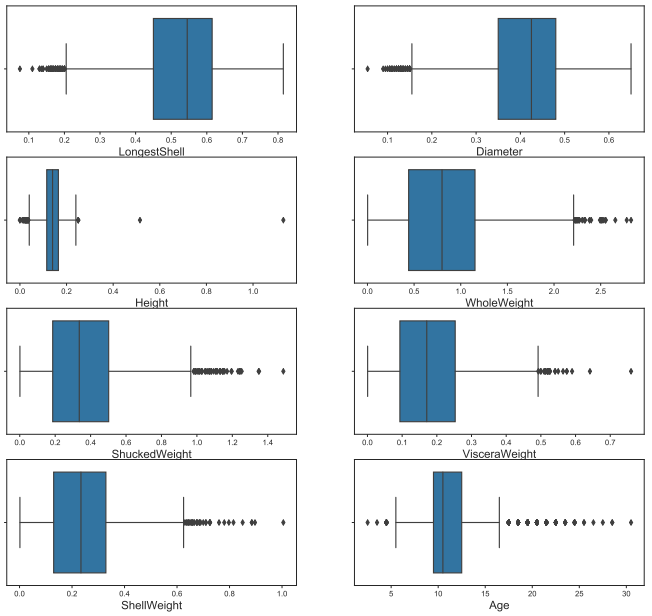

Fig. 4. Boxplot of unconditional variables

TABLE II
AVERAGE OF CONDITIONAL VARIABLES

Variable	Average-M	Average-F	Average-I
LongestShell	0.561391	0.579093	0.427746
Diameter	0.439287	0.454732	0.326494
Height	0.151381	0.158011	0.107996
WholeWeight	0.991459	1.046532	0.431363
ShuckedWeight	0.432946	0.446188	0.191035
VisceraWeight	0.215545	0.230689	0.092010
ShellWeight	0.281969	0.302010	0.128182
Age	12.205497	12.629304	9.390462

It is observed that LongestShell and Diameter present a greater number of values below the average. The other variables present a greater occurrence of values above the average. In particular, Height has the highest asymmetry value, indicating that a possible greater occurrence of values well above the average. It is worth mentioning that the presence of outliers can alter the asymmetry values. To check the presence of this type of point, boxplot-type graphs are used, which consists of a graphical tool capable of representing the variation of a given set of observed data of the numerical type by means of quartiles.

The previous Figure shows that the presence of outliers in the data, especially the Height variable, presents two points much higher than the average, which explains the value obtained for its asymmetry. However, it is not correct to discard these points as information related to measurements would be lost.

Calculating the asymmetries of all variables unconditionally, the following table was obtained:

TABLE III
ASYMMETRY OF VARIABLES

Variable	Asymmetry
LongestShell	-0.639873
Diameter	-0.609198
Height	3.128817
WholeWeight	0.530959
ShuckedWeight	0.719098
VisceraWeight	0.591852
ShellWeight	0.620927
Age	1.114102

The following table shows the asymmetry for each variable in each class

TABLE IV
ASYMMETRY OF CONDITIONAL VARIABLES

Variable	Asymmetry-M	Asymmetry-F	Asymmetry-I
LongestShell	-0.913565	-0.528735	-0.346951
Diameter	-0.923321	-0.506289	-0.292925
Height	0.417547	10.925682	-0.058515
WholeWeight	0.406007	0.368498	0.974459
ShuckedWeight	0.632451	0.546770	0.865294
VisceraWeight	0.506076	0.393427	1.066459
ShellWeight	0.487628	0.691757	1.001923
Age	1.255072	1.474022	1.326831

In terms of asymmetries (γ), it is observed that the abalones of the infantile class present a higher value, of asymmetry, for example, the variables LongestShell, Diameter, Height in the male and female classes suggest a greater occurrence of values below the average, whereas for the child class this also occurs, however, in a less accentuated way. WholeWeight, Shucked-Weight, VisceraWeight and ShellWeight in the abalones in infantile abalones it is clear a greater occurrence of values above the average, so the distributions are much 'heavier' to the right in relation to male and female abalones.

After analyzing the variables individually, the next step was to perform a bivariate analysis. The following is the scatter plot between all pairs of variables.

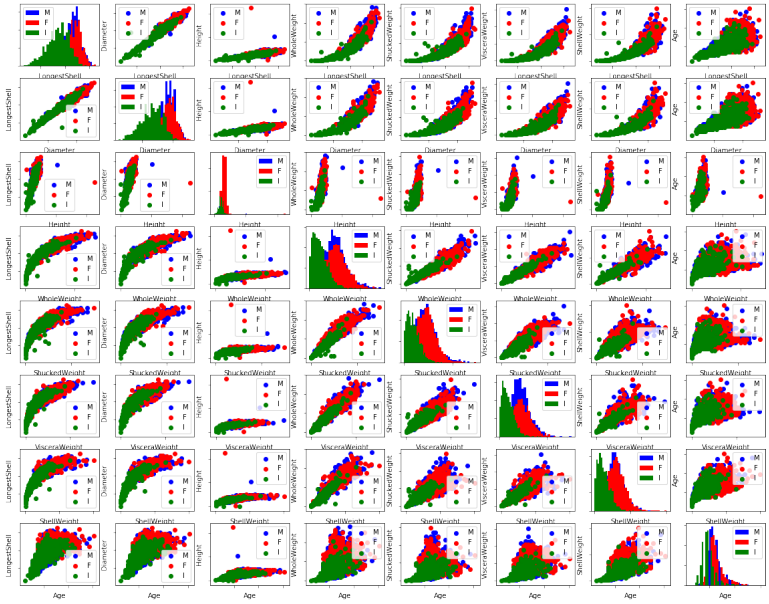

Fig. 5. Scatter plot of all pairs.

a non-linear behavior, such as the pair ShuckedWeight and Diameter, for example. While others have a linear behavior. correlation matrix makes this visualization clearer, its content consists of the correlation coefficients rho with $rho(i,j) = rho_{di,dj}$ and each pair, where i and j represent the rows and columns of the matrix.

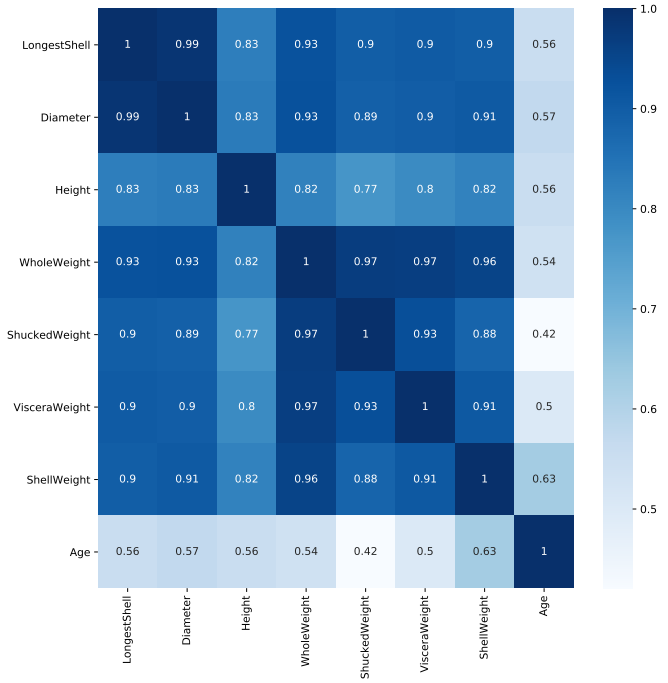

Fig. 6. Correlation matrix

It is verified, then, that there is in fact a linear relationship between some pairs studied. The next step is to minimize the number of variables studied, to reduce the dimensions of the dataset. However, it is first necessary to perform a pre-processing on the data to solve asymmetry problems, as previously discussed. For this, the Yeo-Johnson transformation is used.

TABLE V
NEW ASYMMETRIES FOR EACH VARIABLE

Variável	Assimetria
LongestShell	-0.066669
Diameter	-0.065977
Height	-0.115483
WholeWeight	-0.003980
ShuckedWeight	0.027351
VisceraWeight	0.039780
ShellWeight	0.023390
Age	0.023390

It appears that the asymmetries are much smaller and close to zero, then there are the respective histograms, separated by each class. It is noticeable visually that both distributions have a more apparent symmetry.

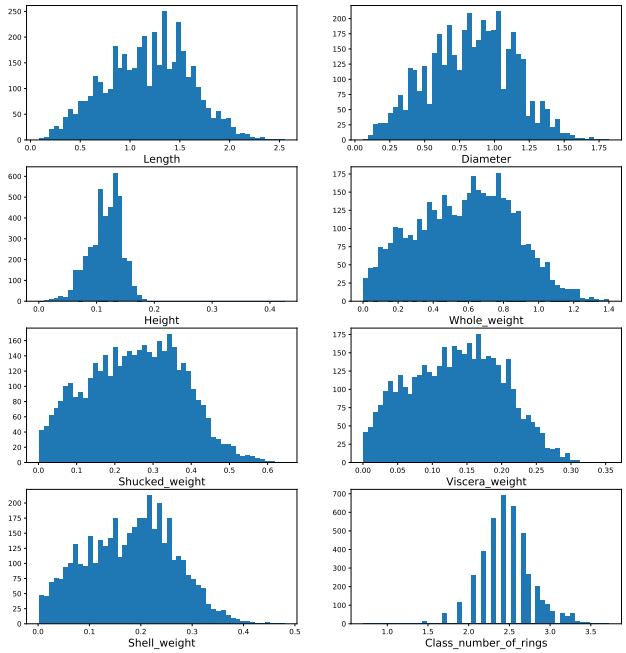

Fig. 7. Conditional histograms of the variables after the application of the Yeo-Johnson transformation.

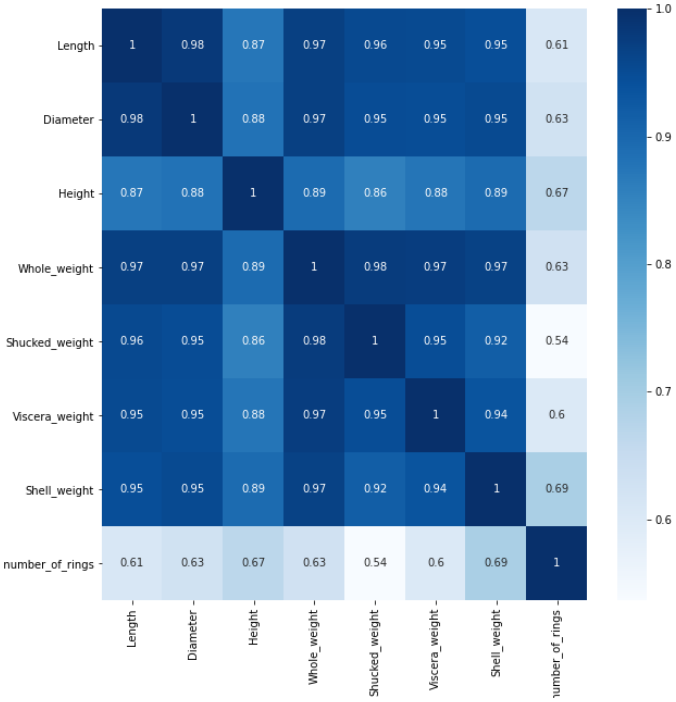

Fig. 8. Correlation matrix after applying the Yeo-Johnson transformation.

It is possible to observe an increase in the correlation of age with the other variables. Such improvement can be explained by the application of the Yeo-Johnson transformation, since one of its characteristics is the increase in the correlation coefficient between two variables that are subjected to this transformation. Figure 5 shows that there are some pairs that have a non-linear behavior, due to the logarithmic nature of the transformation, exponential relationships tended to behave as linear.

Now it performs a reduction of the dimensionality through the PCA, applying the PCA with 4 components, according to the number of variables to 4, there is a variation explained by each component generated in the following Figure:

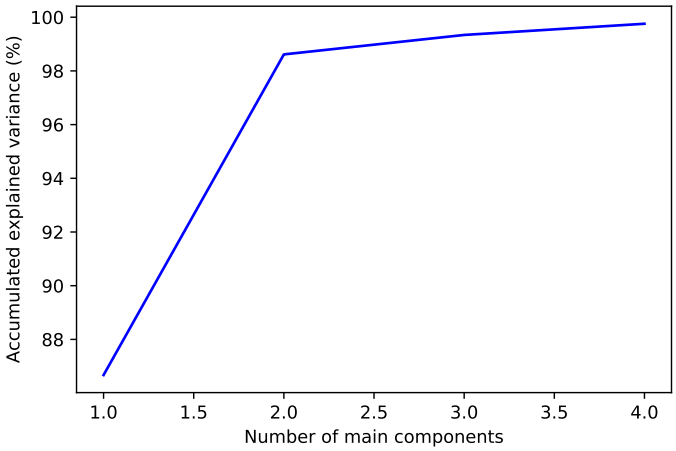

Fig. 9. variance explained by each major component

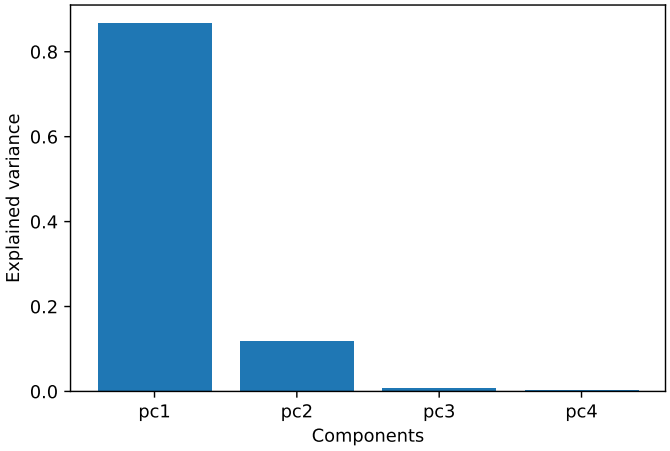

Fig. 10. Variance explained by each component

It turns out that there is a main component that explains almost all of the variance with about 98 text %. The explanation for this is not that there is a more important main component, but that the $PC1$ component is a combination of all the original variables, so the component with the highest degree of explanation is the one that combines all the variables. In other words, there is no elimination of any variable, but the creation of new variables that are combinations of the original ones that make it possible to reduce the dimensions of the original dataset. As the PCA components are orthogonal to each other and are not correlated, it is expected that it will be possible to view the male, female and child classes separately according to the following scatter plot:

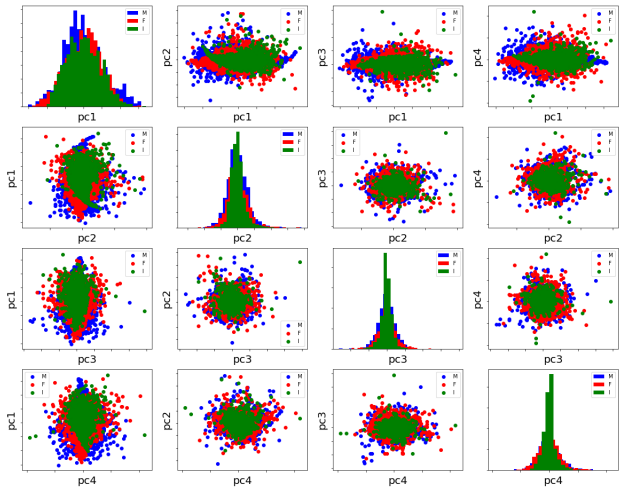

Fig. 11. Scatter plot of the main components.

As can be seen, there is a high overlap between the main components, which makes their separation unfeasible. In particular, the children's class is slightly separated from the male and female, following that a possible separation would occur between adults and children. Looking at the graphs that contain component 1, it appears that the children's class is relatively separate from the male and female classes, which in turn have a high degree of overlap and, therefore, are more difficult to separate.

The PCA is an unsupervised method, the interpretation of the components is not simple, since they are a combination of the original variables. Next, the heat map is observed as a way of verifying how the variables combine to generate the components.

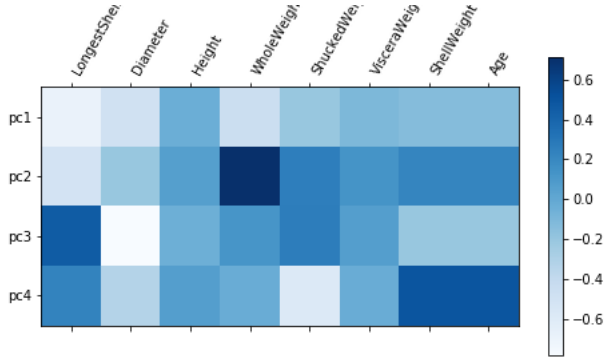

Fig. 12. Correlation between the main components and the original variables - Male

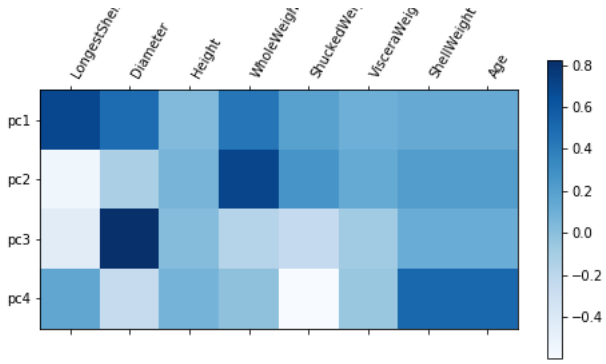

Fig. 13. Correlation between the main components and the original variables - Female

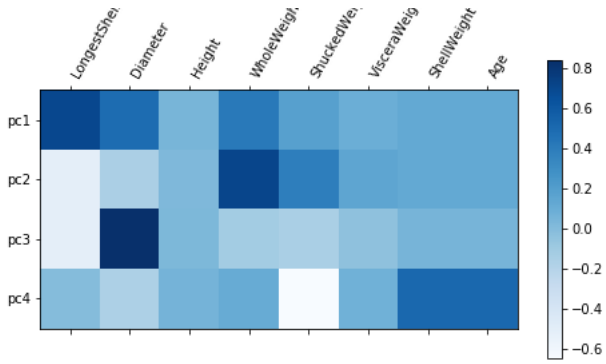

Fig. 14. Correlation between the main components and the original variables - Child

REFERENCES

[1] Marcos Nascimento Magalhães, Antônio Carlos Pedroso De Lima , Noções de Probabilidade e Estatística. Edusp, 2007.

[2] Gareth James, Daniela Witten, Trevor Hastie e Robert Tibshirani, An Introduction to Statistical Learning: with Applications in R.

[3] Minka, T. P.. "Automatic choice of dimensionality for PCA". In NIPS, pp. 598-604.