Estudo de métodos de análise de dados

Antônio Wellington Lima Cavalcante Júnior
Dep. de Engenharia de Computação
Universidade Federal do Ceará
Fortaleza
wellingtonjuniorg5@gmail.com

Igor Torquato Maia Sales
Dep. de Engenharia de Computação
Universidade Federal do Ceará
Fortaleza
igormaiasales@gmail.com

José Moacir Holanda Neto
Dep. de Engenharia Elétrica
Universidade Federal do Ceará
Fortaleza
moacirhn10@gmail.com

Kaique Thompson Pereira Vieira
Dep. Engenharia de Teleinformática
Universidade Federal do Ceará
Fortaleza
kaique.thompson.p@gmail.com

***Abstract* — Using a database with more than 4000 abalone samples, separated into 3 classes and containing 8 predictors, several analyzes were made: unconditional mono-variate analysis, class-conditional mono-variate analysis, unconditional bi-variate analysis and unconditional mult- variate analysis. With the exception of the height predictor, which showed asymmetry on the left due to two outliers, the Male and Female classes showed great similarity, different from the Infant class. In addition, after performing the PCA, we realized that the first major component represented almost the entire percentage of influence of the data.**

Palavras-chave — métodos, classe, variate, mono, bi, multi

I Introdução

Com base em um banco de dados contendo informações de 4177 moluscos, mais conhecidos como abalones (tanto no inglês como no português), divididos em três classes (Male, Female e Infant) onde consta o comprimento, diâmetro, altura, peso total, peso descascado, peso das vísceras, peso da casca e quantidade de anéis.

Através dos métodos que serão vistos posteriormente, verificaremos esses dados, bem como uma relação entre eles quando existente, a fim de identificar padrões ou falta deles. Por fim, apresentaremos os resultados dessa análise.

II Métodos

II.A Unconditional mono-variate

O primeiro método utilizado foi o unconditional mono-variate, onde geramos histogramas que relacionam um preditor com a frequência de ocorrência, independente da classe do abalone. O objetivo dessa análise é verificar a disposição das características da espécie em geral, sem considerar a classe.

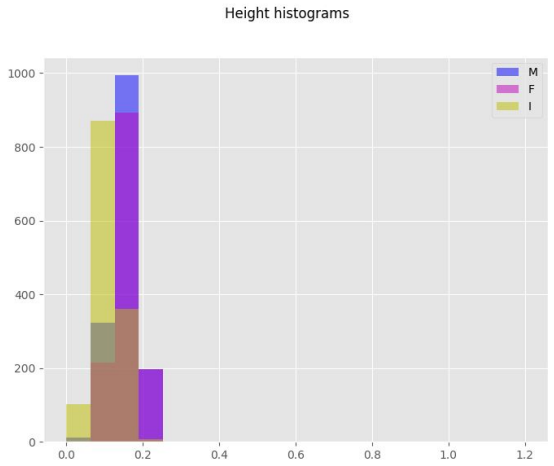
Esse método permite o entendimento de características gerais do conjunto sem revelar aspectos mais específicos relacionados às possíveis classes presentes. Entretanto, observando as distribuições, podemos visualizar quais métodos de pré-processamento se fazem necessários para a utilização da base de dados na criação de modelos preditivos.

II.B Class-conditional mono-variate

O segundo método foi o class-conditional mono-variate, no qual também geramos histogramas como o método anterior, mas diferenciando-os pelas classes. No caso dos abalones, as classes identificadas foram Male, Female e Infant. Isso nos mostra uma disposição mais específica das características.

Diferente do método anterior, esse método permite enxergar mais informações, pois podemos verificar comportamentos diferentes ou não, de acordo com a classe.

Nessas amostras de abalones, identificamos que as classes male e female têm padrões parecidos, no entanto, o infantil já demonstra características diferentes como é possível observar no exemplo abaixo.



No gráfico dos preditores Height é possível observar também a presença de Outliers que “espremem” o gráfico para a esquerda.

Podemos dizer também que o método unconditional mono-variate tem uma maior utilidade se comparado junto do conditional mono-variate, pois é possível analisar se o comportamento da espécie é comum as classes e, se não, quanto cada classe influencia no geral, tendenciando o formato dos histogramas.

II.C Unconditional bi-variate

O terceiro método foi o unconditional bi-variate. Nele deixamos de utilizar os histogramas e passamos para a plotagem, além de que é feita uma relação entre diferentes preditores ao invés de analisar cada um separadamente. Os gráficos podem ser visualizados no Anexo.

É possível perceber que alguns preditores possuem uma alta correlação, tornando a inferência de um gráfico médio das observações uma tarefa trivial, como é o caso dos preditores Length e Diameter. Entretanto, também existem os casos onde essa relação é escassa e dificilmente é possível visualizar uma tendência no gráfico. Como por exemplo, O preditor Rings se relaciona muito confusamente com o todos os tipos de preditores Weight para se inferir qualquer informação do gráfico.

II.D Unconditional multi-variate

Por fim, utilizamos o método unconditional multi-variate, que visa reduzir os preditores avaliados através do procedimento de PCA, onde, no nosso caso, é utilizado apenas os dois preditores com maior autovalores para representar uma relação entre as classes.

Diferente da plotagem do método anterior, este gera apenas um gráfico, que pode ser encontrado no Anexo, pois sua principal função é reduzir a quantidade de dados analisados, mantendo a maior representação possível de acordo com a dimensão necessária, portanto, uma amostragem que antes tinha 7 preditores, agora é representada com apenas dois.

III Resultados

Tabela de Valores de Interesse

Preditores	Valores de Interesse		
	Média	***Desvio Padrão***	***Assimetria***
Length	0.52399201	0.12009291	-0.63987326
Diameter	0.40788125	0.09923986	-0.60919814
Height	0.13951640	0.04182706	3.12881738
Whole weight	0.82874216	0.49038902	0.53095856
Shucked weight	0.35936749	0.22196295	0.71909792
Viscera weight	0.18059360	0.10961425	0.59185215
Shell weight	0.23883086	0.13920267	0.62092682
Rings	9.93368446	3.22416903	1.11410190

Unconditional mono-variate table

Através da análise dos gráficos obtidos (todos disponíveis no anexo ao fim do artigo), podemos identificar um comportamento semelhante entre as classes Male e Female, visto nos gráficos do class conditional mono-variate e reforçado na diagonal do unconditonal bi-variate, concluindo que as divergências no gráfico do unconditional mono-variate se devem aos dados da classe Infant.

Observamos também uma distribuição assimétrica à direita no preditor Heigth. Essa distorção é causada por dois outliers presentes nos dados desses preditores, um na classe Male e outro na classe Female.

Ao realizar o procedimento de PCA para o método unconditional multi-variate, observamos na tabela da variância que o primeiro componente principal sozinho já representava quase toda a porcentagem de influência nos dados. Já no gráfico do PCA, vemos que os pontos da classe Infant estão concentrados mais à esquerda do gráfico e as classes Male e Female mais a direita, conservando o comportamento visto nos outros métodos e mostrando que o PCA teve um bom êxito.

IV Referências

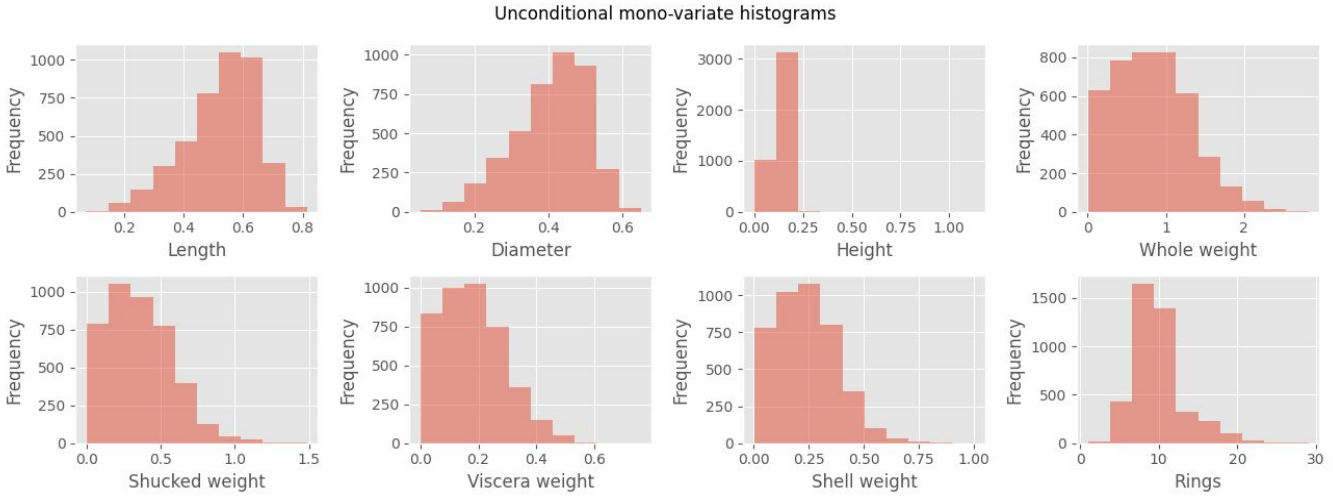
1 Material de sala de aula

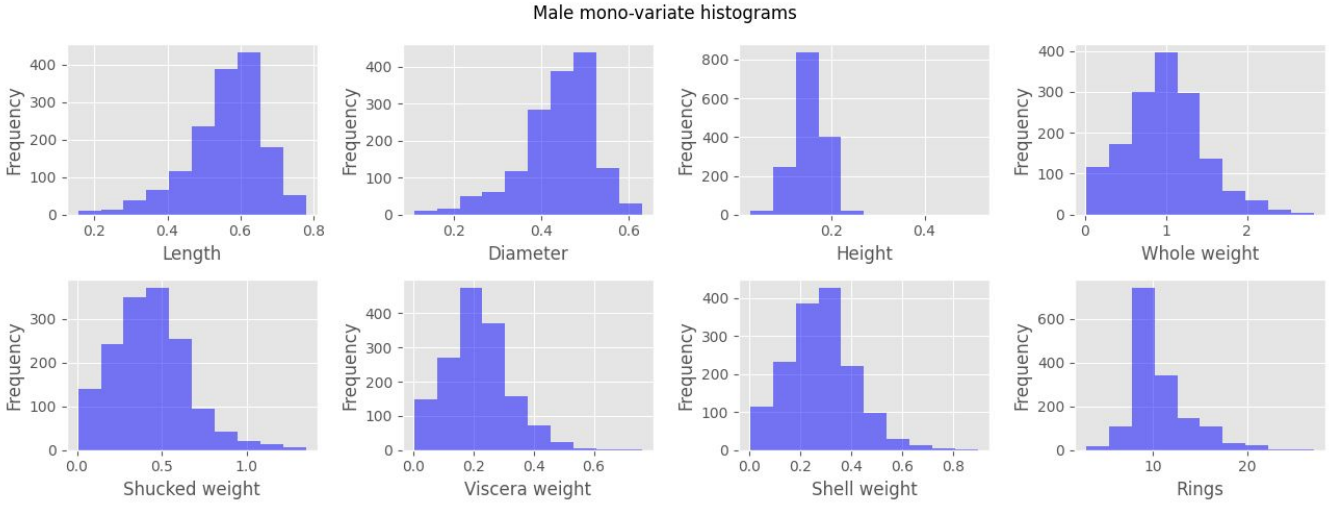
2 Michael Galarnyk, PCA using Python (scikit-learn), 2017. https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60 <Ultimo acesso: 10/01/2021>

3 Matplotlib: Visualization with Python. https://matplotlib.org/ <Ultimo acesso: 10/01/2021>

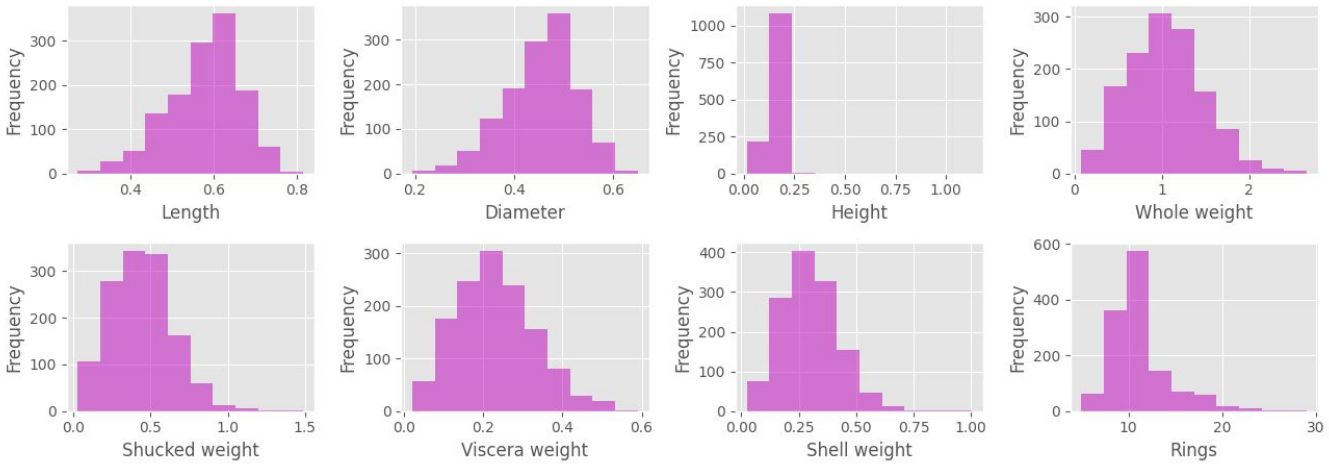
4 Seaborn: statistical data visualization. https://seaborn.pydata.org/ <Ultimo acesso: 10/01/2021>

V ANEXO

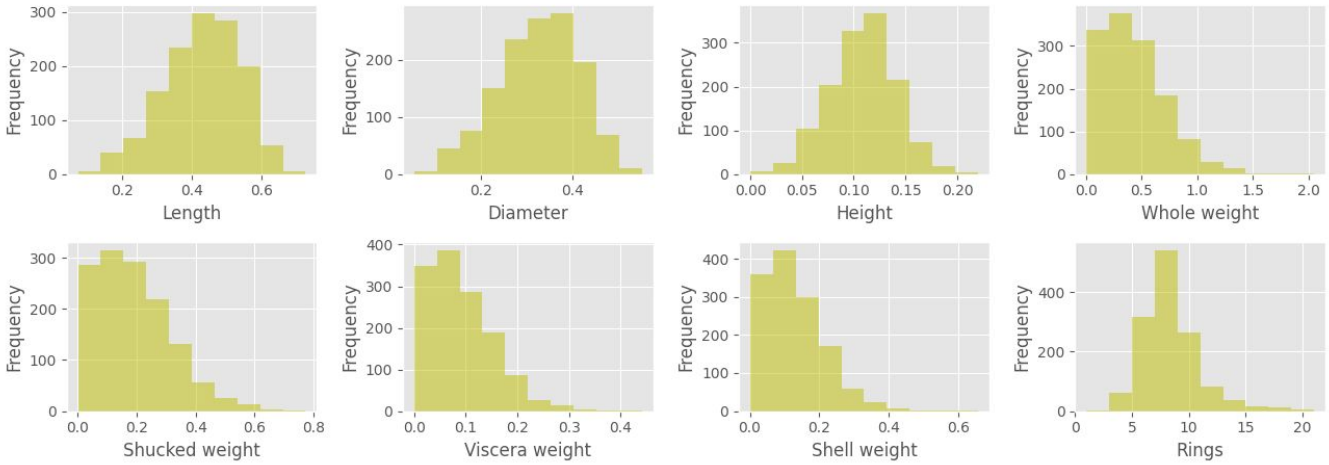




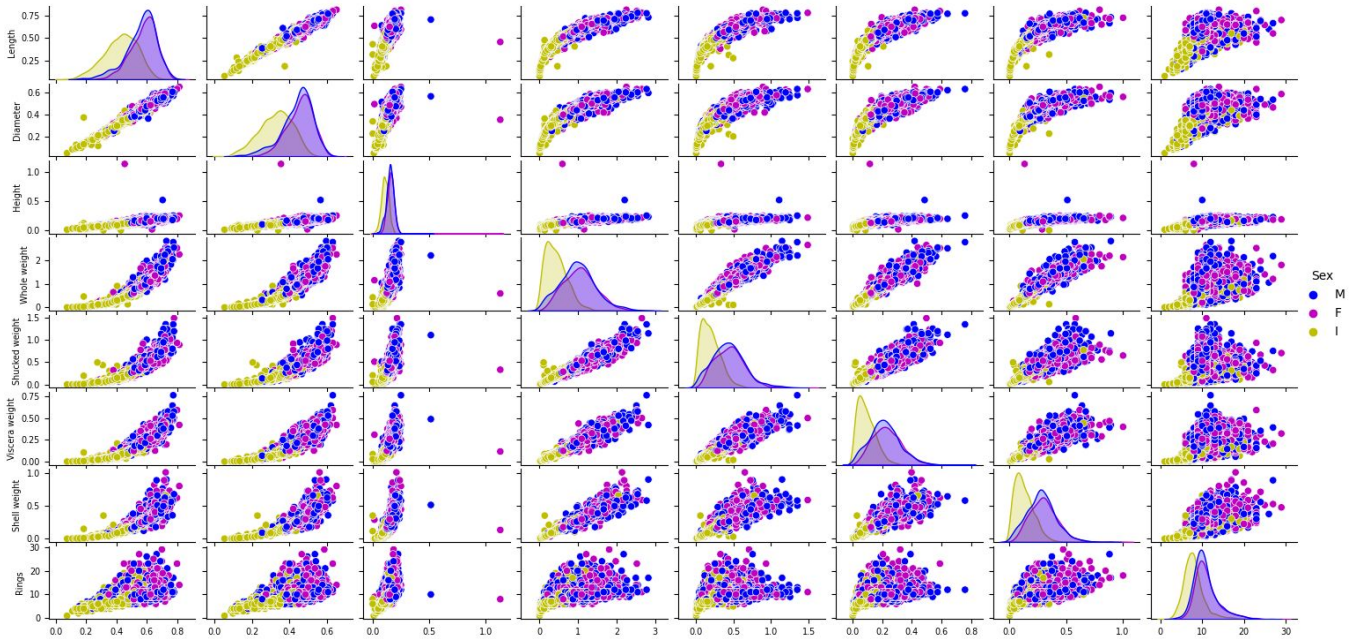
Female mono-variate histograms

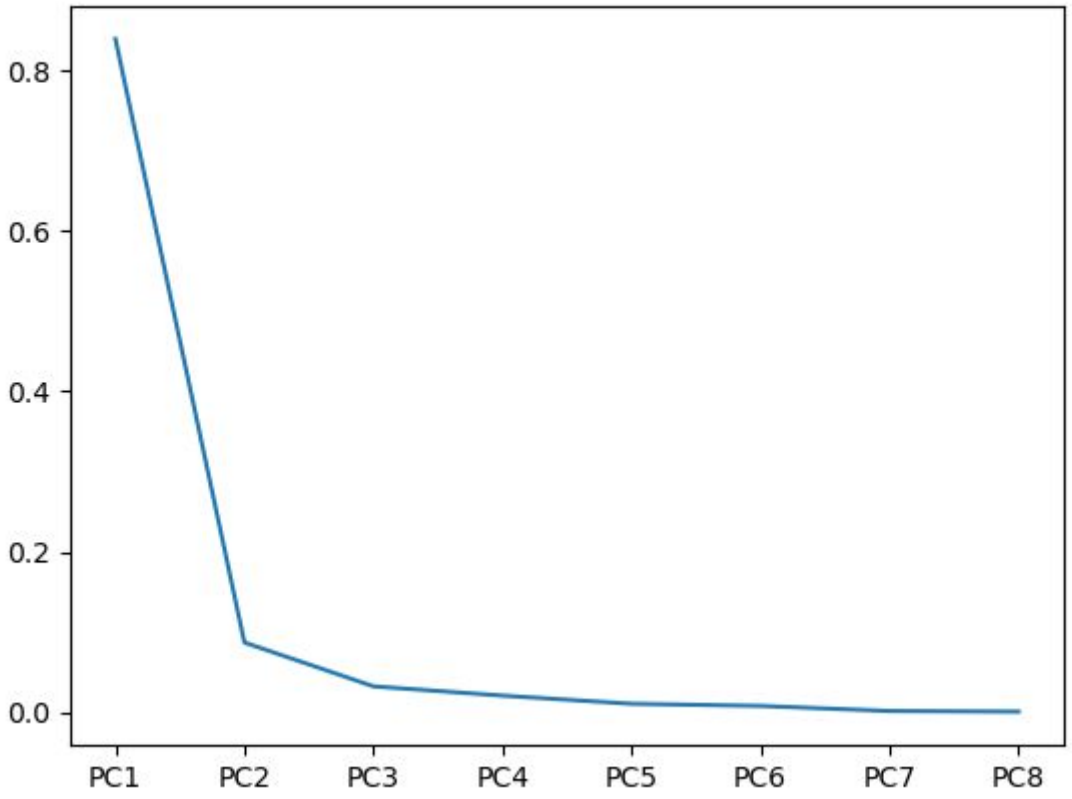


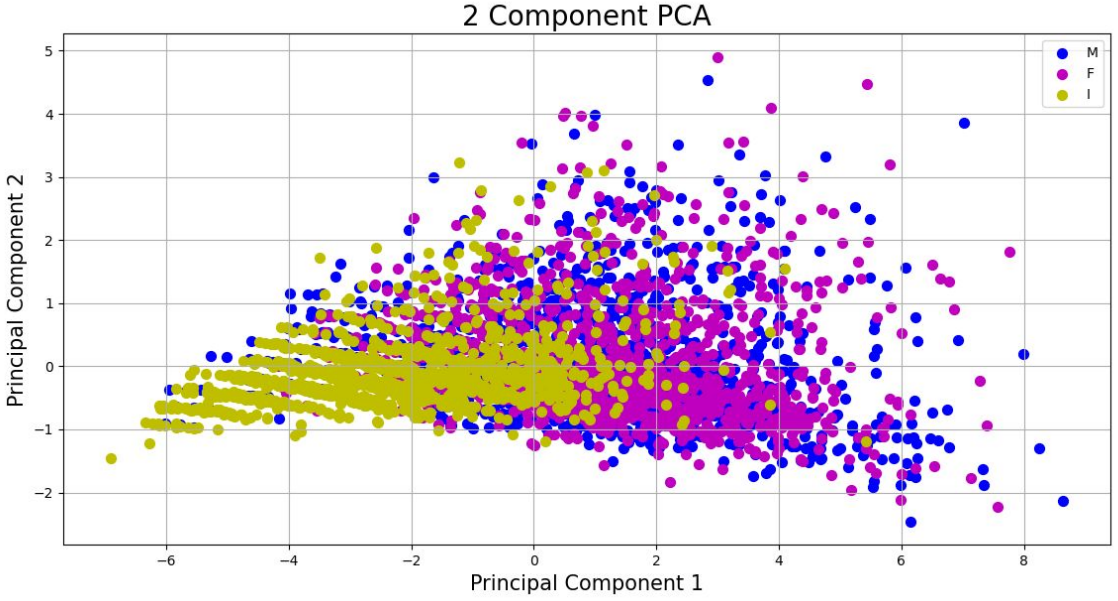
Infant mono-variate histograms



Unconditional bi-variate analysis




0.8
0.6
0.4
0.2
0.0
PC1
PC2
PC3
PC4
PC5
PC6
PC7
PC8


2 Component PCA
Principal Component 2
Principal Component 1
M
F
I