Prevendo a idade do abalone – um estudo

Alice Mateus da Silva
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
alicemateus13@hotmail.com

Andrea Sampaio Elias
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
andrease@alu.ufc.br

Arnaldo Aguiar Portela Filho
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
arnaldoaguiarp@gmail.com

Italo Viana Severo
Departamento de Teleinformática line
Universidade Federal do Ceará
Fortaleza, Brasil
italo_1002@live.com

***Resumo* — Este artigo trata do estudo e da modelagem de dados do molusco abalone. Tem-se como objetivo obter a previsão de número de anéis do molusco e, consequentemente, idade, para saber a hora de colocá-los à venda.**

Palavras-chave – abalone, análise, regressão linear, ridge, lasso, PCR, python, cross-validation, previsão, anéis, idade, venda, consumo.

I. INTRODUÇÃO

No mercado de venda de moluscos é necessário saber algumas informações sobre o animal para otimizar os lucros e cumprir a legislação. Há leis que proíbem a venda de animais para consumo quando estes animais são ainda filhotes. Além disso, a venda de animais filhotes acaba por trazer perda de lucro, pois o animal tem potencial de crescer mais e ficar maior e mais pesado, podendo gerar um maior valor na venda por quilo.

Pensando nisso, são feitos estudos para se poder observar o melhor momento de vender um produto de origem animal. Para tanto, é necessário ter conhecimento de alguns aspectos biológicos e estatísticos sobre o animal.

No caso do abalone, objeto do presente estudo, é necessário ter conhecimento de em que momento o molusco já se encontra adulto e, segundo a biologia, pode-se saber a idade do abalone por seu número de anéis. Para melhor aferência do número de anéis, pode-se usar outras medidas como, por exemplo diâmetro, altura ou peso. Assim, faz-se necessário um modelo preditivo para que através de medidas do abalone possa-se obter seu número de anéis e, consequentemente, sua maturidade.

Para a criação de um bom modelo preditivo faz-se uso de alguns métodos, como o de regressão linear ordinária, L2 (Ridge) e PCR.

II. MÉTODOS

O método utilizado para fazer a modelagem de dados foi o de regressão linear. Inicialmente foi feito o tratamento dos dados. Verificou-se se havia dados ausentes, o que não foi encontrado, como mostrado:

```
Type             0
LongestShell     0
Diameter         0
Height           0
WholeWeight      0
ShuckedWeight    0
VisceraWeight    0
ShellWeight      0
Rings            0
dtype: int64
```

```
Type             4177
LongestShell     4177
Diameter         4177
Height           4177
WholeWeight      4177
ShuckedWeight    4177
VisceraWeight    4177
ShellWeight      4177
Rings            4177
dtype: int64
```

Fig. 1. – Na imagem à esquerda, resultado da verificação da presença de dados nulos. Na imagem à esquerda, resultado da verificação da quantidade de dados por cada coluna.

A seguir fez-se a verificação de outliers (discrepâncias nos dados). Um outlier é um ponto ou conjunto de pontos que são bastante diferentes de outros pontos e, muitas vezes eles podem ser muito altos ou muito baixos. Portanto, é necessário remover os outliers já que estes são um dos principais motivos para um modelo menos preciso. Para removê-los foi usada a técnica IQR, que consiste basicamente em achar a média dos valores e, então, dividir a lista de valores entre os maiores que a média (onde está Q3) e os menores que a média (onde está Q1). Então, pega-se a média da lista de valores maiores (Q3) e a média da lista de valores menores (Q1) e subtrai-se Q3 por Q1. Com o valor da subtração é possível identificar valores que estão muito discrepantes e, assim, poder removê-los. Para todas as colunas de dados foi feito a plotagem desse tipo de dado com os outliers e depois de realizada a técnica de remoção, a plotagem sem os outliers, como demonstrado no exemplo abaixo:

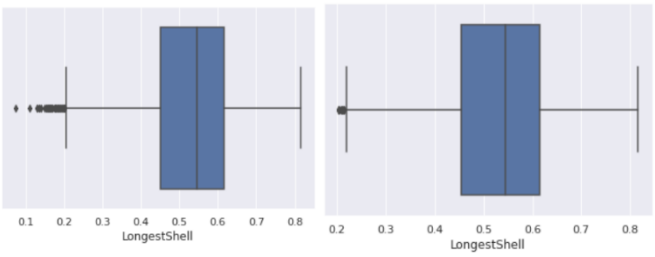

Fig. 2. – Na imagem à esquerda plotagem dos dados de 'LongestShell' com outliers. Na imagem à direita plotagem dos dados de 'LongestShell' sem outliers.

Ao final da remoção de outliers, obteve-se uma tabela de dados com 3.773 linhas; inicialmente tínhamos um conjunto de dados com 4.177 linhas.

	Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
0	M	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.1500	15
1	M	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.0700	7
2	F	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.2100	9
3	M	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.1550	10
4	I	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.0550	7
...	...	...	...	...	...	...	...	...	...
4172	F	0.565	0.450	0.165	0.8870	0.3700	0.2390	0.2490	11
4173	M	0.590	0.440	0.135	0.9660	0.4390	0.2145	0.2605	10
4174	M	0.600	0.475	0.205	1.1760	0.5255	0.2875	0.3080	9
4175	F	0.625	0.485	0.150	1.0945	0.5310	0.2610	0.2960	10
4176	M	0.710	0.555	0.195	1.9485	0.9455	0.3765	0.4950	12

3773 rows × 9 columns

Fig. 3. – Tabela após a remoção de outliers

Foi realizada também a busca por correlações entre os dados, para melhor entendimento dos mesmos, como demonstrado na imagem seguinte:

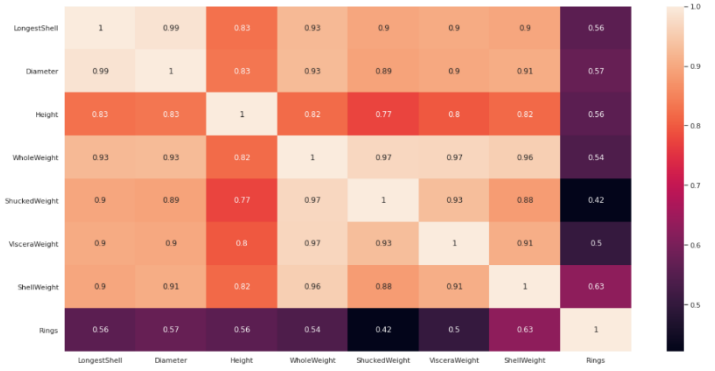

Fig. 4. – Tabela de correlação entre os dados.

Fez-se, também, a descrição dos dados com alguns valores obtidos:

	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight	Rings
count	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000	4177.000000
mean	0.523992	0.407881	0.139516	0.828742	0.359367	0.180594	0.238831	9.933684
std	0.120093	0.099240	0.041827	0.490389	0.221963	0.109614	0.139203	3.224169
min	0.075000	0.055000	0.000000	0.002000	0.001000	0.000500	0.001500	1.000000
25%	0.450000	0.350000	0.115000	0.441500	0.186000	0.093500	0.130000	8.000000
50%	0.545000	0.425000	0.140000	0.799500	0.336000	0.171000	0.234000	9.000000
75%	0.615000	0.480000	0.165000	1.153000	0.502000	0.253000	0.329000	11.000000
max	0.815000	0.650000	1.130000	2.825500	1.488000	0.760000	1.005000	29.000000

Fig. 5. – Tabela de descrição dos dados

Feito o pré-processamento de dados, passou-se à modelagem dos dados.

Inicialmente fez-se a escolha entre diferentes conjuntos de dados, verificando-se a acurácia de cada um na regressão linear. Escolhe-se usar o que apresentou melhor acurácia, ou seja, o dataset 2.

Dataset 1	Dataset 2	Dataset 3
MSE = 4.90	MSE = 2.73	MSE = 3.00
RMSE = 2.142	RMSE = 1.653	RMSE = 1.73
R^2 = 0.53	R^2 = 0.49	R^2 = 0.29

Fig. 6. – Dataset 1 - com outliers e tipos M, F e I (4177 linhas de dados); Dataset 2. - sem outliers e tipos M, F e I (3773 linhas de dados); Dataset 3 - sem outliers e tipos M e F (2512 linhas de dados).

O conjunto de dados contém uma coluna com os tipos de abalones, ou seja, ‘M’, ‘F’ e ‘I’, representando, masculino, feminino e infantil, respectivamente. Como são valores não numéricos, os mesmos não podem ser modelados numa regressão linear, portanto, foi necessário a substituição destes por valores numéricos. ‘M’, ‘F’ e ‘I’ foram substituídos pelos valores, 0, 1 e 2, respectivamente.

Em seguida, passamos a dividir o conjunto de dados em parâmetros de entrada e saída. À saída (Y - outcome), foi atribuída o conjunto de dados da coluna ‘Rings’, ou seja, o número de anéis dos abalones. Já às entradas (X - income), foi atribuído o restante do conjunto de dados ("Type", "LongestShell", "Diameter", "Height", "WholeWeight", "ShuckedWeight", "VisceraWeight", "ShellWeight").

	Type	LongestShell	Diameter	Height	WholeWeight	ShuckedWeight	VisceraWeight	ShellWeight
0	0	0.455	0.365	0.095	0.5140	0.2245	0.1010	0.150
1	0	0.350	0.265	0.090	0.2255	0.0995	0.0485	0.070
2	1	0.530	0.420	0.135	0.6770	0.2565	0.1415	0.210
3	0	0.440	0.365	0.125	0.5160	0.2155	0.1140	0.155
4	2	0.330	0.255	0.080	0.2050	0.0895	0.0395	0.055

Fig. 7. – Tabela com os parâmetros de entrada

Passou-se então aos procedimentos de criação do modelo de regressão linear ordinário, que funciona minimizando a soma do quadrado das diferenças entre os valores preditos e os valores reais. Para tanto, foi necessário dividir o conjunto de dados, tanto entradas, quanto saídas, em conjuntos de treino e conjuntos de teste. A partir daí foi feito o modelo de regressão linear ordinário; o modelo foi, então, treinado com o conjunto de dados de treino usando-se validação cruzada (cross-validation – 10 ‘kfolds’); logo depois, o modelo foi testado com o conjunto de dados de teste. Foram encontrados resultados de acurácia tanto para o conjunto de dados de treino quanto para o conjunto de dados de teste.

Buscando testar mais métodos de regressão linear com o conjunto de dados, aplicou-se a modelagem dos dados com o uso do modelo de regressão linear L2 (Ridge). Nesse tipo de regressão, a função de custo é alterada adicionando uma penalidade equivalente ao quadrado da magnitude dos coeficientes.

Isso é equivalente a dizer a minimização da função de custo sob uma condição específica.

Portanto, a regressão Ridge restringe os coeficientes (w). O termo de penalidade (lambda) regulariza os coeficientes de forma que se os coeficientes assumirem valores grandes, a função de otimização será penalizada. Portanto, a regressão L2 (Ridge) reduz os coeficientes e ajuda a reduzir a complexidade e a multicolinearidade do modelo.

Nessa regressão foi feita a modelagem e usou-se os dados de treino, com a ajuda da validação cruzada, para testar o modelo. Obteve-se, então, o valor ótimo de lambda(λ), a partir do qual o modelo fica otimizado. Como se demonstra, o valor ótimo de λ encontrado foi de 1, pois foi com esse valor de λ que se apresentou o maior valor de R^2 e o menor valor de RMSE, ou seja, valores de melhor acurácia.

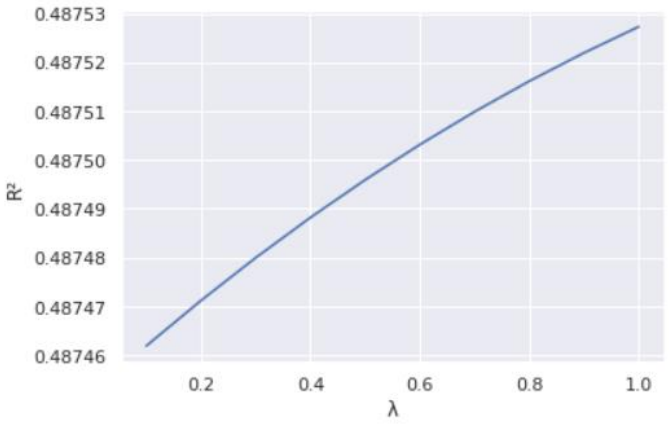

Fig. 8. – Gráfico de R^2 *x* λ.

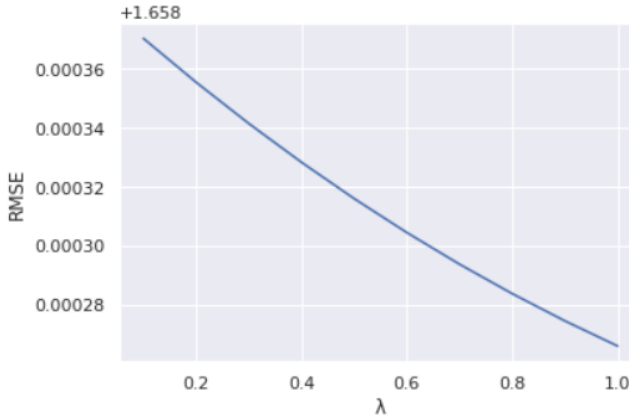

Fig. 9. – Gráfico de R^2 *x* RMSE.

Logo depois, os dados de teste foram usados para testar a acurácia do modelo obtido.

Testamos, ainda, o conjunto de dados com o uso do modelo de regressão PCR (Principal Component Regression), que é um método que faz uso do PCA (Principal Component Analysis) para reduzir a dimensionalidade do conjunto de dados de modo a ser melhor a visualização das correlações entre os dados e que também evita erros decorrentes das dependências entre as variáveis independentes usadas na regressão.

Para o uso desse modelo de regressão, é feita a modelagem e, logo após, faz-se o teste, usando, ainda, a validação cruzada, de com quantos componentes o modelo obtém melhores resultados. Como o conjunto de dados em análise contém 8 (oito) tipos de entradas, o modelo foi testado para cada uma das possibilidades de quantidade de componentes usados (de 1 a 8), como demonstrado a seguir:

```
Número de componentes: 1
R²: 0.3550129483175688
RMSE: 1.866361335672905

Número de componentes: 2
R²: 0.3623165911644402
RMSE: 1.8556144638193182

Número de componentes: 3
R²: 0.4410160502078098
RMSE: 1.73697952880622

Número de componentes: 4
R²: 0.44090799276325154
RMSE: 1.737115917876598

Número de componentes: 5
R²: 0.4883462913991937
RMSE: 1.6618630993110615

Número de componentes: 6
R²: 0.490625754771756
RMSE: 1.6581531512283187

Número de componentes: 7
R²: 0.4910373346420526
RMSE: 1.6574786792398408

Número de componentes: 8
R²: 0.4974951145640516
RMSE: 1.6468509950687074
```

Fig. 10. – Testagem para obter o valor ótimo do número de componentes

Fez-se as plotagens dos valores medidores de acurácia pelo valor do número de componentes:

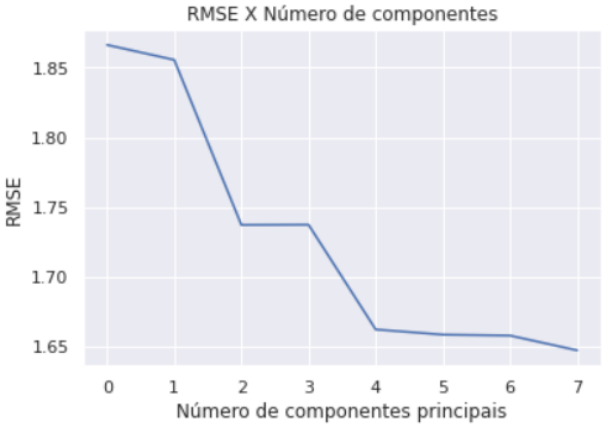

Fig. 11. – Gráfico de RMSE *x* Número de componentes principais

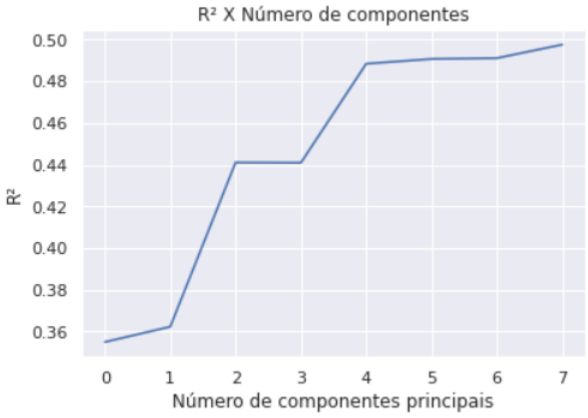

Fig. 12. – Gráfico de R^2 *x* Número de componentes principais

Como observado, obteve-se o valor ótimo de 8 componentes, pois foi com o uso desse número de componentes que resultou o menor RMSE e o maior R^2.

Foi feito então, o treinamento do modelo com o uso do conjunto de dados de treino e o teste do modelo com o uso do conjunto de dados de teste.

III. RESULTADOS

Para a regressão linear ordinária obteve-se os seguintes valores de acurácia para o conjunto de dados de treino com o uso da validação cruzada:

```
KFold 1
Valor do R2: 0.4948836303298054
Valor do RMSE: 1.6444748996577132

KFold 2
Valor do R2: 0.5241085163665278
Valor do RMSE: 1.5993284531439838

KFold 3
Valor do R2: 0.4655532154396691
Valor do RMSE: 1.6976094626234326

KFold 4
Valor do R2: 0.4440528862771522
Valor do RMSE: 1.6722060431469152

KFold 5
Valor do R2: 0.5120561546003586
Valor do RMSE: 1.5675536720407484

KFold 6
Valor do R2: 0.5036951160046317
Valor do RMSE: 1.6738662949880398

KFold 7
Valor do R2: 0.5371896515256871
Valor do RMSE: 1.5758232004344592

KFold 8
Valor do R2: 0.421224519484209
Valor do RMSE: 1.7418897437639418

KFold 9
Valor do R2: 0.4565094334337375
Valor do RMSE: 1.700368759636524

KFold 10
Valor do R2: 0.5152472392086533
Valor do RMSE: 1.7107405013293842
```

Fig. 13. – Resultados de acuraria do modelo de regressão linear ordinário utilizando os dados de treino.

Podemos comentar que para um bom modelo de regressão linear devemos ter valores de RMSE o mais próximo de 0 possível e valores de R^2 o mais próximo de 1 possível.

Quanto mais próximo de 1 o valor de R^2 melhor o modelo de regressão linear. R^2 varia de 0 a 1 e se temos um valor de R^2 de 0,5, por exemplo, concluímos que o modelo de regressão linear pode prever 50% do que está no mundo real. Quanto ao RMSE, quanto mais próximo de 0, melhor o modelo de regressão linear. RMSE é uma medida do erro médio do resultado do modelo de regressão linear em relação ao mundo real. Por exemplo, se o RMSE é 1,6 concluímos que o modelo está errando em 1,6 anéis a mais ou a menos do valor do mundo real.

Nosso modelo de regressão linear apresentou os valores:

- Para os dados de treino:
  - RMSE: 1.6583861030765141 (≈1,66)
  - R^2: 0.4874520362670432 (≈49%)

- Para os dados de teste:
  - RMSE: 1.5920600935181664 (≈1,60)
  - R^2: 0,5398930152430355 (≈54%)

Os valores de acurácia obtidos (RMSE e R^2) entre os resultados com os dados de treino e de teste são bem próximos, com os valores obtidos com os dados de teste sendo ainda melhores.

A seguir faz-se a plotagem dos resultados do modelo de regressão linear ordinário mostrando o gráfico de número de anéis preditos pelo modelo *versus* número de anéis do mundo real.

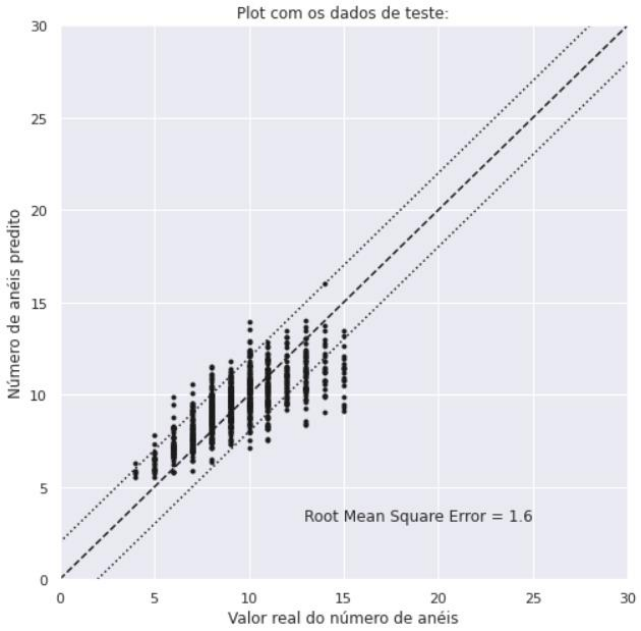

Fig.14. – Gráfico Número de anéis predito *x* número de anéis real.

No método de regressão linear L2 (Ridge) foi obtido o valor ótimo de λ igual 1 e a partir daí fez-se o modelo de Ridge, em que foram obtidos os seguintes valores de acurácia:

- RMSE: 1.5920600935181664
- R^2: 0.5398130502452981

Os valores de acurácia obtidos na regressão linear Ridge foram bem parecidos com os valores obtidos anteriormente com a regressão linear ordinária.

Para a regressão PCR, no momento da criação do modelo, foi feito o teste para obter o valor ótimo do número de componentes. Obteve-se que a partir de 5 (cinco) componentes a acurácia era quase a mesma; porém mesmo que com valores de diferença bem pequena, o número de componentes e que obteve melhores resultados de acurácia foi 8, ou seja, com o uso de todos os componentes. Com o treinamento do modelo com os dados de treino e o teste do modelo com os dados de teste, obteve-se os seguintes resultados medidores da acurácia do modelo:

- RMSE: 1.5892169501858608
- R^2: 0.5075327688683663

Comparando os valores de acurácia obtidos com o PCR com aqueles obtidos com a regressão linear ordinária e a regressão linear Ridge observa-se que os valores de RMSE para os dados de teste são praticamente idênticos nos três tipos de regressão. No entanto, ao se observar os valores obtidos para R^2, nota-se que o valor obtido no PCR é menor que os obtidos nos dois outros modelos. Conclui-se, portanto, que para o conjunto de dados em análise, dentre os três modelos propostos, deve-se utilizar ou o método de regressão linear ordinário ou o método de regressão linear L2 (Ridge).

Com esses modelos preditivos, será, então, possível prever a idade do molusco abalone de modo a suprir a necessidade dos produtores e vendedores. Com o número de anéis, saída do modelo preditivo, consegue-se saber a idade do abalone, pois a mesma é igual a 1,5 vezes o número de anéis desse molusco. Assim, com medidas como peso, altura ou diâmetro é possível saber se o abalone em análise está maduro o suficiente para venda ou não.

Link para o código do estudo: https://colab.research.google.com/drive/1gTgwJdptLQas8XsgrBQSNrRUOhha1uyO?authuser=1#scrollTo=IkM52sGiddDJ

IV REFERÊNCIAS

[1] G. James, D. Witten, T. Hastie e R. Tibshirani, An Introduction to Statistical Learning with applications in R. Springer, 2013.

[2] M. Kuhn, K. Johnson, Applied Predictive Modeling. Springer, 2013.

[3] Business Case: Predicting Abalone Age. Disponível em: https://operational-machine-learning-pipeline.workshop.aws/assets/Model_Framing_Example.html. Acesso em 11 fev.2021

[4] Abalone Shell Data: Linear Models. Disponível em: https://charlesreid1.github.io/circe/Abalone%20-%20Linear%20Models.html. Acesso em 13 fev.2021

[5] Introduction to scikit-learn. Disponível em: https://bids.github.io/2015-06-04-berkeley/intermediate-python/03-sklearn-abalone.html. Acesso em 15 fev.2021

[6] Abalone Shell Data: Linear Models. Disponível em: https://charlesreid1.github.io/circe/Abalone%20-%20Linear%20Models.html. Acesso em 15 fev.2021

[7] Como avaliar seu modelo de classificação. Disponível em: https://medium.com/data-hackers/como-avaliar-seu-modelo-de-classificação-34e6f6011108. Acesso em 15 fev.2021.

[8] How to remove outliers using box-plot?, 2021. Disponível em: https://datascience.stackexchange.com/questions/54808/how-to-remove-outliers-using-box-plot. Acesso em 20 fev.2021

[9] Exploratory data analysis in Python, 2021. Disponível em: https://towardsdatascience.com/exploratory-data-analysis-in-python-c9a77dfa39ce. Acesso em 23 fev.2021

[10] Linear Regression Example, 2021. Disponível em: https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html. Acesso em 25 fev.2021

[11] How to Develop Ridge Regression Models in Python, 2021. Disponível em: https://machinelearningmastery.com/ridge-regression-with-python/. Acesso em 26 fev.2021

[12] Partial Least Squares Regression in Python, 2021. Disponível em: https://nirpyresearch.com/partial-least-squares-regression-python/. Acesso em 26 fev.2021