


Abalone Age Prediction

Alexandre Freire da Silva Osorio
Felipe Esmerino Gomes
Weld Lucas Cunha



Agenda

1. Objetivo
2. Contextualização do problema
3. EDA
4. Infraestrutura da solução
5. Projeto dos modelos de regressão
6. Resultados
7. Conclusões

Objetivo / contextualização

Objetivo

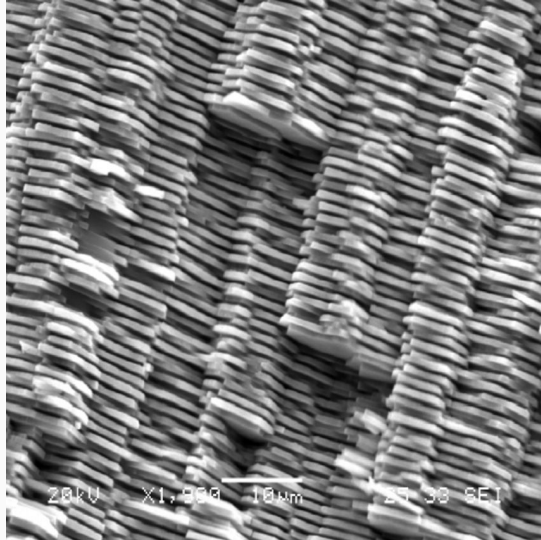
Analisar o dataset *Abalone*, em que a idade biológica do abalone é determinada contando o número de anéis em um microscópio. Somando o número de anéis + 1.5, consegue-se a idade aproximada.

Treinar um regressor que possibilite identificar a idade do abalone de acordo com as features disponíveis, visando melhoria no processo e ganho de tempo.

O regressor foi desenvolvido em computação de nuvem, utilizando serviços da AWS e o pacote PySpark.

Contextualização do problema

Os abalones são moluscos que vivem preferencialmente em águas frias. Em sua concha vemos uma seqüência de poros, seguidos por uma seqüência de pequenas elevações. Na face interior encontra-se uma madrepérola.



A idade do abalone é determinada cortando a concha através do cone, colorindo-a e contando o número de anéis em um microscópio - uma tarefa enfadonha e demorada.

Corte transversal da concha do abalone. A espessura de cada camada nacarada é de cerca de $0,5 \mu\text{m}$. (TAN, T. L.; WONG, D.; LEE, Paul. Iridescence of a shell of mollusk *Haliotis Glabra*. **Optics express**, v. 12, n. 20, p. 4847-4854, 2004.)

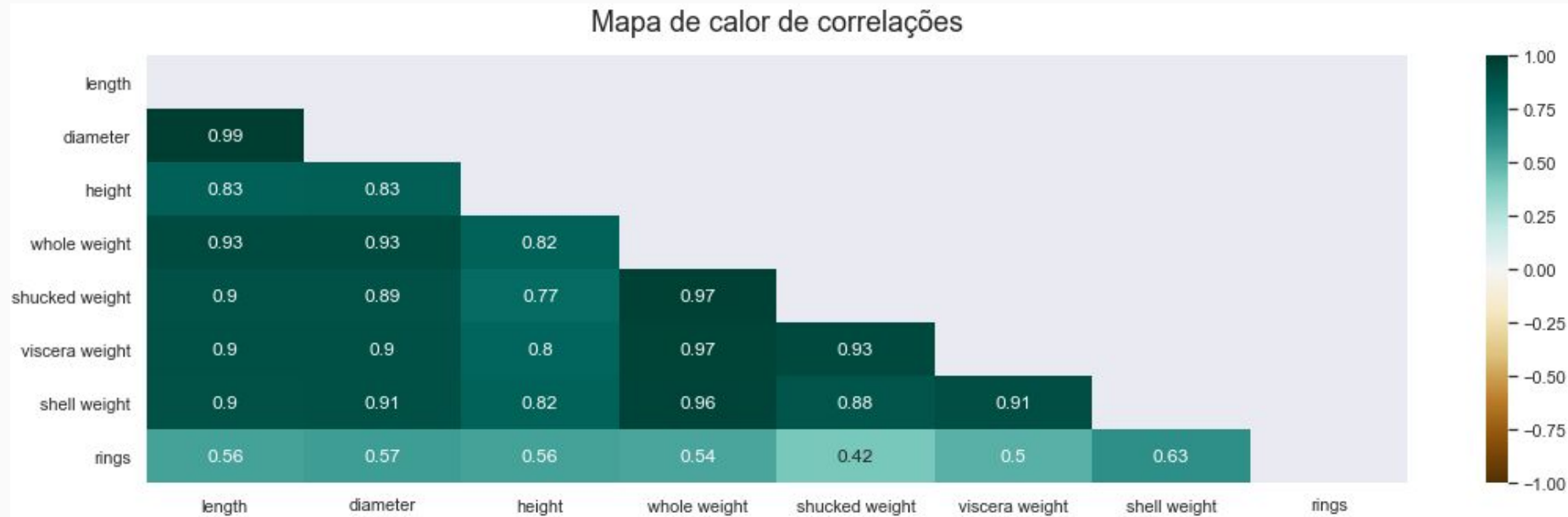
EDA

Descrição dos dados

O dataset possui 4178 amostras e 9 *features*

Feature	Tipo de dado	Amostra	Unidade	Descrição
sex	Categórico nominal	F(female), M(male), I(Infant)		
length	Numérico contínuo	0.075 a 0.815	mm	Maior medida da concha
diameter	Numérico contínuo	0.055 a 0.65	mm	Diâmetro perpendicular ao <i>length</i>
height	Numérico contínuo	0 a 1.13	mm	Altura com carne na concha
whole weight	Numérico contínuo	0.134 a 0.594	g	Peso todo o abalone
shucked weight	Numérico contínuo	0.0575 a 0.332	g	Peso somente da carne
viscera weight	Numérico contínuo	0.0285 a 0.116	g	Peso intestinal (após sangramento)
shell weight	Numérico contínuo	0.3505 a 0.1335	g	Peso da concha após drenada
rings	Numérico discreto	1 a 29		+1,5 dá a idade em anos

Análise de correlação de variáveis

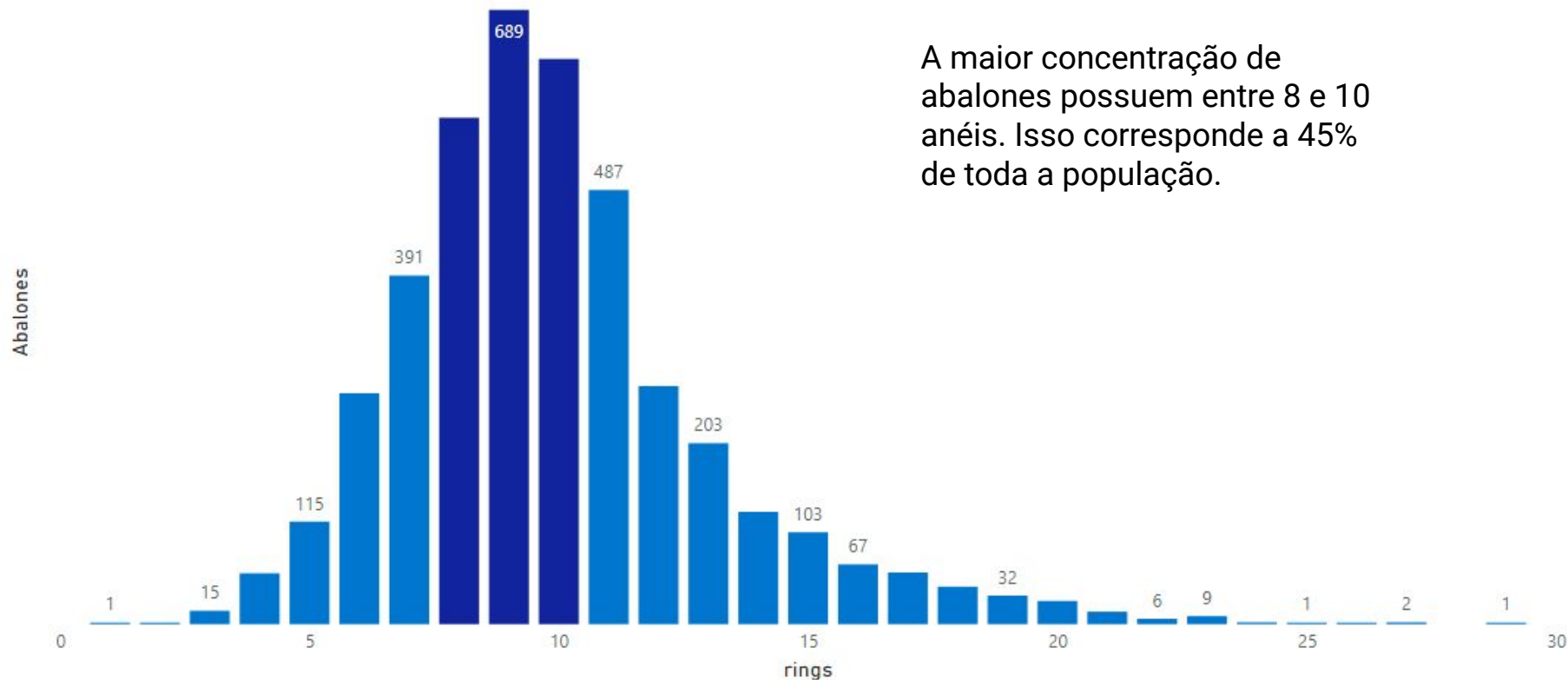


Existe uma relação linear positiva entre todas as variáveis.

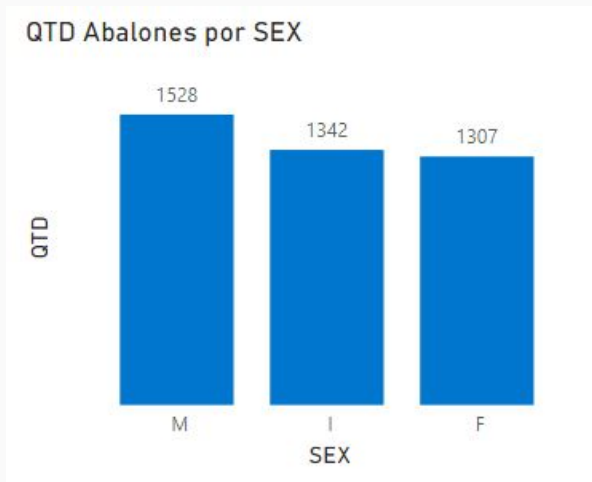
Rings tem correlação moderada com todas as variáveis.

Distribuição da quantidade de anéis

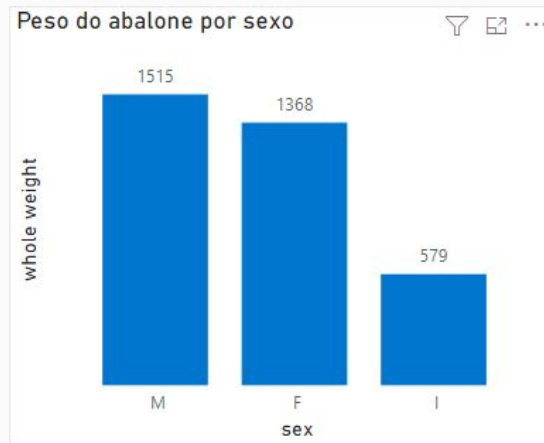
Contagem de abalone por rings



Distribuição por sexo

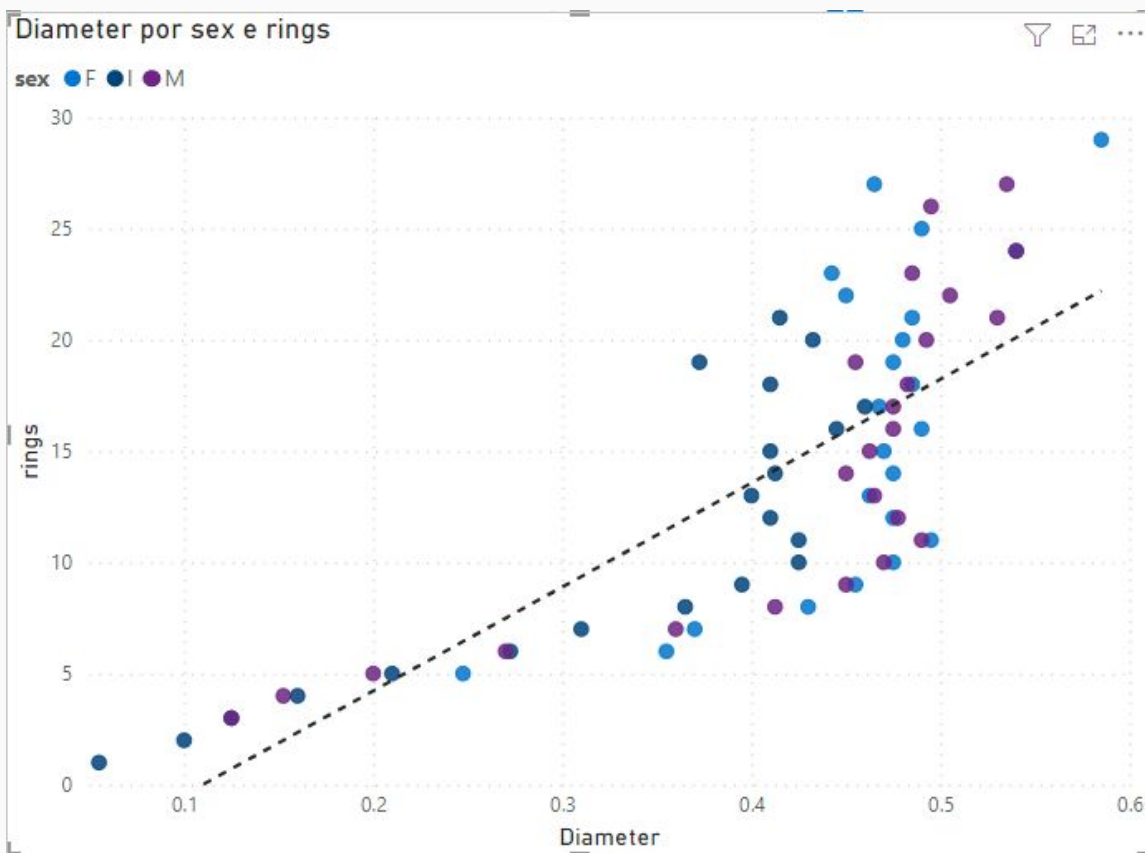


O gênero M(Machos) representa a maior quantidade de abalones representando 36% da amostra, seguido de I(Infantil) com 32% da amostra e F(Fêmea) com 31%.



Os machos além de serem a maior quantidade são os que possuem maior tempo de vida. Eles também são os mais indicados para consumo, levando em consideração o peso sem casca.

Dispersão por sexo e quantidade de anéis

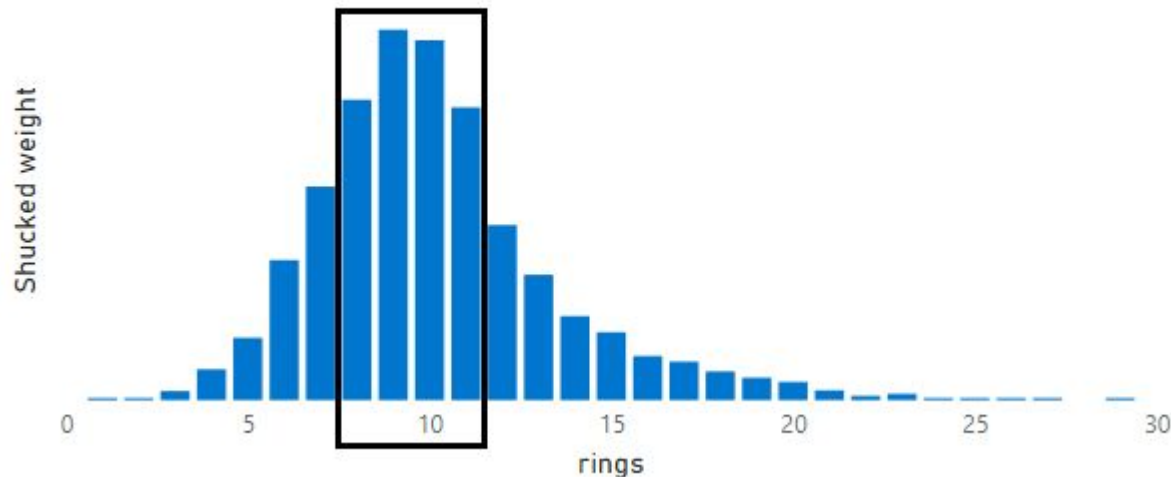


No gráfico de dispersão a esquerda, podemos notar que o sexo do abalone é identificado para todos os casos em que o diametro é maior que 0.5

Isso nos leva a percepção que o sexo do abalone é mais facilmente identificado a partir de uma certa idade/nro.rings.

Distribuição do peso

Distribuição do peso sem casca (com maior QTD carne)

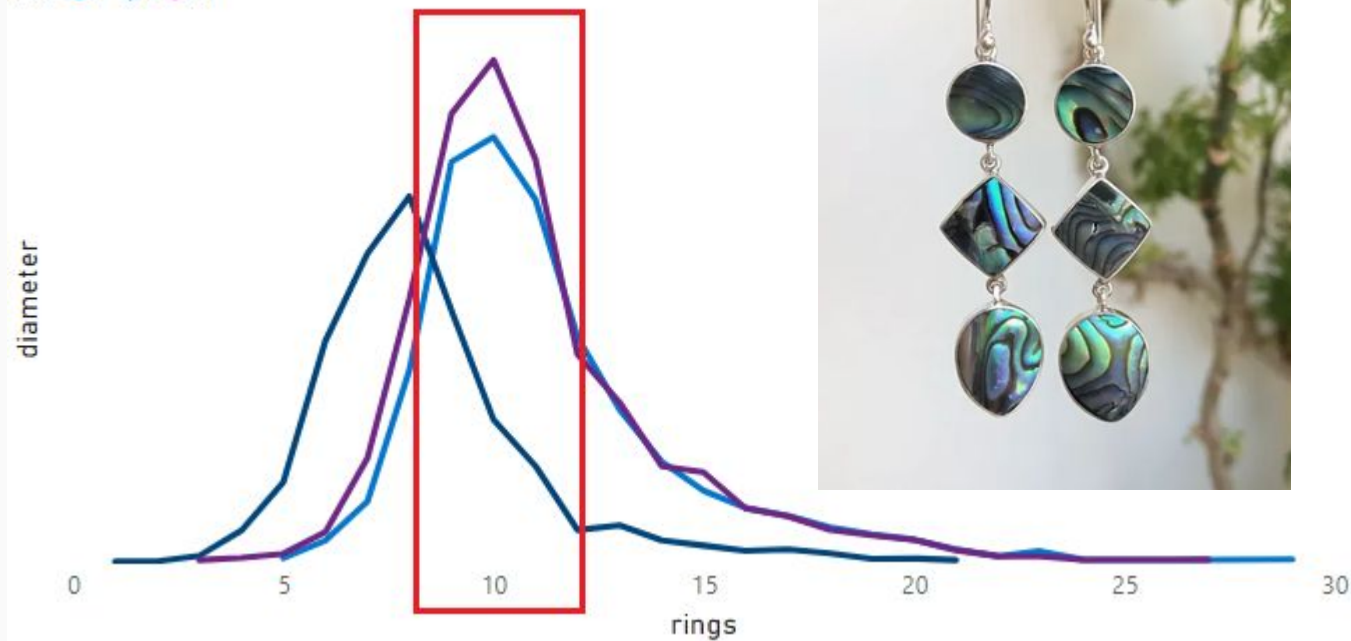


Considerando o consumo de abalone nos países asiáticos e controle da espécie. O gráfico indica que abalones com idade entre 8 e 11 são mais recomendados para consumo e preservação da espécie.

Distribuição do diâmetro

Distribuição do diâmetro por sexo e idade

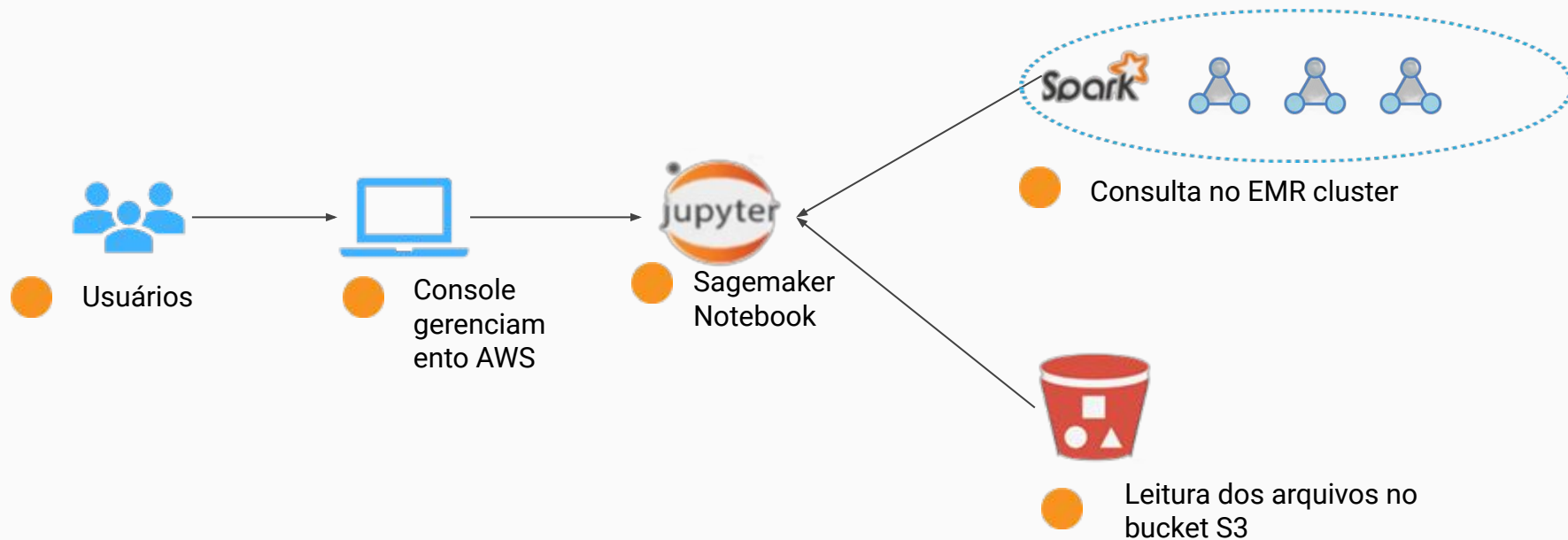
sex ● F ● I ● M



Considerando o uso de abalone para jóias, também sugerimos que seja usado com idade entre 8 e 11 anos, uma vez que as peças são polidas, cortadas e ajustadas para se adequarem.

Infraestrutura da solução

Workflow / Pipeline de Arquitetura



Infraestrutura

Config. Cluster

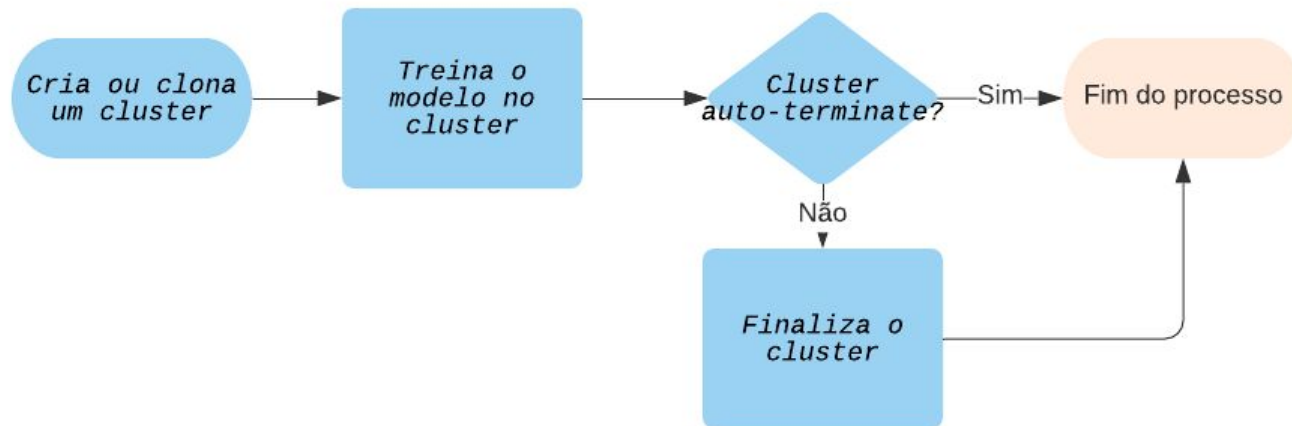
- Release label: emr-5.33.1
- Hadoop distribution: Amazon
- Applications: Hive, Pig, Hue, JupyterEnterpriseGateway e Spark.

Network & Hardware

- Região: us-east-2b
- Master: 1 mr.xlarge
- Core: 2 m5.xlarge

Node type & name	Instance type
MASTER Master - 1	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB
CORE Core - 2	m5.xlarge 4 vCore, 16 GiB memory, EBS only storage EBS Storage: 64 GiB

Ciclo de vida de um cluster



A Amazon recomenda seguir esse processo para diminuir o custo que um cluster ligado gera.

Projeto dos modelos

Projeto do regressor

A estratégia adotada foi a de projetar um regressor simples, chamado de *baseline*, e compará-lo com dois outros regressores, cujos hiperparâmetros foram ajustados usando *grid search*.

baseline: **linear regression** com os hiperparâmetros *default*

regressor 1: **linear regression** com ajuste fino dos seguintes hiperparâmetros

- **regParam** [0.5, 0.1, 0.05, 0.01, 0.005, 0.001, 0.0]: parâmetro de regularização. Quanto maior regParam, menor a variância (menos *overfitting*), ao custo de um maior *bias*.
- **fitIntercept** [False, True]: Se deve calcular o *intercept* para este modelo (False = *intercept* não será usado nos cálculos - ou seja, espera-se que os dados sejam centralizados)
- **elasticNetParam** [0.0, 0.01, 0.05, 0.1, 0.5, 0.8, 1.0]: 0: penalidade L2 | 1: penalidade L1

regressor 2: **random forest**, com ajuste fino dos seguintes hiperparâmetros

- **numTrees** [1,2,3,5,10,15,20]: número de árvores para treinar
- **maxDepth** [1,2,3,5,10]: profundidade máxima da árvore
- **minInstancesPerNode** [1,2,3,5,10]: Número mínimo de instâncias que cada ramo deve ter após a divisão.

Projeto do regressor

Baseline

- Normalização dos dados usando StandardScaler
- *fit* usando hiperparâmetros default

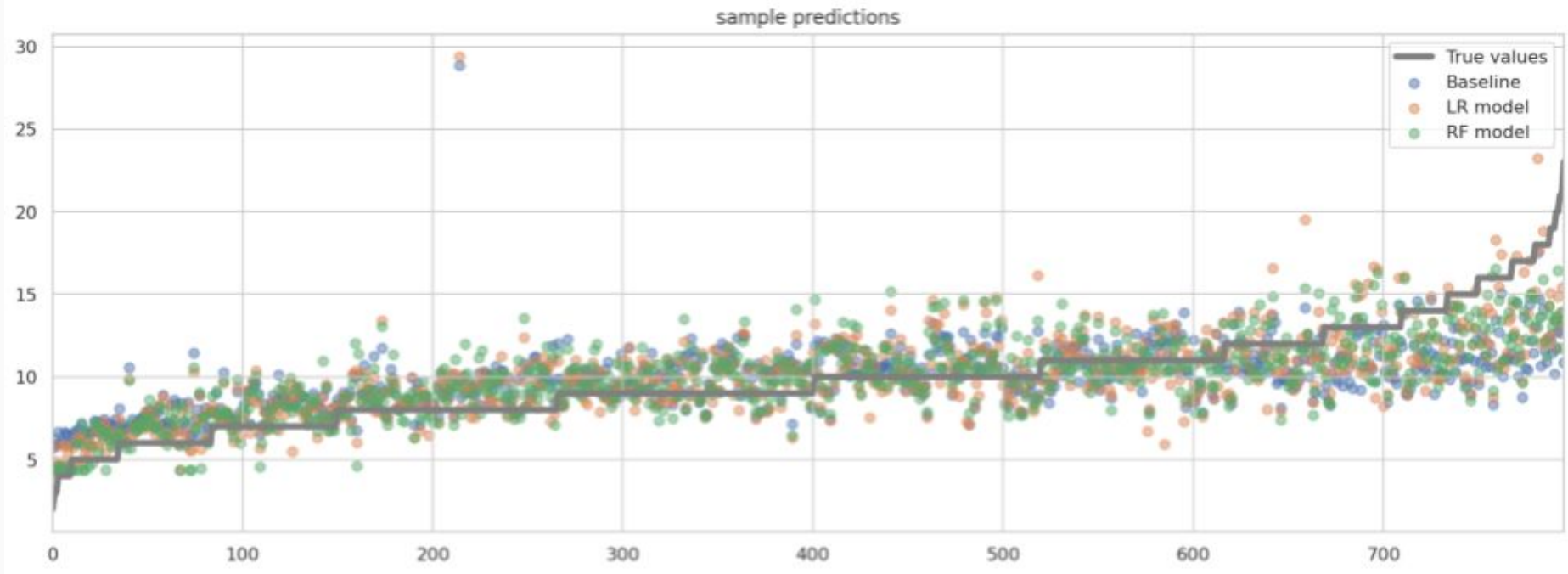
Modelos com ajuste fino

Pipeline

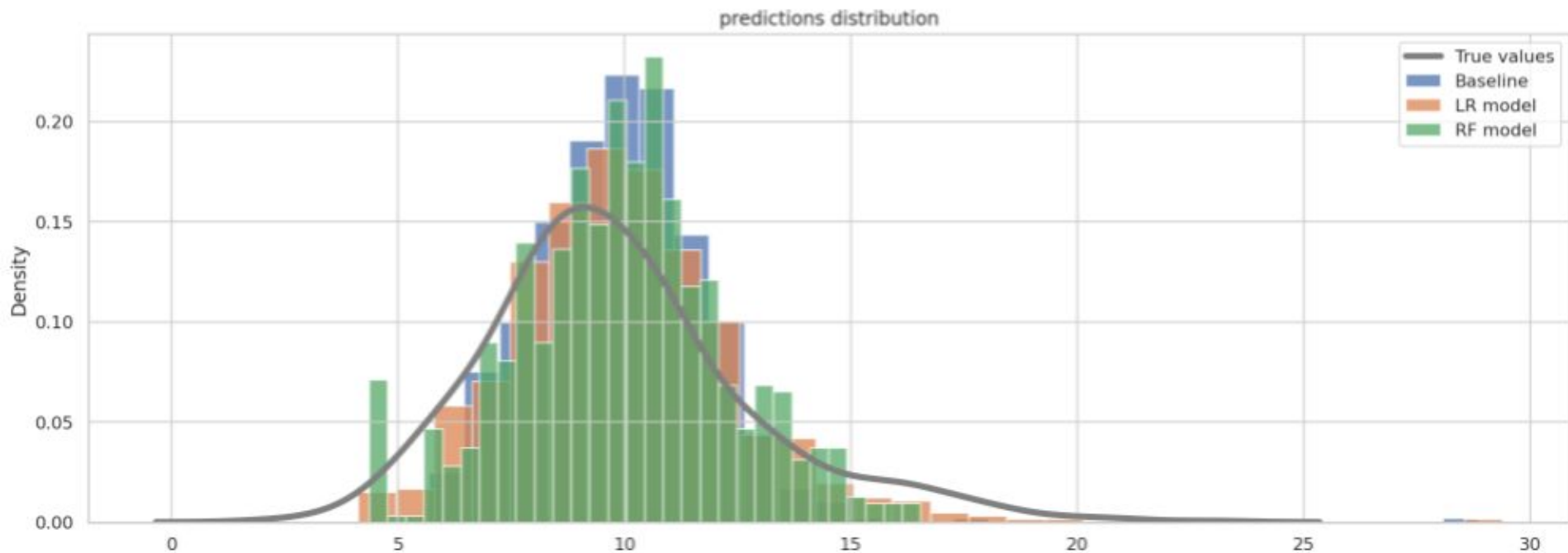
- estimator (Linear Regression e Random Forest)
- cross-validation com 5 *folds*
- avaliação - métricas MSE, MAE, RMSE, R2

Resultados

Resultados



Resultados



Comparando as principais métricas

Modelo	MAE	MSE	RMSE	R2
Baseline	1.7322	5.6597	2.3790	0.3619
Linear Regression	1.5670	4.8005	2.1910	0.4588
Random Forest	1.5103	4.1596	2.0395	0.5310

Conclusão

Comparando o uso do *PySpark* com o *sklearn*, chamou atenção a diferença na organização do dataset exigida pelo *PySpark*.

No *sklearn*, utilizado juntamente com o pacote *Pandas*, os dados são manipulados de forma mais amigável ao usuário, enquanto que o *Pyspark* força certas manipulações para que o dado fique no formato adequado ao seu padrão.

No notebook do *Sagemaker*, foi possível ler os dados diretamente do *S3*, os quais foram processados e posteriormente os modelos foram treinados através do *PySpark*.

Principais dificuldades: configurações dos sistemas (*EMR*, *Sagemaker*) na *AWS* e o aprendizado e a utilização do *Pyspark* em um ambiente cloud.

Muito Obrigado!

“In God we trust. All
others must bring data.”
W. Edwards Deming

