

Machine Learning na Prática

Modelos em Python

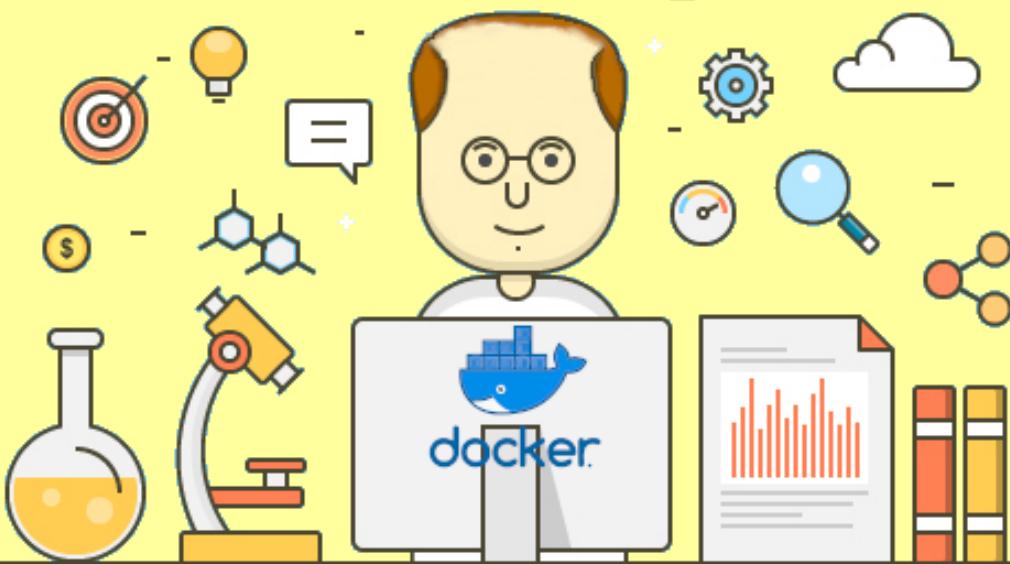
Fernando Anselmo

Copyright © 2020 Fernando Anselmo - v1.0

PUBLICAÇÃO INDEPENDENTE

<http://fernandoanselmo.orgfree.com>

É permitido a total distribuição, cópia e compartilhamento deste arquivo, desde que se preserve os seguintes direitos, conforme a licença da *Creative Commons 3.0*. Logos, ícones e outros itens inseridos nesta obra, são de responsabilidade de seus proprietários. Não possuo a menor intenção em me apropriar da autoria de nenhum artigo de terceiros. Caso não tenha citado a fonte correta de algum texto que coloquei em qualquer seção, basta me enviar um e-mail que farei as devidas retratações, algumas partes podem ter sido cópias (ou baseadas na ideia) de artigos que li na Internet e que me ajudaram a esclarecer muitas dúvidas, considere este como um documento de pesquisa que resolvi compartilhar para ajudar os outros usuários e não é minha intenção tomar crédito de terceiros.

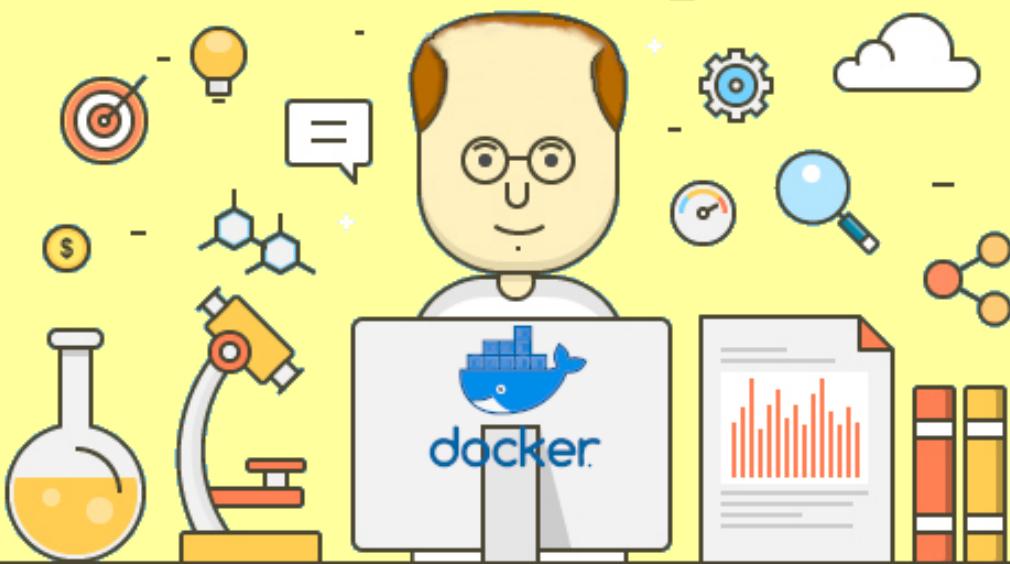


Sumário

1	Entendimento Geral	5
1.1	Do que trata esse livro?	5
1.2	O que é Machine Learning?	6
1.3	Formas de Aprendizado	7
1.3.1	Algoritmo Não Supervisionado	8
1.3.2	Algoritmos Supervisionados	8
1.3.3	Algoritmos de Aprendizagem por Reforço	9
1.4	Montagem do Ambiente	9
2	Conceitos Introdutórios	15
2.1	Termos da Estatística	15
2.2	Termos que devemos saber	16
2.3	Roteiro	18
2.4	Bibliotecas Utilizadas	19

2.5	Distribuição Gaussiana	20
2.6	Distribuição de Poisson	21
2.7	Distribuição Binomial	23
2.8	Feature Selection	25
2.8.1	Coeficientes de Coorelação	27
2.8.2	Chi Squared	27
2.8.3	RFE	27
2.8.4	Ensembles Methods	28
2.9	K Fold Cross Validation	29
2.10	Matriz de Confusão	31
2.10.1	Em valor ou percentual?	33
2.10.2	Na prática	34
2.11	Curva ROC e Valor AUC	36
2.11.1	Na prática	38
2.12	Terminamos?	40
3	EDA	41
3.1	Passos da EDA	41
3.2	Passo 1 - Entender os Dados	42
3.2.1	Localizar os Outliers	44
3.2.2	Tratar Atributos Categóricos	45
3.3	Passo 2 - Limpar os Dados	46
3.4	Passo 3 - Relacionamento entre os Atributos	49
3.5	Conclusão	52
4	Modelos Iniciais	53
4.1	K-Means	53

4.2	Aplicação da Técnica	54
4.3	Plotagem do Resultado do Modelo	56
4.4	K-Nearest Neighbors	57
4.4.1	Predição com K-Nearest Neighbors	59
4.5	Análise de Cluster	60
4.6	Clusterização Hierárquica	63
4.6.1	Clusterização Hierárquica versus K-Nearest Neighbors	66
4.7	Régressão Linear	67
4.7.1	Aplicar a Régressão Linear	68
4.7.2	Régressão Linear com mais de um Preditor	69
4.7.3	Régressão Linear e Limpeza dos Dados	70
4.7.4	Separação e treino	72
A	Considerações Finais	75
A.1	Sobre o Autor	76



1. Entendimento Geral

F Machine Learning é a habilidade de localizar padrões em Dados. (Gil Weinberg - Founding Director of Georgia Tech Center for Music Technology)

1.1 Do que trata esse livro?

Cada vez que leio um livro de **Machine Learning** tenho a sensação que o autor quer mostrar para seus leitores o quanto ele é um bom Matemático, coloca um monte de fórmulas com demonstrações (muitas delas parecem escritas em grego e usam inclusive letras gregas) é isso me faz pensar: *Será que quando indicar um livro para meus alunos vou querer que eles aprendam fórmulas ou que tenham uma base em que possam praticar?*

E graças a isso publiquei uma série de artigos sobre os mais variados modelos de *Machine Learning* na rede **Linkedin** e esses artigos foram a base para esse livro, não espere encontrar aqui muitos conceitos, apesar de ser obrigado a abordá-los ou muita coisa se perderia, mas a ideia aqui é ser totalmente prático.

Já ouvimos isso de prático tantas vezes, existe duas séries de livros especialistas denominadas: "*Hands On*" e "*Cookbook*". Aprecio e tenho muitos desses livros, porém sempre que desejo algo no primeiro caso é muito difícil encontrar bem separado e exposto da forma como queria e no segundo está fragmentado demais. O prático aqui será: Um modelo em linguagem Python, ler bases de dados, com o uso de suas possibilidades, seu treino e melhor forma de obter resultados. Sendo assim não espere encontrar nesse aulas básicas de Python.

1.2 O que é Machine Learning?

De forma bastante genérica, algoritmos de *Machine Learning* (Aprendizado de Máquina e doravante apenas **ML**) são uma mudança de paradigma da “programação tradicional” onde precisamos passar toda a heurística explicitamente para um novo conceito, onde ao invés de escrever cada ação que o algoritmos deve realizar, apenas passamos diversos exemplos e deixamos que o computador resolva (ou aprenda) quais são as “melhores” (menor custo) decisões. É também chamada de *Statistical Learning* (Autores Hastie, Tibshirani & Friedman 2009) utilizada para extrair um modelo a partir de um sistema de observações ou medidas. Sendo um campo relativamente novo da ciência composto de uma variedade de métodos (algoritmos) computacionais e estatísticos que competem entre si.

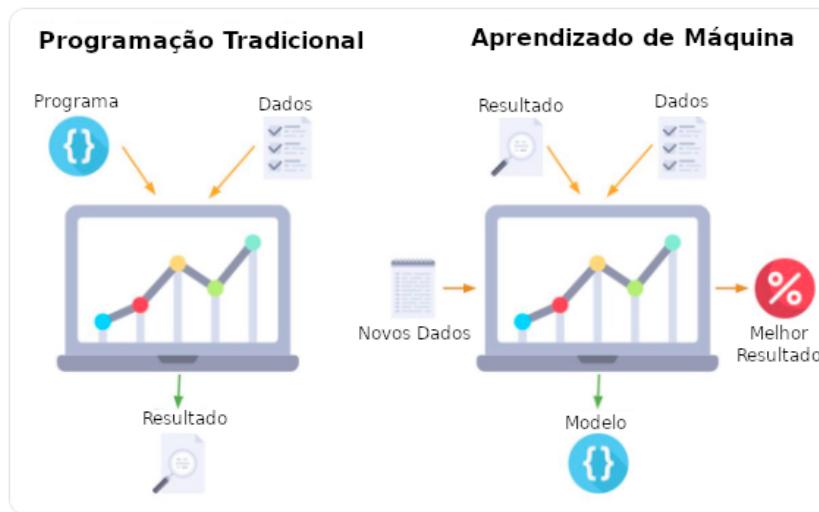


Figura 1.1: Programação Tradicional X Machine Learning

Ou seja, treinamos e melhoramos as respostas até que possam receber **novas observações**, transformamos isso em resultados sem que sejam explicitamente programáveis, isso é chamado "Modelo de ML". Interpretar esses modelos é entender como podemos transformar dados em informação útil. Em geral são classificados em:

- **Clusterização:** Encontrar uma estrutura de dados, summarização.
- **Régressão e Estimação:** Predição de valores contínuos.
- **Classificação:** Predição de um item de um caso de classe/categoria.
- **Associação:** frequentemente ocorre entre itens e eventos.

É muito importante conhecer vários deles e assim podemos decidir qual se ajusta melhor as observações que temos para treiná-los. Devemos ter em mente que ML pode ser bem diferente de **Estatística**. Aqui não estamos preocupados com inferência, causalidade e exogeneidade. ML está mais preocupada, quase que exclusivamente, em melhorar predições ou rapidamente localizar em um mar de informações a resposta de um problema.

Assim podemos pegar a mesma função, por exemplo **Régressão Logística** e analisar do ponto de vista da estatística, que interpretaria se os "betas" são significativos, se os "resíduos" têm uma distribuição normal ou analisar isso do ponto de vista de ML e descobrir como está a relação entre **Precisão** e **Recall**, ou qual

a ROC ou AUC do modelo¹.

1.3 Formas de Aprendizado

Quanto as formas de aprendizado se dividem em:

- **Não Supervisionado**, corresponde a um vetor de observações que é utilizado para observar padrões, tendências, verificar estruturas e descobrir relações.
- **Supervisionado**, além do vetor de observações, existe também uma resposta associada a cada questão.
- **Aprendizagem por Reforço**, uma ação ocorre e as consequências são observadas, assim a próxima ação considera os resultados da primeira ação. É um algoritmo dinâmico que parte do princípio "tentativa e erro".

Entender a diferença entre os dois tipos é bem simples, enviamos ao computador uma série de imagens sobre pratos de comida e não informamos absolutamente nada sobre elas, o máximo que acontecerá é a separação dessas em grupos similares. Imaginemos que em seguida mostramos a seguinte imagem:

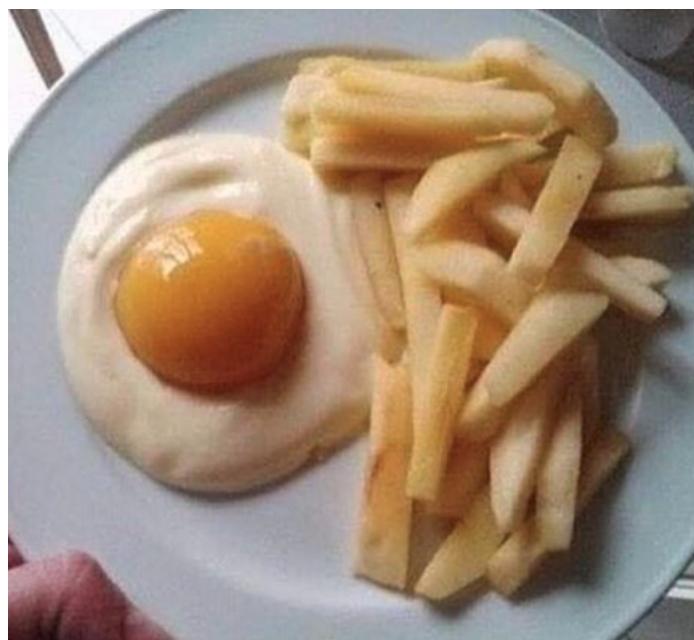


Figura 1.2: Nova informação

Qual será o prato de comida que o computador associará? Exatamente "Ovo frito e batatas fritas"². No segundo tipo de aprendizado além das imagens dizemos o que cada uma vem a ser, porém ao mostrar o mesmo tipo de imagem não pense que o computador consegue classificá-la corretamente, provavelmente pode se confundir mais uma vez (assim como nos confundimos com a imagem mostrada).

¹As curvas ROC e AUC estão entre as métricas mais utilizadas para avaliação em um modelo de ML.

²Apesar que se olhar mais de perto vemos que isso é iogurte, meio pêssego e tiras de maçã

A resolução com problemas de imagem é bem complicada³, pois conseguimos identificar a diferença devido a nossa experiência não apenas visual mas com informações sobre textura, forma e outras que o computador ainda está engatinhando. Por isso o estudo desses algoritmos é algo tão rico e como vimos anteriormente. É muito importante conhecer boa parte dos modelos e decidir qual trabalha melhor com as observações que temos para treiná-lo.

1.3.1 Algoritmo Não Supervisionado

Não envolve um controle direto, aqui o ponto principal do requisito são desconhecidos e ainda precisam ser definidos. Normalmente são usados para: explorar uma estrutura de informação; extrair informações desconhecidas sobre as observações e detectar padrões.

k-means: é o mais popular, usado para segmentar em categorias exclusivas com base em uma variedade de recursos, por exemplo clientes como seus hábitos de consumo ou casas com seu preço, localidade e área.

t-Distributed Stochastic Neighbor Embedding: t-SNE é utilizado para redução de dimensionalidade que é particularmente adequada para a visualização de conjuntos que possuem dados com alta dimensão.

Principal Component Analysis: PCA é usado para enfatizar a variação e destacar padrões fortes em um conjunto de dados, utilizado para facilitar a exploração e visualização dos dados.

Associate Rule Mining: ARM para encontrar padrões frequentes, associações, relações e correlação, normalmente utilizado em proporcionar análises para cesta de compras. Em termos gerais, é aplicado em várias situações como encontrar associação ou determinar um padrão frequente nos conjuntos de observações.

1.3.2 Algoritmos Supervisionados

Neste tipo temos os atributos alvo rotulados e do ponto de vista da máquina, esse processo é uma rotina de "conectar os pontos" ou achar similaridades entre as observações e o que ela representa. Para "alimentar" o algoritmo determinamos que tipo de resultado é desejado ("sim/não", "verdadeiro/falso", a projeção do valor das vendas, a perda líquida de crédito ou o preço da habitação).

Régressão linear: muito utilizado para prever resultados numéricos contínuos, como preços de casas ou ações, umidade ou temperatura de um local, crescimento populacional.

Régressão logística: É um classificador popular utilizado especialmente no setor de crédito para prever inadimplências de empréstimos.

k-Nearest Neighbors: KNN é um algoritmo usado para classificar as observações em duas ou mais categorias (*cluster*) amplamente usado na separação como preços de casas em grupos, por exemplo com base em preço, área, quartos e toda uma gama de outros preditores.

Support Vector Machines: SVM é um classificador popular utilizado na detecção de imagens e faces, além de aplicativos como reconhecimento de manuscrito.

³ Assim como análise da linguagem, chamada de NLP - *Natural Language Processing*

Tree-Based Algorithms: Algoritmos baseados em árvores, como *Random Forest* (florestas aleatórias) ou *Decision Tree* (árvores de decisão), são usados para resolver problemas de classificação e regressão.

Naive Bayes: Utiliza um modelo matemático de probabilidade para resolver problemas de classificação.

1.3.3 Algoritmos de Aprendizagem por Reforço

Tratam de situações onde a máquina começa a aprender por tentativa e erro ao atuar sobre um ambiente dinâmico. Desta maneira, não é necessário novos exemplos ou um modelo a respeito da tarefa a ser executada: a única fonte de aprendizado é a própria experiência do agente, cujo objetivo formal é adquirir uma política de ações que maximize seu desempenho geral.

Q-Learning: é um algoritmo de aprendizado baseado em valores. Esses tipos atualizam uma função com base em uma equação (particularmente neste caso de *Bellman*) matemática, ou então, com base em políticas, estima a função de valor para uma política gananciosa obtida a partir do último aprimoramento. Q-learning é um algoritmo de políticas. Significa que aprende o valor ideal, independentemente das ações do agente. Por outro lado, um aprendiz de política aprende o valor que está sendo executada pelo agente, incluindo as etapas de exploração e encontrará uma política ideal, leva em consideração a exploração inerente dessa.

Temporal Difference: TD é um agente que aprende por meio de episódios sem conhecimento prévio desse ambiente. Isso significa que a diferença temporal adota uma abordagem de aprendizado sem modelo ou supervisão. Podemos considerar isso como uma tentativa e erro.

Monte-Carlo Tree Search: MCTS é um método geralmente usado nos jogos para prever o caminho (movimentos) que a política deve seguir para alcançar a solução final vencedora. Jogos como Cubo de Rubik, Sudoku, Xadrez, Go, ou um simples Jogo da Velha têm muitas propriedades em comuns que levam ao aumento exponencial do número de possíveis ações que podem ser executadas. Esses passos aumentam exponencialmente à medida que o jogo avança. Idealmente, podemos prever todos os movimentos possíveis e seus resultados que podem ocorrer no futuro e assim aumentarmos a chance de ganhar.

Asynchronous Advantage Actor-Critic: A3C é um dos algoritmos mais recentes a serem desenvolvidos no campo de Reforço Profundo. Foi desenvolvido pelo *DeepMind* do Google e implementa um treinamento no qual vários *workers*, em ambiente paralelo, atualizam independentemente uma função de valor global - portanto "assíncrona". Um dos principais benefícios de ter atores assíncronos é a exploração eficaz e eficiente do espaço de estados.

1.4 Montagem do Ambiente

Vemos atualmente uma grande revolução em torno de *Data Science* (falamos em inglês para parecer algo chique) principalmente em torno as ferramentas que tem se atualizado a uma velocidade assustadora. Porém, essas atualizações constantes muitas vezes carregam problemas que podem afetar o seu Sistema Operacional. A pergunta é: Como ficar atualizado e seguro ao mesmo tempo? A única resposta coerente que consegui encontrar foi: Usar a conteinerização para resolver o problema.

Então o ideal é partir atrás de imagens prontas, visitar sites como HubDocker (<https://hub.docker.com>) e rezar para encontrar a imagem que nos atenda ou fazer algo melhor e personalizar a imagem.

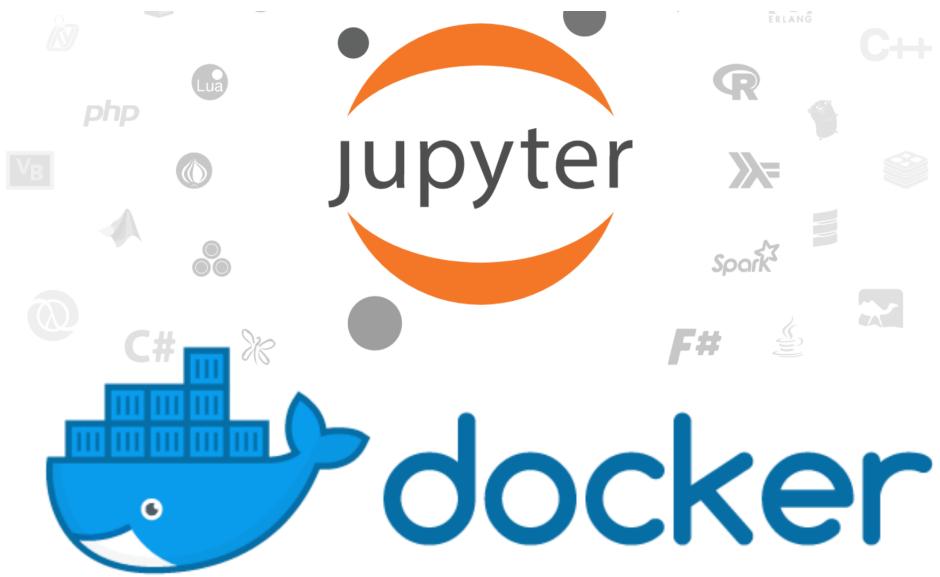


Figura 1.3: Docker e Jupyter para Data Science

Essa segunda alternativa é bem mais próxima a realidade de qualquer um que deseje trabalhar de modo efetivo com Ciência de Dados. A personalização de uma imagem no Docker não é um bicho de sete cabeças (no qual a partir do momento que cortamos uma das cabeças nasce mais duas, então sempre imaginei que seriam mais do que sete) mas algo que pode ser facilmente aprendido por qualquer um que "ao menos" saiba usar o terminal de comando do Linux.

Dica 1.1: Não sabe nem por onde começar com o Docker? Não se desespere, baixe um paper sobre o Docker gratuitamente na minha página no Academia.edu (<https://iesbpreve.academia.edu/FernandoAnselmo>).

Minha primeira personalização de Imagem⁴ surgiu quando descobri que as versões **Jupyter** não me atendiam por completo e **Anaconda** era grande demais. Iremos então trabalhar em uma imagem que possa conter um Jupyter mais adequado e uma Anaconda mais controlada. Com o seguinte resultado do arquivo **Dockerfile** (que obrigatoriamente deve ter esse nome):

```

1 # Base da Imagem
2 FROM ubuntu:19.10
3
4 # Adiciona o metadata para a imagem com o par: chave, valor
5 LABEL maintainer="Fernando Anselmo <fernando.anselmo74@gmail.com>"
6 LABEL version="1.2"
7
8 # Variaveis de Ambiente
9 ENV LANG=C.UTF-8 LC_ALL=C.UTF-8 PATH=/opt/conda/bin:$PATH
10
11 # Execucoes iniciais:
12 # Cria a pasta de ligacao
13 RUN mkdir ~/GitProjects && \
14 # instala os pacotes necessarios

```

⁴O objetivo não é termos uma imagem pequena, mas uma PERSONALIZÁVEL, se não deseja isso recomendo que acesse o tutorial disponível em <https://jcrist.github.io/conda-docker-tips.html>.

```
15 apt-get update && apt-get install --no-install-recommends --yes python3 && \
16 apt-get install -y wget ca-certificates git-core pkg-config tree freetds-dev \
    apt-utils && \
17 # Limpeza basica
18 apt-get autoclean -y && \
19 rm -rf /var/lib/apt/lists/* && \
20 # Configuracao do Jupyter
21 mkdir ~/.ssh && touch ~/.ssh/known_hosts && \
22 ssh-keygen -F github.com || ssh-keyscan github.com >> ~/.ssh/known_hosts && \
23 git clone https://github.com/bobbywlindsey/dotfiles.git && \
24 mkdir ~/.jupyter && \
25 mkdir -p ~/.jupyter/custom && \
26 mkdir -p ~/.jupyter/nbconfig && \
27 cp /dotfiles/jupyter/jupyter_notebook_config.py ~/.jupyter/ && \
28 cp /dotfiles/jupyter/custom/custom.js ~/.jupyter/custom/ && \
29 cp /dotfiles/jupyter/nbconfig/notebook.json ~/.jupyter/nbconfig/ && \
30 rm -rf /dotfiles && \
31 # Instalar o Anaconda
32 echo 'export PATH=/opt/conda/bin:$PATH' > /etc/profile.d/conda.sh && \
33 wget --quiet https://repo.anaconda.com/archive/Anaconda3-2020.02-Linux-x86_64.sh -O \
    ~/anaconda.sh && \
34 /bin/bash ~/anaconda.sh -b -p /opt/conda && \
35 rm ~/anaconda.sh && \
36 conda uninstall anaconda-navigator && \
37 conda update conda && \
38 conda update anaconda && \
39 conda install nodejs && \
40 conda update --all && \
41 # Limpeza basica no Anaconda
42 find /opt/conda/ -follow -type f -name '*.a' -delete && \
43 find /opt/conda/ -follow -type f -name '*.pyc' -delete && \
44 find /opt/conda/ -follow -type f -name '*.js.map' -delete && \
45 find /opt/conda/lib/python*/site-packages/bokeh/server/static -follow -type f -name \
    '*.js' ! -name '*.min.js' -delete && \
46 # Instalar os temas para o Jupyter
47 pip install msgpack jupyterthemes && \
48 jt -t grade3 && \
49 # Instalar os pacotes
50 conda install scipy && \
51 conda install pymssql mkl=2018 && \
52 pip install SQLAlchemy missingno json_tricks \
53 gensim elasticsearch psycopg2-binary \
54 jupyter_contrib_nbextensions mysql-connector-python \
55 jupyter_nbextensions_configurator pymc3 apyori && \
56 # Habilitar as extensoes do Jupyter Notebook
57 jupyter contrib nbextension install --user && \
58 jupyter nbextensions_configurator enable --user && \
59 jupyter nbextension enable codefolding/main && \
60 jupyter nbextension enable collapsible_headings/main && \
61 # Adicionar a extensao do vim-binding
62 mkdir -p $(jupyter --data-dir)/nbextensions && \
63 git clone https://github.com/lambdalisue/jupyter-vim-binding $(jupyter \
    --data-dir)/nbextensions/vim_binding && \
64 cd $(jupyter --data-dir)/nbextensions \
65 chmod -R go-w vim_binding && \
66 # Remover o que nao eh necessario
```

```

67 conda clean -afy && \
68 apt-get remove -y wget git-core pkg-config && \
69 apt-get autoremove -y && apt-get autoclean -y && \
70 # Adicionar o Git ao Jupyter Lab
71 pip install --upgrade jupyterlab-git && \
72 jupyter lab build
73
74 # Configurar o acesso ao Jupyter
75 WORKDIR /root/GitProjects
76 EXPOSE 8888
77 CMD jupyter lab --no-browser --ip=0.0.0.0 --allow-root
    --NotebookApp.token='data-science'

```

Tento manter sempre o script o mais documentado possível deste modo posso remover ou adicionar propriedades sem me incomodar muito. Para criarmos a imagem é muito simples, supondo que a localização do script esteja em uma pasta chamada docker-data-science, então na pasta anterior digitar o comando:

```
$ docker build -t fernandoanselmo/docker-data-science docker-data-science
```

Obviamente que "fernandoanselmo/docker-data-science" pode ser alterado para o nome que desejar, porém na essência isso cria uma imagem que contém além do sistema operacional uma versão completa do Jupyter (com a inclusão de várias funcionalidades) para a realização do nosso trabalho como Cientista de Dados.

Dica 1.2: Socorro. Mas isso é muito complicado e não entendo nada disso! Sem problemas, basta saltar toda essa parte de criação e construção para os próximos tópicos. Esta imagem foi publicada no **DockerHub** e pode ser baixado sem problemas.

Agora o comando:

```
$ docker run -d -t -i -v --privileged /dev/ttyACM0:/dev/ttyACM0 -v
~/Aplicativos/ipynb:/root/GitProjects --network=host
--name meu-jupyter fernandoanselmo/docker-data-science
```

Realiza a criação de um contêiner. Vejamos os detalhes: após a opção -v aparece a expressão "~/Aplicativos/ipynb", essa se refere a uma determinada pasta no sistema operacional onde serão armazenados os Notebooks produzidos. Abrir o navegador no endereço <http://localhost:8888> e obtemos o seguinte resultado:

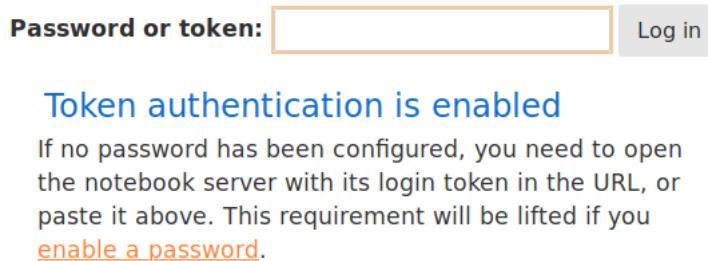


Figura 1.4: Jupyter solicitando o Token

A senha do token está definida na última linha do Script como "data-science", após informá-la o jupyter

está pronto para trabalharmos:

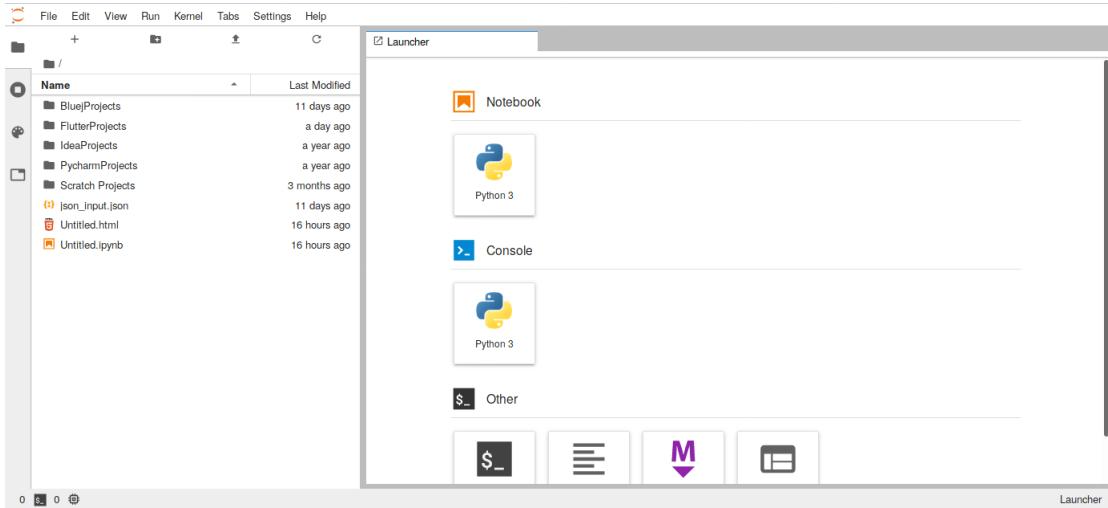


Figura 1.5: Jupyter Lab pronto

A maior vantagem que agora possuímos um ambiente com o **Jupyter Lab** completamente controlado incluindo temas e sincronização com o **Github**. Os próximos comandos são bem mais simples, tais como:

```
$ docker stop meu-jupyter
$ docker start meu-jupyter
```

Respectivamente para encerrar e iniciar o contêiner. Oh não! Agora estou preso, não posso mais fazer atualizações. Devemos entender que os contêineres são **dinâmicos**. Precisamos instalar o Keras/TensorFlow, acessar o contêiner (com ele já iniciado):

```
$ docker exec -it meu-jupyter /bin/bash
```

E instalamos normalmente, como se estivesse em uma máquina com o Ubuntu (com os poderes de superusuário):

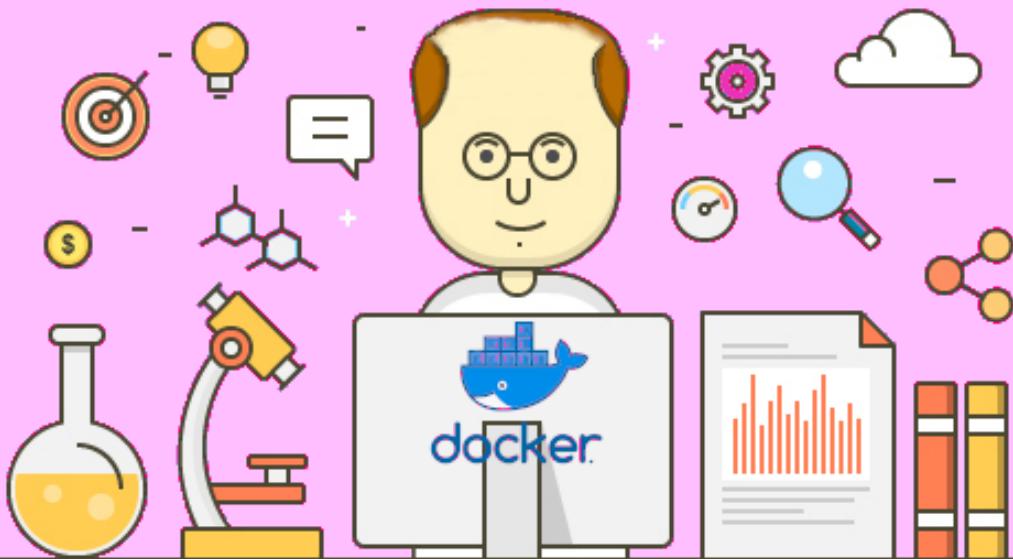
```
# conda install -c conda-forge keras
```

Pronto uma vez testado podemos optar por manter só nesse contêiner, ou então modificar o Script para ao criarmos novos contêineres e estes serão criados com essa funcionalidade embutida.

Verificar qual a versão do Python que está à nossa disposição, em uma célula do Notebook digite:

```
1 !python --version
```

Ao pressionarmos Ctrl+Enter será mostrada a versão 3.7.7. Agora podemos testar e executar qualquer ferramental sem nos preocuparmos em corromper a máquina. Talvez, no máximo, perdemos um contêiner.



2. Conceitos Introdutórios

F A melhor forma de prever o futuro é cria-lo. (Peter Drucker - Escritor e Pai da Administração Moderna)

2.1 Termos da Estatística

Devemos sempre possuir um vocabulário comum, e mesmo que não se entenda absolutamente nada de estatística (e tenha um Estatístico a sua disposição), o Cientista de Dados deve conhecer os seguintes termos:

- **População:** o conjunto constituído por todos os indivíduos que representam pelo menos uma característica comum, cujo comportamento interessa analisar (inferir). Por exemplo, se em uma empresa o diretor gostaria de saber se os funcionários estão satisfeitos com os benefícios oferecidos, a população de estudo são todos os funcionários dessa empresa. O conceito de população depende do objetivo de estudo.
- **Amostra:** um subconjunto, uma parte selecionada da totalidade de observações abrangidas pela população, através da qual se faz inferência sobre as características da população. Por exemplo, uma rádio tem o interesse de saber como está sua audiência com os ouvintes no trânsito. Não é possível perguntar a todos os motoristas que ouvem rádio qual é aquela que eles preferem. Então buscamos uma parte representativa dessa população, isto significa, perguntar somente a alguns motoristas qual rádio preferem escutar enquanto dirigem. Uma amostra tem que ser representativa, sua coleta bem como seu manuseio requer cuidados especiais para que os resultados não sejam distorcidos.
- **Elemento ou Variável de Interesse:** característica a ser observada em cada indivíduo da amostra. Componentes sobre o qual serão observadas ou medidas as características. Onde cada característica

corresponde a um tipo de variável. Por exemplo, se queremos estudar o índice de massa corporal (IMC) de alunos do ensino médio de uma cidade, tomaremos uma amostra dessa população, e mediremos a altura e o peso de cada aluno, já que o IMC é calculado como uma razão entre peso e o quadrado da altura do indivíduo. Nesse caso: peso e altura são as variáveis de interesse.

Tendo em vista as dificuldades de várias naturezas para se observar todos os elementos da população, tomaremos alguns deles para formar um grupo a ser estudado. Este subconjunto da população, em geral com dimensão menor, é denominado **amostra**.

Outros termos de interessante são as "Medidas Resumo", são elas:

- **Média:** É a soma das observações dividida pelo total. Este é o mais comum indicador de uma tendência central de uma variável.
- **Mediana:** Em uma lista ordenada de dados (rol), a mediana é o termo central.
- **Moda:** Refere-se ao termo que mais aparece em uma coleção. Sendo que: Amodal - rol que não tem nenhum valor que se repete; Bimodal - existem 2 valores que se repetem na mesma frequência, N-modal - existem n valores que se repetem na mesma frequência.
- **Variância:** Medida de dispersão dos dados para uma média. Quanto maior for mais distantes da média estão, e quanto menor for mais próximos estão da média.
- **Desvio Padrão (std):** É o resultado positivo da raiz quadrada da variância. Indica como fechado estão os dados em torno da média.
- **Amplitude:** Medida de dispersão da amostra, sendo uma simples diferença entre o menor e maior valor.
- **Coeficiente de Variação** usado para analisar a dispersão em termos relativos a seu valor médio quando duas ou mais séries de valores apresentam unidades de medida diferentes. Dessa forma, podemos dizer que o coeficiente de variação é uma forma de expressar a variabilidade dos dados excluindo a influência da ordem de grandeza da variável.

Probabilidade é uma medida que nos mostra qual o grau de um evento ocorrer novamente. Muita para da ciência de dados é baseada na tentativa de medir a probabilidade de eventos, desde as chances de um anúncio ser clicado até a falha de uma determinada peça em uma linha de montagem.

Dica 2.1: Para saber mais. Quer entender mais sobre Estatística, recomendo este interessante livro online e em constante evolução <http://onlinestatbook.com/2/index.html>.

2.2 Termos que devemos saber

Como toda ciência, existe um vocabulário comum que os cientistas falam e que devemos conhecer, palavras como: treinar um modelo, MSE, *Overfitting* ou bisbilhotagem fazem parte desse vocabulário. Então mesmo sendo este um livro prático, precisamos saber do que se tratam.

feature (atributo) pode ser considerado como **preditor** (ou explicativo) e **alvo** (ou dependente). Atributos preditores são valores de entrada para um algoritmo enquanto que os dependentes de saída (resultado).

Um problema que pode ser trabalhado com técnicas de ML é quando existe um padrão e não é fácil defini-lo. Fazendo uma analogia com estatística, é preciso usar um conjunto de observações (amostra) para descobrir um processo subjacente (probabilidade). Os dados são selecionados e aplicado ao modelo que possui um grau de aprendizado (acurácia). Descobrir um padrão não é memorizar *Overfitting*, pois ao injetar novos dados o modelo deve ser capaz de prever o resultado.

Training (treinar) um modelo significa ensinar o modelo a determinar bons valores para todos os pesos e o viés de exemplos rotulados. *Loss* (perda) é a penalidade por uma má previsão em um único exemplo, que pode ser quantificada por métricas como *Mean Square Error* (MSE). Esse é um cálculo conhecido como *loss function* (função de perda) sendo que representa uma medida de quão bom um modelo de previsão faz em termos de ser capaz prever o resultado correto.

Existem três princípios em ML:

- **Navalha de Occam:** o modelo mais simples que se ajusta aos dados também é o mais plausível.
- **Viés de amostragem:** se os dados são amostrados de forma tendenciosa, então a aprendizagem também produz resultados tendenciosos.
- **Bisbilhotagem de dados:** se um conjunto de dados afeta qualquer etapa do processo de aprendizado, então a capacidade do mesmo conjunto de dados avaliar o resultado foi comprometida (se usar propriedades adequadas do conjunto de dados, pode assumir relações antes de começar a escolha de modelos).

Teoria Vapnik–Chervonenkis (VC) tenta explicar o processo de aprendizagem do ponto de vista estatístico. A dimensão VC é uma medida da complexidade a um espaço de funções que pode ser aprendido por um algoritmo de classificação estatística, é definido como a cardinalidade do maior conjunto de pontos que o algoritmo pode quebrar. Se tiver poucos dados, modelo mais simples funcionam melhor e complexo é um desastre.

Desigualdade de Hoeffding fornece um limite superior na probabilidade que a soma das variáveis independentes não se desvie do valor esperado acima de uma certa quantia. Quando generalizada, modelos muito sofisticados (com grande número de hipóteses) perdem a ligação entre uma amostra e o total, isso implica em memorizar ao invés de aprendizado.

Na aprendizagem supervisionada, um algoritmo é construído através de muitos exemplos e tenta encontrar um modo de minimizar a perda das ligações. Uma parte dos dados deve ser separada para treinamento e outra para teste. Essa separação ocorre para estimar o erro de predição do modelo e ter a certeza de que não está superajustado. Geralmente, é adotada uma razão de 80% dos dados para treinamento e 20% teste.

Gradient Descent (SGD) calcula o gradiente da curva de perda (inclinação) e indica qual o caminho para minimizar o erro (utilizar um gráfico da perda em função do peso), redefine o peso - quando há vários pesos, o gradiente é um vetor de derivadas parciais com relação aos pesos. Esse vetor gradiente (com direção e magnitude) é multiplicado por um escalar conhecido como taxa de aprendizado (*learning rate* ou *step size*) para determinar o próximo ponto. A heurística de buscar mínimo local começa de diferentes valores para então encontrar o melhor dentre todos. A aleatoriedade permite localizar outros mínimos locais, e assim podemos determinar o melhor, que pode ser o mínimo global.

Step (etapa) é o número total de iterações do treinamento. Uma etapa calcula a perda de um lote e usa esse valor para modificar os pesos do modelo uma vez. *Batch size* (tamanho do lote) é o número total de exemplos (escolhidos aleatoriamente) que foram usados para calcular o gradiente em uma única etapa (conjunto de dados inteiro).

Epoch (época) é um ciclo completo de treinamento e um gráfico de erro/perda indica como é a evolução do modelo. Cada *epoch* pode conter resultados piores, mas o melhor é guardado e exibido sempre *pocket algorithm*, a não ser que surja um melhor. Esse gráfico também pode ser construído destinado a um conjunto de dados para teste. Se a curva de perda em função das *epoch* cair muito para o treinamento e não acontecer o mesmo com os dados de teste, revela que o modelo está treinado em excesso para essas amostras de treino e não consegue prever um novo conjunto de amostras.

Overfitting acontece quando o ruído é ajustado também, só que ruído não tem padrão a ser descoberto. O ruído pode ser aleatório ou determinístico, relacionado a informação que existe, mas os dados, ou o conjunto de hipóteses, não conseguem promover o aprendizado dessa informação. *Weight decay* (decaimento de peso) é uma constante regularizadora, que restringe os pesos na direção da função alvo e melhora o ajuste e reduz o *overfitting*.

Normalization (normalizar) significa converter valores de recursos numéricos (como, 100 a 900) em um intervalo padrão (como, 0 a 1 ou de -1 a +1). Um exemplo desse tipo de cálculo é: $valor - min \div max - min$. Se o conjunto de dados é composto de múltiplos atributos, a normalização gera benefícios, como ajudar a descida gradiente a convergir mais rapidamente, evitar que um número torne-se NaN por exceder seu limite de precisão e ajudar o modelo a aprender os pesos apropriados para cada atributo. A normalização deve ocorrer depois de separar os dados de treinamento e teste.

Standard (padronização) é um cálculo importante para comparar medidas com diferentes unidades. Um exemplo desse tipo de cálculo é o "Z score": $(valor - media) \div std$. A padronização deve ser feita antes da normalização.

2.3 Roteiro

Devemos saber que um roteiro para ML passa pelas seguintes fases:

- Entender a questão de negócio (falar a mesma língua).
- Identificar a causa raiz.
- Coletar as observações.
- Realizar uma estatística descritiva e compreender suas características.
- Aplicar uma limpeza: entender os atributos e possíveis valores faltantes.
- Ler os dados, se necessário renomear ou corrigir os atributos.
- Procurar por incoerências.
- Criar novas variáveis para modelar melhor o fenômeno, se necessário.
- Levantar Hipóteses sobre o Comportamento do Negócio.
- Definir o tipo do problema: Regressão, Classificação ou Clusterização.
- Realizar uma Análise Exploratória de Dados.
- Quais hipóteses são falsas e quais são verdadeiras?
- Quais as correlações entre os atributos preditores e alvo?
- Verificar se as variáveis possuem o mesmo peso, em termos de importância, para o modelo.
- Aplicar diferentes algoritmos de ML.
- Comparar os modelos, sob a mesma métrica de performance, o mais apropriado para análise.
- Garantir que o modelo não possui *Overfit*, ou seja, memorização ao invés de aprendizado.
- Escrever os valores de previsão e seu intervalo de confiança do arquivo de teste.
- Descrever uma breve explicação do raciocínio da solução.
- Anotar as respostas encontradas.
- Detalhar as possíveis soluções para o problema.

São vários passos, devemos segui-los para garantir o sucesso das nossas análises. Algumas partes são bem preocupantes, entre elas, as que envolvem os conceitos mais básicos, a tendência é sempre saltá-los sem dar muita importância. Porém são cruciais.

2.4 Bibliotecas Utilizadas

Já temos o *Jupyter* e devemos conhecer o ferramental que iremos trabalhar. Abrir uma célula e digitar os seguintes comandos:

```
1 import numpy as np
2 import pandas as pd
3 import matplotlib
4 import pylab
5 import matplotlib.pyplot as plt
6 from scipy import stats
7 from scipy.stats import norm
8 from numpy.random import seed
9 from numpy.random import randn
10 import matplotlib.colors
11 import scipy
12 import sklearn
13 import seaborn as sns
14 import statsmodels.api as sm
15 %matplotlib inline
```

Na próxima célula digitar:

```
1 print('numpy: {}'.format(np.__version__))
2 print('pandas: {}'.format(pd.__version__))
3 print('scipy: {}'.format(scipy.__version__))
4 print('matplotlib: {}'.format(matplotlib.__version__))
5 print('sklearn: {}'.format(sklearn.__version__))
6 print('seaborn: {}'.format(sns.__version__))
7 print('statsmodel: {}'.format(sm.__version__))
8 print('\nMais informações:\n\n', sklearn.show_versions())
```

E obtemos como resposta as versões das principais bibliotecas que utilizaremos:

- **NumPy** *Numerical Python*. Seu recurso mais poderoso é a matriz n-dimensional. Também contém funções básicas de álgebra linear, transformações de *Fourier*, recursos avançados de números aleatórios e ferramentas para integração com outras linguagens de baixo nível, como Fortran, C e C++.
- **Pandas** *Python and Data Analysis* para operações e manipulações de dados estruturados. Amplamente utilizada para coleta e preparação de dados.
- **SciPy** *Scientific Python*. Tem por base a NumPy. É uma das bibliotecas mais úteis para diversos módulos de ciência e engenharia de alto nível, como matrizes discretas, álgebra linear, otimização e dispersão.
- **Matplotlib** *Math Plotting Library*. para gerar uma grande variedade de gráficos, tais como histogramas, gráfico de linhas e mapa de calor.

- **SkLearn ou Scikit Learn** *SciPy Toolkit Learn.* Tem por base a NumPy, SciPy e Matplotlib. Contém muitas ferramentas para aprendizado de máquina e modelagem estatística, incluindo classificação, regressão, clusterização e redução de dimensionalidade.
- **Seaborn Statistical Data Visualization.** Geração de gráficos atraentes e informativos em Python. Tem por base a Matplotlib. Visa tornar a visualização uma parte central da exploração e compreensão dos dados.
- **Statsmodels Statistical Models.** Permite explorar dados, estimar modelos estatísticos e executar testes. Uma lista extensa de estatísticas descritivas, funções de plotagem e de resultados estão disponíveis para diferentes tipos de dados e para cada estimador.

Caso exista a necessidade de uma atualização, por exemplo da biblioteca **Scikit-learn**, abrir uma nova célula e digitar o comando:

```
!pip install --upgrade scikit-learn
```

Dica 2.2: Bibliotecas Utilizadas. Obtenha uma boa referência sobre essas pois não serão tratadas em nível básico neste livro. Obviamente suas funções serão comentadas mas partiremos do pressuposto que sabemos lidar com estas.

2.5 Distribuição Gaussiana

A Distribuição Gaussiana (também conhecida como Curva Normal) é uma curva em forma de sino e supõe-se que, durante qualquer valor de medição, siga uma distribuição normal de um número igual de medidas acima e abaixo do valor médio. Para entendermos a distribuição normal, é importante conhecer as definições de média, mediana e moda. Se uma distribuição for normal, seus valores são os mesmos. No entanto, o valor da média, mediana e moda podem ser diferentes resultando em uma distribuição inclinada (não gaussiana).

Abrir uma nova célula e iremos gerar um gráfico com uma Distribuição Gaussiana ideal:

```
1 xAxis = np.arange(-3, 3, 0.001) \\
2 yAxis = norm.pdf(xAxis, 0, 1) \\
3 plt.plot(xAxis, yAxis) \\
4 plt.show()
```

Obtemos uma Distribuição Normal perfeita. Normalmente será muito difícil na realidade a curva ser tão perfeita assim, então podemos nos aproximar mais da realidade e gerar a partir de uma amostra com dados aleatórios:

```
1 data = 5 * randn(10000) + 50 \\
2 plt.hist(data, bins=100) \\
3 plt.show()
```

Obtemos a seguinte Distribuição Normal:

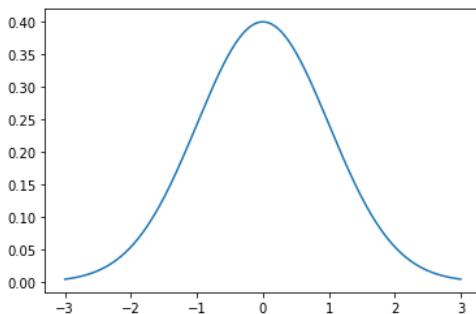


Figura 2.1: Curva da Distribuição Normal

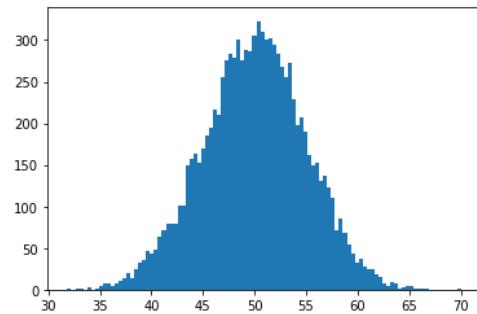


Figura 2.2: A partir de dados randômicos

Também podemos a partir dessa segunda, calcular seus valores principais:

```

1 print('Média: %.3f' % np.mean(data))
2 print('Mediana: %.3f' % np.median(data))
3 print('Moda:', stats.mode(data))

```

2.6 Distribuição de Poisson

Pronuncia-se *Poassom*, é uma distribuição de Probabilidade Discreta para um atributo X (qualquer) que satisfaça as seguintes condições:

- O experimento consistem em calcular quantas vezes (k) que um evento ocorre em um dado intervalo.
- A probabilidade de ocorrer é a mesma em cada intervalo.
- O experimento resulta em resultados que podem ser classificados como **sucessos** ou **fallhas**.
- O número médio de sucessos (μ) que ocorre em uma região especificada é conhecido.
- A probabilidade de um sucesso é proporcional ao tamanho da região.
- A probabilidade de um sucesso em uma região extremamente pequena é praticamente zero.
- O número de ocorrências em um intervalo é independente.

Na prática é uma função de ponto percentual, *Poisson* não existe na forma fechada simples é computado numericamente. Essa é uma distribuição discreta, definida apenas para valores inteiros de X, a função de ponto percentual não é suave da mesma forma que a função de ponto percentual normalmente é para uma distribuição contínua. Começamos com a importação das bibliotecas necessárias:

```

1 import numpy as np
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 %matplotlib inline

```

A NumPy já nos provê a função necessária para vermos como é o gráfico:

```

1 dp = np.random.poisson(lam=3, size=(1000))
2 plt.hist(dp)
3 plt.show()

```

Geramos 1.000 números aleatórios com uma ocorrência 3 e obtemos como resultado:

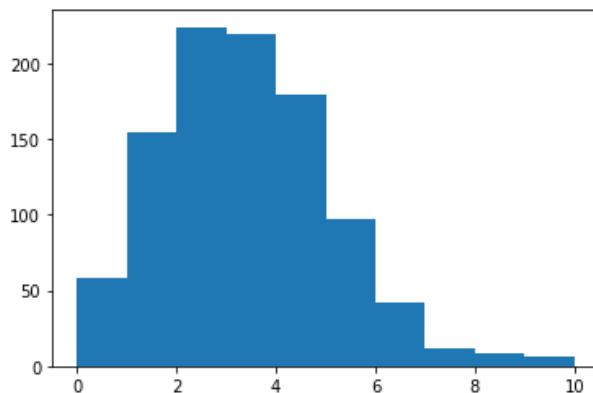


Figura 2.3: Distribuição de Poisson na MatPlotLib

É chamado "Variável Aleatória de Poisson" o número de sucessos resultantes em um experimento de Poisson. Porém esse gráfico fica bem mais apresentável com o uso da Seaborn:

```
1 sns.distplot(dp)
2 plt.show()
```

E obtemos:

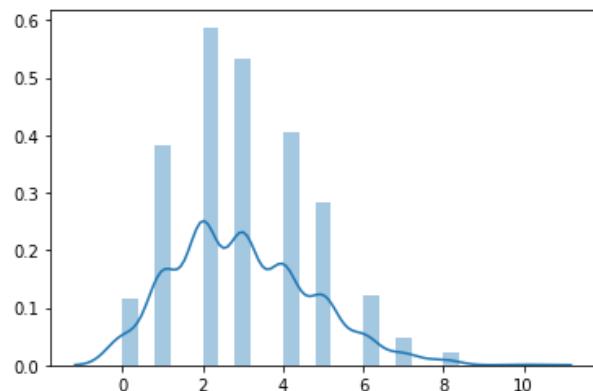


Figura 2.4: Distribuição de Poisson na Seaborn

Dica 2.3: Historicamente falando. 1946, o estatístico britânico RD Clarke publicou "*Uma Aplicação da Distribuição de Poisson*", com sua análise dos acertos de bombas voadoras (mísseis V-1 e V-2) em Londres durante a II Guerra Mundial. Algumas áreas foram atingidas com mais frequência do que outras. Os militares britânicos desejavam saber se os alemães estavam atacando esses distritos (os acertos indicavam grande precisão técnica) ou se a distribuição era por acaso. Se os mísseis fossem de fato apenas alvos aleatórios (dentro de uma área mais geral), os britânicos poderiam simplesmente dispersar instalações importantes para diminuir a probabilidade de serem atingidos.

Se acertarmos a escala, surpreendentemente, ao colocar uma Distribuição Normal e uma de Poisson sobrepostas:

```
1 sns.distplot(np.random.normal(loc=50, scale=7, size=1000), hist=False, label='normal')
```

```

1 sns.distplot(np.random.poisson(lam=50, size=1000), hist=False, label='poisson')
2 plt.show()

```

Obtemos como resultado:

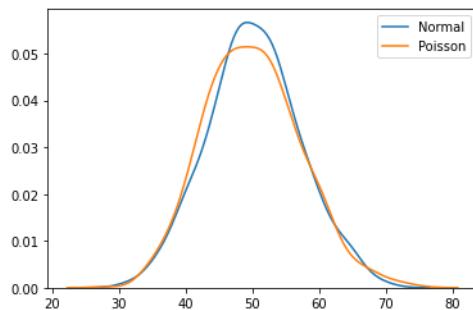


Figura 2.5: Distribuição Normal e de Poisson

E claramente percebemos a similaridade entre ambas. Um caso curioso que ocorre com a Distribuição de Poisson é a "cauda longa", sabemos que em qualquer comércio existe os produtos que mais possuem saída e aqueles outros que estão ali para "compor estoque". Por exemplo:

```

1 ax = sns.distplot(np.random.poisson(lam=3, size=10000), bins=27, kde=False)
2 ax.set(xlabel='Mercadorias', ylabel='Unidades Vendidas')
3 plt.show()

```

Obtemos como resultado:

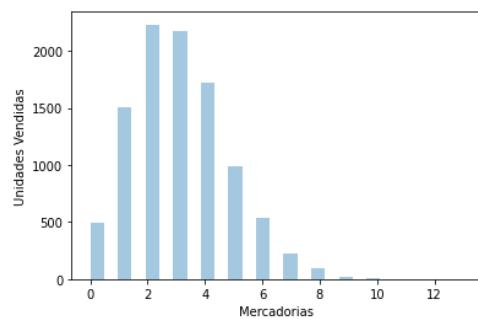


Figura 2.6: A Cauda Longa

A partir do número 6 temos um decrescimento constante, o que demonstra exatamente o que acontece com algumas mercadorias.

2.7 Distribuição Binomial

Neste tipo de distribuição apenas dois resultados são possíveis: sucesso ou fracasso, ganho ou perda, vitória ou perda e a probabilidade é exatamente a mesma para todas as tentativas. No entanto, os resultados

não precisam ser igualmente prováveis, e cada estudo é independente um do outro. Podemos ver sua curva com 1.000 lançamentos de 10 possibilidades

```
1 sns.distplot(np.random.binomial(n=10, p=0.5, size=1000), hist=True)
2 plt.show()
```

Obtemos como resultado:

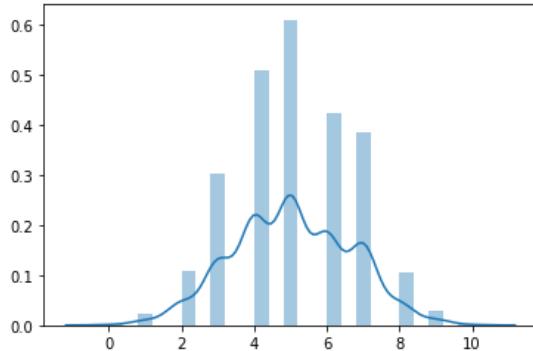


Figura 2.7: Distribuição Binomial

A principal diferença para a normal é que essa é contínua, enquanto que a binomial é discreta, mas se houver pontos de dados suficientes, serão bastante semelhantes (inclusive com a Poisson).

```
1 sns.distplot(np.random.normal(loc=50, scale=7, size=1000), hist=False, label='Normal')
2 sns.distplot(np.random.poisson(lam=50, size=1000), hist=False, label='Poisson')
3 sns.distplot(np.random.binomial(n=100, p=0.5, size=200), hist=False, label='Binomial')
4 plt.show()
```

Obtemos como resultado:

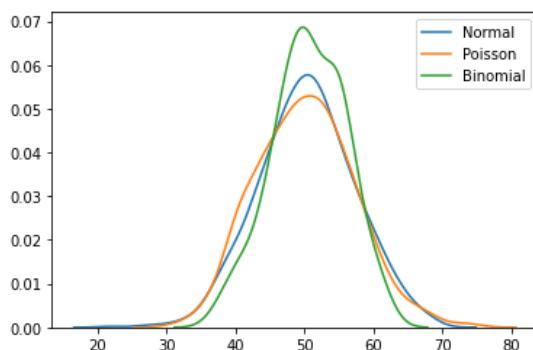


Figura 2.8: Mesmo gráfico 3 Distribuições

Mas e na prática? Imaginemos que 90% dos passageiros reservados chegam de fato para sua viagem. Suponhamos um transporte que pode conter 45 assentos. Muitas vezes as companhias praticam "excesso de reservas"(conhecido por *Overbooking*) isso significa que vende mais passagens do que os assentos disponíveis. Isso se deve ao fato de que às vezes os passageiros não aparecem um assento vazio representa perda. No entanto, ao reservar em excesso corre o risco de ter mais passageiros do que assentos.

Com esses riscos em mente, uma companhia decide vender mais de 45 bilhetes. Supondo que desejam manter a probabilidade de ter mais de 45 passageiros para embarcar no voo abaixo de 0,4 quantas passagens podem vender?

Para resolvermos isso precisamos da SciPy, e procedemos a seguinte codificação:

```
1 from scipy.stats import binom_test
2 for i in range(45, 51):
3     print(i, "-", binom_test(x=45, n=i, p=0.9, alternative='greater'))
```

Obtemos como resultado:

```
45 - 0.008727963568087723
46 - 0.04800379962448249
47 - 0.13833822255419037
48 - 0.27986215181073293
49 - 0.449690866918584
50 - 0.6161230077242769
```

O parâmetro *alternative* que indica a hipótese, podemos usar:

- *greater* maior que
- *less* menor que
- *two-sided* maior e menor que (valor padrão)

E podemos vender **48 passageiros**. E agora sabemos porque todas as companhias de transporte praticam *overbooking*. Para vermos isso graficamente:

```
1 sns.distplot(np.random.binomial(n=48, p=0.9, size=400), label='Binomial')
```

Obtemos como resultado:

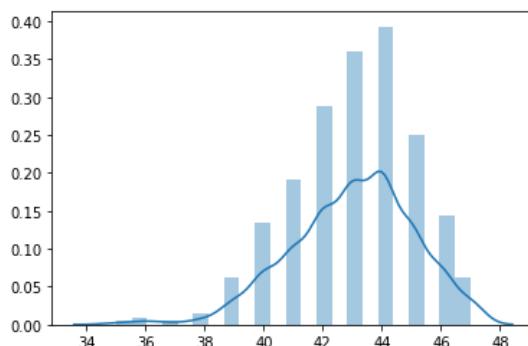


Figura 2.9: Gráfico do Overbooking

2.8 Feature Selection

Seleção de Atributos, Variáveis, Recursos ou Subconjuntos de Variáveis, são os vários nomes do processo que realiza a seleção de atributos relevantes para a construção de modelos.

O objetivo da *Feature Selection* é o de selecionar os atributos que funcionam melhor como **preditores**.

Essa etapa ajuda a reduzir o *overfitting*, aumenta a acurácia do modelo e reduz o tempo de treinamento.

São os seguintes métodos que podemos utilizar:

- **Filter Methods:** usam medidas estatísticas para atribuir uma pontuação para cada atributo. Estas são classificadas de modo a serem mantidas ou removidas do modelo. Normalmente usamos testes univariados que consideram a independência do atributo explicativo com o dependente. Exemplo: *chi squared*, pontuações com Coeficiente de Correlação.
- **Wrapper Methods:** selecionam um conjunto de atributos, onde diferentes combinações são preparadas, avaliadas e comparadas. Um modelo preditivo é usado para avaliar a combinação de atributos e dar uma nota a partir da acurácia do modelo. Exemplo: RFE.
- **Embedded Methods:** aprendem quais atributos contribuem melhor para a acurácia do modelo no momento de sua construção. Exemplo: Métodos de Penalização, Algoritmos Lasso, *Elastic NET* e *Ridge Regression*.

Dica 2.4: Bases de Dados. Para TODOS os exemplos neste livro, utilizamos bases de dados que estão disponibilizadas em <https://github.com/fernandoans/machinelearning/tree/master/bases>. Salvo qualquer observação em contrário.

Importação das bibliotecas utilizadas:

```

1 import pandas as pd
2 from sklearn.feature_selection import SelectKBest
3 from sklearn.feature_selection import f_classif, mutual_info_classif
4 from sklearn.feature_selection import chi2
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.feature_selection import RFE
7 from sklearn.ensemble import RandomForestClassifier
8 %matplotlib inline

```

Os classificadores estão contidos na Scikit-Learn e utilizaremos como nossa fonte de dados o arquivo chamado **pima-indians-diabetes.csv**. Criar um DataFrame para esta:

```

1 colnames = ['gest', 'glic', 'sang', 'skin', 'insul', 'mass', 'familia', 'idade',
   'conf']
2 df = pd.read_csv('pima-indians-diabetes.csv', names=colnames)
3 df.head()

```

Esta base são dados do **Instituto Nacional de Diabetes, Doenças Digestivas e Renais** sendo de pacientes do sexo feminino. Consiste de vários atributos que são considerados como preditivos (explicativos) e um atributo alvo (dependente). Os explicativos incluem o número de gestações que a paciente teve, seu IMC, nível de insulina, idade e outras informações como possível causa da Diabetes em Pacientes. Essa base possui 798 linhas sem quaisquer presença de nulos, verificar com: `df.info()`

O atributo *conf* é alvo, se o paciente em questão teve ou não diabetes. Precisamos isolá-lo:

```

1 X = df.drop(['conf'], axis=1)
2 y = df['conf']

```

Para os outros desejamos conhecer: qual (ou quais) atributos se comportam melhor para o nosso modelo?

2.8.1 Coeficientes de Coorelação

Como primeira forma aplicamos os testes estatísticos:

```
1 f_classif1 = SelectKBest(score_func=f_classif, k=4)
2 fit1 = f_classif1.fit(X,y)
```

Os tipos para o **SelectKBest**, neste caso, são:

- **f_classif**: mais adequado quando os atributos são numéricos e o dependente é categórico.
- **mutual_info_classif**: mais adequando quando não existe uma dependência linear entre os preditores e o alvo.
- **f_regression**: para resolver problemas de regressão.

Visualizar os atributos selecionados:

```
1 cols = fit1.get_support(indices=True)
2 df.iloc[:,cols]
```

Indica que número de gestações (*gest*), concentração de glicose no plasma (*glic*), índice de massa corporal (*mass*) e *idade* seriam os melhores candidatos.

2.8.2 Chi Squared

Essa é uma outra forma de medir a dependência, "elimina"os atributos com a maior probabilidade de serem independentes da classe e, portanto, irrelevantes para a classificação:

```
1 test2 = SelectKBest(chi2, k=4)
2 fit2 = test2.fit(X, y)
```

Visualizar os atributos selecionados:

```
1 cols = fit2.get_support(indices=True)
2 df.iloc[:,cols]
```

Percebemos que aconteceu uma mudança nos atributos: *glic*, *mass* e *idade* continuam, porém ao invés de *gest* aparece a quantidade administrada de insulina (*insul*).

2.8.3 RFE

Recursive Feature Elimination é um método que remove o(s) recurso(s) mais fraco(s) até que o número especificado de recursos seja atingido. Sendo assim é necessário informar ao RFE o número de atributos caso contrário reduz pela metade esse valor de acordo com o número de atributos das observações:

```

1 model = LogisticRegression(max_iter=2000, solver='lbfgs')
2 rfe = RFE(model, n_features_to_select=4)
3 fit3 = rfe.fit(X, y)

```

Visualizar as variáveis selecionadas:

```

1 cols = fit3.get_support(indices=True)
2 df.iloc[:,cols]

```

Como resultado obtemos quase a mesma combinação anterior e aparece mais a variável *familia* que indica se existem casos de diabetes na família.

2.8.4 Ensembles Methods

Métodos de agrupamento¹, como o algoritmo *Random Forest*, podem ser usados para estimar a importância de cada atributo, e retorna um valor para cada um, quanto mais alto esse, maior sua importância.

Aplicar o algoritmo:

```

1 model = RandomForestClassifier(n_estimators=10)
2 model.fit(X, y)
3 RandomForestClassifier(bootstrap=True, class_weight=None,
4   criterion='gini', max_depth=None, max_features='auto',
5   max_leaf_nodes=None, min_impurity_decrease=0.0, min_impurity_split=None,
6   min_samples_leaf=1, min_samples_split=2,
7   min_weight_fraction_leaf=0.0, n_estimators=10,
8   n_jobs=None, oob_score=False, random_state=None,
9   verbose=0, warm_start=False)

```

E podemos gerar uma visualização:

```

1 feature_importancia = pd.DataFrame(model.feature_importances_,
2   index = X.columns, columns=['importancia']).sort_values('importancia',
3   ascending=False)
3 feature_importancia

```

Porém é preferível vê-las através de um gráfico:

```

1 feature_importancia.plot(kind='bar')

```

Obtemos como resultado:

¹São métodos que utilizam vários algoritmos de aprendizado para obter um melhor desempenho preditivo.

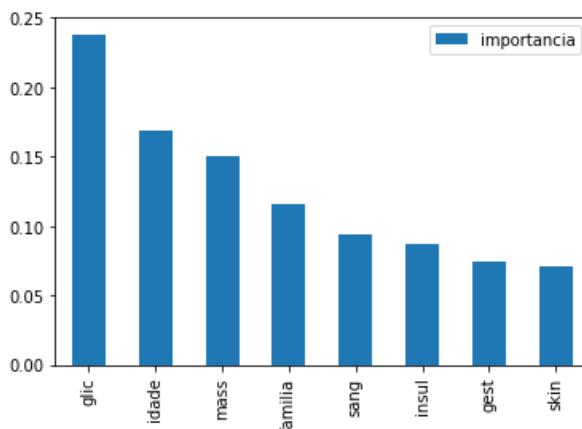


Figura 2.10: Grau de Importância dos Atributos

E afinal de contas com tudo o que vimos. Qual Método utilizar?

- Usar *RFE* caso tenha recursos computacionais para isso.
- Ao trabalhar com Classificação e os atributos forem numéricos, utilizar *f_classif* ou *mutual_info_classif*.
- Ao trabalhar com Regressão e os atributos forem numéricos, utilizar *f_regression* ou *mutual_info_regression*.
- Ao trabalhar com atributos categóricos utilizar *Chi Squared*.

2.9 K Fold Cross Validation

Vimos como encontrar os melhores atributos preditores para trabalhar, porém verificamos um pequeno problema: qual a forma em se escolher o modelo ideal² para nossa base de dados? O que fazemos é testar um modelo várias vezes obtendo pedaços diferentes de treino e teste a cada vez.

Usaremos a mesma base com Casos de Diabetes. Realizar a importação das bibliotecas:

```

1 from pandas import read_csv
2 import matplotlib.pyplot as plt
3 from sklearn.model_selection import KFold
4 from sklearn.model_selection import cross_val_score
5 from sklearn.linear_model import LogisticRegression
6 from sklearn.tree import DecisionTreeClassifier
7 from sklearn.neighbors import KNeighborsClassifier
8 from sklearn.discriminant_analysis import LinearDiscriminantAnalysis
9 from sklearn.naive_bayes import GaussianNB
10 from sklearn.svm import SVC
11 %matplotlib inline

```

Em seguida colocar as observações em um *DataFrame*:

```

1 colnames = ['gest', 'glic', 'sang', 'skin', 'insul', 'mass', 'familia', 'idade',
             'conf']

```

²"Ideal"é apenas um conceito para designar qual terá uma melhor acurácia (não usemos como sinônimo para perfeito ou o melhor)

```
2 df = read_csv('pima-indians-diabetes.data.csv', names=colnames)
3 df.head()
```

Para não gerar códigos repetitivos, criar uma lista com todos os modelos que executaremos:

```
1 models = []
2 models.append(('LR', LogisticRegression()))
3 models.append(('LDA', LinearDiscriminantAnalysis()))
4 models.append(('KNN', KNeighborsClassifier()))
5 models.append(('DT', DecisionTreeClassifier()))
6 models.append(('NB', GaussianNB()))
7 models.append(('SVM', SVC()))
```

Trabalharemos com Regressão Logística (LR), Análise Descriminante (LDA), *K-Nearest Neighbors* (KNN), Árvore de Decisão (DT), *Gaussian* tipo NB e SVM. Avaliamos a acurácia de cada um dos modelos:

```
1 acuracia = []
2 for sigla, modelo in models:
3     kfold = KFold(n_splits=10, random_state=7, shuffle=True)
4     resultado = cross_val_score(modelo, X, Y, cv=kfold, scoring='accuracy', n_jobs=-1)
5     acuracia.append(resultado.mean())
6     print("%s: %f (%f)" % (sigla, resultado.mean(), resultado.std()))
```

O objeto **KFold** criado executa 10 vezes (definido em *n_splits*) cada um dos modelos mantendo um estado de aleatoriedade de 7 registros. O método *cross_val_score* realiza todo o trabalho que tivemos para executar um modelo, dividir a base de dados em treino e teste obtendo o score de cada execução e obtemos o mesmo resultado final, a cada resposta podemos definir o número de vezes a executar (definido em *n_jobs*) neste caso será executado somente 1 vez (o valor é -1).

Obtemos como resultado a lista de 10 valores para cada um dos modelos, para facilitar, trazemos a média (que também será adicionada a lista *acuracia*) e o desvio padrão:

LR: 0.778640 (0.047350)
 LDA: 0.766969 (0.047966)
 KNN: 0.710988 (0.050792)
 DT: 0.696770 (0.046741)
 NB: 0.759142 (0.038960)
 SVM: 0.760458 (0.034712)

Também podemos ter isso de forma gráfica:

```
1 fig = plt.figure(figsize = (8,4))
2 axes = fig.add_axes([0.0, 0.0, 1.0, 1.0])
3 axes.set_title('Comparação dos Modelos');
4 axes.bar([item[0] for item in models], acuracia)
5 plt.show()
```

E obtemos como resultado:

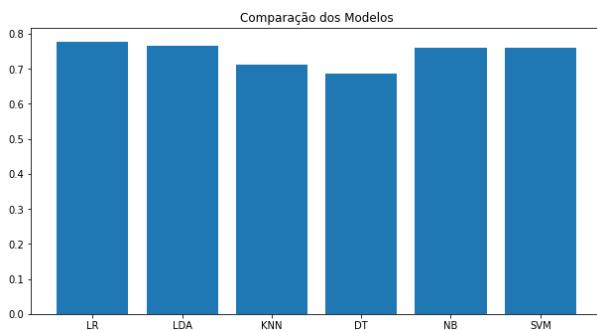


Figura 2.11: Melhor Acurácia dos Modelos

2.10 Matriz de Confusão

A Matriz de Confusão nos auxilia a medir a precisão do nosso modelo. A acurácia é uma boa medida porém um tanto falha, suponhamos que estejamos para realizar um trabalho com dados de pacientes que desejam saber a possibilidade de desenvolver uma determinada doença ou não. São quatro respostas possíveis que o nosso modelo pode prover:

- **Verdadeiro Positivo (TP - true positive)**: no conjunto da classe real, o resultado foi correto. O paciente desenvolveu a doença e o modelo previu que iria desenvolver. (T & T)
- **Falso Positivo (FP - false positive)**: no conjunto da classe real, o resultado foi incorreto. O paciente desenvolveu a doença e o modelo previu que não iria desenvolver. (T & F - Erro tipo 2)
- **Falso Negativo (FN - false negative)**: no conjunto da classe real, o resultado foi incorreto. O paciente não desenvolveu a doença e o modelo previu que iria desenvolver. (F & T - Erro tipo 1)
- **Verdadeiro Negativo (TN - true negative)**: no conjunto da classe real, o resultado foi correto. O paciente não desenvolveu a doença e o modelo previu que não iria desenvolver. (F & F)

Os casos 2 e 3 são erros no modelo, sendo que o 2 é considerado um pior tipo. Para começar com a nossa prática importar as bibliotecas necessárias:

```

1 import numpy as np
2 import pandas as pd
3 import matplotlib.pyplot as plt
4 import seaborn as sn
5 from sklearn.model_selection import train_test_split
6 from sklearn.metrics import confusion_matrix
7 from sklearn.linear_model import LogisticRegression
8 %matplotlib inline

```

A biblioteca `sklearn.metrics` na versão 0.22 ganhou o método **confusion_matrix**. Somente para entendermos como funciona a Matriz de Confusão, suponhamos que foi realizado um teste com base em características de pacientes, existe uma possibilidade de desenvolver (D) ou não (S) uma determinada doença:

```

1 y_true = pd.Series(['D', 'D', 'D', 'D', 'D', 'D', 'S', 'S', 'S', 'S', 'S', 'S'])
2 y_pred = pd.Series(['D', 'D', 'S', 'D', 'D', 'D', 'S', 'D', 'S', 'S', 'D', 'S'])

```

A série contendo `y_true` é o resultado real (ou verdadeiro) e `y_pred` foi o resultado que o modelo disse que iria ocorrer. Se fossemos somente pela acurácia obtemos 13 respostas sendo 4 delas erradas, porém como saber se o modelo está realmente agindo bem e o mais importante onde está se confundindo?

```
1 conf = confusion_matrix(y_true, y_pred)
2 print(conf)
```

Ao usarmos o método `confusion_matrix` obtemos a seguinte Matriz que nos auxilia a responder essas questões:

```
[[5 1][3 4]]
```

O eixo **X** da Matriz representa o que foi predito e **y** o que realmente aconteceu pelo modelo. Nas diagonais obtemos as corretas, cinco que estão Doentes e quatro que estão sadios e o modelo acertou. Porém, três não estão doentes mas foi previsto que estariam (se pensarmos a notícia não é tão ruim - do ponto de vista para o paciente, por isso esse é o erro tipo 1) e um está doente porém prevemos que não estaria (péssima notícia, por isso esse é o erro tipo 2). Visualizar resultados assim pode ser bem complexo, então tentemos uma adaptação do Mapa de Calor da biblioteca **Seaborn**:

```
1 data = {
2     'Ocorreu': y_true,
3     'Predito': y_pred
4 }
5 df = pd.DataFrame(data, columns=['Ocorreu', 'Predito'])
6 conf = pd.crosstab(df['Ocorreu'], df['Predito'], rownames=['Ocorreu'],
7                     colnames=['Predito'])
7 res = sn.heatmap(conf, annot=True, fmt=' .0f', annot_kws={"size":12},
8                   cmap=plt.cm.Blues)
8 plt.show()
```

E obtemos como resultado:

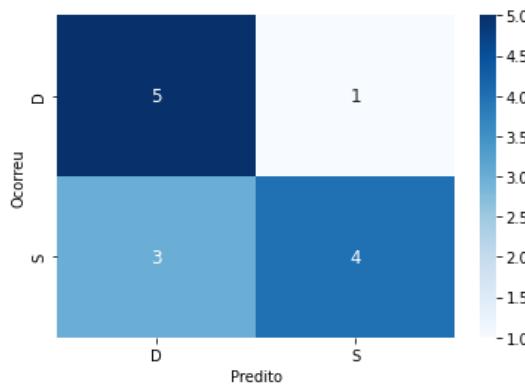


Figura 2.12: Mapa de Calor mostrando a Matriz de Confusão

Quanto mais escuro, maior a quantidade de elementos, assim podemos rapidamente avaliar como nosso modelo se comporta (o ideal é que as diagonais fiquem bem escuras enquanto que as extremidades claras).

2.10.1 Em valor ou percentual?

Um detalhe muito comum de acontecer é: devemos mostrar o valor em decimal ou percentual? Ou seja as quantidades reais das amostras ou um percentual do todo? Usar o seguinte código para gerar uma matriz de um teste realizado com imagens:

```

1 conf_arr = np.array([[88,14,4],[12,85,11],[5,15,91]])
2 sum = conf_arr.sum()
3 df_cm = pd.DataFrame(conf_arr,
4     index = [ 'Cão', 'Gato', 'Coelho'],
5     columns = [ 'Cão', 'Gato', 'Coelho'])
6 res = sn.heatmap(df_cm, annot=True, vmin=0.0, vmax=100.0, cmap=plt.cm.Blues)
7 plt.yticks([0.5,1.5,2.5], [ 'Cão', 'Gato', 'Coelho'], va='center')
8 plt.title('Matriz de Confusão')
9 plt.show()
```

E obtemos como resultado:

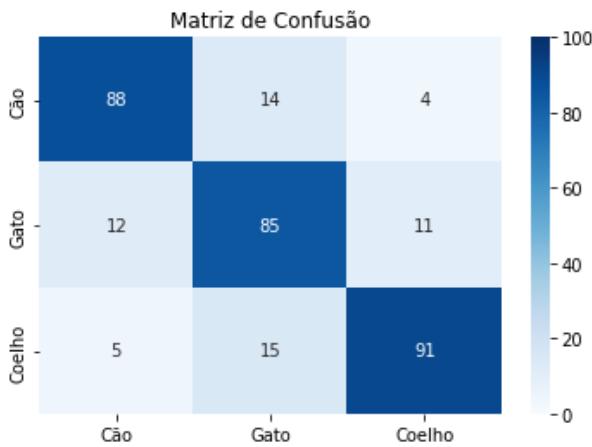


Figura 2.13: Comparativo numérico na Matriz de Confusão

E ao realizarmos um ajuste:

```

1 conf_arr = conf_arr * 100.0 / ( 1.0 * sum )
2 conf_arr /= 100
3 df_cm = pd.DataFrame(conf_arr,
4     index = [ 'Cão', 'Gato', 'Coelho'],
5     columns = [ 'Cão', 'Gato', 'Coelho'])
6 res = sn.heatmap(df_cm, annot=True, vmin=0.0, vmax=0.3, fmt='%.2%', cmap=plt.cm.Blues)
7 plt.title('Matriz de Confusão (em %)')
8 plt.show()
```

Obtemos a seguinte matriz agora com o resultado em percentual:

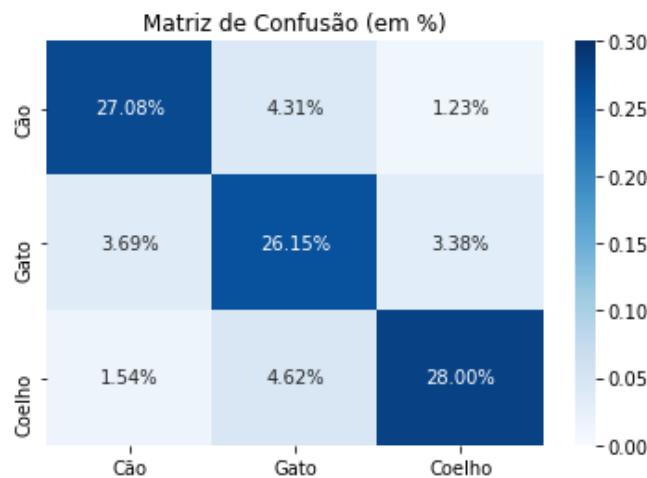


Figura 2.14: Comparativo percentual na Matriz de Confusão

Qual dos dois é melhor? A resposta seria: que fica mais fácil para que o Cientista de Dados consiga explicar o resultado para seu público com a maior clareza, não existe uma regra definida para isso.

2.10.2 Na prática

Voltamos ao nossos dados sobre Diabetes, ler os dados:

```

1 colnames = ['gest', 'glic', 'sang', 'skin', 'insul', 'mass', 'familia', 'idade',
   'conf']
2 df = pd.read_csv('pima-indians-diabetes.data.csv', names=colnames)
3 df.head()

```

Deixar somente os atributos que nos interessam:

```

1 df = df.drop(columns=['sang'],axis=1)
2 df = df.drop(columns=['insul'],axis=1)
3 df = df.drop(columns=['gest'],axis=1)
4 df = df.drop(columns=['skin'],axis=1)
5 df.head()

```

Dica 2.5: 100% de Precisão. Sempre que ocorrer 100% podemos ligar todas as antenas pois existe um erro, nenhum modelo possui essa previsão tão perfeita. Considere também que abaixo de 70% não é preditivo.

Agora acontece um erro clássico, a base contém *conf* o atributo alvo. Se seguimos em frente para separar as bases em teste e treino o modelo provavelmente acusa uma acurácia (errada) com 100% de precisão, é como se ele estivesse colando as respostas. Então devemos isolar esse atributo:

```

1 target = df['conf']
2 df = df.drop(columns=['conf'],axis=1)

```

E assim podemos separar as observações em treino e teste:

```
1 X_train, X_test, y_train, y_test = train_test_split(df, target, test_size = .25)
```

Usamos uma medida de 25%, ou seja 75% para treino e o restante para testar nosso modelo. Existem 768 registros não nulos no total, temos como resultado final: 576 para o modelo treinar e 192 para testar. Normalmente deixamos 25% para teste e verificação da acurácia do modelo esse número pode ser aumentado ou diminuído conforme seus dados, não existe uma regra definida.

Conforme nossa verificação do melhor modelo, vimos que devemos trabalhar com **Regressão Logística**, então executar o modelo e verificar a acurácia:

```
1 clf = LogisticRegression(max_iter=10000)
2 clf.fit(X_train, y_train)
3 print('Acurácia:', clf.score(X_test, y_test))
```

Esse resultado pode variar mas está em torno dos 76%, o problema é, onde esse modelo está errando? Executar a matriz de confusão para descobrir:

```
1 y_pred = clf.predict(X_test)
2 conf = confusion_matrix(y_test, y_pred)
3 data = {
4     'Ocorreu': y_test,
5     'Predito': y_pred
6 }
7 df2 = pd.DataFrame(data, columns=['Ocorreu', 'Predito'])
8 conf = pd.crosstab(df2['Ocorreu'], df2['Predito'], rownames=['Ocorreu'],
9                     colnames=['Predito'])
9 res = sn.heatmap(conf, annot=True, fmt='.0f', annot_kws={"size":12},
10                   cmap=plt.cm.Blues)
10 plt.show()
```

E obtemos como resultado:

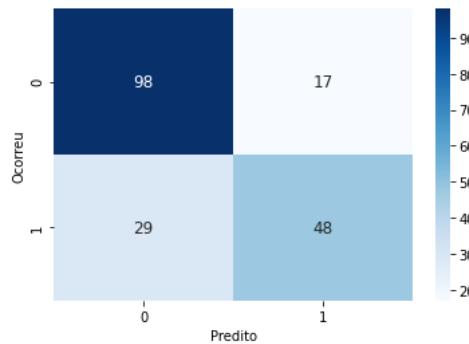


Figura 2.15: Resultado da Matriz de Confusão

Vemos que o modelo se comporta bem nos casos que o paciente não tem diabetes, acerta 85,2% e erra 14,8% das vezes. Já quando tem a doença acerta 62,3% e erra 37,7% das vezes. Ou seja, para melhorarmos a acurácia precisamos de mais dados com pacientes que desenvolveram a diabetes.

2.11 Curva ROC e Valor AUC

Entre todas as teorias vistas para Ciência de Dados este é um dos conceitos mais simples e ao mesmo tempo mais complicado (acho que é equivalente a Matriz de Confusão e inclusive depende dela).

Na teoria **ROC** (*Receiver Operating Characteristic*, algo como Característica de Operação do Receptor) é uma curva de probabilidade. É criada ao traçar a taxa de verdadeiro-positivo (TPR - true positive rate) contra a taxa de falsos-positivos (FPR - false positive rate). Que taxas são essas? Voltando ao conceito de Matriz de Confusão a TPR e FPR são calculadas pelas fórmulas:

$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}}$$

Figura 2.16: Fórmulas para o cálculo da ROC

Ou seja, número de vezes que o classificador acertou a predição contra o número de vezes que errou. **AUC** (*Area Under the Curve*) representa a área da ROC, considera-se como o grau ou medida de separabilidade. Quanto maior o valor, melhor é o modelo em prever ou (por exemplo) em distinguir entre pacientes com e sem uma determinada doença.

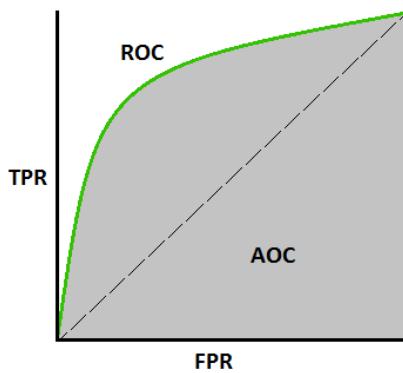


Figura 2.17: Fórmulas para o cálculo da ROC

Vejamos um simples exemplo para entendermos como esse processo funciona:

```

1 import pandas as pd
2 from sklearn import metrics
3 from sklearn.model_selection import train_test_split
4 import matplotlib.pyplot as plt
5 from sklearn.model_selection import cross_val_score
6 from sklearn.datasets import make_classification
7 from sklearn.linear_model import LogisticRegression
8 from sklearn.ensemble import RandomForestClassifier
9 %matplotlib inline

```

Apenas para entendermos como aquilo tudo que foi escrito no início da seção funciona, usar a função `make_classification` para produzir alguns dados aleatórios:

```

1 X, y = make_classification(n_samples = 10000, n_features=10, n_classes = 2, flip_y =
   0.5)
2 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .25)

```

Foram geradas 10.000 amostras com 10 campos e 2 classes (traduzindo para o português isso significa valores 0 e 1), separamos 25% para testar nosso modelo. Usar o Modelo de Regressão Logística para analisar esses dados:

```

1 model = LogisticRegression(solver='liblinear', penalty='l2', C=0.1)
2 model.fit(X_train, y_train)
3 print('Acurácia', model.score(X_test, y_test))

```

Como os dados são randômicos o resultado pode variar, mas chegamos em torno de 70%. E agora podemos calcular o **AUROC** (*Area Under the Receiver Operation Characteristics*):

```

1 y_prob = model.predict_proba(X_test)[:,1]
2 fpr, tpr, _ = metrics.roc_curve(y_test, y_prob)
3 auroc = float(format(metrics.roc_auc_score(y_test, y_prob), '.8f'))
4 print(auroc)

```

Existe 72% de área preenchida. Qual o motivo da diferença? Estamos levando em consideração as taxas TPR e FPR, não apenas acertou/errou. **AUC** resume a curva **ROC** num único valor que é o cálculo da “área sob a curva”, porém apresentar a informação assim não tem graça, colocar de forma gráfica:

```

1 plt.plot(fpr, tpr, label='Curva ROC (area = %0.2f)' % auroc)
2 plt.plot([0, 1], [0, 1], 'k--')
3 plt.xlabel('Taxa de Falso Positivo')
4 plt.ylabel('Taxa de Verdadeiro Positivo')
5 plt.title('Exemplo do ROC')
6 plt.legend(loc="lower right")
7 plt.show()

```

E obtemos como resultado:

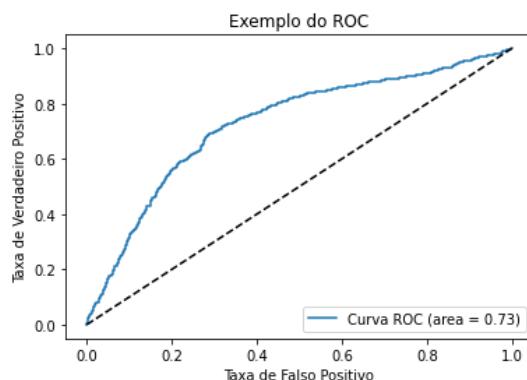


Figura 2.18: Visão da ROC

2.11.1 Na prática

Voltamos ao nossos dados sobre Diabetes, ler os dados:

```
1 colnames = ['gest', 'glic', 'sang', 'skin', 'insul', 'mass', 'familia', 'idade',
   'conf']
2 df = pd.read_csv('pima-indians-diabetes.data.csv', names=colnames)
3 df.head()
```

Deixar somente os atributos que nos interessam:

```
1 df = df.drop(columns=['sang'], axis=1)
2 df = df.drop(columns=['insul'], axis=1)
3 df = df.drop(columns=['gest'], axis=1)
4 df = df.drop(columns=['skin'], axis=1)
5 df.head()
```

Isolar o atributo alvo e retirá-lo dos dados.

```
1 target = df['conf']
2 df = df.drop(columns=['conf'], axis=1)
```

Separar as observações em treino e teste:

```
1 X_train, X_test, y_train, y_test = train_test_split(df, target, test_size = .25)
```

Testamos dois modelos de agrupamento para saber qual o melhor comportamento com esses dados. Para facilitar nossa vida, criar um método que retorna a acurácia de um determinado modelo:

```
1 def score(mdl, Xtrn, Xtst, ytrn, ytst):
2     mdl.fit(Xtrn, ytrn)
3     return float(format(mdl.score(Xtst, ytst), '.8f'))
```

Recebemos o modelo a ser treinado, e o conjunto de atributos separados em X e y (tanto para treino como teste) e devolvemos a acurácia do modelo. De modo semelhante:

```
1 def auroc(ytst, yprob):
2     fpr, tpr, _ = metrics.roc_curve(ytst, yprob)
3     auc = float(format(metrics.roc_auc_score(ytst, yprob), '.8f'))
4     return fpr, tpr, auc
```

Esse outro método recebe o y de teste e o de probabilidades e retorna o **FPR**, **TPR** e **AUC**. Os modelos de agrupamento para realizarmos nosso teste são **Regressão Logística** e **Floresta Aleatória**. Para cada um desses, basicamente, são os mesmos comandos.

Regressão Logística:

```
1 clfRL = LogisticRegression(max_iter=1000)
2 print("Acurácia RL:", score(clfRL, X_train, X_test, y_train, y_test))
3 y_probRL = clfRL.predict_proba(X_test)[:, 1]
```

```

4 fprRL, tprRL, aucRL = auroc(y_test, y_probRL)
5 print("AUC RL", aucRL)

```

Floresta Aleatória:

```

1 clfRF = RandomForestClassifier(n_estimators=1000)
2 print("Acurácia RF:", score(clfRF, X_train, X_test, y_train, y_test))
3 y_probRF = clfRF.predict_proba(X_test)[:,1]
4 fprRF, tprRF, aucRF = auroc(y_test, y_probRF)
5 print("AUC RF:", aucRF)

```

Ambos os modelos trabalham com grupos de 1.000, sendo que para a **Regressão Logística** isso é considerado de iterações realizadas enquanto que para a **Floresta Aleatória** o número de árvores de decisão utilizadas.

Dica 2.6: Como esses modelos trabalham? Não devemos nos preocupar nesse momento como esses modelos processam, em capítulos subsequentes trataremos separadamente cada um deles. Aqui veremos apenas os conceitos que serão necessários para a melhor escolha e avaliação dos modelos.

E obtemos o seguinte resultado, que pode variar de acordo com o que foi separado de treino e teste: A **Regressão Logística** atingiu um resultado melhor tanto na acurácia (73,95%) quanto na curva com quase 80% de área ocupada enquanto que a **Floresta Aleatória** ficou com quase 71% de acurácia e 77,35% de área. Às vezes o resultado de acurácia pode até ser similar, mas raramente a área ocupada será igual:

```

1 plt.plot([0, 1], [0, 1], 'k--')
2 plt.plot(fprRL,tprRL,label="RL " + str(aucRL))
3 plt.plot(fprRF,tprRF,label="RF " + str(aucRF))
4 plt.xlabel('Taxa de Falso Positivo')
5 plt.ylabel('Taxa de Verdadeiro Positivo')
6 plt.title('Detecção da Diabetes')
7 plt.legend(loc="lower right")
8 plt.show()

```

E obtemos como resultado:

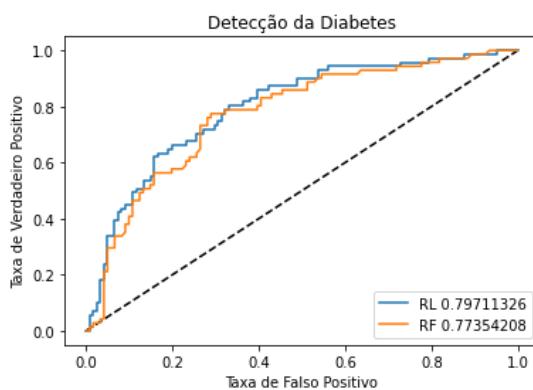


Figura 2.19: Performance dos Modelos

2.12 Terminamos?

Como conceitos sim, mas como prática permita-me deixar um exercício. A **Scikit-Learn** nos oferece uma base sobre de cistos (Câncer de Mama) encontrados em pacientes de *Wisconsin* com classificação se é malígnio (212 amostras) ou benigno (357 amostras). Para usá-la importar a biblioteca:

```
1 from sklearn.datasets import load_breast_cancer
```

Carregamos os dados:

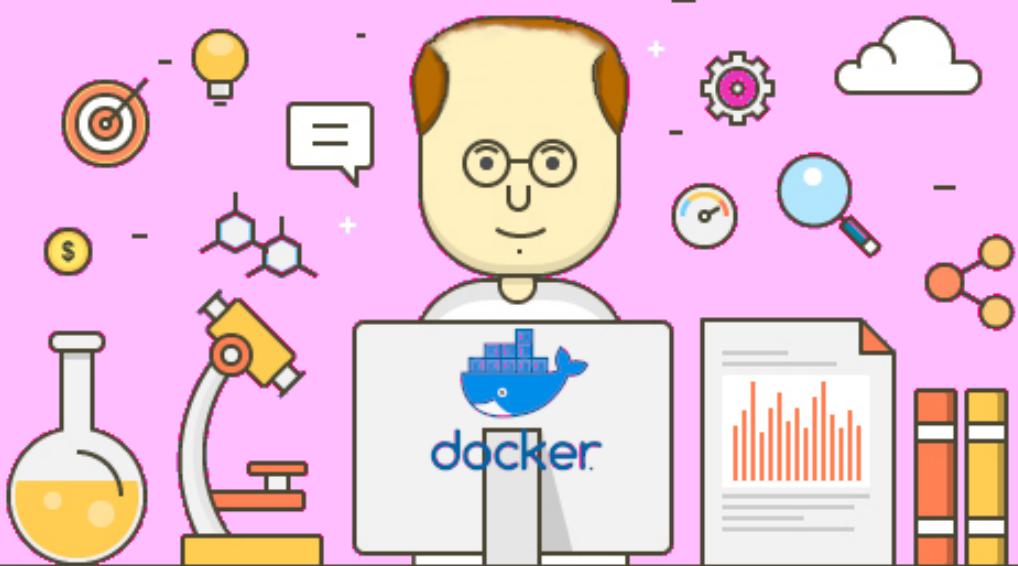
```
1 cancer = load_breast_cancer()
```

E obtemos um objeto *Bunch* da Scikit-Learn com os 569 casos registrados. Para criar as nossas bases de treino e teste usar o comando:

```
1 X_train, X_test, y_train, y_test = train_test_split(cancer.data, cancer.target,
test_size = .25)
```

A variável *data* contém os atributos preditores enquanto que *target* se a paciente teve um tipo malígnio ou não (atributo dependente). Pronto, agora é com você. Aplicar os conhecimentos que vimos nesse capítulo para definir quais são os melhores atributo a usar e o modelo colhendo todos os resultados possíveis.

No endereço https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html está disponibilizada a documentação sobre esta base.



3. EDA

F Nós torturamos os dados até eles confessarem. (Ricardo Cappra - Cientista de Dados)

3.1 Passos da EDA

EDA é fundamental para entender qualquer conjunto de observações. É aqui podemos obter informações e fazer descobertas. Aqui colocamos o conhecimento para trabalhar. Acrônimo para *Exploratory Data Analysis* (Análise Exploratória de Dados), desempenha um papel crítico na compreensão do quê? por que? e como? na declaração do problema.

É a primeira ação realizada na ordem das operações que um Cientista de Dados deve executar ao receber uma nova fonte de observações e a declaração de problema. Tratamos EDA como uma série de técnicas utilizadas como forma de entendermos os diversos aspectos que temos para trabalhar.



Figura 3.1: Passos da EDA

A preparação dos dados para análise é inevitável, e a maneira como fazemos isso define a sua qualidade. Na prática o que faremos nesse capítulo é compreendermos o que temos a nossa disposição para trabalhar.

Normalmente as observações se dividem em atributos **Preditores** (Entradas) e **Alvo** (saída). Uma vez que os localizemos, devemos identificar seu tipo e categoria.

3.2 Passo 1 - Entender os Dados

Nesta fase precisamos compreender o que temos a nossa disposição. Começamos o processo com a importação das bibliotecas:

```
1 import pandas as pd
2 import matplotlib.pyplot as plt
3 import seaborn as sns
4 %matplotlib inline
```

Vamos trabalhar com três bibliotecas básicas, que como já mencionamos devemos conhecê-las a fundo: **Pandas** para análise, **MatPlotLib** e **SeaBorn** para mostrar em forma gráfica.

Agora precisamos dos dados, para isso usaremos o arquivo *StudentsPerformance.csv*:

```
1 df = pd.read_csv('StudentsPerformance.csv')
```

Nessa fase compreendemos melhor o que temos na nossa mão, **Pandas** é ideal para essa tarefa. Seu funcionamento é como um "Editor de Planilha", dessa forma que devemos encarar essa biblioteca, sua diferença básica é a nomenclatura de como o *DataFrame* (e não Planilha) é visualizado:

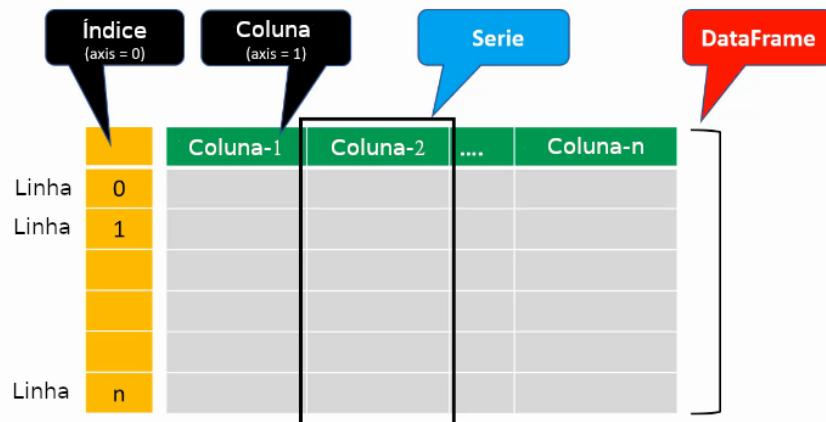


Figura 3.2: Visão Pandas

Uma coluna aqui é vista como uma *Serie* e *index* é o que mantém a "cola" das series juntas. Dois comandos são básicos para visualizarmos o *DataFrame*:

```
1 df.head()
```

Porém nesse livro não utilizaremos o termo "coluna" e sim "atributo" (consideremos ambos como sinônimos). Mostra as primeiras observações, como parâmetro podemos passar a quantidade. E:

```
1 df.tail()
```

Mostra as últimas observações e também como parâmetro podemos passar a quantidade. O que temos até o momento? Sabemos é uma base sobre estudantes e as linhas são: gênero, etnicidade, nível de escolaridade

dos pais, forma de alimentação, realizou um teste de preparação do curso, nota de matemática, nota de leitura e nota de escrita.

Então só com esses dois comandos já podemos saber sobre qual assunto iremos tratar: estudantes que realizaram provas e em quais condições. Quantos registros temos a nossa disposição? Ou quais são os nomes dos atributos?

```
1 print("Tamanho: ", df.shape)
2 print("Nome dos Atributos: ", df.columns)
```

As variáveis *shape* e *columns* do *DataFrame* respondem aos questionamentos. De forma mais completa podemos usar:

```
1 df.info()
```

Nos mostra inclusive o tipo de cada atributo e se contém ou não elementos nulos. Temos 3 atributos que são do tipo inteiro (*int64*) e podemos analisá-los com o comando:

```
1 df.describe()
```

Nos fornece as informações estatísticas básicas como média, desvio padrão, menor valor, máximo, 1º quartil (25%), 2º quartil ou mediana (50%), 3º quartil (75%) e o maior valor. Ou seja, as informações para a montagem de um **BoxPlot**. Vamos montá-lo para melhor visualizar as observações:

```
1 fig, axes = plt.subplots(1, 3, figsize=(10,4))
2 axes[0].boxplot(df['math score'])
3 axes[0].set_title("Matemática")
4 axes[1].boxplot(df['reading score'])
5 axes[1].set_title("Leitura")
6 axes[2].boxplot(df['writing score'])
7 axes[2].set_title("Escrita")
8 plt.show()
```

E obtemos como resultado:

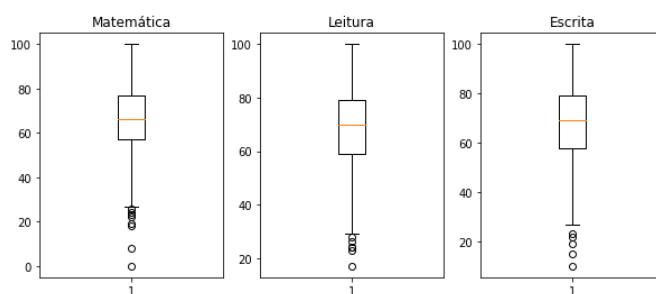


Figura 3.3: BoxPlot das Notas

Boxplot¹ é um gráfico que avalia a distribuição das observações. É formado exatamente com os atributos

¹Diagrama de Caixa se prefere, foi atribuída ao matemático **John W. Tukey** (1915 –2000), curiosamente algumas literaturas chamam de "Tukey BoxPlot", mas se realizar uma pesquisa ninguém sabe ao certo quem criou realmente esse diagrama.

que mostramos na função `describe()`. Porém suas hastes (inferiores e superiores) se estendem do quartil inferior (ou superior) até o menor valor não inferior (ou superior) ao limite. São calculados da seguinte forma:

- Limite inferior: $Q_1 - 1,5 \times (Q_3 - Q_1)$.
- Limite superior: $Q_3 + 1,5 \times (Q_3 - Q_1)$.

Resumidamente é formado da seguinte maneira:

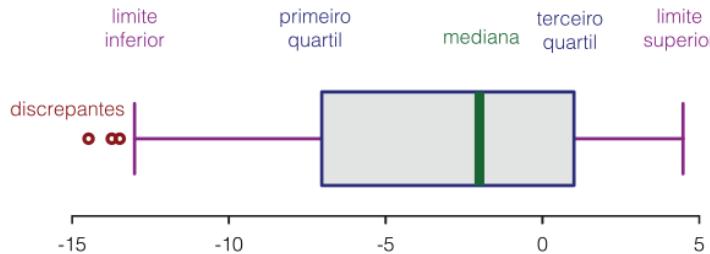


Figura 3.4: Estrutura do BoxPlot

Esses pontos "discrepantes" podem ocorrer acima ou abaixo dos limites, são chamados de *Outliers*. Não é necessariamente um erro, podemos classificá-lo como uma anomalia curiosa e que merece nossa atenção.

3.2.1 Localizar os Outliers

Para achar esses *Outliers* isolamos os três atributos numéricos:

```
1 X = df.iloc[:, 5:8].values
```

E criamos um novo *DataFrame* somente com a modificação de nome por um número:

```
1 pd.options.display.float_format = '{:.1f}'.format
2 xDF = pd.DataFrame(X)
```

Para quê isso serve? A função `describe()` cria um *DataFrame*, podemos percorrê-lo, porém fica muito mais simples se cada atributo for um numeral, pois assim podemos usar um comando `for` para isso:

```
1 z = xDF.describe()
2 for t in z:
3     iqr = z[t][6] - z[t][4]
4     extMenor = z[t][4] - (iqr * 1.5)
5     extMaior = z[t][6] + (iqr * 1.5)
6     print('Para o índice %d valores devem estar abaixo de %.2f e acima de %.2f' %
           (t, extMenor, extMaior))
```

E obtemos o seguinte resultado:

Para o índice 0 valores devem estar abaixo de 27.00 e acima de 107.00

Para o índice 1 valores devem estar abaixo de 29.00 e acima de 109.00

Para o índice 2 valores devem estar abaixo de 25.88 e acima de 110.88

Pelo BoxPlot todos os valores estão abaixo, então para localizá-los:

```

1 matOutliers = (X[:,0] < 27)
2 df[matOutliers]
```

E assim mostramos todas as observações que a nota de matemática (índice 0) é abaixo do valor 27. Proceder de mesmo modo para as notas de leitura (índice 1) e escrita (índice 2) e assim desvendar quais são os *Outliers*.

Podemos também analisar graficamente e visualizar a Distribuição Normal de cada atributo, por exemplo para a nota de Escrita:

```

1 sns.kdeplot(df['writing score'], shade=True)
2 plt.show()
```

Obtemos como resultado:

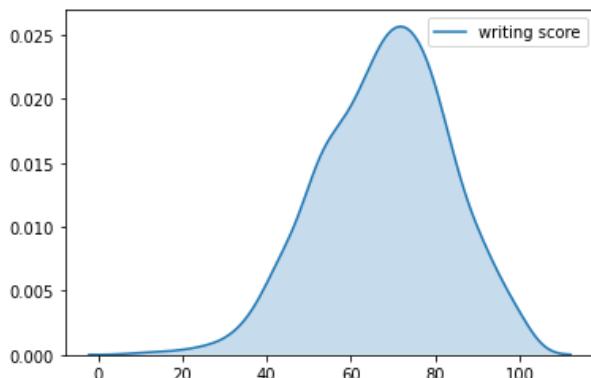


Figura 3.5: Distribuição das observações para Nota de Escrita

E assim verificamos como cada atributo numérico se comporta.

3.2.2 Tratar Atributos Categóricos

Sabemos que os primeiros cinco atributos do *Dataframe* são categóricos, porém conforme a função *info()* o tipo delas estás *object*. É interessante mudarmos para o tipo caractere para evitarmos quaisquer problemas futuros.

```

1 df['gender'] = df['gender'].astype(pd.StringDtype())
2 df['race/ethnicity'] = df['race/ethnicity'].astype(pd.StringDtype())
3 df['parental level of education'] = df['parental level of
   education'].astype(pd.StringDtype())
4 df['lunch'] = df['lunch'].astype(pd.StringDtype())
5 df['test preparation course'] = df['test preparation course'].astype(pd.StringDtype())
```

E ao aplicarmos uma nova chamada a função *info()* vemos que os tipos agora estão corretos. Quantos tipos únicos existem para cada atributo?

```
1 df.nunique()
```

Mostra a quantidade de valores não repetidos de cada atributos (inclusive os numéricos). E agora sabemos que temos: 2 gêneros, 5 etnicidades, 6 níveis de escolaridade dos pais, 2 formas de alimentação e 2 tipos para teste de preparação do curso. Mas quem são?

```
1 print("Gênero: ", df['gender'].unique())
2 print("Etnicidade: ", df['race/ethnicity'].unique())
3 print("Escolaridade dos Pais: ", df['parental level of education'].unique())
4 print("Refeição: ", df['lunch'].unique())
5 print("Realizou Preparatório: ", df['test preparation course'].unique())
```

3.3 Passo 2 - Limpar os Dados

A limpeza dos dados trata de muitos problemas como informação repetida, valores faltantes (que podem ser descobertos por associação) e inconsistentes. Para esse último tipo o pior caso são os nulos. (In)felizmente essa base está horrível para essa fase e assim pegamos um outro arquivo **titanic.csv**:

```
1 df = pd.read_csv('titanic.csv')
2 df.head()
```

Repetimos todo o processo da fase anterior para descobrirmos de que se tratam as observações e descobrimos que são os passageiros (sobreviventes ou não - atributo *Survived* - sendo este atributos alvo) do famoso **RMS Titanic**, este foi pensado para ser o navio mais luxuoso e seguro de sua época e supostamente "inafundável". Como sabemos em sua viagem inaugural de *Southampton* para *Nova Iorque* afundou no dia 14 de abril de 1912 com mais de 1.500 pessoas a bordo. Porém esta base contém apenas 891 registros.

Ao aplicarmos a função `info()` percebemos que os atributos *Age* (idade), *Cabin* (número da cabine) e *Embarked* (local de Embarque) possuem valores faltantes. Que valores são esses?

```
1 print(df.isnull().sum())
```

Sabemos que faltam: 177 em **Age**, 687 em **Cabin** e 2 em **Embarked**. Também podemos mostrar exclusivamente os que faltam, isso é útil para quando temos muitos atributos no modelo:

```
1 null_value_stats = df.isnull().sum(axis=0)
2 null_value_stats[null_value_stats != 0]
```

Ou ainda criar uma função personalizada que retorna um *Dataframe* com a informação mais completa o possível (inclusive com seu percentual):

```
1 def mostrarNulos(data):
2     null_sum = data.isnull().sum()
3     total = null_sum.sort_values(ascending=False)
4     percent = (((null_sum / len(data.index))*100).round(2)).sort_values(ascending=False)
5     df_NULL = pd.concat([total, percent], axis=1, keys=['Tot.Nulo', 'Perc.Nulo'])
6     df_NULL = df_NULL[(df_NULL.T != 0).any()]
7     return df_NULL
```

E ao chamá-la:

```
1 df_Age = mostrarNulos(df)
2 df_Age.head()
```

Obtemos como resultado:

	Tot.Nulo	Perc.Nulo
Cabin	687	77.10
Age	177	19.87
Embarked	2	0.22

Figura 3.6: Nulos e Perceitual da Base Titanic

Lidar com esses tipos de nulos é complicado pois não temos como consultar e o máximo que podemos fazer é podá-los da nossa base ou atribuir um valor genérico que não afete nosso resultado (como o caso de *Embarked*). Porém **Número da Cabine** é um dado relevante? Essa é a principal pergunta que nos devemos fazer, por exemplo existe algum modelo preditivo que possa nos dizer que se estivéssemos em determinada cabine no navio sobreviveríamos ou não? Entretanto **Idade** é um dado relevante (lembra da frase: mulheres e crianças primeiro), então essa é uma característica que pode ser essencial.

Criar uma função com um gráfico para mostrar, por idade, como estão as observações:

```
1 def executarGrafico():
2     try:
3         sns.distplot([df['Age']])
4         plt.show()
5     except ValueError as err:
6         print(err)
```

Agora a cada vez que chamarmos essa função obtemos como resultado:

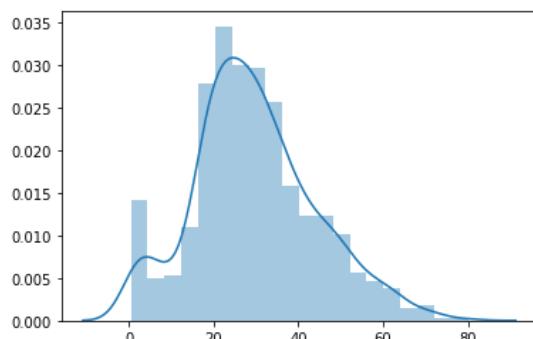


Figura 3.7: Gráfico de Idade do Titanic

Dica 3.1: Imputação ou retirada de valores. Como tratamos de adicionar ou retirar elementos na base a cada vez devemos ler novamente as observações contidas no arquivo CSV.

Porém em algumas versões da *SeaBorn* este pode apresentar erro devido a presença dos nulos, é ideal que os retiremos do *DataFrame* para evitarmos problemas. Em muitas biografias encontramos algo do tipo: "atribuir um valor (preferencialmente *outlier*) para estes tipos". Tentaremos essa técnica com os seguintes comandos:

```
1 df['Age'].fillna(-25, inplace=True)
2 executarGrafico()
```

Obtemos como resultado:

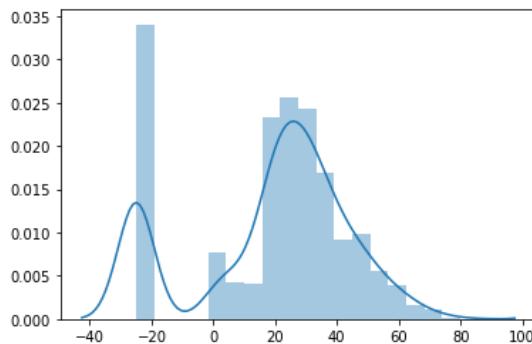


Figura 3.8: Gráfico de Idade do Titanic com Outliers

Nosso gráfico de idade ganhou uma nova barra, que sabemos com valores não existentes, também podemos atribuir qualquer outro valor como por exemplo a média:

```
1 df['Age'] = df['Age'].fillna(df['Age'].mean())
2 executarGrafico()
```

Ou a mediana (função `median()`) que resultaria em um gráfico completamente esquisito. Sendo assim vamos cortar esses valores:

```
1 df = df.dropna(axis=0)
2 executarGrafico()
```

E teremos a seguinte situação:

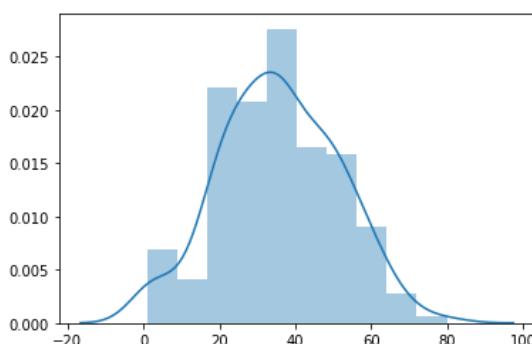


Figura 3.9: Gráfico de Idade do Titanic sem nulos

O que aconteceu? O comando executado eliminou todas as linhas que possuíam valores nulos, e o atributo *Cabin* interferiu e nos deixou, conforme pode ser mostrado com a função *info()*, somente 183 registros no total. Ou seja, o corte que devemos aplicar deve ser cirúrgico e somente no atributo que representa a idade.

```
1 df[‘Age’] = df[‘Age’].dropna(axis=0)
2 executarGrafico()
```

O que nos resulta no mesmo gráfico mostrado no início desta e 891 registros. Como citamos, podemos retirar o atributo *Cabin* para que este não interfira mais em futuras análises:

```
1 df = df.drop([‘Cabin’], axis=1)
```

Dica 3.2: Ferramenta para Limpeza dos Dados. Conhece o **OpenRefine?** é uma ferramenta gratuita dedicada a limpeza e tratamento das observações, baixe uma apostila gratuitamente na minha página do Academia.edu (<https://iesbpreve.academia.edu/FernandoAnselmo>).

3.4 Passo 3 - Relacionamento entre os Atributos

Vamos retomar nossa base de **Estudantes** e verificarmos como os atributos se relacionam:

```
1 df = pd.read_csv(‘StudentsPerformance.csv’)
2 df.corr()
```

E temos um valor que corresponde ao grau de relacionamento, um intervalo de -1 a 1, sendo quanto mais próximo do mínimo menor é seu grau de relacionamento. Porém é muito mais fácil de visualizarmos esse resultado com um Mapa de Calor:

```
1 rel = df.corr()
2 sns.heatmap(rel, xticklabels=rel.columns, yticklabels=rel.columns, annot=True)
3 plt.show()
```

Obtemos como resultado:

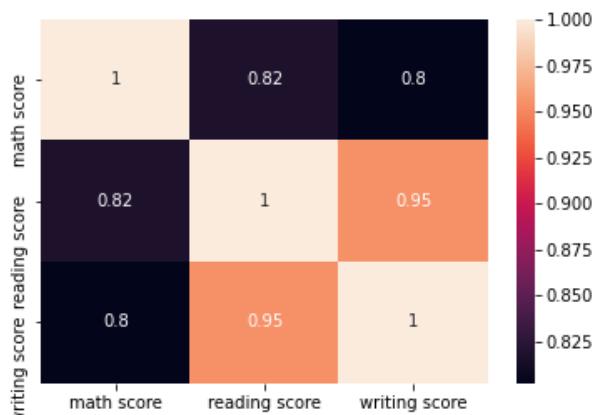


Figura 3.10: Mapa de Calor dos Relacionamentos

Vemos que as notas de Escrita e Leitura possuem um forte grau de relacionamento, como se uma fosse a responsável pela outra. Já a de matemática interfere mais na nota de leitura.

Curiosamente se aplicarmos isso na base do **Titanic** vemos que os atributos mais importantes para *Fare* (sobreviveu) que é nosso alvo são: *Fare* que é o valor pago pela passagem e *Parch* que se refere a quantidade de pais. Ou seja, os mais ricos e se a criança tinha ou não os pais a bordo de modo a colocá-las no bote salva vidas.

Outra maneira de visualizarmos, também de forma gráfica, é através da dispersão de valores:

```
1 sns.pairplot(df)
2 plt.show()
```

Obtemos como resultado:

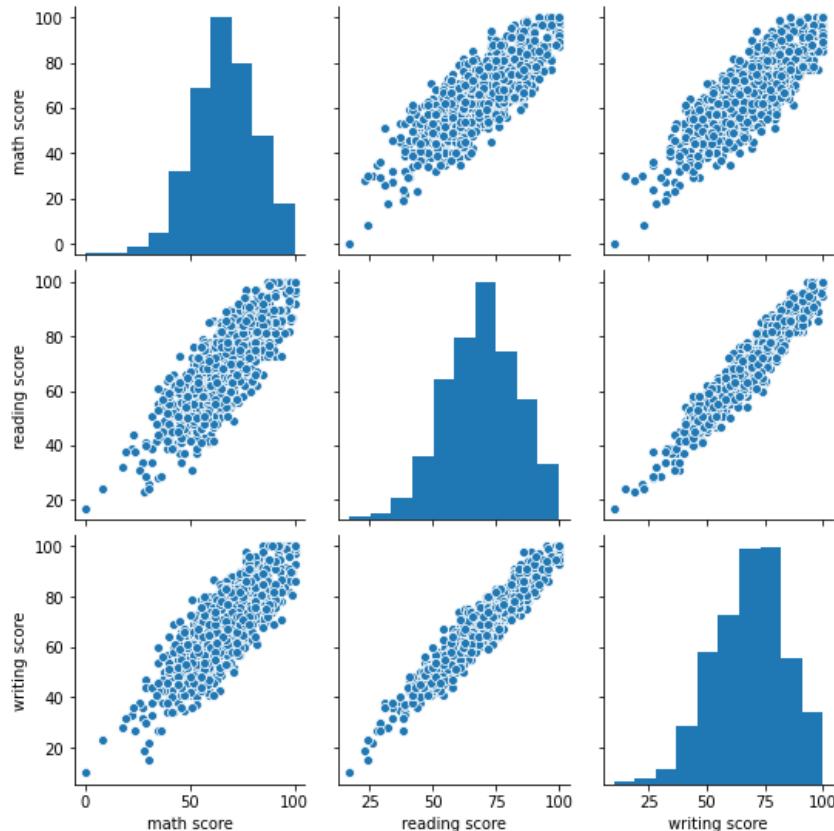


Figura 3.11: Dispersão Associada

Quanto mais juntos aparecem os pontos mais relacionadas estão. Podemos isolar as notas de Escrita e Leitura em um único gráfico, por exemplo:

```
1 sns.regplot(x='writing score', y='reading score', data=df)
2 plt.show()
```

Esta função executa um ajuste e plotagem simples com base no modelo de Regressão Linear. E obtemos como resultado:

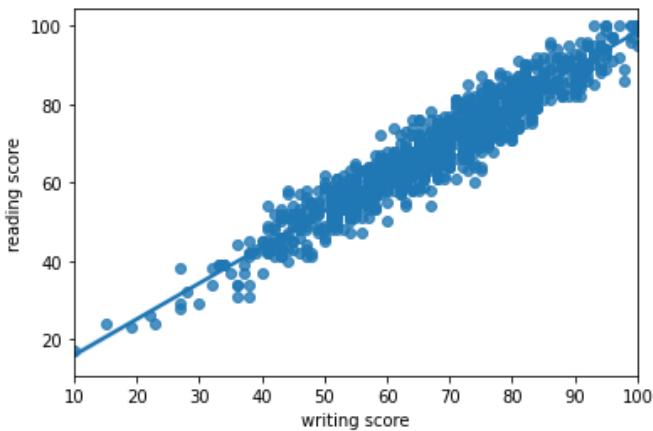


Figura 3.12: Notas de Leitura e Escrita

Porém o mais interessante é colorir os pontos de forma diferente com base em um atributo categórico que pode ser uma causa (para uma nota alta ou baixa), por exemplo o quanto a alimentação interferiu na nota:

```

1 sns.lmplot(x='writing score', y='reading score', hue='lunch', data=df)
2 plt.show()

```

A função *lmplot()* combina *regplot()* com a classe **FacetGrid**. Esta auxilia visualizar a distribuição de um determinado atributos, bem como o relacionamento entre os vários separadamente dentro de subconjuntos das observações. Obtemos como resultado:

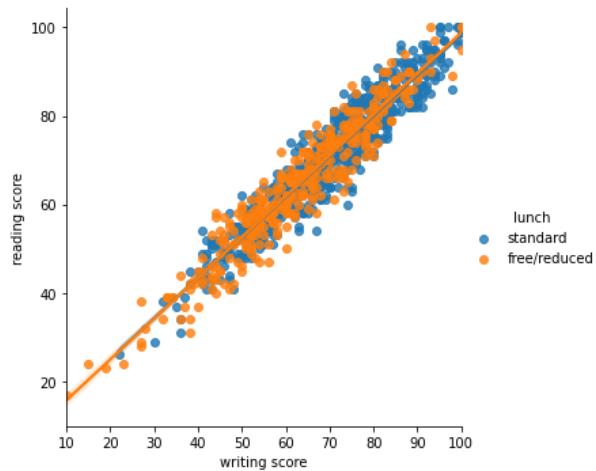


Figura 3.13: Nota associada a Alimentação - RegPlot

Uma melhor forma de visualizar é usar a função *relplot()* que fornece acesso a várias funções diferentes no nível de eixos que mostram o relacionamento entre dois atributos com mapeamentos semânticos de subconjuntos:

```

1 sns.relplot(x='writing score', y='reading score', hue='lunch', data=df)
2 plt.show()

```

Obtemos como resultado:

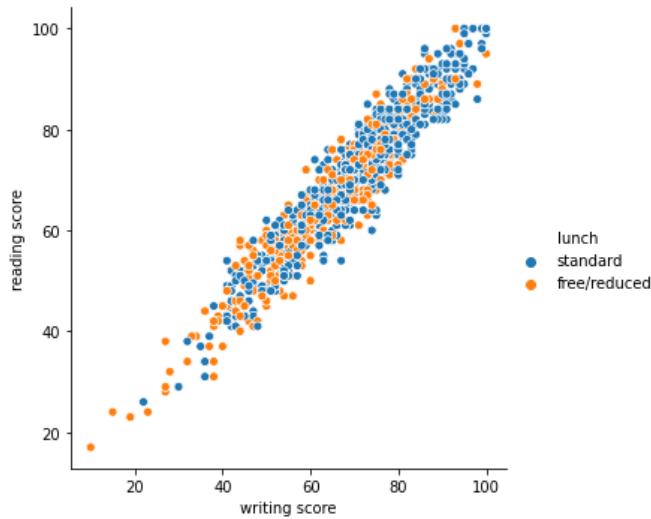


Figura 3.14: Nota associada a Alimentação - RelPlot

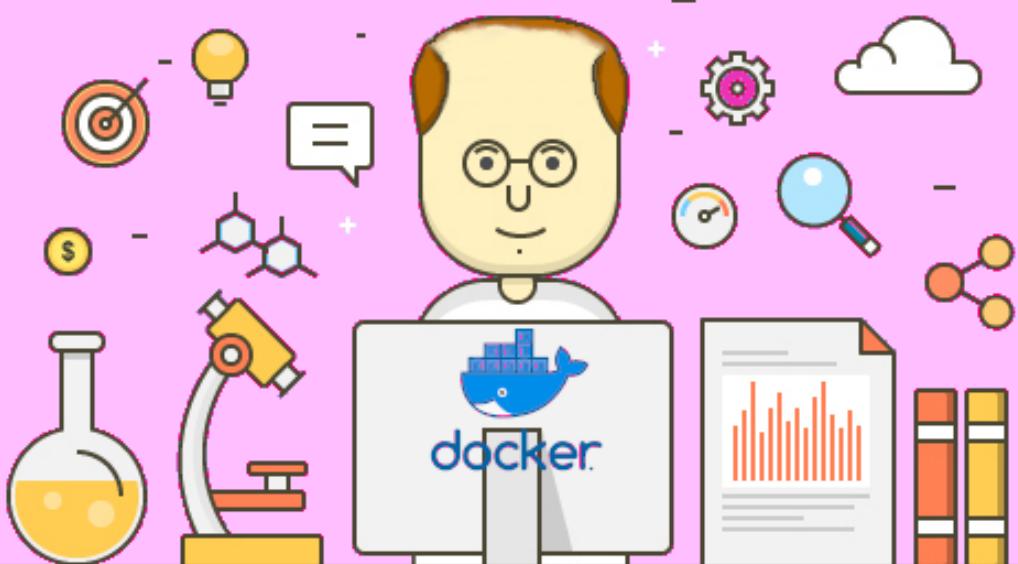
Ou seja, podemos responder várias perguntas apenas com a verificação do relacionamento entre os atributos. Como forma de fixar o conhecimento procure realizar o mesmo teste com outros atributos categóricos e descobrir como se comportam em relação a nota, se existe ou não interferência.

3.5 Conclusão

Mantemos em mente que EDA é um aspecto central da *Data Science*, que às vezes é esquecido. O primeiro passo de qualquer ação que tomemos é conhecer as observações: entendê-las e familiarizar-se. Quais são as respostas que estamos tentando obter? Quais são os atributos e o que significam? Como é a aparência de uma perspectiva estatística? As observações estão formatadas corretamente? Possuem valores ausentes? duplicados? E quanto aos *outliers*? Conhecemos eles? Ou seja, devemos responder a esses questionamentos.

É necessário muito trabalho de preparação, pois no mundo real dados raramente são limpos e homogêneos. Costumamos dizer que 80% do tempo valioso em um Cientista de Dados é utilizado com a localização, limpeza e organização das observações. Os 20% restantes são destinados a realizar as análises.

Agora estamos prontos para começarmos a explorar diversas observações com a utilização dos modelos.



4. Modelos Iniciais

F Na vida, não existe nada a temer, mas a entender. (Marie Curie - Cientista e Vencedora 2 vezes do Prêmio Nobel)

4.1 K-Means

Acredito que K-Means seja o modelo mais simples para começarmos, este é um algoritmo de Aprendizado Não Supervisionado, ou seja, não necessita de atributos alvo para agir, sua função é de separar as observações em grupos de modo que possamos observar melhor os dados.

Sendo assim, nosso problema para usar esse algoritmo é exatamente achar esse k ideal de modo que os grupos sejam separados coerentemente. Para isso existe uma técnica interessante chamada "Técnica do Cotovelo"(Elbow Technique).

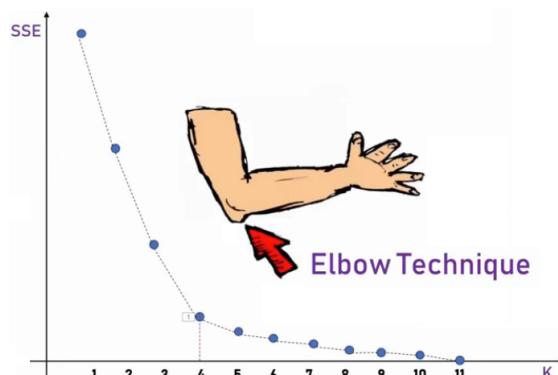


Figura 4.1: Técnica do Cotovelo

Exatamente na posição 4 existe uma "quebra" para passar ao próximo valor, usamos para definir essa quebra o SSE¹ (*Sum Squared Error*).

4.2 Aplicação da Técnica

Para achar o k ideal vamos ativar nosso JupyterLab personalizado que criamos com o Docker e na primeira célula importamos as bibliotecas necessárias:

```

1 import pandas as pd
2 import numpy as np
3 from sklearn.preprocessing import scale
4 from sklearn.cluster import KMeans
5 from matplotlib import pyplot as plt
6
7 %matplotlib inline

```

Importamos a biblioteca Pandas e a Numpy para manipularmos os dados, a Scikit-Learn para usarmos o modelo K-Means e Matplot para vermos o resultado em um gráfico. A última linha é utilizada para mostrar os gráficos no Jupyter. Próximo passo consiste em ler os dados, baixamos o arquivo **gameML.csv** e na posição do nosso arquivo **.ipynb** criamos uma subpasta chamada **bases** e nesta colocamos o arquivo.

```

1 df = pd.read_csv('bases/gameML.csv', delimiter=';')
2 df.head()

```

E como resultado da execução dessa célula devemos ter:

	Nome	Idade	Salário
0	Daenerys Targaryen	27	70000
1	Jon Snow	29	90000
2	Gregor Ciegane	29	61000
3	Arya Stark	28	60000
4	Tyrion Lannister	42	150000

Figura 4.2: Idades e Salários da Empresa GameML

No arquivo existem 3 campos: nome do funcionário, idade e salário, se plotarmos os dados entre idade e salário em gráfico:

```
1 plt.scatter(df['Idade'], df['Salário'])
```

Obtemos como resultado:

¹Soma Residual dos Quadrados, é a soma dos resíduos elevado por 2. É uma medida da discrepância entre os dados e um modelo de estimativa. Um valor pequeno SSE indica um ajuste apertado do modelo aos dados.

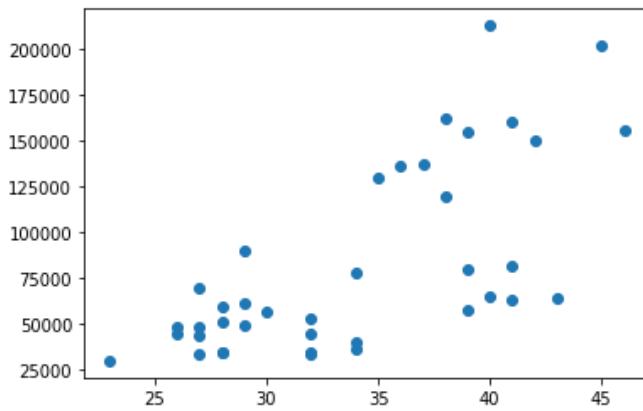


Figura 4.3: Idades e Salários da Empresa GameML

Quantos grupos de dados podemos distinguir? Para localizarmos a quantidade ideal aplicamos a técnica do cotovelo que consiste de:

```

1 k_rng = range(1,10)
2 sse = []
3 for k in k_rng:
4     km = KMeans(n_clusters=k)
5     km.fit(df[['Idade', 'Salário']])
6     sse.append(km.inertia_)
7 plt.xlabel('K')
8 plt.ylabel('SSE (Sum Squared Error)')
9 plt.plot(k_rng, sse)

```

Criar um range de 1 a 10 (um simples número máximo de possíveis *clusters*), para cada valor treinamos o modelo com as variáveis e obtemos o valor do atributo **inércia**. O algoritmo agrupa dados e procura separar amostras em n grupos de igual variação, minimizando um critério conhecido como inércia ou **RSS** dentro do *cluster*. O que estamos fazendo na prática é colocar o valor 1 para o **k** e guardar esse valor, em seguida o valor 2 e assim sucessivamente. Por fim plotamos esse valor em um gráfico e obtemos como resultado:

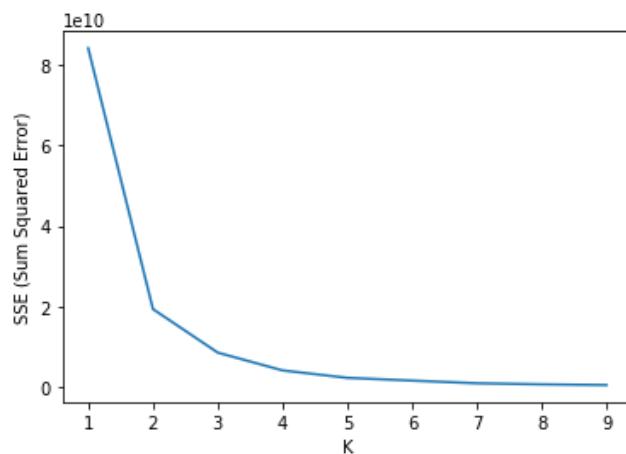


Figura 4.4: Gráfico com os valores de Inércia

E vemos nosso "cotovelo" da curva bem na posição 3, marcando assim o número ideal de clusters.

4.3 Plotagem do Resultado do Modelo

Um detalhe interessante que para usarmos o algoritmo K-Means, devemos colocar os dados em "escala", vamos tentar usar o modelo sem proceder dessa forma:

```
1 km = KMeans(n_clusters=3)
2 y_predict = km.fit_predict(df[['Idade', 'Salário']])
3 df['ypred'] = y_predict
4 df.head()
```

Já sabemos que o valor de 3 clusters é o ideal, então realizamos o treinamento com os atributos Idade e Salário para montarmos um novo atributo com o resultado dessa predição (somente para que o gráfico apareça separado por cores). E plotamos o gráfico:

```
1 cores = np.array(['green', 'red', 'blue'])
2 plt.scatter(x=df['Idade'],
3 y=df['Salário'],
4 c=cores[df.ypred], s=50)
```

Obtemos como resultado:

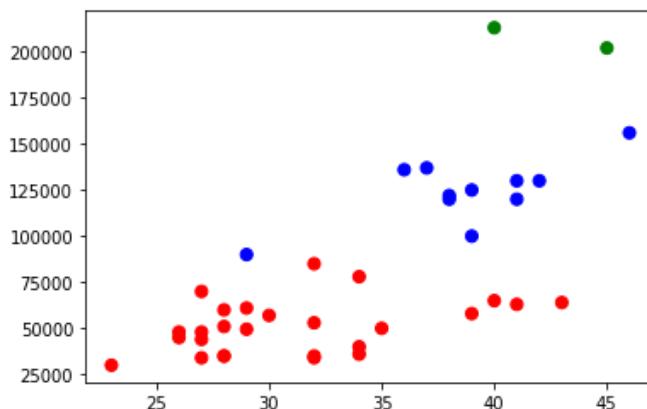


Figura 4.5: Separados por Grupo

E parece que temos algo bem errado com alguns *outliers* aparecendo, observamos o ponto azul no meio dos vermelhos e um outro azul isolado perto dos verdes. Então antes de treinarmos esse algoritmo devemos colocar os dados na mesma escala, isso é feito assim:

```
1 df['Salário'] = scale(df.Salário)
2 df['Idade'] = scale(df.Idade)
3 df.head()
```

Os atributos **idade** e **salário** possuem valores bem diferentes e distantes e isso gera problemas para nosso resultado final, colocar em escala e aproximar (sem modificar o resultado final) os valores seria algo criar

um modelo de um prédio porém mantendo as mesmas proporções do prédio original.

A função da **Scikit-Learn** que realiza este processo é chamada *scale()* e colocamos em escala os atributos se visualizarmos nossos dados agora veremos que o atributo **idade** possui valores entre -2 e 2 enquanto que **salário** entre -1.5 e 3 (são diferentes exatamente para manter a proporcionalidade). Retornamos ao mesmo processo de treinamento:

```

1 km = KMeans(n_clusters=3)
2 y_predict = km.fit_predict(df[['Idade', 'Salário']])
3 df['ypred'] = y_predict
4 df.head()

```

Plotamos novamente o gráfico e agora como resultado teremos:

```

1 cores = np.array(['green', 'red', 'blue'])
2 plt.scatter(x=df['Idade'],
3 y=df['Salário'],
4 c=cores[df.ypred], s=50)

```

E obtemos como resultado:

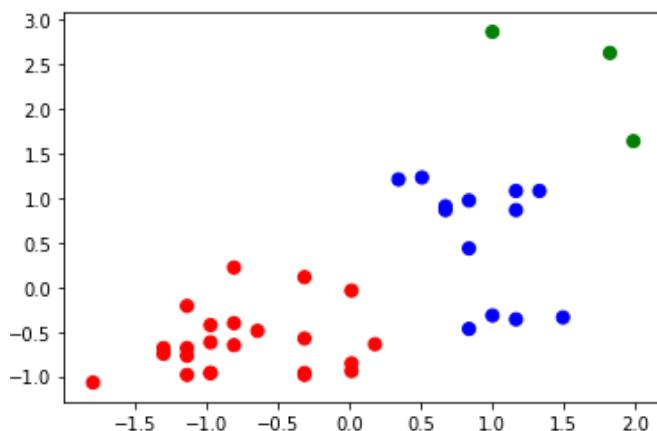


Figura 4.6: Separados por Grupo em Escala

Que é um resultado bem mais coerente.

4.4 K-Nearest Neighbors

Ou simplesmente KNN. Modelos assim existem pois muitas pessoas pensam que separar em *clusters* não auxilia na predição, pois bem nosso próximo modelo é um supervisionado e destinado a Predição por Clusterização (ou se prefere por proximidade dos grupos). KNN que normalmente é usado para a predição de imagens como: Isso é um Gato? Ou não é um Gato? Porém ao invés de imagens, vamos usar uma base bem conhecida chamada **Flores Íris** para entendermos seu comportamento.

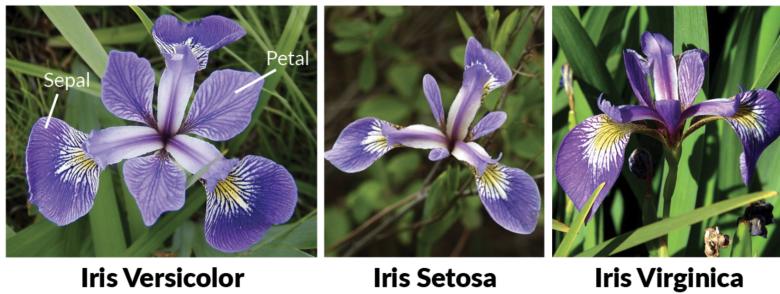


Figura 4.7: Flores Iris

Nessa base existem três espécies separadas: Versicolor, Setosa e Virgíñica. E para distingui-las utilizamos 2 medidas da sépala e da pétala (largura e altura de cada). O problema é que algumas espécies causam as maiores confusões em nossos modelos. Para realizarmos uma predição sobre essa base importamos nossas bibliotecas:

```

1 import numpy as np
2 from matplotlib import pyplot as plt
3 from sklearn import datasets
4 from sklearn.model_selection import train_test_split
5 from sklearn import neighbors
6
7 %matplotlib inline

```

Usamos a **NumPy** para gerenciamento dos dados. **Matplotlib** para plotarmos os gráficos. Da **Scikit-Learn** obtemos os nossos dados através do pacote **datasets** e para separar uma massa de teste contamos com o *train_test_split*. E a *neighbors* contém o nosso algoritmo. O próximo passo consiste na preparação dos dados:

```

1 iris = datasets.load_iris()
2 X, y = iris.data, iris.target
3
4 X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
   random_state=1234)

```

O método *load_iris()* traz a nossa base em uma matriz de dados. Nossa base está dividida em *data* que contém os *features* preditores (tamanho e largura da sépala e tamanho e largura da pétala, que colocaremos em *X*) e *target, feature* que contém a definição da espécie (0 representa **Setosa**, 1 para **Versicolor** e 2 para **Virgíñica** que colocaremos em *y*). Usamos o método *train_test_split* para retirar 20% dos dados como amostra de teste e assim teremos quatro agrupamentos:

- **X_train**, com os dados para treino do algorítimo.
- **X_test**, com os dados para teste.
- **y_train**, com o resultado para o treino.
- **y_test**, com o resultado para o teste.

Com nossos dados preparados vamos treinar o modelo:

```

1 clf = neighbors.KNeighborsClassifier()
2 clf.fit(X_train, y_train)

```

```
3 print(clf.score(X_test, y_test))
```

E conseguimos uma boa acurácia com incríveis 96% de precisão, agora é vermos na prática como isso funciona.

4.4.1 Predição com K-Nearest Neighbors

Primeiro vamos mostrar os dados:

```
1 cores = np.array(['green', 'red', 'blue'])
2 subplot1 = plt.scatter(x=x[:, 0], y=x[:, 1], c=cores[y], s=50)
```

Pegamos as duas primeiras variáveis tamanho e largura da sépala e obtemos como resultado:

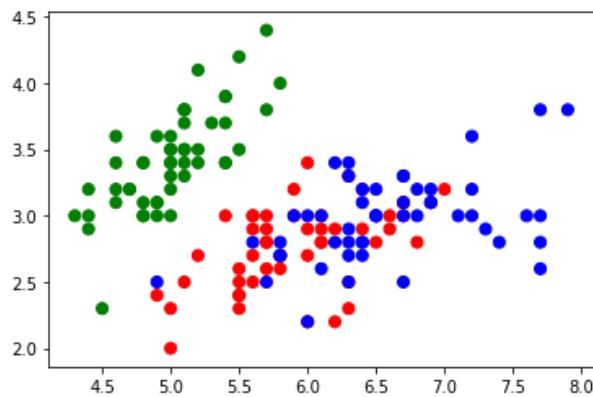


Figura 4.8: Comparar tamanho e largura da Sépala

Não nos perdemos nas cores **Verde** é Setosa, **Vermelho** é Versicolor e **Azul** é Virgínica. Agora vamos pensar em um ponto qualquer nesse espaço, por exemplo:

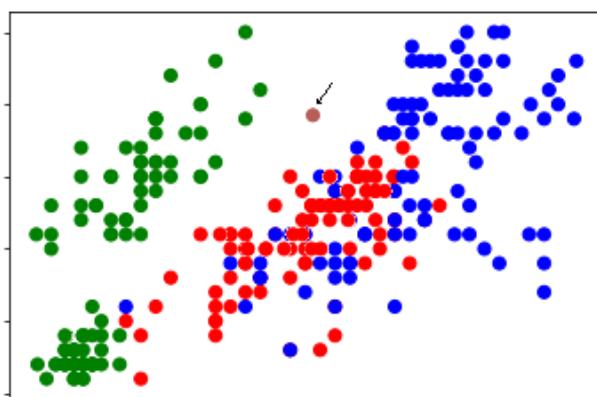


Figura 4.9: Localizar o Ponto Roxo

O ponto roxo fica na interseção do 4º valor de X e y qual cor real ele seria? Observamos no gráfico anterior que os pontos são 6,0 e 4,0 porém nos falta o valor para mais dois atributos tamanho e largura da pétala:

```

1 cores = np.array(['green', 'red', 'blue'])
2 subplot1 = plt.scatter(x=X[:, 2], y=X[:, 3], c=cores[y], s=50)

```

E obtemos como resultado:

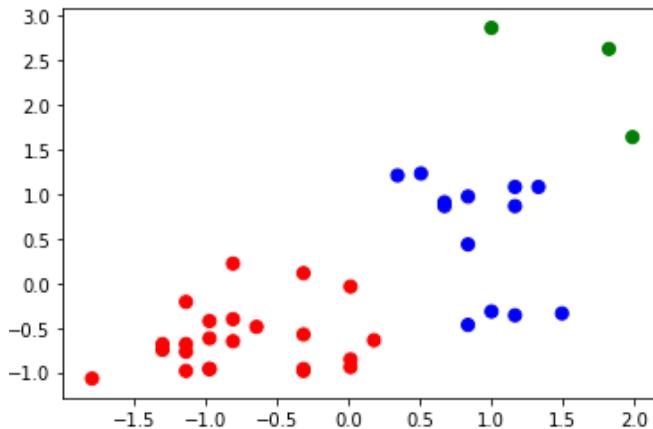


Figura 4.10: Comparar tamanho e largura da Pétala

E verificamos que na interseção do 4º valor de X e y temos os valores 4,0 e 2,0. Agora que obtemos os quatro valores podemos realizar uma predição:

```

1 predicao = clf.predict([[6.0, 4.0, 4.0, 2.0]])
2 print(predicao)

```

E resulta que o modelo prevê que é do tipo [1], ou seja, um ponto vermelho da espécie **Versicolor**.

4.5 Análise de Cluster

Então sabemos agora que ambos modelos K-Means e KNN trabalham utilizando *clusters* (agrupamentos) sendo que o primeiro é do tipo não supervisionado destinado a separação com base em um número de centroides (k) presentes e os valores médios mais próximos (isso representa uma distância Euclidiana entre as observações). Porém é necessário colocar os dados em escala para verificar se não ocorre nenhuma perturbação nesse centroide. Vamos importar algumas bibliotecas para realizarmos mais testes:

```

1 import numpy as np
2 import pandas as pd
3 import seaborn as sns
4 import matplotlib.pyplot as plt
5 from sklearn import datasets
6 from sklearn.preprocessing import scale
7 from sklearn.cluster import KMeans
8 from sklearn.metrics import classification_report
9
10 %matplotlib inline

```

Já passamos por todas e não desejo ser repetitivo porém dessa vez vamos utilizar a Pandas para manipular os dados e a classe *metrics* da SciKit-Learn para mostrar o comportamento do nosso modelo. Iremos continuar usando a base Iris e construímos um *DataFrame* somente com os dados dos atributos preditores porém guardaremos o atributo alvo para verificar como nosso modelo se comportou:

```

1 iris = datasets.load_iris()
2 X = scale(iris.data)
3 y = pd.DataFrame(iris.target)
4 y.columns = ['Targets']
5 variable_names = iris.feature_names
6 iris_df = pd.DataFrame(iris.data)
7 iris_df.columns = variable_names
8 iris_df.head()

```

E nosso *DataFrame* se apresenta da seguinte maneira:

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2

Figura 4.11: DataFrame com os dados dos Atributos Preditores

O próximo passo é construir e treinar nosso modelo:

```

1 clustering = KMeans(n_clusters=3, random_state=5).fit(X)

```

Normalmente para treinar um modelo passamos dois conjuntos de dados, porém o K-Means só recebe um único conjunto, exatamente por não realizar previsões precisa apenas dos dados para separá-los em conjuntos. Mas como será que foi seu comportamento? Descobrimos isso comparando dois gráficos:

```

1 cores = np.array(['green', 'red', 'blue'])
2 relabel = np.choose(clustering.labels_, [1, 0, 2]).astype(np.int64)
3 plt.figure(figsize = [15, 5])
4
5 plt.subplot(1, 4, 1)
6 plt.scatter(x=iris_df['petal length (cm)'],
7 y=iris_df['petal width (cm)'],
8 c=cores[iris.target], s=50)
9 plt.title('Real (Pétala)')
10
11 plt.subplot(1, 4, 2)
12 plt.scatter(x=iris_df['petal length (cm)'],
13 y=iris_df['petal width (cm)'],
14 c=cores[relabel], s=50)
15 plt.title('KMeans (Pétala)')
16
17 plt.subplot(1, 4, 3)
18 plt.scatter(x=iris_df['sepal length (cm)'],
19 y=iris_df['sepal width (cm)'],
20 c=cores[iris.target], s=50)

```

```

21 plt.title('Real (Sépala)')
22
23 plt.subplot(1, 4, 1)
24 plt.scatter(x=iris_df['sepal length (cm)'],
25 y=iris_df['sepal width (cm)'],
26 c=cores[relabel], s=50)
27 plt.title('KMeans (Sépala)')
28
29 plt.show()

```

Usamos os mesmos conjuntos de cores para cada espécie, teremos quatro gráficos comparativos: 1º largura e altura da Pétala e a cor será mostrada com base em nosso atributo alvo (ou seja o valor real), 2º o que o modelo achou que seria o correto, 3º largura e altura da Sépala e o 4º novamente como o modelo separou. E obtemos como resultado:

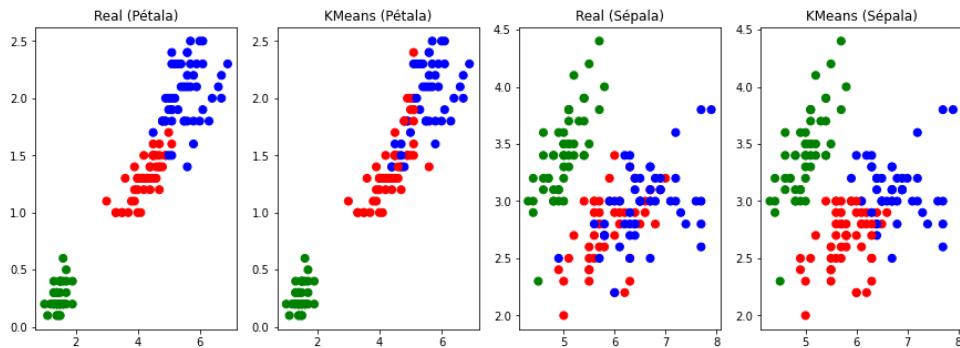


Figura 4.12: Comparativo entre o Real e o KMeans

Para pétala o **K-Means** quase acertou a posição de cada espécie, porém para Sépala aconteceram as maiores confusões, isso se deve ao fato do centroide. Para melhor avaliarmos nosso modelo precisamos de mais medidas: *Precision* (precisão) é a medida de relevância do modelo, *Recall* (revocação ou sensibilidade) se trata da medida de completude do modelo e *F1 Score* se trata de uma média ponderada entre *precision* e *recall*. Podemos obtê-las da seguinte forma:

```

1 metricas = classification_report(y, relabel)
2 print(metricas)

```

Obtemos como resultado:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	50
1	0.74	0.78	0.76	50
2	0.77	0.72	0.74	50
accuracy			0.83	150
macro avg	0.83	0.83	0.83	150
weighted avg	0.83	0.83	0.83	150

Figura 4.13: Relatório de Performance do K-Means

Precision - É a razão entre as observações positivas previstas corretamente e o total de observações positivas previstas. Calculada com a fórmula: $TP \div (TP + FP)$.

Recall - É a razão entre as observações positivas previstas corretamente e todas as observações da classe real. Calculada com a fórmula: $TP \div (TP + FN)$

F1 Score - Essa pontuação leva em consideração tanto os falsos positivos quanto os negativos. Intuitivamente, não é tão fácil entender como precisão, mas F1 é geralmente mais útil que *precision*, especialmente se estivermos com uma distribuição de classe desigual. $2 \times (recall \times precision) \div (recall + precision)$

Acurácia funciona melhor se os falsos positivos e negativos tiverem um custo semelhante. Se o custo for muito diferente, é melhor olharmos essas métricas.

4.6 Clusterização Hierárquica

Este é um modelo alternativo ao particionamento de *cluster* no conjunto de dados, pode ser aplicado para encontrar a distância entre cada ponto e seus vizinhos mais próximos e conectá-lo de forma ideal. Podemos mostrar o número de subgrupos com o auxílio de um Dendrograma².

É útil pois não existe necessidade de especificar o número de *clusters* (ou K) antes da análise e o dendrograma fornece uma representação visual desses. Vamos trazer para o conjunto de bibliotecas visto anteriormente mais três:

```

1 from scipy.cluster.hierarchy import dendrogram, linkage
2 from sklearn.cluster import AgglomerativeClustering
3 from sklearn.metrics import accuracy_score

```

Para este exemplo vamos utilizar outra base que está contida no arquivo **mtcars.csv** (trazer essa para a subpasta **/base**). E carregamos os dados do seguinte modo:

```

1 carros = pd.read_csv('bases/mtcars.csv')
2 carros.columns = ['nome', 'mpg', 'cil', 'disp', 'hp', 'drat', 'wt', 'qsec', 'vs',
   'am', 'gear', 'carb']
3 X = carros[['mpg', 'disp', 'hp', 'wt']].values
4 y = carros['am'].values

```

Essa base contém 32 modelos de carros com os seguintes atributos: Nome, quilometragem, número de cilindros, deslocamento (medida de poder do carro em polegada cúbica), cavalos de força, relação do eixo traseiro, peso (em libras), comparativo de eficiência de gasto de combustível (por 1/4 milha), motor (0 = V-shaped, 1 = straight), câmbio (0 = automática, 1 = manual), total de marchas e carburadores. Porém para não trabalharmos com tantos atributos vamos usar somente: consumo de gasolina (mpg), deslocamento (disp), cavalos de força (hp) e peso (wt) e o nosso objetivo e descobrir se o carro possui um câmbio manual ou automático.

Podemos montar o dendrograma do seguinte modo:

```

1 z = linkage(X, 'ward')
2 dendrogram(z, truncate_mode='lastp', p=12, leaf_rotation=45, leaf_font_size=15,

```

²É um gráfico em formato de árvore que mostra visualmente os relacionamentos entre as observações.

```

show_contracted=True)

3
4 plt.title('Dendograma')
5 plt.xlabel('Tamanho do Cluster')
6 plt.ylabel('Distância')
7 plt.axhline(y=500)
8 plt.axhline(y=150)
9 plt.show()

```

Obtemos como resultado:

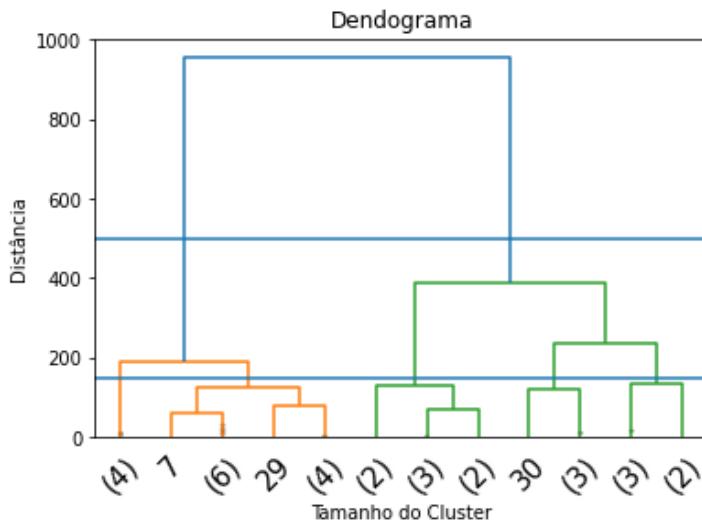


Figura 4.14: Dendrograma dos tamanhos do Cluster

O dendrograma mostra como cada *cluster* é composto e desenha um link em forma de U entre cada cluster e seus filhos. A parte superior indica uma mesclagem. Cada perna indica quais foram mesclados. O comprimento das pernas e do U representa a distância entre os filhos.

Para mesclar recursivamente o par de *clusters* e aumentar minimamente a distância de ligação utilizamos a função `AgglomerativeClustering()`. Essa possui dois parâmetros básicos: *affinity* e *linkage*.

affinity: métrica utilizada para calcular a ligação. Possui as seguintes opções:

- *euclidean* - é o único que aceita o parâmetro *linkage* como *ward*. Refere-se a distância euclidiana que pode ser provada pela aplicação repetida do teorema de Pitágoras.
- *l1* - critério de erro absoluto.
- *l2* - critério de erros quadrados (lembremos do RSS).
- *manhattan* - distância euclidiana ao quadrado.
- *cosine* - também chamada de Similaridade do Cosseno. É a distância do cosseno entre duas variáveis.
- *precomputed* - necessita de uma matriz de distância (em vez de similaridade) como entrada para o método de ajuste, pois X será considerado uma matriz.

linkage: define qual o critério de ligação usar. Determina qual distância usar entre os conjuntos de observação. Possui as seguintes opções:

- *ward* - minimiza a variação dos *clusters* que estão sendo mesclados.
- *average* - média das distâncias de cada observação dos conjuntos.
- *complete* - distâncias máximas entre todas as observações dos dois conjuntos.
- *single* - mínimo das distâncias entre todas as observações dos dois conjuntos.

Como escolher os parâmetros ideais? Fácil, testemos várias combinações e veremos qual possui uma melhor acurácia para os dados que estamos tratando:

```

1 hclusters1 = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
2     linkage='ward').fit(X)
3 print('Método 1:', accuracy_score(y, hclusters1.labels_))
4
5 hclusters2 = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
6     linkage='complete').fit(X)
7 print('Método 2:', accuracy_score(y, hclusters2.labels_))
8
9 hclusters3 = AgglomerativeClustering(n_clusters=2, affinity='euclidean',
10    linkage='average').fit(X)
11 print('Método 3:', accuracy_score(y, hclusters3.labels_))
12
13 hclusters4 = AgglomerativeClustering(n_clusters=2, affinity='manhattan',
14    linkage='single').fit(X)
15 print('Método 4:', accuracy_score(y, hclusters4.labels_))
16
17 hclusters5 = AgglomerativeClustering(n_clusters=2, affinity='manhattan',
18    linkage='complete').fit(X)
19 print('Método 5:', accuracy_score(y, hclusters5.labels_))
20
21 hclusters6 = AgglomerativeClustering(n_clusters=2, affinity='manhattan',
22    linkage='average').fit(X)
23 print('Método 6:', accuracy_score(y, hclusters6.labels_))
24
25 hclusters7 = AgglomerativeClustering(n_clusters=2, affinity='cosine',
26    linkage='single').fit(X)
27 print('Método 7:', accuracy_score(y, hclusters7.labels_))
28
29 hclusters8 = AgglomerativeClustering(n_clusters=2, affinity='cosine',
30    linkage='complete').fit(X)
31 print('Método 8:', accuracy_score(y, hclusters8.labels_))
32
33 hclusters9 = AgglomerativeClustering(n_clusters=2, affinity='cosine',
34    linkage='average').fit(X)
35 print('Método 9:', accuracy_score(y, hclusters9.labels_))

```

Obtemos como resultado:

Método 1: 0.78125
 Método 2: 0.4375
 Método 3: 0.78125
 Método 4: 0.625
 Método 5: 0.71875
 Método 6: 0.71875
 Método 7: 0.3125
 Método 8: 0.28125

```
Método 9: 0.1875
```

Assim para esse caso *Euclidian/Ward* ou *Manhattan/Complete* são os que melhor responderam ao nosso conjunto de dados com uma acurácia de 78,12%. Podemos inclusive tirar um relatório mais completo (como já vimos):

```
1 print(classification_report(y, hclusters1.labels_))
```

Obtemos como resultado:

	precision	recall	f1-score	support
0	0.88	0.74	0.80	19
1	0.69	0.85	0.76	13
accuracy			0.78	32
macro avg	0.78	0.79	0.78	32
weighted avg	0.80	0.78	0.78	32

Figura 4.15: Relatório de Performance da Clusterização Hierárquica

Só que ficou uma pergunta no ar, esse método se comporta melhor que um modelo de clusterização preditivo como o KNN?

4.6.1 Clusterização Hierárquica versus K-Nearest Neighbors

Para verificar como o KNN se comporta com os dados dos carros adicionamos mais quatro bibliotecas:

```
1 from sklearn.neighbors import KNeighborsClassifier
2 from sklearn import preprocessing
3 from sklearn.model_selection import train_test_split
4 from sklearn.metrics import classification_report
```

Como já temos nossos dados, vamos apenas separá-los em bases de treino e teste:

```
1 X = preprocessing.scale(X)
2 X_treino, X_teste, y_treino, y_teste = train_test_split(X, y, test_size=.20,
   random_state=17)
```

Porém devemos sempre lembrar que os modelos de clusterização trabalham melhor quando os dados estão em escala, assim acertamos os atributos preditores antes de realizar a separação de 80% dos dados para treino e 20% para teste.

```
1 clf = KNeighborsClassifier()
2 clf.fit(X_treino, y_treino)
```

Treinamos nosso modelo e podemos avaliar o resultado:

```
1 y_predito = clf.predict(X_teste)
2 print(classification_report(y_teste, y_predito))
```

Obtemos como resultado:

	precision	recall	f1-score	support
0	0.80	1.00	0.89	4
1	1.00	0.67	0.80	3
accuracy			0.86	7
macro avg	0.90	0.83	0.84	7
weighted avg	0.89	0.86	0.85	7

Figura 4.16: Relatório de Performance do K-Nearest Neighbors

E na média percebemos que este se comporta melhor pois atinge resultados acima dos 80%.

4.7 Regressão Linear

A regressão linear tenta modelar o relacionamento entre dois atributos, através de ajustes sob uma equação linear dos dados observados. Um atributo é considerado **explicativo** e a outro **dependente**. Para simplificar um pouco, é uma técnica que utiliza valores de entrada para predizer os de saída (como por exemplo, prever o crescimento da população de um País) através da aplicação dos coeficientes (também chamados de peso) da equação linear.

Comecemos com a importação das bibliotecas que necessitamos:

```

1 import pandas as pd
2 import numpy as np
3 from matplotlib import pyplot as plt
4 from sklearn.linear_model import LinearRegression
5
6 %matplotlib inline

```

A classe *Linear Model* da *Scikit-Learn* o método *LinearRegression* para realizar nosso trabalho. Baixar a base de dados **PopBrasil.csv** que contém as observações de Crescimento da População Brasileira.

```

1 df = pd.read_csv('bases/PopBrasil.csv')
2 df.head()

```

Obtemos como resultado:

Ano	Populacao
0	1960
1	72179226
2	1961
3	74311343
4	1962
5	76514328
6	1963
7	78772657
8	1964
9	81064571

Figura 4.17: Dados da População Brasileira

Devemos saber que o modelo trabalha com a relação entre atributos numéricos: explanatórios X e

dependentes y . Utiliza somente esse tipo devido aos ajustes matemáticos que são realizados e os pesos criados conforme a função minimiza os erros. Nossas observações são bem simples: temos atributos numéricos, "Ano" e "População". Para entendermos o relacionamento entre os atributos, plotamos esses em um gráfico:

```

1 plt.xlabel('Ano')
2 plt.ylabel('Quantidade da População')
3 plt.scatter(df.Anو, df.Populacao, color='red', marker='+')
```

Obtemos como resultado:

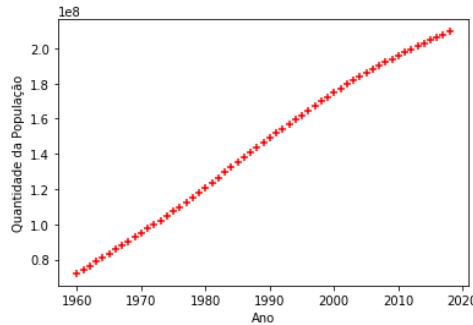


Figura 4.18: Dados da População Brasileira

E esta é a parte mais importante na execução desse modelo, a medida que alteramos o valor de "Ano" o valor de "População" também é afetado, ou seja, existe um relacionamento linear. Essa é a premissa básica para se usar este algoritmo, o relacionamento forte entre os atributos deve existir.

4.7.1 Aplicar a Regressão Linear

Agora que temos nossos atributos conferidos, basta treinarmos nosso modelo e obtermos nossa previsão:

```

1 reg = LinearRegression()
2 reg.fit(df[['Ano']], df.Populacao)
3 prev = reg.predict([[2020]])
4 print("Previsão 2020 é: %d" % prev)
```

E teremos a previsão da população brasileira para o ano de 2020, que é 221.322.254 de habitantes. Como a magia acontece? Pura matemática que é fornecida pela seguinte fórmula:

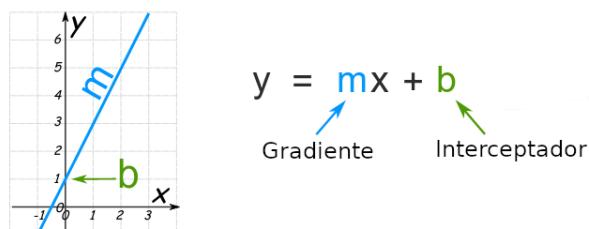


Figura 4.19: Base da Regressão Linear

Dica 4.1: Para saber mais. Se deseja conhecer mais sobre o assunto, visite a página: <https://www.mathsisfun.com/algebra/linear-equations.html> aonde se obtém uma explicação mais completa.

E podemos reproduzir esse resultado pois o objeto treinado nos fornece tanto o valor do Gradiente (`coef_[0]`) quanto do Interceptador (`intercept_`). Então:

```
1 m = reg.coef_[0]
2 b = reg.intercept_
3 prev2020 = m * 2020 + b
4 print("Previsão 2020 é: %d" % prev2020)
```

E temos exatamente o mesmo resultado. Podemos traçar a "Reta da Regressão Linear", pois o modelo consegue predizer os resultados de cada ano:

```
1 plt.xlabel('Ano')
2 plt.ylabel('Quantidade da População')
3 plt.scatter(df.Ano, df.Populacao, color='red', marker='+')
4 plt.plot(df.Ano, reg.predict(df[['Ano']]), color='blue')
```

Obtemos como resultado:

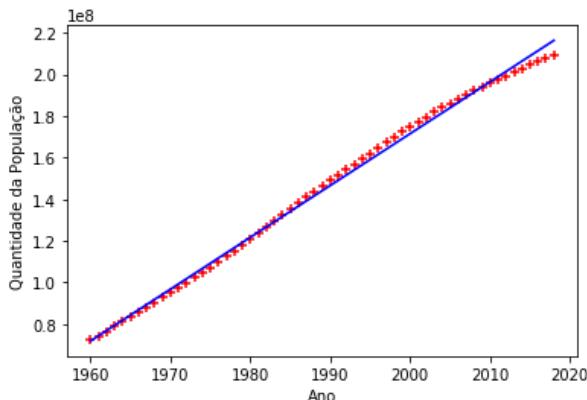


Figura 4.20: Dados da População Brasileira com a Previsão

Vamos praticar nossos novos "poderes de futurólogo", junto a essa base encontramos outra chamada ExpecVida.csv, com ela, tente prever qual será a Expectativa de Vida do brasileiro no ano de 2020.

4.7.2 Regressão Linear com mais de um Preditor

Vimos como usar o modelo de Regressão Linear, porém apenas a título de facilitação do entendimento, somente um atributo preditor. Mas o que acontece quando o alvo é influenciado por mais de um preditor? Vamos entender na prática como isso acontece.

Pensemos em um caso do Varejo, vamos utilizar um conjunto de observações chamado **marketSales.csv** que como o nome sugere, são compostos por transações de vendas. Sabemos que várias coisas influenciam

a saída de um determinado produto, tais como, o grau de visibilidade, peso, se possui muita ou pouca quantidade de gordura, tamanho do mercado ou outros.

Começamos com a importação da bibliotecas necessárias:

```

1 import pandas as pd
2 from sklearn.linear_model import LinearRegression
3 from sklearn.preprocessing import LabelEncoder
4 from sklearn.model_selection import train_test_split
5 from matplotlib import pyplot as plt
6
7 %matplotlib inline

```

E ler nossa base de dados:

```

1 df = pd.read_csv('bases/marketSales.csv')
2 df.head()

```

Até o momento nada de novo, nosso problema começa ao repararmos nas observações:

	Item_Identifier	Item_Weight	Item_Fat_Content	Item_Visibility	Item_Type	Item_MRP	Outlet_Identifier	Outlet_Establishment_Year
0	FDA15	9.30	Low Fat	0.016047	Dairy	249.8092	OUT049	1999
1	DRC01	5.92	Regular	0.019278	Soft Drinks	48.2692	OUT018	2009
2	FDN15	17.50	Low Fat	0.016760	Meat	141.6180	OUT049	1999
3	FDX07	19.20	Regular	0.000000	Fruits and Vegetables	182.0950	OUT010	1998
4	NCD19	8.93	Low Fat	0.000000	Household	53.8614	OUT013	1987

Figura 4.21: Observações sobre Vendas de Produtos

Sabemos que os modelos de regressão só trabalham com tipos numéricos, muito pior existe o caso de nulos entre algumas outras inconsistências nessas 14.204 observações.

4.7.3 Regressão Linear e Limpeza dos Dados

Sejamos francos, maior parte de trabalho do Cientista de Dados é arrumar os dados que sofridamente conseguiu para realizar o trabalho, então começaremos a compreender como uma parte disso funciona. Primeiro detalhe vamos tratar os atributos indesejáveis, nulos e que não contribuem em absolutamente em nada para o aumento/diminuição das vendas. Atributos como o código identificador do produto (*Item_Identifier*) e código identificador do mercado (*Outlet_Identifier*) - por esse motivo que o Cientista de Dados deve entender do negócio.

Ao verificarmos a função *info()* descobrimos ainda que o atributo alvo (*Item_Outlet_Sales*) que indica a quantidade de produtos vendidos possui dados nulos (ou seja, também não servem para previsão).

```

1 df = df.drop(df[df['Item_Outlet_Sales'].isnull()].index)
2 df = df.drop(columns=['Item_Identifier', 'Outlet_Identifier'], axis=1)

```

Cuidado pois se aplicamos um corte seco como: `df.dropna(how='any', inplace=True)` teremos somente 4.650 observações (devido a eliminação dos valores nulos contidos em outros atributos) - ou seja

perdemos quase 10.000 observações. Lembrar que o tratamento dos nulos deve ser cirúrgico e criterioso. Ao aplicar o corte corretamente somente do atributo alvo ficamos com 8.523 observações. Além disso removemos os preditores que não serviam.

Nosso próximo problema com nulos é nos atributos: peso do item (*Item_Weight*) e tamanho da loja (*Outlet_Size*). Em um caso de dados real devemos procurar preencher esses valores solicitando a informação necessária aos responsáveis, porém para fins desse trabalho iremos remover essa colunas também.

```
1 df = df.drop(columns=['Item_Weight', 'Outlet_Size'], axis=1)
```

Não temos mais a presença de nulos, mas ainda temos problemas, precisamos verificar os atributos não numéricos das observações, isto é: conteúdo de gordura (*Item_Fat_Content*), tipo do item (*Item_Type*), localização da loja (*Outlet_Location_Type*) e tipo da loja (*Outlet_Type*). Para isso:

```
1 print("Gordura:", df['Item_Fat_Content'].unique())
2 print("Tipo:", df['Item_Type'].unique())
3 print("Loc. Loja:", df['Outlet_Location_Type'].unique())
4 print("Tipo Loja:", df['Outlet_Type'].unique())
```

O atributo *Item_Fat_Content* possui uma faixa com os seguintes valores: '*LF*', '*Low Fat*', '*Regular*', '*low fat*' ou '*reg*'. Obviamente só existem dois tipos: '*Low Fat*' e '*Regular*' os outros três são variações desses valores. Para corrigir isso e realizar sua conversão:

```
1 df['Item_Fat_Content'] = df['Item_Fat_Content'].map({'LF': 1, 'Low Fat': 1, 'low
   fat': 1, 'reg': 2, 'Regular': 2})
2 df['Item_Fat_Content'] = df['Item_Fat_Content'].astype(pd.Int64Dtype())
3 df['Outlet_Location_Type'] = df['Outlet_Location_Type'].map({'Tier 1': 1, 'Tier 2':
   2, 'Tier 3': 3})
4 df['Outlet_Location_Type'] = df['Outlet_Location_Type'].astype(pd.Int64Dtype())
5 df['Outlet_Type'] = df['Outlet_Type'].map({'Supermarket Type1': 1, 'Supermarket
   Type2': 2, 'Supermarket Type3': 3, 'Grocery Store': 4})
6 df['Outlet_Type'] = df['Outlet_Type'].astype(pd.Int64Dtype())
```

Criamos um dicionário com as faixas, repetimos os mesmos valores para os tipos que são semelhantes e realizamos a troca dos elementos no *DataFrame*. Aplicamos também a mesma prática para a localização e tipo da loja que possui poucos valores. Porém ainda temos o caso de tipo do item que vamos trocá-lo de uma forma diferente (é ideal quando existem muitos valores diferentes).

Cada atributo tem um tipo determinado, por exemplo, *float* aceita números com pontos decimais, *int* numéricos inteiros, *string* caracteres, além disso *Python* trabalha com um tipo especial denominado *category*. Corresponde a uma determinada faixa de valores. Converter tipo do item em atributo categórico:

```
1 df['Item_Type'] = df.Item_Type.astype('category')
```

Uma vez realizado esse processo podemos "codificá-lo":

```
1 le_Item_Type = LabelEncoder()
2 df['Item_Type'] = le_Item_Type.fit_transform(df['Item_Type'])
3 df.head()
```

Para cada valor categorizado é atribuído um valor numérico (ou seja o mesmo trabalho que tivemos para o mapa). Usamos as funções `info()` e `describe()` e podemos partir para a próxima etapa sem quaisquer problemas com os dados, pois agora são todos numéricos e não possuem qualquer valor nulo.

4.7.4 Separação e treino

Separar em treino e teste (para avaliarmos nosso modelo) e remover o atributo alvo:

```
1 target = df['Item_Outlet_Sales']
2 df = df.drop(columns=['Item_Outlet_Sales'], axis=1)
3 X_train, X_test, y_train, y_test = train_test_split(df, target, test_size = .2)
4 print('Amostra de Treino:', X_train.shape)
5 print('Amostra de Teste:', X_test.shape)
```

Usamos um valor de 20% para nosso teste e temos: 6.818 observações para treino e 1.705 de teste. Treinamos nosso modelo e verificamos seu resultado:

```
1 clf = LinearRegression()
2 clf.fit(X_train, y_train)
3 print('Acurácia: ', clf.score(X_test, y_test))
```

E obtemos uma acurácia aproximada de 42% e qual o motivo dessa discrepância tão grande? Simples, estamos cada vez mais perto da realidade e podemos verificar que realizar previsões com altos scores e poucas ações não existe. Pois se fosse assim: Corramos para treinar um modelo que nos dará os seis números da MegaSena. Ou ao menos nos dizer quando vai chover corretamente com muito pouco trabalho. Por fim podemos ver como os dados estão bem discrepantes em relação ao que foi predito e o real:

```
1 y_pred = model.predict(X_test)
2 plt.plot(y_test, y_test)
3 plt.scatter(y_test, y_pred, c = 'red', marker='+')
4 plt.ylabel('Real')
5 plt.xlabel('Predito')
6 plt.show()
```

Obtemos como resultado:

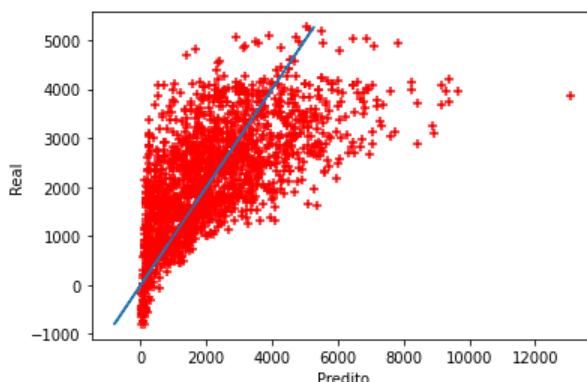
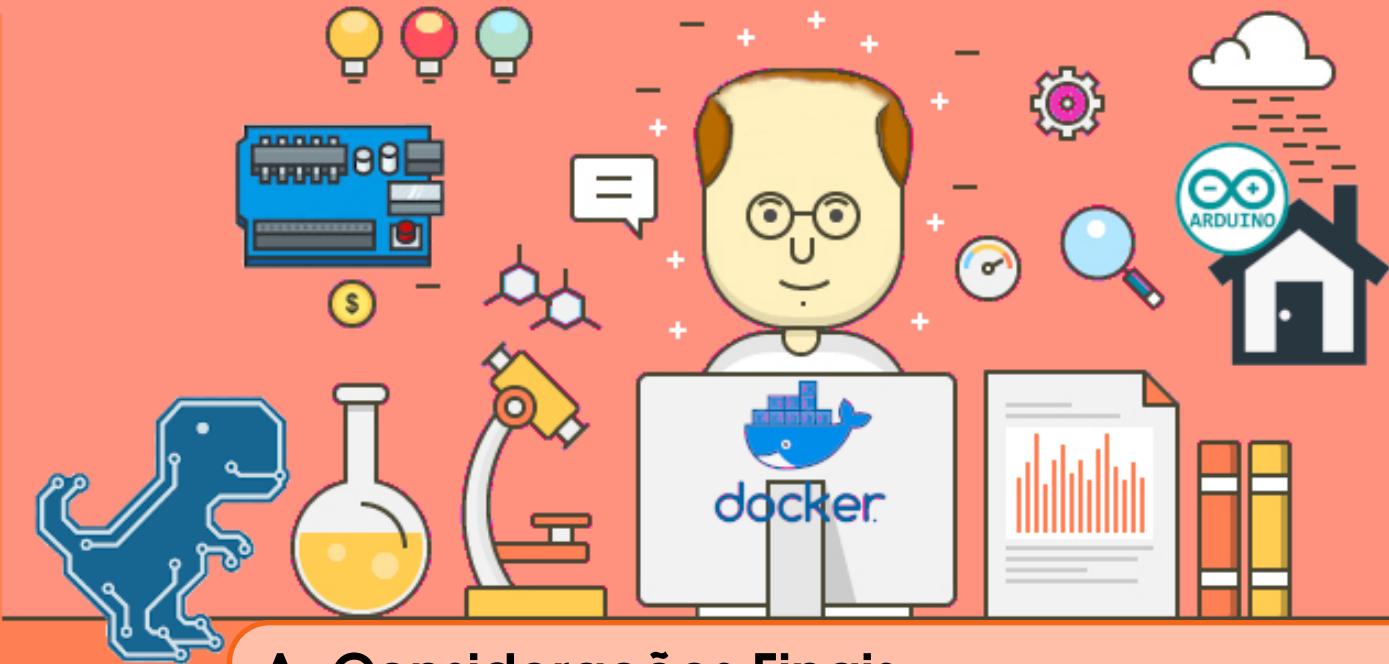


Figura 4.22: Regressão Linear aplicada a vários atributos

Em vermelho são a relação entre o valor real e o que foi predito, a linha azul mostra a Reta da Regressão. Verificamos que temos um ponto bem isolado? Pode ser um *outlier*? Exatamente por esse motivo que passamos um bom tempo em EDA.



A. Considerações Finais

F Você não pode ensinar nada a ninguém, mas pode ajudar a pessoas a descobrirem por si mesmas.
(Galileu Galilei - Físico)

Os artigos deste livro foram selecionados das diversas publicações que fiz no Linkedin e encontradas em outros sites que foram nesta obra explicitamente citadas. Acredito que apenas com a prática podemos almejar o cargo de Cientista de Dados, então segue uma relação de boas bases que podemos encontrar na Internet:

- 20BN-SS-V2: <https://20bn.com/datasets/something-something>
- Actualitix: <https://pt.actualitix.com/>
- Banco Central do Brasil: <https://www3.bcb.gov.br>
- Banco Mundial: <http://data.worldbank.org>
- Censo dos EUA (População americana e mundial): <http://www.census.gov>
- Cidades Americanas: <http://datasf.org>
- Cidade de Chicago: <https://data.cityofchicago.org/>
- CIFAR-10: <https://www.cs.toronto.edu/~kriz/cifar.html>
- Cityscapes: <https://www.cityscapes-dataset.com/>
- Criptomoedas: <https://pro.coinmarketcap.com/migrate/>
- Dados da União Europeia: <http://open-data.europa.eu/en/data>
- Data 360: <http://www.data360.org>
- Datahub: <http://datahub.io/dataset>
- DBpedia: <http://wiki.dbpedia.org/>
- Diversas áreas de negócio e finanças: <https://www.quandl.com>
- Diversos assuntos: <http://www.firebaseio.com>
- Diversos países (incluindo o Brasil): <http://knoema.com>
- Fashion-MNIST: <https://www.kaggle.com/zalando-research/fashionmnist>
- Gapminder: <http://www.gapminder.org/data>

- Google Finance: <https://www.google.com/finance>
- Google Trends: <https://www.google.com/trends>
- Governo do Brasil: <http://dados.gov.br>
- Governo do Canadá (em inglês e francês): <http://open.canada.ca>
- Governo dos EUA: <http://data.gov>
- Governo do Reino Unido: <https://data.gov.uk>
- ImageNET: <http://www.image-net.org/>
- IPEA: <http://www.ipeadata.gov.br>
- IMDB-Wiki: <https://data.vision.ee.ethz.ch/cvl/rrothe/imdb-wiki/>
- Kinetics-700: <https://deepmind.com/research/open-source/kinetics>
- Machine Learning Databases: <https://archive.ics.uci.edu/ml/machine-learning-databases/>
- MEC Microdados INEP: <http://inep.gov.br/microdados>
- MS coco: <http://cocodataset.org/#home>
- MPII Human Pose: <http://human-pose.mpi-inf.mpg.de/>
- Músicas: <https://aws.amazon.com/datasets/million-song-dataset>
- NASA: <https://data.nasa.gov>
- Open Data Monitor: <http://opendatamonitor.eu>
- Open Data Network: <http://www.opendatanetwork.com>
- Open Images: <https://github.com/openimages/dataset>
- Portal de Estatística: <http://www.statista.com>
- Públicos da Amazon: <http://aws.amazon.com/datasets>
- R-Devel: <https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/00Index.html>
- Reconhecimento de Faces: <http://www.face-rec.org/databases>
- Saúde: <http://www.healthdata.gov>
- Statsci: <http://www.statsci.org/datasets.html>
- Stats4stem: <http://www.stats4stem.org/data-sets.html>
- Stanford Large Network Dataset Collection: <http://snap.stanford.edu/data>
- Vincent Rdatasets: <https://vincentarelbundock.github.io/Rdatasets/datasets.html>
- Vitivinicultura Embrapa: <http://vitibrasil.cnpuv.embrapa.br/>

Esse não é o fim de uma jornada acredito ser apenas seu começo. Espero que este livro possa lhe servir para criar algo maravilhoso e fantástico que de onde estiver estarei torcendo por você.

A.1 Sobre o Autor

Fortes conhecimentos em linguagens de programação Java e Python. Especialista formado em Gestão da Tecnologia da Informação com forte experiência em Bancos Relacionais e não Relacionais. Possui habilidades analíticas necessárias para encontrar a providencial agulha no palheiro dos dados recolhidos pela empresa. Responsável pelo desenvolvimento de dashboards com a capacidade para análise de dados e detectar tendências, autor de 15 livros e diversos artigos em revistas especializadas, palestrante em seminários sobre tecnologia. Focado em aprender e trazer mudanças para a organização com conhecimento profundo do negócio.

- Perfil no Linkedin: <http://www.linkedin.com/pub/fernando-anselmo/23/236/bb4>
- Endereço do Git: <https://github.com/fernandoans/machinelearning>

Machine Learning na Prática

ESTE LIVRO PODE E DEVE SER DISTRIBUÍDO LIVREMENTE

Fernando Anselmo