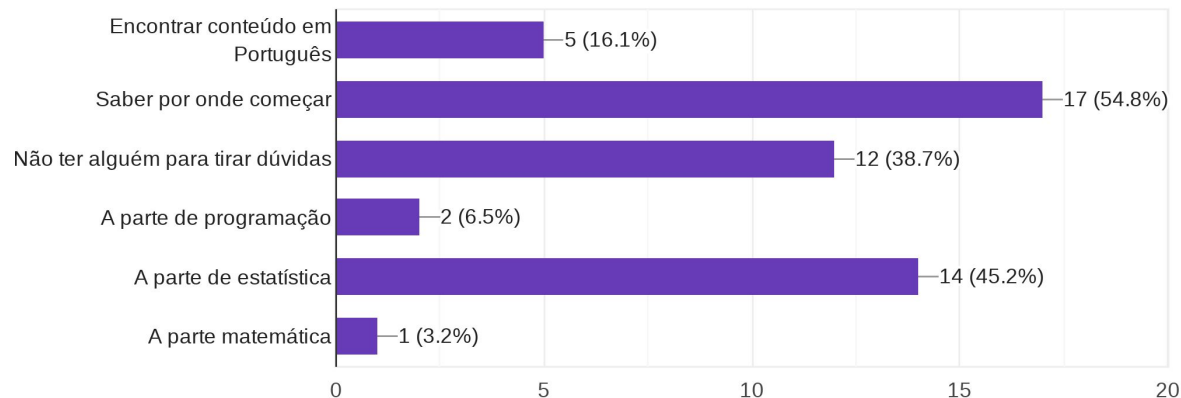

Sessão 1

WiDS Recife Live Coding, 14/12/2019

Dificuldades na área de Data Science

O que você acha mais difícil em data science?

31 respostas



Fonte: pesquisa realizada em novembro/2019

Como serão os live codings?

- Sessões ao-vivo todos os sábados das 14h às 15h
 - Código e slides serão disponibilizados no nosso site
 - O objetivo é treinar para participar do [Datathon](#) em 2020
-

Que ferramentas e tecnologias iremos utilizar?

- [Jupyter Notebook](#) no [Google Colab](#)
- Python
- Pandas
- Scikit-learn

Do que eu preciso para participar?

- Um computador com internet
- Um browser
- Uma conta no gmail (com o Google Colab não precisamos instalar nada no computador)

Começando!

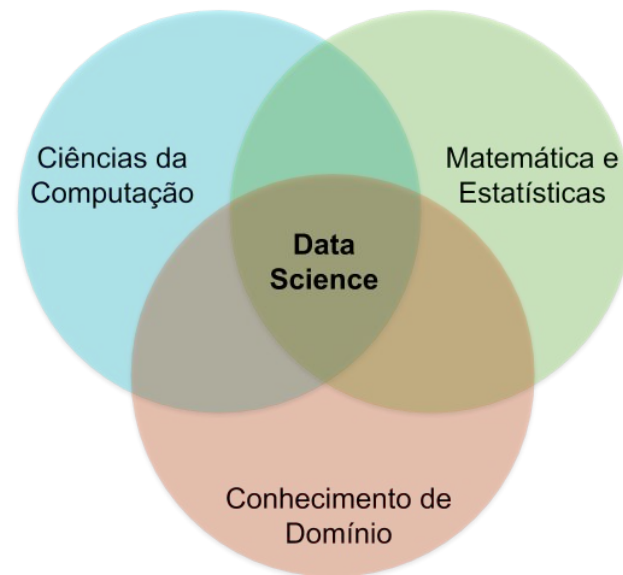
Roteiro

- O que é um modelo em Aprendizagem de Máquina
- O que é um Jupyter Notebook
- Como usar o Google Colab
- Iniciando com o pandas e o scikit-learn
- Treinando nosso primeiro modelo

O que faz uma Cientista de Dados?

Algumas tarefas comuns:

- Programa
- Manipula dados (coletar, limpar e transformar os dados)
- Faz análises com estatística descritiva
- Cria visualizações de dados
- Treina **modelos estatísticos**



O que é um modelo?

- Modelo Preditivo = Algoritmo de aprendizagem de máquina + dados
 - Etapas para treinar um modelo
 - a. Importar os dados
 - b. Limpar e transformar os dados
 - c. Passar dados para o algoritmo de aprendizagem de máquina
 - d. Salvar o modelo
-

O que é um modelo?

- Etapas para fazer previsões com um modelo
 - a. Importar o dado
 - b. Limpar e transformar o dado
 - c. Passar dado para o modelo
 - d. Obter o resultado
-

Treinando nosso primeiro modelo

O problema

- Nosso conjunto de dados é composto por atributos de vestidos
- Queremos treinar um modelo onde a gente envie os atributos do vestido e ele diga **qual a melhor época do ano para usá-lo**

O problema

Classificação

- Esse é um problema de **classificação**:
 - Dados usados para prever uma **categoria**
 - Nossa categoria será a **época do ano**

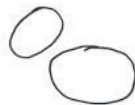
Etapas para resolver o problema

1. Importar os dados
 2. Explorar os dados
 3. Dividir os dados em um conjunto de treinamento e um conjunto de teste
 4. Treinar o modelo com o conjunto de treinamento
 5. Avaliar o modelo com o conjunto de testes
-

Como a gente
aprende?

HOW TO: DRAW A HORSE

BY VAN OKTOP



① DRAW 2 CIRCLES



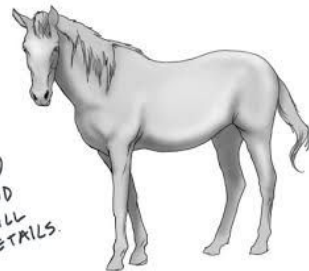
② DRAW THE LEGS



③ DRAW THE FACE



④ DRAW THE HAIR



⑤
ADD
SMALL
DETAILS.

Como a gente aprende?

- Você não apenas pensa sentada, você pensa falando, escrevendo, se comunicando com os outros etc.
- Quando estamos iniciando precisamos de instruções passo-a-passo, **não se preocupe em saber como exatamente tudo funciona**

Como a gente aprende?

- Como fazer uma cirurgia no cérebro:
 - Se você acha que pode ficar mais inteligente você efetivamente se torna mais inteligente
 - Se você acha que consegue fazer algo você vai conseguir fazer

Como a gente aprende?

- É preciso definir metas e objetivos (coisas pra fazer a fim de alcançar as metas). Traçar detalhadamente as metas e os objetivos. Setar prazos e não deixar os estudos pra "quando tiver tempo livre"

Como a gente aprende?

- Fingir que você pode errar faz você acertar. Se permita falhar

**Colocando a mão na
massa**

1. Importar os dados

- [https://archive.ics.uci.edu/ml/datasets/Dresses Attribute Sales](https://archive.ics.uci.edu/ml/datasets/Dresses+Attribute+Sales)
 - https://github.com/widsrecife/dados/blob/master/Sessao_1/Attribute%20DataSet.csv
-

2. Explorar os dados

- 2. Explorar os dados:
 - Normalmente a gente chama cada coluna de **feature**
 - Precisamos separar a feature que será usada como resultado da “saída” do modelo (no nosso caso é a época do ano)
 - É uma boa ideia também ver a distribuição dos dados de cada feature

3. Dividir os dados

- 3. Dividir os dados
 - https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

4. Treinar o modelo

- 4. Treinar o modelo
 - Precisamos converter os dados de String para números
 - Vamos usar o scikit-learn para fazer isso

—

Scikit-learn

Scikit-learn

- A API do scikit
 - **estimator**: interface para construir e treinar modelos, expõe o método `fit()`
 - **predictor**: interface para realizar predições, expõe o método `predict()`
 - **transformer**: interface para transformar dados, expõe o método `transform()`
-

Scikit-learn

- A API do scikit
 - Métodos principais
 - `fit()`
 - `transform()`
 - `fit_transform()`
 - `predict()`
-

Scikit-learn

- OrdinalEncoder()
 - Queremos transformar strings em números
 - Exemplo:
 - ["Vermelho","Roxo"] -> [1,2]
 - ["Vermelho","Vermelho"] -> [1,1]
-

Tarefinha de casa

Tarefinha de casa

Responder as perguntas ao lado

1. O que é o Pandas e como ele nos ajuda a manipular dados?
 2. O Pandas introduz dois novos tipos de dados: DataFrame e Series. Qual a diferença entre os dois?
 3. O que é o Scikit-learn e que métodos você precisa chamar para converter dados de string para número?
-

Obrigada!

E até semana que vem!

Referências

- <https://medium.com/trainingcenter/data-science-um-panorama-geral-87edbbd35885>
- [https://archive.ics.uci.edu/ml/datasets/Dresses Attribute Sales](https://archive.ics.uci.edu/ml/datasets/Dresses+Attribute+Sales)
- <https://docs.microsoft.com/pt-br/dotnet/machine-learning/resources/glossary#classification>
- <https://pragprog.com/book/ahptl/pragmatic-thinking-and-learning>
- <https://usablelearning.com/the-book/>
- <https://www.deborahmesquita.com/2017-02-03/por-onde-comecar-um-projeto-de-data-science>