Introdução à Ciência de Dados com Python

Marcos Wenneton Araújo

UEA UNIVERSIDADE DO ESTADO DO AMAZONAS

PyData MANAUS

INDT

Juliany Raiol

pyladies Manaus

PyData MANAUS

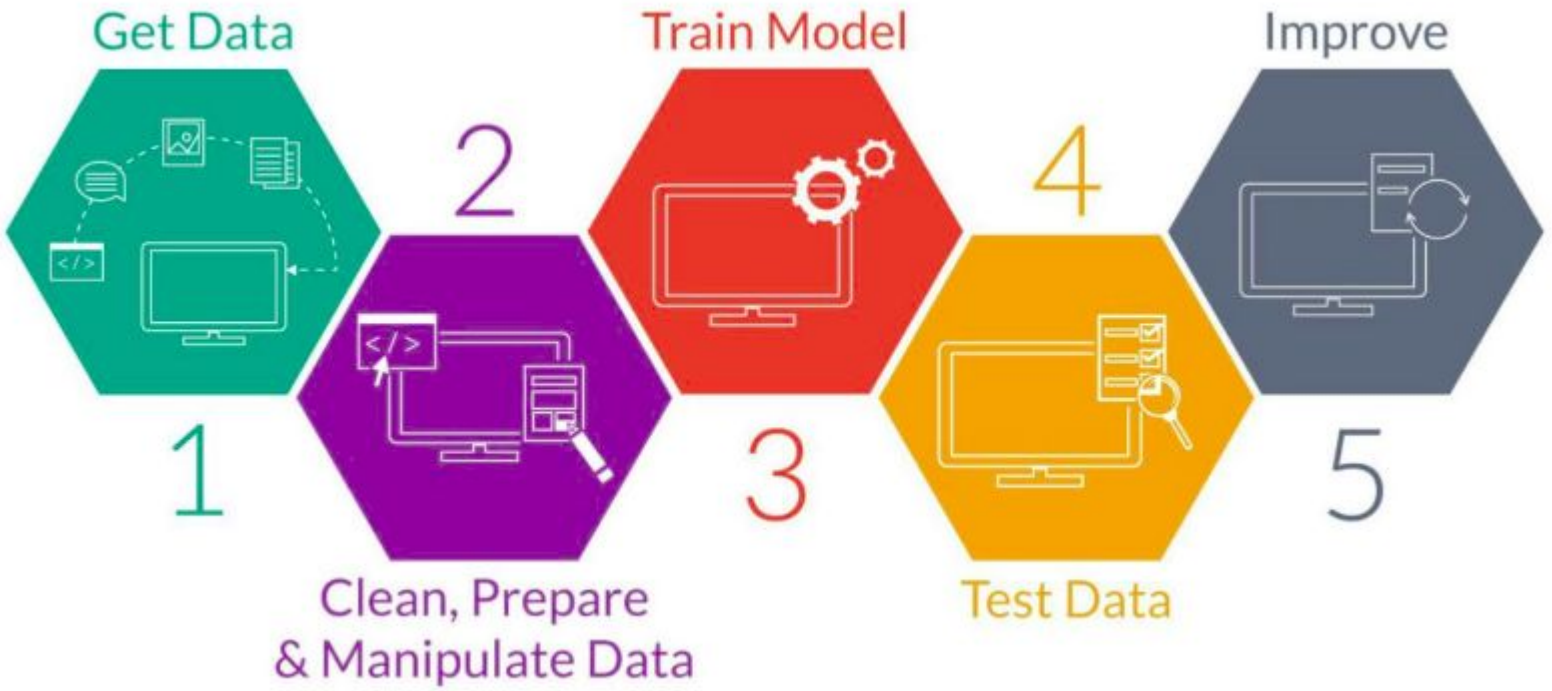
Ciência de Dados

"Estudo e Análise de Dados que visa a extração de conhecimento ou *insights* para possíveis tomadas de decisão"

Análise de Dados

Transformar um conjunto de dados com o objetivo de poder verificá-lo melhor.

Fluxo de trabalho do Aprendizado de Máquina



#1 Encontrar ou Coletar um Conjunto de Dados

IMDb

460,723 images

Wikipedia

62,328 images

#2

Limpeza

```
In [4]: df.head(5)
```

Out[4]:

	dob	photo_taken	full_path	gender	name	face_location	face_score	second_face_score	celeb_id
0	693726	1968	[01/nm0000001_rm124825600_1899-5-10_1968.jpg]	1.0	[Fred Astaire]	[[1072.926, 161.838, 1214.7839999999999, 303.6...	1.459693	1.118973	6488
1	693726	1970	[01/nm0000001_rm3343756032_1899-5-10_1970.jpg]	1.0	[Fred Astaire]	[[477.184, 100.352, 622.592, 245.76]]	2.543198	1.852008	6488
2	693726	1968	[01/nm0000001_rm577153792_1899-5-10_1968.jpg]	1.0	[Fred Astaire]	[[114.96964308962852, 114.96964308962852, 451....	3.455579	2.985660	6488
3	693726	1968	[01/nm0000001_rm946909184_1899-5-10_1968.jpg]	1.0	[Fred Astaire]	[[622.8855056426588, 424.21750383700805, 844.3...	1.872117	NaN	6488
4	693726	1968	[01/nm0000001_rm980463616_1899-5-10_1968.jpg]	1.0	[Fred Astaire]	[[1013.8590023603723, 233.8820422075853, 1201....	1.158766	NaN	6488

Dados Originais

#2

Limpeza

	image_name	class_id
0	imdb/45/nm0876645_rm1932442112_1959-3-14_1999.jpg	40
1	imdb/45/nm1295245_rm2855187968_1985-3-16_2011.jpg	26
2	imdb/56/nm0005256_rm858435072_1971-4-28_2012.jpg	41
3	imdb/32/nm0852132_rm676826112_1971-7-30_2004.jpg	33
4	imdb/27/nm0001427_rm3004406528_1963-6-17_2000.jpg	37

Dados a serem utilizados

#3

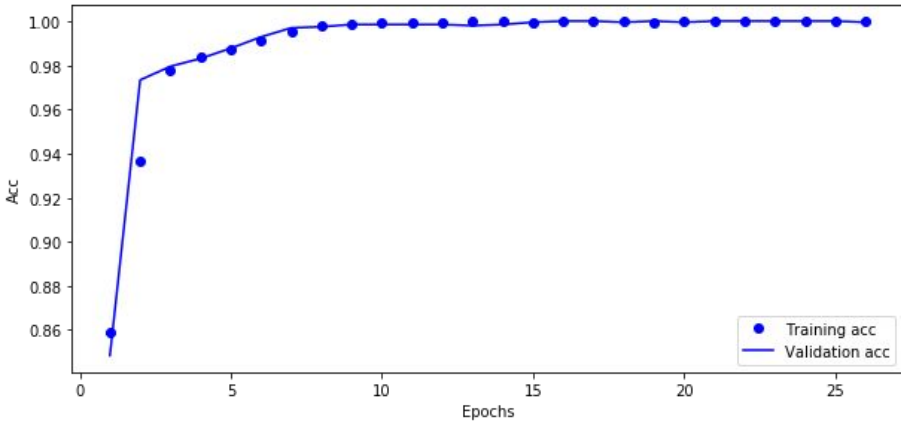
Treinar

```
[ ]  learn = cnn_learner(data, model,
                         metrics=[accuracy, FBeta(average='micro', beta=1)],
                         callback_fns=[partial(EarlyStoppingCallback, monitor='valid_loss', patience=20)],
                         pretrained=False
                        )
```

```
[ ]  learn.fit_one_cycle(200)
```

18.50% [37/200 1:07:32<4:57:34]

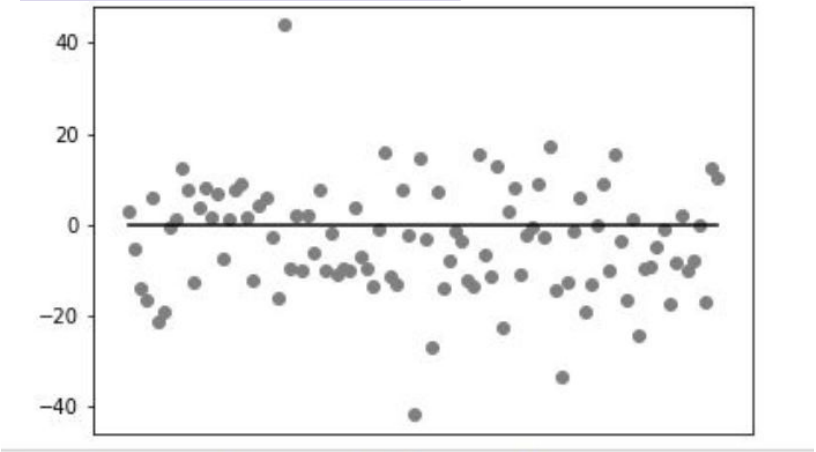
epoch	train_loss	valid_loss	accuracy	f_beta	time
0	0.587283	0.462057	0.832049	0.832049	01:49
1	0.430720	0.341835	0.864920	0.864920	01:51
2	0.374721	0.434729	0.845917	0.845917	01:50
3	0.337999	0.273992	0.891628	0.891628	01:50
4	0.314837	0.426782	0.848998	0.848998	01:50
5	0.300933	0.263492	0.888033	0.888033	01:51

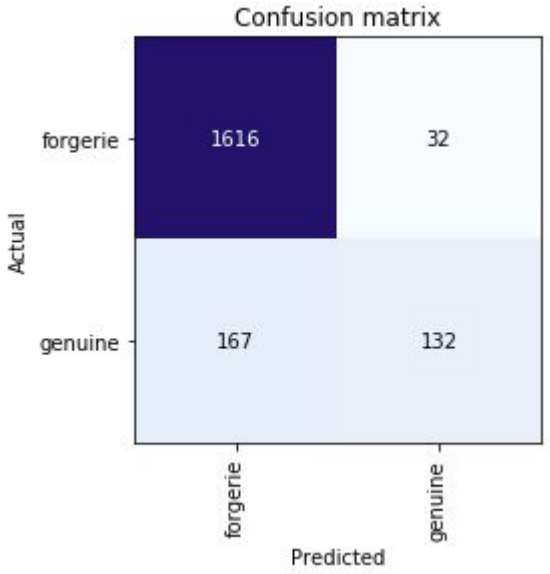


#4

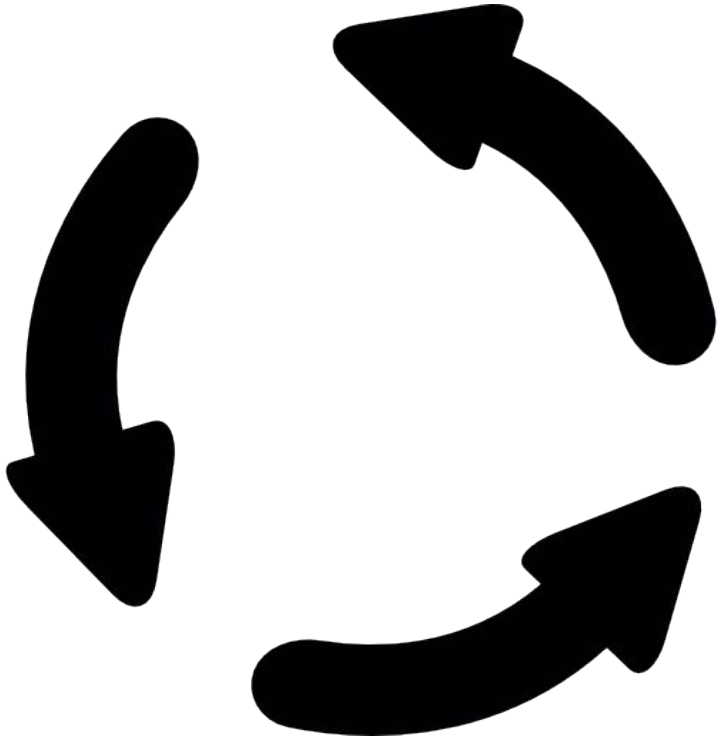
Testar

```
In [19]: fig = df_results.plot(x=df_results.index, y='Zero', color='black')
         fig.axes.scatter(x=df_results.index, y=df_results['Ŷ Previsto - Y Real'],color='grey')
         fig.axes.get_xaxis().set_visible(False)
         fig.axes.get_legend().set_visible(False)
```



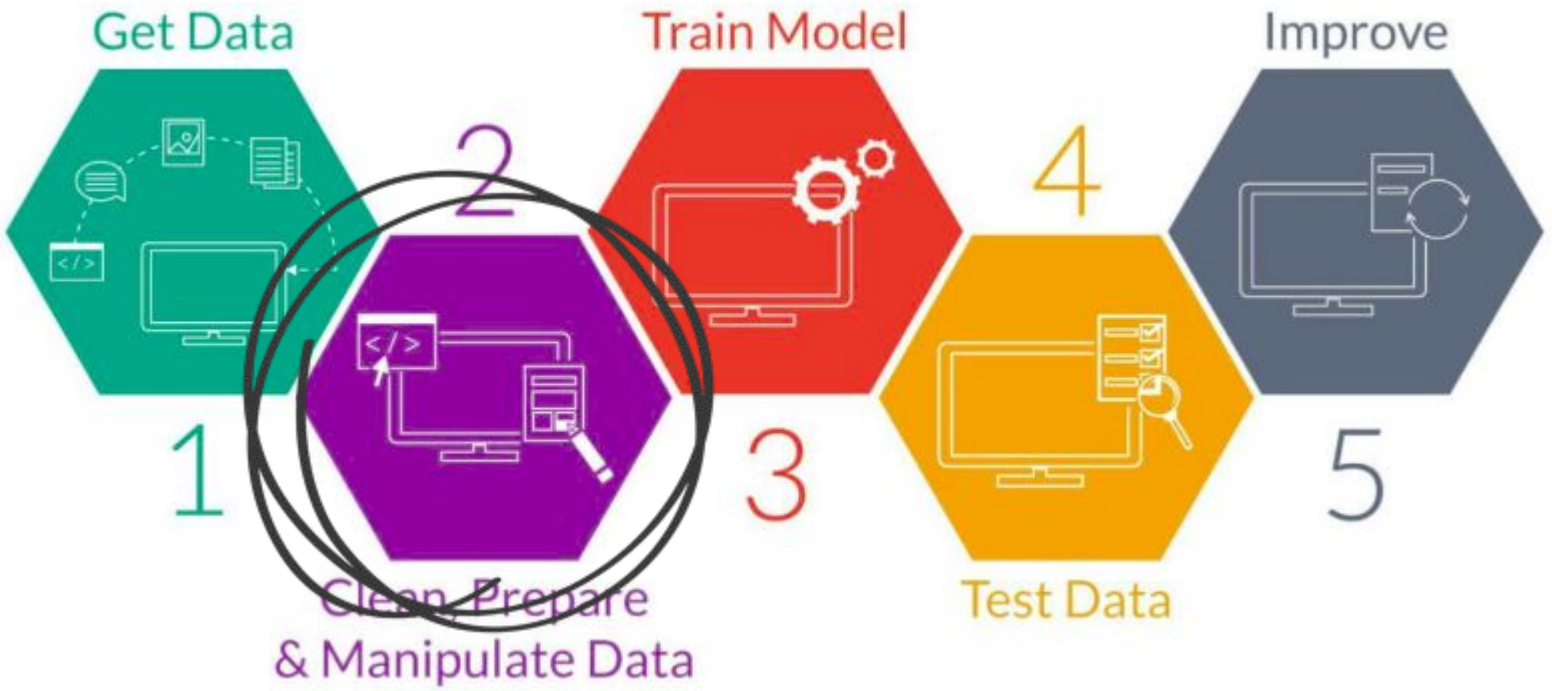


#5 Repetir até conseguir bons resultados



Tarefas de Aprendizado

Escopo deste Minicurso



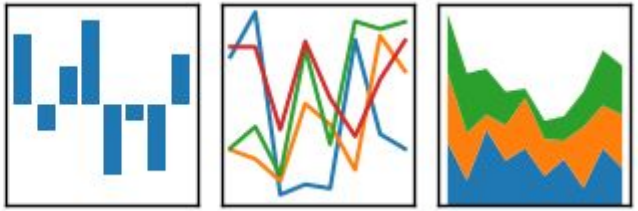
Descrição dos Dados Utilizados

Conjunto de dados com informações sobre os apps disponíveis na Google Play Store Apps, retirado da plataforma kaggle

Pandas

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

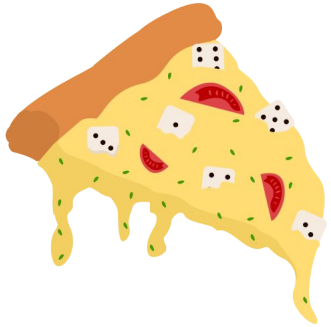


Biblioteca open-source em Python que fornece ferramentas de análise de dados e estruturas de dados de alta performance e **fáceis** de usar**.**

Ferramentas

- Para a conclusão do minicurso iremos utilizar o **Jupyter Notebook**;
- Permite a execução de código em células, permitindo um bom uso em tarefas de análise de dados.

Próximos passos



Podcast: Pizza de Dados

Programas Nanodegree

UDACITY

Data Science Academy

Siraj Raval no YouTube

coursera

Obrigado!
Dúvidas?


t.me/pydatamanaus


fb.com/pydatamanaus


@pydatamanaus


@pydatamanaus