mulheres em inteligência artificial

AULA 5 - REGRESSÃO

#BootcampMIA2022 #SomosMIA

Quem somos?

Bárbara Barbosa

Data Manager na Orderchamp

- Mestra em Sistemas de Informação - com foco em Inteligência Artificial e NLP
- Organizadora do Rails Girls SP, Women Dev Summit e Women in Data Science SP 2019/2020/2021

bahbbc

bahbbc

Quem somos?

Vivian Yamassaki

Lead Data Scientist na Creditas

- Mestra em Sistemas de Informação pela USP com pesquisa em inteligência artificial e área de aplicação em bioinformática
- Co-fundadora da MIA

vivianyamassaki

vivianyamassaki

Agora que já vimos sobre análise exploratória na aula passada...

Finalmente vamos começar a falar sobre MACHINE LEARNING!!!

Uma inspiração...

"O sucesso só é significativo e prazeroso se você sente que ele lhe pertence”.

(Michelle Obama)

Inteligência Artificial

Classificação

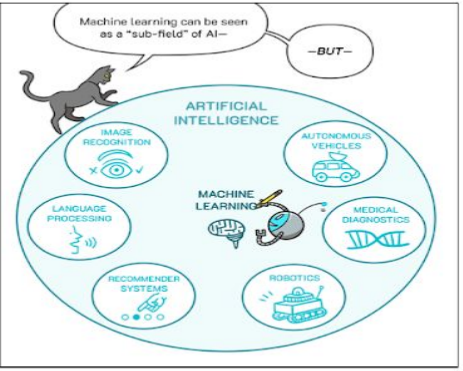
Aprendizado supervisionado

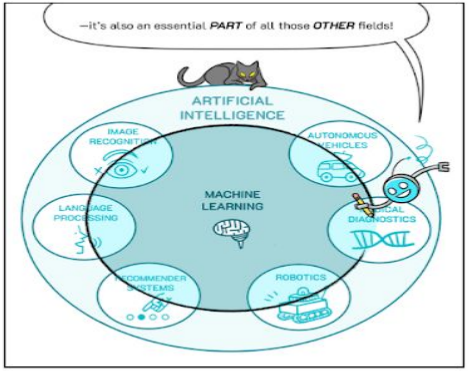
Regressão

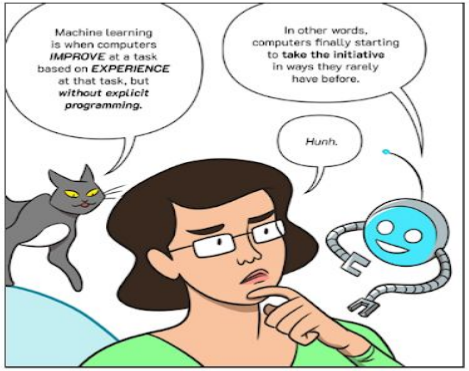
Machine Learning

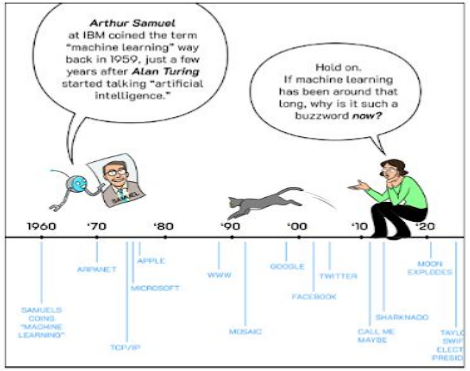
Afinal o que é MACHINE LEARNING?

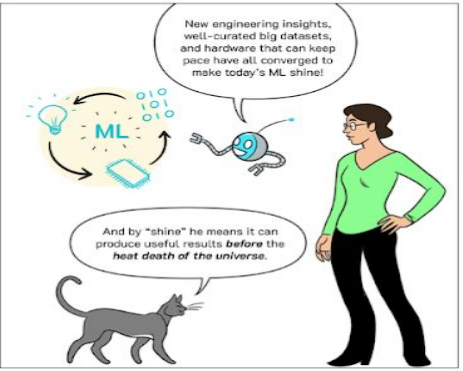
“Aprendizado de Máquina ou Machine Learning é um conjunto de regras e procedimentos, que permite que os computadores possam agir e tomar decisões baseados em dados ao invés de ser explicitamente programados para realizar uma determinada tarefa”

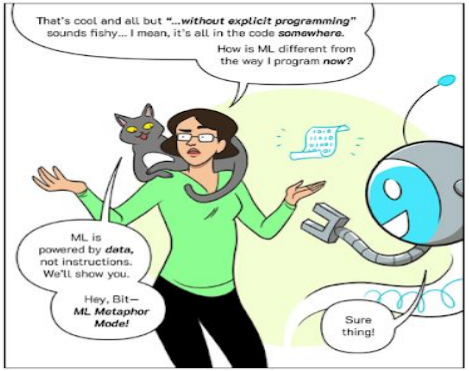


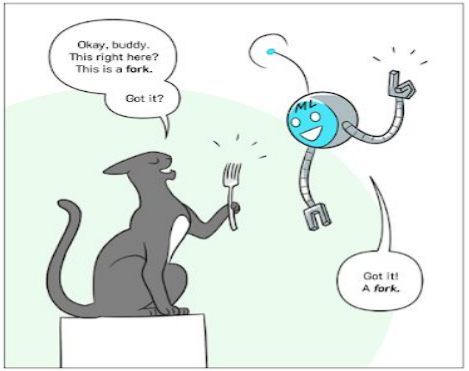












Aqui você consegue ver o quadrinho completo :)

Categorias

- Aprendizado Não-Supervisionado (em que não há labels):
  - Clusterização
- Aprendizado Supervisionado (em que há labels):
  - Classificação
  - Regressão

* Há também o aprendizado semi-supervisionado (em que há alguns dados com labels e a maioria não, mas são métodos menos utilizados e que ainda estão surgindo) e o aprendizado por reforço (em que o programa tenta na verdade encontrar “caminhos”).

Aprendizado supervisionado e não supervisionado

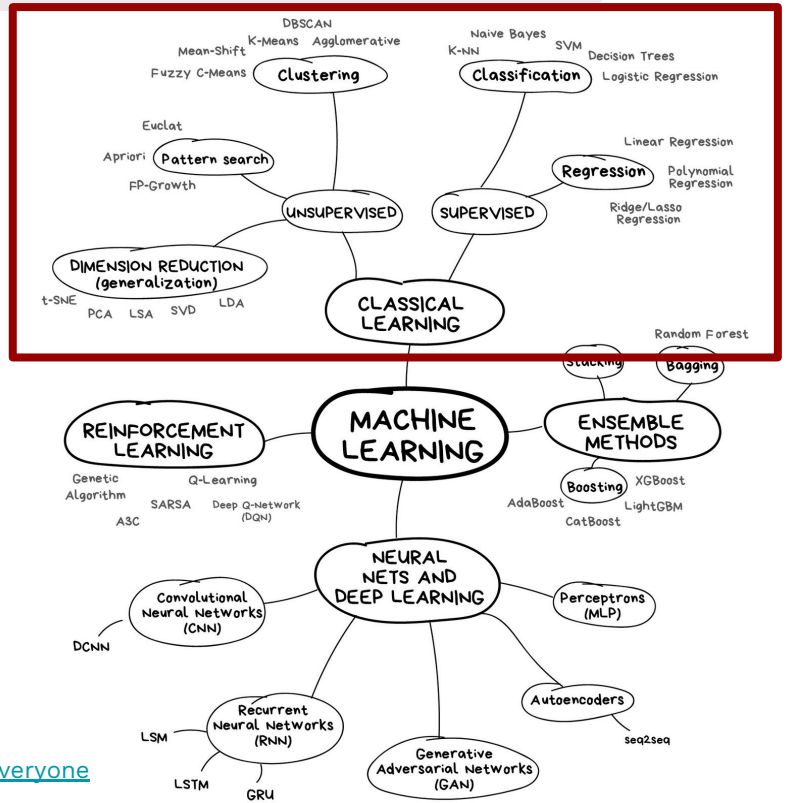


Diagrama retirado do [Machine Learning For Everyone]

Aprendizado supervisionado e não supervisionado

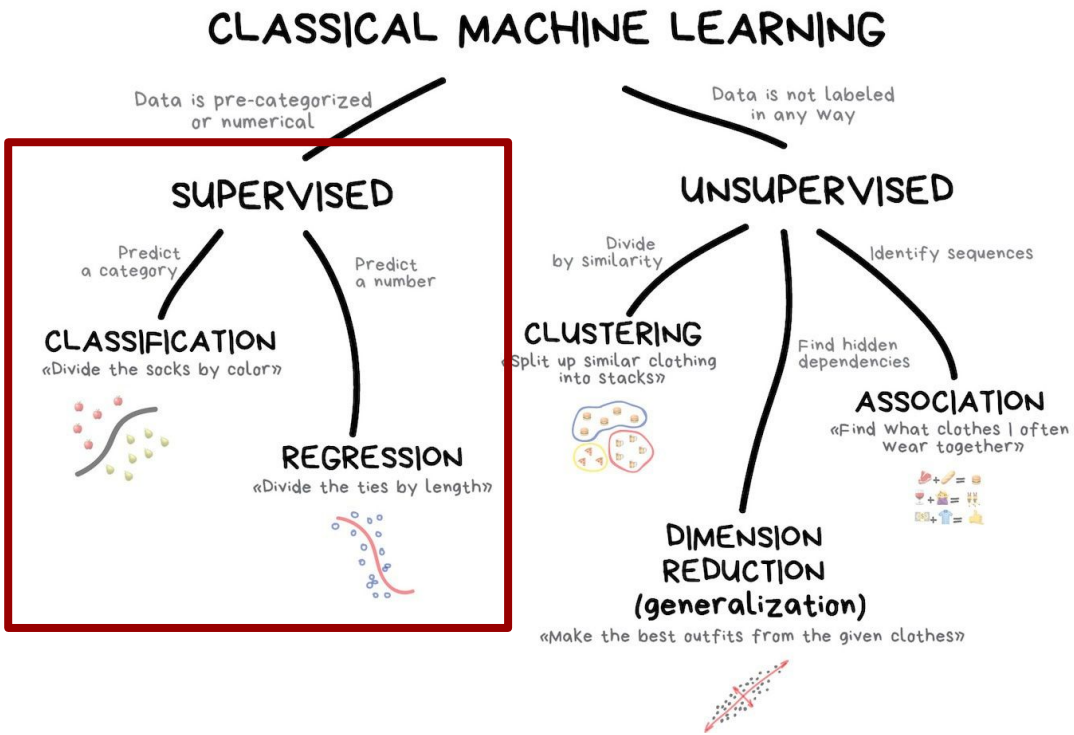


Diagrama retirado do [Machine Learning For Everyone]

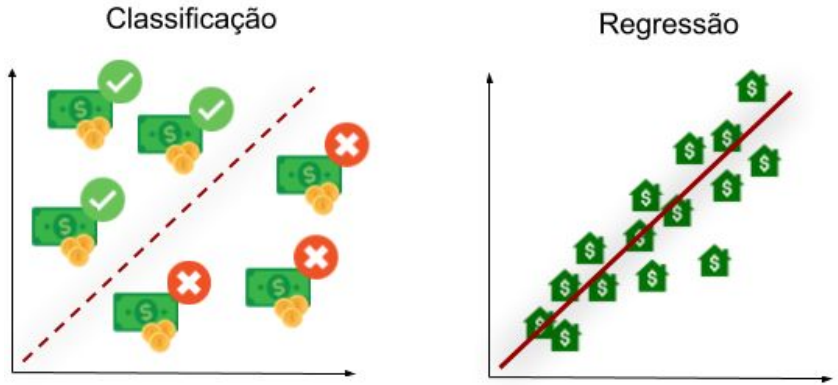
Aprendizado supervisionado e não supervisionado

Aprendizado supervisionado

- Regressão
- Classificação

Aprendizado não supervisionado

- Clustering



Regressão vs Classificação

Classificação vs Regressão

Regressão

Que temperatura fará amanhã em São Paulo?

Classificação

Amanhã fará frio ou calor em São Paulo?

Regressão

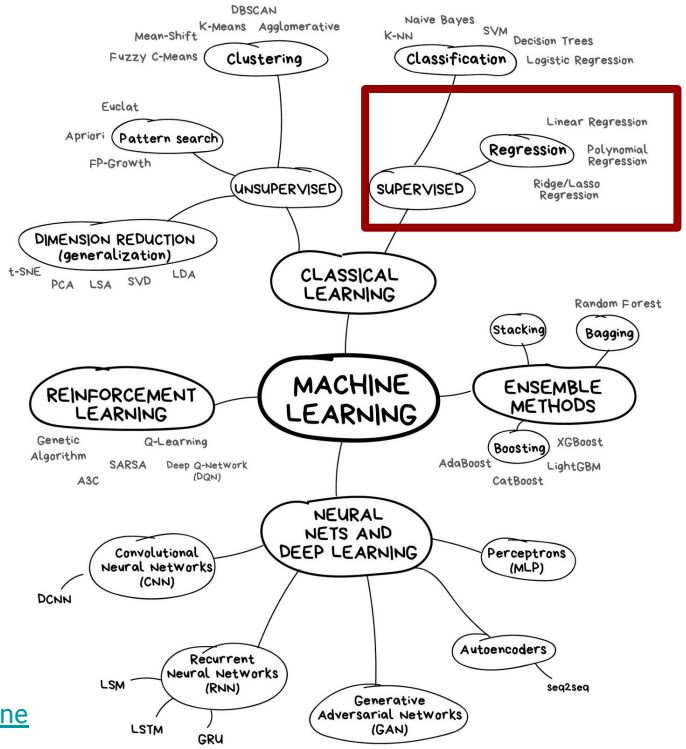


Diagrama retirado do [Machine Learning For Everyone]

Regressão

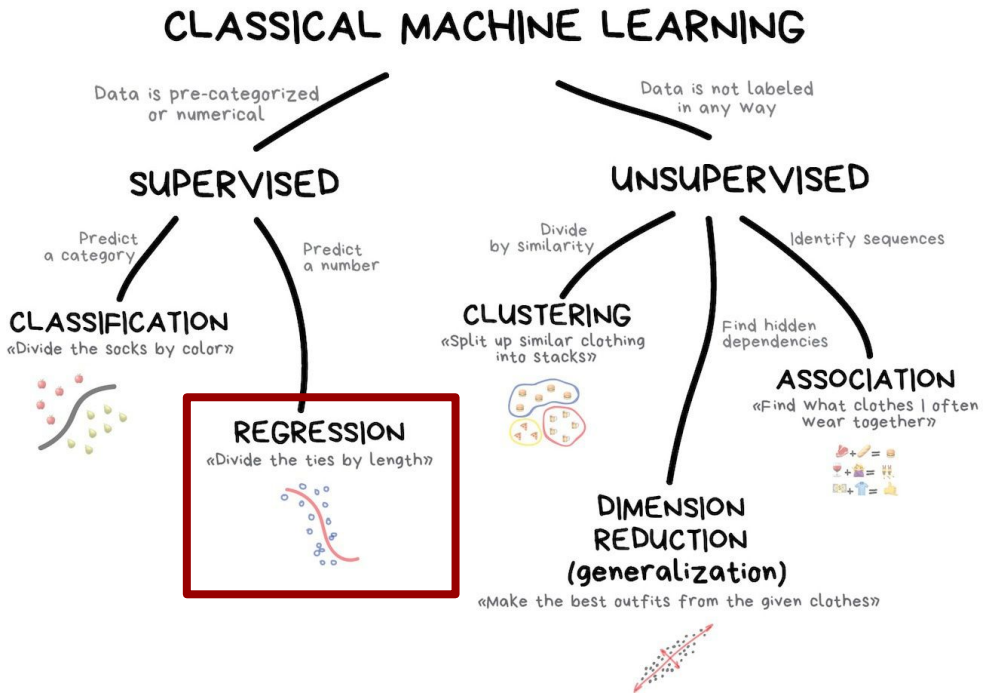


Diagrama retirado do [Machine Learning For Everyone]

Regressão

Na tarefa de regressão, o objetivo é que nosso modelo seja capaz de **predizer um número contínuo**, que ele vai ter aprendido com base em um conjunto de dados já *rotulado*.

	Atributos (features)			Target
	Idade	Renda	Possui dívidas	Limite aprovado do cartão de crédito
Exemplo	18	1.000	Não	1.500
	25	2.500	Sim	1.250
	50	4.500	Sim	4.000
	42	10.000	Não	25.000
	33	6.000	Não	7.500
	27	5.700	Não	10.000

Regressão

Existem diversos algoritmos para resolver esses problemas de regressão. Um deles, que é muito utilizado, é a **Regressão Linear**.

Regressão Linear

Como funciona a Regressão Linear?

Linear regression equation

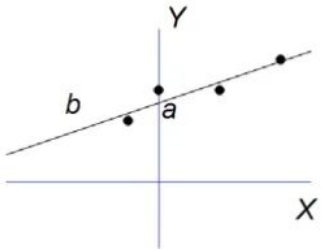
(without error)

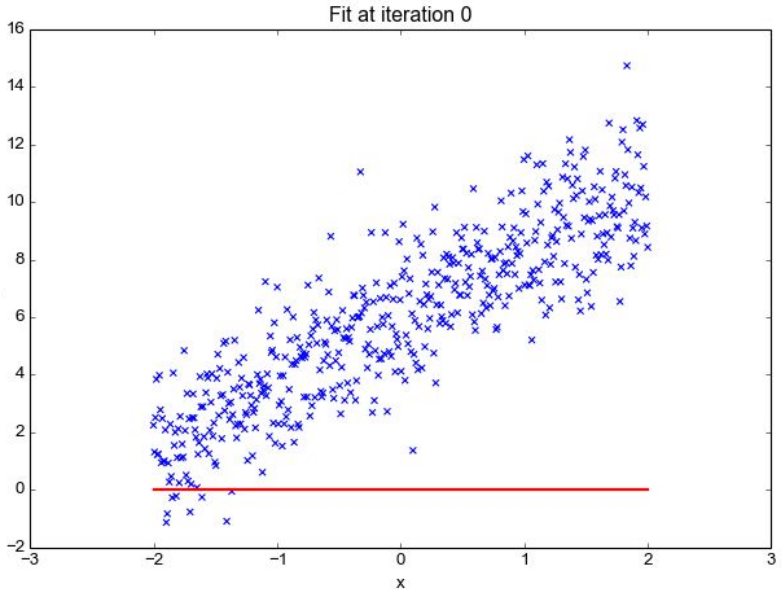
$$\hat{Y} = bX + a$$

predicted values of *Y*

b = slope = rate of predicted ↑/↓ for Y scores for each unit increase in *X*

Y-intercept = level of *Y* when X is 0

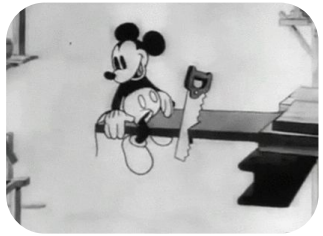


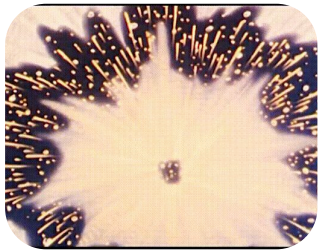


Vamos construir nosso primeiro modelo de regressão?

Seguiremos o seguinte fluxo:

Análise exploratória → **Feature Engineering** → **Modelagem** → **Avaliação**





Pré-processamento dos dados

Nosso primeiro modelo de Regressão

Na aula de hoje, vamos criar um modelo de Regressão para **predizer o valor a ser pago em uma corrida de táxi**.

Feature Engineering

Antes de seguir para a modelagem, precisamos parar para falar um pouquinho melhor sobre a **Feature Engineering**.

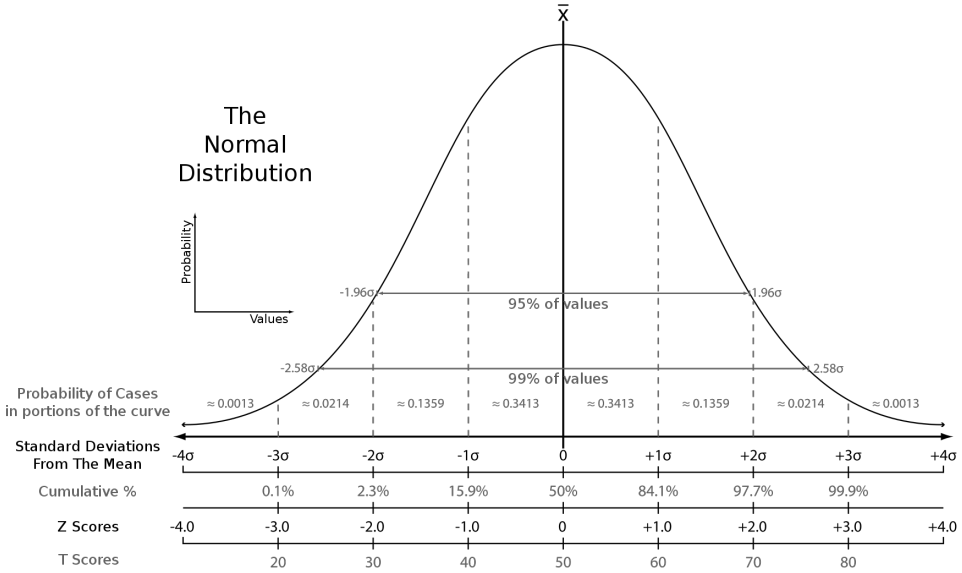
Feature Engineering

A **Feature Engineering** é uma etapa crucial para a criação de modelos de Machine Learning (não só de regressão) e é feita junto com a Análise Exploratória de Dados (EDA).

Algumas coisas que podem ser feitas nessa etapa são:

- Tratamento de valores faltantes
- Tratamento de outliers
- Normalização de dados
- Criação de variáveis

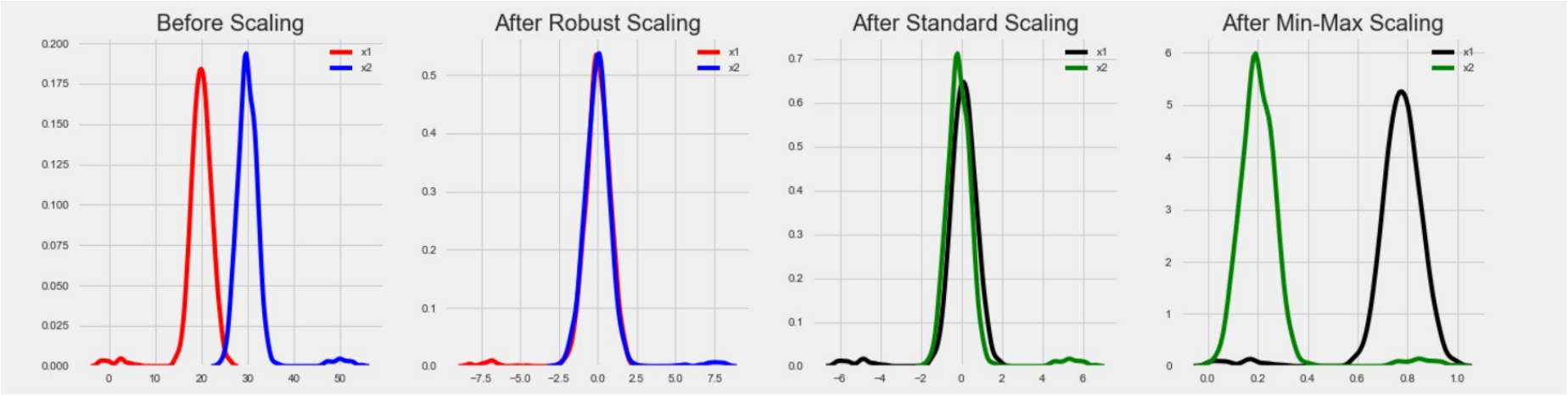
Padronização e Normalização de dados



A Gaussiana é uma distribuição muito importante para a estatística e está ligada a Teoria do limite central

Veja mais: https://www.inf.ufsc.br/~andre.zibetti/probabilidade/normal.html

Padronização e Normalização de dados



Veja mais: https://scikit-learn.org/stable/auto_examples/preprocessing/plot_all_scaling.html#sphx-glr-auto-examples-preprocessing-plot-all-scaling-py

Conjunto de dados

Finalizada a Feature Engineering e a seleção das variáveis que serão utilizadas (mais detalhes sobre isso serão dados na próxima aula!), **temos o nosso conjunto de dados para a modelagem**! \o/

Conjunto de dados

Mas antes de criar o modelo em si, precisamos separar o nosso conjunto de dados em 2:

- **Conjunto de treinamento** é utilizado pelo modelo para aprender
- **Conjunto de teste** é utilizado para avaliar o desempenho do modelo para dados não vistos

* Pode ainda existir um conjunto de validação, conforme explicado aqui.

Pronto! Agora vamos para a prática!!!

Avaliação do modelo

Avaliação do modelo

Existem diversas métricas para avaliar o erro da predição do nosso modelo. Hoje iremos ver as seguintes:

- MSE (Mean Squared Error)
- MAE (Mean Absolute Error)
- R^2

Aqui tem um artigo falando sobre outras métricas de avaliação.

MSE (Mean Squared Error)

O MSE calcula o erro quadrático médio das predições do nosso modelo. **Quanto maior o MSE, pior é o modelo**.

Exemplo	Predição	Target
1	500	600
2	1000	2
3	750	690

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

MSE = ⅓ *((600-500)²+(2-1.000)²+(690-750)²)

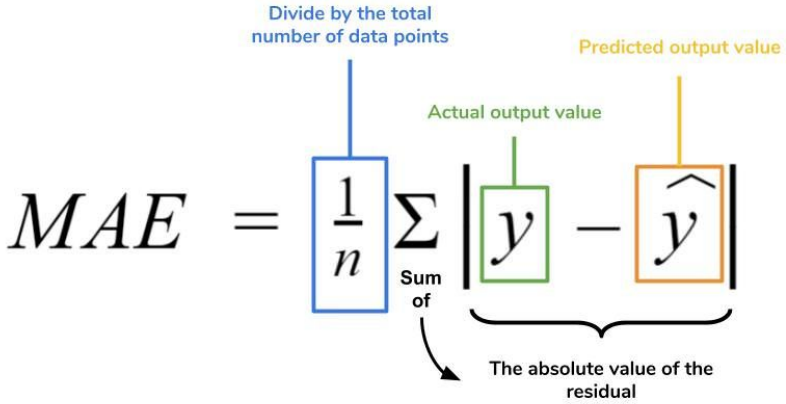
MSE = ⅓ *(10.000+ 996.004+3.600) = ⅓ *1.009.604

MSE = 336.534,66

MAE (Mean Absolute Error)

O MAE calcula a média da diferença absoluta entre o valor predito e o valor real. **Quanto maior o MAE, pior é o modelo**.

Exemplo	Predição	Target
1	500	600
2	1000	2
3	750	690



MAE = ⅓ *(|600-500|+|2-1.000|+|690-750|)

MAE = ⅓ *(100+ 998+60) = ⅓ *1.158

MAE = 386

R^2

O R^2 é uma métrica que indica o quão bom o nosso modelo está em comparação com um modelo ingênuo que faz a predição com base no valor médio do target. **Quanto maior seu valor, melhor é nosso modelo com relação a esse modelo mais simplista**.

Exemplo	Predição	Target
1	500	600
2	1000	2
3	750	690

$$R^2 = 1 - \frac{\frac{1}{n}\sum_i |y_i - \hat{y}_i|^2}{\frac{1}{n}\sum_i |y_i - \bar{y}_{train}|^2},$$

R^2 = 1 - ((⅓*(600-500)²+(2-1.000)²+(690-750)²)/
(⅓*(600-430,66)²+(2-430,66)²+(690-430,66)²))

R^2 = -2.60

Agora que já sabemos avaliar nosso modelo, vamos voltar para a prática!

Quais as vantagens da Regressão Linear?

- Simples de entender e de ser implementado
- Ideal para problemas em que sabemos que as variáveis e o target possuem uma relação linear

E quais as desvantagens do Regressão Linear?

- Muitos problemas do “mundo real” não possuem uma clara relação linear entre as variáveis e o target
- Pode sofrer com outliers

Para praticar...

- Participar da competição do Kaggle da qual pegamos uma amostra do conjunto de dados para nossa aula e tentar obter um modelo melhor (melhorando também a Feature Engineering)!

- Participar de outras competições no Kaggle sobre problemas de regressão (como esse para predizer valores de imóveis). Praticar tanto a utilização de modelos de regressão quanto fazer a Feature Engineering 😉

Nossos contatos


mulheres.em.ia@gmail.com


mulheres-em-ia


@mulheres.em.ia


@mulheres.em.ia


@MulheresemInteligenciaArtificial

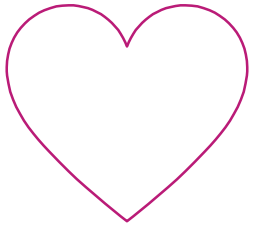

Canal: Mulheres em IA

Linktree

https://linktr.ee/mulheres.em.ia

Grupo Telegram para Mulheres

https://t.me/mulheres_em_ia



Muito obrigada!

Dúvidas? Podem nos procurar! 😉