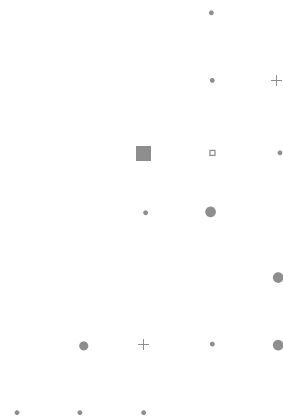


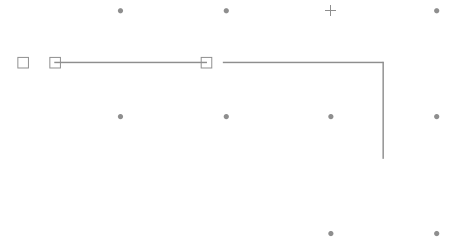
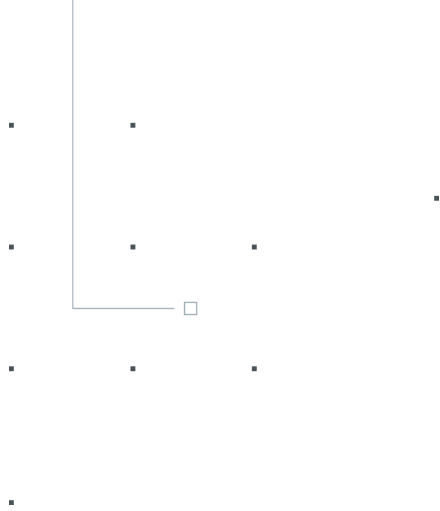


FIAP

Versão original gentilmente cedida pelo professor
FELIPE GUSTAVO SILVA TEODORO



FEATURE ENGINEERING





THIAGO NASCIMENTO NOGUEIRA

PROFESSOR


- Bacharel, mestre e doutor em física pela USP
- Trabalha com ambiente Linux desde 1995
- Atuação em sistemas distribuídos desde 2007
- Big Data desde 2016
- Ciência de Dados desde 2017
- Computação Quântica desde 2021
- IA Generativa desde 2020

 /thnogueira profthiago.nogueira@fiap.com.br



OBJETIVOS DA DISCIPLINA

Trabalhar com conceitos relacionados ao processo de extração, transformação e seleção de características para problemas de Machine e Deep Learning.



PROGRAMA TRADICIONAL VS IA

Programa tradicional



Machine Learning



FEATURE ENGINEERING

A arte de transformar dados brutos em características significativas

- Feature engineering é a prática de criar características úteis a partir de dados brutos para melhorar o desempenho dos modelos de machine learning.
- Essas características são representações específicas que ajudam os algoritmos a entender padrões nos dados.
- A qualidade das características é crítica para o sucesso dos modelos.
- Envolve transformar dados em formatos mais informativos e relevantes para a modelagem.
- É uma combinação de habilidades técnicas, conhecimento de domínio e criatividade.
- Aspecto mais criativo de Ciência de Dados!!!

IMPORTÂNCIA DA FEATURE ENGINEERING

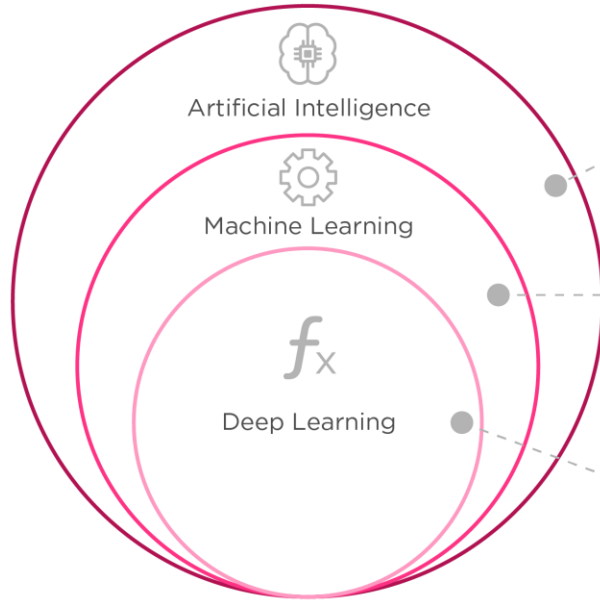
POR QUE É CRUCIAL?

- A qualidade das características é crítica para o sucesso dos modelos de ML.
- Boas características levam a previsões mais precisas.
- Características bem projetadas simplificam a interpretação dos dados.
- A engenharia de características extrai informações úteis dos dados brutos.
- Essencial para resolver problemas do mundo real em várias áreas.

Coming up with features is difficult, time-consuming, requires expert knowledge. "Applied machine learning" is basically feature engineering.

— Andrew Ng, Machine Learning and AI via Brain simulations

MACHINE E DEEP LEARNING



ARTIFICIAL INTELLIGENCE

A technique which enables machines to mimic human behaviour

MACHINE LEARNING

Subset of AI technique which use statistical methods to enable machines to improve with experience

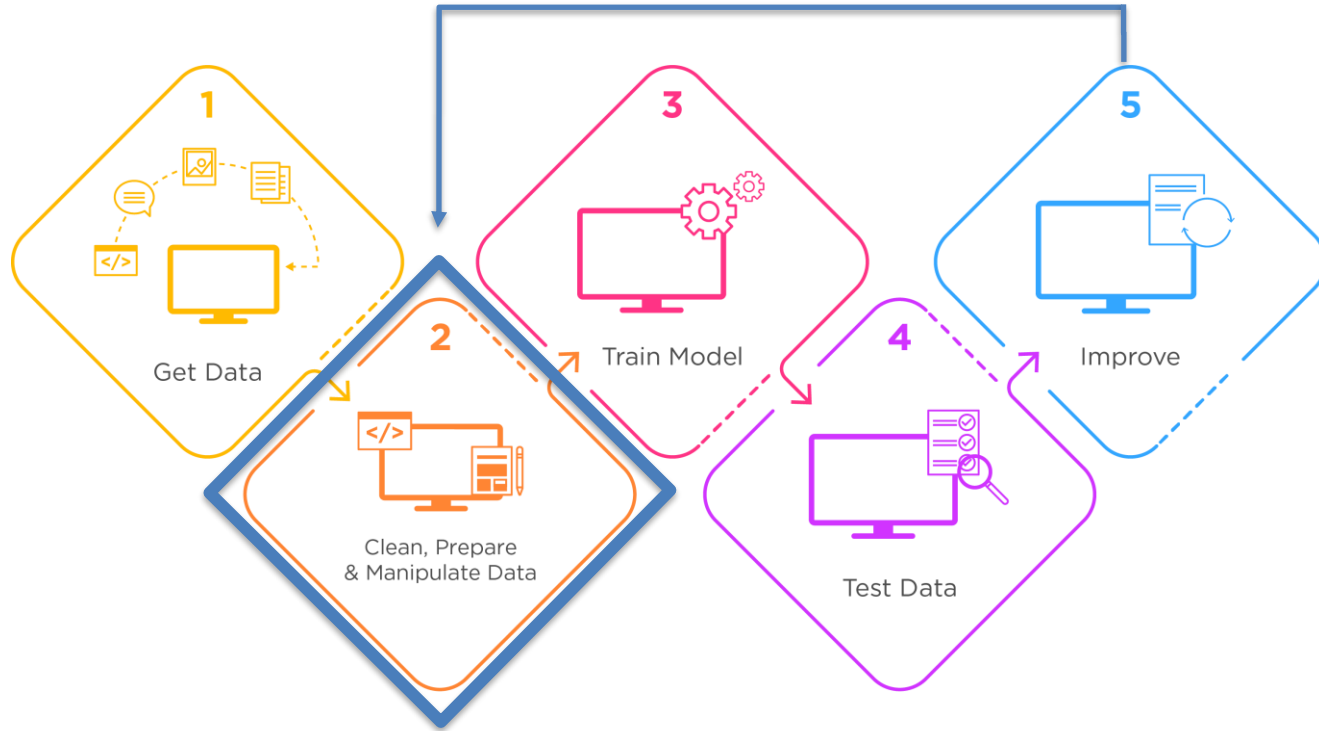
DEEP LEARNING

Subset of ML which make the computation of multi-layer neural network feasible

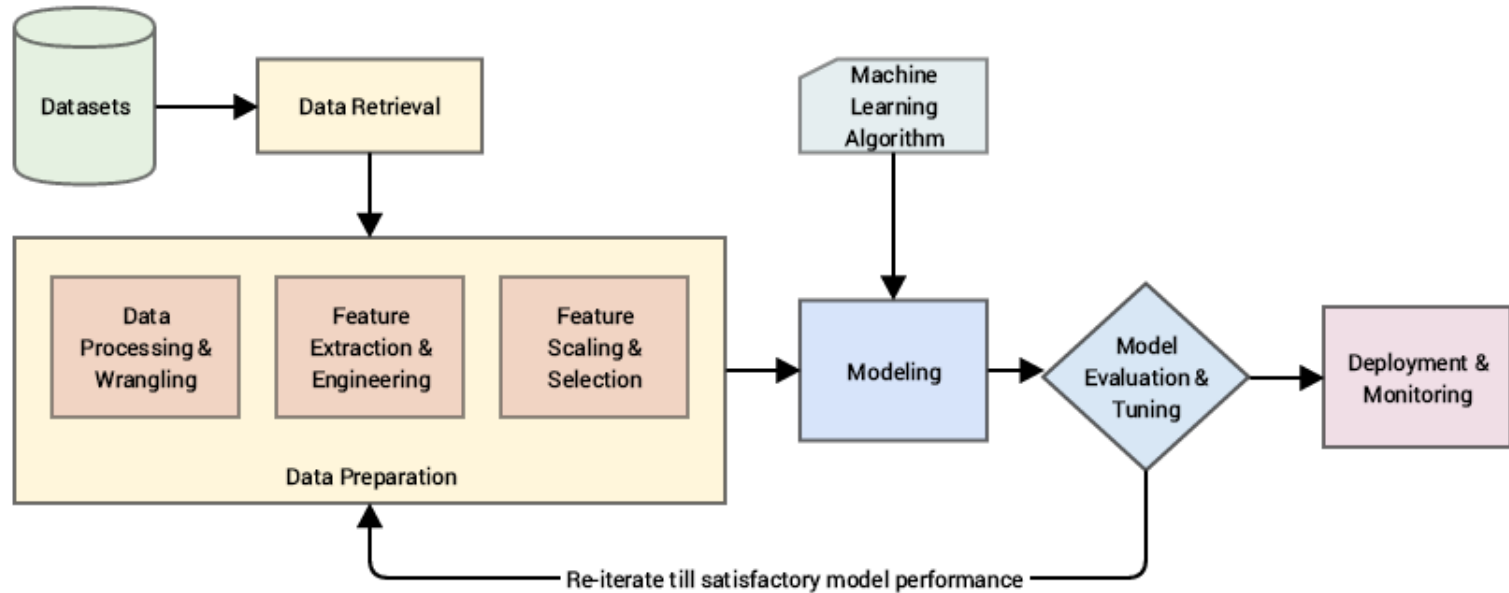
DIA A DIA DE UM CIENTISTA DE DADOS



PIPELINE DE MODELOS DE ML e DL

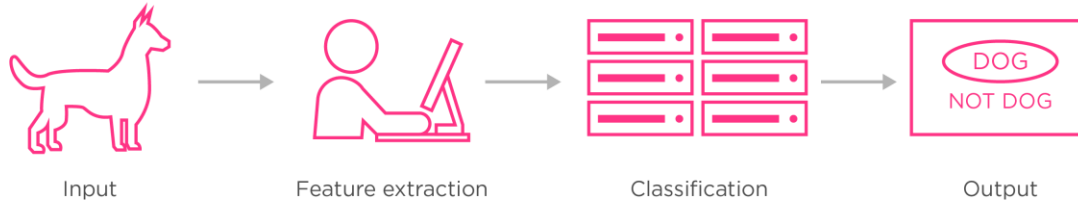


PIPELINE DE MODELOS DE ML e DL



MACHINE & DEEP LEARNING

TRADITIONAL MACHINE LEARNING



DEEP LEARNING



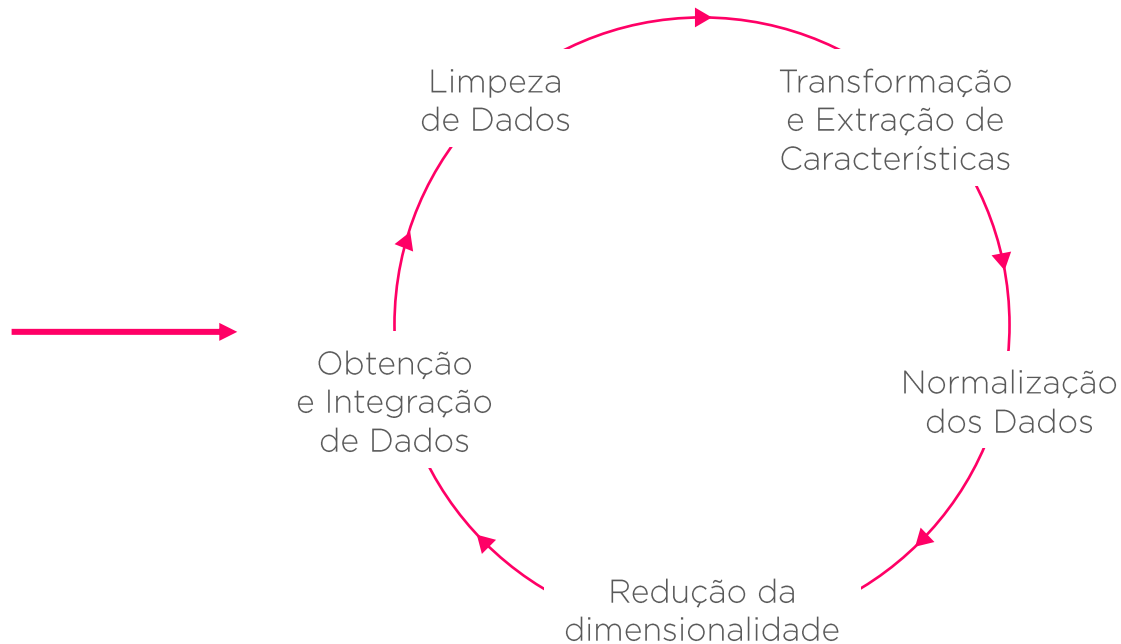
NOSSA JORNADA

Manipulação e
Pré-
Processamento

Extração e
Transformação
de
Características

Seleção de
Características

PRÉ-PROCESSAMENTO DOS DADOS

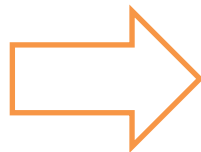


ATIVIDADE

FILTRO INICIAL DE FEATURES

FEATURES

- Nome
- Idade
- Endereço
- Renda R\$
- Renda USD
- Dívidas em outros bancos
- Serasa
- CPF
- Gênero



SEGMENTOS

- Varejo
- Uniclass
- Personalité
- Private

CASE: ESTUDO DE DOENÇAS CRÔNICAS

OBJETIVO

Entendimento da evolução de doenças crônicas a partir do histórico de internações dos pacientes

FONTE PRINCIPAL DADOS

Base de internações do DATASUS

CASE: ESTUDO DE DOENÇAS CRÔNICAS

FEATURES

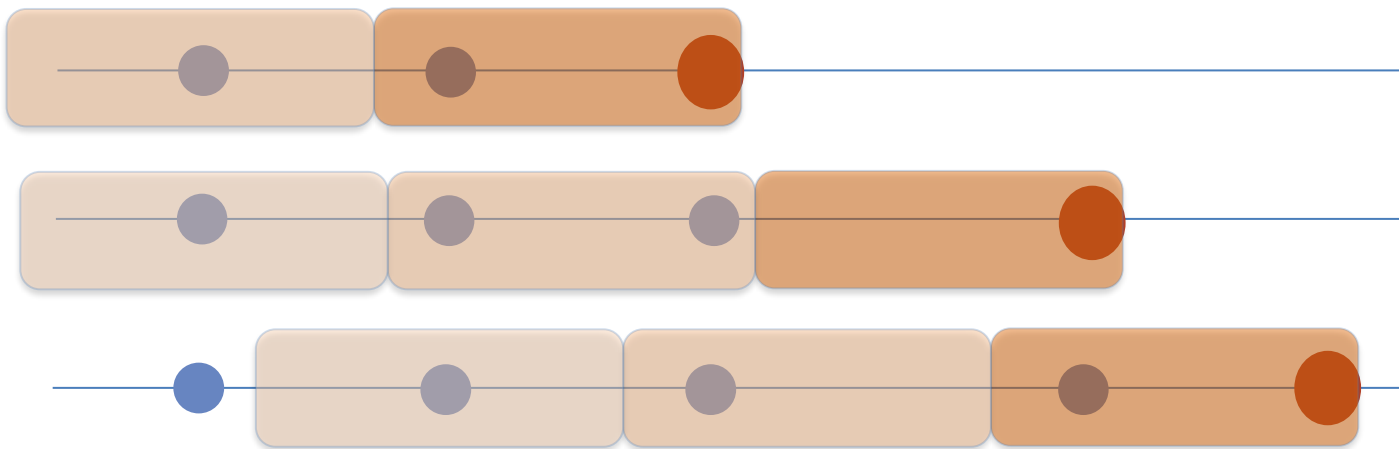
- | | | | | | |
|--------------|--------------|--------------|--------------|--------------|-------------|
| • UF_ZI | • UTI_INT_IN | • VAL_TRANSP | • COD_IDADE | • INSC_PN | • FAEC_TP |
| • ANO_CMPT | • UTI_INT_AN | • VAL_OBSANG | • IDADE | • SEQ_AIH5 | • REGCT |
| • MES_CMPT | • UTI_INT_AL | • VAL_PED1AC | • DIAS_PERM | • CBOR | • RACA_COR |
| • ESPEC | • UTI_INT_TO | • VAL_TOT | • MORTE | • CNAER | • ETNIA |
| • CGC_HOSP | • DIAR_ACOM | • VAL_UTI | • NACIONAL | • VINCPREV | • SEQUENCIA |
| • N_AIH | • QT_DIARIAS | • US_TOT | • NUM_PROC | • GESTOR_COD | • REMESSA |
| • IDENT | • PROC_SOLIC | • DT_INTER | • CAR_INT | • GESTOR_TP | |
| • CEP | • PROC_REA | • DT_SAIDA | • TOT_PT_SP | • GESTOR_CPF | |
| • MUNIC_RES | • VAL_SH | • DIAG_PRINC | • CPF_AUT | • GESTOR_DT | |
| • NASC | • VAL_SP | • DIAG_SECUN | • HOMONIMO | • CNES | |
| • SEXO | • VAL_SADT | • COBRANCA | • NUM_FILHOS | • CNPJ_MANT | |
| • UTI_MES_IN | • VAL_RN | • NATUREZA | • INSTRU | • INFEHOSP | |
| • UTI_MES_AN | • VAL_ACOMP | • GESTAO | • CID_NOTIF | • CID ASSO | |
| • UTI_MES_AL | • VAL_ORTP | • RUBRICA | • CONTRACEP1 | • CID_MORTE | |
| • UTI_MES_TO | • VAL_SANGUE | • IND_VDRL | • CONTRACEP2 | • COMPLEX | |
| • MARCA_UTI | • VAL_SADTSR | • MUNIC_MOV | • GESTRISCO | • FINANC | |

CASE: ESTUDO DE DOENÇAS CRÔNICAS

IDENTIFICAÇÃO DO PACIENTE

- Definição das features para construção da chave

CONSTRUÇÃO DE HISTÓRICO DO PACIENTE



[illegible]

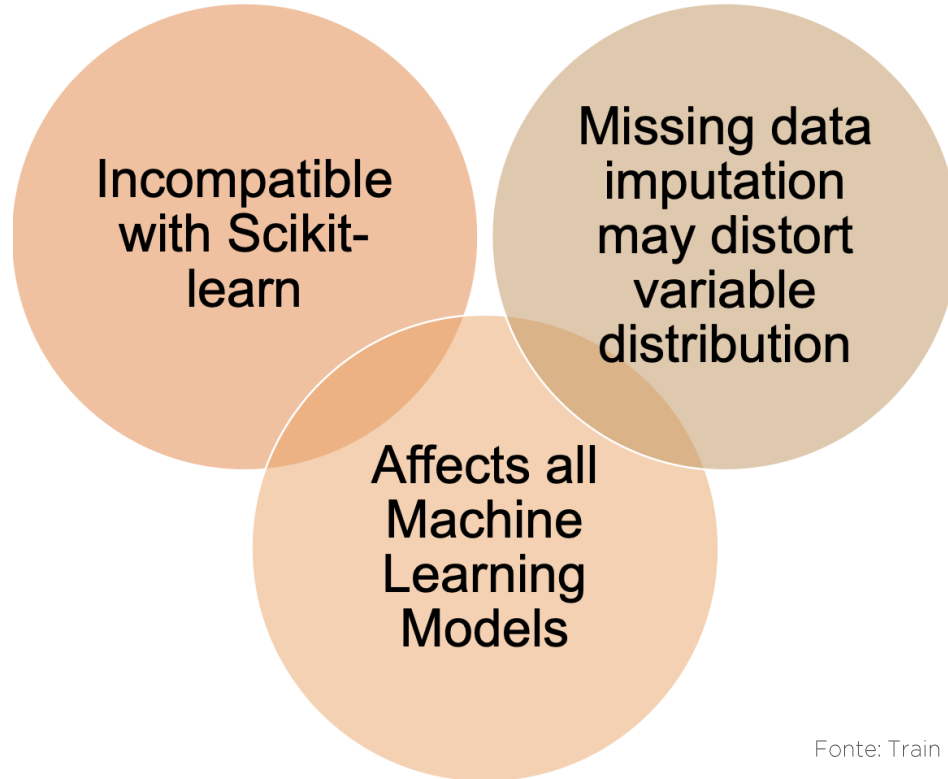
CARACTERISTICAS COMUNS DAS **FEATURES**



MISSING DATA

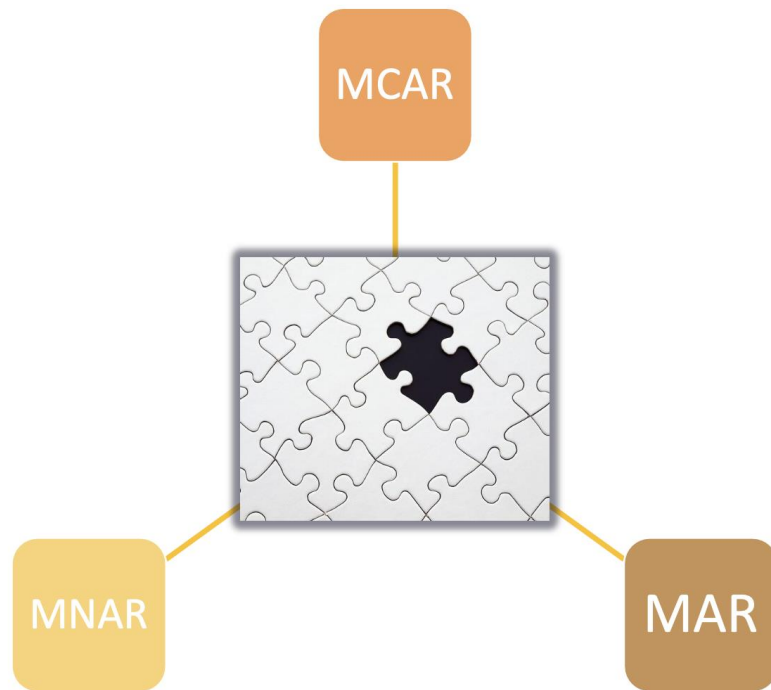
- Dados ausentes, ou valores ausentes, ocorrem quando nenhum dado é armazenado para uma determinada observação em uma variável.
- Dados ausentes são uma ocorrência comum na maioria dos conjuntos de dados
- A falta de dados pode ter um efeito significativo nas conclusões que podem ser tiradas dos dados.

MISSING DATA: IMPACTOS



Fonte: Train in Data

MISSING DATA: TIPOS DE DADOS FALTANTES



MISSING **DATA**: Missing completely at random (MCAR)

- Os dados são ditos do tipo MCAR quando “a probabilidade de estarem ausentes é independente de qualquer observação no dataset”.
- Ou seja, imagine um dataset com 10.000 respostas de entrevistados. Encontramos 100 pessoas que não responderam ao item 'Você prefere o inverno ou verão?' por exemplo. Como não conseguimos identificar alguma relação que explique essa ausência, podemos descartar essas 100 pessoas e realizar a modelagem do problema apenas com as 9900 respostas completas.

MISSING DATA: Missing completely at random (MCAR)

Complete data	
Age	IQ score
25	133
26	121
29	91
30	105
30	110
31	98
44	118
46	93
48	141
51	104
51	116
54	97

Incomplete data	
Age	IQ score
25	
26	121
29	91
30	
30	110
31	
44	118
46	93
48	
51	
51	116
54	

No exemplo ao lado, conseguimos visualizar que não há relação entre a idade e os dados faltantes na tabela da direita. Temos amostras com idade 30 que possuem dados preenchidos e outra não. O mesmo para 51 anos.

Ou seja, não é possível identificar uma relação entre as amostras sem preenchimento e aquelas completamente preenchidas.

MISSING DATA: Missing at random (MAR)

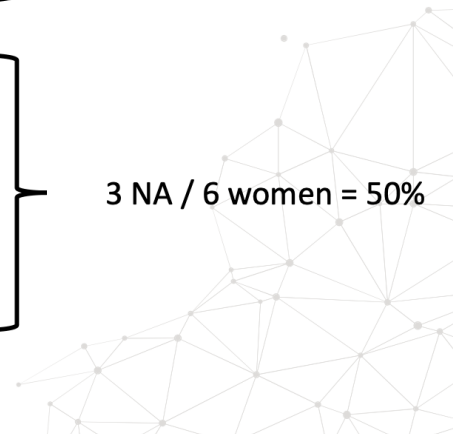
A probabilidade de uma observação estar faltando depende da informação disponível

A abordagem de dados do tipo MAR assumem que as observações com dados faltantes não respeitam uma distribuição aleatória como as amostras com dados observados. Isso significa que precisamos modelar o comportamento das amostras com dados faltantes.

Gender	Weight
Male	60 kg
Male	NA
Male	NA
Male	77 kg
Male	80 kg
Male	62 kg
Female	NA
Female	NA
Female	60 kg
Female	55 kg
Female	NA
Female	58 kg

2 NA / 6 men = 33%

3 NA / 6 women = 50%



MISSING **DATA**: Missing at Not random (MNAR)

Existe um mecanismo ou uma razão pela qual os valores ausentes são introduzidos no conjunto de dados.

Target = depression	No of clinic visits	No sports classes weekly
Yes	1	NA
Yes	NA	NA
Yes	NA	0
Yes	4	2
Yes	NA	1
Yes	3	NA
No	0	0
No	NA	5
No	1	2
No	1	1
No	2	1
No	NA	2

More NA overall for depressed patients

Less NA for non-depressed patients

MISSING DATA: Mas como lidar com dados faltantes?

Variáveis Numéricas



☐ Média e Mediana

☐ Valor Arbitrário

☐ End of Tail

Variáveis Categóricas



☐ Categoria mais frequente

☐ Categoria Faltante

Ambas



☐ Complete Case Analysis

☐ Amostra Aleatória

☐ Similaridade

MISSING DATA: Preenchimento por média e mediana?

Price
100
90
50
40
20
100
60
120
200

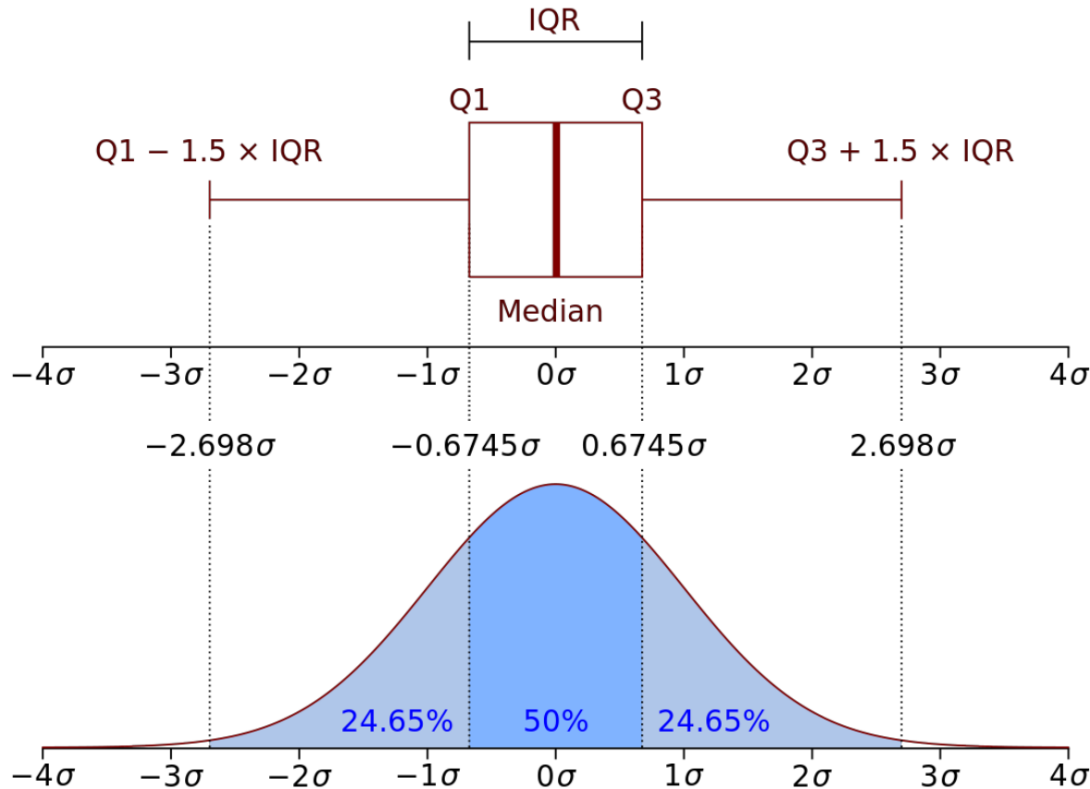
Mean = 86.66

Median = 90



Price
100
90
50
40
20
100
86.66
60
120
86.66
200

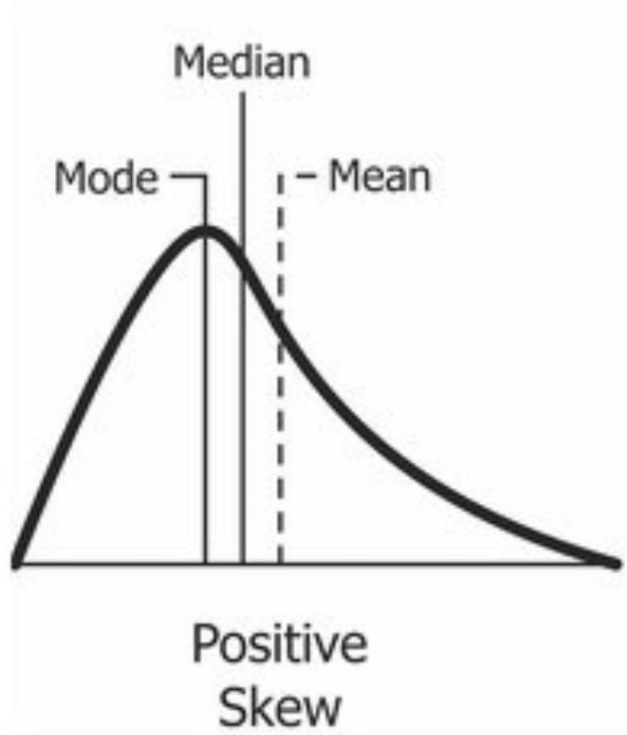
MISSING **DATA**: Preenchimento por média e mediana?



Média ou Mediana?

Se a variável tiver distribuição normal, a média e a mediana são aproximadamente as mesmas

MISSING DATA: Preenchimento por média e mediana?



Média ou Mediana?

Se a variável for assimétrica, a mediana é uma representação melhor.

MISSING DATA: Preenchimento por média e mediana

Faltam dados completamente aleatórios

Não mais de 5% da variável contém dados ausentes

Fácil de implementar

Maneira rápida de obter conjuntos de dados completos

Pode ser integrado na produção (durante a implantação do modelo)

Distorção da distribuição da variável original

Distorção da covariância com as variáveis restantes do conjunto de dados

Quanto maior a porcentagem de NA, maiores as distorções

MISSING DATA: Arbitrary value imputation

Price
100
90
50
40
20
100
60
120
200

Arbitrary = 999



Price
100
90
50
40
20
100
999
60
120
999
200

MISSING DATA: Arbitrary value imputation

Price
100
90
50
40
20
100
60
120
200

~~Arbitrary = 99~~



Price
100
90
50
40
20
100
999
60
120
999
200

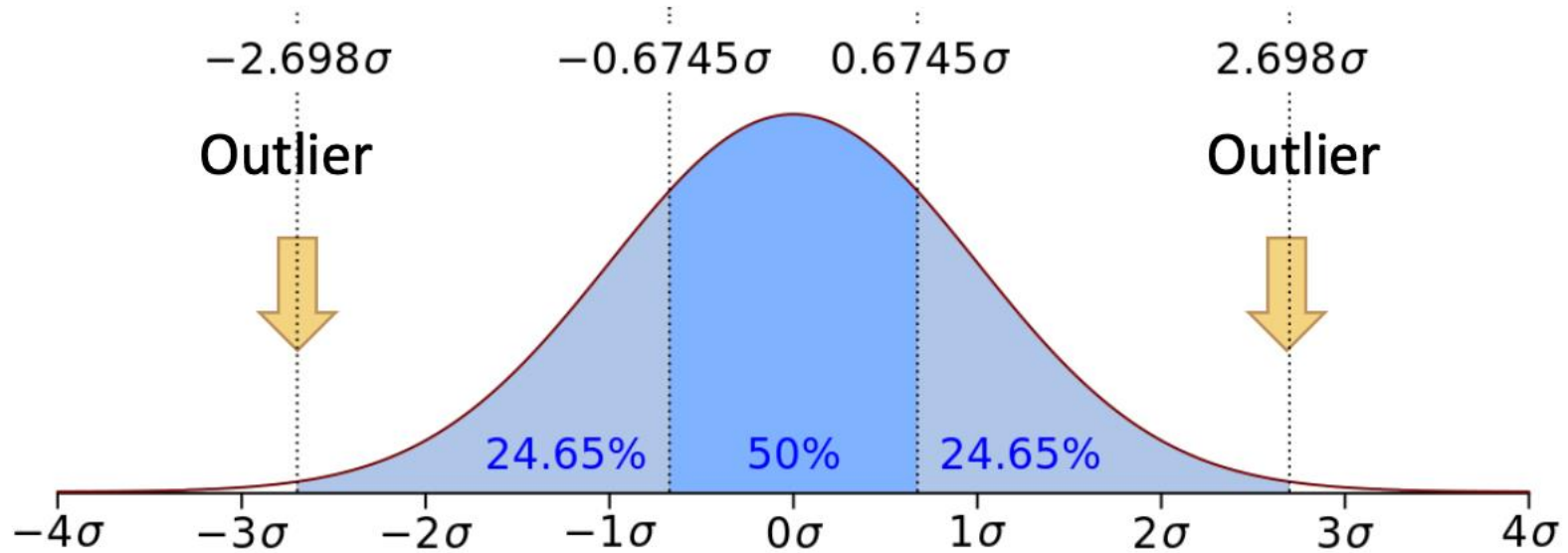
MISSING DATA: End of Tail

A imputação de final de cauda é equivalente à imputação de valor arbitrário, mas selecionando automaticamente valores arbitrários no final das distribuições de variáveis.

Se a variável for normalmente distribuída, podemos usar a média mais ou menos 3 vezes o desvio padrão.

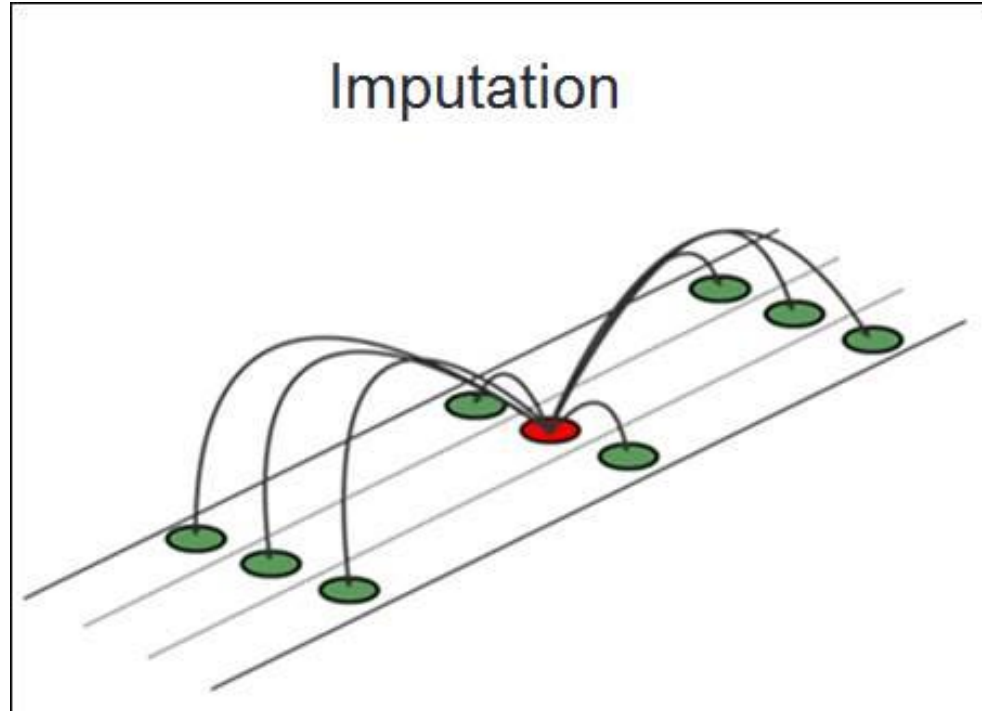
Se a variável for assimétrica, podemos usar a regra de proximidade IQR.

MISSING DATA: End of Tail

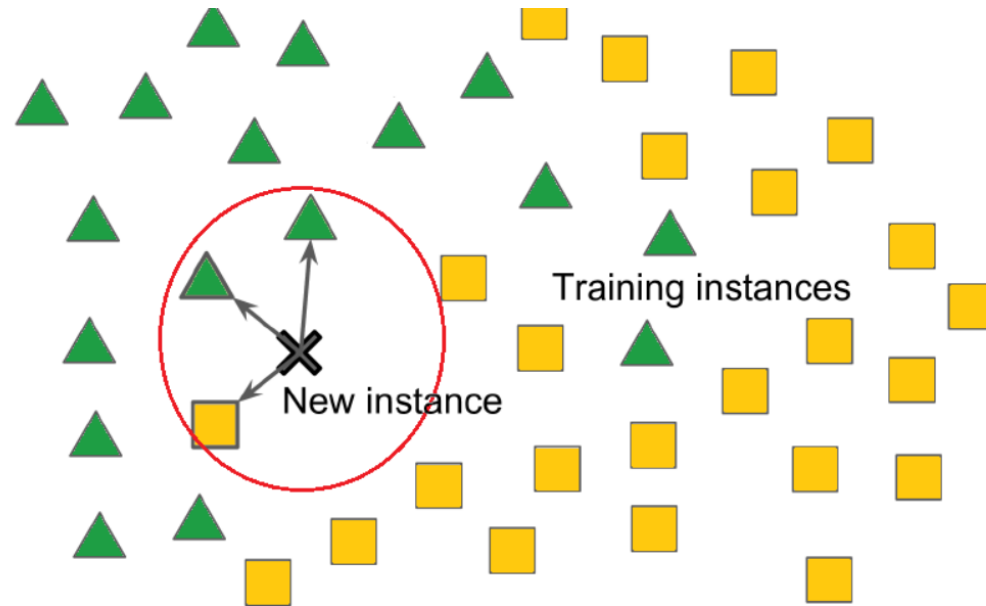


Method	Definition	Pros	Cons
Listwise Deletion	excluding all cases (listwise) that have missing values	preserve distribution if MCAR	1. may discard too much data and hurt the model 2. may yield biased estimates if not MCAR (as we keep a special subsample from the population)
Mean/Median/Mode Imputation	replacing the NA by mean/median/most frequent values (for categorical feature) of that variable	good practice if MCAR	1. distort distribution 2. distort relationship with other variables
End of distribution Imputation	replacing the NA by values that are at the far end of the distribution of that variable, calculated by mean + 3*std	Captures the importance of missingness if there is one	1. distort distribution 2. may be considered outlier if NA is few or mask true outlier if NA is many. 3. if missingness is not important this may mask the predictive power of the original variable
Random Imputation	replacing the NA by taking a random value from the pool of available observations of that variable	preserve distribution if MCAR	not recommended in business settings for its randomness (different result for same input)
Arbitrary Value Imputation	replacing the NA by arbitrary values	Captures the importance of missingness if there is one	1. distort distribution 2. typical used value: -9999/9999. But be aware it may be regarded as outliers.
Add a variable to denote NA	creating an additional variable indicating whether the data was missing for that observation	Captures the importance of missingness if there is one	expand feature space

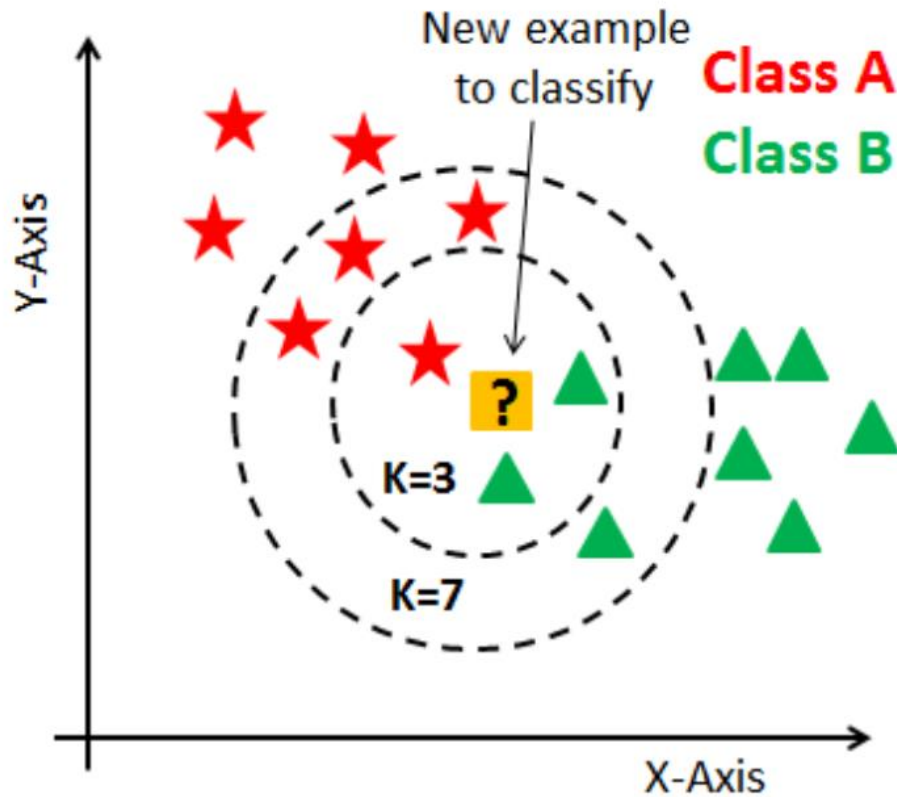
MISSING DATA: Similaridade – KNN Imputer



K NEAREST NEIGHBOR



K NEAREST NEIGHBOR



PRÉ-PROCESSAMENTO DE DADOS

Normalização MinMax

Transformar um conjunto de dados que estão em diferentes grandezas e escalas em um conjunto de dados padronizados.

Normalization Formula

$$X \text{ normalized} = \frac{(X - X \text{ minimum})}{(X \text{ maximum} - X \text{ minimum})}$$

PRÉ-PROCESSAMENTO DE DADOS

Padronização

Centraliza a variável em zero e define a variância como 1.

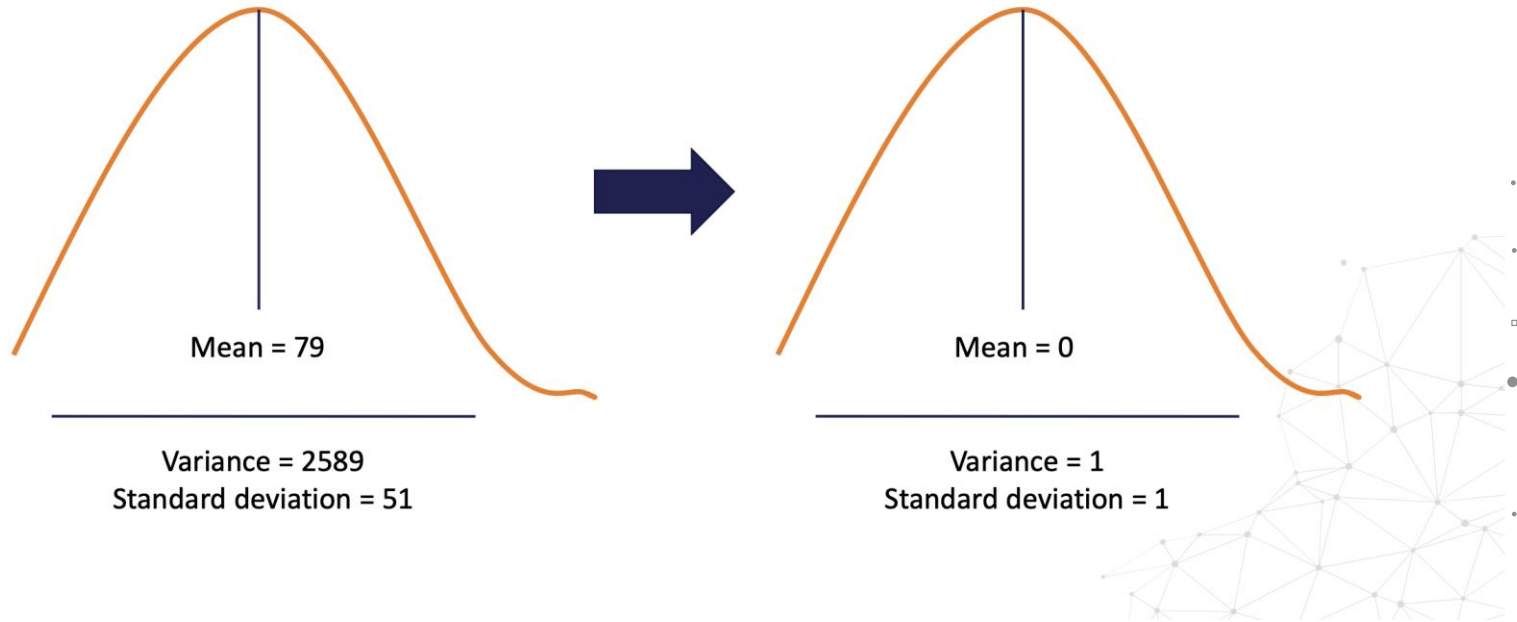
Standardisation Formula

$$X_{\text{standard}} = \frac{X - \text{mean}(X)}{\text{std}(X)}$$

PRÉ-PROCESSAMENTO DE DADOS

Padronização

Centraliza a variável em zero e define a variância como 1.



PRÉ-PROCESSAMENTO DE DADOS

Robust Scaling

Transformar um conjunto de dados que estão em diferentes grandezas e escalas em um conjunto de dados padronizados.

Formula

$$X \text{ normalized} = \frac{X - \text{median}(X)}{75^{\text{th}} \text{ quant } X - 25^{\text{th}} \text{ quant}(X)}$$

PRÉ-PROCESSAMENTO DE DADOS

Robust Scaling

Price
100
90
50
40
20
100
50
60
120
40
200

Median = 60
25th quantile = 45
75th quantile = 100
IQR = 100 - 45 = 55

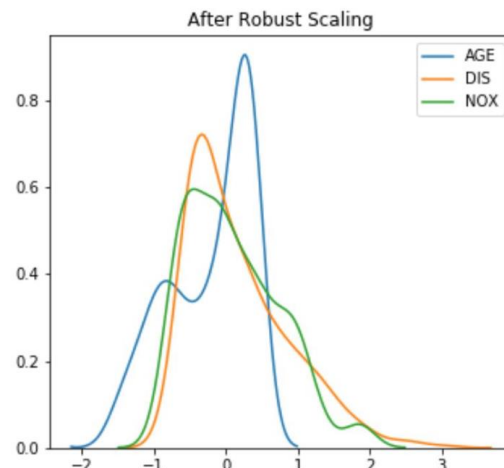
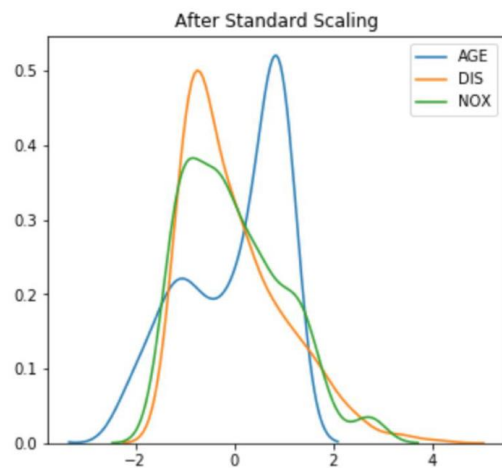
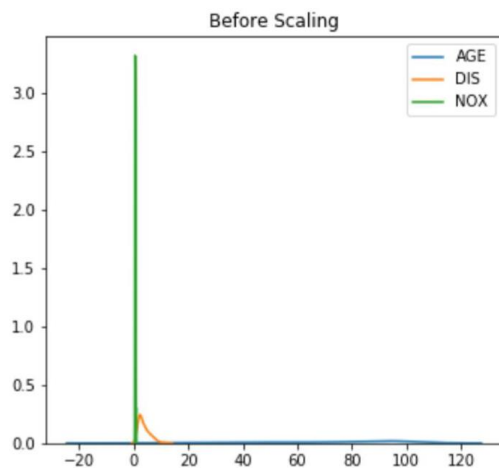


$$\frac{\text{Obs.} - \text{Median}}{\text{IQR}}$$

Price
0.73
0.55
-0.18
-0.36
-0.73
0.73
-0.18
0.00
1.09
-0.36
2.55

PRÉ-PROCESSAMENTO DE DADOS

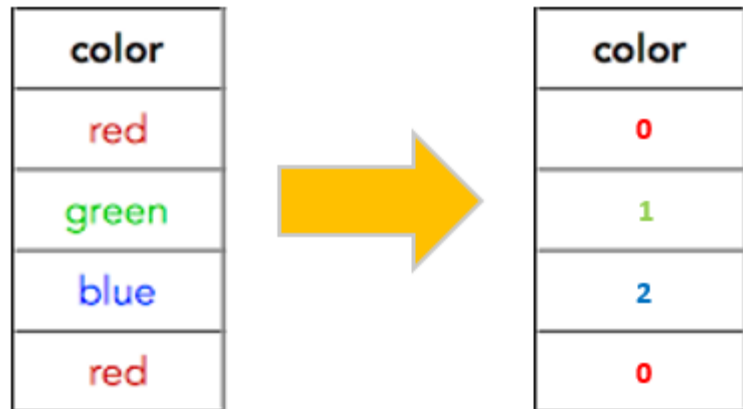
Robust Scaling vs Standard Scaler



<div> <div>••</div> <div>••</div> <div>•+</div> <div>+•</div> <div> </div> <div>+</div> </div>	1	Original	Standardization	Max-Min Scaler	Rubost Scaler
	2	6.9314183	-0.2244971	0.0000003	0.8283487
	3	2.6674115	-0.2244979	0.0000001	0.0690181
	4	7.7248183	-0.2244970	0.0000003	0.9696367
	5	5.7388433	-0.2244973	0.0000002	0.6159760
	6	0.8965615	-0.2244982	0.0000000	-0.2463333
	7	4.5147618	-0.2244975	0.0000002	0.3979926
	8	2.9934144	-0.2244978	0.0000001	0.1270724
	9	4.8708377	-0.2244975	0.0000002	0.4614023
	10	4.2797819	-0.2244976	0.0000002	0.3561476
	11	1.0085616	-0.2244982	0.0000000	-0.2263885
	12	5.5166580	-0.2244974	0.0000002	0.5764094
	13	1.1171326	-0.2244981	0.0000000	-0.2070542
	14	0.4069897	-0.2244983	0.0000000	-0.3335159
	15	5.0536949	-0.2244975	0.0000002	0.4939654
	16	8.4068370	-0.2244969	0.0000003	1.0910900
	17	8.9588050	-0.2244968	0.0000003	1.1893840
	18	0.9543401	-0.2244982	0.0000000	-0.2360442
	19	94750.5292279	-0.2079018	0.0037104	16872.6857158
	20	2051.2433203	-0.2241390	0.0000803	364.8776314
	21	25536631.9371928	4.2485000	1.0000000	4547540.7645023

CATEGORICAL ENCODING

Label Encoding



CATEGORICAL ENCODING

One-Hot Encoding

color		color_red	color_blue	color_green
red		1	0	0
green		0	0	1
blue		0	1	0
red		1	0	0

CATEGORICAL ENCODING

Count/Frequency Encoding

Count

Height	Height
Short	2
Tall	1
Short	1
Medium	1

Frequency

Height	Height
Short	0.4
Tall	0.2
Short	0.2
Medium	0.2

CATEGORICAL ENCODING

Rare Labels Encoding

CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

CITY	CITY_COUNT
A	56
B	84
C	54
D	2
E	12
F	60
G	3
H	5
K	25
L	1
M	36
Z	45

CITY	CITY_COUNT
A	56
B	84
C	54
F	60
K	25
M	36
Z	45
RARE (D,E,G,H,L)	23

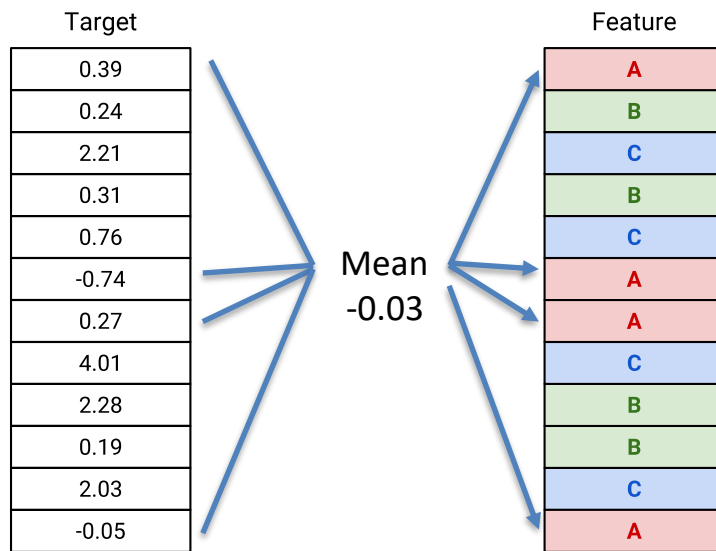
CATEGORICAL ENCODING

Target Encoding – Classificação

color	target		Encoding	target
red	1		1.00	1
green	0		0.5	0
blue	0		0.5	0
red	1		1.00	1

CATEGORICAL ENCODING

Target Encoding - Regressão



CATEGORICAL ENCODING

Target Encoding

- Em uma tarefa de **classificação binária**, a nova representação numérica corresponde a probabilidade do alvo ser de uma classe dado a categoria assumida pela variável categórica (probabilidade de $alvo_y=1$ dado a $cor_x='branco'$ por exemplo).
- Em uma tarefa de **regressão**, a nova representação numérica corresponde ao valor esperado para o alvo y dado a categoria assumida pela variável categórica ($preço_carro_y=12.7$ dado que $cor_x='branco'$ por exemplo).

CATEGORICAL ENCODING

Target Encoding – Target Leakage

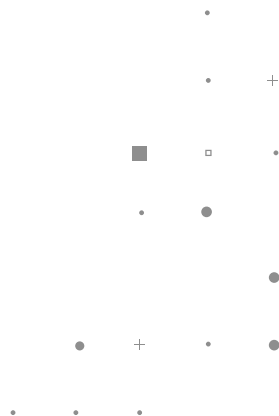




OBRIGADO

FIAP

Copyright © 2020 | Professor Felipe Gustavo Silva Teodoro
Todos os direitos reservados. A reprodução ou divulgação total ou parcial deste documento é expressamente proibida sem consentimento formal, por escrito, do professor(a)/autor(a).





FIAP

