EPI5717: Machine learning para predições em saúde

Aula 3

Prof. Dr. Alexandre Chiavegatto Filho

PREDIÇÃO COM MACHINE LEARNING

- Desenvolver algoritmos que façam boas predições em saúde.
- Principais razões pelas quais algoritmos às vezes não apresentam boa performance preditiva:
  - Extrapolação inadequada dos resultados.
  - Pré-processamento inadequado dos dados.
  - Sobreajuste (mais importante).
  - Validação inadequada da qualidade dos algoritmos.

Extrapolação inadequada

- Desenvolver os algoritmos para uma população e esperar que funcionam corretamente para outra diferente.
  - Importar algoritmos dos EUA/Europa: nossas características genéticas e socioeconômicas são muito diferentes.
  - Extrapolação para períodos diferentes (cuidado com doenças sazonais).

PRÉ-PROCESSAMENTO DOS DADOS

- Técnicas de pré-processamento de dados
    - Seleção das variáveis.
    - Vazamento de dados.
    - Padronização.
    - Redução de dimensão.
    - Colinearidade.
    - Valores missing.
    - One-hot encoding.

PRÉ-PROCESSAMENTO DOS DADOS

Preditores plausíveis:

- Pré-selecionar variáveis que sejam preditoras plausíveis (bom senso do pesquisador).

- Coincidências acontecem em análises de big data e pode ser que o algoritmo dê muita importância para associações espúrias.

PRÉ-PROCESSAMENTO DOS DADOS

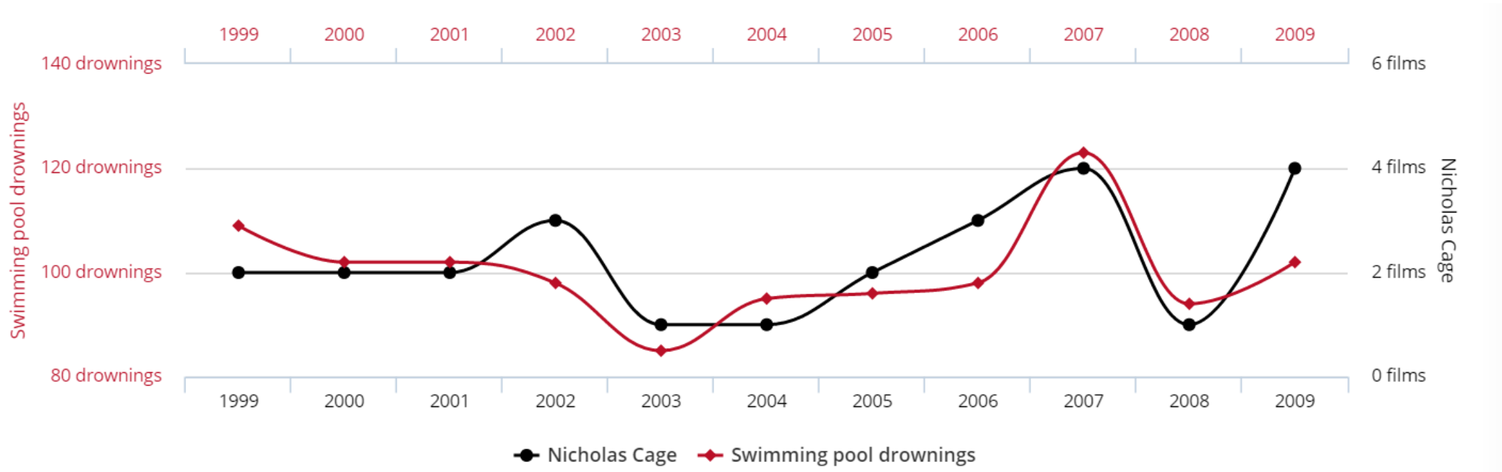
tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Filmas Nicolas Cage appeared in

Correlation: 66,6% (r=0,666004)



PRÉ-PROCESSAMENTO DOS DADOS

Cuidado com vazamento de informação (“data leakage”).

- Acontece quando os dados de treino apresentam informação escondida que faz com que o modelo aprenda padrões que não são do seu interesse.

- Uma variável preditora tem escondida o resultado certo:
    - Não é a variável que está predizendo o desfecho, mas o desfecho que está predizendo ela.

JOURNAL OF MEDICAL INTERNET RESEARCH Chiavegatto Filho et al

Letter to the Editor

Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on “Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning”

Alexandre Chiavegatto Filho, PhD; André Filipe De Moraes Batista, MSc, PhD; Hellen Geremias dos Santos, MPH, PhD

Department of Epidemiology, School of Public Health, University of São Paulo, São Paulo, Brazil

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo

Incluir o número identificador do paciente como variável preditora

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo

Incluir o número identificador do paciente como variável preditora

Problema

Se pacientes de hospital especializado em câncer tiverem números semelhantes.
Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes.
Esse algoritmo aprendeu algo interessante para o sistema de saúde?

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo	Problema	Motivo
Incluir o número identificador do paciente como variável preditora	Se pacientes de hospital especializado em câncer tiverem números semelhantes. Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes. Esse algoritmo aprendeu algo interessante para o sistema de saúde?	Motivo pelo qual os dados e os algoritmos de machine learning precisam ser abertos. Sempre analisar importância preditora das variáveis (Shapley).

PADRONIZAÇÃO

- A escala das variáveis pode afetar muito a qualidade das predições.

- Alguns algoritmos darão preferência para utilizar variáveis com valores muito alto.

PADRONIZAÇÃO

- Padronizar as variáveis contínuas para todas terem média de 0 e desvio-padrão de 1.

$$z_i = \frac{x_i - \mu}{\sigma}$$

- Ou seja, é feita a subtração da média e a divisão pelo desvio padrão dos valores da variável.

REDUÇÃO DE DIMENSÃO

- Quanto maior a dimensão dos dados (número de variáveis) maior o risco de o algoritmo encontrar e utilizar associações espúrias.

REDUÇÃO DE DIMENSÃO

- Análise de Componentes Principais

Técnica de aprendizado não supervisionado.

O objetivo é encontrar combinações lineares das variáveis preditoras que incluam a maior quantidade possível da variância original.

O primeiro componente principal irá preservar a maior combinação linear possível dos dados, o segundo a maior combinação linear possível não correlacionada com o primeiro componente, etc.

VARIÁVEIS COLINEARES

Uma das razões pela qual a ACP é tão utilizada, é o fato de que cria componentes principais não correlacionados.

- Na prática, alguns algoritmos conseguem melhor performance preditiva com variáveis com baixa correlação.

VARIÁVEIS COLINEARES

Uma outra forma de diminuir a presença de variáveis com alta correlação é excluí-las.

- Variáveis colineares trazem informação redundante (tempo perdido).
- Além disso, aumentam a instabilidade dos modelos.
- Estabelecer um limite de correlação com alguma outra variável (0,75 a 0,90).

VARIÁVEIS MISSING

É importante entender por que valores de uma variável estão faltantes.

Motivo sistemático → INFORMAÇÃO PREDITIVA.

Grande diferença em relação a estudos de inferência, em que valores missing devem ser evitados.

Não conseguiu responder a uma pergunta sobre o seu passado → INFORMAÇÃO PREDITIVA.

Pode ajudar na predição de problemas cognitivos graves no futuro

Em variáveis categóricas adicionar uma categoria para missing.

Imputação com machine learning para valores contínuos (adicionar nova variável indicativa de missing).

ONE-HOT ENCODING

Alguns algoritmos têm dificuldade em entender variáveis que têm mais do que uma categoria.

Acham que é uma variável contínua (0, 1, 2, 3...) → porém não têm significado contínuo.

A solução é transformar todas as categorias em uma variável diferente de valores 0 e 1 (one-hot encoding).

Variável com n categorias → criadas n variáveis.

ONE-HOT ENCODING

Pode trazer problemas em alguns modelos, como na regressão linear.

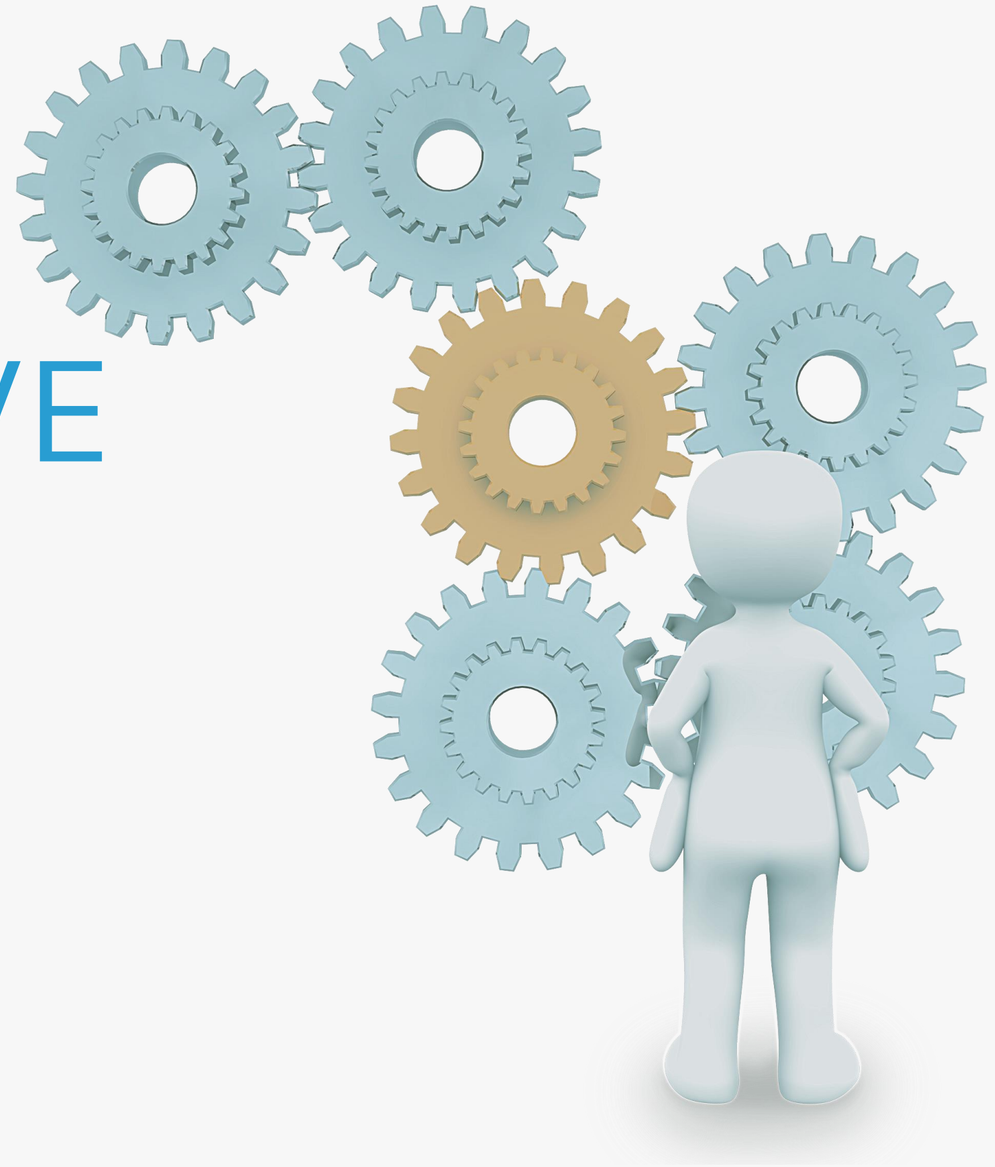
Solução: criar dummies.

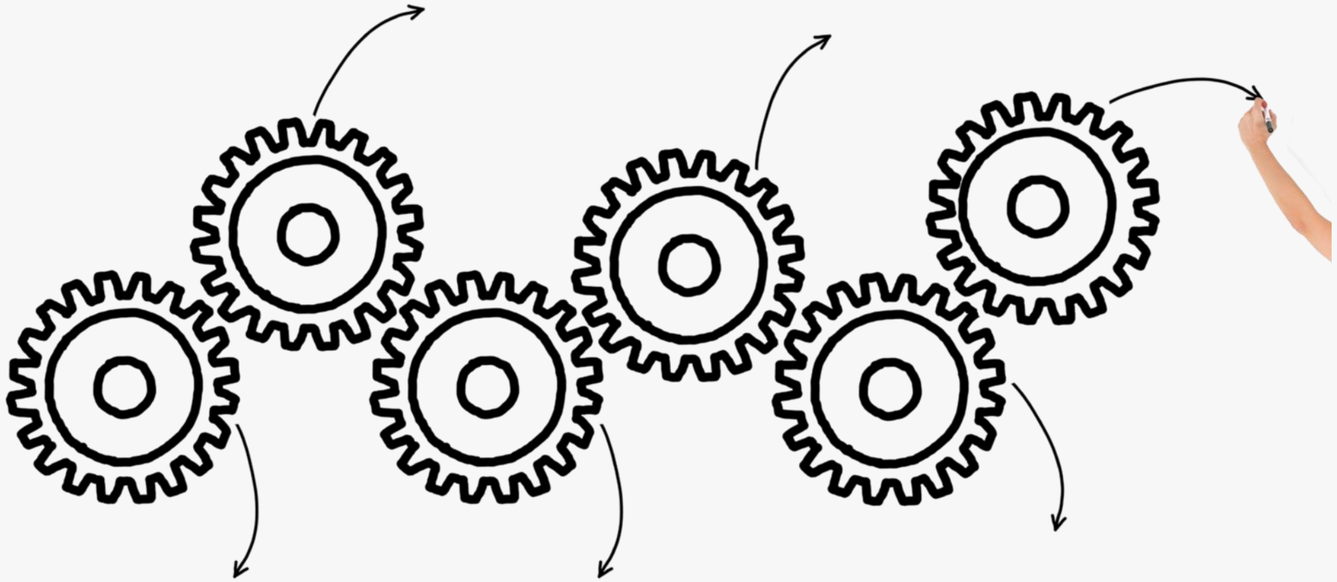
n-1 variáveis (deixar a mais frequente como categoria de referência).

neptuneblog

A COMPREHENSIVE GUIDE TO DATA PREPROCESSING

Author Samadrita Ghosh





O que é o pré-processamento de dados?

Método de análise, filtragem, transformação e codificação para que o algoritmo possa compreender os dados.

Parte considerável de um projeto: cerca de 80% do tempo

Por que o pré-processamento é necessário?

- Algoritmos são equações que aprendem com os dados
- "Se entra lixo, sai lixo"
- Necessidade de dados de alta qualidade
- Dados de mundo real sempre possuem ruídos e valores ausentes
- Solução: Pré-processamento (tratamento dos dados)

Ferramentas e bibliotecas

Para a execução do processo

PYTHON

Scikit Learn

6.3 **Preprocessing**

6.4 Impute

Automunge
(ferramenta em python)

Dados tabulares para ML

R

Framework muito utilizado por pesquisadores.

Diversos pacotes para pré-processamento.

WEKA

Software com suporte para mineração e pré-processamento embutidas no modelo de ML.

RAPIDMINER

Similar ao Weka.
Com várias ferramentas para pré-processamento.

Finalidade

Tendências e inconsistências

1

OBTENHA UMA VISÃO GERAL

2

IDENTIFIQUE DADOS AUSENTES

3

IDENTIFIQUE OUTLIERS E ANOMALIAS

4

REMOVA INCONSISTÊNCIAS

Finalidade

1

OBTENHA UMA VISÃO GERAL

- Entenda o formato dos dados
- Entenda a estrutura em que eles estão armazenados
- Média
- Mediana
- Quantis
- Desvio-padrão

Finalidade

IDENTIFIQUE DADOS AUSENTES

- Problema comum
- Pode interromper padrões
- Pode levar a perda de outros dados (linhas e colunas)
- Alguns algoritmos não aceitam dados ausentes

Finalidade

3

IDENTIFIQUE OUTLIERS E ANOMALIAS

- São pontos fora do padrão
- Podem precisar ser descartados
- Com o descarte há uma maior precisão
- Só devem ser mantidos para detecção da anomalia

Finalidade

4

REMOVA INCONSISTÊNCIAS

- Problema comum
- Ex.: colunas e linhas preenchidas incorretamente
- Ex.: dados duplicados
- Podem ser tratadas por meio de automação
- Geralmente precisam de um check-up manual

Manipulação de valores ausentes

01 Excluir observações

Apenas útil quando conta-se com uma grande base de dados.

02 Substituir por zero

Apenas quando o conjunto de dados é independente do seu efeito.

03 Substituir pela média, mediana ou moda

São aproximações mais coerentes, quando comparadas a substituição por zero.

Manipulação de valores ausentes

04 Interpolar

Gera valores dentro de uma faixa de valores da distribuição dessa variável.

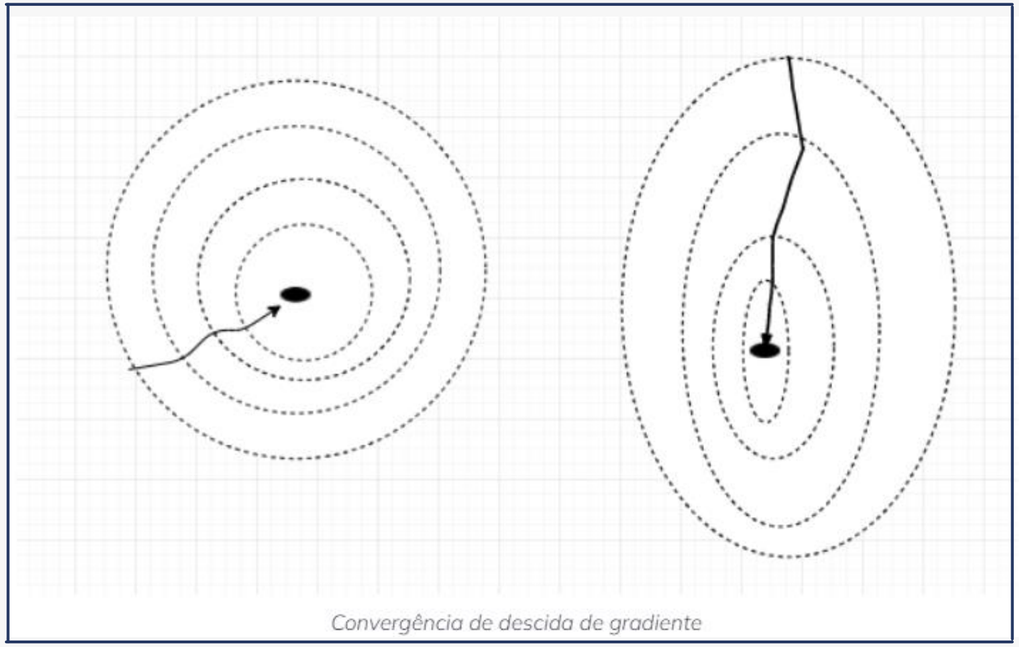
05 Extrapolar

Gera valores além de uma faixa.
Precisa do auxílio de outra variável como referência guiada (em geral o desfecho).

06 Predizer os valores ausentes

Algoritmo estuda as demais variáveis (exceto a com valores faltantes) e prediz seus valores.
Variável com dados ausentes usada como desfecho.

Escala



Convergência de descida de gradiente

01 Min-Max Scaler

Reduz os valores da variável entre uma faixa pré-definida.

02 Standard Scaler

Reduz a escala para ter desvio padrão de 1 e
distribuição centrada em 0.

03 Robust Scaler

Funciona melhor quando há outliers.
Dimensiona os dados em relação ao intervalo interquartil após a remoção da mediana.

04 Max-Abs Scaler

Similar ao Min-Max, mas ao invés de determinar um intervalo, a variável é dimensionada para o valor máximo absoluto. Preserva a dispersão dos dados.

Outliers

01 O que são outliers?

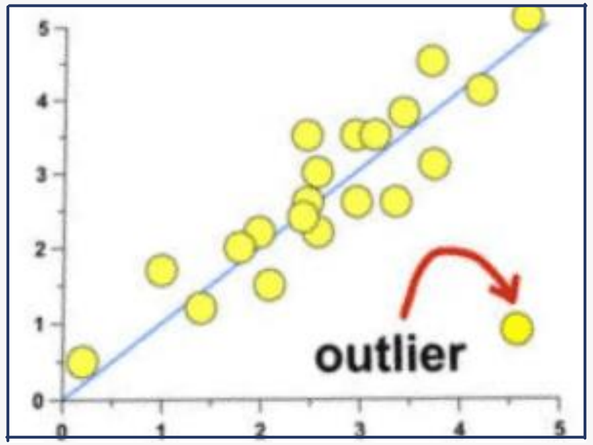
Pontos fora do padrão

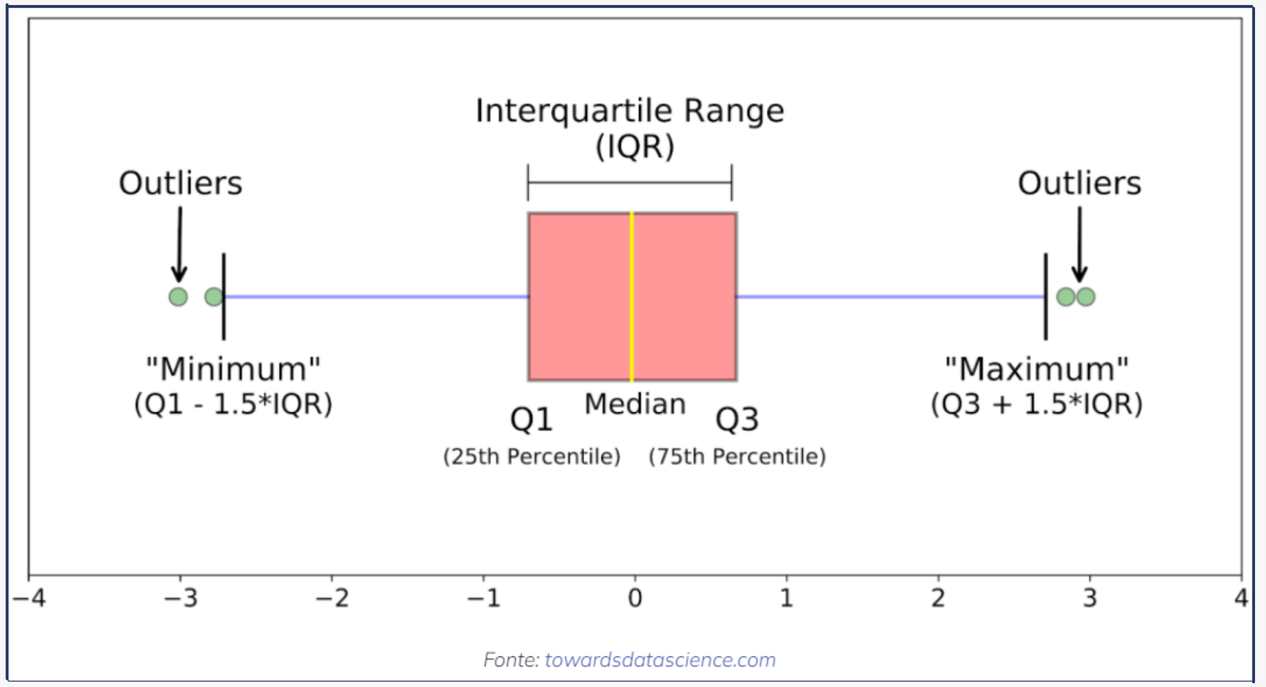
02 Como detectar?

Gráfico box-plot.

03 Como tratar?

Remoção se for erro.

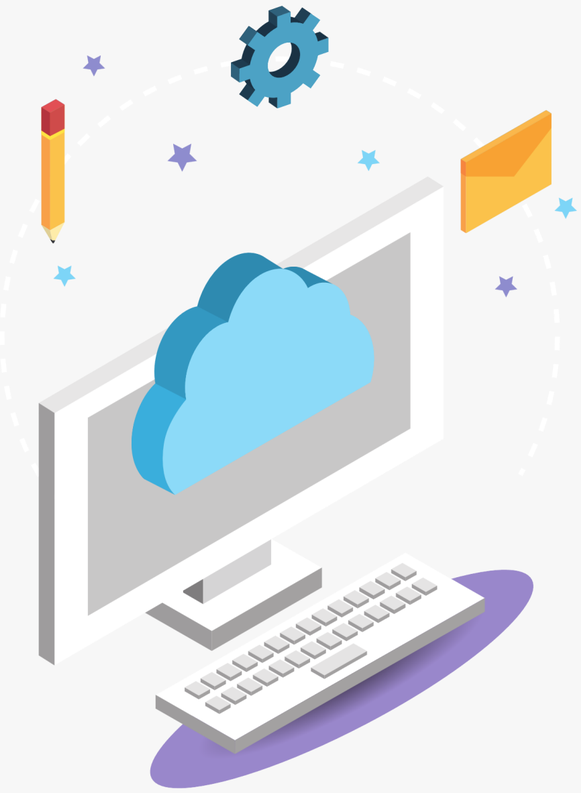




Fonte: towardsdatascience.com

Categorização

Ex.: valores de string em uma coluna.



01 Label/Ordinal Encoding

Atribui para a variável valores inteiros de 1 a n de forma ordinal.

02 One hot encoding

Gera uma coluna binária para cada categoria da variável. (um x todos)

Encoders bayesianos

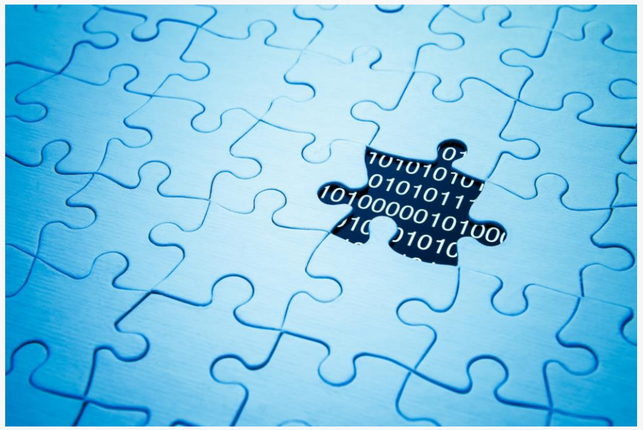
Usa informações da variável dependente nas codificações.

01 Target Encoding

A média do valor alvo (desfecho) por categoria.
Deve-se manter conjunto de teste separado.

02 Weight of Evidence Encoding

Medida em que um valor ou evidência, suporta ou nega uma hipótese.



03 Leave One Out Encoding

Similar ao Target Encoding, mas exclui o atual desfecho da linha quando calcula a média para cada categoria.
Evita outliers e anomalias.

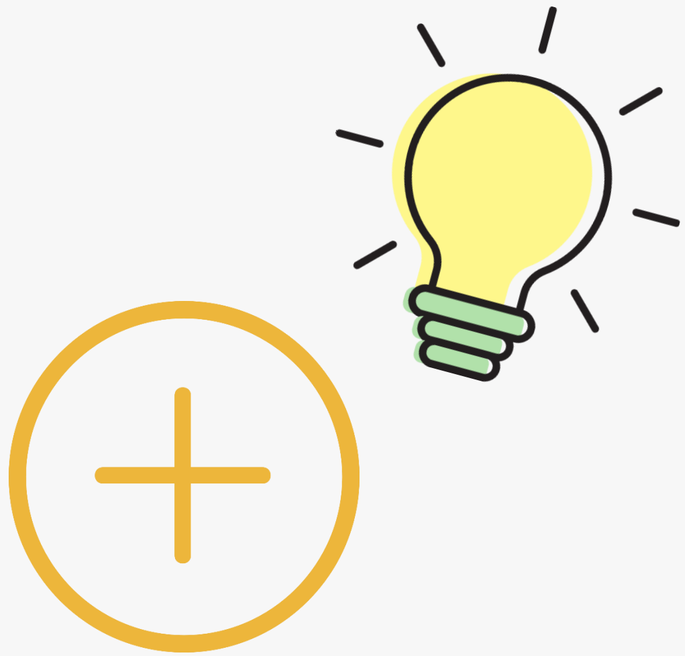
04 James-Stein Encoding

Usa a média ponderada correspondente ao desfecho juntamente com a média de todo o desfecho.
Pesos de acordo com a variância estimada dos valores.
Variância alta indica que essa média não é muito confiável.

Criação e Agregação de variáveis

01 Criar variáveis a partir de outras variáveis

Ex.: Com as variáveis "tempo total" e "distância total" pode-se criar a variável de "velocidade".



02 Construção intuitiva

Captura complexidade. Ex.: quilómetros rodados dependendo do dia da semana.

03 Agregar variáveis

Reduzir a dimensão e criar informação útil.
Ex.: Modelo de série temporal com dados diários de chuva.
O total ou a média diária são úteis, mas a quantidade de chuva a cada hora (dado fracionado) não tanto.

Redução de dimensionalidade

Vantagens:

01 Tempo de processamento mais rápido

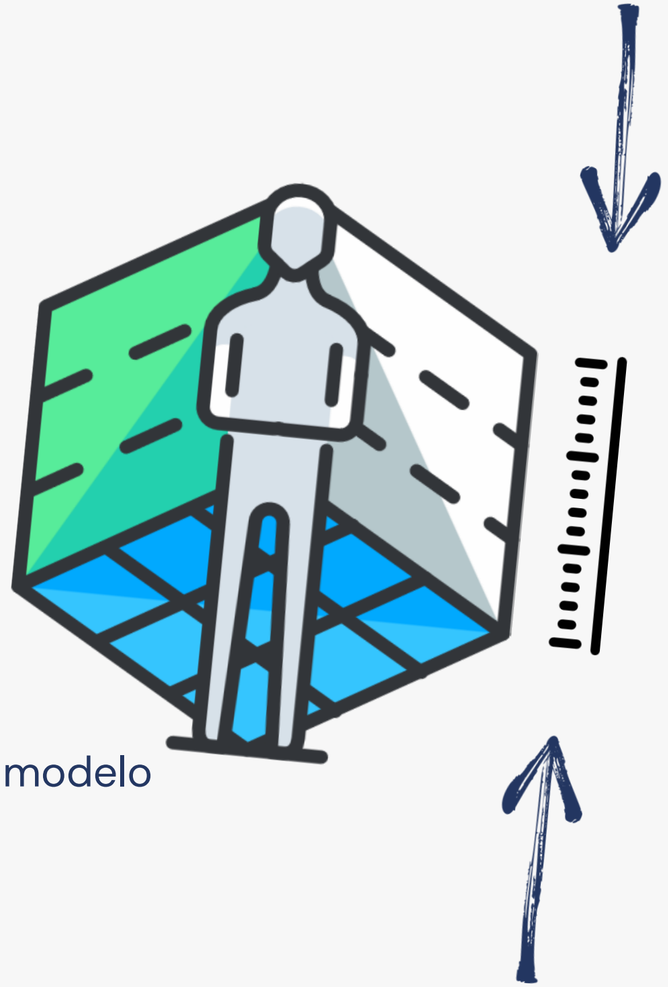
Menor volume de dados = mais rápido treinamento e predição.

02 Acurácia melhorada

Não há variáveis irrelevantes que o modelo possa considerar.

03 Overfitting reduzido

Menos variáveis irrelevantes = Menos propagação de ruído nas decisões do modelo



Seleção univariada

01 Variância

Medida de variabilidade.

Sem variação nos dados, não há padrão a ser absorvido pelo modelo.

Quando há classe minoritária de desfecho, mesmo com baixa variância, ainda é possível que a variável seja uma boa preditora.

USE MÉTODOS DE CORRELAÇÃO!

02 Correlação

Detecta relações lineares entre duas variáveis.

Não capta bem as relações não lineares.

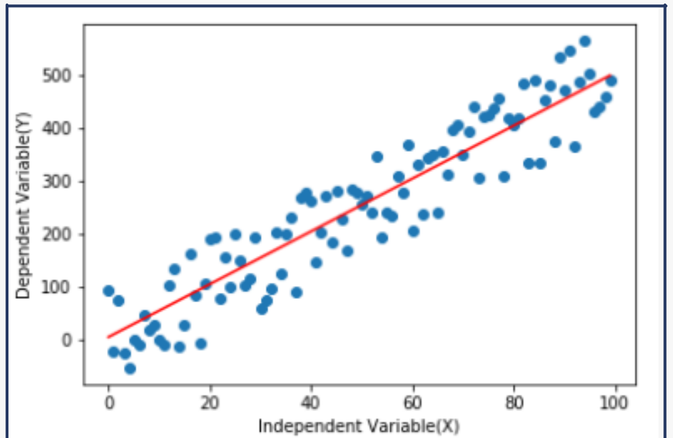
Algumas técnicas populares:

- Pearson: têm de ter relação linear e ambas distribuição normal.
- Spearman Rank: variáveis são medidas um uma escala ordinal.

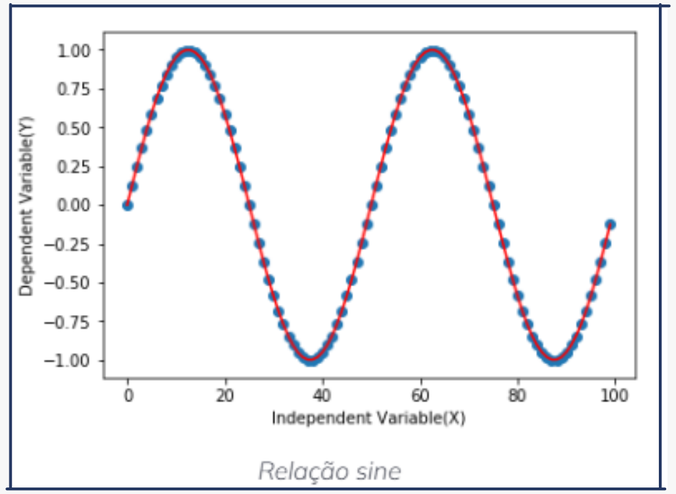
Correlação de medição baseada na variabilidade.

- Kendall Rank: correlação de medição baseada na probabilidade.

TESTE A CORRELAÇÃO ENTRE AS VARIÁVEIS INDEPENDENTES. Retirar a com menos correlação com o desfecho.



Relação linear



Relação sine

	sample	sine	linear
sample	1.000000	-0.389355	0.935135
sine	-0.389355	1.000000	-0.381610
linear	0.935135	-0.381610	1.000000

Escores de correlação

Informações mútuas

Captura a informação não linear.

Responde questões como:

- Quanta informação sobre uma variável pode ser extraída de outra?

- Quanto movimento (aumento ou diminuição) de uma variável pode ser rastreado usando outra variável?

Chi-quadrado

Teste estatístico usado em grupos com variáveis categóricas que avalia a correlação ou probabilidade de associação, com auxílio das distribuições de frequência.

Seleção multivariada

01 Forward Selection

Começa medindo o desempenho do modelo com o mínimo de variáveis, e adiciona outra variável a cada iteração com base no melhor desempenho.

02 Backwards Elimination

Semelhante à anterior, mas na direção inversa. Elimina variáveis a cada iteração com desempenho ruim.
(costuma ser preferido, comparado ao anterior)

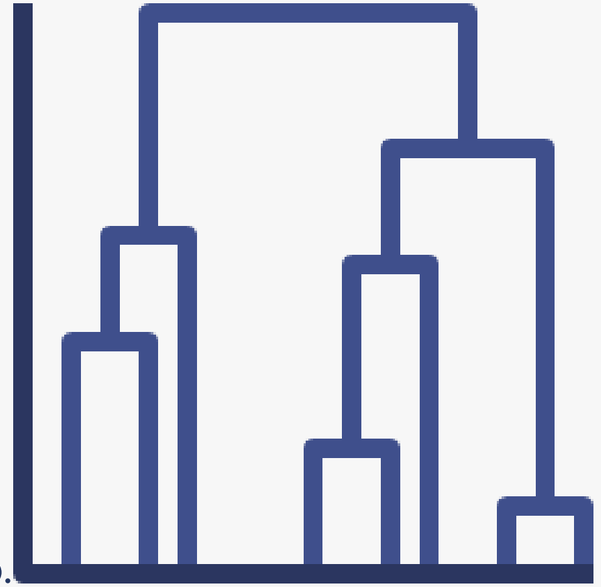
03 Recursive Feature Elimination

Semelhante à anterior, mas substitui a iteração pela recursão.

Seleção multivariada

04 Linear Discriminant Analysis (LDA)

Encontra a combinação linear de caracteristicas que separa duas ou mais classes de uma variável categórica.



05 ANOVA

Conhecida como "Análise da variância".
Teste estatístico para saber se a média de dois grupos é diferente ou não.

06 Embedded methods

Alguns modelos de ML vem com métodos de seleção de variáveis embutido.
Aplicam coeficientes às variáveis com base na importância para o desempenho.
Ex.: Modelo de Ridge e Lasso.

Google Colab

Site:

https://neptune.ai/blog/data-preprocessing-guide

- Autor Samadrita Ghosh
- Atualizado em 16 de agosto de 2021

LABDAPS

LABORATÓRIO DE BIG DATA E ANÁLISE PREDITIVA EM SAÚDE

Obrigado!

Alexandre Chiavegatto Filho

http://labdaps.fsp.usp.br

@SaudenoBR

@labdaps

alexdiasporto@usp.br