Seleção de Features

Diversidados

Quem Somos?

Tatyana Zabanova

Estevão Uyra

Warm up: como vocês fariam seleção de variáveis para um modelo?

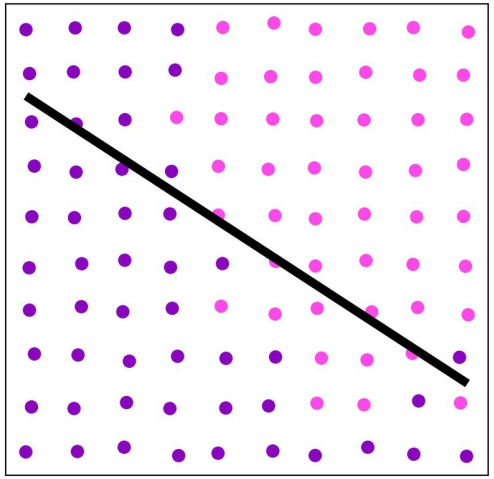
Agenda

1. Revisão
2. Motivação para seleção de variáveis
3. Visão geral dos métodos de seleção
4. Filtragem de variáveis
5. Mão na massa

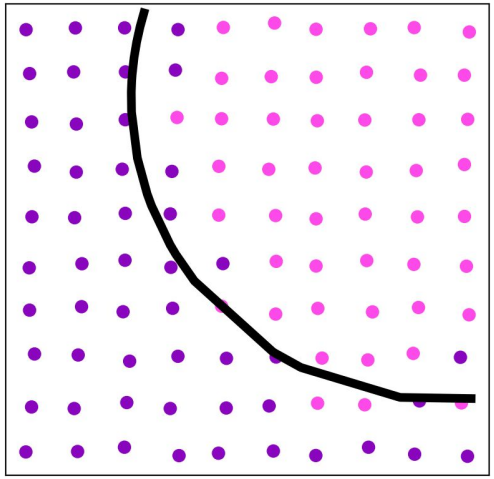
Revisão:
Subajuste vs Sobreajuste

Pontos coloridos segundo uma regra desconhecida
Tarefa: separar os pontos por cor

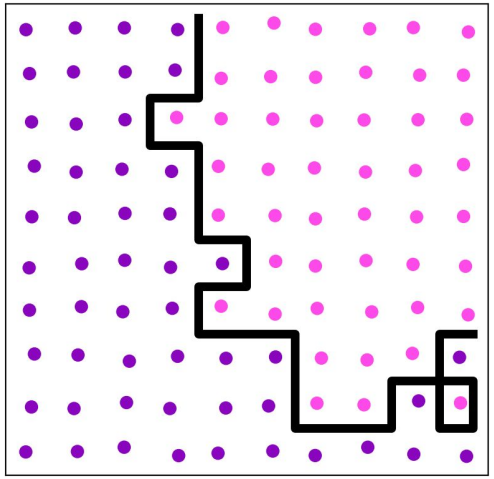
78/100



91/100



100/100

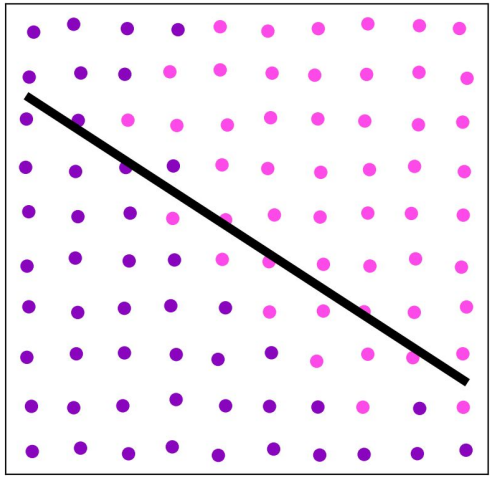


Essas regras continuam funcionando se aplicarmos elas a um conjunto novo de pontos (mas com a mesma lógica)?

“Os modelos *generalizam*?”

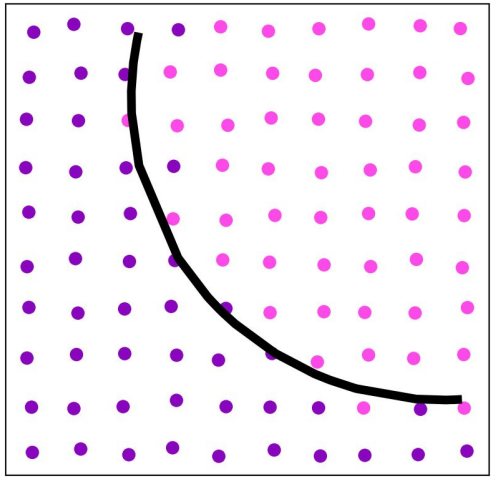
Essas regras continuam funcionando se aplicarmos elas a um conjunto novo de pontos (mas com a mesma lógica)?

81/100



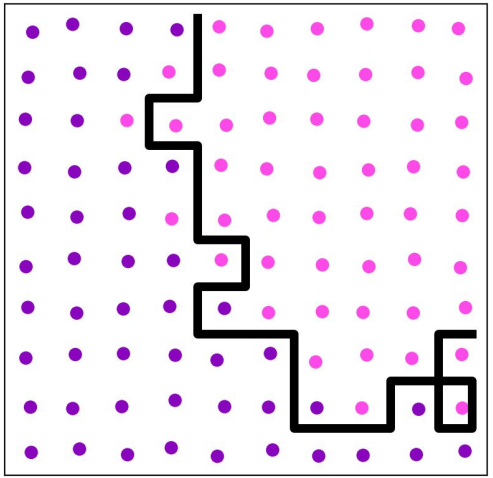
Continua na mesma

94/100



Também continua na mesma

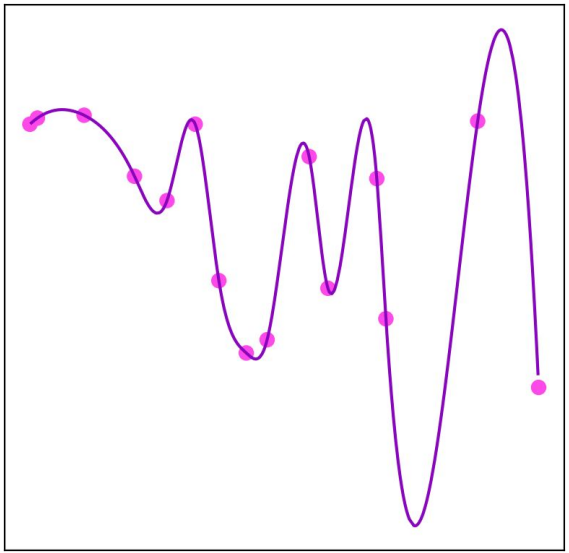
93/100



Piorou

Qualidade do ajuste

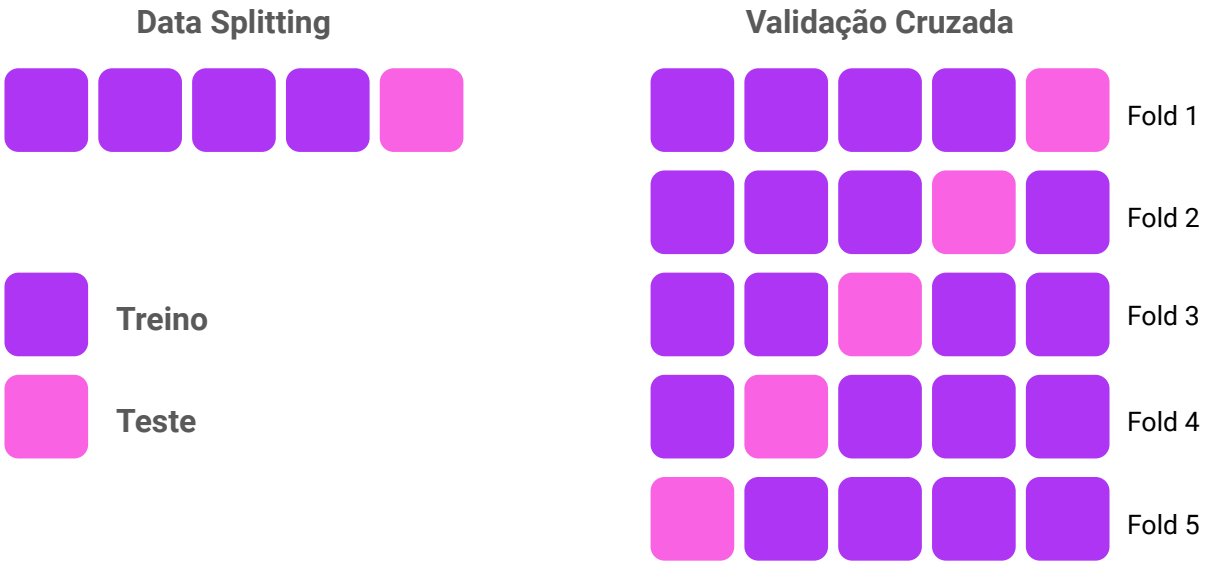
- Dados contém:
  - Relações entre variáveis e a resposta. Exemplo: objetos grandes tendem a pesar mais que os pequenos
  - Ruído, coisas que são específicas dessa amostra e não dos seus dados de forma geral. Exemplo: a minha mochila azul é absurdamente pesada
- **Subajuste (underfitting)**: o modelo não consegue explicar a resposta da melhor forma possível
- **Sobreajuste (overfitting)**: aprende coisas que são muito específicas da amostra, mas não se aplicam ao geral



Revisão:
Data Splitting e Validação Cruzada

Como sei se a relação é geral ou específica?

Separamos o conjunto de dados em várias partes, na esperança de que as relações específicas da amostra não se mantenham entre partes.



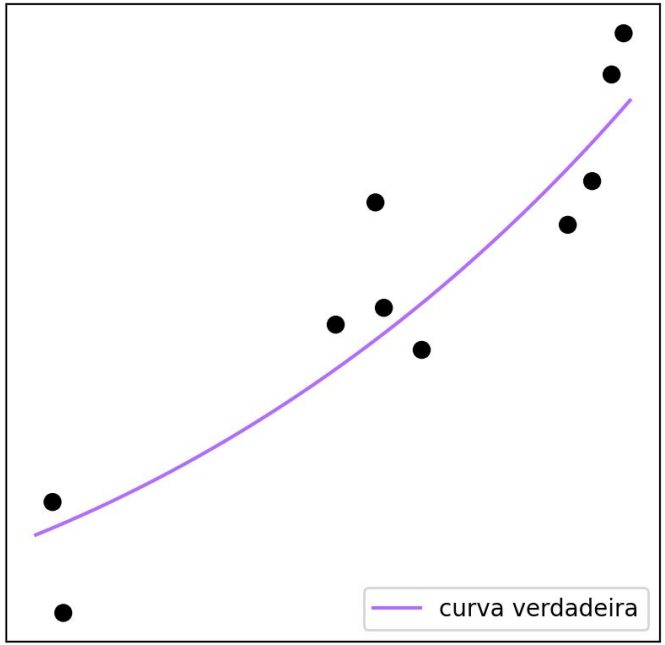
Seleção de Features: Motivação

Motivação principal: Generalizar

Quando treinamos um modelo, queremos aprender as relações gerais:

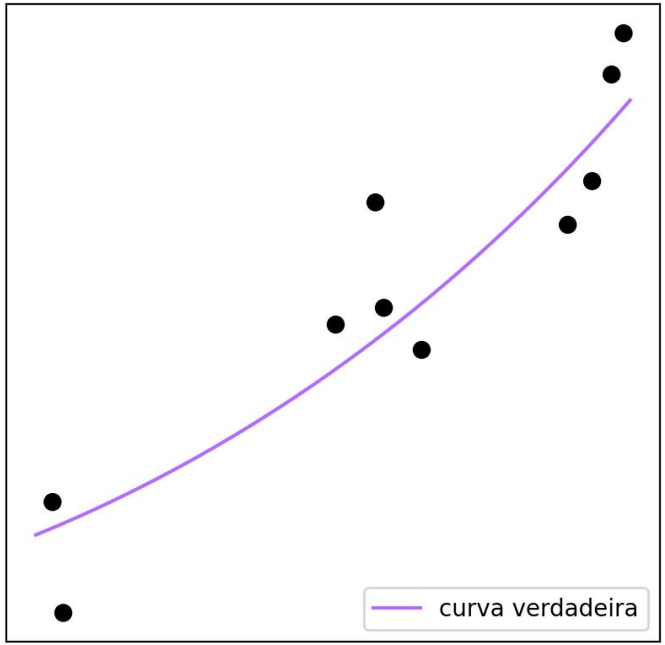
- Predizer a resposta para novas observações
- Entender e quantificar as relações entre variáveis e a resposta

Motivação principal: Generalizar



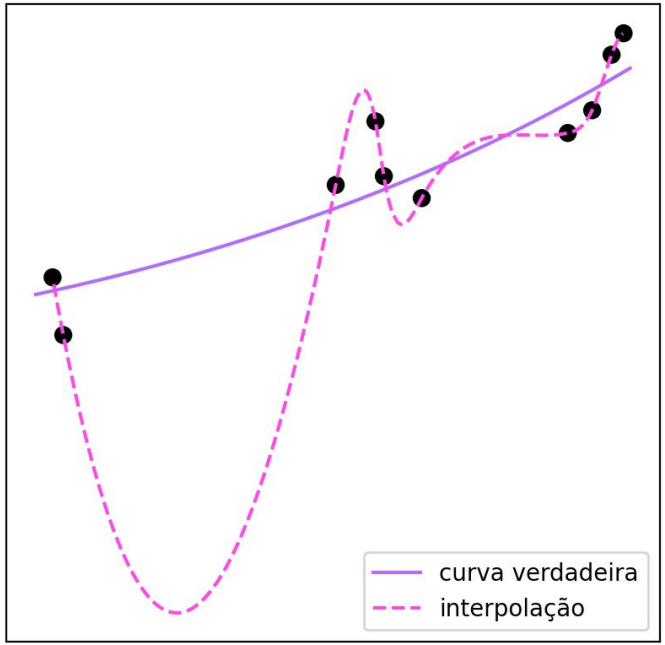
Um exemplo clássico de seleção de features:
Interpolação Polinomial

- Pontos: gerados por polinômio de grau 5 (+ ruído)



Uma curva que passa por todos os pontos é um ajuste bom?

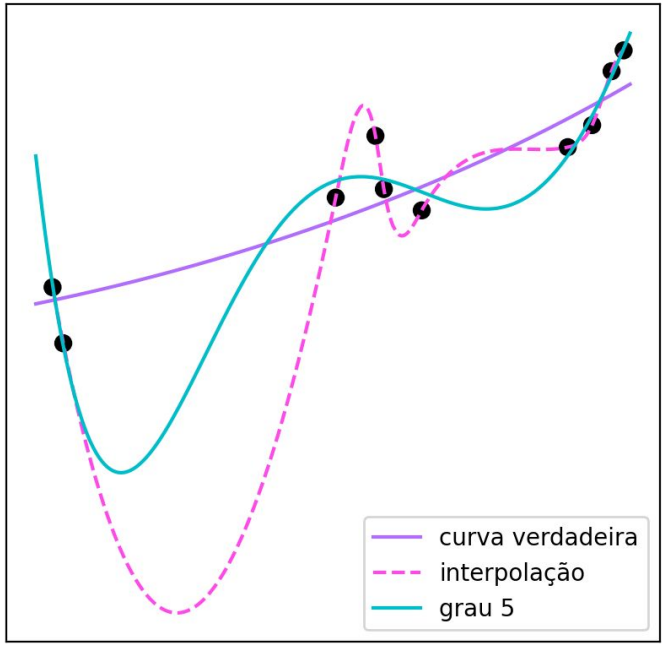
Motivação principal: Generalizar



Um exemplo clássico de seleção de features:
Interpolação Polinomial

- Pontos: gerados por polinômio de grau 5 (+ ruído)
- Podemos usar um polinômio que vai passar por **todos** os pontos (9 variáveis)
- Claramente não estamos generalizando bem aqui

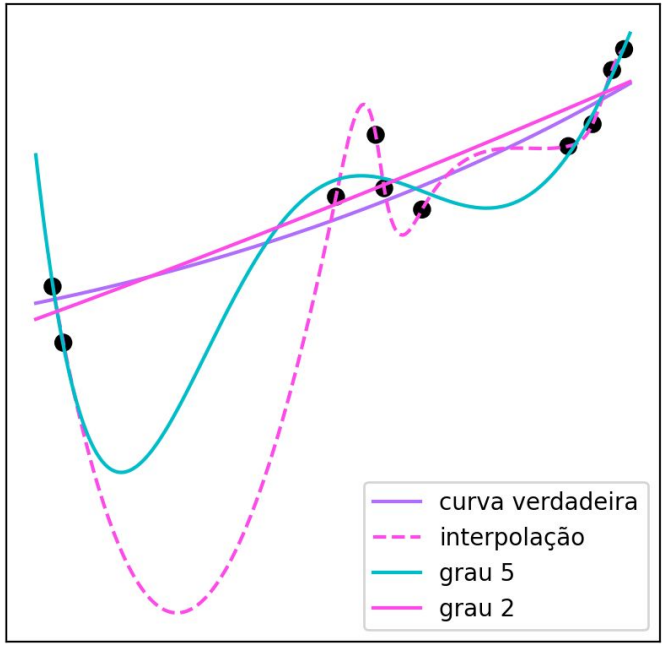
Motivação principal: Generalizar



Um exemplo clássico de seleção de features:
Interpolação Polinomial

- Pontos: gerados por polinômio de grau 5 (+ ruído)
- Podemos usar um polinômio que vai passar por **todos** os pontos
- Mesmo se ajustamos um polinômio de grau 5, ele "aprende" parte do ruído dos dados

Motivação principal: Generalizar

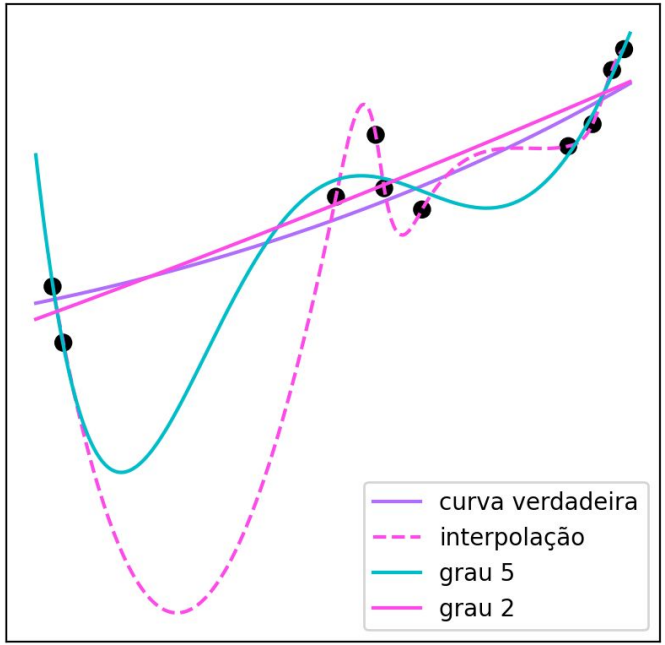


Um exemplo clássico de seleção de features:
Interpolação Polinomial

- Pontos: gerados por polinômio de grau 5 (+ ruído)
- Podemos usar um polinômio que vai passar por **todos** os pontos
- Mesmo se ajustamos um polinômio de grau 5, ele "aprende" parte do ruído dos dados
- Um polinômio de grau 2 parece capturar melhor a tendência por trás dos pontos

PS: Sim, a especificação correta do modelo não garante melhor poder preditivo

Motivação principal: Generalizar



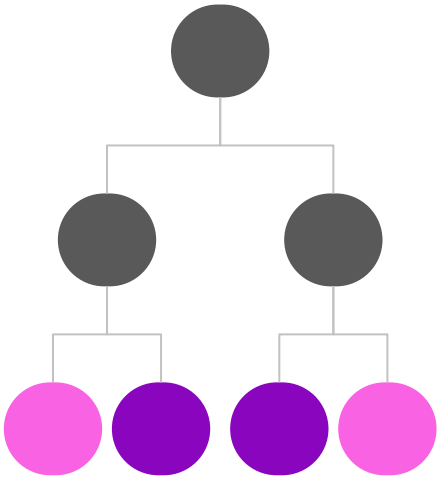
Variáveis que “não fazem sentido” podem ser usadas pelo modelo **por causa de coleta de dados enviesada**

- Variáveis que evidentemente não deveriam afetar o modelo (exemplo: signo)
- Variáveis que interagem com condições experimentais, no caso de um experimento
- Variáveis que são representativas do conjunto de treino mas não do domínio de aplicação

Esse tipo de problema é mais complicado, e geralmente envolve bom-senso e conhecimento do domínio de aplicação do modelo.

Motivações secundárias

- **Parcimônia**: você pode acabar colocando a mesma variável duas vezes em modelos de árvores ou boosting se não ficar de olho.
- **Eficiência computacional**: com menos variáveis, processamento dos dados, treino do modelo, cálculo das predições ficam mais rápidos.
- Menos coisas para **quebrar**, menos coisas para **monitorar**.
- **Interpretabilidade**: quanto mais variáveis, mais *caixa-preta* é o modelo.



Vocês conseguem pensar em alguma outra motivação para remover variáveis do modelo?

Atenção

Até aí, falamos principalmente de aspectos de performance, mas seleção de variáveis também pode se basear em **outros critérios**...

- **Custo de obtenção**. Exemplo: exige uma medição muito cara
- **Dificuldade de obtenção**. Exemplo: leva um período muito longo para medir
- **Qualidade**. Exemplo: a variável provém de uma fonte mal documentada, e você não confia na reprodutibilidade daqueles valores.
- **Questões éticas**. Exemplo: variáveis como gênero ou raça, variáveis obtidas de forma anti-ética, etc.
- **Questões legais**. Exemplo: modelos de crédito

Não vamos abordar esses aspectos no curso, mas é algo que vocês **precisam ter em mente** ao desenvolver modelos.

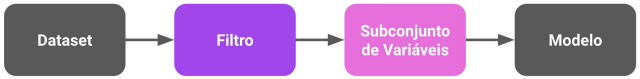
Seleção de Features: Visão Geral

Tipos de seleção de features

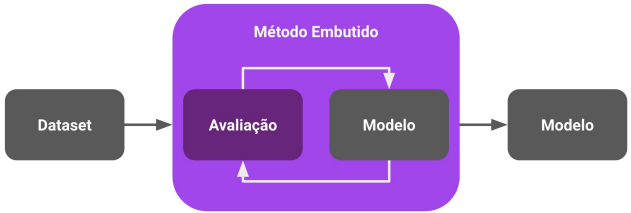
Três principais grupos de abordagens de seleção de features:

- Baseados em **filtros**: a ideia é pensar em algum critério e descartar (filtrar) todas as variáveis que não satisfazem esse critério
- Métodos "**embutidos**": alguns algoritmos já incluem seleção de feature na sua lógica
- **Wrappers**: seleção de features como problema de busca, testando várias combinações de features

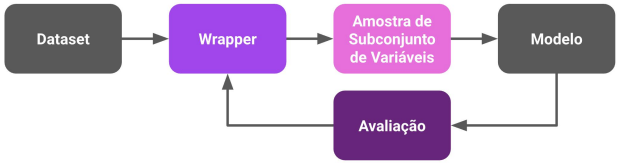
Filtro



Seleção embutida

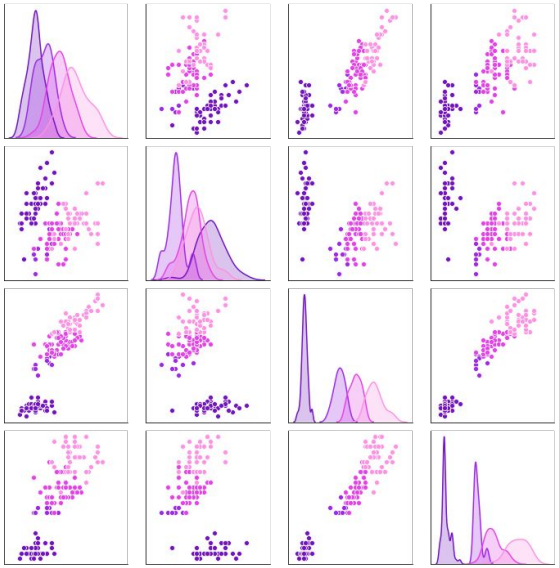


Wrapper



Seleção de Features: Filtros

Filtros

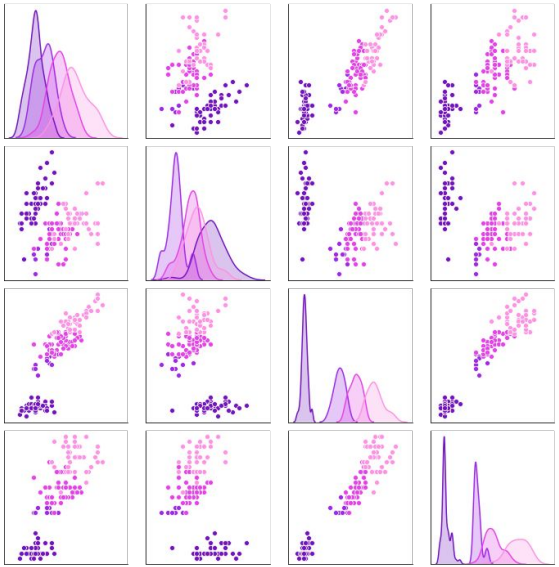


Ideia geral: definir algum critério para eliminar variáveis da análise que permita jogar boa parte delas fora antes de chegar na modelagem.

Alguns filtros comuns:

- Pouca variabilidade
- Alta correlação com outras variáveis
- Baixa correlação com a resposta
- Comportamento estranho
- Computacionalmente ineficiente

Filtros



Vantagens do método:

- **Agnóstico ao modelo** (pelo menos parcialmente)
- **Rápido e eficiente**: permite cortar muitas variáveis rapidamente

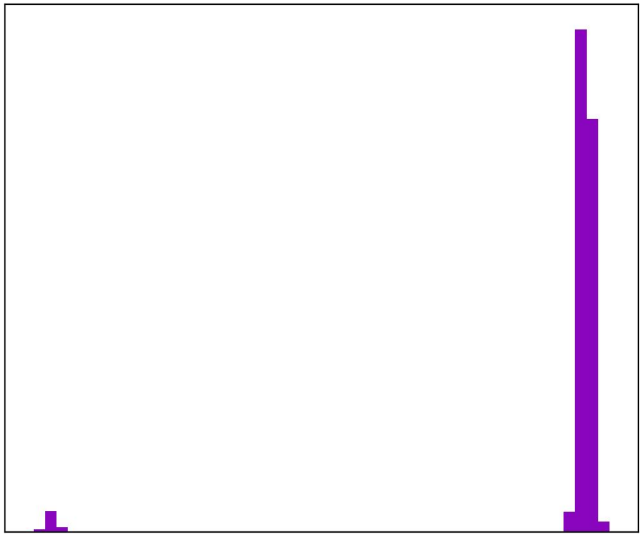
Consequência: filtragem é o primeiro passo da seleção, removendo de forma grosseira as variáveis que claramente não ajudam ou que são redundantes.

Baixa variabilidade

Intuição: e se uma variável assumir só um valor? Como ela é constante, não vai explicar nada da resposta, e por isso não tem motivo de botar essa variável no modelo.

Exemplos:

- Variável com muitos nulos
- Variável categórica ou numérica que tem o mesmo valor para a maioria absoluta das observações
- Variável com variância muito baixa



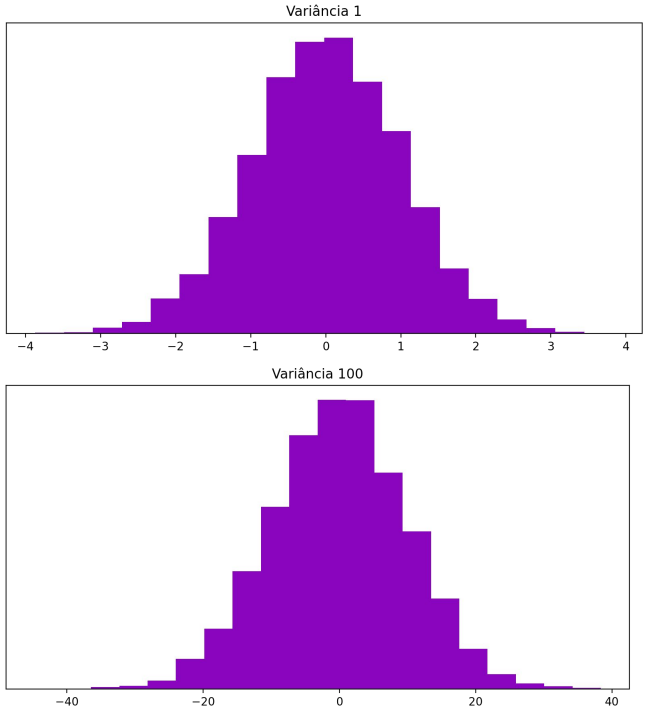
Como vocês acham que isso pode dar errado?

Baixa variabilidade

Regra do dedão: se você não sabe por onde começar, use 95% como ponto de corte.

Para tomar cuidado:

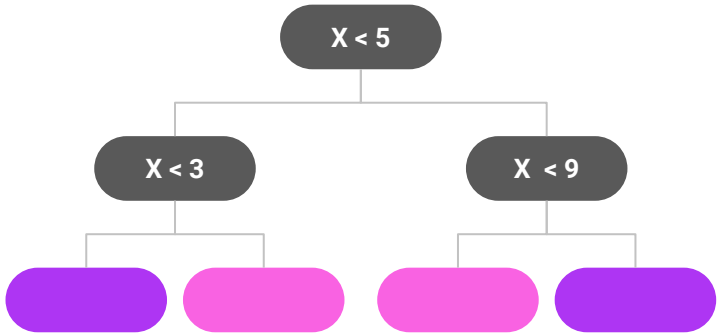
- Os poucos valores diferentes podem ser justamente aqueles que explicam a resposta para algum subgrupo. Você pode jogar fora informação relevante
- A variância das variáveis depende da sua escala. Se você for usar variância como corte, coloque tudo na mesma escala primeiro (não normalize)

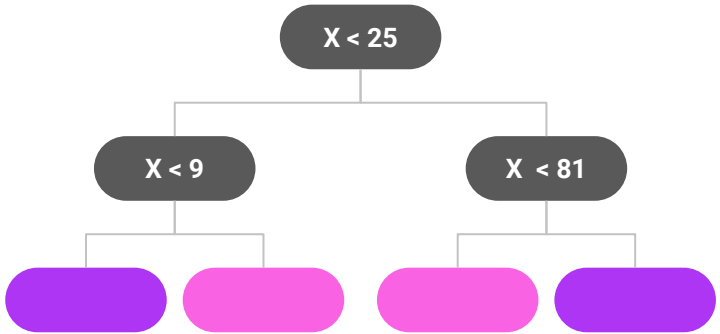


Correlação

Essa parte é **impactada pela escolha do modelo**.

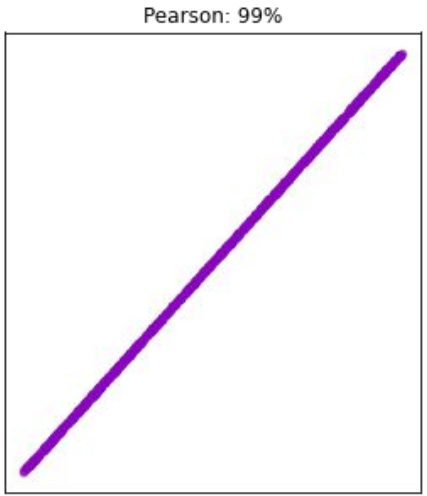
Exemplo: **regressão linear** vs modelo baseado em **árvore** (decision tree, random forest, boosting), e variáveis **X** e **X^2**. Para regressão, essas variáveis são diferentes, mas para uma árvore de decisão, elas são iguais.

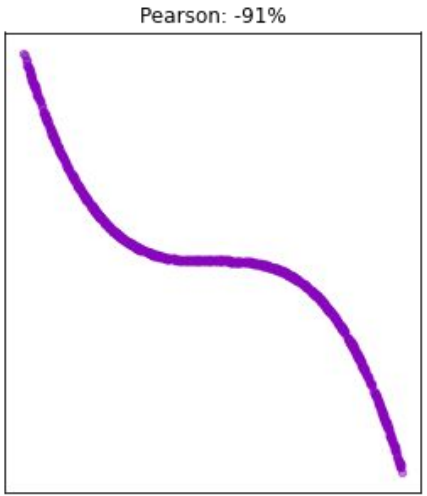


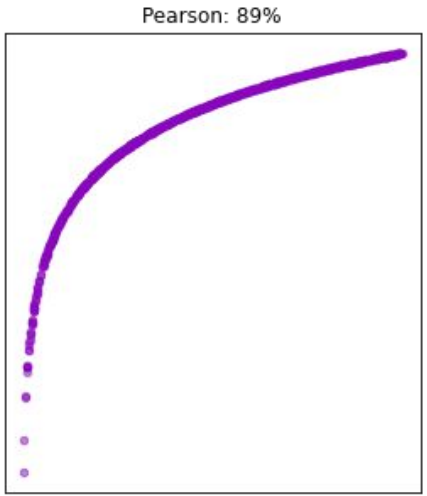


Correlação: Pearson

- Variáveis numéricas
- Correlação linear
- Modelos: Regressão Linear, Logística, Lasso, etc

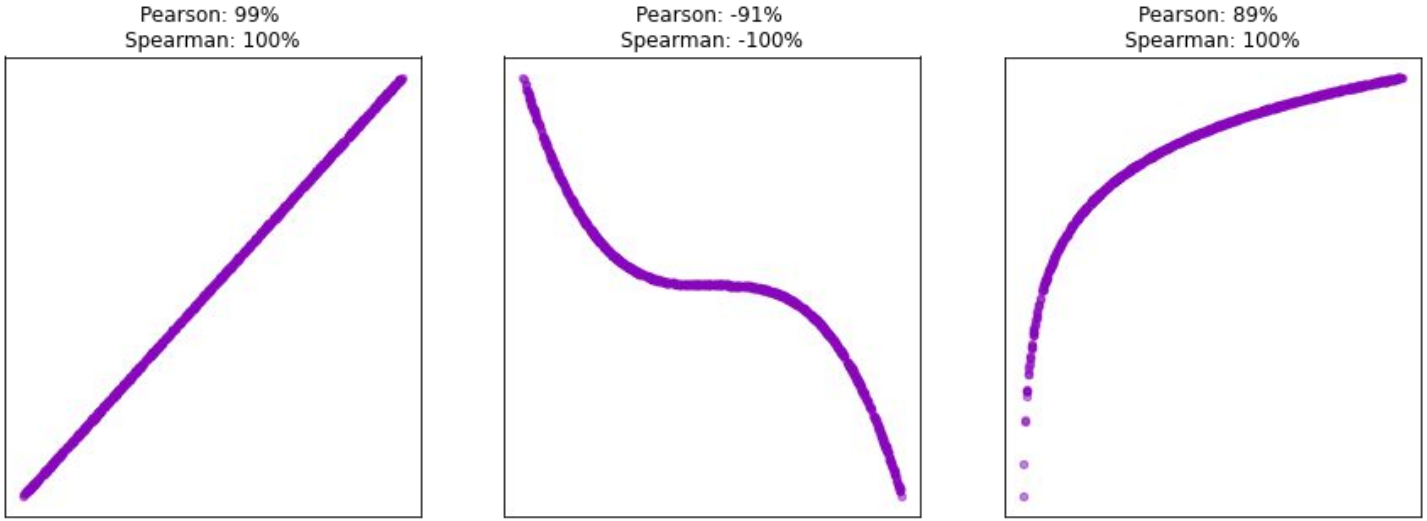






Correlação: Spearman

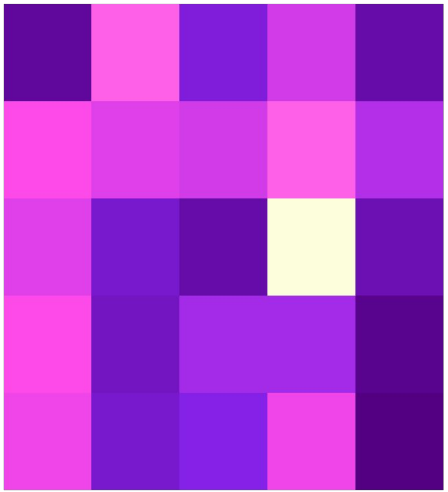
- Variáveis numéricas
- Correlação ordinal (relação monotônica entre duas variáveis)
- Modelos: Árvore de Decisão, Floresta Aleatória, Boosting, etc



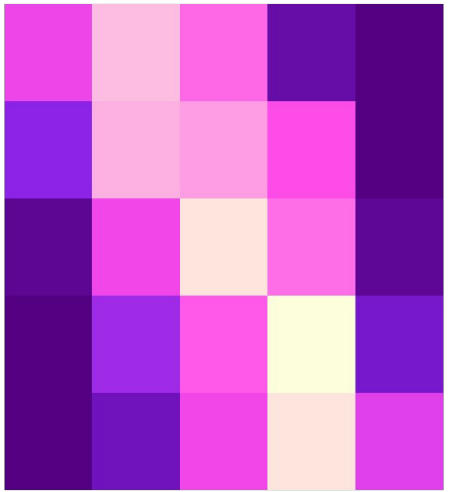
Correlação: Qui Quadrado

- Variáveis categóricas
- Testa se a distribuição de uma variável é a mesma para todas as categorias da outra

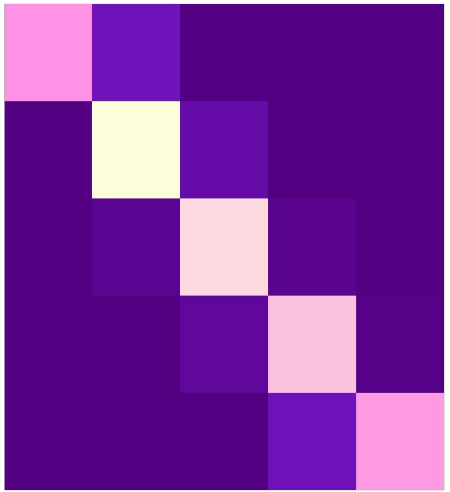
Qui Quadrado: 22



Qui Quadrado: 497



Qui Quadrado: 2818



Baixa correlação com a resposta

Se uma variável explica a resposta, vai existir correlação entre essa variável e a resposta.

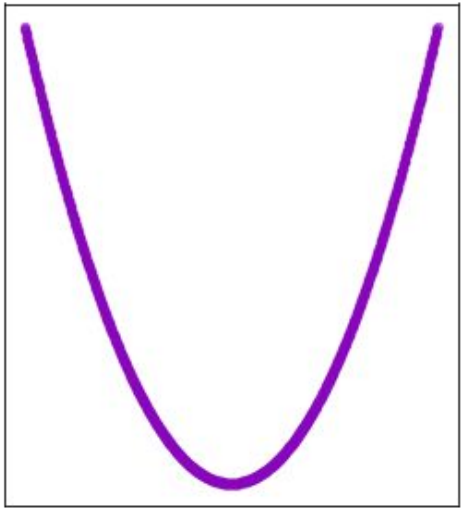
- Pearson, Spearman e outras medidas de correlação
- Entropia e Ganho de Informação (Information Gain)

Como vocês acham que isso pode dar errado?

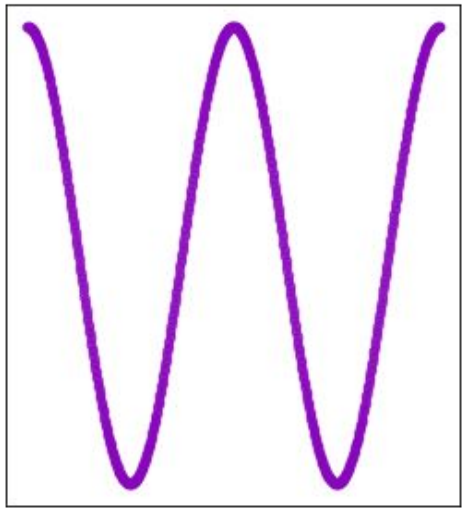
Baixa correlação com a resposta

Problema: se não conseguimos detectar correlação, não significa que de fato ela não existe.

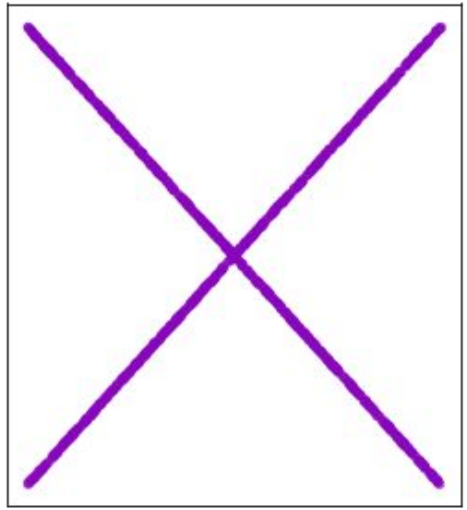
Pearson: 0%
Spearman: 0%



Pearson: 0%
Spearman: 0%



Pearson: 0%
Spearman: 0%



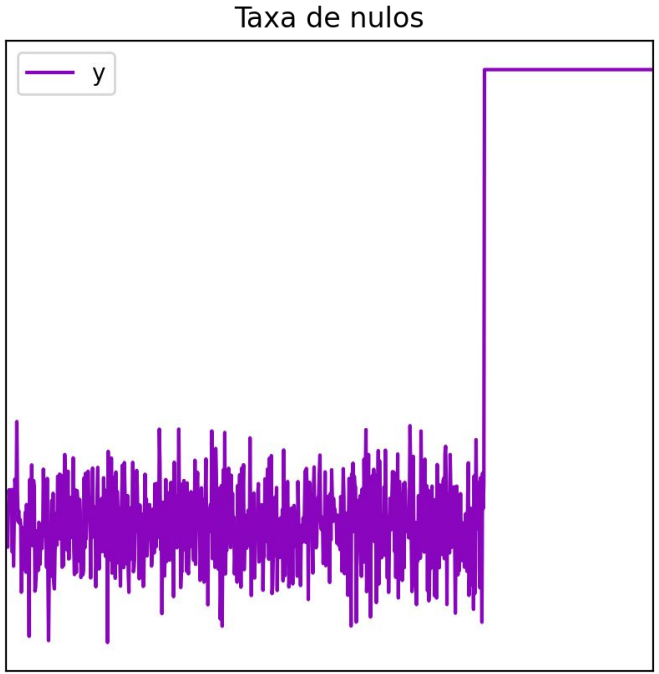
Comportamento estranho

Num mundo ideal, as nossas variáveis são perfeitas, medidas e calculadas corretamente, não quebram, não ficam indisponíveis, etc

Na prática, seleção de features envolve uma **análise descritiva exploratória** (*Exploratory Descriptive Analysis* ou EDA em inglês). Não existe um workflow, pois as analises dependem do seu problema, mas as coisas abaixo são sempre úteis:

- Histograma
- Taxa de nulos ao longo do tempo
- Média / Percentis / Distribuição por categoria ao longo do tempo

Exemplo 1: Nulos



A taxa de nulos de uma variável vai para 100% a partir de uma certa data.

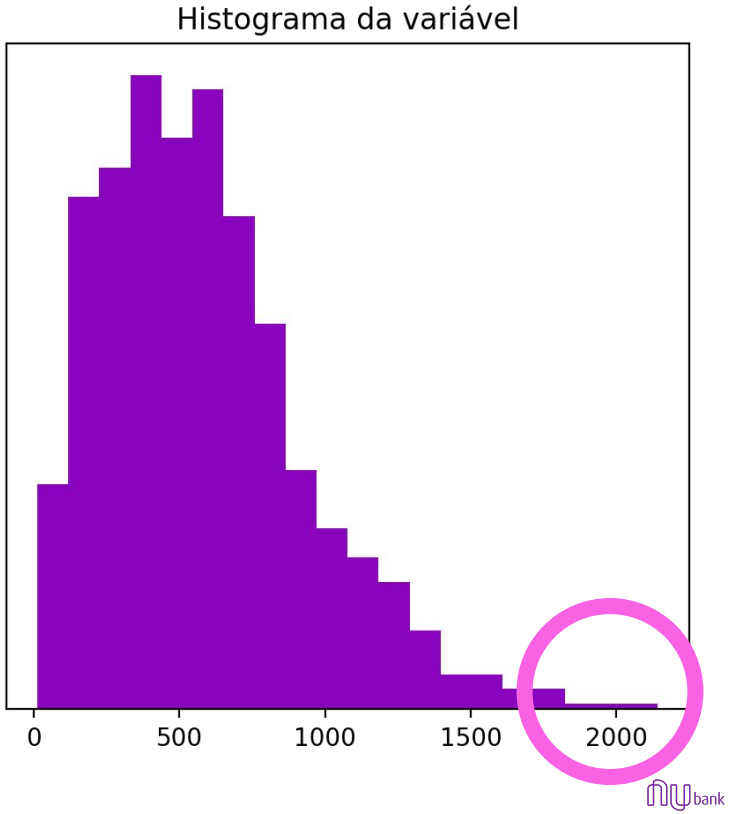
Investigando, descobrimos que esta variável não é mais medida, mas continua na base de dados para se ter o registro.

Não podemos usar ela num modelo preditivo, porque não teremos valores dela para dados novos.

Exemplo 2: Valores incorretos

Sabemos que um determinado dado deve ser armazenado na base de dados por 5 anos, e depois descartado. No nosso dataset temos uma variável definida por número de dias desde a primeira entrada do id na base de dados. Por definição, esta variável não pode passar de 5.

Na prática, observamos valores acima de 5 anos (1825 dias) por alguma razão. Isso é algum bug? E os demais valores, estão corretos?

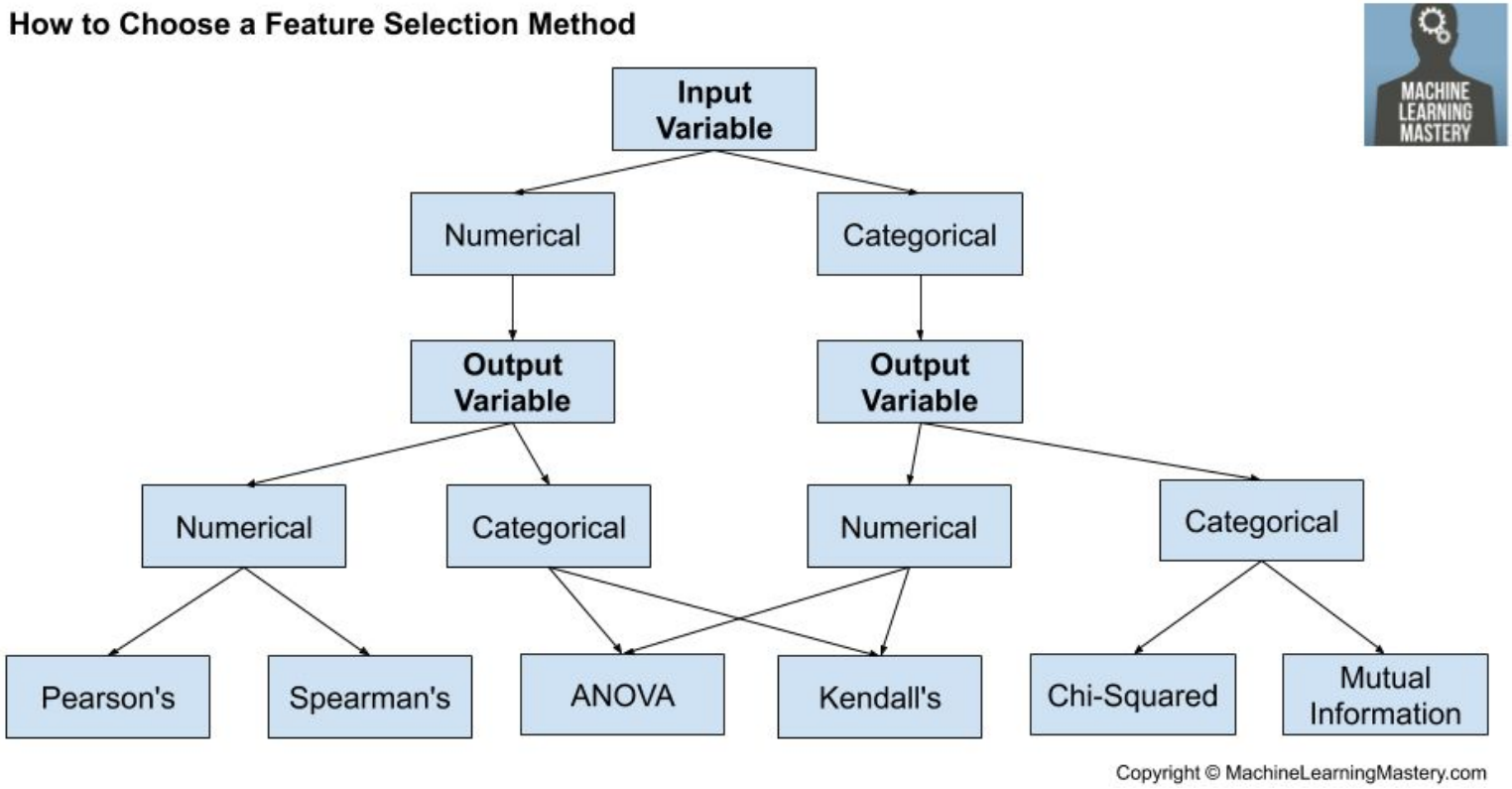


Conseguem pensar em outros exemplos de comportamento estranho?

Dúvidas?

Colinha

Correlações para usar em filtragem



CHAPTER

DATA SCIENCE