



Data Science & Business Analytics

Machine Learning Models

David Issá

davidribeiro.issa@gmail.com

1. Feature Selection

1. Feature Selection

Corresponde ao processo de **selecionar um subconjunto de variáveis relevantes a utilizar** na construção do **modelo preditivo**.

O objetivo é o de **reduzir o número de variáveis de entrada** para aquelas que se acredita serem **mais úteis para o modelo** prever a variável-alvo:

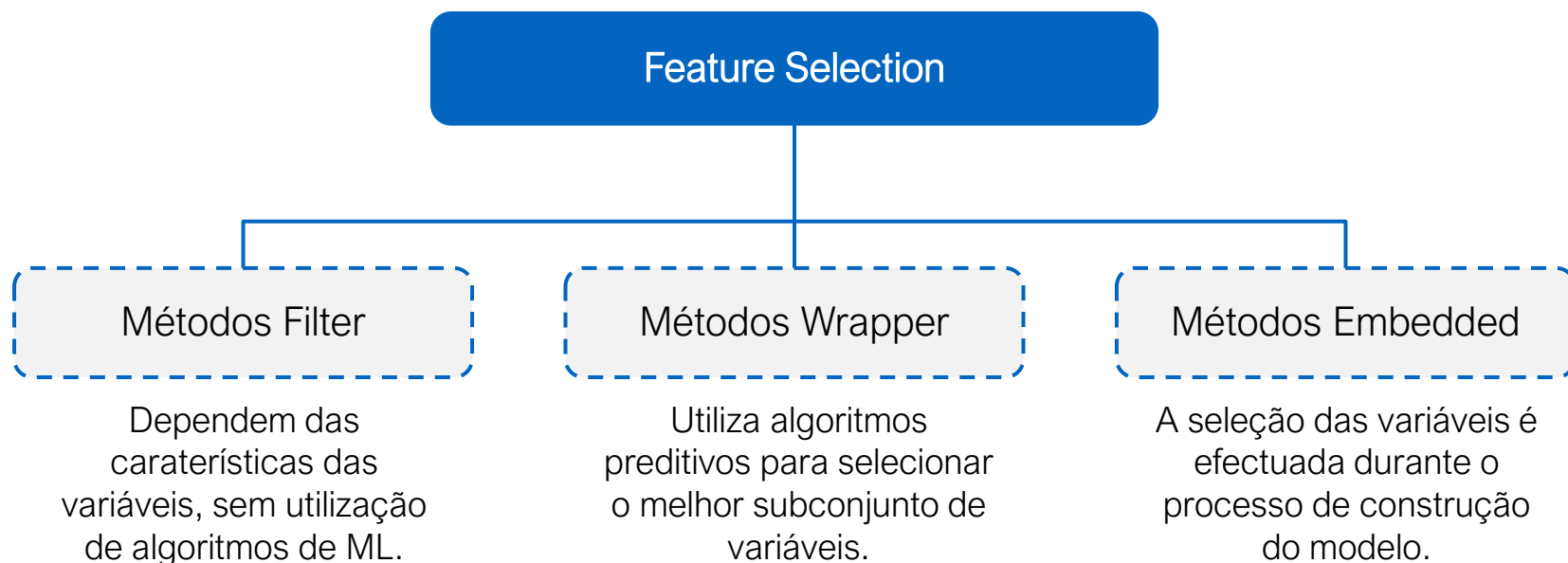
- Melhoria do desempenho: redução de overfitting;
- Redução do custo computacional: tempo de treino mais curto;
- Modelos simples são mais fáceis de compreender;
- Evitar redundância de variáveis.

1. Feature Selection

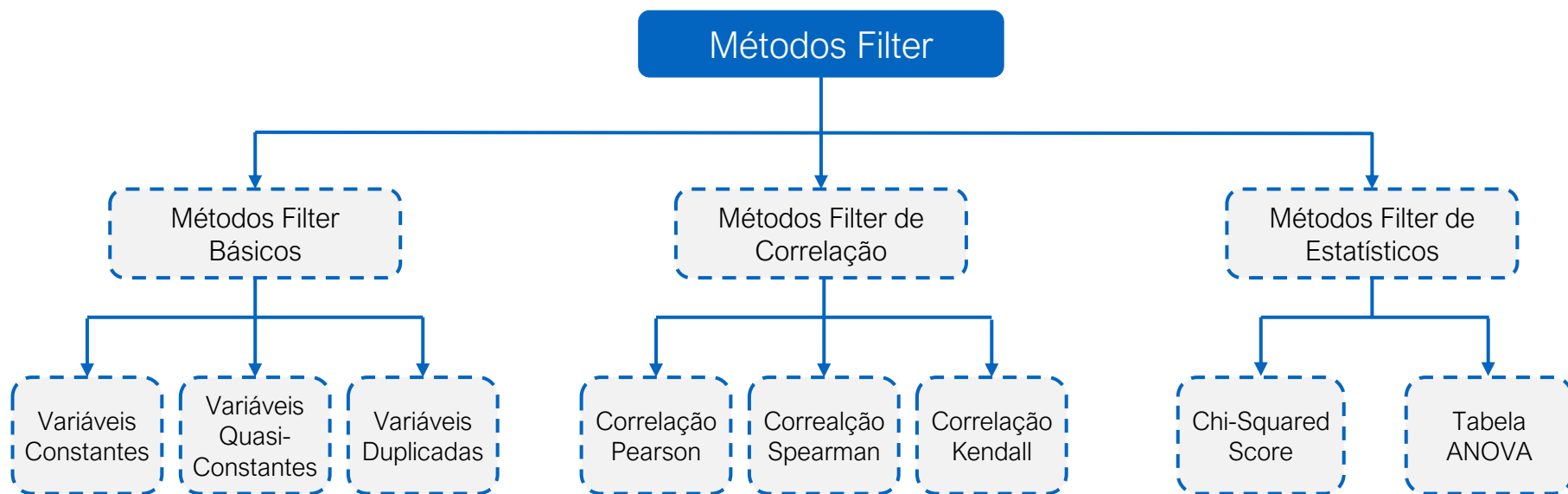
Por vezes, confundem-se os conceitos de Feature Selection, Feature Engineering e Dimensionality Reduction:

- **Feature Engineering:** criar novas variáveis a partir das originais, com o objetivo de criar modelos mais eficazes e com melhor desempenho.
- **Dimensionality Reduction:** uso de algoritmos de unsupervised learning para reduzir o número de variáveis num conjunto de dados. Estas técnicas modificam ou transformam as variáveis para um espaço dimensional inferior.
- **Feature Selection:** permite selecionar variáveis de todas as variáveis disponíveis para criação de melhores modelos e mais eficientes.

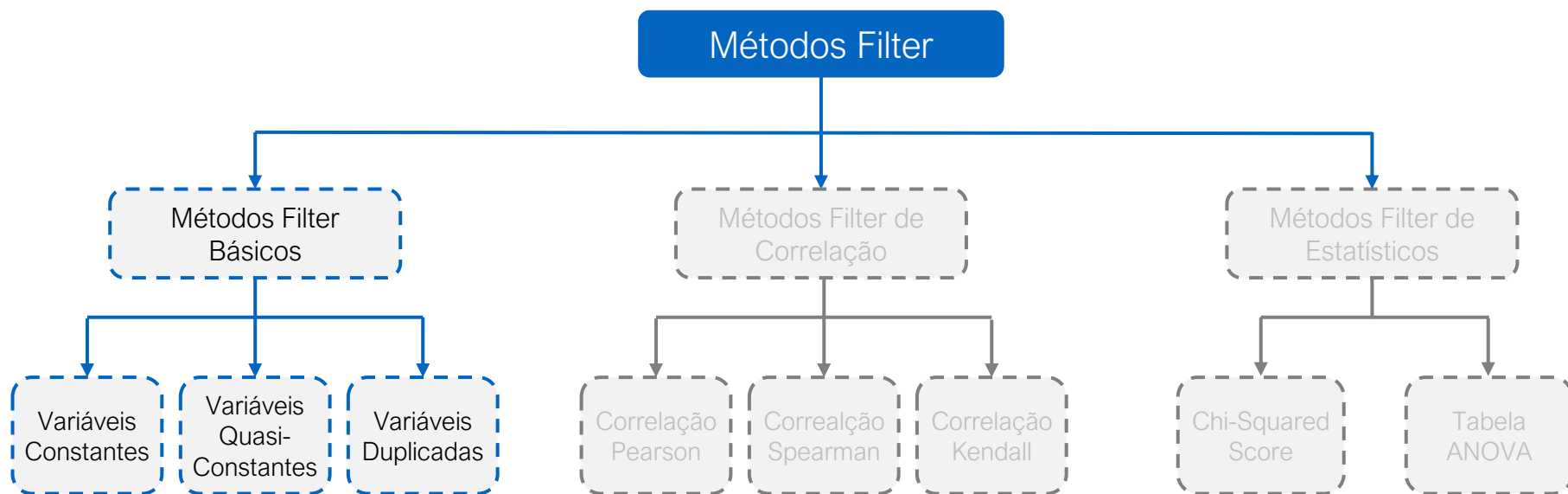
1. Feature Selection



1.1 Feature Selection – Métodos Filter



1.1.1 Feature Selection – Métodos Filter Básicos



1.1.1 Feature Selection – Métodos Filter Básicos

Métodos Filter Básicos – Exemplo:

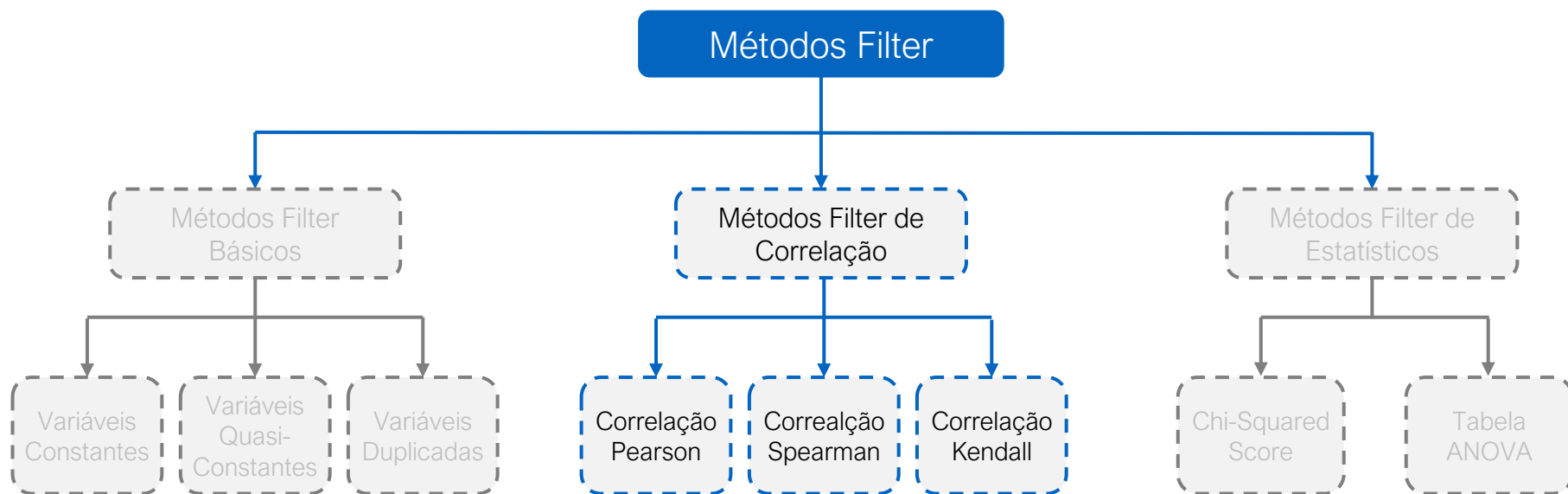
ID	var1	var2	var3	var4	target
1	Female	Yes	Portugal	PT	0
2	Female	Yes	Spain	SP	1
3	Female	Yes	Portugal	PT	1
4	Female	Yes	France	FT	1
5	Female	Yes	Germany	DE	0
6	Female	Yes	Portugal	PT	0
7	Female	No	United Kingdom	UK	0
8	Female	Yes	Spain	SP	1
9	Female	Yes	France	FR	1

Variável
Constantes

Variável
Quasi-Constantes

Variáveis
Duplicadas

1.1.2 Feature Selection – Métodos Filter Correlação



1.1.2 Feature Selection – Métodos Filter Correlação

A correlação mede o grau em que duas variáveis se movem uma em relação à outra.

Mas as variáveis correlacionadas são sempre más?

- Se dois preditores estiverem altamente correlacionados, então fornecem informação redundante. Podemos usar apenas uma dessas variáveis.
- **MAS** se um preditor estiver altamente correlacionado com o alvo, esta deve ser usada no subconjunto final de variáveis escolhidas.

1.1.2 Feature Selection – Métodos Filter Correlação

A correlação mede o grau em que duas variáveis se movem uma em relação à outra.

Mas as variáveis correlacionadas são sempre más?

- Se dois preditores estiverem altamente correlacionados, então fornecem informação redundante. Podemos usar apenas uma dessas variáveis.
- **MAS** se um preditor estiver altamente correlacionado com o alvo, esta deve ser usada no subconjunto final de variáveis escolhidas.

Se a variável A e B estiverem altamente correlacionadas entre si, **devemos manter a variável A ou B?**

- Devemos utilizar mais técnicas de feature selection para compreender o peso de cada variável no objetivo.

1.1.2 Feature Selection – Métodos Filter Correlação

Correlação Pearson (variável independente e variável alvo ambas numéricas)

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

O resultado varia entre 1 e -1:

- Se > 0 , então a relação entre as 2 variáveis é positiva;
- Se < 0 , então a relação entre as 2 variáveis é negativa;
- Quanto mais perto dos extremos, mais forte é a relação.

1.1.2 Feature Selection – Métodos Filter Correlação

Correlação Pearson (variável independente e variável alvo ambas numéricas)

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \sqrt{\sum (y_i - \bar{y})^2}}$$

ID	x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
1	87	241	10	27	270	100	729
2	46	160	-31	-54	1674	961	2916
3	72	210	-5	-4	20	25	16
4	64	195	-13	-19	247	169	361
5	123	285	46	-71	3266	2116	5041
6	97	233	20	19	380	400	361
7	50	174	-27	-40	1080	729	1600
Avg	77	214	NA	NA	NA	NA	NA
Sum	NA	NA	NA	NA	6937	4500	11024

$$r_{xy} = \frac{6937}{\sqrt{4500}\sqrt{11024}} = 0.985$$

1.1.2 Feature Selection – Métodos Filter Correlação

Correlação Spearman (variável independente e variável alvo numéricas ou categóricas ordinais)

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i corresponde à diferença no ranking entre as 2 variáveis para a observação i .

O resultado varia entre 1 e -1:

- Se > 0 , então a relação entre as 2 variáveis é positiva;
- Se < 0 , então a relação entre as 2 variáveis é negativa;
- Quanto mais perto dos extremos, mais forte é a relação.
- Ao contrário da correlação pearson, não assumimos que a relação entre as 2 variáveis é linear.

1.1.2 Feature Selection – Métodos Filter Correlação

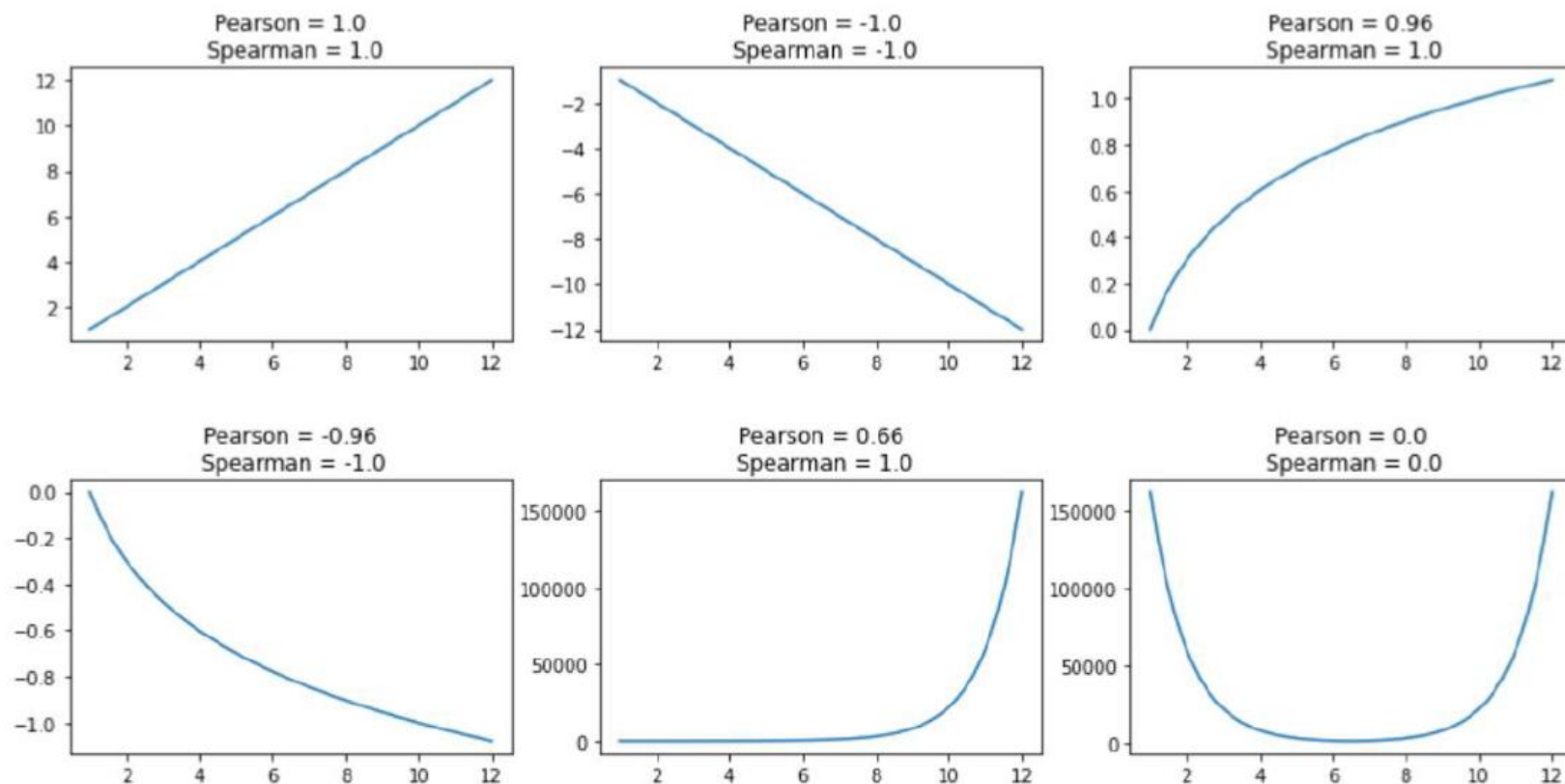
Correlação Spearman (variável independente e variável alvo numéricas ou categóricas ordinais)

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

ID	x	y	Rank x	Rank y	d_i^2
1	87	241	5	6	1
2	46	160	1	1	0
3	72	210	4	4	0
4	64	195	3	3	0
5	123	285	7	7	0
6	97	233	6	5	1
7	50	174	2	2	0
Sum	NA	NA	NA	NA	2

$$\rho = 1 - \frac{6 \times 2}{7(7^2 - 1)} = 0.964$$

1.1.2 Feature Selection – Métodos Filter Correlação



1.1.2 Feature Selection – Métodos Filter Correlação

Correlação Kendall (variável independente e variável alvo numéricas ou categóricas ordinais)

$$\tau = \frac{2(n_c - n_d)}{n(n-1)}$$

n_c corresponde ao número de pares de observações concordantes, e n_d ao número de pares discordantes.

A correlação de Kendall é um coeficiente que representa o **grau de concordância entre 2 variáveis**:

- Um par de observações é concordante se $(X_i > X_j \text{ e } Y_i > Y_j)$ ou $(X_i < X_j \text{ e } Y_i < Y_j)$
- Um par de observações é discordante se $(X_i > X_j \text{ e } Y_i < Y_j)$ ou $(X_i < X_j \text{ e } Y_i > Y_j)$
- Tal como o Pearson e Spearman, o coeficiente varia entre -1 e 1.

ID	x	y
1	87	210
2	46	160
3	72	241

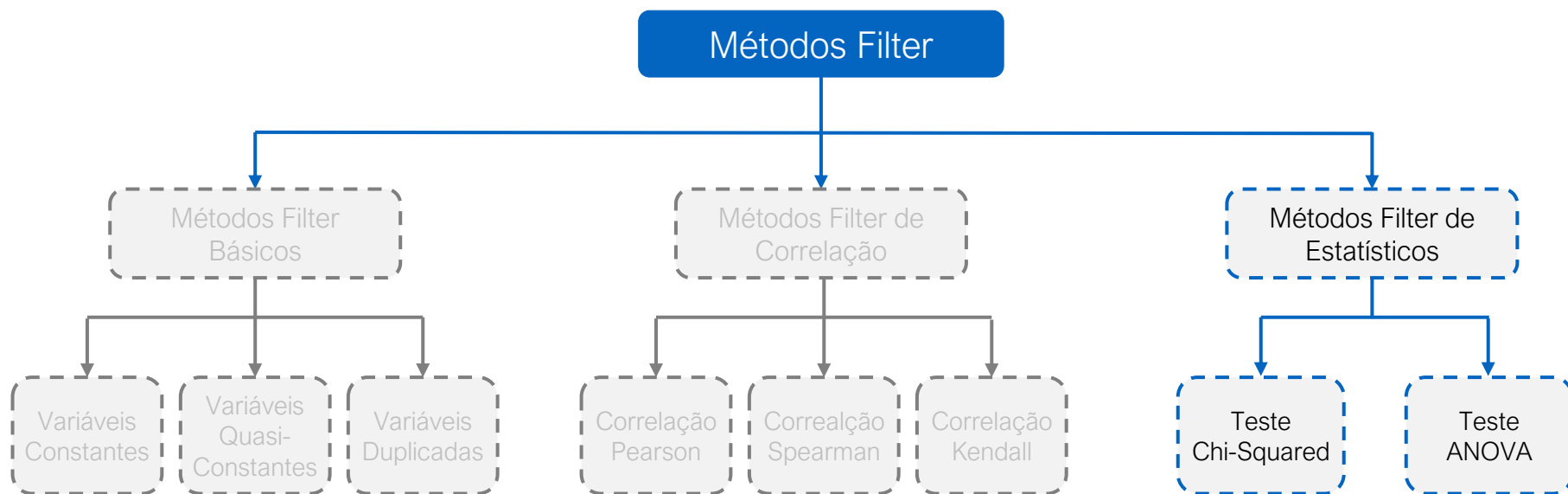
$(X_1 > X_2 \text{ e } Y_1 > Y_2) \rightarrow$ par concordante

$(X_1 > X_3 \text{ e } Y_1 < Y_3) \rightarrow$ par discordante

$(X_2 < X_3 \text{ e } Y_2 < Y_3) \rightarrow$ par concordante

$$\tau = \frac{2(2 - 1)}{3(3 - 1)} = 0.11$$

1.1.3 Feature Selection – Métodos Filter Estatísticos



1.1.3 Feature Selection – Métodos Filter Estatísticos

Teste Chi-Square (variável independente e variável alvo ambas categóricas)

O teste Chi-Square corresponde a um método estatístico utilizado para testar a independência de dois acontecimentos.

Neste caso, pretendemos determinar a relação entre um preditor e a variável alvo:

- Dadas 2 variáveis, podemos obter a **contagem observada “O”** e a **contagem esperada “E”**, sendo que a contagem esperada é calculada assumindo independência.
- Quando 2 variáveis são independentes, a contagem observada é próxima da contagem esperada, pelo que teremos um valor de Chi-Squared menor.
- Um valor **Chi-Squared mais elevado implica que a variável independente é mais dependente do alvo**, e pode ser selecionada para o treino do modelo.

1.1.3 Feature Selection – Métodos Filter Estatísticos

Teste Chi-Square (variável independente e variável alvo ambas categóricas)

1º Passo: Definir hipóteses

- Hipótese nula (H0): As duas variáveis são independentes
- Hipótese alternativa (H1): As duas variáveis não são independentes

2º Passo: Definir a tabela de contingência para os valores observados

Gender	Disease
F	Yes
M	No
F	No
...	...
M	Yes



		Disease		
		Yes	No	Total
Gender	F	380	1780	2160
	M	440	1400	1840
	Total	820	3180	4000



Mostra a distribuição de uma variável em linhas e outra em colunas. Utilizada para estudar a relação entre duas variáveis.

$$\text{Degrees of freedom} = (\text{rows} - 1) \times (\text{cols} - 1) = 1$$

1.1.3 Feature Selection – Métodos Filter Estatísticos

Teste Chi-Square (variável independente e variável alvo ambas categóricas)

3º Passo: Calcular os valores esperados usando probabilidades (assumindo independência)

- $P(Yes \cap Male) = P(Yes) \times P(Male) = \frac{820}{4000} \times \frac{1840}{4000} = 0.0943 \rightarrow 4000 \times 0.0943 = 377$
- $P(Yes \cap Female) = P(Yes) \times P(Female) = \frac{820}{4000} \times \frac{2160}{4000} = 0.1107 \rightarrow 4000 \times 0.1107 = 443$
- $P(No \cap Male) = P(No) \times P(Male) = \frac{3180}{4000} \times \frac{1840}{4000} = 0.3657 \rightarrow 4000 \times 0.3657 = 1463$
- $P(No \cap Female) = P(No) \times P(Female) = \frac{3180}{4000} \times \frac{2160}{4000} = 0.4293 \rightarrow 4000 \times 0.4293 = 1717$

Valores esperados

		Disease		
		Yes	No	Total
Gender	F	443	1717	2160
	M	377	1463	1840
	Total	820	3180	4000

Valores observados

		Disease		
		Yes	No	Total
Gender	F	380	1780	2160
	M	440	1400	1840
	Total	820	3180	4000

1.1.3 Feature Selection – Métodos Filter Estatísticos

[Teste Chi-Square](#) (variável independente e variável alvo ambas categóricas)

4º Passo: Calcular o valor do teste Chi-Squared

$$\chi^2 = \sum \frac{(\text{Observed Value} - \text{Expected Value})^2}{\text{Expected Value}}$$

Gender, Disease	<i>O</i>	<i>E</i>	<i>(O – E)</i>	<i>(O – E)²</i>	$\frac{(O - E)^2}{E}$
Male, Yes	440	377	63	3969	10.53
Male, No	1400	1463	-63	3969	2.71
Female, Yes	380	443	-63	3969	8.96
Female, No	1780	1717	63	3969	2.31
Teste Chi-Squared	NA	NA	NA	NA	24.51

1.1.3 Feature Selection – Métodos Filter Estatísticos

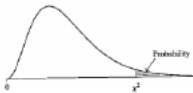
Teste Chi-Square (variável independente e variável alvo ambas categóricas)

5º Passo: Aceitar ou rejeitar a hipótese nula

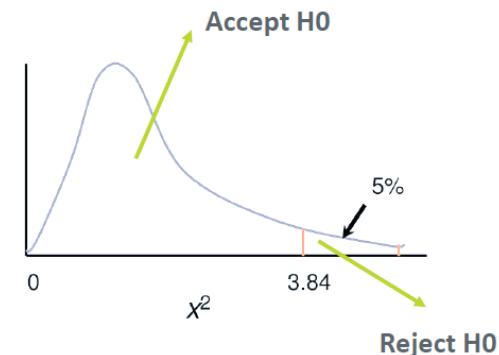
- Testar a hipótese nula de independência com 95% de confiança;
- Degrees of freedom (df) = 2.



TABLE c: Chi-Squared Distribution Values for Various Right-Tail Probabilities



df	0.250	0.100	0.050	0.025	0.010	0.005	0.001
1	1.32	2.71	3.84	5.02	6.63	7.88	10.83
2	2.77	4.61	5.99	7.38	9.21	10.60	13.82
3	4.11	6.25	7.82	9.35	11.34	12.84	16.27
4	5.39	7.78	9.49	11.14	13.28	14.86	18.47
5	6.63	9.24	11.07	12.83	15.09	16.75	20.52
6	7.88	10.64	12.59	14.45	16.81	18.55	22.46
7	9.04	12.02	14.07	16.01	18.48	20.28	24.32
8	10.22	13.36	15.51	17.53	20.09	21.96	26.12
9	11.39	14.68	16.92	19.02	21.67	23.59	27.88
10	12.55	15.99	18.31	20.48	23.21	25.19	29.59
11	13.70	17.28	19.68	21.90	24.72	26.76	31.26
12	14.85	18.55	21.03	23.34	26.22	28.30	32.91
13	15.98	19.81	22.36	24.74	27.69	29.82	34.53
14	17.12	21.06	23.68	26.12	29.14	31.32	36.12
15	18.25	22.31	25.00	27.49	30.58	32.80	37.70
16	19.37	23.54	26.30	28.85	32.00	34.27	39.25
17	20.49	24.77	27.59	30.19	33.41	35.72	40.79
18	21.60	25.99	28.87	31.53	34.81	37.16	42.31
19	22.72	27.20	30.14	32.85	36.19	38.58	43.82
20	23.83	28.41	31.41	34.17	37.57	40.00	45.32
25	29.84	34.38	37.65	40.65	44.31	46.92	52.62
30	34.80	40.26	43.77	46.98	50.89	53.67	59.70
40	43.82	51.80	55.76	59.34	63.69	66.77	73.40
50	56.33	63.17	67.50	71.42	75.15	79.49	86.56
60	66.98	74.40	79.08	82.29	86.58	91.95	99.61
70	77.58	85.53	90.53	93.02	100.4	106.2	114.3
80	88.13	96.58	101.8	104.6	112.3	118.3	126.4
90	98.65	107.6	113.1	116.1	124.1	128.3	137.2
100	109.1	118.5	124.3	129.6	135.8	140.2	148.5



Conclusão: uma vez que o valor do teste Chi-Square (24.51) é superior ao valor crítico da distribuição Chi-Squared para 95% de confiança e $df = 1$ (3.81), rejeitamos a hipótese nula e usamos a variável no modelo.

1.1.3 Feature Selection – Métodos Filter Estatísticos

Teste ANOVA (variável independente numérica e variável alvo categórica)

- Método estatístico utilizado para **verificar se as médias de uma variável independente para cada categoria da variável alvo são significativamente diferentes entre si.**

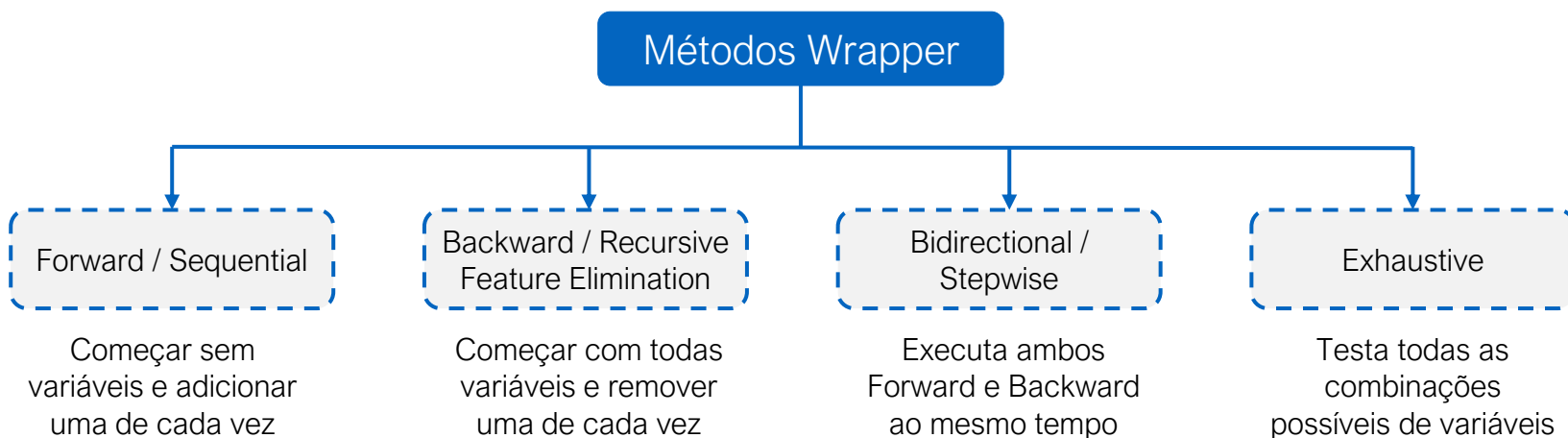
H0: todas as categorias têm a mesma média

H1: pelo menos 1 das categorias distingue-se das outras

- **Se a H0 for rejeitada**, isso significa que existe uma **variância entre os grupos que indica que a variável independente tem impacto na variável alvo**, pelo que devemos incluir essa variável no modelo.

1.2 Feature Selection – Métodos Wrapper

Os métodos Wrapper utilizam algoritmos para selecionar as variáveis. Uma estratégia de pesquisa é processada através do espaço de características possíveis: cada subconjunto é avaliado com base na qualidade do desempenho de um determinado algoritmo.



1.2 Feature Selection – Métodos Wrapper

[Forward / Sequential](#)

Passo 1: Avaliar todas as variáveis individualmente;

Passo 2: Selecionar a que resulta no melhor desempenho do modelo;

Passo 3: Testar todas as combinações possíveis da variáveis selecionada com cada uma das restantes variáveis;

Passo 4: Adicionar a variável cuja combinação produz o melhor desempenho do modelo;

Passo 5: Continuar o ciclo, adicionando uma variável de cada vez em cada iteração até que o critério pré-definido seja atingido.

1.2 Feature Selection – Métodos Wrapper

[Backward / Recursive Feature Elimination \(RFE\)](#)

Passo 1: Começar com todas as variáveis do nosso conjunto de dados;

Passo 2: Avaliar o desempenho do algoritmo;

Passo 3: Remover uma variável de cada vez e avaliar o desempenho do modelo;

Passo 4: Remover permanentemente a variável que menos afeta os resultados do modelo de entre as restantes disponíveis;

Passo 5: Continuar o ciclo, removendo uma variável de cada vez em cada iteração até que o critério pré-definido seja atingido.

1.2 Feature Selection – Métodos Wrapper

Bidirectional / Stepwise

Passo 1: Começar sem variáveis e avaliar todas as características individualmente;

Passo 2: Selecionar a que resulta no melhor desempenho do modelo;

Passo 3: Testar todas as combinações possíveis da variáveis selecionada com cada uma das restantes variáveis;

Passo 4: Adicionar a variável cuja combinação produz o melhor desempenho do modelo;

Passo 5: Verificar se a remoção de alguma das variáveis aumenta o desempenho do modelo, ou seja, se a significância de uma variável foi reduzida para um nível de tolerância pré-definido.

Se for encontrada uma variável não significativa, removê-la do modelo e continuar para o passo 3. Caso contrário, passar diretamente para o passo 3. Parar quando o critério pré-definido for atingido.

1.2 Feature Selection – Métodos Wrapper

Exhaustive

Passo 1: Criar todas as combinações de variáveis possíveis;

Passo 2: Para cada subconjunto, construir um modelo;

Passo 3: Selecionar o subconjunto cujo o modelo teve o melhor desempenho;

Apesar da solução deste método ser mais significativo para o nosso trabalho, tem um grande custo computacional, especialmente quando o número de variáveis é elevado.

1.3 Feature Selection – Métodos Embedded

Os métodos Embedded são técnicas utilizadas para selecionar as variáveis adequadas para o nosso modelo [durante o treino do modelo](#).

O tipo [mais comum são os métodos de regularização](#), também designados por métodos de penalização:

- Introduzem restrições adicionais na otimização de um algoritmo de previsão;
- Inclina o modelo para uma menor complexidade (menos coeficientes);
- Exemplos: Lasso, ElasticNet e Ridge Regression.

Outros algoritmos permitem-nos [obter a “feature importance”](#):

- Define quais as variáveis mais importantes para fazer previsões exactas sobre o alvo;
- Exemplo: Decision Trees e Random Forests.

1.3 Feature Selection – Métodos Embedded

Algumas das principais vantagens são que:

- Os métodos Embedded **tomam em consideração a interação das variáveis** como os métodos Wrapper;
- **Mais rápidos** do que os métodos Wrapper;
- **Mais exactos** do que os métodos Filter;
- Encontram as variáveis com mais propensão a aumentar o desempenho do algoritmo que está a ser treinado.

1.4 Feature Selection - Conclusão

Como entendemos, existem muitos métodos... Cada um pode apresentar resultados diferentes! Como devemos proceder?

1.4 Feature Selection - Conclusão

Como entendemos, existem muitos métodos... Cada um pode apresentar resultados diferentes! Como devemos proceder?

Uma abordagem para identificar as variáveis a manter no modelo é **aplicar diferentes métodos, e combinar os seus resultados**:

Variável	Corr. Spearman (Filter)	RFE (Wrapper)	Lasso (Embedded)	Decision Tree (Embedded)	Contagem	Decisão
Variável 1	Sim	Sim	Sim	Sim	4	Manter
Variável 2	Sim	Sim	Sim	Sim	4	Manter
Variável 3	Sim	Sim	Não	Não	2	Testar com e sem
Variável 4	Não	Não	Sim	Não	1	Remover
Variável 5	Não	Não	Não	Não	0	Remover

Obrigado!