Feature Selection

Seleção de Recursos

**Sumário**

[1. Introdução 3](#_Toc80711665)

[2. Estudos dos livvros 3](#_Toc80711666)

[3. Estudar e analisar bigrama e trigrama 3](#_Toc80711667)

[4. Estudar a biblioteca Gensim com Python 3](#_Toc80711668)

[5. Testar exemplos 3](#_Toc80711669)

[6. Resumo 4](#_Toc80711670)

[7. Bibliografia 5](#_Toc80711671)

[8. Referências 5](#_Toc80711672)

# **Introdução**

xxx

Um Sistema de recuperação de Informação (SRI) deve dentre outras funções ter uma base (corpus) com “todos” os documentos que se deseja trabalhar e deve possuir alguns rótulos específicos.

# **Estudos dos livvros**

**Livro:** Manning, Raghavan, Schutze - Introduction to information retrieval (2008), páginas 251 a 258 (275-)

**Livro:** Aggarwal - Machine Learning for Text (2018), páginas 117 a 122

# **Estudar e analisar bigrama e trigrama**

**xxx**

# **Estudar a biblioteca Gensim com Python**

**xxx**

# **Testar exemplos**

from gensim.models.phrases

from gensim.models.phrases import Phrases

modelo\_fraseador = Phrases(frases, min\_count = min\_c, threshold = threshold,

scoring = scoring)

min\_c = 120, threshold = 0.6, scoring = 'npmi'

trabalhar com notícias de 2016 a 2021

[segunda-feira 15:53] Frederico Shu

Depois de treinar o modelo fraseador:

fraseador = Phraser(modelo\_fraseador) # exportar e fechar o modelo para economizar RAM

fraseador.save(arqfr) # gravar o modelo treinado

fraseador = Phraser.load(arqfr) # carregar fraseador

documento\_analisado = fraseador[documento]

A ideia então seria implementar o fraseador e estudar seleção de atributos

# **Resumo**

Porque fazer Seleção de Features?

1. Tornar a interpretação do fenômeno mais simples.

Menor quantidade de atributos possíveis para descrever um carro por exemplo

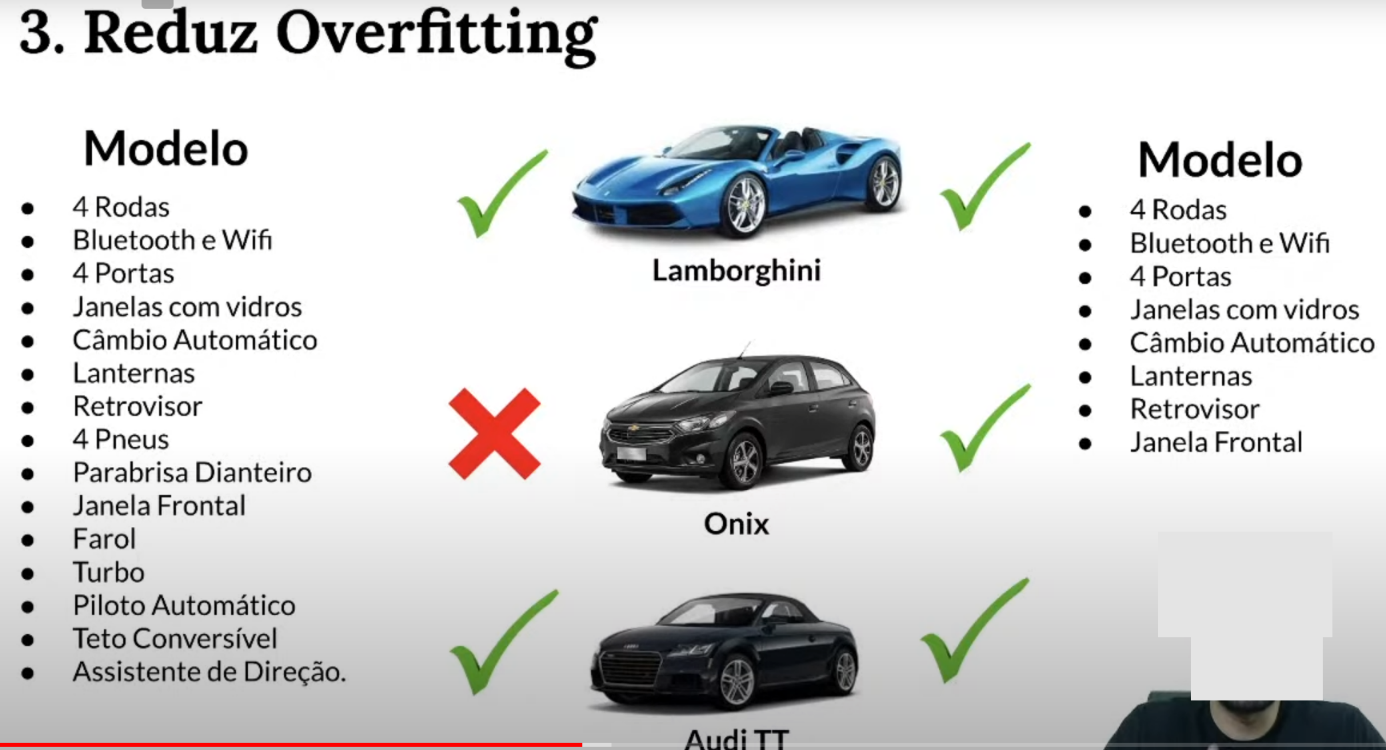


1. Reduzir o tempo de treinamento do modelo
2. Evitar a maldição da dimensionalidade

Quanto mais enxuto o modelo, menos esparso fica dos dados e não se perde o fenômeno.

1. Reduzir a chance de Overfitting (super aprendizado de um fenômeno)

Ter um modelo mais simples, poucas features, pois quanto mais detalhes, determinados carros mais simples não conseguiram ser identificados



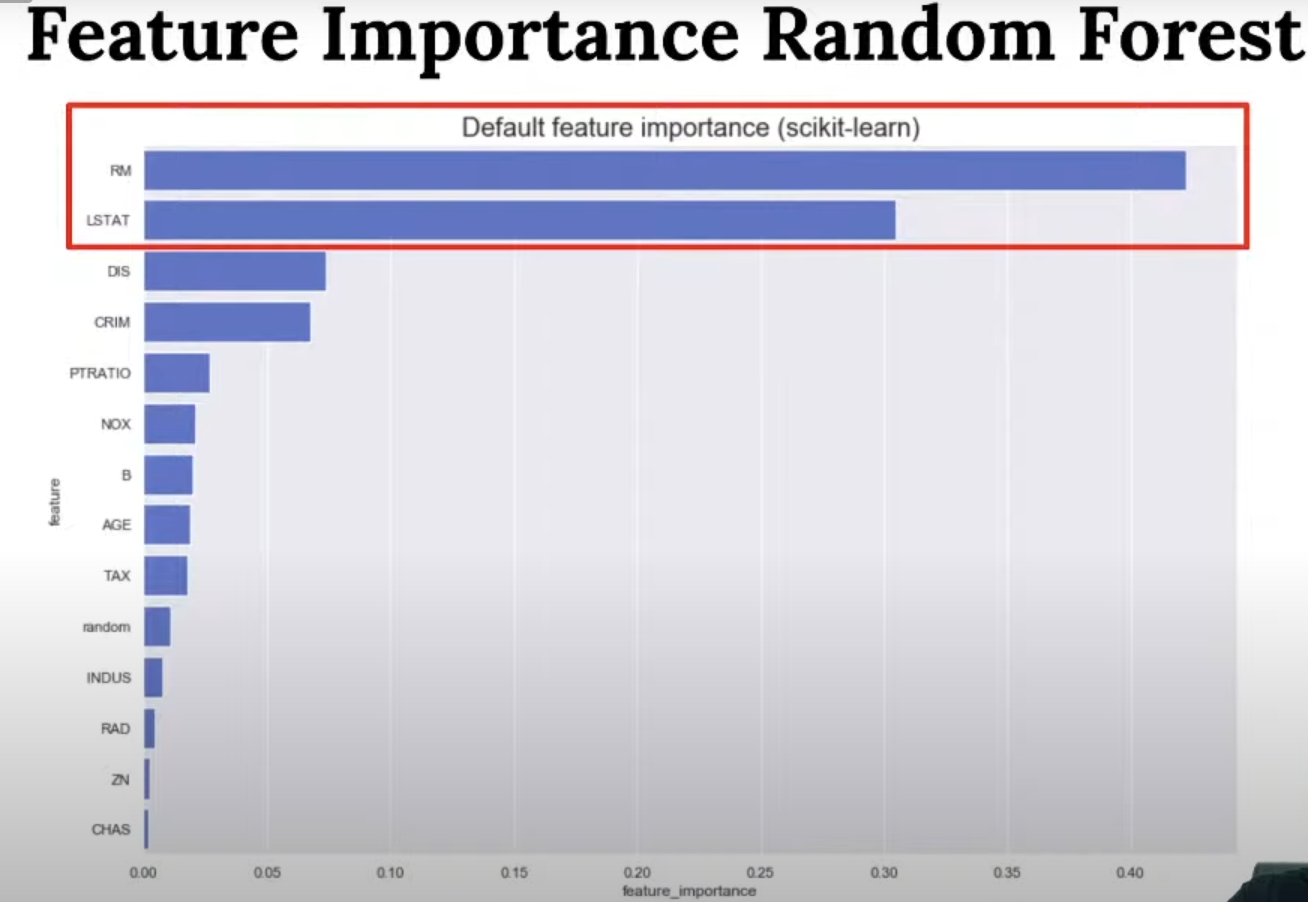
“A explição mais simples sobre um fenômeno observado, deveria prevalecer sobre explicações mais complexas.” (Occam`s Razor)

As 4 formas de fazer seleção de features

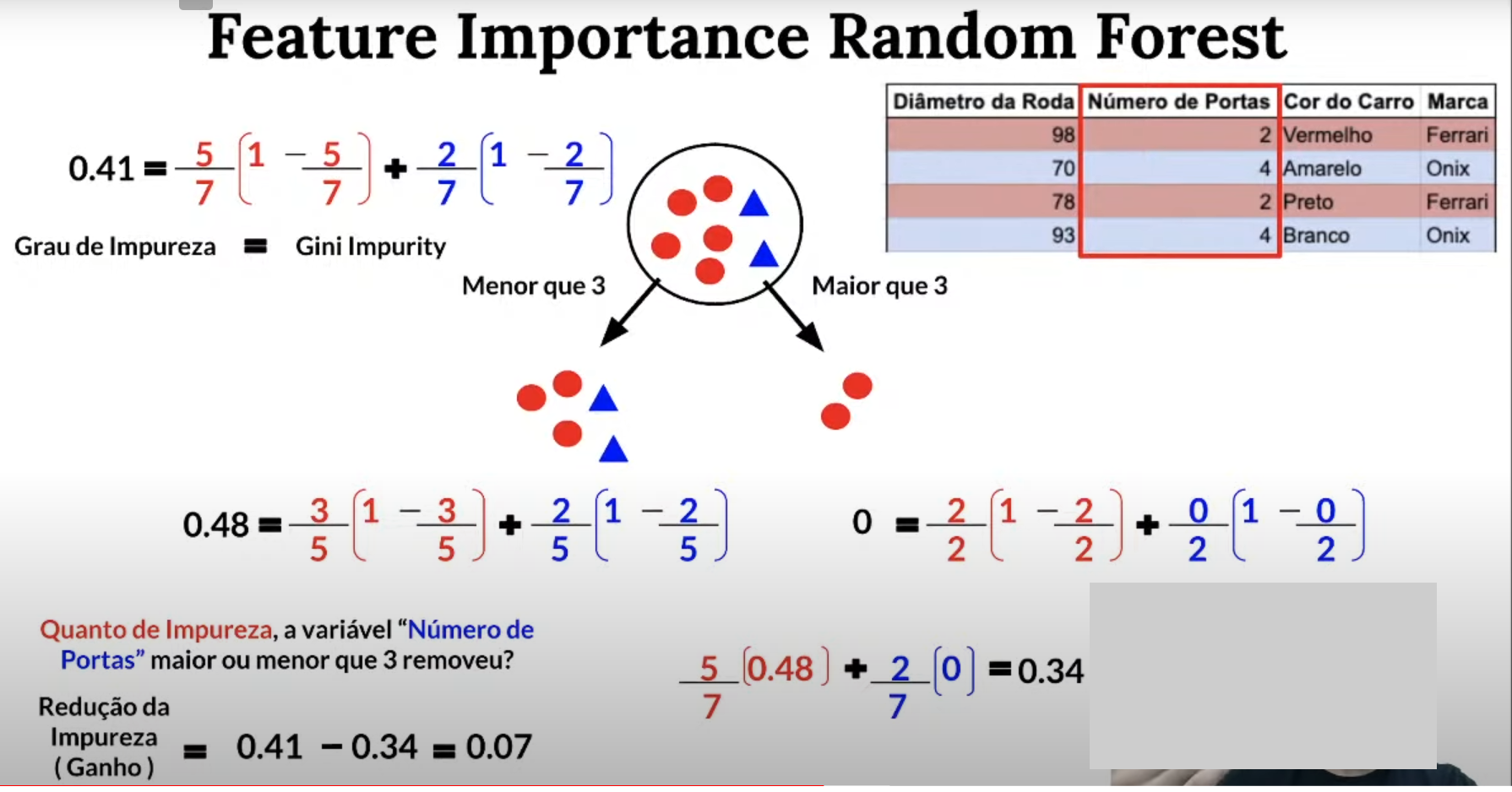
1. Importância das features

Geralmente feito com algoritmos de arvores

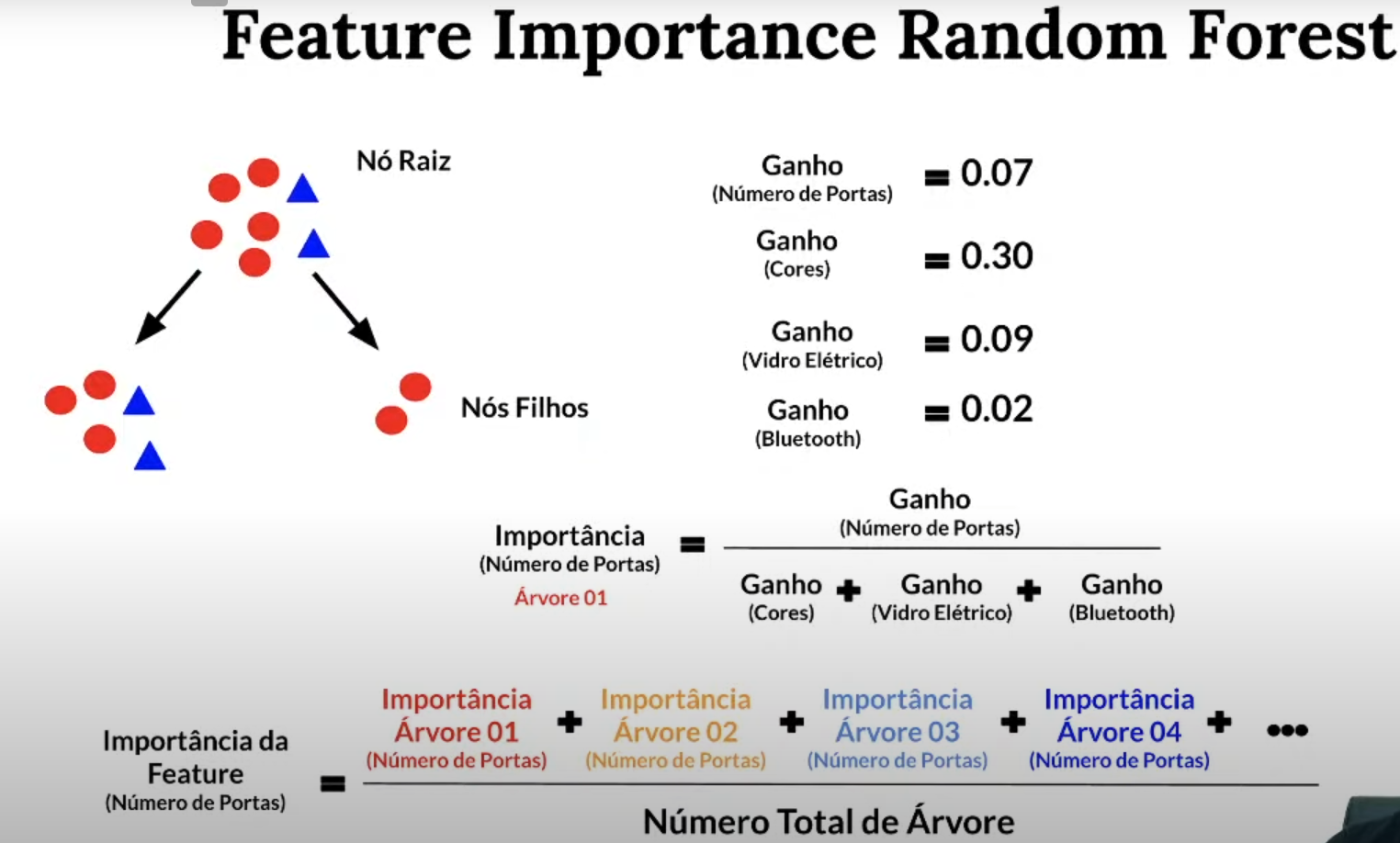
Abaixo gráfico com as features e a importância delas



Com a execução destes gráficos, olhando, manualmente se escolhe as “feature\_importance”. Este calculo é realizado de cada atributo separando nos nós das árvores.

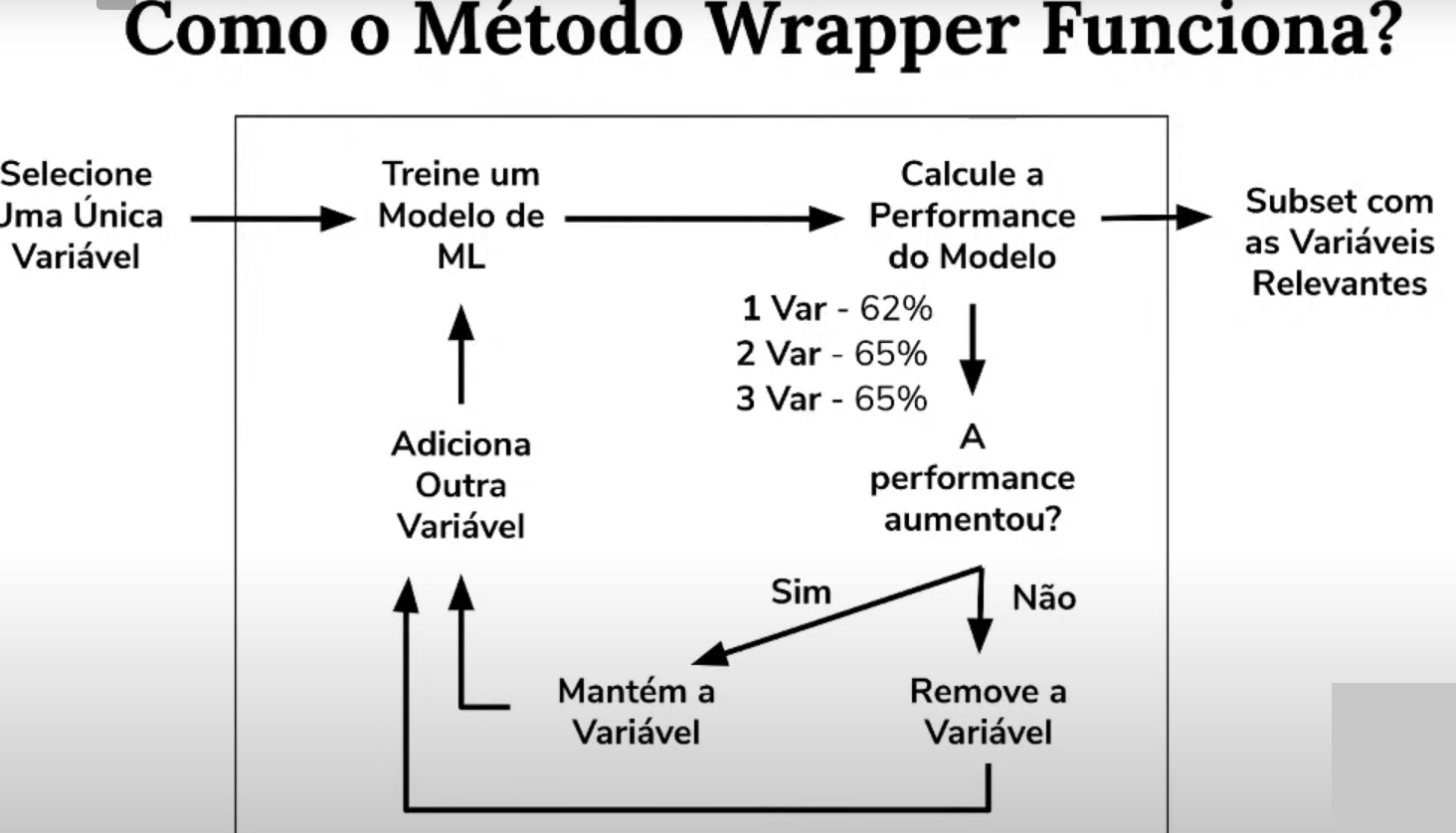


Identificando com a Random Forest



O resultado da figura acima, representa o gráfico acima do sklearn, que é o número que aparece em baixo, da importância das arvores.

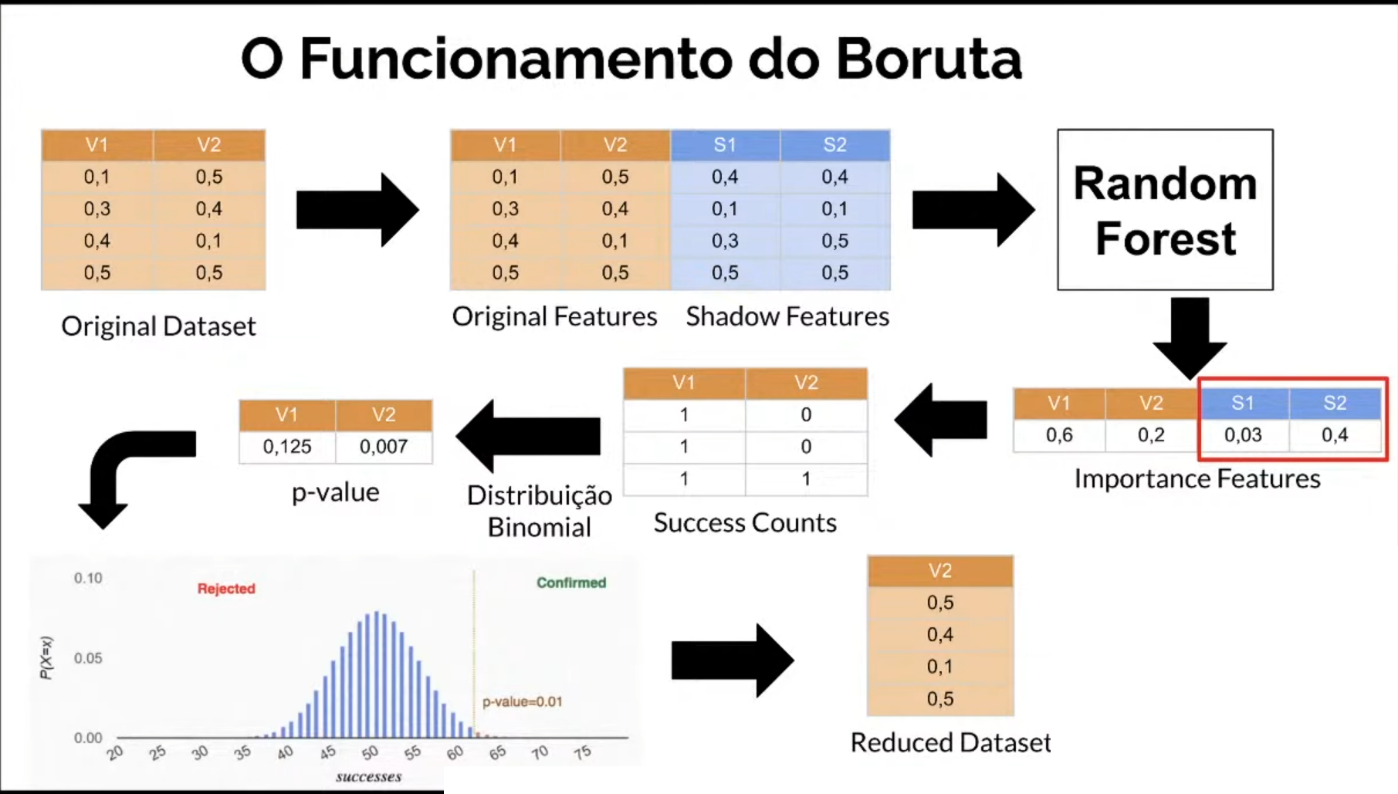
1. Métodos de filtragem
2. Métodos Wrappers (empacotar)



O gráfico acima é implementado pelo algoritmo Boruta (método warapper).

A variável copiada (shadow) é embaralhada para reduzir a correlação, usa modelo de árvore, se dataset grande usar o exaTrees (mais rápido), senão o random forest.

Com o Boruta não precisa escolher as features manualmente (visualmente) ele já decide, porém ele demora mais para rodar.

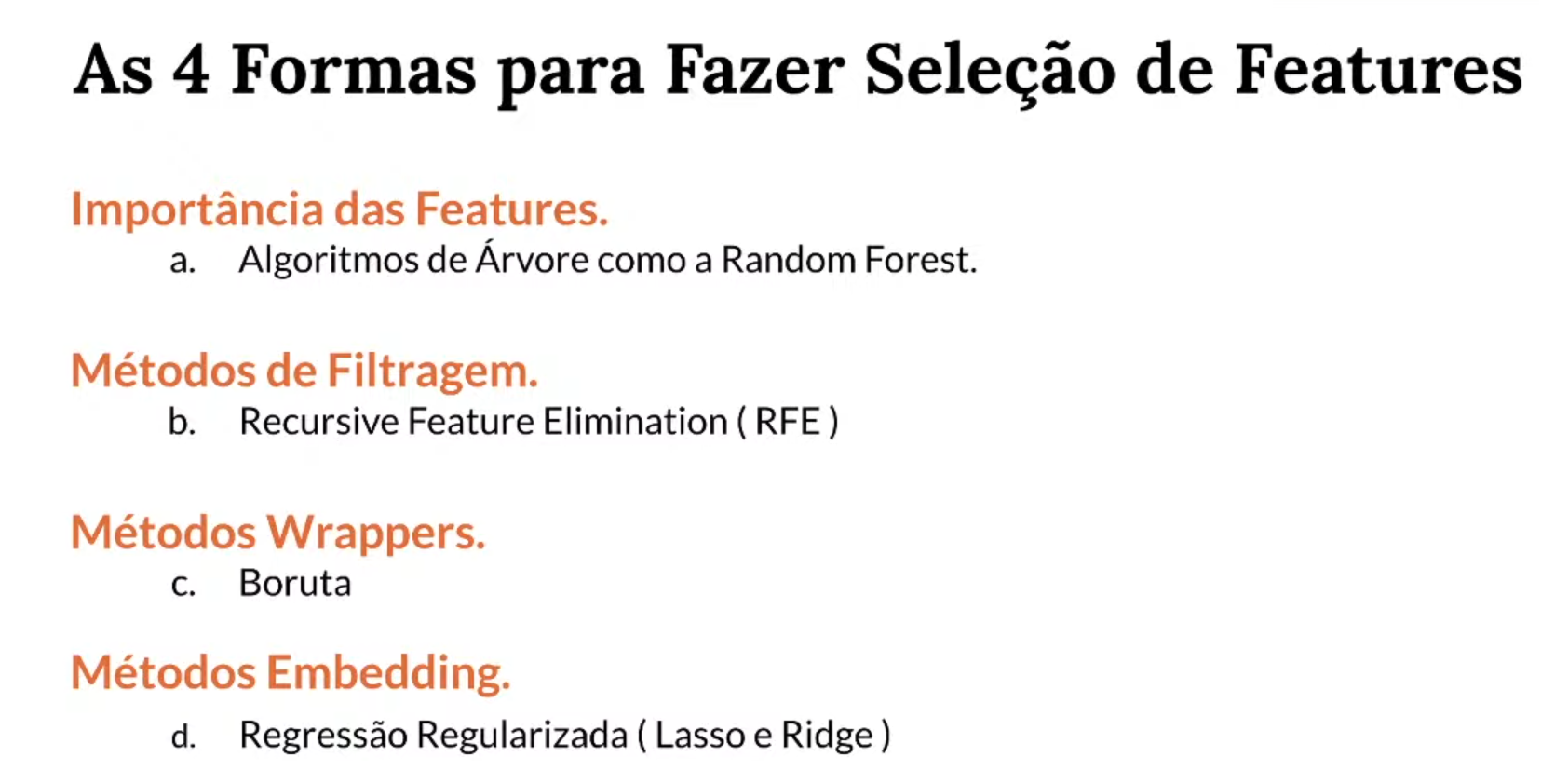


1. Métodos Embedding

Logo:



--



# **Bibliografia**

**SRI** - Sistema de Recuperação da Informação

**Feature** – Para este caso é sinônimo de atributo, varável, coluna e recurso

**Dimensionalidade** –

1 dimensão - pode-se dizer de uma reta, ou seja, um e apenas um atributo em um plano (Ex: atributo idade)

2 dimensões – um plano x e y (Ex: atributos idade e salário)

A cada variável adicionada, aumenta a dimensão, o que pode deixar os dados bem esparsos. O ideal é em se aumetando os atributos, deve-se aumentar os registros (linha) em grande quantidade.

**Esparsidade** – quanto mais se coloca atributos/coluna no plano, menor fica a densidade dos dados, logo, mais esparsos/distantes fica os dados, com isso tem-se a dissolução do fenômeno.

**Densidade** – acúmulo de valores em uma certa posição em uma dimensão. 

**Dissimilaridade dos dados** –

Entropia -

**Fenômeno** –

Gini impurity / Grau de impureza -

# **Referências**

<https://www.amazon.com.br/Introduction-Information-Retrieval-English-Christopher-ebook/dp/B00AHTN5JM>

ver <https://www.youtube.com/watch?v=0LyIGw8MToY&t=3270s>

<https://youtu.be/Bcn5e7LYMhg>

Introdução à Recuperação da Informação

* <https://youtu.be/lJVw0-qbqEM>
* <https://www.youtube.com/watch?v=uwW_NTRpzC8>
* <https://www.youtube.com/watch?v=qWZbHr_GJgE>

ver COMO REALIZAR A SELEÇÃO DAS MELHORES VARIÁVEIS PARA SUA MÁQUINA PREDITIVA - FEATURE SELECTION <https://youtu.be/0bPBxDhvqlI>

ver <https://www.youtube.com/watch?v=4RGT2YRHERY&t=1258s>

vre <https://www.youtube.com/watch?v=Gzn6srbzU30>