
# Seleção de Modelos de Aprendizado de Máquina

**Bootcamp Engenheiro(a) de Machine Learning**

Karen Enes

2021

# Seleção de Modelos de Aprendizado de Máquina

Bootcamp Engenheiro(a) de Machine Learning

Karen Enes

## Sumário

Capítulo 1. Introdução a Seleção de Modelos ..... 7

Avaliação de modelos ..... 7

Case: Predição de Infecção por COVID-19 ..... 7

Métricas de Desempenho ..... 9

Técnicas de Validação ..... 9

Sintonia de Hiperparâmetros ..... 10

Conclusão ..... 10

Capítulo 2. Métricas de Desempenho para Regressão ..... 11

Case: Previsão do número de casos de COVID-19 por semana no Brasil. ..... 11

Métrica: $R^2$ (R-quadrado) ..... 12

Métrica: $R_a^2$ ($R^2$ ajustado) ..... 13

Mean Squared Error - MSE ..... 14

Erro médio quadrático ..... 14

Root Mean Squared Error - RMSE ..... 15

Raiz do Erro médio quadrático ..... 15

Root Mean Squared Logarithmic Error – RMSLE ..... 16

Raiz do Erro médio quadrático considerando o log ..... 16

Mean Absolute Error - MAE ..... 17

Erro médio absoluto ..... 17

Mean Absolute Percentage Error - MAPE ..... 17

Considerações Finais ..... 18

Capítulo 3. Métricas de Desempenho para Classificação ..... 19

Case: Predição de mortalidade de pacientes infectados por COVID-19..... 19

Matriz de Confusão ..... 20

Taxa de Erro ..... 20

Acurácia ..... 21

Falso/Verdadeiro Positivos e Negativos ..... 21

Precisão ..... 22

Revocação ..... 23

Recall, Sensitivity, True Positive Rate..... 23

Precisão x Revocação ..... 23

F-measure..... 24

F-score, F1, Medida F ..... 24

Macro e Micro F1 ..... 24

Curva ROC ..... 25

Receiver Operating Characteristic Curve - ROC Curve ..... 25

AUC – Area Under the ROC Curve..... 25

Área abaixo da curva ROC ..... 25

Conclusão ..... 26

Capítulo 4. Métricas de Desempenho para Clusterização ..... 27
Modelos de Clusterização ..... 27
Previsão de chegada das novas vacinas para COVID-19 ..... 27
Validação de modelos não supervisionados ..... 28
Medidas Internas ..... 29
Medidas Externas ..... 29
Coeficiente de Silhueta ..... 30
Coeficiente de Silhouette, método da Silhueta, Silhouette Coefficient ..... 30
Pureza ..... 31
Purity ..... 31
Índice de Jaccard ..... 32
Jaccard Index ..... 32
Conclusão ..... 32
Capítulo 5. Técnicas de Validação ..... 33
Etapas de desenvolvimento de Modelos ..... 33
Divisão Treino e Teste ..... 34
Train-Test Split, Hold-out validation e Validação Simples ..... 34
Validação Cruzada ..... 35
Variação: Stratified k-fold Cross Validation ..... 37
Leave-one-out ..... 38

Considerações Finais ..... 39

Capítulo 6. Sintonia de Hiperparâmetros ..... 41

Case: Predição de mortalidade em pacientes infectados com COVID-19 ..... 41

Feature Engineering, Feature Selection and Modelo Tuning ..... 41

Sintonia de Hiperparâmetros ..... 42

Sintonia de Hiperparâmetros, Otimização de Hiperparâmetros, Combinação de Hiperparâmetros e Model Tuning ..... 42

Tipos de Hiperparâmetros ..... 42

Otimizando Hiperparâmetros ..... 43

Método Força Bruta ..... 44

Método Random Search ..... 44

Método Grid Search ..... 45

Referências ..... 47

# Capítulo 1. Introdução a Seleção de Modelos

Durante esse módulo, falaremos sobre várias métricas e técnicas que nos auxiliam no processo de avaliação e seleção de modelos de aprendizado de máquina. Nesse capítulo vamos começar a conhecer os conceitos fundamentais para selecionar os melhores modelos. Discutiremos o porquê de selecionar e avaliar, e além disso, também falaremos sobre algumas perguntas básicas que essas métricas e técnicas são capazes de responder.

## Avaliação de modelos

Por que avaliar modelos?

- Assim como quando vamos ao supermercado e escolhemos a marca ou o tipo do arroz que melhor atende as nossas necessidades, o mesmo ocorre com modelos de aprendizado.
- Precisamos saber se o modelo proposto é o ideal.

Como saber se o modelo proposto é o ideal é o principal objetivo dessa disciplina. Ao final do modulo, responderemos perguntas como:

- Como avaliar o desempenho de um modelo?
- Como obter estimativas confiáveis?
- Como comparar o desempenho relative entre diferentes modelos?

## Case: Predição de Infecção por COVID-19

Laura e João Pedro são colaboradores de duas empresas distintas e trabalham no desenvolvimento de novos modelos.

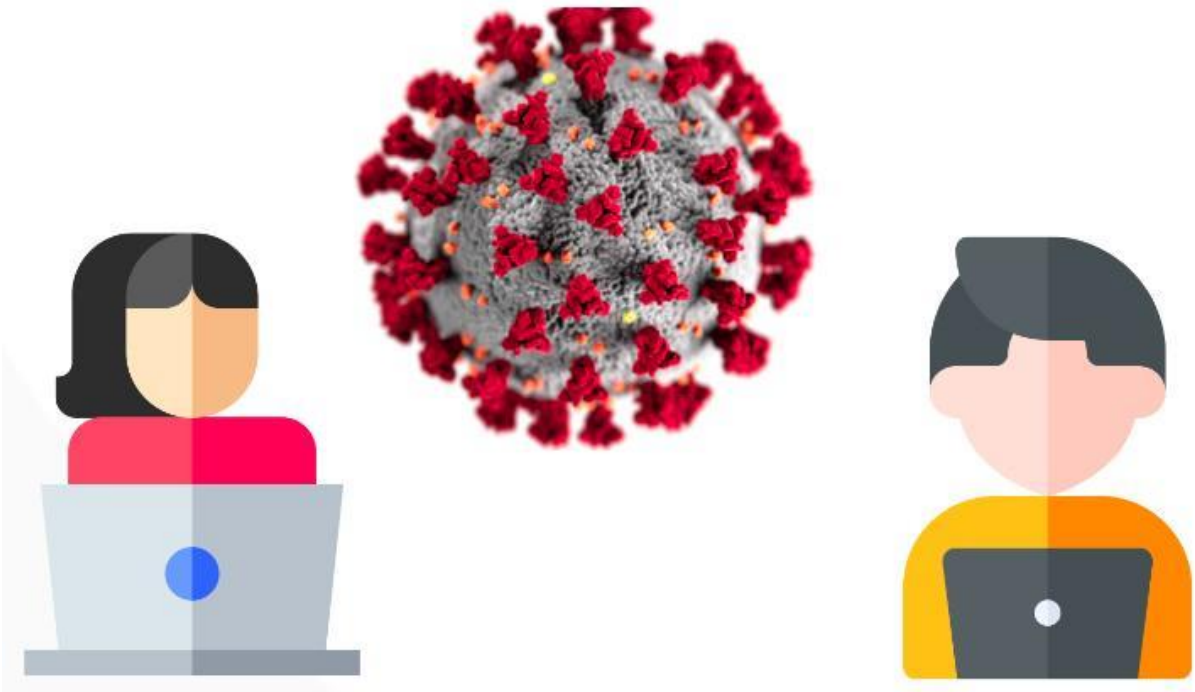
– Eles receberam o mesmo conjunto de dados com informações sobre diversas pessoas que foram infectadas com COVID-19.

– Essas informações podem incluir dados como idade, gênero, altura, peso, comorbidades existentes, local de moradia, cidade, tipo de moradia e acesso a planos de saúde.

– Laura e João tem a missão de implementar um modelo que seja capaz de predizer a chance de infecção de uma nova pessoa não constante da base de dados.

– **Como saber qual o modelo é melhor? O de Laura ou o de João?**

**Figura 1 – Ilustração: Predição de Infecção por COVID-19.**



**Fonte: Ilustração do Coronavírus criada pelo Centro de Controle e Prevenção de Doenças. Ícones programadores -> Flaticon.**

## Métricas de Desempenho

- Analisam a qualidade do modelo desenvolvido.
- Indicam se o modelo está apresentando bons resultados para o conjunto de dados analisado.
- Mostram indícios de como o modelo está funcionando.
- Indica onde podemos agir para melhorar o modelo.
- Variam de acordo com a tarefa realizada e o tipo de aprendizado do modelo.
    - Para o aprendizado supervisionado, nas tarefas de regressão e classificação, as métricas de desempenho calculam medidas de erro do modelo.
- No caso do aprendizado não supervisionado, na tarefa de clusterização, as métricas de desempenho avaliam a qualidade interna e externa dos clusters formados, já que não há informações sobre rótulo das amostras.

## Técnicas de Validação

- As técnicas de validação são responsáveis por medir a capacidade de generalização do modelo.
- Durante a etapa de treinamento, as técnicas de validação auxiliam na construção de um conjunto de validação com amostras selecionadas de forma inteligente, a partir do conjunto de treinamento. O conjunto de validação formado é responsável por simular a etapa de teste do modelo ainda durante a fase de treinamento.
- O objetivo principal dessas técnicas é estimar o quão preciso o modelo é na prática.

## Sintonia de Hiperparâmetros

- Refinamento do modelo para melhoria das métricas.
- Métodos de sintonia de hiperparâmetros são responsáveis por testar diferentes combinações de parâmetros dos modelos de aprendizado.
- Melhores combinações de parâmetros refletem em uma melhor capacidade de generalização e, consequentemente, em um modelo de maior qualidade.

## Conclusão

Em todos os cenários, é preciso conhecer as capacidades e as limitações dos modelos aplicados para a solução de um problema. Todo modelo requer uma calibragem e uma avaliação.

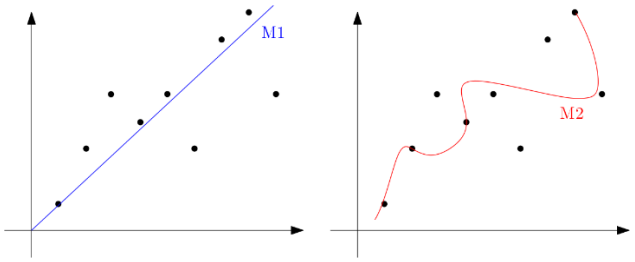
## Capítulo 2. Métricas de Desempenho para Regressão

Começaremos a discutir as métricas de desempenho de modelos de aprendizado supervisionado. Para esses modelos os dados são rotulados, isto é, existe a informação de classe para as amostras do conjunto de dados. Assim é possível medir o quanto um modelo acerta ou erra. Vamos começar falando de métricas de desempenho para a tarefa de regressão.

### Case: Previsão do número de casos de COVID-19 por semana no Brasil.

- Laura e Pedro tem mais uma missão pela frente!
  - Dessa vez eles receberam dados sobre fatores que levam a variações no número de casos semanais no Brasil.
  - A ideia é que eles criem um modelo capaz de prever o número de casos por semana.
  - Para isso, ambos implementam um modelo de regressão. Será que é possível identificar se um dos modelos é melhor do que o outro?

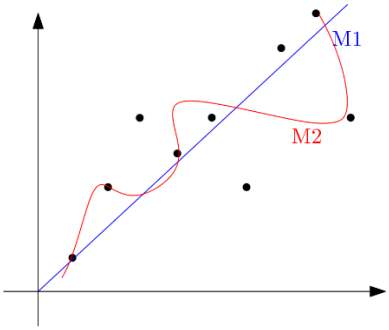
**Figura 2 –Previsão do número de casos de COVID-19 por semana no Brasil. Qual modelo prevê melhor o número de infectados?**



## Métrica: R² (R-quadrado)

- Também chamado de Coeficiênte de Determinação.
- Mede o quanto da variabilidade dos dados é explicado pelo modelo gerado.
  - Ex.: o modelo M2 explica 95% dos dados.
- **Na prática:** mede o quão melhor meu modelo é, em relação a uma regressão media.

**Figura 3 – Qual modelo prevê melhor o número de infectados? Sobreposição dos modelos M1 e M2.**

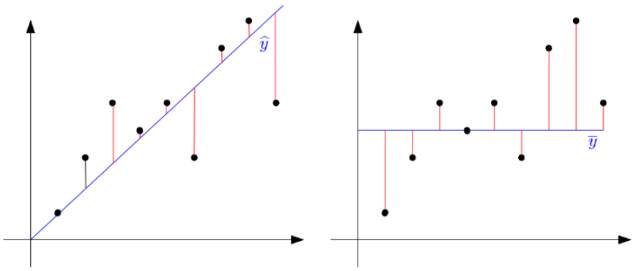



- Vantagens:
  - Métrica de fácil interpretação.
- Desvantagens:
  - Não consegue identificar viés nos dados.
  - Overfitting.
  - Modelos com apenas uma variável.

**Figura 4 – Fóruma R².**

$$R^2 = 1 - \frac{VarianciaResidual}{VarianciaTotal} = 1 - \frac{\Sigma_{i=1}^{n}(y_i - \widehat{y}_i)^2}{\Sigma_{i=1}^{n}(y_i - \overline{y})^2}$$

**Figura 5 – Interpretação $R^2$.**




## Métrica: $R_a^2$ ($R^2$ ajustado)

- Versão modificada para reduzir o viés criado com o acréscimo de features.
- Usado na avaliação de modelos com diferentes números de features.
- Na prática: calculamos a porcentagem de variação na resposta que é explicada pelo modelo, considerando o número de features em relação ao número de amostras.
- **Vantagens:**
  - Avaliação mais precisa dos modelos de regressão
  - Aplicável a modelos multivariados
- **Desvantagens:**
  - Mais dificíl de interpretar.

**Figura 6 – Fórmula $R_a^2$.**

$$R_a^2 = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

- $R^2$ e $Ra^2$ são amplamente utilizadas para avaliar a relação existente entre dois modelos.
- Usadas na avaliação de modelos mais simples, geralmente lineares.
- Existem outras métricas mais objetivas!

## Mean Squared Error - MSE

*Erro médio quadrático*

- Métrica muito utilizada para avaliar a qualidade de regressores.
- O valor mínimo é 0 e quanto maior o MSE, pior é o modelo, no cenário avaliado.
- Diferenças menores influenciam menos no cálculo do MSE, enquanto diferenças maiores influenciam mais.
- **Na prática:** é a diferença entre o valor predito e o valor real elevado ao quadrado para todas as amostras, dividido pelo número total de amostras.
- **Vantagens:**
  - Entendimento simples.
  - Ideal para modelos que não toleram erros de grandes proporções.
- **Desvantagens:**
  - Enviesada erros de grandes proporções.
  - Difícil interpretação.
  - Unidade de medida $u^2$.

**Figura 7 – Fórmula MSE.**

$$MSE = \frac{\Sigma_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}$$

## Root Mean Squared Error - RMSE

*Raiz do Erro médio quadrático*

- Amplamente utilizada! Métrica mais utilizada para avaliar a qualidade de regressores.
- Alternativa para melhorar a interpretabilidade do MSE.
- Resolve a questão das unidade de medida do MSE.
- Medida análoga ao desvio padrão obtido para médias
- **Na prática:** é a raiz quadrada da diferença entre o valor predito e o valor real elevado ao quadrado para todas as amostras, dividido pelo número total de amostras.
- **Vantagens:**
  - Simples de entender e interpretar.
  - Unidade de medida é a mesma da variável que se quer prever.
  - Ideal para modelos que não toleram erros de grandes proporções.
- **Desvantagens:**
  - Enviesada erros de grandes proporções.

**Figura 8 – Fórmula RMSE**

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\Sigma_{i=1}^{n}(\widehat{y_i} - y_i)^2}{n}}$$

## Root Mean Squared Logarithmic Error – RMSLE

*Raiz do Erro médio quadrático considerando o log*

- Evita a penalização das diferenças elevadas.
- Resolve a questão dos erros de grandes proporções encontrados no MSE e no RMSE.
- Bastante utilizada quando não há normalização dos dados.
- **Na prática:** a inclusão do log vai evitar os casos em que o Valor predito >> valor real e o Valor predito << valor real, que geram grande influência no cálculo do erro.
- **Vantagens:**
  - Menos sensível a ruídos nos dados.
  - Unidade de medida é a mesma da variável que se quer prever.
  - Pode ser usada para identificar se o conjunto de dados tem muitos outliers.

**Figura 9 – Fórmula RMSLE.**

$$RMSLE = \sqrt{\frac{\Sigma_{i=1}^{n}(log(y_i+1) - log(\widehat{y_i}+1))^2}{n}}$$

## Mean Absolute Error - MAE

*Erro médio absoluto*

- Métrica simples de fácil entendimento, muito utilizada na previsão de dados sazonais.
- O valor mínimo é 0 e quanto maior o erro, maior é o valor do MAE.
- Menos sensível a valores discrepantes quando comparado com MSE
- **Na prática:** é a média das distâncias entre o valores predito e real.
- **Vantagens:**
  - Entendimento simples e interpretação bem intuitiva.
  - Unidade de medida é a mesma do valor que se quer prever.
- **Desvantagens:**
  - Enviesada erros de grandes proporções

**Figura 10 – Fórmula MAE.**

$$MAE = \frac{\sum_{i=1}^{n} |\widehat{y_i} - y_i|}{n}$$

## Mean Absolute Percentage Error - MAPE

- Métrica amplamente utilizada devido a fácil interpretação resultados.
- Por se tratar de valores percentuais, o MAPE varia de 0 a 100%. Quanto maior o erro do modelo, maior o valor do MAPE.

- **Na prática:** é a porcentagem relacionada a razão da diferença entre o predito e o real em relação ao valor real.
- **Vantagens:**
  - Entendimento simples, extremamente intuitiva.
  - A comunicação de um resultado é amplamente facilitada pela porcentagem.
- **Desvantagens:**
  - Não lida bem com modelos que preveem um intervalo muito grande de valores.

**Figura 11 – Fórmula MAPE.**

$$MAPE = \frac{1}{n} \Sigma_{i=1}^{n} \left| \frac{y_i - \widehat{y_i}}{y_i} \right|$$

## Considerações Finais

- Todas as métricas possuem vantagens e desvantagens que dependem dos dados analisados e das perguntas que queremos responder.
- A utilização de várias métricas é muito comum e permite analisar os modelos sob vários aspectos distintos.
- **Não existe métrica certa ou errada!** Algumas respondem melhor ou pior ao que se está perguntando.

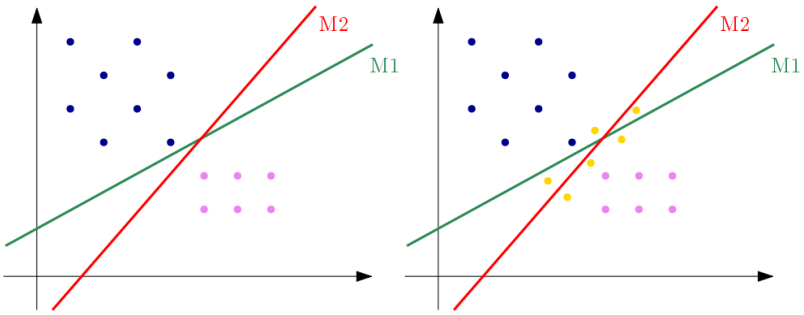
## Capítulo 3. Métricas de Desempenho para Classificação

Vamos agora falar um pouco sobre métricas de desempenho para Classificação!

### Case: Predição de mortalidade de pacientes infectados por COVID-19

- Laura e Pedro não param!
  - Eles agora estão trabalhando em um modelo que seja capaz de prever a chance de mortalidade de uma pessoa recém infectada por COVID-19.
  - Eles usaram os dados que foram fornecidos para criar um modelo de classificação que indica se existe um padrão de morte ou não.
  - **Como ter uma ideia de qual será o comportamento de M1 e M2 perante novas amostras? Como avaliar o desempenho?**

**Figura 12 – Ilustração: Predição de mortalidade de pacientes infectados por COVID-19. Qual modelo responde melhor?**



## Matriz de Confusão

- Técnica visual, simples e objetiva para apresentar o desempenho de um modelo de classificação.
- A matriz mostra os exemplos que foram classificados correta e incorretamente pelo modelo.
- Muito utilizada para identificar cenários no qual o modelo está favorecendo alguma classe.
- Aplicada à classificação binária ou multiclasse.

**Figura 13 – Exemplo de Matriz de Confusão.**

| M1 | | Classe Predita | |
|---|---|---|---|
| | | Morre | Sobrevive |
| Classe Real | Morre | 2 | 1 |
| | Sobrevive | 0 | 3 |

## Taxa de Erro

- Técnica simples e direta que apenas conta a quantidade de erros que os modelos tiveram, independente da classe.
- Métrica de análise superficial, que não permite identificar a distribuição do erro em relação às classes.
- Não permite fornecer pesos diferentes para erros diferentes.
- Aplicada à classificação binária ou multiclasse.

**Figura 14 – Fórmula Taxa de Erro.**

$$Taxa\ de\ Erro = \frac{Incorretamente\ classificados}{Total\ de\ amostras}$$

## Acurácia

- Contrário da taxa de erro.
- Apenas conta a quantidade de acertos que os modelos tiveram, independente da classe.
- Não permite fornecer pesos diferentes para erros diferentes.
- **Acurácias elevadas não necessariamente indicam modelos bons.**
- Aplicada à classificação binária ou multiclasse.

**Figura 15 – Fórmula Acurácia.**

$$Acuracia = \frac{Corretamente\ classificados}{Total\ de\ amostras}$$

## Falso/Verdadeiro Positivos e Negativos

**Figura 16 – Ilustração Falso/Verdadeiro Positivos e Negativos.**


**Fonte:** https://medium.com/@neeraj.kumar.iitg/statistical-performance-measures-12bad66694b7.

**Figura 17 – Falso/Verdadeiro Positivos e Negativos para predição de mortalidade em pacientes infectados por COVID-19.**

| | | Classe Predita | |
|---|---|---|---|
| | | Morre | Sobrevive |
| Classe Real | Morre | Verdadeiro Positivo | Falso Positivo |
| | Sobrevive | Falso Negativo | Verdadeiro Negativo |

## Precisão

- **Precision, Specificity e True Negative Rate.**
- Métrica muito comum amplamente usada para avaliação de modelos. Geralmente usada em conjunto com a Revocação e F1-measure.
- Compara os valores de Verdadeiro Positivos com os erros cometidos para a classe de positivos.
- Ênfase em erros do tipo 1 - Falso Positivo.
- Quanto maior a precisão, mais apurado é o modelo.
- Aplicada à classificação binária ou multiclasse.

**Figura 18 – Fórmula Precisão.**

$$Precisao = \frac{VP}{VP+FP}$$

## Revocação

### *Recall, Sensitivity, True Positive Rate*

- Métrica amplamente usada para avaliação de modelos em conjunto com a Precisão.
- Na prática: de todos os exemplos positivos, quantos foram classificados corretamente?
- Quanto maior a revocação, mais apurado é o modelo.
- Ênfase em erros do tipo 2 - Falso Negativo.
- Aplicada à classificação binária ou multiclasse.

**Figura 19 – Fórmula Revocação**

$$Revocacao = \frac{VP}{VP+FN}$$

## Precisão x Revocação

- Precisão e revocação precisam ser equilibradas entre si para que o modelo proposto tenha uma boa avaliação.
- Por vezes, o aumento da revocação implica em uma diminuição da revocação.
- **Na prática:** quanto mais aumentamos os acertos (melhorar a precisão), menos estamos dispostos a errar (aumentar a revocação).

## F-measure

### *F-score, F1, Medida F*

- Métrica amplamente usada para avaliação de modelos.
- Métrica de fácil interpretação e aponta uma visão geral da qualidade do modelo.
- Combina precisão e revocação em uma única medida.
- Quanto maior a F1 mais apurado é o modelo.
- Aplicada à classificação binária ou multiclasse.
- **Desvantagem:** modelos enviesados para uma das classes podem apresentar F1 alta.

**Figura 20 – Fórmula F1**

$$F1 = \frac{2 * precisao * revocacao}{precisao + revocacao}$$

## Macro e Micro F1

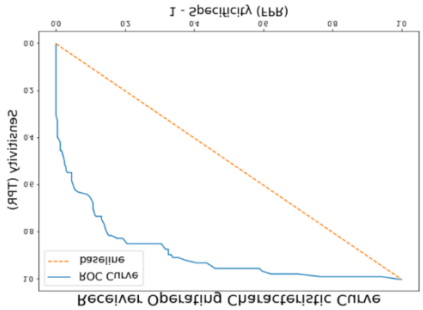
- **Macro F1:**
  - Média aritmética da F1 para cada classe.
- **Micro F1 - F-measure:**
  - Calculada globalmente, contando o total de verdadeiros positivos, falsos negativos e falsos positivos.
- **Weighted F1:**
  - Média ponderada por frequência de classe para cada classe.
  - Retira o viés da F1 tradicional.

## Curva ROC

*Receiver Operating Characteristic Curve - ROC Curve*

- Métrica amplamente usada para avaliação de um modelo, em diferentes pontos da validação.
- Apresentada de forma gráfica, a curva ROC plota os valores de precisão (eixo x) e revocação (eixo y) para os diferentes pontos da validação.
- Combina precisão e revocação em uma única medida.
- Aplicada à classificação binária ou multiclasse.

**Figura 21 – Ilustração: Curva ROC**



**Fonte:** https://medium.com/turing-talks/como-avaliar-seu-modelo-de-classifica%C3%A7%C3%A3o-acd2a03690e.

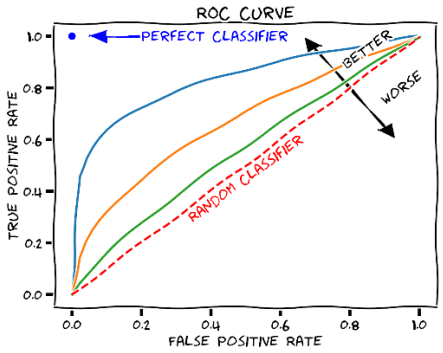
## AUC – Area Under the ROC Curve

*Área abaixo da curva ROC*

- Métrica muito útil baseada na curva ROC, usada para medir a qualidade de diferentes modelos.
- Interpretação complicada.

- Essa métrica não é sensível ao desbalanceamento entre as classes do problema.
- Aplicada à classificação binária ou multiclasse.

**Figura 21 – Ilustração: Curva ROC.**




**Fonte:** https://medium.com/turing-talks/como-avaliar-seu-modelo-de-classifica%C3%A7%C3%A3o-acd2a03690e.

## Conclusão

- O sucesso de um projeto de Machine Learning que envolve a proposta de um classificador, depende fortemente do uso de métricas adequadas para avaliação do modelo proposto.
- A escolha da métrica mais adequada para avaliação do modelo deve levar em consideração as características de distribuição dos dados do problema em questão.
- Não existe uma métrica que avalia corretamente todos os modelos existentes!

# Capítulo 4. Métricas de Desempenho para Clusterização

Vamos discutir as métricas de desempenho de modelos de aprendizado não supervisionado. Para esses modelos os dados não são rotulados, isto é, não existe a informação de classe para as amostras do conjunto de dados. Em geral, nesse cenário não é possível avaliar um modelo a partir dos acertos ou erros. Devemos comparar a qualidade dos agrupamentos realizados para definir se um modelo é melhor do que o outro ou não.

## Modelos de Clusterização

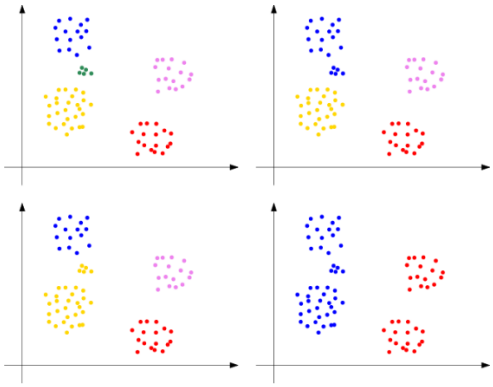
- Um cluster é um subconjunto dos dados originais de um problema, no qual todos os pontos desse subconjunto apresentam características similares.
- Um algoritmo de agrupamento ou clusterização visa identificar os pontos do conjunto de dados original, que devem ser agrupados.
- Modelos de clusterização são parte da tarefa de aprendizado supervisionado.
- O rótulo das amostras não está disponível na base de dados.
- Os modelos trabalham na comparação das amostras e agrupam aquelas que são similares.

## Previsão de chegada das novas vacinas para COVID-19

- Laura e Pedro estão trabalhando em identificar o padrão de transporte das novas vacinas para COVID-19.
- Não existe um prazo fixo estabelecido para a entrega da nova vacina. No entanto, eles conseguiram dados do comportamento de prazo de entrega de outras vacinas.

- Pedro e Laura estão aplicando modelos de clusterização para identificar um padrão no comportamento de entrega e tentar estabelecer um prazo razoável para o recebimento da vacina nova.

**Figura 22 – Ilustração: Previsão de chegada das novas vacinas para COVID-19.**



## Validação de modelos não supervisionados

- Por que avaliar modelos de agrupamento?
  - Característica não aleatória dos dados.
  - Modelos diferentes, clusterings diferentes.
- **Como avaliar modelos de clusterização?**
  - Como não existe a classe para comparar, é difícil avaliar se um modelo está correto para operacionalização ou não.
  - Normalmente, os modelos são avaliados através de comparações e estimativas.
  - Existem métricas para avaliar a similaridade das amostras dentro de um mesmo grupo (métricas intragrupo) e métricas para avaliar a similaridade entre diferentes grupos (métricas intergrupo).

- **Na prática:** um bom modelo de clusterização possui alta similaridade intra cluster e baixa similaridade entre clusters.

## Medidas Internas

- Medidas internas utilizam apenas as informações obtidas através do agrupamento realizado pelo modelo de clusterização.
- Em geral, são baseados em medidas de distância.
- Índices comumente utilizados, já que não há informações sobre os rótulos das amostras.
- **Medidas intragrupo** analisam a compacidade:
  - Avaliam a qualidade do grupo formado.
  - Em geral, essas medidas buscam identificar se os elementos de um mesmo grupo seguem um padrão similar.
- **Medidas intergrupo** analisam a separabilidade:
  - Avaliam a qualidade da dissimilaridade dos grupos formados.
  - Em geral, essas medidas buscam identificar se os elementos de grupos diferentes seguem padrões diferentes.

## Medidas Externas

- Partem do princípio que existe alguma informação sobre o rótulo das amostras.

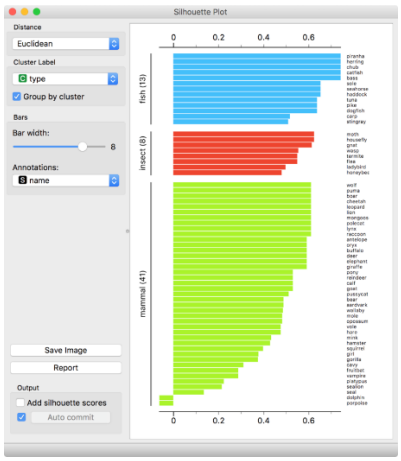
- Implica que possíveis informações sobre a classe das amostras vêm de fonte confiável.
- Aproxima o problema de clusterização a um problema de classificação.

## Coeficiente de Silhueta

*Coeficiente de Silhouette, método da Silhueta, Silhouette Coefficient*

- Medida interna.
- Essa métrica calcula um valor que indica o grau de pertencimento de cada amostra dentro de um grupo.
- O Coeficiente de Silhueta CS varia entre -1 e +1. -1 indica um grau de pertencimento baixo e +1 indica um grau de pertencimento alto.

**Figura 23 – Ilustração: Coeficiente de Silhueta.**




**Fonte: Retirado de:**

**https://en.wikipedia.org/wiki/Silhouette_(clustering)#/media/File:Silhouette-plot-orange.png.**

**Figura 24 – Fórmula Coeficiente de Silhueta.**

$$CS(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

- **i** é a amostra avaliada.
- **a** é a distância média da amostra **i** para todas as outras amostras do cluster.
- **b** é a menor distância média entre **i** e todos os outros clusters.

## Pureza

*Purity*

- Essa métrica calcula a razão do tamanho da classe dominante no cluster em relação ao tamanho do próprio cluster.
- Medida Externa.
- **Na prática:** mede o quão puro um cluster é em relação às amostras dele.

**Figura 25 – Fórmula Pureza.**

$$P(C) = \frac{1}{n} \underset{h}{max}(n^{(h)})$$

- n é o número de amostras do cluster C.
- n(h) é o número de amostras em C que pertencem à classe h=1, 2, …, k.

## Índice de Jaccard

### *Jaccard Index*

- Essa métrica calcula a razão entre o número de elementos comuns a dois clusters e o número de elementos únicos nos dois clusters.
- Medida Externa.
- **Na prática:** quantifica a semelhança entre dois conjuntos.

**Figura 26 – Fórmula Jaccard.**

$$Jaccard(A,B) = \frac{|A \cap B|}{A \cup B} = \frac{VP}{VP + FP + FN}$$

- Assume valores entre 0 e 1.
- Jaccard(A,B) = 0, indica que os clusters não tem elementos em comum
- Jaccard(A,B) = 1, indica que os clusters são idênticos.

## Conclusão

- **Em muitos casos, a qualidade dos grupos está diretamente ligada a quantidade correta de grupos.**
- Qualidade boa implica em quantidade correta de grupos.
- Como estamos falando em aprendizado não supervisionado, a operacionalização é feita considerando bons valores de qualidade.

# Capítulo 5. Técnicas de Validação

Por que precisamos validar modelos? Não basta apenas operacionalizar! Precisamos entender qual modelo vai se comportar melhor na presença de novas amostras. Nesse capítulo, falaremos um pouco sobre técnicas de validação de modelos.

## Etapas de desenvolvimento de Modelos

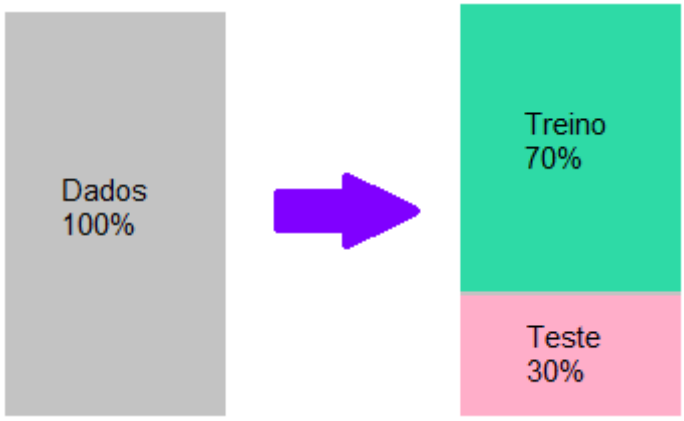
1. **Treino:**
   - A etapa de treinamento consiste na etapa de construção do modelo.
   - Comumente chamada de etapa de aprendizado.
   - É nessa etapa que o modelo é ajustado ao conjunto de dados do problema da melhor forma possível.
2. **Validação:**
   - Após o treinamento do modelo, é preciso entender como o modelo se comporta na presença de dados não observados durante o treinamento.
   - Essa etapa é responsável por um teste controlado do modelo, uma validação da modelagem.
   - É nessa etapa que as métricas vistas anteriormente são aplicadas.
3. **Teste:**
   - Após a validação do modelo, ele pode ser operacionalizado.
   - Nessa etapa, o modelo passa a receber dados que não fazem parte do conjunto de dados original, isto é, dados novos.

## Divisão Treino e Teste

*Train-Test Split, Hold-out validation e Validação Simples*

- Técnica simples, porém pouca utilizada na prática.
- Facilmente enviesada pela determinação da porcentagem de partição.
- Consiste em dividir o conjunto original de dados em duas partes, geralmente de tamanhos distintos.
- Um dos conjuntos é usado para treinamento do modelo e o outro é usado para validar o modelo proposto.

**Figura 27 – Divisão Treino e Teste.**




- **Variação:** Treino, validação e teste.
- Adaptação da técnica original para simular o processo completo de construção até a operacionalização do modelo.
- Pouco utilizada por apenas simular a operacionalização.
- **Na prática:** todas as amostras vem do conjunto original de dados e todas são rotuladas.

**Figura 28 – Divisão Treino, Validação e Teste.**

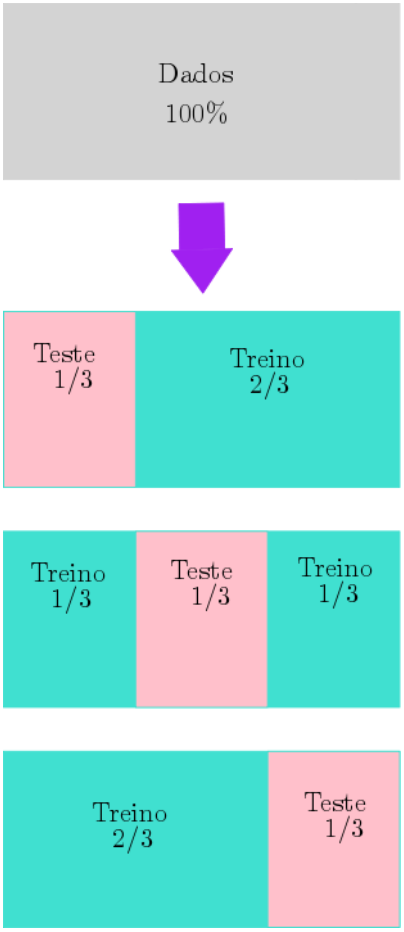

- **Vantagens:**
  - Técnica simples, de fácil implementação.
  - Útil quando a amostra é suficientemente grande.
- **Desvantagens:**
  - No caso de amostras pequenas, a chance de gerar um resultado enviesado é alta.
  - Geralmente não apresenta um bom resultado em cenários de dados desbalanceados.

## Validação Cruzada

- Técnica estatística mais utilizada para avaliação e validação de modelos de aprendizado de máquina.
- Reduz drasticamente o viés da técnica de Divisão de Treino e Teste, e reduz a variabilidade das medidas de erro.

- A validação dos modelos é feita de forma robusta, evitando problemas de aleatoriedade.
- O algoritmo mais utilizado para a validação cruzada é o **k-fold cross-validation.**
- A técnica consiste em dividir o conjunto de dados original em k partes aleatórias e combinar k-1 partes para o treinamento do modelo e a parte restante para teste. O processo é repetido k vezes e o resultado final é a média das execuções.
- Valores de k mais comuns: 3, 5 e 10.

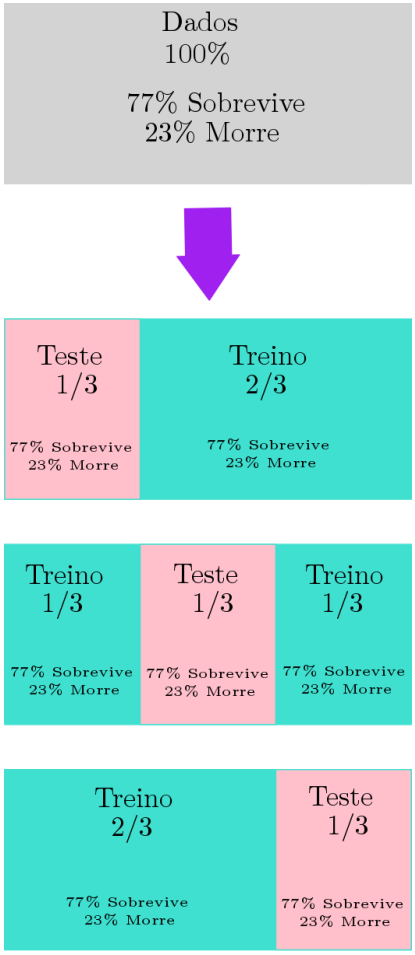
**Figura 29 – Validação Cruzada.**



*Variação: Stratified k-fold Cross Validation*

- Adaptação da técnica original, que considera a distribuição das amostras por classe durante a divisão.
- **Na prática:** o processo de validação é estratificado na maior parte das vezes.
- **Na prática:** em dados desbalanceados, pode ser que algumas amostras precisem ser repetidas para que a estratificação funcione.
- **Na prática:** quanto maior o valor de k, mais robusta é a avaliação e menor é o viés.

**Figura 30 – Validação Cruzada Estratificada.**



- **Vantagens:**
  - Técnica robusta, apresenta bons resultados e de fácil aplicabilidade.
  - Pode ser usada em conjunto com várias métricas de desempenho.
  - Fornece estimativas bem precisas sobre o comportamento do modelo.
- **Desvantagens:**
  - A execução pode ser lenta se o conjunto de dados for muito grande ou se o valor de k for alto.

## Leave-one-out

- Variação da técnica de k-fold cross validation na qual k = número de amostras do conjunto de dados (n).
- Nesse processo, o conjunto de dados é dividido, separando uma única amostra para teste e as demais para treino. O processo é repetido para todas as amostras do conjunto inicial, isto é, por n vezes.
- Ao final do processo, teremos a média de n execuções. Amplamente utilizada quando o conjunto de dados é muito restrito.
- Em geral oferece estimativas menos tendenciosas, mas a técnica **deve ser usada com cautela**, já que a variância dos resultados pode ser aumentada.
- Não existe estratificação. Produz bons resultados na análise de problemas com dados desbalanceados.
- **Vantagens:**
  - Alta robustez.

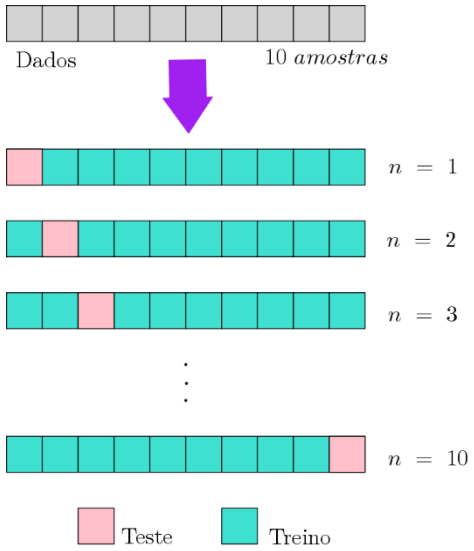
  - Estimador praticamente não enviesado, por considerar quase todos os dados disponíveis para treino.

- **Desvantagens:**

  - Alto custo computacional, principalmente quando n é grande.

  - Alta variabilidade.

**Figura 31 – Leave-one-out.**




## Considerações Finais

- Assim como quando falamos de métricas de desempenho, não existe uma técnica de validação que seja a melhor opção em 100% das situações.

- Para cada problema, é importante avaliar a métrica que mais se adequa às características dos dados que queremos modelar.

- Técnicas de validação são utilizadas para trazer confiança no modelo proposto.

- Além disso, na maioria dos casos, podemos identificar a necessidade de sintonia de hiperparâmetros a partir dos resultados das técnicas.

| | Como funciona? | Quando usar? | Cuidados |
|---|---|---|---|
| **Teste e Treino** | Uma execução da divisão treino e teste. | Conjunto de dados **muito** grandes. | Facilmente enviesada. |
| **Validação Cruzada** | O conjunto de dados é particionado em k partes e recombinado por k vezes. | Basicamente em qualquer situação. | Utilizar divisão estratificada pode reduzir o viés consideravelmente. |
| **Leave-one out** | A cada iteração uma única amostra é separada para teste e o restante para treino. O processo é retido de acordo com o número de amostras. | Conjuntos de dados pequenos; Conjuntos de dados amplamente desbalanceados. | Alto custo computacional. |

# Capítulo 6. Sintonia de Hiperparâmetros

Nesse capítulo vamos falar um pouco sobre como refinar modelos de aprendizado a partir de métodos de sintonia de hiperparâmetros.

## Case: Predição de mortalidade em pacientes infectados com COVID-19

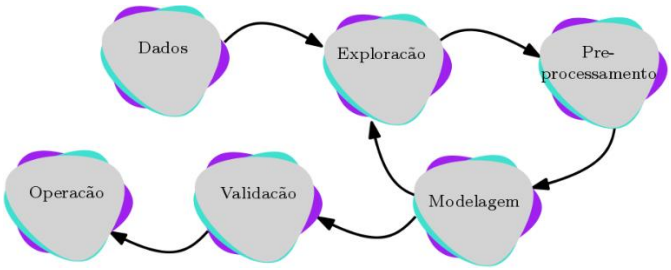
- O modelo da Laura foi escolhido para prever a mortalidade em pacientes infectados com COVID-19. Entretanto, será que o modelo dela pode ser melhorado?
- A resposta para o modelo da Laura e para todos os outros modelos é, em geral, sim.

## Feature Engineering, Feature Selection and Modelo Tuning

Os resultados da modelagem, em geral, podem ser melhorados de duas formas:

- Aprimorando o conjunto de dados, nas etapas de Exploração, pré-processamento e modelagem.
  - Feature Engineering.
  - Feature Selection.
- Aprimorando o modelo proposto, nas etapas de Modelagem e Validação.
  - Model Tuning.

**Figura 32 – Etapas de solução do problema de aprendizado.**




## Sintonia de Hiperparâmetros

*Sintonia de Hiperparâmetros, Otimização de Hiperparâmetros, Combinação de Hiperparâmetros e Model Tuning*

- A grande maioria dos modelos de aprendizado de máquina contam com um ou mais hiperparâmetros responsáveis pela calibragem dos modelos.
- Parâmetros x Hiperparâmetros:
  - Parâmetros são variáveis ajustadas durante o processo de aprendizagem do modelo, de forma automática.
  - Hiperparâmetros são variáveis do algoritmo, definidas antes da etapa de treinamento do modelo.
  - Influenciam diretamente na performance dos modelos.

## Tipos de Hiperparâmetros

- Os hiperparâmetros seguem os tipos característicos das variáveis.
  - Integer – Valores inteiros.
  - Double – Valores de ponto flutuantes.

  - Categorical – Strings.
  - Discrete – Intervalo fixo de valores.

- Hiperparâmetros discretos:
  - São especificados pela escolha de um conjunto fixo de valores.
  - Ex.: número de Árvores no Random Forest.
- Hiperparâmetros contínuos:
  - São especificados por uma distribuição em um intervalo de valores contínuos.
  - Ex.: taxa de aprendizado de Redes Neurais.

## Otimizando Hiperparâmetros

Existem duas formas de otimizar os hiperparâmetros de um modelo:

- Manualmente:
  - O programador define um conjunto de valores, definidos de forma arbitrária, que testa cada um dos valores e analisa os resultados.
  - Claramente não é a melhor forma de tunar um modelo.
- Algoritmos de otimização de hiperparâmetros:
  - Um algoritmo testa diferentes cenários de combinação de parâmetros e retorna o melhor cenário encontrado.
- **Como saber se o modelo com os parâmetros otimizados de fato está apresentando um bom resultado?**

  - Baseline! O resultado retornado é normalmente comparado com o modelo, rodando sobre os mesmos dados com a configuração de hiperparâmetros padrão.

## Método Força Bruta

- Método exaustivo de busca de hiperparâmetros.
- O método de força-bruta consiste em testes de todos os valores possíveis para a otimização de um único hiperparâmetro.
  - Ex.: número de árvores no Random Forest.
  - Todos os valores do conjunto dos números naturais.
- Método pouco utilizado. Computacionalmente custoso e a interpretação dos resultados depende da busca exaustiva de vários parâmetros.
- Método Tentativa e Erro.
- “Se você não sabe quais valores tentar, tente todos.”

## Método Random Search

- Método de busca aleatória de hiperparâmetros.
- O método consiste em testar m combinações aleatórias de hiperparâmetros, e retornar aquela que apresentou melhor resultado de validação.
  - m é definido pelo programador.
  - Ex.: número de árvores no Random Forest.
  - m = 3: 10, 1231, 12000.

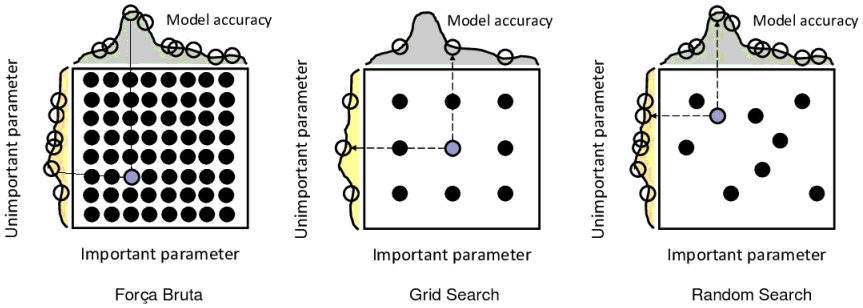
- Método muito utilizado. Pouco custoso e, em geral, apresenta bons resultados.
- **Vantagens:**
  - Tempo de processamento.
  - Muito útil quando o conjunto de dados é muito grande.
- **Desvantagens:**
  - Não garante que a melhor combinação possível vai ser encontrada.

## Método Grid Search

- Busca em Grades:
- Método bem simples e ingênuo. O algoritmo recebe um intervalo de valores para cada parâmetro a ser testado e um fator de divisão, para que entenda como deve ser dividido o intervalo. Ao final, retorna a melhor combinação de parâmetros encontrada no intervalo determinado.
  - Ex.: parâmetro C do SVR.
  - ([1, 5], 1) = 1, 2, 3, 4 e 5.
  - ([1, 1.5], 0.1) = 1, 1.1, 1.2, 1.3, 1.4, 1.5
- Admite independência entre os hiperparâmetros.
- **Vantagens:**
  - Tempo de processamento.
  - Muito útil quando o conjunto de dados é pequeno.
  - Garante a combinação ótima.

- **Desvantagens:**
  - Tempo de processamento.
  - Computacionalmente lento.

**Figura 32 – Comparativo de Força Bruta, Random Search e Grid Search.**




**Fonte: Adaptado pela própria autora de: https://www.oreilly.com/content/evaluating-machine-learning-models/.**

## Referências

BISHOP, Christopher M. *Pattern Recognition and Machine Learning*. Springer, 2nd printing, 2011.

HASTIE, Trevos; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The Elements of Statistical Learning*: Data Mining, Inference, and Prediction. 2. ed. Springer, 2001.

SCIKIT-LEARN. *Model Selection and Evaluation*. Disponível em: <https://scikit-learn.org/stable/model_selection.html#model-selection>. Acesso em: 17 jun. 2021.

ZAKI, Mohammed J.; MEIRA JR, Wagner. *Data Mining and Machine Learning*: Fundamental Concepts and Algorithms. 2. ed. Cambridge University Press, March 2020. Disponível em: <https://dataminingbook.info/>. Acesso em: 17 jun. 2021.