

# ***Machine Learning***



**Cruzeiro do Sul Virtual**  
Educação a distância



# Material Teórico



**Conceitos Logísticos**

**Responsável pelo Conteúdo:**

Prof. Me. Orlando da Silva Junior

**Revisão Textual:**

Prof.<sup>a</sup> Dr.<sup>a</sup> Selma Aparecida Cesarin





- Introdução à Avaliação de Modelos Descritivos;
- Avaliação da Mineração de Padrões Frequentes;
- Avaliação de Agrupamentos;
- Aplicações em *R*;
- Aplicações em *Python*.



### OBJETIVOS DE APRENDIZADO

- Apontar critérios para validação de Modelos Descritivos;
- Avaliar métodos de Aprendizagem Não Supervisionada.





# Orientações de estudo

Para que o conteúdo desta Disciplina seja bem aproveitado e haja maior aplicabilidade na sua formação acadêmica e atuação profissional, siga algumas recomendações básicas:



## Assim:

- ✓ Organize seus estudos de maneira que passem a fazer parte da sua rotina. Por exemplo, você poderá determinar um dia e horário fixos como seu “momento do estudo”;
- ✓ Procure se alimentar e se hidratar quando for estudar; lembre-se de que uma alimentação saudável pode proporcionar melhor aproveitamento do estudo;
- ✓ No material de cada Unidade, há leituras indicadas e, entre elas, artigos científicos, livros, vídeos e sites para aprofundar os conhecimentos adquiridos ao longo da Unidade. Além disso, você também encontrará sugestões de conteúdo extra no item **Material Complementar**, que ampliarão sua interpretação e auxiliarão no pleno entendimento dos temas abordados;
- ✓ Após o contato com o conteúdo proposto, participe dos debates mediados em fóruns de discussão, pois irão auxiliar a verificar o quanto você absorveu de conhecimento, além de propiciar o contato com seus colegas e tutores, o que se apresenta como rico espaço de troca de ideias e de aprendizagem.

# Introdução à Avaliação de Modelos Descritivos

Ao contrário dos Métodos de Aprendizagem Supervisionada, que podem ser avaliados por meio de critérios objetivos, na Avaliação de Modelos Descritos, a avaliação torna-se mais subjetiva.

Enquanto os Modelos Preditivos podem se beneficiar do rótulo presente em cada exemplo do conjunto de treinamento para determinar a eficácia do Processo de Aprendizagem, os métodos não Supervisionados requerem atenção especial do especialista para avaliar os resultados dos algoritmos.

Apesar da Avaliação Não Supervisionada ser mais subjetiva, isso não quer dizer que não existam critérios que possam auxiliar o especialista nessa tarefa.

Por continuarmos trabalhando com a Ciência (nesse caso, a Ciência dos Dados), precisamos garantir a efetividade e a utilidade dos resultados, bem como a sua reproduzibilidade.

Procuraremos tornar a análise o mais objetiva possível, utilizando índices estatísticos que quantificam a informação a respeito da qualidade das regras de associação geradas (para a mineração de padrões frequentes) ou dos grupos formados (para a análise de agrupamentos).

Nesta unidade, vamos conhecer alguns desses índices e critérios, e aprender a trabalhar com eles na prática.

## Avaliação da Mineração de Padrões Frequentes

As regras de associação geradas pelo processo de mineração de padrões frequentes podem ser validadas por meio de diferentes medidas de interesse.

A maior parte dessas medidas se baseia em dois requisitos para quantificar sua utilidade, sendo eles:

- **Suporte:** corresponde ao número de transações para as quais a regra realiza uma predição correta;
- **Confiança:** corresponde ao número de transações que ela prediz corretamente e é proporcional às transações para as quais ela se aplica.

A partir desses conceitos, podemos construir **medidas de interesse** que avaliam as características das regras de associações. Entre as principais medidas, estão o próprio suporte e confiança.

O **suporte** de uma regra de associação corresponde a uma medida que indica a frequência de ocorrência da regra.

Regras com baixo suporte são regras que ocorrem ocasionalmente, tornando-se pouco interessantes para o negócio.

Podemos calcular o suporte de uma regra conforme a seguinte fórmula (de CASTRO; FERRARI, 2016):

$$Suporte(A \rightarrow B) = P(A \cup B) = \frac{\sigma(A \cup B)}{n}$$

Já o propósito da medida de **confiança** é verificar a ocorrência da parte consequente da regra em relação ao antecedente para determinar o grau de confiança entre os itens.

$$Confiança(A \rightarrow B) = P(B|A) = \frac{\sigma(A \cup B)}{\sigma(A)}$$

Para a leitura de ambas as medidas, temos:

- *Suporte(regra)* retorna a probabilidade de *regra* ser encontrada no conjunto total de transações;
- *Confiança(regra)* retorna o grau de confiança de *regra*;
- $A \rightarrow B$  é a regra de associação em que  $A$  implica  $B$ ;
- $\sigma(x)$  é a contagem do suporte da regra correspondendo à quantidade de transações de um determinado conjunto de itens; e
- $n$  é a quantidade total de transações da base de dados.

Ambas as medidas são muito importantes para a avaliação da aplicação de mineração de padrões frequentes. Enquanto a confiança é uma medida de acurácia da regra, o suporte representa a sua significância estatística.

Em aplicações, devemos estabelecer um valor mínimo para cada uma delas a fim de que a regra, satisfazendo esses limiares, possa fazer parte da composição geral das regras da aplicação. Além disso, no caso de termos muitas regras interessantes, esse valor mínimo ajudará na tomada de decisão, auxiliando a eliminar as menos potenciais.

Outras medidas utilizadas em menor grau na avaliação das aplicações de mineração de padrões frequentes são:

- **Lift:** corresponde à razão entre a confiança da regra e a contagem do suporte do consequente da regra.

Pode ser medida pela fórmula:

$$Lift(A \rightarrow B) = \frac{Confiança(A \rightarrow B)}{\sigma(B)} = \frac{\sigma(A \cup B)}{\sigma(A) \cdot \sigma(B)}$$

- **Convicção:** corresponde à razão da frequência esperada de ocorrência do antecedente sem o consequente se eles forem independentes entre si pela frequência de predições incorretas.

Podemos medir a convicção pela fórmula:

$$\text{Convicção}(A \rightarrow B) = \frac{1 - \text{Suporte}(B)}{1 - \text{Confiança}(A \rightarrow B)}$$

Observe que as novas medidas *lift* e convicção são medidas baseadas nas conhecidas anteriormente, suporte e confiança.

Conforme mencionado, muitas outras medidas podem ser utilizadas na avaliação dessas aplicações. Um ponto importante a ser considerado é o estabelecimento de medidas que sejam também baseadas ou permitam uma correlação direta com aquelas duas, vez que a literatura da área está fortemente baseada nelas.

## Avaliação de Agrupamentos

A avaliação dos resultados em análise de agrupamentos deve levar em consideração duas perspectivas:

- A avaliação e comparação dos algoritmos de agrupamento; e
- A validação das estruturas encontradas pelos algoritmos de agrupamento.

Em relação à comparação dos algoritmos de agrupamento, a avaliação pode envolver tanto algoritmos experimentados isoladamente quanto modelos múltiplos.

Independentemente do caso, a avaliação deve garantir a eficácia, a validade e a reproduzibilidade dos experimentos conduzidos pelo especialista.

A realização de uma análise exploratória prévia pode auxiliar o especialista em suas escolhas. Assim, as técnicas de estatística descritiva e visualização de dados podem ser muito úteis nesta tarefa.

Quando falamos da avaliação das estruturas encontradas pelo algoritmo, devemos nos lembrar de que, apesar de estarmos trabalhando com métodos não supervisionados (ou seja, sem rótulo e sem um resultado correto), é importante termos mecanismos para verificar se o método de aprendizagem está realmente buscando uma estrutura apropriada.

Nesta situação, a avaliação deve ser feita com relação à habilidade do algoritmo encontrar estruturas conhecidas.

Para conduzir o Processo de Avaliação dos Métodos de Agrupamento, iremos nos basear em procedimentos estatísticos rigorosos que utilizam índices estatísticos para mensurar a qualidade das estruturas de grupo encontradas pelo algoritmo.

A utilização e a leitura do índice escolhido devem ser guiadas por um critério de validação.

Os critérios de validação expressam a estratégia utilizada para validar uma estrutura de agrupamento. Ele indica a maneira pela qual um índice é aplicado para validar um agrupamento realizado (FACELLI *et al.*, 2011).

Existem três tipos de critérios de validação:

- Critérios relativos;
- Critérios internos; e
- Critérios externos.

Esses critérios são empregados na avaliação de estruturas hierárquicas, particionais e de agrupamentos individuais. Você conhecerá mais sobre cada um deles a seguir.

## Critérios Relativos

O objetivo dos critérios relativos é comparar diversos agrupamentos em relação a alguma característica. Podemos utilizá-los para comparar algoritmos distintos ou para determinar o valor mais adequado de algum hiperparâmetro. Por exemplo, você pode utilizar um critério relativo para determinar o melhor número de grupos para o *k-means*.

Podemos usar alguns índices como critérios relativos. Embora um índice possa ser utilizado tanto como um critério relativo quanto para outro tipo de critério, a utilização do índice depende de como ele é aplicado.

Em geral, a aplicação de um índice como critério relativo considera calcular o seu valor para vários agrupamentos que estão sendo comparados, a fim de obter uma sequência de valores.

Dessa forma, o melhor agrupamento é definido pelo valor que se destaca na sequência, sendo ele mínimo, máximo ou uma inflexão gráfica.

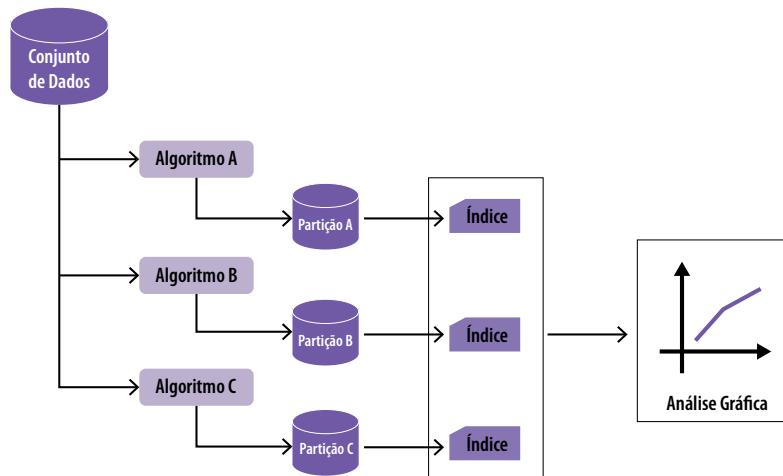


Figura 1 – Critério relativo para validação de agrupamentos

A Figura 1 apresenta uma abordagem para a utilização de critérios relativos em validação de agrupamentos. Observe que múltiplos algoritmos podem ser aplicados às partições selecionadas do conjunto de dados ou um único algoritmo com diferentes configurações pode ser aplicado.

Os índices são extraídos a partir de cada aplicação do algoritmo na partição e, ao final, um gráfico é gerado para melhor análise dos resultados.

Os índices de validação são variáveis aleatórias, cuja distribuição descreve a frequência relativa com a qual seus valores são gerados a partir de uma determinada hipótese.

Para escolhê-los bem, devemos levar em conta alguns aspectos (JAIN; DUBES; 1988):

- **Conceituação do índice:** o índice deve ser intuitivo, ter base teórica e ser computável;
- **Distribuição de probabilidade-base:** o índice deve ser criado a partir de uma distribuição populacional sem estrutura; e
- **Teste de verificação de estrutura:** o índice deve ser capaz em recuperar uma estrutura conhecida.

Entre os índices que cumprem esses aspectos, destacamos:

- **Variância intragrupo:** mede a qualidade de um agrupamento em termos da compactação dos grupos. Apresenta valores no intervalo  $[0, \infty]$ , sendo os valores mais próximos de zero melhores para a partição. Pode ser computado pela seguinte equação, onde  $\bar{x}^{(k)}$  corresponde ao centroide do grupo e  $\pi$  é a partição avaliada:

$$var(\pi) = \sqrt{\frac{1}{n} \sum \sum d(x_i, \bar{x}^{(k)})}$$

- **Conektividade:** reflete o grau com que os exemplos vizinhos são colocados no mesmo grupo. Pode ser calculada pela seguinte equação, onde  $v$  é o número de vizinhos mais próximos que contribuem para a conectividade e  $nn_{ij}$  é o  $j$ -ésimo vizinho mais próximo do exemplo:

$$con(v) = \sum \sum f(x_i, nn_{ij})$$

- **Silhueta:** é baseada na proximidade entre os exemplos de um grupo e na distância dos exemplos de um grupo ao grupo mais próximo. Pode ser usada para avaliar uma partição e tem como valores o intervalo  $[-1, 1]$ , em que a melhor partição tem valores mais próximos de 1. O método da silhueta permite que os valores sejam expressos graficamente. Para isso, devemos calcular o valor  $s(x)$  da silhueta de cada exemplo  $x$ , onde  $a(x)$  é a dissimilaridade média do exemplo em relação a todos os exemplos do mesmo grupo e  $b(x)$  é a dissimilaridade média entre o exemplo e os exemplos do grupo mais próximo:

$$s(x) = \frac{b(x) - a(x)}{\max(a(x), b(x))}$$

## Critérios Internos

---

Os critérios internos tem como meta mensurar a qualidade do agrupamento usando apenas os dados originais, que podem ser os exemplos de treinamento ou a matriz de similaridade. Eles medem o grau em que uma partição obtida por um algoritmo de agrupamento representa uma estrutura presente nos dados, medindo o ajuste entre a partição gerada e os dados utilizados (FACELLI *et al.*, 2011).

Muitas vezes, esse tipo de validação quer mostrar o número de grupos ideal do processo de agrupamento.

Existem muitas dificuldades na aplicação de índices de critérios internos. Uma delas é que podemos tirar conclusões erradas a respeito do número correto de grupos, vez que os índices mais comuns apresentam valores menores para conjuntos de dados que realmente formam grupos do que para dados aleatórios.

Outra dificuldade que vale à pena destacar é a dependência que esses índices têm em valores relacionados às características dos dados, como a quantidade de exemplos e o número de grupos.

## Critérios Externos

---

Os critérios externos buscam avaliar o agrupamento formado conforme uma estrutura previamente estabelecida. Em geral, essa estrutura reflete a visão do especialista sobre a análise dos dados. Por exemplo, o especialista pode conhecer o domínio da aplicação e saber, antecipadamente, a quantidade ideal de grupos.

Nesse sentido, o objetivo da validação externa é medir o quanto o processo de agrupamento confirma uma hipótese previamente definida. Utilizamos os testes estatísticos de hipótese a partir de uma distribuição de referência.

Para utilizar os índices na validação externa, vamos considerar inicialmente as seguintes asserções:

- A = número de pares de exemplos pertencentes a um mesmo grupo e a uma mesma partição;
- B = número de pares exemplos pertencentes a um mesmo grupo e a partições diferentes;
- C = número de pares de exemplos pertencentes a grupos diferentes e à mesma partição; e
- D = número de pares de exemplos pertencentes a grupos diferentes e a partições diferentes.

Assim, entre os índices aplicados em validação externa, os mais utilizados na análise de grupos para as partições  $X$  e  $Y$ , considerando o intervalo  $[0,1]$ , são:

- **Índice de Rand:** calcula a probabilidade de dois exemplos pertencerem ao mesmo grupo ou fazerem parte de dois grupos diferentes nas partições  $X$  e  $Y$ . Pode ser medido pela seguinte equação:

$$Rand(X, Y) = \frac{A + D}{A + B + C + D}$$

- **Índice de Jaccard:** calcula a probabilidade de dois exemplos pertencentes ao mesmo grupo em uma das partições também pertencerem ao mesmo grupo em outra partição. Computamos o índice de Jaccard pela equação:

$$Jaccard(X, Y) = \frac{A}{A + B + C}$$

- **Índice de Folkes e Mallows:** determina a similaridade entre dois grupos  $X$  e  $Y$ , indicando, com valores mais altos, a força do grau dessa semelhança. Podemos encontrar o índice de Folkes e Mallows pela equação:

$$FM(X, Y) = \sqrt{\frac{A}{A + B} * \frac{A}{A + C}}$$

## Aplicações em R

Para realizar os experimentos, você precisará do ambiente  $R$  e do ambiente de desenvolvimento  $RStudio$ .



- Ambiente  $R$ , disponível em: <https://bit.ly/2EM5UmN>
- $RStudio$ , disponível em: <https://bit.ly/3juu9F2>

Ambos os softwares são gratuitos e facilmente instaláveis no *Microsoft Windows*.

A Figura 2 apresenta a interface do software  $RStudio$ . Trabalharemos exclusivamente com ele para a linguagem  $R$ .

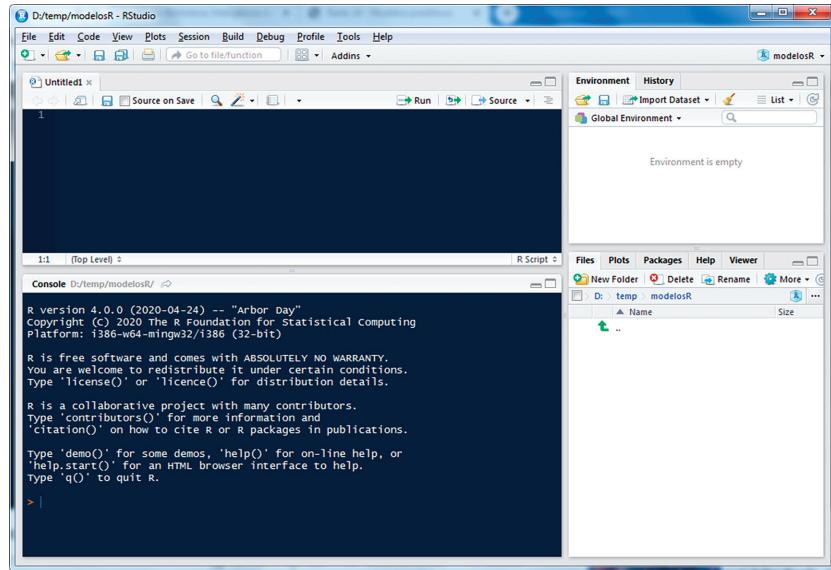


Figura 2 – Interface inicial do software Rstudio

Fonte: Acervo do Conteudista

Inicialmente, devemos identificar os principais métodos de *Machine Learning* dentro da linguagem *R*.

A Tabela 1 indica os nomes das bibliotecas e as funções que poderão ser carregadas e utilizadas, respectivamente.

Tabela 1 – Bibliotecas da linguagem *R* para *Machine Learning*

Método	Pacote	Função
<i>Apriori</i>	arules	apriori
FP-Growth	rCBA	fpgrowth
<i>k-means</i>	stats	kmeans
DBSCAN	fpc	dbscan

Para exemplificar, você construirá uma aplicação descritiva de *Machine Learning* em *R* usando uma base de dados *USAArrests* (estatísticas de crimes cometidos nos EUA em 1973) e o algoritmo *k*-means.

Usando o *RStudio*, comece executando o seguinte código para definir a semente de aleatoriedade do experimento e carregar os conjuntos de dados, padronizando-os logo em seguida:

```
# Define a semente do experimento
set.seed(42)

# Carrega o conjunto de dados
# E padroniza o valor dos atributos
dados <- scale(USAArrests)
```

Continue o seu script *R* com a aplicação do algoritmo *k*-means, usando *k=4*:

```
# Aplica o algoritmo k-means
# Assumindo k = 4
# E imprime o resultado do processo
k <- 4
model <- kmeans(dados, k, nstart=25)
print(model)
```

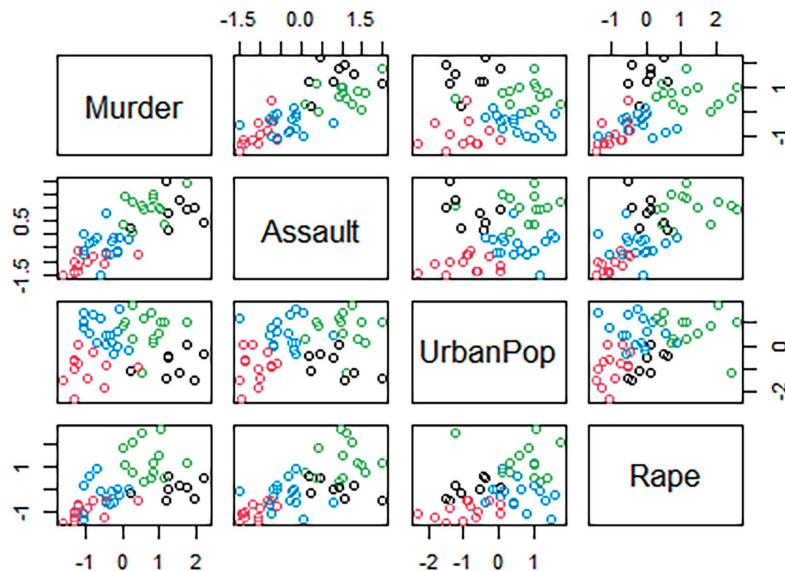


Figura 3 – Agrupamentos formados pelo *k-means*

Fonte: Acervo do Conteudista

Por fim, visualize os grupos formados utilizando o seguindo código:

```
# Plota o gráfico com os grupos formados
with(USArrests, pairs(dados, col=c(1:4)[model$cluster]))
```

Observe, na Figura 3, como os grupos são apresentados. Veja como cada ponto de dado do conjunto de dados foi representado por uma das 4 cores (vermelho, preto, azul e verde) destinada à representação de um grupo formado.

Na figura, você também pode observar a relação entre duas variáveis e o agrupamento realizado por ela.

## Aplicações em Python

Para construir e validar modelos de *Machine Learning* usando a Linguagem *Python*, você precisará de um ambiente que execute os programas que iremos construir.

A distribuição Anaconda é uma das mais populares para Ciência de Dados e pode ser utilizada para essa finalidade, uma vez que ela permite a reprodução dos experimentos por meio da construção de notebooks usando *Jupyter*.



Distribuição Anaconda, disponível em: <https://bit.ly/3cMY88F>

Na aplicação que você vai construir, o objetivo é analisar a cesta de Mercado, buscando por relações entre os produtos comprados pelos consumidores.

Usaremos os dados disponibilizados pelo Kaggle, que você deverá baixar em formato CSV.



Dados Kaggle, disponível em: <https://bit.ly/3l1kmXC>

Não se esqueça de deixar o arquivo no mesmo diretório do seu *notebook*. Usaremos também o algoritmo *Apriori*, aquele mesmo responsável por construir regras de associações.

Para utilizá-lo, comece importando a biblioteca *Apyori* (talvez você tenha que instalá-la):

```
from apyori import apriori
```

Em seguida, carregue os dados na memória:

```
# Lê os dados do computador em formato CSV  
  
dados = pd.read_csv("Market_Basket_Optimisation.csv", header=None)
```

Com os dados carregados, vamos construir a lista de transações a partir do conjunto de dados:

```
# Constrói a lista de transações  
# A partir do conjunto de dados  
transacoes = []  
for i in range(0, len(dados)):  
    transacoes.append([str(dados.values[i,j]) for j in range(0, 20)])
```

Com esse pré-processamento realizado, basta executarmos o algoritmo *Apriori* e geramos a lista de associações relacionadas:

```
# Executa o algoritmo Apriori  
# E retorna um conjunto de associações  
associacoes = apriori(transacoes, min_length = 2,  
                      min_support = 0.003,  
                      min_confidence = 0.3,  
                      min_lift = 3)  
  
# Lista de associações  
associacoes_lista = list(associacoes)
```

Para visualizar os resultados, faça como na Figura 4, imprimindo uma ou mais listas geradas.

```
1 print(associacoes_lista[1])  
  
RelationRecord(items=frozenset({'pasta', 'escalope'}),  
support=0.005865884548726837, ordered_statistics=[Order  
edStatistic(items_base=frozenset({'pasta'}), items_add=  
frozenset({'escalope'}), confidence=0.3728813559322034,  
lift=4.700811850163794)])
```

Figura 4 – Observação da primeira lista de associações

Fonte: Acervo do Conteudista

Na figura, você observa a associação entre os produtos “pasta” (massa italiana) e “escalope” (bife de carne) com nível de confiança de 37,29%.

Para cada regra que você visualizar, encontrará uma associação e nível de confiança diferentes.

Tanto o conjunto de regras quanto o nível de confiança podem mudar conforme o ajuste dos hiperparâmetros do algoritmo.

# Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:

▶ Vídeos

**Curso de C++ – Aula 95 – Agrupamento (*Clustering*) – K-Means**

<https://youtu.be/fyGGhuVR-ik>

📄 Leitura

**Índice de Jaccard – *Jaccard index***

<https://bit.ly/30p8jeP>

***Customer Cluster Analysis***

<https://bit.ly/36oPyfs>

**Biodiversidade e o Índice de Jaccard no R**

<https://bit.ly/3l66aNc>

# Referências

CASTRO, L. N. de; FERRARI, D. G. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

FACELI, K. *et al.* **Inteligência Artificial:** Uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data.** EUA: Prentice-Hall Inc., 1988.

SILVA, L. A. da; PERES, S. M.; BOSCAROLI, C. **Introdução à mineração de dados:** com aplicações em R. São Paulo: Grupo GEN LTC, 2017.





**Cruzeiro do Sul**  
Educacional