Inteligência Artificial Aplicada às Relações Internacionais

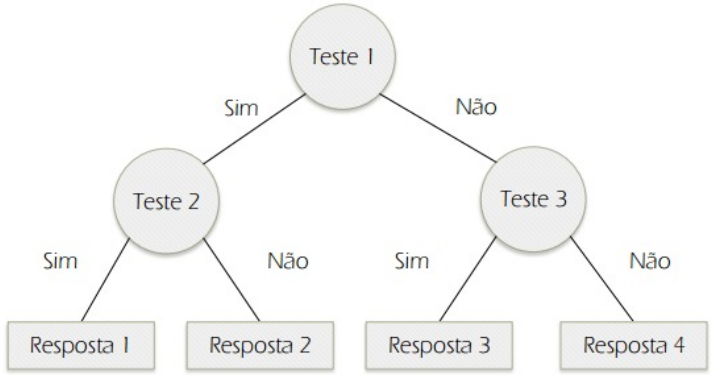
Árvores de Decisão

Prof. Dr. Antonio Marcos SELMINI - antonio.selmini@espm.br
Prof. Dr. Humberto Sandmann – humberto.sandmann@espm.br

TECH | ESPM

Árvores de Decisão

A **Árvore de Decisão** é um tipo de algoritmo de **aprendizagem de máquina supervisionado** que se baseia na ideia de divisão dos dados em grupos homogêneos, podem ser utilizadas em um cenário de classificação ou regressão.

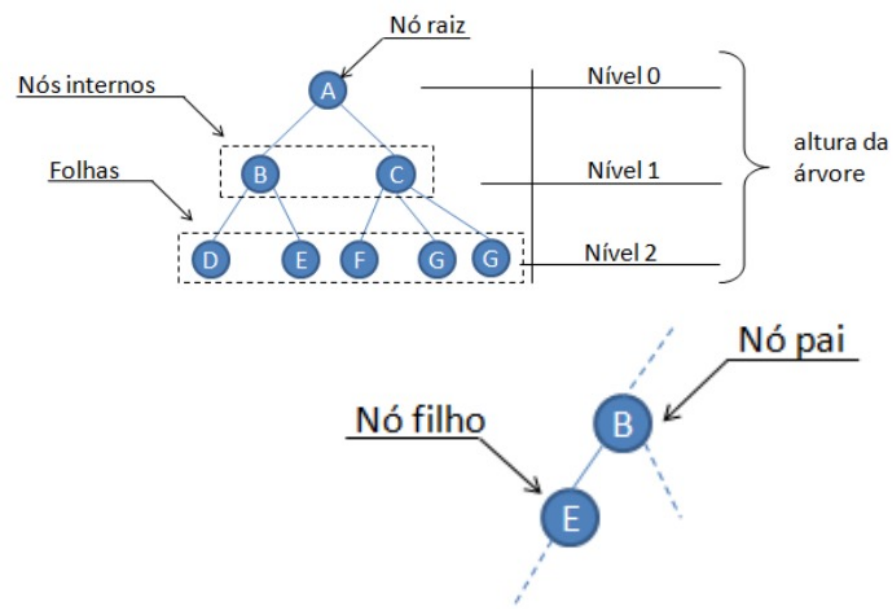


Árvores de Decisão

Pontos que devem ser considerados na utilização de árvores de decisão:

- Possui um fácil entendimento, pois não requer nenhum conhecimento estatístico para a sua interpretação.
- Aceita tanto dados categóricos (dados não numéricos) quanto numéricos diminuindo a necessidade da limpeza de dados em comparação com outros modelos.
- São instáveis, pequenas alterações nos dados de treino produzem novas árvores.

Árvores de Decisão



Fonte: https://saulo.arisa.com.br/wiki/index.php/%C3%81rvores. Acesso em 01/09/2021.

Árvores de Decisão

Fonte: https://didatica.tech/como-funciona-o-algoritmo-arvore-de-decisao/. Acesso em 01/09/2021.

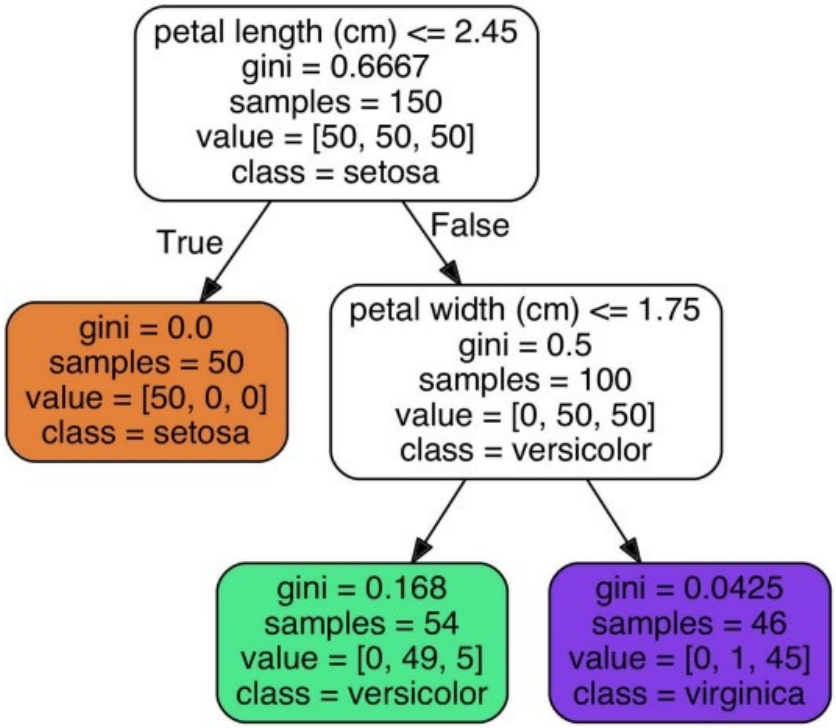
Árvores de Decisão

O objetivo de uma **árvore de decisão** é encontrar o atributo que gera a melhor divisão dos dados, subconjunto com **maior pureza**.

Existem algumas métricas para a definição de pureza, ou seja, qual será a métrica utilizada para decidir qual é o melhor atributo que divide os nossos dados gerando a partição mais pura.

Essas métricas são o **Índice Gini**, **Chi-Square**, **Information Gain**, **redução da variância** e **entropia**.

Árvores de Decisão



Gini (coeficiente de Gini) representa o índice que foi utilizado como medida de impureza. **O valor zero indica que o nó é puro**.

O atributo **samples** de um nó indica a quantidade de instâncias de treinamento a que se aplica.

O atributo **value** de um nó diz a quantas instâncias de treinamento de cada classe o nó se aplica.

Árvores de Decisão

Para a implementação de **árvores de decisão** para classificação usando o **Scikit-Learn**:

biblioteca que deve ser importada

```
from sklearn import tree

classificador = tree.DecisionTreeClassifier()

classificador.fit(x_train, y_train)

tree.plot_tree(classificador)

resultado = classificador.predict(x_test)
```

define o classificador com os parâmetros padrão

imprime a estrutura de árvore

testa a árvore treinada

Árvores de Decisão

Função para medir a qualidade de uma divisão. Os critérios suportados são: `"gini"` para a impureza e, `"entropy"` para o ganho de informação.

```
DecisionTreeClassifier(criterion="gini")
```

Árvores de Decisão

Coeficiente de Gini ou **Entropia**?

Por **padrão** aplica-se o **coeficiente de Gini** para a medida de impureza de um nó, mas pode-se utilizar também a medida de entropia. O coeficiente de Gini para um conjunto é zero (conjunto puro) quando o conjunto contém instâncias de apenas uma classe.

Entropia é um termo aplicado na termodinâmica como uma medida da desordem molecular. Em aprendizado de máquinas, a entropia de um conjunto é zero quando contém instâncias de apenas uma classe.

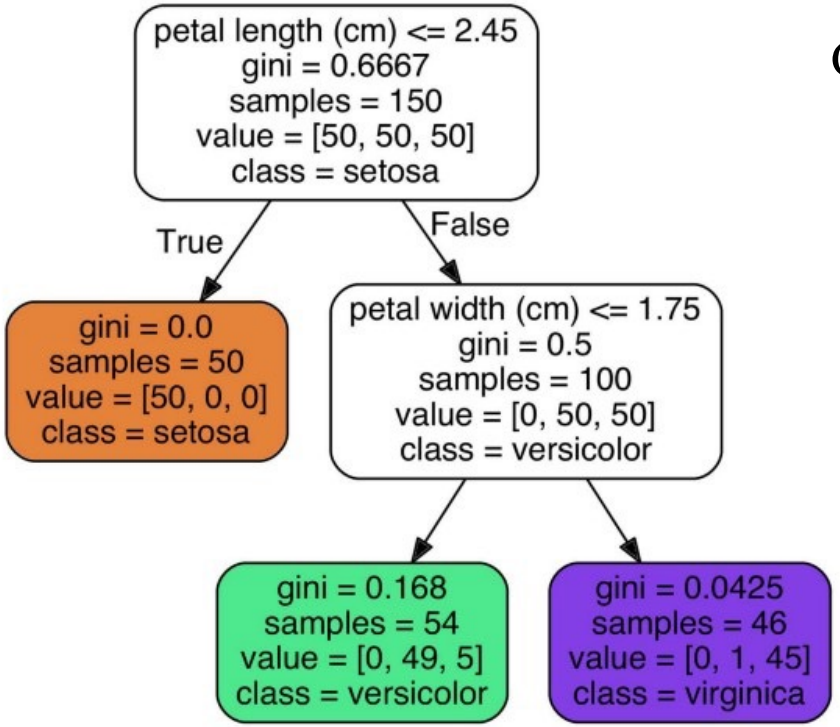
Árvores de Decisão

Expressão para o cálculo do **coeficiente de Gini**:

$$g_i = 1 - \sum_{k=1}^{n} \left(p_{i,k}\right)^2$$

$p_{i,k}$ é a média das instâncias da classe k entre as instâncias de treinamento no nó i.

Árvores de Decisão



Cálculo do **coeficiente de Gini** para o nó esquerdo de profundidade 2:

$$g = 1 - \left(\frac{0}{54}\right)^2 - \left(\frac{49}{54}\right)^2 - \left(\frac{5}{54}\right)^2$$

$$g \approx 0.168$$

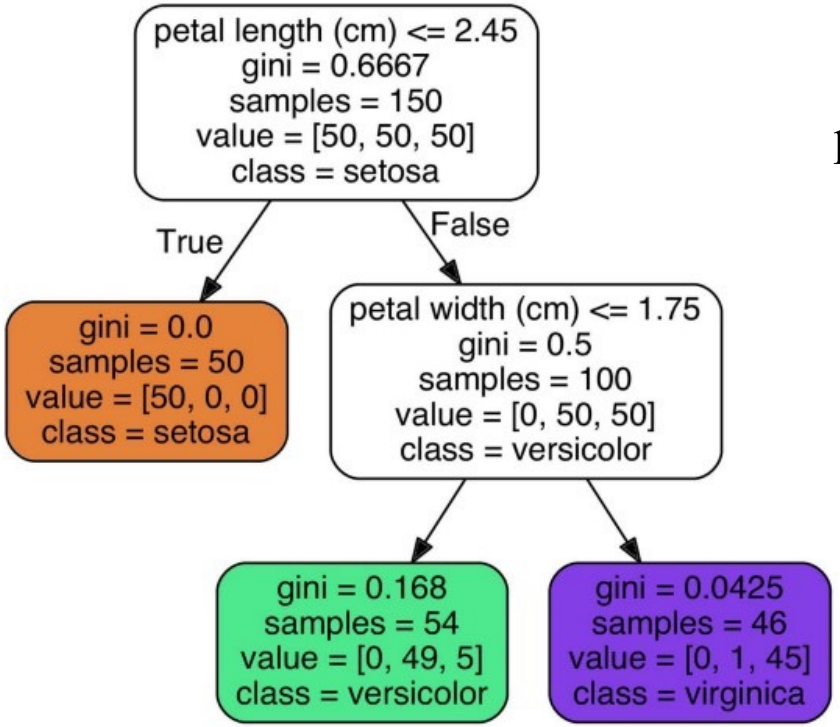
Árvores de Decisão

Expressão para o cálculo do **coeficiente da entropia**:

$$h_i = -\sum_{k=1}^{n} p_{i,k} \log(p_{i,k})$$

$p_{i,k}$ é a média das instâncias da classe k entre as instâncias de treinamento no nó i.

Árvores de Decisão



Cálculo do **coeficiente de entropia** para o nó esquerdo de profundidade 2:

$$h = -\frac{49}{54}\log\left(\frac{49}{54}\right) - \frac{5}{54}\log\left(\frac{5}{54}\right)$$

$$h \approx 0{,}31$$

Árvores de Decisão

Coeficiente de Gini ou **Entropia**?

Na maioria das vezes não faz diferença: os dois coeficientes levam a árvores semelhantes. **O coeficiente de Gini é um pouco mais rápido para calcular, então é um bom padrão**.

Quando os coeficientes diferem, o **coeficiente de Gini** tende a isolar a classe mais frequente em seu próprio ramo da árvore, enquanto que a **entropia** tende a produzir árvores ligeiramente mais equilibradas.

Referências Bibliográficas

- Lenz, Maikon Lucian, et al. Fundamentos de Aprendizagem de Máquina. Grupo A, 2020. Disponível em: https://integrada.minhabiblioteca.com.br/books/9786556900902. Acesso em 20/08/2021.
- Árvore de Decisão. Minerando Dados. Disponível em: https://minerandodados.com.br/arvores-de-decisao-conceitos-e-aplicacoes/. Acesso em 01/09/2021.
- Árvore de Decisão. Medium. https://medium.com/machine-learning-beyond-deep-learning/%C3%A1rvores-de-decis%C3%A3o-3f52f6420b69. Acesso em 01/09/2021.

Referências Bibliográficas

- RUSSEL, S., NORVIG, P. Inteligência Artificial. 3a ed. Grupo GEN, 2013. Disponível em: https://integrada.minhabiblioteca.com.br/reader/books/9788595156104/epubcfi/6/22[%3Bvnd.vst.idref%3Dch01.xhtml]!/4. Acesso em 20/08/2021.