Terminologia de Árvores de Decisão

Jéssica, Luís Felipe, Maria Luiza, Rayssa

Departamento de Estatística
Universidade de Brasília

1 - 2024

Árvore de Decisão

- Uma árvore de decisão é o resultado de uma sequência ordenada de perguntas, e o tipo de pergunta feita em cada etapa da sequência depende das respostas às perguntas anteriores. A sequência termina em uma previsão da classe.
- A origem dessa família de técnicas é do trabalho de Morgan e Sonquist(1963).

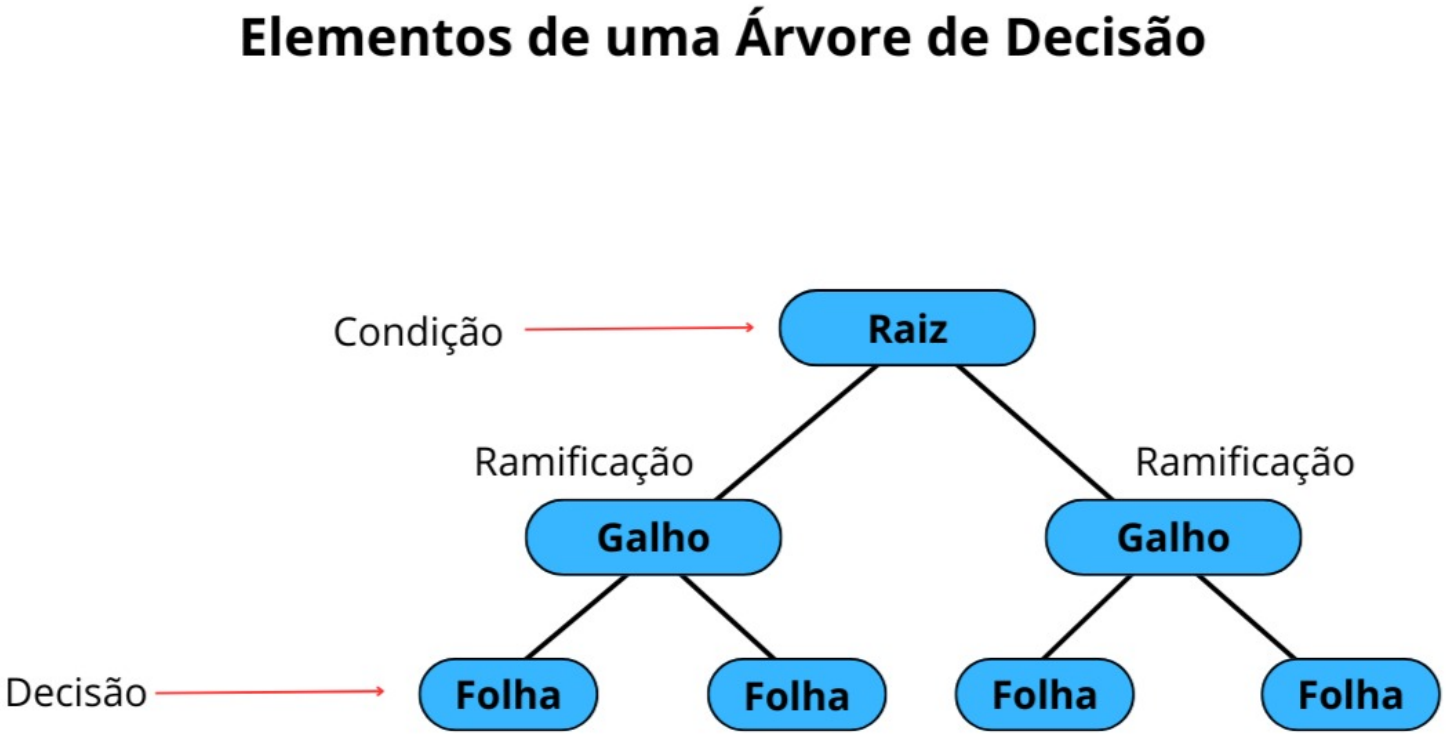
Árvore de Decisão

- Modelos de Árvores de Classificação e Regressão (CART –*Classification And Regression Trees*), (Breiman et al. 1994):
  a) Considere uma partição do espaço gerado pelas variáveis preditoras, X, em M regiões, $R_1, ..., R_M$.
  b) Para cada elemento pertencente a R_j, o previsor de Y (que designaremos $\hat{Y}_{Rj}$) será a moda (no caso discreto) ou a média (no caso contínuo); no caso categorizado, o classificador de Y será a classe com a maior frequência entre os pontos com valores de $X_1, ..., X_n$ em R_j.

Terminologia

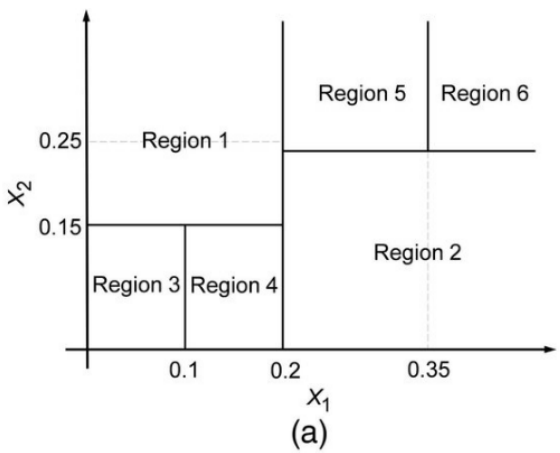
- **Raiz:** Nó que origina a árvore de decisão, ou seja, são todos os dados de treinamento no topo da árvore. Também pode ser referenciado como nível zero da árvore.
- **Nó:** É uma subdivisão dos dados, podendo ser intermediário ou final.
  - **Intermediário (galho):** Um nó intermediário é um nó que se divide em duas categorias. Tal divisão binária é determinada por uma condição booleana no valor de uma única variável, onde a condição é satisfeita ("sim") ou não satisfeita ("não") pelo valor observado dessa variável.
  - **Terminal (folha):**Um nó que não se divide é chamado de nó final e recebe uma classe (classificação). Cada observação em L cai em um dos nós terminais. Pode haver mais de um nó terminal com o mesmo rótulo de classe.

Terminologia

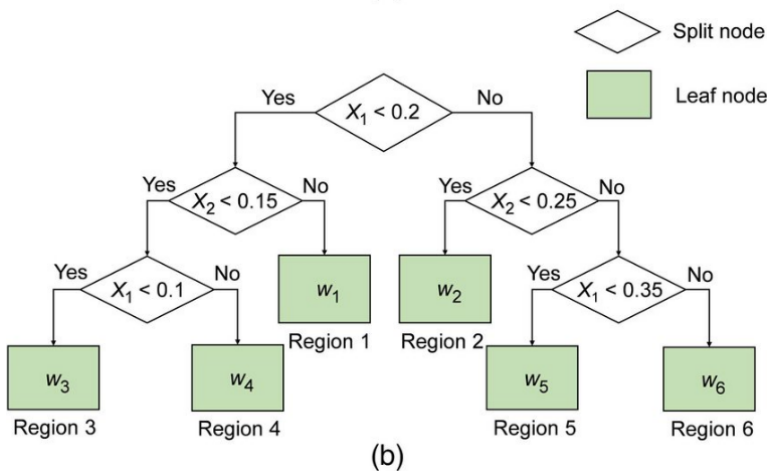


Partições

- Partições são subconjuntos dos dados. Elas se formam a partir da aplicação de condições nos valores observados de uma variável e são determinadas quando se encontram em um nó indivisível (folha).



(a) Partição de um espaço preditivo bidimensional



(b) Estrutura da Árvore associada

Fonte: Wang et al. (2020).

Partições

- **Partições Livres:** Todas as possíveis partições fazem parte do conjunto admissível. Ou seja, são partições sem condições restritivas. Costuma ser o caso de dados com variáveis qualitativas nominais.
- **Partições Restritas:** Nem todas as partições são admissíveis. Alguns exemplos de restrições:

  (a) **Monotônicas**

  (b) **Circulares**

  (c) **Flutuantes**

Partições Restritas

- **Monotônicas:** A variável regressora possui pelo menos um nível ordinal de mensuração que deve ser considerado na partição. Nesse caso, apenas partições que não violem essa ordem são admissíveis.
- **Circulares:** São indicadas para variáveis circulares, como horário e ângulos. Nesse caso, a primeira classe é consecutiva à última classe.
- **Flutuantes:** Alguns valores da variável preditora são ordenados (ou circulares) e os demais são livres; as classes não sujeitas a restrições são denominadas flutuantes. Esse tipo de partição é usado, por exemplo, quando uma das classes corresponde a uma situação de não resposta.

Árvores de Classificação vs. Árvore de Regressão

Ambos são tipos de modelos de árvores de decisão, porém aplicados a problemas diferentes e utilizam critérios distintos para construir e avaliar as árvores.

- Quando a variável resposta Y é categorizada (usualmente referida como **árvore de classificação**), o objetivo é identificar a classe mais provável associada aos valores das variáveis preditoras $X = (x_1, ..., x_p)^\top$. O classificador de Y, nesse caso, será a classe com maior frequência entre os pontos com valores $x_1, ..., x_p$ numa determinada região.
- Já em situações com variável resposta quantitativa Y, as **árvores de regressão** visam prever um valor numérico Y que será a moda (no caso discreto) ou a média (na caso contínuo).

Critérios de Partição

Aqui, discutiremos os métodos ou medidas utilizadas para decidir como dividir os dados em nós nas árvores de decisão durante seu processo de construção.

1. Árvore de Regressão

- **Soma de quadrados da árvore**:

$$\text{SQ(A)} = \sum_{k=1:i\in R_k(A)}^{n_A} (y_i - \bar{y}_k)^2$$

Em que A é a árvore, $R_k(A), k = 1, ..., n_A$ são as k regiões e n_A o número de nós terminais da árvore; y_i são os valores de cada i-ésima observação da resposta no nó k; $\bar{y}_k$ é a média amostral das observações de Y que pertencem à região $R_k(A)$, definida a partir do nó terminal k. A SQ(A) é uma medida do ajuste dos dados da árvore aos dados. Valores pequenos dessa soma de quadrados indicam um bom ajuste.

Critérios de Partição

- **Poda**: normalmente, árvores com muitos nós terminais apresentam bom desempenho no conjunto de treinamento, mas podem estar sujeitas a *overfiting*, e não produzir boas classificações/previsões no conjunto de validação. Uma possibilidade de definir o tamanho é com a poda. Seja A_0 uma árvore com um grande número de nós terminais e A uma subárvore construída a partir de A_0. Seja também $\alpha \geq 0$. Uma função de custo associada a A pode ser construída como:

$$C_\alpha(A) = SQ(A) + \alpha n_A$$

Dado um valor de α, pode-se definir que a melhor árvore, $A \subseteq A_0$, é a que minimiza C_α, conhecido como **parâmetro de complexidade**.

2 Árvores de Classificação

- O critério de partição possui algumas alternativas, como por exemplo, por meio da minimização de:

1. **Erro de classificação**

$$E_k = 1 - \hat{p}_{kj_k}$$

em que $\hat{p}_{kj_k}$ é a proporção de observações de uma partição R_k que pertencem ao nível j de Y.

2. **Índice Gini**

$$G_k = \sum_{j=1}^{q} \hat{p}_{kj}(1 - \hat{p}_{kj}) = 1 - \sum_{j=1}^{q} \hat{p}_{kj}^2$$

em que $\hat{p}_{kj}$ é a proporção de classificação correta da partição R_k.

3. **Índice de Entropia**

$$H_k = -\sum_{j=1}^{q} \hat{p}_{kj} ln(\hat{p}_{kj}) *$$

*Considera-se $p\ln(p) = 0$ se $p = 0$.

Objetivo: Tentar reduzir o erro de generalização (variância, viés).

- **Erro de generalização:** diferença entre o desempenho de um modelo em dados de treinamento e seu desempenho em dados novos ou não vistos, i.e, uma medida de quão bem um modelo se generaliza para dados fora do conjunto de treinamento.

- **Bagging:** Algoritmo que cria conjuntos de treinamento adicionais por amostragem uniforme e com substituição do conjunto de treinamento original.
  bootstrap aggregation $\rightarrow$ (b agg) ing
- **Boosting:** Algoritmo preditivo de *generalized gradient boosting*, que minimiza uma função de perda. No caso de árvores de regressão, a função de perda pondera (pesos) para cada folha dentro da árvore de forma diferencial, de modo que as folhas que fazem melhor predição são bem recompensadas e as que não o fazem são punidas.
- **Florestas:** Algoritmo preditivo que cria de forma aleatória várias Árvores de Decisão (*Decision Trees*) e combina o resultado de todas elas para chegar no resultado final.

Exemplos

Usaremos o conjunto de dados `iris` para criar uma **árvore de regressão** e prever o comprimento da pétala (`Petal.Length`) com base nas outras variáveis.

```r
# Pacotes necessários
library(rpart)
install.packages("rpart.plot")
library(rpart.plot)

# Dados necessários
data(iris)

# Construção da árvore
modelo <- rpart(Petal.Length ~ Sepal.Length + Sepal.Width + Petal.Width + Species,
                data = iris)

summary(modelo)

rpart.plot(modelo)


# Geração das previsões na amostra de validação

set.seed(2)
n <- dim(iris)[1]
indices <- sample(n, size = 75)
Desenv <- iris[indices, ] #amostra de desenvolvimento
Valida <- iris[-indices, ] #amostra de validação

Valida$Prev.length <- predict(modelo, Valida)

# EQM, EAM, EAMR
Media <- mean(Valida$Petal.Length)
EQM <- mean((Valida$Petal.Length-Valida$Prev.length)^2)
(EAM <- mean(abs(Valida$Petal.Length-Valida$Prev.length)))
(EAMR <- mean(abs(Valida$Petal.Length - Valida$Prev.length)/Valida$Petal.Length))
```

Exemplos

```
Call:
rpart(formula = Petal.Length ~ Sepal.Length + Sepal.Width + Petal.Width +
    Species, data = iris, method = "anova")
  n= 150

          CP nsplit  rel error     xerror        xstd
1 0.85149596      0 1.00000000 1.01549761 0.064775974
2 0.08987576      1 0.14850404 0.15260613 0.020726900
3 0.01902604      2 0.05862828 0.05978536 0.008359640
4 0.01137403      3 0.03960224 0.04112205 0.005579793
5 0.01000000      4 0.02822821 0.03820106 0.005454373

Variable importance
 Petal.Width      Species Sepal.Length  Sepal.Width
          31           31           23           15

Node number 1: 150 observations,    complexity param=0.851496
  mean=3.758, MSE=3.095503
  left son=2 (50 obs) right son=3 (100 obs)
  Primary splits:
      Species      splits as  LRR,      improve=0.8514960, (0 missing)
      Petal.Width  < 0.8  to the left,  improve=0.8514960, (0 missing)
      Sepal.Length < 5.55 to the left,  improve=0.6904508, (0 missing)
      Sepal.Width  < 3.35 to the right, improve=0.2612332, (0 missing)
  Surrogate splits:
      Petal.Width  < 0.8  to the left,  agree=1.000, adj=1.00, (0 split)
      Sepal.Length < 5.45 to the left,  agree=0.920, adj=0.76, (0 split)
      Sepal.Width  < 3.35 to the right, agree=0.833, adj=0.50, (0 split)

Node number 2: 50 observations
  mean=1.462, MSE=0.029556
```

Exemplos

Tabela: Métricas de avaliação do desempenho do modelo de árvore de regressão

Medida	**Valor**
Erro Quadrático Médio (EQM)	0.09
Erro Absoluto Médio (EAM)	0.24
Erro Absoluto Médio Relativo (EAMR)	0.07

Exemplos

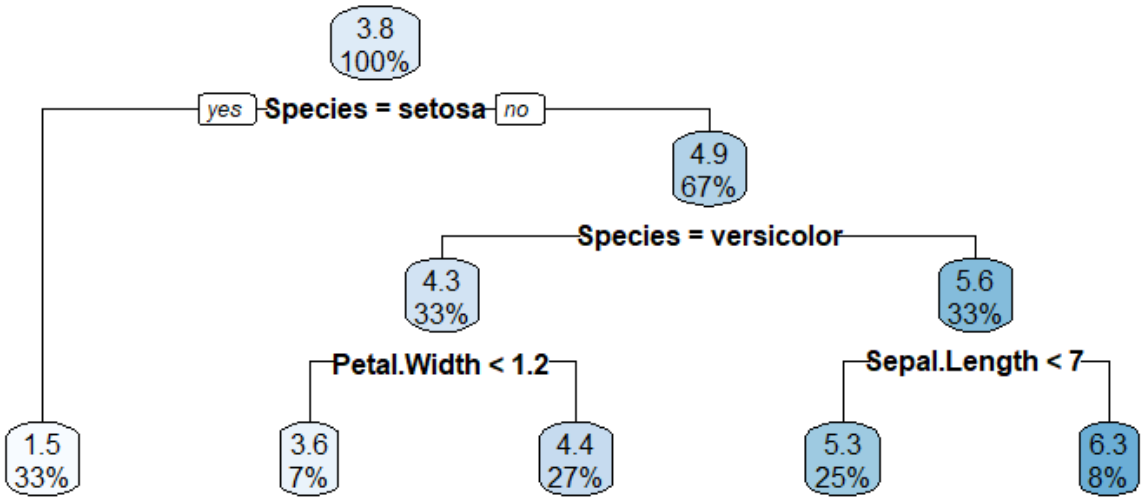


Figura: Árvore de Regressão

Exemplos

Para uma árvore de classificação, utilizaremos novamente o banco Iris, desta vez para classificar os elementos de acordo com todas as variáveis de Sépala e Pétala disponíveis.

```
procedure hpsplit data=sashelp.iris;
    class Species;
    model Species = Petal: Sepal:;
    grow gini;
    /* CHAID para ambos os tipos de variáveis */
    /* Para variáveis respostas categóricas, seguem:
    grow chisquare
    grow entropy
    grow fastchaid
    grow gini
    grow IGR (Information Gain Ratio)*/

    /* Para variáveis respostas contínuas, seguem:
    grow Ftest
    grow RSS
    grow Variance*/
    prune cc;
run;
```

Figura: Comando SAS para árvore de Decisão

Exemplos

Resultado gerado pelo procedimento; há algum problema?

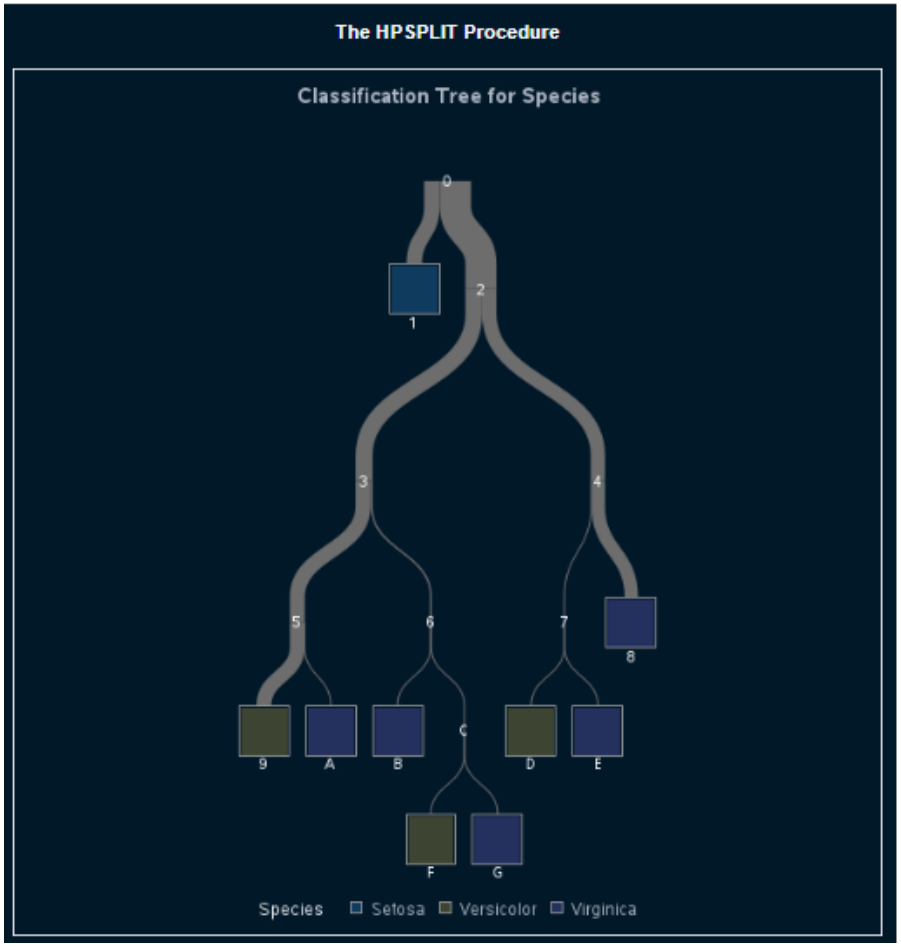


Figura: Árvore de Decisão por HPSPLIT

Exemplos

Verificamos que o parâmetro de custo-complexidade está indo a zero já a partir da divisão em 8 folhas, sendo anulado e se igualando à SQ(A).

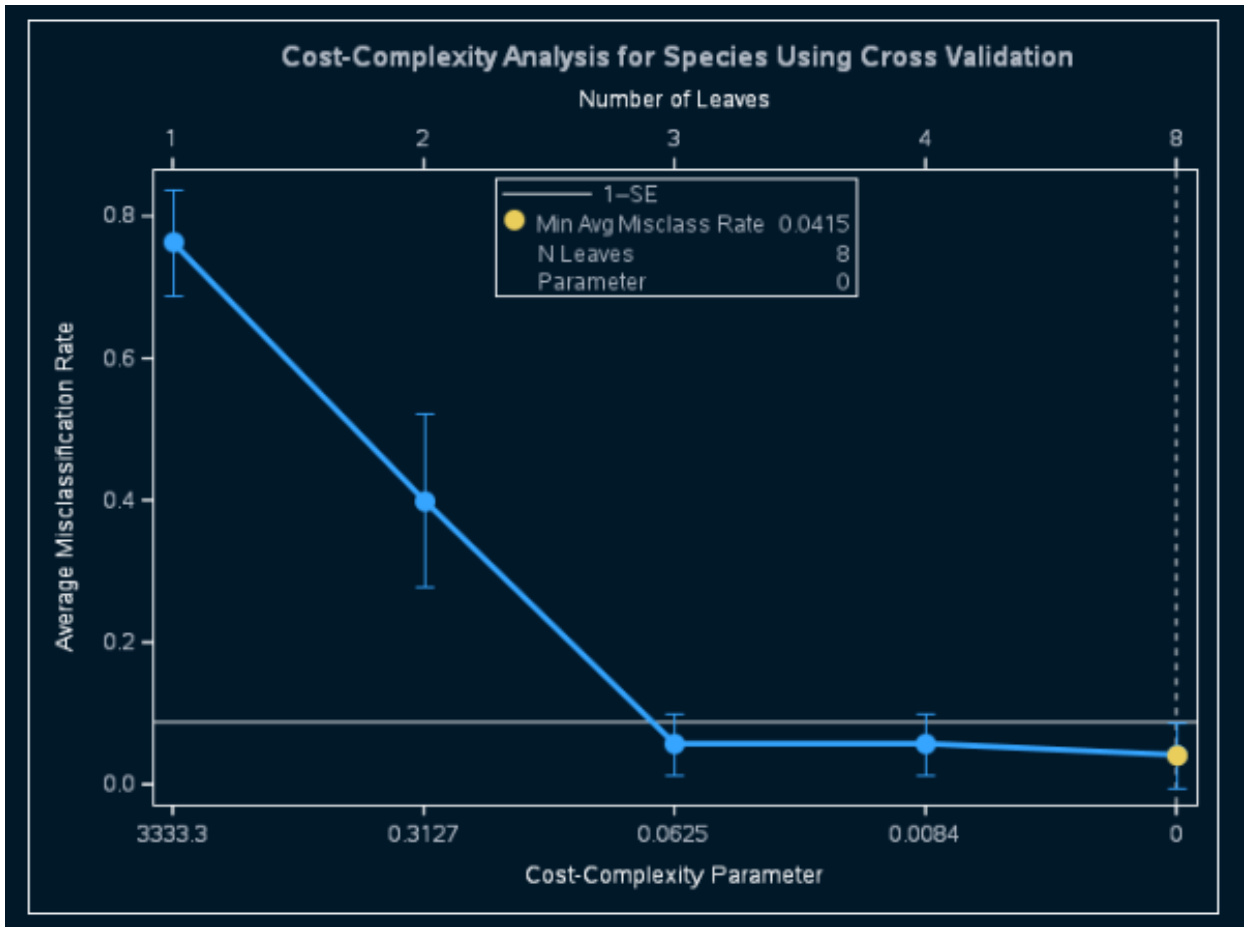


Figura: Gráfico de Custo-Complexidade via validação cruzada

Exemplos

Utilizando a opção Leaves=3, temos uma nova árvore.

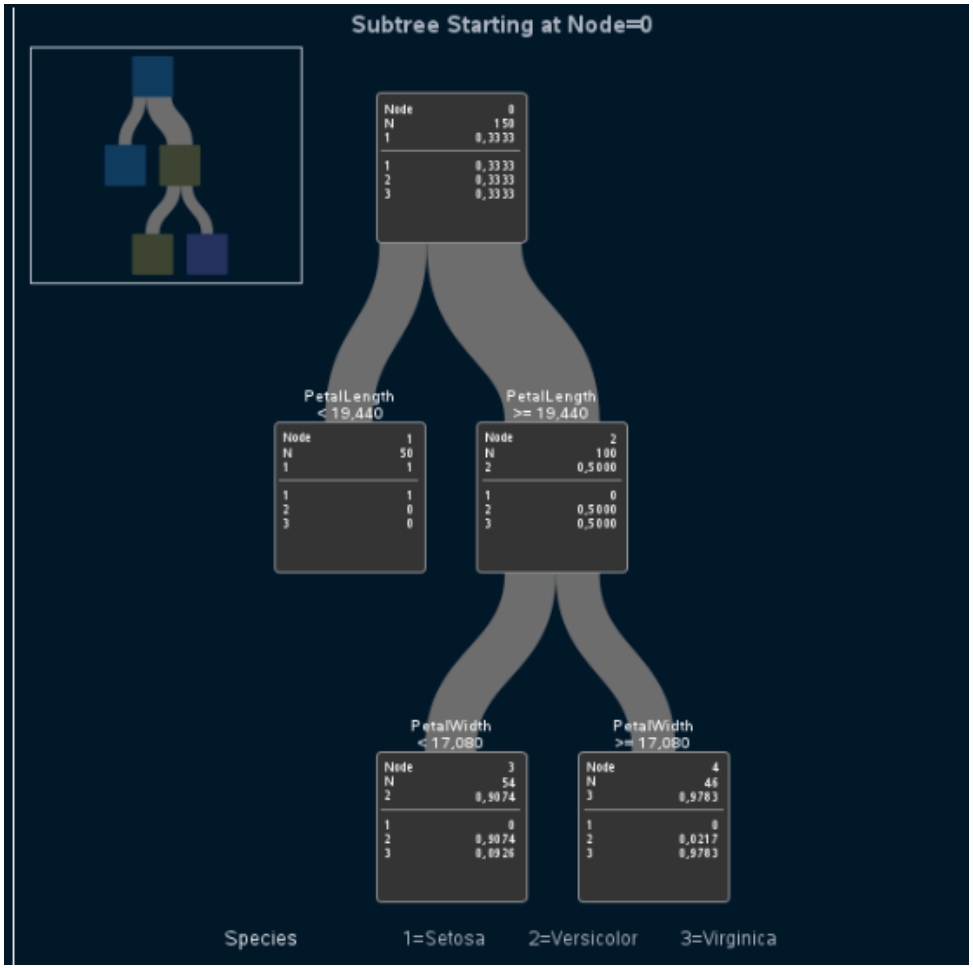


Figura: Árvore de Decisão com 3 folhas por HPSPLIT

Exemplos

Para a nova árvore, teremos a seguinte matriz de confusão.

Model-Based Confusion Matrix				
	Predicted			Error
Actual	Setosa	Versicolor	Virginica	Rate
Setosa	50	0	0	0.0000
Versicolor	0	49	1	0.0200
Virginica	0	5	45	0.1000

Figura: Matriz de Confusão para 3 folhas

Referências

- ARTES, Rinaldo e BARROSO, Lucia. **Métodos multivariados de análise estatística**. São Paulo: Blucher, 2023.
- IZENMAN, Alan J. **Modern Multivariate Statistical Techniques**. 2. ed. Springer, 2013.
- MORETTIN, Pedro A. e SINGER, Julio M. **Estatística e Ciência de Dados**. Universidade de São Paulo, 2021.
- WANG, Mao-Xin; HUANG, Duruo; WANG, Gang e LI, Dian-Qing. **SS-XGBoost: A Machine Learning Framework for Predicting Newmark Sliding Displacements of Slopes**. Journal of Geotechnical and Geoenvironmental Engineering, 2020.