Statistica Classification Problems

Alyson Benoni Matias Pereira - 2019102724

Classificação

- É o ato de prever a qual classe um item pertence;
- Alguns classificadores são binários (sim ou não; 0 ou 1; true ou false);
- Outros classificadores são multi-classes;
  - Permitem classificar um item em diversas categorias (cor, material...);
- Muito comum em Machine Learning;
  - Filtro de spam, reconhecimento de fala, reconhecimento de imagem;

Classificação

- Árvores de decisão;
  - Um dos métodos mais comuns de classificação;
  - Utilizado no software Statistica;
  - Utiliza uma estrutura de árvore para classificar os pontos;
  - São construídas de cima para baixo e os atributos no topo possuem maior impacto nas decisões;
  - Modela decisões complexas e a intepretação de resultados é bastante intuitiva;

Statistica Classification Problem

- O software Statistica possui um recurso importante para classificação;
  - No geral é simples para ser utilizado;
  - Para esse exemplo, será utilizado uma planilha que relaciona o tempo com a possibilidade de se ir jogar golfe ou não;
  - Primeiramente é necessário inserir na planilha os dados colhidos;
  - Nesse caso são 6 dados;
    - Dia;
    - Clima (ensolarado, chuvoso ou nublado);
    - Temperatura;
    - Umidade do ar;
    - Vento (sim ou não);
    - Jogar golf (sim ou não);

Statistica Classification Problem

- A planilha inicial para o problema é a seguinte;

	1 Dia	2 Clima	3 Temperatura	4 Umidade do Ar	5 Vento	6 Jogar Golf?
1	1	ensolarado	29,4	85	FALSE	nao
2	2	ensolarado	26,6	90	TRUE	nao
3	3	nublado	27,7	78	FALSE	sim
4	4	chuvoso	21,1	96	FALSE	sim
5	5	chuvoso	20	80	FALSE	sim
6	6	chuvoso	18,3	70	TRUE	nao
7	7	nublado	17,7	65	TRUE	sim
8	8	ensolarado	22,2	96	FALSE	nao
9	9	ensolarado	20,5	70	FALSE	sim
10	10	chuvoso	23,8	80	FALSE	sim
11	11	ensolarado	23,8	70	TRUE	sim
12	12	nublado	22,2	90	TRUE	sim
13	13	nublado	27,2	75	FALSE	sim
14	14	chuvoso	21,6	80	TRUE	nao

Statistica Classification Problem

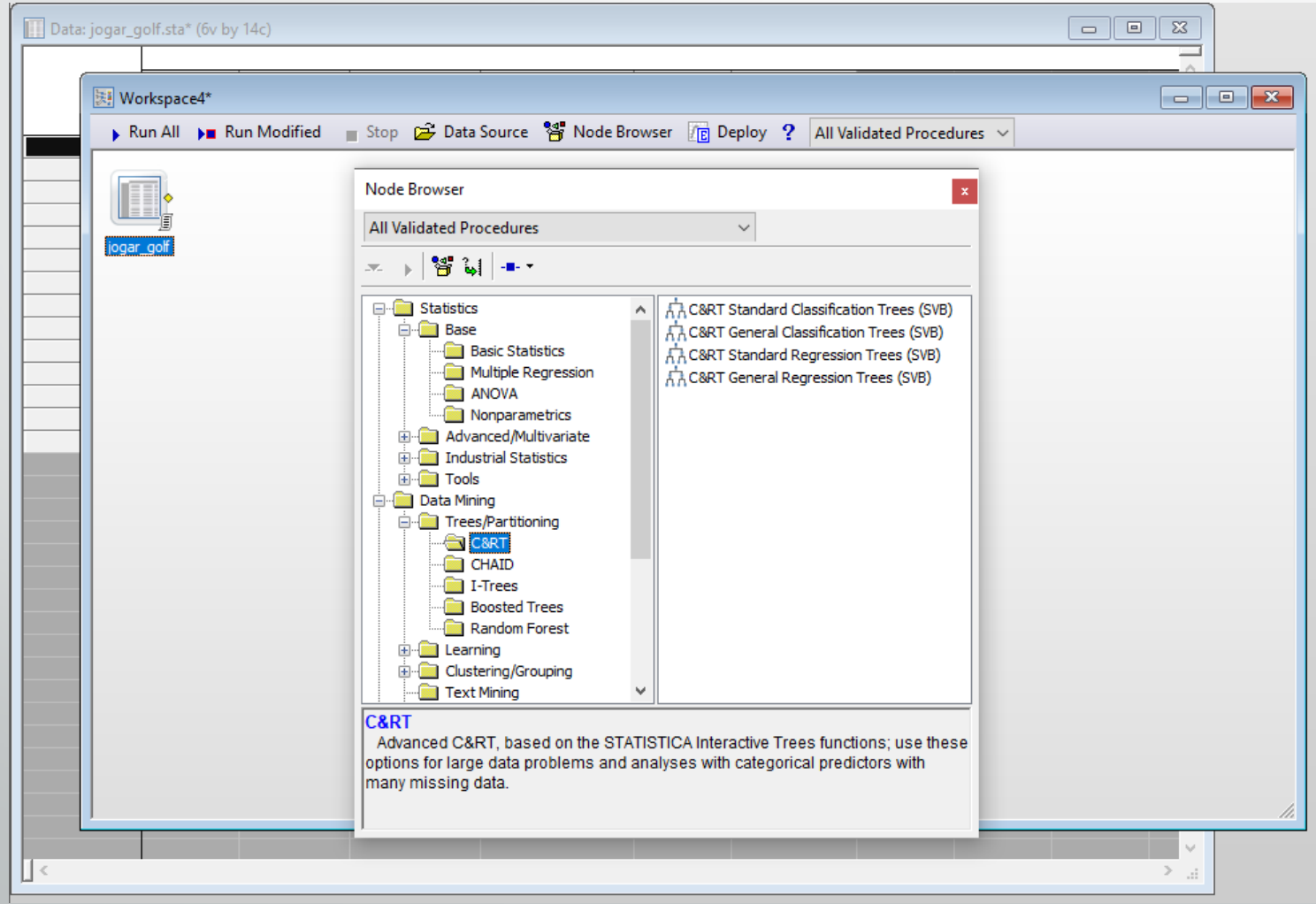
- Após a criação da planilha é necessário criar também um novo Workspace;
- Para isso vá em New -> Worskpace -> All validated Procedures;
- Escolha a planilha que será utilizada para a análise;
- Nesse exemplo a planilha possui o nome de “jogar_golf.sta”;

Statistica Classification Problem

- Agora dentro do workspace é necessário criar uma estrutura de árvore de decisão;
- Basta clicar em node browser -> Data Mining -> Trees/Partitioning -> C&RT;
  - Dentro dessa pasta haverá algumas estruturas;
  - A estrutura utilizada aqui será a C&RT Standard Classification Trees (SVB);

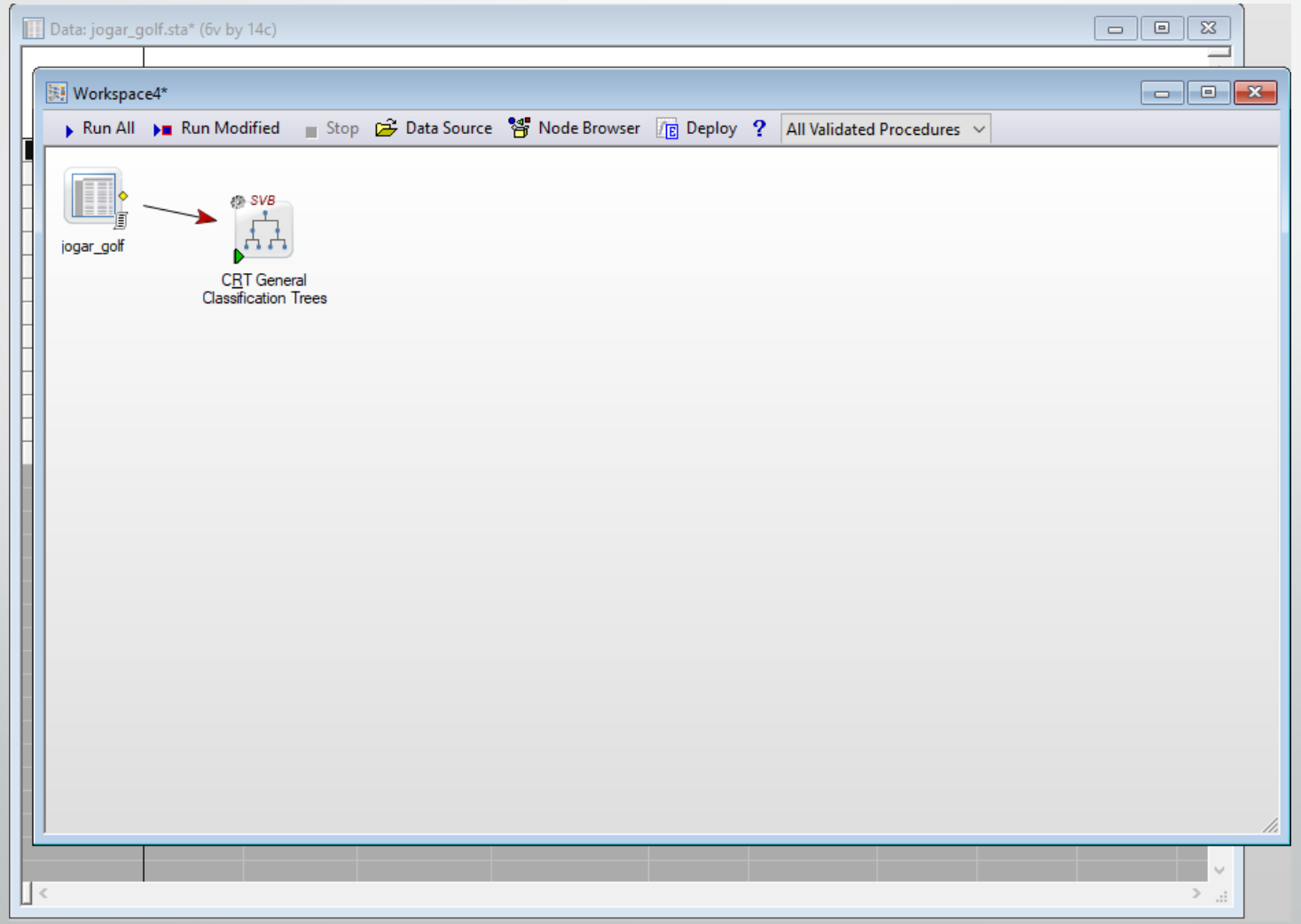
Statistica Classification Problem

- A estrutura utilizada aqui será a C&RT Standard Classification Trees (SVB);



Statistica Classification Problem

- Ao criar essa estrutura, ela aparecerá no workspace;

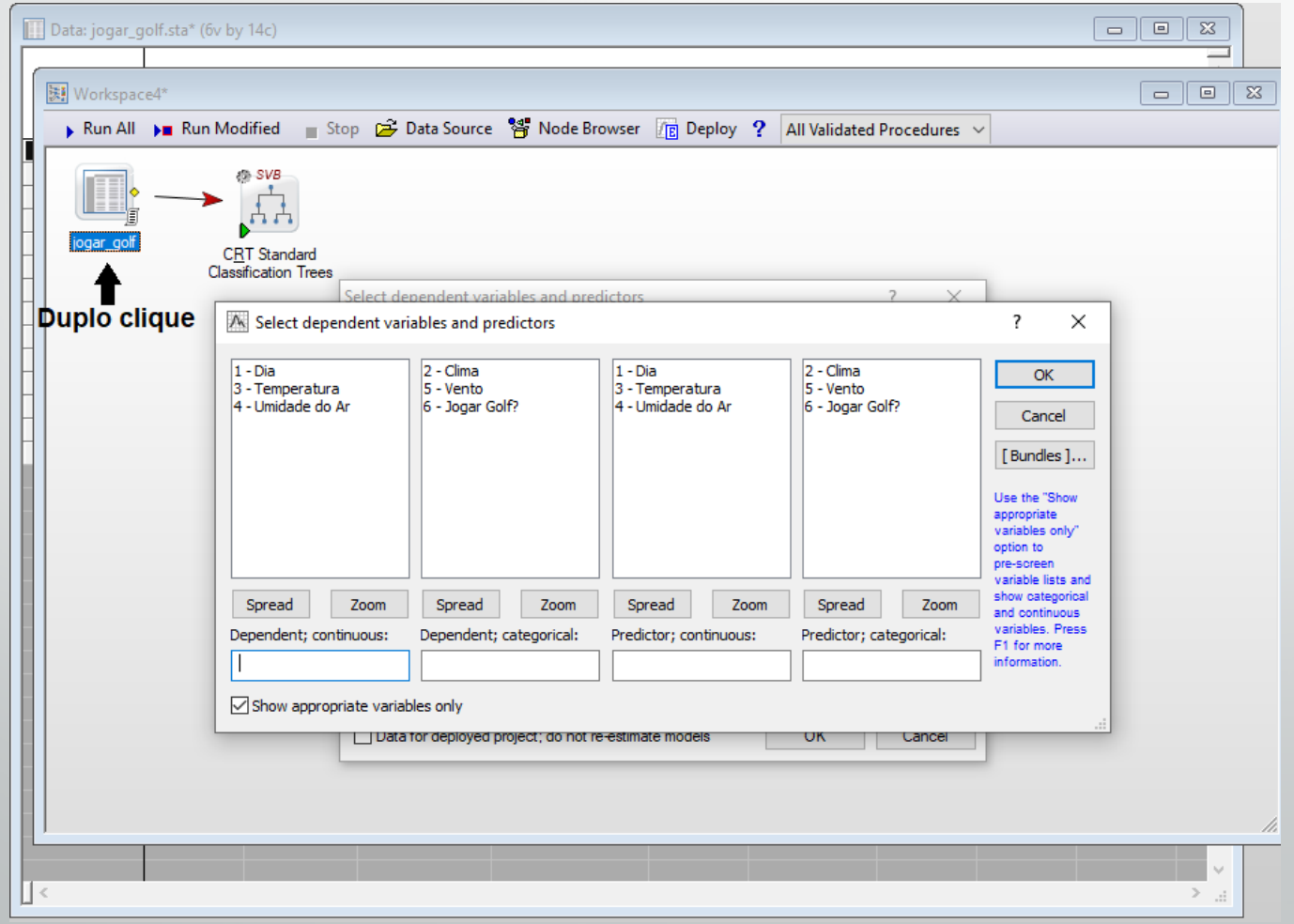


Statistica Classification Problem

- Após isso é necessário inserir as variáveis que se dividem em 4 grupos;
  - Dependentes e contínuas – variáveis contínuas (temperatura, tamanho, valor) que serão previstas;
  - Dependentes e categóricas – variáveis de categorias (sim ou não; quente, morno ou frio);
  - Preditoras e contínuas – variáveis contínuas que servirão pra prever as dependentes;
  - Preditoras e categóricas – variáveis de categorias que servirão para prever as dependentes;
- Para inserir essas variáveis, basta um duplo clique na planilha principal do worspace;

Statistica Classification Problem

- Ao dar um duplo clique em jogar_golf, uma nova janela será exibida;



Statistica Classification Problem

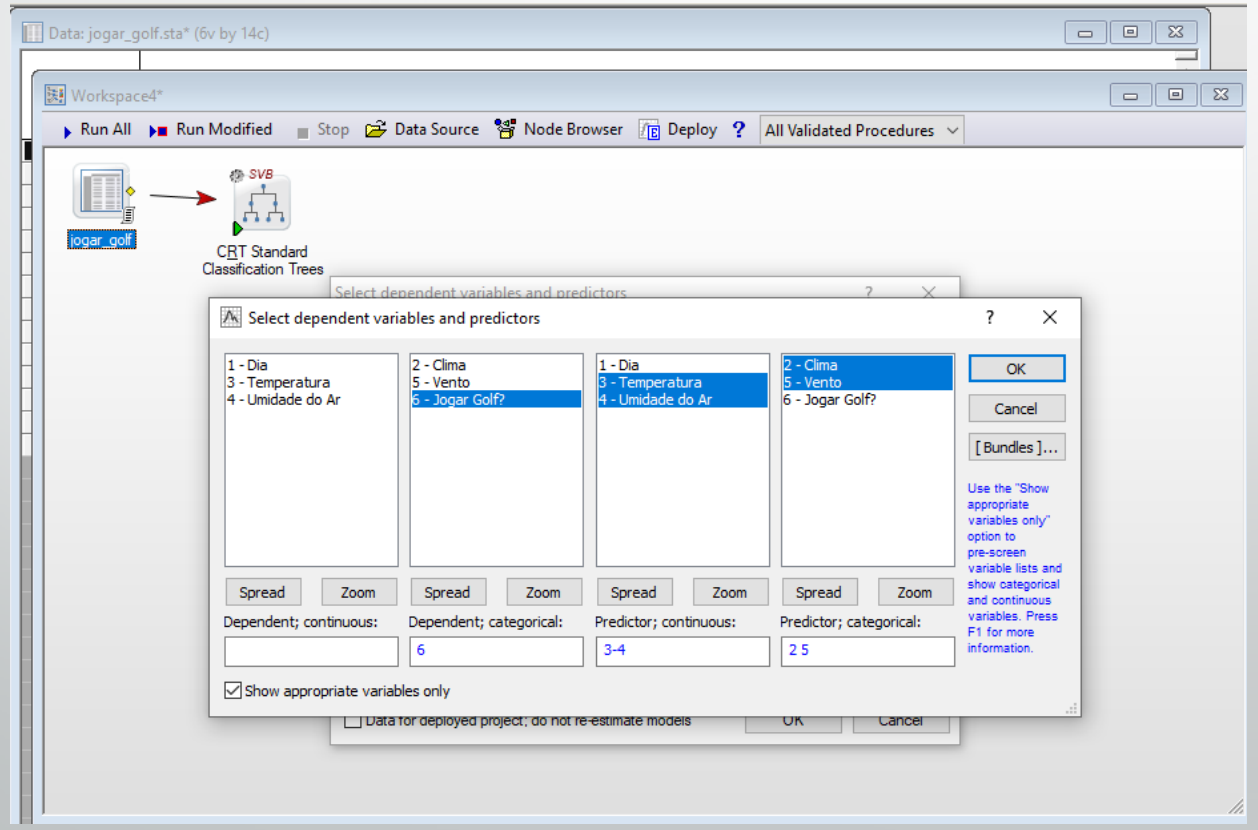
- Nessa nova janela existem 4 espaços para inserir as 4 variáveis mostradas no slide 10;
- Ao clicar no checkbox inferior na janela, somente variáveis apropriadas serão mostradas nos espaços;
- No primeiro espaço devem ser inseridas variáveis dependentes e contínuas;
  - Nesse exemplo não há nenhuma variável dependente e contínua;
- No segundo espaço devem ser inseridas variáveis dependentes e categóricas;
  - Nesse exemplo, a variável "jogar golf?" é uma variável categórica e dependente;
  - Essa variável será a resposta do problema;

Statistica Classification Problem

- No terceiro espaço devem ser inseridas as variáveis preditoras e contínuas;
  - Nesse exemplo existem duas variáveis preditoras e contínuas;
  - Temperatura e umidade do ar;
  - Ambas as variáveis vão influenciar se a resposta será jogar golfe ou não;
- No quarto espaço devem ser inseridas as variáveis preditoras e categóricas;
  - Nesse exemplo há duas variáveis categóricas e preditoras;
  - Clima e vento;
  - Ambas as variáveis vão influenciar a resposta de jogar golfe ou não;

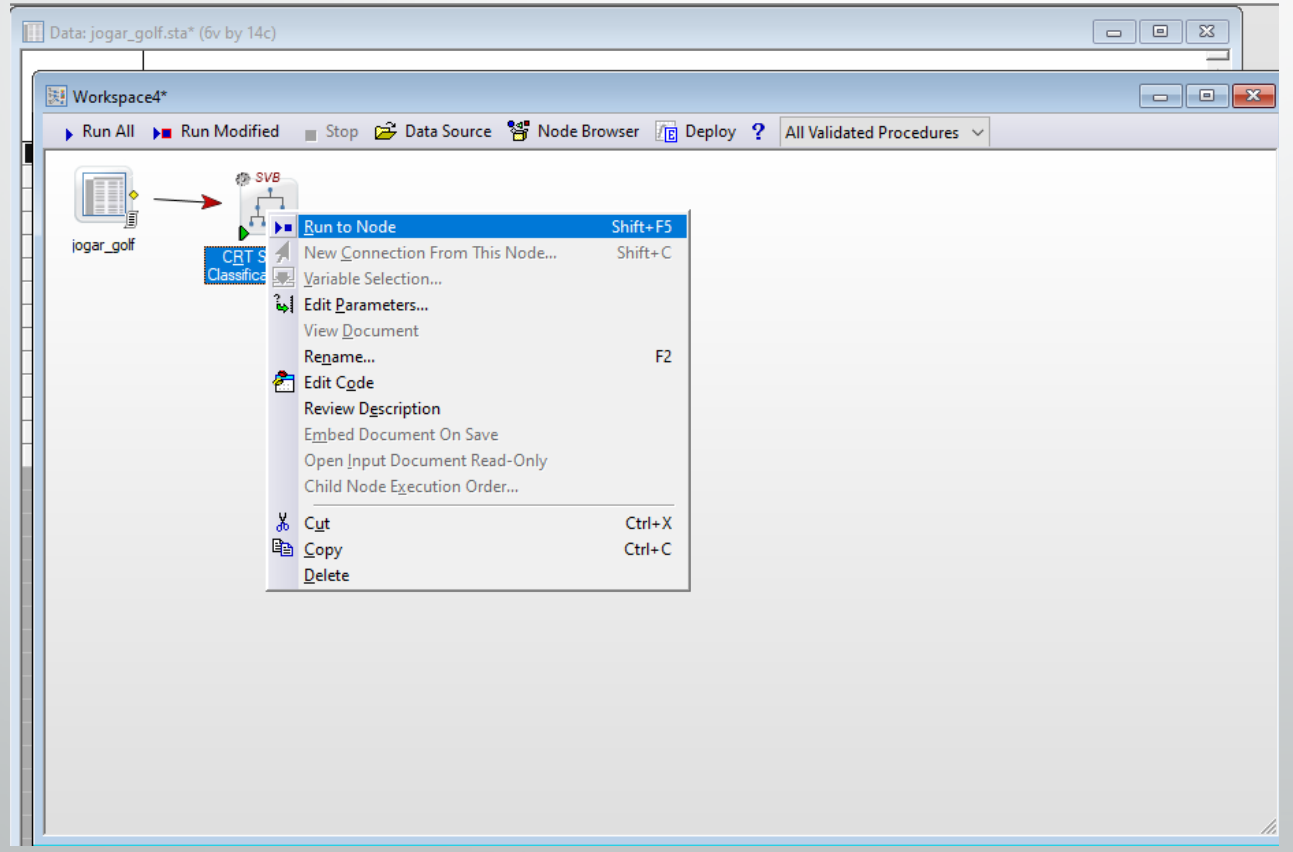
Statistica Classification Problem

- As variáveis selecionadas em cada espaço são destacadas;
- Os números das variáveis aparecem na caixa de texto embaixo dos espaços;
- Após todos os valores inseridos, pressione “OK”;



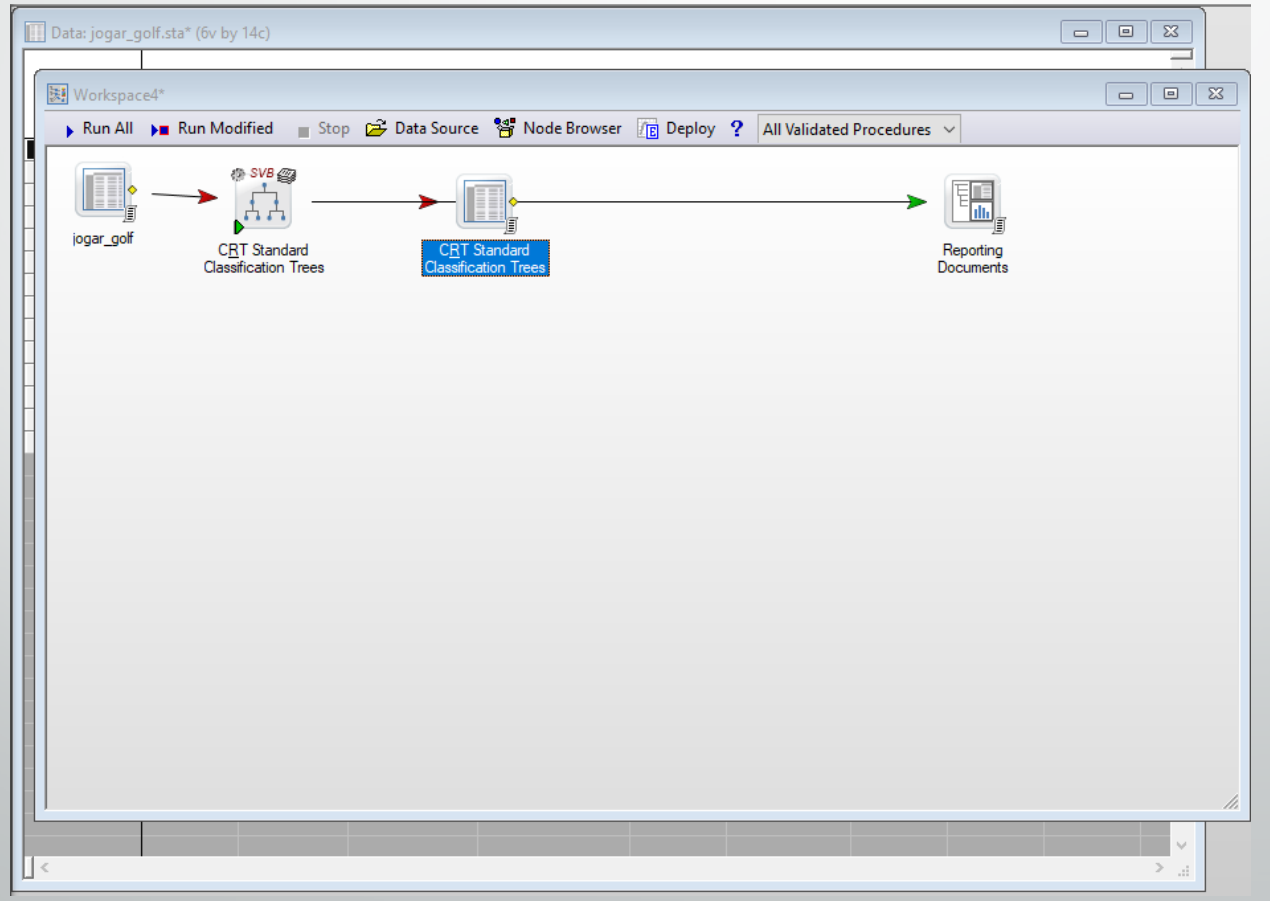
Statistica Classification Problem

- Assim é possível rodar a aplicação;
- Basta clicar com o botão direito sobre CRT Standard Classification Trees e clicar em "run to node"



Statistica Classification Problem

- O software gera dois relatórios;
  - O primeiro relatório é mostrado ao se clicar na planilha CRT Standard Classification Trees;

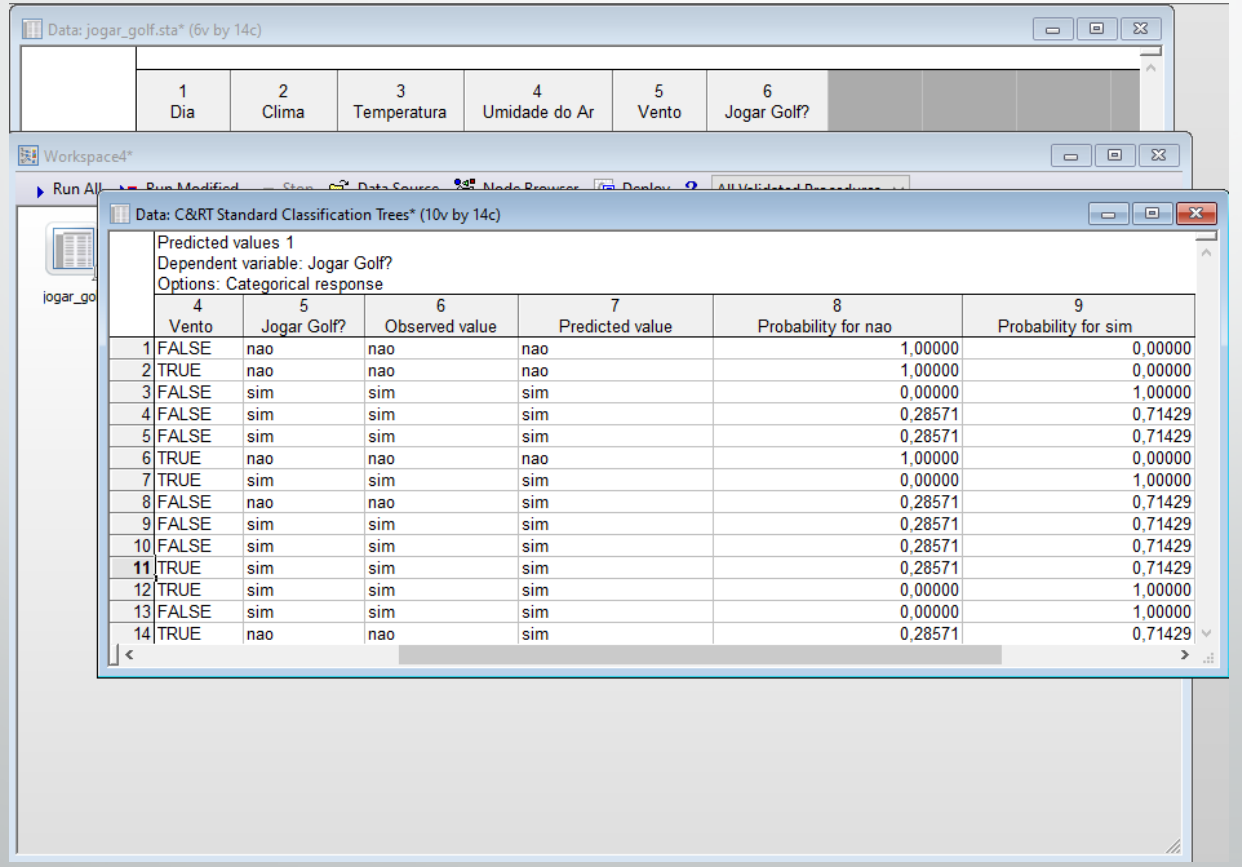


Statistica Classification Problem

- O primeiro relatório mostra uma planilha com todos os valores inseridos e os valores dependentes;
  - Entre os valores mais importantes estão:
  - Valor observado – mostra o valor observado da variável preditora;
  - Valor predito – mostra o valor predito da variável preditora;
  - Probabilidade para não – mostra a probabilidade de não se jogar golfe;
  - Probabilidade para sim – mostra a probabilidade de se jogar golfe;
- Quando a probabilidade para sim ou não é 1, existe certeza de que irá ou não jogar golfe;
- Quando a probabilidade de sim ou não é diferente de 1, são necessários outros dados para se tomar a decisão;
- O próximo slide mostra a planilha gerada pelo software;

Statistica Classification Problem

- Planilha gerada pelo software Statistica;



Data: jogar_golf.sta* (6v by 14c)

	1 Dia	2 Clima	3 Temperatura	4 Umidade do Ar	5 Vento	6 Jogar Golf?

Workspace4*

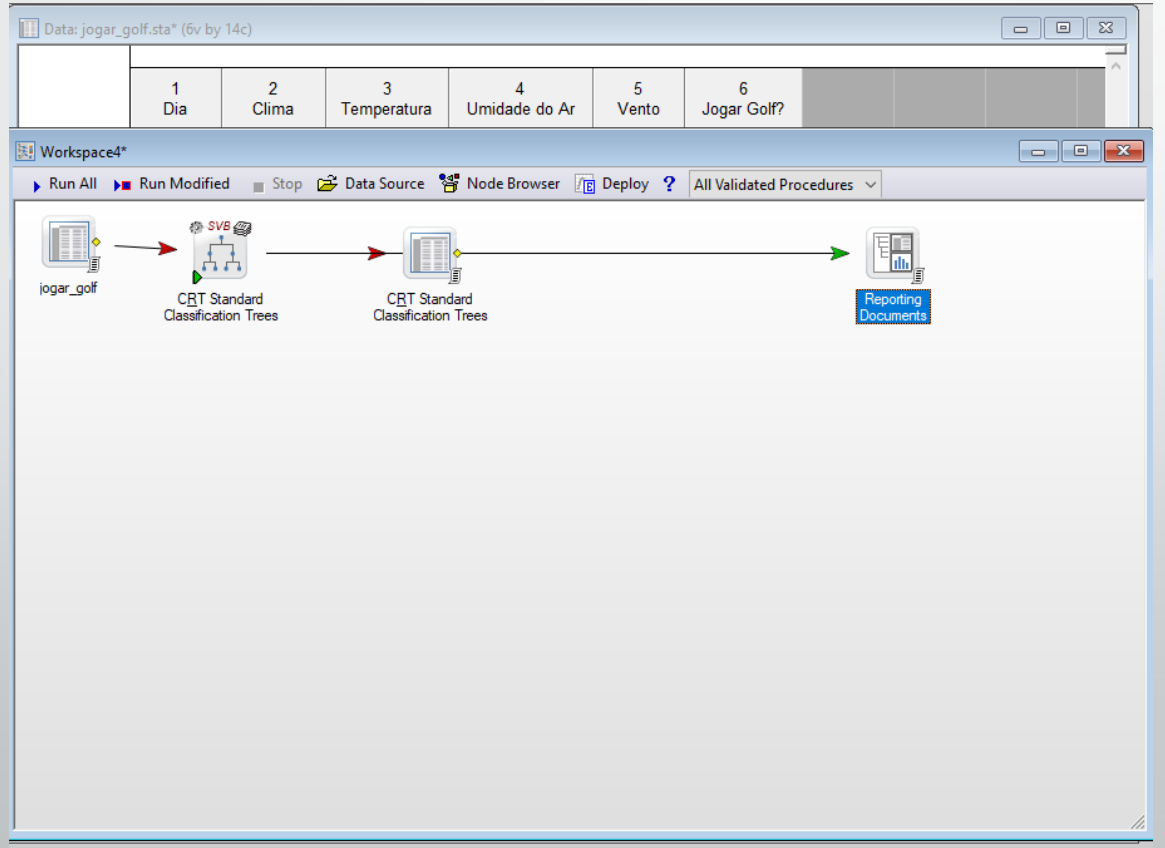
Data: C&RT Standard Classification Trees* (10v by 14c)

Predicted values 1
Dependent variable: Jogar Golf?
Options: Categorical response

	4 Vento	5 Jogar Golf?	6 Observed value	7 Predicted value	8 Probability for nao	9 Probability for sim
1	FALSE	nao	nao	nao	1,00000	0,00000
2	TRUE	nao	nao	nao	1,00000	0,00000
3	FALSE	sim	sim	sim	0,00000	1,00000
4	FALSE	sim	sim	sim	0,28571	0,71429
5	FALSE	sim	sim	sim	0,28571	0,71429
6	TRUE	nao	nao	nao	1,00000	0,00000
7	TRUE	sim	sim	sim	0,00000	1,00000
8	FALSE	nao	nao	sim	0,28571	0,71429
9	FALSE	sim	sim	sim	0,28571	0,71429
10	FALSE	sim	sim	sim	0,28571	0,71429
11	TRUE	sim	sim	sim	0,28571	0,71429
12	TRUE	sim	sim	sim	0,00000	1,00000
13	FALSE	sim	sim	sim	0,00000	1,00000
14	TRUE	nao	nao	sim	0,28571	0,71429

Statistica Classification Problem

- O software gera dois relatórios;
  - O segundo relatório é mostrado ao se clicar na planilha Reporting Documents;

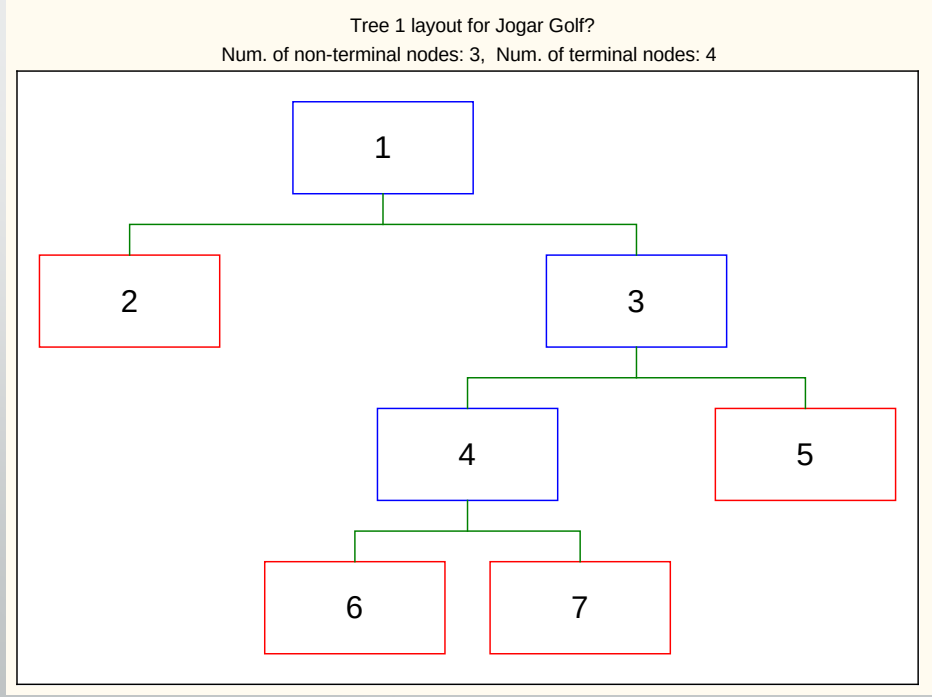


Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Entre eles os mais importantes são;
  - Tree 1 layout for Jogar Golf?;
  - Tree 1 graph for Jogar Golf?;
  - Tree structure 1;
  - Result of terminal nodes;
  - Classification matrix 1 Dependent variable: Jogar Golf? Options: Categorical response, Tree number 1, Analysis sample

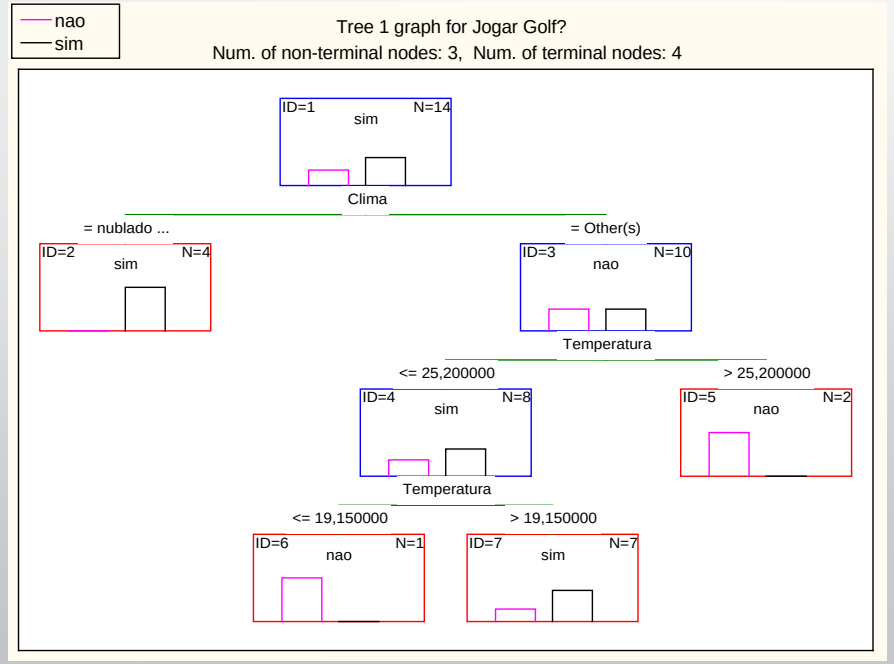
Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Tree 1 layout for jogar Golf;
  - Mostra os nós terminais e não terminais;
  - Nós em vermelho são terminais, nós em azul não são terminais;



Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Tree 1 graph for jogar Golf;
  - Mostra os nós terminais e não terminais;
  - Mostra também a probabilidade da resposta ser sim ou não;



Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Tree structure 1;
  - Mostra uma planilha com a estrutura da árvore;

Tree structure 1 (jogar_golf)
Dependent variable: Jogar Golf?
Options: Categorical response, Tree number 1

Node #	Left branch	Right branch	Size of node	N in class nao	N in class sim	Selected category	Split variable	Split constant	Split category
1	2	3	14	5	9	sim	Clima		nublado
2			4	0	4	sim			
3	4	5	10	5	5	nao	Temperatura	25,2	
4	6	7	8	3	5	sim	Temperatura	19,2	
6			1	1	0	nao			
7			7	2	5	sim			
5			2	2	0	nao			

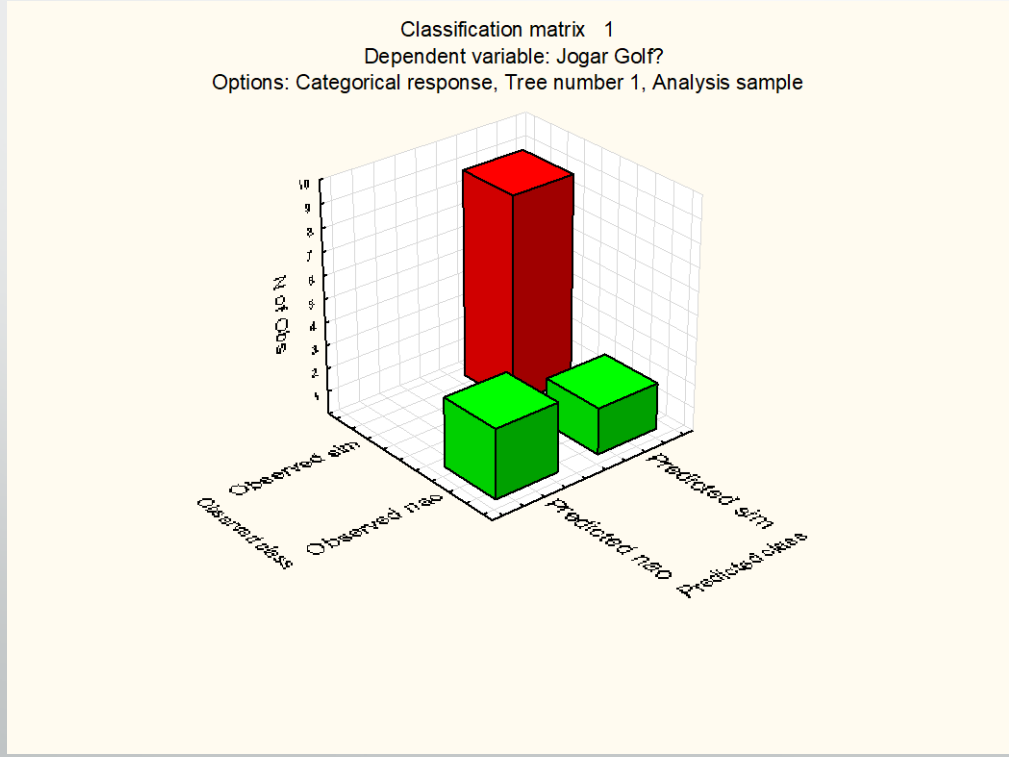
Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Result of terminal nodes;
  - Mostra uma planilha com os resultados dos nós terminais;

Node #	Result of terminal nodes 1 (jogar_golf) Dependent variable: Jogar Golf? Options: Categorical response, Tree number 1 **Class nao**	Class sim	Gain
2	0	4	4,000000
6	1	0	1,000000
7	2	5	7,000000
5	2	0	2,000000

Statistica Classification Problem

- O segundo relatório possui 8 documentos;
  - Classification matrix 1 Dependent variable: Jogar Golf? Options: Categorical response, Tree number 1, Analysis sample;
  - Mostra um gráfico com o resumo das respostas preditas e observadas;



Conclusão

- Nesse exemplo é possível prever quando uma pessoa irá jogar golfe ou não;
- Olhando o relatório do slide 22 é possível perceber que;
  - Se o dia estiver nublado, existe 100% de chance de se jogar golfe;
  - Se não estiver nublado, vai depender da temperatura;
    - Se a temperatura for maior que 25,2 graus, existe 100% de chance de não se jogar golfe;
    - Se a temperatura for menor que 25,2 ainda existe uma chance de se jogar golfe;
      - Se a temperatura for menor que 19,15 graus, existe 100% de chance de não jogar golfe;
      - Se a temperatura for maior que 19,15 existe uma chance de se jogar golfe;