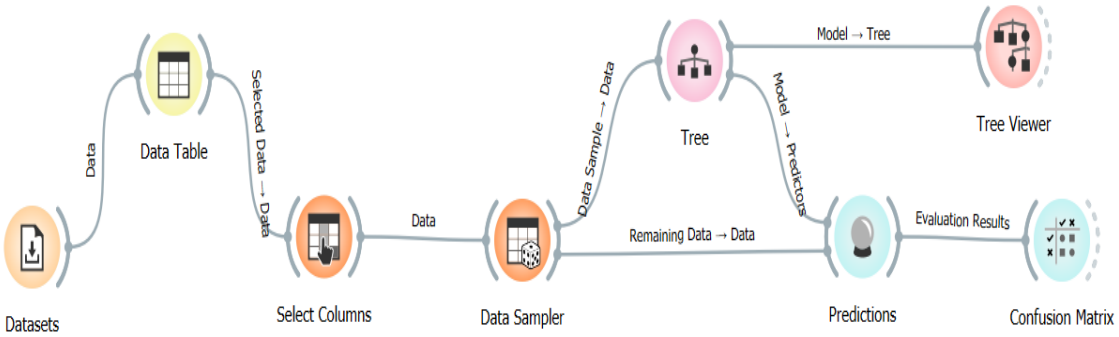
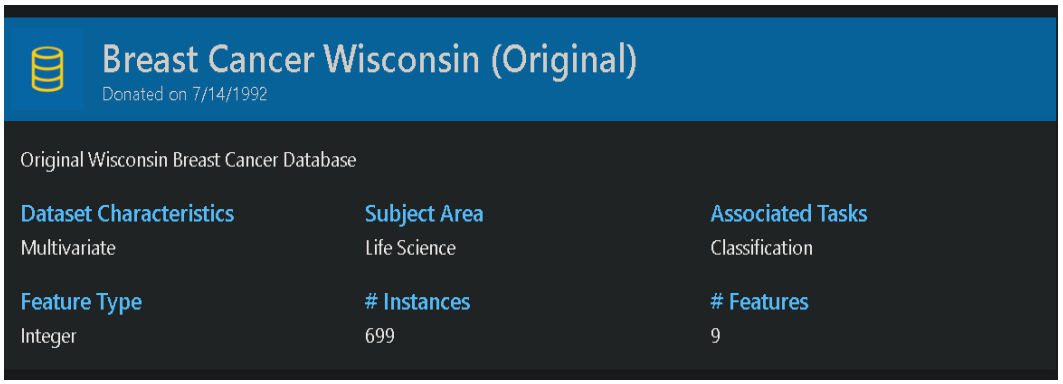
Disciplina: Introdução à Computação
Professor: Adriano Lorena Inacio de Oliveira
Título: Relatório - Projeto Inteligência Artificial
Grupo: Guilherme Siqueira, Walter Crasto, Ivo Neto, Pedro Fischer e Niltton Szpak

Neste trabalho, mostraremos o desenvolvimento da criação de um algoritmo de árvore de decisão utilizando o Orange Data Mining. Inicialmente, escolhemos o conteúdo que seria estudado pela nossa árvore, ou seja, buscamos um dataset que possuísse informações suficientes para o aprendizado da máquina. O dataset escolhido foi sobre a identificação do caráter do câncer (benigno ou maligno), que possui 699 exemplos, de forma que dividimos 80% para o aprendizado do nosso algoritmo e 20% para testes e previsões.

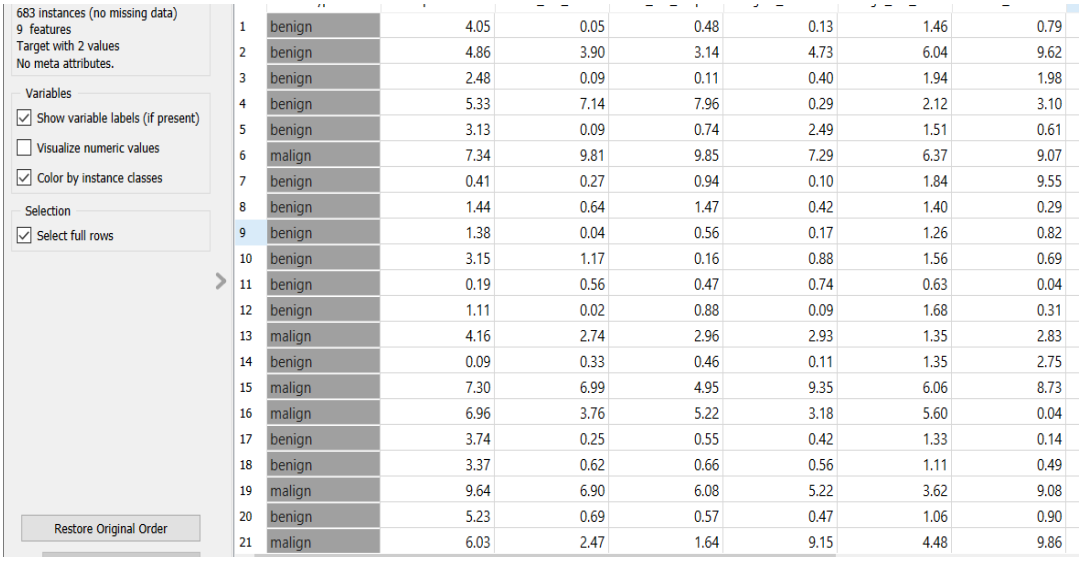


No Orange Data Mining, criamos esse esquema, que funciona da seguinte forma:

1) Primeiro, importamos o nosso dataset escolhido:
   Breast Cancer Wisconsin (Original) - UCI Machine Learning Repository

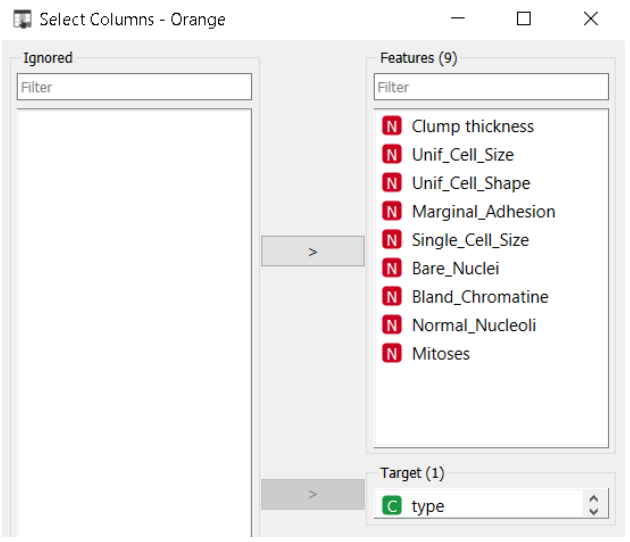


2) Em seguida, na aba de Data Table, podemos verificar os dados de forma mais organizada, no formato de tabela:

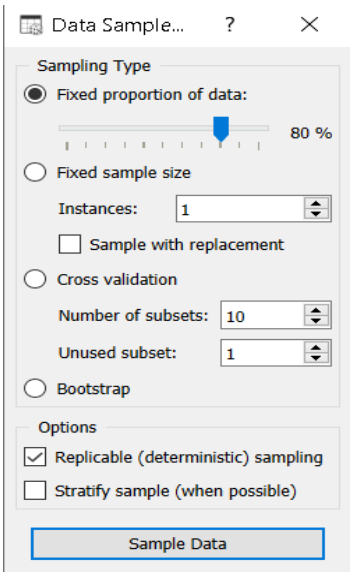


1	benign	4.05	0.05	0.48	0.13	1.46	0.79
2	benign	4.86	3.90	3.14	4.73	6.04	9.62
3	benign	2.48	0.09	0.11	0.40	1.94	1.98
4	benign	5.33	7.14	7.96	0.29	2.12	3.10
5	benign	3.13	0.09	0.74	2.49	1.51	0.61
6	malign	7.34	9.81	9.85	7.29	6.37	9.07
7	benign	0.41	0.27	0.94	0.10	1.84	9.55
8	benign	1.44	0.64	1.47	0.42	1.40	0.29
9	benign	1.38	0.04	0.56	0.17	1.26	0.82
10	benign	3.15	1.17	0.16	0.88	1.56	0.69
11	benign	0.19	0.56	0.47	0.74	0.63	0.04
12	benign	1.11	0.02	0.88	0.09	1.68	0.31
13	malign	4.16	2.74	2.96	2.93	1.35	2.83
14	benign	0.09	0.33	0.46	0.11	1.35	2.75
15	malign	7.30	6.99	4.95	9.35	6.06	8.73
16	malign	6.96	3.76	5.22	3.18	5.60	0.04
17	benign	3.74	0.25	0.55	0.42	1.33	0.14
18	benign	3.37	0.62	0.66	0.56	1.11	0.49
19	malign	9.64	6.90	6.08	5.22	3.62	9.08
20	benign	5.23	0.69	0.57	0.47	1.06	0.90
21	malign	6.03	2.47	1.64	9.15	4.48	9.86

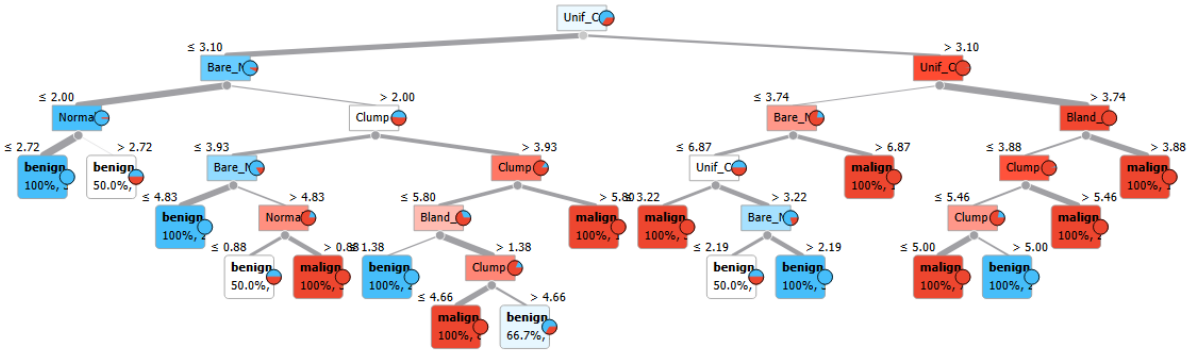
3) Como terceiro passo, partimos para a seção de Select Columns, em que escolhemos os atributos que serão utilizados como *features* e o *target* da árvore, ou seja, o que será usado como parte do estudo e a variável que será o objetivo do nosso estudo. Nesse caso, estamos em busca de prever o tipo do tumor: benigno ou maligno. E como ferramentas de avaliação usaremos o tamanho da célula, formato, número de mitoses, espessura do tumor, etc:



4) Em seguida, usando a aba do Data Sampler, dividimos a nossa amostra em Treinamento e Teste, fazendo uma divisão de 80% para treino e 20% para teste.

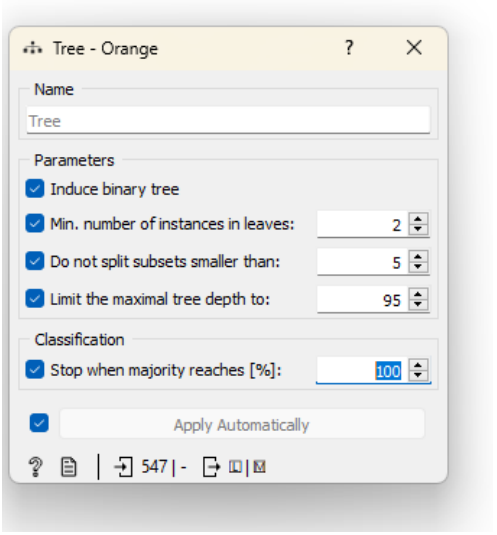


5) Agora, após o tratamento dos dados, podemos criar a nossa árvore, que pode ser observada usando o Tree Viewer:



Ou seja, temos uma árvore que vai tomando decisões em cada nó e chega na conclusão se o tumor é benigno ou maligno. Nessa árvore, temos 33 nós e 17 folhas, ou seja, existem 33 nós a serem percorridos e 17 “destinos finais” em que se existe uma conclusão.

Vale ressaltar que essa árvore foi criada utilizando os seguintes parâmetros:



Ou seja, no primeiro nós induzimos a criação de uma árvore que se divide de forma binária, gerando 2 nós-filhos de cada nó. Em seguida, temos o mínimo de casos que precisam ser colocados em cada folha, de forma que precisamos de no mínimo 2 casos da amostra de teste se encaixando naquela folha para a criação dela. Depois, temos a proibição da divisão de nós com menos de 5 casos em questão, ou seja, se houver menos de 5 casos, aquele ramo da árvore se encerra ali mesmo. E o último parâmetro é a limitação de níveis da árvore, em que nesse caso, colocamos 95, mas ela não chega atingir isso.

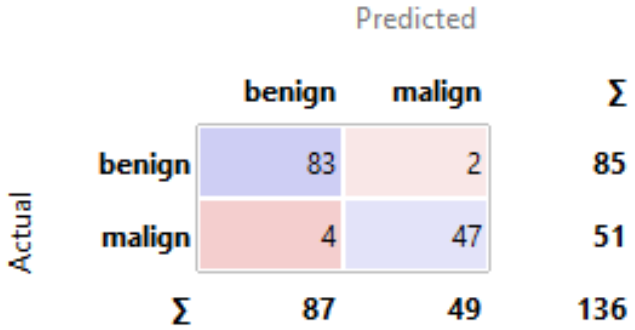
6) Depois da árvore de decisão montada, podemos testar o funcionamento dela na aba Predictions, em que ligamos nossa base de dados e nosso algoritmo e observamos o rendimento dele.

	Tree	error	type	Clump thickness	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adhesion	Single_Cell_Size	Bare_Nuclei	Bland_Chromatin	Normal_Nucleoli	Mitoses
1	1.00 : 0.00 → ben...	0.000	benign	0.37	1.83	1.17	0.11	1.65	0.16	0.68	0.39	0.57
2	1.00 : 0.00 → ben...	0.000	benign	4.68	0.30	0.58	0.31	1.55	0.62	1.34	0.50	0.44
3	0.00 : 1.00 → mal...	0.000	malign	2.24	5.85	5.32	5.56	4.99	9.69	5.92	7.65	2.85
4	1.00 : 0.00 → ben...	0.000	benign	0.50	1.81	1.09	0.87	1.74	0.20	0.43	0.58	0.42
5	1.00 : 0.00 → ben...	0.000	benign	3.44	0.54	0.65	0.93	1.62	1.06	2.93	1.90	0.26
6	1.00 : 0.00 → ben...	0.000	benign	5.34	0.26	0.34	0.50	1.64	0.83	1.80	0.01	0.51
7	0.00 : 1.00 → mal...	0.000	malign	9.67	9.95	9.97	7.14	5.18	0.38	7.97	8.08	0.33
8	0.50 : 0.50 → ben...	0.500	benign	3.87	2.05	0.66	0.52	1.18	0.10	3.94	7.60	0.14
9	1.00 : 0.00 → ben...	0.000	benign	1.12	0.28	0.74	0.76	1.52	0.87	2.60	0.41	0.70
10	1.00 : 0.00 → ben...	0.000	benign	4.18	0.92	2.97	0.89	1.36	0.73	2.93	0.56	0.57
11	1.00 : 0.00 → ben...	0.000	benign	4.75	0.75	0.94	0.28	1.28	0.29	0.58	0.90	0.62
12	0.50 : 0.50 → ben...	0.500	benign	4.20	3.71	4.45	0.11	7.34	0.08	2.17	5.64	0.53
13	1.00 : 0.00 → ben...	0.000	benign	5.00	0.02	0.11	0.16	1.80	0.39	2.41	0.08	0.96
14	0.00 : 1.00 → mal...	0.000	malign	7.92	3.50	3.92	0.64	1.73	8.96	2.24	2.52	0.58
15	1.00 : 0.00 → ben...	0.000	benign	2.29	0.72	0.17	0.30	1.85	0.18	0.89	0.99	0.62
16	0.00 : 1.00 → mal...	0.000	malign	7.71	3.20	3.52	0.47	5.56	9.68	1.91	4.19	1.66
17	0.00 : 1.00 → mal...	0.000	malign	4.99	9.10	9.06	9.49	4.87	1.02	7.93	4.49	0.10
18	1.00 : 0.00 → ben...	0.000	benign	0.60	0.18	0.55	0.41	1.27	0.23	2.02	0.80	0.53
19	1.00 : 0.00 → ben...	0.000	benign	3.63	1.96	1.76	0.24	1.61	0.63	1.93	0.18	0.23
20	0.00 : 1.00 → mal...	0.000	malign	9.34	9.41	9.13	5.30	7.29	3.96	7.92	4.12	0.40
21	0.00 : 1.00 → mal...	0.000	malign	3.10	9.73	7.91	4.64	3.09	0.11	9.62	0.94	0.26
22	0.00 : 1.00 → mal...	0.000	malign	6.94	4.34	5.67	2.18	2.58	7.54	6.25	3.05	0.56
23	1.00 : 0.00 → ben...	0.000	benign	3.13	0.09	0.74	2.49	1.51	0.61	2.43	0.80	0.36
24	1.00 : 0.00 → ben...	0.000	benign	0.26	0.68	0.20	0.32	1.37	0.57	0.89	0.53	0.93
25	1.00 : 0.00 → ben...	0.000	benign	4.66	0.46	0.88	0.41	1.17	0.83	2.20	0.56	0.26
26	0.00 : 1.00 → mal...	0.000	malign	9.51	2.96	4.39	0.98	9.80	4.81	2.33	9.12	1.21
27	0.00 : 1.00 → mal...	0.000	malign	8.12	7.86	7.79	8.08	5.21	2.20	3.68	0.88	0.64
28	0.00 : 1.00 → mal...	0.000	malign	9.52	9.89	9.93	6.59	9.06	9.07	7.63	1.64	0.25
29	1.00 : 0.00 → ben...	0.000	benign	0.50	0.41	0.31	0.64	0.49	0.82	1.22	0.21	0.37
30	0.00 : 1.00 → mal...	0.000	malign	6.61	7.18	6.12	1.57	3.68	7.71	2.66	7.26	1.98
31	1.00 : 0.00 → ben...	0.000	benign	1.49	0.91	0.97	1.46	1.91	0.92	2.03	0.15	0.02
32	0.00 : 1.00 → mal...	0.000	malign	9.38	9.42	8.05	2.75	6.41	4.97	2.51	4.13	0.43
33	1.00 : 0.00 → ben...	0.000	benign	3.97	0.52	0.76	0.05	1.66	0.08	1.92	0.05	0.84
34	1.00 : 0.00 → ben...	0.000	benign	1.68	0.44	0.41	0.57	1.62	0.35	1.38	0.92	0.11
35	1.00 : 0.00 → ben...	0.000	benign	1.95	2.71	0.92	0.00	1.29	0.50	1.79	0.05	0.88
36	0.00 : 1.00 → mal...	0.000	malign	8.47	7.09	7.28	4.87	5.98	1.44	3.42	9.57	3.23
37	1.00 : 0.00 → ben...	0.000	benign	0.15	0.56	0.74	0.55	1.26	0.41	0.43	0.77	0.54
38	1.00 : 0.00 → ben...	0.000	benign	2.34	0.77	0.12	0.62	1.45	0.29	2.00	0.71	0.62

Show perfomance scores Target class: (Average over classes)

Model	AUC	CA	F1	Precision	Recall
Tree	0.966	0.956	0.956	0.956	0.956

Em que nas informações da árvore, podemos observar que obtemos uma precisão de 0.956, ou seja, 95,6% de precisão na análise dos dados de teste. Para observar de maneira mais clara, podemos observar uma matriz de performance para avaliar o desempenho do nosso algoritmo:

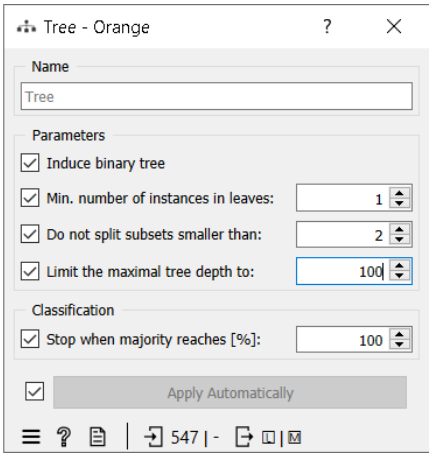


Em que temos:

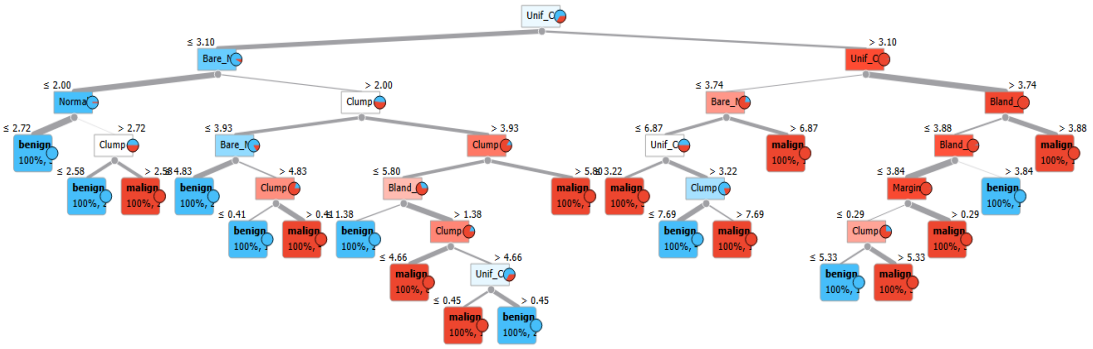
- Dos 85 benignos, acertamos 83
- Dos 51 malignos, acertamos 47

Em seguida, iremos alterar alguns valores dos parâmetros para observar o comportamento da árvore e como essas mudanças alteram o desempenho da mesma.

1) Vamos alterar os parâmetros da árvore em busca de uma precisão extremamente grande, de forma que as folhas poderão representar casos isolados e os nós podem se dividir mesmo que tenham poucos casos a serem resolvidos. Além disso, iremos colocar o limite de níveis da árvore em 100, dando liberdade para ela crescer bastante.

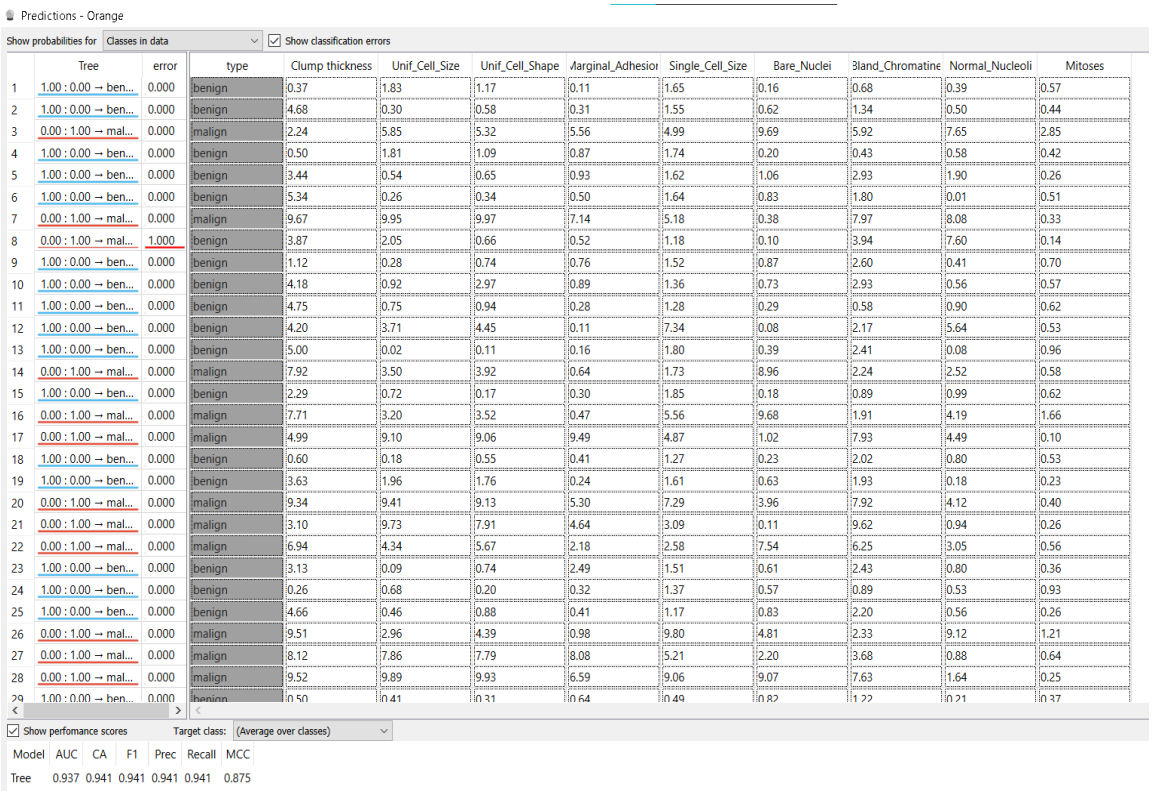


Utilizando esses parâmetros vamos obter a árvore:



Que é claramente uma árvore maior (temos 39 nós e 20 folhas) que busca o veredito utilizando diversos detalhes. O que se espera dos resultados de uma árvore maior? Que a precisão seja superior a todas as outras possíveis árvores, visto que estaríamos analisando minuciosamente cada atributo que influencia no veredito final. Mas, por incrível que pareça, a precisão diminui

em relação aos parâmetros utilizados na árvore do início do projeto, tendo apenas 94,1% de precisão (menos que o 95,6% obtido anteriormente)



Predictions - Orange

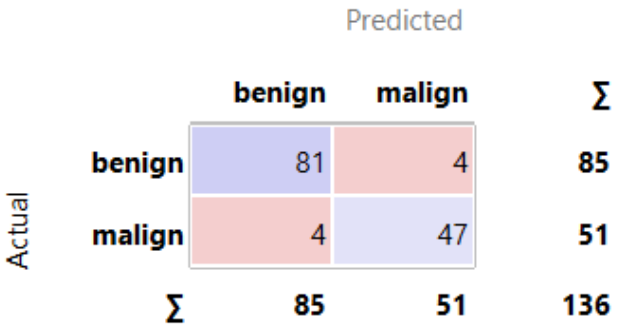
Show probabilities for: Classes in data | Show classification errors

	Tree	error	type	Clump thickness	Unif_Cell_Size	Unif_Cell_Shape	Marginal_Adhesion	Single_Cell_Size	Bare_Nuclei	Bland_Chromatine	Normal_Nucleoli	Mitoses
1	1.00 : 0.00 → ben...	0.000	benign	0.37	1.83	1.17	0.11	1.65	0.16	0.68	0.39	0.57
2	1.00 : 0.00 → ben...	0.000	benign	4.68	0.30	0.58	0.31	1.55	0.62	1.34	0.50	0.44
3	0.00 : 1.00 → mal...	0.000	malign	2.24	5.85	5.32	5.56	4.99	9.69	5.92	7.65	2.85
4	1.00 : 0.00 → ben...	0.000	benign	0.50	1.81	1.09	0.87	1.74	0.20	0.43	0.58	0.42
5	1.00 : 0.00 → ben...	0.000	benign	3.44	0.54	0.65	0.93	1.62	1.06	2.93	1.90	0.26
6	1.00 : 0.00 → ben...	0.000	benign	5.34	0.26	0.34	0.50	1.64	0.83	1.80	0.01	0.51
7	0.00 : 1.00 → mal...	0.000	malign	9.67	9.95	9.97	7.14	5.18	0.38	7.97	8.08	0.33
8	0.00 : 1.00 → mal...	1.000	benign	3.87	2.05	0.66	0.52	1.18	0.10	3.94	7.60	0.14
9	1.00 : 0.00 → ben...	0.000	benign	1.12	0.28	0.74	0.76	1.52	0.87	2.60	0.41	0.70
10	1.00 : 0.00 → ben...	0.000	benign	4.18	0.92	2.97	0.89	1.36	0.73	2.93	0.56	0.57
11	1.00 : 0.00 → ben...	0.000	benign	4.75	0.75	0.94	0.28	1.28	0.29	0.58	0.90	0.62
12	1.00 : 0.00 → ben...	0.000	benign	4.20	3.71	4.45	0.11	7.34	0.08	2.17	5.64	0.53
13	1.00 : 0.00 → ben...	0.000	benign	5.00	0.02	0.11	0.16	1.80	0.39	2.41	0.08	0.96
14	0.00 : 1.00 → mal...	0.000	malign	7.92	3.50	3.92	0.64	1.73	8.96	2.24	2.52	0.58
15	1.00 : 0.00 → ben...	0.000	benign	2.29	0.72	0.17	0.30	1.85	0.18	0.89	0.99	0.62
16	0.00 : 1.00 → mal...	0.000	malign	7.71	3.20	3.52	0.47	5.56	9.68	1.91	4.19	1.66
17	0.00 : 1.00 → mal...	0.000	malign	4.99	9.10	9.06	9.49	4.87	1.02	7.93	4.49	0.10
18	1.00 : 0.00 → ben...	0.000	benign	0.60	0.18	0.55	0.41	1.27	0.23	2.02	0.80	0.53
19	1.00 : 0.00 → ben...	0.000	benign	3.63	1.96	1.76	0.24	1.61	0.63	1.93	0.18	0.23
20	0.00 : 1.00 → mal...	0.000	malign	9.34	9.41	9.13	5.30	7.29	3.96	7.92	4.12	0.40
21	0.00 : 1.00 → mal...	0.000	malign	3.10	9.73	7.91	4.64	3.09	0.11	9.62	0.94	0.26
22	0.00 : 1.00 → mal...	0.000	malign	6.94	4.34	5.67	2.18	2.58	7.54	6.25	3.05	0.56
23	1.00 : 0.00 → ben...	0.000	benign	3.13	0.09	0.74	2.49	1.51	0.61	2.43	0.80	0.36
24	1.00 : 0.00 → ben...	0.000	benign	0.26	0.68	0.20	0.32	1.37	0.57	0.89	0.53	0.93
25	1.00 : 0.00 → ben...	0.000	benign	4.66	0.46	0.88	0.41	1.17	0.83	2.20	0.56	0.26
26	0.00 : 1.00 → mal...	0.000	malign	9.51	2.96	4.39	0.98	9.80	4.81	2.33	9.12	1.21
27	0.00 : 1.00 → mal...	0.000	malign	8.12	7.86	7.79	8.08	5.21	2.20	3.68	0.88	0.64
28	0.00 : 1.00 → mal...	0.000	malign	9.52	9.89	9.93	6.59	9.06	9.07	7.63	1.64	0.25
29	1.00 : 0.00 → ben...	0.000	benign	0.50	0.41	0.31	0.64	0.49	0.82	1.22	0.21	0.37

Show perfomance scores | Target class: (Average over classes)

Model	AUC	CA	F1	Prec	Recall	MCC
Tree	0.937	0.941	0.941	0.941	0.941	0.875

2) Esse fenômeno é conhecido como *overfitting* que é o ajuste demasiado dos dados de treinamento de forma que dados que possuem algum tipo de “ruído” acabam perdendo precisão em estruturas de árvores mais complexas.

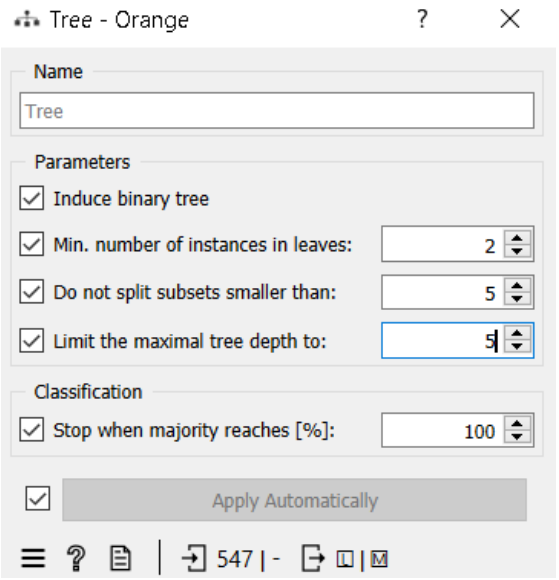


		Predicted: benign	Predicted: malign	∑
Actual	benign	81	4	85
	malign	4	47	51
	∑	85	51	136

Aqui podemos perceber que:

- Dos 85 benignos, acertamos 81 (2 acertos a menos que na árvore original)
- Dos 51 malignos, acertamos 47 (mantém)

Além disso, testando variações de parâmetros, pudemos perceber que a máxima precisão obtida pode ser atingida com a seguinte configuração:



Desse modo, encontramos uma precisão de 97,1%, o que é bastante curioso pois seria esperado que colocando um limite maior para a profundidade da árvore obteríamos resultados melhores. Porém isso não acontece devido ao *overfitting*.

		Predicted		
		benign	**malign**	**∑**
Actual	**benign**	83	2	**85**
	malign	2	49	**51**
	∑	**85**	**51**	**136**

Percebe-se que:

- Dos 85 benignos, acertamos 83 (mantém)
- Dos 51 malignos, acertamos 49 (2 acertos a mais que na árvore original)
- Totalizando uma precisão de 97,1% (3% a mais do que na árvore com *overfitting* que buscava a melhor precisão)