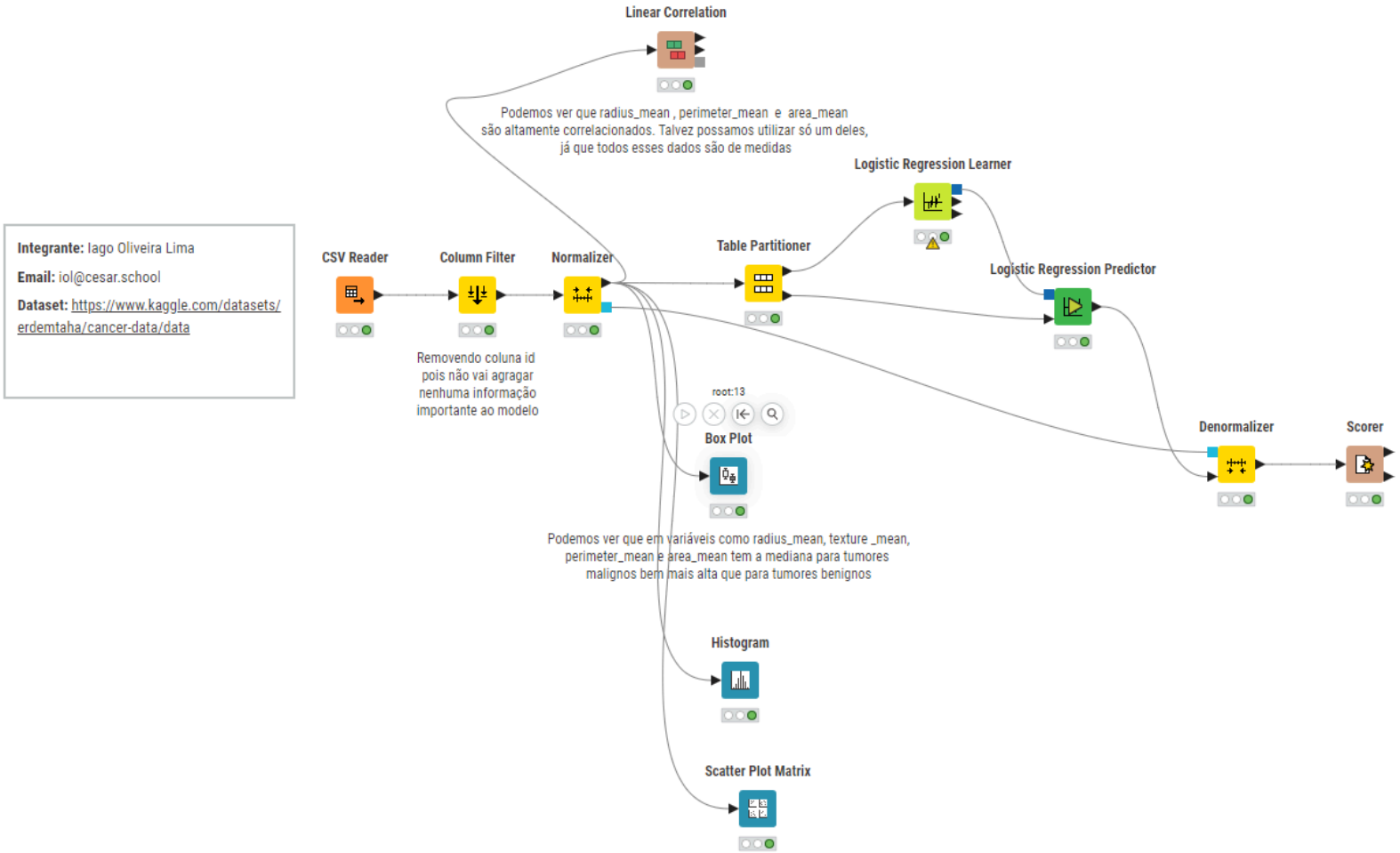
Trabalho Final - Data Science e Inteligência Artificial

Iago Oliveira Lima
iol@cesar.school
Dataset: https://www.kaggle.com/datasets/erdemtaha/cancer-data/data

1. Workflow



2. Análise de performance e conclusão

Rows: 2 | Columns: 2

#	RowID	M *Number (Integer)*	B *Number (Integer)*
1	M	65	4
2	B	0	102

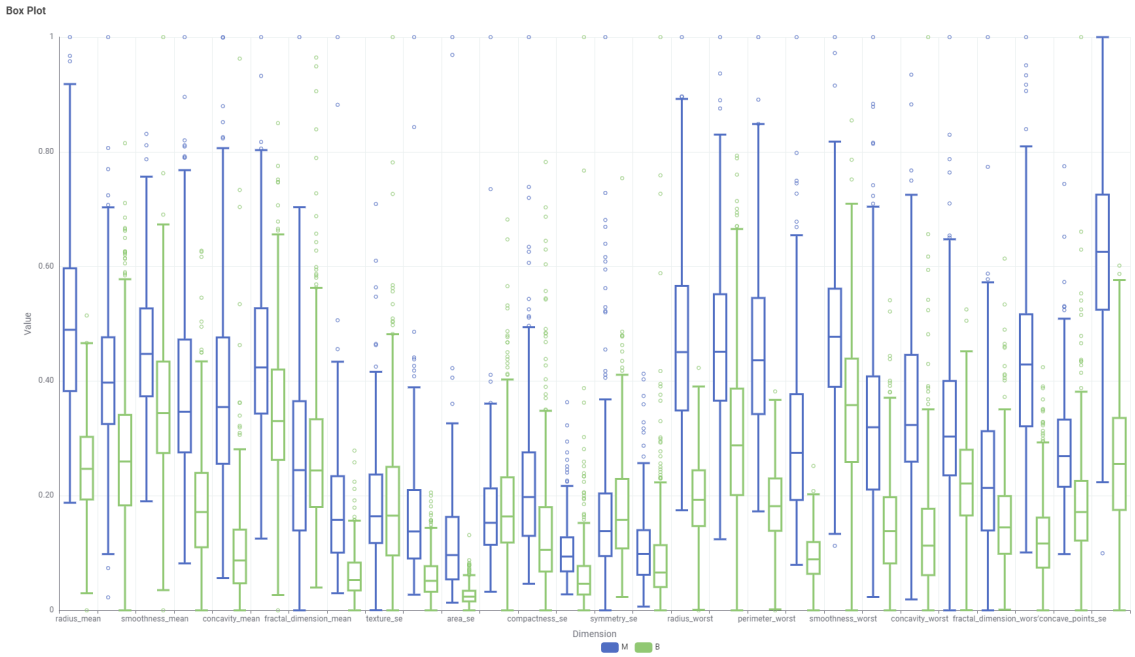
Pela matriz de confusão do nó “score” podemos observar que das 171 amostras que passaram pelo preditor, 167 foram classificadas corretamente e somente 4 foram classificados erroneamente como benignos.

Correlation measure (Table)

Rows: 435 | Columns: 5

#	RowID	First column name *String*	Second column name *String*	Correlation value *Number (Float)*	p value *Number (Float)*	Degrees of freedom *Number (Integer)*
2	Row1	radius_mean	perimeter_mean	0.998	0	567
371	Row37	radius_worst	perimeter_worst	0.994	0	567
3	Row2	radius_mean	area_mean	0.987	0	567
58	Row57	perimeter_mean	area_mean	0.987	0	567
372	Row37	radius_worst	area_worst	0.984	0	567
391	Row39	perimeter_worst	area_worst	0.978	0	567
227	Row22	radius_se	perimeter_se	0.973	0	567
75	Row74	perimeter_mean	perimeter_worst	0.97	0	567
18	Row17	radius_mean	radius_worst	0.97	0	567
73	Row72	perimeter_mean	radius_worst	0.969	0	567
20	Row19	radius_mean	perimeter_worst	0.965	0	567
99	Row98	area_mean	radius_worst	0.963	0	567
102	Row10	area_mean	area_worst	0.959	0	567
101	Row10	area_mean	perimeter_worst	0.959	0	567
228	Row22	radius_se	area_se	0.952	0	567
76	Row75	perimeter_mean	area_worst	0.942	0	567
21	Row20	radius_mean	area_worst	0.941	0	567
265	Row26	perimeter_se	area_se	0.938	0	567
180	Row17	concavity_mean	concave_points_mean	0.921	0	567
47	Row46	texture_mean	texture_worst	0.912	0	567
434	Row43	concave_points_mean	concave_points_worst	0.91	0	567

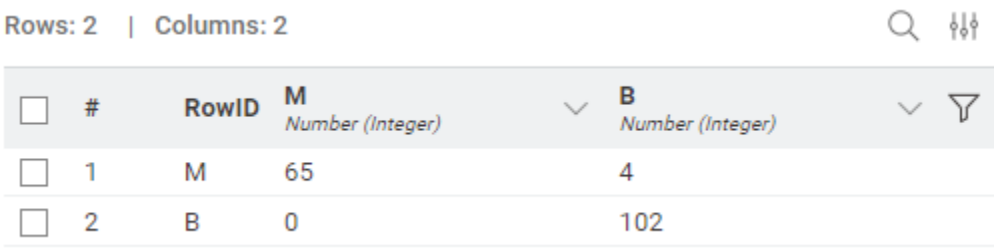
Utilizando o nó de correlação linear e ordenando os resultados da maior correlação para a menor, podemos perceber que alguns atributos têm uma correlação quase 1 como radius_mean, perimeter_mean, areas_mean. Nos dando a entender que talvez possamos utilizar somente um desses atributos já que todos eles são médias de medidas de área, perímetro e raio.



Podemos ver que os atributos como radius_mean, texture_mean, perimeter_mean e area_mean tem a mediana para tumores malignos bem mais alta que para tumores benignos, podendo se tornar bons atributos discriminativos para o modelo.

Na matriz de scatter plot acima, podemos identificar que quase todos os tumores malignos ficam em uma região do gráfico distinta dos benignos, o que nos confirma que são bons bons atributos discriminativos para o modelo.

Sendo assim, resolvi filtrar as colunas perimeter_mean e area_mean, dada a correlação alta radius_mean e obtive exatamente o mesmo resultado no score, conforme imagem abaixo. O que nos mostra que podemos simplificar nosso dataset removendo esses dois atributos pois caracterizam coisas similares ao radius_mean.



Rows: 2 | Columns: 2

#	RowID	M *Number (Integer)*	B *Number (Integer)*
1	M	65	4
2	B	0	102

3. Dificuldades

Particularmente, foi fácil identificar que a técnica de regressão logística se aplica bem a esse tipo de problema, pois é um problema de classificação binária. A minha maior dificuldade foi conseguir identificar quais atributos poderiam ser removidos do nosso dataset, a fim de simplificá-lo sem causar um impacto negativo na acurácia do modelo.

Único ajuste que precisei fazer é que o dataset originalmente veio com uma coluna a mais o que quebrava a leitura do KNIME, mas consertei o CSV na mão mesmo dado que era um problema fácil. O dataset que upei no GitHub já é o dataset corrigido.