

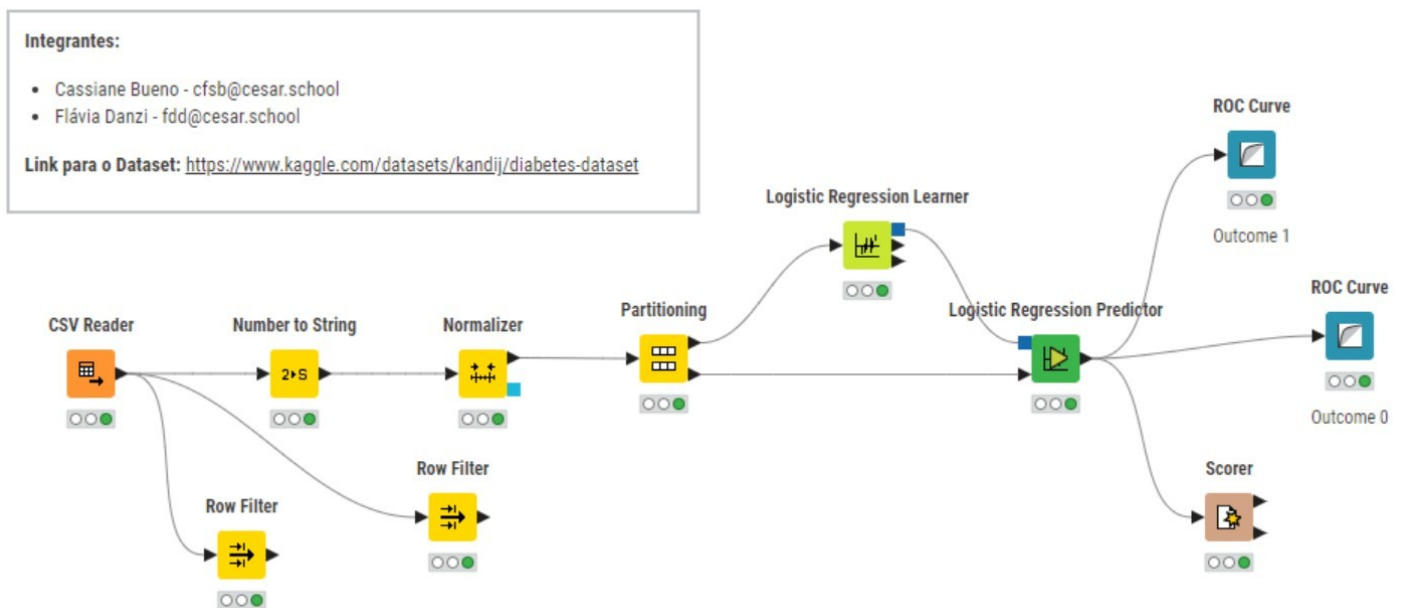
Integrantes (nome - email)

- Cassiane Bueno – cfsb@cesar.school
- Flávia Danzi – fdd@cesar.school

Link para o Dataset

- <https://www.kaggle.com/datasets/kandij/diabetes-dataset>

Workflow



Rows: 768 | Columns: 9

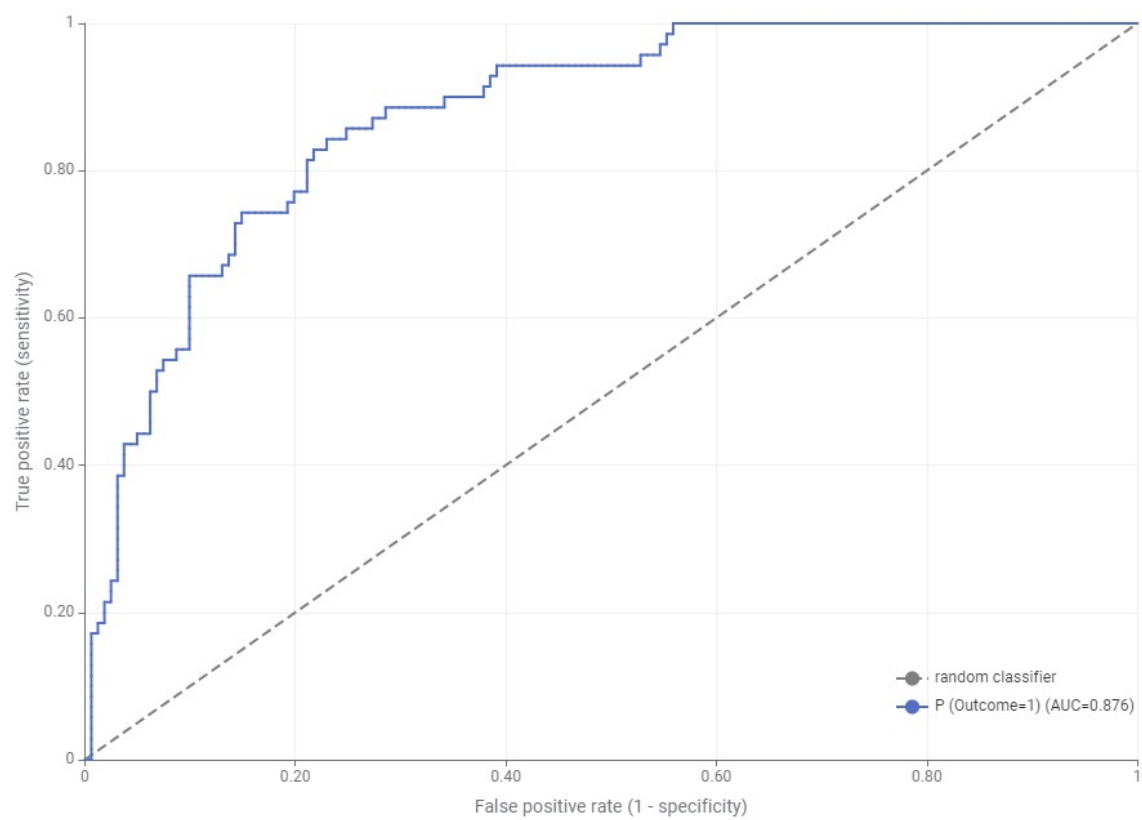
Table Statistics

<input type="checkbox"/>	#	RowID	Pregnancies Number (integer)	Glucose Number (integer)	BloodPress... Number (integer)	SkinThickn... Number (integer)	Insulin Number (integer)	BMI Number (double)	DiabetesPe... Number (double)	Age Number (integer)	Outcome Number (integer)
<input type="checkbox"/>	1	Row0	6	148	72	35	0	33.6	0.627	50	1
<input type="checkbox"/>	2	Row1	1	85	66	29	0	26.6	0.351	31	0
<input type="checkbox"/>	3	Row2	8	183	64	0	0	23.3	0.672	32	1
<input type="checkbox"/>	4	Row3	1	89	66	23	94	28.1	0.167	21	0
<input type="checkbox"/>	5	Row4	0	137	40	35	168	43.1	2.288	33	1
<input type="checkbox"/>	6	Row5	5	116	74	0	0	25.6	0.201	30	0
<input type="checkbox"/>	7	Row6	3	78	50	32	88	31	0.248	26	1
<input type="checkbox"/>	8	Row7	10	115	0	0	0	35.3	0.134	29	0
<input type="checkbox"/>	9	Row8	2	197	70	45	543	30.5	0.158	53	1
<input type="checkbox"/>	10	Row9	8	125	96	0	0	0	0.232	54	1

O dataset possui 768 linhas e 9 colunas numéricas com nomes de fácil compreensão, onde a coluna Outcome indica se a pessoa tem diabetes ou não.

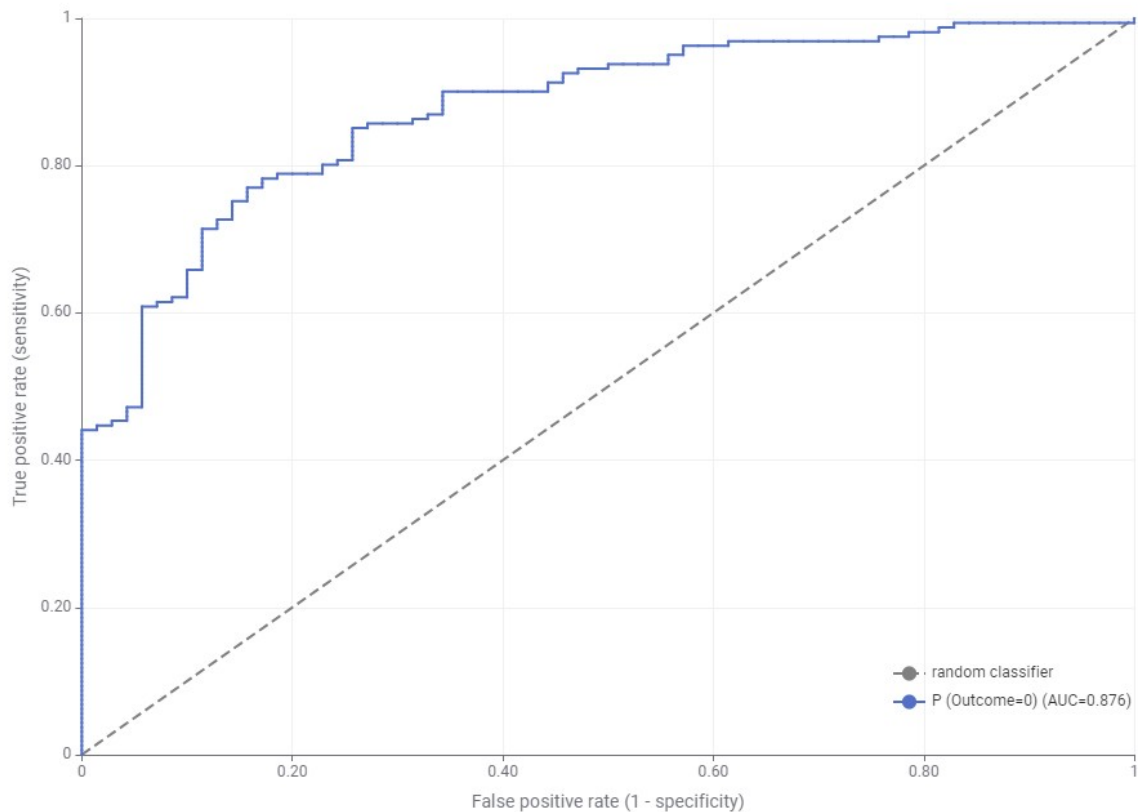
Análise de Performance

ROC Curve



ROC Curve Outcome = 1

ROC Curve



ROC Curve Outcome = 0

A área sob a curva ROC (AUC) de 0.876 representa um bom desempenho do modelo e sugere que o modelo é interessante para discriminar pessoas com ou sem diabetes.

► 1: Confusion matrix ► 2: Accuracy statistics Flow Variables

Rows: 3 Columns: 11

Table Statistics

<input type="checkbox"/>	#	RowID	TruePosit... Number (inte...)	FalsePosi... Number (inte...)	TrueNega... Number (inte...)	FalseNeg... Number (inte...)	Recall Number (dou...)	Precision Number (dou...)	Sensitivity Number (dou...)	Specificity Number (dou...)	F-measure Number (dou...)	Accuracy Number (dou...)
<input type="checkbox"/>	1	1	46	17	144	24	0.657	0.73	0.657	0.894	0.692	
<input type="checkbox"/>	2	0	144	24	46	17	0.894	0.857	0.894	0.657	0.875	
<input type="checkbox"/>	3	Overall										0.823

► 1: Confusion matrix ► 2: Accuracy statistics Flow Variables

Rows: 2 Columns: 2


Table Statistics

<input type="checkbox"/>	#	RowID	1 Number (integer)	0 Number (integer)
<input type="checkbox"/>	1	1	46	24
<input type="checkbox"/>	2	0	17	144

O dataset não é muito equilibrado, das 768 linhas, 500 são com Outcome 0 e apenas 268 são com Outcome 1, o que nos leva a entender o porquê da matriz de confusão acima ilustrar o resultado da RowID 1 consideravelmente inferior ao resultado da RowID 0.

► 1: Confusion matrix

► 2: Accuracy statistics

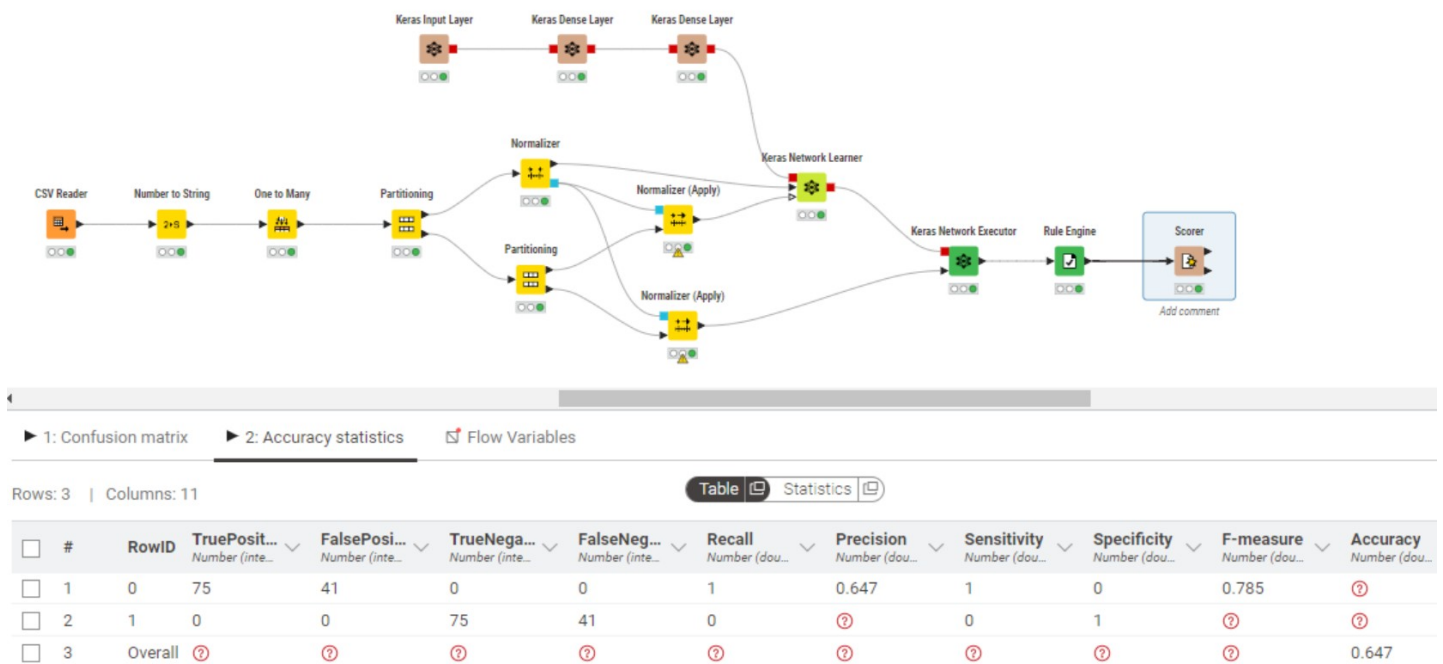
 Flow Variables

Count: 6

Owner ID	Data Type	Variable Name	Value
3:12	DoubleType	Cohen's kappa	0.5675934803451583
3:12	IntType	#False	41
3:12	IntType	#Correct	190
3:12	DoubleType	Error	0.1774891774891775
3:12	DoubleType	Accuracy	0.8225108225108225
	StringType	knime.workspace	C:\Users\cassi\knime-workspace

A acurácia foi de 82.3% com 10.000 épocas. Como o dataset é pequeno, ele não convergiu mesmo após 10.000 épocas. Mesmo com 20.000 ou 30.000 épocas ele não convergiu.

Foi feita uma tentativa para o uso de Keras com 3 camadas, sendo duas camadas densas, uma com função de ativação ReLU e outra com Softmax, mas a acurácia não passou de 64.7%, conforme figura abaixo.



Os dados foram divididos com 70% para treinamento e 30% para teste. Fizemos ajustes também na normalização excluindo um campo que supomos não fazer sentido normalizar por ter valores muito pequenos (DiabetesPedigreeFunction), porém aparentemente não surtiu efeito. Achamos estranho também ter algumas linhas com valores zerados nas colunas BloodPressure e Glucose, pois sabemos que esses valores nunca são zero. Tivemos dificuldade em fazer o modelo convergir com este dataset, tivemos que ajustar o número de épocas e as partições para melhorar a acurácia e diminuir a perda.

Assim como vimos durante o curso, foi interessante podermos entender melhor com este exemplo prático de dataset que a regressão logística é útil para situações nas quais você deseja prever a presença ou a ausência de uma característica ou resultado com base em valores de um conjunto de variáveis preditoras e que a variável dependente é binária ou dicotômica (tem diabetes ou não tem diabetes).