

CESAR SCHOOL
Pós Graduação em Tech Lead

Thiago Neves Santos

MODELO DE PREDIÇÃO DE CÂNCER DE TIREÓIDE
Papilífero e Micropapilífero

Recife
2025

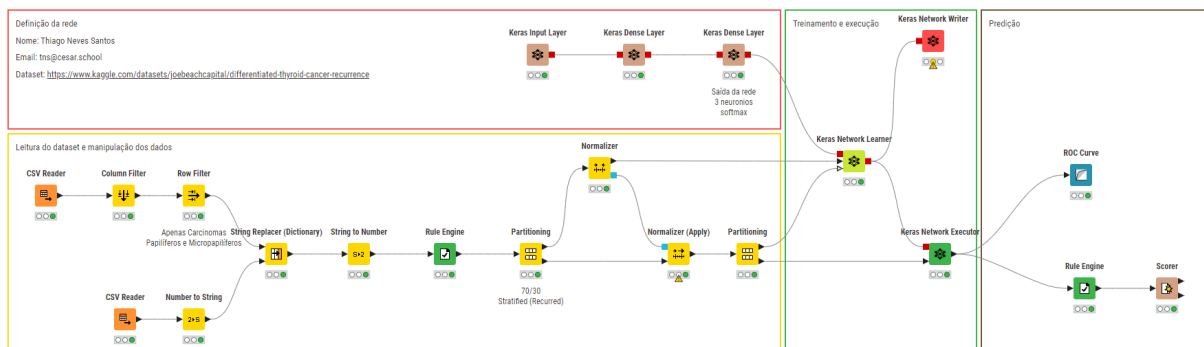
1. Identificação	3
2. Imagens	3
a. Visão geral	3
b. Definição da rede	3
c. Leitura do dataset e manipulação dos dados	4
d. Treinamento e execução	4
e. Predição	5
3. Resultados	5
a. Análise de performance	5
b. ROC Curve	6
4. Conclusão	6

1. Identificação

Nome: Thiago Neves Santos
Email: tns@cesar.school
Dataset: [Differentiated Thyroid Cancer Recurrence](https://www.kaggle.com/datasets/joebeachcapital/differentiated-thyroid-cancer-recurrence)

2. Imagens

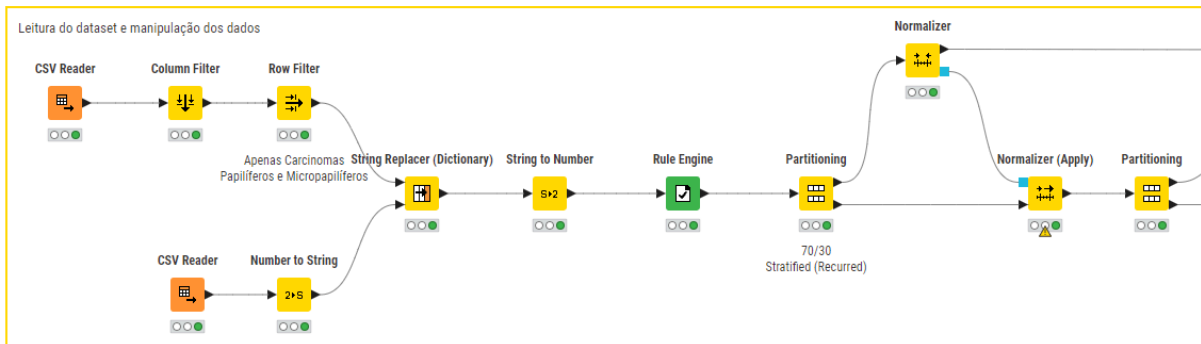
a. Visão geral



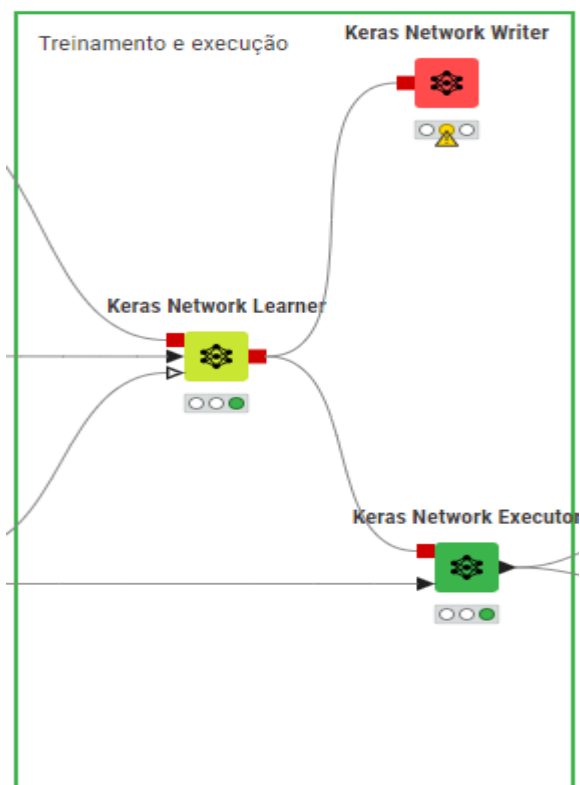
b. Definição da rede



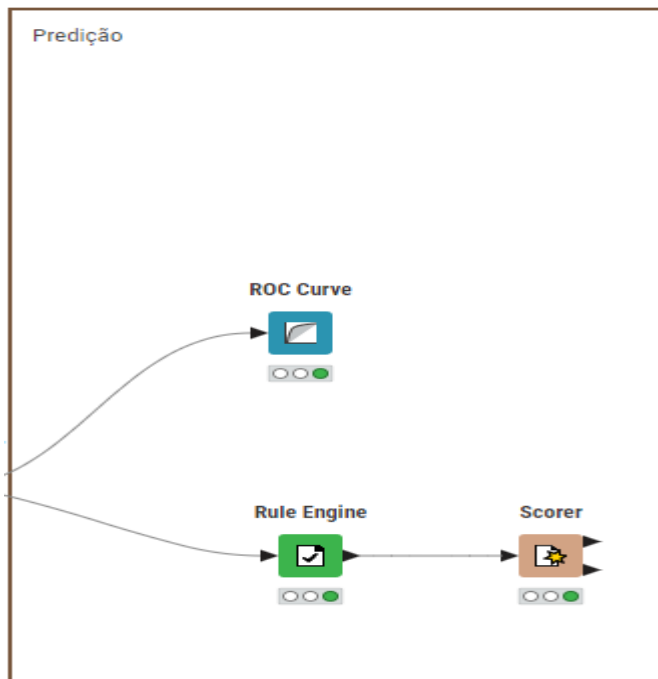
c. Leitura do dataset e manipulação dos dados



d. Treinamento e execução



e. Predição



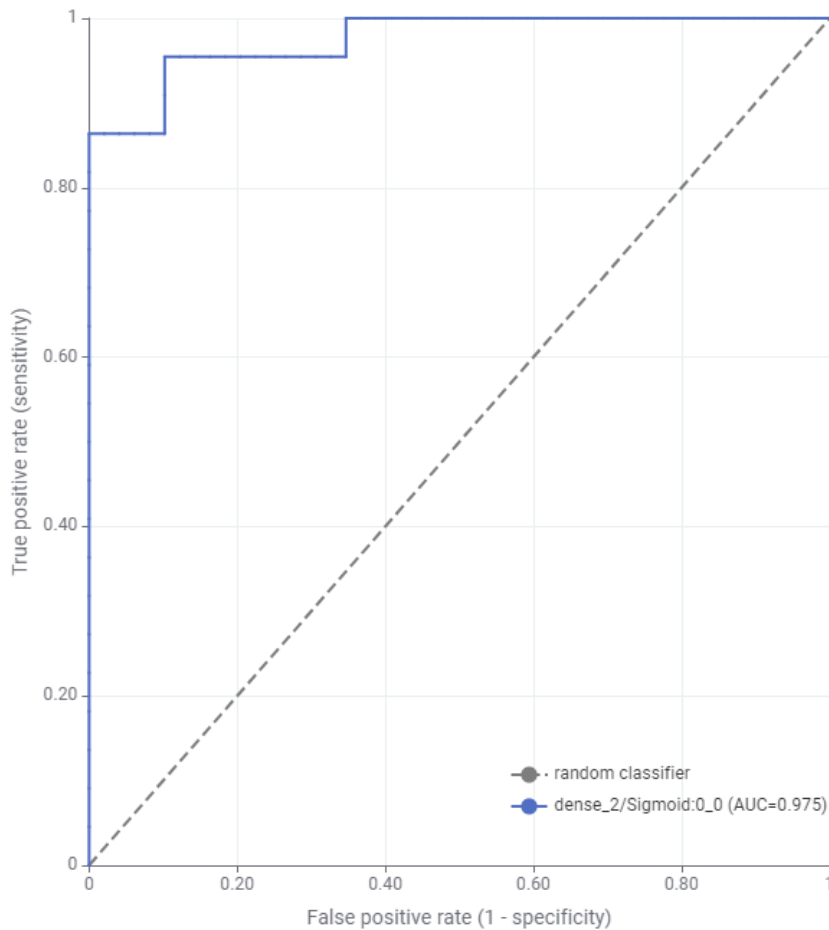
3. Resultados

a. Análise de performance

Confusion Matrix - 3:47 - Scorer				
File	Hilite			
Recurred \...	No	Yes		
No	48	1		
Yes	3	19		
Correct classified: 67			Wrong classified: 4	
Accuracy: 94,366%			Error: 5,634%	
Cohen's kappa (κ): 0,865%				

b. ROC Curve

ROC Curve



4. Conclusão

Eu fiz 3 workflows no Knime: um usando regressão logística sem fazer muito parse do input, outro também usando regressão logística, porém fazendo algumas substituições nos dados de input. Nesses dois workflow, os resultados foram bons, acima de 80% de precisão.

Porém, por se tratar de uma previsão do usuário vir a ter uma recidiva de uma doença maligna, eu procurei aperfeiçoar ao máximo de modo a obter a maior precisão.

Daí eu refiz todo o workflow utilizando a extensão keras deep learning. Criei uma rede neural densa com apenas uma camada oculta.

A princípio eu quebrei muito a cabeça para montar a rede, até que percebi que como meu problema se tratava de uma classificação binária, só precisava de apenas uma saída da

rede (o campo de recurred do dataset original), e a função de ativação deveria ser a sigmoid no nó de saída.

Ajustado o modelo, quebrei um pouco mais a cabeça pois no nó scorer, só estava aparecendo a coluna vinda do rule engine (prediction). Precisei fazer um ajuste no tratamento dos dados. Feito estes ajustes, obtive uma precisão de mais de 90%.

Eu queria aumentar ainda mais a precisão. Para isso, analisei o dataset original e percebi que havia retirado dos inputs, uma variável que tinha alta correlação com o resultado (response). Após os devidos ajustes, consegui a precisão de 94%.

