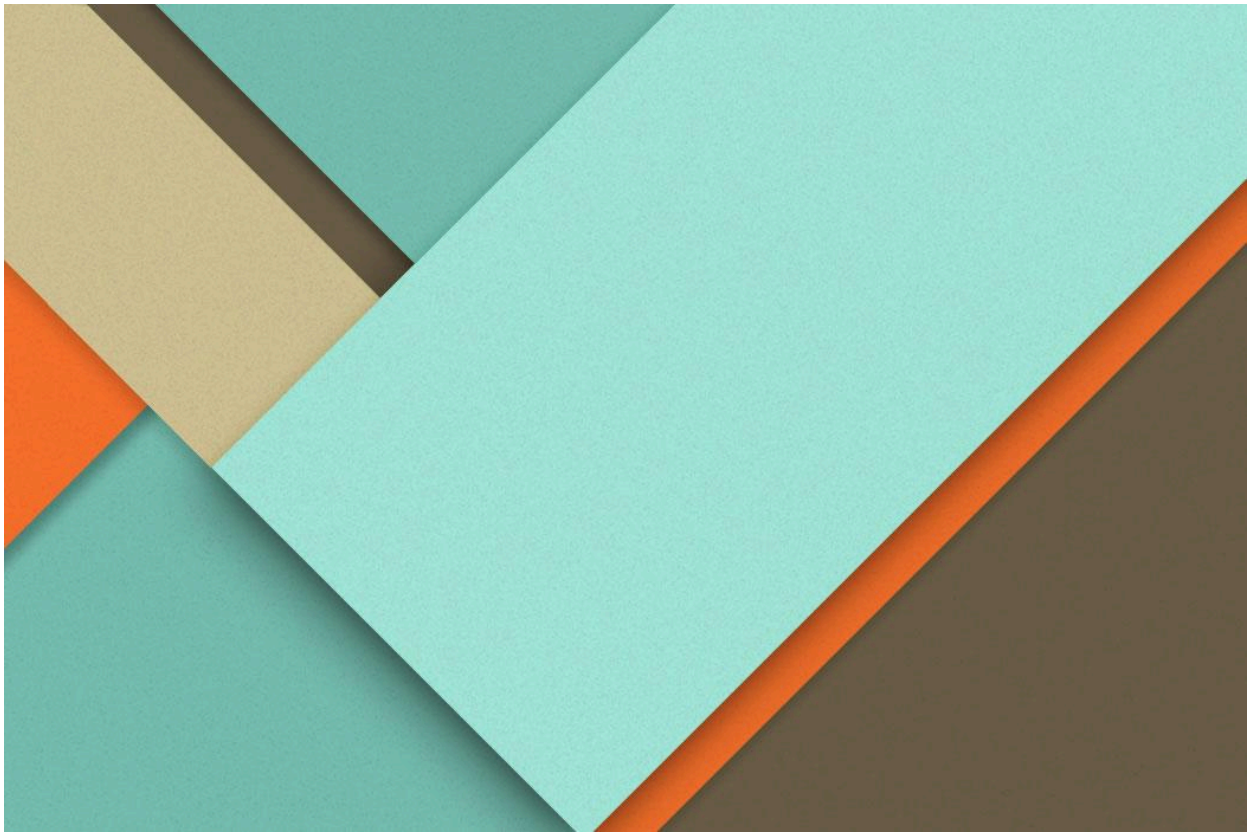

Projeto Final

Tech Lead - Data Science e IA

—

Rafael Winter
rwt@cesar.school
Dataset: https://www.kaggle.com/datasets/zeesolver/social-network

Turma 2024.02

Workflow

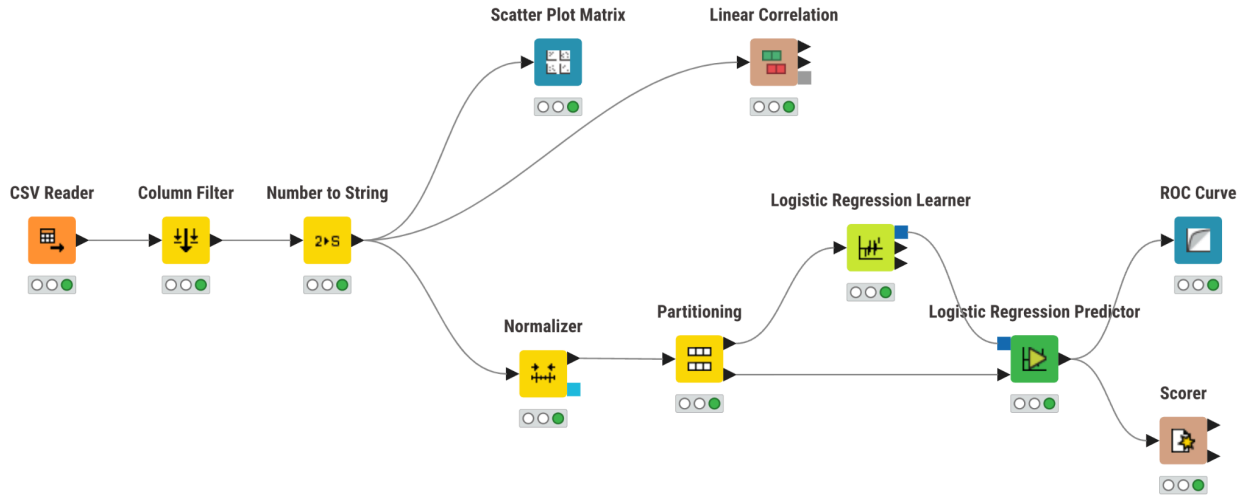

Figura 1 - Capura de tela do Knime mostrando o Workflow do projeto

Processo

Foi escolhido o dataset “Customer Purchasing Power”, que tenta determinar se a pessoa pode comprar um carro de acordo com sua idade e salário estimado, utilizando regressão logística.

Para isso o arquivo CSV foi carregado, e foi removida a coluna User ID, que não é útil na previsão de valor. A coluna Purchased foi convertida para String, pois é a categoria que será prevista pelo modelo. Ao analisar a correlação usando o nó Linear Correlation, o valor de correlação entre Purchased (compra) e Gender (gênero) é 0,042, considerado baixo, o que indica que o gênero não é determinante para a decisão de compra.

Foi feita também uma análise gráfica usando Scatter Plot Matrix, onde parece não haver uma relação forte entre idade e salário estimado. Além disso, os dados parecem estar bem distribuídos, sem evidências de outliers. Ao que parece há relevância nas variáveis idade e salário estimado para a compra, conforme o gráfico a seguir.

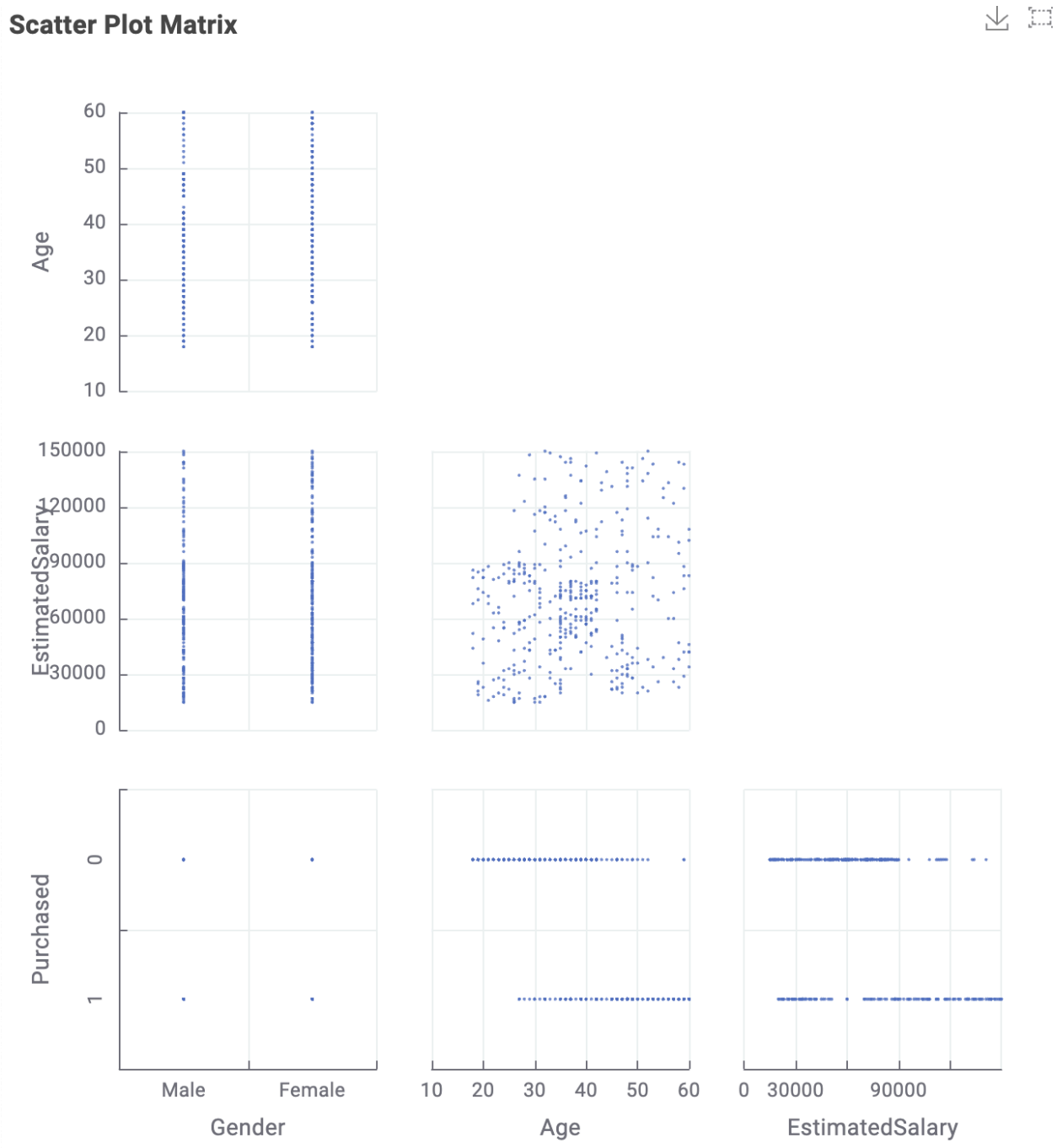

Figura 2 - Scatter Plot Matrix

Foi utilizado um nó Logistic Regression Learner, que recebeu uma partição dos dados já tratados. Ele foi configurado para executar 30.000 épocas, com learning rate de 0,1. As estatísticas mostram no nó Scorer que o modelo tem acurácia de 74,2%. Na saída do Logistic Regression Predictor também foi adicionado um nó ROC Curve, para que fossem

analisadas as contribuições das variáveis na predição, onde fica claro que idade e salário estimado contribuem relevantemente para compra, conforme o gráfico a seguir:

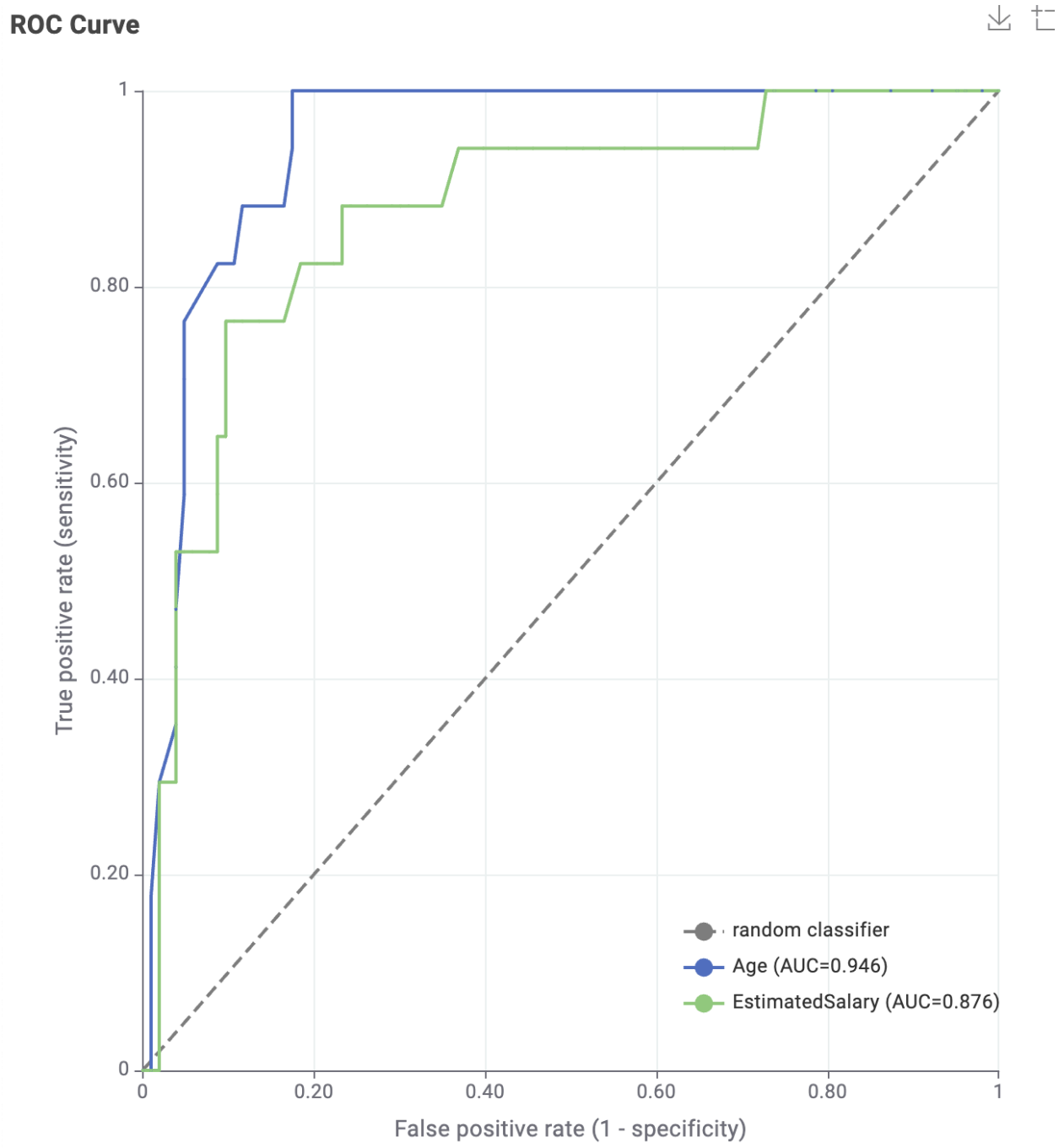

Figura 3 - Gráfico ROC Curve