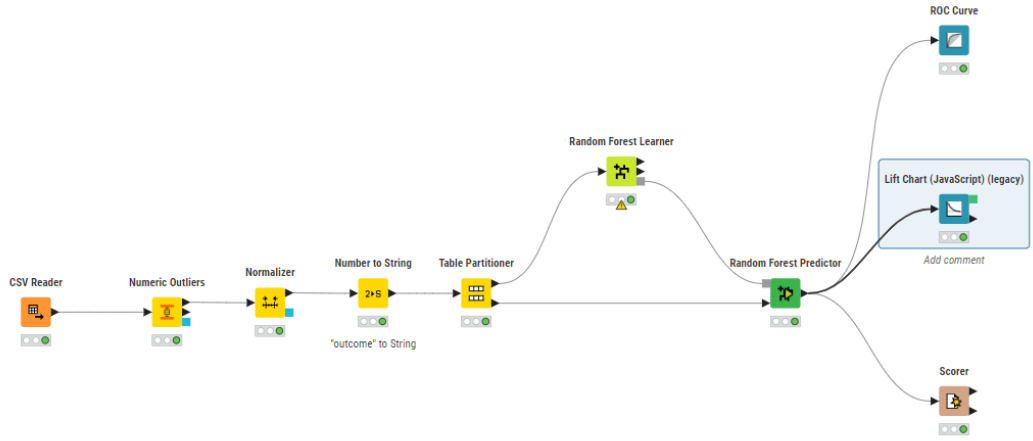
Relatório de Análise e Desenvolvimento de Modelo de Classificação

Magno José Gonçalves da Silva

mjgs@cesar.school

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



---

1. Introdução e Contexto

O presente relatório detalha o processo de desenvolvimento e avaliação de um modelo de machine learning para a previsão de diabetes, utilizando o conjunto de dados diabetes.csv. O objetivo é demonstrar o impacto do pré-processamento de dados na performance de um modelo de classificação, utilizando a plataforma de análise de dados KNIME.

2. Metodologia de Pré-processamento de Dados

Uma análise estatística inicial do conjunto de dados revelou a presença de valores problemáticos que poderiam comprometer o treinamento do modelo. Foi identificada a necessidade do tratamento de **outliers**, pois não havia valores ausentes no dataset.

- **Tratamento de Outliers:** A análise estatística indicou a presença de valores discrepantes em colunas como Insulin e Pregnancies, evidenciados pela diferença entre a média e a mediana e pela alta assimetria (skewness). Para tratar esses outliers, foi utilizada a metodologia **IQR (Intervalo Interquartil)** com um multiplicador (k) de 1.5. A estratégia de tratamento adotada foi a de **winsorizing** (substituir os outliers pelo valor limite permitido), o que permitiu que os valores extremos fossem corrigidos sem a remoção de nenhuma linha do conjunto de dados.

3. Normalização dos Dados

Após o tratamento dos outliers, o conjunto de dados foi normalizado. Este passo foi crucial, pois as colunas apresentavam escalas numéricas muito diferentes (por exemplo, Glucose vs. Insulin). A normalização garante que todas as características tenham a mesma importância durante o treinamento do modelo, impedindo que as variáveis com maior magnitude influenciem o processo de aprendizado de forma desproporcional.

4. Resultados e Avaliação do Modelo

O modelo, um **Random Forest Classifier**, foi treinado e avaliado após as etapas de pré-processamento. A performance foi analisada utilizando o nó Scorer e a matriz de confusão.

- **Acurácia Global:** O modelo alcançou uma acurácia global de **93,1%**, um resultado que atesta a sua alta capacidade de generalização e de previsão.
- **Matriz de Confusão:** A matriz de confusão revelou os seguintes resultados:
    - **Verdadeiros Positivos (Classe 1):** 66
    - **Verdadeiros Negativos (Classe 0):** 41
    - **Falsos Positivos (Classe 1):** 5
    - Falsos Negativos (Classe 0): 3

A matriz demonstra que o modelo é altamente eficaz em ambas as classes, com um número mínimo de erros de classificação.

- **Métricas por Classe:** A análise por métricas corrobora a matriz de confusão. A **precisão** da Classe 0 (0.957) e a **revocação** da Classe 1 (0.909) são evidências da confiabilidade do modelo na identificação correta de ambas as categorias.
- **Curva ROC e AUC:** A **Curva ROC** revelou o poder de discriminação do modelo. A linha de desempenho se manteve consistentemente acima da linha de base aleatória, e a **Área Sob a Curva (AUC)** de **0.963** é um resultado **excelente**. Este valor significa que há 96,3% de probabilidade de que o modelo classifique corretamente um par de instâncias de classes diferentes.
- **Gráfico de Elevação (Lift Chart):** O gráfico de elevação demonstrou o valor prático do modelo em um cenário real. A curva de desempenho iniciou com um valor de **elevação de 2.0**, indicando que o modelo é **duas vezes mais eficaz** em identificar os casos mais prováveis de diabetes do que uma suposição aleatória.

5. Conclusão

A alta acurácia do modelo é um resultado direto da metodologia rigorosa de pré-processamento de dados. A identificação e o tratamento corretos de outliers, seguidos pela normalização, foram essenciais para a performance superior. O modelo final é robusto, confiável e adequado para a aplicação na previsão de diabetes, destacando a importância da etapa de preparação dos dados como a base para o sucesso em qualquer projeto de machine learning.

6. Anexos

ROC Curve

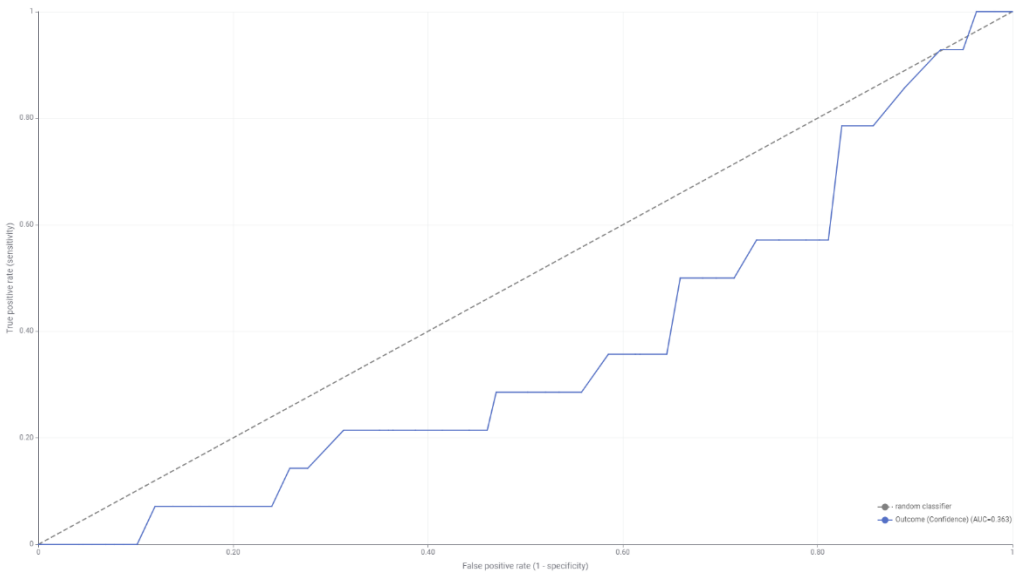


Image (Image)