Treinamento realizado no KNIME usando o dataset Titanic

Aluno: Ariano Batista Neves

Email: ariano.neves@gmail.com

No treinamento, foram aplicados dois algoritmos de aprendizado supervisionado: **Random Forest** e **Regressão Logística**. Ambos foram usados para prever a sobrevivência dos passageiros no Titanic com base em suas características.

---

1. Etapas Realizadas no Treinamento

1. **Carregamento do Dataset**:
    - O dataset Titanic foi importado no KNIME usando o node CVS Reader;
2. **Seleção de Colunas Relevantes**:
    - As colunas utilizadas no modelo foram:
        - `Survived` (variável dependente);
        - `Pclass`, `Age`, `Sex`, `SibSp`,`Fare`,`PassengerId`, `Embarked` e `Parch` (variáveis independentes).
    - Colunas irrelevantes como `Name`,`Cabin` e `Ticket` foram descartadas.
3. **Valores Ausentes**
    - Realizado pré-processamento para lidar com valores ausentes (ex.: Age)
4. **Modelagem com Random Forest**:
    - Utilizado o nó **Random Forest Learner** para criar o modelo:.

    4.1 **Transformação de Número para String**

    - Realizado a transformação da coluna `Fare` de inteiro para String;

    4.2 **Definição dos limites (ou "domínios") de valores das colunas**

    - Foi utilizado o Node Domain Calculator para limitar a coluna Age

    4.3 **Divisão do Dataset**:

    - O dataset foi dividido em treinamento (81%) e teste (19%) usando um nó de Partitioning.

    4.4 O nó **Random Forest Predictor** foi usado para fazer as previsões.

5. **Modelagem com Regressão Logística**:

    5.1 **Normalização**

    - Foi utilizado o node Normalizer para realizar a normalização das colunas

    5.1 **Transformação de Número para String**

- Realizado a transformação da coluna `Fare` de inteiro para String;

5.2 **Definição dos limites (ou "domínios") de valores das colunas**

- Foi utilizado o Node Domain Calculator para limitar a coluna Age

5.3 **Divisão do Dataset**:

- O dataset foi dividido em treinamento (81%) e teste (19%) usando um nó de Partitioning.

5.4 O nó **Logistic Regression Learner** foi configurado para treinar outro modelo.

- Previsões foram feitas com o **Logistic Regression Predictor**.

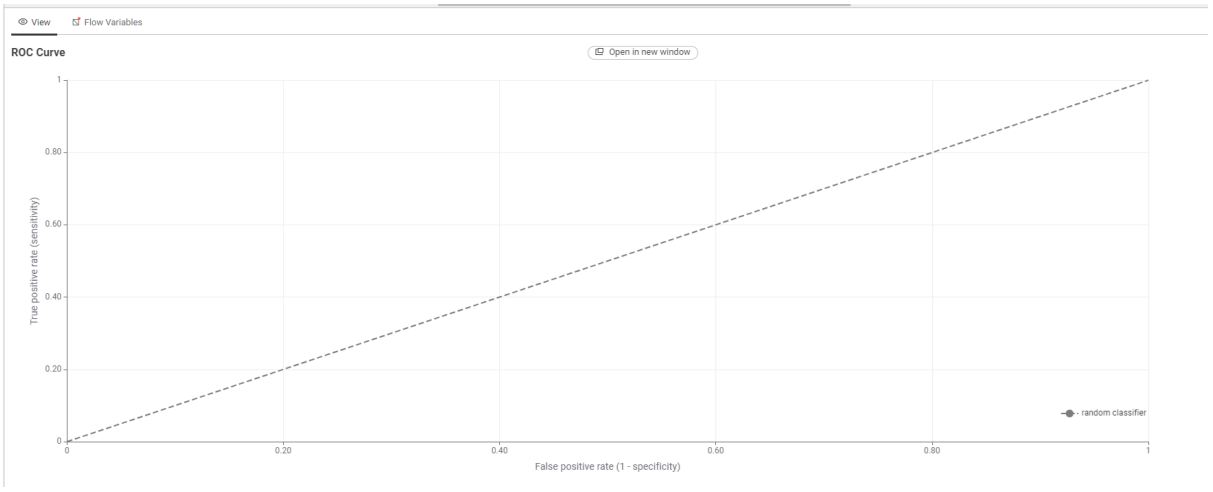
6. **Avaliação**:
   - Ambos os modelos foram avaliados usando métricas como:
     - Acurácia;
     - Matriz de confusão;
     - Curva ROC AUC.

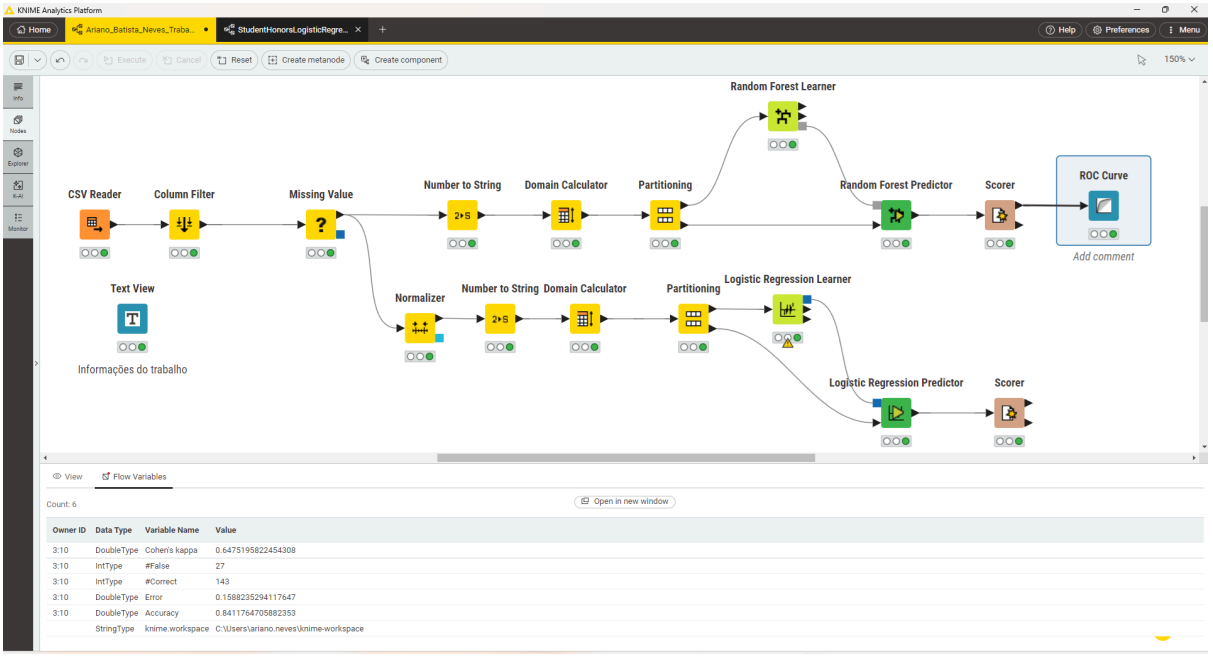
---

7. Considerações Finais

Neste trabalho, utilizei os algoritmos Random Forest e Regressão Logística para analisar os dados. O Random Forest se mostrou bem eficiente, especialmente porque combina várias árvores de decisão, o que ajuda a entender relações mais complexas nos dados. A Regressão Logística, por sua vez, se destacou pela simplicidade e rapidez. Foi mais fácil de entender como ela funciona e de interpretar os resultados.

No fim, os dois algoritmos chegaram a resultados bem parecidos. Mostrando que ambos são boas opções para o dataset escolhido.





View | Flow Variables

Count: 6 — Open in new window

Owner ID	Data Type	Variable Name	Value
3:10	DoubleType	Cohen's kappa	0.6475195822454308
3:10	IntType	#False	27
3:10	IntType	#Correct	143
3:10	DoubleType	Error	0.1588235294117647
3:10	DoubleType	Accuracy	0.8411764705882353
	StringType	knime.workspace	C:\Users\ariano.neves\knime-workspace