

## ***Construção de um Modelo preditivo Naive Bayes utilizando a plataforma analítica KNIME***

Neste trabalho sobre Machine Learning (aprendizado de máquina supervisionado) farei uma pequena demonstração usando a plataforma analítica KNIME. Estarei criando um fluxo de trabalho KNIME que utiliza o método de Naive Bayes para treinar um modelo no conjunto de dados formação de adultos. Neste fluxo utilizarei as operações de preparação, limpeza e manipulação de dados no conjunto de dados de treinamento adult.csv, este conjunto de dados pode ser encontrado no repositório de aprendizagem de máquina UCI, no link abaixo : <http://archive.ics.uci.edu/ml/datasets/Adult> -(Este conjunto de dados foi desenvolvido por Barry Becker e foi extraído do banco de dados do Censo americano de 1994.)

O objetivo dessa predição, utilizando o método de Naive Bayes, se resume a determinar quanto o modelo é capaz de aprender a prever o rendimento anual de uma pessoa ( $\leq 50K$  ou  $> 50K$ ), ou seja, se uma pessoa ganha acima de 50K ( $> 50$ ) ou ganha igual ou menos que 50 K ( $\leq 50$ ) por ano.

Sem entrar em detalhes do Teorema de Bayes , Naive Bayes é um classificador probabilístico baseado no Teorema de Bayes, que mostra como determinar a probabilidade de um evento condicional através da probabilidade inversa. Para facilitar a computação, este classificador assume que a presença (ou ausência) de um atributo não tem relação alguma com qualquer outro atributo.

Para dar início ao fluxo de trabalho KNIME utilizarei o nó leitor de arquivos para ler o conjunto de dados de treinamento a partir de

<http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data>.

Veja abaixo uma pequena descrição dos nós que utilizarei nesta demonstração :

**File Reader** – leitor de arquivos em vários formatos (texto, csv , etc ) pode ler o arquivo localmente ou ler da Web fornecendo a URL.

**Partitioning** - Particiona o arquivo em dois conjuntos de dados (70% para treinamento e 30% para teste. Estes percentuais podem ser modificados , mas o ideal para treinamento de modelo e utilizar 66% ou 70% do conjunto de dados. Para que os resultados possam ser reproduzidos exatamente como estes , caso alguém deseje reproduzi-los , este nó foi configurado com uma semente aleatória (Seed) =1489768553791.

**Naive Bayes Learner** - O nó cria um modelo bayesiano a partir dos dados de treinamento, calcula o número de linhas por valor de atributo por classe para atributos nominais e a distribuição Gaussiana para atributos numéricos. Este nó pode ser configurado para desprezar valores em falta no conjuntos de dados. Nesta demonstração o nó foi configurado para desprezar linhas que possuam valores em falta em qualquer coluna, pois o referido conjunto de dados possui valores em falta em algumas linhas.

**Naive Bayes Predictor** - Prevê a classe por linha com base no modelo aprendido. A probabilidade de classe é o produto da probabilidade por atributo e a probabilidade do próprio atributo de classe. Ao observar o fluxo, pode-se verificar que neste nó entra os 30 % do conjunto de dados que será usado para teste pelo modelo aprendido, modelo este criado pelo nó **Naive Bayes Learner**.

**Scorer** - Este nó deve ser adicionado no final do fluxo de trabalho, a fim de medir o desempenho dos classificadores.

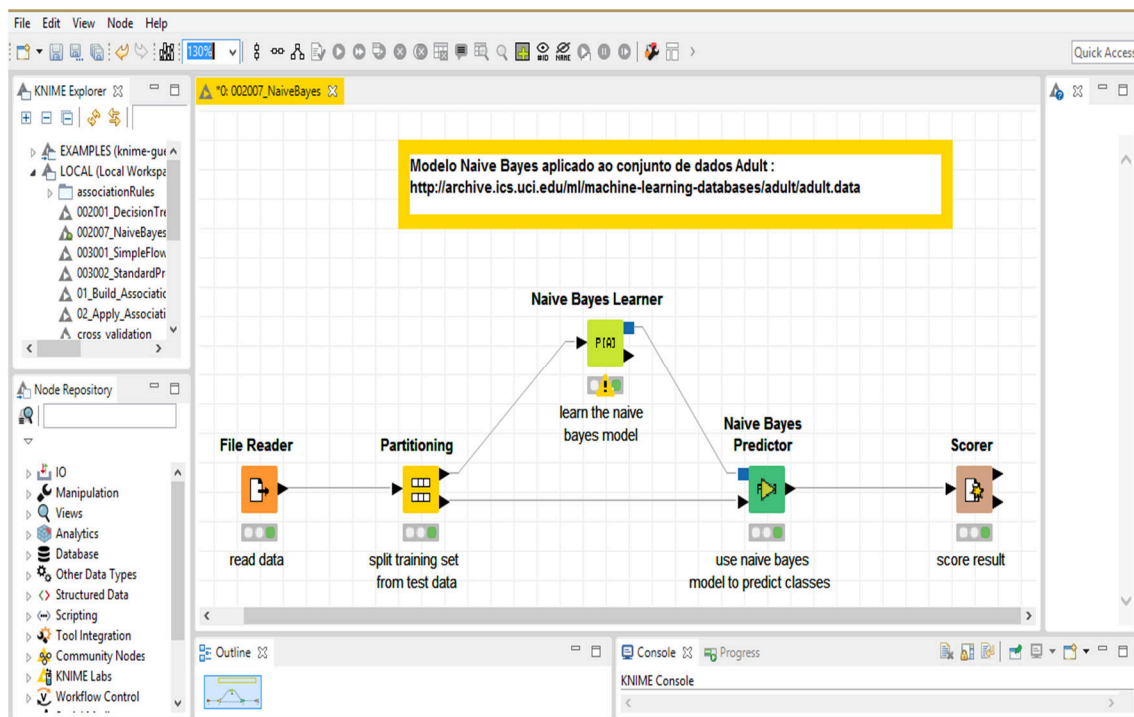
Ele compara duas colunas por seus pares de valor de atributo e mostra a matriz de confusão(Confusion Matrix), ou seja, quantas linhas de qual atributo e sua classificação correspondem. A saída do nó é a matriz de confusão com o número de correspondências em cada célula. Além disso, ele reporta uma série de estatísticas de precisão, como Verdadeiro-Positivos, Falso-Positivos, Verdadeiro-Negativos, Falso-Negativos, Precisão, Sensibilidade, Especificidade, F-measure, bem como a precisão geral e o kappa de Cohen.

## ***Iniciando o fluxo de trabalho no KNIME.***

Após fazer as devidas configurações em cada nó (que aliás , é bastante simples) pode-se executar cada nó, um após outro, na sequência do fluxo; ou pode-se executar automaticamente todo o fluxo pressionando-se as teclas (SHIFT + F7).

Após a execução do fluxo pode-se verificar no nó “**Scorer**”, o resultado do modelo criado **Naive Bayes Predictor** , aplicado ao conjunto de dados.

## ***FLUXO KNIME***



## ***Conclusão***

O modelo foi treinado com um subconjunto de dados com 22.792 linhas, ou seja 70 % do conjunto total 32.561 linhas.

O desempenho do modelo **Naive Bayes Predictor** em relação à predição para as pessoas que tenham o rendimento anual de ( $\leq 50K$  ou  $> 50K$ ) é o seguinte :

## CONFUSION MATRIX

File Hilitte		
Col14   Pre...	<=50K	>50K
<=50K	14506	2798
>50K	1279	4210

Correct classified: 18.716	Wrong classified: 4.077
Accuracy: 82,113 %	Error: 17,887 %
Cohen's kappa (κ) 0,553	

A aba **Confusion Matrix** do Nó **“Scorer”** demonstra que :

Das **22.793** linhas do conjunto de dados, **18.716** (foram corretamente classificadas , ou seja , **4.210** como **Verdadeiro-Positivos** que ganham acima de 50 K e **14.506** como **Verdadeiro-Positivos** que ganham igual ou menos que 50 K. Por outro , **4.077** foram erradamente classificadas, ou seja, 2.798 como **Falso-Positivos** que ganham acima de 50 K e 1.279 como **Falso-Positivos** que ganham igual ou menos que 50 K.

Assim, o desempenho do modelo é o seguinte:

18.716 linhas classificadas corretamente com índice de acerto de 82,113 % de acerto do modelo.

4.077 linhas classificadas erradamente com índice de erro de 17,887 % de erro do modelo.

## **ACCURACY STATISTICS**

File											
Table "default" - Rows: 3   Spec - Columns: 11   Properties   Flow Variables											
Row ID	TruePo...	FalsePo...	TrueNe...	FalseNe...	D Recall	D Precision	D Sensitivity	D Specifity	D F-meas...	D Accuracy	D Cohen...
<=50K	14506	1279	4210	2798	0.838	0.919	0.838	0.767	0.877	?	?
>50K	4210	2798	14506	1279	0.767	0.601	0.767	0.838	0.674	?	?
Overall	?	?	?	?	?	?	?	?	?	0.821	0.553

Já a aba **Accuracy statistics** do Nó "**Scorer**" demonstra que temos:

- I - A sensibilidade(**sensitivity**) de **76%** - 0,767 (**VP** / **VP** + **FN**).  
 - **Sensitivity** é a proporção de verdadeiros positivos, ou seja, a capacidade do modelo classificar um indivíduo como **VP** dado que realmente ele é **VP**.
- II- A especificidade (**specificity**) de **83%** - 0.838 (**VN** / **FP** + **VN**).  
 - **Specificity** é a proporção de verdadeiros negativos, isto é, avalia a capacidade do modelo predizer um indivíduo como **VN** dado que ele realmente é **VN**.
- III - A acurácia (accuracy) no modelo de **82%** - 0.821 (**VP** + **VN**) / (**VP**+**FP**+**FN**+**VN**).  
 - **Accuracy** - é a proporção de predições corretas, sem considerar o que é positivo e o que é negativo e sim o acerto total.