Modelo Naive Bayes com plataforma analítica KNIME

Neste texto sobre Machine Learning (aprendizado de máquina supervisionado) farei uma pequena demonstração usando a plataforma analítica KNIME. Neste trabalho estou criando um fluxo de trabalho KNIME que utiliza o método de Naive Bayes para treinar um modelo no conjunto de dados formação de adultos.

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.

O objetivo dessa predição utilizando o método de Naive Bayes é para determinar se uma pessoa ganha mais de 50K por ano. Usarei o conjunto de dados que pode ser encontrado no repositório de aprendizagem de máquina UCI 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.)

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 baixado da Internet 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 reproduzí-los , foi configurado neste nó 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 possue valores em falta em algumas linhas.

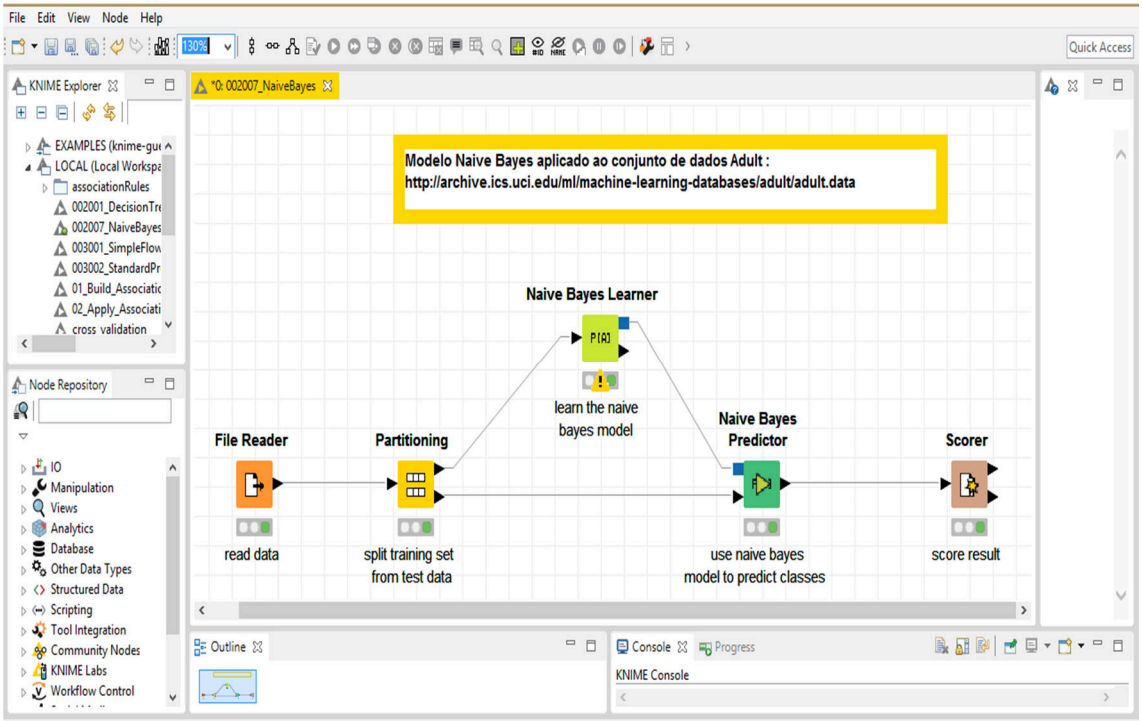
Naive Bayes Predictor - Prevêe a classe por linha com base no modelo aprendido. A probabilidade de classe é o produto da probabilidade por atributo e a probabilidade do próprio atributo de classe. Ao observar o fluxo, pode-se verificar que neste nó entra os 30 % do conjunto de dados que será usado para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

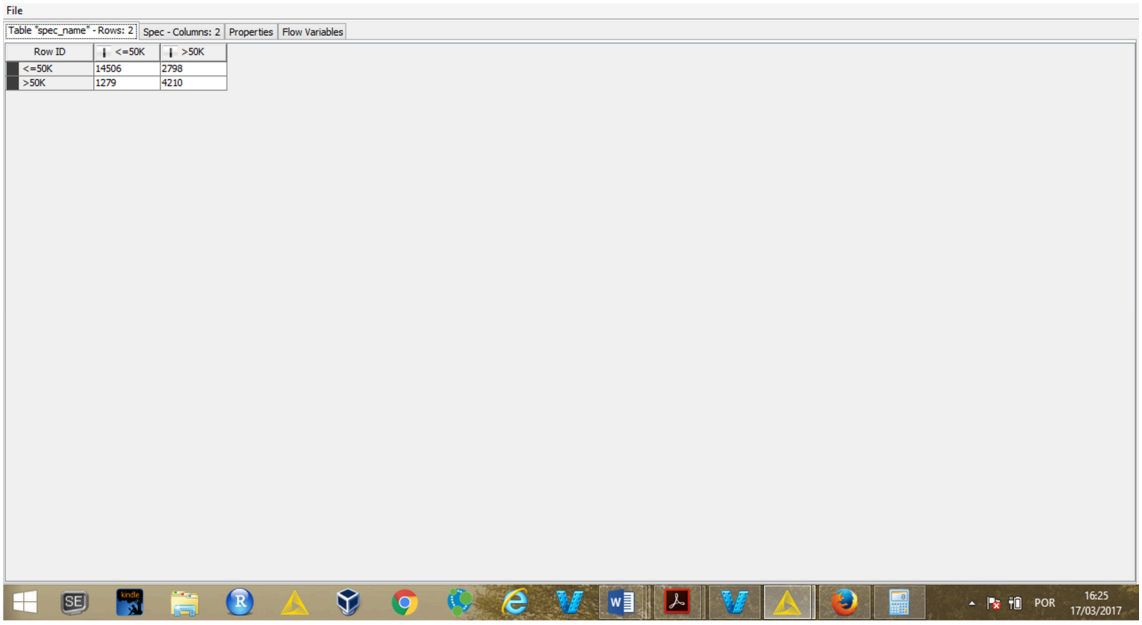
Após a execução do fluxo pode-se verificar no nó ***Scorer,*** o resultado do modelo criado ***Naive Bayes Predictor ,*** aplicado ao conjunto de dados.



Conclusão

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

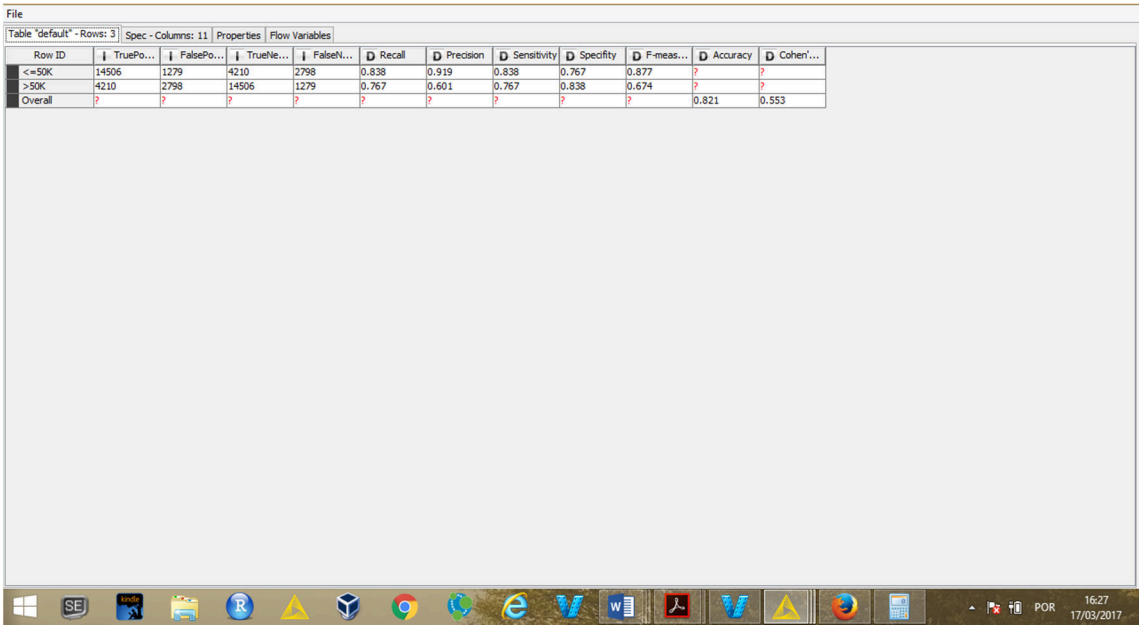
O desempenho do modelo ***Naive Bayes Predictor*** em relação à predição para as pessoas que ganham acima de 50k (>50 k) por ano é o seguinte :



A aba Confusion Matrix do Nó Scorer demonstra que temos

4.210 - Verdadeiro-Positivos , 2.798 - Falso-Positivos

14.506 - Verdadeiro-Negativos , 1.279 - Falso-Negativos.



Já a aba Accuracy statistics do Nó Scorer demonstra que temos

Temos uma sensibilidade(sensitivity) de 76% - 0,767 (VP / VP + FN) , uma especificidade (specificity) de 83% - 0.838 (VN / FP + VN), uma acurácia (accuracy) de 82% - 0.821 (VP + VN) / (VP+FP+FN+VN).