ENGENHARIA DO CONHECIMENTO

Curso de Engenharia Informática, Regime Diurno e Pós-Laboral

3° Ano, 2° Semestre, Ano Letivo 2019/2020

Tutorial de Introdução ao RapidMiner

1. Introdução

Considere que um *data miner* ou *data scientist* está a realizar um projeto de investigação para determinar o melhor medicamento a utilizar no tratamento de uma doença. Ao longo da investigação foram recolhidos dados de um conjunto de pacientes que sofreram da doença em estudo. Durante o tratamento cada paciente reagiu positivamente a um de cinco medicamentos incluídos no estudo.
Uma parte do trabalho do projeto de investigação, que deve agora efetuar, consiste em usar técnicas de *data mining* para determinar qual dos medicamentos poderá ser utilizado com mais sucesso num caso idêntico que ocorra no futuro.

Os dados de cada paciente a ter em conta neste projeto são os seguintes:

- Age – Idade: Number
- Sex – Sexo: M ou F
- BP (Blood Pressure) – Pressão Arterial: HIGH, NORMAL ou LOW
- Cholesterol – Nível de colesterol no sangue: NORMAL ou HIGH
- Na – Concentração de sódio no sangue: Numérico
- K – Concentração de potássio no sangue: Numérico
- Drug – Medicamento prescrito ao qual o paciente reagiu positivamente: Nominal

Nota: Neste tutorial deverá utilizar a ferramenta computacional RapidMiner. Os ficheiros necessários à realização do tutorial encontram-se na página da unidade curricular, plataforma Moodle.

2. Criar um Projeto e Importar Dados

Para a realização deste exercício iremos começar por utilizar um ficheiro CSV (*Comma Separated Values*), pelo que depois de criar um novo projeto deve utilizar o operador *Read CSV*, que se encontra na árvore dos operadores do RapidMiner (*Data Access -> Files -> Read*).

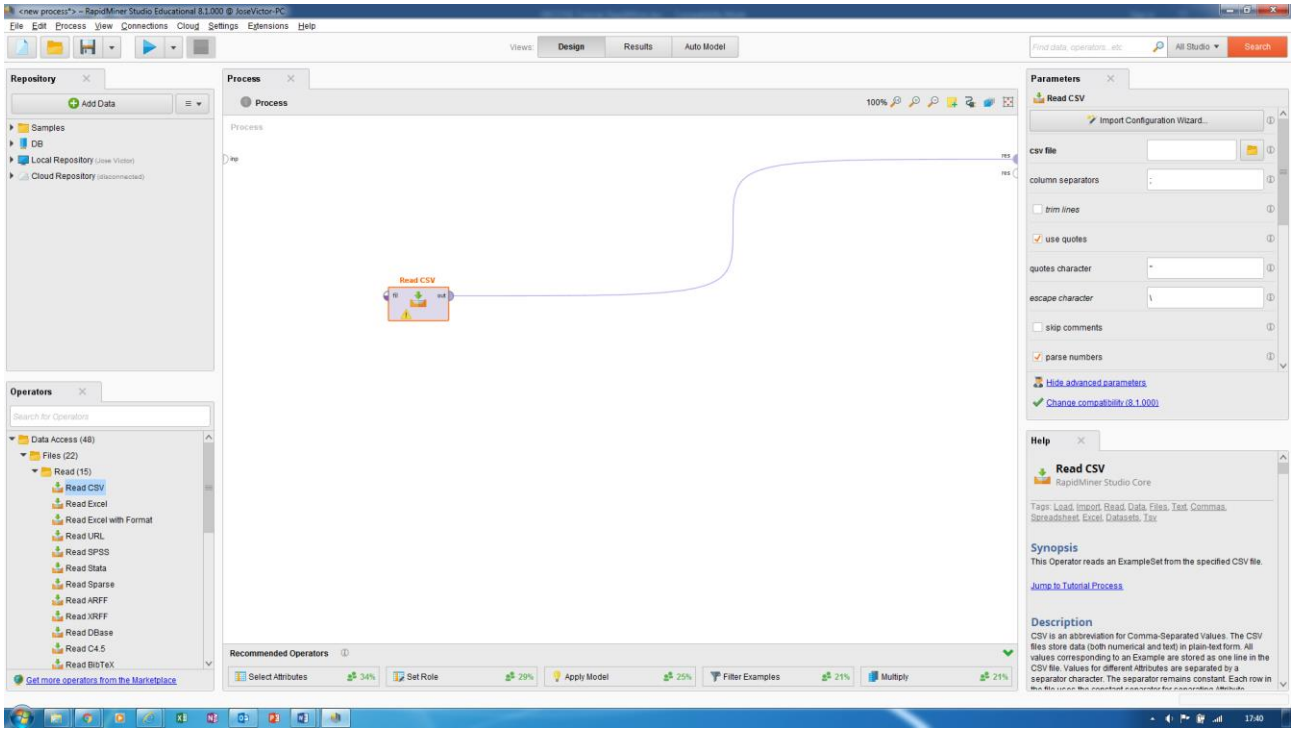

Figura 1: Importar os dados do ficheiro CSV.

- Configurar o operador *Read CSV*, através do *Import Configuration Wizard*, selecionar o ficheiro *DRUG1n.csv*, e seguir os passos do *Wizard*.

Após importar os dados deverá executar o processo de modo a verificar se os dados foram importados corretamente.

3. Visualização dos Dados

Na opção *Data* é possível observar os dados importados na forma de uma tabela com os vários atributos ou características nas colunas e os registos ou amostras nas linhas, Figura 2.
De seguida, na opção *Statistics*, verifique os valores que os vários atributos podem tomar, a sua proporção e a concordância com o tipo de dados apresentado na descrição do problema.

4. Exploração dos Dados

A representação visual dos dados é uma grande ajuda na interpretação do seu significado, como tal é frequente recorrer-se a gráficos para facilitar a sua análise.

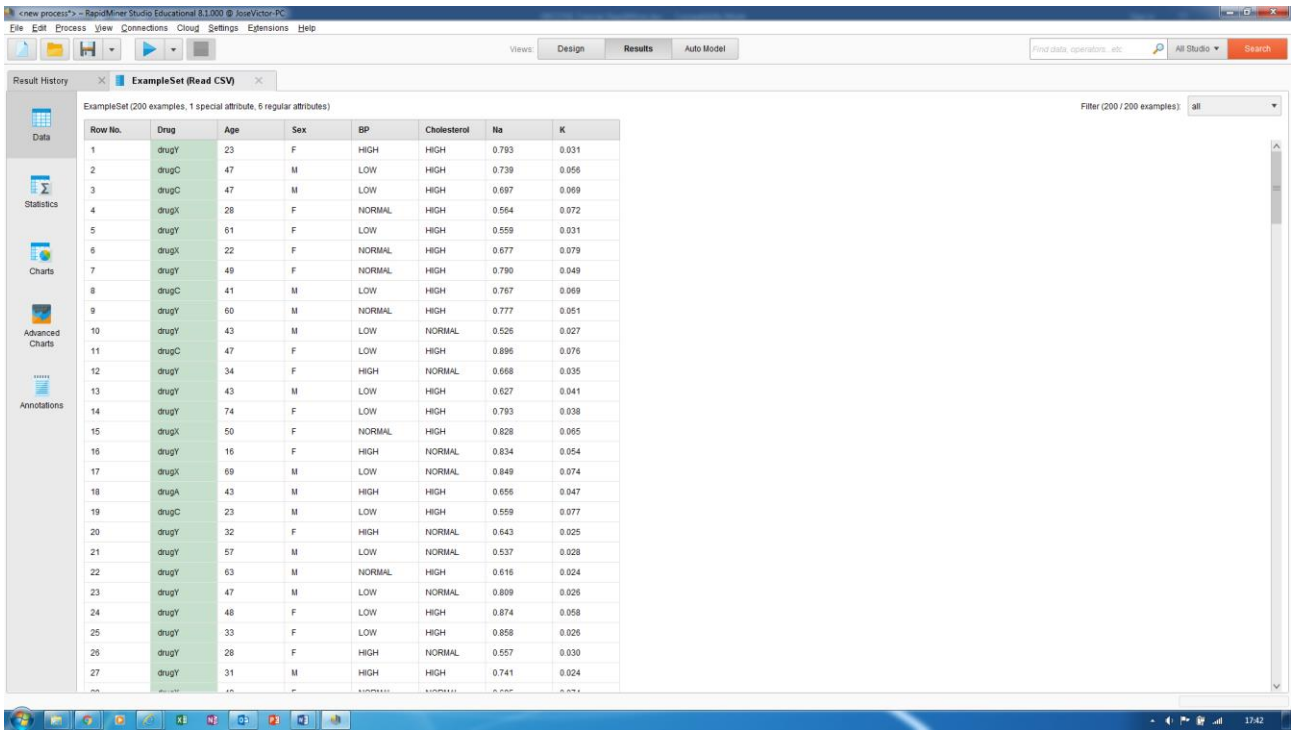

Figura 2: Resultado da importação dos dados do ficheiro CSV.

O RapidMiner dispõe de ferramentas que permitem visualizar graficamente os dados em análise. Antes de avançar para o passo seguinte, explore os diversos atributos ou características, através da opção *Charts*. No parâmetro *Chart style* deverá indicar a opção *Scatter Matrix* e no *Plots* selecionar o atributo *Drug*, Figura 3.

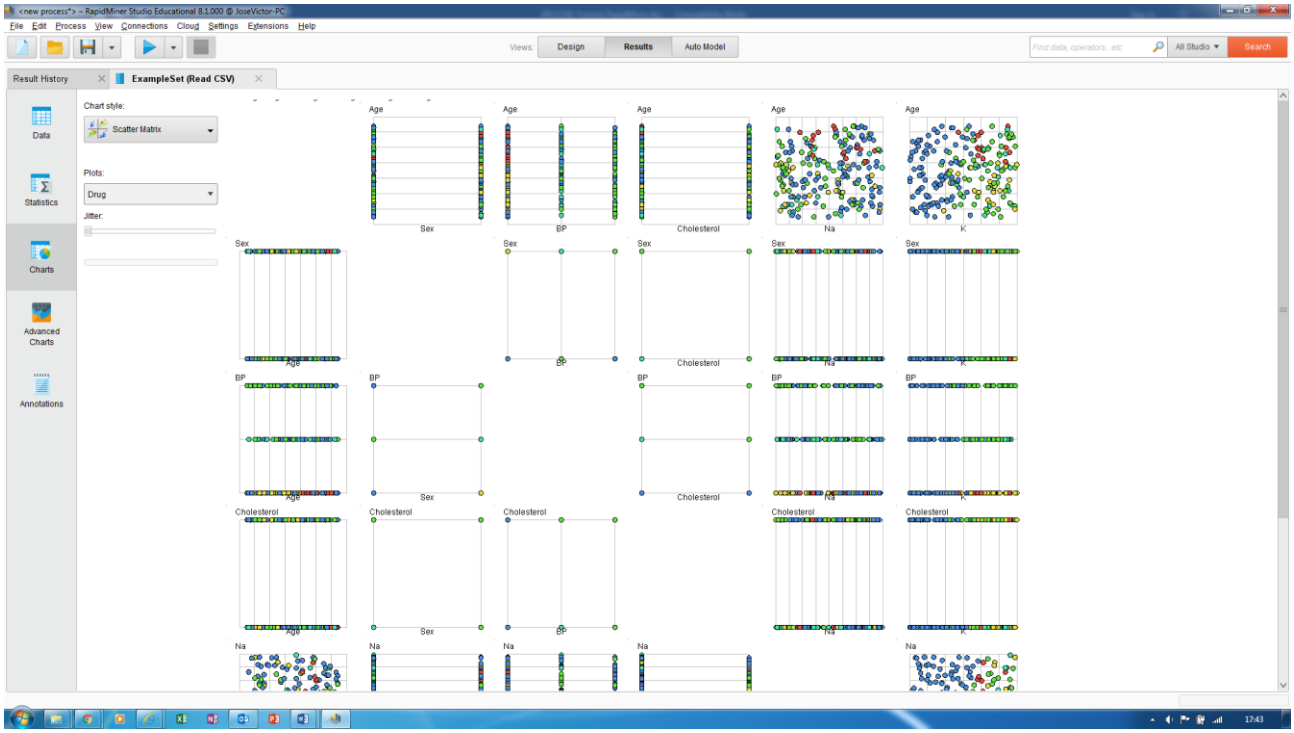

Figura 3: Visualização dos dados.

Aqui é possível determinar algumas coisas, como por exemplo a proporção de pacientes que reagiram a cada medicamento.

- Verifique qual o medicamento a que os pacientes reagiram maioritariamente.

Após esta etapa vamos tentar perceber quais os fatores que podem ter influenciado a característica objetivo *Drug*. Conhecendo o domínio do problema sabemos que as concentrações de sódio e potássio podem ser fatores importantes que influenciam a experiência. Assim, vamos utilizar um novo gráfico que nos permita perceber de que forma estas variáveis influenciam a escolha do medicamento.

No *Chart style* selecione a opção *Scatter* e no *x-Axis* o atributo Na (Sódio), no *y-Axis* o atributo K (Potássio) e no *Color Column* a característica objetivo *Drug*, Figura 4.

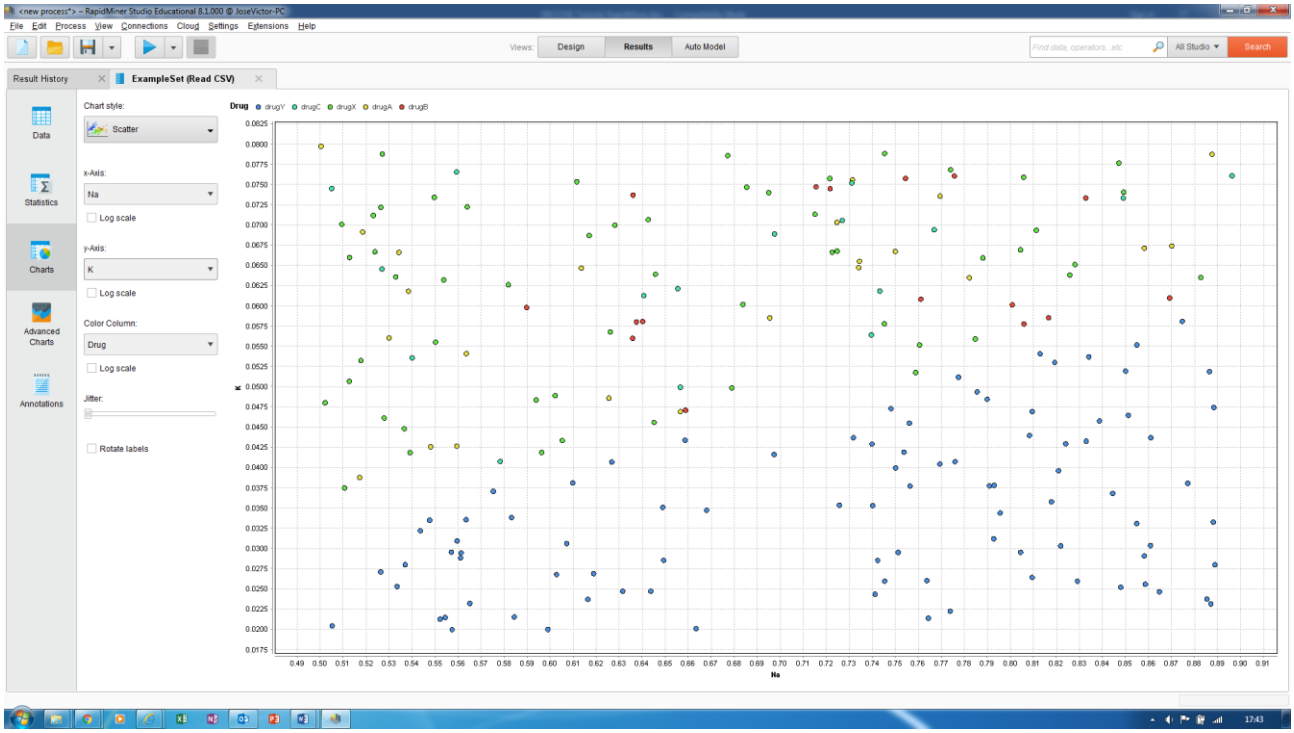

Figura 4: Visualização dos dados que relacionam os atributos Sódio e Potássio.

É possível alterar o tamanho de todos os gráficos, o tamanho dos pontos e "espalhar" os pontos de forma a conseguir visualizar pontos sobrepostos (opção *Jitter*). A opção *Jitter* é muito útil quando estamos a visualizar dados de atributos cujos valores não são numéricos.

- Visualize o gráfico que relaciona o Sódio (Na) vs Potássio (K) e procure estabelecer uma relação entre a concentração de potássio e sódio no sangue e a influência que esta exerce sobre o sucesso do medicamento. Tenha em atenção que no parâmetro *Color Column* deverá estar selecionada a característica objetivo *Drug*.

5. Preparação dos Dados

Até ao momento, apenas foram executadas algumas operações sobre os dados existentes utilizando tabelas e gráficos. No entanto, esta análise simples já conseguiu estabelecer um ponto de partida para a análise em foco, conseguindo-se obter uma relação entre dois atributos (Na e K). A partir daqui vamos proceder à preparação dos dados para a etapa de *data mining*.

Da análise do gráfico anterior concluímos que existe uma relação de rácio entre o Sódio e o Potássio (Na/K). Devido à informação contida nesta relação é útil e necessário criar um novo atributo que represente este rácio para cada registo. Deste modo, numa fase posterior podemos utilizar este valor na construção de um modelo de predição que permita saber em que situações utilizar cada um dos cinco medicamentos.

Para proceder a alterações nos dados é necessário recorrer a operadores específicos que permitem transformar os dados. Estes operadores permitem manipular os dados ao nível dos atributos, dos registos e dos valores dos atributos. Os operadores encontram-se organizados por categorias numa árvore e para selecionar um operador é necessário arrastá-lo para a área de trabalho ou desenho.

Neste exercício, será utilizado um operador para criar um novo atributo, resultante de atributos já existentes. Esta é uma etapa muito comum nos projetos de *data mining*.

Para o efeito arraste para a área de desenho do projeto o operador *Generate Attributes*, que se encontra na árvore dos operadores do RapidMiner (*Blending* -> *Attributes* -> *Generation*) e proceda à sua configuração, Figura 5.

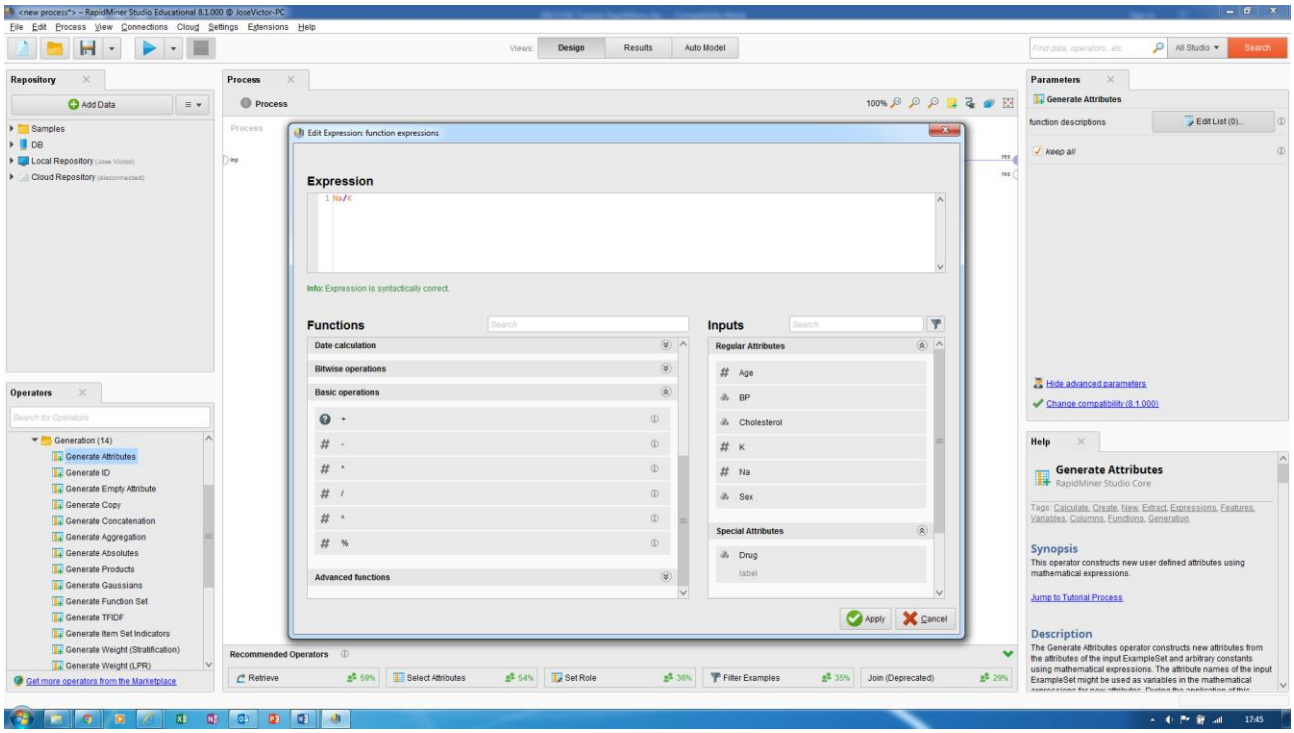

Figura 5: Seleção da configuração do filtro.

Na configuração do operador realize os seguintes passos:

- Atribua o nome *Na_K* ao atributo derivado;
- Preencha a fórmula correspondente (*Na/K*), selecionando os atributos e o respetivo operador.

Execute o processo e verifique que na área de resultados surge o novo atributo Na_K. De seguida no separador *Charts* crie um histograma, opção *Bars* no *Chart style*, com o número de ocorrências de cada medicamento em relação ao valor obtido com o campo derivado anteriormente, e tire as conclusões relativas à interpretação dos dados.

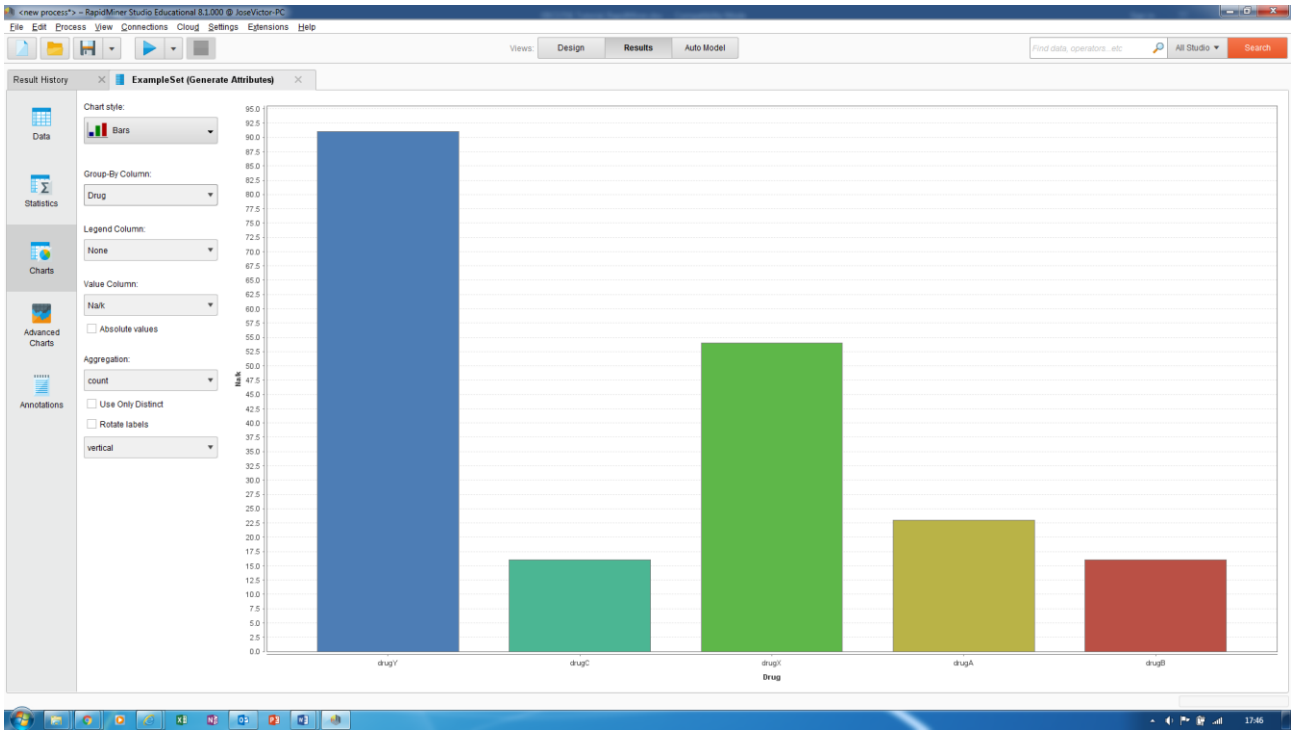

Figura 6: Histograma com o atributo Na_K em função da característica objetivo.

Até ao momento, explorando e manipulando os dados existentes, já foi possível obter algumas conclusões em relação aos dados que nos permitem formular algumas hipóteses. Por exemplo, o rácio entre o sódio e o potássio parece afetar a escolha do medicamento a aplicar. No entanto ainda não é possível explicar todas as relações existentes.

6. Criação do Modelo de Predição

Recapitulando, até ao momento já foi possível verificar a existência de alguns padrões nos dados. O rácio Na/K afeta a escolha do medicamento, no entanto as relações entre os dados ainda não estão completamente claras. Neste passo vamos construir um modelo que nos permita obter as respostas pretendidas. Para o efeito vamos utilizar um algoritmo baseado em árvores de decisão (*Decison Tree*).
Para gerar um modelo ainda é necessário efetuar alguma preparação adicional aos dados. Como estamos a utilizar um atributo derivado (*Na_K*), é preciso filtrar os atributos originais, desta forma estes não serão usados duas vezes no algoritmo de *data mining*.

- Remover os atributos de Na e K.

Para o efeito arraste para a área de desenho do projeto o operador *Select Attributes*, que se encontra na árvore dos operadores do RapidMiner (*Blending -> Attributes -> Selection*) e proceda à sua configuração, Figura 7.

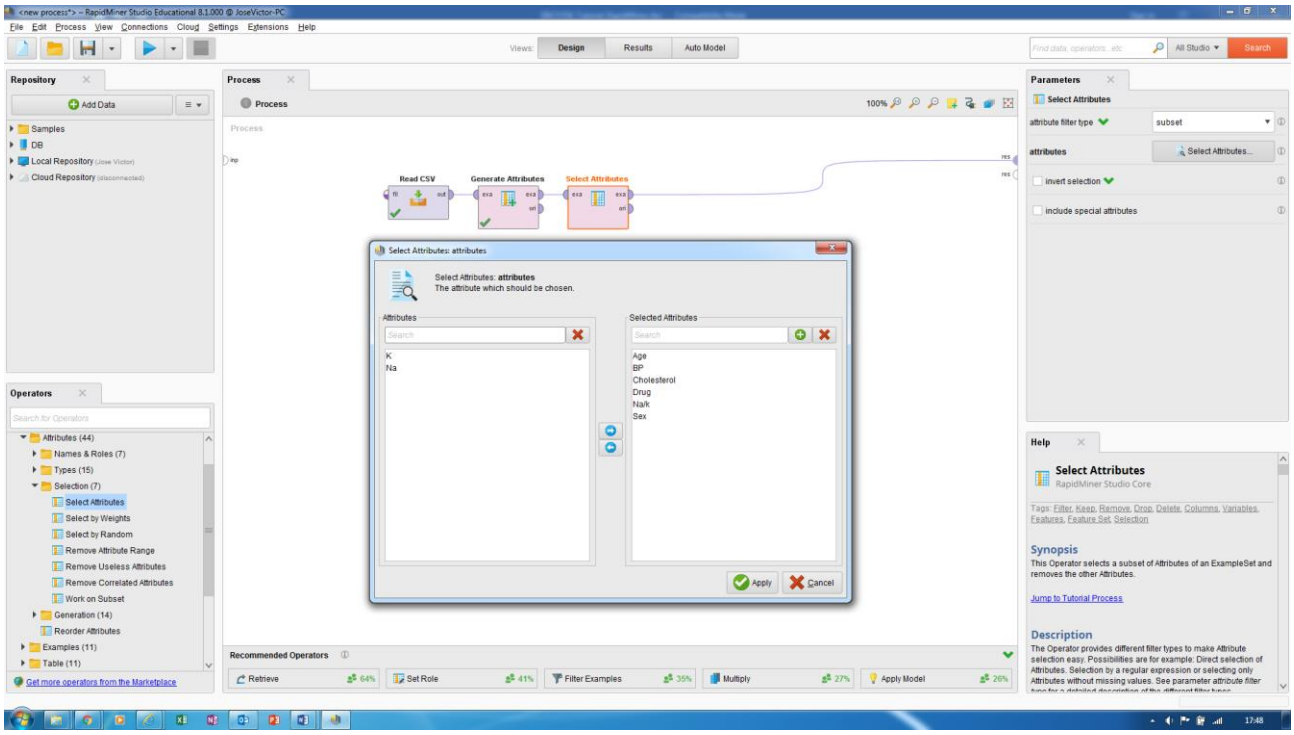

Figura 7: Remoção dos atributos Na e K.

Para construir um modelo, utilizando um algoritmo de árvores de decisão, é necessário acrescentar vários operadores, a saber:

- *Decision Tree* (*Modeling -> Predictive -> Trees*)
- *Apply Model* (*Scoring*)
- *Performance Classification* (*Validation -> Performance -> Predictive*)

Com a ajuda do docente do seu turno desenhe a solução do projeto e configure os vários operadores, Figura 8.

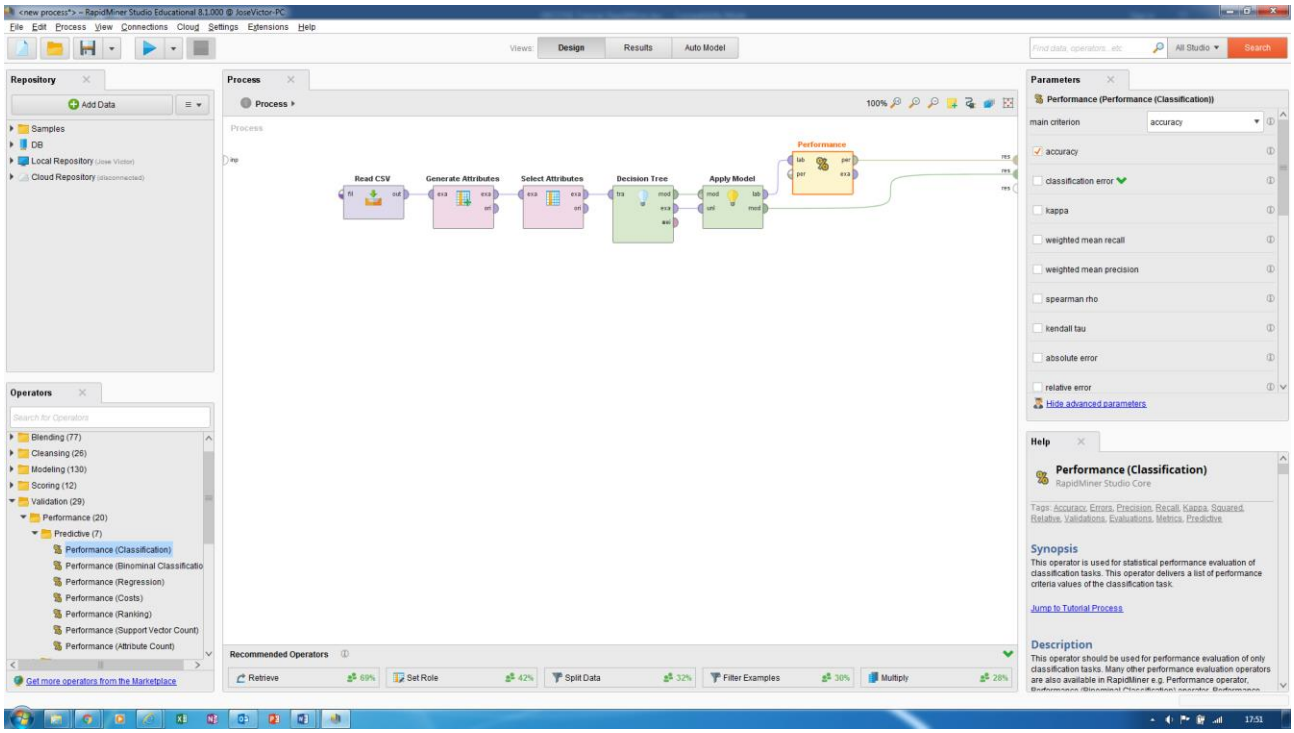

Figura 8: Diagrama para a construção do modelo.

Execute o processo e analise os resultados obtidos, em modo gráfico e texto, nos seguintes separadores *Performance Vector* (*Performance*) e *Tree* (*Decision Tree*), Figura 9.

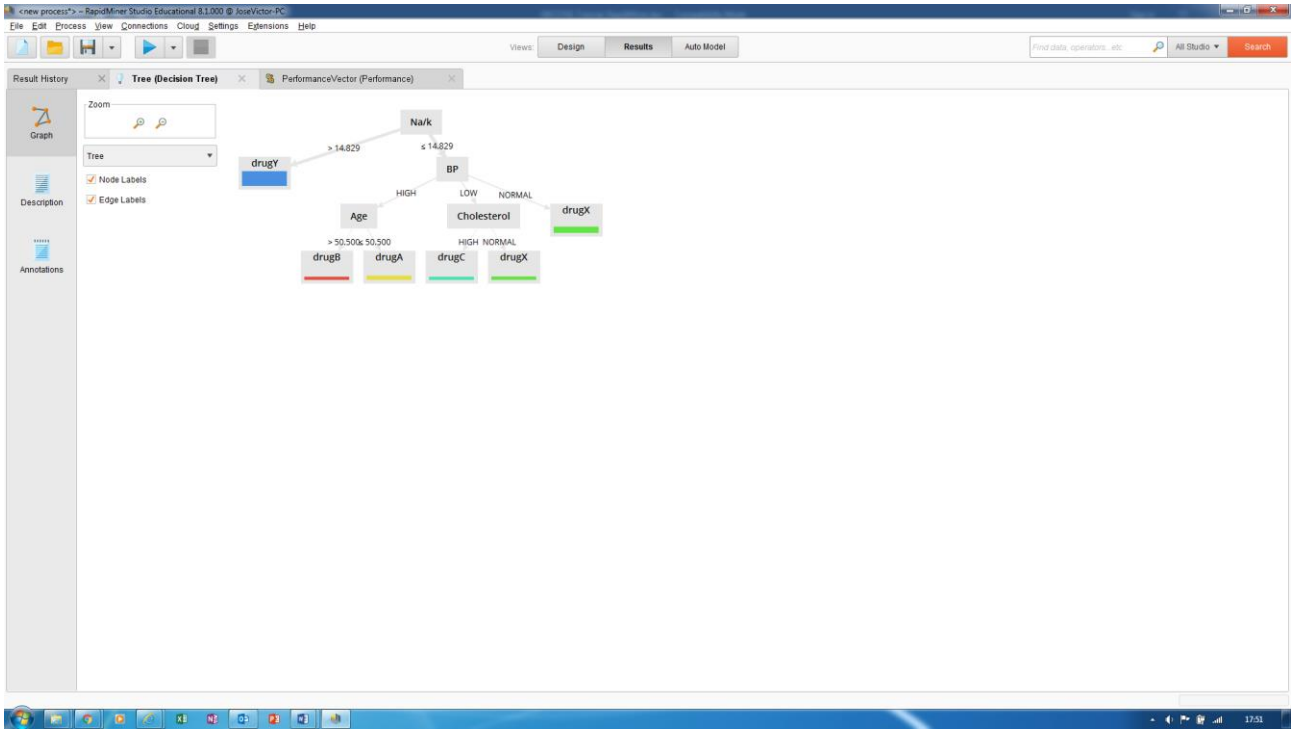

Figura 9: Visualização do modelo obtido.

O conjunto de regras obtido fornece as peças que faltavam ao puzzle para a informação obtida anteriormente fazer sentido.

- Interprete as regras obtidas.
- Visualize estas regras na sua representação gráfica e em texto (repare que aqui é possível observar com maior facilidade o número de casos em que cada medicamento foi bem sucedido).

Repare que neste caso o modelo tem uma precisão de 100%. Isto acontece pois este *data set* é académico, tendo sido elaborado apenas para a execução deste exercício introdutório, e porque na construção do modelo fizemos "batota". Em projetos reais dificilmente se verifica esta situação, sendo a avaliação uma questão fulcral para determinar a utilidade ou não de um modelo.

Experimente agora utilizar a opção de teste *Cross Validation*, para separar os dados de treino (utilizados na construção do modelo) dos dados de teste (utilizados na validação do modelo), Figura 10 e Figura 11.

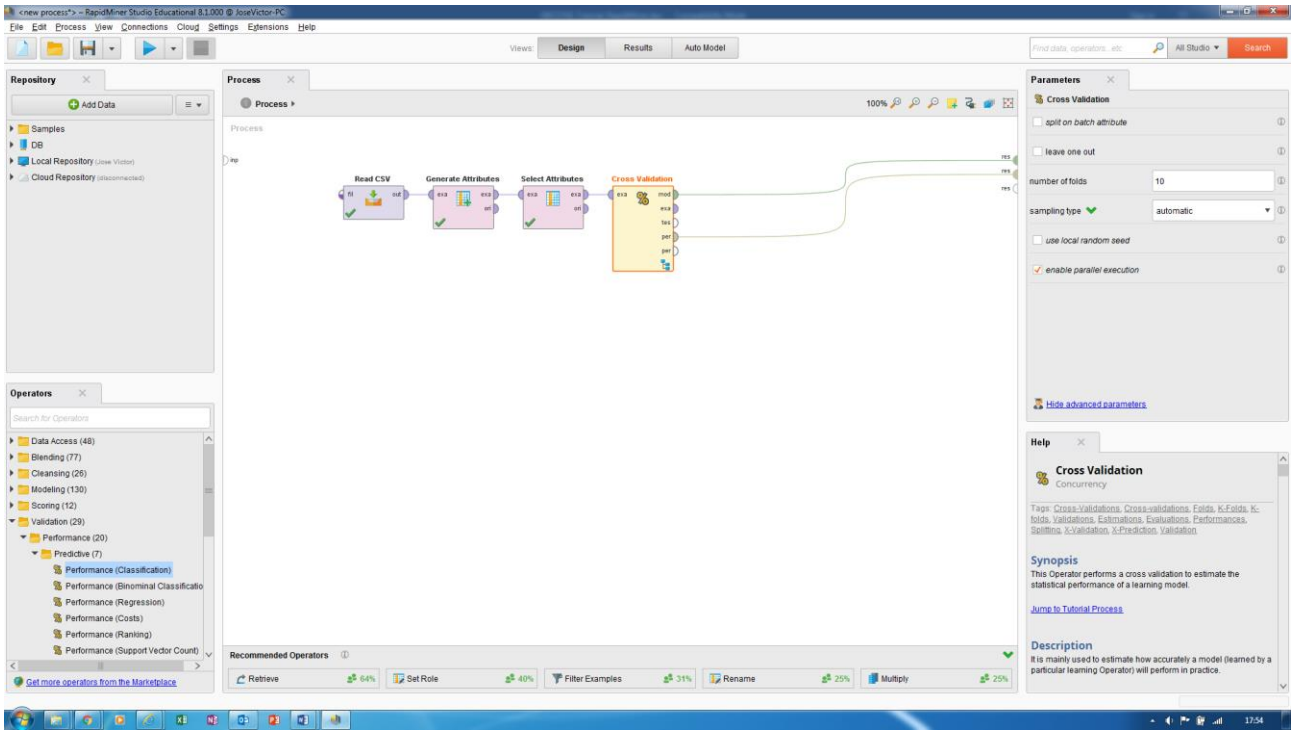

Figura 10: Diagrama para a construção do modelo.

A precisão do modelo baixou para 99%, o que significa que há duas amostras que foram mal classificadas.

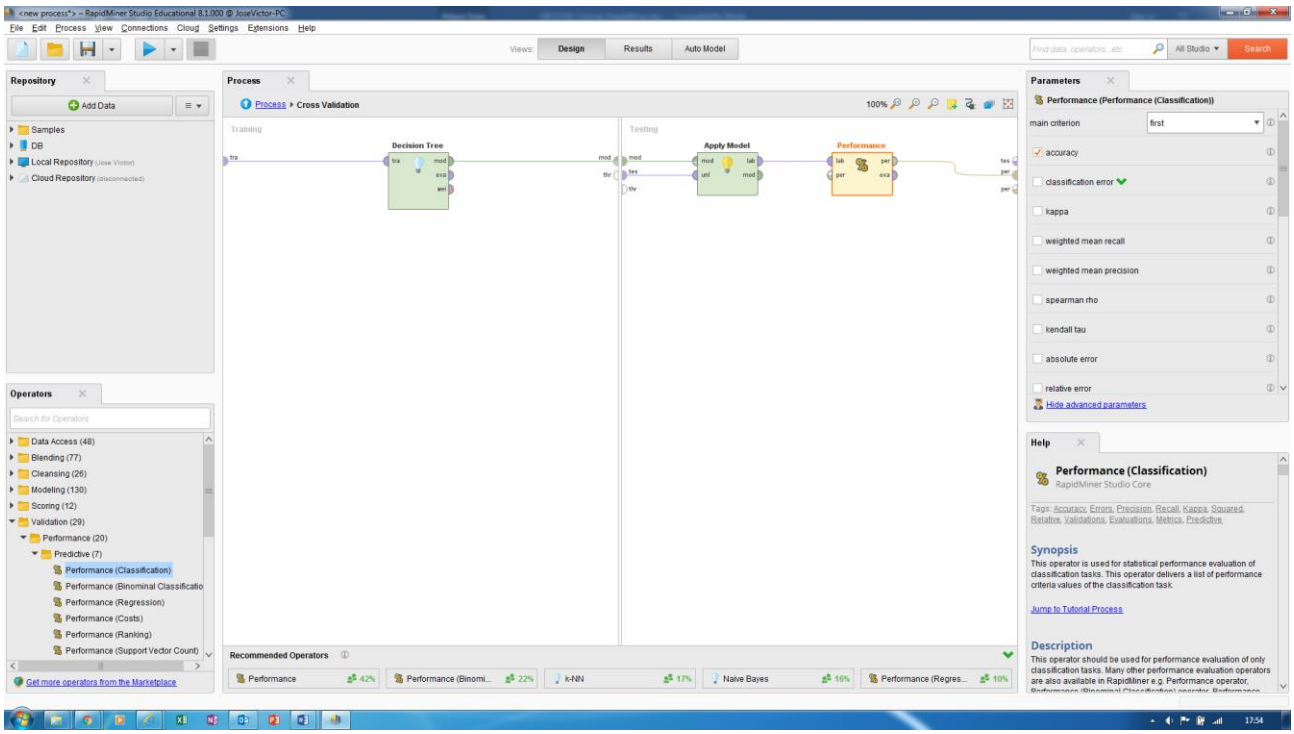

Figura 11: Configuração do operador *Cross Validation*.

Para concluir o tutorial de introdução ao RapidMiner construa um projeto de modo a que sejam utilizados ficheiros de dados separados para a construção e avaliação do modelo, Figura 12.

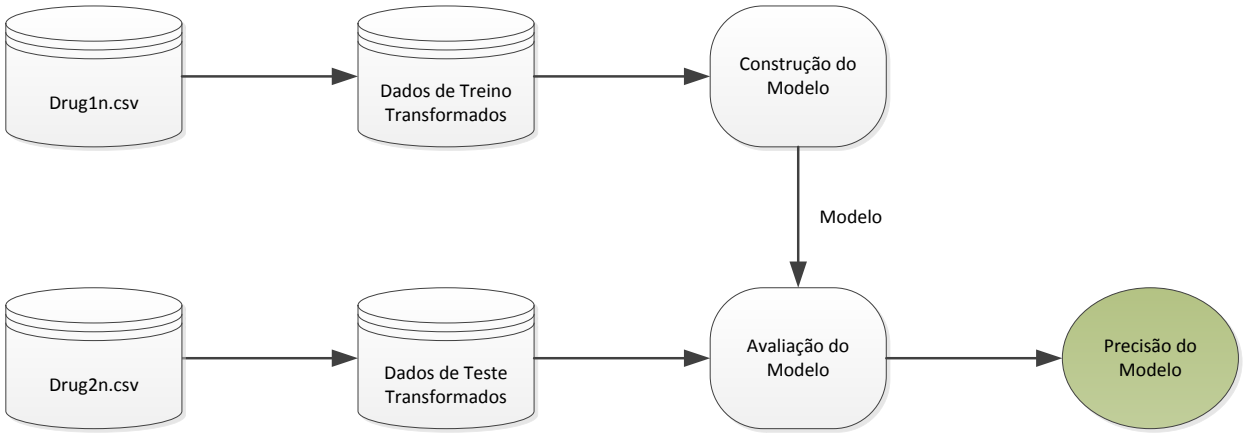

Figura 12: Esquema de construção e avaliação do modelo.

Nota: Todas as transformações levadas a cabo nos dados têm de ser realizadas em ambos os ficheiros.