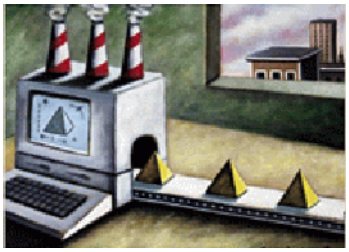
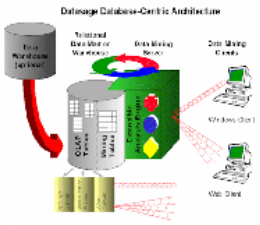
AID (Clementine – 3.ª Parte)

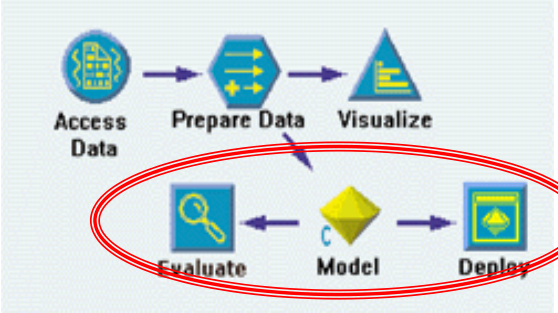


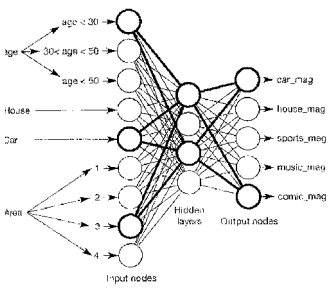


AID (Clementine – 3.ª Parte)

Redes Neuronais

- *Introdução ao Nó Train Net*
- *Criar uma Rede Neuronal*
- *Introduzir a paleta Modelos Gerados*
- *Investigar e interpretar os resultados*
- *Avaliar o modelo*





Introdução à Rede Neuronal Retropropagada

O Nó Train é utilizado para a criar este tipo de rede neuronal

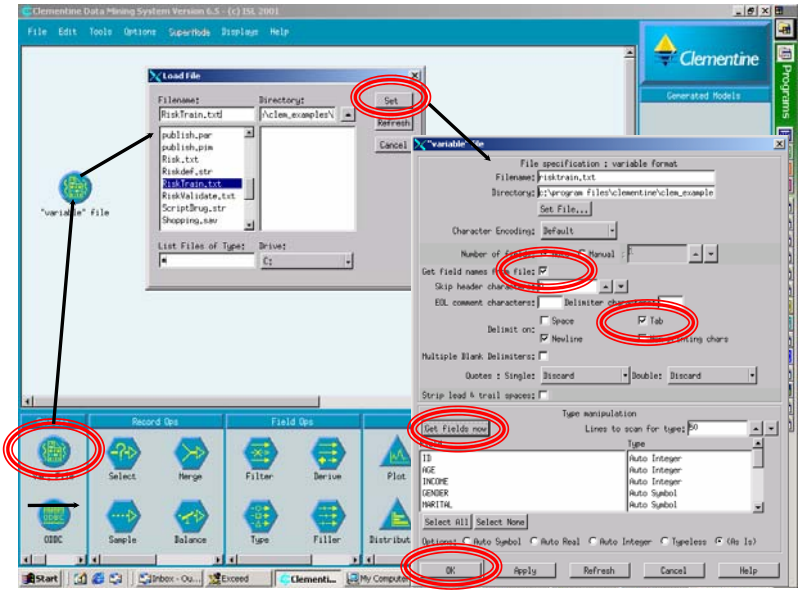
Uma vez a rede treinada, surge na paleta Modelos Gerados um nó Generated Net

- **representa a rede neuronal criada**
- **as suas propriedades podem ser visualizadas**
- **para gerar predições -> passar os novos dados pelo nó gerado**

Antes de um fluxo de dados poder passar pela Nó Train, ou outro da paleta Modelos, deve passar primeiro pelo Nó Type (para aqui se indicar o tipo e direcção de cada campo).

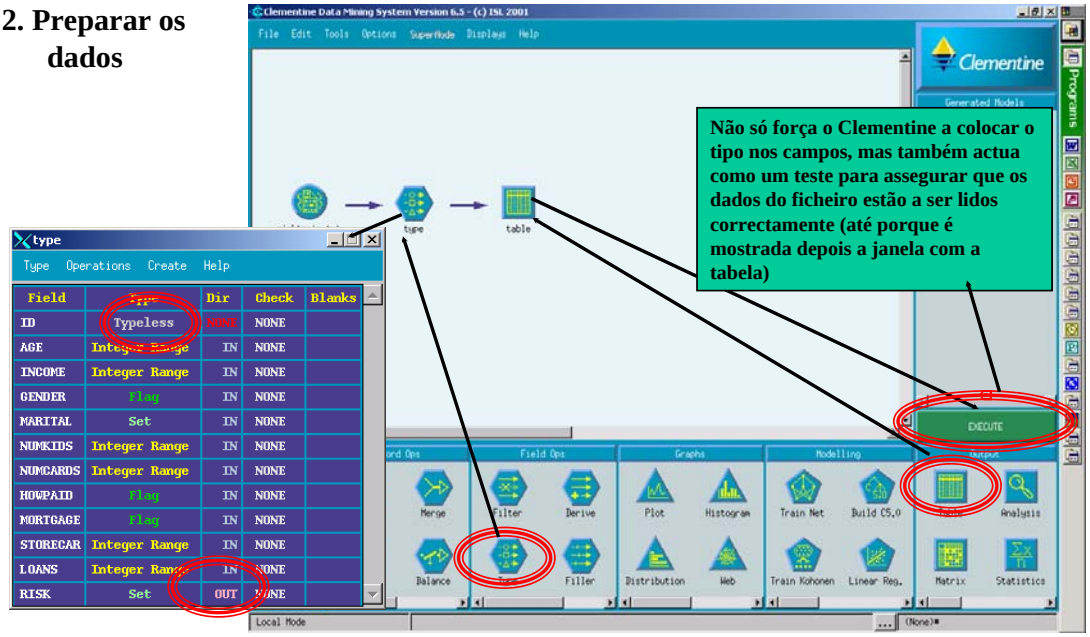
Nó Rede Neuronal

1. Ler os dados



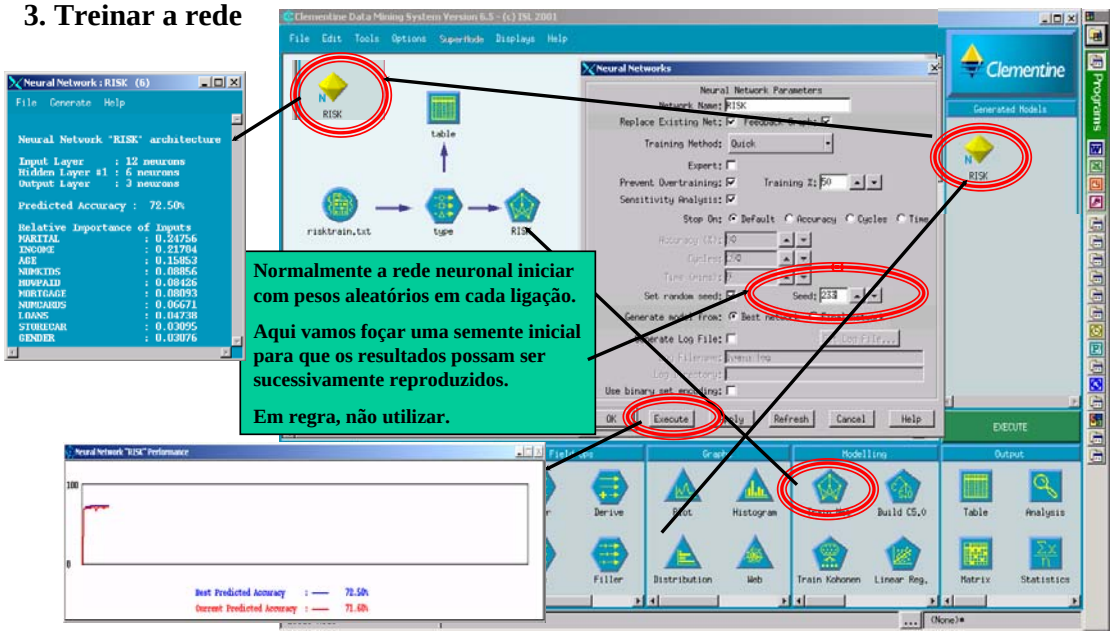
Nó Rede Neuronal

2. Preparar os dados



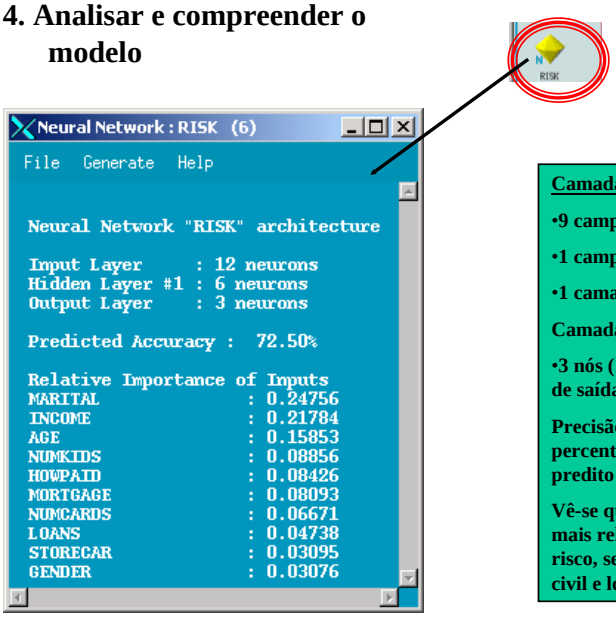
Nó Rede Neuronal

3. Treinar a rede



Nó Rede Neuronal

4. Analisar e compreender o modelo



Camada de entrada:

- **9 campos numéricos ou flag (9 nós)**
- **1 campo set com 3 valores (3 nós)**
- **1 camada escondida com 6 nós**

Camada de saída:

- **3 nós (1 por cada valor da variável de saída – Risco)**

Precisão: 72.50% mostrando a percentagem do conjunto de teste predito correctamente

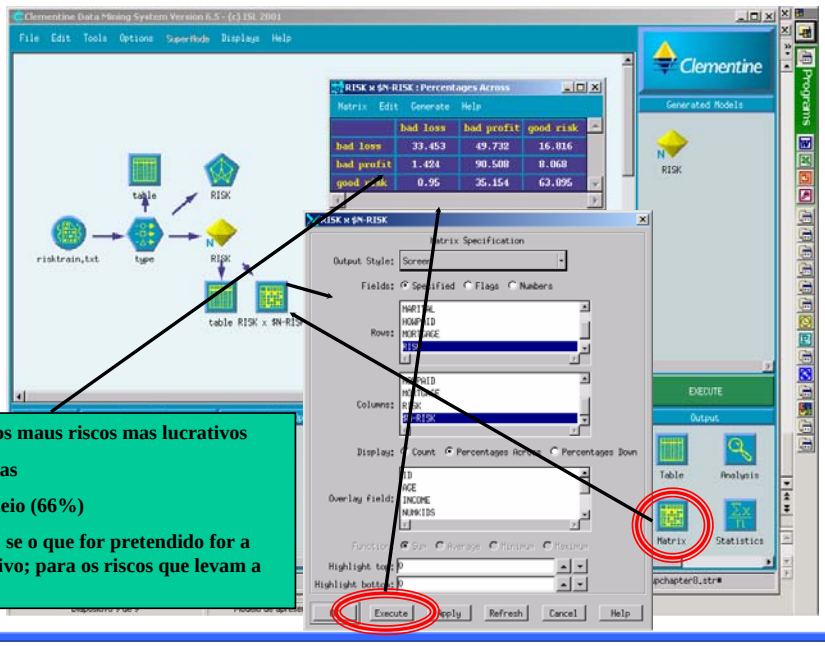
Vê-se que o campo rendimento é o mais relevante para a predição do risco, seguido de perto pelo estado civil e logo a seguir, a idade.

Nó Rede Neuronal

4. Analisar e compreender modelo

Comparar valores preditos com os valores reais:

- **Utilizar uma matriz que nos mostrará as percentagens de acerto do modelo para cada tipo de risco**



Correcção de predição em 90% dos maus riscos mas lucrativos

Só 33% dos maus que geram perdas

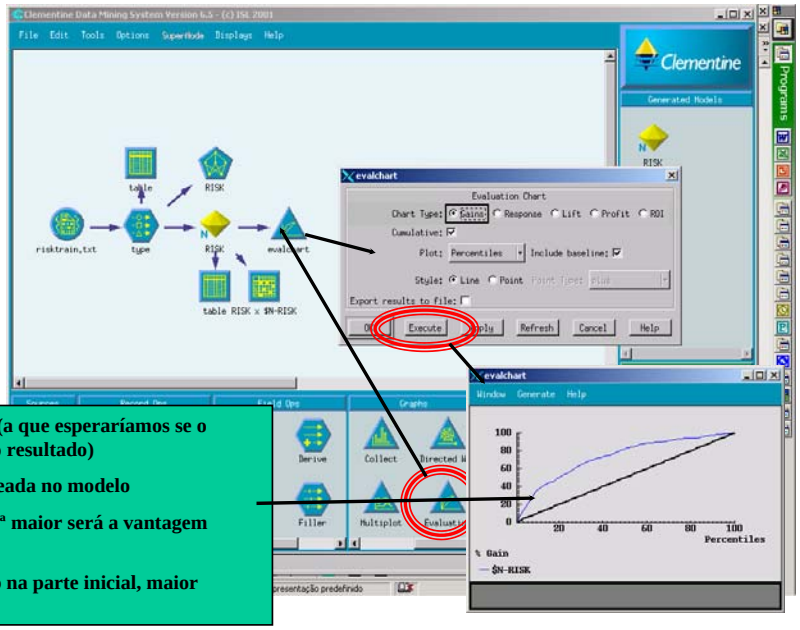
Os bons e lucrativos, algures no meio (66%)

Conclusão: o modelo é satisfatório se o que for pretendido for a predição de risco mau, mas lucrativo; para os riscos que levam a perdas, só cerca de 1/3.

Nó Gráfico de Avaliação

Oferece uma forma <u>fácil de avaliar e comparar</u> modelos preditivos por forma a escolher o melhor modelo para uma dada aplicação.

Mostram o desempenho dos modelos na predição de resultados dados.



A linha diagonal mostra a taxa base (a que esperaríamos se o modloe não fosse capaz de predizer o resultado)

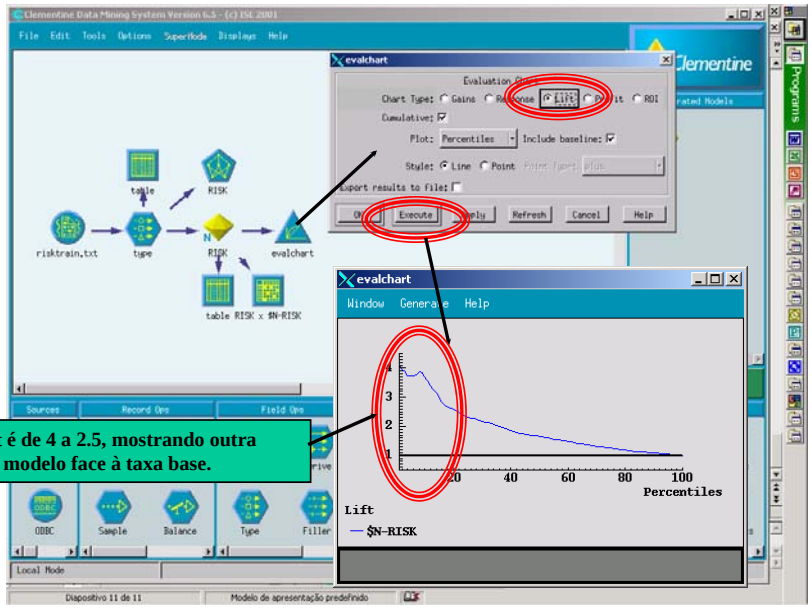
A linha curva representa a linha baseada no modelo

Quão mais a 2.ª linha se afastar da 1.ª maior será a vantagem obtida no uso do modelo.

Quão mais a 2.ª linha for do tipo step na parte inicial, maior sucesso terá o modelo na predição.

Nó Gráfico de Avaliação

Gráfico Lift – mostra o gráfico da relação da percentagem de registos em qualquer quantil que são acertos divididos pela percentagem total de acertos nos dados de treino.

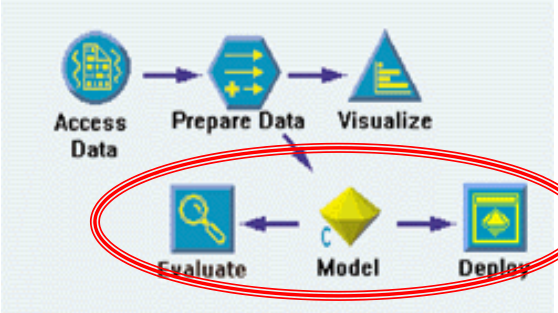


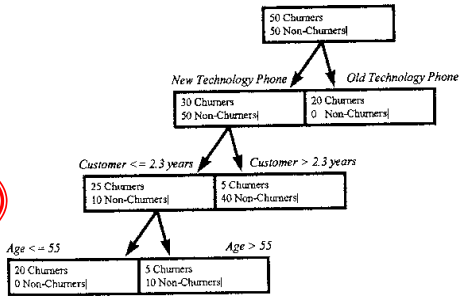
Nos 20 primeiros percentis, o lift é de 4 a 2.5, mostrando outra medida da vantagem relativa do modelo face à taxa base.

AID (Clementine – 3.ª Parte)

Árvores de Decisão (Regras de Indução)

- *Introdução aos dois nós de Árvores de Decisão*
- *Criar uma Árvore C5.0*
- *Analisar e Interpretar os resultados*
- *Criar um conjunto de regras para visualizar as regras de indução de uma forma diferente*





Árvores de Decisão: Algoritmos Disponíveis

Dois algoritmos:

Build C5.0

C&R Tree

- **Ambos constróem uma árvore de decisão através da divisão recursiva dos dados em subgrupos definidos através de campos preditores pela forma como se relacionam com a saída.**
- **Ambos geram grandes árvores e invocam a "poda" posteriormente, ainda que com critérios diferentes.**
- **Ambos permitem valores em falta, embora usem diferentes métodos para os suportarem.**

Árvores de Decisão: Diferenças

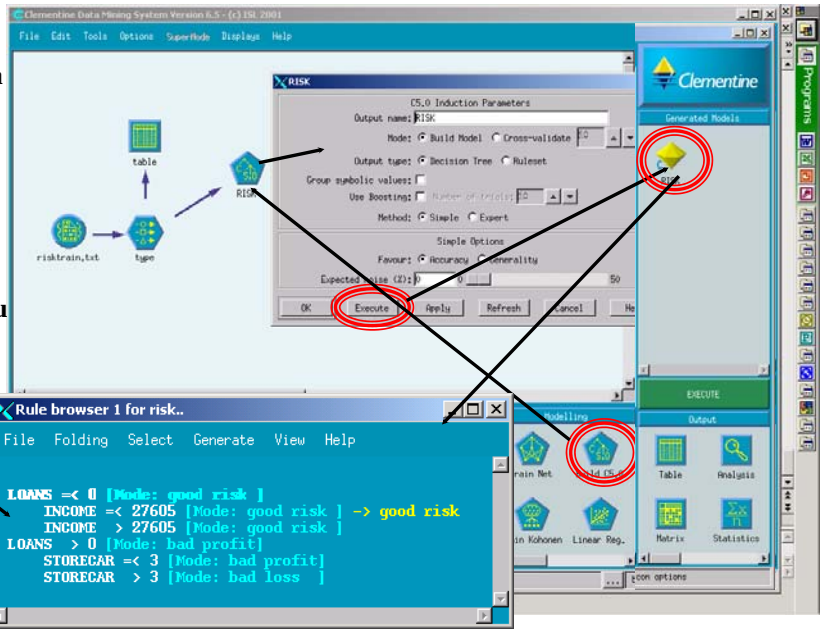
- **C5.0:**
  - **apenas permite output simbólicos**
  - **produz soluções sob a forma de árvores de decisão ou sob a forma de um conjunto de regras**
  - **Suporta divisões com 2 ou mais subgrupos para campos preditores simbólicos**
  - **critério utilizado para a divisão: taxa de ganho de informação**
- **C&R Tree:**
  - **saída simbólica e numérica (daí se chamar tb. regressão)**
  - **só produz árvores de decisão**
  - **só suporta divisões binárias (2 grupos)**
  - **critério utilizado para a divisão: coeficiente Gini**

Instituto Superior Politécnico de VISEU
Escola Superior de Tecnologia
Departamento de Informática

Nó C5.0

Uma vez gerado o modelo, o resultado é colocado na paleta dos Modelos Gerados.

Contém o modelo das regras induzidas e pode ser mostrado sob a forma de árvore de decisão ou conjunto de regras.



Nem todos os ramos estão visíveis

Um novo ramo é mostrado através de indentação no texto

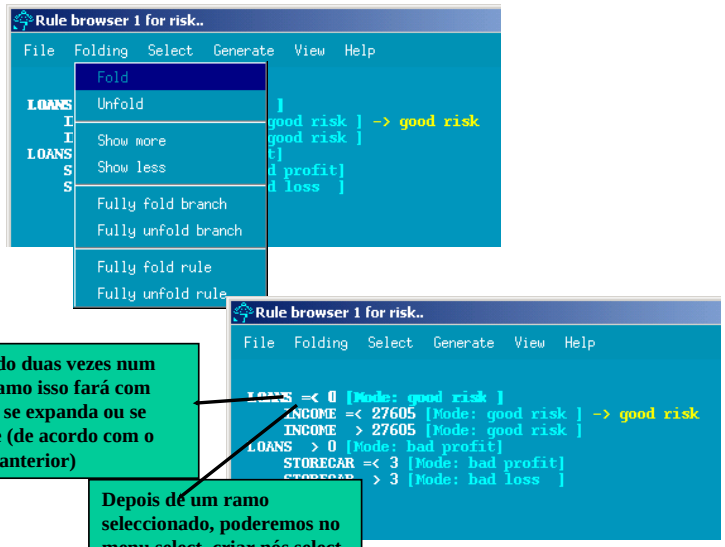
Condições aparecem a branco

Conclusões, a amarelo

Nó C5.0 (inspeccionar a árvore)

Depois de fazer o browse do modelo gerado, pode-se ver a árvore em maior ou menor detalhe:

- Fold (esconde o conteúdo de um dado ramo)
- Unfold (mostra a próxima divisão do ramo)
- Show more (expande todo o ramo)
- Show less (colapsa todo o ramo)
- Fully unfold rule (mostra toda a extensão da árvore de decisão)



Clicando duas vezes num dado ramo isso fará com que ele se expanda ou se colapse (de acordo com o estado anterior)

Depois de um ramo seleccionado, poderemos no menu select, criar nós select ou filter

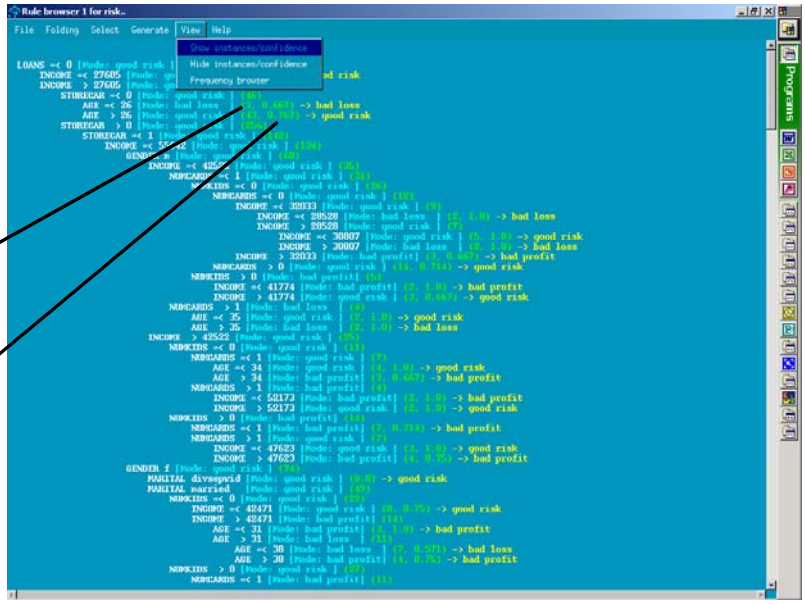
Nó C5.0 (Instâncias e Confiança)

Ao lado mostra-se a árvore na sua expansão máxima

No view poderemos indicar que pretendemos que sejam mostradas as:

Instâncias – número de registos usados para gerar um dado ramo

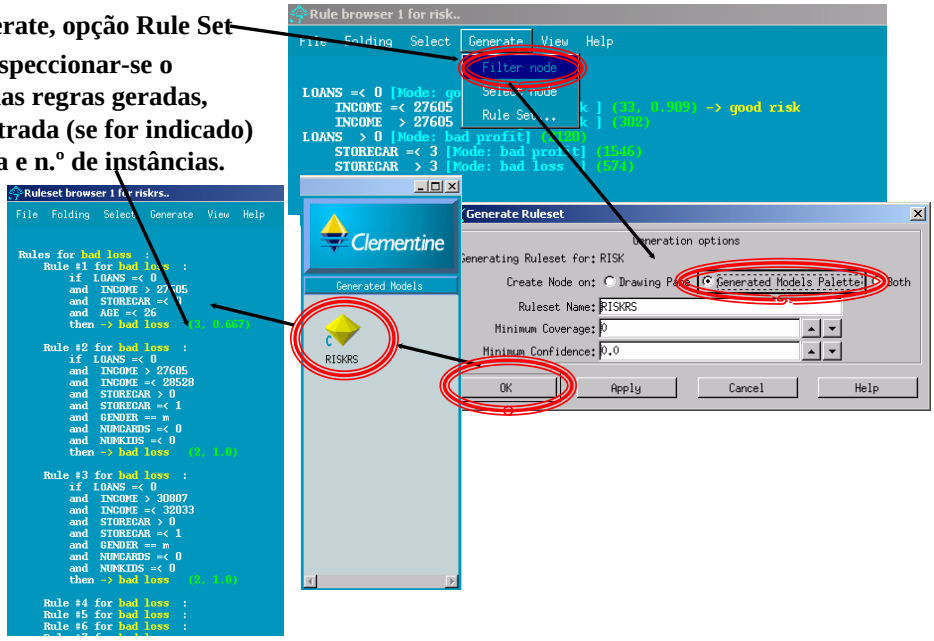
Confiança – grau de confiança de uma dada predição



Nó C5.0 (Gerar um Conjunto de Regras)

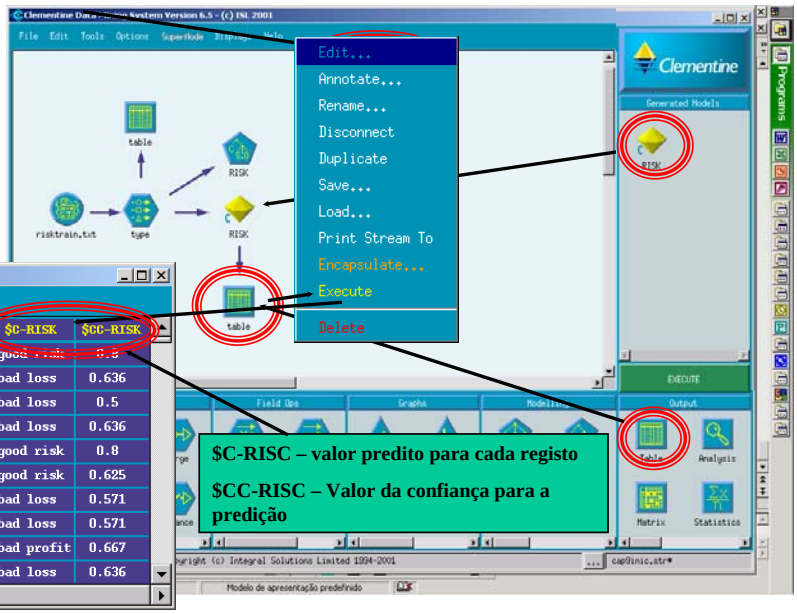
No menu Generate, opção Rule Set

Pode depois inspeccionar-se o conjunto das regras geradas, sendo mostrada (se for indicado) a confiança e n.º de instâncias.



Compreender as Regras e Determinar a Precisão

Contrariamente ao Nó Rede Neuronal, a precisão da predição no modelo de indução de regras não é dado directamente no nó C5.0.



MORTGAGE	STORECAR	LOANS	RISK	$C-RISK	$CC-RISK
y	2	0	good risk	good risk	0.8
y	1	0	bad loss	bad loss	0.636
y	1	1	bad loss	bad loss	0.5
y	1	0	good risk	bad loss	0.636
y	2	0	good risk	good risk	0.8
y	1	1	good risk	good risk	0.625
y	2	1	bad loss	bad loss	0.571
y	2	1	good risk	bad loss	0.571
y	1	1	bad profit	bad profit	0.667
y	1	0	bad profit	bad loss	0.636

$C-RISC – valor predito para cada registo

$CC-RISC – Valor da confiança para a predição

Comparar o Valor Real e o Predito

Utilizar um nó matrix e depois avaliar o modelo.

O modelo prediz correctamente:

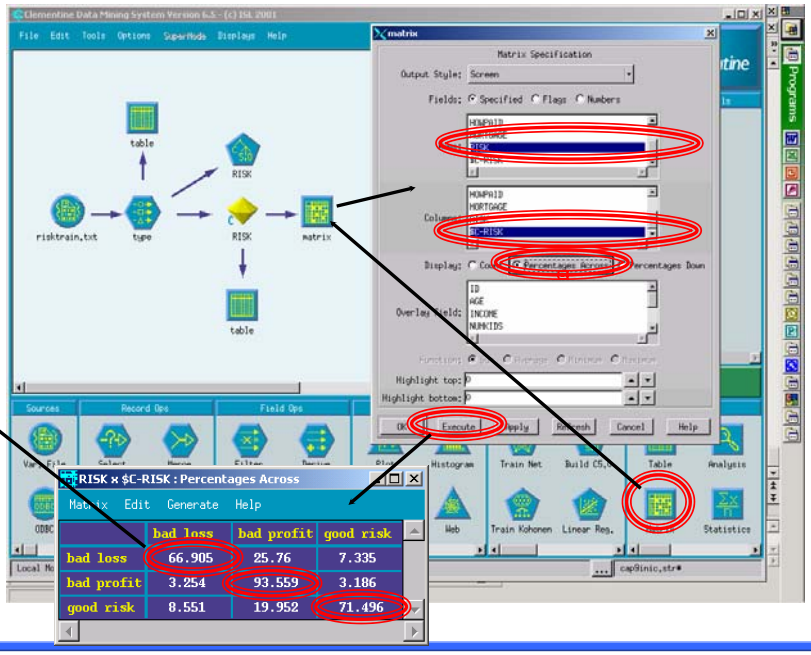
•71% de risco bom

•93.5 de risco mau, mas lucrativo

•67% do risco mau

Bastante melhor do que a rede neuronal, vista anteriormente

Se esta precisão se mantiver numa amostra de validação, será de preferir o modelo de árvore de decisão ou teremos de trabalhar para melhorar a rede neuronal



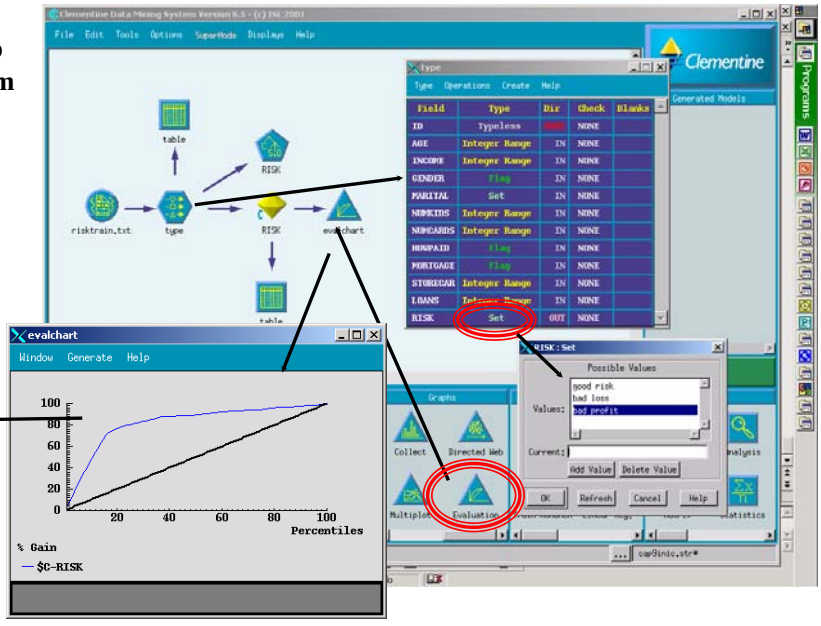
Instituto Superior Politécnico de VISEU
Escola Superior de Tecnologia
Departamento de Informática

Comparar o Valor Real e o Predito

Vamos agora avaliar o modelo em termos do valor de predição bom lucro.

Para o risco tipo good profit, a linha do ganho também cresce rapidamente, semelhante à linha do de ganho do gráfico anterior.

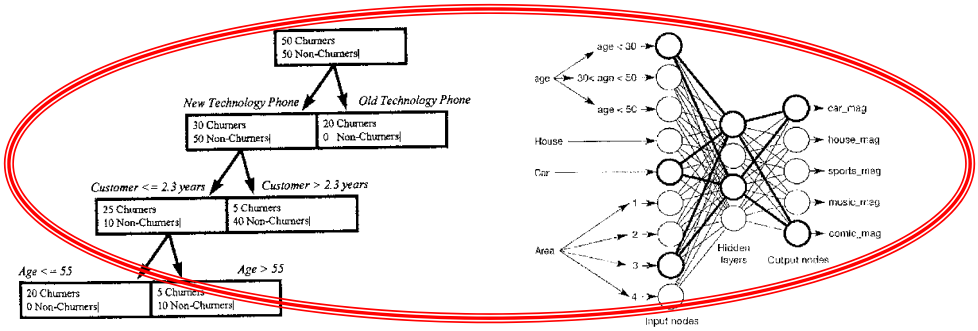
Neste caso os 20 percentis do topo, contêm cerca de 70% do risco bad loss.



AID (Clementine – 3.ª Parte)

Combinar Árvores de Decisão com Redes Neuronais

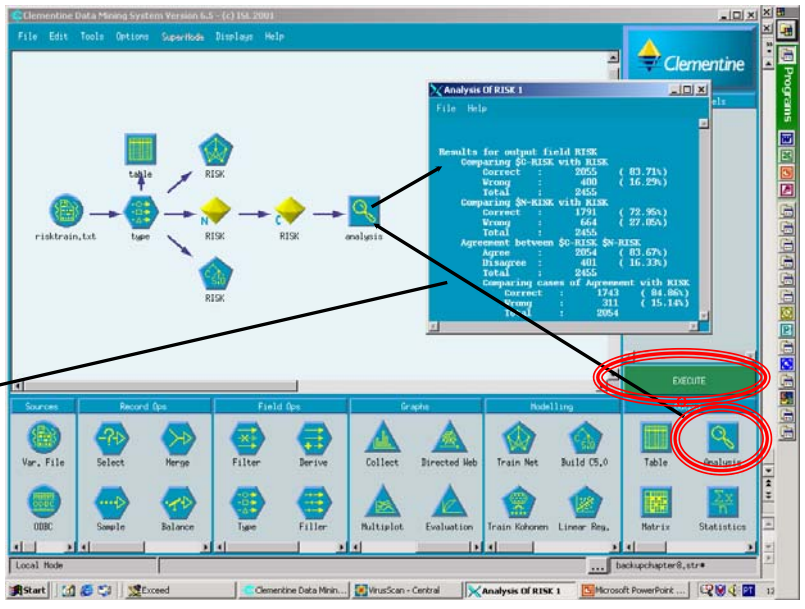
- *Introdução ao Nó Análise*
- *Usar Árvores de Decisão antes das Redes Neuronais*
- *Usar Redes Neuronais ante das Árvores de Decisão*



Comparação de Modelos

Para podermos comparar modelos, <u>eles devem estar no mesmo stream e ligados à mesma fonte de dados.</u>

Nó Análise: avalia o desempenho em termos de correcção de um modelo face aos valores reais.



1.ª secção compara o valore real do risco e o predito com o modelo da árvores de decisão

2.ª secção compara o valore real do risco e o predito com o modelo da rede neuronal

3.ª secção compara o nível de acordo dos dois modelos.

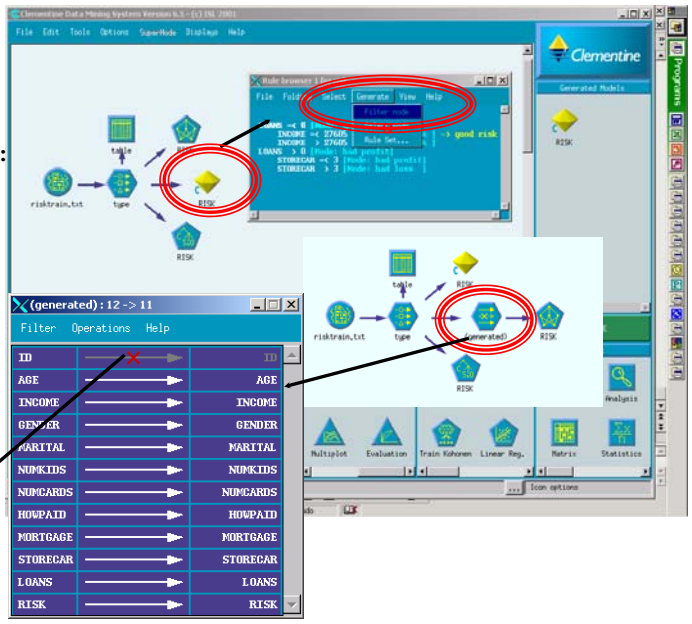
Combinação de Métodos

Árvores de Decisão antes de Redes Neuronais

Problema:

- **Redes neuronais (sem opção prune) utilizam todos os campos de entrada:**
- **Maior tempo de treino**
- **Campos sem impacto na saída, continuam no modelo**

Solução: utilizar árvores de decisão como pré-processamento e reduzir o número de campos à entrada da rede neuronal.



Só o campo ID foi eliminado (já o era anteriormente).

Neste caso a rede neuronal anterior já estava a trabalhar com os campos relevantes.

Instituto Superior Politécnico de VISEU
Escola Superior de Tecnologia
Departamento de Informática

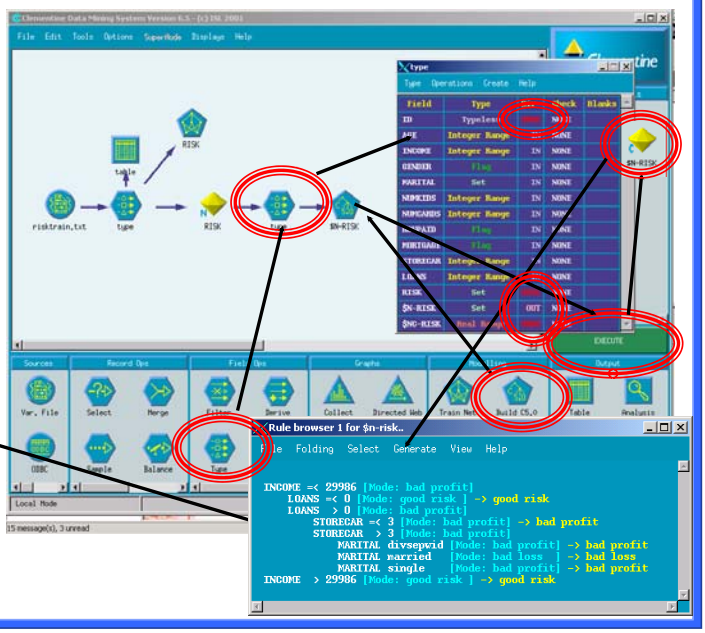
Combinação de Métodos

Árvores de Decisão depois de Redes Neuronais

Problema:

- **Redes neuronais de difícil interpretação dos resultados.**

Solução: utilizar árvores de decisão como pós-processamento e utilizar as suas eminentes capacidades descritivas.



Parece que a RN prediz como bom risco de crédito os clientes que tenham um rendimento acima de 29,986 ou abaixo disto sem empréstimos.

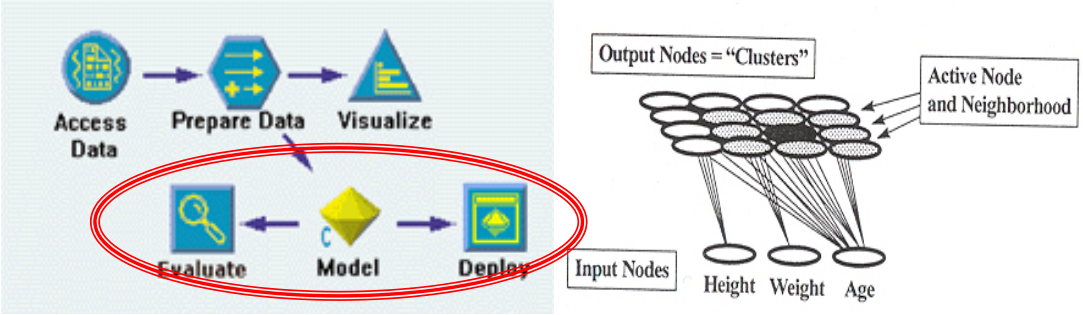
Aqueles abaixo de 29,986 e com empréstimos, serão maus risco de crédito.

Quanto aos maus mas lucrativos ou maus, serão aqueles registos com um grande número de cartões de armazém e de acordo com o seu estado civil.

AID (Clementine – 3.ª Parte)

Redes Kohonen

- *Introdução ao Nó Train Kohonen*
- *Construir uma rede Kohonen*
- *Interpretar os resultados*



Introdução às redes Kohonen

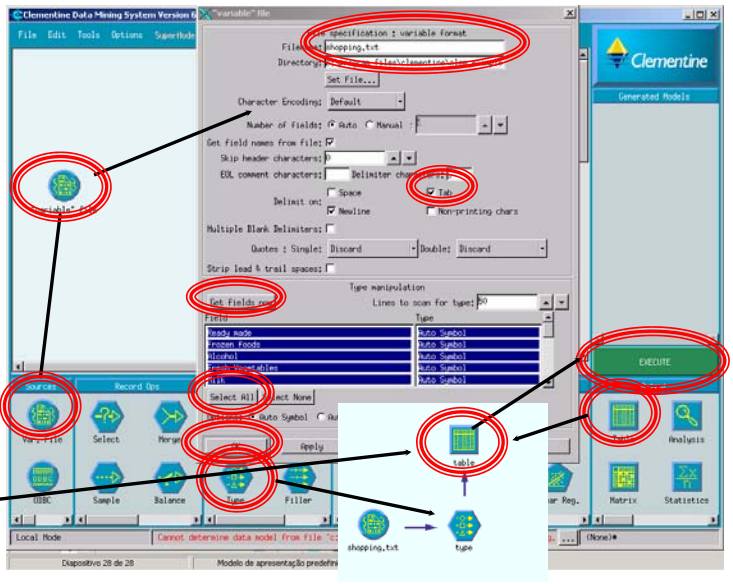
- **Efectuam aprendizagem não supervisionada**
- **Não é dado um campo a predizer, ou seja o campo OUT não é especificado**
- **Tentam encontrar relacionamentos e estruturas genéricas nos dados**
- **A saída da rede é um conjunto de coordenadas (X,Y) que podem ser utilizadas para visualizar grupos de registos e que podem ser combinadas para criar um código dos "pertencentes ao membro"**
- **Espera-se que os "grupos de clusters" ou "segmentos" sejam distintos uns dos outros e que contenham registos semelhantes em alguns aspectos**

Nó Train Kohonen

É gerada uma rede Kohonen na paleta Modelling

Inspeccionando a rede gerada, são-nos dadas informações acerca da estrutura da rede criada

Outra informação pode ser obtida através da manipulação de dados ou utilizando técnicas gráficas.

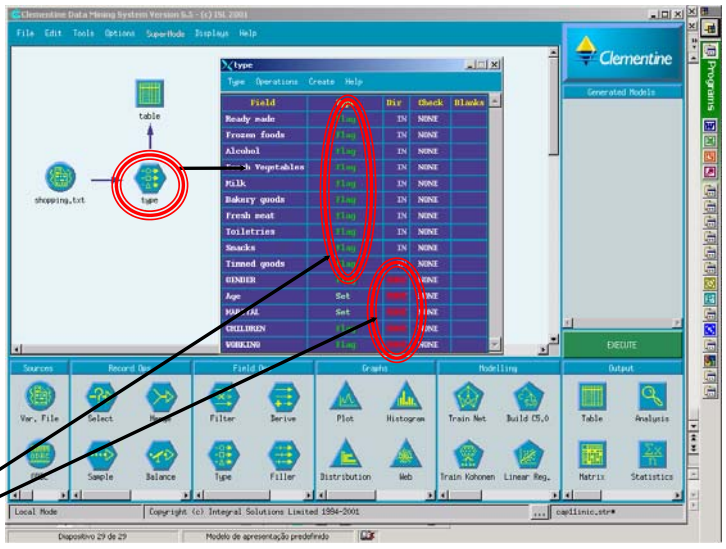


Colocar um nó type e table e forçar a execução, para que os campos sejam tipificados automaticamente.

Nó Train Kohonen

Pretendemos gerar uma rede Kohonen que aparecerá na paleta Modelling

- **Inspeccionando a rede gerada, são-nos dadas informações acerca da estrutura da rede criada**
- **Outra informação pode ser obtida através da manipulação de dados ou utilizando técnicas gráficas.**



Só estamos interessados na segmentação do comportamento de compra, assim os campos não directamente relacionados são colocados com direcção em >NONE para não serem utilizados pelo algoritmo.

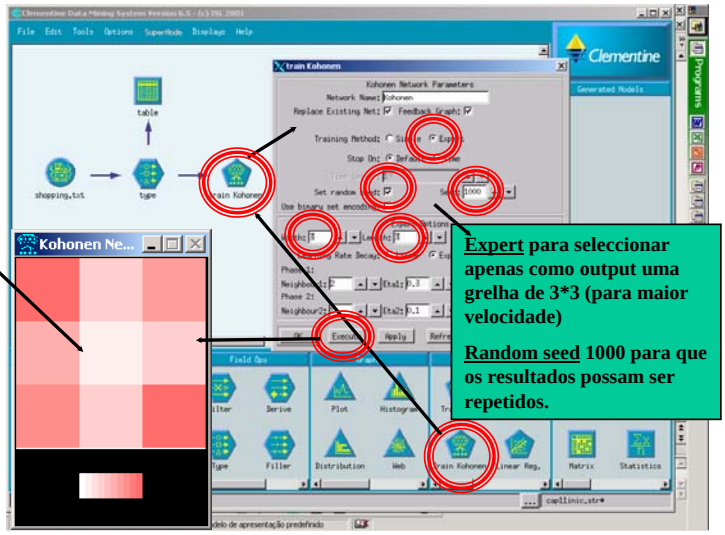
Note-se que todos os campos relacionados com as compras são do tipo Flag.

Nó Train Kohonen (treino)

À medida que a rede Kohonen é treinada surge um gráfico sob a forma de uma grelha. Cada célula da grelha representa um nó de saída

Cada vez que um nó de saída ganha um registo, torna-se mais escura

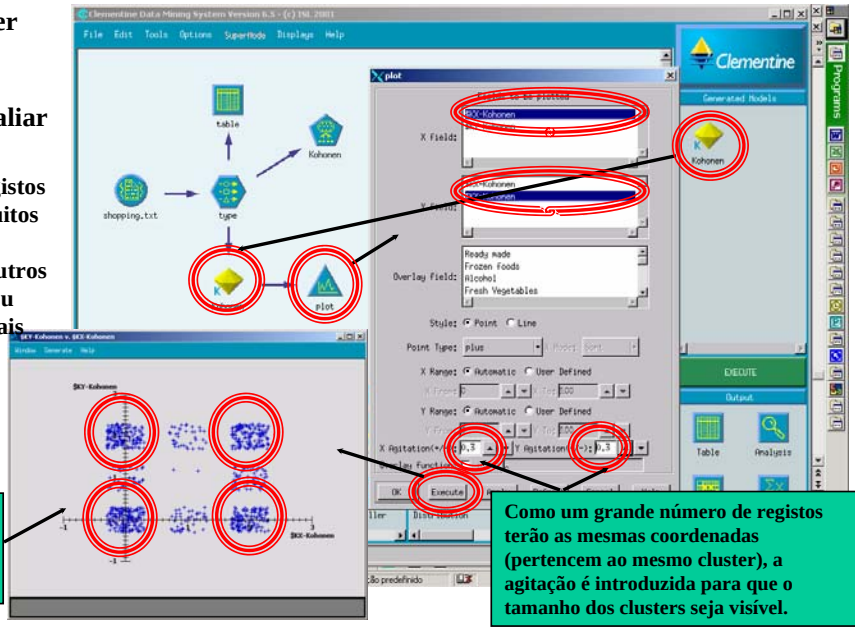
Quanto mais escura a célula, maior será o número de registos no cluster que representa.



Compreender a Rede Kohonen Gerada

O primeiro passo é ver quantos clusters distintos foram encontrados e avaliar da sua utilidade.

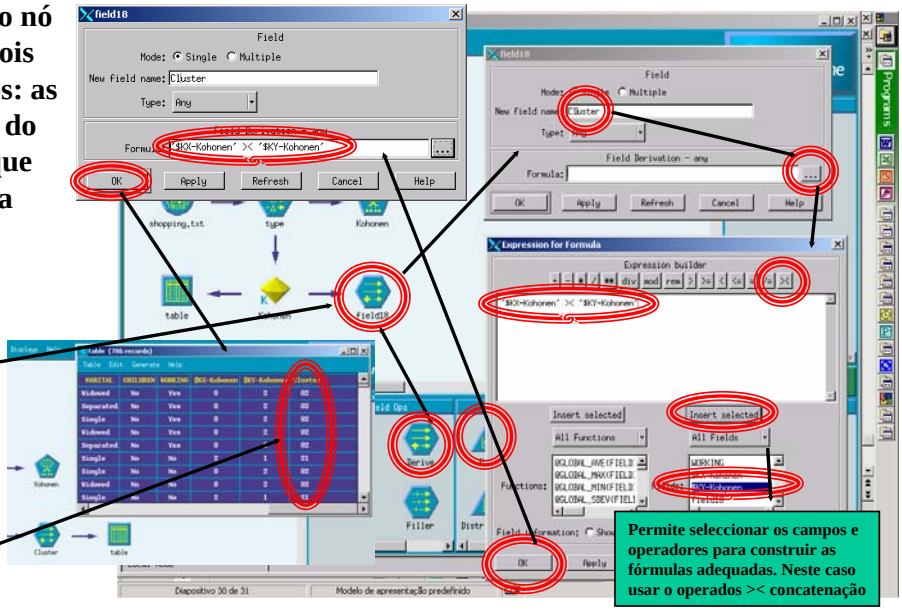
Clusters com poucos registos ou extremos, em muitos casos, não terão interesse, mas em outros (detecção de erros ou fraudes) serão os mais relevantes.



Criar um Código de Referência p/ cada Cluster

Já sabemos que o nó criado cria dois novos campos: as coordenadas do segmento a que pertence cada registos.

Vamos agora criar um campo com as coordenadas de cada registo, através da concatenação dos dois campos gerados.

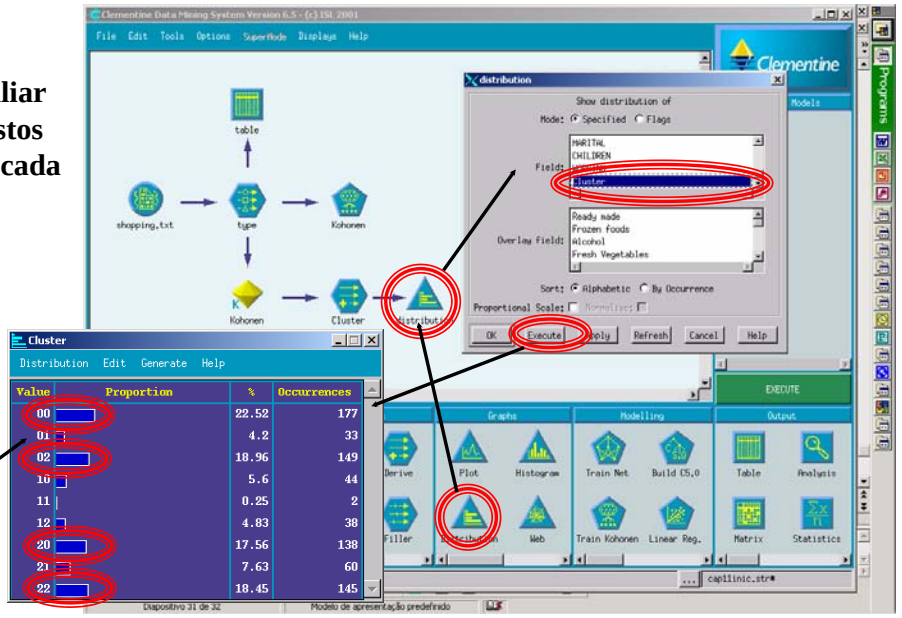


Colocando o nó Table, é mostrado um novo campo (Cluster), criado com o nó Derive

Permite seleccionar os campos e operadores para construir as fórmulas adequadas. Neste caso usar o operados >< concatenação

Focar nos Segmentos Principais

Usando um nó distribuição, podemos avaliar quantos registos pertencem a cada segmento.



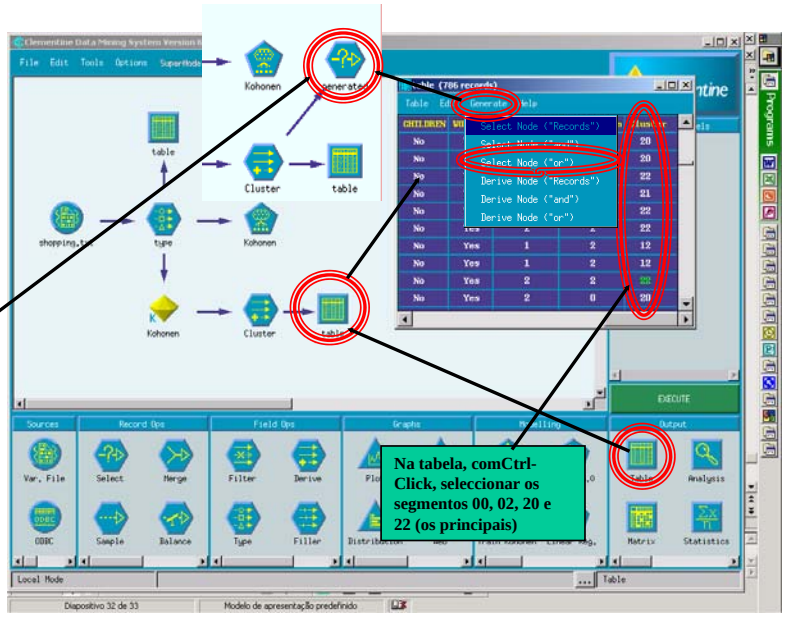
Os 4 grupos principais são o cluster 00, 02, 20 e 22, que cobre 78% dos registos.

No entanto, os grupos mais pequenos podem ter interesse (alvos para grupos de produtos especializados)

Focar nos Segmentos Principais

Estamos interessados nos segmentos principais: com a opção *Generate* da janela table, vamos gerar um nó *Select* para seleccionar os respectivos registos.

Nó que irá disponibilizar os registos que pertencem aos 4 clusters seleccionados



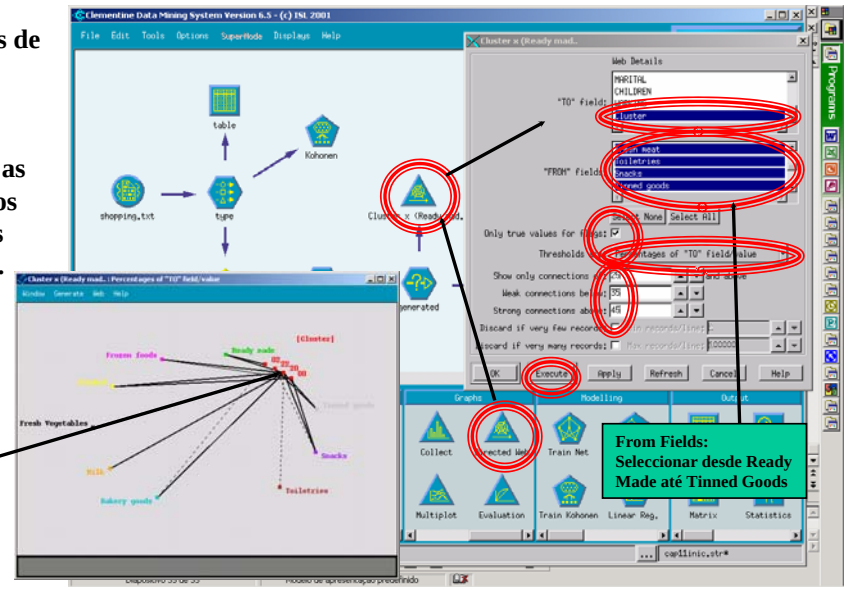
Explorar o Perfil de cada Segmento

Para vermos qual o perfil de compras de cada segmento, vamos usar o nó *Direct Web*, de forma a mostrar as ligações dos vários tipos de produtos com cada cluster.

O gráfico Web directp está um tanto confuso: os 4 clusters estão juntos, de forma que é difícil veras compras se relacionam com cada grupo.

Com o raro, arrastar cada cluster, de forma a ficarem mais separados

Ver figura do acetato seguinte.



Explorar o Perfil de cada Segmento

Cluster 00

Grupo associado a alcohol, sncks e frozen Foods

Cluster 02

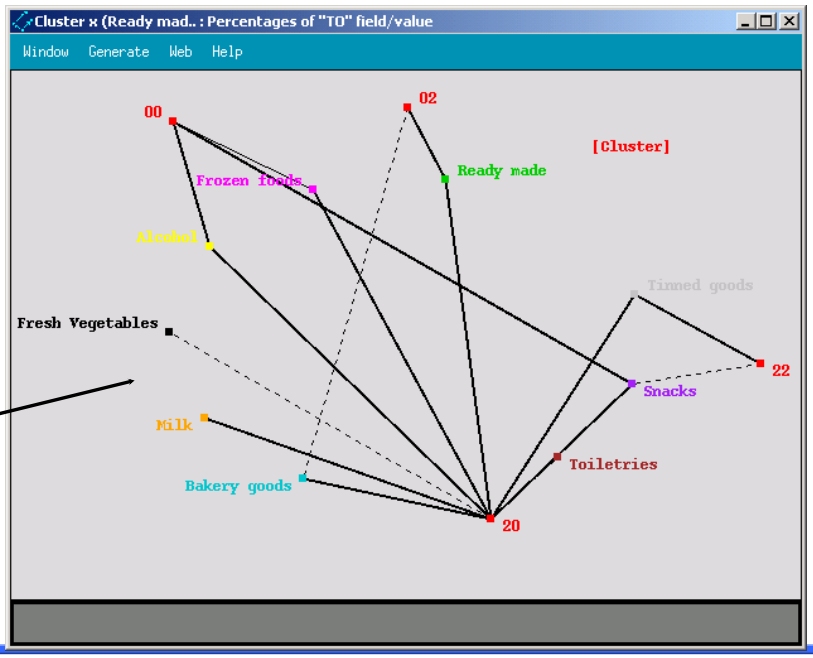
Grupo associado a ready-made

Cluster 20

Grupo associado à maioria dosdiferentes grupos (tinned googs, snacks, bakery goods, alcohol, frozen foods, milk, ready-made)

Cluster 22

Grupo fortemente associado a tinned goods



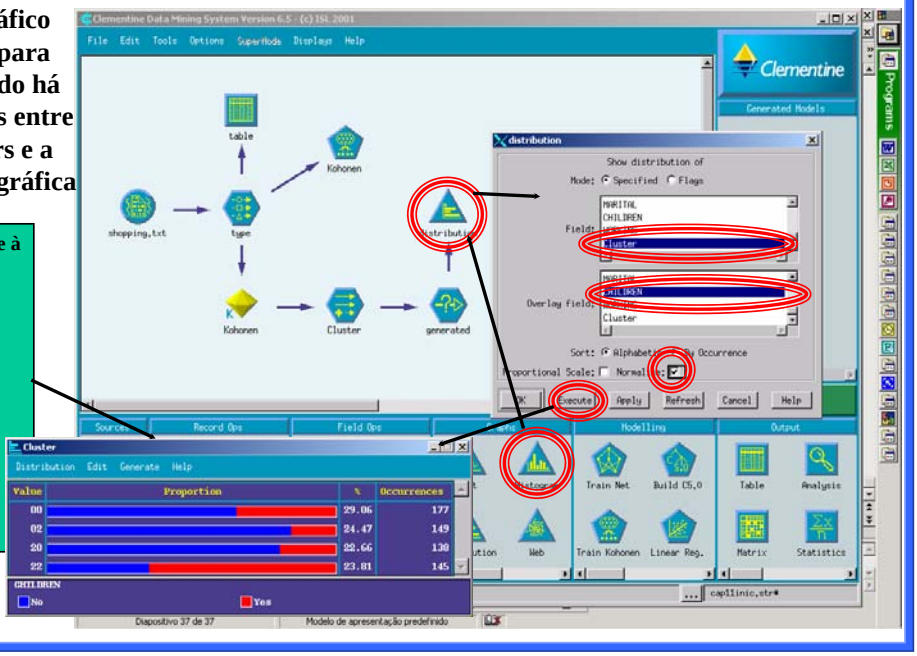
Mais Campos p/ Explorar o Perfil de cada Segmento

Para utilizar um gráfico de distribuição para investigar quando há relacionamentos entre os vários clusters e a informação geográfica existente.

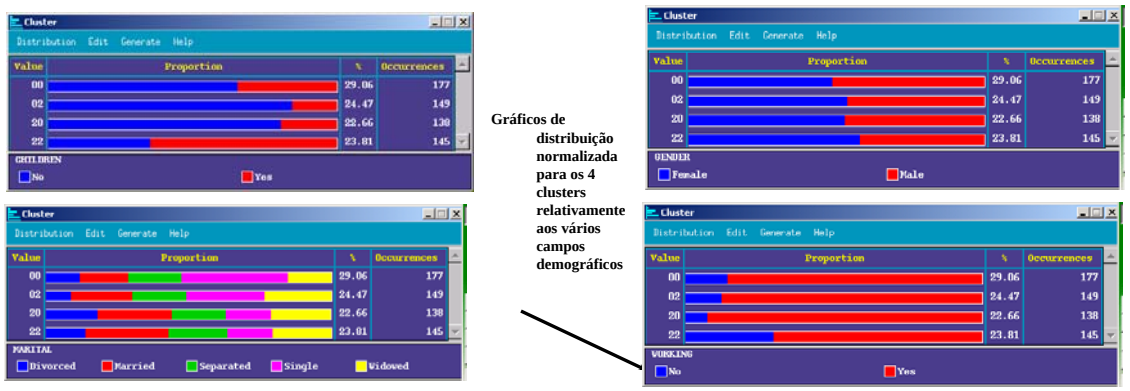
Gráfico dos clusters face à existência de crianças.

O cluster 02 3 20 é composto de indivíduos maioritariamente sem crianças e o 22 de indivíduos maioritariamente com crianças.

Podem criar-se gráficos de distribuição para outros campos



Mais Campos p/ Explorar o Perfil de cada Segmento



Gráficos de distribuição normalizada para os 4 clusters relativamente aos vários campos demográficos

Cluster 00: associado com snacks, alcohol e frozen foods. A maioria dos indivíduos com 40 ou menos anos. A maior proporção de pessoas solteiras e também pessoas a não trabalharem superior à média.

Cluster 02: Compram comidas congeladas já feitas, de idade baixa (grande proporção abaixo dos 30) sem crianças e que trabalham.

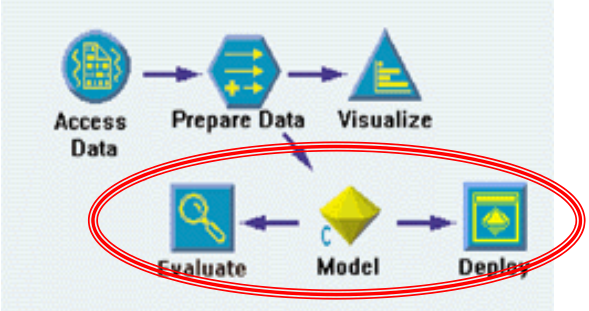
Cluster 20: Convida-se o aluno a tirar as conclusões

Cluster 22: Idem ao cluster 20

AID (Clementine – 3.ª Parte)

Regras de Associação

- *Introdução aos dois métodos de gerar regras de associação*
- *Utilizar o nó APRIORI para construir um conjunto de regras de associação*
- *Interpretar os resultados*



Regra:
"90% das mulheres possuidoras de carros de desporto vermelhos e cães pequenos, usam Chanel Nº5"

Regras de Associação: Introdução

Quando as pessoas compram cigarros, tendem a comprar chocolate ou cerveja?
As pessoas que têm um alto nível de colesterol, tendem a ter pressão sanguínea elevada?
Uma pessoa que contrata um seguro de automóvel, contratará também o seguro para a casa?

As respostas as estas questões serão a base para o posicionamento de um linha de produtos, marketing directo e publicidade.

Como o fazer?

- **Os algoritmos de detecção de associações dão-nos regras mostrando que valores em campos ocorrem tipicamente em conjunto.**

Formato da regra:

Conclusão <= Condição1 & Condição2 & ... & Condição N

Por cada regra é dada a cobertura e precisão

- **Cobertura: proporção de registos no data set que verificam as condições**
- **Precisão (ou confiança): proporção de registos que verificam a conclusão relativamente a todos os registos que verificam as condições**

Regras de Associação: Introdução

Disponíveis no Clementine:

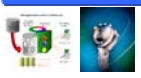
- **Generalized Rule Induction (GRI) – procura as regras independentes mais interessantes, utilizando uma medida de interesse denominada [medida J] e pode tratar campos numéricos como entrada.**
- **APRIORI –abordagem ligeiramente mais eficiente, mas só aceita campos simbólicos. Pode utilizar várias medidas de critério para guiar a geração de regras.**

Ambos os procedimentos produzem modelos não refinados, que podem ser inspeccionados para visualizar o conjunto de regras.

Contudo os modelos criados não podem ser colocado no painel stream ou ser atravessados por dados fazer parte dum stream).

Exemplo de regra: Jornais <= Combustíveis & Chocolate (2051:15%, 0.71)

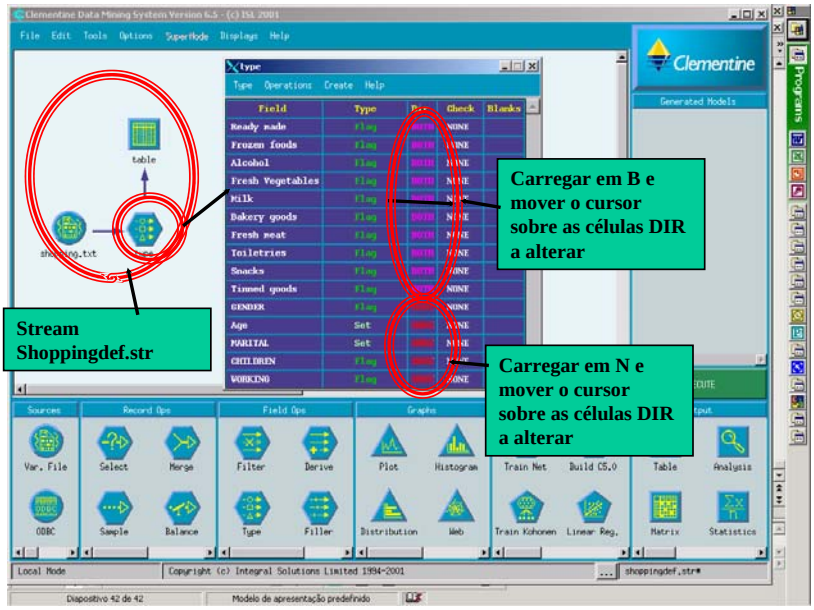
15% dos clientes (2051 indivíduos) compraram combustível e chocolate. Destes 2051, 71% também compraram jornais



NÓ APRIORI

Dado que os nossos campos de dados são simbólicos, vamos utilizar o algoritmo de detecção de regras de associação APRIORI

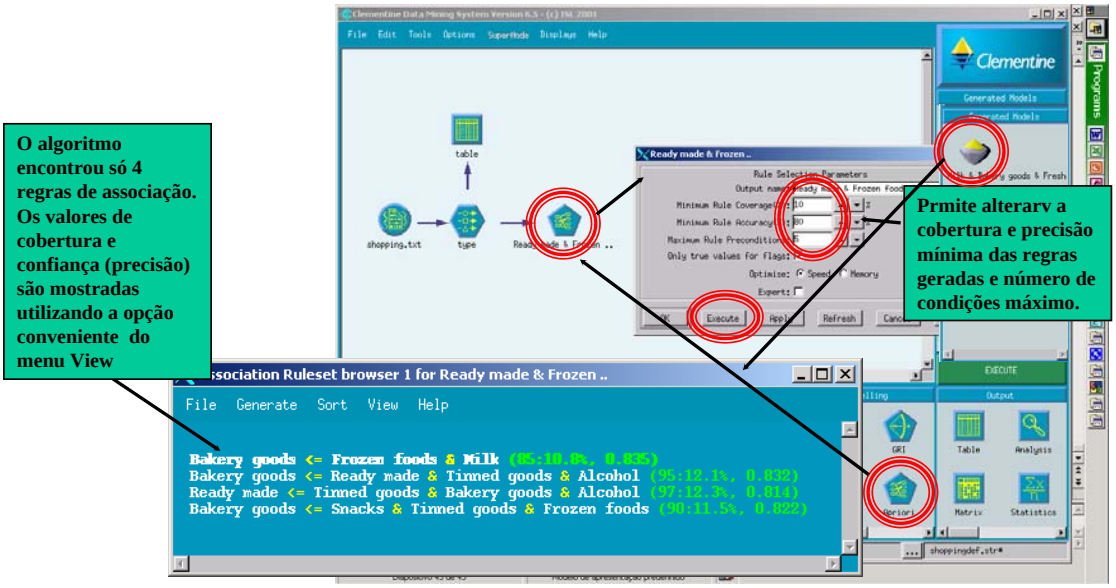
Depois do processo de geração das regras, é gerado um modelo na paleta de Modelos Gerados, que pode ser inspeccionado, mas não colocadas no painel stream ou fazer passar dados por ele.



NÓ APRIORI

O algoritmo encontrou só 4 regras de associação. Os valores de cobertura e confiança (precisão) são mostradas utilizando a opção conveniente do menu View

Prmite alterarv a cobertura e precisão mínima das regras geradas e número de condições máximo.

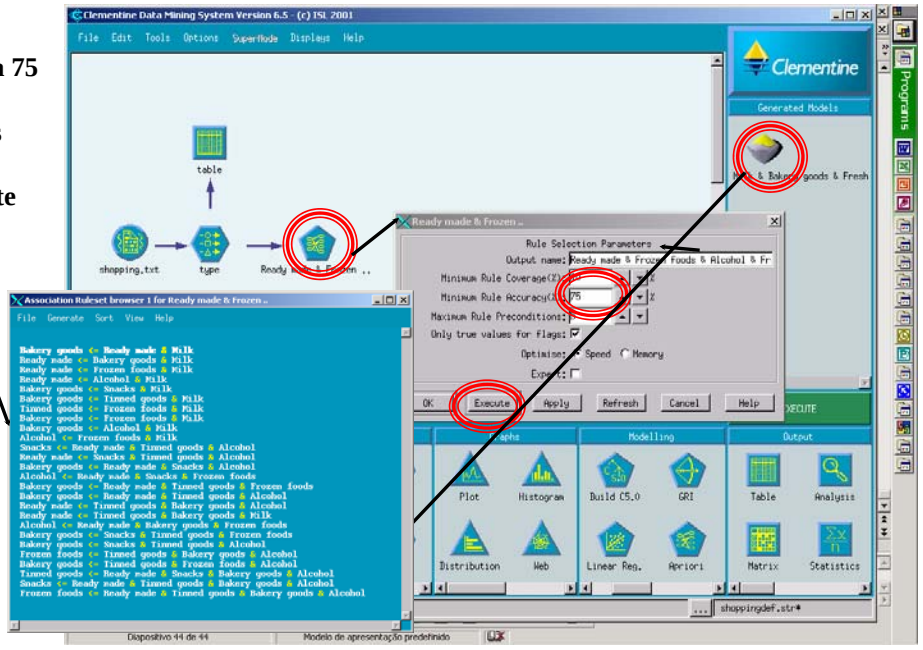


NÓ APRIORI (mais regras)

Vamos baixar a precisão para 75 % para que outras regras possam eventualmente passar e ser mostradas.

O algoritmo já encontrou um conjunto de regras bastante mais rico.

Agora o desafio é saber quais destas regras poderão ser úteis no contexto do negócio ou investigação.



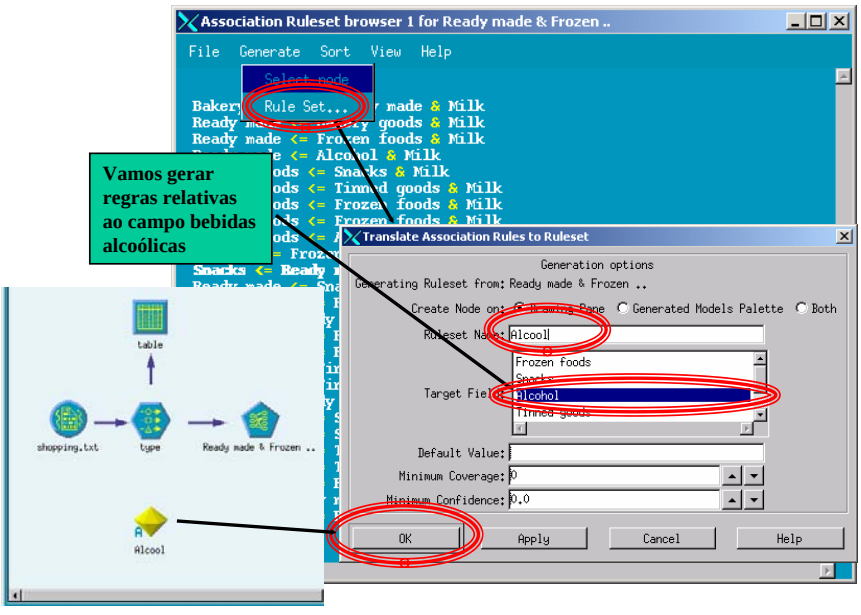
Utilizar as Associações (APRIORI ou GRI)

Uma limitação é que os nós-modelos criados não são passíveis de operar sobre dados (daí a pepita/diamante criado estar meia em bruto);

Mas...

Podemos criar um conjunto de regras do tipo das regras de indução, através do menu Generate.

Vamos gerar regras relativas ao campo bebidas alcoólicas

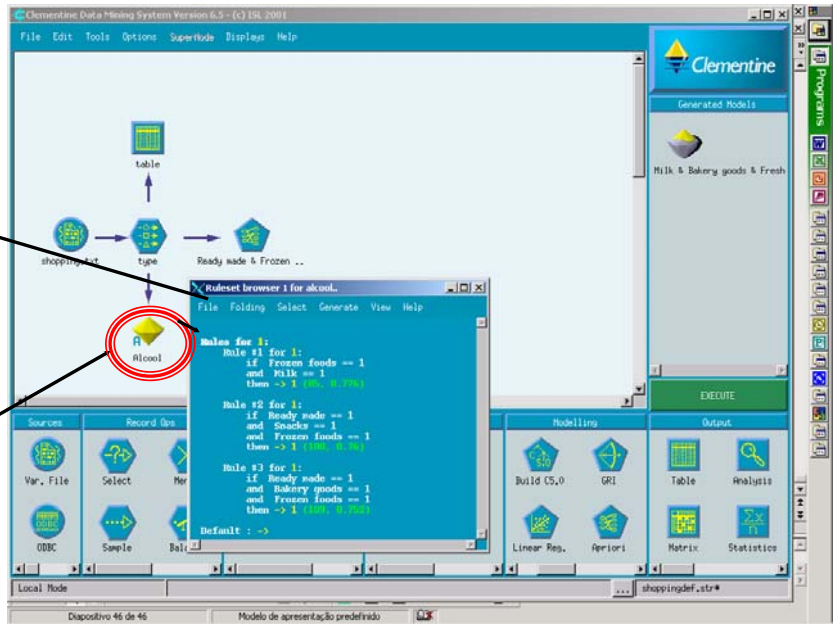


Utilizar as Associações (APRIORI ou GRI)

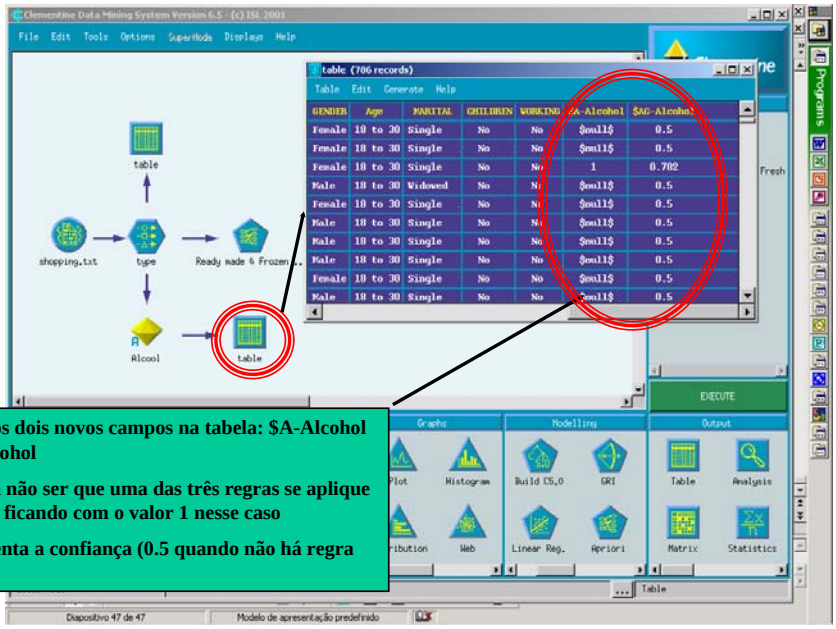
O conjunto de regras contém 3 regras cuja conclusão é a aquisição de produtos alcoólicos.

A primeira associa comida congelada e leite

Modelo gerado para as regras relativas a bebidas alcoólicas.



Utilizar as Associações (APRIORI ou GRI)



São criados dois novos campos na tabela: $A-Alcohol e $AC-Alcohol

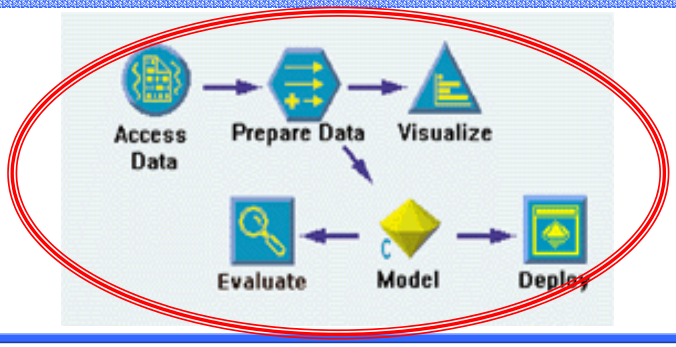
1.º $null, a não ser que uma das três regras se aplique ao registo, ficando com o valor 1 nesse caso

2.º representa a confiança (0.5 quando não há regra aplicável)

AID (Clementine – 3.ª Parte)

Estratégia para o Data Mining e Desenvolvimento de Modelos

- *Discutir estratégias para Data Mining*
- *Sugerir métodos para melhoria dos modelos*
- *Opções para aplicação de modelos em novos dados*



Estratégia p/ Data Mining

O Data Mining é um processo Iterativo, podendo-se começar pela manipulação, visualização e a criação de modelos, mas:

é muito possível que se tenha de voltar a estágios anteriores.

As operações a executar apresentam-se abaixo:

- **Pré-processamento dos dados utilizando técnicas visuais e manipulação de dados. Isto pode envolver a purificação dos dados e, se necessário, transformá-los para uma forma apropriada ao Data Mining;**
- **Procurar padrões e relacionamentos utilizando estatísticas e análises gráficas; estas técnicas poderão também revelar-se úteis para a identificação de um conjunto de atributos promissores a utilizar em técnicas de modelação.**
- **Dividir os dados em conjuntos de treino e teste para permitir testar os modelos gerados**
- **Criar os modelos aplicando as facilidades de modelação e estatísticas ao conjunto de treino.**
- **Testar os modelos resultantes no conjunto de teste para verificar da sua precisão.**
- **Analisar os resultados e refinar o modelo, se necessário.**

CRISP-DM

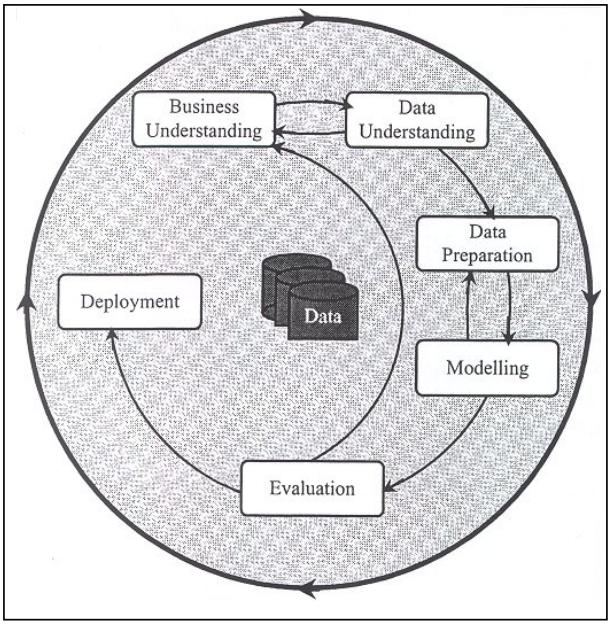
Metodologia de processos para Data Mining, não proprietária e desenvolvida através da colaboração entre várias organizações.

CRISP-DM é o acrónimo de Cross-Industry Standard Process for Data Mining

CRISP-DM pretende decompor um projecto de DM em fases e tarefas, onde cada tarefa deve proporcionar um produto ou componente tangível.

- **Inclui questões de negócio;**
- **Reconhece que o processo é não-linear e iterativo;**
- **A ideia é que todos sigam um processo standard e utilizem termos comuns de forma a que os projectos de DM sejam mais facilmente replicados.**

CRISP-DM



Processo CRISP-DM: fases principais num processo de Data Mining bem sucedido: Compreensão do negócio, compreensão dos dados, preparação dos dados, modelação, avaliação e utilização.

O circulo exterior representa o facto de que todo o processo é iterativo

A ordem das tarefas normal é no sentido dos ponteiros do relógio

As setas em sentido contrário, indicam sequências de tarefas executadas eventualmente de forma repetida.

Sugestões p/ Melhorar o Desempenho do Modelo (1)

1. Teste e Treino - Dividir a amostra em duas partes:

- **1.ª para treinar o modelo**
- **2.ª para avaliar do desempenho do modelo criado (prevenir o problema da sobre-adaptação e tornar o modelo mais genérico)**

2. Balancear os dados

- **Se os dados contém um número desproporcionado de casos em que um dos valores do campo de saída excede largamente o outros, isso levará a grandes dificuldades em criar modelos precisos de predição do valor que ocorre muito pouco. O Clementine tem um nó balance que pode ser utilizado para tratar este problema.**

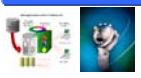
3. Transformar os dados

- **As técnicas de machine learning (redes neuronais em particular) têm melhor desempenho quando os campos numéricos têm uma distribuição uniforme. Uma distribuição não uniforme de valores de saída pode ser transformada com um nó Derive para produzir uma distribuição mais uniforme.**

Sugestões p/ Melhorar o Desempenho do Modelo (2)

4. Combinar métodos de modelação

- **Tal como foi mostrado atrás, as redes neuronais e árvores de decisão podem ser combinadas para ajudar a refinar os modelos. As 2.ªs podem ajudar na identificação dos campos de entrada mais úteis à rede neuronal; também as predições de um modelo podem ser utilizadas como entrada para um outro modelo, o que pode melhorar a precisão preditiva.**



Utilização dos Modelos

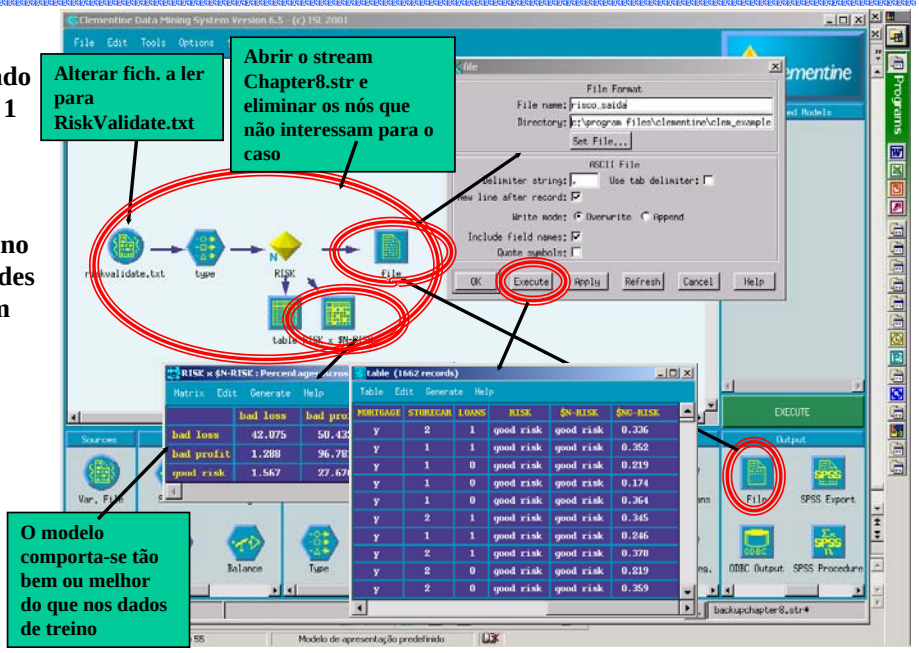
Depois do modelo gerado, há que aplicá-lo a novos dados:

1. **Modificar o stream usado para modelar os dados e adicionar ao nó do modelo gerado um nó de output (escrever para um ficheiro ou base de dados via ODBC). Editar o nó fonte por forma a que aponte para um novo ficheiro e executar o stream, o que fará com que o campo de saída seja predito para os novos registos, depois enviados para o ficheiro ou base de dados de saída. Este método será utilizável na máquina que contenha o Clementine, não conveniente para processamento online ou genérico para a empresa.**
2. **O stream que contém o modelo gerado pode ser publicado (se dispuser do produto Clementine Solution Publisher). Os ficheiros publicados (um fich. imagem e um de parâmetros)podem ser utilizados como entrada no motor Runtime do Clementine que executa streams publicados. Modificando o ficheiro de parâmetros, o stream publicado pode ser direccionado para uma nova fonte de dados. Adicionalmente, também é permitido que outros programas possam controlar a execução do modelo e assim serem incorporados em sistemas que necessitem dos seus resultados.**
3. **O modelo gerado pode ser exportado como código C (ou, em certos modelos, código XML). Isto permite embutir directamente o código respectivo nas aplicações.**

Utilização dos Modelos dentro do stream Clementine (1)

Vamos utilizar o modelo segundo a abordagem 1 do acetato anterior.

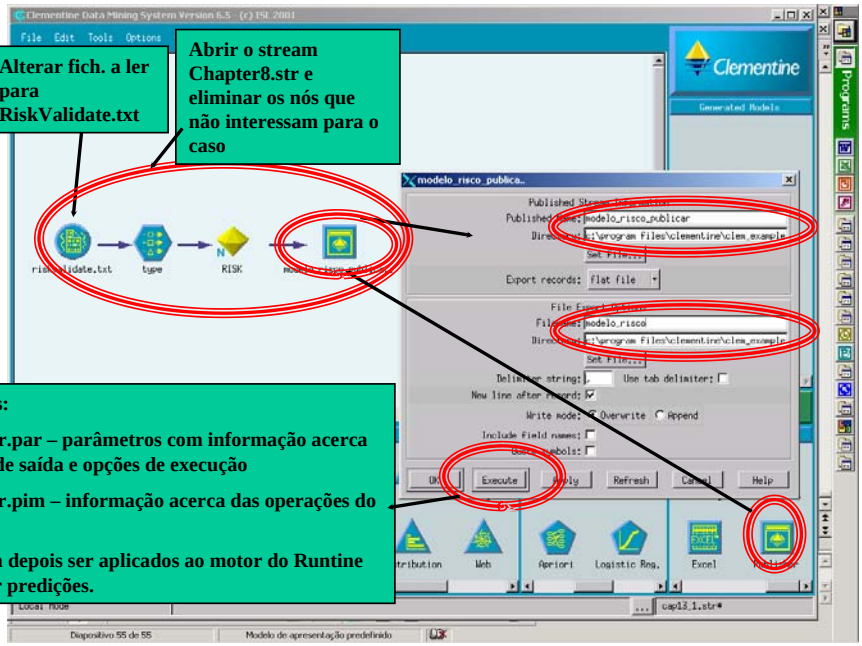
Utilizar o modelo criado atrás (no estudo das redes neuronais) em novos dados.



Utilização dos Modelos dentro do stream Clementine (2)

Vamos utilizar o modelo segundo a abordagem 2 do acetato anterior.

Utilizar o modelo criado atrás (no estudo das redes neuronais) em novos dados.



São criados 2 ficheiros:

modelo_risco_publicar.par – parâmetros com informação acerca fontes de dados, fich. de saída e opções de execução

modelo_risco_publicar.pim – informação acerca das operações do stream

Estes ficheiros podiam depois ser aplicados ao motor do Runtine Clementine para fazer predições.

Algumas Referências Adicionais

1. **Berry, Michael J. And Linoff. (1997) *Data Mining Techniques: For Marketing, Sales and Customer Support*. New York: Wiley.**
2. **Berry, Michael J. And Linoff. (2000) *Masteriong Data Mining: The Art and Science of Customer Relationship Management*. New York; Wiley.**
3. **Berson, Alex and S.J. Smith. (1997) *Data Warehousing, Data Mining & OLAP*. New York. McGraw Hill.**
4. **Fayyad, Usama M., Piatetsky-Shariro, G., Smyth, P. And R. Uthurusamy. (1996) *Advances in Knowledge Discovery and Data Mining*, Menlo Park, CA.; Cambridge, Mass: AAAI Press and MIT Press.**
5. **Han, Jiawei and M. Kamber. (2000) *Data Mining: Concepts and Techniques*. San Francisco: Morgan Kaufman.**
6. **SPSS Inc. (1999) *Data Mining with Confidence*. Chicago: SPSS Inc.**
7. **Westphal, Christopher and T. Blaxon. (1998) *Data Mining Solutions*. New York: Wiley.**
8. **Versão 1.0 do CRISP-DM (Cross-Industry Standard Process for Data Mining). Documento pode ser acedido em www.crisp-dm.org**
9. **Informação bastante exaustiva sobre Data Mining e Estatística em http://www.statsoftinc.com/textbook/stathome.html**