Curso de Engenharia de Sistemas e Informática - 5º Ano

Análise Inteligente de Dados

Ficha de Trabalho N.º 10

Objectivo: Avaliar da Integridade dos dados e proceder à respectiva manipulação; familiarizar-se com as ferramentas do Clementine que permitem procurar relacionamentos nos dados.

I - Integridade dos Dados.

Vamos utilizar o Stream criado na ficha de trabalho anterior.

1. Carregue o Stream criado na ficha de trabalho anterior, ExercTp9, utilizando a opção do menu File...Load Stream.
2. Coloque um nó Type da paleta Field Ops no painel Stream e coloque-o entre os nós Source e Table. Execute o stream por forma a que o Clementine actualize automaticamente os tipos dos campos.
3. Edite o nó Type e verifique se está de acordo com os tipos que o Clementine especificou. Altere os tipos, se necessário. Investigue as definições para blanks para alguns dos campos.
4. Coloque um Nó Quality ao Nó Type e execute esta secção do Stream. Os campos são válidos? Há que ter algumas preocupações com os dados?
5. Seleccione um dos campos simbólicos (do tipo set ou flag) e examine a sua distribuição utilizando o Nó Distribuição.
6. Seleccione um dos campos numéricos (inteiros ou reais) e examine a sua distribuição através do Nó Histograma.
7. Utilizando o Nó Estatística, examine as propriedades estatísticas relacionados com os gastos e visitas pré e pós campanha.
8. Guarde uma cópia actualizada do seu trabalho.

II - Manipulação dos Dados

Vamos utilizar o Stream criado no ponto anterior e executar alguma manipulação nos campos dos dados.

1. Carregue o stream guardado no ponto anterior.
2. Crie um histograma relativo ao campo Total Spend.
3. Vamos gerar de forma automática um **Nó Derive** que crie um novo campo que contenha quatro intervalos relativos ao campo Total Spend. Utilize o botão esquerdo do rato para colocar 3 linhas no histograma onde pretender dividir os dados. Crie o Nó Derive, utilizando o menu **Generate**.
4. Ligue o **Nó Derive** criado ao nó Type. Edite o **Nó Derive** e altere o nome do novo campo para “Intervalos de Gastos Totais”. Acrescente um **Nó Table** ao **Nó Derive** e execute essa parte do stream. Visualize o campo criado utilizando um nó Table.
5. Guarde o stream agora criado com o mesmo nome.

6. Adicionalmente tente criar manualmente um **Nó Select** que seleccione os registos das mulheres com 50 ou mais anos (lembre-se que o CLEM é sensível a letras maiúsculas ou minúsculas).
7. Coloque um **Nó Table** depois do **Nó Select** acabado de criar verifique da correcção da operação do **Nó Select**.

III - Procura de relacionamentos nos dados.

Vamos utilizar o Stream criado no ponto anterior e efectuar uma investigação preliminar dos relacionamentos simples existentes nos dados. Nas fichas futuras tentaremos predizer o campo "resposta à campanha" utilizando diversos tipos de modelos, focando na procura de relacionamentos entre esse campo e outros.

1. Utilizar o stream do exercício anterior, ligue um **Nó Web** da paleta Gráficos ao **Nó Type**.
2. Edite o **Nó Web** por forma a que o gráfico produzido mostre os relacionamentos entre os campos seguintes:
   - Resposta à Campanha
   - Visitas e gastos em pré-campanhas (em categorias)
   - Grupos de idades
   - Sexo

   Devido ao número substancial de registos de dados, indicar para as conexões mostradas, conexões fracas e fortes, 200, 300 e 400 respectivamente. **Execute** o Nó.
3. Edite o Nó Web escondendo conexões irrelevantes. Quais são as três conexões mais fortes com o valor de resposta ..? Que grupos de idade estão maioritariamente associados com os não respondentes?
4. Investigue o relacionamento entre os campos numéricos de "Gastos de Pré-Campanha" e as "Visitas de Pré-Campanha" utilizando o **Nó Plot**. Parecerá haver um relacionamento entre esses dois campos?
5. Utilizando um histograma com um overlay, investigue se há um relacionamento entre os gastos de Pré-Campanha e o campo resposta à campanha. Tente normalizar o gráfico por forma a tornar o relacionamento mais claro. Parecerá haver um relacionamento entre esses dois campos? Se isso é verdade, isso é consistente as conclusões tiradas acerca do gráfico Web?
6. Guarde uma cópia do stream sob o nome Visual.str.