ENGENHARIA DO CONHECIMENTO

Curso de Engenharia Informática, Regime Diurno e Pós-Laboral

3° Ano, 1° Semestre, Ano Letivo 2014/2015

Enunciado do Projeto

1. Introdução

O projeto de *data mining* tem por objetivo consolidar os conhecimentos adquiridos na unidade curricular de Engenharia do Conhecimento. Para tal pretende-se que os estudantes demonstrem o domínio das diferentes etapas do processo de *data mining*, bem como a utilização de ferramentas computacionais que suportam o processo, na resolução de um problema real de *data mining*.

2. Descrição do Projeto

Para a realização do projeto foi selecionado um *data set* (conjunto de dados) relativo a um caso de estudo real na área das vendas online, devendo na elaboração/organização do trabalho de projeto ser seguida a metodologia CRISP-DM (*CRoss Industry Standard Process for Data Mining*), constituía pelas seguintes etapas:

- Compreensão do negócio
- Compreensão dos dados
- Preparação dos dados
- Modelização
- Avaliação
- Colocação em produção

A maioria das etapas do processo de *data mining* é suportada pelas ferramentas computacionais utilizadas no âmbito da unidade curricular (Microsoft Excel, RapidMiner e Weka), pelo que se pretende que em cada etapa sejam realizadas um conjunto de tarefas adequadas ao caso de estudo, de acordo com a metodologia CRISP-DM (Síntese na Tabela 1, em anexo). Para o melhor modelo obtido, ou seja, o modelo a colocar em produção, deverá descrever detalhadamente no relatório de projeto as tarefas e os resultados obtidos em cada etapa, bem como entregar os ficheiros gerados, sendo estes os resultados do projeto.

O projeto inclui também a realização de um estudo comparativo das técnicas de *data mining* utilizadas na resolução do problema. Nesse estudo deverá ser feita uma avaliação quantitativa (precisão), qualitativa (transparência) e uma análise de custo-benefício para os melhores modelos obtidos, colocando em evidência as suas vantagens/desvantagens.

3. Caso de Estudo

O caso de estudo escolhido para a realização deste projeto de *data mining* consiste num conjunto de dados de uma loja de vendas online. O *data set* apresenta dados de 451495 compras, sendo a compra de cada produto caracterizada pela data de compra, data de envio e vários atributos do produto e do cliente, num total de 14, incluindo a característica objetivo (se o produto foi ou não devolvido). Está também disponível um dicionário de dados com a descrição de cada uma das características (atributos).

Pretende-se como objetivo principal de *data mining* construir um modelo para prever se um produto adquirido na loja online será, ou não, ser devolvido pelo cliente. Atualmente estudos apontam que uma em cada três compras online é devolvida, porque os produtos não satisfazem as expetativas dos clientes, o que representa para as lojas custos significativos, já que geralmente estas assumem os custos de devolução.

Um bom modelo trará uma vantagem competitiva à loja de vendas online pois permitirá adotar medidas preventivas, baseadas na probabilidade do produto ser devolvido, como por exemplo restringir as opções de pagamento, ajustar os custos de envio ou a gama de tamanhos dos produtos disponíveis.

4. Aspetos de Funcionamento

O projeto deve ser realizado de preferência em grupos de dois estudantes e entregue, impreterivelmente, até à data definida na calendarização da avaliação contínua/periódica.

O relatório, em formato digital, e os ficheiros necessários à avaliação do projeto devem ser submetidos através da plataforma Moodle. Para que o projeto seja avaliado o relatório impresso terá também de ser entregue.

Os estudantes poderão esclarecer dúvidas e obter apoio na realização do projeto durante as aulas, no horário de gabinete dos docentes, na plataforma Moodle, ou por e-mail, através dos seguintes endereços beatriz.piedade@ipleiria.pt e jose.ramos@ipleiria.pt.

5. Avaliação

Os parâmetros globais de avaliação do projeto, e respetivos pesos, são os seguintes:

Parâmetros de avaliação	**Peso**
Compreensão do Negócio	5%
Compreensão dos Dados	10%
Preparação dos Dados	40%
Modelização	20%
Avaliação	5%
Estudo Comparativo	10%
Relatório do Projeto	10%

A defesa do projeto é individual, por amostragem, podendo alterar a classificação. Por exemplo, se um estudante obtiver 15,0 valores no projeto e 95% na defesa, a classificação será 0,95 x 15,0 = 14,25 valores.

Anexo: Metodologia CRISP-DM

Compreensão do Negócio	Compreensão dos Dados	Preparação dos Dados	Modelização	Avaliação	Colocação em Produção
Determinar objetivos de negócio *Descrição do negócio;* *Objetivos;* *Critérios de Sucesso.*	**Recolha dos Dados Iniciais** *Documentar como foram obtidos os dados.*	**Seleção dos Dados** *Documentar os motivos de seleção/exclusão.*	**Escolha de Técnicas de Modelização** *Técnicas selecionadas;* *Assunções de modelização.*	**Avaliar os Resultados** *Avaliar os resultados tendo em conta os critérios de sucesso definidos;* *Modelos escolhidos.*	**Planeamento da colocação em produção** *Plano de colocação em produção.*
Avaliar o cenário *Inventário de recursos, requisitos, assunções e restrições;* *Riscos e contingências;* *Terminologia;* *Custos e Benefícios.*	**Descrição dos Dados** *Documentar a descrição dos dados iniciais.*	**Limpeza dos Dados** *Documentar as tarefas de limpeza efetuadas.*	**Planificação de Testes** *Separação dos dados em conjuntos de treino e teste.*	**Revisão do Processo** *Revisão do processo.*	**Planeamento da monitorização e manutenção** *Plano de monitorização e manutenção.*
Objetivos de *Data Mining* *Determinar objetivos de Data Mining;* *Critérios de sucesso de Data Mining.*	**Exploração dos Dados** *Documentar as tarefas de exploração dos dados iniciais.*	**Derivar Novos Dados** *Derivar atributos e gerar registos.*	**Construção do Modelo** *Valores de parametrização* *Modelos obtidos;* *Descrição dos modelos.*	**Determinar Ações Futuras** *Recomendação de implementação do modelo, ou redefinição do problema.*	**Elaboração do relatório final** *Relatório final;* *Apresentação final.*
Plano de Projeto *Plano e metodologia;* *Determinação inicial de ferramentas e técnicas.*	**Qualidade dos Dados** *Documentar a qualidade dos dados iniciais.*	**Integrar Dados** *Junção de dados adicionais.*	**Avaliar o Modelo** *Avaliação do modelo obtido;* *Afinação de parâmetros.*		**Revisão do projeto** *Documento com a experiência adquirida no projeto.*
		Formatar Dados *Formatar os novos dados.*			
		Criação do *Data set* *Descrição do Data set.*			

Tabela 1: Tarefas genéricas (bold) e principais resultados (itálico) para cada etapa da metodologia CRISP-DM (Adaptado de CRISP-DM 1.0, *Step-by-step Data Mining Guide*).

Nota: Dependendo da natureza do projeto de *data mining*, algumas das tarefas poderão não ser aplicadas.