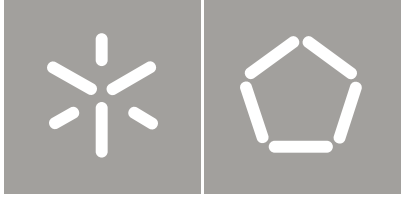
Universidade do Minho
Escola de Engenharia

Joel Frederico Azevedo Costa

Um Ambiente Gráfico para Facilitar Tarefas de Data Mining via Ferramenta R

Um Ambiente Gráfico para Facilitar Tarefas de Data Mining via Ferramenta R
Joel Frederico Azevedo Costa
UMinho | 2011

Outubro de 2011



Universidade do Minho
Escola de Engenharia

Joel Frederico Azevedo Costa

Um Ambiente Gráfico para Facilitar Tarefas de Data Mining via Ferramenta R

Tese de Mestrado
Tecnologias e Sistemas de Informação
Engenharia e Gestão de Sistemas de Informação

Trabalho efectuado sob a orientação do
Professor Doutor Paulo Alexandre Ribeiro Cortez

Outubro de 2011

"It has become appallingly obvious that our technology has exceeded our humanity."

Albert Einstein

Agradecimentos

Quero agradecer em primeiro lugar ao meu orientador, Professor Doutor Paulo Cortez, pelos seus ensinamentos e incentivos ao longo desta etapa da minha formação académica. Certamente, sem o seu apoio, não teria sido possível a realização deste trabalho.

Ao Eng. Joaquim Tinoco, que apesar de não o conhecer pessoalmente, foi incansável na altura em que eram necessários *feedbacks* exteriores em relação ao projeto.

À minha namorada, Patrícia Machado, por todo o apoio, incentivo, dedicação, e ajuda na elaboração desta dissertação. De certa forma existe trabalho seu neste projeto.

À Isa Coxixo, uma amiga e *designer*, que foi responsável pela elaboração do logótipo da aplicação.

Por último agradeço à minha família por todo o investimento em mim, e pelos incentivos constantes na minha formação.

A todos um muito obrigado!

Resumo

Com o rápido crescimento no uso das Tecnologias de Informação nas organizações, os dados organizacionais começaram a crescer a ritmo alucinante, tornando difícil a sua análise.. Este facto fez com que surgisse uma área denominada de *Data Mining*. Esta área utiliza técnicas provenientes da Inteligência Artificial, Estatística, Matemática e Bases de Dados, com o objetivo de extrair conhecimento útil a partir de dados em bruto.

Atualmente existe uma ferramenta *open-source,* chamada R, muito popular entre os analistas de *Data Mining*. Apesar da ferramenta não ser orientada especificamente para o *Data Mining*, pode ser adaptada para tal, através de instalação de *packages*. Em particular, o *package* rminer facilita a utilização de algoritmos de *Data Mining* de aprendizagem supervisionada, tal como Redes Neuronais Artificiais (RNAs) e Máquina de Vetores de Suporte (MVSs), em problemas de Classificação e Regressão. Contudo, o facto desta biblioteca funcionar via um conjunto de comandos, em modo de consola, exige uma certa curva de aprendizagem por parte dos utilizadores não especializados. Assim, neste trabalho é proposto um ambiente gráfico para o rminer, de modo a facilitar a sua adoção/uso por parte de utilizadores que não tenham um conhecimento especializado na linguagem R. Como resultado final, obteve-se um *Graphical User Interface (GUI)*, denominado de Jrminer, que mostra ser bem aceite pela comunidade do rminer, devido ao facto de ser simples, intuitiva e com um *design* coerente.

Palavras-Chave: *Data Mining,* R, rminer, Jrminer, Classificação, Regressão, Redes Neuronais Artificiais, Máquina de Vetores de Suporte

Abstract

With the fast growth of the use of Information Technologies in organizations, organizational data began to grow at very fast pace, making it impossible to analyse using classical statistical methods. This fact motivated the development and spread of the field of Data Mining. This area uses techniques from several disciplines, such as Artificial Intelligence, Statistics, Mathematics and Databases, in order to extract useful knowledge from raw data.

Currently, the R open-source tool is very popular among Data Mining analysts. Although, R is not specifically oriented for Data Mining, the tool mining capabilities can be increased through the installation of packages. In particular, the rminer package facilitates the use of supervised learning Data Mining algorithms, such as Neural Networks and Support Vector Machines, in classification and regression tasks. Yet, rminer works under a console command mode, thus requiring a certain learning curve by non-expert R users. Hence, in this work, we propose a Graphical User Interface (GUI) for the rminer library, in order to facilitate its use by non-specialized R users. As the final result of this work, we achieved a GUI, termed Jrminer, that is well accepted by the rminer community, mainly due to its simplicity, intuitively and coherent design.

Keywords: Data Mining, R, rminer, Jrminer, Classification, Regression, Artificial Neural Networks, Support Vector Machine

Conteúdo

Lista de Acrónimos

AD	Árvores de Decisão
BD	Bases de Dados
BI	*Business Intelligence*
CRISP-DM	*CRoss Industry Standard Process for Data Mining*
DCBD	Descoberta de Conhecimento em Base de Dados
DM	*Data Mining*
GUI	*Graphical User Interface*
LDA	*Linear Discriminant Analysis*
LR	*Logistic Regression*
MVSs	Máquinas de Vetor de Suporte
NB	*Naive Bayes*
QDA	*Quadratic Discriminant Analysis*
RF	*Random Forest*
RNAs	Redes Neuronais Artificiais
SO	Sistema Operativo

Lista de Figuras

Lista de Tabelas

1. Introdução

Este capítulo começa por fazer um enquadramento do tema, seguido da motivação, objetivos, e organização da dissertação.

1.1. Enquadramento

Hoje em dia, as organizações produzem um conjunto de dados muito elevado, e sabem que essas informações são vitais para a sua sobrevivência, uma vez que, podem trazer vantagens competitivas. Contudo, o seu processamento já não é possível com recurso a folhas de cálculo ou até mesmo a consultas *ad-hoc*, pois estas ferramentas apenas utilizam métodos estatísticos clássicos, limitando o tipo de análises que se podem efetuar [Goebel e Gruenwald, 1999]. Para ultrapassar essa dificuldade, surgiu a área da **Descoberta de Conhecimento em Base de Dados (DCBD)/*Data Mining (DM)***. Segundo [Fayyad et al., 1996], a DCBD é todo o processo que envolve a descoberta de conhecimento útil a partir de dados em bruto. Esse processo é constituído por diferentes etapas, das quais se destaca o DM, onde se aplicam técnicas oriundas de diversas disciplinas, tais como, Estatística, Bases de Dados, Matemática e Inteligência Artificial, com o objetivo de extrair e identificar o conhecimento útil para o suporte à tomada de decisão [Turban et al., 2010]. Embora [Fayyad et al., 1996] distinga formalmente os termos DCBD e DM, estes serão utilizados como sinónimos, uma vez que, com o passar dos anos, o termo DM passou a ser mais conhecido, sendo muitas das vezes utilizado para definir todo o processo de DCBD. De referir que existem diversos algoritmos de DM, tais como**, Árvores de Decisão (AD)**, **Redes Neuronais Artificiais (RNAs)**, **Máquina de Vetores de Suporte (MVSs)** e o algoritmo de segmentação *k-means clustering*. Estas técnicas podem ser aplicadas a diversos objetivos de DM, como a Classificação, a Regressão e a Segmentação.

Para melhorar o sucesso na execução de projetos de DM, foram desenvolvidas metodologias DM. Uma das mais populares tem a designação ***CRoss Industry Standard Process for Data Mining (CRISP-DM)***, tendo sido

concebida por um conjunto de empresas, como SPSS[1] e a DaimlerChryslyer [Chapman et al., 2000]. Esta metodologia compreende um conjunto de seis etapas e tem a vantagem de ser neutra, no que diz respeito à adoção de ferramentas DM.

Quanto às ferramentas de DM, dada a importância desta temática, existe no mercado um conjunto enorme.. No que diz respeito às ferramentas gratuitas, de *open-source,* destaca-se o ambiente R. Trata-se de um ambiente que adota uma programação orientada aos objetos e que foi desenvolvido como uma ferramenta estatística para análise de dados. Tem a vantagem de correr em múltiplas plataformas, tais como *Mac OS*, *Windows* e *Linux*. Além disso, esta aplicação pode ser facilmente estendida pela instalação de bibliotecas (*packages*). Atualmente, a ferramenta R tem disponíveis milhares de bibliotecas no repositório CRAN[2]. Entre este conjunto alargado de bibliotecas, destaca-se o rminer [Cortez, 2010].

1.2. Motivação

A biblioteca rminer[3] foi desenvolvida por [Cortez, 2010] com o objetivo de facilitar a utilização de algoritmos de aprendizagem supervisionada e de DM na ferramenta R. Em particular, esta biblioteca está talhada para utilizar RNAs e MVSs para a resolução de problemas Classificação e Regressão. Contudo, o facto do rminer funcionar via um conjunto de comandos, em modo de consola, exige uma certa curva de aprendizagem por parte dos utilizadores não especializados, sendo que por vezes existe mesmo uma "aversão" inicial à adoção desta ferramenta para quem está habituado a ambientes gráficos. Para resolver este problema, foi proposto no âmbito desta dissertação a construção de um ***Graphical User Interface (GUI)*** para a ferramenta rminer, de modo a facilitar o seu uso por utilizadores que não conheçam a linguagem R. Convém referir que no repositório

[1] http://www.spss.com/
[2] http://cran.r-project.org/
[3] http://www3.dsi.uminho.pt/pcortez/Home.html

CRAN[4] já existe uma GUI denominada de *Rattle*[5] [Williams, 2009] que permite trabalhar problemas de DM. Contudo, a sua instalação não é simples. Além disso, apresenta diferentes características quando comparado com a biblioteca rminer. Em particular, o *Rattle* tem diversas limitações no que diz respeito à utilização de RNAs e MVSs em tarefas de Classificação e Regressão [Chapman et al., 2000].

1.3. Objetivos

Face ao que foi exposto nas secções anteriores, a questão de investigação a abordar nesta dissertação é: "**como implementar um ambiente gráfico para a biblioteca rminer, de modo a facilitar o seu uso em tarefas de DM por parte de utilizadores não especializados?**".

Como a própria questão sugere, o objetivo deste trabalho é desenvolver um ambiente gráfico que permita reduzir a curva de aprendizagem da ferramenta R/rminer por parte dos utilizadores não especializados em R. Então será necessário investigar qual o melhor caminho a seguir, pois o ambiente gráfico terá que ter um conjunto de características para que seja uma mais-valia para o utilizador comum. Assim a elaboração deste trabalho terá que cumprir as seguintes metas:

1. Registar informação, através de questionário, sobre os utilizadores em relação ao *package* rminer;
2. Construir um ambiente gráfico tendo em conta as opiniões retiradas do questionário;
3. Enviar a aplicação para avaliação por parte dos utilizadores, e registar a sua opinião em questionário;
4. Demonstrar se os objetivos foram atingidos.

[4] http://cran.r-project.org/
[5] http://rattle.togaware.com/

1.4. Organização

Esta dissertação está organizada em seis capítulos:

- No Capítulo 1, **Introdução**, é feito um enquadramento do tema na atualidade. Também é apresentada a motivação, os objetivos, e a organização da dissertação;
- No Capítulo 2, ***Data Mining***, apresenta-se uma revisão de literatura sobre os conceitos básicos associados à temática do DM;
- No Capítulo 3, **Ferramentas *Data Mining***, apresenta-se as características que as aplicações DM devem possuir. Introduz-se o R e a biblioteca rminer. Além disso, também é apresentada uma lista das ferramentas mais usadas pela comunidade DM;
- No Capítulo 4, **Jrminer: um ambiente gráfico para o rminer**, são apresentados os resultados de um questionário sobre o rminer. Posteriormente é apresentado todo o processo de implementação do GUI, incluindo, apresentação das funcionalidades disponíveis e um exemplo de utilização. No final é possível observar se os objetivos foram cumpridos, com recurso a outro questionário;
- No Capítulo 5, **Conclusão**, é apresentada uma síntese da dissertação, uma discussão, e ideias para trabalho futuro.

2. *Data Mining*

Este capítulo apresenta o estado da arte do DM. Inicialmente é apresentada uma introdução ao tema, seguida das metodologias existentes. Neste capítulo também é possível encontrar, embora de forma sucinta, alguns objetivos DM e técnicas existentes no package rminer.

2.1. Introdução

As organizações, na conjuntura atual, estão em constantes mudanças, o que leva a existência de pressões cada vez maiores em relação ao modo como operam. Decisões seguras, rápidas e estratégicas são necessárias por parte dos gestores para permitir que esta seja competitiva no mercado onde atua. Por outro lado, nas últimas décadas, o custo de armazenamento da informação tem diminuído consideravelmente, levando as organizações a investir em tecnologias de base de dados. De facto, a produção constante de informação ultrapassou a capacidade de uma análise humana. Assim, essa abundância e fácil acessibilidade aos dados, faz com que nos dias de hoje área da DCBD/DM seja de considerável importância para as organizações [Maimon e Lior, 2010].

O DCBD foi definido por [Fayyad et al., 1996] como sendo um processo não trivial que permite a identificação de padrões úteis a partir de um conjunto de dados. Ainda segundo o mesmo autor, a descoberta de conhecimento é um processo iterativo e interativo. Iterativo na medida em que, pode ser necessário despender mais tempo numa etapa que inicialmente se pensava terminada. Interativo, pois o processo requer a participação do utilizador sempre que necessária a tomada de decisão.

Existe um conjunto alargado de objetivos de DM. A taxonomia do DM (ver Figura 1) permite-nos compreender a relação que existe entre estes objetivos e as técnicas/algoritmos de DM. Segundo [Maimon e Lior, 2010], existem fundamentalmente dois tipos de DM:

- **orientado à verificação** (o sistema verifica as hipóteses do utilizador); e
- **orientado à descoberta** (o sistema descobre novas regras e padrões automaticamente).

Quanto à orientação à descoberta, esta pode subdividir-se em:

- **previsão**, onde o objetivo é construir um modelo comportamental capaz de prever o valor de uma ou mais variáveis relacionadas com itens novos (para os quais o modelo não foi treinado); e
- **descrição**, onde o foco é a compreensão de como os dados subjacentes se relacionam com as suas partes.

A grande maioria das técnicas de DM orientadas à descoberta (quantitativa em particular) é baseada na aprendizagem indutiva (i.e., os modelos treinados são aplicados a exemplos futuros desconhecidos), onde os modelos são construídos, explicitamente ou implicitamente, pela generalização de um número suficiente de exemplos de treino [Maimon e Lior, 2010]. Por outro lado, os métodos de verificação permitem lidar com a avaliação de uma hipótese proposta por uma fonte externa (e.g., um utilizador). Esses métodos incluem as técnicas mais comuns da estatística tradicional, porém estão menos associados à área do DM, pois nesta área a maioria dos problemas estão preocupados na descoberta de novas hipóteses, descartando assim hipóteses já conhecidas.

Os autores [Maimon e Lior, 2010] referem que aprendizagem máquina[6] utiliza outra terminologia comum, isto é, os métodos de previsão estão direcionados para a **aprendizagem supervisionada**, que se refere a técnicas que tentam descobrir a relação entre os atributos de entrada (i.e., variáveis independentes) e o atributo de saída (i.e., variável dependente). A relação descoberta é representada em modelos, que normalmente permitem explicar fenómenos que estão ocultos num conjunto de dados, e também podem ser utilizados para prever uma variável de saída a partir dos valores dos atributos de entrada de um dado item. Estes métodos são aplicados num conjunto variado de domínios (e.g., *Marketing*, Finanças, entre

[6] Tradução adotada para o termo *Machine Learning*.

outros). Existem dois tipos de modelos supervisionados e que correspondem a dois objetivos de DM [Rocha et al., 2007]:

- **Classificação**. Criar uma associação entre os atributos de entrada a uma das classes pré-definidas (e.g., classificação de células para o diagnóstico de cancro, classificar clientes com hipotecas bancárias como "bons" ou "maus");
- **Regressão**. Mapeamento entre um vetor de entrada e um valor real (e.g., previsão do mercado de ações, número de vendas de um determinado produto dada as suas características).

Por outro lado, na aprendizagem máquina existe também a **aprendizagem não supervisionada**, que se refere a técnicas que agrupam instâncias, de acordo com medidas de similaridade ou distância. Assim a aprendizagem não supervisionada cobre apenas uma porção dos métodos descritos na Figura 1 (e.g., métodos de *segmentação,* mas não de visualização).

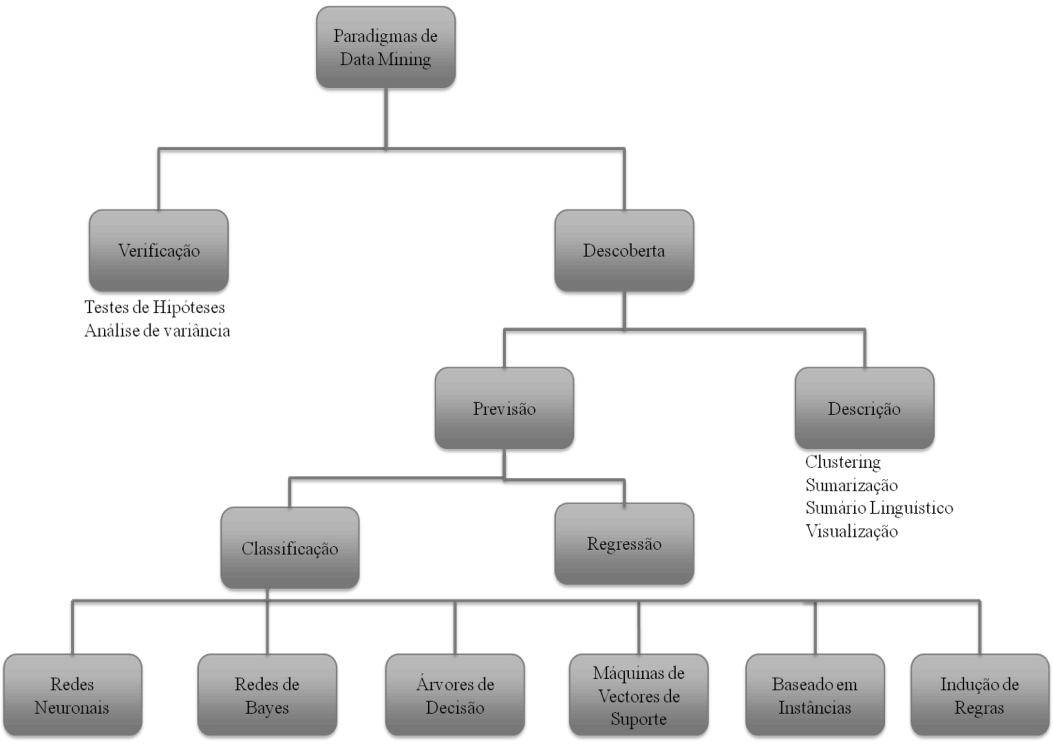


Figura 1. A taxonomia do DM, Adaptado de [Maimon e Lior, 2010]

2.2. Metodologias de *Data Mining*

Para facilitar o sucesso dos projetos de DM, foram desenvolvidas metodologias, das quais se destacam: a SEMMA *(Sample, Explore, Modify, Model, Assesment)*, criada pela SAS; e a CRISP-DM, que atualmente é a mais usada. Segundo [Maimon e Lior, 2010] a CRISP-DM é uma metodologia que surgiu na tentativa de padronizar o processo de DM. Foi construída não só tendo por base o conhecimento académico dos seus autores, mas também a experiência que eles adquiriam ao longo dos anos. Quando comparada com a SEMMA, a CRISP-DM tem a vantagem de ser neutra no que diz respeito à adoção de ferramentas de DM. Assim, nesta dissertação será adotada a metodologia CRISP-DM, sendo que a mesma já é considerada pelo rminer [Cortez, 2010].

Como podemos observar na Figura 2, o ciclo de vida da metodologia CRISP-DM, que consiste em seis fases [Chapman et al., 2000], [Maimon e Lior, 2010]:

- **Estudo do negócio**. Foca-se na compreensão dos objetivos do projeto e seus requisitos, segundo a perspetiva de negócio. Esse conhecimento é utilizado na definição do problema de DM e na criação de um plano preliminar para atingir os objetivos;
- **Estudo dos dados**. Inicia-se com uma recolha inicial dos dados e estende-se com atividades que permitem ao utilizador familiarizar-se com os dados;
- **Preparação dos dados**. Esta fase cobre todas as atividades que permitem a construção final do *dataset*. Nesta etapa, tabelas, registos, e atributos, são transformados e "limpos" para serem utilizados por ferramentas de modelação, de forma a formular hipóteses;
- **Modelação**. São selecionadas e aplicadas várias técnicas (e.g., AD, RNAs, MVSs), e os seus parâmetros são calibrados de forma a atingir o valor ótimo;

- **Avaliação**. São avaliados os modelos para verificar se permitem cumprir os objetivos propostos inicialmente. Essa verificação é realizada com recurso a técnicas (e.g., Matriz Confusão, Custo de erro e Curva ROC);
- **Implementação**. O conhecimento extraído dos modelos é organizado para ser posteriormente apresentado aos clientes. O projeto é documentado e resumido em relatórios.

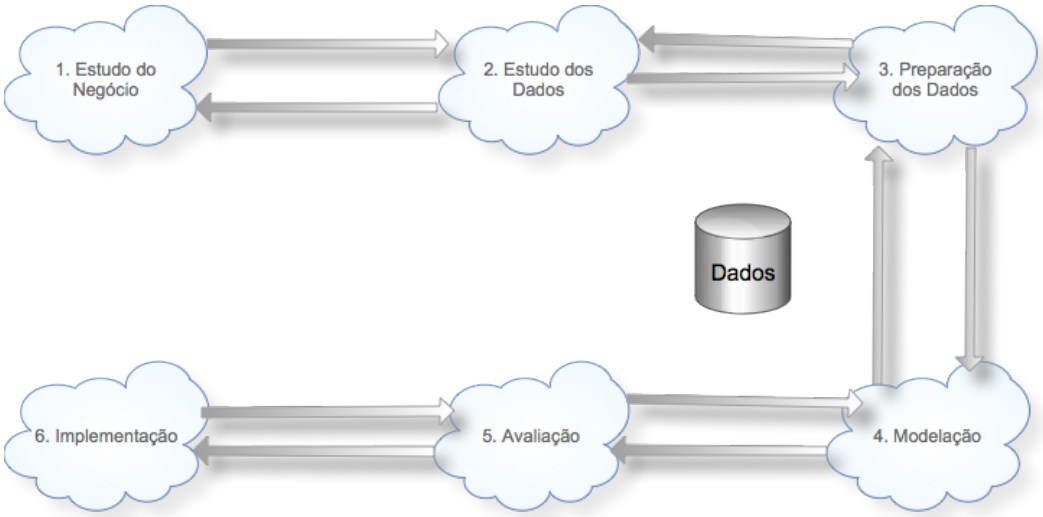


Figura 2. Ciclo de vida do CRISP-DM. Adaptado de [Maimon e Lior, 2010]

2.3. Classificação

A Classificação é, talvez, o objetivo mais comum no mundo do DM, sendo as AD e as RNAs algumas das técnicas mais usadas. O objetivo da Classificação é analisar informação histórica existente num repositório de dados, e, automaticamente, gerar um modelo que permita prever uma classe para um novo item. As técnicas utilizadas usam um conjunto de dados de treino para criar modelos com classes pré-definidas, para assim posteriormente ser possível aplica-lo a dados não classificados [Turban et al., 2010].

Na Figura 3 podemos observar um exemplo de Classificação referente a empréstimos bancários. Este exemplo baseia-se em 23 casos de pedidos de empréstimos, e são considerados como atributos o rendimento da pessoa que pede o empréstimo e o valor do mesmo. Os dados são classificados em duas classes: o X que representa os pagadores de risco e o 0 que representa os pagadores seguros. Através dessas duas classes o banco poderá decidir sobre a atribuição dos empréstimos [Fayyad et al., 1996].

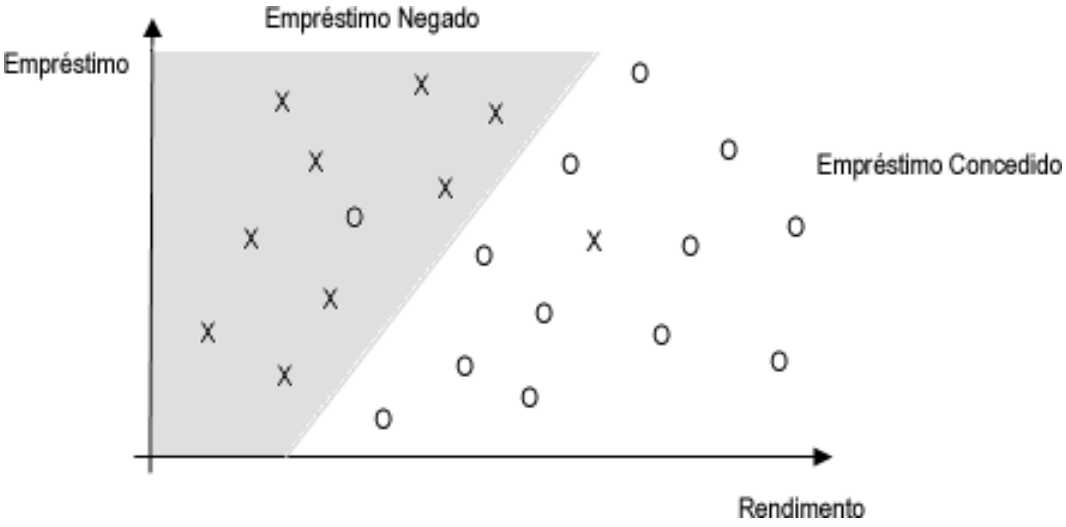


Figura 3. Exemplo de uma classificação. Adaptado de [Fayyad et al., 1996]

2.4. Regressão

A Regressão tem como objetivo encontrar uma função que permita prever uma variável, ou seja, consiste na aprendizagem de uma função que represente de uma forma aproximada o comportamento de variáveis. As regressões são aplicadas a diversos casos (e.g., prever a quantidade de biomassa presente numa floresta, estimar probabilidade de um paciente sobreviver) [Fayyad et al., 1996].

Na Figura 4 podemos observar uma regressão linear simples, sendo a dívida definida como uma função linear do rendimento.

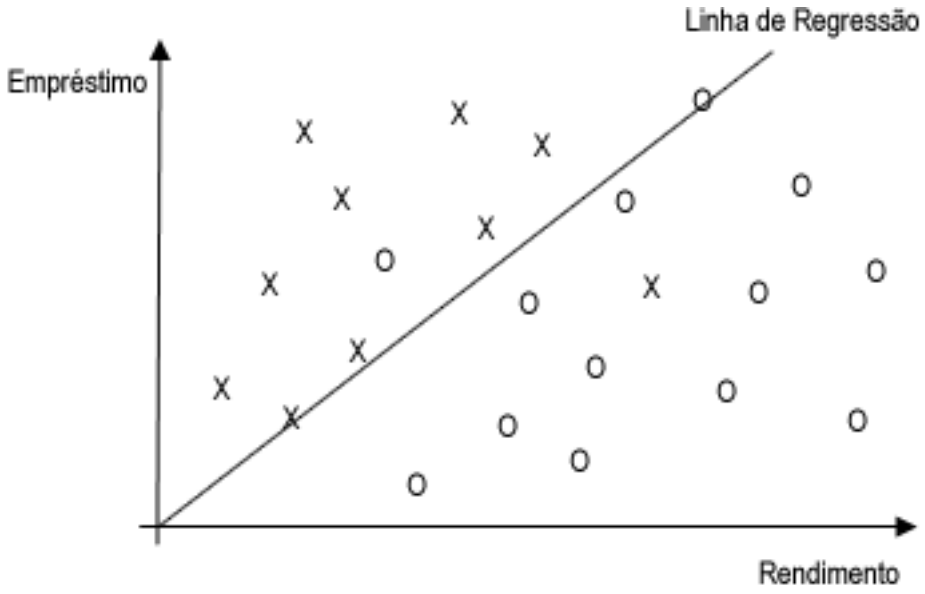


Figura 4. Exemplo de regressão linear. Adaptado de [Fayyad et al., 1996]

2.5. Algoritmos de *Data Mining*

De seguida, descreve-se, de modo sucinto, alguns dos algoritmos de aprendizagem que são utilizados pela biblioteca rminer.

2.5.1. Redes Neuronais

As RNAs são uma importante técnica de DM para modelação quantitativa. Começaram a ganhar popularidade a partir de meados da década de 80, e neste momento são utilizadas nas mais diversas áreas (e.g., ciência, indústria). O crescente interesse dos profissionais desta área, reside no facto de as RNAs serem sistemas flexíveis que aproximam com precisão diversos tipos de funções. O elevado sucesso em vários problemas (e.g., reconhecimento de padrões, controlo de processos, controlo de robôs) faz desta técnica umas das mais eficazes na área de DM [Costa e Simões, 2008].

Segundo [Maimon e Lior, 2010] as RNAs são modelos computacionais para o processamento de informação, e são fundamentalmente importantes na identificação de relações num conjunto de variáveis ou na descoberta de padrões

nos dados. Esta técnica de DM é definida como uma estrutura computacional composta por simples unidades de processamento, designadas por neurónios, que estão conectados entre si através de ligações de entrada e saída, sendo que cada ligação tem um peso associado. Esta interligação permite o envio de sinais entre os vários nodos, possibilitando a este modelo uma propensão natural para armazenar conhecimento empírico, e torná-lo acessível ao utilizador [Cortez, 2002]. Esta estrutura computacional é baseada em algumas capacidades do cérebro humano, sendo que segundo [Maimon e Lior, 2010], elas partilham duas características: processamento paralelo da informação, e aprendizagem e generalização da experiência. Já [Haykin, 1999] diz que as RNAs apresentam dois aspetos semelhantes ao cérebro humano:

- O conhecimento é adquirido através de um processo de aprendizagem, sendo esse processo afetado pelo meio que o rodeia;
- O conhecimento adquirido é guardado nas ligações entre os neurónios.

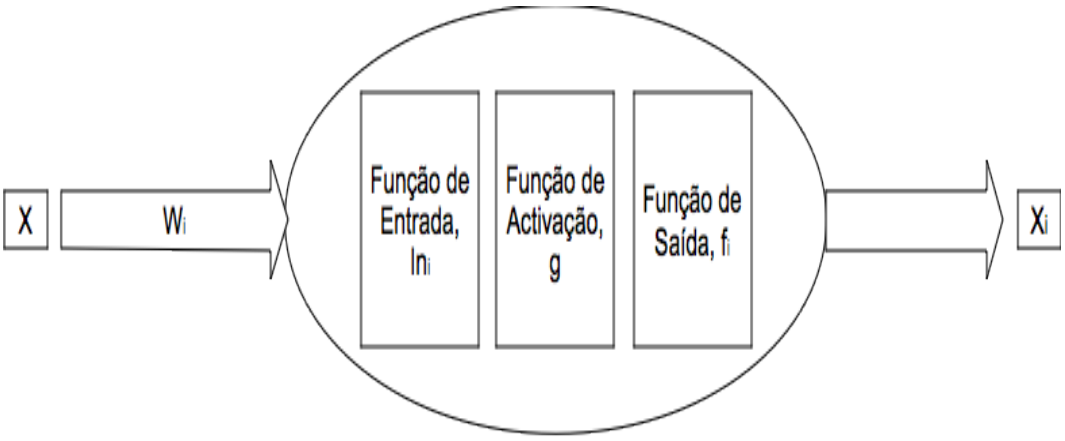


Figura 5. Modelo de um neurónio artificial. Adaptado de [Costa e Simões, 2008]

Na figura 5 é apresentado um neurónio artificial onde $X=<x_0, x_1,....., x_n>$ representa o valor de entrada.. A força de ligação entre as entradas e o neurónio é representado por um vetor $W_i=<w_{i0}, w_{i1,} w_{i2},....., w_{in}>$. Os sinais pesados da entrada são combinados e produzem o sinal total de entrada. A técnica mais comum de combinar os sinais de entrada é através da sua soma pesada, representada pela seguinte função [Costa e Simões, 2008]:

$$In_i = W_i * X = \sum_{k=0}^{n} W_{jk} * X_k$$

Após obtermos o sinal total de entrada determina-se o nível de ativação do neurónio com recurso à função g, representada a seguir:

$$a_i = g(In_i) = g\left(\sum_{k=0}^{n} W_{jk} * X_k\right)$$

Com o nível de ativação calculado é possível calcular-se a saída X_i. Normalmente essa saída é igual a a_i. Contudo o cálculo é efetuado com recurso à seguinte função:

$$X_i = f_i(a_i)$$

Em problemas de Classificação e Regressão, onde o objetivo principal é obter uma boa capacidade de previsão, as RNAs assumem-se como uma boa técnica de modelação. Contudo, [Berry e Linoff, 2000] afirmam que esta técnica tem dificuldades em trabalhar com um número elevado de padrões de entrada, isto é, um grande número de padrões pode implicar um longo tempo de treino que pode ter dificuldade em convergir para uma boa solução.

2.5.2. Árvores de Decisão

Basicamente uma AD permite dividir recursivamente um conjunto de dados de treino até que cada divisão forneça uma classificação para a instância. As AD (ver Figura 6) consistem em nós que formam uma árvore, o que significa que, existe um nó-raiz que não tem ramos de entrada, ao contrário dos restantes nós. Cada nó intermédio específica um teste para o atributo, e cada ramo descendente desse nó corresponde ao valor possível desse atributo. Este conjunto de regras é seguido até ser atingido o nó-terminal ou folha [Maimon e Lior, 2010], [Turban et al., 2010].

Segundo [Michalski, 1998] as árvores geradas seguem a seguinte estrutura:

- Folhas. Correspondem às classes;
- Nós. São os atributos nos quais estão ligadas subárvores;
- Ramos: São os valores dos atributos.

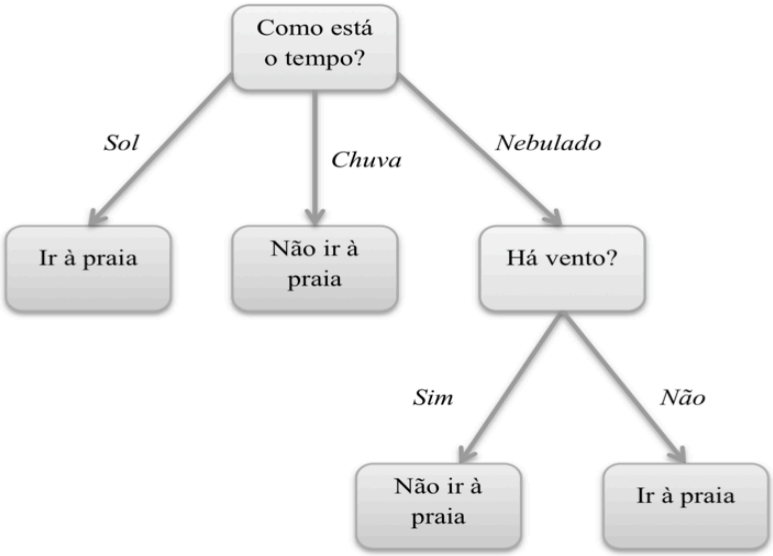


Figura 6. Exemplo de uma Árvore de Decisão. [Michalski, 1998]

[Quinlan, 1998] e [Berry e Linoff, 2000] afirmam que existe dois tipos de AD:

- Classificação, que servem para qualificar e associar os registos com a classe determinada;
- Regressão, que realizam uma estimativa de valor de uma determinada variável.

2.5.3. Máquinas de Vetores de Suporte

As MVSs são especialmente vocacionadas para uma aprendizagem supervisionada, sendo nos dias de hoje muito utilizadas em problemas de

Classificação. Desde o seu aparecimento, as MVSs ganharam uma elevada notoriedade, muito por culpa do bom suporte teórico [Maimon e Lior, 2010].

As técnicas tradicionais são muito focadas para a minimização do risco empírico (i.e., tentar otimizar o desempenho de um conjunto treino). O risco depende da complexidade do conjunto de funções escolhidas, bem como do conjunto treino. Já as MVSs tentam reduzir o risco estrutural (i.e., probabilidade de ocorrer classificação errada de padrões segundo uma probabilidade de distribuição de dados fixa, mas desconhecida). Este facto faz com que os resultados da aplicação deste técnica sejam comparáveis e muito vezes superior aos obtidos por outros algoritmos de aprendizagem (e.g., RNAs) [Pontil e Verri, 1998], [Maimon e Lior, 2010].

O processo de treino das MVSs consiste na obtenção de valores para os pesos que definem um "hiperplano" ideal de mapeamento, de forma a minimizar a função de custo, num processo semelhante ao que ocorre nas RNAs [Pontil e Verri, 1998]. A diferença, é que no caso das MVSs é garantido que os pesos são ótimos, sendo que a não linearidade é garantida via uma função de *kernel*, que transforma o espaço original das entradas num espaço imaginário, de modo a que neste espaço a separação seja do tipo linear.

2.5.4. Random Forest

***Random Forest* (RF)** é uma combinação de previsões onde cada árvore prevista depende de vetores aleatórios independentes e com a mesma distribuição para todas as árvores. Além da construção de cada árvore ser efetuada com recurso a uma amostra aleatória (*bootstrap*) dos dados, o algoritmo RF altera a forma de construção das árvores de classificação. Nas árvores tradicionais, cada nó é dividido usando a melhor divisão entre todas as variáveis, enquanto que no algoritmo de RF o nó é dividido usando o melhor entre um subconjunto de indicadores escolhidos aleatoriamente naquele nó.

Esta estratégia um pouco contraintuitiva acaba por funcionar bem em comparação com outros classificadores, incluindo as AD e discriminantes lineares. Além disso este algoritmo é *"user-friendly",* uma vez que só possui dois parâmetros (o número de variáveis no subconjunto aleatório em cada nó e o número de árvores da floresta) [Breiman, 2001].

2.5.5. Naive Bayes

Um classificador ***Naive Bayes*** **(NB)**, é um simples classificador probabilístico baseado no teorema de Bayes. Ele pode prever a probabilidade de uma associação de classes, tal como, a probabilidade de um determinado tuplo pertencer a uma classe particular. Os classificadores NB assumem que o efeito de um valor de um atributo em uma determinada classe, é independente do efeito dos valores dos outros atributos. Esta hipótese é chamada de *Class Conditional Independence* [Han e Kamber, 2006].

2.5.6. Multiple Regression

O objetivo principal da Regressão Múltipla é modelar uma relação linear entre diversas variáveis independentes (ou entradas), e uma variável dependente (variável de saída). Por exemplo, um agente imobiliário pode registar para cada lista, o tamanho da casa, o número de quartos, a renda média de uma casa naquela zona, e uma avaliação subjetiva da casa. Uma vez esta informação compilada, seria interessante verificar se/e como essas medidas se relacionam com o preço de modo linear com o preço de venda. Com isto, é possível saber, por exemplo, que o número de quartos é melhor indicador do preço de uma casa, do que a estética dela. Também é possível detetar *"outliers",* isto é, casas que possuem um preço mais elevado devido à sua localização e características [StatSoft-Inc, 2011].

2.5.7. K-Nearest Neighbor

O algoritmo K-Nearest Neighbor é um algoritmo que se insere na aprendizagem supervisionada e tem como fundamento a ideia de que os objetos mais próximos têm mais probabilidade de serem do mesmo tipo, ou seja, consiste em encontrar os K objetos classificados como mais próximos dos objetos que se pretendem classificar. Após a procura estar concluída, os novos objetos serão classificados conforme os atributos dos vizinhos. Este tipo de algoritmo requer pouco esforço computacional durante a etapa de treino, contudo, o esforço poderá aumentar na etapa de classificação de novos objetos, pois, na pior das hipóteses, cada objeto não classificado deverá ser comparado com todos os exemplos contidos no conjunto treino [Aha et al., 1991] [StatSoft-Inc, 2011].

2.5.8. *Linear* e *Quadratic Discriminant Analysis*

Linear Discriminant Analysis **(LDA)** é um dos métodos mais eficazes na extração de características, sendo muito popular na área da estatística, e do reconhecimento de padrões. Os métodos baseados em LDA têm a capacidade de extrair características discriminantes maximizando o chamado critério de Fisher.

Existem estudos que mostram LDA como um caso especial do classificador bayesiano ótimo, quando a distribuição dos dados das classes condicionais seguem a função gaussiana, com uma estrutura de covariância idêntica. Nesses casos, os limites das classes resultantes são lineares. Contudo, em muito problemas de reconhecimento de padrões, a distribuição é mais complicada. Nesse casos, o desempenho de técnicas LDA diminui drasticamente. Para esse tipo de problemas as técnicas não lineares parecem ser mais eficazes a lidar com a distribuição de dados complexos.

Quadratic Discriminant Analysis **(QDA)** é uma das técnicas, não lineares, mais utilizadas no reconhecimento de padrões. Na técnica QDA, a distribuição da classe condicional segue a função gaussiana, porém, com um valor

para as matrizes de covariância diferentes. Nesses casos, um limite quadrático mais complexo pode ser formado. Portanto, QDA adapta-se melhor a uma estrutura de dados real.

2.5.9. Regressão Logística (*Logistic Regression*)

O algoritmo de Regressão Logística, ***Logistic Regression*** **(LR),** aplica uma função logística a uma regressão linear, sendo muito utilizado em situações onde a variável de saída é binária ou dicotómica.. Por exemplo, num contexto escolar podemos usar uma busca dicotómica para definir sucesso/fracasso de um alunos. De modo similar, num ambiente médico o resultado pode ser doente/não doente. De referir que o LR é muito utilizado em áreas como a Medicina, Ciências Sociais e o *Marketing* (e.g., previsão da propensão de um cliente para comprar um determinado produto) [Dayton, 1992]

3. Ferramentas *Data Mining*

Neste capítulo é feita uma introdução às ferramentas DM, sendo apresentada uma lista das ferramentas mais utilizadas, e, por fim, introduz-se a ferramenta R/rminer

3.1. Introdução

A escolha da ferramenta ideal de DM é bastante complexa, na medida em que tem que ser tomado em conta diversos fatores, como por exemplo o domínio do problema (i.e., conjunto condições que influenciam a escolha de uma ferramenta), sistema operativo, custo, e licença. Em particular, uma aplicação DM multiplataforma apresenta diversos pontos fortes, sendo mais natural o facto de se poder adaptar a vários tipos de ambientes computacionais. Outra característica importante nas ferramentas é o tipo de uso, isto é, pode ser utilizada para fins académicos ou comerciais. De facto, todos os produtos de *software* de DM têm uma licença de uso, ou seja, aplicações *open-source* permitem o uso e alteração ao produto sem restrições, enquanto que no licenciamento comercial o produto é fechado e foi construído com o objetivo de proporcionar à empresa que o desenvolveu retornos financeiros. Por sua vez, a crescente expansão da quantidade de dados juntamente com a necessidade da descoberta de conhecimento, levaram as organizações a investir nas tecnologias de DM. Tais tecnologias, operam muitas das vezes separadas dos dados exigindo esforço e tempo em tarefas de exportação e importação. Com isto, torna-se desejável uma ligação das aplicações aos Sistemas de Gestão de Bases de Dados existentes.

[Goebel e Gruenwald, 1999] definem um conjunto de características determinantes para o desempenho das diversas ferramentas de DM. São elas:

- **Capacidade de aceder às fontes de dados**. Em grande parte dos casos, os dados a serem analisados estão dispersos por diferentes **BD (Bases de Dados)**. Isto implica a realização de trabalhos de importação, verificação, e exportação antes do processamento;

- **Acesso *online/offline* aos dados**. O acesso *online* significa que as consultas são executadas diretamente na BD, enquanto, em simultâneo, são realizadas transações. Já no acesso *offline* é realizada uma cópia dos dados implicando a exportação/importação para um formato exigido pela ferramenta DM. Esta questão do *online/offline* torna-se importante quando se tem que lidar com situações em que a informação está em constante mudança;
- **Modelos de dados**. A maioria das ferramentas adequam-se ao modelo hierárquico e relacional, sendo este último mais utilizado. Modelos de dados orientados aos objetos e também não padronizados, tais como, multimédia, espacial ou temporal, são incomuns nas tecnologias de DM;
- **Número de tabelas/linhas/atributos**. Estes são os limites, na teoria, da capacidade de processamento da ferramenta DM;
- **Tamanho da BD com que a ferramenta pode lidar**. A quantidade de dados previstos para análise deve ser um fator a ter em conta na escolha de uma ferramenta, pois a quantidade de tabelas/linhas/atributos influência os recursos de um computador (e.g., memória, processamento, espaço em disco). Uma ferramenta que coloque toda a informação na memória principal não é apropriada a um conjunto alargado de dados;
- **Consultas.** Algumas ferramentas permitem ao utilizador executar consultas às hipóteses geradas através de uma interface, ou então recorrendo à linguagem SQL, ou similar;
- **Tipo de atributos com que a ferramenta DM pode lidar.** Isto significa que as aplicações podem ter restrições em relação aos tipos de atributos com que lidam (e.g., a RNAs requerem atributos de entrada numéricos).

[Goebel e Gruenwald, 1999] apresentam no seu estudo critérios de exclusão de ferramentas, tais como:

- Se uma ferramenta funcionar como um servidor de dados para outras ferramentas DM;
- Se não foi desenvolvida para tarefas de DM, embora possua algumas caraterísticas para tal;
- Ferramentas que só servem para visualização de informação.

Mais recentemente, [Piatetsky-Shapiro, 2010] apresentam um estudo sobre as ferramentas DM mais utilizadas pela comunidade DM. Esta análise é realizada com base em inquéritos *on-line* apresentados no site *kdnuggets.com.* Os resultados podem ser observados através da Tabela 1, e apesar de a informação não possuir rigor científico, este é o estudo mais atual sobre ferramentas DM.

Este questionário foi respondido por 912 visitantes, e mostra claramente que as ferramentas de código aberto se encontram no topo das preferências dos utilizadores (ver Tabela 1). As principais são o RapidMiner, R e o KNIME. Em relação às ferramentas comerciais destacam-se o SAS, Matlab, SPSS e IBM Modeler (ex- Clementine) [Piatetsky-Shapiro, 2010]. Ainda segundo um estudo da [RexerAnalytics.com, 2010], nos últimos anos a ferramenta R ultrapassou as outras ferramentas *open-source*, para se tornar na ferramenta mais usada em DM (43%). Ao nível de análise estatísticas, R recebeu uma classificação semelhante a ferramentas como o SPSS e IBM Modeler. Os factos evidenciados mostram que o R está a tornar-se numa aplicação cada vez mais completa tanto na área do DM como da estatística.

Tabela 1. Ferramentas mais usadas na comunidade DM. Adaptado de [Piatetsky-Shapiro, 2010]

Ferramenta	Percentagem de utilização
RapidMiner	**37.8 %**
R	**29.8 %**
Excel	24.3 %
KNIME	**19.2 %**
Código Próprio	18.4 %
Pentaho/Weka	**14.3 %**
SAS	12.0 %
MATLAB	9.2 %
IBM SPSS Statistics	7.9 %
Outras Ferramentas Livres	7.3 %
IBM SPSS Modeler (former Clementine)	7.3 %
Microsoft SQL Server	6.9 %
Statsoft Statistica	6.2 %
Other commercial tools	6.1 %
SAS Enterprise Miner	5.5 %
Zementis	3.7 %
Orange	2.7 %
Oracle DM	2.1 %
KXEN	2.1 %
Salford CART Mars other	1.6 %
VisuaLinks	1.3 %
Viscovery	1.1 %

3.2. Ambientes gráficos *Open-Source*

3.2.1. Weka

A ferramenta *Waikato Environment for Knowledge Analysis* (Weka) surgiu em 1993 proveniente de financiamentos do governo da Nova Zelândia. A aplicação foi construída na linguagem Java com intuito de aplicá-la à economia do país. Hoje em dia a Weka é aceite em todo o mundo, tanto pelas Universidades como pelas empresas.

Atualmente, a ferramenta pertence a uma organização e está licenciada ao abrigo da *General Public License*. Na Figura 7 podemos ver uma imagem da aplicação.

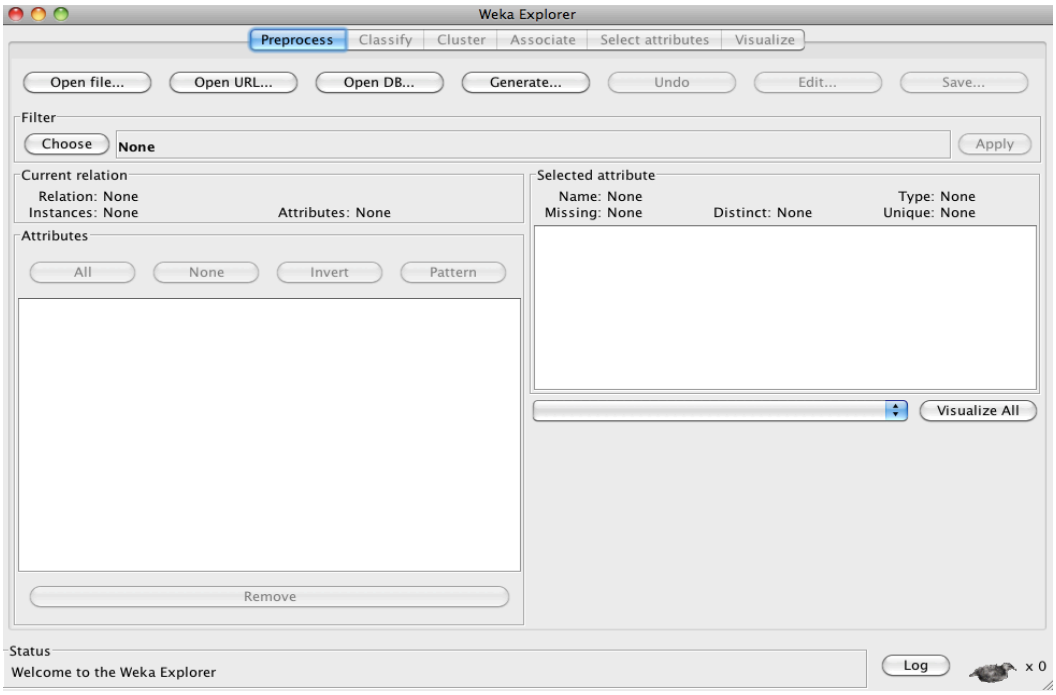


Figura 7. Exemplo da aplicação Weka

3.2.2. Rattle

Rattle (the R Analytical Tool to Learn easily) é uma aplicação gráfica (ver Figura 8), *open-source* que foi construída em cima da ferramenta R. Foi

desenvolvida especialmente para facilitar a transição do DM básico, normalmente oferecida pelas GUI de DM mais tradicionais, para uma análise de dados mais sofisticada, utilizando a poderosa linguagem R.

Esta GUI pode ser usada por utilizadores sem experiência em R e veio permitir a resolução de um conjunto alargado de problemas de DM. Atualmente, segundo [Williams, 2009], esta GUI é utilizada no ensino por várias Universidades e consultores de todo o mundo.

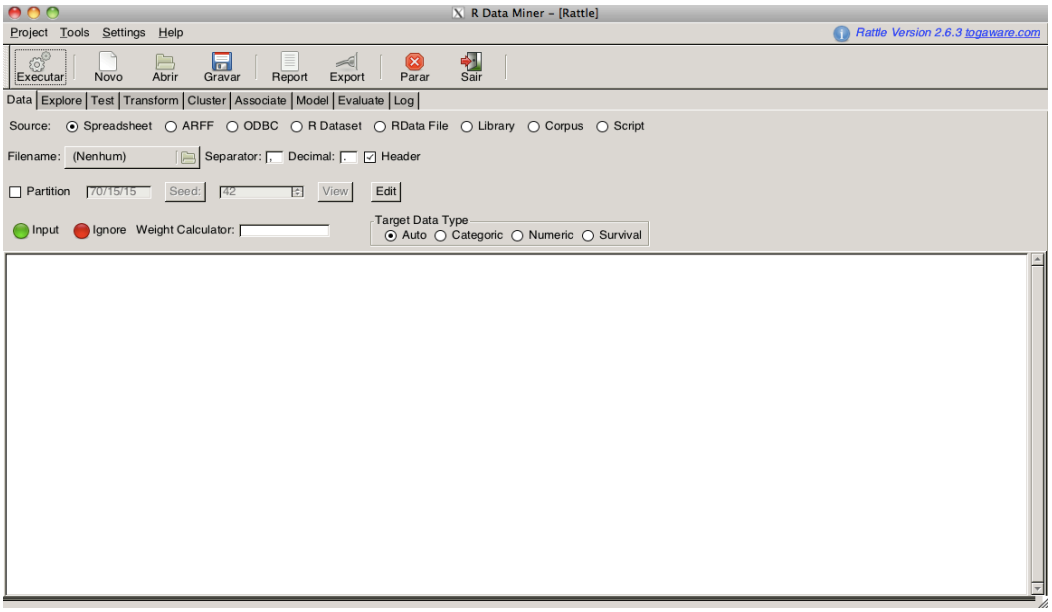


Figura 8. Exemplo da aplicação Rattle

3.2.3. RapidMiner

O RapidMiner é uma aplicação líder mundial dos sistemas *open-source* para DM. Esta ferramenta está disponível como uma aplicação *stand-alone* para análises de dados, e como um motor de DM para a integração dos seus próprios produtos.

O RapidMiner apresenta um conjunto de características, como:

- Integração de dados, ETL (*Extract, Transform, Load*), análise de informação e produção de relatórios, tudo numa única *suite*;
- Tem uma parte gráfica poderosa e intuitiva para análises de processos;
- Reconhecimento de erros *on-the-fly* e correções rápidas;

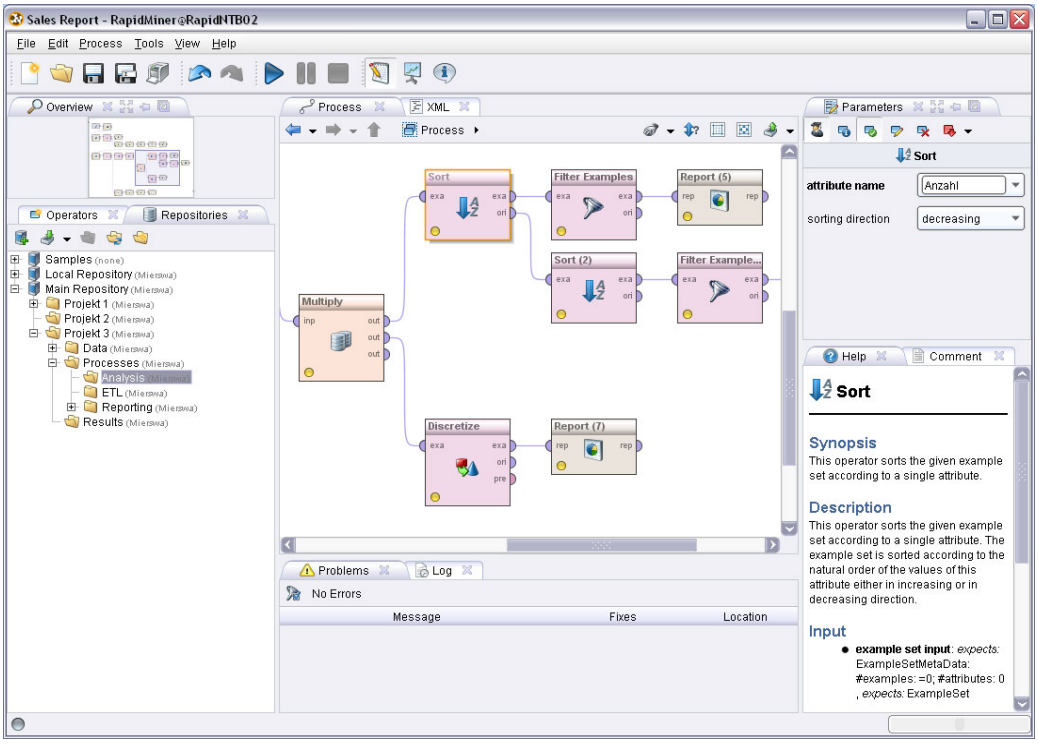


Figura 9. Exemplo da aplicação RapidMiner

3.2.4. KNIME

KNIME (Konstanz Informação Mineiro) é uma aplicação *open-source* que permite a integração, processamento, e análise de dados. Atualmente o software é usado por mais de 6000 profissionais em todo o mundo. Através do *knime.com* é possível fazer o suporte e manutenção da aplicação, além de usufruir de um conjunto de *add-ons*.

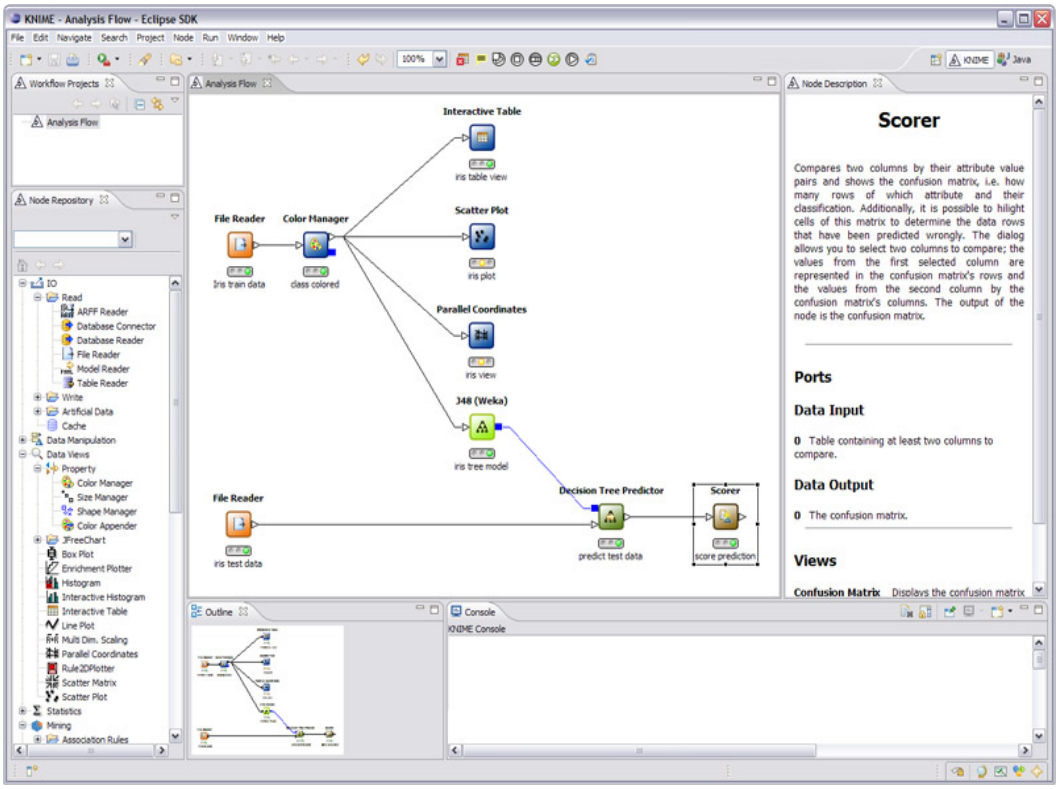


Figura 10. Exemplo da aplicação KNIME.

3.3. Ferramenta R/rminer

O ambiente de programação R é um projeto GNU e foi desenvolvido por Ross Ihaka e Robert Gentleman no Departamento de Estatística da Universidade de Auckland, Nova Zelândia. Este sistema foi criado com intuito de permitir uma programação direcionada para a estatística e análise de dados. Pode ser instalado em qualquer plataforma (e.g., *Windows, Mac OS, Linux*) e possuí a grande vantagem de podem ser estendido, nomeadamente via os milhares de *packages* existentes no repositório CRAN.

A linguagem foi descrita pela primeira vez em 1996, como a junção das linguagens S e Scheme. Essa junção, teve como finalidade juntar estes dois ambientes num só, de forma potenciar as suas vantagens [Ihaka e Gentleman,

1996]. Segundo [Chambers, 2008] R é um dialeto da Linguagem S e utiliza um estilo de programação em funções.

Este ambiente é um conjunto integrado de software que permite a manipulação de dados, cálculo e exibição gráfica. Em particular, com o R consegue-se:

- um tratamento eficaz dos dados e facilidade no seu armazenamento;
- um conjunto de operadores para cálculos;
- inúmeras ferramentas intermédias, que podem ser descarregadas do CRAN conforme o interesse do utilizador e que permitem efetuar análises sofisticadas aos dados;
- facilidades gráficas para o utilizador;
- Uma programação simples, por objetos e eficaz, baseada em funções.

A comunidade do R é muito ativa e constantemente são criadas novas bibliotecas para resolução de novos problemas. No repositório CRAN existe um conjunto de bibliotecas (ver Tabela 2) que permite aos utilizadores integrarem no ambiente R aplicações escritas noutras linguagens de programação, tais como o Java, C, C++ e Perl, possibilitando assim, o aumento das potencialidades.

Atualmente, segundo [Williams, 2009], o R é usado em muitos trabalhos de DM, pois oferece uma amplitude e profundidade nas análises estatísticas, muito para além do que é conseguido com as aplicações proprietárias. [Miller, 2006] diz que o R, quando direcionado para o ***Business Intelligence (BI)***, apresenta vantagens e desvantagens em relação a outras ferramentas de BI (e.g., SAS), nomeadamente:

Vantagens:

- Orientado aos objetos, sendo extensível;
- Integração de estatística, programação, gráficos;
- As diversas bibliotecas permitem uma expansão das capacidades do R.

Desvantagens:

- O R armazena todos os seus dados em memória, o que o torna limitado em grande volumes de dados;
- Uma vez que funciona via "comandos" a sua aprendizagem torna-se mais difícil;
- Apesar de existir uma comunidade grande, é evidente uma falta de apoio técnico ao nível da formação/manutenção.

Tabela 2. Bibliotecas de ligação ao R. Adaptado de [Chambers, 2008]

Linguagem/Sistema	Aplicações	Biblioteca R	Repositório
Perl	WWW, Interfaces	RSPerl	Omegahat
Python	Similar ao Perl	RSPython, rpy	Omegahat, Sourceforge
Java	**Gráficos, Interface**	**rJava, JRI**	**CRAN, RForge**
Oracle, MySQL	Base de dados	ROracle, RmySQL	CRAN

No que diz mais respeito a esta dissertação, a biblioteca rminer começou a ser desenvolvida em 2006, estando atualmente na sua versão 1.1. O seu desenvolvimento iniciou-se para colmatar a falta de funções coerentes em R para aplicar algumas técnicas de DM a um conjunto variado de áreas (e.g., qualidade de vinhos [Cortez et al., 2009a] e deteção de spam [Cortez et al., 2009b]). Em particular, muitos dos alunos que trabalharam com o orientador desta dissertação, oriundos de diversas áreas, tais como Sistemas de Informação, Gestão e Engenharia Civil, apresentaram uma forte resistência inicial ao uso da linguagem R em tarefas de DM. Com a existência do *package* rminer, esta resistência foi-se atenuando.

O rminer permite resolver tarefas do tipo:

- **Classificação.** O objetivo é encontrar uma função que permita associar um conjunto de casos a possíveis classes pré-definidas. Neste modelo o *output* é a probabilidade p(c) para cada classe, sendo que $\sum_{c=1}^{Nc} p(c) = 1$;
- **Regressão**. O objetivo é estimar um valor real através de um conjunto de atributos.

Existe um conjunto enorme de algoritmos DM que permitem resolver estas tarefas. Contudo, o rminer centra-se sobretudo nas RNAs e MVSs. Ambos os modelos são flexíveis e admitem uma modelação não linear dos dados. Não obstante esta especialização, o rminer também permite lidar com diversos outros algoritmos de DM, incluindo por exemplo as AD e o NB.

Em [Cortez, 2010], argumenta-se que o rminer oferece as seguintes funcionalidades:

- Apresenta um conjunto reduzido e coerente de funções, simplificando o uso de algoritmos supervisionados na ferramenta R;
- Realiza uma seleção automática de modelos, ou seja, ajuste ideal dos hiper-parâmetros das RNAs/MVSs;
- Permite o cálculo de diversas métricas e gráficos que são úteis para o DM, incluindo procedimentos de análise de sensibilidade para extrair informação a partir de modelos treinados.

De referir que esta biblioteca funciona à base de comandos, tal como a maior parte das bibliotecas do R, o que faz com que a curva de aprendizagem para utilizadores não especializados com a linguagem R seja considerável. No entanto, depois de dominar o ambiente R/rminer, o utilizador obtém um melhor controlo e compreensão da informação que está a ser executada. Assim, existe uma lacuna que foi identificada pelo autor desta biblioteca e que será trabalhada nesta dissertação: a falta de um ambiente gráfico que facilite a utilização da biblioteca rminer por parte de utilizadores não especializados com o ambiente R. Em

particular, este ambiente gráfico é considerado como potencialmente mais relevante numa fase inicial, para domínio dos comandos rminer.

4. Jrminer: um ambiente gráfico para o rminer

Este capítulo começa por apresentar os resultados referentes ao questionário enviado aos utilizadores do rminer. Numa fase seguinte, é apresentada a aplicação Jrminer, incluindo as tecnologias usadas no seu desenvolvimentos, bem como os packages necessários, instruções de instalação, e explicação do seu funcionamento com recurso a exemplos. No final são apresentados os resultados de um questionário referente à utilização do Jrminer,

4.1. Opinião dos utilizadores sobre o rminer

Numa fase inicial deste trabalho, foi construído um inquérito *on-line* com 9 questões, de modo a permitir ter uma visão sobre o perfil e as opiniões dos utilizadores em relação ao rminer. Importa referir que embora possa existir uma comunidade maior de utilizadores, já que o *package* rminer está disponível no repositório CRAN, o inquérito foi somente enviado a uma lista restrita de 27 utilizadores, que trabalharam diretamente com o orientador deste trabalho ou que manifestaram interesse (e.g., via email) no uso do rminer. Certamente seria ideal ter uma amostra maior de utilizadores, mas tal não foi possível dentro dos limites temporais definidos para esta dissertação, ficando tal estudo alargado para trabalho futuro.

O inquérito foi disponibilizado de modo eletrónico para os tais 27 utilizadores, estando disponível no sítio: http://www.surveymonkey.com/s/B3V7HYL sendo que apenas se obteve aproximadamente 26% de respostas (7 utilizadores). Pretendeu-se construir um questionário pequeno, devido às limitações que a ferramenta *web* impunha para utilizadores não pagos, com questões que cobrissem três diferentes áreas relacionadas com o utilizador. A primeira pretendia obter um perfil mais "pessoal" do utilizador. A segunda tinha o objetivo de obter a sua experiência em relação ao R e ao rminer. A terceira permitia conseguir opiniões sobre a aplicação que viria a ser desenvolvida no futuro.

O questionário não apresenta uma escala semelhante para todas as perguntas. Assim, a possibilidade de resposta de questão para questão varia entre campo de texto livre e resposta de escolha múltipla. Em seguida apresenta-se todas as questões efetuadas:

1. Qual a sua idade?
2. Quais as suas habilitações?
3. Qual a sua experiência na Linguagem R?
4. Qual a sua experiência no *package* rminer?
5. Porque razão usa o *package* rminer?
6. Quanto tempo demorou a dominar o rminer?
7. Acredita que um ambiente gráfico para o rminer era uma mais valia?
8. Na sua opinião, o que deve ser incluído no ambiente gráfico (funcionalidade, aspetos gráfico, etc)?
9. Por favor, deixe o seu e-mail para contacto, se necessário.

Algumas destas questões não têm como objectivo uma análise estatística (e.g., 5, 8 e 9), sendo antes utilizadas no questionário para se poder obter uma informação qualitativa, ou seja, para ter uma opinião mais pessoal sobre o rminer e também abrir a possibilidade de o utilizador contribuir com ideias para o GUI. Isto significa que, essas questões não aparecem nos resultados apresentados a seguir.

De seguida, serão analisadas as respostas do pequeno grupo de utilizadores. Ressalva-se desde já que com tal pequena amostra não se pode garantir um rigor estatístico, pelo que se pretende aqui pelo menos mostrar uma caracterização de alguns dos utilizadores do *package*. De seguida, serão mostrados os resultados obtidos para as questões mais relevantes.

Questão 1. "Qual a sua idade?"

Como se pode observar na Figura 11, são utilizadores entre os 18 e os 30 que utilizam mais a ferramenta R, confirmando de certo modo a noção já adquirida que esta ferramenta é mais utilizada no âmbito académico.

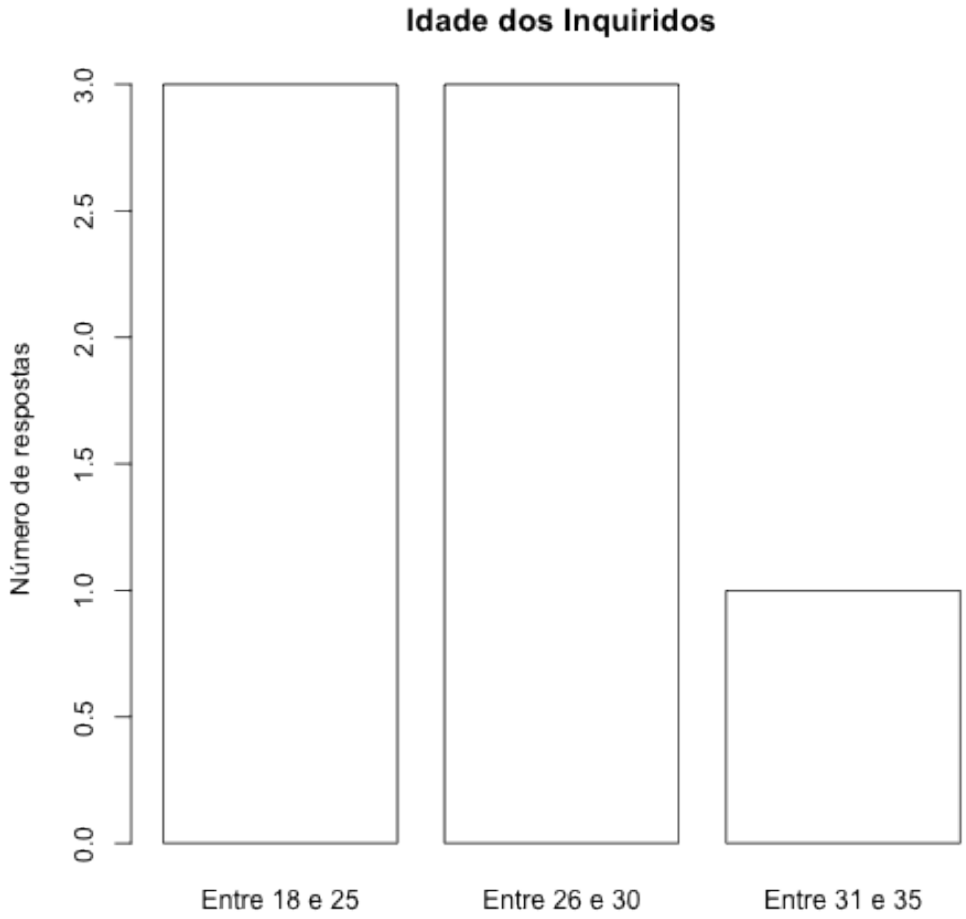


Figura 11. Idade dos utilizadores do rminer

Questão 2. "Quais as suas qualificações?"

Esta é uma questão com relativa importância, uma vez que, permite conhecer as qualificações dos utilizadores em relação ao rminer. Como se pode

verificar na Figura 12, o rminer foi experimentado em todos os graus de ensino superior, tendo o Mestrado levado alguma vantagem.

Figura 12. Qualificação dos utilizadores do rminer

Questão 3 e 4. "Qual a experiência dos utilizadores na Linguagem R e no *package* rminer?"

Procedeu-se à integração destas duas questões para efeitos de análise, uma vez que, de certa forma estão interligadas. Como podemos observar na Figura 13 existe uma certa coerência dos utilizadores em relação à experiência adquirida na utilização do R e do rminer. Este facto é indicador de que a maioria dos inquiridos começou a utilizar o R/rminer como proposta por parte do orientador deste

trabalho. A restante minoria, são utilizadores que já apresentavam "bons" conhecimentos no R, tendo posteriormente utilizado o rminer.

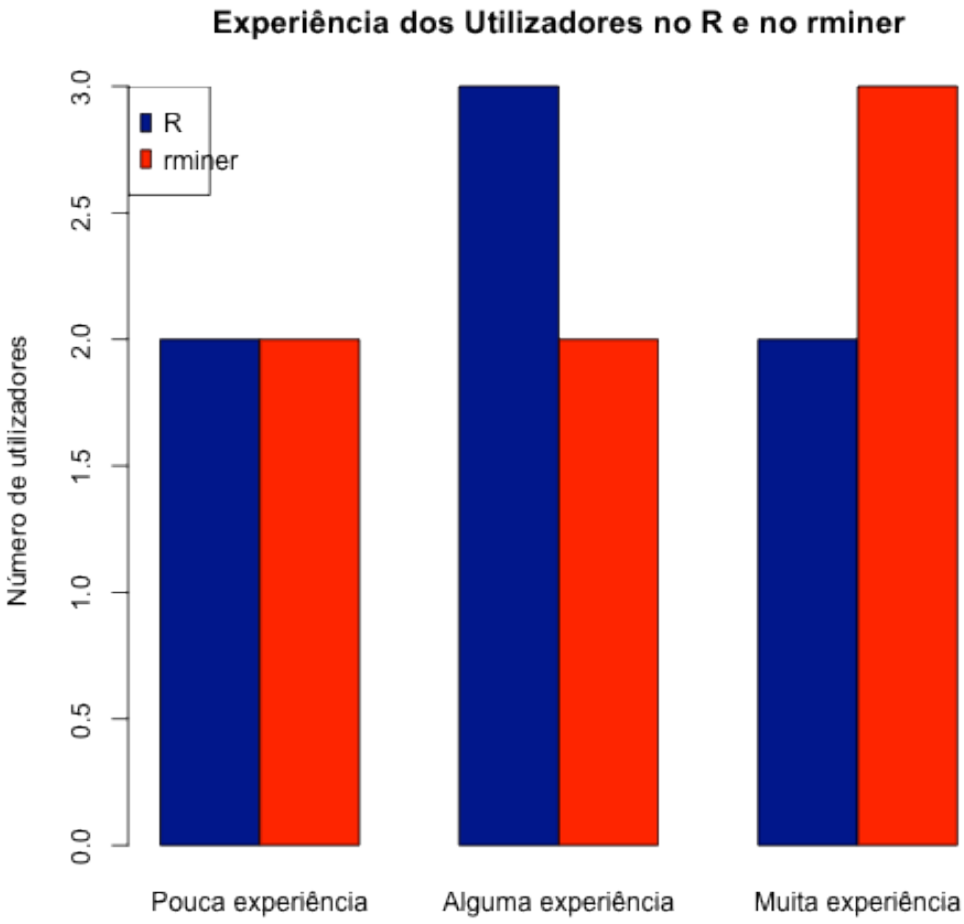


Figura 13. Experiência dos utilizadores no R e no rminer

Questão 6. "Quanto tempo demorou a dominar o rminer?"

Esta é uma questão considerada essencial para a este trabalho. Como já foi referido neste documento, este estudo tem como objetivo reduzir a curva de aprendizagem do rminer, com recurso a uma interface gráfica. Isto significa que, é fulcral para apresentação dos resultados uma comparação entre tempos pré e pós interface gráfica. Na Figura 14 é possível ver uma discrepância entre os tempos, isto é, a maior parte dos utilizadores inquiridos ou demora algum tempo (entre 0 –

20 horas) no domínio do rminer ou então **demora muito tempo** (mais de 40 horas).

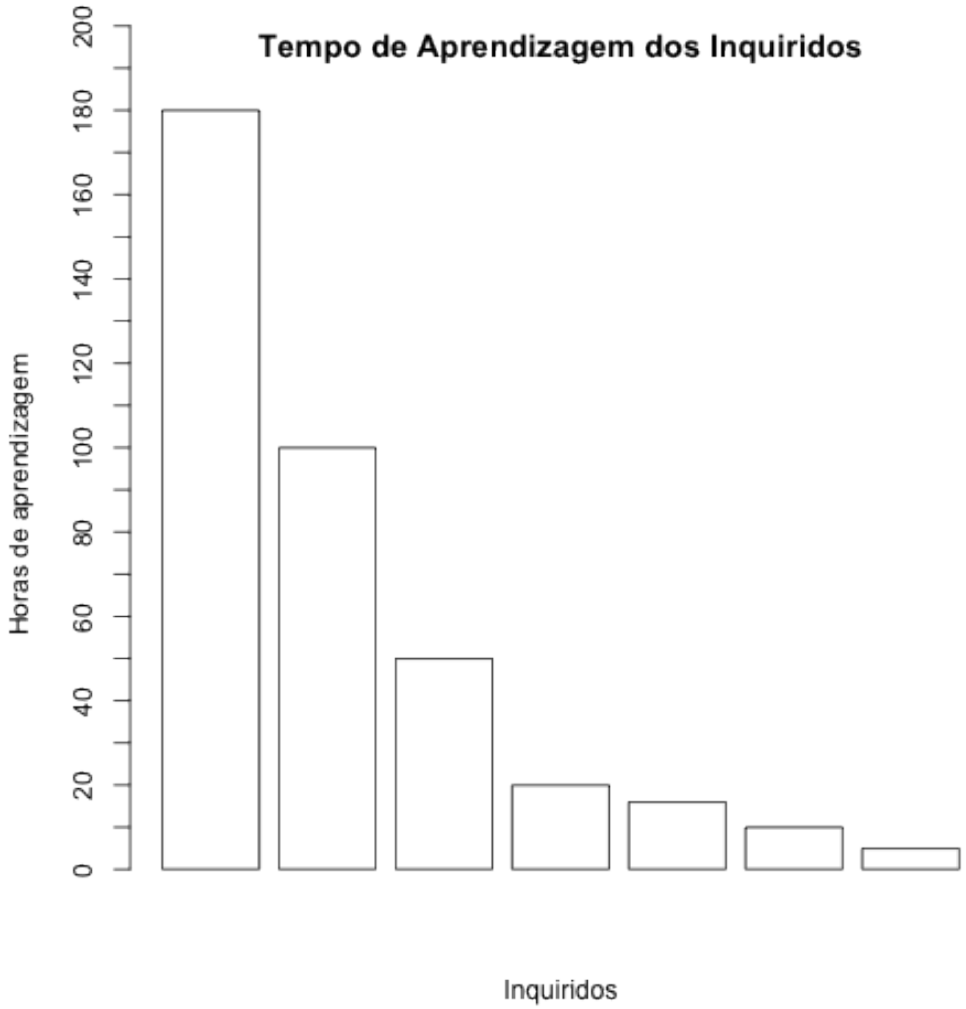


Figura 14. Tempo de aprendizagem do rminer

Como podemos observar na Figura 15 os utilizadores com idade compreendida entre os 18 e os 25 anos apresentam um tempo de aprendizagem com uma média abaixo das 20 horas. Os utilizadores que se encontram com idades entre os 26 e os 30 apresentam uns tempos de aprendizagem mais heterogéneos. Por exemplo, existe um utilizador que necessita de mais de **150 horas,** enquanto existe outro utilizador que precisa de menos de 20 horas.

Em relação à qualificação versus horas de aprendizagem, ver Figura 16, podemos verificar que os utilizadores que apresentam só um grau de licenciatura sentem mais dificuldade de aprendizagem, provavelmente devido ao facto de esta ferramenta ser mais conhecida/usada em áreas de investigação académica.

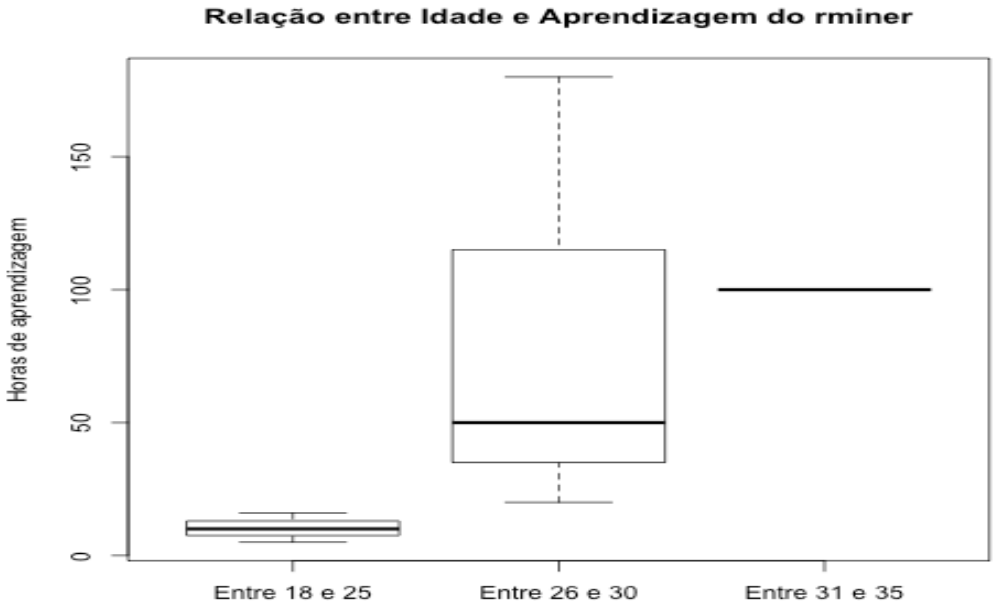


Figura 15. Relação existente entre idade e horas de aprendizagem

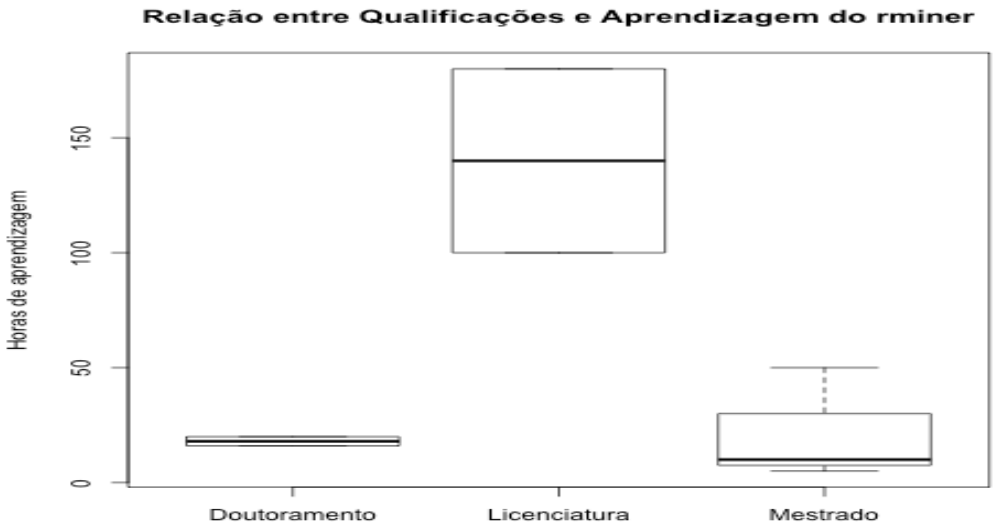


Figura 16. Relação existente entre qualificação e horas de aprendizagem

Questão 7. "Acredita que um ambiente gráfico para o rminer será uma mais valia?"

Como podemos verificar através da Figura 17 quase 100% dos utilizadores acredita que um ambiente gráfico é um valor acrescentado para o rminer. Isto evidencia, claramente, que nos tempos modernos os utilizadores de tecnologias sentem cada vez mais necessidade de utilizar ambientes gráficos. Este facto sugere que o rminer necessita de um GUI para poder ter uma maior aceitação por parte dos utilizadores DM.

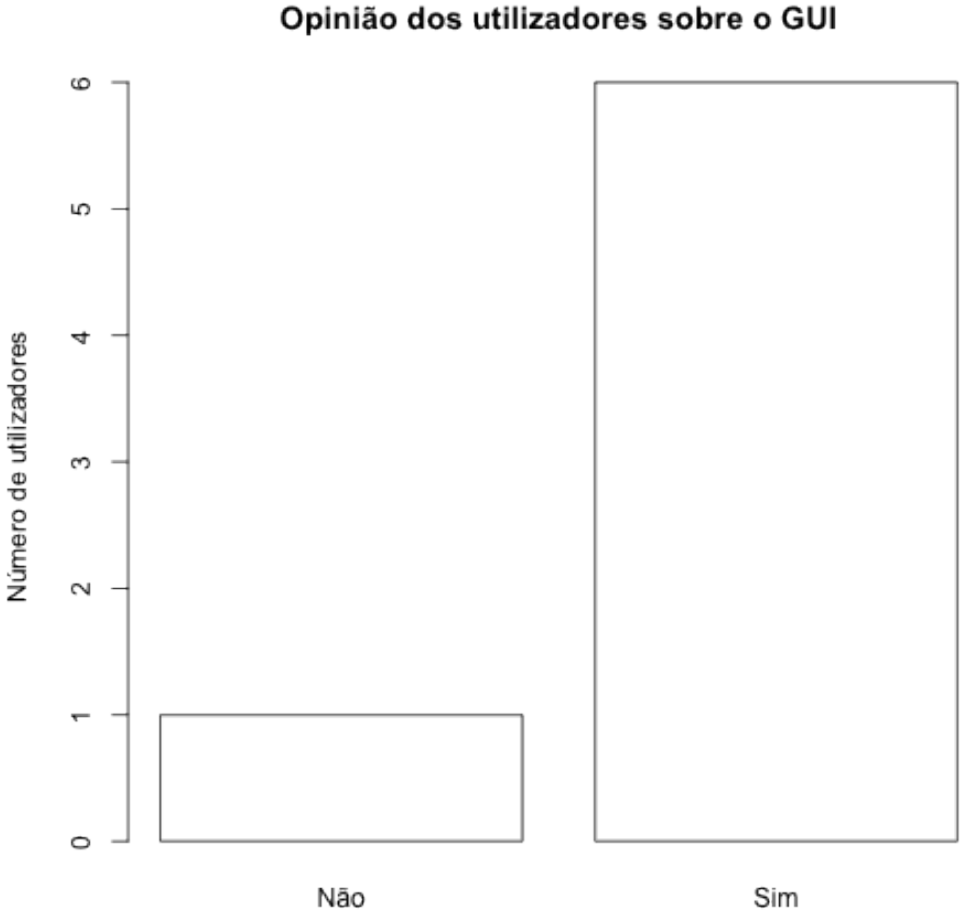


Figura 17. Opinião sobre a criação de um GUI

4.2. Implementação

O desenvolvimento do Jrminer [7] foi sustentado num conjunto de bibliotecas já existentes no mundo *open-source* das tecnologias de DM. Como já foi referido em capítulos anteriores, a interface foi construída para interagir com a tecnologia R, mais concretamente o *package* rminer. A sua construção foi efetuada na linguagem Java, via o *Integrated Development Environment* (IDE) Netbeans, pois, este é muito completo e permite criar aplicações gráficas de uma maneira acelerada e profissional [Netbeans.org, 2011]. A escolha da linguagem Java tem sobretudo a ver com a portabilidade, já que esta linguagem corre em diversas plataformas (e.g., *Windows*, *Mac, Linux*), sendo que também inclui um conjunto de bibliotecas que facilitam o desenvolvimento de aplicações gráficas.

De seguida, procedeu-se à escolha das bibliotecas que permitissem tanto conectar o Java ao R, como para facilitar a construção dos gráficos. A escolha recaiu sobre os *packages* rJava e o JavaGD, devido ao seu conjunto de funções coerentes e documentação, apesar de existirem outras bibliotecas (e.g, SJava, RCaller).

Escolhidas as tecnologias, procedeu-se à sua instalação. A primeira ferramenta a instalar foi o R. Para isso, efetuou-se o download em http://www.r-project.org, versão 2.12.2, sendo esta versão a mais atual na altura da preparação do ambiente de desenvolvimento. Em seguida, abriu-se o ambiente R e procedeu-se à instalação dos Bibliotecas rJava e JavaGD, com recurso ao comando *install.packages("nome da Biblioteca")*.

De seguida, procedeu-se à instalação do IDE Netbeans 7.0. Mais uma vez, esta era a versão mais atual da ferramenta na altura da sua instalação. Posteriormente, abriu-se a aplicação Netbeans e importaram-se as bibliotecas JRI.jar e o JavaGD.jar, para efetuar testes de comunicação entre a linguagem Java e a ferramenta R. Esses testes foram efetuados com recurso a pequenos algoritmos Java. Por exemplo, no caso do JRI.jar usou-se o código da Figura 19. Este foi o

[7] Nome oficial da aplicação. Download disponível em http://jrminer.pt.vu

primeiro teste que se efetuou entre o Java e o R, sendo que nas primeiras vezes tivemos um conjunto de erros, devido à não configuração das variáveis de ambiente. Após este teste ser positivo, passamos para os testes ao JavaGD, com o código da Figura 20.

Durante o desenvolvimento do Jrminer foi necessário instalar um conjunto de bibliotecas além das acima referidas, ver Secção 4.3.

O Jrminer tem um arquitetura de funcionamento semelhante à descrita na Figura 18, onde, o JRI.jar, é a *framework* de comunicação entre o Java e o R. Além dessa comunicação, o JRI.jar funciona como uma camada que permite ao utilizador do Jrminer uma abstração do R. Isto significa que, o utilizador, após uma pequena configuração inicial, não precisará de abrir R para utilizar o Jrminer.

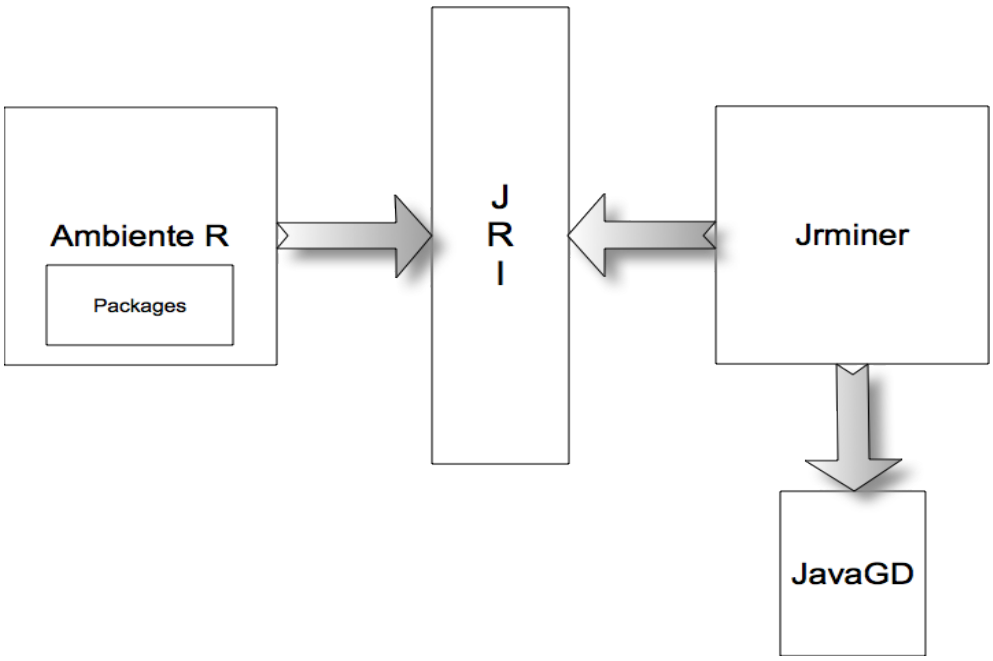


Figura 18. Arquitetura de alto nível de funcionamento Jrminer.

```
import org.rosuda.JRI.Rengine;
import org.rosuda.JRI.RMainLoopCallbacks;
public class ConnectR implements RMainLoopCallbacks {
   private static Rengine engine;
   public ConnectR() {
      startR();
   }
   public void shutdownR() {
      engine.end();
   }
   public Boolean startR() {
      engine = new Rengine(new String[]{"--vanilla"}, false, this);
      if (!engine.waitForR()) {
         return false;
      }
      return true;
   }
   (...)
         }
```

Figura 19. Exemplo do código de ligação do R ao Java

```
import javax.swing.JFrame;
import org.rosuda.JRI.Rengine;
import org.rosuda.javaGD.GDCanvas;
import org.rosuda.javaGD.GDInterface;
public class MyJavaGD1 extends GDInterface {
   public JFrame f;
   @Override
   public void gdOpen(double w, double h) {
      f = new JFrame("JavaGD");
      c = new GDCanvas(w, h);
      f.add((GDCanvas) c);
      f.pack();
      f.setVisible(true);
      f.setTitle("R plot");
      f.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
   }
   public static void main(String[] args) {
      Rengine re = new Rengine(new String[]{"--vanilla"}, false, null);
      re.eval("library(JavaGD)");
      re.eval("Sys.putenv('JAVAGD_CLASS_NAME'='rminer/MyJavaGD1')");
      re.eval("JavaGD()");
      re.eval("plot(c(1,5,3,8,5), type='l', col=2)");
      re.end();
   }

}
```

Figura 20. Exemplo do código de utilizção de gráficos R em Java

4.3. Bibliotecas necessárias

Em seguida apresentamos as bibliotecas necessárias para o funcionamento do Jrminer. São elas:

- **rJava:** é uma interface que providencia uma ponte de baixo nível entre o R e o Java. Ela permite criar objetos, chamar métodos ou aceder a campos de um objeto Java através do R. Em sentido contrário existe o JRI que possibilita chamar o R através de uma aplicação Java. Neste momento o JRI faz parte da biblioteca rJava, contudo pode ser usado de forma separada, especialmente quando em desenvolvimento de aplicações.
- **JavaGD:** é uma biblioteca que que permite criar gráficos R em uma classe Java. A implementação *default* fornece um objeto da classe Canvas, que pode ser usada em qualquer aplicação Java.
- **rminer:** facilita o uso de algoritmos DM em objetivos de classificação e regressão, com recurso a um conjunto de funções simples e coerentes. Pode ser usado um conjunto alargado de algoritmos, mas esta biblioteca está mais adaptada para lidar com RNAs e MVSs.
- **scatterplot3d:** permite criar um gráfico em 3 dimensões. Este facto é importante, pois permite elaborar gráficos com três variáveis.
- **randomForest:** implementa o algoritmo RF Breiman's para classificação e regressão.
- **mda:** biblioteca necessária para usar o algoritmo *Additive Spline Model*.
- **MASS:** permite aplicar estatística moderna a um determinado conjunto de dados. Esta biblioteca implementa os algoritmos LDA e QDA.
- **foreign:** permite aplicar funções de leitura/escrita de ficheiro (e.g., ARFF - ficheiros Weka -).

Estas foram as bibliotecas instaladas/usadas para o funcionamento/construção do Jrminer. Contudo, além das acima descritas, existe um grande conjunto que são secundárias, ou seja, são carregadas automaticamente pelas outras bibliotecas (e.g, rpart, kknn, nnet).

4.4. Instalação

O Jrminer é um pouco diferente das restantes aplicações do mesmo tipo existentes no mundo R. Ao contrário, por exemplo, do Rattle, o Jrminer funciona sem ser necessário ter o ambiente R aberto, permitindo assim ao utilizador uma abstração do mundo R, uma utilização mais reduzida da memória RAM e também uma melhoria de performance no uso da aplicação.

O GUI desenvolvido encontra-se disponível em: o http://jrminer.pt.vu. Em seguida, será explicado todo o processo necessário à instalação funcionamento da aplicação:

- Se não tiver instalado, instalar a ferramenta R disponível em http://www.r-project.org. Selecionar o servidor CRAN e escolher o R conforme o sistema operativo que usa;
- Se não tiver instalado, instalar o ambiente Java *(Java Run Environment)* disponível em http://www.java.com/en/download/manual.jsp;
- De seguida, é necessário instalar o *Biblioteca* rJava. Esta é a única interação do utilizador com a ferramenta R. Para isso, utilizar o comando: *install.packages*("rJava");
- Editar o ficheiro wizard.jar para configurar as variáveis de ambiente necessárias ao funcionamento da aplicação. Este *wizard* cria um ficheiro .bat ou .sh, conforme o sistema operativo.

- Por fim, o utilizador deve correr o ficheiro criado pelo wizard.jar. As restantes bibliotecas necessárias serão instaladas automaticamente na primeira utilização.

O Jrminer é uma aplicação multiplataforma e foi testado em diferentes ambientes, isto é, foram efetuados testes de funcionamento nos ambientes *Windows* XP/7, no *Mac OS* e em diferentes versões do R (2.12, 2.13).

4.5. Funcionalidades disponíveis

O Jrminer é uma aplicação que está estruturada de maneira a que o seu funcionamento siga as várias etapas da metodologia CRISP-DM. De seguida, descreve-se todo o conjunto de funcionalidades apresentadas pelo Jrminer.

Ao abrir a aplicação, é logo apresentada a tabulação *Data Understanding* (fase 2 do CRISP-DM). Aqui, o utilizador poderá efetuar duas tarefas, carregamento do ficheiro de dados (*Import Data*) e exploração do dados (*Explore*), ver Figura 21 e 22. No *Import Data* é possível carregar ficheiros de três formatos, CSV, TXT, ARFF. Conforme a escolha, é necessário a configuração dos parâmetros de identificação, separação dos dados (*Separator*), identificação do caráter decimal (*Decimal*), se o nome das variáveis vêm na primeira linha (*Header*), e o caráter que identifica os valores omissos (*Missing Values*). Quando o ficheiro estiver carregado o utilizador poderá descartar as variáveis que entender e escolher a variável de saída (*Target*). Conforme a saída, o sistema automaticamente selecionará o objetivo de DM mais apropriado.

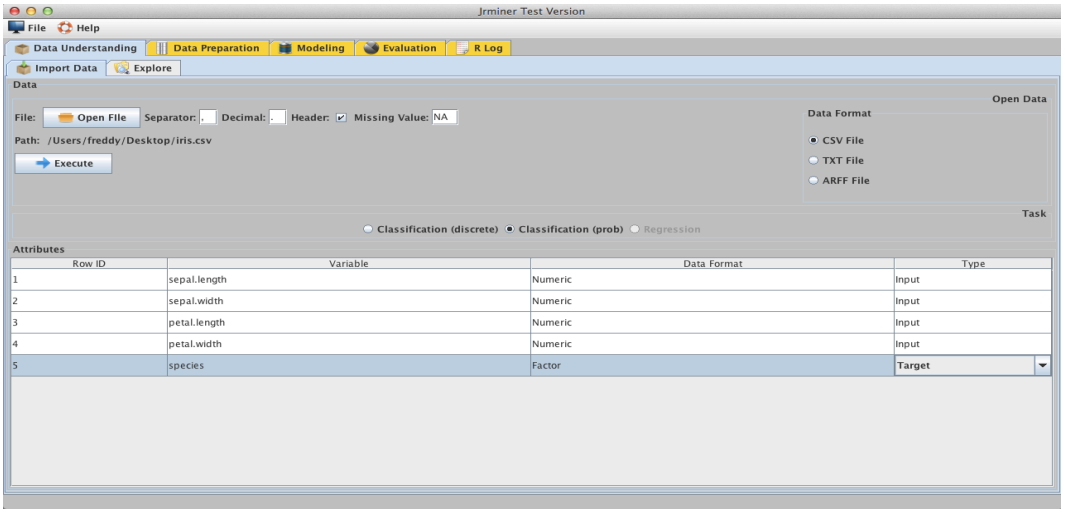


Figura 21. Exemplo do *Import Data* do Jrminer

Na tabulação *explore* é possível fazer uma exploração dos dados. O utilizador tem ao seu dispor funções que permitem ver os valores omissos (*Missing Values*), ver a estrutura dos atributos (*Structure*), e aceder a um sumário dos dados (*Summary*) (i.e., ver os mínimos, máximos, medias, entre outros, de cada atributo). Além destas funções existe a possibilidade de utilizar gráficos para uma melhor visualização, ver Figura 23. É possível usar *Histogram, Box Plot, Scatter, Density,* entre outros. Estes gráficos podem ser configurados com recurso a alguns atributos, como, *Color, Name, X/Y/Z Lab.*

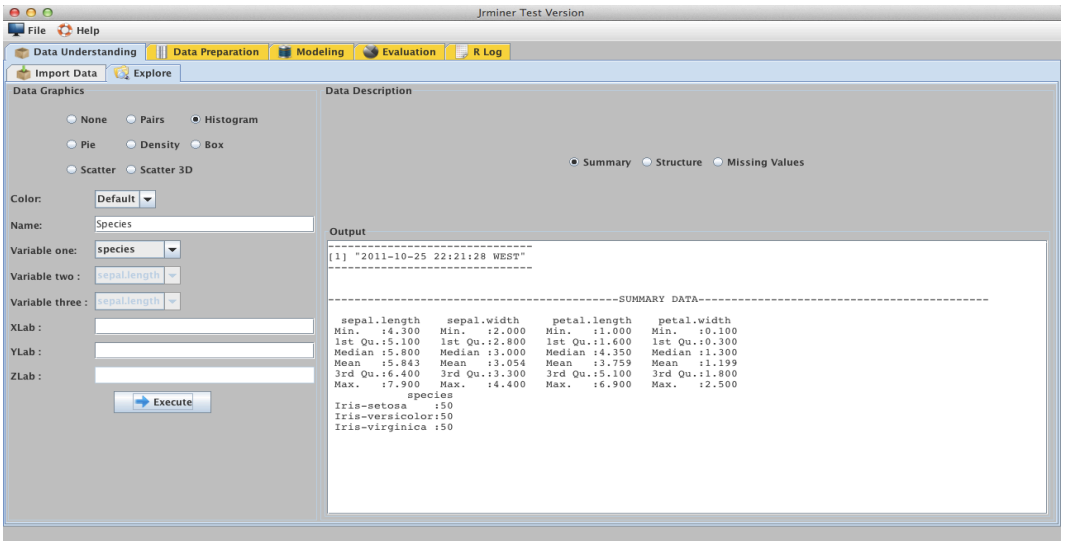


Figura 22. Exemplo do *Explore* do Jrminer

Figura 23. Exemplo de um gráfico *Histogram* no Jrminer

Na tabulação *Data Preparation* (fase 3 do CRISP-DM) o utilizador poderá efetuar um conjunto de transformações nos dados, ver Figura 24. Entre as várias funcionalidades de transformações temos o *Handling Missing Data,* onde é possível substituir os valores omissos pela Moda, Mediana, Média, Zero, *Hot Deck* (i.e., procura do exemplo mais similar), *Other* (i.e., valor à escolha do utilizador), ou então eliminar todos os valores nulos existentes. Também existe a possibilidade de aplicar um escalonamento (*Rescale*) a atributos numéricos. Entre as diversas funções temos o *Log(x+1), Recenter, Scale[0-1], -Median/MAD.* Além disso o utilizador também poderá aplicar outras transformações, como, *Delevels* (i.e., substituir um determinado nível de um ou mais fatores por outro), *Leveling* (i.e., converter um atributo numérico em um fator, através da definição de níveis), ou então *Deleted Selected* (i.e., apagar um determinado atributo). Neste menu de opções também é possível editar os dados ou exportá-los para o formato CSV ou TXT.

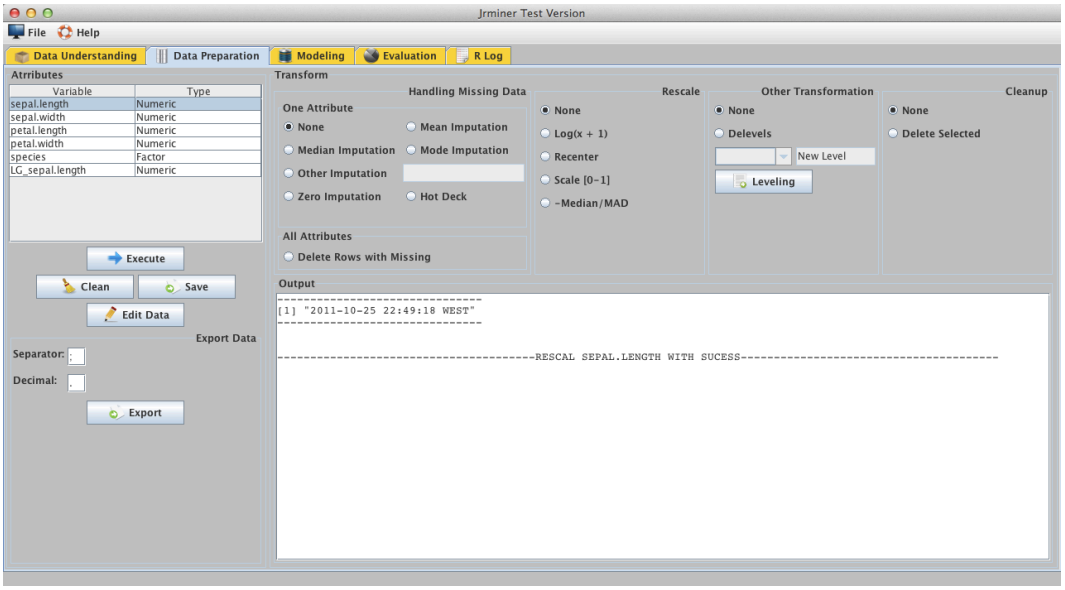


Figura 24. Exemplo do Data Preparation do Jrminer

Na tabulação *Modeling* (fase 4 do CRISP-DM) o utilizador poderá criar um modelo DM, ver Figura 25, através dos dados carregados no *Data Preparation -> Import Data.* Entre as várias funcionalidades existentes nesta terceira tabulação do menu principal do Jrminer temos os *Algorithms*, onde o utilizador poderá escolher, num vasto leque, o algoritmo que pretende aplicar na construção do seu modelo (e.g., *Neural Networks*, *Support Vetor Machines*, *Decision Trees*, entre outros). Conforme a escolha do algoritmo, é possível a configuração de determinados parâmetros (e.g., *Search, vmethods, vpar,* entre outros). Uma das configurações importantes é a escolha da função *fit*, onde só existe a possibilidade de correr e testar o modelo uma vez, e o *mining*, onde é possível definir o número de vezes (*Runs*) que o modelo será corrido e testado. De referir que a função *mining* do rminer executa diversas modelações (*fit*) e previsões (*predict*). Além disso, o utilizador poderá configurar o conjunto de dados de teste e treino, nos parâmetros do *Validation Method,* e as diretorias onde as previsões serão guardadas, nos campos do *Save Directory.*

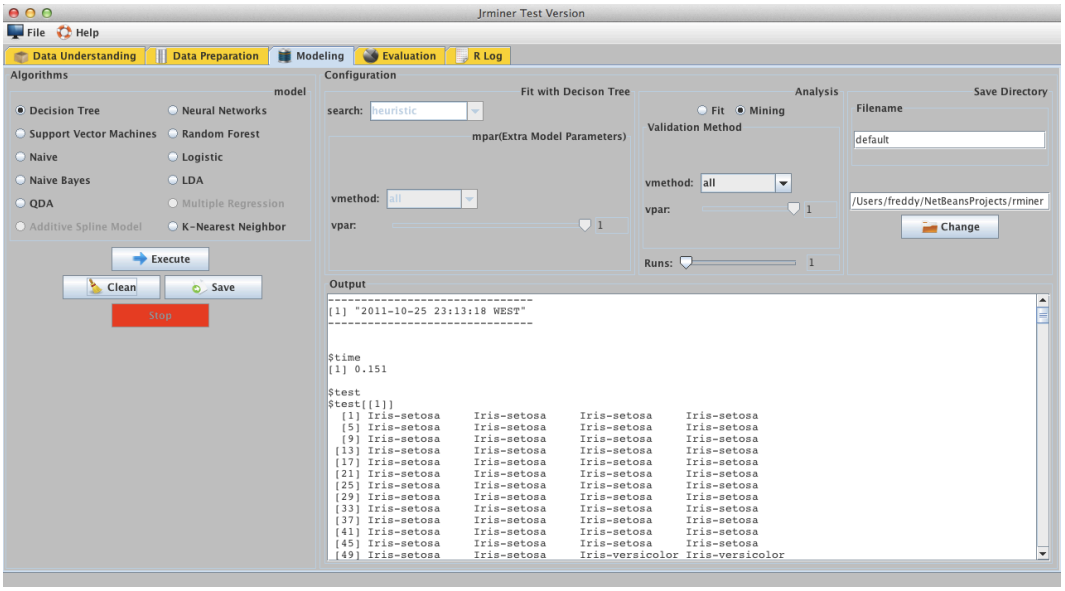


Figura 25. Exemplo do Modeling do Jrminer

No conjunto de funcionalidades oferecidas pela tabulação *Evaluation* (fase 5 do CRISP-DM), encontra-se a possibilidade de estudar os resultados com recurso a gráficos e métricas, ver Figura 26 e 27. Esta tabulação do Jrminer é especial, pois é possível fechar a aplicação e estudar os resultados mais tarde. Para isso, basta carregar o ficheiro no *Load File*. O Jrminer também possibilita comparar dois modelos semelhantes, e estudá-los. Contudo, só é possível em análises via a função *mining*.

Em relação aos gráficos/métricas, estes serão selecionados conforme o objetivo de DM (e.g., em objetivos de Regressão não é possível usar Curva ROC, e em objetivos de Classificação não é possível usar a Curva REC). Este fato é controlado automaticamente pelo sistema, ou seja, só estarão disponíveis para o utilizador os gráficos/métricas permitidos para aquele objetivo. Muitos gráficos/métricas podem ser configurados com recurso a parâmetros, como, *Target Class* (i.e, identifica a classe para o qual o gráfico é construído. -1 é o valor default), *Auxiliar Value,* e *Label.*

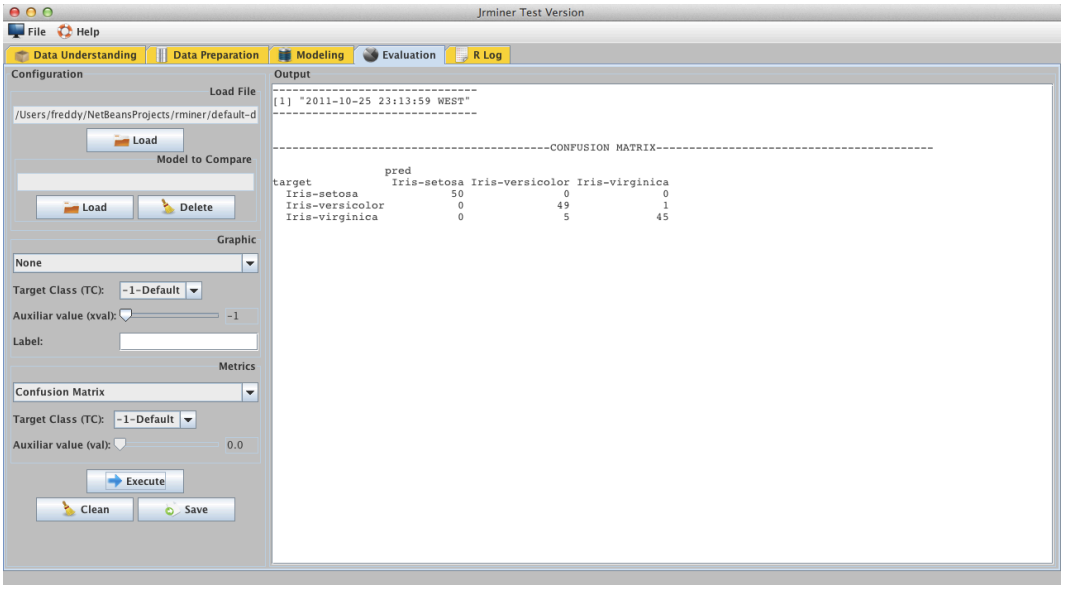


Figura 26. Exemplo do Evaluation do Jrminer

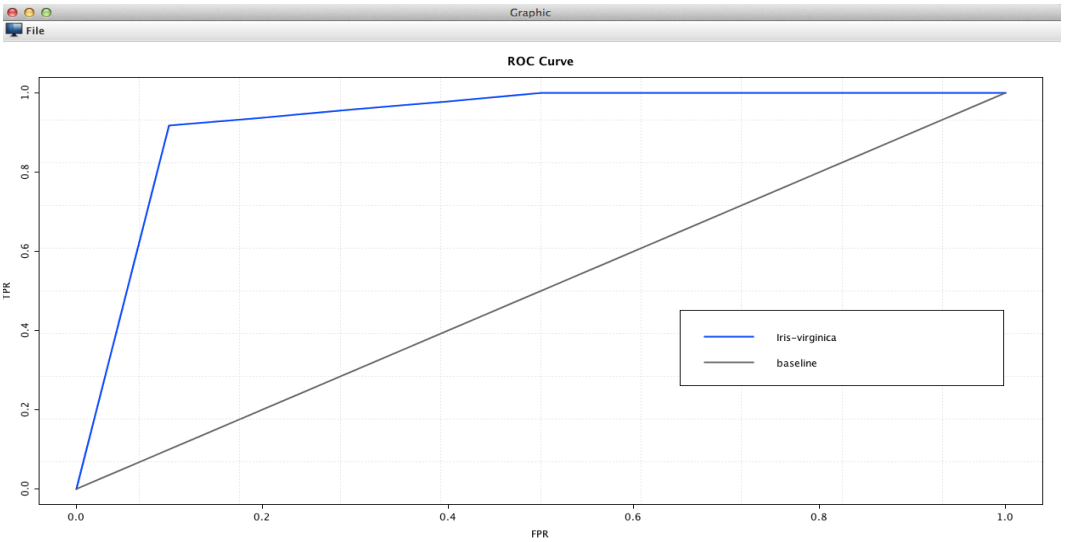


Figura 27. Exemplo de um curva ROC no Jrminer

O Jrminer ainda possibilita o registo dos acontecimentos, na forma de linguagem R, na tabulação *R Log*. Assim, um utilizador poderá ter uma noção clara de quais os comandos R/rminer que estão a ser executados. Além disso é possível exportar os comandos do *R Log* para um ficheiro TXT. Este *R Log* assume-se assim como uma funcionalidade que facilita a transição de uma aprendizagem inicial via a aplicação Jrminer para um uso intensivo e posterior do R/rminer via consola. Ainda, conforme é visível na Figura 28, é possível a criação de uma *Working Path* (i.e., diretório *default* onde serão registados todos os ficheiros).

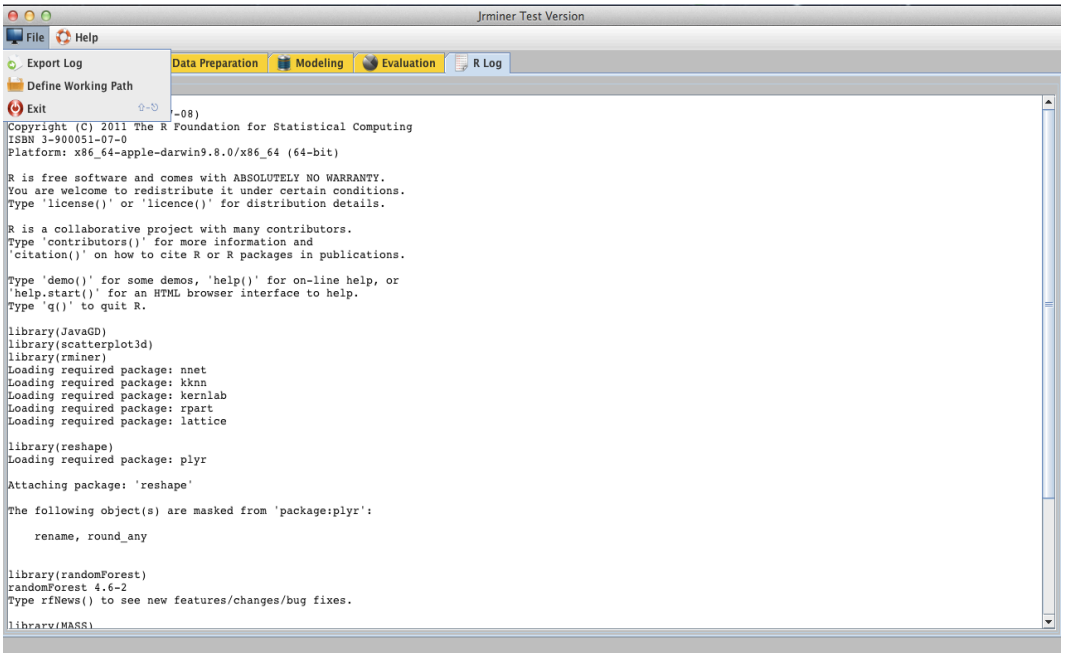


Figura 28. Exemplo do R Log do Jrminer

4.5.1. Exemplo de funcionamento

Nesta subsecção é possível observar um exemplo do funcionamento do Jrminer com recurso a um *dataset* retirado do repositório *UC Irvine Machine Learning Repository* (UCI)[8]. Entre as centenas de *datasets* existentes, escolhemos o iris, pois é o mais popular no repositório. O iris apresenta 150 registos com informação, como, largura/comprimento das sépalas e largura/comprimento das pétalas, das plantas *Setosa*, *Versicolor* e *Virginica*. É importante referir que existe um vídeo disponível no sítio: http://www.youtube.com/watch?v=N9YdfP7vehw que mostra as funcionalidades do Jrminer com recurso a um *dataset*, contudo é importante documentar um exemplo de funcionamento neste trabalho.

Para utilizar o Jrminer foi necessário, em primeiro lugar, carregar um ficheiro com os dados, ver Figura 21. Para isso, seguiram-se os seguintes passos: Escolheu-se o formato CSV; Inseriu-se o "," como caráter separador dos dados; Selecionou-se o ficheiro iris.csv através do botão *Open File*; Procedeu-se ao

[8] http://archive.ics.uci.edu/ml/

carregamento do ficheiro com recurso ao botão *Execute*. Após completar os passos anteriores, uma tabela foi mostrada com os dados carregados. Por fim escolheu-se o atributo *Species* como *Target*.

Numa segunda fase, procedeu-se à exploração dos dados. Para isso escolheu-se a função *Summary*, ver Figura 22. Além disso, escolheu-se também o gráfico *Histogram*, ver Figura 23, para o atributo *Species*. Esse gráfico mostra-nos que todas as classes (i.e., Iris-setosa, Iris-versicolor, Iris-virginica) estão distribuídas igualmente (i.e., 50 casos para cada um).

Numa terceira fase, procedeu-se à transformação dos dados. O iris é um *dataset* em que os seus dados já vêm tratados, logo numa situação natural, não seria necessário utilizar esta etapa. Contudo, uma vez que estamos a documentar um exemplo de funcionamento, vamos aplicar a função *Rescale Log(x+1)* ao atributo *Sepal.Length*. Para isso, procedeu-se à escolha do atributo, seguido da escolha da função. Após isso, é só executar, e o resultado aparece no *Output*. Um novo atributo, *LG_Sepal.Length*, é criado e adicionado à tabela *Atributes.*

Agora entramos numa das fases mais importantes da demonstração do Jrminer, a modelação dos dados. Em primeiro lugar escolhemos a análise *mining*, seguido do algoritmo *Decision Tree*. Os algoritmos do Jrminer já vêm com uma configuração *default*, logo, após a sua escolha, só foi necessário executar. Um modelo foi treinado e testado, sendo que os resultados obtidos aparecem no *Output* (ver Figura 25).

Entramos na última fase de demonstração. Nesta fase é possível avaliar o modelo criado. Assim vamos apresentar uma métrica (Matriz Confusão) e um gráfico (Curva ROC). Em relação à métrica basta selecioná-la na *Combo Box* referente às métricas (ver Figura 26) e executar. O resultado da métrica aparece no *Output.* Como podemos observar na Figura 26, este modelo tem uma elevada capacidade de previsão, pois apenas houve seis classificações erradas. Em relação ao gráfico, basta selecioná-lo na *Combo Box* referente aos gráficos (ver Figura 26) e executar. A Curva ROC também mostra que o modelo criado é de qualidade, pois apresenta pontos de elevado *true positive rate* e baixo *false positive rate* (ver

Figura 27). Contudo, há que refletir que nesta dissertação optou-se pela validação "all" do rminer, que treina e testa o modelo com todos os dados. Num cenário mais realista, poder-se-ia escolher uma validação cruzada ("kfold").

4.6. Opinião dos utilizadores sobre o Jrminer

Após o desenvolvimento do Jrminer, a fase final deste trabalho foi relativa a uma recolha da opinião dos utilizadores do rminer. Para isso, foi construído um novo questionário com 10 questões (este encontra-se disponível em: http://www.surveymonkey.com/s/6HPGB7B). Mais uma vez, pretendeu-se construir um questionário pequeno devido às limitações impostas pela ferramenta *web* para utilizadores não pagos. De seguida, pediu-se aos utilizadores, via email, para testarem a aplicação e posteriormente preencher o questionário. Esse questionário foi enviado a uma lista de 42 utilizadores. De notar que este segundo questionário foi criado meses depois do primeiro, sendo que durante esse tempo aumentou a lista de emails de utilizadores conhecidos do rminer. Contudo, apesar de um envio superior de pedidos de preenchimento, apenas se obtive aproximadamente 21,4% de respostas (total de 9 questionários preenchidos). De seguida, apresentam-se um resumo dos resultados obtidos para estas 9 respostas. Importa novamente referir que dado o número reduzido de respostas, não garantimos que estas refletem (com rigor estatístico), as opiniões dos utilizadores do rminer.

Para a maior parte das questões, cada utilizador tinha que escolher uma de entre 5 opções, segundo uma escala de *Likert*: 1 – Discordo Totalmente (DT), 2 – Discordo (D), 3 – Sem Opinião (SO), 4 – Concordo (C) e 5 – Concordo Totalmente (CT). Em seguida apresenta-se todas as questões efetuadas:

1. Qual a sua idade?
2. Quais as suas habilitações?
3. Participou no primeiro estudo sobre o R/rminer?
4. Sou um *expert* na utilização da ferramenta R.
5. Sou um *expert* na utilização do *package* rminer.

6. Acima de tudo acredito que o GUI Jrminer é uma ferramenta interessante para o *package* rminer.
7. O GUI Jrminer ajuda utilizadores R não especializados a aplicar a ferramenta R/rminer em objetivos de DM, como, Classificação e Regressão.
8. O Jrminer é mais fácil de aprender do que a ferramenta R/rminer.
9. Por favor, reporte erros e bugs.
10. Por favor, insira sugestões para melhorar o Jrminer.

Algumas desta questões não têm efeitos estatísticos (e.g., 9 e 10), uma vez que, apenas foram inseridas no questionário para ter uma opinião qualitativa dos utilizadores sobre o Jrminer. Além disso, uma vez que não se podia garantir que os utilizadores que responderam ao primeiro questionário respondessem ao segundo, foi necessário criar três perguntas para avaliar o perfil do utilizador (e.g., 1, 2 e 3). Os resultados obtidos são sumarizados na Tabela 3:

Tabela 3. Frequência de respostas ao questionário sobre o Jrminer

Questão	DT	D	SO	C	CT
4		5	1	2	1
5		4	1	3	1
6			2	**5**	**2**
7			1	**6**	**2**
8			2	**6**	**1**
Total	**0**	**9**	**7**	**22**	**7**

No geral, a grande maioria das respostas situa-se no lado direito da tabela, logo podemos deduzir que o resultado deste questionário foi bastante positivo. Olhando com mais atenção para a Tabela 3, pode-se verificar que as questões

referentes à experiência no R/rminer por parte dos utilizadores (i.e., questão 4 e 5) é um pouco baixa, contando com 9 respostas negativas num universo de 18 possíveis, ou seja, 50% dos inquiridos não se considera um *expert* na ferramenta R/rminer. Os restantes 50% distribuem-se pelas áreas positivas/neutras do questionário. Estas questões foram as únicas com respostas negativas, e mostram claramente a dificuldade de alguns utilizadores perante ferramentas em que é necessário conhecimentos de programação.

Nas restantes questões, referentes ao Jrminer, pode-se observar claramente um resultado muito positivo. No âmbito deste trabalho essas três questões (i.e., questões 6,7 e 8) são as mais importantes, pois são elas que vão medir, embora sem rigor científico, a satisfação dos utilizadores perante o GUI Jrminer. Assim, com 22 respostas positivas (Concordo), num universo de 27 possíveis, os utilizadores mostram claramente que o Jrminer é uma ferramenta interessante, fácil de aprender, e que vai permitir aos utilizadores, não especializados, aplicar o *package* rminer em objetivos de DM. Esta resposta é a que mais se evidencia nos nossos resultados e vem de encontro àquilo que era julgado pelo orientador deste trabalho quando definiu este tema de dissertação. Para dominar a biblioteca rminer é preciso conhecimentos de programação R, com uso de comandos via consola, sendo que em geral utilizadores informáticos não especializados (e.g. Gestores, Engenheiros Civis) estão mais habituados a ferramentas gráficas. Além disso, ainda existem 7 respostas (Concordo totalmente), que permitem reforçar ainda mais a ideia da importância do Jrminer. Contudo, ainda existe uma pequena minoria que não tem opinião sobre o assunto. Como é uma amostra muito pequena, cerca de 1,89% do universo possível, não influência os resultados

Como já foi referido anteriormente, não foi possível garantir que os utilizadores que responderam ao primeiro questionário respondessem ao segundo, logo na Figura 29 pode-se observar que a maior parte dos inquiridos situa-se entre os 18 e 25 anos, mas também com mais de 36 anos. Além disso, também é possível verificar, através da Figura 30, que os inquiridos na sua maioria apresentam a habilitação de Mestrado. É de realçar que nestas respostas não aparece nenhum utilizador só com Licenciatura.

Apesar de terem sido obtidas poucas respostas (de 9 utilizadores), os resultados obtidos com este inquérito **validam (dentro do que foi possível realizar) o trabalho desenvolvido**. A aplicação gráfica Jrminer, apesar de não ser tão versátil, em termos de funcionalidades, como é o R/rminer em modo consola, é mais simples de utilizar e engloba as funcionalidades mais relevantes oferecidas pelo rminer. Tal é relevante, principalmente no contexto de ensino universitário e de investigação académica relacionado com o supervisor deste trabalho.

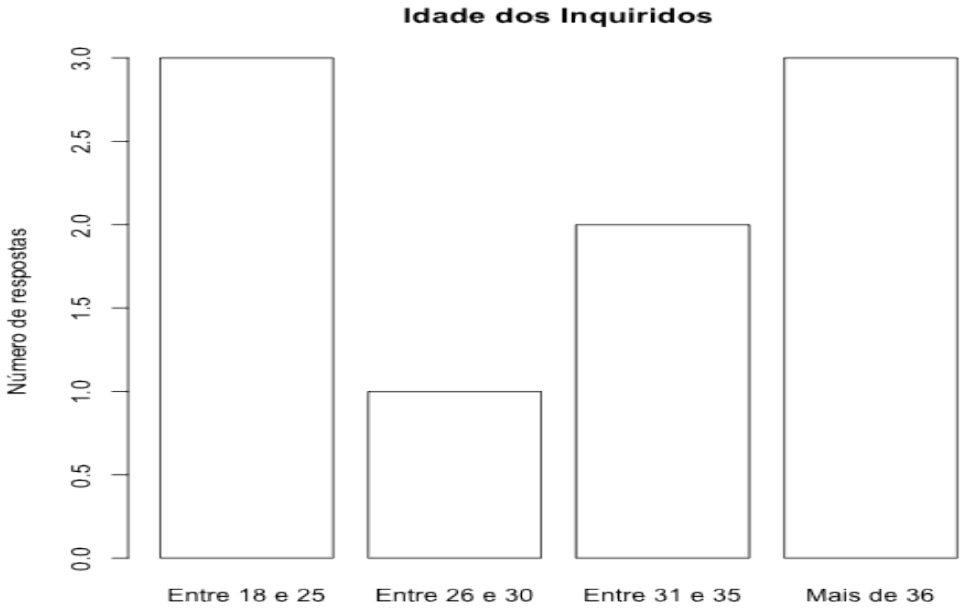


Figura 29. Idade dos inquiridos sobre o Jrminer

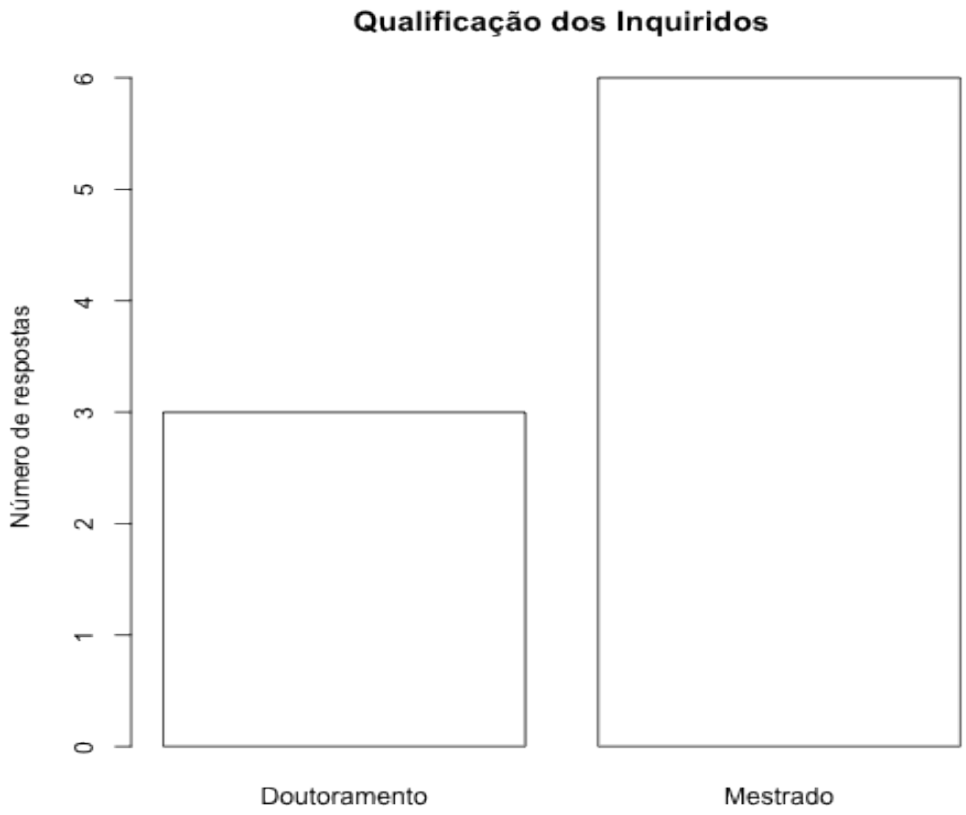


Figura 30. Qualificação dos inquiridos sobre o Jrminer

5. Conclusão

Este capítulo apresenta uma síntese do trabalho realizado, uma discussão do trabalho e uma secção sobre trabalho futuro.

5.1. Síntese

Estamos numa era em que as organizações sabem que a informação que produzem é fulcral para a sua sobrevivência, pois nela estão contidos conhecimentos que podem criar vantagens competitivas. Contudo, a extração desse conhecimento não é possível com técnicas de estatísticas simples. Para resolver esse problema desenvolveu-se a área DCBD/DM. É nesta área que são aplicadas técnicas provenientes de diferentes disciplinas, tais como, Inteligência Artificial, Matemática, Estatística e Base de dados, com o objetivo de extrair conhecimento de dados em bruto.

Nesta dissertação foram abordados dois objetivos DM: Classificação e Regressão. Ambos os objetivos estão inseridos na aprendizagem supervisionada, que permite descobrir uma relação entre os atributos de entrada e o atributo de saída. A relação descoberta é representada em modelos, e permite explicar fenómenos que estão ocultos nos dados. Assim, e possível criar modelos com recurso a algoritmos, como, RNAs, AD, MVSs, entre outros. Para facilitar o sucesso dos projetos DM, foram criadas metodologias. A das mais populares é a CRISP-DM, e possui a vantagem de ser neutra na escolha de tecnologias DM.

São as Tecnologias de Informação que permitem às organizações aplicar os conhecimentos existentes na área DM. O desempenho deste tipo de ferramentas é muito importante, sendo apresentado nesta dissertação uma lista de tópicos aos quais uma ferramenta deve responder. Atualmente, existem inúmeras ferramentas (e.g., Weka, RapidMiner, Knime, Rattle). Entre as mais usadas pelos analistas encontra-se o R. Contudo, o R não é orientado ao DM, mas é possível aumentar as suas funcionalidades através da instalações de bibliotecas (e.g., rminer).

O rminer foi desenvolvido por [Cortez, 2010] com o objetivo de facilitar a utilização de algoritmos DM na ferramenta R. Contudo, para utilizar esta biblioteca é necessário ter conhecimentos de um conjunto de comandos R. Este tipo de interface não corresponde àquilo que os utilizadores, em geral, estão mais habituados, pois estão mais familiarizados aos ambientes gráficos.

Nesta dissertação é apresentado o ambiente gráfico Jrminer que permite aos utilizadores o uso da biblioteca rminer sem grandes conhecimentos da linguagem R. Esta GUI é intuitiva, simples, com um design coerente, o que a torna, segundo inquéritos que foram efetuados, de mais fácil aprendizagem quando comparado à linha de comandos do rminer.

5.2. Discussão

Neste trabalho procedeu-se à criação de um ambiente gráfico que permitisse ao utilizador comum (não especializado) a utilização da biblioteca rminer em objetivos de Classificação e Regressão. O rminer no seu estado natural apresenta muitas dificuldades para o utilizador dito "comum", pois é necessário conhecimentos na linguagem R para utilizar a ferramenta. Ora, este facto faz com que a aplicação desenvolvida no âmbito desta dissertação seja de considerável importância. Além do Jrminer, existe um largo conjunto de ferramentas DM no mercado, sendo que os analistas estão a começar a preferir as ferramentas *open-source,* não só pelo baixo custo, mas também pelo facto de ser suportada por grandes comunidades. O Jrminer alia-se a este contexto, já que é baseado numa filosofia de *open-source.*

Para além da construção da aplicação foi necessário a opinião dos utilizadores do rminer sobre o Jrminer. Os resultados obtidos aqui, são satisfatórios, apesar da pequena amostra utilizada, sendo que quase a totalidade dos utilizadores analisados acredita que o Jrminer facilita a aprendizagem no uso do *package* rminer para a execução de tarefas DM na ferramenta R. Assim, pode-

se afirmar que a questão de investigação definida na secção 1.3 é respondida de modo positivo com este trabalho.

Dado os limites temporais associados a esta dissertação, existem limitações relativas ao trabalho que foi produzido. Embora a maior parte das funcionalidades do *package* rminer estejam disponíveis na aplicação Jrminer, existem certas configurações (e.g., seleção concreta do número de nodos intermédios de uma RNA) que não estão disponíveis. Por outro lado, a avaliação da ferramenta foi efectuada somente via inquéritos disponibilizados *on-line*, sendo que só foram analisadas 9 respostas. Contudo, há que referir que a aplicação Jrminer foi criada do "zero", ou seja, uma grande parte desta dissertação consistiu no desenvolvimento da ferramenta gráfica Jrminer

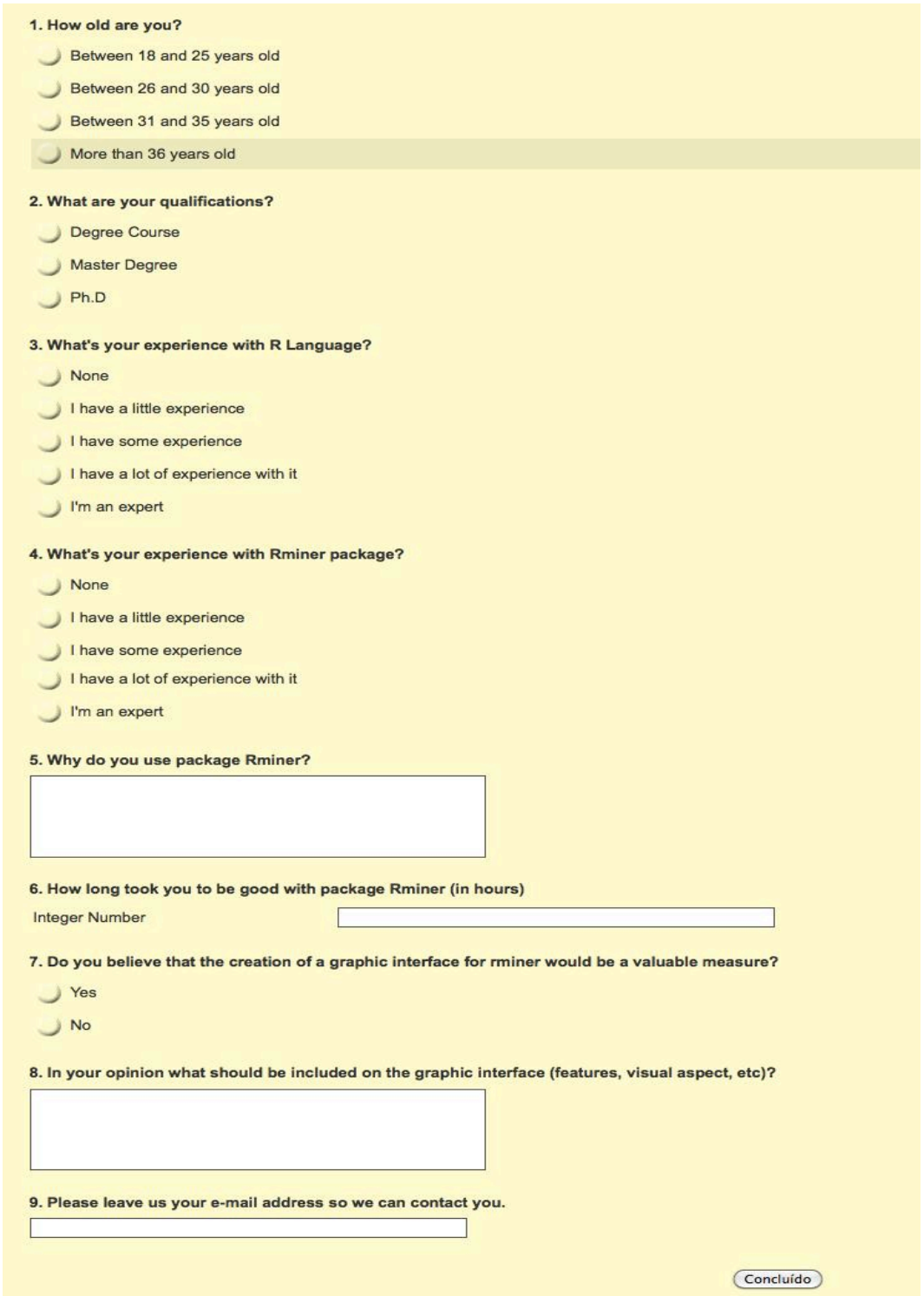
5.3. Trabalho Futuro

Este trabalho apresenta diversas perspetivas de trabalho futuro, tais como:

- Permitir carregamento de dados através de outros formatos de base de dados (e.g., MySQL, SQL Server, Access). Este processo permitirá ao utilizador uma versatilidade maior no carregamento, ou seja, não será necessário exportar os dados através do SGBD para um ficheiro;
- Fazer um estudo sobre o rminer e o Jrminer com uma amostra maior de utilizadores, de forma a validar com mais rigor o impacto da ferramenta Jrminer;
- Aplicar a ferramenta Jrminer em diversas aplicações do mundo real (e.g., análise de dados financeiros).

Anexo A

A.1 – Questionário para registar as opiniões dos utilizadores sobre o rminer



1. How old are you?

- Between 18 and 25 years old
- Between 26 and 30 years old
- Between 31 and 35 years old
- More than 36 years old

2. What are your qualifications?

- Degree Course
- Master Degree
- Ph.D

3. What's your experience with R Language?

- None
- I have a little experience
- I have some experience
- I have a lot of experience with it
- I'm an expert

4. What's your experience with Rminer package?

- None
- I have a little experience
- I have some experience
- I have a lot of experience with it
- I'm an expert

5. Why do you use package Rminer?

6. How long took you to be good with package Rminer (in hours)

Integer Number

7. Do you believe that the creation of a graphic interface for rminer would be a valuable measure?

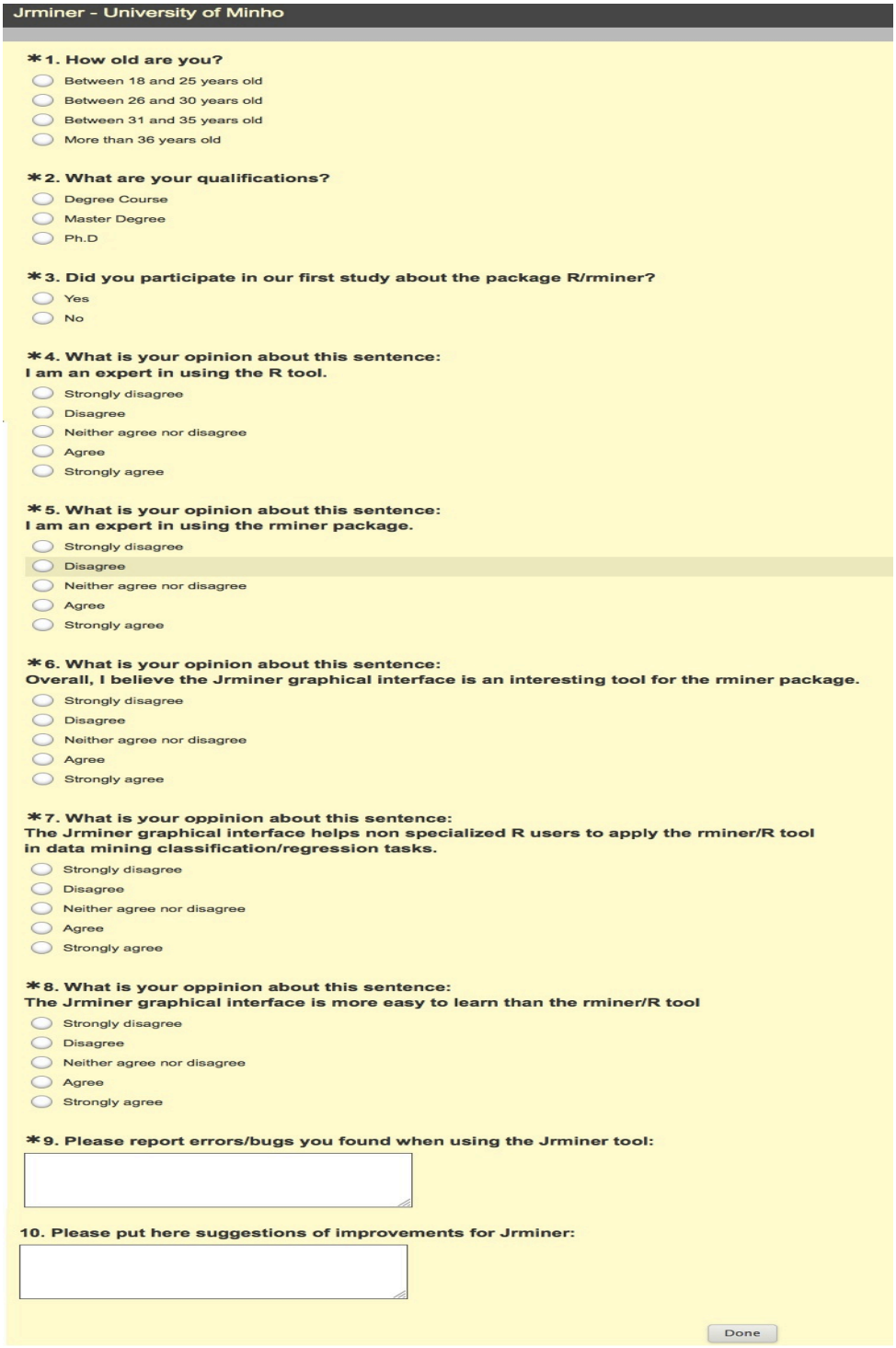
- Yes
- No

8. In your opinion what should be included on the graphic interface (features, visual aspect, etc)?

9. Please leave us your e-mail address so we can contact you.

Concluído

A.2 – Questionário para registar as opiniões dos utilizadores sobre o Jrminer



Jrminer - University of Minho

*1. How old are you?
- Between 18 and 25 years old
- Between 26 and 30 years old
- Between 31 and 35 years old
- More than 36 years old

*2. What are your qualifications?
- Degree Course
- Master Degree
- Ph.D

*3. Did you participate in our first study about the package R/rminer?
- Yes
- No

*4. What is your opinion about this sentence:
I am an expert in using the R tool.
- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

*5. What is your opinion about this sentence:
I am an expert in using the rminer package.
- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

*6. What is your opinion about this sentence:
Overall, I believe the Jrminer graphical interface is an interesting tool for the rminer package.
- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

*7. What is your oppinion about this sentence:
The Jrminer graphical interface helps non specialized R users to apply the rminer/R tool
in data mining classification/regression tasks.
- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

*8. What is your oppinion about this sentence:
The Jrminer graphical interface is more easy to learn than the rminer/R tool
- Strongly disagree
- Disagree
- Neither agree nor disagree
- Agree
- Strongly agree

*9. Please report errors/bugs you found when using the Jrminer tool:

10. Please put here suggestions of improvements for Jrminer:

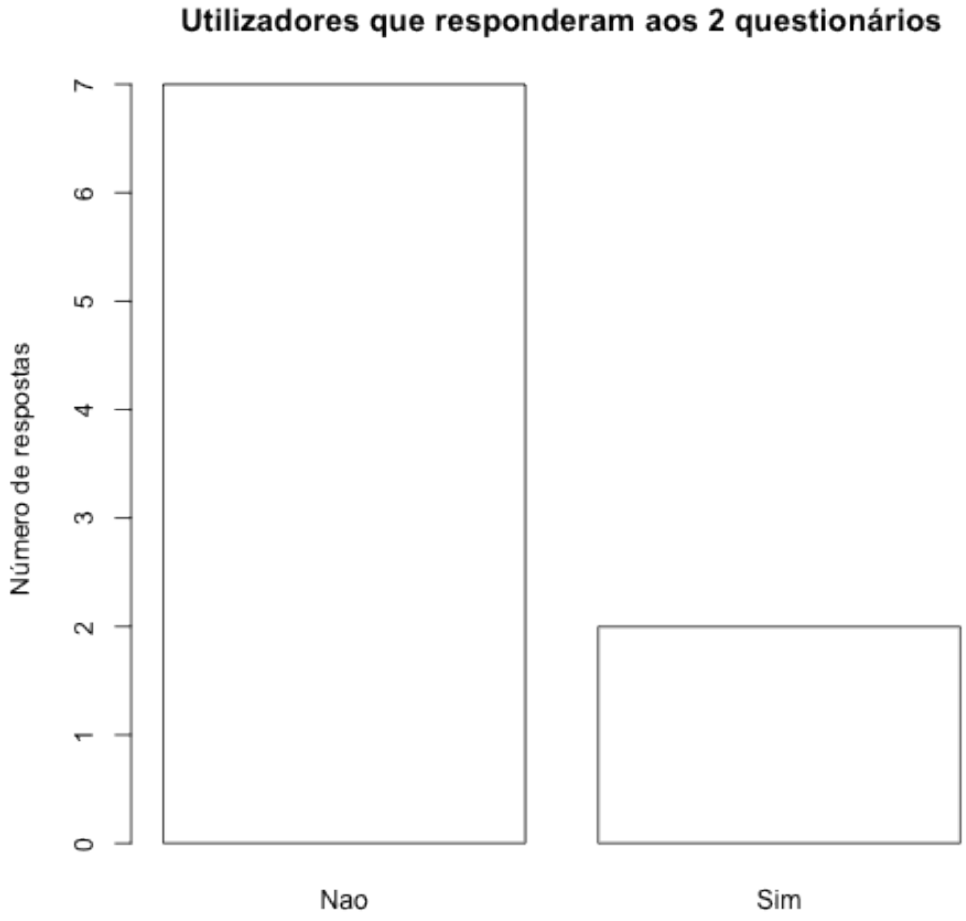
Done

Anexo B

B.1 – Logótipo do Jrminer

Anexo C

C.1 – Número de utilizadores que responderam aos 2 questionários



Bibliografia

[Aha et al., 1991]Aha, W., Kibler, D. & Albert, K. 1991. Instance-Based Learning Al- Gorithms. *Machine Learning* 6**,** 37–66.

[Berry e Linoff, 2000] Berry, J. & Linoff, G. 2000. *Mastering Data Mining: The Art and Science of Customer Relationship Management,* USA, Jonh Wiley & Sons, Inc.

[Breiman, 2001] Breiman, L. 2001. Random Forests. *Machine Learning,* 45**,** 5-32.

[Chambers, 2008] Chambers, J. M. 2008. *Software for Data Analysis - Programming with R*, Springer.

[Chapman et al., 2000] Chapman, P., Clinton, J., Randy, K., Thomas, K., Thomas, R., Colin, S. & Rudiger, W. 2000. Crisp-Dm 1.0: Step by Step Data Mining Guide. SPSS.

[Cortez, 2002] Cortez, P. 2002. *Modelos Inspirados Na Natureza Para a Previsão De Séries Temporais*. ph.D, Universidade do Minho.

[Cortez, 2010] Cortez, P. 2010. Data Mining with Neural Networks and Support Vector Machines Using the R/Rminer Tool. *In:* Perner, P. (ed.) *Advances in Data Mining -- Applications and Theoretical Aspects, 10th Industrial Conference on Data Mining*. Berlin, Germany: LNAI 6171, Springer.

[Cortez et al., 2009a] Cortez, P., Cerdeira, A., Almeida, F., Matos, T. & Reis, J. 2009a. Modeling Wine Preferences by Data Mining from Physicochemical Properties. *Decision Support Systems,* 47**,** 547-553.

[Cortez et al., 2009b] Cortez, P., Lopes, C., Sousa, P., Rocha, M. & Rio, M. 2009b. Symbiotic Data Mining for Personalized Spam Filtering. *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence (WI-09)***,** 149-156.

[Costa e Simões, 2008] Costa, E. & Simões, A. 2008. *Inteligência Artificial - Fundamentos E Aplicações*, FCA - Editora.

[Dayton, 1992] Dayton, C. M. 1992. Logist Regression Analysis. Maryland: University of Marylan.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI Magazine,* 17.

[Goebel e Gruenwald, 1999] Goebel, M. & Gruenwald, L. 1999. A Survey of Data Mining and Knowledge Discovery Software Tools. *ACM SIGKDD Explorations,* Vol. 1.

[Han e Kamber, 2006] Han, J. & Kamber, M. 2006. *Data Mining - Concepts and Techniques*, The Morgan Kaufmann.

[Haykin, 1999] Haykin, S. 1999. *Neural Networks - a Compreensive Foundation,* New Jersey, Prentice-Hall.

[Ihaka e Gentleman, 1996] Ihaka, R. & Gentleman, R. 1996. R: A Language for Data Analysis and Graphics. *Computational and Graphics Statics,* 5**,** 299-314.

[Maimon e Lior, 2010] Maimon, O. & Lior, R. 2010. *Data Mining and Knowledge Discovery Handbook,* New York - USA, Springer.

[Michalski, 1998] Michalski, R. S., Bratko, Ivan and Miroslav, Kubat 1998. Machine Learning and Data Mining Methods and Applications. England: John Wiley & Sons, Inc.

[Miller, 2006] Miller, S. 2006. *R You Ready for Open Source Statistics?* [Online]. http://www.information-management.com/news/1065015-1.html.

Netbeans.Org. 2011. Http://Netbeans.Org/Features/Index.Html [Online].

[Piatetsky-Shapiro, 2010] Piatetsky-Shapiro, G. 2010. Data Mining Tools Used Poll. http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html.

[Pontil e Verri, 1998] Pontil, M. & Verri, A. 1998. Properties of Support Vector Machines. *Neural Computation,* 10**,** 955-974.

[Quinlan, 1998] Quinlan, J. R. 1998. C4.5 Programs for Machine Learning. USA: Morgan Kaufmann Publishers, Inc.

[RexerAnalytics.com, 2010] Rexeranalytics.Com. 2010. *2010 Data Miner Survey* [Online].

[Rocha et al., 2007] Rocha, M., Cortez, P. & Neves, J. 2007. Evolution of Neural Networks for Classification and Regression. *Neurocomputing,* 70**,** 2809-2816.

[StatSoft-Inc, 2011] Statsoft-Inc 2011. Electronic Statistics Textbook. Tulsa, Ok: Statsoft.: http://www.statsoft.com/textbook/.

[Turban et al., 2010] Turban, E., Sharda, R. & Delen, D. 2010. *Decision Support and Business Intelligence Systems*, Pearson.

[Williams, 2009] Williams, G. J. 2009. Rattle: A Data Mining Gui for R. *The R Journal,* 1**,** 45-55.