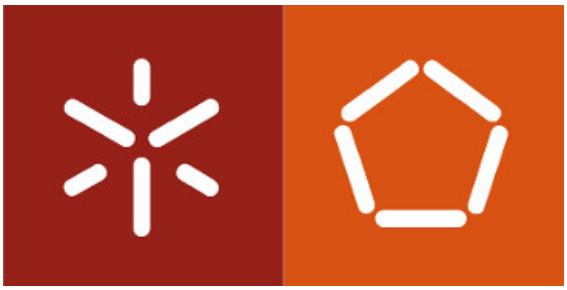

Universidade do Minho

Mestrado Integrado em Engenharia Informática

Engenharia do Conhecimento

Descoberta do Conhecimento

Resolução da Ficha de Exercícios 07

Gonçalo Pinto, A83732

João Diogo Mota, A80791

José Gonçalo Costa, PG42839

José Nuno Costa, A84829

7 de Maio de 2021

Conteúdo

Lista de Figuras

Capítulo 1

Introdução

No 2º semestre do 1º ano do Mestrado em Engenharia Informática da Universidade do Minho, existe uma unidade curricular enquadrada no perfil de Engenharia do Conhecimento denominada por Descoberta do Conhecimento, que tem como objectivo a introdução ao conceito de descoberta do conhecimento.

Nesta ficha pretende-se que cada grupo execute os processos de *Business Understanding, Data Understanding, Data Preparation, Modeling e Evaluation* enquadrados na metodologia do *CRISP-DM* que suporta processos de *Data Mining* num determinado caso de estudo. A metodologia utilizada de *Data Mining* foi baseada em regressão linear, uma abordagem de modelação estatística que calcula uma relação entre uma resposta escalar (ou variável dependente) e uma ou mais variáveis explicativas (ou variáveis independentes) e que depois usa essa relação para efectuar a previsão.

Para este efeito, foi fornecido um conjunto de dados pela equipa docente acerca de preços de casas e suas características, o objectivo deste trabalho é tentar determinar, através dessas características a sua influência no preço final de cada casa.

Capítulo 2

Parte 1

1. ***Que tipo de dados a regressão linear espera para todos os atributos? Qual o tipo de dados do atributo previsto quando este for calculado?***
   A regressão linear espera valores numéricos para todos os seus atributos de forma a entender a relação entre o *input* e o *output* destas variáveis.
   Dado que a regressão linear assume uma relação linear entre as variáveis de *input* e *output*, ou seja, que o *output* é calculado combinando as diferentes variáveis de *input*, o tipo de dados do atributo previsto quando este for calculado é, de igual forma, numérico.

2. ***Porque é que os intervalos de atributos são tão importantes ao realizar data mining através de regressão linear?***
   Ao realizar *Data Mining* através de regressão linear, um aspecto muito relevante a ter em conta é os intervalos de atributos, na medida que os intervalos para os atributos de *scoring* devem estar dentro dos intervalos dos atributos de treino para garantir previsões válidas.

3. ***O que são coeficientes de regressão linear? O que significa 'peso', neste contexto?***
   Os coeficientes de regressão linear representam a capacidade que um dado atributo tem na previsão do atributo designado *label* - este é o atributo alvo, aquilo que se pretende prever. Assim, neste contexto o peso indica quanto um atributo do conjunto de dados se relaciona com o atributo alvo, além disto, um atributo com pouco peso, significa que este não tenha grande influência no atributo que se quer prever.

4. ***Qual é a fórmula matemática de regressão linear e como é organizada?***
   A fórmula matemática de regressão linear é:

   $$y = mx + b \tag{2.1}$$

   Nesta equação, y representa o que se quer prever (*label*), m representa a variável independente, x representa o coeficiente do atributo e por fim, b representa uma constante.

5. ***Como é que resultados da regressão linear podem ser interpretados/avaliados?***
   A análise da regressão linear gera uma equação para descrever a relação entre um ou mais atributos e o atributo resposta. Depois de usar o *RapidMiner*, devemos interpretar os valores-p e o coeficiente de cada atributo que aparece na saída da análise da regressão linear. Estes valores permitem saber quais os atributos que melhor influenciam o atributo que se quer prever, neste caso, o atributo do tipo *label*.

Capítulo 3

Parte 2

1. ***Importe o dataset definido para a FE07, fe07-dataset.csv para o RapidMiner. Execute a componente de Data Understanding e DataPreparation.***
   A primeira etapa do processo de *Data Mining* passa pelo processo de *Data Understanding*, que tem como objectivo a familiarização com os dados. O objectivo deste trabalho é determinar qual o valor de mercado de uma determinada habitação, tendo em consideração diversas características destas. O conjunto de dados utilizado possui 10 atributos, sendo estes:

   - **house_sqft**, tamanho da casa, em *square foot*.
   - **num_of_bedrooms**, número de quartos que a casa têm.
   - **num_of_bathrooms**, número de casas de banho.
   - **year_built**, ano em que a casa foi construída.
   - **tax_assessed_value**, valor fiscal da casa.
   - **last_sold_price**, ultimo preço da casa.
   - **rate_per_sqfoot**, avaliação do *square foot*.
   - **city**, cidade onde a casa se localiza.
   - **home_type**, tipo da casa.
   - **school_rating_1to10**, classificação das escolas nas imediações da casa.

   A segunda etapa do processo de *Data Mining* passa pelo processo de *Data Preparation*, que diz respeito ao processamento dos dados de modo que os dados transformados estejam numa forma mais adequada para os algoritmos de *Data Mining*. O conjunto de dados fornecido apresenta 75 valores desconhecidos no atributo *school_rating_1to10*, assim recorrendo a ferramenta "*Replace Missing Values*"que o software utilizado, *Rapid Miner*, fornece foi substituído os valores em falta pela média deste atributo. Seguidamente como se pretende efectuar um modelo baseado na regressão linear que necessita apenas de atributos numéricos descartou-se o atributo *city*, uma vez que este valor não traz conhecimento dado que é sempre igual, e também o atributo *home_type* (tal como foi sugerido) porque se trata de um atributo nominal, assim com a ferramenta "*Select Attributes*"foi gerado um novo subconjunto de dados.

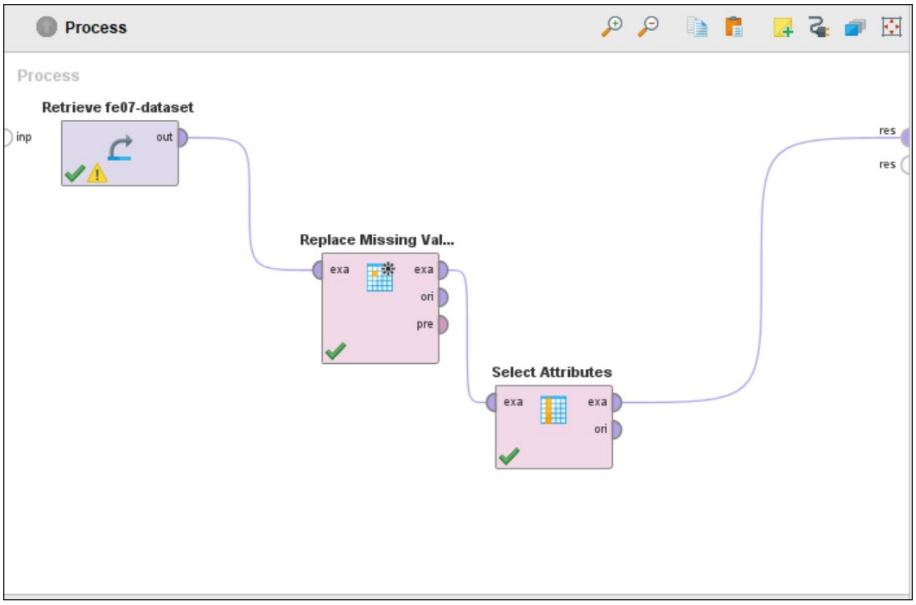

Figura 3.1: Processo de *Data Preparation*.

2. ***Repita os passos no RapidMiner tal como descritos nos slides da aula e após executar o seu modelo, na secção dos resultados, examine os coeficientes dos atributos e as previsões para os custos das casas.***
   A terceira etapa do processo de *Data Mining* passa pelo processo de *Modeling*, nesta fase pretende-se usar algoritmos de regressão linear disponíveis no *RapidMiner*. Na figura 3.2 abaixo apresenta-se o processo de modelação construído.

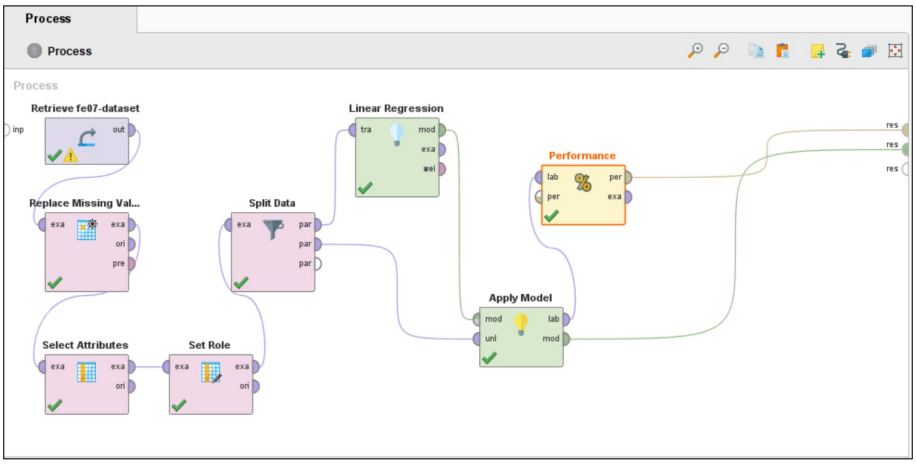

Figura 3.2: Processo de *Modeling*.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value ↓	Code
tax_assessed_value	-0.001	0.001	-0.017	0.982	-1.288	0.202	
num_of_bedrooms	8010.720	3272.382	0.064	0.789	2.448	0.017	**
num_of_bathrooms	-11579.229	3706.492	-0.087	0.660	-3.124	0.003	***
school_rating_1to10	-10315.272	2989.294	-0.103	0.709	-3.451	0.001	****
house_sqft	110.845	2.953	1.045	0.762	37.538	0	****
rate_per_sqfoot	2796.331	72.225	0.918	0.974	38.717	0	****
(Intercept)	-236900.703	12091.660	?	?	-19.592	0	****

Figura 3.3: Resultados da Regressão Linear no conjunto de dados fornecido.

(a) ***Execute a avaliação e documentação dos seus resultados.***
Tal como se pode observar na figura 3.3, existem dois atributos que apresentam o seu *p-value* a 0, sendo estes o atributo *rate_per_sqfoot* e *house_sqft*. Este valor pode indicar alguma incongruência estatística que impede o cálculo do mesmo. Na tentativa de solucionar esta questão, foi removido o campo *school_rating_1to10*, dado que se acrescentou informação que não estava presente no conjunto de dados fornecido. Com esta remoção, os dados obtidos são apresentados na figura 3.4, verificando-se assim que o problema não se encontra neste campo.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value ↓	Code
(Intercept)	699255.456	822952.224	?	?	0.850	0.398	
num_of_bedrooms	3499.134	3174.477	0.028	0.782	1.102	0.274	
year_built	-497.568	428.083	-0.065	0.211	-1.162	0.249	
tax_assessed_value	-0.002	0.001	-0.022	0.981	-1.572	0.120	
num_of_bathrooms	-20569.944	2869.625	-0.154	0.601	-7.168	0.000	****
house_sqft	119.939	4.652	1.130	0.663	25.780	0	****
rate_per_sqfoot	2808.209	179.235	0.922	0.859	15.668	0	****

Figura 3.4: Resultados da Regressão Linear sem o atributo *school_rating_1to10*.

(b) ***Que atributos têm maior peso?***
Considerando os resultados obtidos na figura 3.4 podemos concluir que o atributo com maior peso, é o *num_of_bathrooms*, com um *p-value* de 0.000, uma vez que este valor é inferior a 0.05 sendo este o valor de referência para definir o quão importante um atributo é. De realçar ainda que os atributos *rate_per_sqfoot* e *house_sqft* têm um *p-value* de 0, o que significa que o *dataset* têm algumas inconsistências.

(c) ***Algum atributo foi removido do conjunto de dados por não ter uma boa capacidade de previsão? Em caso afirmativo, quais e por que você acha que eles não eram eficazes na previsão?***
Além do atributo *city* e *home_type*, foi removido também o atributo *school_rating_1to10* pois tal como foi apresentado apresentava a maioria dos valores em falta.

(d) ***Que outros atributos acha que ajudariam o seu modelo a prever melhor o preço da casa?***
Outros atributos que ajudariam o modelo construído a prever melhor o preço da casa, poderia passar pela avaliação dos transportes públicos na proximidade, ou a presença de polícia, bombeiros ou ainda um hospital relativamente perto a cada casa, por fim o grupo pensa que uma informação interessante acrescentar ao conjunto de dados seria a indicação se a casa possui cobertura de fibra óptica.

Capítulo 4

Parte 3

1. ***Encontre um dataset online que cumpra os requisitos para executar um modelo de regressão linear. De forma sucinta execute novamente todo o processo executado até aos resultados, avaliando os modelos atingidos.***
   O *dataset* escolhido para aplicar um modelo de regressão linear foi um relacionado com as variantes tinto do vinho "Vinho Verde"português [1] que é parte do mesmo conjunto de dados que o grupo está a realizar o trabalho prático.
   O conjunto de dados contém um total de 12 variáveis com 1599 entradas. O objectivo é então utilizar um modelo de regressão para determinar como diferentes variáveis independentes ajudam a prever a variável dependente, a qualidade. Saber como cada variável afectará a qualidade do vinho tinto ajudará os produtores, distribuidores e empresas da indústria de vinho tinto a avaliar melhor sua produção, distribuição e estratégia de preços.
   O conjunto de dados utilizado possui 12 atributos, sendo estes:

   - *fixed acidity*, quantidade de acidez fixa.
   - *volatile acidity*, quantidade de acidez volátil.
   - *citric acid*, quantidade de ácido cítrico.
   - *residual sugar*, quantidade de açúcar residual.
   - *chlorides*,, quantidade de cloretos.
   - *free sulfur dioxide*, quantidade de dióxido de enxofre livre.
   - *total sulfur dioxide*, quantidade de dióxido de enxofre total.
   - *density*, densidade do vinho.
   - *pH*, nível do pH do vinho.
   - *sulphates*, quantidade de sulfatos.
   - *alcohol*, nível de álcool do vinho.
   - *quality (class)*, qualidade do vinho.

   Neste conjunto de dados não foi necessário limpar, nem preparar os dados para análise, uma vez que não existem valores em falta, bem como, todos os atributos são numéricos o que é ideal para um modelo de regressão linear.

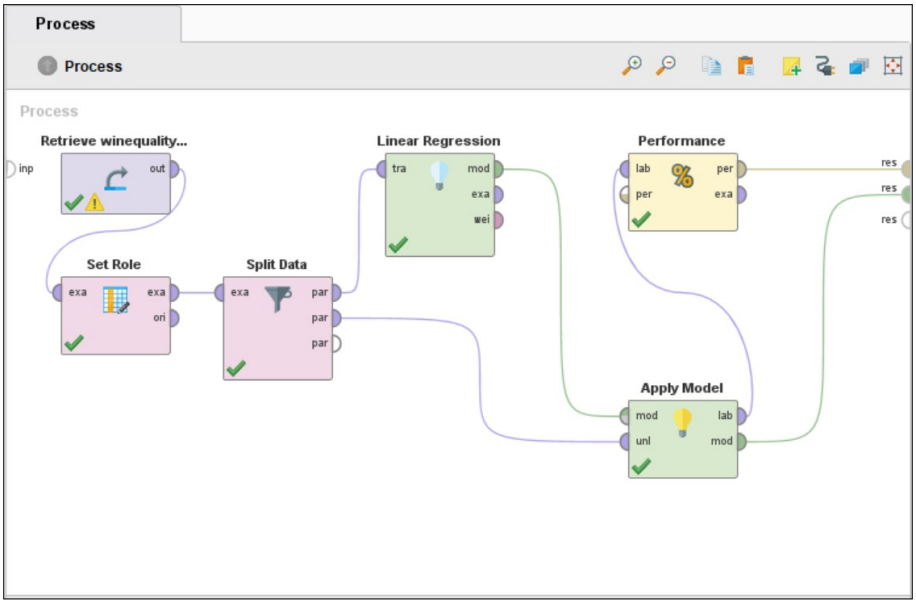

Figura 4.1: Processo de *Modeling* para o *dataset* escolhido.

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value ↓	Code
fixed acidity	0.018	0.020	0.038	0.956	0.921	0.357	
citric acid	-0.250	0.179	-0.060	0.815	-1.402	0.161	
residual sugar	0.021	0.015	0.037	1.000	1.471	0.142	
total sulfur dioxide	-0.002	0.001	-0.072	0.961	-2.714	0.007	***
pH	-0.520	0.186	-0.100	0.998	-2.800	0.005	***
chlorides	-2.013	0.487	-0.122	0.994	-4.138	0.000	****
(Intercept)	4.408	0.728	?	?	6.058	0.000	****
sulphates	0.899	0.133	0.187	0.971	6.767	0.000	****
volatile acidity	-1.239	0.143	-0.269	0.826	-8.653	0	****
alcohol	0.301	0.021	0.396	0.883	14.474	0	****

Figura 4.2: Resultados da Regressão Linear no conjunto de dados escolhido.

Tal como se pode observar na figura 4.2, existem dois atributos que apresentam o seu *p-value* a 0, sendo estes o atributo *volatile acidity* e *alcohol*. Este valor pode indicar alguma incongruência estatística que impede o cálculo do mesmo.

LinearRegression

```
  0.018 * fixed acidity
- 1.239 * volatile acidity
- 0.250 * citric acid
+ 0.021 * residual sugar
- 2.013 * chlorides
- 0.002 * total sulfur dioxide
- 0.520 * pH
+ 0.899 * sulphates
+ 0.301 * alcohol
+ 4.408
```

Figura 4.3: Fórmula da Regressão Linear no conjunto de dados escolhido.

root_mean_squared_error

```
root_mean_squared_error: 0.640 +/- 0.000
```

Figura 4.4: *Root Mean Square Error* obtido no conjunto de dados escolhido.

Considerando os resultados obtidos na figura 4.2 podemos concluir que os atributos com maior peso, são o *total sulfur dioxide* e *pH*, com um *p-value* de 0.007 e 0.005, respectivamente, e ainda os atributos *chlorides* e *sulphates*, ambos com um *p-value* de 0.000. Nestes atributos o *p-value* é inferior a 0.05 sendo este o valor de referência para definir o quão importante um atributo é. Outros atributos que ajudariam o modelo construído a prever melhor a qualidade do vinho poderia ser o ano da colheita ou tempo de fermentação.

Capítulo 5

Conclusão

O presente relatório descreveu, de forma sucinta, o processo de exploração e estudo de um modelo de regressão Linear no *RapidMiner* num caso de estudo específico.

Após a realização deste problema, foi possível compreender processo de regressão linear, bem como, desenvolver um modelo de regressão linear no *RapidMiner*.

Por fim, esperamos que os conhecimentos obtidos e consolidados sejam de enorme utilidade tendo uma perspectiva futura.

Bibliografia

[1] *Red Wine Quality*. 2018. URL: https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009 (ver p. 7).