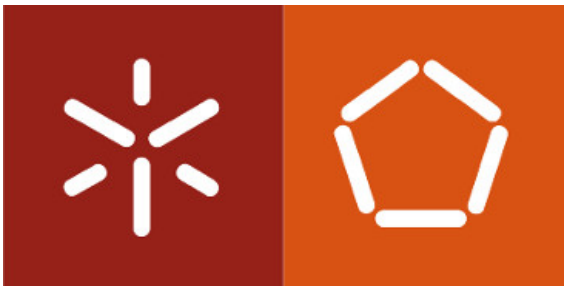

Universidade do Minho

Mestrado Integrado em Engenharia Informática

Engenharia do Conhecimento

Descoberta do Conhecimento

Resolução da Ficha de Exercícios 05

Gonçalo Pinto, A83732

João Diogo Mota, A80791

José Gonçalo Costa, PG42839

José Nuno Costa, A84829

23 de abril 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 objetivo a introdução ao conceito de descoberta do conhecimento.

Nesta ficha pretende-se que cada grupo execute os processos de *Data Understanding* e *Data Preparation* enquadrados na metodologia CRISP-DM de *Data Mining*, de forma a perceber quais os atributos que poderão ter mais influencia sobre outros através da sua correlação.

Para este efeito, foi fornecido um conjunto de dados pela equipa docente sobre consumo/eficiência de combustível num determinado veículo, de forma a tratar estes atributos e obter a matriz de correlação entre eles.

Capítulo 2

Parte 1

1. ***Quais as principais limitações de modelos de correlações?***
   Uma das maiores limitações de um modelo de correlação é que, a nível estatístico, a correlação é uma medida de uma relação entre duas variáveis e não a prova que existe causalidade entre elas.
   Outra das limitações acontece nas correlações não lineares, em que as variáveis não variam constantemente conforme outra aumentam ou diminuem o seu valor num único sentido.

2. ***O que é um coeficiente de correlação e como é interpretado?***
   Um dos componentes mais significativos no mundo da estatística, e bastante utilizado em *data mining* é o uso de matrizes de correlação. Uma correlação representa o nível de relação entre duas variáveis, visando entender como uma variável se comporta consoante a evolução de outra.
   Uma correlação é um valor número entre -1 e 1. Se um par de variáveis tiver uma correlação positiva, significa que quando os valores de uma delas aumentam, os da outra variável têm tendência a aumentar, estando estas mais relacionadas quanto mais próximo de 1 este valor for. Se o valor for negativo, o significado é oposto, ou seja, os valores de uma variável terão tendência a diminuir quando os valores da outra aumentam.
   Desta forma, tendo em conta as correlações existentes entre as variáveis, a forma mais intuitiva de as demonstrar é a construção de uma matriz que permite juntar todas as correlações numa só tabela.

3. ***Qual a diferença entre uma correlação negativa e uma correlação positiva?***
   Correlações positivas significam, que à medida que o valor de um atributo aumenta ou diminui, o valor do outro atributo aumenta ou diminui também. Correlações negativas significam que à medida que o valor de um atributo aumenta o valor do outro atributo diminui, ou vice-versa.

   A ***Se dois atributos diminuem essencialmente à mesma taxa é uma correlação positiva e negativa? Explique.***
   Se dois atributos diminuem essencialmente à mesma taxa então vai haver uma correlação positiva entre ambos, sendo que devido ao facto de ambos diminuírem pode ser que haja uma relação entre ambos, tornando assim a relação entre os dois atributos uma correlação positiva.

4. ***Como é medida a força de uma correlação? Quais os limites para essa força?***
   A correlação é uma técnica estatística que pode mostrar se e com que intensidade os pares de atributos estão relacionados. Uma correlação é um número entre -1 e +1 que mede o grau de associação entre dois atributos. Desta forma, a força de uma correlação é considerada forte quanto mais próximo um coeficiente de correlação estiver de 1 ou de -1. Contudo esta abordagem tem limitações como por exemplo a suposição de que uma correlação prova causalidade é perigosa e muitas vezes falsa, isto é, embora estatisticamente exista uma correlação entre dois atributos, pode não existir nenhuma razão lógica para esse efeito.

5. ***Consegue pensar em atributos que poderiam ser interessantes incluir no dataset estudado no exemplo da aula?***
   Ao *dataset* estudado no exemplo da aula, seria interessante acrescentar os seguintes atributos:

   - Avg_Usage: média de horas diárias que, numa casa, o aquecimento doméstico está ligado;
   - Indoor: média diária da temperatura interna da casa;
   - Year: ano de construção da casa.

Capítulo 3

Parte 2

1. ***Aceda ao ficheiro mpg_ dataset.csv.***
   Feito no *RapidMiner*.
2. ***Execute a operação de Data Understanding.***
   - *cylinders*, traduz o número de cilindros no motor. Trata-se de um valor numérico que varia entre 3 e 8.
   - *displacement*, representa a cilindrada do motor. Trata-se de um atributo numérico, podendo assumir um valor entre 68 e 455.
   - *horsepower*, significa a potência do motor. É um valor numérico que varia entre 100 e 980.
   - *weight*, traduz o peso do veículo. Trata-se de um valor numérico que varia entre 1613 e 5140.
   - *acceleration*, representa a aceleração do veículo, i.e., tempo em segundos para acelerar de 0 a 60. Trata-se de um valor numérico entre 8 e 30.
   - *model year*, significa o ano do modelo do veículo no anos 1900s. Trata-se de um valor numérico entre 70 e 82.
   - *origin*, traduz a origem do carro. Trata-se de um atributo numérico, podendo ser de origem Americana, Europeia ou Japonesa, estando representados no dataset por 1, 2 e 3 respetivamente.
   - *mpg*, representa o consumo/eficiência de combustível (miles per gallon - mpg). Trata-se de um valor numérico entre 9 e 46.6.
3. ***Execute a etapa de Data Preparation no Weka ou no RapidMiner.***
   O processo de *Data Preparation* é um processo essencial em *Data Mining*, pois é nesta fase que são tratados os atributos que poderão causar problemas numa fase de análise, tal como, o tratamento dos atributos que não façam sentido para o contexto pretendido ou atributos vazios. Para isto, recorreu-se ao software *RapidMiner* para fazer este processo. Nesta ferramenta, foi importado o *dataset* a bruto, e recorreu-se ao operador *"Replace Missing Values"* para tratar dos casos referidos a cima, estando demonstrados na figura 3.1 as configurações utilizadas.

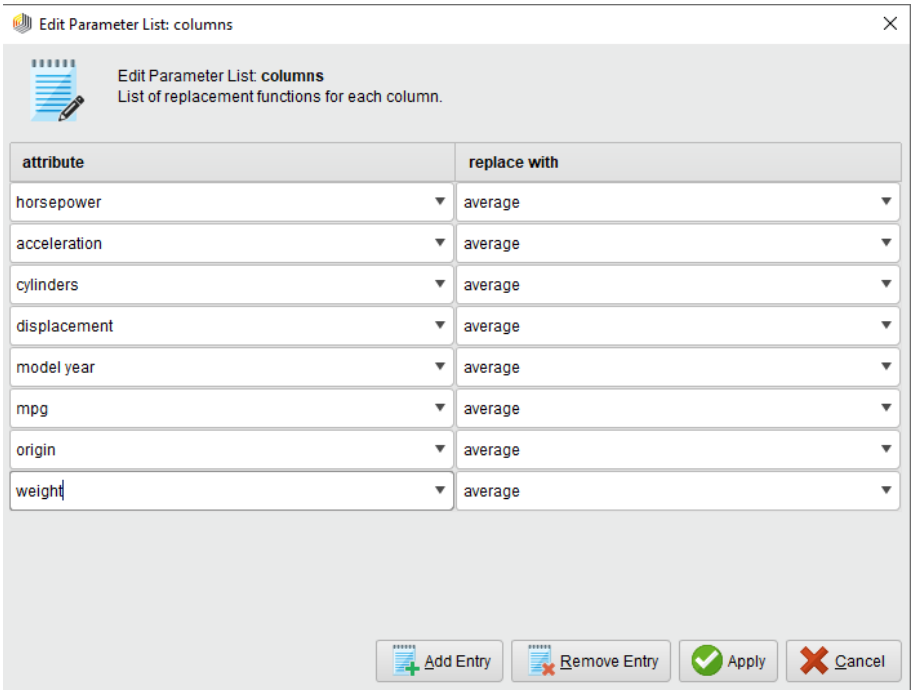

Figura 3.1: *Data Preparation.*

4. ***Documente quais os atributos que podem influenciar ou explicar o consumo/eficiência de combustível num determinado veículo (mpg).***

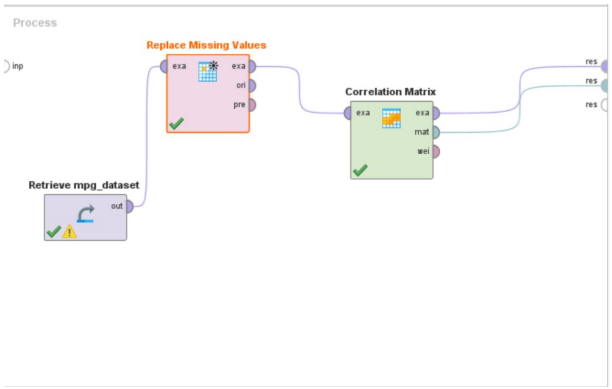

Figura 3.2: Estabelecimento das ligações.

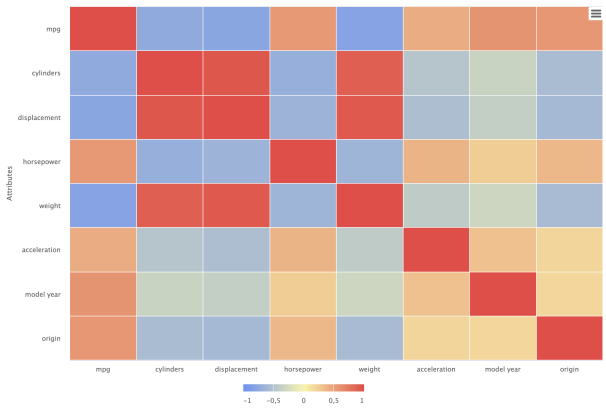

Figura 3.3: Matriz de Correlação por cores.

Attribu...	mpg	cylinde...	displac...	horsep...	weight	acceler...	model ...	origin
mpg	1	-0.749	-0.790	0.531	-0.820	0.414	0.566	0.546
cylinders	-0.749	1	0.946	-0.701	0.892	-0.482	-0.350	-0.570
displac...	-0.790	0.946	1	-0.651	0.933	-0.546	-0.380	-0.606
horsepo...	0.531	-0.701	-0.651	1	-0.639	0.375	0.227	0.355
weight	-0.820	0.892	0.933	-0.639	1	-0.418	-0.314	-0.580
acceler...	0.414	-0.482	-0.546	0.375	-0.418	1	0.302	0.184
model y...	0.566	-0.350	-0.380	0.227	-0.314	0.302	1	0.178
origin	0.546	-0.570	-0.606	0.355	-0.580	0.184	0.178	1

Figura 3.4: Matriz de Correlação.

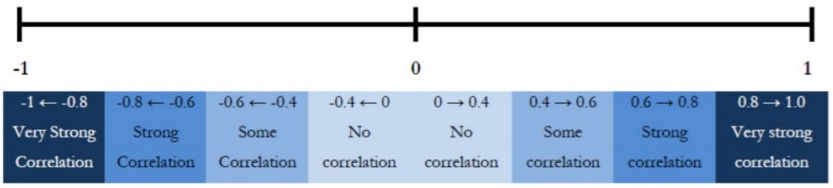

Figura 3.5: Avaliação da correlação.

Como se pode observar na figura 3.4, existem 3 atributos que influenciam bastante o consumo/eficiência de combustível num determinado veículo sendo estes o número de cilindros no motor (*cylinders*), o peso do veículo (*weight*) e a cilindrada do motor (*displacement*), dado que são estes valores que apresentam uma maior correlação, respectivamente com os seguintes valores -0.749, -0.820 e -0.790.

Através da imagem 3.5 podemos observar que existe efetivamente influência por parte destes atributos, tendo o *cylinders* e o *displacement* uma correlação negativa forte, por estarem no intervalo de valores de -0.8 a -0.6, e o *weight* uma correlação negativa muito forte, por estar no intervalo de valores de -1.0 a -0.8. Com estes dados podemos concluir que quanto maiores os valores destes atributos forem, menor será o consumo/eficiência de combustível e vice-versa.

Capítulo 4

Conclusão

O presente relatório descreveu, de forma sucinta, o processo de exploração e estudo dos modelos de correlação no *RapidMiner* num caso de estudo específico.

Após a realização deste problema, foi possível compreender o conceito de correlação, reconhecer o formato necessário dos dados para executar uma correlação, desenvolver um modelo de correlação no *RapidMiner* e ainda foi possível compreender os coeficientes numa matriz de correlação e explicar o seu significado e preponderância.

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