



**Data
Science**

Introdução à Engenharia e Ciência de Dados

Meta 1 - Ciência de Dados

Jorge Henriques, Mauro Pinto

jh@dei.uc.pt, mauropinto@dei.uc.pt



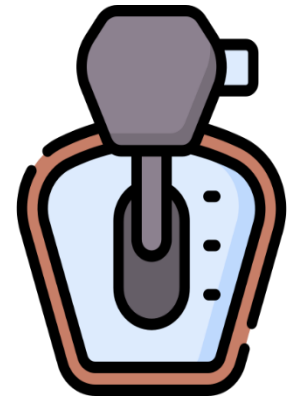
Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia
UNIVERSIDADE DE COIMBRA

- **Objetivo**
- Trabalho - Etapas
- Avaliação



▲ Objetivo

- Implementar e analisar alguns dos métodos estudados em IECD relativos à análise de dados, na classificação do tipo de transmissão de um automóvel: **manual / automática**



Classificação do tipo de transmissão de um automóvel

- Para o efeito conhece-se um conjunto de variáveis/características relativas a um automóvel.
- O desafio é, com base nessas variáveis, desenvolver um sistema que seja capaz de decidir se o automóvel tem **mudanças manuais ou automáticas**.

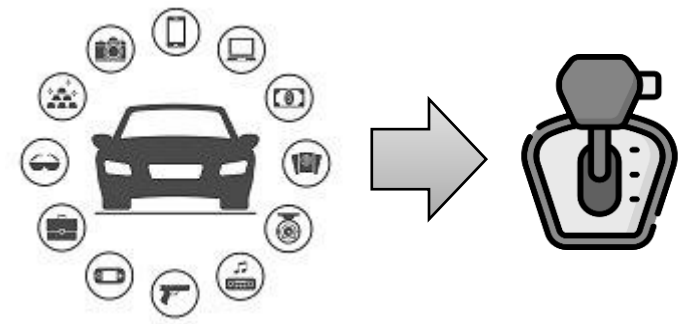


- Objetivo
- **Trabalho - Etapas**
- Avaliação



Trabalho - Etapas

- 1 | Aquisição de dados
- 2 | Pré – processamento de dados
- 3 | Transformação de Dados
- 4 | Modelização & Validação



1 | Aquisição de dados

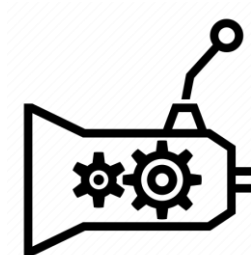
- Para cada automóvel admite-se que são conhecidas as seguintes características

Motor

■ X1	<i>Height</i>	<i>Altura</i>
■ X2	<i>Length</i>	<i>Comprimento</i>
■ X3	<i>Width</i>	<i>Largura</i>
■ X4	<i>Horsepower</i>	<i>Potencia</i>
■ X5	<i>Torque</i>	<i>Torque</i>
■ X6	<i>Hybrid</i>	Se motor elétrico ou combustão interna



1 | Aquisição de dados



Transmissão

- X7 *Transmission* Se transmissão automática ou manual

Outras

- X8 *Driveline* Tração (4 rodas, à frente, atrás)
- x9 *nb_gears* Número de mudanças
- x10 *Make* Fabricante
- x11 *city_mpg* Consumo em cidade (milhas por galão)
- x12 *highway_mpg* Consumo em autoestrada (milhas por galão)
- x13 *year* Ano de fabrico
- x14 *fuel_type* Tipo de combustível



1 | Aquisição de dados

- OS dados são disponibilizados num tabela, de dimensão (N,M)
 - N=5076 (numero de carros)
 - M=14 (número de características)

X1	X2	X2	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14
..						

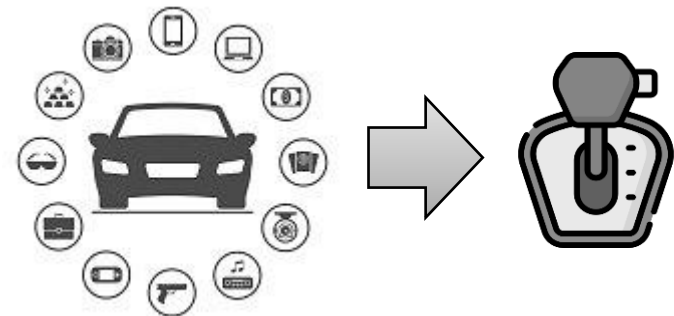


1 | Aquisição de dados

- x1 *Height* *Continua [1, 255]*
 - x2 *Length* *Continua [2, 255]*
 - x3 *Width* *Continua [1, 254]*
 - x4 *Horsepower* *Continua [100, 638]*
 - x5 *Torque* *Continua [98, 774]*
 - x6 *Hybrid* {0,1}= Um motor híbrido tem uma parte elétrica e outra a combustão
- | | | |
|------|---------------------|------------------------------|
| ▪ x7 | <i>Transmission</i> | {0,1}= {Manual, automática } |
|------|---------------------|------------------------------|
- x8 *Driveline* {0,1,2}={Quatro, Frente, Trás}
 - x9 *nb_gears* {4,5,6,7,8}
 - x10 *Make* {0,1,2,3,4,5,6,7,8}={BMW, Hyundai, Mazda, MINI, Mercedes, Mazda, Volvo, Saab, Kia}
 - x11 *city_mpg* *Continua [8, 38]*
 - x12 *highway_mpg* *Continua [8, 223]*
 - x13 *year* {2009, 2010, 2011, 2012}
 - x14 *fuel_type* {0,1}={Diesel, Gasolina}

▲ Etapas

- 1 | Aquisição de dados
- 2 | Pré – processamento de dados
- 3 | Transformação de Dados
- 4 | Modelização & Validação



2 | Pré – processamento de dados

■ 2.1 | Valores em falta

- Alguma das características do motor são desconhecidas.
- Assim, as variáveis de altura, largura e comprimento do motor têm, cada uma, 20 valores com valor 0 (valor em falta).

■ 2.2 | Outliers

- Alguns dos valores são “estranhos”
- Em concreto, o consumo em cidade tem cerca de 20 valores com valores na gama 90-100.

2 | Pré-processamento de dados

Trabalho a fazer

Missing data: substituir valores em Falta

- Media dos existentes (?)
- Valores similares – vizinho mais próximo (?)
- Modelo regressivo (?)

Outliers

- Detecção
- Substituição

2.1 | Valores em falta

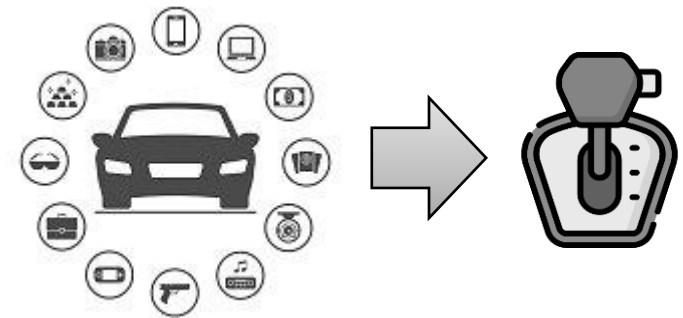
- Deve substituir os valores em falta (*características do motor*)
- **Método:** o que achar mais adequado

2.2 | Outliers

- Deve identificar e substituir os valores anómalos (*consumo em cidade*)
- **Método:** o que achar mais adequado

▲ Etapas

- 1 | Aquisição de dados
- 2 | Pré – processamento de dados
- **3 | Transformação de Dados**
- 4 | Modelização & Validação



3 | Transformação dos dados

3.1 | Seleção de variáveis

- Eventualmente algumas das variáveis não tem muito sentido ser adquiridas (ou tidas em conta para a decisão), uma vez que pouco contribuem para a decisão final

3.2 | Resumir os dados / extração de características

- Seria interessante perceber qual a relação entre alguns parâmetros, obtidos das variáveis de entrada e a decisão final
- Em concreto, pretende-se determinar a média de cada uma das variáveis relativamente aos carros com mudanças manuais e automáticas
 - **Exemplo:** para a potencia do motor, seria possível concluir algo do género ?
 - Média potência motor (transmissão automática) = 545
 - Média potência motor (transmissão manual) = 175

3 | Transformação dos dados

Trabalho a fazer

Seleção variáveis

- Com base na correlação
- *Forward* (começar com uma variável e ir acrescentando)

Médias

- Media/desvio padrão de cada classe (manual/automático)

3.1 | Seleção de variáveis

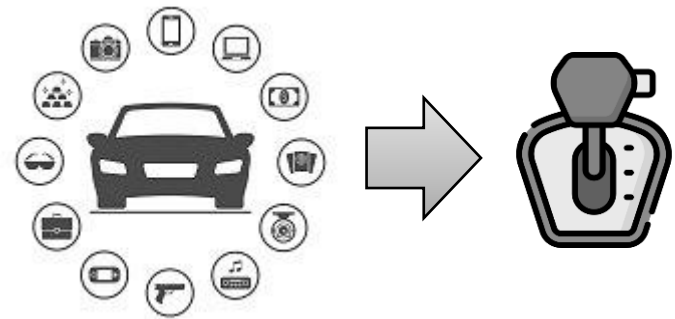
- Selecionar apenas as variáveis mais relevantes (4 ou 5 ?)
- **Método:** o que achar mais adequado.

3.2 | Resumir os dados / extração de características

- Determinar a média e desvio padrão para cada variável, em função da decisão
 - Ex. **PotênciaMédia** (manual)=175
 - Ex. **PotênciaMédia** (automatico)=545

▲ Etapas

- 1 | Aquisição de dados
- 2 | Pré – processamento de dados
- 3 | Transformação de Dados
- 4 | **Modelização & Validação**



▲ 4 | Modelização & Validação

■ 4.1 | Modelo de classificação

- Tendo em conta as variáveis selecionadas e os parâmetros extraídos (médias), deve construir um/vários modelos de classificação

■ 4.2 | Validação

- De forma a avaliar a qualidade do(s) classificador(es), os resultados obtidos pelos modelos devem ser quantificados usando métricas adequadas

4 | Modelização & Validação

Trabalho a fazer

Classificação: modelos

- Regressivo
- KNN
- Regras individuais (usando as variáveis selecionadas)

Validação

- Sensibilidade, especificidade, Fscore

4.1 | Modelo de classificação

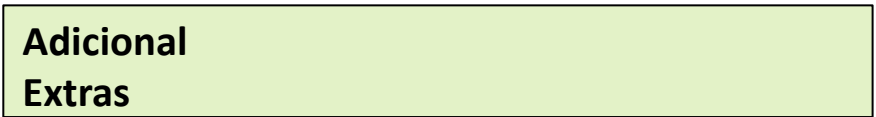
- 1 | Deve implementar um modelo regressivo
- 2 | Deve implementar o método KNN
- 3 | Deve implementar um método que use regras individuais (ou a sua combinação)
 - Exemplo** : Seja **PT a potência** de um motor
 - Se a potência **PT** estiver mais perto da **potênciaMédia** (manual) ➔ **manual**
 - Se a potência **PT** estiver mais perto da **potênciaMédia** (automatico) ➔ **automático**

4.2 | Validação

- Deve quantificar os resultados de cada um dos três modelos usando
 - SE – Sensibilidade, SP – Especificidade, Fscore**



Trabalho a fazer



**Adicional
Extras**

Extras

- Valoriza-se a aplicação de outras técnicas em qualquer uma das fases !
 - Dados de treino/validação (?)
 - Valores em falta: uso do Kmeans (?)
 - Uso de várias variáveis em simultâneo na geração de regras para a classificação (?)
 - Outros métodos (?)
 -

- Objetivo
- Trabalho - Etapas
- **Avaliação**



▲ Elementos de avaliação

- **Entrega – 12 MAio**
- Realizados em grupos de 2 alunos
- A defesa é individual
- Avaliação: 50% (relatório + código) + 50% defesa (individual)

Deve submeter ficheiro zip

- **Nome** | numeroAluno1_numeroAluno2 Ex. **201234124_20144233.zip**
- **Incluir** | código + relatório

■ Código

- **IMPORTANTE:** É apenas permitido o uso das bibliotecas numpy e matplotlib
- **Soluções que façam uso de outras bibliotecas não serão consideradas na avaliação**

▲ Elementos de avaliação

■ Relatório

- Documento pdf resumindo o trabalho apresentado desenvolvido
- Sugestão: 4/5 páginas
 - Não se espera código no relatório nem a sua explicação
 - Deve explicar que métodos implementou em cada fase e o porquê da sua escolha
 - Deve apresentar e discutir/comentar os resultados obtidos

■ Defesa

- Em data a combinar
- Haverá defesa **presencial, individual e obrigatória** do trabalho.