# *Entropia, Informação Mútua e Codificação de Huffman*

Teoria da Informação – Novembro de 2023

*Nuno Batista, Miguel Martins, André Albuquerque*

Universidade de Coimbra | Licenciatura em Engenharia Informática

# ÍNDICE

**1.** Introdução e Objetivos

**2.** Exercícios

**2b.** Relação entre MPG e as restantes variáveis

**7b.** Valor médio teórico de bits por símbolo

**8b)** Comparação do número médio de bits por símbolo após a codificação de huffman.

**8c)** Variância dos comprimentos.

**10b)** Relação da informação mútua com os coeficientes de correlação de Pearson.

**11b)** Comparação entre os valores reais de MPG e os valores estimados de MPG com base nas outras variáveis.

**11e)** Influência das variáveis com mais ou menos informação mútua na estimativa de MPG.

**3.** Conclusões

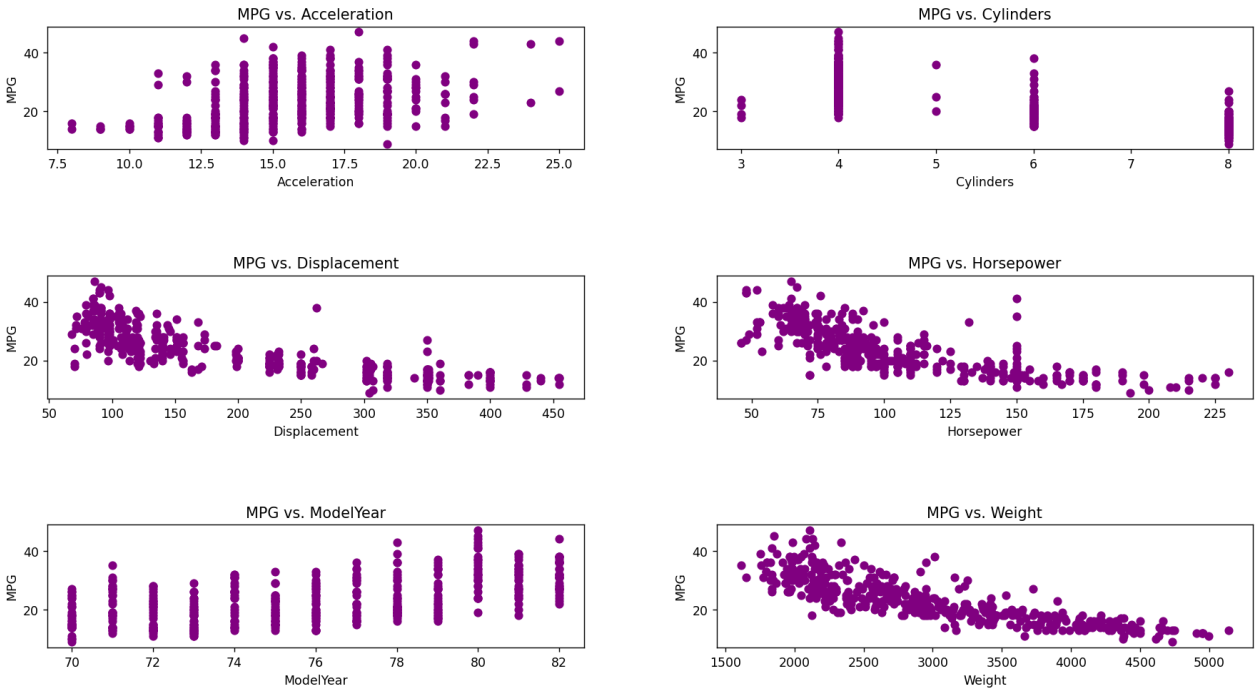
## 1. INTRODUÇÃO E OBJETIVOS

O objetivo deste relatório é explicar o processo de pensamento por trás das soluções usadas para implementar os vários exercícios da ficha TP1 - Entropia, Informação Mútua e Codificação de Huffman num programa de Python.

## 2. EXERCÍCIOS

**2d) Relação entre MPG e as restantes variáveis.**

Após fazer a representação gráfica da variável MPG em função de cada uma das outras variáveis, é evidente que são traçadas algumas relações.




Nomeadamente, quanto maior é o valor de **MPG**, menor é o valor de weight, **displacement**, **horsepower** e **cylinders**. Em contrapartida, o aumento de **MPG** traduz-se para um aumento de **acceleration** e **model year**.

**7c) Valor médio teórico de bits por símbolo.**

Neste exercício foi calculada a entropia de cada uma das variáveis:

$$H(X) = -\sum P(X = xi) \log 2P(X = x_i)$$

| | Acceleration | Cylinders | Displacement | Horsepower | Model Year | Weight | MPG |
|---|---|---|---|---|---|---|---|
| ***H(X)*** | **3.496423557** | **1.590435690** | **4.874068785** | **4.583748555** | **3.690642511** | **6.040364750** | **4.835799622** |

Verificou-se que quanto mais dispersas estão as distribuições de probabilidades maior é o número médio teórico de bits por símbolo, como era de esperar pela definição de entropia.

**8b) Comparação do número médio de bits por símbolo após a codificação de huffman.**

| | Acceleration | Cylinders | Displacement | Horsepower | Model Year | Weight | MPG |
|---|---|---|---|---|---|---|---|
| ***L (C,X)*** | **3.535626535** | **1.729729729** | **4.911547115** | **4.614250614** | **3.727272727** | **6.076167076** | **4.86977886** |

Verifica-se o Teorema da Codificação de Fonte de Shannon, $H(X) \leq L(C, X)$.

Os valores do número médio de bits por símbolo após a codificação de Huffman de uma variável é ligeiramente maior que a entropia dessa mesma variável.

**8c) Variância dos comprimentos.**

Reduzir a variância da codificação dos símbolos é importante para evitar que um buffer de comunicação seja maior do que o necessário. Para reduzir a variância, pode usar-se uma modificação no algoritmo de construção da árvore de Huffman que prioriza os nós que têm menos folhas.

**10b) Relação da informação mútua com os coeficientes de correlação de Pearson.**

A informação mútua entre MPG e as outras variáveis é dada por

$$I(Var; MPG) = H(Var) - H(Var|MPG)$$

quanto maior for o valor absoluto do coeficiente de correlação de Pearson, menor será $H(Var|MPG)$ e mais próxima estará a informação mútua do valor de $H(Var)$.

**11b) Comparação entre os valores reais de MPG e os valores estimados de MPG com base nas outras variáveis.**

| Tendo em conta todas as variáveis | | Tendo em conta todas as variáveis, exceto a com maior informação mútua | | Tendo em conta todas as variáveis, exceto a com menor informação mútua | |
|---|---|---|---|---|---|
| **Valor de MPG previsto** | **Valor de MPG real** | **Valor de MPG previsto** | **Valor de MPG real** | **Valor de MPG previsto** | **Valor de MPG real** |
| **15.4060** | **18** | **36.0796** | **18** | **17.1580** | **18** |
| **14.2504** | **15** | **36.0391** | **15** | **16.0024** | **15** |
| **16.0505** | **18** | **36.1695** | **18** | **17.6564** | **18** |
| **15.8628** | **16** | **35.9818** | **16** | **17.6148** | **16** |
| **15.8474** | **17** | **36.1729** | **17** | **17.4533** | **17** |
| **10.8705** | **15** | **36.3880** | **15** | **12.3305** | **15** |
| **10.9825** | **14** | **36.5000** | **14** | **12.2965** | **14** |
| **11.1515** | **14** | **36.4862** | **14** | **12.4655** | **14** |
| **10.2443** | **14** | **36.3341** | **14** | **11.7043** | **14** |
| **13.7536** | **15** | **36.4687** | **15** | **15.0676** | **15** |
| **[...]** | | | | | |

**11e) Influência das variáveis com mais ou menos informação mútua na estimativa de MPG.**

Utilizando todas as variáveis na estimativa de MPG, o erro de precisão calculado pela função MAE é 2.5721.

Quando se retira da equação o termo envolvendo a variável que apresentou o menor valor de MI, o erro de precisão aumenta pouco, passa a ser 3.0999. Por outro lado, ao retirar o membro envolvendo a variável com maior MI, o erro de precisão aumenta muito, passando a ser 17.1541.

## 3. CONCLUSÕESj

Após cumprir os objetivos deste trabalho, verificou-se que o valor médio teórico de bits por símbolo é fundamental para a compressão de dados e que a codificação de Huffman aproxima-se bastante do limite teórico da média de bits por símbolo. Descobrimos também que os coeficientes de Pearson estão diretamente relacionados à informação mútua de variáveis. Por fim observou-se a influência de variáveis com mais ou menos informação mútua durante o processo de estimativa de MPG.