



**UNIDADE CURRICULAR:**

**CÓDIGO:** 21097

**DOCENTE:** Joaquim Neto

**A preencher pelo estudante**

**NOME:** Pedro Nuno Esteves de Abreu

**N.º DE ESTUDANTE:** 2300485

**CURSO:** Licenciatura em Engenharia Informática

**DATA DE ENTREGA:** 12/05/2026

## **APRENDIZAGEM AUTOMÁTICA: USANDO BIBLIOTECAS EM R**

# **RELATÓRIO**

## 1. Introdução

O presente trabalho tem como objetivo aplicar técnicas de aprendizagem supervisionada em linguagem R, nomeadamente árvores de decisão, k-vizinhos mais próximos (k-NN) e redes neurais, utilizando um *dataset* gerado individualmente através da função disponibilizada no enunciado.

O objetivo consiste em construir modelos capazes de prever se um utilizador de uma plataforma de *ecommerce* efetua ou não uma compra, com base em diferentes variáveis relacionadas com o seu comportamento no website. Para tal, serão avaliados e comparados diferentes modelos de classificação, discutindo-se o respetivo desempenho e adequação ao problema.

## 2. Descrição do Dataset

O *dataset* gerado possui:

- 75 observações;
- 5 variáveis preditoras;
- 1 variável-alvo (Classe).

A distribuição da variável de classe encontra-se equilibrada:

- 37 observações da classe Compra;
- 38 observações da classe NaoCompra.

### 2.1 Variável-alvo

A variável-alvo é aquilo que os modelos pretendem prever. Neste caso é:

- Classe

Representa:

- Compra
- NaoCompra

### 2.2 Variáveis preditoras

Variável	Descrição
TempoSite	Tempo passado no website
Paginas	Número de páginas visitadas
Promocoes	Existência de promoções visualizadas
Avaliacoes	Avaliação dos produtos
Frequencia	Frequência de utilização do website

### 3. Análise Exploratória dos Dados

Foi realizada uma análise estatística inicial através das funções:

- `summary()`, para perceber médias, mínimos, máximos e distribuição;
- histogramas, para perceber distribuição dos valores, detetar assimetrias e identificar concentrações; e
- `boxplots`, para comparar variáveis entre classes, identificar separação entre grupos e visualizar possíveis diferenças discriminativas.

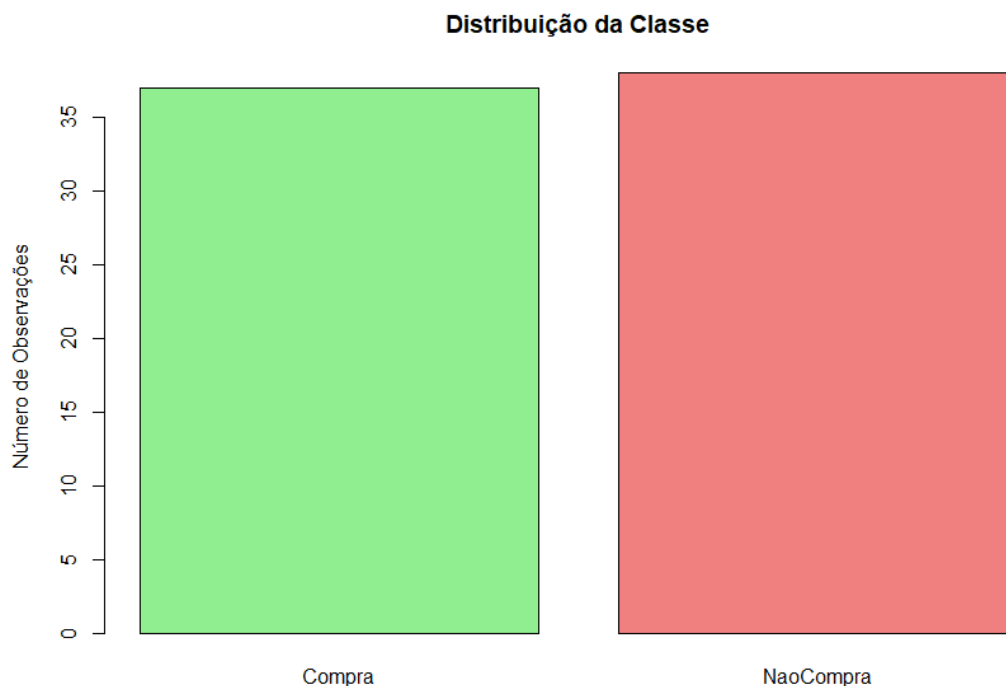
#### 3.1 Estatísticas descritivas

Variável	Média	Mínimo	Máximo
TempoSite	14.56	1	23.9
Paginas	11.49	1	22
Promocoes	0.39	0	1
Avaliacoes	3.67	1	5
Frequencia	6.33	1	12

A variável `Promocoes` é binária (0-ausência; 1-presença).

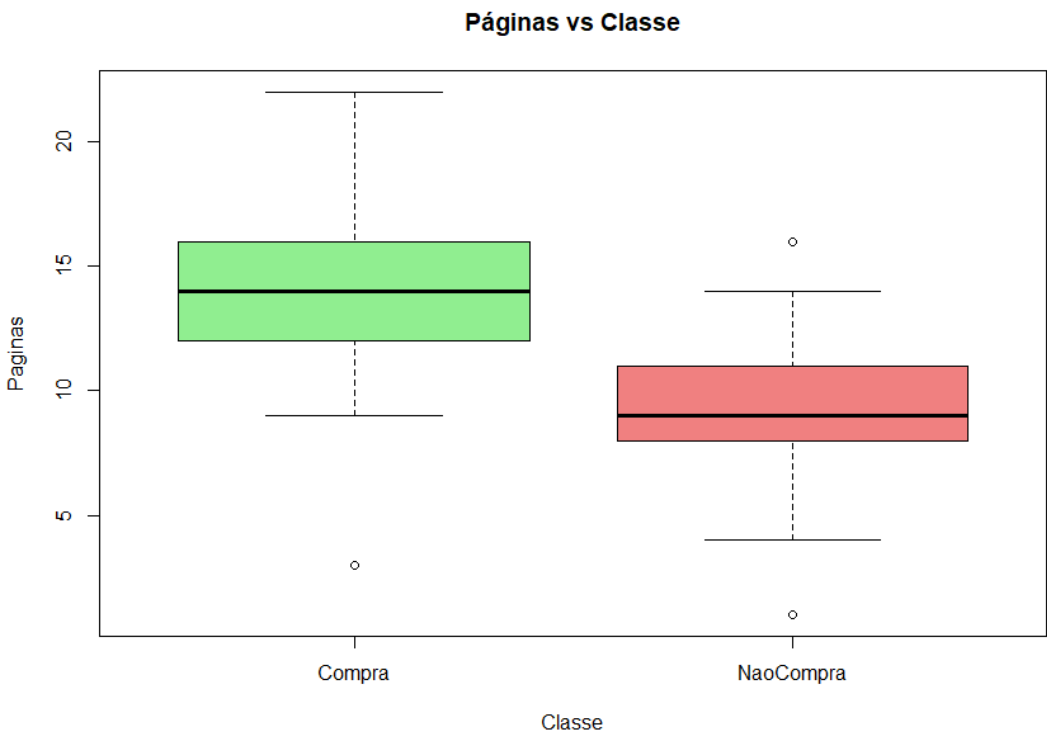
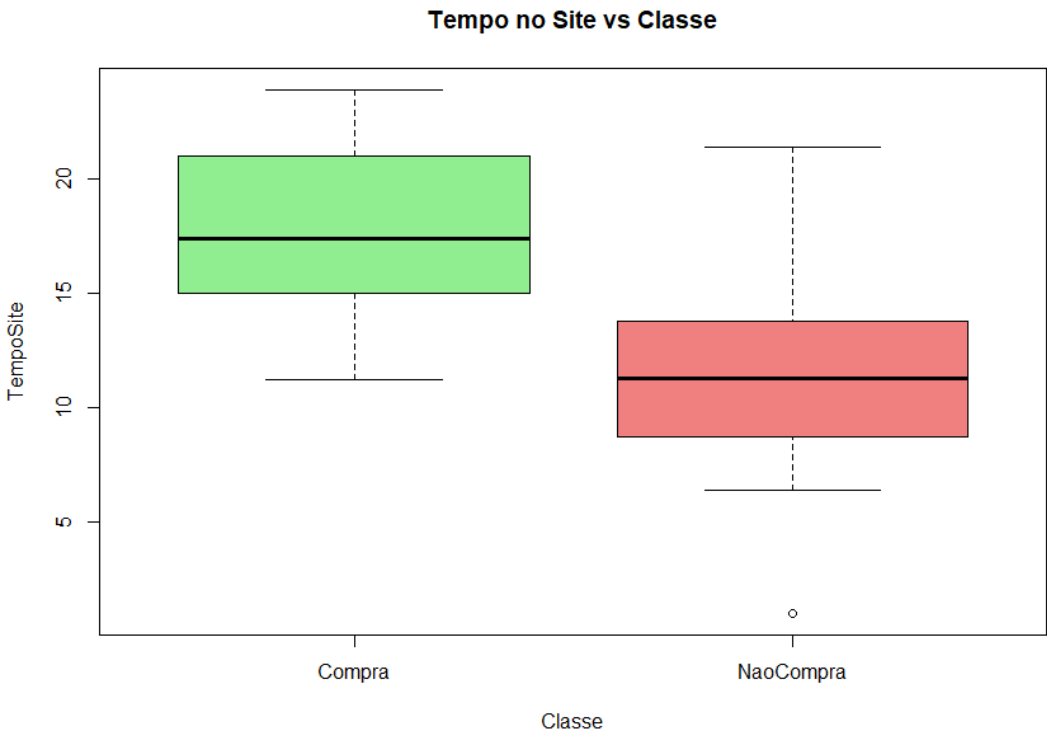
Não foram identificados valores em falta no *dataset*.

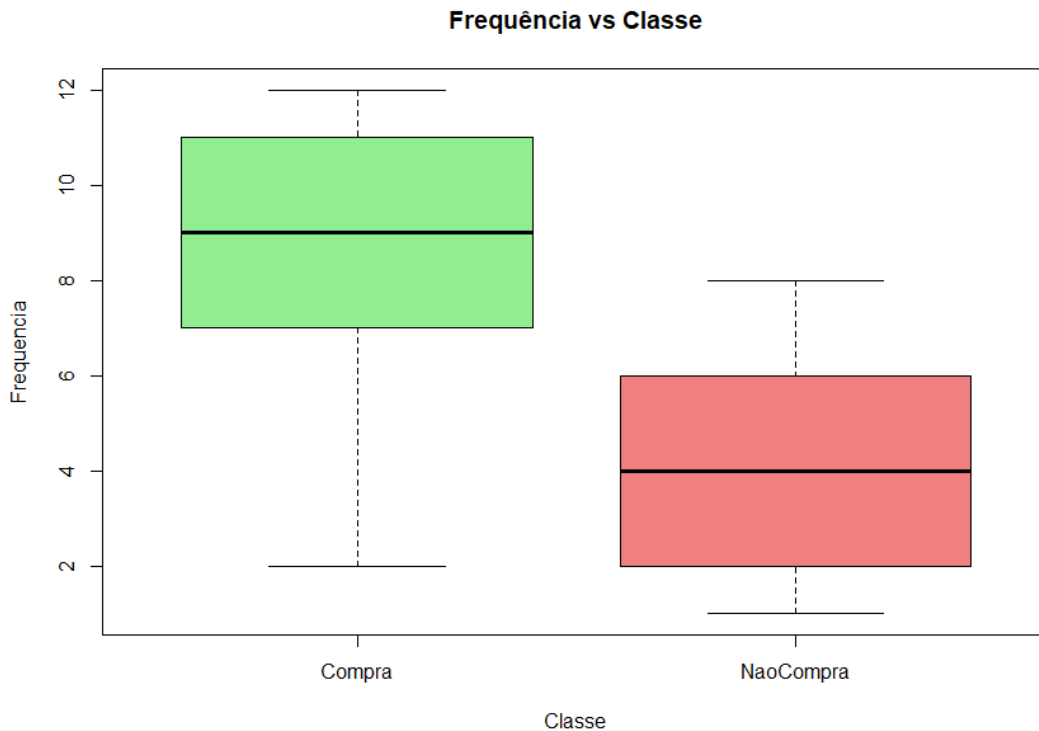
#### 3.2 Distribuição da Classe



A distribuição das classes apresenta um equilíbrio quase perfeito, o que é positivo para a construção dos modelos de classificação.

3.3 Relação entre variáveis e classe





A análise dos *boxplots* sugere que os utilizadores que efetuam compras tendem a:

- passar mais tempo no site;
- visitar mais páginas;
- utilizar o website com maior frequência.

Estas diferenças indicam que as variáveis escolhidas poderão ter capacidade preditiva relevante.

#### 4. Divisão dos Dados

O *dataset* foi dividido em: 70% para treino; 30% para teste. Esta divisão permite avaliar a capacidade de generalização dos modelos em dados não utilizados durante o treino.

A seleção das observações foi realizada de forma aleatória através da função `sample()`, reduzindo possíveis enviesamentos na distribuição dos dados.

Foi utilizada a função `set.seed(123)` para garantir a reprodutibilidade dos resultados, permitindo obter sempre a mesma divisão entre treino e teste em diferentes execuções do programa.

O conjunto de treino ficou composto por 52 observações e o conjunto de teste por 23 observações.

### 1 - Excerto do Treino

TempoSite	Paginas	Promocoes	Avaliacoes	Frequencia	Classe
17.5	11	1	3	9	Compra
13.8	8	0	4	1	NaoCompra
9.7	13	1	4	7	NaoCompra
20.9	9	0	3	12	Compra
22.5	20	1	2	9	Compra

### 2 - Excerto do Teste

TempoSite	Paginas	Promocoes	Avaliacoes	Frequencia	Classe
17.1	13	0	4	8	Compra
11.4	13	1	4	2	Compra
11.8	10	1	4	3	NaoCompra
19.4	14	0	3	12	Compra
13.9	9	1	5	6	Compra

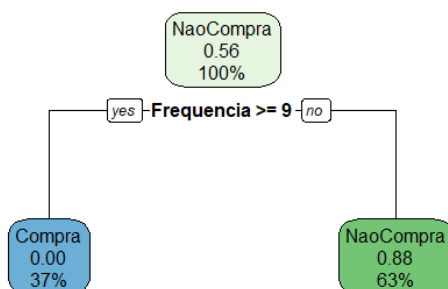
## 5. Árvore de Decisão

### 5.1 Construção do Modelo

Foi utilizado o algoritmo de árvores de decisão através da biblioteca `rpart`.

O modelo foi treinado utilizando o conjunto de treino previamente definido, tendo como objetivo prever a variável `Classe`.

### 5.2 Árvore Gerada



A árvore de decisão gerada utiliza a variável `Frequencia` como principal critério de separação.

O modelo indica que utilizadores com frequência de utilização superior ou igual a 9 apresentam maior probabilidade de realizar compras.

### 5.3 Avaliação do Modelo

A matriz de confusão obtida foi a seguinte:

Previsto	Compra	NaoCompra
Compra	5	0
NaoCompra	9	9

A *accuracy* obtida foi de aproximadamente:

- 60.87%.

### 5.4 Interpretação dos Resultados

O modelo apresentou um desempenho razoável, embora relativamente limitado.

A árvore obtida é bastante simples, utilizando apenas uma variável para a classificação. Isto sugere que o modelo poderá estar a simplificar excessivamente o problema, reduzindo a capacidade de generalização.

Contudo, a árvore apresenta elevada interpretabilidade, permitindo compreender facilmente as regras de decisão utilizadas<sup>1</sup>.

## 6. k-NN

### 6.1 Normalização dos Dados

O algoritmo k-NN baseia-se no cálculo de distâncias entre observações<sup>2</sup>.

Assim, foi necessário proceder à normalização das variáveis numéricas, de forma a evitar que variáveis com escalas maiores tivessem influência excessiva nos cálculos de distância.

Para tal, foi utilizada a fórmula *min-max normalization*:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

A normalização foi aplicada às variáveis:

- TempoSite;
- Paginas;
- Promocoões;
- Avaliacoões;
- Frequência.

---

<sup>1</sup> “One important property of decision trees is that it is possible for a human to understand the reason for the output of the learning algorithm” (Russell & Norvig, 2010, p. 707)

<sup>2</sup> “the very word ‘nearest’ implies a distance metric” (Russell & Norvig, 2010, p. 738).

## 6.2 Construção do Modelo

Foi utilizado o algoritmo k-NN através da biblioteca *class*.

Foram testados diferentes valores de k:

- k = 3;
- k = 5;
- k = 7;
- k = 9.

O objetivo consistiu em verificar a influência do número de vizinhos considerados no desempenho do modelo.

## 6.3 Avaliação

A matriz de confusão obtida foi a seguinte:

	Previsto Compra	Previsto NaoCompra
Real Compra	6	1
Real NaoCompra	8	8

A *accuracy* obtida foi de aproximadamente:

- 60.87%.

Verificou-se que todos os valores de k testados produziram exatamente o mesmo resultado.

## 6.4 Interpretação

O modelo k-NN apresentou um desempenho semelhante ao da árvore de decisão, obtendo igualmente uma *accuracy* de aproximadamente 60.87%.

A matriz de confusão demonstra que o modelo apresentou maior dificuldade em identificar corretamente os casos reais da classe Compra, uma vez que apenas 6 dos 14 casos foram corretamente classificados, correspondendo a uma taxa de identificação de aproximadamente 42.86%.

O facto de diferentes valores de k produzirem os mesmos resultados sugere que a estrutura do *dataset* é relativamente estável para este algoritmo, não existindo grande sensibilidade à escolha do número de vizinhos considerados.



## 7. Redes Neurais

### 7.1 Construção do Modelo

Foi utilizada uma rede neuronal artificial através da biblioteca `nnet`.

Tal como no modelo k-NN, foram utilizados os dados normalizados, uma vez que redes neurais apresentam melhor desempenho quando as variáveis se encontram na mesma escala.

O modelo foi treinado utilizando:

- 5 neurónios na camada escondida;
- máximo de 200 iterações.

### 7.2 Avaliação do Modelo

A matriz de confusão obtida foi a seguinte:

	Previsto Compra	Previsto NaoCompra
Real Compra	7	2
Real NaoCompra	7	7

A *accuracy* obtida foi de aproximadamente:

- 60.87%.

### 7.3 Interpretação dos Resultados

A rede neuronal apresentou um desempenho semelhante aos modelos anteriores, obtendo igualmente uma *accuracy* de aproximadamente 60.87%.

Apesar da *accuracy* global ser igual à dos restantes modelos, a distribuição dos erros foi ligeiramente diferente. O modelo conseguiu identificar corretamente mais casos da classe Compra (50%), embora também ainda tenha produzido muitos falsos positivos.

As redes neurais possuem maior capacidade de modelar relações complexas entre variáveis<sup>3</sup>. Contudo, neste caso, o reduzido número de observações poderá limitar a capacidade de aprendizagem do modelo<sup>4</sup>.

---

<sup>3</sup> “two hidden layers are enough to represent any function and a single layer is enough to represent any continuous function” (Russell & Norvig, 2010, p. 762)

<sup>4</sup> “...and from estimation error of not having enough training examples to limit variance” (Russell & Norvig, 2010, p. 712)

Comparativamente à árvore de decisão, a rede neuronal apresenta menor interpretabilidade, sendo mais difícil compreender diretamente as regras utilizadas para classificação.

## 8. Comparação dos Modelos

Modelo	Accuracy	% Acerto Compra	% Acerto NaoCompra
Árvore de decisão	60.87%	35,7%	100%
k-NN	60.87%	42,9%	88,9%
Rede neuronal	60.87%	50,0%	77,8%

Os três modelos apresentaram exatamente a mesma *accuracy* no conjunto de teste, obtendo um valor aproximado de 60.87%.

Apesar da igualdade na *accuracy* global, verificaram-se diferenças relevantes na capacidade de identificação das diferentes classes.

A árvore de decisão destaca-se pela elevada interpretabilidade, permitindo compreender facilmente as regras utilizadas pelo modelo, e pela excelente capacidade de identificação da classe NaoCompra, embora tenha apresentado maior dificuldade na identificação da classe Compra.

O k-NN implicou a normalização dos dados, dado que é um algoritmo fortemente dependente da proximidade entre observações.

A rede neuronal apresentou maior complexidade estrutural e capacidade teórica de modelar relações não lineares, embora o reduzido tamanho do *dataset* possa ter limitado o seu desempenho.

Os resultados demonstram que a *accuracy*, isoladamente, nem sempre é suficiente para avaliar completamente o comportamento de modelos de classificação, sendo importante analisar também o desempenho individual em cada classe.

## 9. Análise Crítica

Os resultados obtidos demonstram que os três modelos de aprendizagem supervisionada utilizados apresentaram desempenhos relativamente semelhantes, obtendo todos uma *accuracy* aproximada de 60.87%.

Apesar da igualdade na *accuracy* global, verificaram-se diferenças na distribuição dos erros entre classes. Este resultado demonstra que diferentes algoritmos podem alcançar desempenhos quantitativos semelhantes utilizando estratégias de classificação distintas.

A árvore de decisão destacou-se pela simplicidade e elevada interpretabilidade. O modelo permitiu identificar facilmente a variável Frequencia como principal fator de decisão, demonstrando que utilizadores mais frequentes apresentam maior probabilidade de realizar compras. Contudo, a simplicidade excessiva da árvore poderá indicar *underfitting*, reduzindo a capacidade de generalização do modelo.

O algoritmo k-NN demonstrou a importância da normalização dos dados em métodos baseados em distância. Apesar de terem sido testados diferentes valores de k, os resultados mantiveram-se inalterados, sugerindo alguma estabilidade estrutural do *dataset* relativamente a este algoritmo.

Relativamente à rede neuronal, embora este modelo possua maior capacidade teórica para representar relações complexas e não lineares, o reduzido número de observações poderá ter limitado a sua capacidade de aprendizagem. Redes neurais tendem geralmente a beneficiar de *datasets* significativamente maiores.

O *dataset* utilizado apresenta algumas limitações importantes:

- número reduzido de observações;
- possível simplicidade excessiva das relações entre variáveis;
- ausência de variáveis mais diversificadas;
- possível sobreposição entre classes.

Como trabalho futuro, seria interessante:

- aumentar o número de observações;
- utilizar validação cruzada;
- testar outros algoritmos de classificação;
- incluir novas variáveis explicativas;
- analisar métricas adicionais para além da *accuracy*.

Apesar destas limitações, o trabalho permitiu aplicar corretamente diferentes técnicas de aprendizagem supervisionada em R e comparar os respetivos desempenhos num problema de classificação binária (Compra ou Não Compra).

## 10. Conclusão

O presente trabalho permitiu aplicar técnicas fundamentais de aprendizagem supervisionada utilizando a linguagem R, nomeadamente árvores de decisão, k-vizinhos mais próximos (k-NN) e redes neuronais.

Os modelos desenvolvidos tiveram como objetivo prever a realização de compras numa plataforma de *ecommerce* com base em variáveis relacionadas com o comportamento dos utilizadores no website.

Os resultados obtidos demonstraram que os três modelos apresentaram desempenhos semelhantes, alcançando uma *accuracy* aproximada de 60.87%. Apesar da igualdade nos resultados globais, verificaram-se diferenças ao nível da interpretabilidade, complexidade e distribuição dos erros entre classes.

A árvore de decisão revelou-se o modelo mais facilmente interpretável, enquanto a rede neuronal apresentou maior complexidade estrutural. O algoritmo k-NN destacou a importância da normalização dos dados em métodos baseados em distância.

O trabalho permitiu consolidar conhecimentos relacionados com:

- preparação e análise de dados;
- divisão treino/teste;
- construção de modelos de classificação;
- avaliação de desempenho;
- interpretação crítica de resultados.

De forma global, o trabalho permitiu aplicar corretamente técnicas fundamentais de aprendizagem automática supervisionada e compreender as diferenças práticas entre diferentes modelos de classificação.

## **Bibliografia**

Russell, S., & Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3rd ed.). Prentice Hall.