

**Predictive Modelling for *TechScape***

**Machine Learning | MDSAA 2021/22**



**Group 60**António Fonseca - r2018…  
João Carvalho - r2018…  
Lucas Lopes - r2018…  
 Mariana Ferreira - r20181071  
Tiago Oom - r2018…

**Professors:** Roberto Henriques, Carina Albuquerque and Lara Oliveira

**Index**

[**1.** **Introduction** 2](#_Toc89547072)

[**2.** **Data preparation** 2](#_Toc89547073)

[**3.** **Dataset exploration** 2](#_Toc89547074)

[*3.1.* *Descriptive statistics* 2](#_Toc89547075)

[*3.2.* *Coherence checking* 3](#_Toc89547076)

[*3.3.* *Missing values* 3](#_Toc89547077)

[*3.4.* *Outliers* 3](#_Toc89547078)

[**4.** **Feature engineering** 3](#_Toc89547079)

[**5.** **Feature selection** 3](#_Toc89547080)

[**6.** **Model training** 3](#_Toc89547081)

[*6.1.* *Data partition* 3](#_Toc89547082)

[*6.2.* *Data preprocessing* 3](#_Toc89547083)

[*6.3.* *Models* 3](#_Toc89547084)

[**7. Model deployment** 3](#_Toc89547085)

[**8. Discussion** 3](#_Toc89547086)

[**9. Conclusions** 3](#_Toc89547087)

[**10. References** 3](#_Toc89547088)

[**11. Appendices** 3](#_Toc89547089)

**Index**

[**1.** **Introduction** 2](#_Toc89547054)

[**2.** **Data preparation** 2](#_Toc89547055)

[**3.** **Dataset exploration** 3](#_Toc89547056)

[*3.1.* *Descriptive statistics* 3](#_Toc89547057)

[*3.2.* *Coherence checking* 3](#_Toc89547058)

[*3.3.* *Missing values* 3](#_Toc89547059)

[*3.4.* *Outliers* 3](#_Toc89547060)

[**4.** **Feature engineering** 3](#_Toc89547061)

[**5.** **Feature selection** 3](#_Toc89547062)

[**6.** **Model training** 3](#_Toc89547063)

[*6.1.* *Data partition* 3](#_Toc89547064)

[*6.2.* *Data preprocessing* 3](#_Toc89547065)

[*6.3.* *Models* 3](#_Toc89547066)

[**7. Model deployment** 3](#_Toc89547067)

[**8. Discussion** 3](#_Toc89547068)

[**9. Conclusions** 3](#_Toc89547069)

[**10. References** 3](#_Toc89547070)

[**11. Appendices** 3](#_Toc89547071)

## **Introduction**

(Introduzir o tema, objetivos e metodologia)

## **Data preparation**

(Dataset inicial e alteração nos tipos de variáveis)

## **Dataset exploration**

## *Descriptive statistics*

(Análise das estatísticas descritivas e dos gráficos)

## *Coherence checking*

(Explicar o raciocínio)

## *Missing values*

(Explicar que não há)

## *Outliers*

(Metodologia adotada para lidar com outliers)

## **Feature engineering**

(Explicar a criação das novas variáveis e alterações noutras para garantir a coerência e valores não omissos)

## **Feature selection**

(Processo de decidir as variáveis, explicar o raciocínio do k-folds)

## **Model training**

## *Data partition*

(Explicar o objetivo do k-folds)

## *Data preprocessing*

(Tratamento do train depois da partição)

## *Models*

(Desenvolvimento dos modelos)

## **7. Model deployment**

(Implementação do melhor)

## **8. Discussion**

(Discussão dos resultados – comparação de vários modelos?)

## **9. Conclusions**

(Contextualização dos resultados no problema e limitações)

## **10. References**

## **11. Appendices**