Trabalho de grupo IMD

Licenciatura em Ciência de Dados (2º ano)

17 de Novembro de 2022

Diana Mendes (diana.mendes@iscte-iul.pt)
Conceição Figueiredo (conceicao.figueiredo@iscte-iul.pt)
Ricardo Correia (ricardo_correia@iscte-iul.pt)

Deadline entrega: 18 de dezembro de 2022

Tema: Previsão dos preços de reservas em hotéis/alojamentos - desenvolver um modelo que permita estimar o preço a que os quartos são colocados à venda no Airbnb (http://insideairbnb.com/get-the-data/).

- Os docentes atribuem a cada grupo uma cidade (ver em **Anexo**)
- Os dados podem ser importados do site **Inside Airbnb**, http://insideairbnb.com/get-the-data/
- **Apenas fazer o download** do ficheiro listings.csv da **cidade do seu grupo**.
- Em média cada ficheiro tem 4000 linhas e 18 colunas (variáveis)
- Tirar tempo para procurar o máximo de informação de contexto que permita a **compreensão** dos **dados** e do **domínio** de onde provêm (*Business Understanding*).
- Neste trabalho recomenda-se usar o software R (script em RStudio ou Notebook em Jupyter Notebook, IRkernel).
- **Caso for necessário, os trabalhos de grupo podem ser sujeitos a uma apresentação oral.**

Pontos principais:

O desenvolvimento do projeto e do relatório final deve seguir a metodologia CRISP-DM.

1. Com base nos dados e a sua pesquisa sobre os dados (*Business Understanding*), definir o(s) *problema(s)*/*questões* que gostariam de *resolver*/*responder* e justificar a escolha das variáveis.
2. Fazer a limpeza dos dados.
3. Analisar gráficos e estatísticas descritivas das variáveis.
4. Analisar correlação e causalidade entre as variáveis.
5. Pré-processamento dos dados e manipulação de *features* (agrupar, juntar, eliminar, transformar as variáveis).
6. Usar algoritmos de aprendizagem supervisionada (regressão linear, regressão polinomial, interação de variáveis, regressão não-linear) sobre o seu conjunto de dados.
7. Dividir o dataset em conjuntos de treino e de teste.
8. Validar o modelo escolhido e fazer a previsão da variável dependente/alvo (sobre o conjunto de teste).
9. Avaliar a performance da previsão feita (sobre o conjunto de teste).
10. Interpretação/explicação dos resultados obtidos a partir dos dados (prós e contras).
11. **Relatório**: podem usar *dashboards*, editores de texto, ou outras ferramentas.
12. O relatório final deve ser enviado por e-mail em formato pdf. Devem ainda submeter um ficheiro zippado com o dataset final e o ficheiro com o código utilizado.
13. O **código** tem de poder ser executado AS-IS (diretamente) e deverá estar **devidamente comentado**.
14. Deverá incluir todas as experiências feitas, incluindo a análise e preparação de dados, modelação, e avaliação do modelo.
15. A interpretação dos resultados deve ser efetuada de forma crítica e não apenas factual.
16. O relatório tem um **limite de 25 páginas (sem anexos)**.

Anexo: Grupos e Cidade

Grupo 1: Amsterdam (listings.csv)

104825	Hugo Fontan
104809	Miguel Ferreira
104474	Alexandre Magalhães
105156	Alexandre Alves

Grupo 2: Boston (listings.csv)

104745	Diogo Catarino
104532	André Silvestre
104936	Rita Matos
104944	Francisco Gomes

Grupo 3: Prague (listings.csv)

104826	Ricardo Ângelo
104954	Ana Rodrigues
104756	Margarida Morais
104757	Mariana Campelo

Grupo 4: Dublin (listings.csv)

105146	Bernardo Arcão
105136	Gonçalo Sepúlveda
104869	Afonso Peças
104575	Santiago Taylor
104543	Guilherme Soças

Grupo 5: Edinburgh (listings.csv)

104716	Maria João Lourenço
103303	Eliane Susso
99239	Umeima Mahomed
110451	Marco Esperança

Grupo 6: Oslo (listings.csv)

104920	Eduarda Costa
104857	Margarida Matos
103950	Leonardo Medina
104900	Ricardo Martins

Grupo 7: Sevilla (listings.csv)

98601	Pedro Machado
104841	Diogo Freitas
104782	João Botas
103380	Allan Kardec

Grupo 8: Stockholm (listings.csv)

104675	Marta Lourenço
104867	Rosarinho Carvalho
104801	Diana Edral

Grupo 9: Bologna (listings.csv)

104835	Gustavo Veloso
105356	José Antunes
105176	Tiago Coelho

Grupo 10: San Francisco (listings.csv)

105289	André Plancha
105208	Afonso Silva
104914	Rui Chaves
105220	Tomás Ribeiro

Grupo 11: Munich (listings.csv)

104987	Douglas Zachhau
105144	Miguel Rodrigues
104753	Daniel Ramalhete
105185	Daniel Castro

Grupo 12: Northern Rivers (listings.csv)

104940	Carolina Azevedo
105172	Márcia Rita
104779	Maria Pedras
105188	João Madeira

Grupo 13: Portland (listings.csv)

105251	Simão Fonseca
105877	Margarida Pereira
105427	Francisco Rodrigues
104765	Margarida Carvalho

Grupo 14: Dallas (listings.csv)

105875	Rui Bernardo
104848	André Silva
105848	Inês Cabral
104935	Matilde Lemos

Grupo 15: Santiago (Chile) (listings.csv)

99369	Diogo Cancela
105351	Vasco Guerreiro
99240	David Serrão
98584	Antonio Coutinho
98573	Carlota Brejo

Grupo 16: Nashville (listings.csv)

104887	Melissa Mateus
104820	Leonor Pimentel
104588	Tomás Guia
105166	Pedro Rebelo

Grupo 17: Bristol (listings.csv)

104570	Luiz Vieira
103685	Ana Marta Escudeiro
105285	Ricardo Galvão

Grupo 18: Menorca (listings.csv)

105944	Miguel Gonçalves
99064	Simão Miguel
104939	Vasco Mestrinho

Grupo 19: Brussels (listings.csv)

104810	Leonor Pereira
106579	Ariana Meia-Via
104845	Diogo Marques

Grupo 20: Naples (listings.csv)

104787	Bernardo Arjones
100507	Fabiana Marques
100269	Beatriz de Moura
103954	Enderson Santos

Grupo 21: Denver (listings.csv)

104837	Vasco Martinho
105923	Gonçalo Duarte
104930	Gonçalo Verissimo
98991	Manuel Fonseca

Grupo 22: Belize (listings.csv)

105399	Alice Rocha
104990	Francisca Niehus
105840	Carlota Pereira