Universidade da Beira Interior
Departamento de Informática
Licenciatura em Inteligência Artificial e Ciência de Dados
© João C. Neves (jcneves@di.ubi.pt), 2024/25

Elementos de Inteligência Artificial e Ciência de Dados

Trabalho Prático

A PORDATA é uma base de dados estatísticos da Fundação Francisco Manuel dos Santos que reúne e organiza informação oficial sobre diversas áreas da sociedade portuguesa. Disponibilizada gratuitamente, apresenta dados objetivos e contextualizados sobre Portugal, municípios e países europeus, facilitando a análise e compreensão da realidade nacional.

Neste trabalho prático pretende-se realizar a recolha e análise de várias estatísticas sobre Portugal com vista a extrair conhecimento a partir das relações entre as mesmas. Para isto, será necessário realizar as seguintes fases.

- **Recolha de Dados**. Nesta fase pretende-se extrair os dados referentes a um conjunto de estatísticas disponíveis ao nível dos municípios [1]. O conjunto de estatísticas a considerar no trabalho deve ser escolhido em função dos temas que se pretendem analisar.
- **Integração de Dados**. Esta fase prevê a integração dos dados referentes a cada estatística num único conjunto de dados. Este conjunto deve estar estruturado no formato de tabela, sendo as primeiras colunas o Ano e Território e as restantes cada uma das estatísticas selecionadas. Um exemplo deste formato esta disponível neste link.
- **Análise Exploratória de Dados**. Nesta fase espera-se a realização da análise dos dados recolhidos com vista à determinação de características redundantes, valores em falta, existência de *outliers* e correlações entre as variáveis.
- **Limpeza e Preprocessamento de Dados**. Com base nas conclusões extraídas na fase anterior, pretende-se realizar a limpeza do conjunto de dados através da remoção de características desnecessárias, correção ou remoção de exemplos contendo *outliers* ou valores em falta.
- **Análise Descritiva**. Os dados recolhidos devem ser analisados com a ajuda de métodos de aprendizagem não supervisionada, de forma a encontrar ou confirmar padrões nos mesmos que suportem a extração de conhecimento a partir das várias estatísticas recolhidas.

No final, o conjunto de dados obtido deve ser usado para elaborar uma narrativa sobre os dados em forma de apresentação, ou seja, devem ser aplicadas técnicas de análise de dados, (com base na transformação, sumarização e visualização de dados) de forma a extrair conclusões sobre os mesmos, que por sua vez devem ser organizadas de forma a construir uma narrativa sobre os dados.

[1] https://prod2.pordata.pt/municipios

Elementos a Entregar:

No final do projeto, devem ser entregues através do Moodle os seguintes elementos:

- Repositório Git contendo o código Python usado nas diferentes fases do projeto
- Relatório organizado de acordo com as fases do projeto descrevendo as técnicas usadas em cada uma delas, a justificação para o seu uso, e os resultados obtidos. O relatório deve obrigatoriamente fazer referência à função ou script usada para implementar os processos descritos no mesmo, não sendo necessário incluir blocos de código no mesmo.
- Apresentação narrativa do conjunto de dados, contendo as principais conclusões que se podem extrair dos dados recolhidos e processados.

Avaliação:

A classificação final do projeto será obtida através da seguinte fórmula:

$$Nota_{TP1} = 0.8A + 0.2B,$$

sendo A o repositório e relatório desenvolvidoss, e B a apresentação narrativa do conjunto de dados.

O repositório e o relatório serão avaliados de acordo com as técnicas usadas em cada uma das fases, tendo em conta a sua correta aplicação e complexidade. Cada fase contribui para a classificação final de acordo com as seguintes percentagens:

- Recolha de Dados [10%]
- Integração de Dados [15%]
- Análise Exploratória de Dados [20%]
- Limpeza e Preprocessamento de Dados [15%]
- Análise Descritiva [40%]

O desempenho na construção da narrativa a partir do conjunto de dados obtido será medido pela relevância das conclusões obtidas e pela profundidade e complexidade das técnicas usadas para obter essas mesmas conclusões.

