Licenciatura em Ciência de Dados – 2022/23
Métodos de Aprendizagem Não Supervisionada
Enunciado do Trabalho de Grupo

Base de dados

1. A base de dados a utilizar neste desafio: STU_QQQ_5.csv
2. Variáveis de INPUT (utilizadas na PCA/clustering): a escolher pelo grupo.
3. Variáveis de PROFILE (utilizadas na caracterização dos clusters): a escolher pelo grupo.

Estrutura do relatório final

O relatório final deve ter a seguinte estrutura (secções):

1. **Introdução:** Deverá definir de um modo claro o objetivo da análise, o contexto do problema e a sua relevância substantiva, sempre na perspetiva de uma análise não supervisionada num contexto de tomada de decisão (e.g., contexto de recomendação de políticas educativas);
2. **Dados**: Caracterização sumária da base de dados (não mais do que dois parágrafos); a sua dimensão e as variáveis ativas (INPUT) e passivas (PROFILE) utilizadas na análise e na caracterização, respetivamente;
3. **Identificação das dimensões da análise**: A Análise em Componentes Principais utilizando as variáveis de INPUT. Dê um nome a cada uma das novas variáveis;
4. **Identificação da heterogeneidade na base de dados**: Utilizando as novas dimensões (componentes principais), proceda a uma análise de clustering das observações. Dê um nome a cada um dos clusters e proceda à sua caracterização utilizando as variáveis PROFILE;
5. **Conclusão:** Apresente um breve resumo dos resultados principais e uma breve discussão das principais implicações do estudo realizado no apoio à decisão;
6. **Bibliografia**
7. **Anexos**

Para além desta estrutura, devem respeitar-se as seguintes indicações:

1. Número de alunos por grupo: 4 ou 5;
2. O relatório deve ser um documento de análise e de síntese interpretativa. Toda a informação de natureza estritamente técnica deve ser remetida para anexos, que deverão estar claramente organizados, titulados e paginados. O relatório deve ter no máximo 15 páginas (excetuando os anexos e a folha de rosto que contém o título, nomes e números de aluno);
3. O número de secções e conteúdos devem manter-se de modo a permitir a avaliação relativa dos trabalhos, mas a designação das secções pode naturalmente mudar.

Após a definição do objetivo da análise – definir um problema não supervisionado num contexto de tomada de decisão – e identificação das variáveis de INPUT/PROFILE, a realização do trabalho de grupo processa-se em duas fases:

Fase I: Análises PCA

Esta fase é opcional, contudo, recomendável, pois permite garantir que o trabalho vai bem encaminhado. Nesta fase procederá à transformação de um conjunto de variáveis INPUT em agregados a serem utilizados como input numa análise de clustering. Deverá apenas enviar o script de R e resultados no corpo de um email. O script pode ter comentários que ajudam a suportar a decisão e que posteriormente serão discutidos no relatório. Concluída esta fase, as análises correspondentes à Secção 3 do relatório final devem estar concluídas.

Fase II: Envio do relatório final

1. As datas de entrega do trabalho são: fase I – até duas semanas depois da conclusão da lecture2; fase II – coincide com o dia do teste/Exame de 1ª época (até às 23:59);
2. Apenas o relatório final será avaliado;
3. O ficheiro zip, cujo nome deve incluir o último apelido de todos os elementos do grupo, deverá conter deverá conter **três ficheiros**: a base de dados final utilizada (formato csv com headers), o script de R que permite replicar e obter todos os resultados constantes no relatório final (e compatível com a base de dados enviada) e o pdf do relatório final. Se por alguma razão houver problemas de confidencialidade, utilize apenas uma parte aleatória da base de dados original, mas esta deverá coincidir com o script de R e resultados no relatório final.

José G. Dias & Jorge Sinval, 5 maio 2023