

**Curso:** Mestrado Integrado em Informática – Engenharia do Conhecimento  
**U.C.:** Descoberta do Conhecimento

Folha de Exercícios FE07	
Docente	Cristiana Neto
Tema	RapidMiner – K-means Clustering
Turma	PL
Ano Letivo	2019-20 – 2º Semestre
Duração da aula	2 horas

## 1. Parte I

- [1] O que significa o 'k' em k-Means clustering?
- [2] Como se identificam os *clusters*? Qual é o processo que o RapidMiner usa para definir e colocar as observações num determinado *cluster*?
- [3] O que revela a Centroid Table ao utilizador? Como se interpretam os valores dessa tabela?
- [4] Como é que a presença de *outliers* nos atributos de um *dataset* influencia a utilidade de um modelo de k-Means clustering? O que poderia ser feito para resolver este problema?

## 2. Parte II

- [1] Pense num problema que possa ser resolvido agrupando observações em *clusters*. Procure na internet um *dataset* que possa ser utilizado e aplicado a um modelo de k-Means. Sugestão: ir ao [website](https://archive.ics.uci.edu/ml/) da [UCI – Machine Learning Repository](https://archive.ics.uci.edu/ml/) e escolher um *dataset* cuja *Default Task* seja *Clustering*.
  - (a) Importe os dados para o RapidMiner. Não se esqueça de garantir que estes estejam no formato CSV. Execute a etapa de *Data Understanding*.
  - (b) Efectue a etapa de *Data Preparation*. Pode incluir componentes de inconsistência de dados, *missing values*, ou alteração do tipo de dados;
  - (c) Adicione um operador de k-means clustering ao *dataset* no RapidMiner e altere os parâmetros de acordo com a necessidade (sobretudo o valor **k**, para adequar ao problema em questão);
  - (d) Estude a *Centroid Table*, *Folder View*, e outras ferramentas de avaliação;
  - (e) Reporte todos os passos anteriores e as evidências encontradas, bem como de que forma o que foi encontrado permite responder ao problema inicial.

[2] Experimente o mesmo *dataset* com diferentes operadores de k-Means como o Kernel ou Fast. Em que medida diferem do modelo original. Estes operadores mudam os *clusters* originais? Se sim, em que medida?