ISCTE-IUL
Mestrado em Ciência de Dados

# Análise de Comportamento de utilizadores em E-commerce: uma abordagem não supervisionada

Trabalho de Grupo realizado no âmbito da Unidade Curricular de Processamento e Modelação de Big Data do 1º ano do Mestrado em Ciência de Dados

Diogo Freitas, 104841, MCD-LCD-A1
Diogo_Alexandre_Freitas@iscte-iul.pt

João Francisco Botas, 104782, MCD-LCD-A1
Joao_Botas@iscte-iul.pt

Miguel Gonçalves, 105944, MCD-LCD-A1
Miguel_Goncalves_Pereira@iscte-iul.pt

Ricardo Galvão, 105285, MCD-LCD-A1
Araujo_Galvao@iscte-iul.pt

23 de abril 2025
Versão 1.0.0 (Entrega Parcial)

## Índice

# 1. Descrição do problema

Para este projeto foi utilizada uma base de dados de eCommerce, proveniente de uma loja com múltiplas categorias e tipos de produtos. Na nossa análise consideramos os meses de outubro e novembro de 2019, os quais contêm registos detalhados (*logs*) das ações efetuadas pelos utilizadores no *site* da loja, tais como visualizações de produtos, adições ao carrinho e compras[1].

O foco principal da nossa análise é identificar possíveis padrões de consumo dos utilizadores, com base nas suas interações no *site*, de forma a agrupá-los segundo preferências e comportamentos semelhantes. Por exemplo, poderá emergir um grupo de clientes $X$ com maior interesse em produtos da categoria $\alpha$ de uma gama mais elevada (produtos com preço superior). Assim, o problema enquadra-se no domínio da **aprendizagem não supervisionada**, sendo abordado através de técnicas de *clustering*.

A componente de *Clustering* consiste na aplicação de diferentes algoritmos, como K-Means e Gaussian Mixture Models (GMM), sobre um conjunto de variáveis selecionadas a partir das interações dos utilizadores. Com esta abordagem, pretendemos responder a duas questões principais:

- Como podemos caracterizar os grupos formados (*clusters*), através das *features* a utilizar? É possível observar padrões e grupos coesos?
- Como varia a complexidade temporal dos algoritmos utilizados, especialmente quando aplicados a subconjuntos de dados com dimensões distintas?

Numa fase seguinte é então desenvolvido um sistema de recomendações simples com base em similaridade item-item. Embora esta abordagem seja independente do *clustering*, ambas as análises são complementares. Enquanto o *clustering* ajuda a revelar padrões de consumo e perfis de utilizador, a recomendação foca-se em associar/"agrupar" produtos semelhantes aos já visualizados ou comprados, a partir da co-ocorrência entre itens.

---

Para a entrega final ponderamos falar dos seguintes tópicos em cada uma das fases:

# 2. Preparação de dados

- Explicar o joining dos dados e formato a utilizar (parquet);
- Perceber o porquê de utilizarmos as "views" para o *Clustering*;
- Divisão em categorias e sub-categorias;
- Pivot table para o propósito do problema;
- (...).

# 3. Experiências e testes realizados

- Inicialmente para 10000 dados apenas para fins de testes;
- Incrementação de dados e testagem para o "*dataset full*";
- Tentar imaginar qual seria o tempo computacional se o conjunto de dados fosse 400% do disponível, por exemplo (curva de análise temporal);
- (...).

# 4. Resultados

- Comentar a formação dos *clusters*;
- Interpretar a análise de complexidade.

Link do GitHub com o desenvolvimento do projeto:

[1] Na secção Secção 2 serão detalhadas todas as assunções tomadas relativamente ao conjunto de dados utilizado

# Anexos

## Anexo A - (...)