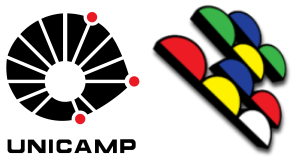


**UNIVERSIDADE ESTADUAL DE CAMPINAS**
**ENGENHARIA E ADMINISTRAÇÃO DE SISTEMAS DE BANCO DE DADOS**
**CT0611 - MINERAÇÃO DE DADOS**

**Docente Responsável: Prof. Dr. Julio C. dos Reis** [dosreis@unicamp.br]

**Monitor:** Eryck Pedro da Silva [eryck@unicamp.br]

# Atividade 02: Agrupamento

## Objetivo

Esta atividade objetiva aplicar  conceitos de Mineração de Dados, com foco em algoritmos de agrupamento, utilizando o método K-Means. Visamos segmentar clientes com base em seus padrões de consumo e perfis demográficos. Serão avaliadas as habilidades de:

- Leitura e tratamento de dados;
- Visualização e exploração de dados;
- Implementação e interpretação do K-Means;
- Avaliação dos agrupamentos obtidos;
- Discussão dos resultados e recomendações.

## Cenário

Uma empresa de marketing deseja segmentar seus clientes para criar campanhas mais eficientes. Com base no dataset "marketing_campaign.csv", contendo informações como idade, renda e hábitos de compra, o(a)s aluno(a)s devem identificar diferentes grupos de consumidores, auxiliando na tomada de decisão para campanhas personalizadas.

## Tarefas

Desenvolva sua solução com base nas instruções das seguintes tarefas em detalhes a seguir:

### 1. Leitura e Tratamento dos Dados (20%)

- Carregar o dataset "marketing_campaign.csv" usando Pandas.

- Exibir informações sobre o dataset (número de linhas e colunas, tipos de dados, valores ausentes).
- Tratar valores ausentes e outliers, justificando as escolhas.
- Normalizar atributos numéricos para padronização.

## 2. Exploração e Visualização dos Dados (20%)

- Criar histogramas para explorar a distribuição de idade e renda dos clientes.
- Criar um gráfico de dispersão entre renda e gasto total.
- Utilizar boxplots para visualizar distribuições de consumo em diferentes segmentos.

## 3. Uso e Execução do Algoritmo de Agrupamento (25%)

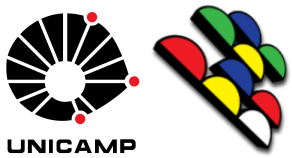
- Usar o algoritmo K-Means implementado no **Scikit-Learn**:
    - Aplicar o método do cotovelo e testar para valores variados de *K* que façam sentido com o resultado do método
    - Instanciar versões distintas do parâmetro *init* para comparar as versões KMeans e KMeans++
- Gerar visualização dos clusters formados.

## 4. Avaliação dos Agrupamentos Obtidos (20%)

- Utilizar o índice de *Silhueta* para medir a coesão dos clusters.
- Interpretar os agrupamentos formados e analisar diferenças entre eles.

## 5. Discussão e Recomendações (15%)

- Resumir os padrões identificados nos agrupamentos.
- Sugerir estratégias de *marketing* baseadas nos resultados.
- Discutir limitações do método e possíveis melhorias.




# Arquivos e Ferramentas

- **Arquivo:** "marketing_campaign.csv"
- **Ferramentas:** Python 3.10+, Pandas, NumPy, Matplotlib, Scikit-Learn

# Submissão

- Esta tarefa pode ser realizada em dupla.
- Apenas um arquivo por dupla deve ser submetido.
- Você deve entregar um arquivo Jupyter Notebook (.ipynb), feito e baixado pelo *Google Colab*, subdividido de forma correspondente às Tarefas propostas nesta atividade.
- Apenas um integrante da equipe deve submeter o arquivo com a solução documentada.
- O arquivo deve ser nomeado da seguinte forma:

atividade02-agrupamento-<nome_dos_integrantes>.ipynb

- [Exemplo: `atividade02-agrupamento-rafael-juliana.ipynb`];
- Esta entrega tem peso de **40%** da nota final desta disciplina.
- A entrega deve ser feita até **05/04/2025 (Sábado)** às 23:59 via Moodle.

# Critérios de Avaliação

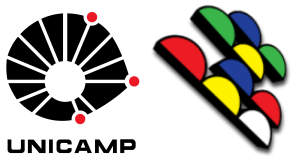
- **Leitura e Tratamento de Dados (20%)**: Correta manipulação e tratamento dos dados, justificando decisões.

- **Visualização e Exploração (20%)**: Uso adequado de gráficos e interpretação dos padrões.

- **Uso e execução do K-Means (25%)**: Execução correta dos algoritmos (KMeans e KMeans++) e definição dos diferentes *K*s abordados, com justificativa para o melhor *K*.

- **Avaliação dos Agrupamentos(20%)**: Métricas corretamente calculadas e analisadas, comparando diferentes instâncias para diferentes parâmetros (valores de k e entre KMeans e KMeans++)

- **Discussão e Recomendações (15%)**: Argumentação clara e bem fundamentada sobre os resultados.



# Informações sobre o Dataset

A Análise de Personalidade do Cliente é uma análise detalhada dos clientes ideais de uma empresa. Ela auxilia o negócio a compreender melhor seus clientes, facilitando a adaptação de produtos às necessidades, comportamentos e preocupações específicas de diferentes tipos de clientes.

A análise de personalidade do cliente permite que uma empresa ajuste seus produtos com base nos clientes-alvo de diferentes segmentos. Por exemplo, em vez de gastar dinheiro promovendo um novo produto para todos os clientes do banco de dados da empresa, a empresa pode analisar qual segmento de clientes tem maior probabilidade de comprá-lo e direcionar a divulgação apenas para esse segmento específico.

No dataset que vocês vão utilizar, há um total de 29 colunas. O objetivo principal é identificar segmentos de preferências de compras dos consumidores por meio de técnica de agrupamento. As colunas podem ser entendidas da seguinte forma:

*Pessoas*

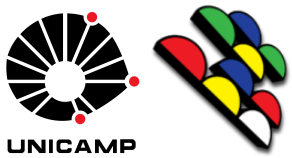
- **ID**: Identificador único do cliente
- **Year_Birth**: Ano de nascimento do cliente
- **Education**: Nível de escolaridade do cliente
- **Marital_Status**: Estado civil do cliente
- **Income**: Renda anual familiar do cliente
- **Kidhome**: Número de crianças no domicílio do cliente
- **Teenhome**: Número de adolescentes no domicílio do cliente
- **Dt_Customer**: Data de cadastro do cliente na empresa
- **Recency**: Número de dias desde a última compra do cliente
- **Complain**: 1 se o cliente fez uma reclamação nos últimos 2 anos, 0 caso contrário

*Produtos*

- **MntWines**: Quantia gasta com vinhos nos últimos 2 anos
- **MntFruits**: Quantia gasta com frutas nos últimos 2 anos
- **MntMeatProducts**: Quantia gasta com carne nos últimos 2 anos
- **MntFishProducts**: Quantia gasta com peixes nos últimos 2 anos
- **MntSweetProducts**: Quantia gasta com doces nos últimos 2 anos
- **MntGoldProds**: Quantia gasta com ouro nos últimos 2 anos

*Promoção*

- **NumDealsPurchases**: Número de compras feitas com desconto
- **AcceptedCmp1**: 1 se o cliente aceitou a oferta na 1ª campanha, 0 caso contrário
- **AcceptedCmp2**: 1 se o cliente aceitou a oferta na 2ª campanha, 0 caso contrário
- **AcceptedCmp3**: 1 se o cliente aceitou a oferta na 3ª campanha, 0 caso contrário
- **AcceptedCmp4**: 1 se o cliente aceitou a oferta na 4ª campanha, 0 caso contrário
- **AcceptedCmp5**: 1 se o cliente aceitou a oferta na 5ª campanha, 0 caso contrário
- **Response**: 1 se o cliente aceitou a oferta na última campanha, 0 caso contrário




*Local*

- **NumWebPurchases**: Número de compras feitas pelo site da empresa
- **NumCatalogPurchases**: Número de compras feitas por catálogo
- **NumStorePurchases**: Número de compras feitas diretamente em lojas físicas
- **NumWebVisitsMonth**: Número de visitas ao site da empresa no último mês