
# Escola Politécnica de Pernambuco

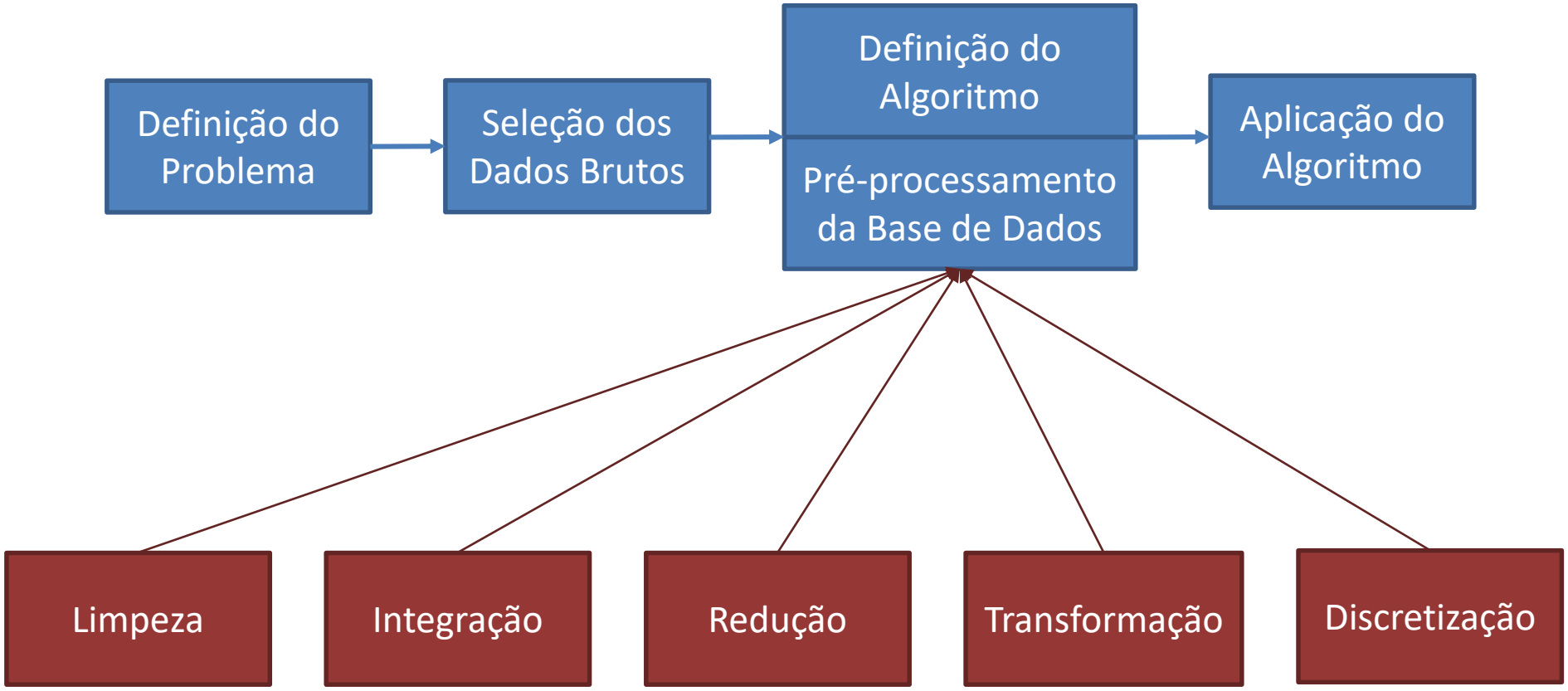
*Engenharia da Computação*

## Mineração de Dados
## Aula 4

Prof. Dr. Alexandre Maciel

***amam@ecomp.poli.br***

# PRÉ-PROCESSAMENTO DOS DADOS



# LIMPEZA DOS DADOS

- Imputar valores AUSENTES

- Suavizar RUÍDOS

- Identificar OUTLIERS

- Corrigir INCONSISTÊNCIAS

# LIMPEZA DOS DADOS

- **Dados AUSENTES**

| CPF | Nome | Sexo | Data_nasc | Est_civil | Num_dep | Renda | Despesa | Tp_res | Bairro_res |
|---|---|---|---|---|---|---|---|---|---|
| 99999999999 | José | 1 | 5/5/1289 | C | 1 | | 1000 | P | Centro |

  - Espaço em branco ou símbolo (?)
  - Implica em perda relevante
  - Imputação não deve enviesar a base
  - Alguns algoritmos não conseguem trabalhar
  - Tratamento incorreto pode promover erro

# LIMPEZA DOS DADOS

- **Imputar valores AUSENTES:**
  1. Ignorar objeto.
  2. Imputar manualmente.
  3. Usar constante global.
  4. Imputar por similaridade ou distância entre objetos.
  5. Imputar de acordo com última observação.
  6. Usar média ou moda do atributo.
  7. Usar média ou moda de todos objetos da mesma classe.
  8. Usar modelos preditivos.

# LIMPEZA DOS DADOS

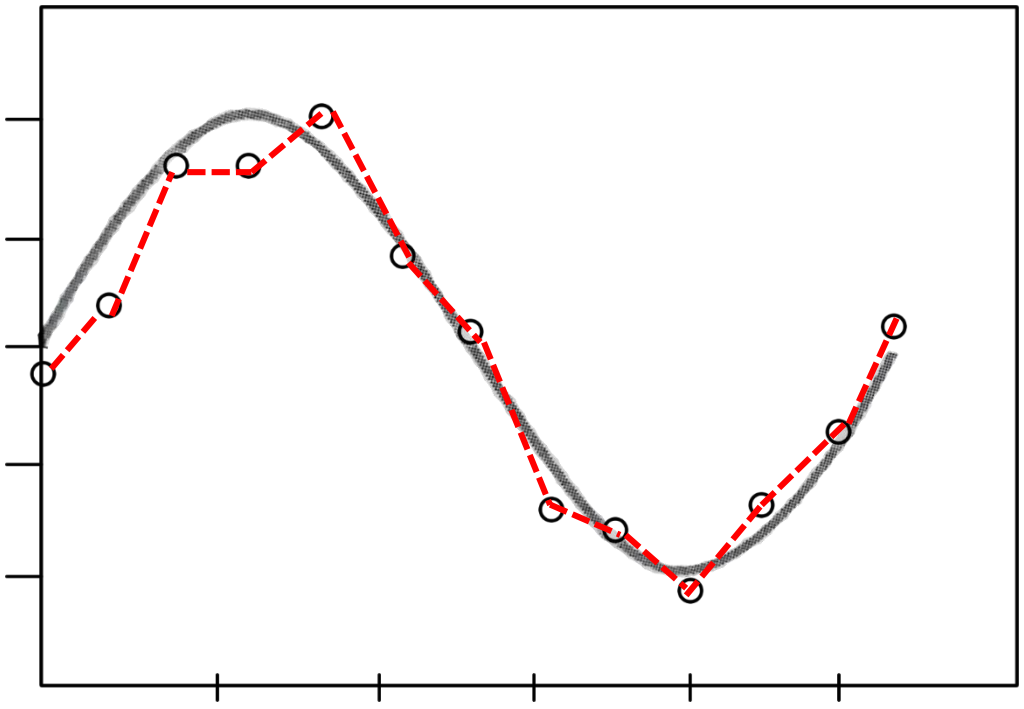
- **Dados RUIDOSOS**

| CPF | Nome | Sexo | Data_nasc | Est_civil | Num_dep | Renda | Despesa | Tp_res | Bairro_res |
|---|---|---|---|---|---|---|---|---|---|
| 99999999999 | José | 1 | 5/5/1289 | C | 1 | | 1000 | P | Centro |

  - Erros de medição e entrada do dados
  - Erros acumulados
  - Distorção da realidade
  - Não existe um padrão consistente
  - Não deve ser considerado pelo algoritmo

# LIMPEZA DOS DADOS

- **Dados RUIDOSOS**



Função seno com ruído

# LIMPEZA DOS DADOS

- **Suavizar dados RUIDOSOS:**

1. Encaixotamento

   - Mesma largura:

| 3 | 11 | 14 | 16 | 19 | 23 |
|---|---|---|---|---|---|

Caixa 1 [3,13]:

| 3 | 11 |
|---|---|

Caixa 2 [13,23]:

| 14 | 16 | 19 | 23 |
|---|---|---|---|

Suavização pela média da caixa:

| 7 | 7 |
|---|---|

| 18 | 18 | 18 | 18 |
|---|---|---|---|

   - Mesma frequência:

| 79 | 141 | 185 | 199 | 220 | 341 |
|---|---|---|---|---|---|

Caixa 1 [79,185]:

| 79 | 141 | 185 |
|---|---|---|

Caixa 2 [199,341]:

| 199 | 220 | 341 |
|---|---|---|

Suavização pelos extremos:

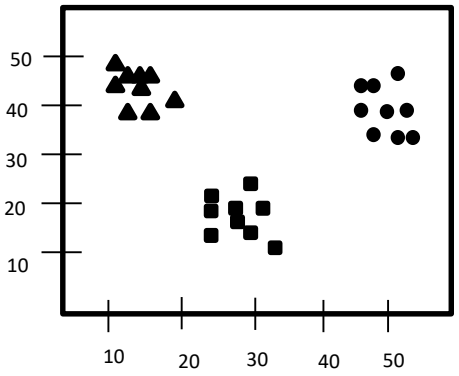
| 79 | 185 | 185 |
|---|---|---|

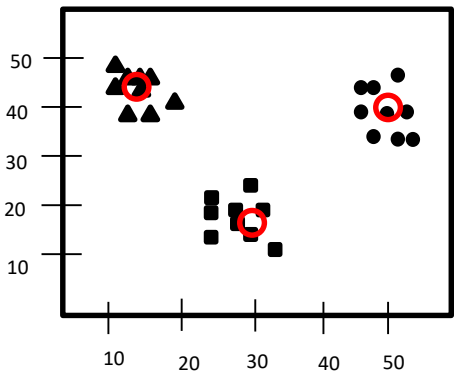
| 199 | 199 | 341 |
|---|---|---|

# LIMPEZA DOS DADOS

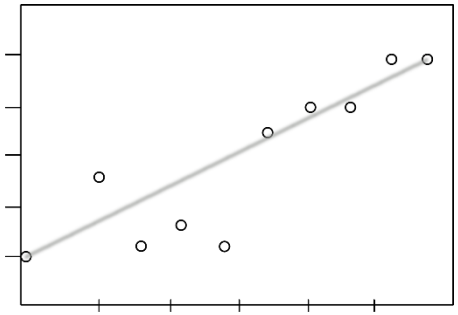
- **Suavizar dados RUIDOSOS:**

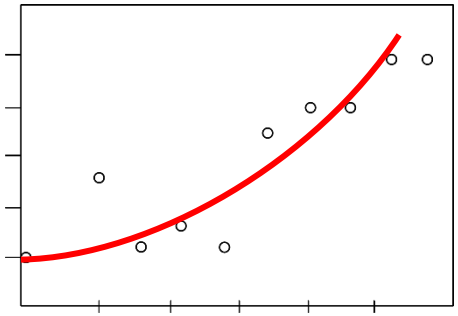
2. Agrupamento




3. Aproximação





# LIMPEZA DOS DADOS

- **Dados INCONSISTENTES**

| CPF | Nome | Sexo | Data_nasc | Est_civil | Num_dep | Renda | Despesa | Tp_res | Bairro_res |
|---|---|---|---|---|---|---|---|---|---|
| 99999999999 | José | X | 5/5/1289 | C | 1 | | 1000 | P | Centro |

  - Discrepância em relação a outros dados.
  - Influencia na validade, utilidade e integridade da aplicação.
  - Diferentes nomes de atributos em diferentes sistemas.

- **Corrigir dados INCONSISTENTES:**
  - Análise manual.
  - Utilização de gráficos.
  - Participação de especialista de domínio.

# INTEGRAÇÃO DE DADOS

- **REDUNDÂNCIA ...** pode ser obtido por vários atributos da base.

- **DUPLICIDADE ...** atributos aparecem repetidos na base.

- **CONFLITOS ...** para mesma entidade, diferentes valores na base.

# REDUÇÃO DOS DADOS

- SELEÇÃO de atributos
- COMPRESSÃO de atributos
- AMOSTRAGEM dos dados
- APROXIMAÇÃO  de dados
- DISCRETIZAÇÃO

# Seleção de atributos

- **Redução horizontal:**

1. Segmentação do conjunto de dados.

```
SELECT * FROM CLIENTE WHERE TP_RES = 'P';
```

2. Eliminação direta dos casos.

```
DELETE FROM CLIENTE WHERE TP_RES <> 'P';
```

3. Agregação de informações.

```
SELECT SUM(VALOR) FROM PEDIDO WHERE CPF = '03274271403';
```

# Seleção de atributos

- **Redução vertical:**
  - Escolha de subconjunto relevantes de atributos.
  - Exclusão ou combinação de atributos.
  - Motivação:
    - Pode conduzir a modelos de conhecimento mais concisos.
    - Otimizar tempo de processamento.
    - A exclusão de um atributo é muito mais relevante em termos de tamanho do conjunto de dados do que de um registro.

# Seleção de atributos

- **Redução vertical**

  1. ***Eliminação direta de atributos***

     - Heurísticas:

       – Eliminar atributos com valores constantes
         » País...

       – Eliminar atributos que sejam identificadores
         » Nome, CPF ...

# Compressão de atributos

- **Redução vertical:**

  ***2. Análise de componentes principais***

  – Técnica de baixo custo computacional aplicável a dados numéricos.

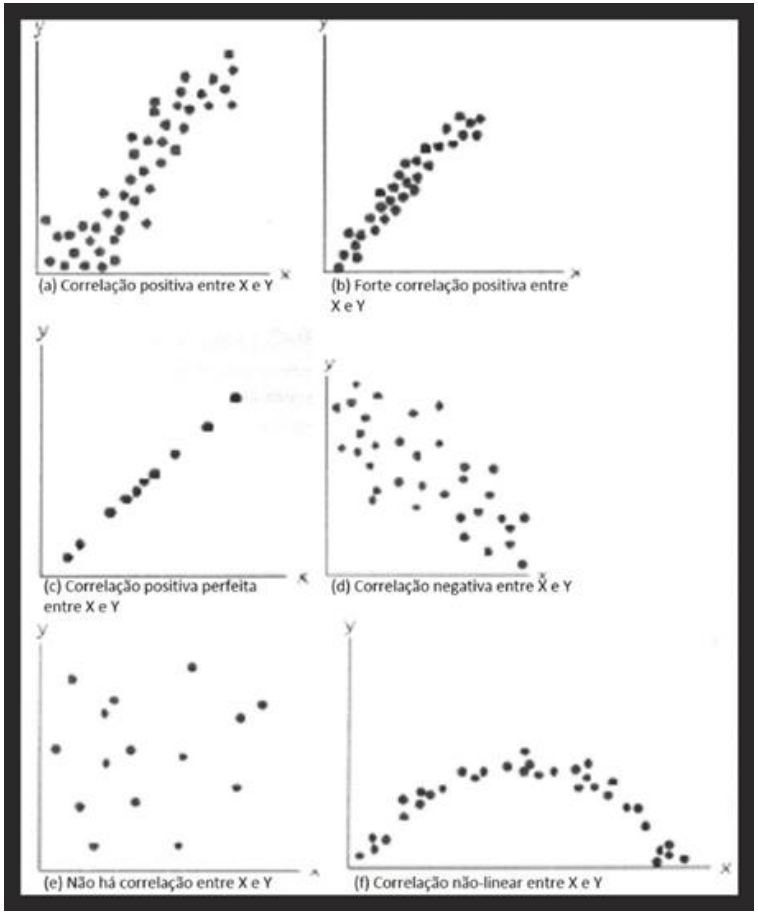
  – Consiste na transformação dos dados para um novo espaço com dimensão inferior ao original.

  – O novo espaço fica caracterizado por um novo conjunto de eixos, ortonormais entre si, ordenados em ordem decrescente de variância.

# Compressão de atributos

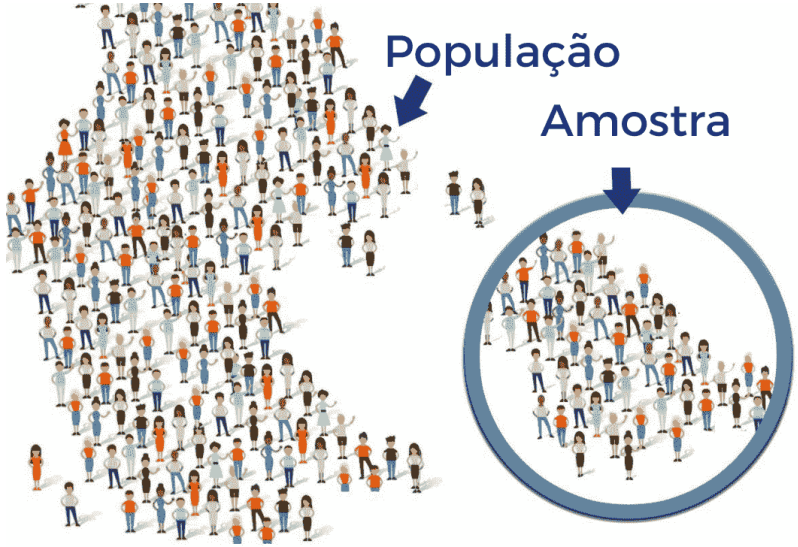
- **Redução vertical:**

  ***3. Análise de correlação***



# Amostragem dos Dados

- **Aleatória sem substituição**
  - n<m objetos retirados.
- **Aleatória com substituição**
  - objetos retirados voltam à base.
- **Sistemática**
  - Ordenação e critério para retirada.
- **Por grupo**
  - Para bases organizadas por grupos.
- **Estratificada**
  - Mantida proporção de classes.



# TRANSFORMAÇÃO DOS DADOS

- PADRONIZAÇÃO

- CODIFICAÇÃO

- NORMALIZAÇÃO

- PARTIÇÃO

# PADRONIZAÇÃO DOS DADOS

- **Capitalização**
  - Maiúscula, minúscula ou ambos.
- **Caracteres especiais**
  - Ferramentas sensíveis ao conjunto de caracteres.
- **Formato dos atributos**
  - Datas, CPF, abreviações...
- **Conversão de unidades**
  - Centímetros-metros, quilômetros por hora-milhas por hora, ...

# CODIFICAÇÃO DOS DADOS

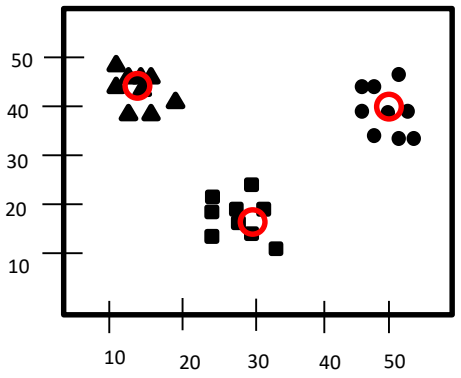
- **Numérica → Categórica**
  - Mapeamento direto
    - 1 -> M
    - 0 -> F
  - Mapeamento em intervalos (discretização)

| Intervalo | Frequência |
|---|---|
| 1000 \|- 1600 | 3 |
| 1600 \|- 4400 | 5 |
| 4400 \|- 5400 | 2 |

***Comprimento determinado pelo usuário***

| Intervalo | Frequência |
|---|---|
| 1000 \|- 2000 | 4 |
| 2000 \|- 3000 | 1 |
| 3000 \|- 4000 | 2 |
| 4000 \|- 5000 | 3 |

***Comprimento em intervalos iguais***




***Divisão por clusterização***

# CODIFICAÇÃO DOS DADOS

- **Categórica -> Numérica**
  - Representação binária

| Valores Originais | Rep. Bin |
|---|---|
| Casado | 001 |
| Solteiro | 010 |
| Viúvo | 100 |
| Divorciado | 011 |
| Outro | 110 |

***Econômica***

| Valores Originais | Rep. Bin |
|---|---|
| Casado | 00001 |
| Solteiro | 00010 |
| Viúvo | 00100 |
| Divorciado | 01000 |
| Outro | 10000 |

***1-N***

| Valores Originais | Rep. Bin |
|---|---|
| Fraco | 0001 |
| Regular | 0011 |
| Bom | 0111 |
| Ótimo | 1111 |

***Temperatura***

# NORMALIZAÇÃO DOS DADOS

- **Ajustar escala de valores, de cada atributo, de modo que sejam restritos a pequenos intervalos.**
  - [-1;1] ou [0;1]

- **Evitar que alguns atributos, por apresentarem uma escala de valores maior, influenciem nos algoritmos de aprendizagem de máquina.**
  - Redes neurais ou métodos baseados em distância.

# Transformação dos dados

- Normalização Max-Min: $A' = \frac{A - Min}{Max - Min}$

| CPF | Despesas Valores Originais | Despesas Normalizadas |
|---|---|---|
| 99999999999 | 1000 | 0,14 |
| 11111111111 | 2000 | 0,43 |
| 33333333333 | 3000 | 0,71 |
| 55555555555 | 1500 | 0,29 |
| 22222222222 | 1500 | 0,29 |
| 00000000000 | 1000 | 0,14 |
| 88888888888 | 3000 | 0,71 |
| 77777777777 | 500 | 0,00 |
| 66666666666 | 4000 | 1,00 |
| 44444444444 | 1000 | 0,14 |

# Transformação dos dados

- Normalização Z-score: $A' = \frac{A-\mu}{\sigma}$

- $\mu = 1850$
- $\sigma = 1131$

| CPF | Despesas Valores Originais | Despesas Normalizadas |
|---|---|---|
| 99999999999 | 1000 | -0,75 |
| 11111111111 | 2000 | 0,13 |
| 33333333333 | 3000 | 1,02 |
| 55555555555 | 1500 | -0,31 |
| 22222222222 | 1500 | -0,31 |
| 00000000000 | 1000 | -0,75 |
| 88888888888 | 3000 | 1,02 |
| 77777777777 | 500 | -1,19 |
| 66666666666 | 4000 | 1,90 |
| 44444444444 | 1000 | -0,75 |

# Transformação dos dados

- Normalização por escala decimal: $A' = \frac{A}{10^j}$

- $\mathbf{j = 4}$

| CPF | Despesas Valores Originais | Despesas Normalizadas |
|---|---|---|
| 99999999999 | 1000 | 0,10 |
| 11111111111 | 2000 | 0,20 |
| 33333333333 | 3000 | 0,30 |
| 55555555555 | 1500 | 0,15 |
| 22222222222 | 1500 | 0,15 |
| 00000000000 | 1000 | 0,10 |
| 88888888888 | 3000 | 0,30 |
| 77777777777 | 500 | 0,05 |
| 66666666666 | 4000 | 0,40 |
| 44444444444 | 1000 | 0,10 |

# PARTIÇÃO DOS DADOS

- ***Holdout***:
  - Treino = $p$ ; Teste = $1 - p$
- ***K-fold Cross Validation:***
  - Divisão do conjunto em $K$ subconjuntos, com $N$ elementos.
- ***Stratifield K-fold Cross Validation:***
  - Considera a proporção de exemplos em cada uma das classes.
- ***Bootstrap***:
  - Conjunto de treinamento gerado a partir de $N$ sorteios aleatórios.
  - Repetido várias vezes a fim de estimar a média de desempenho.

# DINÂMICA

- Vamos rodar o pré-processamento de dados do seu projeto...