Extração e Preparação de Dados

Aula 07 – Dados Ausentes: Tratamento Univariado

IBEMEC - 2026.1 | Prof. Luís Aramis

ibmec

Quem sou eu?

Professor: Luís Aramis dos Reis Pinheiro.

- **Doutorado e Mestrado** em **Ciências Mecânicas** – **UnB – CAPES 7**
- **Graduação** em Licenciatura em Física – **UNIFAP**

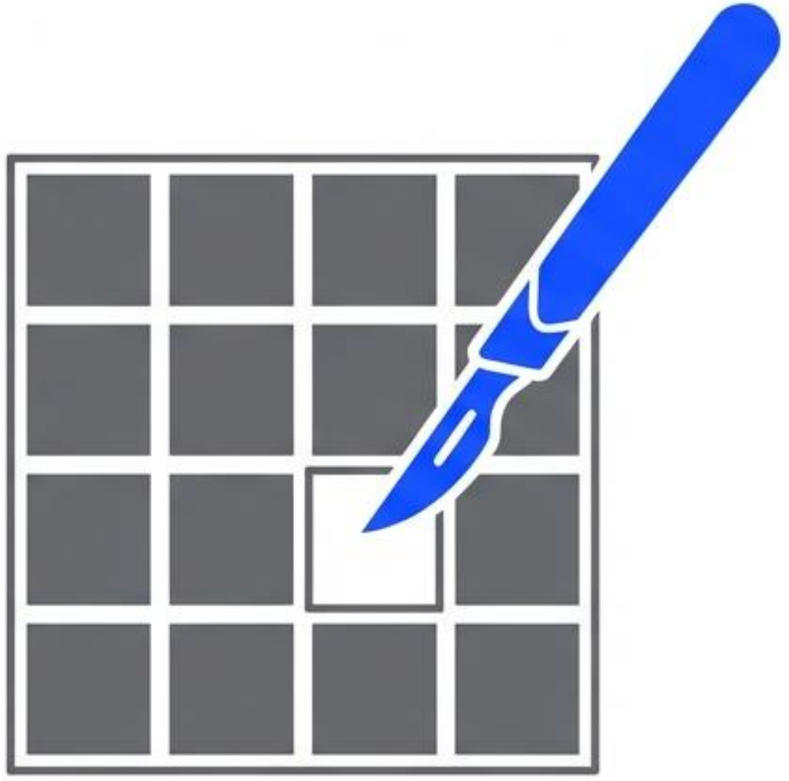
luis.pinheiro@professores.ibemec.edu.br

(96) 99907-5819

@l_aramis

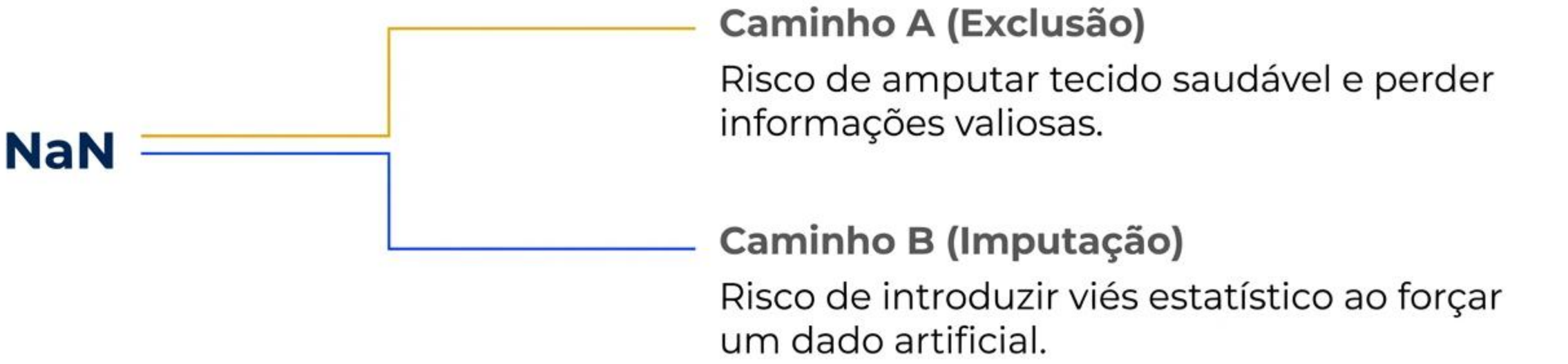
Dados Ausentes: Tratamento Univariado

- Extração e Preparação de Dados (IBM8915)
- A Decisão Crítica: Descartar ou Preencher?
- 05/03 (Quinta-feira)



O Dilema do Cirurgião de Dados

A presença de valores nulos (NaN) exige uma escolha analítica binária.

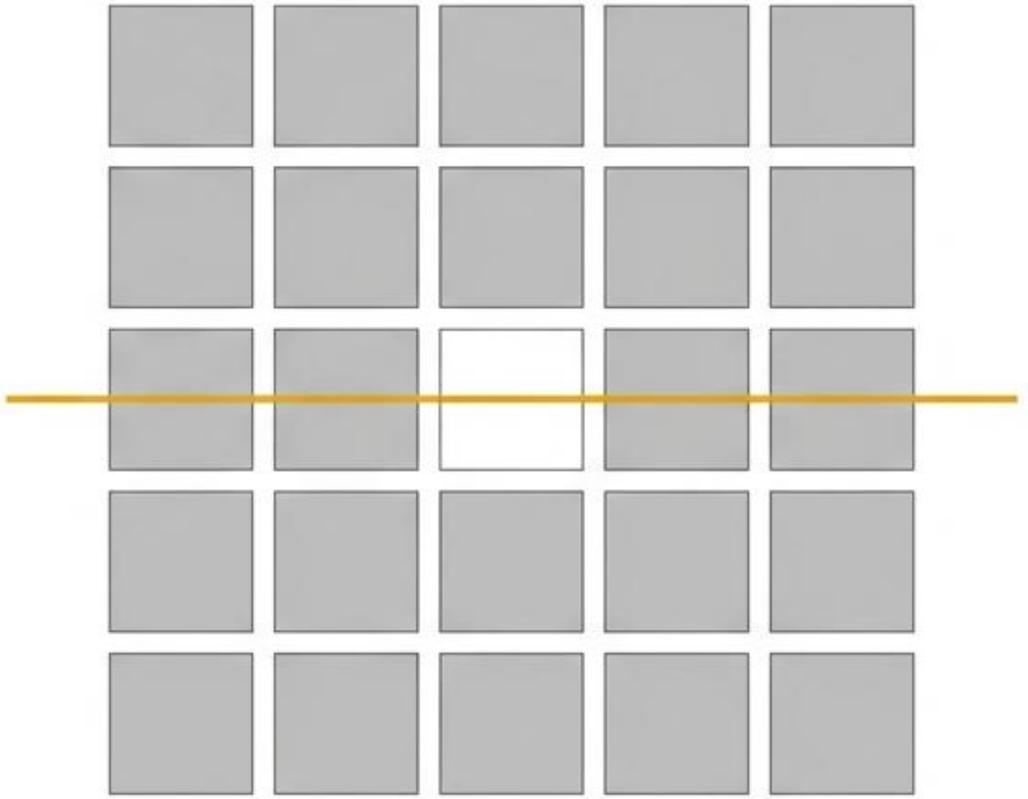


- A decisão depende do mecanismo de ausência (MCAR, MAR, MNAR) diagnosticado na Aula 06.

A Faca: Exclusão com dropna()

- O método dropna() é a ferramenta de remoção nativa do Pandas.
- **Comportamento Padrão**: Um corte cego. Exclui qualquer linha que contenha pelo menos um valor nulo.
- **Exclusão de Colunas**: Utilizando axis=1, descartamos o atributo inteiro.
- **O Perigo**: Aplicado sem parâmetros, pode dizimar um dataset por causa de poucos valores isolados.

```
df_limpo = df.dropna()
```



Precisão Cirúrgica: Preservando Dados

```
df.dropna(how='all')
```

Remove a linha apenas se todas as colunas forem nulas (óbito total do registro).

```
df.dropna(thresh=10)
```

Define um limite mínimo de sobrevida. Exige um número específico de colunas preenchidas.

```
df.dropna(subset=['ID_Paciente'])
```

Foca o corte apenas em colunas vitais específicas.

Negligência: O Perigo do Preenchimento Global

- Imputar significa substituir o valor ausente por um dado estimado.
- O método primário é o fillna().
- O Erro Amador: Executar df.fillna(0) em todo o dataset de forma global.
- A Consequência: Corrupção de variáveis categóricas (inserção de zeros em colunas de nomes ou textos).

ID	Nome	Idade
001	**0**	30

Curativo Específico: Dicionários no fillna()

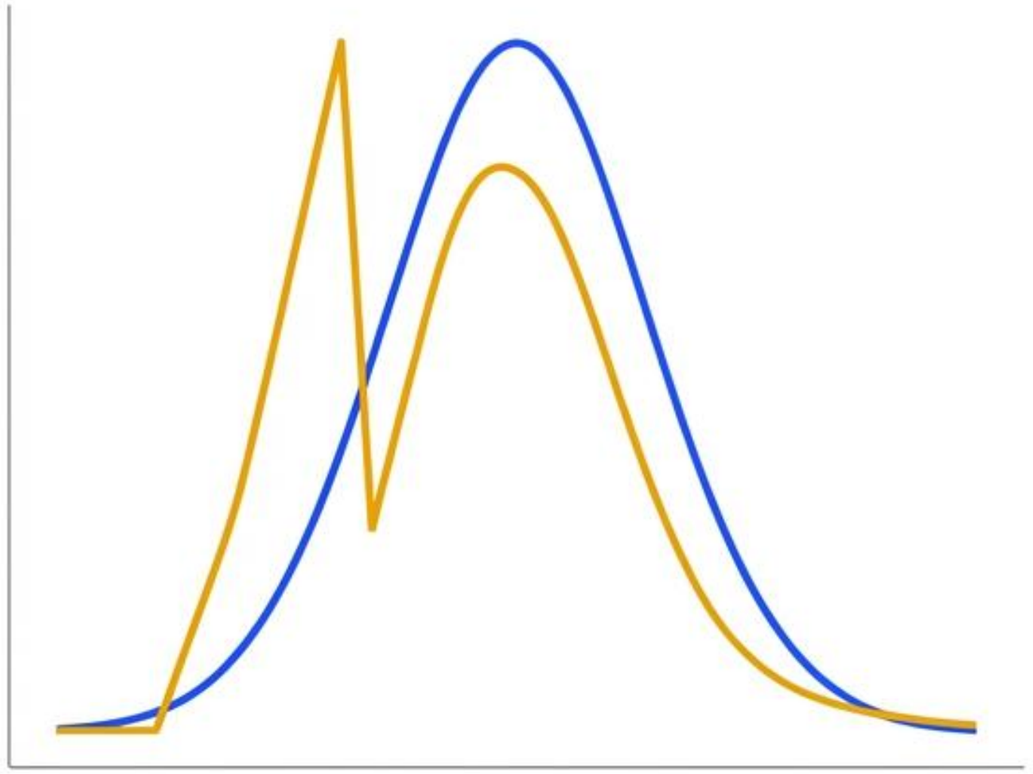
- O Pandas aceita a passagem de Dicionários Python como argumento.
- Permite definir uma regra de preenchimento exclusiva para cada coluna.
- Evita a contaminação cruzada de tipos de dados (dtypes).
- Mantém o rastreamento técnico (metadados) impecável.

```
curativos = {'Idade': 0, 'Cabine': 'Desconhecida'}

df.fillna(value=curativos, inplace=True)
```

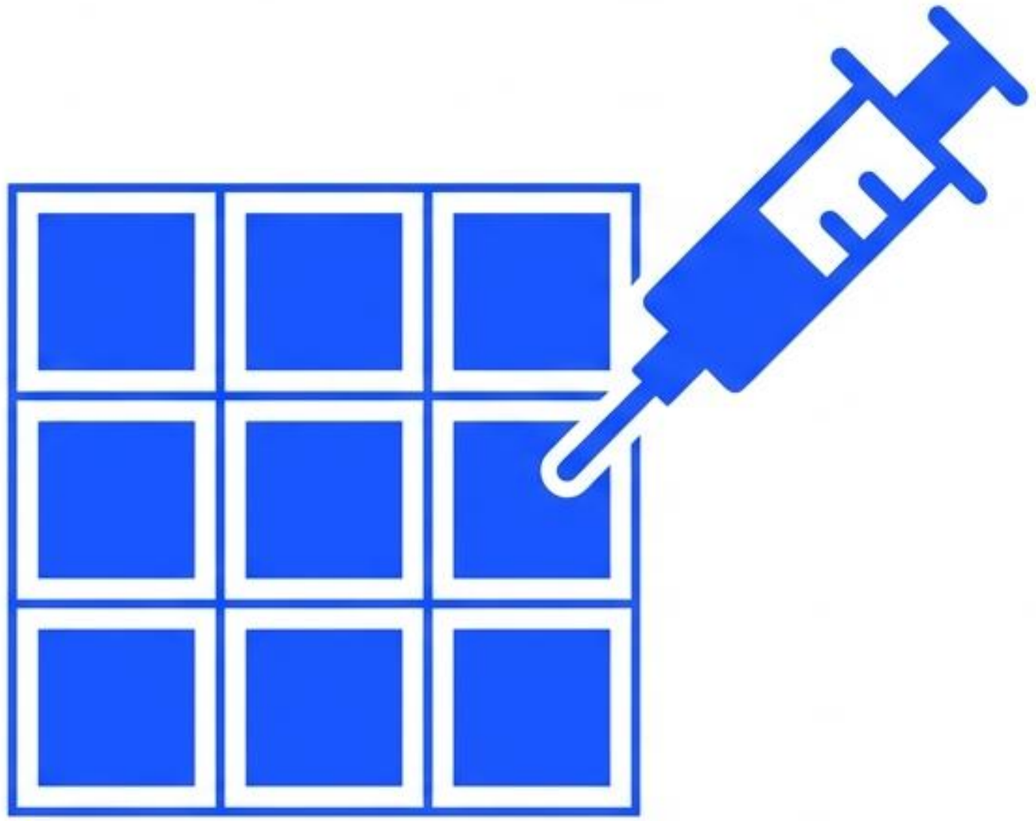
Limitações de Valores Fixos

- Hardcoding (valores fixos como 0) distorce a realidade estatística.
- Zero não é uma idade válida e altera drasticamente a média real do dataset.
- O curativo ideal não é um número arbitrário, mas uma inferência lógica.
- Próximo Passo: Imputação com inteligência estatística.



A "Seringa Inteligente": Imputação Estatística

- Transição de preenchimento manual para preenchimento estatístico.
- A relação matemática por trás do método **.fillna()**.



Imputação Dinâmica com .fillna()

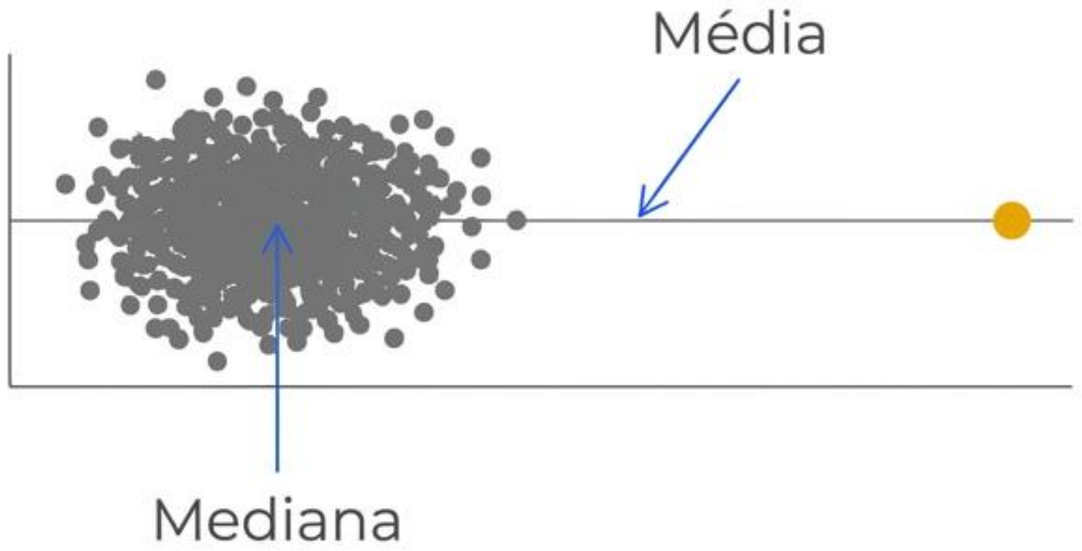
- **Abordagem cega:** Risco grave de distorção da realidade.
- **Abordagem estatística:** Preenchimento baseado na tendência central.
- **Cálculo automatizado:** O Pandas ignora os NaN ao calcular a média.

```
df['coluna'].fillna(0) # Abordagem Cega

df['coluna'].fillna(df['coluna'].mean()) # Abordagem Inteligente
```

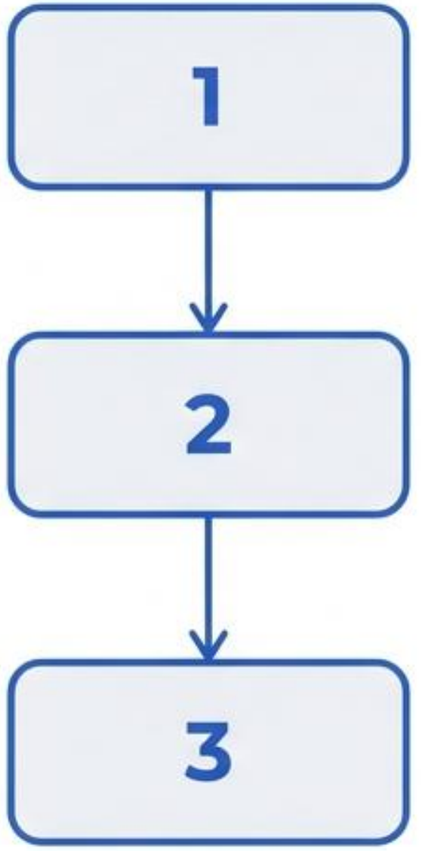
Média vs. Mediana: O Fator Outlier

- **Média**: Altamente sensível a valores extremos (Outliers).
- **Mediana**: Medida de posição robusta, corta o dataset exatamente ao meio.
- Regra de Ouro: Alta variância exige o uso da mediana.



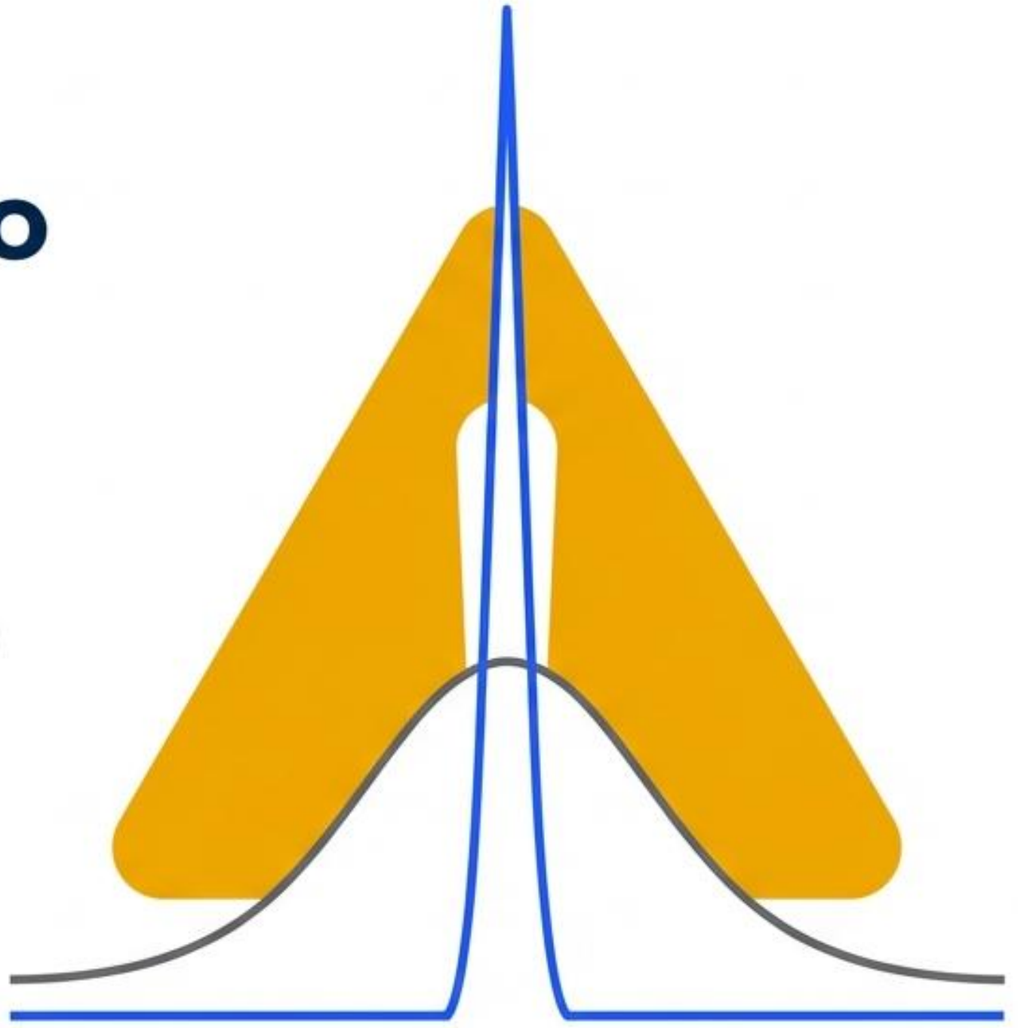
Hands-on: A Cirurgia de Dados (Lab 05)

- Passo 1: Amputação. Descartar colunas com > 70% de dados ausentes (dropna).
- Passo 2: Diagnóstico. Avaliar a presença de outliers nas colunas restantes.
- Passo 3: Intervenção. Preencher vazios utilizando fillna() com a medida estatística adequada.



O Risco Oculto do Tratamento Univariado

- **Cegueira de Correlação:** Analisa a coluna em completo isolamento.
- **Distorção da Variância:** Inserir a mesma média repetidas vezes esmaga a curva de distribuição.
- **Viés Estatístico:** Falsa confiança em modelos de Machine Learning futuros.



Versionamento e Evolução

- Commit do notebook lab_05_imputation.ipynb.
- Mensagens de commit semânticas e profissionais.
- Próxima Aula: Imputação Multivariada (O algoritmo KNN).

```
git add lab_05_imputation.ipynb
git commit -m "feat: imputacao univariada com media e moda"
git push origin main
```

Dúvidas?


CRIATIVIDADE
INOVAÇÃO
¡DEIA!
COMAMUNIDADE
CRIATIVIDADE
BRAINSTORM
INOVAÇÃO

Ibmec

A direção do amanhã.

/ibmec