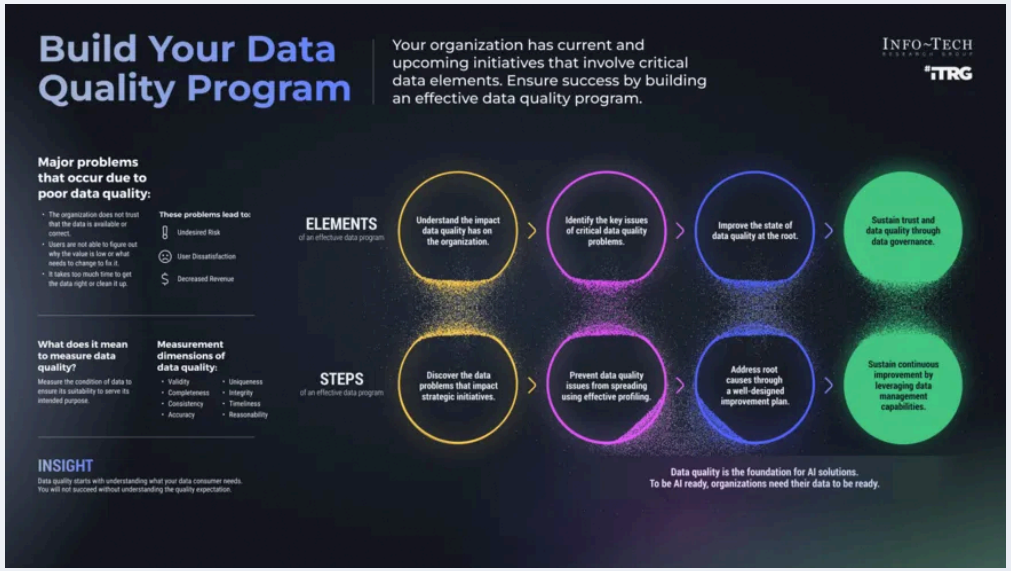
Validação, Viés e Qualidade dos Dados

Garantindo a confiabilidade das suas análises



Aula 09 - UC2

"Um dado ruim é pior do que dado nenhum."

Objetivos da Aula

Compreender a Importância da Validação de Dados

Entender por que a validação é o último controle de qualidade antes da análise e como dados ruins podem comprometer resultados e decisões.

Identificar e Corrigir Problemas Comuns em Datasets

Aprender a detectar e tratar dados ausentes, duplicatas, inconsistências e outliers usando técnicas e ferramentas do Pandas.

Reconhecer Diferentes Tipos de Viés em Dados

Identificar viés de amostra, viés histórico e viés de confirmação, compreendendo seu impacto nas análises e algoritmos.

Aplicar Técnicas de Prevenção de Viés e Documentação

Implementar estratégias para mitigar viés e criar documentação transparente sobre os processos de validação e limpeza realizados.

Ao final desta aula, você estará apto a garantir a qualidade e confiabilidade dos dados em seus projetos de análise.

Introdução e Contexto

> *"Um dado ruim é pior do que dado nenhum."*

A Validação como Último Controle de Qualidade

A validação de dados representa a última linha de defesa antes da análise. É o processo que garante que os dados utilizados são confiáveis e adequados para o propósito pretendido. Sem essa etapa crucial, corremos o risco de basear nossas conclusões em informações incorretas ou tendenciosas.

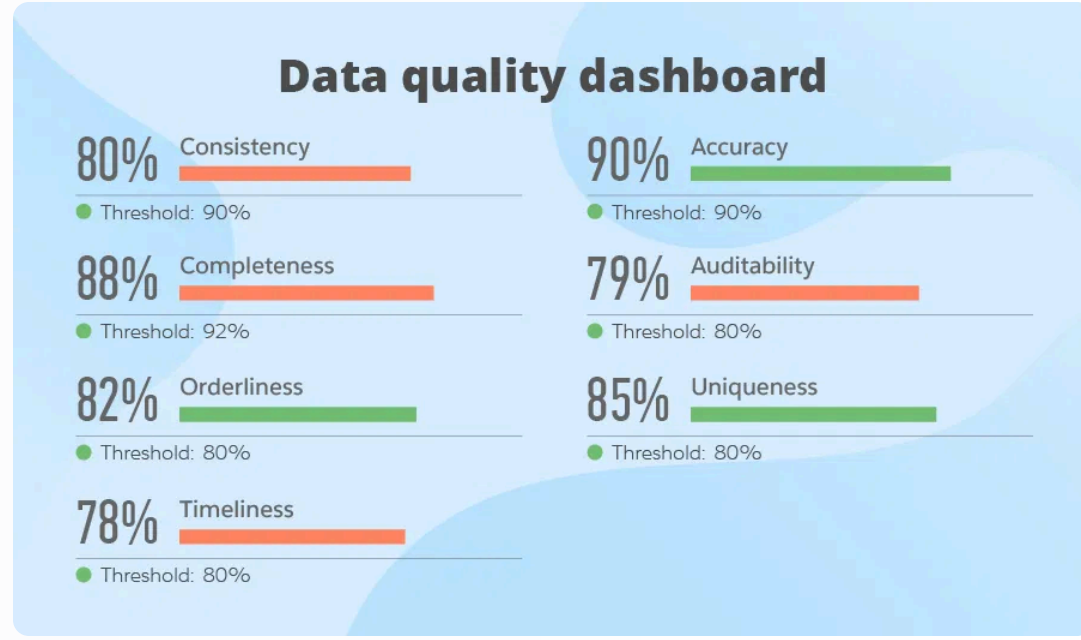
O Que Buscamos Garantir

Através da validação, buscamos assegurar três aspectos fundamentais:

Precisão: Os dados representam corretamente a realidade que pretendem medir

Completude: Não há lacunas significativas que comprometam a análise

Ausência de vieses: Os dados não favorecem sistematicamente certos resultados



Impacto de Dados Ruins nas Conclusões

- ⚠️ Decisões de negócio equivocadas
- ⚠️ Modelos preditivos imprecisos
- ⚠️ Perpetuação de desigualdades sociais
- ⚠️ Perda de credibilidade e confiança

A Tríade dos Problemas de Dados

Problemas de Completude

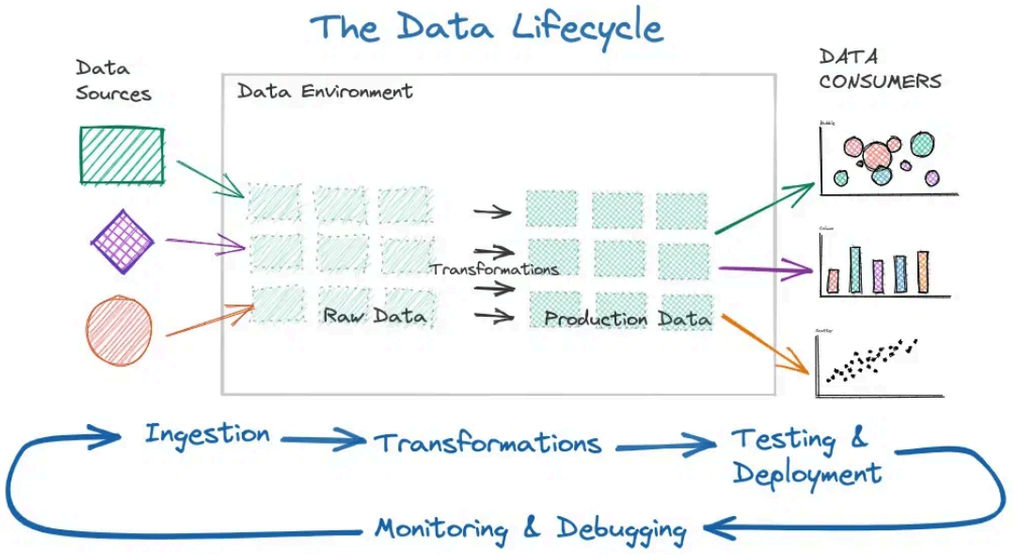
Ocorrem quando há **dados ausentes** (valores nulos ou NaN) no conjunto de dados. Podem ser causados por falhas na coleta, erros no sistema ou quando a informação não é aplicável. Afetam diretamente a representatividade da amostra e podem introduzir viés na análise.

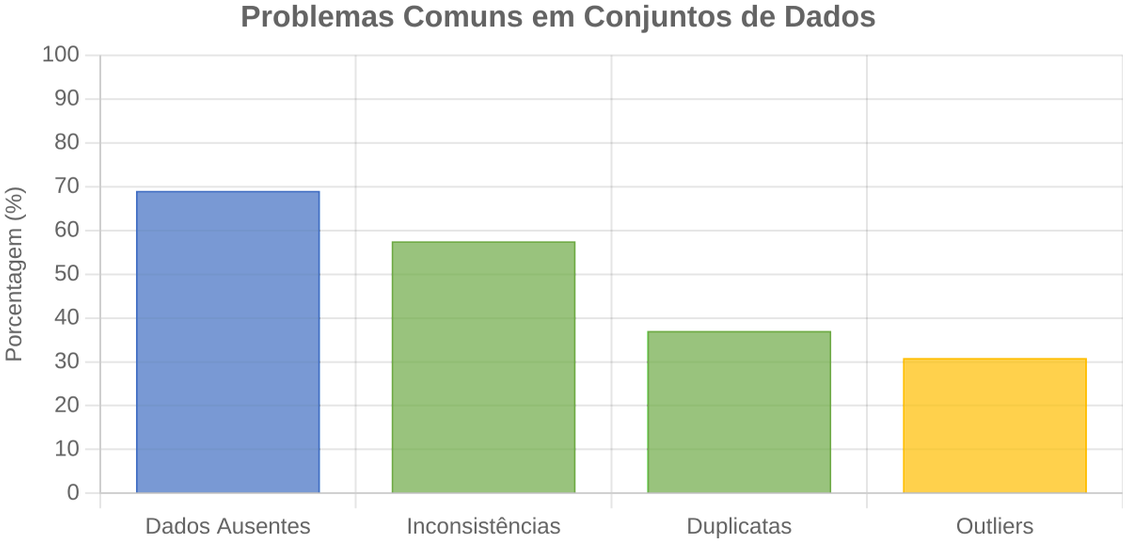
Problemas de Consistência e Precisão

Englobam **dados incorretos e duplicados**. Incluem erros de formatação, tipos de dados incompatíveis, valores fora do domínio esperado e registros repetidos. Comprometem a confiabilidade das análises e podem levar a conclusões equivocadas.

Problemas de Discrepância

Referem-se a **outliers** ou valores atípicos que se distanciam significativamente do padrão observado nos dados. Podem ser erros genuínos ou eventos raros legítimos. Requerem investigação cuidadosa antes de qualquer tratamento para evitar perda de informações valiosas.





Problemas de Completude - Dados Ausentes

O que são valores nulos (NaN)?

Valores nulos representam a ausência de informação em um conjunto de dados. Em Python/Pandas, são representados como `NaN` (Not a Number) ou `None`.

Causas comuns

- **Falha na coleta:** Sensores com defeito, formulários incompletos
- **Erro no sistema:** Falhas de integração entre sistemas
- **Resposta não aplicável:** Campos não relevantes para certos registros

Estratégias de Tratamento

Remoção: Eliminar linhas ou colunas com muitos valores ausentes

Imputação: Substituir por média, mediana ou valores constantes

Detecção de Valores Ausentes

```
# Verificando valores ausentes por coluna
df.isnull().sum()

# Calculando porcentagem de valores ausentes
percent_missing = df.isnull().sum() * 100 / len(df)
```

Tratamento de Valores Ausentes

```
# Remoção de linhas com valores ausentes
df_clean = df.dropna()

# Imputação com média
df['idade'] = df['idade'].fillna(df['idade'].mean())
```

	First Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
20	Lois	NaN	4/22/1995	7:18 PM	64714	4.934	True	Legal
22	Joshua	NaN	3/8/2012	1:58 AM	90816	18.816	True	Client Services
27	Scott	NaN	7/11/1991	6:58 PM	122367	5.218	False	Legal
31	Joyce	NaN	2/20/2005	2:40 PM	88657	12.752	False	Product
41	Christine	NaN	6/28/2015	1:08 AM	66582	11.308	True	Business Development
49	Chris	NaN	1/24/1980	12:13 PM	113590	3.055	False	Sales
51	NaN	NaN	12/17/2011	8:29 AM	41126	14.009	NaN	Sales
53	Alan	NaN	3/3/2014	1:28 PM	40341	17.578	True	Finance
60	Paula	NaN	11/23/2005	2:01 PM	48866	4.271	False	Distribution
64	Kathleen	NaN	4/11/1990	6:46 PM	77834	18.771	False	Business Development
69	Irene	NaN	7/14/2015	4:31 PM	100863	4.382	True	Finance
70	Todd	NaN	6/10/2003	2:26 PM	84692	6.617	False	Client Services
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
939	Ralph	NaN	7/28/1995	6:53 PM	70635	2.147	False	Client Services
945	Gerald	NaN	4/15/1989	12:44 PM	93712	17.426	True	Distribution
961	Antonio	NaN	6/18/1989	9:37 PM	103050	3.050	False	Legal
972	Victor	NaN	7/28/2006	2:49 PM	76381	11.159	True	Sales
985	Stephen	NaN	7/10/1983	8:10 PM	85668	1.909	False	Legal
989	Justin	NaN	2/10/1991	4:58 PM	38344	3.794	False	Legal
995	Henry	NaN	11/23/2014	6:09 AM	132483	16.655	False	Distribution

145 rows × 8 columns

Problemas de Consistência e Precisão

Duplicatas

Registros duplicados ocorrem quando a mesma informação aparece mais de uma vez no conjunto de dados. Podem ser duplicatas exatas (todas as colunas idênticas) ou parciais (chaves principais idênticas).

Inconsistências

Inconsistências são variações na forma como os dados são representados:

- Erros de formatação (datas em formatos diferentes)
- Tipos de dados incorretos (números como texto)
- Valores fora do domínio esperado

Problema	Exemplo	Correção
Variação textual	"SP" vs. "São Paulo"	Padronização
Tipo incorreto	Idade: "30 anos"	Conversão: 30

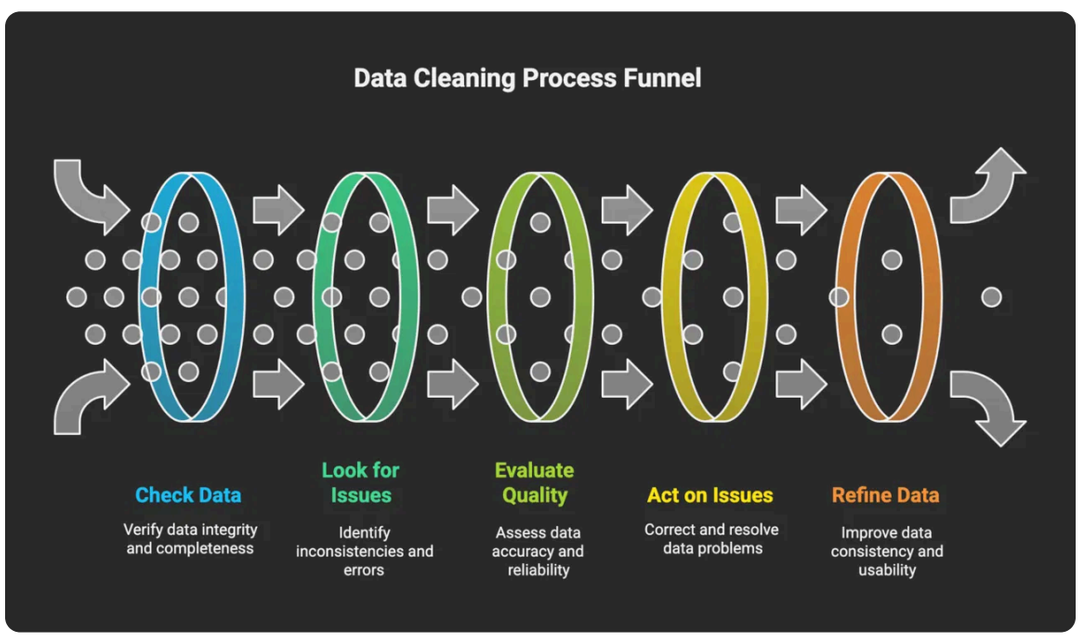
Identificação de Duplicatas

```
# Verificar quantidade de duplicatas
duplicatas = df.duplicated().sum()
print(f"Total de registros duplicados: {duplicatas}")
```

Correção de Inconsistências

```
# Remoção de duplicatas
df_clean = df.drop_duplicates(inplace=False)

# Padronização de texto
df['estado'] = df['estado'].replace({
  'SP': 'São Paulo',
  'RJ': 'Rio de Janeiro'
})
```



Problemas de Discrepância - Outliers

Definição de Outliers

Outliers são valores que se distanciam significativamente do padrão observado nos dados. São pontos que estão muito acima ou abaixo da maioria dos outros valores em uma distribuição.

Tipos de Outliers

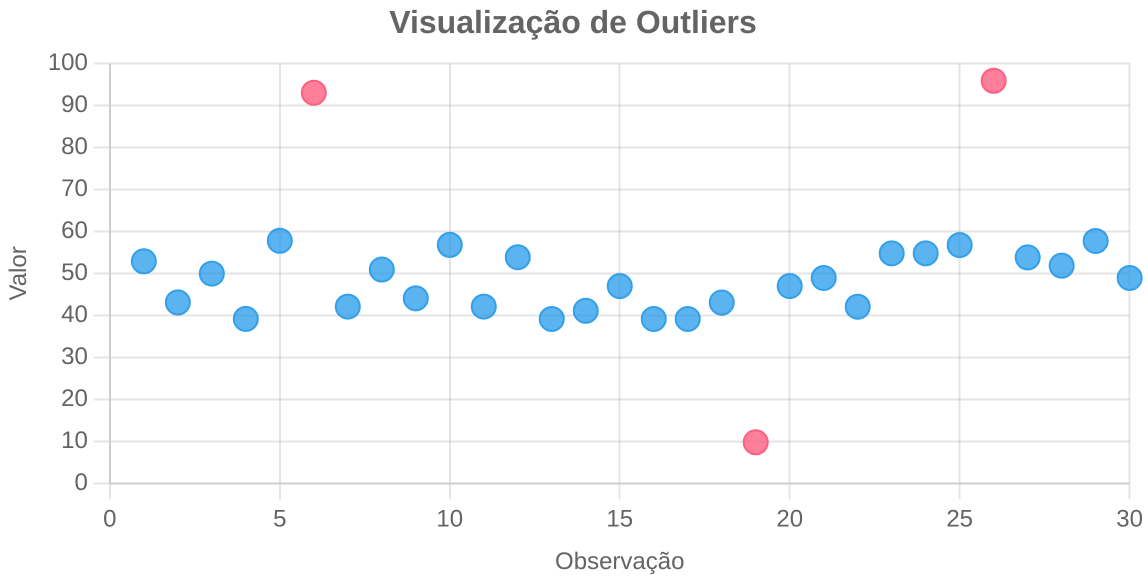
- **Erros de digitação:** Valores incorretos (ex: altura de 250m)
- **Eventos raros legítimos:** Valores extremos mas reais

Estratégias de Identificação

Regra do Desvio Padrão: Valores além de 3 desvios padrão

Método IQR (Boxplot): Valores abaixo de Q1-1.5*IQR ou acima de Q3+1.5*IQR

Importante: *Sempre investigue outliers antes de removê-los.*



Atividade Prática 1 - Detecção e Correção de Problemas

Objetivo

Identificar e corrigir problemas em um dataset com erros comuns, aplicando as técnicas aprendidas para melhorar a qualidade dos dados antes da análise.

Dataset: Vendas de E-commerce

- 10.000 registros de vendas online
- Colunas: ID, Data, Cliente, Produto, Categoria, Quantidade, Preço, Estado
- Problemas: valores nulos, duplicatas, inconsistências e outliers

Passo a Passo

1. **Identificação de valores nulos**

```
df.isnull().sum()
```

2. **Imputação de valores ausentes**

```
df['Quantidade'] = df['Quantidade'].fillna(df['Quantidade'].median())
df['Categoria'] = df['Categoria'].fillna('Não categorizado')
```

3. **Identificação e remoção de duplicatas**

```
duplicatas = df.duplicated().sum()
df.drop_duplicates(inplace=True)
```

4. **Tratamento de inconsistências textuais**

```
df['Estado'] = df['Estado'].replace({
  'SP': 'São Paulo',
  'S. Paulo': 'São Paulo'
})
```

5. **Identificação e análise de outliers**

```
Q1 = df['Preço'].quantile(0.25)
Q3 = df['Preço'].quantile(0.75)
IQR = Q3 - Q1
outliers = df[(df['Preço'] < Q1 - 1.5 * IQR) | (df['Preço'] > Q3 + 1.5 * IQR)]
```

Viés em Dados e Algoritmos

O que é viés?

Viés é uma tendência sistemática que favorece ou desfavorece certos resultados. Em dados e algoritmos, o viés pode levar a conclusões incorretas ou injustas, perpetuando ou amplificando desigualdades existentes.

Viés de Amostra/Seleção

Ocorre quando os dados coletados não representam adequadamente a população que se pretende analisar.

Exemplo: *Pesquisa online sobre hábitos de consumo que exclui pessoas sem acesso à internet.*

Viés Histórico

Acontece quando dados históricos contêm preconceitos ou desigualdades que são perpetuados em modelos preditivos.

Exemplo: *Algoritmo de contratação treinado com dados de uma empresa que historicamente contratou mais homens que mulheres para cargos de liderança.*

Viés de Confirmação

Tendência de interpretar ou coletar dados de forma a confirmar hipóteses ou crenças pré-existentes.

Exemplo: *Pesquisador que ignora resultados contraditórios à sua teoria ou formula perguntas tendenciosas em questionários.*

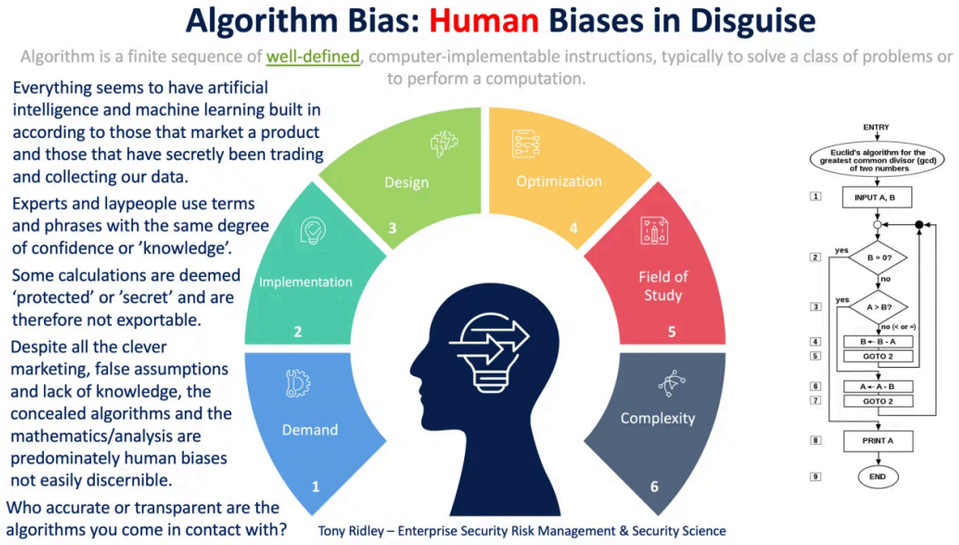


Ilustração de como o viés nos dados de treinamento pode levar a resultados enviesados em algoritmos de IA

Impacto do Viés

O viés em dados e algoritmos pode resultar em:

- Discriminação contra grupos minoritários
- Perpetuação de desigualdades sociais
- Decisões empresariais equivocadas
- Perda de confiança em sistemas automatizados

Atividade Prática 2 - Análise de Casos Reais de Viés

Estudo de Caso: Algoritmo de Saúde

Contexto

Em 2019, pesquisadores da Universidade da Califórnia descobriram que um algoritmo amplamente utilizado por hospitais americanos para identificar pacientes que necessitavam de cuidados adicionais estava sistematicamente subestimando as necessidades de pacientes negros em comparação com pacientes brancos com condições médicas similares.

O Problema

O algoritmo usava custos históricos de saúde como proxy para necessidades médicas. No entanto, devido a disparidades socioeconômicas e barreiras de acesso à saúde, pacientes negros historicamente gastavam menos em cuidados médicos, mesmo quando apresentavam condições mais graves.

Como resultado, o algoritmo atribuía pontuações de risco mais baixas a pacientes negros, reduzindo suas chances de serem selecionados para programas de cuidados especiais, perpetuando desigualdades existentes no sistema de saúde.

Solução Implementada

Após a descoberta, os pesquisadores trabalharam com a empresa desenvolvedora para reformular o algoritmo, substituindo os custos históricos por indicadores diretos de necessidade médica. Essa mudança reduziu o viés racial em 84%.

Tarefa em Grupo

Analise o caso apresentado e trabalhe em grupo para identificar os problemas e propor soluções.

1. **Identificar o tipo de viés envolvido** (Amostra, Histórico ou Confirmação)
2. **Analisar o impacto** desse viés nas decisões tomadas pelo algoritmo
3. **Propor estratégias de mitigação** que poderiam ter evitado o problema
4. **Apresentar conclusões** para a turma (5 minutos por grupo)

Reflexão sobre Responsabilidade Ética

Discuta com seu grupo as seguintes questões:

- Quem deve ser responsabilizado quando algoritmos produzem resultados enviesados?
- Quais medidas preventivas deveriam ser obrigatórias antes da implementação de algoritmos em áreas sensíveis?
- Como equilibrar inovação tecnológica com proteção contra discriminação algorítmica?

Prevenção de Viés e Amostragem Balanceada

Técnicas para Prevenir Viés

Amostragem Estratificada

Divide a população em subgrupos (estratos) com base em características relevantes e seleciona amostras proporcionais de cada estrato, garantindo representatividade.

Balanceamento de Classes

Ajusta a proporção de diferentes grupos nos dados usando técnicas como oversampling (aumentar grupos minoritários) e undersampling (reduzir grupos majoritários).

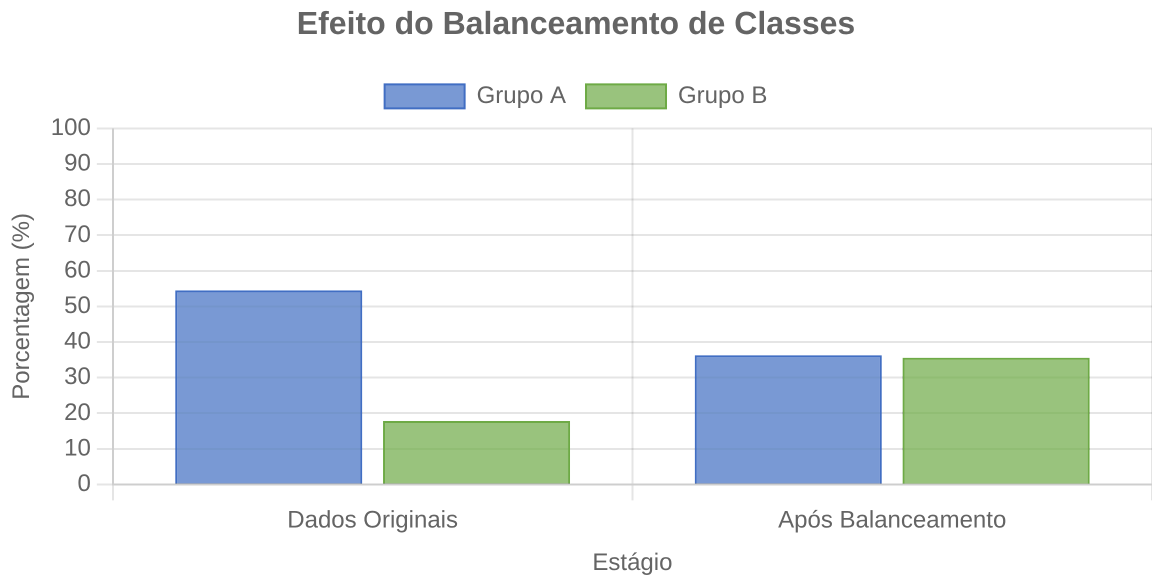
Revisão de Critérios de Filtragem

Examina criticamente os filtros aplicados durante a coleta ou pré-processamento, evitando exclusões que possam introduzir viés sistemático.

Exemplo: Amostragem Estratificada

```
from sklearn.model_selection import train_test_split

# Amostragem estratificada por gênero
X_train, X_test, y_train, y_test = train_test_split(
  df.drop('target', axis=1),
  df['target'],
  stratify=df['genero'],
  random_state=42
)
```



Atividade Prática 3 - Documentação e Apresentação

Importância da Documentação

A documentação transparente do processo de validação e limpeza de dados é essencial para garantir a reprodutibilidade e confiabilidade das análises.

- Permite reproduzir o processo em novos conjuntos de dados
- Facilita a auditoria e revisão por pares
- Cria um histórico das transformações aplicadas

Estrutura do Relatório Explicativo

1. **Problemas Encontrados (Quantificação)**

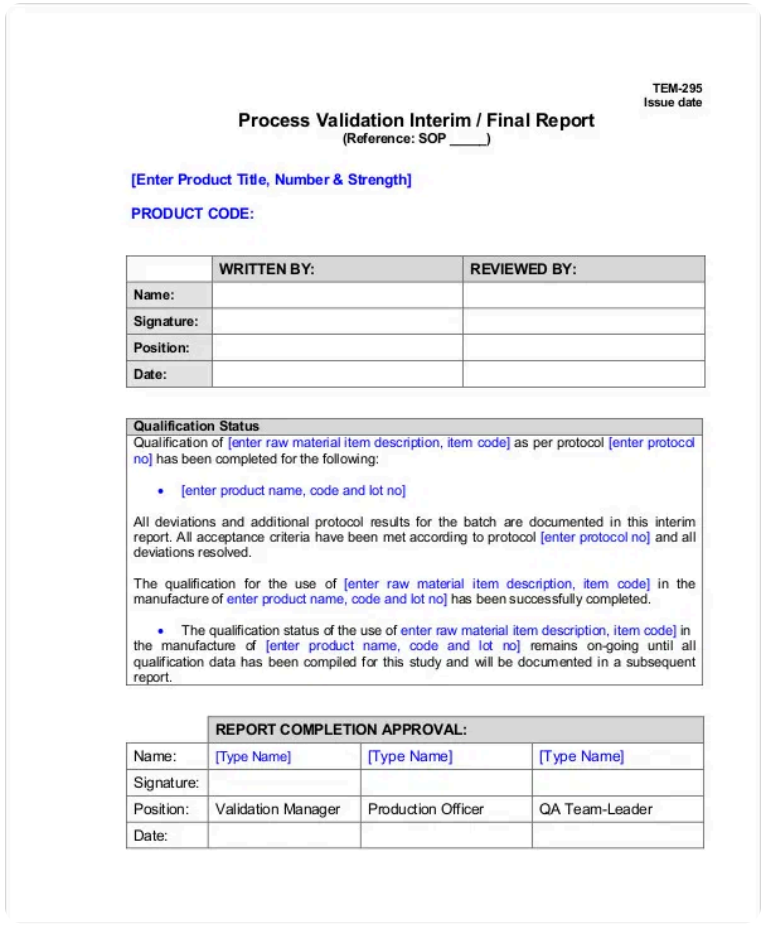
   Documentar todos os problemas identificados no conjunto de dados, incluindo estatísticas sobre valores ausentes, duplicatas e inconsistências.

2. **Ações Corretivas (Funções e Critérios)**

   Detalhar todas as transformações aplicadas, incluindo o código utilizado e os critérios de decisão para cada abordagem escolhida.

3. **Viés e Prevenção (Identificação e Correção)**

   Analisar potenciais vieses nos dados e medidas tomadas para mitigá-los.



TEM-295
Issue date

Process Validation Interim / Final Report
(Reference: SOP ____)

[Enter Product Title, Number & Strength]

PRODUCT CODE:

	WRITTEN BY:	REVIEWED BY:
Name:		
Signature:		
Position:		
Date:		

Qualification Status

Qualification of [enter raw material item description, item code] as per protocol [enter protocol no] has been completed for the following:

- [enter product name, code and lot no]

All deviations and additional protocol results for the batch are documented in this interim report. All acceptance criteria have been met according to protocol [enter protocol no] and all deviations resolved.

The qualification for the use of [enter raw material item description, item code] in the manufacture of enter product name, code and lot no] has been successfully completed.

- The qualification status of the use of enter raw material item description, item code] in the manufacture of [enter product name, code and lot no] remains on-going until all qualification data has been compiled for this study and will be documented in a subsequent report.

	REPORT COMPLETION APPROVAL:		
Name:	[Type Name]	[Type Name]	[Type Name]
Signature:			
Position:	Validation Manager	Production Officer	QA Team-Leader
Date:			

```
# Relatório de Validação e Limpeza de Dados

## 1. Problemas Encontrados

### 1.1 Valores Ausentes
- Coluna 'idade': 120 valores ausentes (1.2%)
- Coluna 'renda': 450 valores ausentes (4.5%)

### 1.2 Duplicatas
- 78 registros duplicados identificados (0.78%)

## 2. Ações Corretivas

### 2.1 Tratamento de Valores Ausentes
```python
# Imputação com mediana para idade
df['idade'] = df['idade'].fillna(df['idade'].median())
```
```

Encerramento e Próximos Passos

Conceitos-Chave

Qualidade de dados é fundamental: Dados ruins levam a análises ruins e decisões equivocadas.

Tríade de problemas: Completude (dados ausentes), Consistência (incorretos/duplicados) e Discrepância (outliers).

Viés em dados: Pode perpetuar desigualdades e levar a conclusões injustas se não for identificado e tratado.

Documentação transparente: Essencial para reprodutibilidade e confiabilidade das análises.

Conexão com Próximas Aulas

Visualização de dados: Aprenderemos a criar visualizações eficazes para comunicar insights.

Modelagem preditiva: Utilizaremos dados validados para construir modelos de machine learning.

Projeto integrador: Aplicaremos todos os conceitos em um projeto de análise de dados completo.

Recursos Adicionais

Livros e Artigos

- "Data Quality: The Accuracy Dimension" - Jack E. Olson
- "Cleaning Data for Effective Data Science" - David Mertz
- "Fairness and Machine Learning" - Solon Barocas, Moritz Hardt, Arvind Narayanan

Ferramentas e Bibliotecas

- PyJanitor: Ferramentas de limpeza de dados para Python
- Great Expectations: Validação e documentação de dados
- Fairlearn: Mitigação de viés em modelos de ML

Dúvidas? Entre em contato: **professor@email.com**