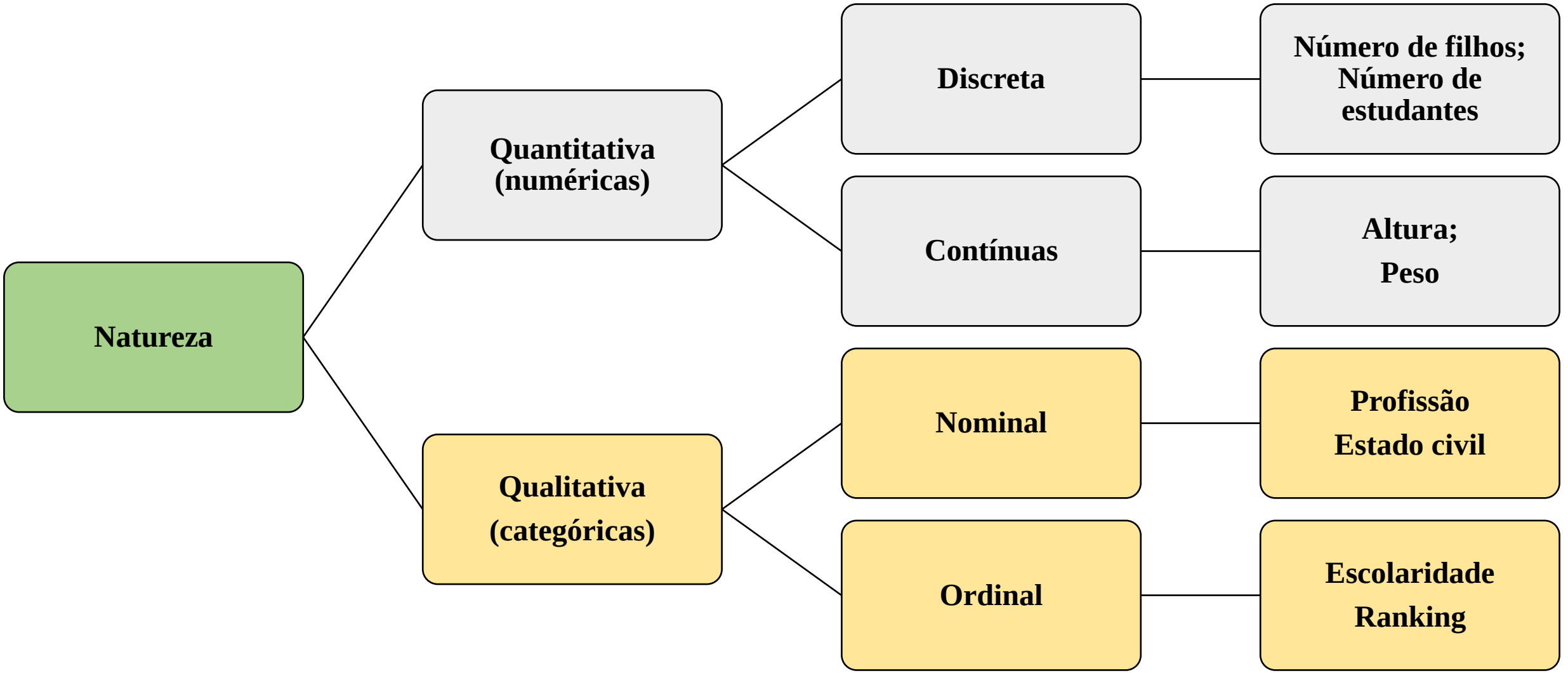
Data Science

UniSenai PR– São José dos Pinhais

Revisão

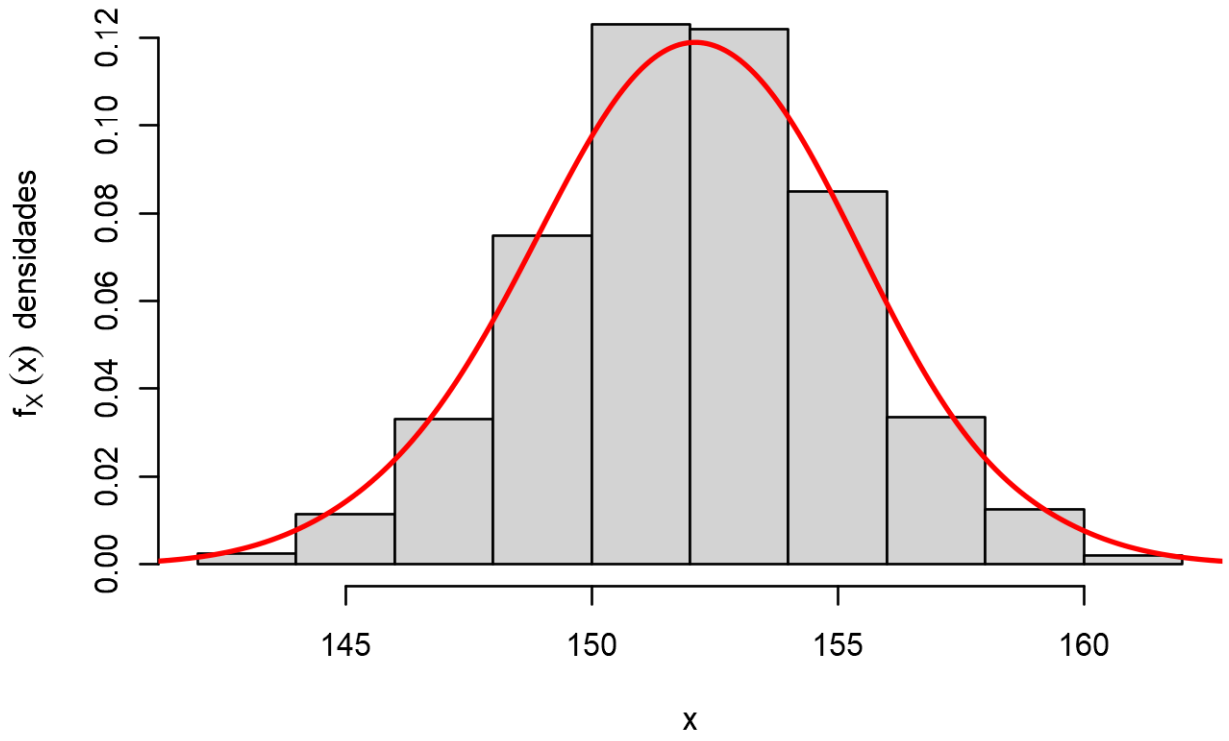

Natureza
Quantitativa (numéricas)
Qualitativa (categóricas)
Discreta
Contínuas
Nominal
Ordinal
Número de filhos; Número de estudantes
Altura; Peso
Profissão Estado civil
Escolaridade Ranking

Medida	Fórmula para Amostras	Fórmula para População
Média	$\bar{x} = \frac{(\sum_{i=1}^{n} x_i)}{n}$	$\mu = \frac{(\sum_{i=1}^{n} x_i)}{\mathrm{N}}$
Mediana	Ordenar os dados e encontrar o valor do meio	Ordenar os dados e encontrar o valor do meio
Moda	Identificar o valor mais frequente	Identificar o valor mais frequente

Medida	Resumo
Média	É sensível a valores extremos, o que pode afetar significativamente o resultado final.
Mediana	Menos afetada por valores extremos.
Moda	Útil para dados qualitativos e pode não ser adequada para representar a tendência central em distribuições contínuas

Medida	Resumo	Utilidade
Amplitude	Medida simples da extensão total dos valores no conjunto de dados.	Não fornece informações sobre a dispersão dos valores em relação à média.
Variância	Medida da dispersão dos valores em relação à média.	Indica a variabilidade dos dados; Maior valor indica maior dispersão; Menor valor indica menor dispersão.
Desvio padrão	Raiz quadrada da variância.	Mede a dispersão dos dados em unidades originais; Facilita a interpretação comparada com os dados originais.
Coeficiente de variação	Medida relativa de dispersão.	Permite comparar a variabilidade entre diferentes conjuntos de dados, independentemente de suas escalas ou médias.

Medida	Definição	Cálculo
Quartis	Pontos que dividem os dados em quartos da distribuição.	Q1 = 25º percentil
		Q2 = 50º percentil (Mediana)
		Q3 = 75º percentil
		Q4 = 100º percentil (Valor máximo)
Intervalo Interquartil	A diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1).	Intervalo Interquartil = Q3 - Q1



Exemplo de distribuição normal

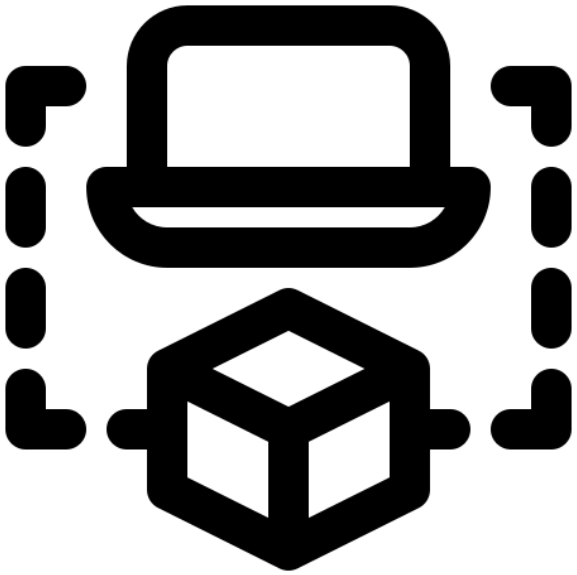
Conceito	Definição
Hipótese nula (H_0)	A afirmação inicial que é testada
Hipótese alternativa (H_1)	A afirmação contrária à nula
Estatística de teste	Medida calculada a partir dos dados amostrais para avaliar a hipótese nula
Nível de significância	Probabilidade máxima que estamos dispostos a aceitar de cometer um erro de Tipo I (rejeitar a hipótese nula quando ela é verdadeira) em um teste de hipótese.
p-valor	A probabilidade de observar a estatística de teste ou algo mais extremo, assumindo que a hipótese nula seja verdadeira

Pré-processamento

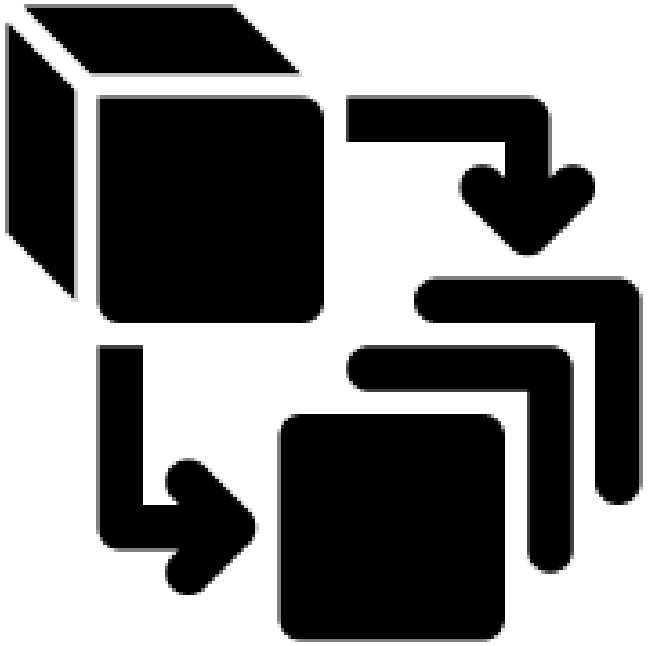
Tópico	Descrição
Pré-processamento de dados	Conjunto de técnicas para preparar, limpar e organizar dados brutos.
Etapas do Pré-processamento	Coleta, Limpeza, Transformação, Redução e Integração de Dados.
Importância	Dados inadequados podem levar a análises errôneas e conclusões falhas.

- Obter os dados de fontes relevantes (banco de dados, APIs, arquivos (csv, txt), etc.);
- Garantir que os dados estejam em um formato adequado e legível.

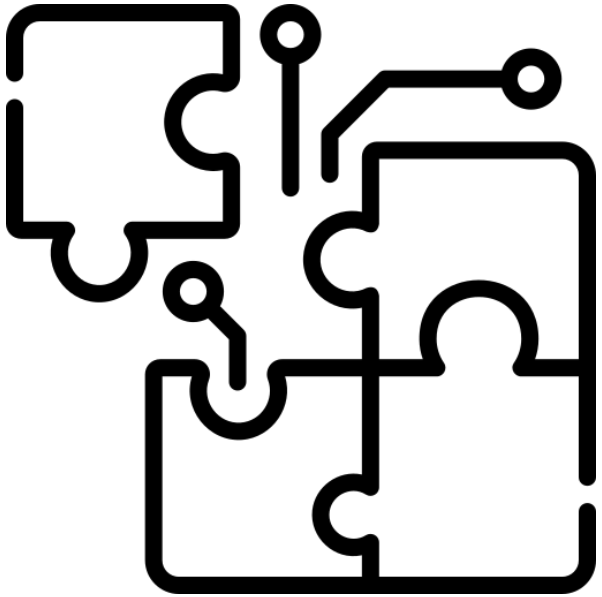
- Identificar e tratar valores faltantes ou nulos;
- Detectar e lidar com *outliers;*
- Padronizar ou corrigir erros de formatação nos dados;
- Remover ou corrigir registros duplicados ou inconsistentes.



- Normalizar ou escalonar os dados para evitar viés em algoritmos sensíveis à escala;
- Codificar variáveis categóricas em formatos numéricos;
- Transformações baseadas em logaritmo ou raiz quadrada para ajustar a distribuição dos dados;
- Criar novas variáveis por meio de agregações, operações matemáticas ou outras técnicas.

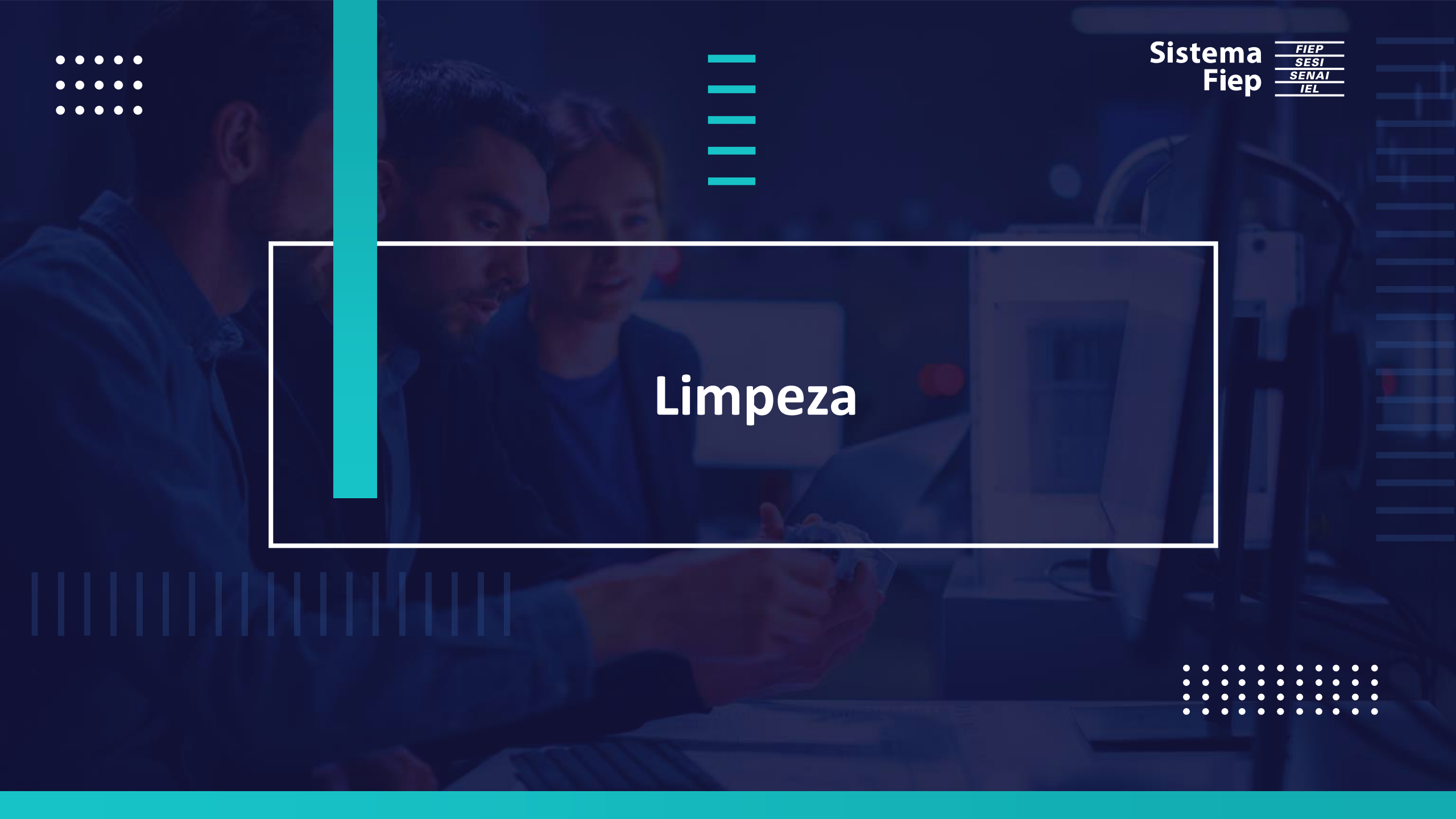


- Selecionar as variáveis mais relevantes e significativas para a análise ou modelo;
- Aplicar técnicas de seleção de características para eliminar variáveis redundantes ou pouco informativas;
- Realizar análise de componentes principais (PCA) para reduzir a dimensionalidade dos dados;



- Combinar dados de diferentes fontes ou tabelas para enriquecer a análise;
- Resolver conflitos de nomenclatura ou formatação entre diferentes conjuntos de dados;
- Assegurar que os dados estejam integrados de maneira coesa e coerente.

Etapa	Descrição
Coleta de Dados	Obter dados de fontes diversas e garantir a legibilidade.
Limpeza de Dados	Identificar e tratar valores faltantes, *outliers*, erros de formatação e registros duplicados.
Transformação de Dados	Normalizar, codificar variáveis, aplicar transformações e criar novas variáveis.
Redução de Dados	Selecionar variáveis relevantes e aplicar técnicas de redução de dimensionalidade.
Integração de Dados	Combinar dados de diferentes fontes e garantir coesão na integração.



Limpeza

Tópico	Descrição
Valores Ausentes	Dados incompletos podem afetar resultados; Estratégias de tratamento.
Métodos de Preenchimento	Preenchimento com média, mediana, valor constante; Algoritmos (*k-Nearest Neighbors* (k-NN), regressão, etc.).
Duplicatas	Dados duplicados podem distorcer análises; Identificação e remoção.
Identificação e Tratamento de outliers	Valores extremos que podem afetar análises; Técnicas de detecção; Remover, transformar ou substituir por valores menos extremos.

Método	Quando Usar	Vantagens	Limitações
Média	Dados contínuos com distribuição aproximadamente normal; ou Quando a média é uma representação adequada	Simplicidade; Preserva a média da distribuição	Sensível a outliers; pode distorcer a distribuição
Mediana	Dados contínuos com distribuição assimétrica, sensíveis a outliers; ou Quando a mediana é uma representação mais robusta	Robusto a *outliers*; Preserva a forma geral da distribuição	Não leva em consideração a relação entre variáveis
Moda	Dados categóricos ou discretos, onde a moda é uma medida relevante	Simplicidade; Útil para variáveis categóricas	Não captura informações sobre a distribuição

Método	Quando Usar	Vantagens	Limitações
k-NN	Dados com valores ausentes onde a relação entre variáveis é importante	Leva em conta a relação entre variáveis	Complexidade computacional; Requer dados completos
Regressão	Dados com valores ausentes e relações complexas entre variáveis	Pode considerar interações entre variáveis	Dependente de um bom modelo de regressão

Técnica	Quando Usar	Vantagens	Limitações
Remover *Outliers*	Valores extremos que são erros de medição ou não representam o comportamento normal dos dados	Pode melhorar a precisão dos modelos; Reduz impacto de *outliers*	Pode levar à perda de informações valiosas; Sensível à escolha do limiar
Substituir *Outliers*	Outliers são preservados, mas seus efeitos são mitigados	Mantém a informação dos *outliers*; Reduz impacto nos modelos	Pode introduzir distorções nos dados; Requer escolha de método adequado

Técnica	Descrição	Vantagens	Desvantagens
Z-score	Mede o desvio de um ponto de dados em relação à média, em termos de desvios padrão.	Sensível a outliers extremos.	Menos adequado para distribuições não normais.
IQR (Intervalo Interquartil)	Calcula a diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1) dos dados.	Mais robusto contra outliers extremos.	Não distingue entre diferentes tipos de outliers.

Z-score (Escores Padrão):

1. Mede quantos desvios padrão um ponto de dados está longe da média;
2. É sensível a outliers extremos, pois leva em consideração a média e o desvio padrão dos dados;
3. Útil quando se deseja identificar pontos de dados que estão muito distantes da média em distribuições normais.
4. Valores com Z-score acima de um limite (geralmente 2 ou 3) são considerados outliers;

IQR (Intervalo Interquartil):

1. Concentra-se nos quartis dos dados, tornando-o mais robusto contra outliers;
2. É menos sensível a outliers extremos do que o Z-score, pois se baseia nos percentis 25 (Q1) e 75 (Q3);
3. Valores abaixo de Q1 - 1,5 * IQR ou acima de Q3 + 1,5 * IQR são frequentemente considerados outliers;
4. É apropriado quando se deseja lidar com valores discrepantes sem remover muitos dados legítimos.

Técnica	Descrição	Vantagens	Desvantagens
Substituição por Limite	Substitui valores acima ou abaixo de limites pré-definidos pelo próprio limite.	Simples de implementar.	Pode distorcer distribuição original.
Substituição por Média	Substitui outliers pela média da variável.	Reduz impacto de valores extremos.	Afeta a média e pode distorcer análises.
Substituição por Mediana	Substitui outliers pela mediana da variável.	Menos sensível a outliers extremos.	Afeta a mediana original.
Substituição por Modelo	Utiliza um modelo estatístico para estimar e substituir outliers com base em padrões dos dados.	Considera relações entre variáveis.	Requer construção e validação de modelo.

Técnica	Descrição	Vantagens	Desvantagens
Transformação de Dados	Transformações matemáticas (logaritmo, raiz, inverso) para reduzir impacto de outliers.	Pode melhorar a normalidade dos dados.	Pode afetar interpretação dos dados.
Truncagem	Definir limites superiores e inferiores e substituir outliers por esses limites.	Controla efeitos de valores extremos.	Reduz variabilidade, mas pode remover informações úteis.


Sistema Fiep
FIEP
SESI
SENAI
IEL
Transformação

- Processo de aplicar funções ou operações matemáticas a conjuntos de dados originais, a fim de modificar suas características e torná-los mais adequados para análises estatísticas ou modelagem;

- Utilizada para melhorar a qualidade dos dados, eliminar ou reduzir distorções, ajustar a distribuição e preparar os dados para a aplicação de algoritmos de aprendizado de máquina ou outras técnicas analíticas

Normalização e Escalonamento de Dados

Técnica	Fórmula	Descrição
Min-Max Scaling	$y = (x - \min) / (\max - \min)$	Transforma os dados para um intervalo específico (geralmente 0 a 1).
Z-score (Padronização)	$y = (x - \mu) / \sigma$	Transforma os dados para ter média zero e desvio padrão igual a um.
Robust Scaling	$y = (x - Q1) / (Q3 - Q1)$	Escalonamento usando primeiro e terceiro quartis (Q1 e Q3), reduzindo impacto de outliers.

Transformação de dados

Técnica	Fórmula	Descrição
Logaritmo	$y = \log(x)$	Reduz valores altos e amplia valores baixos; Útil para dados com distribuição exponencial.
Raiz Quadrada	$y = \sqrt{x}$	Atenua valores altos e amplifica valores baixos; Útil para suavizar assimetria.
Inverso	$y = 1/x$	Transforma valores pequenos em grandes e vice-versa; Pode melhorar interpretabilidade.

Codificação de Variáveis Categóricas

Codificação	Descrição
One-Hot (Dummy)	Transforma cada categoria em uma nova coluna binária; Útil para variáveis categóricas sem ordem.
Ordinal	Atribui números inteiros às categorias com base em uma ordem específica.
Rótulos	Atribui um valor numérico a cada categoria, sem criar novas colunas.

Criação de Novas Variáveis

Técnica	Descrição
Agregações	Criação de novas variáveis a partir de agregações, como soma, média, máximo, mínimo, etc.
Operações Matemáticas	Criação de novas variáveis por meio de operações como adição, subtração, multiplicação, divisão.
Transformações Estatísticas	Criação de variáveis com base em estatísticas, como desvio padrão, coeficiente de variação, etc.
Criação de Indicadores	Criação de variáveis binárias ou categóricas para indicar presença ou ausência de determinado fator.

Redução

Técnica	Descrição
Seleção de Variáveis Relevantes	Identificar e manter as variáveis mais relevantes para a análise ou modelo.
Técnicas de Seleção de Características	Aplicar métodos estatísticos ou algoritmos para eliminar variáveis pouco informativas.

Técnica	Descrição
Correlação	Avaliar a relação linear entre variáveis usando coeficientes de correlação
Testes de Hipóteses	Realizar testes estatísticos para determinar se uma variável tem impacto significativo nos resultados.
Análise de Importância de Variáveis	Utilizar modelos para medir a importância relativa das variáveis na predição.
Métodos de Aprendizado de Máquina	Algoritmos como Árvores de Decisão e *Random Forest* podem avaliar a relevância das variáveis.
Seleção Automatizada de Características	Utilizar algoritmos que selecionam automaticamente as melhores variáveis para o modelo.

Filtro Estatístico

- Avalia a relação variável x resultado para selecionar com base em métricas estatísticas

Wrapper

- Avalia exaustivamente subconjuntos de variáveis usando algoritmos de aprendizado de máquina

Incorporação (*Embedding*)

- Incorpora a seleção de características diretamente no processo de treinamento do modelo (árvores de decisão, por exemplo)

Análise de Componentes Principais (PCA)

- Reduz dimensionalidade, trabalhando com componentes principais

Seleção Recursiva de Características

- Remove iterativamente variáveis menos importantes de acordo com o desempenho do modelo


Sistema Fiep
FIEP
SESI
SENAI
IEL
Integração

Tópico	Descrição
Fusão de Dados	Combinar registros similares de diferentes fontes, geralmente usando chaves de junção.
Junção de Dados	Combinar colunas de diferentes fontes, geralmente usando um campo comum.
Agregação de Dados	Resumir dados de várias fontes, geralmente aplicando funções como média, soma ou contagem.
Integridade Referencial	Garantir que as relações entre os dados de diferentes fontes sejam mantidas.
Consolidação de Dados	Agregar os dados de diferentes fontes em um único repositório centralizado.
Ferramentas e Tecnologias	ETL (*Extract*, *Transform*, *Load*), ferramentas de integração de dados, bancos de dados.

Exemplos