Data Science

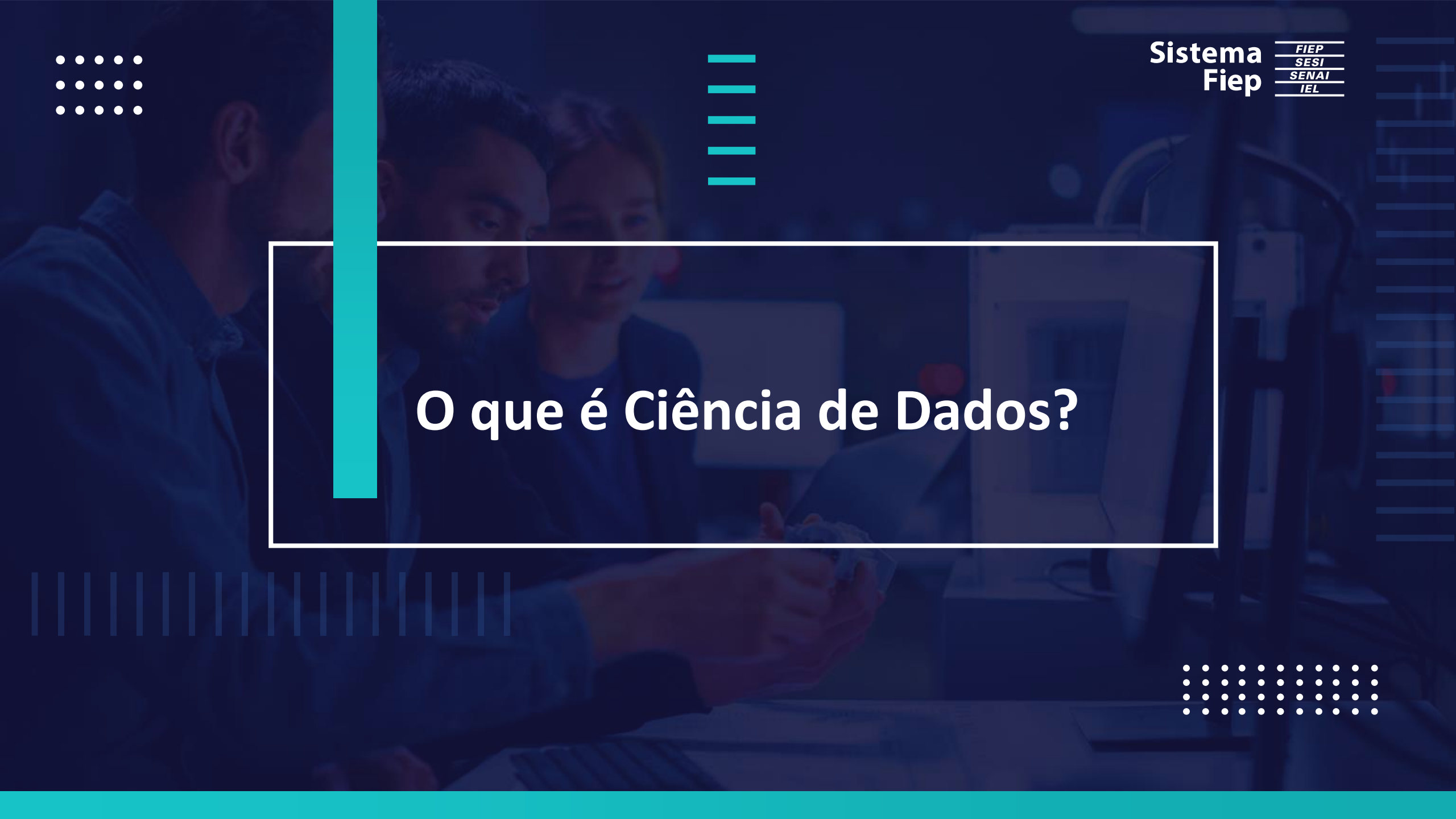
UniSenai PR– São José dos Pinhais

AGENDA

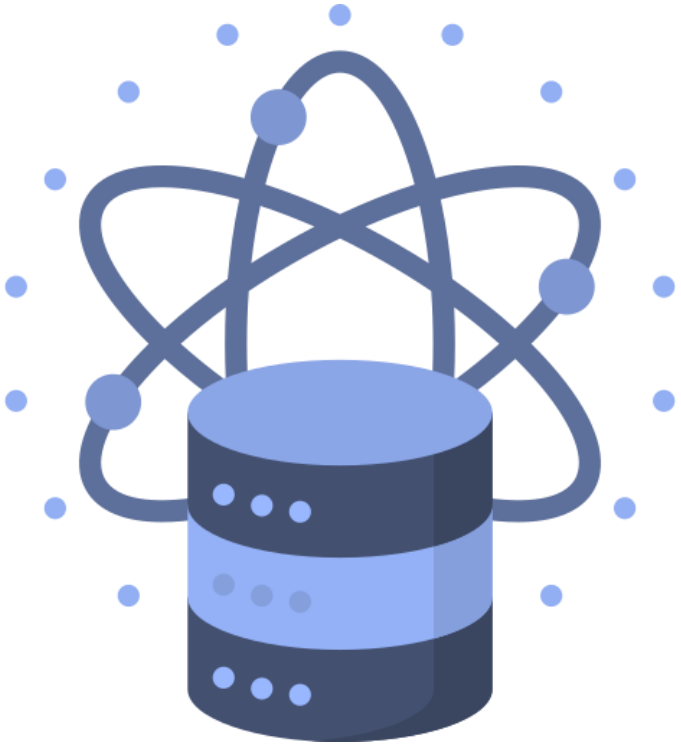
- Apresentação;
- Informações sobre a disciplina:
  - Plano de ensino;
  - Aspectos metodológicos (aulas, avaliações, ferramentas etc.);
- Introdução a conceitos de *data science*.

Sistema Fiep
FIEP SESI SENAI IEL

PLANO DE ENSINO



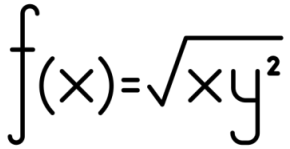
O que é Ciência de Dados?



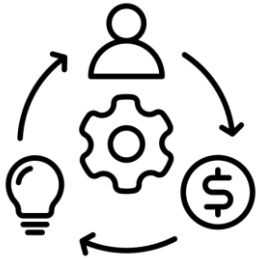
- Envolve **coleta, limpeza, análise e interpretação de dados para a tomada de decisões informadas**;

- O objetivo principal é **descobrir padrões, tendências e informações relevantes que possam ser usadas para resolver problemas** e **fornecer *insights* valiosos para as organizações**.

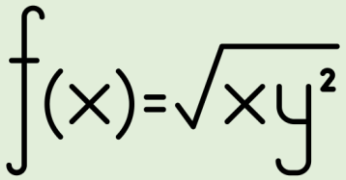
Combina conhecimentos de **Estatística, Matemática, Ciência da Computação e conhecimento de negócio para resolver problemas complexos;**



Requer uma **abordagem colaborativa entre especialistas de diferentes áreas para extrair o máximo de valor dos dados**.

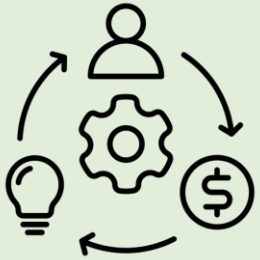


- Fornece as bases teóricas e métodos para análise e interpretação de dados;
- Ajuda a entender incertezas e variabilidades nos dados;
- Contribui para inferências, testes de hipóteses e estimação de parâmetros.



- Fundamenta as técnicas e algoritmos utilizados na Ciência de Dados;
- Aplica conceitos matemáticos em modelagem e resolução de problemas complexos;
- Desenvolve algoritmos de aprendizado de máquina e otimização.

- Habilita a manipulação, processamento e armazenamento eficiente de grandes volumes de dados;
- Desenvolve e implementa algoritmos de análise e modelagem;
- Utiliza linguagens de programação e ferramentas para análise de dados.

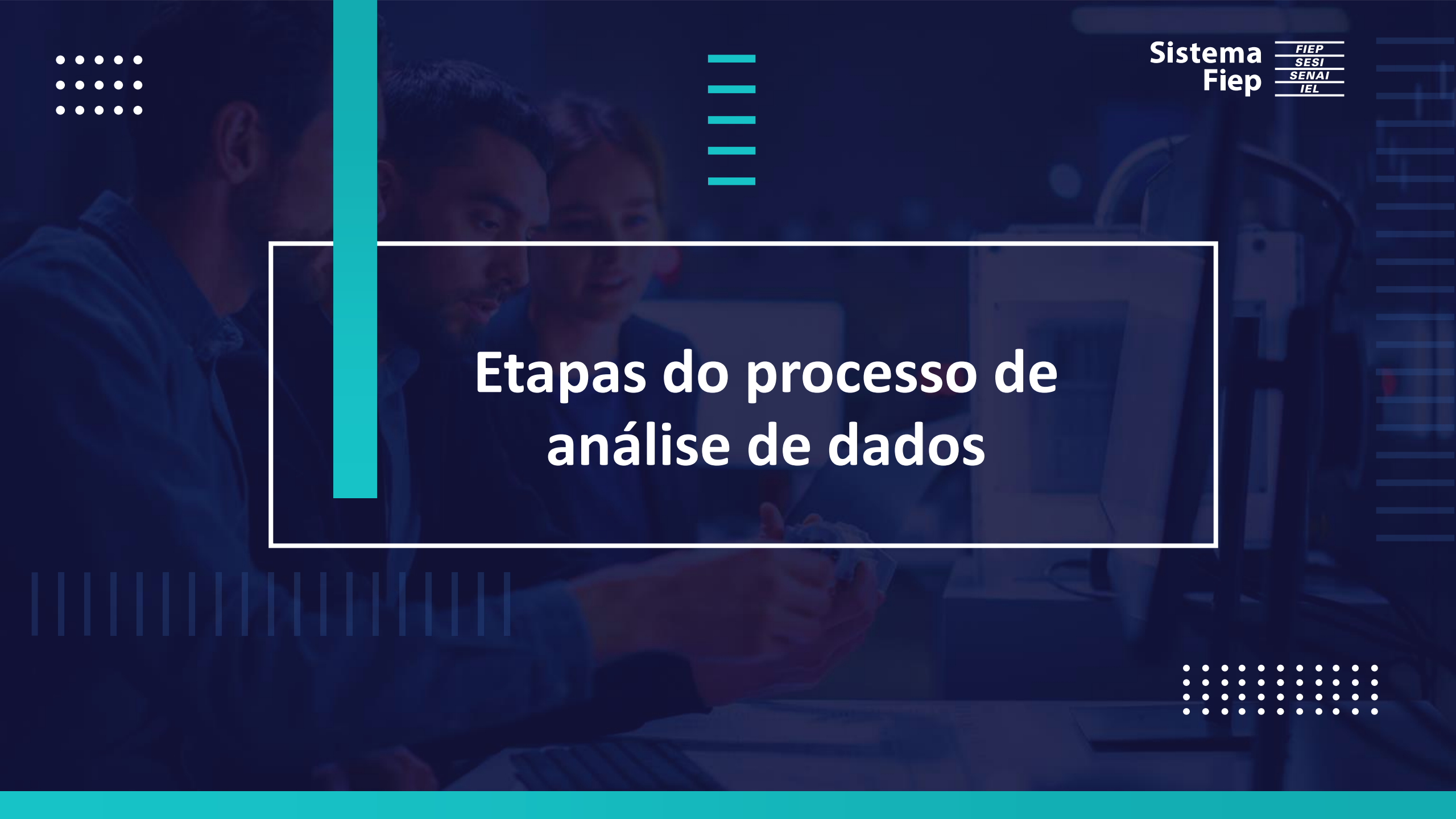


- Conhecimento sobre o contexto e o setor em que os dados estão inseridos;
- Formulação de questões relevantes que podem ser respondidas por meio da análise de dados;
- Colaboração com especialistas do domínio para obter insights significativos.

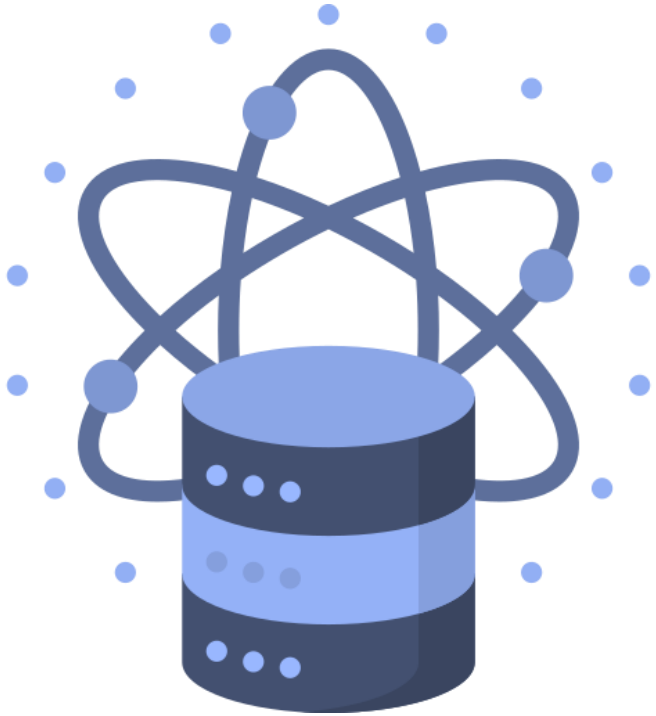
Cientista de Dados

Atua como um **profissional** versátil, capaz de lidar com a **coleta e preparação de dados, aplicar técnicas estatísticas e de aprendizado de máquina, desenvolver modelos preditivos e interpretar os resultados**;

Deve possuir **habilidades analíticas, capacidade de trabalhar em equipe e conhecimento técnico sólido**.



Etapas do processo de análise de dados

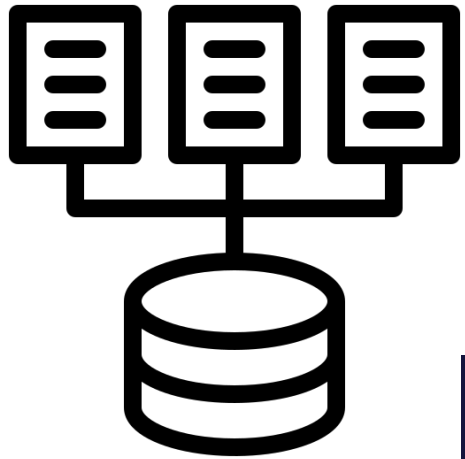


- Coleta
- Limpeza
- Exploração
- Modelagem
- Interpretação

I. **Definir** os **objetivos da análise** e as **perguntas a serem respondidas**;

II. **Identificar** as **fontes de dados relevantes**, como bancos de dados, APIs, arquivos, etc;

III. **Extrair** e **armazenar os dados** em um formato adequado para análise.



I. **Identificar** e **tratar** de **dados ausentes**, ***outliers*** e **valores inválidos**;

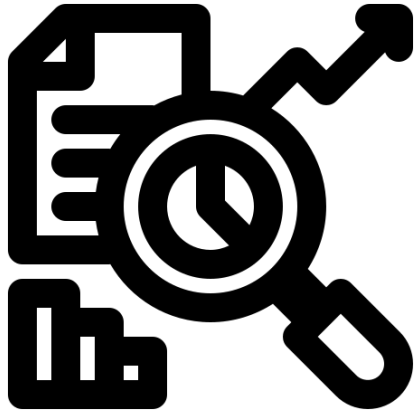
II. **Padronizar** os formatos e tratar os dados duplicados;

III. **Garantir** a consistência e integridade dos dados.

I. Análise descritiva para **compreender a distribuição e características dos dados;**

II. Usar gráficos e visualizações para identificar padrões e tendências;

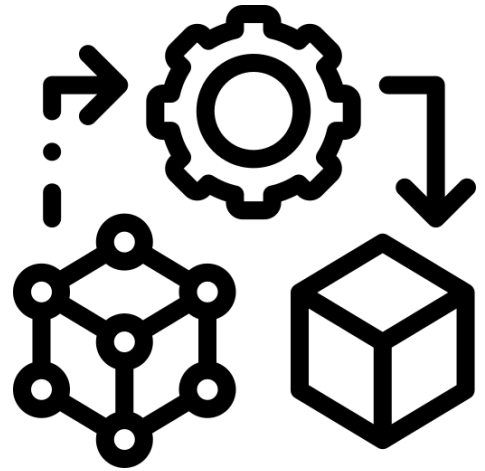
III. Explorar as relações entre variáveis para obter *insights* preliminares.



I. **Escolher técnicas** e **algoritmos** de modelagem adequados aos objetivos do projeto;

II. **Treinar modelos** com os dados de treinamento;

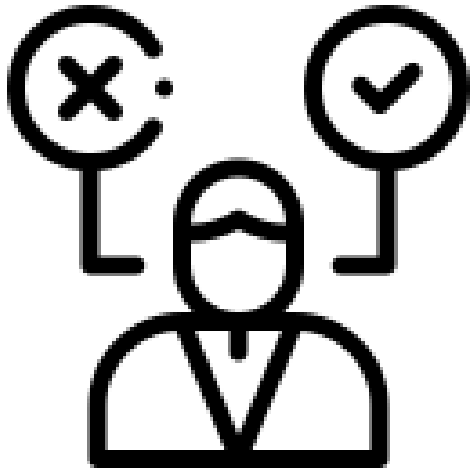
III. **Avaliar** o desempenho dos modelos e **ajustar** os parâmetros.

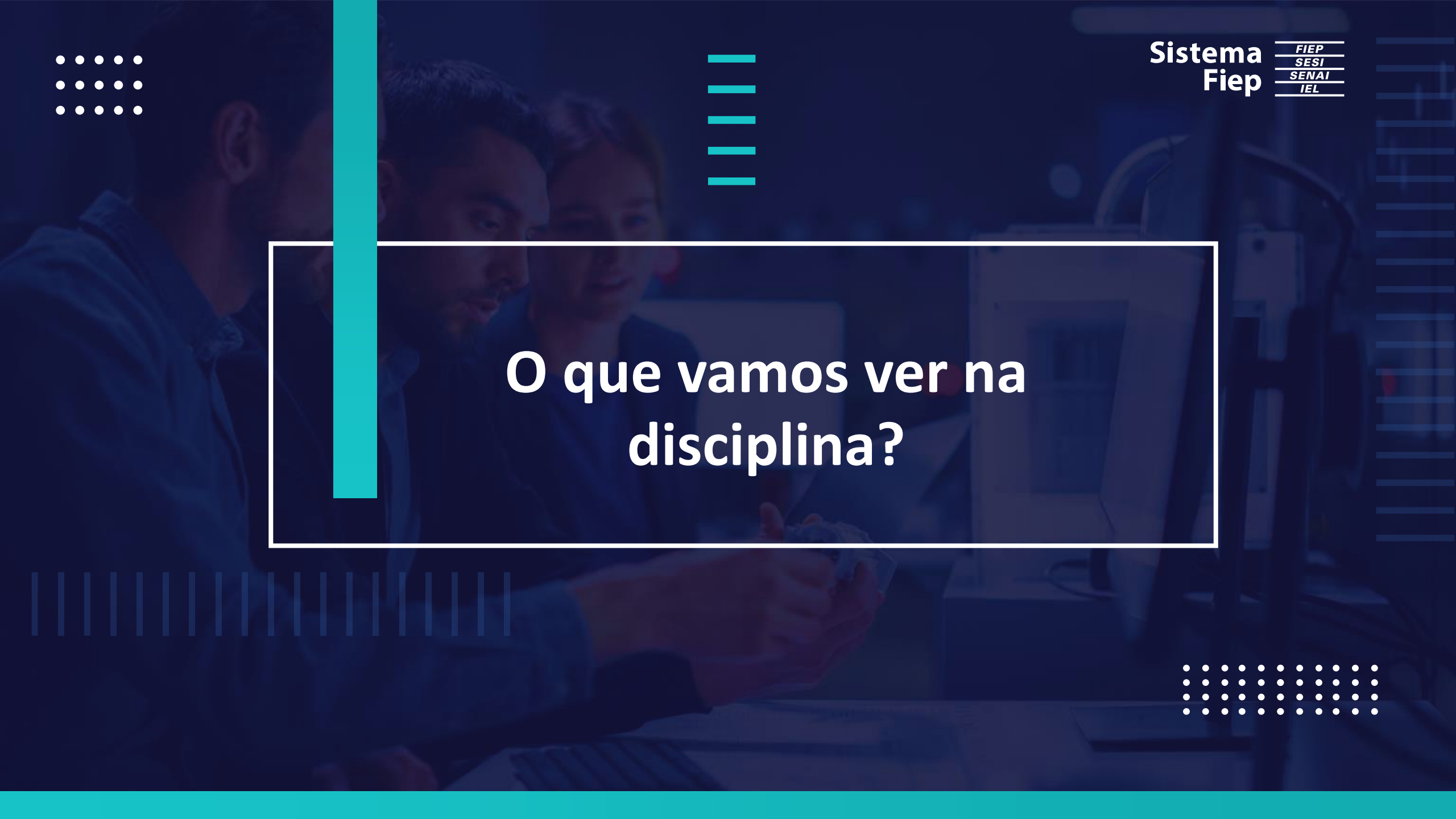


I. **Analisar** os resultados obtidos a partir dos modelos e explorar as suas implicações;

II. **Comunicar** os *insights* para os *stakeholders* e tomadores de decisão;

III. **Discutir** os resultados e recomendar ações futuras.



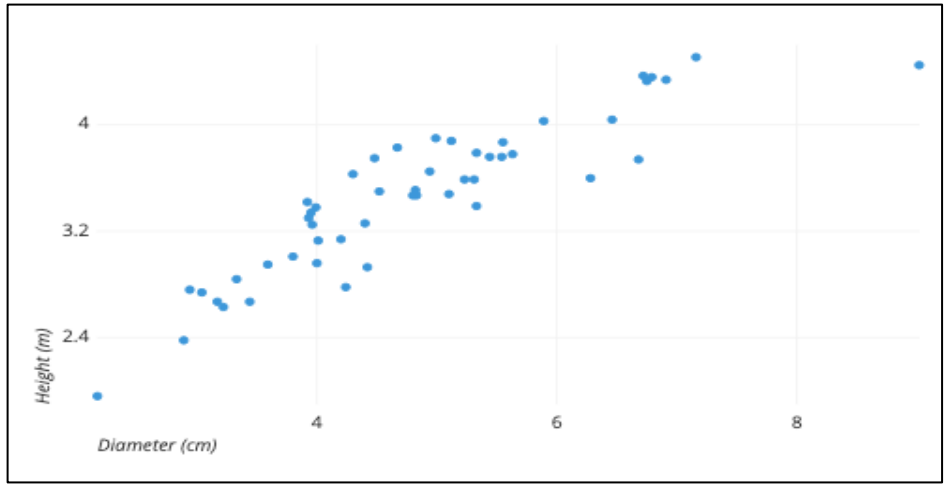


O que vamos ver na disciplina?

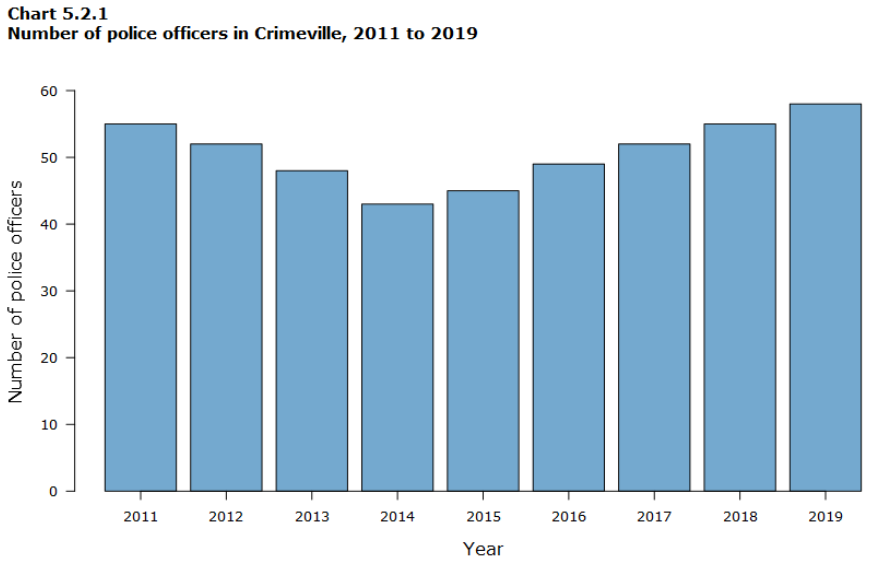
Tipos de dados	Medidas de Tendência Central	Medidas de Dispersão	Distribuição de Frequências	Outros
• Numéricos • Categóricos	• Média • Mediana • Moda	• Variância • Desvio Padrão • Amplitude	• Histograma • Tabela de frequência	• Probabilidade • Correlação • Testes de hipóteses

Aprendizado	Definição	Objetivo
Supervisionado	Algoritmos que usam dados rotulados (entrada e saída conhecida) para aprender a mapear novos dados e fazer previsões ou classificações.	Prever ou classificar novos dados com base nos padrões aprendidos a partir dos dados rotulados.
Não Supervisionado	Algoritmos que exploram padrões e estruturas nos dados não rotulados para realizar agrupamentos ou redução de dimensionalidade.	Descobrir estruturas ocultas nos dados, como agrupamentos naturais ou características relevantes.

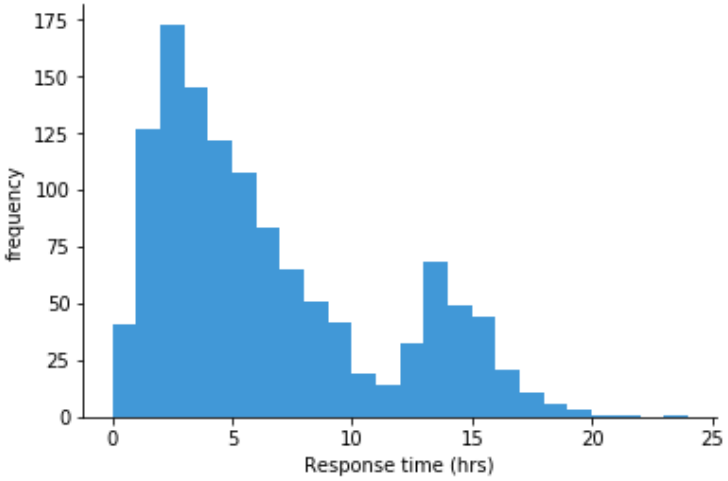
Aprendizado	Algoritmos	Aplicações
Supervisionado	Regressão Linear, Árvores de Decisão, Máquinas de Vetores de Suporte (SVM), Redes Neurais, k-Vizinhos Mais Próximos (KNN)	Classificação de e-mails como spam ou não spam, previsão de preços de imóveis, diagnóstico médico, etc.
Não Supervisionado	K-Means, Agrupamento Hierárquico, Análise de Componentes Principais (PCA), Algoritmos de Redução de Dimensionalidade, etc.	Segmentação de clientes, detecção de padrões de comportamento, redução de dimensionalidade para visualização, etc.

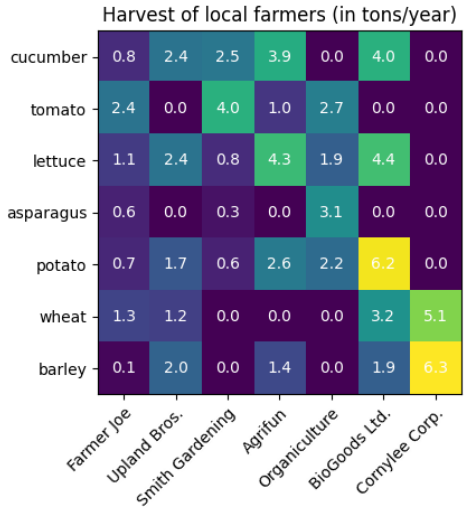


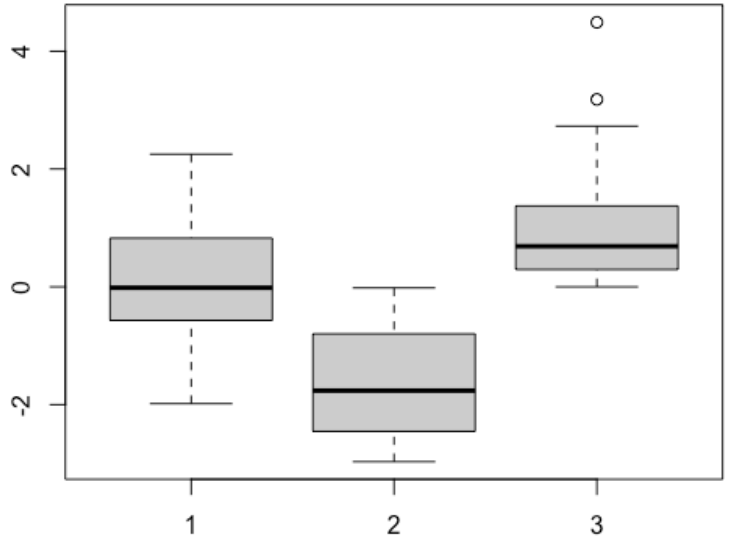
Identificar Relações entre Variáveis



Comparar Categorias



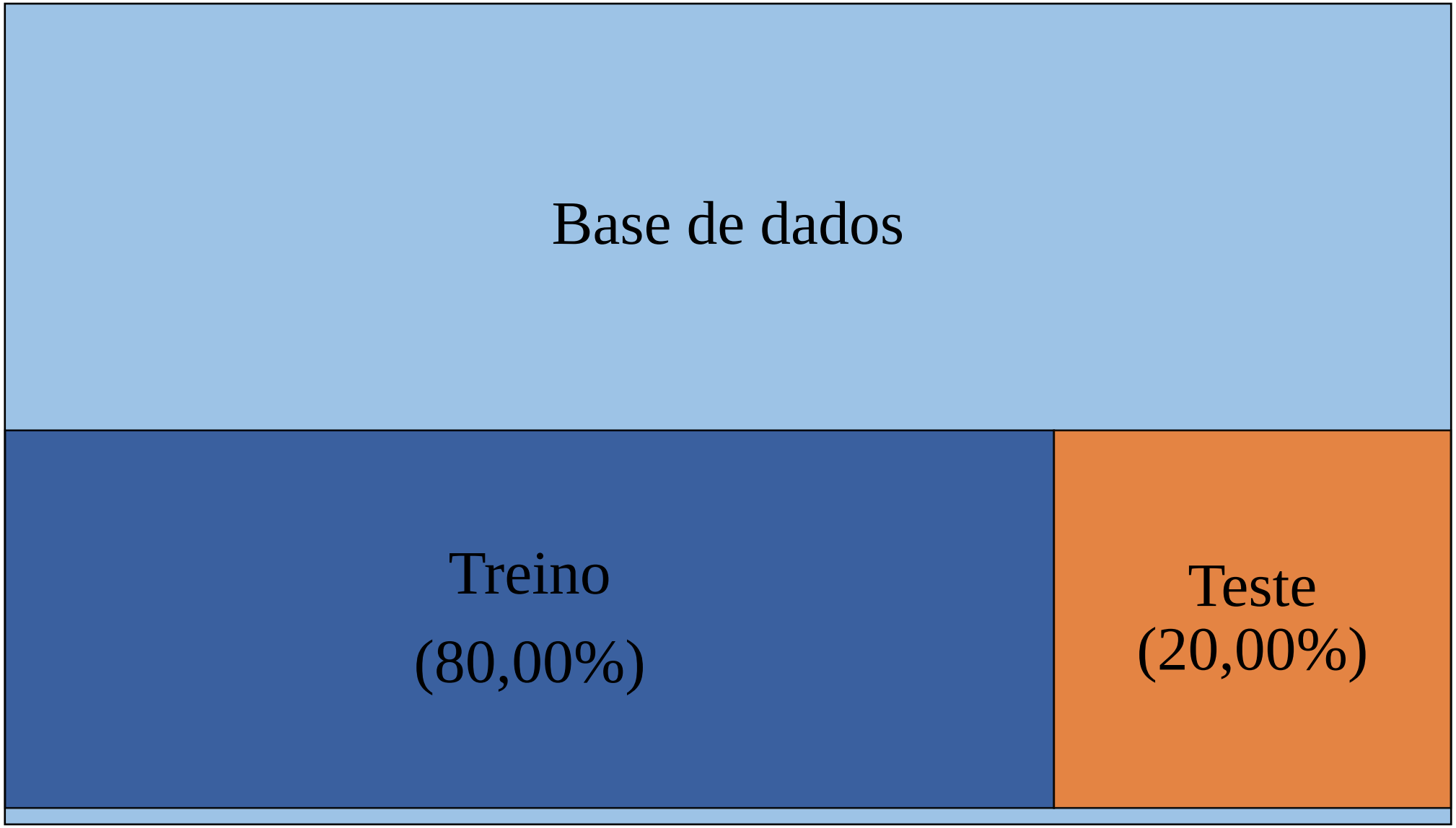


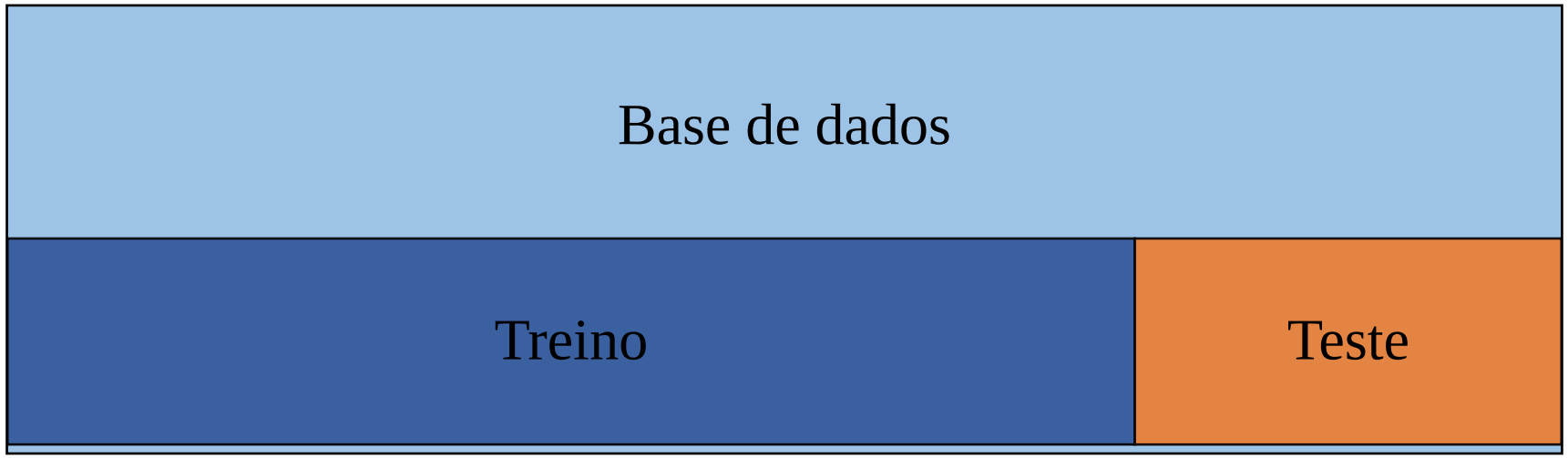


Analisar distribuições

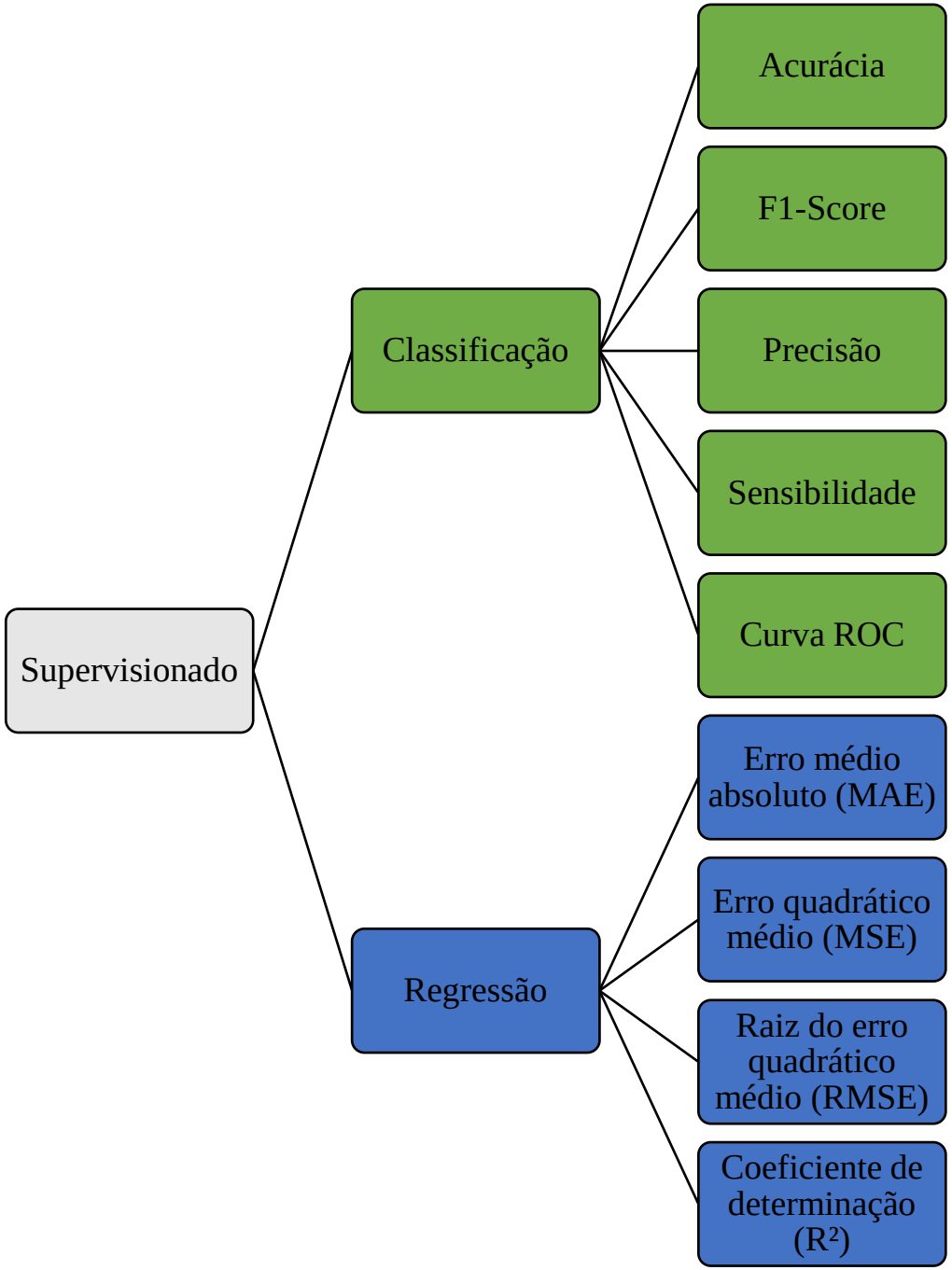
Visualizar correlações

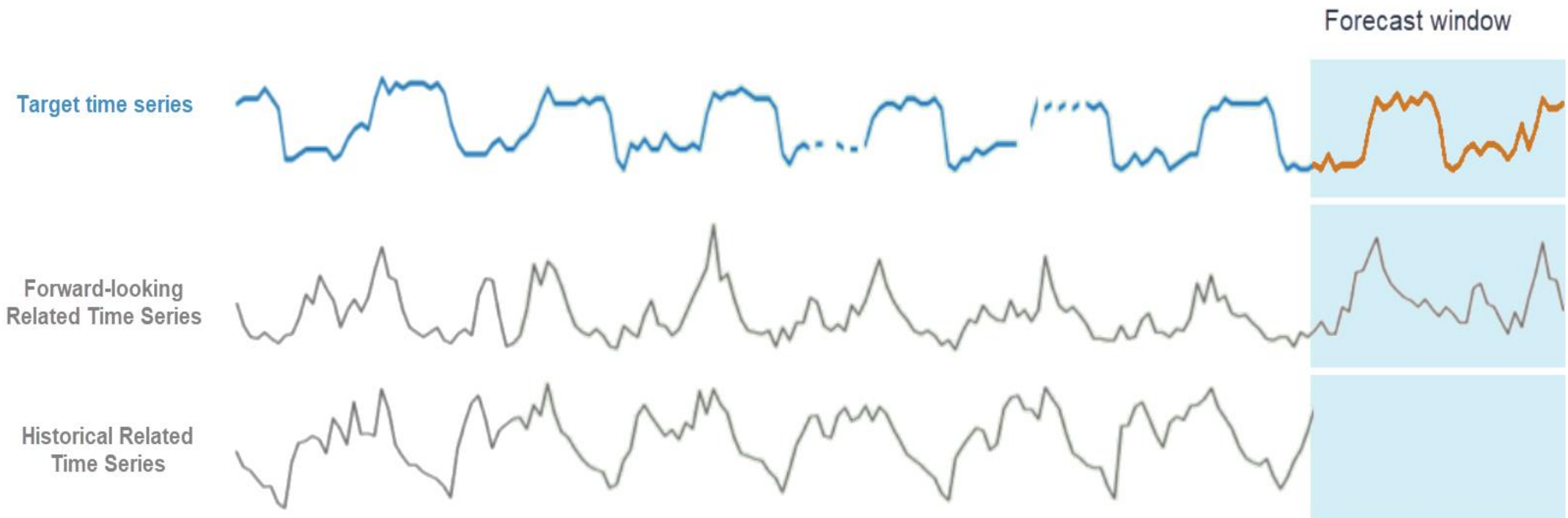
Identificar outliers e variações


Base de dados
Treino
(80,00%)
Teste
(20,00%)



1	**2**	**3**	**4**	**5**
1	**2**	**3**	**4**	**5**
1	**2**	**3**	**4**	**5**
1	**2**	**3**	**4**	**5**
1	**2**	**3**	**4**	**5**


Supervisionado
Classificação
Acurácia
F1-Score
Precisão
Sensibilidade
Curva ROC
Regressão
Erro médio absoluto (MAE)
Erro quadrático médio (MSE)
Raiz do erro quadrático médio (RMSE)
Coeficiente de determinação (R²)


Forecast window
Target time series
Forward-looking
Related Time Series
Historical Related
Time Series

Modelos de Séries Temporais	**Descrição**
Modelos de Médias Móveis (MA)	Consideram a relação entre a observação atual e os erros de previsões passadas.
Modelos Autoregressivos (AR)	Utilizam a relação entre uma observação atual e suas observações anteriores para fazer previsões.
Modelo ARMA (*Autoregressive Moving Average*)	Combina elementos dos modelos AR e MA para lidar com relações autoregressivas e de médias móveis.
Modelo ARIMA (*Autoregressive Integrated Moving Average*)	Incorpora a diferenciação (integração) das observações para tornar a série estacionária e utilizar elementos AR e MA para previsão.
Modelo de Holt-Winters	Incorpora tendência, sazonalidade e componentes de erro para previsão em séries temporais com padrões sazonais.

Técnica	Descrição
One-Hot Encoding	Representação binária em que cada palavra é convertida em um vetor com apenas 0s e 1s.
Bag of Words (BoW)	Vetorização baseada na contagem de frequência das palavras em um texto.
Word Embeddings	Técnicas para mapear palavras em vetores densos e distribuídos, capturando significado.
TF-IDF (*Term Frequency - Inverse Document Frequency*)	Medida que avalia a importância relativa de uma palavra em um documento.

Ferramentas

pandas

NumPy

TensorFlow

kaggle

seaborn

scikit learn

UC Irvine Machine Learning Repository

plotly

NLTK

statsmodels

jupyter

CO