EAD

EDUCAÇÃO A DISTÂNCIA

# UNIDADE I

## Ciência de Dados

Prof. Me. João Cruz

# Definição

## O que é ciência de dados?

- A Ciência de Dados é um campo interdisciplinar que combina técnicas estatísticas, matemáticas e de programação para extrair insights e conhecimentos úteis a partir de conjuntos de dados complexos.

- Ela envolve a coleta, organização, processamento e análise de grandes volumes de dados, com o objetivo de identificar padrões, fazer previsões e tomar decisões embasadas em evidências.

# Etapas da Ciência de Dados

A visão geral da Ciência de Dados abrange várias etapas.

- A primeira delas é a **coleta de dados** (pode ser feita por meio de várias fontes, como sensores, bancos de dados, mídias sociais, entre outros).

- A próxima é o **processo de limpeza** e organização dos dados (remoção de ruídos, erros e inconsistências)
  - Nessa etapa é que os dados são estruturados de forma adequada para análise.

    - A terceira etapa é a **exploração dos dados** (são usadas diferentes técnicas estatísticas e de visualização, aplicadas para entender os padrões e relações presentes nos dados).
      - Nessa etapa também é possível identificar tendências, correlações e outliers e insights.

# Ciência de Dados e Big Data (Quais as diferenças?)

## Ciência de Dados

- A ciência de dados pode ser definida como a disciplina que fornece princípios, metodologias e orientações para transformação, validação, análise e criação de significado a partir de dados.
- O objetivo é extrair conhecimento de conjuntos de dados usando as análises estatísticas tradicionais, algoritmos e ferramentas. As mesmas técnicas podem ser usadas em pequenos e grandes volumes de dados (Big Data).

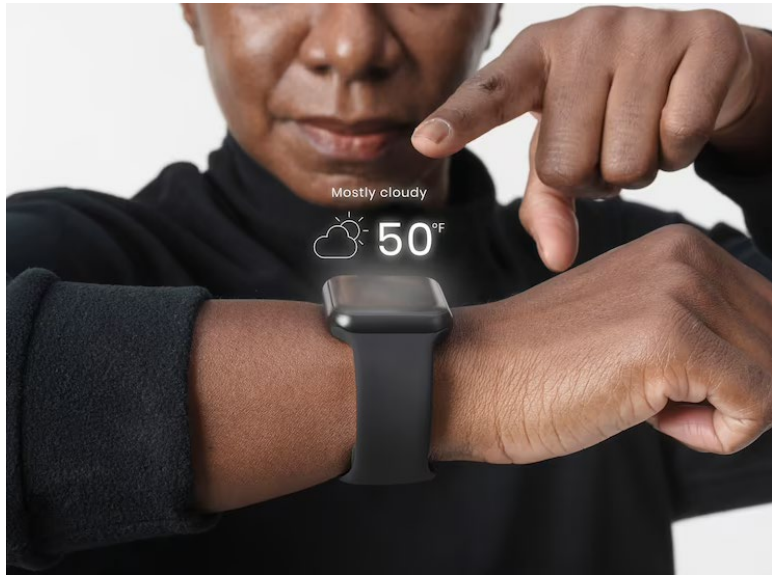
## Big Data

- Refere-se à enorme quantidade de dados gerados a partir de várias fontes, como transações comerciais, mídias sociais, sensores, dispositivos móveis , entre outros.

Podemos classificar uma fonte de dados como Big Data quando utilizamos os 3 Vs:

- **Volume** (grande quantidade de dados).
- **Velocidade** (gerados em alta velocidade).
- **Variedade** (diversidade de tipos e formatos de dados).

# Big Data não tem nada a ver com Ciência de Dados?




Fonte: https://x.gd/9gtO6

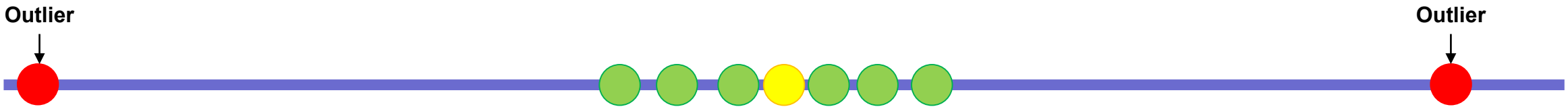
- É importante lembrar que a Ciência de Dados, embora não seja a mesma coisa que Big Data, está intimamente ligada às metodologias para a análise dos dados extraídos.

- No material didático (Big Data e Ciência de dados – além do hype), é possível compreender de forma mais clara como as novas tecnologias implementadas em dispositivos como eletrodomésticos, smartphones, câmeras de segurança e até wearables (dispositivos vestíveis) são compreendidos pela ciência como fontes geradoras de Big Datas.

# Insights e Outliers, para que servem?

## Outliers

- São valores que se diferenciam significativamente do restante dos dados em um conjunto.
- Esses valores extremos estão longe da média ou dos demais valores do conjunto e podem ser causados por erros de medição, comportamentos anômalos ou eventos raros.
- Outliers podem distorcer a análise de dados e afetar negativamente a precisão de modelos e estatísticas descritivas.




Fonte: autoria própria.

- A detecção e o tratamento de outliers são importantes em várias aplicações, pois podem indicar erros de coleta de dados, indicar a presença de eventos incomuns ou fornecer insights valiosos sobre comportamentos excepcionais no conjunto de dados.

# Insights e Outliers, para que servem?

Fonte: https://x.gd/xq1I7

## Insights

- São percepções, entendimentos e conclusões significativas e valiosas obtidas a partir da análise de dados ou informações.
- São descobertas que vão além dos dados brutos, revelando padrões, tendências ou relações ocultas que podem levar a novas ideias, melhorias em processos, estratégias de negócios mais eficientes e tomadas de decisões informadas.
- Podem ser alcançados por meio de diversas técnicas de análise de dados, como estatísticas descritivas, mineração de dados, aprendizado de máquina e visualização de dados.
- Os insights são valiosos para orientar ações e estratégias de negócios, identificar oportunidades e desafios, prever tendências futuras e entender melhor o comportamento dos clientes e usuários.

# Tomada de Decisão Orientada por dados?


Fonte: autoria própria.

- A Tomada de Decisão Orientada por Dados (DOD) é uma prática na qual as decisões são embasadas na análise de dados, em vez de dependerem apenas da intuição dos executivos da alta direção.
- A DOD não é uma prática de “tudo ou nada”, e muitas empresas a adotam em diferentes graus, dependendo das suas necessidades e recursos.
- A DOD se baseia nos dados, em análises e técnicas de estatística e probabilidade para indicar tendências, insights e outliers que somados ao know-how dos executivos da alta direção podem apoiar a tomada de decisão estratégica.

# Problemas e desafios da ciência de dados

## Qualidade dos dados

- A qualidade dos dados é um desafio fundamental na Ciência de Dados. Os dados podem conter erros, estar incompletos, ser inconsistentes ou conter viés.
- A falta de qualidade dos dados pode levar a conclusões errôneas e afetar a confiabilidade dos resultados obtidos.

## Privacidade e ética

- A coleta e o uso de dados envolvem questões de privacidade e ética.
- Ao lidar com dados sensíveis, como informações pessoais dos usuários, é essencial garantir a privacidade e a segurança dos dados.
- Além disso, é necessário considerar o viés nos dados e nos modelos para evitar discriminação ou resultados distorcidos.

# Problemas e desafios da ciência de dados

## Escalabilidade

- A Ciência de Dados lida com conjuntos de dados cada vez maiores (Big Data).
- Desafios da infraestrutura, do armazenamento e do processamento e análise em relação a conjuntos de dados que crescem exponencialmente.
- Novos algoritmos que lidem com a escalabilidade e os grandes volumes são essenciais.

## Complexidade dos algoritmos e modelos

- A escolha e implementação de algoritmos adequados para análise e modelagem de dados específicos é com certeza um grande desafio.
- Existem muitos algoritmos e modelos disponíveis, cada um com suas vantagens e desvantagens, e é necessário entender as características dos dados e os requisitos do problema para selecionar a abordagem mais apropriada.

# Problemas e desafios da ciência de dados

## Interpretação e comunicação dos resultados

- A interpretação correta dos resultados e a capacidade de explicar as descobertas de maneira não técnica são habilidades importantes para garantir que os insights sejam compreendidos e utilizados corretamente.

## Escassez de talentos

- Existe uma demanda crescente por profissionais qualificados em Ciência de Dados, mas há uma escassez de talentos nessa área.
- Encontrar e contratar cientistas de dados, engenheiros de dados e analistas com habilidades técnicas e conhecimentos de negócios é um desafio enfrentado por muitas organizações.

# Problemas e desafios da ciência de dados

**Mudanças rápidas de tecnologia:**

- Como um campo em constante evolução, com avanços tecnológicos e novas técnicas surgindo regularmente, um desafio para Engenheiros e Analistas de Dados é manter-se atualizado com as últimas tecnologias, ferramentas e técnicas, requer um esforço contínuo de aprendizado e desenvolvimento profissional.

# Soluções baseadas em dados

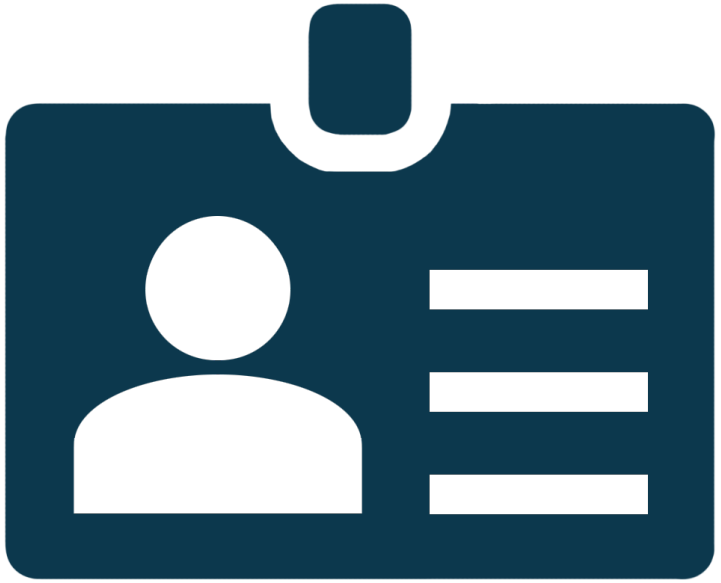
## Soluções da dados

- Limpeza e pré-processamento de dados
- Análise exploratória de dados
- Modelagem preditiva
- Segmentação e personalização
- Detecção de anomalias e fraudes
- Otimização de processos
- Visualização de dados

## Abordagens que também são consideradas soluções baseadas em dados

- Text mining
- Processamento de linguagem natural
- Aprendizado de máquina interpretável
- Aprendizado de reforço
- Dados em tempo real
- Automação de processos

# Habilidades de um profissional de Ciência de Dados



- Conhecimento em programação
- Estatística e matemática
- Aprendizado de máquina e mineração de dados
- Conhecimento de bancos de dados
- Visualização de dados
- Domínio do negócio
- Pensamento analítico e resolução de problemas
- Comunicação e habilidades interpessoais

Fonte: autoria própria.

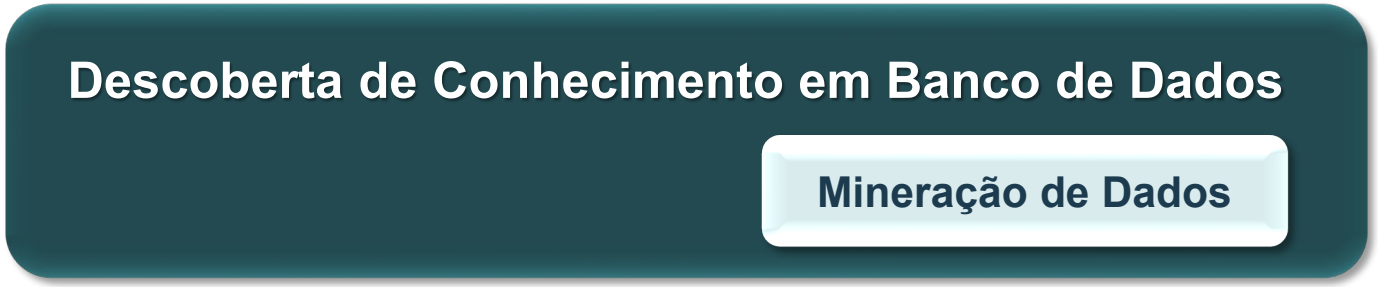
# (KDD) Descoberta de Conhecimentos em Banco de Dados

- **Descoberta de Conhecimento em Bancos de Dados** ou KDD (Knowledge Discovery in Databases) refere-se ao processo de identificar padrões, conhecimentos úteis e informações ocultas em grandes volumes de dados.

- Esse processo abrange várias etapas, incluindo seleção e pré-processamento de dados, transformação, mineração de dados para descoberta de padrões, avaliação dos resultados e interpretação dos achados.

- O objetivo é transformar dados brutos em informações significativas e conhecimento acionável.

Fonte: livro-texto.




  - O termo "**mineração de dados**" (Data Mining) refere-se ao estágio de descoberta do processo de KDD.

# Etapas do KDD

**Seleção de dados**

- Nesta etapa, os **dados relevantes** são **identificados** e **selecionados** para a análise.
- Isso envolve a **definição** de **critérios de inclusão** e **exclusão** e a obtenção dos conjuntos de dados adequados para o problema em questão.

**Pré-processamento de dados**

- Os **dados brutos** podem ser complexos, inconsistentes ou conter **ruído**.
- Nesta etapa, ocorre **a limpeza** e a **transformação dos dados**, incluindo a remoção de dados ausentes ou duplicados, normalização, discretização e outras técnicas de preparação dos dados para análise.

# Etapas do KDD

## Transformação de dados

- A transformação de dados é feita para representar os dados em uma forma mais adequada para análise.
- Isso pode envolver a agregação de dados, a criação de novos atributos ou a redução da dimensionalidade por meio de técnicas como análise de componentes principais (PCA) ou seleção de recursos.

## Mineração de dados

- A mineração de dados é a etapa central do processo de descoberta de conhecimento.
- Nessa etapa, são aplicadas técnicas e algoritmos de aprendizado de máquina, estatística e visualização de dados para identificar padrões, tendências, associações ou relações interessantes nos dados.
- Isso pode incluir técnicas como classificação, regressão, clusterização, regras de associação, redes neurais, entre outras.

# Etapas do KDD

**Avaliação e interpretação dos resultados**

- Após a aplicação das técnicas de mineração de dados, os resultados obtidos são avaliados e interpretados.
- Isso envolve a análise dos padrões descobertos, a validação dos modelos construídos e a interpretação dos insights obtidos em termos do problema ou domínio específico em questão.

**Utilização e aplicação dos conhecimentos**

- Os conhecimentos e insights descobertos durante o processo são utilizados para tomar decisões informadas, desenvolver estratégias, resolver problemas e promover melhorias nos negócios ou em outras áreas de aplicação.

# Vantagens do KDD

- Descobertas de preferências de mercado.
- Associações e tendências ocultas nos dados
- Tendências de vendas
- Insights de comportamentos de clientes
- Decisões em dados e evidências sólidas.
- Maior visão dos riscos financeiros
- Melhoria da produtividade
- Maior eficiência operacional

Fonte: autoria própria.

# Interatividade

Analise as afirmações a seguir:

I. O constante crescimento das bases de dados (Big Data) é um problema, porque torna os algoritmos desenvolvidos pela Ciência de dados defasados muito rapidamente.

II. O DOD vem se tornando cada vez mais um diferencial para as grandes corporações, pois permite que a tomada de decisão não se baseie apenas no know-how dos executivos seniores.

III. Os *Big Datas* são uma ciência diferente da Ciência de Dados, mesmo estando intimamente ligadas, uma vez que a primeira estuda apenas bases de dados com grande volume, velocidade e variedade, enquanto a segunda bases de dados estruturadas e corporativas.

Analisando as três afirmações, qual das alternativas abaixo está correta?

a) Apenas a afirmação I é verdadeira.
b) Apenas a afirmação II é verdadeira.
c) Apenas a afirmação III é verdadeira.
d) Apenas as afirmações II e III são verdadeiras.
e) Todas as afirmações são verdadeiras.

# Resposta

Analise as afirmações a seguir:

I. O constante crescimento das bases de dados (Big Data) é um problema, porque torna os algoritmos desenvolvidos pela Ciência de dados defasados muito rapidamente.

II. O DOD vem se tornando cada vez mais um diferencial para as grandes corporações, pois permite que a tomada de decisão não se baseie apenas no know-how dos executivos seniores.

III. Os *Big Datas* são uma ciência diferente da Ciência de Dados, mesmo estando intimamente ligadas, uma vez que a primeira estuda apenas bases de dados com grande volume, velocidade e variedade, enquanto a segunda bases de dados estruturadas e corporativas.

Analisando as três afirmações, qual das alternativas abaixo está correta?

a) Apenas a afirmação I é verdadeira.
b) Apenas a afirmação II é verdadeira.
c) Apenas a afirmação III é verdadeira.
d) Apenas as afirmações II e III são verdadeiras.
e) Todas as afirmações são verdadeiras.

# CRISP-DM (Cross Industry Standard Process for Data Mining)

- O CRISP-DM é, segundo Chapman (2000), um modelo de processo muito utilizado na área de mineração de dados para guiar projetos de análise de dados.

- Ele fornece uma estrutura flexível e abrangente para a condução de projetos de mineração de dados, permitindo que as equipes enfrentem desafios complexos e tomem decisões informadas ao longo de todo o ciclo de vida do projeto.

- O CRISP-DM é composto por seis etapas interconectadas.

# CRISP-DM (Cross Industry Standard Process for Data Mining)

## Entendimento do Negócio

- Nesta fase inicial, a equipe trabalha para compreender os objetivos e requisitos do projeto, identificando como a mineração de dados pode contribuir para as metas de negócios.

## Entendimento dos Dados

- Nesta fase, os dados disponíveis são explorados e analisados para identificar sua qualidade, relevância e potencial para atender aos objetivos do projeto. Isso envolve a realização de análises exploratórias e a compreensão das características dos dados.

## Preparação dos Dados

- Aqui os dados são limpos, transformados e preparados para análise. Isso inclui lidar com valores ausentes, normalização, seleção de atributos relevantes e outras tarefas de preparação.

# CRISP-DM (Cross Industry Standard Process for Data Mining)

## Modelagem

- Nesta fase, são desenvolvidos modelos de mineração de dados, como algoritmos de aprendizado de máquina, para explorar os padrões e relacionamentos nos dados. Diferentes abordagens são testadas e avaliadas para encontrar a mais adequada.

## Avaliação

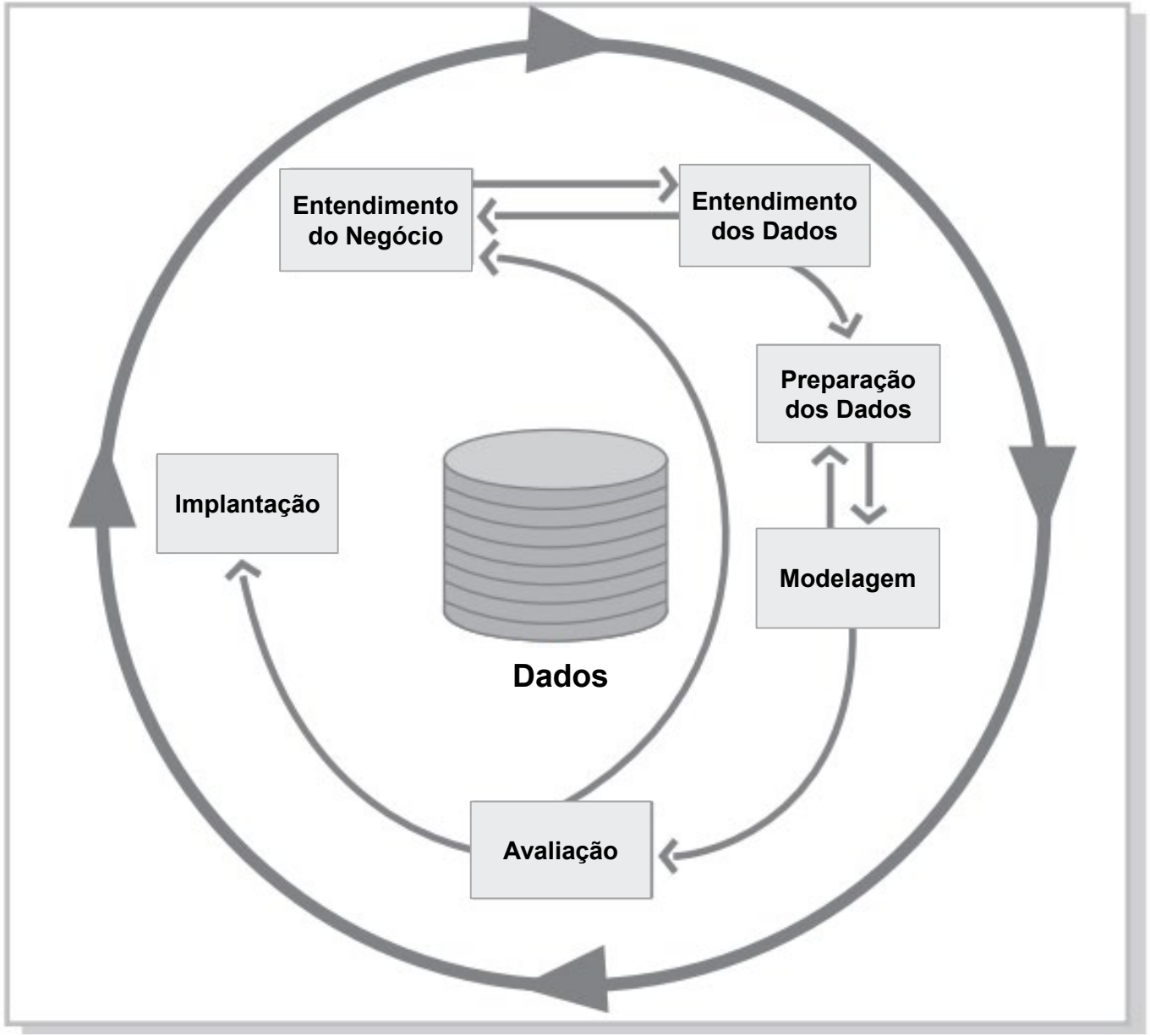
- Os modelos construídos na fase de modelagem são avaliados para garantir que eles atendam aos critérios de sucesso do projeto. Isso pode envolver testes de desempenho, validação cruzada e outros métodos para garantir que os modelos sejam robustos e generalizáveis.

# CRISP-DM (Cross Industry Standard Process for Data Mining)

## Implantação

- Nesta última fase, os resultados da análise são apresentados aos stakeholders e são tomadas medidas para implementar os insights obtidos no ambiente de negócios. Isso pode envolver a criação de relatórios, integração com sistemas existentes ou outras formas de utilização prática.

# CRISP-DM (Cross Industry Standard Process for Data Mining)




Fases do modelo de referência CRISP-DM

Fonte: Chapman (2000, p. 12).

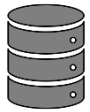
# Extração de conhecimento

- A extração de conhecimento envolve a aplicação de algoritmos e técnicas de descoberta de padrões, associações e tendências nos dados para identificar informações valiosas e conhecimento útil.

Destacam-se:

- Mineração de Dados.
- Aprendizado de Máquina.
- Processamento de Linguagem Natural.
- Visualização de Dados.

# Fontes de Dados



Bases de Dados Estruturadas

Bases de Dados Não Estruturadas

Dados de Sensores e Dispositivos IoT

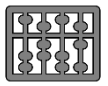
Dados de Mídias Sociais

Dados de Texto e Documentos

Dados de Streaming

Dados de Fontes Externas

Dados Geoespaciais



Dados Transacionais

Fonte: Microsoft Power Point.

# Visão Geral sobre Aprendizado de Máquina

## Definição

- Aprendizado de Máquina, também conhecido como Machine Learning, é uma subárea da inteligência artificial que se concentra no desenvolvimento de algoritmos e modelos capazes de aprender e tomar decisões a partir dos dados, sem serem explicitamente programados.

- O objetivo principal do Aprendizado de Máquina é permitir que os sistemas “aprendam” automaticamente a partir dos dados e melhorem seu desempenho ao longo do tempo, sem a necessidade de regras ou instruções específicas.

# Categorias do Machine Learning

## Aprendizado Supervisionado

- Os algoritmos são **treinados** utilizando um conjunto de dados de entrada pré-rotulados (dados de treinamento).
- O objetivo é aprender a **relação entre as entradas e as saídas** correspondentes, para que o modelo seja capaz de fazer **previsões** ou **tomar decisões** em novos dados não vistos anteriormente.

## Aprendizado Não Supervisionado:

- Os algoritmos são aplicados a conjuntos de dados **sem informações prévias sobre as saídas** desejadas.
- O objetivo é descobrir estruturas, padrões ou grupos intrínsecos nos dados, fornecendo uma visão mais profunda dos dados e insights sobre seu comportamento.

# Categorias do Machine Learning

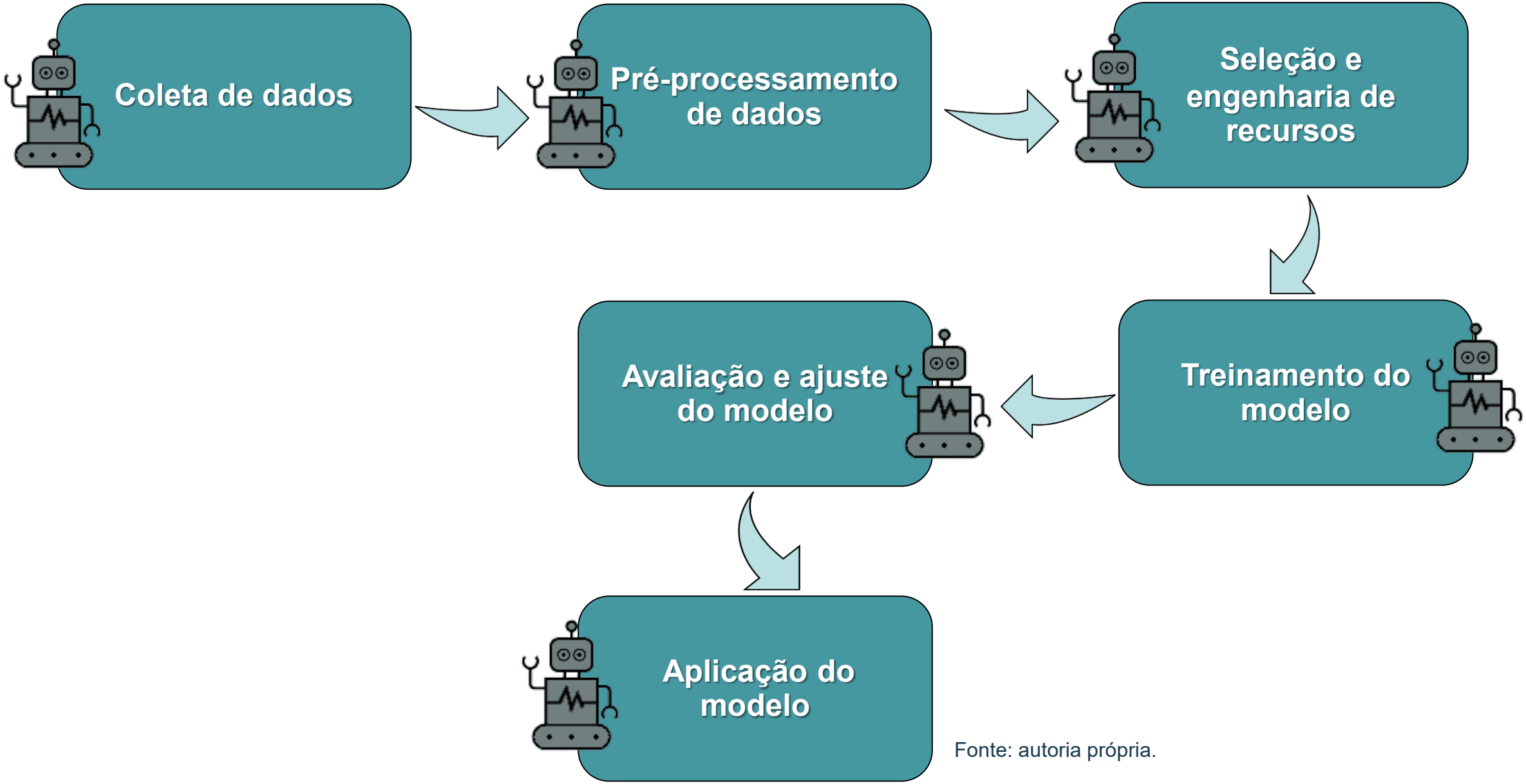
## Aprendizado semissupervisionado

- É usado para as mesmas finalidades que o aprendizado supervisionado, porém **envolve** tanto **dados com rótulos como sem rótulos** para treinamento.
- Ele pode ser aplicado a tarefas como classificação, regressão e previsão. É vantajoso quando rotular todos os dados é caro demais.

## Aprendizado por Reforço

- Envolve o treinamento de algoritmos por meio de interações com um ambiente. O **agente de aprendizado** toma ações em um ambiente e recebe recompensas ou punições com base no desempenho de suas ações.
- O objetivo é **maximizar as recompensas ao longo do tempo**, aprendendo a melhor política de ação.
- Isso é frequentemente aplicado em jogos, robótica e otimização de processos.

# Etapas processo de Aprendizado de Máquinas




Fonte: autoria própria.

# Overfiting e Underfitting

## Overfitting (sobreajuste)

- Ocorre quando o modelo **se torna excessivamente complexo** e se **ajusta** perfeitamente aos **dados de treinamento**, capturando até mesmo o ruído presente nesses dados.
- Como resultado, o **modelo memoriza** os exemplos de treinamento **em vez de aprender os padrões** subjacentes que permitem generalizar para novos dados.
- Isso pode levar a um desempenho pobre na etapa de teste, em que o modelo falha em fazer previsões precisas em dados não vistos.

**Sinais de Overfitting incluem**:

- O desempenho do modelo é excelente nos dados de treinamento, mas ruim nos dados de teste.
- O modelo possui uma complexidade excessiva em relação ao tamanho dos dados disponíveis.
- O modelo captura o ruído presente nos dados de treinamento, resultando em uma precisão excessivamente alta nesses dados, mas não em novos dados.

# Overfiting e Underfitting

## Underfitting (subajuste)

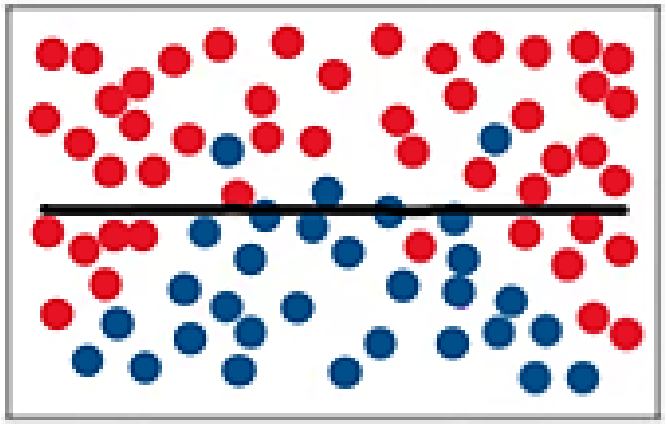
- Ocorre quando o modelo é muito simples ou não é capaz de capturar os padrões presentes nos dados de treinamento.
- Nesse caso, o modelo não consegue se ajustar adequadamente aos dados e acaba subestimando a complexidade do problema.
- O resultado é um desempenho insatisfatório tanto nos dados de treinamento quanto nos dados de teste.

**Sinais de Underfitting incluem:**

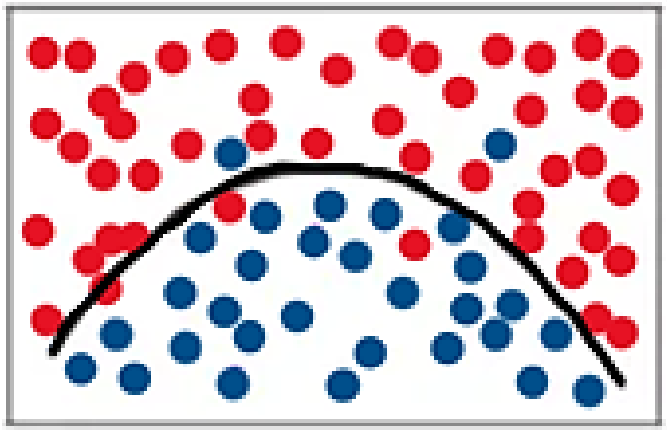
- O desempenho do modelo é ruim tanto nos dados de treinamento quanto nos dados de teste.
- O modelo não consegue capturar os padrões e relações importantes presentes nos dados.
- O modelo é muito simples em relação à complexidade do problema.

# Overfiting e Underfitting

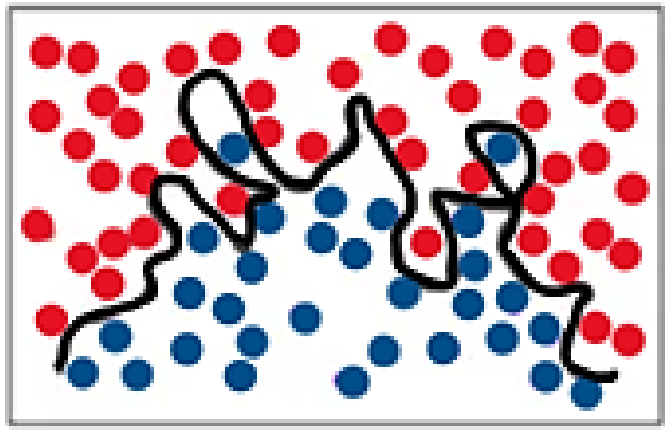
**Underfitting**



**Equilíbrio**



**Overfitting**



Fonte: livro-texto.

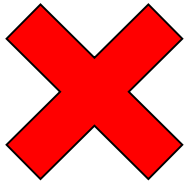
**Outras técnicas para evitar Underfitting e Overfitting:**

- Ajuste de hiperparâmetros.
- Ensemble Learning.
- Cross-Validation.
- Aumento de dados (Data Augmentation).

# Conceitos relacionados a Machine Learning

- Balanço entre Viés e Variância em modelos de Machine Learning.

O viés de um modelo é a simplificação ou suposições errôneas que ele faz sobre os dados



A variância de um modelo refere-se à sensibilidade excessiva do modelo aos dados de treinamento

O objetivo é encontrar um equilíbrio entre viés e variância, em que o modelo seja suficientemente complexo para capturar os padrões importantes nos dados, mas não seja excessivamente complexo a ponto de se ajustar ao ruído presente nos dados de treinamento

Fonte: autoria própria.

# Conceitos relacionados a Machine Learning

## Sistemas de Aprendizado

- Um sistema de aprendizado de Machine Learning é um conjunto de componentes e algoritmos que permitem que uma máquina aprenda a partir dos dados e faça previsões ou tomadas de decisão com base nesse aprendizado.
- O objetivo é capacitar a máquina a reconhecer padrões, extrair informações úteis e melhorar seu desempenho ao longo do tempo, sem ser explicitamente programada para cada tarefa específica.

### Tipos de aprendizagem

- Aprendizado Supervisionado.
- Aprendizado Não Supervisionado.
- Aprendizado por Reforço.
- Aprendizado Semissupervisionado.

# Conceitos relacionados a Machine Learning

## Espaço de Hipóteses

- O espaço de hipóteses se refere ao conjunto de todas as possíveis funções ou modelos que um algoritmo de aprendizado pode escolher como solução para um determinado problema.

- Essas hipóteses são expressas por meio de parâmetros, pesos ou estruturas específicas, dependendo do algoritmo e do tipo de aprendizado utilizado.

    - O espaço de hipóteses define as restrições sobre o conjunto de soluções possíveis que o algoritmo de aprendizado pode explorar durante o processo de treinamento.

# Conceitos relacionados a Machine Learning

## Espaço de Hipóteses

- É importante destacar que a escolha do espaço de hipóteses pode afetar o desempenho do modelo.

- Se o espaço de hipóteses for muito restrito, o modelo pode não ter capacidade suficiente para capturar a complexidade dos dados.

    - Por outro lado, se o espaço de hipóteses for muito amplo, o modelo pode se tornar excessivamente complexo e se ajustar demais aos dados de treinamento, resultando em um sobreajuste.

# Interatividade

Quando falamos em Machine Learning, existem duas categorias de algoritmos utilizados para implementação dessa técnica de extração de conhecimento.

Analise as frases a seguir e indique quais delas são verdadeiras em relação a este contexto:

I. O objetivo do aprendizado supervisionado é que a máquina compreenda a relação entre os dados de entrada e saída no treinamento para poder prever futuras saídas quando receber dados não treinados.

II. Quando falamos de aprendizado supervisionado, não existirá a fase de treinamento, a máquina aprenderá com os feedbacks dos usuários.

III. Em ambos os modelos de aprendizado, mesmo após o treinamento, os modelos podem ser mantidos, proporcionando um aprendizado contínuo.

a) As afirmações I e II são verdadeiras.
b) As afirmações I e III são verdadeiras.
c) As afirmações II e III são verdadeiras.
d) Apenas a afirmação II está correta.
e) Nenhuma das afirmações está correta.

# Resposta

Quando falamos em Machine Learning, existem duas categorias de algoritmos utilizados para implementação dessa técnica de extração de conhecimento.

Analise as frases a seguir e indique quais delas são verdadeiras em relação a este contexto:

I. O objetivo do aprendizado supervisionado é que a máquina compreenda a relação entre os dados de entrada e saída no treinamento para poder prever futuras saídas quando receber dados não treinados.

II. Quando falamos de aprendizado supervisionado, não existirá a fase de treinamento, a máquina aprenderá com os feedbacks dos usuários.

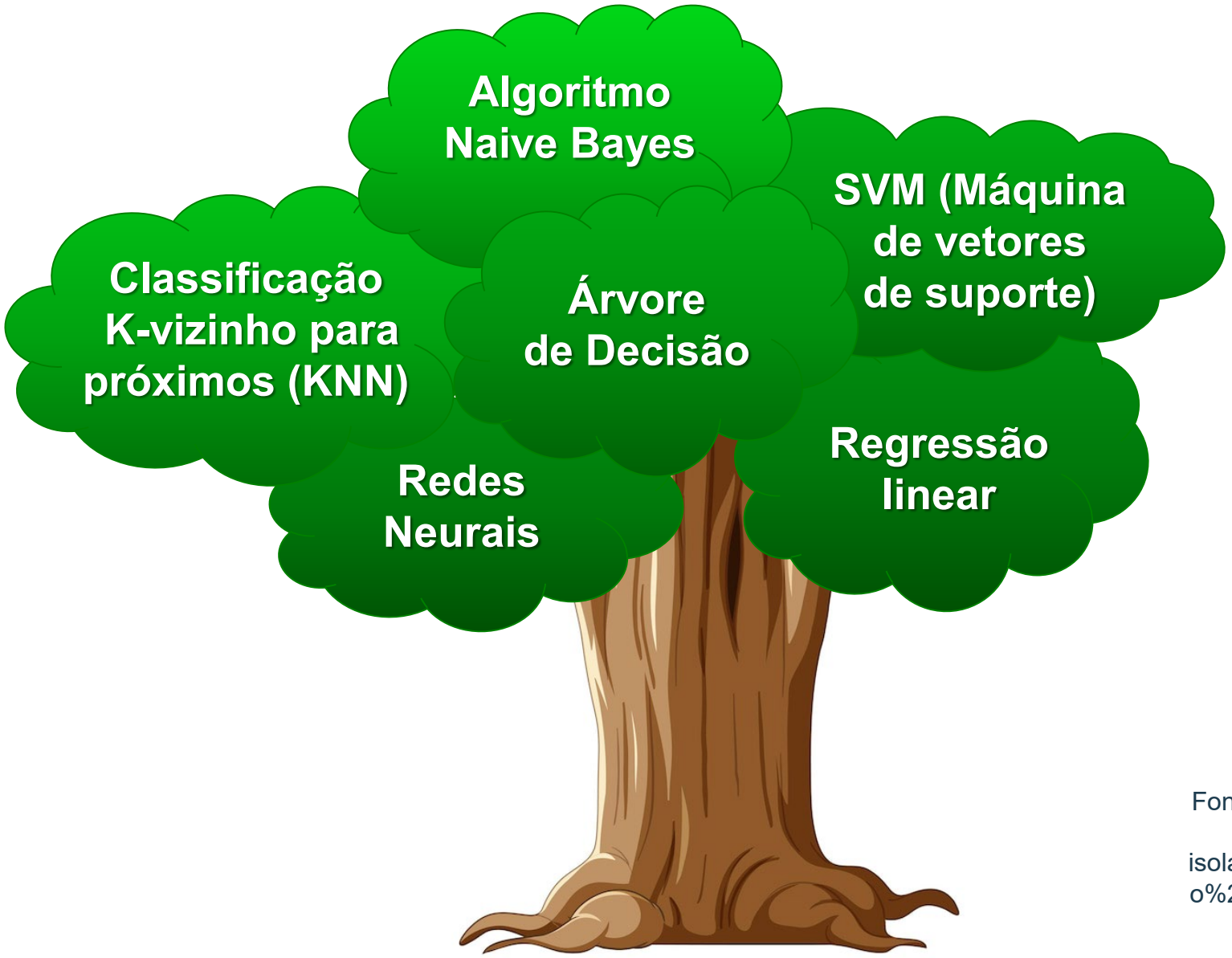
III. Em ambos os modelos de aprendizado, mesmo após o treinamento, os modelos podem ser mantidos, proporcionando um aprendizado contínuo.

a) As afirmações I e II são verdadeiras.
b) As afirmações I e III são verdadeiras.
c) As afirmações II e III são verdadeiras.
d) Apenas a afirmação II está correta.
e) Nenhuma das afirmações está correta.

# Separação didática entre aprendizados de máquina


Fonte: livro-texto.

# Técnicas de algoritmos de aprendizado preditivo

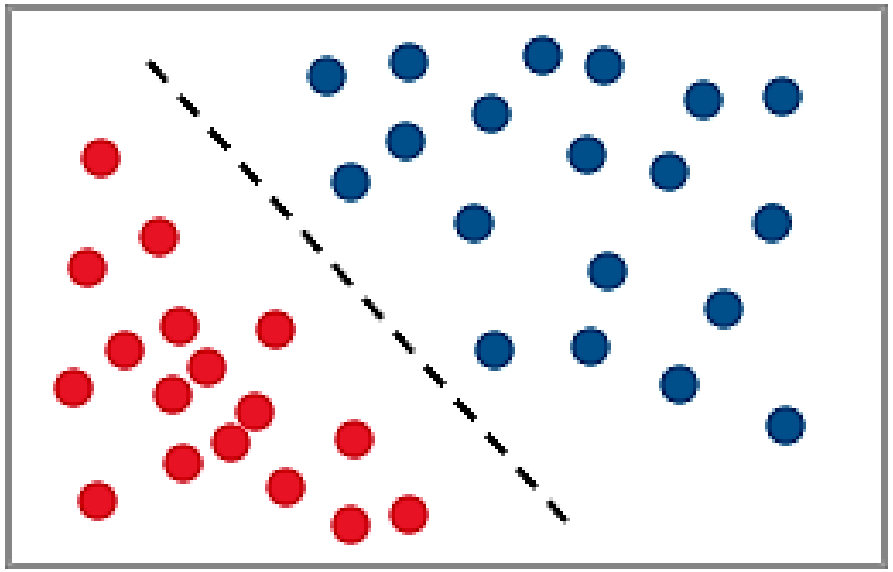



Fonte: https://br.freepik.com/vetores-gratis/tronco-e-raizes-de-arvore-isolados_18218824.htm#query=tronco%20de%20arvore&position=8&from_view=search&track=ais

# Aprendizado Supervisionado – Classificação

## Definição

- A classificação é uma técnica muito importante no campo do aprendizado de máquina, em que o **objetivo é atribuir rótulos** ou **categorias a diferentes instâncias de dados** com base em suas características.
- É uma forma de aprendizado supervisionado em que o algoritmo de aprendizado é treinado usando um conjunto de dados rotulados, em que as classes ou categorias já são conhecidas.



Fonte: livro-texto.

# Exemplos de algoritmos de classificação

- Existem diferentes tipos de algoritmos de classificação, cada um com suas próprias suposições e métodos de classificação.

Vamos ver os mais populares:

**Árvores de decisão**

- Cria uma **estrutura em forma de árvore** que **representa decisões** e **condições** baseadas nas características dos dados.
- Cada **nó da árvore** representa uma característica e cada **ramo** representa uma decisão.
- As **folhas da árvore** correspondem às classes ou categorias finais.

# Exemplos de algoritmos de classificação

**Máquinas de vetores de suporte (SVM)**

- Buscam encontrar um **hiperplano de separação** que maximize a margem entre as classes.
- Ele **mapeia as características de entrada** em um espaço de dimensões superiores e realiza a **classificação com base na posição** dos exemplos nesse espaço.

**k-vizinhos mais próximos (k-NN)**

- Classifica as instâncias de acordo com a classe das instâncias vizinhas mais próximas.
- A distância entre as instâncias é calculada com base nas características e os k-vizinhos mais próximos são considerados para determinar a classe do exemplo em questão.

# Exemplos de algoritmos de classificação

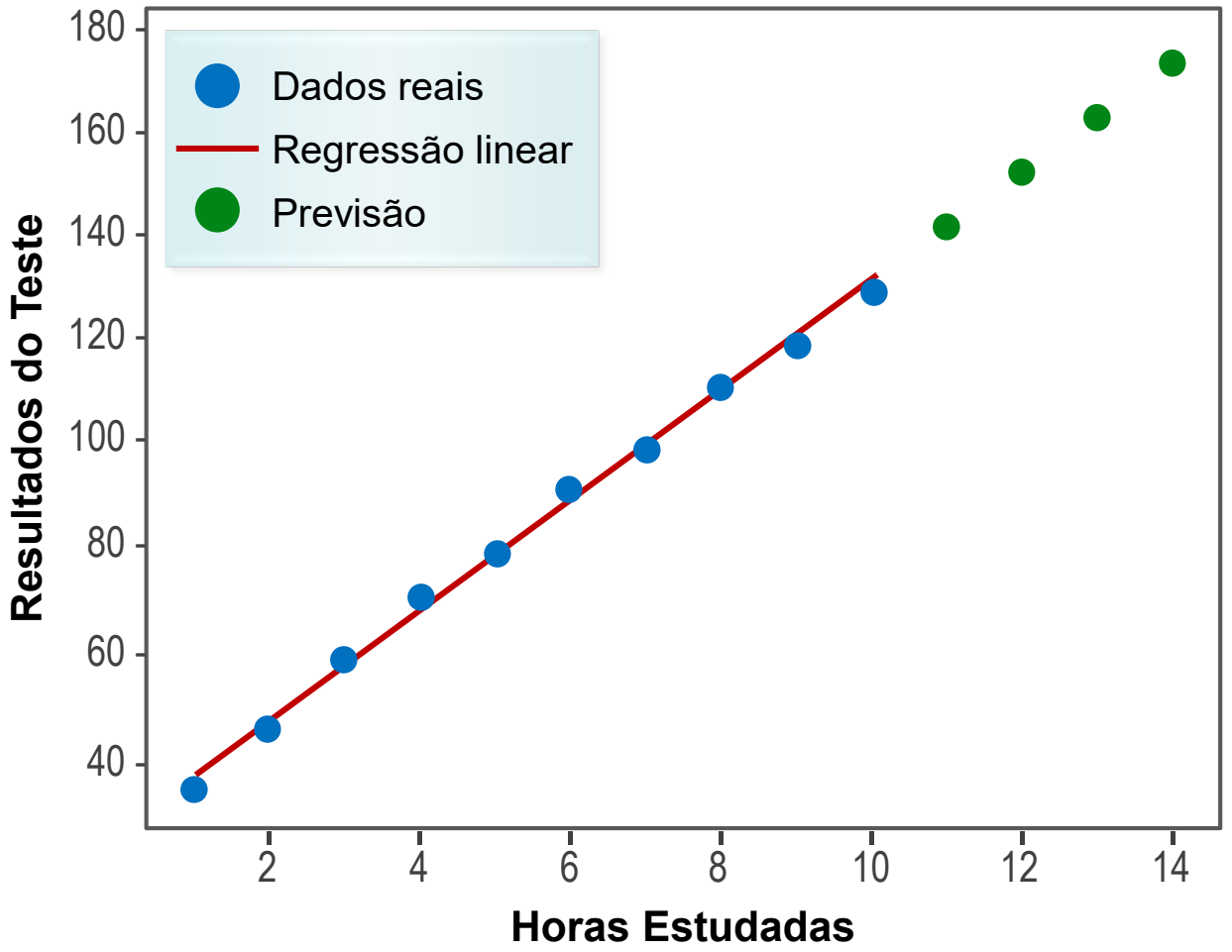
## Redes Neurais

- São modelos inspirados no funcionamento do cérebro humano, compostos por camadas de neurônios interconectados.
- Cada neurônio processa as características de entrada e a rede aprende os pesos das conexões para realizar a classificação.

# Regressão

- É uma **técnica estatística** utilizada no campo do aprendizado de máquina para modelar e **prever relações entre variáveis**.

- A regressão busca prever um valor numérico contínuo com base em um conjunto de variáveis independentes.




Fonte: livro-texto.

# Regressão

- O objetivo da regressão é encontrar uma **função matemática ou estatística** que relacione as variáveis independentes (também chamadas de características ou variáveis de entrada) a uma variável dependente (também conhecida como variável de saída ou variável alvo).

- Essa **função é chamada de modelo de regressão** e é **utilizada para fazer previsões** sobre o valor da variável dependente para novos exemplos de dados.

# Tipos de regressão

- Existem vários tipos de regressão, cada um adequado para diferentes tipos de problemas e dados.

Alguns dos principais tipos de regressão incluem:

**Regressão Linear**

- É um tipo de regressão que assume uma relação linear entre as variáveis independentes e dependentes. O modelo de regressão linear encontra a melhor linha reta que representa essa relação, minimizando a soma dos quadrados dos erros entre os valores reais e os valores previstos.

**Regressão Logística**

- É usada quando a variável dependente é binária, ou seja, possui apenas duas classes. O modelo de regressão logística utiliza uma função logística para estimar a probabilidade de um exemplo pertencer a uma das classes.

# Tipos de regressão

## Regressão Polinomial

- É uma extensão da regressão linear, em que a relação entre as variáveis é modelada usando um polinômio.
- Isso permite capturar relações não lineares entre as variáveis e aumentar a flexibilidade do modelo.

## Regressão de Séries Temporais

- É usada para prever valores futuros com base em padrões temporais nos dados.
- A regressão de séries temporais leva em consideração a dependência temporal dos dados e utiliza métodos como médias móveis, Arima (AutoRegressive Integrated Moving Average) e modelos baseados em suavização exponencial.

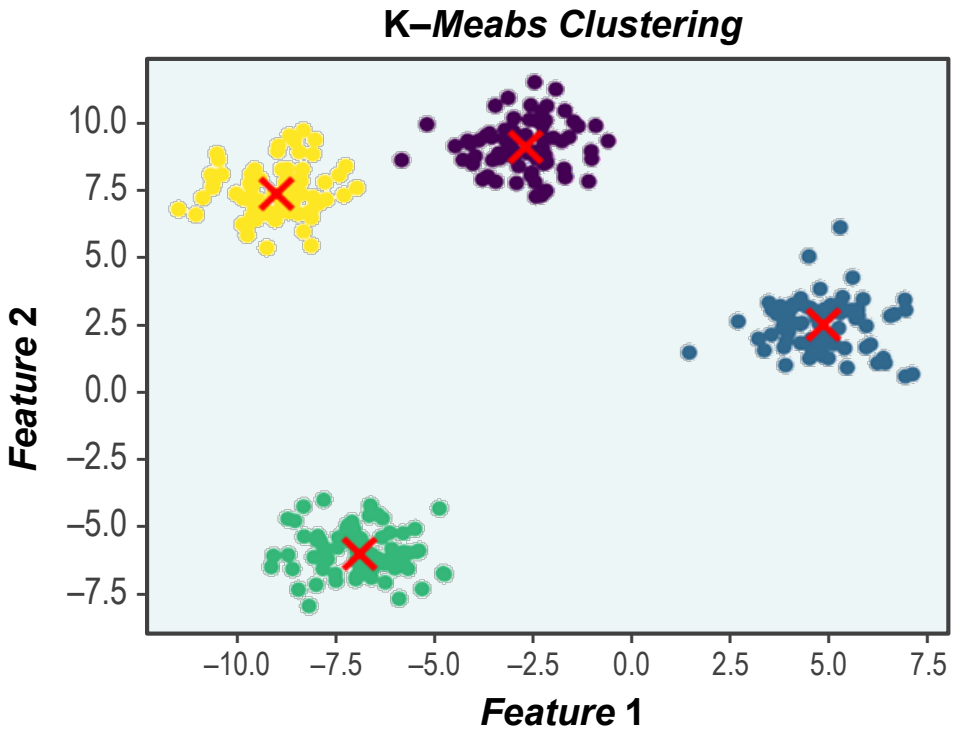
# Aprendizado não supervisionado

- O **aprendizado não supervisionado** é uma das abordagens de aprendizado de máquina, em que o **objetivo é extrair informações úteis** a partir de **dados não rotulados**. Ele busca identificar padrões e estruturas nos dados sem uma definição prévia das classes.

- O sucesso do aprendizado não supervisionado **depende de uma escolha cuidadosa das técnicas e algoritmos utilizados**, bem como da **qualidade dos dados de entrada**.

Fonte: https://br.freepik.com/fotos-gratis/aluno-em-sala-de-aula-olhando-para-o-curso_12977081.htm#query=e-learning&position=0&from_view=search&track=sph

# Aprendizado não supervisionado – Agrupamento




Fonte: livro-texto.

- O agrupamento, também conhecido como clustering, busca identificar grupos ou clusters de objetos similares em um conjunto de dados.
- O objetivo do agrupamento é encontrar estruturas e padrões nos dados sem a necessidade de rótulos ou categorias predefinidas.
- O processo de agrupamento envolve a divisão dos dados em grupos de tal forma que objetos dentro do mesmo grupo sejam mais semelhantes entre si do que com objetos de outros grupos.
- A semelhança é geralmente medida com base nas características ou atributos dos objetos.

# Aprendizado não supervisionado – Agrupamento (Algoritmos)

Alguns dos algoritmos de agrupamento mais comuns incluem:

**K-Means**

- É um algoritmo de particionamento que divide os dados em K grupos, em que K é um valor predefinido.
- Ele inicializa os centroides dos grupos de forma aleatória e, em seguida, itera alternando entre atribuir objetos ao grupo mais próximo e atualizar os centroides com base nos objetos atribuídos.

**Hierárquico**

- É uma abordagem que constrói uma estrutura hierárquica de clusters.
- Existem dois tipos principais: **aglomerativo**, em que cada objeto começa como um cluster e os clusters são combinados de forma iterativa; e **divisivo**, em que todos os objetos começam em um único cluster e são divididos em subclusters.

# Aprendizado não supervisionado – Agrupamento (Algoritmos)

## DBSCAN

- É um algoritmo baseado em densidade que agrupa os objetos com base na densidade local.
- Ele identifica regiões densas de objetos conectados e atribui esses grupos como clusters, enquanto objetos isolados são considerados ruídos.

## Mean Shift

- É um algoritmo que busca iterativamente o centro de massa dos pontos em uma vizinhança definida por uma janela de busca.
- Ele move os pontos em direção aos centros de massa até atingir uma convergência, formando assim os clusters.

# Aprendizado não supervisionado – Associação

- Associação é usada para descobrir padrões interessantes ou relações entre itens em um conjunto de dados.

- A técnica de associação envolve encontrar conjuntos de itens que ocorrem juntos com frequência em um conjunto de dados. Esses conjuntos são chamados de “itemsets frequentes”.

# Aprendizado não supervisionado – Sumarização

- A sumarização é uma técnica que visa resumir informações em um conjunto de dados de forma concisa, mas informativa.

- Ela é usada para extrair as principais ideias, características ou padrões de um conjunto de dados extenso, reduzindo-o para um resumo mais compacto e fácil de entender.

# Aprendizado não supervisionado – Sumarização

Existem dois tipos principais de sumarização:

**Sumarização manual**

- É feita por seres humanos, que leem, analisam e selecionam as informações mais importantes do conjunto de dados para criar um resumo.
- Isso é comum em áreas como jornalismo, em que os profissionais resumem e sintetizam informações de várias fontes para criar notícias ou artigos resumidos.

**Sumarização automática**

- Realizada por algoritmos e técnicas de processamento de linguagem natural (PLN) que analisam o texto ou os dados brutos e extraem informações relevantes para gerar o resumo.

Existem várias abordagens para a sumarização automática, dentre elas as que mais se destacam são:

# Aprendizado não supervisionado – Sumarização

**Sumarização extrativa:**

- Nessa abordagem, as frases ou trechos mais importantes do texto são identificados e selecionados para formar o resumo.
- Geralmente, são considerados critérios como relevância, importância e coerência para determinar quais partes do texto devem ser incluídas no resumo.

**Sumarização abstrativa:**

- Nessa abordagem, o sistema de sumarização gera frases sinteticamente que capturam o significado do texto original, em vez de simplesmente extrair frases do texto original.
- Isso envolve a compreensão do texto, a interpretação do significado e a geração de frases com base nesse entendimento.

# Interatividade

Ao falarmos de modelos descritivos de aprendizagem não supervisionada, quais das técnicas abaixo **não** são recomendadas?

a) Classificação e Regressão.
b) Agrupamento e Associação.
c) Agrupamento e Sumarização.
d) Associação e Sumarização.
e) A fusão das técnicas das alternativas b e c.

# Resposta

Ao falarmos de modelos descritivos de aprendizagem não supervisionada, quais das técnicas abaixo **não** são recomendadas?

a) Classificação e Regressão.
b) Agrupamento e Associação.
c) Agrupamento e Sumarização.
d) Associação e Sumarização.
e) A fusão das técnicas das alternativas b e c.

# Mineração de Dados

- A mineração de dados é uma área de estudo que visa extrair informações valiosas e significativas a partir de grandes conjuntos de dados.

- É um processo iterativo e interdisciplinar que envolve a aplicação de técnicas e algoritmos para descobrir padrões, tendências, relações e conhecimentos ocultos nos dados.


Fonte: https://br.freepik.com/vetores-gratis/ilustracao-do-conceito-de-processamento-de-dados_12219361.htm#query=minera%C3%A7%C3%A3o%20de%20dados&position=0&from_view=search&track=ais

# Etapas da mineração de dados

**Seleção de Dados**

- Envolve a identificação dos dados relevantes para o problema em questão. Essa etapa inclui a definição dos critérios de seleção e a obtenção dos dados necessários.

**Pré-processamento**

- É a fase em que os dados brutos são limpos, organizados e preparados para análise. Isso pode envolver a remoção de dados ausentes, correção de erros, normalização e transformação dos dados.

**Transformação**

- Nesta etapa, os dados são convertidos em uma forma adequada para análise. Isso pode incluir a redução de dimensionalidade, a extração de características relevantes e a aplicação de técnicas estatísticas ou algoritmos de processamento de dados.

# Etapas da mineração de dados

## Mineração de Dados

- É a fase central do processo, em que são aplicados algoritmos e técnicas de mineração de dados para descobrir padrões, relações e conhecimentos nos dados. Isso pode envolver a aplicação de técnicas de aprendizado de máquina, análise estatística, visualização de dados e outras abordagens.

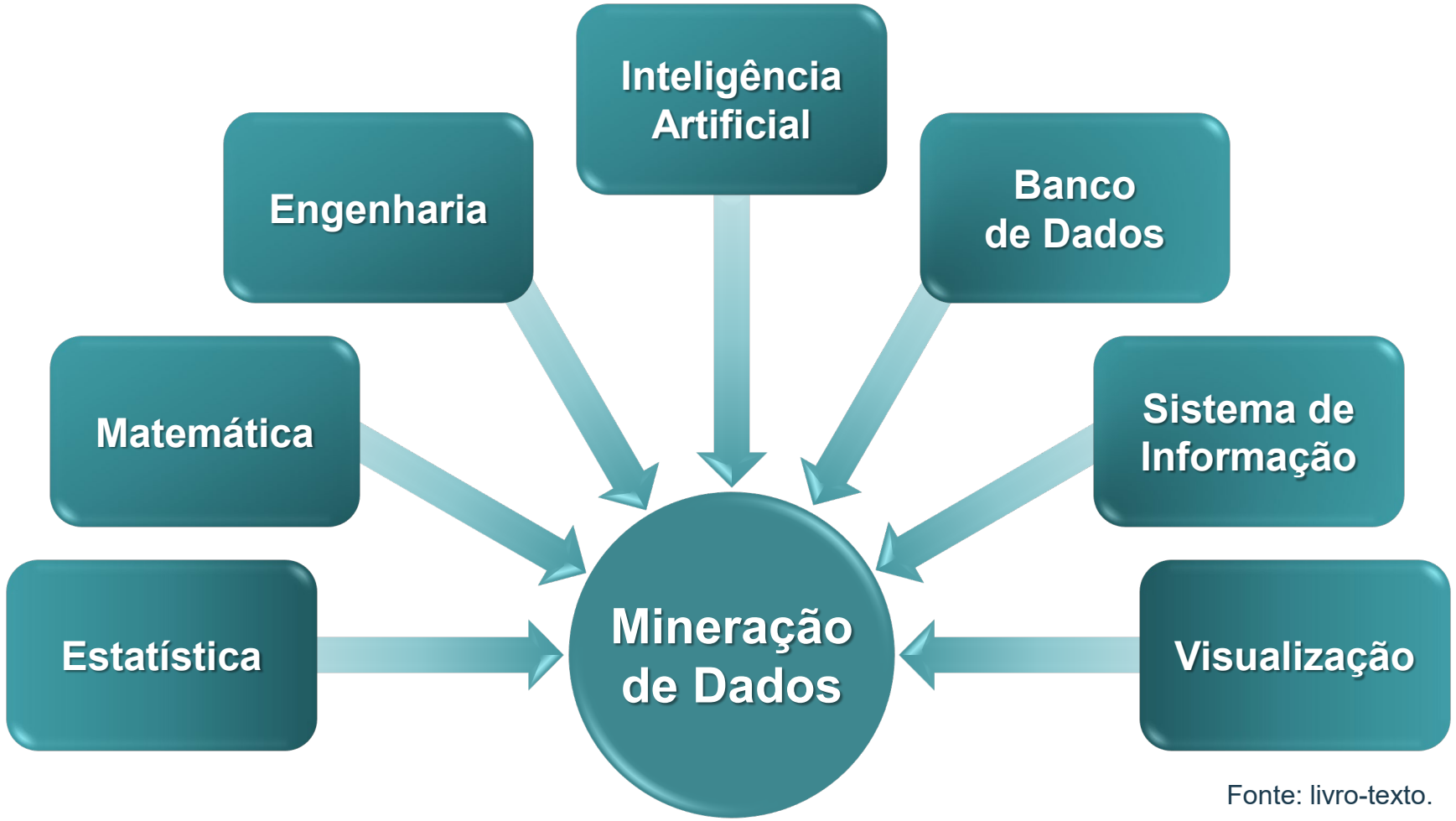
## Avaliação e Interpretação

- Os resultados da mineração de dados são avaliados quanto à sua relevância, qualidade e utilidade. Os padrões e conhecimentos descobertos são interpretados para extrair informações significativas e compreender seu impacto no problema em questão.

## Utilização do Conhecimento

- Os resultados e insights obtidos são utilizados para tomar decisões informadas, desenvolver estratégias, resolver problemas e gerar valor para a organização ou área de estudo.

# Onde aplicamos a mineração de dados




Fonte: livro-texto.

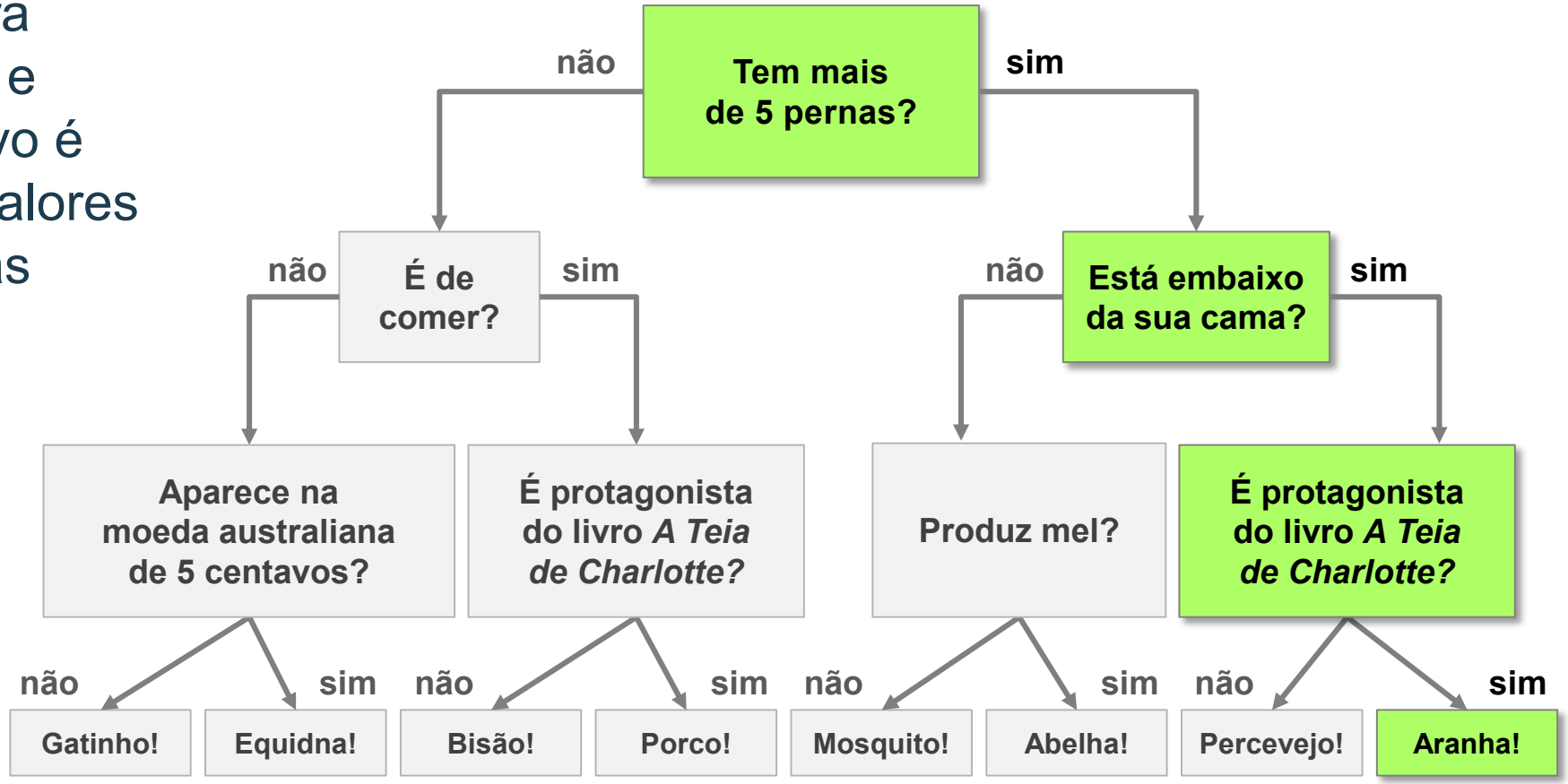
# Modelos de Machine Learning usados em Mineração de Dados

- Os modelos de aprendizado de máquina são representações matemáticas ou estatísticas que capturam os padrões e relações nos dados.
- Existem diversos tipos de modelos, cada um com suas características e aplicações específicas. A escolha do modelo correto depende do problema em questão, dos dados disponíveis e dos objetivos do projeto.
- A construção de um modelo envolve também o treinamento adequado, a avaliação de desempenho e a consideração de outros fatores importantes para o sucesso do projeto.
    - Regressão Linear
    - Regressão Logística
    - Árvores de Decisão
    - Random Forest
    - Redes Neurais Artificiais
    - Máquinas de Vetores de Suporte (SVM)
    - Naive Bayes
    - Algoritmos de Agrupamento (Clustering)

# Árvore de Decisão

- A árvore de decisão é um modelo de aprendizado de máquina que representa uma estrutura hierárquica de decisões e suas possíveis consequências.

- Essa técnica é utilizada para problemas de classificação e regressão, em que o objetivo é tomar decisões ou prever valores com base em características ou atributos dos dados.




Fonte: livro-texto.

# Naive Bayes

- O Naive Bayes é um modelo de aprendizado de máquina baseado no Teorema de Bayes e na suposição de independência condicional entre os recursos.

- Ele é comumente usado para problemas de classificação, especialmente em tarefas de processamento de linguagem natural, como análise de sentimentos, detecção de spam e categorização de documentos.

- O modelo Naive Bayes é chamado “ingênuo” porque assume que todas as características são independentes entre si, ou seja, não há correlação entre elas.

# Naive Bayes

- Descreve a probabilidade de um evento ocorrer dado o conhecimento prévio sobre o evento.

No caso do Naive Bayes, a probabilidade de uma classe dadas as características é calculada usando a seguinte fórmula:

$$P(Classe \mid Características) = P(Classe) * P(Características \mid Classe) / (Características)$$

**Em que:**

- P (Classe | Características) é a probabilidade da classe dadas as características observadas.
- P (Classe) é a probabilidade da classe ocorrer independentemente das características.
- P (Características | Classe) é a probabilidade das características ocorrerem dada a classe.
- P (Características) é a probabilidade das características ocorrerem independentemente da classe.

# K-Vizinhos mais próximos (KNN)

- O algoritmo dos k-vizinhos mais próximos, conhecido como KNN (do inglês K-Nearest Neighbors), é um método de aprendizado de máquina utilizado para classificação e regressão.

- Ele é baseado no princípio de que amostras com características semelhantes tendem a ter rótulos ou valores de saída semelhantes.

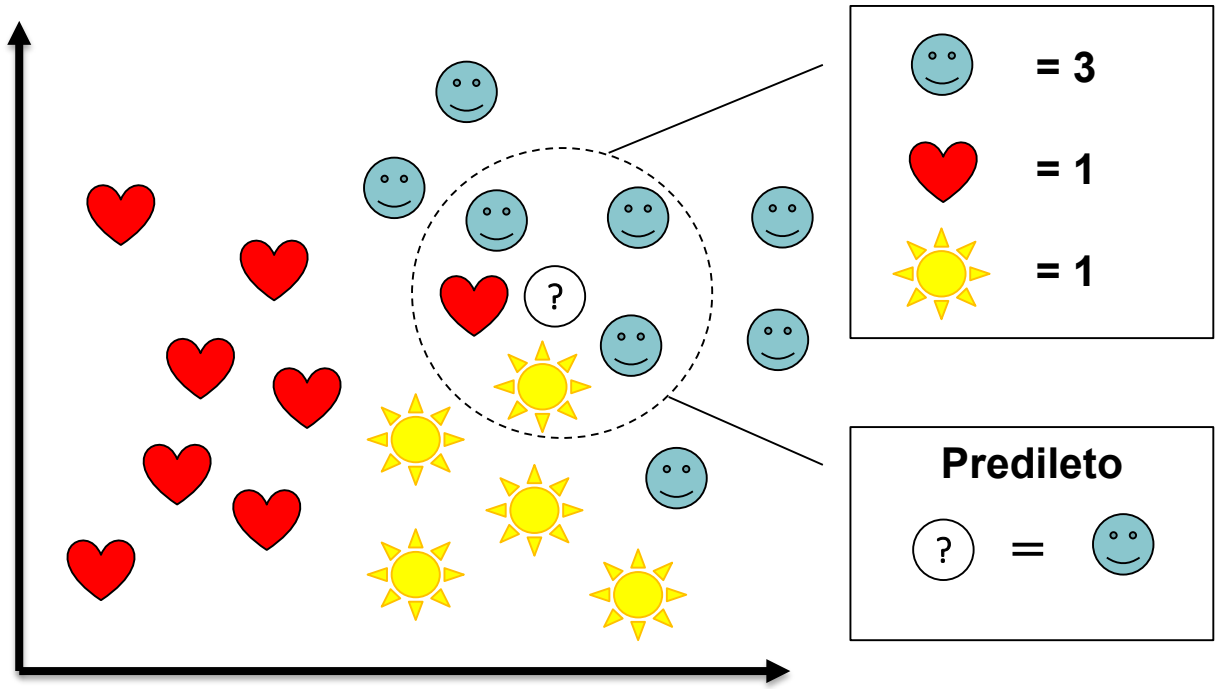
- No KNN, o objetivo é classificar uma nova amostra ou prever seu valor de saída com base nas informações dos k-vizinhos mais próximos presentes no conjunto de treinamento.

    - A distância entre as amostras é geralmente medida usando métricas como a distância Euclidiana ou a distância de Manhattan.

# K-Vizinhos mais próximos (KNN)

- O k-NN é um algoritmo simples de aprendizado de máquina baseado em instância que faz previsões com base nos vizinhos mais próximos no espaço de características.

- A eficácia do k-NN pode ser influenciada pelo valor de k escolhido e pela natureza do conjunto de dados. Experimente diferentes valores de k para ver como isso afeta a precisão do classificador.

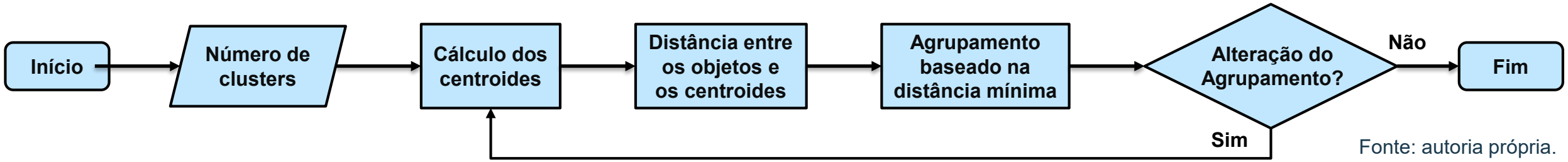



Fonte: autoria própria.

# K-Means

- O K-Means (K-Médias) é um método de aprendizado de máquina não supervisionado utilizado para realizar agrupamento de dados.
- Ele é utilizado em problemas de mineração de dados e análise exploratória, em que o objetivo é encontrar grupos ou clusters de amostras que sejam similares entre si.

**Fluxograma do algoritmo K-Means**




Fonte: autoria própria.

- O objetivo do K-Médias é minimizar a soma dos quadrados das distâncias entre as amostras e seus centroides correspondentes.
- Essa métrica é chamada de “inércia” e representa a coesão dentro de cada cluster.
- Quanto menor a inércia, mais compactos e bem-definidos são os clusters.

# Interatividade

Analise as descrições das etapas da mineração de dados e indique qual das alternativas nomeia de forma correta cada uma das etapas.

**I.** Nesta etapa os padrões e conhecimentos descobertos são interpretados para extrair informações significativas e compreender seu impacto no problema em questão.

**II.** Nesta etapa os dados são convertidos em uma forma adequada para análise.

**III.** É a etapa em que os dados brutos são limpos, organizados e preparados para análise.

a) **I** – Avaliação e Interpretação, **II –** Transformação e **III** – Pré-processamento.

b) **I –** Seleção de Dados, **II** – Pré-processamento e **III** – Mineração de Dados.

c) **I** – Mineração de Dados, **II –** Transformação e **III** – Avaliação do Conhecimento.

d) **I** – Utilização do Conhecimento, **II –** Utilização do Conhecimento e **III** – Seleção de Dados.

e) **I** – Pré-processamento, **II** – Mineração de Dados e **III –** Transformação.

# Resposta

Analise as descrições das etapas da mineração de dados e indique qual das alternativas nomeia de forma correta cada uma das etapas.

**I.** Nesta etapa os padrões e conhecimentos descobertos são interpretados para extrair informações significativas e compreender seu impacto no problema em questão.

**II.** Nesta etapa os dados são convertidos em uma forma adequada para análise.

**III.** É a etapa em que os dados brutos são limpos, organizados e preparados para análise.

a) **I** – Avaliação e Interpretação, **II –** Transformação e **III** – Pré-processamento.

b) **I –** Seleção de Dados, **II** – Pré-processamento e **III** – Mineração de Dados.

c) **I** – Mineração de Dados, **II –** Transformação e **III** – Avaliação do Conhecimento.

d) **I** – Utilização do Conhecimento, **II –** Utilização do Conhecimento e **III** – Seleção de Dados.

e) **I** – Pré-processamento, **II** – Mineração de Dados e **III –** Transformação.

# Referências


- CASTRO, L. N.; FERRARI, D. G. *Introdução à mineração de dados*: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.
- CHAPMAN, P. *et al*. *CRISP-DM 1.0*: step-by-step data mining guide. 2000. Disponível em: https://tinyurl.com/3mn8j4xk. Acesso em: 29 ago. 2023.
- FACELI, K.; LORENA, A. C.; GAMA, J.; DE CARVALHO, A. C. P. L. F. *Inteligência artificial:* uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.
- FILATRO, A. C. *Data science na educação* : presencial, a distância e corporativa. São Paulo: Saraiva, 2020.
- MITCHELL, T. M. *Machine Learning*. Portland: McGraw-Hill, 1997.

# Referências


- NORVIG, P. *Inteligência artificial*. Rio de Janeiro: Grupo GEN, 2013.
- PIATETSKY, G. CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets*, October, 2014. Disponível em: https://tinyurl.com/8b4tevy7. Acesso em: 29 ago. 2023.
- TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introduction to Data Mining*. Boston: Pearson Addison-Wesley, 2006. Disponível em: https://tinyurl.com/mr45rp3p. Acesso em: 29 ago. 2023.

# ATÉ A PRÓXIMA!