Frameworks de Big Data

UNIDADE 06

Explorando os fundamentos da aprendizagem de máquina

Vamos começar esta semana explorando os fundamentos da aprendizagem de máquina. Durante nossas aulas, mergulharemos nas bases essenciais dessa área dinâmica da ciência da computação, desde os princípios teóricos até as aplicações práticas que moldam nosso mundo digital. Preparem-se para uma imersão profunda em algoritmos, modelos e técnicas que capacitam as máquinas a aprender e a tomar decisões inteligentes com base em

dados. Esperamos que essa jornada seja repleta de insights e conquistas enquanto exploramos os fundamentos que impulsionam a aprendizagem de máquina. Vamos começá-la juntos!

Você já ouviu falar sobre aprendizado de máquina? Já pensou em como a aprendizagem de máquina e o *big data* estão interligados? Nesta semana, vamos explorar essa relação e entender como essas tecnologias se complementam e descobrir como podemos utilizar essas ferramentas para extrair *insights* valiosos de conjuntos massivos de dados.

REFLEXÃO

Imagine que você é parte de uma equipe de análise de uma empresa de comércio eletrônico. Essa empresa possui milhões de clientes e registra uma vasta quantidade de dados de interações, como histórico de compras, comportamento de navegação, e *feedback* dos clientes. Agora, a empresa deseja entender melhor o comportamento de compra dos clientes para otimizar suas estratégias de *marketing* e recomendação de produtos.

Como você poderia resolver esse problema? É aqui que entra a aprendizagem de máquina. Usando algoritmos de aprendizagem de máquina, podemos analisar esses dados para identificar padrões e prever comportamentos futuros dos clientes. Por exemplo, podemos desenvolver um modelo que preveja quais produtos é mais provável um cliente específico comprar com base em seu histórico de compras e navegação no *site*.

Entendendo os fundamentos da aprendizagem de máquina

Definição de aprendizagem de máquina e suas aplicações

O aprendizado de máquina (ML), também conhecido como *machine learning*, é um ramo da inteligência artificial (IA) que permite aos computadores aprenderem e melhorarem a partir de dados, sem que sejam explicitamente programados para isso. Uma das definições de ML é:

“A capacidade de melhorar o desempenho na realização de alguma tarefa por meio da experiência”. (Mitchell, 1997)

Portanto, o aprendizado de máquina fundamenta-se na concepção de que os computadores podem aprender a reconhecer padrões nos dados e empregar esses padrões para efetuar previsões ou tomar decisões.

O quadro 1 a seguir condensa os tipos de aprendizagem e as características descritas por NORVIG, proporcionando uma síntese essencial.

Quadro 1: Tipos de aprendizagem de máquina

TIPO	**DESCRIÇÃO**
Supervisionado	O computador é treinado com um conjunto de dados que contém exemplos de entrada e saída esperados. O computador aprende a associar as entradas às saídas esperadas.
Não supervisionado	O computador é treinado com um conjunto de dados que não contém exemplos de entrada e saída esperados. O computador aprende a identificar padrões nos dados sem a ajuda de exemplos.

Fonte: Norvig, 2022.

Há ainda o **aprendizado por reforço**. Nele, o agente aprende a partir de uma série de reforços — recompensas ou punições. Por exemplo, a falta de gorjeta ao final de uma corrida dá ao agente do táxi a indicação de que algo saiu errado, ou os dois pontos de vitória, no final de um jogo de xadrez, informam ao agente que fez a coisa certa.

Cabe ao agente decidir qual das ações anteriores ao reforço foram as maiores responsáveis por isso. Porém, para esta aula, iremos nos concentrar apenas nos tipos de aprendizagem **supervisionado e não supervisionado.**

Na figura 1 abaixo, apresento uma representação gráfica do processo de aprendizagem de máquina descrito no quadro 1.

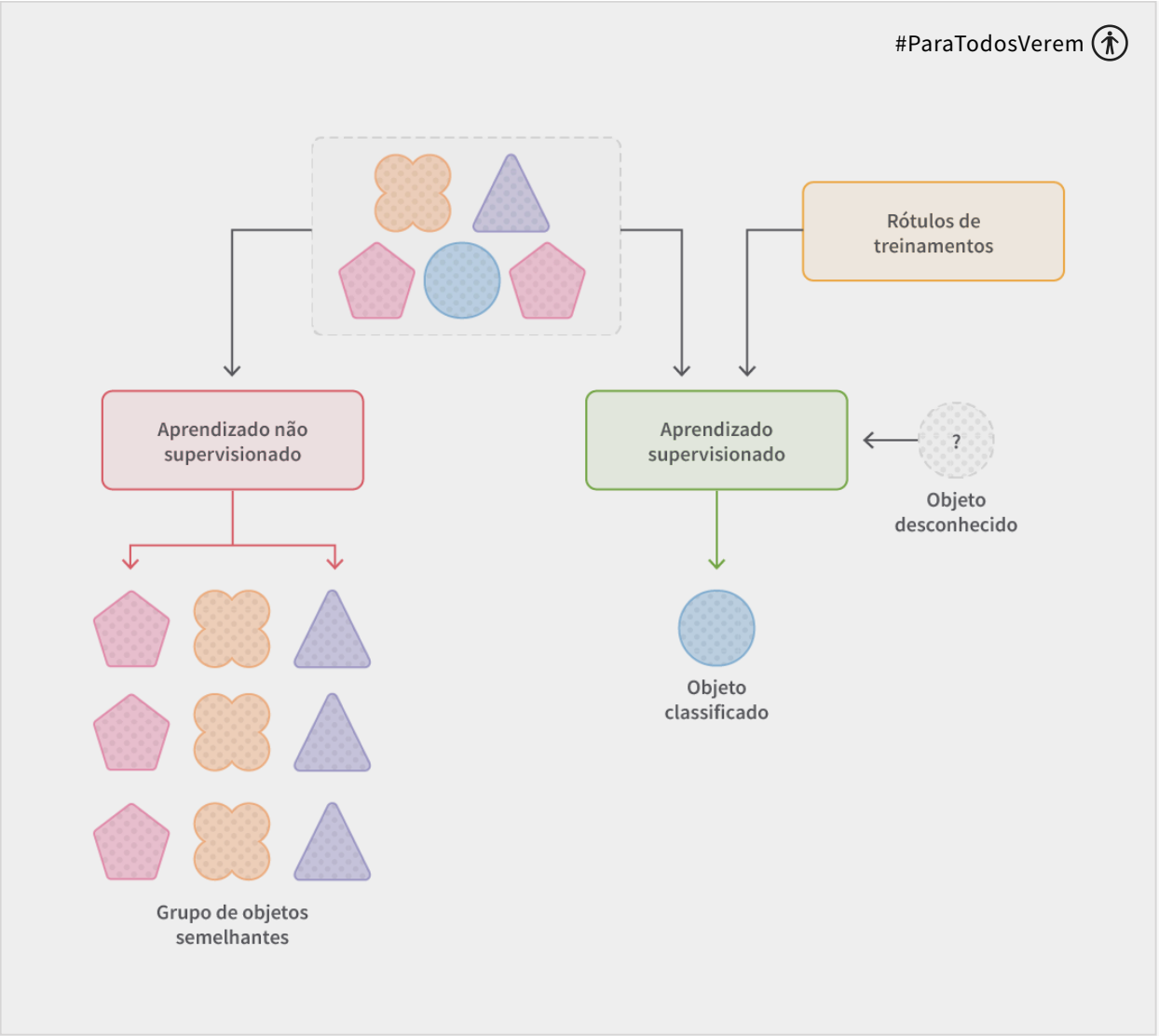

Figura 1: Fluxo de aprendizagem de máquina. Fonte: Adaptado de JUGAL K. Kalita, Dhruba K. Bhattacharyya and Swarup Roy (2023).

E como o aprendizado de máquina tem sido utilizado?

REFLEXÃO

O **aprendizado de máquina** tem uma ampla gama de aplicações em diversas áreas, incluindo:

- **Recomendação de produtos:** o ML é usado para recomendar produtos aos clientes com base em seus dados de compra.
- **Filtragem de *spam*:** o ML é usado para identificar e remover *spam* de *e-mails*.
- **Detecção de fraudes:** o ML é usado para identificar fraudes financeiras.
- **Diagnóstico médico:** o ML é usado para ajudar os médicos a diagnosticarem doenças.

O aprendizado de máquina é uma área em **rápida evolução** que tem o potencial de transformar a forma como interagimos com o mundo ao nosso redor. À medida que os dados continuam a se tornar mais abundantes e acessíveis, o aprendizado de máquina se torna cada vez mais importante em nossas vidas.

Principais algoritmos de aprendizagem de máquina

No quadro 2, apresentamos um guia abrangente dos principais algoritmos de aprendizagem de máquina. Cada algoritmo possui suas nuances e vantagens, tornando-os ferramentas valiosas para a resolução de diversos desafios.

Quadro 2: Principais algoritmos de aprendizagem de máquina

Algoritmo	**Descrição**
Regressão linear	Um algoritmo supervisionado que aprende a mapear um conjunto de variáveis de entrada (preditores) para uma variável de saída contínua. É usado para prever valores numéricos, como preço de uma casa ou temperatura média mensal.
Árvores de decisão	Um algoritmo supervisionado que constrói uma árvore lógica para tomar decisões com base em dados. É fácil de interpretar e pode ser usado para tarefas de classificação e regressão.
Random Forests	Um algoritmo de aprendizado de máquina de conjunto que combina várias árvores de decisão para melhorar a precisão e reduzir o *overfitting*. É robusto a *outliers* e pode ser usado para tarefas de classificação e regressão.
K-Means clustering	Um algoritmo de aprendizado não supervisionado que agrupa dados em um número pré-definido de *clusters*. É usado para identificar padrões em dados não rotulados.

Redes neurais artificiais	Um algoritmo de aprendizado de máquina inspirado no cérebro humano. É capaz de aprender padrões complexos em dados e pode ser usado para tarefas de classificação, regressão e processamento de linguagem natural.

Fonte: O autor (2024).

É importante ressaltar que cada algoritmo possui suas nuances e vantagens, e a escolha ideal dependerá das características específicas do problema a ser resolvido. A experimentação e o aprendizado contínuo são essenciais para dominar essa área em constante evolução.

Ao aprofundar seus conhecimentos nos recursos mencionados nas referências, você estará apto a navegar pelos desafios e oportunidades que a aprendizagem de máquina apresenta. Explore diferentes algoritmos, domine as técnicas de pré-processamento e construa modelos que gerem *insights* valiosos e impulsionem o progresso.

Preparação e pré-processamento de dados

A preparação e o pré-processamento de dados são etapas essenciais no processo de análise de dados. ***Essas etapas visam garantir a qualidade dos dados e prepará-los para a análise.**

+ Importância da qualidade dos dados

A **qualidade dos dados** é fundamental para o sucesso de qualquer análise. Dados de baixa qualidade podem levar a resultados imprecisos ou inválidos. Por isso, é importante dedicar tempo e atenção à preparação e ao pré-processamento de dados.

+ Exploração de dados e identificação de padrões

A **exploração de dados** é uma etapa importante do processo de preparação de dados. Essa etapa consiste em analisar os dados para identificar padrões e tendências. A exploração de dados pode ajudar a identificar problemas nos dados, como dados ausentes ou

valores discrepantes.

+ Tratamento de dados ausentes e normalização

Os **dados ausentes** são uma ocorrência comum em conjuntos de dados. Esses dados podem ser causados por erros de entrada, falha de *hardware* ou outros fatores. O tratamento de dados ausentes é uma tarefa importante para garantir a qualidade dos dados.

No quadro 3 abaixo, temos os tipos de técnicas de preparação e pré-processamento de dados mais comuns no processo de aprendizagem de máquina.

Quadro 3: Principais técnicas de preparação e pré-processamento de dados

Técnica	Descrição	Exemplos
Exploração de dados	Análise dos dados para identificar padrões e tendências.	Análise de distribuição de dados, identificação de *outliers*, análise de correlação.
Tratamento de dados ausentes	Preenchimento de dados ausentes com valores estimados ou exclusão de dados ausentes.	Preenchimento por média, preenchimento por regressão, exclusão de registros.
Normalização	Transformação dos dados para que tenham uma distribuição uniforme.	Normalização *z-score*, normalização mín.-máx.
Discretização	Refere-se ao processo de transformar variáveis contínuas em categorias ou intervalos discretos.	Discretização de dados de temperatura, em que os valores contínuos em graus Celsius podem ser agrupados em categorias como "frio", "morno" e "quente".

Fonte: Norvig, 2022.

Por exemplo, segundo Faceli, para iniciar o processo de aprendizado, os dados brutos são coletados e submetidos à etapa de pré-processamento, que é tipicamente dividida em subconjuntos, conforme vemos na figura 2.

O conjunto de treinamento é utilizado para instruir o modelo, enquanto o conjunto de validação é empregado para avaliar o desempenho do modelo em treinamento e refiná-lo. Se o principal objetivo é prever a classe de uma instância com a máxima precisão possível, especialmente quando aplicado a uma instância de teste sem rótulo, o método a escolher seria o do **aprendizado supervisionado.**

Nesse caso, durante a fase de treinamento, o conjunto de validação é essencial para avaliar o desempenho e ajustar os parâmetros do modelo. Já o conjunto de teste entra em cena após o treinamento completo do modelo, permitindo avaliar seu desempenho em dados inéditos.

Uma vez que atinge um desempenho satisfatório, o modelo aprendido está pronto para prever a categoria de objetos não previamente encontrados. A flexibilidade desse processo permite ajustes finos ao treinamento do modelo, seja ajustando parâmetros de algoritmos de aprendizado de máquina ou reiniciando o processo com diferentes combinações de etapas.

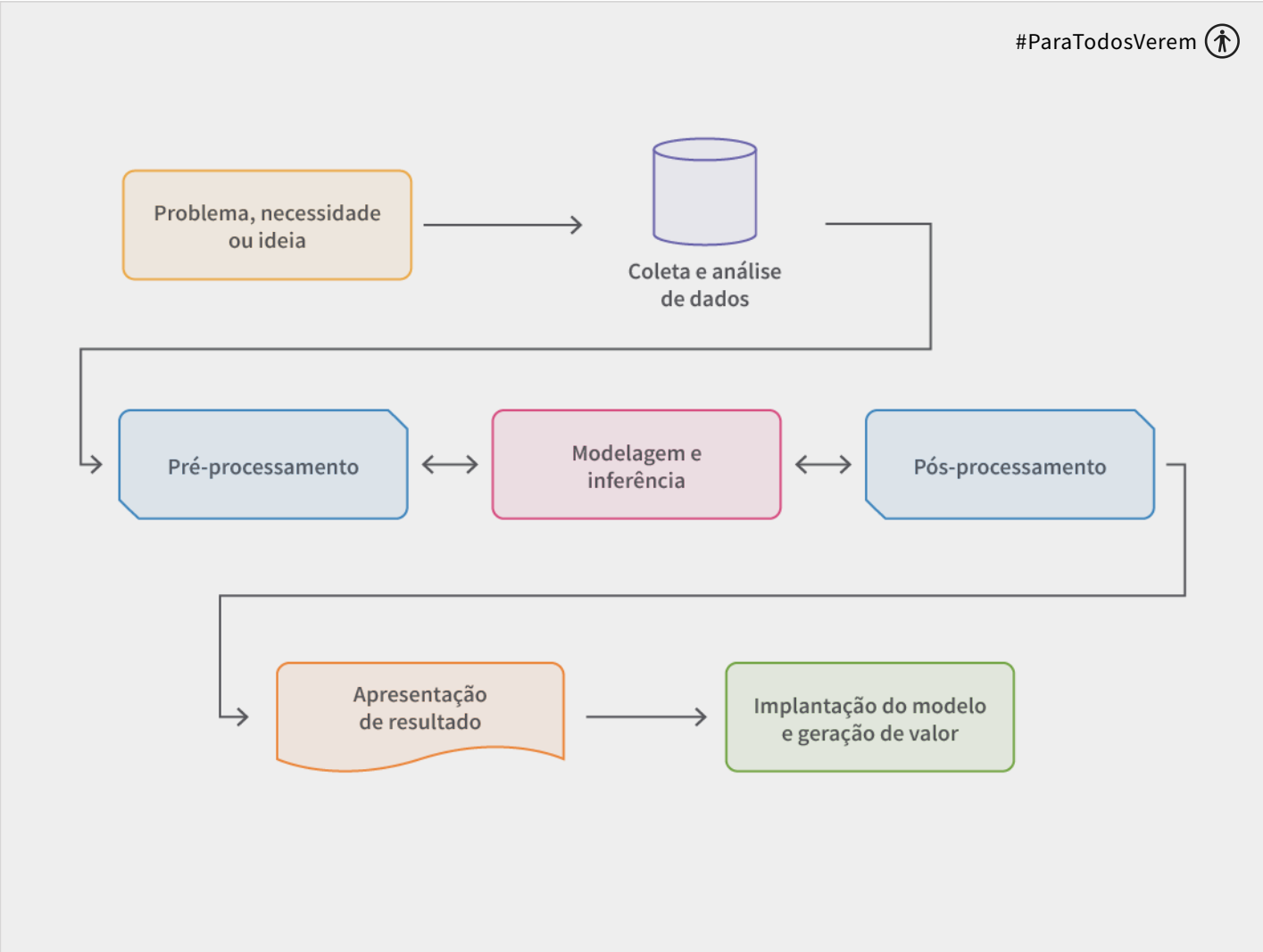

Figura 2: Fluxo modelo supervisionado de aprendizagem. Fonte: Adaptado de Escovedo & Koshiyama, 2020.

Big data e aprendizagem de máquina

No vasto e dinâmico ecossistema da tecnologia, duas forças têm se destacado cada vez mais: ***big data* e aprendizagem de máquina.** São como os pilares de uma ponte que conecta o presente ao futuro, impulsionando inovações e transformações em diversas esferas da sociedade.

O *big data* é o oceano de informações que banha nossas vidas diárias. É o fluxo constante de dados gerados por cada interação digital, seja ao navegar na internet, interagir em redes sociais, realizar transações financeiras ou utilizar dispositivos inteligentes. Essa imensa quantidade de dados não estruturados representa um tesouro de *insights* e oportunidades, mas sua magnitude desafia as capacidades humanas tradicionais de análise.

É aqui que entra a aprendizagem de máquina, uma poderosa ferramenta que permite extrair significado e valor desse dilúvio de dados. Baseada em algoritmos complexos e modelos matemáticos, a aprendizagem de máquina capacita os sistemas a identificarem padrões, fazerem previsões e tomarem decisões com base nos dados disponíveis. Seja reconhecendo padrões de consumo para personalizar recomendações de produtos ou detectando anomalias em sistemas de segurança, a aprendizagem de máquina abre portas para uma infinidade de aplicações práticas. Na figura 3, temos um *framework* de como as aplicações de *big data* e aprendizagem de máquina atuam.

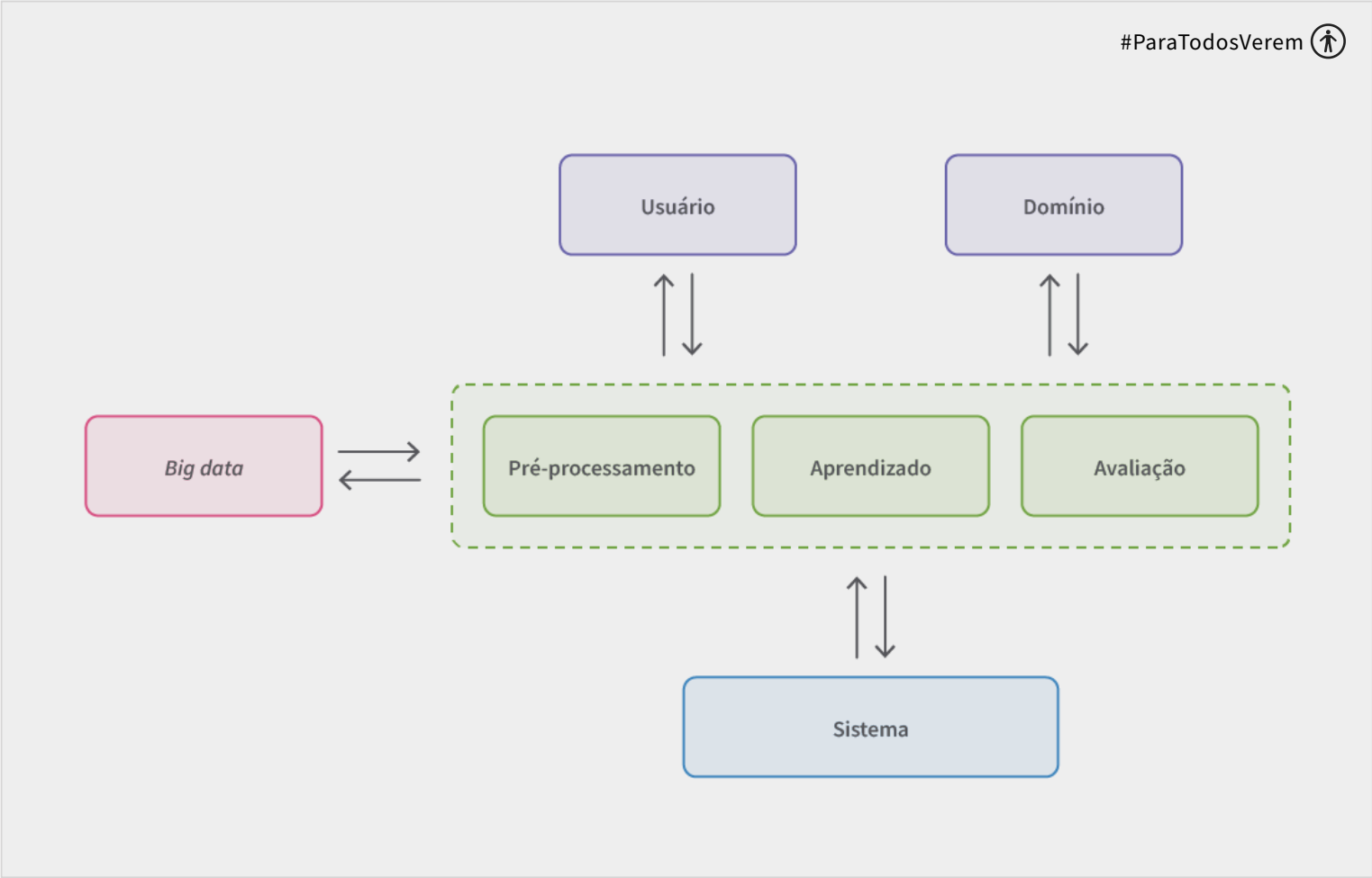

Figura 3: Abordagem de big data em conjunto com aprendizado de máquina. Fonte: Adaptado de Zhou, L., Pan, S., Wang, J., & Vasilakos (2017).

Juntas, *big data* e aprendizagem de máquina formam uma dupla imbatível, impulsionando avanços em áreas tão diversas quanto medicina, finanças, manufatura e *marketing*. Na medicina, por exemplo, a análise de grandes conjuntos de dados de pacientes pode levar a diagnósticos mais precisos e personalizados, enquanto na indústria automotiva, algoritmos de aprendizagem de máquina estão revolucionando a condução autônoma, tornando-a mais segura e eficiente.

No entanto, com grande poder vem grande responsabilidade. O uso ético e transparente do *big data* e da aprendizagem de máquina é crucial para evitar abusos e garantir que essas tecnologias sirvam ao bem comum. Questões de privacidade, viés algorítmico e equidade devem ser cuidadosamente consideradas em todas as fases do desenvolvimento e implementação dessas soluções.

À medida que navegamos pelo vasto universo do *big data* e da aprendizagem de máquina, devemos lembrar que essas são ferramentas poderosas, mas são os seres humanos que moldam seu uso e impacto. Com sabedoria e discernimento, podemos aproveitar todo o potencial dessas tecnologias para construir um mundo mais inteligente, justo e sustentável.

Conclusão

Vamos voltar às questões apresentadas no começo deste material.

A aplicação de técnicas de aprendizado de máquina representa uma revolução na forma como os usuários interagem com plataformas de comércio eletrônico, especialmente no contexto de sugestões de produtos. Ao empregar algoritmos inteligentes, essas plataformas têm a capacidade de analisar padrões complexos no comportamento do usuário. Aspectos como preferências de gênero, histórico de visualização, feedback de avaliações e até mesmo o momento do dia em que o usuário assiste a algo podem ser cruciais para a personalização eficaz das recomendações.

A compreensão desses detalhes permite aos algoritmos oferecerem sugestões altamente personalizadas, antecipando as preferências individuais dos usuários. No entanto, essa sofisticação não está isenta de desafios éticos. A coleta e análise de dados comportamentais podem levantar preocupações sobre privacidade e segurança. O equilíbrio delicado entre personalização eficaz e respeito à privacidade do usuário é um dos desafios éticos primordiais. Garantir a transparência nos processos e oferecer opções claras de controle aos usuários tornam-se essenciais para mitigar essas preocupações.

A capacidade de extrair informações valiosas a partir de dados e aplicar algoritmos de aprendizagem para resolver problemas do mundo real é uma habilidade poderosa que continuará a desempenhar um papel crucial em diversas áreas. A aprendizagem de máquina não é apenas uma disciplina emocionante, mas também uma ferramenta valiosa para impulsionar a inovação em muitos setores.

Vamos ver mais sobre isso....

Referências Bibliográficas

ESCOVEDO, T. KOSHIYAMA, A. **Introdução à data science** — algoritmos de machine learning e métodos de análise. São Paulo, Ed. Casa do Código, 2020.

FACELI, K.; LORENA, A. C.; GAMA, J. *et al.* **Inteligência artificial** – uma abordagem de aprendizado de máquina. Grupo GEN, 2021.

KALITA, K. J.; K.; BHATTACHARYYA, D. K.; ROY, S. **Fundamentals of data science**: theory and practice presents basic and advanced concepts in data science along with real-life applications. [*S.l.*]: Elsevier, 2023

LENZ, M. L.; NEUMANN, F. B.; SANTARELLI, R. *et al*. **Fundamentos de aprendizagem de máquina**. Porto Alegre: Grupo A, 2020.

MITCHELL, T. M. **Aprendizagem de máquina**. 1. ed. Rio de Janeiro: McGraw-Hill, 1997.

RUSSELL, S. J.; NORVIG, P. **Inteligência artificial**: uma abordagem moderna. Rio de Janeiro: Grupo GEN, 2022.

ZHOU, L.; PAN, S.; WANG, J.; VASILAKOS, A. V. Machine learning on big data: opportunities and challenges. **Neurocomputing**, n. 237, p. 350–361. Disponível em: doi:10.1016/j.neucom.2017.01.0. Acesso em: 12 jun. 2024.