

PUC
CAMPINAS

PONTIFÍCIA UNIVERSIDADE CATÓLICA

PROJETO DE MACHINE LEARNING

PUC Campinas

Todos os direitos reservados.

Proibida a reprodução total ou parcial em qualquer mídia sem a autorização.

Autor

Marcelo Henrique dos Santos

Revisão técnica

Uol EdTech

Projeto gráfico

Uol EdTech

Biblioteca PUC (Dados de publicação na fonte)

Santos, Marcelo Henrique dos .

Projeto de Machine Learning / Marcelo Henrique dos Santos. 1 ed.

Belo Horizonte, 2021.

1. As fases de um projeto de aprendizado de máquina. 2. Construção de um Modelo Preditivo. 3. Visualização e apresentação dos resultados de um modelo preditivo.

I. Título.

Catalogação: Biblioteca Central Anisio Teixeira – PUC Campinas

2021 · Proibida a reprodução total ou parcial. Os infratores serão processados na forma da lei.
As imagens e ilustrações utilizadas nesta apostila foram obtidas no site <https://www.shutterstock.com>

FIGURA 01	Ilustração da dinâmica para identificação das frutas (maçã e laranja)	14
FIGURA 02	Representação do processo de coleta de dados	19
FIGURA 03	Representação do processo de preparação de dados	23
FIGURA 04	Representação do processo de desenvolvimento de software	26
FIGURA 05	Representação da utilização dos princípios da aprendizagem de máquina na área de transporte e logística	30
FIGURA 06	Representação da modelagem preditiva	39
FIGURA 07	Representação da utilização dos princípios da análise preditiva no controle de estoque	45
FIGURA 08	Representação de um proprietário de salão de beleza utilizando os recursos da aprendizagem de máquina	48
FIGURA 09	Representação da utilização de modelos preditivos na medicina	49
FIGURA 10	Representação de um conjunto de treinamento	60
FIGURA 11	Representação do processo de apresentação dos resultados de um modelo preditivo	63
FIGURA 12	Representação da utilização da aprendizagem de máquina na área médica	69
FIGURA 13	Representação do funcionamento de um <i>chatbots</i>	70
FIGURA 14	Representação do funcionamento de um sistema de <i>trading</i>	71

UNIDADE 1	AS FASES DE UM PROJETO DE APRENDIZADO DE MÁQUINA	11
	APRESENTAÇÃO	12
	1.1 As fases de um projeto de aprendizado de máquina	12
	1.1.1 Princípios da coleta de dados.	18
	1.1.2 Pré-processamento dos dados.	22
	1.1.3 Planejamento do desenvolvimento de um projeto que utilize os princípios do aprendizado de máquina	25
	1.1.3.1 Impacto das alterações de recursos e orçamento no cronograma.	26
	1.1.3.2 Desenvolvimento de projetos que utilizam os princípios da aprendizagem da máquina.	27
	BIBLIOGRAFIA COMENTADA	31
	MAPA CONCEITUAL	32
	CONCLUSÃO	33
UNIDADE 2	CONSTRUÇÃO DE UM MODELO PREDITIVO	34
	APRESENTAÇÃO	35
	2.1 Construção de um modelo preditivo	35

2.1.1 Definições sobre o modelo preditivo	36
2.1.2 Construção de um modelo preditivo	42
2.1.2.1 Aplicativos de análise preditiva e aprendizado de máquina	44
2.1.2.2 Modelos e algoritmos de análise preditiva	45
2.1.3 Validação de um modelo preditivo	48
2.1.3.1 Validando um algoritmo de aprendizado de máquina na área médica	51
2.1.3.2 Validando um algoritmo de aprendizado de máquina na área financeira	53
BIBLIOGRAFIA COMENTADA	55
MAPA CONCEITUAL	55
CONCLUSÃO	56
UNIDADE 3	
VISUALIZAÇÃO E APRESENTAÇÃO DOS RESULTADOS DE UM MODELO PREDITIVO	57
APRESENTAÇÃO	58
3.1 Visualização e apresentação dos resultados de um modelo preditivo	58
3.1.1 Visualização de um modelo preditivo	60

3.1.2 Apresentação dos resultados de um modelo preditivo	62
3.1.2.1 Práticas para a implantação do modelo preditivo	66
3.1.3 O que é possível fazer a partir do modelo preditivo? Exemplos práticos	68
BIBLIOGRAFIA COMENTADA	72
MAPA CONCEITUAL	73
CONCLUSÃO	73
REFERÊNCIAS	75



Atenção
Para saber



Atividades de
aprendizagem



Saiba mais
Onde pesquisar
Leitura complementar
Dicas



Curiosidades



Glossário



Questões



Mídias integradas



Áudios



Anotações



Citações



Exemplos



Downloads

Biodata do autor

Marcelo Henrique dos Santos

Mestre em Educação pela Universidade Interamericana (2019), com dissertação sobre a utilização dos jogos digitais no Ensino Superior. Especialista em Negócios em Mídias Digitais pelas Faculdades Metropolitanas Unidas – FMU (2018), MBA em Marketing em Vendas pela FMU (2016), especialista em Games: Produção e Programação pelo Centro Universitário Senac (2011) e bacharel em Sistemas de Informação pela Universidade de Mogi das Cruzes – UMC (2008). Atualmente é microempreendedor na área de jogos digitais e professor universitário na área de tecnologia na Universidade Paulista – UNIP e Universidade Estácio de Sá.

Tem experiência na área de Computação, com ênfase em desenvolvimento de aplicativos, jogos digitais, desenvolvimento web e TV digital, e sistemas interativos (trabalhando por muitos anos na empresa TV SBT Canal 4 de São Paulo). Além disso, possui experiência na área de Educação, com ênfase no desenvolvimento de games e hipermídia, atuando principalmente nas áreas de fundamentação e desenvolvimento de metodologias e protótipos. É autor de diversos livros técnicos na área de Computação e Jogos Digitais

Justificativa

À medida que o aprendizado de máquina se prolifera, mais profissionais buscam carreiras como engenheiros de aprendizado de máquina. Uma das melhores maneiras de começar é compreender sobre o processo de desenvolvimento e planejamento de projeto de *machine learning*. O entendimento sobre as fases de um projeto de aprendizado de máquina, a construção de um modelo preditivo e a dinâmica da visualização e apresentação dos resultados de um modelo preditivo se tornam fundamentais para o profissional conseguir desenvolver projetos que apliquem os princípios do aprendizado de máquina.

Engajamento

Aprendizado de máquina é o processo de permitir que os computadores modifiquem ou adaptem suas ações para que esse recurso possa ser preciso. Com os modelos de aprendizado de máquina apropriados, as organizações têm a capacidade de prever continuamente as mudanças nos negócios para que sejam capazes de prever o que vem a seguir.

A modelagem preditiva é a subparte da análise de dados que usa a mineração de dados e probabilidade para prever resultados. Cada modelo é construído pelo número de preditores que são altamente favoráveis para determinar decisões futuras. Uma vez que os dados são recebidos para um preditor específico, um modelo analítico é formulado. Um modelo pode aplicar uma equação linear simples ou uma estrutura neural complexa delineada pelo software em questão, também se houver dados adicionais disponíveis, o modelo analítico é revisado.

A partir desse princípio, espera-se que você consiga responder às seguintes perguntas:

- Como funciona e quais são as vantagens da utilização do pré-processamento dos dados?
- De que forma se deve realizar o planejamento do desenvolvimento de um projeto que utilize os princípios do aprendizado de máquina?
- Quais as características de um modelo preditivo?
- Quais recursos devem ser realizados na validação de um modelo preditivo?
- O que é possível fazer a partir do modelo preditivo?

Apresentação da disciplina

Olá! Começarão seus estudos sobre projeto de machine learning.

Ao longo desta jornada, será possível compreender sobre a gestão e organização de todas as fases de um projeto de aprendizado de máquina, compreendendo a coleta e o pré-processamento dos dados, a construção de um modelo preditivo, a validação desse modelo e a visualização e apresentação dos resultados.

Você verá sobre as fases de um projeto de aprendizado de máquina e os princípios da coleta de dados. Será discutido sobre o pré-processamento dos dados e o processo de planejamento do desenvolvimento de um projeto que utilize os princípios do aprendizado de máquina.

Ao longo do processo de aprendizagem, você compreenderá sobre o processo de visualização e apresentação dos resultados de um modelo preditivo e refletirá sobre o que é possível fazer a partir do modelo preditivo (com exemplos práticos).

Objetivos da disciplina

Ao final desta disciplina, esperamos que você seja capaz de:

- Descrever sobre os princípios da coleta de dados.
- Analisar sobre o pré-processamento dos dados.
- Aplicar os conceitos do aprendizado de máquina.
- Descrever o processo de construção de um modelo preditivo.
- Avaliar sobre a dinâmica da validação de um modelo preditivo.
- Aplicar os conceitos do modelo preditivo de forma prática.
- Analisar sobre o processo de visualização de um modelo preditivo.
- Descrever sobre o processo de apresentação dos resultados de um modelo preditivo.
- Aplicar os princípios do modelo preditivo.

AS FASES DE UM PROJETO DE APRENDIZADO DE MÁQUINA

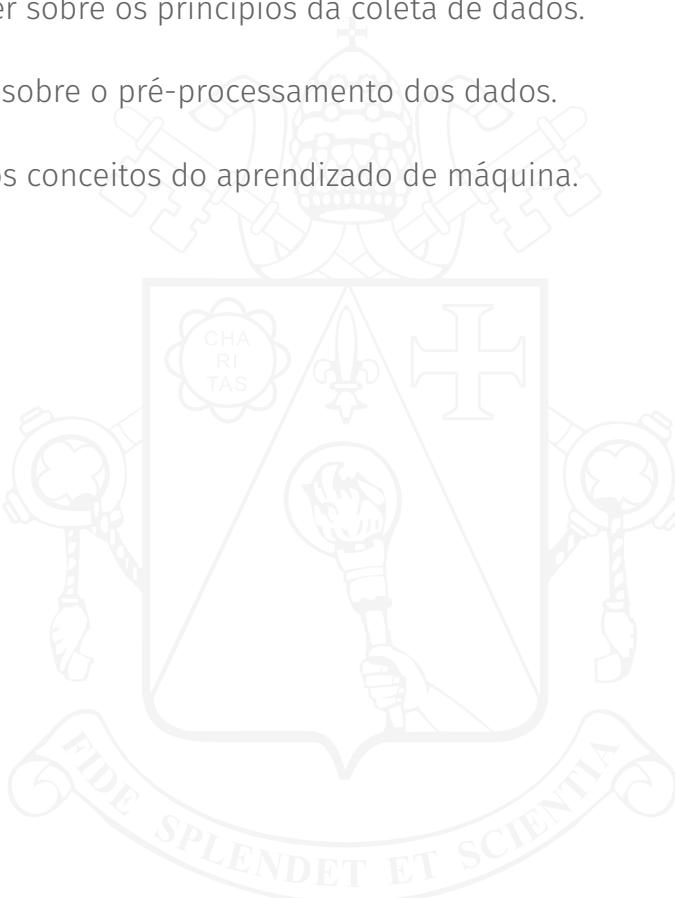
OBJETIVOS

Ao final desta unidade, esperamos que possa:

Descrever sobre os princípios da coleta de dados.

Analizar sobre o pré-processamento dos dados.

Aplicar os conceitos do aprendizado de máquina.



1

univ.
erit
adap
tive

Apresentação

Olá! Nesta unidade, você começará seus estudos sobre as fases de um projeto de aprendizado de máquina.

O objetivo principal é trazer para você os conceitos sobre os princípios da coleta de dados, refletir sobre o pré-processamento dos dados e aplicar os conceitos do aprendizado de máquina de forma prática.

Você verá que os dados podem ser armazenados em seu banco de dados de aplicativo ou por outros provedores de serviços terceirizados. Se você deseja, por exemplo, analisar o comportamento de gastos de seus usuários, pode ser necessário retirar os registros de compras de seu próprio banco de dados. Por outro lado, se você deseja entender os interesses do usuário, pode ser necessário encontrar provedores de serviços terceirizados especializados na geração desse conteúdo.

Além disso, ao longo desta unidade será discutido que o pré-processamento de dados muitas vezes pode ter impacto significativo no desempenho de generalização de um algoritmo de aprendizagem de máquina. A eliminação de instâncias de ruído é um dos problemas mais difíceis. Normalmente, as instâncias removidas têm instâncias excessivamente divergentes que possuem muitos valores nulos de recursos. Esses recursos excessivamente divergentes também são chamados de outliers. Além disso, uma abordagem comum para lidar com a inviabilidade da aprendizagem de conjuntos de dados muito grandes é selecionar uma única amostra do conjunto de dados. A falta de tratamento de dados é outro problema frequente tratado nas etapas de preparação de dados (TENG, 1999).

Desse modo, você começará agora uma disciplina que fará diferença em sua vida profissional. Bons estudos.

1.1 As fases de um projeto de aprendizado de máquina

O aprendizado de máquina é uma área de grande interesse entre os entusiastas de tecnologia. Visto como um ramo da Inteligência Artificial (IA), é basicamente um algoritmo ou modelo que se aprimora por meio do “aprendizado” e, como resultado, torna-se cada vez mais

proficiente no desempenho de sua tarefa. As aplicações de aprendizado de máquina são amplamente difundidas, pois estão se tornando rapidamente parte integrante de diferentes campos, como medicina, e-commerce, bancos, etc.

A crescente complexidade dos problemas a serem computacionalmente tratados, e da velocidade e volume de dados gerados por diferentes setores, motivou o desenvolvimento de ferramentas computacionais mais sofisticadas e autônomas, mais independentes da intervenção humana, para a aquisição de conhecimento. A maioria dessas ferramentas é baseada em Aprendizado de Máquina (AM), uma subárea da IA que faz parte de várias das tecnologias atualmente utilizadas (...) O maior crescimento ocorre em empresas, onde o uso de IA e AM varia desde sua adoção como estratégia de negócio, como é o caso da companhia Google, até para o desenvolvimento de aplicações marginais ao negócio, como os assistentes automáticos comuns nos aplicativos e sites de diversos bancos. A Netflix, por exemplo, utiliza IA no sistema de recomendação e para identificação de padrões de gosto dos seus usuários para a produção de conteúdo próprio (GAMA et al., 2011, p. 1).

De acordo com Martinez (2020), o processo de aprendizado de máquina pode ser dividido em sete etapas:

- Etapa 1: coleta de dados.
- Etapa 2: preparando os dados.
- Etapa 3: escolhendo um modelo.
- Etapa 4: treinamento.
- Etapa 5: avaliação.
- Etapa 6: ajuste de parâmetros.
- Etapa 7: previsão.

Etapa 1: coleta de dados

Para desenvolver o modelo de aprendizado de máquina, a primeira etapa seria coletar dados relevantes que podem ser usados para diferenciar os objetos entre si, como duas frutas. Parâmetros diferentes podem ser usados para classificar uma fruta como laranja ou maçã. A primeira característica seria a cor da própria fruta, a segunda seria a forma da fruta. Usando esses recursos, espera-se que o modelo possa diferenciar com precisão entre as duas frutas.

Um mecanismo seria necessário para reunir os dados para os dois recursos escolhidos. Por exemplo, para a coleta de dados de cor, pode-se usar um espectrômetro e, para os dados de forma, pode-se usar fotos dos frutos para que sejam tratados como figuras 2D (ou 3D, de acordo com o escopo do projeto). Para fins de coleta de dados, se tentaria obter o máximo possível de tipos diferentes de maçãs e laranjas para criar diversos conjuntos de dados para os recursos.

FIGURA 01 Ilustração da dinâmica para identificação das frutas (maçã e laranja)



Fonte: SHUTTERSTOCK.COM, 2021.

A etapa de coleta de dados é a base do processo de aprendizado de máquina. Erros, como escolher os recursos incorretos ou focar em tipos limitados de entradas para o conjunto de dados, podem tornar o modelo completamente ineficaz. Por isso, é imperativo que as considerações necessárias sejam feitas ao coletar dados, pois os erros cometidos nessa etapa só se ampliam à medida que se avança para as etapas posteriores.

Etapa 2: preparando os dados

Depois de coletar os dados para os dois recursos, a próxima etapa seria preparar os dados para outras etapas. Um foco principal desse estágio é reconhecer e minimizar quaisquer tendências potenciais nos conjuntos de dados para os dois recursos.

Primeiro, se poderia randomizar a ordem dos dados para as duas frutas. Isso ocorre porque não se quer que o pedido tenha qualquer influência nas escolhas do modelo. Além disso, se examinariam os conjuntos de dados para qualquer distorção em relação a uma fruta em particular. Novamente, isso ajudaria a identificar e retificar um viés potencial, pois significaria que o modelo seria apto a identificar uma fruta corretamente, mas poderia ter dificuldades com a outra.

Aplicações baseadas em AM utilizam heurísticas que buscam por modelos capazes de representar o conhecimento presente em um conjunto de dados. Em geral, os conjuntos de dados são estruturados em formato tabular, uma matriz atributo-valor, em que cada linha representa um objeto (instância ou exemplo) e cada coluna representa um atributo

(característica ou variável). Os atributos podem ser divididos em atributos preditivos, cujos valores descrevem características dos objetos, que formam um vetor de entrada, e atributo alvo, cujo valor rotula o objeto, com uma classe ou valor numérico. Essas denominações têm por origem o frequente uso dos valores dos atributos preditivos de um objeto para predizer o valor de seu atributo alvo. Nem todos os conjuntos de dados possuem atributo alvo. Quando possuem, são chamados de conjuntos de dados rotulados (GAMA et al., 2011, p. 2).



Dados bem preparados para seu modelo podem melhorar sua eficiência. Isso pode ajudar a reduzir os pontos cegos do modelo, o que se traduz em maior precisão das previsões. Portanto, faz sentido deliberar e revisar seus conjuntos de dados de modo que possam ser ajustados para produzir resultados melhores e significativos.

Etapa 3: escolhendo um modelo

Existem vários modelos desenvolvidos por cientistas de dados que podem ser usados para diferentes fins. Esses modelos são projetados com objetivos diferentes em mente. Por exemplo, alguns modelos são mais adequados para lidar com textos, enquanto outro modelo pode ser mais bem equipado para lidar com imagens.

As opções de modelos de aprendizado de máquina podem ser exploradas em três categorias amplas. A primeira categoria são os modelos de **aprendizagem supervisionada**. Em tais modelos, o resultado é conhecido, portanto refina-se continuamente o próprio modelo até que a saída alcance o nível de precisão desejado. O modelo de regressão é um exemplo de aprendizado supervisionado.

Se o resultado for desconhecido e seja necessária a classificação, a segunda categoria, definida como sendo a **aprendizagem não supervisionada**, deve ser utilizada. Exemplos de aprendizagem não supervisionada incluem K-means e algoritmo Apriori.

A terceira categoria é a aprendizagem por reforço. Ela se concentra em aprender a tomar melhores decisões com base na tentativa e erro. Esse modelo é frequentemente usado em ambientes de negócios. O processo de decisão do Markov é um exemplo de aplicação.

Etapa 4: treinamento

No centro do processo de aprendizado de máquina, está o treinamento do modelo. Grande parte do “aprendizado” é feito nessa fase. Aqui, usa-se a parte do conjunto de dados alocado para o treinamento para ensinar o modelo a diferenciar entre as duas frutas.

Se abordar o modelo em termos matemáticos, as entradas, ou seja, os dois recursos teriam coeficientes. Esses coeficientes são chamados de pesos dos recursos, isso é conhecido como o viés do modelo. O processo de determinação de seus valores é de tentativa e erro. Inicialmente, escolhem-se valores aleatórios para eles e se fornecem entradas. A saída alcançada é comparada com a saída real e a diferença é minimizada tentando diferentes valores de pesos e vieses. As iterações são repetidas usando entradas diferentes do conjunto de dados de treinamento até que o modelo alcance o nível desejado de precisão.



O treinamento requer paciência e experimentação. Também é útil ter conhecimento da área onde o modelo seria implementado. Por exemplo, se um modelo de aprendizado de máquina for usado para identificar clientes de alto risco para uma seguradora, o conhecimento de como a indústria de seguros opera aceleraria o processo de treinamento, já que suposições mais fundamentadas podem ser feitas durante as iterações. O treinamento pode ser altamente recompensador se o modelo começar a ter sucesso em sua função.

Etapa 5: avaliação

Com o modelo treinado, ele precisa ser testado para ver se funcionaria bem em situações do mundo real. É por isso que a parte do conjunto de dados criado para avaliação é usada para verificar a proficiência do modelo. Isso coloca o modelo em um cenário no qual se depara com situações que não fizeram parte de seu treinamento. Nesse caso, pode significar tentar identificar um tipo de maçã ou laranja completamente novo para o modelo. Porém, por meio de seu treinamento, o modelo deve ser capaz de extrapolar as informações e julgar se a fruta é uma maçã ou uma laranja.

A avaliação se torna muito importante quando se trata de aplicações comerciais. A avaliação permite que os cientistas de dados verifiquem se os objetivos que se propõem a atingir foram alcançados ou não. Se os resultados não forem satisfatórios, as etapas anteriores precisam ser revisadas para que a causa raiz por trás do baixo desempenho do modelo possa ser identificada e, posteriormente, retificada.

Caso a avaliação não seja feita corretamente, o modelo pode não se destacar no cumprimento da finalidade comercial desejada. Portanto, a avaliação do modelo é essencial para evitar os efeitos nocivos mencionados. Se a avaliação for bem-sucedida, passa-se para a etapa de ajuste dos parâmetros.

Quando são utilizados ou propostos métodos de aprendizado de máquina para extração de conhecimento, devem ser realizados estudos experimentais para avaliar o desempenho desses métodos. Para uma boa avaliação e fundamental projetar criteriosamente os experimentos a serem realizados, identificando os critérios de avaliação e variáveis independentes, além da neutralização de variáveis aleatórias que fogem ao controle do experimentador. Quanto a identificação das variáveis independentes, a abordagem frequentemente empregada inclui a utilização de um ou mais métodos de controle para a comparação com novos métodos ou algoritmos propostos. Outros fatores, tais como a acomodação do algoritmo a um conjunto de dados em particular, são geralmente tratados pela repetição da execução do algoritmo várias vezes em diferentes (sub) amostras dos dados (BERNARDINI, 2006, p. 33-34).

Etapa 6: ajuste de parâmetros

Essa etapa tenta melhorar os resultados positivos alcançados durante a etapa de avaliação. Para o exemplo, seria conferido se é possível tornar o modelo ainda melhor no reconhecimento de maçãs e laranjas. Existem diferentes maneiras de melhorar o modelo.

Um deles é revisitar a etapa de treinamento e usar várias varreduras do conjunto de dados de treinamento para treinar o modelo. Isso pode levar a uma maior precisão, pois a maior duração do treinamento fornece mais exposição e melhora a qualidade do modelo. Outra maneira de fazer isso é refinar os valores iniciais dados ao modelo. Valores iniciais aleatórios geralmente produzem resultados ruins, pois são gradualmente refinados por tentativa e erro. No entanto, se for possível chegar a valores iniciais melhores ou talvez iniciar o modelo usando uma distribuição em vez de um valor, os resultados podem ser melhores.

Etapa 7: previsão

A etapa final do processo de aprendizado de máquina é a previsão. Esta é a fase em que se considera o modelo pronto para aplicações práticas. O modelo de fruta agora deve ser capaz de responder à pergunta se a fruta dada é uma maçã ou uma laranja. O modelo ganha independência da interferência humana e tira suas próprias conclusões com base em seus conjuntos de dados e treinamento. O desafio para o modelo permanece se ele pode superar ou pelo menos igualar o julgamento humano em diferentes cenários relevantes.

A etapa de previsão é o que o usuário final vê quando utiliza o modelo de aprendizado de máquina em seu respectivo setor. Os modelos de aprendizado de máquina podem processar e vincular grandes quantidades de dados, ou seja, ao utilizar o modelo de aprendizado de máquina como uma ferramenta, pode-se tomar decisões melhores com muito menos esforço e com maior precisão.

1.1.1 Princípios da coleta de dados

Aprendizado de máquina é o processo de permitir que os computadores modifiquem ou adaptem suas ações para que esse recurso possa ser preciso. Imagine que você está jogando um game contra um computador. Você pode vencê-lo todas as vezes no início, mas depois de utilizar algumas vezes, a inteligência do inimigo começa a vencê-lo, até que finalmente você nunca ganha. Ou você está piorando ou o computador está aprendendo a vencer. Tendo aprendido a vencê-lo, ele pode continuar e usar as mesmas estratégias contra outros jogadores, de modo que não comece do zero com cada novo jogador; esta é uma forma de generalização.

Os dados podem ser armazenados em seu banco de dados de aplicativo ou por outros provedores de serviços terceirizados. Se você deseja, por exemplo, analisar o comportamento de gastos de seus usuários, pode ser necessário retirar os registros de compras de seu próprio banco de dados. Por outro lado, se você deseja entender os interesses do usuário, pode ser necessário encontrar provedores de serviços terceirizados especializados na geração desse conteúdo.

De acordo com Tang (2019), os dados podem ser categorizados em dois tipos: **estruturados** e **não estruturados**. Dados estruturados referem-se a tipos de dados bem definidos que são armazenados em bancos de dados fáceis de pesquisar, enquanto os dados não estruturados são todos os elementos que se pode coletar, mas que não são fáceis de pesquisar.



Segue abaixo a relação e exemplificação desses tipos de dados:

Dados estruturados: números, datas, textos, etc.

Dados não estruturados: arquivos de texto e e-mails, arquivos de mídia (vídeos, músicas, fotos), outros arquivos grandes, etc.

Os dados podem ser vistos como coleções de pontos de dados. Talvez o componente mais importante de qualquer problema (e método) de aprendizado de máquina sejam os dados. De acordo com Jung (2020), os dados são considerados como coleções de indivíduos e pontos de dados que são unidades atômicas de “contêineres de informação”. Pontos de dados podem representar documentos de texto, amostras de sinais de séries temporais geradas por sensores, séries temporais inteiras geradas por coleções de sensores, quadros

em um único vídeo, vídeos em um banco de dados de filmes, árvores dentro de uma floresta, florestas dentro de uma coleção de florestas, etc.

Pode-se utilizar o conceito de pontos de dados de maneira altamente abstrata e, portanto, muito flexível. Os pontos de dados podem representar coleções de objetos muito diferentes. Talvez o único requisito conceitual seja que os pontos de dados representem objetos de um tipo semelhante. Outro requisito para uma boa escolha de pontos de dados é que se deve ter acesso a muitos deles. Em termos gerais, a maioria dos métodos de aprendizado de máquina depende de médias estatísticas para obter previsões; quanto mais pontos de dados usados para a média, melhor.

FIGURA 02 Representação do processo de coleta de dados



Fonte: SHUTTERSTOCK.COM, 2021.

Segundo Qualcomm Technologies (2021), no estágio de preparação de dados, estão as etapas de coleta e pré-processamento de dados. Coletar dados para treinar o modelo do aprendizado de máquina é a etapa básica no pipeline (processo que será realizado) de aprendizado de máquina. As previsões feitas pelos sistemas só podem ser tão boas quanto os dados com os quais eles foram treinados. A seguir, estão alguns dos problemas que podem surgir na coleta de dados:

- **Dados imprecisos:** os dados coletados podem não estar relacionados à declaração do problema.
- **Dados ausentes:** subdados podem estar faltando. Isso pode assumir a forma de

valores vazios em colunas ou imagens ausentes para alguma classe de previsão.

- **Desequilíbrio de dados:** algumas classes ou categorias nos dados podem ter um número desproporcionalmente alto ou baixo de amostras correspondentes. Como resultado, eles correm o risco de serem sub-representados no modelo.
- **Viés de dados:** dependendo de como os dados, os assuntos e os próprios rótulos são escolhidos, o modelo pode propagar preconceitos inerentes a gênero, política, idade ou região, por exemplo. O enviesamento de dados é difícil de detectar e remover.

Qualcomm Technologies (2021) apresenta várias técnicas que podem ser aplicadas para resolver esses problemas. Entre os elementos, pode-se destacar os seguintes pontos:

- **Conjuntos de dados pré-limos e disponíveis gratuitamente:** se a declaração do problema (por exemplo, classificação de imagem, reconhecimento de objeto) estiver alinhada com um conjunto de dados limpo, pré-existente e devidamente formulado, aproveite a experiência existente de código aberto.
- **Rastreamento e busca pela web:** ferramentas automatizadas e bots (aplicação de software que realiza tarefas que foram previamente programadas) podem rastrear em busca de dados pela Internet.
- **Dados privados:** os profissionais que trabalham com a aprendizagem de máquina podem criar seus próprios dados. Isso é útil quando a quantidade de dados necessária para treinar o modelo é pequena e a declaração do problema é muito específica para generalizar em um conjunto de dados de código aberto.
- **Dados personalizados:** as agências podem criar ou terceirizar os dados mediante o pagamento de uma taxa.

De acordo com Smola e Vishwanathan (2008), é útil caracterizar os problemas de aprendizagem de acordo com o tipo de dados que eles usam. Isso é uma grande ajuda ao encontrar novos desafios, já que, frequentemente, problemas em tipos de dados semelhantes podem ser resolvidos com técnicas muito semelhantes. Por exemplo, o processamento de linguagem natural e a bioinformática utilizam ferramentas semelhantes para sequências de texto em linguagem natural e para sequências de DNA. Os vetores constituem a entidade mais básica que se pode encontrar no trabalho. Por exemplo, para uma seguradora de vida pode ser interessante obter o vetor de variáveis (pressão arterial, frequência cardíaca, altura, peso, nível do colesterol, gênero) para inferir a expectativa de vida de um cliente potencial.

Um agricultor pode estar interessado em determinar a maturação da fruta com base em dados específicos, como tamanho, peso e dados espectrais. Um engenheiro pode querer encontrar dependências em pares (tensão, corrente). Da mesma forma, pode-se querer representar documentos por um vetor de contagens que descreve a ocorrência de palavras. O último é comumente referido como recursos de pacote de palavras. Um dos desafios em lidar com vetores é que as escalas e unidades de diferentes coordenadas podem variar amplamente. Por exemplo, pode-se medir a altura em quilogramas, libras, gramas, toneladas e pedras, todos os quais totalizariam as mudanças multiplicativas. Da mesma forma, ao representar temperaturas, tem-se uma classe completa de transformações afins, dependendo se são representadas em termos de Celsius, Kelvin ou Fahrenheit. Uma maneira de lidar com esses problemas de forma automática é normalizar os dados.

Um requisito importante para algoritmos de AM é que sejam capazes de lidar com dados imperfeitos. Muitos conjuntos de dados apresentam algum tipo de problema, como presença de ruídos, dados inconsistentes, dados ausentes e dados redundantes. Algoritmos de AM devem, idealmente, ser robustos a esses problemas, minimizando sua influência no processo indutivo. Entretanto, dependendo de sua extensão, esses problemas podem prejudicar ou inviabilizar o aprendizado. Técnicas de pré-processamento permitem identificar e reduzir, ou até mesmo eliminar, esses problemas (GAMA et al., 2011, p. 4).

Smola e Vishwanathan (2008) apresentam meios de realizar esses problemas de forma automática. Entre as soluções, pode-se destacar a organização a partir de **listas**. Em alguns casos, os vetores que são obtidos podem conter um número variável de recursos. Por exemplo, um médico pode não necessariamente decidir realizar uma bateria completa de testes de diagnóstico se o paciente parecer saudável. Os conjuntos podem aparecer em problemas de aprendizagem sempre que houver muitas causas potenciais de um efeito, que não são bem determinadas. As matrizes são um meio conveniente de representar relacionamentos entre pares. Por exemplo, em aplicativos de filtragem colaborativa, as linhas da matriz podem representar usuários, enquanto as colunas correspondem a produtos. Somente em alguns casos se terá conhecimento sobre uma determinada combinação (usuário, produto), como a avaliação do produto por um usuário.

A utilização das *strings* (tipo de dado para textos e/ou caracteres) ocorre com frequência, principalmente nas áreas de bioinformática e processamento de linguagem natural. Elas podem ser a entrada para os problemas de estimativa, por exemplo, ao classificarem um e-mail como spam, ao tentarem localizar todos os nomes de pessoas e organizações em um texto ou ao modelarem a estrutura do tópico de um documento. Da mesma forma, elas podem constituir o resultado de um sistema. Por exemplo, pode-se querer realizar o resumo de documentos, tradução automática ou tentar responder a perguntas de linguagem natural.

Estruturas compostas são os objetos de ocorrência mais comum. Ou seja, na maioria das situações, tem-se uma combinação estruturada de diferentes tipos de dados. Por exemplo, uma página da web pode conter imagens, texto e tabelas, que, por sua vez, contêm números e listas, todos os quais podem constituir nós em um gráfico de páginas da web vinculadas entre si. Uma boa modelagem estatística leva em consideração tais dependências e estruturas, a fim de adaptá-la em um modelo flexível.

1.1.2 Pré-processamento dos dados

Segundo García *et al.* (2015), o pré-processamento de dados é uma etapa frequentemente negligenciada, mas importante no processo de mineração de dados. A coleta de dados é geralmente um processo vagamente controlado, resultando em elementos fora do intervalo de valores, por exemplo, combinações de dados impossíveis (como, Sexo: masculino; Grávida: sim), valores ausentes, etc. Analisar os dados que não foram cuidadosamente selecionados para tais problemas pode produzir resultados enganosos. Assim, a representação e qualidade dos dados são fundamentais antes de executar uma análise. Se houver informações irrelevantes e redundantes presentes ou dados ruidosos e não confiáveis, então a descoberta de conhecimento é mais difícil de conduzir.

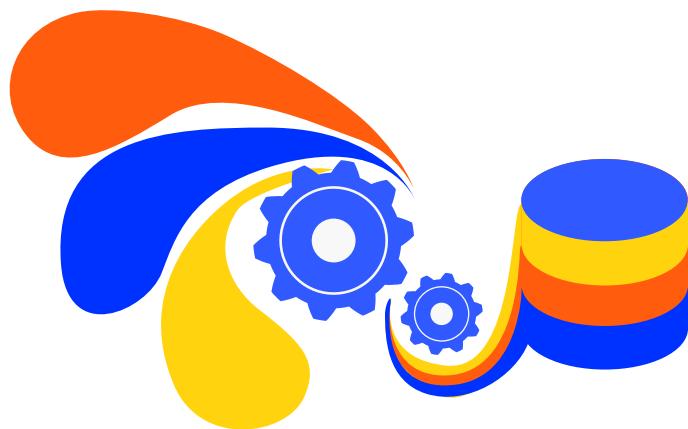


No livro “Extração de conhecimento de dados data mining”, os autores João Gama, André Ponce de Leon Carvalho, Katti Faceli, Ana Carolina Lorena e Márcia Oliveira apresentam sobre os princípios e as tendências nas áreas de aprendizagem automática, reconhecimento de padrões e análise de dados.

O ambiente computacional DLE tem como principal objetivo prover um framework para que novas técnicas de pré-processamento de dados possam ser facilmente e rapidamente implementadas e avaliadas experimentalmente. O ambiente computacional DLE é composto por dois módulos distintos: a biblioteca de classes Discover Object Library (DOL), cujo objetivo principal é ser uma biblioteca de métodos de pré-processamento de dados, e o ambiente para gerenciamento de experimentos SNIFFER, cujo objetivo principal é automatizar avaliações e comparações experimentais de algoritmos de aprendizado. Para utilizar a biblioteca de classes DOL, foi proposta uma sintaxe padrão para representação de dados, denominada Discover Dataset Syntax (DSX) (BERNARDINI, 2006, p. 60).

A preparação de dados pode levar uma quantidade considerável de tempo de processamento. O pré-processamento de dados inclui a preparação de dados, composta pela integração, limpeza, normalização e transformação de dados. Também inclui tarefas de redução de dados, tais como seleção de recurso, seleção de instância, discretização, etc. O resultado esperado após um encadeamento confiável de tarefas de pré-processamento de dados é um conjunto de dados final, que pode ser considerado correto e útil para algoritmos de mineração de dados adicionais.

FIGURA 03 Representação do processo de preparação de dados



Fonte: SHUTTERSTOCK.COM, 2021.

O pré-processamento de dados muitas vezes pode ter impacto significativo no desempenho de generalização de um algoritmo de aprendizagem de máquina. A eliminação de instâncias de ruído é um dos problemas mais difíceis. Normalmente, as instâncias removidas têm instâncias excessivamente divergentes que possuem muitos valores nulos de recursos. Esses recursos excessivamente divergentes também são chamados de outliers. Além disso, uma abordagem comum para lidar com a inviabilidade da aprendizagem de conjuntos de dados muito grandes é selecionar uma única amostra do conjunto de dados. A falta de tratamento de dados é outro problema frequente tratado nas etapas de preparação de dados (TENG, 1999).

García *et al.* (2015) afirmam que, nos últimos anos, essa área se tornou de grande importância porque a aprendizagem de máquina e os algoritmos de mineração requerem dados significativos e gerenciáveis para operar corretamente e para fornecer conhecimento útil, previsões ou descrições. É bem sabido que a maioria dos esforços feitos em um aplicativo de descoberta de conhecimento é dedicado a dados e tarefas de preparação e redução. Teóricos e praticantes estão constantemente pesquisando as técnicas de pré-processamento de dados para garantir resultados confiáveis e precisos, negociando com eficiência e complexidade de tempo.

Os algoritmos de aprendizado lógico e simbólico são capazes de processar dados simbólicos e categóricos apenas. No entanto, no mundo real os problemas envolvem recursos simbólicos e numéricos. Portanto, há uma questão importante para discretizar os recursos numéricos (contínuos). O agrupamento de valores de aprendizado simbólico é um processo útil; é um problema conhecido que recursos com muitos valores são superestimados no processo de seleção dos recursos mais informativos, tanto para a construção da indução de árvores de decisão quanto para derivar as regras de decisão. Além disso, em dados do mundo real, a representação de dados costuma usar muitos recursos, mas apenas alguns deles podem ser relacionados ao conceito de alvo. Pode haver redundância, em que certos recursos são correlacionados, de modo que não é necessário incluir todos eles na modelagem; e interdependência, na qual dois ou mais recursos entre eles transmitem importantes informações que são obscuras se algum deles estiver incluído em seu próprio conjunto (GUYON; ELISSEEFF, 2003).

A seleção do subconjunto de recursos é o processo de identificação e uma dinâmica para remover o máximo de informações irrelevantes e redundantes. Isso reduz a dimensionalidade dos dados e pode permitir que algoritmos de aprendizagem operem mais rápido e mais efetivamente. Em alguns casos, a precisão na classificação futura pode ser melhorada; em outros, o resultado é mais compacto, pois a representação pode ser facilmente interpretada do conceito de destino. Além disso, o problema de interação de recursos pode ser abordado pela construção de novos recursos a partir do recurso básico (GUYON; ELISSEEFF, 2003).

De acordo com Qualcomm Technologies (2021), os dados e imagens brutos do mundo real costumam ser incompletos, inconsistentes e carentes de certos comportamentos ou tendências. Também é provável que contenham muitos erros. Assim, uma vez coletados, eles devem ser pré-processados em um formato que o algoritmo de aprendizado de máquina possa usar para o modelo. O pré-processamento inclui uma série de técnicas e ações:

Limpeza de dados: essas técnicas, manuais e automatizadas, removem dados incorretamente adicionados ou classificados.

Imputações de dados: a maioria das estruturas do aprendizado de máquina inclui métodos e APIs para balancear ou preencher dados ausentes. As técnicas geralmente incluem a imputação de valores ausentes com desvio padrão, média, mediana, entre outras técnicas.

Integração de dados: combinar vários conjuntos de dados para obter um grande corpus pode superar a incompletude em um único conjunto de dados.

Normalização de dados: o tamanho de um conjunto de dados afeta a memória e o processamento necessários para iterações durante o treinamento. A normalização reduz o tamanho, reduzindo a ordem e a magnitude dos dados.

1.1.3 Planejamento do desenvolvimento de um projeto que utilize os princípios do aprendizado de máquina

Com a contínua popularização e o desenvolvimento da tecnologia da informação e tecnologia inteligente, o desenvolvimento de *software* e a engenharia da computação tornaram-se gradualmente um pilar da indústria para promover o rápido desenvolvimento da economia nacional. O aumento de empresas de tecnologia da informação em todo o mundo marca uma nova etapa no desenvolvimento da indústria de *software*. O controle do progresso no gerenciamento de projetos de *software* é baseado na avaliação científica e no controle efetivo da mão de obra, recursos, cronograma, risco, etc., e o *software* correspondente ao plano de desenvolvimento é formulado.

Os recursos existentes são alocados racionalmente para garantir que o investimento mínimo possa criar os maiores lucros, de modo a se tornarem uma das principais prioridades na gestão do dia a dia das empresas de *software*. Para pequenas e médias empresas, o método tradicional de trabalho é também usado no gerenciamento de cronograma de projeto de *software*, o que não só torna a limitação do método de gestão, como também aumenta o custo do trabalho de desenvolvimento de *software*.



Existem muitos fatores que podem influenciar o desenvolvimento do projeto de *software*, fatores técnicos, fatores ambientais, fatores de financiamento, fatores humanos, etc. O fator humano é o mais importante na implementação de projetos de desenvolvimento de *software*, e os fatores técnicos são, em outras palavras, fatores humanos.

O desenvolvimento de projetos de *software* geralmente requer mudanças frequentes. Provavelmente porque os usuários pensam que as mudanças de *software* requerem apenas que os programadores alterem o código e que o custo de modificação não é grande. Mas isso não é o caso. Mudanças nos pensamentos do usuário resultam em uma mudança nos requisitos, e o gerente de projeto é incapaz de rejeitar as alterações do usuário por vários motivos. Isso afetará invisivelmente o andamento do projeto. Se a qualidade de um subprojeto concluído não é a esperada, por exemplo, muitas vezes há lacunas, dificuldades

de manutenção e assim por diante, esse subprojeto deve ser reescrito, o que desperdiçaria os recursos humanos e outros insumos. Portanto, a qualidade do subprojeto afeta o progresso de todo o projeto, e o subprojeto com maior prioridade afetará a prioridade para os demais (HULETT, 2016).

FIGURA 04 Representação do processo de desenvolvimento de *software*



Fonte: SHUTTERSTOCK.COM, 2021.

1.1.3.1 Impacto das alterações de recursos e orçamento no cronograma

Os recursos aqui se referem a recursos humanos. Em alguns subprojetos, pode haver membros da equipe insuficientes ou uma pessoa é responsável por vários subprojetos ao mesmo tempo. Outro recurso se refere aos recursos de informação. Os padrões legais de cada país são diferentes. A renda dos cidadãos em cada cidade é diferente. Os padrões de cada setor não são uniformes. Esses recursos de informação raramente são fornecidos pelos clientes. Além disso, outros recursos referem-se a equipamentos de desenvolvimento de *software* de ambiente. Esses recursos não afetarão o progresso de todo o projeto no tempo (LAVALLÉE, 2015).

O valor do orçamento que afeta os outros recursos acabará por afetar o progresso geral. Por exemplo, um ambiente de desenvolvimento de alto orçamento que usa um ambiente de alto desempenho acelerará a conclusão do projeto e vice-versa (LAVALLÉE, 2015).

1.1.3.2 Desenvolvimento de projetos que utilizam os princípios da aprendizagem da máquina

O aprendizado de máquina envolve muitas disciplinas, como a probabilidade, estatística, teoria da complexidade algorítmica e assim por diante. O aprendizado de máquina é dedicado a como fazer os computadores simularem o comportamento de aprendizagem humana. É usado para ajudar a máquina a adquirir novos conhecimentos de informação, o que a torna mais completa. A inteligência artificial, o reconhecimento de face e de carros sem motorista, que frequentemente são visualizados na literatura, também utilizam a tecnologia de algoritmo de aprendizado de máquina (JORDAN, 2015).

1.1.3.2.1 Etapas gerais na aplicação do aprendizado de máquina

Selecione um modelo adequado, o que geralmente depende do problema. Para diferentes problemas e tarefas, você precisa escolher o modelo apropriado. O modelo é uma coleção de funções. Para julgar a qualidade de uma função, é necessário determinar uma métrica, que é a **função de perda**. A determinação da função de perda também depende do problema específico. Por exemplo, o problema de regressão geralmente adota a distância euclidiana (KYROLA, 2014).

A classificação de problemas geralmente utiliza uma função de custo de entropia cruzada. Descobrir a melhor função, como encontrar a melhor entre as muitas funções. Essa etapa é a maior dificuldade. Frequentemente não é uma tarefa fácil de fazer com rapidez e precisão. Comumente utilizados, os métodos são algoritmo de descida gradiente, método de mínimos quadrados e outros truques (KYROLA, 2014).

1.1.3.2.2 Aplicativos emergentes de aprendizado de máquina

O aprendizado de máquina está tendo um impacto dramático no caminho do desenvolvimento de *software*, que é projetado para acompanhar as mudanças nos negócios. O aprendizado de máquina é tão dramático porque ajuda você a usar dados para conduzir as regras e a lógica de negócios. Com os modelos tradicionais de desenvolvimento de *software*, os programadores escreveram a lógica com base no estado atual do negócio e, em seguida,

adicionam dados relevantes. O valor do aprendizado de máquina é que ele permite que você continuamente aprenda com os dados e preveja o futuro. Esse poderoso conjunto de algoritmos e modelos está sendo usado em todos os setores para melhorar os processos e obter *insights* sobre padrões e anomalias dentro dos dados (HURWITZ; KIRSCH, 2018).

Mas o aprendizado de máquina não é um empreendimento solitário; é um processo de equipe que requer cientistas de dados, engenheiros de dados, analistas de negócios e líderes de negócios para colaborar. O poder do aprendizado de máquina requer a colaboração, então o foco está na solução de problemas.

De acordo com Hurwitz e Kirsch (2018), aprendizado de máquina, Inteligência Artificial (IA) e computação cognitiva estão dominando as conversas sobre como as análises avançadas emergentes podem fornecer às empresas uma vantagem competitiva. Não há debate que os líderes empresariais existentes estão enfrentando novos e imprevistos concorrentes. Essas empresas estão procurando novas estratégias que podem prepará-las para o futuro. A partir desse princípio, o aprendizado de máquina se tornou um dos tópicos mais importantes dentro de organizações de desenvolvimento, que buscam soluções inovadoras e maneiras de aproveitar os ativos de dados para ajudar a empresa a atingir um novo nível de compreensão.

Com os modelos de aprendizado de máquina apropriados, as organizações têm a capacidade de prever continuamente as mudanças nos negócios para que sejam capazes de prever o que vem a seguir. Como os dados são constantemente adicionados, os modelos de aprendizado de máquina garantem que a solução seja constantemente atualizada. O valor é direto: se você usar as fontes de dados mais adequadas e em constante mudança no contexto do aprendizado de máquina, você tem a oportunidade de prever o futuro.

O aprendizado de máquina é uma forma de IA que permite que um sistema aprenda os dados, em vez de programação explícita. Contudo, o aprendizado de máquina não é um processo simples. O aprendizado de máquina usa uma variedade de algoritmos que iterativamente aprendem com os dados para melhorar a sua performance e descrever os dados para prever resultados. À medida que os algoritmos ingerem dados de treinamento, é possível produzir modelos mais precisos com base nesses dados. Um modelo de aprendizado de máquina é a saída gerada quando você treina o algoritmo de aprendizagem de máquina com dados. Após o treinamento, quando você fornece um modelo com uma entrada, você receberá uma saída.



Um algoritmo preditivo criará um modelo preditivo. Então, ao fornecer dados ao modelo preditivo, você receberá uma previsão com base nos dados que treinaram o modelo. O aprendizado de máquina agora é essencial para a criação de modelos analíticos. Você provavelmente interage com aplicativos de aprendizado de máquina sem perceber. Por exemplo, quando você visita um site de comércio eletrônico e começa a ver produtos e a ler comentários, provavelmente você verá outros produtos semelhantes que pode achar interessantes. Essas recomendações não são codificadas por um exército de desenvolvedores. As sugestões são veiculadas no site por meio de um modelo de aprendizado de máquina. O modelo ingere seu histórico de navegação junto com os dados de navegação e a compra de outros usuários para apresentar outros produtos semelhantes que você queira comprar.

De acordo com The Royal Society (2017), embora o aprendizado de máquina já esteja dando suporte a uma gama de sistemas de uso comum, seu potencial deve avançar nos próximos anos. Em áreas da saúde à educação e de transporte até serviços sociais, há sinais de que o aprendizado de máquina pode oferecer suporte a melhorias na eficácia dos produtos ou serviços por meio de maior precisão ou melhor adaptação das intervenções.

Em uma variedade de indústrias – em que há dados suficientes disponíveis para habilitar a aprendizagem de máquina, esses dados são usados de forma eficaz e há acesso ao poder de computação suficiente –, o aprendizado de máquina pode oferecer suporte a uma mudança de etapa na entrega de produtos ou serviços. Como uma tecnologia com potencial disruptivo, o aprendizado de máquina pode mudar como as empresas são organizadas.

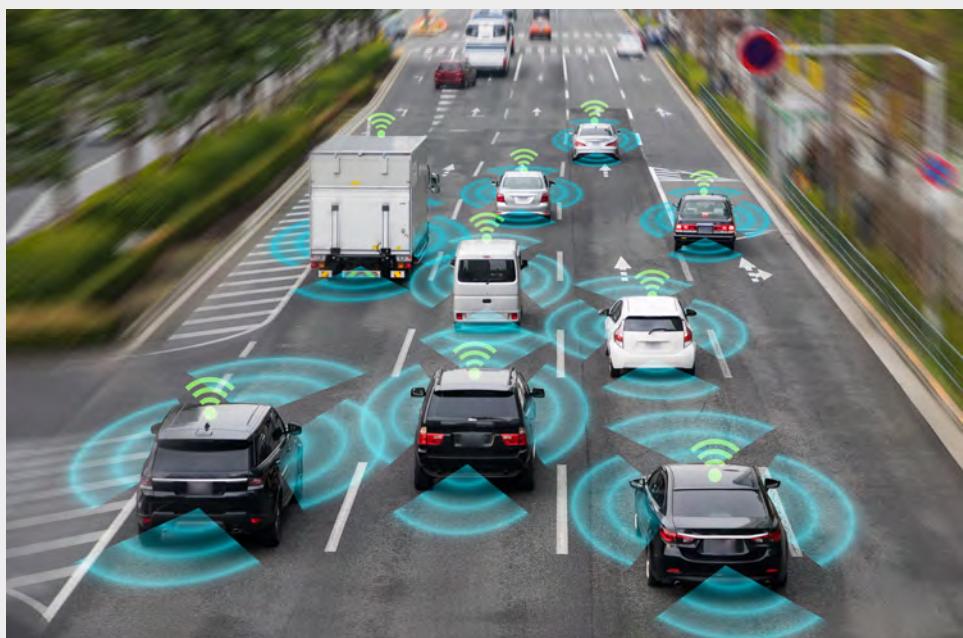
A chave para esse potencial disruptivo é a velocidade de mudança em alguns campos, enquanto em outras áreas as melhorias serão mais graduais devido ao aprendizado de máquina.



Transporte e logística

Para operar com segurança nas estradas, os veículos autônomos precisam ser capazes de reconhecer uma gama de características ambientais, incluindo: obstáculos, sinais de trânsito, pedestres e outros veículos.

FIGURA 05 Representação da utilização dos princípios da aprendizagem de máquina na área de transporte e logística



Fonte: SHUTTERSTOCK.COM, 2021.

O alcance e a variabilidade desses recursos significam que não é possível criar regras codificadas especificando com o que o veículo entrará em contato e como deveria responder em diferentes situações. A aprendizagem de máquina permite que o veículo se adapte a uma gama de recursos e responda de acordo com os eventos. Novos desenvolvimentos na sofisticação de veículos autônomos têm aplicações em uma ampla gama de configurações e indústrias.

Embora os veículos autônomos possam ter um alto perfil das aplicações potenciais do aprendizado de máquina em transporte, a tecnologia poderia oferecer suporte a uma série de funções. Os algoritmos podem analisar o histórico (os dados sobre fluxos de tráfego em uma área) usando essas informações para otimizar o sistema e prever como responderá a diferentes pressões em diferentes horas do dia. Esses *insights* podem então ser usados para reduzir o congestionamento.

Com as informações adequadas, também seria possível avaliar os fluxos de tráfego em tempo real e fazer ajustes dinâmicos para melhorá-lo. Usando o aprendizado de máquina, em vez das técnicas tradicionais de modelagem preditiva, os controladores são capazes de melhorar a precisão de suas previsões de congestionamento.

O aprendizado de máquina também pode desempenhar um papel na otimização da logística e dos processos associados. Isso pode ser feito pela recomendação de como as instalações de armazenamento devem ser estabelecidas, de modo que os produtos podem ser recuperados de forma mais eficiente, ou por meio da previsão de quanto combustível será exigido por diferentes vans de entrega, com base em sua rota provável e com o conhecimento dos fluxos de tráfego.

Esses algoritmos já estão em uso em algumas empresas, contribuindo para melhorias na eficiência e produtividade dos negócios.

Bibliografia comentada

Veja a seguir indicações de leitura que poderão complementar seus estudos. Estas leituras não são obrigatórias, mas poderão ampliar seu conhecimento em relação ao assunto abordado na unidade.

Géron (2019), no Capítulo 15 – Autoencoders, apresenta sobre as representações eficientes de dados, pré-treinamento não supervisionado utilizando autoencoders empilhados, autoencoders de remoção de ruídos e autoencoders esparsos.

Ao término desse capítulo, Harrison (2019) apresenta sobre o autoencoders variacionais e outros autoencoders, além de exemplos práticos de aplicação e utilização.

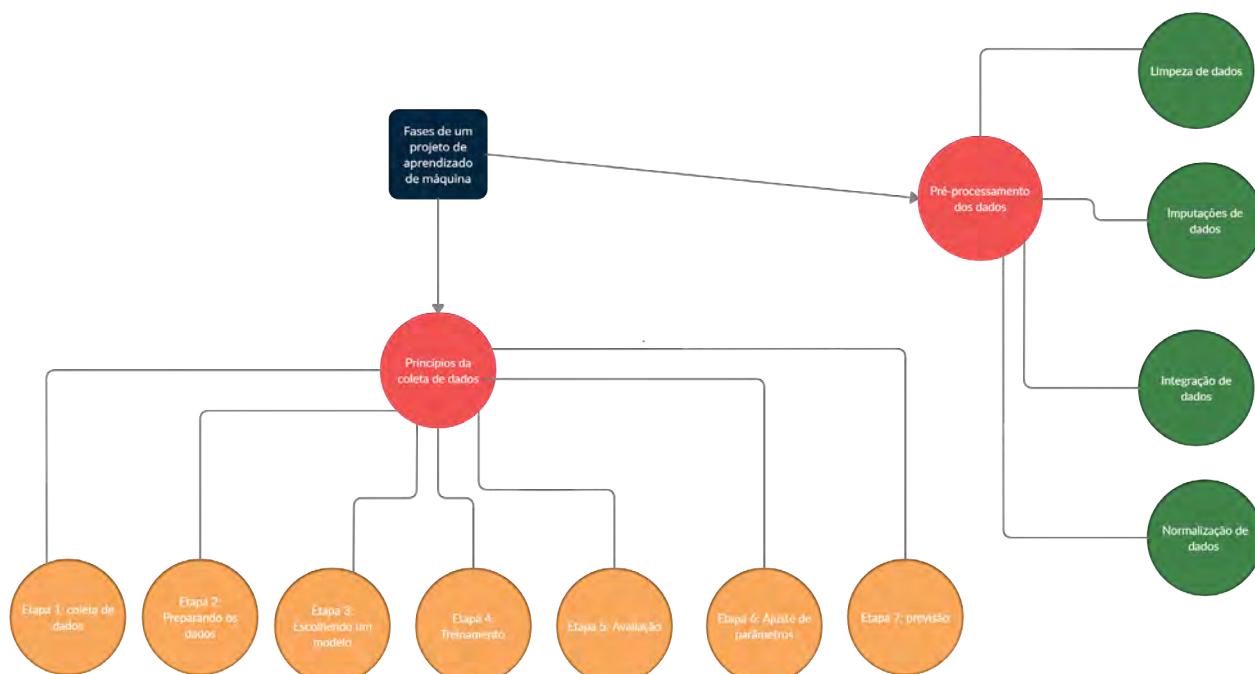
GÉRON, Aurélien. Mão à obra: aprendizado de máquina com Scikit-Learn & TensorFlow.
 Rio de Janeiro: Alta Books, 2019.

Izbicki (2020), no Capítulo 8 – Métodos de classificação, apresenta sobre os classificadores plug-in, especificando sobre os métodos de regressão, regressão logística e bayes ingênuo, bem como análise discriminante (e as suas variações: análise discriminante linear e análise discriminante quadrática).

Ao término desse capítulo, Izbicki (2020) aborda sobre os princípios de Support Vector Machines (SVM) e árvores de classificação, sendo conceitos importantes que você poderá se aprofundar para a construção de um modelo preditivo.

IZBICKI, Rafael; SANTOS, Tiago Mendonça. Aprendizado de máquina: uma abordagem estatística. 2020.

Mapa conceitual



Conclusão

Nesta unidade, você viu que o aprendizado de máquina envolve muitas disciplinas, como a probabilidade, estatística, teoria da complexidade algorítmica e assim por diante. O aprendizado de máquina é dedicado a como fazer os computadores simularem o comportamento de aprendizagem humana. É usado para ajudar a máquina a adquirir novos conhecimentos de informação, o que a torna mais completa. A inteligência artificial, o reconhecimento de face e de carros sem motorista, que frequentemente são visualizados na literatura, também utilizam a tecnologia de algoritmo de aprendizado de máquina.

Você viu que o aprendizado de máquina é uma área de grande interesse entre os entusiastas de tecnologia. Considerado como um ramo da Inteligência Artificial (IA), é basicamente um algoritmo ou modelo que se aprimora por meio do “aprendizado” e, como resultado, torna-se cada vez mais proficiente no desempenho de sua tarefa. As aplicações de aprendizado de máquina são amplamente difundidas, pois estão se tornando rapidamente uma parte integrante de diferentes campos, como medicina, e-commerce, bancos, etc.

Como visto nesta unidade, o pré-processamento de dados é uma etapa frequentemente negligenciada, mas importante no processo de mineração de dados. A coleta de dados é geralmente um processo vagamente controlado, resultando em elementos fora do intervalo de valores, por exemplo, combinações de dados impossíveis (como, Sexo: masculino; Grávida: sim), valores ausentes, etc. Analisar os dados que não foram cuidadosamente selecionados para tais problemas pode produzir resultados enganosos. Assim, a representação e qualidade dos dados são fundamentais antes de executar uma análise. Se houver informações irrelevantes e redundantes presentes ou dados ruidosos e não confiáveis, então a descoberta de conhecimento é mais difícil de conduzir.

Somente com a prática você conseguirá chegar à proficiência nessa técnica. Bons estudos!

CONSTRUÇÃO DE UM MODELO PREDITIVO

OBJETIVOS

Ao final desta unidade, esperamos que possa:

Descrever o processo de construção de um modelo preditivo.

Analizar sobre a dinâmica da validação de um modelo preditivo.

Aplicar os conceitos do modelo preditivo de forma prática.



2

união
pela
educação

Apresentação

Olá, aluno, seja muito bem-vindo! A análise de dados se refere ao processo que envolve várias ferramentas e técnicas de pesquisa qualitativa e quantitativa. Com esses dados acumulados, são produzidos alguns resultados usados para melhorar o desempenho, rendimento, redução de risco e para aumentar a produtividade do negócio. A análise de dados varia de empresa para empresa, dependendo das necessidades, portanto vários modelos de dados foram projetados para atender aos requisitos.

A análise preditiva é uma categoria de análise de dados destinada a fazer previsões sobre resultados futuros com base em dados históricos e técnicas analíticas, como modelagem estatística e aprendizado de máquina. A ciência da análise preditiva pode gerar percepções futuras com um grau significativo de precisão.

Ao longo desta unidade, será discutido que o aprendizado de máquina e a análise preditiva abordam um problema de maneira diferente. Eventualmente, a análise preditiva provavelmente se fundirá com uma aplicação de aprendizado de máquina. O aprendizado de máquina é mais adaptável, mais recente e tem maiores graus de liberdade, de modo que pode se dar ao luxo de ser mais flexível na abordagem de um problema. A análise preditiva existe há mais tempo e é mais processual em seu uso.

Desse modo, nesta unidade você verificará que os dados são caracterizados para identificar e analisar tendências. As ferramentas de análise preditiva são acionadas por vários modelos e algoritmos diferentes que podem ser aplicados a uma ampla variedade de casos de uso. Determinar quais técnicas de modelagem preditiva são melhores para a empresa e/ou projeto é a chave para obter o máximo de uma solução de análise preditiva e aproveitar os dados para tomar decisões criteriosas. Também verá como aplicar os conceitos aprendidos em alguns exemplos práticos. Bons estudos!

2.1 Construção de um modelo preditivo

A manutenção preditiva visa estimar ou prever falhas de um sistema ou seus componentes com base na experiência, leis físicas ou técnicas de aprendizado de máquina.

De acordo com Voronov (2020), as pessoas estavam interessadas em prever o tempo de vida de vários sistemas desde os tempos antigos. Por exemplo, carrinhos são usados para o transporte de mercadorias a partir do momento em que a roda foi inventada. É um sistema mecânico muito simples, no entanto existem vários componentes que podem quebrar, como um eixo ou uma roda, o que levará a uma falha ou parada. Naquela época, as pessoas usavam a inspeção visual e experiência pessoal para decidir se determinada peça deveria ser substituída antes do início da viagem. A tecnologia evoluiu, os sistemas se tornaram mais complexos, muitos novos sistemas surgiram e não é mais possível utilizar apenas a avaliação visual para decidir de forma confiável quando um ou outro componente falhará.

O modelo preditivo é a base da manutenção preditiva. A aplicação de manutenção preditiva e os modelos de prognóstico aos modernos sistemas industriais variam de eletrônica, aeronáutica, automotivo para maquinário industrial e mais aplicações são implementadas todos os dias. Por exemplo, os autores Batzel e Swanson (2009) apresentaram uma estrutura para prever falhas elétricas em um gerador de energia de aeronave que permite evitar falhas inesperadas e reduzir despesas com manutenção do sistema. Já os autores Miao *et al.* (2013) apresentam o aprendizado de máquina para prever falhas em baterias de íon de lítio em que as investigações mostram que o tempo de falhas pode ser estimado de forma confiável.

Os rolamentos são componentes importantes em sistemas mecânicos. Ali *et al.* (2015) sugerem uma abordagem de prognóstico baseada em dados para prever com precisão a avaria dos rolamentos para reduzir o custo de manutenção dos sistemas mecânicos. Os modelos preditivos de degradação em aplicações industriais podem ser integrados em novas tecnologias, como a Internet das Coisas (IoT), que pode melhorar o processo de fabricação, aumentando o tempo de atividade da infraestrutura da planta. Por exemplo, oportunidades e sugestões sobre como os modelos preditivos podem ser implementados junto com a IoT.

Segundo Voronov (2020), atualmente as áreas de interesse atual estão atreladas aos sistemas autônomos e veículos autônomos. O objetivo principal são os sistemas que podem executar as tarefas necessárias sem qualquer supervisão humana. Por exemplo, os veículos autônomos devem entregar mercadorias aos clientes ou funcionar de forma autônoma nos canteiros de obras (nas operações autônomas de sistemas dinâmicos com alto tempo de atividade e requisitos de segurança) sem o feedback direto do usuário. Essa implementação é uma tarefa desafiadora na qual a manutenção preditiva é de suma importância.

2.1.1 Definições sobre o modelo preditivo

O aprendizado de máquina e a análise preditiva abordam um problema de maneira diferente. Eventualmente, a análise preditiva provavelmente se fundirá com uma aplicação de aprendizado

de máquina. O aprendizado de máquina é mais adaptável, mais recente e tem maiores graus de liberdade, de modo que pode se dar ao luxo de ser mais flexível na abordagem de um problema. A análise preditiva existe há mais tempo e é mais processual em seu uso.

Diante desse cenário, Johnson (2020) afirma que tanto o aprendizado de máquina quanto a análise preditiva são usados para fazer previsões sobre um conjunto de dados sobre o futuro. A análise preditiva utiliza a modelagem preditiva, que pode incluir aprendizado de máquina. A análise preditiva tem um propósito muito específico: usar dados históricos para prever a probabilidade de um resultado futuro. O público da análise preditiva tende a ser profissional, adicionando um nível extra de comunicação e interpretabilidade necessárias ao seu trabalho. Para os analistas preditivos, o aprendizado de máquina é uma extensão de sua prática, outra ferramenta em sua caixa de ferramentas, que os ajuda a fazer melhor seu trabalho. Usando a aprendizagem de máquina, os analistas preditivos podem:

- Fornecer respostas, com confiança, para problemas mais complexos.
- Oferecer respostas em tempo real a perguntas que persistem ao longo do tempo com dados em constante mudança.
- Explorar tipos inteiramente novos de problemas.

Por meio do aprendizado de máquina, a análise preditiva pode expandir a forma como conduz sua análise de sentimento para ver o quanto satisfeitos seus clientes e funcionários estão.

De acordo com Tyagi (2020), a quantidade de dados consumidos está aumentando exponencialmente. Observa-se um grande volume de dados acumulado nas organizações, isso pode estar relacionado a parceiros de negócios, consumidores, aliados de aplicativos, executivos internos e externos, visitantes, etc.

Os dados são caracterizados para identificar e analisar tendências. Por outro lado, a análise de dados se refere ao processo que envolve várias ferramentas e técnicas de pesquisa qualitativa e quantitativa que utilizam esses dados acumulados e produzem alguns resultados que são usados para melhorar o desempenho, rendimento, redução de risco e para aumentar a produtividade do negócio. A análise de dados varia de empresa para empresa, dependendo das necessidades, portanto vários modelos de dados foram projetados para atender aos requisitos.

Um ensemble é um conjunto de classificadores cujas individuais são combinadas de alguma forma para classificar um novo caso. Nessa abordagem, primeiramente são retiradas, por exemplo, L amostras (subconjuntos) do conjunto de exemplos (dados)

disponível para realizar o aprendizado. Logo após, cada um desses L subconjuntos é submetido a algum algoritmo de aprendizado, induzindo assim L classificadores (hipóteses), os quais podem ser construídos em paralelo. Apesar de melhorar o poder preditivo dos algoritmos de aprendizado, os métodos de construção de ensembles normalmente geram classificadores grandes, o que pode ser indesejável (BERNARDINI, 2006, p. 32).

A **modelagem preditiva** é a subparte da análise de dados que usa a mineração de dados e probabilidade para prever resultados. Cada modelo é construído pelo número de preditores que são altamente favoráveis para determinar decisões futuras. Uma vez que os dados são recebidos para um preditor específico, um modelo analítico é formulado. Um modelo pode aplicar uma equação linear simples ou uma estrutura neural complexa delineada pelo software em questão, também se houver dados adicionais disponíveis, o modelo analítico é revisado.

Segundo Tyagi (2020), a modelagem preditiva também emprega diferentes algoritmos de regressão e análises ou estatísticas para estimar a probabilidade de um evento usando a teoria de detecção e é amplamente empregada nos campos de Aprendizado de Máquina (ML) e Inteligência Artificial (IA).



A modelagem preditiva geralmente é uma técnica estatística praticada para prever resultados futuros. Estas são soluções em termos de tecnologia de mineração de dados para analisar dados passados e recentes e produzir um modelo para identificar o comportamento futuro dos dados.

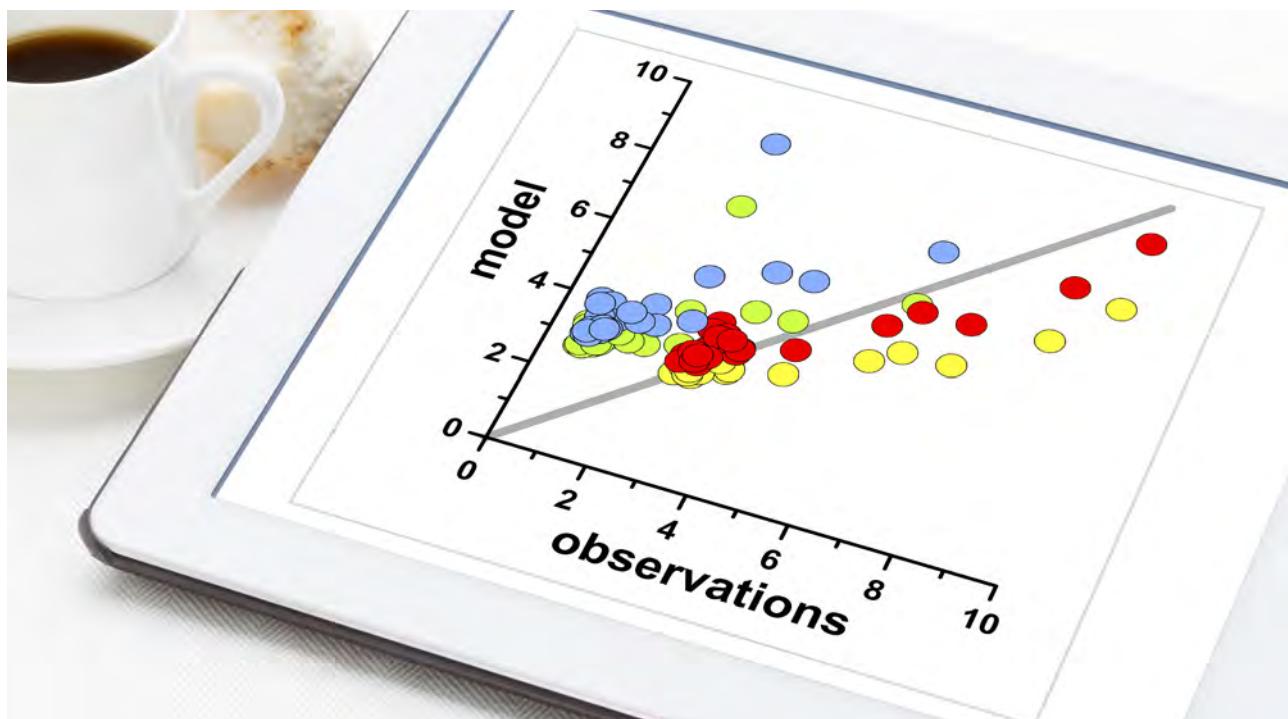
Segundo Wakefield (2021), a análise preditiva ajuda a entender possíveis ocorrências futuras analisando o passado. O aprendizado de máquina evoluiu do estudo de reconhecimento de padrões e explora a noção de que algoritmos podem aprender e fazer previsões sobre os dados. E, conforme eles começam a se tornar mais “inteligentes”, esses algoritmos podem superar as instruções do programa para tomar decisões baseadas em dados e altamente precisas.

Para ilustrar a relação entre AM e indução de modelos, supor um conjunto de dados de pacientes de um hospital. Nesse conjunto, aqui denominado hospital, cada dado (também chamado objeto, exemplo, padrão ou registro) corresponde a um paciente, representado por uma tupla formada pelos valores de características (também chamadas de campos, variáveis ou atributos) e observações do estado do paciente.

Essas características podem ser código de identificação, nome, idade, sexo, estado de origem, além de alguns sintomas e resultados de exames clínicos. Exemplos de sintomas podem ser pressão sanguínea, manchas na pele, peso e temperatura do corpo (GAMA et al., 2011, p. 4).

Wakefield (2021) afirma que a análise preditiva é conduzida por modelagem preditiva. É mais uma abordagem do que um processo. A análise preditiva e o aprendizado de máquina andam de mãos dadas, já que os modelos preditivos geralmente incluem um algoritmo de aprendizado de máquina. Esses modelos podem ser treinados ao longo do tempo para responder a novos dados ou valores, entregando os resultados de que o negócio precisa. A modelagem preditiva se sobrepõe amplamente ao campo do aprendizado de máquina.

FIGURA 06 Representação da modelagem preditiva



Fonte: SHUTTERSTOCK.COM, 2021.

Tyagi (2020) relaciona dois tipos de modelagem preditiva: o modelo paramétrico e o modelo não paramétrico:

- **Modelo paramétrico:** as suposições são a parte crucial de qualquer modelo de dados, além de tornar o modelo mais fácil, melhora as previsões, então os algoritmos que consideram suposições e simplificam a função são conhecidos

como algoritmos paramétricos da aprendizagem de máquina e um modelo de aprendizagem que compila dados com diferentes parâmetros de um tamanho predeterminado, independentemente do número de variáveis de treinamento, é denominado como modelo paramétrico.

- **Modelo não paramétrico:** os algoritmos de aprendizagem de máquina que permitem fazer suposições fortes em termos da função de mapeamento são chamados de algoritmos modelos não paramétricos. Os modelos não paramétricos são adequados para uma grande quantidade de dados sem nenhum conhecimento prévio.



Nos principais aspectos dos benefícios, a modelagem preditiva reduz o custo necessário para que as empresas prevejam resultados de negócios, fatores econômicos e ambientais, circunstâncias de mercado, etc. Mas não significa que os benefícios apareçam sem objetivo, mesmo a modelagem preditiva mostra o número de desafios também. Alguns benefícios e desafios são apresentados a seguir: previsão de custo e demanda nos negócios; previsão de fatores externos influenciados; identificação do oponente, preservação e conservação de equipamentos, entre outros recursos.

De acordo com Tyagi (2020), o modelo preditivo incorpora a execução de algoritmos na execução de dados para previsão. Esse processo é iterativo por natureza, pois treina o modelo para obter as informações mais adequadas para fins de negócios, como vários aplicativos em análises de negócios. Para mergulhar no processo de modelagem preditiva, é fundamental estabelecer os seguintes elementos:

1. **Coleta e purificação de dados:** os dados são acumulados de todas as fontes para extrair as informações necessárias; limpam-se os dados com algumas operações que eliminam os desnecessários para obter estimativas precisas. Várias fontes estão incluídas, como dados de transações e assistência ao cliente, dados de pesquisa e econômicos, dados demográficos e geográficos, dados gerados por máquina e na web, etc.
2. **Transformação de dados:** os dados precisam ser transformados por meio de processamento preciso para obter dados normalizados. Os valores são escalados em uma faixa fornecida de dados normalizados, elementos estranhos são removidos por análise de correlação para concluir a decisão final.

3. **Formulação do modelo preditivo:** qualquer modelo preditivo frequentemente emprega técnicas de regressão para projetar um modelo preditivo usando o algoritmo de classificação. Durante esse processo, os dados de teste são reconhecidos, as decisões de classificação são implementadas nos dados de teste para determinar o desempenho do modelo.
4. **Inferências ou conclusão:** por fim, são feitas inferências a partir do modelo, para isso é realizada a análise de cluster.

As ideias centrais por trás da formulação da modelagem preditiva são os dados que estão sendo gerados diariamente ou os dados históricos que podem conter as informações mais relevantes para os cenários de negócios atuais, a fim de obter o máximo de lucro com modelos adequados e previsões precisas. O processo de modelagem preditiva envolve a tarefa fundamental de extrair informações necessárias de dados estruturados ou não estruturados.

Com todos esses dados, diferentes ferramentas são componentes para extrair inferências e padrões, como as técnicas de aprendizado de máquina são necessárias para identificar tendências nos dados e o modelo de design que estima conclusões futuras. Existe uma variedade de algoritmos de aprendizagem de máquina que estão disponíveis para a implementação da modelagem preditiva, regressão linear e não linear, redes neurais, árvores de decisão, etc. De acordo com Wakefield (2021), os modelos preditivos mais usados são:

Árvores de decisão: as árvores de decisão são uma forma simples, mas poderosa de análise de múltiplas variáveis. Elas são produzidas por algoritmos que identificam várias maneiras de dividir dados em segmentos semelhantes a ramificações. As árvores de decisão particionam os dados em subconjuntos com base em categorias de variáveis de entrada, ajudando você a entender o caminho de decisões de alguém.

Régressão (linear e logística): a regressão é um dos métodos mais populares em estatística. A análise de regressão estima as relações entre as variáveis, encontrando os principais padrões em grandes e diversos conjuntos de dados e como eles se relacionam entre si.

Redes neurais: padronizadas após a operação de neurônios no cérebro humano, as redes neurais (também chamadas de redes neurais artificiais) são uma variedade de tecnologias de aprendizado profundo. Elas são normalmente usadas para resolver problemas complexos de reconhecimento de padrões – e são incrivelmente úteis para analisar grandes conjuntos de dados. São ótimas para lidar com relacionamentos não lineares em dados – e funcionam bem quando certas variáveis são desconhecidas

Outros classificadores: os algoritmos de série temporal representam dados sequencialmente e são úteis para prever valores contínuos ao longo do tempo. Os algoritmos de *clustering*

organizam os dados em grupos cujos membros são semelhantes. Já os algoritmos de detecção de valores discrepantes se concentram na detecção de anomalias, identificando itens, eventos ou observações que não estão em conformidade com um padrão ou padrão esperado em um conjunto de dados. Os modelos Ensemble usam vários algoritmos de aprendizado de máquina para obter melhor desempenho preditivo do que o que poderia ser obtido com um único algoritmo. A análise fatorial é um método usado para descrever a variabilidade e visa encontrar variáveis latentes independentes. O classificador Naïve Bayes permite prever uma classe/categoria com base em determinado conjunto de características, usando a probabilidade. As máquinas de vetores de suporte são técnicas de aprendizado de máquina supervisionadas que usam algoritmos de aprendizado associados para analisar dados e reconhecer padrões. Cada classificador aborda os dados de uma maneira diferente, portanto, para que as organizações obtenham os resultados de que precisam, elas precisam escolher os classificadores e modelos corretos.

2.1.2 Construção de um modelo preditivo

Raschka (2014) apresenta que a modelagem preditiva é o conceito geral de construção de um modelo capaz de fazer previsões. Normalmente, esse modelo inclui um algoritmo de aprendizado de máquina que aprende certas propriedades de um conjunto de dados de treinamento para fazer essas previsões. A modelagem preditiva pode ser dividida em duas subáreas: **regressão e classificação de padrões**. Os modelos de regressão são baseados na análise de relações entre variáveis e tendências, a fim de fazer previsões sobre variáveis contínuas, por exemplo, a previsão da temperatura máxima para os próximos dias na previsão do tempo. Em contraste com os modelos de regressão, a tarefa da classificação de padrões é atribuir rótulos de classe discretos a observações particulares como resultados de uma previsão. Para voltar ao exemplo acima: uma tarefa de classificação de padrões na previsão do tempo pode ser a previsão de um dia ensolarado, chuvoso ou com neve.



As tarefas de classificação de padrões podem ser agrupadas em duas subcategorias principais: aprendizagem supervisionada e não supervisionada.

No aprendizado supervisionado, os rótulos de classe no conjunto de dados, que são usados para construir o modelo de classificação, são conhecidos. Por exemplo, um conjunto de dados para filtragem de spam conteria mensagens de spam, bem como mensagens sem spam. Em um problema de aprendizado supervisionado, se saberia qual mensagem no conjunto de treinamento é spam ou não e se usariam essas informações para treinar o modelo a fim de classificar novas mensagens não vistas.

Segundo Wakefield (2021), embora o aprendizado de máquina e a análise preditiva possam ser excelentes para qualquer organização, implementar essas soluções ao acaso, sem considerar como elas se encaixam nas operações diárias, prejudicará drasticamente sua capacidade de fornecer os *insights* de que a organização precisa. Para obter o máximo da análise preditiva e do aprendizado de máquina, as organizações precisam garantir que tenham a arquitetura para oferecer suporte a essas soluções, bem como dados de alta qualidade para alimentá-los e ajudá-los a aprender. A preparação e a qualidade dos dados são os principais capacitadores da análise preditiva. Os dados de entrada, que podem abranger várias plataformas e conter várias fontes de big data, devem ser centralizados, unificados e em um formato coerente.

Um algoritmo de AM preditivo é uma função que, dado um conjunto de exemplos rotulados, constrói um estimador. O rótulo ou etiqueta toma valores em um domínio conhecido. Se esse domínio for um conjunto de valores nominais, tem-se um problema de classificação, também conhecido como aprendizado de conceitos, e o estimador gerado é um classificador. Se o domínio for um conjunto infinito e ordenado de valores, tem-se um problema de regressão, que induz um regressor. Um classificador (ou regressor), por sua vez, também é uma função, que, dado um exemplo não rotulado, atribui esse exemplo a uma das possíveis classes (GAMA et al., 2011, p. 53).

Wakefield (2021) afirma que, para conseguir isso, as organizações devem desenvolver um programa de governança de dados sólido para policiar o gerenciamento geral dos dados e garantir que apenas dados de alta qualidade sejam capturados e registrados. Além disso, os processos existentes precisarão ser alterados para incluir análises preditivas e aprendizado de máquina, pois isso permitirá que as organizações impulsionem a eficiência em todos os pontos do negócio. Por fim, as organizações precisam saber quais problemas procuram resolver, pois isso as ajudará a determinar o modelo melhor e mais aplicável a ser usado.

Normalmente, os cientistas de dados e especialistas em TI de uma organização têm a tarefa de desenvolver a escolha dos modelos preditivos corretos – ou construir seus próprios modelos para atender às necessidades da organização.

A primeira etapa no desenvolvimento de um modelo preditivo, ao usar a análise de regressão tradicional, é selecionar o candidato relevante e as variáveis preditoras para a uma possível

inclusão no modelo. No entanto, não há consenso sobre a melhor estratégia a utilizar. Uma abordagem de eliminação reversa começa com todas as variáveis candidatas, e os testes de hipótese são aplicados sequencialmente para determinar quais variáveis devem ser removidas do modelo final, enquanto uma abordagem de modelo completo inclui todos os candidatos e as variáveis para evitar o *overfitting* e o viés de seleção (ROYSTON *et al.*, 2009).



De acordo com Greenland (1989), as variáveis preditoras significativas devem ser incluídas normalmente no modelo final, independentemente de sua significância estatística. Mas o número de variáveis incluídas é geralmente limitado pelo tamanho da amostra do conjunto de dados. A seleção inadequada de variáveis é uma das causas comuns de baixo desempenho do modelo nessa situação.

A seleção de variável é um dos problemas ao utilizar as técnicas de aprendizado de máquina, uma vez que muitas vezes não se baseiam apenas em hipóteses predefinidas. São várias outras questões importantes relacionadas ao gerenciamento de dados ao desenvolver um modelo preditivo, como lidar com dados perdidos e com a transformação de variável.

2.1.2.1 Aplicativos de análise preditiva e aprendizado de máquina

Para organizações que transbordam de dados, mas que lutam para transformá-los em *insights* úteis, a análise preditiva e o aprendizado de máquina podem fornecer a solução. Não importa quantos dados uma organização tenha, se ela não puder usar esses dados para aprimorar os processos internos e externos e atender aos objetivos, os dados se tornam um recurso inútil. A análise preditiva é mais comumente usada para segurança, marketing, operações, risco e detecção de fraude (WAKEFIELD, 2021).

Aqui estão apenas alguns exemplos de como a análise preditiva e o aprendizado de máquina são utilizados em diferentes setores:

Serviços bancários e financeiros: nesse setor, a análise preditiva e o aprendizado de máquina são usados em conjunto para detectar e reduzir fraudes, medir o risco de mercado, identificar oportunidades e muito, muito mais.

Segurança: as instituições de segurança geralmente usam análises preditivas para melhorar os serviços e o desempenho, mas também para detectar anomalias, fraudes, entender o comportamento do consumidor e aprimorar a segurança dos dados.

FIGURA 07 Representação da utilização dos princípios da análise preditiva no controle de estoque



Fonte: SHUTTERSTOCK.COM, 2021.

Varejo: os varejistas estão usando análises preditivas e aprendizado de máquina para entender melhor o comportamento do consumidor; quem compra o que e onde? Essas perguntas podem ser prontamente respondidas com os modelos preditivos e conjuntos de dados corretos, ajudando os varejistas a planejar com antecedência e estocar itens com base na sazonalidade e nas tendências do consumidor – melhorando significativamente o *Return over Investment*, ou Retorno sobre Investimento (ROI).

2.1.2.2 Modelos e algoritmos de análise preditiva

Segundo Parthasarathy (2021), as ferramentas de análise preditiva são acionadas por vários modelos e algoritmos diferentes que podem ser aplicados a uma ampla variedade de casos de uso. Determinar quais técnicas de modelagem preditiva são melhores para a empresa e/ou projeto é a chave para obter o máximo de uma solução de análise preditiva e aproveitar os dados para tomar decisões criteriosas.



Por exemplo, considere um varejista que procura reduzir a rotatividade de clientes. Eles podem não ser atendidos pelos mesmos modelos de análise preditiva usados por um hospital que prevê o volume de pacientes admitidos na sala de emergência nos próximos dez dias.

O modelo de classificação é, de certa forma, o mais simples dos vários tipos de modelos de análise preditiva, pois ele coloca os dados em categorias com base no que aprende com os dados históricos. Os modelos de classificação são melhores para responder sim ou não a perguntas, fornecendo uma análise ampla que é útil para orientar ações decisivas. A amplitude de possibilidades com o modelo de classificação – e a facilidade com que ele pode ser retrainado com novos dados – significa que pode ser aplicado a muitos setores diferentes. Já o modelo de armazenamento em cluster (conjunto de computadores que trabalham de forma paralela) classifica os dados em grupos inteligentes aninhados separados com base em atributos semelhantes. Se uma empresa de calçados de comércio eletrônico está procurando implementar campanhas de marketing direcionadas para seus clientes, ela pode passar por centenas de milhares de registros para criar uma estratégia personalizada para cada indivíduo. Mas este é o uso mais eficiente do tempo? Provavelmente não. Usando o modelo de *clustering*, a empresa pode separar rapidamente os clientes em grupos semelhantes com base em características comuns e desenvolver estratégias para cada grupo em uma escala maior (PARTHASARATHY, 2021).

Na aplicação de algoritmos de AM a problemas reais, em geral, o conhecimento que se tem do domínio sob investigação é provido unicamente pelo conjunto de exemplos, a partir do qual a indução de um modelo preditivo/descriptivo é então realizada. Os capítulos anteriores apresentaram várias técnicas de AM que podem ser utilizadas na indução de modelos de classificação e/ou de regressão a partir de um conjunto de exemplos rotulados. De maneira geral, pode-se afirmar que não existe técnica universal, ou seja, não é possível estabelecer a priori que uma técnica de AM em particular se sairá melhor na resolução de qualquer tipo de problema.

Em certos casos, as próprias características das técnicas existentes e do problema que está sendo solucionado podem ser consideradas para auxiliar na escolha da técnica a ser utilizada sobre um novo conjunto de dados. Por exemplo, em domínios em que os exemplos possuem alta dimensionalidade, as SVMs são boas candidatas, enquanto o algoritmo k-NN usando a distância euclidiana pode, a princípio, não parecer uma escolha adequada. Caso seja necessário que o modelo obtido seja interpretável, técnicas simbólicas, como as árvores de decisão, podem ser preferíveis a modelos “caixa-preta” como os gerados pelas RNAs e pelas SVMs (GAMA et al., 2011, p. 148).

Parthasarathy (2021) afirma que o modelo de previsão (um dos modelos de análise preditiva mais amplamente usados) lida com predição de valor métrico, estimando valor numérico para novos dados com base em aprendizados de dados históricos. Esse modelo pode ser aplicado onde quer que haja dados numéricos históricos disponíveis.



Os cenários incluem:

Um *call center* pode prever quantas chamadas de suporte receberá por hora.

Uma loja de calçados pode calcular quanto estoque deve manter em mãos para atender à demanda durante um determinado período de vendas.

De acordo com Parthasarathy (2021), o modelo de outliers (ou seja, dados que se diferenciam de todos os outros) é orientado em torno de entradas de dados anômalos em um conjunto de dados. Ele pode identificar figuras anômalas por si mesmas ou em conjunto com outros números e categorias. O modelo de outlier é particularmente útil para análises preditivas em varejo e finanças.



Por exemplo, ao identificar transações fraudulentas, o modelo pode avaliar não apenas a quantidade, mas também a localização, a hora, o histórico de compras e a natureza de uma compra.

O modelo de série temporal compreende uma sequência de pontos de dados capturados, usando o tempo como parâmetro de entrada. Por exemplo, os casos de uso para esse modelo incluem o número de ligações diárias recebidas nos últimos três meses, as vendas dos últimos 20 trimestres ou o número de pacientes que compareceram a um determinado hospital nas últimas seis semanas. É um meio potente de compreender a maneira como uma métrica singular está se desenvolvendo ao longo do tempo com um nível de precisão além das médias simples. Ele também leva em consideração as estações do ano ou eventos que podem impactar a métrica (PARTHASARATHY, 2021).



Se o proprietário de um salão deseja prever quantas pessoas provavelmente visitarão sua empresa, ele pode recorrer ao método bruto de calcular a média do número total de visitantes nos últimos 90 dias. No entanto, o crescimento nem sempre é estático ou linear, e o modelo de série temporal pode modelar melhor o crescimento exponencial e alinhar melhor o modelo à tendência de uma empresa. Ele também pode prever vários projetos ou várias regiões ao mesmo tempo, em vez de apenas um de cada vez.

FIGURA 08 Representação de um proprietário de salão de beleza utilizando os recursos da aprendizagem de máquina



Fonte: SHUTTERSTOCK.COM, 2021.

2.1.3 Validação de um modelo preditivo

Existem várias técnicas diferentes para desenvolver algoritmos preditivos, usando uma variedade de métodos analíticos de previsão com diversas ferramentas e/ou softwares. Alguns exemplos incluem redes neurais, que suportam máquinas de vetores e árvores de decisão. As árvores de decisão, por exemplo, utilizam técnicas, como classificação e árvores de regressão, para prever vários resultados. Modelos preditivos usando algoritmos de aprendizado de máquina podem, portanto, facilitar o reconhecimento de risco clinicamente

importante e variáveis em pacientes com vários fatores de risco iniciais que podem não ser identificados de outra forma.

FIGURA 09 Representação da utilização de modelos preditivos na medicina



Fonte: SHUTTERSTOCK.COM, 2021.

Para um modelo de previsão ser valioso, ele não deve apenas ter a capacidade preditiva na derivação, mas também deve realizar a validação de forma adequada (*validation cohort*). O desempenho de um modelo pode diferir substancialmente entre *cohort* (grupo de um conjunto de dados) de derivação e validação por vários motivos, incluindo o *overfitting* (sobreajuste) do modelo, falta de importantes variáveis preditoras, variabilidade de preditores que levam a erros de medição e diferenças na combinação de casos de cohort. Portanto, o desempenho do modelo na cohort de derivação pode ser excessivamente otimista e não é uma garantia de que o modelo funcionará igualmente bem. A validação pode ser realizada usando os recursos internos ou externos (ALTMAN et al., 2009).

De acordo com Buskirk et al. (2018), em comparação com os métodos estatísticos tradicionais, as técnicas de aprendizado de máquina são mais propensas a realizar o *overfitting* dos dados, ou seja, a detectar padrões que podem não se generalizar para outros dados. O desenvolvimento do modelo em aprendizado de máquina, portanto, geralmente depende da chamada validação cruzada como um método para reduzir o risco de *overfitting*. A validação cruzada pode ser implementada de maneiras diferentes, mas a ideia geral é

usar uma subamostra dos dados, conhecida como amostra de treinamento ou estimativa, para desenvolver um modelo preditivo. A amostra restante, não incluída na subamostra de treinamento, é referida como um teste ou amostra de validação e é usada para avaliar a precisão do modelo preditivo desenvolvido usando a amostra de treinamento.

Algumas técnicas de aprendizado de máquina usam uma terceira subamostra para fins de ajuste, ou seja, a amostra de validação, para encontrar os parâmetros de ajuste que geram a previsão mais ideal. Nesses casos, uma vez que um modelo foi construído usando a amostra de treinamento e refinado usando a amostra de validação, seu desempenho geral é então avaliado usando a amostra de teste. A precisão preditiva para algoritmos de aprendizado de máquina aplicados a resultados contínuos (por exemplo, problemas de regressão) é geralmente quantificada usando uma estatística de erro quadrático médio que compara o valor observado do resultado a um valor previsto (BUSKIRK *et al.*, 2018).

Em diversos casos, tem-se apenas um conjunto com n objetos, o qual deve ser empregado na indução do preditor e em sua avaliação. Calcular o desempenho preditivo – em termos de taxa de acerto ou de erro, por exemplo – do modelo nos mesmos objetos empregados em seu treinamento produz estimativas otimistas, uma vez que todos os algoritmos de AM tentam melhorar de alguma forma o seu desempenho preditivo nesses objetos durante a fase indutiva. O uso do mesmo conjunto de exemplos no treinamento e na avaliação do preditor é conhecido como ressubstituição. Em geral, o erro/acerto obtido nesse tipo de avaliação é denominado aparente.

Devem-se então utilizar métodos de amostragem alternativos para obter estimativas de desempenho preditivo mais confiáveis, definindo subconjuntos de treinamento e de teste. Os dados de treinamento são empregados na indução e no ajuste do modelo, enquanto os exemplos de teste simulam a apresentação de objetos novos ao preditor, os quais não foram vistos em sua indução. Esses subconjuntos são disjuntos para assegurar que as medidas de desempenho sejam obtidas a partir de um conjunto de exemplos diferente daquele usado no aprendizado (GAMA *et al.*, 2011, p. 150).

Em problemas de classificação, a precisão preditiva pode ser estimada usando uma série de estatísticas, incluindo: sensibilidade, especificidade e precisão geral. Geralmente, o cálculo dessas medidas relacionadas de precisão é baseado em uma matriz de confusão, que é simplesmente uma tabela de tabulação cruzada com as linhas denotando o valor real da variável de destino para cada amostra ou caso no conjunto de teste e as colunas que representam os valores do nível previsto da variável de destino para cada amostra ou caso no conjunto de teste (BUSKIRK *et al.*, 2018).

2.1.3.1 Validando um algoritmo de aprendizado de máquina na área médica

De acordo com Kakarmath *et al.* (2018), as soluções de big data, particularmente algoritmos preditivos de aprendizado de máquina, demonstraram a capacidade de desbloquear o valor dos dados em tempo real em muitas configurações fora da área de saúde. Os métodos de aprendizado de máquina podem ser usados para construir modelos preditivos flexíveis, personalizados e automatizados para otimizar a alocação de recursos e melhorar a eficiência e a qualidade da assistência médica. No entanto, esses modelos estão sujeitos a problemas de sobreajuste, confusão e deterioração no desempenho preditivo ao longo do tempo. É, portanto, necessário avaliar modelos preditivos baseados em aprendizado de máquina em um conjunto de dados independentes antes que eles possam ser adotados na prática clínica.

Kakarmath *et al.* (2018) apontam que o setor de saúde é geralmente considerado um adotante tardio de soluções de big data. Nos Estados Unidos, pelo menos um dos motivos que contribuíram para esse atraso é a adoção relativamente baixa de prontuários eletrônicos (Electronic Medical Records – EMRs) entre os hospitais. Em 2008, o número de hospitais que possuíam um sistema EMR básico era de 9%; em 2015, esse número cresceu para 96%. O rápido crescimento na adoção de EMR, juntamente com a mudança no sistema de saúde dos EUA de uma estrutura de reembolso baseada em volume para uma baseada em valor, estimulou investimentos em soluções baseadas em Inteligência Artificial (IA) para problemas de saúde.

A taxa de readmissão hospitalar é uma das métricas utilizadas para medir a qualidade da assistência prestada por um hospital. Naturalmente, os hospitais começaram a implementar várias intervenções para reduzir as taxas de readmissão. Para otimizar o uso de intervenções dispendiosas de transição de cuidados, uma das estratégias adotadas pelos hospitais tem sido focar nos pacientes com maior risco de readmissão. A estratificação de pacientes internados com base no risco de readmissão pode oferecer aos profissionais de saúde uma visão valiosa para modificar as intervenções, como o planejamento de alta e a oportunidade de influenciar os resultados por meio do gerenciamento proativo de pacientes de alto risco (KAKARMATH *et al.*, 2018).

Modelos de previsão de risco de readmissão hospitalar têm sido tradicionalmente desenvolvidos usando métodos estatísticos baseados em hipóteses desde a década de 1980. A partir de 2015, pelo menos 94 modelos únicos foram descritos na literatura publicada. Embora esses modelos de previsão de risco sejam ferramentas úteis para a tomada de decisão, sua utilidade é limitada por considerações de generalização, adaptabilidade e desempenho preditivo absoluto. Em primeiro lugar, a maioria desses modelos foi desenvolvida usando dados de alta qualidade de pacientes selecionados e, portanto, podem

ter validade externa inconsistente em outras configurações e populações de pacientes, na configuração de dados ausentes e ao longo do tempo. Em segundo lugar, esses modelos exigem que o pessoal de saúde calcule a pontuação de risco para cada paciente, criando barreiras para sua adoção. Finalmente, esses modelos muitas vezes não podem ser adaptados para incorporar informações que podem ter valor preditivo em diferentes populações de pacientes, resultando em desempenho preditivo abaixo do ideal. Em contraste, métodos analíticos de aprendizado de máquina podem ser usados para construir modelos preditivos flexíveis, customizados e automatizados usando as informações disponíveis em EMRs. A promessa de extrair *insights* preditivos em tempo real de dados EMR complexos e volumosos alimentou muito entusiasmo em torno da aplicação de métodos preditivos baseados em aprendizado de máquina em saúde, onde até mesmo um aumento marginal no desempenho pode se traduzir em ganhos significativos de eficiência e qualidade (KAKARMATH *et al.*, 2018).

Modelos preditivos desenvolvidos a partir de dados EMR usando métodos de aprendizado de máquina têm sua própria parcela de desafios de generalização. Primeiro, os modelos que são desenvolvidos usando um grande número de preditores relativos ao número de eventos de resultado são propensos a sobreajuste. Um exemplo bem conhecido disso é o Google Flu Trends, que previu o dobro do número real de consultas médicas relacionadas à influenza em 2013. Em segundo lugar, os modelos desenvolvidos usando dados EMR estão sujeitos a vieses resultantes da autosseleção do paciente, confundindo por indicação e disponibilidade inconsistente de dados de resultados. Finalmente, a própria prática da medicina evolui, impactando a precisão das previsões ao longo do tempo. Dadas essas limitações, é necessário validar o desempenho preditivo de modelos baseados em aprendizado de máquina em um conjunto de dados independentes antes que ele possa ser adotado na prática clínica (KAKARMATH *et al.*, 2018).

Um modelo baseado em aprendizado de máquina para prever reinternações de 30 dias em pacientes com insuficiência cardíaca foi desenvolvido no **Partners HealthCare System** (Boston, MA, EUA) em colaboração com **Hitachi Ltd** (Tóquio, Japão). Esse modelo foi treinado usando dados de prontuários longitudinais não identificados de 11.510 pacientes com insuficiência cardíaca que receberam alta com vida após uma internação nos anos entre 2014 e 2015. Houve 27.334 internações e 6.369 reinternações de 30 dias durante esse período. O modelo final incluiu 3.512 variáveis compreendendo dados demográficos, encontro, diagnóstico, procedimento, medicação e informações laboratoriais, bem como extratos selecionados de notas de visitas ambulatoriais e resumos de alta. Redes unificadas profundas – uma nova estrutura de rede em malha de aprendizado profundo com conexões verticais e horizontais de neurônios para evitar *overfitting* – foram usadas para desenvolver o modelo de previsão de risco. A validação cruzada de dez vezes foi usada para validar o modelo internamente. O modelo apresentou capacidade discriminativa moderada com estatística de concordância de 0,71 (KAKARMATH *et al.*, 2018).

Segundo Kakarmath *et al.* (2018), esse estudo (desenvolvido por **Partners HealthCare System**) descreve o protocolo para validação independente e prospectiva desse modelo baseado em aprendizado de máquina, treinado para prever o risco de readmissão em 30 dias em pacientes com insuficiência cardíaca. O principal objetivo desse estudo é validar prospectivamente um modelo preditivo baseado em aprendizado de máquina para internações em pacientes com insuficiência cardíaca, comparando suas previsões de risco para reinternações de 30 dias com resultados observados prospectivamente em uma coorte independente de pacientes.

2.1.3.2 Validando um algoritmo de aprendizado de máquina na área financeira

Desde o início do mercado de capitais, os investidores tentaram obter vantagem competitiva sobre outro mercado e ser capazes de prever com precisão as séries temporais. Dado o crescimento das fontes de dados disponíveis e o aumento da interconectividade dos investidores, a tomada de decisões está se tornando mais importante do que nunca. Os algoritmos de aprendizado de máquina oferecem recursos de aproximação e funções não lineares, lidando com dados ruidosos e não estacionários e descobrindo padrões latentes em conjuntos de dados.

As ferramentas de pontuação de crédito que usam aprendizado de máquina são projetadas para acelerar as decisões de empréstimo, enquanto potencialmente limitam o risco incremental. Os credores há muito confiam nas pontuações de crédito para fazer os empréstimos e tomar as decisões para empresas e clientes de varejo. Dados financeiros sobre transações e histórico de pagamentos das instituições serviram historicamente como a base da maioria dos modelos de pontuação de crédito. Esses modelos utilizam ferramentas, como regressão, árvores de decisão e análise estatística, para gerar uma pontuação de crédito usando quantidades limitadas de dados estruturados. No entanto, os bancos e outros credores estão cada vez mais recorrendo a fontes de dados adicionais, não estruturadas e semiestruturadas, incluindo a atividade nas mídias sociais, uso do telefone celular e atividade de mensagens de texto, para capturar uma visão mais matizada para melhorar a precisão da classificação dos empréstimos.

De acordo com Financial Stability Board (2017), aplicar o aprendizado de máquina para essa constelação de novos dados permitiu a avaliação de fatores qualitativos, como o comportamento de consumo e a disposição a pagar. A capacidade de aproveitar dados adicionais sobre tais medidas permite uma segmentação maior, mais rápida e mais barata da qualidade do mutuário. Em última análise, leva a uma decisão de crédito mais rápida. Além de facilitar uma avaliação segmentada e potencialmente mais precisa da qualidade

de crédito, o uso de algoritmos de aprendizado de máquina na pontuação de crédito pode ajudar a permitir maior acesso ao crédito. O mutuário com boa capacidade de crédito, muitas vezes, não consegue obter crédito e construir um histórico de crédito. Com o uso de fontes de dados alternativas e a aplicação de algoritmos de aprendizado de máquina para ajudar a desenvolver uma avaliação da capacidade e disposição para reembolsar, os credores podem ser capazes de chegar ao crédito que anteriormente seria impossível.

Existem várias vantagens e desvantagens em usar IA em modelos de pontuação de crédito. Os princípios da inteligência artificial permitem que grandes quantidades de dados sejam analisadas muito rapidamente. Um exemplo da aplicação de big data para pontuação de crédito pode incluir a avaliação de pagamentos de contas que não são de crédito, como o pagamento pontual de telefone celular e outras contas de serviços públicos, em combinação com outros dados. A disponibilidade de dados históricos em uma gama de mutuários e produtos de empréstimo é a chave para o desempenho dessas ferramentas. Da mesma forma, a disponibilidade, qualidade e confiabilidade dos dados sobre o desempenho do produto do mutuário em uma ampla gama de circunstâncias financeiras também é fundamental para o desempenho desses modelos de risco (FINANCIAL STABILITY BOARD, 2017).



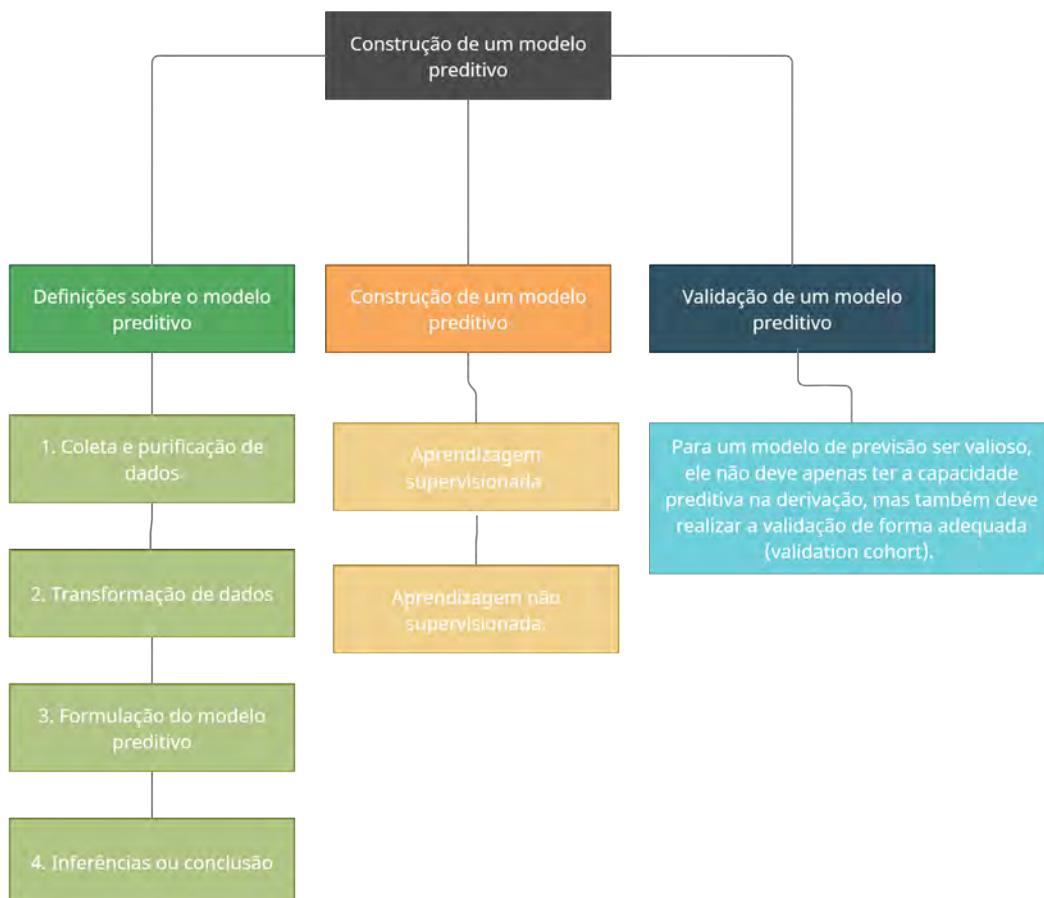
Os modelos de previsão de defeitos ajudam as equipes de garantia de qualidade de *software* a efetivamente concentrar seus recursos limitados nos módulos de *software* mais sujeitos a defeitos. Esses modelos de previsão de defeitos são normalmente treinados usando métricas extraídas de dados históricos de desenvolvimento que é registrado em repositórios de *software*. Os modelos de previsão podem fornecer uma estimativa otimista irreal do desempenho do modelo quando (re)aplicados para a mesma amostra com a qual foram treinados. Para resolver esse problema, técnicas de validação de modelo são usadas para estimar a atuação do modelo.

Bibliografia comentada

No livro Artificial intelligence: a modern approach, dos autores Norvig e Russell, é apresentada uma introdução abrangente e atualizada da teoria e prática da Inteligência Artificial (IA). A obra explora toda a amplitude e profundidade do campo da IA, abordando de forma atualizada sobre as tecnologias mais recentes, apresenta conceitos de uma maneira unificada e oferece cobertura nova ou expandida de aprendizado de máquina, aprendizado profundo, aprendizado de transferência, sistemas multiagentes, robótica, processamento de linguagem natural, causalidade, programação probabilística, privacidade e segurança dos projetos que utilizam os princípios de IA.

NORVIG, Peter; RUSSELL, Stuart. Artificial intelligence: a modern approach. 4. ed. Pearson, 2020.

Mapa conceitual



Conclusão

A unidade começou discutindo que o modelo preditivo é a base da manutenção preditiva. A aplicação de manutenção preditiva e os modelos de prognóstico aos modernos sistemas industriais variam de eletrônica, aeronáutica, automotivo para maquinário industrial e mais aplicações são implementadas todos os dias.

Em seguida, foi mostrado que a análise preditiva e o aprendizado de máquina andam de mãos dadas, já que os modelos preditivos geralmente incluem um algoritmo de aprendizado de máquina. Esses modelos podem ser treinados ao longo do tempo para responder a novos dados ou valores, entregando os resultados de que o negócio precisa. A modelagem preditiva se sobrepõe amplamente ao campo do aprendizado de máquina.

Você viu também que existem várias técnicas diferentes para desenvolver algoritmos preditivos, usando uma variedade de métodos analíticos de previsão com diversas ferramentas e/ou softwares. Alguns exemplos incluem redes neurais, que suportam máquinas de vetores e árvores de decisão. As árvores de decisão, por exemplo, utilizam técnicas, como classificação e árvores de regressão, para prever vários resultados. Modelos preditivos usando algoritmos de aprendizado de máquina podem, portanto, facilitar o reconhecimento de risco clinicamente importante e variáveis em pacientes com vários fatores de risco iniciais que podem não ser identificados de outra forma.

Por fim, cabe aqui uma recomendação importante: lembre-se de que apenas a prática o levará à fluência sobre o processo de construção de um modelo preditivo. Bons estudos! Nesta unidade veremos que a modelagem preditiva é o processo de ajustar ou treinar os parâmetros do modelo de desenvolvimento, usando um algoritmo de mineração de dados para ajustar um conjunto de instâncias do conceito da melhor maneira possível.

VISUALIZAÇÃO E APRESENTAÇÃO DOS RESULTADOS DE UM MODELO PREDITIVO

OBJETIVOS

Ao final desta unidade, esperamos que possa:

Descrever o processo de visualização de um modelo preditivo.

Descrever a apresentação dos resultados de um modelo preditivo.

Aplicar os princípios do modelo preditivo.



3

Unidade 3:
Visualização e Apresentação dos Resultados de um Modelo Preditivo

Apresentação

Nesta unidade veremos que a modelagem preditiva é o processo de ajustar ou treinar os parâmetros do modelo de desenvolvimento, usando um algoritmo de mineração de dados para ajustar um conjunto de instâncias do conceito da melhor maneira possível.

Discutiremos que podem ser necessárias técnicas especializadas para visualização de dados durante o trabalho com conjuntos de dados muito grandes, como costumamos fazer na análise preditiva. Veremos nesta unidade que as técnicas de transparência parcial podem ajudar e que gráficos hexbin são geralmente melhores do que gráficos de dispersão para mostrar os relacionamentos entre as variáveis.

Deste modo, nesta unidade você vai estudar sobre os princípios da visualização de um modelo preditivo e a apresentação dos resultados de um modelo preditivo, e refletir sobre o que é possível fazer a partir do modelo preditivo. Bons estudos!

3.1 Visualização e apresentação dos resultados de um modelo preditivo

A modelagem preditiva é a tarefa de construir um modelo de conceito que expressa a variável-alvo em função das variáveis explicativas (ou seja, a partir da entrada de dados, o modelo deve verificar e encontrar a resposta para o processamento pretendido). O objetivo da modelagem preditiva é minimizar a diferença entre os valores previstos e reais. Uma representação de modelo consiste em um conjunto de parâmetros (atributos, operadores e constantes) organizadas em algum tipo de estrutura.

Uma grande quantidade de informações úteis pode ser extraída de um conjunto de dados por meio de sua análise ou exploração. Informações obtidas na exploração podem ajudar, por exemplo, na seleção da técnica mais apropriada para pré-processamento dos dados e para aprendizado. Uma das formas mais simples de explorar um conjunto de dados é a extração de medidas de uma área da estatística denominada estatística descritiva. A estatística descritiva resume de forma quantitativa as principais características de um conjunto de dados. Muitas dessas medidas são calculadas rapidamente. Por exemplo, no conjunto de dados de pacientes, duas medidas estatísticas podem ser facilmente calculadas: a idade média dos pacientes e a porcentagem de pacientes do sexo masculino. As medidas da estatística descritiva

assumem que os dados são gerados por um processo estatístico. Como o processo pode ser caracterizado por vários parâmetros, as medidas podem ser vistas como estimativas dos parâmetros estatísticos da distribuição que gerou os dados. Por exemplo, os dados podem ter sido gerados por uma distribuição normal com média 0 e variância 1. Essas medidas permitem capturar informações como:

- Frequência;
- Localização ou tendência central (por exemplo, a média);
- Dispersão ou espalhamento (por exemplo, o desvio padrão);
- Distribuição ou formato (GAMA et al., 2011, p. 15).

As instâncias que são usadas para construir o modelo são, consequentemente, chamadas de **conjunto de treinamento**. Um modelo pode ter uma estrutura estática predefinida ou pode ser desenvolvida dinamicamente durante o treinamento. O quanto bem um modelo se encaixa em um conjunto de treinamento específico é definido como a diferença ou o erro entre os valores previstos e o valor real, e é calculado para o treinamento definido pelo uso de uma métrica de erro. Quando o erro de treinamento é suficientemente pequeno, o treinamento para o modelo pode ser usado para fazer previsões para dados novos invisíveis. Normalmente, algumas das instâncias de treinamento são reservadas antes da realização do treinamento, criando assim um conjunto de teste que é usado para avaliar a qualidade do modelo e verificar se funcionará em dados novos (KÖNIG, 2008).

De acordo com Bari et al. (2021), para testar o modelo de análise preditiva que estamos construindo, devemos dividir o conjunto de dados em dois grupos: **conjuntos de dados de treinamento e teste**. Esses conjuntos de dados devem ser selecionados aleatoriamente e possuir uma representação da população real. Os dados semelhantes devem ser usados para os conjuntos de dados de treinamento e teste. Normalmente, o conjunto de dados de treinamento é significativamente maior do que o conjunto de dados de teste.

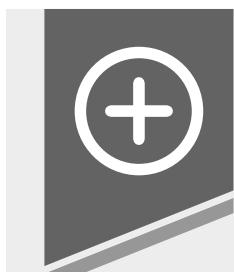
Utilizar o conjunto de dados de teste ajuda a evitar erros, como o *overfitting*. O modelo treinado é executado em dados de teste para ver como o modelo será executado. Alguns cientistas de dados preferem ter um terceiro conjunto de dados com características semelhantes às dos dois primeiros: um **conjunto de dados de validação**. A ideia é que, se você estiver usando ativamente seus dados de teste para refiná-lo, deverá usar um (terceiro) conjunto separado para verificar a precisão do modelo. Ter um conjunto de dados de validação, que não foi usado como parte do processo de desenvolvimento do modelo, ajuda a garantir uma estimativa neutra da precisão e eficácia do modelo.

FIGURA 10 Representação de um conjunto de treinamento


Fonte: SHUTTERSTOCK.COM, 2021.

3.1.1 Visualização de um modelo preditivo

De acordo com König (2008), um modelo treinado nunca é melhor do que os dados usados para o treinamento, portanto, é importante garantir que os dados corretos estejam disponíveis antes do início da modelagem. Se o domínio não é familiar, é importante consultar especialistas de domínio para garantir que nenhum atributo potencialmente importante esteja ausente, uma vez que nenhuma técnica pode modelar relacionamentos que não estão presentes nos dados. Em geral, é melhor incluir mais dados do que menos, pois as técnicas de mineração de dados devem ser capazes de encontrar as relações importantes e desconsiderar atributos supérfluos.



Um conjunto de dados contém um conjunto de instâncias em que cada instância consiste em valores específicos dos atributos pensados para descrever o alvo. A maior parte dos algoritmos de mineração de dados é baseada na suposição de que nenhum dado está faltando e exige que os valores ausentes sejam removidos antes da modelagem.

De acordo com Miller (2014), podem ser necessárias técnicas especializadas para visualização de dados durante o trabalho com conjuntos de dados muito grandes, como costumamos fazer na análise preditiva. Como dito na apresentação desta unidade, os gráficos hexbin são geralmente melhores do que gráficos de dispersão para mostrar os relacionamentos entre as variáveis, pois as técnicas de transparência parcial podem ajudar na execução de todo o processo. Seguem alguns princípios fundamentais que devemos levar em consideração a partir desse contexto:

- **Finding out about (Descobrindo sobre):** Esta é a primeira coisa que fazemos – pesquisar por informações para descobrir o que outros fizeram antes, aprendendo com a literatura. É importante basear-se em trabalhos de acadêmicos e profissionais em muitos campos de estudo, que contribuem para análises preditivas e ciência de dados.
- **Looking at data (Olhando para os dados):** Começamos cada projeto de modelagem a partir da análise de dados e visualização de dados para fins de descoberta. Preparamos os dados para análises posteriores.
- **Predicting how Much (O que prever):** Muitas vezes somos solicitados a prever quantas unidades do produto serão vendidos, o preço dos títulos financeiros ou imobiliários, entre outros. As técnicas de regressão são úteis para fazer essas previsões.
- **Predicting yes or no (Predizendo sim ou não):** Muitos problemas de negócios são problemas de classificação, como prever se uma pessoa vai comprar ou não, observar a inadimplência em um empréstimo ou não, clicar em uma página da Web ou não. Utilizamos as técnicas de classificação na previsão do comportamento do consumidor e na classificação do texto.
- **Testing it out (Testando):** Examinamos modelos com gráficos de diagnóstico. Neste ponto, observamos quão bem um modelo desenvolvido em um conjunto de dados funciona em outros conjuntos de dados.
- **Playing what-if (Jogando):** Podemos precisar manipular variáveis-chave para ver o que acontece com nossas previsões. Na gestão de operações, empregamos os testes de sensibilidade de modelos de programação matemática, pois dessa forma é possível observar como os novos valores para variáveis de entrada afetam os resultados ou recompensas.
- **Explaining it all (Explicando tudo):** Dados e modelos nos ajudam a entender o mundo. Transformamos o que aprendemos em uma explicação que todos possam

compreender e apresentamos os resultados do projeto de forma clara e concisa. Essas apresentações se beneficiam de visualizações de dados bem construídas (design com um propósito).

3.1.2 Apresentação dos resultados de um modelo preditivo

De acordo com Brownlee (2014), depois de encontrar e ajustar um modelo viável de seu problema, é hora de fazer uso desse modelo. Você pode precisar rever o porquê e se lembrar de que forma precisa de uma solução para o problema que está resolvendo, que não é resolvido até que faça algo com os resultados. Dependendo do tipo de problema que está tentando resolver, a apresentação dos resultados será muito diferente.

Quando dados a serem utilizados em uma aplicação de AM são oriundos de diferentes fontes, estando organizados em diferentes conjuntos de dados, esses conjuntos devem ser integrados antes do início do uso da técnica de AM. Nesse caso, é possível que cada conjunto de dados represente diferentes atributos de um mesmo grupo de objetos. Assim, na integração, é necessário identificar quais são os objetos que estão presentes nos diferentes conjuntos a serem combinados. Esse problema é conhecido como problema de identificação de entidade. Essa identificação é realizada por meio da busca por atributos comuns nos conjuntos a serem combinados. Como exemplo, conjuntos de dados médicos podem ter um atributo que identifica o paciente. Assim, os objetos dos diferentes conjuntos que possuem o mesmo valor para o atributo que identifica o paciente são combinados em um único objeto do conjunto integrado. É fácil ver que o(s) atributo(s) utilizado(s) para combinação deve(m) ter um valor único para cada objeto. Alguns aspectos podem dificultar a integração. Por exemplo, atributos correspondentes podem ter nomes diferentes em bases de dados distintas. Além disso, os dados a serem integrados podem ter sido atualizados em momentos diferentes. Para minimizar esses problemas, é comum o uso de metadados em bases de dados. Os metadados são dados sobre dados que, ao descrever as suas principais características, podem ser utilizados para evitar erros no processo de integração. O processo de integração origina um depósito ou repositório de dados (data warehouse), que funciona como uma base de dados centralizada (GAMA et al., 2011, p. 29).

FIGURA 11 Representação do processo de apresentação dos resultados de um modelo preditivo



Fonte: SHUTTERSTOCK.COM, 2021.

De acordo com Brownlee (2014), existem duas facetas principais para usar os resultados de seu esforço de aprendizado de máquina:

- Relate os resultados
- Operacionalização do sistema

Resultados do relatório

Brownlee (2014) afirma que, depois de descobrir um bom modelo e um resultado bom o suficiente (ou não, conforme o caso), você desejará resumir o que foi aprendido e apresentá-lo às partes interessadas. Pode ser você, um cliente ou a empresa para a qual trabalha. Utilize um modelo de PowerPoint e dirija-se às seções que devem ser listadas de forma resumida e objetiva. Seguem abaixo algumas seções que podem ser apresentadas nesse momento:

- **Contexto:** Defina o ambiente no qual o problema existe e defina a motivação para a pergunta de pesquisa.
- **Problema:** Descreva de forma concisa o problema (a partir de uma pergunta).
- **Solução (Resposta):** Descreva de forma concisa a solução como uma resposta à pergunta que você fez na seção anterior (seja específico).
- **Descobertas:** listas com marcadores de descobertas que você fez ao longo do caminho que interessam ao público. Elas podem ser descobertas nos dados, nos métodos que funcionaram ou não, ou nos benefícios de desempenho do modelo que você alcançou ao longo de sua jornada.
- **Limitações:** Considere onde o modelo não funciona ou questões que o modelo não responde. Não se intimide com essas questões, pois definir onde o modelo se destaca é mais confiável se você puder definir onde ele não se sobressai.
- **Conclusões:** Revisite o porquê, a pergunta de pesquisa e a resposta que você descobriu em um pacote compacto que seja fácil de lembrar e repetir para você e para os outros.

O tipo de público para o qual você está apresentando definirá a quantidade de detalhes em que você entrará. Ter disciplina para concluir seus projetos com um relatório de resultados, mesmo em pequenos projetos paralelos, acelerará seu aprendizado em campo.

Operacionalize

Brownlee (2014) afirma que, ao encontrar um modelo que é bom o suficiente para resolver o problema que você enfrenta, é fundamental desenvolver três aspectos principais para operacionalizar um modelo que você pode considerar cuidadosamente antes de colocar um sistema em produção. Essas três áreas nas quais você deve pensar cuidadosamente são a **implementação do algoritmo, o teste automatizado do seu modelo e o rastreamento do desempenho do modelo** ao longo do tempo. Essas três questões muito provavelmente influenciarão o tipo de modelo que você escolher.

- **Implementação de Algoritmo em bibliotecas:** Normalmente os desenvolvedores utilizam uma biblioteca de pesquisa para descobrir o método de melhor desempenho. As implementações de algoritmos em bibliotecas de pesquisa podem

ser excelentes, mas também podem ser escritas para o caso geral do problema, e não para o caso específico com o qual você está trabalhando. Considere localizar uma biblioteca de nível de produção que ofereça suporte ao método que você deseja usar. Pode ser necessário repetir o processo de ajuste do algoritmo se você alternar para uma biblioteca de nível de produção neste ponto.

- **Implementação de Algoritmo por conta própria:** Esta opção pode apresentar risco, dependendo da complexidade do algoritmo que você escolheu e dos truques de implementação que ele usa. Mesmo com o código-fonte aberto, pode haver uma série de operações complexas que podem ser muito difíceis de internalizar e reproduzir com segurança.
- **Teste automatizado do seu modelo:** Escreva testes automatizados que verifiquem se o modelo pode ser construído e atingir um nível mínimo de desempenho repetidamente. Também escreva testes para quaisquer etapas de preparação de dados. Você pode desejar controlar a aleatoriedade usada pelo algoritmo (sementes de números aleatórios) para cada teste de unidade executado para que os testes sejam 100% reproduzíveis.
- **Rastreamento do desempenho do modelo:** Adicione uma infraestrutura para monitorar o desempenho do modelo ao longo do tempo e acione alarmes se a precisão cair abaixo de um nível mínimo. O rastreamento pode ocorrer em tempo real ou com amostras de dados ao vivo em um modelo recriado em um ambiente separado. Um alarme disparado pode ser uma indicação de que a estrutura aprendida pelo modelo nos dados mudou (desvio de conceito) e que o modelo pode precisar ser atualizado ou ajustado.

De acordo os itens apresentados anteriormente, Brownlee (2014) afirma que existem alguns tipos de modelos que podem realizar o aprendizado *online* e se atualizar. Em algumas circunstâncias, pode ser mais sensato gerenciar o processo de atualização do modelo e alternar os modelos (sua configuração interna), pois eles são verificados para ter melhor desempenho.

Um projeto não é considerado concluído até que você entregue os resultados. Os resultados podem ser apresentados para você ou para seus clientes, por esse motivo é muito importante se atentar para as três preocupações ao usar um modelo aprendido em um ambiente de produção, especificamente a natureza da implementação do algoritmo, os testes do modelo e o rastreamento contínuo.

3.1.2.1 Práticas para a implantação do modelo preditivo

Ao adicionar algumas etapas antes da implantação e publicação do modelo preditivo, podemos construir a confiança de que o modelo que está implantando é sustentável e permanece preciso a longo prazo.

A qualidade das hipóteses induzidas pelos atuais sistemas de aprendizado de máquina supervisionado depende da quantidade dos exemplos no conjunto de treinamento. Por outro lado, muitos dos sistemas de aprendizado de máquina conhecidos não estão preparados para trabalhar com uma grande quantidade de exemplos. Grandes conjuntos de dados são tópicos em mineração de dados. Uma maneira para resolver este problema consiste em construir ensembles de classificadores. Um ensemble é um conjunto de classificadores cujas decisões de classificação são combinadas de alguma maneira para classificar um novo caso. Ensembles de classificadores são capazes de melhorar o poder de predição dos algoritmos de aprendizado; entretanto podem ser compostos por muitos classificadores, o que pode ser indesejável. Ainda, apesar de ensembles classificarem novos exemplos melhor que cada classificador individual, eles se comportam como caixas pretas, no sentido de não oferecer ao usuário alguma explicação relacionada à classificação por eles fornecida (BERNARDINI, 2006, p. 32).

De acordo com Brownlee (2016), podemos relacionar cinco etapas de melhores práticas que podemos executar ao implementar seu modelo preditivo na produção:

- 1.** Especificar os requisitos de desempenho.
- 2.** Separar o algoritmo de predição dos coeficientes do modelo.
- 3.** Desenvolver testes automatizados para seu modelo.
- 4.** Desenvolver infraestrutura de teste.
- 5.** Alterar e depois testar as atualizações do modelo.

Especificando os requisitos de desempenho:

Você precisa definir claramente o que constitui um bom e um mau desempenho. Isso pode ser devido a precisão, falsos positivos ou quaisquer métricas importantes para o negócio. Sem a definição correta do desempenho, você não poderá configurar os testes de que precisará para determinar se o sistema está se comportando conforme o esperado, por esse motivo, não prossiga até que você tenha concordado com o mínimo, a média ou uma expectativa de faixa de desempenho (BROWNLEE, 2016).

Separar o algoritmo de previsão dos coeficientes do modelo:

Você pode utilizar uma biblioteca para criar seu modelo preditivo (por exemplo, R, scikit-learn ou Weka), escolher implantar o modelo usando essa biblioteca ou reimplementar o aspecto preditivo do modelo em seu *software*. Você pode até mesmo querer configurar seu modelo como um serviço da web. Independentemente disso, é uma boa prática separar o algoritmo que faz previsões dos recursos internos do modelo. Esses são os coeficientes ou estrutura específicos dentro do modelo aprendidos com seus dados de treinamento (BROWNLEE, 2016). Frequentemente, a complexidade de um algoritmo de aprendizado de máquina está no treinamento do modelo, não em fazer previsões. Por exemplo, fazer previsões com um algoritmo de regressão é bastante direto e fácil de implementar no idioma de sua escolha. Este seria um exemplo de um algoritmo óbvio para reimplementar, em vez da biblioteca usada no treinamento do modelo. De acordo com Brownlee (2016), os números ou a estrutura aprendida pelo modelo podem ser chamados de coeficientes. Esses dados não são configurações para seu aplicativo. Armazene-o em um arquivo externo com o projeto de *software* e trate a configuração como código porque pode facilmente interromper seu projeto. Muito provavelmente você precisará atualizar esta configuração no futuro à medida que aprimora seu modelo.

Desenvolver testes automatizados para seu modelo:

Você precisa de testes automatizados para provar que seu modelo funciona conforme o esperado. É fundamental coletar uma pequena amostra de dados para fazer previsões, então utilize o código e a configuração do algoritmo de produção para fazer previsões, e confirme se os resultados são esperados no teste. Esses testes são seu alarme de alerta precoce. Se eles falharem, seu modelo está com problemas e você deve buscar a correção o mais rápido possível (BROWNLEE, 2016). De acordo com Brownlee (2016), é fundamental realizar os testes e observar se os requisitos mínimos de desempenho do modelo são atendidos. Para realizar esse processo de forma adequada, devemos executar os testes de regressão após cada alteração de código e antes de cada versão.

Desenvolver infraestrutura de teste:

O modelo mudará, assim como o *software* e os dados sobre os quais as previsões estão sendo feitas. Como desenvolvedores, desejamos automatizar a avaliação do modelo de produção com uma configuração especificada em um grande corpo de dados. Isso permitirá que você faça um teste eficiente das alterações no modelo em dados históricos e determine se você realmente fez uma melhoria ou não. Este não é o pequeno conjunto de dados que você pode usar para ajuste de hiperparâmetros, este é o conjunto completo de dados disponíveis, talvez particionado por mês, ano ou alguma outra demarcação importante (BROWNLEE, 2016). Brownlee (2016) afirma que devemos executar o modelo operacional atual

para o desempenho de linha de base e que após essa dinâmica é importante executar novos modelos, competindo por um lugar para entrar nas operações. Uma vez configurado, devemos executar todas as noites ou semanas e fazer com que ele gere relatórios automáticos.

Alterar e depois testar as atualizações do modelo:

Uma mudança menor e mais gerenciável seria nos coeficientes do modelo. Por exemplo, talvez você configure uma grade ou pesquisa aleatória de hiperparâmetros de modelo que é executada todas as noites e exibe novos modelos candidatos. Teste o modelo e seja altamente crítico. Avalie o desempenho do novo modelo, reveja os resultados cuidadosamente e avalie a mudança usando o teste de regressão, como uma verificação automatizada final (BROWNLEE, 2016). Brownlee (2016) afirma que devemos aceitar o novo modelo assim que estiver satisfeito de que ele atende aos requisitos mínimos de desempenho e melhora os resultados anteriores. Como uma catraca, considere a atualização incremental dos requisitos de desempenho à medida que o desempenho do modelo melhora.

3.1.3 O que é possível fazer a partir do modelo preditivo? Exemplos práticos

A indústria de seguros está usando aprendizado de máquina para analisar dados complexos para reduzir custos e melhorar a lucratividade. Desde a análise de dados para impulsionar a formação de preços, o núcleo do negócio de seguros, a tecnologia relacionada a seguros, muitas vezes depende da análise de *big data*. Muitos aplicativos envolvem melhorias para o processo de subscrição, auxiliando os agentes na classificação através de vastos conjuntos de dados que as seguradoras coletam para identificar casos que apresentam maior risco, reduzindo potencialmente os sinistros e melhorando a lucratividade. Algumas seguradoras estão usandoativamente o aprendizado de máquina para melhorar os preços ou marketing de produtos de seguro, incorporando dados altamente granulares em tempo real (FINANCIAL STABILITY BOARD, 2017).

A questão-chave em uma das primeiras definições de Aprendizado de Máquina, “*Self constructing or self-modifying representations of what is being experienced for possible future use*”, é a palavra self: sistemas que se modificam para melhor se acomodarem ao ambiente. O futuro da área de AM aponta para sistemas autônomos que podem incorporar o conhecimento, aprender a partir de dados não estacionários distribuídos em ambientes dinâmicos, com capacidade de transferir o conhecimento entre os problemas de aprendizado. O objetivo do aprendizado automático é a construção de modelos computacionais que descrevem sistemas complexos a partir da observação do comportamento do sistema. Nas últimas duas décadas, a investigação e a prática

do aprendizado automático centram-se no aprendizado a partir de conjuntos de dados relativamente pequenos, que podem ser carregados na totalidade em memória. Regra geral, os algoritmos percorrem o conjunto de treinamento várias vezes, de forma a compensar o reduzido número de observações (GAMA et al., 2011, p. 359).

As empresas geralmente têm acesso a esses dados por meio de parcerias, aquisições ou atividades não relacionadas a seguros. Com a utilização de IA e aplicativos de aprendizado de máquina, podemos observar um aumento substancial no setor de seguros, como a subscrição e o processamento de sinistros. Esses aplicativos podem aprender com um conjunto de treinamento de reivindicações anteriores para destacar as principais considerações para os tomadores de decisão. Técnicas de aprendizado de máquina podem ser usadas para determinar os custos de reparo e categorizar automaticamente a gravidade dos danos do acidente de veículo. Além disso, a IA pode ajudar a reduzir o tempo de processamento de sinistros e os custos operacionais. As seguradoras também estão explorando como IA, aprendizado de máquina e sensores remotos que podem detectar e, em alguns casos, prevenir incidentes seguráveis antes que ocorram, como o derramamento de produtos químicos ou acidentes de carro (FINANCIAL STABILITY BOARD, 2017).

FIGURA 12 Representação da utilização da aprendizagem de máquina na área médica



Fonte: SHUTTERSTOCK.COM, 2021.

Os *chatbots* são assistentes virtuais que ajudam os clientes a realizar transações ou resolver

problemas. Estes programas automatizados utilizam PNL para interagir com clientes em linguagem natural (por texto ou voz), utilizando algoritmos de aprendizado de máquina para melhorar ao longo do tempo. Os *chatbots* estão sendo introduzidos por uma série de empresas, muitas vezes em seus aplicativos móveis ou mídia social. De acordo com o Financial Stability Board (2017), a geração atual de *chatbots* em uso por empresas de serviços financeiros é simples e geralmente fornece informações de saldo ou alertas para os clientes ou respondendo a perguntas simples. Vale a pena observar que o uso crescente de *chatbots* está relacionado com o aumento do uso de aplicativos de mensagens.

FIGURA 13 Representação do funcionamento de um *chatbots*



Fonte: SHUTTERSTOCK.COM, 2021.

As empresas de *trading* estão buscando os recursos da IA e aprendizado de máquina para usar dados para melhorar sua capacidade de vender para os clientes. Por exemplo, analisar o comportamento de negociação passado pode ajudar a antecipar o próximo pedido. A negociação gera grandes quantidades de dados, e essa escala é normalmente exigida pelas ferramentas de aprendizagem de máquina para trabalhar de forma eficaz.

O Financial Stability Board (2017) apresenta que a IA e o aprendizado de máquina podem gerenciar de forma mais proativa as exposições a riscos. Para grandes empresas comerciais, como bancos, o uso de uma carteira central de negociação de risco ou técnicas de gestão de risco baseadas em análise de *big data* permitiu que essas empresas gerenciem riscos e otimizem seu uso de capital por centralizar os riscos que surgem de várias partes de seus negócios. A IA e o aprendizado de máquina podem ajudar na conformidade com as regulamentações comerciais.

FIGURA 14 Representação do funcionamento de um sistema de *trading*



Fonte: SHUTTERSTOCK.COM, 2021.

No gerenciamento de portfólio, ferramentas de IA e aprendizado de máquina estão sendo usadas para identificar novos sinais sobre os movimentos de preços e fazer um uso mais eficaz da vasta quantidade de dados disponíveis e pesquisa de mercado do que com os modelos atuais. As ferramentas de aprendizado de máquina funcionam com os mesmos princípios das técnicas analíticas existentes usadas em investimentos sistemáticos. A principal tarefa é identificar sinais a partir de dados sobre os quais as previsões relativas ao nível de preços ou volatilidade podem ser feitas ao longo de vários horizontes de tempo para gerar retornos maiores e não correlacionados (FINANCIAL STABILITY BOARD, 2017).

O desenvolvimento de novas tecnologias da informação e comunicações alterou drasticamente os processos de coleta, transformação e processamento de dados. O surgimento de novas tecnologias para redes de todos os tipos (nomeadamente, redes sem fios) e os avanços em miniaturização e tecnologia de sensores possibilitam a coleta de informação espaço-temporal nos mais diversos domínios, com um nível de detalhe antes impensável (GAMA et al., 2011, p. 359).

De acordo com o Financial Stability Board (2017), a proporção de negociações realizadas por fundos de quantia entre 2013 e 2016 aproximadamente dobrou de 13% para 27%. Por sua vez, parte da negociação é baseada no aprendizado de máquina. É difícil quantificar exatamente qual proporção utiliza o aprendizado de máquina por dois motivos principais:

- As empresas hesitam em compartilhar informações proprietárias.
- Quando as empresas compartilham informações sobre o uso do aprendizado de máquina, nem sempre há uma definição padrão ou compreensão do que está incluído no conceito.

Além do uso por gestores de fundos, empresas especializadas estão disponibilizando aos gestores de ativos ferramentas de aprendizado de máquina para obter insights do vasto volume de notícias e pesquisas de mercado. Em outros casos, os próprios gestores de ativos estão construindo indicadores, usando capacidade de IA fornecidos por terceiros.

Bibliografia Comentada

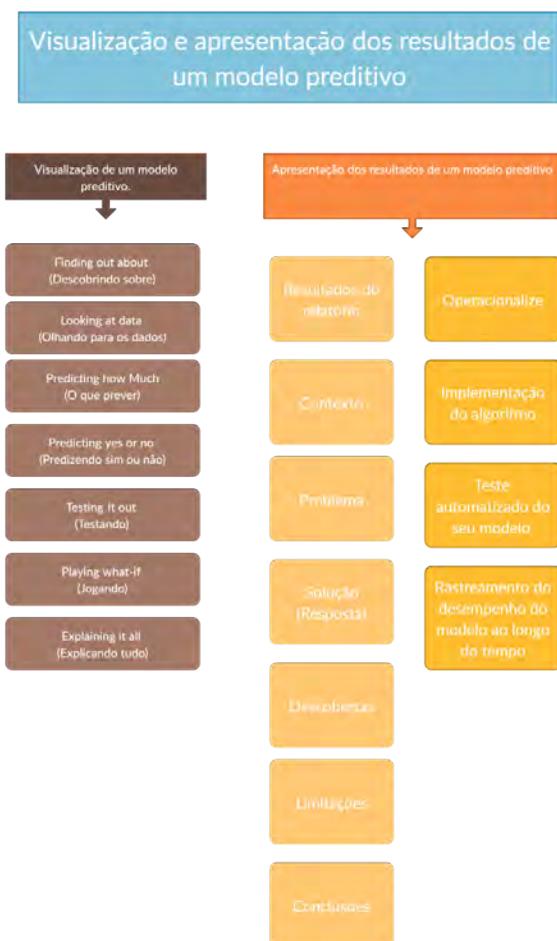
Veja a seguir uma indicação de leitura que poderá complementar seus estudos. Essa leitura não é obrigatória, porém poderá ampliar seu conhecimento em relação ao assunto abordado na unidade.

No livro “**Aprendizado de Máquina Para Leigos**”, dos autores **Mueller e Massaron**, lançado em 2019, são apresentados os princípios gerais sobre o desenvolvimento e a integração dos princípios de aprendizado de máquina. Na parte 5 da obra, intitulada “Aplicação de Aprendizado em problemas reais”, são apresentados Classificação de Imagens, Pontuação para Opiniões e Sentimentos, e processo de desenvolvimento de um sistema que aplique a aprendizagem de máquina para realizar a recomendação de produtos e filme.

Ao ler a obra será possível refletir sobre as possibilidades que temos para o desenvolvimento dos princípios da aprendizagem de máquina e do modelo preditivo.

MUELLER, John Paul; MASSARON, Luca. Aprendizado de Máquina Para Leigos. Alta Books, 2019.

Mapa conceitual



Conclusão

A modelagem preditiva é a tarefa de construir um modelo de conceito que expressa a variável-alvo em função das variáveis explicativas. O objetivo da modelagem preditiva é minimizar a diferença entre os valores previstos e reais. Uma representação de modelo consiste em um conjunto de parâmetros (atributos, operadores e constantes) organizados em algum tipo de estrutura.

Discutimos ao longo do material que é fundamental coletar uma pequena amostra de dados para fazer previsões. Esses testes são basicamente o nosso alarme de alerta precoce. Se eles falharem, o modelo está com problemas e você deve buscar a correção o mais rápido possível.

Ao longo do material foi possível observar que é fundamental realizar os testes e observar

se os requisitos mínimos de desempenho do modelo são atendidos. Para conseguir realizar esse processo de forma adequada, devemos executar os testes de regressão após cada alteração de código e antes de cada versão.

Vimos ainda o que é possível fazer a partir do modelo preditivo, com exemplos práticos. A análise preditiva prevê resultados futuros com base em dados históricos e atuais. Ela utiliza várias técnicas estatísticas e de modelagem de dados para analisar dados anteriores, identificar tendências e ajudar a tomar decisões de negócios. Embora anteriormente o aprendizado de máquina e a análise preditiva fossem vistos como dois conceitos totalmente diferentes e não relacionados, as demandas crescentes de análises de dados eficazes trouxeram os algoritmos de aprendizado de máquina para se entrelaçar com a análise preditiva. Hoje, a análise preditiva utiliza extensivamente o aprendizado de máquina para modelagem de dados devido à sua capacidade de processar com precisão grandes quantidades de dados e reconhecer padrões.

A partir desse cenário, destacamos a importância de se aprofundar no desenvolvimento e aplicação das técnicas do aprendizado de máquina a partir da utilização dos princípios do modelo preditivo.

Referências

ALI, J. B.; BRIGITTE C. M.; LOTFI, S.; MALINOWSKI, S.; FNAIECH, F. Accurate bearing remaining useful life prediction based on Weibull distribution and artificial neural network. *Mechanical Systems and Signal Processing*, n. 56, p. 150-172, 2015.

ALTEXSOFT. Preparing your dataset for machine learning: 10 basic techniques that make your data better. 2021. Disponível em: <<https://www.altexsoft.com/blog/datascience/preparing-your-dataset-for-machine-learning-8-basic-techniques-that-make-your-data-better/>>. Acesso em: 25 mar. 2021.

ALTMAN, D. G.; VERGOUWE, Y.; ROYSTON, P. Prognosis and prognostic research: validating a prognostic model. *BMJ*, 2009.

BANSAL, S. Ultimate guide to understand and implement natural language processing (with codes in Python). 2017. Disponível em: <<https://www.analyticsvidhya.com/blog/2017/01/ultimate-guide-to-understand-implement-natural-language-processing-codes-in-python/>>. Acesso em: 26 mar. 2021.

BARI, A.; CHAOUCHI, M.; JUNG, T. How to Test the Predictive Analysis Model. 2021. Disponível em: <<https://www.dummies.com/programming/big-data/data-science/how-to-test-the-predictive-analysis-model/>>. Acesso em: 3 mai. 2021.

BATZEL, T. D.; DAVID, C. S. Prognostic health management of aircraft power generators. *IEEE Transactions on Aerospace and Electronic Systems*, n. 45, v. 2, p. 473-482, 2009.

BERNARDINI, F. Combinacão de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. São Carlos, 2006.

BERNARDINI, F. Combinação de classificadores simbólicos utilizando medidas de regras de conhecimento e algoritmos genéticos. Tese de Doutorado — ICMC/USP. 2006.

BHARDWAJ, A. et al. Collaborative data analytics with datahub. *PVLDB*, v. 8, n. 12, p. 1.916-1.919, ago. 2015.

BROWNLEE, J. Deploy Your Predictive Model To Production. 2016. Disponível em: <<https://machinelearningmastery.com/deploy-machine-learning-model-to-production/>>. Acesso em: 16 abr. 2021.

BROWNLEE, J. How to Use Machine Learning Results. 2014. Disponível em: <https://machinelearningmastery.com/how-to-use-machine-learning-results/>. Acesso em: 16 abr. 2021.

BUSKIRK, T. D.; KIRCHNER, A.; ECK, A.; SIGNORINO, C. S. An introduction to machine learning methods for survey researchers. 2018. Disponível em: <<https://www.surveypartice.org/article/2718-an-introduction-to-machine-learning-methods-for-survey-researchers>>. Acesso em: 16 abr. 2021.

CARDIE, C. A case-based approach to knowledge acquisition for domain-specific sentence analysis. In: Proceedings of the Eleventh National Conference on Artificial Intelligence, p. 798-803. AAAI Press/The MIT Press, 2001.

CHRAIBI, Chakib. Artificial Intelligence in the U.S. Government: A Mission-Oriented Journey from Dark Data to Open Data and Data Innovation. Medium. 27 mai. 2020. Disponível em: <<https://medium.com/firmai/artificial-intelligence-in-the-u-s-fcf594ee6976>>. Acesso em: 28 Março 2021.

CHRIS, M.; ZITTING, J. Tika in action. Manning Publications Co., 2011.

CLEVELAND, W. S. Visualizing data. Hobart Press, 1993.

CONWAY, D. The Data Science Venn Diagram. Drew Conway Data Consulting. 30 set. 2010. Disponível em: <<http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>>. Acesso em: 27 mar. 2021.

COWIE, J.; LEHNERT, W. Information extraction. Communications of the ACM, v. 39, n. 1, p. 80-91, 1996.

DELOITTE. Business impacts of machine learning. 2017. Disponível em: <https://www2.deloitte.com/content/dam/Deloitte/tr/Documents/process-and-operations/TG_Google%20Machine%20Learning%20report_Digital%20Final.pdf>. Acesso em: 28 mar. 2021.

DEZYRE. 10 awesome machine learning applications of today. 2021. Disponível em: <<https://www.dezyre.com/article/10-awesome-machine-learning-applications-of-today/364>>. Acesso em: 25 mar. 2021.

EUROPEAN UNION. A European approach to Artificial intelligence. Policy. 9 mar. 2021. Disponível em: <<https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>>. Acesso em: 28 mar. 2021.

FAN, W.; BIFET, A. Mining big data: current status, and forecast to the future. ACM SIGKDD Explorations Newsletter, v. 14, n. 2), p. 1-5, 2013.

FINANCIAL STABILITY BOARD. Artificial intelligence and machine learning in financial services. 2017. Disponível em: <<https://www.fsb.org/wp-content/uploads/P011117.pdf>>. Acesso em: 16 abr. 2021.

GAMA, J.; FACELI, K.; LORENA, A. C.; CARVALHO, A. C. P. L. F. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2011.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep learning. The MIT Press, 2016.

GOSSETT, S. These 11 Startups Are Working On Data Privacy In Machine Learning. Builtng. 22 out. 2020. Disponível em: <<https://builtin.com/machine-learning/privacy-preserving-machine-learning>>. Acesso em: 28 mar. 2021.

GREENLAND, S. Modeling and variable selection in epidemiologic analysis. Am J Public Health, n. 79, p. 340-349, 1989.

GREYLING, C. Fundamentals of chatbot information extraction & visualization. 2019. Disponível em: <<https://cobsusgreyling.medium.com/fundamentals-of-chatbot-information-extraction-visualization-cc4a42e69c62>>. Acesso em: 26 mar. 2021.

HALEVY, A. Y. et al. Managing Google's data lake: an overview of the goods system. IEEE Data Eng. Bull., v. 39, n. 3, p. 5-14, 2016.

HARRIS, H. The Data Products Venn Diagram. Data Community DC. 19 set. 2013. Disponível em: <<http://www.datacommunitydc.org/blog/2013/09/the-data-products-venn-diagram>>. Acesso em: 27 Março 2021.

HASTINGS, P.; LYTINEN, S. The ups and downs of lexical acquisition. In: Proceedings of the Twelfth National Conference on Artificial Intelligence, p. 754-759. AAAI Press/The MIT Press, 1994.

HOCHSTER, M. What is data science?. Quora, 2014. Disponível em: <<https://www.quora.com/What-is-data-science>>. Acesso em: 27 mar. 2021.

IMPORT.IO. What is data visualization and why is it important? 2019. Disponível em: <<https://www.import.io/post/what-is-data-visualization/>>. Acesso em: 26 mar. 2021.

JOHNSON, J. Predictive analytics vs machine learning: what's the difference? 2020. Disponível em: <<https://www.bmc.com/blogs/machine-learning-vs-predictive-analytics/>>. Acesso em: 14 abr. 2021.

KAKARMATH, S.; GOLAS, S.; FELSTED, J.; KVEDAR, J.; JETHWANI, K.; AGBOOLA, S. Validating a machine learning algorithm to predict 30-day re-admissions in patients with heart failure: protocol for a prospective cohort study. 2018. Disponível em: <<https://www.researchprotocols.org/2018/9/e176/>>. Acesso em: 16 abr. 2021.

KIM, J.; MOLDOVAN, D. Acquisition of semantic patterns for information extraction from Corpora. In: Proceedings of the Ninth IEEE Conference on Artificial Intelligence for Applications, p. 171-176. Los Alamitos, CA: IEEE Computer Society Press, 1993.

KÖNIG, R. Predictive Techniques and Methods for Decision Support in Situations with Poor Data Quality. 2008. Disponível em: <<http://hb.diva-portal.org/smash/get/diva2:876907/FULLTEXT01.pdf>>. Acesso em: 16 abr. 2021.

KULES, B.; SHNEIDERMAN, B. Users can change their web search tactics: design guidelines for categorized overviews. *Information Processing & Management*, v. 44, n. 2, p.463-484, mar. 2008.

LACHANA, Z. *et al.* Identifying The Different Generations Of Egovernment: An Analysis Framework. In: The 12th Mediterranean Conference on Information Systems (MCIS), 2018.

LEE, K. K. Y.; TANG, W. C.; CHOI, K.S. Alternatives to relational database: comparison of NOSQL and XML approaches for clinical data storage. *Computer Methods and Programs in Biomedicine*, v. 110, n. 1, p.99-109, abr. 2013.

LEEK, J. T.; PENG, R. D. Statistics: what is the question? *Science*, v. 347, n. 6228, p. 1314–5, mar. 2015. Disponível em: <<https://science.sciencemag.org/content/347/6228/1314.summary>>. Acesso em: 27 mar. 2021.

MANNING, C.; SCHÜTZE, H. Foundations of statistical natural language processing. Cambridge: MIT Press, 1999.

MASON, H.; WIGGINS, C. A Taxonomy of Data Science. *Dataists*. 25 set. 2010. Disponível em: <<http://www.dataists.com/2010/09/a-taxonomy-of-data-science/>>. Acesso em: 27 mar. 2021.

McKINNEY, W. Python para análise de dados: tratamento de dados com pandas, NumPy e IPython. Novatec Editora, 2019.

MCQUILLAN, A. G. Honesty and foresight in computer visualizations. *Journal of forestry*, v. 96, n. 6, p. 15-16, jun. 1998.

MIAO, Q.; XIE, L.; CUI, H.; LIANG, W.; PECHT, M. Remaining useful life prediction of lithium-ion battery with unscented particle filter technique. *Microelectronics Reliability*, n. 53, v. 6, p. 805-810, 2013.

MILLER, T. W. Modeling Techniques in Predictive Analytics Business Problems and Solutions with R. 2014. Disponível em: <<https://ptgmedia.pearsoncmg.com/images/9780133412932/samplepages/0133412938.pdf>>. Acesso em: 16 abr. 2021.

MONTOYA, Laura. Latin American Government AI Readiness Meta-Analysis. 8 jul. 2019. Disponível em: <<https://medium.com/latinixinai/latin-american-government-ai-readiness-meta-analysis-ed66f114182f>>. Acesso em: 28 mar. 2021.

NILSSON, N. J. Introduction to machine learning. 2001. Disponível em: <<http://robotics.stanford.edu/people/nilsson/mlbook.html>>. Acesso em: 25 mar. 2021.

NTKL. NTKL 3.6 documentation. 2021. Disponível em: <<https://www.nltk.org/>>. Acesso em: 07 abr. 2021.

PARTHASARATHY, S. Predictive analytics: top 5 predictive analytics models and algorithms. 2021. Disponível em: <<https://www.logjanalytics.com/predictive-analytics/predictive-algorithms-and-models/>>. Acesso em: 15 abr. 2021.

POLYZOTIS, N.; ROY, S.; WHANG, S. E.; ZINKEVICH, M. Data lifecycle challenges in production machine learning: A survey. SIGMOD Rec., v. 47, n. 2, p. 17-28, jun. 2018.

POWERDATA. ¿Qué es Datahub, Data Lake y Datawarehouse? 2018. Disponível em: <<https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/que-es-datahub-data-lake-y-datawarehouse>>. Acesso em: 30 mar. 2021.

PRICEWATERHOUSECOOPERS. What's next for the data science and analytics job market?. 2019. Disponível em: <<https://www.pwc.com/us/en/library/data-science-and-analytics.html>>. Acesso em: 27 mar. 2021.

RASCHKA, S. Predictive modeling, supervised machine learning, and pattern classification: the big picture. 2014. Disponível em: <https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html>. Acesso em: 14 abr. 2021.

RILOFF, E. Automatically generating extraction patterns from untagged text. In: Proceedings of the 13th National Conference on Artificial Intelligence (AAAI), p. 1044-1049, 1996.

ROH, Y.; HEO, G.; WHANG, S. E. A survey on data collection for machine learning. 2019. Disponível em: <<https://arxiv.org/pdf/1811.03402.pdf>>. Acesso em: 24 mar. 2021.

RONZHYN, A.; WIMMER, M. A. Scientific foundations training and entrepreneurship activities in the domain of ICT-enabled governance. In: Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age. ACM. 2018.

ROYSTON, P.; MOONS, K. G.; ALTMAN, D. G. Prognosis and prognostic research: developing a prognostic model. *BMJ*, 2009.

SEBASTIANI, F. Machine learning in automated text categorization: a survey. Technical Report IEI-B4-31-1999. Istituto di Elaborazione dell'Informazione, 1999.

SILBERZAHN, R. et al. Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, v. 1, n. 3, p. 337–356, 2018.

SODERLAND, S.; FISHER, D.; ASELTINE, J.; LEHNERT, W. C. Inducing a conceptual dictionary. In: *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, p. 1314-1319, 1995.

TÉLLEZ-VALERO, A.; MONTES-Y-GÓMEZ, M.; VILLASEÑOR-PINEDA, L. A machine learning approach to information extraction. In: Gelbukh A. (ed.). *Computational linguistics and intelligent text processing*. Lecture Notes in Computer Science, v. 3406. Springer, Berlin, Heidelberg: CICLing, 2005 Disponível em: <https://link.springer.com/chapter/10.1007/978-3-540-30586-6_58>. Acesso em: 26 mar. 2021.

TERRIZZANO, I. G.; SCHWARZ, P. M.; ROTH, M.; COLINO, J. E. Data wrangling: the challenging journey from the wild to the lake. *CIDR*, 2015.

TEXTBLOB. TextBlob: simplified text processing. 2021. Disponível em: <<https://textblob.readthedocs.io/en/dev/>>. Acesso em: 07 abr. 2021.

TYAGI, N. Defining predictive modeling in machine learning. 2020. Disponível em: <<https://medium.com/analytics-steps/defining-predictive-modeling-in-machine-learning-887c23b7a278>>. Acesso em: 14 abr. 2021.

VORONOV, S. Machine learning models for predictive maintenance. Linköping University, 2020. Disponível em: <<http://www.diva-portal.org/smash/get/diva2:1377581/FULLTEXT01.pdf>>. Acesso em: 13 abr. 2021.

WAKEFIELD, K. Predictive analytics and machine learning. 2021. Disponível em: <https://www.sas.com/en_gb/insights/articles/analytics/a-guide-to-predictive-analytics-and-machine-learning.html>. Acesso em: 14 abr. 2021.

ZAMIR, O.; ETZIONI, O.; MADANI, O.; KARP, R. M. Fast and intuitive clustering of web documents. In *Proc. of KDD '97*, p. 287-290, 1997.



PUC
CAMPINAS
PONTIFÍCIA UNIVERSIDADE CATÓLICA