## 1. Introdução: Mineração, Ciência de Dados e o Aprendizado de Máquina

Nesta trilha você vai aprender:

* O que é Mineração, Ciência de Dados e o Aprendizado de Máquina e sua importância
* Algumas aplicações e exemplos importantes de aplicação desses conceitos
* O que são aprendizado Supervisionado e não Supervisionado

Você certamente já teve experiência ou contato com a Mineração de Dados, ou mais modernamente o que chamamos Ciência de Dados, mesmo que não tenha conhecimento disso. Se você já empregou um site como a Netflix ou efetuou compras na Amazon você já teve a experiência de ver recomendados títulos de filmes ou de livros com base nas suas compras anteriores. Esse é apenas um dos usos comuns que podemos fazer da Ciência de Dados e das técnicas de Aprendizado de Máquina e essa, *transformação digital*, está trazendo uma série de mudanças nos negócios e na vida das pessoas em geral.

Outros usos menos evidentes você talvez quase não perceba. Quando você faz uso de cartões de crédito, ou faz operações bancárias na Web ou através de um celular, suas operações certamente estão sendo analisadas por um sistema inteligente, cuja base está na Ciência de Dados e nos Algoritmos de Aprendizado de Máquina, para certificar a autenticidade ou fraude de suas operações. Talvez você até já tenha tido a experiência de um cartão bloqueado por uma suspeita de fraude.

A Ciência de Dados e os Algoritmos de Aprendizado de Máquina não se restringem unicamente ao comércio eletrônico e operações bancárias. Eles também têm aplicações na Ciência, na Medicina diagnóstica, na Robótica e veículos autônomos, em Jogos Digitais e são a base de boa parte dos negócios digitais. Suas soluções são ainda ofertadas pelos principais provedores de serviços em Nuvem como Amazon, Azure, Google Cloud e IBM Bluemix, e por empresas de análise de dados mais tradicionais de dados SAS e Tableau.

Um grande número de fatores tem contribuído para crescente uso da Ciência de Dados e de algoritmo de Aprendizado de Máquina, seja do lado da demanda, com novas aplicações (aplicações móveis, IoT, redes sociais, games, marketing digital, diagnósticos por imagem, aplicações geoespaciais), seja do lado da capacidade das soluções (aumento da capacidade computacional com o uso da computação em nuvem, tecnologias de armazenamento, GPU; grandes avanços no aprendizado de máquinas - *deep learning*, *graphical models*). Técnicas mais elaboradas, como *Deep Learning*, vêm trazendo um impacto massivo no segmento de tecnologias de consumo e você pode encontrar o Google e Baidu empregando essas técnicas para o reconhecimento de imagens, em *search engines* e em soluções de tradução automática, e Apple (Siri), Amazon (Alexa), Microsoft (Cortana) e Samsung (Bixby) empregando-as para *speech recognition*.

Isso deve ser suficiente para você entender a importância dessas soluções, e essa *transformação digital* está trazendo uma série de mudanças nos negócios e na vida das pessoas em geral.

# O que é a Ciência de Dados?

A Ciência de Dados é uma disciplina que trata de como extrair e aplicar insights de informações a partir de dados. É uma ciência interdisciplinar e emprega conceitos de Estatística, Aprendizado Máquina e diferentes técnicas Computacionais de aquisição e armazenamento para obter insights dos dados dando suporte à decisões que podem ser automatizadas ou não.

Ciência de Dados, Aprendizado de Máquina, Mineração e Análise de Dados são muitas vezes igualmente empregados para designar essa busca de *insights* ou de conhecimento dos dados, e apresentam de fato muitos pontos e técnicas em comum, embora possamos fazer algumas distinções entre eles.

O *Aprendizado de Máquina* concentra-se nos algoritmos para extração de padrões dos dados, sendo uma parte importante da Inteligência Artificial em geral. Já a a *Mineração e Análise de Dados* está mais associada à análise de dados estruturados com técnicas mais tradicionais de análise. A *Ciência de Dados* é mais recente e inclui todos esses aspectos incorporando novos desafios como a captura e transformação de grandes volumes de dados (big data), manipulação de dados não estruturados (web, social networks), serviços de computação em nuvem, aspectos de segurança e privacidade etc.

Dareamos aqui mais ênfase na Ciência de Dados, com a aplicação de Algoritmos de Aprendizado de Máquina, que à Mineração e Análise de dados mais tradicionais por se tratar de uma abordagem mais inovadora e poderosa no tratamento dos dados. Isso não quer dizer que técnicas como Aprendizado de Máquina ou Dados Big Data irão substituir todas as formas de análise tradicional e dados estruturados das empresas. Esses ainda são responsáveis pela maior parte do trabalho de análise de dados na maior parte das empresas.

|  |  |  |
| --- | --- | --- |
| **Tipo** | **Ciência de Dados** | **Análise Tradicional** |
| Tipo de Dados | Não estruturados | Estruturados (linhas e colunas) |
| Volume de Dados | Big Data (centenas de Terabytes) | Dezenas de Terabytes ou menos |
| Fluxo de Dados | Contínuo, Big Data | Estático |
| Métodos de Análise | Machine Learning | Visual, Hypothesis-Based |
| Propósito | Data-base Products | Internal Support Decision |

**Diferenças entre técnicas mais atuais de Ciência de Dados e modelos mais tradicionais de Análise**

# Inteligência Artificial e Ciência de Dados

Um termo que também aparece bastante relacionado à Ciência de Dados é o termo Inteligência Artificial. Embora Inteligência Artificial, Ciência de Dados e Machine Learning tenham muito em comum não são exatamente disciplinas idênticas.

A Inteligência Artificial envolve muitos outros temas como, Representação do Conhecimento, Agentes Inteligentes, Visão Computacional, Robótica, Processamento de Linguagem Natural e muitos outros paradigmas.

# Breve Histórico da Ciência de Dados

O ser humano analisa dados desde o primeiro momento em que passa a criá-los. De qualquer modo, para os nossos propósitos, você pode ter esses como alguns marcos importantes da Ciência de Dados até os nossos dias:

1. **17xx - 18xx** - Summary (desciptive) statistics. Fundamentos da Teoria das Probabilidades (Bernoulli, Bayes, Pascal). Distribuição de Probabilidades, Bayes's Rule (Laplace, Gauss). Least Square Method (Gauss). Bar, line, area and pie charts (William Playfair).
2. **19xx - 20xx** - Hypothesis Test (Pearson). Multivariate Analysis (Fisher). Relational Data Model, SQL (Edgar F. Codd, 1970). Turing Machine (Allan Turing). Information Theory (Shannon, 1948). Artificial Neural Model (MCCulloch & Pitts, 1943). Discriminatory Analysis (=classification, Evelyn Fix & Hodges, 1951). First AI Workshop (1956, Dartmouth College).Machine Learning (kmeans, knn, Decision Trees, 1960-70). Computers. Early Internet. Early AI. Perceptron
3. **20xx - 20xx** - Humans generate and store the same amount of data produced since inventing of writing every day. Search Engines. Social Networks. Google, Facebook, Spotify, Netflix. SmartPhones. IoT. Big Data (HDFS, Mapreduce). Deep Learning. t-SNE. Indústria 4.0. NoSQL databases. Cloud Computing. GPU. **Pela primeira vez o valor das Empresas de Informação (Google, Facebook etc.) superam o valor das empresas de Petróleo (2018)**

# Aprendizado de Máquina: Um novo paradigma

O Aprendizado de Máquina ou *Machine Learning* introduz um novo paradigma de como construir automação ou programas, a partir unicamente de dados.

Diagrama

Descrição gerada automaticamente

Figura 1. Diferentes Paradigmas de Programação: sem o Aprendizado de Máquina e com o Aprendizado de Máquina. (Fonte: Goulet, J.-A. (2020), Probabilistic Machine Learning for Civil Engineers, The MIT press)

Em um modo tradicional programas (automações) são construídos a partir do conhecimento que o programador possui ou adquiriu. Por exemplo, um programador conhecendo a regra de conversão de graus Celsius para Fahrenheit pode fazer um programa para fazer exatamente cálculo repetidas vezes. Do mesmo modo ele pode fazer um programa para resolver um sistema de equações lineares e mesmo decidir se o sistema não tem solução se ele souber, por exemplo, o método de *Cramer* de solução de sistemas lineares. Em qualquer caso há uma regra, método ou padrão conhecido que pode ser codificado pelo programador.

A ideia do Aprendizado de Máquina é *capturar* esses padrões a partir de dados nos casos em que nós não conhecemos esses padrões e onde esses padrões são em geral probabilísticos, e não exatos como nos casos anteriores. Você pode imaginar um sistema de detecção de fraudes de cartão de crédito ou o sistema de recomendação de filmes em uma plataforma de streaming. Nós não conhecemos a *regra* ou o padrão por traz dessas decisões, *que valores, horários ou tipos de compra indicariam uma fraude?*, *que filmes podem ser mais recomendados para este ou aquele usuário?*, mas temos um grande conjunto de dados em que podemos buscar um certo padrão para reproduzir, como um programa, esse padrão. Não é esperado, entretanto, que esse resultado seja exato, uma vez que esses não são processos *determinísticos*. Seu resultado será, então, de que *existe uma maior probabilidade de fraude em uma operação* ou de que *este filme tem maior probabilidade de ser uma boa recomendação para este usuário* o que é certamente um resultado útil.

Isso permite revisitarmos nosso conceito de Ciência de Dados:

*A Ciência de Dados é a exploração e análise de grandes quantidades de dados, a fim de descobrir padrões válidos, novos, potencialmente úteis e, finalmente, compreensíveis em dados.*

* **Válido**: os padrões são válidos em geral. Você espera que exista um certo padrão, por exemplo um padrão de Compras com Cartão de Crédito ou de Produtos e Filmes consumidos, mesmo que você não conheça exatamente qual ele é. Você deseja *capturar* esse padrão e espera que ele seja válido em novos casos.
* **Novos**: Nós não conhecemos o padrão de antemão. Se o padrão é conhecido você não precisaria buscá-lo nos dados. Você pode ter milhares de dados sobre o movimento de um satélite, mas se você quer buscar a sua posição, digamos daqui a uma hora, você provavelmente empregará apenas alguns dados e um modelo físico, um padrão já estabelecido por fórmulas que envolvem velocidade, massa, atrito etc. e isso fará muito mais sentido que buscar um novo padrão para o movimento do satélite unicamente através de dados.
* **Útil**: podemos decidir ações a partir da aplicação dos padrões encontrados. Eles permitirão fazer predição de novos casos ou ainda fornecer a resposta mais provável a partir da qual podemos tomar uma decisão como a de bloquear um cartão de crédito com transações suspeitas ou recomendar um filme a um usuário de uma plataforma de streaming.
* **Compreensível**: podemos interpretar e compreender os padrões. Os padrões extraídos podem aplicados a novos dados e serem mensurados na sua eficiência, compreendidos a sua precisão etc. embora diferentes técnicas de aprendizado de máquina apresentem diferentes níveis de compreensão dos modelos.

# Tipos de Padrões e Aplicações

Que *tipos de padrões* válidos, novos, potencialmente úteis e compreensíveis podemos esperar extrair dos dados com algoritmos de aprendizado de máquina? Uma forma útil de você entender os tipos de padrões que podemos obter através do aprendizado máquina é entendermos algumas as tarefas que esses algoritmos são capazes de resolver e suas aplicações.

|  |  |
| --- | --- |
| **Tarefa** | **Exemplos** |
| Classificação | Fraud/not Fraud, Churn/Not Churn |
| Regressão | Preços de Imóveis, Aluguéis, Veículos |
| Clusterização | Segmentação de Clientes, Produtos, Documentos |
| Regras de Associação | Pacotes de produtos |
| Anomaly Detection | Mal funcionamento de dispositivos, Ataques cibernéticos |
| Matching | Recomendação de produtos ou amigos de uma rede social |

Tarefas de Aprendizado de Máquina e Aplicações

Podemos empregar esses padrões para classificar automaticamente Clientes sujeitos a Crédito/ou Não; classificar emails como Spam/Not Spam ou Apólices de Seguro em diferentes classes de risco; ou ainda Peças produzidas com diferentes níveis de qualidade (**Classificação**). Um padrão também pode fornecer uma estimativa de valores para formação de preços, como preços de Máquinas produzidas com peças de diferentes fornecedores ou de imóveis; ou ainda a previsão volumes de Produção, Consumo ou o aumento de casos de uma doença (**Regressão**). Podemos também identificar grupos de Clientes com os mesmos *padrões* de comportamento e gostos (**Clusterização**, como aplicada na segmentação de clientes e produtos); produtos que são adquiridos frequentemente juntos (**Regras de Associação**); comportamentos anômalos, como um tráfego de internet anormal ou o mal comportamento de uma planta de produção (**Detecção de Anomalias**); e ainda, recomendar produtos a um usuário com base em padrões de semelhança com produtos ja adquiridos ou um *par perfeito* em um site de relacionamentos (**Matching**, e sistemas de recomendação em geral).

Tabela

Descrição gerada automaticamente

Figura 2. Exemplo de uma tarefa de classificação (Fonte: Vasant Dhar. Data Science and Prediction (ACM 2013))

# Exemplos de Aplicações

Você vai encontrar abaixo alguns casos clássicos da aplicação da Ciência de Dados bem como do Aprendizado de Máquina em diferentes segmentos da indústria.

## Clássicos da Análise de Dados

Existem muitos casos clássicos que passaram a fazer parte da história e mesmo do folclore da Ciência de Dados dando origem a livros, filmes e uma série de estórias, nem sempre muito fiéis, mas que ainda ilustram o potencial e os resultados da aplicação da Análise de Dados. Vale a pena você conhecer alguns desses casos.

Logotipo

Descrição gerada automaticamente

* [How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did](https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#52ff65166668) | <https://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/#52ff65166668>

Neste case clássico um pai fica indignado ao receber em casa folhetos de promoção de produtos para gestante e bebês julgando que eles expunham sua filha adolescente. Ele não sabia, mas sua filha estava grávida e o Target (uma grande loja de departamentos americana) identificou isso pelas compras que haviam sido feitas com o cartão de crédito do pai.

Uma imagem contendo Ícone

Descrição gerada automaticamente

* [You’ll Never Guess What Surprising Product Walmart Stocks up on Before Big Storms](https://www.southernliving.com/news/walmart-strawberry-pop-tarts-hurricane) | <https://www.southernliving.com/news/walmart-strawberry-pop-tarts-hurricane>

Nos períodos de fortes chuvas ou mesmo tornados é comum nos Estados Unidos os cidadãos terem que ficar dias trancados em casa sem poder sair. O Walmart analisando dados dos períodos de chuva de anos anteriores conseguiu verificar que nesses períodos havia uma demanda extraordinária de tortas de maçã (e provavelmente identificou também outros produtos) e pôde, então, se antecipar e provisionar tortas para atender a demanda, potencializando suas vendas e sem causar a falta do produto para os clientes.

Uma imagem contendo Ícone

Descrição gerada automaticamente

* [Beer and Diapers (Nappies)](https://bigdatabigworld.wordpress.com/2014/11/25/beer-and-nappies/) | <https://bigdatabigworld.wordpress.com/2014/11/25/beer-and-nappies/>

Em mais um case clássico do Wall Mart, analisando dados do comportamento dos consumidores o Wall Mart conseguiu observar uma forte correlação de compras de seus clientes: às sextas-feiras, muitos clientes que compravam fraldas também compravam cerveja! Isso levou o Wall Mart a colocar esses produtos mais próximos para facilitar a compra dos seus clientes. Mais tarde, descobriu-se que os clientes eram jovens pais que, uma vez que já tinham que ficar sexta à noite cuidando de seus filhos ou em casa com suas esposas, aproveitavam para relaxar assistindo os jogos da temporada na TV.



* [The Story of Moneyball Proves Importance of Both Big Data and Big Ideas](https://datasciencedegree.wisconsin.edu/blog/moneyball-proves-importance-big-data-big-ideas/) | <https://datasciencedegree.wisconsin.edu/blog/moneyball-proves-importance-big-data-big-ideas/>

Esse talvez seja o caso mais clássico e por isso tornou-se até mesmo um premiado filme estrelado pelo ator Brad Pitt: Moneyball (O Homem Que Mudou o Jogo). O filme conta a estória de um caso real do impacto do uso de dados para a tomada de decisões. Na temporada de Beisebol de 2001 o Oakland Athletics era um time muito pequeno, passando por dificuldades e sem recursos para contratar grandes estrelas para fazer frente ao campeonato. O diretor do time decide então adotar uma estratégia completamente inovadora. No lugar de tomar decisões com base em conselheiros e *olheiros* de beisebol como todos times faziam à época, ele contrata um estatístico de Yale e passa a decidir sobre a contratação de jogadores e mesmo sobre as posições em que jogariam com base em dados do histórico dos jogadores (quantidade de arremessos com sucesso, pontos por partida etc.). A estratégia conseguiu supreendentemente levar o pequeno time de Oakland até às finais da temporada (algo como levar um time de futebol da terceira divisão à final da libertadores). A partir daí todos os grandes times de Beisebol passaram a empregar estatísticos e analistas de dados para decidir sobre jogadores e estratégias de jogo, e a mesma estratégia foi adotada nos anos seguintes para outros esportes como o Futebol Americano criando uma indústria milionária de análise de dados esportivos.

## Casos Gerais

Basta dar um Google com 'Data Science', 'Ciência de Dados' ou 'Machine Learning' e você encontrará inúmeros cases de aplicação de Ciência de Dados e do Aprendizado de Máquina. O site:

Placa azul com letras brancas

Descrição gerada automaticamente com confiança média

[Machine Learning and Data Science Applications in Industry](https://github.com/ashishpatel26/Real-time-ML-Project) | <https://github.com/ashishpatel26/Real-time-ML-Project>

|  |  |  |
| --- | --- | --- |
|  |  |  |
| [Accommodation & Food](#accommodation) | [Agriculture](#agriculture) | [Banking & Insurance](#bankfin) |
| [Biotechnological & Life Sciences](#biotech) | [Construction & Engineering](#construction) | [Education & Research](#education) |
| [Emergency & Relief](#emergency) | [Finance](#finance) | [Manufacturing](#manufacturing) |
| [Government and Public Works](#public) | [Healthcare](#healthcare) | [Media & Publishing](#media) |
| [Justice, Law and Regulations](#legal) | [Miscellaneous](#miscellaneous) | [Accounting](#accounting) |
| [Real Estate, Rental & Leasing](#realestate) | [Utilities](#utilities) | [Wholesale & Retail](#wholesale) |

Dá uma ideia das várias áreas de aplicação e traz detalhes dos diversos *cases* em cada segmento. Você pode explorar e focar em apenas alguns desses casos que forem mais do seu interesse. Relaciono ainda casos da indústria digital que emprega massivamente essa tecnologia para criar novos produtos e serviços, e da manufatura.

* Casos de Uso da Indústria Digital
  + [Netflix](https://data-flair.training/blogs/data-science-at-netflix/) | <https://data-flair.training/blogs/data-science-at-netflix/>
  + [Twitter](https://data-flair.training/blogs/data-science-at-twitter/) | <https://data-flair.training/blogs/data-science-at-twitter/>
  + [Facebook](https://data-flair.training/blogs/data-science-at-facebook/) | <https://data-flair.training/blogs/data-science-at-facebook/>
* Casos de Uso na Indústria (Manufatura)
  + [How is Data Science Used in Manufacturing?](https://www.sensrtrx.com/how-is-data-science-used-in-manufacturing-companies) | <https://www.sensrtrx.com/how-is-data-science-used-in-manufacturing-companies>
  + Desempenho, garantia de qualidade e rastreamento de defeitos
  + Manutenção preditiva e condicional
  + Previsão de demanda e rendimento
  + Cadeia de suprimentos e relações com fornecedores
  + Preços de mercado global
  + Automação e projeto de novas instalações
  + Novos processos e materiais para desenvolvimento de produtos e técnicas de produção
  + Sustentabilidade e maior eficiência energética

## 

## Casos da Indústria Nacional

São bastante conhecidos casos de aplicação de Ciência de Dados e Aprendizado de Máquina no varejo (Magazine Luiza, Mercado Livre) e Mercado Financeiro (Bradesco e bancos em geral). Vamos então apenas destacar então algumas outras aplicações menos conhecidas.

### Engenharia Aeronáutica, EmbraerX

**A chegada dos carros voadores eVTOL** Projetos impulsionam desenvolvimento de protótipos de eVTOLs no Brasil e no mundo, os veículos têm a proposta de voos autônomos baseados no aprendizado de máquina.

Mapa

Descrição gerada automaticamente

Figura 3. Estudo da Embraer-Uber para o desenvolvimento de veículos aéreos autônomos. (Fonte: <https://revistapesquisa.fapesp.br/2019/12/03/a-chegada-dos-carros-voadores/>)

### Engenharia Florestal

**Identificação de madeiras** Sistemas inteligentes classificam a qualidade das lâminas de madeira e a espécie da árvore à qual pertencem de modo automático para direcionar o uso da lâminas.

Diagrama

Descrição gerada automaticamente

Figura 4. Classificação de Lâminas de Madeira empregando Aprendizado de Máquina (Fonte: <https://revistapesquisa.fapesp.br/2017/07/18/identificacao-de-madeiras/>)

### 

### Energia, Petrobrás

Uma plataforma marítima moderna pode ser equipada com até 80 mil sensores, capazes de gerar aproximadamente 15 petabytes, o equivalente a 15 milhões de gigabytes de informação, durante a vida útil da unidade. São informações que abrangem dados de produção, de segurança operacional, de comportamento de reservatório, entre outras. Imagine o quanto esse volume de dados estratégicos uma vez registrados, interpretados e combinados aos demais dados de outras plataformas pode gerar de valor para uma companhia de petróleo. Essa é a visão da Petrobrás que também entende que “o gerenciamento de dados é o novo “óleo” da indústria. O maior desafio é como conectar, integrar e extrair valor desse volume de dados imenso". Por isso ela investe em inúmeros projetos que envolvem Ciência de Dados e Aprendizado de Máquina, como alguns dos projetos resumidos abaixo.

* **Metodologias Big Data aplicadas a modelagem na otimização de processos:** emprega técnicas de Big Data e Deep Learning para geração e atualização de modelos em tempo real para serem utilizados em monitoração, controle e otimização das unidades operacionais de refino e exploração e produção.
* **Caracterização expressa de petróleos:** para uma avaliação de petróleo completa, em tempo real, a partir de um conjunto limitado de características medidas em laboratório, empregando algoritmos bioinspirados e tecnologias em Deep Learning e Machine Learning.
* **Sistema supervisório de SMS baseado em Inteligência Artificial:** sistema de monitoramento de imagens em ambientes de risco operacional com uso de redes neurais profundas, tecnologia conhecida como Intelligent Video Analytics (IVA), detectando riscos e desvios ocupacionais para reduzir a taxa de acidentados registráveis.
* **Busca semântica e identificação de análogos de reservatório:** investiga o potencial de uma solução tecnológica baseada em tecnologias digitais para identificar similaridades entre projetos exploratórios e projetos de desenvolvimento da produção de reservatórios de petróleo.
* **Desenvolvimento de ferramenta computacional (SimCAP), que contribui para o aumento de produção de asfalto a partir de petróleos do pré-sal:** desenvolvida com base em dados experimentais e inteligência artificial, por meio do uso de técnicas de aprendizagem computacional otimiza o elenco de petróleos para redução das restrições de alocação de pré-sal nas refinarias durante campanha de produção de asfalto.

# O Processo de Ciência de Dados e os Modelos de Aprendizado de Máquina

Mas esse processo que permite as empresas extrair conhecimento útil dos dados para resolver problemas práticos de seus negócios não é em geral um processo simples e, antes de você entender e aprender a construir modelos de Aprendizado de Máquina, é importante que você entenda o processo completo de construção desses modelos.

Esse um processo que pode envolver uma série de técnicas que vão da estatística ao uso de técnicas eficientes de armazenamento e processamento e muitos outros recursos para que se possa aplicar modelos de aprendizado computacional. Para lidar e organizar a complexidade desse processo, um processo de fases é geralmente aplicado. O **CRISP-DM, Cross Industry Standard Process for Data Mining** (Processo Padrão Inter-Indústrias para Mineração de Dados) é um processo de fases bastante aceito na indústria para representar um ciclo completo de Ciência e Análise de Dados, incluido a aplicação de modelos de Aprendizado de Máquina. Esse processo apresenta algumas variantes ao longo do tempo, como por exemplo a versão do Google em The Process for Data Preparation and Feature Engineering (ver <https://developers.google.com/machine-learning/data-prep/process>), mas de fato elas mudam bastante pouco o modelo CRISP e a essência do processo CRISP-DM é em geral preservada e é empregada na maior parte dos casos.

Diagrama

Descrição gerada automaticamente

Figura 5. Fases do CRISP-DM.

# As Fases do CRISP-DM

As fases do modelo de referência CRISP-DM encontram-se resumidas no quadro abaixo.

Diagrama

Descrição gerada automaticamente com confiança média

Figura 6. Guia Visual do Ciclo CRISP-DM. (Fonte: Pete Chapman, et. al. CRISP DM 1.0 Step-by-step data mining guide)

Você pode também querer consultar a documentação original desse modelo de referência.

* [CRISP Visual Guide](http://meusite.mackenzie.br/rogerio/TIC2021S1/T1/crisp_visualguide.pdf) | Acesse aqui <http://meusite.mackenzie.br/rogerio/TIC2021S1/T1/crisp_visualguide.pdf>
* [CRISP DM 1.0](http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf) | Acesse aqui <http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf>

De qualquer modo, você encontra a seguir, uma descrição resumida de cada uma dessas fases.

## Entendimento do negócio

Esta fase inicial concentra-se em compreender os **objetivos e requisitos do projeto ou problema** de uma perspectiva de negócios, em seguida, convertendo este conhecimento em uma definição de problema de mineração de dados e um plano preliminar projetado para alcançar os objetivos.

Como qualquer projeto é uma fase fundamental para o sucesso do projeto, com o alinhamento das expectativas, definição dos recursos etc.

## Entendimento dos dados

A fase de entendimento dos dados começa com a coleta de dados e segue com atividades que permitem que você familiarize-se com os dados, identificar problemas de qualidade de dados, os primeiros insights sobre os dados e / ou detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas.

É uma fase de modelos e hipóteses simples, mas que já pode fornecer muitas respostas relevantes ao problema tratado.

## Preparação de dados

A fase de preparação de dados incluir todas as atividades necessárias para tornar os dados adequados para a modelagem. As tarefas de preparação de dados provavelmente serão realizadas várias vezes e não em qualquer ordem prescrita. As tarefas incluem seleção de dados e atributos, transformações como normalização e recodificação dos dados, limpeza de dados (dados faltantes, ruídos, *outliers* etc.), merge de dados etc.

*É importante destacar que, na prática, a atividade de modelagem tem* ***80% do esforço concentrado nas fases de Entendimento e Preparação dos Dados****. Isto é, não obstante o eventual glamour do Cientista de Dados, seu trabalho consiste, na maior parte do tempo, em coletar e manusear grandes conjuntos de dados.*

## Modelagem

Nesta fase, várias técnicas de modelagem são selecionadas e aplicadas, e seus parâmetros são calibrados para otimizar valores. Normalmente, existem várias técnicas para o mesmo tipo de problema de mineração de dados. Algumas técnicas têm requisitos na forma de dados. Portanto, voltar à fase de preparação de dados geralmente é necessário. Essa é uma característica importante desse modelo de referência: **as fases não necessariamente seguem sempre em uma única direção**, pois achados em uma etapa posterior podem sempre implicar na necessidade de voltar em tarefas das etapas anteriores.

## Avaliação

Nesta fase do projeto, você construiu um modelo (ou modelos) que parece ter alta qualidade a partir de uma análise de dados perspectiva. Antes de prosseguir para a implantação final do modelo, é importante avaliá-lo completamente e revisar as etapas executadas para criá-lo, para ter certeza de que o modelo atinja adequadamente os objetivos de negócios.

Para isso é necessário **empregar métricas** que permitam avaliar os modelos. Um exemplo conhecido é o empregar o coeficiente de determinação () para avaliação de diferentes modelos de regressão linear, existindo métricas como acuracidade, erro etc. para outros tipos de modelos.

## Distribuição

A criação do modelo geralmente não é o fim do projeto. O resultado do modelo precisa ser organizado e apresentado de forma que possa ser usado. Isso frequentemente envolve a inclusão de modelos dentro dos processos de tomada de decisão de uma organização.

Isso pode envolver uma atividade simples de geração de um relatório em uma página Web até a implantação de diversos artefatos de software para coletar dados de compras de clientes e gerar recomendações automáticas no aplicativo de vendas para outros clientes.

*Note que essas fases não são fases estanques e acontecem de forma interativa. Elas muitas vezes são fases que acontecem de modo concorrente, uma realimentando a outra. Por exemplo, durante a Preparação dos Dados podemos fazer descobertas que não haviam sido feitas durante a fase de Entendimento dos Dados, e mesmo na Modelagem podemos encontrar a necessidade de retornar ao Entendimento dos Dados para verificar algum padrão que não havíamos notado antes.*

# Aprendizado Supervisionado e Não Supervisionado

O aprendizado supervisionado é uma abordagem de aprendizado de máquina definida pelo uso de conjuntos de dados rotulados. Por exemplo você pode ter dados de transações de crédito que são previamente rotulados como fraude ou não fraude, dados de exames de tumores rotulados como malignos ou benignos, ou ainda dados de imóveis *rotulados* com os seus preços. Esses dados funcionam como um conjunto de exemplos para o aprendizado de máquina, e são empregados para treinar ou *supervisionar* os algoritmos para classificar novos dados ou prever valores. O algoritmo busca um padrão para aproximar as entradas e saídas rotuladas e medir a precisão dessa aproximação.

Em geral o aprendizado supervisionado trata de dois tipos de problemas, ou tarefas: classificação ou regressão. A classificação de novas transações de crédito como fraude ou não fraude ou de dados de tumores como maligno ou benigno são do primeiro tipo e Árvores de Decisão e Regressão Logística são exemplos de algoritmos que podem ser empregados. A partir do conjunto de exemplos eles podem criar *mapeamento* (como uma função) do conjunto de dados de entrada (variáveis preditoras) e as respectivas categorias (os rótulos, fraude/não fraude ou maligno/benigno) que pode então ser aplicado a novos casos para a predição das categorias. A predição de valores (e não de categorias), como a predição do preço de imóveis, é um exemplo de problema de regressão e algoritmos como os de regressão linear ou uma rede neural, podem mapear entradas e os valores de saída (rótulos) de um conjunto de exemplos para estimar valores de novos imóveis que não fazem parte do conjunto inicial de *treinamento*.



Figura 7. Acima um mesmo conjunto de dados para o aprendizado supervisionado de classificação e regressão. são os rótulos associados a cada instância de dados com variáveis preditoras . Os modelos supervisionados criam, a partir dos exemplos do conjunto de treinamento, uma função que reproduz do mesmo padrão de mapeamento das entradas e saídas do conjunto de treinamento.

# Aprendizado não Supervisionado

O aprendizado não supervisionado usa algoritmos de aprendizado de máquina para analisar e agrupar conjuntos de dados não rotulados. São modelos mais analíticos que os modelos supervisionados que, em geral, buscam fornecer uma predição de valores ou categorias. A ideia desses algoritmos é encontrar padrões ocultos nos dados sem a necessidade de qualquer intervenção humana ou *rótulos* pré-definidos e, por isso *não supervisionados*.

Modelos de aprendizagem não supervisionados são empregados vários tipos de tarefas ou problemas e os principais são: agrupamentos (ou clusterização), associações, redução de dimensionalidade e detecção de anomalias.

Clustering é uma técnica de mineração de dados para agrupar dados não rotulados com base em suas semelhanças ou diferenças. Algoritmos de agrupamento como o K-médias ou Clusterização hierárquica permitem identificar dados semelhantes em grupos que se diferenciam dos demais e são empregados para segmentação de clientes, produtos etc.

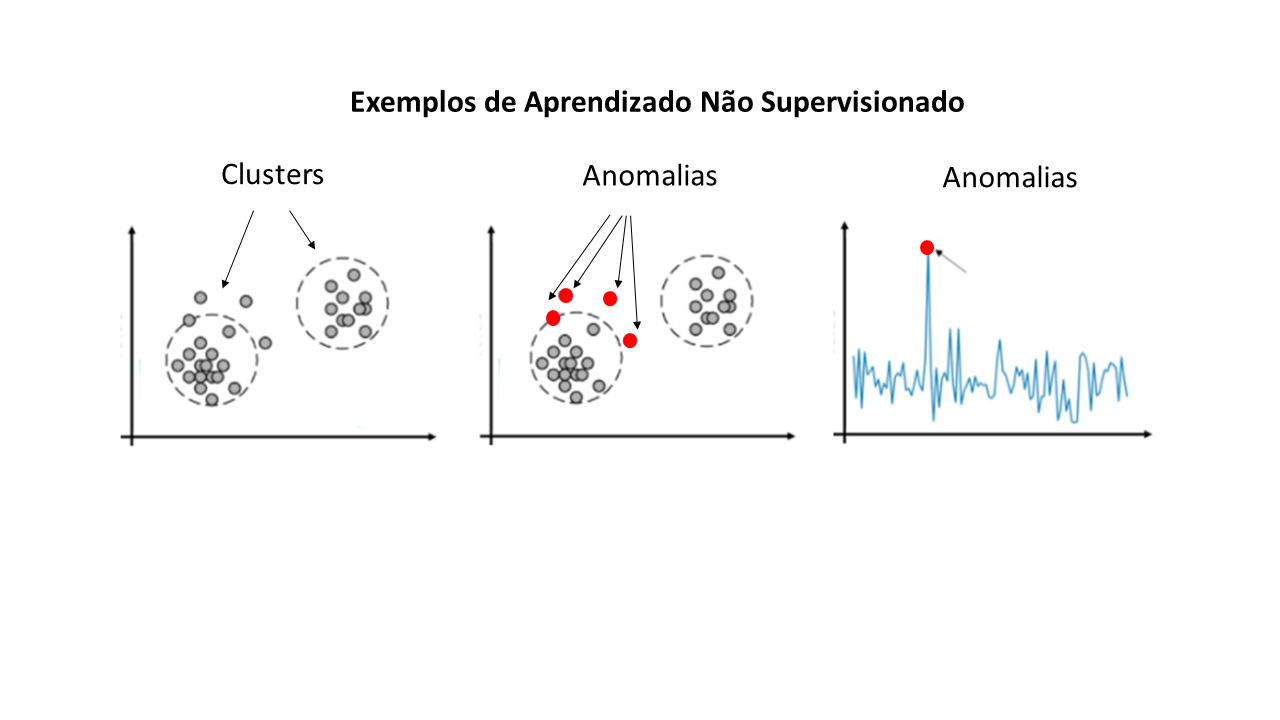
Algoritmos de Associação, como o *Apriori*, podem ser usados ​​para análises de cesta de compras ou sistemas de recomendação e permitem identificar que clientes que compram um determinado item também compram outro. Algoritmos como PCA (*Principal Component Analysis*), permitem reduzir a dimensionalidade de um grande número de atributos para facilitar a análise. E modelos de detecção de anomalia, como *SVM One-Classification*, são empregados para identificar uma anomalia, por exemplo no tráfego de internet de um site de compras ou em uma imagem de câmera que detecta movimento.

Figura 8. Exemplos de Aprendizado Supervisionado. Você mesmo, apenas olhando as figuras, poderia identificar tanto os agrupamentos como as anomalias sem, entretanto, ter qualquer exemplo ou rótulos indicando esses grupos e esses pontos atípicos.

# Aprendizado Supervisionado *vs* Não Supervisionado

A principal diferença entre aprendizagem supervisionada e não supervisionada é a presença de um conjunto de treinamento com esses dados rotulados para o aprendizado. O aprendizado supervisionado usa dados de entrada e saída rotuladas (classes ou valores), enquanto o aprendizado não supervisionado emprega somente os dados de entrada.

Os modelos de aprendizado supervisionado sempre requerem algum tipo de intervenção humana inicial para rotular os dados o que é, em muitos casos um grande problema e você pode imaginar o esforço necessário para classificar (rotular) imagens ou notícias da internet, e a solução desse problema fundamental é uma forte área de pesquisa. Os modelos de aprendizagem não supervisionada por outro lado obtêm por conta própria a estrutura inerente dos dados não rotulados, mas requerem a intervenção humana para validar as variáveis ​​de saída sendo modelos de certo modo analíticos e não preditivos como os modelos Supervisionados. Por exemplo, uma clusterização pode identificar grupos distintos de clientes, mas não fornece qualquer informação de *quem* está nesses grupos e se algum desses grupos é útil para uma campanha de marketing ou uma promoção de vendas. O objetivo dos modelos supervisionados é prever resultados para novas instâncias de dados enquanto no aprendizado não supervisionado o objetivo é obter *insights* sobre os dados existentes.

Outras técnicas de aprendizado que envolvem **aprendizado semi-supervisionado**, que utiliza apenas uma parte dos dados rotulados. Por exemplo, em uma rede social um usuário pode postar uma foto com uma outra pessoa e um algoritmo semi-supervisionado passa a reconhecer essa pessoa em outras fotos a partir do momento em que o usuário identifica essa pessoa em algumas fotos. Já técnicas de **aprendizado por reforço** empregam um sistema com um fator de recompensa para o aprendizado. Nesse tipo de aprendizado o treinamento é feito para tomar uma sequência de decisões e um *agente* aprende a atingir uma meta em um ambiente incerto e potencialmente complexo. O agente efetua tentativas e erros para buscar encontrar uma solução a um problema recebendo recompensas ou penalidades a cada ação que realiza. Seu objetivo é maximizar a recompensa total que pode ser, por exemplo, obter um máximo de pontos em um jogo. Esse tipo de aprendizado é empregado principalmente em robótica e veículos autônomos e, aspiradores autônomos implementam em geral modelos simples desse tipo de aprendizado.

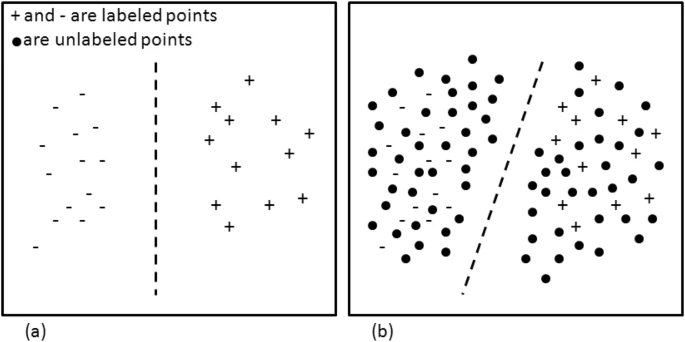


Figura 9. Pontos não previamente classificados podem ser classificados a partir de uma classificação, ou rotulação, parcial dos dados (Fonte: Peikari,2018).

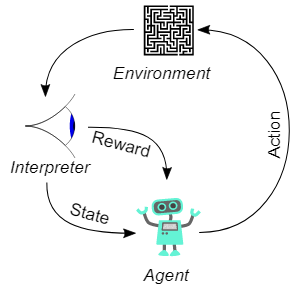


Figura 10. Esquemático do aprendizado por reforço. Um agente interpreta o ambiente e busca maximizar uma função de recompensa. (Fonte: Wikipedia)

# Síntese

Nesta trilha você tomou contato os principais conceitos Mineração, Ciência de Dados e o Aprendizado de Máquina. A partir de uma série de exemplos você pôde entender a importância e as mudanças que esses conceitos e as tecnologias que eles envolvem estão trazendo. Você aprendeu sobre o caráter interdisciplinar da Ciência de Dados e sobre o ciclo de vida (CRISP-DM) que envolve o processo de análise de dados, incluindo aqueles que empregam de modelos de aprendizado de máquina. Por último, você tomou conhecimento de algumas das principais tarefas que podem ser realizadas pelo Aprendizado de Máquina e seus tipos e nas trilhas seguintes você aprenderá a criar diferentes modelos de aprendizado Supervisionado e Não Supervisionado.

# Para Saber Mais

* Que tal explorar alguns cases de aplicação do aprendizado de máquina dentro da sua área de interesse e compartilhar com os colegas? Acesse aqui \_\_\_.**Machine Learning and Data Science Applications in Industry** Disponível em: <https://github.com/ashishpatel26/Real-time-ML-Project> e explore alguns desses cases!
* Nesta curso você entender e aplicar vários modelos de aprendizado, começando pelos modelos de aprendizado supervisionado. Que tal entender mais sobre esses modelos? Acesse, então \_\_\_.**What is Machine Learning?** Disponível em: <https://c3.ai/introduction-what-is-machine-learning/supervised-learning/> Acesso em: 02 de Novembro de 2021. Esse texto trás e explica exmplos de aprendizado supervisionado para classificação e regressão de dados.

# Referências

Pete Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rüdiger Wirth [**CRISP DM 1.0**](http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf) **Step-by-step data mining guide** Disponível em: <http://meusite.mackenzie.br/rogerio/TIC2021S1$/T1/crisp-dm.pdf> Acesso em: 29 de Agosto de 2021.

Delua, Julianna **Supervised vs. Unsupervised Learning: What’s the Difference?** Disponível em: <https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning> Acesso em: 02 de Novembro de 2021.

Mwiti, Derrick **10 Real-Life Applications of Reinforcement Learning** Disponível em: <https://neptune.ai/blog/reinforcement-learning-applications> Acesso em: 02 de Novembro de 2021.

Sayad, Saed. **An Introduction to Data Science**. Disponível em: <http://www.saedsayad.com/>. Acesso em: 02 de Novembro de 2021.

Peikari, M., Salama, S., Nofech-Mozes, S. et al. **A Cluster-then-label Semi-supervised Learning Approach for Pathology Image Classification**. Sci Rep 8, 7193 (2018). <https://doi.org/10.1038/s41598-018-24876-0>

Kotu, Vijay; Deshpande, Balachandre **Data Science: concepts and practice**. 2nd ed. Cambridge, [England]: Morgan Kaufmann, c2019. E-book (570 p.) ISBN 9780128147627 (electronic bk.). Disponível em: <http://pergamum.mackenzie.br:8080/pergamumweb/vinculos/00003c/00003cef.jpg>.

Dhar, V. (2013). **Data science and prediction**. Communications of the ACM, 56(12), 64–73. <doi:10.1145/2500499>

Alpaydin, E. **Machine Learning** (The MIT Press Essential Knowledge). The MIT Press. 2019.

kelleher,John D.; Tierney, Brendan. **Data Science** (The MIT Press Essential Knowledge). The MIT Press. 2018.

\_\_\_.**What is Machine Learning?** Disponível em: <https://c3.ai/introduction-what-is-machine-learning/supervised-learning/> Acesso em: 02 de Novembro de 2021.

\_\_\_.**Machine Learning and Data Science Applications in Industry** Disponível em: <https://github.com/ashishpatel26/Real-time-ML-Project> Acesso em: 02 de Novembro de 2021.