

O que é ciência de dados?

Um guia para iniciantes em Ciência de Dados

Hemant Sharma

Fonte: <https://www.edureka.co/blog/what-is-data-science/>

À medida que o mundo embarcou na era dos *grandes dados* (big data), a necessidade para armazenamento de dados também cresceu. Este foi um dos principais desafios para as organizações até 2010. O foco principal até então estava em se construir *infraestrutura e soluções* para **armazenar** dados.

Atualmente, como as *plataformas de software de computação distribuída*¹ voltadas para processamento de grandes volumes de dados, como por exemplo a *Hadoop*² além de outras tecnologias, resolveram de forma bem-sucedida o *problema de armazenamento*, o foco agora se voltou para o **processamento** desses dados.

Nesta perspectiva, a Ciência de Dados (ou Data Science) é o *ingrediente* secreto atual. Todas as fantasias que se veem nos filmes de ficção científica de Hollywood podem realmente se transformar em realidade por meio da *data science*. A Ciência de Dados é o futuro da Inteligência Artificial (AI – Artificial Intelligence).

Portanto, é importante se entender *o que é ciência de dados* e como essa tecnologia pode *agregar valor* aos negócios.

Neste pequeno texto introdutório serão abordados os tópicos:

- A necessidade de ciência de dados;
- O que é ciência de dados;
- Qual a diferença entre *Inteligência de Negócios* (BI - Business Intelligence) e *Análise de Dados* (**Analytics**); e, por fim,
- O **Ciclo de vida** de ciência de dados, com o apoio de um *caso de uso*³.

Ao final deste texto, será possível entender o que é ciência de dados e qual o seu papel na *extração de insights*⁴ a partir de grandes e complexos conjuntos de dados (*big data*) que estão em nosso entorno via aplicativos e outros softwares.

Entendendo por que precisamos de Ciência de Dados

Tradicionalmente, os dados utilizados nos sistemas e aplicações são, em sua maioria, *estruturados* e de pequeno porte, os quais podem ser analisados utilizando-se ferramentas simples de *Inteligência de Negócios* (**BI**).

Entretanto, ao contrário dos dados utilizados nos sistemas tradicionais, que são principalmente estruturados, atualmente os dados se apresentam preferencialmente de forma *não-estruturada* ou *semiestruturada*.

¹ Um sistema de processamento distribuído ou paralelo é um sistema que interliga vários nós de processamento (computadores individuais, não necessariamente homogêneos) de maneira que um processo de grande consumo seja executado no nó "mais disponível", ou mesmo subdividido por vários nós. Conseguindo-se, portanto, ganhos óbvios nestas soluções: uma tarefa qualquer, se divisível em várias subtarefas pode ser realizada em paralelo.

Fonte: https://pt.wikipedia.org/wiki/Sistema_de_processamento_distribu%C3%ADdo

² Hadoop é uma plataforma de software de computação distribuída voltada para clusters e processamento de grandes volumes de dados.

³ Especificações de casos de uso são narrativas em texto, descrevendo a interação entre um usuário (humano ou máquina) e um sistema, não descrevendo como um software deverá ser construído, mas sim como ele deverá se comportar. Fonte: https://pt.wikipedia.org/wiki/Caso_de_uso.

⁴ Um insight é um acontecimento cognitivo que pode ser associado a vários fenômenos podendo ser sinônimo de compreensão, conhecimento, intuição. Fonte: <https://www.significados.com.br/insight/>

Dados estruturados são aqueles que estão organizados em uma estrutura rígida, a qual foi previamente planejada para armazená-los. Geralmente os dados estruturados estão organizados em forma de colunas e linhas, como em uma tabela ou planilha eletrônica, por exemplo.

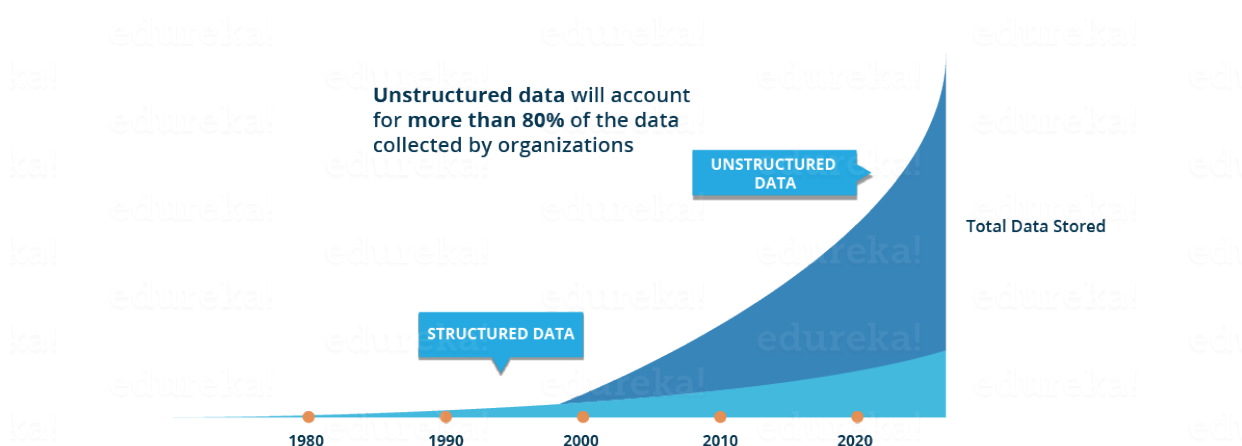
Já os **dados não-estruturados** são aqueles que são armazenados utilizando-se uma estrutura flexível e dinâmica ou sem uma organização definida. O exemplo mais comum de dado não-estruturado é um documento ou arquivo contendo imagens (gráficos e fotos) misturado com textos.

As *redes sociais*, que apresentam elevado volume de dados criados diariamente pelos usuários, representam outro exemplo de **dados não estruturados**.

Atualmente, mais de 80% do conteúdo digital gerado no mundo é do tipo **não estruturado**⁵ (Figura 1).

Os **dados semiestruturados** apresentam uma representação *heterogênea*, ou seja, possuem estrutura, mas esta é flexível. Esta representação de dados agrega um pouco dos formatos estruturado e não-estruturado em termos de benefícios. Facilita o controle por ter um pouco de estrutura, mas também permite uma maior flexibilidade.

Figura 1. Linha do tempo para tendência dos tipos de dados



Como pode ser visualizado na Figura 1, a projeção é que, *a partir de 2020, a maior parte* dos dados disponíveis serão *não-estruturados*. Esses dados são gerados de diferentes fontes, como registros financeiros, arquivos de texto, formulários multimídia, sensores e instrumentos. Em vista disso, as tradicionais ferramentas de BI não serão capazes de processar esse enorme volume e variedade de dados.

Portanto, serão necessárias ferramentas e algoritmos analíticos mais complexos e inovadores para poder processar, analisar e extrair conhecimentos (*insights*) relevantes.

Esta não é a única razão para a ciência de dados ter se tornado tão popular. A Ciência de Dados está sendo utilizada em vários setores. Alguns exemplos:

1. As organizações podem entender exatamente as necessidades de seus clientes a partir dos dados existentes, tais como histórico de navegação, histórico de compras, idade e renda do cliente. Sem dúvida, as organizações já possuíam todos esses dados anteriormente, mas agora com devido a grande quantidade e variedade de dados, as empresas podem *treinar* modelos com mais eficiência e recomendar produtos aos seus clientes com mais precisão, pois isto poderá atrair mais negócios à organização.
2. Para entender o papel da ciência de dados no processo de tomada de decisões, pode se utilizar um cenário diferente. E se um automóvel tivesse a inteligência necessária para

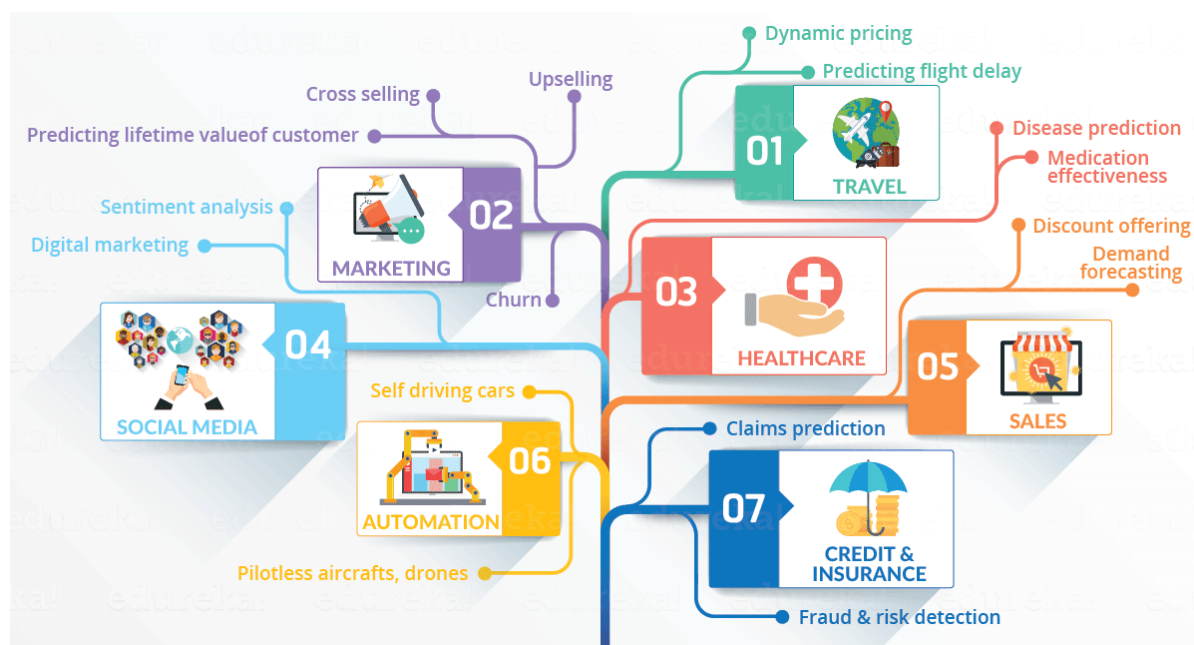
⁵ Fonte: <https://universidadetecnologia.com.br/dados-estruturados-e-nao-estruturados/>

levar as pessoas automaticamente para casa? Os *carros autônomos* coletam informações de dispositivos (sensores) em tempo real (tais como radares, câmeras e lasers), para criar um mapa de sua localização. Com base nesses dados, o *carro pode tomar decisões*, como por exemplo, quando acelerar ou frear, quando ultrapassar e realizar um retorno, utilizando-se de algoritmos de *aprendizado de máquina* (ML - *Machine Learning*).

3. A ciência de dados pode ser usada para realizar *análise preditiva*. Análise preditiva pode ser definida como *uma técnica analítica que utiliza dados, algoritmos e aprendizado de máquina para antecipar tendências e fazer projeções em diversas áreas*. Um clássico exemplo de análise preditiva é o de *previsão do tempo*. Informações sobre navios, aeronaves, radares, satélites são coletados e analisados para **construir modelos estatísticos**. Esses modelos estatísticos não apenas podem prever o clima, mas também ajudarão a antecipar a ocorrência de qualquer calamidade natural. Portanto, com esses dados as autoridades podem antecipar medidas e salvar vidas.

No infográfico da Figura 2 pode-se observar todos os domínios em que a Data Science está influenciando diretamente.

Figura 2. Áreas de atuação da Ciência de Dados



Após esta breve apresentação sobre as principais áreas de atuação da ciência de dados, pode-se definir, de forma geral, o que é ciência de dados.

O que é Ciência de Dados

O uso do termo *Ciência de Dados* é cada vez mais comum, mas o que exatamente significa esta expressão? Quais são as habilidades necessárias para se tornar cientista de dados? Qual a diferença entre BI e Ciência de Dados? Como as decisões e as previsões são realizadas em Ciência de Dados?

Estas são algumas das perguntas que serão respondidas e discutidas neste tópico.

Em primeiro lugar, vamos definir o que é ciência de dados. A Ciência de dados é a combinação de ferramentas, algoritmos e princípios de aprendizado de máquina (ML) com o objetivo de *descobrir padrões ocultos* a partir de *dados brutos*.

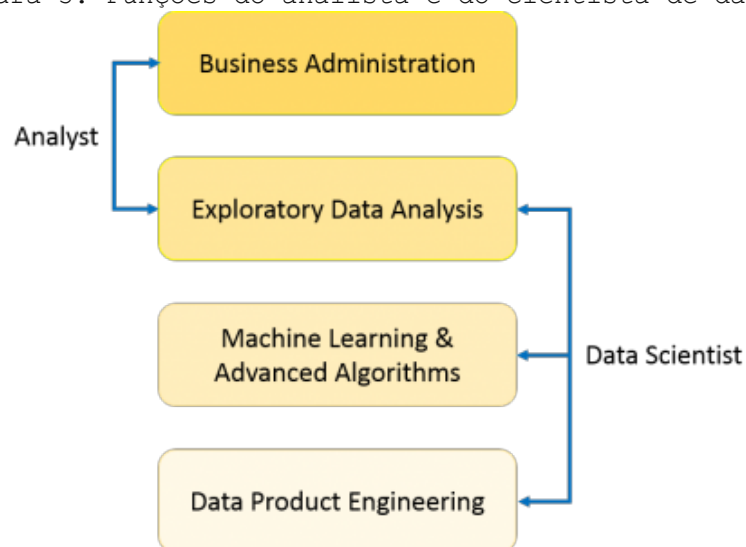
Portanto, como esta combinação de fatores se diferencia do que os estatísticos já vêm fazendo há anos? A resposta está na diferença entre ***explicar*** - tornar (uma ideia, situação ou problema) mais compreensível para alguém, descrevendo-a com mais detalhes ou revelando fatos ou ideias relevantes - e ***prever*** - dizer ou calcular que algo específico irá acontecer no futuro ou se será consequência disso.

Como pode-se observar na Figura 3, um *analista de dados* geralmente *explica* o que está acontecendo processando *dados históricos* - dados coletados sobre eventos e circunstâncias passados - pertencentes a um determinado assunto.

Por outro lado, o *cientista de dados* não apenas faz *análise exploratória* para descobrir informações, mas também se utiliza de vários algoritmos inovadores de aprendizado de máquina (ML) para *identificar a ocorrência de um evento específico no futuro*.

Um cientista de dados analisará os dados segundo diferentes perspectivas, às vezes sob pontos de vista desconhecidos anteriormente.

Figura 3. Funções do analista e do cientista de dados.



Portanto, a Ciência de Dados é utilizada principalmente para tomar decisões e realizar previsões, utilizando *análise causal preditiva*, *análise prescritiva* (ciência preditiva mais ciência de decisão) e aprendizado de máquina (ML).

- **Análise causal preditiva** - É um modelo utilizado para antecipar as probabilidades de ocorrer um evento específico no futuro. Por exemplo, se um banco empresta dinheiro a crédito, a possibilidade de os clientes efetuarem pagamentos futuros no prazo é uma preocupação da instituição financeira. Portanto, o banco pode criar um modelo para realizar análises preditivas utilizando o histórico de pagamentos do cliente para prognosticar se os pagamentos futuros serão pontuais ou não.
- **Análise prescritiva:** É um modelo que tem inteligência para tomar suas próprias decisões e a capacidade de modificá-las por meio de parâmetros dinâmicos. Este campo relativamente novo tem como objetivo principal fornecer *recomendações*. Em outras palavras, a análise prescritiva oferece recomendações específicas para alterar o futuro tendo em vista que este modelo não apenas prevê, mas sugere uma série de ações específicas e fornece visualizações dos possíveis resultados a serem alcançados. O melhor exemplo deste modelo é o **carro autônomo da Google**. Os dados coletados (de radares, sensores, etc.) por veículos comuns podem ser usados para *treinar* o comportamento dos carros autônomos. Desta forma, pode-se executar algoritmos nestes dados coletados para

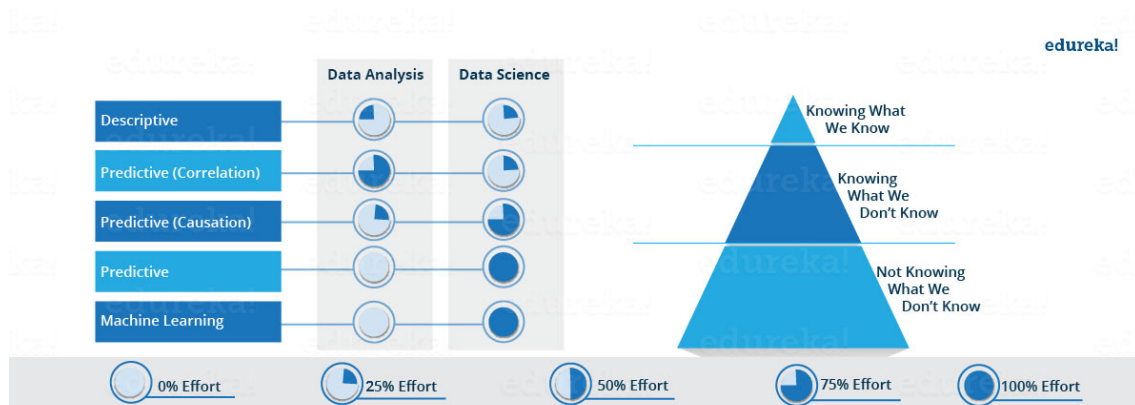
fornecer inteligência aos carros autônomos. Com isso, os carros podem tomar decisões sobre quando virar, qual caminho deve seguir, quando frear ou acelerar.

- **Aprendizado de máquina para fazer previsões.** Quando uma empresa financeira possui dados transacionais e precisa criar um modelo para produzir uma tendência futura, os algoritmos de aprendizado de máquina são a melhor aposta. Isso se enquadra no *paradigma de aprendizagem de máquina supervisionada*. É chamada de **supervisionada** porque os dados já existem e o algoritmo possui modelos preestabelecidos, isto é, o algoritmo já conhece as categorias definidas anteriormente. Com base nestes dados e algoritmos a organização pode treinar suas máquinas. Por exemplo, um modelo de detecção de fraude pode ser treinado utilizando o histórico de compras fraudulentas dos seus clientes.
- **Aprendizado de máquina para descoberta de padrões.** Quando não se possui os parâmetros-base para fazer as previsões, torna-se necessário descobrir padrões ocultos no conjunto de dados (dataset) para fazer previsões importantes. Essa é a característica de um *modelo não supervisionado*, pois inicialmente não se possui categorias predefinidas para se agrupar os dados. O algoritmo mais comum usado para a descoberta de padrões é o *clustering*. **Clustering**, ou análise de agrupamento de dados, é um conjunto de técnicas de prospecção de dados (*data mining*) que visa agrupar dados segundo o seu grau de semelhança.

Um exemplo de descoberta de padrões é o de uma companhia telefônica que necessita implementar uma rede instalando torres de telecomunicações em uma região. A empresa pode usar a técnica de agrupamento (cluster) para definir os locais de posicionamento das torres em que todos os usuários recebam o nível de sinal ideal.

Na Figura 4 podemos observar em que proporção as abordagens descritas anteriormente são diferentes na *Análise de Dados* e na *Ciência de Dados*. Como se pode ver na Figura 4, a **Análise de Dados** incorpora *análise descritiva e previsão* até um determinado ponto. Por outro lado, a **Ciência de Dados** compreende mais *análise preditiva (causal)* e *aprendizado de máquina*.

Figura 4. Diferenças entre Análise de Dados e Ciência de Dados.



Outra questão importante com relação à utilização ambígua entre os conceitos, é que, normalmente, a Ciência de Dados é confundida com Inteligência de Negócios (*Business Intelligence*, ou **BI**). Neste artigo, serão mostrados alguns contrastes concisos e compreensíveis entre os dois conceitos, os quais poderão colaborar para uma melhor compreensão e distinção entre os mesmos.

Business Intelligence (BI) vs. Data Science

O BI analisa basicamente dados passados (dados históricos) para compreender uma situação ou determinado evento *após a sua ocorrência* de forma a obter *insights* e apresentar tendências nos

negócios. O BI permite coletar dados de fontes externas e internas, prepará-los, realizar consultas e criar **dashboards**⁶ para responder a perguntas como “como está a análise de receita trimestral” ou “mostrar quais problemas de curto e/ou longo prazo nos negócios”. O BI pode avaliar o impacto de determinados eventos em curto prazo.

A Ciência de Dados é uma abordagem mais prospectiva, de forma exploratória, com foco na *análise de dados passados ou atuais e na previsão de resultados futuros* com o objetivo de tomar decisões com conhecimento. A Ciência de Dados busca responder *perguntas abertas* como “**quais**” e “**de que forma**” os eventos ocorrem.

O Quadro 1 apresenta um resumo das diferenças entre BI e Ciência de Dados

Quadro 1. BI e Ciência de Dados – diferenças.

Características	Inteligência de Negócios (BI)	Ciência de Dados
Fontes de Dados	Estruturada (normalmente SQL, algumas vezes Data Warehouse)	Estruturada e mão-estruturada (logs de dados, dados em nuvem, SQL, NoSQL, texto)
Abordagem	Estatística e Visualização	Estatística, Aprendizado de Máquina, Análise de Gráficos, Programação Neurolinguística (PNL)
Foco	Passado e Presente	Presente e Futuro
Ferramentas	Pentaho, Microsoft BI, QlikView, R	RapidMiner, BigML, Weka, R

Definido o que é Ciência de Dados, no próximo tópico será apresentado o conceito de *Ciclo de Vida de Ciência de Dados*.

Um erro comum cometido em projetos de ciência de dados é apressar-se a coleta e análise de dados, sem entender os requisitos ou mesmo compreender corretamente qual é o problema do negócio.

Portanto, para garantir o bom funcionamento de um projeto em ciência de dados, é importante que se compreenda e obedeça a todas as fases do seu ciclo de vida.

Ciclo de vida de Ciência de Dados

Na Figura 5 é exibida uma visão geral das principais fases do ciclo de vida de ciência de dados. Cada uma das fases do ciclo de vida será apresentada e discutida neste tópico.

Fase 1 - Descoberta



Antes de se iniciar um projeto de ciência de dados, deve-se compreender as diversas especificações, requisitos, prioridades e saber qual o orçamento necessário.


O fundamental é ser capaz de **fazer as perguntas certas**.

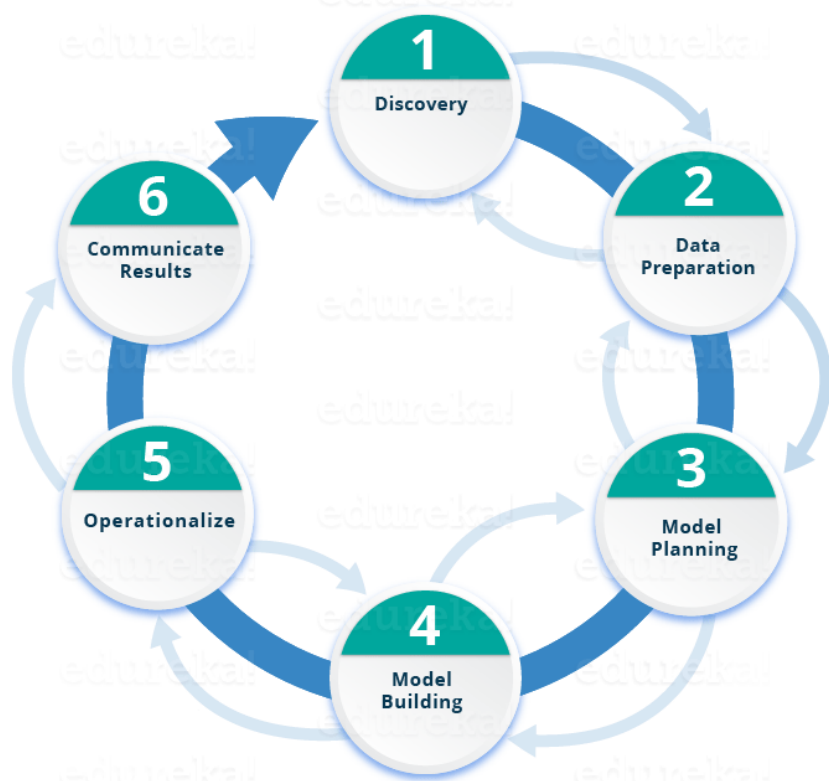
Nesta fase, avalia-se se o projeto possui os recursos necessários em termos de pessoas, tecnologia, tempo e dados para se realizá-lo.

Nesta etapa também é necessário buscar entender **qual é o problema do negócio** e formular as hipóteses iniciais (IH) para testá-lo.

Figura 5. Ciclo de vida de Ciência de Dados.

⁶ **Dashboards** são painéis que mostram métricas e indicadores importantes de forma visual, facilitando a compreensão das informações geradas. O objetivo dos **dashboards**, é possibilitar o monitoramento dos resultados distribuídos em diversos indicadores. Para chegar a esse patamar é preciso responder perguntas essenciais para ter as respostas desejadas.

Fonte: <http://marketingpordados.com/analise-de-dados/o-que-e-dashboard->



Fase 2 Preparação dos dados



Nesta fase, o uso de uma área de teste analítica isolada (**analytical sandbox**) - um ambiente separado de testes que faz parte da arquitetura/repositório geral de todos os dados brutos) na qual seja possível executar análises durante toda a duração do projeto é um recurso importante.

Este ambiente é utilizado nesta etapa para explorar, pré-processar e acondicionar os dados antes da modelagem. Será realizado o processo denominado ETLT (Extrair, Transformar, Carregar e Transformar) para obter dados no sandbox.

Este fluxo de análise estatística é mostrado na Figura 6.

Pode-se usar a *linguagem R* para executar a *limpeza*, *transformação* e *visualização* de dados.

"R" é um ambiente computacional e uma linguagem de programação voltada especificamente para a manipulação, análise e visualização gráfica de dados⁷.

Figura 6. Fluxo de análise estatística.



"*Limpeza e organização de dados*" envolve a preparação dos dados com o objetivo de facilitar sua organização e posterior leitura pelos aplicativos. Muitas vezes, o processo de coleta e

⁷ [https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))

tabulação envolve erros o que gera imprecisões na leitura dos dados pelos softwares estatísticos. O processo de limpeza e transformação é utilizado para remover essas imprecisões e tornar os dados mais precisos.

Esse processo também auxilia na identificação dos **outliers** e ajuda a estabelecer o relacionamento entre as variáveis. Um **outlier** é um valor que foge da normalidade e que pode (e provavelmente irá) causar anomalias nos resultados obtidos por meio de algoritmos e aplicativos de análise.

Somente após a limpeza e organização, é possível realizar uma análise exploratória nos dados (AED). Em estatística, *análise exploratória de dados* (AED) é uma abordagem de análise dos **datasets** de modo a resumir suas principais características, frequentemente utilizando métodos visuais.

Fase 3 - Planejamento do modelo



Nesta etapa do ciclo de vida são definidos os métodos e as técnicas para extrair os relacionamentos entre as variáveis.

Esses relacionamentos servirão de base para a construção dos algoritmos que serão implementados na construção do modelo (Fase 4).

Para tanto, será utilizada a Análise Exploratória de Dados (EDA) utilizando diversas fórmulas estatísticas e ferramentas de visualização.

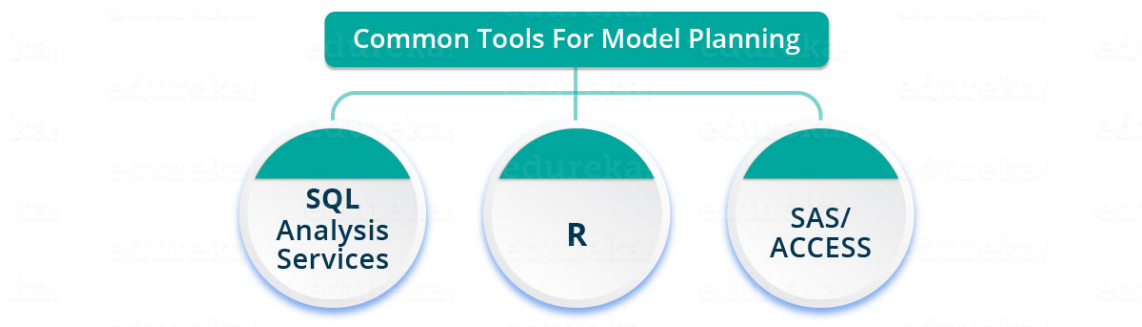
Algumas das ferramentas utilizadas para o planejamento de modelos, são: a linguagem R, os Serviços de Análise SQL e o SAS/Access. Tais ferramentas são apresentadas na Figura 7.

1. A linguagem R possui um conjunto completo de recursos para modelagem e fornece um bom ambiente para a construção de modelos interpretativos.
2. Os serviços SQL Analysis podem realizar análises no banco de dados usando funções de *mineração de dados* e *modelos preditivos básicos*.
3. SAS / ACCESS pode ser usado para acessar dados do **Hadoop** e é utilizado para criar *diagramas de fluxo de modelo* repetíveis e reutilizáveis.

Embora muitas ferramentas estejam disponíveis no mercado, a linguagem R é a ferramenta mais utilizada pelos cientistas de dados por ser gratuita e contar com grande número de colaboradores na criação dos **pacotes**.

Pacotes em R são funções e conjuntos de dados (datasets) desenvolvidos pela comunidade que ampliam o poder da linguagem pois melhoram ou adicionam novas funcionalidades.

Figura 7. Ferramentas para planejamento de modelo.



Assim, após compreender a natureza dos dados coletados e decidir quais algoritmos poderão ser usados, na Fase 4 (construção do modelo), o algoritmo será implementado e os modelos construídos.

Fase 4 - Construção do modelo

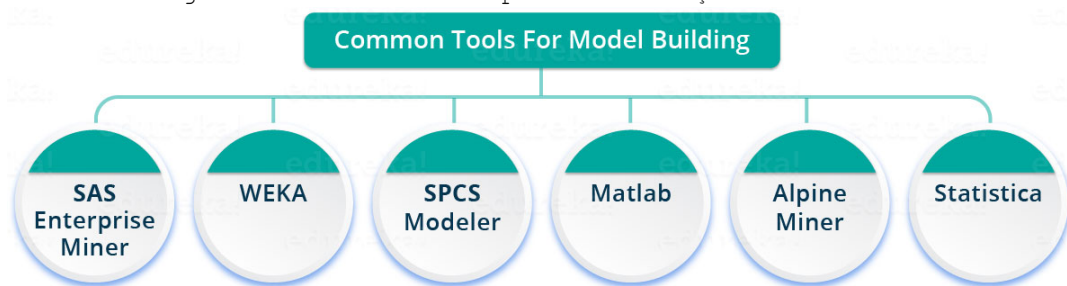


Nesta etapa do ciclo de vida, do conjunto de dados inicial são criados dois **datasets**: um para fins de treinamento e outro para testes.

Nesta etapa o cientista de dados deverá verificar se as ferramentas existentes no seu ambiente de desenvolvimento serão suficientes para executar os modelos gerados ou se será necessário um ambiente mais robusto, utilizando recursos de processamento rápido e processamento paralelo, por exemplo. Várias técnicas de machine learning tais como, classificação, associação e clustering (agrupamento) são avaliadas para construir o modelo final.

Algumas ferramentas para construção de modelos podem ser visualizadas na Figura 8.

Figura 8. Ferramentas para construção de modelos.



Fase 5 – Operacionalização




Na fase de operacionalização são entregues os relatórios finais e documentos técnicos, reuniões são realizadas para a transmissão do conjunto de instruções ou documento para o desenvolvimento das tarefas (briefings), além da implementação dos códigos desenvolvidos.

Em algumas situações, um projeto piloto também é implementado em um ambiente de produção em tempo real. O projeto piloto fornecerá uma imagem clara em pequena escala do desempenho esperado e das restrições antes da implantação completa.

Fase 6 - Divulgação dos resultados

Após a operacionalização do projeto, é importante avaliar se o mesmo conseguiu atingir o objetivo definido na primeira fase do ciclo de vida.

	<p>Portanto, nesta última etapa, devem ser informadas as principais apurações feitas e os resultados obtidos devem ser divulgados às partes interessadas (stackholders).</p> <p>Além disso, nesta fase também deve ser determinado se o projeto foi um sucesso ou um fracasso, com base nos critérios estabelecidos na Fase 1.</p>
---	--

Para explicar as diferentes fases do ciclo de vida de um projeto de ciência de dados percorridos anteriormente, será utilizado um *estudo de caso*⁸ buscando prever a ocorrência de diabetes e quais medidas tomar com antecedência para evitá-la.

⁸ Um **estudo de caso** é uma metodologia de pesquisa comumente usada em ciências sociais. É uma estratégia de pesquisa e uma investigação empírica que investiga um fenômeno em seu contexto da vida real. Os estudos de caso são baseados em uma investigação aprofundada de um único indivíduo, grupo ou evento para explorar as causas dos princípios subjacentes. Uma pesquisa de estudo de caso pode ser estudos de caso único ou múltiplo, inclui evidência quantitativa, conta com várias fontes de evidência e se beneficia do desenvolvimento prévio de proposições teóricas.

Estudo de caso: prevenção de diabetes

E se pudéssemos prever a ocorrência de diabetes e tomar as medidas apropriadas com antecedência para evitá-la?

Nesse caso de uso, buscaremos prever a ocorrência de diabetes usando todo o ciclo de vida discutido anteriormente. Serão percorridas as várias etapas.

Passo 1:

Primeiro, serão coletados os dados com base no histórico médico do paciente, conforme discutido na Fase 1. Os dados da amostra estão na Tabela 1.

Tabela 1. Dados brutos da amostra.

	npreg;glu;bp;skin;bmi;ped;age,income
1;	6;148;72;35;33.6;0.627;50
2;	1;85;66;29;26.6;0.351;31
3;	1;89;80;23;28.1;0.167;21
4;	3;78;50;32;31;0.248;26
5;	2;197;70;45;30.5;0.158;53
6;	5;166;72;19;25.8;0.587;51
7;	0;118;84;47;45.8;0.551;31
8;	1;103;30;38;43.3;0.183;33
9;	3;126;88;41;39.3;0.704;27
10;	9;119;80;35;29;0.263;29
11;	1;97;66;15;23.2;0.487;22
12;	5;109;75;26;36;0.546;60
13;	3;88;58;11;24.8;0.267;22
14;	10;122;78;31;27.6;0.512;45
15;	4;97;60;33;24;0.966;33
16;	9;102;76;37;32.9;0.665;46
17;	2;90;68;42;38.2;0.503;27
18;	4;111;72;47;37.1;1.39;56
19;	3;180;64;25;34;0.271;26
20;	7;106;92;18;39;0.235;48
21;	9;171;110;24;45.4;0.721;54

Como se pode observar, na figura da Tabela 1, temos vários *atributos* nesta *amostra*⁹.

Atributos são as características de um objeto. Essas características também são conhecidas como **variáveis**, utilizando o exemplo da amostra mostrada na Tabela 1, são os dados colhidos no histórico médico do paciente, tais como o número de gravidezes, concentração de glicose no sangue, índice de massa corporal, dentre outras informações.

Para se estabelecer relação entre os nomes das variáveis coletadas e seu conteúdo, é necessária a definição de um dicionário de dados dos seus atributos.

Um **dicionário de dados** (do inglês *data dictionary*) é uma coleção de *metadados* que contém definições e representações de cada elementos de dados (variável).

Os principais atributos definidos para a amostra são:

⁹ **Amostragem** é o processo de selecionar um grupo de indivíduos de uma população, a fim de estudar e caracterizar a população total.

npreg - número de gravidezes
glu - concentração plasmática de glicose
bp - pressão arterial
skin - espessura das dobras cutâneas do tríceps
bmi - IMC - Índice de Massa Corporal
ped - função de pedigree do diabetes
age - idade
income – renda

Passo 2:

Após obter os dados de uma determinada fonte (um questionário, por exemplo) e colocá-los no formato de colunas ou tabelas (tabulá-los), é necessário limpá-los e organizá-los para realizar a *análise exploratória dos dados* (AED).

Observando a Tabela 1, cujo conteúdo pode ser definido como os "**dados brutos**" (**raw data**), verifica-se que os dados coletados apresentam diversas inconsistências, como valores ausentes, colunas em branco, valores não esperados e formato de dados incorreto que precisam ser limpos.

Na figura da Tabela 2, organizou-se os dados em uma "tabela padrão" (no caso, utilizando-se de uma planilha no aplicativo Microsoft Excel ou LibreOffice Calc) com seus diferentes atributos - tornando-os mais *estruturados* e **visíveis**.

Tabela 2. Dados planilhados com inconsistências.

	npreg	glu	bp	skin	bmi	ped	age	income
1	6	148	72	35	33.6	0.627	50	
2	1	85	66	29	26.6	0.351	31	
3	1	89	6600	23	28.1	0.167	21	
4	3	78	50	32	31	0.248	26	
5	2	197	70	45	30.5	0.158	53	
6	5	166	72	19	25.8	0.587	51	
7	0	118	84	47	45.8	0.551	31	
8	one	103	30	38	43.3	0.183	33	
9	3	126	88	41	39.3	0.704	27	
10	9	119	80	35	29	0.263	29	
11	1	97	66	15	23.2	0.487	22	
12	5	109	75	26	36	0.546	60	
13	3	88	58	11	24.8	0.267	22	
14	10	122	78	31	27.6	0.512	45	
15	4		60	33	24	0.966	33	
16	9	102	76	37	32.9	0.665	46	
17	2	90	68	42	38.2	0.503	27	
18	4	111	72	47	37.1	1.39	56	
19	3	180	64	25	34	0.271	26	
20	7	106	92	18		0.235	48	
21	9	171	110	24	45.4	0.721	54	

Estes dados apresentam algumas inconsistências:

- Na coluna **npreg**, "one" está escrito com palavras, e deveria estar na forma numérica como 1.
- Na coluna **bp**, um dos valores é 6600, o que é impossível (pelo menos para humanos), pois o bp não pode atingir um valor tão grande.
- Como se pode observar, a coluna **income** está totalmente em branco e não faz sentido termos *renda para prever diabetes*. Portanto, é redundante tê-la aqui e a mesma deve ser removida da tabela.

- Portanto, deverá ser realizada uma limpeza e um pré-processamento¹⁰ nestes dados removendo os valores discrepantes, preenchendo os valores nulos e normalizando o tipo de dados. Deve ser observado que estas são ações da segunda etapa do ciclo de vida, que, notadamente, se intitula *pré-processamento de dados*.

Finalmente, obtém-se os *dados limpos*, conforme mostrado na Tabela 3, que poderão ser utilizados para análise.

Tabela 3. Dados da amostra já consistentes.

	npreg	glu	bp	skin	bmi	ped	age
1	6	148	72	35	33.6	0.627	50
2	1	85	66	29	26.6	0.351	31
3	1	89	80	23	28.1	0.167	21
4	3	78	50	32	31	0.248	26
5	2	197	70	45	30.5	0.158	53
6	5	166	72	19	25.8	0.587	51
7	0	118	84	47	45.8	0.551	31
8	1	103	30	38	43.3	0.183	33
9	3	126	88	41	39.3	0.704	27
10	9	119	80	35	29	0.263	29
11	1	97	66	15	23.2	0.487	22
12	5	109	75	26	36	0.546	60
13	3	88	58	11	24.8	0.267	22
14	10	122	78	31	27.6	0.512	45
15	4	97	60	33	24	0.966	33
16	9	102	76	37	32.9	0.665	46
17	2	90	68	42	38.2	0.503	27
18	4	111	72	47	37.1	1.39	56
19	3	180	64	25	34	0.271	26
20	7	106	92	18	39	0.235	48
21	9	171	110	24	45.4	0.721	54

Passo 3:

Nesta etapa, serão realizadas algumas análises, tal como apresentadas anteriormente, na Fase 3.

- Primeiramente, os dados são carregados no "sandbox analítico" (ambiente virtual de proteção analítica) e neles serão aplicadas diferentes funções estatísticas. A linguagem R possui pacotes de funções como *describe*, que fornece o número de valores ausentes e valores exclusivos. Também pode-se utilizar a função *summary* (de resumo), que dará informações estatísticas, tais como *média*, *mediana*, *intervalo* e *valores mínimo e máximo*.
- Na sequência serão utilizadas *técnicas de visualização* de dados tais como *histogramas*, *gráficos de linhas* e diagramas de caixas (*box plots*) para se obter uma boa ideia da *distribuição dos dados* (Figura 9).

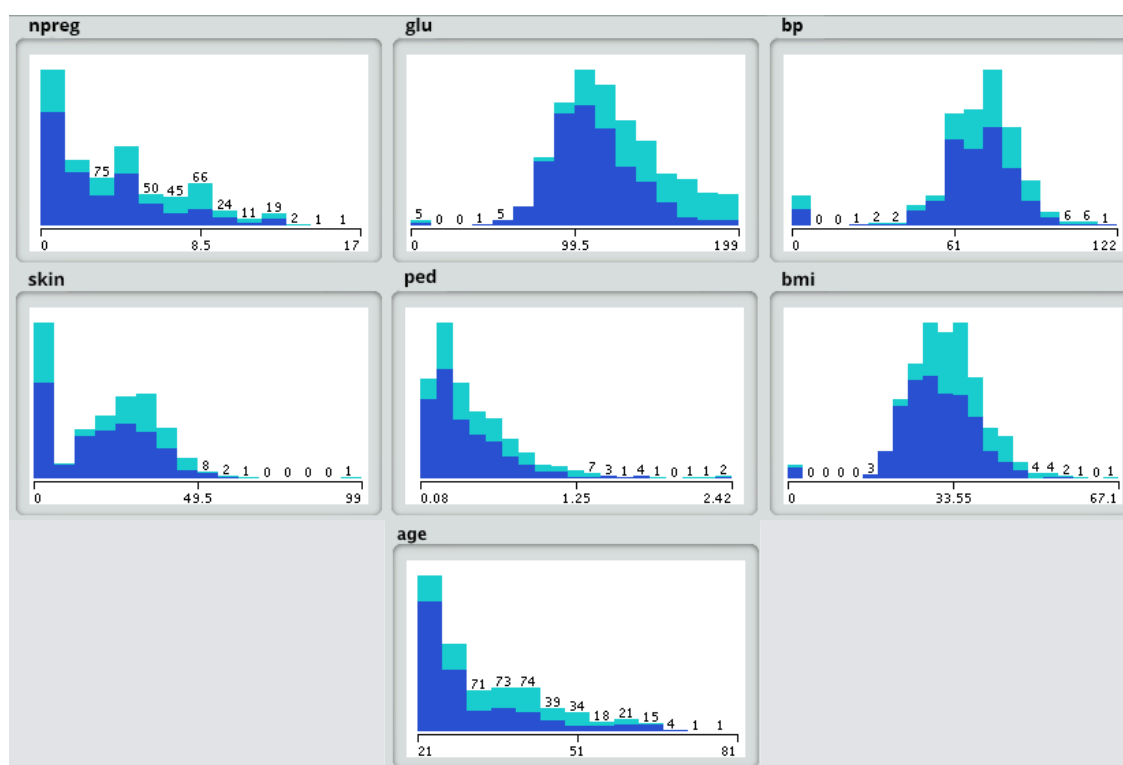
Passo 4:

Com base nas ideias (*insights*) obtidas na etapa anterior, a melhor opção para esse tipo de problema neste estudo de caso é a *árvore de decisão*.

¹⁰ O **pré-processamento** é um conjunto de atividades que envolvem preparação, organização e estruturação dos dados. Trata-se de uma etapa fundamental que precede a realização de análises e previsões. Essa etapa é de grande importância, pois será determinante para a qualidade final dos dados que serão analisados.

Fonte: <https://www.datageeks.com.br/pre-processamento-de-dados/>.

Figura 9. Gráficos para visualização dos dados.



Uma árvore de decisão é um mapa dos possíveis resultados de uma série de escolhas relacionadas⁹. Permite que um indivíduo ou organização compare possíveis ações com base em seus *custos, probabilidades e benefícios*.

Uma árvore de decisão geralmente começa com um único nó, que se divide em possíveis resultados. Cada um desses resultados leva a nós adicionais, que se ramificam em outras possibilidades. Assim, cria-se um gráfico em forma de árvore.

Existem *três tipos de nós*: **nós de probabilidade**, **nós de decisão** e **nós de término**. O nó de probabilidade, representado por um círculo, mostra as probabilidades de certos resultados. Um nó de decisão, representado por um quadrado, mostra uma decisão a ser tomada, e um nó de término mostra o resultado de um caminho de decisão.

Como já temos os principais atributos para análise como **npreg** (número de gravidezes), **bmi** (IMC), etc., usaremos a **técnica de aprendizado supervisionado**¹⁰ para criar um modelo.

- Neste estudo de caso, será utilizada a árvore de decisão porque esta técnica leva em consideração todos os atributos (colunas) de uma só vez, como os que têm um **relacionamento linear** e os que têm um **relacionamento não linear**. No nosso caso, temos uma *relação linear entre npreg e age*, e uma *relação não linear entre npreg e ped*.
- Os modelos de árvore de decisão também são muito robustos, pois podemos usar as diferentes combinações de atributos para criar várias árvores e, finalmente, implementar aquela com a máxima eficiência.

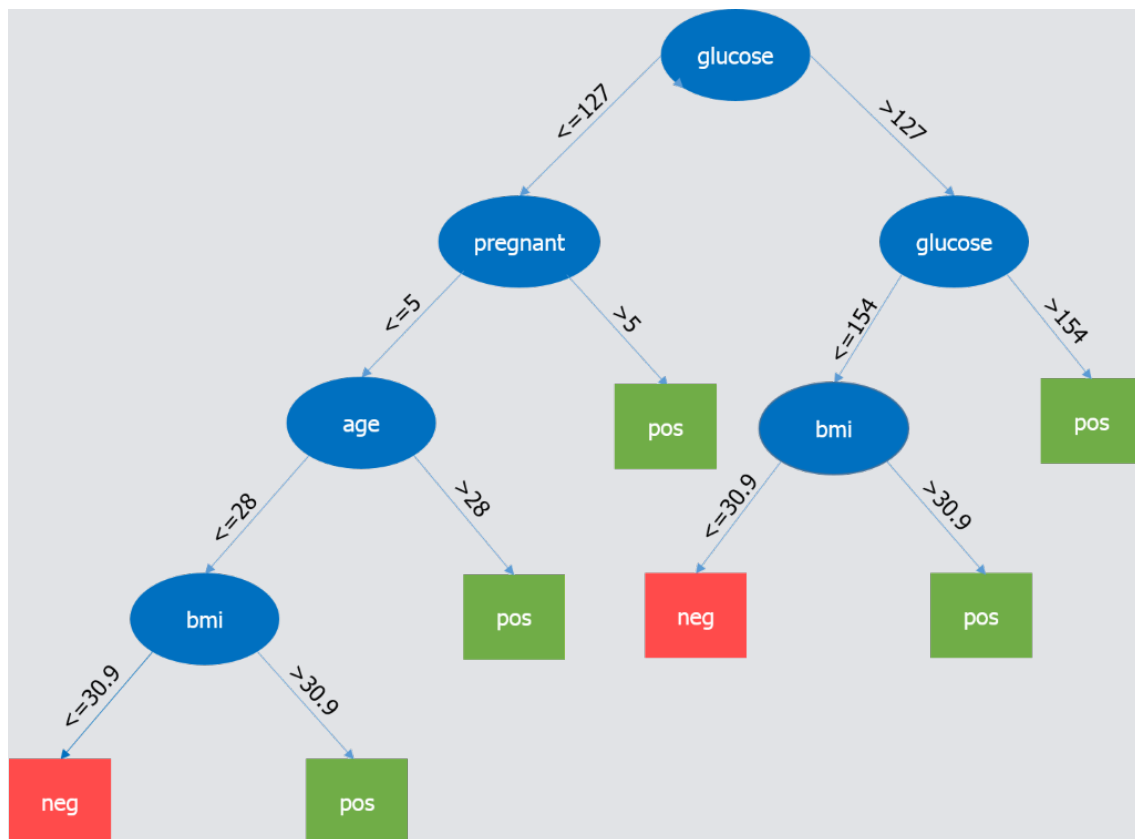
A árvore de decisão resultante pode ser vista na Figura 10.

Na árvore de decisão resultante, o parâmetro mais importante é o nível de **glicose (glucose)**, por isso esta variável (ou atributo ou coluna) é o **nó raiz**.

Como pode ser observado na Figura 10, **nó raiz** e o seu valor determinam o próximo parâmetro importante a ser obtido. Esse processo continua até obtermos o resultado em termos de *pos* ou *neg*. **Pos** significa que a tendência de ter diabetes é **positiva** e **neg** significa que a tendência de ter diabetes é **negativa**.

Para se aprofundar mais sobre implementação da árvore de decisão, consulte o blog [How To Create A Perfect Decision Tree](#).

Figura 10. Árvore de decisão resultante.



Passo 5:

Nesta fase, um pequeno projeto piloto será desenvolvido para verificar se os resultados obtidos são os adequados. Também serão procuradas restrições de desempenho, se houverem. Se os resultados não forem precisos, é preciso replanejar e reconstruir o modelo.

Passo 6:

Depois de executar o projeto com sucesso, a saída deverá ser compartilhada para implantação completa.

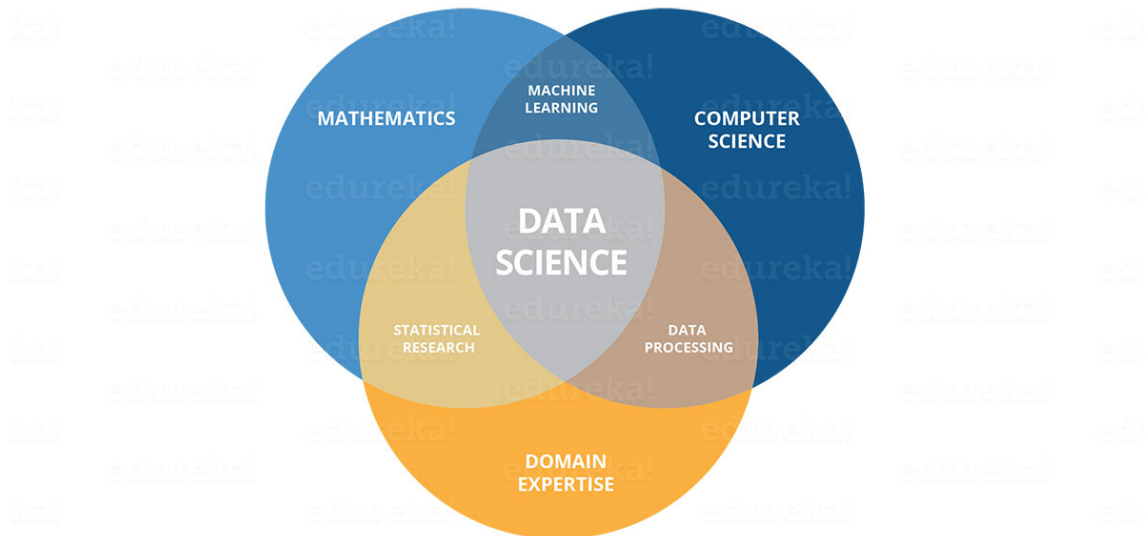
A Função de Cientista de Dados

Ser um cientista de dados é mais fácil dizer do que fazer.

Sendo assim, o que é necessário para ser um cientista de dados?

Um cientista de dados precisa apresentar habilidades em basicamente três áreas principais, como mostrado na Figura 11.

Figura 11. Habilidades necessárias para um cientista de dados



Como se pode observar na Figura 11, para se tornar um cientista de dados, o profissional precisa possuir *habilidades técnicas* (**hard skills**) e *habilidades sociocomportamentais* (**soft skills**).

Conhecimento em estatística e matemática para analisar e visualizar dados também são requisitos básicos necessários.

Desnecessário dizer que **Machine Learning** (*ML - Aprendizado de Máquina*) forma o coração da Ciência de Dados e exige que o cientista também tenha conhecimento dessas técnicas.

Além disso, um entendimento sólido da área em que se está trabalhando para entender claramente os problemas de negócios também é uma necessidade.

Com todas essas características, o cientista de dados também deve ser capaz de implementar diversos tipos de algoritmos, o que exige boas habilidades de codificação.

Por fim, após tomar decisões importantes sobre o projeto, é importante que os resultados obtidos sejam adequadamente entregues às partes interessadas (**stackholders**).

Portanto, uma boa capacidade de comunicação definitivamente adicionará muitos pontos às habilidades sociocomportamentais de qualquer cientista de dados.

Em conclusão, não será errado dizer que o futuro pertence aos cientistas de dados. Existe atualmente uma demanda genérica por cientistas de dados. Nisso se englobam engenheiros de dados, estatísticos, cientistas da computação etc.

Atualmente, o mercado e a academia começaram a moldar o que de fato seria o cientista de dados¹¹. Mais e mais dados fornecerão oportunidades para conduzir as principais decisões de negócios. Em breve, isso mudará a maneira como vemos o mundo inundado de dados ao nosso redor.

Assim, um cientista de dados deve ser altamente qualificado e motivado para resolver os problemas mais complexos.

¹¹ Fonte: <https://exame.abril.com.br/carreira/cientista-de-dados-a-profissao-do-futuro-continua-em-alta/>