ROTEIRO

O que é Ciência de Dados?

A ciência de dados, ou data science, é o estudo dos dados para **extração de dados significativos** para os negócios. É um tema quente entre profissionais qualificados e organizações que se concentram na coleta de dados e na elaboração de interpretações. Ela é uma abordagem multidisciplinar que **combina princípios e práticas das áreas de matemática, estatística, inteligência artificial e engenharia da computação** para **analisar grandes quantidades de dados**.

Qual a importância?

A ciência de dados é importante porque combina ferramentas, métodos e tecnologia para gerar significado com base em dados. As organizações modernas são inundadas com dados; há uma proliferação de dispositivos que podem coletar e armazenar informações automaticamente. Sistemas online e portais de pagamento capturam mais dados nas áreas de comércio eletrônico, medicina, finanças e todos os outros aspectos da vida humana. Temos dados de texto, áudio, vídeo e imagem disponíveis em grandes quantidades.

Além disso tudo, a ciência de dados está revolucionando o modo como as empresas operam. Muitas empresas, independentemente do porte, precisam de uma estratégia robusta de ciência de dados para impulsionar o crescimento e manter uma vantagem competitiva. Alguns dos principais benefícios incluem:

Descobrir padrões transformadores desconhecidos - permite que as empresas descubram novos padrões e relacionamentos que têm o potencial de transformar a organização. Ela pode revelar alterações de baixo custo no gerenciamento de recursos para obter o máximo impacto nas margens de lucro.

Inovar novos produtos e soluções - A ciência de dados pode revelar falhas e problemas que, de outra forma, passariam despercebidos. Mais insights sobre decisões de compra, feedback de clientes e processos de negócios podem impulsionar a inovação em operações internas e soluções externas.

Otimização em tempo real - É muito desafiadora para as empresas, especialmente as de grande porte, responder às mudanças nas condições em tempo real. Isso pode causar perdas significativas ou interrupções na atividade empresariais. A ciência de dados pode ajudar as empresas a prever mudanças e reagir de maneira ideal a diferentes circunstâncias.

Visto isto, também é significativa a ajuda de uma plataforma de ciência de dados pois assim como já foi dito anteriormente, para auxiliar os cientistas de dados, utiliza armazenamento automático de informações:

Onde este **reduz a redundância e impulsiona a inovação**, produz resultados completos, como o compartilhamento de **códigos**, **resultados** e **relatórios**. Ele remove gargalos no fluxo de trabalho, simplificando o gerenciamento e incorporando as melhores práticas.

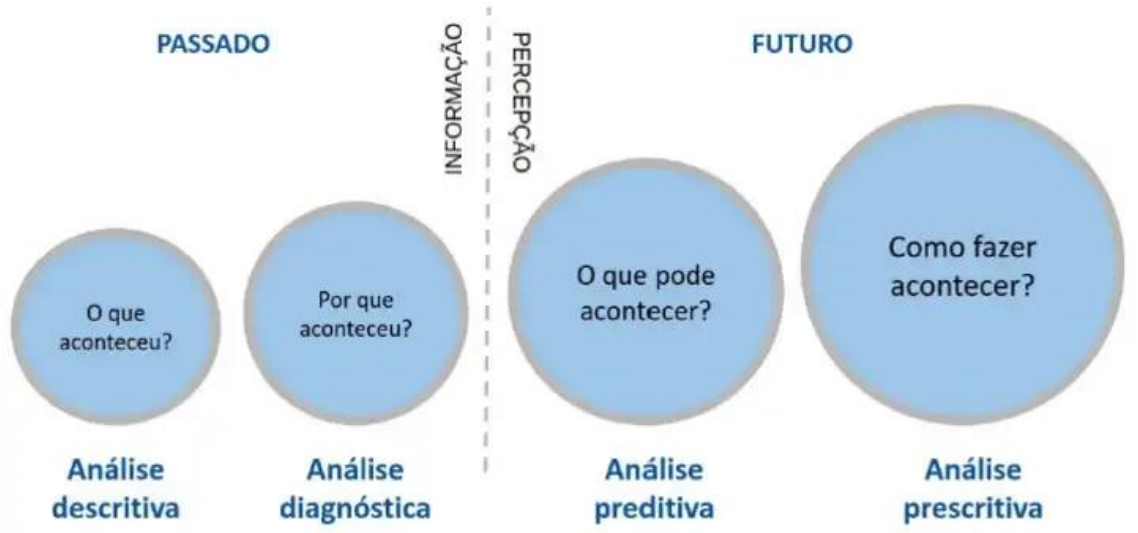
O que faz um cientista de dados?

O profissional de data science tem forte participação no rumo que determinado negócio precisa tomar. Com base na análise de dados, o cientista de dados precisa tentar ter uma **previsão do futuro com base matemática** e quais serão as ações necessárias para alterar (se preciso) a direção daquela empresa.

A função e o trabalho diário de um cientista de dados variam de acordo com o tamanho e os requisitos da organização. Embora eles normalmente sigam o processo de ciência de dados, os detalhes podem variar. Em equipes maiores de ciência de dados, um cientista de dados pode trabalhar com outros analistas, engenheiros, especialistas em machine learning e técnicos de estatísticas para garantir que o processo de ciência de dados seja seguido de ponta a ponta e que as metas de negócios sejam alcançadas.

As funções de um cientista de dados podem incluir o **desenvolvimento de estratégias para analisar dados**, **preparar** dados para análise, **explorar**, analisar e **visualizar dados que sejam úteis a um determinado contexto**, construir modelos com dados usando linguagens de programação, como Python e R, e implementar modelos em aplicativos.

Análises da Ciência de Dados?



A ciência de dados é usada para estudar dados de quatro maneiras principais:

Análise descritiva - A análise descritiva analisa os dados para obter insights sobre o que aconteceu ou o que está acontecendo no ambiente de dados. Ela é caracterizada por visualizações de dados, como gráficos de pizza, gráficos de barras, gráficos de linhas, tabelas ou narrativas geradas.

Análise diagnóstica - A análise diagnóstica é uma análise aprofundada ou detalhada de dados para entender por que algo aconteceu. Ela é caracterizada por técnicas como drill-down, descoberta de dados, mineração de dados e correlações. Várias operações e transformações de dados podem ser realizadas em um determinado conjunto de dados para descobrir padrões exclusivos em cada uma dessas técnicas.

Análise preditiva - A análise preditiva usa dados históricos para fazer previsões precisas sobre padrões de dados que podem ocorrer no futuro. Ela é caracterizada por técnicas como machine learning, previsão, correspondência de padrões e modelagem preditiva.

Em cada uma dessas técnicas, os computadores são treinados para fazer engenharia reversa de conexões de causalidade nos dados.

Análise prescritiva - A análise prescritiva leva os dados preditivos a um novo patamar. Ela não só prevê o que provavelmente acontecerá, mas também sugere uma resposta ideal para esse resultado. Ela pode analisar as potenciais implicações de diferentes escolhas e recomendar o melhor plano de ação. A análise prescritiva usa análise de gráficos, simulação, processamento de eventos complexos, redes neurais e mecanismos de recomendação de machine learning.

Processos da Ciência de Dados

O: Obter dados - Os dados podem ser pré-existentes, recém-adquiridos ou um repositório de dados que pode ser baixado da Internet.

S: Suprimir dados - A supressão de dados, ou limpeza de dados, é o processo de padronização dos dados de acordo com um formato predeterminado. Alguns exemplos de supressão de dados são: Alterar todos os valores de data para um formato padrão comum, corrigir erros de ortografia ou espaços adicionais. corrigir imprecisões matemáticas ou remover vírgulas de números grandes.

E: Explorar dados - A exploração de dados é uma análise de dados preliminar que é usada para planejar outras estratégias de modelagem de dados. Os cientistas de dados obtêm uma compreensão inicial dos dados usando estatísticas descritivas e ferramentas de visualização de dados. Em seguida, eles exploram os dados para identificar padrões interessantes que podem ser estudados ou acionados.

M:Modelar dados - Os algoritmos de software e machine learning são usados para obter insights mais profundos, prever resultados e prescrever o melhor plano de ação.

N: Interpretar resultados - Os cientistas de dados trabalham em conjunto com analistas e empresas para converter insights de dados em ação. Eles fazem diagramas, gráficos e tabelas para representar tendências e previsões. A sumarização de dados ajuda as partes interessadas a entender e implementar os resultados de forma eficaz.

Técnicas mais utilizadas

Classificação - Classificação é a ordenação de dados em grupos ou categorias específicos. Os computadores são treinados para identificar e classificar dados. Conjuntos de dados conhecidos são usados para criar algoritmos de decisão em um computador que processa e categoriza rapidamente os dados. Por exemplo:

Classificar produtos como populares ou não populares·

Classificar as aplicações de seguro como de alto risco ou baixo risco·

Classificar comentários de mídias sociais em positivos, negativos ou neutros.

Regressão - A regressão é o método de encontrar uma relação entre dois pontos de dados aparentemente não relacionados. A conexão geralmente é modelada em torno de uma fórmula matemática e representada como um gráfico ou curvas. Quando o valor de um ponto de dados é conhecido, a regressão é usada para prever o outro ponto de dados. Por exemplo:

A taxa de propagação de doenças transmitidas pelo ar.

A relação entre a satisfação do cliente e o número de funcionários.

A relação entre o número de quartéis de bombeiros e o número de feridos em decorrência de um incêndio em um determinado local.

Clustering - Clustering é o método de agrupar dados intimamente relacionados para procurar padrões e anomalias. O clustering é diferente da classificação porque os dados não podem ser classificados com precisão em categorias fixas. Portanto, os dados são agrupados em relações mais prováveis. Novos padrões e relações podem ser descobertos com o clustering. Por exemplo:

Agrupar clientes com comportamento de compra semelhante para melhorar o atendimento ao cliente.

Agrupar o tráfego de rede para identificar padrões de uso diário e identificar um ataque à rede mais rapidamente.

Agrupar artigos em diversas categorias de notícias diferentes e usar essas informações para encontrar conteúdo de notícias falsas.

O princípio básico por trás das técnicas de ciência de dados - Embora os detalhes variem, os princípios subjacentes por trás dessas técnicas são:

Ensinar uma máquina a classificar dados com base em um conjunto de dados conhecido. Por exemplo, palavras-chave de amostra são fornecidas ao computador com seus respectivos valores de classificação. “Feliz” é positivo, enquanto “Ódio” é negativo.

Fornecer dados desconhecidos à máquina e permitir que o dispositivo classifique o conjunto de dados de forma independente.

Permitir imprecisões de resultados e lidar com o fator de probabilidade do resultado.

Uso em Tecnologias

Os profissionais de ciência de dados trabalham com tecnologias complexas, como:

Inteligência artificial: modelos de machine learning e software relacionado são usados para análises preditivas e prescritivas.

Computação em nuvem: as tecnologias de nuvem deram aos cientistas de dados a flexibilidade e a capacidade de processamento necessárias para análise de dados avançada.

Internet das Coisas: IoT refere-se a vários dispositivos que podem se conectar automaticamente à Internet. Esses dispositivos coletam dados para iniciativas de ciência de dados. Eles geram grandes quantidades de dados que podem ser usados para mineração de dados e extração de dados.

Computação quântica: computadores quânticos podem fazer cálculos complexos em alta velocidade. Cientistas de dados qualificados os usam para criar algoritmos quantitativos complexos.

Referências:

https://aws.amazon.com/pt/what-is/data-science/

https://www.oracle.com/br/what-is-data-science/

https://www.tecmundo.com.br/mercado/228839-ciencia-dados.htm

https://www.cienciaedados.com/8-conceitos-estatisticos-fundamentais-para-data-science/

https://www.cienciaedados.com/probabilidade-e-estatistica-os-fundamentos-para-cientistas-de-dados-parte-1/

Extras caso perguntem:

Quais são as diferentes ferramentas de ciência de dados?

A AWS tem uma série de ferramentas para oferecer suporte a cientistas de dados em todo o mundo:

Armazenamento físico de dados

Para data warehousing, o Amazon Redshift pode executar consultas complexas em dados estruturados ou não estruturados. Analistas e cientistas de dados podem usar o AWS Glue para gerenciar e pesquisar dados. O AWS Glue cria automaticamente um catálogo unificado de todos os dados no data lake, com metadados anexados para torná-los detectáveis.

Machine learning

O Amazon SageMaker é um serviço de machine learning totalmente gerenciado executado no Amazon Elastic Compute Cloud (EC2). Ele permite que os usuários organizem dados, criem, treinem e implantem modelos de machine learning e escalem operações.

Análises

O Amazon Athena é um serviço de consultas interativas que facilita a análise de dados no Amazon S3 ou no Glacier. Ele é rápido, com tecnologia sem servidor e funciona usando consultas SQL padrão.

O Amazon Elastic MapReduce (EMR) processa big data usando servidores como Spark e Hadoop.

O Amazon Kinesis permite agregação e processamento de dados de transmissão em tempo real. Ele usa sequências de cliques em sites, logs de aplicações e dados de telemetria de dispositivos de IoT.

O Amazon OpenSearch permite pesquisa, análise e visualização de petabytes de dados.