O que é ciência de dados?

A ciência de dados é o estudo dos dados para extrair insights significativos para os negócios. 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. Essa análise ajuda os cientistas de dados a fazer e responder perguntas como o que aconteceu, por que aconteceu, o que acontecerá e o que pode ser feito com os resultados.

Por que a ciência de dados é importante?

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.

Infelizmente, dados brutos não têm valor a menos que possam ser acionados. Os cientistas de dados podem converter dados brutos em recomendações significativas. Eles podem descobrir e solucionar problemas que as empresas nem sabiam que existiam. As organizações podem usar essas recomendações para deixar os clientes mais satisfeitos, otimizar a cadeia de suprimentos ou lançar novos produtos.

História da ciência de dados

Embora o termo ciência de dados não seja novo, os significados e conotações mudaram ao longo do tempo. A palavra apareceu pela primeira vez na década de 1960 como um nome alternativo para estatísticas. No final da década de 1990, profissionais de ciência da computação formalizaram o termo. Uma definição proposta para ciência de dados o viu como um campo separado com três aspectos: design de dados, coleta e análise. Ainda levou mais uma década para que o termo fosse usado fora da academia.

Futuro da ciência de dados

A inteligência artificial e as inovações de machine learning tornaram o processamento de dados mais rápido e eficiente. A demanda do setor criou um ecossistema de cursos, diplomas e cargos na área da ciência de dados. Devido ao conjunto de habilidades multifuncionais e à experiência necessária, a ciência de dados mostra um forte crescimento projetado nas próximas décadas.

Para que serve a ciência de dados?

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

1. 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. Por exemplo, um serviço de reserva de voos pode registrar dados como o número de bilhetes reservados a cada dia. A análise descritiva revelará picos de reservas, quedas nas reservas e meses de alta performance para este serviço.

2. 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. Por exemplo, o serviço de voo pode fazer drill-down em um mês particularmente de alta performance para entender melhor o pico de reserva. Isso pode levar à descoberta de que muitos clientes visitam uma determinada cidade para assistir a um evento esportivo mensal.

3. 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. Por exemplo, a equipe de serviço de voo pode usar a ciência de dados para prever padrões de reserva de voo para o próximo ano no início de cada ano. O programa de computador ou algoritmo pode analisar dados anteriores e prever picos de reservas para determinados destinos em maio. Tendo previsto as futuras necessidades de viagem de seus clientes, a empresa poderia iniciar a publicidade direcionada para essas cidades a partir de fevereiro.

4. 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.

Voltando ao exemplo de reserva de voo, a análise prescritiva pode analisar campanhas de marketing históricas para maximizar a vantagem do próximo pico de reservas. Um cientista de dados pode projetar resultados de reservas para diferentes níveis de gastos de marketing em vários canais de marketing. Essas previsões de dados dariam à empresa de reservas de voos mais confiança para tomar suas decisões de marketing.

Quais são os benefícios da ciência de dados para os negócios?

A ciência de dados está revolucionando a maneira 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

A ciência de dados 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. Por exemplo, uma empresa de comércio eletrônico usa ciência de dados para descobrir que muitas consultas de clientes estão sendo geradas após o horário comercial. As investigações revelam que os clientes são mais propensos a comprar se receberem uma resposta imediata em vez de uma resposta no próximo dia útil. Ao implementar o atendimento ao cliente 24 horas por dia, sete dias por semana, a empresa aumenta sua receita em 30%.

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. Por exemplo, uma solução de pagamento online usa ciência de dados para coletar e analisar comentários de clientes sobre a empresa nas mídias sociais. A análise revela que os clientes esquecem as senhas durante os períodos de pico de compra e estão insatisfeitos com o sistema atual de recuperação de senhas. A empresa pode inovar uma solução melhor e ver um aumento significativo na satisfação do cliente.

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. Por exemplo, uma empresa de transporte de caminhões usa ciência de dados para reduzir o tempo de inatividade quando os caminhões quebram. Elas identificam as rotas e os padrões de mudança que levam a avarias mais rápidas e ajustam as programações dos caminhões. Elas também configuram um inventário de peças de reposição comuns que precisam ser substituídas com frequência para que os caminhões possam ser reparados mais rapidamente.

O que é o processo de ciência de dados?

Um problema de negócios normalmente inicia o processo de ciência de dados. Um cientista de dados trabalhará com as partes interessadas do negócio para entender quais são as necessidades do negócio. Uma vez definido o problema, o cientista de dados pode solucioná-lo usando o processo de ciência de dados OSEMN:

O: Obter dados

Os dados podem ser pré-existentes, recém-adquiridos ou um repositório de dados que pode ser baixado da Internet. Os cientistas de dados podem extrair dados de bancos de dados internos ou externos, software de CRM da empresa, logs de servidores da Web, mídias sociais ou comprá-los de fontes confiáveis de terceiros.

S: Suprimir dados

A supressão de dados, ou limpeza de dados, é o processo de padronização dos dados de acordo com um formato predeterminado. Ela inclui lidar com a ausência de dados, corrigir erros de dados e remover quaisquer dados atípicos. 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. Técnicas de machine learning, como associação, classificação e clustering, são aplicadas ao conjunto de dados de treinamento. O modelo pode ser testado em relação a dados de teste predeterminados para avaliar a precisão dos resultados. O modelo de dados pode ser ajustado várias vezes para melhorar os resultados.

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.

Quais são as técnicas de ciência de dados?

Os profissionais de ciência de dados usam sistemas de computação para acompanhar o processo de ciência de dados. As principais técnicas usadas pelos cientistas de dados são:

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.

Os profissionais de ciência de dados usam sistemas de computação para acompanhar o processo de ciência de dados.

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.

Quais são as diferentes tecnologias de ciência de dados?

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

1. Inteligência artificial: modelos de machine learning e software relacionado são usados para análises preditivas e prescritivas.
2. Computação em nuvem: as tecnologias de nuvem deram aos cientistas de dados a flexibilidade e a capacidade de processamento necessárias para análises de dados avançadas.
3. 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.
4. Computação quântica: computadores quânticos podem fazer cálculos complexos em alta velocidade. Cientistas de dados qualificados os usam para criar algoritmos quantitativos complexos.