O Que É Ciência De Dados (I)

Yuri Vacelh Zamulhak Zdebski

Descrição da atividade

A atividade propôs assistir a 2 vídeos sobre ciência de dados, o primeiro, em português do canal Nerdologia, apresentado por Atila e o segundo, em ingles, do canal Joma Tech, apresentado por Joma.

O que é ciência de dados | Nerdologia Tech

O primeiro vídeo aborda a ciência de dados de maneira descontraída, utilizando o exemplo do matemático Abraham Wald. Durante a Segunda Guerra Mundial, Wald percebeu que as áreas perfuradas por balas nos aviões sobreviventes não eram as que precisavam ser blindadas, pois, se a aeronave ainda voava, essas partes não necessitavam de proteção adicional. Esse tipo de raciocínio, que vai além do óbvio e lê nas entrelinhas, é fundamental para um cientista de dados, que deve sempre buscar uma interpretação profunda e evitar conclusões enganosas.

Newton, ao afirmar "Se eu vi mais longe, foi por estar sobre os ombros de gigantes", reconheceu a importância dos trabalhos anteriores, como os de Johannes Kepler. Utilizando as leis do movimento planetário de Kepler, que descrevem as órbitas elípticas dos planetas, Newton foi capaz de formular a lei da gravitação universal. Esse processo de construir sobre o conhecimento existente é de certa forma semelhante ao trabalho de um cientista de dados, que utiliza dados e descobertas anteriores para gerar novos insights (talvez seja meio extremo se comparar com grandes mentes como Newton, mas o raciocínio é parecido).

O montante de dados que temos disponíveis hoje em dia é assustadoramente grande (Big-Data), é inimaginável a quantidade de informação que os usuários do Instagram, por exemplo, geram por dia. Esses números enormes possibilitam a criação de modelos que reconhecem padrões nos dados (machine learning), porém nem todos os dados são limpos, no mundo real, coisas como padrão de tempo, utilização do ponto ou virgula para casas decimais, utilização de palavras com o mesmo significado (feminino, F, fem, mulher), existem dentro dos datasets, além claro de dados não estruturados como imagens, áudios, textos contendo opinião. Garantir a qualidade dos dados, que eles englobam tudo o que se quer analisar (como o exemplo dos aviões, e o da identificação de motos), que estão limpos, também estão no escopo do cientista de dados.

O que realmente é Ciência de Dados? Explicado por um Cientista de Dados

O segundo video trata o tema de um ponto de vista de mais empresarial, diferente do primeiro que era uma introdução generalista do assunto, esse é um vlog de uma pessoa que trabalha como cientista de dados trazendo sua visão com base em suas experiencias. A definição dada por Joma é que, um cientista de dados não é alguém que faz modelos complicados, ou gráficos bonitos, mas sim alguém que utilizando dados, resolve problemas reais, independente da ferramenta utilizada.

O vídeo começa relatando brevemente a história do surgimento da ciência de dados, no início dos anos 2000, coincidindo com a ascensão da web 2.0. Durante esse período, a internet

evoluiu de uma plataforma estática para um espaço interativo, permitindo que os usuários postassem, dessem likes e compartilhassem conteúdo. Isso foi marcado pelo surgimento de sites como YouTube e Facebook. Com o aumento exponencial do volume de dados gerados, William S. Cleveland teve a visão de combinar Data Mining com Ciência da Computação, aplicando cálculos estatísticos a esses novos dados. Essa abordagem inovadora elevou o Data Mining a um novo patamar. A evolução foi tão rápida que, poucos anos depois, o volume de dados tornou-se imenso, dando origem ao conceito de Big Data. Os avanços tecnológicos em áreas como infraestrutura de redes, computação paralela e hardware (como processadores, placas de vídeo e armazenamento) foram cruciais para o crescimento dessa área, permitindo que mais dados fossem gerados, processados e transmitidos.

Segundo o Journal of Data Science (JDS), uma definição de ciência de dados é que qualquer atividade relacionada a dados—como coleta, análise e modelagem—se enquadra nesse campo. No entanto, a parte mais importante da ciência de dados são suas aplicações. Uma das aplicações mais famosas é o Deep Learning, cujas realizações só foram possíveis graças ao Big Data. Apesar de sua notoriedade, o foco excessivo no Deep Learning ofusca outras áreas da ciência de dados, tornando a definição do campo mais complexa para o público em geral. Ser um cientista de dados não é sobre o quão avançado é o seu modelo, é sobre resolver problemas.

Por fim, o vídeo termina definindo algumas funções que um cientista de dados pode realizar, dependendo do tamanho da empresa em que está inserido. Em startups, o cientista de dados tende a desempenhar uma variedade de funções, desde a coleta de dados até o deployment de modelos simples de machine learning. Nessas situações, é comum que o time de dados seja pequeno, devido a restrições orçamentárias.

Em empresas de porte médio, a hierarquia é geralmente mais estruturada. A coleta de dados é responsabilidade dos engenheiros de software, enquanto os engenheiros de dados cuidam da arquitetura (armazenamento e estruturação) e da limpeza dos dados (detecção de anomalias e padronização). Os cientistas de dados, por sua vez, focam na análise propriamente dita.

Em grandes empresas, as funções são mais especializadas e compartimentadas. A coleta de dados continua sendo feita pelos engenheiros de software, enquanto a limpeza e o pipeline de dados são gerenciados pelos engenheiros de dados. A análise de dados é então realizada por cientistas de dados, seguida pelo trabalho de analistas de dados e, finalmente, por pesquisadores especializados em IA/ML, que desenvolvem modelos avançados e soluções de inteligência artificial.

Conclusões

Ser um cientista de dados é mais do que apenas gerar métricas estatísticas sobre um dataset (apesar de que elas são sim importantes), mas sim ver além deles. Com o aumento da geração de dados, o campo expandiu para incluir tecnologias avançadas, transformando a maneira como lidamos com informações complexas. As funções dos profissionais de ciência de dados variam conforme o tamanho e as necessidades das empresas, destacando a versatilidade e a importância dessa área no cenário atual. Essa diversidade de papéis reflete a importância e a adaptabilidade da ciência de dados na resolução de problemas complexos.