UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho Final (TF)

Ciência de Dados

Renato Britto Araujo
180027239

Brasília, DF
2023

Introdução

Ciência de dados é um campo interdisciplinar que combina conhecimentos de estatística, matemática, ciência da computação e domínio do assunto para extrair insights e conhecimentos úteis a partir de grandes conjuntos de dados. Seu objetivo principal é descobrir padrões, tendências e informações ocultas nos dados, a fim de tomar decisões baseadas em evidências, resolver problemas complexos e melhorar processos em diversos setores.

A ciência de dados tem se mostrado uma ferramenta poderosa em diversos campos e setores. Além das aplicações mencionadas anteriormente, essa disciplina tem sido utilizada em áreas como finanças, energia, transporte, educação, governo e muito mais. Ao analisar e interpretar grandes volumes de dados, as organizações podem obter insights valiosos que apoiam a tomada de decisões estratégicas, aprimoram processos e impulsionam a inovação. A ciência de dados oferece um conjunto diversificado de técnicas e metodologias que podem ser adaptadas para atender às necessidades específicas de cada setor, trazendo benefícios significativos para empresas, instituições e órgãos governamentais.

Ciência de dados (Data Science) é um campo amplo que abrange todas as etapas do ciclo de vida dos dados, desde a coleta até a interpretação dos resultados. Envolve habilidades em estatística, programação, aprendizado de máquina e conhecimento do domínio para resolver problemas complexos.

Data Analytics, porém, concentra-se principalmente na análise exploratória de dados para descobrir insights e padrões que possam ser usados para tomar decisões de negócios. É uma subárea da ciência de dados, mas tende a se concentrar mais em aspectos descritivos e preditivos do que em aspectos prescritivos.

Já **Business Analytics** tem como objetivo aplicar técnicas analíticas para resolver problemas específicos de negócios, como otimização de processos, análise de mercado e tomada de decisões estratégicas. Embora compartilhe semelhanças com a ciência de dados e a análise de dados, é mais orientado para a aplicação prática em um contexto de negócios.

Ciência de dados faz forte uso de **estatística**, que fornece a base teórica para a modelagem estatística e inferência a partir dos dados; **matemática**, que oferece ferramentas matemáticas para análise de dados, como álgebra linear, cálculo e teoria das probabilidades; **ciência da computação**, que são habilidades de programação e algoritmos são fundamentais para lidar com grandes volumes de dados e implementar soluções tecnológicas e; o domínio do negócio – o **conhecimento específico do domínio** em que os dados estão sendo analisados é essencial para interpretar corretamente os resultados.

Várias técnicas são comunmente utilizadas para se realizar ciência de dados:

A **coleta de dados** é uma etapa crucial, pois envolve não apenas a obtenção, mas também a preparação cuidadosa dos dados necessários para análise. Isso pode incluir a extração de dados de diferentes fontes, como bancos de dados, sistemas de registro ou até mesmo APIs externas. Além disso, é importante realizar a limpeza e o pré-processamento dos dados, o que implica em tratá-los, removendo valores ausentes ou inconsistentes e transformando-os em um formato adequado para análise posterior.

Após a coleta e o pré-processamento dos dados, entra em cena a **análise exploratória de dados**, que tem como objetivo aprofundar o entendimento dos dados por meio de técnicas de visualização e resumos estatísticos. Essa etapa permite **identificar padrões**, tendências e possíveis anomalias nos dados, fornecendo insights valiosos para a próxima fase.

A **modelagem e a análise estatística** são componentes cruciais do próximo passo, envolvendo a aplicação de técnicas estatísticas e algoritmos de aprendizado de máquina para construir modelos preditivos ou descritivos. Esses modelos são capazes de fazer previsões, classificações ou até mesmo segmentações com base nos dados disponíveis. É fundamental **selecionar e ajustar adequadamente** os modelos estatísticos e algoritmos de aprendizado de máquina, levando em consideração as características dos dados e os objetivos específicos do projeto.

Por fim, a **comunicação dos resultados** é uma etapa essencial para a ciência de dados. É preciso apresentar os insights e descobertas obtidos de forma clara, compreensível e relevante para as partes interessadas. Isso pode envolver a criação de relatórios, visualizações interativas, painéis de controle

ou até mesmo apresentações orais. A comunicação efetiva dos resultados é fundamental para garantir que as informações sejam utilizadas para tomada de decisões informadas e obtenção de valor a partir dos dados. **Uma falha em comunicar** dados não apenas seria **descartar todo o esforço** realizado, mas também pode causar confusão nos recipientes e fazer com que uma **decisão incorreta** seja tomada por falta de entendimento, e portanto se faz absolutamente crucial que seja feito de forma correta, sem o risco de falha de comunicação.

Linguagens de programação como o Python e R são amplamente utilizados para análise de dados, devido à sua rica biblioteca de pacotes e ferramentas estatísticas. Ferramentas de visualização de dadoscomo Tableau, Power BI e Matplotlib, são empregadas para criar visualizações claras e informativas. Algoritmos de aprendizado também podem utilizar estatística avançada e moderna para organizar e detectar informação complexa de alto nível em conjuntos de dados.

Objetivos

A **descoberta de padrões** para identificar insights significativos e padrões ocultos nos dados. Isso envolve a aplicação de técnicas estatísticas e algoritmos de aprendizado de máquina para explorar os dados, revelando relações, correlações e tendências relevantes. A descoberta desses insights pode ajudar as empresas a entender melhor seu público-alvo, prever comportamentos futuros e tomar decisões estratégicas embasadas em dados.

A **tomada de decisões baseada em evidências** buscando fornecer informações confiáveis e objetivas para apoiar a tomada de decisões. Ao analisar dados relevantes, as organizações podem ter uma compreensão mais clara dos desafios e oportunidades que enfrentam. Isso permite que tomem decisões informadas e embasadas em evidências, minimizando a incerteza e o viés subjetivo.

E a **otimização de processos e desempenho**, que é aplicada para otimizar processos e melhorar o desempenho em diferentes áreas. Por meio da análise de dados, é possível identificar gargalos, identificar oportunidades de melhoria e implementar estratégias eficientes. Isso pode levar a um aumento da produtividade, redução de custos, maior satisfação do cliente e melhor

alocação de recursos.

Vantagens

A ciência de dados fornece uma habilidade ampliada de análise que alcança além dos limites humanos rudimentares para se **tomar decisões estratégicas**, fornecendo uma corretude e encontro de correlações incomparável com métodos tradicionais. Isso permite que decisões possam ser ideais dentro de contexto que elas são tomadas e os dados que pode usar.

Um líder (seja definido como um ator que toma decisões baseado no mundo que observa) irá analisar os dados que possui seja de forma social, psicológica, geopolítica, econômica entre outros para conseguir tomar uma decisão apropriada. Porém as limitações impostas sobre este líder pela condição humana impossibilita que este **compreenda o sistema complexo** que este está lidando, e portanto um sistema auxíliar externo que faz uma análise extensiva e completa de todas as variáveis que afetam suas decisões se torna imprescidível. Ciência de dados tem o mesmo papel que um **oráculo, braço direito ou conselheiro real** teve no passado para lideranças em seu comando.

Ela impulsiona a inovação e oferece uma **vantagem competitiva** significativa. Ao explorar e analisar dados, as empresas podem descobrir oportunidades de melhoria, identificar lacunas no mercado e desenvolver soluções inovadoras. A capacidade de extrair insights valiosos dos dados e traduzi-los em ações estratégicas permite que as organizações se destaquem da concorrência, ofereçam produtos e serviços diferenciados e se adaptem às mudanças de mercado, em guerras, no governo e outros de forma mais ágil.

Além disso ela permite uma **iteração, melhora e evolução de sistemas atualmente existes**, haja visto que pode analisar de forma exaustiva (porém rudimentar, se comparado a sabedoria humana) todos os fatores que afetam um sistema, permitindo assim se isolar variáveis e entender os motivos pelo quais gargalos e erros ocorrem.

Desvantagens

Embora a ciência de dados traga consigo várias vantagens benéficas, é importante reconhecer e considerar também algumas desvantagens que podem estar associadas a essa abordagem. Quatro aspectos merecem atenção:

Privacidade e ética é um ponto crítico a ser considerado ao utilizar a ciência de dados diz respeito à privacidade e à ética envolvidas no manuseio de grandes volumes de informações. A coleta e o processamento de dados em larga escala podem apresentar riscos à privacidade das pessoas, caso não sejam implementadas medidas adequadas de segurança e anonimização. Além disso, a utilização de algoritmos de aprendizado de máquina pode gerar preocupações éticas, como discriminação algorítmica e vieses indesejados, que podem impactar negativamente determinados grupos ou indivíduos.

A q**ualidade dos dados** também representa um fator de crucial importância que merece atenção na ciência de dados. A existência de dados incompletos, inconsistentes ou imprecisos pode resultar em análises enviesadas ou conclusões equivocadas. É fundamental dedicar recursos e expertise na fase de limpeza e pré-processamento dos dados, a fim de assegurar sua integridade e confiabilidade. Somente assim é possível obter resultados confiáveis e significativos.

Complexidade e interpretação dos modelos na aplicação de técnicas avançadas de aprendizado de máquina e modelagem estatística na ciência de dados pode gerar modelos complexos e de difícil interpretação. Modelos como redes neurais profundas podem ser considerados "caixas-pretas", dificultando a compreensão dos fatores que influenciam suas previsões ou decisões. Essa falta de transparência pode limitar a capacidade de explicar e justificar os resultados obtidos, especialmente em contextos regulatórios ou em situações em que a tomada de decisões críticas exige uma compreensão clara dos processos envolvidos.

Existe também o risco de ***overfitting*** dos modelos, que podem ser causados por falhas técnicas mas também uma super confiança de que os dados atuais representam um conjunto abrangente o suficiente para abarcar o assunto que está sendo analisado.

No mais, também ocorre a **dependência excessiva de dados,** pois todo o processo que foi mencionado requer uma quantidade significativa de dados para construir modelos robustos e confiáveis. No entanto, em determinados domínios ou setores que possuem acesso limitado a dados de alta qualidade, essa dependência excessiva pode se tornar uma desvantagem. Além disso, a obtenção e a preparação dos dados podem ser processos complexos e custosos, especialmente quando envolvem a integração de fontes de dados heterogêneas. A escassez de dados relevantes e confiáveis pode impactar negativamente a eficácia da ciência de dados em determinados contextos, exigindo abordagens alternativas e cuidadosas.

Dados também podem ser **muito caros** para se acumular, e portanto depender de ciência de dados para certas decisões pode se tornar um gasto excessivo e não deve ser utilizado sempre, sem se considerar o valor que pode ser extraído desse processo.

Exemplos

A Netflix é um exemplo de sucesso na aplicação da ciência de dados. A empresa revolucionou a indústria do entretenimento ao utilizar análise de dados para personalizar as recomendações de conteúdo para seus usuários. Através de algoritmos avançados de aprendizado de máquina, a Netflix analisa o histórico de visualização, preferências e classificações dos usuários para sugerir filmes e séries que sejam de seu interesse. Isso é possível graças à coleta e análise de dados em tempo real, o que permite à Netflix compreender o comportamento dos usuários e criar perfis personalizados para cada um deles. Essas recomendações altamente relevantes aumentam a satisfação do usuário e incentivam sua permanência na plataforma.

Outro exemplo de sucesso na aplicação da ciência de dados é a Amazon, uma empresa de comércio eletrônico. A Amazon utiliza ciência de dados em várias áreas de seu negócio, incluindo recomendações de produtos, logística e otimização de operações. A empresa coleta uma grande quantidade de dados sobre seus clientes e produtos e utiliza algoritmos de aprendizado de máquina para compreender padrões de compra, comportamento do usuário e preferências individuais. Essa abordagem permite oferecer recomendações personalizadas de produtos, melhorar a eficiência de suas operações logísticas e prever a demanda de produtos, garantindo a disponibilidade de estoque.

Referências

BUSINESS Intelligence, Analytics, and Data Science: A Managerial Perspective. [*S. l.*: *s. n.*], 2016.

Aalst, W. van der, Process Mining: Discovery, Conformance and Enhancement of Business Processes, Springer-Verlag Berlin Heidelberg, 2011.

Agrawal, R., Srikant, R., Fast algorithms for mining association rules, Proceedings of the 20th International Conference on Very Large Data Bases, 487-499, 1994.

Alexandre, J., Cavique, L., NoSQL no suporte à análise de grande volume de dados, Revista de Ciências da Computação, 8, 37-48, 2013.

Breiman, L., Statistical modeling: the two cultures, Statistical Science, 16, 199-231, 2001

Base de Dados

A base de dados pode ser encontrada em: https://relational.fit.cvut.cz/dataset/GOSales

Em nenhum momento durante a atividade foi mencionado que o o projeto deveria ser realizado em português, portanto os **nomes originais** do dataset foram mantidos **em inglês**.

Nela é descrita as vendas, produtos e fornecedores (bem como outros dados) das vendas de uma empresa fictícia.

O objetivo primário para qual este dataset foi criado é para se tentar prever a quantidade de vendas que foram realizadas, ou serão realizadas.

Este objetivo se relaciona com a ciência de dados pois permite uma tomada de decisões, como por exemplo a do fornecedor cujo qual os empresários desta empresa iriam focar, visto que o custo de fornecimento e venda do produto traria maior lucro.

Diagrama Entidade-Relacionamento (DE-R)

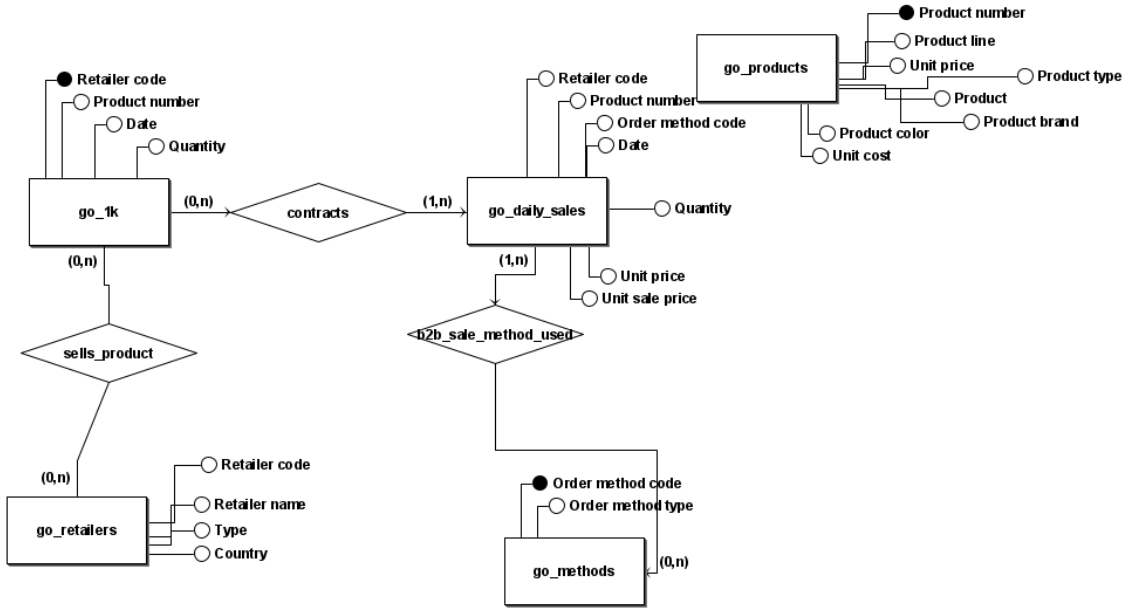

Diagrama Lógico de Dados (DLD)