UNIVERSIDADE DE BRASÍLIA
Faculdade do Gama

Sistemas de Banco de Dados 2

Trabalho final (TF)

Mineração de dados

Gabriel Costa de Oliveira - 190045817

Brasília, DF

2023

Definição

A sociedade contemporânea é caracterizada pela imensa quantidade de dados gerados, representando um volume sem precedentes de armazenamento. Com a emergência da internet, a quantidade de dados armazenados aumentou rapidamente, chegando à ordem de grandeza dos zettabytes em 2015.

Diante desse cenário, vários setores da sociedade reconheceram a necessidade de extrair novas informações desses dados armazenados. Por exemplo, empresas varejistas podem se beneficiar ao identificar tendências de mercado e perfis de clientes com base nos produtos vendidos. Da mesma forma, estudos científicos podem analisar os padrões de saúde dos pacientes para identificar grupos de risco de determinadas doenças.

Segundo Dias (2004), embora as empresas possuam uma vasta quantidade de informações sobre seu negócio, elas enfrentam dificuldades na hora de extrair novos conhecimentos a partir dessas informações existentes.

De acordo com (Goebel e Gruenwald, 1999), a mineração de dados é o processo de descoberta de padrões, conhecimento e informações úteis em grandes volumes de dados. Envolve a aplicação de técnicas estatísticas e algoritmos de aprendizado de máquina para identificar relações e tendências ocultas nos dados, permitindo a extração de insights valiosos e a tomada de decisões informadas. A mineração de dados busca explorar os dados de forma automatizada, identificando padrões e relações que podem não ser facilmente percebidos pelo olhar humano. O objetivo é transformar os dados brutos em conhecimento acionável, auxiliando organizações e pesquisadores a obterem vantagens competitivas e fazerem descobertas significativas.

Objetivos

A mineração de dados surgiu como uma abordagem eficaz para explorar a imensa quantidade de informações disponíveis, utilizando técnicas multidisciplinares que envolvem estatística e ciência de computação. Seu objetivo é descobrir padrões, tendências e correlações relevantes nos conjuntos de dados, transformando os dados brutos em conhecimento útil.

Segundo (Dias 2002) as técnicas de mineração de dados podem ser aplicados a tarefas como classificação, estimativa, associação, segmentação e sumarização. e acordo com Dias (2002), a mineração de dados abrange uma variedade de técnicas que podem ser aplicadas a diversas tarefas, tais como **classificação**, **estimativa**, **associação**, **segmentação** e **sumarização**.

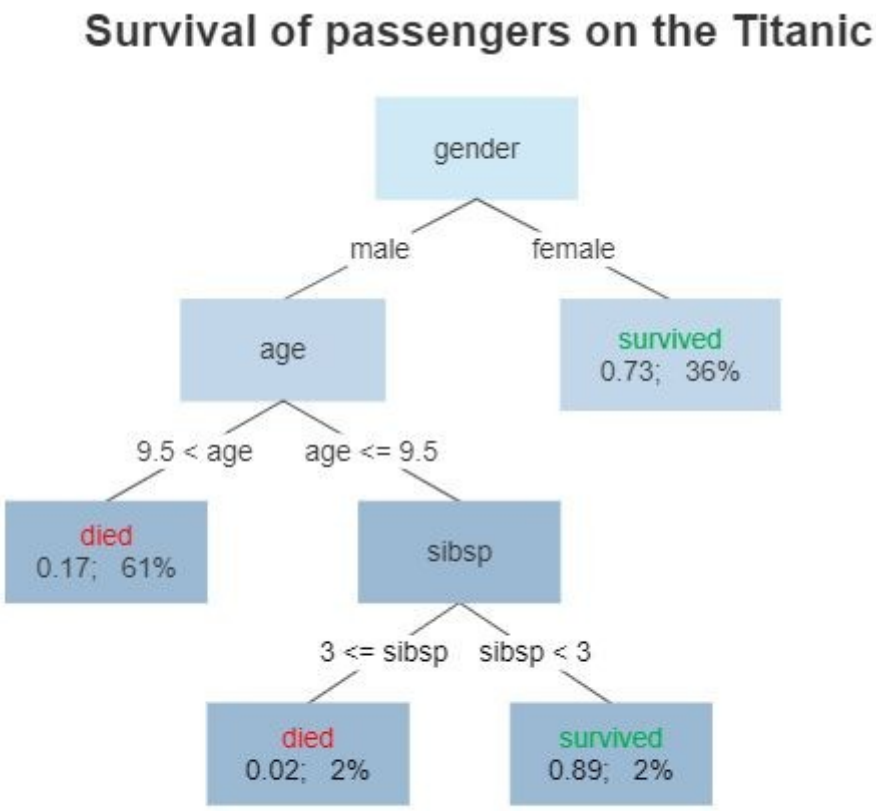
A **classificação** consiste na atribuição de exemplos a classes predefinidas, permitindo a criação de modelos capazes de classificar novos dados com base em características conhecidas.

Na Figura 1 é apresentada uma representação gráfica em forma de árvore que ilustra a relação entre a sobrevivência dos passageiros a bordo do Titanic e suas características, sendo uma forma de classificação dos mesmos. Nessa representação, é considerada a variável "sibsp" que representa o número de cônjuges ou irmãos presentes no navio. Os números exibidos abaixo das folhas indicam a probabilidade de sobrevivência e a porcentagem de observações que se enquadram em cada folha. De forma resumida, observa-se que as chances de sobrevivência eram favoráveis para indivíduos que se enquadrassem em dois grupos distintos: (i) mulheres e (ii) homens com idade inferior a 9,5 anos e menos de 3 irmãos. Essa árvore de segmentação, ou árvore de decisão, permite uma análise visual e compreensão das relações entre as variáveis e suas influências na classificação da sobrevivência dos passageiros no evento do naufrágio do Titanic.

Já a **estimativa**, também conhecida como regressão, busca prever valores numéricos contínuos, permitindo estimativas precisas com base em atributos relevantes.

A tarefa de **associação** envolve a descoberta de relações e padrões interessantes entre itens em um conjunto de dados, possibilitando a identificação de associações frequentes e regras de implicação.

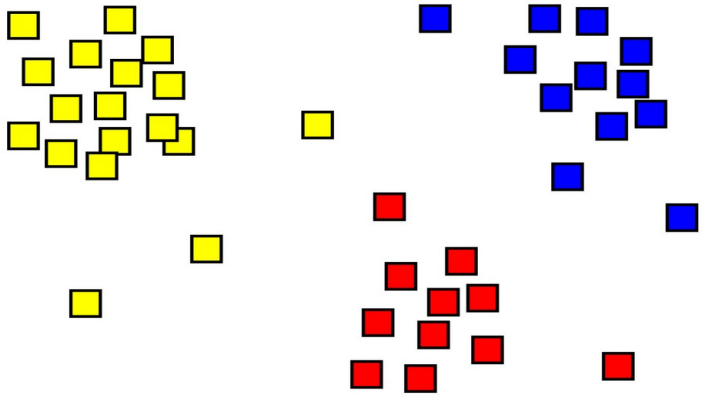
Figura 1 - classificação com arvore de decisão



Fonte - Wikipédia

Por sua vez, a **segmentação**, também chamada de clusterização, agrupa dados com base em similaridades, formando clusters que compartilham características comuns e permitindo a identificação de padrões intrínsecos nos dados.

Figura 1 - O resultado de uma análise de cluster mostrado como a coloração dos quadrados em três clusters



Fonte - Wikipédia

Por fim, a **sumarização** visa extrair informações essenciais e relevantes dos dados, gerando resumos concisos que destacam os principais aspectos e tendências.

Vantagens

A utilização de técnicas de mineração de dados oferece diversas vantagens e benefícios significativos em diferentes campos científicos e aplicados. A mineração de dados é uma abordagem exploratória e analítica que envolve a extração de informações valiosas e conhecimento útil a partir de grandes volumes de dados não estruturados ou estruturados. Essas técnicas são fundamentadas em métodos estatísticos, aprendizado de máquina e inteligência artificial, permitindo a descoberta de padrões,

relacionamentos e tendências ocultas nos dados.Uma das principais vantagens da mineração de dados é a capacidade de lidar com conjuntos de dados complexos e heterogêneos, auxiliando na tomada de decisões informadas. Essas técnicas são capazes de identificar correlações sutis e insights não triviais nos dados, fornecendo uma visão aprofundada e abrangente dos fenômenos estudados. Além disso, a mineração de dados oferece a capacidade de explorar grandes volumes de dados de forma eficiente, identificando padrões previamente desconhecidos e revelando conhecimento oculto

Desvantagem

A utilização de técnicas de mineração de dados apresenta desvantagens e desafios que exigem uma abordagem cuidadosa. Essas limitações podem afetar a eficácia, confiabilidade e interpretação dos resultados obtidos. Alguns dos principais problemas e desvantagens associados à mineração de dados são o **viés** e a **interpretação dos dados**. O viés e a qualidade dos dados desempenham um papel crítico na mineração de dados. Dados incompletos, inconsistentes ou com erros podem levar a resultados distorcidos e conclusões equivocadas.

Além disso, os dados utilizados na mineração podem apresentar viés, refletindo desigualdades existentes na sociedade. Esses viéses podem surgir durante a coleta, amostragem ou anotação dos dados, resultando em resultados enviesados e injustos.

A interpretação e validade dos resultados da mineração de dados são desafiadoras. A descoberta de correlações ou padrões em dados não implica necessariamente em causalidade. A inferência de relações de causa e efeito requer uma análise cuidadosa e conhecimento especializado do domínio. Além disso, a validade dos resultados pode ser questionada se o conjunto de

dados utilizado na mineração não for representativo ou se os algoritmos e parâmetros escolhidos forem inadequados.

Exemplo de uso interessante

Um exemplo interessante de uso da tecnologia de banco de dados é a empresa Amazon. A Amazon é conhecida por sua plataforma de comércio eletrônico, mas também é uma das pioneiras no uso intensivo de dados para impulsionar suas operações e tomar decisões estratégicas.

A Amazon utiliza técnicas avançadas de mineração de dados em seus bancos de dados para analisar o comportamento de compra dos clientes, identificar padrões e preferências, e personalizar recomendações de produtos. Isso permite que a empresa ofereça uma experiência de compra altamente personalizada e aumente a satisfação do cliente.

Além disso, a Amazon utiliza técnicas de análise de dados para otimizar suas operações logísticas e de armazenamento. Por meio do processamento de grandes volumes de dados, a empresa consegue prever demandas, gerenciar estoques de forma eficiente e reduzir custos operacionais.

Uma desvantagem significativa da mineração de dados é a possibilidade de que os dados utilizados possam refletir preconceitos e desigualdades existentes na sociedade, resultando em programas e algoritmos que reproduzem e perpetuam viéses e discriminação.

Referência

ESCOVEDO, Tatiana; KOSHIYAMA, Adriano. Introdução a Data Science: Algoritmos de Machine Learning e métodos de análise. 1. ed. Casa do Código, 2020.

AWARI. Marketing de Banco de Dados Eficaz: Melhores Práticas. Disponível em: https://awari.com.br/marketing-de-banco-de-dados/?utm_source=blog&utm_campaign=projeto+blog&utm_medium=Marketing%20de%20Banco%20de%20Dados%20Eficaz:%20Melhores%20Pr%C3%A1ticas. Acesso em:Acesso em: 11 de junho de 2023.

BARRA, Guilherme. Análise preditiva: o que é, como funciona e exemplos práticos. Disponível em: https://rockcontent.com/br/blog/analise-preditiva/. Acesso em: Acesso em: 11 de junho de 2023.

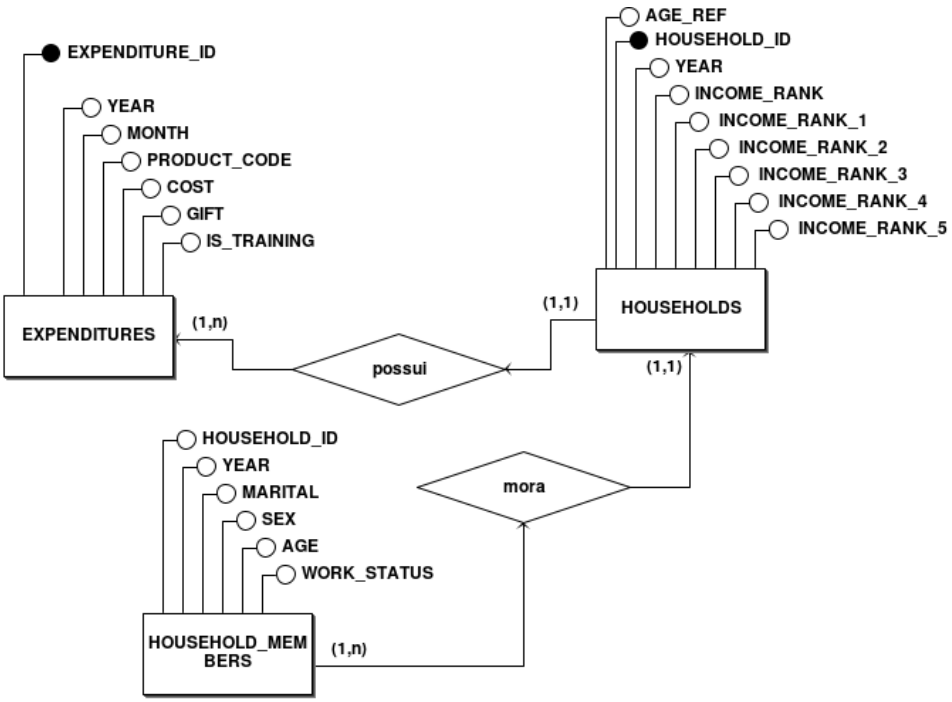
DIAS, Maria Madalena. Parâmetros na escolha de técnicas e ferramentas de mineração de dados. Acta Scientiarum. Technology, v. 24, p. 1715-1725, 2002.

GOEBEL, M.; GRUENWALD, L. A survey of data mining and knowledge discovery software tools. ACM SIGKDD, San Diego, v. 1, n. 1, p. 20-33, 1999.

Base de dados Pesquisadas

Base disponível em https://relational.fit.cvut.cz/dataset/ConsumerExpenditures

DER



DLD