

14º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2023

INTELIGÊNCIA ARTIFICIAL GENERATIVA NO CONTEXTO DO MÉTODO KDD: EXPLORANDO O PRÉ-PROCESSAMENTO DE DADOS

EDUARDO. H. GOMES¹, DOUGLAS R. R. SANTOS², BENILDES F. MENEZES³, BRUNA H. S. SANTOS⁴, MARCELO M. L. JUNIOR⁵, ALBERTO L. FERREIRA⁶, MARCO A. P. MARQUES⁷

¹ Mestre em Engenharia da Informação, Professor do IFSP - Câmpus Cubatão, ehgomes@ifsp.edu.br.

² Graduando em Análise e Desenvolvimento de Sistemas, IFSP, Câmpus Cubatão, douglas.reis@aluno.ifsp.edu.br.

³ Graduando em Análise e Desenvolvimento de Sistemas, IFSP, Câmpus Cubatão, menezes.f@aluno.ifsp.edu.br.

⁴ Graduando em Análise e Desenvolvimento de Sistemas, IFSP, Câmpus Cubatão, bruna.helena@aluno.ifsp.edu.br.

⁵ Graduando em Análise e Desenvolvimento de Sistemas, IFSP, Câmpus Cubatão, marcelo.modesto@aluno.ifsp.edu.br.

⁶ Mestre em Ensino de Ciências, Professor do IFSP - Câmpus Cubatão, alberto@ifsp.edu.br.

⁷ Mestre em Educação, Professor do IFSP - Câmpus Cubatão, mmarques@ifsp.edu.br.

Área de conhecimento (Tabela CNPq): 1.03.03.04-9 Sistemas de Informação

RESUMO: Este trabalho se propõe a explorar a integração da Inteligência Artificial Generativa (IAG) no processo de pré-processamento de dados dentro da metodologia de Descoberta de Conhecimento em Bases de Dados (KDD). Utilizando a capacidade de modelos de linguagem, como o GPT da OpenAI, buscamos entender como essas ferramentas podem auxiliar nas etapas de limpeza e preparação de dados, fases vitais para a eficácia do KDD. Utilizando um estudo de caso sobre a seleção otimizada de rações para pets, demonstramos as potencialidades da utilização da IAG nesse contexto. As descobertas indicam que, enquanto o GPT pode ser uma ferramenta valiosa, a sua eficácia depende de uma compreensão adequada das suas capacidades e limitações. Este estudo, portanto, visa não somente contribuir para o discurso acadêmico, mas espera uma melhoria na qualidade e precisão dos dados após o pré-processamento. Isso, por sua vez, potencializa as fases subsequentes de mineração e pós-processamento, resultando em *insights* mais precisos e confiáveis. Além disso, pode acelerar significativamente o tempo de pré-processamento, permitindo uma análise mais rápida e eficiente de grandes volumes de dados. Em resumo, a adoção de técnicas de IA Generativa tem o potencial de elevar a mineração de dados a um novo patamar de eficácia e eficiência.

PALAVRAS-CHAVE: gpt; inteligência artificial; mineração de dados; openai.

GENERATIVE ARTIFICIAL INTELLIGENCE IN THE CONTEXT OF THE KDD METHOD: EXPLORING DATA PRE-PROCESSING

ABSTRACT: This study aims to explore the integration of Generative Artificial Intelligence (GAI) into the data pre-processing stage within the Knowledge Discovery in Databases (KDD) methodology. Leveraging the capabilities of language models, such as OpenAI's GPT, we seek to understand how these tools can assist in the data cleaning and preparation phases, which are crucial for the effectiveness of KDD. Using a case study on optimized pet food selection, we demonstrate the potentials of using GAI in this context. Findings suggest that while GPT can be a valuable tool, its effectiveness is contingent upon a proper understanding of its capabilities and limitations. This study, therefore, aims not only to contribute to academic discourse but also hopes for an improvement in data quality and accuracy post-pre-processing. This, in turn, enhances the subsequent phases of mining and post-processing, resulting in more accurate and reliable *insights*. Moreover, it can significantly speed up the pre-processing time, allowing for quicker and more efficient analysis of large data volumes. In summary, the adoption of Generative AI techniques has the potential to elevate data mining to a new level of effectiveness and efficiency.

KEYWORDS: gpt; artificial intelligence; data mining; openai.

INTRODUÇÃO

A mineração de dados tem sido uma área em rápida evolução nas últimas décadas. O seu propósito principal é extrair informações úteis de grandes conjuntos de dados, convertendo-as em estruturas compreensíveis para tomada de decisão futura. O processo KDD (*Knowledge Discovery in Databases*) representa uma sequência sistemática de etapas, que inclui pré-processamento, mineração e pós-processamento de dados. Com o avanço da Inteligência Artificial (IA), novas técnicas, particularmente aquelas baseadas em IA generativa, começaram a surgir e têm potencial para revolucionar a forma como encaramos cada etapa do KDD.

Uma das fases críticas no processo KDD é a de pré-processamento de dados, onde a seleção de atributos e a limpeza de dados desempenha um papel fundamental. Dados brutos, especialmente em grandes bases de dados, frequentemente contêm ruídos ou caracteres indesejados que podem distorcer análises posteriores, além de comprometer a integridade e confiabilidade de todo o processo KDD, levando a decisões mal informadas ou conclusões errôneas.

Ao implementar a IA Generativa no processo KDD, podemos esperar uma melhoria significativa na qualidade e precisão dos dados após o pré-processamento. Isso, por sua vez, potencializa as fases subsequentes de mineração e pós-processamento, resultando em *insights* mais precisos e confiáveis. Além disso, a automação desse processo através da IA pode acelerar significativamente o tempo de pré-processamento, permitindo uma análise mais rápida e eficiente de grandes volumes de dados. Em resumo, a adoção de técnicas de IA Generativa tem o potencial de elevar a mineração de dados a um novo patamar de eficácia e eficiência.

PROCESSO KDD E MINERAÇÃO DE DADOS

A técnica de descoberta de conhecimento, também conhecida por sua sigla em inglês KDD (*Knowledge Discovery in Databases*), formalizado por Fayyad (1989), refere-se ao conjunto de técnicas e processos aplicados para descobrir padrões e conhecimentos úteis em grandes volumes de dados. No contexto atual da era da informação, a capacidade de extrair *insights* valiosos dos dados é de vital importância para diversas áreas, incluindo negócios, medicina, educação, entre outras.

O objetivo central da descoberta de conhecimento é identificar e extrair padrões e informações significativas "ocultas" dentro dos dados. Estas informações, por sua vez, podem ser utilizadas para melhorar a tomada de decisão em diversas situações. Esse processo é interdisciplinar e combina conhecimentos de diversas áreas científicas, incluindo Estatística, Banco de Dados, Inteligência Artificial, Aprendizado de Máquina e Reconhecimento de Padrões. (WITTEN, 2011).

Sem o uso de ferramentas computacionais específicas, é praticamente impossível para seres humanos analisar essas grandes quantidades de dados. Portanto, é imperativo desenvolver e implementar ferramentas que possam automatizar análises, interpretações e associações de dados, permitindo assim a geração de *insights* valiosos que auxiliem nas decisões e na criação de estratégias relevantes. (GOLDSCHIMIDT, 2005).

Alguns autores, como Han (2011), usam os termos KDD e Mineração de Dados de forma equivalente, porém, a maioria entende que a mineração é apenas uma componente do processo KDD. Fayyad (1996) destaca que enquanto KDD abrange o processo completo de extrair conhecimento de bases de dados, a mineração de dados representa a fase principal desse processo, que é o reconhecimento de padrões conforme mostrado na Figura 1.

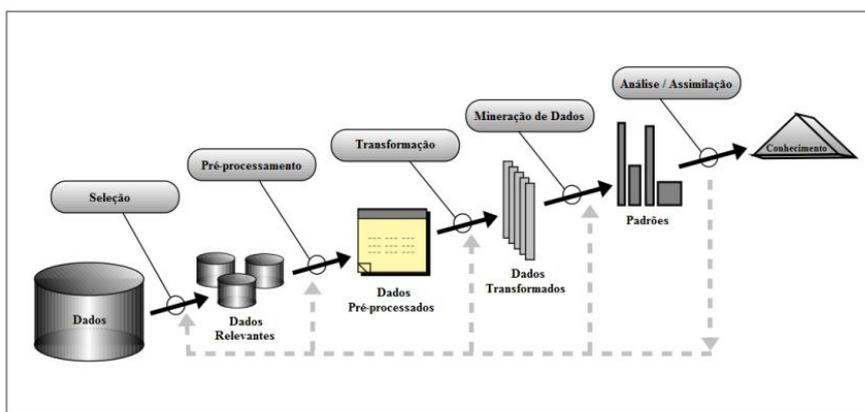


FIGURA 1. Etapas do Processo de KDD (FAYYAD, 1996).

O processo de KDD, como delineado por Fayyad (1996), é uma série de etapas iterativas e interativas, projetadas para identificar padrões novos, válidos e compreensíveis nos dados que são as seguintes: (1) seleção dos dados; (2) pré-processamento dos dados; (3) transformação dos dados; (4) mineração de dados e (5) interpretação e avaliação dos resultados.

O pré-processamento de dados é uma das etapas mais demoradas e laboriosas no processo de KDD. Estima-se que cerca de 80% do tempo gasto no KDD seja dedicado a essa atividade. (PYLE, 1999). Isso acontece porque, muitas vezes, há questões específicas para cada aplicação que exigem soluções personalizadas. Além disso, há uma escassez de ferramentas abrangentes para soluções de pré-processamento que o analista de dados possa explorar e testar. (BATISTA, 2003). Entre as ferramentas disponíveis para auxiliar no processo KDD, encontra-se a Weka, que é amplamente adotada no meio acadêmico.

INTELIGÊNCIA ARTIFICIAL GENERATIVA

A inteligência artificial (IA) tem sido objeto de pesquisa e desenvolvimento contínuo desde sua concepção. Uma das subáreas emergentes e de rápida evolução dentro da IA é a Inteligência Artificial Generativa (IAG) que refere-se a modelos e algoritmos de IA projetados para gerar dados que se assemelham a dados reais. Em contraste com os sistemas de IA "discriminativos", que tentam classificar ou diferenciar dados existentes, os sistemas generativos visam criar novos dados que se ajustem a padrões ou distribuições conhecidas. O objetivo principal é, muitas vezes, produzir saídas que sejam indistinguíveis dos dados reais para um observador humano ou para outros modelos de IA.

Ela se baseia em algoritmos de heurística computacional, utilizando técnicas avançadas de *machine learning* como redes neurais, *deep learning* e sustentada por redes neurais generativas. Uma das abordagens mais destacadas na IAG é a utilização de Redes Adversariais Generativas (GANs). Propostas por Goodfellow (2014), esses sistemas têm a capacidade de criar hipóteses a partir de dados padronizados e aprender de forma autônoma por meio do reconhecimento de padrões em várias camadas de processamento. (MOURA, 2023).

Os modelos de linguagem, especialmente os de grande escala como o GPT (*Generative Pre-trained Transformer*) da OpenAI (2021), conforme evidenciado por Brown (2020), têm demonstrado uma capacidade notável de aprender com poucos exemplos, sendo classificados como "*few-shot learners*". No contexto da Inteligência Artificial generativa, essa habilidade traduz-se na capacidade do modelo de gerar conteúdos linguísticos coerentes e contextualmente relevantes com base em um pequeno conjunto de exemplos fornecidos. No caso do GPT, pode adaptar-se a novas tarefas com apenas uma breve exposição a elas. Esta característica coloca o GPT em uma posição de destaque no campo da IA generativa, abrindo portas para uma ampla gama de aplicações em processamento de linguagem natural e áreas correlatas. (BROWN, 2020).

Conforme relatado por De Oliveira (2021), o GPT tem sido usado em várias áreas no desenvolvimento de sistemas, uma delas é como desenvolvimento conjunto do programador (*pair programming*), o autor também alerta para a falta de pesquisas nessa área onde percebe-se a necessidade de mais estudos na temática. Neste contexto a utilização do GPT no pré-processamento de dados dentro do método KDD torna-se mais uma área a ser explorada nos próximos anos devido a estes modelos,

classificados como "*few-shot learners*", demonstrarem uma capacidade notável de interpretar, limpar e estruturar grandes volumes de dados linguísticos com base em poucos exemplos fornecidos.

A eficiência no tratamento de informações é especialmente relevante para a fase de pré-processamento do KDD, onde a qualidade e a estruturação dos dados são cruciais para as etapas subsequentes de análise. A habilidade do GPT de se adaptar a novas tarefas com mínimo ajuste destaca seu potencial como uma ferramenta robusta para automatizar e otimizar o pré-processamento de dados, facilitando assim a descoberta de padrões e *insights* valiosos em conjuntos de dados complexos.

MATERIAL E MÉTODOS

O conjunto de metodologias de investigação utilizadas neste projeto se fundamentam em duas práticas de pesquisa: a descritiva e a experimental. A descritiva foi utilizada no levantamento do referencial teórico das técnicas de descoberta de conhecimento (KDD). A metodologia experimental foi utilizada pela própria natureza do projeto. O presente estudo de caso foi conduzido com um grupo de alunos do curso superior de Análise e Desenvolvimento de Sistemas. O objetivo central do projeto era desenvolver uma ferramenta digital que auxiliasse proprietários de animais domésticos na seleção otimizada de rações, ponderando aspectos relativos à saúde e economia. Nesse contexto, a elaboração de um modelo analítico baseado nos níveis de garantia presentes em rações para pets foi identificada como primordial.

Para orientar o desenvolvimento do projeto, o grupo formulou questões específicas para serem direcionadas ao modelo GPT (*Generative Pre-trained Transformer*):

- 1. Me fale sobre níveis de garantia em rações para pets.*
- 2. Nos rótulos aparecem muitas informações com diversos níveis de porcentagem. Eu poderia utilizar alguma técnica de Inteligência artificial para saber se uma ração é mais saudável que outra?*
- 3. Preciso que me oriente com detalhes, forneça um passo a passo técnico para fazer essa análise de dados.*
- 4. Quero que você assuma a função de cientista de dados desse projeto e forneça detalhes de como implementar esse projeto.*
- 5. Quais dados devemos coletar? Poderia fornecer algum exemplo prático. (pré-processamento)*
- 6. Qual técnica poderíamos utilizar para identificar padrões? Forneça detalhes técnicos sobre ferramentas, softwares e os algoritmos que deverão ser utilizados nesse projeto.*

RESULTADOS E DISCUSSÃO

Em resposta aos questionamentos formulados, o GPT forneceu diretrizes aprofundadas cobrindo todas as fases do método KDD, desde a aquisição de dados via *webscraping* até a mineração de dados utilizando o algoritmo K-means. A partir dessa orientação, foram identificados clusters de rações categorizadas como estereótipos A, B e C. No escopo deste trabalho, focaremos exclusivamente na etapa de pré-processamento dos dados.

De acordo com as orientações fornecidas pelo modelo GPT em resposta à quinta questão, os principais atributos a serem considerados são provenientes das informações nutricionais estampadas nos rótulos das rações. O GPT forneceu os seguintes detalhes:

TABELA 1. Respostas do GPT.

Resposta

Para coletar os dados necessários para a análise de qualidade nutricional das rações para pets, é preciso ter em mente que as informações nutricionais fornecidas nos rótulos das embalagens são de grande importância.

Para ilustrar esses dados, podemos dar um exemplo prático. Suponha que você queira avaliar a qualidade nutricional de rações para cães adultos de pequeno porte. Para isso, você pode coletar informações nutricionais de diferentes marcas de ração e registrar as seguintes informações:

- *Composição: frango, milho, arroz, trigo, gordura animal, palatabilizantes, antioxidantes, vitaminas e minerais.*
- *Nutrientes: 25% de proteína bruta, 15% de gordura bruta, 3% de fibras brutas, 10% de umidade, 1% de cinzas e 0,3% de cálcio.*
- *Calorias: 3500 kcal/kg.*
- *Aditivos: corantes artificiais, conservantes.*
- *Fabricante: marca X.*

Com base nessas informações, é possível realizar a análise de qualidade nutricional da ração, comparando com outras marcas de ração para cães adultos de pequeno porte.

Após a definição dos atributos-chave, buscou-se o auxílio do modelo GPT para orientar o processo de coleta de dados a partir de websites de fabricantes e revendedores de rações, como mostrado na figura 2.

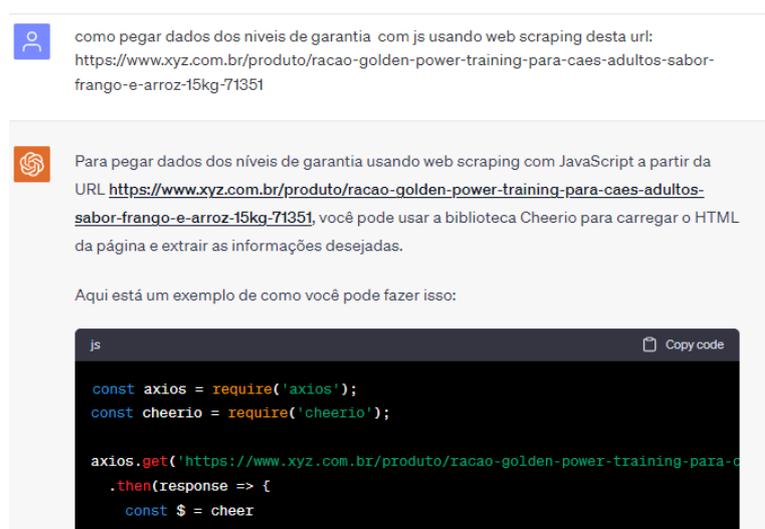


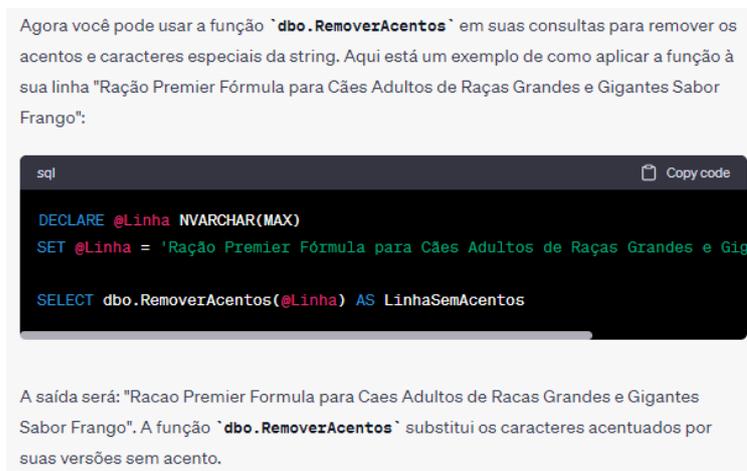
FIGURA 2. Trecho de código para Aquisição de dados (seleção de atributos) fornecido pelo GPT.

Uma vez coletados, os dados brutos, originários de tabelas de níveis de garantia de diferentes rações, lotes e categorias, necessitaram passar por um processo de limpeza dos dados. O objetivo era assegurar que os dados estivessem adequados para análise subsequente. Para isso, recorreu-se novamente ao modelo GPT, conforme ilustrado na figura 3.



FIGURA 3. Trecho de código para limpeza dos dados fornecido pelo GPT.

Na primeira abordagem, a utilização da função *REPLACE* não resultou como esperado, exigindo uma reavaliação. Uma solução alternativa foi então proposta pelo modelo GPT, focada na remoção de caracteres acentuados. Esta nova abordagem foi aplicada com sucesso, resultando em um conjunto de dados limpo e pronto para as etapas posteriores de análise, conforme mostrado na Figura 4.



Agora você pode usar a função `dbo.RemoveAcentos` em suas consultas para remover os acentos e caracteres especiais da string. Aqui está um exemplo de como aplicar a função à sua linha "Ração Premier Fórmula para Cães Adultos de Raças Grandes e Gigantes Sabor Frango":

```
sql Copy code
DECLARE @Linha NVARCHAR(MAX)
SET @Linha = 'Ração Premier Fórmula para Cães Adultos de Raças Grandes e Gigantes Sabor Frango'
SELECT dbo.RemoveAcentos(@Linha) AS LinhaSemAcentos
```

A saída será: "Racao Premier Formula para Caes Adultos de Racas Grandes e Gigantes Sabor Frango". A função `dbo.RemoveAcentos` substitui os caracteres acentuados por suas versões sem acento.

FIGURA 4. Trecho de código com função para remover acentos fornecido pelo GPT.

CONCLUSÕES

A partir dos resultados obtidos, é possível concluir que a integração da IAG, usando o modelo GPT, no processo KDD, particularmente na fase de pré-processamento, mostrou ser uma abordagem viável e eficiente. O modelo não apenas forneceu orientações detalhadas e práticas sobre como coletar, selecionar e limpar os dados, mas também demonstrou adaptabilidade quando confrontado com desafios, como a necessidade de ajustar abordagens durante a limpeza dos dados.

Embora este estudo tenha revelado resultados promissores, também destaca a necessidade de mais pesquisas na integração da IAG e KDD. Questões como a ética da automação, a qualidade dos *insights* gerados e a eficácia a longo prazo do uso de modelos como o GPT no KDD precisam ser abordadas em pesquisas futuras.

REFERÊNCIAS

BATISTA, G. E. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2003.

BROWN, Tom et al. **Language models are few-shot learners**. Advances in neural information processing systems, v. 33, p. 1877-1901, 2020.

DE OLIVEIRA S., J. V.; PACHECO, G. O.; PUGLIESI, J. B. **O Modelo De Inteligência Artificial Gpt-3 Na Programação E Suas Vantagens E Desvantagens No Desenvolvimento Junto Ao Programador**. Revista Eletrônica de Computação Aplicada, v. 2, n. 1, 2021.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. **From Data Mining to Knowledge Discovery in Databases**. AI Magazine, [S.l.], v. 17, n. 3, p.37-54. 1996.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining: Um Guia Prático**. Rio de Janeiro: Elsevier, 2005.

GOODFELLOW, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. **Generative adversarial nets**. In **Advances in neural information processing systems** (pp. 2672-2680). 2014.

HAN, J.; KAMBER, M.; PEI, J. **Data mining: Concepts and Techniques**. 3.ed. Waltham: Morgan Kaufmann Publishers Inc., 2011.

MOURA, M. V. **A Inteligência Artificial Generativa como autora de invenções patenteáveis: um estudo analítico do "Caso DABUS"**. 2023. Disponível em: <https://repositorio.animaeducacao.com.br/handle/ANIMA/36309>. Acesso em 15 ago. 2021.

OPENAI; PILIPISZYN, A. **GPT-3 Powers the Next Generation of Apps**, 2021. Disponível em: <https://openai.com/blog/gpt-3-apps/>. Acesso em 15 ago. 2023.

PYLE, D. **Data Preparation for Data Mining**. San Francisco, CA: Morgan Kaufmann. 38, 42, 48. 1999.

WITTEN, I.; FRANK, E.; HALL, M. **Data Mining: Practical Machine Learning Tools and Techniques**. Morgan Kaufmann Publishers. 2011.