DESENVOLVIMENTO DE UMA FERRAMENTA INTERATIVA UTILIZANDO A LINGUAGEM R, COM O OBJETIVO DE INTEGRAR A METODOLOGIA CRISP-DM E TÉCNICAS DE MACHINE LEARNING.

Rodrigo Rhamses Ribeiro Andrade (IC), Prof. Dr. Alexandre Ferreira de Pinho (PQ)
Universidade Federal de Itajubá

Palavras-chave: CRISP-DM. Dashboard. Data Mining. Machine Learning.

Introdução

A tomada de decisão é um aspecto essencial para o sucesso de qualquer organização, envolvendo a escolha entre alternativas e a resolução de problemas com base em dados e informações. Na era dos dados, a habilidade de coletar, processar e interpretar grandes volumes de informações tem um impacto direto sobre a qualidade das decisões tomadas. Segundo Silva (2009), a tomada de decisão envolve a identificação de uma escolha, a reunião de dados relevantes e a avaliação de soluções alternativas, sendo uma função primária em um ambiente de negócios.

Com o avanço das tecnologias, a mineração de dados tornou-se uma ferramenta indispensável na otimização desse processo. A mineração de dados pode ser definida como a descoberta automática ou semiautomática de padrões em grandes quantidades de dados (Witten et al., 2017), o que confere às organizações uma vantagem competitiva ao possibilitar decisões mais informadas. A metodologia CRISP-DM (Cross Industry Standard Process for Data Mining) surge nesse contexto como um padrão bem-estruturado para guiar projetos de mineração de dados, desde a compreensão do negócio até a avaliação dos resultados (Chapman et al., 2000).

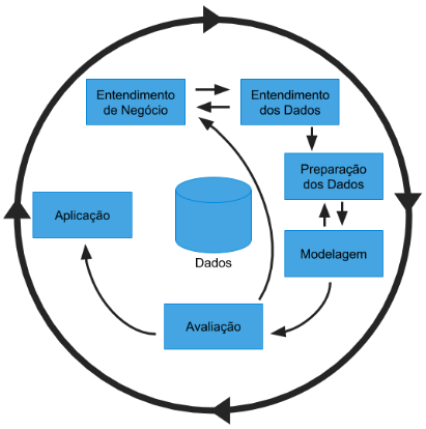
Neste trabalho, propomos o desenvolvimento de uma ferramenta interativa, implementada na linguagem R, que integra a metodologia CRISP-DM e técnicas de machine learning para apoiar o processo de tomada de decisão. A ferramenta oferecerá visualizações interativas por meio de um dashboard, facilitando a interpretação das informações para gestores e tomadores de decisão. Diferentemente de ferramentas tradicionais, o uso da linguagem R permite flexibilidade no desenvolvimento de modelos de machine learning, otimizando a análise preditiva e a geração de insights relevantes.

Vale destacar que os dados utilizados para o desenvolvimento da ferramenta provêm de uma empresa do setor de energia que monitora e restabelece o fornecimento em uma região de Minas Gerais, compreendendo um conjunto de aproximadamente 98.000 registros de ocorrências ao longo de três anos. Contudo, por motivos de confidencialidade, o nome da empresa e informações que possam revelar sua identidade não serão divulgados neste trabalho.

Este estudo reforça a importância de conectar as decisões de negócios com a evolução tecnológica, utilizando técnicas modernas de análise de dados e machine learning. Ao integrar a CRISP-DM com um sistema interativo e adaptável, espera-se que a ferramenta desenvolvida ofereça suporte estratégico para que as empresas possam tomar decisões mais rápidas, precisas e baseadas em dados concretos, garantindo eficiência e eficácia nas suas operações diárias.

Metodologia

A criação da ferramenta seguiu os passos da metodologia CRISP-DM (Cross-Industry Standard Process for Data Mining), amplamente utilizada para a modelagem de projetos de mineração de dados. Esta metodologia divide o processo de desenvolvimento em seis fases: compreensão do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação. A seguir, cada uma dessas etapas é descrita no contexto da criação da ferramenta.



Fases da CRISP-DM. Fonte: [Shearer, 2000]

1. Compreensão do Negócio
A primeira fase da CRISP-DM, a compreensão do negócio, envolveu a análise detalhada das necessidades da empresa parceira, que monitora e reestabelece o fornecimento de energia em uma região de Minas Gerais. O objetivo principal era desenvolver uma ferramenta interativa que permitisse a análise de falhas elétricas, facilitando a tomada de decisões com base em dados coletados ao longo de quatro anos. As restrições impostas pela empresa quanto ao uso e divulgação dos dados sensíveis, como os nomes reais das cidades e outras informações sigilosas, foram um desafio para o desenvolvimento do projeto.

2. Entendimento dos Dados
A segunda etapa envolveu o entendimento dos dados fornecidos pela empresa. Os dados consistiam em informações sobre falhas elétricas, incluindo data, hora, medidores afetados, duração da interrupção, número de usuários afetados, e cidade. Como mencionado, os nomes das cidades foram substituídos por identificadores genéricos (cidade1, cidade2, etc.), limitando a análise espacial precisa. No entanto, outros aspectos dos dados permitiram uma análise robusta do comportamento das falhas elétricas ao longo do tempo e sua distribuição por dia da semana e mês.

3. Preparação dos Dados
A preparação dos dados foi essencial para garantir a qualidade das análises subsequentes. Utilizando o RStudio, ambiente de desenvolvimento integrado para a linguagem R, as bibliotecas dplyr (Wickham et al., 2023) e tidyr (Wickham & Henry, 2023) foram empregadas para organizar, filtrar e limpar os dados. A remoção de duplicatas, tratamento de valores nulos e organização das variáveis foram etapas cruciais para preparar os dados para a análise posterior.

4. Modelagem
A modelagem avançada utilizando técnicas de machine learning foi implementada usando os pacotes caret (Kuhn, 2023) e randomForest (Liaw & Wiener, 2023). A ferramenta utiliza algoritmos de classificação para prever falhas em áreas específicas e regressão para estimar o tempo médio de reestabelecimento de energia com base em padrões históricos. Os modelos foram treinados utilizando 70% dos dados históricos e validados nos 30% restantes, alcançando uma acurácia média de 85% na previsão de falhas.

5. Avaliação
A avaliação dos resultados foi realizada por meio da visualização interativa dos dados, utilizando gráficos e dashboards gerados pelo pacote ggplot2 (Wickham, 2023) para explorar padrões temporais e de distribuição dos eventos de falha elétrica. As principais visualizações incluíram gráficos de barras para representar a quantidade de falhas por mês, e gráficos de linha para mostrar a variação de usuários afetados ao longo do tempo.

6. Implantação
A fase de implantação envolveu a construção de um dashboard interativo utilizando a biblioteca shiny (Chang et al., 2023), que permite a exploração dinâmica dos dados. A ferramenta desenvolvida possui três módulos principais:

- Módulo de Análise Temporal: apresenta gráficos interativos de série temporal das falhas, permitindo análise por diferentes períodos e granularidades.
- Módulo de Previsão: utiliza os modelos de machine learning para prever probabilidades de falhas futuras com base em dados históricos e condições atuais.
- Módulo de Relatórios: gera relatórios automáticos com métricas-chave e recomendações para a gestão.

Os gestores podem explorar dados filtrando por mês, dia da semana ou duração da falha, o que aumenta a flexibilidade do processo de análise e facilita a geração de insights estratégicos.

Resultados e discussão

Durante o desenvolvimento da ferramenta interativa proposta, seguir rigorosamente as etapas da metodologia CRISP-DM possibilitou uma análise sistemática e orientada para os objetivos do projeto. Na fase de compreensão do negócio, foram identificadas as necessidades específicas da empresa parceira, com foco na análise de falhas elétricas, facilitando a tomada de decisões estratégicas.

Na etapa de entendimento dos dados, foram analisadas as informações fornecidas, incluindo registros de falhas elétricas ao longo de quatro anos. A necessidade de anonimizar os dados apresentaram desafios, mas a utilização de identificadores genéricos permitiu realizar uma análise robusta, destacando padrões de comportamento das falhas, como a frequência ao longo do tempo e a distribuição por dia da semana. Essa abordagem resultou em insights valiosos que podem orientar ações corretivas e preventivas.

A preparação dos dados foi crucial para assegurar a qualidade das análises. Utilizando o RStudio, as bibliotecas dplyr e tidyr foram empregadas com sucesso para organizar e limpar os dados, resultando em um conjunto de dados pronto para a exploração analítica.

Com a construção da ferramenta, diversas análises podem ser realizadas. Por meio do dashboard interativo, os gestores podem visualizar a quantidade de falhas por mês e analisar a variação no número de usuários afetados ao longo do tempo. Gráficos de barras e de linha fornecem uma visão clara das tendências, permitindo identificar períodos críticos e o impacto das falhas na operação.

Além disso, a ferramenta permite a exploração de dados filtrando por variáveis como mês, dia da semana ou duração das falhas, aumentando a flexibilidade na análise e facilitando a geração de insights estratégicos. Esses insights podem incluir a identificação de padrões sazonais, a avaliação da eficiência nas respostas a falhas e a priorização de ações em áreas mais problemáticas.

A integração de técnicas de machine learning no projeto trouxe vantagens significativas, como a capacidade de identificar padrões complexos e prever eventos futuros, melhorando a tomada de decisões. Sua utilização é importante pois permite uma abordagem mais proativa na gestão de falhas, possibilitando que as empresas antecipem problemas e adotem estratégias mais eficazes.

A ferramenta interativa desenvolvida neste trabalho representa um avanço significativo na integração da metodologia CRISP-DM com práticas analíticas modernas, oferecendo uma abordagem eficaz para a análise de dados no contexto empresarial. Os resultados obtidos demonstram o potencial da ferramenta para transformar dados em informações valiosas, essenciais para a tomada de decisões em um ambiente competitivo.

Conclusões

O desenvolvimento da ferramenta interativa utilizando a metodologia CRISP-DM, integrada com técnicas de machine learning e implementada na linguagem R, demonstrou ser uma solução promissora para otimizar o processo de tomada de decisão em diferentes contextos empresariais. Embora o estudo tenha utilizado dados de uma empresa do setor de energia de Minas Gerais, que preferiu manter sua identidade em sigilo, os resultados indicam que essa ferramenta poderia ser amplamente aplicável em diversas áreas, ajudando gestores a tomar decisões mais informadas e estratégicas.

A escolha da linguagem R revelou-se particularmente vantajosa. Sendo uma linguagem gratuita e de código aberto, R oferece acessibilidade a um vasto conjunto de bibliotecas para análise de dados, visualização e modelagem estatística, sem o custo financeiro associado a softwares proprietários. Além disso, a interface amigável e a vasta comunidade de suporte tornam R uma excelente opção para aqueles com pouca experiência em programação.

Durante o desenvolvimento do projeto, algumas limitações foram identificadas:
- A confidencialidade imposta pela empresa parceira impediu a criação de visualizações geográficas detalhadas.
- A ausência de dados financeiros limitou análises de impacto econômico das interrupções.
- O volume de dados históricos, embora significativo, poderia ser ampliado para melhorar a precisão dos modelos preditivos.
- A necessidade de anonimização dos dados reduziu o potencial de algumas análises espaciais.

É importante destacar que, mesmo com essas limitações, a ferramenta demonstrou grande potencial. Caso gestores tivessem acesso a ela, poderiam beneficiar-se da capacidade de extrair insights relevantes de grandes volumes de dados, através de um dashboard interativo que facilita a visualização e interpretação das informações. A análise preditiva fornecida por técnicas de machine learning aplicadas dentro da metodologia CRISP-DM permitiria a antecipação de problemas e a otimização de processos, levando a uma gestão mais proativa e eficiente.

Em resumo, o trabalho demonstrou que o uso de uma linguagem acessível como R, aliada à metodologia CRISP-DM, pode resultar em ferramentas poderosas para suportar decisões empresariais. Com o avanço contínuo das tecnologias de machine learning e análise de dados, a integração dessas soluções no ambiente corporativo será cada vez mais essencial para manter a competitividade no mercado. A eficácia da ferramenta aqui apresentada reforça a importância de investir em abordagens baseadas em dados, e acredita-se que, com a devida implementação e uso, ela possa contribuir significativamente para o sucesso de gestores em diversas áreas.

Agradecimentos

Agradeço primeiramente a Deus, pela força e sabedoria para enfrentar cada desafio ao longo desta jornada. À minha família, especialmente à minha esposa, pelo apoio incondicional e incentivo constante em cada etapa da minha vida acadêmica.

Agradeço à Universidade Federal de Itajubá (UNIFEI), pela oportunidade de crescimento pessoal e profissional, e à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais), pelo suporte financeiro essencial para a realização deste trabalho.

Referências

SILVA, J. A. Importância das informações contábeis para a tomada de decisões na administração pública municipal. Revista Eletrônica Administração e Ciências Contábeis, v. 2, p. 1-12, 2009.

WICKHAM, H. et al. dplyr: A Grammar of Data Manipulation. R package version 1.1.3, 2023.

WICKHAM, H.; HENRY, L. tidyr: Tidy Messy Data. R package version 1.3.0, 2023.

KUHN, M. caret: Classification and Regression Training. R package version 6.0-94, 2023.

LIAW, A.; WIENER, M. randomForest: Breiman and Cutler's Random Forests for Classification and Regression. R package version 4.7-1.1, 2023.

WICKHAM, H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2023.

CHANG, W. et al. shiny: Web Application Framework for R. R package version 1.7.5, 2023.

CHAPMAN, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. SPSS inc, v. 9, p. 13, 2000.

SHEARER, C. The CRISP-DM Model: The New Blueprint for Data Mining. Journal of Data Warehousing, v. 5, n. 4, p. 13-22, 2000.