PONTIFÍCIA UNIVERSIDADE CATÓLICA DE CAMPINAS

GABRIEL DE ANTONIO MAZETTO

MATEUS PEREIRA ALVES

ASSISTENTE DE ANÁLISE DE DADOS COM LLM

CAMPINAS

2025

1. TEMA DO PROJETO

O projeto consiste no desenvolvimento de uma aplicação web interativa que opera de forma totalmente local, com o objetivo de auxiliar o usuário em tarefas de análise e exploração de dados. A ferramenta utiliza um Modelo de Linguagem Grande (LLM) da API do Google Gemini como um planejador e editor de código, mas todo o processamento dos dados, desde a ingestão até a visualização, é realizado no ambiente do usuário.

A construção do sistema é pautada na criação de uma base de conhecimento com no mínimo 100 amostras de código. Essa base será estruturada para conter uma breve descrição do que o código faz, o próprio código formatado como função, e uma lista de bibliotecas necessárias para sua execução. O LLM terá acesso inicial apenas às descrições, escolhendo a mais relevante antes de ter acesso ao código. A compilação e organização da base de dados serão realizadas a partir de fontes confiáveis, como documentações oficiais do Pandas, Seaborn e Scikit-learn, complementadas por notebooks do Kaggle e trabalhos acadêmicos prévios da própria equipe do projeto.

Para a avaliação do sistema, será realizada uma abordagem envolvendo uma análise tanto quantitativa quanto qualitativa. A análise quantitativa utilizará testes unitários automatizados em uma base de dados de teste para comparar os resultados do código gerado com os resultados das funções de referência da base. Um ponto central da avaliação será a análise comparativa do desempenho com e sem a consulta à base de conhecimento, medindo a taxa de sucesso de primeira tentativa e a contagem de erros. Complementarmente, uma discussão qualitativa abordará a legibilidade e a qualidade do código gerado, além de uma comparação subjetiva dos resultados, com e sem a consulta à base de códigos.

2. JUSTIFICATIVA

A escolha deste projeto é orientada pela busca de resolver o conflito entre a utilização de LLMs na ciência de dados e as preocupações com a privacidade dos dados. A necessidade de enviar informações sensíveis para servidores de terceiros, que gera preocupações com a conformidade com regulamentações como a Lei Geral de Proteção de Dados (LGPD), é uma limitação que a nossa solução, uma aplicação local, visa mitigar.

A seguir, o **Quadro 1** resume os pontos relacionados as certezas, suposições e dúvidas.

Quadro 1 – Matriz CSD (Certezas, Suposições, Dúvidas)

Categoria	Tópicos
Certezas (C)	• O sistema é local e garante a privacidade dos dados. • A aplicação é viável com Python, Pandas e Streamlit. • A API do Gemini pode gerar código. • A execução do código gerado pelo LLM é segura com o aval do usuário.
Suposições (S)	• O LLM pode gerar consistentemente códigos corretos. • A execução de pip install via subprocesso é confiável em todos os sistemas operacionais. • A base de dados de código será suficiente e efetiva. • A performance do sistema será adequada para *datasets* grandes.
Dúvidas (D)	• Qual a robustez da geração de código para tarefas complexas? • Quais os limites técnicos e de custo computacional? • O sistema pode lidar com dados brutos que requerem tratamento ético? • Quais os desafios de segurança para evitar a execução de código malicioso?

Fonte: Elaborado pelos autores (2025).

As certezas do projeto nos fornecem uma base técnica viável e confiável. O uso de um sistema local é eficaz para garantir a privacidade dos dados do usuário. A aplicação em Python, com bibliotecas como Pandas e Matplotlib, oferece uma base

sólida para o processamento de dados. A execução de comandos “pip install” via subprocesso para instalar dinamicamente as dependências necessárias é um processo conhecido. A API do Gemini pode gerar código funcional, e a confirmação do usuário para a execução garante um nível adicional de segurança.

O projeto é guiado pela necessidade de validar uma série de suposições cruciais para a aplicação de LLMs. Uma delas é que o LLM pode gerar consistentemente códigos corretos. A viabilidade de construir a base de dados de código com base em fontes já existentes, como a documentação de bibliotecas, é também uma suposição que será validada ao longo do desenvolvimento. Acreditamos que, ao alimentar o LLM com descrições e códigos pré-organizados, ele será capaz de gerar respostas mais consistentes. Além disso, supomos que a execução de “pip install” via subprocesso será confiável em diferentes sistemas operacionais, e que a performance do sistema será adequada para *datasets* grandes.

Por fim, as dúvidas do projeto constituem os principais desafios acadêmicos a serem explorados. Questões como a robustez da geração de código para tarefas complexas, os limites técnicos e de custo computacional e as implicações éticas, incluindo o potencial de viés no código gerado e a responsabilidade no seu uso, terão foco da nossa discussão crítica.

Apesar dos desafios e das questões em aberto, este projeto representa uma abordagem para democratizar a análise de dados, oferecendo uma ferramenta poderosa que preserva a privacidade e a segurança do usuário.