Assistente de Análise de Dados com LLM

Autores:

- Gabriel de Antonio Mazetto
- Mateus Pereira Alves

Disciplina: Tópicos em Ciência de Dados

Instituição: Pontifícia Universidade Católica de Campinas

Definição do Problema

Conflito

O uso de LLMs para análise de dados versus a necessidade de garantir a privacidade e segurança da informação, conforme a LGPD.

Solução

Usar uma API de LLM para planejar tarefas e gerar código, enviando apenas metadados não sensíveis, enquanto todo o processamento dos dados do usuário ocorre em um ambiente 100% local.

Objetivo Geral

O quê

Desenvolver uma aplicação web interativa e totalmente local que auxilia usuários em tarefas de análise e exploração de dados.

Pilar

A aplicação será construída sobre uma base de conhecimento (Knowledge Base - KB) com mais de 100 amostras de código de alta qualidade, que guiará o LLM na geração de soluções mais precisas e seguras.

Abordagem Proposta

01

Aplicação Local

Uma ferramenta web onde o usuário faz o upload do seu dataset, que nunca deixa sua máquina.

02

Interação com LLM

O sistema usa um LLM (Google Gemini) que recebe metadados do dataset e as descrições das funções disponíveis na base de conhecimento para planejar a tarefa solicitada.

03

Base de Conhecimento (kb.jsonl)

O núcleo do sistema, contendo funções Python reutilizáveis, tipadas e com documentação.

Cada função é catalogada em um arquivo JSONL com metadados como id, categoria, descricao, bibliotecas e o codigo_funcao.

04

Execução Segura

O código gerado pelo LLM só é executado após a revisão e aprovação explícita do usuário.

Metodologia (1/2) - Construção da Base

Fonte dos Códigos

Documentações oficiais (Pandas, Seaborn), exemplos do livro "Projetos de Ciência de Dados com Python" e códigos de projetos pessoais.

Fluxo de Incorporação

1. ### Extração

   Identificar trechos de código úteis e generalizáveis.

2. ### Adaptação

   Converter o código em uma função padronizada, com tipagem e docstrings no padrão Google.

3. ### Teste de Incorporação

   Validar a função em dois datasets distintos (dataset usado pelo livro e um segundo dataset como o Titanic).

4. ### Ciclo de Correção

   Em caso de falha, realizar até 3 tentativas de correção. Se o erro persistir, a função é movida para um registro de rejeitadas.

5. ### Incorporação Final

   Se aprovada, a função é adicionada ao arquivo kb.jsonl.

Metodologia (2/2) - Avaliação do Sistema

Análise Quantitativa

- Execução de testes unitários para comparar o resultado do código gerado pelo LLM com o resultado das funções de referência da base.
- Análise comparativa de desempenho com e sem a consulta à base de conhecimento, medindo a taxa de sucesso na primeira tentativa.

Análise Qualitativa

- Discussão sobre a legibilidade, qualidade e manutenibilidade do código gerado.
- Comparação subjetiva dos resultados obtidos com e sem o auxílio da base de códigos.

Resultados Esperados

1

Produto Final

Uma aplicação funcional que permite a análise de dados de forma segura e privada.

2

Base de Conhecimento

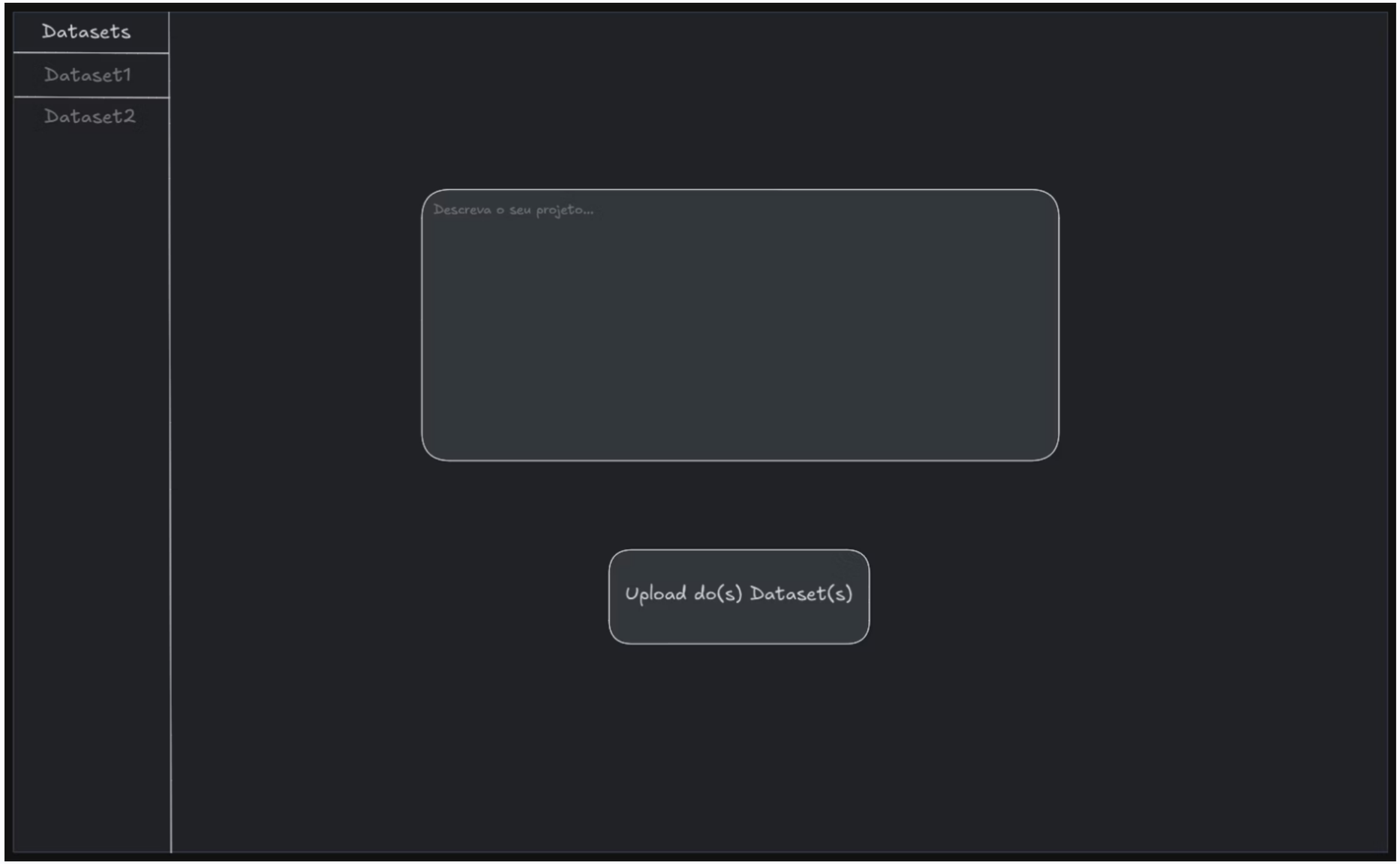
Um dataset com no mínimo 100 funções de análise de dados testadas e prontas para uso.

3

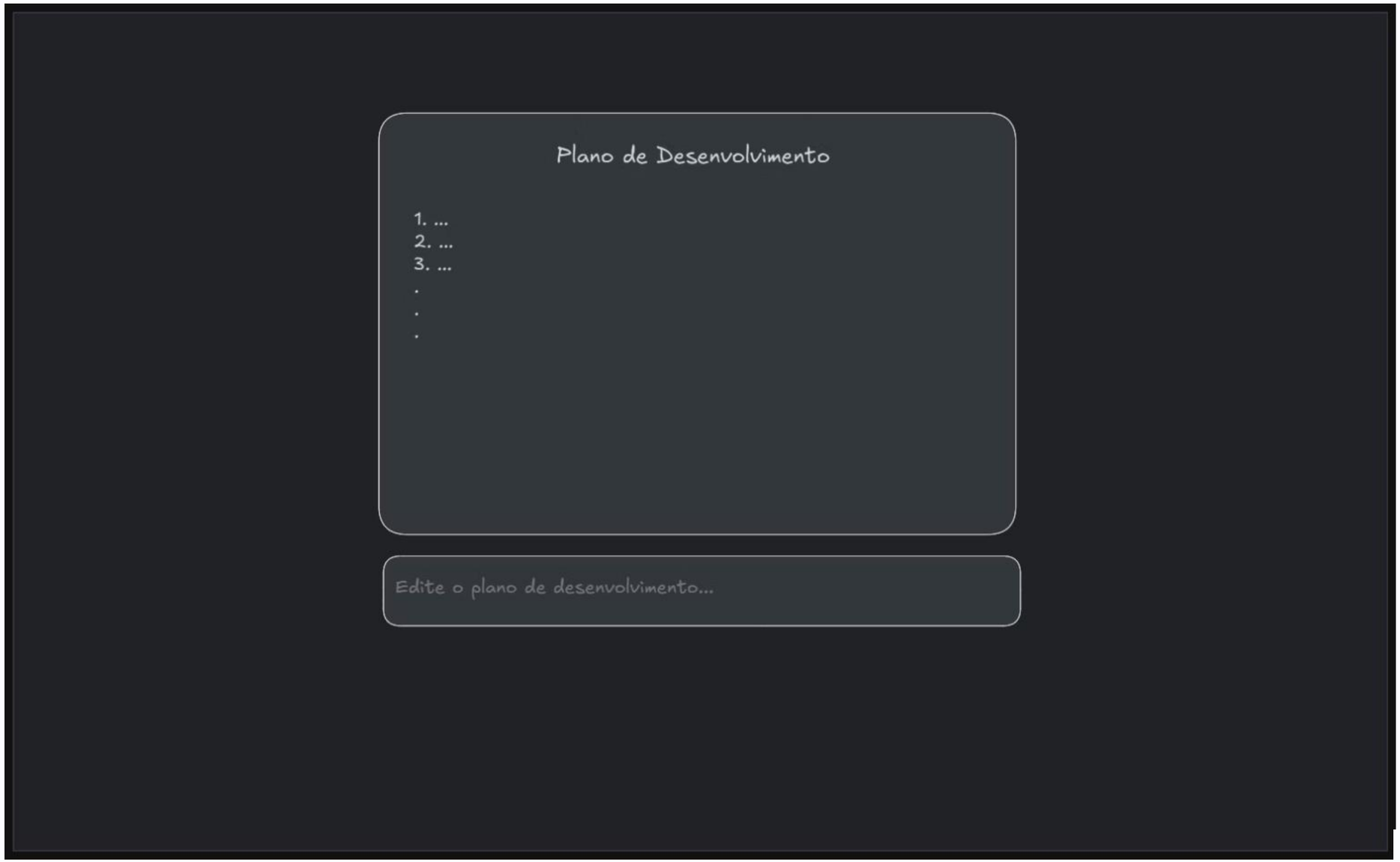
Validação da Hipótese

Evidências quantitativas e qualitativas de que uma base de conhecimento melhora a precisão e a confiabilidade de LLMs para tarefas de programação em ciência de dados.

Esboços do software



Esboços do software



Esboços do software