1. Caracterização do Dataset

Descrição Geral

O dataset analisado reúne informações de *Pull Requests (PRs)* coletadas de repositórios open source escritos em Python e hospedados no GitHub, utilizados no estudo “Impactos da Inteligência Artificial no Desenvolvimento de Software na Linguagem Python”.
Os dados foram obtidos por meio da API pública do GitHub, representando atividades de contribuição e colaboração entre desenvolvedores em projetos de código aberto.

O foco é avaliar mudanças antes e depois da popularização dos LLMs (Large Language Models), considerando dimensões de produtividade, qualidade e colaboração.

Composição do Dataset

Categoria	Descrição
Unidade de Análise	Pull Requests (PRs)
Origem dos Dados	GitHub REST API
Linguagem dos Projetos	Python
Período Coletado	2018 a 2025
Formato do Arquivo	CSV (**`dataset_pulls.csv`**)
Principais Campos	*title, author, created_at, merged_at, is_merged, closed_at, state, base_ref, user_login*

Características Principais

1. **Quantidade de registros (PRs):** número total de pull requests coletados no período.
2. **Quantidade de repositórios analisados:** número total de projetos Python incluídos no estudo.
3. **Intervalo temporal:** período entre o PR mais antigo e o mais recente.
4. **Distribuição temporal:** PRs abertos por ano e por mês.
5. **Status dos PRs:** proporção entre PRs *merged*, *closed* sem merge e *open*.
6. **Tempo médio de integração:** tempo médio entre created_at e merged_at.
7. **Colaboradores:** número de autores únicos e distribuição de contribuições por autor.

Particionamento Temporal para Análise Comparativa

O dataset é dividido em dois subperíodos para fins de comparação longitudinal:

Período	Descrição	Marco Temporal
Pré-LLMs	Período anterior à popularização de LLMs (GitHub Copilot, ChatGPT, etc.)	até dezembro de 2022
Pós-LLMs	Período posterior à adoção massiva de copilotos de IA em ambientes de desenvolvimento	janeiro de 2023 em diante

Essa divisão permite comparar indicadores de produtividade, qualidade e colaboração antes e depois da difusão das ferramentas baseadas em IA.

Variáveis-Chave por Dimensão de Análise

Dimensão	Métricas	Fonte
Produtividade	Nº de PRs/mês, commits/mês, tempo médio de ciclo de PR	created_at, merged_at
Qualidade	Taxa de merge (%), tempo até merge, número de revisões e comentários	state, comments
Colaboração	Nº de autores únicos, PRs por autor, diversidade de contribuidores	author, user_type