MAYARA PERRONI

CIENTISTA DE DADOS

mayara.perroni9@gmail.com
/in/mayara-perroni-50920615b/
/in/mayara-perroni-50920615b/
São Carlos, São Paulo, Brasil

Resumo

Engenheira de Dados e Cientista de Dados Acadêmica com experiência em processamento de linguagem natural (NLP), aprendizado de máquina e engenharia de dados escalável. Atuação em projetos que envolvem integração de dados, desenvolvimento de modelos de machine learning e recomendação, além da implementação de soluções eficientes para análise e predição de eventos. Familiaridade com LLMs para estruturar dados não estruturados em contextos biomédicos.

Hard-skills

- Linguagens de Programação: Python (Pandas, NumPy, Scikit-Learn, PySpark), SQL, R
- Machine Learning e Inteligência Artificial: Modelagem de dados, Regressão, Classificação, Redes Neurais
- Processamento de Linguagem Natural (NLP): Pré-processamento de texto, Modelos de embeddings, LLMs (GPT, BERT), PyTorch, TF-IDF,
- Big Data e Engenharia de Dados: Apache Airflow, AWS S3, Glue, Dremio
- Bancos de Dados: PostgreSQL, BigQuery, MySQL
- ETL e Pipelines de Dados: Transformação de dados no OMOP-CDM, Airbyte
- Visualização de Dados: Power BI, Dashboards interativos, Metabase
- Boas Práticas de Engenharia de Software: Testes unitários, Clean Code, versionamento com Git (GitHub, GitLab, Bitbucket)

Experiências

Ciência de Dados | Projeto de Mestrado | Programa de Pós-Graduação Interunidades em Bioengenharia (PPGIB) - EESC-USP - FMRP-USP - IQSC-USP ***2025-2026***

Título: Integração de Dados Clínicos Não Estruturados para Predição de MACE com Machine Learning

- *A*plicação e avaliação de técnicas avançadas de Processamento de Linguagem Natural (PLN) para estruturar dados não estruturados provenientes de RES, a fim de aprimorar a predição de Eventos Cardiovasculares Adversos Maiores (MACE) no modelo PreCare ML
- Exploração de diversas abordagens de PLN, incluindo embeddings baseados em Large Language Models (LLMs), modelos de tópicos, regras linguísticas e arquiteturas híbridas que combinem Term Frequency-Inverse Document Frequency (TF-IDF) e redes neurais

Engenharia de Dados & Ciência de Dados | Trabalho de conclusão de curso ***2024***

Título: Desenvolvimento e avaliação de modelos de machine learning para predição de eventos cardiovasculares adversos - MACE)

- Desenvolvimento de modelos de NLP e LLMs para estruturar campos livres de registros eletrônicos de saúde, aumentando a disponibilidade de variáveis para predição de MACE.
- Implementação de pipelines de pré-processamento de dados em PySpark e SQL, permitindo escalabilidade no processamento de grandes bases de dados médicas padronizadas no OMOP-CDM.
- Construção de modelos de machine learning supervisionados e não supervisionados para análise preditiva.
- Integração de dados anonimizados de múltiplos centros médicos em um repositório único no BigQuery.
- Extraction, Transformation and Loading (ETL), padronizando dados de registros eletrônicos de saúde (RES) utilizando o Observational Medical Outcomes Partnership (OMOP) - Common Data Model (CDM)
- LLM focado na interpretação semântica e contextual dos textos
- TF-IDF focado na identificação de padrões de palavras-chave associados ao tabagismo
- Modelos testados e avaliados em termos de precisão, recall e capacidade de generalização, considerando a variabilidade dos dados
- Estudo que contribui para o desenvolvimento de classificadores automáticos que podem ser integrados a sistemas de prontuários eletrônicos, aprimorando a categorização de informações críticas, como o histórico de tabagismo, e servindo de base para futuras análises de fatores de risco em registros padronizados

MAYARA PERRONI

CIENTISTA DE DADOS

mayara.perroni9@gmail.com
/in/mayara-perroni-50920615b/
/in/mayara-perroni-50920615b/
São Carlos, São Paulo, Brasil

Experiências

Engenharia de Dados | Somativa

julho de 2023 - o momento

(Implementação de Arquiteturas de Dados e Integração de Sistemas de diversas fontes de dados)

- Desenvolvimento de pipelines de ingestão e transformação de dados utilizando Airbyte, AWS S3, Airflow e AWS Glue, garantindo um fluxo contínuo e confiável de dados entre sistemas distintos.
- Criação de modelos de dados no Drêmio para análise em Power BI, otimizando dashboards para insights e geração de valor.
- Identificação de inconsistências entre fontes de dados (Sapiens, Salesforce e outras bases de dados), permitindo maior precisão nas análises.
- Modelagem de Dados: Experiência na construção de modelos de dados eficientes para representar de forma precisa a realidade do negócio.
- RAW, Trusted e Refined: Expertise na criação de camadas de dados, desde a fase inicial e bruta (RAW), passando pela confiável (Trusted), até alcançar o estado refinado (Refined), garantindo integridade e qualidade em todas as etapas do processo.
- Desenvolvimento de Dashboards: Criação de dashboards impactantes para diversos setores da empresa, traduzindo dados de diferentes fontes (SQL Server, Postgress, Excel, Salesforce) em insights visuais acessíveis, proporcionando uma compreensão mais abrangente e facilitando a tomada de decisões estratégicas.

Cientista de Dados - Carefy

Janeiro de 2023 - Junho de 2023

- Manutenção da ETL na estrutura com e sem a utilização do Prefect;
- Manutenção de modelo de NLP;
- Formulação de projetos iniciais de IA pensando no produto da empresa;
- Painéis (Dashboards) construídos no Metabase que estão disponíveis para mais de 100 usuários diariamente;
- Python para manipulação de dados;
- Manutenção no banco de dados;
- Manutenção de Dashboards.

Pesquisador Técnico - FAPESP

Dezembro de 2021 - Abril de 2023

- Composição de uma imagem médica digital;
- Segmentação de imagens médicas com o uso do 3D Slicer;
- Ambientes de realidade mista (aumentada e virtual);
- Análise de diferentes de segmentação semi-automatizada;
- Python para segmentação automatizada;
- Reconhecimento de padrão em imagens médicas;
- Entrega de modelos 3D para posterior impressão.

Data Anottator - Serasa Experian

Junho de 2021 - Agosto de 2023

- Familiaridade com Técnicas e Ferramentas de Anotação;
- Precisão e Consistência nos Dados Anotados.

Educação

- Instituto Federal de Educação, Ciência e Tecnologia de São Paulo - Técnico em Tecnologia ***2017 - 2019***
- Universidade de São Paulo - Bacharel - Informática Biomédica ***2021 - 2024***

Idiomas

Português - Nativo
Inglês - Intermediário