

Israel Souza Ferreira

Cientista de Dados | Machine Learning

Fortaleza, CE | Remoto | +55 85 98433-2790 | israel.souza.dev@gmail.com

[LinkedIn](#) | [GitHub](#) | [Portfólio: israelsouza.speculummaius.com.br](#)

PERFIL

Cientista de dados com atuação no ciclo completo do problema: extração e tratamento de dados em escala, modelagem estatística e preditiva, desenho de experimentos e comunicação de resultados. Experiência em P&D com bolsa CNPq aplicando **Teoria de Resposta ao Item**, **séries temporais**, **modelos de linguagem** e avaliação rigorosa de modelos a problemas de educação e saúde pública. Formação em Ciência da Computação com base sólida em estatística, otimização e SQL. O diferencial está na avaliação: medir antes de concluir e desconfiar do próprio resultado.

EXPERIÊNCIA PROFISSIONAL

Tieta Artificial Intelligence

Cientista de Dados / Pesquisador de P&D — Bolsista CNPq RHAÉ

Remoto, Brasil

Abr 2025 – Jun 2026

- Modelei parâmetros psicométricos latentes de itens educacionais com **Teoria de Resposta ao Item**, validando as estimativas dos modelos contra parâmetros de referência.
- Desenhei **avaliações pareadas** sobre os mesmos itens para comparar abordagens concorrentes sob o mesmo protocolo, quantificando a correlação de cada sinal estimado com os parâmetros de referência e o custo de inferência associado.
- Construí pipelines de extração de sinais com **modelos de linguagem** orquestrados por **DSPy**.
- Testei arquiteturas de **regressão multitarefa** sobre dados textuais e visuais.
- Estruturei experimentos reproduzíveis, métricas de avaliação e documentação técnica para sustentar as conclusões do ciclo de pesquisa.

Tieta Artificial Intelligence

Pesquisador de P&D — Programa de Formação em Pesquisa e Inovação

Remoto, Brasil

Jul 2026 – Presente

- Desenvolvo um pipeline de **segmentação de questões** de provas escaneadas via OCR, extraíndo itens estruturados para análise educacional.

Consultoria autônoma — Instituição de ensino

Desenvolvedor Full Stack e Engenheiro de Machine Learning

Remoto, Brasil

2026 – Presente

- Entreguei **individualmente** um sistema contratado de registro de ponto por reconhecimento facial, funcional e em fase final de homologação para implantação, cobrindo interface, modelagem, dados, backend, testes e deploy.
- Treinei um classificador EfficientNet-B0 para triagem automática de atestados com **busca de hiperparâmetros** (Optuna), validação cruzada estratificada e decisão híbrida CNN + OCR; diagnostiquei o subajuste por **ablação controlada** e elevei a acurácia de validação em **13 pontos percentuais**.
- Implementei **busca vetorial** 1:N e verificação 1:1 sobre embeddings faciais em $\mathbb{R}^{512}$ com InsightFace e pgvector, agregando decisões *multi-frame* para reduzir falsos positivos, com *liveness* e proteção anti-replay.
- Construí a camada de dados e o serviço de inferência com PostgreSQL, SQLAlchemy, Alembic e FastAPI; entreguei infraestrutura Dockerizada com HTTPS, GitHub Actions e **142 testes automatizados**.

Universidade Federal do Ceará — Campus Quixadá

Monitor de Algoritmos, Cálculo e Pré-Cálculo

Presencial, Brasil

2022 – 2024

- Ministrei apoio em estruturas de dados, algoritmos, análise de complexidade e Cálculo, traduzindo conteúdo quantitativo para turmas de graduação.

PROJETOS SELECIONADOS

- Previsão de Casos de Dengue (séries temporais):** previsão de notificações a partir dos microdados do SINAN, com ETL, modelos LSTM e PLE, rastreo de experimentos em MLflow e serviço de inferência; reduzi o **RMSE de ≈ 98 para 83** na série do Ceará (queda de $\approx 15\%$). PostgreSQL, PyTorch, FastAPI e Docker.
- Auxílio Emergencial — Análise de Desempenho em 31,6 GB:** trabalho acadêmico retomado por iniciativa própria e reconstruído para escala real sobre **31,6 GB** de dados públicos; a instrumentação por etapa correlacionou o crescimento de memória aos identificadores de deduplicação retidos ($r = 0,99$), isolando a causa raiz do gargalo. Ingestão em blocos com COPY binário assíncrono, PostgreSQL, Pandas e Streamlit.
- Triple Roman Domination in Graphs (TCC):** primeira abordagem meta-heurística apresentada na literatura para o problema, com Algoritmo Genético, MAX-MIN Ant System e RVNS, usando Programação Linear Inteira como *baseline* exato de validação.

HABILIDADES TÉCNICAS

Linguagens e consultas: Python, SQL, C++ e Bash.

Análise e estatística: Pandas, NumPy, SciPy, Scikit-learn, testes de hipótese, análise de correlação, validação cruzada, calibração de modelos, análise de ablação e Teoria de Resposta ao Item.

Modelagem e machine learning: PyTorch, Optuna, DSPy, FAISS, InsightFace, OpenCV; séries temporais (LSTM, PLE), regressão multitarefa e multimodal, MoE/MMoE, atenção, busca vetorial e meta-heurísticas.

Dados, MLOps e comunicação: PostgreSQL, pgvector, SQLAlchemy, Alembic, MLflow, Docker, Git, GitHub Actions, GitLab CI/CD, FastAPI, pytest, mypy, Streamlit, visualização de dados e documentação técnica.

FORMAÇÃO ACADÊMICA

Universidade Federal do Ceará — Campus Quixadá

Bacharelado em Ciência da Computação

Brasil

Conclusão prevista: Jul 2026

IDIOMAS

Português: nativo | Inglês: leitura técnica fluente de artigos e documentações