Abdoulfatah Abdillahi

San Jose, CA
317-828-1996 abdoulsfsu@gmail.com LinkedIn Portfolio

Education

San Francisco State University **Jan. 2024 – May 2026**
Master of Science in Data Science and Artificial Intelligence *San Francisco, CA*

San Francisco State University **Sep. 2018 – Dec 2023**
Bachelor of Science in Computer Science *San Francisco, CA*

Relevant Coursework

- Statistical Learning
- Probability & Statistics
- Artificial Intelligence
- Deep Learning
- NLP
- Data Visualization
- Data Structures
- Software Development

Experience

GCEA **Aug 2026 – Present**
Lead, Data Analytics Intern *Oakland, CA*

- Leading a data quality and standardization initiative across the organization's Google Sheets-based systems, resolving missing values, duplicate records, and inconsistent formats to establish clean, reliable data foundations.
- Conduct exploratory data analysis to surface trends, patterns, and data-quality gaps, translating findings into actionable insights for non-technical leadership.
- Assessing current spreadsheet-based workflows and architecting a centralized relational database to replace fragmented spreadsheets, improving data organization, consistency, scalability, and accessibility across the organization.

CodeLab, San Francisco State University **Jan 2024 – May 2026**
Graduate ML Researcher *San Francisco, CA*

- Developed an end-to-end machine learning system to predict ciprofloxacin resistance from whole-genome sequencing data across **588 *E. coli* isolates**; achieved **0.943 balanced accuracy and 0.987 ROC-AUC** on a held-out test set with Random Forest.
- Designed two genomic feature-engineering pipelines in **Python and Pandas** that transformed a **5.7M-position genome alignment** into independent SNP and gene presence/absence (GPA) representations, implementing coordinate mapping, gap-based encoding, MAF and missingness filtering, one-hot encoding, and feature selection.
- Built a leakage-resistant training and evaluation framework with **scikit-learn, XGBoost, and Optuna**, using an 80/20 stratified split, 3-fold cross-validation, and TPE-based hyperparameter optimization while reserving 118 genomes exclusively for final testing.
- Benchmarked **5 classical ML algorithms and retrieval-augmented TabR** across both genomic representations, demonstrating that GPA features improved balanced accuracy for **5 of 6 models**; Random Forest reached 0.943 balanced accuracy while XGBoost achieved the highest ROC-AUC of **0.989**.
- Applied **Mutual Information and MultiSURF** to interpret genomic predictors, recovering established ciprofloxacin-resistance signals in *parC* and *gyrA* and identifying consistent GPA signals, with **13 of the top 20 genes** shared across both feature-selection methods.
- Developed parallel **SLURM/HPC** workflows to process and annotate **40K+ gene clusters**, diagnosed computational bottlenecks in the initial approach, and redesigned the workflow to support large-scale genomic analysis.

Genentech **Jun 2023 – Sep 2023**
Bioinformatics Intern *South San Francisco, CA*

- Migrated a research Neural-SDE modeling framework from TensorFlow to PyTorch, implementing custom training loops and model components to support continued experimentation and downstream machine learning research.
- Collaborated with ML engineers and scientists in weekly cross-functional reviews to communicate technical progress, experimental findings, and implementation blockers; presented final research to a broader technical audience at Genentech's intern symposium.

Projects

AI Voice Testing Platform | GitHub

- Built an end-to-end **Generative AI voice testing platform** that autonomously conducts outbound calls to evaluate healthcare scheduling agents through multi-turn patient simulations.
- Architected a modular evaluation framework integrating **Vapi, LLM prompting, speech-to-text, and voice synthesis**, separating scenario management, call orchestration, and API integrations for reusable testing workflows.
- Created **10 patient personas** and scenario-based evaluations covering appointment scheduling, medication refills, provider routing, conversational state management, and privacy guardrails.

Energy Consumption Forecasting | GitHub **Feb 2025 – May 2025**

- Built an end-to-end regression pipeline on **19,735 records and 29 features**, benchmarking Linear, Ridge, Lasso, Decision Tree, Random Forest, and tuned ensemble models with ANOVA-based feature selection.
- Reduced forecasting error from **67% to 30% MAPE** versus a linear baseline and applied clustering to identify energy-consumption patterns and behavioral segments.

Shopping Trends & Customer Behavior Dashboard | GitHub **Jan 2024 – May 2024**

- Built an interactive data visualization dashboard in **Python, Dash, and Plotly** to analyze shopping behavior across approximately **3,900 customer records**.
- Developed interactive product, seasonal, geographic, and demographic visualizations with age, gender, and product filters to surface customer purchasing trends and behavior patterns.

Technical Skills

Languages: Python, SQL, R, Bash, Java, C/C++, JavaScript
Machine Learning & AI: scikit-learn, PyTorch, TensorFlow/Keras, XGBoost, Optuna
Data & Visualization: Pandas, NumPy, Dask, Plotly, Dash
ML Engineering & Backend: FastAPI, Flask, SLURM/HPC, Jenkins
Developer Tools: Git, GitHub, Linux, Jupyter, Docker

Honors & Awards

- **NIH – UCSF/SFSU Bridges to the Doctorate Fellowship (2024 – 2026)** – Two-year NIH-funded biomedical research fellowship with approximately **$32K** in support.
- **BMC Scholarship, SFSU (Sep 2024)** – Awarded **$1,750** for academic excellence and research commitment.
- **NDiSTEM Travel Scholarship, SACNAS (Oct 2024)** – Received full airfare, lodging, and registration support for the 2024 NDiSTEM Conference.