AFRAH FATHIMA S

Boston | MA | +1 617-397-8937 | shahabuddin.af@northeastern.edu | LinkedIn | Github

SUMMARY: Graduate student in Computer Science specializing in Data Science and AI, with hands-on experience in data analysis, visualization, machine learning, and building AI-powered solutions. Proficient in Python, SQL, and modern data/ML tools. Passionate about applying AI and data-driven approaches to solve real-world problems through impactful internships and collaborative projects.

EDUCATION

Northeastern University, Boston, MA — Sep 2024 – Present
Master of Science in Computer Science
Relevant Coursework: Algorithms, Machine Learning , Essentials of Data Science, DBMS, Cloud Computing, Web Development

Anna University, Chennai, India — Aug 2020- May 2024
Bachelor of Engineering in Computer Science and Engineering

TECHNICAL SKILLS

Programming Languages:	Python, SQL,C, R, Bash, Java **(for data pipelines and production deployment)**
ML and NLP:	Scikit-learn, TensorFlow, PyTorch, Keras, SpaCy, CNNs , A/B Testing, Prompt Engineering
Statistical Methods:	Hypothesis Testing, Statistical Modeling, Experiment Design, Probability Distributions
Data Engineering & APIs:	Apache Spark, Airflow, FastAPI, Flask, REST APIs, PostgreSql, MongoDB, MySQL
AI & LLM Tools:	OpenAI APIs, Hugging Face Transformers, LangChain
Analysis/Visualization:	Pandas, NumPy, Matplotlib, Seaborn, Plotly, Tableau, Power BI, Excel
Cloud & DevOps:	Docker, Git, AWS(S3, SageMaker,Lambda,EC2), GCP, Terraform, GitHub Actions, Postman

WORK EXPERIENCE

Brainvault Technologies, India | **Software Developer Intern** — Jan 2024 – Jun 2024

- Addressed time-consuming travel planning at TravelNet Solutions by developing a smart assistant using **OpenAI, Hugging Face, FastAPI, MySQL, Beautiful Soup,** and **Nominatim API** for automated itinerary creation and location mapping.
- Delivered features such as automated itineraries, surprise trip planning, and nearby attraction suggestions, **reducing planning time by 80–90%** and improving client engagement..
- Used **LLMs (Gemini, LLaMA)** and **PDFPlumber** to achieve **90–95% field-level accuracy,** streamlining workflows and significantly reducing manual effort via **infrastructure automation**.
- Collaborated cross-functionally to scope, prototype, and deploy AI-powered tools for process automation and internal efficiency.

Waycool Food and Products Pvt. Ltd, India | **Machine Learning Intern** — Sep 2022 – Nov 2022

- Traditional methods for analyzing grain size were slow and error-prone; designed a **CNN**-based deep learning model to identify, count, and measure grain sizes with **85–90% accuracy**, significantly improving precision and consistency.
- Enhanced model performance by applying advanced feature extraction, adaptive thresholding, and custom data preprocessing, **reducing computational time by 85%** without compromising accuracy.
- Applied custom preprocessing, statistical validation, and performance benchmarking to deliver actionable insights from vision data.
- Gained deep expertise in computer vision and **model optimization**, earning top performance scores within the intern cohort and commendation from senior data scientists for quality and impact.

PROJECTS

NLP Agent for Legal Document Understanding | Github-Repo — Jan 2024 – Jun 2024

- Built an automated **legal document extraction system for a legal client**, using **Gemini, LLaMA Parse, spaCy, and regex** to reduce manual review and structure legal data, supporting **ML workflow automation and model deployment using FastAPI.**
- **Fine-tuned NLP models** to boost accuracy, achieving **85–90% extraction precision** and improving processing efficiency for the client.
- Optimized model performance by tuning inference speed and memory usage for multi-page document inputs.

Multilingual Real-Time Profanity Detection with Audio & NLP | Github-Repo — Jan 2024 – May 2024

- Developed a real-time multilingual profanity filter for **Rajalakshmi Engineering College** using **STT, NLP, and deep learning**, achieving **95% accuracy in live audio censorship**.
- **Enhanced processing speed by 30%** through parallelized text analysis and TTS, enabling real-time delivery of clean transcripts or filtered audio output with **API deployment and real-time system monitoring**.

Analysis To Find Relationship Between Mental Trauma, Rape, and Suicide | Github-repo — Sep 2023 – Dec 2023

- **Conducted EDA for a research project** under faculty guidance for **Rajalakshmi Engineering College**, analyzing trauma, rape, and suicide data using Python to uncover key correlations.
- Preprocessed and visualized datasets into **interactive business reports** improving interpretability by **50%** for use by policymakers and mental health researchers.

Predictive Modeling for Loan and Insurance Cost Estimation | Github-Repo — Sep 2023 – Dec 2023

- **Developed ML models for Rajalakshmi Engineering College** to predict loan eligibility and insurance premiums using financial and health data.
- **Boosted accuracy by 10% (loans) and 8% (insurance)** by optimizing **SVM, Decision Trees, Random Forests, and Gradient Boosting.**
- Conducted **feature importance analysis and hypothesis testing** to simulate **A/B scenarios**, optimizing risk and pricing decisions. Applied statistical validation to guide model improvements and business strategies.