SAJID SHAIKH

(213) 756-9690 | sajidsha@usc.edu | in/connectwithsajid | connectwithsajid.github.io | Los Angeles, CA

EDUCATION

University of Southern California, Los Angeles **Aug 2024 – May 2026**
Master of Science in Computer Science, Courses: Analysis of Algorithms, LLM(RAG), NLP, Machine Learning *GPA 3.67/4*

MIT-WPU, Pune **Jul 2017 – Jul 2021**
Bachelor of Technology in Computer Science and Engineering, Courses: Data Structures, Big Data, Data Science *GPA 9.12/10*

EXPERIENCE

Kognitic Inc **Jun 2025 – Aug 2025**
Data Science Intern *New Jersey, USA*

- Automated & integrated data pipeline processes for structured clinical trials data, enhancing reliability and scalability.
- Applied fuzzy matching algorithms and similarity scoring techniques to identify and resolve potential duplicate trials, improving data accuracy and reducing redundancy.
- Designed and implemented de-duplication logic to ensure consistent and unique trial records across multiple datasets.
- Documented data pipeline architecture, including workflow logic, metadata rules, and integration points for reproducibility and future development.
- Coordinated with data quality teams to establish & enforce rules that improved consistency and compliance across datasets.
- Collaborated with engineering and domain teams to validate data quality and ensure reliable data ingestion from diverse sources.

ZS Associates **Jul 2021 – May 2024**
Business Technology Solutions Associate/ Consultant *Pune, India*

- Implemented Blue-Green deployment on Docker containers leveraging AWS CI/CD, leading 0 application downtime.
- Optimized query performance using Sort-Merge Join on large-scale dataset, resulting 20% reduction in processing time.
- Leveraged Spark connector to aggregate data generating time-based and patient behavior metrics for enhanced analysis.
- Engineered and implemented a Python Flask API for automated data ingestion from FTP to external systems, reducing manual data transfer efforts by 70%
- Engineered scalable web applications adopting Python Flask Microservices, supporting over 50 live users.
- Devised a common data model using advanced SQL and Pyspark, optimizing disk space and data retrieval speed by 30%
- Dynamically populated PPTX slides, transforming SQL query results into JSON to insert into tables using ‘tr‘ and ‘td‘ tags.
- Implemented token-based authentication using Flask wrapper, validating security via SQL/HTML injection checks.

PROJECTS

A Multimodal Assistive Tool for Visually Impaired | **Skills:** TensorFlow, Object Detection, Image Processing, Python.

- Designed AI assistive vision using, Faster R-CNN, OCR, BLIP-2 for object detection, image captioning, & auditory feedback.
- Implemented VQA, NLP-based descriptions using COCO, Visual Genome, and evaluated with mAP, IoU, CIDEr.

Sentiment Analysis on Amazon Product Reviews | **Skills:** NLTK, Scikit-learn, Feature Extraction, Supervised Learning.

- Loaded Amazon Office Products reviews dataset, created a balanced set of 200,000 reviews (100,000 positive, 100,000 negative) with binary labels, and split into 80% train (160,000) and 20% test (40,000).
- Cleaned reviews by converting to lowercase, removing HTML, non-alphabetical characters, and extra spaces; preprocessed by removing stop words and lemmatizing, reducing average length from 317 to 196 characters.
- Extracted TF-IDF features and trained Perceptron, SVM, Logistic Regression, and Naive Bayes models, achieving test accuracies of 85.7%, 89.0%, 89.2%, and 89.2% respectively.

Vertebral Column KNN Classification | **Skills:** Python, Scikit-learn, Pandas, Matplotlib, KNN, EDA.

- Built a binary classifier for Vertebral Column data using KNN with Euclidean, Manhattan, Minkowski, Chebyshev, and Mahalanobis metrics. Additionally, visualized via scatter/boxplots, and optimized k.
- Evaluated with accuracy, confusion matrix, and F1-score; achieved lowest test errors: 0.06 (Minkowski, p=0.6), 0.10 (weighted Euclidean/Manhattan)

Time Series Classification | **Skills:** NumPy, Matplotlib, Seaborn, SciPy, Time-Series Feature Extraction, Bootstrap.

- Loaded and preprocessed AReM dataset (88 instances, 6 time series each) from 7 activity folders.
- Extracted time-domain features (min, max, mean, median, std, Q1, Q3) for each time series; used bootstrap to estimate 90% CI for feature standard deviations.
- Selected mean, median, and standard deviation as key features based on small confidence intervals from bootstrap analysis for reliable classification.

TECHNICAL SKILLS

Data Analytics: MySQL, PySpark, Tableau, PostgreSQL, Databricks | **Programming Languages:** Python, Java, SQL
Certifications: AWS Fundamentals, Deep Learning, Python Specialization | **Languages:** English(Fluent), Spanish(Beginner)
Cloud Platforms: AWS (EC2, S3, CI/CD, Step Functions, Lambda, Glue, RDS, MWAA)
DevOps & Automation Tools: AWS CI/CD, Docker, Jenkins, Kubernetes, Git, JIRA, VS Code, SonarQube
Frameworks: Flask, Django, React Native, Ionic, Angular, Pytorch, CUDA
Operating Systems: Linux, Unix, Ubuntu