Hamza Riaz

[hamzakamboh68@gmail.com](mailto:hamzakamboh68@gmail.com) | +92 3030595057|

# EDUCATION

## FAST NUCES Islamabad, Pakistan

*Bachelors in Data Science Expected Graduation, june 2026*

*Current GPA: 2.67*

# PROFILE SUMMARY

Aspiring data scientist with a strong grasp of AI, machine learning, and big data ecosystems, focused on crafting practical, high- impact solutions through intelligent data utilization.

* Proficient in Python, Scikit-learn, TensorFlow, SQL, OpenCV, with hands-on experience in data mining, statistical modeling, image processing, and time-series analysis.
* Skilled in using Hadoop, Spark, Kafka, and MongoDB for scalable data processing and real-time data handling.
* Experienced in end-to-end workflows—from data cleaning and feature engineering to model training, evaluation, and deployment.

# PROJECTS

## Electric Load Forecasting Using Data Mining Techniques:

* Developed a predictive system using Python (Pandas, Scikit-Learn, XGBoost, LSTM) to forecast hourly electricity demand based on weather and time features.
* Built clustering models (K-Means, DBSCAN), implemented anomaly detection, and deployed results via a React-based interactive web dashboard.

## End-to-End Big Data Analytics Pipeline:

* Engineered a PySpark-based movie recommender leveraging Spark SQL and MLlib’s ALS algorithm to ingest, clean, and preprocess FilmTrust ratings data. Performed exploratory analysis on rating distributions and top users/movies via DataFrames and RDD transformations for scalable preprocessing. Trained and evaluated collaborative-filtering models—reporting precision/recall metrics—and documented memory- vs. model-based approach trade-offs through modular, notebook-driven code.
* Developed a Python-on-Hadoop MapReduce pipeline using HDFS and YARN to process Margie Travel’s passenger and airport CSVs for fault-tolerant, large-scale ingestion. Calculated per-airport flight counts, identified unused airports, and converted timestamps into HH:MM format via custom mapper/reducer functions. Computed line-of-sight distances (nautical miles) for each flight, aggregated total miles per passenger, and determined the top-mileage traveler in the dataset.
* Conducted a critical evaluation of PySpark (Spark SQL, DataFrames, MLlib) vs Hadoop MapReduce (HDFS, YARN) focusing on scalability, performance, and development efficiency. Analyzed resource utilization and bottlenecks during large-scale data processing, identifying performance trade-offs and optimization opportunities. Compiled detailed reflections and recommendations to guide future enterprise-scale deployments of big data solutions.

## Food Delivery Routing Optimization:

* Implemented in C++ using STL adjacency lists on an N×N grid to model restaurants/customers with weighted edges, ran multi-source Dijkstra with time-window constraints.
* Applied a greedy heuristic to assign riders and sequence deliveries minimizing travel while respecting prep/customer deadlines, and built modules to parse inputs, simulate traffic delays via updated edge weights, and output optimized schedules with cumulative distance/time metrics.

# SKILLS

**Programming & Concepts:** Python, R, C++, SQL, HTML/CSS, JavaScript, [d3.js](http://d3.js/), OOP, Data Structures and Algorithms. **Big Data & Distributed Systems:** Apache Spark, PySpark, Hadoop, Kafka, Airflow, Hive, MongoDB, PyTorch, TensorFlow, Keras, Scikit-learn, OpenCV

**Data Science & ML:** Deep Learning, Time-Series Analysis, Statistical Modeling, Image Processing, Natural Language Processing (NLP), Clustering, Feature Engineering, Model Evaluation, Association Rule Mining.

**System Design & Deployment:** Real-Time Data Processing, Scalable Model Deployment, End-to-End Pipeline Design