Mubashira Khan

Email: mubashira.eocean@gmail.com **Phone:** (+92) 3163909751

GitHub: https://github.com/Mubashirakhan03

ABOUT ME

Detail-oriented data analyst with strong skills in data warehousing, machine learning, and natural language processing, and hands-on experience with large-scale and multimodal datasets. Proficient in Python and experienced in developing predictive and classification models to extract insights from complex data. Motivated to contribute to research-driven projects through a Master's program in Data Science with a rigorous approach to data analysis and interpretation.

EDUCATION AND TRAINING

[01/05/2021 – 12/05/2025]

Bachelor in Software Engineering

University Of Sindh, Pakistan

| **Final grade:** 3.68/4.00 | **Thesis:** HealthCare Vision Assistant for Automated Food Image Analysis and Dietary Recommendation

PROJECTS

HealthCare Vision Assistant – Final Year Project

Developed a multimodal data science system to reduce the gap between generic dietary advice and condition-aware health decisions by integrating image-based analysis with user-provided medical data. The project applies computer vision, NLP, and retrieval-augmented generation to convert unstructured inputs into interpretable, data-driven health recommendations for real-world decision support.

View on GitHub

Real-Time Twitter Sentiment Analysis

Built a scalable big data solution to address the challenge of extracting real-time sentiment insights from high-volume Twitter streams. The system analyzes 5,000+ tweets per minute, integrating Kafka for data ingestion, Spark Streaming for processing, NLP-based sentiment classification, and MongoDB with a Django dashboard for visualization, demonstrating applied skills in big data analytics and real-time machine learning workflows.

View on GitHub

RAG Pipeline – Qdrant, HuggingFace & Gemini

Developed an intelligent, context-aware chatbot to reduce factual inaccuracies and context loss in large language models. The system retrieves relevant information from an indexed document base using Qdrant Vector Store and HuggingFace/Gemini embeddings, achieving a 40% reduction in query latency compared to standard LLM lookups, and generates factually grounded responses via Google Gemini.

View on Github

WORK EXPERIENCE

EOCEAN PVT LTD (KARACHI)

[09/2024 – Current] **Associate SQA Automation Engineer**

- Pioneered automated data validation for ML datasets, catching anomalies 72 hours earlier than manual methods and safeguarding dataset integrity.
- Designed a real-time QA dashboard, providing actionable insights to improve system reliability and pipeline quality.
- Integrated BDD workflows into data-intensive QA processes, enhancing ML pipeline efficiency and reducing regression testing time by 40%.

INVENTION AND INNOVATION (REMOTE)

[09/2023 – 06/2024] **Junior Data Science Engineer**

- Contributed to a large-scale e-commerce sentiment project, processing thousands of reviews per minute using Kafka, Spark Streaming, and NLP for real-time insights.
- Implemented data validation and governance protocols, reducing post-deployment errors by 30% by ensuring clean ML datasets.
- Automated ML model deployments (MLOps), reducing time-to-production by 25% and enhancing real-time pipeline reliability.

PUBLICATIONS

Effectiveness of Transposition on the Throughput in Internet of Medical Things Routing Protocol

Journal Name: City University Research Journal (CUR), HEC Recognized Journal

SKILLS

Machine Learning & AI

Predictive Modeling | ML Pipe line Automation | Sentiment Analysis | NLP & LLM | Multimodal Data Processing | Computer Vision (Image Analysis)

Big Data & Distributed Systems

Apache Spark Streaming | Distributed Processing | Data Ingestion

Programming & Tools

Python | MATLAB | Docker | Streamlit

CERTIFICATIONS

Comprehensive Data Science & AI Training - Udemy

Deep Learning - Simplilearn | SkillUp

Machine Learning with Python - freeCodeCamp

Generative AI for Beginners – Simplilearn | SkillUp

HONOURS AND AWARDS

[12/2024] **Certificate Of Best Performance Awarding institution:** ORIC CUST & Knowledge Streams

Recognized for strong applied understanding of Deep Learning and Machine Learning in data science systems.

[12/2024] **Best Project Award (SIC AI COURSE) Awarding institution:** ORIC CUST & Knowledge Streams

Awarded for designing a Retrieval-Augmented Generation (RAG) system using LLMs and vector databases (Qdrant).