SUPREET SHARMA

Website | Github | Linkedin | supreet.shm523@gmail.com

PROFILE

- I completed my **master's in data science** in November 2024, specializing in **Machine Learning**, with a focus on **Graph Learning** and a strong interest in **Data Engineering** and **Generative AI.**
- I bring over **4.5 years** of experience spanning both **Data Science and Software Engineering**.
- I also hold a **bachelor's degree in computer science** and have **3.5+ years** of experience in Java backend development, working in domains ranging from Payments, Insurance, Life Sciences and Product Lifecycle Management (PLM).

SKILLS AND COMPETENCES

Data Engineering: Data acquisition (Web Scraping, API, Documents) & transformation **|** ETL Pipelines **|** Big Data (Spark, Distributed DBs)

Programming: Python (Pyspark, Numpy, Pandas, TensorFlow, PyTorch, scikit-learn, Langchain) **|** R (shiny, Tidyverse, dplyr) **|** Java

Database: Vector DB - Elasticsearch, Pinecone**|** Relational DB - MySQL, PostgresSQL, SQL Server **|**NoSQL - MongoDB, Neo4J, Cassandra

Machine Learning: Deep Learning (DL) **|** Graph Learning, Natural Language Processing (NLP) **|** MLOps practices (ML pipelines, Continuous delivery and monitoring of ML models)

Generative AI: Integration and fine-tuning of Large Language Models (LLMs) **|** LLMs & Retrieval-Augmented Generation (RAG) **|** Prompt Engineering **|** Hugging Face **|** Llama, Mistral

Dashboard design and visualization: Tableau, Power BI, Grafana

Deployment & Automation tools:

Cloud - AWS services (Redshift, EC2, S3, SageMaker, Lambda, IAM, EventBridge, Cloudwatch, EMR, API Gateway, Kinesis) **|** **Workflow Management** - Apache Airflow, Dagster (Orchestration), Apache Spark, Kafka, ZenML, MLFlow **| Version Control Systems -** Github Actions, CI/CD (GitLab) **| IaaC -** Docker, Kubernetes, Terraform

WORK EXPERIENCE

December 2023 – July 2024 **Generative AI Developer (Student Assistant), E. ON Energy Research Center, Aachen, Germany**

- Integrated attack graph generation using **LLMs** & **Retrieval-Augmented Generation (RAG)** in a Cyber Threat Detection pipeline for energy grids, leveraging **vector embeddings** & **Prompt Engineering**.
- Developed a Python-based Cyber Threat Detection system from **ground up**, utilizing **Neo4j** for network data and implementing **DevOps** practices such as **Docker** and **GitLab CI/CD** to automate deployment.

April 2023 – October 2023 **Data Engineer, Munich RE (Corporate Underwriting), Munich, Germany**

- Built expertise in Insurance and Actuarial Science by developing ETL data pipelines to process Cyber Insurance data, enabling comprehensive risk analysis for financial modeling and pricing analysis.
- **Evaluated, Processed, Cleaned & Transform** large unstructured Data sets to elevate **accuracy** and **quality**.
- Integrated a new database schema and requirements into the existing pipeline and refactored code, ensuring compatibility with **1000+** existing ETL scripts, and streamlined data processing, resulting in a **20-30% speed improvement**.
- Collaborated with underwriters to develop Power BI dashboards that provided detailed analysis on claims and exposure data, supporting insurance portfolio analysis for precise pricing and tariff calculations.
- Integrated **Python into R** pipeline to enhance performance and **Cross-Language functionality**.
- Implemented enhancements for Software Engineering practices, including **testing**, **version control, dependency management** & **CI/CD** processes, to ensure higher **code quality** & smoother deployments.

October 2022 – June 2023 **Student Assistant, RWTH Aachen (Lehrstuhl für Prozessleittechnik), Aachen, Germany**

- Maintained **Web Service APIs** using **Flask** and wrote unit tests for BaSyx Python SDK.
- Optimized and refactored deserialization code, **reducing the codebase by 80%** & increasing efficiency.
- Designed and Implemented a Proof of Concept for **predictive analysis** by leveraging **vector embeddings** with **Graph Neural Networks** for AAS objects.

February 2018 – March 2020 **Research and Development Engineer, Dassault Systèmes, Pune, India**

- Developed and maintained various components of **ENOVIA**, Dassault's **PLM software.**
- Optimized performance by implementing **parallel streams**, reducing query execution time by **60%**.
- Transformed application services into **standalone REST APIs** with authentication and error handling.
- Supported migration from in-house source control to **GIT** and **mentored** team members.

May 2016 - February 2018 **Assistant System Engineer, Tata Consultancy Services (Citi-Bank Client), Pune, India**

- Migrated legacy system to **Java/GWT-based** platform for Worldlink **B2B Cross-border Payment System**.
- Optimized database performance, **reducing workflow runtime** from **Minutes to Seconds**.
- Developed Microservices with encryption for secure payment data exchange between isolated units.

EDUCATION

2020 - 2024 **Master's in data science, RWTH Aachen University, Aachen, Germany**
Lab: Graph Learning
Seminar: Foundations of Supervised Machine Learning with Graphs
Thesis: **Learning Heuristics for Counting Problems with Graph Neural Networks**

2012 - 2016 **Bachelor's in computer science, SRM University, Chennai, India**

PROJECTS

For a detailed overview of all my projects, please explore my portfolio: https://supreetshm947.github.io/portfolio/

★ **GenAIStackOverflow, 2024,** GitHub Link
- Implemented a data pipeline that retrieves relevant, answered StackOverflow posts via its API, generates vector embeddings, and constructs Knowledge Graphs in a Neo4j database to enhance information context.
- The RAG system retrieves relevant posts and combines them with user queries for LLM-driven responses using Gemini.
- The frontend is built with Streamlit, and Docker with GitLab CI/CD ensures streamlined deployment and scalability.

★ **End-to-End MLOps Pipeline with Monitoring using MLflow, Airflow, and Kubernetes, 2024** Github Link
- Implemented a Machine Learning pipeline orchestrated using Apache Airflow that automates model training, experiment tracking using MLflow, model artifact storage in MinIO, and deployment of the trained model as a FastAPI service.
- Additionally, it provides CI/CD automation for deploying Docker images to Kubernetes, and monitoring with Prometheus, InfluxDB, and Grafana.

★ **Learning Heuristics for Counting Problems with Graph Neural Networks - Master's Thesis, August 2024** Github Link
- Built on ANYCSP for Constraint Satisfaction Problems, extending it to counting problems in Graph Coloring and Boolean Satisfiability, efficiently sampling solutions for variable sizes up to 200, where sharpSAT fails to perform.
- Introduced a novel Constraint Value Graph processed through a Recurrent GNN Model.
- Reinforcement learning to capture complex patterns through training on synthetically generated, diverse data examples.

★ **Machine Learning (MLOps) Workflow and Deployment using ZENML for an Image Classification Model, 2023** Github Link
- Created a holistic MLOps pipeline with ZenML for data ingestion, training, evaluation, and deployment of Image classification model.

★ **Crypto Data Pipeline (ETL), September 2024 *(Ongoing)** Github Link
- Developed a holistic pipeline for data acquisition of real-time cryptocurrency price data and Reddit posts on crypto.
- Processed and stored all data in a Datalake using a MinIO container, utilizing Parquet and Delta formats.
- Developed Airflow DAGs to orchestrate regular data updates and used Spark to persist data in the DataLake.
- Developed an inference pipeline incorporating sentiment analysis with BERT and using sentiment scores and historical coin prices to predict future coin prices.

★ **Semantic Analyzer, 2024** Github Link
- Implemented a Natural Language Processing (NLP) model in Pytorch for sentiment analysis.
- The model is trained and evaluated on the IMDB movie review which contains binary labels for movie reviews.

★ **AI Snake Game, 2022** Github Link
- Developed a self-playing Snake game using Reinforcement Learning and the Bellman Equation for model updates.

SOFT SKILLS

- Strong **analytical skills** and the ability to **effectively communicate** technical concepts to diverse audiences.
- Experience in **technical consulting**, providing expertise and guidance to support project goals.
- Ability to **collaborate** with internal teams and external customers to achieve shared objectives.

LANGUAGE

English (Business Proficiency - C1)
German (Basic - A1)
Hindi, Punjabi (Mother Tongue)

INTERESTS

Cooking, Baking, Coding, Running, Strength Training, Movies & listening to podcasts