**Gokul S**  
8124303097 • gokuls13102001@gmail.com • www.linkedin.com/in/gokul-s-0b3729207 • https://github.com/gokuls13

**Professional Summary**

Data Science and Machine Learning enthusiast with 3 years of hands-on experience across EDA, data preprocessing, statistical analysis, and deploying ML/DL models. Proficient in Python, Pandas, NumPy, Scikit-learn, Pytorch, Keras, PySpark, SQL, and visualization tools like Tableau, Matplotlib, and Seaborn.

I’ve worked on supervised and unsupervised learning, optimization algorithms, graph algorithms and evaluation metrics . Skilled in deep learning (Neural Networks), NLP, and LLM/RAG using Hugging Face, Groq, LangChain, Chroma, and Generative AI (GenAI), including model fine-tuning.

Experienced with model deployment using Streamlit, Flask, AWS, Azure, Docker, and version control (GitHub, GitLab). Also worked on TigerGraph (GSQL), Spark MLlib, Hadoop, and Control-M. Successfully built and deployed a chatbot and applied end-to-end data science workflows in real-world projects.

**Technical Skills**

**Programming & Libraries:** Python, Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn, PyTorch, Keras (TensorFlow), Java  
**Statistics & ML Concepts:** Descriptive Statistics, Inferential Statistics ( Probability, Distributions, Hypothesis Testing), EDA, Data Preprocessing  
**Machine Learning:** Supervised (Linear Regression, KNN, Decision Trees, Random Forest, AdaBoost), Unsupervised (K-Means, Hierarchical Clustering)  
**Optimization:** Gradient Descent, Batch GD, SGD, Mini-Batch GD, ADAM  
**Deep Learning:** Vanilla NN, CNN, RNN (LSTM, GRU), GAN, SARIMA  
**NLP & GenAI:** Transformers, LLM, RAG, Hugging Face, Groq, LangChain, Chroma  
**Deployment:** Streamlit, Flask, Docker, AWS, Azure  
**Big Data & Graph:** PySpark, Spark MLlib, Hadoop, TigerGraph (GSQL), Graph Algorithms (WCC, APSP), PostgreSQL  
**Visualization Tools:** Tableau  
**Version Control:** GitHub, GitLab  
**Workflow Orchestration / Scheduling Tools:** Control-M

### ****Professional Experience****

**Mastech Infotrellis** – Full-time Paid Internship  
**Jan 2022 – Sep 2022**

* Contributed as a shadow resource on real-time Data Science projects, supporting data preparation, preprocessing, and ML pipeline development.
* Gained exposure to production-grade data workflows, Big Data tools, and enterprise-level analytics.

**Mastech Infotrellis** – Data Science Associate  
**Oct 2022 – Present**

* Worked on multiple machine learning, deep learning, and GenAI tasks addressing real-world business challenges.
* Involved in full ML lifecycle — from data preprocessing and modeling to deployment on AWS/Azure using Streamlit and Flask.
* Hands-on experience with Big Data tools (PySpark, Hadoop), graph technologies (TigerGraph), workflow tools (Control-M), and MLOps practices.
* Contributed to automation, internal tooling, and GenAI-based chatbot development for enterprise productivity.

**Projects**

1. **Internal project**

**MIT NOVA (Chatbot)**  
 Tech Stack: LLaMA3 (Groq), LangChain, SentenceTransformers, ChromaDB, Flask, Streamlit, Docker, AWS

* Built an internal enterprise chatbot capable of answering employee queries related to company policies.
* Used **Retrieval-Augmented Generation (RAG)** with ChromaDB and sentence-transformers for document retrieval and context generation.
* Developed the frontend using **Streamlit**, backend using **Flask**, containerized with **Docker**, and deployed on **AWS**.
* Improved onboarding and internal support efficiency, and demonstrated the potential of GenAI within the organization.

1. **Dell Project**

Dell Inc. is an American technology company that develops, sells, repairs, and supports personal computers (PCs), servers, data storage devices, network switches, software, computer peripherals including printers and webcams among other products and services. In this project, we handled a large-scale customer identity data platform that receives millions of records weekly. Our team was responsible for **processing, maintaining, and updating customer records** to ensure high data quality and availability for downstream business functions. In parallel, we applied **data science techniques** to identify **insights and recommend improvements** in the data resolution and enrichment processes — helping Dell optimize how customer identities are managed across platforms. I worked in datascience team.

**1. Probabilistic Matching Engine (PME)**  
Tools: Python, TF-IDF, n-grams, KNN, Scikit-learn, postgreSQL

* Designed and developed a **Probabilistic Matching Engine** to enhance resolution of customer identity records which couldn’t be fully resolved by TigerGraph.
* Applied advanced data preprocessing using **n-grams** and **TF-IDF vectorization** to represent input and reference datasets.
* Built a **K-Nearest Neighbors model**, tuned for best K and threshold values, to match unresolved inputs with reference records.
* This solution significantly increased data resolution accuracy, enabling better customer identity tracking and decision-making — and received strong appreciation from the business for its impact.

**2. Graph-Based Identity Resolution & Translation**  
Tech Stack: TigerGraph, GSQL, Graph algorithms

* Implemented **graph algorithms** like **All Pairs Shortest Path** and **Single Source Shortest Path** to solve complex identity resolution and translation use cases.
* Optimized the implementation to execute in under **6 seconds** on billions of records, enabling real-time data transformation and retrieval.
* Helped the client instantly access cookie-to-user mappings, improving targeting and personalization outcomes..

**3. GitLab Repository Sync Automation**  
Tech Stack: Python, GitLab API

* Automated one-way synchronization of GitLab repositories to replicate production business repositories for backup and deployment.
* Streamlined version control practices and reduced manual intervention in production syncs.

1. **Labcorp Project :**

**Labcorp Holdings Inc., operating under the brand name Labcorp, headquartered in Burlington, North Carolina, provides laboratory services used for diagnosis and healthcare decisions.**

**Legacy Data Cleaning for Labcorp**  
Tech Stack: SQL, Python

* Worked on legacy structured data tables to ensure only accurate and high-quality records are retained for downstream usage.
* Focused on **data cleaning and preprocessing**, removing duplicates, outdated entries, and conflicting records to ensure consistency.
* Helped improve **data reliability** and reduced data-related issues in analytical and operational systems, contributing to better decision-making and compliance with data governance policies.

**Education**

**MCA CS&IT**, Jain (Deemed-to-be University)  
*Jan 2023 – Dec 2024*

Grade: 8.9 , Excellent

**BSC Computer Technology**, PSG College of Arts & Science   
*Jul 2019 – Jun 2022*

Grade: 8.4 , Excellent

**Certifications:**

* Microsoft Certified: Azure Fundamentals – Microsoft
* Databricks Lakehouse Fundamentals - Databricks
* Graph Algorithms for Machine Learning Certification – Tigergraph
* TigerGraph Associate Certificate – Tigergraph
* MTA: Introduction to Programming Using Java – Microsoft
* Gitlab 101 & 201 - Gitlab
* Gitlab Solutions Architect Core Verified Associate Certificate – Gitalb
* Getting Started with Cloud Computing – Udemy
* Get started with Microsoft Fabric - Microsoft
* Pyspark Internal Certification– Mastech Infotrellis
* Azure data lake and azure data factory Internal Certification – Mastech Infotrellis
* Kafka Fundamentals Internal Certification – Mastech Infotrellis
* DataScience Internal Certification – Mastech Infotrellis
* Data Modelling for warehouses Internal certification – Mastech Infotrellis
* NPTEL Certified Soft Skill Development – NPTEL

### ****Awards & Recognition****

* **Dynamic Team Player Award**, Mastech Infotrellis – H2 2023  
  Recognized for consistently demonstrating collaboration, initiative, and high-quality contribution across data science projects.
* **Consistent Client Appreciation**, Dell Technologies – 2024,2025  
  Received back-to-back positive feedback for impactful contributions, ownership, and solution delivery.