Nishith Kumar Gupta

+91-8576934666 nishith9900@gmail.com Linkedin Github

Education

Birla Institute of Technology and Science, Pilani — **Hyderabad, Telangana**
B.E in Mechanical Engineering, Minor in Computing and Intelligence — *November 2020 - July 2024*

Academic Heights Public School — **Gorakhpur, UP**
CBSE, Percentage - 90.6 — *April 2018 - August 2020*

Relevant Coursework

- Data Structures
- Machine Learning
- Operating System
- Data Mining
- OOP
- Deep Learning
- Database Management
- NLP

Experience

Data Analyst – Integration & Data Management — **Jun 2024 - Feb 2025**
Veritas Prime — *Hyderabad, Telangana*

- Engineered and optimized end-to-end data integration solutions for **SAP** SuccessFactors Employee Central, leveraging **Python**, **Java**, **Groovy**, and **RESTful APIs** to automate global payroll data processing and built **10** + scalable interfaces
- Designed robust data pipelines using **SQL** and **ETL** methodologies for complex transformation of enterprise-scale datasets
- Integrated **REST** and **SOAP** APIs to synchronize HRIS platforms with external systems, reducing manual interventions
- Deployed integration solutions in cloud environments, managing version control with **Git**, **CI/CD** pipelines, **Jira**, and automated testing frameworks to ensure minimal downtime and seamless production releases

Machine Learning Intern — **Jul 2023 - Dec 2023**
Distributed Energy — *Pune, Maharashtra*

- Developed an end-to-end **Machine Learning** model to forecast Solar Tariff/PPA prices, improving pricing accuracy by **15%**
- Fine-tuned ML algorithms **(Random Forest, XGBoost, and Neural Networks)**, reducing **MAE** by **12%** from the baseline
- Built an automated **ETL** pipeline leveraging Pandas, NumPy, and **SQL** for data ingestion, preprocessing, and transformation
- Deployed the model using **Docker** and **GCP/AWS**, integrating a **Flask API** with an interactive web-based dashboard
- Major Tech Stack Used: Python, NumPy, Pandas, Flask, SQL, Docker, AWS/GCP

Projects

Calorie Burn Prediction | *ML Supervised Regression* **Kaggle** — **May 2025**

- Built a **supervised ML** pipeline using **ensemble techniques** (XGBoost, LightGBM, CatBoost) to predict calories burned from biometric data, securing a top **20% rank** out of **4300+** submissions in a Kaggle competition
- Engineered features (BMI, MET score, etc.) and optimized models with **K-Fold** cross-validation, achieving **0.058 RMSLE**
- Applied **SHAP** for model interpretability and feature importance analysis; leveraged **Optuna** for hyperparameter tuning to enhance model performance

LLM ChatBot with RAG | *GenAI* **GitHub** — **March 2025 - Apr 2025**

- Developed an **LLM**-powered **Retrieval-Augmented Generation (RAG)** chatbot using **LangChain** and **OpenAI**, enabling PDF-based question answering by retrieving relevant document chunks before generating responses
- Built a retrieval pipeline with text extraction, chunking, OpenAI embeddings, **FAISS-based** vector storage, and similarity search
- Designed a modular pipeline with PDF processing, vector store creation, and QA chain logic integrated into a **Streamlit** interface

Chest X-ray Image Classification | *Deep Learning, MLOps* **GitHub** — **Dec 2024 - Jan 2025**

- Orchestrated a **VGG-16 CNN**-based image classification using **TensorFlow/Keras** for detecting chest cancer from medical scans
- Modularized data ingestion, transformation, training, and evaluation processes with DVC and YAML-based config files
- Logged experiments using **MLflow** integrated with **DAGsHub** for seamless collaboration and model versioning
- Containerized the app with **Docker**, automated **CI/CD** with GitHub Actions, and deployed on **AWS EC2** using **ECR**

Hospital Readmission Risk Classifier | *Python, LightGBM, Scikit-learn* **GitHub** — **Oct 2024 - Dec 2024**

- Built a machine learning model to predict hospital readmissions, addressing a critical healthcare challenge that costs **U.S.** hospitals over **$41 billion** annually due to frequent readmissions within **30 days** of discharge
- Tested and optimized multiple machine learning algorithms, including **LightGBM (LGBM)**, **XGBoost**, and **Random Forest**, leveraging hyperparameter tuning and ensemble methods to enhance accuracy, with **LGBM** achieving the highest precision
- Performed **EDA**, uncovering critical insights into patient demographics & comorbidities to reduce readmissions & legal penalties

Technical Skills

Languages: Python, Java, C/C++, SQL, Groovy, HTML/CSS
Developer Tools: AWS, Docker, Snowflake, Tableau, Power BI, GitHub, VS Code, MongoDB, MS Excel
Frameworks: TensorFlow, Apache Spark (PySpark), Flask, Django
Libraries: Pandas, NumPy, NLTK, Matplotlib, ScikitLearn
Statistical Techniques: Hypothesis Testing, A/B Testing, Regression Analysis, Root Cause Analysis

Certifications

Machine Learning Specialization | *Coursera*
• Supervised • Unsupervised • Reinforcement Learning