Harshaditya Kumar Mallipudi

📞 585-825-0412 ✉ hm3799@g.rit.edu 🔗 LinkedIn

EDUCATION

Rochester Institute of Technology — **Rochester, NY**
Master of Science in Data Science — *Aug 2022 – May 2025*
Courses: ML,AI, Data Mining Analytics, Visual Analytics, Mathematical and Statistical Fndn. for DS, Databases - Design Implementation, Big Data Tools.

SRM Institute of Science and Technology — **Chennai, India**
Bachelor of Engineering, Computer Science and Engineering, — *Aug 2018 – May 2022*

RELEVANT EXPERIENCE

Venture Creations Incubator — **Rochester, NY**
Data Engineer/Scientist, Intern — *Feb 2023 – Present.*
- **Collaborated weekly** with **C-suite executives** to identify operational pain points, translating insights into actionable data-driven solutions, resulting in a 25% improvement in project delivery timelines.
- Implemented cloud infrastructure onboarding using Ansible, CloudFormation, and AWS Lambda, ensuring 100% seamless project deployment on EKS with Load Balancers.
- Designed an ML pipeline with Docker, Kubernetes, and Terraform (AWS CI/CD via Jenkins), optimizing model deployment and scalability.
- Developed a real-time data pipeline with BeautifulSoup, Selenium, and AWS ECS (Fargate, Kinesis, S3), improving data ingestion by 40% and lead qualification by 35%.
- Transformed 25+ Tableau dashboards, enhancing SQL-powered athlete performance tracking by 15%.

Verzeo — **Bangalore, India**
Data Analyst, Intern — *Aug 2020 - Oct 2020*
- Conceptualized designed interactive dashboards in R using Shiny and ggplot2, enabling sales teams to monitor KPIs and make data-driven decisions.
- Refined data preprocessing tasks using Pandas and Dask, including outlier detection, missing value imputation, and feature scaling, resulting in cleaner datasets for downstream analysis.

PROJECTS

Visual Analytics for Crime Investigation – (Python, NetworkX, Gephi, Tableau, Apache Airflow)
- Examined 1M+ email records from the 2014 VAST Challenge, leveraging NetworkX for anomaly detection and Gephi for interactive network analysis.
- Established an ETL pipeline with Apache Airflow to automate data extraction and developed Tableau dashboards visualizing hidden communication patterns.

India vs Pakistan Cricket Sentiment Analysis – (Python, NLP, Hugging Face, AWS Lambda, Tableau)
- Extracted and processed 200K+ YouTube comments using AWS Lambda, classifying sentiment with Hugging Face transformers (78% accuracy).
- Assembled Tableau dashboards to visualize fan sentiment across match highlights, identifying trends in excitement, controversy, and national pride.

Amazon Sentiment Analysis – (Python, NLP, TensorFlow, Apache Spark, Azure Data Factory, Power BI)
- Architected a real-time sentiment analysis pipeline for 150K+ Amazon reviews, integrating Azure Data Factory for data ingestion and Apache Spark for distributed processing.
- Trained a TensorFlow model (85% accuracy) for sentiment classification and developed Power BI dashboards tracking customer sentiment trends.

Health Tales: Predicting Health Risks – (Python, Scikit-learn, AWS S3, Glue, Athena, QuickSight)
- Processed 991K+ health records using AWS Glue & Athena, predicting diabetes and cardiovascular risks via Random Forest (72.1% accuracy).
- Launched an AWS-based ETL pipeline, enabling real-time health analytics in QuickSight, identifying key risk factors.

Predicting Tech Sales Success (NPS) – (R, Tidyverse, Snowflake, Random Forest, Power BI)
- Structured a Snowflake-based data warehouse, processing 5K+ tech sales rep records, predicted high Net Promoter Scores (NPS ≥ 9) with 75% accuracy.
- Enhanced hyperparameter tuning on Random Forest, identifying certifications, feedback, and salary as key NPS drivers, visualized in Power BI dashboards.

House Price Analysis in Ames, Iowa – (Python, SciPy, Apache Spark, Regression Analysis, SAS, JMP)
- Assessed 1,460 house sales from the Ames Housing Dataset, leveraging Apache Spark for scalable processing and SciPy for regression modeling.
- Validated findings using SAS & JMP, revealing homes with central air sell for 12% higher, assisting real estate pricing strategies.

SKILLS

Languages	Python, PySpark, SQL, NoSQL, Bash, R, C++, JavaScript
Technologies & Tools	PyTorch, scikit-learn, Linux, Tableau, Excel, Git, Jenkins, Ansible, Docker, Kubernetes, VSCode, SAP S4 HANA, SAP Master Data Management, Advanced Excel, Tableau.
Cloud & Databases	AWS (EC2, RDS, S3, Lambda, CloudFormation, CloudWatch, Kinesis), Kafka, OpenSearch
DevOps	Terraform, Kubernetes, CI/CD Pipelines (GitHub Actions, Jenkins, Ansible), Infrastructure as Code (IaC), Monitoring & Logging (CloudWatch, ELK Stack, Prometheus, Grafana)
Core Competencies	Statistical Data Analysis, Data Visualization, Machine Learning, Natural Language Processing, Deep Learning, Computer Vision, Supply and Demand Planning, Logistics and Warehouse Management
Certifications	**AWS CERTIFIED AI PRACTITIONER**