SAI SINDHURA KOLLEPARA

Bloomington, IN | sindhurakollepara29@gmail.com | +1 930-333-4191 | LinkedIn | GitHub

EDUCATION

Master's in Data Science, Indiana University Bloomington **Aug 2023 - May 2025**
Coursework: Data Visualization, Advanced Database Concepts, Applied ML, Data Mining, Cloud Computing-AWS, Statistics
B. Tech. in Electronics and Communication Engineering, G.V.P College of Engineering **Aug 2017 - Jul 2021**

SKILLS

Programming Languages & Tools: Python, R, JavaScript, SQL, C#, Power BI, Tableau, Flask, Streamlit
Database Technologies: MySQL, PostgreSQL, MongoDB, NoSQL
Data Engineering & Modeling: ETL/ELT, Apache Spark, Hadoop, Data Warehousing, Data Lake Architecture, Databricks, Data Modeling
Machine Learning & AI: NLP, Time Series Analysis, Transformers (BERT, MiniLM), Deep Learning, Sentiment Analysis, Feature Engineering, A/B Testing
Data Science Libraries: Numpy, pandas, scikit-learn, Matplotlib, Seaborn, Plotly, TensorFlow, PySpark
Cloud & DevOps: AWS, Docker, Git, GitHub, CI/CD, Agile (Scrum)
Certifications: AWS Certified Cloud Practitioner, Microsoft Certified: Azure Fundamentals, Microsoft Certified: Power BI Data Analyst Associate (PL-300)

WORK EXPERIENCE

Graduate Research Assistant - DSAIL, Kelley School of Business, Indiana University **Jan 2025 - Present**

- Designed data pipelines to clean and ingest 89 GB of diverse sensor data, ensuring 98% data integrity for mental health analysis.
- Created 3 interactive Python Plotly Dash dashboards to visualize key behavioral patterns linked to depression risk.
- Applied ARIMA, LSTM, and Prophet time-series forecasting models to forecast behavioral changes, improving early intervention capabilities by 25%, with 85% accuracy in depression risk prediction.

Project Engineer - Wipro Limited, India **Dec 2021 - Aug 2023**

- Streamlined Oracle SQL and Python ETL pipelines for Mastercard's Bill Pay Application, automating processes and improving query efficiency by 40% while ensuring data consistency.
- Automated CI/CD pipelines with Jenkins, automating the deployment of data pipelines and reducing deployment failures by 30%, ensuring faster and more reliable releases.
- Containerized data workflows with Docker, boosting pipeline reliability by 35% and enabling efficient parallel processing for high-volume data operations.
- Built Splunk BI dashboards for real-time system performance insights, improving decision-making by 25%.
- Created Oracle SQL, Python and Splunk reports and visuals on transactional data, aiding strategic planning.
- Performed ad-hoc analysis on large-scale transactional data to identify trends, uncover business insights, and answer critical business questions, improving data-driven decision-making.

ACADEMIC PROJECTS

Per-Capita Greenhouse Gas Emissions Analysis

- Processed 42K records from 200 countries using ELT in BigQuery, enhancing data cleaning and aggregation efficiency.
- Developed automated Python data pipelines for emissions analysis, achieving 96% accuracy in predictive modeling.
- Crafted interactive Looker Studio visuals and animated choropleth maps on GCP with Plotly, delivering real-time insights and visual storytelling.

Stack Overflow Users Survey Analysis

- Developed an interactive Power BI dashboard to analyze 100K+ Stack Overflow users' demographics, professional background, engagement trends, and technology preferences.
- Delivered 10+ interactive visualizations to analyze employment status, developer roles, and technology adoption.
- Optimized DAX measures and drill-through features, reducing query response time by 40% for interactive filtering.

Reddit Comments to Post Relevance Analysis

- Built a hybrid NLP model combining LDA and transformers (BERT, MiniLM) to classify 500K+ Reddit comments into 5 relevance levels, improving accuracy by 20%.
- Designed a custom relevance metric integrating semantic similarity, cluster coherence, sentiment, and engagement, boosting scoring consistency by 30%.
- Optimized K-Means clustering to reduce manual moderation by 50% and accelerate relevant comment retrieval by 5x.