AMRITA NEOGI

(520) 427 - 1767 | neogiamrita111@gmail.com | linkedin.com/in/amritaneogi | https://github.com/AmritaNeogi

SUMMARY

Data Scientist with 6+ years' experience in healthcare and finance, transforming Medicaid and multi-site EHR (Electronic Health Record) data into actionable insights while safeguarding PII in high-volume financial pipelines. Recognized with the 'Contextual Master Award' at TCS for secure, large-scale data solutions that improved data quality, compliance, and decision-making.

WORK EXPERIENCE

Data Science Manager, ARID Lab, University of Arizona, Tucson, AZ **Feb 2024 - Present**

- Develop causal-inference and ML models — logistic and clustering on binary features — on multi-site post-COVID data, delivering Tableau and Power BI dashboards and achieving 83% accuracy in survival and utilization prediction.
- Build Python + LLM (Large Language Model) pipelines to detect anomalies in clinical text, improving extraction accuracy by 25% and enhancing data quality for downstream reporting and analysis.
- Build reproducible ETL pipelines in R/SQL/SAS to integrate patient and Medicaid/CHIP data from 10+ sites (~500k records) into the OMOP Common Data Model; clean, deduplicate, and geocode (lat/long, Census tract, FIPS), cutting prep time by 60%.

Graduate Research Assistant, Department of Pediatrics, University of Arizona, Tucson, AZ **Nov 2022 - Dec 2023**

- Automated REDCap–Amazon MTurk integration for congenital heart defect surveillance, boosting data completeness by 25% and improving reporting efficiency.
- Optimized PostgreSQL for multi-site Medicaid data, reducing runtimes ~80% while maintaining secure access.
- Standardized death-certificate records to population-level datasets for anomaly detection and time-trend analysis.

Software Development Engineer, Tata Consultancy Services, Mumbai, India **Mar 2018 – Jul 2022**

- Re-engineered PCI-compliant ETL workflows from Informatica PowerCenter to PySpark, managing a 12-member team and ensuring secure PII handling with Protegrity tokenization while processing 3M+ daily transactions across enterprise data warehouses (EDW).
- Built Python/SQL pipelines to prepare and analyze 10M+ financial transaction records, deploying fraud detection models that reduced false positives by 20% and supported multi-year trend analysis for customer insights.
- Automated CI/CD workflows for ETL and model deployment across mainframe and EDW systems, improving release cadence by 25% and ensuring accurate, repeatable production updates.

PROJECTS

Global Regulatory Insights & Risk Prioritization **Jul 2025**

- Designed Airflow–BigQuery pipelines to automate GDPR/CCPA risk forecasting, integrating fine-tuned T5 and LSTM/Prophet models for scoring, trend analysis, and reporting.
- Built Tableau and Power BI dashboards to track risk signals, analyze patterns, and present findings for faster decision-making.
- Reduced monitoring time by 75% and delivered forecasts with <15% error, flagging top 5 daily risks with 90%+ accuracy, strengthening regulatory compliance

House Pricing Profiler using Snowflake Database **Oct 2023**

- Designed end-to-end data pipeline integrating 60k+ housing records scraped via Bright Data into Snowflake, improving data accessibility and query response times by 40%.
- Applied geocoding and automated translation workflows (Python + APIs) to standardize addresses and convert Polish-to-English text, reducing manual processing by 80%.
- Built Snowflake queries and analytical models to profile regional house pricing trends with 95% accuracy, delivering actionable insights that supported pricing strategy and market comparisons.

RESEARCH

Insurance at Birth & Infant Outcomes, ARID Lab, University of Arizona **Sep 2025**

- Built causal inference and logistic regression models on multi-site EHR data to assess payer-type effects on infant survival.
- Quantified payer disparities in infant survival, with uninsured/self-pay infants at highest mortality risk, Medicaid showing a modest ~10% survival gain, and private coverage the strongest protection at ~70%, highlighting persistent coverage gaps.
- Built reproducible modeling pipelines for subgroup analyses across demographics, improving reporting clarity and supporting equity-focused comparisons.

Healthcare Utilization & Guideline Adherence, ARID Lab, University of Arizona **Aug 2025**

- Applied ML/statistical models (logistic regression, random forests, survival analysis) to 50k+ patient records to evaluate compliance with national care guidelines.
- Identified 3 distinct utilization patterns and the top 5 predictors of continuous care, generating insights to improve outcomes and resource allocation.
- Analyzed 10-year trajectories across pediatric/adult cohorts; delivered equity-focused insights and scalable sequence models.

EDUCATION

MS in Data Science, University of Arizona, Tucson, Arizona **Aug 2022 – Dec 2023**

Coursework: Advanced ML & Deep Learning, Data Visualization, Generative AI, Computer Vision GPA: 4 / 4

BTech in Electrical Engineering, University of Engineering and Management, India **Aug 2013 – May 2017**

Coursework: DBMS, Signal Processing, Optimization & Control Systems, Linear Algebra GPA: 7.66 / 10

SKILLS

- **Programming & Data:** Python, R, SQL (PostgreSQL, Teradata), SAS, SPSS, Excel, ArcGIS (Geocoding)
- **ML & Analytics:** Classification, Regression, Clustering, Cohort Definition & Metrics, Time-Series & Trend Analysis, Causal Inference, Deep Learning (CNN/RNN/LSTM), Anomaly Detection, Summary Statistics, Data Quality & Validation
- **Visualization & Reporting:** Tableau, Power BI, Looker, Clear Reports and Presentations
- **Data Management:** Snowflake, GCP (BigQuery), ETL/ELT (Airflow), HIPAA/PII/PHI, Medicaid/CHIP Data, Git/GitHub
- **Collaboration & Leadership:** Stakeholder Communication, Cross-Functional Collaboration, Mentorship and Code Reviews