

AMRITA NEOGI

 (520) 427 - 1767

 amritaneogi@arizona.edu

 [linkedin.com/in/amritaneogi](https://www.linkedin.com/in/amritaneogi)

 [amritaneogi.github.io](https://github.com/amritaneogi)

 github.com/AmritaNeogi

 AZ-85719

EDUCATION

Master of Science in Data Science

University of Arizona, Tucson, Arizona

Aug 2022 – Dec 2023

GPA: 4 / 4

Bachelor of Technology in Electrical Engineering

University of Engineering and Management, Jaipur, India

Aug 2013 – May 2017

GPA: 7.66 / 10 (equivalent to magna cum laude)

SKILLS

- **Programming Languages:** Python, R, SQL/NoSQL, Core Java, HTML/CSS
- **Tools & Platforms:** Tableau, Looker, RStudio, Jupyter Notebook, MySQL, PostgreSQL, Teradata, Git, Informatica Power Center, Jira, Scrum, Kanban, AWS Athena, Apache Kafka, Google Cloud Platform(GCP), Big Data, Amazon MTurk, Protegrity, Google BigQuery, Snowflake, Apache Airflow
- **Data Science and Engineering:** Data Science Pipeline (cleaning, wrangling, visualization, modeling, interpretation), Machine Learning (Classification, Regression, Clustering), RFM (Recency, Frequency, Monetary) Analysis, Statistics, A/B Testing, Hypothesis Testing, Excel & Spreadsheet Analysis, CICD, Relational and Non-Relational Databases, Strategic Planning & Decision Making, Critical Thinking & Problem Solving.

EXPERIENCE

University of Arizona (Department of Pediatrics), Tucson, AZ

Feb 2023 – Present

Graduate Research Assistant (Focused on Data Wrangling, Data Analysis and Data Pipeline Mapping)

- Led database optimization efforts, ensuring secure dataflow and seamless data queries.
- Orchestrated data cleaning and linkage pipelines, maintaining detailed logs for transparent data refinement.
- Designed data pipelines for MariaDB to PostgreSQL OMOP tables with Python, boosting transfer speed by 30% via Amazon Athena.
- Streamlined post-survey tasks, reducing processing time by 20% by integrating MTurk tasks and REDCap surveys.
- Utilized advanced statistical tests and logistic regression to reveal crucial care patterns, demographics, and engagement factors in patient data.

Tata Consultancy Services, Mumbai, India

Mar 2018 – Jul 2022

Systems Engineer (ETL Developer)

- **Extract, Transfer and Load (ETL) Expertise:**
 - Enhanced productivity by 25% through the implementation of Informatica PowerCenter 9.x for ETL tasks from Mainframe, EDW sources.
 - Designed ETL components, cutting deployment time by 25% and ensuring versatility for seamless database migration.
 - Established and maintained data pipelines, resulting in a 30% increase in efficiency for business intelligence, reporting, and analytics.
- **Optimization and Cloud Computing:**
 - Enhanced CI/CD performance by tuning source, target, mapping, and session levels, reducing bottlenecks.
 - Achieved 50% faster processing and improved scalability with AWS cloud computing for ETL.
 - Streamlined data integration, cutting errors by 28% and lowering infrastructure costs by 2.5%.
- **SDLC Involvement and Project Management:**
 - Led end-to-end SDLC phases, collaborating with Data Science and Investment Science teams, resulting in successful software deployment.
 - Collaborated with cross functional teams, conducted unit testing, and tracked the project in JIRA, boosting visibility and efficiency by 20%.
 - Managed ETL repositories with continuous integration practices, reducing defects by 15% and enhanced development cycles.
 - Presented findings and insights in a clear and influential manner, driving informed decisions and project success.
- **Leadership and Cross-Functional Expertise:**
 - Led a 12-member ETL development team, focusing on data transformation precision, quality, and performance enhancements.
 - Guided junior developers in data modeling best practices and fundamental data integration, fostering holistic skill development.

PROJECTS

- **Identifying Leaf Phenology of Deciduous Broadleaf Forests from PhenoCam Images** In Progress
 - Develop a tool for predicting leaf phenology in deciduous broadleaf forests across various sites using deep learning (CNN Regression).
 - Three methods being employed: AlexNet, ResNet-50, ResNet-101.
- **YouTube Data Pipeline using Apache Airflow** Oct 2023
 - Automated data extraction and transformation for 5000 YouTube channel IDs, reducing manual effort by 30% while enhancing data reliability.
 - Successfully integrated Apache Airflow for scheduling the ETL processes, achieving a 20% increase in data processing efficiency.
 - Implemented robust error handling mechanisms, reducing downtime by 25% and providing clear notifications for quick issue resolution.
- **House Pricing Profiler using Snowflake Database** Oct 2023
 - Extracted and analyzed house price data from Otodom website using Snowflake Database.
 - Leveraged Bright Data for accurate extraction, optimized query response times, and translated Polish text to English.
 - Provided actionable insights, improving decision-making with 95% accuracy in 60,000+ records.
- **Breast Cancer Prediction** Oct 2023
 - Created a breast cancer detection tool with extensive data analysis and model optimization.
 - Implemented Neural Networks using TensorFlow and Keras, iterated on architectures, monitored training, and adjusted hyperparameters.
 - Achieved 92.9% accuracy with Logistic Regression and a remarkable 97.3% accuracy with Neural Network.
- **Uber Data Analytics** Aug 2023
 - Developed a robust and efficient data pipeline and analytics system for Uber data.
 - Extracted and transformed data at a rate of 500 records per second using Google Cloud Storage and Mage ETL.
 - Boosted data accessibility by 30% with an interactive Looker dashboard translating raw data into actionable visuals.
- **Credit Card Fraud Detection** Aug 2023
 - Developed a fraud detection model to identify and mitigate fraudulent activities effectively.
 - Utilized SMOTE Technique to address data imbalance; ML algorithms used: Decision Tree, Logistic Regression, Random Forest, Naïve Bayes.
 - Achieved an accuracy of 92% and precision of 97% for the best model.

CERTIFICATIONS & AWARDS

- Google Business Intelligence; Coursera (Ongoing), Google Project Management; Coursera (Jul 2023), Google Data Analytics; Coursera (Jun 2023)
- 0.5 FTE (Full Time Equivalent) with Complete Tuition Remission; University of Arizona (Feb 2023, Aug 2023)
- Contextual Master Award for achieving a 30% code execution time reduction, enhancing performance, and integrating 2.5 million customers for PNC Bank; Tata Consultancy Services (Mar 2022)