

Md Raisul Islam Roni

iamronim1@gmail.com · [Phone](#) · [rironi.com](#) · [Github](#) · [LinkedIn](#) · U.S. Permanent Resident

SKILLS

Languages - Python, R, SQL, SAS, MATLAB.

Libraries & Tools - Power BI, Git, pandas, NumPy, SciPy, scikit-learn, XGBoost, statsmodels, PyTorch, SHAP, imbalanced-learn (SMOTE), matplotlib, seaborn.

Methods - Machine Learning, Regression & Classification, Time Series Forecasting, Multivariate Statistics, NLP & Topic Modeling, Feature Engineering, Model Validation & Cross-Validation, Experimental Design.

Selected Coursework - Predictive Analytics, Machine Learning, Artificial Intelligence, Data Management, Time Series Analysis, Advanced Data Analysis, Multivariate Statistics, Statistical Methods I & II, Linear Algebra, Optimization Methods.

EXPERIENCE

[Statistical Consulting Laboratory, ETSU](#)

Johnson City, TN

Graduate Statistical Consultant

Aug 2025 - May 2026

- Provided free statistical consulting to ETSU faculty, staff, and graduate students across multiple disciplines, supporting 30+ research projects from study and survey design through data collection, management, and analysis.
- Advised on experimental design, model selection, statistical inference, and data collection/management strategy; carried out analyses in R, SAS, Python, and Excel and communicated results to non-statistician researchers.
- Delivered introductory workshops and one-on-one training on statistical methods and data analysis software (R, SAS, Python) over typical 2–4 week client engagements.

[Oak Ridge National Laboratory](#) [\[GitHub\]](#)

Oak Ridge, TN

Data Science Practicum (ETSU Collaboration)

Aug 2025 - May 2026

- Analyzed high-throughput, multi-modal imaging data from a poplar (bioenergy crop) genotype trial at ORNL's Advanced Plant Phenotyping Laboratory (APPL); built an end-to-end Python pipeline (pandas, NumPy, SciPy, scikit-learn) integrating multi-source imaging data for 446 plants across multiple genotypes.
- **Morphological growth data:** engineered 26 first-14-day growth features and benchmarked 5 models (Random Forest, Gradient Boosting, Ridge, Lasso, MLP) with genotype-grouped cross-validation to predict final biomass (Lasso CV $R^2 = 0.74 \pm 0.06$); Leave-One-Genotype-Out CV exposed genotype-specific overfitting on unseen genotypes.
- **RGB / spectral color data:** computed the NGRDI vegetation index from RGB imaging and built a genotype-stratified, grouped cross-validation classifier that flags top-performing plants by Day 13–14 ($AUC \geq 0.70$) using only standard RGB cameras.
- Fit non-linear growth-curve models, estimated broad-sense heritability (Final Plant Area $H^2 = 0.52$), and applied SHAP (bar/beeswarm/force plots) to interpret predictions for non-technical stakeholders.

[Niswonger Foundation](#) [\[GitHub\]](#)

Johnson City, TN

Data Science Practicum (ETSU Collaboration)

Jan 2026 - May 2026

- Standardized 3 years of fragmented advisor record-keeping from the Foundation's START college-access program into a normalized 8-table relational schema covering 95 students and 1,456 advising interactions.
- Built a Python (openpyxl) ETL pipeline to parse inconsistent source formats such as color-coded meeting logs, hover-comment annotations, and weekly attendance grids, recovering 180+ free-text case notes stored as Excel cell comments that standard pandas ingestion misses.
- Ran EDA (pandas, matplotlib, seaborn) on intake surveys across 3 cohorts ($n=81$), producing 21 visualizations and surfacing high-need students hidden in cohort-level averages.

[Eastman Chemical Company](#)

Kingsport, TN

Data Science Practicum (ETSU Collaboration)

Jan 2024 - May 2024

- Developed an NLP pipeline to automatically assess the quality of Eastman's technical call reports (unstructured free-text logs), scoring readability, relevance, and content density to help standardize reporting.
- Applied Latent Dirichlet Allocation (LDA) topic modeling to convert unstructured call-log text into structured topics, separating relevant from irrelevant content to drive the relevance and density scores.
- Built custom topic-coherence evaluation metrics and evaluated word-embedding (Word2Vec) and LLM / prompt-engineering approaches to interpret and score the unlabeled call-log text.

Dep. of Mathematics and Statistics, ETSU

Graduate Teaching Assistant

Johnson City, TN

Aug 2023 - Aug 2025

- Served as Instructor of Record for MATH 1710: Precalculus (Algebra), delivering lectures, assessments, and academic support to over 80 undergraduate students.
- Served as Teaching Assistant for MATH 1530: Probability and Statistics, supporting instruction, grading, and learning for over 200 undergraduate students.

PROJECTS

Credit Card Fraud Detection [\[GitHub\]](#)

Python, scikit-learn, XGBoost, SMOTE, SHAP

- Built an end-to-end fraud detection pipeline on a benchmark dataset of 284,807 anonymized European cardholder transactions with extreme class imbalance (0.17% fraud); addressed imbalance with SMOTE and compared Logistic Regression, Random Forest, and XGBoost (best F1 = 0.82 with Random Forest).
- Tuned models with RandomizedSearchCV and used SHAP (TreeExplainer) to explain individual predictions and surface the top fraud-driving features.

Time Series Sales Forecasting [\[GitHub\]](#)

Python, SAS, statsmodels

- Built a SARIMA(1,1,1)(0,1,0)[12] model on 48 months of a retail superstore's furniture sales to forecast 12 months ahead, identifying upward trend and Q4 seasonality through iterative differencing.
- Benchmarked against Seasonal Naive and Holt-Winters baselines (RMSE 168 vs. 186) and validated residuals with Durbin-Watson, Kolmogorov-Smirnov, and Anderson-Darling tests (all passed at $\alpha = 0.05$).

Cardiometabolic Risk Analysis [\[GitHub\]](#)

R, Multivariate Statistics

- Applied PCA, Hotelling's T^2 , and one-way MANOVA to 10,616 records from NHANES (the U.S. CDC's national health survey) across 8 biomarkers; built data-driven risk tiers from PC1 and confirmed statistically distinct group separation (Wilks' $\Lambda = 0.24$, $p < 2.2e-16$).
- Used discriminant analysis to identify HbA1c as capturing 96%+ of the multivariate signal separating cardiometabolic risk groups.

Predicting Reading Proficiency [\[GitHub\]](#)

Python, scikit-learn, XGBoost

- Built an end-to-end regression pipeline predicting reading proficiency across 1,217 Iowa public schools: merged school performance data with public-library funding, staffing, and usage statistics (aggregated to ZIP code) into a 140-feature dataset.
- Trained and tuned Lasso, Ridge, Random Forest, and XGBoost models (scikit-learn Pipeline, cross-validated Grid-SearchCV/RandomizedSearchCV); best model reached test $R^2 = 0.74$.
- Ran feature-importance analysis identifying student socioeconomic factors, not library funding, as the dominant driver of reading outcomes.

RESEARCH

Post-Processing GAN-Generated Synthetic Time

Dec 2025

Series with Dynamic Time Warping. [\[GitHub\]](#)

- Developed a two-phase framework integrating Generative Adversarial Networks (GANs) with Soft-DTW alignment to enhance the temporal alignment of synthetic time series.
- Validated alignment performance through correlation, spectral similarity, and autocorrelation metrics to confirm improved temporal coherence.

EDUCATION

East Tennessee State University

Johnson City, TN

Master of Science: Applied Data Science

May 2026

East Tennessee State University

Johnson City, TN

Master of Science: Mathematical Sciences

Dec 2025

University of Dhaka

Dhaka, Bangladesh

Bachelor of Science: Mathematics

Dec 2021