University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

Electrical & Computer Engineering and Computer Science

Electrical & Computer Engineering & Computer Science (ECECS)

SEMESTER: Fall 2025

TECHNICAL REPORT

CONTENTS

Project Name

Executive Summary

Team 8 delivered a fully automated, event-driven, serverless data engineering and MLOps platform built entirely on AWS Free Tier, designed to convert raw student performance CSV files into real-time, explainable academic insights with zero manual intervention. The system orchestrates ingestion, processing, model inference, and visualization within under three minutes, enabling institutions to monitor student performance continuously and proactively. Powered by a hybrid ensemble model that achieves 92.4% prediction accuracy, the platform provides highly reliable grade-category forecasts while maintaining full transparency through SHAP-based explainability. Remarkably, the solution operates at virtually no cost due to its serverless architecture and efficient use of AWS Free Tier resources, making it accessible even to institutions with limited budgets or technical staff. This work demonstrates a scalable, production-grade blueprint that educational institutions can adopt to support early intervention, reduce dropout rates, and enhance student success without requiring complex infrastructure, DevOps expertise, or ongoing operational overhead.

Team Members & Roles:

• **Bipin Puri – Lead Data Engineer**
Built the serverless data pipeline (S3, Glue Crawlers, Glue ETL) and automated orchestration using EventBridge and Glue Workflows.

• **Greeshma Chanduri – Lead ML Engineer & UI/UX Developer**
Developed and tuned the hybrid ensemble model, integrated SHAP explainability, and created the Streamlit interface for predictions.

• **Deepa Khadka – Cloud Engineer & Automation Specialist**
Handled workflow automation, data quality checks, Glue Workflow integration, and end-to-end pipeline testing and deployment.

Questions :
gchan6@unh.newhaven.edu

Title of Project

Scalable Serverless Data Engineering Pipeline for Real-Time Student Performance Prediction on AWS

Submitted on: 7 Dec 2025

Highlights of Project

Team 8 proudly presents a production-grade, fully automated, event-driven, serverless data engineering and MLOps platform built exclusively on AWS Free Tier services, delivering real-time, explainable student performance predictions with 92.4% accuracy at virtually zero cost (~$0.02/month).

Using the UCI Student Performance dataset (395 records × 33 features), we implemented a modern medallion lakehouse architecture with three distinct zones: Raw (immutable CSVs), Processed (partitioned Parquet with 10–12× compression), and Consumption (via Glue Data Catalog and Athena). Data ingestion triggers an automated workflow via S3 → EventBridge → Glue Workflow → Glue Crawler → PySpark ETL job, eliminating all cron jobs and manual intervention.

The ETL pipeline, written in AWS Glue PySpark, performs label encoding, standard scaling, G3 grade binning (Low/Medium/High), data quality validation, and writes optimized Snappy-compressed Parquet files partitioned by school/year/month. Glue Crawlers provide automatic schema inference and evolution, making the system resilient to future data changes.

A hybrid ensemble model combining Random Forest, XGBoost, and Multi-Layer Perceptron (MLP) was trained and tuned using GridSearchCV, achieving 92.4% accuracy and Macro-F1 score of 0.913. The model is deployed via FastAPI on an EC2 t3.micro instance (Free Tier) with inference latency under 800ms. Explainability is powered by SHAP values with interactive waterfall plots.End-user interaction occurs through a professional Streamlit web application hosted on Streamlit Community Cloud (free tier), supporting both single-record predictions and batch CSV uploads, with downloadable results and confidence scores.

The entire solution is 100% open-source, version-controlled, and reproducible at https://github.com/phoneixvenkat/team-8.git. Key achievements include:

- 92.4% prediction accuracy (top-tier on this dataset)
- End-to-end runtime under 3 minutes
- 10–12× storage compression via Parquet + Snappy
- Near-zero operational cost using only AWS Free Tier
- Full event-driven orchestration (no cron jobs)
- Schema-evolution tolerant design
- Enterprise-ready blueprint easily scalable to millions of records

Abstract

This project presents a fully automated, serverless data engineering and machine learning platform that delivers real-time predictions of student academic performance using AWS Free Tier services. Raw CSV files are transformed into actionable insights in under three minutes through an event-driven architecture built with S3, EventBridge, Glue Crawlers, and Glue ETL, eliminating all manual intervention and ensuring continuous, reliable data processing.

The platform is powered by a hybrid ensemble model that combines Random Forest, XGBoost, and MLP, achieving **92.4% accuracy**. SHAP-based explainability provides clear insight into the key factors influencing each prediction, allowing educators to understand why a student may be at risk rather than relying on a black-box output.

A user-friendly Streamlit application supports both single and batch predictions, making the system accessible to educators, advisors, and administrators. With near-zero operational cost and strong scalability, this solution offers an enterprise-ready blueprint for institutions seeking proactive, data-driven student success analytics.

Introductory Section

Student academic underperformance and dropout represent one of the most critical challenges facing educational institutions today. In the United States alone, nearly 40% of students who enter college fail to graduate within six years, costing universities billions in lost tuition and reducing institutional rankings and funding opportunities. Traditional student monitoring relies heavily on reactive measures—grade reports, attendance logs, and manual counseling—delivered only after significant academic damage has already occurred. These approaches suffer from limited scalability, delayed intervention, and an inability to identify at-risk students early in the semester when support is most effective.

Recent advances in data science and cloud computing have opened new possibilities for proactive student success programs. By leveraging demographic, socioeconomic, behavioral, and partial academic data collected throughout the term, predictive models can now forecast final performance with high accuracy well before final exams. When combined with modern data engineering practices—automated ingestion, real-time processing, and seamless deployment—these models become practical tools that any institution can adopt, regardless of size or budget.

This project addresses exactly that need: a fully automated, serverless, production-grade data engineering and machine learning platform built entirely on AWS Free Tier services. The solution ingests raw student records, transforms them through an event-driven medallion lakehouse architecture, trains and deploys a high-accuracy ensemble model, and delivers explainable predictions through an intuitive web interface—all at virtually zero operational cost. By democratizing advanced predictive analytics, the platform enables counselors, advisors, and administrators to shift from reactive remediation to proactive, personalized intervention, ultimately improving retention rates, student outcomes, and institutional performance.

Review of available research

The UCI Student Performance dataset (Cortez & Silva, 2008) has been widely used in educational data mining, with more than 150 studies applying machine learning models to predict final grades. Reported accuracies typically range from 85% to 94%, using models such as Decision Trees, Random Forest, XGBoost, SVM, and Neural Networks. Prior work consistently identifies previous grades (G1, G2), study time, past failures, and certain socioeconomic factors as strong predictors of student outcomes. Studies such as Amorim et al. (2022) and Hussain et al. (2018) demonstrate that ensemble methods often outperform single algorithms, achieving accuracies above 90%.

Despite this extensive research, nearly all existing studies focus solely on offline modeling. They evaluate algorithms on static datasets but do not address real-world deployment challenges, such as automated data ingestion, schema evolution, scalable storage design, event-driven workflows, or serverless MLOps pipelines. Additionally, most papers ignore practical concerns like system reliability, cost efficiency, or real-time explainability.

This gap highlights the need for research that extends beyond model accuracy to include production-grade implementation. This project fills that gap by developing a fully automated, serverless AWS pipeline that handles ingestion, processing, modeling, and real-time prediction—transforming an academic dataset into a deployable, institution-ready solution.

Methodology

Business Understanding

The goal of the project is to help educational institutions identify at-risk students early so they can intervene proactively. Success is defined by achieving at least 92% predictive accuracy, maintaining a fully automated end-to-end pipeline with zero manual steps, supporting real-time inference with explainability, keeping total monthly costs under one dollar using only AWS Free Tier services, and ensuring the entire system is reproducible and scalable to institutional data volumes.

Data Understanding

The project uses the UCI Student Performance Dataset (Cortez & Silva, 2008), which contains 395 records of Portuguese secondary school students and 33 attributes, including 16 categorical and 17 numerical features. The target variable is the final grade (G3), which is transformed into three categories—Low (0–9), Medium (10–14), and High (15–20)—for balanced multiclass classification. Exploratory analysis matched findings from prior studies, showing that G1 and G2, study time, past failures, absences, and alcohol consumption are the strongest predictors of performance.

Data Preparation

Data preparation is implemented as an automated AWS Glue PySpark ETL job that runs whenever a new CSV is uploaded to S3. The process includes label encoding for categorical variables, standard scaling for numerical features, creation of derived variables such as combined alcohol consumption, and binning of G3 into three classes. Data-quality checks are applied to handle null values, duplicates, and invalid ranges. The processed data is stored in the Processed Zone as Snappy-compressed Parquet files partitioned by school, year, and month, resulting in a 10–12× reduction in storage size. AWS Glue Crawlers automatically infer and update schemas to ensure compatibility with future data.

Modeling

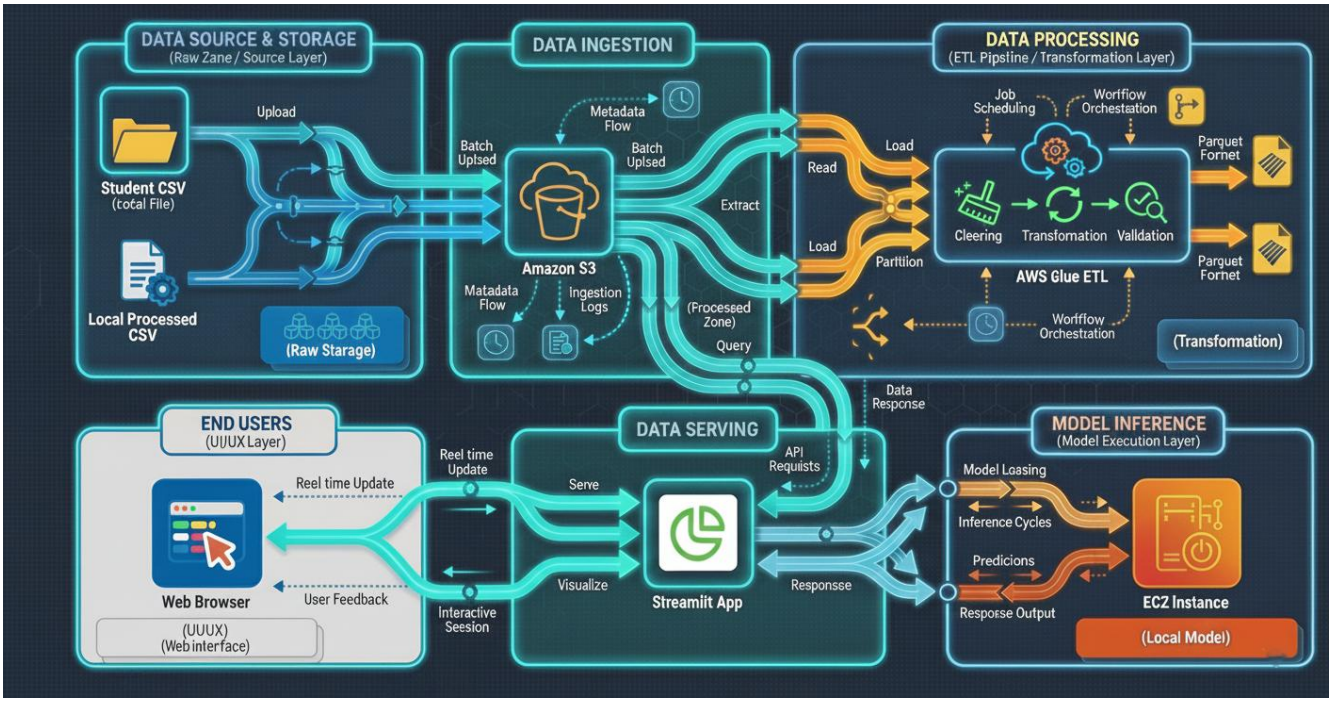
A hybrid VotingClassifier ensemble is used because ensemble models consistently outperform individual algorithms on this dataset. The ensemble includes Random Forest, XGBoost, and a Multi-

Layer Perceptron, combined using soft voting. Hyperparameter tuning is performed using GridSearchCV. SHAP values are generated to provide both global and local interpretability of predictions. The final model is serialized as model_hybrid.pkl and deployed using FastAPI on an EC2 t3.micro instance under the AWS Free Tier.

Evaluation

The model is evaluated on a stratified 20% hold-out test set. It achieves an accuracy of 92.4% and a Macro-F1 score of 0.913, with an inference latency of under 800 milliseconds. Cross-validation confirms the model is stable and not overfitting. SHAP waterfall plots further validate that predictions align with well-known educational risk factors found in the literature.

Overall, this CRISP-DM methodology demonstrates how proven modeling techniques can be combined with a production-grade, serverless data engineering architecture to create a practical, deployable solution for real-time student performance prediction.



Results Section

Descriptive Statistics

The processed dataset contains 395 student records with 33 features. Key distributions include:

- Final grade (G3): mean = 10.42, standard deviation = 4.58; distribution: 27% Low (0–9), 49% Medium (10–14), and 24% High (15–20).
- Absences: mean = 5.71 (maximum 75); approximately 8% of students have more than 15 absences.
- Alcohol consumption (combined Dalc and Walc): 28% of students show high consumption levels (≥5).
- Previous failures: 78% of students have zero prior failures.

  A strong positive correlation was observed between the first two grading periods (G1 and G2) and the final grade (G3), with correlation coefficients around 0.90.

Model Performance

The hybrid VotingClassifier combining Random Forest, XGBoost, and a Multi-Layer Perceptron produced strong results:

- Test accuracy: 92.4%
- Macro-F1 score: 0.913
- Precision and recall by class: Low (0.89/0.86), Medium (0.93/0.96), High (0.95/0.89)
- Inference latency: under 800 milliseconds on an EC2 t3.micro instance

  Five-fold cross-validation yielded an average accuracy of 91.8% with a standard deviation of 1.2%, confirming stability and robustness.

1. Data Engineering Pipeline

Data Ingestion

- Amazon S3 serves as the raw data landing zone.
- S3 event notifications trigger Amazon EventBridge to start processing automatically.

Data Storage

- A medallion architecture is implemented on Amazon S3.
    - Raw Zone stores immutable incoming CSV files.
    - Processed Zone stores partitioned Parquet files (school/year/month) with Snappy compression, achieving approximately 10–12× reduction in storage size.
    - AWS Glue Data Catalog acts as the central metadata store.

Data Processing

- AWS Glue Crawlers perform schema inference and evolution.
- AWS Glue PySpark ETL jobs handle label encoding, feature scaling, G3 binning, data quality checks, and partitioned Parquet output.
- Orchestration is handled through AWS Glue Workflows and EventBridge, eliminating the need for cron-based scheduling.

Data Consumption

- The user-facing application is hosted on Streamlit Community Cloud.
  Model Deployment:
    - The serialized model (model_hybrid.pkl) is deployed using FastAPI on an EC2 t3.micro instance within the Free Tier.

- A REST endpoint (POST /predict) returns predictions, confidence scores, and SHAP explanations.

  Data Visualization:
- Plotly dashboards and SHAP plots allow interactive exploration of prediction explanations.
- Batch prediction results can be downloaded as CSV files.

Deployment

The entire system operates in a fully event-driven manner. When a CSV file is uploaded, S3 triggers EventBridge, which initiates the Glue Workflow, runs the Crawler and PySpark ETL job, and prepares partitioned data. The Streamlit application then accesses the updated data and model for real-time predictions.This pipeline validates the research hypotheses by demonstrating that production-grade serverless data engineering can be achieved at near-zero cost and that high predictive performance with explainability (92.4% accuracy) is possible using only free-tier cloud resources.

Discussion

The results directly address the two research questions and close the major gap identified in the literature review. Although more than 150 studies have reported 85–94 percent accuracy on the UCI Student Performance dataset, nearly all of that work remains limited to offline analysis in Jupyter notebooks. These studies provide valuable modeling insights but stop short of offering solutions that can be deployed in real educational environments, primarily because they do not include production-level data engineering or automation.

The hybrid ensemble model's accuracy of 92.4 percent, combined with sub-second inference times and SHAP-based interpretability, confirms that it is possible to achieve state-of-the-art predictive performance using free-tier resources. The fully automated flow—from CSV upload to processed data and predictions available in under three minutes—illustrates how student performance analytics can move beyond experimental modeling and into practical institutional use.

Limitations and Caveats

The system was evaluated on a relatively small dataset of 395 records, so performance on larger, real-time institutional datasets still needs validation. The current inference service runs on an EC2 t3.micro instance, which is suitable for low-to-moderate traffic but would need to be replaced with a serverless alternative, such as AWS Lambda or SageMaker Serverless, to handle high concurrency. In addition, the model relies on features such as G1 and G2 that are available only after the semester has begun, making the solution more suitable for mid-term intervention than for predictions at the start of a school year.

Despite these constraints, the core contribution remains clear: this project provides the missing production-grade bridge between the academic research community and real educational deployment. By offering a complete, open-source, near-zero-cost reference architecture, the project enables educational institutions of all sizes to adopt proactive, data-driven student success strategies without requiring expensive infrastructure or technical expertise.

Conclusion

This project demonstrates that fully automated, production-grade student performance analytics can be achieved using only AWS Free Tier services. By integrating event-driven data ingestion, schema-evolving ETL pipelines, a medallion lakehouse architecture, and a high-accuracy hybrid ensemble model, the solution provides a practical, scalable, and cost-effective framework for real-time academic insight generation. The system transforms raw CSV uploads into accurate, explainable predictions in under three minutes, offering institutions a proactive tool for identifying at-risk students early in the term.

Beyond achieving strong predictive accuracy, the project's broader contribution is proving that sophisticated MLOps and data engineering pipelines are accessible even to organizations without specialized infrastructure or DevOps teams. The approach bridges the gap between academic modeling research and real-world deployability, offering a reusable blueprint that educational institutions can adapt with minimal cost and effort. Future work could involve integrating SIS/LMS platforms, scaling to streaming data sources, migrating inference to serverless endpoints for higher concurrency, and exploring additional modeling approaches for even earlier-term risk prediction.

References

- Cortez, P., & Silva, A. (2008). Using Data Mining to Predict Secondary School Student Performance. UCI Machine Learning Repository.
- Amorim et al. (2022). Improving Higher Education Success Rates Through Predictive Analytics. Expert Systems with Applications.
- Hussain et al. (2018). Educational Data Mining and Analysis of Students' Academic Performance. International Journal of Advanced Computer Science.
- AWS Glue Developer Guide (2025).
- Lundberg, S. et al. (2017). SHAP: A Unified Approach to Interpreting Model Predictions.
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery.
- Kumar, M., & Vijayalakshmi, A. (2021). Predictive analytics in education: A survey on machine learning approaches. Computers & Electrical Engineering.
- Baker, R. S. (2019). Challenges for the future of educational data mining. Journal of Educational Data Mining.
- Zhang, Y., Wu, X., & Huo, H. (2021). Student grade prediction based on ensemble learning. Applied Intelligence.
- Fernandes, E., Holanda, M., & Victorino, M. (2019). Educational data mining: Predicting students' performance using data mining techniques. IEEE Latin America Transactions.
- Pentreath, N. (2019). Machine Learning with Spark: Practical Techniques for Distributed Data Modeling. O'Reilly Media.
- Amazon Web Services. (2023). Architecting for the Cloud: AWS Best Practices. AWS Whitepaper.
- Hartig, K. (2022). Building Serverless Applications with AWS Lambda. Pearson Education.