Activity 2 — Model Training

Student Dropout Classification

Logistic Regression vs Decision Tree • Results with Synthetic Dataset

Data Mining — Universidad de la Costa • Team: Juan Esteban Bolívar Ferrer

Executive Summary: Results and Recommendations

Goal

Flag at-risk students early; operate with a capacity-aware threshold (policy first, models second).

Approach

Supervised binary classification (dropout 1/0) with identical preprocessing; unsupervised supports cohorts/anomaly/QA.

Synthetic Dataset Results (1,000 rows; 20% dropouts)

- Decision Tree leads on F1 = 0.309 with recall = 0.325 → finds more dropouts today.
- Logistic Regression has better ranking (ROC-AUC = 0.738) but is too conservative at the default 0.50 threshold (recall = 0.125, F1 = 0.213).

Action

Deploy DT as current champion; tune LR (class_weight + threshold sweep) to trade precision for recall.

Attached Dataset (4,424 rows) Status

- Initial run produced dropout_bin = 0 for all rows → label encoding issue (not true class balance).
- Fix: robust recode of Target → dropout_bin (see code). After correct binarization, expect ~30–35% positives (per source dataset), then re-run the same pipeline and threshold policy.

Policy & Governance

Choose an operating threshold to capture ~80% of true dropouts within counselor capacity; show the confusion matrix at that point; run parity checks (gender/origin/SES) and keep a human-in-the-loop with explanations.

Decision Ask (4-Week Pilot Approval)

Approve a 4-week pilot: deploy Decision Tree now on the synthetic-like cohort; in parallel fix & rerun the attached dataset and tune LR.

Deliverables include:

- Risk list + explanations
- Capacity-aware threshold
- ROI & fairness report

Problem and Data Overview

Model Objective

Predicting **dropout within the first academic year** (binary classification 1/0) for early interventions.

Key Features

- **Demographics:** gender, origin, age at entry
- **Academic:** high school GPA, admission score, first semester GPA
- **Financial:** socioeconomic status (SES), scholarship, loan, financial aid

EDA Findings

Strongest correlation: **first_semester_gpa ↔️ dropout (|r| ≈ 0.27)**.
Minor effects: high_school_gpa (~0.10), SES (~0.07).

Datasets used: Synthetic (1,000 records) for initial validation; attached dataset will be processed with an identical pipeline.

Modeling Approach and Preprocessing

01

Main Paradigm

Supervised classification as the primary method. Generates probabilities allowing policy thresholds to be adjusted based on operational capacity.

02

Role of Unsupervised Learning

Supporting analysis: cohort identification, anomaly detection for QA, and feature drift monitoring.

03

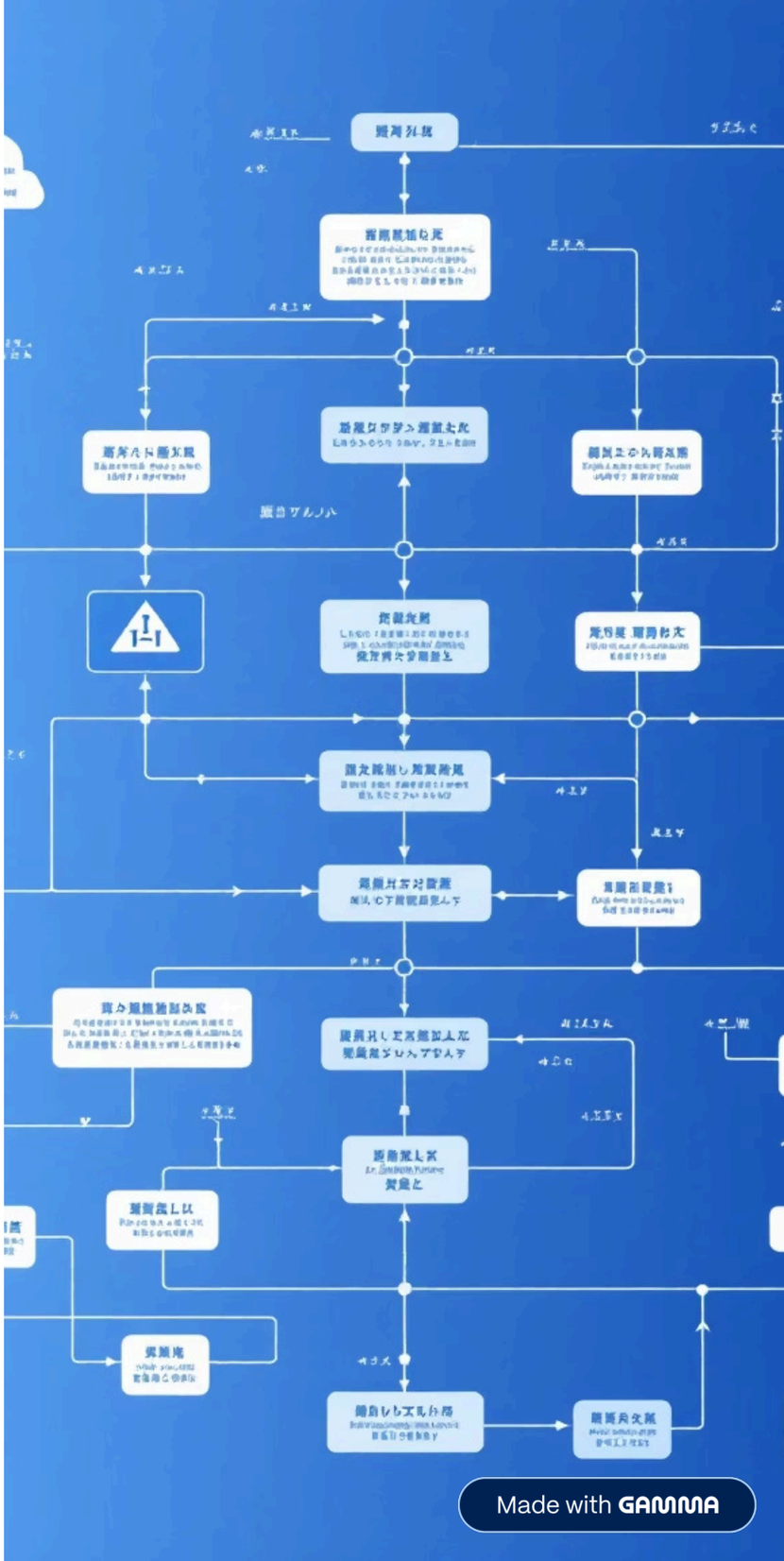
Transformation Pipeline

Numerical variables: median imputation + standardization. **Categorical** variables: mode imputation + one-hot encoding using ColumnTransformer.

04

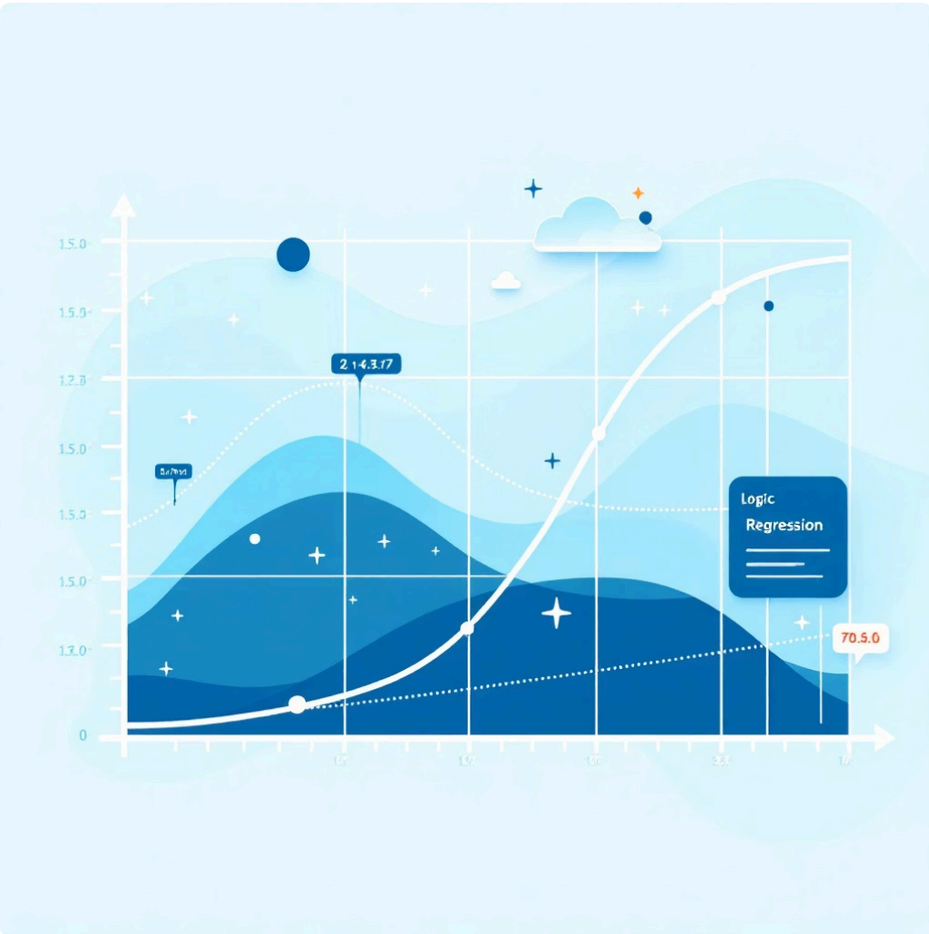
Validation Strategy

80/20 stratified split. Training with **5-fold cross-validation** optimizing F1 score for stability in imbalanced classes.



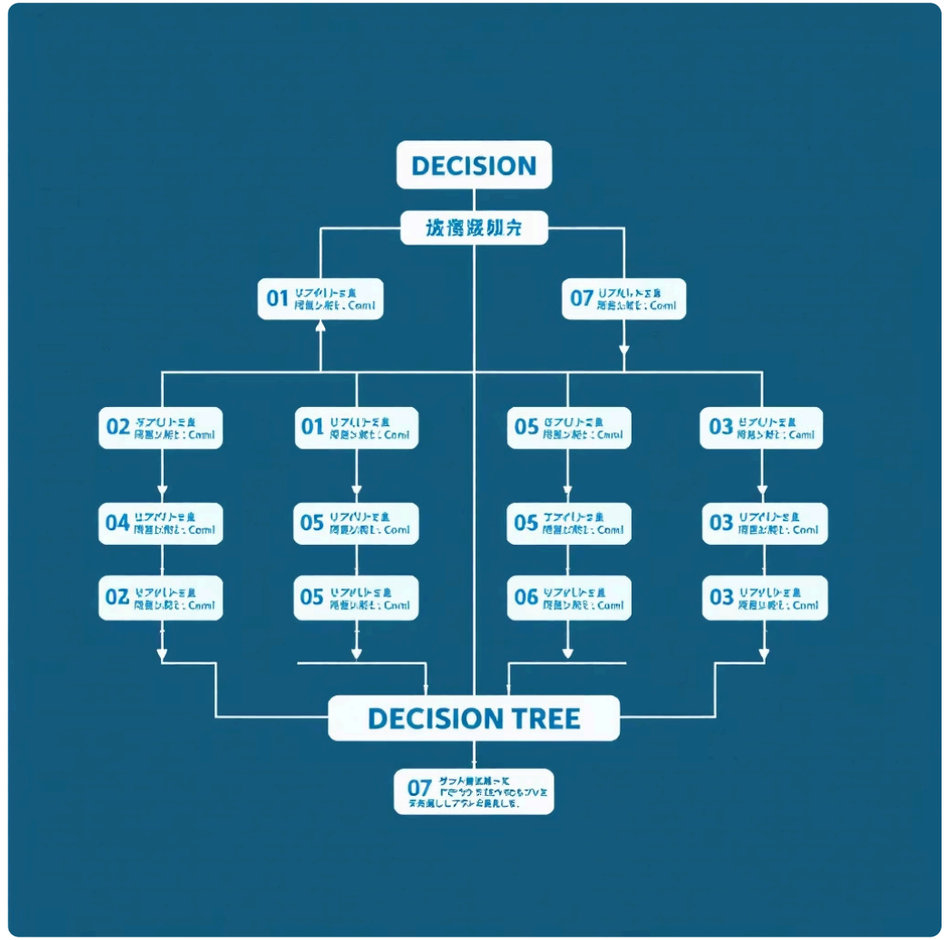
Candidate Model Mechanics

Logistic Regression

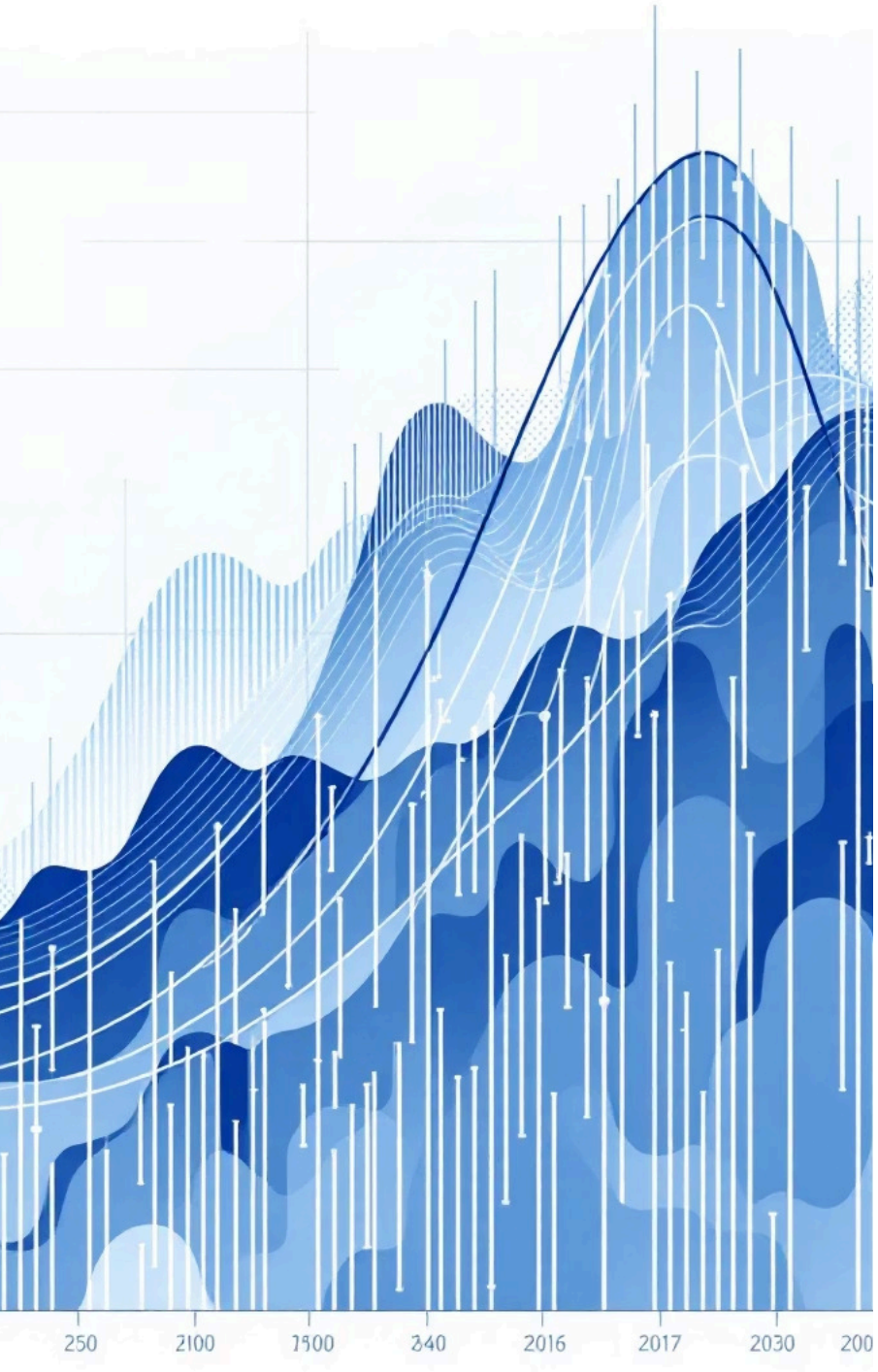


- **Log-odds** model with logistic link function
- Optimizes **log-loss** with L2/L1 regularization
- Generates interpretable **calibrated probabilities**
- Coefficients offer **direct transparency**

Decision Tree



- **Recursive partitioning** of the feature space
- Uses impurity (Gini/entropy) for optimal splits
- Learns non-linear **if-then rules**
- Risk of **overfitting** without proper pruning



Validation Framework and Prioritized Metrics

Training

5-fold stratified cross-validation optimizing **F1** to ensure stability across minorities.

Test Report

Complete metrics: Accuracy, Precision, **Recall, F1, ROC-AUC**. Documentation of confusion matrices.

Recall-Centric View

Prioritization of **F1 and Recall@Top-K**. **False negatives are costly**: at-risk students not identified.

ROC curves allow visualizing the precision-recall trade-off across multiple decision thresholds.

Results: Test Performance Comparison

0.738

ROC-AUC: Logistic Regression

Better ranking capability, but threshold needs optimization

0.325

Recall: Decision Tree

Identifies more students at risk with current configuration

0.309

F1: Decision Tree

Better precision-recall balance without additional tuning

Logistic Regression

- Accuracy: **0.815**
- Precision: **0.714**
- Recall: **0.125** ⚠️
- F1: **0.213**
- CV F1 (train): **0.149 ± 0.028**

Decision Tree

- Accuracy: **0.710**
- Precision: **0.295**
- Recall: **0.325** ✓
- F1: **0.309** ✓
- CV F1 (train): **0.300 ± 0.072**

F1 Selection: Decision Tree leads in current configuration due to superior recall, critical for risk identification.

Threshold Optimization: Capacity-Aware Approach

Threshold Sweep

Explore multiple cutoff points to find the optimal precision-recall balance.

Operational Constraint

Select threshold that delivers **~80% recall** within counselor capacity.

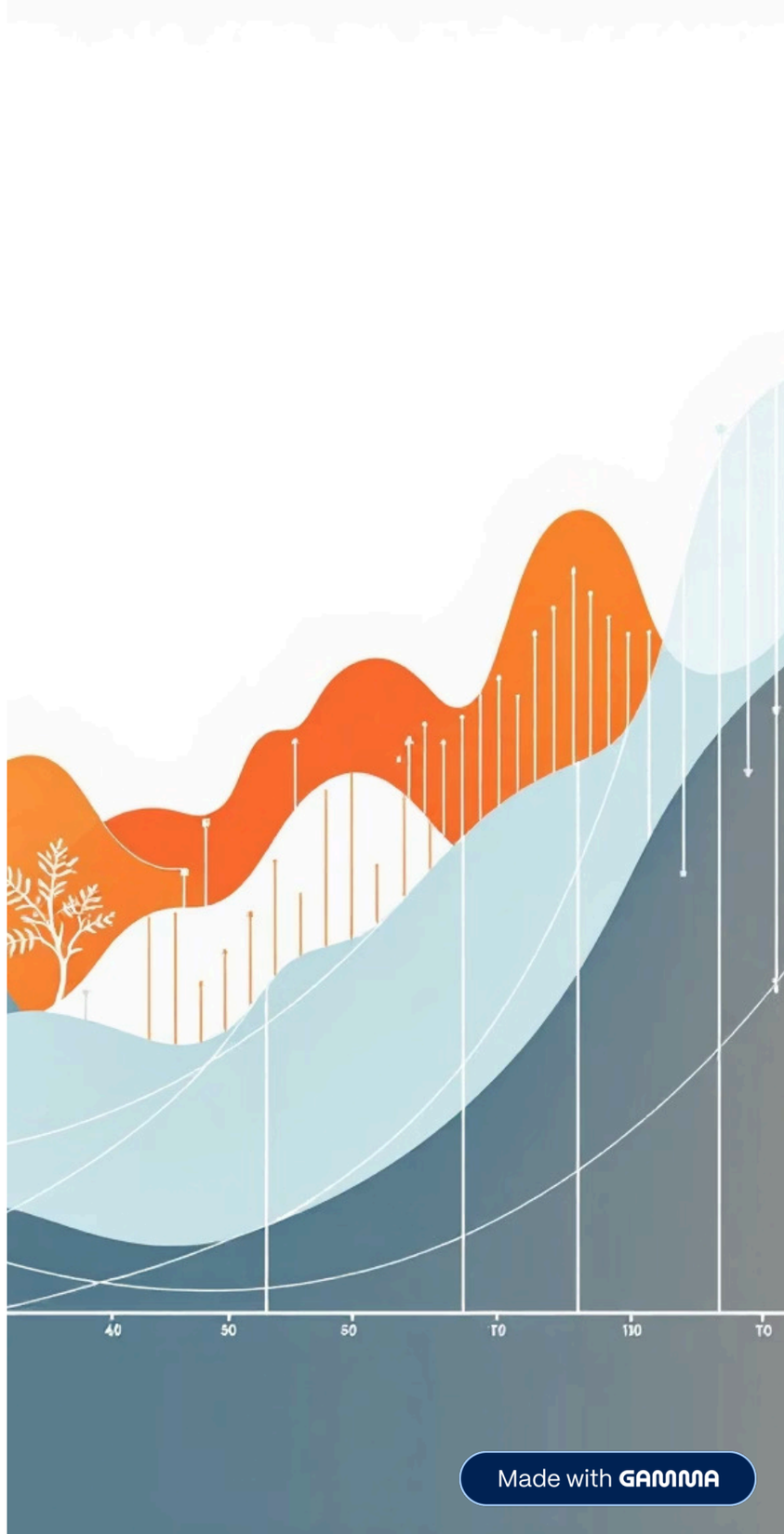
Adjustment for LR

Implement **class_weight='balanced'** + threshold **≈ 0.25-0.40** to trade precision for recall.

Policy Report

Document confusion matrix, signaled students, and **Expected ROI** at the chosen threshold.

> "The optimal threshold is not an isolated technical metric, but a strategic decision that balances operational resources with student impact."



Equity and Model Governance

Equity Framework

Parity Audits

Verify metrics at the **selected policy threshold** segmented by gender, origin, and socioeconomic level.

Data Privacy

Minimize **Personally Identifiable Information (PII)**, apply strict access controls, and maintain a **complete audit log**.

Explainability

Provide **SHAP explanations at the student level** for transparency in individual decisions.

Human Oversight

Maintain **human in the loop**: mandatory review by counselors before interventions.

Deployment, Next Steps, and Required Decision

1 Pilot Phase (4 weeks)

Deploy **Decision Tree** into production. Simultaneously, **adjust RL** with class_weight and threshold sweeping.

2 Weekly Operation

Weekly scoring → **dashboard** with risk list + explanations. Continuous monitoring of feature drift.

3 Comparative Evaluation

If RL achieves recall parity with better AUC/precision, **promote RL**. Otherwise, retain DT and evaluate **Gradient Boosting**.

4 Annual Recalibration

Scheduled retraining with updated data to maintain model relevance and precision.

Decision Requested

Approve 4-week pilot + key deliverables:

- Weekly **risk list** with scoring
- Confusion matrix at policy threshold
- Analysis of **expected ROI**
- **Equity audit** by demographic segments