# PREDICTING STUDENT DROPOUT RISK USING LOGISTIC REGRESSION IN SPSS

## 1. Background and Problem Statement

A regional community college in the U.S. reported increasing dropout rates, particularly among first-year students enrolled in associate degree programs. The administration lacked a data-driven mechanism to identify at-risk students early and deploy intervention programs proactively. While demographic and performance data were regularly collected, there was no established model to predict dropout likelihood. This project involved the use of logistic regression in SPSS to build a predictive model that classifies students into high-risk and low-risk dropout categories based on pre-enrollment and early academic performance indicators.

## 2. Objectives

- To develop a logistic regression model in SPSS that predicts the probability of student dropout based on key academic and demographic variables
- To identify significant predictors influencing dropout decisions
- To provide a classification report that enables academic counselors to target intervention efforts
- To build an easy-to-update SPSS workflow that can be reused each semester

## 3. Methodology

### 3.1 Data Source and Variables

- Dataset: 2,350 first-year students enrolled between Fall 2021 and Fall 2023
- Variables used in the model:
    - Age
    - Gender
    - Ethnicity
    - High school GPA
    - First-semester credit load
    - Attendance in orientation

- Use of tutoring services
- First-semester GPA
- Number of failed courses
- Financial aid status
- Dropout status (binary: 1 = Dropped out, 0 = Retained)

### 3.2 Data Preparation in SPSS

- Cleaned missing values and coded categorical variables using dummy variables
- Performed correlation analysis to avoid multicollinearity
- Normalized skewed numeric variables

### 3.3 Logistic Regression Model in SPSS

- Used SPSS's Binary Logistic Regression procedure
- Model fit assessed using Nagelkerke $R^2$, Hosmer–Lemeshow test, and classification table
- Checked odds ratios and confidence intervals for all predictor variables

## 4. Results and Interpretation

**Model Accuracy**

- Correctly classified: 83.1% of cases
- Area under ROC curve (AUC): 0.79
- Hosmer–Lemeshow significance: 0.423 → good model fit

**Significant Predictors ($p < 0.05$):**

- High school GPA (lower GPA associated with higher dropout risk)
- First-semester GPA (strongest predictor; OR = 0.31)
- Attendance at orientation (non-attendees had 2.7x higher odds of dropping out)
- Use of tutoring services (non-users more likely to drop out)
- Financial aid status (students with unstable aid more prone to dropout)

## 5. Recommendations

- Make orientation mandatory for all new students

- Set GPA-based flags in the student management system to alert advisors
- Expand outreach for tutoring services and link them with at-risk profiles
- Include model predictions in weekly advisor dashboards using SPSS syntax automation
- Conduct follow-up surveys with dropout students to validate predictive variables further

## 6. Deliverables

- SPSS .sav file with labeled data and regression outputs
- Final logistic regression model syntax with interpretation guide
- Predictive accuracy report with ROC curve and confusion matrix
- One-page summary sheet highlighting top dropout risk indicators

## 7. Stakeholder Relevance

**Academic:**

- Useful for institutional research departments to inform policy and retention strategy
- Demonstrates practical use of logistic regression in education research settings
- Relevant for coursework in education analytics, data science, and public policy

**Corporate (EdTech/Consulting):**

- Template model for retention-focused analytics platforms
- Applicable for community colleges and online learning providers in the U.S.