Project title: Student Retention Capstone – Sprint 2

Name: Harshitha Koppala

Course: INST414

Instructor: Huy Nghiem

Date of submission: October 22, 2025

1.) Data Acquisitions & Description

a. Through the acquisition, cleaning, and exploration of the primary dataset, this sprint concentrated on moving from planning to execution. The Predict Student Dropout and Academic Achievement dataset from Kaggle serves as the main source of data. It contains more than 4,500 student records with academic, socioeconomic, and demographic information from a Portuguese university. IPEDS (Integrated Postsecondary Education Data System), a secondary dataset, supplements the primary dataset with demographic and retention data at the institutional level.

b. A single student is represented by each row in the dataset. Age, gender, prior and first semester GPA, scholarship status, attendance rate, credits earned, and dropout status (the target variable) are among the variables. The dataset keeps about 4,480 records after initial cleaning.

Variable	Type	Description	Missing %	Relevance
Dropout	Binary	1=dropout, 0=persist	0%	★ Target
First_sem_gpa	Continuous	GPA after first semester	3%	★
Prior_GPA	Continuous	GPA prior to admission	1%	★
Scholarship_status	Binary	Receiving financial aid	2%	★
Attendance_rate	Continuous	Percentage of classes attended	5%	★
Age_at_enrollment	Continuous	Age when enrolled	0%	

2.) Data Quality Assessment & Cleaning

a. Handling missing values, eliminating duplicates, identifying outliers, and developing derived variables like GPA change (gpa_delta) were all part of the

data cleaning process. Categorical variables were consistently encoded, and numerical columns with low missingness were subjected to median imputation. Extreme outliers were confirmed and eliminated, such as those with ages greater than 100 or GPAs greater than 4.0.

Variables like financial_stress (a binary flag for students with tuition balance and no scholarship) and gpa_delta (difference between first semester and prior GPA) were produced through feature engineering. There are 25 variables and 4,480 records in the final cleaned dataset.

3.) Explanatory Data Analysis

a. Important information regarding academic achievement and dropout probability was uncovered by EDA. The majority of students, according to univariate plots, have GPAs between 2.5 and 3.5, with dropout students grouped at the lower end. According to bivariate analysis, higher dropout rates are linked to lower first-semester GPAs and a lack of scholarships. The best indicators of dropout, according to a correlation heatmap, were attendance rate and GPA variables.

b. Placeholder Figures:

   i. Figure 1: First-Semester GPA Histogram (dropout vs. persist);
   ii. Figure 2: GPA by Dropout Status Boxplot;
   iii. Figure 3: Numerical Features Correlation Matrix
   iv. Figure 4: Scholarship Status-Related Dropout Rate

Every figure has descriptive captions, distinct titles, and labeled axes.

4.) Refined Problem Statement & Analytic Plan

a. "Use early academic and financial indicators to predict whether a first-year student is at risk of dropping out" is the revised problem statement. Early identification during the first two semesters is now the main focus of the analysis. To balance interpretability and predictive power, the modeling will make use of Random Forest, Gradient Boosting, and Logistic Regression. AUC-ROC, precision, and recall (for at-risk detection) are evaluation metrics. A 60/20/20 train-validation-test split will be used when applying cross-validation.

5.) Progress Tracking & Next Steps

a. Sprint 2 achievements included: obtaining and cleaning the primary dataset and performing exploratory data analysis using visualizations. Developed baseline modeling methodology. Added notebooks and README to the GitHub repository
b. Sprint 4 will concentrate on model tuning, interpretability, and final recommendations for universities to enhance student retention, while Sprint 3 will deploy baseline models and assess preliminary findings.
c. Self-evaluation: The project is proceeding as planned. Maintaining class balance between persisters and dropouts is the primary challenge. Cost-sensitive algorithms and oversampling are examples of mitigation. To improve generalizability, more institutional data from IPEDS might be utilized.

6.) Limitations
a. Dataset is from a single Portuguese university, limiting generalizability
b. Some variables may have reporting errors or missing values
c. Institutional data may not capture all student-level risk factors

7.) References

a. Kaggle: *Predict Student Dropout and Academic Success Dataset*
b. U.S. Department of Education. *Integrated Postsecondary Education Data System (IPEDS)*
c. Scikit-learn Documentation: *Model Evaluation Metrics*
d. Pedregosa et al. (2011). *Scikit-learn: Machine Learning in Python.*

8.) Feature Engineering Details
a. Predictive modeling can be greatly enhanced by feature engineering.
b. The difference between the first semester's and previous GPAs is known as gpa_delta.
c. Determines if a student's academic performance is getting better or getting worse.
d. Higher dropout rates are frequently correlated with negative values.
e. For students with unpaid tuition and no scholarships, financial_stress is a binary flag.
f. Determines which students are most likely to face financial difficulties.
g. Attendance_rate_category: Low, medium, and high attendance categories
h. Simplifies analysis and draws attention to risk levels in visuals

i. Credit_completion_ratio: Earned credits divided by attempted credits

j. Aids in identifying persistence risk and academic progress.

9.) Preliminary Findings from EDA

a. Dropout and GPA:

i. The highest dropout rate (~40%) is found among students with a first-semester GPA of less than 2.5.

ii. Persistence has a positive correlation with GPA improvement from before the first semester (gpa_delta).

b. Financial aid and scholarships:

i. Dropout rates are higher among students who lack scholarships or who are under a lot of financial stress.

c. Attendance:

i. Dropout rates are highly associated with low attendance (less than 70%), even for students with respectable GPAs.

d. Populations:

i. While age at enrollment has little bearing, older students may encounter greater difficulties when paired with other risk factors.

10.) Actionable Recommendations for Universities

a. Early observation of pupils who are at risk:

i. Monitor for signs of financial stress, attendance, and first-semester GPA.

b. Specific financial assistance:

i. Scholarships or financial aid for students who have been identified as having financial difficulties

c. Programs for academic support:

i. Study sessions, tutoring, and early warning notifications for students whose GPA is dropping

d. Initiatives for engagement:

i. Promote attendance in class through peer programs, interactive learning, and mentoring.

e. Preliminary EDA offers practical tactics, but complete recommendations will follow modeling:

i. These early insights can already assist to universities in planning preventive interventions.