Student Academic Map: ML-Powered Student Profile, Trajectory, and Risk Visualization

1. INTRODUCTION

The goal of this project is to develop a machine learning system that predicts academically at-risk students and visualizes their academic performance. By analyzing past GPA, attendance, credit load, and failed courses, this system can help educators identify students needing early intervention. The project uses a synthetic student dataset to demonstrate the methodology.

2. DATASET DESCRIPTION

The dataset contains academic records for 100 students across 6 semesters. Each record includes the following features:
- Student_ID: Unique student identifier
- Semester: Academic semester (1–6)
- Previous_GPA: GPA from the previous semester
- Current_GPA: GPA for the current semester
- Attendance_Percent: Student's attendance percentage
- Credit_Load: Number of credits registered
- Failed_Courses: Number of courses failed

The dataset contains no missing values and provides sufficient information for analyzing student performance and identifying risk.

3. EXPLORATORY DATA ANALYSIS & VISUALIZAITON

EDA was conducted to understand the dataset and identify patterns in student performance. The analysis included:
- GPA distribution, which revealed most students perform moderately, with a subset at risk.
- GPA trends over semesters, highlighting improving, stable, and declining trajectories.

- Attendance versus GPA, showing that higher attendance is generally associated with higher GPA.
- Attendance versus GPA, indicating that students with low attendance tend to have a lower GPA.

A simple risk rule was applied: students with Current_GPA < 2.0 or Failed_Courses >= 2 were classified as academically at risk.

Visualizations were created to support the analysis and interpretation of results:

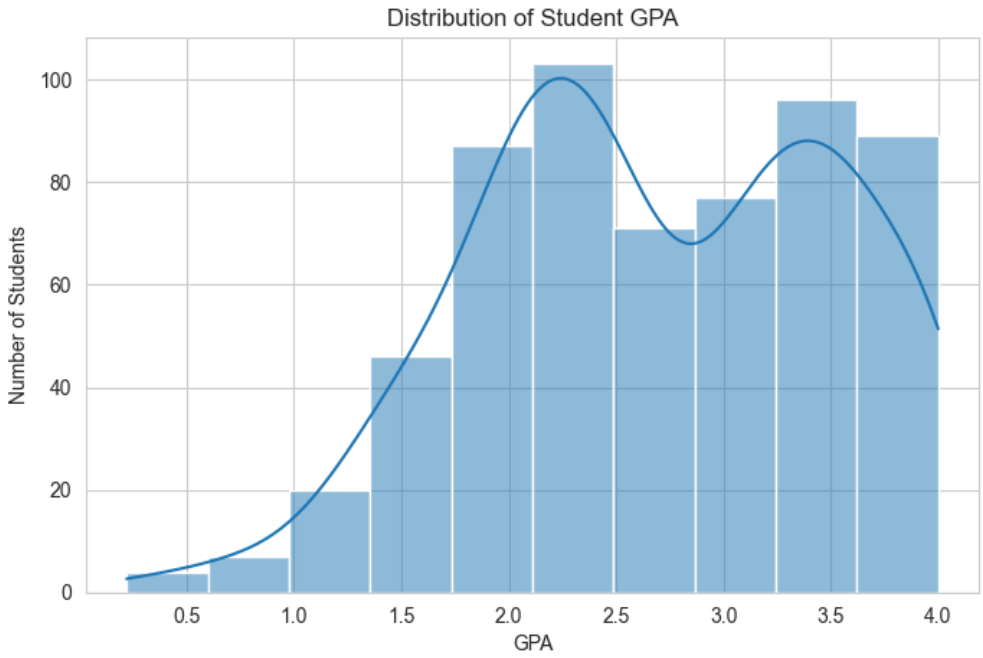


Figure 3 - GPA distribution histogram

This chart shows how students' GPA is distributed across the dataset, providing an overall view of academic performance and helping identify low-performing students. It shows how many students have low, average, or high GPA.

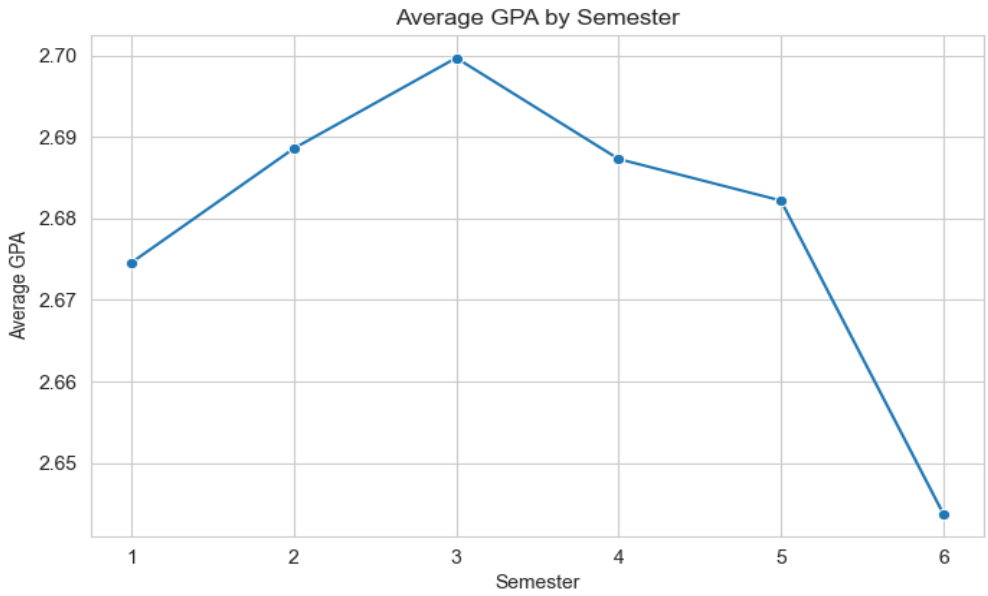


Figure 4 - GPA trend line by semester

This chart illustrates the change in average GPA across semesters, showing whether student performance improves or declines over time. It tracks students' academic progress across semesters.

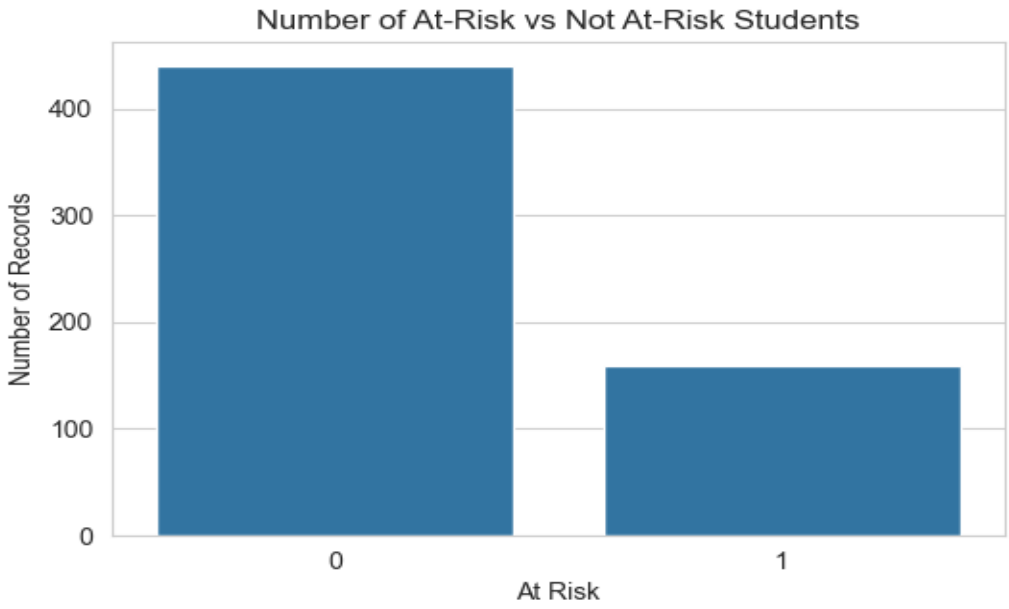
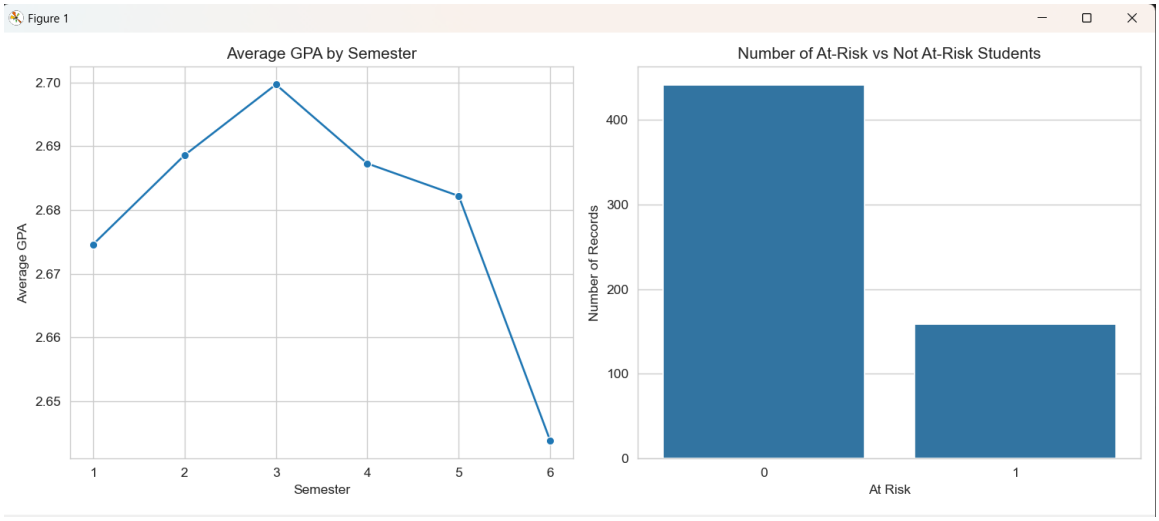


Figure 5 - Risk count bar chart

This chart compares the number of at-risk and not-at-risk students, giving an overview of academic risk within the dataset. It shows how many students need academic support.



Figuure 7 - Mini dashboard combining GPA trend and risk count

This mini dashboard combines GPA trend and risk count. The line chart shows how average GPA changes across semesters, while the bar chart shows how many students are at risk. Together, they help us understand both academic progression and the level of risk among students.

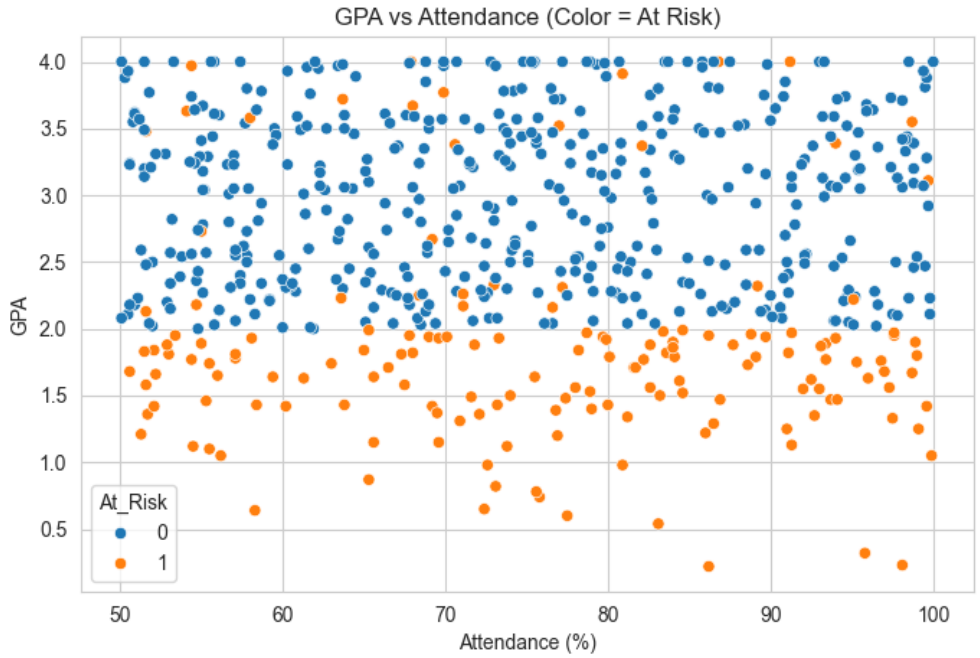


Figure 6 - GPA vs attendance scatter plot

This chart compares the number of at-risk and not-at-risk students, giving an overview of academic risk within the dataset. Students who attend more classes usually perform better.

4. PROBLEM FORMULATION

The project formulates the academic risk prediction as a binary classification task.
- Target variable: At_Risk (1 = at risk, 0 = not at risk)
- Features: Previous_GPA, Attendance_Percent, Credit_Load, Failed_Courses
- Objective: Predict which students are at risk based on their academic history.

5. MACHINE LEARNING MODEL

A Logistic Regression model was selected for its simplicity and effectiveness in binary classification. The model was trained on 80% of the dataset and tested on the remaining 20%. The training process involved learning patterns from previous GPA, attendance, credit load, and failed courses.

Python Code

```
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X = df[["Previous_GPA","Attendance_Percent","Credit_Load","Failed_Courses"]]
y = df["At_Risk"].astype(int)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random

model = LogisticRegression(max_iter=1000)
model.fit(X_train, y_train)
y_pred = model.predict(X_test)
```

6. MODEL EVALUATION

The model's performance was evaluated using accuracy and a confusion matrix.
- Accuracy measures how often the model predicts correctly.
- Confusion matrix shows where the model makes errors:
  - True Positive (TP): Correctly identified at-risk students

- True Negative (TN): Correctly identified not-at-risk students
- False Positive (FP): Incorrectly predicted at-risk students
- False Negative (FN): Missed at-risk students

Python Code

```
from sklearn.metrics import accuracy_score, confusion_matrix, classification_re

accuracy = accuracy_score(y_test, y_pred)
cm = confusion_matrix(y_test, y_pred)
print("Accuracy:", accuracy)
print("Confusion Matrix:\n", cm)
print(classification_report(y_test, y_pred))
```

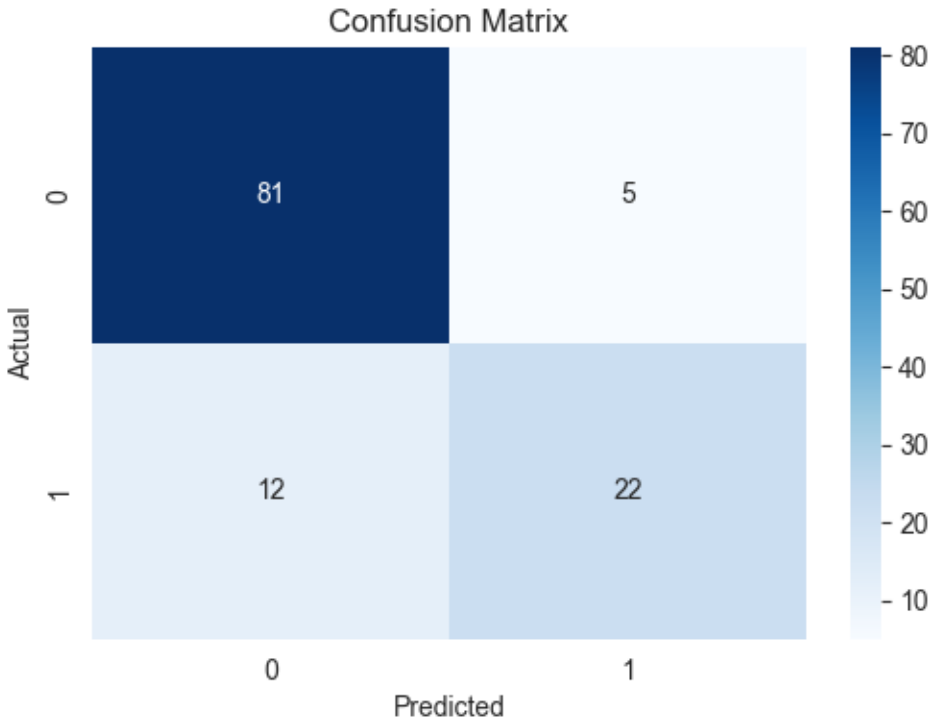


Figure 1 - Confusion matrix heatmap

The confusion matrix shows how many students the model classified correctly and incorrectly.

The diagonal boxes represent correct predictions, while the others represent mistakes.

From this chart, we can see that the model performs well in identifying at-risk students.

```
(student_map_env) PS C:\Users\DELL\Desktop\NCAIR\Data science
dvanced project\DS Advanced.py"
              precision    recall  f1-score   support

           0       0.87      0.94      0.91        86
           1       0.81      0.65      0.72        34

    accuracy                           0.86       120
   macro avg       0.84      0.79      0.81       120
weighted avg       0.86      0.86      0.85       120
```

Figure 2 - Precision, recall, F1-score table

Precision tells us how accurate the model is when it predicts a student is at risk.
Recall tells us how many of the actual at-risk students the model was able to find.
The F1-score combines both precision and recall into one value.

8. CONCLUSION

The project demonstrates how machine learning can be applied to student academic data to identify at-risk students. Logistic Regression effectively predicts students at risk using features such as GPA, attendance, credit load, and failed courses. Visualization of trends and risk categories supports decision-making for early intervention.

Future work could include additional features, larger datasets, or more advanced models to improve prediction accuracy.