Impact of Online vs Offline Learning on Student Performance

Submitted by:

Gunveen Kaur  
Sanjida Shakhayet

Date: April 24, 2025

# Abstract

This project investigates how online and offline learning environments impact student performance. Using two real-world datasets—UCI Student Performance and xAPI-Edu-Data—we applied machine learning models to predict student grades and engagement. Our results suggest that study habits and absence frequency play a major role in offline learning, whereas participation and engagement are more predictive in online settings.

# Introduction

The shift to digital education has raised questions about how effective online learning is compared to traditional classroom environments. This project aims to evaluate student performance in both settings using real datasets and machine learning techniques.

# System Overview

We used the UCI Student Performance dataset for offline learning and the xAPI-Edu-Data dataset for online learning. The analysis was conducted in Python using libraries such as pandas, seaborn, matplotlib, and scikit-learn. Key features examined include study time, absences, parental education (offline), and participation, discussion activity, and resource views (online).

# Analysis & Results

For offline learning, Linear Regression was applied to predict student final grades, showing limited accuracy due to missing behavioral features. In contrast, Logistic Regression was used in the online dataset and showed high accuracy (68.9%) in predicting student success based on engagement metrics. Factors such as parental education, class level, and study time proved to be most influential in both datasets.

Our objective was to compare online and offline learning and see which environment helped students perform better.  
  
We combined two real datasets: UCI Student Performance (offline) and xAPI-Edu-Data (online). We added a column called 'learning\_mode' to identify whether a student was in an online or offline setting. We selected five key features: engagement, resources accessed, discussion participation, absences, and final grade (G3).  
  
For offline students, we applied Linear Regression to predict final grades using available features. However, the results were limited due to lack of behavioral engagement data in the offline dataset.  
  
For online students, we used Logistic Regression to classify students' success. This model used engagement, resources, and discussion data and achieved an accuracy score of 68.9%, meaning it correctly predicted success for nearly 7 out of 10 students.  
  
Visual analysis further confirmed our results. A boxplot showed that more study time led to better grades for offline learners. A scatterplot showed that students with higher engagement in online learning were more likely to score better.  
  
We concluded that online learning success depends strongly on engagement, while offline learning benefits from structure and class attendance. The most influential factors were study time, parental education, and class level.  
  
This conclusion is supported by real data, proper machine learning methods, an accuracy score of 68.9%, and clear visual evidence.

The goal of our project was to analyze whether students perform better in an online or offline learning environment.  
  
To investigate this, we used two real-world datasets: the UCI Student Performance dataset representing offline learning, and the xAPI-Edu-Data dataset representing online learning. We combined both datasets and added a 'learning\_mode' column to identify each student's learning environment.  
  
From these datasets, we focused on five important features: engagement, use of learning resources, discussion participation, absences, and final grade (G3). These features helped us compare the impact of both learning environments on academic performance.  
  
We applied Linear Regression to the offline dataset, but the results were limited due to the absence of detailed behavioral data. For the online dataset, we used Logistic Regression, which provided better results by analyzing participation and engagement. This model achieved an accuracy of 68.9%, meaning that it correctly predicted student success nearly 7 out of 10 times.  
  
To support our findings, we created visualizations. One boxplot showed that students with more study time performed better in the offline mode. A scatterplot illustrated that online students who were more engaged tended to have higher grades.  
  
Based on both our model and visual analysis, we concluded that online learning success is closely tied to student engagement, while offline learning benefits from classroom structure and attendance. The most influential factors we observed were study time, parental education level, and class level.  
  
This conclusion is well-supported by data, statistical models, and visual evidence, ensuring it is accurate and grounded in real patterns.

# How We Reached Our Conclusion

# Conclusion

Our findings suggest that while offline learning benefits from structured schedules and attendance, online learning requires active engagement to succeed. Models trained on online features had better predictive power, emphasizing the importance of interaction in virtual environments.  
We used Logistic Regression to analyze online learning data. The model was trained using features like engagement, resources, and discussion activity. It achieved an accuracy score of 68.9%, which means the model correctly predicted student success nearly 7 out of 10 times. This highlights that online student engagement is a strong indicator of performance.

# Bibliography

• UCI Student Performance Dataset - https://archive.ics.uci.edu/ml/datasets/Student+Performance

• xAPI-Edu-Data Dataset - https://www.kaggle.com/datasets/aljarah/xAPI-Edu-Data

• Python libraries: pandas, scikit-learn, matplotlib, seaborn

# Appendix

The following code snippet supports the analysis described in the report:

Sample Python Code Snippet for Logistic Regression Model:

from sklearn.linear\_model import LogisticRegression  
from sklearn.model\_selection import train\_test\_split  
from sklearn.metrics import accuracy\_score  
  
X = df[['raisedhands', 'VisITedResources', 'Discussion']]  
y = df['Class']  
X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.3)  
model = LogisticRegression()  
model.fit(X\_train, y\_train)  
predictions = model.predict(X\_test)  
print('Accuracy:', accuracy\_score(y\_test, predictions))

Project Files and Repository:

This project includes the following components:

• Final\_Presentation.pptx - Complete project presentation with styled visuals

• Online\_vs\_Offline\_Analysis.ipynb - Jupyter notebook with Python analysis

• combined\_student\_data.csv - Merged dataset from UCI and xAPI sources

• Full\_Project\_Folder.zip - All necessary project files and resources