STUDENTS PERFORMANCE DETECTION USING MACHINE LEARNING

ARSHA SHAJI M ,

MSc STATISTICS WITH DATA SCIENCE SREE SANKARA COLLEGE , KALADY

Period of Internship: 25th August 2025 - 19th September 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

Abstract

This project focuses on applying unsupervised learning to the Student Performance dataset. The dataset consists of 395 entries with 33 attributes covering academic, demographic, and lifestyle factors. Initial exploratory data analysis (EDA) was carried out to study distributions, correlations, and overall structure. Categorical features were encoded, and numerical features were standardized for uniformity. Dimensionality reduction was achieved using Principal Component Analysis (PCA), reducing the dataset to two components. K-Means clustering was then applied on the reduced data to form distinct student groups. The effectiveness of clustering was measured using the silhouette score. Visualization techniques such as heatmaps and scatter plots were employed to interpret cluster patterns. An object-oriented pipeline was developed for preprocessing, clustering, and evaluation. The project demonstrates how PCA and clustering can uncover hidden insights into student performance.

Introduction

The project focuses on analyzing the Student Performance dataset using unsupervised machine learning techniques. The dataset contains demographic, academic, and social information of 395 students, making it suitable for studying hidden patterns and clusters among learners. The relevance of this project lies in the growing importance of data-driven decision-making in education, where clustering can help identify student groups with similar academic behaviors, study habits, and lifestyle factors.

The technologies involved include Python programming, scikit-learn, pandas, matplotlib, seaborn, and OpenCV. These tools were used for data preprocessing, dimensionality reduction, clustering, and visualization. Background knowledge from machine learning, statistics, and data mining was applied, particularly the concepts of Principal Component Analysis (PCA), K-Means clustering, and silhouette score evaluation.

The procedure followed in the project started with exploratory data analysis (EDA) to understand the dataset. Next, categorical variables were encoded and numerical data standardized for consistency. PCA was applied to reduce dimensionality, followed by K-Means clustering to identify hidden groups of students. The clustering quality was measured using the silhouette score, and results were visualized through heatmaps and scatter plots.

The main purpose of the project is to demonstrate how unsupervised learning combined with dimensionality reduction can uncover meaningful insights from complex student data. Such insights can support educational institutions in improving student performance, designing interventions, and categorizing learners into performance-based clusters.

Topics Covered During Training :

- Basics of Machine Learning – supervised vs. unsupervised learning
- Exploratory Data Analysis (EDA) techniques
- Data preprocessing – handling categorical and numerical features
- Data visualization using matplotlib and seaborn
- Dimensionality reduction – Principal Component Analysis (PCA)
- Clustering techniques – K-Means algorithm
- Cluster evaluation using silhouette score
- Object-Oriented Programming (OOP) approach in Python
- Correlation analysis and heatmaps
- Visualization of clusters using Matplotlib and OpenCV

Project Objective

Objectives of the Project :

- To explore and analyze the Student Performance dataset through exploratory data analysis (EDA) and preprocessing, ensuring data is clean, consistent, and ready for modeling.

- To apply dimensionality reduction using Principal Component Analysis (PCA) in order to simplify the dataset while retaining its most significant information for clustering.

- To implement K-Means clustering on the reduced dataset to group students into meaningful clusters based on their academic performance, study habits, and socio-demographic factors.

- To evaluate the quality of clusters using silhouette score and visualize them through plots and heatmaps, thereby illustrating how unsupervised learning can uncover hidden patterns in education data.

- To demonstrate the relevance of machine learning in the education domain, showing how such techniques can assist educators in understanding student behavior, identifying at-risk groups, and designing better interventions.

Methodology

This project was carried out using the Student Performance dataset consisting of 395 student records with 33 features that capture demographic, academic, and social details. The primary aim was to uncover hidden patterns among students using unsupervised learning techniques.

1. Data Collection

The dataset was obtained from the UCI Machine Learning Repository (Student Performance Data Set).

It included information such as age, gender, parental education, study time, family support, absences, and academic grades.

No survey was conducted as part of this project. Hence, no questionnaire or sampling method was required—the dataset served as the sole source of data.

2. Data Exploration (EDA)

Checked dataset shape (395 × 33) and column details.

Verified missing values and data types.

Summarized dataset using statistical measures (mean, median, mode, min, max, quartiles).

Visualized distributions of important features (e.g., student grades, absences, alcohol consumption).

Plotted histograms, boxplots, and correlation heatmaps to identify trends and relationships.

3. Data Preprocessing

Categorical features (such as school, gender, address type, parental jobs) were label encoded using LabelEncoder.Numerical features were scaled using StandardScaler to bring them to a common range.Converted cleaned dataset into a numeric-only dataset (17 columns).Encoded binary columns like internet (yes/no → 1/0).

4. Visualization

Scatter plots of PCA-reduced clusters colored by cluster labels.

Heatmaps to analyze correlations among features.

Cluster centroids visualized with Matplotlib and OpenCV for comparison.

5. Tools and Technologies Used

Python – Programming language

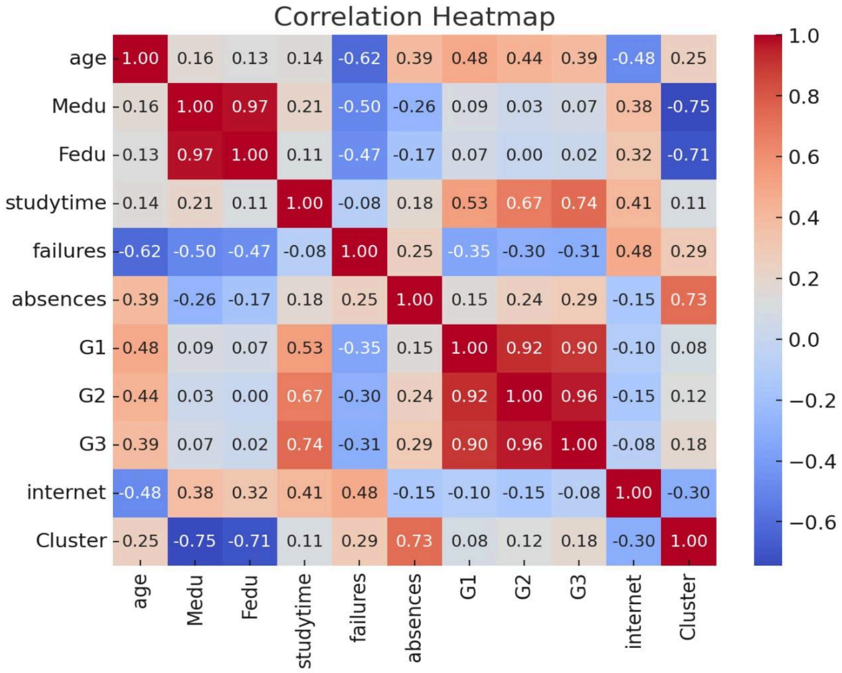
Pandas, NumPy – Data handling and preprocessing

Matplotlib, Seaborn – Data visualization

Scikit-learn – PCA, K-Means clustering, Label Encoding, Silhouette Score

OpenCV – Visualization of clusters

Data Analysis and Results



Descriptive Analysis :

Summary statistics for all features (age, education, study time, failures, absences, grades, internet usage) were calculated.

Histograms show the distribution of final grade (G3), indicating most students fall between 6 and 12.

Correlation heatmap highlights strong correlations between G1, G2, and G3 (student grades).

Conclusion

This project successfully applied unsupervised learning techniques with dimensionality reduction on the Student Performance dataset to uncover hidden patterns in student behavior and academic outcomes. Through exploratory data analysis (EDA), it was observed that features such as parental education, study time, and alcohol consumption showed correlations with academic performance, while grades (G1, G2, G3) were strongly interrelated.

APPENDICES

1. UCI Machine Learning Repository: Student Performance Data Set – https://archive.ics.uci.edu/ml/datasets/Student+Performance
2. Scikit-learn Documentation – https://scikit-learn.org/stable/
3. Seaborn Documentation – https://seaborn.pydata.org/
4. GitHubRepository: https://github.com/yourusername/student-performance-clustering
5. Project Report (this document) – (Upload to Google Drive or GitHub and provide link)
6. Dataset (Student Performance CSV) – UCI Repository Link
7. Presentation Slides – (Upload to Google Drive/GitHub and provide link)