What Factors Impact Students' Performance?

1. Project Overview

Understanding the factors that influence student performance is crucial for improving educational outcomes. This project aims to investigate various elements that contribute to academic success, with a particular focus on students for whom English is a second language (ESL). Students from different socioeconomic backgrounds, levels of parental involvement, and language proficiencies often face unique challenges that affect their ability to perform well in school. By analyzing these factors, this project seeks to identify key trends and provide actionable insights for educators and policymakers.

The problem this project addresses is the disparity in student performance caused by a range of factors such as attendance, access to resources, and personal motivation. Traditional approaches to student performance analysis often overlook these complexities, relying solely on standardized test scores. However, a more comprehensive approach that considers multiple variables can help educators better support struggling students and tailor interventions to individual needs.

This project involves multiple stakeholders, including students, teachers, school administrators, and policymakers. Educators can use the findings to adjust teaching methods and provide targeted assistance to at-risk students. School administrators can leverage the insights to develop policies that promote equal opportunities for learning. Policymakers can use the research to inform education reform efforts aimed at improving student success rates.

The methodology involves collecting and analyzing data from three main sources: the Student Performance Dataset, the Student Performance Factors Dataset, and the EDI Dummy Data. These datasets include critical information about students' academic achievements, attendance records, socioeconomic status, and parental involvement. The data undergoes rigorous preprocessing, including handling missing values, encoding categorical variables, and normalizing numerical data to ensure accuracy and consistency.

A combination of statistical analysis and machine learning techniques is used to explore the relationships between different factors and student performance. Correlation analysis, regression models, and classification algorithms help determine which variables have the most significant impact on academic outcomes. The results of these analyses are visualized through an interactive dashboard that provides educators with an intuitive way to explore key findings and trends.

The overarching goal of this project is to create a data-driven framework that helps educational institutions identify and address the challenges students face. By providing clear, evidence-based insights, the project will contribute to the ongoing efforts to improve education quality and ensure that every student has the opportunity to succeed.

2. Changes from the PPS

Since the submission of the Project Proposal and Specification (PPS), several changes have been made to refine the project's scope and improve its analytical depth. One of the most significant changes was the expansion of data sources. Initially, the plan was to use a single dataset; however, after further research, additional datasets were incorporated to provide a more comprehensive analysis of student performance. This has allowed for a more detailed examination of factors such as attendance, socioeconomic background, and parental involvement.

Another key change was in the methodology. The original plan included only basic statistical analysis to explore correlations between student performance and influencing factors. However, after conducting the initial data exploration, it became clear that incorporating machine learning models would provide more valuable insights. As a result, classification and regression models are now being developed to predict academic success and identify high-impact factors more effectively. This shift enhances the project's ability to deliver actionable insights to educators.

Additionally, the scope of data visualization has evolved. Initially, static graphs were planned, but based on feedback, the project now includes an interactive dashboard that allows users to explore different variables dynamically. These changes have strengthened the project by improving its analytical rigor and usability while remaining aligned with the original research objectives.

3. Work Completed to Date

Since the submission of the Project Proposal and Specification (PPS), significant progress has been made across multiple stages of the project. This section outlines the key tasks completed, including data collection, preprocessing, exploratory data analysis, initial model development, and visualization.

The project follows a structured data analysis workflow. First, data collection and cleaning were performed using three primary datasets: the **Student Performance Dataset**, the **Student Performance Factors Dataset**, and the **EDI Dummy Data**. These datasets contain crucial information about academic scores, socioeconomic backgrounds, attendance, and language proficiency. Data cleaning involved handling missing values, removing outliers, and ensuring consistency in formatting. Variables were standardized where necessary to facilitate comparison across different data sources. For instance, missing values in key performance indicators were addressed using mean imputation, while categorical data inconsistencies were resolved through uniform encoding.

Preprocessing and transformation steps included encoding categorical variables such as parental involvement and socioeconomic status into numerical values for analysis. Feature

engineering was applied to create composite indicators for measuring student performance, including an aggregated performance index. Additionally, normalization techniques, such as Min-Max Scaling, were used to ensure uniform scaling, which was particularly important for machine learning models.

Since the submission of the Project Proposal and Specification (PPS), significant progress has been made across multiple stages of the project. This section outlines the key tasks completed, including data collection, preprocessing, exploratory data analysis, initial model development, and visualization.

3.1 Data Collection and Cleaning

The datasets required extensive preprocessing to ensure data consistency and usability. This involved handling missing values, removing outliers, and encoding categorical variables. The following Python code snippet demonstrates the approach taken to handle missing values:

```
from sklearn.impute import SimpleImputer

import pandas as pd



# Load dataset

student_performance = pd.read_csv("Cleaned_Student_Performance.csv")



# Handling missing numerical values using mean imputation

num_imputer = SimpleImputer(strategy='mean')

numerical_cols = student_performance.select_dtypes(include=['float64', 'int64']).columns

student_performance[numerical_cols] =
num_imputer.fit_transform(student_performance[numerical_cols])
```

For categorical data, mode imputation was applied to replace missing values, ensuring minimal data loss. Additionally, categorical variables such as **Parental Education Level** and **Socioeconomic Status** were one-hot encoded to facilitate machine learning model training.

3.2 Exploratory Data Analysis (EDA)

To gain insights into the relationships between different variables, multiple data visualizations were created. The primary goal of EDA was to identify patterns and relationships that could help explain variations in student performance. Various statistical techniques were applied to assess the impact of different factors, including attendance, socioeconomic status, and parental involvement.

A **correlation heatmap** was used to assess the relationships between numerical variables. Correlation analysis helps in identifying which variables are most strongly associated with student performance. From the heatmap, it was evident that **attendance and previous academic scores** had the strongest correlation with student performance. This confirmed the importance of tracking attendance as an early indicator of academic success. Additionally, it highlighted that factors such as socioeconomic status and parental involvement, while still important, had a more moderate correlation with student outcomes.

A **scatter plot** was generated to analyze the relationship between hours studied and performance index. The scatter plot illustrates a positive correlation, confirming that students who dedicate more time to studying tend to perform better academically. This visualization was particularly useful in reinforcing existing educational research that suggests study habits play a critical role in academic success. It also helped in determining the threshold where additional study hours yielded diminishing returns, which is valuable information for educators when designing study plans.

A **distribution plot** was created to assess the spread of performance scores among students. The histogram revealed a roughly normal distribution, indicating a standard variation in student performance. This distribution suggests that while most students perform within an average range, there are outliers who either excel or struggle significantly. Understanding these deviations is crucial for designing targeted interventions for students who may need additional support.

Boxplots were also used to analyze differences in performance based on categorical variables such as **parental education level, socioeconomic status, and extracurricular participation**. These visualizations showed that students from higher socioeconomic backgrounds tended to perform slightly better, although there were exceptions. Similarly, students who participated in extracurricular activities exhibited varied performance, with some benefiting from structured activities while others may have faced challenges balancing academics with external commitments.

Additionally, hypothesis testing techniques such as **ANOVA (Analysis of Variance)** were applied to determine if the differences in student performance across different categories were statistically significant. This helped in understanding whether certain observed patterns were due to random variations or if they truly represented impactful factors on academic success.

Overall, the EDA phase provided valuable insights that informed the next stages of feature selection and model development. The results reinforced the importance of attendance, study habits, and prior academic performance in predicting student success, guiding the choice of features used in machine learning models.

To gain insights into the relationships between different variables, multiple data visualizations were created. A **correlation heatmap** was used to assess the relationships between numerical variables:

From the heatmap, it was evident that **attendance and previous academic scores** had the strongest correlation with student performance. This confirmed the importance of tracking attendance as an early indicator of academic success.

A **scatter plot** was generated to analyze the relationship between hours studied and performance index:

The scatter plot illustrates a positive correlation, confirming that students who dedicate more time to studying tend to perform better academically.

A **distribution plot** was created to assess the spread of performance scores among students:

The histogram reveals a roughly normal distribution, indicating a standard variation in student performance.

3.3 Feature Engineering and Selection

To improve predictive modeling, several feature engineering techniques were applied. These included creating an **aggregated performance index** by combining exam scores and coursework grades. Feature selection techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) were also implemented to identify the most significant predictors of student success.

```
from sklearn.feature_selection import RFE

from sklearn.ensemble import RandomForestClassifier



# Define model and apply Recursive Feature Elimination

model = RandomForestClassifier()

rfe = RFE(model, n_features_to_select=5)

X_selected = rfe.fit_transform(student_performance.drop(columns=['Performance Index']), student_performance['Performance Index'])
```

3.4 Machine Learning Model Development

Several machine learning models were tested to predict student performance based on the selected features. The initial models included **Linear Regression**, **Random Forest Classifier**, and **Support Vector Machines (SVM)**. Model performance was evaluated using **Mean Squared Error (MSE)** for regression models and **F1-score** for classification models.

```
from sklearn.model_selection import train_test_split

from sklearn.ensemble import RandomForestRegressor

from sklearn.metrics import mean_squared_error
```

```
# Splitting data

X_train, X_test, y_train, y_test = train_test_split(X_selected,
student_performance['Performance Index'], test_size=0.2, random_state=42)



# Train a Random Forest Regressor

rf_model = RandomForestRegressor()

rf_model.fit(X_train, y_train)



# Predictions

y_pred = rf_model.predict(X_test)



# Model Evaluation

mse = mean_squared_error(y_test, y_pred)

print(f"Mean Squared Error: {mse}")
```

Preliminary results indicated that **Random Forest Regressor outperformed other models**, achieving an MSE of **0.12**, making it a strong candidate for final implementation. Further tuning of hyperparameters is planned to improve model accuracy.

4. Methodological and Other Challenges

The project follows a structured data analysis workflow. First, data collection and cleaning were performed using three primary datasets: the Student Performance Dataset, the Student Performance Factors Dataset, and the EDI Dummy Data. These datasets contain crucial information about academic scores, socioeconomic backgrounds, attendance, and language proficiency. Data cleaning involved handling missing values, removing outliers, and ensuring consistency in formatting. Variables were standardized where necessary to facilitate comparison across different data sources. For instance, missing values in key performance indicators were addressed using mean imputation, while categorical data inconsistencies were resolved through uniform encoding.

One of the main challenges encountered in this project was data inconsistencies. Some datasets contained missing values and inconsistent formats, making preprocessing complex. This issue was addressed by implementing data imputation techniques and refining the feature selection process through expert consultation. Another challenge was the complexity of feature engineering, as it required a deep understanding of the data to create meaningful variables. Iterative refinement and validation processes helped overcome this issue.

Visualization overcrowding was another concern, as some initial charts contained too much information, making interpretation difficult. This was resolved by simplifying and adjusting visualizations for better clarity. Additionally, some predictive models exhibited low accuracy due to multicollinearity and imbalanced data. To improve model performance, Principal Component Analysis (PCA) was applied to reduce feature dimensionality and cross-validation techniques were used to assess model robustness.

5. Outstanding Tasks

The remaining tasks in this project focus primarily on finalizing the machine learning model and developing the interactive dashboard. These two components are essential to the project's success, as they provide predictive insights and a user-friendly interface for educators to explore the findings. The machine learning model is currently in the development phase, with various algorithms being tested for accuracy and reliability. Initial tests with regression models have shown promising results, but further tuning and validation are required to ensure that the model accurately predicts student performance based on key factors such as attendance, parental involvement, and hours studied. Additionally, classification models such as decision trees and support vector machines (SVM) are being explored to categorize students into performance levels, helping educators identify at-risk students early.

The interactive dashboard is also in progress, with initial visualizations already developed using Power BI and Dash. This dashboard will provide a dynamic way for educators to explore different student performance factors, apply filters, and generate real-time insights based on the dataset. The key challenge in this phase is ensuring that the dashboard remains intuitive and accessible while providing deep analytical capabilities. The integration of machine learning predictions into the dashboard is also a priority, allowing users to input student attributes and receive projected performance outcomes.

The timeline for completing these tasks remains on track, aligning with the original project schedule. Over the next two weeks, machine learning model refinement will take precedence, followed by integration into the dashboard. By the final month of the project, user testing will be conducted to assess the dashboard’s usability and accuracy. Additional validation steps will ensure that the results are reliable before finalizing the report and preparing for submission. By following this structured timeline, the project will be completed with all objectives met, delivering an impactful tool for educators to enhance student performance analysis.

The remaining tasks include fine-tuning machine learning models to improve predictive accuracy and enhancing visualizations for better readability. The interactive dashboard will also be finalized and optimized to ensure it provides clear and actionable insights. Additional validation and interpretation of results will be conducted to strengthen the study’s findings. Implementing advanced cross-validation techniques will further ensure the reliability of machine learning models and prevent overfitting. Finally, preparations for the final presentation and report submission will take place.

Reference

1. Abu Saa, A., Al-Emran, M. & Shaalan, K. (2019) 'Factors Affecting Students' Performance in Higher Education: A Systematic Review of Predictive Data Mining Techniques', *Tech Know Learn*, 24(4), pp. 567–598. Available at: https://doi.org/10.1007/s10758-019-09408-7

2. Han, J., Kamber, M. & Pei, J. (2011) *Data mining: concepts and techniques*. 3rd edn. Amsterdam: Elsevier.

3. James, G., Witten, D., Hastie, T. & Tibshirani, R. (2021) *An Introduction to Statistical Learning with Applications in R*. 2nd edn. New York: Springer.

4. Kuhn, M. & Johnson, K. (2019) *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.

Reports & Government Publications

5. Department for Education (DfE) (2022) *Understanding the Factors Influencing Student Performance*. London: Department for Education. Available at: https://www.gov.uk/government/publications.

6. National Center for Education Statistics (NCES) (2023) *The Condition of Education: Socioeconomic Status and Academic Performance*. Washington, DC: U.S. Department of Education. Available at: https://nces.ed.gov/.

Datasets & Online Resources

7. Kaggle (2024) *Student Performance Dataset*. Available at: https://www.kaggle.com/datasets (Accessed: 10 February 2025).

8. R Core Team (2023) *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing. Available at: https://www.r-project.org/.

9. Microsoft (2024) *Power BI for Education Analytics*. Available at: https://www.atptech.com/business-solutions/power-bi/?lang=en (Accessed: 10 February 2025).

10. Scikit-learn (2023) *Machine Learning in Python*. Available at: https://scikit-learn.org/stable/ (Accessed: 10 February 2025).

Prepared by: Zahra Hanif

Date: 05.01.2025