Machine learning based Innovative System to Forecast Secondary School Students' Academic Accomplishments

MS.SHRAVANI P.PAWAR, DR.SHEETAL DESHMUKH, DR.PRIYA CHANDRAN

Abstract: *Predicting secondary school students' academic accomplishments is crucial for early intervention and personalized learning strategies. This study develops a machine learning-based system to forecast student performance, including grade and percentage prediction, while analyzing the impact of various socio-economic, educational, personal, and technological factors. The dataset was collected through a structured survey, incorporating aspects such as family income, parental education, access to private tutoring, school infrastructure, learning environment, mental health, career guidance, geographic constraints, government policies, and digital literacy. Pre-processing was done using five machine learning algorithms: Random Forest, Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Gradient BoostingTo assess model performance, various evaluation metrics, such as accuracy and root mean squared error, were utilized. The findings suggest that machine learning methods are capable of accurately forecasting student performance, with the Random Forest algorithm demonstrating the greatest level of precision. This study lays the groundwork for AI-based educational resources aimed at recognizing students who are at risk and facilitating focused interventions.*

Keywords: Machine Learning, Academic Performance Prediction, Student Achievement Forecasting, Random Forest, Education Analytics, Socio-Economic Impact.

*Ms.Shravani P.Pawar (pawarshravani81@gmail.com) Assistant Professor ,BVIMIT,CBD Belapur,Navi Mumbai.
Dr.Sheetal Deshmukh (svdeshmukh2727@gmail.com) Assistant Professor ,Bharati Vidyapeeth's YMIM,Karad.
Dr.Priya Chandran (priyaci2005@gmail.com) Associate Professor ,BVIMIT,CBD Belapur,Navi Mumbai.
* Corresponding Author: Ms.Shravani P.Pawar

Introduction

Education is a fundamental pillar of human development, shaping an individual's career trajectory, social mobility, and overall quality of life. Academic performance, particularly in secondary school, plays a crucial role in determining higher education opportunities and professional success. However, student achievements are influenced by a multitude of factors that extend beyond classroom learning, including socio-economic background, parental support, school infrastructure, psychological well-being, and access to educational resources. This study utilizes machine learning methods to forecast the academic achievements of secondary school students, encompassing their grades and percentage scores, while concurrently examining the influence of diverse socio-economic and educational variables. To build a robust predictive framework, a structured survey was conducted to collect comprehensive data on students' academic and non-academic attributes. The study encompasses key parameters such as family income, parental education, availability of private tutoring, school infrastructure, teacher qualifications, access to study materials, learning environment, student motivation, mental health, career guidance, and digital literacy. Additionally, geographic constraints, government educational policies, and technological accessibility were considered to ensure a holistic assessment of the elements influencing student performance.

Methodology

This research follows a structured and data-driven approach to develop a predictive system for forecasting secondary school students' academic accomplishments using machine learning techniques. The methodology is designed to ensure comprehensive data collection, preprocessing, model selection, training, evaluation, and analysis, allowing for accurate predictions and insightful interpretations of academic performance. The study begins with data collection, where a combination of student academic records and structured survey responses is gathered to create a rich dataset. The dataset encompasses multiple key indicators that influence student performance, including subject-wise marks total score, percentage, and final grade. To ensure a holistic analysis, behavioral factors such as attendance, study hours, class participation, and parental support are incorporated. Additionally, socio-economic factors including family income, parental education, access to private tutoring, and availability of study materials are considered, as these elements have a profound impact on a student's learning environment. Psychological and health-related factors such as student motivation, mental well-being, stress levels, and access to career guidance play a significant role in academic outcomes and are thus included in the dataset.

Furthermore, school infrastructure and learning environment parameters, such as availability of trained teachers, school facilities (libraries, laboratories, and computer access), and the medium of instruction, are examined. Geographic constraints, transportation availability, government education policies, access to scholarships, and technological resources like e-learning platforms are also incorporated to understand their effect on student performance. The collected data is structured and stored systematically, ensuring consistency, completeness, and usability for further analysis. Following data collection, data preprocessing is conducted to refine the dataset and prepare it for model training. Missing values are handled through imputation techniques, while numerical features are normalized to ensure uniformity in scale. Categorical variables are encoded to facilitate machine learning model processing. Techniques for feature selection, including correlation analysis and recursive feature elimination, are utilized to determine the most important attributes that influence academic performance. This step is crucial in enhancing model efficiency and preventing overfitting.

The process of selecting and training the model encompasses the application of five machine learning algorithms: K-Nearest Neighbors (KNN), Decision Tree Random Forest, Support Vector Machine (SVM,), and Gradient Boosting. These algorithms are chosen based on their ability to handle structured educational data, interpret complex relationships between multiple features, and provide reliable classification and regression outputs. Each model undergoes hyperparameter tuning using grid search and cross-validation techniques to optimize predictive accuracy.

Classification models are assessed using metrics including accuracy, precision, recall, and F1-score, while regression models, which forecast percentage scores, are evaluated based on mean squared error (MSE), mean absolute error (MAE), and R-squared (R^2) values. The final stage of the methodology involves result interpretation and impact analysis.

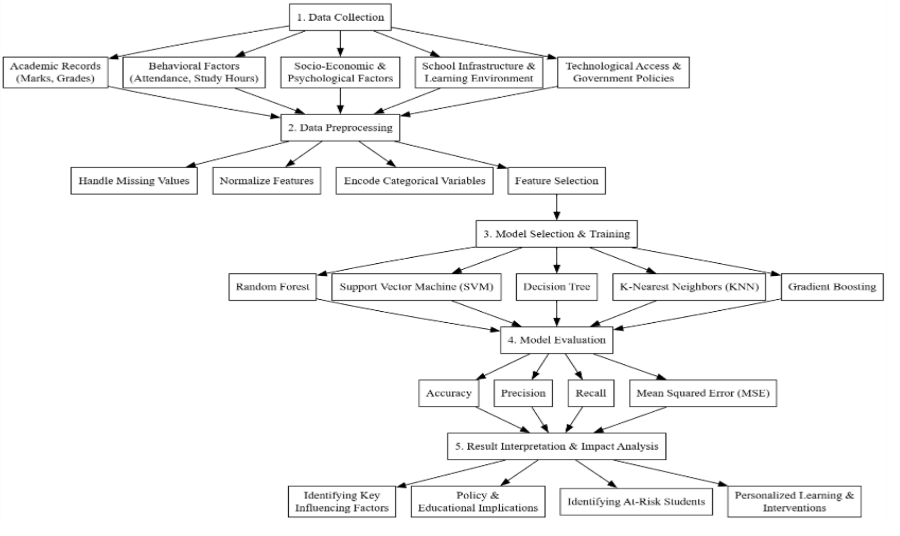

Figure 1.1: Methodology

Data Analysis

The data analysis phase is essential for understanding the factors impacting students' academic achievements. It includes various attributes like academic scores and behavioral indicators, alongside socio-economic elements such as family income and parental education. Psychological factors, including motivation and exam-related stress, are also considered to evaluate their effects on academic performance.

Exploratory Data Analysis (EDA) techniques such as statistical summaries, histograms, box plots, and correlation heatmaps are employed to uncover underlying patterns and dependencies among the variables.

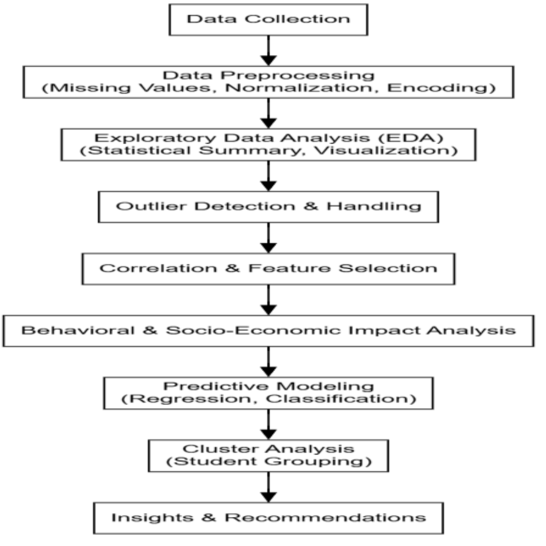

Figure 1.2: Analysis System Architecture

Data Overview

The research dataset identifies critical factors influencing student academic performance, including academic, behavioral, socio-economic, and psychological elements. It emphasizes the significance of school infrastructure, such as trained teachers and digital resources, alongside psychological influences like motivation and mental health. Geographic factors like distance to school also affect attendance, while government education policies and financial aid for low-income students are crucial for accessibility.

Table 1.1: Data Overview

	Student Name	Eng	Hindi	Marathi	Maths-1	Maths-2	Sci-1	Sci-2	His	Geo	Total	Percentage	Grade	Attendance
0	Daki Shreya 36	14	2.0	19	16.0	1.0	0	2.0	5	130	72	45.000	C-1	73
1	Patil Shreya 70	20	13.0	5	19.0	7.0	2	10.0	1	19.0	96	60.000	B-1	74
2	Daki Shreya 16	20	3.0	8	4.0	18.0	9	14.0	5	17.0	98	61.250	B-1	79

3	Patil Shreya 44	11	0.0	10	15.0	6.0	20	14.0	10	1.0	87	54.375	B-2	60
4	Joshi Kimaya 82	6	2.0	3	7.0	6.0	0	15.0	18	18.0	75	46.875	C-1	72

STUDY HOURS	PARENTAL SUPPORT	CLASS PATICIPATION	SOCIO-ECONOMIC STATUS	HEALTH	MOTIVATION
9	Low	Low	Lower	Poor	Low
5	Moderate	Moderate	Middle	Average	Medium
14	Moderate	Moderate	Middle	Average	Medium
7	Moderate	Moderate	Lower	Average	Medium
7	Low	Low	Lower	Poor	Low

Data Pre-Processing

Data pre-processing is an essential phase that guarantees the dataset is organized, free of errors, and appropriate for machine learning algorithms..The collected dataset contains various attributes, including academic scores, behavioral factors, socio-economic details, and psychological indicators. Before applying predictive algorithms, the data undergoes several pre-processing steps to enhance accuracy and reliability.

Table 1.2: Summary of Data Columns

#	Column name`	Non-Null Count	Data type
0	ENG	910	int64
1	HINDI	910	float64
2	MARATHI	910	int64
3	MATHS-1	910	float64
4	MATHS-2	910	float64
5	SCI-1	910	int64
6	SCI-2	910	float64
7	HIS	910	int64
8	GEO	910	float64
9	TOTAL	910	int64
10	PERCENTAGE	910	float64
11	GRADE	910	object
12	ATTENDANCE	910	int64
13	STUDY HOURS	910	int64
14	PARENTAL SUPPORT	910	object
15	CLASS PARTICIPATION	910	object
16	SOCIO-ECONOMIC STATUS	910	object
17	HEALTH	910	object
18	MOTIVATION	910	object

Third, data normalization and standardization are done to bring all numerical variables to a common scale. Fourth, encoding categorical variables is necessary to change non-numeric data into a machine-readable format. Fifth, feature selection and extraction techniques are applied to determine the key factors influencing student performance. Correlation analysis is conducted to check dependencies between academic performance and external factors, and feature importance scores from machine learning models help refine the selection. Finally, the processed dataset is divided into training and testing datasets to assess the performance of the model. The dataset is generally partitioned in an 80-20 or 70-30 ratio to guarantee effective model generalization. This organized pre-processing method ensures that the data is clean, free from bias, and adequately prepared for predictions of academic performance using machine learning.

Preliminary Analysis and Descriptive Statistics

Descriptive statistics are essential for analyzing student performance, highlighting distribution, central tendency, and variability via measures like mean, median, and mode. Standard deviation and variance reflect score spread, with high variance indicating diverse capabilities among students and low variance indicating consistency. Performance analysis by subject reveals areas of strength and difficulty, with a notable link between attendance, study hours, and achievement. Psychological factors, such as motivation and stress levels, significantly affect academic results. These insights emphasize the need for career counseling and mental health support, aiding in the development of predictive models and policy enhancements to boost academic success across various student demographics.

Table 1.3: Statistical Overview

	ENG	HINDI	MARATHI	MATHS-1	MATHS-2	SCI-1	SCI-2	HIS	GEO
Count	910.000000	910.000000	910.000000	910.000000	910.000000	910.000000	910.000000	910.000000	910.000000
Mean	9.497802	9.494505	10.184615	10.103297	9.964835	9.816484	9.806593	9.972527	9.768132
Std	5.956156	6.145600	5.991467	6.078356	5.875391	6.213328	6.207271	5.916016	5.983850
Min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	5.000000	4.000000	5.000000	5.000000	5.000000	4.000000	4.000000	5.000000	5.000000
50%	9.000000	9.000000	11.000000	10.000000	10.000000	10.000000	10.000000	10.000000	9.000000
75%	15.000000	15.000000	15.000000	15.000000	15.000000	15.000000	16.000000	15.000000	15.000000

max	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000	20.000000

Data Visualization

In order to achieve a more profound comprehension of the data and its fundamental relationships, a collection of visualizations was developed. These graphical representations, which encompass scatter plots, line charts, and correlation matrices, offered a lucid and intuitive perspective on the trends, distributions, and correlations among the different features within the dataset.

Correlation Matrix.

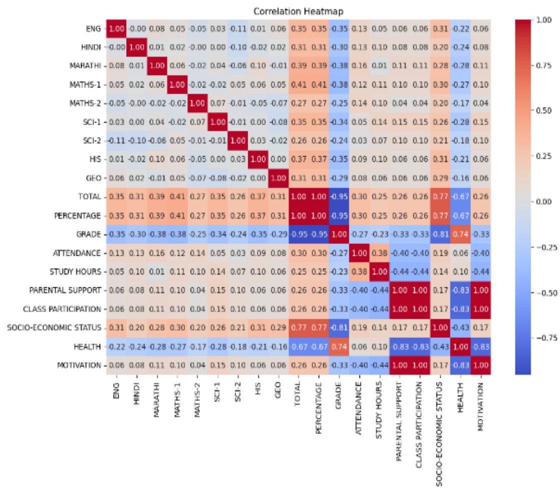

Fig 1.3: Correlation Matrix

Percentage Distribution

The percentage distribution visualization provides a clear understanding of how students' academic performance is spread across the dataset.

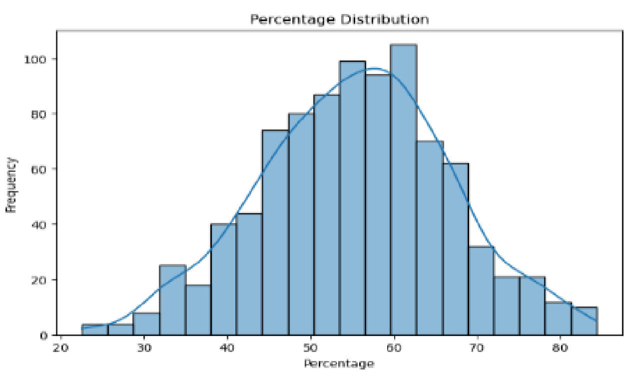

Fig 1.4: Pair Plot

Count Plot

The color hue, based on grades, provides an additional layer of insight into how different grades are distributed across attendance levels.

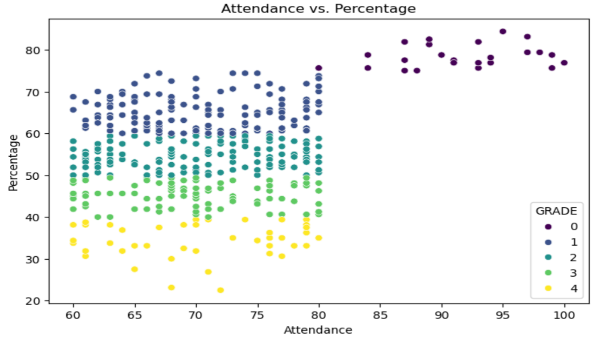

Fig 1.5: Visualization of Attendance Vs. Percentage

Box Plot

The box plot visualization of Study Hours by Grade provides insights into how students' study habits vary across different grades.

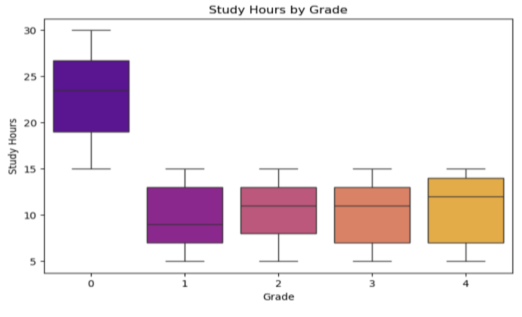

Fig 1.6: Visualization of Study Hours by Grade

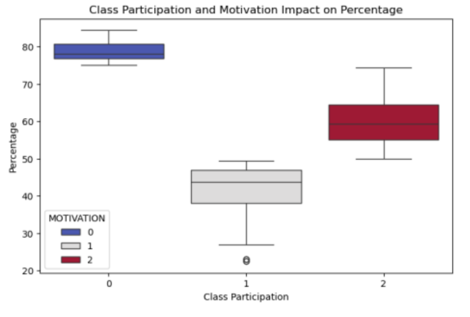

Fig 1.6: Visualization of Class Participation

Categorical Variable Distribution Analysis

Figure 1.7 illustrates the distribution of key non-numeric factors influencing student performance through bar charts. These include grade, parental support, socio-economic status, health, motivation, and class participation, allowing for the identification of patterns and trends. Each subplot is a count plot with categories on the x-axis and student counts on the y-axis.

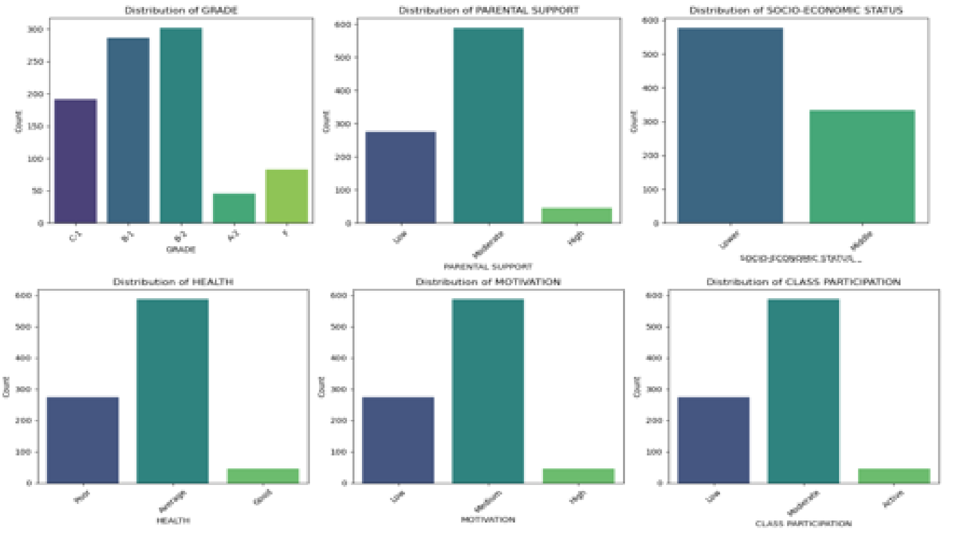

Fig 1.7: Categorical Variable Distribution Analysis

Result and Discussion

This study examines key factors influencing student academic success through data-driven insights and machine learning forecasts. Various supervised learning algorithms were employed to predict student grades using academic and socio-behavioral factors such as attendance, study hours, parental support, socio-economic status, health, and motivation. The models were evaluated on accuracy, mean absolute error, and root mean square error. Results revealed that active class participation, motivation, and study hours are critical determinants of academic performance. Higher engagement in class and parental support correlate with elevated grades, whereas decreased engagement, poor health, and insufficient study hours negatively impact performance.

A correlation analysis highlighted a strong positive relationship between study hours and academic achievement, underscoring the importance of consistent study habits. Socio-economic status also emerged as a vital factor, with students from affluent backgrounds benefiting from better resources and tutoring, leading to higher performance. Additionally, students with high motivation levels achieved superior grades, illustrating the importance of psychological factors in education. Box plot analysis further confirmed that class participation and motivation significantly affect academic outcomes. The study suggests that interventions like personalized mentoring, health initiatives, and enhanced access to digital resources can boost student performance. Schools and policymakers are encouraged to utilize these insights for developing data-driven educational strategies that address learning disparities and foster an inclusive environment.

Conclusion

The study on student academic performance through machine learning identified key factors influencing educational outcomes, revealing strong links between success and elements like study hours, class participation, motivation, and parental support. It showed that engaged students tend to perform better, while socio-economic status affects achievement levels. The research utilized machine learning models to assess the impact of these predictors, emphasizing the importance of student engagement strategies. It concluded that early interventions, personalized learning, and targeted support could enhance academic success, highlighting the role of predictive analytics in identifying at-risk students.

References

Jerich Faddar et al., “Perspectives on educational effectiveness in science and mathematics: The role of non-cognitive measures in TIMSS. Introduction to a special issue” ELSEVIER, Volume 75, 2022.

Utkarsh Verma et al., Prediction of students’ academic performance using Machine Learning Techniques, IEEE, 2020 — practical ML comparison useful for baseline models and methodology.

Adel Ismail Al-Alawi et al., Predicting Student’s Academic Performance Using Data Mining Methods: Review Paper, IEEE, 2023 — recent review summarizing methods, features and evaluation metrics.

Sagardeep Roy et al., Predicting academic performance of student using classification techniques, IEEE, 2017 — classic classification approaches and feature sets for student data.

Meizar Raka Rimadana et al., Predicting Student Academic Performance using Machine Learning and Time Management Skill Data, IEEE, 2019 — introduces behavioral/time-management features relevant to secondary students.

Bo Guo et al., Predicting Students Performance in Educational Data Mining, IEEE, 2015 — foundational EDM techniques and preprocessing advice.

Divya Thakur et al., Predicting Student's Performance using Data Mining Algorithm, IEEE, 2022 — recent experiments on common algorithms and feature engineering.

[8] Praveena Chakrapani et al., Academic Performance Prediction Using Machine Learning: A Comprehensive & Systematic Review, IEEE, 2022 — systematic overview useful for literature background and gaps.

Nehal Eleyan et al., Predicting Student Performance Using Educational Data Mining, IEEE, 2022 — applied EDM work with secondary/higher education datasets.

D. K. Arun et al., Student Academic Performance Prediction using Educational Data Mining, IEEE, 2021 — practical pipeline and evaluation best practices.

Neeta Sharma et al., Predicting Academic Performance of Students Using Machine Learning Models, IEEE, 2023 — up-to-date ML models (including ensemble/deep approaches) and comparative results.

Gomathy Suganya Ramaswami et al., Predicting Students Final Academic Performance using Feature Selection Approaches, IEEE, 2020 — feature selection techniques important for building interpretable school-level models.

Aws Khudhur et al., Students' Performance Prediction Using Machine Learning Based on Generative Adversarial Network, IEEE, 2023 — addresses small/imbalanced datasets via GAN augmentation — useful for sparse secondary-school data.

Essa Alhazmi et al., Early Predicting of Students Performance in Higher Education, IEEE, 2023 — methods for early-warning prediction (transferable to earlier secondary grades).

Roselyne Ayienda et al., Predicting Students Academic Performance using a Hybrid of Machine Learning Algorithms, IEEE, 2021 — hybrid/ensemble strategies that often improve forecasting accuracy.

Mohamed Mohsen Elsaid Khoudier et al., Prediction of student performance using machine learning techniques, IEEE, 2023 — another recent empirical study demonstrating contemporary pipelines.

Isreal M. Ogundele et al., Prediction of Student Academic Performance Based on Machine Learning Model, IEEE, 2024 — very recent work (2024) showing latest practices and possibly relevant datasets/methods for secondary schools.

Wentong Liu et al., Student Performance Prediction by LMS Data and Classroom Videos, IEEE, 2022 — multimodal data usage (LMS + videos) that can inspire innovative feature sources for richer forecasting.

Xiaochen Lai et al., Predicting the academic performance of students with GPcSAGE, IEEE, 2023 — advanced graph / GNN-style approaches for relational data (peers, classes) which can boost secondary-school forecasts.

Luis Vives et al., Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks, IEEE, 2023 — time-series / sequence modeling (LSTM) valuable when you have temporal records across semesters.