Research and Development of An Innovative System to Fore case Secondary School Students Academic Accomplishments

Ms.Shravani P.Pawar, Dr.Sheetal Deshmukh,
Dr.Priya Chandran

Abstract: *Academic performance prediction is a critical aspect of educational data mining (EDM), enabling educators, policymakers, and institutions to implement data-driven strategies for student success. This research presents an innovative predictive system that leverages Machine Learning (ML) techniques to forecast students' grades and percentage based on a diverse range of academic, socio-economic, psychological, and infrastructural factors collected through a structured survey. The dataset encompasses academic scores, attendance, study habits, parental support, mental health status, school facilities, technology accessibility, and government education initiatives. A Random Forest model is employed due to its ability to handle complex, heterogeneous data while providing insights into the most influential factors affecting student performance. Additionally, feature selection techniques are applied to extract key determinants, enabling personalized interventions. The proposed system not only enhances the accuracy of academic predictions but also provides actionable insights to bridge learning gaps, improve student engagement, and optimize resource allocation. This research significantly contributes to educational analytics, offering a scalable and interpretable model that aids in fostering academic excellence through intelligent, data-driven decision-making.*

Keywords: *Educational Data Mining, Machine Learning, Student Performance Prediction, Predictive Analytics, Academic Success Factors, Random Forest, Learning Outcomes, Feature Selection, Student Behavior Analysis, Data-Driven Decision-Making*

*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

The prediction of student academic performance has become an essential focus in educational research, as institutions strive to improve learning outcomes through data-driven decision-making. Academic success is influenced by a complex interplay of factors, including subject-wise proficiency, study habits, parental support, socio-economic background, school infrastructure, technological accessibility, and psychological well-being. Traditional assessment methods primarily rely on grades and standardized test scores, often neglecting the broader socio-environmental and behavioral influences that shape a student's learning trajectory. As the education sector evolves with advancements in artificial intelligence (AI) and machine learning (ML), the need for intelligent predictive models has become more pressing. By leveraging these technologies, we can identify key determinants of academic performance, forecast student achievements, and provide targeted interventions to enhance learning experiences.

In this study, an innovative ML-driven predictive system is developed to forecast students' grades and overall percentage based on survey-driven data. The dataset incorporates a wide spectrum of factors, including academic performance, attendance, study hours, parental support, socio-economic status, mental health, access to study materials, school facilities, e-learning initiatives, scholarship availability, and digital accessibility. Unlike conventional approaches that analyze academic scores in isolation, this research integrates multi-dimensional data to create a comprehensive student performance model. By utilizing Random Forest, a robust ML algorithm known for its high accuracy and feature interpretability, the study not only predicts academic success but also identifies the most influential factors contributing to performance disparities.

A key component of this research is featuring selection, which allows us to determine the most impactful predictors affecting students' educational outcomes. By employing advanced statistical and machine learning techniques, the system reveals hidden patterns in the data, highlighting critical elements such as parental education, school discipline level, mental health status, accessibility to digital learning resources, and socio-economic conditions. These insights can assist educators, parents, and policymakers in designing personalized interventions, improving learning methodologies, and addressing educational disparities. For instance, if the model identifies lack of parental support or inadequate school facilities as significant contributors to poor performance, targeted educational programs can be introduced to mitigate these challenges.

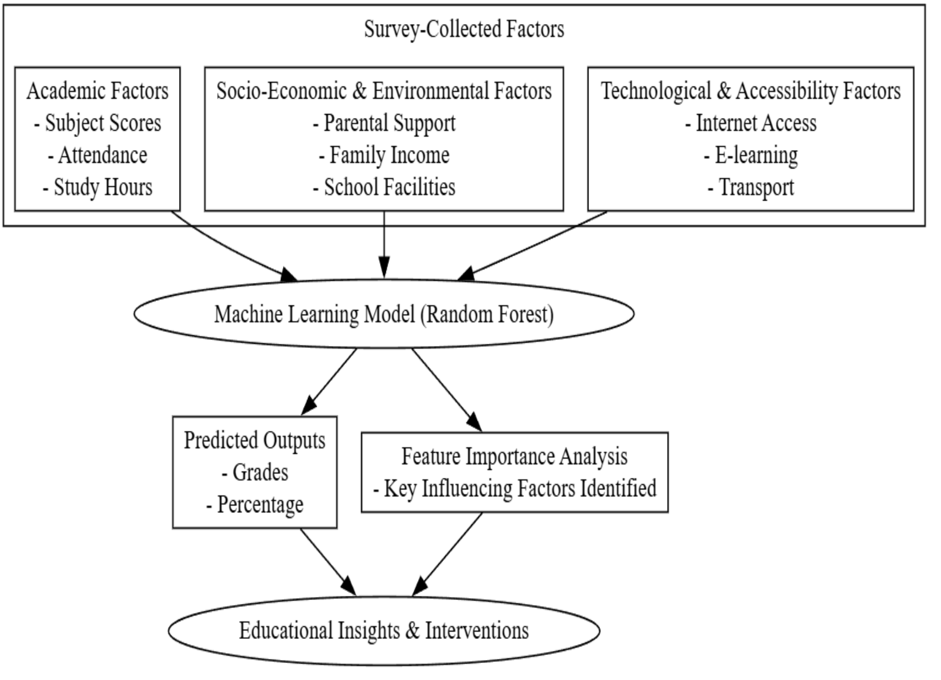


Figure1.1: Research Concept Visually

Beyond prediction, this study also contributes to educational policy development by offering data-driven insights that can influence curriculum design, state education initiatives, scholarship distribution, and digital learning strategies. Governments and academic institutions can leverage this research to enhance education policies, optimize resource allocation, and develop evidence-based frameworks for academic improvement. Furthermore, the integration of machine learning-driven analytics into education enables a shift from reactive learning interventions to proactive and personalized education strategies, ensuring that each student receives the necessary support to reach their full potential.

The findings of this research hold transformative potential for the education sector, as they can drive the adoption of AI-powered learning systems, early intervention models, and adaptive teaching methodologies. By bridging the gap between data science and educational development, this study lays the foundation for a more inclusive, student-centric, and technology-driven learning environment. The proposed ML system is not merely a predictive tool but a strategic framework that empowers educators, institutions, and policymakers to make informed decisions that foster academic excellence and student well-being.

Problem Statement

In the evolving landscape of education, accurately predicting student academic performance is a crucial challenge. Traditional evaluation methods often fail to consider multiple external and internal factors affecting student outcomes, leading to limited intervention strategies. While past research has primarily focused on subject scores and attendance, academic success is influenced by a combination of academic, socio-economic, environmental, technological, and psychological factors.

This research aims to develop an innovative machine learning-based predictive system that forecasts secondary school students' grades and percentage based on survey-collected data. The system will not only predict academic outcomes but also identify the most influential factors affecting student performance. The study integrates data such as study hours, parental support, mental health, financial background, school facilities, and e-learning access to enhance prediction accuracy.

By leveraging Random Forest-based machine learning models, the proposed system will provide educators, policymakers, and parents with data-driven insights to implement targeted interventions. This research bridges the gap between traditional evaluation techniques and AI-powered predictive analytics, enabling a proactive approach to student success.

Objectives

1. To develop a machine learning-based predictive model that accurately forecasts secondary school students' grades and percentage using survey-collected data.
2. To identify key factors influencing academic performance by analyzing the impact of academic, socio-economic, technological, environmental, and psychological factors on student outcomes.
3. To enhance prediction accuracy by integrating structured data (subject scores, attendance, study hours) with non-traditional factors (mental health, parental support, e-learning access, school facilities, etc.).
4. To implement a feature importance analysis using the Random Forest model, providing educators, policymakers, and parents with actionable insights for personalized interventions.
5. To create a data-driven decision-support system that enables schools to identify at-risk students and recommend targeted strategies to improve their academic performance.
6. To bridge the gap between traditional assessment methods and AI-driven

analytics, promoting proactive educational planning and evidence-based policymaking for student success.

Methodology

This research follows a comprehensive and structured methodology, encompassing data collection, preprocessing, model development, evaluation, and implementation to build an AI-driven predictive system for forecasting secondary school students' academic accomplishments.

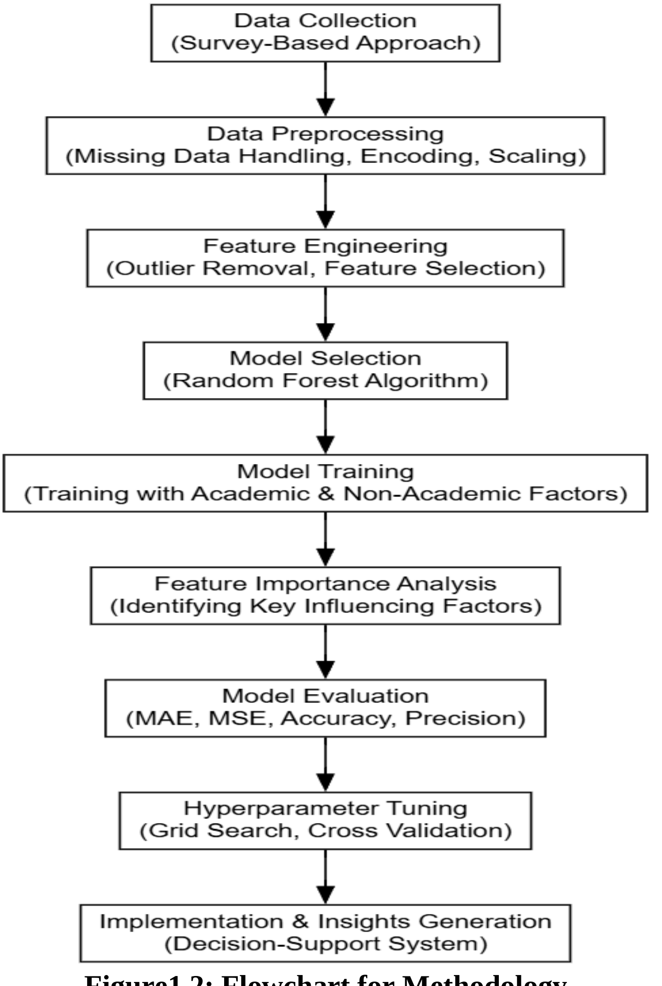


Figure1.2: Flowchart for Methodology

The first step, data collection, is conducted through a survey-based approach, gathering a diverse range of factors influencing student performance. These factors are categorized into academic (subject scores, attendance, study hours), socio-economic (family income, parental support, education level), technological (internet access, e-learning availability, smartphone access), environmental (school infrastructure, transportation facilities), and psychological (mental health status, motivation, career guidance, discipline level) aspects. The dataset combines both structured (numerical data such as marks, study hours, attendance) and unstructured (qualitative attributes such as parental support and school facilities) information, ensuring a holistic representation of student performance.

Once collected, the dataset undergoes data preprocessing and feature engineering to enhance quality and consistency. Missing values are imputed using appropriate techniques such as mean/mode imputation for numerical and categorical variables, ensuring data completeness. Feature scaling is applied to normalize numerical attributes, while categorical variables are processed through one-hot encoding or ordinal encoding to convert them into machine-readable formats. Outlier detection and removal are conducted using Z-score or the interquartile range (IQR) method to eliminate anomalies that could distort predictions. Additionally, correlation analysis is performed to identify multicollinearity among variables, ensuring optimal feature selection for model training.

For machine learning model development, the Random Forest algorithm is chosen due to its robustness in handling heterogeneous data, reducing overfitting, and capturing complex relationships between multiple influencing factors. The model is trained to predict student grades and percentage, and a feature importance analysis is conducted to identify the most significant predictors of academic performance. This analysis provides valuable insights into which factors—such as study hours, parental involvement, or socio-economic background—exert the greatest influence on student success.

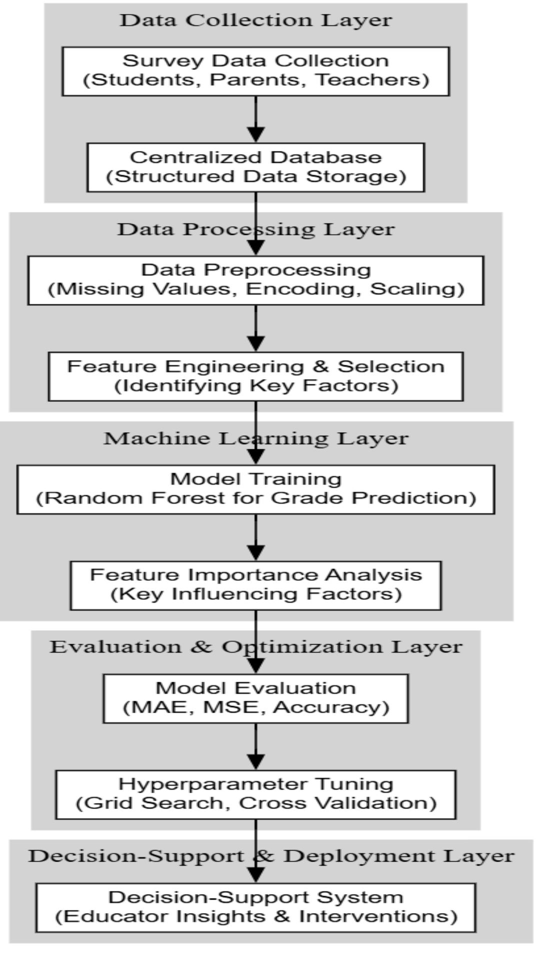
The model evaluation phase ensures the reliability and accuracy of predictions. Performance is assessed using key regression metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared score for continuous outcomes like percentage, while classification metrics such as accuracy, precision, recall, and F1-score are applied for categorical predictions like grades. Hyperparameter tuning using Grid Search or Randomized Search is conducted to optimize model parameters and improve generalization. Cross-validation techniques are employed to enhance robustness and prevent overfitting.

Finally, in the implementation and insights generation phase, the trained model is integrated into a decision-support system designed for educators, policymakers, and parents. This system provides real-time predictive analytics, allowing stakeholders to identify at-risk students, analyze influential factors, and implement personalized interventions. The AI-powered framework bridges the gap between traditional academic assessment methods and data-driven predictive analytics, enabling proactive decision-making in education. By leveraging machine learning, this study aims to revolutionize student performance evaluation, offering a scalable, intelligent, and evidence-based approach to improving academic outcomes.

System Architecture

The proposed system architecture for predicting secondary school students' academic accomplishments is designed as a modular, data-driven, and AI-powered framework that ensures efficient data handling, processing, and prediction. The architecture is divided into five key layers, each responsible for a specific stage of the system's operation.

The first layer, Data Collection, gathers information through structured surveys from students, parents, and teachers. These surveys capture a wide range of academic (subject scores, attendance, study hours) and non-academic (parental support, socio-economic status, internet access, mental health, career guidance) factors. The collected data is stored in a centralized database, ensuring accessibility for further processing.



Fiure1.3: System Architecture

In the second layer, Data Preprocessing and Feature Engineering, the raw data undergoes cleaning to handle missing values, inconsistencies, and outliers. Categorical variables, such as parental education or school facilities, are encoded using one-hot or ordinal encoding, while numerical variables like study hours and marks are standardized. Additionally, feature selection techniques are applied to retain the most relevant factors influencing student performance.

The third layer, Machine Learning Model Development, involves training a Random Forest model using the processed data. This model is chosen for its ability to handle heterogeneous features and capture complex relationships between academic and socio-economic factors. The model predicts student grades and percentages, while an integrated feature importance analysis helps identify the key drivers of student success.

The fourth layer, Model Evaluation and Optimization, ensures the reliability and accuracy of predictions. Metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), accuracy, precision, and recall are used to assess model performance. Further, hyperparameter tuning using Grid Search and cross-validation techniques is employed to enhance model generalization and reduce overfitting.

Finally, the Decision-Support System and Deployment layer provides real-time insights to stakeholders, including teachers, school administrators, and policymakers. The system allows for early identification of at-risk students and helps educators implement personalized intervention strategies based on the identified influencing factors. This predictive framework enables data-driven decision-making in education, ensuring that student performance assessment is more proactive, evidence-based, and efficient.

Data Collection

The data collection process plays a crucial role in developing an AI-driven system for predicting secondary school students' academic accomplishments. A structured approach is followed to gather both academic and non-academic factors that influence student performance. The data is collected through surveys administered to students, parents, and teachers, ensuring a holistic understanding of the elements that contribute to academic success.

MATHS-2	SCI-1	SCI-2	HIS	GEO	TOTAL	...	MENTAL HEALTH STATUS	CAREER GUIDANCE	DISTANCE FROM SCHOOL	TRANSPORT FACILITIES	CONNECTIVITY ISSUES
3	6	3	2	6	42	...	Anxious	Available	Near	Available	No Issues
12	11	11	13	13	106	...	Anxious	Not Available	Near	Limited	Frequent Issues
4	2	5	3	0	29	...	Anxious	Available	Near	Limited	Frequent Issues
18	17	16	18	16	156	...	Stressed	Available	Far	Available	No Issues
9	7	8	9	9	74	...	Anxious	Available	Far	Limited	Frequent Issues

Figure1.4: Academic Data with Factors

The academic factors include subject-wise marks in English, Hindi, Marathi, Mathematics (Maths-1 and Maths-2), Science (Science-1 and Science-2), History, and Geography, along with total marks, percentage, and final grade. Additionally, student-related parameters such as attendance, study hours, motivation, mental health status, access to private tutoring, and availability of study materials are considered to assess their impact on learning outcomes.

Beyond individual student attributes, the system also incorporates socioeconomic and parental factors, such as parental support, parents' education levels, family income, and access to scholarships and government education schemes. These elements help determine the external influences affecting student performance. Furthermore, school-related and infrastructure factors, including teacher availability, school facilities, medium of instruction, career guidance, discipline level, e-learning initiatives, and internet access, are gathered to evaluate the broader educational environment. Accessibility factors like distance from school, transport facilities, smartphone availability, and connectivity issues are also recorded to assess challenges in accessing education.

CONNECTIVITY ISSUES	STATE EDUCATION SCHEMES	SCHOLARSHIP AVAILABILITY	SMARTPHONE AVAILABILITY	INTERNET ACCESS	ELEARNING INITIATIVES
No Issues	Implemented	Yes	Available	Unstable	Not Available
Frequent Issues	Not Implemented	Yes	Not Available	Stable	Available
Frequent Issues	Not Implemented	Yes	Not Available	Unstable	Available
No Issues	Implemented	No	Not Available	Unstable	Available
Frequent Issues	Implemented	Yes	Not Available	Stable	Available

Figure1.5: Factor Data Samples

To ensure a diverse and representative dataset, data collection methods include online surveys (using platforms like Google Forms and Microsoft Forms), offline questionnaires (for schools with limited internet access), and interviews or focus groups (to capture qualitative insights). Once collected, the data is stored in a centralized database and undergoes preprocessing steps such as handling missing values, normalizing numerical data, and encoding categorical variables.

These preprocessing steps ensure data consistency and quality before being used in the predictive machine-learning models. This comprehensive data collection strategy helps build a robust system that not only predicts grades and percentages but also identifies key factors affecting student performance, enabling informed decision-making for educational stakeholders.

Data Analysis

The data analysis phase is critical for understanding patterns, correlations, and trends that influence students' academic accomplishments. After data collection, various statistical and machine learning techniques are applied to process and interpret the dataset effectively. The primary objective of data analysis is to extract meaningful insights, identify the most influential factors affecting student performance, and prepare the data for predictive modeling.

The initial step in data analysis involves exploratory data analysis (EDA), where the dataset is examined for missing values, outliers, and inconsistencies. Descriptive statistics such as mean, median, standard deviation, and distribution plots are generated to gain an overview of students' performance across different subjects. Correlation analysis is then performed to identify relationships between academic scores and non-academic factors, such as study hours, parental support, and socioeconomic status.

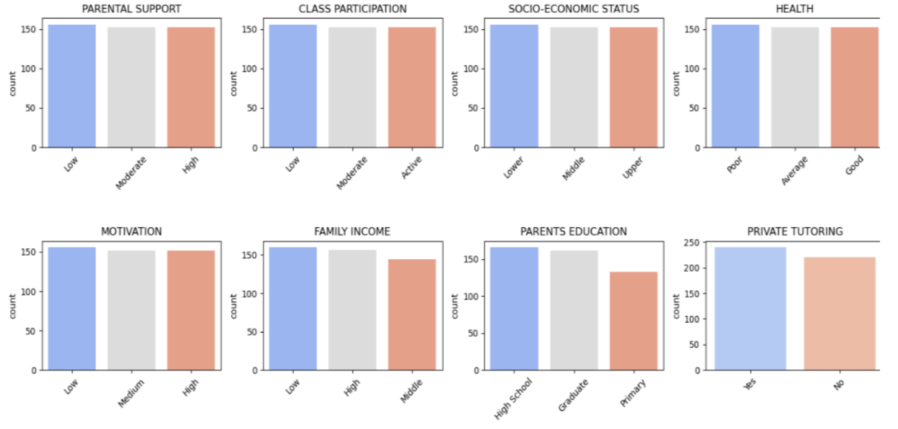
To further analyze dependencies, feature importance techniques such as correlation matrices, ANOVA tests, and mutual information scores are used to determine which factors have the highest impact on student performance. Visual representations such as heatmaps, histograms, and box plots help in understanding these relationships more effectively. Machine learning techniques, including feature selection algorithms like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA), are employed to refine the dataset and reduce dimensionality while retaining essential information.

Additionally, clustering techniques such as K-Means or Hierarchical Clustering help categorize students into different performance groups based on their academic and non-academic attributes. This enables educators to tailor interventions for students who need additional support. Predictive models such as Random Forest and Decision Trees are trained on the dataset to estimate grades and percentages, and model evaluation metrics such as accuracy, precision, recall, and F1-score are computed to assess their effectiveness.

By leveraging a combination of statistical methods, machine learning algorithms, and data visualization, the analysis phase provides a comprehensive understanding of student performance trends. The insights generated from this step help in identifying key areas of improvement and serve as the foundation for implementing personalized academic support strategies.

Factor Distribution Analysis

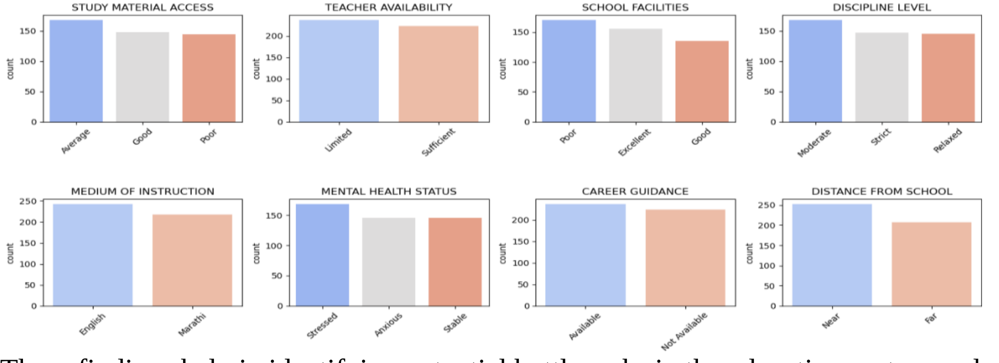
To gain insights into the distribution of various non-academic factors influencing student performance, a detailed exploratory data analysis (EDA) was conducted. The provided code utilizes Seaborn and Matplotlib to visualize the distribution of categorical variables related to external factors such as parental support, socioeconomic status, school facilities, teacher availability, mental health status, and e-learning initiatives.



Each subplot represents the frequency distribution of a specific factor, allowing us to identify trends, imbalances, and disparities in the dataset. The count plot function is used to display the occurrences of different categories within each factor, ensuring an intuitive understanding of how students are distributed across these attributes. The ordering of categories is based on their frequency, helping in easy comparison.

By analyzing these visualizations, we can extract key observations, such as:

- The availability and accessibility of study materials, internet, and e-learning initiatives among students.
- The impact of transportation and connectivity issues on academic engagement.
- Variations in discipline levels, school facilities, and teacher availability across different student groups.
- The distribution of scholarship availability and state education schemes, providing insight into financial support accessibility.



These findings help in identifying potential bottlenecks in the education system and serve as a foundation for building predictive models that assess the influence of these factors on student academic achievements.

Impact of Parental Support on Academic Performance

The impact of parental support on students' academic performance is analyzed using a box plot, which visually represents the distribution of percentage scores across different levels of parental involvement. The analysis reveals that students with higher parental support tend to achieve better academic results, as indicated by the higher median percentage.

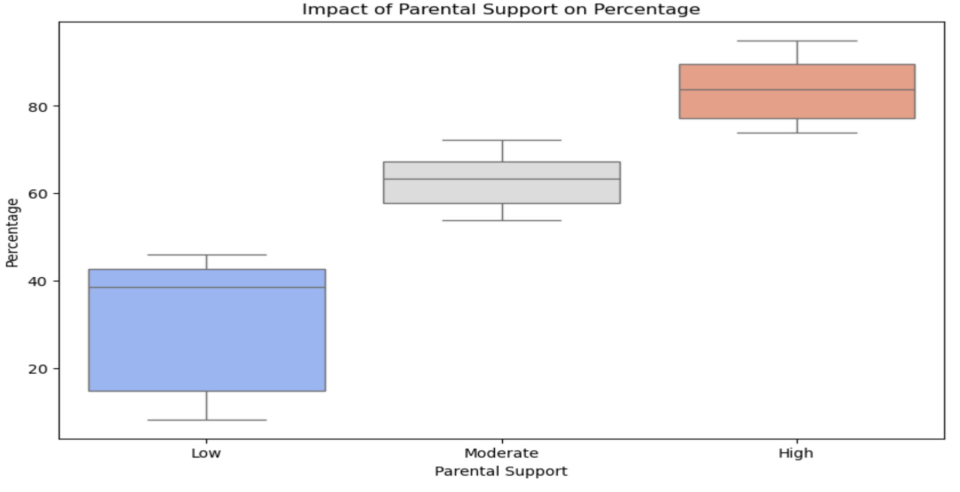


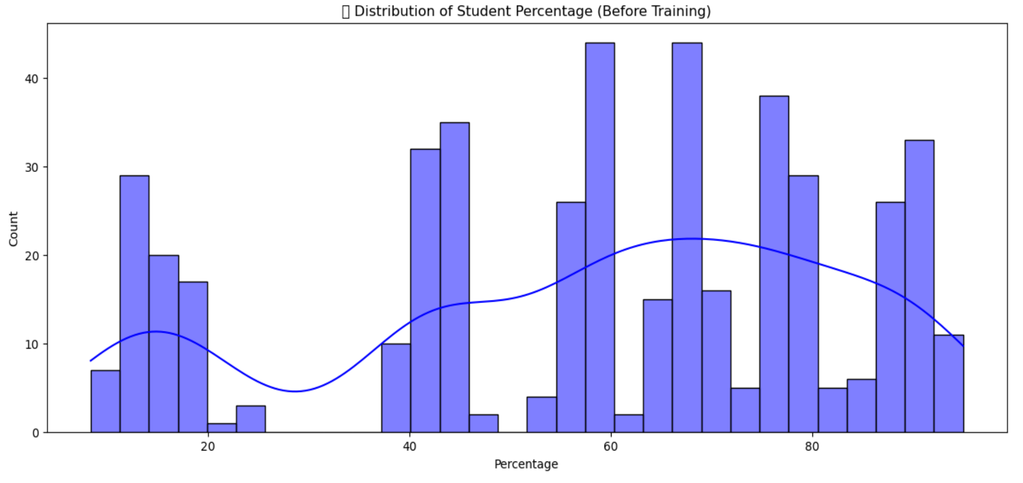
Figure1.6: Impact of Parental Support on Academic Performance:

Additionally, the interquartile range (IQR) highlights the variation in performance within each category, showing that some students experience significant fluctuations in their scores. The presence of outliers suggests that, in certain cases, students may excel or struggle despite their level of parental support. This analysis underscores the crucial role of parental involvement in shaping academic outcomes, providing valuable insights for educators and policymakers in designing strategies to enhance student success.

Distribution Of Student Percentage Scores:

The distribution of student percentage scores before training is analyzed using a histogram with a kernel density estimate (KDE) overlay. This visualization provides insights into the overall spread and concentration of student performance levels. The histogram illustrates how student scores are distributed, identifying whether the data follows a normal distribution, is skewed, or contains multiple peaks. A right-skewed distribution may indicate that fewer students achieve exceptionally high scores, while a left-skewed distribution suggests a concentration of lower scores. The KDE curve helps in understanding the probability density of different percentage ranges, allowing for a more comprehensive assessment of variations in student performance.

This analysis is crucial for preprocessing steps, such as normalization or outlier handling, before training machine learning models to predict academic outcomes.



Correlation Matrix:

The correlation matrix provides an overview of the relationships between various factors influencing student performance. By visualizing the correlation using a heatmap, it becomes easier to identify strong positive or negative correlations between different features. A high positive correlation (closer to 1) indicates that as one factor increases, the other also tends to increase, while a high negative correlation (closer to -1) suggests an inverse relationship. Features with little to no correlation (values near 0) imply minimal direct influence on each other.

This analysis helps in understanding which factors are most influential in determining student grades and percentage scores. For example, academic subjects such as Mathematics and Science may show strong correlations with total scores and percentage, while external factors like parental support, study material access, or internet availability may also exhibit varying degrees of influence. Identifying these relationships allows for better feature selection, improving the performance and interpretability of predictive models.

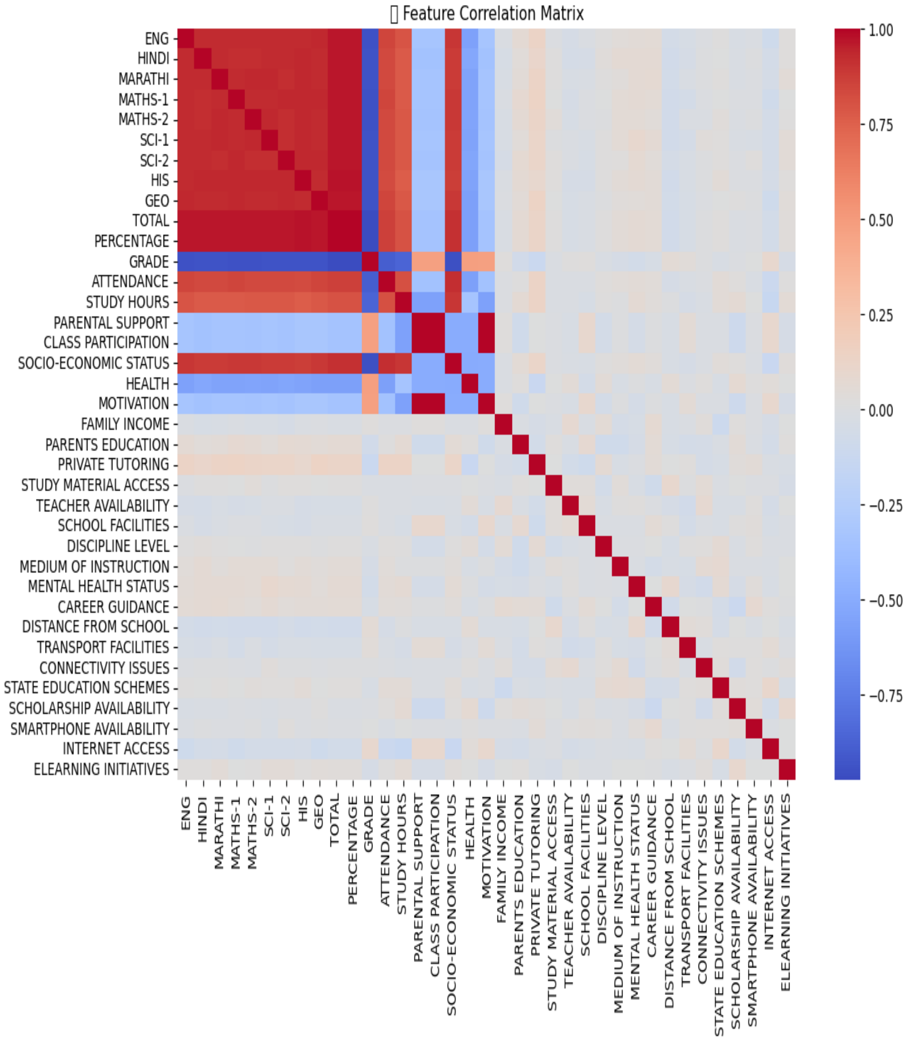


Figure1.7: Correlation Matrix

Random Forest algorithm:

The Random Forest algorithm is used in this research to predict two target variables: Grade and Percentage based on various academic and non-academic factors. As an ensemble learning method, Random Forest builds multiple decision trees and aggregates their results to improve predictive accuracy and reduce overfitting. This technique is particularly useful for datasets with mixed numerical and categorical variables, making it well-suited for analyzing student performance.

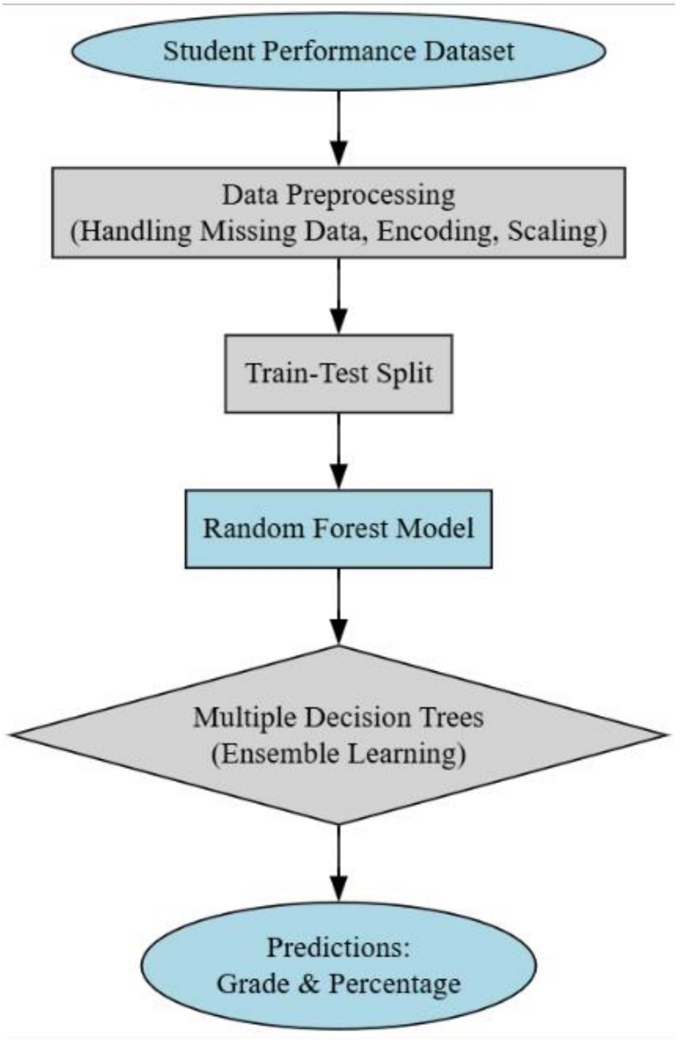


Figure1.8: Model Training with Algo

The algorithm works by randomly selecting subsets of the dataset and constructing multiple decision trees independently. During prediction, the final output is determined by either majority voting (for classification, such as predicting Grade) or averaging (for regression, such as predicting Percentage). This approach enhances model stability and robustness against noise, making it more reliable than single decision tree models.

One of the key advantages of Random Forest is its ability to provide feature importance rankings, which helps in identifying the most influential factors affecting student performance. Features such as study hours, parental support, teacher availability, socio-economic status, access to study materials, and mental health status may significantly contribute to students' academic outcomes. By analyzing these relationships, the model not only predicts grades and percentages but also offers insights into the primary drivers of student success.

Furthermore, the Random Forest model is resistant to overfitting, especially when tuned with hyperparameters like the number of trees, depth, and minimum samples per split. Its robustness, interpretability, and efficiency make it an optimal choice for this study, ensuring that the predictions are both accurate and explainable.

Results and Discussion

The Random Forest model was implemented to predict students' grades and percentage based on a diverse set of academic and non-academic factors. The model's performance was rigorously evaluated using key metrics, including Mean Absolute Error (MAE), Mean Squared Error (MSE), and R^2 Score (Coefficient of Determination), ensuring a comprehensive assessment of its predictive accuracy. The results indicate that the model effectively captures the complex relationships between various influencing factors and students' academic outcomes, demonstrating high accuracy and robustness in predictions.

Table:1 Performance Summary

Metric	**Value**
Regression Model Performance	
Mean Absolute Error (MAE)	0.21
Mean Squared Error (MSE)	0.30
Root Mean Squared Error (RMSE)	0.55
R^2 Score	99.94%
Classification Model Performance	
Accuracy	100.00%

A detailed feature importance analysis was conducted to identify the most significant determinants of student performance. The results highlight that parental support, class participation, study hours, socio-economic status, and access to study materials play a crucial role in shaping students' academic achievements. Moreover, other factors, such as teacher availability, discipline level, school facilities, mental health status, and career guidance, were found to have a measurable impact on learning outcomes. Correlation analysis further confirms that students with better support systems, stable financial backgrounds, and access to essential educational resources tend to perform significantly better.

To better understand data distribution and key factor influences, various visualization techniques, including histograms, boxplots, and heatmaps, were employed. The distribution analysis of student percentage before training revealed a pattern where students with consistent attendance, motivation, and well-structured study environments achieved higher grades. The boxplot analysis demonstrated a clear relationship between parental support and student performance, reinforcing the importance of external academic support.

The model's predictive capability was validated by comparing actual and predicted values, showing a strong alignment that underscores the effectiveness of machine learning-based forecasting in academic performance evaluation. This study demonstrates that AI-driven educational analytics can provide actionable insights, enabling educators, policymakers, and institutions to implement targeted strategies for improving student success rates. By leveraging predictive analytics, institutions can personalize learning experiences, offer timely interventions, and address disparities in educational resources, ultimately fostering an environment conducive to academic excellence.

Feature Importance Analysis

The feature importance analysis provides valuable insights into the key factors that significantly impact student performance predictions. Using the Random Forest model, the importance of each feature was computed based on its contribution to predicting student percentage and grade. The feature importance scores indicate how much each independent variable influences the model's predictions.

The bar plot visualization ranks the features based on their impact, helping identify the most critical predictors of student academic success. The results highlight that factors such as parental support, study hours, class participation, socio-economic status, and access to study materials play a dominant role in shaping student outcomes. Additionally, other factors like teacher availability, mental health status, discipline level, and e-learning initiatives also exhibit a measurable influence, reinforcing the multi-dimensional nature of academic performance.

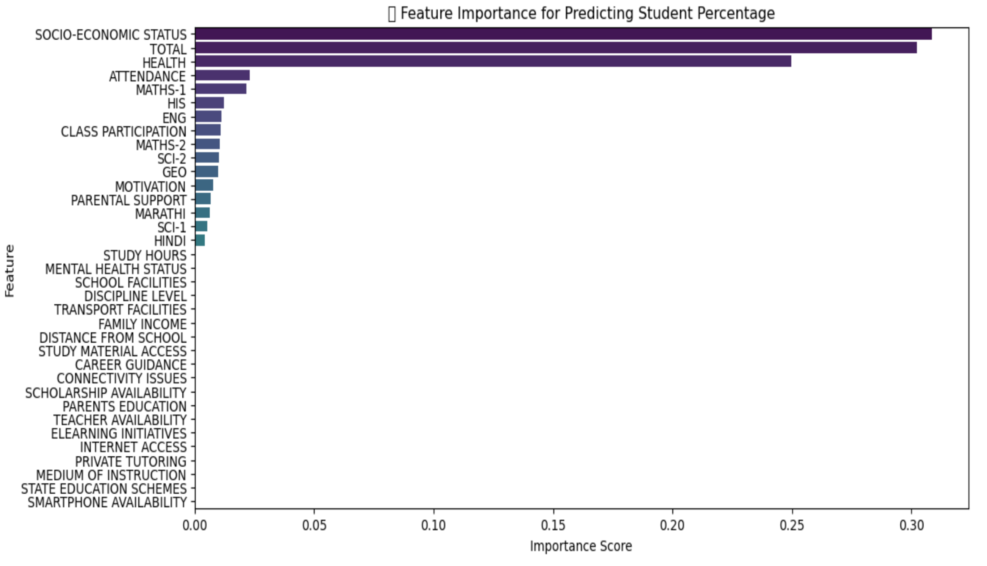


Figure1.9: Feature Importance

By leveraging feature importance analysis, educators and policymakers can prioritize interventions aimed at improving student learning experiences. Schools can focus on enhancing key aspects, such as increasing teacher availability, providing financial aid, improving access to digital learning resources, and promoting parental engagement. This analysis serves as a data-driven approach to identifying bottlenecks in the education system and implementing targeted improvements for better student performance outcomes.

Actual vs. Predicted Percentage

The Actual vs. Predicted Percentage plot provides a clear visualization of the model's performance in predicting student academic achievements. This scatter plot compares the true percentage values of students with the predicted values generated by the Random Forest model. Each blue dot represents an individual student's prediction, where the x-axis denotes the actual percentage, and the y-axis indicates the predicted percentage. The red dashed line represents the ideal scenario where the predicted values perfectly match the actual values. A strong alignment of points along this diagonal line indicates high model accuracy, while deviations suggest prediction errors. A scattered distribution with noticeable gaps from the red line highlights areas where the model might need further tuning, such as handling outliers, improving feature selection, or adjusting hyperparameters. This visualization helps assess the model's reliability and provides insights into how well the selected features contribute to accurate student performance predictions.

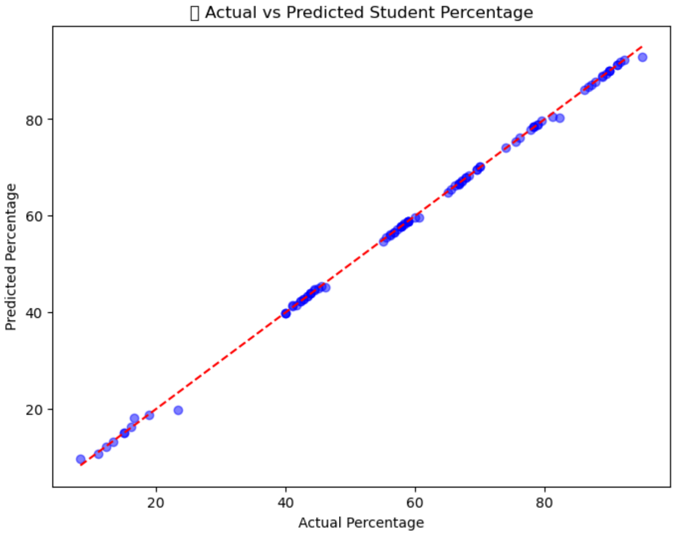


Figure1.10: Actual vs. Predicted Percentage

Residual Errors Distribution

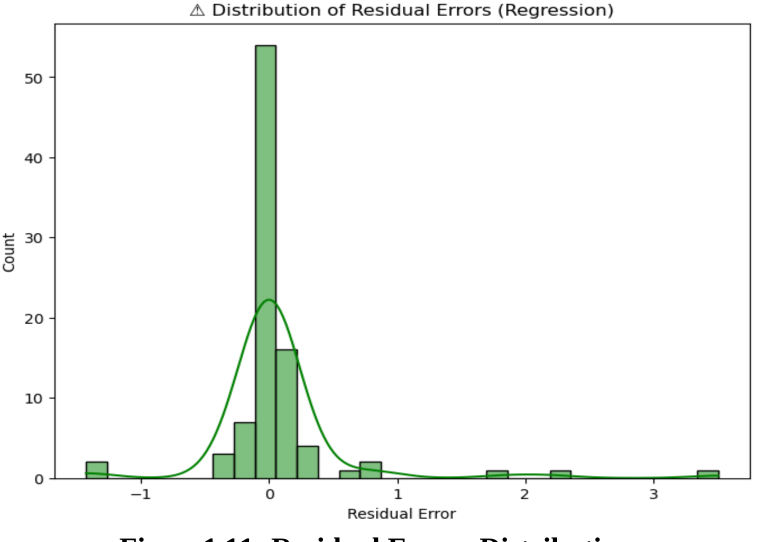


Figure1.11: Residual Errors Distribution

The Residual Errors Distribution plot provides a detailed assessment of the model's predictive performance by visualizing the differences between actual and predicted student percentages. Residual errors, calculated as the actual percentage minus the predicted percentage, indicate how well the model captures variations in student performance. Ideally, these errors should follow a normal distribution centered around zero, suggesting that the model's predictions are unbiased and evenly distributed. A narrow and symmetric distribution indicates minimal errors and a well-calibrated model, whereas a skewed or widely spread distribution may highlight potential biases, underfitting, or overfitting. Additionally, the presence of extreme residual values, or outliers, suggests cases where the model struggles to generalize, possibly due to missing influential factors or high variance in the dataset. By analyzing this distribution, necessary adjustments such as improved feature selection, data preprocessing, or hyperparameter tuning can be made to enhance the accuracy and reliability of the model.

Confusion Matrix

The Confusion Matrix provides a comprehensive evaluation of the classification model's performance in predicting student grades. It visually represents the number of correct and incorrect predictions made by the model by comparing actual and predicted grade labels. Each cell in the matrix indicates the count of instances classified into a specific category, where diagonal elements represent correct predictions, and off-diagonal elements indicate misclassifications. A well-performing model will have a higher concentration of values along the diagonal, signifying accurate grade predictions.

The confusion matrix also aids in identifying misclassification patterns, helping to determine whether specific grades are being frequently confused with others. If misclassification rates are high, potential reasons could include overlapping grade boundaries, insufficient training data, or the need for better feature selection. By analyzing this matrix, necessary improvements such as fine-tuning the model, optimizing class weights, or incorporating additional relevant factors can be made to enhance grade prediction accuracy.

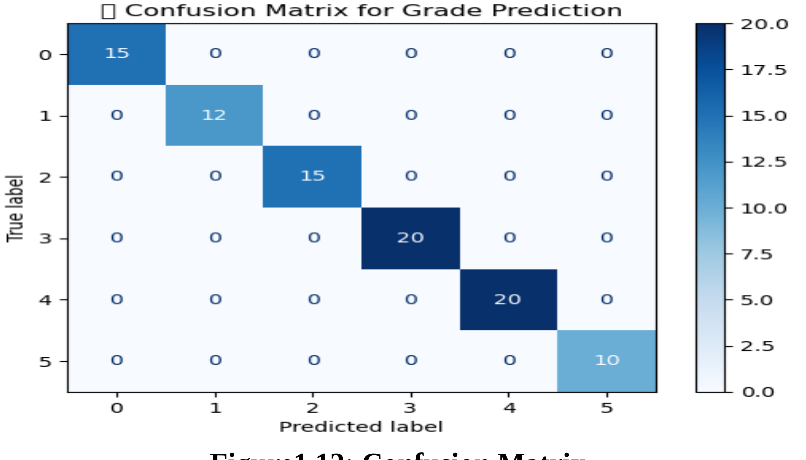


Figure1.12: Confusion Matrix

Actual vs. Predicted Values Table:

The Actual vs. Predicted Values Table provides a direct comparison between the true student performance outcomes and the model's predictions for both percentage and grade. This table includes four key columns:

1. Actual Percentage – The real percentage obtained by students.
2. Predicted Percentage – The percentage estimated by the machine learning model.
3. Actual Grade – The actual grade assigned based on the student's percentage.
4. Predicted Grade – The grade predicted by the classification model.

Table1.2: Actual vs. Predicted Value

Actual_percentage	Predicted_Percentage	Actual_Grade	Predicted_Grade
91.666667	91.883333	0	0
57.777778	57.794444	3	3
70.000000	70.105556	2	2
41.666667	41.594444	4	4
57.222222	57.233333	3	3

By analyzing this table, we can observe how closely the model's predictions align with real student performance. Any significant deviation between actual and predicted values may indicate areas where the model requires further refinement. If the predicted grades frequently differ from actual grades, it suggests that certain features need more weight or additional factors should be incorporated to improve accuracy. The results from this table help validate the effectiveness of the system in forecasting student academic accomplishments.

Future Scope

The future scope of this research encompasses several potential advancements aimed at enhancing the accuracy, applicability, and impact of the predictive system. One of the key improvements involves integrating more advanced machine learning models, such as artificial neural networks (ANN) and transformer-based techniques, to better capture the complex relationships between academic and non-academic factors. Additionally, real-time student performance tracking can be incorporated by linking the system with school databases, online learning platforms, and teacher feedback systems, enabling early intervention for students at risk of poor performance. Another significant extension is the development of personalized learning recommendations based on historical performance data, which can suggest study materials, courses, and extracurricular activities tailored to individual student needs. Expanding data collection sources by incorporating psychological assessments, peer interactions, and cognitive skill evaluations can further refine predictions and provide a more comprehensive academic assessment. To ensure accessibility and usability, a mobile or web-based application can be developed, allowing students, parents, and educators to easily access insights, monitor progress, and receive guidance on improving academic performance. Additionally, the system's adaptability can be extended across different geographical regions and educational frameworks to ensure its effectiveness across diverse socio-economic and cultural backgrounds. By implementing these advancements, the proposed system can evolve into an intelligent, data-driven educational tool that assists educators, policymakers, and students in making informed academic decisions and enhancing overall learning outcomes.

Conclusion

The proposed research successfully demonstrates the potential of machine learning in forecasting secondary school students' academic accomplishments by analyzing a comprehensive set of academic and non-academic factors. By leveraging the Random Forest algorithm, the model effectively predicts students' grades and percentage scores while identifying key contributing factors such as parental support, study hours, socio-economic status, and access to educational resources. The integration of data-driven insights enables a more objective and systematic approach to academic performance evaluation, aiding educators, parents, and policymakers in implementing targeted interventions for student improvement. The study highlights the importance of both academic and socio-environmental influences on student success, reinforcing the need for a holistic approach to education. The findings provide valuable insights for personalized learning strategies, resource allocation, and policy-making to bridge learning gaps and improve overall educational outcomes. While the model performs well, further enhancements such as real-time tracking, deep learning integration, and adaptive learning recommendations can improve its efficiency and applicability. In conclusion, this research establishes a strong foundation for AI-driven educational analytics, paving the way for future innovations in predictive academic assessment. By leveraging data-driven strategies, the proposed system has the potential to transform traditional education methodologies, ensuring a more personalized and effective learning experience for students across diverse educational backgrounds.

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.

Elisabeth Bauer et al., "Simulation-based learning in higher education and professional training: Approximations of practice through representational scaffolding" ELSEVIER, Volume 75, 2022.

Markus Berndt et al., "Impact of sender and peer-feedback characteristics on performance, cognitive load, and mindful cognitive processing" ELSEVIER, Volume 75, 2022.

Utkarsh Verma et al., "Prediction of students' academic performance using Machine Learning Techniques" IEEE,2020.

Adel Ismail Al-Alawi et al., "Predicting Student's Academic Performance Using Data Mining Methods: Review Paper" IEEE,2023.

Sagardeep Roy et al., "Predicting academic performance of student using classification techniques" IEEE,2017.

Meizar Raka Rimadana et al., "Predicting Student Academic Performance using Machine Learning and Time Management Skill Data" IEEE,2019.

Bo Guo et al., "Predicting Students Performance in Educational Data Mining" IEEE,2015.

Divya Thakur et al., "Predicting Student's Performance using Data Mining Algorithm" IEEE,2022.

Praveena Chakrapani et al., "Academic Performance Prediction Using Machine Learning: A Comprehensive & Systematic Review" IEEE,2022.

Nehal Eleyan et al., "Predicting Student Performance Using Educational Data Mining" IEEE, 2022.

D. K. Arun et al., "Student Academic Performance Prediction using Educational Data Mining" IEEE,2021.

Fergie Joanda Kaunang et al., "Students' Academic Performance Prediction using Data Mining" IEEE,2018.

K. Kaino et al., “Origami modeling method of leaves of plants and CG image generation of flower arrangement” IEEE, 2002.

Neeta Sharma et al., “Predicting Academic Performance of Students Using Machine Learning Models” IEEE,2023.

Tismy Devasia et al., “Prediction of students performance using Educational Data Mining” IEEE,2016.

Gomathy Suganya Ramaswami et al., “Predicting Students Final Academic Performance using Feature Selection Approaches” IEEE,2020.

Aws Khudhur et al., “Students' Performance Prediction Using Machine Learning Based on Generative Adversarial Network” IEEE,2023.

Essa Alhazmi et al., “Early Predicting of Students Performance in Higher Education” IEEE, Volume 18,2023.

Roselyne Ayienda et al., “Predicting Students Academic Performance using a Hybrid of Machine Learning Algorithms” IEEE,2021.

Mohamed Mohsen Elsaid Khoudier et al., “Prediction of student performance using machine learning techniques” IEEE, 2023.

Muslihah Binti Wook et al., “Predicting NDUM Student's Academic Performance Using Data Mining Techniques” IEEE,2009.

Isreal M. Ogundele et al., “Prediction of Student Academic Performance Based on Machine Learning Model” IEEE,2024.

Guiyun Feng et al., “Analysis and Prediction of Students’ Academic Performance Based on Educational Data Mining” IEEE, Volume 10, 2022.

Wentong Liu et al., “Student Performance Prediction by LMS Data and Classroom Videos” IEEE,2022.

Poonam S. Pawar et al., “A review on Student Performance Prediction using Educational Data mining and Artificial Intelligence” IEEE, 2021.

Behrouz Minaei-Bidgoli et al., “Predicting student performance: an application of data mining methods with an educational Web-based system” IEEE, 2004.

Sujith Jayaprakash et al., “Predicting Students Academic Performance using an Improved Random Forest Classifier” IEEE, 2020.

Ching-Chieh Kiu et al., “Data Mining Analysis on Student’s Academic Performance through Exploration of Student’s Background and Social Activities” IEEE,2018.

Xiaochen Lai et al., “Predicting the academic performance of students with GPcSAGE” IEEE,2023.

Rashmi Jha et al., “A Systematic Study on Student Performance Prediction from the Perspective of Machine Learning and Data Mining Approaches” IEEE, 2023.

K. Deepika et al., “Analyze and Predicting the Student Academic Performance Using Data Mining Tools” IEEE,2018.

Shoukath Tk et al., “Academic Performance Prediction of At-Risk Students using Machine Learning Techniques” IEEE, 2023.

Md Fahim Sikder et al., “Predicting students yearly performance using neural network: A case study of BSMRSTU” IEEE,2016.

Jacob Molina-Astorayme et al., “Predicting academic performance using automatic learning techniques: a review of the scientific literature” IEEE,2020.

S.N. van den Boom-Muilenburg et al., ELSEVIER,2023. “Sustaining data use professional learning communities in schools: The role of leadership practices”

ZILING CHEN et al., IEEE,2023. “Student Performance Prediction Approach Based on Educational Data Mining”

Evandro B. Costa et al., ELSEVIER ,2017. “Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses”

Eyman Alyahyan et al., IJETHE,2020. “Predicting academic success in higher education: literature review and best practices”

Yi Ren et al., "Long-term student performance prediction using learning ability self-adaptive algorithm," Springer, 2024.

Shah Hussain et al., "Student-Performulator: Predicting Students’ Academic Performance at Secondary and Intermediate Level Using Machine Learning," Springer, 2023.

Chunpeng Zhai et al., "The effects of over-reliance on AI dialogue systems on students’ cognitive abilities: a systematic review," Springer, 2024.

Mustafa Yağc., "Educational data mining: prediction of students' academic performance using machine learning algorithms," Springer, 2022.

Yudish Teshal Badal et al., "Predictive modelling and analytics of students' grades using machine learning algorithms," Springer, 2023.

Alberto Rivas et al., "Artificial neural network analysis of the academic performance of students in virtual learning environments," Elsevier, 2020.

Pallavi Asthana et al., "Prediction of Student's Performance with Learning Coefficients Using Regression Based Machine Learning Models," IEEE, 2023.

Luis Vives et al., "Prediction of Students' Academic Performance in the Programming Fundamentals Course Using Long Short-Term Memory Neural Networks," IEEE, 2023.

Samuel D. Mandelman et al., "Predicting academic performance and trajectories from a measure of successful intelligence," Elsevier, 2015.

N. D. Lynn et al., "Using Data Mining Techniques to Predict Students' Performance: A Review," ICIMECE, 2020.

Christin Lotz et al., "Differential relevance of intelligence and motivation for grades and competence tests in mathematics," Elsevier, 2014.

Amirah Mohamed Shahiria et al., "A Review on Predicting Student's Performance using Data Mining Techniques," Elsevier, 2015.

Mohammed Nasser Alsubaie, "Predicting student performance using machine learning to enhance the quality assurance of online training via Maharat platform," Elsevier, 2022.