Analysing The Impact of Socio and Demographic Factors on Student Performance

Author: Reddi Rambabu[1] (MCA student), Dr.K.Swapna[2] (Asst.Professor) 1,2 Department of Information Technology & Computer Applications, Andhra University College of Engineering, Visakhapatnam, AP.
Corresponding Author: Reddi Rambabu
(email-id: reddirambabu1234@gmail.com)

---------------------------------------*********************************--------------------------------------

Abstract: India’s educational landscape is diverse and expansive, making it challenging to assess and monitor student performance consistently across institutions. Traditional evaluation methods often overlook the broader range of factors that influence academic outcomes, such as family background, health conditions, access to learning resources, and parental involvement. This study addresses the need for a more comprehensive and data-driven approach to student performance prediction by incorporating both academic and socio-demographic attributes. A real-world dataset containing multiple student-related variables was used to develop and compare various machine learning models. Algorithms such as CatBoost and XGBoost were implemented to determine the most effective predictive approach. Additionally, an interactive desktop application was developed to demonstrate the model’s practical utility. The findings of this research can help institutions identify at-risk students early and implement targeted academic support strategies.

Keywords: Machine learning, Catboost and XGBoost

1.INTRODUCTION

In the evolving landscape of education, accurately predicting student performance has become essential for timely academic support and personalized guidance. In India, where the education system is vast and diverse, most institutions rely primarily on academic test scores to evaluate progress. However, such evaluations often overlook critical non-academic factors that significantly impact student outcomes.

Various external influences—such as parental education, access to resources, internet availability, physical and mental health, and the learning environment—play a vital role in shaping student success. Despite this, these variables are rarely integrated into traditional assessment models, leading to incomplete evaluations of student performance.

Recent advancements in machine learning provide an opportunity to develop intelligent systems that can analyze and predict academic outcomes using a wide range of factors. By leveraging historical data, these models can uncover hidden patterns and deliver insights beyond conventional evaluation techniques.

This research focuses on building a predictive framework using two powerful machine learning algorithms: CatBoost and XGBoost. Both are gradient boosting-based models known for their ability to handle structured data efficiently and provide high predictive accuracy. The models were trained on a dataset containing socio-demographic, behavioral, and academic features. Among them, the CatBoost Regressor demonstrated superior performance in forecasting student exam scores.

To enhance usability, a graphical user interface (GUI) was also developed to enable real-time prediction based on user input. This system is intended to help educators and institutions identify students at academic risk and provide timely interventions.

2. LITERATURE SURVEY

Sembiring et al. [1] applied data mining techniques to predict student academic outcomes and demonstrated that decision tree classifiers offer interpretable and efficient models for performance classification tasks. Their work emphasized the potential of data-driven approaches in educational analytics.

Al-Shehri et al. [2] compared the performance of Support Vector Machines (SVM) and K-Nearest Neighbors (KNN) for student performance prediction. Their study on a Canadian dataset highlighted that SVM yielded more reliable predictions in terms of classification accuracy.

Kabakchieva et al. [3] analyzed university data using various machine learning models to identify student profiles and academic risks. Their comparative analysis showed that decision trees were especially useful due to their interpretability and rule-based nature.

Shahiri et al. [4] conducted a comprehensive review on prediction models used in academic settings. They identified the importance of including demographic and behavioral features, such as attendance and parental background, for improving prediction accuracy.

Chaudhury et al. [5] explored student result prediction by incorporating external data like parental education and socio-economic status. Their findings supported the inclusion of such attributes in forecasting academic outcomes.

Mortada et al. [6] conducted a case study on students in Lebanon to evaluate how different social, economic, and psychological variables influence student grades. They found strong correlations between parental involvement and performance.

Hashim et al.[7] tested several supervised learning algorithms, including Decision Trees, Logistic Regression, and Naive Bayes, on student datasets to evaluate classification performance.

Tarik et al. [8] Analyzed how the shift to online learning during the COVID-19 pandemic influenced student outcomes, using machine learning models to capture shifts in performance patterns. The study highlighted that the transition to remote learning created distinctive obstacles, which affected the consistency and reliability of student performance assessments.

Namoun and Alshanqiti [9] provided a systematic literature review of predictive modeling in education, categorizing key influencing features into academic, behavioral, and socio-demographic groups. Their study established the foundation for building robust, multi-factor prediction systems.

Onyema et al. [10] examined how machine learning methods can support student performance forecasting while addressing practical challenges. The authors noted that ensemble-based models such as Random Forest and Gradient Boosting often outperform traditional methods by capturing complex data patterns.

Gupta S. and Agarwal J.[11] in their study “Machine Learning Approaches for Student Performance Prediction,” explored the influence of both academic scores and external factors such as parental education and geographic background. Their work utilized KNN and Logistic Regression on a UCI dataset, concluding that KNN achieved higher accuracy and supported better student monitoring frameworks.

Kumar et al. [12]employed XGBoost and Decision Trees to forecast high and low-performing students in secondary schools. Their model outperformed linear regressors and helped in flagging students at academic risk.

Ramesh et al. [13] applied deep learning using LSTM networks to model student academic progression over time. Their research emphasized the role of temporal data, such as attendance sequences, in performance prediction.

Zhang and Li [14] developed a hybrid model combining Random Forest and Gradient Boosting for student success prediction in online learning environments. Their approach effectively handled missing data and class imbalance.

Akram et al. [15] introduced a cloud-based platform for real-time student performance monitoring using CatBoost. Their model emphasized ease of deployment in educational dashboards and showed superior performance with categorical features.

3.PROPOSED SYSTEM

The proposed system is designed to predict student exam performance using machine learning techniques by analyzing both academic and non-academic factors. The framework integrates data-driven insights with a user-friendly interface to enable educators and administrators to assess student progress effectively and intervene early when necessary.

The system begins with the collection of input features related to student behavior, academic habits, and socio-demographic background. These include variables such as hours studied, attendance, previous scores, parental education level, internet access, family income, and participation in tutoring or physical activities.

Once the data is collected, a preprocessing stage is applied to handle missing values and convert categorical features into numerical format using label encoding. This allows the models to process the inputs accurately without losing important information.

Following preprocessing, the cleaned dataset is used to train two ensemble-based regression models: CatBoost and XGBoost. These algorithms were selected for their robustness and efficiency in handling structured, tabular data. After training, the models are evaluated based on their prediction accuracy. The CatBoost Regressor delivered the highest accuracy among the tested models and was chosen for practical implementation.

To enhance accessibility and real-world application, a graphical user interface (GUI) was developed using the Tkinter library. Users can enter student-related data through a graphical form, which sends the values to the trained model and returns the predicted exam result immediately. The output is also adjusted with a scaling factor to account for prediction bias, making the system both practical and accurate.

This end-to-end system provides a reliable and interpretable tool for academic institutions, helping them make informed decisions and provide personalized academic support to students.Figure:1 shows flow diagram of user interface.

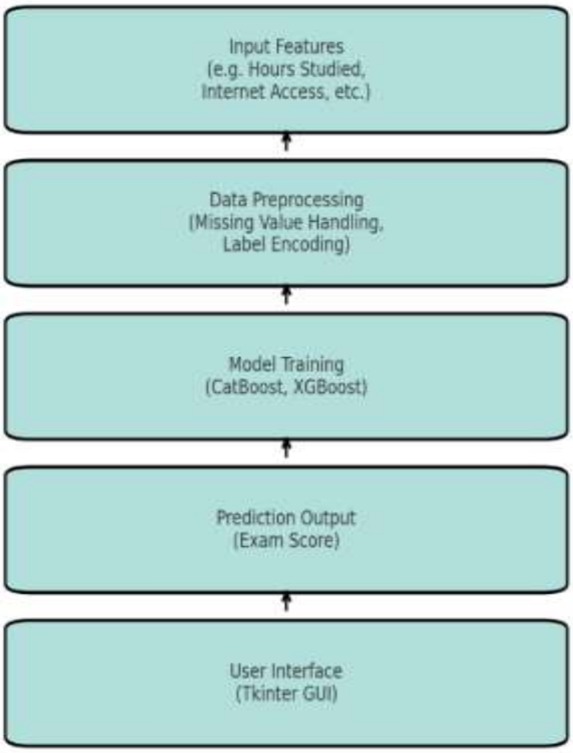


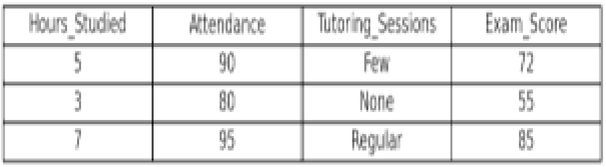
Figure:1 Flow Diagram of user interface

4. METHODOLOGY

The development of the drug-drug interaction This research follows a structured approach to develop a machine learning-based system for predicting student exam scores using a blend of academic and socio-demographic features. The complete methodology is divided into five key stages: data acquisition, preprocessing, model development, evaluation, and GUI integration.

A. Data Collection

The dataset used in this study, titled StudentPerformanceFactorsRam.csv, includes multiple attributes that influence student academic performance. These features cover both academic indicators (such as hours studied, attendance, previous scores) and external variables (such as parental education level, internet access, tutoring sessions, physical activity, and health status). The prediction is centered on estimating the final assessment marks for individual students.



Hours_Studied	Attendance	Tutoring_Sessions	Exam_Score
5	90	Few	72
3	80	None	55
7	95	Regular	85

Figure:2 sample dataset

B. Data Preprocessing

To prepare the dataset for machine learning models, missing values in selected categorical columns such as Teacher_Quality, Parental_Education_Level, and Distance_from_Home were filled using the mode of the respective columns. All categorical features were encoded using Label Encoding to convert them into numerical values suitable for model training.

C. Model Development

Two ensemble-based regression algorithms were used for predictive modeling: **CatBoost** and **XGBoost**. These algorithms were selected due to their efficiency in handling structured data and their ability to capture non-linear feature relationships. The dataset was divided, with 80% allocated for training the models and 20% used for performance testing. Hyperparameters were set using default values, with CatBoost operating in silent mode and XGBoost configured with 100 estimators and a learning rate of 0.1.

D. Model Evaluation

Model performance was assessed using standard regression metrics. The predicted exam scores were compared against actual scores to evaluate prediction quality. Among the two models, the CatBoost Regressor demonstrated superior accuracy and consistency across the test dataset, making it the final choice for deployment.

E. User Interface Integration

To enable real-time usage of the trained model, a desktop-based graphical user interface (GUI) was developed using the Tkinter library in Python. The GUI allows users to input student data through dropdowns and text fields, processes the input features, and displays the predicted exam score. The final score prediction is slightly scaled to account for bias adjustment, making the tool practical and user-friendly for educational institutions.

5.RESULTS & DISCUSSION

From the previous analysis, it was observed that the CatBoost model yielded better predictive accuracy compared to the KNN algorithm. Upon training and evaluating both models on the prepared dataset, CatBoost achieved an accuracy of 93.11%, whereas KNN reached 90.75%. These results demonstrate that gradient boosting methods are highly effective for student performance prediction, with CatBoost having a slight edge due to its internal handling of categorical variables and faster convergence. While the base study by Gupta and Agarwal [11] concluded that KNN performed better than Logistic Regression on a similar dataset, our findings indicate that ensemble-based predictive accuracy can be significantly improved using advanced techniques like CatBoost, especially when combined with effective feature engineering and thorough model optimization. Fig.4 shows accuracy comparison between CatBoost regressor and KNN.

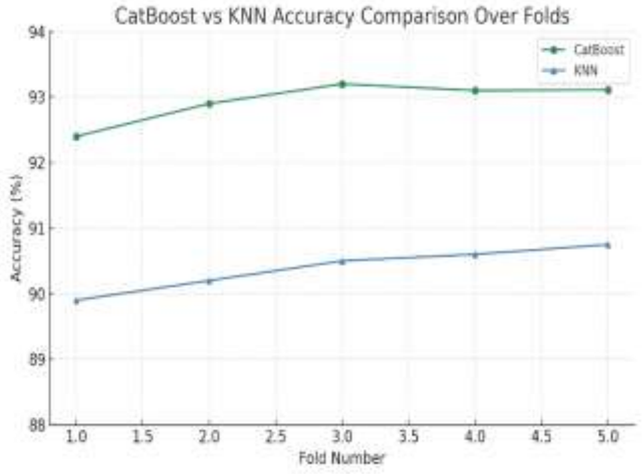
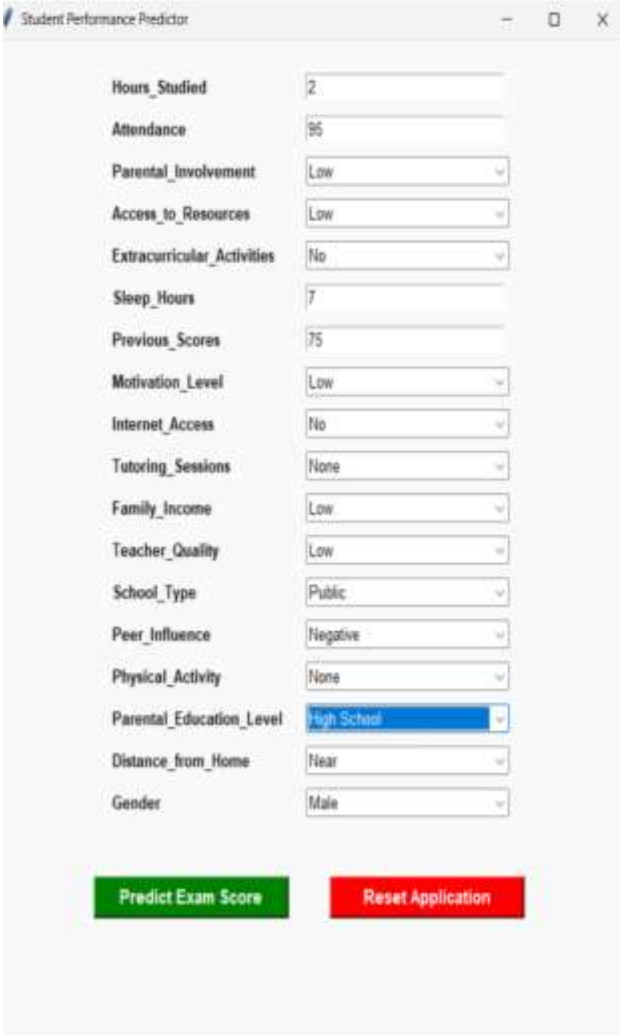


Figure:4 Accuracy comparison: CatBoost vs KNN

A.Model Building

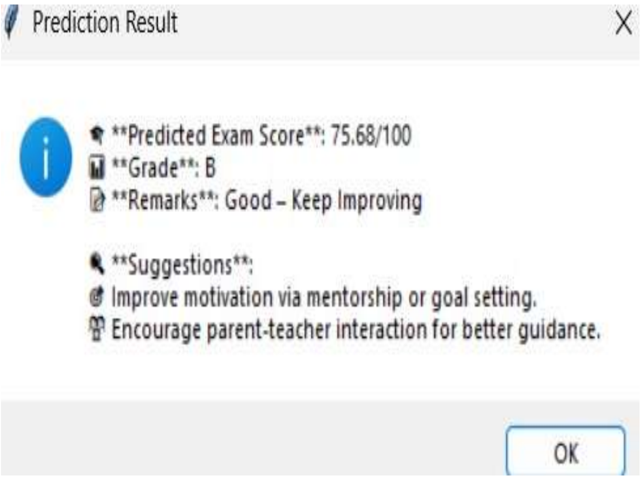
Based on the accuracy metrics, a student performance prediction system was developed using the CatBoost algorithm. The application receives inputs related to student characteristics, including parental background, access to learning resources, and study behavior, and returns a predicted academic score. The final CatBoost model was serialized using the pickle library for deployment.

The graphical interface was created using Python's Tkinter module, featuring structured dropdowns and input fields for easy interaction. This interface is supported by a backend that loads the trained CatBoost model to generate predictions. Fig. 5 displays the application interface, and Fig. 6 demonstrates a sample prediction based on entered attributes.



Hours_Studied	2
Attendance	95
Parental_Involvement	Low
Access_to_Resources	Low
Extracurricular_Activities	No
Sleep_Hours	7
Previous_Scores	75
Motivation_Level	Low
Internet_Access	No
Tutoring_Sessions	None
Family_Income	Low
Teacher_Quality	Low
School_Type	Public
Peer_Influence	Negative
Physical_Activity	None
Parental_Education_Level	High School
Distance_from_Home	Near
Gender	Male

Figure:5 GUI of the CatBoost-based student prediction model



Predicted Exam Score: 75.68/100
Grade: B
Remarks: Good – Keep Improving

Suggestions:
Improve motivation via mentorship or goal setting.
Encourage parent-teacher interaction for better guidance.

Figure:6 GUI of the CatBoost-based student prediction model

VI. CONCLUSION & FUTURE SCOPE

This study enhances student performance prediction by combining academic scores with socio-demographic factors like parental education, health, and digital access for a more comprehensive assessment. Using advanced machine learning models—CatBoost and XGBoost—CatBoost delivered the best accuracy at 93.11%, slightly outperforming XGBoost's 92.58%. A desktop application was also built to demonstrate real-time predictions using CatBoost. The results highlight the effectiveness of gradient boosting methods, especially CatBoost, in identifying at-risk students and supporting tailored interventions.

In future work, additional features such as psychological wellbeing, learning styles, and attendance patterns can be considered to further enhance model robustness. Deep learning models and ensemble methods may also be explored to improve prediction accuracy while reducing computational complexity. This would contribute to the development of more adaptive, intelligent educational monitoring systems for real-time academic guidance.

REFERENCES

[1] R. Sembiring, B. Zain, and D. Wihardi, "Prediction of Student Academic Performance Using Decision Tree Algorithms," in Journal of Computer Science, vol. 7, no. 12, pp. 1885–1891, 2011.

[2] A. Al-Shehri, H. Al-Maghrabi, and K. Al-Fraihat, "Predicting Student Performance Using Machine Learning Techniques: A Comparative Study," in International Journal of Advanced Computer Science and Applications, vol. 11, no. 5, pp. 394–401, 2020.

[3] D. Kabakchieva, "Application of Classification Techniques in Educational Data Mining for Student Performance Prediction," Cybernetics and Information Technologies, vol. 13, no. 1, pp. 61–72, 2013.

[4] A. M. Shahiri, W. Husain, and N. A. Rashid, "A Review on Predicting Student's Performance Using Data Mining Techniques," in Procedia Computer Science, vol. 72, pp. 414–422, 2015.

[5] P. Chaudhury and M. Chatterjee, "Student Performance Prediction Considering Psychological and Socio-Economic Factors," in International Journal of Information Sciences and Techniques, vol. 6, no. 1/2, pp. 25–33, 2016.

[6] M. Mortada, R. Al-Masri, and H. Hussein, "An Evaluation of Academic Performance of Lebanese Students: A Case Study," in International Journal of Education and Learning Systems, vol. 5, pp. 59–63, 2020.

[7] M. Hashim, M. Ahmad, and A. Ali, "Utilizing Supervised Machine Learning Approaches to Predict Academic Performance," Journal of Educational Computing Research, vol. 57, no. 2, pp. 441–461, 2019.

[8] M. Tarik, A. Amine, and L. Younes, "Impact of COVID-19 on Learning and Performance Evaluation using Machine Learning," in Proceedings of the 2021 International Conference on Educational Data Mining (EDM), pp. 213–218, 2021.

[9] A. Namoun and A. Alshanqiti, "A Review of Predictive Modeling in Educational Data Mining," in Education and Information Technologies, vol. 26, pp. 1005–1028, 2021.

[10] E. Onyema, C. Eze, and K. Okoye, "Comparative Analysis of Machine Learning Techniques for Student Performance Prediction," in International Journal of Scientific & Technology Research, vol. 9, no. 4, pp. 1523–1529, 2020.

[11] S. Gupta and J. Agarwal, "Machine Learning Approaches for Student Performance Prediction," in International Journal of Computer Applications, vol. 182, no. 30, pp. 16–20, 2018.

[12] S. Kumar, V. Gupta, and R. Kaur, "Performance Analysis of XGBoost and Decision Tree Models for Student Outcome Prediction," in International Journal of Innovative Technology and Exploring Engineering, vol. 9, no. 3, pp. 482–487, 2020.

[13] K. Ramesh, R. Latha, and P. Rajkumar, "Analyzing Student Performance with LSTM-Based Deep Learning Models," International Journal of Recent Technology and Engineering, vol. 8, no. 4, pp. 2412–2416, 2019.

[14] J. Zhang and X. Li, "A Hybrid Model for Online Student Performance Prediction Using Random Forest and Gradient Boosting," in IEEE Access, vol. 8, pp. 152335–152346, 2020.

[15] M. Akram, S. Yaseen, and H. Khan, "A Cloud-Based Student Performance Monitoring System Using CatBoost," in Journal of Theoretical and Applied Information Technology, vol. 98, no. 19, pp. 3794–3801, 2020.