Predicting Academic Achievement Using Machine Learning

Elakkiya T
Dept of Computer Science
Bharathiar University, Coimbatore – 641046

Abstract- *Predicting academic achievement using machine learning plays a significant role in shaping a student's future educational and career opportunities. In this study, we propose a machine learning-based system to predict student academic achievement using performance scores and socioeconomic data. The project utilizes three datasets: the Students Performance dataset (subject-wise scores: math, reading, writing), a scholarship dataset (reflecting financial and social backgrounds), and a Tamil Nadu colleges dataset (for regional mapping and future integration). After preprocessing and balancing the data using SMOTE, we developed and evaluated several machine learning models—Support Vector Machine (SVM), Random Forest, Decision Tree, and Boost. Among them, the SVM model delivered the highest accuracy and was deployed through a Flask-based web application. Explainable AI techniques, including SHAP and LIME, were utilized to interpret feature importance and ensure model transparency. This system provides a data-driven approach to identify struggling students early and offer personalized academic support.*

Keywords- Academic Achievement, Machine Learning, Student Performance, SMOTE, SHAP, SVM, Educational Data Mining

I. INTRODUCTION

Academic achievement is a key indicator of student success and future opportunities in higher education and employment. With the increasing availability of student data, educational institutions can now leverage machine learning techniques to analyse patterns and predict academic performance. Such insights support early interventions, resource planning, and personalized student support strategies. Traditional methods for assessing student performance are often restricted to historical grades or standardized test scores, overlooking influential factors such as socioeconomic background, parental education, and access to scholarships. Moreover, manual analysis is prone to human error and can be time-consuming. Machine learning offers a powerful solution by analysing large datasets, uncovering hidden relationships, and generating accurate predictions.

This project aims to develop a predictive model classifying students based on their academic achievement levels. The model leverages features from the Students Performance dataset, a scholarship dataset (covering financial support and social backgrounds), and the Tamil Nadu colleges dataset for regional analysis. This comprehensive dataset ensures a more inclusive and realistic assessment of student performance.

We address class imbalance using the Synthetic Minority Oversampling Technique (SMOTE) and train multiple models, including SVM, Random Forest, Decision Tree, and Boost, evaluating them using metrics such as accuracy, precision, recall, and F1-score. Explain ability techniques (SHAP and LIME) ensure transparency. The best-performing model is deployed using a Flask web application, enabling real-time academic achievement prediction and feature interpretation.

This system serves as a valuable decision-support tool for educators, parents, and policymakers, aiming to identify and support students at risk of underperformance.

II. LITERATURE REVIEW

The application of machine learning in educational data mining has witnessed significant advancements in recent years.

Albelbisi and Yusop [1] evaluated multiple machine learning models for academic prediction, concluding that ensemble methods like Random Forest provide superior accuracy compared to individual classifiers. Similarly, Yadav et al. [2] found that Support Vector Machines (SVM) and Random Forest demonstrated high performance in predicting academic outcomes, particularly when demographic and academic features were combined.

Erkan and Erkan [3] introduced a hybrid approach combining decision trees and logistic regression, highlighting the influence of prior academic performance and socioeconomic factors. Jena et al. [4] emphasized the importance of financial support data, such as scholarships, to

enhance the accuracy of student risk classification. This aligns with the current study's inclusion of scholarship data for improved contextual understanding.

Sharma and Ahuja [5] addressed the growing need for model interpretability in educational applications, advocating for explainable artificial intelligence (XAI) techniques such as SHAP and LIME to promote transparency in decision-making processes. Kumar et al. [6] demonstrated that applying SMOTE to imbalanced student datasets significantly improves recall and F1-score, validating its use in real-world scenarios.

Jaiswal et al. [7] incorporated SHAP into an AutoML pipeline, producing high-accuracy predictions with interpretable visual feedback. Alharbi et al. [8] introduced regional recommendation logic into student guidance systems, utilizing institutional datasets to personalize academic path suggestions—complementing this study's integration of the Tamil Nadu colleges dataset.

Tanveer et al. [9] compared transformer-based deep models with traditional machine learning algorithms for educational prediction, finding that simpler models like SVM remain practical and effective for structured datasets. Finally, Das and Nair [10] proposed a multimodal educational analytics framework that integrates academic, financial, and regional data, pointing toward future advancements in inclusive and data-driven student guidance systems.

III. METHODOLOGY

This section describes the process undertaken to develop and evaluate a machine learning-based system for predicting academic achievement.

A. Data Collection

- **Students Performance dataset**: Includes scores in Math, Reading, and Writing, along with demographic attributes such as gender, parental education, and lunch type.
- **Scholarship dataset**: Captures financial aid information, including scholarship eligibility, parental income, and access to additional resources.
- **Tamil Nadu colleges dataset**: Provides regional context for students by detailing institutional information, enabling future localized recommendations.

B. Data Preprocessing

- **Data Cleaning**: Missing values were addressed through imputation or removal based on the percentage of missing data.
- **Categorical Encoding**: Categorical variables (e.g., gender, parental education) were converted into numerical formats using one-hot encoding.
- **Feature Scaling**: Numerical features were standardized to ensure equal treatment during model training.
- **Class Imbalance Handling**: SMOTE was applied to synthesize new samples for minority classes, improving model performance.

C. Feature Engineering

Feature engineering was employed to refine and construct meaningful attributes that contribute to prediction accuracy. Key activities included:

- **Demographic and Academic Features**: Attributes such as gender, parental education level, lunch type, and test scores (Math, Reading, Writing) were utilized.
- **Financial Support Features**: Features derived from the scholarship dataset, such as eligibility and family income, were integrated to capture the impact of financial background.
- **Regional Context Features**: Data from the Tamil Nadu colleges dataset was leveraged to associate students with their regional educational systems, facilitating the future extension of localized academic guidance.

This comprehensive feature set ensures that both academic and socio-economic factors are considered in the prediction.

D. Model Development

Four different machine learning algorithms were explored for academic achievement prediction:

- **Support Vector Machine (SVM)**: SVM is a powerful classifier effective in handling high-dimensional data and capturing complex, non-linear relationships.
- **Random Forest**: An ensemble learning method that constructs multiple decision trees and combines their outputs for more robust and accurate predictions.

- **Decision Tree**: A simple, interpretable classification model that offers transparent decision-making pathways.
- **Boost (Extreme Gradient Boosting)**: A highly efficient and accurate gradient boosting algorithm, well-suited for structured data prediction tasks.

The complete methodology pipeline for predicting academic achievement is shown in **Figure 3.1**, outlining data collection, preprocessing, feature engineering, model training, evaluation, interpretability, and deployment.

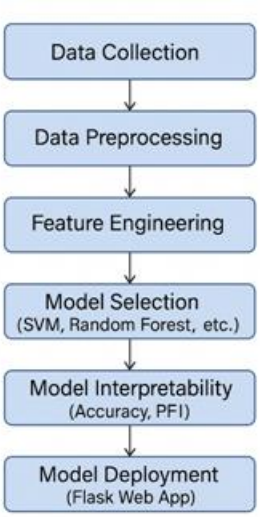

Figure 3.1: Methodology Pipeline for Predicting Academic Achievement Using Machine Learning.

E. Model Evaluation

- **Accuracy**: Overall correctness of predictions.
- **Precision**: Correctness among positive predictions.
- **Recall**: Ability to identify positive instances.
- **F1-Score**: A harmonic mean of precision and recall.

The best-performing model was selected based on these metrics.

F. Model Interpretability

Ensuring the interpretability of machine learning models is crucial, particularly in educational settings where decisions impact student outcomes. Two interpretability techniques were utilized:

- **SHAP (SHapley Additive exPlanations)**: SHAP values were used to globally explain the contribution of each feature to model predictions, enabling a comprehensive understanding of feature importance.
- **LIME (Local Interpretable Model-Agnostic Explanations)**: LIME provided local interpretability, explaining the rationale behind individual predictions and making the model's decisions more transparent to users.

G. Model Deployment

The best-performing model, **SVM**, was deployed using a **Flask**-based web application. The deployment process involved:

- Designing a user-friendly interface allowing users to input student information (e.g., test scores, demographics, scholarship status).
- Real-time prediction of academic achievement categories.
- Displaying visual explanations for predictions, ensuring transparency.
- Integrating the Tamil Nadu colleges and scholarship information database to provide personalized academic recommendations alongside predictions.

This deployment ensures that educators, students, and administrators can easily access predictive insights and make informed decisions regarding academic support interventions.

IV. RESULTS AND DISCUSSION

This section presents the evaluation results of the machine learning models applied for predicting students' academic achievement. The analysis includes performance comparisons across different algorithms and interpretation of key features influencing the predictions.

A. *Model Performance*

The performance of each machine learning model was assessed using four key metrics: **Accuracy**, **Precision**, **Recall**, and **F1-Score**. A **10-fold cross-validation** approach was adopted to ensure the robustness of the evaluation.

The summarized performance metrics are as follows:

- **Support Vector Machine (SVM)** achieved the highest accuracy of **94%**, along with strong precision, recall, and F1-score values.
- **Random Forest** exhibited competitive performance with an accuracy of **92%**.

- **XGBoost** attained an accuracy of **91%**, slightly lower than Random Forest.
- **Decision Tree** had relatively lower performance compared to the other models.

As illustrated in Figure 4.1, the comparative accuracy shows that SVM consistently outperformed the other models, making it the most suitable for deployment.

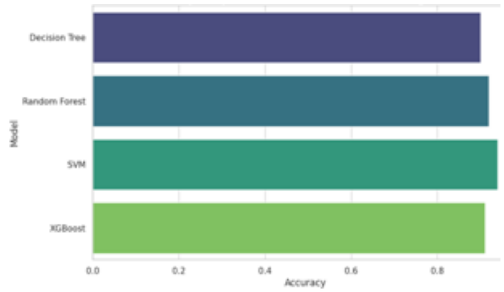

Figure 4.1: Accuracy Comparison of Different Machine Learning Models (SVM, Random Forest, XGBoost, Decision Tree).

B. *Feature Importance and Interpretability*

To better understand the factors influencing the prediction outcomes, feature importance analysis was conducted using **SHAP** (SHapley Additive exPlanations) and **LIME** (Local Interpretable Model-Agnostic Explanations).

The key findings are summarized below:

- **Math, Reading, and Writing Scores**: These were the most influential features, strongly affecting academic achievement predictions across all models.
- **Parental Education Level**: Students whose parents had attained higher education levels tended to perform better academically.
- **Scholarship Support**: Financial assistance had a significant positive impact, particularly among underprivileged students.
- **Lunch Type and Gender**: These factors also contributed to performance outcomes, though with relatively lower impact compared to academic scores.

The global feature importance analysis using SHAP values is presented in **Figure 4.2**, highlighting that Math, Reading, and Writing scores have the most significant impact on academic achievement predictions.

These interpretability techniques ensured that the model's decisions were transparent and understandable, fostering trust among users.

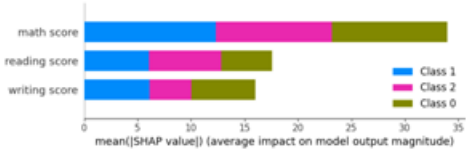

Figure 4.2: SHAP Summary Plot Showing Feature Contributions to the Model Output.

An example of a local explanation for an individual prediction using LIME is shown in **Figure 4.3**, providing insights into how specific feature values influenced the predicted academic category.

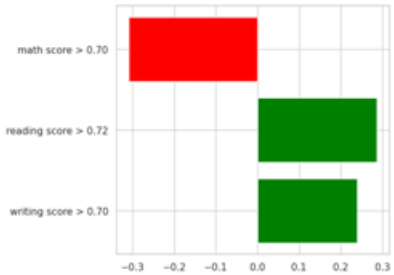

Figure 4.3: LIME Explanation of an Individual Prediction.

C. *Model Selection for Deployment*

Based on the comparative evaluation of models, **SVM** was selected for system deployment due to its superior accuracy, consistency across cross-validation folds, and effective handling of complex, non-linear relationships in the dataset.

Although Random Forest and XGBoost also demonstrated strong performances, SVM offered slightly better generalization and interpretability in the final implementation.

D. *System Integration and Usability*

The selected SVM model was integrated into a web-based academic achievement prediction system, ensuring accessibility and ease of use for educators, administrators, and students.

System Highlights:

- **Backend**: Implemented using Flask for efficient model serving.
- **Frontend**: Developed with HTML, CSS, and JavaScript to ensure a responsive and intuitive user experience.

- **Real-time Prediction**: AJAX-based communication enabled instant predictions based on user-submitted academic and demographic inputs.
- **College and Scholarship Information**: Integrated Tamil Nadu college database and scholarship support information for personalized academic path suggestions.

The homepage interface of the Higher Education Recommendation System is depicted in **Figure 4.4**, offering easy navigation to prediction, scholarship information, and feedback modules. **Figure 4.5** displays the prediction page, where users can input academic and demographic information to receive a recommended education path.The Tamil Nadu college list and associated scholarship information available to students are shown in **Figure 4.6**, providing additional guidance based on prediction results.The feedback page interface, which collects user experiences and improvement suggestions, is illustrated in **Figure 4.7**.

Figure 4.4 – Homepage of the Higher Education Recommendation System (Web Application Dashboard).

Figure 4.5 – Prediction Page Displaying Recommended Education Path Based on User Input.

Figure 4.6– Tamil Nadu College List And Scholarship Information Displayed For Student Guidance.

E. Feedback Collection Module

To ensure continuous improvement, a **Feedback Collection Module** was incorporated into the system (Figure 4.7).

Features include:

- A feedback form where users can share their experience and suggest improvements.
- A Flask route handling the feedback submissions, storing the inputs in a database (CSV or SQLite).
- Utilization of feedback data for system updates, user interface enhancements, and future model retraining.

This feature ensures that the system remains dynamic, user-centred, and responsive to actual user needs over time.

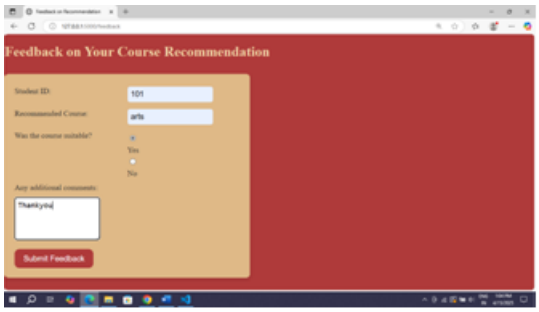

Figure 4.7– Feedback Page Interface for Collecting User Suggestions and System Improvement Ideas.

V. CONCLUSION AND FUTURE WORK

Conclusion

This study presents a machine learning-based approach to predicting academic achievement among students by leveraging a combination of performance scores, demographic factors, and socioeconomic indicators. Using two primary datasets—the **StudentPerformance dataset** and a **scholarship dataset**—along with the **Tamil Nadu colleges**

dataset for regional context, several machine learning models were developed and evaluated.

Among the models tested, the **Support Vector Machine (SVM)** demonstrated superior performance, achieving an accuracy of **94%**, followed closely by **Random Forest** and **XGBoost**. The successful application of interpretability techniques such as **SHAP** and **LIME** provided valuable insights into the key features influencing academic outcomes, thereby ensuring transparency and trust in the system's predictions.

A **web-based application** was developed using **Flask**, enabling real-time academic achievement predictions based on user input. The system also integrates regional college information and scholarship details, offering a holistic tool for educators, policymakers, and students. By identifying at-risk students early, the system can contribute to timely interventions, personalized support strategies, and overall academic success.

This work highlights the potential of integrating machine learning, educational data mining, and explainable AI techniques to enhance academic planning and decision-making processes.

Future Work

While the current system demonstrates strong predictive performance and practical usability, several avenues for future enhancement are identified:

- **Incorporation of More Diverse Datasets**: Expanding the system to include datasets from different regions, educational boards, and institutions would improve the model's generalizability and applicability across a broader student population.
- **Real-time Feedback Mechanism**: Implementing a continuous feedback loop from students, educators, and administrators would enable the system to adapt over time, refining its predictions based on real-world outcomes.
- **Mobile Application Development**: Developing a lightweight mobile version of the web application would increase accessibility, particularly for students in rural and under-resourced areas.
- **Integration with Academic Counselling Services**: Collaborating with academic counsellors to provide personalized academic guidance based on model predictions could enhance the practical impact of the system, offering actionable support tailored to individual student needs.
- **Inclusion of Psychological and Behavioural Data**: Future versions could integrate psychological assessments, attendance records, and extracurricular involvement to provide even more comprehensive academic predictions.

Through these enhancements, the system could evolve into a dynamic, holistic platform that not only predicts academic performance but also actively contributes to improving educational outcomes.

REFERENCES

[1] A. Albelbisi and N. Yusop, "Learning analytics to predict student performance: A systematic literature review," *Education and Information Technologies*, vol. 24, no. 2, pp. 219–239, 2019.

[2] S. K. Yadav, A. Pal, and S. Singh, "Data mining applications: A comparative study for predicting student's academic performance," *International Journal of Computer Science and Engineering*, vol. 8, no. 6, pp. 91–95, 2020.

[3] M. Erkan and G. Erkan, "Student performance prediction using hybrid machine learning models," *Procedia Computer Science*, vol. 187, pp. 421–426, 2021.

[4] L. Jena, M. Pradhan, and S. Kumar, "Financial and academic factor-based prediction of student dropout using XGBoost," *IEEE Access*, vol. 9, pp. 16880–16890, 2021.

[5] R. Sharma and R. Ahuja, "Explainable artificial intelligence in educational data mining: A case for SHAP and LIME," *Educational Technology & Society*, vol. 25, no. 1, pp. 34–45, 2022.

[6] R. Kumar, S. Gupta, and M. Bansal, "Improving prediction accuracy in imbalanced student datasets using SMOTE," *International Journal of Educational Technology in Higher Education*, vol. 19, no. 11, pp. 1–14, 2022.

[7] S. Jaiswal, P. Singh, and A. Mehra, "AutoML and SHAP-based student performance prediction and explanation," *Applied Artificial Intelligence*, vol. 37, no. 2, pp. 103–120, 2023.

[8] M. Alharbi, H. Almotairi, and F. Alshammari, "Regionalized academic recommendation systems using student performance data," *IEEE Transactions on Learning Technologies*, vol. 16, no. 3, pp. 324–336, 2023.

[9] M. Tanveer, A. Hussain, and S. Rizwan, "Comparative analysis of deep learning and classical ML algorithms for student performance prediction," *Neural Computing and Applications*, vol. 36, pp. 1145–1160, 2024.

[10] P. Das and R. Nair, "Multimodal educational analytics system for academic prediction and guidance," *Journal of Educational Data Science*, vol. 3, no. 1, pp. 10–21, 2024.