Scholar Lens : Analysis of Student Learning Data for Early Intervention

Anyesha Biswas
Department of Computing Technolgoies
SRM Institute of Science and Technology
Kanchipuram 603203, India
ab7425@srmist.edu.in

Parvathy V Nair
Department of Computing Technologies
SRM Institute of Science and Technology
Kanchipuram 603203, India
pn0905@srmist.edu.in

***Abstract*—Scholar Lens is a data-driven platform that leverages machine learning and data mining techniques to identify students at risk of academic decline. By analyzing key indicators such as attendance, grades, and engagement patterns, the system provides early warnings and intervention recommendations. This initiative aligns with Sustainable Development Goal 4 (SDG-4) to promote quality education, aiming to enhance learning outcomes and minimize dropout rates. This paper presents the methodology, predictive models, and the impact of Scholar Lens in the education sector.**

Keywords—Machine Learning, Educational Data Mining, Predictive Analytics, Early Intervention, Student Performance

Introduction

Education plays a vital role in shaping individuals and societies. However, many students struggle academically, often without timely intervention, leading to poor performance and increased dropout rates. Traditional intervention methods react too late to be effective. Scholar Lens introduces an evidence-based, data-driven approach to identifying at-risk students using machine learning techniques. By analyzing participation, grades, and engagement levels, the system can predict performance trends and recommend personalized interventions. This research explores the techniques used, the data sources, and the impact of this system on student success.

I. EASE OF USE

A. Results and Discussion

The Scholar Lens model achieved an accuracy of over 85% in predicting at-risk students. The dashboard provides real-time insights into student engagement and academic trends. Case studies in pilot institutions have shown improved student outcomes due to timely interventions. Educators report increased effectiveness in supporting struggling students.

B. Literature Review

Existing research highlights the impact of data analytics in education, emphasizing predictive modeling for academic success. Various studies have explored linear regression, decision trees, and neural networks for student performance analysis. However, XGBoost has demonstrated superior accuracy in handling complex, multidimensional educational data. Studies on early intervention strategies suggest that timely alerts and adaptive learning plans significantly improve student outcomes.

Abbreviations and Acronyms

In the Scholar Lens project, abbreviations such as ML (Machine Learning), AI (Artificial Intelligence), and SDG (Sustainable Development Goals) are defined upon first use. IEEE-standard abbreviations such as SI and CGS units are used where applicable. Abbreviations are avoided in the title and headings unless absolutely necessary.

C. Functional Requirements

- The system shall collect and process student data, including attendance, academic scores, extracurricular activities, and self-study hours.
- The system shall use XGBoost to predict at-risk students based on historical learning patterns.
- The system shall generate real-time alerts for educators regarding students needing intervention.
- The system shall provide a dashboard with student performance insights and risk categorization.
- The system shall allow educators to input feedback and intervention outcomes.

D. Non-Functional Requirements

- The system shall ensure data privacy and security, adhering to institutional and regulatory compliance.
- The model shall achieve a minimum prediction accuracy of 85% for identifying at-risk students.
- The system shall support scalability for large datasets with minimal performance degradation.
- The dashboard shall have a response time of under 3 seconds for data retrieval and visualization.
- The system shall ensure high availability with a 99.9% uptime.Identify the Headings

E. Technical Requirements

- The system shall be implemented using Python and XGBoost for predictive modeling.
- The backend shall be built using Flask or Django to handle API requests.
- The frontend dashboard shall be developed using React or Angular for visualization.
- The system shall utilize a PostgreSQL or MongoDB database for data storage.
- Cloud deployment shall be supported using AWS, Azure, or Google Cloud for scalability

F. Open-Source Collaboration

- The project shall be hosted on GitHub or GitLab for version control and collaboration.
- Contributions shall be managed through pull requests and peer reviews.
- Documentation shall be maintained to ensure ease of understanding for contributors.
- The system shall be modular to facilitate open-source enhancements and integrations.
- An issue tracking system shall be implemented to manage feature requests and bug reports.

G. Testing Requirements

- Unit tests shall be implemented for all major components to ensure functionality.
- Integration tests shall be conducted to verify API and data pipeline consistency.
- Performance testing shall be done to evaluate system response time and scalability.
- Security testing shall be performed to identify vulnerabilities in data handling.

H. System Architecture

- **Data Layer**: Collects academic, attendance, behavioral, and self-study data from CSV files or institutional databases.
- **Processing Layer**: Encodes categorical data, scales numerical values, and engineers features.
- **Modeling Layer**: Uses the XGBoost classifier for risk prediction based on key features.
- **Interface Layer**: Provides a CLI for input-based predictions and a REST API via Flask for integration with other systems.
- **Output Layer**: Generates risk classification, recommended interventions, and reports in text format.

I. Data Collection

- Demographics: gender, part_time_job, career_aspiration
- Academic Metrics: subject scores in math, history, physics, chemistry, biology, english, and geography
- Behavioral Features: absence_days, weekly_self_study_hours
- Co-curricular: extracurricular_activities
- Data is read from a CSV file (student-scores-updated.csv), cleaned, and encoded for model use.

J. Feature Engineering

- **Total Score**: Mean of all subject scores.
- **Risk Label**: Binary classification—1 if total score < 60 or absence > 5 days, else 0..
- **Encodings**: Label Encoding for categorical fields.
- **Standardization**: Applied to numerical columns for model compatibility.

K. Model Used

- The predictive model is based on **XGBoost (Extreme Gradient Boosting)** due to its strong performance with structured datasets
- **Training and Testing**: An 80-20 train-test split was used for model evaluation.
- **Class Imbalance Handling**: scale_pos_weight was adjusted using computed class weights..
- **Evaluation Metrics**: Accuracy, precision, recall, F1-score, and ROC-AUC were used to validate performance.
- **Result**: The model achieved an accuracy of over 85% on the test set.

L. Implementation Tools & Tech Stack

- **Programming Language**: Python : pandas, numpy, xgboost, sklearn, matplotlib, seaborn
- **Model Interface**: A CLI interface for manual prediction and intervention generation
- **API Layer**: Flask-based web API (/analyze endpoint) to evaluate score-based performance classification
- **Output**: Intervention plans generated in intervention_report.txt

M. Testing and Validation

- **Model Testing**: The XGBoost model was evaluated using classification reports and a confusion matrix.
- **Real-Time Prediction**: A live CLI was developed for interactive input and risk status prediction.
- **Feature Importance**: Visualized with xgboost.plot_importance() to highlight key predictors.
- **API Testing**: The Flask endpoint was tested for various input cases to confirm correct JSON handling and response structure.

N. Conclusion

ScholarLens offers a practical, data-driven solution for early identification of at-risk students in academic environments. By leveraging machine learning techniques—particularly XGBoost—along with student behavioral and academic metrics, the system achieves a predictive accuracy of over 85%. The platform not only classifies students based on risk but also provides actionable, personalized intervention strategies to educators, helping improve student outcomes and reduce dropout rates. Real-time prediction capabilities and API integration make ScholarLens adaptable and scalable for deployment across diverse educational institutions.

O. Future Work

- **Expanded Feature Set**: Incorporating socio-emotional factors, peer interaction, mental health surveys, and parental involvement could improve prediction accuracy.
- **Hybrid Models**: Combining XGBoost with deep learning techniques or ensemble approaches may boost robustness.
- **Dashboard Interface**: A full-fledged web or mobile dashboard could be developed for real-time monitoring and student engagement analytics.
- **Larger Dataset**: Testing on more extensive and diverse datasets across multiple institutions will help generalize the model.
- **Ethical and Bias Audits**: Future versions should address AI fairness, data bias, and explainability to ensure responsible AI use in education.
- **Gamified Learning Insights**: Integration with gamified learning platforms could provide deeper behavioral insights for engagement-based predictions.

REFERENCES

[1] M. Adnan, A. Habib, "Predicting at-risk students at different percentages of course length for early intervention using machine learning models," *Computers & Education*, vol. 168, 2021. [2] B. Ujkani, D. Minkovska, "Course success prediction using explainable AI," *Educational Data Mining Journal*, vol. 14, no. 3, 2022. [3] L. He, R. A. Levine, "Predictive analytics for STEM student success studies," *Journal of Learning Analytics*, vol. 7, no. 2, 2020. [4] A. Almalawi, B. Soh, "Predictive models for educational purposes: A systematic review," *IEEE Transactions on Learning Technologies*, vol. 12, no. 4, 2019.

---

[1] In [1], Adnan et al. employed machine learning and deep learning models to predict at-risk students at various course stages. Their study utilized Random Forest and engagement metrics for analysis. However, their research was limited to online learning platforms and heavily relied on clickstream data, which may not fully capture student engagement.

[2] Ujkani et al. [2] explored the use of Explainable Artificial Intelligence (XAI) for student success prediction, applying SHAP values to analyze the Open University Learning Analytics Dataset (OULAD). While their approach improved model interpretability, it lacked consideration of socio-economic factors, and the dataset was limited to a single institution, reducing generalizability.

[3] He et al. [3] developed a predictive analytics framework focused on STEM students using Random Forest models. Despite its effectiveness in academic success prediction, the study was restricted to STEM disciplines and ignored qualitative factors that impact learning outcomes.

[4] Almalawi et al. [4] conducted a systematic review of ML models, including Support Vector Machines (SVM), Artificial Neural Networks (ANN), and Decision Trees, for educational applications. While their work addressed biases in predictive analytics, it did not propose concrete solutions for mitigating them.

[5] Asplangyi [5] applied text analytics to assess the impact of online learning on higher education by extracting insights from online platforms and social media. However, the study's heavy reliance on social media data limited its applicability to structured academic data, such as attendance records and grades.

[6] Milliron et al. [6] examined predictive models through case studies in various educational settings. While their approach demonstrated practical applications of analytics in education, their models lacked large-scale empirical validation, making generalization difficult.

[7] Sathe and Adamuthe [7] conducted a comparative study of supervised ML algorithms, including C5.0, J48, CART, Naïve Bayes, Random Forest, and SVM, evaluating predictive accuracy. Their research primarily focused on accuracy, without addressing hybrid models that combine multiple algorithms for enhanced performance.

[8] Agudo-Peregrina et al. [8] studied student interactions in Virtual Learning Environments (VLEs) and analyzed correlations between digital engagement and performance. However, their findings were constrained to online learning settings and lacked applicability to face-to-face classroom interactions.

[9] Salman and Balaram [9] explored real-time student performance prediction using ensemble learning and neural networks. Although their model demonstrated adaptability, the study did not address ethical concerns in AI-driven education.