

AI-Based Dropout Prediction and Counselling System

Selvarani D¹, Aishwarya², Thejas MR² & Umesh R²

¹Assistant Professor, Department of Information Science and Engineering
Sri Sairam College of Engineering, Bengaluru

²Final Year, B.E (Information Science and Engineering), Sri Sairam College of Engineering, Bengaluru

DOI: doi.org/10.34293/iejcsa.v4i5.177

Abstract - Student dropout is a serious concern for higher education institutions because it affects the academic progress of students, the utilisation of institutional resources, and the overall quality of education. In many institutions, at-risk students are identified only after poor attendance, low marks, or backlogs have already accumulated, and counselling is carried out without structured data or follow-up. This work presents an AI-based dropout prediction and counselling system that combines student academic, attendance, and behavioural data with machine learning models and a PostgreSQL database. The system pre-processes the collected data, predicts the dropout probability of each student, classifies the student into a low, medium, or high risk level, and identifies the factors contributing to the prediction using an explainability technique. Based on the risk level, a counselling recommendation is generated, the concerned mentor or counsellor is alerted, and counselling sessions and follow-up actions are recorded in the same student record. The proposed framework therefore connects early identification of at-risk students with timely counselling support in a single workflow, and provides an organised basis for student retention activities in institutions.

Keywords: Student Dropout Prediction, Machine Learning, Counselling System, Early Warning, Explainable AI, PostgreSQL, Student Retention

INTRODUCTION

Student retention is one of the most important indicators of the quality and effectiveness of a higher education institution. When a student discontinues a programme before completion, the effect is felt by the student, the family, and the institution. Dropout is rarely caused by a single event. It usually develops gradually through factors such as declining attendance, poor internal assessment performance, accumulating backlogs, financial difficulty, lack of engagement, and personal or social problems [1], [2]. Because these signs appear well before a student actually leaves, there is an opportunity to identify the student early and provide support.

In most institutions, however, the information related to a student is distributed across attendance registers, mark sheets, fee records, mentor diaries, and separate digital files. Identifying students who are drifting towards dropout therefore depends on the manual effort and personal observation of mentors and class teachers. Such identification is often delayed, and counselling is provided only after the situation has become serious. Even when counselling is carried out, the discussion, the advice given, and the follow-up are often not recorded in a way that can be reviewed later.

Educational data mining and machine learning have shown that student records can be used to predict academic failure and dropout with useful accuracy [2], [3], [4], [5]. Prediction alone, however, does not retain a student. A prediction becomes useful only when it reaches the right person at the right time and is followed by a structured counselling action. A practical system must therefore connect three activities: collecting and organising student data, predicting the risk of dropout, and supporting the counselling process that follows.

The developed system establishes a digital record for each student and connects it with an AI-based prediction module and a counselling module through a PostgreSQL database. The system supports registration of student information, entry of academic and attendance data, prediction of dropout risk, explanation of the main contributing factors, generation of counselling recommendations, and recording of counselling sessions and follow-up status.

The overall objective of this work is to provide an organised method for identifying students at risk of dropout and supporting them through timely counselling. By combining data-driven prediction, explainable risk factors, database-supported record management, and counselling follow-up, the proposed framework helps mentors and counsellors to identify, understand, and support at-risk students more effectively.

LITERATURE REVIEW

The development of a dropout prediction and counselling system requires consideration of both the reasons for student attrition and the technologies available for early identification and support. The literature reviewed for this work was mainly examined in three areas: dropout prediction using educational data mining, machine learning models and their explainability, and digital management of student information and counselling support. These areas were considered to understand the limitations of existing practices and to determine how they could be combined into a single system for student retention.

A. Dropout Prediction Using Educational Data Mining

Early theoretical work explained dropout as a result of the interaction between a student's background, academic performance, and integration into the academic and social life of the institution [1]. Later studies applied data mining to predict school failure and dropout from academic and socio-economic attributes [2], and to predict dropout in distance learning and e-learning courses using classification techniques [3], [4]. Large-scale studies on university records have also shown that enrolment and early performance data can be used to estimate the risk of dropout [5], [6]. Surveys of the field report that academic performance, attendance, and demographic factors are among the most commonly used predictors [7], [8], [9].

B. Machine Learning Models and Explainability

Tree-based ensemble methods such as Random Forest [11] and gradient boosting [12] are widely used for student performance and dropout prediction because they handle mixed numerical and categorical attributes and generally provide reliable accuracy on tabular data. Since dropout cases are much fewer than non-dropout cases, class imbalance is a common problem, and oversampling techniques such as SMOTE are used to improve the detection of the minority class [13]. For practical use by mentors and counsellors, the reason behind a prediction is as important as the prediction itself, and explanation methods such as SHAP can show how each attribute contributes to an individual prediction [14].

C. Digital Student Records and Counselling Support

Learning analytics systems have been used to give early feedback and alerts to students and instructors, and the use of such alerts has been associated with improved student success [10]. Structured databases such as PostgreSQL can be used to maintain student, academic, and counselling information in one place. For the proposed system, PostgreSQL is used to maintain the records associated with each student, the prediction results, and the counselling history.

D. Review Methodology

The review was conducted by examining existing approaches related to dropout prediction, machine learning model selection and explainability, and learning analytics for student support. Their applications and limitations were considered in relation to the requirements of an institutional counselling process. The findings were used to define the requirements of the proposed AI-based prediction and counselling framework.

RESEARCH GAP

Existing research has shown that machine learning models can predict student dropout and academic failure with useful accuracy [2], [3], [5], [6]. Much of this work concentrates on improving the prediction model, comparing algorithms, or identifying important attributes. Separately, learning analytics tools have demonstrated the value of alerts and feedback to students [10]. However, prediction, explanation, counselling action, and follow-up are often considered independently. A student-level system that connects the predicted risk of a student with the explanation of its causes, a recommended counselling action, and a recorded follow-up within a common digital record is less explored. This work addresses this gap by integrating AI-based dropout prediction, explainable risk factors, PostgreSQL data management, and counselling tracking within a single workflow.

Table 1 Research Gap Identified from Existing Work

Existing Research Focus	Research Gap	Proposed Approach
AI-based dropout prediction	Mainly focuses on model accuracy and algorithm comparison	Combines prediction with counselling action in one workflow
Educational data mining	Limited explanation of risk for individual students	Shows contributing factors for every prediction using explainable AI
Student information systems	Academic, attendance, and counselling data may remain separate	Links all student information through a centralised PostgreSQL database
Early alert and analytics tools	Alerts raised without recorded counselling follow-up	Connects risk alerts with counselling sessions and follow-up status

PROBLEM STATEMENT

Institutions enrol large numbers of students, and each student may be affected by different academic, financial, and personal factors. Managing the information required to identify and support students at risk of dropout becomes difficult when records and counselling activities are maintained through conventional methods. The major problems considered in this work are described below.

Educational institutions lack an integrated system to identify at-risk students early because student information is fragmented across multiple records, delaying timely intervention and support.

Traditional identification of at-risk students depends on the observation of mentors and class teachers. A student is usually noticed only after repeated absence, failure in assessments, or a direct request for a transfer or discontinuation. By this stage, the opportunity for effective early intervention may already have been reduced. Information about a student, such as attendance, internal marks, backlogs, fee status, and mentor observations, is often stored in different registers and files. Collecting the complete picture of a single student requires effort and time, and records may be incomplete or outdated when they are not updated consistently.

PROPOSED SYSTEM

The proposed system connects each student with a digital record that is continuously used for dropout prediction and counselling support. The system is designed to simplify student data management, early identification of at-risk students, explanation of risk factors, and tracking of counselling activities. PostgreSQL is used to store the information associated with each student, while an AI-based prediction module analyses the data and estimates the probability of dropout.

A. AI-Based Integrated Student Monitoring, Dropout Prediction and Risk Assessment System

A unique Student_ID identifies each student in the system. Profile details such as department, semester, and admission category are entered at registration. Academic data

such as internal marks, semester grade point average, and number of backlogs, attendance data, and behavioural indicators such as participation in activities, learning-platform engagement, and fee payment status are updated periodically. Mentors and counsellors can also add their observations to the same record. The prediction module uses supervised machine learning models, such as Logistic Regression, Support Vector Machine, Random Forest [11], and gradient boosting [12], that are trained on historical student data labelled as dropout or continuing. The model that gives the best performance during validation is selected for deployment. For each student, the module produces a dropout probability, which is stored with the corresponding student record. The dropout probability is converted into a risk level using configurable thresholds, so that a student is classified as low, medium, or high risk. To support the counsellor, the system also identifies the attributes that contributed most to the prediction of an individual student using SHAP-based explanation [14]. For example, the explanation may show that low attendance and an increasing number of backlogs are the main reasons for the high risk of a student.

SYSTEM ARCHITECTURE

The overall architecture of the proposed system is shown in the block diagram below.

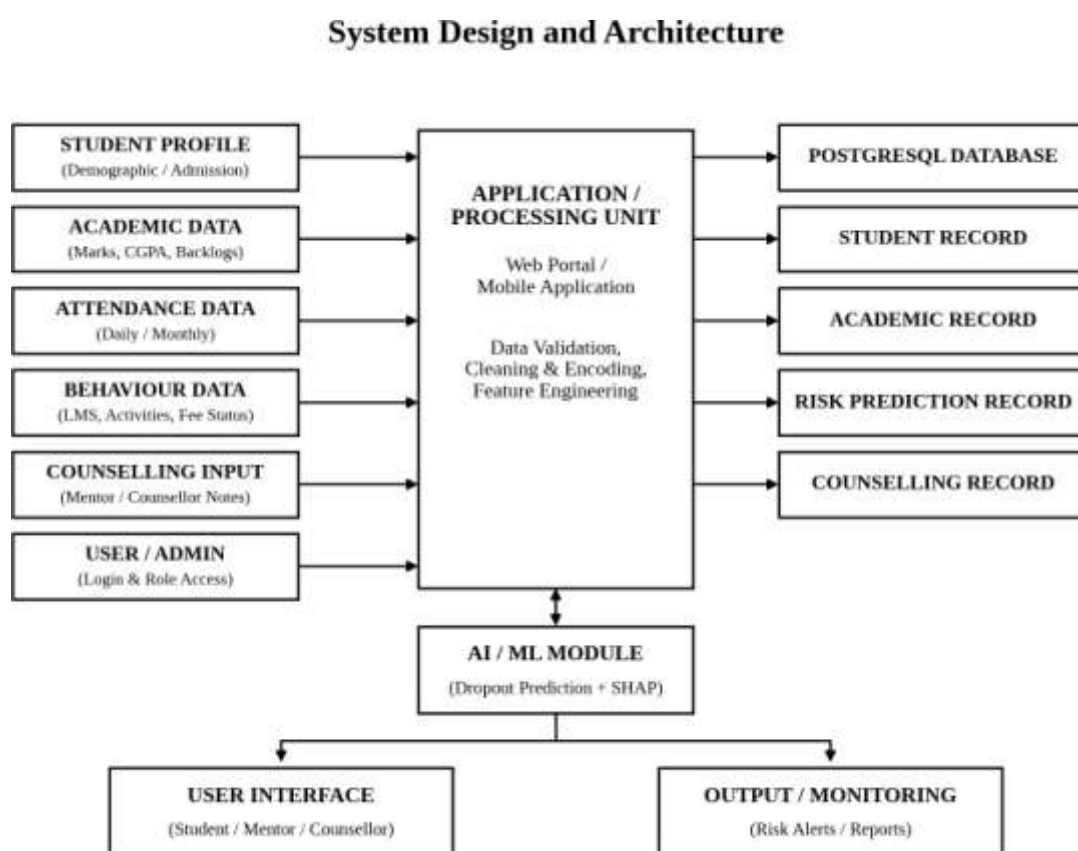


Fig. 1. Block Diagram of Proposed System

The architecture consists of input, processing, storage, intelligence, and output layers. Student profile, academic, attendance, behavioural, and counselling inputs are

entered by authorised users through the web portal or mobile application, which acts as the application and processing unit. This unit validates and pre-processes the data and stores it in the PostgreSQL database as student, academic, risk prediction, and counselling records. The AI/ML module exchanges data with the processing unit in both directions: it receives the processed features, returns the dropout probability and risk factors, and the results are stored in the database. The user interface and the output/monitoring module present risk alerts, dashboards, and reports to students, mentors, counsellors, and administrators.

Table 2 Representative Database Structure

Entity	Representative Fields	Purpose
Student	Student_ID, Name, Department, Semester, Admission_Category	Stores student identity and profile details
Academic_Record	Record_ID, Student_ID, CGPA, Internal_Marks, Attendance_Pct, Backlogs	Maintains academic and attendance information
Risk_Prediction_Record	Prediction_ID, Student_ID, Risk_Score, Risk_Level, Top_Factors, Prediction_Date	Stores AI-generated risk results and explanations
Counselling_Record	Session_ID, Student_ID, Counsellor_ID, Session_Date, Action_Plan, Follow_Up_Status	Stores counselling sessions and follow-up

This architecture provides a digital link between individual students and their associated academic, prediction, and counselling information, supporting organised tracking and AI-assisted counselling within the same workflow.

METHODOLOGY/WORKFLOW

The proposed system connects each student with a digital record that supports prediction and counselling. Initially, the student profile, academic, attendance, and behavioural details are entered into the system and stored in the PostgreSQL database. The data is pre-processed, and the trained model predicts the dropout probability of the student. The risk level and the main contributing factors are stored with the student record.

If the student is classified as medium or high risk, a counselling recommendation is generated and the mentor or counsellor is alerted. The counselling session and its outcome are recorded in the same record. In the next cycle, the data is updated and the student is assessed again, allowing the academic, risk, and counselling information to be maintained together throughout the student's programme.

The proposed system follows the sequence shown in Fig. 2.

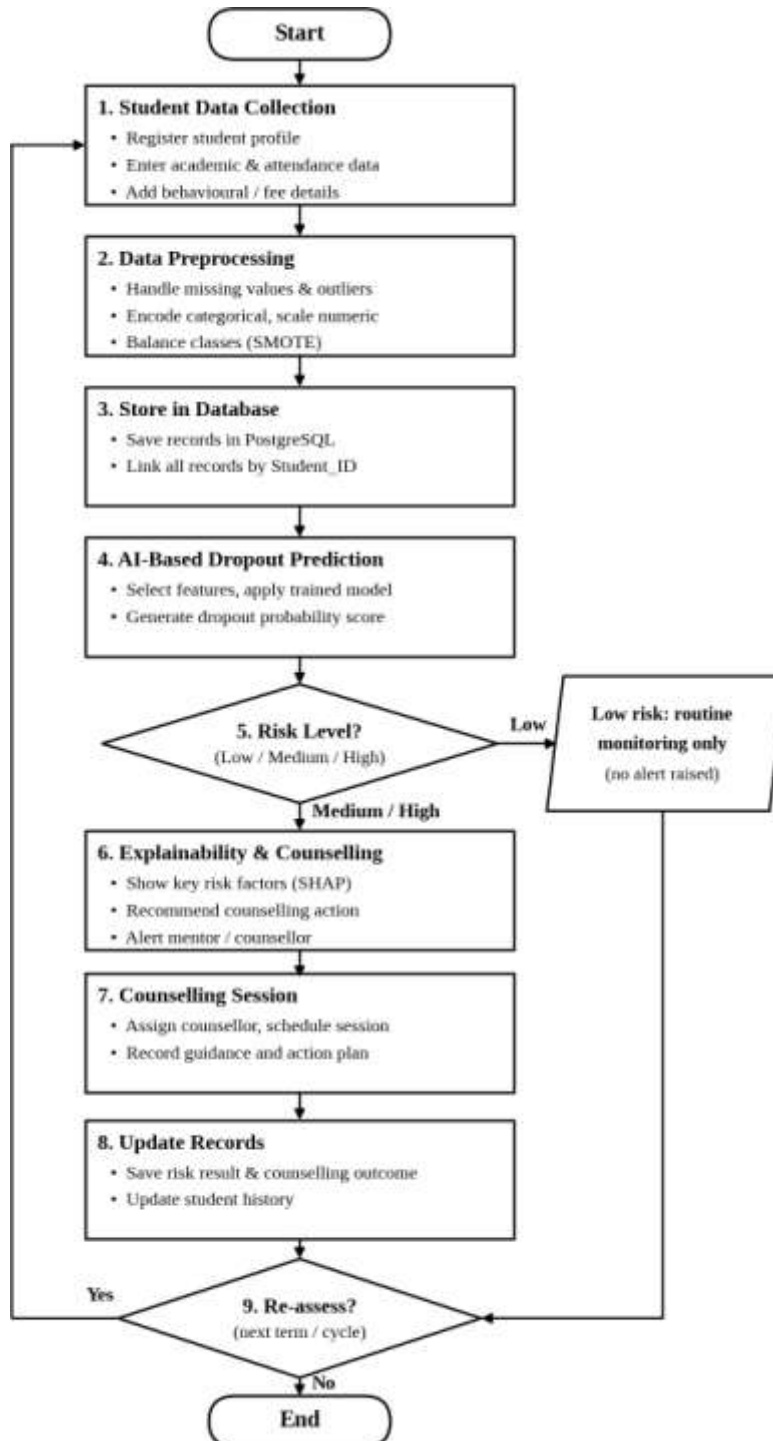


Fig. 2. Flow chart of the proposed solution.

The workflow shown in Fig. 2 is carried out in the following stages.

1. Student data collection: the student profile, academic performance, attendance, and behavioural and fee details are entered or imported into the system.
2. Data pre-processing: missing values and outliers are handled, categorical attributes are encoded, numerical attributes are scaled, and class imbalance in the training data is addressed.

3. Database storage: the processed records are stored in the PostgreSQL database and linked through the Student_ID.
4. AI-based dropout prediction: relevant features are selected and the trained model generates a dropout probability for the student.
5. Risk classification: the probability is converted into a low, medium, or high risk level. Students at low risk continue under routine monitoring, while medium and high risk students are taken to the next stage.
6. Explainability and counselling recommendation: the main factors contributing to the risk are identified, a counselling recommendation is generated, and the mentor or counsellor is alerted.
7. Counselling session: a counsellor is assigned, the session is scheduled, and the guidance and action plan are recorded.
8. Record update: the risk result, counselling outcome, and follow-up status are saved in the record of the same student.
9. Re-assessment: if the next term or assessment cycle begins, the process returns to data collection with updated information; otherwise it ends.

IMPLEMENTATION AND EVALUATION

A. Implementation

The developed system combines student data management, AI-based dropout prediction, explainability, and counselling tracking within a common workflow. The application layer is provided as a web portal that can also be accessed from a mobile device. The prediction module is developed using Python-based machine learning libraries [15], and the relational data is stored in a PostgreSQL database.

The student data module allows authorised users to register students and to enter or update academic, attendance, and behavioural information. Each student is identified by a unique Student_ID, and all records of the student are retrieved through this identifier. The pre-processing component prepares the data for the model by cleaning, encoding, scaling, and deriving additional features from the stored records.

The prediction module is trained on historical student data, as shown in Fig. 3. Candidate models are compared on a held-out portion of the data, and the best-performing model is used to generate the dropout probability and risk level for current students. The explainability component lists the attributes that contribute most to each prediction, and these are displayed along with the risk level.

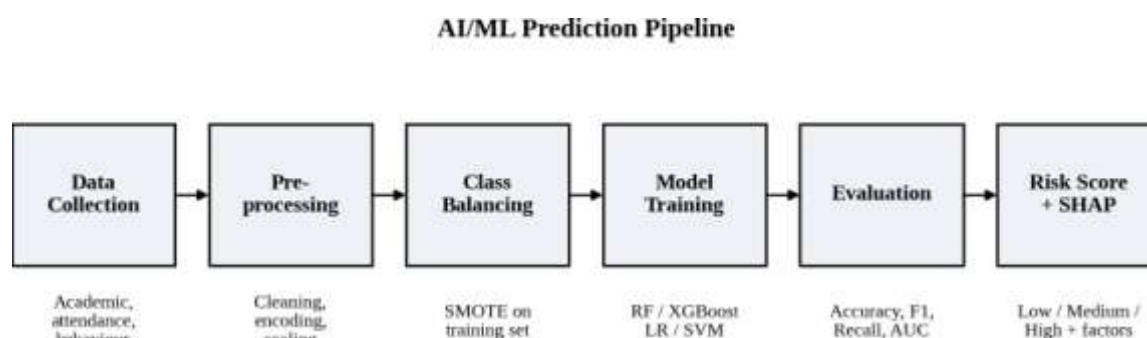


Figure 3 AI/ML prediction pipeline of the proposed system.

The counselling module receives the risk result and generates a recommendation for medium and high risk students. It supports assignment of a counsellor, scheduling of sessions, recording of the discussion and action plan, and updating of the follow-up status. The dashboard provides mentors and administrators with a summary of the number of students in each risk category, the pending counselling sessions, and the progress of students who have been counselled.

B. Evaluation and Discussion

The prediction models are evaluated using a stratified split of the dataset into training and testing portions, with cross-validation used for model selection. Since the cost of missing a student who actually drops out is higher than the cost of an unnecessary counselling session, recall of the dropout class is considered together with accuracy, precision, F1-score, and the area under the ROC curve. Resampling is applied only to the training portion so that the test results are not affected. Table III shows the format in which the comparison of models is reported.

Table 3 Performance Comparison of Candidate Models (to Be Filled from Experimental Results)

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	[]	[]	[]	[]	[]
Support Vector Machine	[]	[]	[]	[]	[]
Random Forest	[]	[]	[]	[]	[]
Gradient Boosting (XGBoost)	[]	[]	[]	[]	[]

The system makes the information of each student easier to access and organise. By keeping academic, prediction, and counselling data under the same identifier, it avoids the need to search through separate registers, and gives the counsellor a more complete view of the student before a session.

The explainability component further extends the system by showing the reasons behind each prediction. Instead of receiving only a risk label, the counsellor can see, for

example, that low attendance and declining internal marks are the main contributors, and can direct the discussion towards these problems. The recorded counselling history also allows the institution to review whether counselling was provided and how the risk level of the student changed in later cycles.

Overall, the proposed framework provides a common platform for student data management, dropout prediction, risk explanation, and counselling support. The AI module provides early identification, while the counselling module ensures that the prediction leads to a recorded action.

C. Scalability and Applicability

The proposed architecture is designed to support different departments, programmes, and institutions while maintaining a common data-management process. The use of a unique Student_ID provides a consistent link between each student and the associated records, allowing the same framework to be applied across batches and departments.

The database structure can accommodate additional information such as co-curricular participation, library usage, hostel details, and learning-platform activity, and the prediction model can be retrained as more data becomes available. This allows the system to be extended to a larger number of students without changing the basic workflow.

D. Ethical and Privacy Considerations

A dropout prediction system uses sensitive personal data, and its output may influence how a student is treated. The risk level is therefore intended only as support for counselling and not as a basis for any disciplinary or admission decision. Access to data is restricted by role, and the data is stored securely. The model should also be checked periodically for unfair behaviour towards particular groups of students, and the final decision on counselling should remain with the human counsellor.

CONCLUSION

The developed system provides a digital approach for identifying students at risk of dropout and supporting them through counselling. Each student is associated with a structured digital record containing profile, academic, prediction, and counselling information. PostgreSQL provides centralised storage for the associated data, while the AI-based prediction module estimates the dropout risk and the explainability component shows the factors behind it.

The main contribution of the system is the connection established between early prediction and counselling action. Instead of treating data collection, prediction, and counselling as separate activities, the developed workflow associates them with a common student record. The framework provides a foundation for validation with institutional data, larger-scale deployment, and integration with existing institutional systems.

Future work will focus on extending the prototype for practical institutional use. The planned improvements include training and validating the model on larger and more diverse multi-year datasets, integrating the system with the institutional ERP and learning management system for automatic data updates, adding automated SMS and e-mail alerts to mentors and parents, introducing a chatbot for first-level student support, and conducting fairness and bias audits of the model. Further work can include analysis of the effectiveness of counselling, department-level retention analytics, and scalable deployment across multiple institutions.

REFERENCES

- [1] V. Tinto, "Dropout from higher education: A theoretical synthesis of recent research," *Review of Educational Research*, vol. 45, no. 1, pp. 89–125, 1975.
- [2] C. Márquez-Vera, C. R. Morales, and S. V. Soto, "Predicting school failure and dropout by using data mining techniques," *IEEE Revista Iberoamericana de Tecnologías del Aprendizaje*, vol. 8, no. 1, pp. 7–14, 2013.
- [3] S. Kotsiantis, C. Pierrakeas, and P. Pintelas, "Preventing student dropout in distance learning using machine learning techniques," in *Knowledge-Based Intelligent Information and Engineering Systems (KES 2003)*, *Lecture Notes in Computer Science*, vol. 2774, Springer, 2003, pp. 267–274.
- [4] I. Lykourantzou, I. Giannoukos, V. Nikolopoulos, G. Mpardis, and V. Loumos, "Dropout prediction in e-learning courses through the combination of machine learning techniques," *Computers & Education*, vol. 53, no. 3, pp. 950–965, 2009.
- [5] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting student dropout in higher education," *arXiv preprint arXiv:1606.06364*, 2016.
- [6] V. Realinho, J. Machado, L. Baptista, and M. V. Martins, "Predicting student dropout and academic success," *Data*, vol. 7, no. 11, Art. no. 146, 2022.
- [7] E. Alyahyan and D. Düştégör, "Predicting academic success in higher education: literature review and best practices," *International Journal of Educational Technology in Higher Education*, vol. 17, Art. no. 3, 2020.
- [8] R. S. J. d. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.
- [9] C. Romero and S. Ventura, "Educational data mining: A review of the state of the art," *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, vol. 40, no. 6, pp. 601–618, 2010.
- [10] K. E. Arnold and M. D. Pistilli, "Course Signals at Purdue: Using learning analytics to increase student success," in *Proc. 2nd Int. Conf. on Learning Analytics and Knowledge (LAK '12)*, 2012, pp. 267–270.
- [11] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

- [12] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
- [13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.
- [14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems 30, 2017, pp. 4765–4774.
- [15] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.