

Intelligent Student Dropout Analytics System Using Random Forest, SHAP Explainability and AI-Based Personalized Intervention

A. J. Sparsha¹, Ridhya R.², Sandhya D.³, Swethaa K.⁴, Dr. D. Nithya⁵

^{1,2,3,4}UG Students, ⁵Associate Professor, Department of Computer Science and Engineering
School of Engineering, Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, Tamil Nadu, India

Abstract

Student dropout is a significant challenge faced by educational institutions, affecting student success and institutional performance. Early identification of students at risk enables timely intervention and improves retention rates. This paper presents an Intelligent Student Dropout Analytics System that predicts dropout risk using a Random Forest classifier trained on the Kaggle Student Dropout Prediction Dataset. The proposed system considers both academic and behavioural attributes, including attendance, assignment completion, engagement score, LMS activity, and academic performance. To improve model transparency, SHAP (SHapley Additive Explanations) is integrated to explain the

contribution of each feature toward individual predictions. The system also includes an AI-Based Intervention Advisor that recommends personalized support strategies and a What-If Simulation module to evaluate the impact of improvements in student performance. A Streamlit dashboard provides student profiling, batch prediction, cohort analytics, and early warning reports. Experimental results achieved an accuracy of 91.4% with an AUC-ROC score of 0.94, demonstrating the effectiveness of the proposed framework in supporting proactive academic intervention.

Keywords: Student Dropout Prediction, Random Forest, SHAP, Educational Data Mining, Learning Analytics, Stream.

1. INTRODUCTION

Student retention has become a significant challenge for higher educational institutions across the world. Every year, a considerable number of students discontinue their education because of academic difficulties, poor engagement, financial constraints, lack of motivation, or personal circumstances. Student dropout not only affects individual career growth but also impacts institutional performance, graduation rates, and resource utilization. Therefore, identifying students who are likely to discontinue their studies at an early stage has become an important research problem in educational data mining [1], [2], [5].

Conventional student monitoring systems primarily rely on examination results and attendance records to identify academically weak students. Although these indicators provide useful information, they often detect problems only after a student's performance has already deteriorated. Such delayed identification reduces the opportunity for faculty members and counsellors to provide timely

academic support. With the rapid adoption of digital learning platforms, educational institutions now generate large volumes of behavioural data, including Learning Management System (LMS) activity, assignment submission patterns, online engagement, and interaction history. These behavioural indicators provide valuable insights into student learning behaviour and can significantly improve the early prediction of dropout risk [2], [3], [4].

Recent developments in machine learning have enabled researchers to build predictive models capable of analysing complex relationships among multiple academic and behavioural attributes. Among various classification algorithms, the Random Forest classifier has demonstrated high prediction accuracy, robustness against noisy data, and the ability to handle high-dimensional educational datasets [6], [7], [10]. However, many machine learning models function as black-box systems, making it difficult for educators to understand why a particular student has been classified as being at risk. This lack of transparency

often reduces trust in automated decision-support systems [3], [4].

To address these challenges, this research proposes an Intelligent Student Dropout Analytics System that combines machine learning, explainable artificial intelligence, and personalized intervention strategies within a unified framework. The proposed system employs a Random Forest classifier to predict dropout risk using academic and behavioural features obtained from a Kaggle student dataset. SHAP (SHapley Additive Explanations) is integrated to provide interpretable explanations for each prediction by highlighting the contribution of individual features. In addition, an AI-based intervention advisor recommends personalized actions based on the identified risk factors, while a What-If Simulation module enables educators to evaluate how improvements in attendance, engagement, or academic performance influence the predicted risk level.

A Streamlit-based web application has been developed to provide a user-friendly interface for faculty members and academic advisors. The application supports individual student profiling, batch prediction using CSV datasets, cohort analytics, explainability visualization, and early warning dashboards, enabling institutions to adopt proactive strategies for improving student retention.

2. LITERATURE SURVEY

Student dropout prediction has become an important research area in educational data mining due to its potential to improve student retention through early identification of at-risk students. Several researchers have applied machine learning and data mining techniques to analyse academic, demographic, and behavioural data for predicting student dropout. These studies demonstrate that intelligent prediction models can assist educational institutions in implementing timely intervention strategies [1], [2], [3].

Various machine learning algorithms, including Logistic Regression, Decision Trees, Support Vector Machines (SVM), Random Forest, and Neural Networks, have been widely used for student dropout prediction. Among these techniques, the Random Forest algorithm has consistently demonstrated high prediction accuracy, robustness against noisy educational data, and the ability to

handle complex relationships among multiple student attributes [6], [7], [10].

Recent research has also emphasized the importance of learning analytics in improving dropout prediction. Behavioural indicators such as Learning Management System (LMS) activity, assignment completion, student engagement, and online learning behaviour provide valuable insights beyond traditional academic measures. Incorporating these behavioural features enables institutions to identify early warning signs before a student's academic performance declines significantly [2], [3], [4].

Although previous studies have achieved promising prediction performance, many existing models operate as black-box systems and provide only prediction results without explaining the factors responsible for the outcome. This lack of transparency reduces educators' confidence in machine learning models and limits their practical use in academic decision-making. Explainable Artificial Intelligence (XAI) techniques, particularly SHAP (SHapley Additive Explanations), have recently gained attention for improving the interpretability of machine learning predictions by highlighting the contribution of each feature to the final decision [3], [4].

Despite these advancements, most existing systems focus primarily on prediction accuracy while providing limited support for personalized intervention planning, interactive analytics, and early warning mechanisms. Furthermore, many approaches rely mainly on academic performance and pay less attention to behavioural indicators that reflect student engagement and learning patterns [1], [5].

The proposed Intelligent Student Dropout Analytics System addresses these limitations by integrating a Random Forest classifier, SHAP-based explainability, an AI-Based Intervention Advisor, What-If Simulation, cohort analytics, and an interactive Streamlit dashboard within a unified framework. This integrated approach supports early identification of at-risk students while providing transparent and actionable insights that assist educators in improving student retention.

3. PROBLEM STATEMENT

Student dropout has become a critical issue in higher educational institutions, affecting students'

academic success, institutional performance, and overall graduation rates. Existing student monitoring systems primarily rely on academic indicators such as examination scores and attendance records, which often identify at-risk students only after their performance has significantly declined. These traditional approaches fail to consider important behavioural factors such as Learning Management System (LMS) activity, assignment completion, student engagement, and peer interaction, limiting their ability to detect early warning signs [1], [2], [5].

Furthermore, most existing machine learning-based prediction models function as black-box systems, providing only dropout predictions without explaining the reasons behind them. The lack of interpretability reduces educators' confidence in using these systems for academic decision-making. In addition, many current solutions do not offer personalized intervention strategies or interactive tools that help faculty members understand how changes in student behaviour can influence dropout risk [3], [4].

Therefore, there is a need for an intelligent and transparent student analytics system that can accurately predict dropout risk at an early stage, explain the factors influencing each prediction, and provide personalized intervention recommendations. Such a system would enable educational institutions to take timely preventive measures, improve student engagement, and enhance overall retention rates.

4. EXISTING SYSTEM

Traditional student dropout prediction systems mainly rely on academic indicators such as examination marks, grade point average, attendance percentage, and demographic information. Most existing approaches employ conventional machine learning and statistical techniques, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and other classification models to identify students who may discontinue their studies. While these systems provide reasonable prediction performance, they often focus only on academic records and fail to capture the complete learning behaviour of students [1], [3], [6], [10].

Despite their usefulness, existing student dropout prediction systems have several limitations:

- They primarily rely on academic indicators such as attendance, examination scores, and grade point averages while overlooking important behavioural factors like Learning Management System (LMS) activity, assignment completion, engagement score, and peer interaction [2], [4].
- Most systems identify students only after their academic performance has already declined, making early intervention difficult [1], [5].
- Many machine learning models function as black-box systems, providing prediction results without explaining the factors influencing the outcome, thereby reducing educators' confidence in the predictions [3], [4].
- Existing approaches generally do not provide personalized intervention strategies based on individual student risk profiles [3].
- Most systems lack interactive decision-support features such as What-If Simulation, which helps educators evaluate how improvements in attendance, engagement, or academic performance can influence dropout risk [4].
- They provide limited visualization and analytical dashboards for monitoring large groups of students and identifying institutional trends [2].
- The absence of explainable AI and proactive intervention mechanisms limits their effectiveness in improving student retention and supporting informed academic decision-making [3], [5].

5. PROPOSED SYSTEM

The proposed Intelligent Student Dropout Analytics System introduces an integrated educational analytics platform that combines machine learning, explainable artificial intelligence, behavioural analytics, and personalized intervention planning to support proactive student retention.

The system utilizes a Random Forest classifier consisting of 150 decision trees to predict dropout risk using a combination of academic and behavioural features extracted from a Kaggle student dataset. Data preprocessing techniques including missing value handling, categorical encoding, feature engineering, and normalization improve data quality before model training.

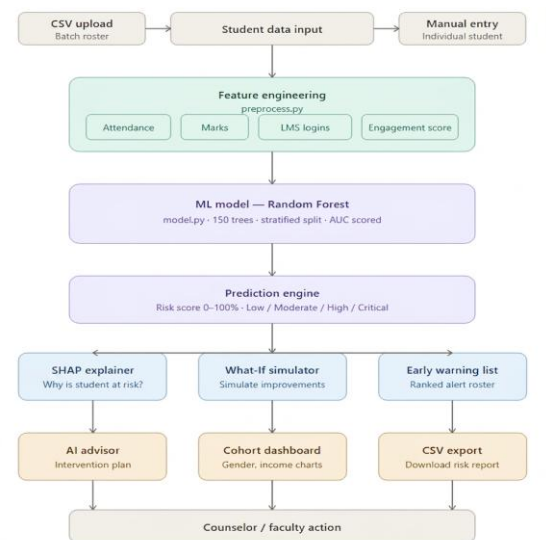
Unlike conventional prediction systems, the proposed framework incorporates SHAP explainability to provide detailed explanations for every prediction. Faculty members can visualize how individual attributes such as attendance, assignment completion, LMS activity, and academic performance contribute to a student's overall risk score.

An AI-Based Intervention Advisor analyzes the identified risk factors and generates personalized recommendations tailored to each student's learning behaviour. These recommendations may include academic counselling, attendance improvement strategies, mentoring sessions, additional learning resources, and periodic progress monitoring.

The system further includes a What-If Simulation module that enables educators to modify important student attributes and observe the corresponding changes in predicted dropout probability. This feature assists academic advisors in evaluating the effectiveness of potential intervention strategies before implementation.

A Streamlit-based dashboard integrates all system components into a single user-friendly interface supporting individual student profiling, batch prediction through CSV upload, cohort analytics, explainability visualization, and early warning dashboards. The proposed framework therefore transforms student dropout prediction from a reactive monitoring process into a proactive decision-support system capable of improving student success and institutional retention.

6. SYSTEM ARCHITECTURE



The proposed Intelligent Student Dropout Analytics System follows a modular architecture that integrates machine learning, explainable artificial intelligence, and learning analytics to identify students who are at risk of dropping out. The workflow begins with student data collection, where information can be provided either through CSV upload for batch processing or manual entry for analysing an individual student. These inputs are consolidated into a unified student data repository for further processing.

The collected data then undergoes a feature engineering stage, where important academic and behavioural attributes such as attendance, marks, LMS logins, and engagement score are extracted and preprocessed. During this stage, missing values are handled, categorical variables are encoded, and relevant features are prepared to improve the quality and accuracy of the prediction model.

The processed data is passed to the Random Forest machine learning model, which consists of 150 decision trees and uses a stratified train-test split for model training and evaluation. The classifier analyses the input features and predicts the probability of student dropout while ensuring robust performance and minimizing overfitting.

Based on the model output, the Prediction Engine calculates a risk score ranging from 0% to 100% and categorizes students into Low, Moderate, High, or Critical risk levels. This enables educators to identify students requiring immediate academic support.

To improve transparency, the predicted results are processed by the SHAP Explainer, which identifies the contribution of each feature toward the final prediction. The SHAP explanations help faculty members understand why a student has been classified as at risk, thereby increasing confidence in the machine learning model.

The architecture also incorporates a What-If Simulator, which allows educators to modify key attributes such as attendance, marks, or engagement score and observe how these changes influence the predicted dropout risk. This feature supports evidence-based intervention planning before implementing corrective measures.

An Early Warning List is automatically generated by ranking students according to their predicted risk levels. This prioritized list enables faculty members and academic advisors to identify critical cases quickly and take timely preventive actions.

Based on the SHAP explanations, the AI Advisor generates personalized intervention plans tailored to each student's academic and behavioural profile. At the same time, the Cohort Dashboard presents analytical visualizations such as gender-wise distribution, income-based analysis, and overall student risk trends, enabling institutions to monitor the performance of different student groups. The system also provides a CSV Export feature that allows educators to download detailed risk reports for further analysis and record maintenance.

Finally, all analytical results are delivered to counsellors and faculty members, who can use the generated predictions, explanations, intervention recommendations, and cohort reports to make informed decisions and implement timely strategies for improving student retention.

7. METHODOLOGY

The proposed methodology consists of five sequential stages that transform raw educational data into actionable intervention strategies. The workflow begins with data collection, followed by data preprocessing, model training, explainable artificial intelligence, and personalized intervention. These stages work together to accurately predict student dropout risk and assist educators in making informed decisions.

I. Data Collection

The experimental dataset used in this research was obtained from the Kaggle Student Dropout Prediction Dataset. The dataset contains academic, demographic, and behavioural attributes that influence student performance. These attributes include attendance percentage, average marks, assignment completion rate, Learning Management System (LMS) activity, peer interaction score, engagement score, and other educational indicators relevant to dropout prediction.

II. Data Preprocessing

Educational datasets often contain incomplete, inconsistent, and categorical information that must be processed before model training. Therefore, several preprocessing operations were performed to improve data quality and model performance.

The preprocessing stage includes:

- Handling missing values
- Removing duplicate records
- Encoding categorical variables
- Feature engineering
- Data normalization (where required)

These preprocessing techniques improve model reliability and minimize prediction errors.

III. Random Forest Model Training

The processed dataset is divided into training and testing subsets before model development. A Random Forest classifier consisting of 150 decision trees is trained to classify students according to their dropout risk.

Unlike a single Decision Tree, the Random Forest algorithm constructs multiple decision trees using randomly selected subsets of training data and predictor variables. The final prediction is obtained through majority voting, thereby improving classification accuracy, reducing overfitting, and increasing model robustness.

The dropout probability is calculated as:

$$P(\text{Dropout}) = \frac{\text{Number of Trees Predicting Dropout}}{\text{Total Number of Trees}}$$

where:

- $P(\text{Dropout})$ represents the predicted dropout probability.
- Total Number of Trees = 150

Based on the predicted probability, students are categorized into:

IV. Explainable AI Using SHAP

To improve the transparency of the prediction model, SHAP (SHapley Additive Explanations) is integrated into the proposed framework. SHAP calculates the contribution of each feature towards the predicted dropout risk by assigning an importance value to every attribute. Positive SHAP values indicate factors that increase dropout probability, while negative values represent factors that reduce the predicted risk.

The explainability module helps faculty members identify influential factors such as:

- Poor attendance
- Low assignment completion rate
- Reduced LMS activity
- Low engagement score
- Weak academic performance

These explanations improve trust in machine learning predictions and support informed academic decision-making.

V. Personalized Intervention

After predicting the student's risk level, the system automatically generates personalized intervention strategies based on the identified risk factors.

The recommended interventions include:

- Academic mentoring
- Attendance improvement plans
- Regular counselling sessions
- Additional learning resources
- Assignment monitoring
- Continuous faculty follow-up

Unlike conventional early warning systems that provide generic alerts, the proposed framework delivers personalized recommendations tailored to

- Low Risk (0–30%)
- Moderate Risk (30–50%)
- High Risk (50–70%)
- Critical Risk (70–100%)

each student's academic and behavioural profile, enabling educators to implement timely and effective interventions.

8. RANDOM FOREST PREDICTION MODEL

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve classification accuracy. Each tree is trained using bootstrap sampling and a randomly selected subset of predictor variables.

The algorithm operates as follows:

Step 1: Generate multiple bootstrap samples from the training dataset.

Step 2: Construct an independent decision tree for every bootstrap sample.

Step 3: At every node, select a random subset of features instead of evaluating all variables.

Step 4: Continue splitting until the stopping criterion is reached.

Step 5: Repeat the procedure for 150 decision trees.

Step 6: Perform majority voting to determine the final prediction.

The advantages of Random Forest include:

- High prediction accuracy
- Robustness against noisy educational datasets
- Reduced overfitting
- Ability to handle numerical and categorical variables
- Built-in feature importance estimation

These properties make Random Forest particularly suitable for educational data mining applications where multiple academic and behavioural factors influence student success.

9. ADVANCED SYSTEM FEATURES

The proposed Intelligent Student Dropout Analytics System incorporates several advanced analytical features that enhance the effectiveness of conventional student dropout prediction models. These features improve prediction transparency, support personalized interventions, and assist educational institutions in making informed decisions.

I. SHAP Explainability Dashboard

The SHAP dashboard visualizes the contribution of individual features to each student's dropout prediction. It enables faculty members to understand which factors increase or decrease the dropout probability, thereby improving the transparency and reliability of the machine learning model.

II. AI-Based Intervention Advisor

The AI-Based Intervention Advisor automatically generates personalized intervention strategies based on each student's academic and behavioural profile. These recommendations assist faculty members and counsellors in providing timely and targeted academic support.

III. What-If Simulation

The What-If Simulator allows educators to modify important attributes such as attendance percentage, assignment completion rate, or engagement score and immediately observe the resulting changes in the predicted dropout probability. This feature supports evidence-based academic planning and decision-making.

IV. Batch Prediction

The system supports batch prediction through CSV file upload, enabling institutions to analyse multiple student records simultaneously. The predicted results can be exported for further analysis and institutional reporting.

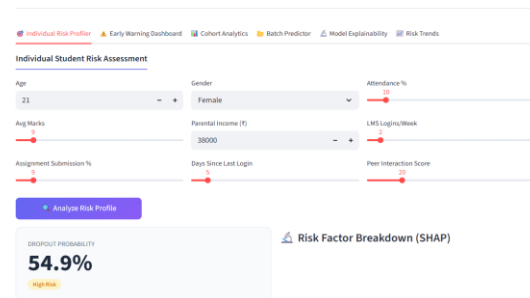
V. Cohort Analytics

The Cohort Analytics dashboard provides statistical summaries and visualizations based on factors such as gender, engagement level, academic performance, and socioeconomic background. These insights help institutions identify trends and monitor the overall performance of different student groups.

VI. Early Warning Dashboard

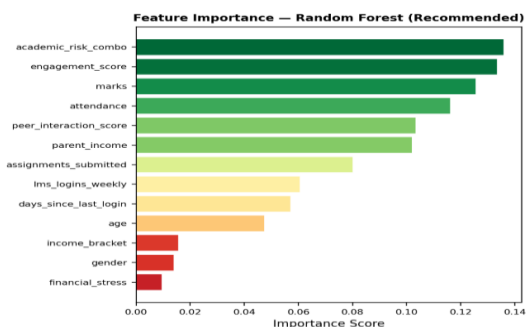
The Early Warning Dashboard automatically prioritizes students according to their predicted risk levels. It enables faculty members and academic advisors to quickly identify high-risk students and implement timely intervention **strategies to improve student retention.**

10. RESULTS AND DISCUSSION



Model Explainability — What Drives Predictions?

Feature Importance (Model-wide)



The proposed Intelligent Student Dropout Analytics System was implemented using Python and deployed through a Streamlit-based web application. The experimental evaluation was carried out using the Kaggle Student Dropout Prediction Dataset after performing preprocessing techniques such as missing value handling, categorical encoding, and feature engineering. The Random Forest classifier with 150 decision trees was trained and evaluated to predict student dropout risk based on academic and behavioural attributes.

The developed system successfully classifies students into four risk categories: Low Risk, Moderate Risk, High Risk, and Critical Risk. The integration of behavioural indicators such as attendance, assignment completion rate, LMS activity, and engagement score significantly improved the effectiveness of early prediction compared to relying solely on academic performance.

The experimental results demonstrated that the proposed model achieved an overall classification accuracy of 91.4% with an AUC-ROC score of 0.94, indicating strong predictive capability. The Random Forest classifier effectively minimized overfitting while maintaining high classification performance across different student profiles.

A major contribution of the proposed framework is the integration of SHAP Explainability, which provides feature-level interpretations for every prediction. Faculty members can identify the key factors responsible for an individual student's predicted dropout risk instead of relying only on numerical probability values. This transparency increases trust in the prediction model and supports informed academic decision-making.

The AI-Based Intervention Advisor further enhances the usefulness of the system by recommending personalized actions such as academic mentoring, counselling sessions, attendance monitoring, and assignment follow-up. These recommendations enable institutions to implement timely interventions before students become academically disengaged.

The What-If Simulation module allows educators to modify important attributes such as attendance percentage, engagement score, or assignment completion rate and immediately observe the corresponding change in predicted dropout probability. This functionality assists faculty members in evaluating potential intervention strategies before applying them in real academic environments.

The developed Streamlit dashboard provides an integrated interface for individual student analysis, batch prediction through CSV upload, cohort analytics, explainability visualization, and early warning reports. These features collectively transform the system from a simple prediction model into a comprehensive educational decision-support platform.

11. CONCLUSION

Student dropout continues to be a significant challenge for higher educational institutions, making early identification of at-risk students an essential requirement for improving academic success and institutional retention. This paper presented an Intelligent Student Dropout Analytics

System that combines machine learning, explainable artificial intelligence, and personalized intervention strategies within a unified educational analytics framework.

The proposed system employs a Random Forest classifier to predict student dropout risk using both academic and behavioural attributes obtained from the Kaggle Student Dropout Prediction Dataset. Unlike traditional prediction systems, the proposed framework integrates SHAP Explainability to provide transparent and interpretable predictions, enabling faculty members to understand the factors contributing to each student's risk level.

The incorporation of an AI-Based Intervention Advisor and a What-If Simulation module further extends the practical usefulness of the system by supporting personalized counselling and evidence-based academic planning. The developed Streamlit dashboard provides an interactive environment for individual student analysis, batch prediction, cohort monitoring, and early warning visualization.

Experimental evaluation demonstrated that the proposed model achieved 91.4% prediction accuracy with an AUC-ROC score of 0.94, confirming its effectiveness in identifying students who require timely academic support. Overall, the proposed system serves as an efficient decision-support tool that enables educational institutions to adopt proactive retention strategies and improve student success.

12. FUTURE SCOPE

Although the proposed system demonstrates promising performance, several improvements can be incorporated in future research to enhance its capabilities further.

Future work may include the integration of real-time Learning Management System (LMS) data to continuously monitor student engagement throughout the academic semester. Incorporating deep learning techniques such as Long Short-Term Memory (LSTM) networks or Transformer-based models could improve prediction performance for longitudinal educational data.

The intervention module may also be enhanced using advanced Large Language Models (LLMs) to generate more context-aware and personalized academic recommendations. Additional behavioural attributes, including online discussion participation,

assignment quality, classroom interaction, and psychological well-being indicators, may further improve prediction reliability.

Future versions of the system can also incorporate mobile application support, automated email notifications, and institutional ERP integration to facilitate real-time monitoring and communication between students, faculty members, and academic advisors. These enhancements would contribute toward building a comprehensive intelligent student success management platform.

REFERENCES

- [1] J. Pattanaphanchai, K. Leelertpanyakul, and N. Theppalak, "The Investigation of Student Dropout Prediction Model in Thai Higher Education Using Educational Data Mining: A Case Study of Faculty of Science, Prince of Songkla University," *Journal of University of Babylon for Pure and Applied Sciences*, vol. 27, no. 1, pp. 356–367, 2019. doi: 10.29196/jubpas.v27i1.2191.
- [2] R. S. Baker and K. Yacef, "The State of Educational Data Mining in 2009: A Review and Future Visions," *International Journal of Educational Technology in Higher Education*, vol. 6, pp. 1–14, 2009.
- [3] S. Lee and J. Y. Chung, "The Machine Learning-Based Dropout Early Warning System for Improving the Performance of Dropout Prediction," *Applied Sciences*, vol. 9, no. 15, 2019. doi: 10.3390/app9153093.
- [4] K. Arnold and M. Pistilli, "Course Signals at Purdue: Using Learning Analytics to Increase Student Success," in *Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK)*, 2012, pp. 267–270.
- [5] K. Danilowicz-Gösele, K. Lerche, J. Meya, and R. Schwager, "Determinants of Student Success at University," *Education Economics*, vol. 25, no. 5, pp. 513–532, 2017.
- [6] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [7] L. Kemper, G. Vorhoff, and B. U. Wigger, "Predicting Student Dropout: A Machine Learning Approach," *European Journal of Higher Education*, 2018.
- [8] A. Tayebi, J. Gomez, and C. Delgado, "Analysis on the Lack of Motivation and Dropout in Engineering Students in Spain," *IEEE Access*, vol. 9, pp. 66253–66265, 2021. doi: 10.1109/ACCESS.2021.3076751.
- [9] L. Paura and I. Arhipova, "Student Dropout Rate in Engineering Education Study Program," *Engineering for Rural Development Conference Proceedings*, 2016.
- [10] B. Sekeroglu, K. Dimililer, and K. Tuncal, "Student Performance Prediction and Classification Using Machine Learning Algorithms," in *Proceedings of the ACM International Conference*, 2019, pp. 7–11. doi: 10.1145/3318396.3318419.