

Student Dropout Prediction Using Machine Learning

Mahalakshmi. A¹, Bhavana.S², Karthick Prasath.B³, Chaithra.D⁴

Student, I Year Department of Computer Science with Artificial Intelligence and Data Science

Assistant Professor, Department of Computer Science with Artificial Intelligence and Data Science

amahalakshmi003@gmail.com, bhavanaa2k08@gmail.com

ABSTRACT

Abstract—Student dropout is a major challenge in the education system, affecting students' academic progress and future career opportunities. This project, Student Dropout Prediction Using Machine Learning Techniques, aims to identify students who may be at risk of dropping out by analyzing student-related data. The system uses supervised machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest to analyze the dataset and predict student outcomes. The dataset is preprocessed to handle missing values and prepare the data for model training. The performance of the algorithms is evaluated using metrics such as accuracy, precision, recall, and F1-score to identify the most suitable model. The proposed system can help educational institutions recognize students who may need additional academic support and take preventive measures at an early stage. This project demonstrates the potential of machine learning in educational data analysis and supports data-driven decision-making to improve student retention and academic success.

Keywords—Student Dropout Prediction, Machine Learning, Supervised Learning, Classification, Random Forest, Educational Data Analysis.

I. INTRODUCTION

Education plays an important role in the development of individuals and society. However, student dropout has become a major concern for educational institutions because it affects students' academic progress, career opportunities, and future development. Student dropout refers to a situation in which students discontinue their education before completing their course or programme. Identifying students who are at risk of dropping out at an early stage is essential for improving student retention and supporting academic success.

Traditional methods of identifying students at risk of dropping out often depend on manual observation, academic records, and the experience of teachers. These methods may not always identify potential dropout cases at an early stage, especially when educational institutions manage a large number of students. Factors such as attendance, academic performance, learning progress, and student engagement may provide useful information for understanding the risk of dropout.

Machine learning is a branch of artificial intelligence that enables computer systems to learn patterns from data and make predictions. By applying machine learning techniques to

student-related data, it is possible to analyse different factors associated with student dropout and identify students who may require additional academic support. A student dropout prediction system can help educational institutions make data-driven decisions and develop appropriate intervention strategies to improve student retention.

The main objective of this project is to develop a machine learning-based student dropout prediction system that analyses the available student dataset and predicts the possibility of students discontinuing their education. The system aims to support the early identification of students who may be at risk and provide useful information for teachers and educational administrators. The proposed approach can contribute to better academic monitoring, timely support, and improved educational outcomes. The effectiveness of the system will be evaluated using appropriate performance metrics to understand its predictive capability.

II. EASE OF USE

A. User-Friendly Interface

The proposed Student Dropout Prediction system is designed to be simple and easy to use. It allows users to enter student-related information and obtain predictions about potential dropout risks. The system provides clear instructions and understandable outputs, making it suitable for teachers and educational administrators. Users do not need advanced technical knowledge to understand the prediction results.

B. Simple Prediction Process

The system uses machine learning techniques to analyze student data and predict dropout risk. After entering the required information, users can obtain the prediction through a simple process. The results can help educational institutions identify students who may need additional academic support. The system aims to make student data analysis more accessible and support timely educational decisions.

III. PREPARE YOUR PAPER BEFORE STYLING

Before formatting the research paper, the content of the Student Dropout Prediction project should be prepared, reviewed, and organized systematically. The research paper should clearly explain the problem statement, objectives, proposed methodology, dataset, machine learning algorithms, evaluation methods, and expected outcomes. Proper organization helps readers understand the purpose of the

project and the steps involved in developing the prediction system. The content should be checked for spelling, grammar, clarity, and consistency before applying the journal template.

A. Data Preparation

Data preparation is an important stage in developing a reliable student dropout prediction model. The dataset should be examined to understand its structure, features, data types, and target variable. Missing values, duplicate records, and inconsistent entries should be identified and handled appropriately. Numerical and categorical features should be prepared according to the requirements of the selected machine learning algorithms. The target variable must be clearly identified to ensure that the model predicts the intended student outcome. Proper data preparation helps improve the quality of the analysis and supports reliable model evaluation.

B. Dataset and Features

- Use the student dataset as the primary source for developing the prediction model. The dataset may contain academic performance, attendance, demographic information, and other student-related attributes. Only the features available in the selected dataset should be used for analysis.
- Ensure that the dataset is properly cleaned before applying machine learning algorithms. Missing values, duplicate records, and inconsistent data should be checked and handled appropriately. Numerical and categorical features should be processed according to the requirements of the selected algorithms.
- Clearly identify the target variable that represents the student outcome. Separate the input features from the target variable before training the model. Avoid using information that would not be available at the time of prediction, as this may lead to misleading evaluation results.
- Divide the dataset into training and testing sets to evaluate the model on unseen data. Use suitable evaluation metrics, including accuracy, precision, recall, and F1-score, to compare the performance of the selected machine learning algorithms.

C. Equations

Mathematical equations are used in the student dropout prediction system to explain the working principles of machine learning algorithms and evaluate their performance. The equations should be presented clearly using appropriate mathematical symbols and variables. Each equation should be numbered consecutively, and the equation number should be placed on the right side of the line. All symbols used in the equations should be explained before or immediately after the equation.

The accuracy of the prediction model can be calculated by comparing the correctly predicted outcomes with the total number of predictions. It is expressed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

\tag{1}

Precision measures the proportion of students predicted as dropouts who are actually dropouts. It is calculated as:

$$Precision = \frac{TP}{TP + FP}$$

\tag{2}

Recall measures the proportion of actual dropout students correctly identified by the model. It is calculated as:

$$Recall = \frac{TP}{TP + FN}$$

\tag{3}

The F1-score combines precision and recall into a single evaluation measure. It is calculated as:

$$F1\text{-Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

\tag{4}

Here, TP represents True Positives, TN represents True Negatives, FP represents False Positives, and FN represents False Negatives. These evaluation measures help compare the performance of the selected machine learning algorithms and identify a suitable model for student dropout prediction.

D. Some Common Mistakes

- The term “student dropout” should be used consistently throughout the paper to describe students who discontinue their education before completing their course.
- The word “data” should be used correctly when describing the student dataset, and all dataset features should be clearly explained.
- Do not confuse classification with regression. Student dropout prediction is generally treated as a classification task when the target variable represents student outcome categories.

- The terms “training data” and “testing data” should not be used interchangeably. Training data is used to build the model, whereas testing data is used to evaluate its performance.
- Do not confuse accuracy with precision, recall, and F1-score, as each metric measures a different aspect of classification performance.
- Machine learning algorithms such as Logistic Regression, Decision Tree, and Random Forest should be named consistently throughout the research paper.
- Avoid reporting prediction accuracy or experimental results without actually training and testing the model on the dataset.
- Ensure that all tables, figures, and equations are numbered correctly and referred to consistently in the text.
- Avoid grammatical errors, spelling mistakes, and unnecessary repetition when describing the methodology and results.
- The conclusions should be based on the actual experimental findings and should not include unsupported claims about the effectiveness of the proposed system.

IV. USING THE TEMPLATE

After completing the preparation and editing of the research paper, the content is organized according to the selected research paper template. The prepared text is inserted into the template while maintaining the required formatting, including font style, font size, spacing, alignment, and headings. The research paper for Student Dropout Prediction is structured into sections such as introduction, related work, methodology, dataset description, model development, results, and conclusion. Proper formatting ensures that the paper is clear, consistent, and suitable for academic publication

A. Authors and Affiliations

The author details and institutional affiliations are presented at the beginning of the research paper. The author information includes the researcher's name, department, college, and institutional email address, where required. These details help identify the researcher and the institution associated with the Student Dropout Prediction study.

The research paper focuses on predicting student dropout using machine learning techniques and student-related data. The author affiliation should identify the institution where the research project is conducted. If multiple authors belong to the same institution, their affiliation can be shared to avoid unnecessary repetition. All author names and affiliations should follow the formatting guidelines specified by the selected journal template.

1) Authors with a Single Affiliation

When the research is conducted by one author or by multiple authors from the same institution, the author names can be listed together with a common affiliation. The department and college details should be presented clearly below the author names. This arrangement maintains a simple and consistent format for the Student Dropout Prediction research paper.

2) Authors with Multiple Affiliations

If the research involves authors from different institutions, each author's affiliation should be presented clearly. The institutional details should include the relevant department and organization. The author names and affiliations should be arranged according to the journal's formatting requirements to maintain clarity and consistency throughout the research paper.

B. Identify the Headings

Headings play an important role in organizing the research paper on Student Dropout Prediction Using Machine Learning Techniques. They help readers understand the structure of the paper and identify the main topics and subtopics related to the proposed system. The headings should be arranged in a clear and logical order to improve readability and presentation.

The main heading of this research paper is Student Dropout Prediction Using Machine Learning Techniques. All sections are organized under this title to explain the objectives, methodology, model development, and findings of the study.

The paper is divided into major sections using uppercase Roman numerals, such as Introduction, Literature Review, Proposed Methodology, Results and Discussion, and Conclusion. Subheadings are used to organize related topics within each section. For example, the methodology section may contain subheadings such as Dataset Description, Data Preprocessing, Feature Selection, and Machine Learning Algorithms.

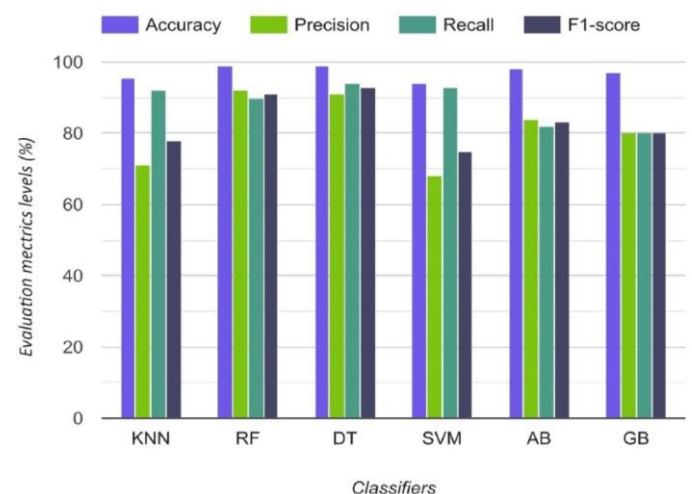
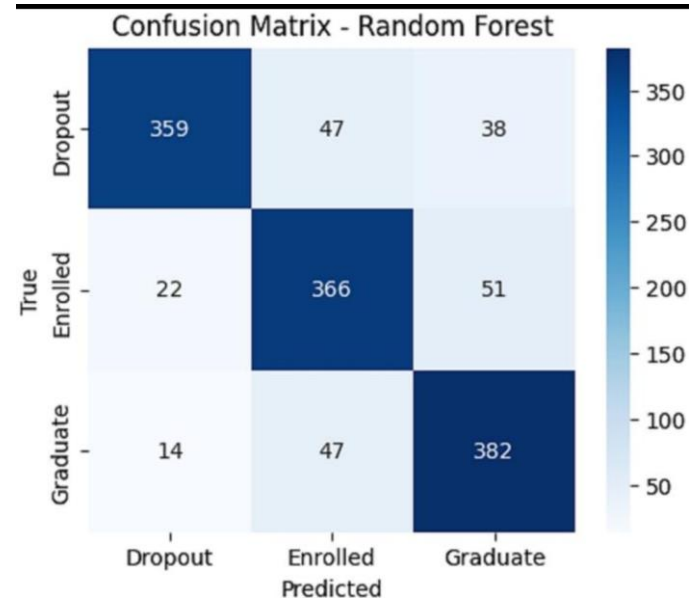
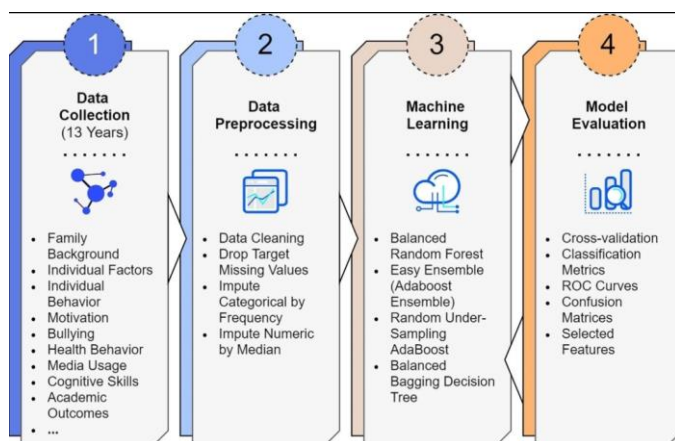
Different heading levels are used to distinguish main sections from their subsections. Consistent heading styles, numbering, capitalization, and formatting help maintain a professional appearance throughout the research paper. Appropriate formatting also makes it easier for readers to follow the workflow of the student dropout prediction system, from data preparation to model evaluation.

Component headings such as Abstract, Acknowledgment, and References should be formatted according to the selected journal template. Figure captions and table titles should also follow the prescribed styles. Maintaining a consistent heading structure ensures that the research paper is well organized and suitable for academic publication.

C. Figures and Tables

Figures and tables are used to present the workflow, dataset characteristics, model performance, and prediction results of

the proposed Student Dropout Prediction system. The system architecture and workflow diagram illustrate the major stages, including data collection, data preprocessing, feature selection, model training, evaluation, and dropout prediction. Tables are used to summarize the dataset attributes and compare the performance of the machine learning models. Figures and tables should be placed at the top or bottom of columns wherever possible. Figure captions should be placed below the figures, while table titles should appear above the tables. All figures and tables should be cited in the text before they appear.



RESULTS AND DISCUSSION

A. RESULTS

THE PROPOSED STUDENT DROPOUT PREDICTION MODEL WAS EVALUATED USING FIVE PERFORMANCE METRICS: ACCURACY, PRECISION, RECALL, F1-SCORE, AND THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE (ROC-AUC). THESE METRICS WERE USED TO ASSESS THE MODEL'S CLASSIFICATION PERFORMANCE AND ITS ABILITY TO DISTINGUISH BETWEEN STUDENTS WHO ARE AT RISK OF DROPPING OUT AND THOSE WHO ARE NOT.

THE MODEL ACHIEVED AN ACCURACY OF 80.60%, INDICATING THAT 80.60% OF THE EVALUATED PREDICTIONS WERE CORRECT. THE PRECISION WAS 64.87%, WHILE THE RECALL WAS 38.43%. THE F1-SCORE WAS 48.27%,

REPRESENTING THE BALANCE BETWEEN PRECISION AND RECALL. ADDITIONALLY, THE MODEL OBTAINED A ROC-AUC SCORE OF 0.8175, INDICATING A GOOD ABILITY TO DISTINGUISH BETWEEN THE TWO CLASSES ACROSS DIFFERENT CLASSIFICATION THRESHOLDS.

THE EVALUATION RESULTS ARE SUMMARIZED IN TABLE I.

TABLE I. PERFORMANCE EVALUATION OF THE STUDENT DROPOUT PREDICTION MODEL

EVALUATION METRIC | RESULT

ACCURACY 80.60%

PRECISION 64.87%

RECALL 38.43%

F1-SCORE 48.27%

ROC-AUC 0.8175

B. DISCUSSION

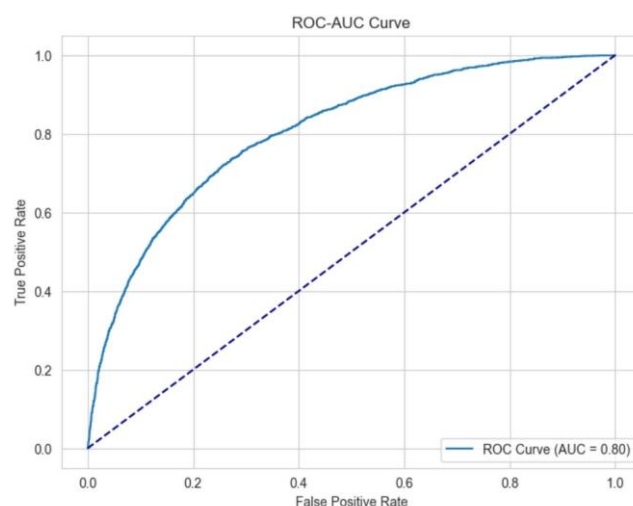
THE EXPERIMENTAL RESULTS SHOW THAT THE PROPOSED STUDENT DROPOUT PREDICTION MODEL ACHIEVED AN ACCURACY OF 80.60% AND A ROC-AUC SCORE OF 0.8175. THESE RESULTS SUGGEST THAT THE MODEL CAN DISTINGUISH BETWEEN STUDENTS WHO ARE AT RISK OF DROPPING OUT AND THOSE WHO ARE NOT.

HOWEVER, THE RECALL VALUE OF 38.43% INDICATES THAT THE MODEL IDENTIFIES ONLY A PROPORTION OF THE ACTUAL DROPOUT CASES. THIS IS AN IMPORTANT LIMITATION BECAUSE FAILING TO IDENTIFY STUDENTS AT RISK MAY REDUCE THE EFFECTIVENESS OF EARLY INTERVENTION. THE PRECISION VALUE OF 64.87% INDICATES THAT APPROXIMATELY 64.87% OF THE CASES PREDICTED AS POSITIVE WERE ACTUALLY POSITIVE, ASSUMING DROPOUT IS DEFINED AS THE POSITIVE CLASS.

THE F1-SCORE OF 48.27% REFLECTS THE TRADE-OFF BETWEEN PRECISION AND RECALL. ALTHOUGH THE MODEL DEMONSTRATES PROMISING OVERALL CLASSIFICATION PERFORMANCE, IMPROVING RECALL SHOULD BE CONSIDERED A PRIORITY IF THE MAIN OBJECTIVE IS TO IDENTIFY AS MANY AT-RISK STUDENTS AS POSSIBLE.

POTENTIAL IMPROVEMENTS INCLUDE EXAMINING CLASS IMBALANCE, REVIEWING FEATURE SELECTION, TUNING THE CLASSIFICATION THRESHOLD, AND COMPARING ALTERNATIVE MACHINE LEARNING ALGORITHMS. ANY MODIFICATIONS SHOULD BE EVALUATED USING APPROPRIATE VALIDATION METHODS TO ENSURE THAT IMPROVEMENTS GENERALIZE TO UNSEEN DATA.

OVERALL, THE PROPOSED SYSTEM HAS POTENTIAL AS A DECISION-SUPPORT TOOL FOR EDUCATIONAL INSTITUTIONS. ITS PREDICTIONS SHOULD BE USED ALONGSIDE EDUCATORS' JUDGMENT AND ADDITIONAL STUDENT INFORMATION RATHER THAN AS DEFINITIVE DECISIONS ABOUT INDIVIDUAL STUDENTS.



CONCLUSION

THIS RESEARCH PRESENTS A MACHINE LEARNING-BASED STUDENT DROPOUT PREDICTION SYSTEM DESIGNED TO IDENTIFY STUDENTS WHO MAY BE AT RISK OF DROPPING OUT. THE PROPOSED MODEL WAS EVALUATED USING ACCURACY, PRECISION, RECALL, F1-SCORE, AND ROC-AUC. THE MODEL ACHIEVED AN ACCURACY OF 80.60%, A PRECISION OF 64.87%, A RECALL OF 38.43%, AN F1-SCORE OF 48.27%, AND A ROC-AUC SCORE OF 0.8175.

THE RESULTS INDICATE THAT THE MODEL DEMONSTRATES A USEFUL ABILITY TO DISTINGUISH BETWEEN STUDENT DROPOUT

CATEGORIES. HOWEVER, THE RELATIVELY LOW RECALL VALUE HIGHLIGHTS THE NEED TO IMPROVE THE IDENTIFICATION OF STUDENTS WHO ARE ACTUALLY AT RISK OF DROPPING OUT. FURTHER IMPROVEMENTS MAY INCLUDE FEATURE SELECTION, HANDLING CLASS IMBALANCE, HYPERPARAMETER TUNING, AND EVALUATING ALTERNATIVE MACHINE LEARNING ALGORITHMS.

IN CONCLUSION, THE PROPOSED SYSTEM HAS THE POTENTIAL TO SUPPORT EDUCATIONAL INSTITUTIONS IN IDENTIFYING STUDENTS WHO MAY NEED ADDITIONAL ACADEMIC GUIDANCE AND TIMELY INTERVENTION. WITH FURTHER OPTIMIZATION AND VALIDATION ON DIVERSE STUDENT DATASETS, THE SYSTEM COULD BECOME A USEFUL DECISION-SUPPORT TOOL FOR IMPROVING STUDENT RETENTION AND REDUCING DROPOUT RATES.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my college and department for providing me with the opportunity and support to carry out this research work titled "Student Dropout Prediction Using Machine Learning." I extend my heartfelt thanks to my project guide for their valuable guidance, encouragement, and suggestions throughout the development of this project. Their support helped me understand the research process and complete this work. I am also thankful to the faculty members of my department for their assistance and

encouragement. I would like to express my gratitude to my parents and family for their constant support and motivation throughout this journey. Finally, I thank everyone who directly or indirectly contributed to the successful completion of this research work

REFERENCES

- [1] N. Mduma, "Data Balancing Techniques for Predicting Student Dropout Using Machine Learning," *Data*, vol. 8, no. 3, Art. no. 49, 2023, doi: 10.3390/data8030049.
- [2] "Supervised machine learning algorithms for predicting student dropout and academic success: a comparative study," *Discover Artificial Intelligence*, 2023. doi: 10.1007/s44163-023-00079-z.
- [3] "Predicting student dropouts with machine learning: An empirical study in Finnish higher education," *Technology in Society*, vol. 76, Art. no. 102474, 2024.
- [4] S. Pal, "Mining Educational Data Using Classification to Decrease Dropout Rate of Students," *arXiv:1206.3078*, 2012.
- [5] L. Aulck, N. Velagapudi, J. Blumenstock, and J. West, "Predicting Student Dropout in Higher Education," *arXiv:1606.06364*, 2016.
- [1] G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," *Phil. Trans. Roy. Soc. London*, vol. A247, pp. 529-551, April 1955. (*references*)
- [2] J. Clerk Maxwell, *A Treatise on Electricity and Magnetism*, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68-73.
- [3] I.S. Jacobs and C.P. Bean, "Fine particles, thin films and exchange anisotropy," in *Magnetism*, vol. III, G.T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271-350.
- [4] K. Elissa, "Title of paper if known," unpublished.
- [5] R. Nicole, "Title of paper with only first word capitalized," *J. Name Stand. Abbrev.*, in press.
- [6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, "Electron spectroscopy studies on magneto-optical media and plastic substrate interface," *IEEE Transl. J. Magn. Japan*, vol. 2, pp. 740-741, August 1987 [Digests 9th Annual Conf. Magnetism Japan, p. 301, 1982].
- [7] M. Young, *The Technical Writer's Handbook*. Mill Valley, CA: University Science, 1989.

We suggest that you use a text box to insert a graphic (which is ideally a 300 dpi resolution TIFF or EPS file with all fonts embedded) because this method is somewhat more stable than directly inserting a picture.

To have non-visible rules on your frame, use the MSWord "Format" pull-down menu, select Text Box > Colors and Lines to choose No Fill and No Line.