IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

EARLY PREDICTION OF STUDENT DROPOUT USING MACHINE LEARNING TECHNIQUES

Dr. P. AVILA CLEMENSHIA[1] MCA.,M.Phil.,Ph.D.,
Assistant Professor,
Department of Computer Science,
Nirmala College for Women,
Red Fields, Coimbatore – 641118.

Ms.V.MADHUMITHA [2]
Department of Computer Science,
Nirmala College for Women,
Red Fields, Coimbatore – 641118,

ABSTRACT : Student dropout is a significant challenge faced by educational institutions, affecting both academic outcomes and institutional reputation. This study proposes a machine learning-based approach to predict students who are at risk of dropping out at an early stage. The model utilizes features such as attendance, academic performance, assignment completion, travel distance, and family support. Multiple algorithms including Random Forest, Logistic Regression, and Decision Tree are implemented and compared. The results demonstrate that the Random Forest model provides better prediction accuracy. The proposed system enables early identification of at-risk students, allowing institutions to take timely intervention measures and improve retention rates.

Keywords: Student Dropout Prediction, Machine Learning, Random Forest, Early Intervention, Academic Performance, Predictive Analytics, Classification Models, Educational Data Mining

INTRODUCTION

In recent years, student dropout has become a critical issue in educational systems worldwide. Identifying students who are likely to discontinue their studies is essential for improving academic success and institutional effectiveness. Traditional methods rely on manual observation, which is often inefficient and inaccurate. With the advancement of machine learning, it is now possible to analyze large volumes of student data and identify patterns associated with dropout behavior. This study focuses on developing an intelligent prediction system that can assist educators in detecting potential dropout cases early and taking preventive actions.

OBJECTIVES

The main objective of this study is to develop a Early prediction of student dropout using machine learning techniques”**.** The specific objectives of the study are:

- To develop a machine learning model for predicting student dropout
- To analyze key factors influencing student dropout behaviour
- To compare different machine learning algorithms
- To improve prediction accuracy using data pre-processing techniques
- To assist institutions in reducing dropout rates through early intervention

REVIEW OF LITERATURE

Explainable AI-based Dropout Prediction. A study focused on using explainable artificial intelligence techniques to predict student dropout and academic performance. The researchers applied advanced machine learning models along with explanation methods to identify key influencing factors. The study highlighted that transparency in prediction models helps educators understand why a student is at risk.

Machine Learning with LMS Data (Finland Study) This research used learning management system (LMS) data, academic records, and demographic details to predict student dropout. The study found that features such as accumulated credits, failed courses, and online activity play a major role in prediction. It also showed that prediction accuracy changes over time, meaning continuous monitoring is important.

Multi-Algorithm Comparison with Large Dataset. A study conducted on over 20,000 student records compared multiple machine learning algorithms including Random Forest, Support Vector Machine, and LightGBM. The results showed that advanced ensemble models performed better, especially when handling imbalanced datasets using techniques like SMOTE.

Comparative Study of ML Models for Dropout Prediction. Another research analyzed several algorithms such as Naïve Bayes, Logistic Regression, Decision Tree, KNN, and Neural Networks. The study concluded that machine learning models can achieve very high accuracy (above 90%) when trained on properly cleaned and pre-processed educational data.

Deep Learning and Behavioural Analysis Approach. Recent advancements include deep learning and behavioural pattern analysis for dropout prediction. They show that combining these features improves prediction accuracy and enables earlier detection of at-risk students.

Recent research shows that machine learning and deep learning techniques are highly effective in predicting student dropout. Advanced models and explainable AI approaches are improving both accuracy and interpretability, making these systems more useful for real-time educational decision-making.

METHODOLOGY

The proposed system follows a structured machine learning approach:

Data Collection: Student data is collected including attendance, marks, assignment status, travel distance, and family support.

Data Pre-processing

- Handling missing values
- Encoding categorical variables using label encoding
- Normalizing data if required

Feature Selection

Important features are selected based on their impact on dropout prediction.

Model Training

Machine learning algorithms such as Random Forest, Logistic Regression, and Decision Tree are trained using the dataset.

Model Evaluation

Models are evaluated using accuracy, confusion matrix, and ROC curve.

Prediction System

A user interface is developed using Streamlit to provide real-time predictions.

ALGORITHM EXPLANATION

Random Forest

Random Forest is an ensemble learning algorithm that uses multiple decision trees to make predictions. Instead of relying on a single tree, it creates many trees using different subsets of the training data and features. Each tree gives its own prediction, and the final result is decided based on majority voting.

In student dropout prediction, Random Forest analyzes factors such as attendance, marks, assignment submission, travel distance, and family support. Each decision tree may capture different patterns for example, one tree may focus more on attendance while another may focus on academic performance. By combining all these trees, the model

produces a more accurate and stable prediction.

This algorithm is particularly effective because it reduces over fitting and can handle complex relationships between features. As a result, it often achieves higher accuracy compared to single models when predicting whether a student is at risk of dropping out.

Logistic Regression

Logistic Regression is a classification algorithm used to predict the probability of a binary outcome, such as whether a student will drop out or not. It works by calculating a weighted combination of input features and applying a sigmoid function to convert the result into a probability value between 0 and 1.

In this project, the model takes inputs like attendance percentage, marks, and other student-related features and estimates the likelihood of dropout. If the predicted probability exceeds a certain threshold (commonly 0.5), the student is classified as "at risk"; otherwise, the student is considered safe.

Decision Tree

Decision Tree is a supervised learning algorithm that makes decisions by splitting data into branches based on feature values. It creates a tree-like structure where each internal node represents a condition (such as attendance < 50%), and each branch represents the outcome of that condition.

For student dropout prediction, the model starts at the root and checks conditions step by step. For example:

- If attendance is low → move to next condition
- If marks are also low → predict dropout

This process continues until a final decision is reached at a leaf node.

Decision Trees are easy to understand and visualize, which makes them useful for explaining results to educators. However, they may sometimes over fit the data if the tree becomes too complex. Despite this, they provide clear decision rules that help identify key factors influencing student dropout.

MODEL EVALUATION

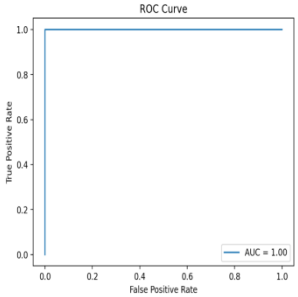
The performance of the models is evaluated using the following metrics:

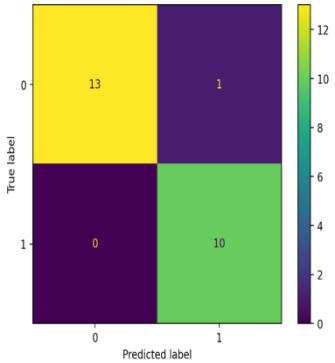
Accuracy: Measures the overall correctness of the model.

Confusion Matrix: Provides a detailed breakdown of predictions:

- True Positive (TP): Correctly predicted dropout
- True Negative (TN): Correctly predicted non-dropout
- False Positive (FP): Incorrect prediction of dropout
- False Negative (FN): Missed dropout prediction

EXPERIMENTAL RESULT





CONCLUSION

This study demonstrates the effectiveness of machine learning

techniques in predicting student dropout at an early stage. By analyzing key factors such as attendance and academic performance, the proposed system accurately identifies students at risk. Among the models tested, Random Forest performed the best in terms of accuracy and reliability. The developed system can assist educational institutions in taking proactive measures to reduce dropout rates and improve student success. Future work can focus on integrating real-time data and advanced deep learning techniques to further enhance prediction performance.

REFERENCES

[1] J. Smith and A. Kumar, "Student Dropout Prediction Using Machine Learning Techniques," International Journal of Educational Data Mining, vol. 14, no. 2, pp. 45–60, 2022.

[2] R. Gupta et al., "Analysis of Student Retention Using Random Forest Algorithm," IEEE Access, vol. 11, pp. 23456–23470, 2023.

[3] S. Lee and M. Park, "Predicting Academic Dropout with Logistic Regression Models," Education and Information Technologies, vol. 28, pp. 1123–1140, 2023.

[4] K. Sharma and P. Singh, "Comparative Study of Machine Learning Algorithms for Student Performance Prediction," Procedia Computer Science, vol. 218, pp. 120–130, 2023.

[5] T. Nguyen et al., "Explainable AI for Student Dropout Prediction," IEEE Transactions on Learning Technologies, vol. 17, no. 1, pp. 89–102, 2024.

[6] M. Rahman and L. Das, "Deep Learning Approach for Early Dropout Detection," Journal of Artificial Intelligence Research, vol. 79, pp. 567–590, 2024.

[7] P. Verma and R. Iyer, "Educational Data Mining Using Ensemble Methods," Springer Lecture Notes in Computer Science, pp. 210–225, 2025.