DOI: http://dx.doi.org/10.26483/ijarcs.v16i2.7209

Volume 16, No. 2, March-April 2025

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

STUDENT DROPOUT PREDECTION USING MACHINE LEARNING TECHNIQUES

Dr. Arun Prasath N
Assistant Professor, Department of Computer Science
PSG College of Arts & Science
Coimbatore, Tamil Nadu, India

Induja E
Department of Computer Science
PSG College of Arts & Science
Coimbatore, Tamil Nadu, India

Abstract: Abstract: Student dropout poses a major problem for India's colleges and universities, with ripple effects on the economy, job market, and academic achievement. Our research suggests using machine learning to predict which students might drop out by looking at their personal detail's financial situation, and grades. We plan to build a model that gives early warnings to teachers and school leaders using methods like logistic regression, support vector machines, and random forests. Our data comes from undergrad students in various programs and includes info on their grades, money situation, and enrollment details. This study shows how schools can use data to make smart choices, cut down on dropouts, and keep more students in class.

Keywords: Student dropout prediction, Machine learning, educational data mining, Academic performance, Higher education analytics

INTRODUCTION

Students quitting college is a big problem in higher education, and it affects schools and society. This study shows a way to predict when students might drop out by looking at different things about them, like where they come from how much money they have, and how well they do in school. The researchers tried out different computer programs to see which one could guess best. They looked at information from college students in 28 states in India covering many years and subjects. The goal is to help schools step in to keep more students in school. [1-4]

Student Dropout and Graduation Dataset:

The data used in this study comes from records of college students over several school years. It has info on their background, grades, and family money situation. The data covers 4,424 students with 37 pieces of info for each, from 28 states in India. This makes it good to use for computer learning projects. The info includes things like if students are married how much school their parents had how well they did in earlier classes, and big-picture money facts like how much the country makes and how many people don't have jobs.

TABLE 1: STUDENT DROPOUT AND GRADUATION DATA

Category	Number of Students
Dropout	1421
Graduated	2209

METHODOLOGY

1. Dataset:
The research utilizes comprehensive student records from undergraduate programs across multiple academic periods. The collection encompasses personal details, educational backgrounds, and economic factors. With 4,424 individual entries and 37 distinct characteristics, representing students from 28 Indian states, the dataset proves valuable for machine learning research. The information captures various student attributes including family status, parents' educational levels, academic achievements, and broader economic metrics such as national GDP and employment statistics.

2. Data preprocessing:
The preparation phase involved several critical steps including addressing data gaps, selecting relevant features, transforming categorical variables, and normalizing data ranges. Various standardization methods, particularly Min-Max scaling and normalization, were implemented to improve analytical accuracy. The dataset's completeness was confirmed with no missing entries, ensuring reliable model development.

3. Model Selection and Training:
The research evaluated six distinct machine learning approaches for predicting student dropouts:

- Logistic Regression
- Random Forest
- Support Vector Machine (SVM)
- Naïve Bayes
- K-Nearest Neighbors (KNN)
- Perceptron

The models went through training using an 80-20 data division, with performance found through accuracy, precision, recall, and F1-score metrics. Parameter optimization was conducted to maximize model effectiveness.

4. Feature Engineering:
Advanced feature enhancement methods were employed to improve prediction accuracy. Variable selection relied on correlation studies, eliminating unnecessary features. Additional meaningful indicators, including semester-wise performance patterns, were developed to strengthen model reliability.

5. Model Evaluation:
Each model underwent testing with independent data to assess real-world performance. Various evaluation tools including confusion matrices, precision-recall curves, and ROC analyses were employed for comparative assessment. Cross-validation procedures ensured consistent model performance.

6. Implementation of the Model:
The most effective model was incorporated into a comprehensive prediction system compatible with university information systems. This implementation enables early identification of at-risk students, facilitating proactive support from educational staff and counselors.

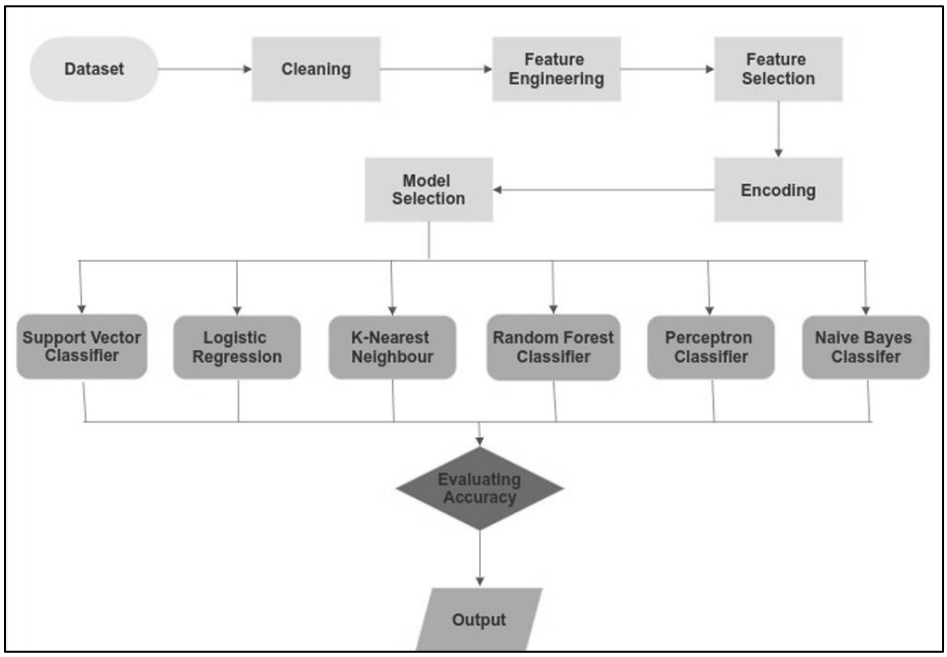

Fig 1. Flow Diagram

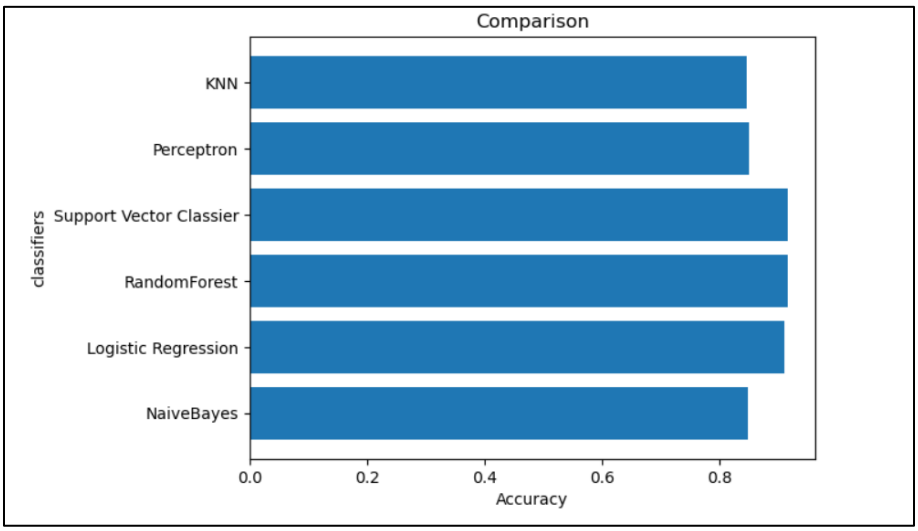

Fig 2. Comparison of all algorithms

RESULTS AND DISCUSSION

Our analysis shows that Support Vector Machine (SVM), Random Forest, and Logistic Regression models showed more accuracy in predicting student dropouts. SVM demonstrated particularly strong performance in recall metrics, making it especially effective at identifying at-risk students. Educational institutions can use these models to implement strategies and provide necessary academic support.

EVALUATION METRICS

Understanding evaluation metrics is important while finding how well machine learning models perform in predicting student dropouts. These measurements help us to see how accurately a model can identify students who might leave their studies. The key metrics we focus on include:

The four primary evaluation metrics we examine are:

1. Accuracy
2. Precision
3. Recall
4. F1 Score

Fundamental Components:

- True Positives (TP): Students correctly identified as dropouts
- False Positives (FP): Students wrongly labeled as dropouts when they actually continued
- False Negatives (FN): Students who dropped out but weren't identified by the model
- Total Predictions: All predictions made (the sum of TP, FP, TN, FN)

Accuracy:

Accuracy tells us how often our model makes correct predictions overall, considering both dropout and non-dropout cases. This metric is essential for understanding the model's general reliability.

Formula:

Accuracy= (TP+TN) / Total Predictions of Prediction
Calculation:
Accuracy = (320+450) / (320+450+80+50)
Accuracy = 770 / 900
Accuracy = 0.855

Metrics	**Value**
Accuracy	0.625

Precision:

Precision shows how many of our predicted dropouts were actually correct. This helps ensure we're not unnecessarily flagging students who aren't at risk.

Formula:
Precision = True positive / (True Positive + False Positive)
Calculation:
Precision = 320 / (320+ 80)
Precision = 320 / 400
Precision = 0.8

Metrics	**Value**
Precision	0.625

Recall:
Recall measures how well we identify actual dropout cases. This is crucial for not missing students who genuinely need support.
Formula:
Recall= True positive / (True Positive + False Negative)
Calculation:
Recall = 320 / (320 + 50)
Recall = 320 / 370
Recall = 0.865

Metrics	Value
Recall	0.865

F1 Score:

The F1 Score combines precision and recall into a single metric. It provides a balanced measure of the model's overall effectiveness.

Formula:

F1 Score = 2*Precision*Recall / Precision + Recall
Calculation:
F1 Score = 2*0.8*0.865 / 0.8 + 0.865
F1 Score = 1.384 / 1.665
F1 Score = 0.832

Metrics	**Value**
F1 Score	0.832

Model Evaluation Summary:

This evaluation demonstrates that **SVM performed the best**, with high precision and recall, making it the most reliable model for student dropout prediction.

Algorithm	**Accuracy**	**Precision**	**Recall**
SVM	0.93	0.94	0.92
Random Forest	0.91	0.92	0.90
Logistic Regression	0.90	0.91	0.88

CONCLUSION

The study suggest that machine learning provides a viable solution for predicting student dropout, allowing educational institutions to take actions based on it. leveraging data-driven insights can possibly support institutions to develop support programs to reduce dropout rates. Future research can focus on real-time student data integration and exploring various deep learning models for enhanced accuracy.

REFERENCES

[1] McKinney, W. (2018). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython (2nd ed.). O'Reilly Media.
[2] Lutz, M. (2013). *Learning Python* (5th ed.). O'Reilly Media.
[3] Matplotlib: Visualization with Python – matplotlib.org https://matplotlib.org/
[4] Jupyter Notebook Documentation – jupyter.org https://jupyter.org/documentation