Indian Journal of Computer Science and Technology
https://www.doi.org/10.59256/indjcst.20240302031
Volume 3, Issue2 (May-August 2024), PP: 213-217.
www.indjcst.com
ISSN No: 2583-5300

Predicting Student Success: A Comparative Examination of Machine Learning Techniques

M. Priyadharshini[1], S. Indra[2], S. Achuthan[3], K. Lokesh[4]
[1,2,3,4]PG Student, Department of Computer Science and Engineering, Bharathidasan Engineering College, Vellore, Tamil Nadu, India.

To Cite this Article: *M. Priyadharshini[1], S. Indra[2], S. Achuthan[3], K. Lokesh[4], "Predicting Student Success: A Comparative Examination of Machine Learning Techniques", Indian Journal of Computer Science and Technology, Volume 03, Issue 02 (May-August 2024), PP: 213-217.*

Abstract: *An essential component of individual and society growth is student education. It could involve innovative curriculum design, efficient teaching techniques, educational technology, and resources. The majority of educational data mining (EDM) research, however, has concentrated on identifying at-risk children so that early, focused interventions can be given, as well as forecasting students' future performance. EDM seeks to create techniques for examining the distinct and progressively larger amounts of data produced by educational environments in order to gain a deeper comprehension of both learners and the environments in which they are taught. It is applicable to study the consequences of educational support and forecast how students will learn in the future. Using student performance datasets, we assessed the suggested feature selection technique's efficacy employing using five machine learning classifiers: Decision Tree (DT), Naïve Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and Logistic Regression (LR). This suggested approach uses the chosen algorithm to verify training duration, analysis, recall, accuracy, and precision. But in order to determine which algorithm possessed the greatest best accuracy, they compared all of them the outcomes of this review will help academic researchers, practitioners, and professionals deal with imbalanced classification, particularly within the field of higher education.*

Keywords: *Educational Data Mining (EDM), Students' Future Performance, Educational Technology, Machine Learning Classifiers, Higher Education.*

I.INTRODUCTION

Many scholars have applied data mining and machine learning approaches to the subject of education due to the nation's growing emphasis on education and the extensive use of large-scale datasets in the sector. The goal of the data mining study field known as "educational data mining" (EDM) is to find patterns, trends, and linkages in the vast amounts of data produced by educational activities by identifying correlations between various variables. This emerging discipline analyses educational Big Data using statistical, machine learning, and data mining techniques, focusing on student performance prediction. It is critical to realize that assessing student performance entails more than just looking at marks; it calls for a thorough evaluation that considers aspects like the complexity of the course and the unique grading standards of each student [2]. Educational establishments gain a great deal from EDM's predictive insights, which allow them to maximize resources and give students individualized help.

There is a strong association between the traits of information thinking and learning effect, according to an examination of the behaviour features of college students' information literacy using the Pearson algorithm. The learning impact of information literacy in college students is categorized and predicted using supervised classification algorithms including Decision Tree, K Nearest Neighbour, Naive Bayes, Neural Net, and Random Forest. It is found that when it comes to learning impact classification prediction, the prediction model known as Random Forest the best [3].

With the huge increase in popularity of online learning in recent years, students everywhere can now choose from a variety of flexible learning options. One issue that online the challenge that educational institutions encounter is the high percentage of student dropouts. Online course providers, educational institutions, and individual students are all quite worried about dropout. But because MOOCs don't have as much of a binding force as traditional classrooms do, a lot of students have dropped out of their courses for personal or professional reasons, wasting educational resources [4].

Data are now seen as a crucial tool for making defensible decisions. Applying cutting-edge techniques is therefore advantageous for obtaining specific, practical knowledge. Machine learning is a component of artificial intelligence (ML), which compiles all techniques that enable a machine to learn and make precise predictions based on historical data. Among the essential elements components in assessing and tracking student achievement in universities (HEI) is student grade prediction. Over the years, this field has drawn a lot of attention from the education sector since a number of investigations have shown the accuracy of student grade prediction with the aid of current machine learning algorithms, hence improving student achievement.

To guarantee that forecast student anxiety, this research explores the use of active and machine learning methodologies. This study investigates the ways in which these Technologies have the potential to comprehend and forecast the anxiety levels of students. To ensure that improve machine learning models' ability to anticipate students' anxiety levels, this study makes use of

active learning techniques. This adaption also highlights the value of active learning approaches in improving machine learning models' accuracy in predicting student anxiety. This research builds predictive models for student anxiety Using methods for machine learning and two datasets including behavioural data. [13]

Support Vector Machine and Random Forest are the most often used algorithms among those in use. Training Higher education institution students are a crucial part is an essential component of creating new leadership. (HEIs). It gives them the tools they must question conventional wisdom, promote fresh perspectives in accordance with modern problems and characteristics, accomplish ongoing, self-directed learning, and adapt to a range of work-related situations. The significance of student graduation is predicated on the previously listed elements. Extra assignments and projects assigned Students that don't do well might work on improving their poorly can help improve their performance. The inability to identify at-risk pupils in a timely manner is a significant issue, though. Several scholars are utilizing machine learning approaches to look at this matter. Furthermore to its many applications, machine learning is also being utilized by educators to assist in identifying students who are at-risk and to offer early intervention.

II.RELATED WORK

An important factor in determining a student's learning level is their performance. It assists students not just with goal-setting and academic planning, but also with early intervention for kids who might experience academic issues, additionally with providing individualized advice and support. In the majority of universities, a student's ability to graduate successfully is also largely determined by how well they do. For teachers to raise the standard of their instruction and the efficiency of their instructional management, accurate student performance prediction is crucial. Regression tasks and classification tasks based on grade-level divisions are the two categories of grade prediction task types. Regression tasks are designed to predict specific numerical grades. Online, behavioural, and academic data are the categories into which the many data types used in grade prediction can be separated. Data created by students on online learning systems is known as online data.

DAOZONG SUN et al [1], with this approach, grades, CGPA, SGPA, and other prior academic performance are accustomed to predict grades. Four machine learning techniques—the multi-layer perceptron (MLP), logistic regression (LR), simple vector machine (SVM), and CART—are compared to this model. Academic data is statistics about how well students performed on assignments, tests, quizzes, and their GPA during their educational journey. Using Multilayer Perceptron (MLP), Logistic Regression (LR), Support Vector Machine (SVM), and Classification and Regression Trees (CART) as baseline models, approaches for machine learning classification were contrasted with the MFAPM model in the comparison experiment. The dataset for the baseline model is student-grade data, and the linear kernel is employed as the SVM's kernel function. The ReLU function is utilized as the activation function for the hidden layers of the multi-layer perceptron, including three layers, each with an equivalent number of neurons in each layer. The PyTorch 1.8.1+cu111 framework and Python 3.7.12 were employed as the foundation for the MFAPM model, while adaptive moment estimate optimization was used as the optimizer. The suggested MFAPM model produced improved predictive outcomes for all four-evaluation measures for student performance prediction tasks when compared to the MLP, LR, SVM, and CART four prediction approaches. Recall, accuracy, and precision were 72.5%, 82.2% and 73.5% respectively, while the F1 score was 73.6%. The MFAPM model outperforms the other four comparative models on these four evaluation parameters, showing improvements over the best baseline model of 9%, 23.5%, 16.6% and 21.1% respectively.

KOUSHIK ROY et al [2], this research employed machine learning to predict student performance and examined the results of feature selection strategies. Pre-processing, feature selection, evaluation, exploratory data analysis, and data cleaning were all part of the methodology. A number of machine learning methods were applied,, and cross-validation was employed to evaluate assess them. The five classifiers for machine learning utilized in this system are LR, KNN, SVM, NB, and DT. The Adaptive Feature Selection Algorithm(AFSA) is the main topic of this proposed comparison with other algorithms. The methods employed to evaluate training time, analysis, and accuracy using the chosen algorithm. Following the evaluation of all five methods, an AFSA model was shown to have a high frequency and accuracy of student performance prediction.

YONG SHI et al [3], to look into the predicted learning effect, this research builds An anticipatory model of learning impact based on information literacy learning behaviour characteristics. It then examines the features of college students' learning behaviours. Data on information literacy learned by 320 college students from a Chinese institution was used in the project. There is a strong association between the traits of information thinking and learning effect, according to an analysis of the learning behaviour features of college students' information literacy students using the Pearson algorithm. The learning effect of information literacy in college students is categorized and predicted using supervised classification algorithms such Decision Tree, KNN, Naive Bayes, Neural Net, and Random Forest. It is found that when it comes to learning impact classification prediction, the Random Forest prediction model performs the best. The findings indicate that the values pertaining to precision, recall, accuracy, F1-Score, and Kapaa coefficient are 92.50%, 84.56%, 94.81%, and 89.39%, respectively. This study provides recommendations for varied interventions and a management decision-making framework.

Table 1 Acronym used in professional contexts.

Acronyms	Description
MLP	Multi-layer Perceptron
LR	Logistic Regression
SVM	Simple Vector Machine
CART	Classification and Regression Trees
KNN	K-Nearest Neighbour

NB	Naïve Bayes
DT	Decision Tree
RF	Random Forest
CBL	Case Based Learning
DNN	Deep Neural Network
ANN	Artificial Neural Network
CNN	Convolutional Neural Network
XAI	Explainable AI
SGD	Stochastic Gradient Descent
XgBoost	eXtreme Gradient Boosting
T-SNE	t-distributed Stochastic Neighbour Embedding
FDT	Fastest Decision Tree

MUSTAPHA SKITTOU et al [5], analysing student performance or forecasting at-risk students has been greatly streamlined using with the aid of help of the Early Warning System (EWS). The goal of our study is to create an early warning system that considers a variety of institutional, sociocultural, and pedagogical elements that directly influence a student's decision to leave school. In order to achieve exhaustiveness and precision in the selection of dropout indicators, we have worked on an original database devoted to this problem. We demonstrate how this might be helpful for lesson planning by use a Django application we made especially for this use to visualize the findings. Our primary tools for analysis were machine learning techniques with a focus on categorization. Thus, the most well-known and effective for this particular analysis aspect are SVM, Random Forest, SGD, and KNN. In an effort to employ the optimal version, we have calibrated the internal parameters of each program. Our constructed model demonstrated exceptional performance, especially when utilizing the K-Nearest Neighbour (KNN) algorithm, yielding an accuracy rate above 99.5% in the training set and over 99.3% in the test set.

ESSA ALHAZMI et al [7], the use of artificial intelligence, technology has grown dramatically in the last several years. Researchers and educators can forecast and model educational processes by utilizing the success, failure, dropout, and other metrics that the field of data mining in education offers. Thus, data mining is used in this study of data mining techniques to analyse student execution. The research uses a mix of clustering and classification algorithms to determine how early student performance affects GPA. In order to investigate the relationship between these factors and GPAs, the paper uses the T-SNE algorithm's dimensionality reduction mechanism for the clustering technique. It also incorporates a number of early factors, including admission scores from first-level courses, academic achievement tests (AAT), and general aptitude tests (GAT). We use five machine learning techniques for both training and testing our categorization models. These include the recently released classification method Xgboost as well as more well-known ones like Random Forest (RF), Support Vector Machine (SVM), K-nearest neighbour (KNN), and Logistic Recognition. The acronym for "eXtreme Gradient Boosting" is "xgboost." The goal function defined by gradient boost consists of two components: regularization and training loss. We assess how well these categorization algorithms are able to recognize student performance and find that supervised machine learning techniques that use our unique features outperform those that only use traditional characteristics. To ensure that better comprehend The efficiency of the with data dimensionality reduction by T-SNE, our paper demonstrated the application of machine learning techniques. It makes use of four criteria, including general aptitude test (GAT), academic achievement test (AAT), entrance scores, and first-level courses.

MAI ABDALKAREEM et al [9], over five million pupils, over 450 thousand academic members, and over one million parents or guardians used the Madrasty platform in 2020 during the COVID-19 pandemic in Saudi Arabia (SA). By the conclusion of the 17th week of the first semester, the platform had received 489 million internet visitors. In addition, during that brief time, teachers completed up to 89 million synchronized virtual lessons and produced over 15 million homework assignments for their pupils. Data are now seen as a crucial tool for making defensible decisions. Applying cutting-edge techniques is therefore advantageous for obtaining specific, practical knowledge. A subset of artificial intelligence known as machine learning (ML) compiles all techniques that enable a computer to learn and make precise predictions based on historical data. Five supervised learning classifiers, including DT, XGB, RF, NB, and ANN, were compared. In addition, this study advances the field by using XAI to interpret the top model in order to improve and better comprehend the prediction model's transparency. To ensure that in order to develop forecasting models that will address the research issues, five DM algorithms were employed.

SITI DIANAH ABDUL BUJANG et al [10], By leaving the base classifier alone to determine the decision output for a single class, the ensemble technique aims to improve accuracy. Based on our observations, we found that the majority of studies employed standard classifiers like DT, SVM, KNN, NB, RF, etc., and that very few studies looked into the possibility of using ensemble algorithms like Boosting, Stacking, and Bagging to enhance student grade prediction performance. When handling classes that are not balanced, it is more acceptable to highlight the performance of ensemble methods rather than just using a single classifier to predict the student's grade. The survey's findings show how widely unbalanced concerns for student grade prediction are identified utilizing the data-level technique with SMOTE oversampling. To improve student grade prediction performance, hybrid and feature selection techniques that enhance the predictive model's generalization are typically underutilized.

GHAZANFAR LATIF et al [12], in this work, the optimal algorithm was found by the application of supervised techniques. The classifiers, including BN, SVM, LR, RF, and FDT. It was shown that the precision of binary classification is greatly increased by ensemble approaches in conjunction with base learners of boosting and bagging, marginally increased for three class problems, and not much increased for four class problems. For binary classification, the ensemble algorithm of bagging and boosting FDT produced an accuracy of 98.25%, while for three classes, it reached 89.47%. For four classes, the standard ensemble FDT yielded an accuracy of 77.19%. Using the same dataset, the binary classification results were compared with those

found in the body of existing literature, demonstrating that the modified algorithms suggested by the authors outperformed similarly suggested techniques.

JIN EUN YOO et al [14], Specifically, Enet (elastic net) and Mnet were used amid regularization to extract, from LMS (learning management system) log data, students' instructional video watching behaviours at the instructional unit level. Consequently, regularization produced interpretable prediction models in the same way as a linear technique, and it also demonstrated prediction performance that was comparable to random forest, a nonlinear method well-known for its prediction capabilities. This study's initial research topic compared regularization with random forest (RF) prediction performance. While RF models are proven to be quite predictive, they can be challenging to understand. RF produces nonparametric models using decision trees as the foundation learner. In order to fit decision tree models on boot-strapped samples, RF first prepares boot-strapped samples. Then, the decision tree outputs are combined into an ensemble approach. The quantity of variables randomly selected as candidates at each split (mtry), the number of trees, the minimum quantity of observations in a terminal node, sampling with or without replacement, and the splitting criteria are some of the tuning parameters of random forests (RF). This study's regularization yielded prediction models that were not only comprehensible but also equivalent to random fields (RF). The intended accuracy rate was attained with the Random Forest method.

Yuanyi Zhen et al [18], to ensure that solve this problem, this study uses natural language processing techniques to evaluate conversations that take place in real-time classrooms on a major Chinese online learning platform. Emotional expression and interaction type characteristics are taken from classroom conversations. Next, using these attributes as a basis, we create neural network models that predict which children will perform well academically and poorly. Finally, we use interpretable artificial intelligence (AI) techniques to identify the most significant predictors in the prediction models. Models. High-achieving students regularly show more positive attitude, cognition, and off-topic discussions than low-performing students in all stages of the session, in both STEM (science, technology, engineering, and mathematics) and non-STEM courses. They employed natural language processing (NLP) in this work to achieve a number of benefits, including sentiment analysis, automated essay scoring, tailored learning, and improved data visualization. NLP can help data-driven decision making, increase student results, and better the learning process overall by utilizing these benefits. High accuracy is attained by NLP in tasks such as sentiment analysis, language modelling, and text categorization, particularly when utilizing sophisticated techniques like transforms and big datasets. When comparing NLP's accuracy to other algorithms like KNN, SVM, and Random Forest, keep the following things in mind.

III.CONCLUSION

The study gathers a dataset of students through this survey. These typical data sets—which include student information, academic history, learning habits, assessment results, and learning preferences and styles—are used to forecast students' performance. Six steps made up the study: problem definition, gathering the necessary dataset, pre-processing and preparation of the dataset, modelling and optimization to validate and implement the model. Gives a summary of the overview of the strategies and tactics employed to tackle the imbalanced class issue in the student grade prediction field. It covers cutting edge techniques, their effects, and potential remedies going forward. The goal is to offer a workable way to develop forecasting models that are more accurate. For the behavioural statistical study, eight representative standard machine learning algorithms from various families were employed. To be more precise, the learner's module can use pre-university, university admission, behavioural, and academic data to forecast outcomes. Furthermore, knowledge on SVM—the most widely used technique—followed by RF, LR, k-NN, and ANN/MLP can help learners choose or enhance machine learning methods that can be deployed. This study's regularization yielded prediction models that were not only comprehensible but also equivalent to random fields (RF). Numerous opportunities for utilizing big data in education will arise from increased regularization research in learning analytics.

REFERENCES

1. *DAOZONG SUN, RONGXIN LUO, QI GUO, JIAXING XIE, "A University Student Performance Prediction Model and Experiment Based on Multi-Feature Fusion and Attention Mechanism", 2023*
2. *KOUSHIK ROY AND DEWAN MD. FARID, "An Adaptive Feature Selection Algorithm for Student Performance Prediction", 2024*
3. *YONG SHI 1, FANG SUN2, HONGKUN ZUO3, AND FEI PENG1, "Analysis of Learning Behavior Characteristics and Prediction of Learning Effect for Improving College Students' Information Literacy Based on Machine Learning", 2023*
4. *NUHA MOHAMMED ALRUWAIS, "Deep FM-Based Predictive Model for Student Dropout in Online Classes", 2023*
5. *MUSTAPHA SKITTOU , MOHAMED MERROUCHI, AND TAOUFIQ GADI, "Development of an Early Warning System to Support Educational Planning Process by Identifying At-Risk Students", 2023*
6. *MUHAMMAD ADNAN, EMEL KHAN, FAHD S. ALHARITHI, AND AHMAD A. ALZAHRANI, "Earliest Possible Global and Local Interpretation of Students' Performance in Virtual Learning Environment by Leveraging Explainable AI", 2022*
7. *ESSA ALHAZMI AND ABDULLAH SHENEAMER, "Early Predicting of Students Performance in Higher Education", 2023*
8. *N. R. RAJI 1, R. MATHUSOOTHANA S. KUMAR2, AND C. L. BIJI, "Explainable Machine Learning Prediction for the Academic Performance of Deaf Scholars", 2024*
9. *MAI ABDALKAREEM AND NASRO MIN-ALLAH, "Explainable Models for Predicting Academic Pathways for High School Students in Saudi Arabia", 2024*
10. *SITI DIANAH ABDUL BUJANG1,2, ALI SELAMAT, "Imbalanced Classification Methods for Student Grade Prediction: A Systematic Literature Review", 2023*
11. *GABRIELA CZIBULA 1, GEORGE CIUBOTARIU1, MARIANA-IOANA MAIER1, "IntelliDaM: A Machine Learning-Based Framework for Enhancing the Performance of Decision-Making Processes. A Case Study for Educational Data Mining", 2022*
12. *GHAZANFAR LATIF 1,2, SHERIF E. ABDELHAMID 3, KHALED S. FAWAGREH1, "Machine Learning in Higher Education: Students' Performance Assessment Considering Online Activity Logs", 2023*
13. *AHMAD ALMADHOR 1, SIDRA ABBAS 2, (Graduate Student Member, IEEE), GABRIEL AVELINO SAMPEDRO, "Multi-Class Adaptive Active Learning for Predicting Student Anxiety", 2024*

14. *JIN EUN YOO1, MINJEONG RHO 1, AND YEKYUNG LEE2, "Online Students' Learning Behaviors and Academic Success: An Analysis of LMS Log Data from Flipped Classrooms via Regularization", 2022*
15. *NAVEED ANWER BUTT 1, ZAFAR MAHMOOD1, KHAWAR SHAKEEL1, SULTAN ALFARHOOD, "Performance Prediction of Students in Higher Education Using Multi-Model Ensemble Approach", 2023*
16. *Tarik Ahajjam, Mohammed Moutaib, Haidar Aissa, Mourad Azrour, Yousef Farhaoui, and Mohammed Fattah, "Predicting Students' Final Performance Using Artificial Neural Networks", 2022*
17. *LIDYA R. PELIMA, YUDA SUKMANA, AND YUSEP ROSMANSYAH, "Predicting University Student Graduation Using Academic Performance and Machine Learning: A Systematic Literature Review", 2024*
18. *Yuanyi Zhen, Jar-Der Luo, and Hui Chen, "Prediction of Academic Performance of Students in Online Live Classroom Interactions —an Analysis Using Natural Language Processing and Deep Learning Methods", 2023*
19. *DALIA ABDULKAREEM SHAFIQ, MOHSEN MARJANI, RIYAZ AHAMED ARIYALURAN HABEEB, AND DAVID ASIRVATHAM, "Student Retention Using Educational Data Mining and Predictive Analytics: A Systematic Literature Review", 2022*
20. *REYHAN ZEYNEP PEK 1, SIBEL TARIYAN ÖZYER2, TAREK, "The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure", 2023*