ISSN: 2633-4828 Vol. 6 No.1, January, 2024

International Journal of Applied Engineering & Technology

A STUDY OF STUDENTS ACADEMIC PERFORMANCE AND DROPOUT PREDICTION IN HEI USING MACHINE LEARNING MODELS

[1]K. Sangeetha and [2]Dr. N. Shanmuga Priya

[1]Research Scholar, Department of Computer Science (Graphics and Creative Design), Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore

[2]Associate Professor and Head, Department of Computer Applications, Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore

ABSTRACT

Educational data mining has become an effective tool for exploring the hidden relationships in educational data and predicting students' academic achievements. Higher education institutions record a significant amount of data about their students, representing a considerable potential to generate information, knowledge, and monitoring. Predicting student performance, preventing failure and identifying the factors influencing student dropout are issues that have attracted a great deal of research interest recently. The main aim of this study is to review the various machine learning models for students academic performance and dropout prediction. Implementing Education Data Mining can help higher education institutions to predict their student's performance. Predicting student's performance can help to predict the weak student and help the higher education institutions to make strategy and decision making related to student's performance improvement. The higher education institutions can take early step to prevent student fail or student drop out. After knowing the prediction, it is also expected to the weak student that they can improve their performance and achieve better score. This study analyzes the various machine learning models and its properties. The result shows the performance of the machine learning models.DoS attack.

Keywords: - Machine Learning, Students Performance, Dropout, Prediction, Classification, Higher Education Institution (HEI).

1 INTRODUCTION

Educational data mining (EDM) is the use of traditional DM methods to solve problems related to education (Baker & Yacef, 2009). EDM is the use of DM methods on educational data such as student information, educational records, exam results, student participation in class, and the frequency of students' asking questions. In recent years, EDM has become an effective tool used to identify hidden patterns in educational data, predict academic achievement, and improve the learning/teaching environment.

Learning analytics has gained a new dimension through the use of EDM (Waheed et al., 2020). Learning analytics covers the various aspects of collecting student information together, better understanding the learning environment by examining and analysing it, and revealing the best student/teacher performance (Long & Siemens, 2011). Learning analytics is the compilation, measurement and reporting of data about students and their contexts in order to understand and optimize learning and the environments in which it takes place. It also deals with the institutions developing new strategies. Modern learning institutions operate in a highly competitive and complex environment. Thus, analyzing performance, providing high-quality education, formulating strategies for evaluating the students' performance, and identifying future needs are some challenges faced by most HEIs today. Student intervention plans are implemented in HEIs to overcome students' problems during their studies. Student performance prediction at entry-level and during the subsequent periods helps the HEIs effectively develop and evolve the intervention plans, where both the management and educators are the beneficiaries of the students' performance prediction plans.

Reducing student dropout rates is one of the challenges facing in the education sector globally. The problem has brought a major concern in the field of education and policy-making communities (Aulck et al., 2016). Finding and implementing solutions to this problem has implications well beyond the benefits to individual students.

Moreover, enabling students to complete their higher education means investing in future progress and better standards of life with multiplier effects. To effectively address this problem, it is crucial to ensure that all students finish their higher education on time through early intervention on students who might be at risk of dropping classes. This requires data-driven predictive techniques that can facilitate determination of at-risk students and timely planning for interventions (Fei and Yeung, 2015).

Machine learning is a promising tool for building a predictive model for a student's performance and dropout. There have been several studies in EDM that focused on the predictions of Students Academic Performance and dropouts which featured the use of data mining techniques like Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT) and its variants, Artificial Neural Network (ANN), k-Nearest Neighbour (kNN), Logistic Regression (LR), Extreme Learning Machine (ELM) and ensemble methods like the Random Forest (RF); Adaptive Boosting (AB) and Bagging for predicting the students' academic performance and dropouts in particular. Other data mining techniques like the use of Nature-Inspired algorithms, Deep Learning and the hybrid methods had also been applied to predict SAP and dropouts even though the studies in this respect are scanty. These studies revealed the eligibility of data mining techniques to efficiently predict the performance of students at different levels of their studies using various academic performance dependent factors. However, there still exists the need for more researches in conducting and producing an improved framework for Students Academic Performance and dropout predictions. This paper is set out to review some related and relevant studies in this area and reveal their strengths and limitations with a view to identifying research gaps.

The rest of the paper is organized as follows. In section 2, reviews the various models or solution related to the problem of Students Academic Performance and dropout prediction. Section 3; explain the system model for the problem statement used in the proposed scheme. In section 4, the existing prediction algorithms and its performance is analyzed. Finally, the proposed work is summarizing in section 5.

2. RELATED WORK

A systematic literature review is performed with a research method that must be unbiased and ensure completeness to evaluate all available research related to the respective field. The objective is to collect, organize, and synthesize existing knowledge relating to machine learning approaches on student academic performance and dropout prediction. The surveyed papers focused on several works which have been done on machine learning in education such as student dropout prediction, student academic performance prediction, student final result prediction etc. The findings of these studies are very useful on understanding the problem and improving measures to address solution. We searched several databases such as ResearchGate, Elsevier, Association for Computing Machinery (ACM), Science Direct, Springer Link, IEEE Xplore, and other computer science journals.

Fernandes et al. (2019) developed a model with the demographic characteristics of the students and the achievement grades obtained from the in-term activities. In that study, students' academic achievement was predicted with classification models based on Gradient Boosting Machine (GBM). The results showed that the best qualities for estimating achievement scores were the previous year's achievement scores and unattendance. The authors found that demographic characteristics such as neighbourhood, school and age information were also potential indicators of success or failure. In addition, he argued that this model could guide the development of new policies to prevent failure.

Cruz-Jesus et al. (2020) predicted student academic performance with 16 demographics such as age, gender, class attendance, internet access, computer possession, and the number of courses taken. Random forest, logistic regression, k-nearest neighbours and support vector machines, which are among the machine learning methods, were able to predict students' performance with accuracy ranging from 50 to 81%.

Xu et al. (2019) determined the relationship between the internet usage behaviors of university students and their academic performance and he predicted students' performance with machine learning methods. The model he proposed predicted students' academic performance at a high level of accuracy. The results suggested that Internet connection frequency features were positively correlated with academic performance, whereas Internet traffic volume features were negatively correlated with academic performance. In addition, he concluded that internet usage features had an important role on students' academic performance.

Waheed et al., (2020) designed a model with artificial neural networks on students' records related to their navigation through the LMS. The results showed that demographics and student clickstream activities had a significant impact on student performance. Students who navigated through courses performed higher. Students' participation in the learning environment had nothing to do with their performance. However, he concluded that the deep learning model could be an important tool in the early prediction of student performance.

Rebai et al. (2020) proposed a machine learning-based model to identify the key factors affecting academic performance of schools and to determine the relationship between these factors. He concluded that the regression trees showed that the most important factors associated with higher performance were school size, competition, class size, parental pressure, and gender proportions. In addition, according to the random forest algorithm results, the school size and the percentage of girls had a powerful impact on the predictive accuracy of the model.

Musso et al., (2020) proposed a machine learning model based on learning strategies, perception of social support, motivation, socio-demographics, health condition, and academic performance characteristics. With this model, he predicted the academic performance and dropouts. He concluded that the predictive variable with the highest effect on predicting GPA was learning strategies while the variable with the greatest effect on determining dropouts was background information.

The literature review suggests that, it is a necessity to improve the quality of education by predicting the academic performance of the students and supporting those who are in the risk group. In the literature, the prediction of academic performance was made with many and various variables. From the literature review, it can be found that EDM models can be divided into two types: descriptive models and predictive models. Descriptive models are used to describe models and provide reference for decision-making, whereas predictive models are mainly used for data-based prediction. The former is mostly used to evaluate students' academic performance and provide a reference for teaching managers to make decisions, whereas the latter is mostly used to predict students' academic performance, help prevent the risk of dropout, and improve students' academic performance.

3 STAGES OF MACHINE LEARNING MODELS

The online machine learning repository provides us with the dataset containing more than 261 student records both in the training and testing dataset with 10 parameters. The dataset contains information of the students in the form of exam, access, test grades, project, assignments, result points, tests, result grades, graduate project grade, and year. We are only concerned with few parameters which affect the drop out value.

A. **Data Collection** - In this phase, we gathered dataset from the Kaggle to evaluate and compare the performance of predictive machine learning techniques used in this project. In this experiment, data for independent features like exam, access, test grades, project, assignments, result points, tests, result grades, graduate project grade, and year were collected from several institutions in the form of a ".csv" file which results in the prediction of the dependent feature i.e., dropout or no dropout. We are concerned mainly with the three parameters like access, results and assignments to predict student drop out.

Pre-Processing of the Data - Data Cleaning & Transformation: One of the crucial steps in the pre-processing of data stage is data cleaning. In this phase, the undesired data is eliminated, and the missing values or NA values are rectified. We then remove a few inaccurate and outlier data points that could lead to mistakes in our prediction models.

B. **Feature Engineering**: Access, assessments, and tests were chosen as significant elements that influence the dropout outcome in this experiment based on the significance of the factors for dropout prediction.

a. Accesses are the total count of a student saw a course throughout the observational period.

b. Assignments indicate the overall score from all the actions that were evaluated throughout the observation time.

c. The results of the tests taken during the semester are the total of the results from the midterm and final exams.

As a final point, the variables access, assignments, and tests were selected because their importance lies in making predictions as early as possible so that intervention can be initiated. It must be emphasized that only the total number of partial tests conducted throughout the observation period was included. In addition to the two variables previously described with the number of accesses & good correlation (tests and projects), may be a more appropriate feature for prediction. The presence or frequency of access may allow us to identify students who do not complete the course in a timely manner, thus making them more likely to fail.

C. Model - Various machine learning methods are used to the dataset during the modelling process. A model is created based on students' current actions and successes in order to predict learner failure and performance in the future. An algorithm or classifier can resolve this typical classification problem by determining whether a student can complete the course. To find a good predictive model, one can utilize one of the many various Machine Learning categorization techniques available today. The needs coming from the paper's primary goal and related works were taken into consideration, and the most widely used classifiers were ultimately used:

1) **Neural network**: This is a paralleled processing method that maximizes predicted accuracy by utilizing the structure and operations of the brain. Data is inputted into a deep learning NN, which subsequently analyses the data to produce the output across multiple layers.

2) **Logistic Regression**: In order to understand the data and characterize the relationship between an independent variable & a dependent variable that could be ordinal, interval in nature, nominal, or one method of prediction analysis called logistic regression is used. Finding the best model to explain the correlation between a group of independent factors & Boolean characteristic of interest and is the aim of this machine learning model, because the result's dependent variable is a binary variable.

3) **Naïve Bayes**: This classifier is a probabilistic straightforward classifier based on the theorem of bayes with stringent feature independence conditions.

4) **Decision Tree**: Decision trees are frequently utilized to solve classification & regression issues. The overarching objective of using decision tree is to construct a model which produces rules of decision from evaluating data sets and forecasts target classes. The decision tree adopts a structure of tree consisting of branches, roots, and leaves. In contrast to leaf nodes, which represent class labels, internal nodes represent decision-making qualities. Compared to other categorization strategies, the DT approach is easier to comprehend.

5) **Support Vector Machine**: It is a linear regression & classification model that may be applied to both nonlinear & linear problems. The algorithm divides the input into categories using a hyper plane. The value of each characteristic is represented by the value of a particular coordinate in this model, and in n dimensional space of points every item of data will be displayed as a point.

6) **Random Forest**: The Random Forest method generates a large number of ensembles of Classification & regression decision trees. A number of trees of decision are printed using a randomly chosen subset of the training datasets. The use of many decision trees increases the accuracy of the results. The algorithm allows missing data and has a relatively short runtime. Random forest is used to randomize the algorithm rather than the training data set. The sort of class that decision trees produce is the decision class.

The main goal of this paper is the comparison of the numerous prediction performances indicators of each used classifier in order to choose the most appropriate prediction model. The grid search method, which is frequently used to determine the best parameter settings, was utilized to modify hyper parameters.

Table 1: Comparative Study of Machine Learning Models

Dataset	Existing Algorithm (Logistic Regression) (%)	Existing Algorithm (XGBoost) (%)	Existing Algorithm (Decision Tree) (%)	Existing Algorithm (Random Forest) (%)
UCI - Student Performance Data Set	91.5	93	92.6	93
Kaggle - Student Performance Data Set	90.8	91.5	92	91.7

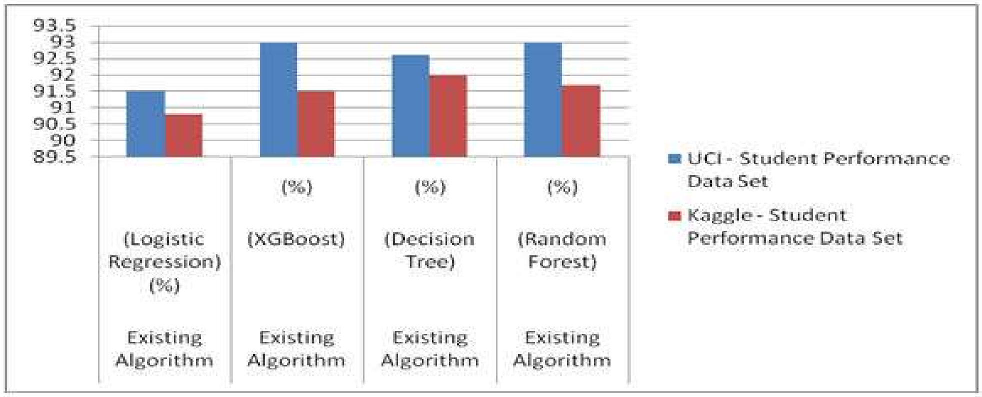

Fig.2: Comparison of Existing Models

4 PROPOSED METHODOLOGY

The aim of the proposed research is to build a model by utilizing appropriate feature selection techniques and classification algorithms to improve accuracy. The four phases of the approach used in this study include data collection, feature selection and pre-processing, data visualization, and classification for the student's class. The phase I deals with data collection and Preprocessing, Phase II is feature extraction techniques. Phase III is concerned with feature selection is to divide the dataset into training and testing. Phase IV is concerned with classification algorithms to build a model for prediction, and the evaluation purposes.

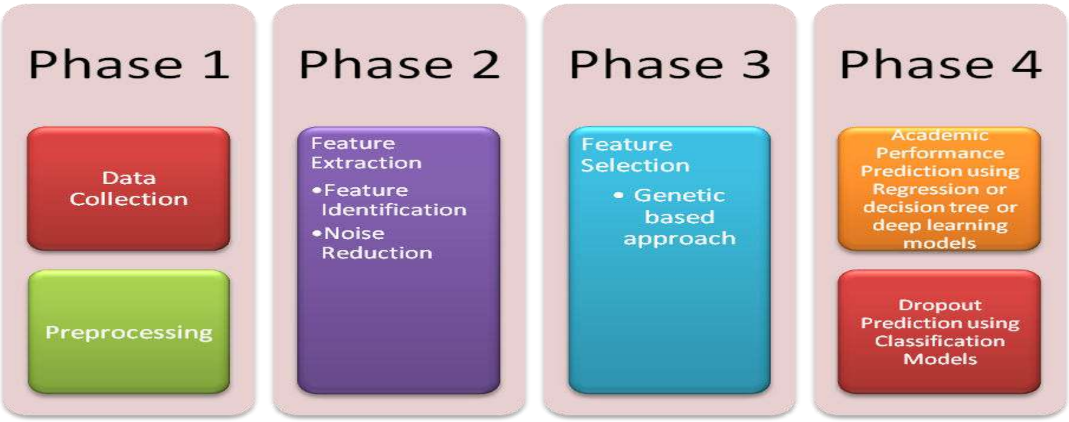

Fig. 2: Proposed Methodology

The proposed methodology provides the following benefits,

- Predictive models for students' academic performance, which can provide early warnings about struggling students.
- A predictive model for identifying students at risk of dropping out, enabling timely intervention.

- Insights into the factors that influence students' academic performance and dropout risk, which can inform institutional strategies and policies

4. EXPERIMENTAL SETUP

Evaluating classifier efficiency is an important part of comparing and selecting the best one. There are several methods for measuring and evaluating the output of machine learning algorithms. This paper employs a variety of assessment methods, including prediction Accuracy, Sensitivity, Precision, and F1-score; additionally, a statistical evaluation technique is employed for more reliable and efficient analyzing and comparing. Analyzing and evaluating the output of the classifiers is an essential technique. While assessment methods are easy to use, the results obtained can be misleading. Seeking the right model or system based on their strengths is therefore a crucial challenge. Evaluating classifier efficiency is an important part of comparing and selecting the best one. There are several methods for measuring and evaluating the output of machine learning algorithms. There are five commonly used different measures for evaluating classification consistency. Details are as follows:

- **CCI (Correctly Classified Instances)**: represents the number of correctly identified instances divided by the total number of instances Precision is a term that is commonly used.
- **ICI (Incorrectly Classified Instances)**: represents the number of instances that were wrongly labelled divided by the total number of instances.
- **Precision**: of algorithm represents the percentage of accurate classified instances from all truly classified instances.
- **Recall**: reflects the division number of correctly classified instances by the total number of all instances.
- **F-Measure**: measured from recall and precision values.

5. CONCLUSION

The primary aim of research is to significantly predict student performance and dropout in order to enhance academic outcomes. This can be done by the use of various educational data mining techniques to provide high-quality education. One way to achieve the highest degree of quality in the higher education sector is by accurate estimation of students' learning in educational institutions. There are numerous prediction models available using different mining techniques. Existing policies have largely been unable to respond to the increasing demands for higher and master training as mandated by the education framework. The current models are reviewed in this paper, and a novel model is proposed to effectively predict student success. This research work aims to specify the challenges and opportunities of quality education in higher education institutions, as well as provide a model for improving education quality.

6. REFERENCES

[1] Baker, R. S., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of educational data mining, 1(1), 3-17.

[2] Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. Computers in Human behavior, 104, 106189.

[3] Long, P., & Siemens, G. (2011). What is learning analytics. In Proceedings of the 1st International Conference Learning Analytics and Knowledge, LAK (Vol. 11).

[4] Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2016). Predicting student dropout in higher education. arXiv preprint arXiv:1606.06364.

[5] Fei, M., & Yeung, D. Y. (2015, November). Temporal models for predicting student dropout in massive open online courses. In 2015 IEEE international conference on data mining workshop (ICDMW) (pp. 256-263). IEEE.

[6] Cruz-Jesus, F., Castelli, M., Oliveira, T., Mendes, R., Nunes, C., Sa-Velho, M., & Rosa-Louro, A. (2020). Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country. Heliyon, 6(6).

[7] Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting student dropout and academic success. Data, 7(11), 146.

[8] Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. Computers in Human Behavior, 98, 166-173.

[9] Rebai, S., Yahia, F. B., & Essid, H. (2020). A graphically based machine learning approach to predict secondary schools performance in Tunisia. Socio-Economic Planning Sciences, 70, 100724.

[10] Musso, M. F., Hernández, C. F. R., & Cascallar, E. C. (2020). Predicting key educational outcomes in academic trajectories: a machine-learning approach. Higher Education, 80, 875-894.

[11] Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learning Environments, 9(1), 11.

[12] Lee, S., & Chung, J. Y. (2019). The machine learning-based dropout early warning system for improving the performance of dropout prediction. Applied Sciences, 9(15), 3093.

[13] Del Bonifro, F., Gabbrielli, M., Lisanti, G., & Zingaro, S. P. (2020). Student dropout prediction. In Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part I 21 (pp. 129-140). Springer International Publishing.

[14] Chung, J. Y., & Lee, S. (2019). Dropout early warning systems for high school students using machine learning. Children and Youth Services Review, 96, 346-353.

[15] Rai, S., Shastry, K. A., Pratap, S., Kishore, S., Mishra, P., & Sanjay, H. A. (2021). Machine learning approach for student academic performance prediction. In Evolution in Computational Intelligence: Frontiers in Intelligent Computing: Theory and Applications (FICTA 2020), Volume 1 (pp. 611-618). Springer Singapore.

[16] Fernández-García, A. J., Preciado, J. C., Melchor, F., Rodriguez-Echeverria, R., Conejero, J. M., & Sánchez-Figueroa, F. (2021). A real-life machine learning experience for predicting university dropout at different stages using academic data. IEEE Access, 9, 133076-133090.

[17] Prenkaj, B., Velardi, P., Stilo, G., Distante, D., & Faralli, S. (2020). A survey of machine learning approaches for student dropout prediction in online courses. ACM Computing Surveys (CSUR), 53(3), 1-34.

[18] Zeineddine, H., Braendle, U., & Farah, A. (2021). Enhancing prediction of student success: Automated machine learning approach. Computers & Electrical Engineering, 89, 106903.

[19] Bhutto, E. S., Siddiqui, I. F., Arain, Q. A., & Anwar, M. (2020, February). Predicting students' academic performance through supervised machine learning. In 2020 International Conference on Information Science and Communication Technology (ICISCT) (pp. 1-6). IEEE.

[20] Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. Computers and Education: Artificial Intelligence, 3, 100066.