International Journal of Early Childhood Special Education (INT-JECS)
DOI:10.9756/INTJECSE/V14I7.60 ISSN: 1308-5581 Vol 14, Issue 07 2022

Investigating Student Learning Process and Predicting Student Performance Using Machine Learning Approaches

E. Sandhya
(Assistant Professor, Department of IT, Sree Vidyanikethan Engineering College, Tirupati)

RamPrakash Reddy Arava
(Assistant Professor, Department of CSE, K.S.R.M Collge of Engineering, Kadapa)

Dr. E. S. Phalguna Krishna
(Assistant Professor, Department of CSE, GITAM School of Technology, Bengaluru)

Dr. K.K. Baseer
(Professor, Department of IT, Sree Vidyanikethan Engineering College, Tirupati)

ABSTRACT

Interpreting and forecasting student grades in today's higher education institutions have become a difficult task. Predicting and assessing student achievement is a critical task at institutions like universities, colleges, and school. The main aim is to concentrate on developing an improved algorithm and application to understand student performance prediction in educational system. The performance measure can be improved by using several classifiers and comparing them to previous research to investigate student performance prediction. The models are more precise than traditional models. The proposed work is evaluated using student academic performance which is taken from Kaggle website. The information gain and entropy values are used to choose the required characteristics from all of the features. The technique used for selecting the features is correlation coefficient which is one among the filter method. The models used are Naïve Bayes, XG Boost, Decision Tree and Hybrid model. Hybrid model is the combination of XG Boost and Random Forest algorithms. An accuracy of 98.39%, 96.42%, 86.55% and 74.13% is obtained for Hybrid Model, XG Boost, Decision Tree and Naïve Bayes. Among all the models the better accuracy is obtained for Hybrid model.

Keywords: SMOTE, Decision Tree, Naïve Bayes.

1. *INTRODUCTION*

Many organizations of higher education are set up across the world. So, education quality is judged by rate of student success and helps to determine their degree to pursue. The capability of students helps them to gain more knowledge and have a secure platform in future. So, Predicting and Interpreting of student performance is an important task in present world. Higher education institutions are concerned about precise estimates and early identification of students who may be destroyed. In terms of improving academic quality and making the best use of available resources for real-world intervention, it is crucial not only for students, but also for educational administrators and institutions.. However, only minor success has been observed in terms of correctness, competence, and reduction in student disruption across the various outlines and models that academics have employed across universities for forecasting performance. Using data mining techniques, we can analyse and interpret performance of students based on prior courses and assist them in improving their academic records. Data mining concept is the one which helps to get hidden information and know the relation between the attributes in huge datasets [8]. There are many accomplishments of Data Mining techniques in many parts such as engineering,

schooling, advertising, remedial, and economic. Data mining technique has the ability to provide alternative solutions while decision making which helps in avoiding the future the problems that arise later in the area of decision making [9]. The investigation data in educational field using DM techniques are called as Educational Data Mining. Education data mining is a wide area that provides machine learning, statistical information and data mining algorithms to find out essential datasets in education. Data Mining in education is a huge platform where it consists of different a application which helps us build applications for student's performance prediction like finding in what category does the student fall and helps faculty in taking decisions. Educational data can be gathered in various forms but mostly it is taken from educational institutions [10].

While making a prediction on student's performance we must take care of various factors such as their financial status, personal interests (what to be studied), and family problems. An assurance should be made that the parent/Guardian and student like the institution policies so that they can have some freedom to think and concentrate on their instruction by having good creative skills. Classification, clustering, and relationship mining are some of the data mining methods that may be used in educational contexts to forecast a undergraduate's presentation [11]. K-Means, Naive Bayes and Decision Tree approaches [7] are used to forecast student performance and inform instructors about vulnerable students so they may get the assistance they need. As a result, building a learning classifier with students' observed records as the training set and matching student past data, which serves as features, with their label, which serves as the actual appearance, is required..

The ultimate goal is for teachers to pay attention to students who are on the verge of dropping out and offer advice for making better performance and improve their graduation rates. Using various data mining approaches to address the challenges of class inequity that necessitate solutions like data augmentation for the minority class, often known as SMOTE [12], to improve the accuracy of student performance models without compromising their interpretability and also by merging two prototype types: bagging and boosting approaches.

For the first time, the XG Boost algorithm was tested on student performance data. We employed a hyperparameter optimization strategy to increase accuracy and minimise model overfitting by selecting the best hyperparameters for the learning process. Our models were trained and evaluated using tenfold cross validation, and the algorithms' performance was analysed using appropriate metrics.

2. *RELATED WORKS*

Accurate forecasting and identification of students at danger of dropout plays a top concerns for Institutions of higher learning. It is critical not only for students, but also for educational institutions to improve academic quality and make effective use of existing resources. However, academics have utilised a variety of frameworks and models to predict performance across institutions [1]. Many data mining techniques are being developed to cite previously unknown material from educational data. The information gathered assists institutions in improving their teaching techniques and educational processes. All of these improvements help to improve undergraduate presentation as well as educational outputs in general. Student performance prediction prototype based on data mining algorithms with novel data characteristics and features, dubbed "student behavioural features." These characteristics have to do with the learner's interaction with the e-learning management system. A group of classifiers, particularly Artificial Neural Networks, Decision Tree and Naive Bayesian [2] assess the routine of the student's prediction model. As a result of the digitalization of theoretical courses, universities are producing a tremendous amount of data influencing students in electronic form. They must translate this massive volume of data into information that will help teachers, administrators, and policymakers make better decisions. It may also help to improve the quality of educational processes by disseminating timely information to various participants. The goal of data mining technologies is to extract useful information from data. Educational Data Mining (EDM) suggests five basic categories or methodologies

for the application of data mining technologies to educational data: prediction, clustering, connection mining, discovery inside models, and data distillation for human judgement [3]. The demand for analysing the data and eliminating the valuable information has become a common concern and a rich academic subject of inquiry for many scholars as massive data warehouses grow more prevalent. Data mining techniques are employed as logical tools to extract hidden information from data warehouses in the form of models. Finance, marketing, the economics, telecommunications, medical, healthcare, and student routine applications are just a few of the domains that use data mining techniques in their systems. Academic institutions, on the other hand, create a vast amount of data influencing students through computerised forms as a result of their position of forecasting student success and digitising university system It is critical for these academic institutions to transmit vast amounts of data and turn it into usable information. Its goal is to make it easier for educators, employees, and authorities to complete their duties [4]. Educational Data Mining is large platforms that contains a variety of applications that enable us construct apps for student performance prediction, such as determining which category a student belongs to and assisting faculty in making judgments. Educational data can be collected in a variety of ways, although it is most commonly obtained via educational institutions. The Data Mining (DM) idea is one that aids in the discovery of hidden information and the understanding of the relationships between variables in large datasets. Data mining techniques have made significant contributions in a variety of fields, including engineering, education, marketing, medicine, finance, and sports. The Data Mining approach has the potential to propose alternative answers while making decisions, which aids in preventing future difficulties in the decision-making domain. Educational Data Mining is the research of data in the educational area using Data Mining methodologies [5]. Academic failure among university students is one of the most serious issues affecting higher education. One solution is anticipating performance of students in order to guide the students in improving their performance. Using data mining we can analyse and interpret information from prior courses. Classification, clustering, and relationship mining are approximately of the data mining methods that may be used in educational contexts for student's performance predication [6]. Clustering and classification and association rules are some of the methodological approaches used for data analysis in Data Mining. One of the prediction strategies is classification, in which data is categorised based on a training set and then the pattern is utilised to predict current information (testing set). Clustering is the process of grouping records into categories that are similar but not identical.

3. *PROPOSED WORK*

According to our research and extensive literature review, Random Forest works well for predicting student performance with high accuracy, but tree-based algorithms like XG Boost can improve it even more. It has the ability to boost performance significantly. We also employed supervised and unsupervised learning techniques. The data is saved on a cloud server, making it accessible from anywhere, and the system generates forecasts and offers advice to the student. The accuracy of the forecast is improved by using ensemble methods. Train and test data are used to pick the primary features. The outcome is analyzed using decision trees as a graphic representation. The techniques utilized are scalable and deliver accurate results, unlike previous machine learning models. It takes less time to train the algorithm since it takes less time to execute. We also addressed the issue of class inequality, which necessitated the use of techniques such as data oversampling, also known as the SMOTE technique. SMOTE was used as a balancing method, oversampling all classes to the majority class's sample size. We employed a hyperparameter optimization strategy to improve precision and avoid model overfitting by identifying a set of optimum hyperparameters for the learning process. Our models were trained and evaluated using ten-fold cross validation. Data Preparation of the raw data is cleaned up by removing null values, outliers, and undesired properties. SMOTE was used to pre-process the data. The relevant characteristics are chosen from all the features based on the information gain and entropy values. The co-relation coefficient, a filter approach, was used to pick the features. Figure1.

depicts the block diagram of proposed work

I. Dataset Collection:

The Kaggle repository contains an opensource dataset named student academic performance [13]. This dataset has 17 features. Among 17 features, Discussion, Announcement Views, Visited Resources, and Raised Hands are numerical features, while the rest are categorical values.

II. Data Preprocessing:

The activities that must be made to modify encoded information so that it may be processed by the computer are referred to as data pre-processing. For a model to be reliable and exact in predictions, the method must be able to quickly interpret the features of the data. Data pre-processing can help with the following issues.

- Outliers and unanticipated data points may cause the model's entire learning to be interrupted, resulting in incorrect predictions, as a result of duplicate or missing values.

III. Data cleaning:

Data cleaning is mostly done as part of data preparation to clean up the data by filling in missing values and smoothing down noisy data, resolving variance, and removing outliers. Clustering is used to reduce noisy data and eliminate a random mistake or variation in a restricted variable. The groups are created from data with similar values. The values that do not belong in the cluster can be saved as noisy data and discarded.

IV. Feature selection

Feature selection is a method of limiting the input variable to your model by using only relevant data and discarding noisy data. It's the process of determining which features to include in your machine learning model based on the type of problem you're trying to solve automatically. It helps us reduce the amount of data we input as well as the level of noise in our data. Following selection of feature, the data will be trained and tested. Training the model entails creating the model, and testing the model entails determining the model's correctness.

V. Multi Classifiers

Here, various algorithms are used for predicting student performance based on accuracy. The algorithms used are Decision trees, XG Boost and K-Means Clustering.

a. Decision Trees

A decision tree is a flowchart-like structure that helps you make decisions and is easy to understand. Decision Trees are a type of supervised learning that can be used to solve classification and regression problems. It's a graphical depiction for acquiring all possible options to a decision in a given situation. It selects the best attribute using the Attribute Selection Measure to split the data. A root tree with no incoming edges and an inside node with outgoing edges make up the decision tree. The remaining nodes are known as leaves or decision nodes.

b. Naive Bayes

The Naive Bayes algorithm is a supervised learning approach based on the Bayes theorem for dealing with classification problems. It's commonly utilized in applications like text categorization that require a big training dataset. The Naive Bayes Classifier is a probabilistic classifier that predicts outcomes

based on the likelihood of an item. It's one of the most straightforward and effective classification methods for creating rapid machine learning models that can make accurate predictions.

c. XG Boost (extreme Gradient Boosting)

In XG Boost, gradient-boosted decision trees are implemented. XG Boost dominates structured or tabular datasets in classification and regression predictive modelling issues. XG Boost offers features such as Gradient Boosting, Stochastic Gradient Boosting, and Regularized Gradient Boosting, as well as System Features for use in a range of computing environments.

d. Hybrid Model

Hybrid model is the combination of two or more algorithms which help in improving the accuracy of the application. Here, two algorithms are composed together in order to rise the accuracy of the project. XG Boost and Random Forest algorithms are combined together here to improve the performance.

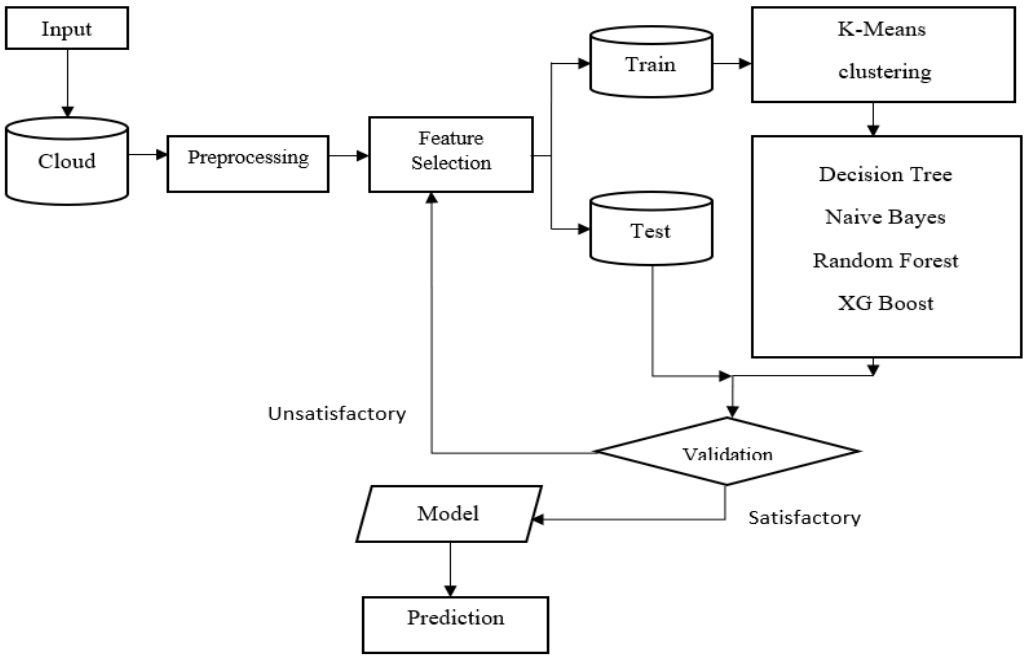

Fig 1. Block diagram of proposed work.

4. *RESULTS*

With XG Boost, Decision Tree, and Nave Bayes, Hybrid Model offers the highest accuracy. As a consequence, our models outperform traditional models in terms of accuracy. The computation models achieve a precision of over 98%, which matches the result obtained in a previous study using the same dataset. These results were acquired using approaches and processes such as SMOTE, hyperparameter optimization, and cross-validation, all of which indicate the robustness of the unique model. Precision and recall are machine learning presentation characteristics that help identify and categorize patterns. The model's ability to classify positive samples is measured using precision. Recall is a measurement of how many positive samples the ML model correctly categorizes.The data show that our models are more accurate than traditional ones. Figure 2 depicts a comparison of various multi classifiers.

The prediction models outperform the findings of a research that employed the same dataset, with an accuracy of over 97.83%. Figure 3 and Table 1 depicts a comparison of various multi classifiers accuracy results before and after employing the SMOTE technique.

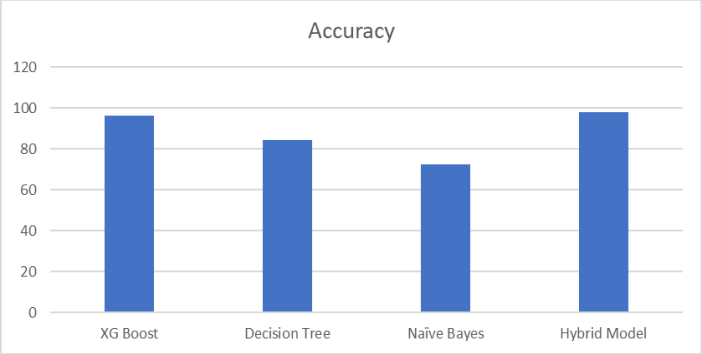

Figure 2: Comparison of various multi-classifiers

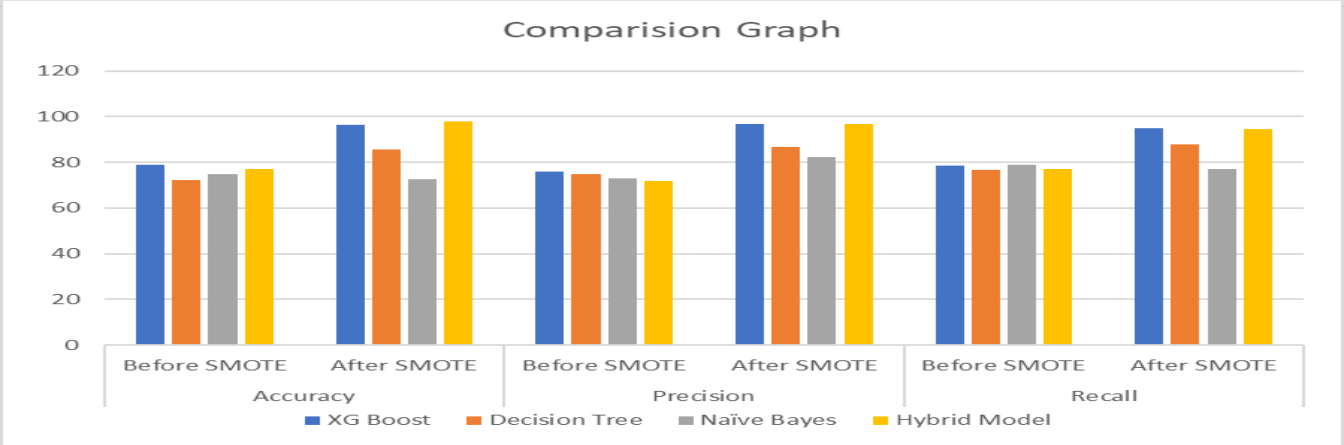

Figure 3: Comparison of various multi-classifiers.

Table 1 Comparison of various Machine Learning algorithms using SMOTE and without SMOTE.

Algorithms	Accuracy		Precision		Recall	
	Before SMOTE	After SMOTE	Before SMOTE	After SMOTE	Before SMOTE	After SMOTE
XG Boost	78.98	96.28	75.89	96.86	78.67	94.89
Decision Tree	72.25	85.52	74.98	86.89	76.84	87.91
Naïve Bayes	74.86	72.45	72.96	82.26	78.92	76.89
Hybrid Model	76.93	97.83	71.94	96.84	76.94	94.45

5. *CONCLUSION AND FUTURE WORK*

Prediction of student performance is an emerging situation which is needed to be focused and make better analysis so that the deserved one can be benefited. Here various algorithms are used to predict the performances and make analysis accordingly. Different approaches provide numerous results when there is a change in the split size. In all the algorithms a good accuracy is obtained only to one algorithm. The

algorithm to which a good accuracy is obtained is the hybrid model which is the combination of XG Boost and Random Forest algorithms. This provides the best accuracy and is very effective to use and also easy to understand. We are currently utilizing a dataset from the Kaggle repository. In the future, we will employ different datasets to expand the scope of identifying student performance utilizing dynamic selection methods in order to improve the model's performance, which might be valuable in educational institutions.

VI. REFERENCES

[1] F. Razaque, N. Soomro, S. A. Shaikh, S. Soomro, J. A. Samo, N. Kumar, and H. Dharejo, ‘‘Using Naïve Bayes algorithm to students’ bachelor academic performances analysis,’’ in Proc. 4th IEEE Int. Conf. Eng. Technol. Appl. Sci. (ICETAS), Dec. 2017, pp. 1–5.

[2] K. B. Bhegade and S. V. Shinde, ‘‘Student performance prediction system with educational data mining,’’ Int. J. Computer. Appl., vol. 146, no. 5, pp. 32–35, Jul. 2016.

[3] E. A. Amrieh, T. Hamtini, and I. Aljarah, ‘‘Mining educational data to predict student’s academic performance using ensemble methods,’’ Int. J. Database Theory Appl., vol. 9, no. 8, pp. 119–136, 2016.

[4] O. Adejo and T. Connolly, ‘‘An integrated system framework for predicting students’ academic performance in higher educational institutions,’’ Int. J. Computer. Science. Information. Technology, vol. 9, no. 3, pp. 149–157, Jun. 2017.

[5] B. Sen, E. Ucar, and D. Delen, ‘‘Predicting and analyzing secondary education placement-test scores: A data mining approach,’’ Expert Syst. Appl., vol. 39, no. 10, pp. 9468–9476, Aug. 2012.

[6] M. Yalcintas and U. A. Ozturk, ‘‘An energy benchmarking model based on artificial neural network method utilizing U.S. Commercial buildings energy consumption survey (CBECS) database,’’ Int. J. Energy Res., vol. 31, no. 4, pp. 412–421, 2007.

[7] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han, and X. Zhao, ‘‘A review of data-driven approaches for prediction and classification of building energy consumption,’’ Renew. Sustain. Energy Rev., vol. 82, pp. 1027–1047, Feb. 2018.

[8] P. Arjunan, K. Poolla, and C. Miller, ‘‘EnergyStar++: Towards more accurate and explanatory building energy benchmarking,’’ Appl. Energy, vol. 276, Oct. 2020, Art. no. 115413.

[9] C. Konstantinou, ‘‘Cyber-physical systems security education through hands-on lab exercises,’’ IEEE Des. Test. Comput., vol. 37, no. 6, pp. 47–55, Dec. 2020.

[10] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, ‘‘Prediction of student academic performance by an application of data mining techniques,’’ in Proc. Int. Conf. Manage. Artif. Intell., 2011, vol. 6, no. 1, pp. 110–114. s

[11] G. Gray, C. McGuinness, and P. Owende, ‘‘An application of classification models to predict learner progression in tertiary education,’’ in Proc. IEEE Int. Advance Comput. Conf. (IACC), Feb. 2014, pp. 549–554

[12] Karthik, M.G., Krishnan, M.B.M. Hybrid random forest and synthetic minority over sampling technique for detecting internet of things attacks. J Ambient Intell Human Comput (2021). DOI:10.1007/s12652-021-03082-3

[13] Kaggle, Kaggle.com[online], Available: https://www.kaggle.com/datasets.