International Journal of Modern Computation, Information, and Communication Technology
2021;4(11-12):49-54.
ISSN: 2581-5954

A Data Mining Approach for the Prediction of Low Performing Students

Yeligeti Raju[1*], LNC.Prakash K[2], Channapragada Rama Seshagiri Rao[3], Narendhar Mulugu[4]

[1]*Department of Information Technology, Vignana Bharathi Institute of Technology, Hyderabad, Telangana, India.*
[2]*Department of Computer Science and Engineering, CVR College of Engineering, Hyderabad, Telangana, India.*
[3]*Department of Computer Science and Engineering, Vignana Bharathi Engineering College, Hyderabad, Telangana, India.*
[4]*Department of Computer Science and Engineering, Malla Reddy Institute of Engineering and Technology, Hyderabad, Telangana, India.*

[*]*Corresponding author: dr.kamalakannan@staffemail.apu.edu.my*

Abstract. By maximising the greatest results and decreasing the number of students who fail, educational institutions aim to improve the quality of education. Improving and developing academic achievement requires the ability to accurately estimate student performance in advance. Using data mining techniques, educational institutions can get valuable insights from the study of educational data. The purpose of this research is to present educational data mining approaches and incorporate them into a web-based platform for forecasting low-performing pupils. Prediction models were compared in a comparative study. High-performance models were then designed to get better results. Hybrid RF, a hybrid random forest algorithm, is the most effective at classifying data. In order to enhance the learning outcomes and intervention methods, a new feature selection approach called MICHI is presented, which is the mixture of similarity measure and chi-square techniques based on ranking feature scores. Using the suggested methodologies for educational data mining, a system for predicting student academic performance will be constructed for academic entrepreneurs to obtain an early forecast of student achievement for prompt treatment. Experiments and surveys demonstrate the utility and efficacy of the academic prediction system that was built. Use of the system aids educators in intervening and enhancing student outcomes.

Keywords: Prediction, data mining, feature extraction, random forest, academic performance

INTRODUCTION

Every country's prosperity and long-term economic strength rely on the quality of its educational system. Undereducation and a scarcity of highly trained workers are directly linked to low performance in emerging countries. As a result, improving students' educational outcomes is a top priority for educational institutions. Academic performance and the quality of the learning experience have been the focus of educational institutions in recent years. It is a primary objective of educational systems across the world to raise the great outcome and lower the failure rate of students who are underachieving [1].

Due to their low performance, educational institutions are concerned that these students are at risk of failing, dropping out, or retaking classes. Since at-risk pupils can only be properly identified effectively enough through prediction, it has lately been one of the most successful strategies for solving this problem. Because of this, researchers believe that using early prediction to spot kids at risk and in need of support and intervention is an effective strategy. Innovation and information technology have established their importance in several fields during the last decade. It is the study of the use of data gathering, computer vision, and analytics in educational contexts to investigate educational data [2].

Data mining, machine learning, statistics, psycho-pedagogy, cognitive psychology, and recommender system methodologies and techniques are just a few of the multidisciplinary disciplines of study that are brought together by EDM to tackle educational problems. Effective EDM strategies were needed in educational settings to discover student learning patterns and provide intervention and build up a strategy to improve academic performance in diverse management settings, planning, and scheduling [3].

Received: 12.10.2021 Revised: 19.11.2021 Accepted: 29.11.2021

Monitoring and forecasting academic achievement have been made easier by the introduction of various analytic tools. To begin, you'll need to gather the relevant information, which can be found in school databases or by filling out surveys. For this reason, the second phase in the preprocessing process is to clean and transform raw data into a form that can be used by a computer. The experiment's endpoint will be reached in the third stage, which involves implementing certain EDM approaches [4].

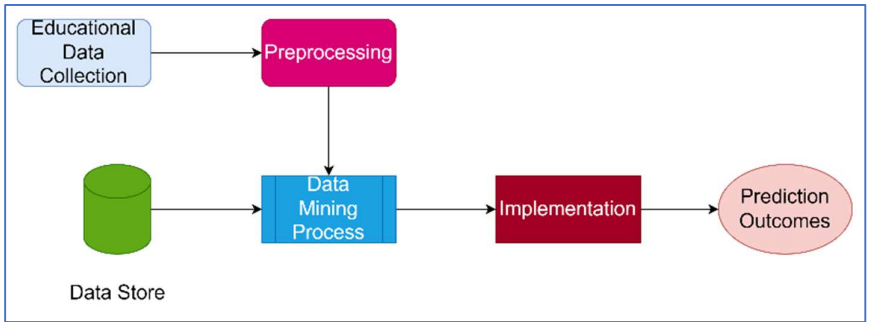

FIGURE 1. Overview of Educational Data Mining

The evaluation of experimental outcomes provides solutions to educational problems and decision-making. Last but not least, either alter the training programme or postpone it until a new study can be undertaken to obtain a more precise outcome. There are many critical issues about education that may be answered by using educational data mining (EDM). EDM approaches have been used in a number of academic performance studies [5]. According to the goals of the research, a variety of methods have been used. In order to distinguish the features of data, as well as its complexity and contribution significance to a project, sophisticated EDM techniques are needed.

EXISTING WORKS

Students in higher education are the focus of most academic research, yet there is a lack of research on student performance evaluation in high school. Students' competence in high school is an important measure of academic sector development since it pertains to students' basic knowledge for upper secondary. Students who are considered at risk of failing in high school must receive the appropriate intervention and improvement in order to improve their low-test scores [6].

Students who have a difficult time in school are more likely to fail the national test and struggle in college. Electronic data management (EDM) is employed in the context of academic underachievement to provide prompt intervention and improvement. Models for EDM were built in this work and used to predict high school pupil achievement on a web-based system. Estrera et al. used high school data to predict student performance for academic ranking in a Philippine institution [7].

This forecast was made using three data mining methods. Decision trees (DT), naive Bayes (NB), and k-nearest neighbour (KNN) models were employed in the study (KNN). The data utilised in the forecast was gathered from the responses of incoming freshmen to a school toward a questionnaire distributed. In this work, the authors offered three feature selection algorithms such as Chi-square (CHI), mutual information (IG), and expectation maximization (GR), all of which were used to better predict and explain student learning patterns. As a consequence, the DT algorithm has the greatest accuracy rate of 90.67 percent [8].

Students in a mixed learning setting were observed by Dimic et al. for their patterns of behaviour. It was necessary to combine data from several sources in order to build a dataset usable by data mining techniques [9]. A total of 226 occurrences were retrieved. During the experiment, we concentrated on the data pre-treatment stages of data mining. Data augmentation, balanced ambiguity, relief, mutual information feature extraction, wrapper, and classifier subset evaluator approaches were all employed to narrow down the list of potential candidates for the most critical characteristics.

In order to calculate the correlations between characteristics, we used information measures (MI). A variety of feature subsets from various feature selection methods were utilised as prediction models for the following prediction models: empirical Bayes, combining one-dependence approximation, decision trees, and support

vector machines. When it came to identifying the ideal feature sets, the REF, wrapper technique, and MI were shown to be the most effective strategies. Blended learning environments can be improved by selecting lower-cardinality subsets of student learning activities, according to the study presented [10].

To improve the accuracy of prediction models for student academic performance, Zaffar and Savita studied the study of feature selection approaches. Six feature selection approaches were used in the study: causative link image classification (Cds), Chi-squared, Screened, discriminant analysis (IG), statistical method (PC), and Alleviation. a total of fifteen different classifiers were employed, including: Bayesian Network (BN), Nave Bayes, Principle components (NBU), Gradient Boosting (MLP), Simple Warehouse management (SL), Metaheuristic Optimization Enhancement (SMO), Rule Base, OneR, PART, JRip, Outcome Tree trunk (DS), J48, Stochastic Shrubland (RF), Random Forest (rf (RT), and the REP Tree (RepT).

Using multiple feature selection sets improved accuracy by 21% to 30%, according to the trials. Students' high school and university records were employed by Saa et al. as factors to predict their academic achievement in college. In the study, a private university provided the data utilised [11]. A total of 56,000 samples including the personal information of students were included in the collection. When it came to making forecasts, he used models including decision trees, artificial neural networks, random forests, naive Bayes, logistic regression, and the generalised linear model (GLM).

Using student information record systems, the researchers hoped to gauge pupils' progress and pinpoint areas of concern and opportunity for improvement. Thus, data augmentation (IG) was employed to identify the most important elements. As a consequence of the research, the RF technique was found to be the best accurate predictor of student performance in the forecasting model.

PROPOSED SYSTEM

The next phase in academic achievement is early warning systems for anticipating student performance. Approaches for predicting early education patterns and hazards, such as retention and drop-out, as well as student results can be described as predictive methods. Students' academic performance may be predicted more accurately with the use of an instructional early warning system. It is feasible to forecast pupils' performance and intervene early enough to prevent them from failing. Data on teaching and learning patterns and key impacting factors are needed before providing help to high school pupils. There is a dearth of statistics on secondary school education in the majority of underdeveloped nations. Even if there is, the vast majority of it is made up of students' names and other identifying information that is of little value for intervention. Because of this, the research team took great care while creating the survey questions for this study.

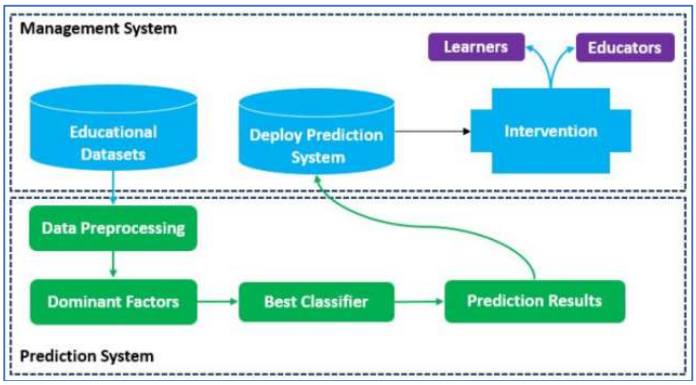

FIGURE 2. Proposed System Architecture

Teachers from a variety of educational institutions, research personnel, and top academics in the education sector were all involved in preparing questionnaires for data collecting. The study's goal is to improve the academic performance of high school pupils. Numerous high schools in India provided the data for this study. It is possible for survey questionnaires to be written and then administered by educators and other interested parties at any time utilising a Google document hosted in an online repository.

However, the timing of the survey is quite important to consider. Prior to the final test of the academic year, it is frequently a good idea to do an intervention. In the first semester of the school year, researchers distributed

questionnaires to all of the pupils. The reason for this is because students have already begun their studies, making this a good opportunity. Overall, they have managed to keep track of their learning preferences, the goals they set for themselves, and how those goals were achieved. Such that the treatment may be applied at the commencement of Semester II, particularly first before final national test is taken, collecting data during this time period is beneficial.

Data preparation is a tedious, yet necessary, step in the data processing process. Each action is designed to improve EDM's ability to forecast the future. Before making a forecast, it's a good idea to take some time to think about it. Feature extraction, data conversion, and data singular value decomposition are all necessary to increase the performance of a model provided in this paper. R Studio, an IDE for the R programming language, was used for data preparation and the experimentation in this study.

Many elements, including external circumstances, motivation, and longitudinal factors, can have an impact on a student's performance, whether they are an adult or a youngster. Students' academic achievement is mostly determined by factors other than intelligence or brilliance, such as self-discipline, enthusiasm, and support from family, teachers, and peers. The predictors are based on three key categories: home or domestic variables, student or individual characteristics, and school factors. In this work, new strategies for selecting useful traits were suggested. It is possible to increase the accuracy of estimation techniques and utilise it as a suggestion to learn students' behaviours for intervention by acquiring the informative characteristics or dominating set. This may be accomplished with the use of a data mining approach known as feature selection (FS).

RESULTS AND DISCUSSIONS

Each of the four suggested classifiers is compared against the FS method's feature set in this section. Experiments using the original dataset are shown in the table. Ibc, Prana, Mu, and Chiu classifiers are shown to perform well on subsets picked by each of the four aforementioned methods. Experiments are conducted on each FS algorithm's dominating set, and the results are compared. For the original dataset, the experiment outcomes of four classifiers with respect to ACC and RMSE are shown in the table. Composite Classifier and Hybrid RF, two tree-based models, produce the greatest ACC and lowest RMSE.

TABLE 1. Performance Analysis

Proposed Models	KNN	Hybrid C 5.0	Hybrid RF	IDBN
ACC (%)	95.95	99.25	99.72	83.14
RMSE	0.261	0.073	0.041	0.759

IG feature selection datasets were used to evaluate the relative performance classifiers. The Hydraulic C5.0 and Hybrid RF versions perform more effectively than the other types in terms of performance. With both chosen sets and dominant sets, hybrid RF yields the greatest ACC and smallest RMSE of any RF design. In terms of KNN performance, employing the dominating sets including the CHI algorithm's top five feature sets represents a substantial improvement. In this classification challenge, the best results were obtained using hybrid Suitable point and RF.

Both hybrid models' performance was enhanced by the dominating sets the ACC and RMSE results of the classification techniques based on the MI algorithm's subsets. All other models are outperformed by the Composite C5.0 and Dual RF. When comparing KNN, Hybrid C5.0, and Composite RF to the dominating sets, their performance is better. T the suggested classifiers perform well when given the input selected features from the MICHI method under consideration. Using the dominating sets improves performance greatly. The best categorization result was achieved using hybrid RF.

This project attempts to provide educational stakeholders with both an ideal prediction model and a dominating collection of data. Combining an optimum approach with a strong set yields reliable data. The ACC and RMSE of hybrid RF are the highest available. Results show that the suggested Hybrid RF using the MICHI algorithm's dominating set achieves the best classification accuracy of 99.98 percent and a root-mean-square error of 0.008. An additional benefit of our suggested feature selection approach MICHI is that it yields the best dominating collection of features in our tests. MICHI is a new feature selection approach based on ranked feature ratings that combines CHI and MI algorithms.

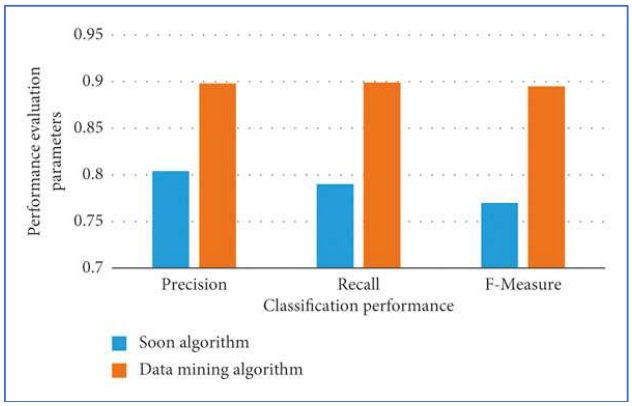

FIGURE 3. Performance Analysis

TABLE 2. Outcome of the Experiments

	Selected Set		Dominant Set		
Models	ACC	RMSE	N	ACC	RMSE
KNN	95.34	0.257	5	99.77	0.047
Hybrid C 5.0	99.85	0.040	29	99.85	0.040
Hybrid RF	99.87	0.038	29	99.89	0.035
IDBN	85.63	0.571	29	87.03	0.525

Using the MICHI algorithm, the most essential aspects in a student's performance are manually ranked. Not only does this set help with the prediction model's performance, but it also identifies the elements and student learning habits that might use some help and attention. An effective way to address an issue is to combine early forecasts with counselling and other forms of intervention. As a result, the MICHI algorithm and the Hybrid RF model are blended and made part of the APPS.

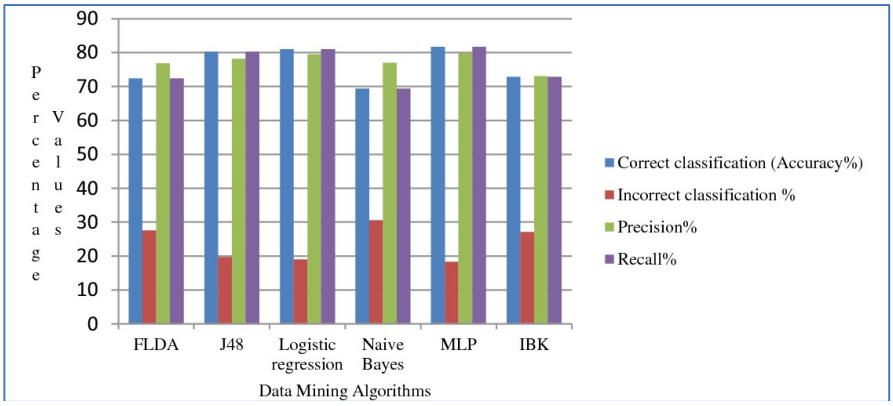

FIGURE 4. Accuracy analysis of Algorithms

UI (user interface) data are uploaded to the network in which the proposed methodology is used to determine the degree of performance of students. The user or client can then see the results on the UI. Educational stakeholders, such as teachers, administrators, or other staff members who have access to a student database, are the end users. Using the data, they've acquired, they can see what happens.

CONCLUSIONS

Based on an academic performance prediction system, this project aims to apply EDM approaches to provide an accurate indication for intervention and to improve student outcomes. Using the technology, users may access real-time data on students' development and student patterns, allowing them to better forecast students' progress levels and improve academic achievements. An excellent feature selection process and a predictive model are at the heart of the APPS, which identifies the most important determinants of student achievement. To acquire the best classification results, we conducted a comparison study of EDM prediction models with those that performed the best in the classification test. Four different FS techniques' feature sets were used in a cross-validation experiment to compare the performance of the classifiers. Classifiers work together to determine which criteria are most important. As a set of highly influential elements to be addressed for management and enhancement of academic achievement, the dominant set derived by the MICHI algorithm considerably enhances prediction models' performance. The experimental findings show that Hybrid RF outperforms the other classification methods in terms of classification accuracy. The web-based forecasting model integrates the created prediction model and the most important elements.

REFERENCES

[1]. G. Akçapınar, A. Altun and P. Aşkar, 2019, "Using learning analytics to develop early-warning system for at-risk students," *Int. J. of Educational Tech. in Higher Education*, **16(1)**, pp.1-20.

[2]. G. Dimić, D. Rančić, I. Milentijević, P. Spalević and K. Plećić, 2017, "Comparative study: feature selection methods in the blended learning environment," *Facta Universitatis. Series: Automatic Control and Robotics*, **16(2)**, pp. 95-116.

[3]. P. J. Estrera, P. E. Natan, B. G. Rivera and F. B. Colarte, 2017, "Student Performance Analysis for Academic Ranking Using Decision Tree Approach in University of Science and Technology of Southern Philippines Senior High School," *Int. J. of Eng. and Tech.*, **3(5)**, pp. 147-153.

[4]. Y. H. Hu, C. L. Lo and S. P. Shih, 2014, "Developing early warning systems to predict students' online learning performance," *Computers in Human Behavior*, **36,** pp. 469-478.

[5]. A. Al Mazidi and E. Abusham, 2018, "Study of general education diploma students' performance and prediction in Sultanate of Oman, based on data mining approaches," *Int. J. of Eng. Business Management*, **10** pp. 1-8.

[6]. A. Peña-Ayala, 2014, "Educational data mining: A survey and a data mining-based analysis of recent works," *Expert systems with applications*, **41(4)**, pp.1432-1462.

[7]. C. Romero and S. Ventura, 2020, "Educational data mining and learning analytics: An updated survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, **10(3)**, pp. 1-9.

[8]. C. Romero and S. Ventura, 2010, "Educational data mining: a review of the state of the art," *IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **40(6)**, pp.601-618.

[9]. A. A. Saa, M. Al-Emran and K. Shaalan, 2019, "Mining student information system records to predict students' academic performance," *In Int. Conf. on Advanced Machine Learning Technologies and Applications,* pp. 229-239.

[10]. P. Thakar and A. Mehta, 2015, "Performance analysis and prediction in educational data mining," *Int. J. of Computer Application*, **110(15)**, pp. 60-68.

[11]. M. Zaffar, M. A. Hashmani, K. S. Savita and S. S. Rizvi, 2018, "A study of feature selection algorithms for predicting students academic performance," *Int. J. of Advanced Computer Science and Applications*, **9(5)**, pp. 541-549.

[12]. S. Khan, X. Liu, K. A. Shakil, and M. Alam, 2017, "A survey on scholarly data: From big data perspective," *Information Processing & Management*, **53(4)**, pp. 923-944.

[13]. E. Fernandes, M. Holanda, M. Victorino, V. Borges, R. Carvalho, and G. Van Erven, 2019, "Educational data mining: Predictive analysis of academic performance of public school students in the capital of Brazil," *J. of Business Res.*, **94**, pp. 335-343.

[14]. R. Asif, A. Merceron, S. A. Ali, and N. G. Haider, 2017, "Analyzing undergraduate students' performance using educational data mining," *Computers & Education*, **113**, pp. 177-194.

[15]. H. Goker, H. I. Bulbul, and E. Irmak, 2013, "The estimation of students' academic success by data mining methods," *In: 12th Int. Conf. on Machine Learning and Applications*. IEEE, pp. 535-539.