www.rsya.org

Journal of Analytical Research

Vol 2 – Issue 2, July-Dec 2022

www.rsya.org/jar

Predicting the Performance of Students by Mining the Education Data

Dipti Chahar, Kaushilenrdra Kumar

ITM University, Gwalior, Madhya Pradesh 474001

Corresponding Author: dipti.chahar270595@yahoo.com

Abstract:
The technique of collecting valuable information from a large volume of data is called as data mining. It is focused with the strategies development for calculate the many forms of data that are generated in educational settings. Its objective is to obtain a better understanding of how students learned and the environments in which students learned in order to enhance educational outcomes and gain insights into and explanations of educational phenomena. When a teacher explains the topic, pupils comprehend it and learn it. Although there is no absolute measure for assessing examination score, knowledge is one scale that reflects students' performance. The use of data mining methods to increase the capability of student's performance in institutions is considered in this study. We report a real-world experiment done at PES University in Bangalore, India, in this study. This approach aids in identifying kids who require more advising or counseling from the teacher in order to receive a high-quality education.
Keywords: LMS, Data Mining, EDM, ID3,

1. Introduction

Humans face a difficult challenge in analyzing vast amounts of data to produce summarized usable knowledge. Data mining is the process of analyzing large amounts of data in order to extract essential or valuable information. Numbers, words, pictures, and facts can all be processed by computers [1]. This task analyses the data based on patterns, associations, and relationships in order to get information. Predicting student performance with high accuracy is useful since it improves in identifying students with low academic credentials at an early point in their academic careers. Student retention at universities is linked to academic achievement and the enrollment method. Because of the advancement of information and communication technology in higher education, the way students study and professors teach has changed dramatically [2]. Face-to-face learning is transformed into online learning when online and blended courses (partially or entirely) use the Internet to offer course information and instructions to learners. One approach to promote online learning is to use Learning Management

Systems (LMSs), which provide online learning resources such as course content, quizzes, assignments, and forums. Supervisors that utilise LMSs may easily manage and supply learning resources, as well as monitor their students' progress, because nearly all of the teachers' and students' actions are documented in such systems. Teachers may optimize teaching and learning by gaining information into students' online activity [3]. It is worth noting, however, that the data recorded by LMSs is primarily raw and does not give reliable information or measures of current theoretical ideas. Furthermore, because many students who use LMSs fail to adjust to the demands of such environments, LMSs provide pedagogical problems for teachers (in addition to their benefits). As a result, it's critical to have a deeper knowledge of the process, as well as if and how this data may be used to improve the learning process [4]. EDM (Educational Data Mining) is a new area that focuses on creating ways for mining educational data to better understand student behavior and perhaps detect individuals with learning disabilities early on . The use of EDM methods to educational data allows teachers to make informed decisions that will improve learning and ultimately lead to increased academic performance [5]. Academic success is influenced by a variety of elements, including personal, social, psychological, and environmental influences. The use of Data Mining is a highly promising method for achieving this goal [6]. In this research focuses and investigates the education field of data mining based on performance of students, psychological, and other contextual criteria. The purpose of this project is to extract information from student databases in order to help students perform better. In this case, data mining techniques like ID3, C4.5, and Bagging are employed.

RELATED WORKS

While the utilization of data mining in superior learning is young topic of research, there is lot of effort in education sector. This is due to its potential for educational institutions. Romero and Ventura [7] conducted 1995-2005 survey on mining of learning data. They found that education data mining is a viable study topic and does not have unique needs in other fields. Work should thus be focused on the educational data mining field. El-Halees [8] conducted case study using learning data mining to assess education conduct of pupils. The objective of His work aims to illustrate how beneficial data mining in higher education may be utilized to progress the performance of students. He used student data from the data base course and collected all accessible data from the e-learning system, include individual data and university documents of students, course records and data. He then used data mining techniques to find various forms of information, including organization rules, and resolution tree categorization rules. He also divided student into EMclustering groups and found all aberration in the data via analysis. Finally, he explained how we may take use of the information to increase the student's performance. Al-Radaideh et al [9] utilized data mining techniques to enhance academic performance of students by evaluating student data to determine key variables that may have an influence on student performance. The systems perspective found are predicated on decision tree approach and the strategic approach extracted are examined and graded. It permits students to predict the final level in subject, evaluation and work.. These associated details will be available to the teacher even before final test is conducted. This study allows educators minimize the failure rates by taking suitable actions and improving student achievement at the correct moment.

Kurniawan & Halim [10] Data analysis and the concept storage server can enable the low-level students to determine the appropriateness of their course or module and adapt the measures to improve their academic school performance. Jacob et al. [11] Research will analyze if a modification in one of the variables results in the other. Decision tree algorithm provide possible results and are utilized in this study to predict student success. In the creation of something like a model with a predictor variables and various independent factors a regression analysis is utilized; when the models suitable, the correlation values is measured through using data from the predictor factors. Al-shargabi &Nusari [12] Specialists have studied and thoroughly assessed the findings and rated them in terms of the various characteristics such as reliability, realism, usefulness and creativity. In furthermore, a quantitative analysis of the correctness of the manufactured model using SQL queries was performed (rules). Moscoso-Zea et al. [13] The analysis demonstrates that random trees are more precise but have limited because of the difficulties in interpretation while the J48 method has superior interpretative-ness of outcomes in the visualization of data categorization and just slightly lower efficiency. Ahuja et al. [14] Educational data mining is one of those developing technologies to evaluate the data acquired from research and education and then apply machine learning and data mining approaches to anticipate student behavior. This study is intended to use the research study quoted to choose the best suitable method based on the demand for EDM clustering or classification. The regular search is aimed at better understanding pupils and learning how to make them more productive.

METHODS OF DATA MINING

Data is a method of creating a set of functions or models that identify and separate data concepts or classes in order to use the model to predict a class of facts whose class label is unknown. The model is derived by analyzing a collection of training data. Data prediction and visualization are 2 types of data analysis that may be useful for derive models that describe to estimate future data trends or key data classes. The categorization of data is a two-step procedure. In the first step is to create a model that describes a predefined collection of ideas or data classes. The model is built by examining information tuples that are characterized by attributes. Every tuple is believed to belong to a predetermined class, as indicated by the class label property, which is one of the attributes. The training data set is made up of the data tuples that were used to create the model. Each tuples that make up the training set are called training trials, and they are chosen at random from the sample population. Classification rules, decision trees, and mathematical formulas are used to express the learnt model. The model is then used to classify the data in the second phase. First, the model's prediction accuracy is calculated. ID3, c4.5, and bagging are the basic approaches utilized in this research.

Iterative Dichotomiser3 (ID3)

Quinlan Ross [15] proposed this decision tree method in 1986. Hunt's algorithm is at the heart of it. There are two stages to the tree's construction. Tree construction and pruning are the two processes. The splitting attribute is chosen by ID3 using the information gain measure. In order to create a tree model, it only accepts categorical characteristics. When there is noise, it does not

produce accurate results. A pre-processing approach must be employed to eliminate the noise. To construct a prediction model, the information gain for each attribute is computed, and the highest information gain attribute is chosen as the root node. The attribute should be labelled as a root node, with arcs representing the property's potential values. After then, all possible result instances are evaluated to see if they belong to the same class. If all of the instances belong to the same class, the node is represented by a single class name; otherwise, the instances are classified using the splitting attribute. The ID3 method may manage continuous attributes by separating or directly evaluating the values to determine the optimal split point by applying a threshold to the attribute values. Pruning is not supported by ID3.

C4.5

This method is the successor of Quinlan Ross's ID3 [16]. It's based on Hunt's algorithm as well. C4.5 creates a decision tree using both categorical and continuous characteristics. To manage continuous attributes, C4.5 divides the attribute values into two divisions based on the specified threshold, with all values over the threshold being assigned to one child and the rest to another. It also deals with attribute values that are missing. C4.5 builds a decision tree using Gain Ratio as an attribute selection metric. When there are numerous result values for an attribute, it reduces the bias of knowledge gain. Determine the gain ratio of each characteristic first. The characteristic with the highest gain ratio will be the root node. To enhance classification accuracy, C4.5 utilises optimistic pruning to remove superfluous branches from the decision tree.

Bagging

Bagging is a method used in prediction data mining to aggregate projected classifications from different models or the similar kind of model with varying learning data. Assume that the goal of data mining is to construct a predicting classification model. We may continually sub-sample the dataset and apply a tree classifier to the subsequent samples. In fact, extremely distinct trees are frequently generated for various samples, demonstrating the instability of models that is typically seen with standard datasets. One technique for obtaining a single prediction is by using every tree discovered in the various samples and does some basic voting.

DATA MINING PROCESS

Data for this study were collected from several degree colleges and institutes connected with PES University in Bangalore, India. To forecast the student's performance, these variables are evaluated using decision trees. In order to use this approach, the following procedures must be followed in order.

Preparations of Data

The data set utilized in this study was gathered from several institutions using the sampling technique for the BBA course of the 2019-20 academic years. The data is initially 200 bytes in

size. After merging process errors were eliminated, data held in various tables was merged in a single table in this phase.

Selection and Transformation of Data

Only the fields necessary for data mining were chosen in this stage. A small number of derived variables were chosen. While some of the variables' information was taken from the database. Table 1 contains a list of all the predictor and response variables generated from the database.

Table 1: Student Related Information

Variable	Representation	Values
Sex	Students sex	{Male, Female}
Cat	Students category	{GEN, OBC, SC, ST}
10th	Students grade in 10th	{O- 90% - 100%, A- 80% - 90%, B – 70% - 80%, C – 60% - 70%, D- 50%-60%, E- 40% - 50%, F < 40%}
12th	Students grade in 12th	{O- 90% - 100%, A- 80% - 90%, B – 70% - 80%, C – 60% - 70%, D- 50%-60%, E- 40% - 50%, F < 40%}
AT	Admission type	{Direct, Test}
CL	College location	{Village, City, District, State}
Hostel	Students live in hostel or not	{Yes, Not}
FI	Father income	{High, Medium, poor, BPL}
FQ	Occupation of father	{Service, Business, Agriculture, NA}
MI	Mother income	{High, Medium, poor, BPL}
MQ	Occupation of mother	{House Wife, Service, Business, Agriculture, NA}
Results	Results of BBA	{First ≥ 60% Second ≥ 45 & < 45%, Fail< 36%}

Implementation of Mining Model

WEKA is a widely used machine learning and data mining toolkit that was created at the University of Waikato in New Zealand. It includes a vast library of cutting-edge machine learning. WEKA includes data pre-processing, regression, classification, clustering, association rules, visualization, and visualization tools. WEKA has grown in popularity among academic and industry researchers, and it is also frequently utilized in classrooms.

To utilize WEKA, the gathered data must be processed and converted to file format so that it can be read by the WEKA data mining tools.

Results and Discussion

To develop the classification model, three different classification approaches were applied to the dataset. The bagging algorithm, the ADT, and ID3 decision tree algorithm are the approaches used.

Table2. Feature of the Students Categorization

Variable	**Representation**	**Values**					
Sex	Students sex	{Male-129, Female- 71}					
Cat	Students category	{GEN-75, OBC-63, SC-40, ST-22}					
10th	Students grade in 10th	O- 81, A- 13	B – 34	C – 16	D- 21,	E- 28,	F- 7
12th	Students grade in 12th	O- 53, A- 42	B – 48	C – 23	D- 12,	E- 20,	F- 2
AT	Admission type	{Direct- 58, Test-142}					
CL	College location	{Village-81, City-96, District-23, }					
Hostel	Students live in hostel or not	{YES-77, Not- 123}					
FI	Father income	{High- 113, Medium-40, poor-33, BPL-14}					
FQ	Occupation of father	{Service-101, Business-59, Agriculture-35, NA-5}					
MQ	Occupation of mother	{House Wife-114, Service-64, Business-22}					
Results	Results of BBA	First - 83	Second- 71	Third-26 Fail-20			

Visually analyze the data and figure out the distribution of values after applying the pre-processing and preparation procedures. The distribution of student values is depicted in table 2. Author conducted some tests to assess the effectiveness and utility of various categorization algorithms for estimating student performance. Table 3 displays the outcomes of the trials.

Table3. Working of the classifiers

Estimate Criteria	Classification		
	ID3	C4.5	Bagging
Timing to develop model (Sec)	0.03	0.03	0.06
Correctly classified instances	155	140	148
Incorrectly classified instances	45	60	52
Accuracy (%)	77.5%	70%	74%

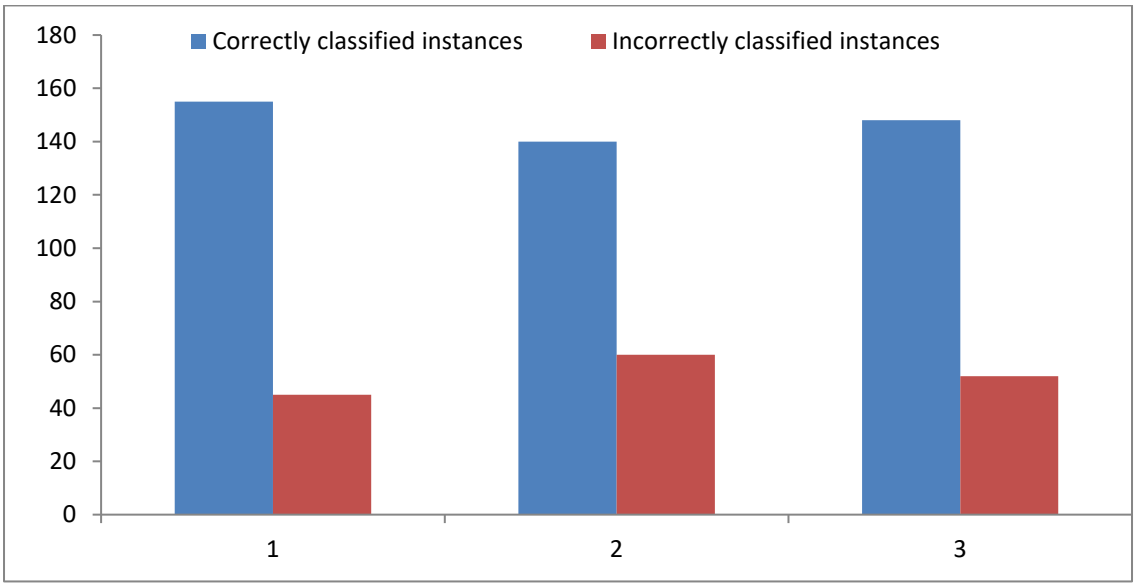

Figure1. Efficiency of various models

The percentage of properly categorized cases is commonly referred to as a model's accuracy. As a result, the ID3 classifier is more accurate than other classifiers. The mean absolute error, Kappa statistic and root mean squared error will only be available in numerical form [17]. For reference and evaluation, we also present the root relative squared and relative absolute error in percentage. The simulation results are presented in Table 4.

Table4. Training and Simulation Error

Estimate Criteria	Classification		
	Bagging	ID3	C4.5
Kappa Statistic	0.5971	0.6885	0.5324
MAE	0.1915	0.1188	0.1903
RMSE	0.3197	0.3211	0.352
RRE	59.4971%	32.7113 %	53.3166%
RAS	75.2754%	78.8587 %	84.8813%

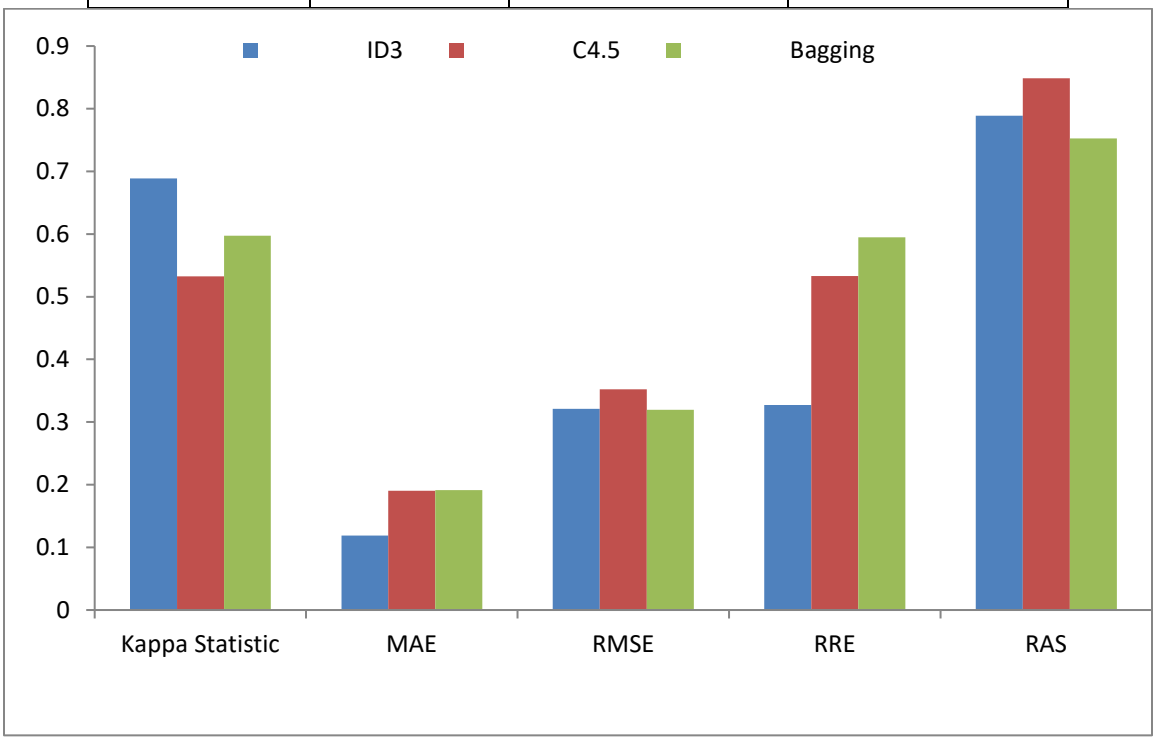

Figure2. Comparison b/w variables

Conclusion

The most commonly used classifiers are investigated, and tests are carried out to determine the optimum classifier for predicting student achievement.

As a result, we have achieved our goal of evaluating student performance using the three specified classification algorithms based on Weka. Based on the performance statistics, the best method is ID3 Classification, which has an accuracy of 77.5 percent and a total time to create the model of 0.03 seconds. In comparison to the others, the ID3 classifier has the lowest average error of 0.16. These findings show that, of the machine learning algorithms evaluated, the ID3 classifier has the potential to considerably enhance the performance of traditional classification approaches. The best decision tree classifier for predicting student success in the BBA test is researched and experiments are done to identify the best classifier. The true positive rate of the model for the FAIL class is 0.84 for ID3 and C4.5 decision trees, indicating that the model is correctly identifying students who are likely to fail, according to the classifiers accuracy. These kids may be eligible for appropriate counseling in order to improve their grades. From past year student data, machine learning techniques such as the C4.5 decision tree algorithm may build successful prediction models. The empirical results demonstrate that by applying predictive models to the data of entering new students, we can create a brief but accurate prediction list for the student. This research will also try to figure out which students require more help.

References

1. Bertino, E., Fovino, I. N., & Provenza, L. P. (2005). A framework for evaluating privacy preserving data mining algorithms. *Data Mining and Knowledge Discovery*, *11*(2), 121-154.
2. Luo, Y., Yao, C., Mo, Y., Xie, B., Yang, G., & Gui, H. (2021). A creative approach to understanding the hidden information within the business data using Deep Learning. Information Processing & Management, 58(5), 102615.
3. Fayyad, U., Shapiro, G. P., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. American Association for Artificial Intelligence, 37-54.
4. Alaa el-Halees, "Mining students data to analyze e-Learning behavior: A Case Study", 2009.
5. Navinchandran, M., Sharp, M. E., Brundage, M. P., & Sexton, T. B. (2021). Discovering critical KPI factors from natural language in maintenance work orders. Journal of Intelligent Manufacturing, 1-19.
6. Romero, C., Ventura, S., Espejo, P. G., & Hervás, C. (2008, June). Data mining algorithms to classify students. In *Educational data mining 2008*..
7. Romero, C., Ventura, S. (2007). Educational Data Mining: A Survey from 1995 to 2005. Expert Systems with Applications 33, 2007, pp.135-146.
8. El-Halees, A. (2008) 'Mining Students Data to Analyze Learning Behavior: A Case Study',The 2008 international Arab Conference of Information Technology (ACIT2008) Conference Proceedings, University of Sfax, Tunisia, Dec 15- 18.

9. Al-Radaideh, Q., Al-Shawakfa, E. and Al-Najjar, M. (2006) 'Mining Student Data Using Decision Trees', The 2006 International Arab Conference on Information Technology (ACIT'2006) – Conference Proceedings.
10. Kurniawan, Y., & Halim, E. (2013, August). Use data warehouse and data mining to predict student academic performance in schools: A case study (perspective application and benefits). In Proceedings of 2013 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE) (pp. 98-103). IEEE.
11. Jacob, J., Jha, K., Kotak, P., & Puthran, S. (2015, October). Educational data mining techniques and their applications. In 2015 International Conference on Green Computing and Internet of Things (ICGCIoT) (pp. 1344-1348). IEEE.
12. Al-shargabi, A. A., & Nusari, A. N. (2010, February). Discovering vital patterns from UST students data by applying data mining techniques. In 2010 The 2nd International Conference on Computer and Automation Engineering (ICCAE) (Vol. 2, pp. 547-551).IEEE.
13. Moscoso-Zea, O., Saa, P., & Luján-Mora, S. (2019). Evaluation of algorithms to predict graduation rate in higher education institutions by applying educational data mining Australasian Journal of Engineering Education, 24(1), 4-13.
14. Ahuja, R., Jha, A., Maurya, R., & Srivastava, R. (2019). Analysis of educational data mining. In Harmony Search and Nature Inspired Optimization Algorithms (pp. 897-907). Springer, Singapore.
15. J. R. Quinlan, "Introduction of decision tree", Journal of Machine learning", : pp. 81-106, 1986.
16. Quinlan, J.R. (1993), C4.5: Programs for machine learning, Morgan Kaufmann, San Francisco.