International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-3, September 2019

Enhancing the Quality Education using Predictive and Descriptive Data Mining Model

S. Jayakumar, R. Parameswari, A.Akila

Abstract: ***Data mining is the trending field used to get relevant knowledge from the database given. This technique consists of subfield called educational data mining is the emerging area used to extract the hidden patterns from the huge data with the help of tools techniques developed by the researchers of the educational data mining. The purpose of extracting patterns from the educational database is to improve the quality of education can be provided to the students for their better feature. The patterns are extracted by using the existing data mining techniques to enhance student performance. Educational data mining techniques such as classification, regression, clustering are available in the field. Classification is defined as the technique used to categorize the data based on the given label and constraints. In this paper, the algorithms like naves Bayes, Random Forest and J48 algorithms used to classify the data instances under the given labels using the constraints given., the classification algorithms like naves Bayes shows best performance accuracy with the given student dataset. Clustering and apriori rule have a strong relationship in student performance. In this paper, predictive data mining used to predict the student's performance to enhance the study level of the students in the organization.***

Keywords: Educational Data mining, Predictive Model, Descriptive Model, Cluster.

I. INTRODUCTION

Educational data mining is the subfield in the data mining used in the educational field to extract the hidden patterns in the educational database. The mining techniques like decision tree, Navies Bayes, K-nearest and K- means clustering techniques etc, are used in the real world. By using the above techniques knowledge extracted from the huge datasets to form the association rules, clustering, classification and prediction.

The WEKA tool used to accept the data taken from the online database and the quality of the given data is enhanced using preprocessing techniques. The preprocessed data is given for the feature selection task to extract the features of the data given. Then the techniques like Classification, Regression, Clustering, visualization techniques used to process the data and to yield the result accordingly.

Manuscript published on 30 September 2019
* Correspondence Author
Mr. S. Jayakumar, Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai, India. Email: jaimca69@gmail.com
Dr. R. Parameswari*, Department of Computer Science , Vels Institute of Science, Technology and Advanced Studies, Chennai, India. Email: dr.r.parameswari16@gmail.com
Dr.A.Akila, Department of Computer Science, Vels Institute of Science, Technology & Advanced Studies, Chennai, India. Email: dr.a.akilaganesh@gmail.com

The data collected by the researchers, was transferred to excel sheet and the required attributes were selected and converted into ARFF file format. The transformations were applied on the data collected with the cleaning and integration of data. The cleaned and integrated data was then classified by using J48, Random Tree, Naive Bayesian classifier, clustering and Apriori rule mining algorithms.

The classified data was visualized in different formats such as graphs and decision trees. The results of the above five algorithms were compared in order to reach the decisions and give final recommendations on different parameters of the study.

The organization consists of a number of students access the test and quiz given by the organization to evaluate the performance of the students. The most popular factor is the score obtained from the students are considered, but the score factor depends on the other factors like semester marks, sessional marks and regular attendance of the particular student. These factors are very difficult to track because these factors strongly depend on the student's marks in the above exams and attendance. The statistics of the student's performance also leads to the wrong conclusion. Soft computing techniques used to overcome the difficulty faced in calculating GPA and CGPA. The main aim of the paper is to predict the performance of the students' using different techniques by adding additional factors for decision making.

II. SYSTEM ARCHITECTURE

The system architecture (fig 1) shows the strong connection between gender and performance. According to the system architecture the attributes like age, gender, GPA, type of institutions, Educational level are collected to predict student learning activities. The prediction of student learning activities is implemented based on the architecture shown in figure-1 and all the students are aware of quality education across the country.

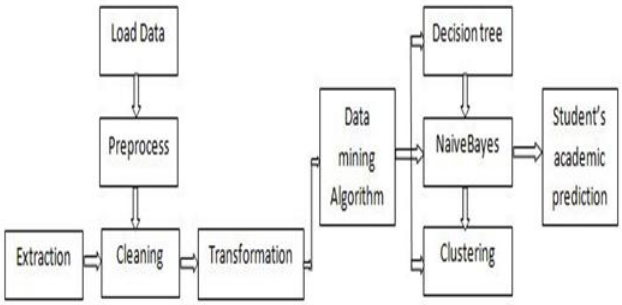

Fig 1: System Architecture

III. PREDICTIVE DATA MINING MODEL IN HIGHER EDUCATION

Predictive model is used to apply neural networks, logistics regression and Classification and Regression Tree (CART) techniques to predict admission process [7]. The probability of correct classification techniques are CART, logistics regression and neural network obtained 74%, 64%, 75% respectively [3]. The author suggested decision trees, neural network and logistics regression to predict the process of application.

In the tested data, the logistics regression obtained 66% of the admitted candidate. However 49% of the enrolled candidates and 78% of non-enrolled candidates [1]. The author discussed based on the student performance the technical course via web can be analyzed and compared using data mining classification. The outcome of the given set shows the significant of accuracy improvement [5]. In student dropout [6] predicts whether the student graduates in six years or not. The actual data set 56 for both academic and attributes. Both methods achieved only 63% accuracy over 5000 data set.

The author used decision tree (CHAID) to analyze the high school dropouts [18]. The classification of dropout and non-dropouts student overall rate was 15.79% and 10.36% [18]. The author analyzed the different factors with effect of academic performance of average mark, retention and desertion [17]. The data set considered of 25000 students consists of 16 attributes of different aspects such as age, gender, faculty, test mark etc. The clustering techniques used to predicting good or poor academic achievement of students.

The author used to survey data to retention problem using data mining. To analyze the retained students are approximately 82% and 88% for non-retained student the outcome of testing data. The actual retention rate of this data set was 82.61% [3].

Machine learning techniques were applied to prevent student dropout ratio in higher education. The different classification techniques are used and compared with different attribute datasets [14]. The classification techniques are naïve bayes, feed forward neural network, support vector machine, decision tree and logistics regression. The overall accuracy results 80% achieved and compared all test cases using naïve bayes classifier techniques. Bayesian networks to predict data mining model for testing the student skills [15]. Based on the skill sets the author creates assessment, online system and the test the knowledge skill of the student.

IV. DESCRIPTIVE DATA MINING MODEL IN HIGHER EDUCATION

The author focused on the use of data mining in educational institution. The result of entrance examination and their performance of studied are classified with the help of cluster analysis and K-means algorithm technique [9]. In academic performance, the author studied higher education data to predict the performance of academic using data mining techniques [8]. The author suggested association rule mining approach for selecting the precise student aim at remedial classes using Gifted Education Programme[9]. The overall result of both naïve bayes and decision trees 47% of score [10].

V. DATAMINING ALGORITHMS IN WEKA TOOL FOR EDM

WEKA is one of the popular data mining tool used to implement data mining techniques. The data mining techniques like clustering and classification techniques by setting the WEKA environment. According to the previous study, the researchers have implemented classification algorithms like J48, Bayes net, PART, Random Forest. The above-mentioned algorithms are implemented by using supervised learning techniques with the input data as test and train data [4].The four classifiers are

A. J48 Classifier

This classifier is the technique categorized under supervised learning algorithm developed by Ross Quantian [4]. This is an important algorithm used to create decision trees by using the C4.5 algorithm. The performance accuracy of this classifier is shown in fig 6. This classifier shows better accuracy then Navies Bayes and also consumes less computational time with a large database when compared to Navies Bayes by using conditional direct graph [4].

B. Random Forest Classifier

As said earlier the supervised learning techniques contains two phases training phase and testing phase. In the training phase, this classifier uses a bootstrap algorithm to construct complete classification trees samples. In the testing phase, the random samples of the bootstrap algorithm are considered as the output [7].

C. Naïve Bayes

The classifier algorithm is the technique mainly used for statistical analysis of the data given. This technique uses Bayes theorem for classification to construct the graph structure with the dataset given. This technique takes the samples of the particular class to find the maximum probabilities to predict the class membership. The Bayes theorem uses two important techniques for prediction. The main thing is Naves Bayes assumes the effect of the attribute is independent to each other [8].

D. Association Rule results

Association rule is to extract the frequent patterns occur in the datasets. The frequent patterns are of if-else format. The rules consist of two parts such as if part followed by the consequent part where the if part consists of association rule which monitors the path of execution. The else part is executed if the given condition is false. Association rules work in the above format to mine the frequent patterns [2].

VI. PERFORMANCE ANALYSIS

In experiment analysis student performance dataset was used. This datasets consists of various parameters for analyzing results. WEKA results with student performance dataset

This section mainly focuses on the analysis of the prediction accuracy of the classification techniques Random forest, Naive Bayes, J48. The purpose of this analysis is to prove the classification scenarios are suitable for the considered prediction model.

Table-I: Result of Weka

Algorithms	Naive Bayes	Decision Tree (J48)	Random Forest
Correctly classified	76.6425	76.3275	75.91375
Incorrectly classified	23.3575	23.80375	24.08625
Kappa Statistic	28	46.625	42.375
Mean absolute error	26.125	20.25	22.125
Root Mean Squared error	34	31	32.125
Relative absolute error	78.8425	59.255	66.07375
Root relative squared	87.07625	78.5925	82.71375

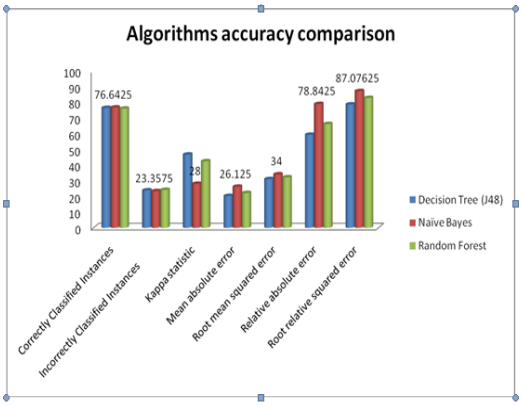

Fig. 2. Accuracy of Algorithms Results

Fig. 2 shows that Decision Tree (J48), Naïve Bayes and Random Forest are most often used by the many researchers to predict the student's performance in their education than other algorithms. Some researchers used combination of these algorithms to forecast the performance of the student in educational data mining. From the above figure Naïve Bayes gives high accuracy than J48 and Random Forest algorithms.

Table-II: Gender comparison result

Gender	
Male	**Female**
187	208

Table-II shows the gender comparison of 395 students among them 187 (47.34%) male and female 208 (52.66%) were responded. When we compare between them female has higher academic performance than male students

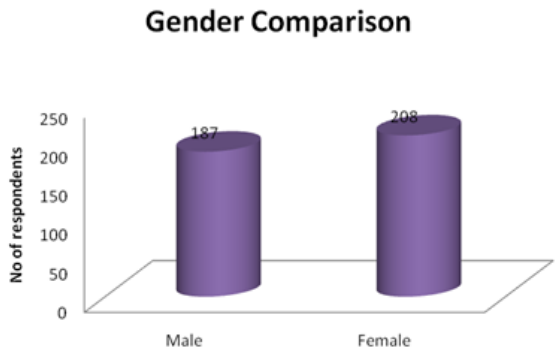

Fig. 3. Gender comparison result

Fig. 3 shows the gender comparison result.

Table-III: College type comparison result

College Type		
Arts & Science	**Diploma**	**Engineering**
238	48	109

Table III shows that in Arts & Science 238 students, from Diploma 48 students and from Engineering 109 students have responded to the survey but Engineering students academic performance is better than Arts & Science and Diploma students. The following graph shows the number students who joined in the particular colleges.

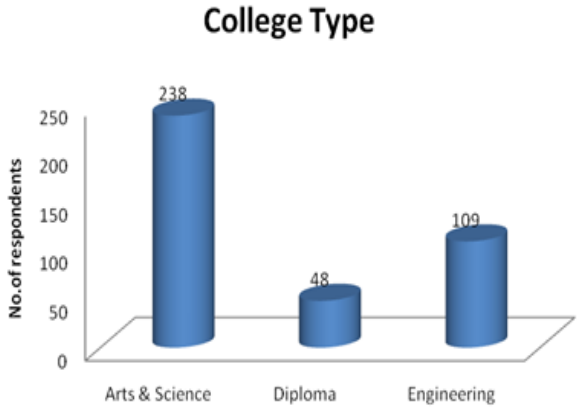

Fig. 4: College type comparison result

Fig. 4 shows the college type comparison result of the students.

Table-IV: Region result

Region	
Rural	**Urban**
214	181

Table IV shows that 214 students have enrolled from rural area and 181 students from urban area.

Table-V: Family income result

Income		
15000-25000	**25000-50000**	**Above 50000**
76	80	239

Table V shows the family income of the students. From the table value 76 student's income is very low, 80 students income is average and 239 students is above 50000 (high).

Table-IV: Students study hours

Study hours (week)			
Above 30 hrs	20 to 30 hrs	10 to 20 hrs	2 to 10 hrs
134	169	80	12

Table VI shows 12 students study 2 to 10 hours, 80 students study 10 to 20 hours, 169 students 20 to 30 hrs and 134 students study above 30 hours in a week. So if student spend more time for their education they can perform well in the academic.

Table-VII: Students grade result

Grade							
A	A+	A++	B	B+	C+	D	D+
71	85	104	10	10	1	108	6

Table VII shows that grade of the students 71 students have got A grade, 85 students got A+ grade, 104 students got A++ grade, 10 students got B grade, 10 students got B+ grade, only one student got C+ grade, 108 students D grade and only 6 students got D+ grade.

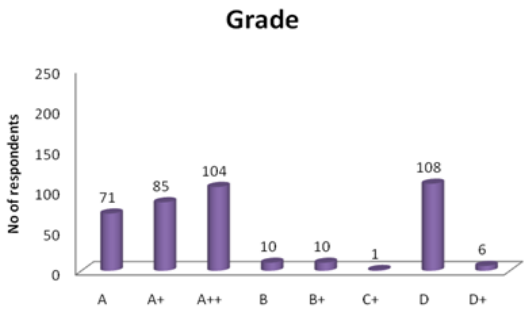

Fig. 5. Students grade result

Fig. 5 shows the grade received by the students in their academic performance. Only one students has got less than 50 marks rest of them have scored average and high grade marks. From the figure it is clear that the majority of students pay attention in their academic performance.

Table-VIII: Students ranking result

Ranking	
First lass	260
First class with distinction	114
Second class	17
Third class	1

Table VIII shows that 374 students have performed well than 18 in their last semester

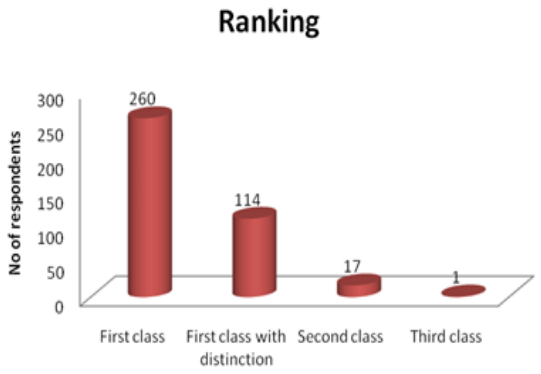

Fig. 6. Students ranking result

Fig.6 shows that ranking list of the students.

The various attributes are examined with the improvement of student's academic performance using data mining algorithms. Statistical analysis, clustering and association rule mining were used to determine the impact of education and from this analysis clearly pointed out that all the algorithms plays an important role of younger generation life impacted positively and strong relationship with their learning activities.

VII. CONCLUSION AND FUTURE WORK

Current experimental results studies have focused EDM techniques on student gender comparison, study hours in a week, last semester GPA as well as problem areas in studies with the goal to improve learning quality. The use of the data collected in prior studies have been able to aid in demonstrating the importance of time studying a subject compared to results, as well as finding problem areas that need more focus in the classroom. The use of EDM to remove noisy data, and find stronger patterns shall discover a correlation between specific environment attributes and the affect it has on a student's learning accomplishments. The data received from the proposed research study should aid in struggling programs to increase average GPA, promote student learning, and in return increase student population. Prediction of student academic performance helps both the teacher and parents to predict about their success and failure in examinations. This paper is to improve the classification accuracy and different factors related to the student performance. At last this paper is observed that Naïve Bayes performs better than J48 and Random Forest classifier algorithms for predicting students' academic performance by taking various measures to get exact accuracy of the result. K-means and Apriori rule mining algorithm show that all the parameters used for the research have strong relationship with the students. In future this algorithm and model can be applied for predicting student performance or new algorithm; new student variable and new data mining tools can be also identified for better prediction based on this study.

REFERENCES

1. Antons, E.N. Maltz (2006), "Expanding the role of institutional research at small private universities: A case study in enrollment management using data mining", New Directions for Institutional Research, 2006(131):69.
2. Arora R K, Badal D (2014), "Placement Prediction through Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), Volume 4, Issue 7, pg 447-51
3. Atwell, W. Ding, M. Ehasz, S. Johnson, M. Wang (2006), "Using data mining techniques to predict student development and retention". In Proceedings of the National Symposium on Student Retention.
4. Bai, S.M. (2006), and S.M. Chen. "Automatically Constructing Grade Membership Functions for Evaluating Students' Evaluation for Fuzzy Grading." In Proceeding of the 6th International Symposium on Soft Computing for Industry. Hungary: IEEE, 2006, 1-6.
5. Baker R.S.J.d, Inventado P.S, "Educational Data Mining and Learning Analytics", In J.A. Larusson, B White (Eds.) Learning Analytics: From Research to Practice. Berlin, Germany: Springer.
6. Barker,T.,Trafalis,,T. R. Rhoads (2004)."Learning from student data" Systems and Information Engineering Design Symposium, pp. 79–86,135.
7. Chang (2006), "Applying data mining to predict college admissions yield: A case study", New Directions for Institutional Research, (131).
8. Dorina Kabakchieva (2013), "Predicting Student Performance by Using Data Mining Methods for Classification", Cybernatics and Information Technologies, Vol 13.
9. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.
10. Erdogan and Timor (2005) A data mining application in a student database. Journal of Aeronautic and space technologies,Volume 2 No:2 (53-57)
11. Erdoğan, M. Timor (2005), "A data mining application in a student database". Journal of aeronautics and space technologies, volume 2, number 2 ,pp.53-57.
12. Hijazi, and R. S. M. M. Naqvi (2006), "Factors affecting student"s performance: A Case of Private Colleges", Bangladesh e-Journal of Sociology, Vol. 3, No. 1.
13. Kalles D., Pierrakeas C.(2004), "Analyzing student performance in distance learning with genetic algorithms and decision trees", Hellenic Open University, Patras, Greece.
14. Kotsiantis (2007), "Supervised Machine Learning: A Review of Classification Techniques", Informatica, Vol. 31, No. 3, pp. 249-268.
15. Pardos, N.T. Heffernan,B. Anderson,C.L.Heffernan (2006), "Using fine grained skill mod1els to fit student performance with Bayesian networks", In 8th International Conference on Intelligent Tutoring Systems (ITS 2006), pp. 5–12.
16. Ma Y, Agnihotri L, Baker R and Mojard S (2016), Ëffect of student ability and question difficulty on duration", Proceedings of 9th International conference on Educational Data Mining ,pg 86-93.
17. Salazar,J.Gosalbez,I.Bosch, R.Miralles,L.Vergara (2004), "A case study of knowledge discovery on academic achievement, student desertion and student retention". Information Technology: Research and Education, ITRE 2004. 2nd International Conference, pp.150–154.

AUTHORS PROFILE

Mr.S.Jayakuamar is a Research Scholar in Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai. His research interest lies in the area of Data Mining.

Dr.R.Parameswari is working as Associate Professor in Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai. She has 13 years of teaching experience. She has completed PhD in Computer Science from St.Peter's University, Chennai. She is presently guiding 8 Ph.D scholars and 1 Mphil Scholar. She has produced 3 M.Phil Scholars. She has published 20 papers in various International journals including journals indexed in Scopus. She has presented many papers in International Conferences and attended many seminars and workshops conducted by various educational Institutions. She is acting as editor and reviewer in many International Journals. Her research interest lies in the area of Cloud Computing, Big data Analytics, Internet of Things.

Dr.AAkila is working as Assistant Professor in Department of Computer Science, Vels Institute of Science, Technology and Advanced Studies, Chennai. She has 13 years of She has completed PhD in Computer Science from Bharathiar University, Coimbatore. She is presently guiding 8 PhD scholars and 1 Mphil Scholar . She has produced 4 M.Phil Scholars. She has received many prestigious awards. She has published more than 40 papers in various International journals including journals indexed in Scopus. She has presented many papers in International Conferences and attended many seminars and workshops conducted by various educational Institutions. She is acting as editor and reviewer in many International Journals. Her research interest lies in the area of Speech Recognition Systems, Data Structures and Neural Networks.