Using Data Mining to Predict Students Performance

Madhumitha S[1], Vaishali Roy[2], Vinitha B[3],Vijayakumar S[4]

[1,2,3]Department of Computer Science and Engineering, R.M.K. Engineering College, Tamil Nadu, India
[4]Associate Professor, Department of Computer Science and Engineering, R.M.K. Engineering College, Tamil Nadu, India

Abstract - *Educational data mining is used to extract student data from dataset and transform it into a meaningful and useful structure for further use. In this project, we are using Logistic Regression algorithm to predict student's performance. The existing system uses decision tree, Naive Bayes algorithms to classify the student data. As the decision tree algorithm makes use of numerical values to analyse the data, so it is inefficient in prediction. On the other hand, though the Naive Bayes approach uses textual attributes there exists dependencies among the variables. Hence these algorithms are comparatively inaccurate. The proposed system uses Logistic Regression which is a statistical method for analysing a dataset in which there are one or more independent variables that determines an outcome. Individual attributes are also analysed that influences students' final performance.*

Key Words: Educational Data Mining, dataset, Predict student performance, Logistic Regression.

1. INTRODUCTION

Analyzing the huge amount of data to form useful information is a tedious task for human. Data Mining is the area which analyses large volume of data to extract necessary or useful information. Computers can process data like texts, images, and numbers. This performs analysis based on patterns, association, relations in these data to get useful information. The prediction in students' performance is beneficial as it helps in identifying students with low academic performance at the early stage of academics. In universities, high students' performance results in job opportunities for students and helps college with increased admission rate for the next academic year. The steps to assist the low academic students with better education are:
(a) Generation of data source which consists of predictive variables. (b) Identification of various factors which affects the performance of student's learning.
(c) Construction of a prediction model with the help of data mining technique on the basis of predictive variables which were identified.
(d) Validation of the model. Data Mining can assist in the field of education to extend our understanding of learning process by identifying variables and evaluating them. Mining in the field of education is known as Educational Data Mining. Students' attendance in college, hours spent on a daily basis after college, grades, study
hours, family income of the students are significantly related to student performance. By means of Logistic Regression model, it has been found that the factors like attendance and study hours are highly correlated with the performance of the student. This helps us to, increase students' academic achievements. Students' performance prediction helps to identify the most effective factors which work with student's test score and then tune these factors to make better student test performance.

2. RELATED WORK

J K Jothi and K Venkatalakshmi conducted the students' performance analysis on the students' data collected from the Villupuram college of Engineering and Technology.
The data included five year period and applied clustering methods on the data to overcome the problem of low score of students, and to increase students' academic performance [1].

Sheik and Gadage have done the analysis related to the student learning behavior by using different data mining models, namely classification, clustering, decision tree, pattern mining and text mining [2].

Mythili M S and Shanavas A R applied classification algorithms to analyze and evaluate school students' performance using weka. They came with various classification algorithms, namely J48, Random Forest, Multilayer perception, IBI and decision tree with data collected from the student management system [3].

Suyal and Mohod applied the association and classification rule to identify the students' performance. They mainly focused to find students who need special attention to reduce failure rate [4].

Noah, Barida and Egerton conducted a study to evaluate students' performance by grouping the grading into various classes using CGPA. They used different methods like Neural network, Regression and K-means to identify the weak performers for improving performance [5].

Backer and Yacef conducted a study for identifying the best model for EDM. They analyzed data and reached the conclusion that most of the papers includes prediction than relationship mining [6].

Angeline D M conducted a study on the students' performance by using Apriori algorithm that extracts set of rules specific to each category and analyze the knowledge to classify the student based on their involvement in assignment, internal assessment test, group action etc. It helps to identify the students' range like average, below average, and good performance [7].

Remesh, Parkavi, and Yasodha conducted a study on the placement chance prediction by applying different techniques such as Naive Bayes Simple, MultiLayerPerception, SMO, J48, and REPTree by its accuracy. From the result they concluded that MultiLayerPerception is more suitable than other algorithms [8].

Bhise, Thorat and Supekar suggested a method using K-means clustering algorithm by describing it step by step. This research focused on reducing drop-out-ratio of the students and improve it by considering the factors like midterm and final exam assignment. They considered clustering techniques namely hierarchical, partitions, and categorical [9].

ElGamal A F presented a study for predicting students performance in a programming related course. Here the data is collected from the department of computer science from Mansoura University and applied extract rules for predicting students' performance in programming related course [10].

3. METHODOLOGY

Data mining is the knowledge discovery process from huge volume of data. The mechanism works using data set where the student performance in the end semester is evaluated.

3.1 Data preparation

Student related data were collected from a large data set. In this step, data is checked for null values, errors and corrected if any.

3.2 Data selection and transformation

In this step only fields which were necessary for the data mining process were selected. The student's 10th, 12th, degree marks in each semester, assignment, study hours, parent's education, income were taken as attribute values for prediction.

3.3 Logistic Regression Algorithm

Logistic regression is a classification method where the dependent variable can have only two possible values. When compared with linear regression, Logistic Regression is preferred because data plotting shows that it does not suit linear regression pattern.
The logistic regression equation will be

$$\Pi(x) = 1/1+ -y \quad (1)$$

Where $y = \beta0 + \beta1X1 + \beta2X2 + \beta3X3 + \beta4X4$ and e being Euler's number

$\beta0$ is the constant

$\beta1$ is coefficient for variable X1

$\beta2$ coefficient for X2

β3 coefficient for variable

β4 coefficient for variable

Sl. No	DESCRIPTION	POSSIBLE VALUES
1	Students' Gender	Male, Female
2	Medium of Teaching	English, Tamil
3	Living Location	Village, City
4	Kind of stay	Home , Hostel , PG
5	Student's marks in 10th	<200,200-299,300-400,>400
6	Student's marks in 12th	<600,600-799,800-1000,>1000
7	Father's Occupation	Farmer, Business, Service, Retired, Not-Applicable
8	Mother's Occupation	Housewife, Business, Service, Retired, Not-Applicable
9	Student Interested In Higher Education	Yes, No
10	Do you use Mobile? If Yes, Since how many Months/Years	Yes, No
11	Do you use the Internet? If Yes, Since how many Months/Years	Yes, No
12	Do you use the social network? If Yes, Since how many Months/Years	Yes, No
13	Reading Habit	Early morning , Night
14	How many hours do you spend on studies per day	1, 2, >3
15	Family status	Nuclear, Joint
16	Family annual income status	Poor, Medium, High

Table -1: Student Related Variables

4. EXPERIMENTAL RESULTS

After testing with selected attributes using Logistic Regression, the accuracy increased to 82.03%, which is the best among all the models we tested so far. By observation from the bar chart it is evident that around 230 students will perform poor, 600 students will fairly perform and 200 students will perform well in their semester examinations. This helps the university to predict the students' performance in a better manner and give appropriate training to improve the academic results.

Table -2: Model Performance

Model	Accuracy
Naive Bayes	80.76%
K-Nearest Neighbor (K=4)	76.96%
Logistic Regression	82.03%
Multi-Layer Perceptron	78.73%
J4.8	81.77%

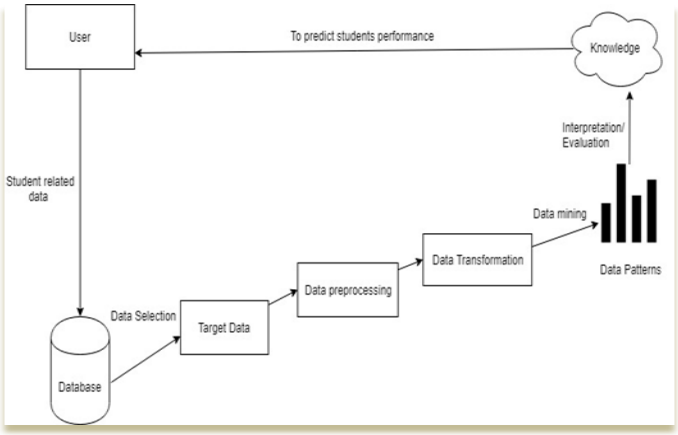

Fig -1: Architecture Diagram

Table- 3: Students Dataset

school	sex	age	address	famsize	Pstatus	Medu	Fedu	Mjob	Fjob	reason	guardian	traveltime	studytime	failures	schoolsup	famsup	paid	activities	nursery	higher
GP	F	18	U	GT3	A	4	4	at_home	teacher	course	mother	2	2	0	yes	no	no	no	yes	yes
GP	F	17	U	GT3	T	1	1	at_home	other	course	father	1	2	0	no	yes	no	no	no	yes
GP	F	15	U	LE3	T	1	1	at_home	other	other	mother	1	2	3	yes	no	yes	no	yes	yes
GP	F	15	U	GT3	T	4	2	health	services	home	mother	1	3	0	no	yes	yes	yes	yes	yes
GP	F	16	U	GT3	T	3	3	other	other	home	father	1	2	0	no	yes	yes	no	yes	yes
GP	M	16	U	LE3	T	4	3	services	other	reputatio	mother	1	2	0	no	yes	yes	yes	yes	yes
GP	M	16	U	LE3	T	2	2	other	other	home	mother	1	2	0	no	no	no	no	yes	yes
GP	F	17	U	GT3	A	4	4	other	teacher	home	mother	2	2	0	yes	yes	no	no	yes	yes
GP	M	15	U	LE3	A	3	2	services	other	home	mother	1	2	0	no	yes	yes	no	yes	yes
GP	M	15	U	GT3	T	3	4	other	other	home	mother	1	2	0	no	yes	yes	yes	yes	yes
GP	F	15	U	GT3	T	4	4	teacher	health	reputatio	mother	1	2	0	no	yes	yes	no	yes	yes
GP	F	15	U	GT3	T	2	1	services	other	reputatio	father	3	3	0	no	yes	no	yes	yes	yes
GP	M	15	U	LE3	T	4	4	health	services	course	father	1	1	0	no	yes	yes	yes	yes	yes
GP	M	15	U	GT3	T	4	3	teacher	other	course	mother	2	2	0	no	yes	yes	no	yes	yes
GP	M	15	U	GT3	A	2	2	other	other	home	other	1	3	0	no	yes	no	no	yes	yes
GP	F	16	U	GT3	T	4	4	health	other	home	mother	1	1	0	no	yes	no	no	yes	yes
GP	F	16	U	GT3	T	4	4	services	services	reputatio	mother	1	3	0	no	yes	yes	yes	yes	yes
GP	F	16	U	GT3	T	3	3	other	other	reputatio	mother	3	2	0	yes	yes	no	yes	yes	yes
GP	M	17	U	GT3	T	3	2	services	services	course	mother	1	1	3	no	yes	no	yes	yes	yes
GP	M	16	U	LE3	T	4	3	health	other	home	father	1	1	0	no	no	yes	yes	yes	yes
GP	M	15	U	GT3	T	4	3	teacher	other	reputatio	mother	1	2	0	no	no	no	no	yes	yes
GP	M	15	U	GT3	T	4	4	health	health	other	father	1	1	0	no	yes	yes	no	yes	yes
GP	M	16	U	LE3	T	4	2	teacher	other	course	mother	1	2	0	no	no	no	yes	yes	yes
GP	M	16	U	LE3	T	2	2	other	other	reputatio	mother	2	2	0	no	yes	no	yes	yes	yes

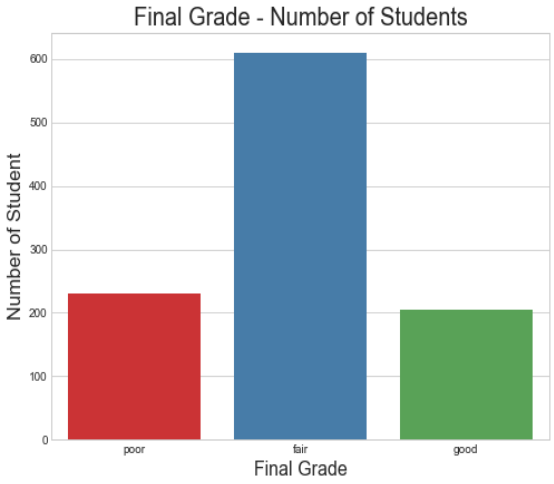

Chart -1: Final Grade Distribution

5. CONCLUSION

In this paper, the Logistic Regression is employed for students' data to predict the students' performance. This study can facilitate the students and the professors to motivate the students of all categories to perform well. This study helps to identify students who require special attention, minimize the failure rate and take relevant action to succeed in the upcoming semester examination.

Future work includes employing big data to help decision makers in education sector to make better and more intelligent decision in education big data environment.

REFERENCES

[1] J.K.Jothi and K.Venkatalakshmi, "Intellectual performance analysis of students by using data mining techniques", International Journal of Innovative Research in Science, Engineering and Technology, vol 3, Special iss 3, March 2014.

[2] Nikitaben Shelke and Shriniwas Gadage, "A survey of data mining approaches in performance analysis and evaluation", International Journal of Advanced Research in Computer Science and Software Engineering , vol 5, iss 4, 2015K.

[3] M.S. Mythili and A.R.Mohamed Shanavas , "An analysis of students' Performance using classification algorithms ", IOSR-JCE, Volume 16, iss1, Jan. 2014.

[4] A.Dinesh Kumar and V.Radhika, "A survey on predicting student performance", International Journal of Computer Science and Information Technologies, Vol. 5, 2014.

[5] E. Osmanbegović and M. Suljić, '"Data mining approach for predicting students performance", Economic Review, vol 10, iss 1, 2012.

[6] Sayali Rajesh Suyal and Mohini Mukund Mohod, "Quality improvisation of student performance using data mining techniques",
International Journal of Scientific and Research Publications, vol 4,iss 4, April 2014.

[7] OTOBO Firstman Noah, BAAH Barida and Taylor Onate Egerton, "Evaluation of student performance using data mining over a given data space", International Journal of Recent Technology and Engineering
(IJRTE) ISSN: 2277-3878, Volume-2, iss 4, September 2013.

[8] Brijesh Kumar Baradwaj and Saurabh Pal, "Mining educational data to analyze Students' performance", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[9] T.Jeevalatha, N. Ananthi and D.Saravana Kumar, "Performance analysis of undergraduate students placement selection using Decision Tree Algorithms", International Journal of Computer Applications (0975- 8887), vol 108, December 2012.

[10] Ryan J.D.B.Baker and Kalina Yacef , "The state of educational data mining in 2009: A review and future revisions", Journal of Educational
Data Mining , Vol.1,No.1, February 2009.

[11] A.F . ElGamal, "An educational data mining model for predicting student performance in programming course", International Journal of
Computer Applications(0975-8887), Vol.70, No.17, May 2013.

[12] D.Magdalene Delighta Angeline, "Association rule generation for student performance analysis using Apriori Algorithm", The SIJ Transactions on Computer Science Engineering And Applications(CSEA), vol.1, March-April 2013.

[13] Bhise R.B, Thorat S.S and Supekar A.K, "Importance of data mining in higher education system", IOSR Journal of Humanities and Social
Science, ISSN: 2279-0837, vol.6, iss 6, January-February 2013.

[14] V.Ramesh, P.Parkavi and P.Yasodha, "Performance analysis of data mining techniques for placement chance prediction", International Journal of Scientific and Egineering Research , Vol.2, iss 8, August 2011.

[15] Mohammed M.Abu Tair and Alaa M.El-Halees, 'Mining Educational Data to Improve Students' Performance: A case Study', International journal of information and Communication Technology Research,
ISSN: 2223-4985, vol.2 no.2, February 2012
.