An Ingenious Methodology for the Collation of Existing Algorithms for the Prognosis of Student Performance

Arya B V, Amritha P B, C V Prasanna Kumar

Abstract: In this proposed research work we use a profound Data mining technique which is an automated procedure of discovering interesting patterns by means of comprehensible predictive models from large data sets by grouping them. Predicting a student's academic performance is very crucial especially for universities. Educational Data Mining (EDM) is an approach for extricating useful data that could possibly affect a firm. Nowadays student's performance is swayed by a lot of aspects. These aspects might involve the academic performance of a student. This subject evaluates numerous factors probably suspected to alter a student's empirical performance in scholastic, and discover a subjective design which classifies and forecast the student's learning outcomes. The intention of this research is to conduct a case study on factors swayed by the student's academic achievements and to dictate greater impact factors. In this paper we focus on the academic achievement evaluation on the basis of correct instances and incorrect instances by means of Naive Bayes and Random Forest algorithms. This paper intends to make a metaphorical assessment of Naive Bayes and random Forest classifier on student data and dictate the best algorithm.

Keywords : Naïve Bayes, Random Forest, Educational Data Mining.

I. INTRODUCTION

In this term of proliferation and machinery rebellion, education is the basic amenities for every human being. It plays an imperative role in enhancing human investment and is correlated with attaining selfhood..It makes sure the retrieval of knowledge facilitates entities to enhance their productivity and promote their materiality of life.The pivotal aspect for instructors is primarily to guide students adequately which eventually helps them to contribute their best towards academics. To attain this, it is fundamental for every educator to perceive the various aspects that direct the academic progress of students. Educational Data Mining is a one among the various areas in the field of data mining and also to discover fruitful specimens and to extract convenient knowledge from the educational information scheme. This helps students to guide their academics finer and enhance their empirical performance. Data mining is appropriate for any kind of information archive. The Naive Bayes is an extensively accepted classification algorithm that is based on the principle of independence between attributes of data points. The Random Forest algorithm is a combo learning scheme that is set off by formulating a stacks of decision trees.

II. BACKGROUND STUDY

In this section, literature related to student's performance prediction is reviewed.

.Dinesh Kumar, K.Sathesh Kumar, R.Pandi Selvam conducted a study on different prediction techniques and prediction tools[9]. For the purpose they reviewed 20 papers that have been published in the year from 2011 to 2017. The algorithms implemented are Bayesian Classification, ID3, CART, C4.5, OneR, C4.5, Multilayer Perceptron, Nearest Neighbor algorithm, Random Forest Tree, Linear Regression. Their work was only different from to improve the classification accuracy and different factors related to the student performance. They also found that most of the research has been applied in weka tool for prediction. Matlab tool is also been used by some researchers. Rapid miner tool is rarely used to analyze the performance. Among the diverse prediction algorithms most of the researchers are often used J48 in prediction. Some researchers employ a combination of varied algorithms in order to prognosticate the academic achievements of a student. They concluded that most of the research has been done on the same prediction algorithms, student variables, and data mining tools.

Amjad Abu Saa conducted comparative research to determine alliance between student's confidential, communal factors and their academic performance in the preceding semester by using data mining technique[5]. A survey was conducted among students of various universities and collected data on personal, communal and academic related to them. They have been implemented Four decision tree algorithms and noticed that the student academic performance is fully dependent on their own academic endeavor. The student performance is diagonalized by the Grade Point Average and is primarily for anticipating academic performance at the end of the semester. They noticed that among the four algorithm CART gives more accuracy which is 40% and the least algorithm accuracy was 33% for ID3. Naive Bayes algorithm is also tested and there is a noticeable potential in the result with an accuracy of 36% which is inevitable

Revised Manuscript Received on March 30, 2020.
* Correspondence Author

Arya B V*, PG Student Assistant Professor Department of Computer Science and IT, Amrita School of Arts And Sciences,Kochi Amrita Vishwa Vidyapeetham, India. bvarya97@gmail.com

Amritha P B , PG Student Assistant Professor Department of Computer Science and IT, Amrita School of Arts And Sciences,Kochi Amrita Vishwa Vidyapeetham, India. pbamrithaamru@gmail.com

C V Prasanna Kumar, PG Student Assistant Professor Department of Computer Science and IT, Amrita School of Arts And Sciences,Kochi Amrita Vishwa Vidyapeetham, India. prasannakumarcv@asas.kh.amrita.edu

Efrem YohannesObsie and Seid Ahmed Adem conducted data mining research to figure out which perspectives perform finer than others. They present a prediction of student's intellectual performance from an informational dataset explicitly using their academic scores, disregarding their financial, communal and psychological aspects. They implemented Linear Regression, Neural Network and Support Vector Regression model to estimate academic performance which is deliberated only by viewing student final year CGPA[8]. The performance of the various procedures was measured using the Root Means Square Error and Coefficient of Correlation. The data for the study is taken from a group of students. The relatable factors of a student were collected through various levels of interaction. The discrete factors that are chosen are Gender, Mobile no, Section, Nationality, Entry year, University Entrance Examination Result, Course scores, Semester GPA, Final CGPA8. They noticed that the LR method showed a slight improvement in diminish the Root Mean Square Error among the expected and intended values. This enables us to estimate the final year CGPA with a correlation coefficient of 0.0047.

Aranga Arivarasan and Dr. M Karthikeyan conducted a Performance exploration using Random forest, J48 and Naive Bayes algorithms. This work deals with provisional perusal of various classifiers in the background of database to boost correct positive rate and reduce incorrect positive rate using WEKA tool. The dataset used for this experiment is studentsheet.csv dataset. It is a UCI repository databases which embrace of 7 trait and 1728 specimens. Classification accuracy and confusion matrix was intended for all the three algorithms. Experimental results show that the Random Forests classification algorithms yields better results by achieving accuracy of 94.05%. F-measure were also intended to check the accuracy of all three classification algorithms. Cast analysis is the same for all the algorithms.

III. METHODOLOGY

Through a pervasive analysis of the publications and debate with professionals on the topic academic performance of a student, a numerous factor was diagnosed that have a significant impact on them.

Naive Bayes Classification

Naive Bayes classification is a simple and probability based classification technique, which suspects that all hardened attributes in a dataset are autonomous from each other. The Naive Bayes classifier suspects that the survival of an attribute in a class does not have any relation with other components. This represents knowledge involving a random variable set. In this classifier model, each node in the graph serves as random variable and the edge between variables represent conditional dependency. The basic assumption made by the classifier is that each attribute is independent and equal. By using this approach posterior probability of event B given A can be calculated with the following formula.

$$P({}^{B}/_{A}) = \frac{(P(B).P({}^{A}/_{B}))}{P(A)}$$

Whereas, A stands for reliant feature vector and B is represented as class variable. $X = (x_1, x_2, x_3, \ldots, x_n)$ In the above formula possible values of A can be $a_1 , a_2 , a_3 ,\ldots a_n$

Random Forest Classification

Random Forest Classification is a classification and regression method which is flexible and easy to implement. A Forest is termed as a collection of decision trees. First they select random data from our sample dataset and create an appropriate decision tree. Secondly, do prediction for each tree and opt the best among the trees based on the majority vote system. In this classifier the tree grows as given below:

a) From the sample training set N number of cases are selected.

b) The input variable symbolizes I, where at each node i is a constant variable that is chosen randomly and split the node accordingly.

c) Without pruning each tree extends as much as possible.

d) The Forest error rate is being depend on two things:

Correlation between any two trees and the strength of each individual tree. Increasing the correlation will enhance the forest error rate and strength of tree will depreciate the forest error rate.

For this analysis, present-day real-world data was accumulated from a group of UG students. An illustrative of 700+ students was cumulated for that. The domain assess for some of the attributes were listed below:

Attribute	Description	Domain
Age	Student's age	{17,18,19,20,21}
Gender	Student's gender	{Male, Female}
FamSize	Family Size	{3,Equal to 4,More than 4}
F Job	Father's Job	{Business, Engineer, Doctor, Professor, Work in abroad, Others}
M Job	Mother's Job	{Homemaker, Teacher, Doctor, Nurse, Others}
Par Status	Marital Status of Parents	{Married, Divorced, Single-Parent}
D/H	Day scholar/Hosteller	Day scholar/Hosteller
MT	Mode of Transport	{Pedestrian, Bicycle/Two wheeler, Car, Public Transport, College bus}
TT	Travel Time	{Less than 15Minutes,15-30minutes,30minutes-1hr,More than 1 hr}

10 th	10 th percentage	{100-90,90-80,80-70,70-60,60-50}
12 th	12 th percentage	{100-90,90-80,80-70,70-60,60-50}
CGPA	Student's CGPA	{10-9,9-8,8-7,7-6,6-5,below 5}
Arrears	Have any arrears	{Yes, No}
Choose college	Why did you choose this college	{Placement, Institute Reputation, Teaching, Nearby home, References }
Time Spend in studies	How much time do you spend every day in studies	{Less than 2 hrs, 2-3 hrs, More than 4 hrs}
Friends	Number of friends	{Less than 3,3-6,More than 6}
ECA	Extracurricular activity	{Yes, No}
EPC	Extra paid class	{Yes, No}
HE	Higher Education	{Yes, No}
Time in SM	Time spend in social media	{1-2,3-6,over 6 hrs }
Whatsapp	Reason for using whatsapp	{Chatting, Academic work, General information, Other}
APP	Most used APPs	{Gaming, Entertainment, Learning, Others}
FH	Food habits	{Veg, Non-veg, Both}
CA	Change in academic	{Teaching method, Event conducted, Syllabus, Other}

IV. RESULT

We have implemented two classification algorithm Naive Bayes algorithm and Random forest algorithm on our dataset studentsheet.csv in weka tool. Weka tool enables embedded algorithms for preprocessing, classification, clustering etc. Fig1 shows the line graph of the attribute CGPA and Fig2 shows the marital status of parents.

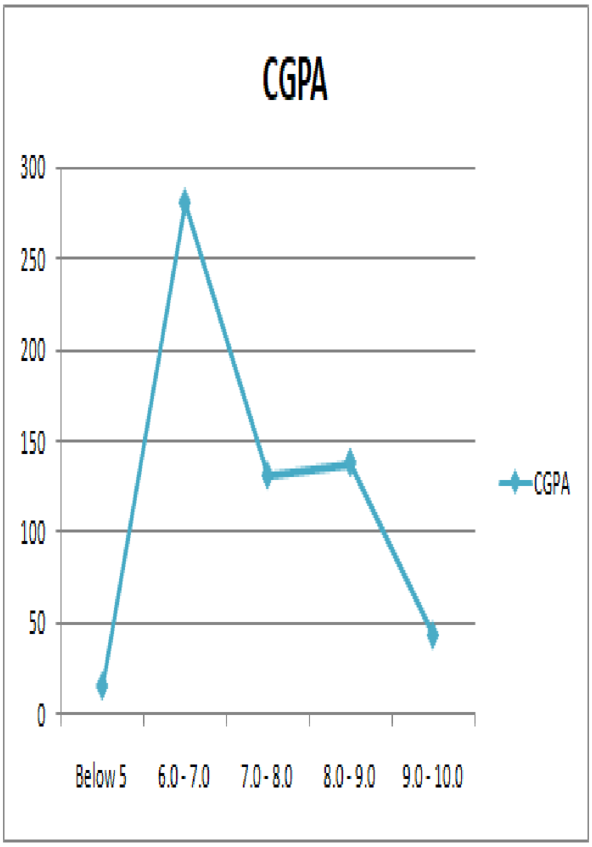

Figure 1.Line Graph of CGPA

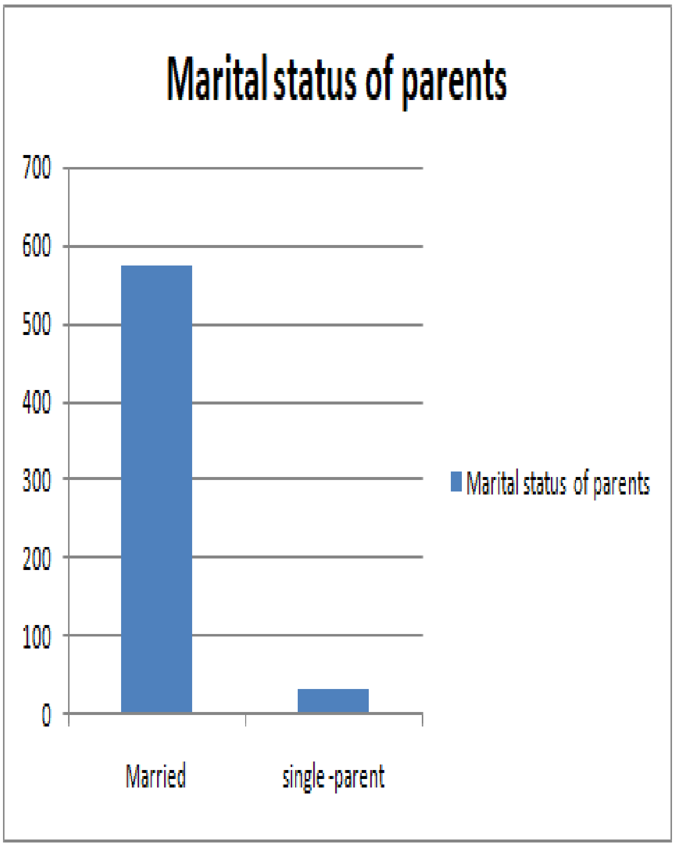

Figure 2. Bar Graph of Marital Status of Parent's

4.1 Accuracy measure of Algorithm

Algorithm	TP Rate	FP Rate	Precision	Recall	F Measure	Class	Accuracy
Naive Bayes Classification	1.000	0.002	0.923	1.000	0.960	Below 5	76.252 %
	0.876	0.144	0.833	0.876	0.854	6.0 - 7.0	
	0.667	0.057	0.765	0.667	0.713	7.0- 8.0	
	0.744	0.072	0.744	0.744	0.744	8.0 - 9.0	
	0.953	0.011	0.872	0.953	0.911	9.0 - 10.0	
	0.958	0.971	0.942	0.958	0.950	Married	90.468 %
	0.029	0.042	0.040	0.029	0.033	Single - Parent	
Random Forest	0.667	0.000	1.000	0.667	0.800	Below 5	76.413 %
	1.000	0.195	0.808	1.000	0.894	6.0 - 7.0	
	0.750	0.0000	1.000	0.750	0.857	7.0 - 8.0	
	0.842	0.002	0.991	0.842	0.911	8.0 - 9.0	
	0.953	0.000	1.000	0.953	0.976	9.0 - 10.0	
	1.000	1.000	0.942	1.000	0.970	Married	93.699 %
	0.000	0.000	0.000	0.000	0.000	Single - Parent	

V. CONCLUSION

In this study, two data mining algorithms were implemented in order to generate a subjective model which anticipate systematically and constructively from a gathered training dataset. In order to predict student's performance two attributes are chosen which has a greater impact on the performance, parent's status and CGPA of student. In the current study, it was noticed that the Random Forest algorithm has more accuracy than Naive Bayes classifier. This study can sway and help many universities or organization to carry out data mining techniques on their dataset periodically to figure out lively outcomes and patterns. This will eventually help both university and students to boost the academic achievement drastically.

REFERENCES

1. Rohit Arora and Suman, "Comparative Analysis of Classification Algorithms on Different Datasets using WEKA". International Journal of Computer Applications, vol. 54, No.13, September 2012
2. Tina R, Patil and Mrs. S. S. Sherekar, "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification". International Journal Of Computer Science And Applications, vol. 6, No.2, April 2013
3. Mrs. M.S. Mythili1, Dr. A.R.Mohamed Shanavas, "An Analysis of students' performance using classification algorithms " . IOSR Journal of Computer Engineering, vol. 16, pp 63-69,2014
4. Raheela Asif, Agathe Merceron and Mahmood K. Pathan,"Predicting Student Academic Performance at Degree Level: A Case Study". I.J. Intelligent Systems and Applications , 2015, 01, 49-61
5. Amjad Abu Saa,"Educational Data Mining & Students' Performance Prediction". International Journal of Advanced Computer Science and Applications, vol. 7, No. 5, 2016
6. Aranga Arivarasan, Dr. M karthikeyan. "Classification based Performance analysis using Naïve-Bayes J48 and Random Forest algorithms". International Journal of Applied Research,3(6), 174-178,2017
7. Mukesh Kumar,Prof. A.J. Singh."Evaluation of Data Mining Techniques for Predicting Student's Performance". I.J. Modern Education and Computer Science, 8, 25-31,2017
8. Efrem Yohannes Obsie,Seid Ahmed Adem,"Prediction of Student Academic Performance using Neural Network, Linear Regression and Support Vector Regression: A Case Study ". International Journal of Computer Applications, vol. 180, No.40, 2018
9. A.Dinesh Kumar , R.Pandi Selvam , K.Sathesh Kumar."Review on Prediction Algorithms in Educational Data Mining ". International Journal of Pure and Applied Mathematics,vol.118,No.8,2018
10. Jasvinder Kumar ,"A Comprehensive Study of Educational Data Mining" . International Journal of Electrical Electronics & Computer Science Engineering, 2015, ISSN : 2348-2273
11. Hafez Mousa1 and Ashraf Maghari,"School Students' Performance Prediction Using Data Mining Classification ".International Journal of Advanced Research in Computer and Communication Engineering ,vol. 6, August 2017
12. Ajay Kumar Mishra and Bikram Kesari Ratha," Study of Random Tree and Random Forest Data Mining Algorithms for Microarray Data Analysis ," International Journal on Advanced Electrical and Computer Engineering, vol. 3,2016
13. Erum Shahzadi and Zahoor Ahmad," A Study on academic performance of university students ".Proc.8th International Conference on Recent Advances in statistics Lahore, February 8 – 9

AUTHORS PROFILE

Arya B V MCA Scholar at Amrita Vishwa Vidyapeetham University, Amrita School of Arts and Sciences, Kochi, Kerala , India bvarya97@gmail.com

Amritha P B MCA Scholar at Amrita Vishwa Vidyapeetham University, Amrita School of Arts and Sciences, Kochi, Kerala , India pbamrithaamru@gmail.com

C V Prasanna Kumar Master Degree in Computer Applications, Post Graduate Diploma in Computer Applications, almost 18 Years of experiences both in abroad and India. Currently working as Asst.Professor in Amrita Vishwa Vidyapeetham. prasannakumarcv@asas.kh.amrita.edu