International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

Student Dropout Prediction & Educational Data Mining

Mahesh Mardolkar, N Kumaran

***Abstract*: Educational data like students performance is very important to study and analyze and to improve the quality of education. The study concerned to data mining techniques with educational data is known as Educational Data Mining (EDM). This study finds knowledge and interesting patterns in educational organization. Students performance are the subject mainly concerned to find the qualitative model based on student's personal and social factors then classify and predict the student performance. Proper counseling to underperforming students can reduce dropout ratio and help them to continue their studies.**

Keywords: Data Mining, Education, Patterns, Performance, Student.

I. INTRODUCTION

The recent trend in Data mining and Knowledge discovery in database (KDD) is Educational Data Mining (EDM). Discovery of knowledge and useful patterns is the main focus in data mining. In Educational Data Mining researchers discover knowledge to help students to manage their education to perform better or help the educational organizations to manage their students. To classify a student, data is analyzed to create association rule or decision tree which supports in decision making and enhance student's performance. Educational data mainly generates specific classification and prediction rules helping the students to enhance their performance.

The most effective and familiar data mining technique is classification, which is used to predict and classify values. Through a questionnaire educational data is collected Educational Data Mining is implemented to classify and predict student's performance. To identify the relationship between personal and social factors and also check student's performance is the objective of this paper. Better enhanced quality of education and instructions can be provided by this newly discovered knowledge which will further aid in identifying underperformers at the beginning of the year and more attention can be given to avoid dropouts. Also students performing better can benefit from this study and they can enhance their performance further. To classify the students there are many different types of classification methods in data mining and every method has its advantages and disadvantages.

In this paper different classification methods are used to verify the results based on accuracy and precision.

II. RELATED WORK

A research on group of 50 students was conducted by Bharadwaj and Pal. The students were enrolled for a course of four years period (2007-2010). "Class Test Grades", "Previous Semester Marks", "Seminar Performance", "Assignments", "Attendance", "Semester Marks", "Lab Work", "General Proficiency" were considered as performance indicators. ID3 decision tree algorithm was used to construct decision tree. As a data mining technique Bharadwaj and Pal selected ID3 decision tree to analyze student performance, since it is very simple decision tree learning algorithm [1].

Similar research on generating classification rules and prediction was conducted by Ahmed and Elaraby, records of the student's behavior and activities was processed and analyzed for a course program across 6 years (2005-2010). Their study was able to help students to improve their performance, also identify the students who require special attention to avoid dropouts [2].

A data mining research using Naïve Bayes classification to predict, classify and analyze students as underperformer or performer was conducted by Pandey and Pal. One of the simple probability classification techniques is Naïve Bayes classification, the name "Naïve" is considered since all the attributes are independent of each other. Their research was able to classify and predict student's grade in upcoming years based on performance in previous year [3].

A significant data mining research was conducted by Bharadwaj and Pal, using Naïve Bayes classification, a questionnaire was used to collect student's records which had personal, psychological and social related questions and was used in the study to identify relation between grades and performance of students. In their study they found that the most influencing factor in student's performance is the performance grade in secondary school, which says that students performing well in secondary school will also continue to perform well in Bachelor study. It was also found that medium of teaching, annual income, family status, living location, parents qualification and many other similar parameters were highly contribute in student performance. Hence student performance can be predicted if the personal and social information about a student is known [4].

Revised Manuscript Received on December 30, 2019.
* Correspondence Author
Mahesh Mardolkar, Research Scholar, Department of Computer & Information Science, Annamalai University, Chidambaram, Tamil Nadu, India. Email: principal@bharateshbca.com
Dr. N. Kumaran, Department of Computer & Information Science, Annamalai University, Chidambaram, Tamil Nadu, India.

III. DATA MINING PROCESS

The relation between students performance, personal and social factors is the main objective of this study. This study also focuses on prediction of student's performance. The study was conducted on students of Bharatesh English Medium School, Belagavi, Karnataka, India studying in class IX.

A. Dataset

A survey is conducted to collect and create dataset. A questionnaire is designed to collect the data covering personal, academic and social details of a student.

Table. 1 shows the different attributes and the possible values in the questionnaire.

Following is the details of how the questionnaire is processed mentioning some of the attributes.

- *LMED:* The language which is spoken by the student much of the time at home or even with friends. Since the language in which is taught if it is different from mother tongue. Then will impact the student performance.
- *PPER:* Is an attribute which tells us about previous academic performance of a student.
- *MFAM:* Options like no parents, single parent, both parents, medium family or large family is checked.
- *AINCOME*: This attribute tells us the financial status of the student's parents. Since the students option for part time job in low income category which again affects their performance at school.
- *NFRI:* The number of friends also affects the performance since much of the time is spent with friends.
- *SATT:* Students spent much of their time in other activities neglecting studies, like working part time, involvement in sports and other competitions, due to shortage of attendance performance is low [5]-[7].

Table- 1: Description and possible values of attributes

Attribute	Description	Values
CATEGORY	Category/ Caste	{SC/ST, Cat-1, OBC, GM, Minority}
LMED	Language of Medium	{different, little similar, similar}
PPER	Previous Performance	{D/40, C/50, B/60, A/70, A+/80 above}
MFAM	Family size	{more, 3, 2, 1, none}
EMEM	Earning members in family	{none, 1, 2, 3, more}
AINCOME	Annual Income	{<50,000, 1 lac, 2lac, 3lac, more}
APER	Academic pressure	{poor, fair, good, very good, excellent}
NFRI	Number of friends	{more, 20, 10, 5, less}
TFRI	Time spent with friends	{more, 3hr, 2hr, 1hr, no}
SATT	Shortage of attendance	{always, often, sometimes, rarely, no}
ICLASS	Classes are interesting	{not interesting, fair, good, very good, excellent}
FEES	Difficulty in paying fees	{always, often, sometimes, rarely, no}

B. Data Exploration

The dataset needs to be explored in statistical manner for better understanding. Further need to implement visualization using diagrams and graphical plots. Visualization is required to apply more complex algorithms and data mining technique.

Table- 2: Range of values in dataset

Attribute	Values
CATEGORY	{SC/ST(2), Cat-1(0), OBC(29), GM(29), Minority(41)}
LMED	{different(87), little similar(14), similar(5)}
PPER	{D/40(2), C/50(4), B/60(19), A/70(33), A+/80 above(48)}
MFAM	{more(96), 3(7), 2(0), 1(1), none(2)}
EMEM	{none(1), 1(77), 2(20), 3(4), more(4)}
AINCOME	{<50,000(58), 1 lac(19), 2lac(13), 3lac(5), more(10)}
APER	{poor(1), fair(7), good(41), very good(31), excellent(26)}
NFRI	{more(33), 20(7), 10(17), 5(32), less(17)}
TFRI	{more(47), 3hr(11), 2hr(17), 1hr(25), no(6)}
SATT	{always(2), often(2), sometimes(33), rarely(20), no(49)}
ICLASS	{not interesting(10), fair(23), good(24), very good(31), excellent(18)}
FEES	{always(5), often(3), sometimes(43), rarely(2), no(53)}

Table- 3: Summary of dataset

Attribute	Mode	Least
CATEGORY	Minority(41)	SC/ST(2)
LMED	different(87)	similar(5)
PPER	A+/80 above(48)	D/40(2)
MFAM	more(96)	2(0)
EMEM	1(77)	none(1)
AINCOME	<50,000(58)	3lac(5)
APER	good(41)	poor(1)
NFRI	more(33)	20(7)
TFRI	more(47)	no(6)
SATT	no(49)	always(2), often(2)
ICLASS	very good(31)	not interesting(10),
FEES	no(53)	often(3)

According to the attributes the range of data in dataset is shown in Table 2, ordered from highest to lowest.

The least frequency value and the highest frequency value and their summary of statistics are shown in Table 3.

C. Implementation and Result

Classification, Association, Clustering, Artificial Intelligence etc are well known techniques in data mining and Knowledge Discovery in Database (KDD). The most widely used data mining technique is classification. Classification is very simple and easy to use technique. Classification predict data object's class or category based on training dataset used on previously learned classes. There are different types of classification techniques - Decision Tree, KNN (k-Nearest Neighbor), Naïve Bayes, Neural Networks etc. The decision tree classification technique is used in data mining process for predicting the student's dropout using grade. The data mining implementation and processing is done on RStudio and Weka. Decision tree is an supervised classification technique which builds top-down tree model for given dataset. It is also predictive modeling technique used for classification or prediction using the available training dataset. The tree includes nodes, internal nodes and the leaf node. The last node is the final suggested class of the data object. In this study C4.5 decision tree and ID3 decision tree is used on the student dataset.

Table- 4: Confusion matrix of C4.5 algorithm

		Actual				Class Precision (%)
		Excellent	*Very Good*	*Good*	*Pass*	
Prediction	*Excellent*	43	2	3	0	46.2
	Very Good	29	4	0	0	66.7
	Good	15	0	4	0	57.1
	Pass	4	0	0	0	-
Class Recall (%)		89.6	12.1	21.1	0	

Table- 5: Confusion matrix of ID3 algorithm

		Actual				Class Precision (%)
		Excellent	*Very Good*	*Good*	*Pass*	
Prediction	*Excellent*	40	2	3	0	43.7
	Very Good	24	3	0	0	62.4
	Good	13	0	4	0	53.2
	Pass	4	0	0	0	-
Class Recall (%)		84.6	10.8	21.1	0	

Ross Quinlan developed C4.5 and ID3 decision tree algorithm. The technique of pruning is used in C4.5 which removes the node that adds no value to the final prediction. Unprimed decision tree is generated in case of ID3 algorithm for producing the decision tree. Following settings are used.

Splitting criterion = Information gain ratio

Minimal leaf size = 1

Minimal split size = 4

Minimal gain = 0.1

The confusion matrix generated after running the C4.5 and ID3 decision tree algorithm is shown in Table 4 and Table 5 respectively. C4.5 algorithm was able to predict 51 objects out of 106 with an accuracy of 48%, also the ID3 algorithm was able to predict 47 objects out of 106 with an accuracy of 43%. In further review of these algorithms in terms of performance and accuracy C4.5 has better accuracy of 48% as compared to ID3 with 43% accuracy, shown in Table 6.[8], [9].

Table- 6: Accuracy comparison of algorithms

Algorithm	Accuracy
C4.5	48%
ID3	43%

IV. CONCLUSION

Multiple data mining tasks to create predictive model for the prediction of grades from student's data is discussed in this paper. A survey was constructed to collect personal, social and academic data related to the students. Finally interesting results were drawn from two well known decision tree algorithms. This study can help educational organizations to identify the students dropping out. With further counseling of such students can help them to continue their studies.

REFERENCES

1. Bharadwaj. B. K. and Pal, Mining Educational Data to Analyze students' performance, International Journal of Advance Computer Science and Applications, Vol2, No.6 2011.
2. Ahmed, A.B.E.D and Elaraby, I.S, Data Mining: A prediction for Students Performance Using Classification Method, World Journal of Computer Application and Technology 2(2), pp 43-47.
3. Pandey U. K. and Pal, S, Data Mining: A prediction of performer or underperformer using classification, International Journal of Computer Science and Information Technologies, Vol 2(2), 2011, 686-690.
4. Bharadwaj B.K and Pal, S, Data Mining: A prediction for performance improvement using classification, International Journal of Computer Science and Information Security, Vol 9, No.4, April 2011.
5. Yadav S. K., Bharadwaj B and Pal, S, Data Mining Applications: A Comparative Study for Predicting Students Performance, International Journal of Innovative Technology & Creative Engineering, ISSN:2045-711, Vol. 1, No.12, Dec 2012
6. Yadav S. K. and Pal,S , Data Mining: A prediction for performance improvement of engineering students using classification, World of Computer Science and Information Technology Journal, ISSN:2221-0741, Vol.2, No.2, 51-56,2012.
7. Amjad Abu Saa, Educational Data Mining & Students Performance Prediction, International Journal of Advance Computer Science and Applications, Vol. 7, No.5, 212-220, 2016.
8. Mahesh Mardolkar, N. Kumaran, School Dropout Analysis with R Programming Charts, International Journal of Research, Vol.5, Issue-04, ISSN: 2348-6848, Feb 2018
9. Mahesh Mardolkar, N. Kumaran, Universal Comparison of School Education in RStudio, International Journal of Management, Technology and Engineering, Vol.8, Issue XII, ISSN:2249-7455, Dec 2018.

AUTHORS PROFILE

Mahesh Mardolkar, pursuing his Ph.D. in Computer Science from Annamalai University, Chidambaram, India, he is Principal at Bharatesh College of Computer Applications, Belagavi, India, he has done research work in the field of Data Mining and the area of focus is Educational Data Mining.

Dr. N. Kumaran, received the B. E degree in Computer Science and Engineering from Adhiparasakthi Engineering College, Melmaruvathure in 2002. He received the M.E degree in Computer Science and Engineering from Annamalai University, Annamalainagar in the year 2004. He worked at Adhiyamaan College of Engineering, Hosur from 2004 to 2006. He has been with Annamalai University, since 2006. He awarded his Ph.D in Computer Science and Engineering at Annamalai University in the year 2016. He published 20 papers in international conferences and journals. His research interest includes Digital Image Processing, Medical Imaging, Content Based Image Retrieval and Genetic Algorithms.