I.J. Education and Management Engineering, **2017, 2, 8-19**
Published Online March 2017 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijeme.2017.02.02

Available online at http://www.mecs-press.net/ijeme

Literature Survey on Educational Dropout Prediction

Mukesh Kumar [1], Prof. A.J. Singh [2], Dr. Disha Handa [3]

[1,2] *Himachal Pradesh University, Summer-Hill, Shimla (H.P) Pin Code: 171005, India.*
[3] *IT Consultant, DesktekTeam*

Abstract

Educational Data Mining (EDM) is one of the crucial application areas of data mining which helps in predicting educational dropout and hence provides timely help to students. In Indian context, predicting educational dropouts is a major problem. By implementing EDM, we can predict the learning habits of the student. At present EDM has not been introduced at higher education level. Due to this we cannot recognize the genuine problems of students during their education. The objective of this analysis is to find the existing gaps in predicting educational dropout and find the missing attributes if any, which my further contribute for better prediction. After that we try to find the best attributes and DM techniques which are frequently used for dropout prediction. Based on the combination of missing attribute and best attribute of student data thus far, a new algorithm can be tested which may overcome the shortcomings of previous work done.

Index Terms: Educational Data Mining, Prediction Techniques, Educational Dropout, Analysis.

1. Introduction to Data Mining in Education

Data mining helps to mine the unique and significant data from the data warehouse. It is used to express knowledge discovery and search for essential relationships among different variables/attribute in the data warehouse. It can be used in the different field of the real world like banking, education, medical, telecommunications and fraud detection etc. EDM has emerged due to the increasing accessibility of educational data and hence the need to analyze this massive data. EDM is a multidisciplinary field of research that is used to analyze educational data using data mining techniques [1]. It is essential in education especially when want to check the performance of the student in the near future with their previous record in their education. It is extremely time-consuming and durable process if we want to analyze the student performance manually. The outcome of the manual process is also not up to the mark. There is a dilemma of school dropout in the education system in India and we just try to find out the reasons for their dropout [2]. For that, work has been done to find out the diverse attribute of the student which helps them to dropout in education. So reducing

Corresponding author. Tel.: 8872671333
E-mail address: mukesh.kumarphd2014@gmail.com

the education dropout in India is one of the challenges that educational institution is dealing with.

Nomenclature	
DM	Data Mining
EDM	Educational Data Mining
CART	Classification and Regression Tree
ID3	Iterative Dichotomiser
ITS	Intelligent Tutoring System
SVM	Support Vector Machine

They aims to enroll more students by providing qualified faculty, best infrastructure, improving laboratories work, sports facilities, and best study program. After enrolment, the main aim of each faculty is to guide each student to effectively complete their studies with the proper knowledge and acquire good skills [2] [3]. Nowadays, however, the deployment of Student Information Systems at the institutional level provides an appropriate infrastructure for student's data organization and storage as well as data acquisition and deeper analyses. This data can help model the behaviour of dropouts, and predict future dropouts, therefore giving a chance to counselors to advise and guide students into success. The demand for education in India has increased as more and more children are now attending their schools. But there is lots of problem with the education system causing many students to drop their study. We lack in good infrastructure, quality teachers and poor delivery of course content in India causing people to drop out. It is a common excuse for the students that they don't have easy access to educational institutions. This problem is very true for that student how migrants for different places due to their family problem. They just face the problem for the issue of transfer certificates, school leaving certificate and other such formalities. “It is our educational system that is not encouraging people, creating more and more formalities for the migrant’s students”. Due to all the formalities needed to fulfill, it looks easier to shift jobs than to shift a schools/colleges and once a child is out of school for too long, admissions become even more difficult.

The objectives of the present study are:

i) To analyze different DM techniques used in the education.
ii) To analyze the work done by different researchers in the field of education with DM till 2016.
iii) To analyze the work done by different researchers to predict the educational dropout with DM.
iv) To spot different attributes and DM techniques which are frequently used for dropout prediction?
v) To find any existing gaps in predicting dropout and find the missing attributes if any, which my further contribute in the better prediction.

Section II describes different data mining techniques which are mostly applied in the education system. In section III describes recent work on educational data mining by different researcher since 2009-2016 and recent work done on educational dropout. Section IV describes different attributes selected by researchers for predicting educational dropout. In last section, a conclusion on the review is given with future scope.

2. Data Mining Technique’s Relevant in Education Surroundings

EDM is creating and using some computer based algorithm to identify patterns in a huge educational database. Without EDM it is impossible to analyze enormous amount of data and hence the patterns. Different types of algorithms and techniques like Classification, Regression, Association Rules mining, Genetic Algorithm, Clustering, Nearest Neighbors method, Decision Trees are used for information retrieval from educational databases. These techniques are further used for predicting student performance, educational dropout, design new curriculum etc. In this section, brief introduction of different data mining techniques are given.

Classification:

Classification techniques are completely based on machine learning. These techniques classify each dataset into predefined classes. To classify data in database some mathematical techniques similar to neural network, decision trees, statistics and linear programming are used. We try to understand the problem of classification with a real life example. Let us take an example of University in which different students are studying. With the help of classification techniques we can predict about those students who may have educational dropout in near future. We can also classify different student according to their performance in their study.

Clustering:

Clustering is used to make cluster of comparatively identical cases or observations. Things in a cluster are comparable to each other. They are also unrelated to things outside the clusters. Let us take an example of university in which lots of students are studying. We can cluster different student according to their attributes like course, grade, activity, age, gender, hosteller, day-scholar, rural, urban area student. With this we can provide different types of facility to different cluster according to their specific requirement.

Prediction:

Prediction is useful to predict the value of an unknown attribute with the help of some known attribute. With the help of these techniques, we can predict the future learning habits of the student. By applying these techniques on the academics data; we can predict about the student's performance in near future. Also we can predict the student result with respect to their sessional test record. In this example, result is a dependent variable and sessional marks are independent variables.

Association Rule mining:

It is well-known and leading technique in data mining. With this technique we can find the hidden patterns between different attributes of a single dataset. This technique is also known as relational data mining technique. The primary use of it is in market-basket analysis. In educational context, it is used to find the association between different attributes of the student's by which the performance affects. P. Sunil Kumar, D. Jena et al presented "Mining the factors affecting the high school dropouts in rural areas". In this paper author considered seven different attributes to find out the relationship between attribute which affect the student to dropout. In his analysis, they found that the students who are not interested in the study are mostly going to be the school dropout as compared to teaching environment and poverty. They also tried AV analysis, Correlation and conviction analysis and found that poverty as well as teaching environment also make student disinterested in the study [16].

3. Recent work done on Educational Data Mining

During last few year a lots of work have been done in EDM, like to predict the student's academic performance and progress, Educational Dropout prediction, Student placement prediction, Student final result prediction etc. The Result of these predictions is very helpful for framing as well as implementation of new rules and regulations, adopting new teaching methodology, improving placement records, improving curriculum in educational institution.

Searched databases: Springer Link, Researchgate, IEEE Xplore, ACM Digital library, Elsevier, Science Direct other computer science journals. Searching sentences and keywords: Predicting student performance, Predicting student performance is using data mining techniques, Application of data mining in education, EDM

methodology or techniques, student dropout prediction using data mining techniques. Publication periods taken into consideration: 2009 to 2016. Types of text searched: Documents, PDF, Full length paper with abstract and keywords. Search Items: Journal articles, Conferences paper, Workshop papers, Expert lectures or talks, topics related blogs, Topic related communities (like Educational data mining community).

In this section, first we discuss about work done by different researcher in educational data mining. In the second part of this section we especially discuss about the work done on Educational Dropout.

3.1. Work done on educational setting by researcher since 2009-2016

EDM is a group of dissimilar area that incorporates student result prediction and classification by using some techniques. The first annual conference on EDM has been started since 2008. After that lots of publications and articles have been published in this area. So we are selecting some publication of high-quality for analysis and present the result of each publication in a table. This section present an organized review of available literature during 2009-2016 based on the extremely cited paper in this province.

Table 1. Table Listing Authors Name with Their Year of Publication, Publisher, and Total Citation Till 2016

S. No	Research paper authors	Publication	Publisher	Citation
1	Ryan S.J.D. Baker, Kalina Yacef	2009	JEDM	699
2	Shu-Hsien Liao, Pei-Hui Chu	2012	Elsevier	148
3	Siti Khadijah Mohamada et al	2013	Elsevier	149
4	Cristobal Romero and Sebastian Ventura	2013	Wiley & Sons	168
5	Alejandro Pena-Ayala	2013	Elsevier	85
6	Muna Al-Razgan, Atheer S et al	2014	Springer	37
7	Kenneth R. Koedinger et al	2015	Wiley & Sons	11
8	Laura Calvet Liñán et al	2015	RUSC	20
9	Laci Mary Barbosa Manhães et al	2015	ACM	116
10	Manuel Ángel, José María Luna et al	2015	Springer	298
11	Vlatko Nikolovski, Riste Stojanov at al	2015	Conference	149
13	Zhi-Ting Zhu, Ming-Hua Yu et al	2016	Springer	28
14	Yasmeen Altujjar et al	May 2016	SDMA2016	1

Ryan S.J.D. Baker, Kalina Yacef presented “The State of Educational Data Mining in 2009: A Review and Future Visions". In this paper, the author discusses the entire rapidly growing field of EDM. In this paper, the author points out that EDM remain used in Australia, New Zealand, North America and some part of Europe, with little participation from remaining part of the world. The impact of DM in education remains same like other related fields like data mining in intelligent tutoring system using artificial intelligence [1].

Cristobal Romero and Sebastian Ventura presented “Data mining in education”. In this paper, researcher brings the different community of educationalist and researcher together like computer scientists and learning scientists. Here author apply different DM techniques to analyze data generated during teaching and learning process in different education practice. After analysis, the result is used for further decision making to improve the educational practice [3].

Muna Al-Razgan, Atheer S et al presented “Educational Data Mining: A Systematic Review of the Published Literature 2006-2013”. The entire researchers on educational data mining domain are working toward the development of educational games, mobile application learning and ITS for learners. The study found that everyone is focusing on working on the educational games, intelligent tutoring and mobile development application [4].

Kenneth R. Koedinger et al presented “Data mining and education”. In this research, authors tries to describe the different, exciting and slowly growing areas of educational data mining. In his view, educational data mining areas are one of the interesting areas to study because it touches the basic research question that how students learn and behave in multiple disciplines of his life. This question is important to answer because it contributes to the development of the student as well as for the development of the society [5].

Laura Calvet Liñán et al presented “Educational Data Mining and Learning Analytics: differences, similarities, and time evolution”. In this paper, the author presented the similarities between two different areas of research. They highlighted the goals, types of methodologies and techniques used in EDM and LA. They also reveal the difference between these two areas of learning like their origins and trends. But at last the outcome of each learning process leads to help the development of the society. As these two learning processes affect the society still they come up with some barriers, easy to understand tools [6].

Laci Mary Barbosa Manhães et al presented “Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs”. In this paper, the author applies the time-varying data collection for the prediction of the student's performance. Most of the student data are stored in the academic registries and no other major external data collections are needed to predict student performance. The experimental results are further used to predict the students who are at risk of dropping out in their study [7].

Manuel Ángel, José María Luna et al presented “Discovering Clues to Avoid Middle School Failure at Early Stages”. In this paper, the author is using WEKA which one of the most used data mining tool for data analysis. They are also using 10 fold classical techniques which address the problem with different perspective and different comparison. They further selected Bayesian probabilistic classifier (NaiveBayes), SVM, Multi-layer Perceptron (MLP), K-nearest neighbour techniques based on rules and decision trees (JRip, OneR, J48 and PART) and a nominal class classifier based on boosting (AdaBoostM1) in different datasets [8].

Vlatko Nikolovski, Riste Stojanov at al presented “Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education”. In his work author find that the higher percentage of accuracy of any classifier algorithm is totally dependent on the quality of attributes and data model which are selected for the data collection. According to analysis of the dataset, helpful attributes are mathematical and programming courses application. Since the results reveal a pattern among the number of examination applications between the mathematical and programming courses, equally important are the demographic characteristics of the students [9].

Zhi-Ting Zhu, Ming-Hua Yu et al presented “A research framework of smart education”. In this paper, author tries to explain the meaning of smart education system less than the smart city design. They afford each individual of the smart city with modified services with faultless teaching-learning understanding. Learning is a lifelong process and generates a lot of behavioral data of the learners, now the problem is how to integrate this data in the smart education system for providing quality educational services. Due to these listed problems, researcher now try to find the interconnection between the smart education system and how they interoperable with each other [10].

Yasmeen Altujjar, Wejdan Altamimi et al presented “Predicting Critical Courses Affecting Students Performance: A Case Study”. They used ID3 classification data mining algorithm on the dataset of bachelor's student of Information Technology. They intend to predict the student feat and hence to recognize significant courses in the bachelor program. They follow this process on each year student and found that feat in the 2nd year is the most perfect [11].

From above discussion, it is clear that different researcher has been working on educational data and finds different result according to their requirement. The different areas of EDM are online course study analysis, result prediction of student, Predicting student placement record, analyzing student web learning habits, MOOC course analysis, predicting educational dropout etc. In most of the research papers, researchers have used classification, clustering and association algorithms of data mining for their prediction in educational setting. These techniques were extensively used in the early period of EDM but still widely used in the different application area. To make prediction students attributes are used. The different attributes used for student performance prediction are academic, social, demographics, personal and family. But in most of the case

student CGPA, internal marks, external marks, parent education and occupation and poor teaching methodology are the main factors which affect the result of the students. We concentrate on three different areas: Different areas of Data Mining in education setting, Main attributes which affect the performance of the students, Different Data Mining Techniques used for Prediction of student performance. We have presenting all our outcome of this study in Table 2.

Table 2. Review of Different Research Papers on Educational Data Mining

Subject/Topic	Possible values
Different application areas of Data Mining in education setting	To analyze online course-study, Predicting student Performance, To predict the student result, Recommending to students, Grouping students, Student modeling, Constructing courseware, Planning and scheduling, Detecting undesirable student behaviors, To predict the student placement record, To analyze student web learning habits, To analyze MOOC course result, To predict educational dropout, To reframe new curriculum for course, Providing feedback for supporting instructors, Social network analysis
Main attributes which affects the progress of the students	Gender, Student CGPA, Internal marks, External marks, Parent's education , Parent's occupation, Poor teaching methodology, Student interest, Attendance, Level of motivation, Score in math's
Different Data Mining Techniques used for Prediction of student progress in education	ID3, JRip, OneR, J48, PART, NaiveBayes, SMO, RBFNetwork , MLP, IBk, CHAID, CAR, C4.5, ICRM, BayesNet(BN) , NB(NaiveBayes), SVM1, SVM2, AdaBoost(AB) , DecisionTable(DT), RandomForest(RF)

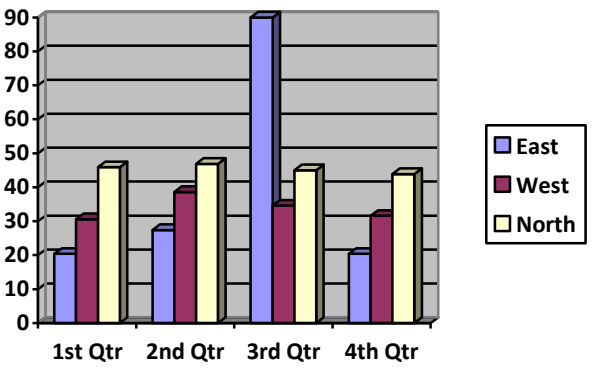
Authors are using different attributes and algorithm for the prediction of the student performance according to their requirement. But we find educational dropout prediction very interested to study because it is biggest problem. In India we don't have any fixed criteria to check the student performance.

3.2. Recent work done on Predicting Educational Dropout

Educational dropout is one of the major problems in India. There are lots of factors which affect the student education, but financial and domestic responsibilities are the major reason for educational dropout. A survey by National Sample Survey office shows that in India 13 out of 100 between the age group of 5-29 years is education dropout because they did not consider education as important in their life. In another survey conducted by the independent agency, that one out of four students is educational dropout due to the similar reason. With EDM, we can find out the reasons of students educational dropout. Further using these reasons we can try to improve the performance and progress of the student. We are selecting some qualities research paper to understand the work done on educational dropout. Table-3 listing authors name, techniques used in research, publication year of research papers and total citations.

Table 3. Table Listing Authors Name, Techniques used in Research, Publication Year and Outcomes.

S.No	Authors	Techniques used for analysis	Publication	Citations
1	Pedro A. Willging, Scott D. Johnson	Logistic regression analysis	2004	329
2	Russell Rumberger, Sun Ah Lim	Review only	2008	312
3	Gerben W. Dekker, et al	CART, C4.5, J48, SL, JRip, RF	2009	179
4	Dr. Saurabh Pal	Naïve Bayesian Algorithm	2012	12
5	P. Sunil Kumar, D. Jena	Association rules mining	2013	04
6	Miguel Gil, Norma Reyes et al	ANN based algorithm	2013	15
7	Sateesh Gouda, Dr. TV Sekher	Logistic Regression	2014	32
8	Allan Sales, Leandro B. et al	NB, C5.0, SVM, LR and MP	2015	25
9	Subitha Sivakumar, Sivakumar et al	Improved ID3 algorithm	2016	04
10	Carlos Márquez-Vera, et al	ICRM2	2016	05



Pedro A. Willging, Scott D. Johnson presented "Factors that influence student's decision to dropout of online courses". They exclusively focused on online master's degree program of University of Illinois. The university award master's degree after the completion of nine HRE Online course. To predict the online course dropout attributes, they use a logistic regression for analysis. They considered seven categorical free variables like age, gender, cohort, ethnicity, occupation, location, and GPA for analysis. After analysis, only GPA is found to be significant variable by the logistic regression analysis [12].

Russell Rumberger and Sun Ah Lim presented "Why Students Drop out of School: A Review of 25 Years of Research". The different nations today are facing the problems of student educational dropout due to the different reasons. Most of the student who entered into a four years diploma course failed to get it. To understand the reason for this, we need to understand the student performance at every step of their education. There are lots of influencing points which help to raise the student dropout rate. Parent's interference in playschool and untimely elementary school is obviously acceptable to decrease the dropout rate. In an early elementary school education, careful experimental evaluations and small class size have proven to decrease the dropout rate and improve the graduation rate [13].

Gerben W. Dekker, et al presented "Predicting Students Drop Out: A Case Study". In this paper, they demonstrate effectiveness of different classification data mining techniques. They also discuss the effectiveness of different cost-sensitive approach on the data set generated by the division of EE at the Eindhoven University of Technology. They are using different data mining algorithm like CART, J48, C4.5, JRip, RandomForest and SimpleLogistic on the selected dataset and find the result with 75-80% accuracies which are very difficult to achieve by any other data mining algorithm [14].

Dr. Saurabh Pal in his paper entitled "Mining Educational Data Using Classification to Decrease Dropout Rate of Students". In this paper, they try to find out the education dropout for the student of IET of VBSPU, Jaunpur. They try to find out the education dropout with the help of data mining techniques. In his experiment, they apply machine learning algorithm to extract the knowledge from existing student data for making a predictive model for future. They use Naive Bayes classification algorithm to analyze the previous year student data. They collected the student academic data like High School grade, Senior Secondary grade, and student's family position etc, to predict the student's performance and find the list of that student who needs special attention [15].

P. Sunil Kumar, D. Jena et al presented "Mining the factors affecting the high school dropouts in rural areas". In this paper authors are considered seven different attributes and find the relationship between them which force the students to dropout. In their analysis, they found that the students who are not interested in the study are mostly the school dropout as compared to teaching environment and poverty. They also tried AV analysis, Correlation and conviction analysis and found that poverty as well as teaching environment also make student disinterested in the study [16].

Miguel Gil, Norma Reyes et al presented "Predicting Early Students with High Risk to Drop out of University using a Neural Network-Based Approach". ". In this paper, the author considered a Black Propagation Artificial Neural-network system, which is used to measure the probability of the education dropout of the university student once they enrolled in some academic program. They also found that the result of the analysis may be inconsistent due to student data filled in the survey. Sexual lives of the student are also being considered as one of the important factors for the prediction [17].

Sateesh Gouda, Dr. T. V. Sekher presented "Factors Leading to School Dropouts in India: An Analysis of National Family Health Survey-3 Data". In this paper authors found that cost of the study, interest of student households work, family income are some factors which force the student to dropout. In India, 6% of the girl students dropout their study when they got married. They categorize the most important reasons given by different households for the education dropout of their children. After grouping all these data an analysis was made and they found that 46% education dropout are due to household factor. Another factor like poor infrastructure, lack of good faculty in school is contributing 4% dropout for boys and 15% for the girl students. So at the end, they emphasize that if we want to decrease the dropout rate then try to improve the school infrastructure, recruit good faculty and improve the quality of education in the school [18].

Allan Sales, Leandro B. et al presented "Predicting Student Dropout: A Case Study in Brazilian Higher Education". In this paper author considered education dropout recognition problem as a classification problem. They apply the classification algorithm on different students of a public university of Brazil who is admitted in 130 different courses. They targeted only the first year students. They are applying some classification model and found some good result considering only some but an informative attribute. After analysis, they found that only STATUS.SEM and MEAN.APPR are one of the leading factors for the student prediction. More than 70% of F-measure can be achieved with these two factors of student data [19].

Subitha Sivakumar, Sivakumar et al presented "Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree". In this paper, the author tries to make an enhanced decision algorithm which is totally derived from the ID3 algorithm. The enhanced decision algorithm is more efficient in the prediction of the student whether the student dropout or may continue with their study in future. With the help of this improved algorithm, the administration of educational institution is forced to make new guidelines and policy for improvement [20].

Carlos Márquez-Vera, Alberto Cano, et al presented "Early dropout prediction using data mining: a case study with high school students". They found in his study that the classification algorithms are important or more trustworthy for the prediction of the student a very early stage or before the middle of the course. They analyze that after step II and step III they got very good result i.e. at first 4 and 6 weeks of the course. They also purposed a new ICRM2 algorithm which outperformed all other classification algorithm used till date [21].

From above discussion, we can conclude that Predicting Educational Dropout is important research topic on which all the research community of EDM working today. At present educational dropout is a major problem in Indian education system. We find educational dropout prediction very important. Because it further affect literacy rate, national development of any country. In this section we found that Age, Gender, Cohort, Ethnicity, Occupation, location, GPA, Parent's interference, careful experimental evaluations, small class size, High School grade, Senior Secondary grade, student's family position, interest, poverty, teaching environment, Sexual lives, cost of the study, households work, family income, got married, household factor, poor infrastructure, lack of good faculty, STATUS.SEM and MEAN.APPR are that most important attributes for predicting educational dropout.

Most of the EDM researcher found Logistic regression analysis, CART, C4.5, J48, SL, JRip, RF, Naïve Bayesian Algorithm, Association rules mining, ANN based algorithm, Logistic Regression, NB, C5.0, SVM, LR and MP, Improved ID3 algorithm and ICRM2 data mining algorithm important. We have presenting all our outcome of this study in Table 4.

Table 4. Review of Different Research Papers on Predicting Educational Dropout

Subject/Topic	Feature Involved
Attributes which are mainly helpful for predicting educational dropout	Age, Gender, Cohort, Ethnicity, Occupation, location, GPA, Parent's interference, careful experimental evaluations, small class size, High School grade, Senior Secondary grade, student's family position, interest, poverty, teaching environment, Sexual lives, cost of the study, households work, family income, got married, household factor, poor infrastructure, lack of good faculty, STATUS.SEM and MEAN.APPR
Different Data Mining Techniques used for Predicting educational dropout	Logistic regression analysis, CART, C4.5, J48, SL, JRip, RF, Naïve Bayesian Algorithm, Association rules mining, ANN based algorithm, Logistic Regression, NB, C5.0, SVM, LR and MP, Improved ID3 algorithm , ICRM2

At the end of this section, it is clear that most of the researcher used student academic, parent's qualification and occupation, institutional level attribute and demographic attribute for their analysis. In most of the cases also classification algorithm like CART, C4.5, J48, SL, JRip and RF are used for prediction.

4. Different Attributes and DMT which are Frequently used for Predicting Educational Dropout

In this section, we examine those attributes which force students' to be educational dropout. These attributes are also supportive for prediction of student's performance using EDM techniques. It has increasing attention and anxiety about the difficulty of student's disappointment and shaping the key factors contributing to this problem. There are lots of factors affecting student performance in their education. The best attribute is selected based on the frequent occurrence in all research paper taken into consideration for our research. In selected attributes, we have filtered 10 best attributes to predict the student dropout in education. The best selected attributes are grade in HSG, SSG and other related education, Gender, Family Grade's student in HSG, SSG and other related education, Gender (M/F), Family structure, Parents Qualification, Parents Occupation, Required for Household work, Addictions (Alcohol, Smoke, Pills, Solvents, Drugs etc), Basic facility in the education institution different for boys and girls, Poor Teaching methodology adopted, Got married depicted in Table 5.

Table 5. Attributes Considered Significant for Educational Dropout Prediction by Researcher

S.No	Important attributes of Student data considered for analysis	Frequency of attribute occurrences in listed papers
1	Grade's student in HSG, SSG and other related education	8
2	Gender(M/F)	6
3	Family Structure	5
4	Parents Qualification	4
5	Parents Occupation	5
6	Required for Household work	3
7	Addictions(Alcohol, Smoke, Pills, Solvents, Drugs etc)	4
8	Basic facility in the education institution different for boys and girls	2
9	Poor Teaching methodology adopted	5
10	Got married	2

Attributes and feature selection are essential aspect of the DM because there are lots of attributes are there for any student but their entire attribute is not as much important than other. For example, roll number is one of the attributes for a student but it does not affect the performance of the student as compared to other attributes like marks in academics. The objection behind the attribute collection is to enhance the feat of the algorithm, providing efficient and profitable prediction for the improved perception of the procedure of data collection which we follow for analysis.

In case of data mining algorithm used for predicting educational dropout, classification and association rule mining are mostly used. About more than 50% of studied research papers used these techniques for prediction of dropout student. Logistic regression analysis is also an important algorithm for prediction.

5. Conclusion and Future Scope

In the nutshell, it is important to say that predicting educational dropout is a major, important and challenging task for every education institution's administrator, policy maker and educators. To deal with these problems, researcher tries to make uses of DM techniques. After studied many significant research papers, we realized that data mining techniques will prove helpful to improve the educational standard. In our study, we found that data mining techniques are useful in an online study program, engineering institution for predicting their placement, predicting educational dropout, the overall result of the student etc. Most of the researchers in educational data mining using Naïve Bayesian Algorithm, Association rules mining, ANN based algorithm, Logistic Regression, CART, C4.5, J48, (BayesNet), SimpleLogistic, JRip, RandomForest, Logistic regression analysis, ICRM2 for the classification of the educational dropout student. With the help of data mining techniques, we can find out all those attributes which help the student to dropout.

The best selected attributes are grade in HSG, SSG and other related education, Gender, Family structure, Parents Qualification, Parents Occupation, Required for Household work, Addictions (Alcohol, Smoke, Pills, Solvents, Drugs etc), Basic facility in the education institution different for boys and girls, Poor Teaching methodology adopted, Got married. These selected features may be increased or may be changed in our future research. Because we are considering these attributes according to their frequency of occurrences in the selected papers which are under consideration.

Acknowledgements

I am grateful to my guide Prof. A.J. Singh and Dr. Disha Handa for all help and valuable suggestion provided by them during the study.

References

[1] Baker, R. S. J. D., & Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. Journal of Educational Data Mining, 1, 3-17.

[2] Siti Khadijah Mohamad, Zaidatun Tasi presented "Educational data mining: A review". The 9th International Conference on Cognitive Science Procedia - Social and Behavioral Sciences 97 (2013) 320 – 324

[3] Cristobal Romero and Sebastian Ventura presented "Data mining in education". WIREs Data Mining Knowl Discov 2013, 3: 12–27 doi: 10.1002/widm.1075

[4] Muna Al-Razgan, Atheer S et al presented "Educational Data Mining: A Systematic Review of the Published Literature 2006-2013". Proceedings of DaEng-2013, DOI: 10.1007/978-981-4585-18-7_80

[5] Kenneth R. Koedinger et al presented "Data mining and education". WIREs Cogn Sci 2015. doi: 10.1002/wcs.1350

[6] Laura Calvet Liñán et al presented "Educational Data Mining and Learning Analytics: differences, similarities, and time evolution". Universities and Knowledge Society Journal, 12(3). pp. 98-112. doi: http://dx.doi.org/10.7238/rusc.v12i3.2515

[7] Laci Mary Barbosa Manhães et al presented "Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs". ACM 978-1-4503-3196-8/15/04$15.00.

[8] Manuel Ángel, José María Luna et al presented "Discovering Clues to Avoid Middle School Failure at

Early Stages". LAK '15, March 16 - 20, 2015, Poughkeepsie, NY, USA Copyright 2015 ACM 978-1-4503-3417-4/15/03 $15.00

[9] Vlatko Nikolovski, Riste Stojanov at al presented "Educational Data Mining: Case Study for Predicting Student Dropout in Higher Education". https://www.researchgate.net/publication/282333827

[10] Zhi-Ting Zhu, Ming-Hua Yu et al presented "A research framework of smart education". Zhu et al. Smart Learning Environments (2016) 3:4 DOI 10.1186/s40561-016-0026-2

[11] Yasmeen Altujjar, Wejdan Altamimi et al presented "Predicting Critical Courses Affecting Students Performance: A Case Study". DOI: 10.1111/exsy.12135, Expert Systems, February 2016, Vol. 33, No. 1, © 2015 Wiley Publishing Ltd

[12] Pedro A. Willging, Scott D. Johnson presented "Factors that influence students' decision to drop out of online courses". JALN Volume 8, Issue 4 - December 2004.

[13] Russell Rumberger and Sun Ah Lim presented "Why Students Drop out of School: A Review of 25 Years of Research". California Dropout Research Project, October 2008.

[14] Gerben W. Dekker, et al presented "Predicting Students Drop Out: A Case Study". Educational Data Mining 2009.

[15] Dr. Saurabh Pal in his paper entitled "Mining Educational Data Using Classification to Decrease Dropout Rate of Students". International journal of multidisciplinary sciences and engineering, vol. 3, no. 5, may 2012.

[16] P. Sunil Kumar, D. Jena t al presented "Mining the factors affecting the high school dropouts in rural areas". (IJACECT), ISSN (Print): 2278-5140, Volume-2, Issue – 3, 2013.

[17] Miguel Gil, Norma Reyes et al presented "Predicting Early Students with High Risk to Drop out of University using a Neural Network-Based Approach". ICCGI 2013, ISBN: 978-1-61208-283-7.

[18] Sateesh Gouda M1, Dr.T.V.Sekher2 presented " Factors Leading to School Dropouts in India: An Analysis of National Family Health Survey-3 Data". (IOSR-JRME) e-ISSN: 2320–7388, p-ISSN: 2320–737X Volume 4, Issue 6 Ver. III (Nov - Dec. 2014), PP 75-83

[19] Allan Sales, Leandro B. et al presented "Predicting Student Dropout: A Case Study in Brazilian Higher Education". 3rd KDMiLe – Proceedings – ISSN 2318-1060, Oct 13-15, 2015 – Petropolis, RJ, Brazil.

[20] Subitha Sivakumar, et al presented "Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree". DOI: 10.17485/ijst/2016/v9i4/87032, January 2016.

[21] Carlos Márquez-Vera, Alberto Cano, et al presented "Early dropout prediction using data mining: a case study with high school students". Expert Systems, February 2016, Vol. 33, No. 1, © 2015 Wiley Publishing Ltd.

[22] B. R.B., T. S.S and S. A.K, "Importance of Data Mining in Higher Education System," Journal Of Humanities And Social Science (IOSR-JHSS), vol. 6, no. 6, pp. 18-21, 2013.

[23] J. Luan, "Data Mining and Knowledge Management in Higher Education -Potential Applications." in Processdings of AIR Forum, Torento, Canada, 2002.

[24] B. Baradwaj and S. Pal, "Mining educational data to analyze student's performance," International Journal of Advanced Computer Science and Applications, vol. 2, no. 6, pp. 63-69, 2012.

[25] A. Kumar and Vijaya lakshmi, "Implication Of Classification Techniques In Predicting Student's Recital," International Journal of Data Mining & Knowledge Management Process, vol. 1, no. 5, pp. 41-51, 2011.

[26] S. Kotsiantis, "Educational data mining: a case study for predicting dropout-prone students," International Journal of Knowledge Engineering and Soft Data Paradigms, vol. 1, no. 2, p. 101, 2009.

[27] D. G. W, P. Mykola and V. J. M, "Predicting Students Drop Out: A Case Study," International Working Group on Educational Data Mining, 2009.

[28] B. Jaroslav, H. Bydzovská, J. Géryk, T. Obsivac and L. Popelinsky, "Predicting Drop-Out from Social Behaviour of Students," International Educational Data Mining Society, 2012.

[29] L. Rokach, Data mining with decision trees: theory and applications, vol. 69, World scientific, 2008.

[30] S. J. Russell and P. Norvig., Artificial Intelligence: A Modern Approach (AIMA), 3rd ed., Prentice Hall,

2009.
[31] J. a. P. Han and Y. Jian and Yin, "Mining frequent patterns without candidate generation," in ACM SIG MOD Record, 2000.

Authors' Profiles

Mukesh Kumar (10/04/1982) has pursuing PhD in Computer Science from Himachal Pradesh University, Summer-Hill Shimla-5. India. My research interest includes Data Mining, Educational Data Mining, Big Data and Image Cryptography.

How to cite this paper: Mukesh Kumar, A.J. Singh, Disha Handa,"Literature Survey on Educational Dropout Prediction", International Journal of Education and Management Engineering(IJEME), Vol.7, No.2, pp.8-19, 2017.DOI: 10.5815/ijeme.2017.02.02