ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Data Mining for Prediction of Students' Performance in the Secondary Schools of the State of Punjab

Sohajbir Singh Ubha, Gaganpreet Kaur Bhalla

M. Tech. Scholar, Dept. of CSE, Rayat Bahra University, Mohali, Punjab, India

Assistant Professor, Dept. of CSE, Rayat Bahra University, Mohali, Punjab, India

ABSTRACT: Education is the backbone of any country and it is very important to improve the educational strength of the country. There are various methods and challenges on the way, use of technologies like computers, smart rooms, projectors, and eBooks. But these recourses are useful only when we know which student needs which type of resource or, in other words, if we are able to predict the results of students, we can improve results and decrease drop out ratio. In our research, we used Data Mining in education to improve the results of the schools. In this research paper, we have explained data mining, its usefulness in education and its results with the use of WEKA Data Mining Tool. We took the real data of 838 students of senior secondary schools of Punjab and performed data transformations to get the useful attributes. After transformation and cleaning of data, we applied decision tree algorithms like J48, Naïve Bayesian, Random Tree, and Decision Stump to find the results.

KEYWORDS: KDD (Knowledge Discovery in Database), J48 Algorithm, Naïve Bayesian Algorithm, Random Tree Algorithm, Decision Stump Algorithm, WEKA data mining tool.

I.INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into meaningful and useful information. It is the practice of automatically searching large stores of data to discover patterns and trends that go beyond simple analysis. Data mining uses sophisticated mathematical algorithms to segment the data and evaluate the probability of future events. It allows the user to analyze data from many different dimensions or angles. Data mining, also popularly known as Knowledge Discovery in Database, refers to extracting or "mining" knowledge from large amounts of data.

Educational Data Mining is concerned with developing new methods to discover knowledge from educational database in order to analyse student trends and behaviours towards education. Educational data mining is defined as the area of scientific inquiry centered around the development of methods for making discoveries within the unique kinds of data that come from educational settings, and using those methods to better understand students and the settings which they learn in.

Data Mining is used to highlight meaningful information and support decision making. In the educational sector, for example, it can be helpful for course administrators and educators for analyzing the usage information and students activities during a course to get a brief idea of their learning. Visualization of information and statics are the two main methods that have been used for this task.

II. RELATED WORK

In [20] the authors identified the factors influencing the performance of students in previous examinations and tried to find out a suitable data mining algorithm to predict the grade of students so as to give them a timely help and improve their performance next year.The authors took the student data of rural and urban primary schools of district Betul, Madhya Pradesh. An experimental as well as survey methodology was adopted to generate a database. They used

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

WEKA data mining tool and three algorithms namely Decision Tree Algorithm, Naïve Bayesian Algorithm and Zero R Classification Algorithm. After analyzing all the results, they concluded that the type of the school doesn't influence the performance of students. In [2] the author said that datamining is very useful in educational field to find important pattern of data and by using prediction method, a model can be developed which can be used to predict students' performance. Different data mining techniques like classification, clustering, association rule mining and regression etc. can be used for this purpose. This prediction will help the teachers to identify the weak students and help them in improving their performance. She concluded that predicting students' performance is one of the major applications of educational data mining and by using decision trees, their performance can be predicted and students with poor performance can not only be warned but management and teachers can also take appropriate steps like giving more attention and taking extra lectures by which their performance can be improved. In [16] the author applied data mining techniques to predict the higher education admissibility of women students at Government Arts and Science Colleges in the state of Tamil Nadu. She studied 690 undergraduate students from Government Arts College, Pudukkottai. The main focus of the research was on the development of data mining models for predicting the students likely to go for higher studies based on their personal, pre-college, and graduate performance characteristics. She applied Decision Tree Classifier and Naïve Bayesian Classifier algorithms on the data set. She found the highest level of accuracy through the Decision Tree Model.

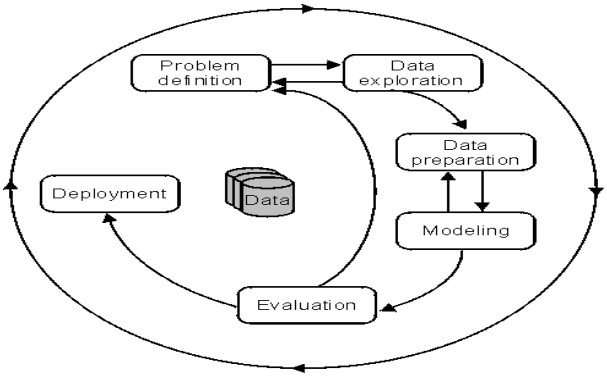
III. COMPONENTS OF DATA MINING

The major components of a data mining system are data source, data warehouse, data mining engine, knowledge base and pattern evaluation module. It shows that our system is using the historical data from the data warehouse server and then training the data after applying various pre-processing techniques.

IV. PHASES OF DATA MINING

Data mining is an iterative process that typically involves the following phases:

- Problem definition
- Data exploration
- Data preparation
- Modelling
- Evaluation
- Deployment



V. OBJECTIVES OF THE STUDY

- To collect, clean and integrate the data collected from the schools.
- To implement J48, Random Tree, Decision Stump and Naïve Bayesian Classifier algorithms on integrated data using WEKA Data Mining Tool.
- To visualize the data in different formats.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

- To compare and evaluate the results of J48, Random Tree, Decision Stump and Naïve Bayesian Classifier algorithms using different parameters.

VI. SCOPE AND SAMPLE OF THE STUDY

The study is based upon the data collected from 15 schools in the state of Punjab. The selection of the schools was made from the schools under the management of Shiromani GurdwaraParbandhak Committee, Sri Amritsar which is running more than 70 schools. It is a charitable body which was formed under the Legislative Act of 1925 and is the apex body of the Sikhs. The Population for the purpose of the study consists of 838 students of the sample selected. Through extensive search of the literature and discussion with experts on student performance, a number of factors that are considered to exert influence on the performance of a student were identified. These influencing factors were categorized as input variables. The output variables, on the other hand, represent some possible grades.

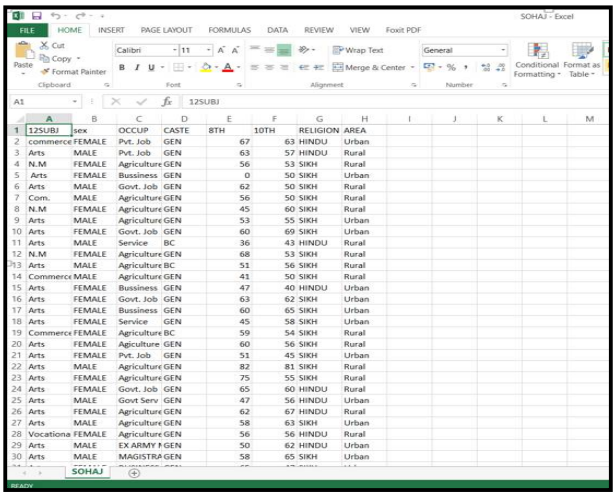
VII. METHODOLOGY OF THE STUDY

In our study, we used WEKA data mining tool. WEKA supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection.

The data collected by the researchers, was transferred to excel sheet and the required attributes were selected and converted into .csv 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, Decision Stump and Naive Bayesian classifier algorithms. The classified data was visualized in different formats such as graphs and decision trees. The results of the above four algorithms were compared in order to reach the decisions and give final recommendations on different parameters of the study.

VIII. CLEANED AND TRANSFORMED DATA

After cleaning and transformation of data the data, with following eight attributes was found suitable for further analysis:



	12SUBJ	sex	OCCUP	CASTE	8TH	10TH	RELIGION	AREA
2	commerce	FEMALE	Pvt. Job	GEN	67	63	HINDU	Urban
3	Arts	MALE	Pvt. Job	GEN	63	57	HINDU	Rural
4	N.M	FEMALE	Agriculture	GEN	56	53	SIKH	Rural
5	Arts	FEMALE	Bussiness	GEN	0	50	SIKH	Urban
6	Arts	MALE	Govt. Job	GEN	62	50	SIKH	Rural
7	Com.	MALE	Agriculture	GEN	56	50	SIKH	Rural
8	N.M	FEMALE	Agriculture	GEN	45	60	SIKH	Rural
9	Arts	MALE	Agriculture	GEN	53	55	SIKH	Urban
10	Arts	FEMALE	Govt. Job	GEN	60	69	SIKH	Urban
11	Arts	MALE	Service	BC	36	43	HINDU	Rural
12	N.M	FEMALE	Agriculture	GEN	68	53	SIKH	Rural
13	Arts	MALE	Agriculture	BC	51	56	SIKH	Rural
14	Commerce	MALE	Agriculture	GEN	41	50	SIKH	Rural
15	Arts	FEMALE	Bussiness	GEN	47	40	HINDU	Urban
16	Arts	FEMALE	Govt. Job	GEN	63	62	SIKH	Urban
17	Arts	FEMALE	Bussiness	GEN	60	65	SIKH	Urban
18	Arts	FEMALE	Service	GEN	45	58	SIKH	Urban
19	Commerce	FEMALE	Agriculture	BC	59	54	SIKH	Rural
20	Arts	FEMALE	Agiculture	GEN	60	56	SIKH	Rural
21	Arts	FEMALE	Pvt. Job	GEN	51	45	SIKH	Urban
22	Arts	MALE	Agriculture	GEN	82	81	SIKH	Rural
23	Arts	FEMALE	Agriculture	GEN	75	55	SIKH	Rural
24	Arts	FEMALE	Govt. Job	GEN	65	60	HINDU	Rural
25	Arts	MALE	Govt Serv	GEN	47	56	HINDU	Urban
26	Arts	FEMALE	Agriculture	GEN	62	67	HINDU	Rural
27	Arts	MALE	Agriculture	GEN	58	63	SIKH	Urban
28	Vocationa	FEMALE	Agriculture	GEN	56	56	HINDU	Rural
29	Arts	MALE	EX ARMY P	GEN	50	62	HINDU	Urban
30	Arts	MALE	MAGISTRA	GEN	58	65	SIKH	Urban

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

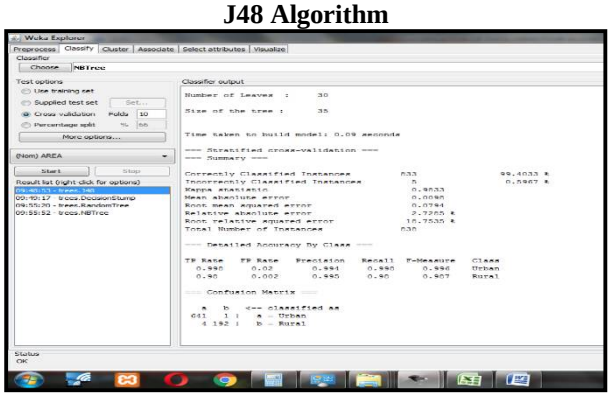
(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

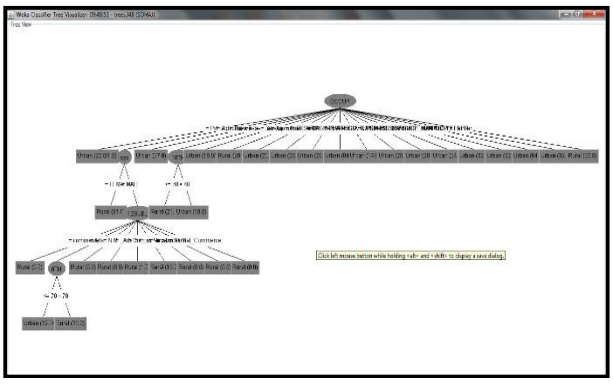
IX. RESULTS OF THE STUDY

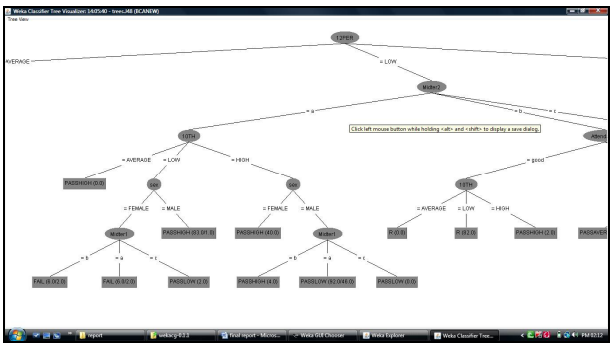
The results of the study have been discussed with the application of the following four algorithms:

J48 Algorithm



For J48 classification, 833 instances are correctly classified out of 838 instances which shows success rate of 99.4%. These results, when interpreted in the form of decision tree, help in identifying the weak students and those students, whose chances of failure in the examination are high. It suggests that we should concentrate more on those students in order to minimize the failure rate and drop out ratio so that the overall performance of the school can be improved. The decision tree can be shown as under:





The decision tree as shown above starts from the root node that is occupation of father/guardian. The occupation of father/guardian has been classified into five categories viz. Government Service, Private Service, Ex-Service Man, Business and Farming. The next sub-tree classifies the data with respect to the location of the students which is either Rural or Urban. The data is further classified with respect to Sex and 10th class percentage in the tree. The last sub-tree

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

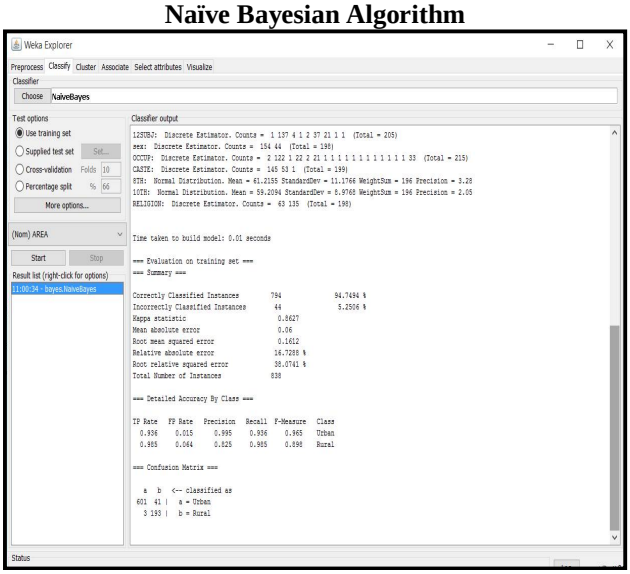
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

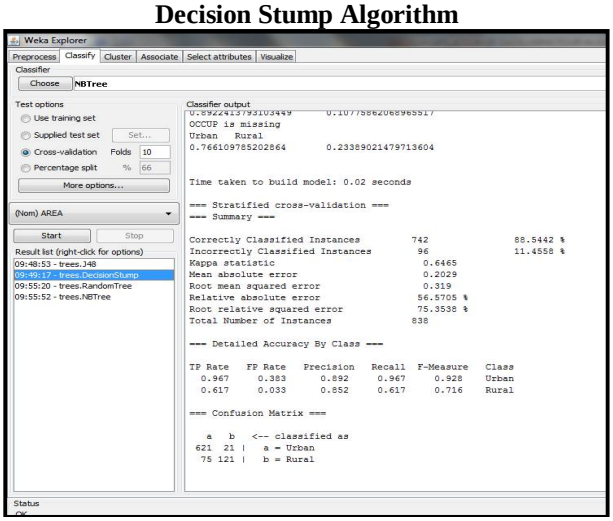
classifies the data on the basis of subjects taken by students in 12th class and it creates sub-tree of the percentage of 8th class students and it signifies those students who are in the risk zone of failure.

Naïve Bayesian Algorithm



Under Naïve Bayesian classification, 794 instances are correctly classified out of 838 instances which shows success rate of 94.7%.

Decision Stump Algorithm



Under Naïve Bayesian classification, 742 instances are correctly classified out of 838 instances which shows success rate of 88.5%.

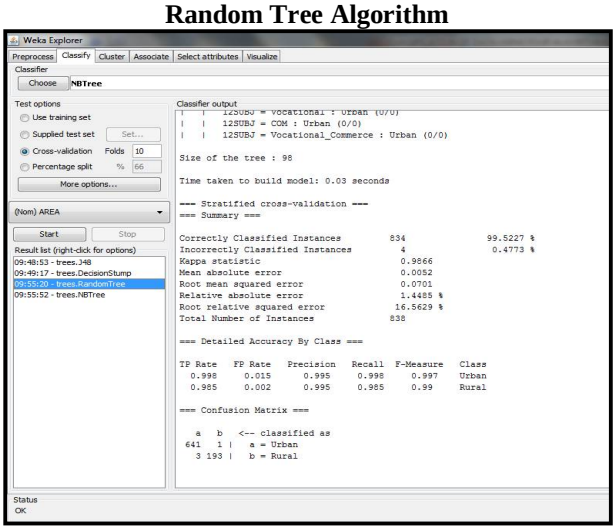
ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

Random Tree Algorithm



Under Naïve Bayesian classification, 834 instances are correctly classified out of 838 instances which shows success rate of 99.5%.

X. COMPARATIVE ANALYSIS OF FOUR ALGORITHMS

The following table highlights the comparative results obtained under four algorithms:

Algorithm	Correctly Classified Instances	Incorrectly Classified Instances	Mean Absolute Error	Root Mean squared Error	Relative Absolute Error	Root Relative Squared Error
J48	833	5	0.0098	0.0794	2.7265	18.7535
Naïve Bayesian	794	44	0.06	0.1612	16.7288	38.0741
Random Tree	834	4	0.0052	0.0701	1.4485	16.5629
Decision Stump	742	96	0.2029	0.319	56.5705	75.3538

The above table makes it amply clear that both the algorithms J48 and Random Tree have the same level of accuracy because they have 833 and 834 correctly classified instances out of 838 total instances. The other two algorithms are far away from the accuracy identification level in comparison to these algorithms. However, it is also visible from the application of the algorithms that Random Tree is unable to give the graphical presentation of the results in the form of Decision Tree. There is not much difference in the identification accuracy of J48 and Random Tree algorithms but because of the provision of graphical presentation of the results in the form of decision tree and sub-trees, it is recommended that J48 is the best option out the four studied alternatives for the prediction of Educational Data Mining in Senior Secondary Schools in the state of Punjab.

XI. CONCLUSION

The current education system does not involve any prediction about fail or pass percentage based on the performance. The system doesn't deal with dropouts. There is no efficient method to caution the student about the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

student about the deficiency in attendance. It doesn't identify the weak student and inform the teacher. Another common problem in larger number of students in class may feel lost in the crowd.

So from the above table it is very much clear that, although there is not much difference in the identification accuracy of J48 and Random Tree algorithms but because of the provision of graphical presentation of the results in the form of decision tree and sub-trees, it is recommended that J48 is the best option out the four studied alternatives for the prediction of Educational Data Mining in Senior Secondary Schools in the state of Punjab.

The results in case of J48 algorithm, when interpreted in the form of decision tree, helps in identifying the weak students and those students, whose chances of failure in the examination are high. It guides that we should concentrate more on those students in order to minimize the failure rate and drop out ratio so that the overall performance of the school can be improved.

XII.FUTURE SCOPE

Data mining is quite a vast field. It is currently implemented in many fields like in business and academics as we have used in our research work. Data mining can be further used in academics for sports analysis, prediction of interesting subjects, predictions of staff and faculty behavior etc. Moreover, it can be implemented with the help of some another tool and can also be implemented in WEKA with improved J48 algorithm. It can also be used in the field of agriculture where it has a huge potential.

REFERENCES

1. Anitha, S., Jasmine, B. and Deepalakshami, S.,"A Study of School Drop-Outs in Theni District: A Data Mining Analysis", International Journal on Engineering Research and Technology, Vol.2, Issue 6, pp. 328-333, 2013.
2. Bansode, J., "Mining Educational Data to Predict Students' Academic Performance", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 4, Issue1, pp. 1-5, 2016.
3. Baradwaj, B.K. and Pal, S., "Data Mining: A Prediction for Performance Improvement using Classification", International Journal of Computer Science and Information Security, Vol. 9, Issue 4, 2011.
4. Baradwaj, B.K. and Pal, S., "Mining Educational Data to Predict Students' Performance", International Journal of Advanced Computer Science and Applications, Vol. 2, Issue 6, 2011.
5. 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, Vol. 6, Issue 6, pp. 18-21, 2013.
6. Elakia, Gayathri, Aarthi and Naren, J., "Application of Data Mining in Educational Database for Predicting Behavioural Pattern of the Students", International Journal of Computer Science and Information Technologies, Vol. 5, Issue 3, pp. 4649-4652, 2014.
7. Hung, J.L., Hsu, Y.C. and Rice, K., "Integrating Data Mining in Program Evaluation of K-12 Online Education", International Forum of Educational Technology and Society, Vol. 15, Issue 3, pp. 27-41, 2012.
8. Jha, J. and Ragha, L., "Educational Data Mining Using Improved Apriori Algorithm". International Journal of Information and Computer Technology, Vol. 3, Issue 4, pp.411-418, 2013.
9. Kalpana, J. and Venkatalakshmi, K.,"Intellectual Performance Analysis of Students by using Data Mining Techniques", International Journal of Innovative Research in Science, Engineering and Technology, Vol.3, Issue 3, 1922-1929, 2014.
10. Kumar, D. A. and Radhika, V.,"A Survey on Predicting Student Performance", International Journal of Computer Science and Information Technologies, Vol.5, Issue 5, pp. 6147-6149, 2014.
11. Kumar, V. and Chadha, A., "An Empirical Study of the Applications Data Mining Techniques in Higher Education", International Journal of Advanced Computer Science and Applications, Vol. 2, Issue 3, pp. 80-84, 2011.
12. Kumar, N.V.A. and Uma, G.V., "Improving Academic Performance of Students by Applying Data Mining Technique", European Journal of Scientific Research, Vol. 34, Issue 4, pp. 526-534, 2009.
13. Manhaes, L.M.B., Cruz, S.M.S.D. and Zimbrao, G., "Evaluating Performance and Dropouts of Undergraduates using Educational Data Mining",Universidade Federal do Rio de Janerio, 2014.
14. Ma, Y., Liu, B., Wong, C.K., Yu, P.S. and Lee, S.M., "Targeting the Right Students using Data Mining", ACM Digital Library, 2000.
15. Nirmala, Devi.R., Deepa, R. and Kalaiarasi, P., "Mining Educational Data for Predicting Higher Secondary School Students' Grade Using ID3 Algorithm", International Journal of Engineering and Computer Science, Vol.4, Issue 1, pp. 10068-10071, 2015.
16. Padmapriya, A., "Prediction of Higher Education Admissibility Using Classification Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 11, pp. 330-336, 2012.
17. Pal, A.K. and Pal, S., "Analysis and Mining of Educational Data for Predicting the Performance of Students", International Journal of Electronics Communication and Computer Engineering, Vol.4, Issue 5, 2013.
18. Romero, C., Ventura, S., Espejo, P.G. and Hervas, C., "Data Mining Algorithms to Classify Students", Computer Science Department of Cordoba University, 2008.
19. Shymala, K. and Rajagopalan, S.P., "Data Mining Model for a better Higher Educational System", Information Technology Journal, Vol. 5, Issue 3, pp. 560-564, 2006.
20. Singh, M., Nagar, H. and Sant, A.,'Using Data Mining to Predict School Student Performance", International Journal of Advanced Research and Innovative Ideas in Education, Vol. 2, Issue 1, pp. 43-46, 2016.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 8, August 2016

21. Singh, M. and Sant, A., "Performance Analysis of Primary School Students using Data Mining Techniques", International journal of Research in Advent Technology, Vol. 2, Issue 6, pp. 100-105, 2014.
22. Sreenivasrao, V. and Yohannes, G., "Improving Academic Performance of Students of Defence University based on Data warehousing and Data Mining", Global Journal of Computer Science and Technology, Vol. 12, Issue 2, 2012.
23. Sumitha, R. and Vinothkumar, E.S., "Prediction of Student Outcomes using Data Mining Techniques", International Journal of Scientific Engineering and Applied Science, Vol.2, Issue, pp. 132-139, 2016.
24. Tair, M.A. and El-Halees, A.M., "Mining Educational Data to improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, Vol. 2, Issue 2, pp. 140-146, 2012.
25. Zahedifard, M., Attarzadeh, I., Pazhokhzadeh, H. and Malekzadeh, J., "Prediction of Students' Performance in High School by Data Mining Classification Techniques", International Academic Journal of Science and Engineering, Vol. 2, Issue 7, pp. 25-33, 2015.

BIOGRAPHY

Sohajbir Singh Ubha is a Research Scholar in the Computer Science and Engineering Department, University School of Engineering and Technology, Rayat and Bahra University, Mohali. He received Bachelorof Technology degree in 2014 from BBSBEC, Fatehgarh Sahib, Punjab, India. Hehas varied interests in the fields of Computing, Management and Literature. He has many research papers published in various national and international conference and journals.

Gaganpreet Kaur Bhalla is an Assistant Professor in the Computer Science and Engineering Department, University School of Engineering and Technology, Rayat and Bahra University, Mohali. She has 7 years of teaching experience. She has many research papers published in various national and international conference and journals.