DOI: 10.15415/jotitt.2016.42007

Analysis of Student's Data using Rapid Miner

SHEENA ANGRA[1], **SACHIN AHUJA**[2]

[1]Ph.D Scholar, Chitkara University, India

[2]Associate Director, CURIN, Chitkara University, India

E-mail: sheena.angra@chitkara.edu.in, sachin.ahuja@chitkara.edu.in

Published Online: December 28, 2016

Abstract*:* Data mining offers a new advance to data analysis using techniques based on machine learning, together with the conventional methods collectively known as educational data mining (EDM). Educational Data Mining has turned up as an interesting and useful research area for finding methods to improve quality of education and to identify various patterns in educational settings. It is useful in extracting information of students, teachers, courses, administrators from educational institutes such as schools/ colleges/universities and helps to suggest interesting learning experiences to various stakeholders. This paper focuses on the applications of data mining in the field of education and implementation of three widely used data mining techniques using Rapid Miner on the data collected through a survey.

Keywords: Educational Data Mining; Data Mining; EDM Objectives;Rapid Miner; EDM data and Stakeholders.

I. INTRODUCTION

Data Mining is an analytical process used to explore data (usually large amounts of data - typically business or market related) in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data. Data Mining has attracted many researchers and scientists to work in this area due to presence of large amount of data which is available in various formats like records,texts,files,sounds,images and many other formats[1].The data which is collected from various applications and repositories requires various data mining techniques to extract useful and novel information from them in order to give accurate results and for the purpose of future prediction. There are various steps which are used to extract information from data. The main purpose of using data mining techniques is to use various algorithms and methods to extract and discover some patterns. Data mining is used in various areas such as data visualization, statistics, machine

Journal on Today's Ideas – Tomorrow's Technologies, Vol. 4, No. 2, December 2016 pp. 109-117

CHITKARA UNIVERSITY

learning, database systems and information retrieval[1].

Data mining covers multitude of application areas including Businesses, banking, Insurance, Scientific, Medical, Weather forecasting etc. which involves huge amount of data for processing. Educational Data Mining is an upcoming area in the field of data mining as educational settings are experiencing the phenomenon of data explosion.Computerized systems collect data about a multitude of everyday transactions in academic institution. Data is related to students' attendance, performance, historical data, demographic data, admissions, accounts, internet usage and many more. The goal of educational data mining is to apply data mining techniques on the available data in terms of educational context and come up with a model using data mining techniques that help in decision making.

II. EDM OBJECTIVES

EDM involves different groups of users or participants such as students, teachers, Course developers, University/Institute Management and policy makers at Governing Bodies. Different groups look at educational information from different angles according to their own mission, vision, and objectives for using data mining. For example, knowledge discovered by EDM algorithms can be used to not only help teachers to manage their classes, understand their students' learning processes and reflect on their own teaching methods, but also to support a learner's reflections on the situation and provide feedback to learners [14]. The Data mining process can be applied in large number of academic activities that can help in decision making. These task can be fulfilled using different data mining techniques on the data collected from various educational settings. Each task has its own specific requirements for the input data and requires thorough knowledge of the underlying data mining concept/algorithm. The activities/task are listed below:-

i. Allotment of resources,
ii. Assessment of the student's learning performance,
iii. Course adaptation based on the student learning behavior,
iv. Learning recommendations based on the student learning behavior,
v. Evaluation of learning material and educational web based courses,
vi. Feedback to both teacher and students in different courses,
vii. Detection of typical students' learning behaviors,
viii. Prediction of students enrollment in particular course,
ix. Recommendations on designing a new course,
x. Recommendation on teaching/Learning methodology for specific courses

III. RAPID MINER

Rapid miner is a data mining analytics tool which is used to analyse data and support various techniques of data mining.It is used for industrial applications ,research,training,application development and education[7].It contains about 100 learning schemes for clustering,classification and regression analysis[15]. It supports around 22 file formats such as .xls,.csv and so on[16].In this information can be imported from various databases for analysis and prediction purposes[7].

IV. EDM DATA AND STAKEHOLDERS

Educational data is collected from various sources such as colleges,universities,schools and by having an eye on the online activities of students and instructors.The type of data which is needed for analysis is student's response,hints requested,wrong answers,correct answers,responses to surveys and questionnaires etc. [3].We have two types of data- offline and online data.

i) *Offline Data*: Offline data is generated from both modern classrooms as well as traditional classrooms.We can get behavioral data, educator's information,student's attendance,course information, student's information[4].

ii) *Online Data*: Online data is generated from stakeholders which are geographically separated,web based education[5],distance education[6],online forums and social networking sites.Online data is also generated from Intelligent Tutoring Systems(ITS) and Learning Management Systems(LMS) [2].

EDM involves different groups of users or participants. Different groups look at educational information from different angles according to their own mission, vision and objectives for using data mining. The stakeholders of EDM can be broadly classified into following categories:-

i) *Educators*: Educators understands the methods and the learning process that can be used to improve the teaching methods.Educators use various EDM applications to determine the best methods to deliver information related to a course and to sructure and organise the syllabus.They also determine the indicators which shows student's engagement and satisfaction of the course[8].

ii) *Learners*: EDM helps to inform parents of students about their child's progress in a particular course[9].It is very important to make learners aware of the online environment.The biggest challenge is to understand these groups and generate some models based on it[10].

iii) *Administrators*: As institutions are considered to be responsible for the

success of students so it is very important to administer EDM applications in educational settings[11]. Sometimes it is very difficult to gather information in an efficient manner in order to administer the applications[11].

iv) *Researchers*: The main focus of the researchers is to evaluate and develop data mining techniques for accuracy and effectiveness.They also focus on different ways to improve student's performance.

V. ANALYSIS OF DATA

Our purpose is to analyse student's data in order to find the factors which are impacting their performance.In order to achieve this we have done an online survey based on the questionnaire which included around 40 factors such as father's income,annual income of parents, father's income,e- mail id,contact number and so on. Out of them some factor such as contact number,e-mail id are irrelevant as they don't effect student's performance and some factors such as father's occupation, 10 percentage have turned out to be influential.

We have applied some techniques on the collected data on Rapid Miner which is a data mining analytics tool.The techniques which we have applied are WJ- 48 algorithm,K-means clustering algorithm and linear regression.We have chosen these techniques as they are widely used techniques and produce better results as compared to other algorithms.

VI. IMPLEMENTATION AND RESULTS

The results of the above mentioned techniques are described below:

i) WJ-48 algorithm:It is a decision tree algorithm which when applied to the data generate some rules based on the label. Decision tree is used to represent a tree like structure in the form of IF-THEN statements[1].Figure 1 shows the implementation of this algorithm and Figure 2 produces the desired results.

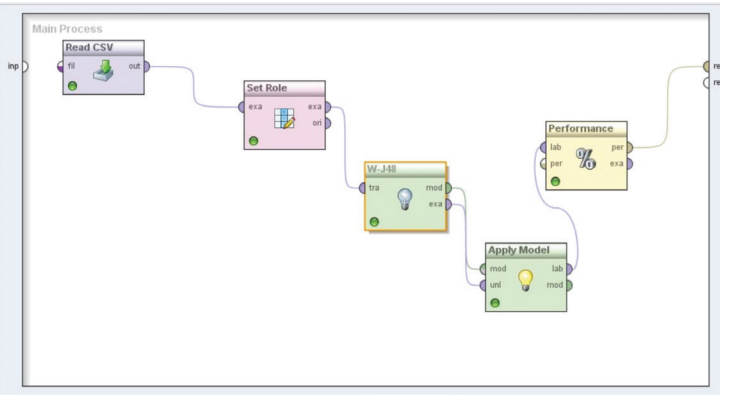

Figure 1: Implementing WJ-48 Algorithm

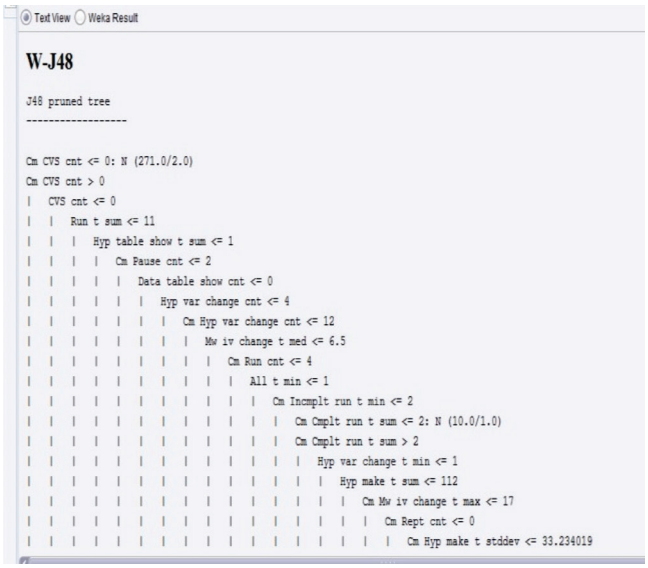

Figure 2: Tree generated after implementing WJ-48 Algorithm

ii) K-means Clustering Algorithm: Clustering is a process which groups similar objects into one cluster (12). It is also termed as unsupervised classification. It means that while classifying data objects clustering does not depend on training tuples and predefined classes [13].K-means is a clustering algorithm which works on Euclidian distance and is used to group similar objects in one cluster. It defines K centers for K clusters that is; one center is defined for each cluster. The main objective of using K-means is to reduce intra cluster variance and to maximize the performance. Figure 3 shows the implementation of K-means algorithm, figure 4 shows the distribution of different items in different clusters and figure 5 describes the calculation of performance vector for all the 4 clusters. Performance of a cluster depends on the number of items per cluster.

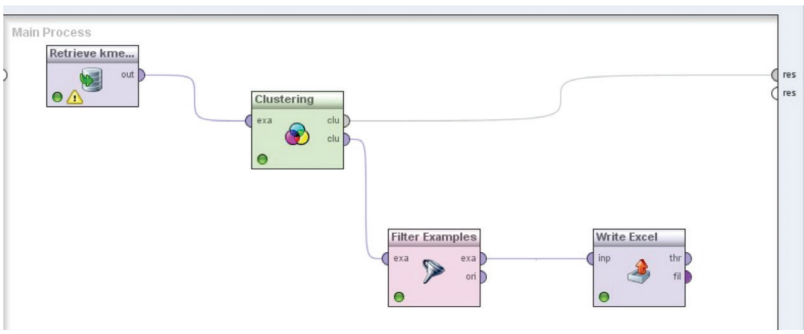

Figure 3:Implementing K-means Clustering algorithm on collected data

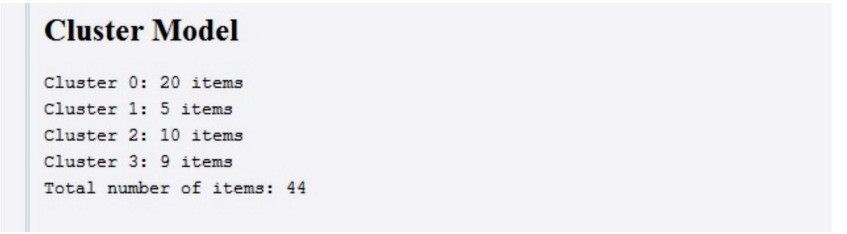

Figure 4: Number of items per cluster

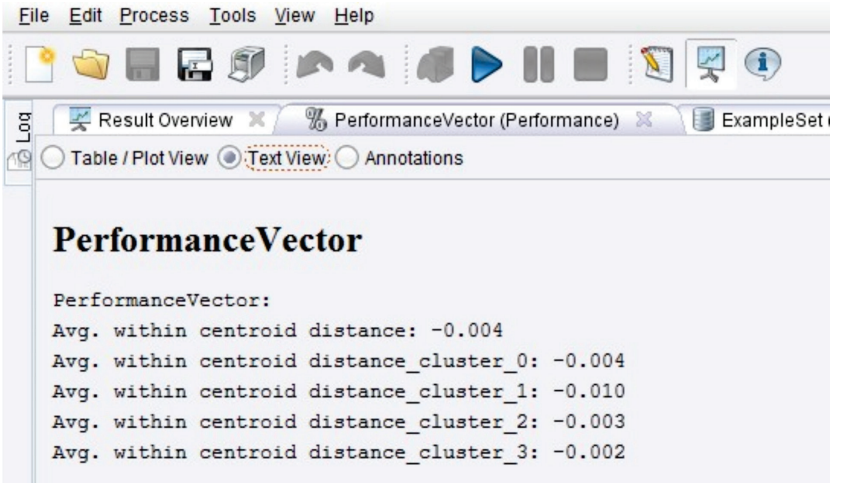

Figure 5: Performance vector calculation

iii) Linear Regression: Linear regression is a predictive analysis technique which is used to find the relationship between one dependent variable and other independent variable. Figure 6 shows the implementation of linear regression on the data and figure 7 shows the root mean squared error in the data which estimates the deviation of data from the expected mean.

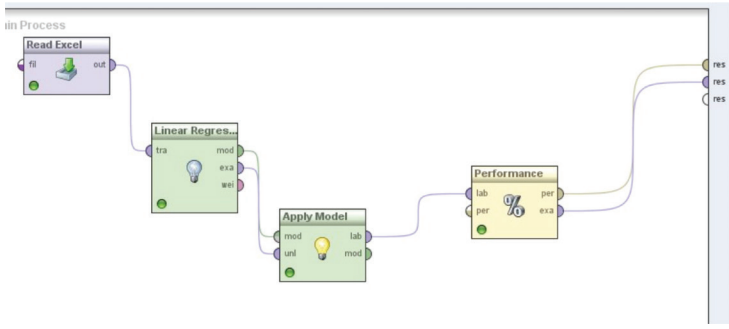

Figure 6: Implementing Linear Regression

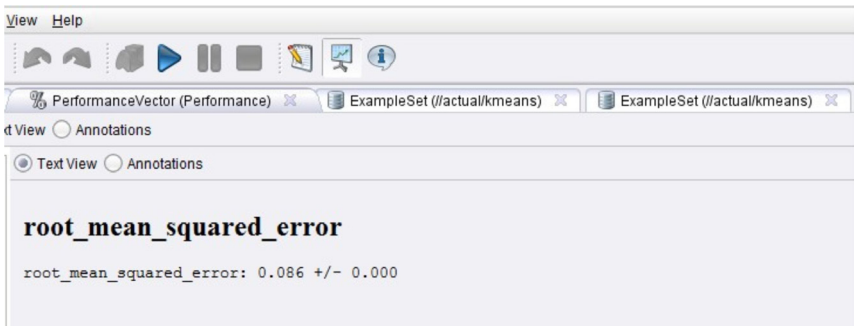

Figure 7: Calculation of Root mean squared error

Our main purpose is to predict the performance of students of next semester based on the results of previous semester.After analysis of data using the mentioned 3 techniques we are also working on the rules which will be generated after applying these techniques.In order to accomplish this we have collected some academic and demographic data of large number of students so that these techniques can be tested on wide range of students also.We have applied decision tree algorithm on this data and we have found that there are some factors that are influencing the performance of students and these factors include 10^{th} percentage,12^{th} percentage,attendance,father's income and 3^{rd} sem SGPA.So we have decided to design a java framework using Netbeans IDE 8.1 to accomplish this task which includes only those factors which influences the student's performance so that only these factors should be taken into consideration while enrolling or judging a student.Figure 8 gives the choice to select a particular algorithm and since we have worked only on decision tree for large amount of data so figure 9 describes the generated rules.

Figure 8: Choice to choose an algorithm

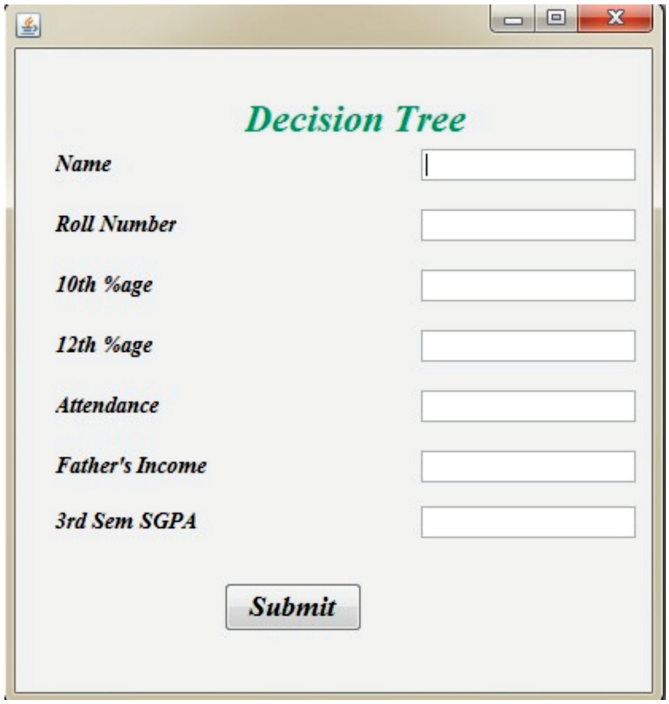

Figure 9: Rules generated after applying Decision Tree Algorithm

VII. CONCLUSION

This paper presentes implementation of 3 data mining techniques on the collected data. EDM is emerging as a great research area in the field of education.It uses various techniques to understand student's behavior and performance.It helps to predict grades of students of one class by analysing the grades of previous classes.It analyses student's offline and online activities and also suggests some methods to insructors/teachers to organise the course content according to the student's needs and performance. The full integration of data mining in the educational environment is not yet witnessed but the future line of research in this area can be a full operational implementation of data mining in educational environment for all the stakeholders.Different techniques works on different parameters such as performance vector and root mean square error.

VIII. FUTURE SCOPE

A software having Java framework can be designed in order to provide facility to the institution to judge student's based on only some influential factors rather than considering all the demographic and academic factors.

REFERENCES

1. Sachin R., Vijay M., A survey and Future Vision of Data mining in Educational Field, In Second International Conference on Advanced Computing & Communication Technologies, 2012.
2. Jindal and Dutta Borah, –A Survey On Educational Data Mining and ResearchTrends, In International Journal Of Database Management Systems,2013, Vol.5, No.3. http://dx.doi.org/10.5121/ijdms.2013.5304.
3. Prakash, Hanumanthappa & Kavitha, Big Data in Educational Data Mining and Learning Analytics, InInternational Journal of Innovative Research in Computer and Communication Engineering,2014, Vol. 2. http://dx.doi.org/10.15680/IJIRCCE.2014.0212044.
4. Romero, C., and Ventura S., Data Mining in Education, WIREsData Mining and Know.Dis.,2013, Vol.3,pp.12-27. http://dx.doi.org/10.1002/widm.1075.
5. Ha, S., Bae, S., and Park, S , Web mining for distance education, In Proc.Int. Conf. On Management of Innovation and Technology, IEEE.,2000, pp.715-719. http://dx.doi.org/10.1109/EmbeddedCom-ScalCom.2009.98.
6. Romero, C., and Ventura, S., Educational Data Mining : A survey from 1995 to 005, Expert Systems with Applications.,2007,Vol. 33, pp.135-146. http://dx.doi.org/10.1016/j.eswa.2006.04.005.
7. Rangra and Bansal., Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering 4(6), June - 2014, pp. 216-223.
8. U.S. Department of Education, Office of Educational Technology. –Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief (PDF). Retrieved 30 March 2014.
9. Assessing the Economic Impact of Copyright Reform in the Area of Technology-Enhanced Learning. Retrieved 6 April 2014.
10. Azarnoush, Bahareh, et al., Toward a Framework forLearner Segmentation. JEDM-Journal of Educational Data Mining 5.2, 2013: 102-126.
11. Huebner, Richard A., A survey of educational datamining research (PDF). Research in Higher Education Journal.Retrieved 30 March 2014.
12. Jain, A. K., Murty, M. N., & Flynn, P. J., Data clustering: A review, ACM Computing Surveys, 31(3), 1999, (pp. 264–323). http://dx.doi.org/10.1145/331499.331504.
13. Yedla,Pathakota & Srinivasa, Enhancing K-Means Clustering Algorithm with Improved Initial Center, International Journal of Computer Science and Information Technologies,2010, Vol. 1 (2) ,121-125.
14. Romero, C. & Ventura, S., Educational Data Mining:A Review of the State-of-the-Art, Systems,Man,and Cybernetics— Part C:Applications and Reviews, IEEE Transactions on 40.6(2010):601-618.
15. Mierswa, I., Wurst, M., Klinkenberg, R., Scholz, M., Euler,T., YALE: Rapid Prototyping for Complex Data Mining tasks, in Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining(KDD-06), pp. 935-940, 2006. http://dx.doi.org/10.1145/1150402.1150531.
16. Ralf Mikut and Markus Reischl Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, pages 431–443, September/October 2011.