ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 1, pp: (123-129), Month: January - March 2015, Available at: www.researchpublish.com

A Review on Role of Data Mining Techniques in Enhancing Educational Data to Analyze Student's Performance

[1]Dr. N. Preethi, [2]Deepak Goswami

[1]Assistant Professor, Jain University, Bangalore, India
[2]PG Student Jain University, Bangalore, India

Abstract: **Educational Data Mining (EDM) is an rising ground search data in educational background by relating different Data Mining (DM) techniques/tools. It gives basic knowledge of teaching and learning method for useful education development. Educational data mining (also referred to as "EDM") is clear as the area of scientific query centered around the growth of methods for making discoveries within the single kinds of data that come from educational background, and using those technique to better identify students and the background which they be trained in. This boost in rapidity and feasibility has had the advantage of making imitation much more possible. This paper addresses the purpose of data mining in educational society to pull out useful information from the vast data sets and providing logical tool to analysis and use this information for decision making processes by taking real life examples.**

Keywords: **Educational Data Mining (EDM); Classification; Knowledge Discovery in Database (KDD); ID3 Algorithm. Higher education,, Data mining, Knowledge discover, Classification, Association rule, Prediction.**

I. INTRODUCTION

Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information. [1] The data can be collected from various educational institutes that reside in their databases. The data can be personal or academic which can be used to understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems , to improve curriculums and many other benefits.[1][2] Educational data mining uses many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others.[3] Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. Educational Data Mining (EDM) is an emerging field exploring data in educational context by applying different Data Mining (DM) techniques/tools. EDM inherits properties from areas like Learning Analytics, Psychometrics, Artificial Intelligence, Information Technology, Machine learning, Statics, Database Management System, Computing and Data Mining. It can be considered as interdisciplinary research field which provides intrinsic knowledge of teaching and learning process for effective education [4].

The main objective of this paper is to use data mining methodologies to study students‟ performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this paper the classification task is used to evaluate student's performance and as there are many approaches that are used for data classification, the decision tree(J48) and Bayesian(NaiveBayes) classification methods is used here. Information's like Gender, Time duration between posts(in min) and duration between posting and replies were collected from the online student's Examination, to predict the performance, overall students knowledge, skill and disposition at the end of the online exam . The results are based on the time taken by the students to give answers and time duration between the replies.

II. DATA MINING DEFINITION AND TECHNIQUES

Data mining methods like prediction, clustering and relationship mining are mostly used in the field of marketing, agriculture and finance etc. These methods can be efficiently applied on educational data . Data mining having many type of techniques Like cluastering, classification, neural network etc but in this paper we are consider only two techniques Clustering techniques. Classification techniques, Predication, Association rule, Neural networks, Decision Trees, Nearest Neighbor Method.

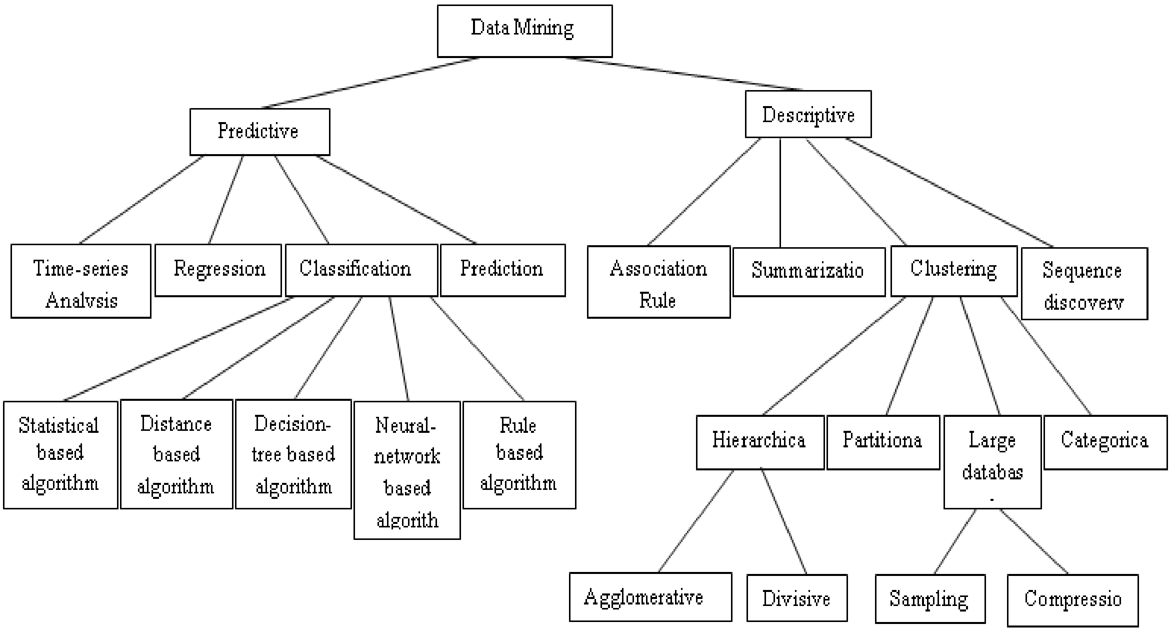


Figure: 1. Data mining techniques [5]

A. Classification:

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier.

B. Clustering:

Clustering can be said as identification of similar classes of objects. By using clustering techniques we can further identify dense and sparse regions in object space and can discover overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes of object but it becomes costly so clustering can be used as preprocessing approach for attribute subset selection and classification.

C. Prediction:

Regression technique can be adapted for predication. Regression analysis can be used to model the relationship between one or more independent variables and dependent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. Unfortunately, many real-world problems are not simply prediction. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to forecast future values. The same model types can often be used for both regression and classification. For example, the CART (Classification and Regression Trees) decision tree algorithm can be used to build both classification trees (to classify categorical response variables) and regression trees (to forecast continuous response variables). Neural networks too can create both classification and regression models.

D. Association rule:

Association and correlation is usually to find frequent item set findings among large data sets. This type of finding helps businesses to make certain decisions, such as catalogue design, cross marketing and customer shopping behavior analysis. Association Rule algorithms need to be able to generate rules with confidence values less than one. However the number of possible Association Rules for a given dataset is generally very large and a high proportion of the rules are usually of little (if any) value.

E. Neural networks:

Neural network is a set of connected input/output units and each connection has a weight present with it. During the learning phase, network learns by adjusting weights so as to be able to predict the correct class labels of the input tuples. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. These are well suited for continuous valued inputs and outputs. Neural networks are best at identifying patterns or trends in data and well suited for prediction or forecasting needs.

F. Decision Trees:

Decision tree is tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID).

G. Nearest Neighbor Method:

A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k is greater than or equal to 1). Sometimes called the k-nearest neighbor technique.

III. EDUCATIONAL DATA MINING

Educational data mining is emerging as a research area with a suite of computational and psychological methods and research approaches for understanding how students learn. New computer-supported interactive learning methods and tools—intelligent tutoring systems, simulations, games—have opened up opportunities to collect and analyze student data, to discover patterns and trends in those data, and to make new discoveries and test hypotheses about how students learn. Just as with early efforts to understand online behaviors, early efforts at educational data mining involved mining website log data [6], but now more integrated, instrumented, and sophisticated online learning systems provide more kinds of data. Educational data mining generally emphasizes reducing learning into small components that can be analyzed and then influenced by software that adapts to the student [7]. An important and unique feature of educational data is that they are hierarchical. Data at the keystroke level, the answer level, the session level, the student level, the classroom level, the teacher level, and the school level are nested inside one another [8][9] . Other important features are time, sequence, and context. Time is important to capture data, such as length of practice sessions or time to learn.

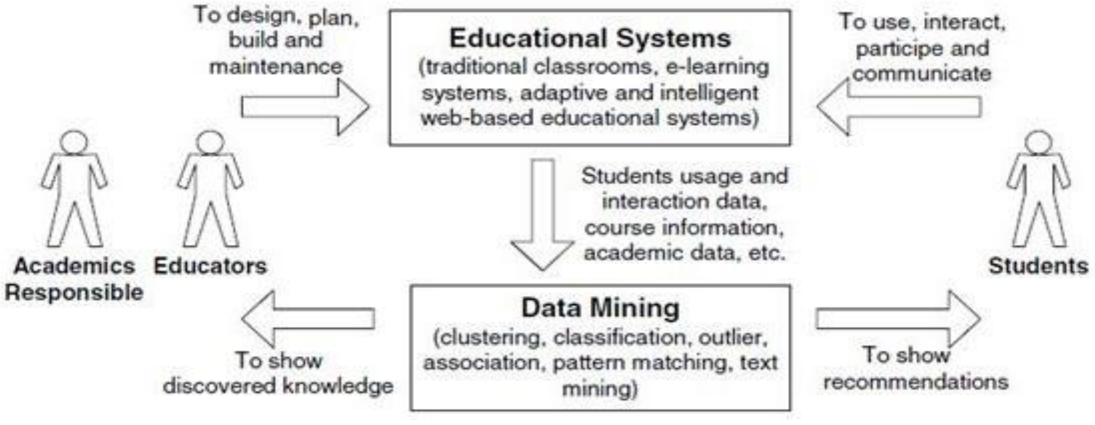
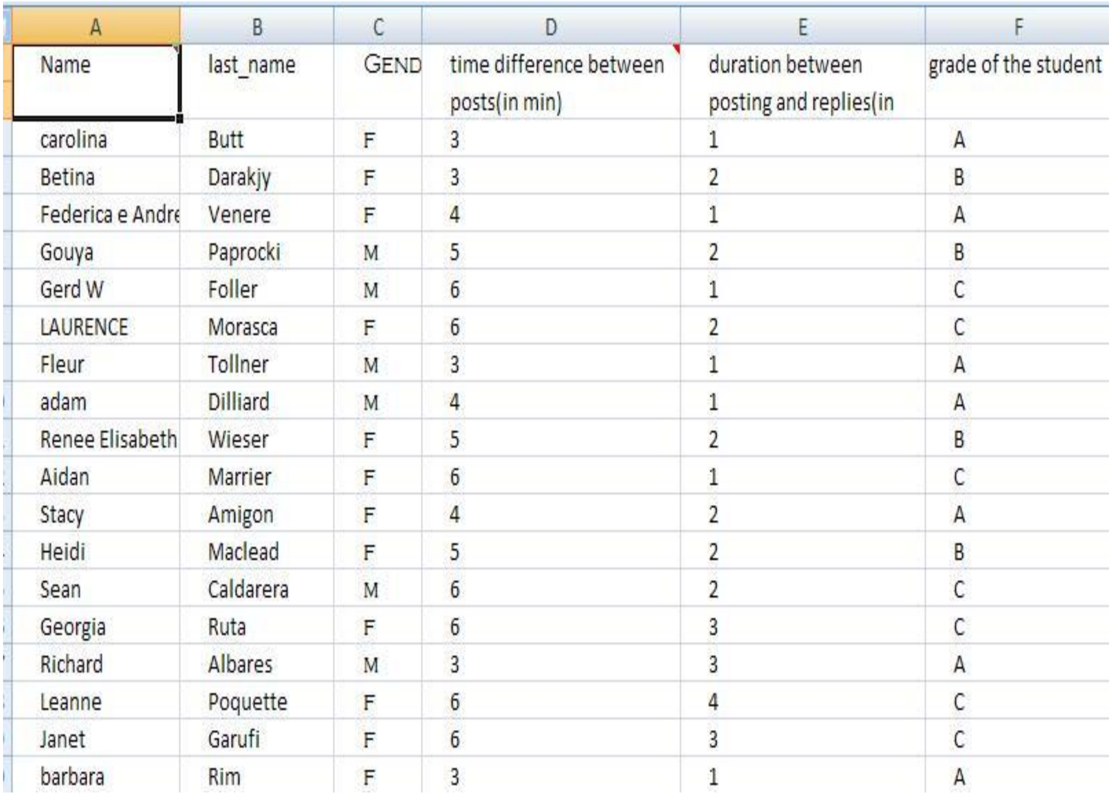

To design, plan, build and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participe and communicate
Academics Responsible
Educators
Students usage and interaction data, course information, academic data, etc.
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure: 2. Data mining applications in the education sector [2]

IV. RESULTS AND DISCUSSION

The data set of 100 students used in this study was obtained from online examination. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This applications of Data Mining in Educational society to extract useful information from the huge data sets and providing WEKA as an analytical tool to view and use this information for decision making processes by taking real life examples. front end design using jsp and servlet and backend design using mongoDB and for the analysis we used weka jar file.

Table I. Sample data set collected from online examinations.



A	B	C	D	E	F
Name	last_name	GEND	time difference between posts(in min)	duration between posting and replies(in	grade of the student
carolina	Butt	F	3	1	A
Betina	Darakjy	F	3	2	B
Federica e Andre	Venere	F	4	1	A
Gouya	Paprocki	M	5	2	B
Gerd W	Foller	M	6	1	C
LAURENCE	Morasca	F	6	2	C
Fleur	Tollner	M	3	1	A
adam	Dilliard	M	4	1	A
Renee Elisabeth	Wieser	F	5	2	B
Aidan	Marrier	F	6	1	C
Stacy	Amigon	F	4	2	A
Heidi	Maclead	F	5	2	B
Sean	Caldarera	M	6	2	C
Georgia	Ruta	F	6	3	C
Richard	Albares	M	3	3	A
Leanne	Poquette	F	6	4	C
Janet	Garufi	F	6	3	C
barbara	Rim	F	3	1	A

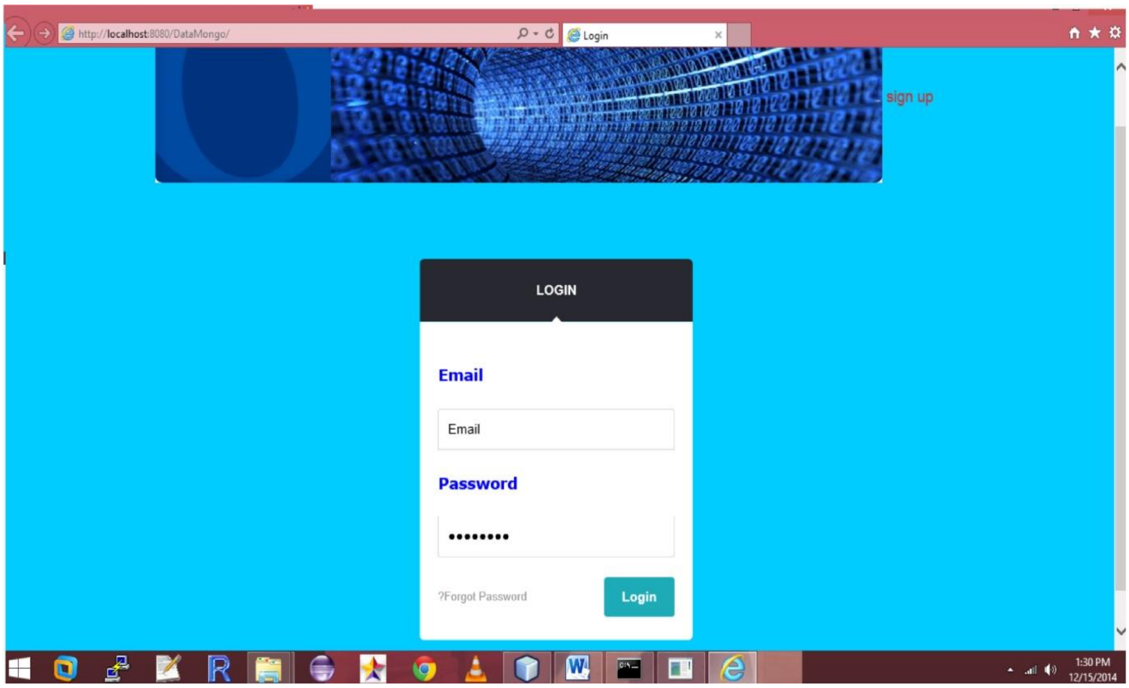


Figure: 3. Login page to start a session by a valid user using their email id and password

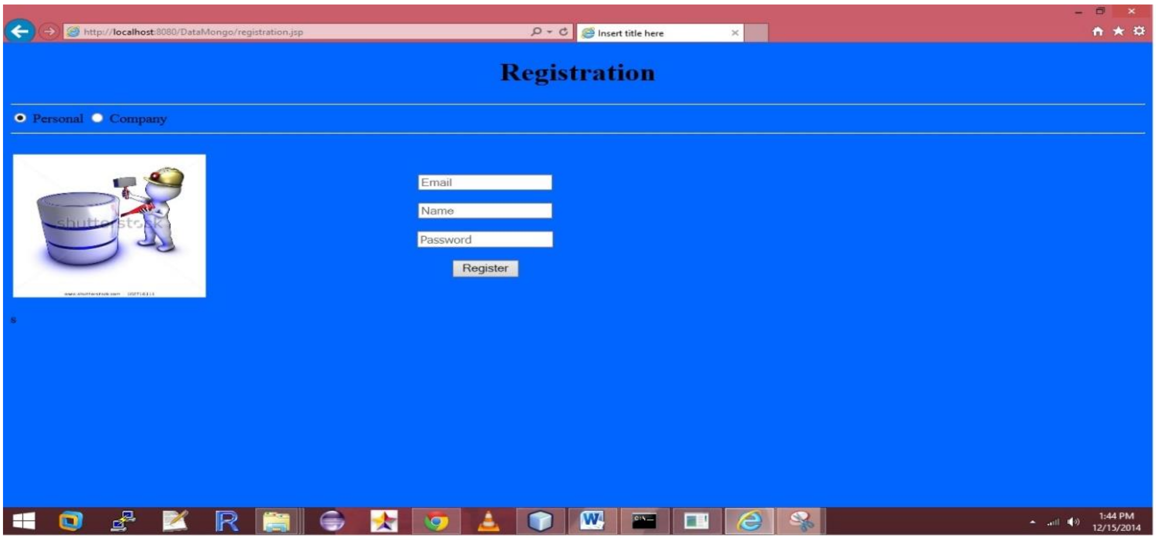
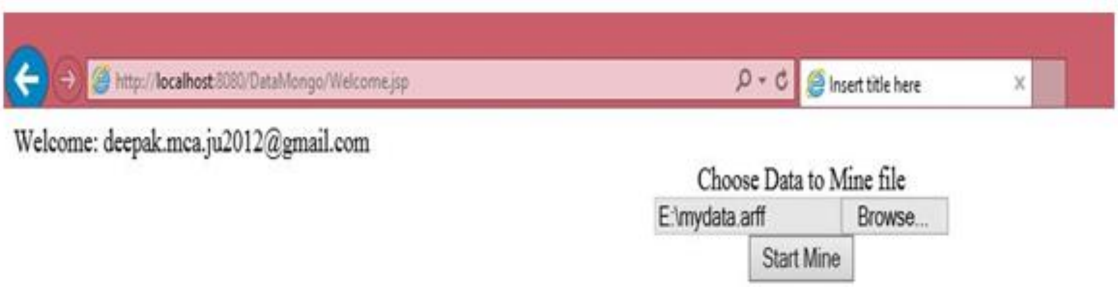


Figure: 4. this is a registration page to create a new user

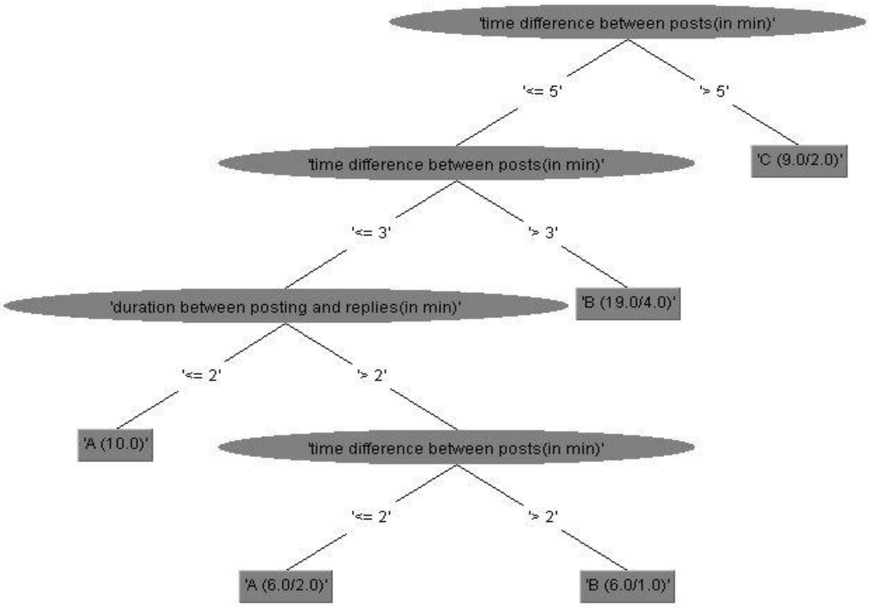


Welcome: deepak.mca.ju2012@gmail.com

Choose Data to Mine file

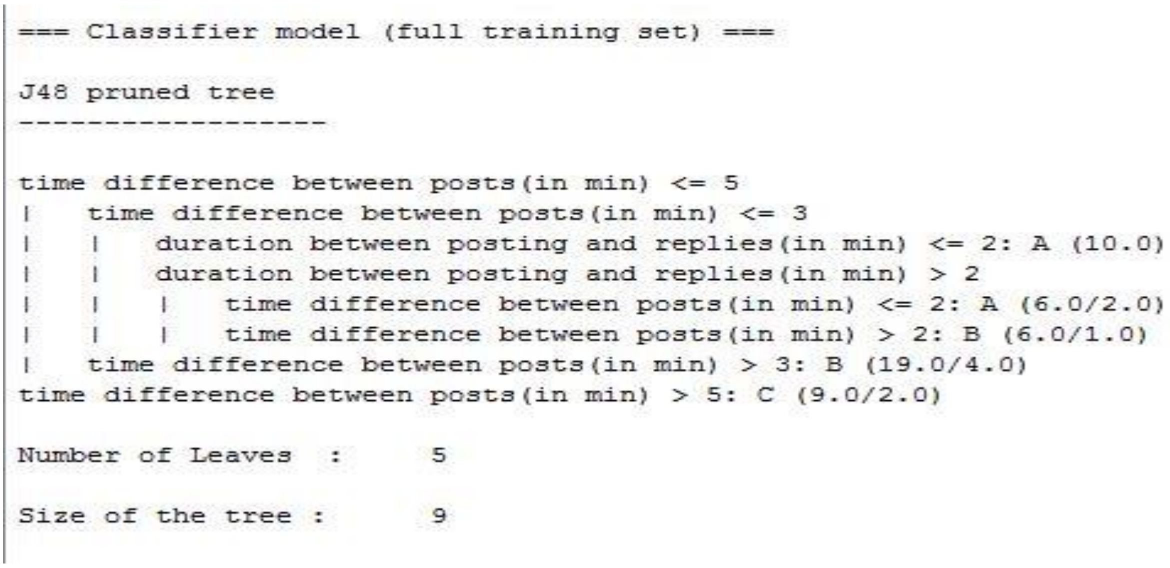
E:\mydata.arff

Figure: 5. for selecting a file for data mining purpose in .arff format only.



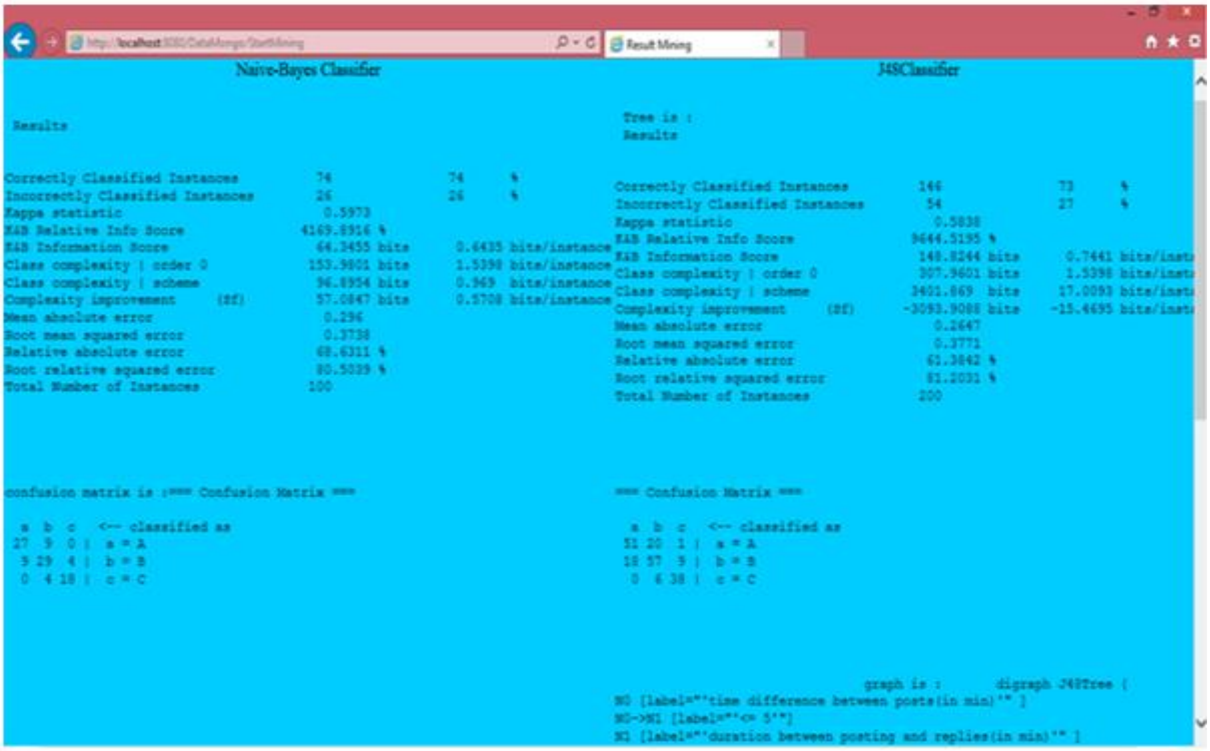
'time difference between posts(in min)'
'<= 5'
'> 5'
'C (9.0/2.0)'
'time difference between posts(in min)'
'<= 3'
'> 3'
'B (19.0/4.0)'
'duration between posting and replies(in min)'
'<= 2'
'> 2'
'A (10.0)'
'time difference between posts(in min)'
'<= 2'
'> 2'
'A (6.0/2.0)'
'B (6.0/1.0)'

Figure: 6. Decision tree obtained by using j48 Tree classifier



```
=== Classifier model (full training set) ===

J48 pruned tree
------------------

time difference between posts(in min) <= 5
|   time difference between posts(in min) <= 3
|   |   duration between posting and replies(in min) <= 2: A (10.0)
|   |   duration between posting and replies(in min) > 2
|   |   |   time difference between posts(in min) <= 2: A (6.0/2.0)
|   |   |   time difference between posts(in min) > 2: B (6.0/1.0)
|   time difference between posts(in min) > 3: B (19.0/4.0)
time difference between posts(in min) > 5: C (9.0/2.0)

Number of Leaves  :     5

Size of the tree :      9
```

Figure: 7. Pruned tree of datasets



```
Results


Correctly Classified Instances          74               74      %
Incorrectly Classified Instances        26               26      %
Kappa statistic                          0.5973
KAB Relative Info Score               4169.8916 %
KAB Information Score                   64.3455 bits      0.6435 bits/instance
Class complexity | order 0             153.9801 bits      1.5398 bits/instance
Class complexity | scheme               96.8954 bits      0.969  bits/instance
Complexity improvement     (Sf)         57.0847 bits      0.5708 bits/instance
Mean absolute error                      0.296
Root mean squared error                  0.3738
Relative absolute error                 68.6311 %
Root relative squared error             80.5039 %
Total Number of Instances              100


confusion matrix is :=== Confusion Matrix ===

  a  b  c   <-- classified as
 27  9  0 |  a = A
  5 29  4 |  b = B
  0  4 18 |  c = C
```

```
Tree is :
Results


Correctly Classified Instances         146               73      %
Incorrectly Classified Instances        54               27      %
Kappa statistic                          0.5838
KAB Relative Info Score               9644.5195 %
KAB Information Score                  148.8244 bits      0.7441 bits/insta
Class complexity | order 0             307.9601 bits      1.5398 bits/insta
Class complexity | scheme             3401.869  bits     17.0093 bits/insta
Complexity improvement     (Sf)      -3093.9088 bits    -15.4695 bits/insta
Mean absolute error                      0.2647
Root mean squared error                  0.3771
Relative absolute error                 61.3842 %
Root relative squared error             81.2031 %
Total Number of Instances              200


=== Confusion Matrix ===

  a  b  c   <-- classified as
 51 20  1 |  a = A
 18 57  5 |  b = B
  0  6 38 |  c = C


                        graph is :       digraph J48Tree {
N0 [label="'time difference between posts(in min)'" ]
N0->N1 [label="'<= 5'"]
N1 [label="'duration between posting and replies(in min)'" ]
```

Figure: 8. The results obtained from the dataset using naïve based classifier and j48 tree (Decision tree)

V. CONCLUSION

Data mining is a broad area that integrates with several fields including machine learning, statistics, pattern recognition, artificial intelligence, and to analysis of large volumes of data etc. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of retainment of faculties in the university, to find the gap between the number of candidates applied for the post, number of applicants responded, number of applicants appeared, selected and finally joined. This work focuses on research trends in Offline, Online and Uncertain data, useful data sources, links etc in an educational context. Different colleges/institutions affiliated to the same University should adopt a single model for academic planning to strengthen the utilization of existing resources. Lastly this work can further be improved for designing Knowledge

Discovery based Decision Support System (KDDS) which will capable of giving right decision for research in Science & Technology based on the demand of the society. “Data mining in Educational Institution” Web Application is help to the present Education System. The various data mining techniques are discussed which can support education system via generating strategic information. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, to find the gap between the numbers of candidates applied for the post, number of applicants responded, number of applicants appeared.

REFERENCES

[1] C. Romero, S. Ventura, E. Garcia, "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008.

[2] C. Romero, S. Ventura "Educational data mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007.

[3] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, “Data Mining Model for Higher Education System”, Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010.

[4] Romero, C., and Ventura S.(2013), “ Data Mining in Education”. WIREs Data Mining and Know.Dis.Vol.3, pp.12-27.

[5] S. Anupama Kumar and M. N. Vijayalakshmi” Relevance of Data Mining Techniques in Edification Sector” International Journal of Machine Learning and Computing, Vol. 3, No. 1, February 2013.

[6] Baker, R. S. J. D., and K. Yacef. 2009. “The State of Educational Data Mining in 2009: A Review and Future Visions.” Journal of Educational Data Mining 1 (1): 3–17.

[7] Siemens, G., and R. S. J. d. Baker. 2012. “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration.” In Proceedings of LAK12: 2nd International Conference on Learning Analytics & Knowledge, New York, NY: Association for Computing Machinery, 252–254.

[8] Baker, R. S. J. d., S. M. Gowda, and A. T. Corbett. 2011. “Automatically Detecting a Student’s Preparation for Future Learning: Help Use Is Key.” In Proceedings of the 4th International Conference on Educational Data Mining, edited by M. Pechenizkiy, T. Calders, C. Conati, S. Ventura, C. Romero, and J. Stamper, 179–188.

[9] Romero C. R. and S. Ventura. 2010. “Educational Data Mining: A Review of the State of the Art.” IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews 40 (6): 601–618.

[10] Witten, I. H. and Frank, E., Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition, Morgan Kaufman Publishers, San Francisco, 2005, p.5.