Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 3, March 2017, pg.196 – 203

An Approach To Suggest Company Specific Placement Opportunities Using Data Mining Techniques

Revathy S
Computer Science and Engineering Rajalakshmi Engineering College Chennai, India
revathyjagan1996@gmail.com

Roopika G
Computer Science and Engineering Rajalakshmi Engineering College Chennai, India
roopikaganesh1996@gmail.com

Rishitha R
Computer Science and Engineering Rajalakshmi Engineering College Chennai, India
rishitharaju5@gmail.com

Revathy P
Computer Science and Engineering Rajalakshmi Engineering College Chennai, India
revathy.p@rajalakshmi.edu.in

Abstract— Educational Data Mining (EDM) is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational organization and by using methods such as data mining for better understanding of students preparation for placement. This paper provides a framework for statistical experiment to identify the number of students those who are likely to be placed from a large database of all students under a particular department of a college containing their academic record. Data mining has become very important technique where we can categorize the students according to their qualifications. A variety of important parameters for measuring student's performance including academic performance, technical skills in various domains, programming skills, quantitative and reasoning skills are considered to capture the desirability and ability of a student for placement in their domain of strength. With the simulation and analysis results, the system also determines the chances of placement for a student based on their logical reasoning, quantitative and programming skills. Using these analysis, classification is done to predict how and where the student can improve in regard to placements. To predict the company name where a student is likely to be placed, the C5.0 algorithm is used for classification. The classifier is then applied to predict a test data set to test for accuracy.

Keywords: Classification, Placement, Data mining

I. INTRODUCTION

Data mining, which deals with the technique of extracting useful information from a dataset through analysis, has recently developed as an emerging discipline, called Educational Data Mining or EDM. These Educational Data mining techniques explore the hidden knowledge about students from their resumes. This helps the student to identify their domain of strength and the chances of getting placed in a company through that domain. This may help the student to improve his/her knowledge and the possibilities of getting hired.

The possible stages in EDM are exactly similar to those involved in Data mining such as:

- Data Collection.
- Data Preprocessing.
- Data Mining.
- Pattern Extraction.
- Knowledge Discovery

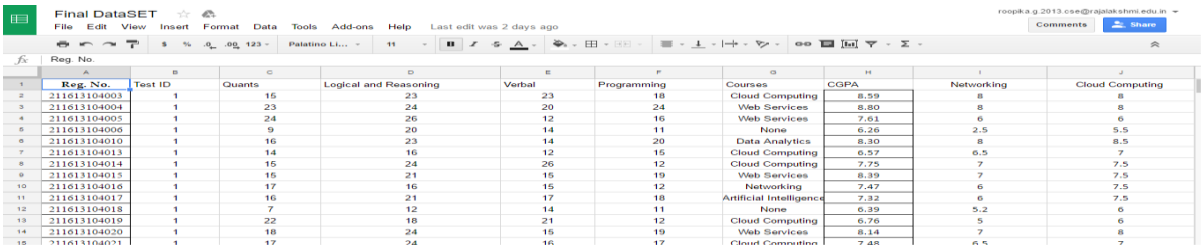
II. RELATED WORKS

Data mining is a practice that has been gaining a lot of popularity in the past couple of decades; however this is not a new idea or practice. Data mining is defined as the process of discovering patterns in large amounts of data or uncovering hidden information in large amounts of data. In a study[1] they used data mining to look at two different datasets – interactions between students and their professors and interactions between fellow students. The study found similarities and differences in the way students interacted with their professors in online questions and the ways that they interacted with their fellow students in online chat messages, which also identified disciplinary differences in the students' online participation. As well as the study found a correlation between the number of questions a student asked and their final grade. This study suggests that using data mining and text mining for online learning data can produce considerable insight into students' learning behaviours.[1].In a study about the article [2], a comparison of CART and C5.0 which produces a better result of mining data of students who are more likely to continue education or not resulted in 42.1% and 77.1% respectively proving that C5.0 is a better algorithm. As a result we have concluded on using the C5.0 algorithm for analyses in our proposed system hoping to yield a better outcome. Previous research has been conducted on classification models using decision tree algorithms on numerical data. In one such study, a professor worked on the data collected through the surveys from senior undergraduate students [3].Decision tree algorithms in the WEKA tool, ID3 and J48 were applied to predict who students are likely to continue their education with the postgraduate degree. The model was applied on two different students data and an accuracy of 88.68% and 71.74% was achieved with C4.5.In another study, it was found that the tree based algorithms outperformed the methods like Neural Networks and SVM. So as previous studies revealed that C4.5 is more accurate for such data [4][5][6] but C5.0 offers a number of improvements like speed, memory usage ,support for boosting etc [7]. In another study [8] ,student details were analysed to predict whether the student is eligible for placement or higher education. So, in this project, C5.0 algorithm has been used for classification. Similar to these articles the system proposed by us maps the domain interests of the student through the courses he has attended to the company listed in that domain. This provides a great opportunity for students to like their career after their learning process.

III. DATA PREPROCESSING

Data of students were collected which consists of the following parametres:

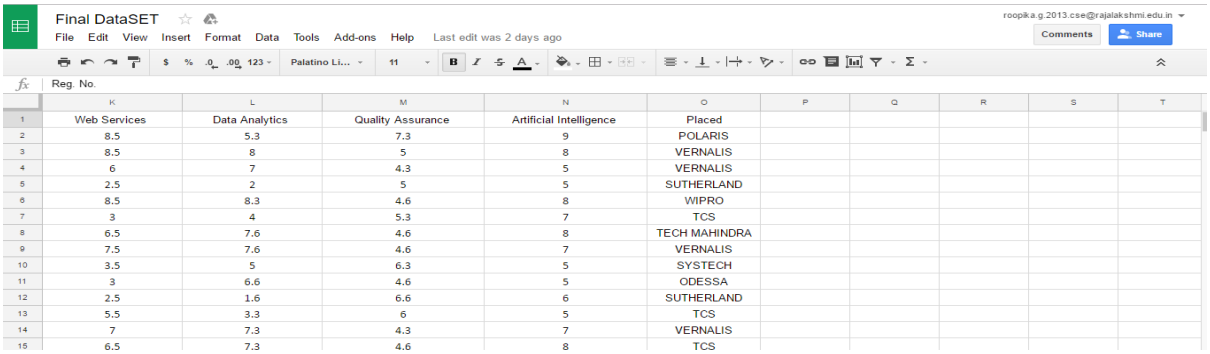
1. Individual CGPA
2. Quants, logical and reasoning, verbal marks obtained during training sessions
3. Programming marks obtained during training sessions
4. Marks in various domains under Computer Science Stream
5. Company in which they were recruited



Final DataSET

Reg. No.	Test ID	Quants	Logical and Reasoning	Verbal	Programming	Courses	CGPA	Networking	Cloud Computing
211613104003	1	15	23	23	18	Cloud Computing	8.59	8	8
211613104004	1	23	24	20	24	Web Services	8.80	8	8
211613104005	1	24	26	12	16	Web Services	7.61	6	6
211613104006	1	9	20	14	11	None	6.26	2.5	5.5
211613104010	1	16	23	14	20	Data Analytics	8.30	8	8.5
211613104013	1	14	16	12	15	Cloud Computing	6.57	6.5	7
211613104014	1	15	24	26	12	Cloud Computing	7.75	7	7.5
211613104015	1	15	21	15	19	Web Services	8.39	7	7.5
211613104016	1	17	16	15	12	Networking	7.47	6	7.5
211613104017	1	16	21	17	18	Artificial Intelligence	7.32	6	7.5
211613104018	1	7	12	14	11	None	6.39	5.2	6
211613104019	1	22	18	21	12	Cloud Computing	6.76	5	6
211613104020	1	18	24	15	19	Web Services	8.14	7	8

Figure1 Preprocessed dataset-1



Final DataSET

Web Services	Data Analytics	Quality Assurance	Artificial Intelligence	Placed
8.5	5.3	7.3	9	POLARIS
8.5	8	5	8	VERNALIS
6	7	4.3	5	VERNALIS
2.5	2	5	5	SUTHERLAND
8.5	8.3	4.6	8	WIPRO
3	4	5.3	7	TCS
6.5	7.6	4.6	8	TECH MAHINDRA
7.5	7.6	4.6	7	VERNALIS
3.5	5	6.3	5	SYSTECH
3	6.6	4.6	5	ODESSA
2.5	1.6	6.6	6	SUTHERLAND
5.5	3.3	6	5	TCS
7	7.3	4.3	7	VERNALIS
6.5	7.3	4.6	8	TCS

Figure 1.1 Preprocessed dataset-2

The marks of several students under Computer Science department was collected and grouped into major domains such as Quality assurance, Web Services ,Networking ,Cloud Computing, Data Analytics and Artificial Intelligence. For eg: Subjects such as Total quality management and Software testing were aggregated to form a domain called Quality assurance. The grades obtained in each subject was then converted into equivalent grade point such as S=10, A=9, B=8, C=7, D=6, E=5 and U=0 to add up to a total mark out of 10.The above dataset is then given into the classification algorithm C5.0 which builds a decision tree that classifies students based on certain threshold for the above mentioned domains along with quants, verbal, logical reasoning, programming marks and their CGPA. So that they will have higher chances of getting placed in certain companies that maps to their domain.

IV. DATA MINING TECHNIQUES

4.1 Classification

The C5.0 algorithm was created by Quinlan which is the slightly improved version of C4.5 that possess the following advantages:

- Speed - C5.0 is proven to be faster than C4.5 (several orders of magnitude)
- Memory usage - C5.0 consumes memory efficiently than C4.5
- The accuracy of the tree is improved with support for boosting.

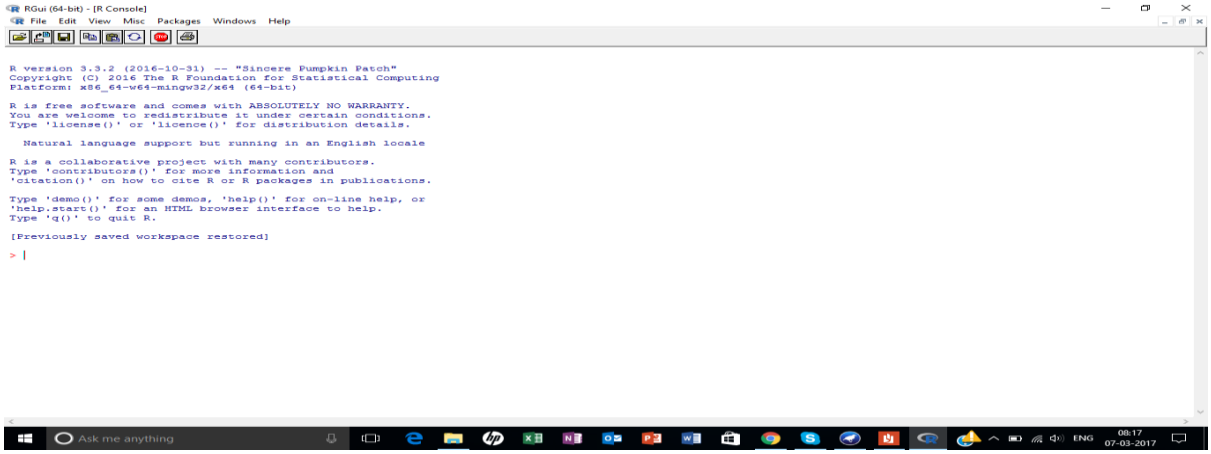
4.2 Prediction in R

Prediction is the task of applying classifier rules on test data to predict the accuracy of the classification algorithm. In a prediction the following takes place:

1. The actual dataset is split into two parts for cross-validation where one part being the training data and the other being test data.
2. The classification algorithm is applied on training data which results in classifier rules.
3. The classifier rules obtained from training data is applied to the test data for prediction which yields the prediction rate of the classification algorithm.

V. DATA ANALYSIS USING R

R (Revolution) is a free software environment that provides support for a wide variety of statistical and graphical techniques and can also be extended easily via packages. Currently there are more than 4800 packages that are available in the CRAN package repository. The strength of R is identified by the well-designed plots that it produces.



Fig*ure 2: R-3.3.*1

Advantages:

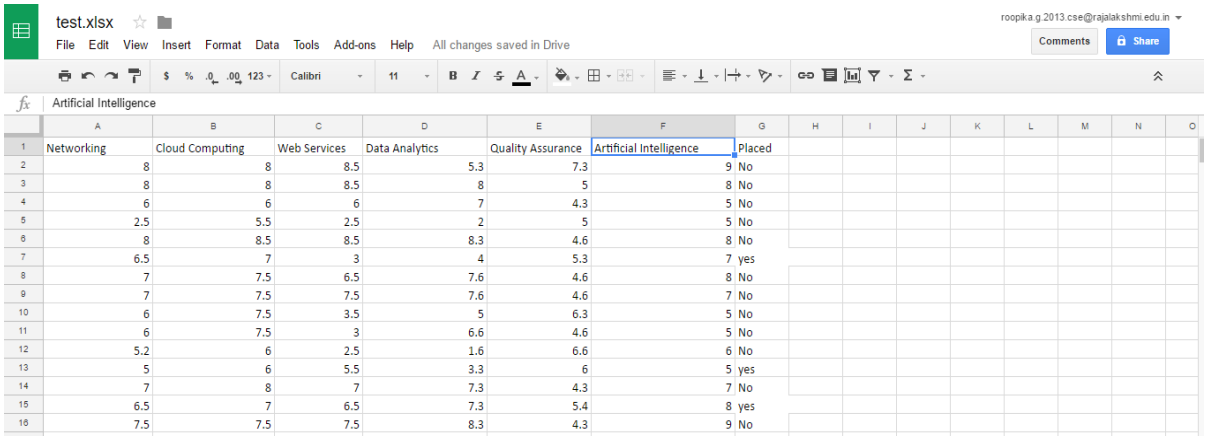
- Statisticians and researchers use R programming language for statistical analysis.
- R provides an outstanding programmable graphical packages.
- R allows anyone to use and modify ,since it is a free open source software.
- R allows us to run it anywhere and anytime ,since it has no license restrictions.
- R welcomes anyone to provide code enhancements and new packages.

Limitations:

- Documentation is sometimes a tedious task However, some standard books are available to fill in the documentation gaps.
- Some packages are not of perfect quality.

5.1. Classification using R: C5.0 Classification algorithm was used to classify the dataset which results in decision tree from which classifiers were built. The classification is applied in two stages for the dataset as mentioned below:

Stage 1: Classification of domain for each company: The dataset was prepared company-wise as shown in Figure 3 from the original dataset. Then , the algorithm was applied for the prepared dataset which resulted in a decision tree as shown in the Figure 4.



test.xlsx

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

roopika.g.2013.cse@rajalakshmi.edu.in

Comments Share

Artificial Intelligence

	A	B	C	D	E	F	G
1	Networking	Cloud Computing	Web Services	Data Analytics	Quality Assurance	Artificial Intelligence	Placed
2	8	8	8.5	5.3	7.3	9	No
3	8	8	8.5	8	5	8	No
4	6	6	6	7	4.3	5	No
5	2.5	5.5	2.5	2	5	5	No
6	8	8.5	8.5	8.3	4.6	8	No
7	6.5	7	3	4	5.3	7	yes
8	7	7.5	6.5	7.6	4.6	8	No
9	7	7.5	7.5	7.6	4.6	7	No
10	6	7.5	3.5	5	6.3	5	No
11	6	7.5	3	6.6	4.6	5	No
12	5.2	6	2.5	1.6	6.6	6	No
13	5	6	5.5	3.3	6	5	yes
14	7	8	7	7.3	4.3	7	No
15	6.5	7	6.5	7.3	5.4	8	yes
16	7.5	7.5	7.5	8.3	4.3	9	No

Figure3: Company specific dataset

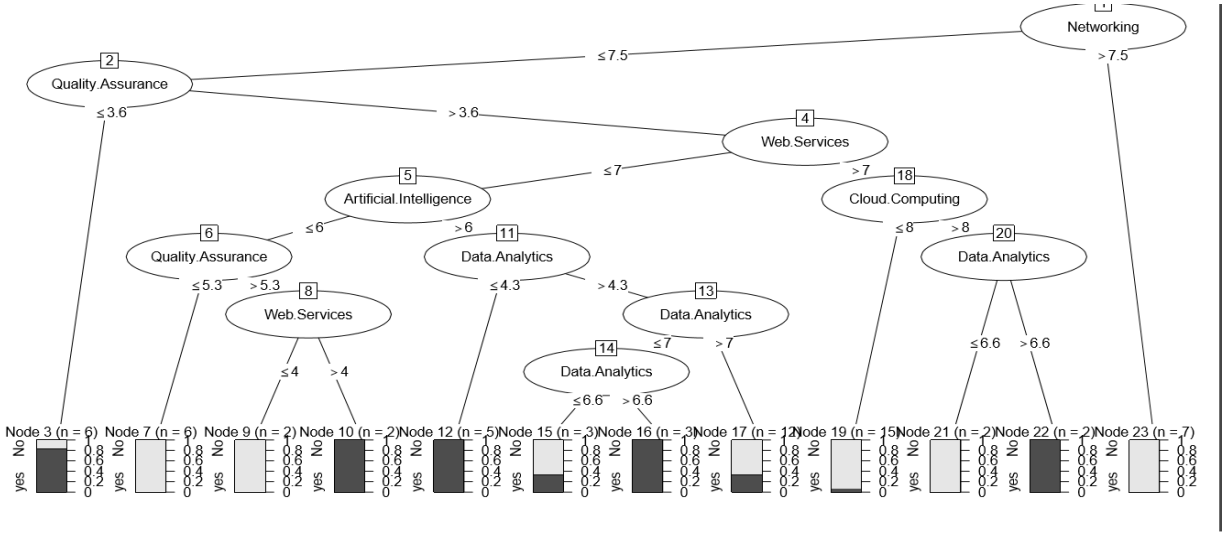
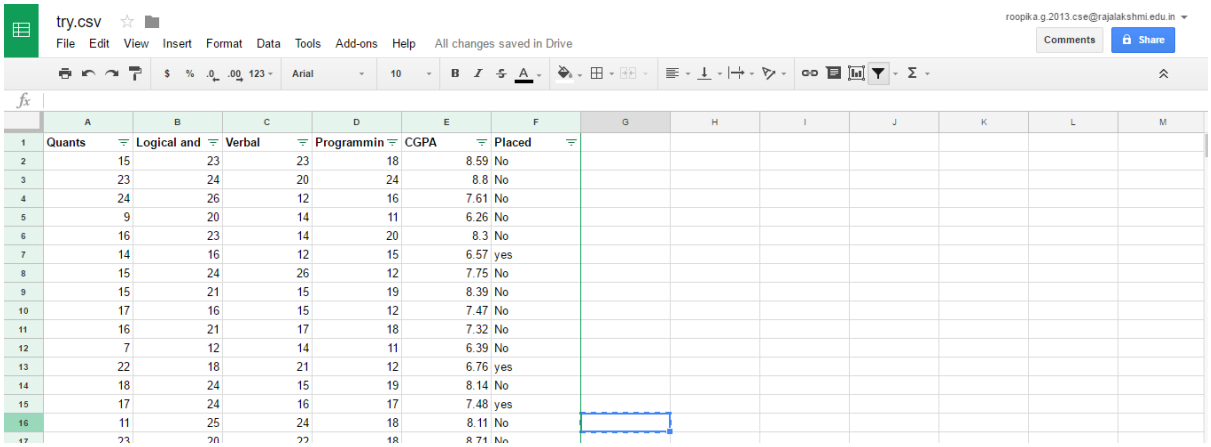


Figure4: Decision tree for the above dataset

From the above decision tree, best classification rules were identified [9] for a single company. These classification rules provides insights about the chances of placement for students for that company through their domain.

Stage2: Classification of remaining data for each company: The dataset which includes Quants, Logical reasoning, Programming, Verbal marks and CGPA was prepared with regard to company as shown in Figure 5 from the original dataset. Then, the algorithm was applied for the prepared dataset which resulted in a decision tree as shown in the Figure 6.



try.csv

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

roopika.g.2013.cse@rajalakshmi.edu.in

Comments Share

	A	B	C	D	E	F
1	Quants	Logical and	Verbal	Programmin	CGPA	Placed
2	15	23	23	18	8.59	No
3	23	24	20	24	8.8	No
4	24	26	12	16	7.61	No
5	9	20	14	11	6.26	No
6	16	23	14	20	8.3	No
7	14	16	12	15	6.57	yes
8	15	24	26	12	7.75	No
9	15	21	15	19	8.39	No
10	17	16	15	12	7.47	No
11	16	21	17	18	7.32	No
12	7	12	14	11	6.39	No
13	22	18	21	12	6.76	yes
14	18	24	15	19	8.14	No
15	17	24	16	17	7.48	yes
16	11	25	24	18	8.11	No

Figure 5: Prepared dataset

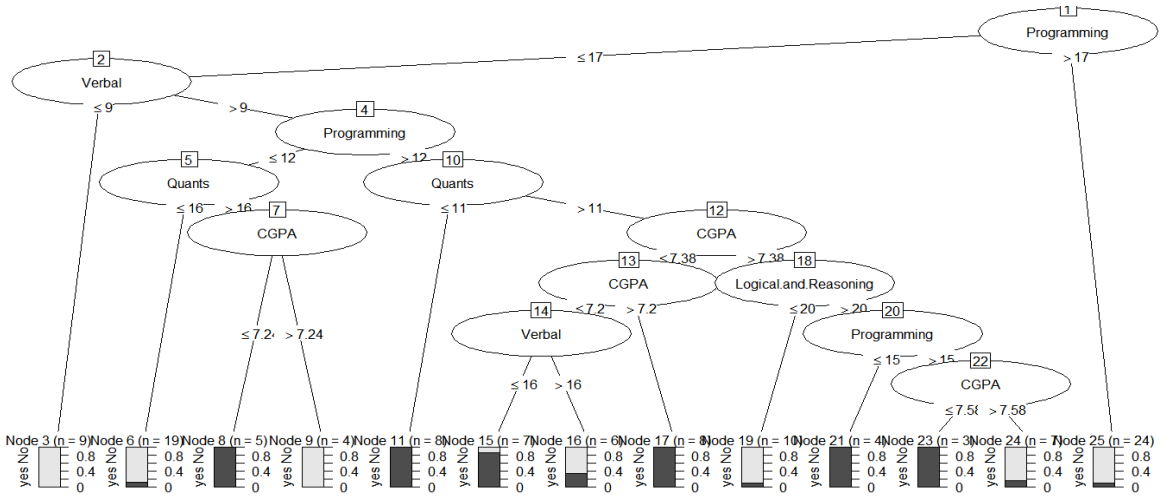


Figure 6: Decision tree for the above dataset

Classification rules[9] were built from the above decision tree that classifies the above prepared data set for a particular company. In a similar manner, classifiers are built for each company as mentioned in the original dataset. This helps the students to identify their eligible companies beforehand and could enhance placement results.

5.2 Prediction in R : The classifier was then tested against a new data set to calculate the accuracy of the model. The predicted output of the classifier was then plotted as a pie chart for easy visualization.

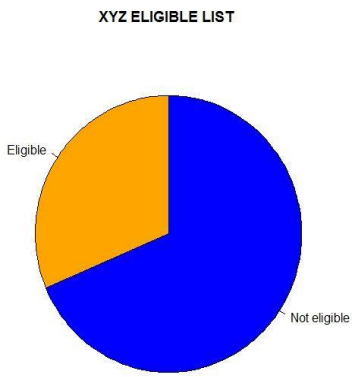


Figure 7: Prediction result for a particular company

The above Figure 7 illustrates the prediction of classifier built for a particular company against the test data which produced an accuracy of 75%.The students eligible for a particular company was grouped and in a similar manner, it was accomplished for other companies.

VI. RESULTS AND DISCUSSION

From the above analysis, the eligible company for a student is identified based on their technical and reasoning skills. The classified output is then clustered which resulted in the list of students who are eligible for each company This analysis was effectively applied for the student dataset under a particular department in an institution that predicts the placement opportunities.

VII. CONCLUSION AND FUTURE WORKS

Using C5.0 algorithm, a classifier was built to predict the eligibility criteria for each company. Prediction is done by applying the classifier rules against the test data to find the group of students who are eligible for each company which yielded a prediction rate of 75%. The future enhancement in this project is developing it as a mobile application to enhance the user experience.

References

1. He W. (2012). Examining Students' Online Interaction in a Live Video Streaming Environment Using Data Mining and Text Mining. Computers in Human Behavior
2. Data Mining and Knowledge Management in Higher Education -Potential Applications.luan,jing.
3. V. P. Bresfelean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment," Proceedings of the ITI 2007 29th Int. Conf. on Information Technology Interfaces, June 25-28, 2007.
4. G. Eason, B. R. R. Kabra, and R. S. Bichkar, "Performance Prediction of Engineering Students using Decision," International Journal of Computer Applications (0975 – 8887) Volume 36– No.11, December 2011.
5. Sudheep Elayidom.M, Sumam Mary Idikkula, and Joseph Alexander, "Applying Statistical Dependency Analysis Techniques In a Data `Mining Domain," International Journal Of Data Engineering (IJDE), Volume (1), Issue (2).
6. V.Ramesh, P.Parkavi, and P.Yasodha, "Performance Analysis of Data Mining Techniques for Placement Chance Prediction," International Journal of Scientific & Engineering Research ISSN 2229-5518 Volume 2, Issue 8, August 2011.

7. Rutvija Pandya,Jayati Pandya ,"C5.0 Algorithm to Improved Decision Tree with Feature Selection anReduced Error Pruning", "International Journal of Computer Applications (0975 -8887) Volume 117 – No. 16, May2015.
8. Data Mining Approach for Suggesting Higher Education Courses Based on Student's Performance. The International Journal Of Science & Technoledge (ISSN 2321 – 919X) .
9. Integration of Rules from a Random Forest . International Conference on Information and Electronic Engineering IPCSIT vol.6 (2011) © (2011) IACSIT Press, Singapore, 2011.
10. http://stackoverflow.com/questions/153776.

