

PLACEMENT PREDICTION DECISION SUPPORT SYSTEM USING DATA MINING

Dissertation submitted in partial fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

in

INFORMATION TECHNOLOGY

By

JOSHITA GOYAL

11601048

Supervisor

SHILPA SHARMA



School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

@ Copyright LOVELY PROFESSIONAL UNIVERSITY, Punjab (INDIA)

November 2017

ALL RIGHTS RESERVED

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program : P173::M.Tech. (Information Technology) [Full Time]

COURSE CODE : INT548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0055
Supervisor Name : Shilpa Sharma **UID :** 13891 **Designation :** Assistant Professor
Qualification : _____ **Research Experience :** _____

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Joshita Goyal	11601048	2016	K1638	9888243037

SPECIALIZATION AREA : Programming-II **Supervisor Signature:** _____

PROPOSED TOPIC : Placement Prediction Decision Support System using Data Mining.

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	Parameter	Rating (out of 10)
1	Project Novelty: Potential of the project to create new knowledge	6.75
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.00
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.25
4	Project Supervision: Project supervisor's is technically competent to guide students, resolve any issues, and impart necessary skills.	7.25
5	Social Applicability: Project work intends to solve a practical problem.	7.25
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	6.75

PAC Committee Members		
PAC Member 1 Name: Kewal Krishan	UID: 11179	Recommended (Y/N): Yes
PAC Member 2 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 3 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 4 Name: Dr. Pooja Gupta	UID: 19580	Recommended (Y/N): Yes
PAC Member 5 Name: Kamlesh Lakhwani	UID: 20980	Recommended (Y/N): NA
PAC Member 6 Name: Dr. Priyanka Chawla	UID: 22046	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): Yes

Final Topic Approved by PAC: Placement Prediction Decision Support System using Data Mining

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024::Amandeep Nagpal

Approval Date: 04 Nov 2017

11/25/2017 11:23:19 AM

DECLARATION BY THE SCHOLAR

I hereby declare that the research work reported in the dissertation proposal entitled "PLACEMENT PREDICTION DECISION SUPPORT SYSTEM USING DATA MINING" in partial fulfillment of the requirement for the award of Degree for Master of Technology in Information Technology at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Shilpa Sharm. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

Signature of Candidate

Joshita Goyal

11601048

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation proposal entitled “Placement prediction decision support system using data mining” submitted by **Joshita Goyal** at **Lovely Professional University, Phagwara, India** is a bonafide record of his / her original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Signature of Supervisor

(Name of Supervisor)

Date:

Counter Signed by:

1) Concerned HOD:

HoD's Signature: _____

HoD Name: _____

Date: _____

2) Neutral Examiners:

External Examiner

Signature: _____

Name: _____

Affiliation: _____

Date: _____

Internal Examiner

Signature: _____

Name: _____

Date: _____

ACKNOWLEDGEMENT

I would like to thank my mentor Ms. Shilpa Sharma who was always there for me whenever I needed. She helped me in every possible way. She directed me and guided me throughout the process. Whenever I found myself in trouble I rushed to her cabin and she resolved my issue.

I would also thank the PAC committee for considering my research as noble and asked me to carry it on.

Finally I would like to thank my family and friends who were always there for me who always supported me and encouraging me to work hard.

Joshita Goyal

11601048

TABLE OF CONTENTS

CONTENTS	PAGE NO.
PAC form	i
Declaration by the Scholar	ii
Supervisor's Certificate	iii
Acknowledgement	iv
Table of Contents	v
List of Figures	vii
Abstract	viii
CHAPTER 1: INTRODUCTION	1
1.1 INTRODUCTION TO DATA MINING	1
1.2 EDUCATION DATA MINING	1
1.3 KNOWLEDGE DISCOVERY PROCESS	2
1.4 POPULAR ALGORITHMS USED IN DATA MINING	2
CHAPTER 2: REVIEW OF LITERATURE	4
CHAPTER 3: PROPOSED WORK	20
3.1 SCOPE OF THE STUDY	20
3.2 OBKECTIVES	20
CHAPTER 4: RESEARCH METHODOLOGY	21
CHAPTER 5: EXPECTED OUTCOME	24

CHAPTER 6: SUMMARY AND CONCLUSION	25
LIST OF REFERENCES	26

LIST OF FIGURES

FIGURE NO.	DESCRIPTION	PAGE NO
Figure 4.1.	Apriori Algorithm	23

ABSTRACT

This is propose work for the research topic “Placement prediction decision support system using data mining.” As the name suggests this research is based on predicting students’ performance in upcoming placement drives in universities. This research is not just helpful to universities in retaining their reputation but it will also help the students to know that how they are likely to perform in upcoming placement drives. This help them in knowing their current status and will direct them to perform better. In this research we’ll we using academic data of students for processing data using association mining technique called apriori algorithm. Apriori algorithm is having limitations and the major one is larger execution time taken by this algorithm in processing bulk data. Through research we aim to develop an improved apriori algorithm with top down approach. Techniques to be used in this algorithm are insertion sort and depth first search. After implementing both the algorithms on same dataset we’ll be able to compare the results of both using weka data mining analysis tool. By this we will be able to process data in lesser execution time then the previous one. This research provide status of students at multiple levels of the placement process such that it will be evaluated that how a student is likely to perform in aptitude test, technical test, group discussion and final interview.

CHAPTER 1

INTRODUCTION

1.1. Introduction to data mining

Data mining is not new field in the area of research. Data mining is the domain which lead to data analytics and creation of many systems which are used across the world in many everyday application. Data mining is application in all fields be it agriculture, science, biology, education and where not. Data mining as the name suggests it's the approach by which we are able to mine the data means we can dug a large database which will help us get useful information contained in it. This information can be framed to form knowledge. Data mining can also be refer to art by which we are able to develop decision support system. These systems have been used in medical, forensics, etc. We are also keeping up the legacy to use data mining in our research which is based on placement prediction system. Placement prediction comes under education data mining which is sub domain of data mining.

Data mining is giving the researcher a lot of opportunities to work on existing problems and to make up new system which can be helpful. Data mining contains a lot of techniques like classification, clustering and association. In this research we'll be using association technique to find out patterns. We will work on education data mining. In this analyzing the existing research we'll do certain improvements to make a system a better and robust.

1.2. Education data mining

Undoubtedly education data mining is the emerging area of research. In this a lot of research have been done. Basically main aim of this area is to make our education system a better system by rigorous research. A lot of researchers have been putting their efforts to make it possible. Education data mining is not different from data mining but it's the sub domain of data mining wherein all the techniques and approaches used in data mining are applicable. Education data mining deals with

every aspect of education be it related to student teacher relationship or analysis of students' performance or behavior of student and many more things comes under this. Researches which have done in this field are related to students' performance, behavior of students' in project submission, how higher schools grades are related to future career and a lot of work have been carried out under this. Education data in not just help to education system but also to online platforms which are providing good certifications to students. It also helps improve business. As we know that now a days education is not limited to classrooms but online platforms are also imparting education to the students.

Education data mining for business: Many websites are hosted on internet to find out what students need these days. They are providing online courses to them. That's how education data mining is helping entrepreneurs to make revenues. According to the students interest they impart knowledge to them. Here in this research we'll be working on to improve the existing placement prediction of students by using our won algorithm in it.

1.3. Knowledge discovery process

Knowledge discovery as we know is what we call data mining. So here im discussing the process used in it. This process contains several phases which are listed below:

- **Cleaning and preprocessing:** In this process, we remove the redundant data and we'll fill the missing values in the dat.
- **Integrating:** Now we'll integrate the data from various database and will form one database.
- **Selection:** Selection of require data is done in it. The data which is required for analysis in chosen in it.
- **Data transformation:** In this we'll transform our data to form which is accepted by our data mining tools which is used for analysis.
- **Application of data mining techniques:** Out of various techniques we'll apply one or required number of techniques to dataset for analysis on data mining tool.
- **Analysis of result and prognosis:** Now we'll analyze the result and will make predictions on what can be done.

1.4. Algorithms used in data mining

There are various algorithms in data mining which are used for analysis. These algorithms are implemented in various data mining tools for analysis. Some of the algorithms are listed below:

- **Classification:** In this algorithm, we classify a particular item to a target class to which it can belong based on its properties.
- **Association:** This is basically to find the relationship or associations between various attributes within the database.
- **Clustering:** By this technique we form clusters of various items. This can be implemented after classification of items.
- **K-means:** This algorithm is a type of clustering in which we classify items in k clusters with little difference in clusters.
- **Apriori:** In this we find frequent itemsets that is we'll find what item can occur together. This algorithm is mostly applicable in business related applications.
- **Regression:** This is another data mining technique which gives numeric value of the dependent variable. Relation of dependent variable with other independent variable is evaluated in this.
- **Support vector machine:** This algorithm is helpful when we know the difference between two itemsets. Then we can express it using hyperplane and equation is $y=mx+b$.
- **K nearest neighbor:** If we have dataset which is clustered and while inputting new item into it then it will choose the clusters based on k neighbors.

CHAPTER 2

REVIEW OF LITERATURE

- **Maria Lydia Fioravanti and Ellen Francine Barbosa,” A Systematic Mapping on Pedagogical Patterns,” 2016 IEEE Frontiers in Education Conference (FIE), 2016, Pages: 1 – 9**

In this paper, authors stated that pedagogical patterns can be a tool to assist in new teaching and learning applications. Authors said that despite having a lot of benefits, existing education application have a lot of challenges and problems which needs to be solved and to address these they have taken pedagogical patterns as a tool. For this they have considered 109 pedagogical patterns from 35 scientific studies. This study was conducted in five steps. Firstly answers to two questions was evaluated that what pedagogical patterns have been used. And what problems will these patterns solve. Primary study was conducted by browsing manually through journals. Then papers were screened and relevant papers were extracted. Then authors focused on abstract of paper and then systematic mapping was done. It was concluded that systematic study was more better approach than other approaches.

- **Liang Ge, Huang Tang, Qing Zhou, Yunheng Tang, Jiangtao Lang,” Classification Algorithms to predict students’ Extraversion-Introversion Traits,” 2016 International Conference on Cyberworlds (CW),2016, Pages: 135 - 138**

In this research authors wished to find out the extraversion and introversion traits of the students. For this they performed an experiment in which they distributed the questionnaire to students in which 79 students participated. The questionnaire consists of various questions and the score of questionnaire was divided into two scales 0 to 50 and 50 to 100. Score of 0-50 were considered introvert and latter were considered extrovert. Now they also collected they browsing history of student on the basis of information in campus card. It contained data in four columns and 25499 rows. This contained student id, type of site visited, year-month, access count. Then they also

contained consumption data that at which site student spends more. This data was processed using naïve bayes, linear SVM and classification and regression trees in sci-kit. It was concluded that students who spends more time online were introvert and rest were extrovert. Linear SVM performed best for this experiment.

➤ **Larian M Nkomo and MuesserNat,” Discovering Students Use of Learning Resources with Educational Data Mining,” 2016 HONET-ICT, 2016, Pages: 98 –102**

In this research authors carried out a research in which they formulated hypothesis that student view the resource at last minute before exam and they procrastinate online project submissions. For this experiment, they collected data from log of MOODLE which contained date, time of resource accessed, name of student, resource accessed and action done. Data collected was related to mid term and case study. It contained four chapters and four case studies. Each chapter and case study had related data regarding its access. Authors used weka for it. They preprocess the data. Converted into csv format as accepted by weka. Then they applied apriori algorithm into it. and hence results were evaluated. This research was just to give insight to the professor to know the behavior of the student. This research was limited to data of post graduate students .

➤ **Pornthep Khongchai and Pokpong Songmuang,” Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique,” 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Yogyakarta, Indonesia,2016,Pages 1-6.**

According to this paper, authors created a system wherein they predict the salaries of the students who are currently enrolled in a degree. This system was created to motivate students to work hard. In this researchers took data of students who are currently pursuing some degree and who are graduated and are doing jobs. Data contained academic records of the students and data of graduated students with their salary. Three students who were top rankers and attained highest salaries were considered as exemplars. Tool used for this was weka. Data mining techniques which were applied

were KNN, naïve bayes, j48, multilayer preceptor and support vector machine. Out of all these KNN gives better accuracy with 84.69%.

- **Houda Essalmi, Mohamed El Far, Mohammed El Mohajir, Mohamed Chahhou,” A novel Approach for mining frequent itemsets: AprioriMin,” 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 2016, Pages: 286 - 289**

While all authors were busy experimenting with existing algorithms on different datasets then these authors attempted to create a new algorithm from the existing one. They create a new improved apriori algorithm which takes less execution time and twice faster than the existing one. Authors optimized the generation and pruning phase of the algorithm.

Apriori is one of the best algorithm when you want to find relation between different items and you want know that how frequent a particular itemset appears based on various parameters. The new algorithm created by the authors is AprioriMin. In this research authors compared the results of this algorithm with existing one, and two other algorithms namely FP growth and Close. Usually the existing algorithm works into steps that is to produce the k itemsets candidates out of k-1 itemsets. And scan transactions to calculate Support candidates. New algorithm is based on Support min. Support min is used in pruning step to find if k itemset can't be frequent on one hand. On the other hand its used mine frequent itemset by not having access to database but by calculating minimum support from candidate set which approximates the metric support. The algorithm was created in java. After implementing it showed drastic difference between the existing algorithms which were apriori, FP growth and close when applied on same datasets.

- **Shyam Kumar Singh and Preetham Kumar,” I2Apriori: An Improved Apriori Algorithm based on Infrequent Count,” 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),2016, Pages: 1281 -1285**

This research is again done to optimize apriori algorithm. Authors in this research attempted to reduce the CPU computational time by reducing the number of transactional scan. They applied the concept of infrequent counts and two way searching. Infrequent counts is based on minimum threshold support. Normally apriori run in a way in which it has to find the minimum support and then generating rules from frequent itemsets. It use level wise and breadth first search to find out frequent itemsets where k-itemsets are used to generate k+1 itemsets. It is having a limitation which is that it has to scan all transactions to check to if a given itemset is frequent. Now the proposed algorithm I2Apriori is based on Support. In this it scans the database from top to bottom and bottom to top until the frequency is not equal to minimum threshold frequency and infrequent support count is equal to total transaction subtracted by min support plus 1. On its basis itemset are declared frequent or infrequent. When this algorithm was implemented it was found that the new one is better than the existing one. As it reduced traversing the database and hence reduced execution time.

➤ **Jiaoling Du, Xiangli Zhang, Hongmei Zhang and Lei Chen,” Research and Improvement of Apriori Algorithm,”2016 Sixth International Conference on Information Science and Technology (ICIST), 2016, Pages: 117 - 121**

In this authors developed a new apriori algorithm called improved DC_Apriori algorithm. Researches have been done to reduced the number of candidate set and execution time. We know how apriori works such that it first scan the original transaction database and then found minimum support from it. Based on this support further candidate sets are formed till we get no frequent itemset. Now author found three limitations in it which are clearly stated in this paper are as follow. First is that it has to repeatedly scan the database. Second is that it has to generate a lot of candiadate sets based on support from previous transaction set. Third is it takes a lot of execution time. Now authors have change the databse structure of how they transaction ids and itemset are stored. Authors have stored the items in structure and each item set has corresponding transaction id in which it occurs. Now support will be based on transaction ids. Second step of improved algorithm is to check if it can from frequent itemsets or not. If yes then they have proceeded to step three wherein candidate sets are

generated. When this algorithm was implemented it was concluded that this algorithm worked better in comparison to Apriori and MC_Apriori algorithm. It was able to overcome all the three problems stated above. It was executing in lesser time and consumed less CPU.

- **Zahyah Alharbi, James Cornford, Liam Dolder and Beatriz De La Iglesia,"Using data mining techniques to predict students at risk of poor performance,"2016 SAI Computing Conference (SAI),2016, Pages: 523 - 531**

In this paper, authors attempted to find out that by what percentage does students who are from different nations are at risk of performing bad in future. They took data from admission cell and first year students of a particular university. The students whose data was missing and discontinued their course were eliminated in the process of cleaning and filtering. Dataset was also based on their personal characteristics that whether students are from overseas or the native country. Then a dataset was taken which contained data of undergraduate students who were from different disciplines and different schools. On both dataset data cleaning was done. Firstly it was found out that how many students from particular location obtained good honor, not good honor. It showed that students who are from overseas perform worst than other students. When analyses were performed on second dataset than the outcome was again same. They accessed each attribute independently using feature selection algorithm with chi square. Then they found that fee status, nationality, gender were all significant attributes. Then certain classification algorithms were used to found out that good honor and not good honor. They used IBM SPSS Modeler v15 software. The auto classifier which were selected for this were C&R Tree, Quest and CHAID. After experimenting they found out that 24% students achieve good honor degrees and 67% will achieve low honor degrees. By this research they found out that which students need more important and remedial classes. This system was developed to help administration to know the future status of the students.

- **Wildan Budiawan Zulfikar, Agung Wahana, Wisnu Uriawan and Nur Lukman,"Implementation of association rules with apriori algorithm f**

or increasing the quality of promotion ,” 2016 4th International Conference on Cyber and IT Service Management, 2016, Pages: 1 – 5

In this paper, authors applied data mining techniques to increase the sales of a company. Basically the authors used a company called XMART. As the basis purpose of this research was to increase the sales of company so they applied apriori algorithm to sales data. Apriori algorithm fits best for this purpose so as to know that which item will occur with item as combination. Frequent dataset is important to know. So they applied apriori algorithm. But the existing algorithm was having a disadvantage that with larger dataset it takes more execution time and CPU consumption is more. And also candidate sets formed by this approach is more. So an improved apriori algorithm devised for this process aims to produce less candidate sets. According to data there were eight clusters which contain frequent item sets. Firstly confidence and support was evaluated. Prior to this antecedent and consequent were determined. They used minimum limit for support and confidence about 60% to produce less rules. Then evaluation was done in which rules were evaluated. By implementation of this algorithm, authors were able to form the association rules which can improve the sales. Efficient rule building was done by use of improved apriori algorithm.

➤ **Shashi Kant Shankar and Amritpal Kaur,”Constraint data mining using apriori algorithm with AND operation,”2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2016, Pages : 1025-1029**

In this authors, worked on to improve the existing apriori algorithm. As apriori algorithm is one of the most oldest association technique used in data mining. So a lot of researches have been done to improve it. This research is also an attempt to do so with constraint data mining algorithm. The major limitation stated in this paper about apriori algorithm is time and space. Existing algorithm takes a lot of execution time as it generates a lot of itemsets which are not useful. So authors applied constraint data mining to it. Constrained data mining algorithm is yet another approach which use parameter as constraint which is passed as a condition for filtering for reducing the generated rules. It allows association rules to generated from the filtered itemsets. Now in proposed algorithm MST and MCT are anded with each other. So that the result of

frequent itemset will be passed from this condition thus removing unused rules. The experiment was performed on sample itemset using MATLAB software. Now the result obtained after implementation of this approach drastically reduced the execution time and space. The authors wanted to extend this work by implementing machine learning algorithms using this approach as future work.

- **Ashish Shah,” Association Rule Mining with Modified Apriori Algorithm using Top down Approach,” 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, Pages: 747 - 752**

In this paper, author have implemented an improved apriori algorithm using top down approach. In this author has used the concept of hash. As number of improvements have been done to apriori algorithm. This is yet another technique to modify apriori algorithm by using hash function. This function divides the frequent itemsets into buckets. He has also proposed a technique by which infrequent itemsets can be removed from candidate sets. Earlier traditional method used to scan the transactional database and produce itemsets which were infrequent. Some limitations of traditional apriori algorithm has been specified in this paper. It says the computation starts from smallest frequent itemset and reaches to the largest. When itemset is longer and frequent it takes larger processing time. So author used hash based technique which divides the dataset into two parts. If we use the propose technique we will see that all itemsets are frequent and if some itemset is not frequent we can prune it. Thus it saves a lot time and CPU consumption. By this technique author was able to reduce processing time to greater extent. This technique provided an another technique to improve the apriori algorithm.

- **Maocai Cheng, Kaiyong Xu and Xuerong Gong,” Research on audit log association rule mining based on improved Apriori algorithm,” 2016 IEEE International Conference on Big Data Analysis(ICBDA), 2016, Pages: 1-7**

In this paper, authors have proposed an improved apriori algorithm called E-Apriori algorithm, which improve like other proposed algorithm the performance. The proposed

algorithm aims to reduce the problem of low utilization of audit logs of the security audit system using association rule mining. E-apriori algorithm will make the running time less and improve the efficiency. It will also identify the attacks. Improved algorithm also scans the database like apriori but in this support count has one more variable which has inverse relation with itemset. By this time complexity decreases in the improved algorithm. The experiment was carried out in weka platform. When it was implemented the improvement was shown clearly in the form of graphs. The attempt was successful as the proposed technique was able to reduce the execution time thus improving the performance. Efficiency was more by the proposed techniques over the existing one in terms of execution time.

➤ **Anupam Khan and Soumya K. Ghosh, “Analysing the Impact of Poor Teaching on Student Performance ,“ 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)**

In this age of information, educational data mining (EDM) is useful for finding the undiscovered relationships in a large dataset. In fact, EDM is amplifying its' popularity in recent time. In this paper, authors' objective was to validate the association between poor teaching quality and performance degradation by the student. This study successfully establishes the proportional relationship between poor teaching and student performance degradation. It shows how the final grade in a course actually co-relates with teaching quality. They collected data from an institute which is of national importance. They gathered data by following means: they gave students the questionnaire through online interface to evaluate teaching quality, data of student performance evaluation by teachers were done, declaration of Result and Publication of Teaching Evaluation, and relevant data was taken. Then they pre-processed the data. They considered the minimum final CGPA required for passing graduation. From preceding CGPA they evaluated the students' performance. Then teaching quality for the courses was evaluated. The student performance degradation factor (λ), a measure of negative value addition by a student, is significant while actual grade is less than expected in a course. It is the difference between expected and actual grade converted to a common scale then after data pre-processing they converted the data for analysis.

Apriori algorithm of association rule mining was used to discover correlation, implication, causality among the set of items or attributes. As the area of interest of their study is the degradation, where the student performs worse than expected. Therefore, they have extracted the association metrics for the cases where the value of D is true. They have calculated support, confidence and lift by considering two attributes teaching quality(T) and degradation (λ). They checked confidence of degradation against teaching quality for all student types. The result shows that the confidence of degradation gradually increases with poorer teaching quality. This is applicable for all student types. It is also important to mention that the degradation confidence decreases in better student, while teaching is of same quality. This suggests that performance degradation is inversely related to both student and teaching quality. This observation differs the literature which suggests that the performance is primarily related to student quality only. However, as per their observation the performance is dependent on teaching quality as well. This study proposed a quantifiable measure that indicates the degradation with respect to expected performance of a student.

- **Senthil Kumar Thangavel , Divya Bharathi P and Abijith Sankar,” Student Placement Analyzer: A Recommendation System Using Machine Learning,” 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2017), Jan. 06 – 07, 2017, Coimbatore, INDIA**

The authors have developed a placement prediction system using decision tree classifier. This system can tell us that which student is likely to get placed in company and who requires hard work. They have taken past data of students. In this they have taken academic details and also evaluated that which student is interested in placements. By this system, a student will get one of the five placement statuses which are mass recruiter, core company, dream company, not eligible and not interested. Data was taken from placement cell of their institution. Data was loaded in python code and macros were attached to it for easy processing. Then decision tree classifier was applied on it using scikit libraries. Scikit is an open source machine learning module in python which contain various machine learning algorithms. Now they compared the results with other tools such that weka and data miner by processing the data using decision

tree. It was concluded that the system they formed was more efficient than weka and data miner. They also compared the efficiency of decision tree by applying other algorithms on same data using weka. It was concluded that decision tree worked with more accuracy than other algorithms with least execution time.

- **Ravi Kumar Rathore and J. Jayanthi,” Student prediction system for placements training using fuzzy inference system. ICTACT Journal of Soft Computing, April 2017, VOLUME: 07, ISSUE: 03**

This paper predicted that whether a student is eligible for placements or not. MATLAB was tool used for this. Techniques used for this was fuzzy logic and inference system. Rules were made using fuzzy logic. If-else statements were produced to give true or false value which will tell us whether a student is eligible for placements. The author wanted to extend the research using other classification algorithms to build a better system.

- **Ankita Kadambande, Snehal Thakur, Akshata Mohol, Prof A.M.Ingole,” Predict Student Performance by Utilizing Data Mining Technique and Support Vector Machine”, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 05 | May -2017**

These authors works on to build a student performance prediction system. They build a system considering three modules of student, staff and admin. Wherein students will be appearing for test and can view their performance, staff can view the performance of students and admin can upload the events like test etc. This research was to perform a new system over an existing system which was that data was processed in weka using j48 algorithm. In proposed system the data stored in apache database was processed using support vector machine. And final result was evaluated and students performance was predicted. It’s main aim was to improve the quality of education system.

- **Vinicius Gottin, Haydée Jiménez, Anna Carolina Finamore, Marco A. Casanova, Antonio L. Furtado and Bernardo P. Nunes,” An Analysis of Degree Curricula through Mining Student Records”, 2017 IEEE**

**17th International Conference on Advanced Learning Technologies,
2017 Pages: 276 – 280**

In this paper, researchers collaborated to find out that degree curriculum affects the students' performance to what extent. They have considered course of students as items and number of courses as transaction. They implemented this research by taking two types of degree curricula law and architecture. They categorized the students in three categories such that failed, abandoned or enrolled in. Now they evaluated as though how many students abandoned in particular semester. Based on this they formed assumptions that which student enroll in which course. Finally they evaluated that in which course students perform well and in which not. And group of students failed in which subject. This gives an administrator an insight into what courses should be recommended to students and which not. What courses the students are lagged in. This system can help to build a production degree curricula for students. For this research, researchers use frequent itemset mining as data mining technique.

- **Dijana Oreški, Mario Konecki and Luka Milić,” Estimating profile of successful IT student: data mining approach”, 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),2017, Pages: 723 - 727**

In this research, authors worked to estimate the profile of IT student. They worked by formulating hypothesis that high school grades and type of high school are associated with university grades. They used clustering for this. They distributed questionnaires for the students. Questionnaire was collected. And eight clusters were formed on the basis of data. It was suggested that grades above 4.0 will be considered as students who can be successful in IT. Clusters formed on the basis of age, gender, year of passing, school and subjects taken. This research also focused on another issue which states are guys are more successful in IT than girls. This research proved it wrong. This research can be a guideline to the students that what they can do in future. And what course they should Opt for. It tells whether a student will be able to fit in IT career or not.

- **K.Nasaramma, M.Bangaru Lakshmi, D.kiranmayi,” Prediction and Comparative Analysis of Students Placements Using C4.5 &C5.0”,**

K.Nasaramma et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 8 (3) , 2017,367-368

These authors collaborated to find out that which algorithm in better in evaluating the performance of students. They took academic data and then preprocess it. Feature selection was applied on it. Then data was processed in weka using C4.5 and C5.0 classifiers. It was found out that C4.5 gave accuracy of 88.8% while C5.0 gave accuracy of 95%. Thus C5.0 performed better than the previous one.

- **Jin Soung Yoo, Yei-Sol Woo and Sang Jun Park,” Mining Course Trajectories of Successful and Failure Students: A CaseStudy,” 2017 IEEE International Conference on Big Knowledge (ICBK), 2017, Pages: 270 – 275**

This paper aimed at the difficulties faced by student in choosing the right course. Most students enroll in degrees and they drop out owing to various reasons. In this paper, authors discussed this problem. Researchers used relationship mining to address two problems in education which are identifying the course trajectory students take to earn a degree and finding specific courses which may led student to change the course or to drop out. This is one of the most difficult challenge a student faces as it lays the foundation of his future. There are various types of relationship mining which are as follow : association rule mining ,sequential pattern mining, correlation mining and casual data mining. In association rule mining there can be any relationship between the variables. In correlation mining, there is linear correlation between variables. In sequential pattern mining, there is relationship between sequential occurrences of variables and it finds out if there exists any specific order of sequence. In casual rule mining, there is casual relationship between variables. In this paper, authors have used sequential rule mining and found out courses which students often enroll in. enrollment in major and minor courses were also considered. Data considered was of undergraduate students. Students who left the course were excluded in selection process. In second task data of students were taken who changed their course. Sequential pattern miming framework an open source framework was used for preprocessing and data cleaning. And further the data was fed to PrefixSpan program.

Sequential patterns were obtained with a support count. According to results computer science students mostly follow the courses which are advised to them.

- **A. Parkavi, K. Lakshmi,” Predicting the Course Knowledge Level of Students using Data Mining Techniques,” 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017, Pages: 128 – 133**

In this paper, author devised an algorithm to find out the course level knowledge of the student. The algorithm was created using association rule mining and regression analysis. This paper aims to find out that how a student is performing in a particular subject. Data of student was taken based on internal tests. Based on the performance it was suggested that whether a student needs a remedial class or not. For analysis purpose, authors used R programming and python. According to the algorithm stated in the paper it says that bloom level taxonomy was used. Firstly, excel sheet was read which contained marks of students in each subject. Then each question was analyzed as per the ranges of blooms taxonomy. Based on this skills were evaluated and best skill was evaluated. At the end it was concluded that whether a student is good at remembering, understanding, implementing, analyzing, evaluating and creative. Then chi square test for two subject were conducted which showed that there was no significant variation in knowledge levels. When regression analysis was applied on two different exam, it was concluded that student performance didn't vary significantly from blooming level. Hence it was proved that the devised algorithm is good for students' subject level knowledge analysis.

- **Wongpanya Nuankaew, Praty Nuankaew, Sittichai Bussaman, Passakorn Tanasirathum, "Hidden Academic Relationship between Academic achievement and Higher Education Institutions," 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), 2017, Pages: 308 - 313.**

In this paper, authors wished to find out the hidden relationship between academic achievements and higher education institutes. In this paper authors used data from two sources one from certain website and another is from questionnaire which was distributed to the students with rating on basis of likert scale. Usually it is found that students who opt for higher institute doesn't take admissions on the basis of their abilities, aptitude and marks but they are likely to follow their friends. These are some reasons that they are not able to perform well and usually abandon studies. These are referred to as the hidden relationship which author is trying to found out. Firstly data is preprocessed and data cleaning is done. Then to find out the relationship author implemented two algorithm techniques which were decision tree and feature selection. Feature selection contained further two types that is forward selection and backward elimination. While evaluating a confusion matrix was formed. Students' attitude was evaluated from questionnaire. After application of both the algorithms it was found out that decision tree worked with better efficiency than attribute selection. The accuracy which was given by decision tree was 82.99%. This paper was an effort to evaluate that how students choose to get admission into higher institutes. This research can be helpful in guiding students properly that what they can choose in future as their career option.

- **Febrianti Widyahastuti and Vinay Utami Tjhin,"Predicting Student's performance in final examination using linear regression and multilayer perceptron," 2017 10th International Conference on Human System Interactions(HIS), 2017, Pages: 188-192**

In this paper, authors have utilized online discussion forums. Basic purpose of this study was to help teacher know that how students are likely to perform in their final examination based on discussion form. In this authors collaborated to help institutes improve their students' performance. For this purpose authors gathered data from log of discussion forum. The data gathered two things one was what students are posting and another was their attendance in discussion forum. Weka tools was used for data analysis. The techniques which were used were Linear regression and multilayer preceptor. The data was analyzed based on three factors which were correlation coefficient, mean absolute error and root mean squared error. Correlation coefficient tells the degree of correctness between actual value and value predicted. Mean absolute error tells the closeness of prediction with respect the actual outcome as stated in the

paper and the last factor used for comparing two algorithm was root mean squared error which value of difference between the values predicted and available output. After evaluation of two approaches based on the criteria specified above the result was evaluated. Multilayer preceptor gave better accuracy than linear regression.

- **Mohammad-Hossein Nadimi-Shahraki and Mehdi Mansouri,” Hp-Apriori: Horizontal parallel-apriori algorithm for frequent itemset mining from big data,” 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017, Pages : 286 – 290**

In this paper, authors attempted to develop an algorithm which can faster the execution of processing database in big data environment. When we process high dimensional transactional database, then apriori algorithm takes a lot of time in execution. Maximum time taken by apriori algorithm is not in scanning database frequently for candidate itemsets but for calculating of support of first itemset as stated in this paper. Authors have taken help of distributed processing and parallel processing to implement it by modified apriori algorithm. Authors named their algorithm as horizontal parallel apriori. For use in this algorithm authors have divided the transactional database into four subsets. Firstly authors partitioned the database into horizontal dataset and then vertically for parallel processing. Authors used parallel data mining because of two things that is performance improvements of existing techniques and concurrent analysis. The algorithm was performed on four datasets which were retail, accidents, kosarak and webdocs. Proposed algorithm used four sub processes to get frequent itemsets. Out of four sub process first was used to generate candidate itemsets then next two sub processes were used to find the support count of candidate itemsets. And fourth task was to combine the results and determines that how many frequent itemset are there adding the to global FI list. This was implemented and results were evaluated and compared with another algorithm called count distribution which was performed on same dataset. Results obtained showed that proposed algorithm was better than the latter. So this can be helpful for processing data on big data.

- **Jun Yang, Haoxiang Huang and Xiaohui Jin,” Mining Web Access Sequence with Improved Apriori Algorithm,” 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),Year: 2017, Volume: 1, Pages: 780 - 784**

As number of researches have been done to optimize the apriori algorithm of data mining. This is yet another apriori algorithm proposed in this paper which is implemented to find out continuous sequential pattern. Apriori algorithm is used to find frequent itemsets from transactional database. Normally it is applicable for business or in retail sector to find out buying patterns of customers. But here authors have implemented to mine web log. The main purpose of this research is to predict that when user will be accessing the website for the next time. Generally traditional apriori algorithm is not applicable here due to its limitations. Major limitation that has been stated in this paper is that it takes a lot of processing time and traversal time. AC-Apriori algorithm was formed by combining aproiri algorithm with Aho-Corasick automation to reduce the scanning transaction database. In Aho-Corasick automation, during support calculation, the trie tree and failure pointer are constructed for candidate sequences and then candidate sequences are converted to AC automation. When implement this algorithm it reduced the execution time. And it worked better than existing algorithm inefficiency.

3.1. SCOPE OF THE STUDY

1. We would like to extend this research by using other machine learning approaches for better system. With Machine learning algorithms we would make our system a robust system. We would also integrate other algorithms like classification and other data mining algorithms which can make our system a better system and we'll see which best fits our requirements.
2. We would extend this system by associating the hidden relationship of performance in high school with future placements. Schools grades plays important role while choosing a discipline for your career. Also the choice of chosen discipline matters a lot in our future. So while it directly affects our placements. So we would integrate these factors in our future work.

3.2. OBJECTIVES

1. To implement the Association Rule Mining Algorithm i.e Apriori Algorithm on academic information data set.
2. To propose an update Apriori algorithm by using Quick Sort and Depth first techniques on existing education information data set.
3. To implement an update Apriori algorithm on existing academic information dataset to get more meaningful Association Rules.
4. To analyze the results of proposed algorithm and compare with existing Apriori Algorithm in the terms of Following Comparison parameters:
 1. Quality of Association Rules.
 2. No. of Item Sets.

CHAPTER 4

RESEARCH METHODOLOGY

- **First of all review of literature is to be done. It will provide an insight into what has been done in association mining. It will also contain the study of various approaches used in to obtained candidate set in various applications.**

In this research, we focuses on to develop a system which aim to help a lot in making our education system a better from the existing one. For this we have gone through a lot of research papers so that we can come to know that what is going on in our existing system and what updations have been done so that we can be directed to the right path. In our system like for any other system a lot of researches have been done. People have worked harder to construct an intelligent education system which focuses on many aspect. We will develop a system based on association mining. In our first phase we have done rigorous literature review regarding what researches have been done in education system. From this we found that existing research have been done using existing techniques. So we moved forward to develop a proposed algorithm. In this phase following steps were undertaken:

- Reviewing literature
 - Finding the broad topic.
- **In second stage we shall develop the proposed algorithm and implement the same. It will contain the introduction to proposed approach and how to apply proposed approach to the data set.**

During the second phase, as told earlier in the first phase we will propose an improved apriori algorithm which will aim to improve the speed of existing system. Our algorithm will be developed keeping in mind the drawback of the existing one which is that it takes alot of execution time as it needs to scan the original transaction database again and again. We will propose this algorithm and then we'll see that how to apply this algorithm on our dataset. In this we will see that what MST is required and if the itemset is frequent or not. Different steps to be followed in second phase are as follow:

- Analysis of existing research.

- Analysis of limitations in existing algorithm
 - Formation of objectives
 - Proposal of improved algorithm.
- **In third stage we shall do the Result analysis in the term of lift matrix and rules of the data obtained by Apriori algorithms. These will be then used to detect hidden rules.**

In this phase we will analyze the result in the terms of lift matrix. We'll find different rules formed by the algorithm. Lift matrix will give us measure of the frequency of together occurrence of antecedent and consequent. Rules which will be formed will lead us to our result. Rules will be based on frequent itemsets. All infrequent itemsets will be eliminated in this process. Then we'll compare our proposed technique with the existing one. And we'll see that whether it is able to make the process faster or has proved our hypothesis true or false. If our result is true such that we met our objective then our research will end. Summing up steps in few lines which are as follow:

Implementation of proposed algorithm in existing research using weka.

- Analysis of results of improved algorithm.
 - Compare the results with the existing one.
- **In the end conclude and suggest future work for Association Rule in Data Mining.**
- Now when we will be done with our work we will make conclusion out of what we have done. Our conclusion will include our results or the statistics obtained from this whole process. Basically we'll sum up everything we done in our research in a brief summary including our proposed technique and result in form of graphs and statistics. Then we'll suggest that what extra can be done in future from this research. This would be lying pathway for another researcher to come up with new researches. These steps will be followed:
- Make conclusion from results.
 - Suggest future work.

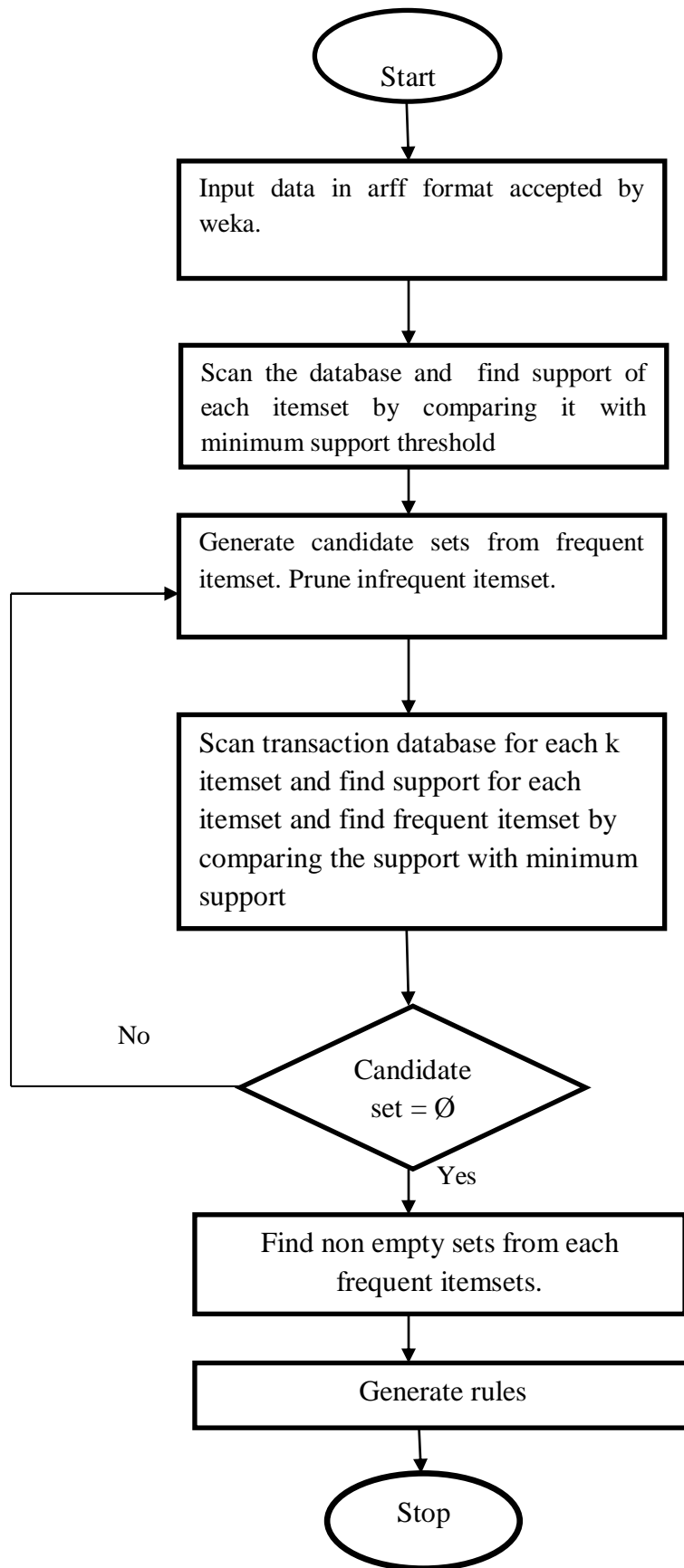


Fig 4.1. Apriori Algorithm

CHAPTER 5

EXPECTED OUTCOME

After implementing the methodology, we have we focuses on the outcome. Our expected outcome is based on the algorithm which has been used. In this research we have come implemented aproiri algorithm which repeated scans the transaction database. And certain rules are formed from it based on frequent itemsets. Now as we have used updated apriori algorithm, it will takes less time than the existing aproiri algorithm. Now after comparing existing algorithm with the proposed one it it is expected to overcome the weakness in existing system. Also it will predict that what levels in the interview a student will clear.

1. The methodology adopted will be generate rules based on the candidate sets.
2. Updated algorithm will take less execution time on the because of use of top down approach.
3. The proposed system will be able to predict that which students are likely to get placed and what rounds are they likely to clear in the placement process.

CHAPTER 6

SUMMARY AND CONCLUSIONS

In this research, we aim to develop a placement prediction system which will help retain education institutes' reputation. From the previous studies it was found that they have been using existing approaches to develop similar systems but in our research we be using updated data mining methodology for processing the data. The updated apriori algorithm which we will be using will process larger data sets faster than the existing one. And with better accuracy. With the proposed system the administrator will come to know about the future performance of students in placement drives and what phase of placement process a student is likely to pass. This is novel approach and can be used by any education institute. Also its not just good for institutes but also for students. Because with its use a student will come to know his/her status in placements and can work accordingly.

LIST OF REFERENCES

- [1] Maria Lydia Fioravanti and Ellen Francine Barbosa," A Systematic Mapping on Pedagogical Patterns," 2016 IEEE Frontiers in Education Conference (FIE), 2016, Pages: 1 – 9
- [2] Liang Ge, Huang Tang, Qing Zhou, Yunheng Tang, Jiangtao Lang," Classification Algorithms to predict students' Extraversion-Introversion Traits," 2016 International Conference on Cyberworlds (CW),2016, Pages: 135 - 138
- [3] Larian M Nkomo and MuesserNat," Discovering Students Use of Learning Resources with Educational Data Mining," 2016 HONET-ICT, 2016, Pages: 98 –102
- [4] Pornthep Khongchai and Pokpong Songmuang," Implement of Salary Prediction System to Improve Student Motivation using Data Mining Technique," 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS), Yogyakarta, Indonesia,2016,Pages 1-6.
- [5] Houda Essalmi, Mohamed El Far, Mohammed El Mohajir, Mohamed Chahhou," A novel Approach for mining frequent itemsets: AprioriMin," 2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), 2016, Pages: 286 - 289
- [6] Shyam Kumar Singh and Preetham Kumar," I2Apriori: An Improved Apriori Algorithm based on Infrequent Count," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT),2016, Pages: 1281 -1285
- [7] Jioling Du, Xiangli Zhang, Hongmei Zhang and Lei Chen," Research and Improvement of Apriori Algorithm,"2016 Sixth International Conference on Information Science and Technology (ICIST), 2016, Pages: 117 – 121
- [8] Zahyah Alharbi, James Cornford, Liam Dolder and Beatriz De La Iglesia," Using data mining techniques to predict students at risk of poor performance,"2016 SAI Computing Conference (SAI),2016, Pages: 523 – 531
- [9] Wildan Budiawan Zulfikar, Agung Wahana, Wisnu Uriawan and Nur Lukman," Implementation of association rules with apriori algorithm for increasing the quality of promotion ," 2016 4th International Conference on Cyber and IT Service Management, 2016, Pages: 1 – 5
- [10] Shashi Kant Shankar and Amritpal Kaur,"Constraint data mining using apriori algorithm with AND operation,"2016 IEEE International Conference on Recent Trends

- in Electronics, Information & Communication Technology (RTEICT), 2016, Pages : 1025-1029
- [11] Ashish Shah,” Association Rule Mining with Modified Apriori Algorithm using Top down Approach,” 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), 2016, Pages: 747 – 752
- [12] Maocai Cheng, Kaiyong Xu and Xuerong Gong,” Research on audit log association rule mining based on improved Apriori algorithm,” 2016 IEEE International Conference on Big Data Analysis(ICBDA), 2016, Pages: 1-7
- [13] Anupam Khan and Soumya K. Ghosh, “Analysing the Impact of Poor Teaching on Student Performance ,“ 2016 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)
- [14] Senthil Kumar Thangavel , Divya Bharathi P and Abijith Sankar,” Student Placement Analyzer: A Recommendation System Using Machine Learning,” 2017 International Conference on Advanced Computing and Communication Systems (ICACCS -2017), Jan. 06 – 07, 2017, Coimbatore, INDIA
- [15] Ravi Kumar Rathore and J. Jayanthi,” Student prediction system for placements training using fuzzy inference system. ICTACT Journal of Soft Computing, April 2017, VOLUME: 07, ISSUE: 03
- [16] Ankita Kadambande, Snehal Thakur, Akshata Mohol, Prof A.M.Ingole,” Predict Student Performance by Utilizing Data Mining Technique and Support Vector Machine”, International Research Journal of Engineering and Technology (IRJET), Volume: 04 Issue: 05 | May -2017
- [17] Vinicius Gottin, Haydée Jiménez, Anna Carolina Finamore, Marco A. Casanova, Antonio L. Furtado and Bernardo P. Nunes,” An Analysis of Degree Curricula through Mining Student Records”, 2017 IEEE 17th International Conference on Advanced Learning Technologies, 2017 Pages: 276 – 280
- [18] Dijana Oreški, Mario Konecki and Luka Milić,” Estimating profile of successful IT student: data mining approach”, 2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO),2017, Pages: 723 - 727
- [19] K.Nasaramma, M.Bangaru Lakshmi, D.kiranmayi,” Prediction and Comparative Analysis of Students Placements Using C4.5 &C5.0”, K.Nasaramma et al, / (IJCSIT)

International Journal of Computer Science and Information Technologies, Vol. 8 (3) ,
2017,367-368

- [20] Jin Soung Yoo, Yei-Sol Woo and Sang Jun Park, "Mining Course Trajectories of Successful and Failure Students: A Case Study," 2017 IEEE International Conference on Big Knowledge (ICBK), 2017, Pages: 270-275
- [21] A. Parkavi, K. Lakshmi," Predicting the Course Knowledge Level of Students using Data Mining Techniques," 2017 IEEE International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), 2017, Pages: 128 – 133
- [22] Wongpanya Nuankaew, Praty Nuankaew, Sittichai Bussaman, Passakorn Tanasirathum, "Hidden Academic Relationship between Academic achievement and Higher Education Institutions," 2017 International Conference on Digital Arts, Media and Technology (ICDAMT), 2017, Pages: 308 - 313.
- [23] Febrianti Widyahastuti and Vinay Utami Tjhin,"Predicting students performance in final examination using linear regression and multilayer perceptron," 2017 10th International Conference on Human System Interactions(HIS), 2017, Pages: 188-192
- [24] Mohammad-Hossein Nadimi-Shahraki and Mehdi Mansouri," Hp-Apriori: Horizontal parallel-apriori algorithm for frequent itemset mining from big data," 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017, Pages : 286 – 290
- [25] Jun Yang, Haoxiang Huang and Xiaohui Jin," Mining Web Access Sequence with Improved Apriori Algorithm," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC),Year: 2017, Volume: 1, Pages: 780 – 784