International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR)
ISSN(P): 2249-6831; ISSN(E): 2249-7943
Special Edition, Apr 2015, 1-10

MINING INTERESTING AND USEFUL PATTERNS FROM STUDENT ASSESSMENT DATA

VINAY JAIPRAKASH YADAV

Shah And Anchor Kutchhi Engineering College, Mahavir Education Trust Chowk, W. T Patil marg, Next to Duke's Company, Chembur, Mumbai, Maharashtra, India

ABSTRACT

The motivation for this project came from the difficulties faced by the educational institution to analyze the Students performance effectively and then arrive at important results. The main idea behind this project is to minimize the unnecessary clerical time taken by Educational institutions from time to time to analyze the Students performance and that too in an ineffective way by replacing this trend with a well defined and easy system that provides us with more accurate results. In this Project we have collected a large amount of student data around 2 lakh records from the university database that have been generated over the period of the years by some or many educational Institutions combined. In this Project the primary focus is to make the task of the tutors and other higher authorities easy by showing them graphically the changing scenarios of the education. This project can be used to see the data associated with the students growth in the performance very closely by the tutors and based upon the observation the tutors or the higher authorities can take some important decision for improving the educational scenarios so that there is growth in students performing Behavior in their academics and also improving the standards and the needs of the students and areas or subjects in which in past the results were not impressive.

KEYWORDS: Educational Institutions, Academics, Changing Senarios, Tutors, Higher Authorities, Primary Focus, university Database

INTRODUCTION

Universities Today are operating in a very competitive and complex environment. This is mainly because of the increase in the transport and other facilities and the students are travelling from different educational background to take admission and learn from various reputed institutes. As a result the modern universities have a very desperate requirement of analyzing their performance, To analyze uniqueness and also build a planned structure for future development and actions. University management need to pay attention to the profile of the new admitted students, the academic behavior of the particular type of students based on their educational backgrounds, the characteristics of the students based on the received data. This will help the Institutes to take very important and necessary decision like which professors are more efficient for which subjects, which professors can be more expressive for some particular kinds of students based on their behavior and educational background, The rules needs a change based on the changing scenario, Which facilities should be enhanced and should be provided more efficiently. This will also help the universities to take important business decisions based on the arrived conclusion. On the other hand, the use of Internet in education has created a new context known as e-learning or web-based education in which large value of information about teaching–learning interaction are generated and made available. All this information provides a rich educational data. Educational data mining seeks to use these data

repositories to better understand learners and learning, and to improve computational approaches that combine data and theory to enhance practice to benefit learners.

LITERATURE REVIEW

- Keel (Knowledge Extraction Based on Evolutionary Learning):

This tool in Datamining can be used for importing and exporting the data from the KEEL format to other Formats and Vice-versa. In this software we need to select the type of the algorithms such as classification, regression etc. In educational scenario it can be used to design experiments that can be used to display the learning process of students or certain education data model. This Tool also has online mode offering educational support to learn the operation of the algorithm specified by the user[3].

- Moodle Datamining Tool:

In this they have developed a specific Moodle data mining tool oriented for use by on-line instructors. It has a simple interface (see Figure 1) to facilitate the execution of data mining techniques. Integration of this tool into the Moodle environment itself. In this way, instructors can both create/maintain courses and carry out all data mining processing with the same interface. they can thus directly apply feedback and results derived by data mining back into Moodle courses. We have implemented this tool in Java using the KEEL framework [3] which is an open source framework for building data mining models. The instructors have to created training and test data files starting from the Moodle database. Select one or several courses and one Moodle table (mdl_log, mdl_chat, mdl_forum, mdl_quiz, etc.) or create a summary table. Then, data files will be automatically preprocessed and created. Next select one of the available mining algorithms and the location of the output directory over a summary file and the decision tree is obtained. We can see that the results files (traand. test files with partial results and .txt file with the obtained model) appear in a new window [2].

- The Author heresuggests EDM subjects, tasks and applications dealing with the assessment of the student's learning performance, applications that provide course adaptation and learning recommendations based on the student's learning behavior, approaches dealing with the evaluation of learning material and educational web-based courses, applications that involve feedback to both teacher and students in e-learning courses, and developments for detection of a typical students' learning behaviors[10].
- The author has tried to divide the students into three categories based on the risk that they will pass the current academic year or not. High risk, Medium Risk and Low risk, the students with high risk needs to be given more time and guidance. The data was collected manually and by distributing a Questionnaire to all the students of academic year 2003 and 2004 of Belgian university and a French university. Each student had 42 questions to be answered and a total of 148 variables. Based on the answers of the students the database was created. Later it was found that all students who performed bad in January session failed at the end of the year and those who obtained total marks of 70% in January session passed at the end of the year. Thus approach here is more manual and effective for small number of students [8].

Project Implementation

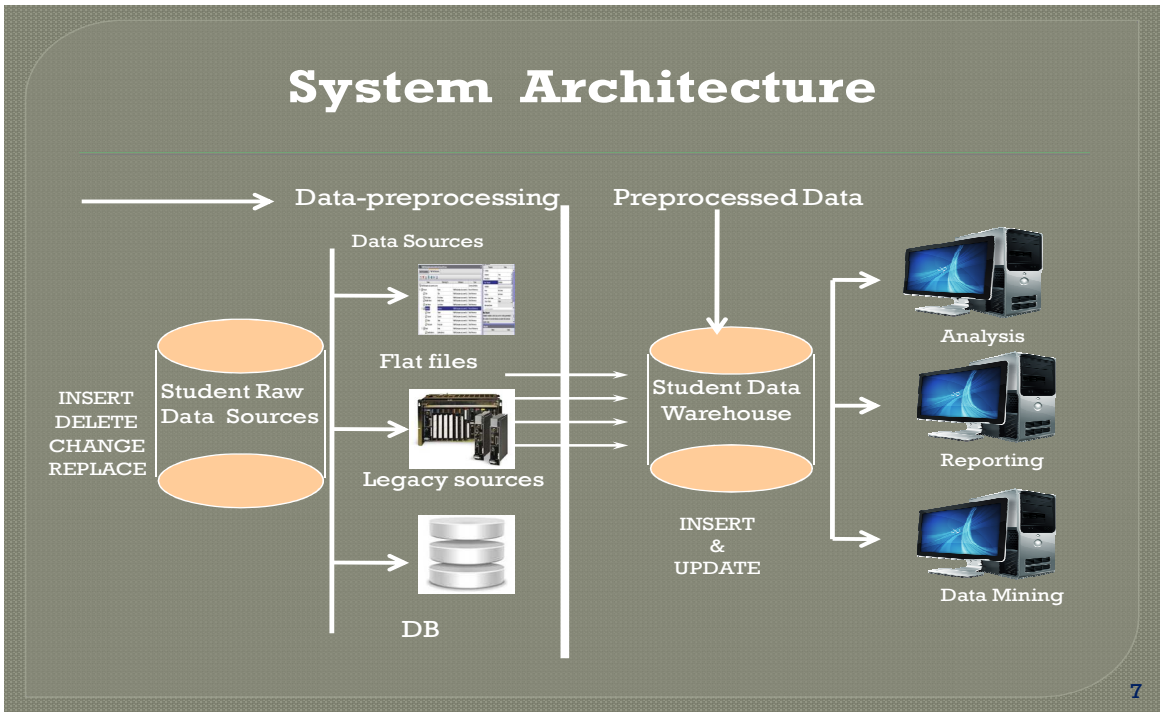


Figure 1: Project Implementation Strategy

FOLLOWING ARE THE STEPS OF THE PROJECT IMPLEMENTATION STRATEGY

Preprocessing

Data Transformation Data Transformation is the process in which the data is transformed from one format to another and also to Consolidate into format standard for the System. Initially the Exam data of the students was available in the .dbf format which was Converted from .dbf format to .xls format. This .xls format document is used to load the data into the database to create Datawarehouse. The Conversion from .dbf to .xls was done using the software called DBF converter.

Data Cleaning

- **Deleting the Duplicate Records**

After loading the data from .xls file into the Database the data had many replicas. The Problem with the replica was that there were many students for which the entry of the same record for the same semester and same subject were entered twice or thrice and as a result when the query was fired into the database the results for the marks were added for all the three or two duplicate entries and as result of this the marks that should have been for one entry only were added twice or thrice and so the results had errors arriving because of the aggregation of the duplicate records. Let us say we have students information as follows in the tabular form which has the information like students roll no, name, semester, subject, marks, address, year and phone no have been included. We have duplicate and exactly same records for some students. Because of which the marks for this students get added up in the database And as result the actual marks are added multiple times.

Table 1: Dummy Example of Student Details

Roll no	Name	Semester	Subject	Marks	Address	Year	Phone no
1	Ram	5	Math	95	Bhandup	2003	2566078
2	Daniel	4	Java	67	Thane	2004	2566987
1	Ram	5	Math	95	Bhandup	2003	2566078
4	Rajesh	6	DSA	56	Mulund	2005	2591089
2	Daniel	4	Java	67	Thane	2004	2566987
6	Pankaj	3	MEIT	48	Kalyan	2008	2591087
7	Prakash	5	OS	74	Nerul	2009	2591007

In the above table we have duplicate records for the Roll no 1 and 2 having the name as Ram and Daniel now suppose when we try to find the students who scored highest marks in let us say in year 2003 and if Ram is one of the highest marks getter in that year for maths subject, Then his actual marks that should be reflected should be 95, but because of the duplicate record both the marks in each record gets added up that is 95+95 =190 and getting 190 marks out of 100 is not possible and that's how the error occurred. As result it became very important to remove the duplicate records in order to get the correct result for a particular year and a particular semester.

Solution- Select *, count(*) as cnt

From dbo.Fact_marks

Having Count(*) >1

Group by Roll no, Name, Semester, Subject, marks, Address, Year, Phone no. This above command can be used to find the duplicate records and then. In order to delete this records we can replace the select statement with The delete statement

- **Removal of All the Special Characters**

Initially when the .dbf file was converted to .xls the .xls file had many special characters like "COMMAS(,) and DOLLARS($)" and so as a result for many columns were the data represented the marks of the student was actually a fact but because there were some special characters the entire data type of that column changed to String. As a result when we queried the database the marks were treated as String. So we removed all the special characters manually

DATAWAREHOUSE DESIGN FOR THIS PROJECT

Software used- IBM Cognos Framework Manager-10, SQL SERVER 2005, VM warehouse

Operating System Used- Windows XP service pack-3

There are Two Layers Present in this Design Namely

Database layer

Dimensional Layer

Database layer snapshot-

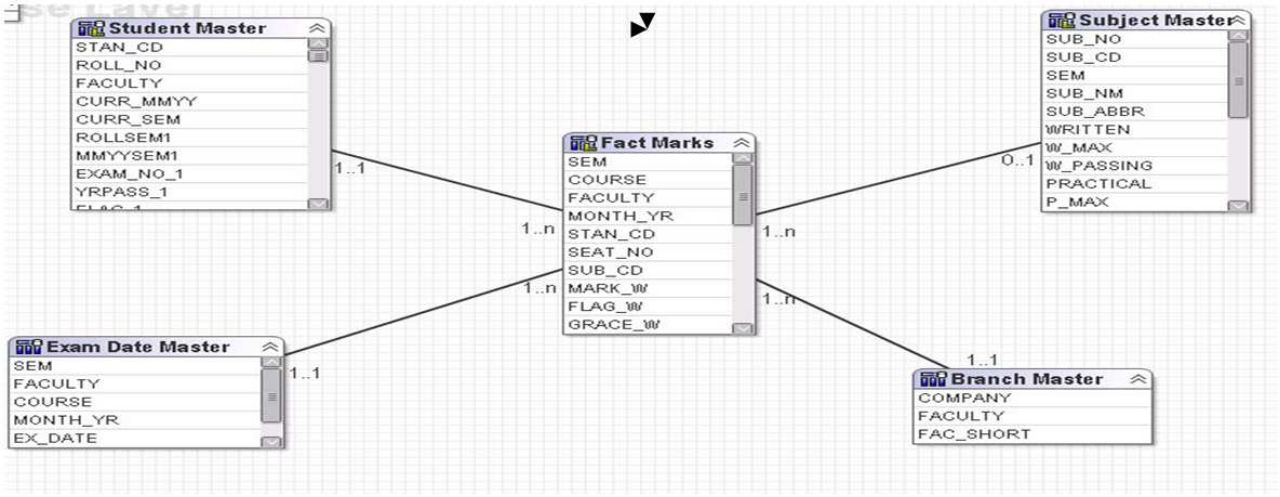


Figure 2: Database Layer

Database Layer Description-

Fact Table – The fact table in the above diagram called "Fact Marks" is present in the centre. This fact table consist of numeric data or quantitative value of the students for 18 years From 1996 to year 2014, Thus the fact table has very huge historical data and this data can be used to analyze Students performance very effectively.

Dimensional Table-Dimension table is used to describe the entities in detail we have three Four Dimension Tables namely Student Master, Exam Date Master, Branch Master, Subject Master. Let us describe the dimensions one by one.

- **Student Master**- Describes the Student in Detail Like the Semester, Subject Name, Theory Marks, Practical Marks, Oral Marks, It also describes history of the student like the Father's Name, Mother's Name, Previous attended College etc.
- **Subject Master-** Describes the subject in detail like Sub No, Subject Name, Maxium Theory, Practical, Oral Marks, Minimum Marks to pass theory, Practical, Oral etc.
- **Exam Date Master**- It describes the exam related parameters in detail like company, Faculty etc.
- **Branch Master-** It describes the Branch related detail of all the Branches.

In the above Figure we have "Fact marks" Fact table in the center and we have dimension Tables connected to the fact table by using the connecting parameter or fields. This example represents a Simple Star Schema.

Dimensional Layer Snapshot:

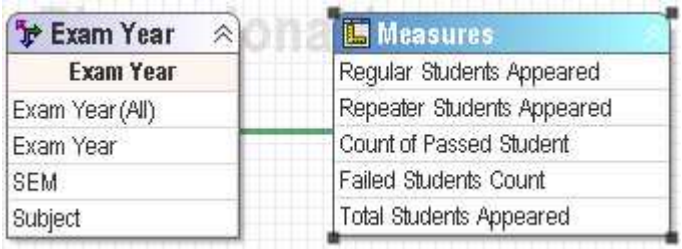


Figure 3 Dimensional Layer

Dimensional Layer Description: The dimensional layer is used by us to create the hierarchy to use the drill up and drill down facility.

Drill Down is a phenomenon in which we move from the summary into a greater detail and try to analyze the data in its more finer form.

Drill up is a phenomenon in which from the detailed view we move towards more summarized view. It is exactly opposite to that of the drill up concept.

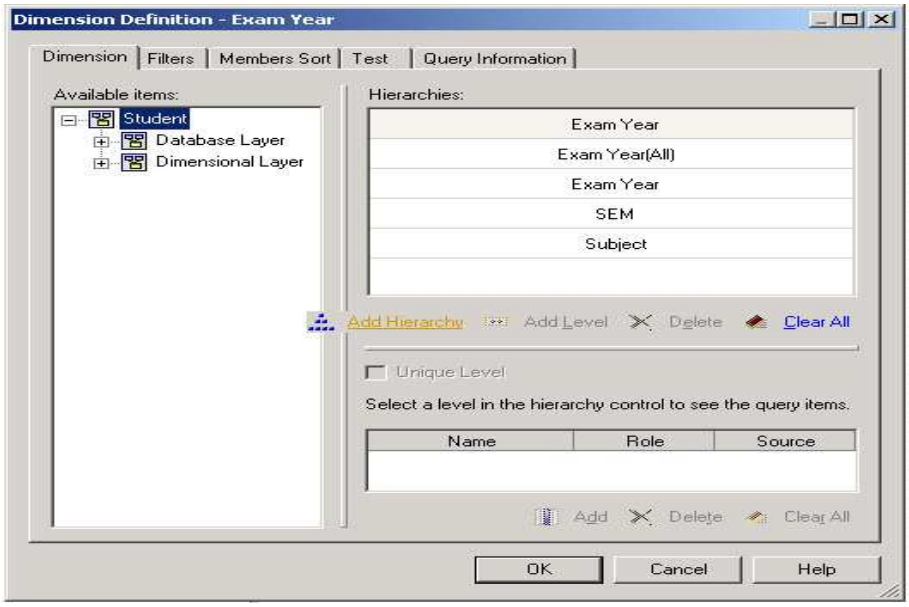


Figure 4: Dimensional Drill Hierarchy

In this project we are using year as the entity to drill down. As you can see in the above figure :

- The top most Hierarchy is the year
- The next Hierarchy is the semester. As we drill down from year we go to the semester.
- The Next Hierarchy is the subject as we drill down the Semester we move towards more finer details that is the individual subjects.
- We can add many such details depending upon the availability of the data.
- The advantage of the Dimensional layer is that it gives us both the summarized view as well as the detail view of the entities of which the hierarchy is formed.

REPORTING AND ANALYSIS

Software Used – IBM Cognos 10

- **Operating system used – Windows Xp Service pack-3**

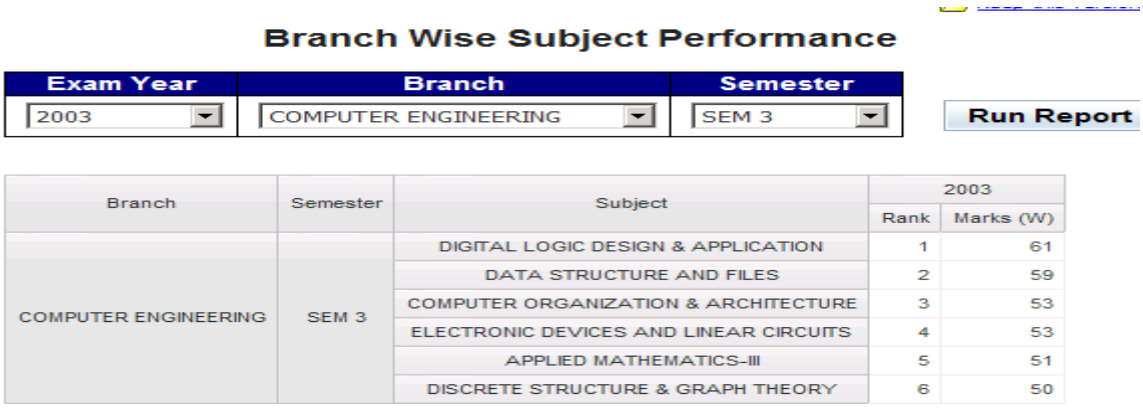
This is the final stage of the project in this after the Datawarehouse is created the Datawarehouse is used to fetch the data and used by the reporting tool to show the graphic view of the result derived or an important representation of the conclusion of the data fetched.

Various reports can be developed from the Datawarehouse data and various views like graphs, charts, bars, circular balls, pie charts, lists etc can be used to show the graphical view of the data or the results that can derived some of the example reporting ideas are as follows

Report-1- Top Five Students

- This Report shows the Branch wise Performance of the Top five Students for a particular year and particular Semester in a summarized manner.
- We have also used here a concept called “Drill through definitions”. Which is used to get a detailed level information of each student in the top five cadre.
- By using Drill through definitions we navigate from summarized level to detailed Level of information.
- Drill through definitions allows us to pass the parameter from one report to another Report.

Report-2- Branch Wise Student Performance

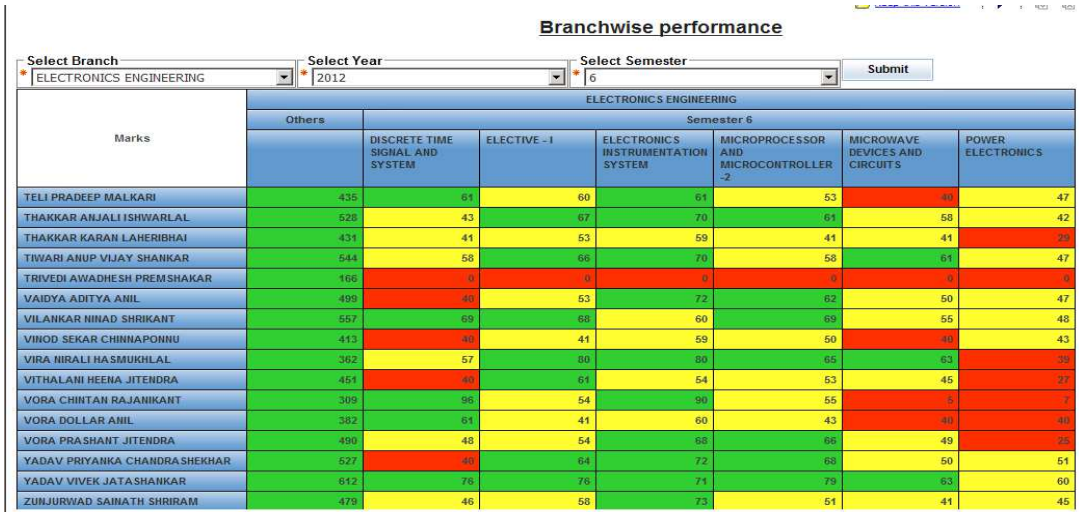


Branch	Semester	Subject	2003	
			Rank	Marks (W)
COMPUTER ENGINEERING	SEM 3	DIGITAL LOGIC DESIGN & APPLICATION	1	61
		DATA STRUCTURE AND FILES	2	59
		COMPUTER ORGANIZATION & ARCHITECTURE	3	53
		ELECTRONIC DEVICES AND LINEAR CIRCUITS	4	53
		APPLIED MATHEMATICS-III	5	51
		DISCRETE STRUCTURE & GRAPH THEORY	6	50

Figure 5: Branch wise Subject Performance

- This Report shows the Branch wise average performance of each particular Subject, for a particular year and for a particular semester.
- This report also has column called rank which takes the value dynamically depending upon the performance of the Students in that subject.
- Thus the subject which has highest average marks takes the rank =1And as the average marks for each subject decreases the rank also decreases.

Report-3- Branch Wise Performance of All Students



Branchwise performance

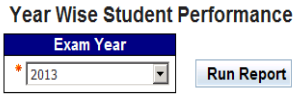
Select Branch: ELECTRONICS ENGINEERING | Select Year: 2012 | Select Semester: 6 | Submit

Marks	ELECTRONICS ENGINEERING						
	Others	Semester 6					
		DISCRETE TIME SIGNAL AND SYSTEM	ELECTIVE - I	ELECTRONICS INSTRUMENTATION SYSTEM	MICROPROCESSOR AND MICROCONTROLLER -2	MICROWAVE DEVICES AND CIRCUITS	POWER ELECTRONICS
TELI PRADEEP MALKARI	435	61	60	61	53	40	47
THAKKAR ANJALI ISHWARLAL	528	43	67	70	61	58	42
THAKKAR KARAN LAHERIBHAI	431	41	53	59	41	41	29
TIWARI ANUP VIJAY SHANKAR	544	58	66	70	58	61	47
TRIVEDI AWADHESH PREMSHAKAR	166	0	0	0	0	0	0
VAIDYA ADITYA ANIL	499	40	53	72	62	50	47
VILANKAR NINAD SHRIKANT	557	69	68	60	69	55	48
VINOD SEKAR CHINNAPONNU	413	40	41	59	50	40	43
VIRA NIRALI HASMUKHLAL	362	57	80	80	65	63	39
VITHALANI HEENA JITENDRA	451	40	61	54	53	45	27
VORA CHINTAN RAJANIKANT	309	96	54	90	55	5	7
VORA DOLLAR ANIL	382	61	41	60	43	40	40
VORA PRASHANT JITENDRA	490	48	54	68	66	49	25
YADAV PRIYANKA CHANDRASHEKHAR	527	40	64	72	68	50	51
YADAV VIVEK JATASHANKAR	612	76	76	71	79	63	60
ZUNJURWAD SAINATH SHRIRAM	479	46	58	73	51	41	45

Figure 6 Branch Wise Performance of All Students

- This report shows in Detail the performance of each student for a particular branch, for a particular year and for a particular Semester.
- The Green Block or color Indicates the Student is getting marks greater than equal to 60 marks.
- The Yellow Block indicates the student is getting marks between 40 and 60.
- The Red Block Indicates the student is getting marks less than or equal to 40.

Report-4 – Year wise Student Performance (Drill up, Drill Down)



Year Wise Student Performance

Exam Year: 2013 | Run Report

SEM 3	Total Students Appeared	Regular Students Appeared	Repeater Students Appeared	Count of Passed Student	Failed Students Count
ENGINEERING MATHEMATICS	32	0	32	15	17
BASIC OF ELECTRONICS CIRCUITS	32	0	32	24	8
DIGITAL SYSTEM DESIGN-I	32	0	32	28	4
ELECTRICAL NETWORK ANALYSIS AND SYNTHESIS	32	0	32	23	9
CONTROL SYSTEM	32	0	32	28	4
PRESENTATION AND COMMUNICATION TECHNIQUES	32	0	32	0	32
APPLIED MATHEMATICS - III	45	0	45	21	24
DATA STRUCTURE AND ALGORITHMS	45	0	45	45	0
ELECTRONIC DEVICES AND CIRCUITS	45	0	45	33	12
DIGITAL LOGIC DESIGN AND APPLICATIONS	45	0	45	34	11
G U I AND DATABASE MANAGEMENT	45	0	45	42	3
COMMUNUICATION & PRESENTATION TECHNIQUES	45	0	45	0	45
	77	0	77	77	0
Overall - Total	539	0	539	370	169

Figure 7: Subject Wise Drill Down Summarized Report

- This report is used to view the year wise performance of the students in a summarized manner. This report gives us the overview of the students appearing the exam and the number of students passing the exam in that year.
- This report can also be used to find the students who Appeared for the first time and the students who are repeaters.

- This report gives us the facility of drilling down the Order and then from year if we drill down we will go to the semester. From semester if we drill down we will move towards the individual subjects in that semester and this report shows us the performance of the students for each individual subject regarding the number of students appearing for that subject and number of students passing the subjects. Also the students who appeared for the first time and the students who are repeaters.

CONCLUSIONS

Thus we have successfully Loaded the historical data of the students in the system in a standard format and created a Datawarehouse by using Dimensional modeling from the Raw student data and also created the relationship between the tables to join them in order to obtain some meaningful output. This has leaded us to create two layers in Datawarehouse called database layer and dimensional layer. Dimension layer was used to show drill down and drill up facility and their importance. In the end we have also created some of the reports to get both the summarized result as well as the detailed result. Thus we have achieved our objective of showing some meaningful patterns in changing Educational scenario for a Educational university or college. These patterns can be used effectively to guide the Students to make progress in their academics and can also help Teachers and professors to make necessary changes in their own teaching Styles. This kind of model can be used in various e-learning systems to continuously study the rapid dynamic behavior of the students and then accordingly suggest important changes in student academic and study pattern.

Future Scope

This Model can be enhanced in the future ahead to add a predictive model. The current model can show graphically the performance of the students in the reports that we generate. The current model helps us to analyze how the data is changing over the period of years and thus helping us to keep a check on the students exam data and come out with useful and important conclusion, but this model is not helping us to predict the future performance of the students based on the current or the past data that we are seeing, Thus adding a predictive analysis model will add a very impressive feature to this project and will help us overcome the problem related to the students performance before the problem actually occurs.

REFERENCES

1. Baker, R. S. J. d. (in press) Data Mining for Education. To appear in McGaw, B., Peterson, P., Baker, E. (Eds.) *International Encyclopedia of Education (3rd edition)*. Oxford, UK: Elsevier
2. Cristobal Romero, Member, IEEE, and Sebastian Ventura, Senior Member, IEEE,‖ *Educational Data Mining: A Review of the State of the Art"*, IEEE Transactions on systems, man, and cybernetics—part c: applications and reviews, vol. 40, no. 6, November 2010.
3. J. Alacala-fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez and F. Herrera, ―*KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework"*, April 20 2010, J. of Mult.-Valued Logic & Soft Computing, Vol. 17, pp. 255–287
4. http://moodle.org.
5. Dougherty, J., Kohavi, M., Sahami, M. ―*Supervised and Unsupervised Discretization of Continuous Features"*. Twelfth international Conf. on Machine Learning, San Francisco, 1995. pages.194–202.

6. Dr. Kanak Saxena et al, *"Result Analysis Using Various Pattern Mining Techniques:- A Recommendation to Strengthen the Standard of Technical Education"*, International Journal on Computer Science and Engineering Vol.1(3), 2009, pages 235-238.

7. J. F. Superby, J.-P. Vandamme, *—Determination of factors influencing the achievement of the first-year university students using data mining methods"*, Production and Operation management Department Catholic University of Mons Chausee de Binche 151, 7000 mons Belgium pages 1-8.

8. Dimensional Modelling: In a Business Intelligence Environment By Chuck Ballard, Daniel M. Farrel, Amit Gupta, Carlos Mazuela, Stanizlav Vohnik(Redbooks). Literature Cited Vinay J. Yadav Reg. No: 10485 Page 102

9. F. Castro, A. Vellido, A. Nebot, and F. Mugica, ―Applying data mining techniques to e-learning problems,‖ in *Evolution of Teaching and Learning Paradigms in Intelligent Environment* (Studies in Computational Intelligence), vol. 62, L. C. Jain, R. Tedman, and D. Tedman, Eds. New York: Springer-Verlag, 2007, pp. 183–221.

10. Ryan S. J. d. Baker, Adriana M. J. A.,‖ *Labeling Student Behavior Faster and More Precisely with Text Replays"*, Human Computer Interaction Institute, Carnegie Mellon University, The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings, pages 38- 45.

11. Agathe Merceron and Kalina Yacef, *—Interestingness Measures for Association Rules in Educational Data"*, The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings, pages 57-66.

12. Cesar Vialardi, Javier Bravo and Leila Shafti, Alvaro Ortigosa, *"Recommendation in Higher Education using datamining Techniques"* Universidad de lima, Educational Datamining 2009, pages 1-10.

13. Dr. Varun Kumar, Anupama Chadha, *—An Empirical Study of the Applications of Data Mining Techniques in Higher Education"*, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, March 2011.

14. Philip I., Pavlik Jr, Hao Cen, Lili Wu, and Kenneth R. Koedinger, *—Using Item-type Performance Covariance to Improve the Skill Model of an Existing Tutor"*, The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings, pages 77-85.

15. Manolis Mavrikis, *"Data-driven modelling of students' interactions in an ILE"*, The 1st International Conference on Educational Data Mining Montréal, Québec, Canada, June 20-21, 2008 Proceedings, pages 87-94.