Volume 4, Issue 10, October 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Descriptive Data Mining for Educational Environment

Dr. Pramod Kumar Rai
Computer Centre, A. P. S. University,
Rewa (M.P.) India

Abstract: ***In recent years, there has been increasing interest in the use of data mining to investigate scientific questions within educational environment. Such type of investigation comes under Educational Data Mining also referred to as EDM. Educational data mining focuses around the e development of methods for making discoveries within the unique kinds of data that come from educational organizations. EDM use these methods to better understand students and the organizations which they learn in. In this paper, experiments have been performed on university examination data. The new knowledge discovered from this may help in improving decision-making procedure of university.***

Keywords: KDD, EDM, DM

I. INTRODUCTION

Data mining (DM) is the process of extracting interesting (non-trivial, implicit, previously unknown and potentially useful) information or patterns from large information repositories such as: relational database, data warehouses, XML repositories, etc. [CHY96]. Also data mining is known as one of the core processes of knowledge discovery in database (KDD). Data mining tasks can be divided into six broad categories – Classification, Estimation, Prediction, Clustering, Association and Description.

Knowledge discovery in database [FPSU96] is the process of identifying useful and ultimately understandable structure in data. This process involve selecting or sampling data from a operational data / data warehouse, cleaning or preprocessing it, transforming or reducing it (if needed), applying data mining techniques to produce a structure, and then evaluating the derived structure.

Knowledge Discovery is a bottom-up approach that starts with the data and tries to get it to tell us something we didn't already know. Knowledge discovery can be either *directed* or *undirected*. We use undirected knowledge discovery to recognize relationship in the data and directed knowledge discovery to explain those relationships once they have been identified. Clustering and association rule-mining comes under undirected knowledge discovery. In undirected knowledge discovery the data mining tool is simply let loose on the data in the hope that it will discover meaningful structure. One common use of undirected knowledge discovery is market basket analysis that asks, "What items sell together". Another application is clustering, where groups of records are assigned to the same cluster if they have some thing in common.

II. EDUCATIONAL DATA MINING

Educational data mining methods often differ from methods from the broader data mining literature. In EDM the multiple levels of meaningful hierarchy in educational data are required to be exploited. Issues of time, sequence, and context also play important roles in the study of educational data.

Data mining techniques are already in use in the private sector. Many of the data mining techniques used in the corporate world, however, are transferable to educational environment.

Data mining in educational environment is a recent research field. A survey on educational data mining between 1995 and 2005 is presented in [RV07]. Also, educational data mining used by [MKKP03] to predict students final grade using data collected from web based system. [BD05] used educational data mining to identify and then enhance educational process in higher educational system which can improve their decision making process.

The application of data mining in an educational environment has been considered by [ET05]. The relationship between students university entrance examination results and their success was studied using cluster analysis and k-means algorithm techniques.

Shyamala and Rajagopalan [SR07], in their study have presented the work of data mining in predicting the drop out feature of students. They applied decision tree technique to choose the best prediction and clustering analysis. Baker and Yacef [BY09] had reviewed the history and current trends in the field of Educational Data Mining and reviewed the key applications of EDM methods.

III. DATA MINING MODELING & EXPERIMENTS

For experimental purpose the examination database of Awadhesh Pratap Singh University, Rewa (MP) India of undergraduate students (B.A., B.Sc. and B.Com.) for the years 2004 to 2008 has been selected. By performing so many preprocessing steps, we have transformed the data in single file graduate with 1,75,216 instances.

The prepared data and attributes are used as the input for the development model. The outcome of the model can be used for academicians and decision makers. The knowledge obtained from data mining techniques gives the managerial decision makers the useful information for decision making. The models are classified in two main categories : (i) descriptive and (ii) predictive models.

(i) Descriptive model describes the data set in a concise and summarized manner and presents the interesting general properties of the data. It explains the patterns in existing data, which may be used to guide decisions.

(ii) Predictive model predicts behavior based on historic data and uses data with known results to build a model that can be later used to explicitly predict values for different data.

For descriptive data mining modeling we used the graduate dataset containing 1,75,216 instances and the dataset is analyzed. The frequency tables of dataset is given in Table 1 to Table 4. For visualization the corresponding bar charts are given in Figure 1 to Figure 3. Other analysis tables are given in Table 5 to 8.

Table 1 : Class wise Frequency Table of graduate dataset

CLASSCODE

	Class	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	BA-I	42077	24.0	24.0	24.0
	BA-II	32812	18.7	18.7	42.7
	BA-III	29315	16.7	16.7	59.5
	BSc-I	20049	11.4	11.4	70.9
	BSc-II	14567	8.3	8.3	79.2
	BSc-III	12315	7.0	7.0	86.3
	BCom-I	10416	5.9	5.9	92.2
	BCom-II	6637	3.8	3.8	96.0
	BCom-III	7028	4.0	4.0	100.0
	Total	175216	100.0	100.0	

Table 2 : Examination Year wise Frequency Table of graduate dataset

EYEAR

	Exam Year	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	2004	37961	21.7	21.7	21.7
	2005	34950	19.9	19.9	41.6
	2006	30898	17.6	17.6	59.2
	2007	34295	19.6	19.6	78.8
	2008	37112	21.2	21.2	100.0
	Total	175216	100.0	100.0	

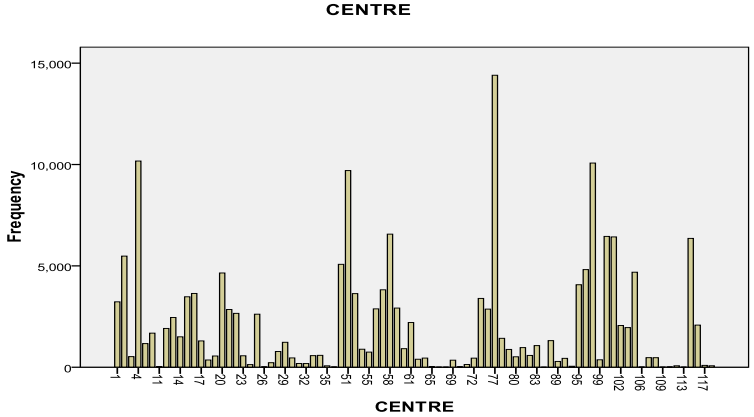

Figure 1 : Examination Centre/College wise Frequency chart

Table 3: Sex wise Frequency Table

SEX

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female Unmarried	90637	51.7	51.7	51.7
	Female Married	1402	.8	.8	52.5
	Male	83177	47.5	47.5	100.0
	Total	175216	100.0	100.0	

Table 4: Caste wise Frequency Table

CASTE

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	General	117247	66.9	66.9	66.9
	SC	16126	9.2	9.2	76.1
	ST	9394	5.4	5.4	81.5
	OBC	32449	18.5	18.5	100.0
	Total	175216	100.0	100.0	

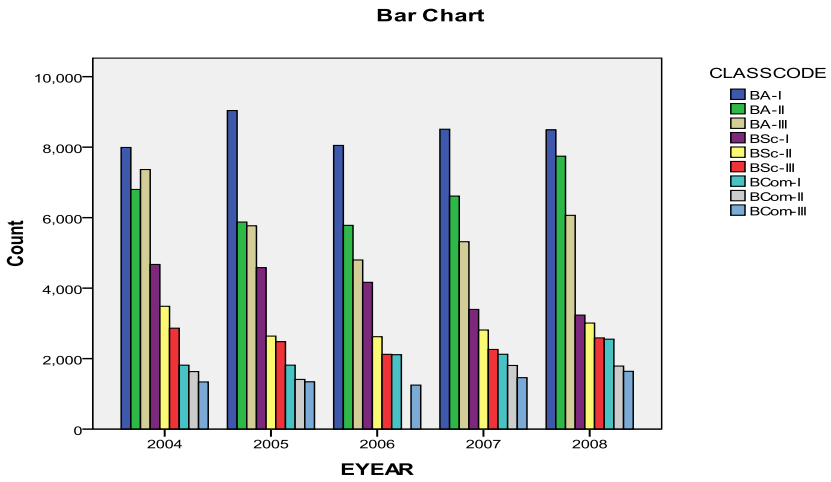

Figure 2 : Cross tabulation chart Examination Year(EYEAR) and Class Code

Table 5: Cross tabulation Table Category and Student Result

CATEGORY * RESN Crosstabulation

			RESN			
			Pass	Supplementary	Fail	Total
CATEGORY	Regular	Count	82192	16794	18319	117305
		% within CATEGORY	70.1%	14.3%	15.6%	100.0%
	Ex-Student	Count	2912	1458	3183	7553
		% within CATEGORY	38.6%	19.3%	42.1%	100.0%
	Private	Count	26575	9147	10168	45890
		% within CATEGORY	57.9%	19.9%	22.2%	100.0%
	Supplementary	Count	2753	0	1712	4465
		% within CATEGORY	61.7%	.0%	38.3%	100.0%
Total		Count	114432	27399	33382	175213
		% within CATEGORY	65.3%	15.6%	19.1%	100.0%

Table 6: Cross tabulation Table Sex and Student Result

SEX * RESN Crosstabulation

			RESN			
			Pass	Supplementary	Fail	Total
SEX	Female Unmarried	Count	65488	13148	11998	90634
		% within SEX	72.3%	14.5%	13.2%	100.0%
	Female Married	Count	848	251	303	1402
		% within SEX	60.5%	17.9%	21.6%	100.0%
	Male	Count	48096	14000	21081	83177
		% within SEX	57.8%	16.8%	25.3%	100.0%
Total		Count	114432	27399	33382	175213
		% within SEX	65.3%	15.6%	19.1%	100.0%

Table 7: Cross tabulation Table Caste and Student Result

CASTE * RESN Crosstabulation

			RESN			
			Pass	Supplementary	Fail	Total
CASTE	General	Count	77205	17930	22109	117244
		% within CASTE	65.8%	15.3%	18.9%	100.0%
	SC	Count	10311	2694	3121	16126
		% within CASTE	63.9%	16.7%	19.4%	100.0%
	ST	Count	5511	1670	2213	9394
		% within CASTE	58.7%	17.8%	23.6%	100.0%
	OBC	Count	21405	5105	5939	32449
		% within CASTE	66.0%	15.7%	18.3%	100.0%
Total		Count	114432	27399	33382	175213
		% within CASTE	65.3%	15.6%	19.1%	100.0%

Table 8: Cross tabulation Table Examination Year, Sex and Student Result

EYEAR * SEX * RESN Crosstabulation

				SEX			
RESN				Un Female Unmarried	Mar. Fem Married	Male	Total
Pass	EYEAR	2004	Count	8860	131	9075	18066
			%	49.0%	.7%	50.2%	100.0%
		2005	Count	10307	153	8693	19153
			%	53.8%	.8%	45.4%	100.0%
		2006	Count	12907	119	9557	22583
			%	57.2%	.5%	42.3%	100.0%
		2007	Count	16014	160	11065	27239
			%	58.8%	.6%	40.6%	100.0%
		2008	Count	17400	285	9706	27391
			%	63.5%	1.0%	35.4%	100.0%
	Total		Count	65488	848	48096	114432
			%	57.2%	.7%	42.0%	100.0%

Supp	EYEAR	2004	Count	2862	69	4647	7578
			%	37.8%	.9%	61.3%	100.0%
		2005	Count	2566	49	3217	5832
			%	44.0%	.8%	55.2%	100.0%
		2006	Count	2250	25	1851	4126
			%	54.5%	.6%	44.9%	100.0%
		2007	Count	2195	32	2047	4274
			%	51.4%	.7%	47.9%	100.0%
		2008	Count	3275	76	2238	5589
			%	58.6%	1.4%	40.0%	100.0%
	Total		Count	13148	251	14000	27399
			%	48.0%	.9%	51.1%	100.0%
Fail	EYEAR	2004	Count	3599	111	8607	12317
			%	29.2%	.9%	69.9%	100.0%
		2005	Count	3078	93	6794	9965
			%	30.9%	.9%	68.2%	100.0%
		2006	Count	2435	20	1734	4189
			%	58.1%	.5%	41.4%	100.0%
		2007	Count	1136	21	1622	2779
			%	40.9%	.8%	58.4%	100.0%
		2008	Count	1750	58	2324	4132
			%	42.4%	1.4%	56.2%	100.0%
	Total		Count	11998	303	21081	33382
			% within EYEAR	35.9%	.9%	63.2%	100.0%

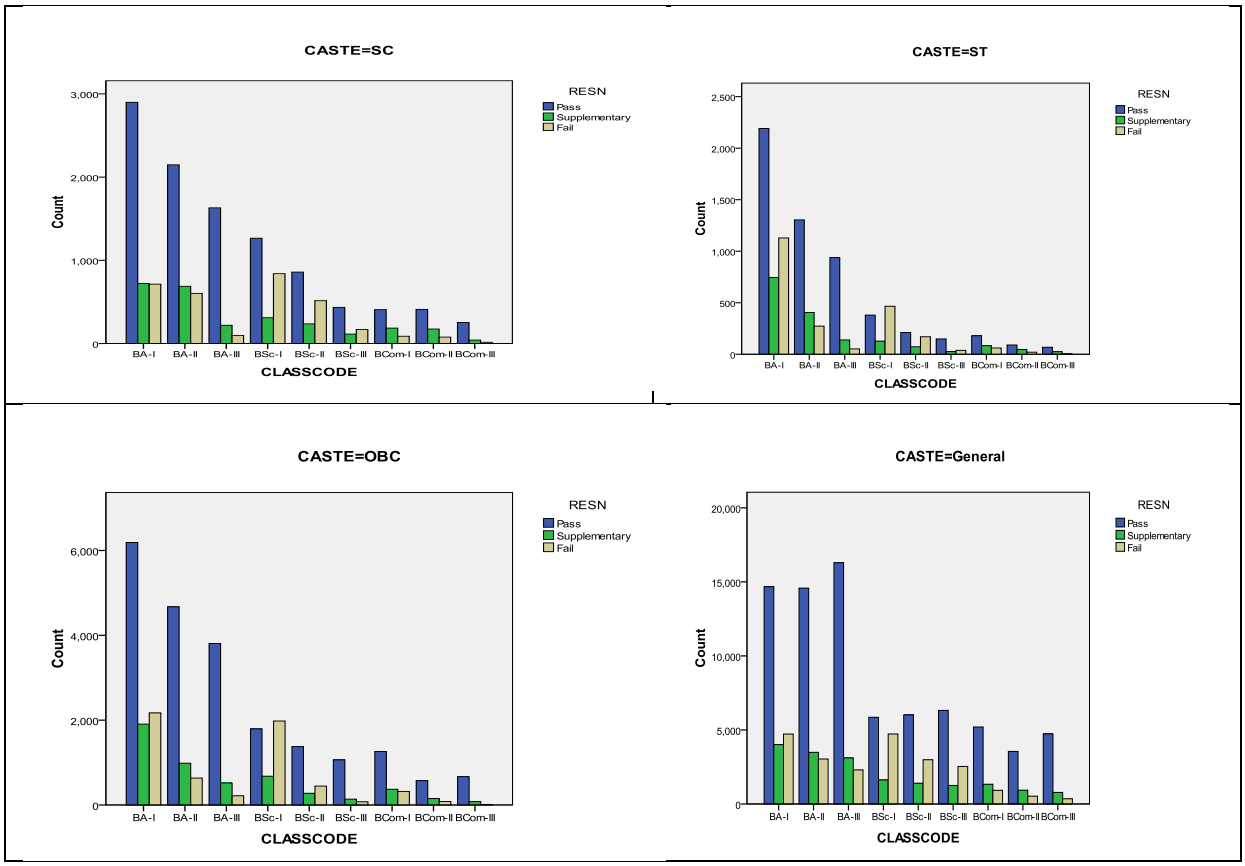

Figure 3: Cross tabulation chart Class Code, Student Result and Caste

IV. ANALYSIS OF EXPERIMENT

The analysis results of experiment are summarized below. It is proposed to further explore the dataset by using some other data mining techniques and tools.

- The dataset contains 59.5% students from BA classes, 26.8% students from B.Sc. classes and 13.7% students from B.Com. classes for the years 2004 to 2008. This indicates the preference of art subjects by the students.
- The percentage population of students in dataset for the year 2004 is 21.7%, for the year 2005 is 19.9%, for the year 2006 is 17.6%, for the year 2007 is 19.6% and for the year 2008 is 21.2%.
- The college/centre with highest number of enrollments are college code 4, 51, 77 and 98. These colleges have approximately 25% enrollment. These colleges are lead colleges of four different districts of the university jurisdiction. This indicates the student's enrollment preference for certain colleges.
- Many colleges have very poor enrollment (for example the college code 11,26, 36, 66, 67 etc.). The necessary steps for increase in enrollment are required by the college / university administration in these colleges to make them viable.
- The percentage of married females is less than 1%. There are total 52.5% females and 47.5% males. This clearly indicates the higher number of female enrollment in graduate classes in comparison to male students. One of the reason for less enrollment of male students may be movement of male students for other professional courses.
- The percentage of regular students is 66.9% and private students are 26.2%. This show the students prefer to study as fulltime regular students.
- The overall pass percentage is 65.3%. The pass percentage of regular category students is 70.1% and for ex-students pass percentage is 38.6%. This indicates the effect of regular classroom teaching to improve the student's skill.
- Female students have performed better than male students. Even the pass percentage of married females is better than male students.
- There are no significant variations in caste wise pass percentage.
- The pass percentage of second year students is almost better than first year and the pass percentage of third year students is almost better than second year for all classes. This indicates the improvement in the student skill in the next higher class for passing the university examination.
- The subject wise result in certain colleges is very poor. Theses colleges and subjects are identified through multiple response cross table. The college/university administration should think for the reasons and proper enhancement in the teaching resources/infrastructure related to the study of these subjects in the identified colleges. For example, the college code 5 and 6 requires the attention in Physics teaching to improve the result of enrolled Physics students in these colleges.

V. CONCLUSION

The present study is an attempt to enhance the traditional educational process via knowledge discovery through data mining technology. The advantages and suitability of this system in higher learning institution has been discussed. It also provides an opportunity to learn the existing area of study for data mining in the educational environment.

This is useful in educational environment to either support the current decision making or help to set new strategies and plans to improve the decision making procedures. For experimental purpose we have used university examination data of five years of undergraduate students. This study utilizes data mining in the field of education. The steps of the data mining process were carried out and explained in detail. The area of application was education, different from the usual data mining studies. The use of the data mining technique in the field of education may provide us more useful findings which can be used to increase the quality of education.

With this study we conclude that the use of DM techniques in educational environment is highly useful for the educational settings through the improvement in decision making. Even though the obtained results had been satisfactory, it is necessary to mention that for future works it will be necessary to use the data from other faculties and universities in order to know consistency and convergence.

ACKNOWLEDGEMENT

The author is thankful to Sri Anjesh Kumar of Govind Ballabh Pant Social Science Institute, allahabad for useful discussion and support. During the visit to the institute, the data mining tool SPSS was used with his help for the experiment.

REFERENCES

[BD05] Beikzadeh, M. and Delavari, N., "A New Analysis Model for Data Mining Processes in Higher Educational Systems", On the proceedings of the 6th Information Technology Based Higher Education and Training, 7-9 July 2005.

[BY09] Baker, R.S.J.D. and Yacef K., "The state of Educational Data Mining in 2009 : A review and Future Visions", JEDM, Vol. 1, 2009

[CHY96] Chen M., Han J. and Yu P.S., "Data Mining: An Overview from a Database Perspective". IEEE

Trans. On Knowledge and Data Engineering, Vol. 8, No. 6, pp. 866-883, December, 1996.

[ET05] Erdogan, S. Z. and Timor, M., "A Data Mining Application in a student database", Journal of Aeronautics and Space Technologies, Vol. 2, No 2, 53-57, July 2005.

[FPSU96] Fayyad U.M., Piatetsky-Shapiro G., Smyth P., Uthurusamy R. (Eds.): "Advances in Knowledge Discovery and Data Mining". MenloPark, CA: AAAI Pres / The MIT Press, 1996.

[MKKP03] Minaei-Bidgoli B., Kashy D., Kortemeyer G., Punch W., "Predicting Student Performance: An Application of Data Mining Methods with an Educational Web-Based System", In the Processing of 33rd ASEE/IEEE conference of Frontiers in Education, 2003

[RV07] Romero,C. and Ventura, S. ,"Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications, vol. 33, 135-146, 2007

[SR07] Shyamala, K. and Rajagopalan, S. P., "Mining student data to characterize drop out feature using clustering and decision tree technique", International Journal of Soft Computing, Vol. 2 (1), 150-156, 2007

