 ISBN: 978-81-927230-0-9

An Academic Analytics Perspective to Understand Impact of Social Ambience on Performance of Students

Parag Bhalchandra
School of Computational Sciences, SRTM University, Nanded, MS, 431606, India
srtmun.parag@gmail.com

Mahesh Joshi
School of Educational Sciences, SRTM University, Nanded, MS, 431606, India
maheshmj25@gmail.com

Hanmant Fadewar
School of Computational Sciences, SRTM University, Nanded, MS, 431606, India
fadewar_hsf@yahoo.com

Aniket Muley
School of Mathematical Sciences, SRTM University, Nanded, MS, 431606, India
aniket.muley@gmail.com

Santosh Khamitkar
School of Computational Sciences, SRTM University, Nanded, MS, 431606, India
s.khamitkar@gmail.com

Pawan Wasnik
School of Computational Sciences, SRTM University, Nanded, MS,431606, India
pawan_wasnik@yahoo.com

***Abstract* - In this study, our objective is to apply academic analytics over educational databases to get new insights for performance analysis otherwise which are invisible. Identify social variables that have inter-relation together which can affect student's performance. This paper demonstrates empirical work over student's dataset and academic insight was gained out with SPSS platform.**

Keywords : Educational Data Mining, Canonical Correlation Analysis, Multivariate Analysis, Performance Analysis

I. INTRODUCTION

The performance of student is merely taken as the studious efforts being out , hard works being taken and number of marks being obtained. However, the literature review has shown remarkable observations where performance is hampered by psychological, economical, personal, social and academic conditions. A review of literature has been conducted prior to hypothesis building. Later, it was decided to initiate a research study to investigate various social attributes shows correlation with the performance of student. Since, we aim for discovery of social attributes which affects performance of students; we primarily investigated literature across the globe to see what other people have done. To understand performance of a student, we underwent discussions with educationalist. The faculties from School of Educational Sciences, of our university had given us orientation on the same. We finally understood that the mere marks in final examination cannot be taken as main indicator of performance. The performance in broader sense is how well a student does in over all courses. For proper understanding the performance terminology, we primarily relied on the work of Shoukat Ali at al.[1,6], Graetz et al. [4,7], Considine and Zappala [5, 8], and Staffolani & Bratti [9,10] . This work is an example of a joint interdisciplinary work undertaken by three Schools of our University, viz, School of Computational Sciences, School of Mathematical Sciences and School of Educational Sciences. The primary objective of this, presented research study is to introduce Academic Analytics to collected dataset comprising various attributes of students. The secondary research objective is to investigate whether studious nature alone contributes to performance of students? If no, what other social variables are related to increase or decrease in performance?

The impact of social variables with the performance of students is invisible one. To make it visible, we have used data mining and statistical analysis approaches [1,15,16]. These two approaches together form the Academic Analytics. Data mining or Knowledge Discovery in Database (KDD) is the process of finding hidden and useful knowledge from large amount of data [1,2,3]. Now days, all educational organizations, institutions or universities have been computerized and they have database systems having all essential data from all vital parts. Using these databases, a real word dataset related to social, economical, personal, and performance related variables of our university students was created using questionnaire and progress reports. A student's dataset was created with 360 records and 46 fields by closed questionnaire method. Academic Analytical algorithms were implemented using SPSS software [4,17].

The database was built from two sources: previous examination's progress reports and our questionnaires. The questions in our questionnaire have predefined options [10,11,18] related to student's attributes as defined in Pritchard and Wilson [5, 18]. The questionnaire consisted of 43 closed type questions . The questionnaire were distributed to students and demonstrated for feedback. Thereafter, we have devised out the dataset of 360 student records and each record consists of 46 fields. Originally, there were 43 fields, equal to the total number of questions in questionnaire. Three additional fields were added for seeking information of students. Microsoft Excel 2007 software is used to record the dataset. Data set values like Yes / No were converted in to numeric values like 1 or 0. Other numerical codes in the range 0,1,2,3,4 …6 were also given depending upon the number of possible answers a question can have. A snapshot of questionnaire and the dataset is as given in Figure 1.

1	Course code	MSc (5) , MCA (6)
2	Your name	
3	Gender (sex)	Male (1) Female (0)
4	Marital status	Married (2) unmarried(3)
5	Age	
6	Home address	Urban(1) rural (2) foreign(3)
7	Mobile no.	
8	Personal email id	
9	Degree passer and percentage	General B.Sc. / (1) B.Sc.(computer CS) / (2) BCA / BCS/ (3) Other / (4) (5)
	Percentage	
10	Degree collage name	
11	Father's Education	Below or SSC/ (1) HSC/ (2) Graduate/ (3) Post Graduate/ (4) other (5)
12	Fathers job and annual income	Service / (1) Business/ (2) Agriculture/ (3) In house/ (4) Other/ (5)
	Income	0-1 lakh (1) , 1.1-2 lakh(2), 2.1-5 lakh(3) , 5lakh - above (4)
13	Mothers education	Below or SSC/ (1) HSC/ (2) Graduate/ (3) Post Graduate/ (4) other (5)
14	Mothers job and annual income	Service / (1) Business/ (2) Agriculture/ (3) In house/ (4) Other/ (5)
	Income	0-1 lakh (1) , 1.1-2 lakh(2), 2.1-5 lakh(3) , 5lakh - above (4)
15	Family size	
16	Family relationship	Excellent / (1) Good/ (2) Satisfactory/ (3) Bad/ (4) Very Bad (5)
17	Family support to your education	Excellent / (1) Good/ (2) Satisfactory/ (3) Bad/ (4) Very Bad (5)
18	Reason to choose this course	Career in IT/ (1) Near to Home/ (2) Reputation of course / (3) Blind Decision/ (4) Parents wish: (5)
19	Travel mode and time needed	Bus/ (1) Railway/ (2) City Bus/ (3 but taken as 1) Rickshaw/ (4) Self Vehicle / (4) walking (6) (5)

Fig. 1. Sample Questionnaire

We are aware that many variables and their interrelations need to be analyzed for characterization of an object. It is always true for questionnaires as they consist of many questions, such that each question contributes for one variable [12]. All implementations are carried out on SPSS data mining platform [17].

II. EXPERIMENTATIONS AND DISCUSSIONS

The core objective is to find relationship personal details with social ambience; we have carried out some set of experiments. The purpose of the first experiment was to examine the relationship between student's details and his/her family background. We have made region wise groups of students. Then we have selected variables like availability of a personal computer, access to internet, accommodation type, travelling mode and free time to study. Then, Canonical correlation analysis is used to find the significant relationship between student's details and his selected variables of social background to determine the associations among two sets of variables. It is suitable in the same situations where manifold regression would be, but where are there are multiple intercorrelated result variables. For the empirical analysis purpose SPSS 22v software is used. The SPSS analysis is carried out using the manova command. The manova command is not available in the point-and-click analysis menu. The manova command is one of SPSS's hidden gems and used with the discrim option, manova will compute the canonical correlation analysis. Below tables show run time snaps during our experiments. Tables 1 to 10 shows the snapshots from SPSS environment.

Table 1. Cross tabulation of Region Vs. Students having Self PC

Region	Self PC –No	Self PC- Yes	Total
Urban	46	135	181
Rural	84	91	175
Foreign	0	3	3
Total	130	229	359

Table 2. Chi-Square Tests Analysis

Analysis	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	21.366[a]	2	.000
Likelihood Ratio	22.504	2	.000
Linear-by-Linear Association	15.360	1	.000
N of Valid Cases	359	NA	NA

NB: *a. 2 cells (33.3%) have expected count less than 5. The minimum expected count is 1.09.*

Table 3. Cross tabulation of Region Vs. Use of internet

REGION	INTERNET Connectivity?		Total
	No	Yes	
Urban	5	176	181
Rural	25	150	175
Foreign	0	3	3
Total	30	329	359

Table 4. Chi-Square Tests Analysis

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15.703[a]	2	.000
Likelihood Ratio	17.056	2	.000
Linear-by-Linear Association	13.262	1	.000
N of Valid Cases	359		

NB*. a. 2 cells (33.3%) have expected count less than 5. The minimum expected*

 ISBN: 978-81-927230-0-9

Table 5. Cross tabulation of Region Vs. place of living

Region	Place of living				Total
	Own Home	Hostel	Shared Room	Relative Home	
Urban	108	43	17	13	181
Rural	72	66	25	12	175
Foreign	0	1	2	0	3
Total	180	110	44	25	359

Table 6. Chi-Square Tests Analysis

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	22.541[a]	6	.001
Likelihood Ratio	20.630	6	.002
Linear-by-Linear Association	7.313	1	.007
N of Valid Cases	359		

NB: *a. 4 cells (33.3%) have expected count less than 5. The minimum expected count is .21.*

Table 7. Place of Living Vs. Travelling modes

PLACE OF LVING	T-MODE					Total
	Bus	Railway	Rickshaw	Self Vehicle	Walking	
Own Home	85	9	10	66	10	180
Hostel	43	3	3	4	57	110
Shared Room	26	1	4	3	10	44
Relative Home	18	0	0	6	1	25
Total	172	13	17	79	78	359

Table 8. Chi-Square Tests Analysis

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	122.684[a]	12	.000
Likelihood Ratio	130.220	12	.000
Linear-by-Linear Association	.862	1	.353
N of Valid Cases	359		

NB*: a. 5cells (25.0%) have expected count less than 5. The minimum expected count is .91.*

Table 9. Cross tabulation of Region Vs. Free Time to Study

REGION	Free Time available for Study					Total
	Excellent	Good	Satisfactory	Bad	Very Bad	
Urban	19	91	64	6	1	181
Rural	14	120	39	2	0	175
Foreign	0	3	0	0	0	3
Total	33	214	103	8	1	359

Table 10. Chi-Square Tests Analysis

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	15.808[a]	8	.045
Likelihood Ratio	17.387	8	.026
Linear-by-Linear Association	6.080	1	.014
N of Valid Cases	359		

NB: *a. 9 cells (60.0%) have expected count less than 5. The minimum expected count is .01*

The red coloured numbers in Table 7 shows inappropriate data, which we did not considered during experimentations.

Our observations gave us significant outcomes. The number of possible canonical variates also known as canonical dimensions, is equal to the number of variables in the smaller set. In this study, the first set has three variables and the second set has eight. This leads to three possible canonical variates for each set, which corresponds to the three columns for each set and three canonical correlation coefficients in the output. The Canonical dimensions are latent variables that are analogous to factors obtained in factor analysis, except that canonical variates also maximize the correlation between the two sets of variables. In general, not all the canonical dimensions will be statistically significant. A significant dimension corresponds to a significant canonical correlation and vice versa.

Based on these results, the correlations between Gender and Age variable in a group and the group's canonical variates shows 0.86 and 0.75 respectively shows strong positive correlation. The variance in dependent variables explained by canonical variables 45.50%, 28.62 % and 25.86% respectively. The variance in dependent variables is accounted by canonical variates 4.13%, 1.30% and 0.32% respectively. While computing correlations between covariates and canonical variables it is observe that, there is significant positive correlation (r=0.85) between fathers income with gender of the student. According to age of students shows significant negative correlation with mothers education (r=-0.63) and fathers education (r=-0.58). Also, Students UG level percentage is positively correlated with mothers income (r=0.79) and mothers education (r=0.50).

III. CONCLUSION

It was openly understood that many social, habitual and economical aspects are associated with performance of students. However, these were invisible and no attempt was made at our University to scientifically visualize them. The study took it as challenge. In this study, only the social background of students was considered. Academic analytics in terms of computing correlations between covariates and canonical variables is demonstrated. The study has observed that, there is significant positive correlation between student's performance and his/her family social background variables.

REFERENCES

1. Margaret Dunham, Data Mining: Introductory and Advanced Topics , by Margaret H. Dunham , , Pearson publications, 2002.

2. Han,J. and Kamber, M., (2006) "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray.
3. Behrouz.et.al., (2003) Predicting Student Performance: An Application of Data Mining Methods With The Educational Web-Based System Lon-CAPA © 2003 IEEE, Boulder, CO.
4. IBM SPSS Statistics 22 Documentation on internet retrieved at www.ibm.com/support/docview.wss?uid=swg27038407.
5. Pritchard, M. E., and Wilson, G. S. (2003). Using emotional and social factors to predict student success. Journal of College Student Development 44(1): 18–28.
6. Shoukat Ali et al , Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus, American Journal of Educational Research, 2013 1 (8), pp 283-289.
7. Graetz, B. (1995), Socio-economic status in education research and policy in John Ainley et al., Socio-economic Status and School Education DEET/ACER Canberra., J Pediatr Psychol. 1995 Apr;20(2):205-16.
8. Considine, G. & Zappala, G. (2002). Influence of social and economic disadvantage in the academic performance of school students in Australia. Journal of Sociology, 38, 129-148.
9. Bratti, M. and Staffolani, S. 2002, 'Student Time Allocation and Educational Production Functions', University of Ancona Department of Economics Working Paper No. 170.
10. Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M. (2000). Targeting the right students using data mining. Paper presented at the Sixth ACM SIGKDD International Conference, Boston, MA (Conference Proceedings; p. 457-464).
11. B. Minaei-Bidgoli, D. A. Kashy, G. Kortemeyer and, W. F. Punch."Predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA" In Proceedings of ASEE/IEEE Frontiers in Education Conference, Boulder, CO: IEEE, 2003.
12. Kotsiantis S. 2009. Educational Data Mining: A Case Study for Predicting Dropout – Prone Students. Int. J.Knowledge Engineering and Soft Data Paradigms, 1(2), 101–111.
13. Nikhil Rajadhayx et al , Data mining in Educational Domain , retrieved from http://arxiv.org/pdf/1207.1535.pdf.
14. Gordon Linoff, Michael J, et al , Data Mining Techniques , 3e, Wiley Publications.
15. Eko Indrato , edited notes on Data Mining, retrieved from www. Http://recommender-systems.readthedocs.org/en/latest/datamining.html.
16. Paulo Cortez and Alice Silva ,Using Data Mining To Predict Secondary School Student Performance , retrieved from http://www.researchgate.net/ publication/ Using_data_mining_to_ predict_s econdary_ school_ student_ performance.
17. Field, A. (2000). Discovering Statistics using SPSS for Windows. London – Thousand Oaks – New Delhi: Sage publications.