

# Academic Analytics Implemented for Students Performance in Terms of Canonical Correlation Analysis and Chi-Square Analysis

Aniket Muley, Parag Bhalchandra, Mahesh Joshi and Pawan Wasnik

**Abstract** In this research study, we were interested to test the significant association between selected variables which otherwise called as invisible and have indirect impact on the performance of the students. We have devised out our own dataset for the experimental purpose. Our study has made these variables and their relationship visible. The results enable us to determine characteristics of learning environment related to performance.

**Keywords** Data mining · Statistical analysis · Patterns

## 1 Introduction

Academic analytics is one branch of modern day's data analysis which uses statistical analysis and data mining methods to reveal and recognize hidden patterns in vast educational databases [1–6]. Such patterns enable us to throw better light on educational aspects related to student behavior, prognostication, student-centric learning, remedial aspects, and learning outcome with high accuracy. This will

---

A. Muley (✉)

School of Mathematical Sciences, S.R.T.M. University, Nanded 431606,  
Maharashtra, India

e-mail: aniket.muley@gmail.com

P. Bhalchandra · P. Wasnik

School of Computational Sciences, S.R.T.M. University, Nanded 431606,  
Maharashtra, India

e-mail: srtmun.parag@gmail.com

P. Wasnik

e-mail: pawan\_wasnik@yahoo.com

M. Joshi

School of Educational Sciences, S.R.T.M. University, Nanded 431606,  
Maharashtra, India

e-mail: maheshmj25@gmail.com

© Springer Nature Singapore Pte Ltd. 2018

D.K. Mishra et al. (eds.), *Information and Communication Technology*,  
Advances in Intelligent Systems and Computing 625,  
[https://doi.org/10.1007/978-981-10-5508-9\\_26](https://doi.org/10.1007/978-981-10-5508-9_26)

definitely increase standards of Indian higher educational system [6]. Due to digitization and effective use of computers, IT and ICT technologies, all educational organizations, institutions, and universities have generated and stored large data in their databases [7–13]. This data can be a key source for futuristic decision making processes if it is being processed through academic analytics. We took it as a challenge to see all the business intelligence, patterns, correlations, and rules embedded in this data. Our work is an interdisciplinary work undertaken by three schools of our university as performance analysis shares sphere with educational pedagogies, statistics, and computer-enabled technologies. The academic analytics was implemented using SPSS software [14, 15].

A closed questionnaire with predefined answers was used for data gathering [16] on A4 size single-sided paper sheet. Performance-related economical, social, and emotional attributes of this questionnaire were selected with the help of School of Educational Sciences and as per theory of Pritchard and Wilson [16, 17]. The questionnaire was modified number of times to reduce the complexity of understanding as well as to increase simplicity of answering. It was tested on subset of students after every revision. An Excel sheet was prepared for the answers using code such as 0, 1, 2. The confidential issues in datasets were properly addressed as dataset carried personal information of students. The error rate during preprocessing was 38% which finally reduced to 5% after proper convincing to students. The questionnaire looks like Figs. 1 and 2.

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)
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 / walking (6) (5)

Fig. 1 Sample questionnaire

FORM_NO	CORSNAME	YRNAME	GENDER	MARRIED	AGE	REGION	MOBNO	UG
1	6	DUGANE SEDHARTH NAGORAO	1	3	24	2	8446461553	2
2	6	SHASHANK H. BALASKAR	1	3	21	2	7709512133	3
3	6	MOTE PRADEEP KASINATH	1	3	22	2	2462229251	3
4	6	PANDIT SWAPNIL RAGHUNATH	1	3	24	2	9970403753	2
5	6	GOTRE NILESHKUMAR NAMDEV	1	3	22	1	9890402292	2
6	6	GADEWAR LEENA PRAMOD	0	3	23	2	9665360530	2
7	6	KOLHE VANITA PANDURANG	0	3	22	1	9579836606	2
8	6	BARDE NISHA P	0	3	21	1	2462229251	2
9	6	GAWANDE SANTOSH PRABHAKAR	1	3	24	2	9028064993	2
10	6	NITIN NARESH DEKATE	1	3	21	2	9011952075	2
11	6	ASHUTOSH V. DONGRE	1	3	23	1	9730834463	2
12	6	SWAMI SANTOSH SHIVLING	1	3	22	2	8793336938	3
13	6	WATHORE ANKUSH NAVNATH	1	3	21	1	2462229251	3
14	6	BAHIWAL AKSHAY BALAPRASAD	1	3	21	1	7709689015	3
15	6	EMEKAR SANDEEP SUBHASH	1	3	23	1	9881177713	2
16	6	GALIPELLI SANDHYA SHANKABABU	0	3	21	1	7709977551	2

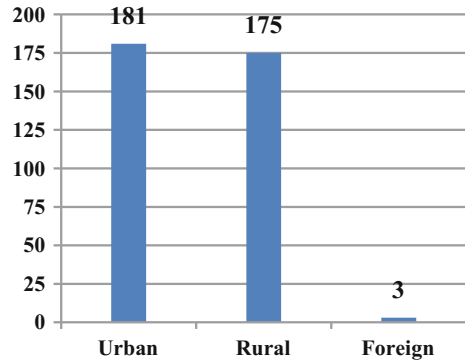
Fig. 2 Data set in MS-Excel

## 2 Experimentations and Discussion

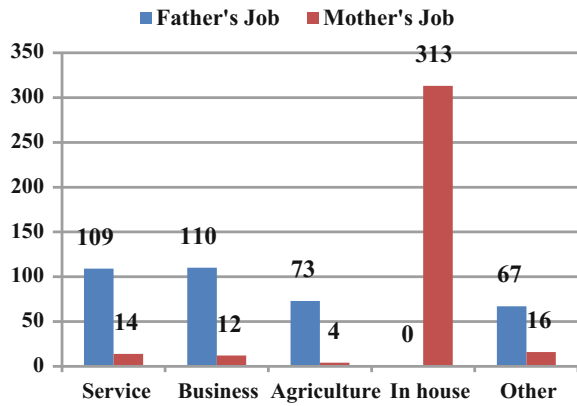
Our aim was to discover invisible attributes related to performance of students. So we had discussions with educationalist and then finally understood that the semester end marks alone cannot be taken as main indicator of student’s performance. The performance is indistinct term. For proper knowledge, surveyed literatures such as Shoukat Ali et al. [4], Graetz [18], Considine and Zappala [19], and Bratti and Staffolani [20]. This analysis is helpful for identifying the personal, social and economic kinds attributes in our study.

With these preliminary investigations and understanding, we decided to identify key variables that accelerates or downgrades educational performance at large. We had thought that economical and social conditions of students can be important variables from our dataset/questionnaire as far as performance was concerned. To do so, many variables and their interrelations needed to be analyzed for proper analysis. It is always true for questionnaires as they consist of many questions, such that each question contributes for one variable [7, 21–23]. Studying all variables and their interrelation may be complicated as they may divert us from the original research focus. For such exploratory analysis, factor analysis has been invented [22]. We have used SPSS22.0v to analyze the data set. The snapshots given below show the evidence of empirical analysis. The descriptive statistics are used through MS-Excel to represent our data in the diagrammatic form. Some of the interesting facts are shown in Figs. 3, 4, 5 and 6. Further, canonical correlation analysis and chi-square testing have been done on the experimental data set.

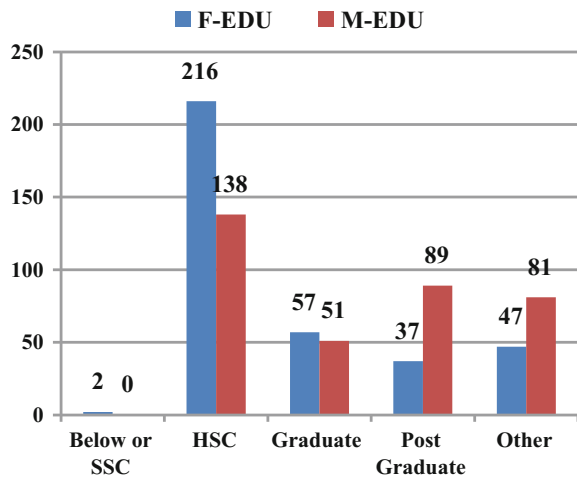
**Fig. 3** Region-wise distribution



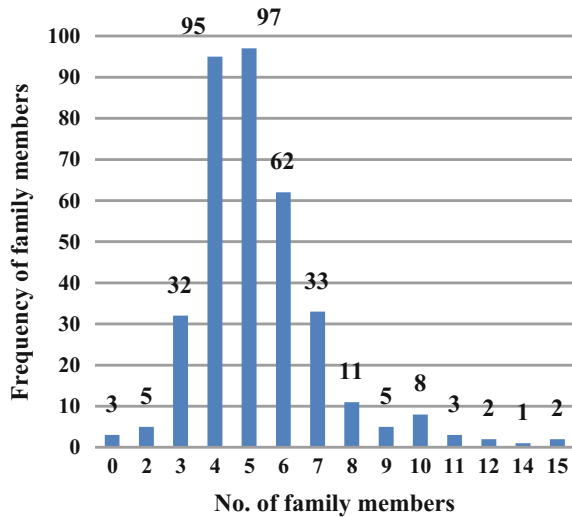
**Fig. 4** Diversity in jobs among parents



**Fig. 5** Parents versus their education level



**Fig. 6** Students versus their family size



### 2.1 Program Code

The SPSS22.0v is used to analyze the data set [16].

```

FREQUENCIES VARIABLES=GENDER MARRIED AGE REGION UG FEDU FJOB
FINCOM MEDU MJOB MINCOM FSIZE FRELATIONS FSUPPORT REASON
TMODE
TTIME STIME FAILURES TUTORIAL SCHOLERSHIP PJOB MM HARDSUB_UG
STUDY_HOME SELFLIB SELFPC PLACELVING INTERNET F_T_STUDY
F_T_FRIEND MOVIEPWEEK CAREERDREM PARALLELOURSE OWN_NOTES
FREE_T_ACC PER_SATISF MATERIAL HLT_STATUS
/ORDER=ANALYSIS.
CROSSTABS
/TABLES=REGION GENDER BY FAILURES STIME SCHOLERSHIP PJOB
SELFLIB SELFPC PLACELVING INTERNET F_T_STUDY F_T_FRIEND
MOVIEPWEEK OWN_NOTES PER_SATISF MATERIAL
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ
/CELLS=COUNT
/COUNT ROUND CELL.
    
```

The use of descriptive statistics has been made using MS-Excel to represent our data in the diagrammatic form. Figures 3, 4, 5 and 6 show the distribution of the data according to region-wise classification, diversity of parents jobs, education-wise, and their family size-wise, respectively. The students came from urban and rural backgrounds are found to be approximately same of Indian students as compared to foreign students. The discrimination in the student’s performance is observed

according to their parent's job and educational background. Also, numbers of family members in student's family were represented through the bar plot. The interesting facts are shown in Figs. 3, 4, 5 and 6.

## **2.2 Canonical Correlation Analysis**

The core objective is to find relationship between personal details with family background. We made two groups for proper analysis. The first group is student's details containing three parameters, viz. gender, age, and UG percentage. The second group is his/her family background and the parameters chosen are: father's education, father's job, father's income, mother's education, mother's job, mother's income, family size, and whether student does any part-time job? Here, Canonical correlation analysis is used to find the significant relationship between student's details and his family background to determine the associations among two sets of variables. Our observations gave us significant outcomes.

## **2.3 Chi-Square Analysis**

Sample analysis using chi-square tests is mentioned here. Similar way, the results were computed and it has been represented in the form of conclusion. Below figures and tables show the use of descriptive statistics. These together show some data regarding diversity of the students according to course-wise, gender-wise, undergraduate background, father's occupation, and their family size. We have applied chi-square test to test the significance among the above objectives and assumptions that there will be significant difference among the variables under study.

Some of the parameters which show significant differences in our study are as scholarship holder students with gender-wise; difference gender-wise about their career dreams; between gender-wise percentages obtained at UG level, between region-wise percentages obtained at UG level by the students; between age group-wise obtained scholarships; between age group-wise obtained UG Percentage; students and their father's education; students and their father's job; between gender-wise and their mother's education; age-wise and their family size; age-wise and part-time job; region-wise and father's education; region-wise and family size; students place of living and self library. Further, we have made analysis using chi-square Tests with the help of SPSS 22.0 software [15] and found some significant results. These are represented in the form of tables. According to region-wise study with respect to variables like place of living, do they have their own PC? Do they use internet? How much free time they have for study? It was surprising to note that there are significant differences with respect to student's living places. These differences came because of student's awareness to use internet. Our students are from computer science field, and hence, it is expected that

they must frequently use internet. From our experimental analysis, it is found to be true. While dealing with students free time for study perspective, it has been observed that there is significant enough good time is available for study. It was assumed that in due course of studentship, he/she may get sufficient time for study rather than doing any other work. This particularly holds true as the Nanded region is not a metro city or an industrial hub. When we did gender-wise study with a variable, how much scholarship they get? It is observed that there is significance difference. Male students get more scholarship than female students. Also, we found significance among gender-wise difference in their place of living. Most of the female students preferred to live at own home or in hostels due to safety issues. Tables 1, 2, 3, 4 and 5 show these results.

**Table 1** Chi-square tests analysis for region versus students having self PC

	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	21.366 <sup>a</sup>	3	0.000
Likelihood ratio	22.504	3	0.000
Linear-by-linear association	15.360	1	0.000
No. of valid cases	359		

<sup>a</sup>2 cells (33.3%) have expected count less than 5 and the minimum expected count is 1.09

**Table 2** Chi-square tests analysis for region versus place of living

	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	15.703 <sup>a</sup>	3	0.000
Likelihood ratio	17.056	3	0.000
Linear-by-linear association	13.262	1	0.000
No. of valid cases	359		

<sup>a</sup>2 cells (33.3%) have expected count less than 5 and the minimum expected count is 0.25

**Table 3** Chi-square tests analysis for region versus free time to study

	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	15.808 <sup>a</sup>	3	0.000
Likelihood ratio	17.387	3	0.000
Linear-by-linear association	6.080	1	0.000
No. of valid cases	359		

<sup>a</sup>9 cells (60.0%) have expected count less than 5 and the minimum expected count is 0.01

**Table 4** Chi-square tests analysis for region versus students having self PC

	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	21.366 <sup>a</sup>	3	0.000
Likelihood ratio	22.504	3	0.000
Linear-by-linear association	15.360	1	0.000
No. of valid cases	359		

<sup>a</sup>0 cells (0.0%) have expected count less than 5 and the minimum expected count is 78.28

**Table 5** Chi-square tests analysis for gender versus place of living

	Value	df	Asymp. sig. (2-sided)
Pearson chi-square	12.996 <sup>a</sup>	3	0.000
Likelihood ratio	13.174	3	0.000
Linear-by-linear association	0.002	1	0.000
No. of valid cases	359		

<sup>a</sup>0 cells (0.0%) have expected count less than 5 and the minimum expected count is 10.93

### 3 Conclusion

The performance of the student is fuzzy terms and it is affected by many parameters. In this study, our data reveal that it is due to the social and economical condition of students. However, no scientific evidences were there for such outcome. The study took it as challenge and it has been discovered that the student's performance mere did not depend on his/her studious nature. This paper shows effective use of academic analytics in terms of descriptive statistics. Here, we have applied canonical correlation analysis and chi-square test to test the significance among the stated objectives and assumptions. We have finally discovered new variables, which otherwise were invisible that hampers performance of students.

### References

1. Dunham Margaret H.: Data Mining: Introductory and Advanced Topics, Pearson publications (2002)
2. Han, J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd edition. The Morgan Kaufmann Series in Data Management Systems, Jim Gray. (2006)
3. Behrouz et.al.: Predicting Student Performance: An Application of Data Mining Methods With The Educational Web-Based System Lon-CAPA IEEE, Boulder, CO. (2003)
4. Shoukat Ali et al.: Factors Contributing to the Students Academic Performance: A Case Study of Islamia University Sub-Campus, American Journal of Educational Research, 1 (8), pp. 283–289 (2013)
5. Gordon Linoff, Michael J, et al.: Data Mining Techniques, 3Ed, Wiley Publications.
6. Eko Indrato: edited notes on Data Mining, retrieved from <http://recommender-systems.readthedocs.org/en/latest/datamining.html>



7. Ma, Y., Liu, B., Wong, C. K., Yu, P. S., & Lee, S. M.: Targeting the right students using data mining. Paper presented at the Sixth ACM SIGKDD International Conference Proceedings, Boston, MA, pp. 457–464 (2000)
8. Minaei-Bidgoli, B., Kashy, D. A., Kortemeyer G. and, Punch, W. F.: 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)
9. Kotsiantis S.: Educational Data Mining: A Case Study for Predicting Dropout – Prone Students. *Int. J. Knowledge Engineering and Soft Data Paradigms*, 1(2), pp. 101–111 (2009)
10. Berkhin Pavel: Survey of Clustering Data Mining Techniques, Accrue Software, available at [www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf](http://www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf)
11. Sasirekha K., Baby, P.: Agglomerative Hierarchical Clustering Algorithm- A Review, *International Journal of Scientific and Research Publications*, 3(3), pp. 83 (2013)
12. Nikhil Rajadhayx et al.: Data mining in Educational Domain, retrieved from <http://arxiv.org/pdf/1207.1535.pdf>
13. Murugesan Keerthiram, Zhang Jun: Hybrid Hierarchical Clustering: An Experimental Analysis, Technical Report: CMIDA-hipsccs#001–11, retrieved from [www.cs.uky.edu/~jzhang/pub/techrep.html](http://www.cs.uky.edu/~jzhang/pub/techrep.html)
14. Field, A.: *Discovering Statistics using SPSS for Windows*. London–Thousand Oaks – New Delhi: Sage publications. (2000)
15. IBM SPSS Statistics 22 Documentation on internet retrieved at [www.ibm.com/support/docview.wss?uid=swg27038407](http://www.ibm.com/support/docview.wss?uid=swg27038407)
16. Cortez Paulo and Silva Alice: Using Data Mining to Predict Secondary School Student Performance, retrieved from [http://www.researchgate.net/publication/Using\\_data\\_mining\\_to\\_predict\\_secondary\\_school\\_student\\_performance](http://www.researchgate.net/publication/Using_data_mining_to_predict_secondary_school_student_performance)
17. Pritchard, M. E., and Wilson, G. S.: Using emotional and social factors to predict student success. *Journal of College Student Development* 44(1): pp. 18–28. (2003)
18. Graetz, B.: 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.* 20 (2):205–216 (1995)
19. Considine, G. & Zappala, G.: Influence of social and economic disadvantage in the academic performance of school students in Australia. *Journal of Sociology*, 38, 129–148 (2002)
20. Bratti, M. and Staffolani, S.: Student Time Allocation and Educational Production Functions, University of Ancona Department of Economics Working Paper No. 170 (2002)
21. Introduction to factor analysis, web resource [www.yorku.ca/ptryfos/f1400.pdf](http://www.yorku.ca/ptryfos/f1400.pdf)
22. Rietveld, T. & Van Hout R.: *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin – New York: Mouton de Gruyter (1993)
23. Habing, B.: *Exploratory Factor Analysis*. Website: <http://www.stat.sc.edu/~habing/courses/530EFA.pdf> (accessed 10 May 2004) (2003)