International Journal of Emerging Engineering Research and Technology
Volume 4, Issue 1, January 2016, PP 12 -21
ISSN 2349-4395 (Print) & ISSN 2349-4409 (Online)

Prognostication of Student's Performance: Factor Analysis Strategy for Educational Dataset

Aniket Muley[2], Parag Bhalchandra[1], Mahesh Joshi[3], Pawan Wasnik[1]

[1]*School of Computational Sciences, S.R.T.M. University, Nanded, MS, 431606, India*
[2]*School of Mathematical Sciences, S.R.T.M. University, Nanded, MS, 431606, India*
[3]*School of Educational Sciences, S.R.T.M. University, Nanded, MS, 431606, India*

ABSTRACT

The emerging field of academic analytics is educational data mining. It is rapidly producing new possibilities in vast databases of educational data by extending their current reporting capabilities of uncover and hidden patterns. The objective of this research work is to analyse the factors affecting on the performance of collected student's educational dataset. Also, to discover academic and social variables that have interrelation together which rigorously affect the student's performance? The result enables us to determine characteristic patterns in palpating student's exploratory learning environment. It can be subsequently used to identify more and less effective prognostication strategies for students. All implementations are carried out on R Miner platform.

Keywords: Educational Data Mining, Factor Analysis, Clustering, Statistical Analysis

INTRODUCTION

Data is a key source of intelligence and has competitive advantage for every higher educational organization. With the explosion of electronic data available to educational organizations and the demand for better and faster decisions [16], the role of data driven intelligence is becoming central in educational organizations. Contrary Data mining or Knowledge Discovery in Database (KDD) is the process of converting the raw data into useful knowledge required supporting decision making [1, 16]. It automates the process of knowledge discovery, making us more productive in our search for useful information than we would be otherwise. It also increases the confidence with which we can make business decisions.

Now days, all educational organizations, institutions or Universities have been computerized and they have database systems capturing all essential data from all vital parts. Even though, one of the biggest challenges that higher educational institutes face is predicting knowledge from databases. Data mining finds applications in educational industry as virtually every educational organization is in the process of exploring and implementing data mining solutions to core problems [16] including student support, course registration processes, alumina associations, designing new courses, etc. In addition to these challenges, traditional issues such as enrollment management and time-to-degree continue to motivate higher education institutions to search for better solutions [17]. These challenges can be addressed by use of Educational data mining algorithms. In Educational data mining, the data mining capabilities to educational databases enable organizations to use their current reporting capabilities to uncover and understand hidden patterns in vast databases [2,3,16]. These patterns are then built into data mining models and used to predict individual behavior with high accuracy. As a result of this insight, institutions are able to allocate resources and staff more effectively [17]. Despite the Government, UGC, etc are pouring lot of funds for improving education standards, academic performance of students from rural Universities of India is not improving. The Academic analytics will help in extracting high level knowledge from raw data and hence offer an interesting automated tool that can aid the educational domain [1, 3].

Address for correspondence:

srtmun.parag@gmail.com

Over the years, University like Swami Ramanand Teerth Marathwada University, Nanded has accumulated a vast amount of data in their databases- information systems. These data typically represent daily operations and transactions within education, administration contexts. It is easy to see that all the business intelligence and rules are, in some way, can be embedded in these data. This is the first most attempts. 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 purpose of this study is to introduce data mining processes on collected datasets. Using these results, one can realize what data has to be captured, what surrounding information has also to be captured, how mining can be carried out, etc for accurate insight. The research question then becomes precise like: How can we mine this vast amount of data in order to learn the embedded business intelligence? How to apply that intelligence to gain intelligence in a more efficient and effective manner? The present work intends to approach student achievements, performance in University level education using above fields. Some real world data set of students has been composed using questionnaire and progress reports. This includes student's personal details like social, intellectual, demographic, habitual, health and economical data. The core attributes where we are finding downfall in performance of students were modeled using factor analysis models.

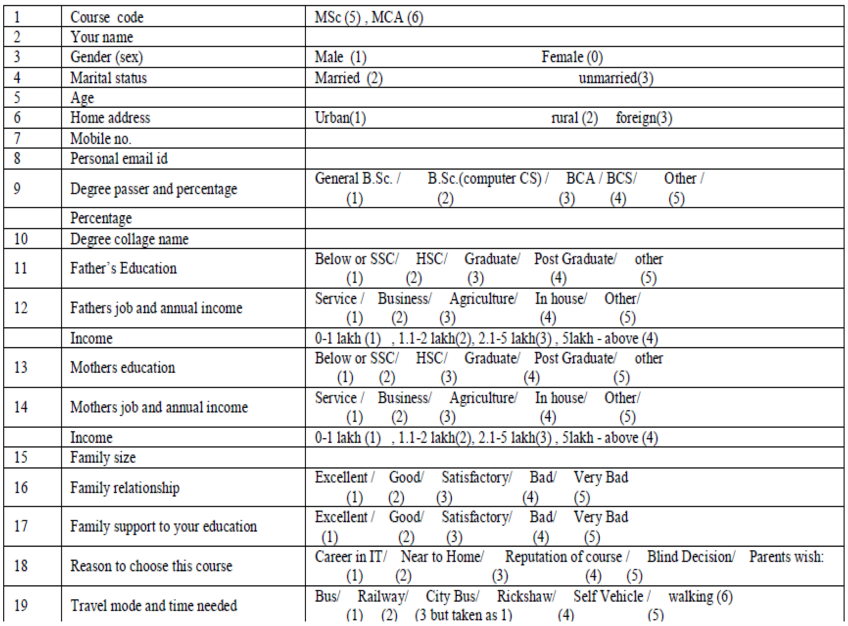
The results show that a good predictive accuracy can be achieved if student's historical information is available. A student's dataset was created with 360 records and 46 fields by closed questionnaire method. Hierarchical clustering algorithms were implemented using R Miner software [26].

RESEARCH METHODOLOGY

In India, the educational system consists of 10+2+ 3 pattern of undergraduate education for no-professional streams and 10+2+4 pattern for undergraduate professional streams. The students admitted to School of Computational Sciences are mainly from the first pattern. Most of the students are under graduate from the public, grant-in-aids sanctioned and nominal expensive education system. There are several courses like BCS, BCA, B.Sc. (CS or IT or CM) under science and technology faculty from where the students mainly come. This study will consider data collected during the 2009-2012. The database was built from two sources: previous examination's progress reports and questionnaires used to complement the previous information. We have also collected information of the fresher.

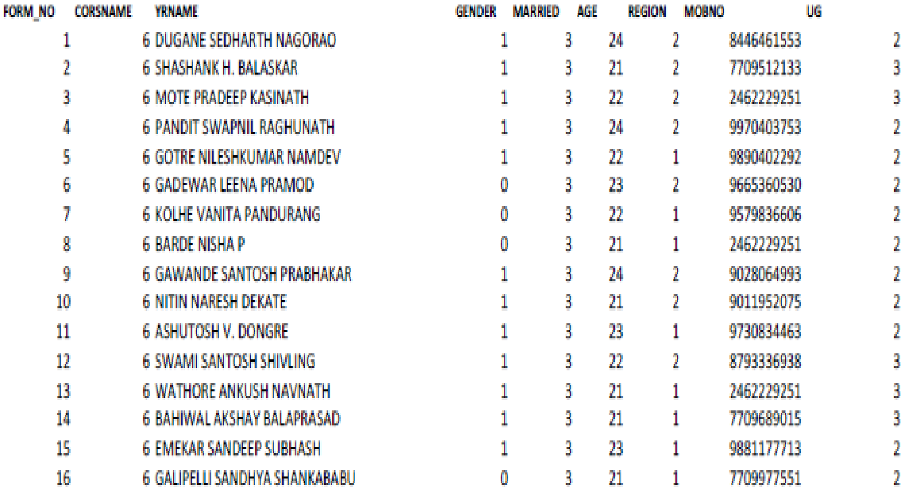
The questionnaire is designed with closed questions [18] (i.e. with predefined options) related to several demographic (e.g. mother's education, family income), social/emotional attributes as defined in Pritchard and Wilson [5, 18] and performance related (e.g. number of past class failures) variables that were expected to affect student performance. The questionnaire was reviewed and tested on a small set of students in order to get a feedback. The final version contained 43 questions in a single one sided paper sheet and it was answered by all students in the School. Finally, the data was integrated into a dataset with 360 student records and each record consists of 46 fields (three additional fields than the total number of questions 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 0r 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. Likewise other answers are also converted into numeric values.

During the pre-processing stage, some features were discarded due to the lack of confusing values like few respondents have not answered about their family income (probably due to privacy issues. There was some false information like all students live at their homes with their parents and have a personal computer at home. This cannot be true as students live in hostels, relative's home or even they share rooms. We got around 18% confusing data or falsely filled data or partly filled questionnaire in first step. In order to correct them, the students were called on one to one basis and convinced to fill the missing data. There were some confidential issues discovered after going through the filled questionnaire. Appropriate actions were taken time to time for example, students shied to disclose their personal mobile numbers. This was very redundant with girl students. Significant number of students got convinced and for remaining students, we put our departmental number as their contact number. A snapshot of questionnaire and the dataset is as given in below figures.



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. / B.Sc.(computer CS) / BCA / BCS/ Other / (1) (2) (3) (4) (5)
	Percentage	
10	Degree collage name	
11	Father's Education	Below or SSC/ HSC/ Graduate/ Post Graduate/ other (1) (2) (3) (4) (5)
12	Fathers job and annual income	Service / Business/ Agriculture/ In house/ Other/ (1) (2) (3) (4) (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/ HSC/ Graduate/ Post Graduate/ other (1) (2) (3) (4) (5)
14	Mothers job and annual income	Service / Business/ Agriculture/ In house/ Other/ (1) (2) (3) (4) (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 / Good/ Satisfactory/ Bad/ Very Bad (1) (2) (3) (4) (5)
17	Family support to your education	Excellent / Good/ Satisfactory/ Bad/ Very Bad (1) (2) (3) (4) (5)
18	Reason to choose this course	Career in IT/ Near to Home/ Reputation of course / Blind Decision/ Parents wish: (1) (2) (3) (4) (5)
19	Travel mode and time needed	Bus/ Railway/ City Bus/ Rickshaw/ Self Vehicle / walking (6) (1) (2) (3 but taken as 1) (4) (5)

Fig1. *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

Fig2. *Data Set in MS-Excel*

Since we aim for discovery of 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 mare 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.[6]. This work is a lucid discussion on performance analysis of students. This work is then taken as main base for our analysis and some terminology is also borrowed from the same. As per this study, the student's academic gain and learning performance are affected by numerous factor including gender, age, father/guardian social economic status, residential area of students, medium of instructions in schools, tuition trend, daily study hour and accommodation as hostelries [6] or shearing room with other friends. This consideration matches with few attributes of our dataset. In the literature review Shoukat Ali at al.[6] cites other references for proper understanding of factors important for performance analysis. A similar works of Graetz et al. [7] suggested that the student educational success contingent heavily on social status of student's parents/ guardians in the society. This also matches with our understanding of proportionate relationship of performance with social and economical conditions of students. The Shoukat Ali et al.[6] also cites that Considine and Zappala [8] have noticed that parent's income or social status positively affects the student test score in examination. They [6] further reviews that the work of Staffolani and Bratti[9] observed that the measurement of students previous educational outcomes are the most important indicators of students future achievements.

It is generally assumed that the students who showed better or higher performance in the starting classes of their studies also performed better in future academic years at degree level. The combined findings that affects performance as per [6] includes identify students' effort, previous schooling, parent's educational background, family income, self motivation of students, age of student, learning preferences and entry qualification of students as important factors that have effect on student's academic performance in different setting. We felt that once we discover these hidden truths, we can undertake corrective measures that improve the academic performance of graduate students [6]. This is the prognostication for improving student's performance.

Further, it was matter of curiosity that whether such investigations were made in past or not. So we started investigating profiling studies for student's performances. In effect, several studies have been found which have addressed related issues. The Ma et al. [9] applied a DM approach based in Association Rules in order to select weak students [18]. The input variables here included demographic attributes e.g. sex, region and performance over the past years and the proposed solution outperformed the traditional allocation procedure. In 2003, Minaei-Bidgoli et al. [10] modeled online student grades from the Michigan State University using three classification approaches (i.e. binary: pass/fail; 3-level: low, middle, high; and 9-level: from 1 - lowest grade to 9 - highest score). Their database included 227 samples with online features (e.g. number of corrected answers or tries for homework) and the best results were obtained by a classifier ensemble (e.g. Decision Tree and Neural Network) with accuracy rates of 94% (binary), 72% (3-classes) and 62% (9- classes)[18]. The Kotsiantis et al. [11,18] applied several DM algorithms to predict the performance of computer science students from an university distance learning program. For each student, several demographic (e.g. sex, age, marital status) and performance attributes (e.g. mark in a given assignment) were used as inputs of a binary pass/fail classifier. The best solution was obtained by a Naive Bayes method with an accuracy of 74%. It was also observed that the past school grades has a much higher impact than demographic variables. More recently, Pardos et al. [12,18] collected data from an online tutoring system regarding USA 8th grade Math tests. The authors adopted a regression approach, where the aim was to predict the math test score based on individual skills. The authors used Bayesian Networks [18] and the best result was a predictive error of 15%.

Keeping the knowledge earned after careful review of above primary sources, in this work, we aim to predict attributes which are directly related to student's performance and if possible to identify the key variables that accelerates or downgrades educational performance at large. The analysis was carried out with some thoughts in mind including social aspects of students, economical favorableness of students to peruse education, personal desires & ambitions and finally resources available with students which can help them to get success. After analysing all above contemporary results, we understood that there is need to discover important variables from our dataset / questionnaire. In routine sense, their interrelationship can give us an idea for prognostication interventions. Many variables and their interrelations needed to be analysed 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 [20]. Studying all variables and their interrelation may be complicated as they may divert us from the original research focus. For such expletory analysis, factor analysis has been invented [21]. Factor analysis attempts to bring inter-correlated variables together under more general, underlying variables. More specifically, the goal of factor analysis is to reduce the dimensionality of the original space and to give an interpretation to the new space, spanned by a reduced number of new dimensions which are supposed to underlie the old ones [21], or to explain the variance in the observed variables in terms of underlying latent factors [22] Thus, factor analysis offers not only the possibility of gaining a clear view of the data, but also the possibility of using the output in subsequent analyses [20,21]. Although there exists many studies on the theories and practical applicability of factor analysis, we rely on the terminology adopted in a lucid note "Introduction to factor analysis" [20], available across the web. As the goal of this paper is to show and explain the use of factor analysis in R Miner, the theoretical aspects of factor analysis will here be discussed from a practical, applied perspective. The starting point of factor analysis is a correlation matrix, in which the inter-correlations between the studied variables are presented [20, 21]. The dimensionality of this matrix can be reduced by looking for variables that correlate highly with a group of other variables, but correlate very badly with variables outside of that group [23]. These variables with high inter-correlations could well measure one underlying variable, which is called a factor. The obtained factor creates a new dimension that can be visualized as classification axes along which measurement variables can be plotted [20, 23]. This projection of the scores of the original variables on the factor leads to two results: factor scores and factor loadings. Factor scores are the scores of a subject on factor [21], while factor loadings are the correlation of the original variable with a factor. The factor scores can then for example be used as new scores in multiple regression analysis, while the factor loadings are especially useful in determining the substantive importance of a particular variable to a factor [20, 21], by squaring this factor loading as it is a correlation basically, and the squared correlation of a variable determines the amount of variance accounted for by that particular variable. This is important information in interpreting and naming the factors. When the data are appropriate, it is possible to create a correlation matrix by calculating the correlations between each pair of variables. In this matrix clusters of variables with high inter-correlations are found. These clusters of variables could well be manifestations of the same underlying variable [20,21]. The data of this matrix could then be reduced down into these two underlying variables or factors. With respect to the correlation matrix, two things are important: the variables have to be inter-correlated, but they should not correlate too highly as this would cause difficulties in determining the unique contribution of the variables to a factor [23]. We have carried out factor analysis in R, a language and environment for statistical computing and graphics. The R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS [25]. It is a GNU project which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, etc) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

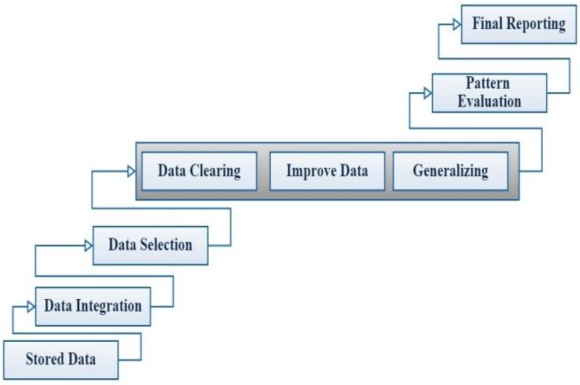


Fig3. *Research Methodology*

Factor or component analysis for array data can be quite straightforward to implement in R [24] by using by Scree plot, Proportion of total variance, Average Eigen value rule, Log-Eigen value diagram, etc [24,25]. In general if your Scree plot is very inconclusive then you just need to pick your poison. There is no absolute right or wrong for any data as in reality the number of PCs to use actually depends on your understanding of the problem. The only dataset you can really know the dimensionality of is the one you constructed yourself. We choose Kaiser's criterion for our analysis [25,26]. It is one of the most commonly used criteria for determining the number of factors or components to include is the latent root criterion, also known as the Eigen value-one criterion or the Kaiser criterion[26,27]. With this approach, one can retain and interpret any component that has an Eigen value greater than 1.0. The rationale for this criterion is straightforward. Each observed variable contributes one unit of variance to the total variance in the data set (the 1.0 on the diagonal of the correlation matrix). Any component that displays an Eigen value greater than 1.0 is accounting for a greater amount of variance than was contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance and is worthy of being retained [26,27]. On the other hand, a component with an Eigen value less than 1.0 is accounting for less variance than had been contributed by one variable. The problem with this criterion is that the number of factors extracted is usually about one third the numbers of items, regardless of whether many of the additional factors are noise. Parallel analysis and the Scree criterion are generally more accurate procedures for determining the number of factors to extract. The overall research methodology is shown in Figure 3.

EXPERIMENTATIONS AND DISCUSSIONS

R Program for Principle Component Analysis

```
A1<-read.table("D:/a/CN.csv", header=T,sep=",");
 A2<-read.table("D:/a/Gen.csv", header=T,sep=",");
A3<-read.table("D:/a/MS.csv", header=T,sep=",");
A4<-read.table("D:/a/Age.csv", header=T,sep=",");
….. so on up to A44
```

Determine Number of Factors to Extract Scree Plot

We use Scree plot to determine the number of factors in the dataset. It plots components on the X axis and the Eigen values on the Y axis. Later connect them with lines. The following procedure to make a Scree plot is needs nFactors package:

```
library(FactoMineR)
   result <- PCA(d1)
ev <- eigen(cor(d1))
library(nFactors)
ap <- parallel(subject=nrow(d1),var=ncol(d1),rep=100,cent=0.05)
nS <- nScree(ev$values, ap$eigen$qevpea)
 plotnScree(nS)
```

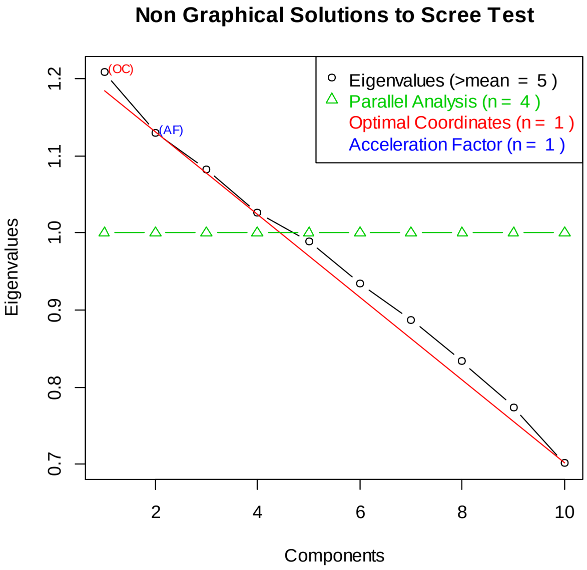


Fig4. *Non Graphical Solution to Screen Test*

The components on the X axis, and the Eigen values on the Y axis and connect them with lines. The Kaiser rule is to discard components whose Eigen values are below 1.0 and with that we can say that four factors playing an important role to check the performance of the student. The Eigen values from the correlation matrix.

```
ev <- eigen(cor(d1))
ev$values
[1] 1.7607207 1.3070246 1.1549692 1.0614750 0.9822735 0.9533021 0.8276333
    0.7148341 0.6893489 0.5484187
```

Here, the significant number of factors should be 4. The intuition of Factor Analysis is to find hidden variables which affect our observed variables by looking at the correlation:

cor(d1)

Factor Analysis is easy to do in R. Let's do Factor Analysis assuming that the number of the

hidden variables are 4.

fa <- factanal(d1, factor=4)

fa

Call:

factanal(x = d1, factors = 4)

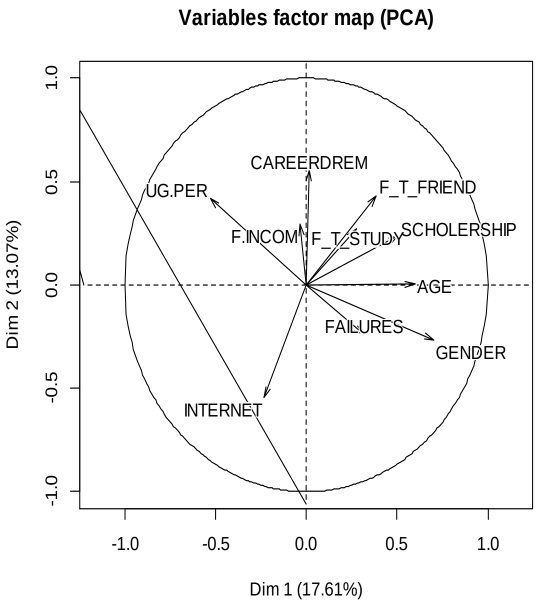


Fig5. *Variable Factor Map*

Thus, all the four factors seem to have some meanings, and that's why we should keep them. The intuition of Factor Analysis is to find hidden variables which affect your observed variables by looking at the correlation.

cor(d1)

Factor Analysis is easy to do in R. Let's do Factor Analysis assuming that the number of the

hidden variables are 4.

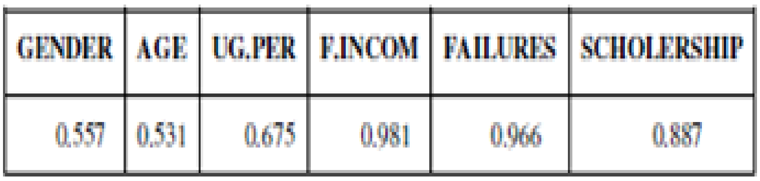
fa <- factanal(d1, factor=4)

fa

Call:

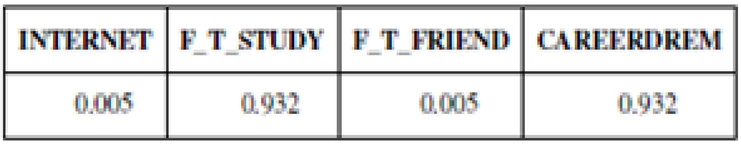
factanal(x = d1, factors = 4)

Table1. *Uniqueness of Six Variables*



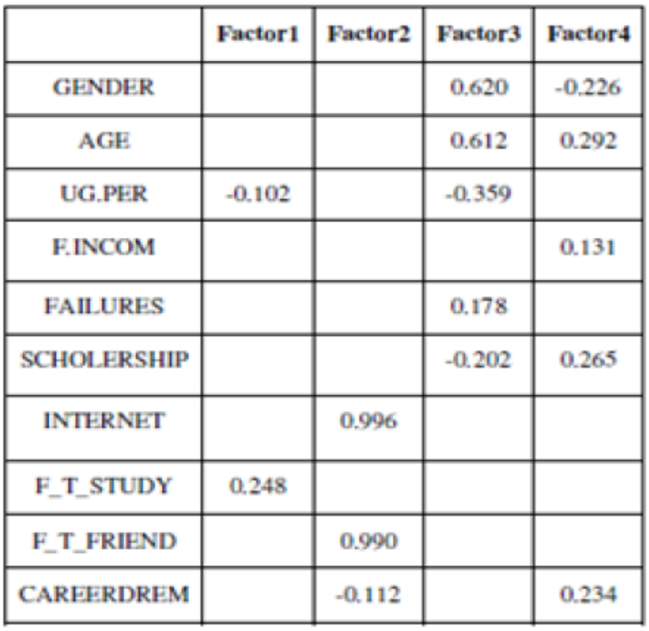
GENDER	AGE	UG.PER	F.INCOM	FAILURES	SCHOLERSHIP
0.557	0.531	0.675	0.981	0.966	0.887

Table2. *Uniqueness of Four Variables*



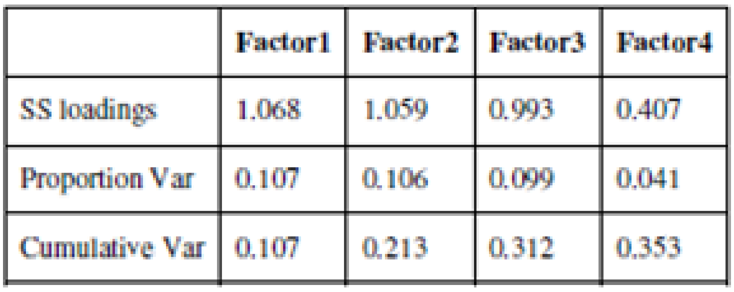
INTERNET	F_T_STUDY	F_T_FRIEND	CAREERDREM
0.005	0.932	0.005	0.932

Table3. *Factor Analysis*



	Factor1	Factor2	Factor3	Factor4
GENDER			0.620	-0.226
AGE			0.612	0.292
UG.PER	-0.102		-0.359	
F.INCOM				0.131
FAILURES			0.178	
SCHOLERSHIP			-0.202	0.265
INTERNET		0.996		
F_T_STUDY	0.248			
F_T_FRIEND		0.990		
CAREERDREM		-0.112		0.234

Table4. *Loading Analysis*



	Factor1	Factor2	Factor3	Factor4
SS loadings	1.068	1.059	0.993	0.407
Proportion Var	0.107	0.106	0.099	0.041
Cumulative Var	0.107	0.213	0.312	0.353

Also, by using the Chi-square statistic with the assumption those 4 factors are sufficient. The chi square statistic is 16.65 on 11 degrees of freedom. The p-value is 0.119. Here, the factor analysis is doing a null hypothesis test in which the null hypothesis is that the model described by the factor we have found predicts the data well. This means, we cannot reject the null hypothesis, so the factor predicts the data well from the statistics perspective. This is why the result says Test of the hypothesis that 4 factors are sufficient.

Kaiser Criterion

The Kaiser rule is to discard components whose Eigen values are below 1.0. The Eigen values from the correlation matrix. Our results are,

```
ev <- eigen(cor(d1))

ev$values

[1] 1.7607207 1.3070246 1.1549692 1.0614750 0.9822735 0.9533021 0.8276333

    0.7148341 0.6893489 0.5484187
```

So, we can determine that the number of factors should be 4. In the results of FA, some coefficients are missing, but this means these coefficients are just too small and not necessary equal to zero.

PROGNOSTICATION IN TERMS OF FACTOR ANALYSIS

From the above propositions, theory, experimentations and analysis we found that negative as well as positive insights in terms of factors that affect the performance of students. The correlation matrix approach necessarily worked as cause and effect analysis is used to find hidden factors for analysis of performance. Thus, by implementing above test we can state that our variables are in a sensible way i.e. The Gender, Age and Failures form positive correlations. The Factor 1 is structured with negative loading of UG Percentage and Free Time for Study. The opposite direction of these two variables indicates that free time availability causes negative impact on UG Percentage. Thus to improve UG Percentage, consumption of free time for study is required for optimum extent. The Factor 2 is structured with positive association of academic use of internet and the leisure time for friends and negative loading of Career dream. This category of Factor-2 reviles to investigate other hidden variables that affect Career dream. From remaining two variables, it may be interpreted that student is not able to take proper decisions about career choice though there is ample use of internet. The Factor 3 is loaded with positive association of Age, Gender and failures in Past examinations. It is also negatively loaded with variables UG Percentage and Scholarships. The category of Factor 3 reviles that older age and male gender have more failures. This factor also needs identification of new variables like stress, family responsibilities etc. which grow with age of male gender may cause failures which may be improved with availing scholarships to them. This holds true in Indian Educational context as scholarships are primarily assumed as financial support. Factor-4 category is formed by positive association of Age, Scholarship, Family income and Career dream. It is negatively loaded with Gender variable. For Career dream, the variables Age and Scholarship have more association with factor than Family income. These four factors are majorly emphasizing prognostications avenues for increasing performance of the students. Thus, all the four factors seem to have some meanings, and that's why we should keep them.

CONCLUSION

The development of analytics around performance and social learning has been important in increasing awareness of the impact of social dimensions about performance. It was openly understood that many social, habitual and economical aspects are associated with performance of students and mere studying cannot be the sole criteria for performance. However these were hidden attributes and no attempt was made at our University level to scientifically visualize them. This study took it as challenge and using data mining algorithms, such insight was obtained. The specific objective of the undertaken research work to find out if there are any patterns in the available data that could be useful for predicting students' performance. We have scientifically made visible the hidden aspects of which four discovered factors are majorly emphasizing on prognostications possibilities to increase the performance of the students. The utility of these studies lies in the need to undertake corrective measures that improve the academic performance of our students in terms of prognostication actions in their weaker aspects.

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

[11] students using data mining. Paper presented at the Sixth ACM SIGKDD International Conference, Boston, MA (Conference Proceedings; p. 457-464).

[12] 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.

[13] 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.

[14] Pavel Berkhin, Survey of Clustering Data Mining Techniques, Accrue Software, available at www.cc.gatech.edu/~isbell/reading/papers/berkhin02survey.pdf.

[15] K.Sasirekha, P.Baby, Agglomerative Hierarchical Clustering Algorithm- A Review , International Journal of Scientific and Research Publications, Volume 3, Issue 3, March 2013 1 ISSN 2250-3153.

[16] Nikhil Rajadhayx et al , Data mining in Educational Domain , retrieved from http://arxiv.org/pdf/1207.1535.pdf.

[17] Gordon Linoff, Michael J, et al , Data Mining Techniques , 3e, Wiley Publications.

[18] Eko Indrato , edited notes on Data Mining, retrieved from www. Http://recommender-systems.readthedocs.org/en/latest/datamining.html.

[19] 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.

[20] Keerthiram Murugesan, Jun Zhang , Hybrid Hierarchical Clustering: An Experimental Analysis , Technical Report: CMIDA-hipsccs #001-11, retrieved from www.cs.uky.edu/~jzhang pub/t echrep.html.

[21] Introduction to factor analysis, web resource www.yorku.ca/ptryfos/f1400.pdf

[22] Rietveld, T. & Van Hout, R. (1993). Statistical Techniques for the Study of Language and Language Behaviour. Berlin – New York: Mouton de Gruyter.

[23] Habing, B. (2003). Exploratory Factor Analysis. Website:

[24] http://www.stat.sc.edu/~habing/courses/530EFA.pdf (accessed 10 May 2004).

[25] Field, A. (2000). Discovering Statistics using SPSS for Windows. London – Thousand Oaks – New Delhi: Sage publications.

[26] Web resource available at http://stats.stackexchange.com/questions/44060/choosing-number-of-principal-components-to-retain

[27] Web resource available at https://www.r-project.org

[28] Web resource available at http://www.utexas.edu/courses/schwab/sw388r7/

AUTHORS' BIOGRAPHY

Dr. Aniket Muley, he is an Assistant Professor in Statistics. He has 10+ papers in International Conferences and Journals.

Dr.Parag Bhalchandra, he is an Assistant Professor in Computer Science. He has 15+ papers in International Conferences and Journals.

Dr. Mahesh Joshi, he is an Assistant Professor in Educational. He has 10+ papers in International Conferences and Journals.

Mr. Pawan Wasnik, he is a Research Scholar in Computer Science. He has 3+ papers in International Conferences and Journals.