Logistics of Data Mining Techniques in Education, Assessing Academic Performance of Self-Financing Arts and Science College Students

R. Senthil Kumar[1] and K. Arulanandam[2]

[1]Research Scholar, Department of Computer Science, Periyar University
Salem, Tamilnadu, India
Email ID : senthil.r.india@gmail.com

[2]Asst. Prof.&Head, Department of Computer Science, GTM College
Gudiyattam, Tamilnadu, India
Email ID : arulatgtmc@gmail.com

ABSTRACT--- ***Data mining methods are often implemented at advanced universities today for analyzing available data and extracting information and knowledge to support decision-making. University management focus more on the profile of admitted students, getting aware of the different types and specific students' characteristics based on the received data. Educational data mining is an emerging field for knowledge discovering from large scale of educational data. To identify the improvement pattern of the academic performance of students studying in self-financing arts and science colleges, data were collected with the information like father's education, mother's education, classification, subject, college location, facilities, etc from 1398 students through questionnaire. Classification analysis of associated factors with academic performance identified the urban residents, higher parental education, science students who utilise college facilities and with higher skilled knowledge, time spending, liking college with more faculty concern including good presentation of teaching materials increased the regular students significant with academic performance. The factor analysis identified the four factors, faculty concern, classification, location of residence and college and parental education which explained 38.9% of total variation. K-means cluster analysis reduced to five clusters the student data, the first cluster composed with maternal education. Students of the other clusters identified facilities like students cognitive factors. College and home location and finally the subject taken was the fifth cluster.***

Keywords--- Educational data mining, Association Function, Factor Analysis, K-means Cluster Analysis

1. INTRODUCTION

Data mining can be implemented in different areas such as Fraud detection, Medical, Education, Banking, Marketing and Telecommunications. Education is the platform on which a society improves the quality of its citizens. To improve on the quality of education, there is a need to be able to predict academic performance of the students [1]. Application of data mining in education sector is an emerging trend . The data mining terms, tasks, techniques and application can be used to develop data mining in education sector [2]. Now Data mining technique is useful to study social problems especially in education, mainly in higher studies and literacy patterns. Methods of data mining usefulness for student ability, skill and understanding, facilities, faculty efforts identify the students, especially in education process are affected in the academic performance by various factors.

The prediction of academic performance of students is really challenging because it depends on various factors like personal, socio-economic, psychological and other environmental attributes. It informs domain experts about the optimal sequencing of instruction in order to achieve the best tutoring for students. This should help researchers in the education research community to better model students' knowledge and performance in intelligent tutoring systems more accurately.

Data mining provides many techniques like Classifications, Clustering, Naïve Bayesian, decision trees, neural networks and Fuzzy rules. In this paper, the association of cross tabulation, factor analysis and clustering techniques were used. Data mining provides many tasks that could be used to study the student's performance. Scope of data mining includes statistics, artificial intelligence and machine learning. The artificial intelligence is based on heuristics and it represents an attempt to approach statistical problems similar to the human way of thinking [3]. Students' abilities, motivation, and behaviour work in tandem to influence their academic performance. If students are lacking in even one of these areas, their performances will be significantly lower. Statistics is the basis of most technologies that are used in the process of knowledge discovery in databases.

2. REVIEW OF LITERATURE

The universities desire to improve their educational quality through the usage of data mining in higher education to help the universities, educators, and students to improve their performance has become more and more attractive to both university managers and researchers [4]. Data Mining can be used in the educational field to enhance our understanding of the learning process to focus on identifying, extracting and evaluating variables related to the learning process of students as described by Alaa el-Halees [5]. Educational Data Mining researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning, computer-adaptive testing. EDM appears to be growing in size rapidly. At this point, educational data mining methods have had some level of impact on education and related interdisciplinary fields [6]. Modeling student individual differences in these areas enables software to respond to those individual differences, significantly improving student learning [7].

Data mining techniques feature selection and classification trees to explore the socio-demographic variables such as age, gender, ethnicity, education, work status, and disability and study environment course programme and course block that may influence persistence or dropout of students, identifying the most important factors for student success and developing a profile of the typical successful and unsuccessful students [8]. EDM methods has been looking for empirical evidence to refine and extend educational theories and well-known educational phenomena, towards gaining deeper understanding of the key factors impacting learning, often with a view to design better learning systems, center around how educational data mining methods can support the development of more sensitive and effective e-learning systems.

Attempts were used to predict student failure by applying and comparing four data mining algorithms − Decision Tree, Random Forest, Neural Network and Support Vector Machine. Uses of decision trees, neural networks and linear discriminant analysis also used to categorise students' performance and to model their performance [9]. Classification algorithms are used to predict the performance of computer science students [10]. Educational data mining methods have prompted the researchers to model relevant student variables in real-time, including higher-level constructs than were earlier possible. Researchers have also been able to extend student modeling even beyond educational software, towards figuring out what factors are predictive of student failure or non-retention in college courses or in colleges altogether [11].

S.No	Variable Name	Description	Domain
1	RESLOC	Residence Location	{Rural, Semi Urban, Urban}
2	COLLOC	College Location	{Rural, Semi Urban, Urban}
3	FATEDU	Father Education	{Primary, Secondary, Higher}
4	MOTEDU	Mother Education	{Primary, Secondary, Higher}
5	CLASSIFICATION	Classification	{Arts, Science}
6	SUBJECT	Subject	{Tamil, English, History, Economics, B.Com, BBA, BCA, Maths, Physics, Chemistry, Botany, Zoology, Comp.Sci.}
7	CANTEEN	Canteen Facility	{Excellent, Very Good, Good, Fair}
8	DRIWATER	Drinking Water	{Excellent, Very Good, Good, Fair}
9	LIKECOL	Liking college	{Enthusiastic, I like it, Neutral}
10	STUMAT	Study material	{By Faculty, Text Book, Reference Book}
11	INTFAC	Internet Facility	{Yes, No}
12	JOBAFF	Job Affect College Work	{ Enhances, Not interfere, Takes some time, Take lot of time}
13	FACCON	Faculty Concern	{Excellent, Very Good, Good, Fair}
14	QUAPRE	Quality of Presentation	{Clear and Informative, Clear}
15	KNOWSKILL	Knowledge and skill	{ For specific job, Very much, Quite a bit, some}
16	UNDYOU	Understanding Yourself	{ Very much, Quite a bit, some}
17	ATTENDANCE	Attendance percentage	{>60, 61-70, 71-80, >90, <60}
18	COMMSKI	Communication Skill	{ Excellent, Very Good, Good, Fair}
19	TIMESPEND	Time spent for reading	{2 hours, 3 hours, 4 hours, 5 hours}
20	LANPRO	Language Proficiency	{ Improved dramatically, Improved somewhat, Not improved, Did n't take the course}

To address the student performance prediction problem, many works have been published but most of them rely on classification/regression methods such as Bayesian networks, logistic regression, linear regression, decision trees, neural networks and support vector machines in predicting student performance.

Significant Attributes were

Non-Significant Attributes were

S.No	Variable Name	Description	Domain
1	GENDER	Gender	{Male, Female}
2	FATOCU	Father Occupation	{Agriculture, Business, Service, Teacher, Others}
3	MOTOCU	Mother Occupation	{Home Maker, Business, Service, Teacher}
4	INCOME	Income	{<Rs. 50,000PA, Rs. 50,000 to 5,00,000, >5,00,000
5	COLINF	College Infrastructure	{Excellent, Very Good, Good, Fair}
6	COLCAM	College Campus	{Excellent, Very Good, Good, Fair}
7	LIBFAC	Library Facility	{Excellent, Very Good, Good, Fair}
8	LABFAC	Lab Facility	{Excellent, Very Good, Good, Fair}
9	MEDFAC	Medical Facility	{Excellent, Very Good, Good, Fair}
10	SECSYS	Security System	{Excellent, Very Good, Good, Fair}
11	TRAVBY	Travel By	{College Bus, Private Bus, Own Vehicle}
12	BUSFAC	Bus Facility	{Excellent, Very Good, Good, Fair}
13	MEDINS	Medium of Instruction	{Tamil, English}
14	LEACEN	Learning Centre	{Regularly, Sometimes, Ever}
15	KNWFAC	Knowledge of Faculty	{Excellent, Very Good, Good, Fair}
16	TEAFAC	Teaching By Faculty	{Excellent, Very Good, Good, Fair}
17	PREMAT	Presentation of Material	{Excellent, Very Good, Good, Fair}
18	INTACT	Interested in Activities	{Yes, No}
19	INTSPO	Interested in Sports	{Yes, No}
20	INDVIS	Industrial Visit	{Yes, No}
21	PLACEM	Placements	{Yes, No}
22	SCHLAR	Scholarship	{Yes, No}

Data mining helps to extract the original and the valuable data from the large amount of dataset. Information such as age, parents' qualification, parents' occupation, academic record, attitude towards university was collected from the students to forecast those students requiring monitoring.

Many studies included a wide range of potential predictors, including personality factors, intelligence and aptitude tests, academic achievement, previous college achievements, and demographic data and some of these factors seemed to be stronger than others, however there is no consistent agreement among different studies. The factors that are associated with student failure or non-retention in courses. Key area of application has been in the improvement of student models. Student models represent information about a student's characteristics or state, such as the student's current knowledge, motivation, meta-cognition, and attitudes.

3. MATERIALS AND METHODS

The descriptive study to assess educational data mining technique using the information collected from 1398 students of self-financing arts and science colleges studying final year under graduation course in Thiruvannamalai district of Tamilnadu, India, were formed for this study. The students of self-financing arts and science colleges identified randomly selected after their colleges were identified by simple random sample selection, the students with the selected subject from nine colleges, not more than 20 students were identified in each of the subject classes among the nine selected colleges.

The questionnaire prepared containing the academic performance of the students along with the associated functions of socio economic and demographic characteristics was subjected to a pilot study and modified. The reliability co-efficient for the questionnaire chronboch alpha was 0.73 which identified a good reliability of the questionnaire.

Data Mining techniques are such as Classification (decision tree), clustering, association rule and statistical methods. It also provides an effective methodology to compare the various classification of the training data and evaluate the test and validation datasets**.** Association functionand data mining techniques were used. Factor analysis and cluster analysis were also performed. Information like father education, mother education, classification, subject, college location, facilities, etc. are collected from students through questionnaire. Clustering and classification both are very useful to improve the performance on education sector [12]. DM tools which are available on the market are usually the products of companies coming from the databases, hardware, statistical analysis or other related fields [13]. Each tool has different features and requirements as SPSS Clementine, Matlab, and ANOVA. SPSS Clementine is one of the most widely used DM software suites. The software provides various DM modeling techniques such as classification (decision tree), clustering, association rule, and statistical methods. It also provides an effective methodology to compare the various classifications of the training data and evaluate the test and validation datasets [14]. Clementine supports the entire DM process, from performing data input, data cleansing, data transformation and producing a result.

Descriptive statistics was used to find the patterns of the socio-demographic and academic performance variables and their distributional aspects. The test of association patterns were studied to describe the data mining of association functions of the study variables. To identify the reduced patterns of important significant features, factor analysis was carried out. The cluster analysis was used by k-means clustering algorithm to identify the similarity of the students.

4. RESULTS AND FINDINGS

Among the respondents by gender the academic performance was not significantly different (p=0.217). An opposite correlation was found among rural students while positive relationship were observed among the small town and urban students. (Figure 1)

The association of students performance with the residence of the student showed significantly associated with performance. In the below 60 mark category poor performance the rural students were the highest (72.6%). The decreasing trend was found from good response to the poor of the marks to small town and rural while the trend in the rural increases in the proportion of the students as compared to decreasing marks.

The relationship between father's education and the students' marks is highly significant (p=0.0001); were the father's education level is higher secondary the pattern of positive relationship is found while for primary educated father the opposite correlation by the marks is observed. Below 60 marks category is the highest (60.7%). Similar pattern is observed by mother's education i.e by education of parents positive by increasing marks were obtained by the students as education level of parents increase. The proportion of better marks were also increasing. There is no significant difference by father's occupation (p=0.228),mother's occupation (p=0.268) and income (p=0.123).

(Figure 2) The classification is significantly associated with good performance (p=0.0001). Higher percentage of students (58.7%) in science classification obtained above 80 of marks than the percentage of arts students (41.3%) significantly. In the below 60 marks poor performance group more arts students (66.2%) were found than the science students (33.8%).

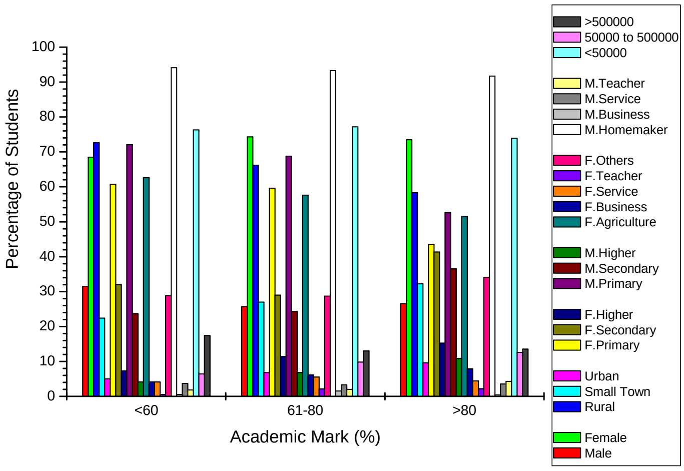


Figure 1: The pattern of socio-economic and demographic characters by academic performance

The subject taken by the students associated the performance significantly (p=0.0001). Among the arts subject students the largest proportion of English students (47.9%) were found in the below 60 marks group. This makes significant difference by classification found earlier between arts and science students. Maths subject students in all the science classification students were with the highest proportion (17.4%) in the 60-80 marks group. Chemistry (15.7%) and Computer Science (12.6%) students also obtained higher proportion followed by science students.

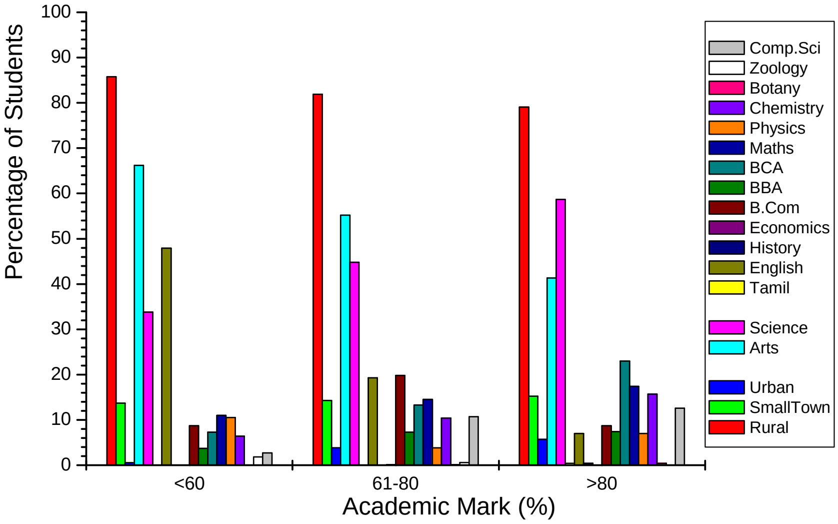


Figure 2: The pattern of college location and subjects undertaken by academic performance

In Figure 3, College location is the factor by which the student performance differs significantly (p=0.046). Poor marks were obtained (below 60%) while the college is located in rural areas (85.8%) in highest proportion, while the college was situated in small town and urban areas, increasing marks were found in the above 80% mark group.

Table 1 : Infrastructural facilities structural availability by academic performance of students

Variable Name	Description	Academic marks (%)								
		Above 80 (230)		61-80 (949)		Below 60 (219)		Total (1398)		Significance
		No	%	No	%	No	%	No	%	
Medium of Instruction	Tamil	37	16.1	99	10.4	26	11.9	162	11.6	0.055^{NS}
	English	193	83.9	850	89.6	193	88.1	1236	88.4	
Liking college	Enthusiastic	41	17.8	92	9.7	17	7.8	150	10.7	0.000^{***}
	I like it	146	63.5	692	72.9	180	72.9	1018	72.8	
	Neutral	43	18.7	165	17.4	22	10.0	230	16.5	
Study material	By faculty	76	33.0	308	32.5	60	27.4	444	31.8	0.003^{**}
	Text books	78	33.9	362	38.2	111	50.7	551	39.4	
	Ref. books	76	33.0	278	29.3	48	21.9	402	28.8	
Learning Centre	Regularly	54	23.5	263	27.7	69	31.5	386	27.6	0.339^{NS}
	Sometimes	142	61.7	566	59.7	127	58.0	835	59.8	
	Ever	34	14.8	119	12.6	23	10.5	176	12.6	
Internet Facility	Yes	106	46.3	375	39.6	106	48.4	587	42.0	0.021^{*}
	No	123	53.7	573	60.4	113	51.6	809	58.0	
Job Affect College Work	Enhances	72	31.3	255	26.9	57	26.0	384	27.5	0.000^{***}
	Not interfere	109	47.4	338	35.7	74	33.8	521	37.3	
	Takes some time	32	13.9	237	25.0	57	26.0	326	23.3	
	Takes lot of time	17	7.4	118	12.4	31	14.2	166	11.9	

By college infrastructure and college campus there is no significant difference in the performance of student in the academic marks. While the canteen facility was fair, the highest performance was obtained in higher proportion (39.6%). In the other better groups the performance is least in higher proportion among below 60% marks group. By library facility and lab facility there is no significant difference in the performance of students in the academic mark. Drinking water facility (p=0.0001) is the factor by which the student performance differs significantly. Negative relationship was found between the drinking water facility and the academic performance. Where the water facility is fair, the higher proportion of the students (25.7%) obtained the highest marks and the proportion is increasing to the higher marks to lower marks group as the water facility level increasing. In the canteen facility also similar results were obtained

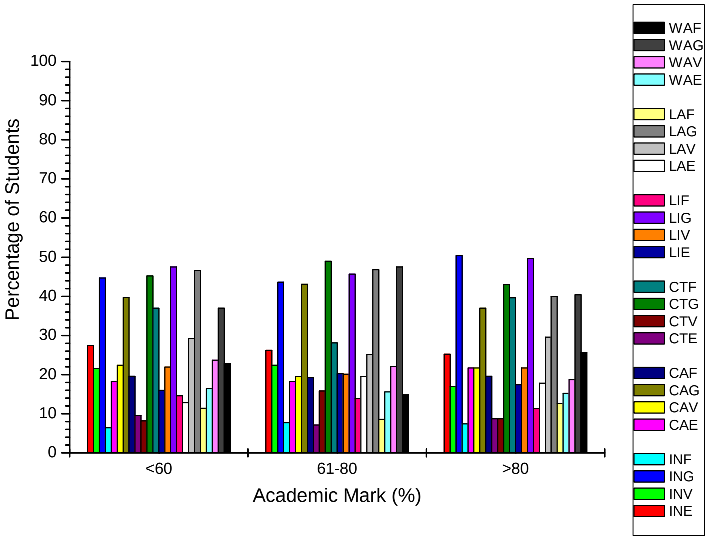


Figure 3: The pattern of college infrastructure and facilities by academic performance

By medical facility, security system, travel by and bus facility there is no significant difference. As shown in Table 1 medium of instruction (p=0.055) and learning centre (p=0.339) were not significant. Study material (p=0.003), liking college (0.0001), internet facility (p=0.021) and job affecting the college work (0.0001) differed significantly. The liking of the college enthusiastically is positively related highest marks (above 80) were obtained in higher proportion (17.8%). Similarly the neutral students also had the best performance (18.7%).

Among the study materials by faculty (33.0%) and reference books (33.0%) significantly the highest performance was found in the higher proportion in the above 80 academic marks. In the text book study material the lowest marks below 60 were obtained in the higher proportion (50.7%).

Internet facility is found to be negatively related to academic performance. Higher proportion of student (48.4%) obtained lowest mark (below 60) when they are not using the internet facility, higher performance (above 80) were obtained (53.7%). By the non-interfere job category in the higher percentage (47.4%) were obtained while taking some time (26.0%) and lot of time (14.2%) categories of job affecting the performance is lower.

Table 2 : Faculty involvement by academic performance of students

Variable Name	Description	Academic marks (%)								
		Above 80 (230)		61-80 (949)		Below 60 (219)		Total (1398)		Significance
		No	%	No	%	No	%	No	%	
Knowledge of Faculty	Excellent	57	24.8	213	22.5	65	29.7	335	24.0	0.342^{NS}
	Very Good	67	29.1	299	31.5	55	25.1	421	30.1	
	Good	96	41.7	387	40.8	88	40.2	571	40.9	
	Fair	10	4.3	49	5.2	11	5.0	70	5.0	
Faculty Concern	Excellent	69	30.0	241	25.4	68	31.1	378	27.1	0.027*
	Very Good	64	27.8	341	36.0	60	27.4	465	33.3	
	Good	86	37.4	321	33.9	73	33.3	480	34.4	
	Fair	11	4.8	45	4.7	18	8.2	74	5.3	
Teaching by Faculty	Excellent	78	33.9	295	31.1	67	30.6	440	31.5	0.724^{NS}
	Very Good	66	28.7	288	30.3	65	29.7	419	30.0	
	Good	81	35.2	336	35.4	76	34.7	493	35.3	
	Fair	5	2.2	30	3.2	11	5.0	46	3.3	
Presentation of Material	Excellent	76	33.0	280	29.5	58	26.5	414	29.6	0.146^{NS}
	Very Good	75	32.6	317	33.4	78	35.6	470	33.6	
	Good	72	31.3	331	34.9	71	32.4	474	33.9	
	Fair	7	3.0	21	2.2	12	5.5	40	2.9	
Quality of Presentation	Clear and informative	94	40.9	498	52.5	110	50.2	702	50.3	0.007**
	Clear	136	59.1	450	47.5	109	49.8	695	49.7	

From Table 2 by knowledge of faculty (p=0.342), teaching by faculty (p=0.724) and presentation of material (p=0.146) are not associated with the academic performance. Faculty concern (p=0.027) and the quality of presentation (p=0.007) were significantly influencing the academic performance. Middle level concern of faculty were related to middle level of performance very good (36%), good (33.9%) fair group were the highest (8.2%) in the lowest mark group (below 60%). While excellent concern is fairly equal proportion in the lowest (31.1%), highest (30%) mark group. When the quality of presentation is clear better performance (above 80%) was found in higher proportion (59.1%) while it is clear and informative medium level of marks were obtained (60-80) in higher proportion (52.5%).

In Table 3, Knowledge and skill (p=0.005), understanding yourself (p=0.010), attendance percentage (p=0.0001), communication skill (p=0.0001), time spent for reading (p=0.0001), language proficiency (p=0.0001) were significantly correlated with the academic performance. When the knowledge and skill is quite a bit (16%) and very much (53%) the academic performance is the least. For some (17%) and specific job (26.8%) category had better performance in higher proportion. By understanding yourself quite a bit (24.2%) and some (12.3%) lead the to lowest performance, while understanding yourself very much leads to better performance (74%).

When the attendance was above 90% the best performance was observed in higher proportion (79.2%) and it is slowly reduced to worse as the attendance decreased the academic performance become less in all the below attendance categories. When the communication skill was fair (10%) and good(49.3%) the lowest mark was obtained in higher proportion. When the communication skill was very good, middle level mark was obtained in higher proportion(20.8%). In the excellent communication group best performance (above 80) marks were obtained in higher proportion (29.6%). This result identifies the positive relationship of the communication skill with the academic performance. Highest time spent (5 hrs) leads to best performance in higher proportion (31.3%), 4 hrs time spending also resulted in similar way but middle level time spending gave the middle level marks in higher proportion (31%) the least time spending (2 hrs or less) produced the worst performance (below 60) marks in higher proportion(58.9%).This shows that positive relationship of time spent for reading is positively related. Those who did not take language proficiency course had the best performance. Among those who under took the course improved dramatically and not improved group performed at medium level. But the improved some what level had the least performance. By interest in activities , interest in sports, industrial visit, placements and scholarship were not related with the academic performance.

Table 3: Self motivation by academic performance of student

Variable Name	Description	Academic marks (%)								
		Above 80		61-80		Below 60		Total		Significance
		No	%	No	%	No	%	No	%	
Knowledge and skill	For specific job	64	27.8	273	28.8	39	17.8	376	26.9	0.005^{**}
	Very much	105	45.7	463	48.8	116	53.0	684	48.9	
	Quite a bit	22	9.6	101	10.6	35	16.0	158	11.3	
	Some	39	17.0	112	11.8	29	13.2	180	12.9	
Understanding Yourself	Very much	172	74.8	702	74.0	139	63.5	1013	72.5	0.010^{**}
	Quite a bit	31	13.5	146	15.4	53	24.2	230	16.5	
	Some	27	11.7	101	10.6	27	12.3	155	11.1	
Attendance percentage	<60	4	1.7	21	2.2	13	5.9	38	2.7	0.0001^{***}
	60-70	5	2.2	23	2.4	15	6.8	43	3.1	
	71-80	3	1.3	73	7.7	19	8.7	95	6.8	
	81-90	45	19.6	231	24.3	56	25.6	332	23.7	
	91-100	173	75.2	601	63.3	116	53.0	890	63.7	
Communication Skill	Excellent	68	29.6	178	18.8	48	21.9	294	21.0	0.0001^{***}
	Very Good	53	23.0	270	28.5	41	18.7	364	26.0	
	Good	101	43.9	441	46.5	108	49.3	650	46.5	
	Fair	8	3.5	60	6.3	22	10.0	90	6.4	
Interested in activities	Yes	193	83.9	761	80.2	177	80.8	1131	80.9	0.436^{NS}
	No	37	16.1	188	19.8	42	19.2	267	19.1	
Interested in Sports	Yes	145	63.0	636	67.1	154	70.3	935	66.9	0.257^{NS}
	No	85	37.0	312	32.9	65	29.7	462	33.1	
Time spent for reading	2 hours	85	37.0	424	44.7	129	58.9	638	45.7	0.0001^{***}
	3 hours	44	19.1	298	31.4	50	22.8	392	28.1	
	4 hours	29	12.6	96	10.1	9	4.1	134	9.6	
	5 hours	72	31.3	130	13.7	31	14.2	233	16.7	
Language Proficiency	Improved dramaticall y	82	35.8	378	39.8	76	34.9	536	38.4	0.0001^{***}
	Improved somewhat	82	35.8	314	33.1	104	47.7	500	35.8	
	Not improved	10	4.4	85	9.0	8	3.7	103	7.4	
	Didn't take the course	55	24.0	172	18.1	30	13.8	257	18.4	
Industrial Visit	Yes	58	25.2	229	24.3	45	20.7	332	23.9	0.477^{NS}
	No	172	74.8	715	75.7	172	79.3	1059	76.1	
Placements	Yes	104	45.6	440	46.5	101	46.5	645	46.4	0.969^{NS}
	No	124	54.4	506	53.5	116	53.5	746	53.6	
Scholarship	Yes	87	38.0	360	37.9	92	42.2	539	38.6	0.495^{NS}
	No	142	62.0	589	62.1	126	57.8	857	61.4	

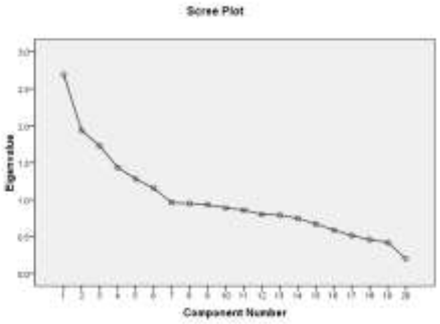
Factor Analysis:

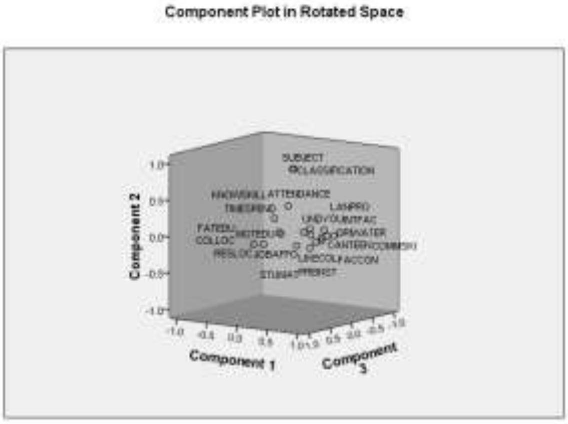
Table 4 : Total Variance Explained

Rotation Sums of Squared Loadings					
Factor No.	Factor Name	Component	Eigen Value	% of Variance	Cumulative %
1	FACCON	0.646	2.551	12.753	12.753
	COMMSKI	0.600			
2	CLASSIFICATION	0.896	1.921	9.606	22.359
	SUBJECT	0.888			
3	COLLOC	0.805	1.690	8.451	30.811
	RESLOC	0.660			
4	FATEDU	0.859	1.627	8.135	38.946
	MOTEDU	0.850			

From Table 4, the factor analysis identified four factors, faculty concern, classification, location and parental education which explained 38.9% of total variation. The highest and nearest factor loadings identified by the first factor are FACCON(0.646) and COMMSKI(0.600) which explained 12.75%. CLASSIFICATION(0.896) and SUBJECT(0.888) formed the second factor which explained 9.6%, while COLLOC(0.805) and RESLOC(0.660) are the third factor explaining 8.4% . The fourth factor explained parental education which FATEDU(0.859) and MOTEDU(0.850) explained 8.14% . The screeplot shown in Fig. 4 with eigen values of all the variables included in the analysis as the number of 20 components of factors obtained.

Fig. 4 Depicting Eigen value by the component number





Cluster Analysis:

By using k-means cluster analysis 5 clusters identified, the number of students in each cluster is shown in Table 5. Table 6 identified the variables which combined together with the five clusters. The final cluster center are shown for the five factors in the Table. The cluster centers identifies the distance between the cluster and the variables. The five clusters identified the factors related to the performance of the number of students in each cluster is shown in the Table 6. The first cluster identified with the lowest distance with the mother education while the second cluster identified with the factor of liking college, students material, and job affecting. The third cluster grouped with communication skill, language proficiency and the knowledge skill is called the cognitive factor. The fouth cluster consisted of presentation, understanding yourself, classification, college location, residence location, subject taken, internet facility and father education. Fifth cluster consists of drinking water, subject taken, attendance and canteen facility. These five clusters converged to five clusters finally among the performance related characteristics.

Table 5: Number of cases in each Cluster

Cluster	1	306.000
	2	147.000
	3	324.000
	4	279.000
	5	333.000
Valid		1.389E3
Missing		9.000

Table 6: Final Cluster Centers

	Cluster				
	1	2	3	4	5
FACCON	2	2	2	2	3
COMMSKI	2	2	2	2	3
DRIWATER	3	3	3	2	3
CANTEEN	3	3	3	3	3
LANPRO	2	2	2	2	3
LIKECOL	2	2	2	2	2
QUAPRE	2	2	1	1	2
KNOWSKILL	2	2	2	2	3
UNDYOU	1	1	1	1	2
JOBAFF	2	2	2	2	2
CLASSIFICATION	1	2	1	2	2
SUBJECT	2	13	5	8	9
ATTENDANCE	3	4	4	4	4
STUMAT	2	2	2	2	2
TIMESPEND	2	2	2	2	2
COLLOC	1	1	1	2	1
RESLOC	2	2	2	2	1
INTFAC	2	1	2	1	2
FATEDU	2	2	1	2	2
MOTEDU	1	1	1	1	1

5. DISCUSSION

The significant result of this study by the patterns of academic performance brings out the important factors in education data mining to carry out research with the associated factors. Many universities are extremely focused on assessment, thus, the pressure on “teach to the test” leads to a significant amount of time spending for preparing and taking standardized tests [15]. Those who study four hours or more has gained more higher marks than those who spent less time significantly the impact of ability for academic performance to be much higher for students who spend more time studying than for those who spend less. Many studies found a strong relationship positively related between students attendance and academic performance our study indicated those who attended more than 90% only had higher percentage of academic performance than those who attended less. Parents' occupation plays a major role in predicting the students grades but in our study parents' occupation failed to be a significant factor.

The most important teaching skill is to establish relationship between teacher and student; the conducted studies with regard to the situation of communication messages have shown that the message sender should not only know the topic of the communication and have enough information about it, but also have information about how to present it. In our study those who had excellent communication skill performed the best three other categories very good, good and fair.

Those who did not take proficiency course for language performed our study indicated that the best performances could avoid the language proficiency course resulted in inverse relationship of language proficiency with the academic performance. The majority of obtained scores in knowledge was less than the mean knowledge score of all the people most studied people had knowledge about effective factors for behavioural change including physical reasoning and emotional skills, individual and network of family and social structures. Those who had some knowledge in our study performed better than those who has better knowledge and skill and also our study showed those who understand self very much performed better that other categories of understanding less.

Water may to a smaller extent boost attention during a mental activity significantly; better performance of all participants in the second measurement session suggests that practice has had a huge effect on the attention test results after drinking water. We don't know enough about the recovery of these functions that is how exactly and at what dynamics water improves these functions following dehydration and tiredness. We may assume that the relationship between the decline in cognitive functions due to dehydration and then recovery due to hydration is not linear, which creates the need for further research in this area [16]. Our study resulted in negative relationship with academic performance.

The huge proportion of urban students are good in programming skill compared to rural students [17] but in our study urban students perform better than the rural students. The parents who do not go beyond elementary or secondary schools are not able to give proper health to their children in the educational problems that children's academic achievements in most cases do not mostly depend on parental educational level but contrary to this our study students positive relationships were found for both father's and mother's educational levels. Few other studies also identified the association of the influence of parental educational level on secondary school students academic achievements. There is no significant difference in the achievement mean score of male and female students in urban areas. The findings established homogeneity among male and female students in terms of academic achievements irrespective of school locations that gender and location do not effect the negative relationship between student problems and academic performance. Our study also shows similar results that by gender the academic performance was not different, irrespective of location. The schools in urban area achieved more than the schools in the rural areas in science subject [18].

6. CONCLUSION

In this study, we conclude that science students' performance, especially, in maths subject is better than the arts subject students. Also that the urban college location and urban residence location influences the academic performances. The factor analysis reduced four factors identified, faculty concern, communication skill of the students plays a leading role in academic performance followed by the students classification of science subject and same classification. Urban location of college and residence along with the better parents' education brings out the better performance of the students.

7. ACKNOWLEDGEMENT

R. Senthil Kumar thanks Dr. K.G. Selvaraj, Rtd. Prof., Department of Biostatistics, Christian Medical College, Vellore, Tamilnadu, India.

8. REFERENCES

[1] Kolo, D.K., S.A. Adepojub and J.K. Alhassanb, "A Decision Tree Approach for Predicting Students Academic Performance", I. J. Education and Management Engineering, 5, 12-19, 2015.

[2] Mohd Maqsood, A., "Role of data mining in education sector", International Journal of Computer Science and

Mobile Computing, 2,4, 374 – 383, 2013.

[3] Sumathi, S. and S.N. Sivanandam, Introduction to Data Mining And Its Applications. Netherlands, Sringer. ISBN 978-3-540-34351-6, 2006.

[4] Nguyen, T.N., L. Drumond, A. Grimberghe and L.S. Thieme, "Recommender System for Predicting Student Performances", Procedia Computer Science, 1, 2811–2819, 2010.

[5] Alaa el-Halees., "Mining Students Data to Analyze e-Learning Behavior", A Case Study, 2009.

[6] Baker,R.S.J.D. and K. Yacef, "The State of Educational Data Mining in 2009, A Review and Future Visions", Journal of Educational Data Mining, 1,1, 3-16, 2009.

[7] Corbett, A.T., "Cognitive Computer Tutors, Solving the Two-Sigma Problem", In Proceedings of the International Conference on User Modeling, 137-147, 2001.

[8] Kovačić, Z., "Early Prediction of Student Success, Mining Students Enrolment Data", In Proceedings of Informing Science & IT Education Conference, 647-665, 2010.

[9] Vandamme, J., N. Meskens and J. Super, "Predicting Academic Performance by DataMining Methods", Education Economics, 15,4, 405-419, 2007.

[10] Kotsiantis, S., C. Pierrakeas and P. Pintelas, "Prediction of Student's Performance inDistance Learning Using Machine Learning Techniques", Applied Artificial Intelligence, 18, 5, 411-426, 2004.

[11] Dekker, G., M. Pechenizkiy and J. Vleeshouwers, "Predicting StudentsDrop Out, A Case Study", In Proceedings of the International Conference on EducationalData Mining, 2009.

[12] Suman and M. Pooja, "A Comparative Study on Role of Data Mining Techniques in Education - A Review", International Journal of Emerging Trends & Technology in Computer Science, 3, 3, 65-69, 2014.

[13] Mihai, A. and C. Daniel, "Commercially Available Data Mining Tools used in the Economic Environment", Database Systems Journal, 1, 2, 45–54, 2010.

[14] Ogor, E. N., and C. Islands, "Student Academic Performance Monitoring and Evaluation Using Data Mining Techniques", In, CERMA '07 Proceedings of the Electronics, Robotics and Automotive Mechanics Conference , 354-359, 2007.

[15] Feng, M., N. Heffernan and K. Koedinger, "Addressing the assessment challenge with an online system that tutors as it assesses", User Modelingand User-Adapted Interaction, 19, 3, 243–266, 2009.

[16] Krecar, I.M., M. Kolega and S. F. Kunac, "The Effects of Drinking Water on Attention", Procedia - Social and Behavioral Sciences, 159, 577–583, 2014.

[17] Ramesh, V., P. Parkavi and K. Ramar, "Predicting Studentperformance", A Statistical and Data Mining Approach, International Journal of Computer Applications, 63, 8, 36-39, 2013.

[18] Agbaje, O. Rashidat, Awodun and O. Adebisi, "Impact of School Location on Academic Achievement of Science Students in Senior Secondary School Certificate Examination", International Journal of Scientific and Research Publications, 4, 9, 1-4, 2014.