IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-5, May 2018 E-ISSN: 2347-2693

Data mining in the academic performance of self – financing arts and science college students using K-Means clustering algorithm

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

[1*]Department of Computer Science, Periyar University, Salem, Tamilnadu, India
[2] Department of Computer Science, GTM College, Gudiyattam, Tamilnadu, India

*Corresponding author: senthil_geetha@rediffmail.com, Tel.: 09443401098

Available online at: www.ijcseonline.org

Accepted: 21/May/2018, Published: 21/May/20182018

Abstract**:** To impart quality education and to improve the quality of managerial decisions are the main objective of any higher educational institution and also to reduce the drop out ratio to a significant level and to improve the performance of students. To apply data mining techniques by weka software for the academic performance related variables are analyzed. To segment students into groups according to their characteristics cluster analysis was used in this study. This includes the student's socio economic characters, skill development characters, motivational characters and infrastructural facilities. The application technique will help to classify the best performance of students. The academic performance of 1398 self – financing arts and science college students were selected during their final year of the study. The useful information and related attributes were stored in Educational database and to extract meaningful information and to develop the significant relationship clustering methods were used in this paper. To enhance the quality of educational system by analyzing and improving student's best performance related characters were identified.

Keyword: - Educational data mining, K-Means clustering, Weka Interface, Academic performance

I. Introduction

One of the basis is to monitor the progression of student's performance in higher education is performance evaluation. It is possible to discover the key characteristics from the student's performance and possibly to use those characteristics for future prediction with the use of data mining techniques, such as clustering.

A very significant technology in data mining is cluster analysis. Datasets are divided into various meaningful groups. A cluster is an aggregation of data items with common similarities based on the measurement of same kind of information. Using simple and efficient analysis tool Weka Interface, K-Means clustering algorithm it is possible to identify student's performance in higher education. Extracting previously unknown, valid, positional useful and hidden patterns from large data sets is a process of data clustering. The extents of data mining in educational databases are increasing rapidly. Clustering technique is most widely used to identify student's performance.

Usually this algorithm is used to analyze different factors such as socio economic, skill development, motivation and infrastructural facilities that affect a student's learning behavior and performance during academic career. The student's academic performance depends on diverse factors such as psychological, environmental, socio-economic, and personal variables. The objective of this paper is to predict student academic performance, cluster groups of students with similar performance and to identify the quality of student using data mining techniques.

Each data point belongs to cluster with the minimum mean value, K-Means clustering algorithm partitions *n* data points of the dataset into *k* clusters. Using Euclidean distance formula centroid mean value can be calculated. Section I contains the introduction of higher education data mining, Section II contain literature of the related work, Section III explain the methodology employed for present study, Section IV describes the results of k-means clustering algorithm, Section V contain discussion regarding the improvement of academic performance, Section VI contain the conclusion of the present research work.

II. Related Work

To reduce dropout ratio and to improve the student's academic performance and to enhance quality of education was proposed with k-means clustering algorithm and weka interface [1]. This investigation study may be helpful for teachers as well as students and it is used to predict student's learning activities [2].

Statistical methods play an important role in analyzing and evaluating the performance in college to make academic decisions. In a study k-means grouping method in clustering algorithm has been used to improve the quality of engineering education [3]. This research paper is used for predicting student's performance based on clustering algorithm. In this study the multiple linear regressions is used. It can be identified to predict only one semester percentage of the student's at a time [4].

A trusted model using data mining technique which extracts required information, so that the present education system may adopt this as a strategic management tool. This simple analysis shows that information retrieval from vast data, which can be used for the process of decision making by the management of an educational institution [5]. They evaluated student's performance on basis of class test, mid test and final test. This information will help professor to judge the students fail chance before final exam. The students are grouped into three categories high, medium and low [6]. Most of the student's performance is good in their academic terms but when there is less percentage in attendance they failed to attain their semester marks. The previous records were analyzed in order to improve the performance of students and to determine their behavioural pattern in academic wise. This identifies to assist the difficulty faced by the students to produce more marks in the semester exams and to enhance their co-curricular activities. This is done by clustering the students based on their performance [7].

Five potential faculties' personal and professional credentials are considered. The result analysis for the subjects they have handled is collected around four consecutive years. An effort is made to map the outcome produced with faculty credentials. Based on the performance of students the inferences are drawn. According the data collected from faculties clearly shows that the performance of students is considerably very good. The faculties' contribution is also very high. Identifying slow learners, conducting remedial classes and continuous monitoring of students are regularly done. This can happen only by dedicated, highly committed and experienced staff. The quality staff plays a significant role in promoting higher education [8]. In this study they have proved that student's performance can be predicted by using a data set that consisted of student's gender, parental education, financial background etc. They used Bayesian networks to predict the student's outcome based on attributes like attendance, performance in class tests and assignments [9]. A system for analyzing students' results based on cluster analysis and uses standard statistical algorithms to arrange their scores data according to the level of their performance is described. In this paper, we also implemented k-mean clustering algorithm for analyzing students' result data [10]. The influence of the Mahatma Gandhi National Rural Employment Guarantee Act on the rural India was studied by employing data mining technique [11]. Compacting data sets provides a new and efficient method for discovering frequent pattern and to keep a record of all resulting subset to avoid duplicated generation [12].

III. Methodology

The descriptive study is to assess educational data mining technique using the information collected from the students. The self-financing arts and science college students studying final year under graduation course in Thiruvannamalai district of Tamilnadu, India formed the basis for this study. Nine colleges were selected using random sampling technique. 1398 students were selected using randomly. A maximum of 20 samples were taken in each course.

The questionnaire was prepared based on the academic performance and all the related variables of the students. The associated functions of socio economic and demographic characteristics was subjected to pilot study and modified. The reliable co-efficient for the questionnaire chronboch alpha was 0.73, which identified a good reliability.

Data clustering method is mostly used to operate on a large data value; it is used to discover the hidden pattern to make decision quickly and efficiently. K number of objects is randomly selected by K-means algorithm which represents a cluster mean. Based on the distance between the data points and cluster mean a data points is assigned to the cluster. K-Means Clustering – Algorithm involves the following steps:

1. Place K points into the space represented by the data points that are being clustered.
2. Assign each data points to the group that has the closest centroid.
3. When all data points have been assigned, recalculate the positions of the K-centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move.

Clusters are the subsets of large set of data which are segmented by cluster analysis. Initially the students are all in a same group. But when K-means clustering is applied on it then it clusters the student's into five major categories.

Clusters are fully dependent on the selection of the initial cluster centroids in K-mean clustering algorithm. The distances of all data elements are calculated by Euclidean distance formula. K-Data elements are selected as initial centers.

Data elements having less distance to centroids are moved to the appropriate cluster. A number of factors that are considered to have influence on the performance of a student were identified. The primary data is collected from a self – financing arts and science colleges. These influencing factors were categorized as input variables.

Weka (Waikato Environment for Knowledge Analysis) is a popular machine learning software written in

Java, it was developed at the University of Waikato, New Zealand. It has four components: Simple CLI, Experimenter, Explorer and Knowledge Flow. To store data in a database ARFF format is used by weka.

The cluster for each variable was considered to be best which were equal to or more than the overall centroid. For clustering 1398 students according to their academic performance, the first step is to load the data set and then choose the number of clusters. After choosing 5 clusters and applying K-Means Algorithm on the given data set of 1398 students in WEKA Tool, an output of 5 clustered instances is obtained.

The best variables in each of the cluster were identified by choosing the greater mean centroid point for each cluster which is above than the overall mean centroid point for each attribute. The variables used in the analysis are presented in the following four domains relating to the students' performance.

Table 1. Socio Economic Status Attributes

S.No	Variable Name	Description	Domain
1	Gender	Gender	{Male, Female}
2	Classification	Classification	{Arts, Science}
3	Subject	Subject	{Tamil, English, History, Economics, B.Com, BBA, BCA, Maths, Physics, Chemistry, Botany, Zoology, Comp.Sci.}
4	Colloc	College Location	{Rural, Small Town, Urban}
5	Resloc	Residence Location	{Rural, Small Town, Urban}
6	Fatedu	Father Education	{Primary, Secondary, Higher}
7	Motedu	Mother Education	{Primary, Secondary, Higher}
8	Foccu	Father Occupation	{Agriculture, Business, Service, Teacher, Others}
9	Moccu	Mother Occupation	{Home Maker, Business, Service, Teacher}
10	Ecostat	Economical Status	{<Rs. 50,000PA, Rs. 50,000 to 5,00,000, >5,00,000
11	Pargra	Parents Graduate	{No, Yes both parents, Yes mother only, Yes father only}

Table 2. Skill Development Attributes

S.No	Variable Name	Description	Domain
1	Commski	Communication Skill	{ Excellent, Very Good, Good, Fair}
2	Jobaff	Job Affect College Work	{ Enhances, Not interfere, Takes some time, Take lot of time}
3	Likecol	Liking college	{Enthusiastic, I like it, Neutral}
4	Knowskill	Knowledge and skill	{ For specific job, Very much, Quite a bit, some}
5	Undyou	Understanding Yourself	{ Very much, Quite a bit, some}
6	Premat	Presentation of Material	{Excellent, Very Good, Good, Fair}
7	Medium	Medium of Instruction	{Tamil, English}
8	Timespend	Time spent for reading	{2 hours, 3 hours, 4 hours, 5 hours}
9	Attendance	Attendance percentage	{>60, 61-70, 71-80, >90, <60}

Table 3. Motivation Attributes

S.No	Variable Name	Description	Domain
1	Subassi	Submit Assignment	{Yes, No}
2	Knowfac	Knowledge of Faculty	{Excellent, Very Good, Good, Fair}
3	Teaqua	Teaching Quality	{Excellent, Very Good, Good, Fair}
4	Faccon	Faculty Concern	{Excellent, Very Good, Good, Fair}
5	Leacen	Learning Centre	{Regularly, Sometimes, Ever}
6	Preinst	Presentation by Instructor	{They are clear and Informative, They are clear}
7	Lanpro	Language Proficiency	{ Improved dramatically, Improved somewhat, Not improved, Did n't take the course}
8	Placement	Placements	{Yes, No}
9	Schlorship	Scholarship	{Yes, No}
10	Indvisit	Industrial Visit	{Yes, No}

Table 4. Infrastructural Facilities Attributes

S.No	Variable Name	Description	Domain
1	Travel	Travel By	{College Bus, Private Bus, Own Vehicle}
2	Colinf	College Infrastructure	{Excellent, Very Good, Good, Fair}
3	Intspo	Interested in Sports	{Yes, No}
4	Exacti	Extracurricular Activities	{Yes, No}
5	Libfaci	Library Facility	{Excellent, Very Good, Good, Fair}
6	Stumat	Study material	{By Faculty, Text Book, Reference Book}
7	Acccomp	Access to Computer	{Yes, No}
8	Lab	Laboratory Notes	{Excellent, Very Good, Good, Fair}
9	Campus	College Campus	{Excellent, Very Good, Good, Fair}
10	Intfac	Internet Facility	{Yes, No}
11	Security	Security System	{Excellent, Very Good, Good, Fair}
12	Driwater	Drinking Water	{Excellent, Very Good, Good, Fair}
13	Canteen	Canteen Facility	{Excellent, Very Good, Good, Fair}
14	Bus	Bus Facility	{Excellent, Very Good, Good, Fair}
15	Medical	Medical Facility	{Excellent, Very Good, Good, Fair}

IV. Results

Table 5. Number of students classified into five Clusters

Cluster No.	No. of Students	Percentage (%)
Cluster 1	368	26
Cluster 2	192	14
Cluster 3	287	21
Cluster 4	201	14
Cluster 5	350	25
Cluster Total	1398	100

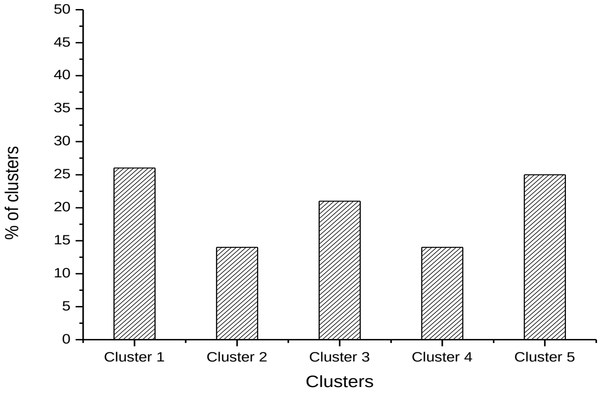


Figure 1. Total Clusters

In Table 5, it is observed that more number of students are classified in cluster 1 (26%).

Table 6. Better Centroid Points (B) among Socio Economic Status characters

S.No	Name of the Attriubute	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Gender	B	B		B	
2	Classification		B	B	B	
3	Subject		B	B	B	
4	Colloc	B	B	B	B	B
5	Resloc		B	B	B	B
6	Fatedu	B			B	
7	Motedu		B	B	B	B
8	Foccu			B	B	
9	Moccu		B		B	B
10	Ecostat		B			B
11	Pargra	B	B		B	
12	Total	4(11.80%)	9(26.50%)	6(17.60%)	10(29.40%)	5(14.70%)

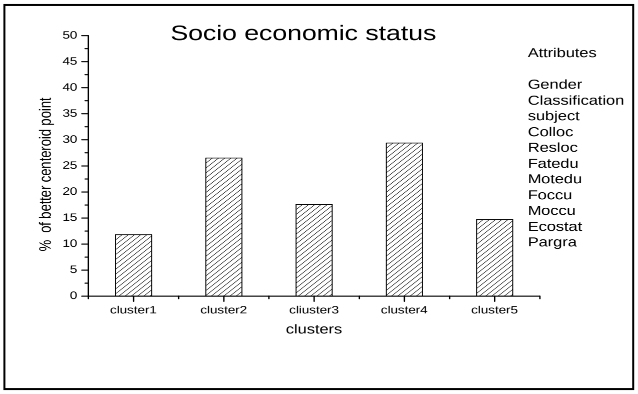


Figure2. Better centroid points among Socio Economic Status

In the Table 6, the socio economic status better centroid points mostly were found in the cluster 4 in almost all the attributes except economical status. Among all the clusters the cluster 4 has maximum percent of better centroid points (29.4%).

Table 7. Better centroid points (B) among Skill Development characters

S.No	Name of the Attriubute	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Commski				B	B
2	Jobaff		B		B	
3	Likecol				B	B
4	Knowskill	B				B
5	Undyou	B	B		B	B
6	Premat				B	B
7	Medium	B	B		B	B
8	Timespend			B		B
9	Attendance		B		B	
10	Total	3(13.6%)	4(18.2%)	1(4.5%)	7(31.8%)	7(31.8%)

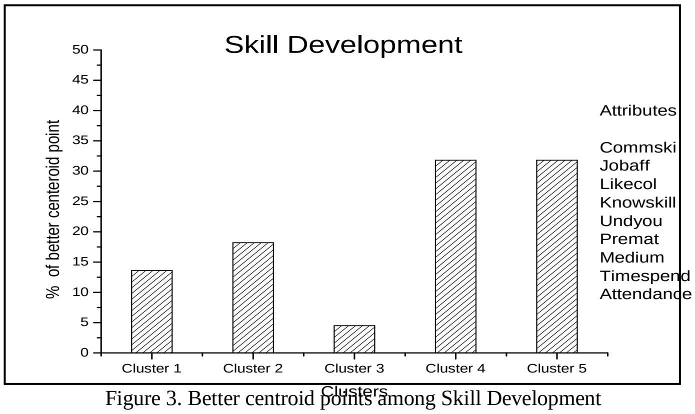


Figure 3. Better centroid points among Skill Development

In the Table 7, the skill development attributes better centroid points mostly were found in the cluster 4 and cluster 5.They are equally best than the other clusters. In the cluster 4 all the attributes are the best except knowledge and skill and time spent for reading. In the cluster 5 all the attributes are the best except job affecting college work and attendance percentage. Among all the clusters, cluster 4 and cluster 5 has maximum percent of better centroid points (31.8%).

Table 8. Better centroid points (B) Motivation Attributes characters

S.No	Name of the Attriubute	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Subassi		B		B	B
2	Knowfac		B		B	B
3	Teaqua				B	B
4	Faccon				B	B
5	Leacen	B				B
6	Preinst				B	B
7	Lanpro	B			B	B
8	Placement			B		B
9	Schlorship			B		B
10	Indvisit					B
11	Total	2(9.1%)	2(9.1%)	2(9.1%)	6(27.3%)	10(45.5%)

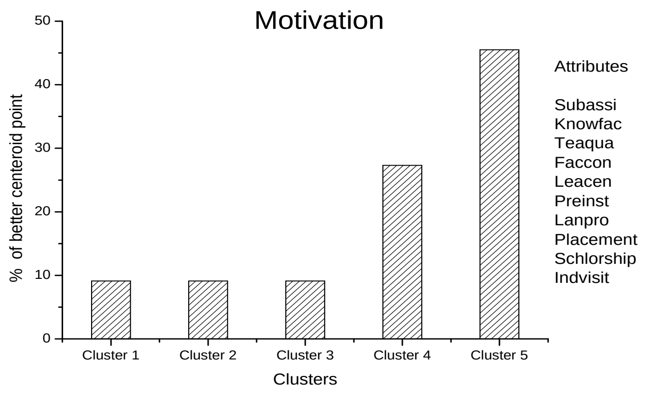


Figure 4.Better centroid points among Motivation

In the Table 8, the motivation attributes better centroid points mostly were found in the cluster 5 in all the attributes. Among all the clusters, the cluster 5 has maximum percentage of better centroid points (45.5%).

Table 9. Best centroid points (B) Infrastructural Facilities characters

S.No	Name of the Attriubute	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
1	Travel	B	B	B	B	
2	Colinf				B	B
3	Intspo		B		B	B
4	Exacti		B		B	B
5	Libfaci				B	B
6	Stumat		B	B	B	
7	Acccomp		B	B		B
8	Lab		B		B	B
9	Campus				B	B
10	Intfac					B
11	Security				B	B
12	Driwater				B	B
13	Canteen				B	B
14	Bus				B	B
15	Medical				B	B
	Total	1(2.80%)	6(16.70%)	3(8.30%)	13(36.10%)	13(36.10%)

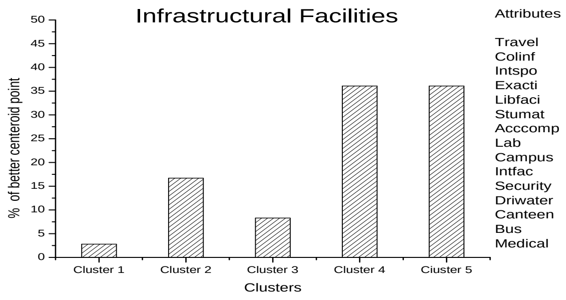
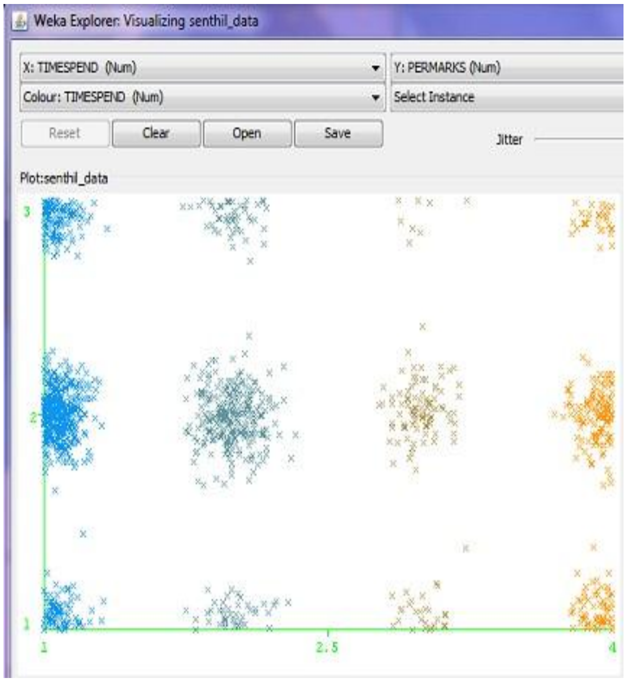


Figure 5. Better centroid points among Infrastructural Facilities



Weka Explorer: Visualizing senthil_data

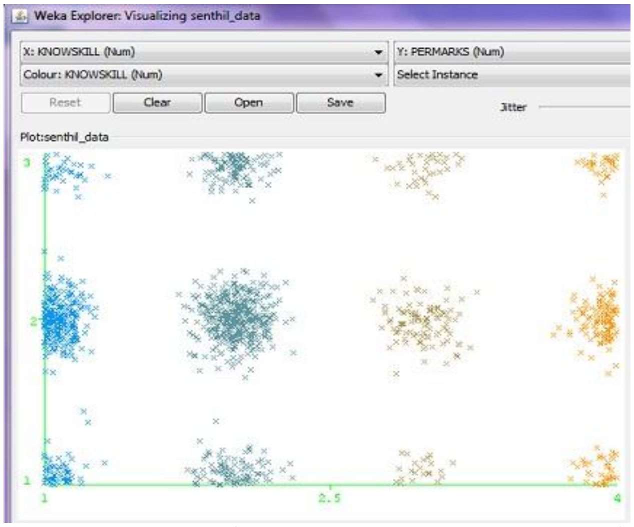
X: TIMESPEND (Num)

Y: PERMARKS (Num)

Colour: TIMESPEND (Num)

Figure 6. Visualization of cluster – Time spent for reading

In Figure 6, Visualization of cluster shows more number of students spent 3hours for reading and they got percentage of marks between 61to 80 category.



Weka Explorer: Visualizing senthil_data

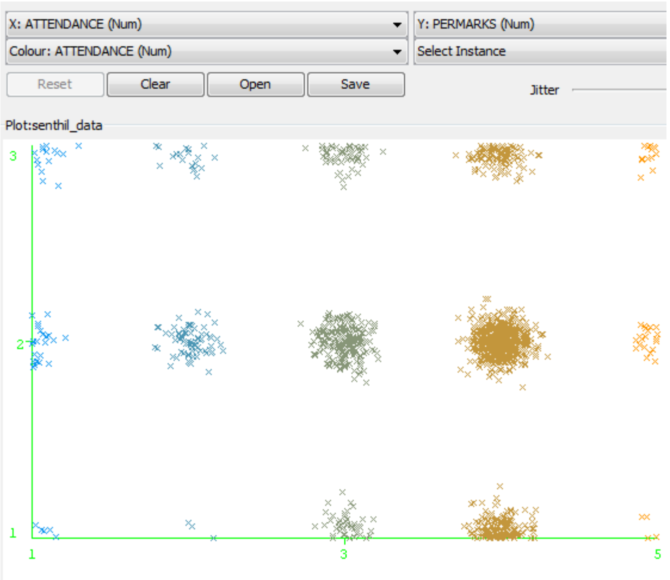
X: KNOWSKILL (Num)

Y: PERMARKS (Num)

Colour: KNOWSKILL (Num)

Figure 7. Visualization of cluster – Acquiring Knowledge and Skill of students

In Figure 7, more number of students acquire very high knowledge and skill but they belong to 61to 80 marks percentage category.



X: ATTENDANCE (Num)

Y: PERMARKS (Num)

Colour: ATTENDANCE (Num)

Select Instance

Plot:senthil_data

Figure 8. Visualization of cluster – Attendance percentage of students

In Figure 8, more number of students' attendance percentage is above 90 but they belong to 61to 80 marks category.

V. Discussion

From the total number of clusters, cluster 1 and cluster 5 has more number of students. Cluster 4 which contains all the attributes was found to be best for socio economic status except for economic status attribute. Cluster 4 and Cluster 5 which contains all the attributes was found to be best for skill development. In cluster 4, except for the time spent for reading attribute and in cluster 5, except for job affecting the college work and attendance attributes Cluster 5 contains all the attributes found to be best in motivation. Cluster 4 and Cluster 5 which contains all the attributes was found to be best in infrastructural facilities. In cluster 4, except for access to computer and internet facility attributes and in cluster 5 travel and student material attributes.

VI. Conclusion

In this research, K-Means clustering algorithm was used by applying weka interface to identify academic performance and to enhance the educational quality of self-financing arts and science college students. To predict the performance in semester examinations, some of the influencing factors were identified.

From this study it is observed that more number of students are classified in cluster 1 (26%). In the Socio economic status better centroid points mostly were found in the cluster 4. In the skill development attributes better centroid points were mostly found in the cluster 4 and cluster 5. In the motivation attributes better centroid points mostly were found in the cluster 4 and cluster 5. In the infrastructural facilities better centroid points mostly were found in the cluster 4. Number of students spent 3 hours of reading, they acquire very high knowledge and skill, there attendance percentage is above 90% but they belong to 61-80 mark category. The information obtained may be used by teachers as well as students. From various data mining techniques clustering is the efficient method for predicting student's performance.

VII. Acknowledgement

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

VIII. References

[1] Mahesh Singh, Anita Rani, Ritu Sharma, *"An optimised approach for student's academic Performance by k-means clustering algorithm using Weka interface",* International Journal of Advanced Computational Engineering and Networking, Vol.2, Issue.7, pp.5-9, 2014.

[2] Md. Hedayetul Islam Shovon, MahfuzaHaque, *"An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree",*International Journal of Advanced Computer Science and Applications,Vol.3, Issue.8, pp.146-149, 2012.

[3] SreedeviKadiyala, Chandra Srinivas Potluri, *"Analyzing the Student's Academic Performance by using Clustering Methods in Data Mining",*International Journal of Scientific & Engineering Research, Vol.5, Issue.6,pp.198-202, 2014.

[4] SnehalBhogan, KedarSawant, PurvaNaik, Rubana Shaikh, OdeliaDiukar, SayleeDessai, *"Predicting student performance based on clustering and classification",*IOSR Journal of Computer Engineering (IOSR-JCE),Vol.19, Issue.3, pp.49-52, 2017.

[5] M. Durairaj, C. Vijitha, *"Educational Data mining for Prediction of Student Performance Using Clustering Algorithms",*(IJCSIT) International Journal of Computer Science and Information Technologies, Vol.5, Issue.4, pp.5987-5991, 2014.

[6] ShashikantPradipBorgavakar, Amit Shrivasrava, *"Evaluating student's performance using k-means clustering",*International Journal of Engineering Research & Technology(IJERT), Vol.6, Issue.5, pp.114-116, 2017.

[7] T. Prabha, D. ShanmugaPriyaa, *"Knowledge discovery of the students academic performance in higher education using intuitionistic fuzzy based clustering",*Journal of Theoretical and Applied Information Technology, Vol.95, Issue.24, pp.7005-7019, 2017.

[8] B. HemaMalini, L Suresh, *"Data Mining in Higher Education System and the Quality ofFaculty Affecting Students Academic Performance: A Systematic Review",* International Journal of Innovations & Advancement in Computer Science (IJIACS), Vol.7, Issue.3, pp.66-70, 2018.

[9] M. Muthalagu, *"Applied on Clustering Algorithm in Edm",* International Journal of Computer Engineering and Applications, Vol.12, Issue.1, pp.71-76, 2018.

[10] O. J. Oyelade, O. O. Oladipupo, I. C. Obagbuwa, *“Application of k-Means Clustering algorithm for prediction of Students' Academic Performance”,* International Journal of Computer Science and Information Security(IJCSIS), Vol.7, Issue.1, pp.292-295, 2010.

[11] Kritika Yadav , Mahesh Parmar, “ *Review Paper on Data Mining and its Techniques and Mahatma Gandhi National Rural Employment Guarantee Act*”, International Journal of Computer Sciences and Engineering , Vol.5,Issue.4, pp. 68-73, 2017.

[12] Nidhi Sethi, Pradeep Sharma, “International Journal of Scientific Research in Computer Science and Engineering, Vol.1, Issue.3, pp.31-34, 2013

Authors Profile

Mr. R. Senthil Kumar pursed Bachelor of Science from NGM College, Bharathiar University, Coimbatore in 1991 and Master of Computer Applications from Madurai Kamaraj University, Madurai in 2002.He pursed Master of Philosophy in Computer Science from Periyar University, Salem in year 2007. He is currently pursuing Ph.D from Periyar University, Salem and currently working as Assistant Professor and Head in Department of Computer Science, RTG Arts and Science College, Polur, Tamilnadu, India since 2010. He is a member of Computer Society of India. He has published one paper in reputed international journal. His main research work focuses on Educational Data Mining. He has 9 years of teaching experience and 4 years of Research Experience.

Dr. K. Arulanandam pursed Bachelor of Science from University of Madras in 1997 and Master of Computer Applications from University of Madras, Chennai in 2001. He pursed Master of Philosophy in Computer Science from Manonmaniam Sundaranar University in 2003. He pursed Doctor of Philosophy in Computer Science from Vinayaka Missions University, Salem and currently working as Assistant Professor and Head in Department of Computer Science, GTM College, Gudiyattam, Tamilnadu, India since 2011. He is a life member of CSI, CSTA, IAENG, ISTE, IACSIT, IJST, ACEEE and ISCA. He has published 36 research papers in reputed national and international journals. His main research work focuses on Computer Networks. He has 17 years of teaching experience and 14 years of Research Experience.