ANALYZING THE SOCIAL FACTORS THAT AFFECTS THE WOMEN EDUCATION USING ADVANCED HIERARCHICAL CLUSTERING ALGORITHM

[1]Dr.A.Sumathi, [2]Dr. P. Sharmila

[1]Associate Professor, [2]Principal

[1] Department of Computer Science,

[1]Navarasam Arts & Science College for Women,

Arachalur, Erode (Dt) – 638 101, Tamil Nadu, India.

Abstract: Implementation of data mining techniques in the field of education system is getting more importance in the recent days, refers to the Educational Data Mining (EDM). Year by year, the number of students' admission to various schools, colleges and universities is getting increased. As the enrolment increases, the size of the dataset gets increases consisting large amount of data. This leads to the decrement in the quality assessment and other education related issues which becomes a critical issue and major concern which needs to be sorted out. The major purpose of retaining the standard of the education is that it contributes a major part in the development of the country. It is mandatory to assess and accredit the facilities provided by all the education centres as the retention and the development of skills becomes very important. In the present scenario, the Government of India implemented many steps in order to improve the women education in all parts of the state. Despite the efforts taken by the concerned bodies, the level of women participation in the education is very low when compared to other states. This paper rely on addressing the major issues that prevails in a society that affects the women education. An attempt to examine of what can be done to balance this inevitable situation is carried out in this paper. The analysis has been carried out using Advanced Hierarchical Clustering Algorithm (AHCA) with R-tool and the performance analysis was made in the mentioned algorithm. This paper concludes with the possible solution that can be provided in order to overcome this problem.

Keywords: Social Factors, Educational Data Mining (EDM), Advanced Hierarchical Clustering Algorithm (AHCA) and R Tool.

I. INTRODUCTION

The Education is a process by which the people learn the unknown skills, gaining knowledge and also understand about them along with the world. The process followed in the education is systematic for the young people and also the importance of the education cannot be emphasized as it brings the knowledge and the development not only for the physical structure but also it shapes the internal character. The outcome being the emotional and the well-being of the people of a surviving society. Wherever we go and wherever we are, there is an increase in the knowledge about ourselves and also the environment.

There are many meaning given for Education by different authors. "Bringing up or training a person mentally, so that the person becomes capable of thinking for himself or herself, the family and the community" – quoted in 'The Chambers Dictionary'. This particularly explains the process of equipping a person, irrespective of sex with a vital societal key with which to open or lock many doors of life. On providing the proper education, the women will not be able to read and write but also will contribute to the development of the nation in many ways.

There are many reasons for hindering the women education in India. In the year 1986, Oshin found that 50% of the population denied to be educated due to the federal and the state laws designed 70 that promote the equality in our society. An early marriage, as a common problem (1986), plays a major role in hindering the education to women children.

As a part of the research work, the problem pertaining to the social factors affecting the women education in the Erode district of Tamilnadu is taken for analysis in this chapter. The analysis has been carried out with the help of various clustering algorithms using R-tool. The performances of both the algorithms have been compared with their earlier versions.

II. SOCIAL FACTORS AFFECTING WOMENS EDUCATION

Various factors have been related to women education that prevails among the students. Among those factors Social Factor (SF), Parent Mentality (PM), Educational Facilities (EF), Family Background (FBG) and Physically Challenged (PC) were identified to be factors that affect the women education. The analysis of the aforesaid factors have been carried out with the Advanced Hierarchical Clustering (AHC) using R-tool.

III. HIERARCHICAL CLUSTERING

Hierarchical clustering involves creating clusters that have a predetermined ordering from top to bottom. For example, all files and folders on the hard disk are organized in a hierarchy. Figure 1 represents the simple formation of the hierarchical clustering.

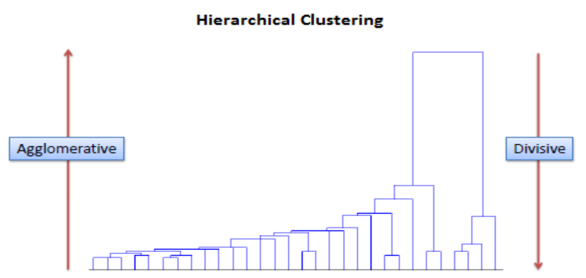


Fig. 1 Representation of Hierarchical Clustering

There are two types of hierarchical clustering listed as Divisive method and Agglomerative method

3.1 Divisive method

In divisive or top-down clustering method we assign all of the observations to a single cluster and then partition the cluster to two least similar clusters. Finally, we proceed recursively on each cluster until there is one cluster for each observation. There is evidence that divisive algorithms produce more accurate hierarchies than agglomerative algorithms in some circumstances but is conceptually more complex.

3.2 Agglomerative method

In agglomerative or bottom-up clustering method we assign each observation to its own cluster. Then, compute the similarity (e.g., distance) between each of the clusters and join the two most similar clusters. Finally, repeat steps 2 and 3 until there is only a single cluster left. The related algorithm is shown below.

Given:

A set of Objects $\{x_1,\ldots..,x_{n\}}$

A distance function dist(c_1,c_2)

for =1 to n

$c_i=\{x_i\}$

end for

$C=\{c_1,\ldots.,c_n\}$

l=n+1

whileC.size> 1 **do**

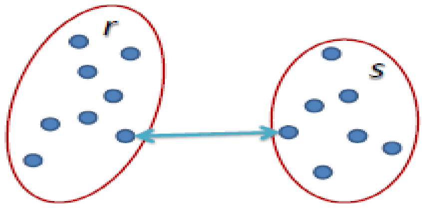
- (c_{min1}, c_{min2})=minimum dist(c_i.c_j) for all c_i,c_j in C
- remove $c_{min1}$and c_{min2} form C
- add $\{c_{min1}$, $c_{min2\}}$ to C
- l=l+1

end while

Before any clustering is performed, it is required to determine the proximity matrix containing the distance between each point using a distance function. Then, the matrix is updated to display the distance between each cluster. The following three methods differ in how the distance between each cluster is measured.

3.3 Single Linkage

In single linkage hierarchical clustering, the distance between two clusters is defined as the shortest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two closest points. Figure 3.7 depicts the Single Linkage Mechanism.

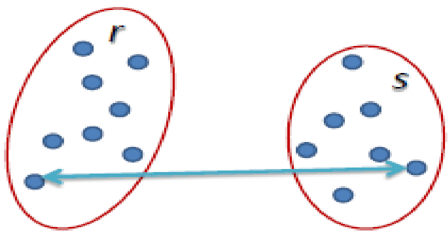


$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

Fig. 2 Representation of Single Linkage Mechanism

3.4Complete Linkage

In complete linkage hierarchical clustering, the distance between two clusters is defined as the longest distance between two points in each cluster. For example, the distance between clusters “r” and “s” to the left is equal to the length of the arrow between their two furthest points. Figure 3.8 depicts the Complete Linkage Mechanism.

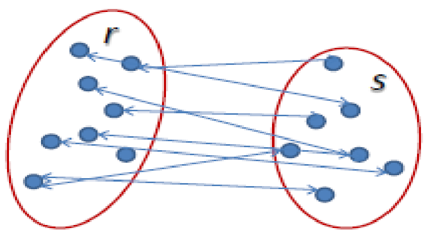


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

Fig. 3 Representation of Complete Linkage Mechanism

3.5 Average Linkage

In average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster. For example, the distance between clusters “r” and “s” to the left is equal to the average length each arrow between connecting the points of one cluster to the other. Figure 4 depicts the Average Linkage Mechanism.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

Fig. 4 Representation of Average Linkage Mechanism

IV. ADVANCED HIERARCHICAL CLUSTERING ALGORITHM (AHCA)

The major objective of defining the Advanced Hierarchical Clustering Algorithm (AHCA) is to cluster the dataset using the item-based clustering approach based on the traditional algorithm. Many of the clustering methods construct the data representation matrix based on the density of the dataset. With this matrix, the right similarity measure between the data is chosen and the similarity is calculated. In the clustering process, the similarity measure plays a major role, as it affects the efficiency of the clustering. In this proposed clustering algorithm, by neglecting the data count process, the similar data are merged together efficiently than the existing method with less computational time.

4.1 ANALYSIS

Initially, the data is loaded in to the R-tool for the purpose of pre-processing. Once it is pre-processed, the clustering of the data in to similar groups takes place. Table 1 presents the clustering of vectors in to similar groups. Figure 5 depicts the dendrogram structure of the clusters classified.

Table 1: Clustering of vectors into similar groups

2[1]	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[42]	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[83]	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[124]	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[165]	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3
[206]	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[247]	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
[288]	3 3 3 3 3 3 3 3 3 3 3 3 3

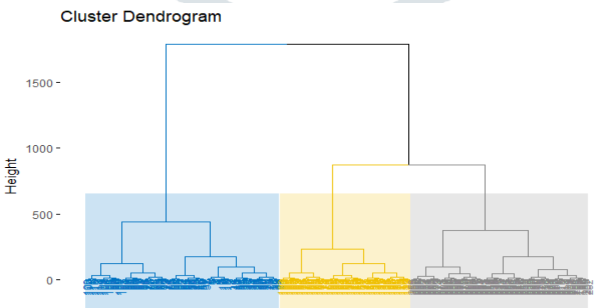


Fig.5 Cluster Dendrogram Representation

In the Figure 5, it is observed that the cluster formation is taken in the x-axis and the height of the cluster formation is along the y-axis. From the Table 2, it is seen that the total number of clusters is divided into 3 groups of each cluster of size 100. The mean value of all the clusters is given in table 3.9.

Table 2:Representation of Cluster Mean Value for various factors

SF	**PM**	**EF**	**FBG**
Mean :1.88	Mean : 1.6	Mean :2	Mean :1.70
Sum of Squares by Cluster			
Cluster 1		83548.84	
Cluster 2		83441.56	
Cluster 2		83441.56	
Average between the sum of squares to the total sum of squares		**88.9%**	

The time taken for the computational process is found to be 1.04 sec, which is less when compared to other algorithms taken for analysis. From the table it is observed that the sum of squares for the different clusters formed is found to be 83548.84, 83441.56 and 83441.56. The average between the sum of the squares to the total sum of squares is found to be 88.9 %.

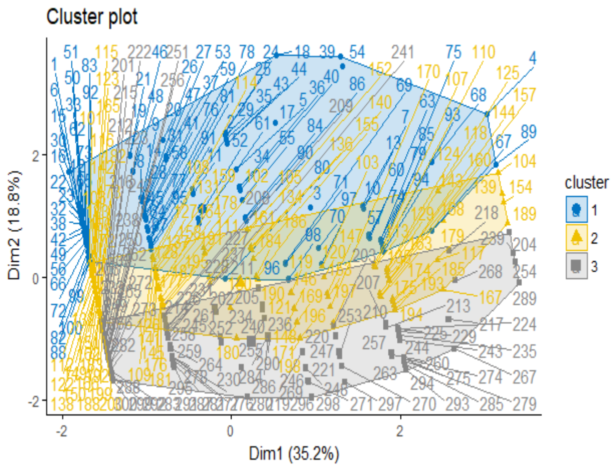


Fig. 6 Cluster formation for the pre-processed dataset

Figure 6 represents the cluster formation group of three with 100 samples each grouped under the category of similarity. Form the diagrammatical representation of cluster formation, it is observed that the similar data values are grouped within the same cluster as depicted in the table 3.8. This increases the accuracy of the clustering process and the quality of the clusters formed also gets increased. The comparison between the algorithms taken for analysis has been presented in the Table 3 in terms of computational time. From the table it is observed that the time taken by the AHCA is less when compared to KMCA and MKMCA.

Table 3: Comparison of Computational time between the algorithms

S. No.	**Algorithm**	**No. of records**	**Computational time (seconds)**
1	K-Means Algorithm (KMA)	300	7.36
2	Modified K-Means Clustering Algorithm (MKMCA)	300	1.38
3	Advanced Hierarchical Clustering Algorithm (AHCA)	300	1.04

Table 4: Percentile Analysis of various factors

Attributes	Code	Value	String	Percentage (%)
Social Factor	SF	Married	1	68
		Unmarried	2	32
Parent Mentality	PM	Supportive	2	60
		Discourage	1	40
Educational Facilities	EF	Good	3	26
		Better	2	48
		Worst	1	26
Family Background	FBG	High	3	6
		Medium	2	58
		Low	1	36
Physical Challenged	PC	No	1	97
		Yes	2	3

The percentile analysis of the various factors that affects the womens education has been presented in the Table 4. Figure 7 depicts the graphical representation of various factors in terms of percentage. It is clearly observed that the social factor (SF) overtakes all the factors except physically challenged with a percentage of 68, which is higher than the other factors. This particular factor makes the women children not to pursue their study till the end of the degree course.

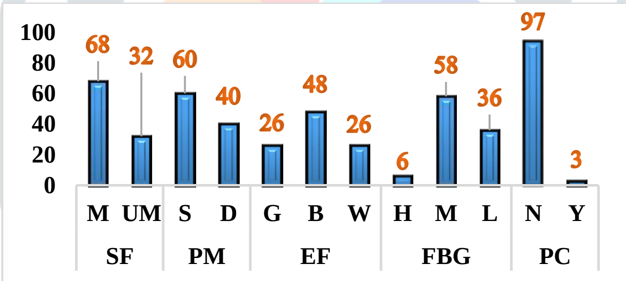


Fig.7 Percentile analysis of various factors

4.2 SOLUTIONS

As per the current scenario, there are nearly 70 million child brides worldwide. A girl's childhood abruptly ends with this marriage. The future prospects falls in to jeopardy after this incident. Mostly in the developing world, millions of girls are forced to wed as it is a part of regular practice that impedes progress on human rights, education, global health and economic development. The evaluation has been carried out for various factors among 300 students from various colleges in erode district of Tamilnadu. From the evaluation, the following solutions were identified in order to delay or prevent the early marriage.

i) Empower girls with information, skills and support networks

The girls can become more knowledgeable and self-confident, by bringing the group of girls together to learn the skills that consists of numerical analysis. Also they should know the way to communicate with others and negotiate with them, how to be more energetic in the reproductive years, how to involve in a team work and also should know how to earn and manage money. The aforesaid solutions will enable the girls to take better decisions and they can find the alternatives to early marriage.

ii) Economic support and incentives to girls and their families

By committing early marriage, parents might get benefited through bride price and dowry. By increasing the economic security of the poor households might curb the early marriage. For the struggling families, financial support to be provided such as arranging loans, an opportunity to learn an income-generated skill that would yield economical relief. In future, the daughters who generate income with the help of the skill they learnt seem to adding more value to their family.

iii) Educate and rally parents and community members

The girls may be educated by arranging various meeting and conferences about how the health and future of the girls will be affected. With the knowledge acquired at a new stage, the attitude and the behavior about the early marriage can shift to a later stage and it might act as a greater challenge.

iv) Enhance girls' access to a high-quality education

Girls with no education are three times as likely to marry before 18 as those with secondary or higher education. Providing incentives such as uniforms, scholarships, necessary skills, support to girls to enroll and remain in college can delay early marriage. Programs aimed at improving the safety and Girl-friendliness of colleges, strengthening the college curricula and making lessons relevant to girls' lives also are very effective. Rather than schools, colleges allow girls to build the social networks as well as skills to better advocate for themselves.

v) Encourage supportive laws and policies

Many countries with high rates of early marriage have passes legislation to prohibit the patience or have established a legal minimum age for marriage. Advocating for the implementation of such laws, and raising awareness among government officials and community leaders and members, helps strengthen and/or better enforce existing initiatives around girls' rights. Where legislation is not on the books, advocating for legal and policy reform is a critical first step.

CONCLUSION

It is well known that there are many nodal centres setup across our country for improving the quality of education for both male and female. The Government of India provides information to all those nodal centres periodically in such a way many advanced methods are implemented from time-to-time. Based on the factors to be analysed, the dataset has to be created or it has to be collected from the particular community of people. After collecting the information, data pre-processing has to be carried out. The purpose of this step is to remove the unwanted or missed out data in a large dataset. If there are few missed information, the database will become unstable. After pre-processing, the original dataset will be framed and the analysis has to be carried out based on the new and reframed dataset.

REFERENCES

1. Alejandro Pena-Ayala. 2013. "Educational data mining: A survey and a data mining-based analysis of recent works", Expert Systems with Applications – Elsevier (In press).
2. AnnisaUswwatunKhasanah. 2018. "A Review on Student's Performance Prediction Using Educational Data Mining Techniques", Journal of Engineering and Applied Sciences, Vol. 13, No. 6, pp. 5302-5307.
3. KM Archana Patel and PrateekThakral. 2016. "The Best Clustering Algorithms in Data Mining". International Conference on Communication and Signal Processing, April 6-8, pp. 2042-2046.
4. Chintan Shan and Anjali Jivani. 2013. "Comparison of Data Mining Clustering Algorithms". International Conference on Engineering: Nirma University, pp. 1-4.
5. M. Durairaj and C. Vijitha. 2014. "Educational Data mining for Prediction of StudentPerformance Using Clustering Algorithms", International Journal of Computer Science and Information Technologies, Vol. 5, No. 4, pp. 5987-5991.
6. Mansi Gera and Shivani Goel. 2015. "Data Mining - Techniques, Methods and Algorithms: A Review on Tools and their Validity", International Journal of Computer Applications, Vol. 113, No. 18, pp. 22-29.
7. Marie Bienkowski, Mingyu Feng and Barbara Means. 2012. "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief". U.S. Department of Education. pp. 1-77.
8. Meenu Sharma and Kamal Borana. 2014. "Clustering In Data Mining: A Brief Review", International Journal of Core Engineering & Management, Vol. 1, No. 5, pp. 117-126.
9. Pooja Thakar, Anil Mehta and Manisha. 2015. "Performance Analysis and Prediction in Educational Data Mining: A Research Travelogue", International Journal of Computer Applications, Vol. 110, No. 15, pp. 60-68.
10. Rajni Jindal and Malaya Dutta Borah. 2013. "A Survey on Educational Data Mining and Research Trends". International Journal of Database Management Systems, Vol.5, No.3, pp. 53-73.

11. Rohaila Abdul Razak, Mazni Omar and Mazida Ahmad. 2018. "A Student Performance Prediction Model Using Mining Techique", International Journal of Engineering & Technology, Vol. 7, pp. 61-63.
12. Siti Khadijah Mohamad and Zaidatun Tasir. 2013. “Educational data mining: A review”. Presented in the 9th international conference on Cognitive Science, Procedia – Social and Behavioral Sciences, Vol. 97, pp. 320-324.
13. VandnaDahiya. 2018. "A Survey on Eduational Data Mining", International Journal of Research in Humanities, Arts and Literature, Vol. 6, No. 5, pp. 23-30.
14. Varun Kumar and Anupama Chadha. 2011. “An Empirical Study of the Applications ofData Mining Techniques in Higher Education”, International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, pp. 80-84.
15. P. Veeramuthu, R. Periyasamy and V. Sugasini. 2014. “Analysis of Student Result Using Clustering Techniques”, International Journal of Computer Science and Information Technologies, Vol. 5, No. 4, pp. 5092-5094.