Proceedings of the Fifth International Conference on Computing Methodologies and Communication (ICCMC 2021)
IEEE Xplore Part Number: CFP21K25-ART

Data Analysis on Student's Performance based on Health status using Genetic Algorithm and Clustering algorithms

Dr V.Preetha
Assistant Professor,
Department of Computer Science
SRI SRNM COLLEGE ,Sattur
Virudhunagar District Tamilnadu
preetha@srnmcollege.ac.in

2021 5th International Conference on Computing Methodologies and Communication (ICCMC) | 978-1-6654-0360-3/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICCMC51019.2021.9418235

Abstract— Data analysis is the emerging research field that relies on methods and techniques to make insights on the data sets. Data analysis on student's academic Performance based on their Health status such as nutritious food intake,hygienic life style and frequency of health issues is the main objective of the research. The datasets were obtained by Questionnaire method and the analysis were carried out initially with clustering algorithms such as K-means algorithm, Hierarchical clustering and EM Method.In the second phase, Genetic search was performed and the outputs were generated. The statistical output representation for the important attributes are given using orange software.The algorithmic Experimental setup was also carried out with weka datamining tool on student's dataset that has 113 instances and 93 attributes.The findings of the research work were that K-means algorithm outperformed well when compared with EM method and Hierarchical clustering. Genetic search method predicted correlated attributes for the selected class attribute and the outputs are generated.The statistical data analysis shows that nutrition and health issues of female students has an impact on the academic performance of students.

Keywords— Genetic Algorithm, Clustering Algorithm,Kmeans,Hierarchical, EM Method,Health status indicator

I. INTRODUCTION

In real time situations, Health status indicator of every individual especially for girls is essential for a prosperous life and career achievement**. "Learning without Burden**" is necessary to achieve the goals of the students. Girl student's performance is high in sports, Education and Extracurricular activities. But the Health issues of girl students become a hindrance for their successful achievement. Girl Students in college level can be categorized as different types such as normal healthy students, Students with lack of Hemoglobin content, Students with malnutrition, differently able girl students, students under medications, students with mild health issues and so on. The research aimed at finding the need of the students and self evaluation of the students' about nutrition and to further increase the academic performance of the students. For the purpose of the research, clustering algorithms such as K-means, Hierarchical clustering and EM applied. Further Genetic algorithm is applied for prediction and data analysis. Many Literature works were made based on Health indicators in existing works. In [1], the article explores the differences between the current health status of adolescents with the obtained target set in Healthy China 2030. In [2], a comparative analysis was made with the upper elementary school children with health problem and suggested the inclusion of physical health education. MANOVA and DISCRA analysis was performed for the analysis. In [3], the physical activity (PA), the intensity of metabolic processes and motor skills of university students were analyzed and the statistical analysis was performed. The study of the article in [4] deals with the nutritional status of toddlers classification with naïve bayes classification based on z-score value. Xing Xu et.al in [5] in their work, predicted the Internet usage behavior of 4000 students with machine learning algorithms. In [6], cluster analysis, principal component analysis or factor analysis and other measures are performed to analyze three health-related behaviors of interest among 11–16 year olds in high-income countries. In the cited work [7], Hybrid genetic algorithm (HGA), K-means is used for clustering. In [8], the authors proposed the authors proposed four traits of interest for hierarchical clustering algorithms as: (1) empirical performance, (2) theoretical guarantees, (3) cluster balance, and (4) scalability for clustering problems. Based on the analysis, it was concluded that student's performance analysis was reviewed in every aspect based on many attributes and also the recent clustering algorithms was widely used in many applications. In [9], the authors compared the clustering techniques in Educational data mining. K means, density based, and hierarchical methods are analysed for clustering analysis in terms of silhouette coefficient and distribution. The authors in [10], predicted the student's performance using modified version of Harris Hawks optimization algorithm. The authors in [11], developed a new algorithm based on genetic algorithm to identify the positive covid19 cases. In the work

[12] entitled, "Student performance analysis and prediction in classroom learning: A review of Educational data mining studies, a systematic review of EDM studies on student performance prediction is analysed. The main aim of the study in [13] is to develop and test a conceptual framework in a university context and investigation of academic success. In [14], adaptive E-learning system was surveyed, and soft computing tools is discussed within the ambit of Educational data mining. In the research work [15], finding out Students' course success in vocational courses was identified by unsupervised machine learning algorithms.

II. CLUSTERING ALGORITHMS

A. K-means algorithm

In K-means clustering algorithm, each object is assigned to precisely one set of clusters. The initial step starts by deciding the number of clusters with k value. The value of 'K' is generally a small integer, such as 2, 3, 4 or 5, but may be larger. The quality of set of clusters is determined by the value of an objective function which is calculated by the sum of the squares of the distances of each point from the centroid. Each cluster will have the centroid point and the objective function is given. Each point will be assigned to the nearest centroid in each cluster. For 'N' number of iterations, all the objects are assigned to 'K' clusters based on the centroids. The centroids are recalculated and the steps are repeated.

ALGORITHM 2.1

Choose a value of k
Select 'K' objects with 'K' centroids
Assign each object to the nearest centroids
Recalculate the centroids
Repeat steps 3 and 4 until the centroids no longer move

B. Hierarchical clustering

In Hierarchical clustering, starting with each object in a cluster of its own, the process repeatedly merge the closest pair of clusters until we end up with just one cluster containing all the objects.

ALGORITHM 2.2

Assign each object to its own single object cluster
Calculate the distance
Merge the closest pair
Calculate the distance between clusters
Recalculate the centroids
Repeat steps 3 and 4 until all the objects are in a cluster

C. Expected Maximization(EM) Method

EM assigns a probability distribution to each instance which indicates the probability of it belonging to each of the clusters. The cross validation performed to determine the number of clusters is done in the following steps:
1.The number of clusters is set to 1
2.The training set is split randomly into 10 folds.
3.EM is performed 10 times using the 10 folds
4.The loglikelihood is averaged over all 10 results.
5.If loglikelihood has increased the number of clusters is increased by 1 and the algorithm continues at step 2.

III. GENETIC ALGORITHM

Some of the attributes likely to be correlated in combination of other attributes as well as the analysis of the behavioral of the attributes can be performed using search space. The most fitness individual can be calculated using the objective function and the significant of the algorithm is that the fitness attribute is more likely to be survived. The selection, crossover, mutation operators are used to select the best fit individual.

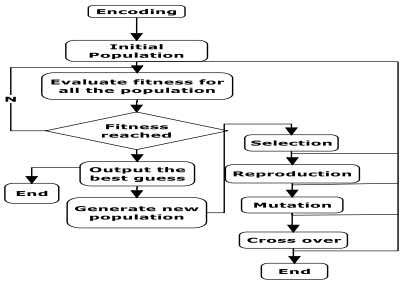

Fig.1. Genetic Algorithm method

4 METHODOLOGY

The methodology of the proposed work is given in the following Fig.2.The dataset is collected from college students of two different courses in google forms.The answers for the optional questions are given as blank statements and the mandatory answers for the questions are given as multiple choice questions.Data preprocessing was minimal as the mandatory questions do not have missing values.For the working of the algorithm, the numeric values and nominal conversions are performed.

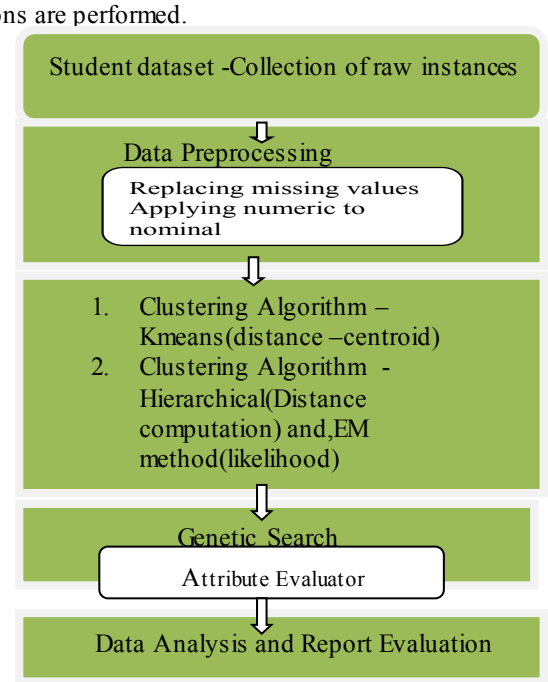

Fig. 2. Methodology of the proposed work

The data set was collected from from113 students of different courses. The question included psychological behavior of students in the aspects of health issues, hygiene, academic performance of the students from earlier studies, awareness of pandemic diseases and mental behavior in the class room learning.

Sample questionarrie:

1. Percentage of marks upto current semester
2. Differently abled student(speaking,listening, hearing,walking problems,partial blindness problems)
3. Sports person - participated in sports activities in school/college level
4. Won prizes in sports in school level
5. Won prizes in sports in college level/state level/national level
6. Participated in extra curricular activities
7. I will drink more water
8. I will eat more amount of nutritious food(iron rich)
9. I will exercise or atleast have a chance to walk
10. I will face minor health issues once in _____
11. Being a Female, my Health is a hindrance(problem) o to my sports/Extracurricular/Academic Achievements
12. My academic Performance increases/decreases gradually
13. I will be dull/active in my classroom all the time/rarely
14. I will be extremely clean/moderately clean in self hygiene and insist my family members to be clean
15. I am aware of Pandemic diseases and taking preventive measures

Experimental analysis was carried out using weka 3.6.13.The explorer module was used for experimental set up.K-means algorithm automatically handled combination of categorical and numerical attributes. The result window displayed the centroid of each cluster and the statistics and percentage of instances assigned to different clusters. Cluster centroids are represented as the mean vectors for each cluster. Hierarchical clustering was also performed by giving classes for cluster evaluation in the training set. Genetic algorithm is performed in search space with different set of attributes and subset evaluator.

IV. RESULT ANALYSIS AND DISCUSSION

The important attributes such as percentage of student's marks from XII standard and the questions answering yes/No based on nutritious factors, sports performance and other self evaluation questions are chosen for clustering. It was tested with K means and Hierarchical clustering. With K means algorithm, the number of iterations made for clustering was 5 and two clusters 0 and 1 were formed. The output of the K-means algorithm was given in the Fig.3.

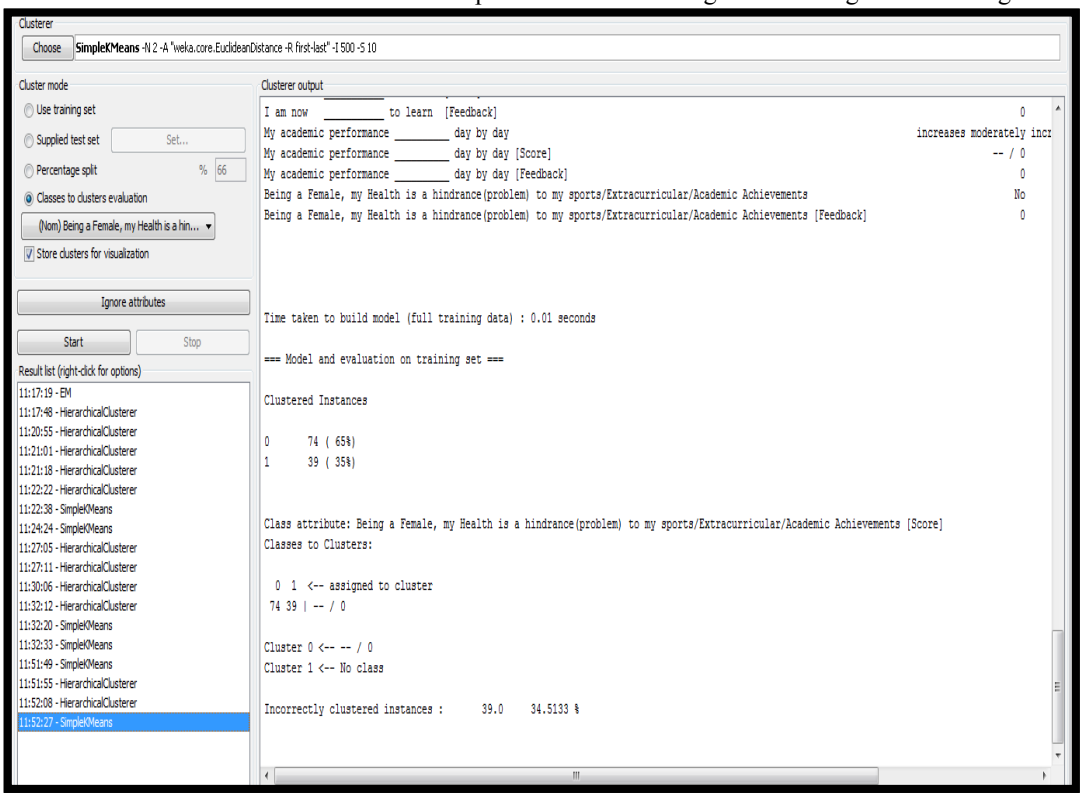

Fig 3. Output of K-means clustering

The total sum of squared errors was 652.85. The clustered instances under two categories were given in TABLE 1.It was observed that Hierarchical clustering grouped all the data inside cluster 0 which was misleading to categorize. The question, “Being a Female, my Health is a hindrance(problem) to my sports/Extracurricular/Academic Achievements “was analyzed by choosing it as the target attribute and the results are analyzed with K-means algorithm and Hierarchical clustering.

TABLE 1.Clustered Instances

S.No	Algorithm	Clustered Instances	
		Cluster 0	Cluster 1
1	K-means	74 (65%)	39 (35%)
2	Hierarchical	112 (99%)	1 (1%)

The incorrectly classified instance for cluster 0 is 39% and cluster 0 was 34%. Hierarchical clustering again categorized as two sets with 99% instances in cluster 0 and 1% classification in cluster 1.EM method was applied to the same target method and was found that 4 clusters were formed by evaluating mean,standard deviation and loglikelihood.The ouput in weka window for EM clustering method is given in the following Fig.4.

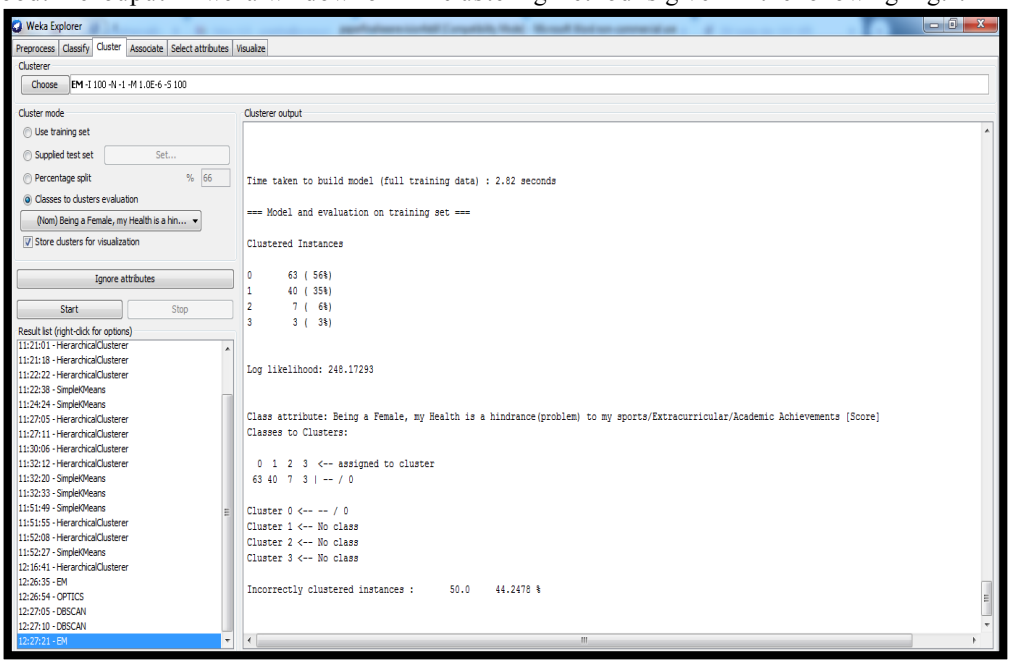

Fig 4. Output of EM clustering

It was observed that most of the datasets falls under only two category of clusters 0 and 1.The results concludes that K-means clustering performance was better when compared that EM method as the incorrectly clustered instance was high in EM method.Also, since Hierarchical clustering outputs only 2 clusters in agglomerative or dendogram hierarchy, the correlation between attribute analysis was difficult to manipulate.

TABLE 2. Incorrectly Clustered Instances

S.No	Algorithm	Incorrectly Clustered Instances	
		Cluster 0	Cluster 1
1	K-means	39%	35%
2	EM method	50%	44%

Further the data interpretation analysis was carried out with Orange Software with the student's dataset. The question under self-evaluation criteria, "*My academic performance increases gradually/ moderately/decreases*" was analyzed and was found that academic performance increases moderately was chosen by most of the students which was given in the following fig.5.

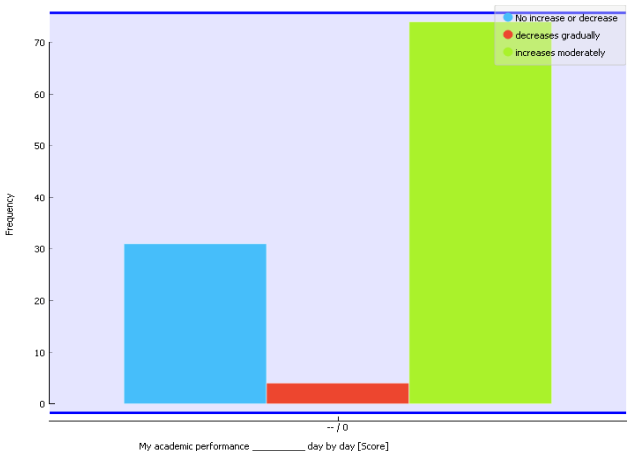

Fig 5. Statistical Representation chart –Result 1

The question *"I will get tired during class hours"* was answered by most of the students as "No" but nearly 45% students answered "yes".It was illustrated in Fig.6.This results in the correlation between the attribute "I will eat more amount of nutritious food" and the statistical measure was tested.

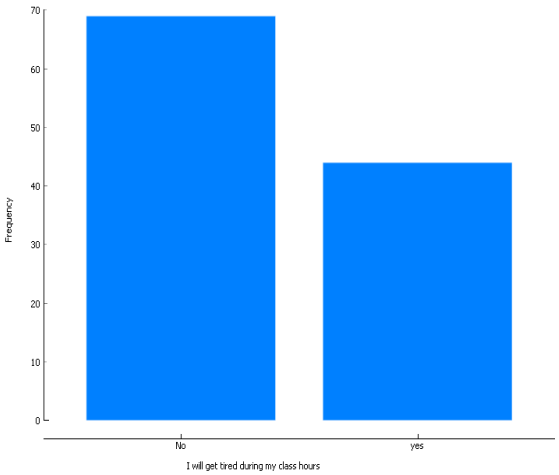

Fig 6. Statistical Representation chart –Result 2

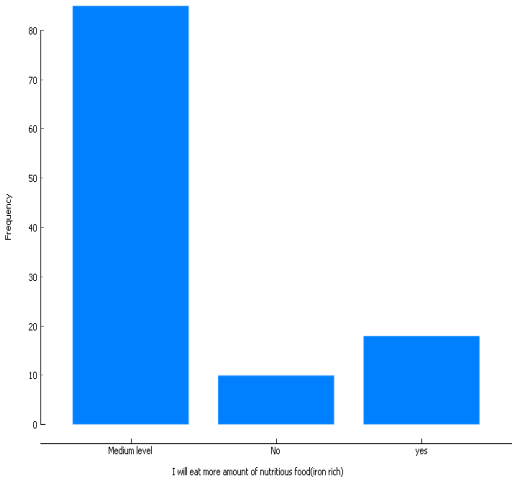

Fig 7. Statistical Representation chart –Result 3

The question "I will eat more amount of Nutritious food(iron rich) was answered by most of the students as "medium level" and 10% students answered as "No".It was illustrated in Fig.7

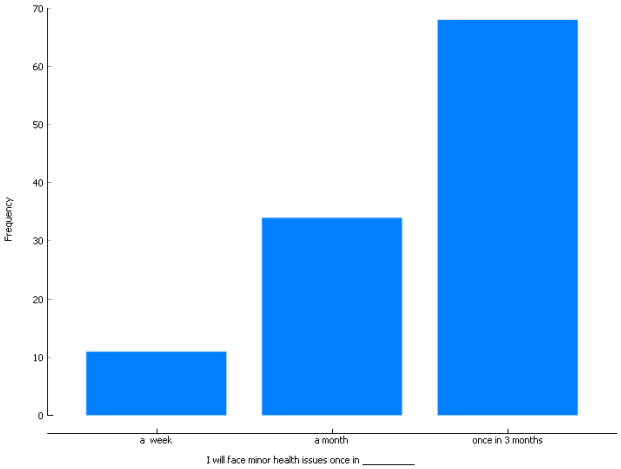

Fig 8. Statistical Representation chart –Result 4

Fig 8 shows the results of the question "I will face minor health issues once in a week/a month/3 months.35% students have chosen once in a month and 10% students have chosen once in a week. *"Being a female my health is a hindrance to achievements*" question was answered by most of the students as "No". Fig.9. depicts the statistics.

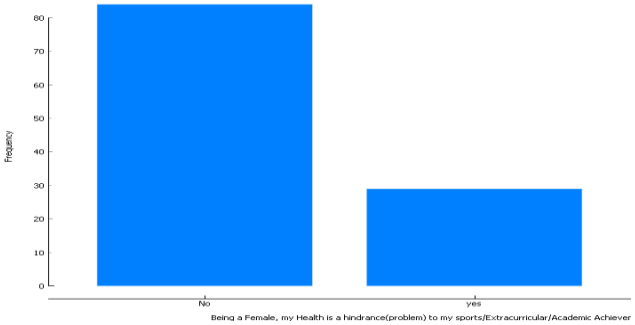

Fig 9. Statistical Representation chart –Result 5

Genetic Algorithm was performed for the correlation of attributes and best search. “Being female my health is a hindrance to my sports/extracurricular/academic activities” are correlated and found 16 attributes are correlated with the selected question as locally predictive attributes. Hence, the data analysis concludes that health issues, nutritions factors were correlated with the academic performance and percentage of marks from xii standard of a student.The experimental for genetic algorithm was crossover probability is 0.6,mutation probability is 0.33,population size 20 with seed value 1 which was given in the fig 10.

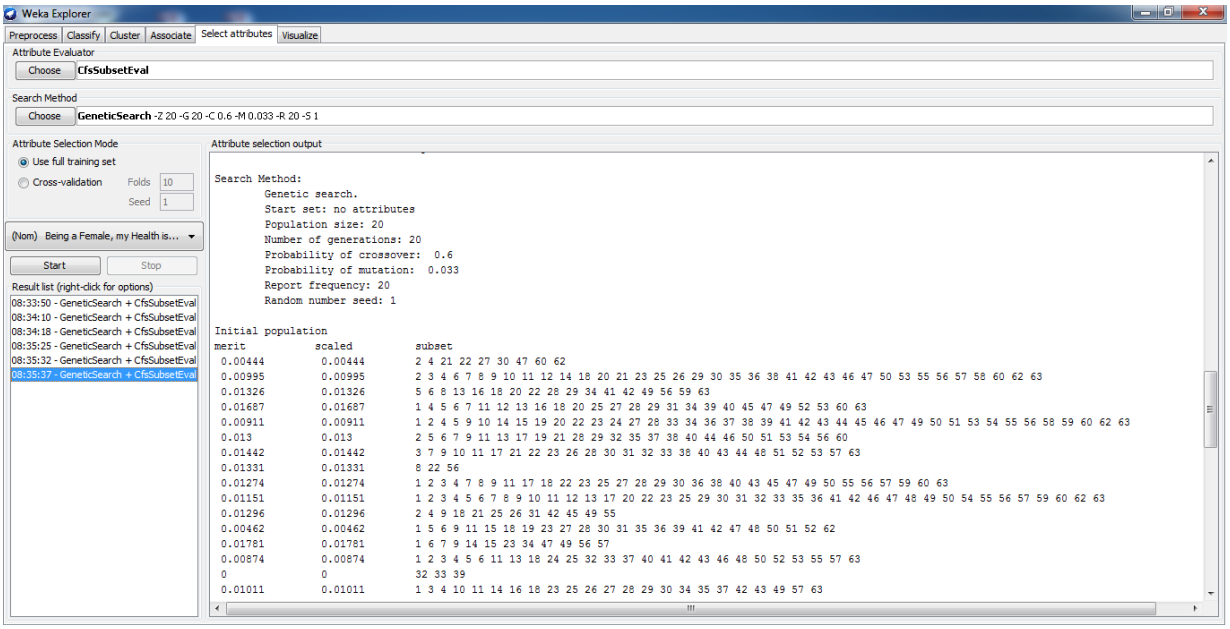

Fig 10. Genetic Algorithm output1

The 16 attributes were listed in the output window of weka which was given in fig.11. From the above data analysis reports, it was concluded that since, more than 40% students selected “yes” for the statement “I will get tired during class hours” and more than 80% students opted medium level for nutritious food, it was concluded that there is an impact on the academic performance. Further the results concludes that more than 30% students have opted “No increase/decrease” in academic performance and more than 25% students feel being female is a hindrance to their sports/Academic achievements.These factors has to be analysed in particular and it should be rectified by direct counseling with the students.

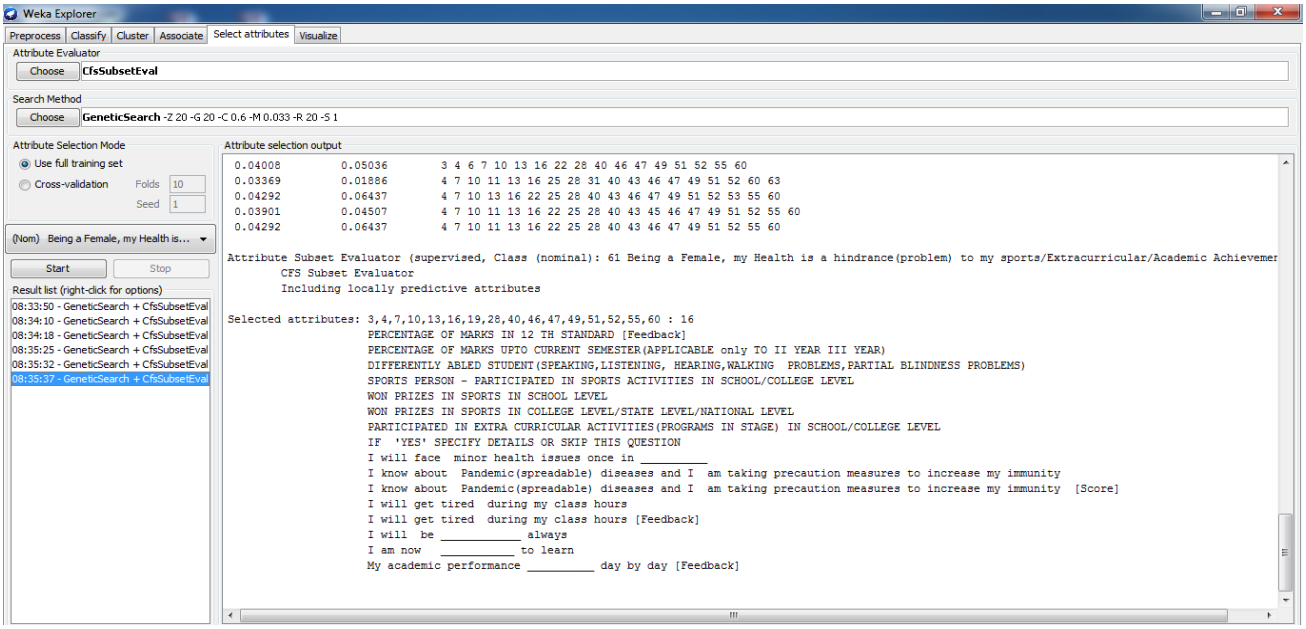

Fig 11. Genetic Algorithm Output2

V. CONCLUSION

Data analysis based on health status indicators and other related aspects were analysed based on clustering algorithms and genetic algorithm.The data analysis yield many facts of interestingness among students performance and psychological behaviour of the students in all aspects.Further the most important clustering algorithms were analysed based on incorrectly classified instances and found that K-means outperformed well in most of the cases.The future enhancement of the work will be performed using more number of instances and the output of the analytic report will be analyzed.

References

[1].Dong, B., Zou, Z., Song, Y., Hu, P., Luo, D., Wen, B., ... & Patton, G. C. (2020). Adolescent health and healthy China 2030: a review. *Journal of Adolescent Health*, *67*(5), S24-S31.

[2].Popović, R., Samouilidou, E., Popović, J., & Dolga, M. (2020). Assessment of the Quality of Life, Health, and Social Wellness in Upper Elementary School Students: Cross-Cultural and Gender Specificity. *Britain International of Humanities and Social Sciences (BIoHS) Journal*, *2*(1), 127-142.

[3].Lvova, M. I., Shvedov, V. V., & Sulimin, V. V. (2020, May). The Use of Information and Communication Technologies in the Analysis of a Healthy Lifestyle of Students at the University. In International Scientific Conference "Digitalization of Education: History, Trends and Prospects"(DETP 2020) (pp. 515-518). Atlantis Press.

[4].Putri, T. E., Subagio, R. T., & Sobiki, P. (2020, November). Classification System Of Toddler Nutrition Status using Naïve Bayes Classifier Based on Z-Score Value and Anthropometry Index. In *Journal of Physics: Conference Series* (Vol. 1641, No. 1, p. 012005). IOP Publishing.

[5].Xu, X., Wang, J., Peng, H., & Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Computers in Human Behavior*, *98*, 166-173.

[6].Whitaker, V., Oldham, M., Boyd, J., Fairbrother, H., Curtis, P., Meier, P., & Holmes, J. (2021). Clustering of health-related behaviours within children aged 11–16: a systematic review. *BMC Public Health*, *21*(1), 1-12.

[7].Islam, M., Rafa, S. R., & Kibria, M. (2021). Early Prediction of Heart Disease Using PCA and Hybrid Genetic Algorithm with k-Means. arXiv preprint arXiv:2101.00183.

[8].Hajiaghayi, M., & Knittel, M. (2021). Improved Hierarchical Clustering on Massive Datasets with Broad Guarantees. arXiv preprint arXiv Datasets with Broad Guarantees. arXiv preprint arXiv:2101.04818.

[9].DeFreitas,K., & Bernard M.(2015).Comparative performance analysis of clustering techniques in Educational data mining.IADIS International journal on computer science & Information systems, 10.(2).

[10].Turabieh,H., Al Azwari,S.Rokaya et.al(2021) Enhanced Harris Hawks optimization as a feature selection for the prediction of students performance,compung, 1-22.

[11].Sharma,S., &jain A. (2021).An algorithm to identify the positive covid19 cases using genetic algorithm(GABFCov19),Journal of interdisciplinary mathematics.

[12].Khan,A., & Ghosh,S.K. (2021).Student performance analysis and prediction in classroom learning A review of Educational data mining studies. Education and information technologies 26(1),205 -240.

[13].Fenollar.P,Roman.S.&Cuestas,P.J(2007).University students' academic performance.An integrative conceptual framework and empirical analysis.British journal of Educational psychology,77(4),873-891.

[14].Sweta.s (2021).Educational data mining Techniques with Modern Approach.In Modern Approach to Educational data mining and its Applications(P.No.25-38),Springer.

[15].Emre,C.A.M., & OZDAG, M.E(2021), Discovery of course success using unsupervised Machine learning Algorithms.Malaysian online Journal of Educational Technology,9(1),26-47.