K-Means, Clustering Algorithm For Student's Selection And Performance Prediction

Osei Wusu Brempong Jnr

Abstract: Machine learning, an application of Artificial intelligence uses computer algorithms designed to make decisions and predict outcomes based on analyzing huge data sets.[1] ML enables systems to automatically change and increase in accuracy without being programmed. One major advantage of ML technology in education is student's selection and prediction of their academic performance. ML beneficial in education is its ability to track learner's progress and also adjust courses to respond to student's needs which helps in increasing student and teacher engagement [2]. ML feedbacks also put instructors in position to analyze and understand student's potential and interest, identify struggling students and provide extra support to struggling student's to overcome learning challenges. GES (Ghana Education Service) has already begun digitalizing the Ghana education system by implementing the computerized school selection and placement system (CSSPS) which is an automated merit base computerized system that uses a deferred acceptance algorithm for assignment[3]. In this system, students are ranked according to their priority levels, they are then proposed as a match to their first choice school in order of their test score ranking. In this paper, we propose a machine learning algorithm K-means clustering to grouping students into ranks of their grades and to analyze their results based on cluster analysis. The student's evaluation factors like average First and Second semester exams, mid-term quizzes are studied. This analysis will enable teachers and school academic administrators to establish prior knowledge of student's grades and predict their performance

Key words : k-means, clustering, academic performance, prediction, GPA.

1 INTRODUCTION

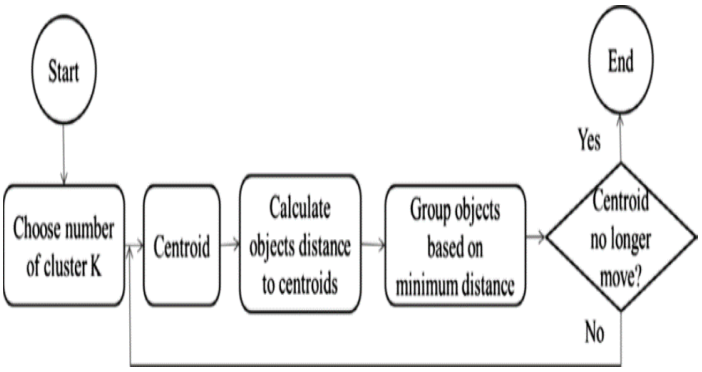
Most Ghanaian students graduating from basic education get assigned to a senior secondary school based on the computer school selection placement system. Senior Secondary schools in Ghana last for of three years in which final examinations taken are conducted by the West Africa Examination Council (WAEC). Ghana Senior Secondary school level courses include compulsory core subjects and electives from the general (arts and science options), agriculture and environmental studies, business, vocational and technical. Eligible candidates from final year (3rd year) secondary schools are registered to sit for the exams in May/June by relevant school authorities[4]. These examinations include multiple-choice questions as well as essays that cover four mandatory core subjects, English, Integrated Science, Mathematics, and Social Sciences. Assessing students to make them eligible to undertake WAEC exams as part of today's educational system in Ghana. It serves as an individual evaluation system and a way to compare the performance of the student body. The main purpose to perform assessments and evaluate students' performance is to gather relevant information about student academic progress, determine student interest to make a judgment about their learning process, and evaluate their readiness to take the WAEC final examinations. With assessment information, teachers and school academic administrators can reflect on each student's level of achievement as well as on specific inclinations of the group to prepare teaching plans to suit their academic needs. Graded Point Average (GPA) is one of the main commonly used indicators to analyze student academic performance. With the help of machine learning such as clustering algorithms, it is possible to discover the key characteristics of students' performance and use them for future predictions.

- *Osei Wusu Brempong Jnr is currently pursuing PhD degree program in Computer Science and Technology in Dalian Maritime University, in Dalian China. kobeoseijnr@hotmail.*

K-means clustering algorithm with Euclidean distance measures where the distance is computed by finding the square of the distance between each score, summing the squares, and finding the square root of the sum. This paper presents k-means clustering a machine learning algorithm as an effective tool to evaluate student's performance, help select students that are eligible for WAEC examinations, and make a future prediction on students' academic progress in Senior Secondary Schools in Ghana. Machine learning is an application of artificial intelligence that enables software applications to become accurate at making predictions without being explicitly programmed, the ML algorithms use data as input to predict new output values[5]. In Unsupervised ML, the algorithms are used to analyze and cluster unlabeled datasets, also discover hidden data without the need for human intervention[6]. Cluster analysis is an unsupervised ML algorithm that provides insight into the data being analyzed by dividing the objects into groups(clusters) of objects, such that objects in the cluster become more similar to each other than to objects in other clusters[7]. Clusters do not use external information such as class labels. K-means clustering one of the oldest and most widely used clustering algorithms is a prototype-based, simple partitioned clustering algorithm that finds K non-overlapping clusters[8]. These clusters are represented by their centroids. This study makes use of a k-means clustering algorithm to segment students into groups according to their Year-1 first semester grades and the average of year -1, first and second-semester grades to evaluate and predict their academic performance. In Ghana, Senior Secondary Schools years are divided into two semesters. This evaluation enables instructors to select students who require additional help in their academics to make them more eligible for WAEC final examinations.

2 METHODOLOGY

The k-means algorithm proposed for this paper is a prototype-based, simple partitional clustering algorithm that attempts to find K non-overlapping clusters[9]. These clusters are represented by their centroids which is the mean of the point in that cluster. The process of clustering a K-means is as follows.



The procedure of k -means algorithm[10]

Step 1: Start with the fixed number of clusters and select an initial partition.

Step2: After determining the cluster centers, we assign each object to the object nearest cluster center.

Step3: Determining the cluster centers, or centroids of the clusters, based on the new partition created by the completion of step 2.

Step4: Repeat steps 2 and 3 until an optimum value of the objective function is achieved.

Step5: If possible adjust the number of clusters through merging and splitting existing clusters. At this time, the removal of cluster outliers can be made.

Repeat Steps 2-5 until cluster membership stabilizes

Given $D = \{x_1, \ldots, x_n\}$ is the data set to be clustered. K-means can be expressed by an objective function that depends on the proximities of the data points to the cluster centroids as follows[11]:

$$min_{\{m_k\},1\le k\le K} \sum_{k=1}^{k} \sum_{x\in c_k} \pi_x dist(x, m_k) \quad (1)$$

Where π_x is the weight of x, n_k is the number of data objects assigned to clusters

$$c_k, m_k = \sum_{x\in C_k} \frac{x_x x}{n_k} \quad (2)$$

is the centroid of the cluster C_k, K is the number of clusters set by the user, and the function 'dist.' computes the distance between object x and centroid $m_k, 1 \le k \le K.$ While the selection of the distance function is optional, the squared Euclidean distance, i.e. $\|x - m\|^2$ is used for this research paper. The K-means algorithm chosen for this paper has some distinct advantages compared with other clustering algorithms. That is, K-means is very simple and robust, highly efficient, and can be used for a wide variety of data types. Its therefore been ranked as the second among the top-10 data mining algorithms and has also become the de facto benchmark method for newly proposed methods.

2.1 DATA- COLLECTION

Sample Dataset collected from Juabeng Senior High, a public senior secondary school in the suburb of Kumasi, Ghana's second-biggest city. A sample of academic records of first-year students in SHS-1 2019, which consist of 3 different classes of pure science students. Data set attributes to contain: the record of student ID number, Student name, an average of semester 1 grades, an average of semester 2 grades, and first and second-semester total grades. The average grades of each semester include scores for all core subjects which are Physics, Chemistry, Biology, Mathematics, English, ICT, and Social Studies.

PERFORMANCE INDEX

90 and Above	EXCELLENT	4.0GPA
80-89	VERY GOOD	3.0GPA
60-79	GOOD	2.5GPA
50-59	FAIR	2.0GPA
Below 45	POOR	1.0GPA

Performance index translating student's percentage scores into GPA as an academic performance indicator. Class one consist of 30 students, class two consist of 25 students and class three 20 students.

CLASS -1

ID	NAME	sem1	sem2	average
101	S.AGYEI	75	70	72.5
102	K.NKANSAH	45	30	37.5
103	O.WUSU	80	85	82.5
104	W.BREMPONG	40	55	47.5

CLASS -2

ID	Name	sem1	sem2	average
301	O.KWABENA	90	95	92.5
302	K.WUSU	88	85	86.5
303	U.PEPRAH	87	90	88.5
304	R.ASHITEY	80	85	82.5

CLASS- 3

ID	NAME	sem1	sem2	average
401	S.ANTWI	50	42	46
402	E.ASANTEWAA	44	50	47
403	A.SAMUEAL	48	45	46.5
404	A.EMMANUEL	60	55	57.5

2.2 DATA PREPROCESSING

Data preparation transforms raw data into an understandable format for ML algorithms. Most data are noisy containing errors, outliers, duplicates, etc. Data preprocessing eliminate noise and errors in the Data. In this paper we practiced with the use of Normalization for data preprocessing, this is done by scaling the data such that each element of the dataset lies in the range $[0:1]$. Each element of the dataset is generated as

$$X_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}, x_{max} \neq x_{min} \quad (3)$$

2.3 EVALUATION METHODS

To evaluate how well our models will be performing based on different K −clusters, 2 metrics were used that may help give us some intuition about K.[12]

Elbow method

The elbow method illustrates an idea of how a good K number of clusters would be based on using SSE between the cluster centroids and the data points. K is picked out at the point where SSE begins to flatten and gradually developing into an elbow.

Silhouette Analysis

This analysis helps to determine the level of separation between the clusters. Is done by computing the average distance from all data in the same cluster (ai) and computing the average distance from all points in the closest cluster (bi)

$$\frac{b^i - a^i}{max(a^i, b^i)} \quad (4)$$

a, b takes the values in the interval [-1,1]

if $0->$ the sample becomes very close to the nearest clusters.

If $1->$ the sample is far from the nearest clusters

If $-1->$ the sample is assigned to the wrong clusters

Since we want the coefficients to be big as possible and close to 1 to have a good cluster.

3 RESULTS

The model was applied to our dataset (academic results of two semesters) of Juabeng Senior High pure science classes. The results for class1 consisting of 30 students are depicted in table-1, figure-1a and figure-1b respectively.

Table -1

Cluster	Student Size	average grade	GPA	Percentage%
1	10	92	4.0gpa	33
2	7	50	2.0gpa	23
3	6	36	1.0gpa	20
4	7	75	2.5gpa	23

In table-1, the number of clusters $= 4$ with cluster 1 consisting of 10 students with an average grade of 92% in their first year. This gives a GPA of 4.0 indicating excellent performance, therefore showing that 33% of students in class1 are predicted to be excellent in their academics and automatically eligible for the final WAEC (West Africa Examination Council) examination. Cluster 2 in class1 consist of 7 students with an average grade of 50% giving a GPA of 2.0 indicating a Fair performance level. Cluster 2 shows that 23% of students in class1 are predicted Fairly eligible for WAEC examinations. Cluster 3 consists of 6 students with an average score of 36% indicating Poor performance with a GPA of 1.0, proving that 20% of students in class1 are below-average level and underperforming in their academics. Making them non-eligible for WAEC examinations. Cluster 4 with 7 student's GPA of 2.5, an average score of 75% indicating Good Performance, making them eligible for WAEC examinations.

As demonstrated in this document, the numbering forsections upper case Arabic numerals, then upper case Arabic numerals, separated by per

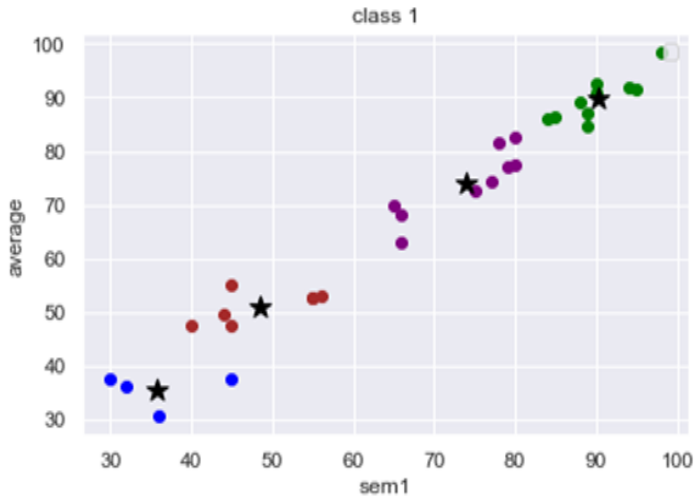


Figure 1a. scatter plot of class1 with K=4. cluster centroid indicated in black asterisk.

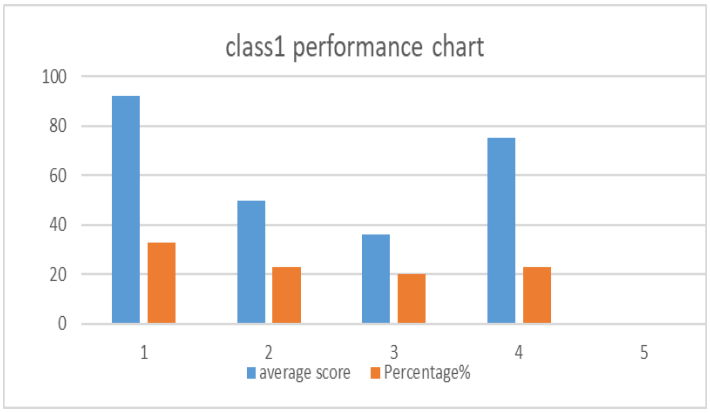


Figure-1b. chart of class1 students' academic performance showing average score versus percentage of student's.

TABLE- 2

Cluster	Student Size	average grade	GPA	Percentage%
1	11	85	3.0gpa	44
2	3	55	2.0gpa	12
3	11	90	4.0gpa	44

The results for class-2 consisting of 25 students are depicted in table-2, figure-2a, and figure-2b respectively

In table class-2, number of clusters $= 3$ with cluster 1 consisting of 11 students with an average grade of 85% in their first year with a GPA of 3.0 indicating very good performance. Cluster 2 in class 2 consists of 3 students with an average grade of 55% giving a GPA of 2.0, Cluster 3 consists of 11 students with an average score of 90% also indicating excellent performance with a GPA of 4.0, these records show that 88% of students in class2 are automatically considered eligible for WAEC examinations due to their good and excellent performance of average scores 85% and 90% respectively. Record also indicates that only 12% of total students in class2 are with average academic performance and fairly eligible for WAC examinations

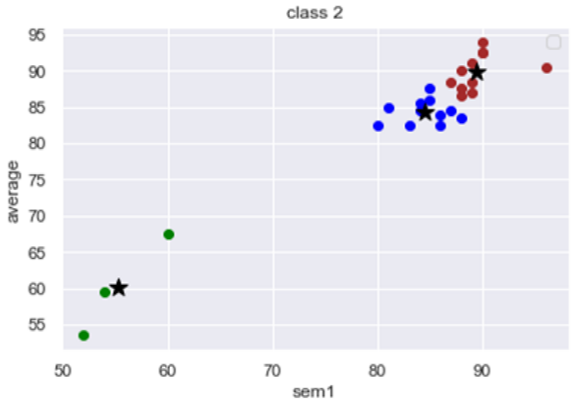


Figure -2a. Scatter plot of class2 students with K=3. Cluster centroid indicated in black asterisk

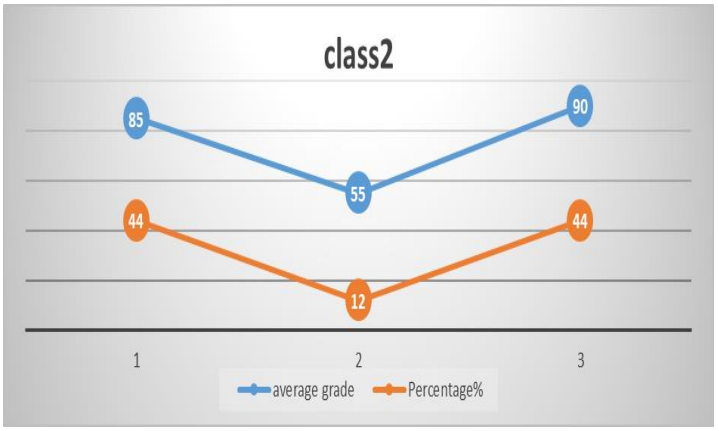


Figure-2b. chart of class2 students' academic performance showing average score versus percentage of student's.

Table -3

Cluster	Student Size	average grade	GPA	Percentage%
1	5	57	2.0gpa	25
2	12	44	1.0gpa	60
3	3	79	2.5gpa	15

The results for class-3 consisting of 20 students is depicted in table-3, figure-3a, and figure-3b respectively

In table class-3, the number of clusters $= 3$, cluster 1 consists of 5 students with an average grade of 57% in their first year with a GPA of 2.0 indicating a fair performance. Cluster 2 in class-3 consist of 12 students with, an average grade of 44% giving GPA of 1.0, Cluster 3 consist of 3 students, an average score of 79% also indicating the good performance of GPA 2.5, these records show that only 15% of students in class-3 may be considered eligible for WAEC examinations based on their good performance score of 79%.

Figure-3a. chart of class3 students' academic performance showing average score versus percentage of student's

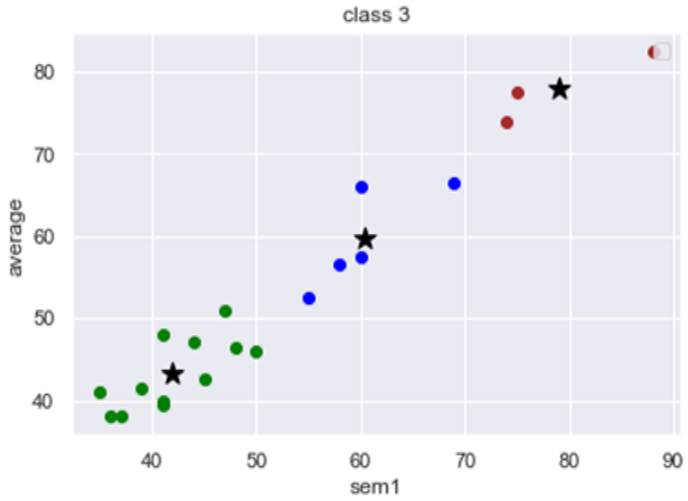


Figure -3b. Scatter plot of class3 students with K=3. Cluster centroid indicated in black asterisk

4 CONCLUSION AND DISCUSSION

This paper has successfully implored the use of the K-means clustering algorithm to group students based on their academic grades. The Euclidean distance used for calculating and measuring the distance between points in datasets was applied. The model was executed with a sample of a dataset from Juabeng Senior High School of first-year students taking pure science classes which consist a total of 95 science students making up class1,2 and 3. Students' academic records from First year semester1 and semester2 were calculated. These records are clustered to produce the percentage of students grouped into their average GPA. Silhouette analysis used to determine the level of separation between the clusters was applied. This research paper recommends a solution to student assessment crises within the Ghana Education Systems of which most secondary schools fail to provide individual evaluation systems which can compare the performance of the student body and to gather relevant information about student academic progress and evaluate their readiness for sitting WAEC final exams. This model has demonstrated that students' academic performance can be predicted from their First-year and selected into groups eligible for final exams in their third year. The model also provides student assessment information of which teachers and school academic administrators can reflect on each student's group level of achievement and solicit teaching plans to suit the academic needs of groups predicted to be ineligible for WAEC exams.

REFERENCES

[1] ED Burns, TechTarget. .in-depth guide to machine learning in the enterprise. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML. (2020 May 15th).

[2] Victoriia Rohalveych, Intellias.. machine learning in education: Benefits and opportunities to explore. https://www.intellias.com/benefits-of-machine-learning-in-education/(2021 January 4th)

[3] Amedahe, F.K; & Asamoah-Gyimah, EIntroduction to educational research. (2005). Cape Coast: Centre for Continuing Education of the University of Cape Coast (CCEUCC)

[4] Mehwish Kamran,Yigu Lian ,WENR.Education in Ghana. https://wenr.wes.org/2019/04/education-in-ghana.. (2019 April 16th).

[5] ED Burns, TechTarget. in-depth guide to machine learning in the enterprise. https://searchenterpriseai.techtarget.com/definition/machine-learning-ML. (2020 May 15th).

[6] Juliana Delua, IBM, supervised vs unsupervised': What's the Difference. https://www.ibm.com/cloud/blog/supervised-vs-unsupervised-learning.. (2021 March 12th)

[7] Auffarth, B, "Clustering by a Genetic Algorithm with Biased Mutation Operator".. (July 18–23, 2010). Wcci Cec. IEEE

[8] M.S. Aldenderfer and R.K. Blashfield. Cluster Analysis. Sage Publications, Los Angeles, 1985

[9] Junjie Wu, (2012). Advance in K-means Clustering, 2012. (pp.1-16): Springer

[10] Jain A K, Murty M N and Flynn P J 1999 Data clustering: a review ACM computing surveys (CSUR) 31 3 264-323

[11] Junjie Wu, Advance in K-means Clustering, 2012. (pp.7-10): Springer

[12] Imad Dabbura, Towards Data Science. k-means Clustering: Algorithm, Applications, Evaluation, Methods and Drawbacks. https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a. (2018 September 12th)