تم استلام الورقة بتاريخ: 9 /2023/12م وتم نشرها على الموقع بتاريخ: 14 /2024/1م

Students' Performance Analysis Using the K-Means Clustering Algorithm

Kamal Ali Albashiri

Department of Data Analysis- Faculty of Accounting
Gharyan University- LIBYA
kamal.albashiri@gu.edu.ly

Abstract

The purpose of educational data mining is to extract important details from the field's existing data in order to find hidden, significant, and valuable information that can be used to enhance student performance.

The research work in this paper uses cluster analysis to visualize the students' performance by grouping them according to certain qualities, comparing the results, and creating representations of the performance.

This paper presents the k-means clustering algorithm as a simple and efficient tool to monitor the progression of students' performance in higher institutions.

The findings will be useful in identifying students who may struggle academically and in determining areas that require modifications to instructional practices in order to provide these students with better support.

Keywords: Data analysis, educational data mining, k-means algorithm, clusters, and performance.

تم استلام الورقة بتاريخ: 9 /2023/12م   وتم نشرها على الموقع بتاريخ: 14 /2024/1م

تحليل أداء الطلاب باستخدام خوارزمية التجميع K-Means

كمال علي البشيري
قسم تحليل البيانات – كلية المحاسبة – جامعة غريان – ليبيا
kamal.albashiri@gu.edu.ly

الملخص

الغرض من التنقيب في البيانات التعليمية هو استخراج تفاصيل مهمة من البيانات الموجودة في هذا المجال من أجل العثور على معلومات مخفية ومهمة وقيمة يمكن استخدامها لتعزيز أداء الطلبة.

يستخدم في هذا البحث التحليل العنقودي لتصور أداء الطلاب الجامعيين من خلال تجميعهم وفقًا لصفات معينة، ومقارنة النتائج، وإنشاء تمثيلات بيانية للأداء.

يقدم هذا البحث خوارزمية التجميع k-means كأداة بسيطة وفعالة لمراقبة تقدم أداء الطلاب في المؤسسات الجامعية.ستكون النتائج مفيدة في تحديد الطلاب الذين قد يعانون أكاديميًا وفي تحديد المجالات التي تتطلب إصلاحات على الممارسات التعليمية من أجل تزويد هؤلاء الطلاب بدعم أفضل.

الكلمات المفتاحية: تحليل البيانات، وتنقيب البيانات التعليمية، وخوارزمية k-means، والتجميع، والأداء الأكاديمي.

1. Introduction

One of the most powerful machine-learning tools is clustering, which is an unsupervised learning technique used to discover patterns and unseen data. K-means clustering is one popular technique in the field of unsupervised learning that involves clustering data into a predefined number of clusters. The benefits of k-means clustering are its low processing cost, ability to build stable and compact clusters, ability to converge after several interactions, and simplicity of execution [1]. A given data collection is categorized and sorted using k-means clustering by first specifying the number of clusters, k. For every k clusters, there are k centroids.

تم استلام الورقة بتاريخ: 9 /2023/12م وتم نشرها على الموقع بتاريخ: 14 /2024/1م

The algorithm is used to locate observational groups in the data that haven't been deliberately selected. The algorithm's goal is to create a better data arrangement so the outcome is that the data has comparable features [2].

Clustering in higher education means it classifies the student by their academic performance. Lack of deep and sufficient knowledge in the higher system may prevent system management from achieving quality objectives. Data clustering methodology can help bridge these knowledge gaps in the higher education system and can help in generating the right decision [3]. There are many studies in the literature that have shown similarities in using unsupervised machine learning, especially clustering algorithms, to analyze and predict student performance [4]. According to these studies, the clustering technique has shown good performance in making predictions and producing interesting patterns when used with students' data.

This paper aims to provide insights related to the student's performance using the k-means clustering algorithm by analyzing data obtained from the databases of a university in Libya[1].

The work model is shown in Figure 1 and described in the following sections. The data manipulation steps provide a description of the information, statistical techniques, and models applied in this work.

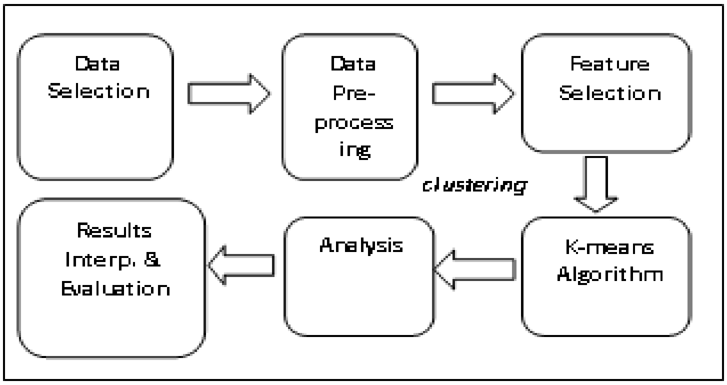


Figure 1. Steps of students' performance analysis.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

Objectives and methodology are presented. The k-means method is also discussed. The results of the work observations are covered in the later sections. The experiments are applied using the sophisticated data mining tool (Orange)² [5].

The remainder of the paper is organized as follows: Section 2 presents the research objectives. Section 3 presents the methodology and steps taken in this paper. Section 4 describes the selected data. In Section 5, data pre-processing steps are explained.

The development of clustering models using the k-means algorithm is described in Section 6. Experiments and results are discussed in Section 7, while the conclusion of this paper and further works are outlined in Section 8.

2. Objective

Tools and approaches that can automatically analyze students' data are needed to derive hidden patterns and knowledge that could be of great use to provide insights into student performance.

The aim of clustering in this paper is to partition students into homogenous groups according to their academic achievements, as measured by their results. These techniques can help both instructors and students improve their quality of education. It can provide students, faculty, and administrative staff with outcomes that are simple to understand. The instructor can analyze different causes of low academic achievement and introduce effective teaching methods. The new outcome may motivate students to study hard and progress in their academic performance.

3. Methodology

The methodology steps taken in this work are:

1. Partition students into homogenous clusters according to their academic achievements (*high*, *medium*, and *low* performing students);
2. Perform a comparison of the results of the students from each cluster;
3. Create visualization views of student performance;

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

4. Derive insights (trends/patterns/facts) from the performance analysis.

4. Data Selection

The dataset selected for this work consists of 338 records regarding university students, with twelve attributes such as demographics, and academic-related data includes year of study, gender, semester, birthday, high school type, and final grades at the end of the year for two conductive semesters. All of them were undergraduate students enrolled in six different departments and four different years for the 2021–2022 academic years. The data portions were chosen to be relatively equal for each department and study year, as depicted in figure 2.

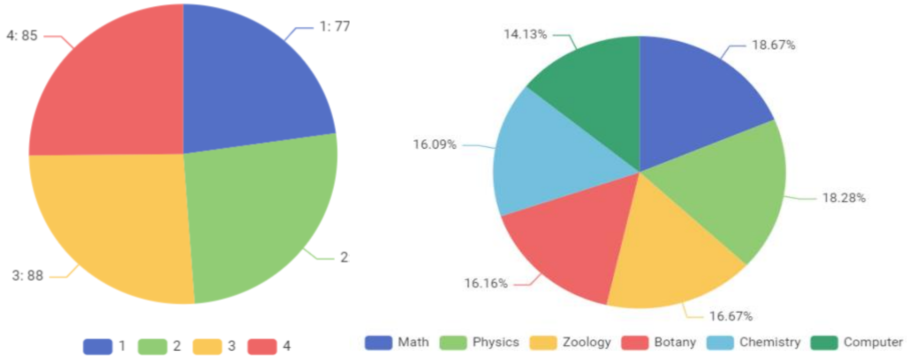


Figure 2. Departmental and year distribution.

5. Data Pre-processing

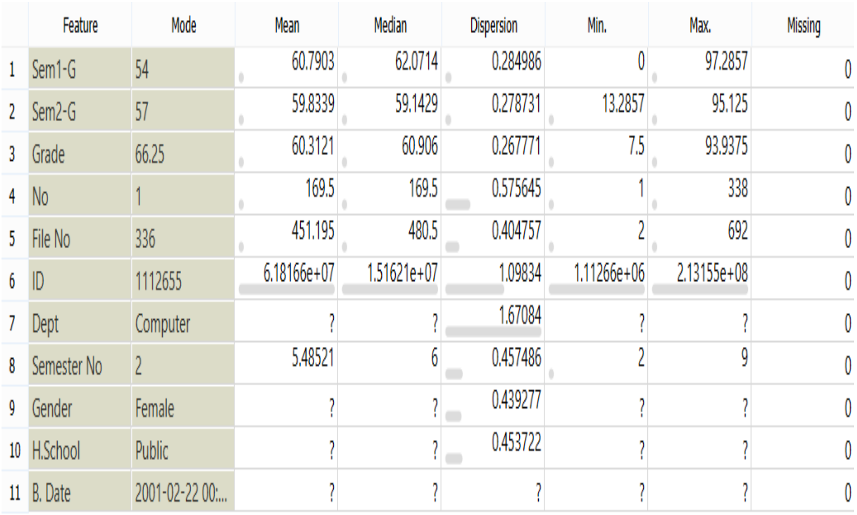
This step focuses on transforming the dataset to ensure it is suitable for the clustering algorithm and the data mining tool. Too many attributes in datasets can cause a curse of dimensionality and difficulties when processing and analyzing the data. Most machine learning algorithms generally require numeric input and output variables [6, 7]. This restriction must be addressed when implementing and developing machine-learning models. This

تم استلام الورقة بتاريخ: 9 /2023/12م   وتم نشرها على الموقع بتاريخ: 14 /2024/1م

implies that all characteristics, including categories or nominal variables, must be transformed into numeric variables before being fed into the clustering model.

All these should be dealt with in the data transformation process where several data pre-processing techniques and tasks are conducted on the raw dataset to ensure the quality of the training data. Table 1 displays the list of attributes and their statistics.

Table 1. List of attributes' statistics.

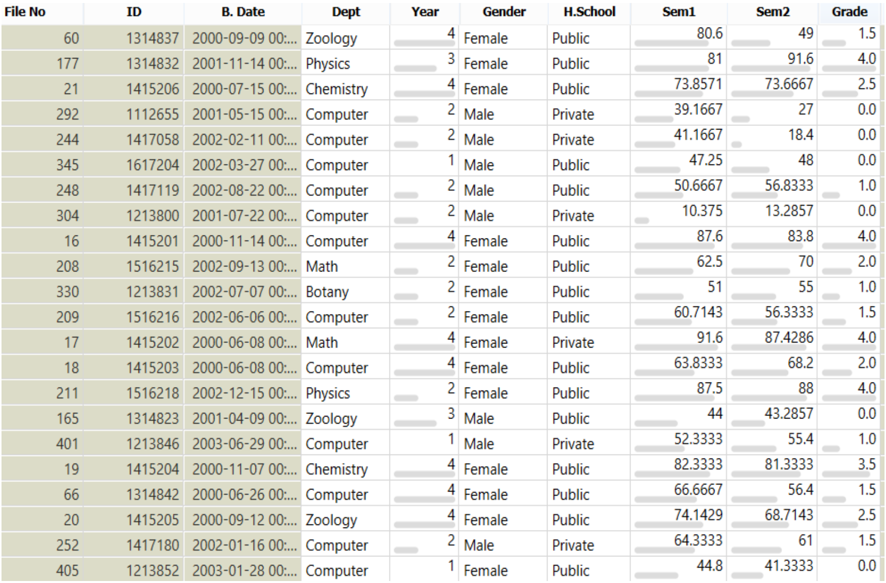


	Feature	Mode	Mean	Median	Dispersion	Min.	Max.	Missing
1	Sem1-G	54	60.7903	62.0714	0.284986	0	97.2857	0
2	Sem2-G	57	59.8339	59.1429	0.278731	13.2857	95.125	0
3	Grade	66.25	60.3121	60.906	0.267771	7.5	93.9375	0
4	No	1	169.5	169.5	0.575645	1	338	0
5	File No	336	451.195	480.5	0.404757	2	692	0
6	ID	1112655	6.18166e+07	1.51621e+07	1.09834	1.11266e+06	2.13155e+08	0
7	Dept	Computer	?	?	1.67084	?	?	0
8	Semester No	2	5.48521	6	0.457486	2	9	0
9	Gender	Female	?	?	0.439277	?	?	0
10	H.School	Public	?	?	0.453722	?	?	0
11	B. Date	2001-02-22 00:...	?	?	?	?	?	0

The dataset was narrowed down to only include records with two consecutive semesters from six majors (departments) and only considered significant attributes, while eliminating unimportant ones like names and addresses. Furthermore, attributes (ID, gender, birth date, year number, file number, and student department attributes) are chosen at the end of the unsupervised feature selection and will be used later in the analysis. Then data cleaning was carried out to clear the attributes with too many missing values. The attributes that contain some missing values are manually replaced with specified values. Outlier values and repetitive variables are also removed from the dataset. In addition, some string

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

values were translated from Arabic to English. Figure 3 shows a sample of the dataset.

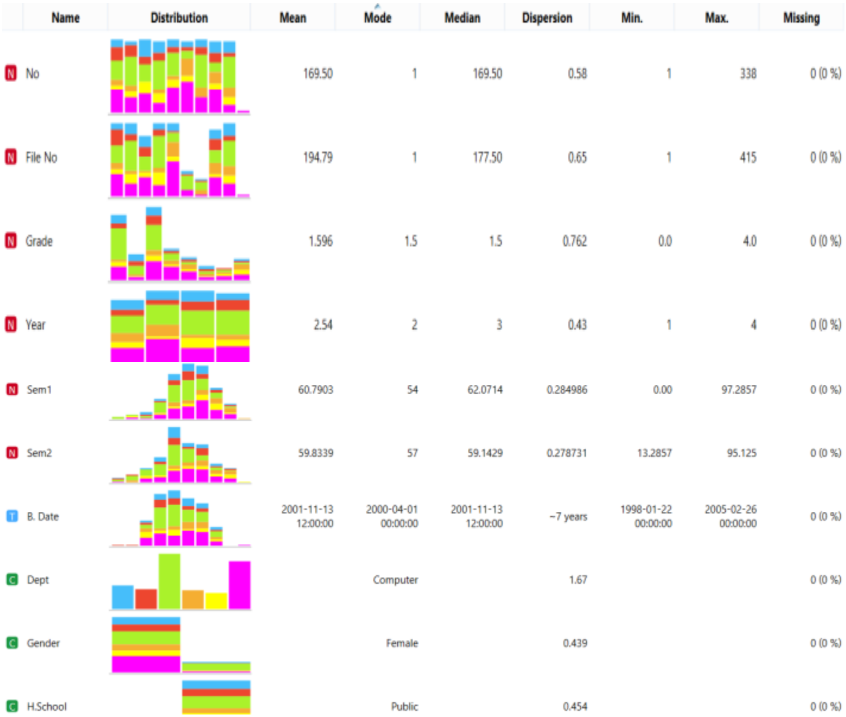


File No	ID	B. Date	Dept	Year	Gender	H.School	Sem1	Sem2	Grade
60	1314837	2000-09-09 00:...	Zoology	4	Female	Public	80.6	49	1.5
177	1314832	2001-11-14 00:...	Physics	3	Female	Public	81	91.6	4.0
21	1415206	2000-07-15 00:...	Chemistry	4	Female	Public	73.8571	73.6667	2.5
292	1112655	2001-05-15 00:...	Computer	2	Male	Private	39.1667	27	0.0
244	1417058	2002-02-11 00:...	Computer	2	Male	Private	41.1667	18.4	0.0
345	1617204	2002-03-27 00:...	Computer	1	Male	Public	47.25	48	0.0
248	1417119	2002-08-22 00:...	Computer	2	Male	Public	50.6667	56.8333	1.0
304	1213800	2001-07-22 00:...	Computer	2	Male	Private	10.375	13.2857	0.0
16	1415201	2000-11-14 00:...	Computer	4	Female	Public	87.6	83.8	4.0
208	1516215	2002-09-13 00:...	Math	2	Female	Public	62.5	70	2.0
330	1213831	2002-07-07 00:...	Botany	2	Female	Public	51	55	1.0
209	1516216	2002-06-06 00:...	Computer	2	Female	Public	60.7143	56.3333	1.5
17	1415202	2000-06-08 00:...	Math	4	Female	Private	91.6	87.4286	4.0
18	1415203	2000-06-08 00:...	Computer	4	Female	Public	63.8333	68.2	2.0
211	1516218	2002-12-15 00:...	Physics	2	Female	Public	87.5	88	4.0
165	1314823	2001-04-09 00:...	Zoology	3	Male	Public	44	43.2857	0.0
401	1213846	2003-06-29 00:...	Computer	1	Male	Private	52.3333	55.4	1.0
19	1415204	2000-11-07 00:...	Chemistry	4	Female	Public	82.3333	81.3333	3.5
66	1314842	2000-06-26 00:...	Computer	4	Female	Public	66.6667	56.4	1.5
20	1415205	2000-09-12 00:...	Zoology	4	Female	Public	74.1429	68.7143	2.5
252	1417180	2002-01-16 00:...	Computer	2	Male	Private	64.3333	61	1.5
405	1213852	2003-01-28 00:...	Computer	1	Female	Public	44.8	41.3333	0.0

Figure 3. dataset sample.

Feature selection is employed in this stage to obtain a suitable dataset for clustering development. The features that have a high variance are the date of birth, the scores of semesters (sem1 and sem2), grade, and ID. On the other hand, there are four attributes with a low variance, namely gender, name, address, and file number. As shown in figure 4, the feature type can be categorical, numeric, time, or string. The histogram shows the distribution of the feature's values. The values of numeric features are split into bins. Further columns show different statistics (mean, minimal, and maximal values, which are computed only for numeric features). Mode shows the most common value for a numeric or categorical feature.

تم استلام الورقة بتاريخ: 9 / 2023/12م   وتم نشرها على الموقع بتاريخ: 14 / 2024/1م



	Name	Distribution	Mean	Mode	Median	Dispersion	Min.	Max.	Missing
N	No		169.50	1	169.50	0.58	1	338	0 (0 %)
N	File No		194.79	1	177.50	0.65	1	415	0 (0 %)
N	Grade		1.596	1.5	1.5	0.762	0.0	4.0	0 (0 %)
N	Year		2.54	2	3	0.43	1	4	0 (0 %)
N	Sem1		60.7903	54	62.0714	0.284986	0.00	97.2857	0 (0 %)
N	Sem2		59.8339	57	59.1429	0.278731	13.2857	95.125	0 (0 %)
T	B. Date		2001-11-13 12:00:00	2000-04-01 00:00:00	2001-11-13 12:00:00	~7 years	1998-01-22 00:00:00	2005-02-26 00:00:00	0 (0 %)
C	Dept			Computer		1.67			0 (0 %)
C	Gender			Female		0.439			0 (0 %)
C	H.School			Public		0.454			0 (0 %)

Figure 4. Feature statistics on the data set.

Dispersion shows the coefficient of variation for numeric features and the entropy for categorical features. From the observation of the histogram distribution of the feature's values, the features of sem1 and sem2 show the highest variance values, so they are a good choice to be used to analyze student performance patterns. Figure 5 shows the scattered plot of the dataset before clustering.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

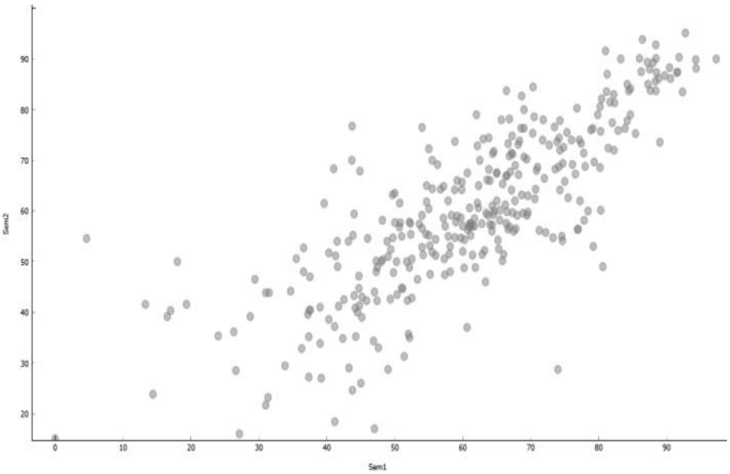


Figure 5. Dataset Scattered Plot before Clustering

6. K-Means Clustering Algorithm

The k-means algorithm is an unsupervised learning and iterative algorithm that tries to partition the dataset into K predefined, distinct, non-overlapping subgroups, called clusters [1].

The following is a summary of the k-means algorithm steps:

1. The algorithm randomly picks k points as the original cluster centers (“means”).
2. Each point in the dataset is assigned to the closed cluster based on the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence means if the output of repeating Steps 2 and 3 does not make a material difference in the definition of clusters or if no observations change clusters.

K-means divides any set of data points into disjoint clusters such that every object in the dataset belongs to only one cluster. This grouping is done on the basis of minimizing the sum-of-squared distances between objects and their respective centroids. The ease of use, simplicity, and satisfactory performance across a wide variety of datasets were the basis for choosing the k-means algorithm.

K-means clustering algorithm uses the Euclidean distance measure, 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.

One measurement is *Within Cluster Sum of Squares* (WCSS), which measures the squared average distance of all the points within a cluster to the cluster centroid. To calculate WCSS, it first finds the Euclidean distance between a given point and the centroid to which it is assigned. Then iterate this process for all points in the cluster, and then sum the values for the cluster and divide by the number of points. Finally, it calculates the average across all clusters. This gives the average WCSS.

The following equation is used to calculate the sum of the squares of the distances between the cluster centers [8].

$$\sum_{\mathrm{i}}^{m}\sum_{\mathrm{j}}^{n}\left|\mathrm{dj}^{(\mathrm{i})} - \mathrm{ci}\right|^{2} \qquad (1)$$

Where

m = total number of students in a cluster

n = the dimension of the dataset *(sem1, and sem2)*

$\left|dj^{i} - ci\right|^{2}$ denotes sum of squares of distances between the cluster centre ci and a data point dj^{i}.

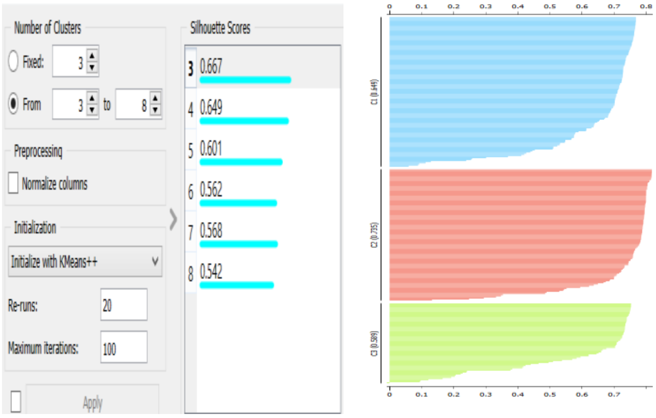
$\sum_{\mathrm{i}}^{\mathrm{m}}\sum_{\mathrm{j}}^{\mathrm{n}}\left|\mathrm{dj}^{(\mathrm{i})} - \mathrm{ci}\right|^{2}$ denotes total distance of all the data points from their corresponding cluster centers.

In general, a cluster that has a small sum of squares is more compact than a cluster that has a large sum of squares.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

There are a number of other metrics for K-means clustering that can help you hone your use of this unsupervised learning method. In this paper the Silhouette Coefficient Method is used [9].

Given the students' dataset comprising 338 records, the k-means algorithm is performed by partitioning the dataset into a fixed number of clusters and, thereafter, searching for the optimum number of clusters that best describe the structure of the dataset. In each phase, the centroids were randomly initialized while constructing k ($n \geq k$) partitions. In the first phase, k was fixed at 3 ($k = 3$) according to the number of clusters and the results presented in figure 6.


Number of Clusters
○ Fixed: 3
● From 3 to 8

Preprocessing
□ Normalize columns

Initialization
Initialize with KMeans++

Re-runs: 20
Maximum iterations: 100

□ Apply

Silhouette Scores	
3	0.667
4	0.649
5	0.601
6	0.562
7	0.568
8	0.542

Figure 6. Silhouette Analysis for the Number of Clusters (k = 3)

6.1 Selecting the number of clusters with silhouette analysis

The silhouette coefficient index is used to evaluate the quality and strength of a group. The high silhouette coefficient value indicates a model with a better batch and signals that an object is well matched to its batch and does not match the adjacent batches. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters [9].

The silhouette analysis is used to choose an optimal value for K clusters [9]. figure 6 shows the silhouette scores of k clusters with values of 3, 4, 5, 6, 7, and 8 and also shows the silhouette plot for k = 3. As it can be seen, three clusters are a good pick for the given dataset.

The plot in figure 6 shows that the three clusters generated by the k-means have the highest index score (0.667) above the silhouette average value line, giving a good picture of the clustering results.

In k-means clustering, data is selected randomly to be the center of the initial cluster according to the number of clusters determined. In the 3-cluster model, data are selected randomly through the average, minimum, and maximum values of the feature attribute values.

Figure 7 shows the three clusters generated by the k-means.

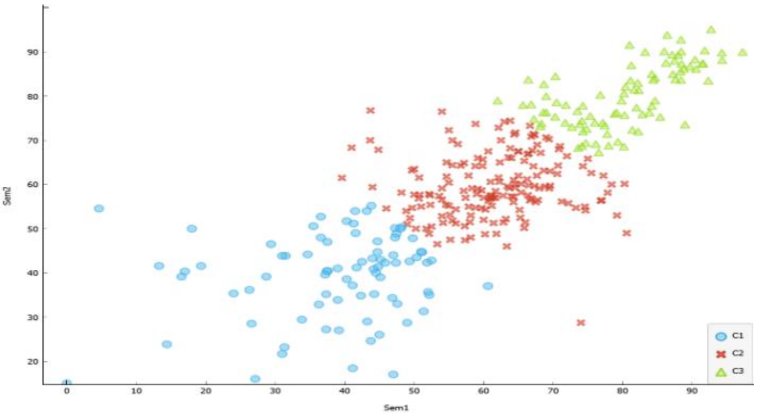


Figure 7. K-mean clustering (k = 3)

7. Experiments and Results

In this research paper, clustering was applied to the obtained dataset and performed on the basis of marks obtained by students in two consecutive semesters in a given academic year. The results were generated by applying the k-means algorithm using Orange

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

software. Three clusters were pre-selected (k = 3), as this can mimic the real proportions of the student performance classification (*high*, *medium*, or *low*).

The k-means algorithm has successfully grouped the data into three desired clusters. A grade, or GPA (Grade Access Point), is calculated for each student for two consecutive semesters. A grade is a score out of 100. Students with a grade below 50 go into *the low* group (cluster 0). Students scoring above 70 are placed in *the high* group (cluster 2), while those scoring between 50 and 70 fall into *the medium* group (cluster 1). Figure 8 depicts the grade percentage of the k-mean clustering (k = 3). According to this performance analysis, 75 (21.55%) students on average of 38.68 fall into *the low* cluster, 103 (33.74%) students on average of 60.55 fall into *the high* cluster, and 160 (44.71%) students on average of 88.23 fall into *the medium* cluster.

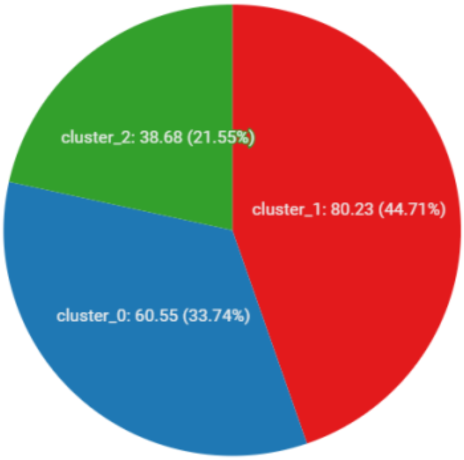


Figure 8. Percentage's of Students Performance

The overall performance is analyzed by applying a number of assessments in each of the clusters. The evaluated results are measured by summing the average of the individual scores in each cluster.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

Visualizing the distributions of the significant features of the dataset provides important insights into the data. Figure 9 displays bar plots of distributions of department, gender, and high school features.

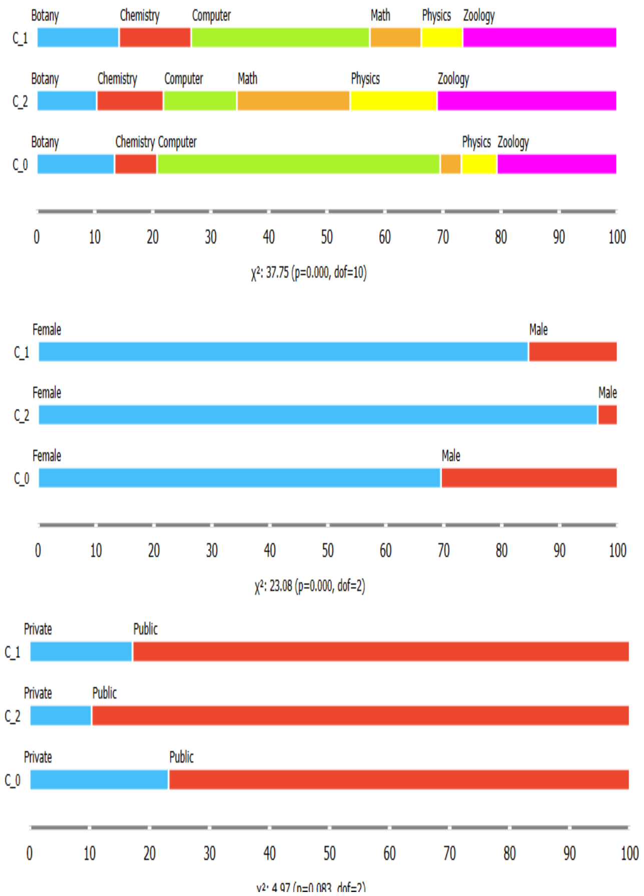


Figure 9. Evaluation results of description variables

Many insights can be derived from the charts shown in figure 9, including:

- The majority of students from the zoology department are in *the high* group.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

- The majority of students in the computer department are in *the low* group.
- The minority of students in the math department are in *the low* group.
- Students from chemistry, botany, and physics are spread almost equally among groups.
- The majority of female students are in *the high* group.
- The majority of male students are in *the low* group.
- The majority of students from private high schools are in *the low* group.

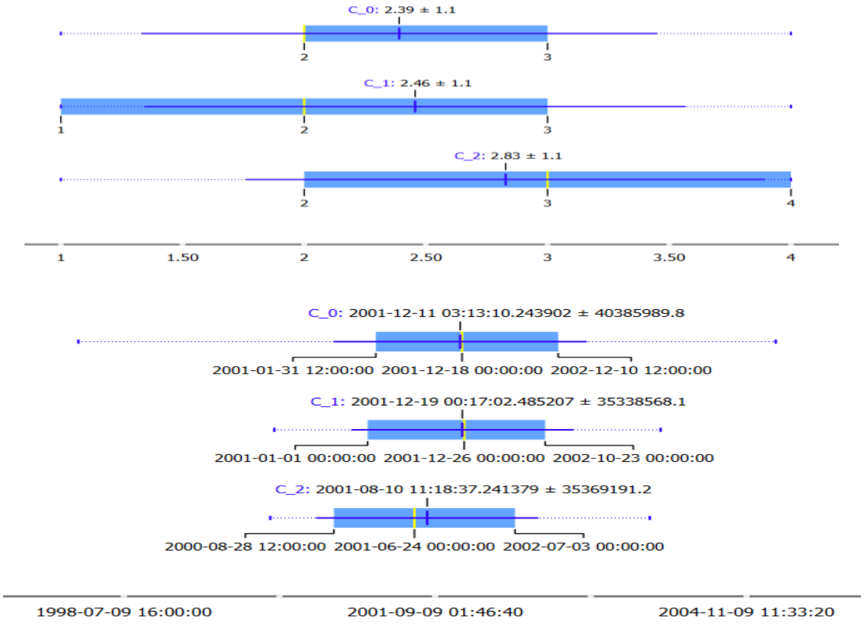


Figure 10. Evaluation results of study year and birthday variables

Furthermore, the evaluation of year number and birthday features is shown in figure 10, can bring more insights derived from the charts include:

- Few of the first-year students are in *the low* group.

- Most of the students in years 2 and 3 are in *the low* group.
- Most of the students in year 4 are in *the high* group.
- The older the student, the more likely they are to be in *the high* group.
- Younger-age students are spread almost equally among groups.

This analysis shows the general level of academic performance and helps to discover hidden facts and trends within the dataset.

8. Conclusion

In this paper k-means clustering algorithm was used for cluster validity analysis and provided a means of clustering and visualizing the various attributes and clusters of students academic performance.

The model was implemented with the Orange tool and tested on a twelve-attribute dataset. The best- performing cluster number was 3, with the average correlation yielding the highest value of 0.667.

The k-mean clustering algorithm serves as a good benchmark to monitor the progression of students’ performance in higher education. It also enhances the decision-making by academic planners to monitor the candidates’ performance year by year by improving future academic results.

As the results show, the attributes or variables that affect student performance are gender (which is why female students score higher than males), age (younger students score less than older students), or department (the majority of zoology students have higher scores). All that might help in taking actions to raise student academic performance in the following academic year. Determining students who are likely to spend another year in an institution or graduate with poor results at an early stage of their studies is of great importance. The findings can also assist student-related departments in providing better services and management, such as psychological consulting and academic guidance.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

In future work, expanding the dataset by adding more attributes related to student performance and carrying out a cluster analysis would result in better cluster performance.

References

[1] MacQueen and J. B., (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1, University of California Press.

[2] PeArez-Ortega J, Almanza-Ortega NN, and Romero D., , (2018). Balancing effort and benefit of K-means clustering algorithms *in Big Data realms*.

[3] Ahuja, R.; Jha, A.; Maurya, R.; Srivastava, R. , (2019). Analysis of Educational Data Mining. *In Harmony Search and Nature-Inspired Optimization Algorithms*, Springer Singapore.

[4] Sunita M. Dol, Dr. P. M. Jawandhiya, , (2023). A Review of Data Mining in the Education Sector, *Journal of Engineering Education Transformations*, Volume No. 36, Special Issue, eISSN 2394-1707.

[5] Demsar J, Curk T, Erjavec A, Gorup C, Hocevar T, Milutinovic M, Mozina M, Polajnar M, Toplak M, Staric A, Stajdohar M, Umek L, Zagar L, Zbontar J, Zitnik M, Zupan B. , (2013). Orange: Data Mining Toolbox in Python, *Journal of Machine Learning Research* 14(Aug): pp. 2349−2353.

[6] S. Alelyani, J. Tang, and H. Liu, , (2013). Feature Selection for Clustering: A Review, in: C. Aggarwal and C. Reddy (eds.), *Data Clustering: Algorithms and Applications*, CRC Press.

[7] Rahman, A.M., Sani, N.S., Hamdan, R., Ali Othman, Z., and Abu Bakar, A., (2021). A Clustering Approach to Identify Multidimensional Poverty Indicators for the Bottom 40 Percent Group. *PLoS ONE*.

[8] Fahim A. M., Salem A. M., Torkey F. A. and Ramadan M. A., (2006). An efficient enhanced k-means clustering algorithm, *Journal of Zhejiang University Science A*., pp. 1626–1633.

تم استلام الورقة بتاريخ: 9 / 2023/12م وتم نشرها على الموقع بتاريخ: 14 / 2024/1م

[9] Shutaywi, M.; Kachouie, N.N. , (2021). Silhouette Analysis for Performance Evaluation *in Machine Learning with Applications to Clustering*. Entropy 23, 759.