Submitted: 27/01/2026 - Accepted: 25/02/2026 - Published: 29/03/2026

Utilizing Cluster Analysis and Discriminant Analysis for Data Classification and Academic Performance Prediction

Mohammad Muosa Al-Shumrani[1*]

[1*]*Taif University Department of Psychology, Taif University, Taif, Saudi Arabia*
m.m.m.a1313@hotmail.com

DOI: 10.26417/24zhfp71

Abstract

This study aims to utilize cluster analysis and discriminant function analysis to classify student data into two groups: high and low academic performance, and to evaluate the accuracy of the cluster classification. The study adopted a descriptive research design, and the sample consisted of a random sample of 62 students. Both cluster analysis and discriminant analysis were applied to the data. The results of the cluster analysis indicated that Aptitude Test scores and Achievement Test scores played a significant role in classifying cases into clusters. In addition, the results of the discriminant function analysis showed that Wilks' Lambda was statistically significant, indicating the discriminant function's ability to distinguish between the two groups. Furthermore, the assessment of classification accuracy revealed that the classification rate based on cluster analysis was very high, confirming the accuracy and effectiveness of the classification process. The study recommends applying cluster analysis to classify cases into homogeneous or closely related clusters within each group.

Keywords: cluster analysis, discriminant analysis, eigenvalue, academic performance, prediction

1. Introduction

Researchers in many studies and research need to classify existing cases into similar and different groups. One of the most well-known research objectives is the construction of classifications, which is one of the reasons that cluster analysis is significant as researchers in all fields need to conduct classifications and continuously review them (Romesburg, 2004).

Ramdeen and Yim (2015) affirm that in everyday life, we attempt to classify similar elements and categorize them into different groups, which is a natural and fundamental way to create a taxonomic system. It is also essential to identify similarities within data to build meaningful groups.

Multivariate statistical analysis techniques, including cluster analysis, can be used to discover a system for organizing observations, where group members share certain common characteristics. It is one of the statistical techniques that classify cases into relatively homogeneous groups internally while being relatively heterogeneous among them.

Cluster analysis and discriminant analysis are widely used multivariate statistical methods for classifying cases or variables. Both are suitable for classification, but they differ in their fundamental concepts while ultimately arriving at results for classifying observations. In this regard, Hardle and Simer (2003) indicate that discriminant analysis focuses on cases where different groups are already known, providing decision rules for classifying cases based on these known groups. Similarly, Ramdeen and Yim (2015) affirm that discriminant analysis classifies new cases into pre-defined groups based on specific criteria.

Despite the existence of many previous studies that have focused on using both methods, their combined application in research still requires further studies, particularly in the educational field. Wilson and Hardgrave (1995) point out that classifying student performance in undergraduate programs and predicting academic success or failure are important issues. They necessitate the use of advanced statistical methods such as cluster analysis and discriminant analysis. Ramdeen and Yim (2015) emphasize that although these techniques have been used in social and health sciences, they have not gained the same widespread popularity as in the natural sciences.

The use of cluster analysis and discriminant analysis in psychology and education remains rare. Therefore, the study aims to address this issue by clarifying how both methods can be used in data analysis and observation classification, determining the extent to which independent variables contribute to the classification process, and developing discriminant functions that can be used to predict the classification of cases into the identified clusters.

1.1. Study Problem and Questions

The problem of the current study lies in answering the following questions:

- How can cases be classified into high and low academic performance using cluster analysis?
- How can discriminant analysis be used to develop a discriminant function that classifies cases into clusters?

1.2. Study Objectives

The current study aims to employ cluster analysis and discriminant function analysis to classify data from a sample of male and female students into homogeneous groups based on high and low academic performance. It also seeks to determine the differences and similarities between these cases and assess the accuracy of classifying the two groups.

1.3. Significance of the Study

The significance of the current study stems from its focus on employing cluster analysis and discriminant function analysis in classifying cases and developing discriminant models. The findings of this study may be beneficial in the following areas:

- Theoretical Significance: By clarifying the methods of cluster analysis and discriminant function analysis.
- Practical Significance: By assisting researchers and professionals in various scientific fields in determining the optimal classification method.

1.4. Study Terminology

Cluster Analysis: A statistical method aimed at classifying a set of cases or variables in specific ways and organizing them into clusters, ensuring that the cases within each cluster are homogeneous concerning certain characteristics (Al-Shafi, 2014).

Discriminant Function: Statistical functions whose number equals the number of dependent variable levels minus one or the number of independent variables, whichever is smaller (Hair et al., 2006).

1.5. Study Delimitations

The current research is limited to studying the application of cluster analysis and discriminant function analysis in analyzing data from a sample of university students. The study was applied to students at Umm Al-Qura University in Makkah, Saudi Arabia in the academic year 2014/2015.

2. Theoretical Framework

2.1. Cluster Analysis

It is a set of tools used to create groups (clusters) from multivariate data, with the aim of forming homogeneous groups from large, heterogeneous sets. Romesburg (2004) mentions that cluster analysis is a general term for a variety of mathematical methods that can be used to classify cases into similar groups.

Cluster analysis develops methods and tools for classifying a group of individuals by grouping "similar" individuals according to certain relevant criteria. Cluster analysis

is applied in many fields such as medical sciences, economics, marketing, natural sciences, and more (Hardle & Simer, 2003).

King (2015) points out that interest in cluster analysis began in the 1960s in the fields of biology and ecology. Two events led to an explosion of interest in cluster analysis. First, the availability and widespread use of large, high-speed computers provided new possibilities for researchers. Additionally, the publication of numerical taxonomy principles by Sokal and Sneath covered three areas: first, a variety of different cluster analysis techniques; second, the use of computers in classificatory research; and third, the experimental approach to classification in the life sciences and the need for cluster analysis in many areas of study.

Cluster analysis is considered a type of data reduction technique, which also includes factor analysis and discriminant analysis. These techniques primarily reduce data. For example, factor analysis reduces the variables in a model and transforms them into factors, while discriminant analysis classifies new cases into predefined groups based on specific criteria. Cluster analysis is unique among these techniques because its goal is to reduce the number of cases or observations by classifying them into homogeneous groups and identifying these groups without needing to know the group membership or the number of potential groups in advance. Cluster analysis also allows for various options regarding the algorithm for merging groups, with each option resulting in a different aggregation structure. Therefore, cluster analysis can be a useful statistical tool for exploring underlying structures in different types of data sets (Ramdeen & Yim, 2015).

2.2. Measures of Similarity or Dissimilarity

Cluster analysis uses various similarity or dissimilarity indices between each pair of observations. The distance measure between two observations is considered the appropriate metric for determining the degree of proximity between them. The most commonly used distance function is the Euclidean distance (Rencher, 2002), which can be calculated using the following formula: $d(x, y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$

Xj represents the values in group X.

Yj represents the values in group Y.

2.3. Types of Cluster Analysis

1. Hierarchical Clustering

Hierarchical clustering methods, along with other clustering algorithms, attempt to identify meaningful groups in the data using a computationally efficient technique (Rencher, 2002). The classification consists of a series of divisions that may range from a single group containing all individuals to n groups, each containing a single

individual (Everitt et al., 2011). Hierarchical clustering does not require prior knowledge of the number of clusters.

$$N(n,g)=\frac{1}{g}\sum_{k=1}^{g}\binom{g}{k}(-1)^{g-k}k^{n}$$

There are several methods that can be used

1-Single Linkage (Nearest Neighbor) Method

In this method, the smallest distance between each pair of observations is calculated. It assumes that the most similar elements form the nucleus of the cluster, and then the remaining elements are added based on their similarity, starting from the most similar to the least (Romesburg, 2004). The linkage process relies on the shortest distance between pairs of cases and links them together using the following formula (Rencher, 2002):

$$D(A,B)=\min\left\{d(y_i,y_j),y_i\in A,y_j\in B\right\}$$

2-Complete Linkage (Farthest Neighbor) Method

This method assumes that the least similar elements between the cases form the nucleus of the cluster. The largest distance between the cases is then calculated, and they are linked together using the following formula:

$$D(A,B)=\max\left\{d(y_i,y_j),y_i\in A,y_j\in B\right\}$$

3-Average Linkage Method

This method uses the average distance between a point from the first cluster and a point from the second cluster. The formula for this method is:

$$D(A,B)=\frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B}d(y_i,y_j)$$

4-Centroid Method

In the Centroid Method, the similarity between two clusters is defined as the similarity between their central points (centroids). The centroid of a cluster is the mean of the coordinates of all the points in that cluster. This method computes the distance between the centroids of two clusters and uses this distance to determine their similarity.

$$D(A,B)=d(\overline{y}_A,\overline{y}_B)$$

5- Wad's Method

The minimum variance clustering method is one of the most used methods. This method follows a series of aggregation steps that begin with clusters, each containing one element and ends with a single cluster containing all the elements. This method uses the sum of squares or variance index (Romesburg, 2004).

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \overline{y}_A)'(y_i - \overline{y}_A)$$

$$SSE_B = \sum_{i=1}^{n_B} (y_i - \overline{y}_B)'(y_i - \overline{y}_B)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \overline{y}_{AB})'(y_i - \overline{y}_{AB})$$

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

2) Nonhierarchical Methods

K-Means Method: One of the most used methods, and it requires determining the number of clusters in advance. This method is based on classifying cases into homogeneous groups in terms of characteristics using algorithms that can handle a large number of cases (Ahmed, 2015).

$$d(x, y) = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2}$$

2.3. Cluster Validity

The cross-validation method can be used to verify or validate the results of clustering. The data is randomly divided into two subsets, A and B, for example, and clustering is performed separately on each of A and B. The results should be similar if the groups are valid (Rencher, 2002).

2.3.1 Assumptions of Cluster Analysis

There are several assumptions or conditions that Shiraz (2015) emphasizes to ensure their fulfillment:

1. The variables require standardized measurement, especially when the measurement levels differ from one variable to another.
2. Consideration of how outliers affect the ability to classify the data.
3. Essential variables should not be neglected, as their inclusion could influence the cluster analysis.

2.4. Discriminant Function Analysis

It is a linear combination of one or more independent variables that distinguishes between two groups of observations or subjects (individuals, companies, etc.) in predefined groups. Discriminant function analysis is used to evaluate how well a classification fits by knowing the membership of sample study elements in each group, as well as predicting the classification location for new, unclassified cases (Huberty, 1994). Hardle and Simer (2003) emphasize that discriminant analysis focuses on cases where the groups are known in advance, and decision rules for classifying cases are based on the known groups. Ahmed (2015) indicates that discriminant analysis aims to classify cases into two or more groups based on a set of variables. The purpose of the analysis is to identify a specific pattern that organizes the cases, which are often individuals or objects and divides them into groups where the elements share common characteristics.

As Hair et al. (2006) indicate, discrimination is verified by calculating the weights for each independent variable to assess the variance between groups relative to the variance within the group. The discriminant function, as proposed by Fisher in 1936, is used to separate two groups and extract a classification model for the membership of new observations. It takes the following form:

$$z_{ik} = a + w_1 x_{1k} + w_2 x_{2k} + \ldots\ldots + w_i x_{ik}$$

- z_{ik} is the discriminant score,
- a is a constant,
- w_i is the discriminant weight,
- x_{ik} is the independent variable.

The number of discriminant functions is determined by the relationship k (-1), where k is the number of groups or levels of the dependent variable. For example, if the data includes a dependent variable with three levels (k =3), the discriminant analysis will yield two discriminant functions, which can be used to classify new cases (Hair et al., 2006).

El-Hanjouri and Hamad (2015) mention that Discriminant Analysis (DA) is a multivariate statistical technique used to build a predictive classification model for distinguishing and classifying each observation into one of the groups. This technique allows individuals to understand the differences between two, three, or more groups in relation to multiple variables at once. It is considered the first multivariate statistical classification method used for decades by researchers and practitioners in developing classification models. In discriminant analysis, multiple quantitative features are used to distinguish between the classification variables. It differs from cluster analysis in that it requires prior knowledge of the classification. The main

purpose of discriminant analysis is to estimate the relationship between a categorical dependent variable and a set of quantitative independent variables.

Klecka (1980) and Kardiyn and Olmus (2016) indicate that discriminant analysis requires certain assumptions to be met. These include the presence of two or more groups, with at least two cases in each group. Additionally, the independent variables must be quantitative and follow a normal distribution. Furthermore, the homogeneity of variance-covariance matrices across all groups must be satisfied.

2.5. Previous Studies

In a study conducted by Al-temimi et al. (2018), the aim was to use cluster analysis and discriminant analysis to classify displaced populations based on epidemic factors. The study relied on reliable data from government sources and international organizations, collected from the Information and Research Administration. The results showed that cluster analysis classified the displaced governorates into homogeneous groups based on camp types. Additionally, the results of discriminant analysis highlighted the contribution of the variables in distinguishing between the governorates.

In a study conducted by Huang et al. (2017), the aim was to use cluster analysis and discriminant analysis to study the relationship between volatile compounds and characteristics of fermented Chinese wheat dough. The sample consisted of seven types of dough with different flavors. The results of the cluster analysis revealed three classifications and the discriminant compounds contributing to the classification were identified.

In a study conducted by Yaqub (2017), multivariate statistical methods, including cluster analysis and discriminant analysis, were used in an applied study on several Iraqi banks to classify and differentiate the degree of homogeneity among the banks. The study was applied to a sample of 20 banks. By applying cluster analysis, the banks were classified into two homogeneous groups based on their nature and performance. Discriminant analysis was then applied to identify the key variables contributing to the heterogeneity between the groups derived from the cluster analysis. The study concluded that a set of indicators contributed more to the differentiation between the banks. A discriminant function was formulated to assess the impact of each variable and predict the membership of any bank in the appropriate group. The results from the discriminant function confirmed that the classification of the 20 banks was accurate.

In a study conducted by Rose et al. (2016), the aim was to use cluster analysis and discriminant analysis to identify distinct neuropsychological factors and their relationship with clinical symptoms in a group of children and adolescents. The sample consisted of 253 girls and adolescents diagnosed with anorexia. The results showed that cluster analysis was used to determine the optimal number of groups,

while discriminant analysis identified the neuropsychological variables that best differentiated between the groups.

A study conducted by Panagopoulos et al. (2016) aimed to use cluster analysis and discriminant analysis to classify complex groundwater systems. The proposed method involved conducting chemical analyses of water samples for key ions from 57 samples collected from wells on the island of Lesvos in Greece. The results of the cluster analysis classified the samples into four distinct groups. The discriminant analysis results confirmed the effectiveness of cluster analysis, as the classification accuracy of the samples was 94.7%.

Kardiyen and Olmus (2016) conducted a study that highlighted the implications of the problem of classifying two groups in various fields, such as medicine, economics, and finance. The study used three classification models and employed simulated samples with different distributions and sample sizes. The results of the study showed that discriminant function analysis had an advantage in classifying the data.

In a study conducted by Tanos et al., (2015), the aim was to use cluster analysis and discriminant analysis to optimize the utilization of a water quality monitoring network on the Tisza River in Hungary, considering seasonal variations. The Tisza River is the second-largest river in Central Europe. The study identified 15 water quality parameters measured at 14 sampling locations over the period from 1975 to 2005. The results of the cluster analysis revealed four classifications, and the results of the discriminant analysis showed the spatial and temporal improvement of the monitoring network using both cluster and discriminant analysis.

In a study conducted by El-Hanjouri and Hamad (2015), multivariate statistical analysis methods were applied, specifically cluster analysis (CA), to identify the disparities in household living standards across Palestinian regions. The study concluded that there was a convergence in the living standards of households in the first three groups, which represented high living standards, and a second group with medium living standards, while the third group consisted of low living standards. Discriminant analysis (DA) was then applied to distinguish the variables that significantly contributed to this disparity among households within the Palestinian regions. The results indicated that variables such as monthly income, assistance, agricultural land, livestock holdings, total expenditure, and calculated rent significantly contributed to the classification.

Ali (2015) conducted a study aimed at the effectiveness of using cluster analysis and discriminant analysis in verifying the discriminative significance of intelligence and personality tests. The study sample consisted of 610 male and female students from the secondary stage in Damascus Governorate, and the study used Raven's Progressive Matrices test and Eysenck's test. The results of the study showed the effectiveness of cluster analysis in identifying homogeneous groups based on the distances for both tests, and the study also showed the effectiveness of discriminant

analysis in distinguishing between the groups and predicting the membership of individuals in the groups generated by cluster analysis.

Jaiswara et al. (2013) conducted a study on the effectiveness of discriminant function and cluster analysis in identifying cricket species. The discriminant function analysis and cluster analysis methods were used to identify and classify cricket species based on their sound signals. The results showed that classification using discriminant analysis achieved an accuracy of 95-100% and was unaffected by the number of species used. The results also demonstrated the accuracy of classification using cluster analysis, which does not require prior knowledge. The findings indicate the effectiveness of cluster analysis in classifying and identifying species. The study also stated that discriminant analysis is considered the stronger and more accurate method, but it requires pre-classification and can only be used with known classifications. On the other hand, cluster analysis is less powerful, and its accuracy is more dependent on the number of variables examined together, but it can be used in cases where the classifications are not previously known. Both methods can be used to develop quantitative and automated tools to determine different classifications.

In a study conducted by Rashid and Mahdi (2011), the aim was to use hierarchical and non-hierarchical cluster analysis methods to analyze the reality of education in Iraq. The study included 25 variables in this field to assess the similarity in service delivery and determine the best methods used for analysis. The study concluded that the hierarchical method was the best method used, as it had the lowest value according to the reduced relationship scale.

In a predictive study conducted by Maroco et al. (2011), several predictive methods, including discriminant analysis, were used. The study utilized real data from a sample of 92 patient records to classify cases into cognitive impairment and other categories. Accuracy and sensitivity criteria were used for comparison, and the study concluded that discriminant analysis demonstrated a high degree of accuracy, sensitivity, and discriminatory power compared to other methods.

In a study conducted by Mustafa (2007), the aim was to compare cluster analysis methods on a sample of data representing a group of Maghreb countries that share common characteristics. The study concluded with a classification of these countries, and the results indicated that the outcomes of the cluster analysis methods did not differ significantly from one another.

Hwang (2001) focuses on the accuracy of group classification through discriminant analysis. The primary goal and purpose of this study is to provide an overview of how discriminant analysis works and how it can help answer various research questions. Additionally, the study explains and clarifies what discriminant analysis is and why it is important and provides illustrative data on how to use discriminant analysis. The study emphasizes the superiority of discriminant analysis in classifying observations.

In a study conducted by Al-Jaouni and Ghanem (2001), the aim was to use one of the multivariate statistical analysis methods, namely cluster analysis and discriminant analysis, to study the determination of the economic and social structure levels of households in the community. This method was applied to six economic and social variables. The study concluded by determining the levels of the economic and social structure using cluster analysis and distinguishing between the defined levels through discriminant analysis to differentiate between the two groups.

Through the presentation of previous studies, it can be observed that most of the earlier research focused on applying these methods to social phenomena, such as Al-temimi et al. (2018) in classifying displaced people, and Yaqub (2017) in classifying banks, as well as Al-Jaouni and Ghanem (2001) in economic and social classification. There are also studies focusing on scientific phenomena, such as Jaiswara et al. (2013) in determining cricket species, Tanos et al. (2015) in water quality monitoring, Al-Raid et al. (2016) in classifying water based on certain minerals, and Huang et al. (2017) in classifying volatile compounds. However, few studies, to the best of the researcher's knowledge, have employed both cluster analysis and discriminant analysis on educational variables. This current study aligns with previous studies in terms of methodology but differs in the type of data used. What distinguishes this study from previous research is its introduction of a new concept: employing both methods of multivariate statistical analysis together to classify students based on academic performance. A topic that has not been explored in educational literature to the researcher's knowledge.

3. Methodology

The study adopted the descriptive method.

3.1. Study Sample

The sample consisted of randomly selected data from 62 students out of 1,722 students at Umm Al-Qura University for the academic year 2014/2015.

3.2. Study Tools

The study relied on the data collected from the student sample.

3.3. Study Variables

Independent Variables:

X1: High school grades

X2: Aptitude test scores

X3: Achievement test scores

X4: Preparatory year GPA

Dependent Variable:

Y: A categorical variable representing cumulative GPA, taking the value (1) if the GPA is high (1.75 or above) and (2) if the GPA is low (below 1.75).

3.4. Data Analysis

The study utilized the Statistical Package for the Social Sciences (SPSS) to process the data and applied cluster analysis and discriminant analysis to answer the research questions.

4. Study Results

The K-Means cluster analysis method was applied, as it requires prior knowledge of the number of clusters used to classify the cases.

Table 1. Distribution of cases into clusters and distance of each case from the cluster center

Case no.	Cluster	Distance	Case no.	Cluster	Distance	Case no.	Cluster	Distance
1	1	12.368	23	1	6.689	45	2	3.286
2	2	1.919	24	2	7.091	46	1	3.174
3	2	8.953	25	2	4.210	47	1	5.494
4	1	10.648	26	1	2.968	48	2	6.741
5	1	14.824	27	2	7.112	49	2	4.431
6	1	5.170	28	2	5.638	50	1	4.423
7	2	6.288	29	2	2.520	51	2	9.830
8	1	7.215	30	1	6.781	52	1	10.298
9	1	3.185	31	1	10.452	53	1	3.274
10	1	2.076	32	2	6.164	54	2	4.966
11	1	5.097	33	2	6.483	55	1	4.952
12	2	7.663	34	1	8.086	56	2	10.155
13	1	5.955	35	1	6.242	57	2	10.797
14	2	5.710	36	2	10.886	58	2	5.755
15	1	6.137	37	2	12.084	59	2	11.727
16	2	6.449	38	2	6.568	60	2	5.761
17	1	19.775	39	2	2.179	61	1	6.309
18	1	4.361	40	2	4.789	62	2	4.348

19	2	4.347	41	1	5.218
20	2	6.508	42	1	5.243
21	1	4.320	43	2	2.745
22	2	15.135	44	2	4.226

Table 1 presents a list of all cases and the clusters to which each case belongs. The analysis results indicate the presence of two cluster groups, as shown in the second, fifth, and eighth columns (Cluster). The third, sixth, and ninth columns display the distance (or deviation) of each case from the cluster center. For example, Case No. (17) belongs to the first cluster and is the farthest case from its center, with 19.775. In contrast, Case No. (26) is the closest to the first cluster center, with 2.968. This means that the distance between the cases and the center of the first cluster ranges from 2.968 to 19.775, comprising 28 cases, as shown in Table 1. For Cluster 2, Case No. (2) is the closest to the second cluster center, with 1.919, while Case No. (37) is the farthest, with 12.084. This indicates that the distance between the cases and the center of the second cluster ranges from 1.919 to 12.084, including 34 cases.

Table 2. For each variable using the two cluster groups

Variable	Cluster		Error		F	Sig.
	Mean square	df	Mean square	df		
X1	12.964	1	11.572	60	1.120	.294
X2	1202.287	1	20.308	60	59.202	.000
X3	900.875	1	25.460	60	35.384	.000
X4	.147	1	.988	60	.149	.701

Table 2 presents the analysis of variance for each of the independent variables using the two cluster groups. The column "Cluster" provides the mean squares between the two groups, while the "Error" column gives the mean squares within the groups.

By examining the significance level for variables (X2) and (X3), we find that they are significant at the 0.01 level, meaning they play an important role in classifying cases into clusters. On the other hand, variables (X1) and (X4) are not statistically significant, indicating that these variables did not contribute to the classification of cases into clusters.

As observed from the cluster analysis results using the K-Means method, the classification of cases into two cluster groups was confirmed, with each cluster containing a number of similar cases.

The classification derived from the cluster analysis results was used to establish the discriminant function. Accordingly, the variables included in the discriminant analysis consist of a binary dependent variable, where (1) refers to cases in the first cluster, and (2) refers to cases in the second cluster. The independent variables are those that contributed to the classification into two clusters and were used to derive the discriminant function for classifying cases based on the dependent variable group.

Table 3. Analysis of variance table and Wilks' Lambda value for the set of means

Variable	Wilks' Lambda	F	df1	df2	Sig.
X1	.982	1.120	1	60	.294
X2	.503	59.202	1	60	.000
X3	.629	35.384	1	60	.000
X4	.998	.149	1	60	.701

It is observed after conducting the discriminant analysis and from the results of Table 3 that the independent variables (X2) and (X3) were statistically significant at the (0.01) level. This means that the differences between the means of the two groups were statistically significant. However, the independent variables (X1) and (X4) were not statistically significant. Additionally, considering the Wilks' Lambda (Wilks' Lambda) value for these variables, it is found to be higher for the non-significant independent variables and close to one (0.982) and (0.998), indicating no statistical significance.

Table 4. The eigenvalues of the discriminant function.

Function	Eigenvalue	% Variance	Cumulative %	Canonical Correlation
1	1.453(a)	100.0	100.0	.770

It is evident from Table 4 that the discriminant function corresponds to an eigenvalue of (1.453) with a canonical correlation of (0.770), indicating the strength of the relationship between the variables included in the analysis. The function explained (100%) of the variance.

Table 5. Wilks' Lambda values for the discriminant function

Discriminant Function	Wilks' Lambda	Chi-Square	df	Sig.
1	.408	52.042	4	.000

It is observed from Table 5 that Wilks' Lambda value is (0.408), and the Chi-Square value is (52.042), which is statistically significant at the (0.01) significance level. This indicates the discriminant function's ability to distinguish between the two groups.

Table 6. Standardized discriminant function coefficients

Variables	Standardized Discriminant Function Coefficients
	1
X1	-.243
X2	.733
X3	.566
X4	-.200

From Table 6, the standardized discriminant function coefficients can be used to construct the standardized discriminant function model. By estimating the discriminant coefficients, we can determine the influence of each variable on the model. The higher the value of the coefficient (whether positive or negative), the greater the contribution of that variable to the discriminant function. Therefore, the standardized discriminant function can be expressed as follows: $y = -0.243x_1 + 0.733x_2 + 0.566x_3 - 0.20x_4$

Table 7. Classification accuracy results

		Cluster Number	Predicted Group Membership		Total
			1	2	
Original	Count	1	28	0	28
		2	1	33	34
	%	1	100.0	.0	100.0
		2	2.9	97.1	100.0
Cross-Validation	Count	1	28	0	28
		2	2	32	34
	%	1	100.0	.0	100.0
		2	5.9	94.1	100.0

98.4% of the original grouped cases were classified correctly.

After deriving the discriminant function, the classification of cases is verified to check if they are correctly assigned to the cluster they were classified into. In other words, is the classification accurate? The results from Table 7 show that when comparing the classification of cases based on the cluster analysis, the classification was correct with an accuracy rate of 98.4%, which is a very high percentage, confirming the accuracy of the classification.

5. Discussion

The K-Means clustering method was applied because this technique requires prior knowledge of the number of clusters, as indicated by Ahmed (2015, p. 54) in using this method for determining the number of clusters. This method relies on Euclidean distance to determine the proximity of a case to the cluster center. Table 1 shows a list of all cases and the clusters to which each case belongs. The results of the analysis revealed two clusters. In the first cluster, the distance between the cases and the cluster center ranged from 2.968 to 19.775, comprising 28 cases. In the second cluster, the distance between the cases and the cluster center ranged from 1.919 to 12.084, comprising 34 cases. This finding aligns with the studies of Al-temimi et al. (2018), Huang et al. (2017), Yaqub (2017), and Rose et al. (2016), which used clustering analysis to identify cases based on their proximity or distance from each cluster .

Moreover, as shown in Table 2, the significance of the independent variables using the two clusters is evident. It can be observed that the significance level of variables (x2) and (x3) was statistically significant at the 0.01 level, indicating that they play an important role in classifying cases into clusters. On the other hand, variables (x1) and (x4) were not statistically significant, meaning these variables did not contribute to classifying cases into clusters.

The classification obtained from the cluster analysis results was relied upon to find the discriminant functions, which determine the contribution of each variable to the classification process. The variables included in the discriminant analysis consisted of a binary dependent variable, where (1) refers to cases classified in the first cluster, (2) refers to the classification of cases in the second cluster and a set of independent variables through which the classification into two clusters was made. These were used to find the discriminant function that classifies the cases according to the dependent variable group.

It was found after performing the discriminant analysis and from the results of Table 3 that the independent variables (X2) and (X3) were statistically significant at the (0.01) level, while the independent variables (X1) and (X4) were not statistically significant. Additionally, the Wilks' Lambda value for these variables was higher for the non-significant independent variables and close to one (0.982, 0.998), which indicates no statistical significance. This agrees with the study of Jaiswara, et al. (2013), which emphasized the need to use both methods to develop quantitative and automated tools to determine different classifications. It also aligns with the study of

Al-Jaouni and Ghanem (2001), which used cluster analysis for classification and discrimination between the two groups through discriminant analysis.

As shown in Table 4, the discriminant function has a canonical correlation of (0.770), which indicates the strength of the relationship between the variables included in the analysis, corresponding to an eigenvalue of (1.453), explaining (100%) of the variance. It is also noted from Table 5 that the Wilks' Lambda value was (0.408), and the chi-square value was (52.042), which is statistically significant at the (0.01) level, indicating the discriminant function's ability to differentiate between the two groups.

From Table 6, the standardized discriminant function was formulated by estimating the standardized discriminant coefficients, which show the extent to which the variables influence the model. Therefore, the standardized discriminant function can be formulated as follows:

$$y = -0.243x_1 + 0.733x_2 + 0.566x_3 - 0.20x_4$$

The study, after obtaining the discriminant function, verified the classification of cases to check if they indeed fell within the cluster they were classified into. The results from Table 7 show that when comparing the classification of the cases based on cluster analysis, the classification was correct with a high percentage of (98.4%), confirming the accuracy of the classification. This result aligns with the findings of the study by Panagopoulos et al. (2016), which confirmed the usefulness of discriminant analysis. In their study, the correct classification rate of the samples was (94.7%).

These results also align with the studies of Huang, et al. (2017) and Al-temimi (2018), where cluster analysis was used to classify cases, and discriminant analysis was employed to determine the discriminating variables contributing to the classification. Similarly, the study by Yaqub (2017) found that cluster analysis led to the classification of cases into two homogeneous groups based on work nature and performance, and discriminant analysis was applied to identify the key variables that led to heterogeneity among the groups obtained from the cluster analysis.

Additionally, these findings are consistent with the study by Rose et al. (2016), which concluded that using cluster analysis to determine the number of clusters, followed by discriminant analysis to identify the variables that best differentiate between the clusters, which is an effective approach.

6. Recommendations

The study recommends

1. Applying cluster analysis to classify cases into homogeneous or closely related clusters within each group.
2. Utilizing discriminant analysis to identify the variables that contributed to the classification and determine the discriminant functions.

3. Incorporating advanced statistical methods in classifying cases into clusters to enhance the precision and reliability of the classification process.
4. Conducting studies to compare the methods used for calculating distances in cluster analysis and assess their effectiveness and efficiency.

References

[1] Ahmed, T. (2015). Classification of Syrian governorates based on household consumption using cluster analysis. *Tishreen University Journal for Research and Scientific Studies, 37*(2).

[2] Ali, K. A. (2015). *The effectiveness of using cluster analysis and discriminant analysis in verifying the discriminatory significance of intelligence and personality tests: A comparative field study in Damascus Governorate* (Unpublished master's thesis). Faculty of Education, Damascus University.

[3] Al-Jaouni, F., & Ghanem, A. (2001). Multivariate statistical analysis (Cluster analysis) in the study of determining the levels of the economic and social structure of family groups in society. *Damascus University Journal, 17*(2).

[4] Al-Shafi, M. M. (2014). *Traditional and advanced statistics in scientific and human research (Book II)*. Riyadh: Al-Rushd Library.

[5] Al-temimi, S., Al-Saffar, R., & Shebib, H. (2018). Classification and identification of IDP camps after Mosul events based on epidemics and other factors using cluster analysis and discriminant analysis. *International Journal of Pharmaceutical Research & Allied Sciences, 7*(3), 63–73.

[6] El-Hanjouri, M., & Hamad, B. (2015). Using cluster analysis and discriminant analysis methods in classification with application on standard of living family in Palestinian areas. *International Journal of Statistics and Applications, 5*(5), 213–222. https://doi.org/10.5923/j.statistics.20150505.05

[7] Everitt, B., Landau, S., Leese, M., & Stahi, D. (2011). *Cluster analysis* (5th ed.). Wiley Series.

[8] Hair, J., Black, W., Babin, B., Anderson, R., & Tatham, R. (2006). *Multivariate data analysis* (6th ed.). Pearson Education.

[9] Hardle, W., & Simar, L. (2003). *Applied multivariate statistical analysis*. Springer.

[10] Huang, M., Li, Y., Zhan, P., Liu, P., Tian, H., & Fan, J. (2017). Correlation of volatile compounds and sensory attributes of Chinese traditional sweet fermented flour pastes using hierarchical cluster analysis and partial least squares-discriminant analysis. *Journal of Chemistry*, ID 3213492. https://doi.org/10.1155/2017/3213492

[11] Huberty, C. (1994). *Applied discriminant analysis*. John Wiley & Sons.

[12] Hwang, D. (2001). Issues in predictive discriminant analysis: Using and interpreting the leave-one-out jackknife method and the improvement-over-change "I" index effect size. Paper presented at the *Annual Meeting of the Southwest Educational Research Association*.

[13] Jaiswara, R., Nandi, D., & Balakrishnan, R. (2013). Examining the effectiveness of discriminant function analysis and cluster analysis in species identification of male field crickets based on their calling songs. *PLoS ONE, 8*(9), e75930. https://doi.org/10.1371/journal.pone.0075930

[14] Kardiyen, F., & Olmus, H. (2016). A comparison of two group classification approaches to fat-tailed and skewed data. *Communication in Statistics-Simulation and Computation, 45*, 17–32. https://doi.org/10.1080/03610918.2015.1074916

[15] King, R. (2015). *Cluster analysis and data mining*. Dulles: Mercury Learning and Information.

[16] Klecka, W. (1980). *Discriminant analysis*. California: Sage Publications.

[17] Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, L., & Mendonça, A. (2011). Data mining methods in the prediction of dementia: A real–data comparison of the accuracy, sensitivity, and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees, and random forests. *BMC Research Notes, 4*, 299. https://doi.org/10.1186/1756-0500-4-299

[18] Mustafa, N. (2007). The use of some cluster analysis methods in classification with practical application. *Al-Taqani Journal, 20*(2).

[19] Panagopoulos, G., Angelopoulou, D., Tzirtzilakis, E., & Giannoulopoulos, P. (2016). The contribution of cluster and discriminant analysis to the classification of complex aquifer systems. *Environmental Monitoring and Assessment, 188*(591). https://doi.org/10.1007/s10661-016-5590

[20] Ramdeen, K., & Yim, O. (2015). Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. *The Quantitative Methods for Psychology, 11*(1). https://doi.org/10.20982/tqmp.11.1.p008

[21] Rashid, A., & Mahdi, N. (2011). Analysis of the reality of education in Iraq using cluster analysis methods (Comparative study). *Al-Qadisiyah Journal for Administrative and Economic Sciences, 13*(2).

[22] Rencher, A. (2002). *Methods of multivariate analysis* (2nd ed.). Canada: Wiley-Interscience.

[23] Romesburg, H. (2004). *Cluster analysis for researchers*. Lulu Press.

[24] Rose, M., Stedal, K., Reville, M., Noort, B., Kappel, V., Frampton, L., Watkins, B., & Lask, B. (2016). Similarities and differences of neuropsychological profiles in children and adolescents with anorexia nervosa and healthy controls using cluster and discriminant function analyses. *Archives of Clinical Neuropsychology, 31*, 877–895. https://doi.org/10.1093/arclin/acw081

[25] Shiraz, M. S. (2015). *Statistical data analysis using SPSS*. Jeddah: Khawarizmi Scientific.

[26] Tanos, P., Kovács, J., Kovács, S., Anda, A., & Hatvani, I. (2015). Optimization of the monitoring network on the River Tisza (Central Europe, Hungary) using combined cluster and discriminant analysis, taking seasonality into account. *Environmental Monitoring and Assessment, 187*, 4777. https://doi.org/10.1007/s10661-015-4777

[27] Wilson, L., & Hardgrave, C. (1995). Predicting graduate student success in an MBA program: Regression versus classification. *Educational and Psychological Measurement, 55*(2), 186–195.

[28] Yaqub, A. A. (2017). Cluster and discriminant analysis in an applied study on some Iraqi banks. *Gulf Economic Journal*, Issue 31.