Al-Qadisiyah Journal for Administrative and Economic Sciences ISSN$_{Online}$: 2312-9883
QJAE, Volume 27, Issue 2 (2025) ISSN$_{Print}$: 1816-9171
DOI: https:// doi.org/ 10.33916/ qjae.2025 .028392

Cluster Analysis and Logistic Regression for Identifying Principal Variables and Predicting Type 2 Diabetes

Abdulsalam Idres Abdulkarim Wisam Wadalluh Saleem
abdulsalam.23csp145@student.uomosul.edu.iq Wisam-stat@uomosul.edu.iq
University of Mosul

Article history:
Received: 20/2/2025
Accepted: 3/3/2025
Available online: 15 /6 /2025

Corresponding Author : Abdulsalam Idres Abdulkarim Wisam Wadalluh Saleem

Abstract: In this paper study focused on analyzing type 2 diabetes data from 237 individuals, both diabetic and non-diabetic, collected from the Community Health Department/ Nineveh Health Directorate. Hierarchical Cluster Analysis (HCA) was used to classify the key factors influencing diabetes using Complete Linkage to measure distances between clusters. The distance matrix was then calculated using Gower Distance due to the presence of both quantitative and qualitative variables. The best regression equation for each cluster's variables was predicted by applying logistic regression to identify the most influential variables and predict the risk of diabetes with high accuracy. The analysis was performed using Python 3, and issues such as multicollinearity were checked using the correlation matrix and the Variance Inflation Factor (VIF). The cluster analysis results identified three main groups of variables, and these three clusters were used as indicators in the logistic regression model to assess the impact of each group on the likelihood of developing the disease. The results showed that the cluster related to blood glucose levels (HbA1C and blood glucose levels) was the most statistically significant factor in predicting diabetes, while metabolic, health, demographic, and behavioral factors did not show a strong association. Diabetes risk was predicted based on the previous two variables using logistic regression to explain and estimate this relationship. The results suggest that combining cluster analysis with logistic regression enhances the model's ability to predict the influential factors more accurately, which contributes to improving decision-making and providing deeper insights into the studied data.

Keywords: type 2 diabetes, Hierarchical Cluster Analysis, Complete Linkage, logistic regression model, HbA1C and blood glucose levels.

INTRODUCTION: Type 2 diabetes is one of the most widespread chronic diseases globally, causing serious health issues that affect patients' quality of life and increase the risk of heart disease, high blood pressure, and other metabolic complications. In recent years, there has been an increasing need for advanced data analysis methods to understand the factors influencing diabetes and develop more accurate predictive models. The researcher [5] conducted a study on analyzing different patterns of diabetes and the associated risk factors using cluster analysis. The study utilized data from large population-based studies to identify different patterns of the disease and distinguish between various types of diabetes based on risk factors. The researchers [8] conducted a study aimed at identifying different subgroups of elderly patients with diabetes and multiple comorbidities. The results showed that elderly patients with diabetes alone had better health conditions, while those with diabetes and multiple comorbidities experienced more health issues, particularly in areas such as depression and diabetes-related distress. The researchers [13] conducted a study that introduced a new classification of diabetes based on clusters, which could provide a potential solution for early prevention and treatment of type 2 diabetes—a major challenge for both patients and physicians. The risk of diabetes complications and comorbidities in each subtype was compared using logistic regression analysis. The study successfully classified newly diagnosed type 2 diabetes patients into four distinct subtypes, each with different clinical characteristics, medication treatments, and varying risks of diabetes-related complications and comorbidities. Hierarchical Clustering Analysis (HCA) was used, which allows grouping similar variables into clusters, helping to reduce complexity and improve data interpretation. Additionally, logistic regression is a powerful tool for understanding the relationship between different factors and predicting the likelihood of developing the disease.

Study Problem:

The study aims to analyze and classify the factors affecting type 2 diabetes by handling a complex dataset containing both qualitative and quantitative variables. Hierarchical clustering analysis will be used to group individuals based on shared patterns, which will help identify the most at-risk categories. Subsequently, a logistic regression model will be developed to predict the risk of disease based on these factors, thereby improving disease understanding and enhancing prevention and early diagnosis strategies.

Study Objectives:

This research aims to:

- Predicting the most influential factors affecting diabetes by identifying the independent variables that impact the incidence of Type 2 diabetes.
- Using Hierarchical Cluster Analysis (HCA) to partition the data into clusters representing groups with similar characteristics and influencing factors.
- Integrating the results of cluster analysis with logistic regression to analyze the impact of different factors on the target variables and predict the most influential factors.
- Providing an improved logistic regression model using the derived clusters, enabling accurate predictions of the dependent variables.

Study significance and contributions:

1. The significance of the study lies in predicting the key independent variables influencing the onset of type 2 diabetes, which helps improve the understanding of the factors that increase the risk of the disease and contributes to making health decisions based on accurate data.
2. Predicting a regression model that includes the key variables influencing the onset of diabetes.
3. Stating the importance of each variable and its impact on the onset of diabetes.

1. Diabetes Mellitus:

Diabetes is a group of metabolic diseases characterized by high blood sugar levels due to defects in insulin secretion, insulin function, or both. Diabetes is a serious global health issue and one of the major chronic diseases affecting humans. The full term "Diabetes Mellitus" originates from the Greek words "Syphon" and "Sugar." The problem arises when there is a deficiency in insulin production and/or an inability of insulin to exert its proper effect, leading to impaired glucose utilization, which increases blood glucose levels. Since glucose is essential for cellular metabolism, it needs to be transported into cells. The pancreas produces insulin, a hormone secreted by beta cells (B cells), which lowers blood glucose levels by transporting glucose from the bloodstream into cells for energy use. [9] [11]

2. Types of Diabetes:

Diabetes is not a single condition but consists of multiple major types, including: [4] [11]

1. **Type 1 Diabetes (T1D):** This type, previously known as insulin-dependent diabetes or juvenile diabetes, is characterized by insufficient or abnormal insulin secretion from pancreatic beta cells, leading to minimal or no insulin production. Patients require daily insulin injections to prevent coma or death. Symptoms include excessive urination, persistent thirst and hunger, weight loss, vision changes, and fatigue. The exact cause of Type 1 diabetes is unknown, and prevention is not possible with current medical knowledge.
2. **Type 2 Diabetes (T2D):** This type, formerly known as non-insulin-dependent diabetes or adult-onset diabetes, occurs due to the body's ineffective use of insulin, often resulting from obesity and physical inactivity. It accounts for about **95%** of all diabetes cases. Symptoms may be similar to those of Type 1 but are often less pronounced, leading to late diagnosis after complications have already developed. Type 2 diabetes primarily affects adults but can also occur in younger individuals. Treatment involves blood sugar-lowering medications and insulin production enhancers to optimize glucose utilization.

3. Cluster Analysis:

Cluster analysis classifies observations into unnamed groups based on specific variable patterns. These methods group objects or variables under study into homogeneous clusters while distinguishing them from other groups. The primary goal of this analysis is to discover patterns that organize observations into clusters with shared properties, making it easier to predict the behavior or characteristics of new objects based on their assigned clusters. Cluster analysis has proven successful in many fields, including public health, medicine, and marketing. [7] [11]

4. Data Standardization:

Distance measurement values are closely linked to the scale of measurement used. Therefore, it is common practice to standardize variables before measuring differences between observations, especially when variables are measured on different scales (e.g., kilometers, kilograms, centimeters). Standardization ensures comparability by transforming

variables so that their mean is **zero** and their standard deviation is **one**. This approach is widely used in **gene expression data analysis** before clustering. [3]
The value of distance measurements is closely related to the scale on which the measurements are taken. Therefore, variables are often standardized (i.e., their units of measurement are unified) before measuring the differences between observations. This procedure is particularly recommended when variables are measured on different scales, for example (kilometers, kilograms, centimeters, ...). Otherwise, the obtained measures of variation will be heavily influenced. [3]
The goal is to make the variables comparable. Generally, variables are scaled so that the **standard deviation equals one** and the **mean equals zero**. Standardizing data is a widely used approach in the context of **gene expression data analysis** before clustering. We may also want to scale data when the mean and/or standard deviation of the variables differ significantly. [7]
If the units of measurement for (X) are different (such as income, number of family members, housing area, etc.), we transform these variables into standardized (Z) variables using the following relationship:

$$Z_i = \frac{x_i - \bar{x}_i}{\sigma i} \qquad \dots \quad (1)$$

Thus, we obtain the standardized variables: $(Z_1, Z_{2,} \dots, Z_p)$, which are characterized by having a mean of zero ($\bar{z}_i = 0$) and a variance of one ($\sigma_i^2 = 1$). We then work with these variables. [3]

5. Components of Cluster Analysis: [11]

1. Cluster: A group of relatively homogeneous cases or observations. The elements within a single cluster are similar to each other, while elements from different clusters are less homogeneous.
2. Element: Numerical values of measurable quantities, referred to as attributes.
3. Distance: The space or gap between two elements. The relationship between similarity and distance is inverse.
4. Graphical Tree (Dendrograms): The hierarchical structure resulting from the clustering process.

Hierarchical Cluster Analysis (HCA):
Hierarchical clustering can be performed using two main methods: [1]
1. **Divisive Method:** Begins with a single large cluster that is progressively divided into smaller clusters.
2. **Agglomerative Method:** Starts with individual points as separate clusters and merges them based on similarity until one large cluster forms.

Both methods use dendrograms to illustrate clustering results, where each node represents an observation, and branches depict the merging process.
The clustering or agglomeration method is the opposite of the partitioning method, as the process starts from the core of the clusters and progresses to the formation of the final cluster tree based on the degree of similarity between the elements. The method begins by merging the most similar observations or those that are closest in distance, then gradually proceeds with the merging process until it can stop when the distances between the clusters exceed a predefined value (d_0), known as the Arbitrary Threshold Level, or when a sudden jump in distances occurs. It is also assumed in this method that each element initially represents a separate subset, and then the similar subsets are gradually grouped into a comprehensive set that includes all the data. [1]
Initially, each observation is considered a separate cluster representing a leaf. Subsequently, the most similar clusters are merged iteratively until a single large cluster is formed, representing the root. Agglomerative clustering operates in a bottom-up manner, where each element starts as an independent cluster consisting of a single leaf. At each step of the algorithm, the two most similar clusters are merged to form a larger cluster (nodes). This process continues until all points are merged into one large cluster representing the root. As shown in Figure (1). [7]

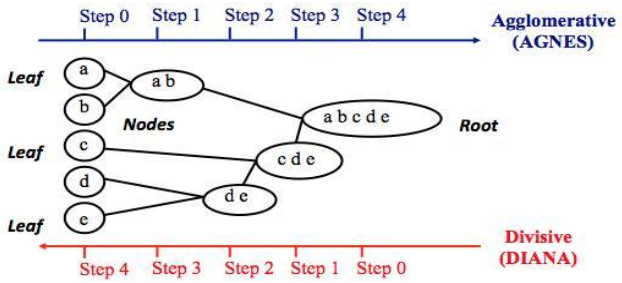


Figure (1) Hierarchical Agglomerative and Divisive Clustering.

To calculate the distance between subgroups, the Complete Linkage method was used, where the group components depend on the maximum distance between them (also known as the Farthest Neighbor Rule) according to the following formula:

$$D_{IJ} = Max\left(d_{ij}\right) \quad \ldots \quad (2)$$

Where i, j represents the elements in clusters i, j respectively.

The following methods are used to measure the quality of clusters and perform cluster analysis: similarity measure, dendrograms, elbow method, and Dunn index. [2] [3]

6. Distance Matrix

The first step in conducting cluster analysis is calculating the Distance Matrix. This matrix is symmetrical, where the number of rows equals the number of columns. The rows and columns represent the elements for which the distance is to be measured, while its elements d_{ij} indicate the measured distance between any two of these elements. Cluster analysis typically begins by constructing this matrix, which serves as one of the distance measures between observations. The core idea is to group similar units into separate clusters. The general form of this matrix can be represented as follows: [2]

$$\mathrm{D} = d_{ij} = \begin{bmatrix} d_{11} & d_{12} \ldots\ldots & d_{1n} \\ d_{21} & d_{22} \ldots\ldots & d_{2n} \\ d_{n1}^{\vdots} & d_{n2}^{\vdots} \ldots\ldots & d_{nn}^{\vdots} \end{bmatrix}$$

d_{ij} : The distance measured between any pair of elements

7. Gower Distance:

Gower Distance is a similarity measure used to calculate the distance between data points that contain different types of variables, such as:

1. Numerical variables: For example, age and salary.
2. Categorical variables: For example, gender or car color.
3. Binary variables: For example, (yes/no) or (true/false).

Gower Distance is calculated by aggregating the partial distances for each variable in the data. Each type of variable is measured using an appropriate metric, and the values are then normalized to fall between 0 and 1. Finally, the average of all partial distances is computed to obtain the total distance between two points. [8]

8. Logistic Regression:

Logistic regression is used in medical studies to analyze categorical dependent variables, such as whether a patient has diabetes (Yes/No). Unlike linear regression, which is designed for continuous dependent variables, logistic regression models the probability of an event occurring using the logit function: [6]

$$p(Y=1) = \frac{e^{(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}}{1+e^{(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_nX_n)}} \quad \ldots (3)$$

where p is the probability of the event occurring. The logistic regression model transforms probabilities into odds ratios, making it suitable for binary outcomes. **[12]**

Transforming the formula into a linear relationship using the natural logarithm (Logit Function).

$$\log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_nX_n \quad \ldots (4)$$

where the left-hand side represents what is known as the natural logarithm of the odds (Log Odds). The values of (β) beta are estimated using Maximum Likelihood Estimation (MLE).

Logistic regression is used to predict the probability of a specific event occurring by fitting the data to a logistic curve. Thus, it is a generalized linear model that takes the form of a logistic function, as shown in Figure (2). Consequently, logistic regression determines the estimation parameters that maximize the likelihood of the event occurring (presence of a distinctive characteristic), unlike linear regression, which determines the parameters that minimize the sum of squared errors. [6] [10]

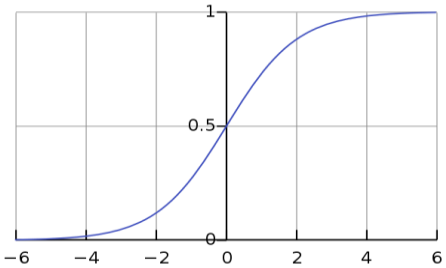


Figure (2): The Logistic Function

9. Model Evaluation Metrics: [6]

1. **Nagelkerke's R^2 & Cox & Snell's R^2:** These are alternative measures to R^2 in linear regression, used to determine the model's goodness-of-fit.
2. **Hosmer-Lemeshow Test:** Evaluates whether the model represents the data well by comparing observed and expected values.
3. **Wald Statistic Test:** Determines the statistical significance of each independent variable's effect on the dependent variable.

10. Practical Application:

The applied aspect included three main components: the first involved collecting data and the variables included in the study, the second focused on the statistical analysis of the variables using cluster analysis, and the third component involved analyzing the data using logistic regression.

Hierarchical cluster analysis (HCA) and logistic regression were applied to analyze factors affecting Type 2 diabetes. The study data were obtained from a hospital in 2024, with a sample size of **237 individuals**. The dataset included:

- **Dependent variable:** Diabetes diagnosis (0 = Non-diabetic, 1 = Diabetic)
- **Independent variables:** Multiple demographic and clinical factors

By integrating **HCA** and **logistic regression**, the study aimed to enhance predictive accuracy and identify the most significant risk factors for Type 2 diabetes.

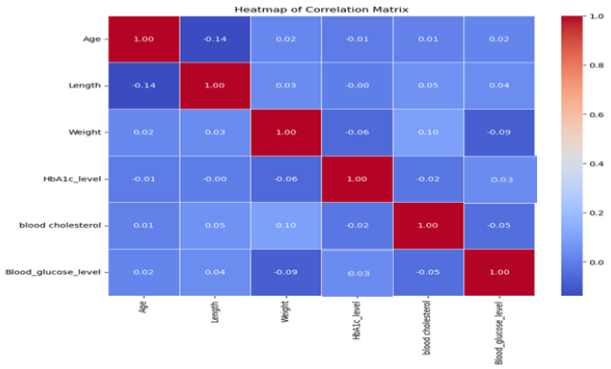
Statistical Analysis Steps:

- The analysis was conducted using Python 3. Before performing cluster analysis on the data, it was necessary to test the correlation between variables. The correlation matrix was calculated, along with the heatmap of the correlation matrix, assess the potential relationships between variables.

	Age	Length	Weight	HbA1c_level	blood cholesterol	Blood_glucose_level
Age	1.000000	-0.137198	0.018322	-0.010927	0.005394	0.015861
Length	-0.137198	1.000000	0.033101	-0.001332	0.051887	0.037611
Weight	0.018322	0.033101	1.000000	-0.060978	0.097083	-0.087553
HbA1c_level	-0.010927	-0.001332	-0.060978	1.000000	-0.022815	0.033101
blood cholesterol	0.005394	0.051887	0.097083	-0.022815	1.000000	-0.046272
Blood_glucose_level	0.015861	0.037611	-0.087553	0.033101	-0.046272	1.000000

From the correlation matrix above, it is evident that the correlations are very weak, indicating the absence of strong linear relationships between the variables.

Through the heatmap of the correlation matrix, no strong correlation between the variables was observed that would



affect the use of cluster analysis. Therefore, the second step is to perform cluster analysis.

To verify the presence or absence of multicollinearity issues, the Variance Inflation Factor (VIF) test was used to assess multicollinearity for each quantitative variable, as shown in the following Table (1):

Table (1): Values of the Variance Inflation Factor (VIF)

Variables	VIF
Age	1.035
Length	1.613
Weight	1.124

HbA1c_level	1.102
Blood_glucose_level	1.235

From Table (1), we observe that the values of the Variance Inflation Factor (VIF) are less than 5, indicating the absence of multicollinearity issues among the variables, If the value is greater than or equal to 10, it indicates a severe multicollinearity problem.

- After confirming the absence of correlation and multicollinearity issues among the variables in the previous steps, hierarchical cluster analysis was applied using Gower distance to measure the distances between clusters, and complete linkage was used to determine the distances between clusters. The results of the clustering were represented using a dendrograms in Figure (3), and the clustering results are shown in Table (2) as follows:

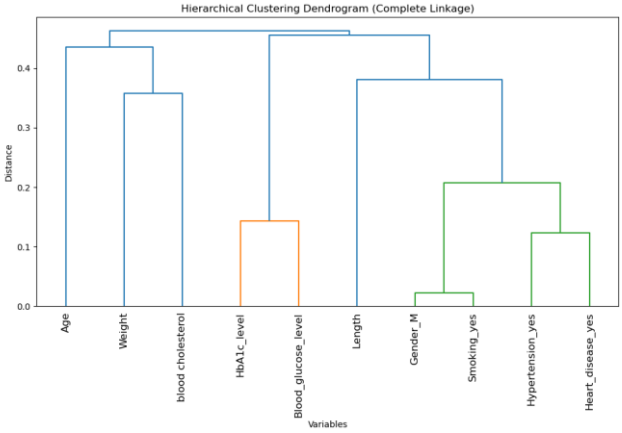


Figure (3): Dendrograms of Clusters

Table (2): Results of Cluster Analysis for the Variables

Step	Number of clusters	Similarity level	Distance level	Clusters joined		New cluster	Number of obs. in new cluster
1	9	98.9305	0.02139	1	5	1	2
2	8	95.2760	0.09448	3	4	3	2
3	7	95.0199	0.09960	8	10	8	2
4	6	54.8603	0.90279	1	3	1	4
5	5	51.7528	0.96494	7	9	7	2
6	4	46.4421	1.07116	1	6	1	5
7	3	20.6687	1.58663	2	7	2	3
8	2	8.3554	1.83289	1	8	1	7
9	1	1.4171	1.97166	1	2	1	10

Table (2) illustrates the clustering steps for the diabetes variables, showing similarity levels and distances. The highest similarity level (98.9305) was observed between clusters (1 and 5), indicating that these clusters are highly similar, which corresponds to the lowest distance level (0.02139), meaning that the merging of these clusters is greater. The "Clusters joined" column represents the cluster numbers that were merged in that step, while "New cluster" indicates the number of the new cluster formed after the merge. Finally, "Number of obs. in new cluster" represents the number of observations in the new cluster.

11. Interpretation of Cluster Analysis Results

Through the dendrogram in Figure (3) and Table (2), the results of the cluster analysis revealed three main groups of variables (three clusters):

1. Cluster 1 (Metabolic and Health Factors):

This cluster includes age, weight, and blood cholesterol levels, which represent metabolic factors that may influence blood glucose levels.

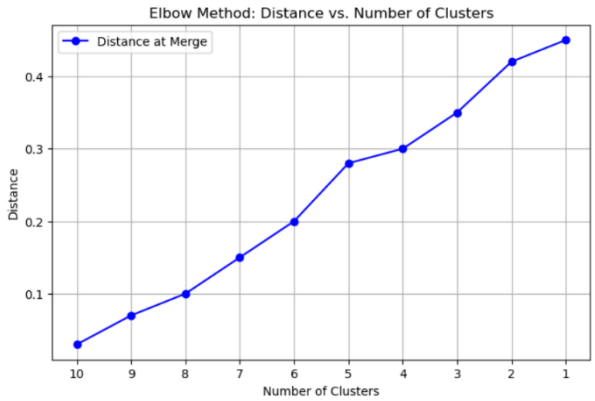
2. Cluster 2 (Diabetes Indicators):

This cluster comprises HbA1C levels and blood glucose levels, as these factors are among the most direct indicators of diabetes.

3. Cluster 3 (Demographic and Behavioral Factors):

This cluster includes gender, smoking, hypertension, and heart disease, highlighting the role of behavioral and demographic factors in increasing the risk of diabetes.
These clusters provide insights into the relationships between different variables and their potential impact on health outcomes, particularly in the context of diabetes risk.
A set of criteria was used to determine the optimal selection of the three clusters, including the Silhouette Score (0.8823). Since this value is very close to 1, it indicates that the clusters are excellent and well-defined. Additionally, the elbow method was used to determine the optimal number of clusters in the cluster analysis. From Figure (4), it appears that the elbow point is at three clusters, meaning:

Figure (4) Elbow Method



As the number of clusters increases from 1 to 3, there is a significant decrease in distance, indicating that dividing the data into 3 clusters greatly improves data grouping.
The Dunn Index value obtained was 2.52, which indicates that the clusters are cohesive, meaning the distances between variables within each cluster are relatively small. This implies that the variables within each cluster are highly similar. Furthermore, the clusters are well-separated, meaning the distances between different clusters are large, indicating that the clusters are clearly distinct from one another. This value suggests that the clusters have excellent quality and that the cluster analysis successfully grouped the variables accurately and efficiently.

- After classifying the variables into clusters, the three clusters were used as indicators in the logistic regression model to predict the presence or absence of diabetes. The analysis also examined the significant and non-significant clusters and variables. The results of the analysis are presented in Table (3) below.

Table (3) Logistic Regression Model Analysis Results

Pseudo R-squ	Log-Likelihood	LL-Null	LLR p-value	BIC
0.8310	-27.7340	-164.1000	7.879E-59	77.3411

	Const	Cluster_1	Cluster_2	Cluster_3	AIC
P>|z|	0.0850	0.6780	0.0000	0.6620	63.4689
Coff	1.0109	-0.2662	5.3883	-0.5970	

From the results of Table (3), it is evident that only Cluster 2 has a significant effect. Therefore, the model will be rebuilt using only Cluster 2, as it is the only explanatory variable with statistical significance in this model and has a strong positive impact on the probability of diabetes. The insignificant variables (Cluster 1 and Cluster 3) will be removed, and a new model will be constructed to determine whether AIC and BIC improve. The performance of the new model will be evaluated using AIC, BIC, and Pseudo R-squared. The results of the logistic regression for Cluster 2, after confirming the significance of its variables, are presented in Table (4) below.

Table (4) Logistic Regression for Cluster 2

Pseudo R-squ	Log-Likelihood	LL-Null	LLR p-value
0.8598	-27.9380	-164.1000	3.519E-61

	Const	Cluster_2	AIC	BIC
P>|z|	0.0480	0.0000	59.8765	66.8126

From the results of Table (4), the coefficient for Cluster 2 is 5.3856, which is highly significant ($P < 0.001$), indicating that this cluster, which includes the variables HbA1c_level and Blood_glucose_level, has a strong impact on diabetes risk. Pseudo R-squared = 0.8598, suggesting that the model explains a large proportion of the variance in the data, showing an improvement compared to the previous model in Table (3). AIC = 59.88 and BIC = 66.81, both relatively low compared to the previous model in Table (3), indicating that this model is superior in terms of simplicity and quality.

After confirming that Cluster 2 is the most significant, all variables from the three clusters were included in the logistic regression model to determine the significance of each variable. Table (5) presents the analysis results.

Table (5) Significance and Non-significance of Cluster Variables

	Pseudo R-squ		Log-Likelihood		LL-Null	LLR p-value				
	0.9285		-11.737		-164.1	2.58E-60				
	Const	Age	Weight	blood cholesterol	HbA1c_level	Blood_glucose_level	Gender	Smoking	Hypertension	Heart_disease
P>|z|	0.28	0.7	0.49	0.98	0.0070	0.0090	0.99	0.966	0.98	0.62

From the results of Table (5), the Pseudo R-squared = 0.9285 indicates that the model explains a large proportion of the variance. Additionally, the values of the variables (Blood_glucose_level = 0.009, HbA1c_level = 0.007) are both smaller than ($P > |z|$) = 0.05, meaning they have a statistically significant effect on diabetes risk. Therefore, only these variables were used in constructing the logistic regression equation to predict the likelihood of diabetes.

A logistic regression model will be built based on the two variables (Blood_glucose_level and HbA1c_level), as they have a significant impact on diabetes risk. Table (6) below presents the results of the analysis based on the variables from Cluster 2 only.

Table (6) Analysis Results Based on Cluster 2 Variables Only

	Const	HbA1c_level	Blood_glucose_level
coef	-3.0585	5.0708	2.6716
P>|z|	0.0049	0.006	0.005
S.E	1.34	1.466	0.82
Wald	5.21	11.98	10.64
R^2 Nagelkerke = 0.80		Hosmer-Lemeshow Chi-square= 5.8366	
R^2 Cox & Snell = 0.85		P-value = 0.9442	

From the results in Table (6), it is evident that Blood_glucose_level and HbA1c_level have a statistically significant and strong impact on the likelihood of disease occurrence, indicating that they are the best variables for predicting health status. Based on the Wald values, all variables are statistically significant. Additionally, the P-value = 0.9442 is much higher than 0.05, which means that the model fits the data well, as indicated by the Hosmer-Lemeshow Chi-square = 5.8366. Furthermore, the values of Nagelkerke $R^2 = 0.80$ and Cox & Snell $R^2 = 0.85$ show that the updated model explains between 80% to 85% of the variance in the data. The graphs in Figure (5) illustrate the impact of Cluster 2 with its variables on health status.

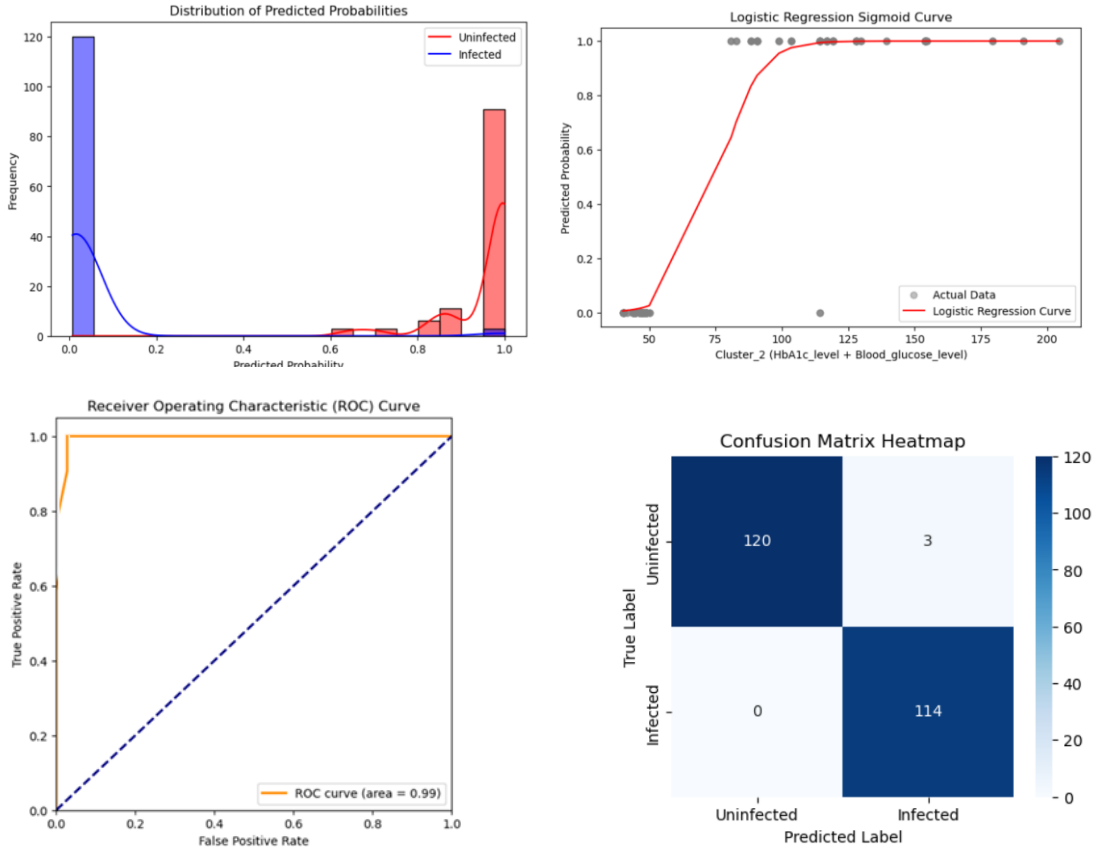


Figure (5) illustrates the effect of Cluster 2 and its variables on health status

From Figure (5)

The distribution of the predicted probability values for the model, where the model predicts diabetes in blue and non-diabetes in red.

From the logistic regression curve, we observe the following:

- When the values of Cluster (2) are less than 60 (approximately), the probability of diabetes is close to zero.
- When the values of Cluster (2) are between 70 and 120 (approximately), a sharp transition occurs in the probability, meaning that individuals with higher levels of HbA1c and glucose are more likely to develop diabetes.
- When the values of Cluster (2) are greater than 120 (approximately), the probability of diabetes is close to 1, indicating that these individuals have a very high likelihood of developing the disease.

The False Positive Rate curve (AUC = 0.99) indicates that the model has a very high discriminatory ability between diabetic and non-diabetic individuals. The AUC value is very close to 1, which means the model almost achieves perfect classification. The closer the curve is to the top-left corner, the better the model's performance.

From the confusion matrix, True Negative (TN) = 120, it shows that the value indicates the correctly classified non-diabetic cases as non-diabetic. These values reflect the model's accuracy in correctly excluding non-diabetic individuals. False Positive (FP) =3 The cases of non-diabetic individuals that were incorrectly classified as diabetic are very few, indicating that the model has a very low error rate in classifying non-diabetic individuals. False Negative (FN)=0, The cases of diabetic individuals that were incorrectly classified as non-diabetic (FN) are nonexistent, meaning the model never failed to identify diabetic individuals, which is excellent. True Positive (TP)=114, The cases of diabetic individuals correctly classified as diabetic show a high value, meaning the model is very accurate in identifying diabetic individuals.

The statistical metrics derived from the confusion matrix include, (Accuracy=0.987), This indicates that the model is very accurate. (Precision=0.974), which means that most of the cases classified as diabetic were indeed diabetic. (Sensitivity=0.1), This means that the model did not miss any diabetic cases, and there are no (False Negatives). (F1-Score=0.987), This means that the model is very well-balanced between accuracy and sensitivity.

Based on the results above, we concluded that the best logistic regression model for studying the key factors influencing Type 2 diabetes is as follows:

$Log_e(0) = -3.0585 + 5.0708X_8 + 2.6716X_{10}$ … (5)

From the logistic regression equation, we find that the variable (HbA1c level) contributes to the effect on the dependent variable (Y) by a factor of (5.0708), meaning that a one-unit change in this variable leads to an increase in the probability of diabetes by (5.0708). As for the second variable, it contributes to the effect on the dependent variable (Y) by a factor of (2.6716), meaning that a one-unit change in this variable leads to an increase in the probability of diabetes by (2.6716).

12. Conclusions

1. The cluster analysis revealed that there are three main groups of factors influencing diabetes.
2. The efficiency and suitability of the logistic regression model in predicting Type 2 diabetes showed that blood glucose levels (HbA1c_level) and blood glucose levels (Blood_glucose_level) are the most influential variables in the risk of developing diabetes.
3. Demographic and behavioral factors, such as smoking, high blood pressure, and heart disease, were not significantly influential in predicting the risk when included alongside the direct indicators of blood glucose and diabetes levels from Cluster 2.
4. This research shows that combining hierarchical cluster analysis with logistic regression is an effective approach for identifying the factors influencing Type 2 diabetes. The results indicate that blood sugar and glucose levels are the strongest predictors of diabetes risk, while factors like age, weight, and smoking were not significantly influential. These results help improve diagnostic and predictive processes, contributing to the development of more effective preventive and therapeutic strategies.
5. The model reflects the strong relationship between Cluster 2 and the likelihood of diabetes, where the risk of diabetes increases as HbA1c and blood glucose levels rise.
6. This analysis can be used to determine the critical threshold levels of HbA1c and blood glucose that indicate a higher risk of developing diabetes.

13. References:

1- Afzal, A., Khan, L., Hussain, M. Z., Hasan, M. Z., Mustafa, M., Khalid, A., & Javaid, A. (2024). Customer segmentation using hierarchical clustering. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-6). IEEE.

2- Almaghri, K. I., & Chakraborty, S. (2016). A Comparative Investigation of K-means and Partition Around Medoid Methods of Clustering-a Case Study with Acute Lymphoblastic Leukemia Data". Palestine University Journal for Research and Studies, 56(4020), 1-21.

3- Feng, F., Duan, Q., Jiang, X., Kao, X., & Zhang, D. (2024). DendroX: multi-level multi-cluster selection in dendrograms. BMC genomics, 25(1), 134.

4- G. Roglic, (2016). "WHO Global report on diabetes: A summary," International Journal of Noncommunicable Diseases, vol. 1, no. 1, p. 3.

5- Huang, J., Wang, L., & Yang, Q. (2022). Application of Cluster Analysis in Identifying Diabetes Subtypes and Their Related Risk Factors. Diabetes & Metabolism Journal, 46(2), 345-357.

6- Joshi, T. N., & Chawan, P. M. (2018). Logistic regression and svm based diabetes prediction system. International Journal For Technological Research In Engineering, 5, 4347-4350.

7- Kassambara, A. (2017). Machine learning essentials: Practical guide in R. Sthda.

8- Liu, P., Yuan, H., Ning, Y., Chakraborty, B., Liu, N., & Peres, M. A. (2024). A modified and weighted Gower distance-based clustering analysis for mixed type data: a simulation and empirical analyses. BMC Medical Research Methodology, 24(1), 305.

9- Maria, A., & Gadelha, J. (2009). Global burden of disease attributable to diabetes mellitus in Brazil Carga global de doença devida e atribuível ao diabetes mellitus no Brasil. 25(6), 1234–1244.

10- Pampel, F. (2021). Logistic regression: A primer. SAGE Publications, Inc.

11- Scott, R. A., Lu, V. I., Grove, N., Patnaik, J. L., & Manoharan, N. (2024). Rates of diabetic retinopathy among cluster analysis—identified type 2 diabetic mellitus subgroups. Graefe's Archive for Clinical and Experimental Ophthalmology, 262(2), 411-419.

12- Supsermpol, P., Huynh, V. N., Thajchayapong, S., Suppakitjarak, N., & Chiadamrong, N. (2025). Predicting post-IPO financial performance: a hybrid approach using logistic regression and decision trees. Journal of Asian Business and Economic Studies.

13- Wang, Y., & Chen, H. (2024). Clinical application of cluster analysis in patients with newly diagnosed type 2 diabetes. Hormones, 1-14.