Al-Qadisiyah Journal for Administrative and Economic Sciences
QJAE, Volume 27, Issue 2 (2025)
DOI: https:// doi.org/ 10.33916/ qjae.2025 .028392

ISSN$_{Online}$: 2312-9883
ISSN$_{Print}$: 1816-9171

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

Abdulsalam Idres Abdulkarim
abdulsalam.23csp145@student.uomosul.edu.iq

Dr. Wisam Wadalluh Saleem
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: This study looked at type two diabetes data from 237 individuals, both diabetic and non-diabetic, that were collected from the Nineveh Health Directorate's Community Health Department. The primary determinants of diabetes were categorized using Hierarchical Cluster Analysis (HCA), with cluster distances measured using Complete Linkage. Because a mix of qualitative and quantitative variables were present, a distance matrix was then computed using Gower Distance. To find the most important variables and accurately predict the risk of diabetes, logistic regression was used to find the right equation for regression for each cluster's variables. Python 3 was used to perform the analysis, and a correlation matrix and VIF (variance inflation factor) were used to check for issues like multicollinearity. Three major groups of variables were identified by the cluster analysis results, and the logistic regression model used these three clusters as indicators to assess each group's impact on the chance of developing the disease. To the results, the cluster tied to the levels of blood glucose (HbA1C and the level of blood glucose) was the most significant risk factor of diabetes, while there was not a significant association identified between metabolic, health, social, and behavioral factors. In order convey and estimate this relationship, diabetes risk was predicted using logistic regression based on the two variables defined above. in regard to the outcomes, cluster analysis and logistic regression interact to improve the model's capacity to predict the significant factors, which helps with decision-making and offers greater comprehension of the data according to study.

Key words: type 2 diabetes, blood glucose levels, HbA1C, logistic regression model, complete linkage, and hierarchical cluster analysis.

INTRODUCTION: One of the oldest and most prevalent diseases in the world, type 2 diabetes increase the risk of cardiovascular disease, high blood pressure, and other metabolic issues as well as causing major health problems that lower a patient's quality of life. Advanced statistical methods have become more and more necessary in recent years in order to better understand the variables influencing diabetes and create accurate prediction models. The researcher [5] used cluster analysis to look at several diabetes patterns and the risk factors that can be linked to them. As a way to distinguish between different types of diabetes based on risk factors and identify different types of the disease, the study utilized data from large population-based studies. So as to identify different types of elderly patients with diabetes and multiple comorbidities, the researchers [8] conducted a study. The results show that while older patients with diabetes and multiple comorbidities had more health problems, in particular in areas like depression and diabetes-related distress, those with obesity alone had better health conditions. In their study, the researchers [13] provided a new clustering classification of diabetes that might provide a way to prevent and treat type 2 diabetes early on, which is a major issue for doctors as well as patients. Logistic regression analysis was used to compare each subtype's risk of complications of diabetes and comorbidities. Four unique subtypes of newly diagnosed type 2 diabetes patients have been successfully identified by the study; these subtypes differ in their clinical features, therapies, and risks of complications and comorbidities tied to diabetes. In order to reduce complexity and improve data interpretation, hierarchical clustering analysis (HCA), which organizes related variables into clusters, was applied. Additionally, logistic regression is a powerful method for understanding out the various factors relate to other factors and predict the chance of developing the illness.

Study Problem:

By experimenting with a complex information set containing a mix of qualitative and quantitative variables, the study seeks to determine and organize the factors that affect type 2 diabetes. The most at-risk groups will be determined by placing people based on common patterns using hierarchical clustering analysis. In order to improve disease understanding and advance prevention and early diagnosis strategies, a logistic regression model will be created to predict the risk of health based on these factors.

Study Objectives:

This research aims to:

- To identify the independent variables that affect the risk of Type 2 diabetes in order to predict the most crucial variables affecting diabetes.
- splitting the data into clusters that represent groups with like characteristics and influences using a hierarchical Cluster Analysis (HCA).
- Using the results of logistic regression and cluster analysis to see the ways different factors affect the target variables to choose the most significant.
- Using the derived clusters, an improved logistic regression model is given, enabling precise dependent variable predictions.

Study significance and contributions:

1. The study's significance rests in its ability to predict the major independent variables that affect the onset of type 2 diabetes, which expands the understanding of the risk factors for this disease and aids in the making of informed health decisions.
2. Developing a regression model which includes the key variables that influence the rise of diabetes.
3. Stating each variable's importance and the way they impact the onset of diabetes.

1. Diabetes Mellitus:

A group of metabolic diseases known as diabetes is characterized by raised blood sugar levels resulted on by defects either in insulin function or secretion, or both. One of the main chronic diseases affecting entire populations is diabetes. "Diabetes Mellitus" originates from the Greek words "Syphon" and "Sugar." Blood glucose levels rise as a result of impaired glucose utilization resulted on by a lack of insulin production or an inability of insulin to operate as meant. Glucose has to get into cells since it is needed for cellular metabolism. 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 by transporting glucose from the bloodstream into cells for energy use, the hormone insulin, which is made by the pancreas and secreted by beta cells (B cells), lowers blood glucose levels. [9] [11]

2. **Types of Diabetes:**

Diabetes has multiple primary types in lieu of being made a single condition, including: [4] [11]

1. **Type 1 Diabetes (T1D):** This kind, previously referred to as juvenile diabetes or insulin-dependent diabetes, is typified by unusual or insufficient insulin released from pancreatic beta cells, which results in little to no insulin production. For patients to avoid coma or death, daily insulin injections are required. Excessive urination, perennial hunger and thirst, fatigue, weight loss, and changes in vision are some of the symptoms. Type 1 diabetes has no known cause, and current medical knowledge does not allow for prevention.
2. **Type 2 Diabetes (T2D):** once called to as adult-onset diabetes or non-insulin-dependent diabetes, this type is brought on by the body's insufficient use of insulin, which often comes on by obesity and a lack of physical activity. Around 95 percent of all cases of diabetes are caused by it. while the symptoms may resemble those of Type 1, they are usually sporadic, ending in a delayed diagnosis after problems have already arisen. Although it can strike younger people as well, type 2 diabetes occurs primarily in adults. In order to maximize glucose utilization, treatment consists of insulin production enhancers and blood sugar-lowering drugs.

3. Cluster Analysis:

Cluster analysis uses specific variable patterns to group observations into unnamed groups. These methods are used to divide the variables or objects under inquiry from other groups while organizing them into homogeneous clusters. Finding patterns which organize observations into clusters with common features is the primary focus of this analysis. This is going to render it simpler to forecast the traits or behavior of new objects based on the clusters to which they are assigned. Various fields, including marketing, medicine, and public health, have found success with cluster analysis.[7] [11]

4. Data Standardization:

The scale of measurement is closely linked to the distance measurement values. Therefore, it is usual to standardize variables before determining differences between observations, especially if variables can be measured on different scales (e.g., kilometers, kilograms, centimeters). Comparability is made certain through standardization, which shifts 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:

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

Thus, we obtain the standardized variables: $(Z_1, Z_2, \ldots, Z_p)$, 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]

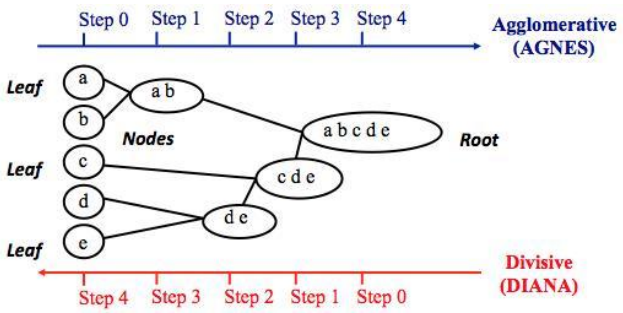


Figure (1) Hierarchical Agglomerative and Divisive Clustering.

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]

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

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]

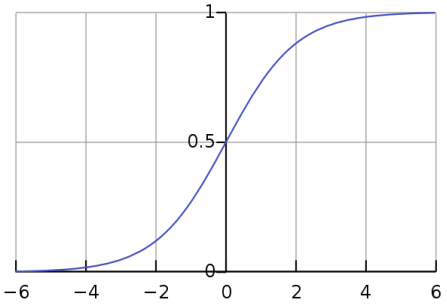


Figure (2): The Logistic Function

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:** Confirms the significance level of the relationship within each independent variable and the dependent variable.

10. Practical Application:

a total of three main elements to the applied aspect: the first consisted of collecting data and variables for the study; the second was using cluster analysis to mathematically analyze the variables; and the third was using logistic regression to analyze the data.

The factors affecting Type 2 diabetes were investigated using logistic regression and hierarchical cluster analysis (HCA). In 2024, a hospital provided the study's data, including **237 individuals**. From the dataset were:

- **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.

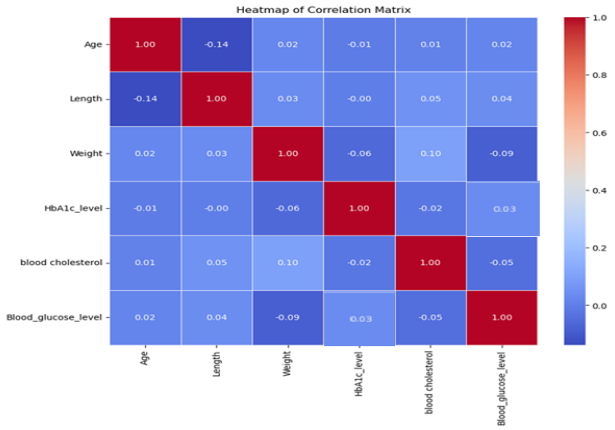
Statistical Analysis Steps:

- Python 3 was used to perform the analysis. The correlation between variables had to be tested before cluster analysis was able to be used to the data. In order to gauge the possible relationships between variables, the correlation matrix and its heatmap were created.

	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

It is clear from the correlation matrix above that there are no strong linear relationships between the variables, as the correlations are very weak.

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



would affect the use of cluster analysis. Thus, cluster analysis is the second step.

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. Dendrograms were used to explain the clustering results in Figure (3), and Table (2) shows the clustering results as follows:

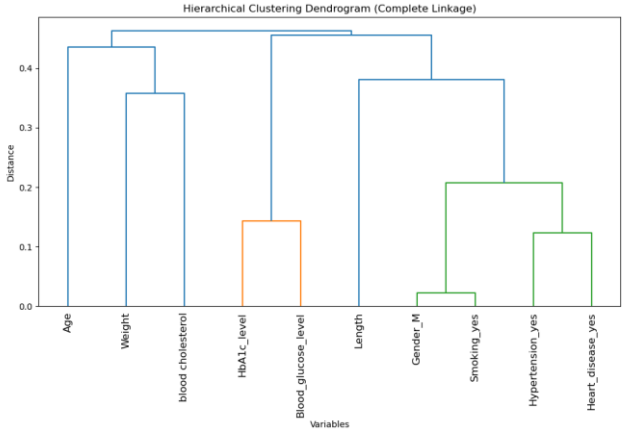


Figure (3): Dendrograms of Clusters

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

The clustering steps for the diabetes variables can be seen in Table (2), along with the distances and similarity levels. The lowest distance level (0.02139) shows that these clusters are merging more, while the greatest similarity level (98.9305) between clusters 1 and 5 indicates that these clusters are highly similar. The cluster numbers that were merged in that step can be seen in the "Clusters joined" column, and the number of new clusters that were formed soon after the merge is shown in the "New cluster" column. Lastly, the number of observations in the new cluster is given by "Number of obs. in new cluster."

11. Interpretation of Cluster Analysis Results

Three primary groups of variables (three clusters) have been determined by the cluster analysis results using the dendrogram in Figure (3) and Table (2):

1. Cluster 1 (Metabolic and Health Factors):

The metabolic factors that can influence blood glucose levels are shown by this cluster, which also includes blood cholesterol levels, age, and weight.

2. Cluster 2 (Diabetes Indicators):

Since blood glucose and HbA1C levels are two of the most prominent markers of diabetes, they are included in this cluster.

3. Cluster 3 (Demographic and Behavioral Factors):

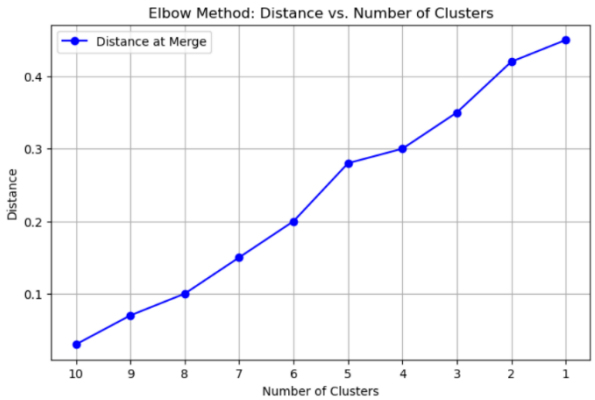
This cluster, comprising heart disease, smoking, gender, and hypertension, points out the manner in which lifestyle and demographic factors raise the risk of diabetes.

These clusters throw light on the relationships between multiple variables and how their can impact health outcomes, particularly as it comes to diabetes risk.

Three clusters were chosen based on a set of criteria, including the Silhouette Score (0.8823). This value's closeness to 1 indicates that the clusters are excellent and clearly defined.

The ideal number of clusters in the cluster analysis was also identified using the elbow method. The elbow point appears to be at three clusters in Figure (4), that demonstrates:

Figure (4) Elbow Method



The distance decreases dramatically as the number of clusters expands from one to three, suggesting that data grouping is significantly enhanced by dividing the data into three clusters.

The clusters are cohesive, meaning that the distances between variables within each cluster are relatively small, corresponding to the Dunn Index value of 2.52. It indicates that each cluster's variables are very similar. likewise, the clusters are well-separated, which means that there have been major gaps between them, proving that they are clearly different from one another. This value shows that the clusters possess an excellent caliber and that the variables were successfully and accurately grouped by the cluster analysis.

- Following the variables' clustering, the logistic regression model utilized the three clusters as indicators to predict whether diabetes would be diagnosed or not. The analysis additionally examined at the variables and clusters that were significant and non-significant. The results of the analysis are shown 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	

It is clear from Table 3 that just Cluster 2 has a significant impact. Since Cluster 2 is the only explanatory variable in this model with statistical significance and has a major beneficial effect on the likelihood of getting diabetes, the model will be rebuilt solely using it. To find out whether AIC and BIC improve, a new model will be built after the insignificant variables (Clusters 1 and 3) are omitted. AIC, BIC, and pseudo R-squared will be used for evaluating the new model's performance. with the evidence of the variables' significance, the logistic regression results for Cluster 2 are shown 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

According to Table (4), Cluster 2's coefficient is 5.3856, which is highly significant (P < 0.001) and shows that this cluster, which includes the variables blood glucose level and HbA1c level, has a major impact on diabetes risk. Pseudo R-squared = 0.8598 indicates that the model is better than the previous model in Table (3) and explains quite a bit of the variance in the data. AIC = 59.88 and BIC = 66.81 are both relatively low when compared to the prior model in Table (3), implying that this model is better in terms of quality and simplicity.
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.

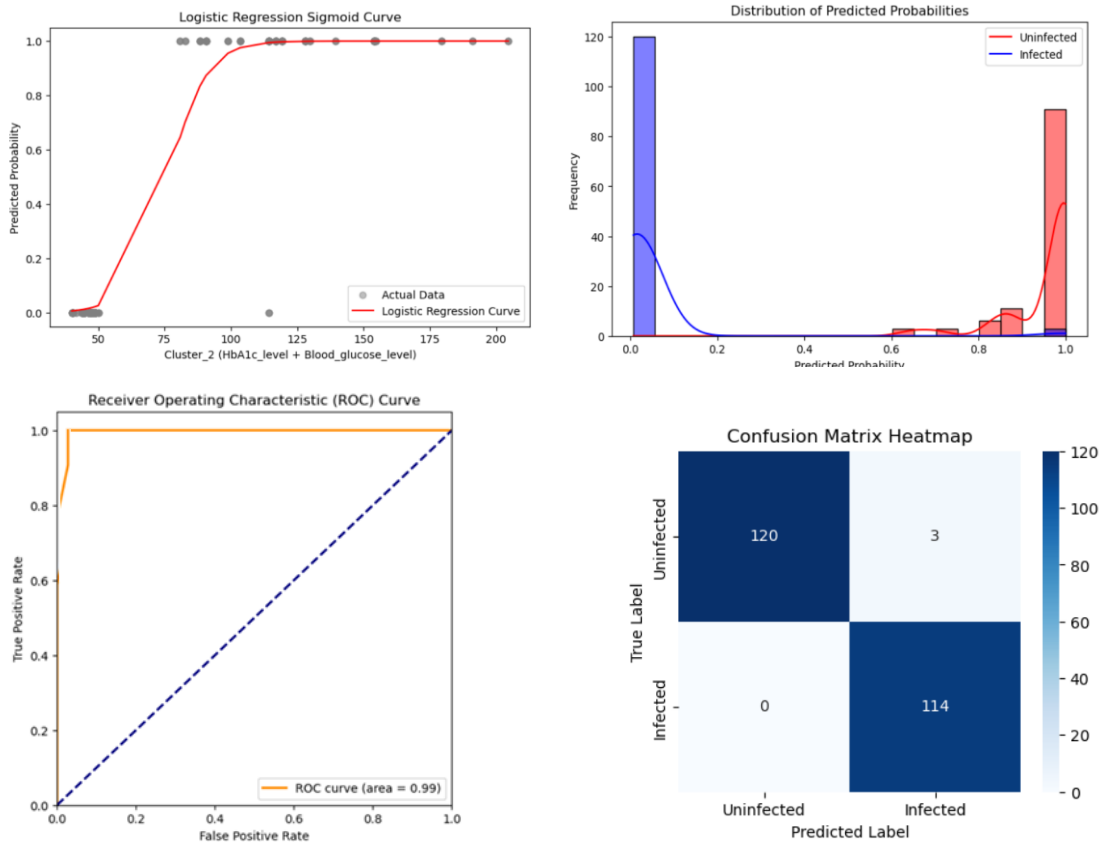


Figure (5) illustrates the effect of Cluster 2 and its variables on health status

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). regard to the second variable, it has a factor of 2.6716 in the effect on the dependent variable (Y), that means that a one-unit change in this variable raises the risk of diabetes by 2.6716.

12. Conclusions

1. Three primary groups of factors influencing diabetes have been found by the cluster analysis.
2. Blood glucose levels (HbA1c_level) and blood glucose levels (Blood_glucose_level) are the most significant factors in the risk of developing diabetes, based on the efficacy and suitability of the logistic regression model in predicting Type 2 diabetes.
3. When combined with the direct indicators of blood glucose and diabetes levels from Cluster 2, behavioral and demographic factors like smoking, high blood pressure, and heart disease did not significantly affect risk prediction.
4. This research shows that the use of logistic regression and hierarchical cluster analysis is an effective approach to identify the variables influencing Type 2 diabetes. The results show that the most accurate signs of diabetes risk are blood sugar and glucose levels, whereas age, weight, and smoking had little to no impact. These results will aid in the development of better therapeutic and preventive strategies by enhancing diagnostic and predictive techniques.
5. Cluster 2 and the risk of diabetes have a strong connection in the model, with the risk of diabetes rising as blood glucose and HbA1c levels do.
6. The critical threshold levels of blood glucose and HbA1c that signify a greater chance of developing diabetes can be determined with this analysis.

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.