Journal of Applied Mathematics and Physics, 2025, 13(9), 3186-3201
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Analysis of Risk Factors and Segment-Specific Strategies for Diabetes Prevention

Aya Patricia Konan[1], Adama Coulibaly[2], Kouassi Bernard Saha[3], Souleymane Oumtanaga[4]

[1]Faculty of Mathematics and Computer Science, Felix Houphouët-Boigny University, Abidjan, Côte d'Ivoire
[2]Institute for Mathematical Research (IRMA), Abidjan, Côte d'Ivoire
[3]Higher Teacher Training School, National Polytechnic Institute Félix Houphouët-Boigny, Yamoussoukro, Côte d'Ivoire
[4]Laboratory of Computer Science and Telecommunications, National Polytechnic Institute, Abidjan, Côte d'Ivoire
Email: scolarite@univ-fhb.edu.ci, Couliba@yahoo.fr, benitosaha@gmail.com, oumtana@gmail.com

How to cite this paper: Konan, A.P., Coulibaly, A., Saha, K.B., and Oumtanaga, S. (2025) Analysis of Risk Factors and Segment-Specific Strategies for Diabetes Prevention. *Journal of Applied Mathematics and Physics*, **13**, 3186-3201.
https://doi.org/10.4236/jamp.2025.139181

Received: September 2, 2025
Accepted: September 25, 2025
Published: September 28, 2025

Abstract

This study proposes a segmented approach to analyzing diabetes risk factors using the dataset *diabete_custom.xlsx* (150 individuals, 14 medical and behavioral variables). The combination of KMeans with logistic regression and KMeans with decision tree enabled the definition of three clusters corresponding to low, moderate, and high risk, while identifying key variables such as blood glucose, BMI, and heredity. The hybrid models improve accuracy and interpretability compared to KMeans alone, with the decision tree being slightly more effective in unbalanced clusters. These findings provide a foundation for personalized interventions, including targeted screening, glycemic and nutritional monitoring, physical activity, and educational campaigns tailored to each risk profile.

Keywords

Diabetes, KMeans, Logistic Regression, Decision Tree, Segmentation

1. Introduction

Diabetes is a multifactorial chronic disease influenced by genetic, biological, and behavioral factors. Each individual presents a unique risk profile, making prevention particularly complex. Existing literature often focuses on global diabetes prediction or the extraction of general rules without considering population segmentation. This is the case, for example, in Sarra S. (2024) "AI-Based Approach for a Diabetes Prediction System" [1], Mohebbi M. A. (2021) "A Machine Learning Approach to Treatment Improvement in Diabetes" [2], and Amani Hamada Bonheur (2024) "Design and Implementation of an Intelligent Web Application for Diabe-

tes Diagnosis" [3]. However, diabetes risk varies according to individual profiles, and a uniform approach does not allow for targeted prevention strategies.

Segmenting the population into homogeneous subgroups enables the identification of key factors and guides personalized interventions. This study adopts a hybrid approach combining KMeans, logistic regression, and decision trees, applied to the *diabete_custom.xlsx* dataset. The objective is to identify the determinant variables for each cluster and extract simple rules to inform clinical and behavioral prevention.

2. Dataset Description

The *diabete_custom.xlsx* dataset comprises 150 individuals and 14 variables, derived and enriched from the Pima Indians Diabetes Dataset [4]. It includes medical factors (age, BMI, blood glucose, HbA1c, blood pressure, cholesterol, waist circumference, family history) as well as behavioral factors (physical activity, smoking, alcohol consumption, BMI category). The target variable, Diabetes, is binary (0 = absence, 1 = presence). The Excel file is directly compatible with standard data analysis tools.

3. Methodology

3.1. Data Preprocessing

The methodology begins with data preprocessing, which consists of separating the explanatory variables X from the target variable Y. Each variable x_j is then normalized to ensure comparability across features, according to the formula:

$$x_j^{norm} = \frac{x_j - \mu_j}{\sigma_j}$$

where μ_j and σ_j represent the mean and standard deviation of variable x_j, respectively. This normalization is essential for distance-based methods such as KMeans, in order to prevent any single variable from dominating the others due to differences in scale [5].

The choice of the optimal number of clusters k is determined using two complementary approaches. The elbow method relies on intra-cluster inertia, defined as:

$$W_k = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$

where μ_i is the centroid of cluster C_i and $\|\cdot\|$ denotes the Euclidean norm. The total inertia decreases as k increases, and the "elbow" of the curve helps identify a trade-off between the number of clusters and the compactness of the groups [6].

The silhouette method complements this analysis by evaluating the cohesion and separation of clusters for each individual i, according to:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

where $a(i)$ is the average intra-cluster distance and $b(i)$ is the average distance to the nearest cluster. The value of $s(i)$ ranges from −1 to 1, with a score close to 1 indicating that the individual is well assigned to its cluster, while a negative score suggests inappropriate clustering [7].

3.1.1. Presentation of the Curves

Figure 1 and Figure 2 show the elbow plot and the silhouette plot, respectively.

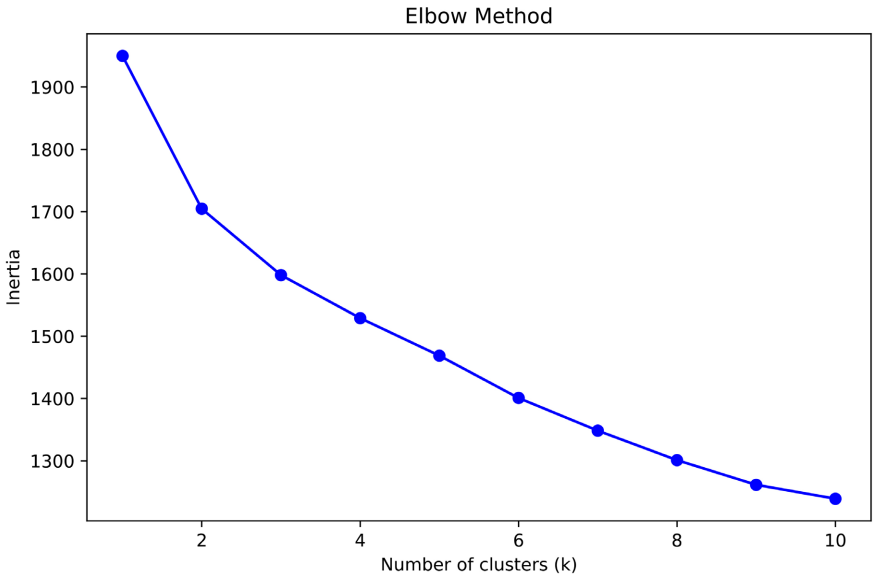

Figure 1. Elbow curve.

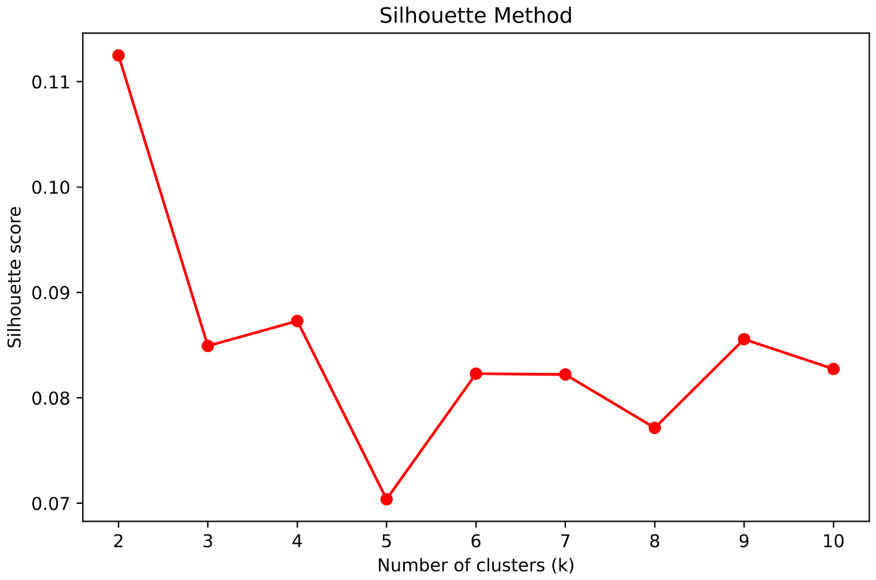

Figure 2. Silhouette curve.

3.1.2. Justification for Choosing $k = 3$

The elbow method shows a sharp decrease in inertia up to $k = 3$, followed by stabilization, indicating an optimal point. Simultaneously, the silhouette score reaches

its highest value at $k = 3$, reflecting strong internal cohesion and clear separation between groups. These convergent results justify the choice of three clusters, consistent with the expected typology of diabetes profiles (non-diabetic, pre-diabetic, diabetic).

3.2. Hybrid Model: KMeans + Logistic Regression

For each cluster C_k, logistic regression is applied:

$$P(y=1|x)=\frac{1}{1+e^{-\left(\beta_0+\sum_{j=1}^{p}\beta_j x_j\right)}}$$

where:

- $y = 1$ indicates the presence of diabetes.
- B_0 is the intercept.
- B_j is the coefficient associated with variable x_j.

Interpretation:

- If $B_j > 0$, the variable x_j contributes to an increased risk of diabetes;
- If $B_j < 0$, it has a protective effect by reducing this risk.

Simplified rules by cluster:

If x_j increases, then the risk of diabetes is proportional to $\propto \beta_j$.

3.3. Hybrid Model: KMeans + Decision Tree

The decision tree aims to partition the data into homogeneous subgroups [8].

- **Impurity criterion used:** Gini index

$$\text{Gini}=1-\sum_{C+1}^{C}p_c^2$$

where p_c is the proportion of observations of class c in the node [9].

- **Decisions are represented as simple rules:**

Example: "If Blood Glucose > 140 mg/dl and BMI > 30 → Diabetes."

3.4. Model Evaluation

The evaluation of the hybrid models' performance is based on two complementary aspects: quantitative indicators and the relevance of the extracted rules.

3.4.1. Quantitative Indicators

To assess the quality of predictions, accuracy is used, which corresponds to the proportion of correct predictions relative to the total number of predictions:

$$\text{Accuracy}=\frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}$$

where TP (True Positives) represents true positives, TN true negatives, FP false positives, and FN false negatives [10].

Recall assesses the model's ability to correctly identify individuals who are actually positive (diabetic):

$$\text{Recall}=\frac{\text{TP}}{\text{TP}+\text{FN}}$$

Finally, the F1-score combines precision and recall into a harmonic mean, providing a single metric that reflects the balance between accuracy and sensitivity:

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}$$

3.4.2. Relevance of the Rules

Beyond numerical metrics, it is essential to evaluate the readability and clinical and behavioral applicability of the rules extracted by the models. The rules should be interpretable, consistent with known risk factors such as blood glucose, body mass index, heredity, and physical activity, and directly actionable to guide prevention and intervention strategies.

4. Results

4.1. Comparison of the Performance of KMeans + Logistic Regression and KMeans + Decision Tree

Table 1 and Table 2 present the baseline performance of KMeans and the cluster distribution, while Table 3 and Table 4 show the enhanced performance achieved through Logistic Regression and Decision Tree per cluster.

Table 1. Overall results of KMeans + mapping → classes.

Metric	Value	Comment
Accuracy	0.71	Fair performance but can be improved, especially for diabetic individuals.
Precision (0)	0.84	The non-diabetic class is well predicted.
Recall (0)	0.70	70% of non-diabetic individuals are correctly identified.
F1-score (0)	0.76	Good balance between precision and recall for non-diabetics.
Precision (1)	0.54	Low precision for diabetic individuals.
Recall (1)	0.71	Acceptable recall, some diabetics are misclassified.
F1-score (1)	0.61	Average score for the diabetic class.
Macro avg	0.69	Balanced average across both classes.
Weighted avg	0.74	Acceptable weighted average.

Legend 1: Overall performance of KMeans after cluster mapping, with class-specific precision and comments on reliability.

Table 2. Cluster-wise results—Unsupervised KMeans.

Cluster	Accuracy	Precision (0/1)	Recall (0/1)	F1-score (0/1)	Comment
0	0.68	0.68/0.00	1.00/0.00	0.81/0.00	Class 1 (diabetic) not predicted, low performance for this class.
1	0.98	0.98/0.00	1.00/0.00	0.99/0.00	Very good for non-diabetics, but no diabetics present to evaluate.

Continued

2	0.54	0.00/0.54	0.00/1.00	0.00/0.70	Conversely, class 0 poorly predicted, overall performance low.

Legend 2: Cluster-wise performance for unsupervised KMeans, indicating the model's ability to predict each class and the observed limitations for certain classes.

Table 3. Cluster-wise results—KMeans + logistic regression.

Cluster	Accuracy	Precision (0/1)	Recall (0/1)	F1-score (0/1)	Comment
0	0.97	1.00/0.93	0.96/1.00	0.98/0.96	Very good performance for both classes.
1	0.98	0.98/0.00	1.00/0.00	0.99/0.00	Diabetic class almost absent, minority class evaluation not possible.
2	0.98	1.00/0.97	0.97/1.00	0.98/0.99	Excellent performance in this cluster, very balanced.

Legend 3: Cluster-wise performance for the KMeans + Logistic Regression model, showing overall effectiveness and evaluation limitations when certain classes are underrepresented.

Table 4. Cluster-wise results—KMeans + decision tree.

Cluster	Accuracy	Precision (0/1)	Recall (0/1)	F1-score (0/1)	Comment
0	0.92	1.00/0.83	0.86/1.00	0.92/0.91	Good balance, slightly underestimated for class 1.
1	0.86	1.00/0.00	0.86/0.00	0.92/0.00	Difficult cluster: no diabetics present, recall and F1 for class 1 not computable.
2	1.00	1.00/1.00	1.00/1.00	1.00/1.00	Perfect cluster for the decision tree, all classes correctly predicted.

Legend 4: Cluster-wise performance for the KMeans + Decision Tree model, highlighting overall effectiveness, class balance, and limitations related to clusters where certain classes are absent.

4.2. Interpretation of Results

Global KMeans with class mapping shows fair performance (Accuracy 0.71), with good prediction for non-diabetic individuals but limited precision for diabetics. At the cluster level, unsupervised KMeans exhibits strong variations: some clusters contain almost exclusively one class, making evaluation of the other class impossible, and prediction is poor for the minority class. KMeans + Logistic Regression achieves excellent performance for balanced clusters, but evaluation is limited when a class is nearly absent. KMeans + Decision Tree efficiently predicts all classes in balanced clusters, maintaining good class balance, whereas clusters where

certain classes are absent show limitations in recall and F1-score. Overall, hybrid models improve prediction compared to KMeans alone, particularly in well-represented clusters.

4.3. Graphs

4.3.1. PCA Visualization of Clusters and Interpretation

After applying KMeans, individuals are grouped into three clusters:

- Cluster 0: High-risk profiles, a mix of diabetics and non-diabetics.
- Cluster 1: Very low-risk profiles, predominantly non-diabetic, homogeneous.
- Cluster 2: Very high-risk profiles, higher proportion of diabetics, heterogeneous profiles.

The PCA projection in Figure 3 shows good separation of the groups for both hybrid models.

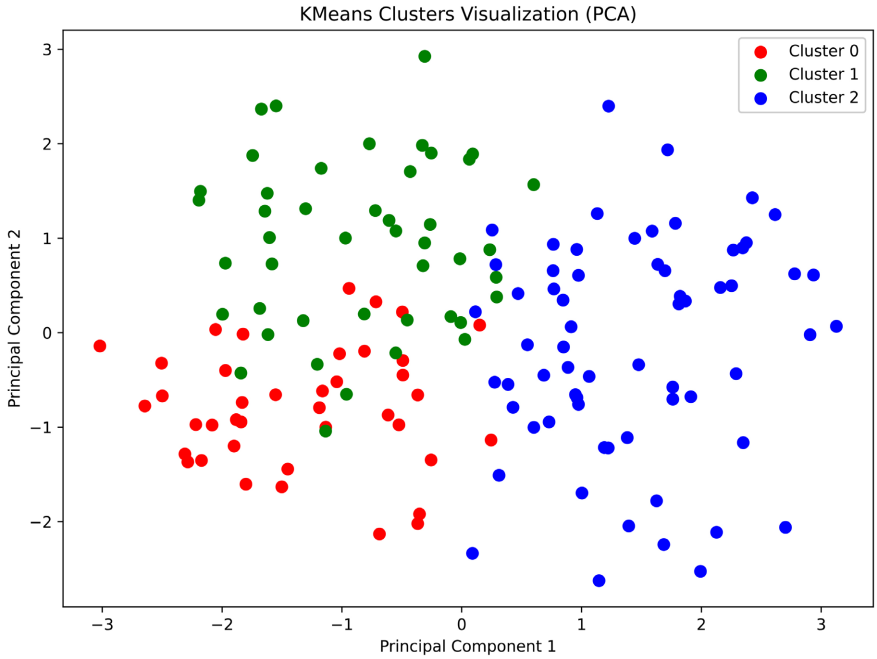

Figure 3. KMeans model on PCA projection.

4.3.2. Interpretations and Explanations of KMeans Clusters

Cluster 0—High Risk

- **Average profile:** Age: 52 years, Blood Glucose: 173 mg/dL, HbA1c: 7.8%, BMI: 25.4, Physical Activity: high, Alcohol Consumption: high.
- **Explanation:**
  - The population is relatively older, which is a classic risk factor for diabetes.
  - Very high blood glucose and HbA1c above 6.5% indicate a pre-diabetic or confirmed diabetic state.
  - BMI is moderate, so the risk is not dominated by obesity.
  - High physical activity may partially mitigate risk, but high alcohol consumption and glycemic profile are aggravating factors.

- **Proportion of diabetics:** 32.5% → indicates that this cluster already contains a significant share of diabetic individuals.
- **Conclusion:** Despite a moderate BMI, the glycemic profile and advanced age place this cluster at high risk, requiring medical monitoring and targeted interventions focused on glycemic control and nutrition.

Cluster 1—Very Low Risk

- **Average profile:** Age: 50 years, Blood Glucose: 110 mg/dL, HbA1c: 6.9%, BMI: 25.1, Physical Activity: low, Alcohol Consumption: low.
- **Explanation:**
  - Blood glucose and HbA1c are close to normal, indicating minimal diabetes risk.
  - Age is slightly lower than that of Cluster 0.
  - Low physical activity could be a risk factor, but it is offset by a generally healthy metabolic profile.
- **Proportion of diabetics:** 2.2% → almost no diabetics in this cluster.
- **Conclusion:** This cluster represents a very low-risk population with an overall healthy profile.

Cluster 2—Very High Risk

- **Average profile:** Age: 46 years, Blood Glucose: 129 mg/dL, HbA1c: 6.8%, BMI: 35.1, Physical Activity: low, Alcohol Consumption: moderate.
- **Explanation:**
  - Despite a younger age, severe obesity (BMI > 35) dominates the risk profile.
  - Blood glucose and HbA1c are moderate but already above normal thresholds.
  - Low physical activity exacerbates metabolic risk.
- **Proportion of diabetics:** 53.8% → this is the most affected cluster.
- **Conclusion:** The combination of severe obesity, elevated blood glucose, and low physical activity places this cluster at very high risk for diabetes. Interventions should prioritize weight management, nutrition, and physical activity.

General Remarks:

- KMeans clustering uses all standardized factors. Therefore, some clusters may exhibit high risk even if a single factor (such as age or blood glucose) is not maximal, because other factors compensate.
- The profiles are consistent: a "young obese cluster with a high proportion of diabetics" (Cluster 2) and an "older cluster with moderate BMI" (Cluster 0) reflect realistic diabetes subpopulations.

4.3.3. Hybrid KMeans + Logistic Regression

The KMeans model combined with logistic regression allowed the identification of three distinct clusters, each exhibiting specific diabetes-related risk profiles. This approach highlights the key variables unique to each group and facilitates targeted analysis.

Cluster 0 (Logistic Regression)

Interpretation:

The model shows excellent performance (accuracy 0.97). The diabetic class (1) is very well detected (recall 1.00) with good precision (0.93). The most influential variables are heredity, BMI, and blood glucose, which significantly increase the risk of diabetes. Paradoxically, HbA1c appears as a protective factor (negative coefficient).

Explanation:

This cluster groups a population where heredity and excess weight play a major role. The counterintuitive result for HbA1c is related to a particular distribution: some individuals with lower HbA1c may still be classified as diabetic due to very high BMI and blood glucose. This illustrates the limitation of global coefficients within a restricted cluster.

Cluster 1 (Logistic Regression)

Interpretation:

The model achieves excellent overall performance (accuracy 0.98), but it does not correctly predict the diabetic class (no individuals of class 1 detected, which skews precision/recall). The determining factors are heredity, blood glucose, BMI, and cholesterol, which increase risk. As in Cluster 0, HbA1c paradoxically appears as a protective factor.

Explanation:

This cluster is unbalanced: almost no diabetics are present (class 1 support = 1). The model is therefore overfitted to non-diabetics, explaining the biased performance. The absence of diabetics prevents a true assessment of model robustness. This cluster mainly illustrates the preventive role of normal variables (low blood glucose and HbA1c) but highlights the statistical limitations of regression on minority groups.

Cluster 2 (Logistic Regression)

Interpretation:

Excellent performance (accuracy 0.98). The model clearly distinguishes between diabetic and non-diabetic individuals. Blood glucose is by far the dominant factor (very high coefficient). Some results appear counterintuitive: age, BMI, and heredity appear as protective factors (negative coefficients), whereas HbA1c and cholesterol increase the risk.

Explanation:

This cluster appears to be characterized by a younger population with an overall high BMI. The predominant weight of blood glucose masks the other variables: even a young obese individual may be classified as non-diabetic if their blood glucose is normal. The "protective" effect of age and BMI does not reflect clinical reality but rather an internal correlation effect: in this cluster, true diabetics are younger and have very high blood glucose/HbA1c, hence this paradox.

Figure 4 shows the variables ranked by importance, followed by a summary of the simplified explanatory rules in Figure 5, which indicate the positive or negative influence of each factor on the probability of belonging to a diabetic profile in Cluster 0.

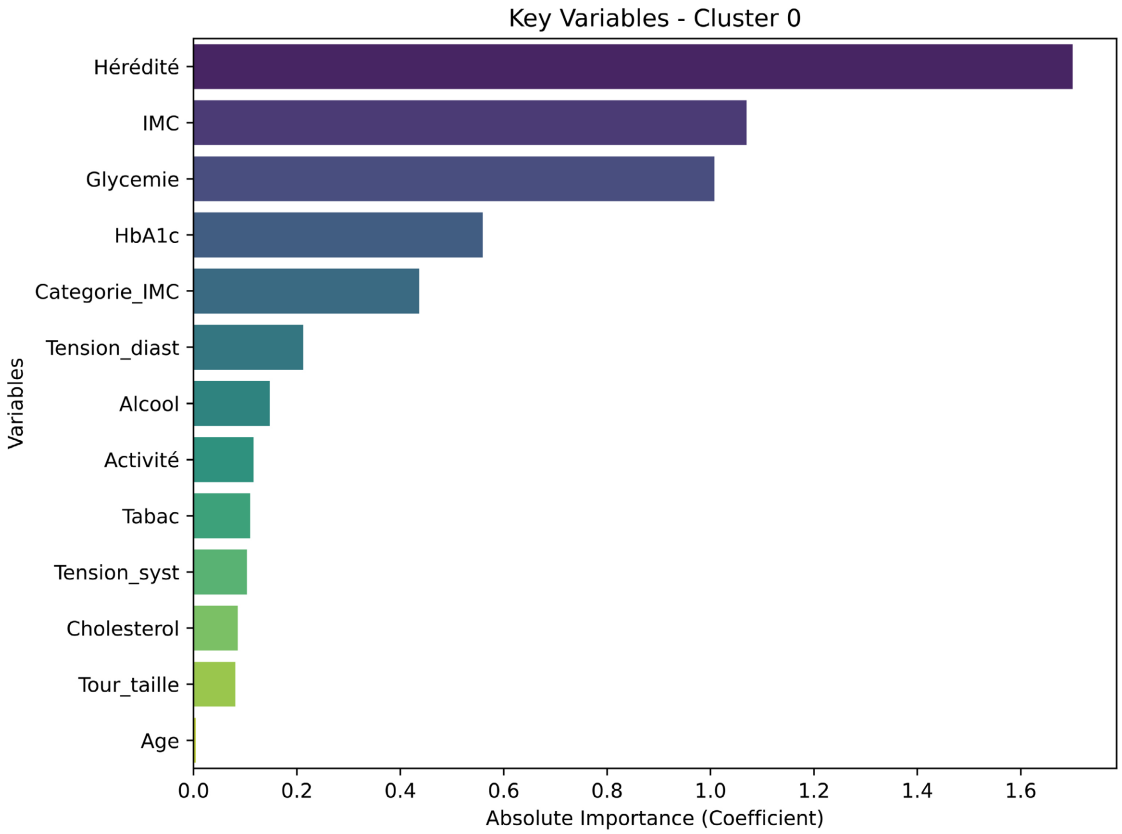

Figure 4. Key variables/cluster 0.

Simplified Explanatory Rules - Cluster 0

- An increase in Hérédité increases diabetes risk
- An increase in IMC increases diabetes risk
- An increase in Glycemie increases diabetes risk
- An increase in HbA1c decreases diabetes risk
- An increase in Categorie_IMC increases diabetes risk
- An increase in Tension_diast increases diabetes risk
- An increase in Alcool decreases diabetes risk
- An increase in Activité increases diabetes risk
- An increase in Tabac increases diabetes risk
- An increase in Tension_syst decreases diabetes risk
- An increase in Cholesterol decreases diabetes risk
- An increase in Tour_taille decreases diabetes risk
- An increase in Age decreases diabetes risk

Figure 5. Explanatory rules/cluster 0.

4.3.4. Hybrid KMeans + Decision Tree

This model highlights simple and interpretable rules to differentiate risk profiles. The decision tree primarily relies on heredity, BMI category, and blood glucose to classify individuals as diabetic or non-diabetic.

Cluster 0 (Decision Tree)

Interpretation: This cluster shows good classification performance (accuracy 92%). The rules indicate that diabetes is mainly associated with BMI (above 1.5, indicating overweight/obesity) and blood glucose (>146.5), especially in cases of positive heredity.

Explanation: Risk is modulated by two key factors: high BMI and blood glucose above the critical threshold increase the probability of diabetes, particularly for individuals with a family history. Conversely, normal BMI and lower blood glucose act as protective factors.

Cluster 1 (Decision Tree)

Interpretation: Classification shows high precision for non-diabetic individuals (accuracy 86%), but no data on diabetic cases (zero support). The tree indicates that blood glucose ≤ 140 is associated with non-diabetics.

Explanation: This cluster groups a predominantly healthy population, characterized by normal blood glucose. The absence of diabetic cases prevents fine-tuning predictions for positives but confirms that blood glucose remains the main discriminating indicator.

Cluster 2 (Decision Tree)

Interpretation: The results are perfect (accuracy 100%). The main rule is based on blood glucose: ≤125 for non-diabetics, >125 for diabetics.

Explanation: This cluster illustrates a clear and robust separation based exclusively on blood glucose. It represents a group where glycemic value is the predominant factor, making classification highly reliable and directly interpretable clinically.

The rules extracted from Figure 6 and Figure 7 illustrate the decision-making process and the key variables of Cluster 0.

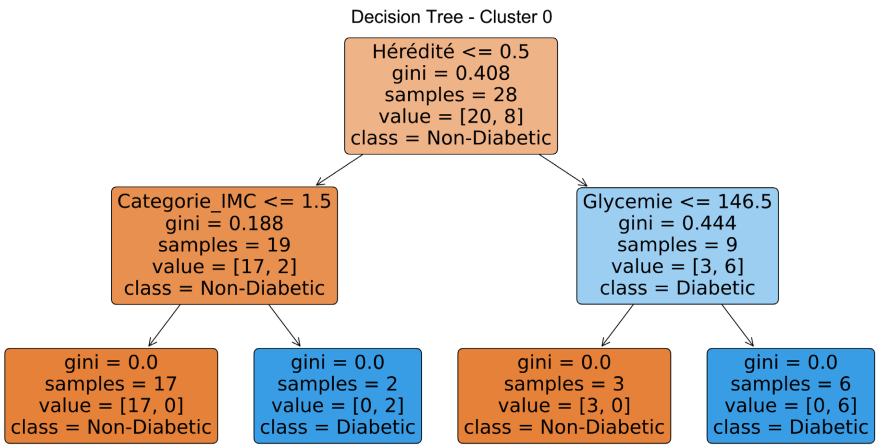

Figure 6. Decision tree—KMeans + Decision tree/Cluster 0.

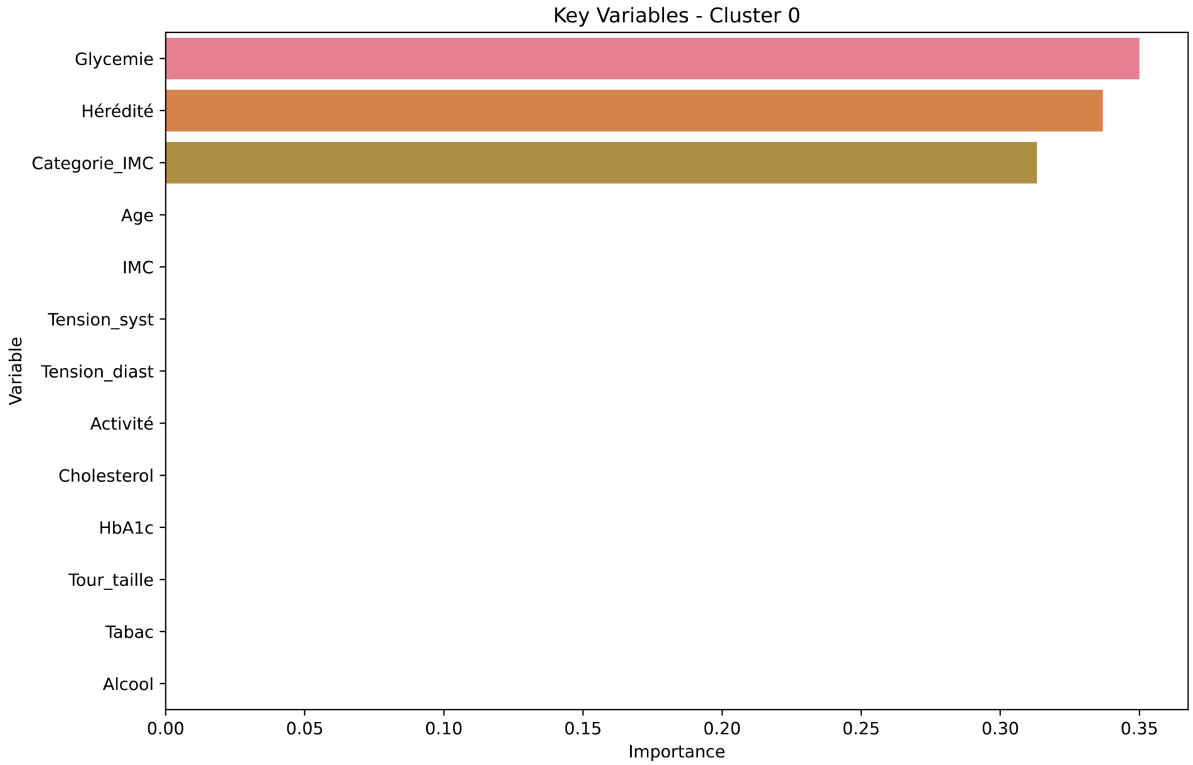

Figure 7. Key variables/cluster 0.

5. Discussion

5.1. Performance Analysis: Strengths and Weaknesses of Each Model

Table 5 summarizes the performance of the KMeans-based hybrid models, highlighting their strengths and limitations. This synthesis, without providing detailed numerical metrics, allows a quick assessment of the overall effectiveness of each approach, while emphasizing the influence of minority classes and the impact of unbalanced clusters on the results.

Table 5. Strengths and weaknesses of each model.

Model	Strengths	Weaknesses
Global Kmeans (mapping → classes)	Good identification of the majority class; acceptable balance between the two classes	Low precision for the minority class; some diabetics misclassified
Unsupervised KMeans by cluster	Some clusters predict the majority class very well	Minority class often absent or poorly predicted; overall poor performance in unbalanced clusters
KMeans + Logistic Regression	Excellent performance in balanced clusters; both classes correctly predicted	Evaluation impossible for the minority class in unbalanced clusters; sensitive to class distribution

Continued

KMeans + Decision Tree	Excellent performance in balanced clusters; captures complex variable interactions; good balance in some partially unbalanced clusters	Clusters with absence of a class prevent evaluation of some metrics; sensitive to small clusters and risk of overfitting

Legend 5: Summary of the strengths and limitations of KMeans alone and hybrid models, depending on cluster balance and diabetic prediction.

5.2. Best Method

By comparing the two approaches, it is observed that both models perform very well on balanced clusters but show limitations when certain classes are underrepresented [11]. KMeans + Logistic Regression offers a good precision/recall trade-off on balanced clusters but becomes limited when the minority class (diabetics) is almost absent [12]. KMeans + Decision Tree remains highly effective on balanced clusters, sometimes achieving perfect balance, but struggles on unbalanced clusters where a class is absent [13]. In this study, the KMeans clusters are unbalanced, making the Decision Tree slightly preferable: it excels on the main cluster, remains generally robust, and provides interpretable rules, whereas Logistic Regression, although reliable, is less suited to highly unbalanced clusters.

5.3. Educational and Medical Actions Based on the KMeans + Logistic Regression Method

Intervention guidelines by cluster, according to clinical, practical, behavioral, and educational dimensions, are as follows:

Cluster 0—Risk modulated by BMI, heredity, and blood glucose:

- **Clinical:** Regular screening is recommended for individuals with a family history, close glycemic monitoring if blood glucose exceeds 146 mg/dL, and weight management.
- **Practical:** Annual medical check-ups and referral to a nutritionist in cases of overweight are advised.
- **Behavioral:** Encourage maintenance of regular physical activity and reduction of alcohol consumption.
- **Educational:** Specific awareness of risks related to heredity and overweight, as well as education on glycemic monitoring, is essential.

Cluster 1—Predominantly healthy population with low diabetes prevalence:

- **Clinical:** Standard screening is sufficient, with glycemic vigilance if values approach the 140 mg/dL threshold.
- **Practical:** Promote a balanced lifestyle without intensive medical interventions.
- **Behavioral:** Encourage increased physical activity to enhance cardiovascular protection.
- **Educational:** General prevention campaigns focusing on healthy living, balanced nutrition, and exercise are appropriate.

Cluster 2—High-risk population with clear distinction based on blood glucose:

- **Clinical:** Close monitoring, including HbA1c and blood glucose tests, is necessary, with prompt management for cases exceeding 125 mg/dL and referral for specialized follow-up.
- **Practical:** Implement intensive weight reduction programs and tailored nutritional monitoring.
- **Behavioral:** Adoption of an appropriate diet, regular and supervised physical activity, and reduction of tobacco and alcohol consumption are recommended.
- **Educational:** Targeted educational programs on obesity-related risks and the importance of glycemic control, as well as practical workshops to facilitate lifestyle changes, are essential.

5.4. Limitations

5.4.1. Study Limitations

Methodologically, the small sample size and the arbitrary choice of the number of clusters may affect the stability of the results, with a risk of overfitting in the decision tree. Clinically, the averages used to define the clusters do not reflect individual variability, and some groups (Cluster 1) are underrepresented. Practically, implementing the recommendations, particularly for intensive monitoring of Cluster 2, may be challenging in resource-limited settings. Behaviorally, the study does not account for actual adherence or social and cultural factors. Educationally, the interventions remain general, and their effectiveness has not been evaluated.

5.4.2. Perspectives

After implementing cluster-specific interventions, the perspectives are as follows:

- **Methodological:** The interventions allow validation and refinement of the KMeans + Decision Tree model, enabling more precise monitoring of at-risk individuals.
- **Clinical:** Targeted screening and glycemic monitoring improve early detection and individualized management.
- **Practical:** Annual check-ups, nutritional follow-up, and weight reduction programs optimize the use of medical resources.
- **Behavioral:** Adoption of regular physical activity, healthy eating, and reduction of alcohol and tobacco consumption decreases overall risk.
- **Educational:** Education and awareness programs are tailored to each cluster, from general prevention for Cluster 1 to intensive programs for Cluster 2.

6. Conclusion

6.1. Key Points

This study highlights the value of a segmented approach for analyzing diabetes risk factors. The combination of KMeans with logistic regression and decision tree methods allowed classification of the population into three distinct clusters, corresponding to low-, moderate-, and high-risk profiles. The hybrid models im-

proved both accuracy and interpretability compared to KMeans alone, emphasizing cluster-specific key variables such as blood glucose, BMI, and heredity. Nevertheless, some unbalanced clusters limited the evaluation of minority classes, highlighting methodological and statistical constraints related to sample size. In this context, the cluster performance makes the hybrid KMeans + Decision Tree method generally preferable.

6.2. Value of the Dataset

The dataset *diabete_custom.xlsx*, derived from the Pima Indians Diabetes Dataset and enriched with behavioral and medical variables, enabled the application of hybrid methods on realistic tabular data. Its multidimensional richness facilitated the identification of risk profiles and the detection of key diabetes factors across different clusters, while providing support for targeted and reproducible analyses.

6.3. Recommendations

Based on the results, the following recommendations are proposed:

- **Clinical:** Adapt screening and monitoring according to the identified profiles, with particular attention to individuals in high-risk clusters.
- **Practical:** Implement regular medical check-ups and programs for nutritional monitoring and physical activity for at-risk clusters.
- **Behavioral:** Promote a healthy lifestyle, emphasizing physical activity, weight management, and reduction of alcohol and tobacco consumption.
- **Educational:** Develop targeted educational campaigns according to the cluster, ranging from general prevention for low-risk populations to intensive interventions for high-risk profiles.

In conclusion, this segmented approach provides a powerful tool for guiding personalized prevention strategies, better adapted to individual profiles, and constitutes a solid foundation for future research aimed at optimizing diabetes prevention and management.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Sarra, S. (2024) Approche basée ia pour un système de prédiction du diabète. Thèse de Doctorat, Université Larbi Tébessi.

[2] Mohebbi, M.A. (2021) A Machine Learning Approach to Treatment Improvement in Diabetes. Thèse de Doctorat, Université Technique du Danemark.

[3] Hamada Bonheur, A. (2024) Conception et réalisation d'une application web de diagnostic intelligent du diabète. Thèse de Doctorat, Université des Sciences et Technologies de l'Université de Constantine 2.

[4] Dua, D. and Graff, C. (2017) UCI Machine Learning Repository: Pima Indians Diabetes Dataset. University of California.

[5] Tan, P.-N., Steinbach, M. and Kumar, V. (2019) Introduction à la fouille de données.

2e Edition, Pearson, 120.

[6] Kaufman, L. and Rousseeuw, P.J. (2005) Clustering par partition et méthodes de validation. Dunod, 82.

[7] Rousseeuw, P.J. (1987) Silhouettes: une méthode graphique pour interpréter et valider des clusters. Masson, 53.

[8] Quinlan, J.R. (1993) C4.5: Arbres de décision pour la classification. Eyrolles, 45.

[9] Hastie, T., Tibshirani, R. and Friedman, J. (2011) Apprentissage statistique: Avec applications en R. Springer, 85.

[10] Chicco, D. and Jurman, G. (2020) The Advantages of the Matthews Correlation Coefficient (MCC) over F1 Score and Accuracy in Binary Classification Evaluation. *BMC Genomics*, **21**, Article No. 6. https://doi.org/10.1186/s12864-019-6413-7

[11] Gupta, S.L., Khandelwal, V., Katria, V., Sharma, D.A. and Pandey, A. (2024) Analyzing the Efficacy of K-Means Clustering and Logistic Regression for Diabetes Prediction. *South Eastern European Journal of Public Health*, **25**, 1255-1262. https://doi.org/10.70135/seejph.vi.2454

[12] ElSeddawy, A.I., Karim, F.K., Hussein, A.M. and Khafaga, D.S. (2022) Predictive Analysis of Diabetes-Risk with Class Imbalance. *Computational Intelligence and Neuroscience*, **2022**, Article ID: 3078025. https://doi.org/10.1155/2022/3078025

[13] Aliyu, H.A. (2024) Optimizing Machine Learning Algorithms for Diabetes Data. ScienceDirect, 5.