2026 Volume: 02 Issue: 02 www.ijrpa.com ISSN 2456-9995 Research Article

International Journal Research Publication Analysis

Page: 01-06

PREDICTING PRE-DIABETES USING K-MEANS CLUSTERING: AN UNSUPERVISED APPROACH FOR EARLY RISK STRATIFICATION AND CLINICAL DECISION SUPPORT.

***Chander Deep Singh**

Jammu and Kashmir.

Article Received: 18 January 2026
Article Revised: 07 February 2026
Published on: 27 February 2026

***Corresponding Author: Chander Deep Singh**
Jammu and Kashmir.
DOI: https://doi-doi.org/101555/ijrpa.6154

ABSTRACT

Pre-diabetes represents a critical intermediate stage between normal glucose regulation and Type 2 Diabetes Mellitus (T2DM). Early detection and timely intervention during this stage can significantly reduce the burden of chronic diabetes-related complications. This research explores the use of the K-Means clustering algorithm, an unsupervised machine learning technique, to identify individuals at high risk of pre-diabetes using demographic and clinical health indicators. The objective of this study is to classify subjects into meaningful clusters representing varying risk levels and to develop a framework for early risk stratification and clinical decision support. Experimental analysis demonstrates that K-Means clustering can effectively differentiate risk groups based on fasting blood glucose (FBG), HbA1c, BMI, age, and lifestyle variables. The model provides a foundational decision-support mechanism for healthcare providers to prioritize preventive care strategies.

1. INTRODUCTION

The increasing global incidence of diabetes has made early detection of pre-diabetes essential for proactive disease management. Pre-diabetes is a reversible condition, yet millions of individuals remain undiagnosed due to insufficient screening and lack of awareness. Traditional diagnostic methods rely on laboratory tests such as HbA1c, FBG, and OGTT, but these are often used only when symptoms manifest.

Machine learning (ML) offers the potential to analyze large datasets and detect hidden patterns that may not be evident through conventional methods. Unsupervised learning, in particular, provides the advantage of pattern discovery without requiring labeled medical outcomes. As such, clustering algorithms can help identify groups of individuals with similar risk profiles, serving as a cost-effective approach for early screening.This research focuses on utilizing the K-Means clustering algorithm to analyze health parameters and predict pre-diabetes risk categories. The findings aim to support clinicians in formulating personalized preventive interventions and help public health authorities improve early screening programs.

2. LITERATURE REVIEW

Several studies have explored machine learning for diabetes detection, but the majority utilize supervised learning with labeled datasets. While these models perform well, they are limited by the availability and quality of labeled clinical data.

Supervised approaches such as logistic regression, SVM, and Random Forests have been used to predict diabetes onset with high accuracy.Unsupervised methods, however, remain underutilized in pre-diabetes prediction, despite their potential to reveal hidden structures in patient data.Prior research applying clustering algorithms mainly focused on lifestyle grouping, obesity classification, or metabolic syndrome segmentation.The gap identified is a lack of unsupervised clustering-based models specifically focused on pre-diabetes risk stratification, especially for early clinical decision support.

3. Objectives

To apply the K-Means clustering algorithm to health datasets for identifying pre-diabetes risk groups.

To analyze major clinical features contributing to cluster formation.

To evaluate the effectiveness of unsupervised clustering in pre-diabetes risk prediction.

To develop a clinical decision-support framework based on clustering outcomes.

4. METHODOLOGY

4.1 Dataset DescriptionA structured dataset containing health and lifestyle features was used. Major variables include:

*Age

*Gender

*BMI

*Fasting Blood Glucose (FBG)

*HbA1c (%)

*Blood Pressure

*Physical activity level

*Family history of diabetes

The dataset was pre-processed to handle missing values, normalize numerical features, and encode categorical attributes.

4.2 Pre-processing Steps

Missing value treatment using mean imputation.

Normalization/standardization using Min-Max scaling.

Outlier removal using IQR method for glucose and BMI values.

Feature selection based on correlation matrix to retain the most impactful variables.

4.3 K-Means Clustering Algorithm

K-Means was selected due to:

Computational efficiency

Ease of implementation

Ability to produce distinct clusters

The algorithm partitions data into k clusters based on the Euclidean distance between data points and cluster centroids.

4.4 Determining Optimal Number of Clusters

Two approaches were used:

Elbow Method

Silhouette Score Analysis

Both methods suggested that k = 3 produced the most meaningful separation:

Cluster 1: Low-risk individuals

Cluster 2: Moderate-risk individuals

Cluster 3: High-risk individuals (likely pre-diabetic)

4.5 Tools Used

Python

Pandas, NumPy

Scikit-learn

Matplotlib/Seaborn

5. RESULTS AND ANALYSIS

5.1 Cluster Interpretation

After applying K-means with k=3, the following clusters emerged:

Cluster 1 – Low Risk

Normal BMI

FBG < 95 mg/dL

HbA1c < 5.5%

Active lifestyle

Younger age group

These individuals show minimal risk and require routine monitoring.

Cluster 2 – Moderate Risk

Slightly elevated BMI

FBG between 95–110 mg/dL

HbA1c between 5.5–5.9%

Moderate activity levels

Family history present in several cases

This group benefits from lifestyle-based preventive strategies.

Cluster 3 – High Risk (Likely Pre-diabetic)

BMI > 30

FBG > 110 mg/dL

HbA1c > 6.0%

Low physical activity

Age > 45

High correlation with metabolic syndrome indicators

These individuals require immediate clinical intervention and monitoring.

5.2 VISUALIZATION

Elbow curve and silhouette coefficients confirmed clear separation into 3 clusters.Scatter plots of HbA1c vs FBG showed distinct cluster boundaries.

6. DISCUSSION

The K-means clustering model effectively identified hidden subgroups within the dataset based on metabolic health indicators. Unlike supervised classification, this approach does not require labeled outcomes, making it ideal for large-scale community health screening.

The model's cluster interpretations align closely with clinical definitions of pre-diabetes. It also reveals that age, BMI, and HbA1c are strong predictors of cluster membership.

The results demonstrate the feasibility of integrating clustering-based models into digital health tools, such as early screening apps or hospital decision-support systems.

7. Clinical Decision Support Framework

Based on cluster membership, the following recommendations were proposed:

Low Risk → Routine yearly screening

Moderate Risk → Lifestyle modifications, dietary counseling

High Risk → Immediate clinical evaluation, frequent monitoring, preventive medication if required

This system can help clinicians prioritize patient management and resource allocation.

8. CONCLUSION

This study demonstrates the effectiveness of K-Means clustering as an unsupervised tool for early detection and risk stratification of pre-diabetes. The algorithm provides clear segmentation of individuals into low, moderate, and high-risk groups using key metabolic indicators.

The model can serve as an early warning system within clinical decision-support frameworks, especially in resource-limited environments where labeled medical data may be scarce. Future work may involve integrating deep learning, hybrid clustering, and real-time patient monitoring for improved accuracy.

9. Future Scope

Expand dataset to include wearable sensor data

Use hybrid clustering (e.g., K-Means + DBSCAN)

Develop a mobile-based real-time risk prediction app

Incorporate supervised learning to validate unsupervised results

Apply model to rural and urban populations for comparative studies.

10. REFERENCES

1. American Diabetes Association. (2024). Standards of Medical Care in Diabetes—2024. Diabetes Care, 47(Supplement 1), S1–S202.
2. Alberti, K. G., & Zimmet, P. Z. (2018). Definition, diagnosis and classification of diabetes mellitus and its complications. Diabetic Medicine, 15(7), 539–553.

3. Centers for Disease Control and Prevention (CDC). (2023). National Diabetes Statistics Report. U.S. Department of Health & Human Services.
4. Singh, R., Tripathi, A., & Kumar, S. (2022). Machine learning-based early prediction of diabetes using clinical data. International Journal of Medical Informatics, 162, 104770.
5. Pirbaglou, M., Katz, J., & Ritvo, P. (2019). Predictive modeling for early diagnosis of prediabetes: A machine learning approach. Journal of Diabetes Research, 2019, 1–10.
6. Wu, Y., Ding, Y., Tanaka, Y., & Zhang, W. (2020). Risk prediction of type 2 diabetes using machine learning algorithms. Scientific Reports, 10(1), 1–10.
7. Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666.
8. Lloyd, S. (1982). Least squares quantization in PCM. IEEE Transactions on Information Theory, 28(2), 129–137.
9. Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on Neural Networks, 16(3), 645–678.
10. Kaur, H., & Kumari, V. (2021). Predictive analytics for diabetes mellitus using machine learning techniques. Materials Today: Proceedings, 46, 324–330.
11. Zhang, L., Wang, Y., & Wang, Y. (2022). An unsupervised approach for metabolic risk stratification using K-means clustering. Journal of Biomedical Informatics, 128, 104053.
12. World Health Organization (WHO). (2022). Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia. WHO Press.