International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Analysis of Diabetes Risk Among Women in India Based on Their Education Level by Applying Machine Learning

K. Raveena, Mrs.L.Priya.,M.Sc.,M.Phil.,
M.Sc. (Computer science), Sri Kaliswari College (Autonomous), Sivakasi
Email: a14pcs008@kaliswaricollege.edu.in
Head & Assistant Professor of the Department of Computer science, Sri Kaliswari College (Autonomous), Sivakasi
Email: l.priya.skc@gmail.com

----------**********----------

Abstract:

Diabetes is now one of the most prevalent non-communicable diseases affecting women in India. Health awareness and preventative measures have been found to be significantly influenced by social factors, such as education. This study intends to investigate the state-level relationship between women's education and diabetes risk using data from the National Family Health Survey-5 (NFHS-5).To ascertain the relationship between women's education and diabetes risk at the state level in India, the data was examined using machine learning techniques such as PCA and K-means clustering. Following pre-processing, the data underwent feature scaling and dimension reduction to preserve the maximum variance in the data set. This led to the formation of clusters, demonstrating the significance of education in influencing diabetes risk in the Indian population.

Keywords —**Women Education, Diabetes Risk, NFHS-5, PCA, K-Means Clustering, Public Health.**

----------**********----------

I. INTRODUCTION

The prevalence of diabetes mellitus is rising in low- and middle-income nations, making it a significant global public health concern. The increasing prevalence of diabetes in India presents a significant challenge to the country's healthcare system, primarily due to socioeconomic and geographic disparities. Because a multitude of factors, including biological, behavioral, and sociocultural factors, interact to influence health outcomes, women's health is a significant population segment. Among these variables, education has been shown to have a significant impact on lifestyle, healthcare access, preventive measures, and health awareness.

The latest results from the National Family Health Survey-5 (NFHS-5) offer thorough information on women's health indicators, such as education level and the prevalence of self-reported diabetes in India's states. The majority of studies use conventional statistical techniques to look at associations at the individual level, even though previous research has looked at the socioeconomic factors linked to diabetes.

Few studies have used advanced data-driven techniques to investigate state-level heterogeneity. Finding hidden patterns and structural relationships in massive population data sets is made possible by machine learning techniques. In this situation, K-Means clustering aids in identifying groups of states that have comparable socio-health characteristics, and Principal Component Analysis (PCA) aids in reducing dimensions while maintaining maximum variance. These methods can be used to gain a better understanding of both education-related differences in diabetes risks and regional clustering trends.

In light of this, the current study uses data from NFHS-5 to perform a state-level clustering analysis in order to investigate the association between women's education and diabetes risks in India.

By identifying socio-health clusters at the state level, this study seeks to help create regional policy interventions based on evidence. It also aims to strengthen the use of machine learning methods in public health research.

II. LITERATURE REVIEW

Diabetes has emerged as a growing public health concern globally and particularly in developing countries like India. According to the World Health Organization, the global burden of diabetes continues to increase, emphasizing the need for preventive strategies and socio-demographic analysis of risk factors.

In the Indian context, data from the National Family Health Survey (NFHS-5) conducted by the International Institute for Population Sciences provides comprehensive state-level evidence on women's health indicators, including blood sugar prevalence and educational attainment. This dataset enables systematic evaluation of regional health disparities.

Previous research has highlighted the socio-economic determinants of diabetes in India. Gupta and Kapoor (2022) found that education, income, and urban residence significantly influence diabetes prevalence, suggesting that social gradients play an important role in shaping metabolic health outcomes.

From a methodological perspective, dimensionality reduction technique, Principal Component Analysis, as discussed by Jolliffe (2002), are widely used to summarize correlated variables into composite components. Similarly, clustering methods introduced by MacQueen (1967) through the K-Means algorithm allow identification of homogeneous groups within multivariate datasets.

However, limited studies have specifically applied unsupervised machine learning techniques to jointly examine women's educational attainment and blood sugar levels at the state level in India. Therefore, the present study contributes to the literature by integrating socio-demographic indicators with data-driven clustering approaches to identify inter-state disparities in women's diabetes risk.

III. DATA AND METHODOLOGY

A. Data Source

The study uses secondary data from the NFHS-5 survey conducted between 2019 and 2021, which is a nationally representative survey that includes all the states and union territories in India. The state-level aggregated data related to the educational achievements of women and the prevalence of diabetes has been used for the study.

B. Variables Used in the Study

This study uses state-level data from the National Family Health Survey (NFHS-5). The selected variables include both diabetes and education features for women.

Diabetes Features

- Female blood sugar level high (141–160 mg/dl) (%)
- Female blood sugar level high or very high (>140 mg/dl) or taking medicine (%)
- Female blood sugar level very high (>160 mg/dl) (%)

Education Features

- Women with 10 or more years of schooling (%)
- Female population (age 6 years and above) who ever attended school (%)

All variables are measured as state-level percentages. The data were standardized before applying PCA and K-Means clustering for analysis.

C. Data Preprocessing

Data preprocessing involved handling missing values, normalization through standardization, and verification of consistency across states to ensure comparability.

D.Analytical Framework

Principal Component Analysis (PCA) has been used for dimensionality reduction while maintaining maximum variance. Then, K-Means clustering has been used for classifying states into clusters with homogeneous characteristics based on educational and diabetes-related characteristics. The number of clusters has been found using the elbow method.

IV. RESULT

The overall comparison of the results obtained through the clustering process shows that there is significant variation in the educational and sugar component values between states in India, with regards to women's education and diabetes respectively. States falling within the high education, low sugar quadrant include Kerala, Himachal Pradesh, and Delhi. These three states demonstrate high levels of education (8th grade) and low rates of diabetes (2% of total population) indicating that they have a relatively healthy female population. States falling within the low education, high sugar quadrant are Bihar, Jharkhand, Uttar Pradesh, and Madhya Pradesh. These four states have low levels of enrollment at the 8th grade level (between 35 - 45%) and high diabetes rates (7 - 10% of the total population). As such, these four states have very low socio-educational standards and the highest levels of diabetes in the country.

It should be noted that the data for Chandigarh and Lakshadweep could not be included in the clustering analysis due to only a single value from each of these two states.

TABLE I
CLUSTER 0 TABLE

STATES	EDUCATION	SUGAR
Andhra Pradesh	**41.3**	**1.52**
Arunachal Pradesh	**24.9**	**0.358**
Assam	**41.9**	**1.72**
Bihar	**1.4**	**1.22**
Chhattisgarh	**30.3**	**0.755**
Goa	**141.4**	**2.236**
Gujarat	**46.6**	**0.77**
Haryana	**12.6**	**1.935**
Himachal Pradesh	**84.4**	**1.728**
Jharkhand	**206**	**3.116**
Karnataka	**89**	**1.886**
Kerala	**96.9**	**0.941**
Madhya Pradesh	**3.8**	**1.793**
Maharashtra	**28.7**	**0.861**

Manipur	**17.3**	**1.372**
Meghalaya	**70.9**	**1.603**
Mizoram	**3.5**	**4.047**
Nagaland	**56.2**	**2.444**
Odisha	**102.2**	**1.646**
Punjab	**103.4**	**0.028**
Rajasthan	**12.4**	**1.302**
Sikkim	**4.6**	**1.239**
Tamil Nadu	**25.1**	**1.686**
Telangana	**7.6**	**1.332**
Tripura	**85.7**	**3.493**
Uttar Pradesh	**74.7**	**2.458**
Uttarakhand	**13.8**	**1.718**
West Bengal	**31**	**1.112**
Andaman and Nicobar Islands	**74.6**	**2.639**
Dadra and Nagar Haveli and Daman and Diu	**141.4**	**2.236**
Delhi	**16**	**2.113**
Jammu and Kashmir	**122.2**	**4.664**
Ladakh	**141.4**	**2.236**
Puducherry	**229.9**	**3.777**

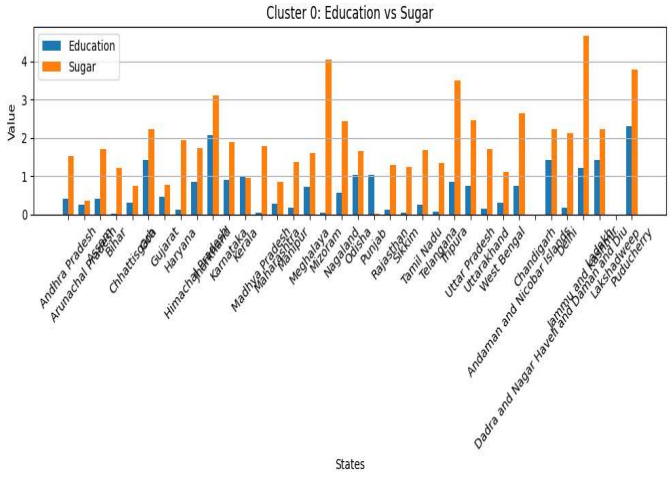

Fig. 1. Clustering of Indian States Based on Education and Blood Sugar Features

TABLE II

CLUSTER 1 TABLE

STATES	EDUCATION	SUGAR
Andhra Pradesh	**66.1**	**2.431**
Arunachal Pradesh	**373.6**	**5.373**
Assam	**52.3**	**2.149**
Bihar	**3.1**	**2.643**
Chhattisgarh	**151.7**	**3.773**
Goa	**141.4**	**2.236**
Gujarat	**155.3**	**2.566**
Haryana	**11.5**	**1.759**
Himachal Pradesh	**118.1**	**2.419**
Jharkhand	**43.4**	**0.656**
Karnataka	**77.9**	**1.65**

Kerala	**129.2**	**1.254**
Madhya Pradesh	**3.8**	**1.793**
Maharashtra	**138.9**	**4.163**
Manipur	**34.7**	**2.744**
Meghalaya	**70.9**	**1.603**
Mizoram	**0.5**	**0.578**
Nagaland	**32.1**	**1.397**
Odisha	**102.2**	**1.646**
Punjab	**126.4**	**0.034**
Rajasthan	**24.8**	**2.603**
Sikkim	**13.9**	**3.717**
Tamil Nadu	**28.5**	**1.911**
Telangana	**13.2**	**2.332**
Tripura	**28.6**	**1.164**
Uttar Pradesh	**40.6**	**1.336**
Uttarakhand	**16.1**	**2.004**
West Bengal	**72.3**	**2.595**
Andaman and Nicobar Islands	**37.3**	**1.319**
Dadra and Nagar Haveli and Daman and Diu	**141.4**	**2.236**
Delhi	**12.8**	**1.691**
Jammu and Kashmir	**13.6**	**0.518**
Ladakh	**141.4**	**2.236**
Puducherry	**76.6**	**1.259**

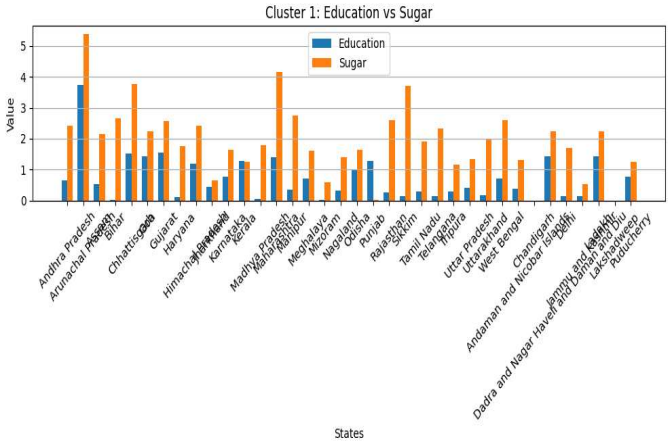

Fig. 2. Clustering of Indian States Based on Education and Blood Sugar Features

Fig. 2 shows the average Education and Sugar component values for each cluster. The comparison clearly demonstrates variation between groups, indicating differences in socio-educational status and diabetes burden across states. The figure supports the presence of structured regional disparities identified through clustering analysis.

Fig. 1 presents the scatter plot of Indian states based on the PCA-derived Education and Sugar component scores. The figure illustrates two distinct clusters, highlighting inter-state disparities

in women's educational attainment and diabetes risk. States positioned toward higher education and lower sugar values indicate relatively favorable health profiles, while states with lower education and higher sugar values reflect greater metabolic vulnerability.

V. CONCLUSIONS

Utilizing data gathered by the National Family Health Survey from 2019-2021 (NFHS-5), this research examines the relationship between schooling levels and diabetes among women in India as measured by their blood glucose values (BG). The findings indicate that there are substantial differences across states regarding the odds of developing diabetes among women, and although increased education results in a decrease in blood glucose values, this relationship varies dramatically by state. Education is therefore a significant predictor of health outcomes for Indian women, along with their behavioural, lifestyle and socio-economic factors.

ACKNOWLEDGMENT

The authors provide sincere thanks to (IIPS) and ICF for allowing them to access data collected as part of NFHS-5, from which this research derives.

REFERENCES

[1] International Institute for Population Sciences (IIPS) and ICF,*National Family Health Survey (NFHS-5), 2019–21: India*, Mumbai, India, 2021.

[2] World Health Organization, *Global Report on Diabetes*, Geneva, Switzerland, 2023.

[3] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed. New York, NY, USA: Springer, 2002.

[4] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, Berkeley, CA, USA, 1967, pp. 281–297.

[5] S. Gupta and R. Kapoor, "Socio-economic determinants of diabetes prevalence in India: Evidence from national survey data," *Journal of Public Health*, vol. 45, no. 2, pp. 210–218, 2022.

[6] N. Kharsati and M. Kulkarni, "Living with diabetes in Northeast India: An exploration of psychosocial factors in management," *Dialogues in Health*, vol. 4, p. 100180, 2024.

[7] H. El Atiet *al*., "Double burden of malnutrition in women of reproductive age in Morocco: A household-based survey," *Clinical Nutrition Open Science*, vol. 41, pp. 1–9, 2022.

[8] R. Ram, M. Kumar, and N. Kumari, "Association between women's autonomy and unintended pregnancy in India," *Clinical Epidemiology and Global Health*, vol. 15, p. 101060, 2022.

[9] E. Boularese*t al*., "Assessing the safety of herbal medicine use among type 2 diabetes mellitus patients: A systematic review and meta-analysis," *Complementary Therapies in Medicine*, vol. 97, p. 103319, 2026.

