STAT 515

Applied Statistics & Visualization for Analytics

DIABETES PREDICTORS

Venkata Lakshmi Prasuna Sai Poojitha Anuvala Setty

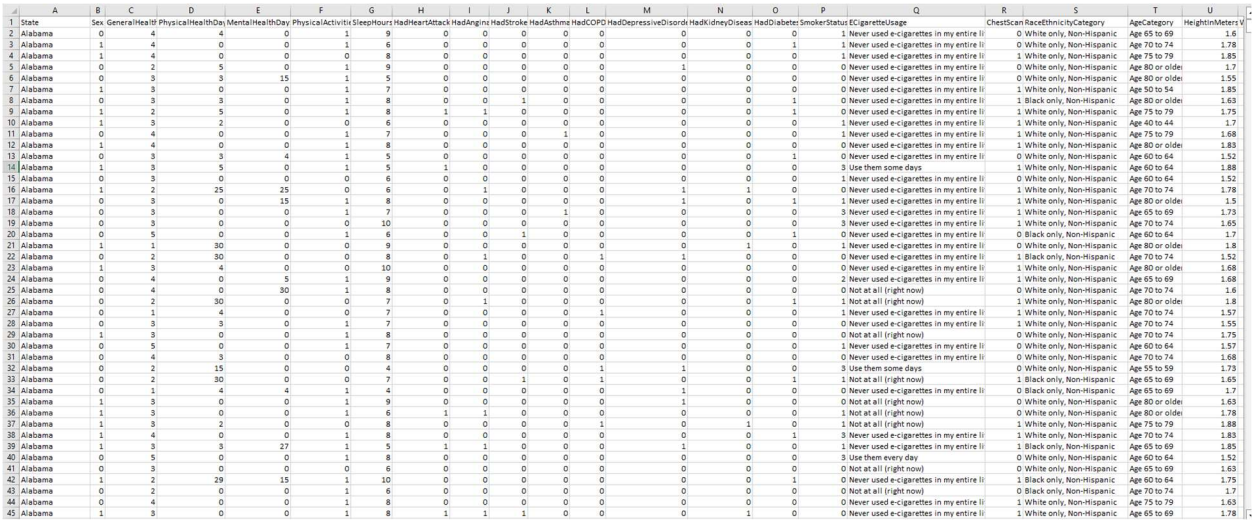
G01421210

I. INTRODUCTION

Diabetes is the eighth leading cause of death in the United States in 2020. This epidemic extends beyond personal health, putting a pressure on healthcare systems and economy. This research aims to find the predictors of diabetes and create targeted interventions for groups in need. This dataset is derived from the Behavioral Risk Factor Surveillance System (BRFSS). It is a national system of health-related telephone surveys that collects data about the residents of U.S regarding their health conditions and habits. It conducts more than 400,000 adult interviews each year. Responses to a person's lifestyle factors and health conditions are extracted into a dataset and are analyzed.

II. DATASET

The research relies on a detailed collection of information that covers a wide range of aspects related to diabetes. This dataset includes details about people's age, gender, habits, and health history. By examining this data closely, we aim to uncover patterns and connections that can help us understand why diabetes occurs.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	State	Sex	GeneralHealt	PhysicalHealthDay	MentalHealthDay	PhysicalActivitie	SleepHours	HadHeartAttack	HadAngina	HadStroke	HadAsthma	HadCOPD	HadDepressiveDisorde	HadKidneyDisease	HadDiabetes	SmokerStatus	ECigaretteUsage	ChestScan	RaceEthnicityCategory	AgeCategory	HeightInMeters
2	Alabama	0	4	4	0	1	9	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 65 to 69	1.6
3	Alabama	1	4	0	0	1	6	0	0	0	0	0	0	0	1	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 70 to 74	1.78
4	Alabama	1	4	0	0	0	8	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 75 to 79	1.85
5	Alabama	0	2	5	0	1	9	0	0	0	0	0	1	0	0	0	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 80 or olde	1.7
6	Alabama	0	3	3	15	1	5	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 80 or olde	1.55
7	Alabama	1	3	0	0	1	7	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 50 to 54	1.85
8	Alabama	0	3	3	0	1	8	0	0	1	0	0	0	0	1	0	Never used e-cigarettes in my entire li	1	Black only, Non-Hispanic	Age 80 or olde	1.63
9	Alabama	1	2	5	0	1	8	1	1	0	0	0	0	0	1	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 75 to 79	1.75
10	Alabama	1	3	2	0	0	6	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 40 to 44	1.7
11	Alabama	0	4	0	0	1	7	0	0	0	1	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 75 to 79	1.68
12	Alabama	1	4	0	0	1	8	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 80 or olde	1.83
13	Alabama	0	3	3	4	1	5	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 60 to 64	1.52
14	Alabama	1	3	5	0	1	5	1	0	0	0	0	0	0	0	3	Use them some days	1	White only, Non-Hispanic	Age 60 to 64	1.88
15	Alabama	0	3	0	0	0	6	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 60 to 64	1.52
16	Alabama	1	2	25	25	0	6	0	1	0	0	0	1	1	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 70 to 74	1.78
17	Alabama	0	3	0	15	1	8	0	0	0	0	0	1	0	1	1	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 80 or olde	1.5
18	Alabama	0	3	0	0	1	7	0	0	0	1	0	0	0	0	3	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 65 to 69	1.73
19	Alabama	0	3	0	0	0	10	0	0	0	0	0	0	0	0	3	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 70 to 74	1.65
20	Alabama	0	5	0	0	1	6	0	0	1	0	0	0	0	1	0	Never used e-cigarettes in my entire li	0	Black only, Non-Hispanic	Age 60 to 64	1.7
21	Alabama	1	1	30	0	0	9	0	0	0	0	0	0	1	0	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 80 or olde	1.8
22	Alabama	0	2	30	0	0	8	0	1	0	0	1	1	0	0	0	Never used e-cigarettes in my entire li	1	Black only, Non-Hispanic	Age 70 to 74	1.52
23	Alabama	1	3	4	0	0	10	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 80 or olde	1.68
24	Alabama	0	4	0	5	1	9	0	0	0	0	0	0	0	0	2	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 65 to 69	1.68
25	Alabama	0	4	0	30	1	8	0	0	0	0	0	0	0	0	0	Not at all (right now)	0	White only, Non-Hispanic	Age 70 to 74	1.6
26	Alabama	0	2	30	0	0	7	0	1	0	0	0	0	0	1	1	Not at all (right now)	1	White only, Non-Hispanic	Age 80 or olde	1.8
27	Alabama	0	1	4	0	0	7	0	0	0	0	1	0	0	0	1	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 70 to 74	1.57
28	Alabama	0	3	3	0	1	7	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 70 to 74	1.55
29	Alabama	1	3	0	0	1	8	0	0	0	0	0	0	0	0	0	Not at all (right now)	0	White only, Non-Hispanic	Age 70 to 74	1.75
30	Alabama	0	5	0	0	1	7	0	0	0	0	0	0	0	0	1	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 60 to 64	1.57
31	Alabama	0	4	3	0	0	8	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	0	White only, Non-Hispanic	Age 70 to 74	1.68
32	Alabama	0	2	15	0	0	4	0	0	0	0	1	1	0	0	3	Use them some days	0	White only, Non-Hispanic	Age 55 to 59	1.73
33	Alabama	0	2	30	0	0	7	0	0	1	0	1	0	0	1	1	Not at all (right now)	1	Black only, Non-Hispanic	Age 65 to 69	1.65
34	Alabama	0	1	4	4	1	4	0	0	0	0	0	0	1	0	0	Never used e-cigarettes in my entire li	0	Black only, Non-Hispanic	Age 65 to 69	1.7
35	Alabama	1	3	0	0	1	9	0	0	0	0	0	0	1	0	0	Not at all (right now)	0	White only, Non-Hispanic	Age 80 or olde	1.63
36	Alabama	1	3	0	0	1	6	1	1	0	0	0	0	0	0	1	Not at all (right now)	0	White only, Non-Hispanic	Age 80 or olde	1.78
37	Alabama	1	3	2	0	0	8	0	0	0	0	1	0	0	1	1	Not at all (right now)	1	White only, Non-Hispanic	Age 75 to 79	1.88
38	Alabama	1	4	0	0	1	8	0	0	0	0	0	0	0	1	3	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 70 to 74	1.83
39	Alabama	1	3	3	27	1	5	1	1	0	0	0	1	0	0	1	Never used e-cigarettes in my entire li	1	Black only, Non-Hispanic	Age 65 to 69	1.85
40	Alabama	0	5	0	0	1	8	0	0	0	0	0	0	0	0	3	Use them every day	0	White only, Non-Hispanic	Age 60 to 64	1.52
41	Alabama	0	3	0	0	0	6	0	0	0	0	0	0	0	0	0	Not at all (right now)	0	White only, Non-Hispanic	Age 65 to 69	1.63
42	Alabama	1	2	29	15	1	10	0	0	0	0	0	0	0	1	0	Never used e-cigarettes in my entire li	1	Black only, Non-Hispanic	Age 60 to 64	1.75
43	Alabama	0	2	0	0	1	6	0	0	0	0	0	0	0	0	0	Not at all (right now)	0	Black only, Non-Hispanic	Age 70 to 74	1.7
44	Alabama	0	4	0	0	1	8	0	0	0	0	0	0	0	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 75 to 79	1.63
45	Alabama	1	3	0	0	1	8	1	1	1	0	0	0	1	0	0	Never used e-cigarettes in my entire li	1	White only, Non-Hispanic	Age 65 to 69	1.78

Fig.1: Snapshot of the dataset

Column Name	Description	Column Name	Description
State	State in which the respondent resides	HadAsthma	0 = No 1 = Yes
Sex	0 = female 1 = male	HadCOPD	0 = No 1 = Yes
GeneralHealth	Would you say that in general your health is on a scale of 1-5: 1 = poor 2 = fair 3 = good 4 = very good 5 = excellent	HadDepressiveDisorder	0 = No 1 = Yes
PhysicalHealthDays	Physical illness or injury days in the past 30 days	HadKidneyDisease	0 = No 1 = Yes
MentalHealthDays	Poor mental health days in the past 30 days	HadDiabetes	0 = No 1 = Yes
LastCheckupTime	About how long has it been since you last visited a doctor for a routine checkup?	SmokerStatus	Four-level smoker status
PhysicalActivities	During the past month, other than your regular job, did you participate in any physical activities or exercises? 0 = No 1 = Yes	AgeCategory	Fourteen-level age category

SleepHours	On average, how many hours of sleep do you get in a 24-hour period?	HeightInMeters	Reported height in meters
HadHeartAttack	(Ever told) you had a heart attack, also called a myocardial infarction? 0 = No 1 = Yes	WeightInKgs	Reported weight in kilograms
HadAngina	0 = No 1 = Yes	BMI	Body Mass Index (BMI)
HadStroke	0 = No 1 = Yes	AlcoholDrinkers	0 = No 1 = Yes

Table 1: Column Names and their Descriptions

III. EXPLORATORY DATA ANALYSIS (EDA)

To understand the dataset, it is important to visualize and interpret data patterns. It forms the basis for informed decision-making and make accurate data-driven insights.

This choropleth map uses data from the dataset to represent the occurrence of diabetes in various states. Each state is color-coded according to the percentage of diabetics, ranging from light blue (lower percentages) to dark blue (greater percentages). The legend on the right-side aids in the interpretation of the color scale, and the map's title is "Percentage of Diabetes by State." This visualization provides a spatial perspective on the prevalence of diabetes, assisting in the identification of regional patterns and variances.

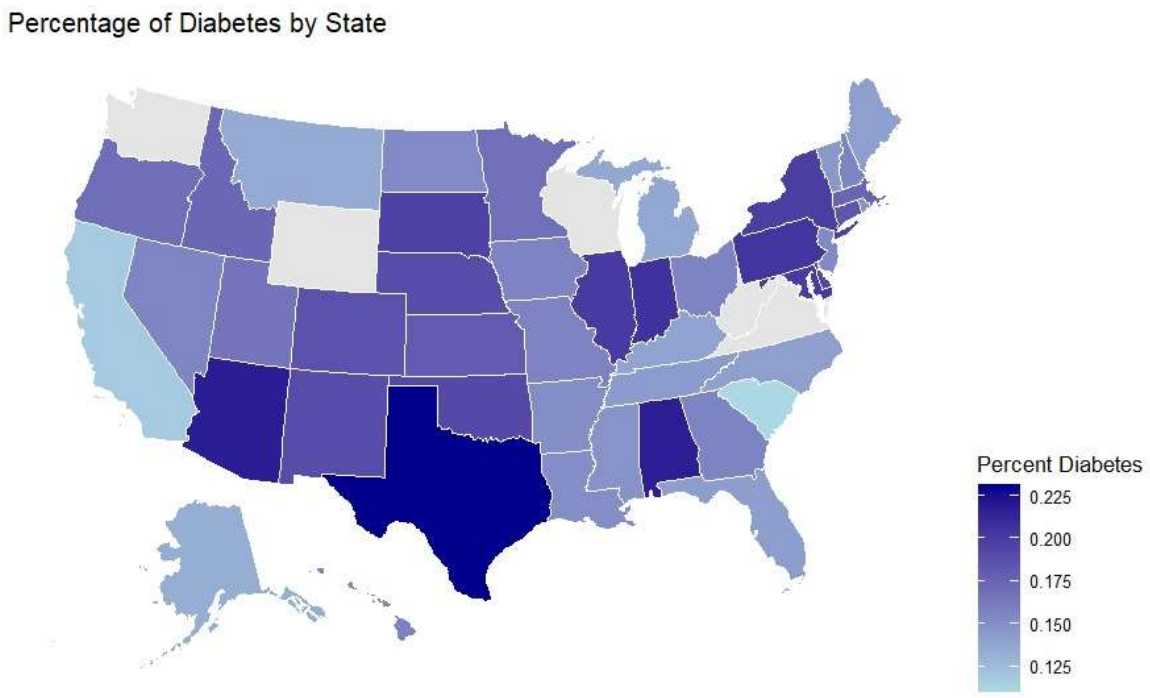


Fig. 2: Percentage of people with diabetes by state according to the survey

The density plot illustrates the distribution of age and body mass index (BMI) by diabetes status. In the age distribution, the prevalence of diabetes rises with age, peaking at 25% for those aged 75-79, indicating a strong association between age and diabetes. Additionally, the plot highlights differences in age distribution between individuals with and without diabetes, emphasizing the older age of those with diabetes. In the BMI distribution, people with diabetes tend to have higher BMI, with a median of 35.26 compared to 28.89 for those without diabetes. This underscores the link between diabetes and overweight or obesity. The density plots provide valuable insights into age and BMI patterns related to diabetes prevalence in the dataset.

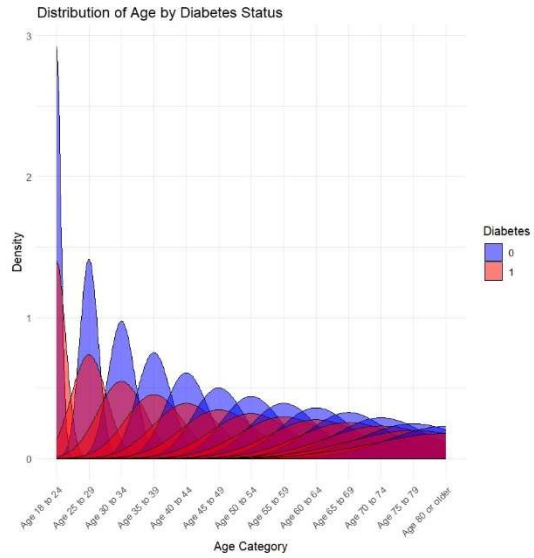


Fig. 3: Distribution of Age by Diabetes Status

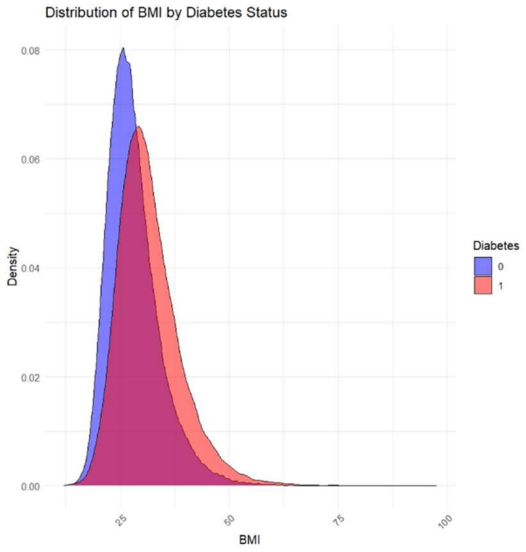


Fig. 4: Distribution of BMI by Diabetes Status

The correlation matrix reveals relationships between key health indicators and the prevalence of diabetes in the dataset. Notably, lower self-rated general health shows a moderate negative correlation with diabetes, suggesting a potential link. People experiencing more days of physical health issues exhibit a slight positive correlation with diabetes. Engaging in more physical activities is weakly correlated with a lower likelihood of diabetes. Higher BMI and weight demonstrate weak positive correlations with diabetes, indicating a potential connection between body composition and diabetes history.

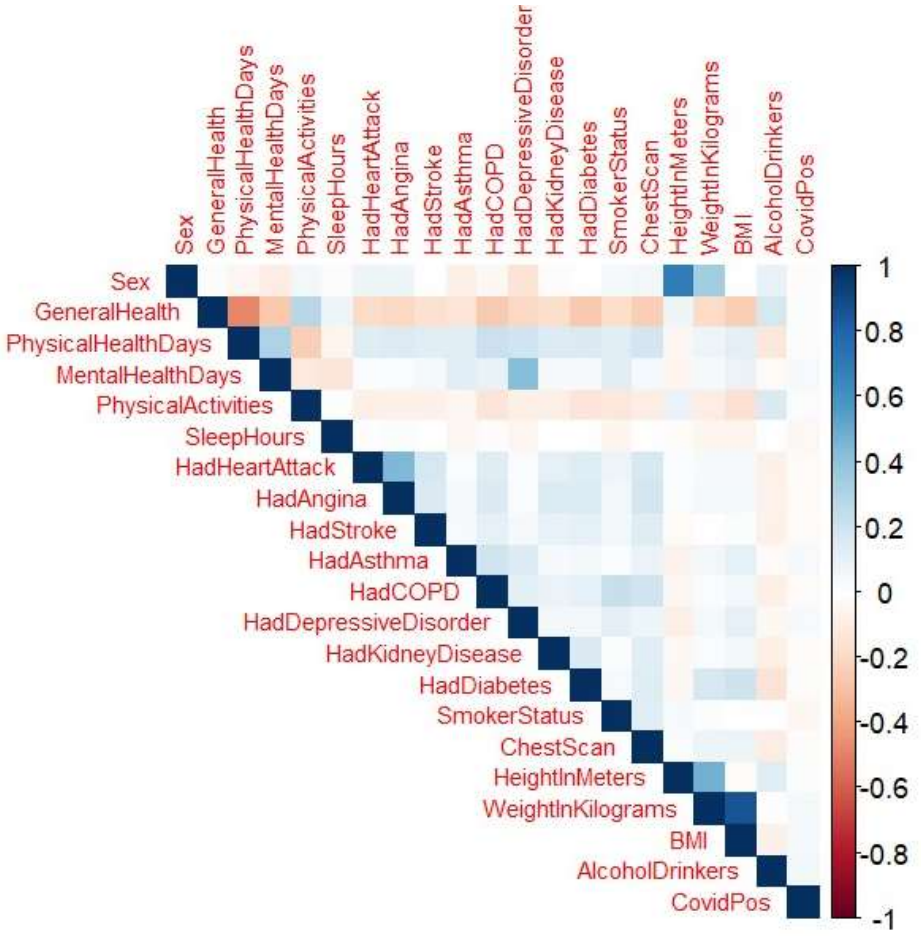


Fig. 5: Correlation Matrix

IV. MODELS

i) Logistic Regression

A logistic model is a statistical method used for binary classification problems, such as predicting whether an event will occur or not. It's particularly useful when the outcome variable is categorical with two possible outcomes. The logistic regression model, designed to predict whether individuals have diabetes or not based on various health indicators, performs reasonably well with an accuracy of around 83.8%. This means it accurately identified individuals with and without diabetes in the test dataset most of the time. The confusion matrix provides a detailed breakdown, showing the number of correct and incorrect predictions. The model

correctly identified a large number of individuals without diabetes but had some misclassifications, especially in predicting individuals with diabetes. The ROC curve visually summarizes the model's overall performance, offering insights into its ability to make accurate predictions. Further adjustments and fine-tuning could enhance the model's effectiveness.

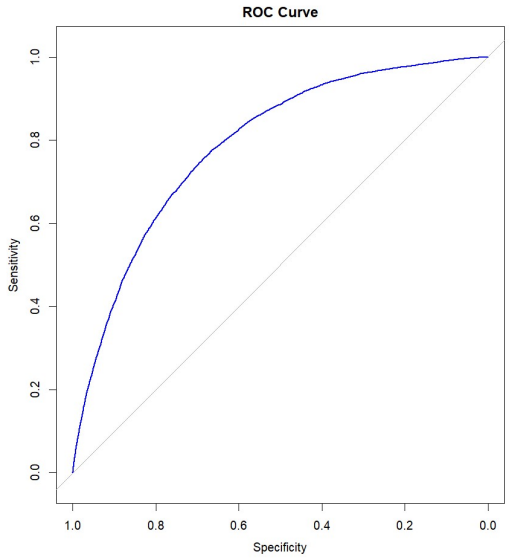


Fig. 6: ROC Curve

ii) K-Means Clustering

K-means clustering is a data segmentation technique dividing individuals into distinct groups based on similar attributes. In our analysis, K-means revealed two clusters with notable health differences. The High Prevalence Cluster (Cluster 3) comprises mainly females reporting lower general health, higher chronic conditions, and a 28.9% diabetes prevalence. Conversely, the Low Prevalence Cluster (Cluster 1), primarily males, exhibits better health indicators, including a 10.9% diabetes prevalence. K-means clustering assists in identifying subpopulations with varying health profiles, informing targeted interventions to address diabetes and related health issues within specific demographic clusters.

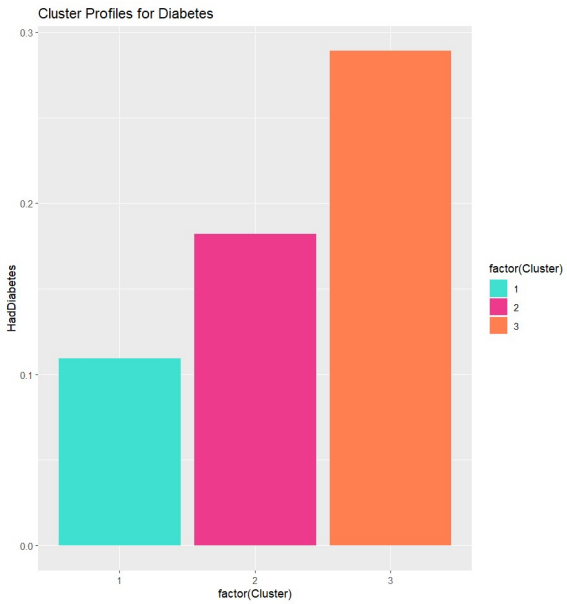


Fig. 7: Cluster Profiles for Diabetes

V. CONCLUSION

The project uncovered important patterns related to diabetes. It showed that as people get older, they are more likely to have diabetes. Factors like body mass index (BMI) and overall health are also linked to diabetes. By grouping similar individuals, the study identified a cluster with a higher diabetes prevalence, mainly in older people with lower general health and higher BMI. The analysis confirmed these connections using statistical models. Overall, the research emphasizes that age and health play crucial roles in diabetes, providing valuable insights for managing and preventing the condition in the studied population.

VI. REFERENCES

[1] "BRFSS Survey Data 2022", August, 2023
Available:https://www.cdc.gov/brfss/annualdata/annual2022.html

[2] Y. Du *et al.*, "Technology-Assisted Self-Monitoring of Lifestyle Behaviors and Health Indicators in Diabetes: Qualitative study," *JMIR Diabetes*, vol. 5, no. 3, p. e21183, Aug. 2020, doi: 10.2196/21183.