Exploratory Analysis of Diabetes Health Indicators

October 2, 2025

1 Diabetes Health Indicators Analysis

1.1 Background

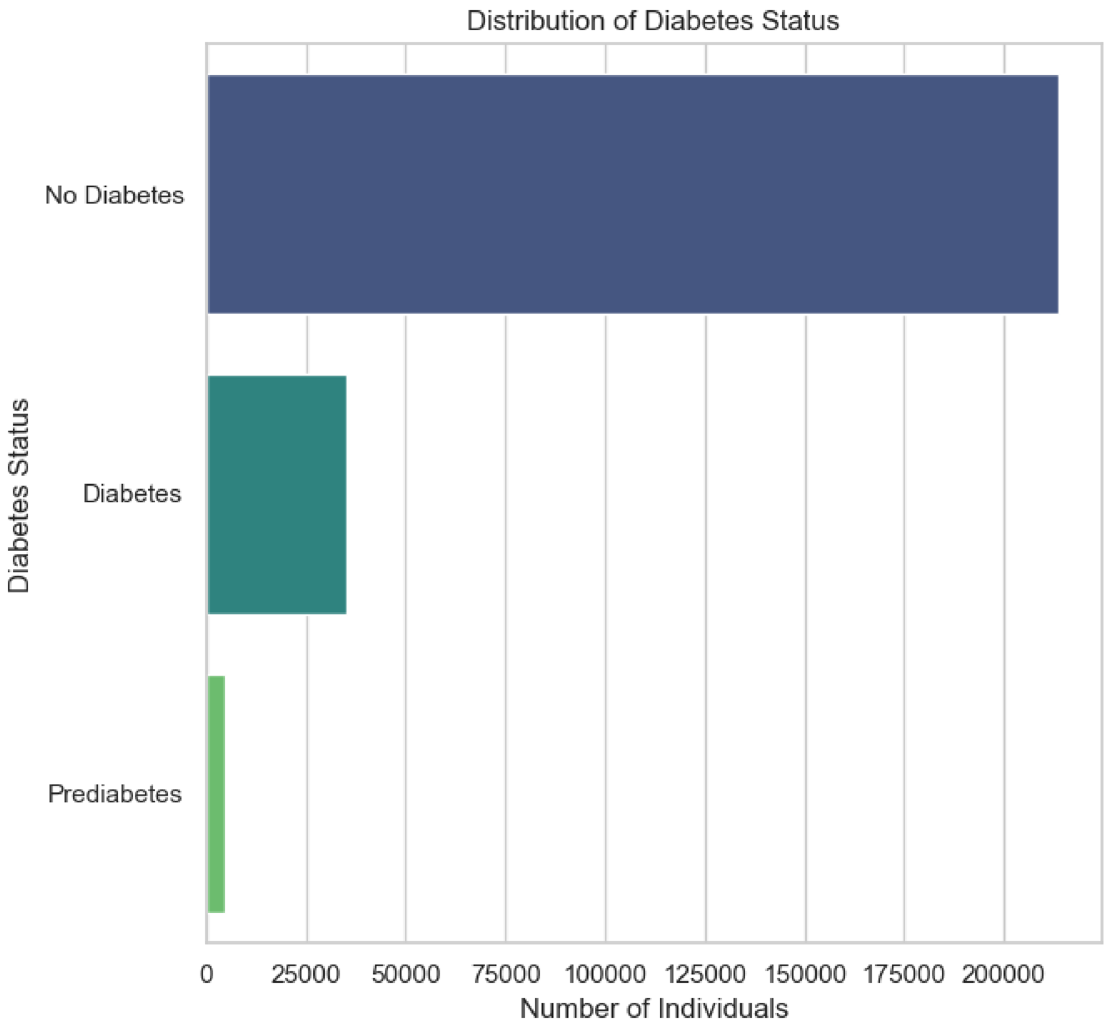
Diabetes is a chronic health condition affecting millions of people worldwide. This project analyzes a dataset of health indicators to understand factors associated with diabetes prevalence and risk.

1.2 Problem Definition

This analysis aims to:

1. Identify which health indicators are most strongly associated with diabetes status
2. Examine how demographic factors correlate with diabetes risk
3. Explore relationships between modifiable risk factors and diabetes
4. Suggest potential intervention points for diabetes prevention

The dataset reveals a significant disparity in diabetes prevalence among the studied population. The majority of subjects, approximately 85%, have no diabetes, while only about 13% have diabetes and a mere 2% have prediabetes. This distribution highlights that while diabetes affects a minority of the population, it still represents a substantial health burden given the sample size. The stark contrast between these groups provides a strong basis for comparative analysis of risk factors and demographic patterns associated with the condition.



BMI distributions show a clear relationship with diabetes status. Individuals with diabetes and prediabetes demonstrate notably higher median BMI values, approximately 30 and 29 respectively, compared to those without diabetes at approximately 25. The density plot reveals that people without diabetes have a peak BMI distribution around 25, while those with diabetes show a broader distribution with higher concentrations in the overweight (BMI 25-30) and obese (BMI greater than 30) ranges. This visualization confirms BMI as a significant risk factor, with higher values strongly associated with diabetes diagnosis.

Age emerges as a critical factor in diabetes prevalence, with a dramatic increase observed in older age groups. The graph below demonstrates that diabetes rates begin climbing noticeably after age 45, with the steepest increases in the 65-69, 70-74, 75-79, and 80+ age brackets. The visualization reveals that while diabetes affects less than 10% of adults under 45, this rate more than doubles to over 20% in the elderly population. This clear age-related progression suggests that screening and intervention strategies should be prioritized based on age, particularly for individuals entering middle age and beyond.

Individuals with diabetes show substantially higher rates of coexisting conditions. Those with diabetes have markedly elevated rates of high blood pressure (73%), high cholesterol (67%), and heart disease (22%) compared to non-diabetic individuals (38%, 35%, and 7% respectively). Interestingly, prediabetic individuals also show higher rates of these conditions than the non-diabetic group, suggesting that these conditions may develop along a continuum with prediabetes representing an intermediate risk state. These patterns underscore the interconnected nature of metabolic and cardiovascular conditions, highlighting the importance of comprehensive care approaches.

Self-reported general health shows a striking correlation with diabetes status. The graph below reveals that individuals without diabetes most frequently report “very good” health (38%), while those with diabetes more commonly report only “good” (27%) or “fair” (28%) health, with very few reporting “excellent” health (3%). This suggests that diabetes significantly impacts perceived well-being and quality of life. Prediabetic individuals show an intermediate pattern, with health ratings falling between the other two groups, further supporting the concept of prediabetes as a transitional state in terms of both physical health and subjective well-being.

The analysis of lifestyle behaviors in the bar chart below reveals meaningful differences across diabetes status groups. People without diabetes show higher rates of positive health behaviors: physical activity (78%), fruit consumption (63%), and vegetable intake (81%) compared to diabetic individuals (63%, 60%, and 75% respectively). Heavy alcohol consumption is generally low across all groups but slightly higher in those without diabetes. These patterns suggest that lifestyle modifications might be both preventive for those at risk and beneficial for those already diagnosed with diabetes or prediabetes, with particular emphasis on increasing physical activity.

The correlation matrix provides a comprehensive view of the relationships between health variables. Diabetes status shows the strongest positive correlations with general health (0.30), high blood pressure (0.27), BMI (0.22), and difficulty walking (0.22). Prior figures confirm these as the top factors associated with diabetes. Physical activity shows a negative correlation (-0.12), indicating its protective effect. This analysis reinforces the complex, interconnected nature of diabetes with various physiological, behavioral, and demographic factors, suggesting that comprehensive assessment and intervention approaches are necessary.

The prevalence of diabetes shows minimal variation between sexes. Both females and males exhibit similar patterns with approximately 13-15% having diabetes and 1-2% having prediabetes. This suggests that biological sex alone may not be a strong independent risk factor for diabetes, though interactions between sex and other risk factors could still be clinically relevant. The comparable rates across sexes indicate that diabetes prevention and management strategies should target both men and women equally, with emphasis on risk factors that transcend sex differences.

Education level demonstrates a clear inverse relationship with diabetes prevalence. The comparison below shows that individuals with higher education levels, specifically those with college education, have significantly lower diabetes rates (10% for those with 4+ years of college) compared to those with less education (27% for those who never attended school). Similarly, higher income levels are associated with lower diabetes prevalence, with rates decreasing from 25% in the lowest income bracket to just 8% in the highest. These socioeconomic trends highlight the social determinants of health and suggest that educational initiatives and economic policies could indirectly impact diabetes prevalence by addressing these underlying disparities.

The combined risk score analysis below illustrates how risk factors accumulate differently across diabetes status groups. Individuals with diabetes show significantly higher median risk scores and wider variability in their risk profiles compared to those without diabetes. This analysis suggests that diabetes is often accompanied by a constellation of risk factors rather than isolated abnormalities. The violin plot of BMI by physical activity further demonstrates how lifestyle factors interact with metabolic parameters across diabetes status groups, with physical activity associated with lower BMI distributions regardless of diabetes status.

The scatter plot exploring the relationship between age, BMI, and diabetes status reveals complex interactions between these variables. While higher BMI values are more frequently associated with diabetes regardless of age, the distribution of points suggests that the BMI threshold for diabetes risk may vary across age groups. This visualization helps identify particularly vulnerable populations, specifically those with both advanced age and elevated BMI, who may benefit most from targeted screening and intervention efforts.

2 Conclusion

This exploratory data analysis reveals diabetes as a complex condition with multiple interrelated risk factors spanning demographics, lifestyle behaviors, comorbidities, and socioeconomic indicators. The clear patterns observed across BMI distributions, age groups, comorbidity rates, and socioeconomic gradients provide valuable insights for developing targeted prevention strategies and personalized interventions. The analysis particularly highlights the importance of addressing modifiable factors such as physical activity and diet, while recognizing the influence of social determinants like education and income on diabetes risk. These findings can inform both clinical approaches to diabetes management and public health policies aimed at reducing diabetes burden in the population.