SugarStats: Using Machine Learning to Analyze Diabetes Risk
CS133 Term Project

Aaminah Mohammad, Akash Hebbar, Justine Mae Legson, Steven Dinh

San Jose State University

CS 133: Data Visualization

Professor Jessica Westfall

November 30, 2025

Table of Contents

Abstract:

This project explores diabetes risk factors using the CDC Diabetes Health Indicators dataset, including health, lifestyle, and demographic responses from over 250,000 individuals. The goal of this project is to investigate which health and lifestyle habits are linked to diabetes and to discover if we can predict diabetes using machine learning. The dataset included many features such as BMI, physical activity, mental health, income, and general health rating. We conducted exploratory data analysis on these features to look for patterns and differences between people with diabetes and those without diabetes. We then trained several machine learning models, such as Logistic Regression, Random Forest, and XGBoost. We compared them using cross-validation to see which model worked the best. Our results showed clear patterns such as higher diabetes rates in people with higher BMI, poorer health ratings, and lower physical activity. The best-performing model was chosen based on accuracy, F1 score, and AUC results. Our findings suggest that machine learning can help identify individuals who may be at risk for diabetes

Background

Diabetes is a disease where the body cannot produce enough insulin or use the insulin effectively. This can lead to elevated blood glucose levels, which over time can damage the heart, kidneys, nerves, eyes, and blood vessels (Hill-Briggs et al., 2020). There are multiple forms of diabetes, most commonly Type 1, Type 2, and prediabetes, the last of which indicates early warning signs before full disease onset. Being able to predict diabetes is important as it allows for lifestyle changes, medical intervention, and monitoring that can prevent severe complications. Catching risk early, especially in pre-diabetic individuals, can significantly reduce long-term costs, improve quality of life, and prevent diabetes from progressing (Eseadi et al., 2023).

The dataset was chosen as it had a large number of samples, no missing data or incorrect value types, and had a lot of relevant features that can be important in predicting diabetes.

The project goals were to be able to predict if someone has diabetes based on a couple of clinical tests, giving the patient quick and accurate results so the doctor can prescribe the correct course of treatment. Diabetes can go undetected for a long time and can lead to long-term health problems as well as expensive procedures for the patient. Instead, if an accurate machine learning algorithm is developed that would allow for a quick 5-10 minute added test in daily checkups and catch diabetes quicker, and would be a fraction of the cost for the patient. The dataset contains 253,680 records with 21 features (UCI Machine Learning Repository, 2015). The target variable in this dataset is Diabetes_012, which has three classes. Class 0 is used for healthy individuals, Class 1 is used for prediabetes, and Class 2 is used for diabetes. Key features in this dataset are BMI, Income, Physical Activity, General Health, and Age.

Since the dataset was already mostly clean there was not much preprocessing required. Some

of the preprocessing steps that we took were converting the target column to type int so that the values are correctly treated for machine learning models. The variable general health is coded as 1-5 in the data set, so the numeric values are mapped to descriptive categories to make the exploratory analysis, plotting, and interpretation easier and more intuitive. The age variable is coded into 13 numerical buckets, and we preprocessed so that it is converted into actual age ranges, making the analysis more interpretable since age groups can now be read as actual age ranges.

Exploratory Data Analysis

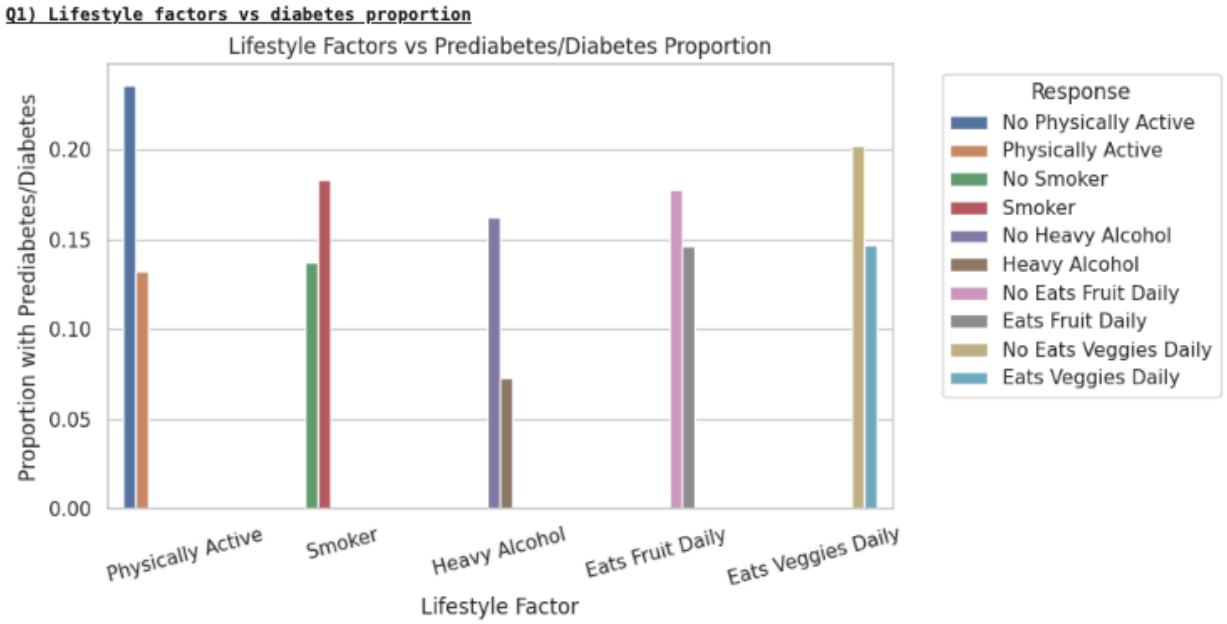
To guide our analysis, we explored five main questions using this dataset:

1. How do lifestyle habits such as smoking, physical activity, and diet relate to diabetes risk?
2. How does BMI differ between non-diabetic, prediabetic, and diabetic groups?
3. How are mental and physical unhealthy days connected to diabetes prevalence?
4. Do socioeconomic indicators such as income and education level influence diabetes risk?
5. Are individuals with healthcare access and regular medical checkups less likely to have diabetes?

These five questions structured our data exploration and led directly into model selection and prediction work.

- Q1: Lifestyle Factors vs Diabetes

  Which lifestyle factors (like diet, smoking, or physical activity) have the biggest impact on diabetes risk?



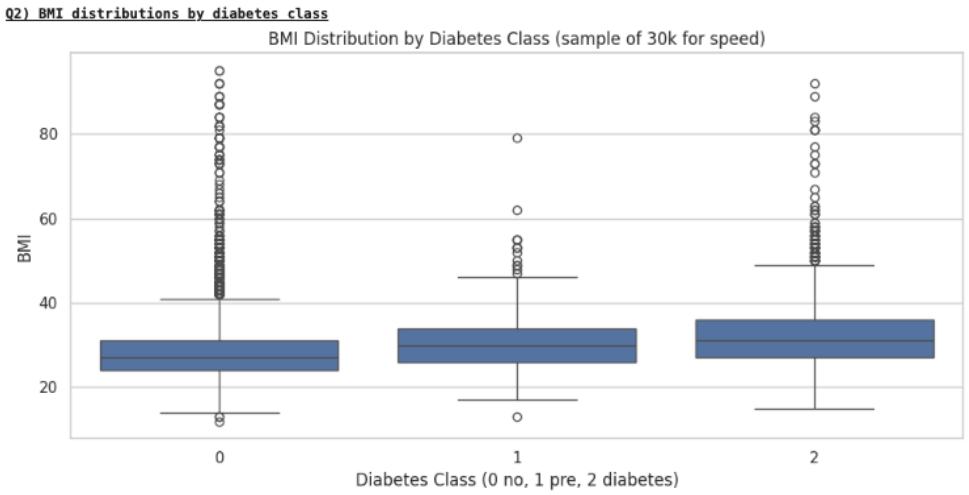
From the graph, physical activity has the biggest impact on diabetes risk. People who

don’t exercise have the highest rates of prediabetes/diabetes. Smoking also makes a difference as smokers tend to have higher diabetes rates than non-smokers. Eating fruits and veggies daily still helps, but the effect is smaller compared to exercise and smoking.

Overall, not being physically active is the strongest risk factor, followed by smoking, with diet having a smaller but still noticeable effect.

- Q2: BMI vs Diabetes Levels

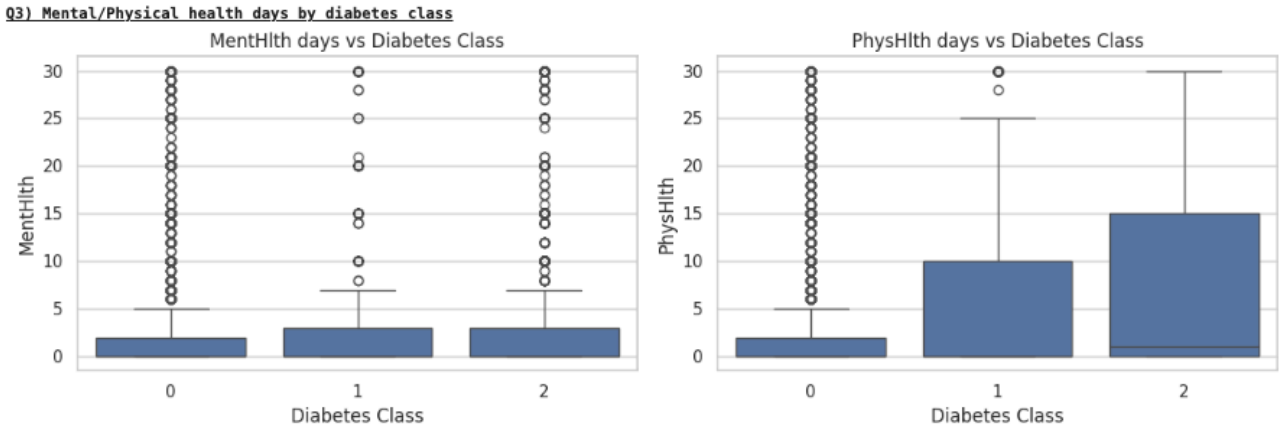
  How does BMI differ among people who are healthy, prediabetic, or diabetic?



The boxplot shows that BMI tends to be higher for people with prediabetes and diabetes compared to those without diabetes. While there’s overlap between the groups, the median BMI increases as you go from healthy (0) → prediabetic (1) → diabetic (2). This suggests that higher BMI is generally associated with a greater risk of developing diabetes.

- Q3: Mental/Physical Health vs Diabetes

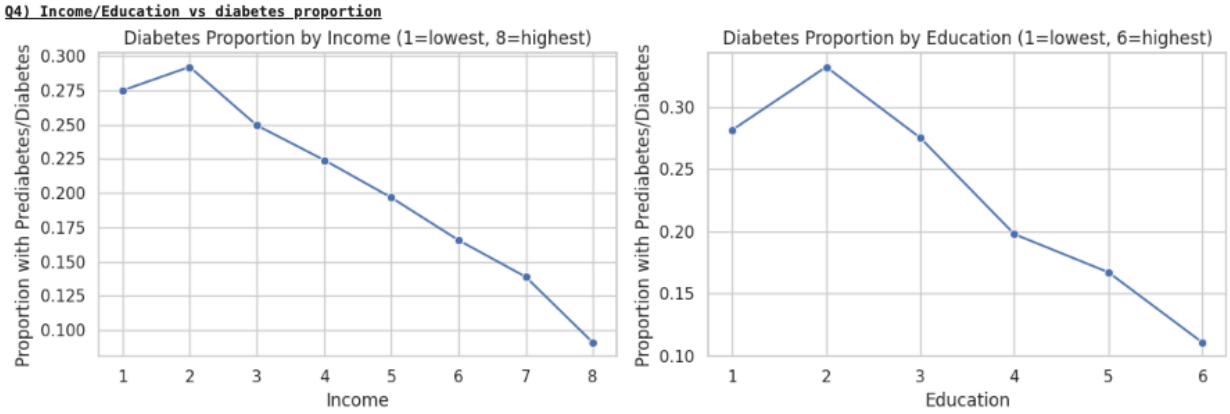
  How do the number of poor mental and physical health days differ across the diabetes classes?



People in the diabetic class tend to report more days of poor physical health compared to the healthy group, and prediabetic individuals fall in between. Mental health days also increase slightly with diabetes class, but the difference is smaller. Overall, worse physical and mental health is more common as diabetes severity increases.

- Q4: Income & Education vs Diabetes Rates

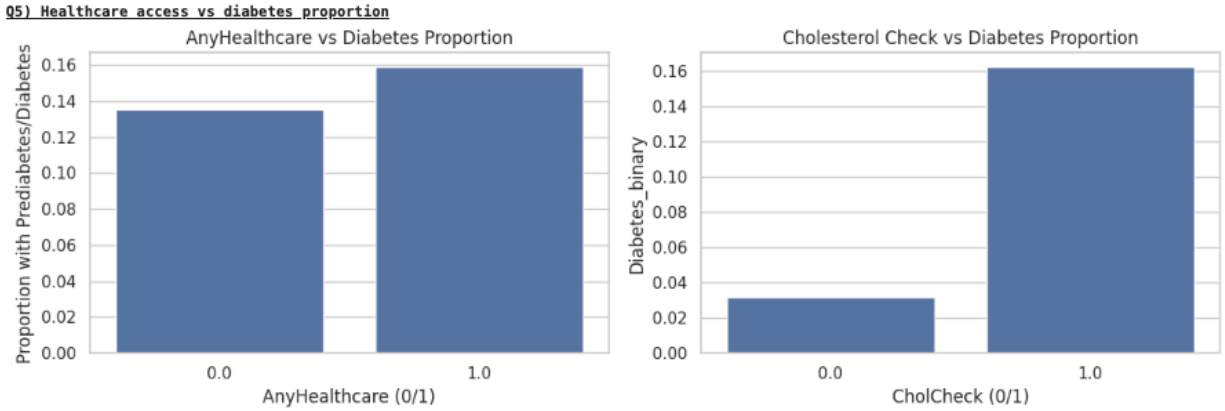
  Does someone's income or education level affect their chances of having diabetes?



From the graph, both income and education appear to be linked to diabetes risk. People with lower income levels have the highest proportion of diabetes or prediabetes, and the rate steadily goes down as income increases. The same pattern shows up with education: individuals with less education have higher diabetes rates, while those with more education show lower rates. Overall, the trend suggests that higher income and higher education are associated with a lower chance of having diabetes.

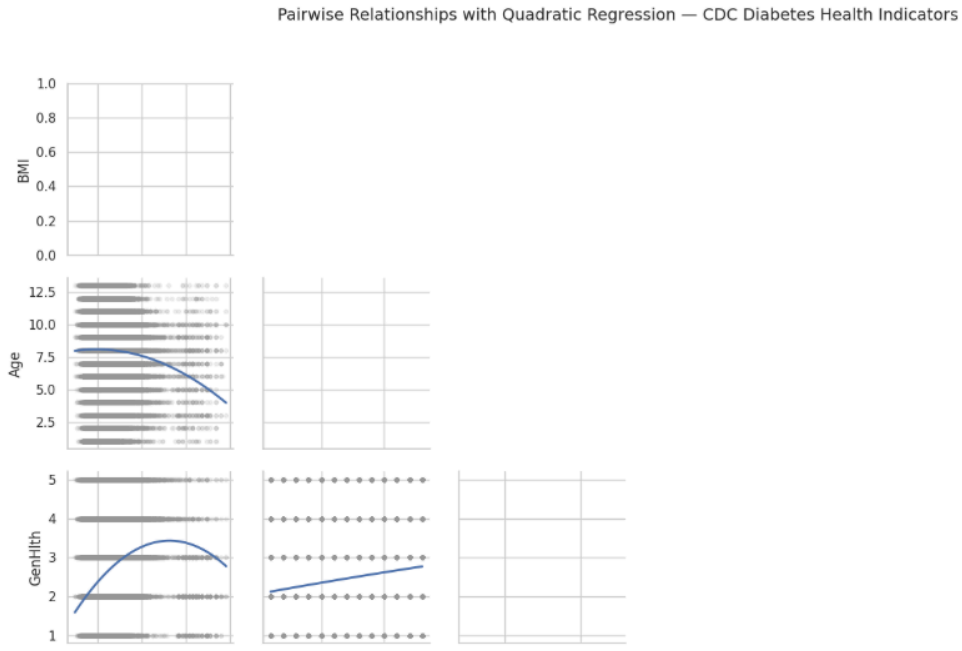
- Q5: Healthcare Access vs Diabetes

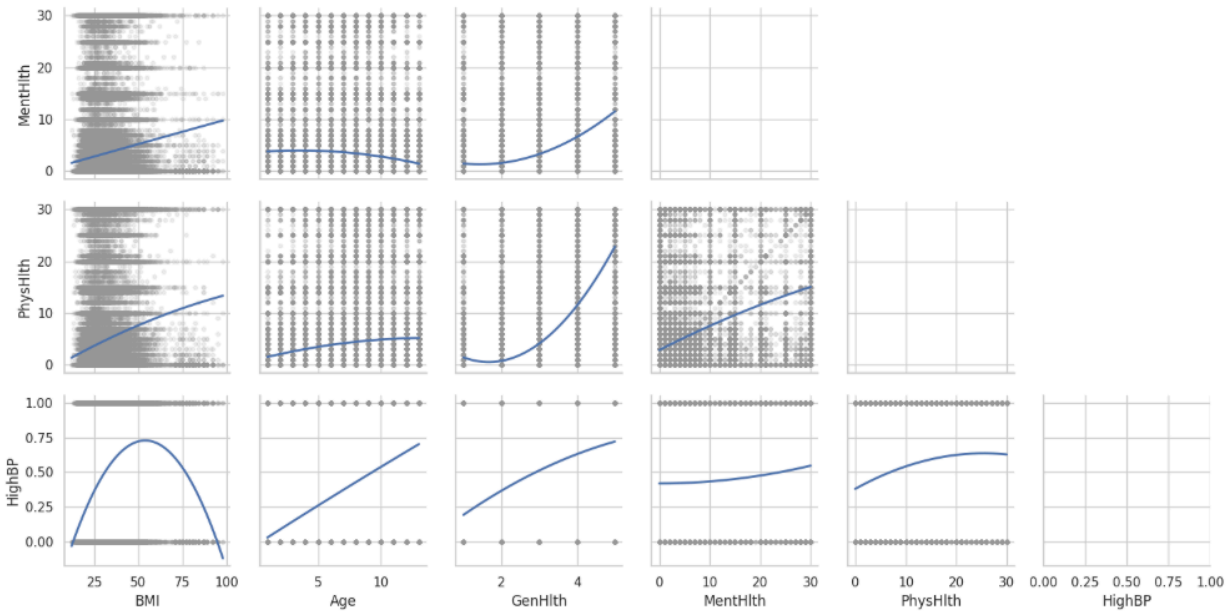
Are people who have healthcare coverage and regular checkups less likely to have diabetes than those who don't?



The graph shows that people with healthcare coverage or who get cholesterol checks actually have higher diabetes rates. This doesn't mean healthcare causes diabetes—it's probably because people who already have health issues (like diabetes or prediabetes) are more likely to seek medical care, have insurance, or get regular checkups. While people without coverage might just not be getting diagnosed.

- Pairwise Polynomial Regression Plot





- Why polynomial was used:

We used polynomial regression because many of these relationships aren't straight lines. Things like BMI, age, or general health don't always increase or decrease at a constant rate. Sometimes they curve up, level off, or even bend in the opposite direction. A polynomial curve can capture those shapes better than a straight line, so it gives a more realistic view of how the health variables relate to each other.

- Relationships observed:
  - BMI and mental/physical health days show an upward curve, meaning higher BMI is linked with more days of poor health.
  - Age vs BMI bends downward a bit, which suggests that BMI may peak and then drop slightly with older age.
  - General health vs BMI or age also forms curves rather than straight lines--worse general health is often linked to higher BMI and maybe mid-range age.
  - High blood pressure (HighBP) tends to rise as age or general health worsens.

- Interactive Plotly Visualization

  - What it shows

The visualization shows how the proportion of people with prediabetes or diabetes changes depending on how they rate their general health. Each bar represents a different health level (like excellent, good, fair, poor), and the age slider lets you see how this pattern shifts across

different age groups. It basically shows that worse self-reported health is linked with higher diabetes rates, and you can explore how that looks for each age band.

The interactive animation is generated automatically when the notebook is run. After Cell 7 executes, an HTML file named sugarstats_interactive.html will be created. The plot can be explored directly inside Google Colab or downloaded for offline viewing using:

```
from google.colab import files
files.download("sugarstats_interactive.html")
```

Machine Learning Pipeline

Machine learning models can identify risk patterns earlier and more efficiently than manual screening (Mujumdar & Vaidehi, 2019).The models used for this project were Logistic Regression, Random Forest, and XGBoost. All three models were used to classify whether a person had no risk of diabetes, prediabetes, or diabetes based on a list of features that the model was trained on .

The dataset was split with 80% being used for training and 20% of the dataset used for testing. By having 20% of the data set aside for testing, we were able to have a solid method to make sure our model was not overfitting or underfitting.

In our machine learning pipeline, we used a 5-fold stratified cross-validation with three metrics. The metrics we used were accuracy, macro F1, and Macro ROC-AUC. After running the machine learning pipeline, we used the model with the best macro F1 score to train fully, test, and run feature importance.
Logistic regression had the highest macro F1 score out of the 3 models, so we used it and were able to get a test accuracy of 0.690.

Results

After looking at the confusion matrix, one can see that the model performs well with class 0 but poorly on classes 1 and 2. In class 0, the model had more than 30,000 true positives, while class 1 had around 100 true positives and class 2 had around 4700 true positives, with most being classified as class 0. This tells us that the model has a large bias for class 0, resulting in poor recall for classes 1 and 2. Key findings from this model were that test accuracy was 0.690 for logistic regression, but this was probably due to the strong performance on class 0. The Macro F1 score of 0.438 tells us that the model struggles with minority classes and that the severe class imbalance affects classes 1 and 2. The features that had the most importance were general health status, age, BMI, and high BP. The features that hadthe most importance

are well-established clinical risk factors for diabetes, which increases confidence in the model. Poor performance on classes 1 and 2 indicated that it was difficult to differentiate between prediabetes and diabetes from healthy individuals using current features. Additional features like glucose and family history could improve prediction (Mujumdar & Vaidehi, 2019). Health-related behaviors had a low impact on whether an individual had diabetes or not, indicating they may not be measured precisely or contribute less to short-term classification.
Limitations of this dataset are that there is severe class imbalance, with class 1 having around 900 samples, making it heavily underrepresented. This led to poor recall and almost no ability to detect prediabetes properly. Another possible limitation is that some of the features are broad, such as general health, which could lead to less precision in the model overall. Another potential limitation is that there could be more clinical features that could be added, such as glucose levels or family history.

Conclusion

The logistic regression model provides a highly interpretable framework for understanding which health factors most strongly influence diabetes risk. While accuracy is relatively high, performance is heavily skewed toward the majority class, resulting in weak detection of early-stage diabetes. General health, age, BMI, blood pressure, and cholesterol emerged as the most influential predictors. The main limitations stem from dataset imbalance, self-reported features, and the linear nature of the model. Future work should incorporate resampling techniques, additional clinical variables, more flexible models, and optimized decision thresholds to achieve more balanced and clinically useful predictions.
There are no special instructions for running the notebook, and running the cells as normal would lead to good results.

References

Eseadi, C., Amedu, A. N., Ilechukwu, L. C., Ngwu, M. O., & Ossai, O. V. (2023). Accessibility and utilization of healthcare services among diabetic patients: Is diabetes a poor man's ailment?. *World Journal of Diabetes*, *14*(10), 1493.

Hill-Briggs, F., Adler, N. E., Berkowitz, S. A., Chin, M. H., Gary-Webb, T. L., Navas-Acien, A., ... & Haire-Joshu, D. (2020). Social determinants of health and diabetes: a scientific review. *Diabetes care*, *44*(1), 258.

Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Procedia Computer Science*, *165*, 292-299.

UCI Machine Learning Repository. (2015). *CDC Diabetes Health Indicators Dataset*. https://doi.org/10.24432/C53919