Project Goals

In this project, we aimed to answer two main questions:

1. Can the data collected here accurately predict whether an individual has diabetes?
2. What risk factors are most impactful/predictive of an individual having diabetes?

To address these questions, we developed a predictive model to identify diabetes cases and analyzed the impact of various health indicators.

Data Overview

The dataset contained over 250,000 records, each representing a patient who completed a health survey. There were 22 columns, encompassing different health indicators related to diabetes. Many of these were binary (yes/no) variables, such as whether an individual has high blood pressure. Given the categorical nature of the data, we utilized correlation matrices to understand relationships between variables.

When we first started looking at the data, we wanted to see how each variable compared to the average rate of diabetes. Bar charts, like the ones used in our analysis (Figure 1), showed that individuals with high blood pressure had a significantly higher rate of diabetes (around 24%) compared to those without high blood pressure (about 6%). Similarly, we observed trends for other factors such as high cholesterol.

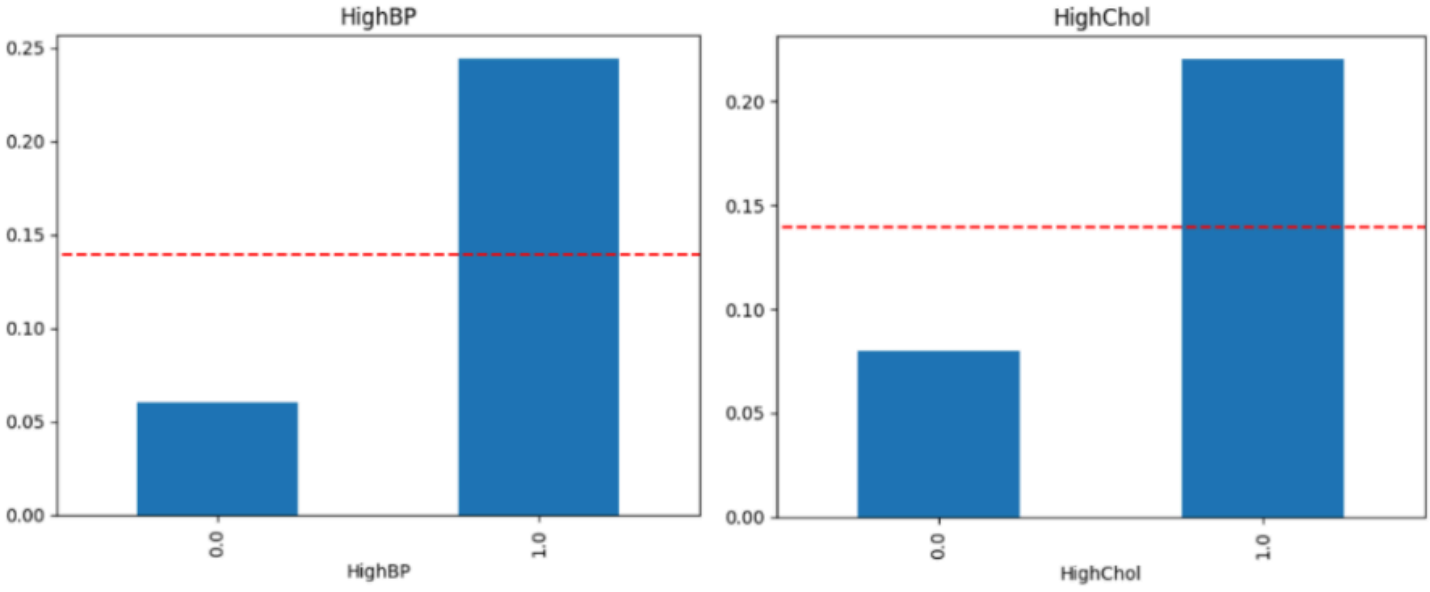

Figure 1

Understanding Diabetes and Key Risk Factors

Diabetes is a chronic disease where the body either does not produce insulin or cannot use it effectively, leading to high blood sugar levels. Our analysis confirmed that high

cholesterol and high blood pressure were among the strongest predictors of diabetes. Additionally, we noticed an interesting inverse relationship between heavy alcohol consumption and diabetes likelihood, suggesting that individuals diagnosed with diabetes may be less likely to consume high amounts of alcohol.

The Model

After identifying key risk factors, we incorporated them into our predictive model, which produced the following correlation matrix (Figure 2). We found that about 25% of the predictions were false positives, meaning our model could predict diabetes fairly well but still had room for improvement. To refine our predictions, we analyzed the weight assigned to each variable and adjusted our approach accordingly (Figure 3).

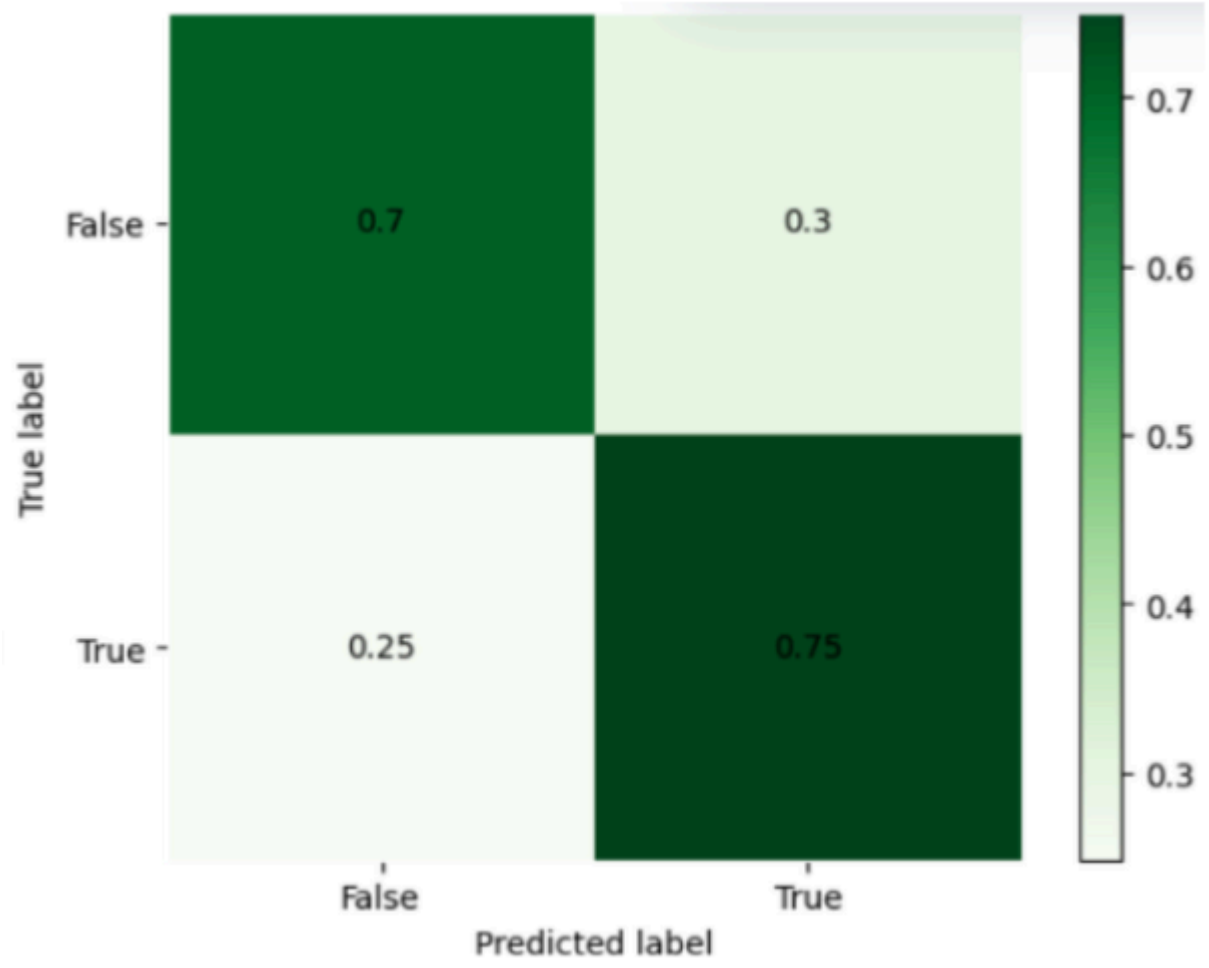

Figure 2

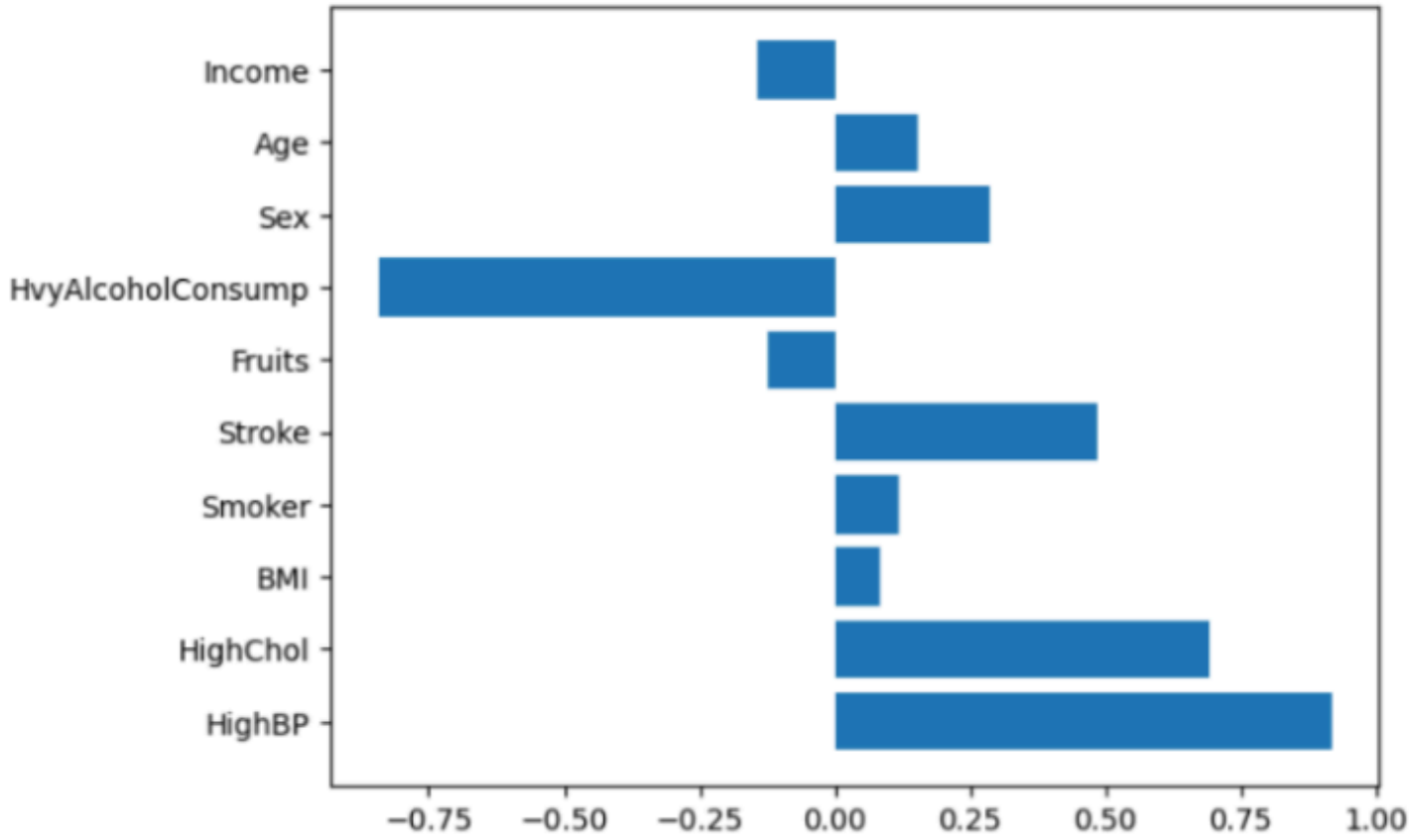

Figure 3

Improving the Model

To enhance model accuracy, we introduced additional variables:

1. High Readings - A combined indicator of high blood pressure and high cholesterol.
2. Healthcare Access - A new column that merged healthcare availability and whether individuals skipped doctor visits.
3. Good Health Indicators - A combination of physical activity levels, stroke history, and heart disease presence.

Despite these improvements, our model's accuracy only increased by 2%, bringing the total accuracy to 74% (Figure 4). To understand misclassifications, we plotted a class separation histogram (Figure 5), highlighting areas where our model struggled the most—particularly within the 0.4-0.6 probability range.

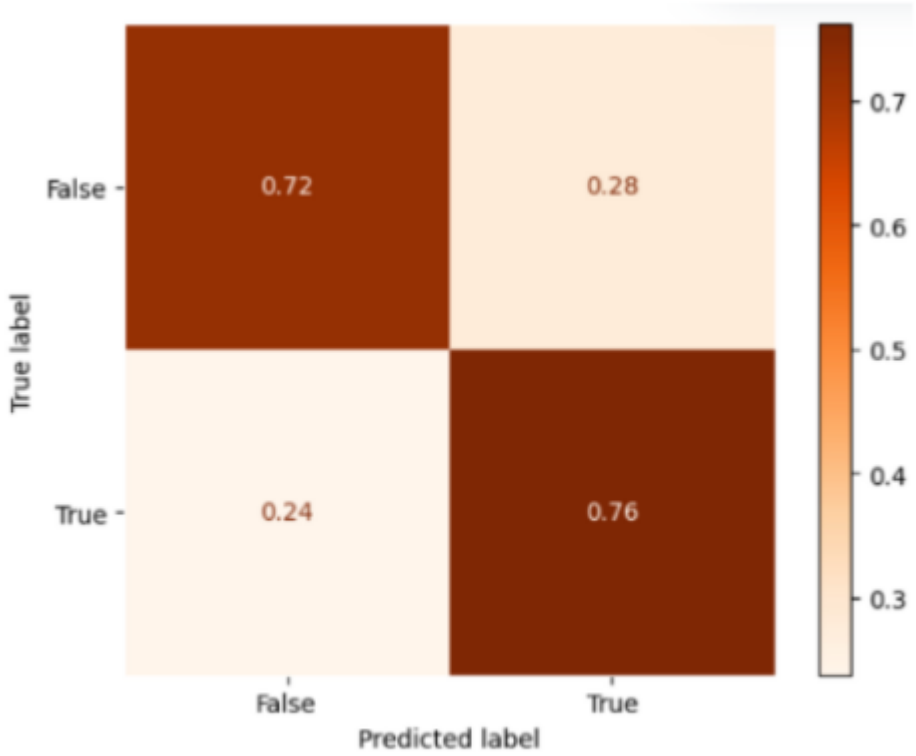

Figure 4

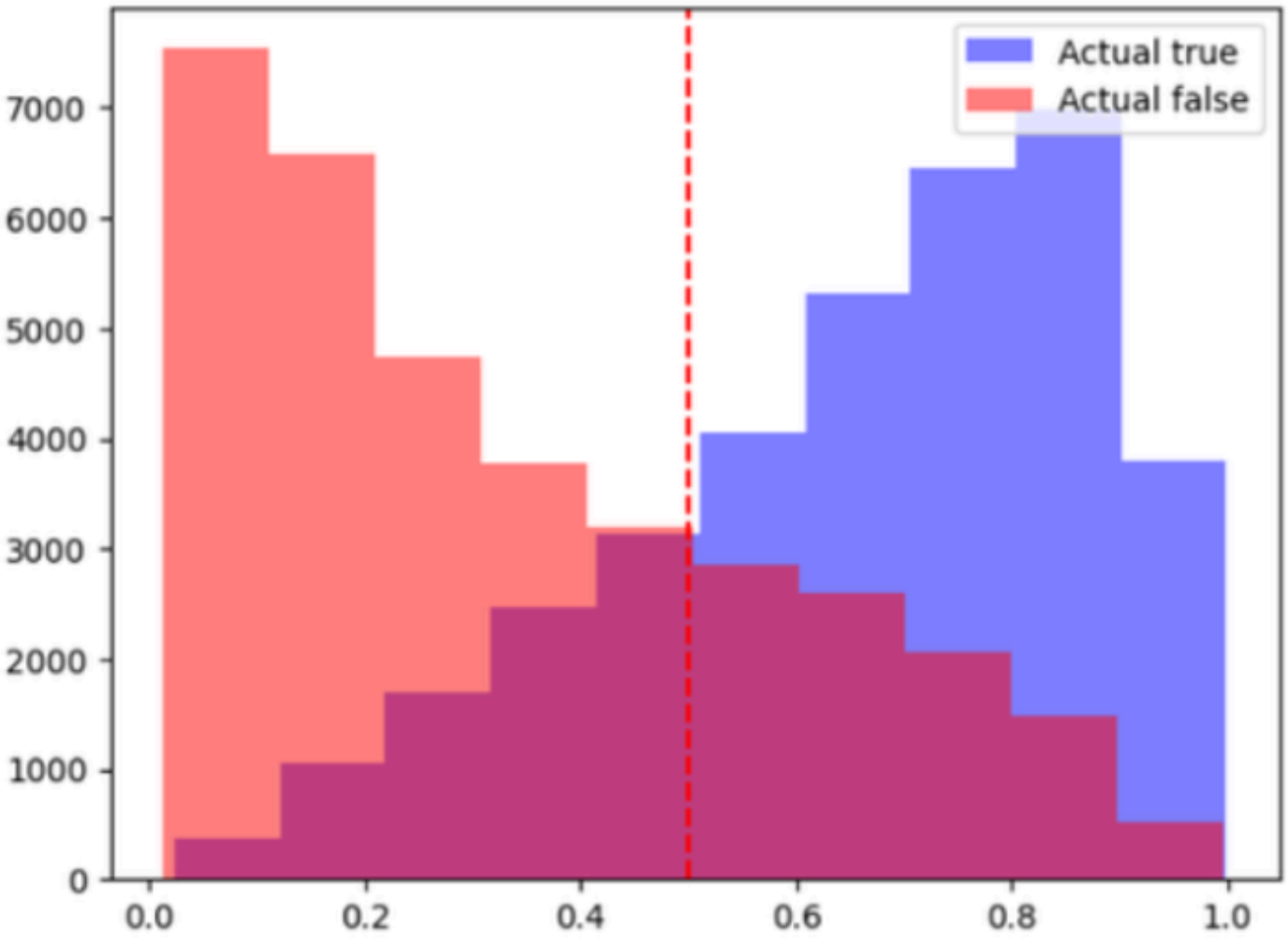

Figure 5

Other Considerations

We explored potential differences in diabetes risk factors across gender, age groups, and income levels, but found no statistically significant disparities. However, we acknowledged that some key dietary and medication-related factors were missing from the dataset, limiting prediction accuracy.

Wrapping Up

Reflecting on our initial questions:

1. Can the data provide accurate predictions? - Yes, with an accuracy of up to 74%.
2. What are the most impactful risk factors? - High blood pressure and high cholesterol were the strongest predictors, while general health was a significant factor in individuals without diabetes.

While 74% accuracy is a solid starting point, incorporating additional data, such as dietary habits and glucose measurements, could further improve the model's effectiveness. Despite these limitations, our analysis demonstrates the power of data-driven healthcare insights and the potential of machine learning in medical research.