Predicting Diabetes: An Exploration

Fantastic Four (Data Analysts Edition):

Meihan Chen, Parnika Dandepally, Catherine Maclean, Likitha Temmanaboyina

1. Define the Problem

Our question is: What features are most important for determining whether an individual has diabetes or not? Answering this question solves the problem of identifying the key characteristics of diabetes and, by doing this, helps with prevention for those who do not yet have diabetes but could get it in the future by detecting early signs, efficient diagnosis for those that have it, and overall healthcare efficiency because diagnosis is now efficient, cutting unnecessary tests and their costs.

The dataset we are using contains 21 different feature variables. These variables include diabetes status, high BP status, high cholesterol status, whether the individual had a cholesterol check in the past 5 years, BMI, whether the individual smoked 100+ cigarettes in their life, if the individual has had a stroke, heart disease/attack status, physical activity, fruit consumption, vegetable consumption, heavy alcohol consumption, the status of health care coverage, if the individual needed to see a doctor but they couldn't afford it in the past 12 months, general health status, mental health status, physical health status, walking difficulty status, sex, age, education, and income. Most of the features are binary or categorical, while the remainder are numerical variables.

The constraints of our project are the limited number of variables we have access to, as we do not know if unaccounted-for outside factors may have an influence. In terms of sample size, it is of a large scope. This particular sample contains 253,680 individual survey responses. We hope to be able to apply our conclusions about diabetes based on this sample to the general population.

The sample was taken by the CDC from people of their own free will, so there are no ethical concerns to be taken up with the sampling method. The data and the conclusions we drew from it will only be used to serve as education for the public, not as something to act against them.

2. Data Collection

The data was collected by the CDC via a telephone survey, which collects data on health-related conditions, behaviors, and preventative services. The original data set was taken from the American population, and responses from 441,455 American individuals were recorded. It was then cleaned and turned into diabetes.csv, which is a clean dataset of 253,680 survey responses to the CDC's BRFSS2015. So, there are 253,680 samples and 21 feature variables, most of which are binary classification variables (i.e., yes or no.) These are the individuals our data set focuses on. The data is of good quality, as it was collected by the CDC, which is a government agency, and we believe it is reliable because of this.

However, it is important to note that this is a self-reported survey, and people may lie about how often they consume alcohol, cigarettes, fruits, vegetables, etc., which could lead to a failure in the ability to correctly identify the characteristics of those with diabetes. The class imbalance in this sample can create a misleading conclusion about the population, but we worked to counteract this by using models with equal sample sizes from each category.

3. Data Preparation

There are no missing values. Using the custom function detect_outliers, we found 9847 BMI outliers, 36208 MentHlth outliers, and 40949 PhysHlth outliers. We decided to keep these because, despite being outliers, they still represent individuals whose responses were recorded and should be considered. Additionally, most of the outliers were likely from the diabetes or pre-diabetes group, so by taking out these outliers, the already small diabetes and pre-diabetes groups could become even smaller. An example of this is BMI, which had many outliers of greater values. The diabetes and pre-diabetes groups both had higher mean BMIs than the no-diabetes group. Additionally, BMI and diabetes also had a

positive correlation to the heatmap we created. There were no inconsistencies or errors in the data that needed to be corrected. We checked if there were missing values and then if there were duplicated rows. There were 23899 duplicate rows, but we did not drop these because each row represents a unique individual. No more data preparation was needed. The data's original format was already ideal for processing and visualization. We did not need to transform or scale any of the variables.

4. Data Exploration

Looking at the distribution, the "No Diabetes" group made up 84.2% of this dataset, "Diabetes" made up 13.9%, and "Pre-diabetes" made up 1.8%. This class imbalance was important to take into consideration when building the models.

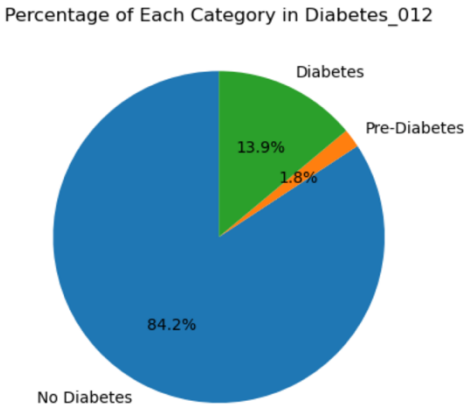

Image 1: The distribution of the data.

During the data exploration for our project proposal, we learned that average BMI, high BP levels, and high cholesterol levels seem to increase from nondiabetic to diabetic people and that people also seem to be more likely to get diabetes as they turn older. We discovered that our findings are supported by data previously published by the CDC, which states that being overweight/obese, being 45 or older, and being physically active less than 3 times a week are risk factors for type 2 diabetes (Diabetes Risk Factors). These risk factors seem somewhat connected to the features we noticed in our exploration.

For example, people who are overweight are more likely to have high cholesterol, and those who are overweight are at more risk for type 2 diabetes (Diabetes Risk Factors). Additionally, we saw that diabetic and prediabetic people also seem to be less likely to see a doctor due to costs. During the data exploration for this write-up, we investigated all the variables to compare them.

On average, the No Diabetes group answered “no” to high BP, high cholesterol, smoking, stroke, and heart disease or attack. Most of them answered “yes” to regular physical activity, regular fruit consumption, regular vegetable consumption, and if they had healthcare. Their income and education are higher than those of the Pre-Diabetes and Diabetes group. They experienced fewer bad mental health days and bad physical health days than those with pre-diabetes and diabetes. They have less difficulty walking. Their BMIs are also lower, on average. Looking at the mean responses of each group, the No Diabetes group appears to score better in almost every health and socioeconomic category. The exception to this is that the No Diabetes group does not have their cholesterol checked on average quite as often as the Pre-Diabetes and Diabetes group.

Diabetes_012	HighBP	HighChol	CholCheck	BMI	Smoker	Stroke	HeartDiseaseorAttack	PhysActivity	Fruits	Veggies	...	AnyHealthcare	NoDocbcCost	GenHlth	MentHlth	PhysHlth	DiffWalk	Sex	Age	Education	Income
0.0	0.371132	0.379171	0.957104	27.742521	0.429680	0.031628	0.071833	0.779077	0.643023	0.821439	...	0.949739	0.079610	2.372391	2.944404	3.582416	0.132282	0.433985	7.786559	5.106629	6.208663
1.0	0.629022	0.620816	0.986612	30.724466	0.492766	0.057223	0.143382	0.678471	0.602246	0.768948	...	0.945152	0.129346	2.975599	4.529907	6.348305	0.277478	0.437702	9.083351	4.784496	5.351112
2.0	0.752674	0.670118	0.993182	31.944011	0.518220	0.092457	0.222882	0.630538	0.585441	0.756408	...	0.959769	0.105868	3.290981	4.461806	7.954479	0.371216	0.479121	9.379053	4.745516	5.210094

Mean values of the data set.

The average person in this dataset does not have diabetes, does not have high BP, does not have high cholesterol, checks their cholesterol regularly, has a high BMI, has not smoked 100+ cigarettes in their life, has not had a stroke, has not had heart disease or attack, exercises regularly, consumes fruit, has healthcare, does not avoid the doctor because of the cost, has good to very good general health, has good mental and physical health most of the time, does not have difficulty walking, has 1-3 years of college, and makes between $35,000 and $75,000 annually. All of these answers are affected by the class imbalance in this dataset, wherein “no diabetes” is the majority. Some outliers can affect the data, as mentioned previously in data preparation.

After analyzing our dataset, we found some correlations among the variables. GenHlth and PhysHlth are correlated, with a score of 0.45 out of 1.00. GenHlth and DiffWalk are correlated, with a score of 0.42 out of 1.00. DiffWalk and PhysHlth are correlated, with a score of 0.41 out of 1.00. Because difficulty walking is correlated with poor physical health, poor physical health may be correlated with poor general health. Other correlations include Age and HighBP, which scored 0.35 out of 1.00, GenHlth and HighBP, which scored 0.30 out of 1.00, and HighCol and HighBP, which scored 0.30 out of 1.00. All other correlations scored less than 0.30 but can be seen on the visualization.

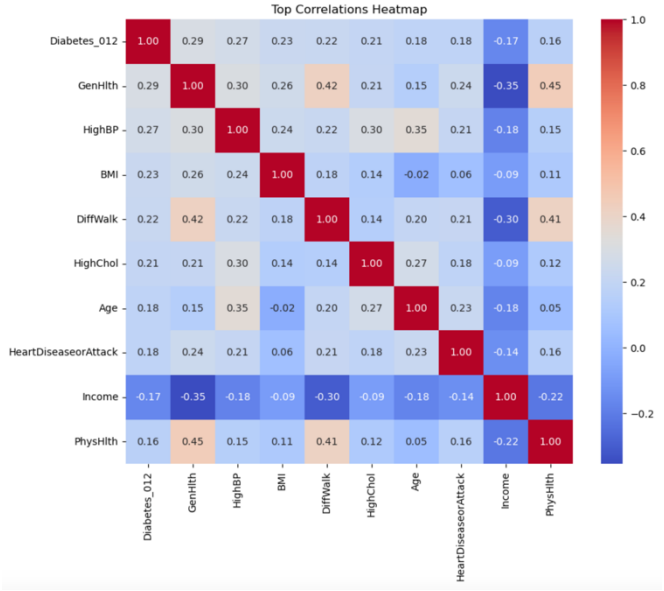

Image 2: The top ten correlations in the data

5 & 6. Model Building & Model Evaluation

Model 1:

We asked a question where we needed to look at what the most important features of those diagnosed with diabetes are. So, we are essentially trying to distinguish the most predictive variables

based on the data available. This is where a classification model is good because we are trying to take characteristics and group a patient into the diabetic or non-diabetic category.

As we noticed earlier in the data exploration stage, there is an imbalance between the amount of data points between classes. The diabetes group had 35,346 data points, and the non-diabetes group had 213,703 data points. So, for the classification model, we have decided to under-sample the non-diabetic group. Since there are fewer data points on those with diabetes, we randomly sampled 35,346 (which is the number of data points that the diabetes class has) from the no diabetes class and made a new dataset. Our new dataset for the model is 50% diabetes and 50% no diabetes. We left out the pre-diabetes group because the sample size was so small, and it can be difficult to predict a pre-diagnosis compared to a full diagnosis (diabetes).

To evaluate the model, we randomly sampled 80% of the data points from the new dataset and used it for training. The rest of the data points were used in the test dataset. To figure out what depth the tree should be to yield the highest accuracy score, we ran the model many times with different max depths. From there, we used a function to determine the depth that yielded the highest accuracy score for the test dataset and used it to make the tree figure diagram. One thing to note is that the tree diagram and the scores change every time it is run based on how the data points are split into the train or test dataset. When we ran the model, the depth that was used was 8. The model indicates that High BP is the most informative feature of the features used in the model to predict whether a person is diabetic or not. The accuracy score for the model is 0.7493987834205686, or approximately 74.9%. From the score, we can say that the model does a decent job of predicting between people with diabetes and no diabetes. If the model was used on new data, then there could be potential problems. If this new data contains people who would be diagnosed as pre-diabetic in real-time, then the model would not be very useful in this case as it is not trained to recognize pre-diabetes.

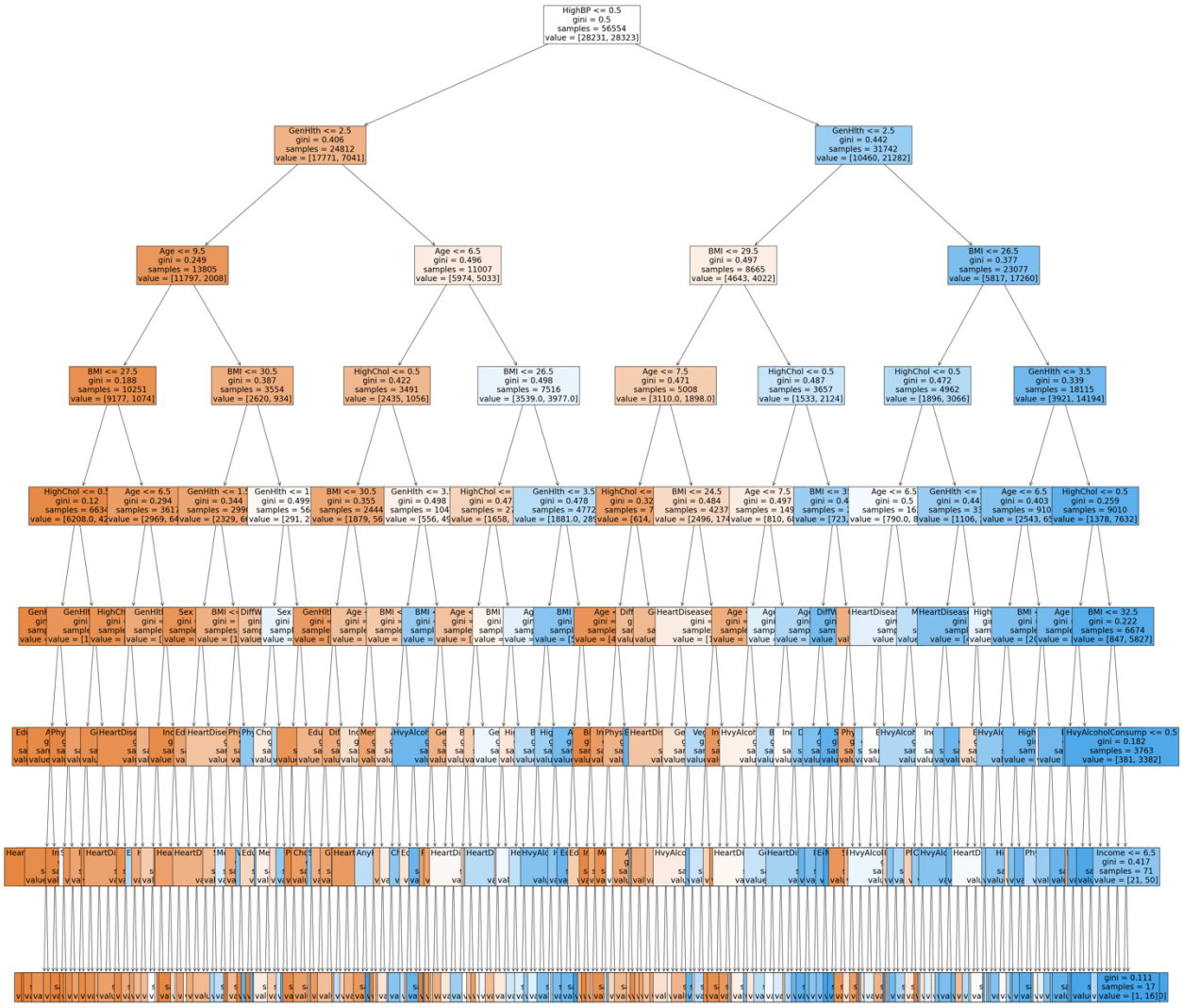

Image 3: Model 1 Tree diagram

Model 2:

We applied the Random Forest classification algorithm, a flexible algorithm capable of handling both binary and continuous variables. This model is particularly well-suited for datasets with mixed predictor types, including features such as blood pressure (HighBP), smoking status (Smoker), and body

mass index (BMI). Random Forest minimizes the risk of overfitting by leveraging an ensemble of decision trees and handles outliers effectively. Furthermore, it offers the added benefit of feature importance scoring, providing insights into which predictors are most influential in determining outcomes, such as diabetes prevalence. The ease of implementation and minimal preprocessing requirements make Random Forest an efficient and effective choice for this classification problem.

The dataset initially exhibited class imbalance, with significantly fewer cases of diabetes compared to non-diabetes cases. To address this imbalance, the Synthetic Minority Oversampling Technique (SMOTE) was applied. SMOTE generates synthetic samples for the minority class, ensuring a balanced class distribution for the training process. It prevents the model from being biased towards the majority class and improves its performance on the minority class. After applying SMOTE, the dataset was split into training and testing sets using an 80-20 split, ensuring that the training data maintains a balanced class distribution. This ensures that the model is trained on sufficient data while retaining a separate set for unbiased evaluation.

The Random Forest classifier was initialized with 100 estimators and a fixed random state for reproducibility. The training dataset was used to fit the model, enabling it to learn patterns and relationships between predictors and the target variable.

	precision	recall	f1-score	support
No Diabetes	0.89	0.96	0.92	42721
Diabetes	0.95	0.88	0.91	42761
accuracy			0.92	85482
macro avg	0.92	0.92	0.92	85482
weighted avg	0.92	0.92	0.92	85482

Image 4: Classification Report for Model 2

The trained Random Forest classifier was evaluated on the test set using a classification report, which provides detailed performance metrics, including precision, recall, f1-score, and support for each class. The evaluation results are as follows:

- Precision:
    - 89% of the data points predicated as “No Diabetes” in the test set are correct.
    - 95% of the data points predicted as “Diabetes” in the test set are correct.
- Recall
    - The model successfully identifies 96% of actual "No Diabetes" cases
    - The model correctly identifies 88% of actual “Diabetes” cases.
- Accuracy
    - The model achieves an overall accuracy of 92% on the test set, meaning 92% of all predictions are correct.

The classification report highlights that the model performs slightly better for the "No Diabetes" class than for the "Diabetes" class. Despite this slight disparity, the overall high accuracy, precision, recall, and F1-scores indicate that the Random Forest classifier is effective at predicting both classes in this balanced dataset. The evaluation metrics, summarized in the classification report, confirm the model’s reliability for this binary classification task, making it suitable for applications requiring accurate identification of diabetes cases while keeping false positives and false negatives to a minimum.

To gain insights into the predictors’ relative contributions, feature importance scores were extracted from the Random Forest model. A bar plot was created to visualize the importance of each feature. Predictors like HighBP, general health (GenHlth), cholesterol levels (HighChol), and BMI emerged as the most significant contributors to the classification task.

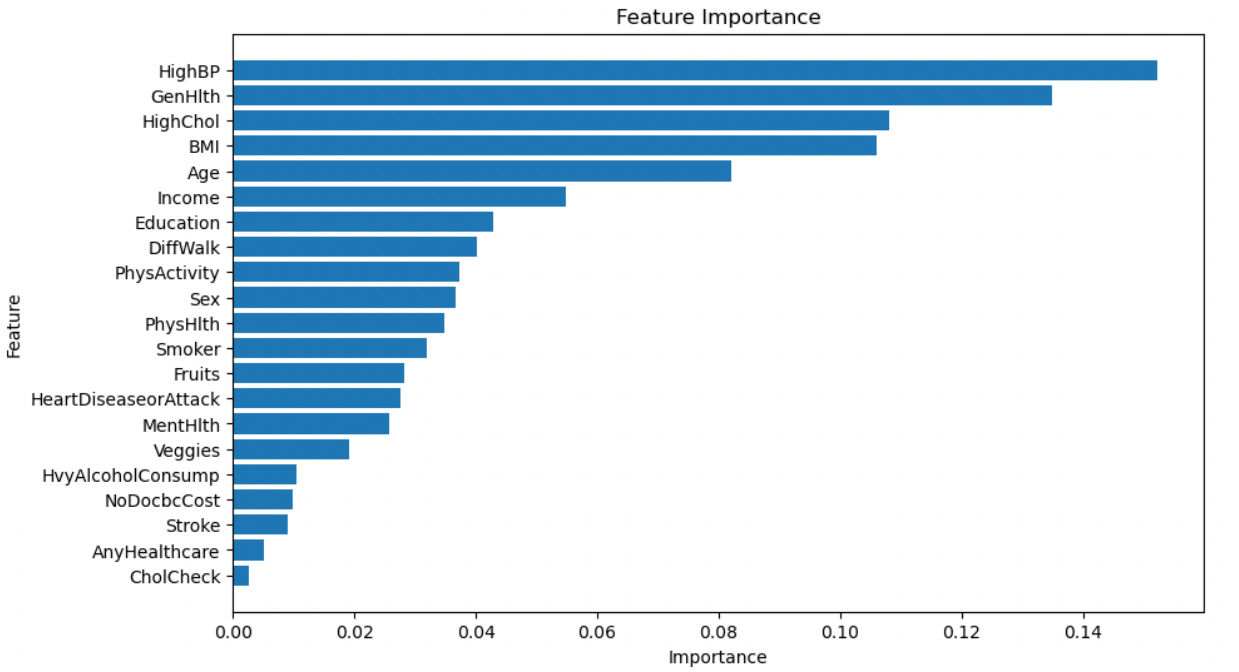

Image 5: Feature importance is ranked by Model 2's results.

7. Model Deployment

Our models determine the most important features for determining whether someone has diabetes or no diabetes. If medical providers know what to check for in someone they believe may have diabetes, they can make efficient diagnoses and avoid other unnecessary tests. If an individual knows that they are not considered healthy in the most important features for determining diabetes, they can know that they should ask to be tested for it when they do to their medical provider and avoid paying for unnecessary testing of other illnesses.

Both the decision tree model and random forest model can be integrated into medical settings and used to predict whether a patient has diabetes or not. When a patient goes to their doctor, after their current medical information is collected by the nurse who checks their BMI, BP, etc., the nurse can use a computer to run the models. The models will be used to predict whether or not the patient has diabetes

based on their information and will notify the medical provider of the results. Of course, a doctor should make the final diagnosis, but the results from the decision tree and random forest can help with diagnosis efficiency.

This data is only taken from a small portion of the American population. Over time, more data should be collected for the model. It is important to have current and not outdated data, as well as large, diverse sample sizes to include a variety of demographics, since diabetes/no diabetes features may be different in different people.

There is always a chance that the models could be incorrect in predicting whether a patient has diabetes or no diabetes. Although the data was collected by the government and should be fairly accurate because of this, responses were given by individuals via the phone with no one to verify the information they shared. The data the models use is also generalized in the form of yes or no questions (e.g., “Is your blood pressure high?” “Is your cholesterol high?”) and does not take into consideration that these feature variables are different numerical values. Having just a yes or a no instead of numerical data for BP, cholesterol, how many days an individual has exercised, etc., may affect the models’ ability to predict diabetes. Some could lie about their responses or make errors in the information they report. Due to this, the data has the potential to be inaccurate and creates the challenge of trying to accurately predict whether a patient has diabetes despite potentially inaccurate data and the potential risk of these predictions being inaccurate because of the inaccurate data.

The models are also incredibly binary at times, with categorical variables being measured as 0s and 1s rather than fully considering their meanings (for example, HighChol, with no high cholesterol individuals being represented as 0s and high cholesterol individuals being represented as 1s. The model takes these into account to do mathematical calculations, despite the high cholesterol number being a cutoff number, and so slightly skews our results. Rather than this, it could be better to have a range scale

based on cholesterol ranges, similar to the Income variable. This issue is similarly seen in other variables such as Stroke, PhysActivity, etc.

Additionally, the models do not take into account most pre-existing medical conditions, which could affect whether an individual has diabetes or not. The model checks for past heart attack/stroke data and whether the individual is a smoker but does not check for other conditions such as cancer, forms of arthritis, etc. This could lead to the model misdiagnosing someone as diabetic when their symptoms are caused by a pre-existing condition or misdiagnosing someone as non-diabetic when some of their symptoms are being overshadowed by other factors. The model is also limited by its number of features, meaning that when a patient tells their doctor about a factor that the model does not take into account, the model is unable to incorporate the new factor.

However, though the models could create inaccurate predictions, this does not matter much in the end since a doctor will have to diagnose the patient themself. All the models will do is predict whether a patient is likely to have diabetes/no diabetes and will not diagnose the patient on their own. If the models predict that someone has diabetes, their results can make a doctor who was previously not thinking about testing for diabetes test for it. The doctor will consider the models' results, and they, as the health care provider, will make the final diagnosis. Based on whether the models accurately predicted whether a patient had diabetes or not, they will have separate success rates that can be measured as they continue to be used. This success rate can be used to measure and evaluate its performance once it is used in medical settings.

8. Summary

In short, we investigated the question: What features are most important for determining whether an individual has diabetes or not? Our decision tree established HighBP as the most important feature,

then split into GenHlth, which then split into BMI and Age. Depending on the previous variable, it then split into BMI, HighChol, Age, or GenHlth. Our random forest model put HighBP as the most important feature—this same feature was ranked as most important by the first model, which should be noted—then GenHlth, HighChol, then very closely after, BMI, then Age. It seems that these are the most important variables for determining whether an individual has diabetes or not, and our models utilize these variables, along with the remaining variables, to predict whether an individual has diabetes or not. Our models can be implemented in medical settings and, with the supervision of human medical professionals, assist with diagnosis efficiency.

Works Cited

"Diabetes Risk Factors." *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, www.cdc.gov/diabetes/risk-factors/index.html. Accessed 7 Dec. 2024.