Identifying Key Risk Factors for Diabetes and Developing Prevention Strategies

David Kim, Jiyoon Moon, Rebekah Northrup, Sahithi Rampally

Team Data-betics

DATA 110-001

Prof. Harlin Lee

December 7, 2024

Data Science Lifecycle

Define the Problem:

Diabetes is a growing health crisis worldwide strongly influenced by lifestyle and demographic factors. This project aims to identify the key risk factors contributing to diabetes to inform preventive strategies that can help individuals mitigate their risk. We focused on the primary risk factors associated with diabetes and how this information can be leveraged to develop effective prevention strategies for the broader American population.

Data Collection:

The dataset comes from the Behavioral Risk Factor Surveillance System (BRFSS), a health-related survey conducted by the CDC. We used a 2015 dataset available on Kaggle. The sample number is 253,680 survey responses and there are 22 features including demographics, diabetes status, BMI, physical activity, general health, and mental health. The dataset is based on self-reported survey data, leading to potential inaccuracies and biases. It excludes populations without telephone access, underrepresenting low-income and rural households. Certain features, like mental and physical health, may be subjective due to individual and cultural variations.

Data Preparation:

To handle missing values, we identified them using isna() function. We found that there were no missing values, so we continued our data preparation. We then made a new table without "pre-diabetics" to specifically address diabetes. The column "Diabetes_bin" was created to convert diabetes from 2 to 1. Finally, the order of the columns was shifted so that the "Diabetes_bin" column was the first in order to make a correlation heatmap.

Data Exploration:

Prior to cleaning or altering the structure of the dataset, we used the shape attribute and info() function in order to consider the data from a high-level view. This helped us to verify the number of samples and features we're working with, as well as the variable type and non-null counts. To get a better understanding of our dataset after data cleaning, we used the describe() function to view key statistics like mean, standard deviation, and range for each feature.

After cleaning the dataset by excluding pre-diabetic entries, we visualized the relationships between variables using a correlation matrix and heatmap (sns.heatmap) to identify the strongest associations with the target variable, Diabetes_bin. Below is our heatmap visualization (Figure 1) for reference:

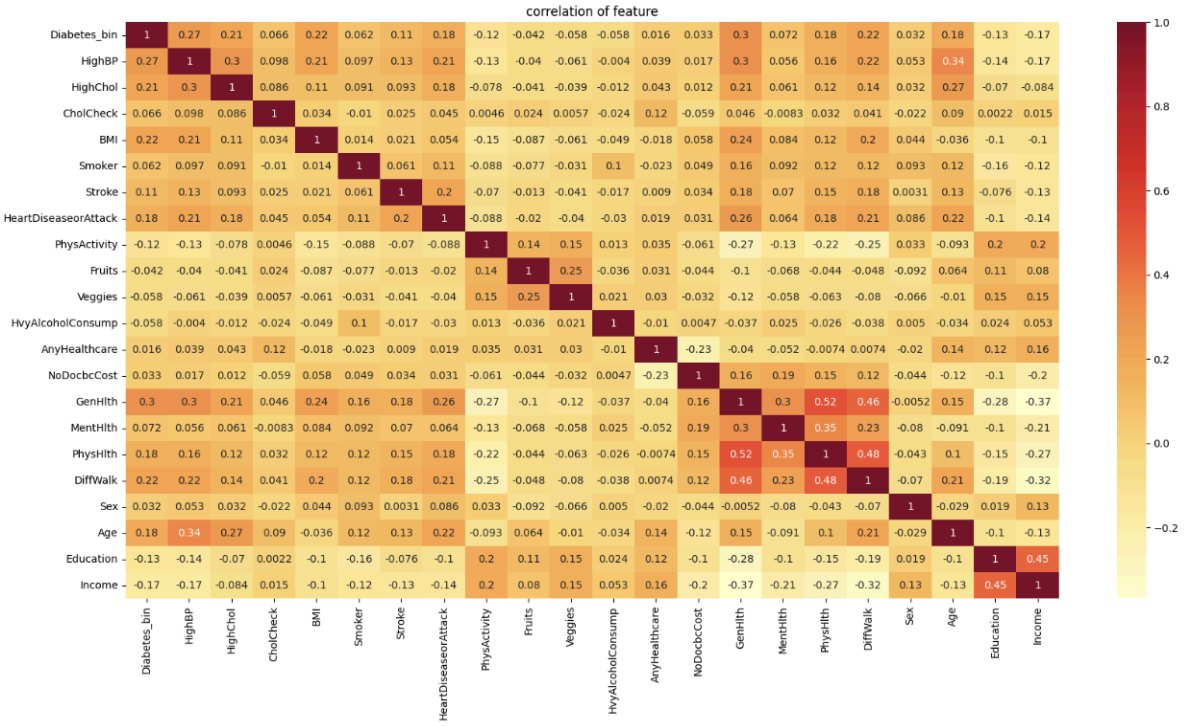

Figure 1: Correlation of Features

The top correlated variables with diabetes were:

- GenHlth (General Health)
- HighBP (High Blood Pressure)
- DiffWalk (Difficulty Walking)
- BMI (Body Mass Index)
- HighChol (High Cholesterol)

We found that these variables had the highest correlation coefficients with Diabetes_bin, with GenHlth having the strongest relationship. Other notable factors included Age, HeartDiseaseorAttack, and PhysHlth (Physical Health).

This analysis provided insights into which features are most significant for predicting diabetes, guiding the next steps in feature selection and modeling. The correlation analysis also highlighted the inverse relationships, such as lower physical activity levels and income being associated with higher diabetes prevalence.

Model Building:

Once we found the features with the highest correlation coefficients, we utilized the top 11 most correlated ones (half of the features) to build our model. As in the conventional modeling procedure, we dropped our target variable column (Diabetes_bin) from the dataset since we're looking to predict this feature. Below is a preview of the dataset after selecting the top 11 variables (Figure 2).

	GenHlth	HighBP	DiffWalk	BMI	HighChol	Age	HeartDiseaseorAttack	PhysHlth	Income	Education	PhysActivity
0	5.0	1.0	1.0	40.0	1.0	9.0	0.0	15.0	3.0	4.0	0.0
1	3.0	0.0	0.0	25.0	0.0	7.0	0.0	0.0	1.0	6.0	1.0
2	5.0	1.0	1.0	28.0	1.0	9.0	0.0	30.0	8.0	4.0	0.0
3	2.0	1.0	0.0	27.0	0.0	11.0	0.0	0.0	6.0	3.0	1.0
4	2.0	1.0	0.0	24.0	1.0	11.0	0.0	0.0	4.0	5.0	1.0
...	...	...	...	...	...	...	...	...	...	...	...
253675	3.0	1.0	0.0	45.0	1.0	5.0	0.0	5.0	7.0	6.0	0.0
253676	4.0	1.0	1.0	18.0	1.0	11.0	0.0	0.0	4.0	2.0	0.0
253677	1.0	0.0	0.0	28.0	0.0	2.0	0.0	0.0	2.0	5.0	1.0
253678	3.0	1.0	0.0	23.0	0.0	7.0	0.0	0.0	1.0	5.0	0.0
253679	2.0	1.0	0.0	25.0	1.0	9.0	1.0	0.0	2.0	6.0	1.0

249049 rows × 11 columns

Figure 2: Dataset with top 11 variables prepared for modeling.

For this project, we chose a Decision Tree Classifier due to its simplicity, interpretability, and ability to handle both numerical and categorical data effectively. Decision trees are especially useful in identifying the most informative variables for classifying samples, which aligns with our goal of understanding the key factors contributing to diabetes.

For simplicity and interpretability, we chose a maximum depth of 3. Below is the resulting visualization of the trained decision tree (Figure 3), which highlights the hierarchical structure and the significant features contributing to predictions.

```
Score on train: 0.8587525534659379
Score on test: 0.855370407548685
```

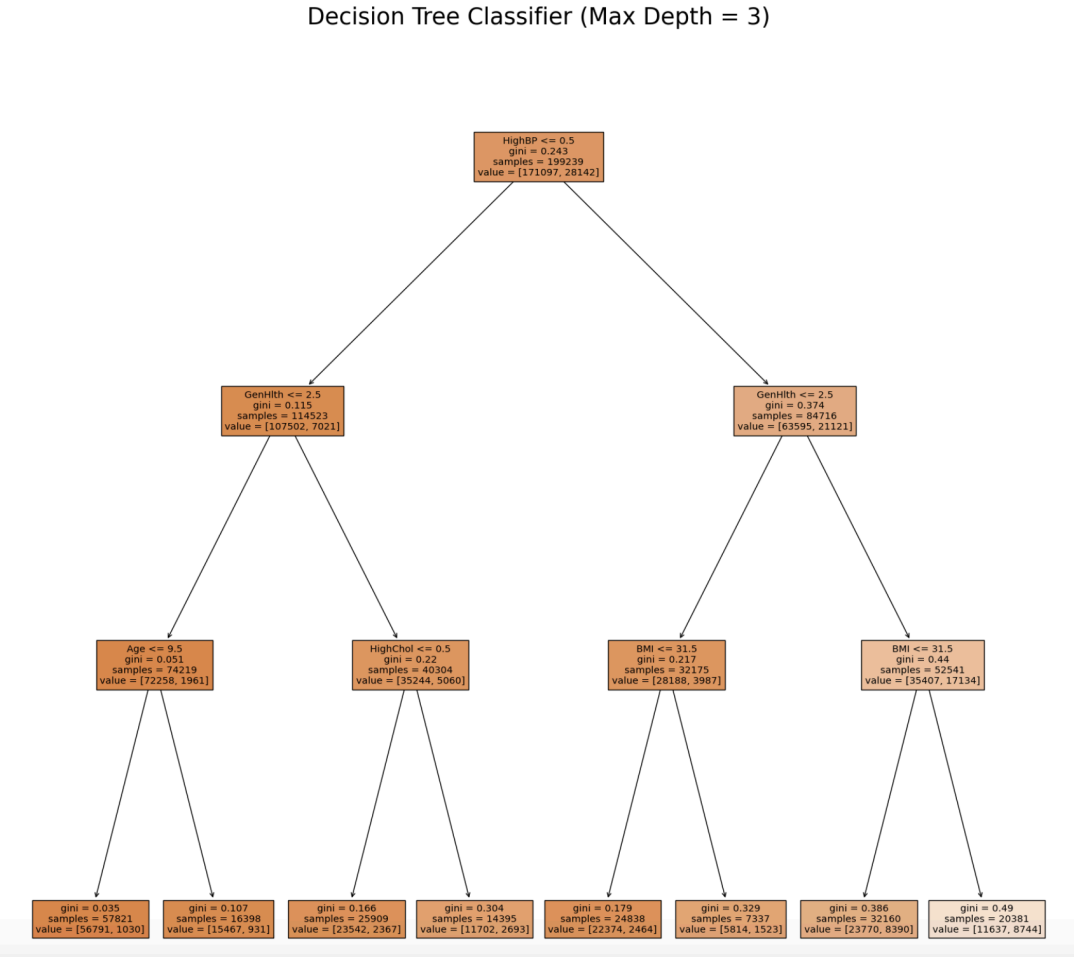

Figure 3: Decision Tree Classifier visualization with max depth = 3.

We tested how prediction scores changed with varying tree depths. As shown in Figure 4, the model's performance plateaued at a tree depth of 5 or 6, confirming diminishing returns for deeper trees.

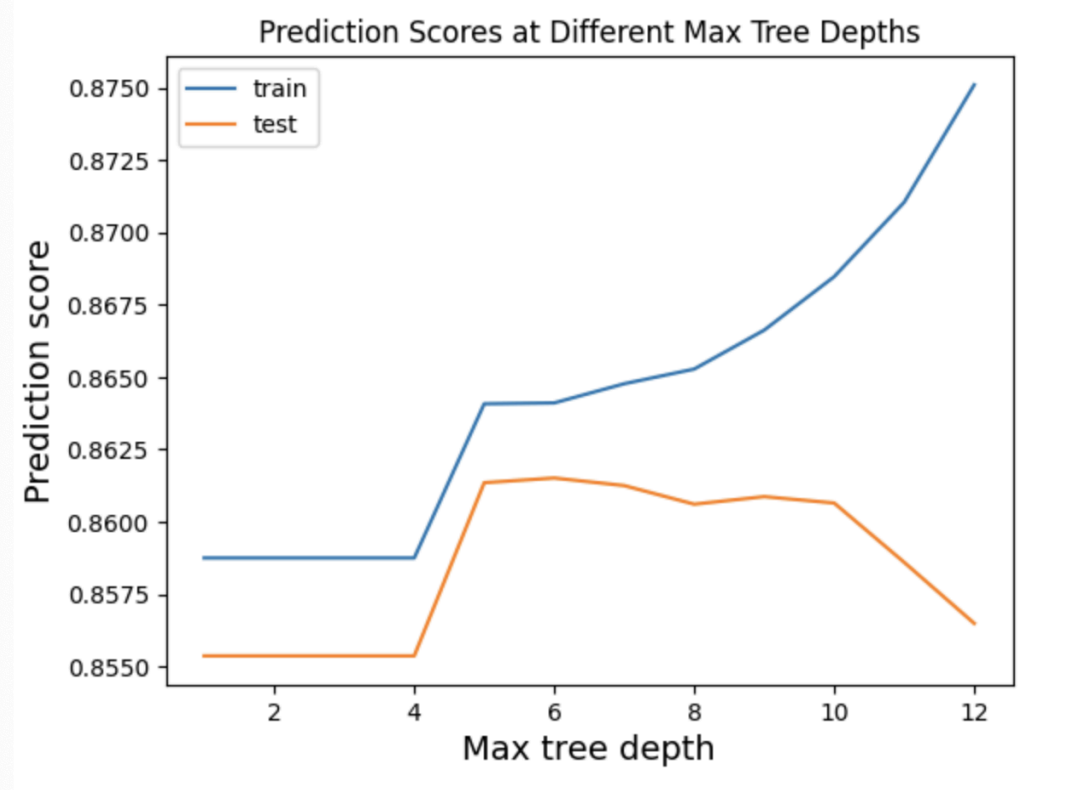

Figure 4: Prediction scores at different max tree depths.

We narrowed our focus to the top 11 variables that demonstrated the highest correlation with Diabetes_bin, including:

- GenHlth (General Health): Highest correlation with diabetes prevalence.
- HighBP (High Blood Pressure): Frequently identified as the root feature in our decision tree models.
- DiffWalk (Difficulty Walking): Significantly correlated with diabetes, suggesting mobility challenges.
- BMI (Body Mass Index): A key metric of physical health.
- HighChol (High Cholesterol): Known risk factor for diabetes.

Other variables included Age, HeartDiseaseorAttack, PhysHlth, Income, Education, and PhysActivity.

To split the dataset, we used 80% of the data for training and 20% for testing. A decision tree with a maximum depth of 3 was chosen to balance visibility, interpretability, and performance.

Although Figure 4 shows that a tree depth of 5 offers better accuracy, the difference was very minimal and may be considered insignificant. We concluded a depth of 3 would be sufficient for us to derive insights without overfitting.

We also explored the effect of randomly selecting subsets of features and identified HighBP as the most frequently occurring root node across multiple iterations, emphasizing its significance in diabetes classification.

Model Evaluation:

The question we were seeking to investigate was: "Can key lifestyle and demographic factors be used to effectively classify individuals as diabetic or non-diabetic?" When answering this question for each of the features, the decision tree model achieved:

- Training accuracy: 85.88%
- Testing accuracy: 85.54%

At first glance, this indicates a well-performing model with minimal overfitting. However, a deeper analysis reveals the potential for overfitting due to the inherent class imbalance within the dataset. ~85% of the patients in the dataset do not have diabetes. This imbalance means that a model could achieve high accuracy simply by predicting all individuals as non-diabetic without actually learning meaningful patterns. For example, if the model predicts all instances as non-diabetic, it would still achieve 85% accuracy, which is highly misleading.

When testing prediction scores across varying tree depths, the accuracy evened out at depths of 5 or 6, suggesting that deeper trees may not provide significant accuracy increases.

Additionally, we conducted multiple runs with random subsets of features to assess the validity of the model. Below is a visualization of the decision tree trained with random subsets of features (Figure 5).

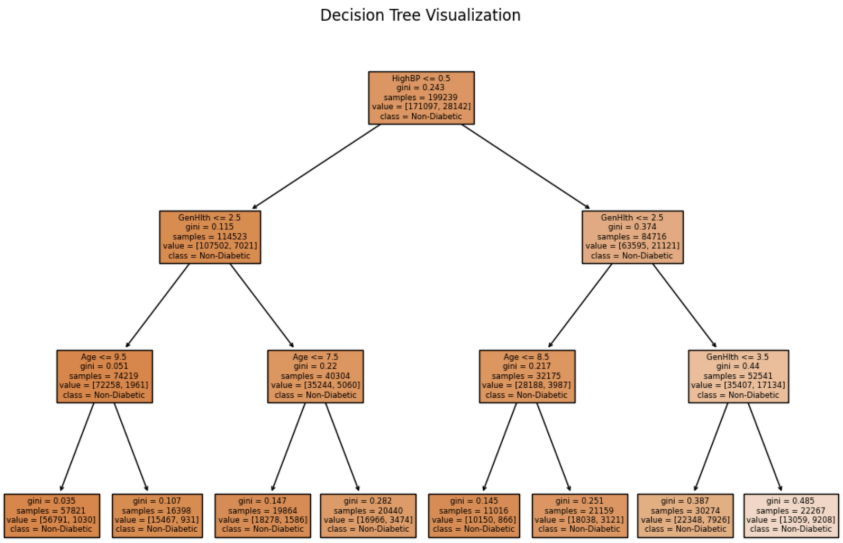

Figure 5: Decision Tree visualization with randomly selected features.

HighBP was consistently the root feature when using all 11 selected variables, reinforcing its position as one of the highest predictors. Other influential variables included GenHlth, DiffWalk, and HighChol, however, we inferred that those variables were the root feature in the cases when HighBP was not a part of the randomly selected features. These outcomes are presented in Figure 6.

	Num_Features	Root_Feature	Frequency
0	11	HighBP	30
1	8	HighBP	20
2	5	GenHlth	12
3	5	HighBP	10
4	8	GenHlth	7
5	5	DiffWalk	4
6	5	HighChol	3
7	8	DiffWalk	3
8	5	BMI	1

Figure 6: Frequency of root feature across random subsets of features.

This evaluation confirms the strong predictive power of HighBP, followed by other influential variables like GenHlth, DiffWalk, and HighChol.

Model Deployment:

The decision tree model could be deployed in healthcare settings to:

- Screen individuals for diabetes risk: Based on their responses to simple lifestyle and health-related questions related to high blood pressure (HighBP), general health metrics (GenHlth), difficulty walking (DiffWalk), and other highly predictive factors.
- Guide targeted interventions: Prioritize high-risk individuals (specifically those who face challenges associated with the predictive features we outlined) for counseling or preventive measures.

Deployment Challenges

- Data Accuracy: The reliance on self-reported data introduces biases and inaccuracies that could impact model reliability.
- Generalizability & Selection Bias: The model is trained on a specific dataset and may require re-training to adapt to other populations. The data was collected with selection bias, which stemmed from the exclusion of specific populations in the collection process itself. As a telephone survey, participants are only made up of people with access to phones, causing certain groups of people (ie. lower-income households, rural populations, etc.) to be left out or underrepresented. This may make the dataset a less accurate representation of the broader population as a whole.
- Ethical Considerations: Misclassification could lead to unnecessary stress or missed interventions, requiring careful communication of predictions to users.
- Class imbalance: Despite efforts to balance the dataset during model training, the original dataset's class imbalance could still influence predictions. For instance, a model trained on imbalanced data may struggle to generalize to new, more balanced populations. Future work could explore advanced techniques like SMOTE (Synthetic Minority Oversampling Technique) to address class imbalance more comprehensively.

By addressing these challenges, this model can serve as a crucial tool for early diabetes detection and prevention strategies.