Detecting Diabetes using Machine Learning

Stephen Butters

1 Introduction

1.1 Problem Statement

The problem addressed in this project is the prediction of diabetes using health-related data. Specifically, the goal is to develop a model that can accurately predict whether an individual is diabetic or not based on factors such as age, body mass index (BMI), physical activity, and other health indicators. Given the widespread prevalence of diabetes and its significant impact on individuals' health, early detection and prediction are crucial for managing and preventing this chronic disease.

1.2 Motivation

Diabetes is a major global health issue, with millions of people living with it, often without knowing they have it. According to the National Institute of Diabetes and Kidney Diseases, "Diabetes raises the risk for damage to the eyes, kidneys, nerves, and heart. Diabetes is also linked to some types of cancer. Taking steps to prevent or manage diabetes may lower your risk of developing diabetes health problems." (NIDDK 1). This means early detection can help prevent or manage the disease more effectively. The challenge, however, lies in the complex interplay of multiple factors that contribute to diabetes. Health indicators like BMI, physical activity, blood pressure, and cholesterol levels are all interconnected, making it difficult to create a model that can accurately predict the risk of diabetes. This project aims to address this challenge by using advanced machine learning techniques to identify the most important factors and make reliable predictions.

1.3 Approach

To address this problem, I used a Gradient Boosting model, a powerful machine learning algorithm known for its high accuracy in classification tasks. Gradient Boosting works by building multiple decision trees sequentially, where each tree corrects the errors of the previous one, leading to improved predictions over time. This approach is well-suited for the task, as it can handle the complexity of the relationships between health factors and diabetes risk. By training the model on a dataset of health indicators, I was able to predict the likelihood of an individual being diabetic and analyze the feature importance to understand which factors have the greatest influence on diabetes prediction.

2 Data

2.1 Introduction of the Data

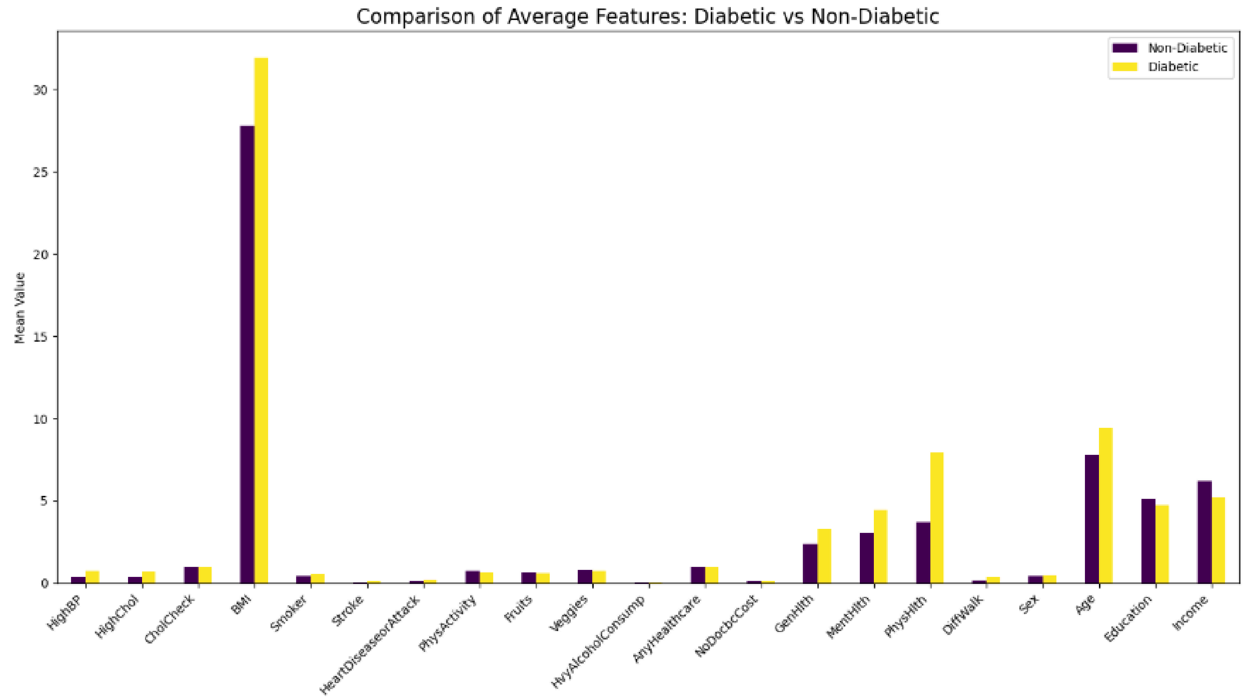
This dataset contains health and demographic information collected from 70,000 individuals as part of the 2015 Behavioral Risk Factor Surveillance System (BRFSS) survey. It includes 21 various health indicators such as blood pressure, cholesterol levels, BMI, physical activity, smoking habits, and more, alongside whether or not the individual has diabetes. The goal of this dataset is to predict whether an individual has diabetes based on these health and lifestyle features.

The target variable is "Diabetes_binary", where 1 indicates the person has diabetes and 0 indicates the person does not have diabetes. The other columns represent different factors that might influence diabetes risk, including health conditions (e.g., HighBP, HighChol), lifestyle factors (e.g., PhysActivity, Smoker), demographic factors (e.g., Age, Income), and access to healthcare (e.g., AnyHealthcare, NoDocbcCost).

2.2 Using Gradient Boosting to Help Predict Diabetes

Gradient Boosting is a machine learning technique that combines multiple decision trees to improve prediction accuracy. It works by training each tree to correct the mistakes made by previous trees, making it very powerful for predictive tasks like this one. In this case, Gradient Boosting can be used to build a model that predicts whether someone is likely to have diabetes based on their health and lifestyle data. By training the model with features such as BMI, smoking habits, physical activity, and more, it learns to identify patterns that indicate whether a person has diabetes or not. Once the model is trained, it can predict diabetes status for new individuals. The model is then evaluated for its performance through metrics such as accuracy, precision, and recall to understand how well it can make predictions.

2.3 Basic Visual Analysis



A bar graph was used because it’s a simple way to compare the average values of different features between diabetic and non-diabetic groups. The graph shows that people in the diabetic group tend to have a higher average BMI, which is a measure of body fat, compared to those without diabetes. It also shows that diabetics are less likely to engage in physical activity, as their average physical activity is lower. More people in the diabetic group report having high blood pressure, which is shown by their higher average for that feature. This makes it easy to see how certain health and lifestyle factors differ between the two groups.

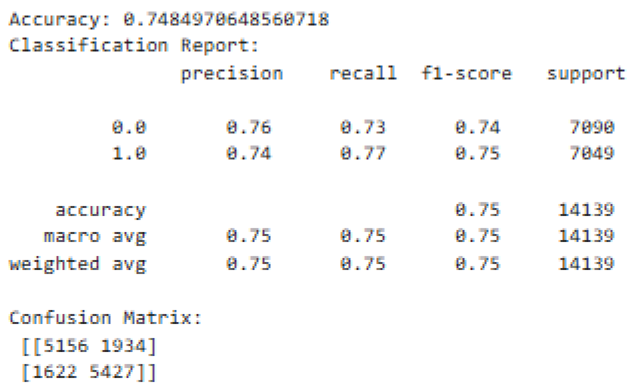
2.4 Preprocessing the Data

Preprocessing this dataset mainly involved separating the features from the target variable, splitting the data into training and testing sets, and handling the data in a way that would allow the Gradient Boosting model to make predictions. Because there were no missing values or categorical variables to deal with, the data was ready for analysis. Preprocessing helps ensure that the model has the best possible data to work with and can make accurate predictions about diabetes.

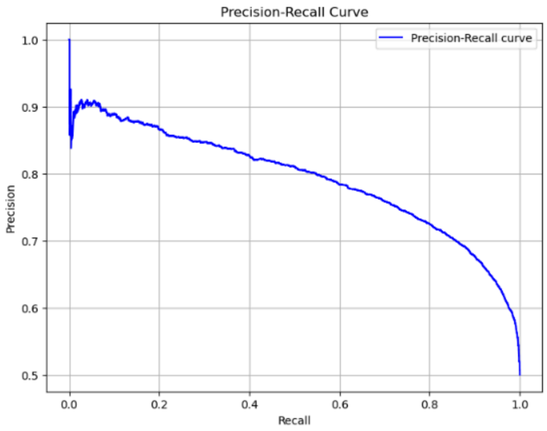
3 Methods and Results

3.1 Logistic regression

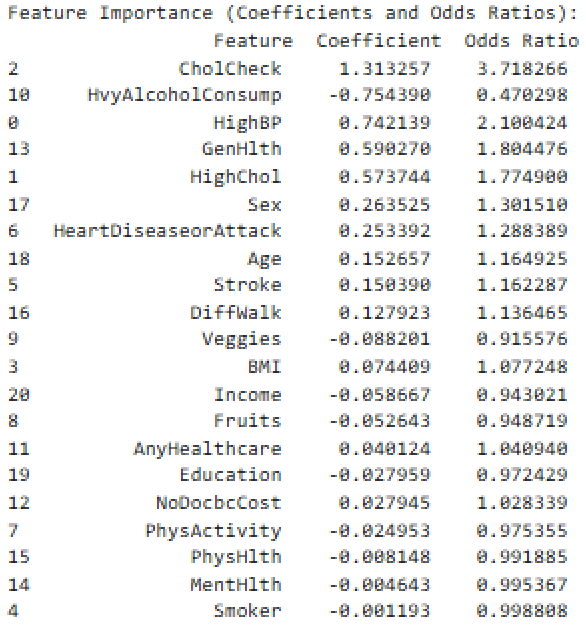
Logistic regression was chosen for this dataset because it is a simple and effective method for binary classification problems like predicting whether or not someone has diabetes. The target variable, "Diabetes_binary," is binary, making logistic regression an ideal fit since it predicts probabilities between 0 and 1 and converts them into class labels. It also provides coefficients for each feature, allowing us to interpret the relationship between factors like BMI, physical activity, and high blood pressure and the likelihood of diabetes. The model is computationally efficient, handles numerical features well, and aligns with the dataset’s assumptions, such as a linear relationship between features and the log-odds of the outcome. It serves as a baseline for comparing the performance of more complex models like Gradient Boosting. Logistic regression’s ability to highlight feature importance through coefficients and odds ratios provides valuable insights into how health and lifestyle factors contribute to diabetes risk.



```
Accuracy: 0.7484970648560718
Classification Report:
              precision    recall  f1-score   support

         0.0       0.76      0.73      0.74      7090
         1.0       0.74      0.77      0.75      7049

    accuracy                           0.75     14139
   macro avg       0.75      0.75      0.75     14139
weighted avg       0.75      0.75      0.75     14139

Confusion Matrix:
 [[5156 1934]
 [1622 5427]]
```



The logistic regression model performs fairly well, with an accuracy of about 75%, meaning it correctly predicts whether a person has diabetes or not 75% of the time. The model is slightly better at predicting diabetic individuals, with a recall of 77%, which means it correctly identifies most of the diabetics. However, it struggles a bit more with non-diabetic individuals, correctly identifying 73% of them. In terms of precision, when the model predicts someone as diabetic, it is correct 74% of the time, and when it predicts someone as non-diabetic, it is correct 76% of the time. The confusion matrix shows that there are some misclassifications, especially among non-diabetic individuals, with 1934 being incorrectly labeled as diabetic.



```
Feature Importance (Coefficients and Odds Ratios):
                 Feature  Coefficient  Odds Ratio
2              CholCheck     1.313257    3.718266
10     HvyAlcoholConsump    -0.754390    0.470298
0                 HighBP     0.742139    2.100424
13               GenHlth     0.590270    1.804476
1               HighChol     0.573744    1.774900
17                   Sex     0.263525    1.301510
6   HeartDiseaseorAttack     0.253392    1.288389
18                   Age     0.152657    1.164925
5                 Stroke     0.150390    1.162287
16              DiffWalk     0.127923    1.136465
9                Veggies    -0.088201    0.915576
3                    BMI     0.074409    1.077248
20                Income    -0.058667    0.943021
8                 Fruits    -0.052643    0.948719
11         AnyHealthcare     0.040124    1.040940
19             Education    -0.027959    0.972429
12           NoDocbcCost     0.027945    1.028339
7           PhysActivity    -0.024953    0.975355
15              PhysHlth    -0.008148    0.991885
14              MentHlth    -0.004643    0.995367
4                 Smoker    -0.001193    0.998808
```

The results of the feature importance analysis show that certain factors are strongly linked to the likelihood of having diabetes. For example, CholCheck (whether a person has had their cholesterol checked) is a very strong predictor, with an odds ratio of 3.72. This means people who have had their cholesterol checked are more than three times more likely to have diabetes compared to those who haven’t. Similarly, HighBP (high blood pressure) also increases the chances of having diabetes by more than twice, with an odds ratio of 2.10. Other factors like general health and high cholesterol also play a significant role in increasing diabetes risk, as they have odds ratios above 1.

On the other hand, some factors seem to reduce the likelihood of having diabetes. For example, HvyAlcoholConsump (heavy alcohol consumption) has an odds ratio of 0.47, which suggests that heavy drinking may lower the likelihood of having diabetes. Similarly, eating vegetables and fruits seems to slightly reduce the chances of developing diabetes, with odds ratios of 0.92 and

0.95, respectively. Physical activity also has a small negative impact on diabetes risk, though its effect is minimal.

Other factors like income, education, and access to healthcare have a smaller influence on diabetes risk, as their odds ratios are close to 1, indicating that they don't have a strong impact in this model. For example, smoking has an odds ratio of 0.99, suggesting it has almost no effect on the likelihood of developing diabetes.

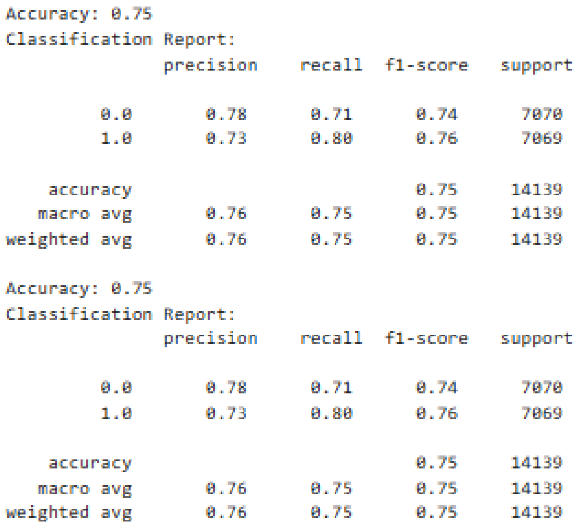
The results show that factors like high blood pressure, high cholesterol, and poor general health are strongly linked to a higher risk of diabetes. On the other hand, healthy habits like eating fruits and vegetables and staying physically active may help lower the risk.

3.2 Gradient Boosting

Gradient Boosting was chosen for this dataset because it is a powerful and widely-used machine learning technique that works well for a variety of problems, especially when the goal is to predict categories (like whether a person has diabetes or not). This method is known for its high performance in tasks where other models might not do as well, particularly when there are complex patterns or interactions between the features.

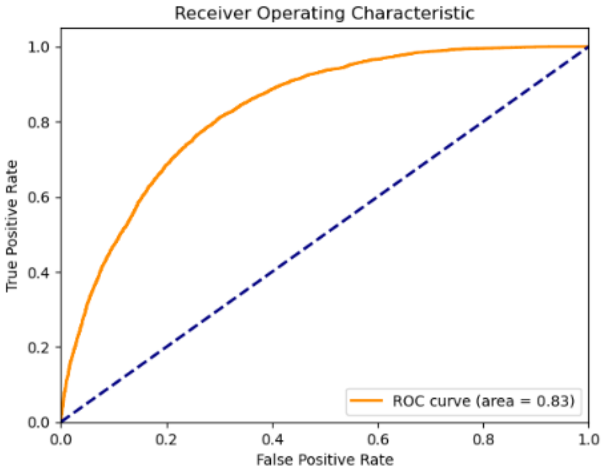
One of the main reasons Gradient Boosting is useful is that it builds multiple decision trees one after the other. Each tree tries to correct the mistakes made by the previous one, which helps the model become more accurate over time. This is called "boosting," and it allows the model to learn from its errors and improve gradually. For a dataset like this, which contains various health-related features (like BMI, physical activity, and age), Gradient Boosting can capture the complex relationships between these features and the likelihood of having diabetes.

Another advantage of Gradient Boosting is that it doesn't just focus on the overall performance of the model but also gives more weight to the harder-to-predict examples. This means that it can better handle situations where certain features or instances in the data are tricky and require more attention. Given that diabetes prediction is a complex task influenced by many factors, Gradient Boosting is a good choice because it can handle such challenges effectively.

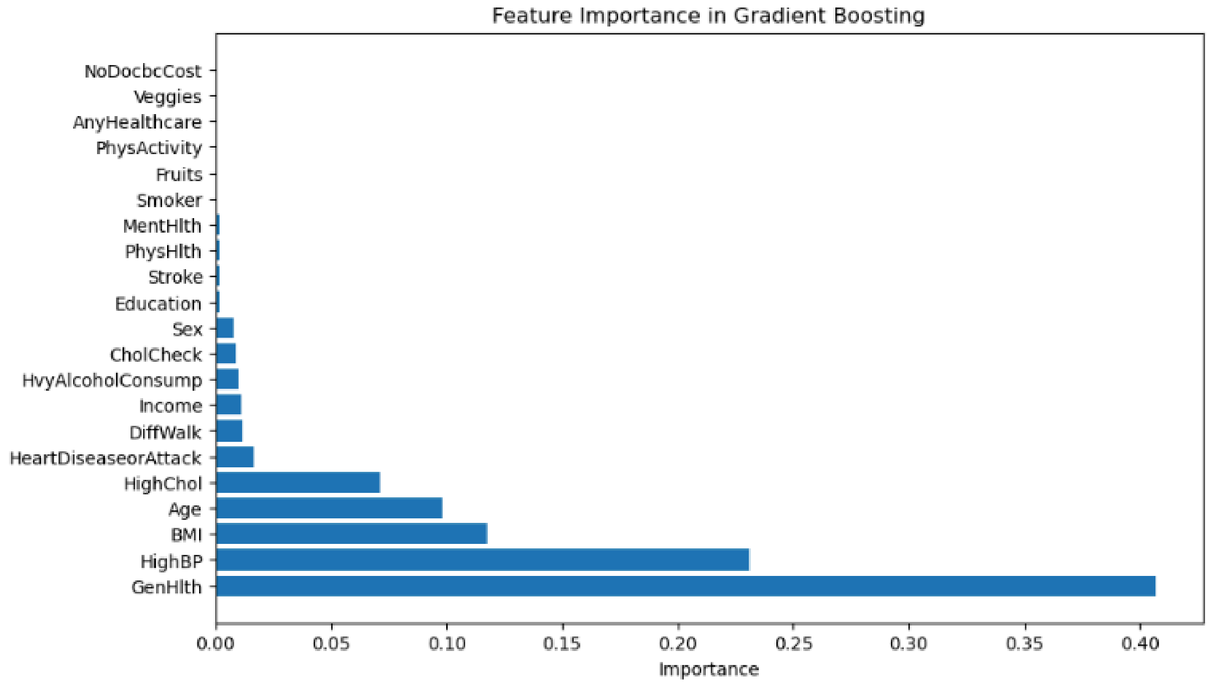


```
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

         0.0       0.78      0.71      0.74      7070
         1.0       0.73      0.80      0.76      7069

    accuracy                           0.75     14139
   macro avg       0.76      0.75      0.75     14139
weighted avg       0.76      0.75      0.75     14139
```

```
Accuracy: 0.75
Classification Report:
              precision    recall  f1-score   support

         0.0       0.78      0.71      0.74      7070
         1.0       0.73      0.80      0.76      7069

    accuracy                           0.75     14139
   macro avg       0.76      0.75      0.75     14139
weighted avg       0.76      0.75      0.75     14139
```



The model has an AUC (Area Under the Curve) of 0.83, meaning that it is quite effective at distinguishing between diabetic and non-diabetic individuals. The AUC is a metric used to measure the performance of a classification model, where values range from 0 to 1. An AUC of 0.83 indicates that the model correctly identifies the classes (diabetic or non-diabetic) 83% of the time on average. This is considered a good result, as it shows the model performs much better than random guessing, which would have an AUC of 0.5.



The feature importance bar graph from the Gradient Boosting model highlights the relative significance of different features in predicting the target variable. GenHlth (General Health) stands out with very high importance, meaning it plays a crucial role in determining whether someone is likely to have diabetes. This suggests that a person's overall perception of their health is strongly linked to the likelihood of being diabetic.

HighBP (High Blood Pressure) also holds a high importance, indicating that individuals with high blood pressure are more likely to develop diabetes. This is consistent with existing medical knowledge, as high blood pressure is a common risk factor for diabetes.

The features HighChol (High Cholesterol), Age, and BMI (Body Mass Index) have some importance, which means they still contribute to the model's predictions, but not as strongly as GenHlth or HighBP. These features are known risk factors for diabetes, so their moderate importance suggests they help, but aren't as decisive in this particular model.

Looking at this overall model, it shows that health-related factors, particularly general health and high blood pressure, are the most influential in predicting diabetes, while other factors like cholesterol, age, and BMI also contribute, but to a lesser degree.

4 Conclusion

This project successfully addressed the problem of predicting diabetes based on a range of health indicators using machine learning techniques. By implementing a Gradient Boosting model, I was able to create a classifier that accurately distinguished between diabetic and non-diabetic individuals, with an AUC of 0.83, demonstrating the model's effectiveness. The feature importance analysis revealed key factors, such as general health and high blood pressure, that play significant roles in diabetes prediction.

Throughout this project, I learned a great deal about machine learning techniques, particularly the power of Gradient Boosting for classification tasks. I also gained a deeper understanding of the health indicators that are closely related to diabetes risk, such as BMI, physical activity, and blood pressure. One of the key takeaways was how to interpret model outputs, including AUC and feature importance, which allowed me to gain valuable insights into which factors influence the prediction most.

One of the challenges I faced was dealing with the complexity of the dataset. There are many features that influence diabetes, and it took some time to figure out which ones were the most important. However, by using techniques like feature importance analysis, I was able to better understand which factors mattered most. Another challenge was making sure the model was trained correctly and evaluated accurately. I overcame this by testing the model with different performance metrics, like accuracy and AUC, to make sure it was working as expected.

This experience not only enhanced my technical skills but also deepened my understanding of how machine learning can be applied to real-world health problems, highlighting its potential for making significant contributions to public health and disease prevention.

5 References

Anshul. “Gradient Boosting Algorithm: A Complete Guide for Beginners.” *Analytics Vidhya*, 26 Nov. 2024, www.analyticsvidhya.com/blog/2021/09/gradient-boosting-algorithm-a-complete-guide-for-beginners/.

“What Is Diabetes? - Niddk.” *National Institute of Diabetes and Digestive and Kidney Diseases*, U.S. Department of Health and Human Services, www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes. Accessed 6 Dec. 2024.

“Provide examples on how to tune hyperparameters for gradient boosting” prompt. *ChatGPT*, GPT-4, OpenAI, 8 Dec. 2024, chat.openai.com/chat

6 Acknowledgement

“Gradient Boosting Algorithm: A Complete Guide for Beginners.” by Anshul was used to help create and evaluate my gradient boosting algorithm. “What Is Diabetes? - Niddk.” by the National Institute of Diabetes and Digestive and Kidney Diseases helped give background to the project on diabetes. ChatGPT helped by providing examples on how to speed up tuning hyperparameters for this large dataset.

7 Source Code

https://github.com/StephenButters13/diabetes-detection-project