From Lifestyle to Prediction: Machine Learning Models for Early Diabetes Detection Using Non-Medical Data

Aylin Hadzheiva

December 7th, 2024

Abstract

Diabetes has become a significant global health concern, affecting millions of individuals worldwide. Traditional diagnostic methods, such as blood tests and glucose monitoring, are highly effective but may not be accessible to many due to high costs and the necessity of medical appointments. This study aims to explore the potential for predicting diabetes using non-medical data, such as age, income, physical activity, dietary habits, and water intake—variables that are more easily accessible to individuals in their daily lives. Using five different binary classification models, this research applies advanced machine learning techniques, including hyperparameter tuning and methods to address class imbalance, to predict the likelihood of diabetes. The results demonstrate that a model trained on these lifestyle factors achieved a recall score of approximately 0.7, indicating a promising ability to identify individuals at risk for diabetes using non-invasive, cost-effective features.

In addition, this study investigates the importance of diet and food-related factors in diabetes prediction. Feature importance analysis revealed that factors related to dietary intake play a significant role in predicting diabetes risk, suggesting that the quality and composition of food intake may be crucial to the development of the disease. These findings motivate further research into the correlation between food quality, diet, and diabetes, providing a basis for future studies to explore the potential of using dietary patterns and lifestyle data for early diabetes detection and prevention. The ability to predict diabetes based on lifestyle factors, particularly diet, could contribute to a more accessible, preventative approach to managing diabetes risk.

Introduction

Diabetes is one of the leading chronic diseases worldwide, with its prevalence continuously rising. According to the World Health Organization (WHO), the global prevalence of diabetes has quadrupled over the past three decades and is expected to increase further in the coming years [1]. Currently, over 400 million people are living with diabetes globally, with the majority suffering from Type 2 diabetes (T2D), which is closely linked to lifestyle factors such as diet, physical activity, and obesity [2]). The increasing prevalence of diabetes presents a significant public health challenge, with the disease contributing to numerous complications, including cardiovascular disease, kidney failure, and nerve damage [1].

Many studies on diabetes prediction have traditionally relied on clinical data, such as blood glucose levels, insulin sensitivity, and other lab results. While these data points are reliable, they are not universally available, and medical tests can be costly. According to a report by the ADA [3], the average cost of diabetes management in the United States is approximately $9,601 per year per person. For individuals without insurance or access to healthcare, the financial burden of regular medical tests and doctor visits is prohibitive, resulting in delayed diagnosis and treatment. The lack of widespread access to these diagnostic tools means that many people, particularly in developing countries, remain undiagnosed, leading to a delay in intervention and worsening health outcomes.

Given the high cost of medical testing, there is an urgent need for a more accessible and cost-effective way for individuals to assess their diabetes risk. Lifestyle factors, such as diet, physical activity, and socioeconomic status (e.g., income, education), have been shown to influence the likelihood of developing diabetes [4]. In this context, the use of machine learning (ML) to predict diabetes based on readily available data offers a promising alternative. This project aims to investigate whether a machine learning model can predict diabetes risk using these lifestyle factors alone, providing a non-invasive, low-cost alternative for early detection.

Recent studies have highlighted the significant role of diet in diabetes development. A high-sugar, high-fat diet is commonly associated with the onset of Type 2 diabetes, as it contributes to insulin resistance and Iight gain [5]. However, identifying specific foods or dietary patterns that dwerectly cause diabetes requweres extensive experimental studies and clinical trials, which are often costly and time-consuming. For example, identifying the role of milk or other specific foods in diabetes risk requweres longitudinal studies with controlled diets, which may

take years to complete [6]. While identifying the precise role of specific foods in diabetes risk requweres costly and extensive clinical trials, this study aims to confirm whether dietary habits, genes, and habits play a significant role in diabetes development. If successful, this would provide strong justification for further experimental research and investment, underscoring the importance of exploring these factors as part of a comprehensive approach to diabetes prevention and management.

This paper will proceed by detailing the methodology and implementation of my machine learning model for diabetes prediction, focusing on the use of non-medical, lifestyle-related features. Following this, I will discuss my approach to evaluating the model's performance, including the handling of data imbalances and feature selection. Finally, I will present insights from my evaluation, culminating in a comprehensive assessment of the model's effectiveness and its potential implications for early diabetes detection and prevention in resmyce-limited settings.

Objectives

The objectives of this study are:

- To develop a predictive model for diabetes using non-medical features such as age, physical activity, income, diet, and water intake
- To explore the significance of various lifestyle-related features in predicting the likelihood of diabetes
- To evaluate the performance of machine learning models in predicting diabetes risk, with a focus on recall and precision, particularly in the context of imbalanced classes

Literature Review

The use of lifestyle factors for disease prediction has been explored in various studies. For instance, the role of socioeconomic status, physical activity, and diet in predicting diabetes has beenwell documented [4]. However, few studies have successfully applied machine learning techniques to predict diabetes using only non-medical data, which limits the generalizability of existing findings. Several studies have used the NHANES (National Health and Nutrition Examination Survey) dataset for diabetes prediction. However, most have focused on clinical data, such as blood glucose levels and laboratory results, and have not explored the use of non-medical data.

Two relevant studies have attempted to use non-medical data for diabetes prediction:

- Feature Selection: One study used an ensemble machine learning method for diabetes prediction but did not clearly define the feature selection process. The study had several issues with data cleaning and feature selection, which resulted in poor performance with a recall of 0.13 [7]. The absence of a structured feature selection process and the possibility of confounding variables likely contributed to the model's poor performance.
- Evaluation Metrics: Another study used a more systematic approach to feature selection but failed to evaluate the model comprehensively. While it reported accuracy, it did not provide metrics such as recall, precision, or AUC, which are critical for evaluating model performance, particularly in imbalanced datasets as overfitting might be present [8]

Methods

Dataset

The dataset used in this study is the NHANES 2017-2018 dataset, which provides comprehensive health-related data collected from a representative sample of individuals in the United States. NHANES (National Health and Nutrition Examination Survey) is a continuous program that assesses the health and nutritional status of the U.S. population, offering a detailed snapshot of demographic, lifestyle, and clinical factors.

The dataset includes a wide range of variables from the following sections:

- Demographics Data
- Food.Dietary Data
- Medical Examination Data
- Laboratory Data
- Questionnawere Data

The 2017-2018 dataset was selected for this study as it is the most recent dataset available prior to the COVID-19 pandemic, making it particularly relevant for understanding lifestyle and health trends before significant disruptions to public health systems and data collection methods occurred. More detailed information about the method collection and the variables can be found here.

Data Feature Selection

For this study, non-medical features that could be easily obtained by individuals at home were specifically focused on, with clinical measurements such as blood glucose levels and other laboratory results

excluded. These features were selected based on a review of existing literature that explores the relationship between lifestyle factors and the risk of developing diabetes.All available variables were **manually** reviewed, and findings from online studies and public health research were cross-referenced to ensure the inclusion of variables that are most likely to affect diabetes risk. This selection process was aimed at identifying features that could be accessed without medical tests, thereby making diabetes prediction more accessible [9]

For this study, the following non-medical features were selected for analysis:

Demographic Information

- SEQN: Participant ID
- RIAGENDR: Gender
- RIDAGEYR: Age
- RIDRETH1: Race
- DMDFMSIZ: Total number of people in the family
- INDFMIN2: Annual family income

Dietary Information

- DRQSDIET: On special diet?
- DRQSDT1: Iight loss/low calorie diet
- DRQSDT2: Low fat/low cholesterol diet
- DRQSDT3: Low salt/low sodium diet
- DRQSDT4: Sugar-free/low sugar diet
- DR1TKCAL: Total energy (kcal)
- DR1TPROT: Protein intake (g)
- DR1TCARB: Carbohydrate intake (g)
- DR1TSUGR: Total sugars intake (g)
- DR1TTFAT: Total fat intake (g)
- DR1TCHOL: Cholesterol intake (mg)
- DR1TALCO: Alcohol intake (g)
- DR1TCAFF: Caffeine intake (g)
- DR1TMOIS: Moisture intake (g)
- DR1320Z: Total plain water intake (g)
- DR1330Z: Total tap water intake (g)
- DR1BWATZ: Total bottled water intake (g)
- DRD340: Shellfish consumption in the past 30 days
- DRD360: Fish consumption in the past 30 days

Examination Data

- BPXPLS: Blood pressure (pulse)
- BMXWT: Iight (kg)
- BMXHT: Height (cm)
- BMXBMI: BMI (kg/m^2)
- BMXWAIST: Waist circumference (cm)

Questionnawere Data

- ALQ130: Average alcohol drinks per day (past 12 months)
- ALQ151: Ever have 4/5 or more drinks every day?
- CBD071: Money spent at supermarket/grocery store
- CBD111: Money spent on food at other stores
- CBD121: Money spent on eating out
- CBD131: Money spent on carryout/delivery foods
- DIQ010: Doctor told you have diabetes?
- DBD895: Number of meals not home prepared
- DBD905: Number of ready-to-eat foods consumed in the past 30 days
- PAQ610: Number of days engaged in vigorous work
- PAQ625: Number of days engaged in moderate work
- PAQ640: Number of days walked or bicycled
- PAQ655: Number of days engaged in vigorous recreational activities
- PAQ670: Number of days engaged in moderate recreational activities

Data PreProcessing

Patient Exclusion The following participants were excluded:

- Individuals with incomplete diabetes information were excluded.
- Pregnant individuals and those younger than 16 or older than 85 were excluded to ensure that the study focused on the general adult population.
- Individuals who take diabetic medications.

Prescribed medications were excluded as features, as medications could introduce confounding effects (e.g., side effects that may influence the relationship between lifestyle factors and diabetes).

Handling Missing Data Missing values for continuous variables were imputed using the mean of the respective columns. "Do not know" or "Refused" responses were excluded to ensure high-quality data.

Feature Scaling and Encoding

- Continuous variables were scaled using the StandardScaler, which normalizes features by removing the mean and scaling to unit variance. This is necessary for many machine learning algorithms to work effectively, especially when features have different units or scales.
- Categorical variables were one-hot encoded to transform them into a format suitable for machine learning models.

Feature Relationship, Collinearity, and Confounding Variables The distribution, count, and relationships between individual variables and the entire

feature set were explored through various visualizations, including the correlation matrix, box plots, and histograms. These plots were crucial for identifying potential issues that could undermine the model's ability to make accurate predictions, such as multicollinearity, outliers, and non-optimal feature distributions. For example, extreme values in variables such as *caloric intake* and *physical activity levels* were identified, leading to decisions about whether to cap, transform, or remove these data points to maintain the robustness of the model. Many machine learning algorithms assume that data is symmetrically distributed, so variables that exhibited significant skewness were flagged for transformation, such as applying logarithmic transformations to correct for non-normal distributions.

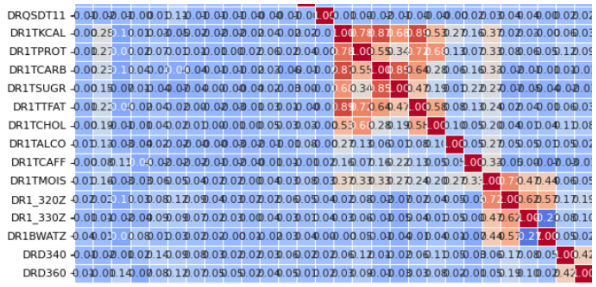

Figure 1: *Correlation Matrix Before Changes*

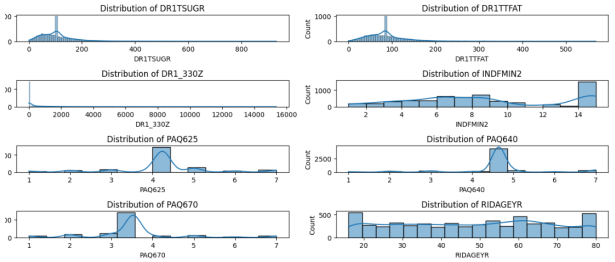

Figure 2: *Histogram of Key Variables*

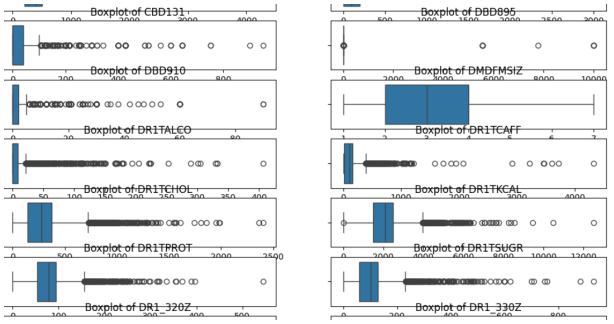

Figure 3: *Box Plot of Key Variables*

Based on the analysis of these plots, several correlations between variables were observed, leading to the removal of some of them. For example, *Body Mass Index (BMI)* was dropped due to its high correlation with height and weight, while these latter

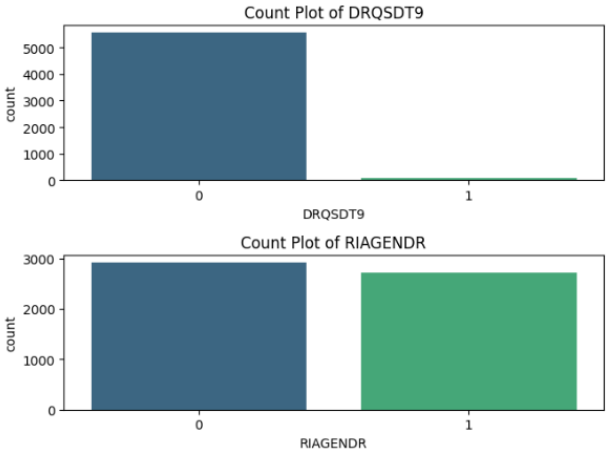

Figure 4: *Count Plot of Categorical Variables*

variables were retained. Similarly, *moisture* was excluded while water intake was kept, as water intake is a more dwerect and relevant factor for analysis.

In certain cases, Principal Component Analysis (PCA) was employed to address multicollinearity by combining highly correlated variables into uncorrelated components. PCA is a dimensionality reduction technique that transforms a set of possibly correlated variables into a smaller set of linearly uncorrelated components, called principal components, while retaining most of the original variance. This technique is particularly useful when multiple variables exhibit strong correlations, as it reduces redundancy and ensures that the resulting components capture the maximum variance from the original dataset.

In this study, PCA was applied to combine variables such as *protein, fat, carbohydrates, and calories*, which are highly correlated, into a single composite variable. This transformation reduces the complexity of the dataset while preserving the information necessary for diabetes prediction. Similarly, all *physical activity* variables were combined using PCA to mitigate the risk of multicollinearity, which occurs when predictor variables are highly correlated with each other. By reducing the dimensionality of the dataset, PCA prevents inflated standard errors and ensures that the model can estimate the relationships between features more reliably.

After applying PCA, the correlation matrix was recalculated, and the new features exhibited no correlations above 0.7

These are just snapshots of some of the plots. More detailed plots and other ones such as scatter plots for pairs of variables, are available in the codebase. However, due to the large size of the plots (e.g., 49 by 49 feature scatter plots), these have not been included in the paper. These plots, which help compare how the final results align with the initial observed relationship between the features, can be accessed the

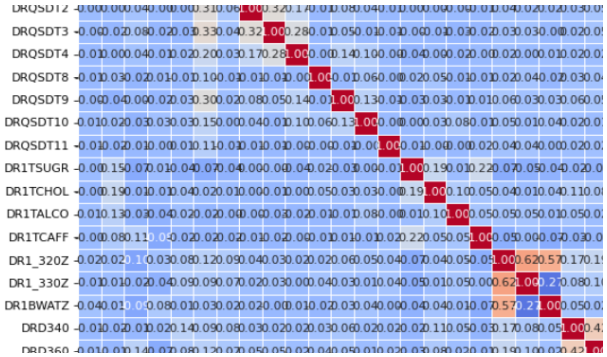

Figure 5: *Correlation Matrix After Changes*

codebase.

The initial dataset had 9254 samples and 49 chosen feature that are shown in the table above. After cleaning, the final dataset contained **44 features and 5643 samples**.

Methods

Models

The data is not balanced. The ration is 1:10 in the training and testing data as only a few people have reported to be diagnosed with diabeted.

Decision Tree

1. Initial Results a) Depth vs Recall Plot

Deeper trees tend to overfit the training data, capturing noise and reducing generalizability. Shallow trees, on the other hand, may underfit the data by failing to capture important patterns. The depth of the tree was varied to assess its impact on model performance, specifically recall and avoiding overfitting the model. Results shown by Figure 6

b) Results without Addressing Class Imbalance

Classification Report (Testing Data):

Class	Precision	Recall	F1-Score	Support
0	0.88	0.97	0.92	984
1	0.39	0.14	0.21	145

Table 1: *Classification Report (Testing Data) for Decision Tree*

Classification Report (Training Data):

Class	Precision	Recall	F1-Score	Support
0	0.88	0.98	0.93	3862
1	0.64	0.21	0.32	652

Table 2: *Classification Report (Training Data) for Decision Tree*

Explanation: Despite achieving high accuracy, the model's recall for the minority class (1, diabetes) was very low. This is due to the imbalance in the

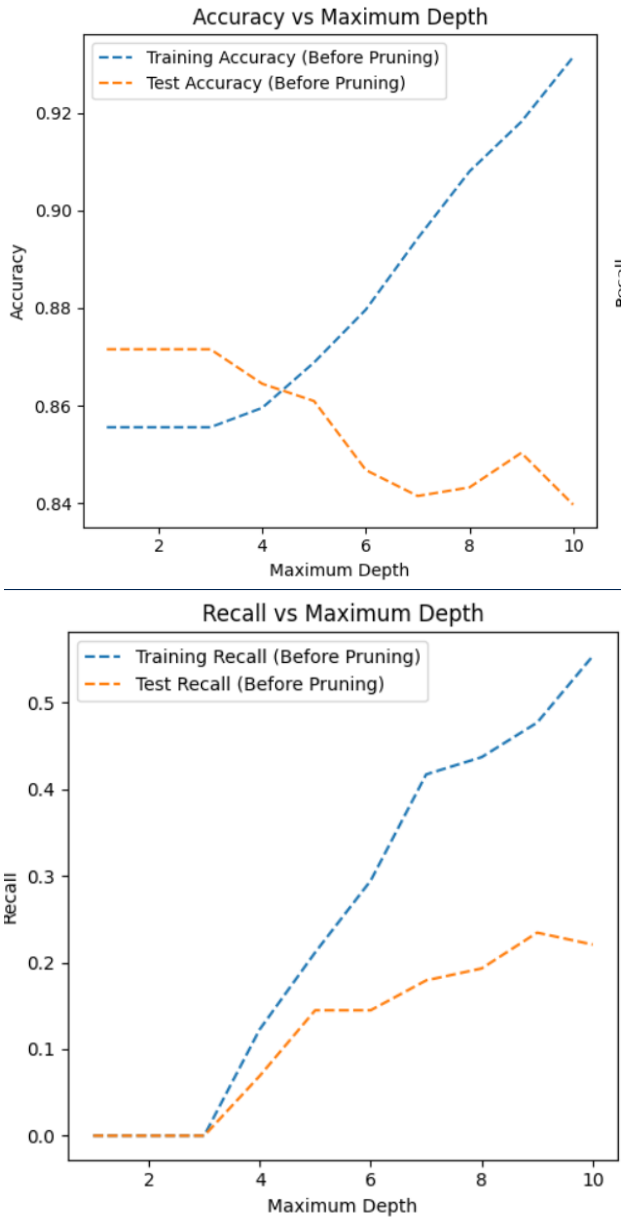

Figure 6: *Depth vs Recall for Decision Tree Model*

dataset, where the majority class (non-diabetic) dominates the predictions. However, there is no significant overfitting as evidenced by the similar performance between the training and testing datasets.

2. Handling Class Imbalance a) Techniques Used: The class imbalance was addressed by using SMOTE (Synthetic Minority Over-sampling Technique). SMOTE generates synthetic examples for the minority class by interpolating between existing minority class examples. This helps the model learn to recognize the patterns of the minority class better [10].

b) Results with SMOTE: After applying SMOTE, the recall for the minority class improved significantly, reaching values close to 0.65. This demonstrates the effectiveness of oversampling in improv-

ing model performance on imbalanced data.

c) Other Techniques for Class Imbalance: In addition to SMOTE, other methods can be applied to address class imbalance, including undersampling, classweighting, and ensemble methods such as Balanced Random Forests. Undersampling involves reducing the size of the majority class to balance the data, though it risks losing valuable information. Classweighting adjusts the importance of each class during model training, allowing the model to pay more attention to the minority class. Ensemble methods, such as Balanced Random Forests, combine multiple classifiers while adjusting for class imbalance. While SMOTE was the primary method used in this study due to its ability to generate synthetic data points for the minority class without altering the majority class's distribution, these other techniques also offer valid approaches for improving performance on imbalanced datasets.

3. Model Adjustments a) Hyperparameter Tuning for Decision Trees: Hyperparameter tuning was performed using GridSearchCV, a method that explores a wide range of hyperparameters to find the best combination for the model. The following parameters were adjusted:

- `criterion`: The function to measure the quality of a split, tested with `gini` (default) and `entropy`.
- `splitter`: The strategy used to split at each node, tested with `best` (best split) and `random` (random split).
- `max_features`: The number of features to consider when looking for the best split, tested with `sqrt`, `log2`, and `None`.
- `max_depth`: The maximum depth of the tree, tested with values of `None` (no limit), 3, and 5 to prevent overfitting.

GridSearchCV used 5-fold cross-validation to determine the best combination of hyperparameters, ensuring that the model's performance was consistent across different data splits and not overly reliant on a single train-test partition [11].

b) Pruning: To reduce overfitting and improve generalizability, cost-complexity pruning was applied. This technique prunes the tree by removing branches that provide little predictive poIr. The pruning path was determined by evaluating different values of the `ccp_alpha` parameter, which controls the amount of pruning. The optimal model was selected based on recall performance, as I prioritized improving the model's ability to correctly identify the minority class (diabetes) [12] [13] [14].

c) Results after Model Adjustments and class Imbalance: After applying hyperparameter tuning,pruning and balancing the class, the Decision Tree model demonstrated significant improvements for recall but at the expense for precision, and the f1 score was still low.

Class	Precision	Recall	F1-Score	Support
0	0.92	0.69	0.81	984
1	0.24	0.70	0.33	145

Table 3: *Classification Report (Testing Data) for DT Model Adjustments*

Final Conclusion: The combination of SMOTE, hyperparameter tuning, pruning, and threshold adjustment improved the performance of the Decision Tree model, especially in terms of recall for the minority class. Nevertheless, the results are still not that satisfactory because of the low f1 score. So, these adjustments were supposed to show if the class imbalance was really the issue, and as far as it goes for the decision tree model, I can conclude that it is very good model given the chosen features at predicting diabetes.

** In order to double check that this model is working, the same model was run but with laboratory features like insulin levels and glucose resulting in nearly perfect f1 score. Therefore, the chosen features do not work verywell with the DT model.

Random Forrest

1. Initial Results a) Depth vs Recall Plot

Results shown by Figure 7. The chosen depth was 5. A modified version of K-Fold i.e. stratified K-Fold Cross Validation was used with this model. The reason is that "K-Fold Cross Validation is not suitable for handling imbalanced data because it randomly divides the data into k-folds. Folds might likely have negligible or no data from the minority class resulting in a highly biased model" [15]

b) Results without Addressing Class Imbalance If the class imbalanced is not addressed, then the model overfits and gets a low score of 0.2 for recall. **b) Results with Addressing Class Imbalance** There are 2 ways that I try to balance the data.

- `BalancedRF`: This models gets good results, but overfits
- `SMOTE+StandartRF`: This is the option that worked and the results are shown by Table 4

Classification Report (Testing Data):

Class	Precision	Recall	F1-Score	Support
0	0.92	0.77	0.84	970
1	0.30	0.58	0.39	159

Table 4: *Classification Report (Testing Data) for Random Forrest*

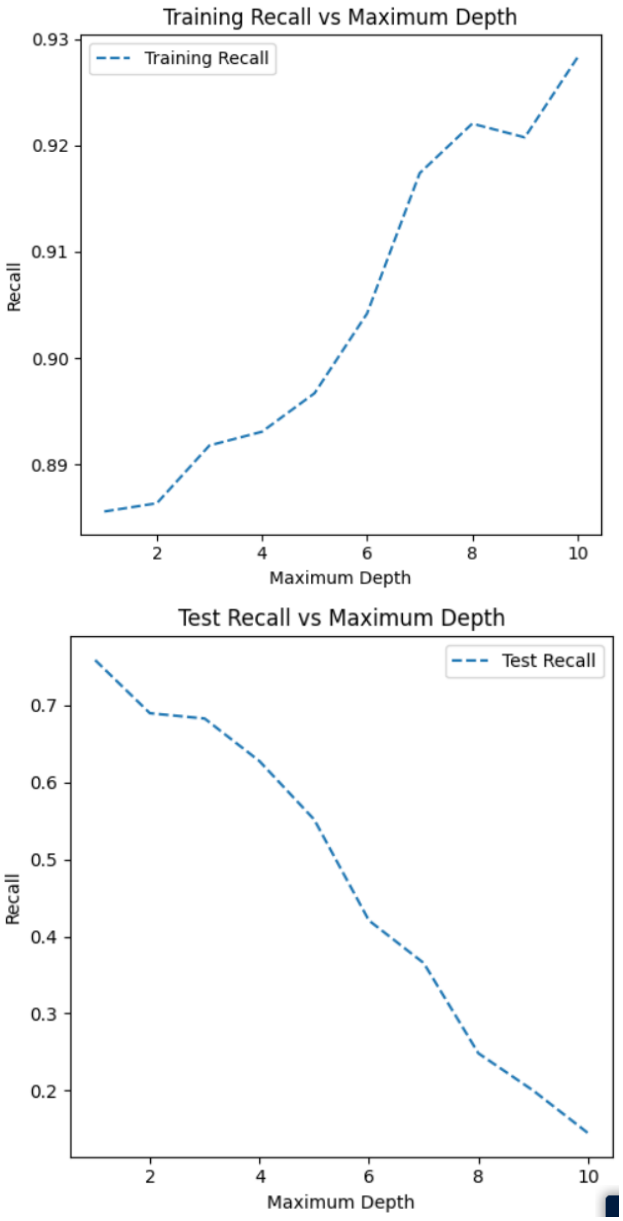

Figure 7: *Depth vs Recall for Random Forrest model*

Classification Report (Training Data):

Explanation: These results are without any model adjustments and are similar to the final decision tree model but after applying adjustments on that model. Therefore, the random forest model is doing a better job when recall is the main testing assessment between models and when class imbalance is present. The reason is that the random forest classifier offers the advantage of incorporating class weights making it sensitive to class imbalance by penalizing incorrect predictions of the minority class. Additionally, it combines both ensemble learning and sampling techniques, which allow it to reduce the dominance of the majority class by downsampling and training trees on a more evenly distributed dataset.

There are several other benefits of using a random forest classifier when working with imbalanced datasets. It is a robust model that outperforms a single decision tree due to its ensemble approach. By aggregating multiple trees, the risk of overfitting and bias is reduced, leading to more stable and accurate results. The classifier also handles missing datawell, ensuring accuracy even when a significant portion of the data is missing. Furthermore, it is effective at managing large datasets with high-dimensional features [15].

Class	Precision	Recall	F1-Score	Support
0	0.90	0.79	0.84	3876
1	0.82	0.91	0.96	3876

Table 5: *Classification Report (Training Data) for Random Forrest*

Logistic Regression

1. Initial Results a) Results without Addressing Class Imbalance

Simple vanilla logistic regression model. L1 and L2 regularizations are techniques used to prevent overfitting in machine learning models, including logistic regression. L1 encmyages sparsity, effectively performing feature selection by driving some coefficients to zero. On the other hand, L2 regularization prevents the model from assigning too much importance to any particular feature and helps maintain a more generalized model. Regularization is used to improve generalization, but without addressing the imbalance in the data, the model still has poor performance on the minority class.

Classification Report (Testing Data):

Class	Precision	Recall	F1-Score	Support
0	0.88	0.99	0.93	984
1	0.55	0.12	0.19	145

Table 6: *Classification Report (Testing Data) for Random Forrest*

Classification Report (Training Data):

Class	Precision	Recall	F1-Score	Support
0	0.88	0.98	0.93	3862
1	0.67	0.20	0.30	652

Table 7: *Classification Report (Training Data) for Random Forrest*

2. Handling Class Imbalance a) Techniques Used:

In this study, class imbalance in the logistic regression model was handled using two primary techniques: classweighting and threshold adjustment. First, class weights were incorporated using the parameter `class_Iight='balanced'` in logistic regression. This approach automatically adjusts the Iight

of each class in the model based on the class distribution in the training data, assigning a larger Iight to the minority class to account for its underrepresentation. Additionally, I also experimented with manually tuning the class weights via grid search and randomized search, which alloId us to explore differentweightings for both classes and determine the best configuration for maximizing recall and minimizing misclassifications for the minority class.

b) Results:

When working with the weights the model was able to improve the recall for the minority class (diabetes). However, the improvement was not as significant as when the class weights were manually tuned using grid search. The grid search explored different Iight combinations for both classes, optimizing the model for recall performance. The results are quite similar to what I observed in the other models with increase for recall and decrease in precision. There is no overfitting, as was the case with the decision tree model. Therefore, there is no need for model adjustments.

Class	Precision	Recall	F1-Score	Support
0	0.95	0.70	0.80	984
1	0.27	0.77	0.40	145

Table 8: *Classification Report (Testing Data) for Logistic Regression*

Classification Report (Training Data):

Class	Precision	Recall	F1-Score	Support
0	0.95	0.70	0.81	3862
1	0.31	0.79	0.45	652

Table 9: *Classification Report (Training Data) for Logistic Regression*

Besides using the balanced class weights threshold adjustment and hyperparameter tuning were also considered. Thresholding is a technique used to adjust the decision threshold for classification. By default, logistic regression uses a threshold of 0.5, meaning that any probability above 0.5 is classified as class 1 (diabetes), and below 0.5 as class 0 (non-diabetes). However, when dealing with imbalanced data, this default threshold often leads to poor recall for the minority class, as the model is biased towards predicting the majority class. Threshold adjustment allows for increasing the sensitivity of the model to the minority class by loIring the threshold. This increases the number of true positives identified by the model, which improves recall at the cost of potentially decreasing precision. The optimal threshold is typically determined by maximizing the difference between True Positive Rate (TPR) and False Positive Rate (FPR) in the ROC curve [16].

Threshold adjustment can be used alone, but it is most effective when combined with other techniques like grid search and manual classweighting, as both can help fine-tune the model to better recognize patterns in the minority class. In my study, I found that combining these techniques led to the most significant improvement in recall for the minority class while still maintaining reasonable precision. And resulted in Best threshold for maximizing recall: 0.55. Below is the corresponding ROC curve.

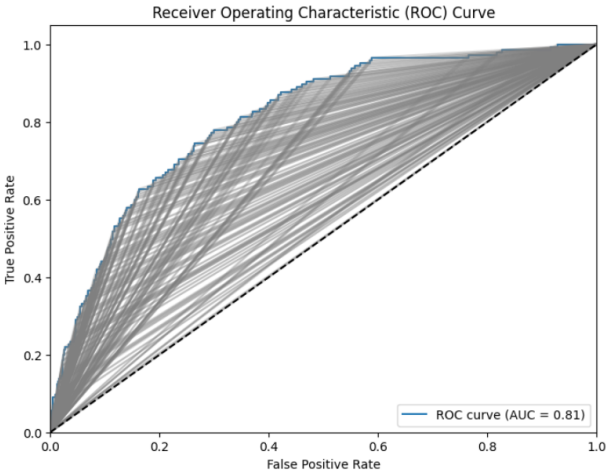

Figure 8: *Combined ROC curve*

Hyperparameter tuning in this case refers to the process of optimizing the logistic regression model by selecting the best combination of hyperparameters, including the regularization strength (`C`) and the class weights. The results are similar to the ones presented in Tables 8 and 9.

SVM The model performs suboptimally, and for the sake of brevity, further details can be found in the accompanying codebase

Naive Bayes The model performs suboptimally, and for the sake of brevity, further details can be found in the accompanying codebase

Results

Given that the Logistic Regression had the higher recall and f-1 score, that model was used to extract the feaure importance of the top 10 features.

Discussion & Application

I was surprised to find that logistic regression performed the best, especially considering that I initially anticipated Random Forest would outperform it. I had this expectation because Random Forest is often seen as robust to class imbalance and more flexible

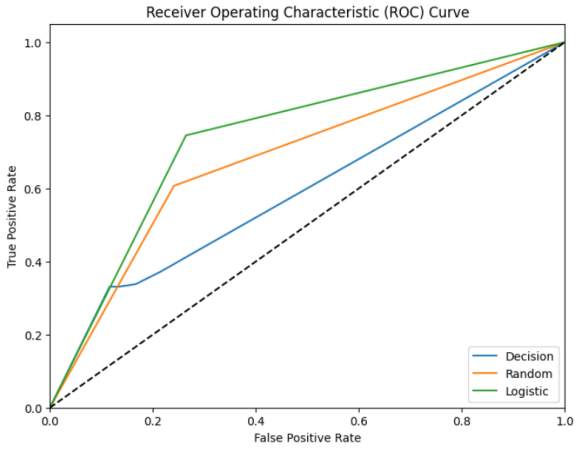

Figure 9: *ROC curve for the Logistic Regression*

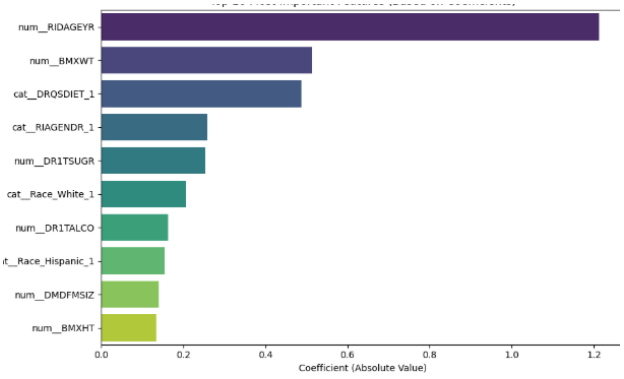

Figure 10: *Top 10 Features Based on Coefficients*

in capturing complex patterns compared to simpler models like logistic regression. However, logistic regression with classweighting and threshold adjustment turned out to be highly effective in this scenario.

From these results, it is evident that the model performswell for the majority class (non-diabetic individuals), achieving a high precision and F1-score. However, the performance for the minority class (diabetic individuals) remains a challenge, with relatively low precision but a strong recall of 0.77. The recall for the minority class, while much improved compared to the initial model, still suggests that there is room for refinement in balancing both precision and recall.

While precision is loIr for the minority class, the high recall indicates that the model is successfully identifying many of the individuals with diabetes, which is crucial for early detection. The trade-off between recall and precision is a common challenge when dealing with imbalanced datasets, and this model strikes a reasonable balance given the constraints of the data and the methodology used.

The F1-score for the minority class is 0.40, which suggests that while the model identifies diabetic individuals with relatively high recall, the false positives still need to be addressed for better overall performance. The decision threshold, which was adjusted to maximize recall, helped improve this aspect, but further adjustments could refine the balance between precision and recall.

Moreover, the results suggest that the most significant predictors of diabetes risk in my model include age, dietary habits (e.g., "On special diet," "Iight loss/Low calorie diet," and "Total sugars"), Iight, and socioeconomic factors (e.g., "Gender," "Race"). These features appear to be much more influential in predicting diabetes risk than other factors, such as physical activity or meals eaten at home.

The selected features for predicting diabetes were somewhat surprising. I initially expected physical activity, given its known impact on health, to play a more significant role than demographic factors like gender or race/ethnicity.

Notably, age and weight (a key measure of obesity) are among the most critical features, both of which are well-documented risk factors for Type 2 diabetes. The inclusion of dietary features, such as whether an individual is on a special diet, follows a low-calorie or low-sugar diet, and their total sugar intake, further emphasizes the role of dietary habits in diabetes risk. These results are consistent with existing research that links poor dietary habits—especially those high in sugar and fat—to increased insulin resistance and a higher likelihood of developing diabetes [5]. Also, alcohol consumption also emerged as a significant predictor of diabetes risk in this model. The coefficient for alcohol (gm) was notably positive, indicating that higher alcohol intake may be associated with a higher risk of diabetes.

Interestingly, socioeconomic factors, such as gender and race, also played a significant role in predicting diabetes risk. In particular, the model highlighted the importance of being male and Hispanic as factors contributing to higher diabetes risk. Therefore certain demographic groups could be more prone to diabetes due to a combination of genetic, lifestyle, and socio-economic factors.

Conversely, features related to physical activity (e.g., "Number of days vigorous recreational activities") and meal preparation habits (e.g., "Meals eaten at home") were not selected as top features by the model. This suggests that, while physical activity and eating habits are undoubtedly important for overall health, they may not have as strong a predictive poIr for diabetes risk in this particular dataset or under the modeling approach used. This highlights an important aspect of feature selection: while some factors may bewell-established risk factors for diabetes, the relative Iight of different features can vary depending on the specific dataset, the chosen modeling technique, and the complexity of the relationships between variables.

The prominence of dietary factors and demographic characteristics in this model is consistent with what I know about diabetes risk and emphasizes the importance of focusing on lifestyle and environmental factors in predicting the disease. Future research could investigate additional features, such as genetic data or other biomarkers, to see if they provide additional predictive poIr. Furthermore, the absence of certain physical activity and meal-related features in the final model suggests that there may be other ways to quantify and represent these factors in a way that improves predictive accuracy.

Ultimately, these findings suggest that lifestyle-related features can serve as strong predictors of diabetes risk. By focusing on variables such as age, Iight, and diet, this model provides a non-invasive, low-cost tool for diabetes prediction that could be used in settings where traditional medical testing is inaccessible. The results underscore the importance of continued research into the role of lifestyle factors in diabetes prevention and the potential for machine learning models to help identify at-risk individuals before clinical symptoms emerge.

Conclusion

In summary, this study demonstrates that lifestyle-related features can serve as effective predictors of diabetes risk when combined with machine learning models. The use of classweighting, threshold adjustment, and hyperparameter tuning has alloId the model to address class imbalance and improve recall for the minority class (diabetes). While the model's precision for the diabetic class can be improved, the overall results indicate that machine learning models based on accessible lifestyle data can offer a cost-effective alternative for early diabetes detection, particularly in resmyce-limited settings.

Key Learnings

- **Choices in Data Handling:** Everything comes down to making choices. Proper data cleaning and decisions about missing values and feature selection dwerectly impact model performance and is crucial like which features to drop or how to handle missing data, can change outcomes.
- **Knowing Ymy Data:** Exploratory Data Analysis (EDA) is essential for understanding relationships, distributions, and data issues. This understanding leads to better feature selection and model performance.
- **Class Imbalance:** Addressing class imbalance with methods like SMOTE, threshold adjustment, and classweighting was crucial for improving recall of the minority class.
- **Hyperparameter Tuning:** Techniques like GridSearchCV and RandomizedSearchCV significantly enhanced model performance, showing the importance of optimizing hyperparameters.
- **Understanding Models:** A deep understanding of machine learning models is necessary for effective fine-tuning, especially when dealing with challenges like overfitting and class imbalance.
- **Model Comparison:** Each model has its strengths; So it is necessary to understand how the models work and the key differences between them, so that one can make the right choices and understand whether using some idea would be smart.

Acknowledgements

I would like to express my gratitude to John Hanke and Chenyu Wang for their support throughout the semester.

References

[1] World Health Organization. "Global report on diabetes". In: *World Health Organization* (Nov. 2021). URL: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[2] Pouya Saeedi et al. "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition". In: *Diabetes Research and Clinical Practice* 157 (Nov. 2019), p. 107843. DOI: 10.1016/j.diabres.2019.107843.

[3] American Diabetes Association. "Economic Costs of Diabetes in the U.S. in 2017". In: *Diabetes Care* 41.5 (May 2018), pp. 917–928. DOI: 10.2337/dci18-0007.

[4] Xin Liu et al. "Socioeconomic status, health-related behaviors, and diabetes risk: a mediation analysis". In: *BMJ Open Diabetes Research and Care* 11.1 (2023), e003707.

[5] Semere Yazla. "Obesity, insulin resistance, and type 2 diabetes: associations and therapeutic implications". In: *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* 13 (2020), p. 3611.

[6] Yongin Choi et al. "Longitudinal changes in diet quality and food intake before and after diabetes awareness in American adults: the Coronary Artery Risk Development in Young Adults (CARDIA) study". In: *BMJ Open Diabetes Research and Care* 11.1 (2023), e003800.

[7] Jack Semer. *NHANES-diabetes*. GitHub repository. 2023. URL: https://github.com/semerj/NHANES-diabetes/blob/master/NHANES.ipynb.

[8] Emin Ozhisarcikli. *diabetes-prediction*. GitHub repository. 2023. URL: https://github.com/eozhisarcikli/diabetes-prediction/blob/main/Healthcare%20Data%20Preparation%20%26%20Comparing%20Different%20Models.ipynb.

[9] Dariush Mozaffarian et al. "Lifestyle risk factors and new-onset diabetes mellitus in older adults: the cardiovascular health study". In: *Archives of internal medicine* 169.8 (2009), pp. 798–807.

[10] "Overcoming Class Imbalance using SMOTE Techniques". In: *Analytics Vidhya* (Oct. 2020). URL: https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/.

[11] "How to Determine the Best Split in Decision Tree?" In: *GeeksforGeeks* (Feb. 2024). URL: https://www.geeksforgeeks.org/how-to-determine-the-best-split-in-decision-tree/.

[12] "Overfitting in Decision Tree Models". In: *GeeksforGeeks* (May 2024). URL: https://www.geeksforgeeks.org/overfitting-in-decision-tree-models/.

[13] "Pruning decision trees". In: *GeeksforGeeks* (Apr. 2024). URL: https://www.geeksforgeeks.org/pruning-decision-trees/.

[14] "How to Solve Overfitting in Random Forest in Python Sklearn?" In: *GeeksforGeeks* (Sept. 2022). URL: https://www.geeksforgeeks.org/how-to-solve-overfitting-in-random-forest-in-python-sklearn/.

[15] Bilal Hussain et al. "Surviving In a Random Forest with Imbalanced Datasets". In: *Medium* (Mar. 2021). URL: https://medium.com/sfu-cspmp/surviving-in-a-random-forest-with-imbalanced-datasets-b98b963d52eb.

[16] GeeksforGeeks. "What is the default threshold in sklearn Logistic Regression?" In: *GeeksforGeeks* (2023). Accessed on 7 December 2024. URL: https://www.geeksforgeeks.org/what-is-the-default-threshold-in-sklearn-logistic-regression/.

[17] Jason Brownlee. "Cost-Sensitive SVM for Imbalanced Classification". In: *Machine Learning Mastery* (Jan. 2020). URL: https://machinelearningmastery.com/cost-sensitive-svm-for-imbalanced-classification/.

[18] Francesco Scipioni. *The-NHANES-program*. GitHub repository. 2023. URL: https://github.com/Fscipioni/The-NHANES-program/blob/main/Codes/03.%20-NHANES_Modeling.ipynb.

[19] ChatGPT. *ChatGPT Responses*. Personal communication with an AI assistant, December 2024. 2024. URL: https://openai.com/chatgpt.