### 

### 

# The Cutting Edge: Machine Learning for Predicting Diabetes

Project Report

Team 2A

Chun-Hui Tsai

Ruicheng Li

Lalitha Bhupalam

Hilary Nguyen

Abhinav Samavedula

### 

# Table of Contents

[1. Introduction 3](#_ry22an41o269)

[1.1. Executive Summary 3](#_mrfqr43sxw7c)

[1.2. Business Implications 3](#_ezkitgnaelql)

[1.3. Objective 4](#_9mqz2xhlzcyz)

[2. Data 5](#_2kcjdovv6r4k)

[2.1. Data Acquisition 5](#_q9kvm98tapcc)

[2.2. Data Description 5](#_7o3in6cxm0js)

[2.3. Exploratory Data Analysis 6](#_i8co72136o1a)

[3. Modeling 10](#_j72qbdlvfbw1)

[3.1. Logistic Regression 10](#_hcbri518b7ii)

[3.2. Decision Tree 15](#_pyve8lmfybem)

[3.3. Random Forest 19](#_qds9dbu20pxf)

[4. Takeaways 22](#_u0ts3qns7q7f)

[4.1. Results 22](#_184bvfhkndxe)

[4.2. What we learn 22](#_ibtunaqq6nse)

[5. Conclusion 23](#_nmhaigwdt6c2)

## 

## 

## Introduction

### Executive Summary

Our data mining analysis of the Diabetes Prediction Dataset revealed several key insights. We use effective data preprocessing in imbalanced datasets to greatly enhance model performance. Techniques like oversampling and undersampling improved recall rates, which is crucial for accurate medical predictions. Decision trees and random forests become the more suitable models for our dataset. These insights highlight the importance of data mining and model selection in the healthcare industry, guiding effective prediction and early intervention strategies in medical analytics.

### Business Implications

Diabetes, a long-term disease affecting millions of people worldwide, poses substantial public health challenges for people around the world. As of 2023, the Centers of Disease Control and Prevention (CDC) reports that 37.3 million Americans, or 11.3% of the population, have diabetes. Of these, 21 million American adults suffer from type 2 diabetes, which is often associated with lifestyle choices such as being overweight, a diet high in sugar and carbohydrates, and a sedentary lifestyle.[[1]](#footnote-0)

Globally, the situation is also alarming. The International Diabetes Federation (IDF) reports that 10.5% of the adult population aged 20 to 79, or 536.6 million people, have diabetes. This number is expected to rise significantly, with projections suggesting that by 2045, about 783 million adults will be living with diabetes. This rise is driven by factors like urbanization, an aging population, decreasing physical activity, and increasing rates of overweight and obesity.[[2]](#footnote-1)

Diabetes prediction models play a crucial role in the medical field by providing early warning to individuals at risk. These models can help diagnose the disease, potentially reducing its impact and preventing complications. As the prevalence of diabetes continues to rise, both in the U.S. and globally, the importance of these models becomes increasingly evident in managing the widespread health challenge.

### Objective

In analyzing our diabetes dataset, the primary aim is to extract any meaningful insights from various health indicators like age, BMI, and blood glucose levels. This detailed examination helps in understanding the complex interplay of these factors in diabetes. By analyzing attributes such as age, which might indicate the prevalence of diabetes in different age groups, or BMI and blood glucose levels, which are direct indicators of diabetes risk, we plan to transform these data points into visible presentations for us to interpret and interact. This process is critical for us to make reliable conclusions and understand the broad implications of diabetes on health.

## 

## Data

### Data Acquisition

The Diabetes Prediction Dataset on Kaggle[[3]](#footnote-2) includes demographic information like gender and age, as well as lifestyle and health indicators such as BMI and smoking. These indicators are pivotal in understanding the difference in diabetes risk among individuals. Our goal is to identify patterns behind all this data, enabling us to signify the potential risk of developing diabetes and allow for early precautions.

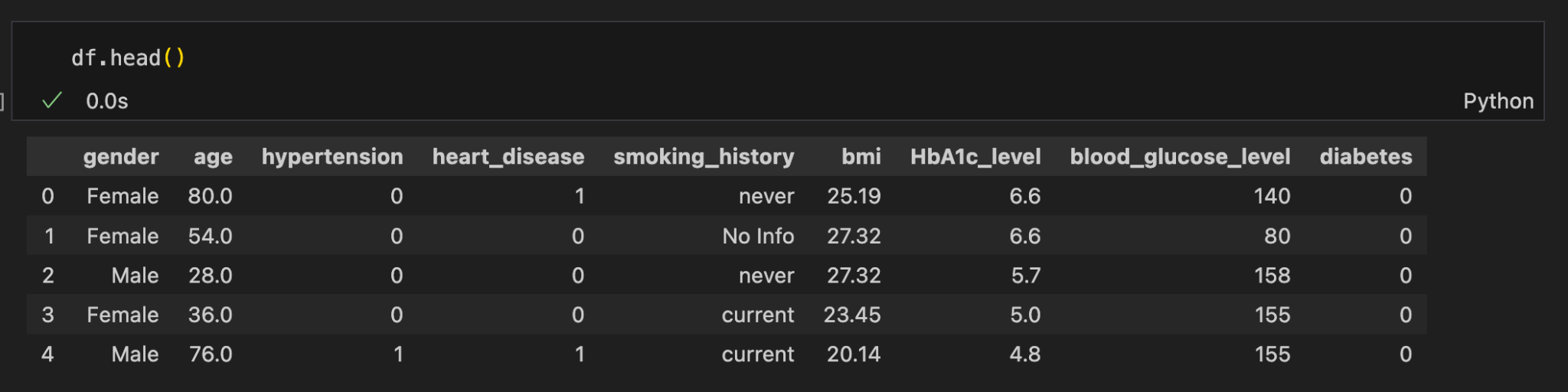
### Data Description

The diabetes dataset consists of various demographic features such as age, gender, body mass index, etc.

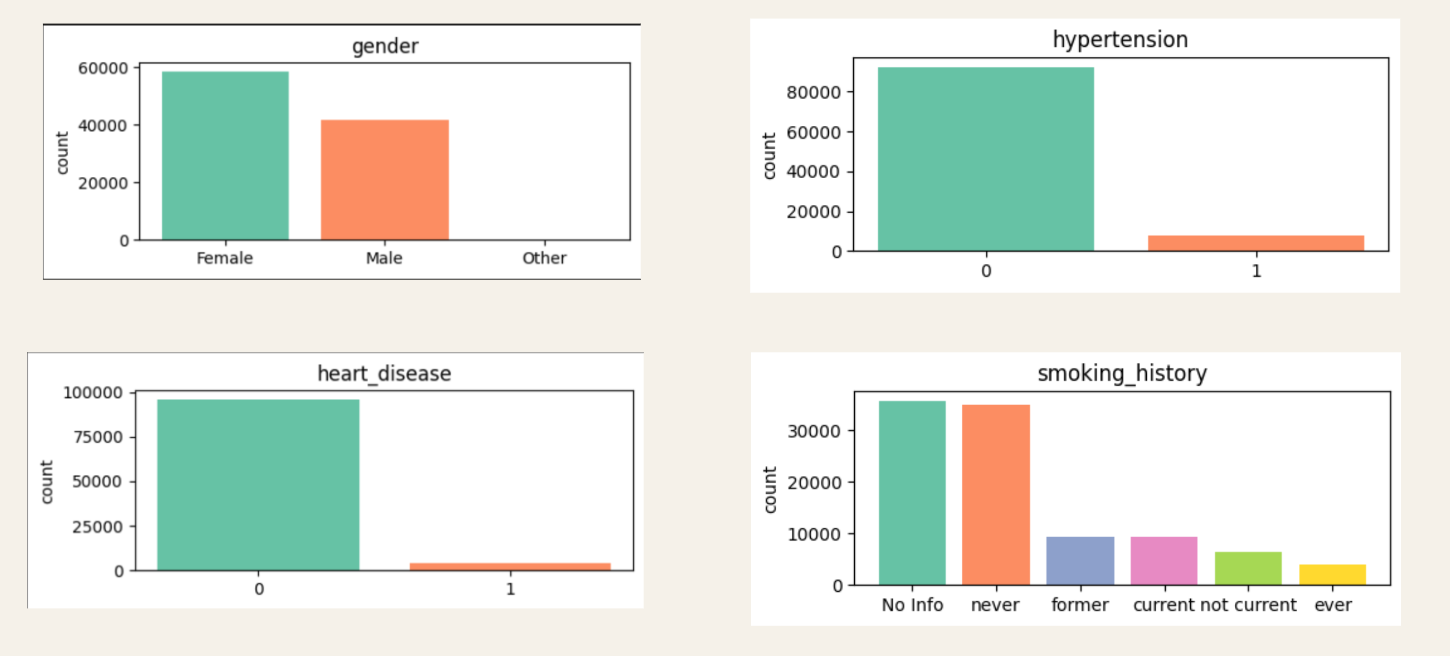
| Gender | Biological sex of the individual can have an impact on their susceptibility to diabetes. |
| --- | --- |
| Age | Important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0-80 in our dataset. |
| Hypertension | A medical condition in which the blood pressure in the arteries is persistently elevated. A value of 0 indicates they don’t have hypertension and 1 means they have hypertension. |
| Heart Disease | A medical condition that is associated with an increased risk of developing diabetes. A value of 0 indicates they don’t have heart disease and 1 means they have heart disease. |
| Smoking History | A risk factor for diabetes and can exacerbate the complications associated with diabetes. In our dataset we have 6 categories: not current, former, No Info, current, never, ever. |
| BMI | Body Mass Index (BMI) is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes. The range of BMI in the dataset is from 10.16 to 71.55. BMI less than 18.5 is underweight, 18.5-24.9 is normal, 25-29.9 is overweight, and 30 or more is obese. |
| HbA1c\_level | HbA1c (Hemoglobin A1c) level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate a greater risk of developing diabetes. Mostly more than 6.5% of HbA1c Level indicates diabetes. |
| blood\_glucose\_level | It refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes. |
| Diabetes | The target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence of diabetes. |

### Exploratory Data Analysis

Our Dataset has a comprehensive range of eight distinct attributes, each of which potentially influences the risk of diabetes. These include gender, age, hypertension, heart disease, smoking history, body mass index (BMI), HbA1c level, and blood glucose level. By analyzing these variables and building models, we aim to discover the relationship between different health indicators and their impact on diabetes risk, offering a more detailed understanding of the disease’s progression.

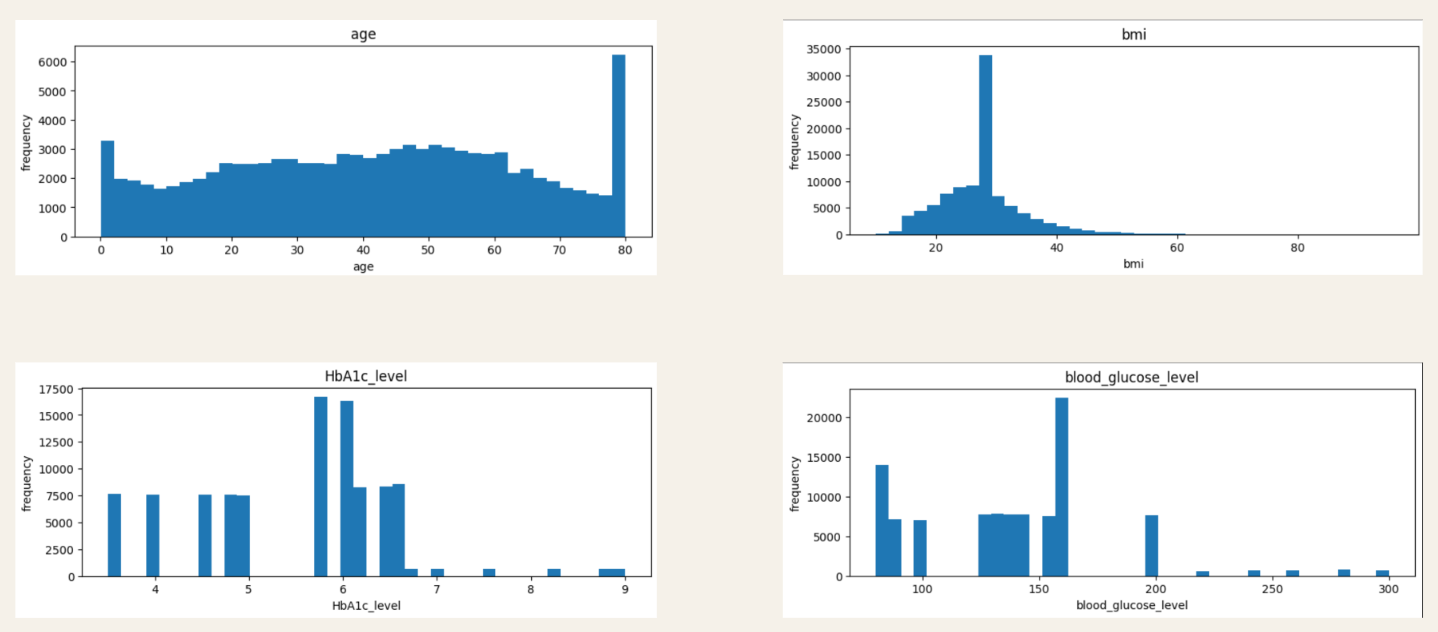


In this analysis, we are examining four distinct categorical graphs, representing a comprehensive dataset of 100,000 data points. It exhibits no missing values across all data categories. Among these categories, the smoking history data stands out due to its somewhat repetitive nature, such as "never," "former," "current," "not current," and "ever." To improve the interpretability of this data, we will undertake a data transformation process at a later stage.

****

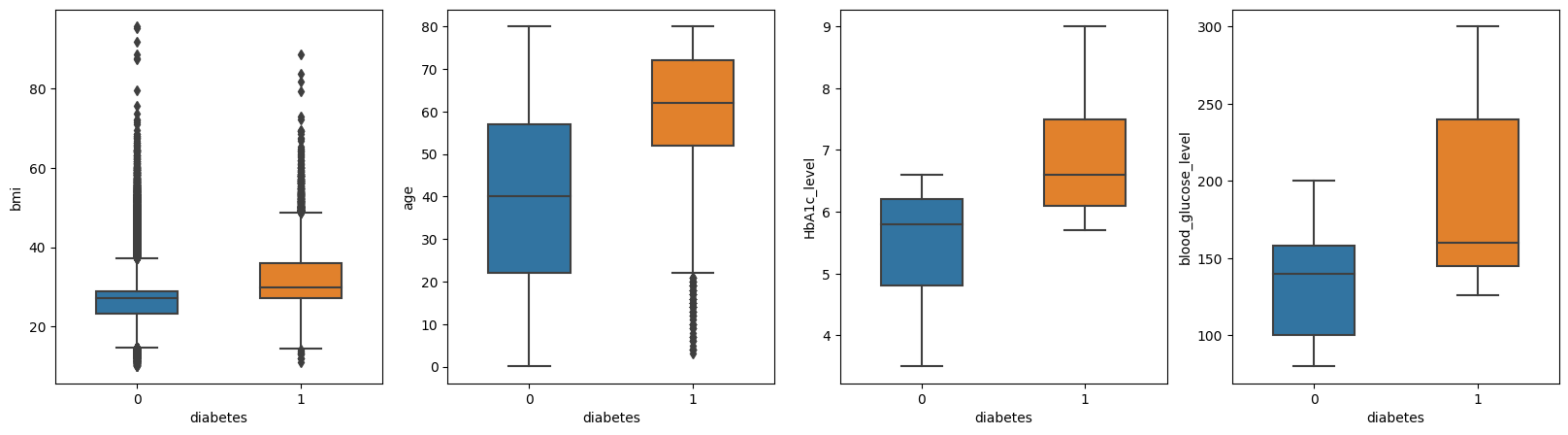
**(Categorical Variable)**

The histograms provided give us an insight into the data distribution of four key health-related variables: age, BMI, HbA1c\_level, and blood glucose level.

****

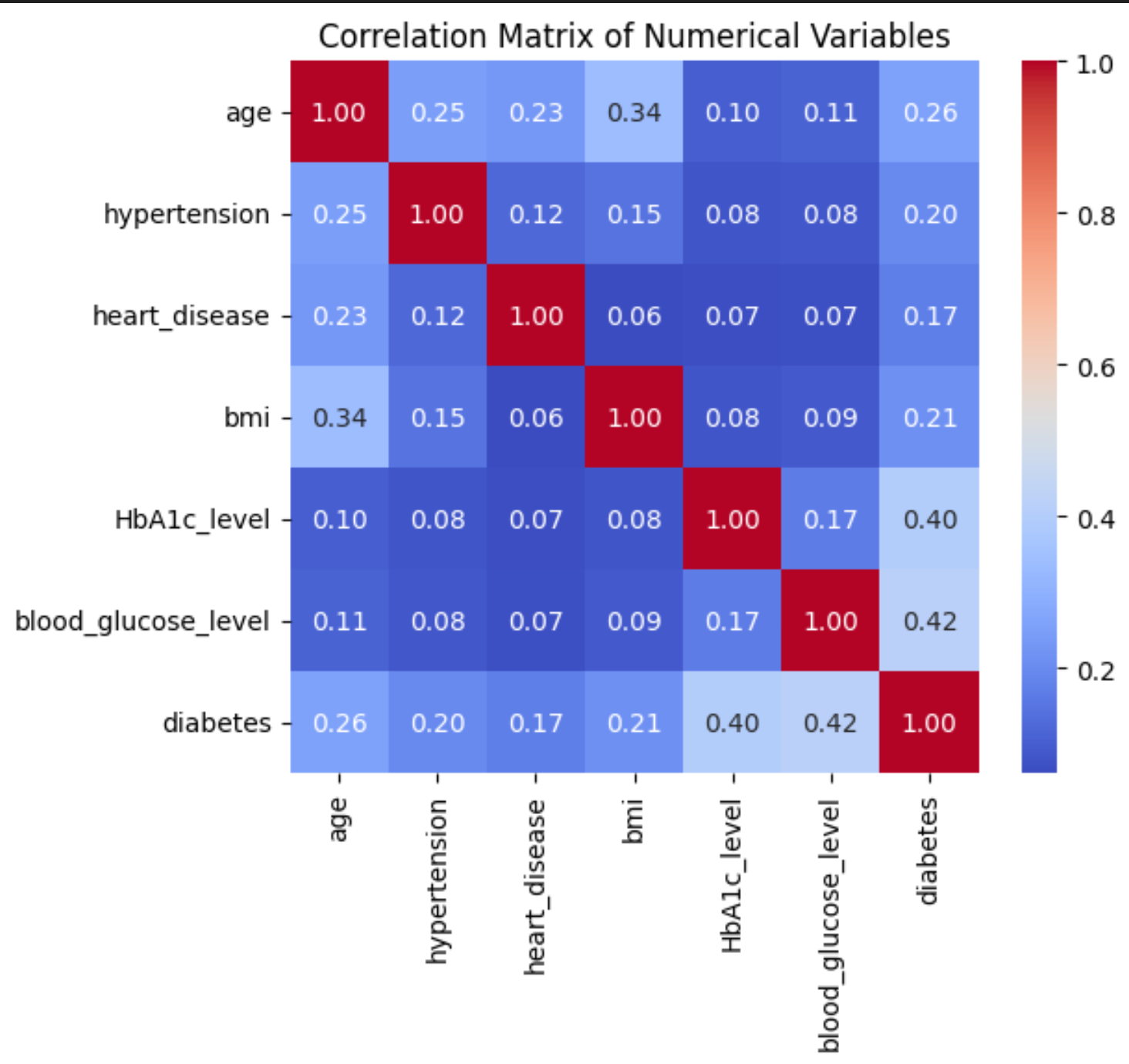
**(Numerical Variable)**

The box plots help understand the dynamics between age, diabetes, and associated blood markers. We put all the numerical values of the data into four different box plots. As we can see in the graph. Elderly people tend to have a high probability of getting diabetes. The trend is further supported by the correlation between diabetes and elevated levels of HbA1c and blood glucose as indicators of the disease.

****

**(Box Plot)**

The correlation matrix provides a visual representation of the relationship between various numerical health-related variables. The HbA1c level and blood and blood glucose level show strong and positive correlations with diabetes. This indicates that as blood glucose increases, so does the likelihood of diabetes.

****

**(Correlation Matrix)**

## 

## Modeling

We aim to predict diabetes in patients based on their medical history and demographic information accurately and evaluate the effectiveness of various analytical approaches when applied to this dataset.

To materialize these insights, we chose three classifier models—logistic regression, decision tree, and random forest. The selection is based on the following considerations:

1. Logistic Regression: Logistic regression was chosen because of the substantial size of our dataset. Its simplicity and computational efficiency make it particularly suitable for handling large datasets effectively.

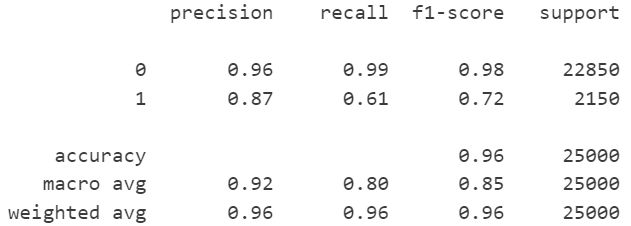
2. Decision tree: As our dataset comprises a mix of numeric and categorical variables, the decision tree is fairly robust across variable types, and can capture complex, non-linear patterns within the data —areas where logistic regression may exhibit weaker performance.

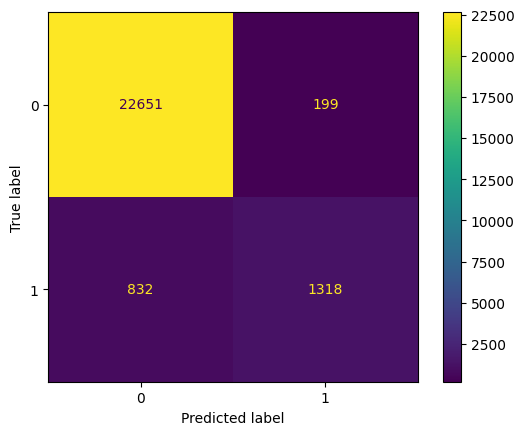
3. Random tree: Recognizing a single tree is prone to overfitting and lacks stability, we adopt a random forest model. This ensemble learning approach could reduce both variance and bias by aggregating predictions from multiple trees to improve generalization performance.

The models are trained on a random 75% subset of our data and tested with the remaining 25%. We evaluate each model using a confusion matrix and classification report demonstrating precision, recall, and f1-score. However, we mainly focus on the recall rate for class 1 because it specifically measures the ability of the classifier to detect all positive diabetes cases. In a healthcare scenario, false negatives signify overlooking potentially critical cases, which can have serious consequences. Therefore, our goal is to select the final classifier with the highest recall rate to minimize this risk.

### Logistic Regression

To accommodate logistic regression's requirement for numeric features, we transform all categorical variables into numerical form through one-hot encoding during model construction. Ｗithout additional preprocessing techniques, the test results reveal a 0.96 accuracy score but a 0.61 recall rate. This implies that only 61% of actual diabetes patients are correctly identified, which clearly doesn’t meet our expectations.





To enhance our results, we implemented the following preprocessing steps:

Step 1 – Remove rows under the gender column where the value is 'other', as it doesn't denote male or female.

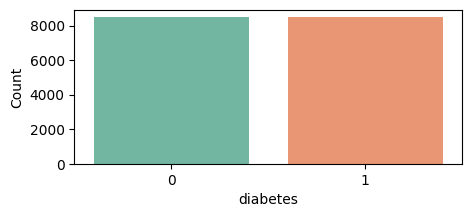
Step 2 – Consolidate categories with similar or overlapping meanings under the smoking\_history column. Transform 'not\_current' into 'former' and 'ever' into 'current'.

Step 3 – Standardize all numeric variables, including 'age', 'bmi', 'HbA1c\_level', and 'blood\_glucose\_level', as logistic regression is sensitive to the scale of input features. This ensures that the model produces reliable results across features with varying scales.

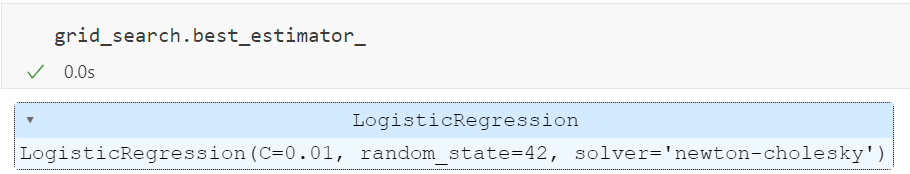
We observe a significant class imbalance in our dataset, with 91.5% belonging to class 0 (non-diabetes) and only 8.5% to class 1 (diabetes). This imbalance leads to a biased model that performs poorly on the minority class. Therefore, to address this problem, we employ two main resampling approaches: random oversampling and random undersampling, known for their simplicity and effectiveness. Furthermore, we utilize GridSearchCV for hyperparameter selection, which helps identify the optimal parameter combination for model performance through cross-validation.

Step 4 – Resampling and hyperparameter selection

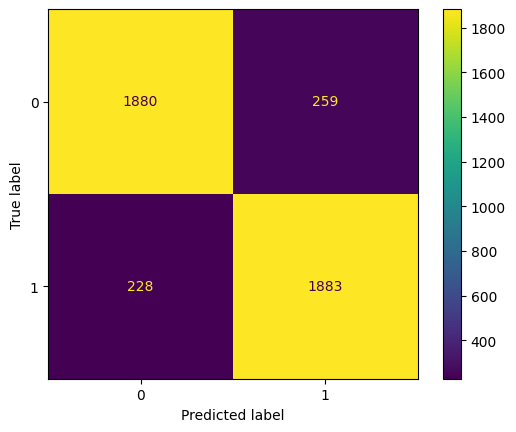
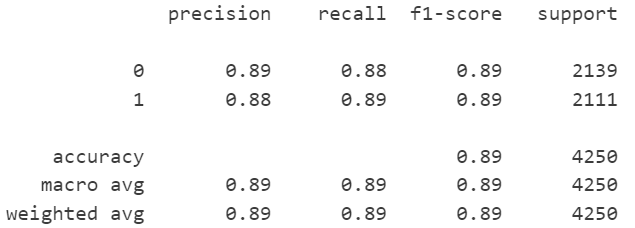
1. RandomUnderSampler from the imbalanced-learn library creates a balanced class with 8,000 counts in both classes, 16,000 instances in total.



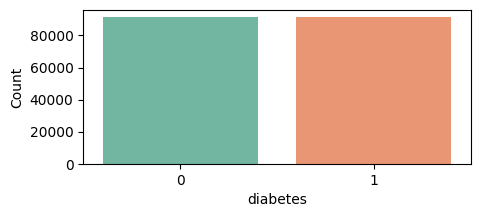
With this new dataset, we run the model, integrating hyperparameter selection that yielded the best estimators. The selected parameters include penalty=L2, C=0.01, and solver='newton-cholesky'. 'Newton-cholesky' is a suitable choice, particularly when the number of samples is significantly greater than the number of features.



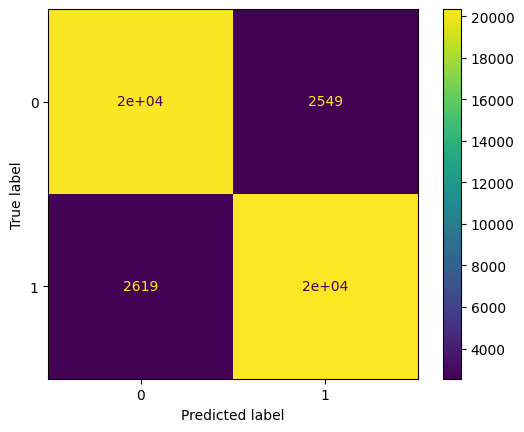
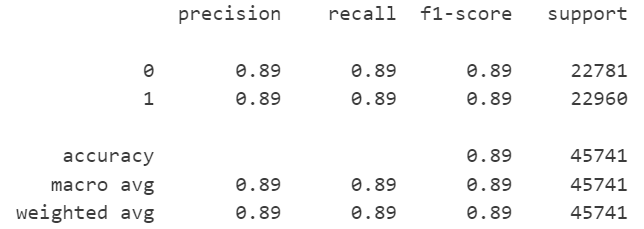
The results are shown below. While the overall accuracy appears lower than that before preprocessing, there is a notable improvement in the recall rate (from 0.61 to 0.89).



1. RandomOverSampler from the imbalanced-learn library creates a balanced class with 80,000 counts in both classes, 160,000 instances in total.



With the same process and parameters selected, we also obtain similar results after oversampling.



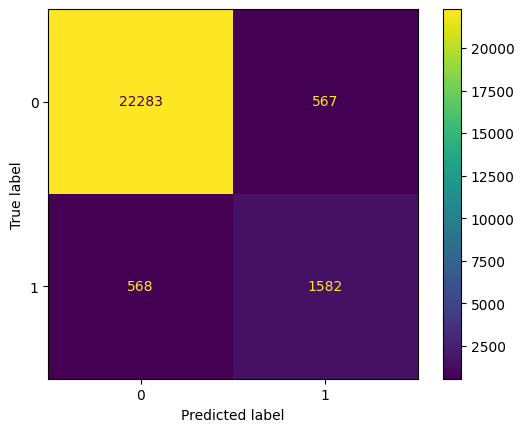
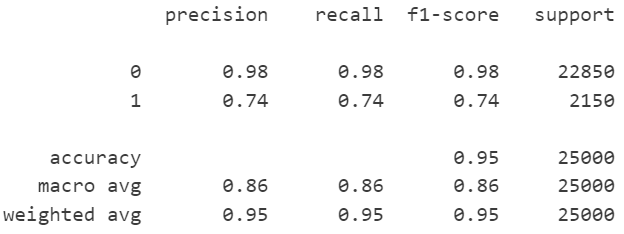
This is the extent to which we believe the model could be improved. Below is a summary of all our logistic models' performance. From 0.61 to 0.89, we improve the recall by 46%.

| **Logistics Regression** | Accuracy score | Precision | Recall | F1-score |
| --- | --- | --- | --- | --- |
| **Without pre-processing** | 0.95876 | 0.87 | 0.61 | 0.72 |
| **Undersampling**  **+hyperparameter tuning** | 0.88541 | 0.88 | 0.89 | 0.89 |
| **Oversampling**  **+hyperparameter tuning** | 0.88702 | 0.89 | 0.89 | 0.89 |

### Decision Tree

In the quest for improved results, we tested the decision tree algorithm.

Before any preprocessing, the testing accuracy is 0.95; however, there is a concern as 26% of true positives are not being identified.



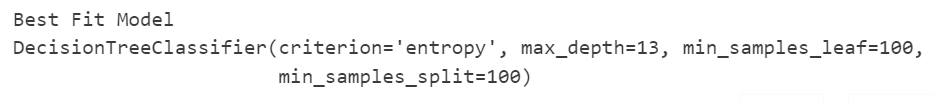
To increase the recall rate, we conduct similar data engineering techniques as what we have done for logistic regression.

Step 1 – Remove rows under the gender column where the value is 'other'

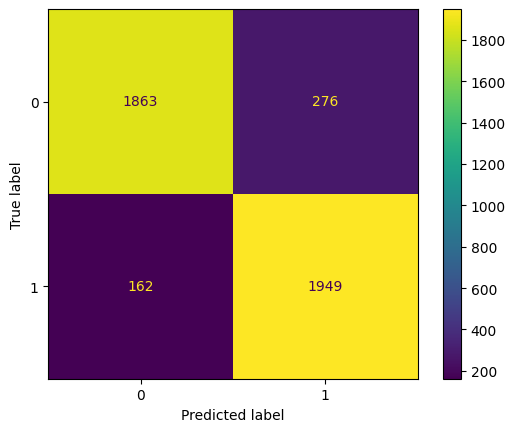
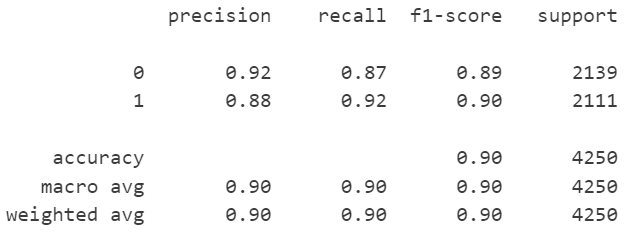
Step 2 – Under the smoking\_history column, transform 'not\_current' into 'former' and 'ever' into 'current'.

Step 3 – Resampling and hyperparameter selection

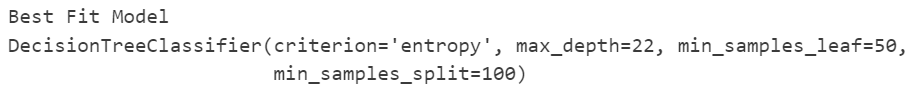
1. After generating 16,000 rows of balanced class data from undersampling using RandomUnderSampler, we start to build our decision tree. We pre-prune the tree using GridSearchCV to get a set of parameters that optimize the model.



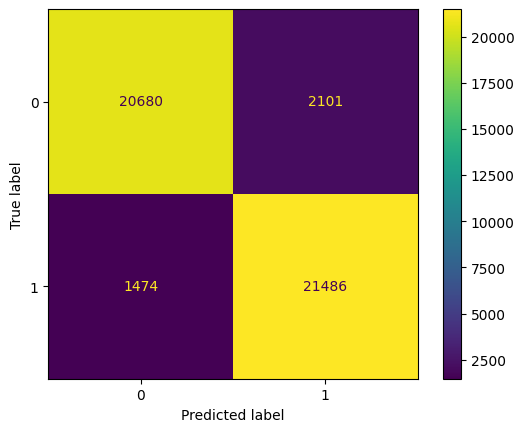
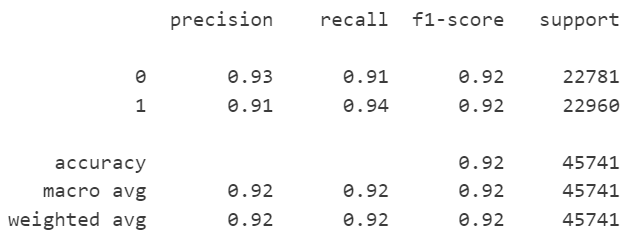
The test performance metrics summary indicates a slightly lower accuracy of 0.90. Nonetheless, the recall rate increases favorably to 92%, surpassing the performance of logistic regression.



1. Also, we constructed another decision tree using 160,000 rows of balanced class data generated through oversampling with RandomOverSampler. Employing GridSearchCV, we pre-pruned the tree to obtain an optimized set of parameters for the model.



The result is even better: all indices show improvement.



By far, the decision tree outperforms logistic regression well, with each rate exceeding 0.90. Below is a summary of all our decision trees' performance. From 0.74 to 0.94, we improve the recall by 27%:

| **Decision Tree** | Accuracy score | Precision | Recall | F1-score |
| --- | --- | --- | --- | --- |
| **Without pre-processing** | 0.95476 | 0.74 | 0.74 | 0.74 |
| **Undersampling**  **+pruning** | 0.89694 | 0.88 | 0.92 | 0.90 |
| **Oversampling**  **+pruning** | 0.92184 | 0.91 | 0.94 | 0.92 |

### 

### 

### Random Forest

Given the remarkable effectiveness of the decision tree, we are intrigued by the potential improvements in predictions with its ensemble learning counterpart - random forests.

Before data preparation, we obtain a very high accuracy score but low ability to classify positive instances.

## 

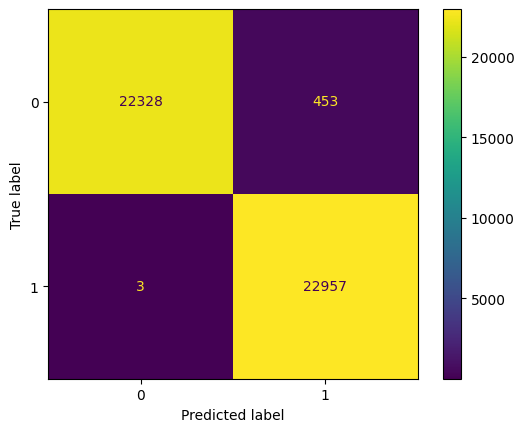
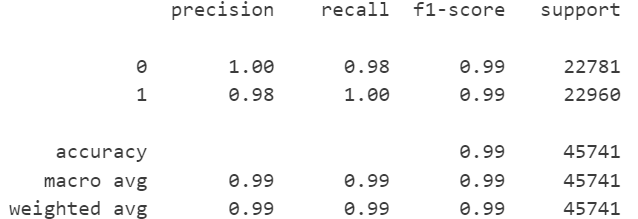
Again, we follow the same data preprocessing steps as what we have done for decision trees.

Step 1 – Remove rows under the gender column where the value is 'other'

Step 2 – Under the smoking\_history column, transform 'not\_current' into 'former' and 'ever' into 'current'.

Step 3 – Instead of conducting both resampling methods, we opt for oversampling exclusively as it produces a better result as suggested previously.

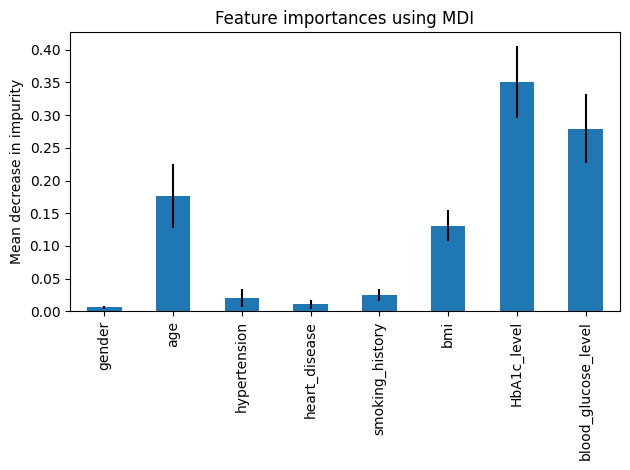
Next, we construct random forests with n\_estimators set to 100 and criterion set to 'entropy'. This results in exceptionally remarkable outcomes, with the prediction accuracy for actual diabetes patients nearing 100%, a level that couldn't be outdone!



Here is an overview of the performance of our random forests. With a boost from 0.68 to 1.00, we achieve a 47% improvement in recall, marking the highest enhancement among our three models.

| **Random Forest** | Accuracy score | Precision | Recall | F1-score |
| --- | --- | --- | --- | --- |
| **Without pre-processing** | 0.96964 | 0.95 | 0.68 | 0.79 |
| **Oversampling** | 0.99003 | 0.98 | 1.00 | 0.99 |

At this stage, we have ultimately identified that random forest is the optimal classifier for diabetes prediction. The key contributors to this promising outcome are the features: “HbA1c\_level”, “blood\_glucose”, and “age”. This inference aligns with the insights gleaned from the presented heatmap.



## Takeaways

### Results

In our analysis of diabetes prediction using various models, we observed distinct outcomes based on the data preprocessing techniques applied. For the logistic regression model, we gain a recall rate of 0.89 after pre-processing, beating the benchmark of 0.61. However, Logistic Regression is the weakest among all our models, so we concluded Logistic Regression is not suitable for our dataset.

With the decision tree model, the pre-processed models yield better performance than the un-processed model. The pre-processed undersampling data with a 0.92 recall rate and pre-processed oversampling data with a 0.94 recall rate both beat the benchmark set by the un-processed model with a 0.74 recall rate.

The random forest model, with its pre-processed oversampling approach, achieved the highest recall rate of 1.0, which beats the benchmark un-processed 0.68 recall rate. The performance indicates the superiority of the random forest model in our dataset analysis.

### What we learn

For our data mining analysis in a scenario with highly imbalanced data, we learned that:

1. In imbalanced datasets, methods like oversampling can significantly improve key metrics, such as the recall rate, ensuring a more accurate representation of minority classes.
2. The choice of models like decision trees and their advanced form, random forests, is crucial when dealing with diverse datasets. These models effectively handle mixed data types, leading to better accuracy in predictions.
3. Addressing data imbalances leads to more accurate and reliable results, essential in healthcare data analysis.

## Conclusion

For patients, early detection of diabetes is crucial. It allows for early intervention, which can significantly reduce the risk of complications such as heart disease, kidney failure, and vision loss. Early diagnosis enables patients to make necessary lifestyle changes, like improving diet and increasing physical activity, which can effectively manage or even reverse the condition in its early stages. This proactive approach greatly improves the overall quality of life and long-term health outcomes for individuals.

For healthcare institutions, early detection of diabetes translates into more efficient resource management and cost savings. By identifying and managing the condition early, healthcare providers can prevent the progression of the disease and avoid the high costs associated with treating advanced diabetes and its complications. This improves patient care in the healthcare system. Early intervention programs can lead to a significant reduction in emergency care and hospital admissions, making the healthcare system more sustainable and effective.

Our project was intended to further investigate any meaningful insights into diabetes risk factors via different key indicators like BMI, age, and blood glucose levels, which are crucial for early diabetes detection. By applying logistic regression, decision tree, and random forest models, this early identification can lead to timely interventions, significantly improving patient outcomes and reducing healthcare system burdens. Our analysis not only aids better disease management but also supports the optimization of healthcare resources. A future expansion of this work could be to use more data about health risk factors and potentially a person’s daily diet, which we believe will yield higher accuracy and more interesting results.

1. Julia, Nina. “Diabetes Statistics: Facts & Latest Data in the US (2023 Update).” *CFAH*, 9 Jan. 2023, cfah.org/diabetes-statistics/#How\_Many\_People\_Have\_Type\_2\_Diabetes. [↑](#footnote-ref-0)
2. https://idf.org/ [↑](#footnote-ref-1)
3. Mustafa, Mohammed. “Diabetes Prediction Dataset.” *Kaggle*, 8 Apr. 2023, www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset?fbclid=IwAR2pSti2A8sk8Er1VC1khdBEZVHgInbLPhxCDpE5DmLYgKpE4oksAOyXBVc. [↑](#footnote-ref-2)