

Diabetes prediction using Machine Learning

CIND820 CAPSTONE PROJECT



Table of Contents

Abstract.....	3
Literature Review	4
Data Description.....	6
Dataset #1	6
Exploratory Data Analysis Dataset #1	8
Dataset #2	19
Exploratory Data Analysis Dataset #2	20
GitHub Link for Code Files	23
Methodology.....	24
Hypothesis Testing	27
T-test:.....	27
Chi-Square Test:.....	27
Results & Discussion for Dataset #1	28
Results & Discussion for Dataset #2:.....	34
Conclusions, Limitation, Ethical Consideration	36
References	38

Abstract

In 2021, International Diabetes Federation (IDF) reported approximately 10.5% of the adult population from age 20 to 79 has diabetes. Almost half of patients are unaware of their diagnosed condition and are living without any awareness and cautionary measure. It is estimated by IDF that by 2045 there will be 1 in 8 adults living with diabetes. Diabetes is expected to more than double by 2050¹. Diabetes can cause long term damage to human function including the following but not limited to blindness, heart attacks kidney failure, and stroke according to World Health Organization article on Diabetes dated 14 November 2024². Diabetes data from the National Health and Nutrition Examination Survey revealed that during 2021- 2023, the total diabetes case was 15.8%, of which 4.5% were undiagnosed diabetes adults from United States³.

The primary objective of this project is to develop a system that accurately predicts an individual's likelihood of developing diabetes based on key health parameters. This involves building a machine learning model to enhance predictive accuracy. Some of the research questions include:

1. What are the **key factors** and correlations that increase the likelihood of developing diabetes?
2. What is the relationship between Body Mass Index (BMI) and the likelihood of a positive diabetes diagnosis? Does BMI above 30 increase the likelihood of diabetes by more than 50% compared to normal-weight individuals? **This question is related to BMI and the likelihood of diabetes.**
3. What is the mean blood glucose level for diabetic vs. non-diabetic individuals? Are blood glucose levels consistently higher in obese/overweight individuals compared to those with a normal BMI? What is the average BMI for individuals with and without hypertension? What is the percentage of individuals with HbA1c (HemoglobinA1c) levels above 6.5% have diabetes? **These questions confirm other clinical thresholds related to diabetic likelihood.**
4. Which machine learning model is most reliable in detecting diabetic individuals? **This question is related to machine learning models and evaluation.**
5. How well does a top-performing diabetes prediction model, trained on a large general dataset, perform when fine-tuned and applied to another dataset? What impact does this transfer have on the model's predictive performance and generalizability? **This question is related to transferability of machine learning models to other datasets.**

The aim of the project is to build predictive modeling techniques from machine learning to be applied to datasets to predict diabetes. The performance assessment of these models will be

evaluated and compared to determine the most effective approach. The proposed algorithms include logistic regression, k-nearest neighbors (KNN) classifier, decision tree classifier, random forest classifier, and support vector machine (SVM) classifier. Python, along with various libraries such as pandas, will be used for performing the modeling and analysis.

Literature Review

According to the World Health Organization (2024), the global prevalence of diabetes has increased four-fold over the past decades. With the global rise in diabetes cases, researchers and data scientists worldwide have explored various approaches to develop methods for early disease prediction. Several researchers used various machine learning (ML) algorithms to predict diabetes using different datasets over the time.

Several scholars used the machine learning (ML) method to predict diabetes using Pima Indian diabetes (PIDD) dataset. In the research study titled "Diabetes Prediction Using Machine Learning" (Rani, 2020). The author utilizes PIDD. The dataset comprised 2,000 instances, each with 8 features. The features included number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. To predicting diabetes, five different machine learning classification algorithms were used: K-Nearest Neighbour, Logistic Regression, Random Forest, Support Vector Machine, and Decision Tree. It was concluded that the Decision Tree algorithm achieved the highest performance by 98% accuracy on the training dataset and 99% on the test dataset.

However, the paper titled "Prediction of diabetes using classification algorithms" (Sisodia, D., & Sisodia, D. S. 2018), the authors applied three machine learning classification algorithms—Decision Tree, Support Vector Machine (SVM), and Naive Bayes—to the similar data source from PIDD to predict the likelihood of diabetes in patients. The dataset included 768 instances and 8 same features. The study evaluated different algorithms (Decision Tree, SVM, Naive Bayes) on this dataset. Their findings indicate that the Naive Bayes classifier outperformed the others, achieving the highest accuracy of 76.30%. These results were further validated using Receiver Operating Characteristic (ROC) curves.

While many studies have relied on PIDD, recent research has explored alternative datasets and advanced ML techniques to enhance predictive accuracy and generalizability. Predicting the Onset of Diabetes with Machine Learning Methods" by Chou et al. (2023)

examines the rising prevalence of diabetes in Taiwan and investigates the effectiveness of machine learning techniques in early disease prediction. The author used the data is from Taipei municipal medical center, analyzed records of 15,000 women aged 20 to 80, collected between 2018 and 2022. The researchers focused on eight key features like PIDD dataset. The following ML models are trained: logistic regression, neural network, decision jungle, and boosted decision tree. Among these, the boosted decision tree model demonstrated superior performance, achieving an area under the curve (AUC) of 0.991, indicating its high predictive accuracy for diabetes onset.

Similarly, "A Comparison of Machine Learning Algorithms for Diabetes Prediction" (Khanam & Foo, 2021) utilized a feature reduction method, retaining only five key features (Pregnancy, Glucose, BMI, Insulin, and Age) from PIDD. The study compared Logistic Regression and Support Vector Machine (SVM) and found that both models performed well for train/test split and K-fold cross-validation methods. Additionally, they developed a Neural Network (NN) model with varying hidden layers and epochs. They used NN models with 1, 2, 3 hidden layers varying the epochs 200, 400, 800. Their findings suggested that Neural Networks with two hidden layers achieved an accuracy of 88.6%, highlighting the potential of deep learning models in diabetes prediction.

Beyond traditional ML models, "DDPIS: Diabetes Disease Prediction by Improvising SVM" (Sharma et al.) introduced an enhanced SVM-based platform for diabetes prediction. The research utilized the UCI Machine Learning Repository's dataset with 16 attributes from both male and female patients. The author achieved 93.26% accuracy using an Improvised SVM model. This study demonstrates the effectiveness of model optimization techniques.

Building on the advancements in ML, deep learning techniques have also been explored. The study "Leveraging Pima Dataset to Diabetes Prediction: Case Study of Deep Neural Network" (Hounguè & Bigirimana) applied Deep Neural Networks (DNN) using the PIDD dataset, similar to Sisodia & Sisodia (2018). They employed 10-fold cross-validation and achieved an accuracy of 89%. Interestingly, their findings suggest that using 10-fold cross-validation may decrease the efficiency of DNN models in diabetes prediction. The study points out the potential of deep learning approaches in improving diabetes risk assessment models.

Based on the literature reviews for this project, the application of machine learning and deep learning in diabetes prediction has evolved significantly, transitioning from traditional classification models using structured datasets to more advanced techniques incorporating feature selection, ensemble learning, and deep neural networks. These advancements have led

to higher predictive accuracy, improved generalizability, and enhanced early detection capabilities of diabetes diagnosis.

Data Description

For this project, ML model will be applied to the two datasets.

Dataset #1

The dataset #1 is a collection of health indicators and demographics and behaviors data from patients, along with binary classification diabetes status (No Diabetes, Diabetes). There are 100,000 rows in this dataset. The health indicators includes bmi, HbA1c level, blood glucose level and demographics and behaviors included gender, age, smoking history. The source of the dataset is Electronic Health Records (EHRs) and download from this link: <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset/data>.

Types of Data Collected:

- Demographic Information: Age, gender, smoking history.
- Clinical Factors: BMI, hypertension, heart disease.
- Laboratory Test Results: HbA1c levels, blood glucose levels.
- Survey-Based Data: Lifestyle and risk factor assessments.

This dataset's specific location or continent and the date of collection is not disclosed due to the confidentiality and privacy. EHRs were collected from multiple healthcare providers and compiled into a single dataset. Therefore, data may not represent the general population as it is collected from specific healthcare settings. The dataset may not include diverse populations from various geographic or socioeconomic backgrounds.

The dataset has total eight features or independent variables and one target feature/dependent variable. The type of data feature is described in Table-1.

Dataset #1	Variable types	Brief description
Gender	Categorical	Categorized as male, female, or other
Age	Numerical	Ranges from 0-80; diabetes is more prevalent in older adults.
Hypertension	Numerical	Binary (0 = no, 1 = yes); high blood pressure increases diabetes risk.
Heart disease	Numerical	Binary (0 = no, 1 = yes); associated with higher diabetes risk.
Smoking History	Categorical	Classified as not current, former, No Info, current, never, or ever; smoking elevates diabetes risk.
BMI (Body Mass Index)	Numerical	Ranges from 10.16 to 71.55; higher BMI correlates with greater diabetes risk. Categories: underweight (<18.5), normal (18.5-24.9), overweight (25-29.9), obese (≥ 30).
HbA1c (Hemoglobin A1c)	Numerical	Measures average blood sugar over 2-3 months; levels >6.5% indicate diabetes.
Blood glucose	Numerical	High levels are a primary diabetes indicator.
Diabetes (Target Variable)	Numerical	Binary (0 = no diabetes, 1 = diabetes).

Table-1: Feature type (Categorical or Numerical)

Exploratory Data Analysis Dataset #1

The explanatory data analysis section is divided into three parts.

1. An initial analysis is performed through descriptive statistics of the features of the dataset.
2. Univariate analysis is performed for each of the independent variables.
3. Bivariate analysis was performed in pair on some of the important variables.

A. Initial Analysis:

Dataset 1 comprises 100,000 observations across 9 variables, with no missing cells (0.0%). There are no missing values were found, eliminating bias concerns. However, it contains 3,085 duplicate rows, accounting for 3.85% of the data. Target Column (diabetes) is binary (0 or 1).

```
Dataset Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
#   Column                Non-Null Count  Dtype
---  -
0   gender                100000 non-null  object
1   age                   100000 non-null  float64
2   hypertension          100000 non-null  int64
3   heart_disease         100000 non-null  int64
4   smoking_history       100000 non-null  object
5   bmi                   100000 non-null  float64
6   HbA1c_level           100000 non-null  float64
7   blood_glucose_level   100000 non-null  int64
8   diabetes              100000 non-null  int64
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB

Duplicate Rows Count: 3854

Total Missing Values in Dataset: 0
```

Table-2: Dataset 1 info.

The key statistics (mean, mode, standard deviation) are summarized in Table-2. The target variable, diabetes, is highly imbalanced, with only 8.5% of individuals diagnosed. Values range between 0 and 1, confirming binary classification. 25th, 50th, and 75th percentiles: All are 0, highlighting that at least 75% of the observations are non-diabetic.

Summary Statistics:

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000	100000.000000
mean	41.885856	0.07485	0.039420	27.320767	5.527507	138.058060	0.085000
std	22.516840	0.26315	0.194593	6.636783	1.070672	40.708136	0.278883
min	0.080000	0.00000	0.000000	10.010000	3.500000	80.000000	0.000000
25%	24.000000	0.00000	0.000000	23.630000	4.800000	100.000000	0.000000
50%	43.000000	0.00000	0.000000	27.320000	5.800000	140.000000	0.000000
75%	60.000000	0.00000	0.000000	29.580000	6.200000	159.000000	0.000000
max	80.000000	1.00000	1.000000	95.690000	9.000000	300.000000	1.000000

Table-3: Statistical measures for numerical features.

B. Univariate Analysis

1. Gender

The dataset contains 58,552 (58.6%) female and 41,430 (41.4%) male instances, with only 18 (<0.1%) entries classified as "Other." Given the negligible proportion of the "Other" category, it can be removed without significantly impacting the final analysis. *The dataset has 58.56% Female individuals, which forms the majority.*

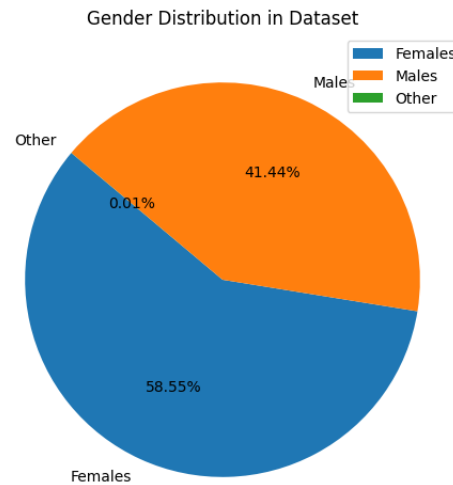


Fig-1: Ratio of male, female and others in the dataset.

2. Age

The age distribution in the dataset ranges from newborn to 80 years. The data appears to be fairly spread across different age groups. The mean age is 41.88 years for this dataset. However, there is an increase in frequency at age 80, suggesting a higher representation of elderly individuals in this dataset.

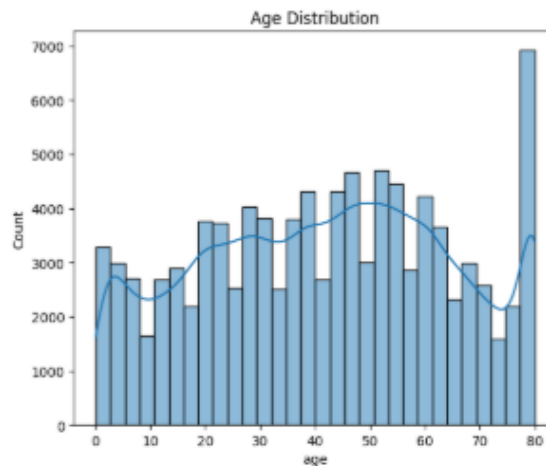


Fig.2: Distribution of "age" in the dataset

3. Hypertension

This dataset shows a high imbalance in the hypertension feature, with 7,485 (7.49%) of patients having hypertension and 92,515 (92.51%) without it. *Hypertension is relatively rare in the dataset, with only about 7.03% of individuals affected.*

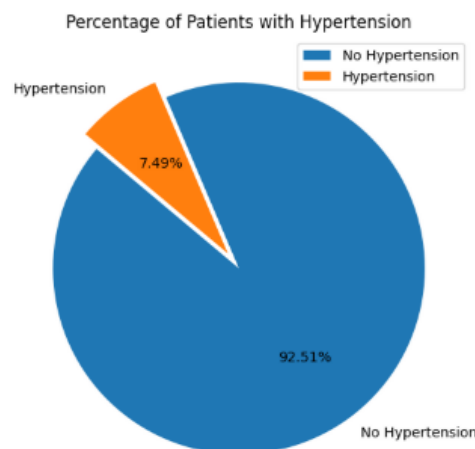


Fig.3: Ratio of "Hypertension" in the dataset

4. Heart Disease

The heart disease feature is highly imbalanced, with only 3,942 (3.94%) patients having heart disease, while 96,058 (96.06%) do not. *This indicates a highly imbalanced dataset, which can negatively impact machine learning models. Accuracy alone is not a good metric Possible Solutions are SMOTE, under sampling, or class weighting technique.*

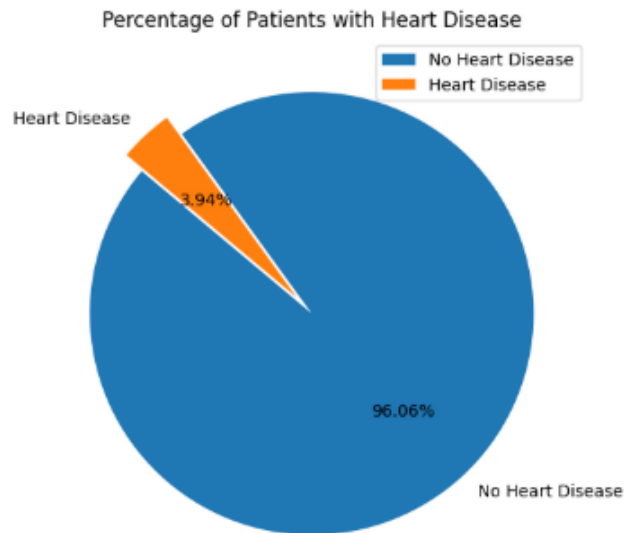


Fig.4: Ratio of "heath disease" in the dataset

5. Smoking History

35.81% of instances (35,816) have no information on smoking history, which may significantly impact results. Current smokers account for 9.28%, while former smokers represent 9.35%.

	Smoking History	Count	Frequency (%)
0	No Info	35810	35.816447
1	never	35092	35.098318
2	former	9352	9.353684
3	current	9286	9.287672
4	not current	6439	6.440159
5	ever	4003	4.003721

Fig 5A: Smoking History Distribution Table

After grouping, grouping smoking categories into broader groups (former, current, Not Current, and ever into one "Smoked" category). This is further reviewed in the model processing steps.

	Smoking Group	Count	Frequency (%)
0	Unknown	35810	35.816447
1	Never Smoked	35092	35.098318
2	Smoked	29080	29.085235

Fig 5B: Grouping Smoking History Distribution Table

6. Body Mass Index (BMI)

According to the Centers for Disease Control and Prevention, these are the BMI Category

- ≤ 18.5 Underweight
- 18.5 – 24.9 Normal
- 25 – 29.9 Overweight
- Greater than 30 Obesity

The BMI distribution shows a right-skewed pattern, indicating that a majority of patients have BMI values in the lower range.

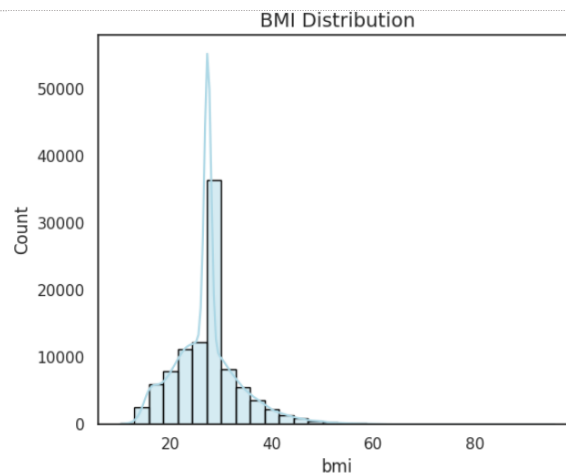


Fig.6: Distribution of BMI in the dataset

The presence of high BMI values suggests outliers, which may require treatment before further modeling. After removing outliers, the dataset size was reduced by 7,085 rows (approximately 7.1%).

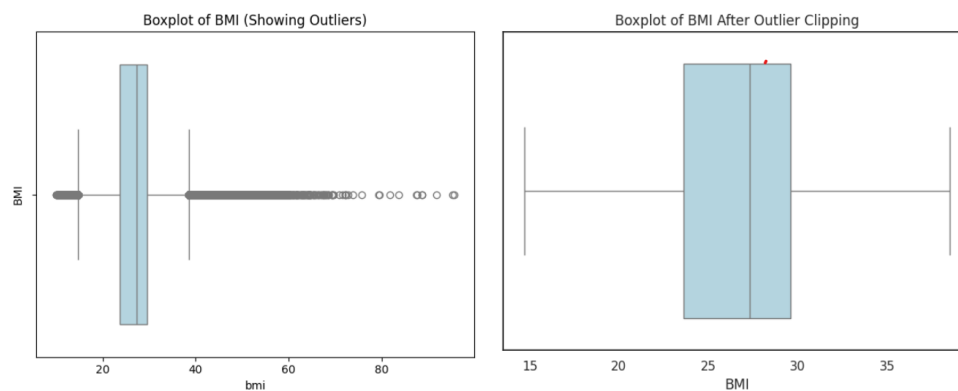


Fig.7: Boxplot for BMI showing outliers and without outlier.

7. HbA1c Level

A hemoglobin A1C (HbA1C) test is a blood test that shows what your average blood sugar (glucose) level was over the past two to three months. we will create a new feature based on the value of (HbA1C). The following ranges are used to diagnose prediabetes and diabetes according to the Centers for Disease Control and Prevention. (2022, September 30):

- Normal: below 5.7%
- Prediabetes: 5.7% to 6.4%

- Diabetes: 6.5% or above

The HbA1c levels in this dataset range from 3.5% to 9%.

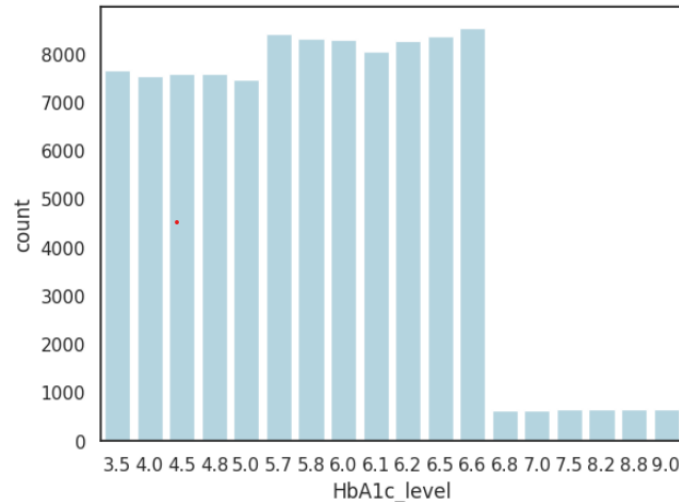


Fig 8: Distribution of HbA1c level in the dataset.

These percentage ranges refer to the Hemoglobin A1c (HbA1c) levels, a widely used biomarker for diagnosing diabetes and prediabetes.

HbA1c levels range from 3.5% to 9.0%.

	HbA1c Category	Count	Percentage (%)
0	≤ 5.7% (Non-Diabetic)	46262	46.270329
1	5.7% - 6.4% (Prediabetic)	32926	32.931928
2	≥ 6.5% (Diabetic)	20794	20.797744

Fig 9: HbA1c Level Table

8. Blood Glucose Level

The blood glucose levels range from 80 mg/dL to 300 mg/dL. These thresholds align with the American Diabetes Association (ADA) guidelines for diabetes screening and diagnosis.

Glucose levels can be categorized as follows:

- ≤ 99 mg/dL: Normal
- 100 – 125 mg/dL: Prediabetic
- ≥ 126 mg/dL: Diabetic

Resources : <https://www.cdc.gov/diabetes/basics/getting-tested.html>

	Blood Glucose Level	Prediction
0	≤ 99	0.00% have diabetes
1	100 – 125	0.00% have diabetes
2	≥ 126	11.83% have diabetes

Fig.10: Distribution blood glucose level of persons in the dataset.

9. Diabetes (Target Feature)

Diabetes is the dependent variable (target feature) in this dataset. 8.5% of the dataset has diabetes. The dataset exhibits an imbalance in diabetes cases. The dataset has an underrepresentation of diabetes cases compared to global estimates but remains within an acceptable range considering unknown geographical factors.

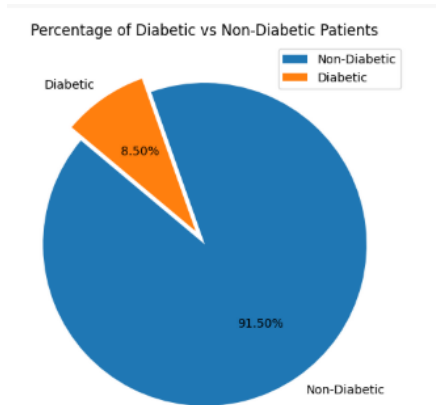


Fig.11: Distribution of Diabetes (Target Feature) in the dataset.

C. Bivariate Analysis

Bivariate analysis examines the relationship between two variables in the dataset.

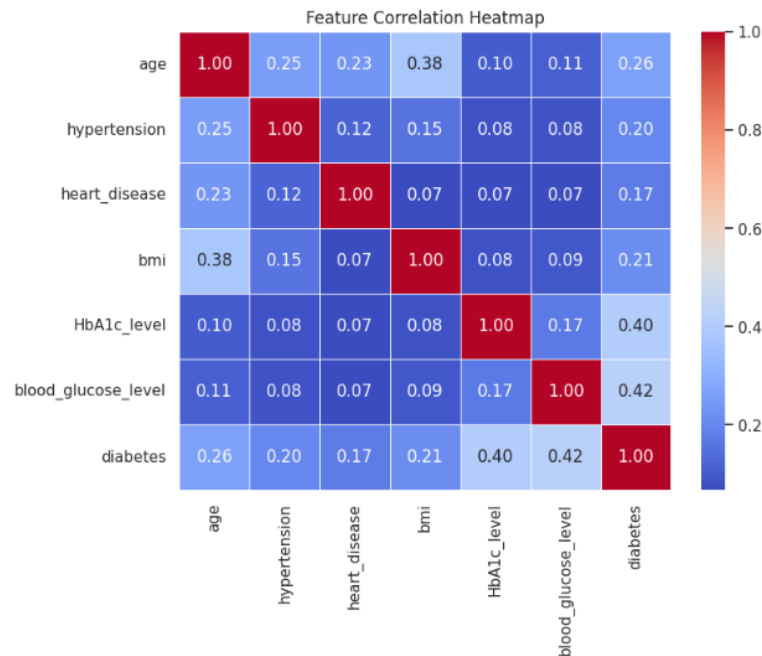


Fig.12: Feature Correlation Heatmap

The following findings are observed within the heatmap:

- Diabetes is positively correlated with both HbA1c level and blood glucose level, confirming their importance in predicting diabetes.
- BMI has a moderate correlation with age, suggesting that older individuals may have a slightly higher BMI.
- There is a moderate correlation between age and diabetes, suggesting that older individuals have a higher likelihood of developing diabetes.
- Hypertension and heart disease exhibit some correlation, which aligns with medical findings that hypertension can increase the risk of heart disease.
- Smoking history shows minimal correlation with most variables, indicating that its impact on diabetes may be limited.

HbA1c_level vs. diabetes:

Diabetes is positively correlated with HbA1c level, confirming their importance in predicting diabetes. What percentage of individuals with HbA1c levels above 6.5% have diabetes?

Percentage of individuals with HbA1c > 6.5% who have diabetes: 36.82%

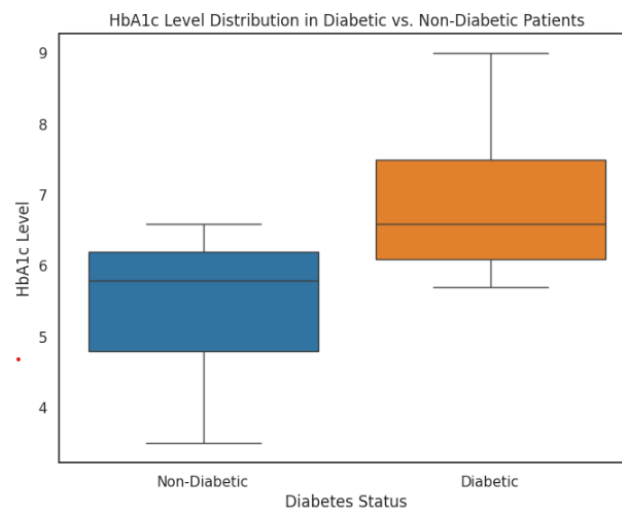


Fig.13: HbA1C level for target variable.

BMI vs diabetes:

According to the ordinal category [underweight, normal, overweight, obesity], as the weight category increases, the percentage of patients with diabetes increases.

BMI Category	Prediction
0 Underweight	0.75% have diabetes
1 Normal	3.84% have diabetes
2 Overweight	7.25% have diabetes
3 Obesity	17.92% have diabetes

Fig.14: Different ordinal category and diabetes

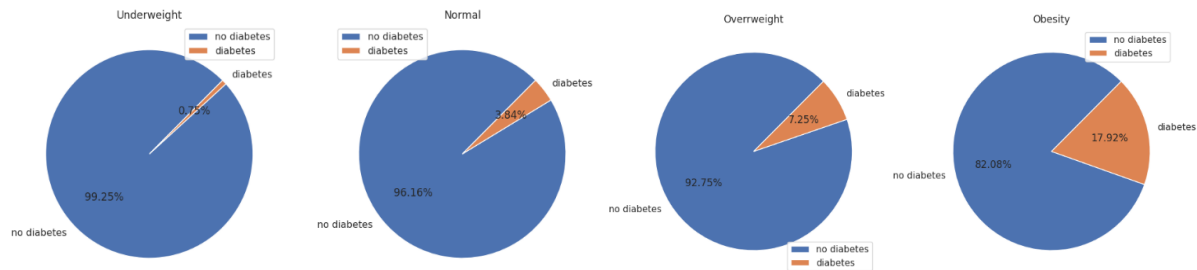


Fig.15: Pie charts for different ordinal category and diabetes

From data, Diabetes rate in obese individuals: 18.03%. Diabetes rate in non-obese individuals: 5.58%

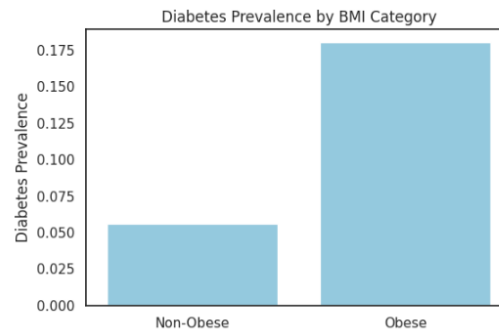


Fig.16: bar chart compares diabetes between Non-Obese and Obese individuals

Blood Glucose Level vs diabetes:

What is the mean blood glucose level for diabetic vs. non-diabetic individuals? Diabetes is positively correlated with blood glucose level, confirming their importance in predicting diabetes. According to the American Diabetes Association (2024), This chart categorizes blood glucose levels based on fasting blood glucose values. It is a clinical reference used to diagnose normal glucose levels, prediabetes, and diabetes. Blood Glucose Level (mg/dL) category are as follow:

- ≤ 99 Normal
- 100 – 125 Prediabetes
- ≥ 126 Diabetes

This confirms in the bar plot shows the blood sugar level of 220 mg/dl and above are diagnosed with diabetes.

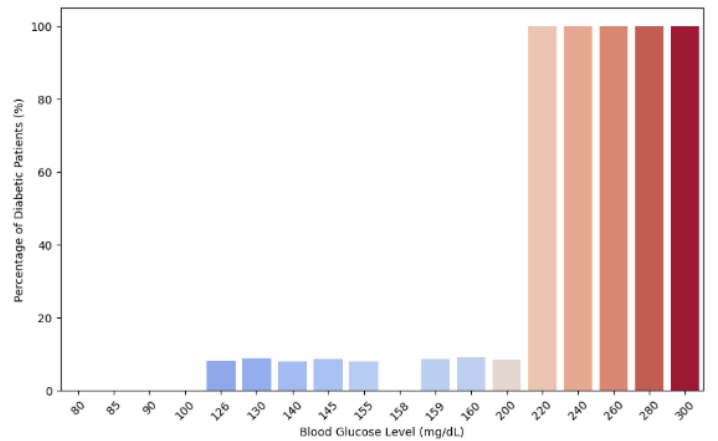


Fig.17: Bar plot for Glucose Level vs percentage of persons with diabetes

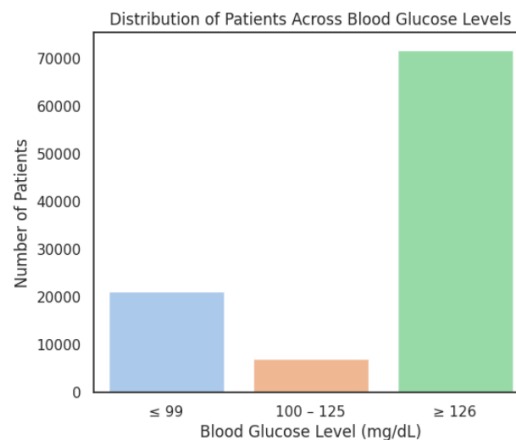


Fig.18: Distribution of Patients Across Blood Glucose Levels

Hypertension vs bmi

What is the average BMI for individuals with and without hypertension?

- *Mean BMI for individuals with Hypertension: 30.01*
- *Mean BMI for individuals without Hypertension: 26.75*
- *Median BMI for individuals with Hypertension: 28.70*
- *Median BMI for individuals without Hypertension: 27.32*

Age vs Heart Disease

- *Age & Heart Disease: 0.266 (Moderate correlation)*
- *Median Age for individuals with heart disease: 70.00*
- *Median Age for individuals without heart disease: 41.00*

Dataset #2

The dataset #2 contains 253,680 entries and 22 columns, all of which are numeric (float type) using Y Profiling. The target variable is Diabetes (0 = No diabetes, 1 = Pre-diabetes, 2 = Diabetes). A large proportion (84.24%) of the dataset consists of non-diabetic individuals. Only 13.93% have diabetes, and 1.83% are classified as pre-diabetic. There is a significant class imbalance, with a small proportion of pre-diabetic and diabetic cases compared to non-diabetic cases. There are no missing values.

The dataset (#2) is a collection of health indicators and demographics data from patients, along with three categories Diabetes status (0 = No, 1 = Prediabetes, 2 = Diabetes). There are 253,680 rows in this dataset. The health indicators includes highbp, highchol, bmi, smoker, stroke, etc. and demographics and behaviors included physical activity, fruits, veggies, sex, age, etc. The source of the dataset is 2015 Behavioral Risk Factor Surveillance System (BRFSS) and download from this link: <https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset>

Types of Data Collected:

Demographic Information: Age, gender, smoking history.

Clinical Factors: BMI, hypertension, heart disease.

Laboratory Test Results: None

Survey-Based Data: Physical Activity, Physical Health, General Health Status etc.

```
Data columns (total 22 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   Diabetes_012                          253680 non-null float64
1   HighBP                                253680 non-null float64
2   HighChol                             253680 non-null float64
3   CholCheck                            253680 non-null float64
4   BMI                                  253680 non-null float64
5   Smoker                               253680 non-null float64
6   Stroke                               253680 non-null float64
7   HeartDiseaseorAttack                 253680 non-null float64
8   PhysActivity                         253680 non-null float64
9   Fruits                              253680 non-null float64
10  Veggies                              253680 non-null float64
11  HvyAlcoholConsump                   253680 non-null float64
12  AnyHealthcare                       253680 non-null float64
13  NoDocbcCost                         253680 non-null float64
14  GenHlth                             253680 non-null float64
15  MentHlth                             253680 non-null float64
16  PhysHlth                             253680 non-null float64
17  DiffWalk                             253680 non-null float64
18  Sex                                  253680 non-null float64
19  Age                                  253680 non-null float64
20  Education                           253680 non-null float64
21  Income                              253680 non-null float64
dtypes: float64(22)
memory usage: 42.6 MB
```

	count
Diabetes_012	
0.0	213703
2.0	35346
1.0	4631

```
dtype: int64
```

Fig.19: Dataset #2 info

Exploratory Data Analysis Dataset #2

The bar chart shows a highly imbalanced distribution of diabetes status, with the majority of individuals classified as non-diabetic (0), and substantially fewer labeled as prediabetic (2) or diabetic (1).

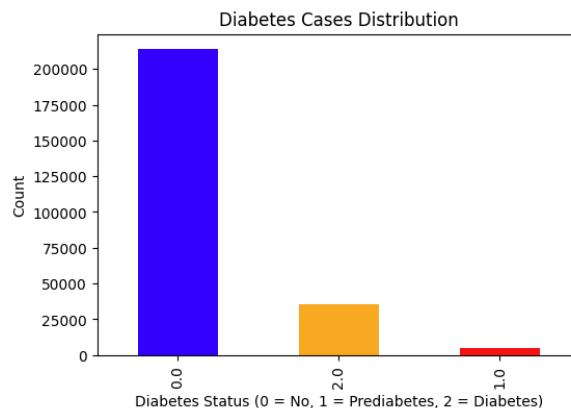


Fig.20: Bar plot for Distribution of Diabetes Status in the Dataset 2

The results below indicate that after removing duplicates and converted into binary target variable. Dataset 2 contains 229,712 records. Feature importance analysis reveals BMI as the most influential predictor for diabetes, followed by General Health, Age, High Blood Pressure, and High Cholesterol

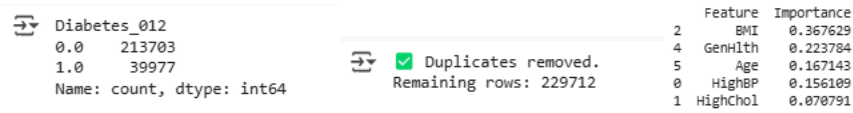


Fig.21: Dataset #2 Binary target and feature importance

In the statistics summary, the median (50th percentile) is 0, confirming that over 50% of the records represent non-diabetic individuals. Mean of 0.139 indicates most individuals are non-diabetic. The target variable shows class imbalance

Summary Statistics:

	Diabetes_012	HighBP	HighChol	BMI	Smoker	GenHlth	Age
count	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000	253680.000000
mean	0.139333	0.429001	0.424121	28.382364	0.443169	2.511392	8.032119
std	0.346294	0.494934	0.494210	6.608694	0.496761	1.068477	3.054220
min	0.000000	0.000000	0.000000	12.000000	0.000000	1.000000	1.000000
25%	0.000000	0.000000	0.000000	24.000000	0.000000	2.000000	6.000000
50%	0.000000	0.000000	0.000000	27.000000	0.000000	2.000000	8.000000
75%	0.000000	1.000000	1.000000	31.000000	1.000000	3.000000	10.000000
max	1.000000	1.000000	1.000000	98.000000	1.000000	5.000000	13.000000

Table-4: Statistical measures for numerical features.

Dataset 2 consists of several health metrics collected from individuals, explicitly including BMI, blood pressure, cholesterol levels, and diabetes status (Diabetes_012). The correlation heatmap for the dataset 2 shows high blood pressure, cholesterol, BMI, and physical activity might show meaningful relationships with diabetes.

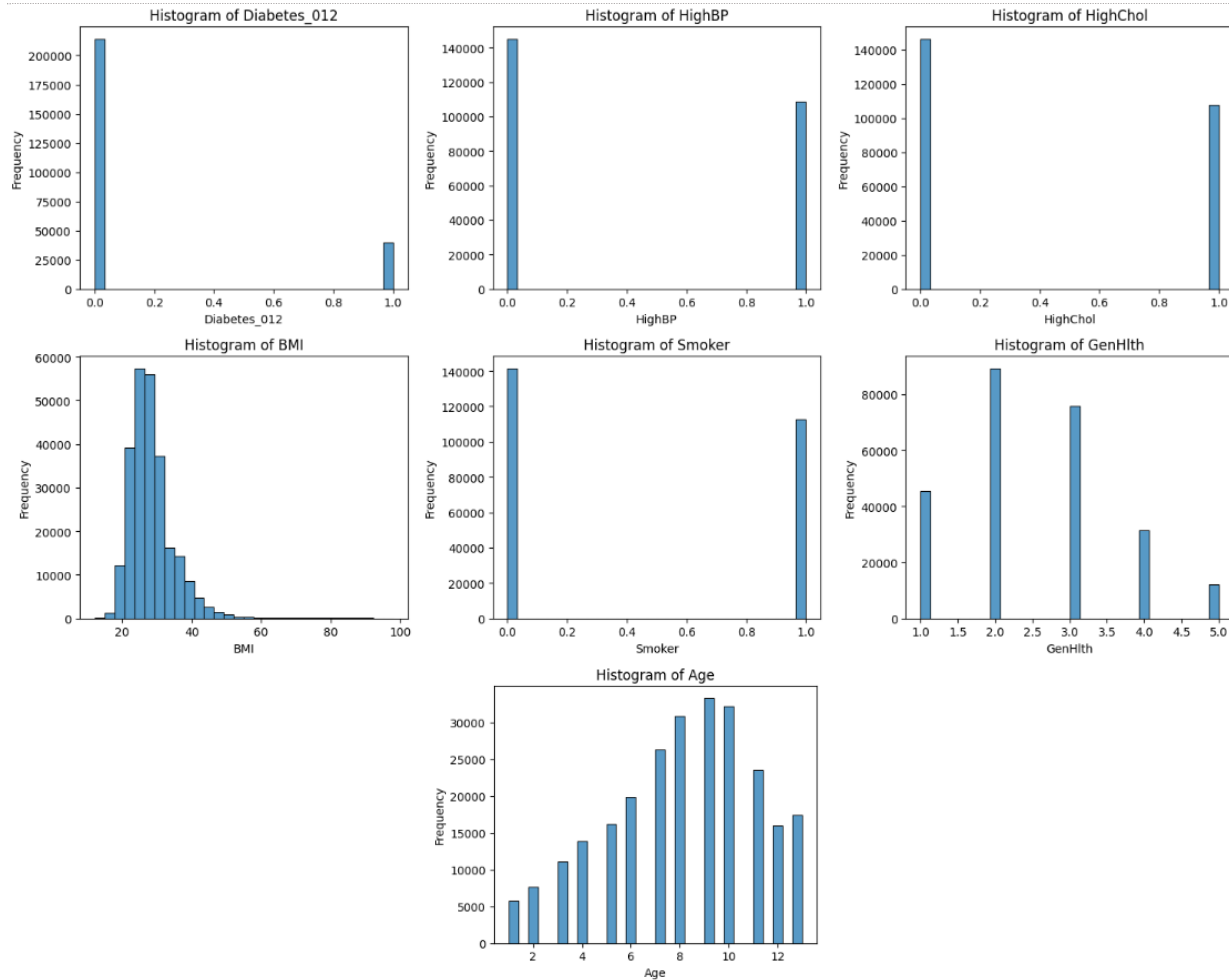


Fig.22: Distribution of variables in this dataset

The histograms reveal clear class imbalance in the target variable Diabetes_012, with a majority of records labeled as non-diabetic. The correlation matrix further shows that high blood pressure (0.27), general health (0.30), and BMI (0.22) have the strongest positive correlations with diabetes, reinforcing their importance as predictive features in the model.

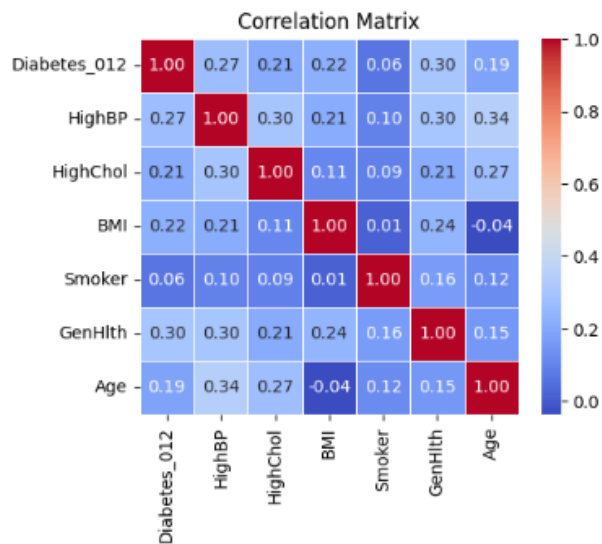


Fig.23: Feature Correlation Heatmap

Calculate percentage of individuals with BMI > 30 who have diabetes (Diabetes_012 == 2)
 Percentage of individuals with BMI > 30 who have diabetes: 27.68%

#High Blood Pressure (HighBP) vs Diabetes
 Percentage of individuals with HighBP who have diabetes: 27.12%

#High Cholesterol (HighChol) vs Diabetes
 Percentage of individuals with HighChol who have diabetes: 24.69%

Among individuals with BMI greater than 30, 27.68% were diagnosed with diabetes, highlighting obesity as a major contributing factor. Similarly, 27.12% of those with high blood pressure and 24.69% of individuals with high cholesterol were diabetic, reinforcing the strong association between these health conditions and diabetes risk. This dataset is valid for training and testing the tuned model/classifier the effectiveness of diabetic prediction.

GitHub Link for Code Files

A link to a repository on GitHub website where codes and results are uploaded:

<https://github.com/AsmaShaikhTMU/Projects>

Methodology

In this project, two datasets of varying sizes are utilized. See Fig.24: Process flowchart of Project Methodology below. Initially, a model is trained on the larger dataset (#1) to capture general patterns and features.

The dataset (#1) undergoes preprocessing to handle missing values and ensure data quality. This step includes data cleaning and imputation techniques to address inconsistencies. Then further exploratory analysis was performed on the dataset. This step confirms that the data is consistent and ready for the next stage. Next, feature selection may be applied to identify the most relevant attributes, improving model accuracy as applicable.

Following this, the machine learning workflow is built and evaluated a diabetes prediction model using multiple classification algorithms. The dataset is trained using train-test split (using sklearn's `train_test_split` function with `random_state=42`). A `ColumnTransformer` was used to apply `StandardScaler` to numerical features and `OneHotEncoder` to categorical features. This was integrated into a Pipeline for each model.

The following five classification models were evaluated using 5-fold cross-validation:

1. Logistic Regression is a simple baseline model for binary classification tasks. This can assist in understanding how features affect the prediction.
2. K-Nearest Neighbors (KNN) is a non-parametric/linear method that makes no assumptions about the underlying data distribution. It makes predictions by looking at similar patients.
3. Decision Tree handles both numerical and categorical data well and shows how decisions are made step-by-step using feature splits.
4. Random Forest is an ensemble of decision trees, reducing overfitting and improving generalization. Hence many trees will improve accuracy.
5. Support Vector Machine (SVM) is high-dimensional spaces and works well with clear margins of separation. This is powerful for finding the boundary between diabetic and non-diabetic cases

After the first iteration of each model, outcomes are evaluated. The following evaluation metrics of the five classifiers/models to evaluate the performance of diabetes prediction:

Accuracy: Measures the overall correctness of the model.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: Measures of how many of the predicted positives are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Measures how many actual positives the model correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: Harmonic mean of Precision and Recall; balances both.

$$F_1 = 2 \times \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

F2 Score: Gives more weight to Recall than Precision, which is useful in medical diagnosis where false negatives are more dangerous.

$$F_2 = 5 \times \frac{\text{Precision} \cdot \text{Recall}}{4 \times \text{Precision} + \text{Recall}}$$

Mean CV Score: Evaluates model consistency on different training/testing subsets.

$$\text{Mean CV Score} = \text{Average accuracy across } k - \text{folds}$$

Test Score: Reflects how well the model performs on unseen data.

$$\text{Test Score} = \text{Accuracy on the held - out test dataset}$$

These metrics provide insights into the effectiveness and reliability of each model in predicting diabetes. Best classifier is then further tuned by The Feature importances, GridSearchCV Tuning and SMOTE. In the final stage, comparative analysis is performed and analyzed the results to identify the best model. The pre-trained classifier model (from dataset #1) is then applied to another dataset (#2), to the target variable—diabetes prediction to test its effectiveness.

Final Process Flowchart

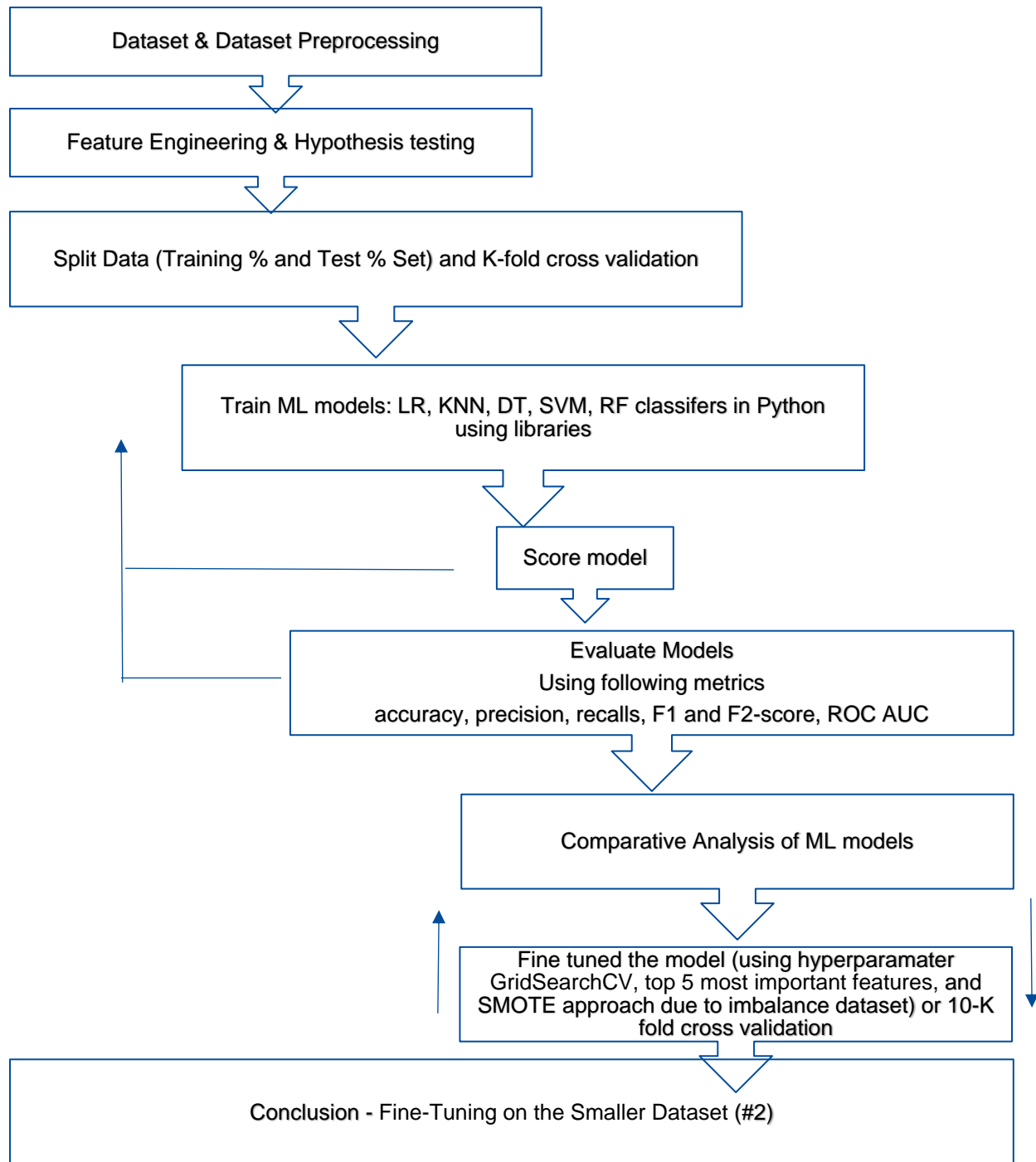


Fig.24: Process flowchart of Project Methodology.

Hypothesis Testing

Hypothesis testing was performed to validate whether the observed relationships between key features (such as glucose, BMI, and age) and the target variable were statistically significant, ensuring that only meaningful patterns were used to guide model development

T-test:

The t-test is a statistical test used to determine whether there is a significant difference between the means of two groups:

Target (diabetic vs. non-diabetic patients).g

Continuous variable (e.g., blood glucose levels, BMI, HbA1c levels).

Question: Compare blood glucose levels between diabetic and non-diabetic individuals.

```
T-Test Results (Blood Glucose vs Diabetes):  
T-Statistic: 94.7966, P-Value: 0.0000  
Significant Difference
```

Since the p-value is much lower than 0.05, we reject the null hypothesis. There is a significant difference in blood glucose levels between diabetic and non-diabetic individuals. Blood glucose is a strong differentiator for diabetes.

Chi-Square Test:

The Chi-Square Test is a statistical test used to determine whether there is a significant association between two categorical variables

Question: if hypertension and diabetes are significantly associated.

```
Chi-Square Test Results (Hypertension vs Diabetes):  
Chi-Square Statistic: 3909.5098, P-Value: 0.0000  
Significant Association
```

There is strong statistical association between hypertension and diabetes. This means that individuals with hypertension are significantly more likely to have diabetes.

Results & Discussion for Dataset #1

The model training and evaluation were conducted in Python using libraries (pandas, scikit-learn, and numpy). This section of the report presents a comparative analysis of five classifiers. The classifier evaluation table below provides a side-by-side comparison of 5 machine learning algorithms.

Results Table:

Model	Mean CV Score	Test Score	Accuracy	Precision	Recall	F1 Score	F2 Score
Logistic Regression	0.960805	0.959494	0.959494	0.879026	0.630423	0.734252	0.668219
KNN	0.960318	0.959444	0.959444	0.901667	0.609577	0.727395	0.651807
Decision Tree	0.951941	0.951843	0.951843	0.722101	0.743662	0.732723	0.739247
Random Forest	0.969732	0.968695	0.968695	0.942263	0.689577	0.796357	0.728658
SVM	0.963631	0.961794	0.961794	0.984660	0.578592	0.728886	0.630603

Table-5: Evaluation metrics of the five baseline models.

Random Forest exhibited the best overall performance across multiple evaluation metrics:

- Highest Mean CV Score: 0.9697
- Highest Accuracy: 0.9687
- Best F1 Score: 0.7964

As shown in Figure 23, the Random Forest classifier demonstrated strong and consistent results across all metrics, including Accuracy, Precision, Recall, F1, and F2 scores. Based on these findings and repeated model iterations, Random Forest is recommended for balanced and reliable predictions.

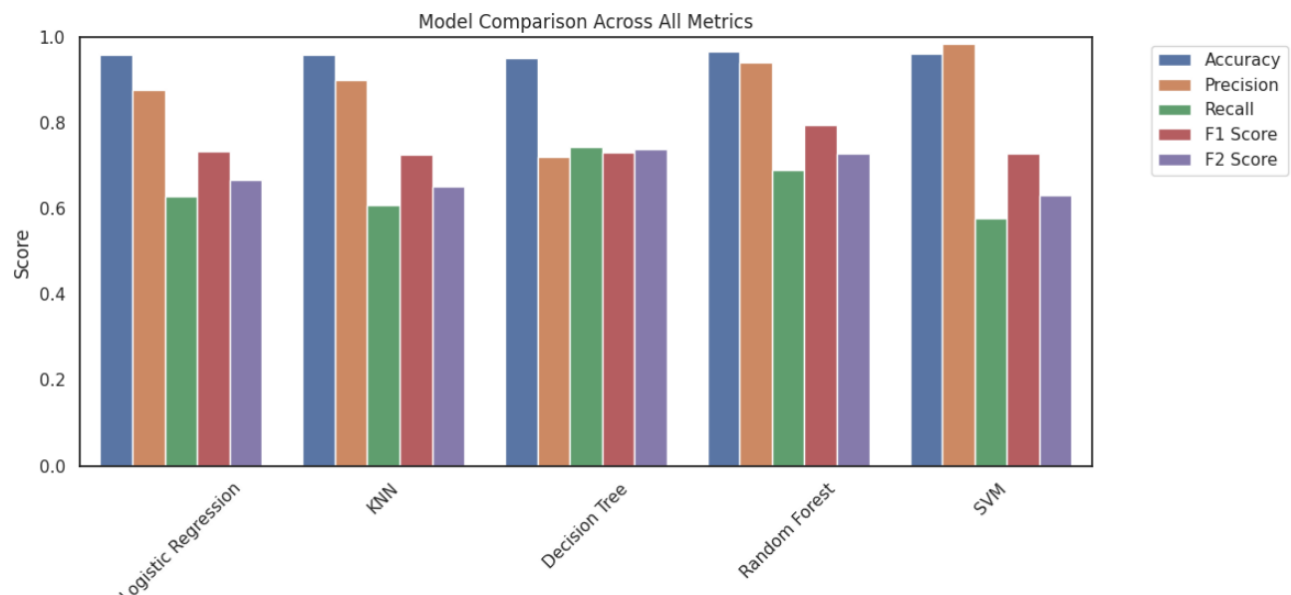


Fig.25: Grouped bar chart comparing the performance of five baseline machine learning models

Random Forest has the highest F1 Score, indicating it maintains a strong balance between Precision and Recall in the bar chart below. The other models have relatively similar F1 Scores, but all slightly lower than Random Forest. In terms of the F2 Score, Random Forest also performed competitively, closely trailing the Decision Tree classifier.

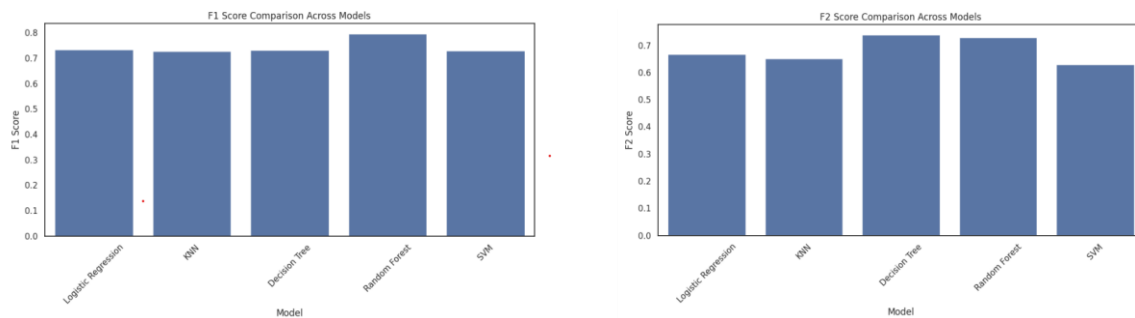


Fig.26: Bar chart comparing F1 & F2 scores of five machine learning models.

The confusion matrix below illustrates the number of true positives, false positives, true negatives, and false negatives on the test set, providing insights into the model's misclassification patterns.

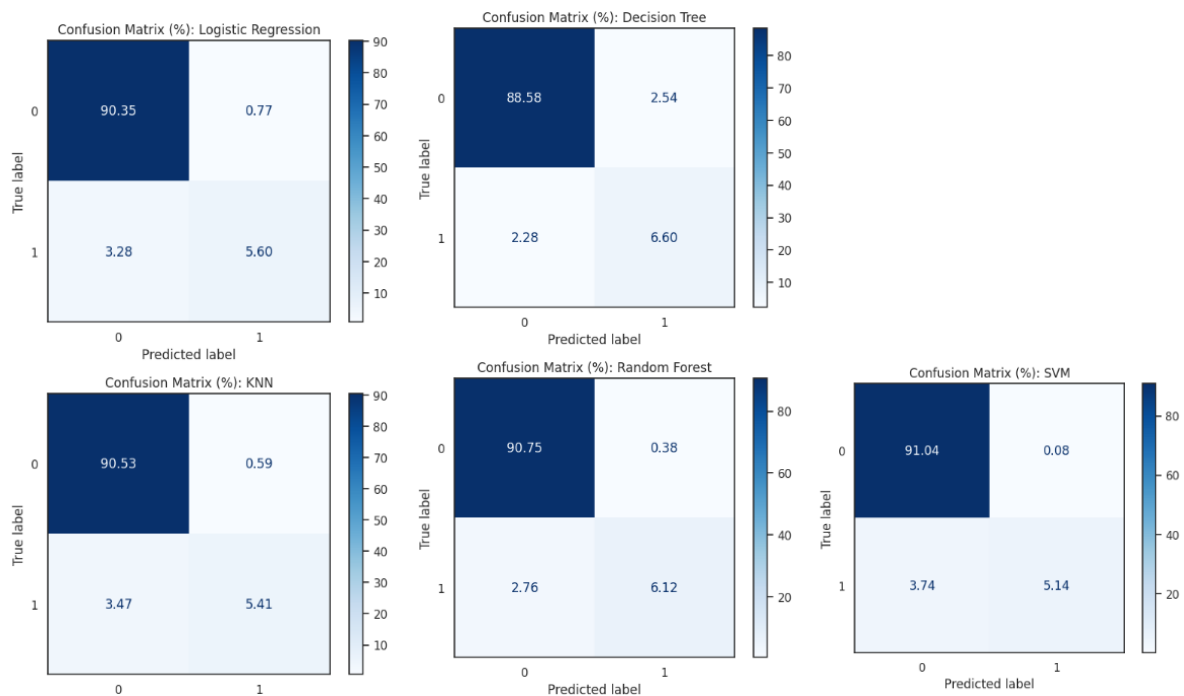


Fig.27: Confusion Matrices with 5-fold CV.

The ROC (Receiver Operating Characteristic) curve plots the True Positive Rate (Sensitivity) against the False Positive Rate, while the AUC (Area Under the Curve) quantifies the model's overall ability to distinguish between classes. An AUC value closer to 1.0 indicates superior model performance.

As shown in the figure below, Random Forest and Logistic Regression both achieved the highest AUC of 0.96 confirming their strong and consistent predictive performance across various classification thresholds. These results highlight Random Forest and Logistic Regression as the top-performing models based on AUC.

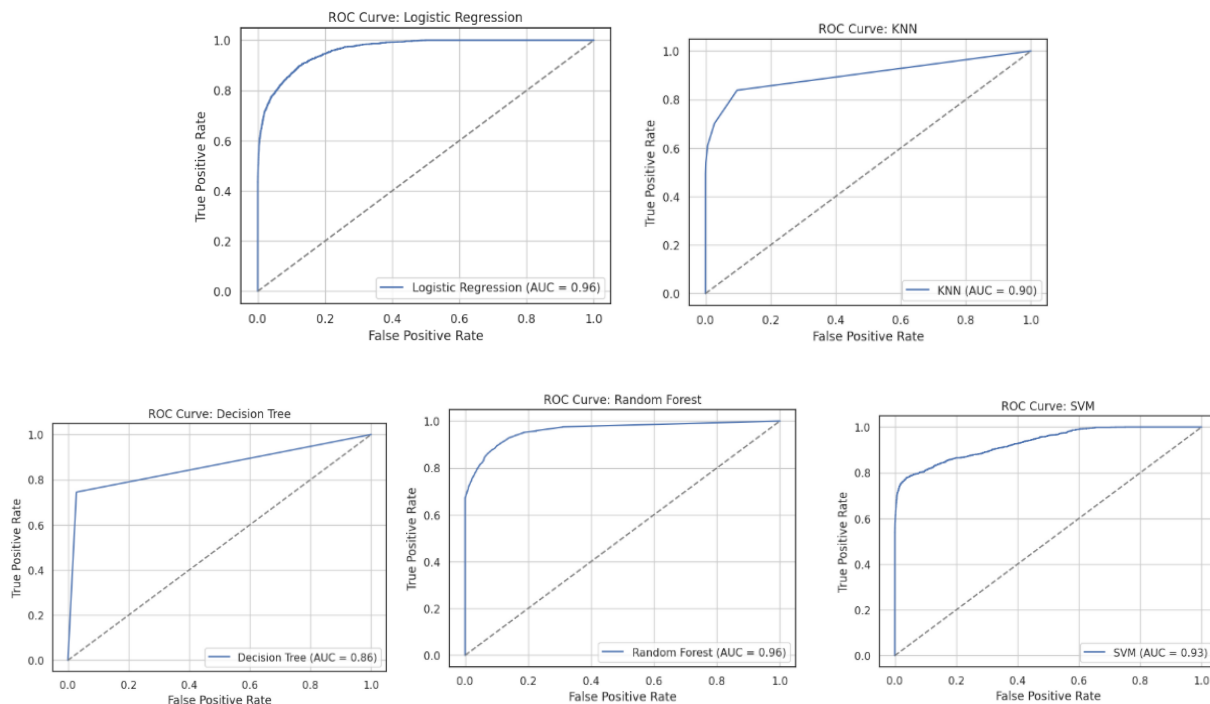


Fig.28: Comparison of ROC-AUC curves with 5 classifiers.

Hyperparameter tuning: Based on the evaluation results, further analysis focused solely on the Random Forest classifier due to its superior performance across multiple metrics. To optimize its predictive capability, hyperparameter tuning was conducted using GridSearchCV. The best parameters identified were: max_depth: 5 and n_estimators: 50

```
Best parameters: {'randomforestclassifier__max_depth': 5, 'randomforestclassifier__n_estimators': 50}
Best cross-val score: 0.9721197724573358
Test Accuracy: 0.9708456268440266
```

Based on previous evaluations, the focus was placed on the Random Forest classifier for further optimization. Hyperparameter tuning was performed using GridSearchCV, which tested 18 parameter combinations over 5-fold cross-validation (90 total fits).

```
Fitting 5 folds for each of 18 candidates, totalling 90 fits

✔ Best Hyperparameters for Random Forest:
{'randomforestclassifier__max_depth': 10, 'randomforestclassifier__min_samples_split': 5, 'randomforestclassifier__n_estimators': 50}

📊 Classification Report (Tuned Random Forest):
```

	precision	recall	f1-score	support
0	0.97	1.00	0.98	18222
1	1.00	0.67	0.80	1775
accuracy			0.97	19997
macro avg	0.98	0.84	0.89	19997
weighted avg	0.97	0.97	0.97	19997

These results above confirm that hyperparameter tuning further improved the Random Forest model's performance and generalization ability.

Feature Importance – Tuned Random Forest: For diabetic cases (Class 1), although precision is perfect (no false positives), recall is lower (67%), meaning the model misses some true positive cases. The overall performance is very high but recall for the diabetic class could still be improved. Therefore, Feature importance is calculated:

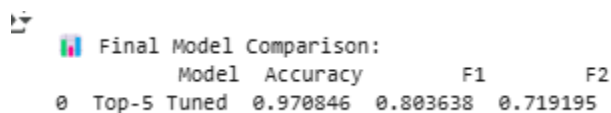
	Feature	Importance
4	HbA1c_level	0.413342
5	blood_glucose_level	0.319270
3	bmi	0.108556
0	age	0.107603
1	hypertension	0.015521

A performance comparison was conducted between the Full Model (using all features) and the Top-5 Model (using only the five most important features identified by feature importance).

Results are below

	Model	Accuracy	F1 Score
0	Full Model	0.968695	0.796357
1	Top-5 Model	0.966595	0.786718

F2 score is a weighted version of the F1 score that puts more emphasis on Recall (i.e., catching all actual diabetic cases).



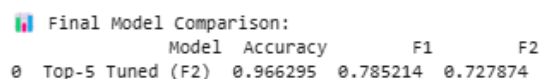
Final Model Comparison:				
	Model	Accuracy	F1	F2
0	Top-5 Tuned	0.970846	0.803638	0.719195

The Top-5 Tuned model, after hyperparameter optimization, slightly outperformed the Full Model in both accuracy and F1 score and additionally delivered a strong F2 score.

To further optimize the model for medical diagnosis, GridSearchCV was re-run with a focus on maximizing the F2 Score, which places greater emphasis on Recall over Precision. This adjustment reflects the clinical priority of minimizing false negatives, where failing to identify diabetic cases presents a significantly higher risk than issuing false positives.

Fitting 5 folds for each of 9 candidates, totalling 45 fits

✓ Best Hyperparameters: {'randomforestclassifier__max_depth': None, 'randomforestclassifier__n_estimators': 50}



Final Model Comparison:				
	Model	Accuracy	F1	F2
0	Top-5 Tuned (F2)	0.966295	0.785214	0.727874

SMOTE + GridSearchCV 5Folds + Top-5 Model (F2 Scoring): To address class imbalance and prioritize Recall—crucial in medical diagnostics—SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset prior to training. Including GridSearchCV with 5-fold cross-validation was conducted to maximize the F2 Score, favoring a reduction in false negatives

Fitting 5 folds for each of 9 candidates, totalling 45 fits

✓ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 200}

📊 Final Model Comparison:

	Model	Accuracy	F1	F2
0	Top-5 Tuned + SMOTE (F2)	0.905786	0.630298	0.770559

This model above demonstrates a significant improvement in F2 Score, reinforcing its suitability for healthcare applications where recall is critical. The 5-fold model is better for maximizing Recall and F2 Score(to avoid missing diabetic cases). This model provides a strong balance between accuracy and recall-driven performance, making it well-suited for diabetes prediction where minimizing false negatives is a top priority.

SMOTE + GridSearchCV (10 Folds) – Top-5 Features (F2 Scoring): To further improve model robustness and generalizability, the number of cross-validation folds was increased to 10, providing a more stable estimate of performance across subsets of the data. To enhance model stability and sensitivity, the final iteration results below:

Fitting 10 folds for each of 9 candidates, totalling 90 fits

✓ Best Hyperparameters: {'classifier__max_depth': 5, 'classifier__n_estimators': 50}

📊 Final Model Comparison:

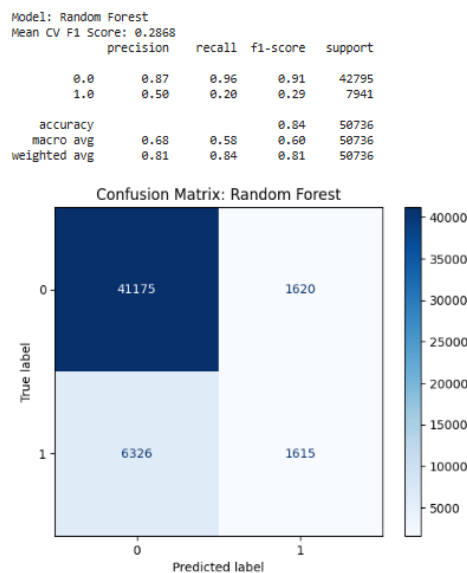
	Model	Accuracy	F1	F2
0	Top-5 Tuned + SMOTE (F2)	0.914187	0.643983	0.764909

10-fold model tuned model is more balanced model with slightly better generalization and robustness

Results & Discussion for Dataset #2:

In this section, top-performing diabetes prediction model (SMOTE + GridSearchCV 5Folds + Top-5 Model (F2 Scoring): and SMOTE + Research 10Folds + Top-5 Model (F2 Scoring) are applied to domain-specific dataset #2.

The initial Random Forest model was trained and evaluated on a large, imbalanced dataset. The results show strong performance on the majority class (non-diabetic), with a precision of 0.87 and recall of 0.96, resulting in an F1 score of 0.91. However, the model struggled significantly with the minority class (diabetic), achieving a recall of only 0.20 and an F1 score of 0.29.



After hyperparameter tuning, the Random Forest model achieved an improved accuracy of 86%. Precision for the diabetic class increased slightly to 0.46, though recall remained low at 0.17, indicating the model still struggles to identify positive diabetes cases despite tuning.

Fitting 5 folds for each of 18 candidates, totalling 90 fits

✓ Best Hyperparameters for Random Forest:
{'randomforestclassifier__max_depth': None, 'randomforestclassifier__min_samples_split': 2, 'randomforestclassifier__n_estimators': 100}

📄 Classification Report (Tuned Random Forest):

	precision	recall	f1-score	support
0.0	0.88	0.97	0.92	43739
1.0	0.46	0.17	0.25	6997
accuracy			0.86	50736
macro avg	0.67	0.57	0.59	50736
weighted avg	0.82	0.86	0.83	50736

The results below F1 score did not improve:

Fitting 5 folds for each of 9 candidates, totalling 45 fits

✓ Best Hyperparameters: {'randomforestclassifier__max_depth': None, 'randomforestclassifier__n_estimators': 50}

Final Model Comparison:

	Model	Accuracy	F1	F2
0	Top-5 Tuned (F2)	0.846224	0.283563	0.222395

The best-tuned Random Forest model using F2 score optimization achieved an accuracy of 84.62%, with an F1 score of 0.284 and F2 score of 0.222.

SMOTE + GridSearchCV (5 Folds) – Top Features (F2 Scoring): Although overall accuracy decreased to 71.89%, the model became much more effective at detecting the minority class, making it a more reliable tool for early diabetes detection.

Fitting 5 folds for each of 9 candidates, totalling 45 fits

✓ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 50}

Final Model Comparison:

	Model	Accuracy	F1	F2
0	Top-5 Tuned + SMOTE (F2)	0.718878	0.463696	0.611488

SMOTE + GridSearchCV (10 Folds) – Top Features (F2 Scoring): After applying SMOTE and tuned Random Forest model (with max_depth=10 and n_estimators=50), the model's F1 score improved to 0.464 and F2 score to 0.611, indicating significantly better performance in identifying diabetic cases below:

Fitting 10 folds for each of 9 candidates, totalling 90 fits

✓ Best Hyperparameters: {'randomforestclassifier__max_depth': 10, 'randomforestclassifier__n_estimators': 100}

Final Model Comparison:

	Model	Accuracy	F1	F2
0	Top-5 Tuned + SMOTE (F2)	0.718701	0.463136	0.610629

The dataset #2 utilizing advanced machine learning techniques, including cross validation (k=5 or K-10) hyperparameter optimization with GridSearchCV and addressing class imbalance with Synthetic Minority Over-sampling Technique (SMOTE), we aim to enhance the predictive accuracy of classification models for diabetes as learned for dataset #1.

Conclusions, Limitation, Ethical Consideration

During the modeling process, several machine learning models demonstrated high accuracy in predicting diabetes, accuracy alone is not sufficient in medical applications. F2 Score becomes a more critical metric, as it emphasizes Recall over Precision, which is crucial to avoid missing true diabetic cases. This is appropriate for healthcare where missing a diagnosis is more dangerous than overdiagnosing.

After multiple model iterations and evaluation rounds, the best-performing algorithm was the Tuned RF Classifier, trained using the top 5 most influential features, with class imbalance handled via SMOTE and hyperparameters optimized using GridSearchCV. The model was validated using both 5-fold and 10-fold cross-validation. This approach not only has strong accuracy but also significantly improved F2 Score, making it more suitable for deployment in clinical screening settings where early detection is paramount.

In terms of computational performance, there were no significant delays observed during the execution of either the baseline Random Forest or the Tuned Random Forest models. efficiency was not a distinguishing factor in model selection, with performance metrics remaining the primary focus for evaluation.

Data preparation was a critical component of the modeling methodology. Dataset did not contain missing values. The model optimization is limited to dataset's top 5 features. Not all variables in the dataset were taken into account. For example, 'smoking history' variable was excluded due to approximately 35% of its values lacking information, which could compromise model reliability.

Additionally, demographic variables such as ethnicity, genetic factors (whether biological parents have diabetes or not), or geographical location are important for capturing the socio-environmental and hereditary influences on diabetes risk. During the modeling process, demographic variables were not taken into consideration. The dataset may have introduced some biases. For example dataset #1 there is an increase in frequency at age 80, suggesting a higher representation of elderly individuals in this dataset. Including demographic variables and features in future datasets could improve the model's robustness and enable more personalized and accurate risk predictions.

While the selected tuned RF model demonstrates strong performance in early diabetes prediction, it is essential to address ethical aspects. Dataset may have introduced bias. Dataset 1 contains features such as gender and age, but it lacks ethnicity data, limiting the ability to

assess or mitigate bias across racial or ethnic groups. Dataset 2 includes similar demographic features but is imbalanced. Both Dataset 1 and Dataset 2 are anonymous and publicly available. The project has not looked into model transparency and interpretability in the clinical setting. Finally, the selected tuned RF model has only been validated on two datasets in this project. It has not been tested in a healthcare environment with real-time data or diverse patient populations.

References

- ¹ International Diabetes Federation. (n.d.). *Diabetes facts & figures*. Retrieved January 27, 2025, from <https://idf.org/about-diabetes/diabetes-facts-figures/>
 - ² World Health Organization. (n.d.). *Diabetes*. Retrieved January 27, 2025, from <https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=Factors%20that%20contribute%20to%20developing,tests%20with%20a%20healthcare%20provider.>
 - ³ Centers for Disease Control and Prevention. (2023). *National Health and Nutrition Examination Survey, 201–2023: Data brief 516*. Retrieved January 27, 2025, from <https://www.cdc.gov/nchs/products/databriefs/db516.htm#:~:text=The%20age%2Dadjusted%20prevalence%20of%20total%20diabetes%20increased%20from%209.7,in%20August%202021%E2%80%93August%202023>
- World Health Organization. (2024, November 13). Urgent action needed as global diabetes cases increase four-fold over past decades. WHO. <https://www.who.int/news/item/13-11-2024-urgent-action-needed-as-global-diabetes-cases-increase-four-fold-over-past-decades>
- Chou, C.-Y., Hsu, D.-Y., & Chou, C.-H. (2023). *Predicting the onset of diabetes with machine learning methods*. *Healthcare*, 11(4), 573. <https://pmc.ncbi.nlm.nih.gov/articles/PMC10057336/>
- Rani, K. J. (2020). Diabetes prediction using machine learning. ResearchGate. https://www.researchgate.net/publication/347091823_Diabetes_Prediction_Using_Machine_Learning
- Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia Computer Science*, 132, 1578–1585. <https://www.sciencedirect.com/science/article/pii/S1877050918308548>.
- Houngué, P., & Bigirimana, A. G. (2022). Leveraging Pima dataset to diabetes prediction: Case study of deep neural network. *Journal of Computer and Communications*, 10(11), 15–28. <https://www.scirp.org/journal/paperinformation?paperid=120929>
- Khanam, J. J., & Foo, S. Y. (2021). A comparison of machine learning algorithms for diabetes prediction. *ICT Express*, 7(4), 432–439. <https://www.sciencedirect.com/science/article/pii/S2405959521000205?via%3Dihub>
- The Guardian. (2024, November 13). More than 800 million people around the world have diabetes, study finds. The Guardian. <https://www.theguardian.com/society/2024/nov/13/diabetes-rates-increase-world-study>
- Nathan, D. M., Kuenen, J., Borg, R., Zheng, H., Schoenfeld, D., & Heine, R. J. (2010). Translating the A1C assay into estimated average glucose values. *Diabetes Care*, 31(8), 1473–1478. <https://pmc.ncbi.nlm.nih.gov/articles/PMC2999978/>
- American Diabetes Association (ADA). (2024). Standards of medical care in diabetes – Classification and diagnosis. <https://diabetes.org>
- According to the Centers for Disease Control and Prevention. (2022, September 1). BMI categories. U.S. Department of Health & Human Services. <https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html>

Centers for Disease Control and Prevention. (2022, September 30). *All about your A1C*. U.S. Department of Health & Human Services. <https://www.cdc.gov/diabetes/managing/managing-blood-sugar/a1c.html>

Centers for Disease Control and Prevention. (2022, May 4). *Diabetes tests*. U.S. Department of Health & Human Services. <https://www.cdc.gov/diabetes/basics/getting-tested.html>