Diabetes prediction using Machine Learning

CIND820 CAPSTONE PROJECT

Literature Review, Data Description, and Approach

Submitted by: Asma Shaikh (052129962)

Supervisor's name: Dr.Tamer Abdou

March 3, 2025

Toronto Metropolitan University

The Chang School of Continuing Education

Table of Contents

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 2050[1]. 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 2024[2]. 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 Status[3].

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. How do age and gender influence the probability of developing diabetes?
3. What is the relationship between Body Mass Index (BMI) and the likelihood of a positive diabetes diagnosis?
4. Are blood glucose levels consistently higher in obese/overweight individuals compared to those with a normal BMI?
5. How does smoking history impact the risk of developing diabetes in the future?
6. What is the accuracy of diabetes prediction models, and which model performs best?
7. How does pre-training a diabetes prediction model on a large, general dataset, followed by fine-tuning on a smaller, domain-specific dataset, affect the model's predictive performance and generalizability?

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 for this project 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 & Descriptive Statistics

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

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 (#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.

Dataset #1

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 3.1% of the data. Since the data is anonymous, duplicates are acceptable and are retained.

Key statistics (mean, mode, standard deviation) are summarized in Table-2.

	age	hypertension	heart_disease	bmi	HbA1c_level	blood_glucose_level	diabetes
count	100000.000000	100000.00000	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-2: 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.

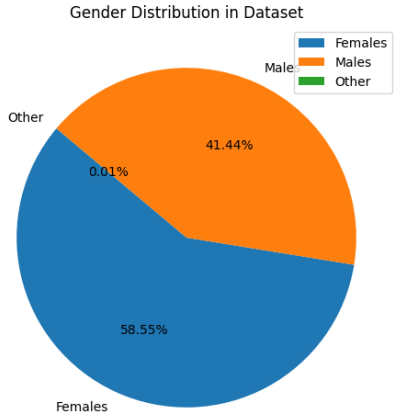

Fig-1: Ratio of male, female and others in the dataset.

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.

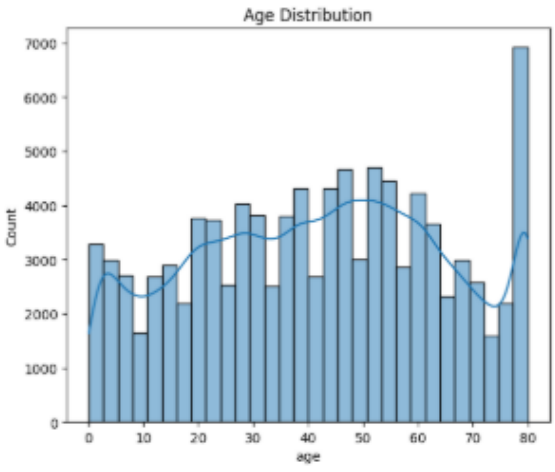

Fig.2: Distribution of "age" in the dataset

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.

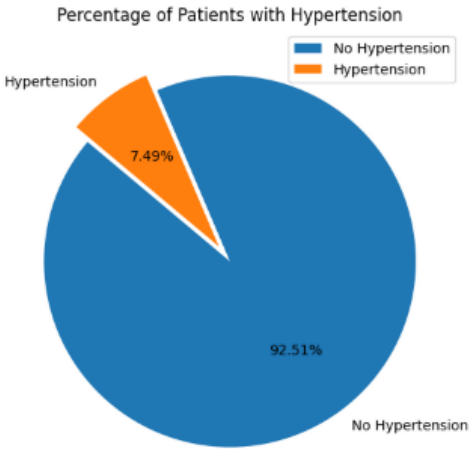

Fig.3: Ratio of "Hypertension" in the dataset

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.

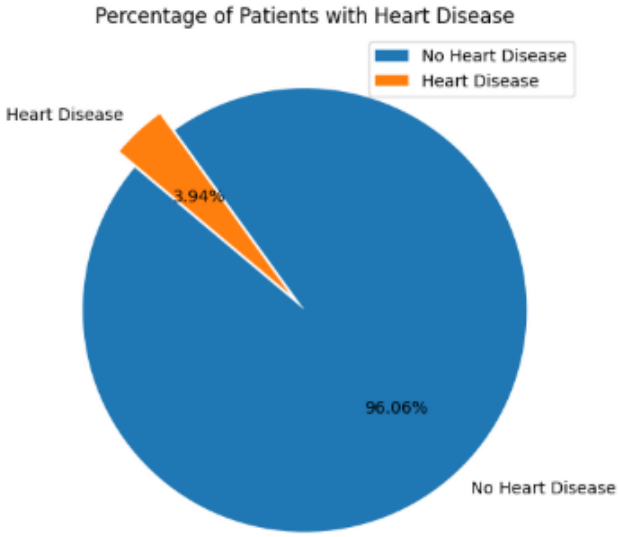

Fig.4: Ratio of "heath disease" in the dataset

5. Smoking History

35.81% of instances (35,816) have no information on smoking history, which may significantly impact results. Current smokers account for 11.1%, while former smokers represent 6.6%.
Given the large proportion of missing data, this column may need to be removed or imputed before further analysis.

Smoking History	Count	Frequency (%)
No Info	35816	35.816
never	35095	35.095
former	9352	9.352
current	9286	9.286
not current	6447	6.447
ever	4004	4.004

Fig 5: Smoking History Distribution Table

6. Body Mass Index (BMI)

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

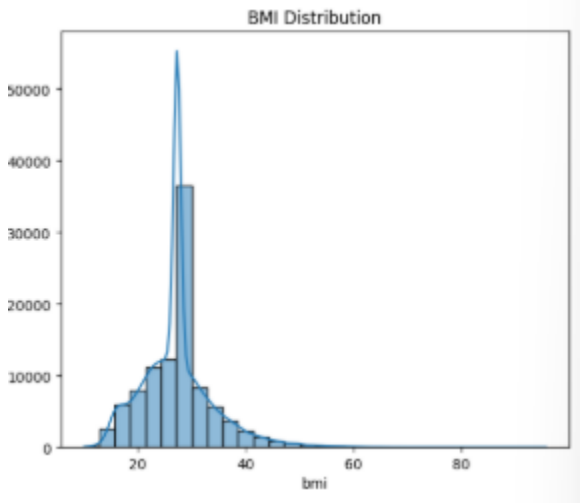

Fig.6: Distribution of HbA1c level in the dataset

The presence of high BMI values suggests outliers, which may require treatment before further modeling. After removing outliers, the dataset size was reduced from 100,000 rows to 92,914 rows, indicating that 7,086 rows (approximately 7.1%) were identified as outliers and excluded from further analysis.

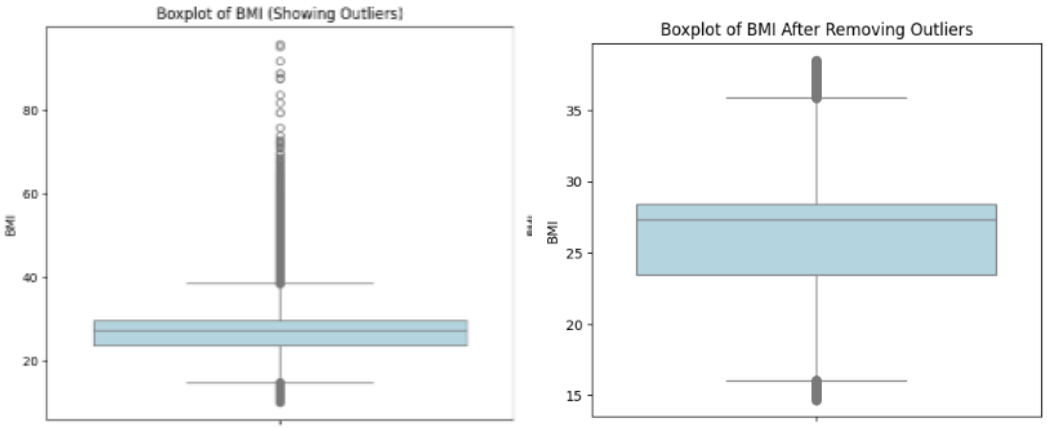

Fig.7: Boxplot for BMI showing outliers and without outlier.

7. HbA1c Level

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

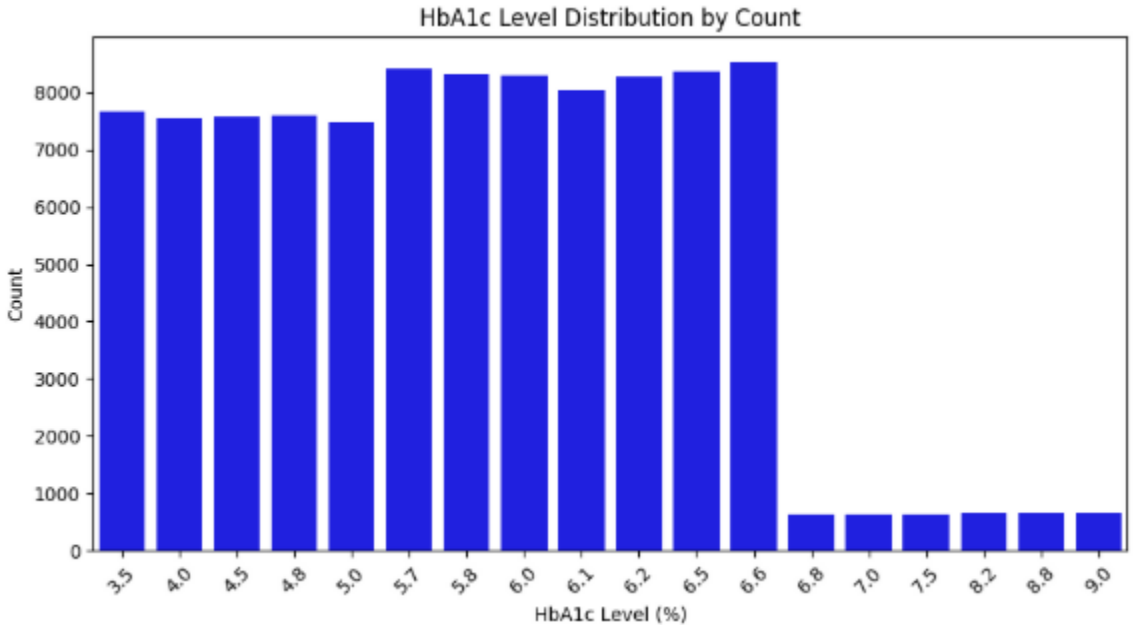

Fig 8: Distribution of HbA1c level in the dataset.

These percentage ranges refer to the Hemoglobin A1c (HbA1c) levels, a widely used biomarker for diagnosing diabetes and prediabetes. These thresholds are based on American Diabetes Association (ADA) guidelines:

HbA1c Category	Count	Percentage (%)
≤ 5.7% (Non-Diabetic)	46270	46.270
5.7% - 6.4% (Prediabetic)	32933	32.933
≥ 6.5% (Diabetic)	20797	20.797

Fig 9: HbA1c Level Table

8. Blood Glucose Level

The blood glucose levels range from 80 mg/dL to 300 mg/dL.

Glucose levels can be categorized as follows:

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

These thresholds align with the American Diabetes Association (ADA) guidelines for diabetes screening and diagnosis.

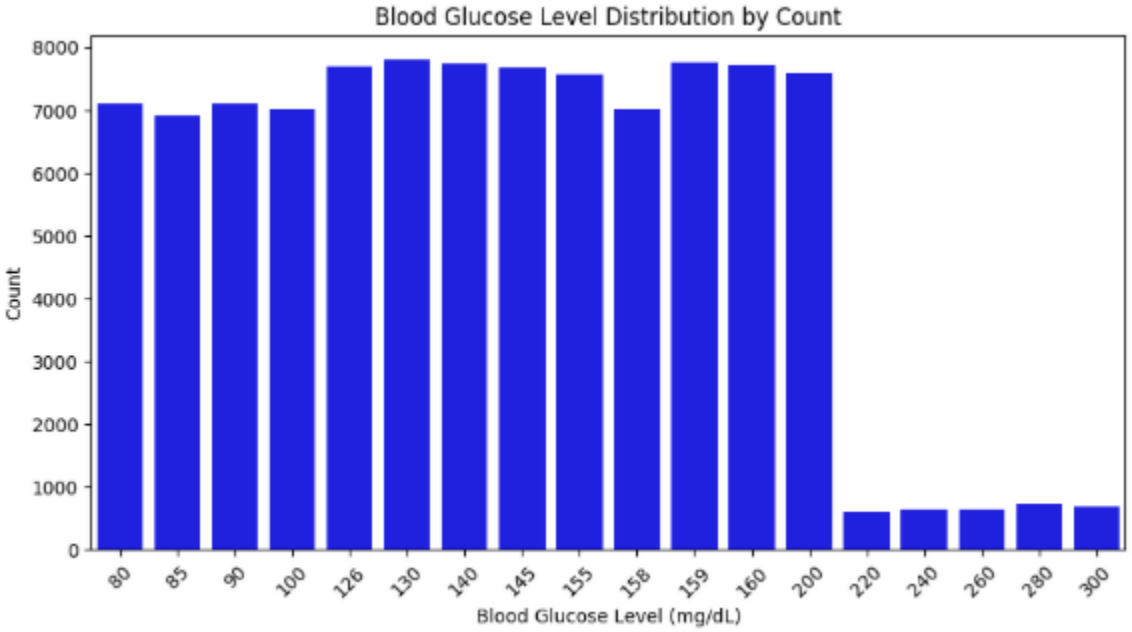

Fig.10: Distribution blood glucose level of persons in the dataset.

9. Diabetes (Target Feature)

Diabetes is the dependent variable (target feature) in this dataset. 8.5% of the dataset has diabetes. According to The Guardian (2024), the global prevalence of diabetes has doubled from 7% in 1990 to 14% in 2022. 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.

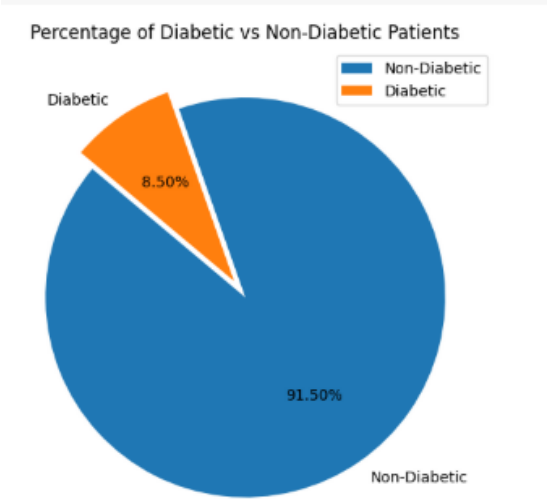

Fig.11: Distribution of Diabetes (Target Feature) in the dataset.

C. Bivariate Analysis

Bivariate analysis examines the relationship between two variables in the dataset. The correlation heatmap represents a form of bivariate analysis using Y Profiling.

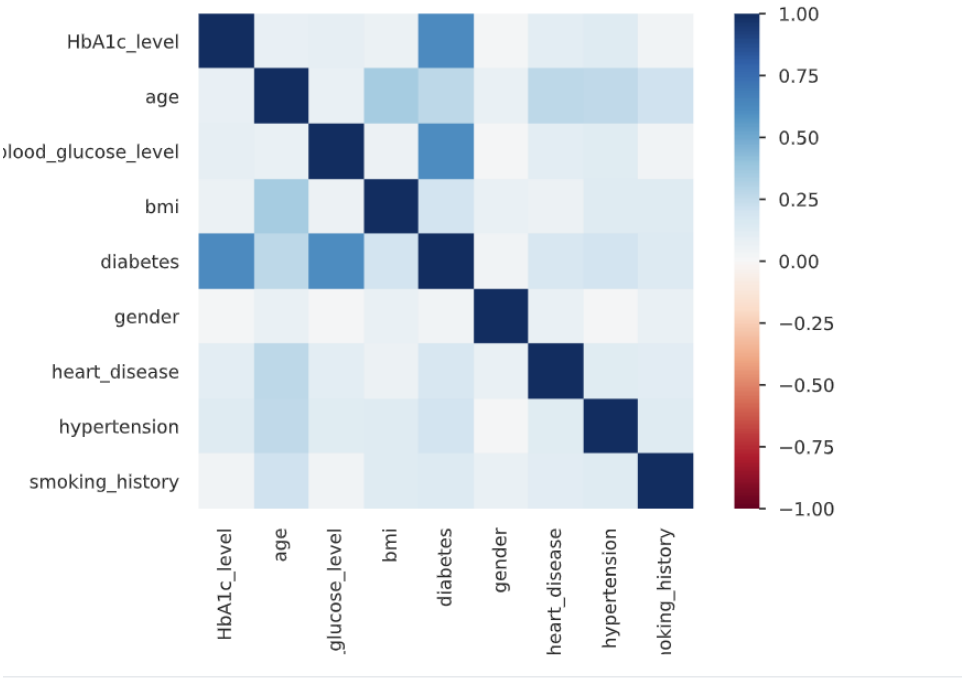

Fig.11: Correlation Heatmap using Y-profiling to the dataset.

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.
- HbA1c level and blood glucose level show a strong positive correlation, indicating that higher blood glucose levels are associated with increased HbA1c.
- 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.
- The heatmap suggests little to no correlation between gender and diabetes, implying that both men and women are similarly affected in this dataset
- 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. HbA1c_level thresholds are endorsed by the American Diabetes Association (ADA) and the World Health Organization (WHO). If the value of HbA1c lies

HbA1c level	Actual diagnosis
< 5.7 Normal	100% no diabetes
5.7 – 6.4 Prediabetes	7.47% have diabetes
>= 6.5 Diabetes	23.64% have diabetes

Table-4: HbA1c level vs diagnosis.

As HbA1c value increases in the dataset, the percentage of actual diagnosis of diabetes increases and for HbA1c level 6.5 or greater, it is 5%.

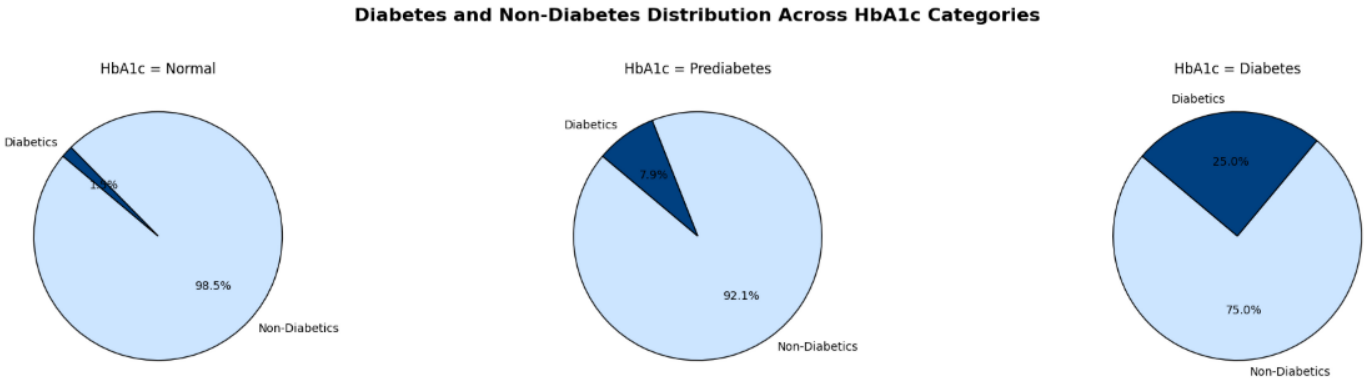

Fig.12: Pie Chart for HbA1C vs Diabetes

Blood Glucose Level vs diabetes:

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 (FBG) values. It is a clinical reference used to diagnose normal glucose levels, prediabetes, and diabetes.

Blood Glucose Level (mg/dL)	Category
=< 99	Normal
100 – 125	Prediabetes
>= 126	Diabetes

Table-3: Standard chart for Blood Glucose Level Categories

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

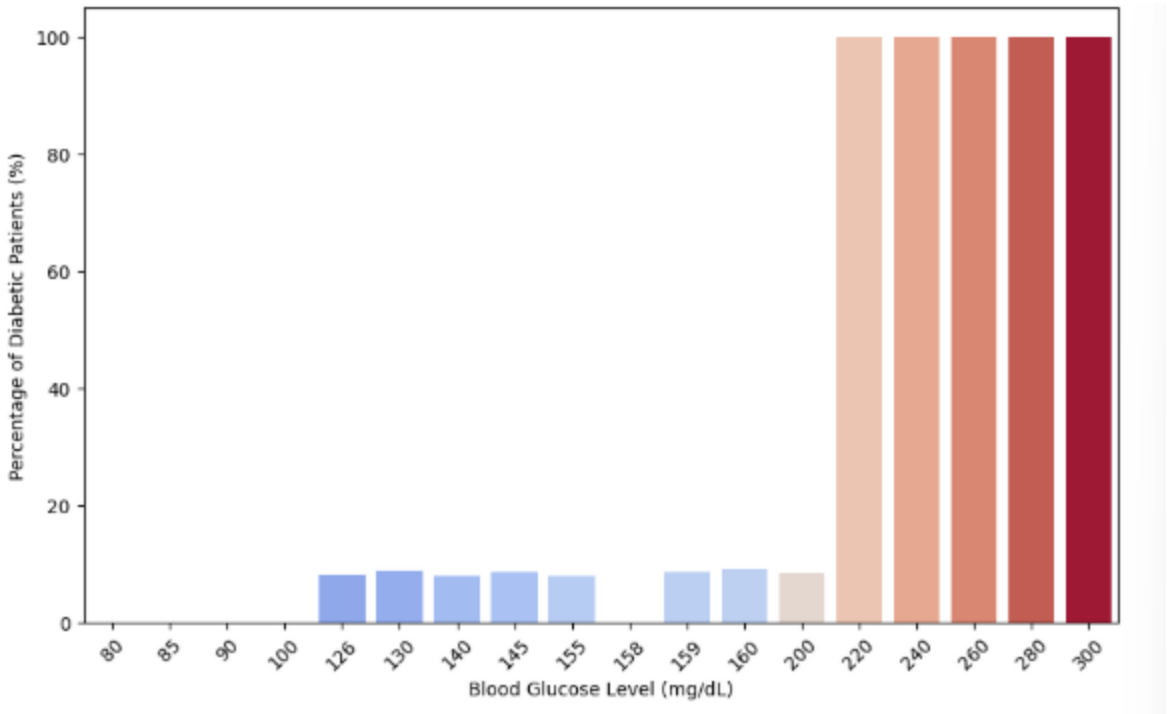

Fig.13: Bar plot for Glucose Level vs percentage of persons with diabetes

BMI vs diabetes:

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

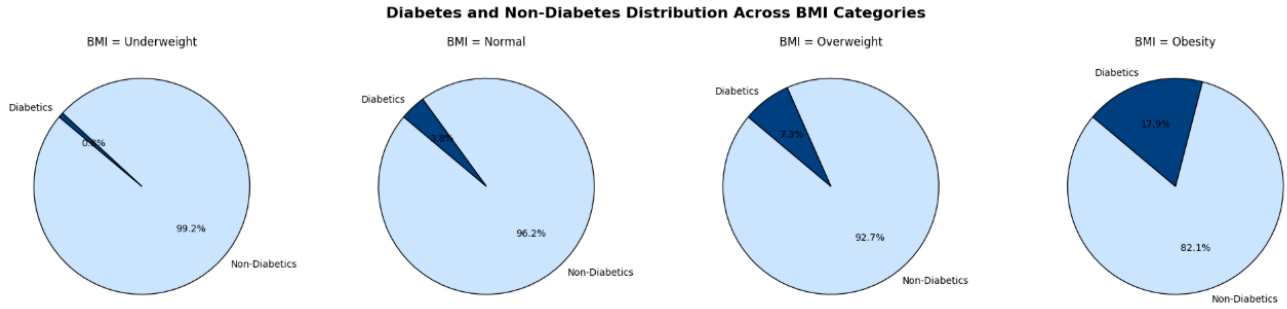

Fig.14: Pie Charts for different ordinal category and diabetes

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). Only Independent Variables to be validated that exist in dataset 1.

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.

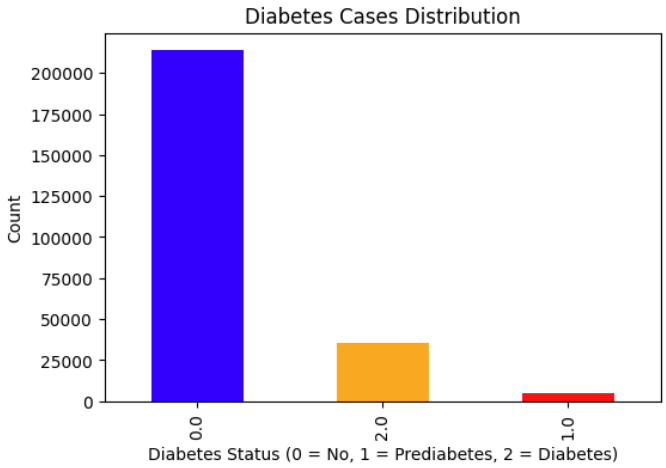

Fig.15: Bar plot for Distribution of Diabetes Status in the Dataset #2

The correlation heatmap for the dataset #2 shows high blood pressure, cholesterol, BMI, and physical activity might show meaningful relationships with diabetes.

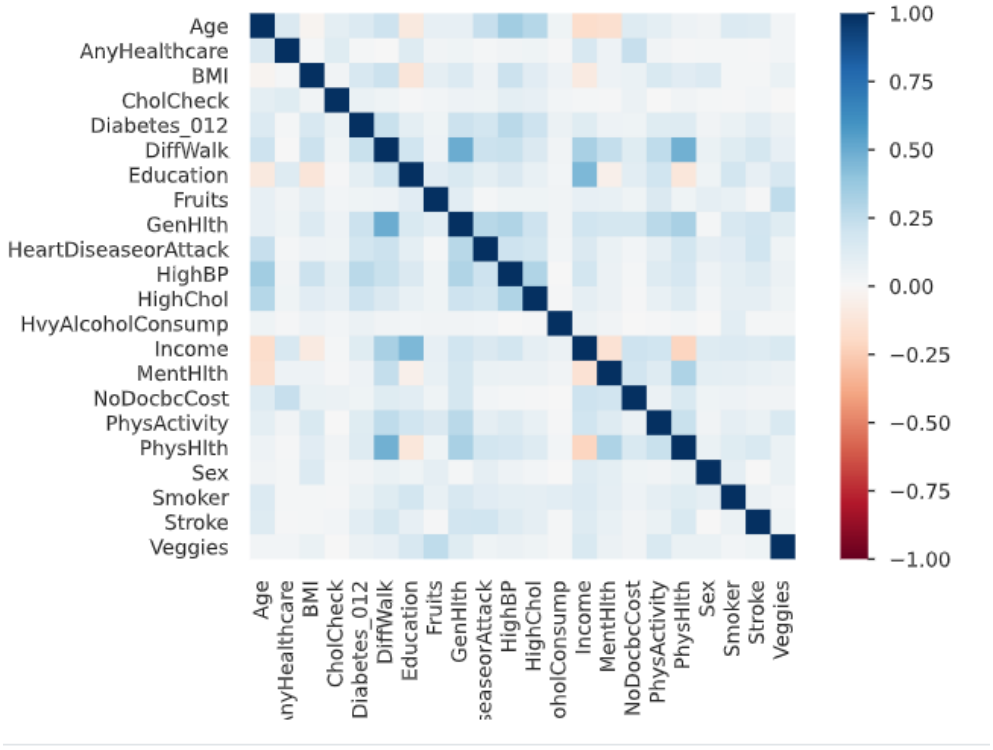

Fig.16: Correlation Heatmap dataset #2 using Y-profiling to the dataset.

High BMI, high blood pressure, and cholesterol are strongly linked to diabetes.There are strongest positive correlations between age and heart disease/High Blood Pressure. Higher BMI is positively correlated with diabetes. People with high blood pressure often have high cholesterol. Poor general health is associated with more physical health problems. Smoking may not be a direct predictor of diabetes in this dataset.

GitHub

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.17: 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 the consistent and ready for the next stage.

Next, feature selection is performed to identify the most relevant attributes, improving model accuracy as applicable. Following this, different machine learning algorithms will be trained using 80%-20% train-test split or K-Fold Cross-Validation. Further scaling on the dataset was performed before applying the five algorithms were: logistic regression, k-nearest neighbor regression, decision tree classifier, random forest classifier, support vector modeling. After the first iteration of each model, outcomes are evaluated.

To evaluate model performance, key metrics such as accuracy, precision, recall, and F1-score are used. These metrics provide insights into the effectiveness and reliability of each model in predicting diabetes. To improve the outcomes, cross-validation may be applied on the model.In the final stage, comparative analysis is performed and analyzed the results to identify the best model.

The pre-trained model (from dataset #1) is fine-tuned using the smaller dataset (#2), to the target variable—diabetes prediction. This process adjusts the model parameters using the smaller dataset to enhance its predictive accuracy for diabetes.

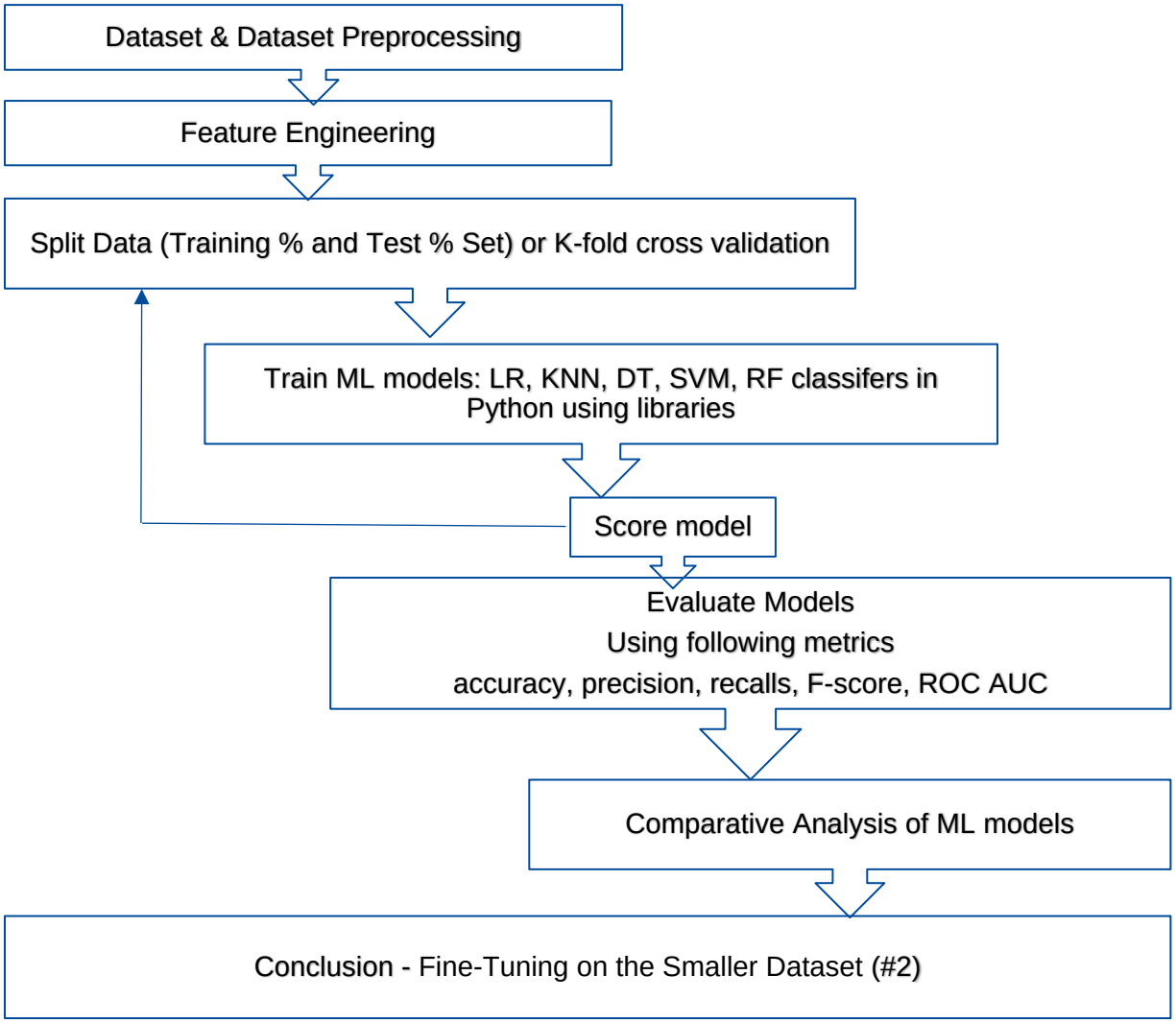

Fig.17: Process flowchart of Project Methodology below.

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, 2021–2023: 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.

Hounguè, 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 comparisn 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