RESEARCH TREATISE

A Dissertation Submitted to

SVKM's NMIMS (Deemed To Be University) In

Partial Fulfilment for the Degree of

M.Sc. Statistics and Data Science 2024-2026

"Analysing Deterministic Mathematical Model of Diabetes Mellitus."

By: Group – 6

Roll No.	Name	SAP ID
A020	Himani Grover	86062400022
A037	Lawrence Mondal	86062400062
A063	Sanya Saxena	86062400046
A068	Prashant srivastav	86062400071
A069	Harsh Tantak	86062400038

Project Supervisor

Dr. Debasmita Mukherjee

Nilkamal School of Mathematics,Applied Statistics and Analytics,
SVKM's Narsee Monjee Institute of Management Studies (Deemed-To-Be-University)

V.L. Mehta Rd, Vile Parle (West), Mumbai – 4000

Analysing Deterministic Mathematical Model of Diabetes Mellitus

1 Acknowledgement

We would like to express our sincere gratitude to our mentor, Dr. Debasmita Mukherjee, for her invaluable support and guidance throughout the course of this project. Her expertise and thoughtful advice, especially in the selection of the project topic, were instrumental in shaping the direction of our work. Her approachability and consistent communication made it easy for us to seek assistance whenever we faced challenges. Her constructive feedback and insights helped us refine our project and achieve our objectives. This project would not have been possible without her encouragement and guidance, and we are deeply grateful for her mentorship.

Contents

2 Abstract

Type 2 Diabetes Mellitus (T2DM) is a type of long-term metabolic syndrome where the body produces too much blood sugar due to an inability to use or create insulin in a normal manner. As such, there are important contributors to its onset such as lifestyle, genetics, glucose and insulin levels, cholesterol levels, and body mass index (BMI). This study utilizes a modeling technique to depict how glucose and other associated factors escalate the course of T2DM. For the visualization of data, box plots, scatter plots, and radar plots are used to demonstrate the observable patterns and relationships. To characteristics of the interrelationships, analysis of linear regression is used to determine which element is most related to, or which has the strongest predictive influence, over diabetes. Such analyses put modeling, data visualization and analysis together providing concepts on how T2DM occurs as well as most appropriate and effective intervention procedures for ty2dm and its management.

3 Introduction

Diabetes mellitus is a chronic metabolic disease in which a person has high levels of glucose in their blood. The most prevalent type of diabetes is Type 2, dependent on lifestyle factors such as diet and exercise, and genetic predisposition. The disease interaction includes the level of glucose, insulin, and different types of cholesterol in the body, intake of food and body mass index (BMI).[1] Mathematical models are very important in understanding the dynamics of these interactions, which generally help in designing more efficient interventions and treatment plans for researchers and healthcare professionals.[2] In this work, a deterministic mathematical approach aims to model such interactions for better insight into the dynamics of diabetes.
Hyperglycemia is a hallmark of diabetes mellitus, and especially of T2DM. A chronic metabolic disease, T2DM is primarily caused by insulin resistance or insufficiency. Lifestyle and genetic factors that play a role with this disease include body composition, food intake, and physical activity. Development of type 2 diabetes is so complex that several biomarkers including glucose and insulin levels, cholesterol, and even BMI needed to be studied by using mathematical models. Models help identify the crucial points for intervention and prepare for management and prevention plans of problems associated with diabetes.[3] In this study, we employ a deterministic mathematical model that integrates key biomarkers and demographics to analyze T2DM dynamics. Data visualization techniques, such as box plots, scatter plots, and radar plots, are used to examine trends and relationships within the data, revealing critical insights into how diabetes and associated health factors manifest across different population groups. Additionally, linear regression analysis is applied to assess the predictive relationships between variables, with a focus on determining which factors most significantly impact diabetes classification. Through this combined approach of modeling, visualization, and regression analysis, this research provides a comprehensive understanding of diabetes progression and the efficacy of potential interventions.[4]

[5],[6],[7],[8],[9],[10],[11],[12],[13],[14],[15],[16],[17],[18],[19],[20],[21]

4 Rationale

Mathematical models provide a systematic methodology to interpret the intricacies of chronic diseases such as diabetes mellitus. The deterministic approach enables the simulation and prediction of the effect of glucose and other factors related to the disease on glucose levels and the general progression of the disease. To identify the primary drivers of disease progression and stability in the system, we limit our study to measurable clinical parameters that include glucose, insulin, HDL, LDL, and BMI. This may be of use in guiding individualized interventions and contributing to the wider scientific understanding of diabetes.

5 literature review

Type 2 diabetes mellitus is still a burden on global health with an increasingly high prevalence of the disease and all its implications for public health and economic resources. Many studies demonstrate that the problems in T2DM management are more complex than that is assumed and that understanding and modulating a disease progression must be multi-dimensional.

Frankum and Ogden (2005) had researched blood glucose self-estimation in diabetic patients and concluded that gaps exist between the expectations of patients to adequately predict their glucose levels. This implies that there is a need for more adequate education and help the patient receives to ensure he regulates his glucose levels properly, as this is fundamentally a component of diabetes management.

The study by AlShurbaji et al. 2023 is the most comprehensive study done on a diabetes mellitus mathematical model, using numerical analysis for evaluation of stability and parameters of the disease. The authors highlight the importance of stability analysis in achieving a better understanding of dynamics in diabetes and forecasting outcomes when scenarios are varied. Their results confirm the utility of deterministic modeling and how the advancement of the disease is modified by various interventions-in this case, dietary changes and treatment.

Studies by Boutayeb et al. (2004) applied population models to better understand the transition from prediabetes into advanced stages of T2DM, with early intervention generally understood as a means to improve health outcomes. As defined by this population-based approach, lifestyle intervention has been proven to be essential in the fight against T2DM: targeted changes in health population strategies can sufficiently slow down the disease.

Forlenza et al. evaluated the safety and efficacy of an automated insulin delivery system named the Minimed 670G in pediatric patients diagnosed with T1DM. As far as this study is T1D focused, it demonstrates how technology adapted through mathematical modeling improved glucose regulation. Such developments represent the potential inclusion of automated systems in the management of diabetes toward maintenance of continuous glucose monitoring and adaptive treatment.

Gyorgy et al (2015) conducted research on adaptive control solutions for T1DM by proposing real-time correction of insulin infusion depending on the glucose value. These studies provide insight into the potential of control systems in minimizing complications regarding diabetic treatment. Their application in T2DM would be invaluable as a direction for future research when combined with mathematical modeling that predicts and corrects the resultant treatment response in real time.

6 Methodology:

- Mathematical modeling
- Data visualization
- Estimation
- Multilinear regression analysis

6.1 Model Development:

In order to explain the dynamics of glucose, insulin, BMI, dietary intake, Low-Density Lipoprotein (LDL), and High-Density Lipoprotein (HDL) in the context of type 2 diabetes, a deterministic mathematical model was created using Ordinary Differential Equations (ODEs). Based on known physiological relationships between these variables, this model was developed with the goal of capturing how blood glucose levels change and are regulated under various physiological circumstances.

6.1.1 Parameter Estimation:

Estimating Parameters A non-linear optimization technique was used for parameter estimation in order to calibrate the model based on observed data. We used the scipy.optimize.minimize function in Python, which employs numerical optimization techniques to find the best-fit parameters that minimize the discrepancy between the simulated outcomes from the ODE model and the observed values in the dataset. By using the sum of squared errors (SSE) between simulated and observed data points as the optimization's objective function, precise parameter estimation that most closely mirrored physiological processes was made possible.

6.1.2 Statistical Analysis:

The data was analyzed with the use of numerous statistical and graphical techniques that sought to explore the relationships between the variables. Moreover, it was possible to use the validation properties of the model developed such as the non-negativity property, existence and uniqueness of solutions, existence of an equilibrium point, and stability analysis to validate the predictive power of the ODE model. To determine relationships between significant variables (dietary intake related to BMI, glucose to insulin concentration etc.) amongst others, linear regression analysis was done. Such linear regression was done within the Python statsmodels library which enabled the use of parameter estimates, the confidence intervals and R-squared values to demonstrate model fit in terms of the picture that the data created.
Additionally, the demographic and physiological traits of people with and without diabetes were compiled using descriptive statistics, which shed light on the distinctions between the two groups.

6.1.3 Model Validation:

By contrasting the optimized ODE outputs with the observed data, the model's validity was evaluated. To make sure the model appropriately captures the behaviour of the biological system for people with type 2 diabetes, we assessed the model's performance using statistical metrics like Mean Squared Error (MSE) and R-squared. Furthermore ,We examined key parameters to determine the model's stability across a variety of situations and to identify critical parameters impacting the system's dynamics. [**?**]

6.1.4 Software and Tools:

The main programming language for modelling, data analysis, and visualisation was Python. Among the important libraries were:

- SciPy: for optimisation and numerical integration
- NumPy: for managing mathematical operations and numerical data
- Pandas: for analysing and manipulating data For producing excellent data and model output visualizations, use Matplotlib and Seaborn. For reproducibility of results, all scripts used for data processing, modelling, and analysis are available and documented.

7 Definitions

1.Beta modeling: Beta modeling refers to the use of rate constants (*beta*) to represent the rates at which transitions or changes occur between different states or compartments in a mathematical model. These rate constants are crucial in compartmental models, such as disease progression models, to quantify the effects of various factors on the movement from one state to another.

2.Deterministic Mathematical Model: A model wherein the outcome is known with precision because the relationships among the states and events are well established without any forms of randomness variability. In this study, a deterministic approach is used in predicting the progression of T2DM on measurable clinical and demographic parameters.

3.Linear Regression: A statistical method to measure the relationship between one dependent variable and one or more independent variables. In this paper, linear regression helps in predicting diabetes classification by studying the influence of factors such as age, BMI, and cholesterol.

4.Stability Analysis: A technique applied in determining the stability of a model's equilibrium states that help understand how small changes in

parameters could influence the overall system. This is an important analysis particularly in modeling chronic diseases and long-term behavior. .

5.Body Mass Index (BMI): Healthcare professionals use the body mass index (BMI) to calculate your body fat percentage based on your height and weight. It can assist in determining the risk factors for specific medical conditions. Body fatness is not always accurately represented by the BMI.

6.HbA1c: 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.

7.High-Density Lipoprotein (HDL) and Low-Density Lipoprotein (LDL): LDL/HDL cholesterol ratio is the ratio of two types of lipids in the blood. LDL stands for low density lipoprotein and HDL stands for high density lipoprotein. LDL is often considered "bad" cholesterol and HDL is "good" cholesterol.

8 frame work

- framing equations considering factors(glucose level,insulin level,cholesterol,age,BMI,Carbohydrate intake)
- framing equations considering factors(glucose level,insulin level,cholesterol,age,BMI,Beta cell intake)
- Data cleaning and plotting
- Using Linear Regression model to predict the diabetes class predictions.

9 Equations

The papers studied so far focus on parameters separately, and the *deterministic model* is considered.In this Paper, various parameters are considered simultaneously. **Parameters:**

- Glucose Level
- insulin level
- dietary intake
- HDL
- LDL
- BMI

$$\frac{dG(t)}{dt} = k - (k_1 - k_3)G(t) \tag{1}$$

$$\frac{dI(t)}{dt} = k_4 G(t) - k_2 I(t) \tag{2}$$

$$\frac{dB(t)}{dt} = k_6 D(t) - k_7 B(t) \tag{3}$$

$$\frac{dH(t)}{dt} = -k_9 D(t) - k_{10} H(t) \tag{4}$$

$$\frac{dL(t)}{dt} = -k_{11} D(t) + k_{12} B(t) - k_{13} L(t) \tag{5}$$

$$\frac{dD(t)}{dt} = -k_{15} I(t) - k_{16} D(t) \tag{6}$$

9.1 Notations

- k: population entering type 2 diabetes
- $k_1 - k_3$:Represents natural glucose decay (use by cells).
- k_2:Represents natural insulin decay over time.
- k_4:Insulin production increases with glucose level G(t) at rate k_4
- k_6:BMI increases with dietary intake, at rate k_6
- k_7:Natural decrease in BMI, possibly due to energy expenditure.
- k_9:HDL decreases with certain dietary intake factors.
- k_{10}:Natural decay of HDL over time.
- k_{11}:LDL decreases with dietary intake at rate k_{11}.
- k_{12}:LDL is influenced by BMI.
- k_{13}:Natural decay rate of LDL.
- k_{15}:Insulin reduces dietary intake as a response to satiety.
- k_{16}:Natural decrease in dietary intake over time.

9.2 Stability Analysis

$$k - (k_1 - k_3)\, G_1(t) = 0 \tag{1}$$
$$k_4\, G_1(t) - k_2\, G_2(t) = 0 \tag{2}$$
$$-k_5\, G_1(t) + k_6\, D_1(t) = 0 \tag{3}$$
$$k_6\, D_1(t) - k_7\, D_2(t) = 0 \tag{4}$$
$$-k_9\, D_1(t) + k_{10}\, B_1(t) = 0 \tag{5}$$
$$-k_{11}\, D_1(t) + k_{12}\, B_1(t) - k_{13}\, L_1(t) = 0 \tag{6}$$

$$G_1^* = \frac{k}{k_1 - k_3}$$
$$G_2^* = \frac{k_4}{k_2} \cdot \frac{k}{k_1 - k_3}$$
$$D_1^* = \frac{k_5}{k_6} \cdot \frac{k}{k_1 - k_3}$$
$$D_2^* = \frac{k_5}{k_7} \cdot \frac{k}{k_1 - k_3}$$
$$B_1^* = \frac{k_9}{k_{10}} \cdot \frac{k_5}{k_6} \cdot \frac{k}{k_1 - k_3}$$
$$L_1^* = \frac{k_{12} \cdot \frac{k_9}{k_{10}} \cdot \frac{k_5}{k_6} \cdot \frac{k}{k_1 - k_3} - k_{11} \cdot \frac{k_5}{k_6} \cdot \frac{k}{k_1 - k_3}}{k_{13}}$$

$$J = \begin{pmatrix} -(k_1 + k_3) & 0 & 0 & 0 & 0 & 0 \\ k_4 & -k_2 & 0 & 0 & 0 & 0 \\ 0 & -k_{15} & -k_{16} & 0 & 0 & 0 \\ 0 & 0 & k_6 & -k_7 & 0 & 0 \\ 0 & 0 & -k_9 & 0 & -k_{10} & 0 \\ 0 & 0 & -k_{11} & k_{12} & 0 & -k_{13} \end{pmatrix}$$

$$\lambda_1 = -(k_1 + k_3)$$
$$\lambda_2 = -k_2$$
$$\lambda_3 = -k_{16}$$
$$\lambda_4 = -k_7$$
$$\lambda_5 = -k_{10}$$
$$\lambda_6 = -k_{13}$$

10 Data Description

The dataset consists of 1,000 records covering various clinical and demographic attributes related to diabetes, including the following columns:

- **Demographics:** ID, No_Patient, Gender, AGE
- **Health Metrics:** Urea, Cr (Creatinine), HbA1c, Chol (Cholesterol), TG (Triglycerides), HDL, LDL, VLDL, and BMI
- **Diabetes Classification:** CLASS

Summary Statistics:

- **Age:** Ranges from 20 to 79, with a mean of around 54 years.
- **HbA1c (Average Blood Sugar):** Varies widely, with a mean of 8.28, indicating many individuals may have elevated levels (above 6.5 typically indicates diabetes).
- **BMI:** Average BMI is about 29.6, which is in the overweight range.
- **Cholesterol, Triglycerides, and Other Lipids:** These values vary widely, indicating a mix of normal and elevated levels among the subjects.

11 Data Visualizations

11.1 Box and Violin Plot

The box and violin plots compare the age distribution across different diabetes classes: N (no diabetes), P (prediabetes), and Y (diabetes).

11.1.1 Box Plot Analysis

The purpose of box plots is to provide a concise summary of the age distribution for each class, highlighting the median, quartiles, and potential outliers.
Observations:

- The median age is higher for the Y (diabetes) class than for N (no diabetes) and P (prediabetes) classes.
- The spread (interquartile range) is wider in the Y class, indicating greater age variability among people with diabetes.
- There are more outliers in the Y group, suggesting some individuals with diabetes are much younger or older than the majority.

11.1.2 Violin Plot Analysis

The purpose of violin plots is to show the entire age distribution, including the density of data points at different age values, providing more detail about the distribution shape than box plots.
Observations:

- The N and P classes have a fairly even distribution across ages, while the Y class shows a peak around the median, indicating that most individuals with diabetes fall within a specific age range.
- The spread of ages in the Y class appears wider at both ends, consistent with the box plot.

Purpose of Using Both Plots:

- Box Plot: Useful for summarizing central tendency, variability, and identifying outliers.
- Violin Plot: Adds information on the shape of the distribution, revealing nuances in density that box plots might miss.

11.2 Scatter Plot Analysis

The scatter plot visualizes the relationship between HDL (High-Density Lipoprotein) and LDL (Low-Density Lipoprotein) levels, with the color intensity representing total cholesterol (Chol) levels.
Axes:

- The x-axis shows HDL levels, while the y-axis shows LDL levels.
- Both HDL and LDL are types of cholesterol; HDL is often called "good" cholesterol, while LDL is known as "bad" cholesterol. Higher HDL and lower LDL are typically desirable for cardiovascular health.

Color Gradient:

- The color gradient indicates total cholesterol levels, with darker red hues representing higher cholesterol values.
- Higher cholesterol values (darker red) seem concentrated at certain LDL levels, indicating individuals with higher LDL tend to have elevated total cholesterol.

Patterns and Insights:

- HDL values are clustered mostly between 0.5 and 2.0, suggesting that most individuals fall within this range.
- As LDL levels increase, total cholesterol generally increases as well, which aligns with the fact that LDL is a significant contributor to total cholesterol.

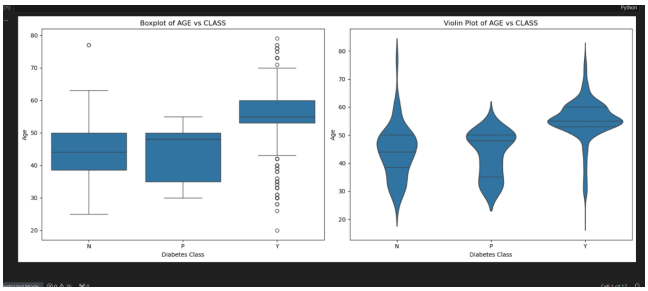

Figure 1: Violin Plot Analysis for Age Distribution Across Diabetes Classes

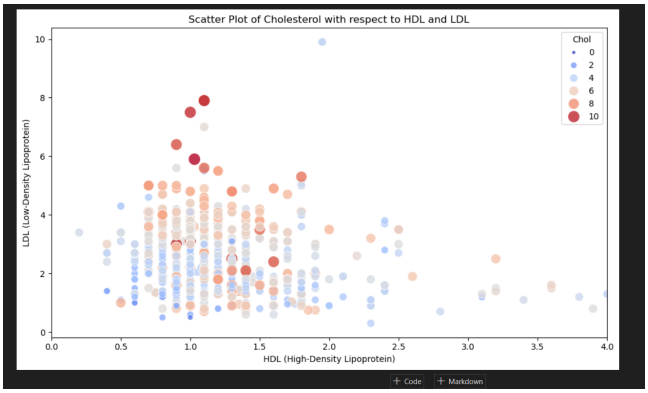

Figure 2: Scatter Plot of HDL vs LDL Levels

- Higher LDL with low HDL may indicate a higher cardiovascular risk.

Purpose: The scatter plot is ideal for examining correlations and clustering patterns between HDL, LDL, and total cholesterol. The color intensity provides a multidimensional view.

11.3 Scatter Plot of Age vs HbA1c

This scatter plot displays the relationship between age (on the x-axis) and HbA1c levels (on the y-axis), with a regression line included to illustrate the trend.
Analysis:

- There is a positive correlation between age and HbA1c levels. As age increases, HbA1c levels tend to increase, indicated by the upward-sloping regression line.
- The regression line indicates the general trend in the data, with the confidence interval suggesting a relatively strong relationship between age and HbA1c levels.
- There are a few outliers with higher HbA1c values, which may represent individuals with poorly controlled diabetes.

11.4 Boxplot of Age vs Diabetes Class

This boxplot shows the distribution of age across three different diabetes classes: N (No Diabetes), P (Prediabetic), and Y (Diabetes).
Observations:

- No Diabetes (N): The median age is around 45, with an interquartile range (IQR) from about 35 to 50. There is an outlier above 70.
- Prediabetic (P): The median age is approximately 55, with the IQR extending from around 42 to 60.
- Diabetes (Y): The median age is around 60, with an IQR from about 50 to 65. More outliers are present in the younger age group, indicating a small but significant number of younger individuals with diabetes.

Trends and Insights:

- The median age increases from the "No Diabetes" group to the "Diabetes" group, suggesting a general trend where older age correlates with a higher likelihood of diabetes or prediabetes.
- The presence of younger outliers in the diabetic category could indicate that lifestyle and genetic factors also play a significant role.

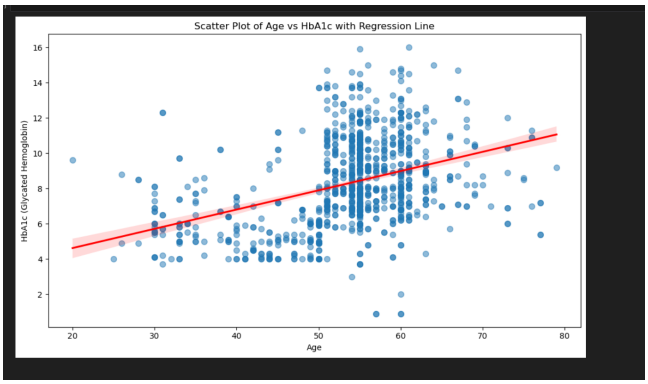

Figure 3: Scatter Plot of Age vs HbA1c Levels

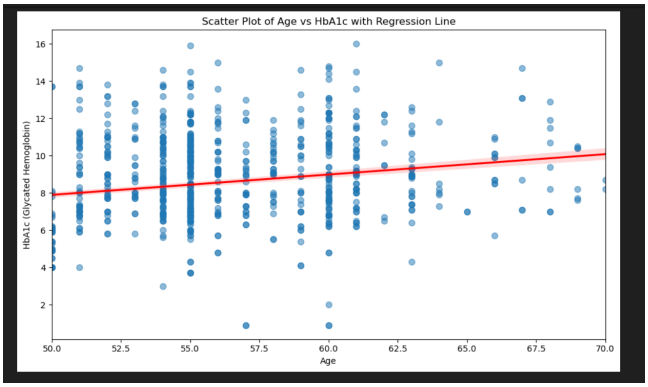

Figure 4: Scatter Plot of Age vs HbA1c Levels

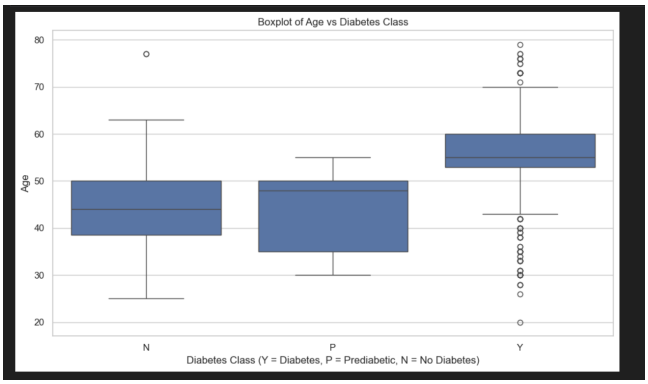

Figure 5: Plot of Health Metrics Across Diabetes Classes

11.5 Gender vs Diabetes

This bar chart shows the relationship between gender and diabetes classification.
Key Observations:

- The largest group is in the diabetic category (Y), with more males than females.
- The "No Diabetes" (N) group has slightly more females than males.
- The "Prediabetic" (P) group has the fewest individuals for both genders.

11.6 Radar Plot of Health Profiles by Diabetes Class (30-50 Age Group)

This radar plot shows the average health profiles for individuals in the 30-50 age group across three diabetes classes: Prediabetic (P), Diabetic (Y), and No Diabetes (N).
Observations:

- The Diabetic group (Y) has the highest values for Chol and HbA1c, followed by the Prediabetic group (P), and the No Diabetes group (N) has the lowest values across all metrics.
- The plot highlights how health metrics like cholesterol and HbA1c vary across diabetes classes.

12 Estimation Using Bootrap:

12.1 Objective:

To identify critical factors associated with the classification of individuals into three categories based on diabetes status:

- Class P: Prediabetic
- Class N: Non-diabetic
- Class Y: Diabetic

12.1.1 Methodology:

Using multinomial logistic regression and bootstrap parameter estimation, coefficients were calculated for each feature in the dataset, enabling the determination of the relationship between each feature and the likelihood of belonging to a specific diabetes class relative to a reference class. The bootstrapped confidence intervals further validate the stability of each feature's influence across samples.

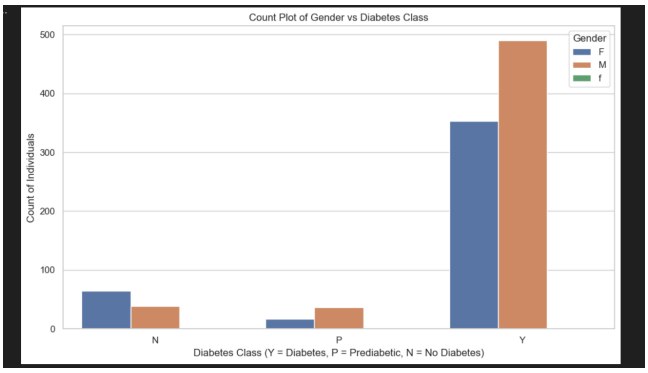

Figure 6: Histogram of Gender vs Diabetes

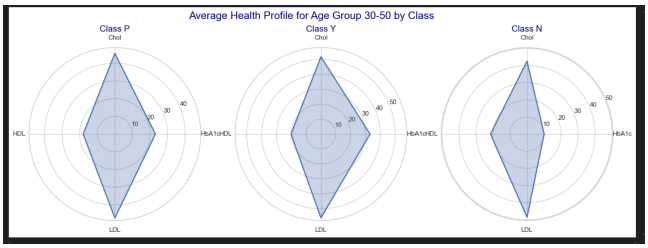

Figure 7: Radar Plot of Health Metrics Across Diabetes Classes

12.1.2 Key Findings:

Important Features for Classification

- 1. The blood glucose control HbA1c test is said to be very precise as a predictor of prediabetes and diabetes diagnosis. The class Y likelihoods of being given a diabetes diagnosis and elevated risk because of extremes in HbA1c values are predictable and confirmed with strong evidence. The prediction of non-diabetes (N class) diagnosis is adverse, just as the prediabetes risk factor case.
- 2. BMI (Body Mass Index) Coefficient Impact : Consistently negative coefficients, especially for those classified as non-diabetic and prediabetic **Interpretation:** The chance of being diagnosed with diabetes or prediabetes is significantly correlated with BMI. This illustrates the known connection between metabolic diseases, which are risk factors for diabetes, and elevated body mass index
- 3. Triglycerides (TG) **Coefficient Impact:** Significant negative coefficients, particularly in relation to Non-diabetic classification. **Interpretation:** A higher risk of prediabetes and diabetes is linked to elevated triglyceride levels. As a prevalent indicator of metabolic syndrome, triglycerides lend credence to its significance as a predictor.
- 4. VLDL (Very Low-Density Lipoprotein) **Coefficient Impact:** VLDL has considerable variability but remains a notable predictor across classes. **Interpretation:** Higher VLDL levels contribute to an increased likelihood of Diabetic and Prediabetic classification, reflecting dyslipidemia as a factor in diabetes development.
- 5. HDL (High-Density Lipoprotein) **Coefficient Impact:** Positive coefficient for Non-diabetic classification. **Interpretation:** Higher HDL is linked to a lower risk of developing diabetes or prediabetes, which is consistent with HDL's protective function in metabolic health.
- 6. Cholesterol and LDL **Coefficient Impact:** Moderate coefficients with less clear differentiation among classes. **Interpretation:** Even though they are secondary, these characteristics nonetheless influence the model's overall prediction, especially when combined with VLDL and triglyceride levels.

12.1.3 Some more Interpretations and results

The model effectively represents the dynamics of the observed data for the corresponding variable when the parameters are tuned, as seen by the closeness of the simulated and observed lines over time points. Significant differences between the simulated and observed curves over particular ranges or at particular times hint to potential areas for model improvement.

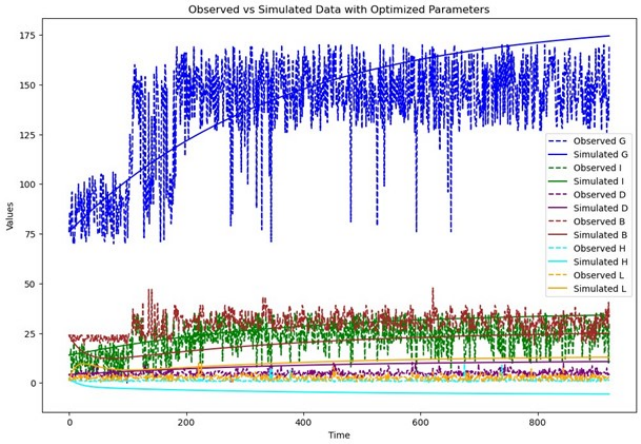

Figure 8: Observed vs Stimulated Data with Optimized Parameters

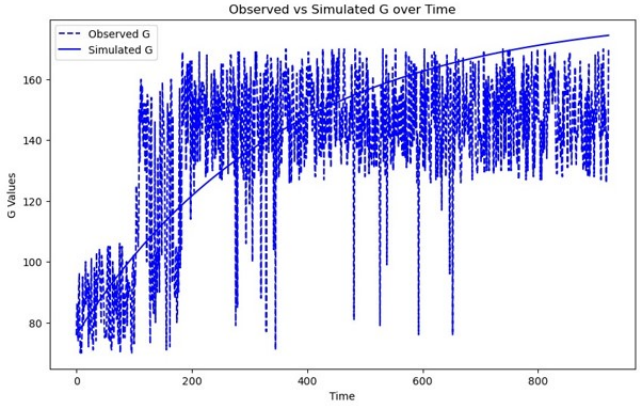

Figure 9: Observed vs Stimulated G over time

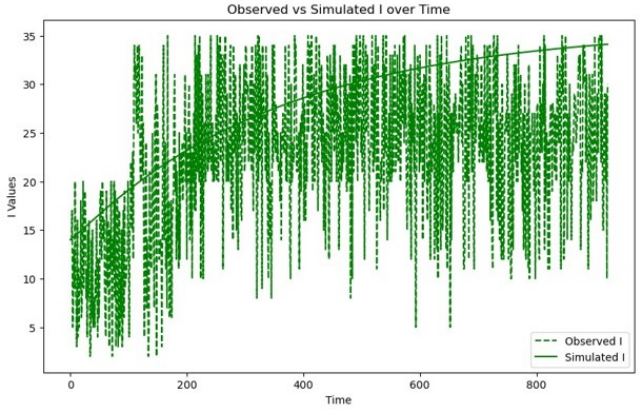

Figure 10: Observed vs Stimulated I over time

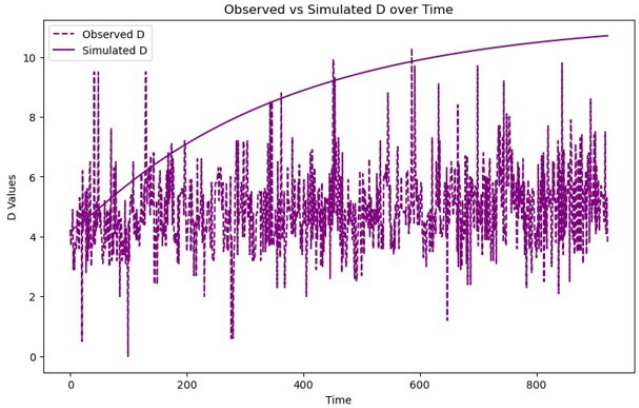

Figure 11: Observed vs Stimulated D over time

12.1.4 Summary of Influential Features by Class

- Prediabetic (Class P): HbA1c, BMI, and TG levels are primary indicators.
- Non-diabetic (Class N): Lower BMI and high HDL values are protective, lowering the risk of being classified as having diabetes.
- Diabetic (Class Y): The categorization of diabetes is favorably correlated with HbA1c, BMI, TG, and VLDL levels, indicating that these are the main risk factors.

12.1.5 Conclusion:

HbA1c, BMI, TG, and VLDL are highlighted in the model's estimation as the most significant predictors of diabetes classification, especially when it comes to identifying those who are Diabetic (Class Y). HDL is one of the secondary characteristics that helps identify people who do not have diabetes. By focusing on the most significant indicators in diabetes risk stratification, this data lays the groundwork for focused health treatments.
Stakeholders can use this report as a reference to select diabetes screening and preventive initiatives according to the most important predictive characteristics.

13 Linear Regression

13.1 Model Overview

This section presents the outcomes of a classification model applied to a diabetes dataset. The key metrics evaluate the model's effectiveness and highlight the significant features contributing to diabetes class predictions.

13.2 1. Model Accuracy

The model's accuracy stands at 90.5%, meaning it correctly classified 90.5% of the cases in the dataset.

13.3 2. Confusion Matrix

The confusion matrix below demonstrates the model's performance across different classes:

- **N (No diabetes)**: 33 predictions were correct, with 3 misclassified as P and 9 as Y.
- **P (Prediabetes)**: 13 correct predictions, with some instances misclassified as N and Y.

- **Y (Diabetes)**: 327 correct predictions, with some misclassified as N or P.

Additionally, the model did not successfully classify two samples with class Y.

13.4 3. Classification Report

The report offers insights into precision, recall, and F1 scores for each class:

- **N Class**: Precision = 0.79, Recall = 0.73, F1-score = 0.76.
- **P Class**: Precision = 0.29, Recall = 0.10, F1-score = 0.15.
- **Y Class**: Precision = 0.93, Recall = 0.98, F1-score = 0.96.

The Y class exhibits the highest precision and recall, suggesting the model performs well at identifying diabetes cases. However, the model has difficulty distinguishing cases in the P (prediabetes) class, which results in lower precision and recall.

13.5 4. Feature Importance

This part of the analysis highlights features that have the greatest impact on the model:

- **AGE** and **LDL** show positive coefficients, indicating an increase in the likelihood of predicting certain diabetes classes.
- **BMI**, **HbA1c**, and **Chol** have negative coefficients, which reduce the likelihood of specific classifications.

Top Features:

- **AGE**: 0.381 (positive influence)
- **LDL**: 0.306 (positive influence)
- **BMI**: -2.28 (negative influence), suggesting a critical but inversely impactful role in classification.

13.6 5. Warnings

The report includes warnings about metrics that are undefined for labels with no predicted samples. This implies that certain classes were not predicted, likely due to data imbalances.

13.7 Summary

Overall, the model performs well, particularly in identifying cases of Y (diabetes), though it has challenges with the P (prediabetes) category. Key influencing factors include AGE and LDL, while BMI has a significant inverse effect. Addressing data imbalances could improve accuracy for less represented classes.

13.8 Feature Coefficients Visualization

The bar chart below illustrates feature coefficients in the classification model, showing the impact of each feature on diabetes classification.

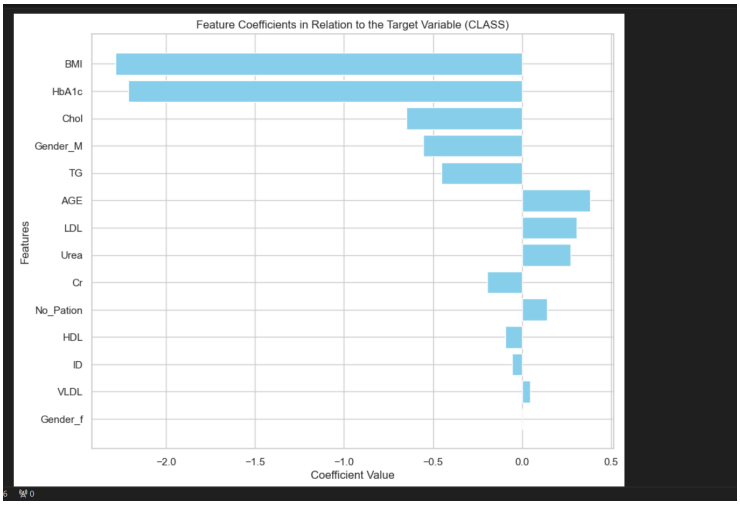

Figure 12: Feature coefficient in Relation to Target variables (Class)

13.8.1 Key Points:

- **Coefficient Value Interpretation**:
  - **Positive Coefficients**: Features with positive coefficients increase the likelihood of the positive class (e.g., diabetes) with higher values.
  - **Negative Coefficients**: Features with negative coefficients decrease the likelihood of the positive class as values increase.
- **Top Features with Major Influence**:
  - **BMI** and **HbA1c**: These features show the largest negative coefficients, indicating a strong influence in reducing the likelihood

of certain classifications. BMI is particularly impactful, followed closely by HbA1c.

  - **Chol** (Cholesterol) and **Gender_M** (Male): Both have negative coefficients, though their influence is less than BMI and HbA1c.

- **Additional Influential Features**:

  - **TG** (Triglycerides) and **AGE**: These features have moderate negative coefficients, suggesting a noticeable but lower impact compared to BMI and HbA1c.
  - **LDL** (Low-Density Lipoprotein): LDL shows a slight positive coefficient, indicating a higher LDL level may raise the likelihood of the positive class.

- **Features with Minor Influence**:

  - Features such as **Urea**, **Cr** (Creatinine), and **No_Pation** have smaller coefficients, implying minimal impact on classification.
  - **HDL** (High-Density Lipoprotein), **ID**, and **VLDL** show near-zero coefficients, suggesting they contribute little to the prediction.

14 Summary

TThe features BMI and HbA1c have the most significant negative influence on classification outcomes, followed by Cholesterol and Gender_M. LDL is among the few features with a positive coefficient, potentially linking it directly to the target class. This visualization aids in identifying influential features, helping guide model optimization and further analysis.

15 Conclusion

This analysis reveals that diabetes classification correlates with age, BMI, and cholesterol levels, with diabetics showing generally higher health metric averages.

16 Future scope

In this paper, the data is not as relevant as the observed data of some parameters is incorrect. So,In further research more refined clinical Data will be taken and estimation of BMI,LDL and HDl will be done.Stronger cross-validation, testing with various datasets, may point out possible overfitting or underfitting situations, thus more effectively suggesting a better parameter estimate.

References

[1] M. AlShurbaji, L. A. Kader, H. Hannan, M. Mortula, and G. A. Husseini, "Comprehensive study of a diabetes mellitus mathematical model using numerical methods with stability and parametric analysis," *Int J Environ Res Public Health*, vol. 20, no. 2, p. 939, 2023.

[2] G. P. Forlenza, O. Pinhas-Hamiel, and D. R. Liljenquist, "Safety evaluation of the minimed 670g system in children 7-13 years of age with type 1 diabetes," *Diabetes Technology and Therapeutics*, vol. 21, no. 1, pp. 11–19, 2019.

[3] R. A. DeFronzo, E. Ferrannini, L. Groop, R. R. Henry, W. H. Herman, J. J. Holst, F. B. Hu, C. R. Kahn, I. Raz, G. I. Shulman, *et al.*, "Type 2 diabetes mellitus," *Nature reviews Disease primers*, vol. 1, no. 1, pp. 1–22, 2015.

[4] R. M. Anderson and R. M. May, *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, 1992.

[5] R. N. Bergman, D. T. Finegood, and M. Ader, "Assessment of insulin sensitivity in vivo," *Endocrine Reviews*, vol. 6, no. 1, pp. 45–86, 1985.

[6] A. Boutayeb and A. Chetouani, "A critical review of mathematical models and data used in diabetology," *Biomedical Engineering Online*, vol. 5, no. 1, p. 43, 2006.

[7] C. Cobelli and A. Mari, "Control-oriented modeling of glucose kinetics for insulin control in diabetes," *Annals of Biomedical Engineering*, vol. 14, no. 3, pp. 219–229, 1986.

[8] A. De Gaetano and O. Arino, "Mathematical modeling of the intravenous glucose tolerance test," *Journal of Mathematical Biology*, vol. 40, no. 2, pp. 136–168, 2000.

[9] M. Derouich and A. Boutayeb, "The effect of physical exercise on the dynamics of glucose and insulin," *Computers in Biology and Medicine*, vol. 32, no. 4, pp. 299–309, 2002.

[10] S. Frankum and J. Ogden, "Estimation of blood glucose levels by people with diabetes: a cross-sectional study," *British Journal of General Practice*, vol. 55, no. 521, pp. 944–948, 2005.

[11] J. E. Gerich, J. M. Davis, and A. A. Baez, "Pathophysiology of type 2 diabetes and implications for therapy," *Endocrinology and Metabolism Clinics of North America*, vol. 27, no. 3, pp. 599–617, 1998.

[12] E. Gyorgy, K. T. Jozef, and K. Levente, "Adaptive control solution for type 1 diabetes control," in *IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*, pp. 215–220, 2015.

[13] H. W. Hethcote and J. A. Yorke, *Gonorrhea Transmission Dynamics and Control*. Springer Science and Business Media, 1984.

[14] H. W. Hethcote, "The mathematics of infectious diseases," *SIAM Review*, vol. 42, no. 4, pp. 599–653, 2000.

[15] A. J. Lotka, "Studies on the mode of growth of material aggregates. part i. analytical solution of equations of growth," *Journal of the American Chemical Society*, vol. 33, pp. 1815–1843, 1911.

[16] A. Makroglou, J. Li, and Y. Kuang, "Mathematical models and software tools for the glucose-insulin regulatory system and diabetes: an overview," *Applied Numerical Mathematics*, vol. 56, no. 3-4, pp. 559–573, 2006.

[17] R. S. Parker and F. J. Doyle, "Closed-loop control of blood glucose concentration," *Annual Review of Biomedical Engineering*, vol. 3, no. 1, pp. 245–267, 2001.

[18] M. Puckett and J. L. Hargrove, "Modeling the progression of type 2 diabetes with underlying obesity," *PLoS Computational Biology*, vol. 19, no. 8, p. e1010914, 2023.

[19] R. M. Souza, L. A. V. Bonamichi, and E. R. Pereira-Filho, "Design of experiments (doe) application in two cases of study in pharmaceutical industries," 2024.

[20] G. M. Steil, K. Rebrin, R. Janowski, *et al.*, "Modeling beta-cell insulin secretion–implications for closed-loop glucose homeostasis," *Diabetes Technology Therapeutics*, vol. 5, no. 6, pp. 953–964, 2003.

[21] B. Topp, K. Promislow, G. deVries, *et al.*, "A model of -cell mass, insulin, and glucose kinetics: pathways to diabetes," *Journal of Theoretical Biology*, vol. 206, no. 4, pp. 605–619, 2000.