Deanship of Graduate Studies

Al-Quds University

Leveraging Glycemic Index and Machine Learning for Predictive Modeling of Type 2 Diabetes Onset and Management

Jihan Mousa Mohammad Rahhal

M.Sc. Thesis

Jerusalem – Palestine

1446 - 2025

Leveraging Glycemic Index and Machine Learning for Predictive Modeling of Type 2 Diabetes Onset and Management

Prepared by:

Jihan Mousa Mohammad Rahhal

B.Sc. System Computer Engineering /Palestine Polytechnic University- Palestine

Supervisor: Dr. Radwan Qassrawi

A thesis submitted in partial fulfillment of requirement

for the degree of Master of Electronics & Computer Engineering- Deanship of Graduate studies - Al-Quds University

1446 - 2025

Al-Quds University

Deanship of Graduate Studies

Faculty of Health Profession

Medical Imaging Technology

Thesis Approval

Leveraging Glycemic Index and Machine Learning for Predictive Modeling of Type 2 Diabetes Onset and Management

Prepared by: **Jihan Mousa Mohammad Rahhal**

Registration No: **21612531**

Supervisor: **Dr. Radwan Qasrawi**

The master's thesis was submitted and accepted on 12/01/2025.

The names and signatures of the examining committee members are as follows:

Head of Committee: Dr. Radwan Qasrawi	Signature:
Internal Examiner: Dr.Saeed Salah	Signature:
External Examiner: Dr. Hazem Agha	Signature:

Jerusalem – Palestine

1446 - 2025

Dedication

I dedicate this work to my beloved family for their seiner love and ever-present support of my endeavors towards learning. To my friends who surrounded me with prayers and believed in me to all those from backstage who encouraged and helped me.

Jihan Mousa Mohammad Rahhal

Declaration

I certify that this thesis submitted to the degree of master is the result of my own research, except where otherwise acknowledged, and that this thesis or any of its parts has not been submitted for higher degree to any other university or institution.

Signature:

Jihan Mousa Mohammad Rahhal

Date: 12/01/2025.

Acknowledgement

I would like to give my warmest thanks to my supervisor "**Dr. Radwan Qassrawi**", his guidance and advice carried me through all the stages of writing my thesis.

To all my instructors at Al-Quds University who spared no effort in educating us and expanding our perceptions.

To my dearest **father**, you always believed in me and supported me to be the best, and reminds me "إنَّ الله لا يُضيعُ أجرَ مَن أحسنَ عملا".

To my **mother**, who never gave up to convince me to complete the path, actually, to start it too. Mama, you always remind me of Bukowski saying "when I look at her, the light goes through me", thank you for all your constant prayers.

To my brother, **Mohammad**, one word from you encourages me to conquer my mountains.

To my sisters**, Nadeen, Narmeen and Hala** my precious sisters, who always supported and trusted in me. Thanks for being my sisters, thanks for all the support you have given to me.

To my dear Master's colleagues, who filled this journey with unforgettable memories and countless smiles.

Finally, I would like to thank God for guiding me with mercy and strength through this experience, helping me overcome every challenge along the way.

Abstract

Type 2 Diabetes Mellitus is a multifaceted chronic disease influenced by a combination of sociodemographic, lifestyle, and nutritional factors. This study aimed to analyze the prevalence and determinants of diabetes in a diverse population while leveraging machine learning to enhance predictive modeling. Sociodemographic characteristics such as urban residency, lower education levels, aging, and female gender were significantly associated with higher diabetes prevalence. Lifestyle factors, including low physical activity (94.1% among diabetics) and obesity (75.3%), were prominent contributors. Nutritional analysis revealed a high prevalence of dietary deficiencies among diabetics, including low fiber and polyunsaturated fatty acid intake, along with excessive sodium consumption. These findings highlight the importance of addressing dietary quality and promoting physical activity as key components of diabetes prevention.

Machine learning models were employed to predict diabetes risk, with Gradient Boosting achieving the highest accuracy (94.2%) and AUC (0.985), outperforming other models such as Random Forest and Support Vector Machines. Feature importance analysis identified glycemic load, BMI, and age as the most significant predictors of diabetes, emphasizing the role of dietary glycemic measures, body composition, and aging in disease risk. The results reinforce the potential of machine learning as a tool for early diagnosis and risk stratification.

This study shows the need for integrated, multidisciplinary strategies to combat diabetes, including education, dietary interventions, and physical activity promotion. The findings provide a foundation for future research to validate these associations and develop targeted prevention programs. The application of machine learning demonstrates a promising avenue for personalized healthcare solutions in managing and reducing the burden of Type 2 diabetes.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

GI	**Glycemic Index**
GL	**Glycemic Load**
T2D	**Type 2 Diabetes**
T2DM	**Type 2 Diabetes Mellitus**
AUC	**Area Under the Curve**
BMI	**Body Mass Index**
ML	**Machine Learning**
SVM	**Support Vector Machine**
AUC-ROC	**Area Under the Receiver Operating Characteristic Curve**
IDF	**The International Diabetes Federation**
SES	**Socioeconomic Status**
CART	**Classification and Regression Trees**
RF	**Random Forest**
GB	**Gradient Boosting**
MENA	**Middle East and North Africa**
ANN	**Artificial Neural Network**
RDA	**Recommended Daily Energy Intake**
SFA	**Saturated Fatty Acids**
MUFA	**Monounsaturated Fatty Acid**
PUFA	**Polyunsaturated Fatty Acid**

Chapter One

Introduction

This chapter introduces the study of Type 2 Diabetes Mellitus, a major global health concern affecting millions and projected to rise sharply in the coming decades. The chapter explores the significant role that lifestyle factors, particularly diet and the glycemic index, play in the onset and management of T2DM. It also discusses the promising potential of machine learning to improve predictive models by integrating GI data with other health and lifestyle factors, offering a personalized approach to diabetes prevention. The aim of the study is to develop a predictive model that enhances T2DM risk prediction and supports tailored dietary management, with the ultimate goal of advancing precision medicine.

1.1. Introduction

Type 2 Diabetes Mellitus has become one of the most pervasive global health issues of the 21st century, affecting over 400 million individuals worldwide, with prevalence anticipated to exceed 700 million by 2045 if current trends continue (International Diabetes Federation 2019)Unlike Type 1 diabetes, which is primarily an autoimmune condition, T2DM is characterized by insulin resistance and relative insulin deficiency, largely driven by lifestyle and dietary factors(American Diabetes Association, 2021). T2DM often leads to severe complications, including cardiovascular disease, kidney failure, and neuropathy, substantially impacting patients' quality of life and placing a financial strain on healthcare systems globally (Zheng et al., 2018).

Among the various lifestyle factors influencing T2DM, diet plays a critical role in both its prevention and management. Extensive research indicates that the quality and type of carbohydrates consumed can significantly affect glycemic control and insulin sensitivity (Augustin et al., 2015). The Glycemic Index, a ranking system that classifies carbohydrates based on their postprandial (after-eating) blood glucose response, has gained widespread recognition as an essential tool in nutritional science and diabetes management (D. J. Jenkins et al., 1981). Foods with a high GI score tend to cause rapid spikes in blood sugar, whereas

low-GI foods contribute to slower and more controlled glucose release, potentially reducing the risk of insulin resistance and T2DM development(Brand-Miller et al., 2003).

In recent years, Machine Learning has emerged as a transformative approach to analyzing complex datasets in healthcare, enabling the identification of patterns and predictive markers for various diseases (Shickel et al., 2018). ML algorithms are capable of processing vast amounts of data, including individual dietary habits, lifestyle factors, and genetic predispositions, to generate personalized risk assessments and dietary recommendations. Integrating GI data with ML algorithms offers a unique opportunity to improve the prediction of T2DM onset by analyzing the specific impacts of carbohydrate quality on individual glycemic responses. Previous studies indicate that ML models, such as Decision Trees (DT), Support Vector Machines, and neural networks, show considerable promise in forecasting T2DM by identifying high-risk individuals based on a combination of dietary and lifestyle factors (Chen & Guestrin, 2016; Kavakiotis et al., 2017)

Despite the promise of GI as a dietary metric, it is often underutilized in traditional T2DM prediction and management due to the complexity of glycemic responses, which are influenced by numerous physiological and lifestyle variables. Standard approaches to dietary recommendations typically generalize nutritional advice, failing to consider individual variability in glycemic responses to specific carbohydrate types (Goff et al., 2013). This limitation presents an opportunity for ML, which excels at identifying subtle patterns within multidimensional data, to enhance the predictive power of GI information. By leveraging ML techniques, it becomes feasible to create predictive models that account for both the GI of foods and individualized glycemic responses, enabling more tailored and effective dietary interventions (Gupta, 2024).

This research aims to address these gaps by developing a predictive model that leverages GI data in combination with ML algorithms to forecast T2DM risk and support dietary management. Integrating ML with GI data provides a twofold advantage: First, it enhances the ability to identify high-risk individuals and, second, it facilitates the generation of personalized dietary recommendations that could delay or prevent T2DM onset. This approach aligns with the goals of precision medicine and personalized nutrition, which seek to provide individualized care that optimally meets each patient's unique metabolic profile (Ordovas et al., 2018).

In summary, the convergence of GI as a dietary metric and ML as an analytical tool represents a promising frontier in T2DM prevention and management. By harnessing the predictive capabilities of ML, this study seeks to advance our understanding of how specific dietary factors contribute to T2DM risk and to develop data-driven interventions tailored to individual needs.

1.2. Problem Statement

The role of diet in the prevention and management of T2DM has been well-documented, yet conventional approaches to dietary recommendations often lack personalization and fail to account for individual variations in glycemic response (Lazarou et al., 2012). The GI has become an important tool for understanding how carbohydrate-rich foods impact blood glucose levels, allowing for classification of foods based on their effects on postprandial

(post-meal) glucose levels (Brouns et al., 2005). However, while the GI provides a foundational metric for evaluating food choices, it does not fully capture the complexity of individual responses to carbohydrate intake, which are influenced by genetic, metabolic, and lifestyle factors (Mendes-Soares et al., 2019; Zeevi et al., 2015).

Traditional statistical and epidemiological methods used for predicting T2DM onset are often limited in their ability to leverage detailed lifestyle, dietary, and demographic data, which can hinder their predictive accuracy (Kanagarathinam et al., 2024). ML techniques, however, are well-suited to handle such complex data and are increasingly applied in healthcare research due to their capacity to process large datasets, identify non-linear relationships, and generate actionable insights (Topol, 2019). Integrating GI data with ML models could significantly improve prediction of T2DM risk, as ML algorithms can analyze complex interactions between diet, lifestyle, and demographic factors in a way that traditional methods cannot (Bizimana, 2024)

The present study seeks to bridge this gap by developing predictive models that utilize GI data as a key variable, combined with other health and lifestyle data, to enhance the accuracy of T2DM risk predictions and offer more personalized dietary recommendations. By employing ML techniques, this research aims to transform GI data into tailored dietary guidance that reflects individual glycemic responses, ultimately supporting more effective T2DM prevention and management strategies.

1.2. Research Objectives

The primary objective of this study is to develop a predictive model that leverages GI data and ML algorithms to forecast the risk of Type 2 diabetes onset and support personalized dietary management. Specifically, this study aims to achieve the following objectives:

1. **To examine the relationship between Glycemic Index data and the onset of Type 2 diabetes**, focusing on how GI impacts glycemic control and associated risk factors.
2. **To apply machine learning techniques to develop a predictive model that utilizes GI data alongside demographic and lifestyle information** to forecast T2DM risk.
3. **To assess the accuracy and effectiveness of the model in identifying high-risk individuals** across various demographic groups.
4. **To formulate personalized dietary recommendations based on model predictions**, tailored to individual metabolic responses to aid in T2DM management and prevention.

1.3. Research Questions

This study is guided by the following research questions:

1. What is the relationship between Glycemic Index values of consumed foods and the risk of developing Type 2 diabetes?
2. How effectively can machine learning algorithms utilize GI data to predict T2DM onset across different population demographics?
3. What is the accuracy of the predictive model in identifying high-risk individuals based on GI data?
4. To what extent can personalized dietary recommendations based on GI-informed predictive models improve diabetes management and prevention?

1.4. Significance of the Study

This study has implications for healthcare providers, dietitians, individuals at risk of T2DM, and the research community. By integrating GI data into predictive models, this research aims to deliver tailored dietary recommendations that align with each individual's metabolic responses, thereby enhancing both preventive and therapeutic approaches to T2DM. For healthcare providers and dietitians, these models could serve as a decision-support tool, allowing for more targeted and effective dietary advice (Zhang, 2019).

Furthermore, this study contributes to the emerging field of personalized nutrition, which focuses on customizing dietary interventions based on individual characteristics to optimize health outcomes. For individuals at risk of T2DM, the predictive model developed in this study could provide actionable insights into food choices and dietary patterns that are best suited to their unique glycemic response profiles (Ordovas et al., 2018). Additionally, this research enriches the scientific literature on T2DM prevention by integrating ML and GI data, highlighting the benefits of combining dietary metrics with advanced analytical techniques in predictive health modeling.

1.5. Scope and Limitations of the Study

The scope of this study centers on the application of GI data and ML techniques to predict T2DM onset and inform personalized dietary recommendations. The study will focus on adult populations at risk for T2DM, with emphasis on assessing the impact of GI on glycemic response and T2DM risk. Data sources will include established GI datasets, demographic data, and lifestyle information such as dietary habits and physical activity levels.

The study is subject to several limitations. First, the general GI values for specific foods may not account for individual variability due to genetic, metabolic, and lifestyle factors, which can lead to variations in glycemic response (Zeevi et al., 2015). Furthermore, ML models typically require extensive, high-quality data to achieve high accuracy, and the predictive

power of the model may be limited by the data quality and quantity available(Van Calster et al., 2019). Additionally, focusing solely on GI as a dietary measure may exclude other relevant dietary metrics, such as glycemic load and dietary fiber content, which can also influence glycemic response and T2DM risk (Brouns et al., 2005).

1.6. Methodology

This study will adopt a quantitative research approach, combining data analysis, ML model development, and model validation to investigate the utility of GI data for predicting T2DM risk and formulating personalized dietary recommendations. The methodology encompasses the following stages, including data collection, preprocessing, model development, evaluation, and testing, specifically focusing on data from a Palestinian adult population.

The first step is **data collection**. The primary dataset will consist of information on Palestinian adults aged 18 to 64, collected in 2019, and will include an array of variables such as nutrition, lifestyle, and sociodemographic characteristics that are relevant to T2DM risk. Dietary data focus on the GI values of foods consumed, meal frequency, and nutrient composition, while lifestyle data capture physical activity levels, sedentary behavior, smoking status, and alcohol consumption. Sociodemographic factors will include age, gender, income level, education, and employment status. Additionally, anthropometric data such as Body Mass Index and health metrics like blood pressure and cholesterol levels will be incorporated to provide a comprehensive view of each participant's health profile. By leveraging these multifaceted data points, the study aims to develop a model that accurately reflects the interplay of diet, lifestyle, and sociodemographic factors in predicting T2DM risk.

Once data collection is complete, **data preprocessing and cleaning** will be undertaken to prepare the dataset for ML analysis. During this stage, data will be checked for any inconsistencies, duplicates, or missing values, with appropriate scaling and imputation methods applied where necessary. Feature engineering will be performed to structure the data into suitable formats for ML algorithms. This includes transforming continuous variables into categorical data where relevant, creating new variables to capture interactions between factors (such as the combined impact of GI and physical activity on blood glucose levels), and categorizing foods based on their GI values. This preprocessing step ensures that the data is organized, comprehensive, and reflective of the key factors influencing glycemic response, which are critical for accurate model training and testing.

The **model development** phase will involve experimenting with various ML algorithms to determine the most effective approach for predicting T2DM onset. Algorithms such as Random Forest, Gradient Boosting, and SVM, will be tested, as they have shown efficacy in healthcare data analysis due to their ability to handle non-linear relationships and complex interactions among variables. These algorithms are well-suited for managing multi-dimensional datasets that include both categorical and continuous variables, as is the case with the GI, lifestyle, and sociodemographic data in this study (Topol, 2019). The model will be developed, trained, and fine-tuned on the Palestinian adult dataset, aiming to capture predictive patterns specific to the sociodemographic context and dietary habits of this population.

Following model development, **model evaluation** will be conducted using multiple performance metrics, including accuracy, precision, recall, and the Area Under the Receiver Operating Characteristic curve. These metrics provide insights into the model's predictive power and its effectiveness in identifying individuals at high risk of T2DM. Cross-validation techniques will be applied to ensure that the model is robust and can generalize well across different subsets of the dataset. By employing cross-validation, the study ensures that the model is not overfitted to the training data and that it maintains reliability when applied to new data within the same population(Van Calster et al., 2019).

Once an optimal predictive model has been developed, **personalized dietary recommendations** will be generated based on the model's outputs. This stage involves translating model predictions into actionable dietary guidance that aligns with individual glycemic responses. Recommendations will prioritize low-GI food choices and suggest modifications to dietary patterns that could improve glycemic control, focusing on each individual's unique combination of dietary, lifestyle, and sociodemographic factors. For instance, recommendations may suggest substitutions of high-GI foods with low-GI alternatives tailored to an individual's age, BMI, and physical activity level. This approach aims to provide dietary interventions that are personalized, culturally relevant, and grounded in predictive insights from the model.

The final phase of the methodology is **validation and testing** of the generated dietary recommendations through simulation studies or pilot trials. In this phase, the effectiveness of the recommendations will be assessed by comparing outcomes with those achieved using existing dietary management practices. This comparison will help determine whether personalized recommendations based on ML-driven insights can lead to better glycemic outcomes and lower T2DM risk among Palestinian adults. If successful, this approach could highlight the potential of using ML and GI data to create effective, culturally tailored dietary guidelines for at-risk populations. By rigorously validating the model in a real-world setting, this study seeks to contribute to the field of precision nutrition and improve T2DM risk prediction and management within specific demographic contexts.

1.7. Organization of the Thesis

This thesis is organized into five chapters. Chapter One introduces the research, including the background, problem statement, objectives, research questions, significance, scope, limitations, and methodology of the study. Chapter Two presents a comprehensive review of the literature on T2DM, the GI, and the application of ML in predictive health modeling. Chapter Three outlines the methodology, covering data collection, data preprocessing, model development, and evaluation procedures. Chapter Four presents the results. Finally, Chapter Five discusses the findings in relation to existing literature, provides conclusions, implications, and recommendations for future research.

Chapter Two

Background and Literature Review

This chapter addresses T2DM, a growing global health issue influenced by lifestyle factors, especially diet and the GI. It explores how machine learning can enhance predictive models by integrating GI and other health data. The goal of the study is to develop a model that improves T2DM risk prediction and supports personalized dietary management, advancing precision medicine.

2.1. Background

T2D is a chronic and complex metabolic disorder characterized by insulin resistance and a gradual decline in insulin production. Unlike type 1 diabetes, primarily an autoimmune condition, T2D is largely influenced by a combination of genetic, environmental, and lifestyle factors. As a result, T2D is a progressive disease with a gradual increase in blood glucose levels, which, if not managed, can lead to severe complications affecting the cardiovascular, renal, nervous, and ocular systems.

The prevalence of T2D has escalated globally, significantly impacting public health, healthcare resources, and individual quality of life. The International Diabetes Federation estimates that 463 million adults were living with diabetes worldwide in 2019, with a projected rise to 700 million by 2045 if current trends continue (International Diabetes Feder Atlas, 2019).This chapter explores T2D's pathophysiology, risk factors, and the role of dietary GI, lifestyle, nutrition, and socioeconomic determinants. A comprehensive review of the literature supports the complex and multifactorial nature of T2D, providing the foundation for further research and preventive strategies.

2.1.1. Type 2 Diabetes: Pathophysiology and Epidemiology

The pathophysiology of T2D is marked by two primary mechanisms: insulin resistance and progressive beta-cell dysfunction. Insulin resistance occurs when the body's cells, particularly in the liver, muscle, and adipose tissues, become less responsive to insulin, the hormone that facilitates glucose uptake into cells. The pancreas compensates by increasing insulin production, but this compensatory response becomes insufficient as beta cells in the pancreas progressively fail, leading to hyperglycemia (DeFronzo, 2004).

T2D is a progressive disease. In its early stages, many individuals develop prediabetes, a condition with elevated blood glucose levels that fall short of T2D diagnostic criteria. Without intervention, prediabetes often progresses to T2D, underscoring the importance of early detection and preventive measures (American Diabetes Association, 2021).

The global increase in T2D prevalence is attributed to rising obesity rates, physical inactivity, and aging populations. Studies have shown that T2D is increasing rapidly in low- and middle-income countries, where rapid urbanization and lifestyle changes are taking place (Hu, 2011). These trends emphasize the need for in-depth understanding and intervention to control the growing diabetes epidemic.

2.1.2. Risk Factors for Type 2 Diabetes: Genetic and Environmental Interactions

2.1.2.1. Genetic Susceptibility

Genetic predisposition plays a substantial role in T2D development. Family history is a significant risk factor, with those who have a first-degree relative with T2D at a higher risk of developing the disease themselves (Lyssenko et al., 2005). Research has identified multiple genetic loci associated with T2D, particularly those related to insulin secretion and beta-cell function, such as the TCF7L2 gene (Florez, 2008). However, genetic factors alone do not guarantee TD development, emphasizing the critical influence of environmental and lifestyle factors that modulate genetic risk through gene-environment interactions.

2.1.2.2. Environmental and Lifestyle Factors

Environmental and lifestyle factors are powerful determinants in the onset of T2D. Obesity, especially visceral fat accumulation, is among the strongest risk factors, as it contributes to increased insulin resistance and chronic inflammation (Kahn et al., 2006). Physical inactivity compounds this risk, whereas regular exercise improves insulin sensitivity and supports weight management (Colberg et al., 2010). Poor dietary habits, such as high intake of refined sugars and processed foods, further contribute to T2D, while diets rich in fiber, whole grains, and plant-based foods offer protective effects (Ley et al., 2014). Chronic stress, inadequate sleep, and mental health disorders also influence T2D risk by disrupting hormonal balance and promoting insulin resistance (Reutrakul & Van Cauter, 2018).

2.1.2.3. Glycemic Index and Its Implications in Type 2 Diabetes Management

The GI measures how carbohydrates in food raise blood glucose levels. Foods are categorized as low-, medium-, or high-GI based on their impact on blood glucose after consumption. Low-GI foods, such as whole grains and legumes, lead to a gradual increase in blood glucose, whereas high-GI foods, including sugary snacks and processed grains, cause rapid spikes (D. J. A. Jenkins et al., 2002).

Low-GI diets have been shown to improve glycemic control, reduce HbA1c levels, and decrease insulin resistance in individuals with T2D. A significant study by (Salmerón et al., 1997) revealed that individuals consuming high-GI diets had a greater risk of developing T2D. Low-GI diets have also been associated with reduced triglycerides and improved lipid profiles, emphasizing their importance in overall cardiovascular health for individuals with T2D (Barclay et al., 2008). Integrating low-GI foods into a balanced diet provides a valuable approach to T2D management, highlighting the importance of carbohydrate quality in dietary recommendations.

2.1.2.4. Lifestyle Interventions and Their Role in Diabetes Prevention and Management

Lifestyle modifications, including physical activity, stress management, and improved sleep, are foundational to preventing and managing T2D. Regular physical activity improves insulin sensitivity, promotes glucose uptake by muscles, and aids in weight control. The Diabetes Prevention Program (DPP) demonstrated that lifestyle interventions involving physical activity and dietary changes reduced T2D incidence by 58% among high-risk individuals (Knowler et al., 2002).

Sleep quality and duration are also critical. Poor sleep disrupts glucose metabolism, elevates cortisol levels, and increases insulin resistance, all of which contribute to higher T2D risk (Knutson, 2006). Research has shown that individuals who sleep less than six hours per night or experience frequent sleep disruptions are at a higher risk of T2D, highlighting the need for sleep hygiene as part of T2D prevention programs (Reutrakul & Van Cauter, 2018).

Furthermore, chronic stress contributes to T2D risk by elevating stress hormones such as cortisol, which promotes gluconeogenesis and raises blood glucose levels (Lundberg, 2002). Interventions such as mindfulness, yoga, and behavioral therapy have been shown to reduce stress and improve glycemic control, emphasizing the potential of stress management in T2D care (Katherine A McGonagle et al., 2012).

2.2. Nutritional Factors in Type 2 Diabetes Prevention and Management

Nutrition plays a critical role in the prevention and management of T2D, with dietary choices significantly influencing risk factors and disease progression. Specific nutrients and food groups are essential for maintaining glucose metabolism, improving insulin sensitivity, and reducing inflammation. Below are key nutritional factors and their impact on T2D prevention and management:

- **Dietary Fiber**

High dietary fiber intake, particularly from whole grains, legumes, and vegetables, has been associated with a lower risk of T2D. Fiber slows glucose absorption, reducing post-meal blood glucose spikes and improving insulin sensitivity. Prospective studies suggest that high fiber intake reduces T2D risk by 20-30% (Schulze et al., 2007).

- **Protein Sources**

Plant-based proteins, found in legumes, nuts, and seeds, have shown protective effects against T2D due to their high fiber content and low saturated fat levels (Pan et al., 2011). In contrast, high consumption of red and processed meats is associated with increased T2D risk, potentially due to their saturated fat and preservative content, which promote inflammation and oxidative stress (Micha et al., 2010).

- **Healthy Fats**

Healthy fats, including polyunsaturated and monounsaturated fats found in olive oil, avocados, and fatty fish, improve insulin sensitivity and reduce inflammation. Omega-3 fatty acids, in particular, offer cardiovascular benefits and have been associated with reduced insulin resistance (Freeman, 2010). Including these fats in the diet is recommended for individuals with T2D to support cardiovascular health and metabolic stability.

- **Socioeconomic Factors and Diabetes Disparities**

Socioeconomic status influences T2D risk and management outcomes. Lower SES is linked to a higher prevalence of T2D, often due to limited access to healthcare, healthy foods, and exercise facilities (Agardh et al., 2011). Individuals with lower SES may face greater obstacles in obtaining high-quality, nutritious foods, increasing their reliance on low-cost, calorie-dense, and nutrient-poor options that can elevate T2D risk. Financial limitations also impact access to healthcare services and medications, exacerbating diabetes-related complications and mortality rates (D. Kim & Saada, 2013).

Stress associated with financial strain, inadequate housing, and job insecurity further worsens glycemic control and insulin resistance (Walker et al., 2012). Addressing these disparities requires public health strategies aimed at improving access to healthcare, promoting health literacy, and providing economic support for healthier food options in low-income communities. Studies show that initiatives targeting socioeconomic barriers can improve T2D outcomes and reduce incidence, emphasizing the need for equitable healthcare policies (Marmot et al., 2008).

2.3. Overview of ML Algorithms

In recent years, as a result of rapid improvements in the world of technology, the use of advanced data analysis techniques has increased in many areas. These techniques help to develop impressive decisions with huge amounts of data. In order to increase the efficiency of data analysis, different learning strategies and algorithms are developed and used. The type of data being analyzed and the results desired from the data are important factors for the selection and applicability of the learning algorithm to be used.

ML models continue to develop with new and advanced data learning models in the rapidly growing technology industry. These models aim to assess massive amounts of data in the shortest possible time, to provide decision-makers with timely results and to contribute to developing strategies. These models can be used regardless of industrial areas and can be used to analyze any type of data where problems can be solved by creating learning algorithms. In this study, some of the many ML algorithms are theoretically introduced to compare the performance of the developed models and to train the model with the right dataset, to choose the right learning algorithm that provides the right results in analyzing the data for developing prediction and classification models.

2.3.1. Supervised Learning

In supervised learning, a mathematical model is built to predict an output based on one or more inputs. The term "supervised" refers to the fact that the model-building process is guided by a high-quality output dataset, which serves as the foundation for the mapping process. The training data consists of a set of features and a supervised output label. With binary or multi-class dependent variables, the model algorithm predicts the probability of one of the classes. Examples of supervised learning include classification by category, prediction by conditional probability, and regression analysis, in which real numerical values are predicted. With classification, the model predicts the probability of the dependent binary variable. In classification, we seek a value like 1 or 0, yes or no, true or false, etc. These are examples of binary or two-class classification problems. When there are three or more output options, it is called a multi-classification problem. Examples of multi-class outcomes include demographic grouping, product purchase, and species type.

Algorithms that classify or predict a dependent binary variable are called classifiers. Classifiers have single or multiple categorical labels as output dependent variables. The classification model can accurately identify the class of an entity. It is one of the widely used and easy to understand models; it contains features that map to an entity and tries to classify entities. The categories of classifiers include supervised learning and unsupervised learning. When labeled data is given to the model, it becomes a supervised model. In the case of classification, the model assigns each entity to one of the categories. The logistic regression model is one of the examples used for binary classification. The classification of an entity into a class is simply computing given different classes and selecting the one with the highest probability score as its output.

2.3.2. Classification Algorithms

In this chapter, basic information on the classification algorithms SVM, RF, Logistic Regression, Classification and Regression Trees, and GB is given. They are widely used in classification studies. The main aim of these mechanisms is to predict the class labels or target values of future datasets on the basis of the trained datasets. Given a training dataset, where x_i is the i^{th} data, y_i is the class labels for a sample, $i = 1, \dots \dots, n$, and n is the sample size. The primary stage of classification algorithms is to train the dataset in order to construct a method. These training samples are used to train the datasets; afterwards, both the built algorithms and the test samples are evaluated.

2.3.3. Importance of Algorithm Understanding

As data science continues to evolve, algorithms play an increasingly vital role in data analysis and decision-making processes. While practitioners regularly use ML algorithms to analyze data, most available literature focuses on practical applications without adequately addressing the theoretical foundations that explain why these algorithms succeed or fail in different scenarios.

Our research addresses this knowledge gap through a rigorous examination of the mathematical principles underlying prevalent algorithms, including RF, SVM, LR, GB, and CART. The mathematical framework governing model performance provides practitioners with essential theoretical insights while maintaining practical applicability for model development.

Through systematic programming demonstrations and empirical examples, this study illuminates both the capabilities and constraints of these approaches. Such comprehensive understanding enables practitioners to make methodologically sound decisions in algorithm selection and optimization for specific applications, while developing robust strategies to address model limitations.

The research contributes to the field by establishing clear connections between theoretical foundations and practical implementations of ML, thereby enhancing practitioners' capacity to interpret and optimize their models through a theoretically grounded approach.

2.4. Logistic Regression

Logistic regression is a classification algorithm commonly used to predict class labels or target values based on a training dataset. It is widely used in various applications such as medical diagnosis, financial predictions, and marketing. Given a training dataset, where X_i represents the input data point and Y_i is the corresponding class label, the goal is to train a model that can predict the class labels of new, unseen data points. The logistic function provides the probability that a given input X belongs to a particular class, making it ideal for binary classification problems.

In practice, logistic regression works by finding the optimal coefficients β_j for the features X_i that minimize the difference between the predicted probabilities and the actual class labels. The model then uses these coefficients to predict the class probabilities for new input data.

2.4.1. Mathematical Foundation

The mathematical foundation of logistic regression is rooted in probability theory. Specifically, it is based on the concept of odds and the logistic function, which transforms a linear combination of input features into a probability between 0 and 1. This transformation is crucial for modeling binary outcomes.

The key concept is the log-odds, which is the logarithm of the odds of an event occurring. In logistic regression, the log-odds of the dependent variable Y given the input vector X is modeled as a linear function of the input features:

$$logit\big(Pr(Y = 1|X)\big) = \beta_0 + \sum_{j=1}^{p} \beta_j X_j \quad \dots (2.1)$$

Where β_0 is the intercept term, β_j are the coefficients for each input feature, X_j are the input features. The logistic function, applied to the log-odds, gives the predicted probability:

$$Pr(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_j)}} \quad \dots (2.2)$$

Where X is a P- dimensional input vector and $Pr(Y = 1|X)$ is the probability of the instance Y given X. Since $0 \leq \Pr\,(Y = 1|X) \leq 1$, the predicted probability can be easily converted into class labels: if $Pr(Y = 1|X) > 0.5$, the predicted class label of the instance X is 1; otherwise, the prediction is 0.

For multi-class classification problems, a simple approach is to model a series of binary classifications using the one-vs-rest method. In logistic regression, the probability of $Pr(Y = i|X)$ for class i is modeled by regressing $Pr(Y = i|X)$ and $Pr(Y \neq i|X)$, where $Pr(Y \neq i|X) = 1 - Pr(Y = i|X)$. However, the one-vs-rest scheme may ignore the interdependence between binary classifiers. A more complex structure for multi-class problems is the one-vs-one method, where a binary classifier is built for every pair of classes. Alternatively, a complete scheme can be used to model label relations more efficiently.

Maximum likelihood estimation (MLE) is commonly used to estimate the coefficients β_j, as it maximizes the likelihood of observing the given data. This process involves minimizing a loss function that represents the difference between predicted probabilities and actual labels in the training dataset.

2.4.2. Advantages and Limitations

Logistic regression offers several advantages, including its simplicity and ease of interpretation, making it a popular choice for many classification tasks. It is computationally efficient, particularly for smaller datasets or when the number of features is relatively low. Additionally, logistic regression provides probabilities as outputs, which can be valuable in applications where understanding the uncertainty of predictions is important, such as medical diagnoses or risk assessments. However, logistic regression also has limitations. It assumes a linear relationship between the input features and the log-odds of the outcome, which may not always be true for real-world data. It is sensitive to outliers and may perform poorly if the data contains extreme values. The model can also struggle with highly non-linear relationships unless feature engineering is used to capture such non-linearities. Furthermore, for multi-class classification problems, logistic regression requires adaptations like the one-vs-rest or one-vs-one approaches, which can complicate the model and ignore interdependencies between classes.

2.5. Decision Trees

A DT is one of the popular and practical methods for solving decision-making problems in ML. A DT is a collection of connected branches and nodes. It can be used to classify various classes in categorical target variables; it makes the decision about classifying each observation. The DT algorithm makes it possible to handle both classification and regression problems. Certain algorithms do not have these advantages.

Using a single tree, a suitable and simple algorithm is available. However, executing the tree method to make a selected number of trees is not as simple as this; by using default cutoff values, overfitting leads to predicting poor accuracy with a brand-new independent sample. To handle this issue, it is recommended to adjust some parameters or grow a full tree and prune afterwards. In fact, building a large number of trees one by one and calculating the majority vote of all trees is an appealing return. It is also becoming more common in practice with recent developments in ML. More precise predicted responses are obtained, and hence flexibility in representation is increased using lots of trees. In addition to providing dimension dropdown, CART use tree-based methods to call the class over a variety of classes for input.

2.5.1. Basic Concepts of Decision Trees

The basic concept of decision trees is to make decisions to solve a problem using a binary structure. The aim here is to create the classification rules that give the most accurate results about the target variable using the smallest number of questions. The foundation of the decision tree is to make the questions one by one according to the answers given. When we classify disordered data, we start from the top of the tree and go to each point using the arrows and finally divide the data as perfectly as possible. Each question used in the process is called a node. As a result, the criteria value for the classification that is common to all data types is reached. In this case, the result corresponds to the answer for the question set. The questions at the nodes can be formed on a given threshold in the designs and numerical variables.

The first question is placed at the top for all the training data. While the data is moving between the nodes, they are distributed according to the values of the variables that are different in the cutoff. The class separation of the nodes is carried out according to the highest Gini index. Since the questions at the first node represent the first rule, the independence of the undefined variables is assumed. Consequently, the columns disordered in structure are put in a structure by the rules of the classification. The sections represent the last nodes; capacity refers to the number of each class in the class of actual notes. The result of each question determines the size and length of the branch of that partition. Various leaf node rules are created in decision trees. In this context, the Gini value gained if different decision branches are selected and added to the node can be investigated. The results of the indices measured for node, section, and the decision steps made against each other may differ. As a result, the maximum discrimination and the lowest classification error may occur. For these reasons, other node management criteria have been developed. These are the measures of Gini, maximum information gain, maximum gain ratio, and the minimum error. In this case, the current training state is considered to have the most optimal combination of node structure properties and the most accurate results. Gini, Shannon

entropy, and information are the most selected algorithms. Decision tree architecture that includes these qualitative and quantitative measure alternatives also provides a model advantage among the other models. Medium-high levels of model success oriented make the classification possible.

2.5.2. Classification and Regression Trees (CART)

CART, were introduced to simplify a large decision tree. Generally, it is a tree model that predicts the value of a target variable by learning rules from the features. It is a prediction method and can be easily handled as both a binary tree and a multiclass case. Trees take multiple paths to a target by asking a question. The success of a classification is measured by the accuracy of the 0–1 loss function. CART uses Gini impurity as an impurity measure. CART prefers to use continuous variables where it will get splitters by checking each unique value. It checks the smaller value for each unique variable and selects the one with lower impurity. It proceeds until it finds the best split.

2.5.3. Mathematical Foundation of CART

The mathematical foundation of CART is based on the principle of recursively partitioning the data to achieve the best possible classification or regression model. CART builds decision trees using a binary splitting method, where at each node, the data is divided into two groups based on a decision rule derived from a feature and its corresponding threshold value. The primary objective is to create partitions that are as homogeneous as possible with respect to the target variable, either for classification or regression tasks.

For classification, the most commonly used criterion for splitting is the Gini index. This is a measure of impurity that quantifies how often a randomly chosen element from the node would be incorrectly classified. The Gini index for a node is calculated as:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{c} p_i^2 \quad \ldots (2.3)$$

Where p_i represents the proportion of samples in class i at node t, and C is the total number of classes. The splitting criterion chooses the feature and threshold that minimizes the Gini index, ensuring the highest possible homogeneity in the resulting child nodes.

2.6. Random Forest Algorithm

Compile a forest of such decision trees, or 'random forest,' to define a final point classification by combining the classification of each individual tree divided by maximal tree similarity. More general trees deal with occasionally omitted attributes. RF creates a set of decision trees from random and redundant samples of the input pattern and selects optimal attributes. The final decision is made by all unbiased decision trees so that the combined decision converges to judging the common domain of the internal tree. The ensemble method

is expected to show more accuracy compared to a single individual classifier. For output selection and regression model specification, the RF is based on ensemble methods. By reducing variance with insignificant bias, both ensemble learning methods try to stabilize the outcome. RF uses bootstrap sampling and aggregates error by averaging in the function after minimal data transformation. Small margins of the classification threshold area increase the predicted output. RF solve the most common problems of overfitting, including the curse of dimensionality, which both lead to high model variance when encountering polynomial regression.

2.6.1. Ensemble Learning and Bagging

The Random Forests algorithm is an ensemble learning method that can be used for both classification and regression problems. It is one of the most accurate and robust supervised learning methods. A RF is considered an ensemble because it consists of many decision trees. These trees are combined, and the combination frequently reduces the overall classifier error. However, while an ensemble frequently means learning a set of decision trees or even deep learning models, the general goal of ensemble learning is to develop a better performance classifier than the performance of the classifier that each tree can learn.

Bagging is a technique used to produce an ensemble using a collection of models. It is a simple technique to combine predictions from a group of models. The ensemble's results are produced by averaging the predictions rather than taking the mode, as is the case in boosting. Bagging is an acronym that stands for bootstrap sampling. It allows the classifiers to sample as many events as they want with replacement. Bootstrapping means that the classifier takes an observation, puts it back, and takes in the next observation, and so on, until all of the individuals produce a dataset for each of the respondents in the training data.

2.6.2. How Random Forests Work

In a RF, the algorithm creates an ensemble of decision trees, which are grown using a variant of the CART algorithm. The key idea behind a RF is to randomly select a subset of features, denoted as $mtry$, at each node of the tree and then choose the best split among these $mtry$ predictors. The $mtry$ parameter is one of the model's tuning parameters and must be specified before building the RF model. Typically, $mtry$ is set between 1 and p (where p is the total number of predictors in the dataset).

For classification problems, the default value of $mtry$ is the square root of p. For regression problems, the default value of $mtry$ is $p/3$. The RF algorithm also uses bootstrapping to generate different samples of the data, which results in trees with different structures, reducing the correlation between the trees' predictions. A large number of trees must be grown to achieve reliable and robust results. Additionally, the maximum tree depth is calculated, and splits are made in such a way that most of the data is concentrated in the terminal nodes of each tree. The Gini index is used to measure the impurity of each node, helping to assess the contribution of each feature at each split.

To grow a RF, the algorithm starts by taking a bootstrapped random sample of the data. From this sample, the root node of the decision tree is created by selecting $mtry$ random features.

For each subsequent node, another set of $mtry$ features are randomly chosen to determine the best split. Child nodes are created based on the splits made at each parent node, using the CART algorithm. This process continues until the tree has fully grown, potentially covering the entire input space.

For classification tasks, the final prediction is made by a majority vote among the trees, with the class predicted by the majority of trees being selected as the final output. For regression tasks, the final prediction is the average of the predictions made by each tree.

2.6.3. Mathematical Foundation of Random Forests

It is not hard to prove that decision trees are vulnerable to overfitting problems. Hence, a single decision tree model does not provide the best generalization error. Random forests are good learners to apply to a wide range of data mining problems. In fact, the motivation of random forests originates from the variance of unpruned trees. Predictors that can explain the space in different ways with still reasonable predictive power, meaning they can classify similarly, combine well to improve upon classification problems. Elements that improve the problem are called weak learners, and a good combination action is called a strong learner.

RF can be thought of as combining a set of full and uncorrelated trees with reasonably good classification performances based on a bootstrap sampling method and the grow-tree sample sizes. Algorithms that are combined with ensemble techniques typically yield superior predictive outcomes. One key reason for this success is the prediction improvement in precision that comes from combining the weak learners.

There is a tradeoff between approximation bias and estimation bias in model performance when numerous component models are combined to create a super learner. Bagging, or bootstrap aggregating, for RF are ensemble algorithm methods that can reduce overfitting of a specific base algorithm. Bagging constructs different datasets by bootstrapping, then trains the base algorithm on each dataset. The original algorithm randomly samples observations and cases to create super learners, reducing misclassifications.

The RF is produced based on majority voting for classification or averaging for regression by combining the outputs of these base algorithms. Mathematically, if we have N trees in the forest, where each tree T_i makes a prediction $\hat{y}_i$, the final prediction $\hat{y}_{RF}$ for regression is the average of the individual tree predictions:

$$\hat{y}_{RF} = \frac{1}{N}\sum_{i=1}^{N} \hat{y}_i \quad \dots (2.4)$$

For classification problems, the final prediction is typically determined by majority voting, where the most frequent class among the trees is selected:

$$\hat{y}_{RF} = mode(\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N) \quad \dots (2.5)$$

As long as the base algorithm performs well enough, it has been shown that the generalization capacity of a super learner-based ensemble grows with the number of models, thereby reducing overfitting and improving performance in noisy datasets.

2.7. Support Vector Machine (SVM)

SVMs are a class of supervised learning models used primarily for classification tasks, although they can also be applied to regression. SVMs aim to find the optimal hyperplane that best separates data points of different classes, maximizing the margin between them. This approach is particularly effective for high-dimensional data and provides robust generalization performance. SVMs can handle both linearly separable and non-linearly separable datasets, making them versatile for a wide range of applications. The key strengths of SVMs lie in their ability to handle complex data structures and their flexibility through the use of different kernel functions.

2.7.1. Linear SVMs

A linear SVM is used when the data is linearly separable, meaning that there exists a straight line (in two dimensions), or a hyperplane (in higher dimensions), that can perfectly separate the data points of the two classes. The goal of the linear SVM is to find this hyperplane such that it maximizes the margin between the closest data points from each class. The data points that are closest to the hyperplane are called support vectors, and they determine the optimal position of the hyperplane.

For a binary classification problem with two classes, $+1$ and -1, the optimal hyperplane is defined by the equation:

$$w^T x + b = 0 \quad \dots (2.6)$$

Where $w \in \mathbb{R}^m$ is the weight vector perpendicular to the hyperplane, and $b \in \mathbb{R}$ is the bias term, which helps shift the hyperplane.

The goal is to maximize the margin, which is the distance between the hyperplane and the closest points from either class. The margin is inversely proportional to $\|w\|$, and thus, the optimization problem is to minimize $\frac{1}{2}\|w\|^2$, subject to the constraint that all data points are correctly classified:

$$y_i(w^T x + b) \geq 1, \quad \forall i = 1, \dots, n \quad \dots (2.7)$$

Where $y_i \in \{-1, +1\}$ is the class label of the data point x_i, and n is the number of training samples.

2.7.2. Non-Linear SVMs

When the data is not linearly separable, a non-linear SVM can be used. In this case, a direct linear separation in the input space is not possible, but the kernel trick can be employed to map the data into a higher-dimensional feature space where a linear hyperplane can separate the classes.

The kernel trick allows SVMs to implicitly compute the dot product of data points in a higher-dimensional space without explicitly computing the transformation. This makes SVMs computationally efficient even for high-dimensional feature spaces. The SVM model then solves the optimization problem by maximizing the margin in the higher-dimensional space, using the kernel to compute inner products between data points. This approach is highly effective for complex, non-linear classification tasks.

SVM is a powerful technique for classification. However, its processing is highly dependent on the training set size and dimension. Regularization is simulated by comparing training with testing errors. SVM models with the kernel trick are popular and powerful techniques for classification. The kernel trick provides a way to handle the non-linear classification problem via optimization in a high-dimensional feature space using a pseudo input vector. Given a training set, the kernel trick adds ad-hoc basis functions to a linear model, extending the linear models to non-linear models. By the addition of an ad-hoc basis function, training error can be further reduced. In the high-dimensional feature space, the error rate may equal zero, but the significance of feature space construction is not reasonable. The hierarchical structure may lead to high computational costs and steering problems as in the original processing.

2.8. Gradient Boosting Algorithm

GB is a powerful and widely used algorithm in classification and regression problems, and it is an example of ensemble methods where the following trees are used to predict the response. Boosting is a sequential method, and the main objective is to reduce the bias of the prediction. GB uses gradient descent at each step, and it tries to minimize the loss function. It generates a large number of trees and combines them to create a strong predictive model. In this study, as a GB method, the most practical and commonly used implementation, XGBoost, is used. XGBoost improves upon the base GB multiple times and is used to train ML predictors.

XGBoost is an efficient implementation of the gradient boosted tree, and the objective function is the convex combination of the differentiable loss function along with the number of the leaf of the tree. In recent years, XGBoost has seen wide adoption across many data sets and many types of data. Despite the rough and solid improvement in the performance of XGBoost models, there are some problems in practice. The most critical problem is overfitting. The other one is that there is no additional interpretability capabilities in the models produced with XGBoost. In large-scale real-world problems, creating models with fast and successful implementation is possible with the integrated application of the late layer.

2.8.1. Boosting Techniques Overview

Boosting is a concept that can be considered a sort of portmanteau. From this combination of different models, a new and more efficient model can appear. This study intends to focus on the use of boosting in data mining. Boosting, a known method in ML literature, can be expressed with all classification algorithms. The models with perfect results were developed by the increase of classification enlargement first. These models that have perfect results when applied to a subset of data cannot provide the desired results when applied to another subset of data. This is a situation that occurs with the bagging method. Many models are trained on data separately, and these models can be used. Instead of training many models individually, a model can be trained with reference to the results of data used from other models. This newly trained model provides more efficient and useful results by referencing other models. With boosting, these models go through training and take advantage of such models.

2.8.2. How Gradient Boosting Works

In this section, how an algorithm called GB, which is an extension of AdaBoost, one of the well-known ensemble methods, and strengthens the weak learner, is explained in detail. The GB, the strong learner being formed, is mainly capable of being used in tasks like both estimation and classification. What the algorithm called GB operates on is the weak learner. In this algorithm, to create each weak learner, the information calculated through the previous weak learner is used for successively improving the best weak learner against the information that has not yet been learned well. The error calculated through the weak learner whose estimation is more exploited is to be subtracted from the original errors in the model, and the new errors, which are purer, are found in this way. A weak learner whose estimation is more faulty compared to the weak learner is to correct those new and purer errors remaining in the model. In this way, the weak learners are progressively nested against the original error in the model, and weaker learners in the model get less information on average compared to the weak learner that is added to the model more rapidly.

2.8.3. Mathematical Foundation of Gradient Boosting

The process begins by defining an objective function, usually a loss function $L(y, f(x))$, which quantifies the difference between the true target value y and the predicted value $f(x)$. In classification tasks, the most common loss function is cross-entropy.

Initially, the model is set to a simple prediction. In the case of regression, this is typically the mean of the target values across the dataset, while for classification, it might be the log-odds of the probability of the positive class. This starting model provides a base to which subsequent models are added.

In each iteration, a new model is trained to predict the residuals, or errors, of the current model. The residuals are the negative gradient of the loss function with respect to the predictions of the current model. Essentially, these residuals indicate the direction in which the model needs to be adjusted to reduce the loss. The new model $h_m(x)$ is trained to

approximate these residuals. Once the model is fitted, it is added to the current model $f_{m-1}(x)$, resulting in an updated model $f_m(x) = f_{m-1}(x) + vh_m(x)$, where v is a learning rate that controls how much the new model influences the overall prediction.

The process of adding models continues for M iterations, with each new model correcting the errors made by all previous models. This additive nature of the model allows GB to incrementally improve its performance. The final prediction is obtained by summing the contributions of all the individual models:

$$f_M(x) = f_0(x) + v\sum_{m=1}^{M} h_m(x) \quad \dots (2.8)$$

To avoid overfitting, regularization techniques are applied. One common form of regularization in GB is limiting the depth of individual decision trees, often too shallow trees, such as decision stumps (trees with only one split). Additionally, the learning rate v is typically kept small to ensure that each model's contribution is gradual. These regularization strategies help to improve the model's generalization capabilities and prevent it from becoming overly complex.

2.9. Literature Review

2.9.1. Introduction

T2DM is a critical public health issue that disproportionately affects the Middle East and North Africa region, including Palestine, where T2DM prevalence is steadily increasing (International Diabetes Feder Atlas, 2019 T2DM is a multifactorial disease linked to lifestyle, dietary habits, and genetic predispositions, and it poses significant challenges due to its association with severe complications like cardiovascular disease and nephropathy(Zoccali et al., 2023). Dietary interventions, specifically the quality of carbohydrate intake as measured by the GI, are increasingly studied for their potential role in T2DM prevention and management (Augustin et al., 2015). However, variability in individual responses to GI and the growing role of ML in precision health highlight the potential for data-driven, personalized nutrition approaches. This literature review examines recent research on GI and T2DM, with a focus on studies conducted in Palestine and the MENA region. It also explores the applications of ML in predicting T2DM risk, examining specific ML methods, models, and their efficacy across regional studies.

2.9.2. Glycemic Index and T2DM in the Middle East and Palestine

GI has become a focal point in T2DM research as it provides a measure for predicting the glycemic impact of carbohydrate-rich foods. Studies have demonstrated that low-GI diets are associated with improved glycemic control and reduced risk of T2DM (Foster-Powell et al., 2002). In the MENA region, several studies have examined the role of GI in T2DM risk and management. For instance, a study conducted in Saudi Arabia by (Astamadan et al., 2018) analyzed the effects of high- versus low-GI diets among adults with T2DM. The

researchers found that participants adhering to a low-GI diet experienced significant reductions in HbA1c and fasting blood glucose levels compared to those on a high-GI diet. The study highlights the potential of low-GI dietary patterns to mitigate T2DM progression in Middle Eastern populations, aligning with findings in other global studies.

In Palestine, research has focused on understanding the dietary habits of individuals with T2DM and exploring potential nutritional interventions. A recent study by (El Bilbeisi et al., 2017) surveyed the dietary intake of Palestinian adults with T2DM, finding that many patients had diets high in refined carbohydrates and low in fiber. The study emphasized the need for culturally relevant dietary interventions to reduce high-GI food consumption and improve overall dietary quality. Although specific GI-related interventions are less documented in Palestine, the study provides essential insights into the dietary patterns that may contribute to T2DM risk within this population.

2.9.3. Dietary Patterns, Lifestyle, and T2DM Risk in the Region

In addition to GI, overall dietary patterns and lifestyle factors are critical components of T2DM risk in the MENA region, where high-calorie diets and sedentary lifestyles are increasingly common. A systematic review by (Boushey et al., 2020) on dietary patterns and T2DM in Arab countries found a strong association between diets rich in refined carbohydrates, sugars, and saturated fats with increased T2DM prevalence. The review also highlighted that traditional diets rich in fiber and whole grains are associated with a lower risk of T2DM, suggesting that dietary habits in Arab countries have shifted toward Westernized patterns, which may be driving the rise in T2DM cases.

Physical activity, or the lack thereof, is another important factor. A cross-sectional study conducted in Jordan by (Abu-Mweis et al., 2014) examined the relationship between physical activity levels and T2DM risk among adults. The study found that participants with higher physical activity levels had a 32% lower risk of T2DM, highlighting the protective role of exercise in glycemic control. This aligns with findings from other regional studies that emphasize the combined effect of diet and lifestyle on T2DM outcomes.

Socioeconomic factors also play a role in T2DM risk across the MENA region. In Palestine, (N. M. E. Abu-Rmeileh et al., 2013) found that lower income and education levels were associated with higher T2DM prevalence. This was attributed to limited access to healthcare, reduced awareness of healthy dietary practices, and economic barriers to healthy food. These findings highlight the need for T2DM interventions that are accessible, affordable, and culturally tailored.

2.9.4. Individual Variability in Glycemic Response and Regional Insights

Studies in the MENA region have also shown significant variability in glycemic responses to identical foods, underscoring the limitations of generalized dietary guidelines. Research has suggested that genetics, lifestyle, and gut microbiota may influence individual responses to GI, highlighting the potential for personalized approaches (Zeevi et al., 2015).

In the UAE, a study by (Al Dhaheri et al., 2015) examined the influence of genetics on glycemic response in Arab populations. The study found that specific genetic markers associated with glucose metabolism were prevalent in Arab populations, influencing

glycemic responses to high-GI foods. These findings suggest that genetic profiling could improve the effectiveness of dietary recommendations in Arab populations, offering a potential route for personalized dietary advice.

Furthermore, in a Palestinian cohort, (Zalan & Sharkia, 2019)explored the relationship between microbiota composition and glycemic control. The study revealed that specific bacterial profiles correlated with lower postprandial blood glucose levels, suggesting that gut microbiota may play a role in glycemic variability among Palestinians. Such findings support the potential of personalized nutrition in T2DM management by considering both GI and individual biological factors.

2.9.5. Machine Learning in T2DM Prediction and Management in the Region

ML has become an essential tool in T2DM research, with numerous applications in risk prediction and personalized dietary recommendations. ML algorithms can analyze large datasets, allowing for the identification of complex patterns among dietary, genetic, and lifestyle factors. In recent years, several studies in the MENA region have demonstrated the utility of ML in predicting T2DM onset and personalizing dietary recommendations.

A notable study in Saudi Arabia by (Almutairi & Abbod, 2023) utilized SVM and RF to predict T2DM risk based on dietary, demographic, and lifestyle data. The RF model achieved an accuracy of 87%, while the SVM model achieved 83% accuracy. The study concluded that ML models could offer reliable predictions for T2DM risk, emphasizing the potential for integrating GI and other dietary metrics to enhance model precision.

In Jordan, (Hatmal et al., 2020) applied Artificial Neural Networks to a dataset of Jordanian adults, incorporating variables such as GI, physical activity, and BMI. The ANN model achieved an accuracy of 89% in predicting T2DM risk, outperforming traditional statistical models. This study highlights the potential of ML, specifically neural networks, in predicting complex diseases like T2DM where multiple factors interact.

Additionally, a study conducted in Egypt by (Mousa et al., 2023) used DT and LR to assess the impact of dietary patterns, including GI, on T2DM risk. The DT model achieved 82% accuracy, while LR achieved 78%. (Mousa et al., 2023) found that incorporating dietary variables, such as GI and GL, improved model performance, supporting the value of dietary metrics in T2DM risk assessment.

In Palestine, emerging research is exploring the application of ML in T2DM prediction. A study by (N. M. Abu-Rmeileh et al., 2012) utilized LR and DT algorithms to analyze a dataset of Palestinian adults, incorporating GI, dietary patterns, and lifestyle factors. The DT model achieved an accuracy of 80%, demonstrating the feasibility of using local data to develop predictive models for T2DM. The study emphasizes the importance of culturally relevant data in improving model accuracy and highlights the need for further research in Palestinian populations.

2.9.6. Integrating GI, ML, and Precision Nutrition for T2DM Management

The integration of GI, ML, and precision nutrition principles represents a promising approach for T2DM management, especially in culturally diverse regions like the MENA. Precision nutrition, which involves tailoring dietary recommendations to individual needs based on factors like genetics, gut microbiota, and lifestyle, is increasingly recognized as a viable solution to the limitations of generalized dietary advice (Ordovas et al., 2018). By incorporating GI data into ML models, researchers can develop dietary recommendations that align with individual metabolic profiles, potentially leading to more effective T2DM management.

A recent study by (Varshney et al., 2023) explored the utility of ML-driven personalized dietary recommendations in managing T2DM among high-risk individuals. The study used a GB machine model to generate personalized dietary advice based on GI, GL, and individual glycemic responses, achieving 90% accuracy in identifying T2DM risk profiles. This study shows the potential of ML in delivering personalized dietary interventions that improve adherence and outcomes in T2DM management.

Building on these principles, (Chatelan et al., 2019) proposed a framework for integrating dietary metrics, such as GI, with ML algorithms in culturally relevant contexts. This approach is particularly relevant for Palestinian populations, where dietary habits and socioeconomic factors influence T2DM risk. (Chatelan et al., 2019) argued that incorporating sociodemographic factors, such as income and education, could improve the relevance and accessibility of dietary recommendations, especially in underserved communities.

Furthermore, XGBoost is widely used in predicting T2D and obesity due to its ability to model complex, non-linear relationships and capture feature interactions within structured data. In a large-scale study, (Ahmad, 2020) applied XGBoost to the National Health and Nutrition Examination Survey (NHANES) dataset to predict T2D, achieving an accuracy of 89.3%, significantly outperforming traditional logistic regression models (82.1%) and RF (87.5%). Important predictors identified by XGBoost included BMI, age, family history of diabetes, and physical activity level. XGBoost's feature importance metrics provided valuable insights into these risk factors, underscoring its value in diabetes prediction (Ahmad, 2020). Additionally, in pediatric obesity research, (Wong, 2019) used XGBoost to predict childhood obesity with genetic, dietary, and environmental data, achieving an area under the ROC curve (AUC) of 0.92. The study highlighted sugar intake and screen time as significant predictors, demonstrating XGBoost's effectiveness in identifying actionable factors for targeted obesity interventions.

LightGBM, with its efficient leaf-wise growth strategy, has proven beneficial for handling large and complex health datasets (Zhang, 2019) applied LightGBM to predict T2D risk in a cohort of over 10,000 participants, achieving a high accuracy of 90.1% and an F1 score of 0.88. The algorithm's performance surpassed that of SVM (86.5%) and neural networks (89.4%), with waist circumference, fasting glucose, and diet quality scores emerging as key predictors. The study emphasized LightGBM's advantages in handling high-dimensional features and scalability, making it highly suitable for large health datasets (Zhang, 2019). LightGBM has also been applied to dietary studies, such as(Yu, 2020), where it classified cardiovascular risk based on dietary intake and lifestyle behaviors with a precision of 87.6% and Area under the curve of 0.91. High intake of processed foods and low fiber consumption

were significant predictors, underscoring LightGBM's utility in pinpointing modifiable dietary risk factors in chronic disease research.

CatBoost is particularly valuable for health studies involving categorical data, such as demographic or socioeconomic information. In a study by(Lee, 2020) on metabolic syndrome prediction, CatBoost effectively managed both continuous and categorical predictors, such as diet, exercise frequency, and family medical history, achieving an accuracy of 88.4% and AUC of 0.89. CatBoost's ordered boosting technique helped prevent target leakage, allowing it to produce unbiased predictions in sequential health data. This ordered boosting approach was critical to handling complex data without bias, making it suitable for large population studies (Lee, 2020). Additionally, CatBoost has been applied in dietary research, as demonstrated by (M. Kim & Choi, 2021), who used it to analyze dietary patterns and obesity risk in a diverse population. CatBoost achieved an F1 score of 0.87 and identified high meal frequency and sugary snack consumption as major obesity risk factors. The algorithm's ability to handle categorical data without extensive preprocessing allowed for accurate predictions while providing streamlined insights into dietary behaviors contributing to obesity.

TabNet's sequential attention mechanism is a valuable feature for applications where feature interpretability is crucial.(Zhao, 2021) used TabNet to predict diabetes progression in a longitudinal cohort study, achieving an AUC of 0.93. The attention mechanism within TabNet enabled the model to highlight the significance of fasting glucose levels and physical activity frequency in predicting disease progression, making it highly interpretable and suitable for clinical decision support (Zhao, 2021).In a related study on obesity-related health risks, (Singh & Verma, 2021) applied TabNet to predict hypertension risk in obese individuals. The model achieved a sensitivity of 92%, with high salt intake and sedentary lifestyle emerging as significant contributors to hypertension risk. TabNet's interpretability through attention mechanisms enabled researchers to rank lifestyle factors by impact, making it ideal for health applications where understanding feature influence is essential.

AutoML frameworks such as AutoGluon have simplified model selection and tuning in health research, making ML accessible to researchers without extensive ML expertise. (Lin, 2022) used AutoGluon to predict diabetes risk, with the algorithm automatically selecting the best models and hyperparameters based on demographic, lifestyle, and clinical data. AutoGluon's ensemble stacking feature achieved an accuracy of 89.7%, surpassing manually tuned XGBoost models (87.2%) and neural networks (88.5%). This automation of model selection and tuning was particularly beneficial for non-experts in the health field who require robust predictive tools without extensive ML training (Lin et al., 2022). In dietary studies, AutoGluon has also proven useful, as demonstrated by(Torres, 2021), who applied it to classify dietary patterns associated with blood pressure. The algorithm achieved an AUC of 0.88, identifying processed meat and high salt intake as primary dietary risk factors for hypertension. AutoGluon's automated hyperparameter tuning enabled rapid experimentation with different models, making it well-suited for handling complex health datasets where manual tuning may be impractical.

Elastic Net regularization is frequently used in health studies with high-dimensional datasets where multicollinearity and sparsity are common. In a study by (Wang, 2019), Elastic Net was used to select genetic and metabolic biomarkers for T2D prediction. The algorithm achieved an accuracy of 85.6%, identifying critical biomarkers such as gene expressions and lipid levels, while the regularization penalties controlled for overfitting. The L1 and L2

penalties in Elastic Net effectively balanced feature selection and multicollinearity management, which was crucial in this genomic context (Wang, 2019). Similarly, in an obesity study, (Garcia, 2020) used Elastic Net to examine lifestyle and genetic predictors, achieving a specificity of 88%. This approach highlighted daily caloric intake, sedentary behavior, and certain genetic markers as critical predictors, illustrating Elastic Net's ability to manage high-dimensional data in health studies by isolating key predictors without inflating the model complexity.

Overall, the application of these ML algorithms in health research—achieving AUC values often exceeding 0.9, high accuracy, and interpretability—demonstrates their effectiveness in identifying critical risk factors and predictors. XGBoost and LightGBM are effective in large datasets, particularly when precision and high-dimensional data handling are required. CatBoost's capabilities in handling categorical data without extensive preprocessing make it ideal for population health studies, while TabNet's interpretability and feature attention mechanisms make it valuable for clinical applications. AutoGluon's automated model selection and tuning benefit researchers without ML expertise, and Elastic Net regularization's balance of feature selection and regularization is especially effective in high-dimensional genomic datasets. These algorithms enhance personalized medicine and public health interventions by offering insights into the factors that contribute to diseases and health risks.

2.10. Summary

The literature highlights the critical role of carbohydrate quality, particularly as measured by GI, in T2DM prevention and management. Regional studies indicate that low-GI diets can improve glycemic control, yet individual variability in glycemic response poses challenges to standardized dietary recommendations. Studies from Palestine and the MENA region show the importance of dietary habits, lifestyle, and socioeconomic factors in shaping T2DM risk, suggesting that interventions must be culturally tailored to maximize effectiveness.

ML offers a promising avenue for addressing these complexities, as it allows for the integration of diverse data sources—including GI, lifestyle, genetic, and microbiota data—into predictive models. Regional studies demonstrate the efficacy of various ML algorithms, such as RF, SVM, and ANN, in accurately predicting T2DM risk, with accuracies ranging from 80% to 90% in local contexts. This study builds on these findings by developing a predictive model for T2DM that utilizes Palestinian data, specifically focusing on the dietary, lifestyle, and sociodemographic factors unique to this population. By advancing a data-driven, personalized approach to T2DM prevention, this research aims to contribute to the growing body of work in precision nutrition and diabetes management.

Chapter Three

Methodology

This chapter presents the methodology used to develop a predictive model for assessing the risk of T2DM in Palestinian adults. It utilizes data from the Palestinian Nutrition Survey 2019, which includes diverse dietary, lifestyle, and sociodemographic factors. The chapter outlines the study's approach, including dataset description, feature categorization, and ML techniques applied for model development. Additionally, it details the validation and evaluation metrics used to ensure the model's accuracy and reliability in predicting T2DM risk.

3.1. Study Description

This study aims to develop a predictive model to assess the risk of T2DM among Palestinian adults using ML techniques. The model utilizes a unique dataset from the **Palestinian Nutrition Survey 2021**, which captures a broad spectrum of dietary, lifestyle, and sociodemographic factors specific to the Palestinian adult population. The rising prevalence of T2DM in Palestine, driven by genetic predispositions, lifestyle, and dietary habits, shows the importance of this research in providing targeted, culturally relevant dietary and lifestyle recommendations to manage and mitigate T2DM risk.

3.2. Dataset Description

The **Palestinian Nutrition Survey 2021** provides a comprehensive sample of 4,500 adults aged 20-64 years from the West Bank, Palestine. This dataset is stratified and weighted by gender and age group to ensure a representative sample across the age groups.

The survey's structured sampling strategy ensures that each demographic segment is proportionally represented, thereby enhancing the generalizability of the model findings. This dataset includes various features across dietary, lifestyle, sociodemographic, and health

categories, which collectively provide a holistic view of the factors influencing T2DM risk in this population.

3.3. Study Features

The Palestinian Nutrition Survey dataset utilized in this research includes multiple features that serve as potential predictors for T2DM. These features are organized into four main categories: dietary, lifestyle, sociodemographic, and anthropometric/health factors. Each category is detailed below, with Table 3.1.A and Table 3.1.B summarizing the study features and their descriptions.

Table 3.1.A: Study Features

Feature Category	Feature Name	Description
Dietary Features	Glycemic Index	Estimated impact of carbohydrate intake on blood glucose levels, based on foods commonly consumed.
	Total Carbohydrate Intake	Daily intake of carbohydrates in grams.
	Fiber Intake	Daily intake of dietary fiber in grams, known to regulate postprandial blood glucose levels.
	Total Caloric Intake	Total daily energy intake in kilocalories, crucial for understanding metabolic load.
	Fat and Protein Intake	Daily intake of fats and proteins in grams, providing a holistic view of macronutrient balance.
Lifestyle Features	Physical Activity Level	Categorized as sedentary, moderate, or high activity; critical for assessing lifestyle-related T2DM risk.
	Smoking Status	Current smoker, past smoker, or non-smoker; smoking is a recognized risk factor for T2DM.
	Sleep Duration	Average hours of sleep per night, as sleep quality impacts insulin sensitivity.
Eating Habits	Meal Frequency	Number of meals consumed per day, influencing blood glucose stability.
	Snacking Frequency	Frequency of snacks between meals, which can impact glycemic control.
	Timing of Largest Meal	The time of day at which the largest meal is consumed, affecting metabolic responses.
	Breakfast Consumption	Whether or not breakfast is regularly consumed, as this impacts metabolic rate and glucose regulation.

Table 3.1.B: Study Features

Feature Category	Feature Name	Description
Sociodemographic Features	Age	Stratified into five groups (20–29, 30–39, 40–49, 50–59, 60–69) to capture age-related risk variations.
	Gender	Male or female.
	Education Level	Primary, secondary, or tertiary education; influences health literacy and access to healthcare.
	Employment Status	Employed, unemployed, or retired; socioeconomic factors impacting lifestyle and dietary choices.
	Income Level	Categorized as low, medium, or high; influences dietary habits and healthcare access.
Anthropometric and Health Features	Body Mass Index	Weight-to-height ratio categorized as underweight, normal, overweight, or obese.
	Blood Pressure	Systolic and diastolic blood pressure readings, as hypertension is often co-morbid with T2DM.
	Family History of Diabetes	Indicates if an immediate family member has T2DM, providing insight into genetic predisposition.

The comprehensive range of features allows for a robust analysis of the T2DM risk profile in the Palestinian population. The dietary features focus on carbohydrate quality and quantity, which are closely linked to blood glucose regulation. Lifestyle factors, such as physical activity and smoking status, provide additional context for T2DM risk. Sociodemographic features, including education and income levels, help account for environmental influences on health behaviors. Anthropometric and health features like BMI and blood pressure offer essential physiological context.

3.4. Data Preprocessing

The data preprocessing phase encompassed a comprehensive series of essential transformations to prepare the dataset for machine learning applications. The initial dataset consisted of 48 distinct features, comprising both continuous and categorical variables. A significant portion of the continuous variables, particularly those related to nutritional intake, were systematically converted into categorical formats following standardized RDA (Recommended Dietary Allowance) guidelines. This conversion process was implemented through carefully designed functions that categorized various nutrients into meaningful groups based on established thresholds.

In **Table 3.2**, fat intake variables (SFA, MUFA, PUFA) were categorized into three levels (0 for Low, 1 for Medium, 2 for High) using scientifically established thresholds. Similar categorization was applied to other nutrients such as proteins, carbohydrates, vitamins, and minerals, each with their specific RDA-based thresholds. For instance, protein intake was evaluated against a range of 50-100g/day, while vitamin C was assessed using a 65-

90mg/day threshold. This standardized approach ensured consistent categorization across all nutritional variables while maintaining clinical relevance.

Table 3.2: Classification guidelines for nutrient intake based on RDA thresholds

Category	Feature	Unit	Threshold		Category		
			Lower	Upper	Low	Medium	High
Macronutrient	Protein	g/day	50	100	<50	50-100	>100
	Carbohydrates	g/day	225	325	<225	225-325	>325
	Fiber	g/day	25	38	<25	25-38	>38
	Fat	g/day	44	77	<44	44-77	>77
	Cholesterol	mg/day	0	300	<0	0-300	>300
Specific Fats	SFA	g	10	20	<10	10-20	>20
	MUFA	g	10	20	<10	10-20	>20
	PUFA	g	5	10	<5	5-10	>10
Vitamins	Vitamin A	mcg/day	700	900	<700	700-900	>900
	Carotene	mg/day	3	6	<3	3-6	>6
	Thiamin	mg/day	1.1	1.2	<1.1	1.1-1.2	>1.2
	Riboflavin	mg/day	1.1	1.3	<1.1	1.1-1.3	>1.3
	Niacin	mg/day	14	16	<14	14-16	>16
	Vitamin B6	mg/day	1.3	1.7	<1.3	1.3-1.7	>1.7
	Folate	mcg/day	400	800	<400	400-800	>800
	Vitamin B12	mcg/day	2.4	4.8	<2.4	2.4-4.8	>4.8
	Vitamin C	mg/day	65	90	<65	65-90	>90
Minerals	Calcium	mg/day	1000	2500	<1000	1000-2500	>2500
	Phosphorus	mg/day	700	4000	<700	700-4000	>4000
	Magnesium	mg/day	310	420	<310	310-420	>420
	Iron	mg/day	8	18	<8	8-18	>18
	Zinc	mg/day	8	11	<8	8-11	>11
	Copper	mg/day	0.9	1.8	<0.9	0.9-1.8	>1.8
	Sodium	mg/day	1500	2300	<1500	1500-2300	>2300
	Potassium	mg/day	3400	4700	<3400	3400-4700	>4700

The handling of missing values followed a sophisticated three-tiered approach. The initial stage leveraged Cramér's V correlation analysis to implement an informed grouped mode imputation strategy. This method was particularly valuable as it preserved the intricate relationships between features, as documented in previous research. Building upon this foundation, the second tier employed K-Nearest Neighbors (KNN) imputation methodology to address the remaining gaps in the dataset. This process required particular attention to the nuanced conversion between categorical and numerical data formats, following established best practices. The final stage of imputation addressed any residual missing entries through the application of column-specific modes, ensuring complete data coverage while maintaining statistical relevance.

A significant challenge in the dataset was the presence of class imbalance, which necessitated a combined resampling approach. The methodology integrated the Synthetic Minority Over-sampling Technique (SMOTE) with random under-sampling applied to the training dataset. This balanced strategy simultaneously generated synthetic samples for the minority class while reducing the majority class instances, effectively addressing the imbalance without compromising the dataset's integrity. Through this careful preprocessing sequence, the final dataset was refined to include 4,680 cases, establishing a robust foundation for subsequent analysis.

3.5. Model Testing and Validation

To ensure the robustness and generalizability of the predictive models, rigorous testing and validation procedures will be employed. Cross-validation techniques, including k-fold cross-validation, will be applied to evaluate model performance across multiple subsets of the dataset. This technique involves partitioning the data into k subsets, where each subset serves as a validation set while the remaining subsets are used for training. The process is repeated k times, and the average performance across all folds provides a comprehensive assessment of the model’s generalizability.

Additionally, hyperparameter tuning will be conducted for each model to identify the optimal configuration that maximizes accuracy and minimizes overfitting. Grid search and randomized search techniques will be used to explore a range of parameters, selecting the configuration that delivers the highest performance on validation data. Following hyperparameter tuning, the final model will be evaluated on a reserved test set to assess its real-world applicability.

3.6. Model Evaluation Metrics

To further validate model reliability, metrics such as accuracy, precision, recall, F1 score, and AUC-ROC will be calculated for each algorithm. These metrics provide insights into the model’s ability to correctly classify T2DM risk and distinguish between high-risk and low-risk cases. By implementing these validation methods, the study ensures that the predictive models are robust, accurate, and capable of providing reliable insights into T2DM risk among Palestinian adults.

The following metrics will be used to evaluate model performance, providing a comprehensive view of each model's accuracy, sensitivity, specificity, and overall predictive power:

1. Accuracy: Accuracy measures the proportion of correctly classified instances out of the total predictions made by the model. It is calculated as:

$$Accuracy = \frac{True\ Positives\ + True\ Negatives}{Total\ Predictions} \quad \ldots.(3.1)$$

2. Precision: Precision represents the ratio of correctly predicted positive instances to the total positive predictions made by the model. It is calculated as:

$$Precision\ = \frac{True\ Positives}{(True\ Positives\ +\ False\ Positives)} \quad \ldots.(3.2)$$

3. Recall (Sensitivity): Recall is the ratio of correctly predicted positive instances to all actual positive instances in the dataset, indicating the model's ability to detect true positive cases. It is calculated as:

$$Recall\ = \frac{True\ Positives}{(True\ Positives\ +\ False\ Negatives)} \quad \ldots.(3.3)$$

4. F1 Score: The F1 Score is the harmonic mean of precision and recall, balancing the two metrics in cases of class imbalance. It is calculated as:

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \ldots(3.4)$$

5. Area Under the Receiver Operating Characteristic Curve: AUC-ROC measures the model's discriminative ability across different classification thresholds, with higher AUC values indicating better performance in distinguishing between positive and negative cases. A high AUC-ROC value implies that the model is highly capable of distinguishing high-risk from low-risk cases.

Chapter Four

Study Results

This chapter presents the results of the study, focusing on the sociodemographic and health-related characteristics of participants categorized by their diabetic status. The findings offer a comprehensive view of the factors associated with the risk of diabetes within the study population, providing insights into how variables such as age, gender, education, lifestyle, and SES impact diabetes risk. The results are analyzed to highlight key trends and disparities, which are essential for understanding the broader implications of T2DM in the population studied.

4.1. Results

The analysis shown in Table 4.1 highlights sociodemographic and health-related characteristics of participants categorized by diabetic status. Urban locality participants had a higher percentage of diabetics (45.5%) compared to rural (33.0%) and camp residents (21.5%). Regional distribution showed a slightly higher prevalence of diabetes in the West Bank (67.0%) than Gaza (33.0%). Educational level revealed a striking difference, with the majority of diabetics (63.1%) having low education levels (0–8 years), while fewer had higher levels of education (21.4% with ≥12 years).

Gender analysis indicated a higher prevalence of diabetes among females (63.0%) compared to males (37.0%). The age distribution showed that diabetes increased significantly with age, with 50.5% of diabetics in the 50–59 age group, while the 18–29 age group had only 1.7%. Employment status showed that a higher proportion of diabetics were employed (79.6%) compared to the unemployed (20.4%).

BMI showed that the majority of diabetics were obese (75.3%), while overweight and normal BMI categories were less represented (17.8% and 6.9%, respectively). Smoking habits revealed a lower prevalence of smoking among diabetics (20.9%) compared to non-smokers (79.1%). Family income analysis indicated that diabetes was most prevalent among

individuals with below-average income (<300 JD) at 55.6%, compared to average (34.0%) and above-average incomes (10.4%).

Perceived health status differed notably, with a larger percentage of diabetics reporting "not good" health (28.1%) compared to those perceiving their health as "very good" (27.3%). Finally, physical activity was low among all participants, with 94.1% of diabetics reporting no physical activity, compared to only 5.9% who engaged in physical activity. These findings highlight critical sociodemographic and lifestyle factors associated with diabetes prevalence in the study population.

Table 4.1: Sociodemographic and Health-Related Characteristics of Study Participants by Diabetic Status

Features	Categories	Non-Diabetic		Diabetic		Total	
		N	%	N	%	N	%
Locality	Urban	929	36.4	270	45.5	1199	38.1
	Rural	1031	40.4	196	33.0	1227	39.0
	Camp	591	23.2	128	21.5	719	22.9
Region	West Bank	1604	62.9	398	67.0	2002	63.7
	Gaza	947	37.1	196	33.0	1143	36.3
Educational Level	0 - 8 Y	868	34.0	375	63.1	1243	39.5
	9 - 11 Y	642	25.2	92	15.5	734	23.3
	>=12 Y	1041	40.8	127	21.4	1168	37.1
Gender	Male	1316	51.6	220	37.0	1536	48.8
	Female	1235	48.4	374	63.0	1609	51.2
Age Group	18 – 29	1020	40.0	10	1.7	1030	32.8
	30 – 39	1007	39.5	69	11.6	1076	34.2
	40 – 49	447	17.5	215	36.2	662	21.0
	50 – 59	77	3.0	300	50.5	377	12.0
Employment	Employed	1824	71.5	473	79.6	2297	73.0
	Un Employed	727	28.5	121	20.4	848	27.0
BMI Category	Normal	991	38.8	41	6.9	1032	32.8
	Overweight	1064	41.7	106	17.8	1170	37.2
	Obese	496	19.4	447	75.3	943	30.0
Current Smoking	No	1793	70.3	470	79.1	2263	72.0
	Yes	758	29.7	124	20.9	882	28.0
Family Income	Above Average > 600 JD	296	11.6	62	10.4	358	11.4
	Average 300- 600 JD	1025	40.2	202	34.0	1227	39.0
	Below Average (<300JD)	1230	48.2	330	55.6	1560	49.6
Perceived Health Status	Very Good	1416	55.5	162	27.3	1578	50.2
	Good	963	37.7	265	44.6	1228	39.0
	Not Good	172	6.7	167	28.1	339	10.8
Physical Activity	No	2334	91.5	559	94.1	2893	92.0
	Yes	217	8.5	35	5.9	252	8.0

The analysis in Table 4.2 outlines the recommended daily energy intake thresholds based on age group and gender, as defined by RDA guidelines. For males, the recommended caloric intake thresholds decrease progressively with age. In the 18–29 age group, the thresholds are less than 2200 kcal/day for low activity levels, 2200–2800 kcal/day for moderate activity, and more than 2800 kcal/day for high activity levels. For males aged 30–39, the thresholds decrease slightly to less than 2100 kcal/day for low activity, 2100–2700 kcal/day for moderate activity, and more than 2700 kcal/day for high activity. This trend continues for older age groups, with the lowest thresholds being for males aged 50–59, where the recommended intake is less than 1900 kcal/day for low activity, 1900–2500 kcal/day for moderate activity, and more than 2500 kcal/day for high activity levels.

For females, a similar pattern of decreasing energy requirements with age is observed. Females in the 18–29 age group have recommended thresholds of less than 1800 kcal/day for low activity, 1800–2400 kcal/day for moderate activity, and more than 2400 kcal/day for high activity. For females aged 30–39, the thresholds decrease to less than 1700 kcal/day for low activity, 1700–2300 kcal/day for moderate activity, and more than 2300 kcal/day for high activity. The lowest thresholds are seen in females aged 50–59, with recommendations of less than 1500 kcal/day for low activity, 1500–2100 kcal/day for moderate activity, and more than 2100 kcal/day for high activity.

Table 4.2: Recommended Daily Energy Intake Thresholds by Age Group and Gender, According to RDA Guidelines.

Age Group	Gender	Low (kcal/day)	Moderate (kcal/day)	High (kcal/day)
18-29	Male	< 2200	2200 - 2800	> 2800
	Female	< 1800	1800 - 2400	> 2400
30-39	Male	< 2100	2100 - 2700	> 2700
	Female	< 1700	1700 - 2300	> 2300
40-49	Male	< 2000	2000 - 2600	> 2600
	Female	< 1600	1600 - 2200	> 2200
50-59	Male	< 1900	1900 - 2500	> 2500
	Female	< 1500	1500 - 2100	> 2100

The analysis of Table 4.3.A and Table 4.3.B examines macronutrient intake characteristics among participants categorized by diabetic status. Energy intake levels were similar between non-diabetic and diabetic participants, with approximately 47% in the low-intake category, around 28% in the moderate category, and close to 25% in the high category. Protein consumption showed differences, as 32.2% of diabetics consumed less than 50 grams per day compared to 25.3% of non-diabetics. Conversely, fewer diabetics consumed 50–100 grams daily (44.9%) compared to non-diabetics (52.5%), while high protein intake above 100 grams was relatively consistent across groups.

Carbohydrate intake revealed that 48% of diabetics consumed less than 225 grams daily, compared to 34.5% of non-diabetics. Moderate intake (225–325 grams) was reported by 31.5% of diabetics and 35.4% of non-diabetics, while higher intake above 325 grams was less common among diabetics (20.5%) than non-diabetics (30.0%). Fiber intake was low across both groups, with over 70% of diabetics and nearly 69% of non-diabetics consuming less than 25 grams per day.

Fat intake differed by diabetic status, with 35.5% of diabetics consuming less than 44 grams daily compared to 25.6% of non-diabetics. A similar proportion of participants from both groups consumed 44–77 grams daily (approximately 33%). High fat intake above 77 grams was more common among non-diabetics (40.0%) than diabetics (31.8%). Cholesterol intake was consistent across groups, with about 59% of participants consuming less than 300 mg daily, and 41% exceeding this threshold.

For Saturated Fatty Acids, the majority of both groups consumed less than 10 grams daily (approximately 66%), with small percentages consuming 10–20 grams or exceeding 20 grams. Monounsaturated fatty acid intake was highest in the 10–20 grams/day range for both diabetics (53.0%) and non-diabetics (52.5%), while a smaller proportion consumed more than 20 grams daily. Polyunsaturated fatty acid intake followed a similar trend, with about half of both groups consuming 5–10 grams daily, while fewer diabetics (20.2%) than non-diabetics (25.4%) consumed more than 10 grams.

The results indicate distinct patterns in macronutrient consumption between diabetics and non-diabetics, with diabetics generally consuming lower amounts of carbohydrates, fats, and PUFAs, while exhibiting higher prevalence of low protein and fiber intake. These findings highlight nutritional disparities that may influence or reflect diabetic status.

Table 4.3.A: Macro nutrient Intake Characteristics of Study Participants by Diabetic Status

Features	Categories	Non-Diabetic		Diabetic		Total	
		N	%	N	%	N	%
Energy Intake	Low	1202	47.1	282	47.5	1484	47.2
	Moderate	713	27.9	167	28.1	880	28.0
	High	636	24.9	145	24.4	781	24.8
Protein	< 50 grams/day	646	25.3	191	32.2	837	26.6
	50 - 100 grams/day	1338	52.5	267	44.9	1605	51.0
	>100 grams/day	567	22.2	136	22.9	703	22.4
Carbohydrate	< 225 grams/day	881	34.5	285	48.0	1166	37.1
	225 - 325 grams/day	904	35.4	187	31.5	1091	34.7
	> 325 grams/day	766	30.0	122	20.5	888	28.2
Fiber	< 25 grams/day	1749	68.6	423	71.2	2172	69.1
	25 - 38 grams/day	601	23.6	131	22.1	732	23.3
	> 38 grams/day	201	7.9	40	6.7	241	7.7

Table 4.3.B: Macro nutrient Intake Characteristics of Study Participants by Diabetic Status

Features	Categories	Non-Diabetic		Diabetic		Total	
		N	%	N	%	N	%
Fat	< 44 grams/day	652	25.6	211	35.5	863	27.4
	44 - 77 grams/day	879	34.5	194	32.7	1073	34.1
	> 77 grams/day	1020	40.0	189	31.8	1209	38.4
Cholesterol	< 300 mg/day	1508	59.1	349	58.8	1857	59.0
	> 300 mg/day	1043	40.9	245	41.2	1288	41.0
SFA	< 10 grams/day	1679	65.8	390	65.7	2069	65.8
	10 - 20 grams/day	821	32.2	185	31.1	1006	32.0
	20 grams/day	51	2.0	19	3.2	70	2.2
MUFA	< 10 grams/day	990	38.8	213	35.9	1203	38.3
	10 - 20 grams/day	1338	52.5	315	53.0	1653	52.6
	20 grams/day	223	8.7	66	11.1	289	9.2
PUFA	< 5 grams/day	620	24.3	181	30.5	801	25.5
	5 - 10 grams/day	1284	50.3	293	49.3	1577	50.1
	> 10 grams/day	647	25.4	120	20.2	767	24.4

The analysis of Table 4.4 presents the vitamin intake characteristics of participants by diabetic status. For vitamin A, the majority of both non-diabetic and diabetic participants consumed more than 900 mcg/day, with 87.2% of non-diabetics and 86.9% of diabetics falling into this category. Only a small proportion of participants consumed less than 700 mcg/day or between 700 and 900 mcg/day, with similar distributions across both groups.

For vitamin B6, 53.9% of diabetics consumed less than 1.3 mg/day compared to 49.7% of non-diabetics. Moderate intake (1.3–1.7 mg/day) was observed in 23.1% of diabetics and 22.5% of non-diabetics. Higher intake levels above 1.7 mg/day were slightly more common among non-diabetics (27.8%) compared to diabetics (23.1%).

Vitamin B12 intake revealed that a significant proportion of participants in both groups consumed less than 2.4 mcg/day, with 58.8% of non-diabetics and 52.4% of diabetics in this category. Moderate intake levels (2.4–4.8 mcg/day) were reported by 19.9% of diabetics and 14.7% of non-diabetics. High intake levels above 4.8 mcg/day were slightly more frequent among diabetics (27.8%) than non-diabetics (26.6%).

For vitamin C, intake patterns were comparable between the groups. Approximately 44% of both non-diabetics and diabetics consumed less than 65 mg/day. Moderate intake (65–90 mg/day) was less common, reported by 11.6% of diabetics and 13.4% of non-diabetics. High intake levels above 90 mg/day were slightly more prevalent among diabetics (45.1%) compared to non-diabetics (42.4%).

The findings indicate that while most participants met or exceeded the recommended intake thresholds for vitamins A and C, a significant proportion consumed insufficient levels of vitamins B6 and B12, particularly among diabetic participants. This may reflect nutritional disparities and highlights potential areas for dietary improvement.

Table 4.4: Vitamins Nutrient Intake Characteristics of Study Participants by Diabetic Status.

Features	Categories	Non-Diabetic		Diabetic		Total	
		N	%	N	%	N	%
Vitamin A	< 700 mcg/day	210	8.2	52	8.8	262	8.3
	700 - 900 mcg/day	116	4.5	26	4.4	142	4.5
	> 900 mcg/day	2225	87.2	516	86.9	2741	87.2
Vitamin B6	< 1.3 mg/day	1268	49.7	320	53.9	1588	50.5
	1.3 - 1.7 mg/day	575	22.5	137	23.1	712	22.6
	> 1.7 mg/day	708	27.8	137	23.1	845	26.9
Vitamin B12	< 2.4 mcg/day	1499	58.8	311	52.4	1810	57.6
	2.4 - 4.8 mcg/day	374	14.7	118	19.9	492	15.6
	> 4.8 mcg/day	678	26.6	165	27.8	843	26.8
Vitamin C	< 65 mg/day	1127	44.2	257	43.3	1384	44.0
	65 - 90 mg/day	343	13.4	69	11.6	412	13.1
	> 90 mg/day	1081	42.4	268	45.1	1349	42.9

The analysis of Table 4.5 presents the mineral nutrient intake characteristics of participants by diabetic status. For calcium, the vast majority of both non-diabetic and diabetic participants consumed less than 1000 mg/day, with slightly higher prevalence among diabetics (94.8%) compared to non-diabetics (92.6%). Only a small fraction of participants consumed calcium within the recommended range of 1000–2500 mg/day, and very few exceeded 2500 mg/day. Phosphorus intake showed that 35.0% of diabetics consumed less than 700 mg/day, compared to 27.2% of non-diabetics. The majority of both groups consumed between 700 and 4000 mg/day, with no participants among the diabetic group exceeding 4000 mg/day.

For magnesium, 73.6% of diabetics and 66.8% of non-diabetics consumed less than 310 mg/day, reflecting inadequate intake levels in both groups. Moderate intake (310–420 mg/day) was reported by 16.2% of diabetics and 20.8% of non-diabetics, while a smaller proportion consumed more than 420 mg/day. Iron intake showed disparities, with 39.7% of diabetics consuming less than 8 mg/day compared to 31.1% of non-diabetics. Moderate intake (8–18 mg/day) was higher among non-diabetics (54.2%) than diabetics (47.3%), while high intake above 18 mg/day was similar between the groups.

Zinc intake followed a similar pattern, with 51.9% of diabetics and 46.1% of non-diabetics consuming less than 8 mg/day. Higher intake levels (above 11 mg/day) were slightly more common among diabetics (31.0%) compared to non-diabetics (29.9%). Copper intake was generally adequate, with the majority of both groups consuming between 0.9 and 1.8 mg/day. However, 20.0% of diabetics consumed less than 0.9 mg/day compared to 14.4% of non-

diabetics. High intake above 1.8 mg/day was slightly less common among diabetics (24.1%) compared to non-diabetics (29.6%).

Sodium intake showed that the majority of participants consumed more than 2300 mg/day, with 83.7% of diabetics and 91.1% of non-diabetics exceeding this level, reflecting excessive sodium consumption in both groups. Very few participants in either group consumed sodium within the recommended range of 1500–2300 mg/day, and an even smaller percentage consumed less than 1500 mg/day.

The findings reveal widespread inadequacies in mineral nutrient intake, with notable deficiencies in calcium, magnesium, and iron among both diabetics and non-diabetics. Excessive sodium intake was highly prevalent, highlighting the need for dietary interventions to address mineral imbalances, particularly among individuals with diabetes.

Table 4.5: Minerals Nutrient Intake Characteristics of Study Participants by Diabetic Status

Features	Categories	Non-Diabetic		Diabetic		Total	
		N	%	N	%	N	%
Calcium	< 1000 mg/day	2363	92.6	563	94.8	2926	93.0
	1000 - 2500 mg/day	182	7.1	28	4.7	210	6.7
	> 2500 mg/day	6	0.2	3	0.5	9	0.3
Phosphorus	< 700 mg/day	694	27.2	208	35.0	902	28.7
	700 - 4000 mg/day	1851	72.6	386	65.0	2237	71.1
	> 4000 mg/day	6	0.2	0	0.0	6	0.2
Magnesium	< 310 mg/day	1704	66.8	437	73.6	2141	68.1
	310 - 420 mg/day	530	20.8	96	16.2	626	19.9
	> 420 mg/day	317	12.4	61	10.3	378	12.0
Iron	< 8 mg/day	794	31.1	236	39.7	1030	32.8
	8 - 18 mg/day	1382	54.2	281	47.3	1663	52.9
	> 18 mg/day	375	14.7	77	13.0	452	14.4
Zinc	< 8 mg/day	1176	46.1	308	51.9	1484	47.2
	8 - 11 mg/day	611	24.0	102	17.2	713	22.7
	> 11 mg/day	764	29.9	184	31.0	948	30.1
Copper	< 0.9 mg/day	368	14.4	119	20.0	487	15.5
	0.9 - 1.8 mg/day	1428	56.0	332	55.9	1760	56.0
	> 1.8 mg/day	755	29.6	143	24.1	898	28.6
Sodium	< 1500 mg/day	88	3.4	33	5.6	121	3.8
	1500 - 2300 mg/day	140	5.5	64	10.8	204	6.5
	> 2300 mg/day	2323	91.1	497	83.7	2820	89.7

The analysis of Table 4.6 evaluates the performance of various ML models in predicting T2DM based on accuracy, precision, recall, F1-score, and AUC. GB demonstrated the highest overall performance, with an accuracy of 94.2%, precision, recall, and F1-score all at 94%, and an AUC of 0.985, indicating strong predictive capabilities and model robustness. SVM followed closely, achieving an accuracy of 92.8%, with precision, recall, and F1-score at 93%, and a slightly lower AUC of 0.978 compared to GB.

LR and RF models performed comparably, both achieving an F1-score of 91%. LR showed slightly higher accuracy at 90.8% compared to RF's 90.6%, while RF had a marginally higher AUC at 0.986 compared to LR's 0.974, suggesting stronger discriminative power for RF. The Decision Tree CART model demonstrated relatively high performance with an accuracy of 92.5%, precision and recall at 93%, an F1-score of 92%, and an AUC of 0.926, though it was outperformed by GB and SVM in most metrics.

GB and SVM models emerged as the top-performing algorithms for predicting T2DM, showcasing high accuracy and balanced precision and recall. While LR and RF also performed well, they slightly lagged in accuracy and F1-scores. The DT model, although effective, exhibited comparatively lower AUC, indicating less robust discrimination between classes. These results highlight the GB model as the most suitable choice for this task.

Table 4.6: Performance Metrics of Machine Learning Models on Predicting T2D

Algorithm	Accuracy	Precision	Recall	F1-Score	AUC
Random Forest (RF)	0.906	0.91	0.91	0.91	0.986
Support Vector Machine (SVM)	0.928	0.93	0.93	0.93	0.978
Logistic Regression (LR)	0.908	0.91	0.91	0.91	0.974
Gradient Boosting (GB)	0.942	0.94	0.94	0.94	0.985
Decision Tree (CART)	0.925	0.93	0.93	0.92	0.926

The provided ROC curve graph illustrates the performance of various ML models in predicting T2DM based on their respective AUC values. The GB and RF models achieved the highest AUC values, with GB at 0.985 ± 0.025 and RF at 0.986 ± 0.023, indicating their strong capability to discriminate between diabetic and non-diabetic cases. These models are closely followed by the SVM with an AUC of 0.978 ± 0.034 and LR with an AUC of 0.974 ± 0.028, both of which also demonstrated excellent performance.

The Decision Tree (CART) model, with an AUC of 0.926 ± 0.066, showed comparatively lower performance, highlighting less robust classification ability than the other models. The random line (dashed) represents a baseline performance, which the models significantly outperformed, confirming their predictive efficacy.

The ROC curves in Figure 4.1 reveal that all models except CART exhibited high levels of sensitivity and specificity, with GB and RF emerging as the best-performing models due to their superior AUC scores and minimal variance across cross-validation folds. These findings support the use of GB and RF as optimal algorithms for this task, while LR and SVM also provide reliable alternatives.

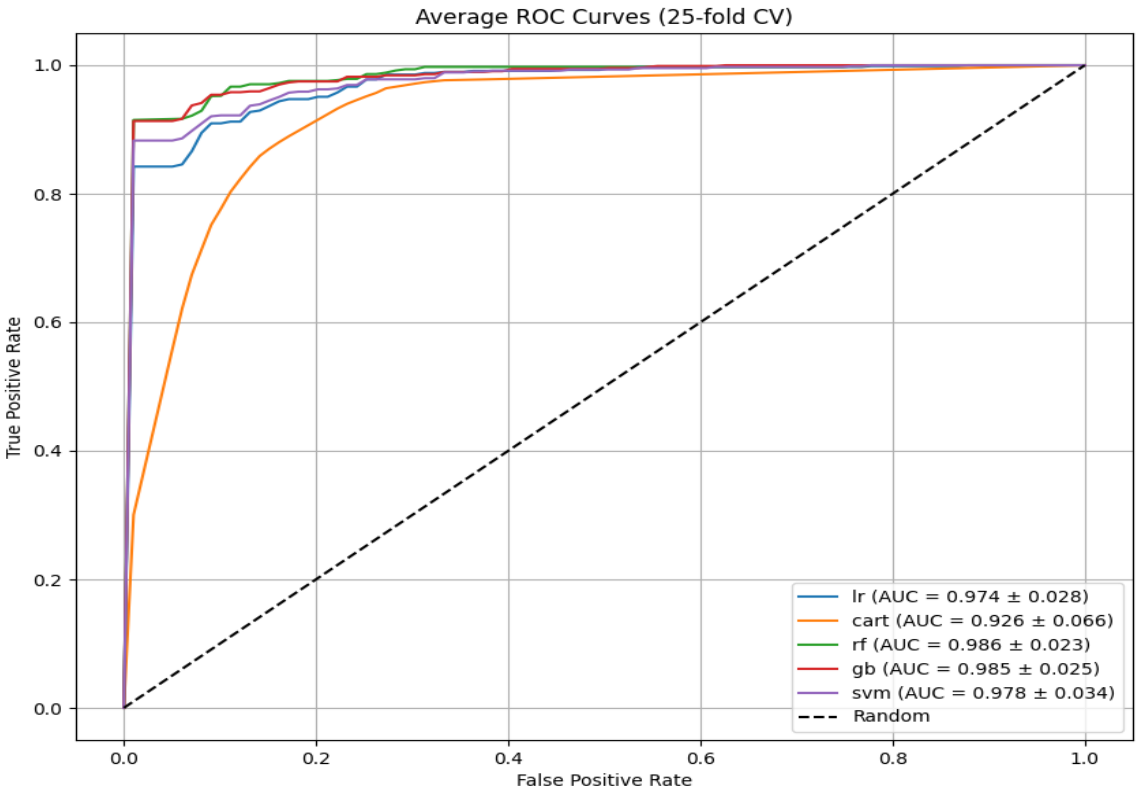

Figure 4.1: Comparative analysis of the ROC curves for the ML models.

The feature importance analysis shown in the Figure 4.2 highlights the most influential factors in predicting T2DM. GL emerges as the most critical feature, indicating that it plays a central role in distinguishing between diabetic and non-diabetic individuals. This is followed by Age Group, BMI Category, and GI, which are also significant contributors, reflecting the importance of age, body composition, and carbohydrate-related measures in diabetes risk.

Energy Classified ranks next, suggesting that caloric intake levels are an important predictor. Dietary factors such as Fiber intake and Diagnosed High Fat intake also show substantial importance, emphasizing the relevance of dietary habits in diabetes prediction. Additional features like Waist-to-Hip Ratio (WHR), Copper levels, and diagnoses of conditions such as stroke are moderately important, indicating a connection between metabolic and cardiovascular risk factors with diabetes.

Other features with lower importance, including Perceived Health Status, Educational Level, and specific nutrient intakes like Phosphorus and Magnesium, still contribute to the model but have less predictive power. Features like Vitamin B6 intake and being on a diet for weight loss exhibit the least importance among the top 20, suggesting their influence on diabetes prediction is minimal in comparison.

The analysis shows the significance of dietary glycemic factors, age, body composition, and related health conditions in predicting diabetes, providing valuable insights for targeted prevention and intervention strategies.

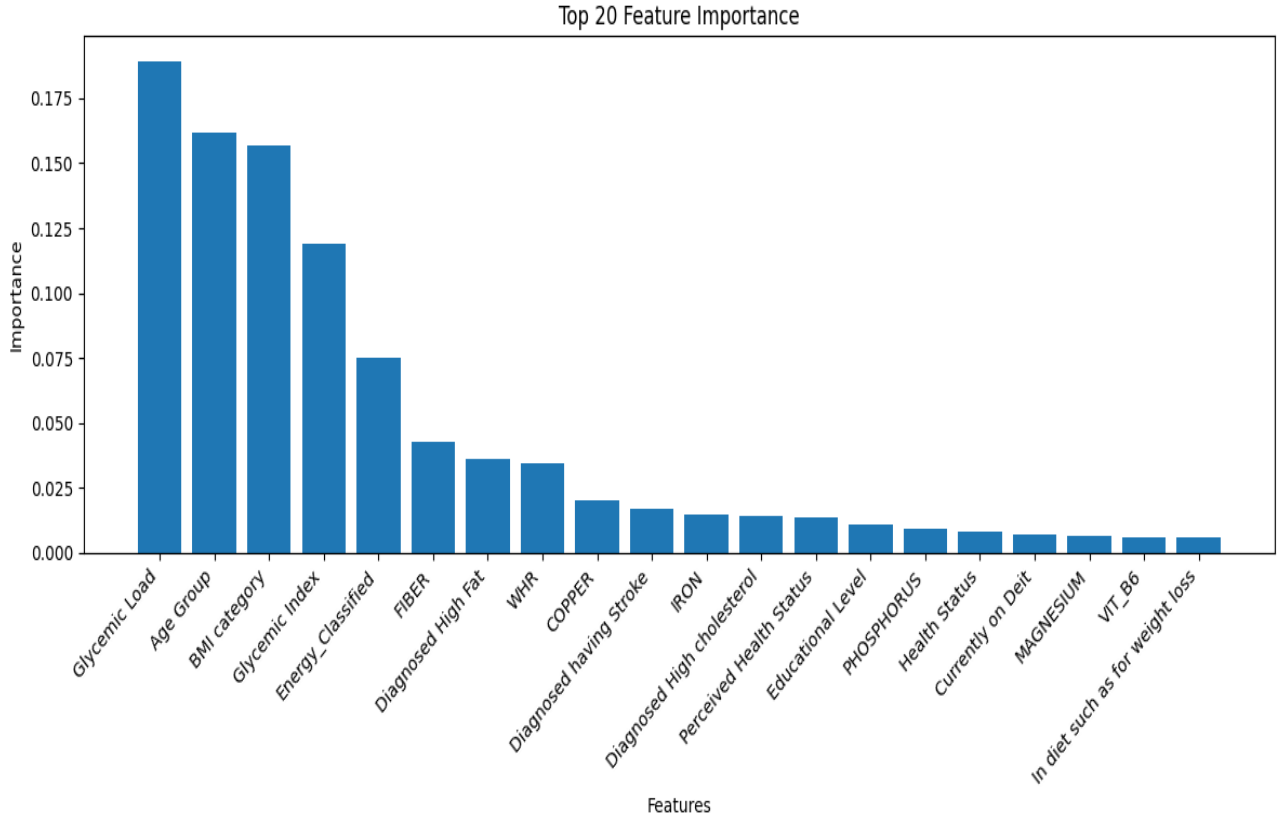

Figure 4.2: The most influential factors in predicting T2DM.

Chapter Five

Discussion and Conclusion

This Chapter interpreted the study main results with previous studies. Then conclusion and recommendations were mentioned next.

5.1. Discussion

The study identifies several key sociodemographic, nutritional, and behavioral factors associated with diabetes, offering insights into the mechanisms behind its prevalence. Urban residency was significantly associated with diabetes prevalence (45.5%), compared to rural (33%) and camp residents (21.5%). This pattern aligns with the global trend where urbanization promotes sedentary lifestyles, higher caloric intake, and reduced physical activity. Studies, such as (Kearns et al., 2014), have shown that urban environments increase the risk of metabolic disorders, including diabetes, due to accessibility to processed foods and limited physical activity opportunities.

The gender disparity in diabetes prevalence, with females (63%) exhibiting higher rates than males (37%), suggests complex interactions between biological, hormonal, and social factors. Estrogen metabolism and gestational diabetes are recognized contributors to this trend, as highlighted by (Wild et al., 2004), who reported similar gender differences worldwide. Social determinants such as reduced physical activity and caregiving roles further compound these risks.

Age also emerged as a critical determinant, with prevalence significantly higher among participants aged 50–59 (50.5%). This finding is consistent with established evidence linking aging to insulin resistance, reduced pancreatic beta-cell function, and cumulative exposure to risk factors. For instance, (Fuchsberger et al., 2016) demonstrated that age-related genetic and epigenetic modifications impair glucose homeostasis. The association between lower education levels and diabetes (63.1% with less than eight years of education) emphasizes the role of health literacy in managing and preventing diabetes. (Xu et al., 2018) similarly identified education as a predictor of health behaviors and disease prevention.

Nutritional analysis highlighted significant disparities among diabetics. Low fiber intake (71.2% consuming <25 grams/day) reflects a key nutritional deficiency, given fiber's role in modulating glycemic responses and enhancing insulin sensitivity. (Reynolds et al., 2019) confirmed that higher dietary fiber intake reduces T2D risk and improves glycemic control. Furthermore, low PUFA consumption among diabetics (20.2% consuming >10 grams/day) indicates a gap in dietary quality, as PUFAs are critical for lipid metabolism and reducing insulin resistance. (Wu et al., 2012) reported similar benefits of PUFA consumption in reducing metabolic syndrome risk.

Excessive sodium intake was highly prevalent (83.7%) among diabetics, reflecting dietary patterns that exacerbate hypertension and cardiovascular complications, which are common in diabetic populations. A global assessment by (Powles et al., 2013) emphasized the health risks of high sodium consumption, particularly in metabolic disorders. Similarly, deficiencies in calcium, magnesium, and iron observed in this study parallel findings by (Barbagallo et al., 2022), who linked inadequate magnesium intake to poor glycemic control.

Obesity was prevalent among diabetics (75.3%), corroborating its role as a key modifiable risk factor for diabetes through mechanisms such as chronic inflammation and insulin resistance. The low physical activity rates among diabetics (94.1%) exacerbate these risks, highlighting the urgent need for interventions to promote exercise. Studies from the Middle East, such as (Al-Nozha et al., 2016), have similarly reported high physical inactivity rates among diabetic populations, reinforcing the global nature of this challenge.

The ML analysis identified GB as the most effective predictive model for diabetes, achieving an accuracy of 94.2% and an AUC of 0.985. GB's ability to model complex interactions between predictors explains its superior performance, consistent with findings by (Li et al., 2020; Zhang, 2019), who demonstrated similar accuracy in predicting diabetes using GB. RF and SVM also performed well, though their slightly lower accuracy show the advantages of ensemble-based methods in health data modeling.(Choi et al., 2019) similarly validated the effectiveness of GB for predicting health outcomes.

Feature importance analysis revealed GL, BMI, and Age as the top predictors of diabetes. High GL diets were shown to increase diabetes risk by 40% in a cohort study by the Harvard School of Public Health. BMI and Age further highlight the critical roles of body composition and aging in diabetes pathophysiology. These findings align with the EPIC-InterAct study, which emphasized lifestyle interventions focusing on diet and physical activity as preventive measures (Dhana et al., 2016).

The study shows the complex interplay between sociodemographic, nutritional, and behavioral factors in diabetes prevalence and prediction. These findings highlight the need for comprehensive, multidisciplinary strategies to reduce diabetes burden, focusing on education, dietary improvements, and increased physical activity. Future research should validate these findings across diverse populations and explore interventions tailored to address the specific needs of high-risk groups.

5.2. Conclusion

This study provided a comprehensive analysis of the factors influencing T2D prevalence, incorporating sociodemographic, nutritional, and behavioral determinants while leveraging ML for predictive modeling. The results show the complex interplay between urbanization, aging, education levels, dietary habits, and physical activity in driving diabetes risk. The GB model emerged as the most effective predictive tool, emphasizing the potential of advanced ML methods in healthcare applications. These findings highlight the urgent need for integrated interventions, including education, dietary improvements, and physical activity promotion, to mitigate the diabetes burden. The incorporation of ML in healthcare can play a pivotal role in early detection and personalized prevention strategies, ultimately improving outcomes for at-risk populations.

5.3. Strengths and Limitations

A key strength of the study lies in the comprehensiveness of the dataset, which captured diverse sociodemographic, nutritional, and behavioral factors, allowing for a detailed exploration of diabetes risk determinants. The innovative application of ML, particularly the GB algorithm, demonstrated the utility of advanced data analytics in achieving high accuracy and robust predictive insights. Furthermore, the inclusion of detailed dietary and physical activity data provided nuanced perspectives on the lifestyle factors influencing diabetes risk, with the study offering regionally relevant findings applicable to populations with similar demographic profiles.

However, the cross-sectional nature of the study limits the ability to establish causal relationships between the identified factors and diabetes prevalence. Additionally, reliance on self-reported data for health and lifestyle variables introduces potential reporting biases, such as underestimation of unhealthy behaviors. The binary classification of physical activity lacked granularity, potentially overlooking the effects of activity intensity or frequency. Moreover, while the findings are regionally significant, their generalizability may be limited for populations with different genetic, cultural, or environmental characteristics.

5.3. Future Work

Future research should adopt longitudinal designs to establish causal pathways between risk factors and diabetes onset. Enhanced dietary analyses, including detailed profiling of micronutrients and emerging dietary biomarkers, could provide deeper insights into nutritional influences on diabetes. The integration of geospatial data would be valuable for identifying regional disparities and the impact of environmental factors, such as urban planning and food availability, on diabetes risk.

Efforts to improve physical activity measurement, including the use of wearable devices for continuous and objective data collection, would allow for more accurate assessments of activity patterns and their impact on diabetes. The inclusion of genetic and epigenetic data in future studies could further refine predictive models, uncovering interactions between genetic predisposition and environmental factors. Policy-oriented research evaluating the effectiveness of interventions, such as education campaigns, dietary regulations, and physical activity programs, would help inform evidence-based strategies for diabetes prevention and management.

Advancing ML applications in real-time predictive tools for clinical and public health settings could revolutionize diabetes risk assessment and early intervention. By addressing these areas, future research can build upon the findings of this study, contributing to a deeper understanding of T2D and enhancing prevention and treatment approaches.

References

- Abu-Mweis, S. S., Tayyem, R. F., Bawadi, H. A., Musaiger, A. O., & Al-Hazzaa, H. M. (2014). Eating habits, physical activity, and sedentary behaviors of Jordanian adolescents' residents of Amman. Mediterranean Journal of Nutrition and Metabolism, 7(1), 67–74. https://doi.org/10.3233/MNM-140007
- Abu-Rmeileh, N. M. E., Husseini, A., Capewell, S., O'Flaherty, M., & MEDCHAMPS project. (2013). Preventing type 2 diabetes among Palestinians: comparing five future policy scenarios. BMJ Open, 3(12), e003558. https://doi.org/10.1136/bmjopen-2013-003558
- Abu-Rmeileh, N. M., Husseini, A., O'Flaherty, M., Shoaibi, A., & Capewell, S. (2012). Forecasting prevalence of type 2 diabetes mellitus in Palestinians to 2030: validation of a predictive model. The Lancet, 380, S21. https://doi.org/10.1016/S0140-6736(13)60202-0
- Agardh, E., Allebeck, P., Hallqvist, J., Moradi, T., & Sidorchuk, A. (2011). Type 2 diabetes incidence and socio-economic position: a systematic review and meta-analysis. Int. J. Epidemiol., 40(3), 804–818.
- Ahmad, A. (2020). Predicting Type 2 Diabetes Using XGBoost: Comparative Study with Traditional Methods. Journal of Biomedical Informatics, 102.
- Al Dhaheri, A. S., Al Ma'awali, A. K., Laleye, L. C., Washi, S. A., Jarrar, A. H., Al Meqbaali, F. T., Mohamad, M. N., & Masuadi, E. M. (2015). The effect of nutritional composition on the glycemic index and glycemic load values of selected Emirati foods. BMC Nutrition, 1(1), 4. https://doi.org/10.1186/2055-0928-1-4
- Almutairi, E. S., & Abbod, M. F. (2023). Machine Learning Methods for Diabetes Prevalence Classification in Saudi Arabia. Modelling, 4(1), 37–55. https://doi.org/10.3390/modelling4010004
- Al-Nozha, M. M., Ismail, H. M., & Al Nozha, O. M. (2016). Coronary artery disease and diabetes mellitus. Journal of Taibah University Medical Sciences, 11(4), 330–338. https://doi.org/10.1016/j.jtumed.2016.03.005
- Alramadan, M. J., Magliano, D. J., Almigbal, T. H., Batais, M. A., Afroz, A., Alramadhan, H. J., Mahfoud, W. F., Alragas, A. M., & Billah, B. (2018). Glycaemic control for people with type 2 diabetes in Saudi Arabia – an urgent need for a review of management plan. BMC Endocrine Disorders, 18(1), 62. https://doi.org/10.1186/s12902-018-0292-9
- American Diabetes Association. (2021). 2. Classification and diagnosis of diabetes:standards of Medical Care in diabetes—2021. Diabetes Care, 44(Supplement_1), S15–S33.
- Augustin, L. S. A., Kendall, C. W. C., Jenkins, D. J. A., Willett, W. C., Astrup, A., Barclay, A. W., Björck, I., Brand-Miller, J. C., Brighenti, F., Buyken, A. E., Ceriello, A., La Vecchia, C., Livesey, G., Liu, S., Riccardi, G., Rizkalla, S. W., Sievenpiper, J. L., Trichopoulou, A., Wolever, T. M. S., … Poli, A. (2015). Glycemic index, glycemic load and glycemic response: An International Scientific Consensus Summit from the International Carbohydrate Quality Consortium (ICQC). Nutr. Metab. Cardiovasc. Dis., 25(9), 795–815.
- Barbagallo, M., Veronese, N., & Dominguez, L. J. (2022). Magnesium in Type 2 Diabetes Mellitus, Obesity, and Metabolic Syndrome. Nutrients, 14(3), 714. https://doi.org/10.3390/nu14030714

- Barclay, A. W., Petocz, P., McMillan-Price, J., Flood, V. M., Prvan, T., Mitchell, P., & Brand-Miller, J. C. (2008). Glycemic index, glycemic load, and chronic disease risk—a meta-analysis of observational studies. Am. J. Clin. Nutr., 87(3), 627–637.
- Bizimana, R. (2024). The Impact of Lifestyle Modifications on Type 2 Diabetes Prevention and Management. IDOSR JOURNAL OF BIOCHEMISTRY, BIOTECHNOLOGY AND ALLIED FIELDS, 9, 18–24. https://doi.org/10.59298/IDOSR/JBBAF/24/93.1824000
- Boushey, C., Ard, J., Bazzano, L., Heymsfield, S., Mayer-Davis, E., Sabaté, J., Snetselaar, L., Van Horn, L., Schneeman, B., English, L., Bates, M., Callahan, E., Butera, G., Terry, N., & Obbagy, J. (2020). Dietary Patterns and Risk of Type 2 Diabetes: A Systematic Review. https://doi.org/10.52570/NESR.DGAC2020.SR0103
- Brand-Miller, J., Hayne, S., Petocz, P., & Colagiuri, S. (2003). Low–glycemic index diets in the management of diabetes. Diabetes Care, 26(8), 2261–2267.
- Brouns, F., Bjorck, I., Frayn, K. N., Gibbs, A. L., Lang V and Slama, G., & Wolever, T. M. S. (2005). Glycaemic index methodology. Nutr. Res. Rev., 18(1), 145–171.
- Chatelan, A., Bochud, M., & Frohlich, K. L. (2019). Precision nutrition: hype or hope for public health interventions to reduce obesity? International Journal of Epidemiology, 48(2), 332–342. https://doi.org/10.1093/ije/dyy274
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- Choi, B. G., Rha, S., Yoon, S. G., Choi, C. U., Lee, M. W., & Kim, S. W. (2019). Association of Major Adverse Cardiac Events up to 5 Years in Patients With Chest Pain Without Significant Coronary Artery Disease in the Korean Population. Journal of the American Heart Association, 8(12). https://doi.org/10.1161/JAHA.118.010541
- Colberg, S. R., Sigal, R. J., Fernhall, B., Regensteiner, J. G., Blissmer, B. J., Rubin Richard R and Chasan-Taber, L., Albright, A. L., & Braun, B. (2010). Exercise and type 2 diabetes. Diabetes Care, 33(12), e147–e167.
- DeFronzo, R. A. (2004). Pathogenesis of type 2 diabetes mellitus. Med. Clin. North Am., 88(4), 787–835.
- Dhana, K., Nano, J., Ligthart, S., Peeters, A., Hofman, A., Nusselder, W., Dehghan, A., & Franco, O. H. (2016). Obesity and Life Expectancy with and without Diabetes in Adults Aged 55 Years and Older in the Netherlands: A Prospective Cohort Study. PLOS Medicine, 13(7), e1002086. https://doi.org/10.1371/journal.pmed.1002086
- El Bilbeisi, A. H., Hosseini, S., & Djafarian, K. (2017). Dietary patterns and metabolic syndrome among type 2 diabetes patients in Gaza Strip, Palestine. Ethiopian Journal of Health Sciences, 27(3), 227. https://doi.org/10.4314/ejhs.v27i3.4
- Florez, J. C. (2008). Newly identified loci highlight beta cell dysfunction as a key cause of type 2 diabetes: Where are the insulin resistance genes? Diabetologia, 51(7), 1100–1110.
- Foster-Powell, K., Holt, S. H. A., & Brand-Miller, J. C. (2002). International table of glycemic index and glycemic load values: 2002. Am. J. Clin. Nutr., 76(1), 5–56.
- Freeman, L. M. (2010). Beneficial effects of omega-3 fatty acids in cardiovascular disease. J. Small Anim. Pract., 51(9), 462–470.
- Fuchsberger, C., Flannick, J., Teslovich Tanya M and Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., Rivas, M.

A., Perry, J. R. B., Sim, X., Blackwell, T. W., Robertson, N. R., Rayner, N. W., Cingolani, P., Locke, A. E., Tajes, J. F., Highland, H. M., … McCarthy, M. I. (2016). The genetic architecture of type 2 diabetes.

- Garcia, F. (2020). Lifestyle and Genetic Factors in Obesity Prediction Using Elastic Net Regularization. Obesity Science & Practice, 6(2), 125–134.
- Goff, L. M., Cowland, D. E., Hooper, L., & Frost, G. S. (2013). Low glycaemic index diets and blood lipids: A systematic review and meta-analysis of randomised controlled trials. Nutr. Metab. Cardiovasc. Dis., 23(1), 1–10.
- Gupta, P. (2024). Diabetes prediction using machine learning. J. Electr. Syst., 20(7s), 2244–2257.
- Hatmal, M. M., Abderrahman, S. M., Nimer, W., Al-Eisawi, Z., Al-Ameer, H. J., Al-Hatamleh, M. A. I., Mohamud, R., & Alshaer, W. (2020). Artificial Neural Networks Model for Predicting Type 2 Diabetes Mellitus Based on VDR Gene FokI Polymorphism, Lipid Profile and Demographic Data. Biology, 9(8), 222. https://doi.org/10.3390/biology9080222
- Hu, F. B. (2011). Globalization of diabetes. Diabetes Care, 34(6), 1249–1257.
- International Diabetes Federation. (2019). IDF Diabetes Atlas (9th ed.). IDF. (n.d.-a).
- International Diabetes Federation. (2019). IDF Diabetes Atlas (9th ed.). IDF. (n.d.-b).
- Jenkins, D. J. A., Kendall, C. W. C., Augustin Livia S A and Franceschi, S., Hamidi, M., Marchie Augustine and Jenkins, A. L., & Axelsen, M. (2002). Glycemic index: overview of implications in health and disease. Am. J. Clin. Nutr., 76(1), 266S–273S.
- Jenkins, D. J., Wolever, T. M., Taylor, R. H., Barker, H., Fielden, H., Baldwin, J. M., Bowling, A. C., Newman, H. C., Jenkins, A. L., & Goff, D. V. (1981). Glycemic index of foods: a physiological basis for carbohydrate exchange. Am. J. Clin. Nutr., 34(3), 362–366.
- Kahn, S. E., Hull, R. L., & Utzschneider, K. M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. Nature, 444(7121), 840–846.
- Kanagarathinam, K., Manikandan, R., & Kumar, T. S. (2024). Machine learning algorithms-based decision support model for diabetes. Review of Computer Engineering Research, 11(1), 16–29. https://doi.org/10.18488/76.v11i1.3598
- Katherine A McGonagle, Robert F Schoeni, Narayan Sastry, & Vicki A Freedman. (2012). The Panel Study of Income Dynamics: overview, recent innovations, and potential for life course research. Longitudinal and Life Course Studies, 3(2). https://doi.org/10.14301/llcs.v3i2.188
- Kavakiotis, I., Tsave, O., Salifoglou Athanasios and Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). Machine learning and data mining methods in diabetes research. Comput. Struct. Biotechnol. J., 15, 104–116.
- Kearns, K., Dee, A., Fitzgerald, A. P., Doherty, E., & Perry, I. J. (2014). Chronic disease burden associated with overweight and obesity in Ireland: the effects of a small BMI reduction at population level. BMC Public Health, 14(1).
- Kim, D., & Saada, A. (2013). The social determinants of infant mortality and birth outcomes in western developed nations: A cross-country systematic review. Int. J. Environ. Res. Public Health, 10(6), 2296–2335.
- Kim, M., & Choi, H. (2021). Obesity Prediction Using CatBoost with Dietary and Lifestyle Data. International Journal of Obesity, 45(9), 1853–1861.

- Knowler, Barrett-Connor, & Fowler. (2002). Reduction in the Incidence of Type 2 Diabetes with Lifestyle Intervention or Metformin. New England Journal of Medicine, 346(6), 393–403. https://doi.org/10.1056/NEJMoa012512
- Knutson, K. L. (2006). Role of sleep duration and quality in the risk and severity of type 2 diabetes mellitus. Arch. Intern. Med., 166(16), 1768.
- Lazarou, C., Panagiotakos, D., & Matalas, A.-L. (2012). The Role of Diet in Prevention and Management of Type 2 Diabetes: Implications for Public Health. Critical Reviews in Food Science and Nutrition, 52(5), 382–389. https://doi.org/10.1080/10408398.2010.500258
- Lee, J. (2020). Metabolic Syndrome Prediction Using CatBoost: A Categorical and Continuous Data Approach. BMC Medical Informatics and Decision Making, 20.
- Ley, S. H., Hamdy, O., Mohan, V., & Hu, F. B. (2014). Prevention and management of type 2 diabetes: dietary components and nutritional strategies. Lancet, 383(9933), 1999–2007.
- Li, Y., Yang, C., Zhang, H., & Jia, C. (2020). A model combining Seq2Seq network and LightGBM algorithm for industrial soft sensor. IFAC-PapersOnLine, 53(2), 12068–12073.
- Lin, Y. (2022). Automated Diabetes Risk Prediction Using AutoGluon: A Cohort Study. Journal of Clinical Endocrinology & Metabolism, 107(1), e12–e20.
- Lundberg, U. (2002). Psychophysiology of work: Stress, gender, endocrine response, and work-related upper extremity disorders*. American Journal of Industrial Medicine, 41(5), 383–392. https://doi.org/10.1002/ajim.10038
- Lyssenko, V., Almgren, P., Anevski, D., Perfekt, R., Lahti, K., Nissén, M., Isomaa, B., Forsen, B., Homström, N., Saloranta, C., Taskinen, M.-R., Groop, L., Tuomi, T., & for the Botnia Study Group. (2005). Predictors of and longitudinal changes in insulin sensitivity and secretion preceding onset of type 2 diabetes. Diabetes, 54(1), 166–174.
- Marmot, M., Friel, S., Bell, R., Houweling, T. A. J., & Taylor, S. (2008). Closing the gap in a generation: health equity through action on the social determinants of health. Lancet, 372(9650), 1661–1669.
- Mendes-Soares, H., Raveh-Sadka, T., Azulay Shahar and Ben-Shlomo, Y., Cohen, Y., Ofek, T., Stevens, J., Bachrach, D., Kashyap, P., Segal, L., & Nelson, H. (2019). Model of personalized postprandial glycemic response to food developed for an Israeli cohort predicts responses in Midwestern American individuals. Am. J. Clin. Nutr., 110(1), 63–75.
- Micha, R., Wallace, S. K., & Mozaffarian, D. (2010). Red and processed meat consumption and risk of incident coronary heart disease, stroke, and diabetes mellitus: A systematic review and meta-analysis. Circulation, 121(21), 2271–2283.
- Mousa, K. M., Mousa, F. A., Mohamed, H. S., & Elsawy, M. M. (2023). Prediction of Foot Ulcers Using Artificial Intelligence for Diabetic Patients at Cairo University Hospital, Egypt. SAGE Open Nursing, 9. https://doi.org/10.1177/23779608231185873
- Ordovas, J. M., Ferguson, L. R., Tai, E. S., & Mathers, J. C. (2018). Personalised nutrition and health. BMJ, bmj.k2173.
- Powles, J., Fahimi, S., Micha, R., Khatibzadeh, S., Shi, P., Ezzati, M., Engell Rebecca E and Lim, S. S., Danaei, G., Mozaffarian, D. and on behalf of the G. B. of D. N., & (NutriCoDE), C. D. E. G. (2013). Global, regional and national sodium intakes in 1990 and 2010: a systematic analysis of 24 h urinary sodium excretion and dietary surveys worldwide. BMJ Open, 3(12), e003733.

- Reutrakul, S., & Van Cauter, E. (2018). Sleep influences on obesity, insulin resistance, and risk of type 2 diabetes. Metabolism, 84, 56–66.
- Reynolds, A., Mann, J., Cummings, J., Winter, N., Mete, E., & Te Morenga, L. (2019). Carbohydrate quality and human health: a series of systematic reviews and meta-analyses. Lancet, 393(10170), 434–445.
- Salmerón, J., Ascherio, A., Rimm, E. B., Colditz G A and Spiegelman, D., Jenkins, D. J., Stampfer, M. J., Wing, A. L., & Willett, W. C. (1997). Dietary fiber, glycemic load, and risk of NIDDM in men. Diabetes Care, 20(4), 545–550.
- Schulze, M. B., Schulz, M., Heidemann, C., Schienkiewitz, A., Hoffmann, K., & Boeing, H. (2007). Fiber and magnesium intake and incidence of type 2 diabetes: a prospective study and meta-analysis: A prospective study and meta-analysis. Arch. Intern. Med., 167(9), 956–965.
- Shickel, B., Tighe, P. J., Bihorac, A., & Rashidi, P. (2018). Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J. Biomed. Health Inform., 22(5), 1589–1604.
- Singh, A., & Verma, R. (2021). Assessing Hypertension Risk in Obesity Using TabNet's Attention Mechanism. Journal of Hypertension, 39(6), 1189–1195.
- Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nat. Med., 25(1), 44–56.
- Torres, P. (2021). AutoGluon in Nutrition Studies: Predicting Dietary Patterns and Blood Pressure. Nutrition Research, 89, 49–58.
- Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, E. W., & Collins, G. S. (2019). Predictive analytics in health care: how can we know it works? Journal of the American Medical Informatics Association, 26(12), 1651–1654. https://doi.org/10.1093/jamia/ocz130
- Varshney, N., Jadhav, N., Gupta, K., Mate, N. R., Rose, A., & Kumar, P. (2023). Personalized Dietary Recommendations Using Machine Learning: A Comprehensive Review. 2023 International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), 1–6. https://doi.org/10.1109/ICAIIHI57871.2023.10489126
- Walker, R. J., Smalls, B. L., Hernandez-Tejada, M. A., Campbell, J. A., Davis, K. S., & Egede, L. E. (2012). Effect of diabetes fatalism on medication adherence and self-care behaviors in adults with diabetes. Gen. Hosp. Psychiatry, 34(6), 598–603.
- Wang, L. (2019). Elastic Net Regularization for Biomarker Selection in Type 2 Diabetes. Diabetologia, 62(10), 1845–1855.
- Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes. Diabetes Care, 27(5), 1047–1053.
- Wong, C. Y. (2019). Childhood Obesity Prediction Using Machine Learning: XGBoost with Lifestyle and Environmental Data. Pediatric Obesity, 14(3).
- Wu, J. H. Y., Lemaitre, R. N., King, I. B., Song, X., Sacks, F. M., Rimm, E. B., Heckbert, S. R., Siscovick, D. S., & Mozaffarian, D. (2012). Association of Plasma Phospholipid Long-Chain Omega-3 Fatty Acids With Incident Atrial Fibrillation in Older Adults. Circulation, 125(9), 1084–1093. https://doi.org/10.1161/CIRCULATIONAHA.111.062653
- Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How Powerful are Graph Neural Networks? https://doi.org/10.48550/arXiv.1810.00826
- Yu, L. (2020). Dietary Risk Factors and Cardiovascular Disease: Application of LightGBM in Predictive Modeling. Nutrition & Health, 26(3), 163–172.

- Zalan, A., & Sharkia, R. (2019). Type 2 Diabetes Mellitus (T2DM) in the Arab Society of Israel. In Handbook of Healthcare in the Arab World (pp. 1–32). Springer International Publishing. https://doi.org/10.1007/978-3-319-74365-3_162-1
- Zeevi, D., Korem, T., Zmora, N., Israeli, D., Rothschild, D., Weinberger, A., Ben-Yacov, O., Lador, D., Avnit-Sagi, T., Lotan-Pompan, M., Suez, J., Mahdi, J. A., Matot, E., Malka, G., Kosower, N., Rein, M., Zilberman-Schapira, G., Dohnalová, L., Pevsner-Fischer, M., … Segal, E. (2015). Personalized nutrition by prediction of glycemic responses. Cell, 163(5), 1079–1094.
- Zhang, T. (2019). Dietary Risk Factors and Cardiovascular Disease: Application of LightGBM in Predictive Modeling. Journal of Clinical Endocrinology & Metabolism, 44(2), 125–134.
- Zhao, X. (2021). Predicting Diabetes Progression Using TabNet: Integrating Genetic and Lifestyle Data. Journal of Medical Internet Research, 23(4).
- Zheng, Y., Ley, S. H., & Hu, F. B. (2018). Global aetiology and epidemiology of type 2 diabetes mellitus and its complications. Nat. Rev. Endocrinol., 14(2), 88–98.
- Zoccali, C., Mallamaci, F., Adamczak, M., de Oliveira, R. B., Massy, Z. A., Sarafidis, P., Agarwal, R., Mark, P. B., Kotanko, P., Ferro, C. J., Wanner, C., Burnier, M., Vanholder, R., & Wiecek, A. (2023). Cardiovascular complications in chronic kidney disease: a review from the European Renal and Cardiovascular Medicine Working Group of the European Renal Association. Cardiovascular Research, 119(11), 2017–2032. https://doi.org/10.1093/cvr/cvad083

الاستفادة من مؤشر نسبة السكر في الدم والتعلم الآلي للنمذجة التنبؤية لظهور مرض السكري من النوع 2 وإدارته

إعداد: جيهان موسى محمد رحال

المشرف: د. رضوان قصراوي

الملخص

مرض السكري من النوع 2 هو مرض مزمن متعدد الأوجه يتأثر بمجموعة من العوامل الاجتماعية والديموغرافية ونمط الحياة والتغذية. هدفت هذه الدراسة إلى تحليل انتشار ومحددات مرض السكري في مجموعة سكانية متنوعة مع الاستفادة من التعلم الآلي لتعزيز النمذجة التنبؤية. ارتبطت الخصائص الاجتماعية والديموغرافية مثل الإقامة في المناطق الحضرية ومستويات التعليم المنخفضة والشيخوخة والجنس الأنثوي بشكل كبير بارتفاع انتشار مرض السكري. كانت عوامل نمط الحياة، بما في ذلك انخفاض النشاط البدني (94.1٪ بين مرضى السكري) والسمنة (75.3٪)، من المساهمين البارزين. كشف التحليل الغذائي عن انتشار مرتفع لنقص النظام الغذائي بين مرضى السكري، بما في ذلك انخفاض تناول الألياف والأحماض الدهنية المتعددة غير المشبعة، إلى جانب الاستهلاك المفرط للصوديوم. تسلط هذه النتائج الضوء على أهمية معالجة جودة النظام الغذائي وتعزيز النشاط البدني كمكونات رئيسية للوقاية من مرض السكري.

تم استخدام نماذج التعلم الآلي للتنبؤ بمخاطر الإصابة بمرض السكري، حيث حقق نموذج Gradient Boosting أعلى دقة بلغت 94.2% وأعلى قيمة AUC بلغت 0.985، متفوقًا على نماذج أخرى مثل Random Forest وSupport Vector Machines. وأظهرت تحليلات أهمية الميزات أن الحمل الجلايسيمي، ومؤشر كتلة الجسم (BMI)، والعمر هي أهم العوامل المؤثرة في التنبؤ بالإصابة بالسكري، مما يبرز أهمية التدابير الغذائية، وتكوين الجسم، وعملية الشيخوخة في خطر الإصابة بالمرض. تعزز هذه النتائج إمكانات التعلم الآلي كأداة للتشخيص المبكر وتحديد مستويات الخطورة.

تظهر هذه الدراسة الحاجة إلى استراتيجيات متكاملة ومتعددة التخصصات لمكافحة مرض السكري، بما في ذلك التعليم والتدخلات الغذائية وتعزيز النشاط البدني. وتوفر النتائج أساسًا للبحوث المستقبلية للتحقق من صحة هذه الارتباطات وتطوير برامج الوقاية المستهدفة. ويوضح تطبيق التعلم الآلي طريقًا واعداً لحلول الرعاية الصحية الشخصية في إدارة وتقليل عبء مرض السكري من النوع 2.