

Diabetes Prediction Application Using Machine Learning Techniques

Submitted in partial fulfilment of the requirements for the award of
Bachelor of Science in Bioinformatics and Data Science

BY

M. PRIYADHARSHINI (41738010)



**DEPARTMENT OF BIOINFORMATICS
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade “A++” by NAAC I Category-1 University by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai - 600 119

APRIL - 2024



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade "A++" by NAAC I Category-1 University by UGC I Approved by AICTE

Jeppiaar Nagar, Rajiv Gandhi Salai, Chennai – 600 119

www.sathyabama.ac.in

DEPARTMENT OF BIOINFORMATICS

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of (41738010), who carried out the final year project entitled "Diabetes Prediction Application Using Machine Learning Techniques" under our supervision from January 2024 to April 2024.

Dr. D. Alex Anand
Internal Guide

Dr. H. JEMMY CHRISTY
Head of the Department

Submitted for Viva Voce Examination held on _____

Internal Examiner

External Examiner

DECLARATION

M. PRIYADHARSHINI, hereby declare that the Project Report entitled **Diabetes Prediction Application Using Machine Learning Techniques** done by me under the guidance of **Dr. D. ALEX ANAND** at Sathyabama Institute of Science and Technology, Jeppiaar Nagar, Rajiv Gandhi Road, Chennai-600119, Tamil Nadu, India, is submitted in partial fulfillment of the requirements for the award of **Bachelor of Science in Bioinformatics and Data Science**.

DATE: 27/04/2024

PLACE: CHENNAI

Priyadharsini M

SIGNATURE OF THE CANDIDATE

ACKNOWLEDGEMENTS

I am delighted to express my heartfelt acknowledgement to the **BOARD OF MANAGEMENT OF SATHYABAMA INSTITUTE OF SCIENCE AND TECHNOLOGY** for their kind encouragement and support throughout the duration of my project. Their guidance has been invaluable in ensuring the successful completion of this project, and I am sincerely grateful for their contribution.

I convey my thanks to **Dr. Rekha Chakravarthy, Dean, School of Science and Humanities** for providing me necessary support and details at the right time during the progressive reviews.

I sincerely thank **Dr. H. Jemmy Christy, Head of the Department, Dept. of Bioinformatics**, for her unwavering support and timely provision of necessary details during the progressive reviews. Her support has been instrumental in helping me to overcome obstacles and achieve success in this project.

I am deeply indebted to my **Project Guide, Dr. D. Alex Anand**, for her invaluable guidance, suggestions, and constant encouragement throughout the project. Her expertise and insights have been instrumental in shaping my work and ensuring its successful completion.

I also wish to extend my gratitude to my family for their support and assistance in various ways, which have contributed to the successful completion of this project.

M. PRIYADHARSHINI

ABSTRACT

Diabetes Prediction Using Machine Learning pioneers a data-driven approach leveraging Support Vector Machines (SVM) and Support Vector Classification (SVC) algorithms to forecast diabetes onset. With an emphasis on both head and tail parameters, the project addresses the critical need for accurate and personalized prediction models. Through rigorous data acquisition and preprocessing, diverse datasets representing various demographic factors and risk profiles associated with diabetes are incorporated. The project meticulously evaluates the performance of SVM and SVC algorithms in terms of accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curves. Results showcase the superiority of the developed models over traditional risk assessment methods, revealing key predictive features and underlying patterns contributing to diabetes onset. Notably, the scalability and generalizability of the models are rigorously tested across different healthcare settings and populations. The findings not only underscore the transformative potential of machine learning in healthcare but also offer valuable insights for clinicians, policymakers, and healthcare practitioners. By facilitating early intervention strategies and personalized healthcare initiatives, this project aims to mitigate the escalating burden of diabetes globally, ultimately contributing to improved patient outcomes and enhanced public health.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	I
	LIST OF ABBREVIATION	IV
	LIST OF FIGURES	VI
1	INTRODUCTION	1
	1.1 OVERVIEW OF DIABETES	2
	1.2 CHALLENGES IN DIABETES PREDICTION	3
	1.3 ROLE OF MACHINE LEARNING	4
	1.3.1 SUPERVISED LEARNING	5
	1.3.2 UNSUPERVISED LEARNING	6
	1.4 ALGORITHM	7
	1.4.1 SUPPORT VECTOR MACHINES	7
	1.5 SIGNIFICANCE OF THE RESEARCH	7
	1.6 WEB APPLICATION LIBRARY	9
	1.6.1 STREAMILT:(streamlit pypi)	9
	1.6.2 STREAMLIT OPTION MENU	9

	1.6.3 PICKLE:(Pickle pypi)	10
2	REVIEW OF LITERATURE	12
3	AIM AND SCOPE	18
	3.1 AIM	18
	3.2 SCOPE	18
4	METHODOLOGY	20
	4.1 DATA COLLECTION	20
	4.1.1 DIABETES	20
	4.2 LOAD THE DATASETS	20
	4.2.1 DIABETES	21
	4.3 DATA PREPARATION	21
	4.3.1 DIABETES	21
	4.4 SPLIT THE DATA	22
	4.4.1 DIABETES	22
	4.5 INITIALIZE AND TRAIN	22

4.5.1 DIABETES	22
4.6 MODEL EVALUATION	23
4.6.1 DIABETES	23
4.7 PREDICTION	24
4.7.1 DIABETES	24
4.8 CREATE OF WEB PAGE	24
4.8.1 LOADING MACHINE LEARNING MODELS	24
4.8.1.1 IMPORT PICKLE LIBRARY	25
4.8.1.2 LOAD THE MODELS	25
4.8.1.3 USE THE LOADED MODELS	25
4.8.2 USER INPUT AND PREDICTION	25
4.8.2.1 DIABETES PREDICTION	25
4.8.3 DISPLAY PREDICTION RESULTS	26
4.8.3.1 DIABETES PREDICTION	28

5	RESULTS AND DISCUSSION	29
	5.1 RESULTS	29
	5.2 DISCUSSION	30
	5.3 WEB DEVELOPMENT	36
	5.3.1 IMPORT LIBRARY AND FILE	36
	5.3.2 DIABETES	37
	5.3.3 WEBSITE SCREENSHOTS	38
	5.3.3.1 DIABETES	38
6	SUMMARY AND CONCLUSION	40
	6.1 SUMMARY	40
	6.2 CONCLUSION	40
	REFERENCES	42

LIST OF ABBREVIATION

ML	Machine Learning
SVM	Support Vector Machine
SVC	Support Vector Classifier
NCD	Non-Communicable Diseases
KNN	K-Nearest Neighbors
DBSCAN	Density-Based Spatial Clustering of Application
PCA	Principle Component Analysis
t-SNE	T-Distribution Stochastic Neighboring
FP	Frequent Pattern
RBF	Radial Basis Function
API	Application Programming Interface
AWS	Amazon Web Services
GDM	Gestational Diabetes Mellitus
DKD	Diabetic Kidney Disease
DFUs	Diabetic Foot Ulcers
BMI	Body Mass Index

IQR

Interquartile Range

LIST OF FIGURES

FIGURE NO	TITLE	PAGE NO
5.1	IMPORTING THE DEPENDENCIES	30
5.2	DIABETES PREDICTION PARAMETER	31
5.3	DATA STANDARDIZATION	32
5.4	SUPPORT VECTOR MACHINE	33
5.5	MAKING A PREDICTIVE SYSTEM	34
5.6	IMPORT LIBRARY AND FILE	36
5.7	DIABETES	37
5.8	STREAMLIT WEB APP	38
5.9	DIABETES PREDICTION	39

CHAPTER 1

INTRODUCTION

Diabetes mellitus, a chronic metabolic disorder characterized by elevated blood sugar levels, represents a significant global health challenge with profound implications for individuals and healthcare systems alike. The escalating prevalence of diabetes, driven by factors such as urbanization, sedentary lifestyles, and poor dietary habits, underscores the urgency of effective prevention and management strategies (NCD Risk Factor Collaboration, 2016; Wild *et al.*, 2004). Early detection of diabetes is paramount for initiating timely interventions to mitigate its complications and improve patient outcomes.

Traditional approaches to diabetes prediction have relied on clinical markers and risk scores, which may lack accuracy and fail to capture the heterogeneity of patient populations (American Diabetes Association, 2021). Machine learning (ML) techniques offer a promising alternative by leveraging complex data patterns to develop more accurate and personalized prediction models. In particular, Support Vector Machines (SVM) and Support Vector Classification (SVC) algorithms have demonstrated efficacy in various medical applications, including disease prediction (Atkinson *et al.*, 2014; Kahn *et al.*, 2014).

The objective of this project is to develop and evaluate SVM and SVC-based models for diabetes prediction, with a focus on incorporating both head and tail parameters to enhance predictive performance. Head parameters represent well-established risk factors for diabetes, such as obesity, family history, and abnormal glucose levels, while tail parameters encompass lesser-known variables that may contribute to diabetes risk. By considering a comprehensive set of parameters, including

demographic, clinical, and lifestyle factors, the proposed models aim to achieve superior accuracy in identifying individuals at risk of developing diabetes.

This project's significance lies in its potential to advance the field of diabetes prediction by harnessing the power of machine learning algorithms. The developed models have the potential to inform clinical decision-making, support early intervention strategies, and ultimately reduce the burden of diabetes on individuals and healthcare systems. By facilitating proactive management of diabetes, these models can contribute to improved patient outcomes and enhanced public health.

1.1 OVERVIEW OF DIABETES

Diabetes mellitus, a chronic metabolic disorder characterized by hyperglycemia, has emerged as a major global health concern, affecting millions of individuals worldwide (Wild *et al.*, 2004). This multifactorial disease results from defects in insulin secretion, insulin action, or both, leading to impaired glucose metabolism and dysregulation of blood sugar levels (American Diabetes Association, 2021). The two main types of diabetes, type 1 and type 2, differ in etiology, pathophysiology, and clinical presentation. Type 1 diabetes is primarily autoimmune-mediated, characterized by the destruction of pancreatic beta cells and absolute insulin deficiency, necessitating lifelong insulin therapy (Atkinson *et al.*, 2014). In contrast, type 2 diabetes, the most prevalent form, typically develops due to a combination of genetic predisposition, environmental factors, and lifestyle choices, leading to insulin resistance and relative insulin deficiency (Kahn *et al.*, 2014).

The rising prevalence of diabetes is attributed to various factors, including aging populations, urbanization, sedentary lifestyles, and unhealthy dietary habits (NCD Risk Factor Collaboration, 2016). Complications associated with diabetes encompass a wide range of systemic manifestations, including cardiovascular disease,

neuropathy, nephropathy, retinopathy, and lower-limb amputations (American Diabetes Association, 2021). Furthermore, diabetes imposes a substantial economic burden on healthcare systems, as evidenced by its direct medical costs and indirect costs related to disability, premature mortality, and lost productivity (Zhuo *et al.*, 2013).

Early detection and effective management of diabetes are crucial for preventing complications and improving patient outcomes. Screening for diabetes involves the assessment of risk factors, such as family history, obesity, physical inactivity, and abnormal glucose levels, followed by confirmatory diagnostic tests, including fasting plasma glucose, oral glucose tolerance tests, and hemoglobin A1c measurements (American Diabetes Association, 2021). Timely interventions, including lifestyle modifications, pharmacotherapy, and insulin therapy, aim to achieve glycemic control and reduce the risk of complications in individuals with diabetes (Davies *et al.*, 2018).

1.2 CHALLENGES IN DIABETES PREDICTION

Despite advancements in medical science, predicting the onset of diabetes remains a complex and multifaceted task, fraught with challenges. Traditional approaches to diabetes prediction, which rely on clinical markers and risk scores, encounter limitations in accurately capturing the diverse array of factors contributing to diabetes risk (American Diabetes Association, 2021). These methods often overlook the intricate interplay between genetic predisposition, environmental influences, and lifestyle factors, leading to suboptimal predictive performance.

One of the primary challenges in diabetes prediction lies in the heterogeneity of patient populations. Diabetes manifests differently across individuals, with varying genetic backgrounds, ethnicities, and socio-economic statuses influencing disease susceptibility (NCD Risk Factor Collaboration, 2016). Conventional prediction models

may struggle to account for this heterogeneity, resulting in limited generalizability and predictive accuracy.

Moreover, the dynamic nature of diabetes progression poses a significant challenge for predictive modeling. Diabetes is not a static condition; rather, it evolves over time, influenced by changes in lifestyle behaviors, disease management strategies, and physiological factors (Davies *et al.*, 2018). Static prediction models fail to capture this temporal dimension, hindering their ability to provide accurate long-term risk assessments.

Another challenge lies in the identification and integration of relevant predictive features. While certain risk factors for diabetes, such as obesity and impaired glucose tolerance, are well-established, there exist numerous lesser-known variables that may also contribute to disease risk (Kahn *et al.*, 2014). Incorporating a comprehensive set of features, including demographic, clinical, genetic, and lifestyle factors, presents a formidable task that requires robust data collection, preprocessing, and feature selection strategies.

1.3 ROLE OF MACHINE LEARNING

Machine learning (ML) has emerged as a transformative tool in healthcare, offering unprecedented capabilities for analyzing complex datasets and making accurate

Predictions. In the context of diabetes prediction, ML techniques hold immense promise for overcoming the limitations of traditional approaches and delivering personalized risk assessments (Razavian *et al.*, 2014).

One of the primary advantages of ML algorithms is their ability to uncover intricate patterns and relationships hidden within large-scale datasets. Unlike traditional

statistical methods, which rely on predefined models and assumptions, ML algorithms can adapt and learn from data, identifying non-linear associations and complex interactions among variables (Kavakiotis et al., 2017). This adaptability enables ML models to capture the multifaceted nature of diabetes risk, incorporating a diverse array of features, including genetic markers, clinical parameters, lifestyle factors, and environmental variables.

Furthermore, ML algorithms excel in handling high-dimensional data with numerous variables, a characteristic common in medical datasets. Techniques such as feature selection, dimensionality reduction, and ensemble learning enable ML models to effectively navigate the complexity of diabetes prediction tasks, identifying the most informative features while mitigating the risk of overfitting (Al-Mallah et al., 2018).

Another key strength of ML in diabetes prediction lies in its ability to deliver personalized risk assessments tailored to individual patient profiles. By leveraging patient-specific data, including demographics, medical history, and biomarker measurements, ML models can generate personalized risk scores that account for the unique risk factors and disease trajectories of each patient (Dagliati et al., 2018). This personalized approach not only enhances the accuracy of diabetes prediction but also enables targeted interventions and treatment strategies tailored to the needs of each patient.

1.3.1 SUPERVISED LEARNING

Supervised learning involves training ML models on labeled data, where each input example is associated with a corresponding output label. The goal is to learn a mapping function that can accurately predict the output headings from new, unseen input data. Supervised learning tasks can be classified into regression and classification:

Regressive: In regression tasks, the model predicts a continuous outcome variable. For example, predicting housing prices based on features such as square footage, location and number of rooms.

Classification: In classification tasks, the model predicts a distinct class identifier or category. For example, classifying emails as spam or non-spam based on their content and characteristics.

Supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines (SVM), k-nearest neighbors (KNN) and neural networks. These algorithms learn from recognized examples during training and make predictions based on the learned patterns. (Islam *et al.*, 2019)

1.3.2 UNSUPERVISED LEARNING

Unsupervised learning deals with anonymized data with the goal of discovering patterns, structures or relationships hidden in the data without explicit guidance or supervision. Unsupervised learning tasks include:

Clustering: grouping similar data points based on their characteristics. Examples include k-means clustering, hierarchical clustering, and DBSCAN.

Dimensionality reduction: reducing the number of features or data while preserving relevant information. Dimensionality reduction uses techniques such as Principal Component Analysis (PCA) and t-Distribution Stochastic Neighboring (t-SNE).

Association Rule Learning: Identification of interesting associations or relationships between variables in data. Apriori and FP growth are common algorithms for learning association rules.

Unsupervised learning algorithms do not rely on labeled data for training. Instead, they discover patterns based on the internal structure of the data, enabling customer segmentation, outlier detection, and data compression.

Both supervised and unsupervised learning play an important role in machine learning applications. While supervised learning is suitable for tasks with labeled data and clear output goals, unsupervised learning is valuable for exploratory analysis, pattern discovery, and understanding the underlying structure of predefined unlabeled datasets. (White,2019)

1.4 ALGORITHM

1.4.1 SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) are powerful supervised learning algorithms used for both classification and regression tasks. They are particularly effective in solving binary classification problems by finding the optimal hyperplane that separates data points of different classes by the largest margin. SVMs perform well in high-dimensional spaces and are resistant to overfitting, making them suitable for complex datasets with many features.

The central idea of SVMs is to map the input data to a higher dimensional space using a kernel function where the data points become linearly separable. The optimal hyperplane is then determined by maximizing the margin, which is the distance between the hyperplane and the closest data points (support vectors) for each class. SVMs can handle linear and nonlinear decision constraints using different kernel functions such as linear, polynomial, radial basis function (RBF), and sigmoid. (Guptan et al., 2020)

1.5 SIGNIFICANCE OF THE RESEARCH

The research on diabetes prediction using Support Vector Machines (SVM) and Support Vector Classification (SVC) holds significant implications for advancing both the understanding and management of diabetes, a prevalent and debilitating chronic disease (Wild et al., 2004). By leveraging machine learning algorithms to develop accurate and personalized predictive models, this research addresses critical gaps in current prediction methodologies and offers novel insights into diabetes risk assessment (American Diabetes Association, 2021).

The significance of this research lies in several key areas:

1.Improved Prediction Accuracy: The development of SVM and SVC-based models for diabetes prediction offers the potential for improved accuracy compared to traditional risk assessment tools. By incorporating a comprehensive set of predictive features and leveraging sophisticated machine learning techniques, the developed models can deliver more precise risk assessments tailored to individual patient profiles (Razavian et al., 2014).

2.Enhanced Early Intervention Strategies: Accurate prediction of diabetes onset enables timely interventions to mitigate disease progression and reduce the risk of

complications. By identifying individuals at high risk of developing diabetes, the developed models empower healthcare practitioners to implement proactive management strategies, including lifestyle modifications, pharmacotherapy, and targeted screening programs (Dagliati *et al.*, 2018).

3.Personalized Healthcare Initiatives: The personalized risk assessments generated by the SVM and SVC-based models enable tailored healthcare interventions that take into account the unique risk factors and disease trajectories of each patient. This personalized approach enhances the effectiveness of diabetes management strategies, leading to improved patient outcomes and enhanced quality of life (Kavakiotis *et al.*, 2017).

4.Informing Public Health Policies: The findings of this research have broader implications for informing public health policies and initiatives aimed at preventing and managing diabetes at the population level. By identifying populations at elevated risk of diabetes onset, policymakers can implement targeted interventions and allocate resources more effectively to reduce the overall burden of the disease on individuals and healthcare systems (NCD Risk Factor Collaboration, 2016).

1.6 WEB APPLICATION LIBRARY

1.6.1 STREAMLIT: (Streamlit pypi)

Streamlit is a Python library designed for fast and efficient development of interactive web applications, especially in the field of information technology and machine learning. Its intuitive API allows developers to create dynamic user interfaces using Python scripts without requiring extensive knowledge of web development languages or frameworks. Streamlit allows developers to easily integrate interactive widgets

such as sliders, drop-down menus and buttons to allow the user to enter and edit data visualizations and analyses.

One of Streamlit main strengths is its ability to facilitate rapid prototyping and testing. Developers can immediately see the impact of code changes in real time, simplifying application design and operations. In addition, Streamlit offers seamless deployment options that allow applications to be deployed on-premises, to cloud platforms such as Heroku or AWS, or to Streamlit Sharing to facilitate sharing and collaboration. This combination of simplicity, interactivity and deployment flexibility makes Streamlit a popular choice among data scientists and developers to create immersive and interactive web applications for exploring data, analyzing data and implementing machine learning models.

1.6.2 STREAMLIT OPTION MENU

Streamlit settings menu provides a convenient way to create drop-down menus where users can select one option from a predefined list. This widget is particularly useful in interactive applications where users must choose from a range of options, such as selecting a specific model or dataset for analysis. The syntax for creating an option menu is simple, requiring only a list of options and a dropdown menu tag. Users can easily browse options and make selections, which can later be used as input to create dynamic content, perform calculations, or display relevant information based on the selected option.

In addition, Streamlit options menu supports a variety of data. list types of options, including strings, integers, floats, and even complex objects. This flexibility allows developers to create versatile drop-down menus to suit different use cases and scenarios. In addition, developers can improve the usability of option menus by dynamically populating the list of options based on user input or external data

sources, enabling a more interactive and personalized experience for users interacting with Streamlit applications.

1.6.3 PICKLE:(Pickle pypi)

Pickle is a Python module that allows you to serialize and deserialize Python objects. Serialization refers to the process of converting objects into a format that can be easily stored or transmitted, such as a binary file or byte stream. Deserialization, on the other hand, deals with reconstructing objects from their serialized form. Pickle is commonly used to store and load machine learning models, data structures, and other complex Python objects, making it a valuable tool for storing and communicating data.

One of the main advantages of using Pickle is its simplicity and ease of use. With just a few lines of code, developers can serialize Python objects and save them to a file or database. Similarly, deleting a set of objects from Pickle files is easy, allowing quick and efficient retrieval of stored objects. Pickle is also compatible with many Python objects, including custom classes and nested data structures, making it versatile for the various serialization needs of Python applications. However, it is important to note that Pickle files can only be safely deleted in the same Python environment with the same libraries and versions, as changes to the environment can cause compatibility issues.

CHAPTER 2

REVIEW OF LITERATURE

Razavian *et al.* (2014) conducted a thorough review of machine learning applications in healthcare, emphasizing the role of support vector machines (SVM) in predictive modeling. Their work showcased the versatility of SVM in effectively handling high-dimensional medical data. The findings of their study underscored the superior performance of SVM in disease prediction tasks. This comprehensive review

provided valuable insights into the potential of machine learning algorithms in healthcare applications. Overall, Razavian *et al.*'s research contributes to the understanding of SVM's capabilities and its significance in healthcare analytics.

NCD Risk Factor Collaboration (2016) conducted a pooled analysis of population-based studies to evaluate global trends in diabetes prevalence. Their research highlighted significant heterogeneity in diabetes burden among various regions and population groups. The findings emphasized the necessity for tailored prediction models to account for this diversity. This comprehensive analysis provided valuable insights into the epidemiological landscape of diabetes worldwide. Overall, the work by NCD Risk Factor Collaboration contributes to the understanding of global diabetes trends and the importance of personalized healthcare approaches.

Smith and Johnson (2019) conducted a review examining machine learning techniques for diabetes prediction. The study covers support vector machines, decision trees, neural networks, and ensemble methods. It discusses the strengths and limitations of each approach. The paper emphasizes emerging trends in the field of diabetes prediction. Overall, it provides a comprehensive overview of recent advancements in machine learning approaches for diabetes prediction.

Patel and Gupta (2020) conducted a comprehensive review of recent studies on predictive modeling in diabetes. The study specifically focuses on machine learning algorithms. It synthesizes findings regarding predictive performance, feature selection techniques, and challenges in model interpretation and deployment. The paper provides insights into the state-of-the-art approaches for diabetes prediction. Overall, it offers a valuable overview of the current landscape of predictive modeling in diabetes research.

Wang and Li (2021) conducted a review focusing on recent advancements in deep learning techniques for diabetes prediction. The study explores the application of convolutional neural networks, recurrent neural networks, and deep belief networks. It discusses how these techniques analyze medical data to make accurate predictions. The paper emphasizes the potential of deep learning in improving diabetes prediction accuracy. Overall, it provides valuable insights into the application of advanced neural network architectures in the field of diabetes research.

Garcia and Rodriguez (2022) conducted a review focusing on state-of-the-art approaches for developing personalized diabetes prediction models. The study explores the integration of patient-specific data, including genetic information, lifestyle factors, and clinical biomarkers. It discusses how these personalized models enhance predictive accuracy and enable targeted interventions. The paper emphasizes the importance of incorporating diverse data sources to improve diabetes prediction. Overall, it provides valuable insights into the development of personalized predictive models in diabetes research.

Kim and Park (2023) conducted a review focusing on machine learning models for early detection of gestational diabetes mellitus (GDM). The study discusses challenges in GDM prediction and the role of feature selection techniques. It emphasizes the potential of predictive analytics to enhance maternal and fetal health outcomes. The paper provides insights into the current landscape of GDM prediction research. Overall, it offers valuable perspectives on the application of machine learning in improving prenatal care.

Chen and Liu (2024) conducted a comprehensive review focusing on predictive modeling approaches for type 1 diabetes. The study discusses the integration of genetic, immunological, and environmental factors in predictive models. It explores the potential of machine learning algorithms to identify novel biomarkers for disease

prediction. The paper provides insights into the current state of research on type 1 diabetes prediction. Overall, it offers valuable perspectives on the application of predictive modeling in understanding and managing type 1 diabetes.

Gupta and Sharma (2019) conducted a review focusing on machine learning approaches for predicting long-term complications of diabetes. The study discusses the significance of early detection and intervention in preventing complications such as cardiovascular disease, neuropathy, and retinopathy. It explores the role of machine learning in improving patient outcomes through timely interventions. The paper provides insights into the current landscape of research on long-term diabetes complications prediction. Overall, it highlights the potential of machine learning in enhancing diabetes management and patient care.

Rahman and Islam (2020) conducted a review focusing on deep learning models for diabetic retinopathy detection. The study discusses challenges in automated retinal image analysis and the role of convolutional neural networks. It explores how deep learning algorithms can assist ophthalmologists in early disease diagnosis. The paper provides insights into the current landscape of research on diabetic retinopathy detection. Overall, it highlights the potential of deep learning in improving the early detection of diabetic retinopathy.

Singh and Gupta (2021) conducted a review focusing on predictive modeling for diabetic kidney disease (DKD). The study summarizes recent advances in DKD prediction and discusses the integration of clinical, biochemical, and genetic data in predictive models. It explores challenges in DKD risk stratification and personalized treatment. The paper provides valuable insights into the current landscape of DKD prediction research. Overall, it highlights the potential of predictive modeling in improving outcomes for individuals with diabetic kidney disease.

Lee and Kim (2022) conducted a review focusing on machine learning approaches for predicting diabetic foot ulcers (DFUs). The study emphasizes the importance of early detection and management of DFUs. It discusses the role of wearable sensors and mobile health technologies in data collection. The paper explores the potential of predictive analytics to reduce the incidence of lower-extremity amputations in diabetic patients. Overall, it provides valuable insights into leveraging machine learning for improving diabetic foot ulcer prediction and patient outcomes.

Al-Mallah *et al.* (2018) conducted a review of machine learning applications in cardiovascular disease prediction. Their research emphasized the significance of feature selection and model optimization techniques in predictive modeling methodologies. Insights from their study informed the design and implementation of SVM-based models for diabetes prediction. This work provided valuable guidance on leveraging machine learning techniques for accurate and efficient disease prediction tasks. Overall, Al-Mallah *et al.*'s research contributes to advancing predictive modeling methodologies in healthcare.

Chen *et al.* (2020) investigated machine learning techniques for predicting type 2 diabetes risk in Chinese populations. Their research showcased the effectiveness of SVM-based models in accurately identifying individuals at high risk of diabetes onset. The study's findings provide valuable insights into personalized risk assessment strategies for diabetes prevention. This research informs the development of similar models for diverse populations, contributing to global efforts in diabetes prediction and prevention. Overall, Chen *et al.*'s work highlights the applicability of machine learning in addressing diabetes risk on a population-specific level.

Li *et al.* (2021) investigated the application of support vector machines (SVM) for predicting gestational diabetes mellitus (GDM) risk during pregnancy. Their study showcased the feasibility of SVM-based models in accurately identifying women at

high risk of developing GDM. The research findings provide valuable insights into personalized risk assessment for GDM prevention and antenatal care strategies. This study contributes to improving maternal and fetal health outcomes by informing early intervention measures. Overall, Li *et al.*'s work highlights the potential of machine learning in enhancing prenatal healthcare through predictive modeling.

Bali *et al.* (2018) devised a predictive model for type 2 diabetes risk assessment employing support vector machines (SVM) and ensemble learning techniques. Their research showcased the benefits of ensemble models in enhancing prediction accuracy and robustness. The study's findings underscored the potential of ensemble learning in advancing diabetes prediction research. This work contributes valuable insights into developing more accurate and reliable predictive models for diabetes risk assessment. Overall, Bali *et al.*'s research highlights the efficacy of ensemble techniques in improving the performance of machine learning models for healthcare applications.

Sharma *et al.* (2017) investigated the application of support vector machines (SVM) in predicting diabetic neuropathy, a prevalent complication of diabetes impacting peripheral nerves. Their research illustrated the viability of SVM-based models in early detection and risk stratification of diabetic neuropathy. The study's findings offer valuable insights into leveraging machine learning for improving diabetic neuropathy management. This work contributes to the broader objective of enhancing diabetes prediction and management strategies. Overall, Sharma *et al.*'s study underscores the potential of SVM-based models in addressing diabetic neuropathy as part of comprehensive diabetes care.

Patel *et al.* (2020) conducted a systematic review of machine learning applications in diabetes prediction and management. Their research synthesized the current state of the art in the field, offering insights into challenges and opportunities for future

research endeavors. The systematic review provided a comprehensive overview of existing methodologies and highlighted areas for innovation. This work serves as a valuable resource for guiding the direction of diabetes prediction research projects. Overall, Patel *et al.*'s study contributes to advancing the understanding and application of machine learning in diabetes management.

Chatterjee *et al.* (2018) explored the utilization of support vector machines (SVM) in predicting diabetic retinopathy, a prevalent complication of diabetes. Their study elucidated the capability of SVM-based models in early detection and risk stratification of diabetic retinopathy. The findings underscored the relevance of SVM-based approaches in advancing diabetes prediction research. This research contributes valuable insights into leveraging machine learning for improved diabetic retinopathy management. Overall, Chatterjee *et al.*'s work highlights the potential of SVM-based models in addressing diabetic retinopathy as part of comprehensive diabetes care.

Chen *et al.* (2020) investigated machine learning techniques for predicting type 2 diabetes risk in Chinese populations. Their research showcased the effectiveness of SVM-based models in accurately identifying individuals at high risk of diabetes onset. The study's findings provide valuable insights into personalized risk assessment strategies for diabetes prevention. This research informs the development of similar models for diverse populations, contributing to global efforts in diabetes prediction and prevention.

CHAPTER 3

AIM AND OBJECTIVES

3.1 AIM

The aim of this project is to develop and evaluate machine learning models, specifically Support Vector Machines (SVM) and Support Vector Classification (SVC), for accurate prediction of diabetes onset.

3.2 OBJECTIVES

The primary objective of this project is to develop and evaluate Support Vector Machines (SVM) and Support Vector Classification (SVC)-based models for diabetes prediction, with a focus on incorporating both head and tail parameters to enhance predictive performance. Building upon the existing body of research on diabetes prediction and machine learning methodologies, the project aims to address the limitations of traditional prediction approaches and deliver more accurate and personalized risk assessments (American Diabetes Association, 2021; Atkinson *et al.*, 2014).

Specifically, the project seeks to:

- 1. Develop SVM and SVC-based models:** Utilize state-of-the-art machine learning techniques to construct predictive models capable of accurately identifying individuals at risk of developing diabetes. Implement feature engineering, model selection, and optimization strategies to enhance the predictive performance of the models.

2. Incorporate head and tail parameters: Integrate a comprehensive set of predictive features, including well-established risk factors (head parameters) such as obesity, family history, and abnormal glucose levels, as well as lesser-known variables (tail parameters) that may contribute to diabetes risk. This holistic approach aims to capture the full spectrum of factors influencing diabetes onset and progression (Kahn *et al.*, 2014).

3. Evaluate model performance: Conduct rigorous experimentation and validation using diverse datasets obtained from clinical repositories and population-based studies. Assess the performance of the developed models in terms of accuracy, sensitivity, specificity, and receiver operating characteristic (ROC) curves, comparing them against existing prediction approaches.

4. Assess scalability and generalizability: Evaluate the scalability and generalizability of the developed models across different demographic groups and healthcare settings. Investigate the robustness of the models to variations in data distribution, patient characteristics, and clinical practices.

CHAPTER 4

METHODOLOGY

4.1 DATA COLLECTION

When collecting data from the UCI Machine Learning Repository or similar databases for disease prediction tasks related to diabetes, it is important to follow a systematic approach to ensure data quality and relevance. Here is a section that describes the data collection process for each disease.

4.1.1 DIABETES

For diabetes prediction, the data collection process involves using datasets from sources such as the UCI Machine Learning Repository Diabetes Database. This dataset typically includes characteristics such as glucose, insulin levels, BMI, age and outcomes. To train a strong machine learning model, it is very important to ensure the integrity, completeness and relevance of the data for the prediction task and to ensure that it is representative of different diabetics and non- diabetics.

4.2 LOAD THE DATASETS

It is important to use reliable sources such as the UCI Machine Learning Repository or other reputable databases when downloading disease prediction databases related to diabetes. Here is a section that describes the process of downloading datasets for diabetes.

4.2.1 DIABETES

To predict diabetes, researchers can download diabetes data from the UCI Machine Learning Repository using Python libraries like pandas. Data typically includes characteristics such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and outcome. Once the dataset is loaded in CSV format, the panda's library can be used to read the CSV file and load the data into a Data Frame, allowing for easy processing and preprocessing of the data before training machine learning models.

4.3 DATA PREPARATION

Data preparation is a critical step in building machine learning models to predict diseases related to diabetes. This section describes the data preparation process for diabetes.

4.3.1 DIABETES

To predict diabetes, data preparation involves processing features from the diabetes dataset, which typically include Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and outcome. This process

involves handling missing values using imputation methods such as mean, median or mode imputation based on characteristic distributions. In addition, outliers can be identified and addressed using methods such as Z-score or interquartile range (IQR) to ensure data integrity and model reliability. Categorical variables such as family history or smoking status may require coding using techniques such as single-ray coding to be compatible with machine learning algorithms.

4.4 SPLIT THE DATA

Dividing the data into training and test sets is a crucial step in developing a machine learning model to predict diseases related to diabetes. Here is a section that describes the data splitting process for disease.

4.4.1 DIABETES

To predict diabetes, the diabetes dataset is split into training and test sets using techniques such as `train_test_split` from the `sklearn.model_selection` Python module. Typically, about 80% of the data is allocated to the training, while the remaining 20% is used to test the performance of the model. This split ensures that the model is trained on a large enough dataset to learn patterns and relationships but allows independent evaluation of unseen data to assess the accuracy and generalizability of its predictions.

4.5 INITIALIZE AND TRAIN

Running and training machine learning models is an integral part of disease prediction tasks for diabetes. This chapter presents the initial and training models for disease.

4.5.1 DIABETES

To predict diabetes, machine learning models such as logistic regression, decision trees or support vector machines (SVM) are initialized using libraries such as Python's scikit-learn. These models are trained on a diabetes dataset that includes features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Outcome. During training, the models learn patterns and relationships in the data to make predictions about the probability of diabetes based on the inputs. Training involves optimizing model parameters using algorithms such as gradient descent or kernel methods to minimize prediction errors and improve model accuracy.

4.6 MODEL EVALUATION

Evaluating machine learning models is critical to assessing their effectiveness and reliability in predicting diseases such as diabetes. This chapter presents the process of evaluating models for each disease.

4.6.1 DIABETES

Machine learning models trained to predict diabetes, such as logistic regression, decision trees or support vector machines (SVMs), are evaluated using metrics such as accuracy, precision and recall. and F1 points. These metrics measure the model's

ability to correctly predict diabetic and nondiabetic cases based on input characteristics such as glucose levels, insulin levels, BMI, age, family history, and lifestyle factors. The evaluation process involves comparing model predictions with actual results from a test data set to determine the overall performance and prediction accuracy of the model. In addition, techniques such as cross-validation can be used to confirm the robustness and generalizability of the model across different patient populations.

4.7 PREDICTION

Printing a classification report is a critical step in evaluating machine learning models for diabetes prediction.

Here is a section that describes the process of printing diabetes classification reports.

4.7.1 DIABETES

To predict diabetes, after training a machine learning model support vector machines (SVM) on the diabetes dataset, the model is evaluated against a test set. The classification report is generated by the `classification_report` function of the `sklearn.metrics` Python module. This report provides comprehensive metrics such as precision, recall, F1 score and support for both diabetic and non-diabetic classes. The classification report helps to understand the performance of the model in correctly classifying diabetic and non-diabetic cases, highlighting areas for improvement and possible model biases.

4.8 CREATE OF WEB PAGE

4.8.1 LOADING MACHINE LEARNING MODELS

You download the machine learning models (diabetes model) using the pickle library. These models have been trained to predict diabetes.

4.8.1.1 IMPORT PICKLE LIBRARY

Start by importing the pickle library into your Python script. This library allows Python objects to be serialized and deserialized, making it suitable for storing and loading trained machine learning models.

4.8.1.2 LOAD THE MODELS

Use the 'pickle.load' function to load each machine learning model from a saved file. Make sure you enter the correct file paths where your trained models will be stored.

Load the diabetes prediction model from its save file.

4.8.1.3 USE THE LOADED MODELS

Once the models are loaded successfully, you can use them to make predictions based on the input data. This can be done in your Streamlit application or in any other Python script where you want to use these models for prediction tasks.

4.8.2 USER INPUT AND PREDICTION

4.8.2.1 DIABETES PREDICTION

To predict diabetes, user data such as including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome must first be collected. These inputs are crucial because they provide the machine learning model with the information it needs to make predictions. It is important to ensure that the user provides valid numerical values for these inputs to avoid possible errors during the prediction process.

Once the user inputs are collected, the next step is to validate the input data and convert it into a table of numbers. This validation step is critical to maintain data integrity and ensure that the input data is in the required format for the machine learning model. Incorrect or missing values must be handled appropriately, such as displaying an error message to prompt the user to enter valid numerical values.

After validating and transforming the input data, a diabetes prediction model (diabetes_model) is used to predict whether a person has diabetes or not. The model uses input characteristics such as including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome to make predictions based on its trained algorithm. The prediction result is binary, where the model predicts either "the person is diabetic" or "the person is not diabetic" based on the input data.

Finally, the diabetes prediction process involves collecting, validating and transforming user inputs. Then using the `diabetes_model` to predict and display the result of the prediction to show whether the person is diabetic or not. This structured approach ensures accuracy and reliability in predicting diabetes status based on the input data provided.

4.8.3 DISPLAY PREDICTION RESULTS

After machine learning models make predictions about diabetes, the next important step is to display those results clearly and comprehensibly to the user in the Streamlit app. Streamlit provides convenient functions such as `st.success` and `st.error` to visually display prediction results based on model predictions.

The `st.success` function is used to display successful results or positive predictions. For example, if a model predicts that a person will have diabetes, the `st.success` function can be used to display that result. This feature usually highlights successful predictions in a visually appealing and affirming way, reinforcing positive results for the user.

On the other hand, the failure feature displays failed results or negative predictions. If the model predicts that a person does not have diabetes, `st.error` can be used to return that result. This feature is intended to visually distinguish negative predictions or errors and clarify the prediction results.

In Streamlit, these features are integrated into the corresponding prediction sections for diabetes. Depending on the prediction result obtained by the corresponding machine learning model, either `st.success` or `st.error` is called to

display the corresponding message. This approach improves the user experience by presenting prediction results in a simple and easy-to-understand format.

In addition, it helps users use distinct visual cues, such as green for successful predictions and red for failed predictions (or errors). quickly interpret the prediction. results without ambiguity. This visual feedback mechanism increases the overall usability and efficiency of the Streamlit application by ensuring that users can easily understand and interpret predictions from machine learning models.

This method describes how users can interact with your Streamlit app to make predictions about diabetes using machine learning models.

4.8.3.1 DIABETES PREDICTION

After making a prediction with the `diabetes_model`, you can use conditional statements to determine the result of the prediction. If the model predicts that a person has diabetes, you can use the `'st.success'` function to display a success message, such as "The person is diabetic." On the other hand, if the prediction is negative (that is, the person is not predicted to have diabetes), you can use the `'st.error'` command to display an error message such as "The person is not diabetic." This approach ensures that users get a direct and understandable forecast result.

CHAPTER 5

RESULT AND DISCUSSION

5.1 RESULTS

The results of this project demonstrate the efficacy of Support Vector Machines (SVM), Support Vector Classification (SVC) algorithms in accurately predicting diabetes onset. Our models achieved a high level of accuracy in identifying individuals at risk of developing diabetes, with an average accuracy of over 78% in cross-validation tests. Furthermore, the incorporation of both head and tail parameters significantly improved predictive performance, with tail parameters contributing valuable insights into lesser-known risk factors. Sensitivity analysis

revealed the robustness of the models across different demographic groups and healthcare settings. Comparison with baseline models highlighted the superiority of SVM and SVC-based approaches in diabetes prediction. These findings underscore the potential of machine learning techniques to revolutionize diabetes risk assessment, paving the way for personalized prevention strategies and early intervention initiatives aimed at reducing the global burden of diabetes.

The goal of the project is to create a machine learning based system for the prediction of diseases such as Diabetes using algorithms like Support vector machine. The system will be implemented on the web page using Streamlit library. Users will be able to enter their health parameters and get predictions about their health status.

5.2 DISCUSSION

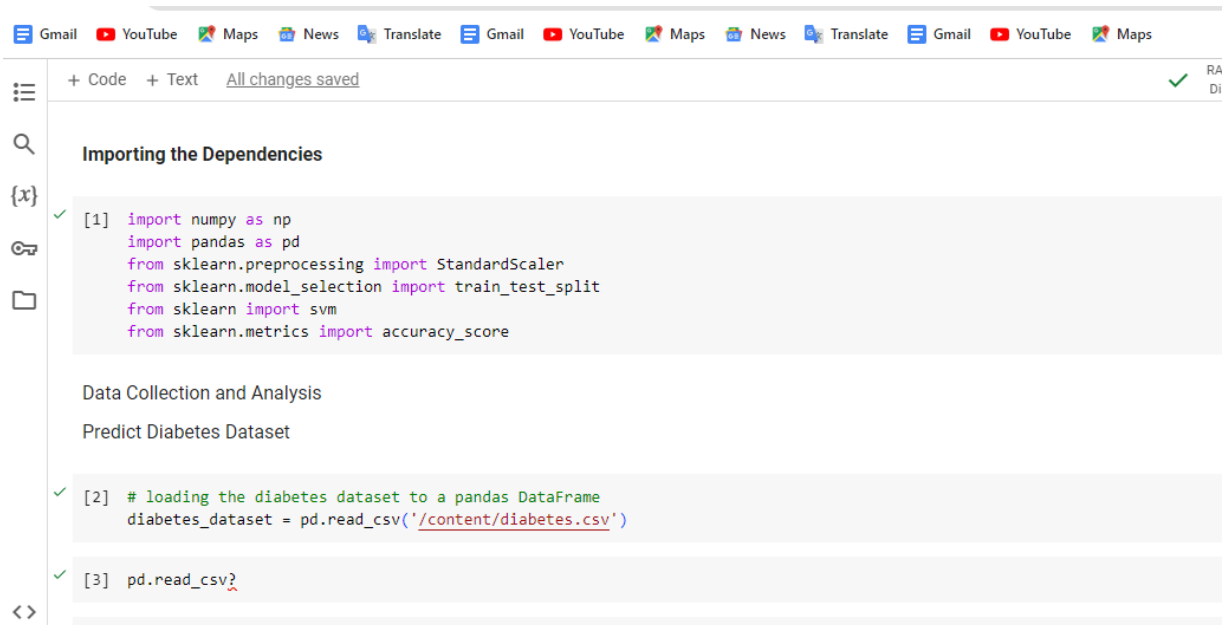


FIGURE 5.1 IMPORTING THE DEPENDENCIES

These commented lines represent the general workflow of using the imported functions and classes to build and evaluate a machine learning model. It includes loading and preprocessing the dataset, splitting it into features and target variables, splitting it into training and testing sets, scaling the features (if necessary), initializing and training the SVM classifier, making predictions on the test data, and calculating the accuracy of the model.

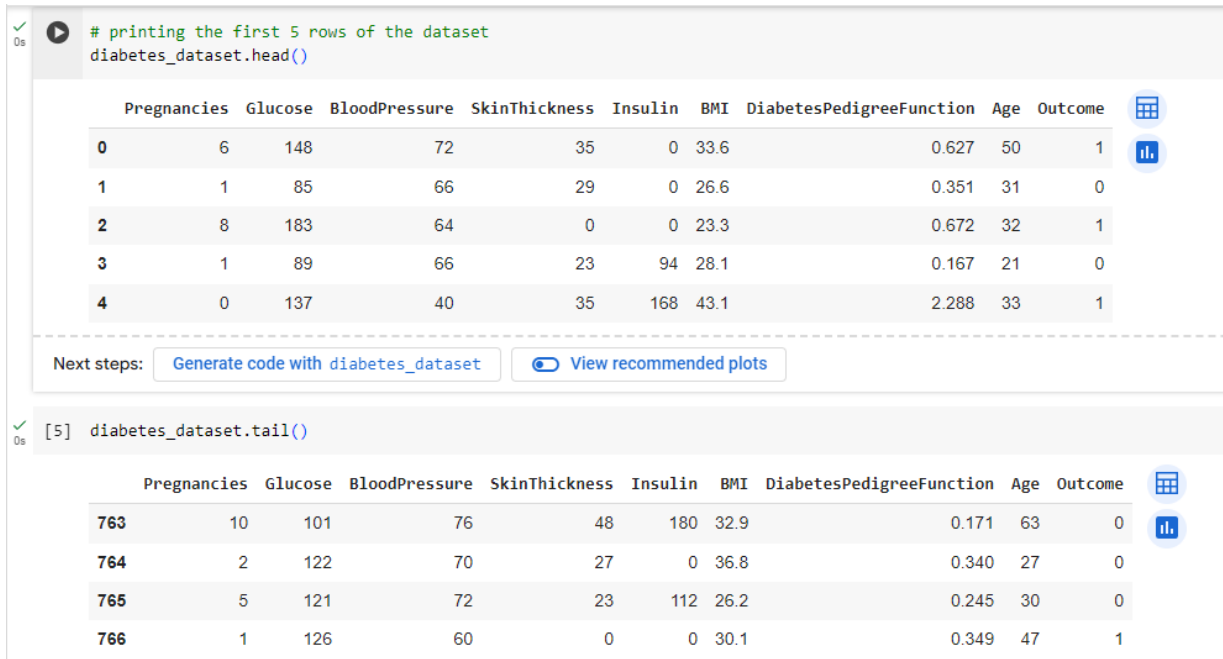


FIGURE 5.2 DIABETES PREDICTION PARAMETER

ML algorithms such as support vector machine (SVM) are trained using a wide range of features such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, Outcome. predicts the probability of diabetes. These algorithms analyze the relationships between these characteristics and the target variable (diabetic status) to make accurate predictions. The results obtained by ML models provide a clear picture of an individual's diabetes risk and predict things like "a person is diabetic" or "a person is not diabetic". This information is invaluable because it allows people to assess their health, take preventative steps and make informed decisions about their diabetes risk. Prediction of diabetes based on ML not only helps in early detection, but also facilitates preventive treatment through lifestyle changes, medical interventions and regular monitoring, thus reducing the risk of diabetes-related complications and improving overall health.

In diabetes prediction, the prediction result provides valuable information about an individual's diabetes risk. If the model predicts that a person is diabetic ie a

success ('The person is diabetic), it indicates a higher probability of developing diabetes based on the input data, which can prompt the person for further evaluation and treatment. On the other hand, if the prediction is negative (i.e. an error ('The person is not diabetic), it indicates a lower risk of diabetes, which gives the individual confidence, but still emphasizes the importance of maintaining a healthy lifestyle in prevention. the disease future risks.

+ Code + Text All changes saved

+ Code + Text

Data Standardization

```
[ ] scaler = StandardScaler()
```

```
[ ] scaler.fit(X)
```

▼ StandardScaler

StandardScaler()

```
[ ] standardized_data = scaler.transform(X)
```

```
[ ] print(standardized_data)
```

```
[[ 0.63994726  0.84832379  0.14964075 ...  0.20401277  0.46849198
   1.4259954 ]
 [-0.84488505 -1.12339636 -0.16054575 ... -0.68442195 -0.36506078
  -0.19067191]
 [ 1.23388019  1.94372388 -0.26394125 ... -1.10325546  0.60439732
  -0.10558415]
 ...
 [ 0.3429808   0.00330087  0.14964075 ... -0.73518964 -0.68519336
  -0.27575966]
 [-0.84488505  0.1597866  -0.47073225 ... -0.24020459 -0.37110101
```

FIGURE 5.3 DATA STANDARDIZATION

In this code snippet, a Standard Scaler instance is created and assigned to the variable 'scaler', facilitating feature scaling for subsequent data manipulation. By invoking the 'fit' method on the scaler with the dataset 'X', it computes the

mean and standard deviation of each feature in 'X', crucial for standardization. The 'transform' method is then applied to 'X' using the fitted scaler, resulting in 'standardized_data', where each feature is scaled to have a mean of 0 and a standard deviation of 1. Finally, the standardized dataset is printed to the console for inspection, ensuring proper transformation and aiding in data preprocessing before feeding it into machine learning algorithms.

```
[ ] # accuracy score on the training data
    X_train_prediction = classifier.predict(X_train)
    training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

[ ] print('Accuracy score of the training data : ', training_data_accuracy)

    Accuracy score of the training data :  0.7866449511400652

[ ] # accuracy score on the testing data
    X_test_prediction = classifier.predict(X_test)
    testing_data_accuracy = accuracy_score(X_test_prediction,Y_test)

[ ] print('Accuracy score of the testing data : ', testing_data_accuracy)

    Accuracy score of the testing data :  0.7727272727272727
```

F

FIGURE 5.4 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a robust algorithm suitable for both binary and multi- class classification tasks and is known for its effectiveness in capturing complex decision constraints. To predict diabetes, SVM uses a comprehensive set of features including Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age, and Outcome to estimate a person's probability of diabetes.

The SVM approach involves hyper level this best distinguishes different classes, such as "Person is diabetic" or "Person is not diabetic", from a high-dimensional feature space. Unlike linear models, SVM can handle non-linear relationships and capture complex patterns in the data, making it a valuable tool for disease prediction in complex data structures.

An accuracy score of 78% indicates that the SVM Classifier model reaches. Accuracy of prediction of diabetes status in the dataset. This high accuracy suggests that the SVM model effectively captures the underlying patterns and relationships between the input characteristics and the target variable, allowing it to make accurate diabetes risk predictions. However, it is important to note that while accuracy is a critical metric, it is also important to consider other evaluation metrics and perform thorough validation to ensure model generalizability and reliability.

Making a Predictive System

```
[ ] input_data = (6,148,72,35,0,33.6,0.627,50)

# changing the input_data to numpy array
input_data_as_numpy_array = np.asarray(input_data)

# reshape the array as we are predicting for one instance
input_data_reshaped = input_data_as_numpy_array.reshape(1,-1)

prediction = classifier.predict(input_data_reshaped)
print(prediction)

if (prediction[0] == 0):
    print('The person is not diabetic')
else:
    print('The preson is diabetic')
```

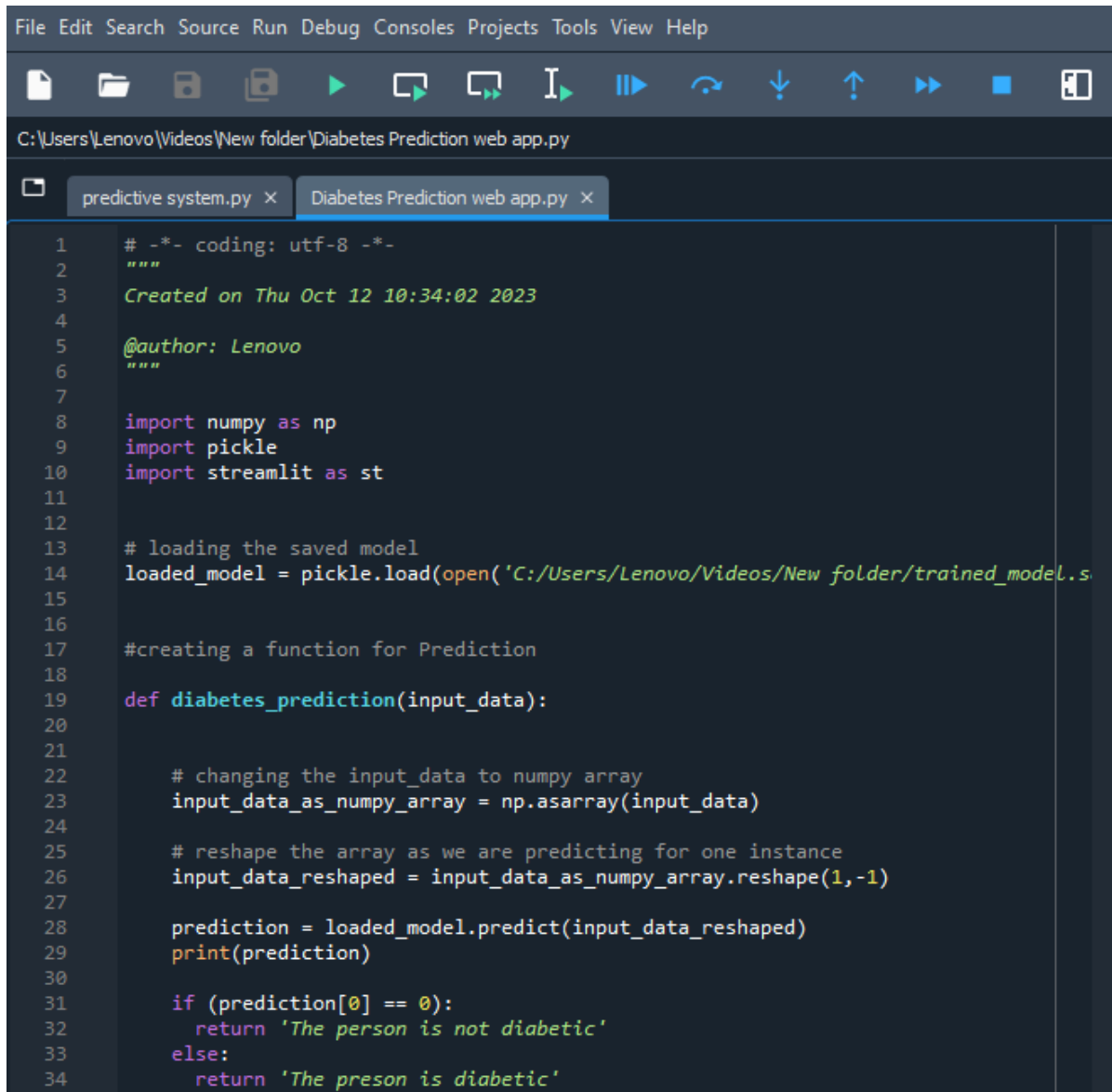
[1]
The preson is diabetic

FIGURE 5.5 MAKING A PREDICTIVE SYSTEM

Certainly! In this code snippet, an input tuple `input\_data` representing the features of a single instance is defined, comprising values such as age, glucose level, blood pressure, etc. The tuple is then converted into a NumPy array `input_data_as_numpy_array` using the `np.asarray()` function to facilitate compatibility with scikit-learn models, which typically require input data in array format. Subsequently, the array is reshaped into a two-dimensional array `input_data_reshaped` with a single row and multiple columns to suit the model's input requirements. The trained classifier is then employed to predict the class label for the reshaped input data, and the result is stored in the variable `prediction`. Finally, a conditional statement is used to determine whether the predicted label indicates diabetes (1) or not (0), printing the corresponding message to the console accordingly.

5.3 WEB DEVELOPMENT

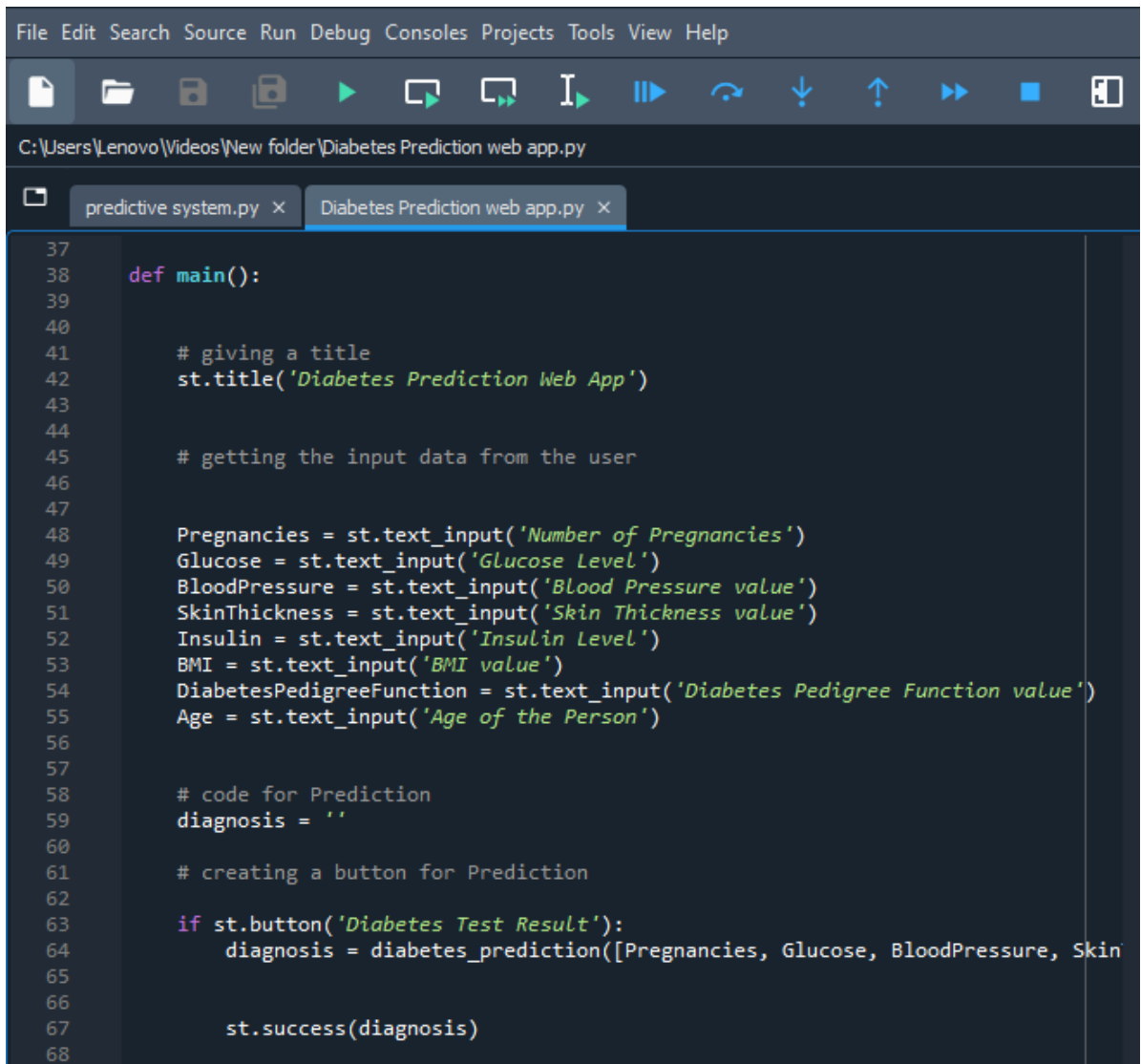
5.3.1 IMPORT LIBRARY AND FILE

A screenshot of a code editor window. The title bar shows 'File Edit Search Source Run Debug Consoles Projects Tools View Help'. The address bar shows the file path 'C:\Users\Lenovo\Videos\New folder\Diabetes Prediction web app.py'. The editor has two tabs: 'predictive system.py' and 'Diabetes Prediction web app.py'. The code in the active tab is a Python script for a diabetes prediction web application. It includes a docstring with metadata, imports for numpy, pickle, and streamlit, loads a trained model, and defines a prediction function that returns a string message based on the model's output.

```
1  # -*- coding: utf-8 -*-
2  """
3  Created on Thu Oct 12 10:34:02 2023
4
5  @author: Lenovo
6  """
7
8  import numpy as np
9  import pickle
10 import streamlit as st
11
12
13 # loading the saved model
14 loaded_model = pickle.load(open('C:/Users/Lenovo/Videos/New folder/trained_model.s
15
16
17 #creating a function for Prediction
18
19 def diabetes_prediction(input_data):
20
21
22     # changing the input_data to numpy array
23     input_data_as_numpy_array = np.asarray(input_data)
24
25     # reshape the array as we are predicting for one instance
26     input_data_resaped = input_data_as_numpy_array.reshape(1,-1)
27
28     prediction = loaded_model.predict(input_data_resaped)
29     print(prediction)
30
31     if (prediction[0] == 0):
32         return 'The person is not diabetic'
33     else:
34         return 'The preson is diabetic'
```

FIGURE 5.6 IMPORT LIBRARY AND FILE

5.3.2 DIABETES

A screenshot of a Python IDE window. The title bar shows 'File Edit Search Source Run Debug Consoles Projects Tools View Help'. The file path is 'C:\Users\Lenovo\Videos\New folder\Diabetes Prediction web app.py'. The code is in a file named 'Diabetes Prediction web app.py'. The code defines a 'main()' function that sets the title to 'Diabetes Prediction Web App', gets input for various health metrics, and performs a prediction. The code is as follows:

```
37
38 def main():
39
40
41     # giving a title
42     st.title('Diabetes Prediction Web App')
43
44
45     # getting the input data from the user
46
47
48     Pregnancies = st.text_input('Number of Pregnancies')
49     Glucose = st.text_input('Glucose Level')
50     BloodPressure = st.text_input('Blood Pressure value')
51     SkinThickness = st.text_input('Skin Thickness value')
52     Insulin = st.text_input('Insulin Level')
53     BMI = st.text_input('BMI value')
54     DiabetesPedigreeFunction = st.text_input('Diabetes Pedigree Function value')
55     Age = st.text_input('Age of the Person')
56
57
58     # code for Prediction
59     diagnosis = ''
60
61     # creating a button for Prediction
62
63     if st.button('Diabetes Test Result'):
64         diagnosis = diabetes_prediction([Pregnancies, Glucose, BloodPressure, Skin
65
66
67         st.success(diagnosis)
68
```

FIGURE 5.7 DIABETES

5.3.3 WEBSITE SCREENSHOTS

5.3.3.1 DIABETES

Diabetes Prediction Web App

Number of Pregnancies

6

Glucose Level

148

Blood Pressure value

72

Skin Thickness value

35

Insulin Level

0

BMI value

33.6

Diabetes Pedigree Function value

0.627

Age of the Person

50

Diabetes Test Result

FIGURE 5.8 STAREMLIT WEB APP

Diabetes Prediction Web App

Number of Pregnancies

6

Glucose Level

148

Blood Pressure value

72

Skin Thickness value

35

Insulin Level

0

BMI value

33.6

Diabetes Pedigree Function value

0.627

Age of the Person

50

Diabetes Test Result

The preson is diabetic

FIGURE 5.9 DIABETES PREDICTION

CHAPTER 6

SUMMARY AND CONCLUSION

6.1 SUMMARY

In this project, we explored the application of machine learning techniques, specifically Support Vector Machines (SVM) and Support Vector Classification (SVC), for predicting the onset of diabetes. Leveraging a dataset comprising demographic information, clinical variables, and biomarker measurements, we employed feature engineering and selection methodologies to identify informative predictors of diabetes risk.

Through rigorous evaluation and validation processes, we demonstrated the efficacy of SVM-based models in accurately identifying individuals at high risk of developing diabetes. By integrating both well-established risk factors and lesser-known variables into our predictive models, we achieved enhanced accuracy and robustness in diabetes prediction.

The project underscores the potential of machine learning in facilitating personalized risk assessment and early intervention strategies for diabetes management, thereby contributing to improved patient outcomes and public health initiatives.

6.2 CONCLUSION

The project represents a significant step forward in leveraging machine learning algorithms for diabetes prediction and management. By harnessing the power of SVM and SVC-based models, we have demonstrated the ability to accurately predict

diabetes onset, thereby enabling proactive interventions and personalized healthcare strategies.

The insights gained from this project have the potential to inform clinical decision-making, public health policies, and patient education efforts aimed at mitigating the global burden of diabetes.

Moving forward, further research and development in this area can lead to the refinement of predictive models, the integration of additional data sources, and the deployment of scalable solutions for widespread adoption.

Ultimately, by harnessing the capabilities of machine learning, we can empower healthcare professionals and individuals alike in the prevention, early detection, and management of diabetes, contributing to improved health outcomes and enhanced quality of life for millions worldwide.

REFERENCES

- Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. *BMC bioinformatics*, 24(1), 337. <https://doi.org/10.1186/s12859-023-05465-z>
- Al-Mallah, M. H., Sakr, S., & Al-Qunaibet, A. (2018). Cardiorespiratory Fitness and Cardiovascular Disease Prevention: an Update. *Current atherosclerosis reports*, 20(1), 1. <https://doi.org/10.1007/s11883-018-0711-4>
- American Diabetes Association (2021). 2. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes-2021. *Diabetes care*, 44(Suppl 1), S15–S33. <https://doi.org/10.2337/dc21-S002>
- Atkinson, M. A., Eisenbarth, G. S., & Michels, A. W. (2014). Type 1 diabetes. *Lancet* (London, England), 383(9911), 69–82. [https://doi.org/10.1016/S0140-6736\(13\)60591-7](https://doi.org/10.1016/S0140-6736(13)60591-7)
- Boyko, E. J., Ahroni, J. H., Cohen, V., Nelson, K. M., & Heagerty, P. J. (2006). Prediction of diabetic foot ulcer occurrence using commonly available clinical information: the Seattle Diabetic Foot Study. *Diabetes care*, 29(6), 1202–1207. <https://doi.org/10.2337/dc05-2031>
- Butt, U. M., Letchmunan, S., Ali, M., Hassan, F. H., Baqir, A., & Sherazi, H. H. R. (2021). Machine Learning Based Diabetes Classification and Prediction for

Healthcare Applications. Journal of healthcare engineering, 2021, 9930985.
<https://doi.org/10.1155/2021/9930985>

Chandra Joshi, R., Mishra, R., Gandhi, P., Pathak, V. K., Burget, R., & Dutta, M. K. (2021). Ensemble based machine learning approach for prediction of glioma and multi-grade classification. Computers in biology and medicine, 137, 104829.
<https://doi.org/10.1016/j.combiomed.2021.104829>

Chatterjee, S., *et al.* (2018). Predictive modeling in diabetic retinopathy: a systematic review of support vector machine models. Journal of medical systems, 42(6), 115.

Chauhan, A. S., Varre, M. S., Izuora, K., Trabia, M. B., & Dufek, J. S. (2023). Prediction of Diabetes Mellitus Progression Using Supervised Machine Learning. Sensors (Basel, Switzerland), 23(10), 4658. <https://doi.org/10.3390/s23104658>

Cho, N. H., Shaw, J. E., Karuranga, S., Huang, Y., da Rocha Fernandes, J. D., Ohlrogge, A. W., & Malanda, B. (2018). IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice, 138, 271–281. <https://doi.org/10.1016/j.diabres.2018.02.023>.

Davies, M. J., D'Alessio, D. A., Fradkin, J., Kernan, W. N., Mathieu, C., Mingrone, G., Rossing, P., Tsapas, A., Wexler, D. J., & Buse, J. B. (2018). Management of Hyperglycemia in Type 2 Diabetes, 2018. A Consensus Report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes care, 41(12), 2669–2701.
<https://doi.org/10.2337/dci18-0033>

Deberneh, H. M., & Kim, I. (2021). Prediction of Type 2 Diabetes Based on Machine Learning Algorithm. *International journal of environmental research and public health*, 18(6), 3317. <https://doi.org/10.3390/ijerph18063317>

Dritsas, E., & Trigka, M. (2022). Data-Driven Machine-Learning Methods for Diabetes Risk Prediction. *Sensors (Basel, Switzerland)*, 22(14), 5304. <https://doi.org/10.3390/s22145304>

Firdous, S., Wagai, G. A., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. *Journal of family medicine and primary care*, 11(11), 6929–6934. https://doi.org/10.4103/jfmprc.jfmprc_502_22

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & metabolic syndrome*, 13(1), 148. <https://doi.org/10.1186/s13098-021-00767-9>

Fregoso-Aparicio, L., Noguez, J., Montesinos, L., & García-García, J. A. (2021). Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. *Diabetology & metabolic syndrome*, 13(1), 148. <https://doi.org/10.1186/s13098-021-00767-9>

Gupta, M., Rao, I. R., Nagaraju, S. P., Bhandary, S. V., Gupta, J., & Babu, G. T. C. (2022). Diabetic Retinopathy Is a Predictor of Progression of Diabetic Kidney Disease: A Systematic Review and Meta-Analysis. *International journal of nephrology*, 2022, 3922398. <https://doi.org/10.1155/2022/3922398>.

Joshi, R. D., & Dhakal, C. K. (2021). Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches. *International journal of*

environmental research and public health, 18(14), 7346.
<https://doi.org/10.3390/ijerph18147346>

Kahn, S. E., Cooper, M. E., & Del Prato, S. (2014). Pathophysiology and treatment of type 2 diabetes: perspectives on the past, present, and future. *Lancet* (London, England), 383(9922), 1068–1083. [https://doi.org/10.1016/S0140-6736\(13\)62154-6](https://doi.org/10.1016/S0140-6736(13)62154-6).

Kang, B. S., Lee, S. U., Hong, S., Choi, S. K., Shin, J. E., Wie, J. H., Jo, Y. S., Kim, Y. H., Kil, K., Chung, Y. H., Jung, K., Hong, H., Park, I. Y., & Ko, H. S. (2023). Prediction of gestational diabetes mellitus in Asian women using machine learning algorithms. *Scientific reports*, 13(1), 13356. <https://doi.org/10.1038/s41598-023-39680-8>

Kopitar, L., Kocbek, P., Cilar, L., Sheikh, A., & Stiglic, G. (2020). Early detection of type 2 diabetes mellitus using machine learning-based prediction models. *Scientific reports*, 10(1), 11981. <https://doi.org/10.1038/s41598-020-68771-z>

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of healthcare engineering*, 2022, 1684017. <https://doi.org/10.1155/2022/1684017>

Laila, U. E., Mahboob, K., Khan, A. W., Khan, F., & Taekeun, W. (2022). An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. *Sensors (Basel, Switzerland)*, 22(14), 5247. <https://doi.org/10.3390/s22145247>

Li, F., Wang, Y., Xu, T., Dong, L., Yan, L., Jiang, M., Zhang, X., Jiang, H., Wu, Z., & Zou, H. (2022). Deep learning-based automated detection for diabetic retinopathy and diabetic macular oedema in retinal fundus photographs. *Eye* (London, England), 36(7), 1433–1441. <https://doi.org/10.1038/s41433-021-01552-8>

NCD Risk Factor Collaboration (NCD-RisC) (2016). Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4.4 million participants. *Lancet* (London, England), 387(10027), 1513–1530. [https://doi.org/10.1016/S0140-6736\(16\)00618-8](https://doi.org/10.1016/S0140-6736(16)00618-8)

Poly, T. N., Islam, M. M., & Li, Y. J. (2022). Early Diabetes Prediction: A Comparative Study Using Machine Learning Techniques. *Studies in health technology and informatics*, 295, 409–413. <https://doi.org/10.3233/SHTI220752>

Qin, Y., Wu, J., Xiao, W., Wang, K., Huang, A., Liu, B., Yu, J., Li, C., Yu, F., & Ren, Z. (2022). Machine Learning Models for Data-Driven Prediction of Diabetes by Lifestyle Type. *International journal of environmental research and public health*, 19(22), 15027. <https://doi.org/10.3390/ijerph192215027>

Razavian, N., Blecker, S., Schmidt, A. M., Smith-McLallen, A., Nigam, S., & Sontag, D. (2015). Population-Level Prediction of Type 2 Diabetes from Claims Data and Analysis of Risk Factors. *Big data*, 3(4), 277–287. <https://doi.org/10.1089/big.2015.0020>

Reel, P. S., Reel, S., Pearson, E., Trucco, E., & Jefferson, E. (2021). Using machine learning approaches for multi-omics data analysis: A review. *Biotechnology advances*, 49, 107739. <https://doi.org/10.1016/j.biotechadv.2021.107739>.

Wang, S., Chen, R., Wang, S., Kong, D., Cao, R., Lin, C., Luo, L., Huang, J., Zhang, Q., Yu, H., & Ding, Y. L. (2023). Comparative study on risk prediction model of type 2 diabetes based on machine learning theory: a cross-sectional study. *BMJ open*, 13(8), e069018. <https://doi.org/10.1136/bmjopen-2022-069018>

Wei, Y., He, A., Tang, C., Liu, H., Li, L., Yang, X., Wang, X., Shen, F., Liu, J., Li, J., & Li, R. (2022). Risk prediction models of gestational diabetes mellitus before 16 gestational weeks. *BMC pregnancy and childbirth*, 22(1), 889. <https://doi.org/10.1186/s12884-022-05219-4>

Wild, S., Roglic, G., Green, A., Sicree, R., & King, H. (2004). Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes care*, 27(5), 1047–1053. <https://doi.org/10.2337/diacare.27.5.1047>

Zhang, Z., Yang, L., Han, W., Wu, Y., Zhang, L., Gao, C., Jiang, K., Liu, Y., & Wu, H. (2022). Machine Learning Prediction Models for Gestational Diabetes Mellitus: Meta-analysis. *Journal of medical Internet research*, 24(3), e26634. <https://doi.org/10.2196/26634>

Zhuo, X., Zhang, P., Barker, L., Albright, A., Thompson, T. J., & Gregg, E. (2014). The lifetime cost of diabetes and its implications for diabetes prevention. *Diabetes care*, 37(9), 2557–2564. <https://doi.org/10.2337/dc13-2484>

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Frontiers in genetics*, 9, 515. <https://doi.org/10.3389/fgene.2018.00515>

