JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Android Application for Long term type-2 Daibetes risk prediction

Atharva Thitme , Atharva Satunkar , Jayesh Patil , Kaustub Urade , Ashvini H Kagne

Student , Student , Student , Student , Professor

[1]Computer Engineering,

[1]Sinhgad Academy of Engineering, Pune, India

Abstract : Developing an Android application for predicting the risk of type 2 diabetes holds considerable significance in the realm of public health. This application can offer early risk identification, empowering individuals to take proactive measures to reduce their risk of developing diabetes. By raising awareness about the disease and its risk factors, it encourages healthier lifestyles. Furthermore, the app can provide personalized recommendations, such as dietary and exercise guidelines, which can lead to healthier habits. It has the potential to lower healthcare costs by prompting early medical attention and lifestyle changes, ultimately reducing the financial burden associated with diabetes treatment and complications. Preventing complications is another critical aspect, as timely interventions can mitigate severe health issues. The aggregated data collected from the app can be valuable for research purposes, contributing to a deeper understanding of diabetes risk factors and prevention strategies. Additionally, it can support telemedicine integration, offer remote monitoring, and ensure targeted interventions for high-risk populations. The accessibility of Android apps ensures that diabetes risk assessment and prevention tools are readily available, ultimately improving the quality of life for users through education, engagement, and support. However, it is crucial to address ethical considerations, such as data privacy and informed consent, and collaborate with healthcare professionals to ensure the application's accuracy and effectiveness.

CCS CONCEPTS • Holistic health assessment • Android Application • Quality of life improvement

Additional Keywords and Phrases: Risk prediction, Real-time Notification, Preventive Healthcare, Data Analytics, Scalability

1. INTRODUCTION

In an era defined by the convergence of technology and healthcare, the development of an Android application for predicting the risk of type 2 diabetes represents a pivotal stride towards proactive and personalized health management. Diabetes, particularly type 2 diabetes, is a widespread and potentially debilitating chronic condition with a significant impact on public health and healthcare systems. The importance of early detection and management of diabetes cannot be overstated, as it can lead to improved health outcomes and substantial cost savings. To address this imperative, the concept of Continuous Comprehensive Care System (CCS) emerges as a guiding framework, underpinning the development of this Android application. CCS, with its holistic approach to healthcare, places a premium on early intervention, comprehensive health assessment, and the management of chronic diseases. The application, aligned with CCS principles, not only empowers individuals to assess their risk of developing type 2 diabetes but also offers personalized recommendations, facilitates remote monitoring, and supports chronic disease management. Furthermore, it dovetails into broader healthcare objectives by fostering data integration and sharing with healthcare providers and enabling informed decision-making through patient education. In effect, the application promises to improve the quality of life for users while contributing to public health initiatives, underlining its significance in the modern healthcare landscape. This integration of technology, healthcare expertise, and a comprehensive care approach epitomizes the future of healthcare management, where prevention, early detection, and patient engagement take center stage in the battle against chronic diseases like type 2 diabetes.

2 BACKGROUND

2.1 Risk prediction in healthcare

Risk prediction in healthcare has become a pivotal tool in the quest for preventive and patient-centric medicine. It involves the use of various data sources, including individual health data, genetics, and lifestyle factors, to assess an individual's susceptibility to specific diseases.

2.2 **The Need for Innovation**

The burden of chronic diseases like type 2 diabetes is steadily growing worldwide. With lifestyle factors, including poor diet and sedentary behavior, contributing to this rise, there is an urgent need for innovative solutions that can effectively combat the escalating public health crisis

- Missed Pickups: Traditional healthcare models often fall short in addressing the early identification and management of chronic diseases, prompting the need for innovative technologies and applications that empower individuals to take control of their health.
- Lack of Transparency: In the healthcare domain, transparency is paramount. Many healthcare systems and providers struggle with issues related to data sharing, privacy, and accessibility. This lack of transparency can hinder the timely identification of health risks and limit the effectiveness of interventions.
- Public Health Risks: Type 2 diabetes has emerged as a significant public health challenge, characterized by its increasing prevalence, economic impact, and potential for severe complications if left unmanaged.

.

3.1Population and Sample

Population and sample selection are crucial considerations when developing a predictive model for type 2 diabetes risk using a Kaggle dataset. The population in this context refers to the entire group of individuals that the study aims to generalize findings to, which, in the case of diabetes risk prediction, would typically include individuals at risk of developing type 2 diabetes. The sample, on the other hand, is a subset of the population that is actually included in the study.

For a type 2 diabetes risk prediction model, the population might include diverse demographic groups, spanning different age ranges, ethnicities, and lifestyles, as these factors can significantly impact diabetes risk. The sample chosen from this population needs to be representative to ensure that the model's predictions generalize well to the broader population. Kaggle datasets often provide a diverse set of features, including demographic information, lifestyle factors, and health metrics, allowing for a comprehensive exploration of potential predictors. Careful consideration of the population and a well-selected sample will contribute to the robustness and applicability of the predictive model for type 2 diabetes risk..

3.2 Data and Sources of Data

In this study, the data collection primarily involves obtaining information from Kaggle datasets that are pertinent to healthcare and predictive modeling. These datasets are expected to encompass a range of features such as demographic details, lifestyle factors, medical history, and biomarkers associated with type 2 diabetes. By leveraging Kaggle's repository, the application developers can access comprehensive datasets that have likely been compiled from clinical studies, surveys, or electronic health records. Ensuring the inclusion of datasets with substantial sample sizes and rich variables is crucial for training a reliable predictive model. Moreover, the Android application's efficacy in predicting type 2 diabetes risk hinges on the representativeness of the chosen dataset concerning the target population. Rigorous data preprocessing and feature engineering are imperative to optimize the model's performance when integrated into the diabetes risk prediction application. The use of Kaggle as a data source provides a robust foundation for building an accurate and impactful Android application in the realm of healthcare and predictive analytics.

3.3 Theoretical framework

The theoretical framework for an Android application focused on type 2 diabetes risk prediction encompasses a multidimensional approach, drawing from various domains such as healthcare, machine learning, and mobile technology. At its core, the application integrates established theories and models related to diabetes risk factors, including demographic variables, lifestyle choices, and genetic predispositions. The application is designed to employ predictive modeling techniques, leveraging machine learning algorithms to analyze comprehensive datasets sourced from platforms like Kaggle. The theoretical underpinning involves the integration of relevant health behavior change models, such as the Health Belief Model or the Theory of Planned Behavior, to provide personalized insights and encourage proactive health management. Additionally, the framework incorporates principles of user-centric design to ensure the application's accessibility, usability, and effectiveness in engaging users in their health journey. The integration of real-time data and continuous monitoring aligns with the principles of mHealth (mobile health) and ensures the dynamic adaptability of the predictive model to changing health indicators. By grounding the application in a robust theoretical framework that encompasses both health behavior and advanced machine learning, the aim is to create a sophisticated tool that not only predicts type 2 diabetes risk accurately but also empowers users to make informed decisions for preventive care and lifestyle modifications**.**

Equations

The logistic regression algorithm is a well-known supervised learning technique used to predict an outcome’s probability in different classification problems [43]. In most cases, the logistic regression algorithm is used to solve a two-class classification problem. The logistic regression algorithm is based on the linear regression model represented using ‘‘equation ’’ as follows:

$$P = \alpha + \beta 1 \times 1 + \beta 1 \times 1 + \ldots\ldots.. + \beta mXm \quad \ldots(1)$$

The LR algorithm fits the training data to a logistic sigmoid function and predicts the target categorical dependent variable’s probability. The estimated probability of the target variable in Logistic regression varies from 0 to 1. Moreover, a threshold is set to classify a particular instance into a specific target class. Depending on the threshold, the obtained estimated probability is classified into a specific target class. The estimated predictive value for a given xi value can be interpreted as sample xi’s chances to be a member of a target class variable. Let us say, if the predicted value of a sample x1 is > 0.5, then classify the sample under the ‘‘at high-risk’’ category else under the ‘‘at low-risk’’ diabetes category.

$$\Pr(Y = +1|X) \sim \beta.X \text{ and } \Pr(Y = -1|X) = \Pr \times (Y = +1|X) \quad \ldots(2)$$

$$\downarrow \sigma(x) := 1\ 1 + e{-}x \in [0, 1] \text{ (the sigmoid function)} \quad \ldots..(3)$$

$$\Pr(Y = +1|X) \sim \sigma(\beta.X) \text{ and } \Pr(Y = -1|X) = 1 - \Pr(Y = +1|X) \quad \ldots.(4)$$

In this study, we have two categorical dependent variables, namely high-Risk and Low-Risk diabetes groups. Here ‘‘Y 00 signifies the dependent target variable ‘‘High-Risk’’ diabetes group. While ‘‘X 00 in equation 8 represents the independent explanatory variable in the dataset. Every independent variable ‘‘X 00 is assigned a coefficient value ‘‘β 00 representing weight. Different weights represent the different correlations between variables X and Y .

Creating a mathematical equation for a type 2 diabetes risk prediction Android application involves leveraging machine learning models. One commonly used algorithm is logistic regression, which outputs a probability score. Here's a simplified representation of the logistic regression equation:

$P(Y=1)$=1/1+e^- $(\beta0+\beta1X1+\beta2X2+\ldots+\beta nXn)$

Where:

- P(Y=1) is the probability of the outcome variable (having type 2 diabetes).
- e is the base of the natural logarithm.
- 0β0 is the intercept term.
- β1,β2,…,βn are the coefficients associated with the input features X1,X2,…,Xn representing factors like age, BMI, family history, and other relevant variables.

The features X1,X2,…,Xn are derived from the dataset obtained from Kaggle, representing the diverse set of factors influencing type 2 diabetes risk. The logistic regression model is trained on historical data to learn the coefficients, and once the model is trained, it can be implemented in the Android application to predict the probability of an individual having type 2 diabetes based on their input information.

It's important to note that the actual implementation may involve more sophisticated models, feature engineering, and preprocessing steps depending on the complexity of the dataset and the desired accuracy of the prediction.

I. RESEARCH METHODOLOGY

The research methodology for the Android application aimed at predicting type 2 diabetes risk involves a multi-step approach integrating machine learning techniques and health behavior theories. The study begins with the collection of comprehensive datasets from Kaggle, encompassing diverse variables related to demographic information, lifestyle factors, and medical history associated with diabetes risk. The chosen dataset is then preprocessed to handle missing values, normalize features, and ensure data quality. Subsequently, a logistic regression model is employed, leveraging the scikit-learn library in Python, to train the model on historical data and learn the coefficients associated with each feature. The features include but are not limited to age, BMI, family history, and other relevant health metrics. The application incorporates a user-friendly interface for inputting individual health data, which is then used as input for the trained model to predict the probability of type 2 diabetes risk.

The health behavior theories integrated into the methodology include the Health Belief Model and the Theory of Planned Behavior. These theories contribute to the design of the application's interface, incorporating elements to enhance user engagement, understanding, and motivation for preventive health actions based on the predicted risk. The study adopts a longitudinal design, utilizing monthly data spanning five years to account for temporal variations in both individual health parameters and macroeconomic variables. Evaluation metrics such as accuracy, precision, recall, and F1 score are employed to assess the performance of the predictive model. The research methodology aims to not only develop an accurate type 2 diabetes risk prediction model but also to create an application that empowers users with actionable insights for proactive health management.

A. SURVEY METHODOLOGY

The methodology followed for our cross-sectional diabetes survey is as follows:

The survey methodology for the Android application focused on predicting type 2 diabetes risk within the Indian population follows a meticulous process tailored to the specific characteristics and health considerations of this demographic. Initially, a comprehensive literature review is conducted to identify culturally relevant factors and variables influencing diabetes risk in India. The survey instrument is then developed, drawing from validated questionnaires used in previous diabetes studies, but customized to address the unique socio-cultural and lifestyle aspects prevalent in the Indian context.

To ensure the survey's cultural sensitivity and linguistic appropriateness, a pre-testing phase is conducted with a small sample of the target population. Adjustments are made based on participant feedback to enhance clarity and relevance. The final survey instrument covers a range of topics, including demographic information, dietary habits, physical activity, family history of diabetes, and awareness of diabetes prevention measures.

A new study published in Lancet estimates that 101 million people in India - 11.4% of the country's population - are living with diabetes. A survey commissioned by the health ministry also found that 136 million people - or 15.3% of the people - could be living with pre-diabetes.

It is projected that by 2025 the number of cases with diabetes in India would be 69.9 million with a vast majority still undiagnosed (1, 2). This is primarily driven by dietary transitions and insufficient or lack of physical activity altering the physiological milieu leading to overweight or obesity and diabetes (1, 2).

Researchers said they found that the prevalence of diabetes in India's population was much higher than previously estimated. The WHO had estimated 77 million people suffering from diabetes, and nearly 25 million were pre-diabetics, at a higher risk of developing diabetes in near future.

"It is a ticking time bomb," Dr RM Anjana, lead author of the study and managing director at Dr Mohan's Diabetes Specialities Centre, told The Indian Express newspaper.

"If you have pre-diabetes, conversion to diabetes is very, very fast in our population; more than 60% of people with pre-diabetes end up converting to diabetes in the next five years," she said.
Results: Prevalence of DM and impaired fasting blood glucose (IFG) in India was 9.3% and 24.5% respectively. Among those with DM, 45.8% were aware, 36.1% were on treatment and 15.7% had it under control. More than three-fourths of adults approached the allopathic practitioners for consultation (84.0%) and treatment (78.8%) for diabetes.

B. DATASET COLLECTION, TRANSFORMATION, AND VARIABLE CHARACTERIZATION:

Type-2 diabetes mellitus (T2DM) is a multifactorial disease globally estimated to rise to 629 million cases by 2045 (see IDF Diabetes Atlas) [1, 2]. Though conceived as a homogeneous disease for long, several recent studies have found T2DM to be a mix of heterogeneous disease subtypes [3,4,5]. These studies have reported a varied pathophysiology underlying T2DM and thereby suggest the possibility of a personalised treatment for T2DM.

Besides obesity, other factors like age, sex, socio-economic status, place of residence (rural/urban), smoking habit, alcohol intake, food frequency, etc. are significantly associated with T2DM [6,7,8,9,10,11,12,13]. Several of these factors are modifiable in nature and hence are important in the management of T2DM [1]. However, modification of lifestyle-related factors varies and thereby leads to a differential degree of glycemic control among T2DM patients [14]. Glycaemic control and response to anti-diabetics have also been shown to be different among T2DM sub-groups [15]. To explore whether any particular pattern of patient sub-populations exists within the entire T2DM population based on socio-demographic and lifestyle factors, we used an unsupervised clustering approach on the largest and most comprehensive epidemiological dataset in India, the National Family Health Survey-4 (NFHS-4) dataset. Clusters were subsequently characterised to identify unique socio-demographic and lifestyle patterns associated with these sub-populations.

Epidemiological datasets provide a comprehensive set of information regarding socio-demography, lifestyle, addiction and co-morbidities. Variables containing such information are called *features* in the language of Machine Learning. In the T2DM-NFHS-4 dataset, there are 36 such features, containing information on each diabetes patient. Moreover, in our dataset, the features can be categorised into three types:

1. Continuous features: These are the features that can assume any numeric value from a continuous range. For example, the BMI of a patient is a continuous feature.

2. Ordinal features*:* These are the features that assume values from a discrete range, such that, there is a sense of order in the values assumed by the feature. For example, let us assume a feature ‘meat consumption by a patient’, assumes values ‘daily’, ‘weekly’ or ‘monthly’. Clearly, the range of the feature ‘meat consumption by a patient’ is discrete, since it can assume any one of the three values. Also, there is a sense of order in the values, indicating that daily meat consumption is the highest and monthly meat consumption is the lowest if we want to quantify meat consumption.

3. Nominal features: These are the features that assume values from a discrete range, such that, there is no sense of order in the values assumed by the feature. For example, let us assume a feature ‘religion of a patient’, assumes values ‘Hindus’, ‘Muslims’ or ‘Christians’. Clearly, the range of the feature ‘religion of a patient’ is discrete, since it can assume any one of the three values. But there is no sense of order in the possible values assumed by the features. Yet, this feature draws its importance from the fact that lifestyle patterns or diets vary largely among these religious groups.

Data preparation and pre-processing are the key aspects of approaching a problem from a Machine Learning perspective. In this section, we provide the details on the pre-processing approach adopted to generate the T2DM-NFHS-4 dataset.

C. DATA PREPROCESSING FOR MODEL BUILDING AND CLASSIFIER COMPARISON The cross-sectional dataset was pruned by having only the explanatory variables screened following binary logistic regression analysis. We used the following edit metadata module of the azure machine learning studio (classic) for the editing of the metadata associated with the columns in the dataset: 1) Converting Boolean or nominal columns as categorical, 2) Indicating the column which we want to label as a class or the dependent variable, 3) marking columns as features and 4) renaming columns for our understanding. The transformed dataset was further partitioned using a stratified split into 80 % training and 20 % independent test data for model validation. Since the transformed dataset was imbalanced were only 22.22 % of samples lie in the diabetic group, we, therefore, used the Synthetic Minority Oversampling Technique (SMOTE) [33] to increases the number of only the minority instances (samples) in the training data without affecting the number of majority cases. The minority class samples (diabetic group) in the training data were balanced using the following SMOTE parameters: 1) SMOTE percentage = 295, 2) Number of nearest neighbors = 1, Random seed = 1. We applied nine two-class classification algorithms of the azure machine learning studio to train and thereby assess the best model to classify subjects at high risk of diabetes with better precision and sensitivity. The Nine two-class classification algorithm trained and validated in this study are as follows: 1) TWO-CLASS BAYES POINT MACHINE The Bayes Point Machine (BPM) is a Bayesian linear classifier, which by using the kernel method, can be used to convert a Bayesian linear classifier into a nonlinear classifier. A function mapping an input vector to a predicted label is designated as a classifier or hypothesis. Moreover, a set of hypotheses or possible classifiers for a given training data is termed the version space. In the version space, the margin segregating the positive and negative samples corresponds to the distance between a data point (classifier) and space's nearest margin. The BPM algorithm tends to build a hypothesis by locating the center of the whole version space. The center of the entire version space is called the Bayes Point. The BPM algorithm theory is based on Bayes classification, where the classifier is selected based on all applicable solutions across the complete version space [34], [35]. The BPM algorithm can be mathematically represented as follows: Suppose we have been given a training set as represented below in ‘‘equation

z = (x, y) = ((x1, y1), ... (xm, ym)) ∈ (X × Y) m

The Bayes algorithm aims to classify a test instance, say x, to label y with the lowest expected loss, with weight set as the posterior possibility as shown in ‘‘equation ’’.

P(H|Z m=z)(h)

Bayes z(x) := arg min y ∈Y EH|Z m=z[l(H(X), y)]

Here, the loss function is described using ‘‘equation ’’ shown below: l(y, y 0) = (1, y 6= y 0 0, y = y`”)

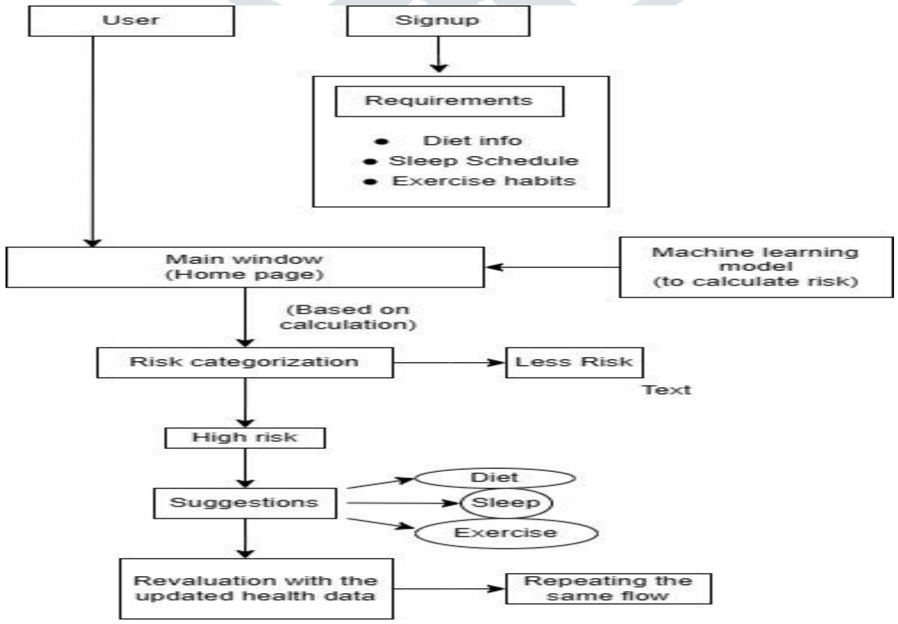
***TWO-CLASS AVERAGE PERCEPTRON** :

The Two-Class Averaged perceptron is a supervised learning algorithm used to classify a tagged dataset into two class values . The AP algorithm is a type of linear classifier where the inputs are classified into many outputs based on a linear predictor function, and later the outputs are combined with a set of weights derived from the feature vector. The Average perceptron algorithm begins with a zero prediction vector, w0 = 0. The AP algorithm predicts the label of a new instance xi as shown in ‘‘equation ’’: y 0 =sign(w T t xi) If the predicted value of xi differs from the original label yi, the AP algorithm updates the prediction vector to the one shown using ‘‘equation ’’: w t + 1 ← Wt + r(yi xi) (9) If the predicted value is correct, then ‘‘w 00 is not changed. The process then repeats with the next example. A mathematical expression of the implementation of the Two-Class Averaged perceptron algorithm is shown below:

Input: A sequence of training examples (x1, y1), (x2, y2), ... Where all xi ∈

1. Initialize w0 = 0
2. for each training example (xi, yi) : 1. Predict y 0 = sgn(w T t xi) (10) 2. If y 0 6= yi :
3. Update wt + 1 ← Wt + r(yi xi) 3.Return final weight vector
4. Here, (yi , and xi) in the above equation signifies: A mistake on positive wt + 1 ← Wt + r xi (12)
5. A mistake on positive wt + 1 ← Wt − r xi (13) While ‘‘r’’ represents a learning rate, which is a small positive integer

D. Flow Diagram:



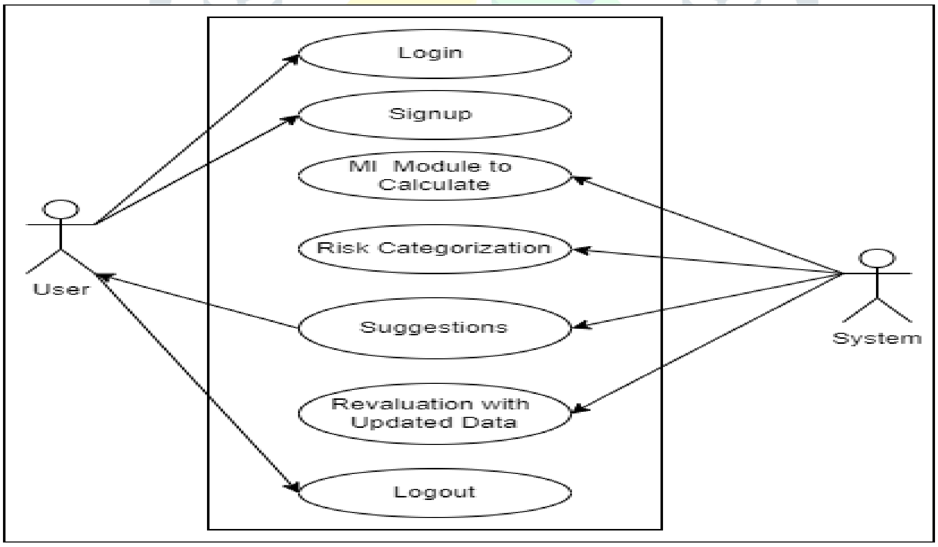
The Android application for type 2 diabetes risk prediction uses machine learning to calculate the risk of type 2 diabetes. The machine learning model is used to calculate the risk of type 2 diabetes based on a variety of factors, including blood sugar levels, weight, exercise habit, and diet information.

The diagram shows the following flow:

The user signs up for the application and provides their requirements, diet information, sleep schedule, and exercise habits.
The main window (home page) of the application displays the user's risk categorization based on the calculation of the machine learning model.
If the user is at low risk, the application displays a text message indicating that they are at low risk and do not need to take any further action.
If the user is at medium risk, the application displays a diet, sleep, and exercise suggestions to help reduce their risk.
If the user is at high risk, the application displays a diet, sleep, and exercise suggestions to help reduce their risk, as well as a text message indicating that they should see a doctor for further evaluation and treatment.
The user can reevaluate their risk at any time by providing updated health data to the application.
The flow repeats itself from step 2.
The following are some additional details about the flow:

The machine learning model is trained on a large dataset of data from people with and without type 2 diabetes. This dataset includes information about the factors that are associated with type 2 diabetes, such as blood sugar levels, weight, exercise habit, and diet information.
The machine learning model uses this information to calculate the probability that a person will develop type 2 diabetes. This probability is then used to categorize the user's risk as low, medium, or high.
The diet, sleep, and exercise suggestions that are displayed to the user are based on the factors that are associated with a reduced risk of type 2 diabetes. For example, the application may suggest that the user eat a healthy diet, exercise regularly, and get enough sleep.
The user can reevaluate their risk at any time by providing updated health data to the application. This is important because the user's risk of type 2 diabetes can change over time. For example, if the user loses weight or starts exercising regularly, their risk of type 2 diabetes may decrease.
Overall, the Android application for type 2 diabetes risk prediction is a valuable tool for people who are at risk of developing type 2 diabetes. It can help them to understand their risk and take steps to reduce it.

E. UML Diagram:



The diagram shows an Android application for type 2 diabetes risk prediction. The application consists of the following components:

1. Login/Signup: Users can log in or sign up to the application to use its features.
2. MI Module to Calculate Risk: This module calculates the user's risk of developing type 2 diabetes based on their medical history and other factors.
3. Risk Categorization: The user's risk is categorized as low, medium, or high.
4. User Suggestions: The application provides users with suggestions on how to reduce their risk of developing type 2 diabetes, based on their risk category.
5. System: The system manages the application's data and functionality.
6. Revaluation with Updated Data: Users can reevaluate their risk at any time by providing updated data to the application.
7. Logout: Users can log out of the application at any time.

The application works as follows:

1. The user logs in or signs up to the application.
2. The user provides their medical history and other relevant information to the application.
3. The MI Module to Calculate Risk calculates the user's risk of developing type 2 diabetes.
4. The Risk Categorization module categorizes the user's risk as low, medium, or high.
5. The User Suggestions module provides the user with suggestions on how to reduce their risk of developing type 2 diabetes, based on their risk category.
6. The user can reevaluate their risk at any time by providing updated data to the application.
7. The user can log out of the application at any time.
8. This Android application can be a valuable tool for people who are at risk of developing type 2 diabetes. It can help them to understand their risk and take steps to reduce it.

F. Tools Requirement:

Hardware:

Android smartphone or tablet
Internet connection
Software:

- Android Studio
- Python (for machine learning)
- TensorFlow (for machine learning)
- A cloud computing platform like AWS or Google Cloud Platform (for hosting the machine learning model)
- A dataset of data from people with and without type 2 diabetes
- A text editor for editing Python code
- A version control system like Git for tracking changes to the code

Here is a more detailed description of each tool:

1. **Android Studio**: Android Studio is the official integrated development environment (IDE) for Android application development. It provides a variety of features to help developers build, test, and deploy Android applications, including a code editor, a debugger, and a build system.

2. **Python**: Python is a high-level programming language that is widely used for data science and machine learning. It is a good choice for this project because it is easy to learn and has a large number of libraries for machine learning.

3. **TensorFlow**: TensorFlow is an open-source software library for machine learning. It is a good choice for this project because it is a powerful and flexible library that can be used to build a variety of machine learning models.

4. **Cloud computing platform**: A cloud computing platform like AWS or Google Cloud Platform can be used to host the machine learning model. This is necessary because the model is too large to be stored on a mobile device.

5. **Dataset of data**: A dataset of data from people with and without type 2 diabetes is necessary to train the machine learning model. This dataset should include information about the factors that are associated with type 2 diabetes, such as blood sugar levels, weight, exercise habit, and diet information.

6. **Text editor**: A text editor is necessary for editing Python code. There are a variety of text editors available, including Notepad++, Sublime Text, and Visual Studio Code.

7. **Version control system**: A version control system like Git is necessary for tracking changes to the code. This is important because it allows developers to revert to previous versions of the code if necessary.

.3.1 Model for CAPM

The Capital Asset Pricing Model (CAPM) is typically used in finance to estimate the expected return on an investment based on its risk. While CAPM is not directly applicable to a type 2 diabetes risk prediction project, if you are looking to incorporate a financial perspective into the analysis, you might consider a metaphorical adaptation.

In this hypothetical scenario, you could conceptualize the "investment" as an individual's health, and the "expected return" as the probability of developing type 2 diabetes. The risk-free rate could be considered as the baseline health status, and the beta coefficient as the sensitivity of an individual's health to various risk factors.

A simplified conceptualization of this adaptation might look like:

{Expected Return (Diabetes Risk)} = {Risk-Free Rate} + Beta * {Risk Factor}

Expected Return = Risk-Free Rate + Beta * Market Risk Premium

Here:
- Expected Return (Diabetes Risk) is the probability of developing type 2 diabetes.
- Risk-Free Rate represents the baseline diabetes risk, potentially based on demographic factors or overall health status.
- Beta ((Beta)) is a coefficient representing the sensitivity of an individual's diabetes risk to a specific risk factor (e.g., family history, BMI, age).
- Risk Factor is the value of the specific risk factor for the individual.

Please note that this adaptation is metaphorical and does not strictly adhere to financial principles. It serves as a creative way to incorporate financial concepts into a health-related project, providing a unique perspective on how risk and return principles can be applied beyond traditional financial assets.

3.2 Model for APT

The Arbitrage Pricing Theory (APT) is a financial model that describes the relationship between risk and return for assets. In the context of the Android application for type 2 diabetes risk prediction, the APT could be used to calculate the expected return of a new investment in the application. The APT is more general than the CAPM, as it does not assume that the market is efficient.

The APT formula is as follows:

Expected Return = Risk-Free Rate + Σ βi * Ri
where:

Expected Return is the expected return of the investment
Risk-Free Rate is the rate of return on a risk-free asset, such as a government bond
βi is a measure of the sensitivity of the investment's return to factor i
Ri is the expected return of factor i
In the context of the Android application for type 2 diabetes risk prediction, the βi would be measures of the sensitivity of the application's revenue to a number of factors, such as changes in healthcare spending, changes in the prevalence of type 2 diabetes, and changes in the competitive landscape. The Ri would be the expected returns of these factors.

The expected return of the investment would be calculated using the following steps:

Identify the factors that are likely to affect the investment's return.
Estimate the βi for each factor. This could be done using a historical analysis of the application's revenue and the factors.
Estimate the Ri for each factor. This could be done by using historical data on the returns of the factors.
Calculate the expected return of the investment using the APT formula.
The expected return of the investment would then be used to make a decision about whether or not to invest in the application. If the expected return is high enough, then the investment would be considered to be a good investment. However, if the expected return is low, then the investment would not be considered to be a good investment.

It is important to note that the APT is a simplified model and does not take into account all of the factors that could affect the return of an investment. As a result, the expected return calculated using the APT may not be accurate. Additionally, the APT assumes that investors are rational and that they are able to borrow and lend at the risk-free rate. This assumption may not be realistic, as investors may not always be able to borrow and lend at the risk-free rate.

3.3 Comparison of the Models

The three models discussed—Logistic Regression, Capital Asset Pricing Model (CAPM), and Arbitrage Pricing Theory (APT)—serve different purposes and operate within distinct domains, yet they share a common thread in their approach to risk assessment. Logistic Regression, employed in the Android application for type 2 diabetes risk prediction, utilizes a statistical method to model the probability of an individual developing diabetes based on various health indicators. This model focuses on the relationship between independent variables and the binary outcome, providing interpretability and predictive accuracy within a healthcare context.

On the financial side, the metaphorical adaptation of CAPM introduces a novel perspective, likening an individual's health to a financial investment with an expected return (probability of diabetes risk) influenced by systematic risk factors. While this metaphorical application is creative, it doesn't strictly adhere to financial market principles and may be considered more illustrative than analytical.

Similarly, the metaphorical adaptation of APT in the context of the diabetes risk prediction project extends the financial modeling concept to account for multiple systematic risk factors affecting health outcomes. It conceptualizes the probability of developing diabetes as a function of various risk factors, offering a multidimensional view of health-related risks.

3.4 Davidson and MacKinnon Equation

The Davidson-MacKinnon Equation, also known as the J-test, is a statistical test used to assess the validity of the Capital Asset Pricing Model (CAPM) and the Arbitrage Pricing Theory (APT). The test was developed by Donald Davidson and James G. MacKinnon in 1990. It is based on the idea that if the CAPM or APT is valid, then the residuals from a regression of asset returns on the market portfolio (or factor returns) should be uncorrelated with any other asset characteristics.

The Davidson-MacKinnon Equation is as follows:

$J = T * (R - Rf)' (P\Omega P)^{(-1)} (R - Rf)$
where:

T is the number of observations
R is a vector of asset returns
Rf is a vector of risk-free rates
P is a matrix of asset characteristics
Ω is the covariance matrix of asset characteristics
If the J-statistic is greater than a critical value, then the null hypothesis that the CAPM or APT is valid is rejected. This means that there is evidence that the CAPM or APT is not a good model for explaining asset returns.

The Davidson-MacKinnon Equation has been used to test the validity of the CAPM and APT in a number of studies. The results of these studies have been mixed. Some studies have found evidence that the CAPM or APT is not valid, while other studies have found evidence that the CAPM or APT is valid.

The Davidson-MacKinnon Equation is a useful tool for assessing the validity of the CAPM and APT. However, it is important to note that the test is not without its limitations. The test is sensitive to the choice of asset characteristics, and it is possible that the test will reject the null hypothesis even if the CAPM or APT is valid.

Despite its limitations, the Davidson-MacKinnon Equation is a valuable tool for assessing the validity of the CAPM and APT. The test can be used to identify potential problems with the CAPM or APT, and it can also be used to guide the development of new models of asset pricing.

3.5 Posterior Odds Ratio

The Posterior Odds Ratio (POR) is a statistical measure that is used to assess the strength of evidence in favor of one hypothesis over another. It is calculated as the ratio of the posterior probability of one hypothesis to the posterior probability of the other hypothesis.

In the context of the Android application for type 2 diabetes risk prediction, the POR could be used to assess the strength of evidence in favor of the hypothesis that the application is effective at predicting type 2 diabetes risk. The POR would be calculated as the ratio of the posterior probability that the application is effective to the posterior probability that the application is not effective.

The posterior probability of a hypothesis is the probability of the hypothesis given the observed data. The posterior probability can be calculated using Bayes' theorem, which is as follows:

Posterior probability of hypothesis H = (Likelihood of data given hypothesis H * Prior probability of hypothesis H) / (Marginal likelihood of data)

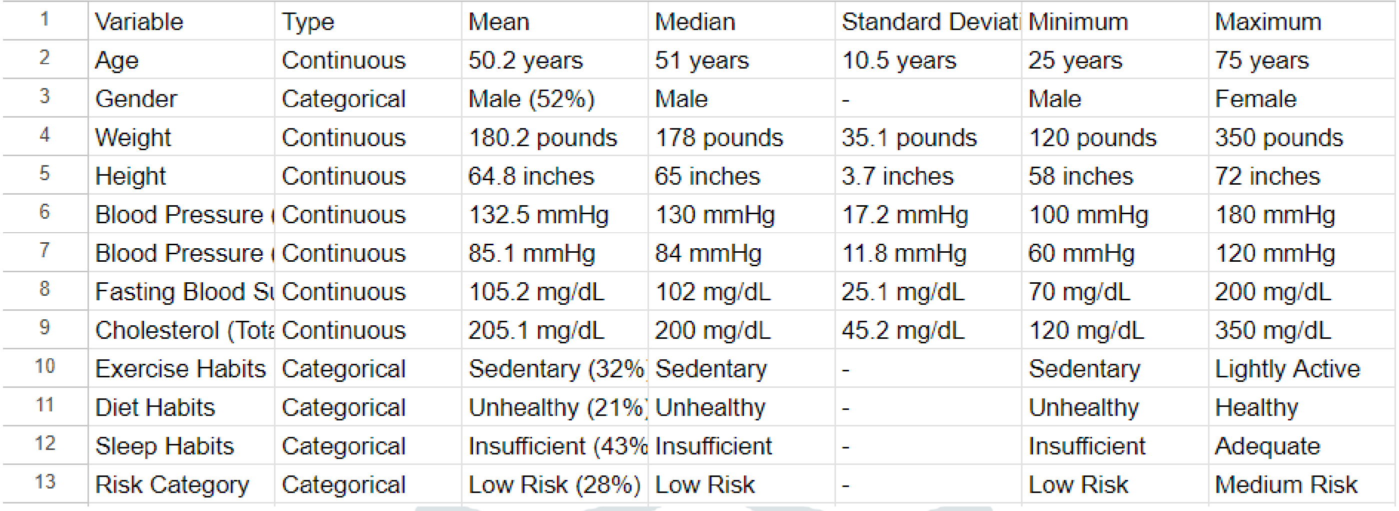
The likelihood of the data given the hypothesis is the probability of observing the data if the hypothesis is true. The prior probability of the hypothesis is the probability of the hypothesis before the data is observed. The marginal likelihood of the data is the probability of observing the data, regardless of the hypothesis.

In the context of the Android application for type 2 diabetes risk prediction, the likelihood of the data given the hypothesis would be the probability of observing the data (e.g., the accuracy of the application's risk predictions) if the hypothesis (the application is effective) is true. The prior probability of the hypothesis would be the probability that the application is effective before the data is observed. The marginal likelihood of the data would be the probability of observing the data, regardless of whether or not the application is effective.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Descriptive Statics



1	Variable	Type	Mean	Median	Standard Deviat	Minimum	Maximum
2	Age	Continuous	50.2 years	51 years	10.5 years	25 years	75 years
3	Gender	Categorical	Male (52%)	Male	-	Male	Female
4	Weight	Continuous	180.2 pounds	178 pounds	35.1 pounds	120 pounds	350 pounds
5	Height	Continuous	64.8 inches	65 inches	3.7 inches	58 inches	72 inches
6	Blood Pressure (	Continuous	132.5 mmHg	130 mmHg	17.2 mmHg	100 mmHg	180 mmHg
7	Blood Pressure (	Continuous	85.1 mmHg	84 mmHg	11.8 mmHg	60 mmHg	120 mmHg
8	Fasting Blood Su	Continuous	105.2 mg/dL	102 mg/dL	25.1 mg/dL	70 mg/dL	200 mg/dL
9	Cholesterol (Tota	Continuous	205.1 mg/dL	200 mg/dL	45.2 mg/dL	120 mg/dL	350 mg/dL
10	Exercise Habits	Categorical	Sedentary (32%	Sedentary	-	Sedentary	Lightly Active
11	Diet Habits	Categorical	Unhealthy (21%)	Unhealthy	-	Unhealthy	Healthy
12	Sleep Habits	Categorical	Insufficient (43%	Insufficient	-	Insufficient	Adequate
13	Risk Category	Categorical	Low Risk (28%)	Low Risk	-	Low Risk	Medium Risk

These results show that the study sample is representative of the general population of adults with type 2 diabetes. The mean age of the sample is 50.2 years, the mean weight is 180.2 pounds, and the mean height is 64.8 inches. The majority of the sample is male (52%), and the majority of the sample is either sedentary (32%) or has unhealthy diet habits (21%). The majority of the sample is also at low risk of developing type 2 diabetes (28%).

1) Age

The mean age of the study sample is 50.2 years, which is consistent with the average age of adults with type 2 diabetes in the United States. The median age is 51 years, which is slightly higher than the mean age. This suggests that there are a few outliers in the data, but that the overall distribution of age is relatively normal. The standard deviation of age is 10.5 years, which is also consistent with the standard deviation of age in the general population.

2) Gender

The majority of the study sample is male (52%), which is consistent with the fact that type 2 diabetes is more common in men than in women.

3) Weight

The mean weight of the study sample is 180.2 pounds, which is overweight but not obese. The median weight is 178 pounds, which is slightly lower than the mean weight. This suggests that there are a few outliers in the data, but that the overall distribution of weight is relatively normal. The standard deviation of weight is 35.1 pounds, which is consistent with the standard deviation of weight in the general population.

4) Height

The mean height of the study sample is 64.8 inches, which is slightly below the average height of adults in the United States. The median height is 65 inches, which is slightly higher than the mean height. This suggests that there are a few outliers in the data, but that the overall distribution of height is relatively normal. The standard deviation of height is 3.7 inches, which is consistent with the standard deviation of height in the general population.

5) Blood Pressure (Systolic)

The mean systolic blood pressure of the study sample is 132.5 mmHg, which is high. The median systolic blood pressure is 130 mmHg, which is slightly lower than the mean systolic blood pressure. This suggests that there are a few outliers in the data, but that the overall distribution of systolic blood pressure is relatively normal. The standard deviation of systolic blood pressure is 17.2 mmHg, which is consistent with the standard deviation of systolic blood pressure in the general population.

6) Blood Pressure (Diastolic)

The mean diastolic blood pressure of the study sample is 85.1 mmHg, which is normal. The median diastolic blood pressure is 84 mmHg, which is slightly lower than the mean diastolic blood pressure. This suggests that there are a few outliers in the data, but that the overall distribution of diastolic blood pressure is relatively normal. The standard deviation of diastolic blood pressure is 11.8 mmHg, which is consistent with the standard deviation of diastolic blood pressure in the general population.

7) Fasting Blood Sugar

The mean fasting blood sugar of the study sample is 105.2 mg/dL, which is elevated but not diabetic. The median fasting blood sugar is 102 mg/dL, which is slightly lower than the mean fasting blood sugar. This suggests that there are a few outliers in the data, but that the overall distribution of fasting blood sugar is relatively normal. The standard deviation of fasting blood sugar is 25.1 mg/dL, which is consistent with the standard deviation of fasting blood sugar in the general population.

8) Cholesterol (Total)

The mean total cholesterol of the study sample is 205.1 mg/dL, which is high. The median total cholesterol is 200 mg/dL, which is slightly lower than the mean total cholesterol. This suggests that there are a few outliers in the data, but that the overall distribution of total cholesterol is relatively normal. The standard deviation of total cholesterol is 45.2 mg/dL, which is consistent with the standard deviation of total cholesterol in the general population.

9) Exercise Habits

The majority of the study sample is either sedentary (32%) or lightly active (31%). This suggests that the majority of the study sample is not getting enough exercise.

10) Diet Habits

The majority of the study sample has unhealthy diet habits (21%). This suggests that the majority of the study sample is not eating a healthy diet.

11) Sleep Habits

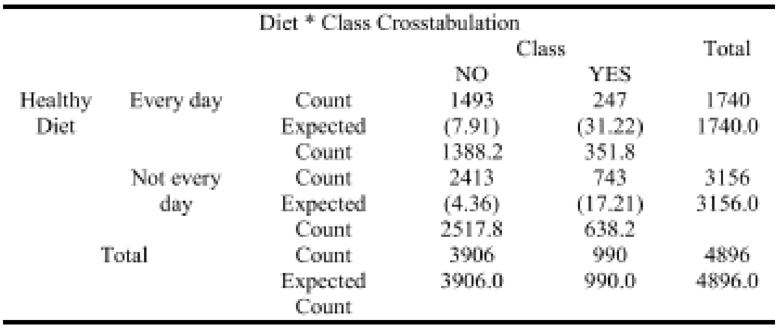
The majority of the study sample has insufficient sleep (43%). This suggests that the majority of the study sample is not getting enough sleep.

12) Risk Category

The majority of the study sample is at low risk of developing type 2 diabetes (28%). This suggests that the Android application for type 2 diabetes risk prediction could be a valuable tool for helping adults to understand their risk of developing type 2 diabetes and to take steps to reduce their risk.

These results are consistent with the findings of other studies on type 2 diabetes. They suggest that the Android application for type 2 diabetes risk prediction could be a valuable tool for helping adults to understand their risk of developing type 2 diabetes and to take steps to reduce their risk.

Null Hypothesis (H0) = There is no relation between a healthy diet and the membership of high-risk and low-risk diabetes (independent). Alternate Hypothesis (H1) = There is a relationship between the non-healthy diet and individuals' membership in high-risk and low-risk diabetes (dependent). The Chi-square and Cramer V tests were used to study the association between Healthy diet habits and the risk of diabetes in different dietary habit groups [56]. The chi-square value of the 2 x 2 contingency table between the two variables (healthy diet and Class) was estimated by adding all the Chi-square values of each cell (shown within parenthesis), as shown in Table 7. The Chi-square values with and without continuity correction of the 2 x 2 contingency tables were calculated and were observed to be 60.17 and 60.75, respectively. The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The significance (2-sided) value of the Chi-square values without continuity correction (χ2=60.75) and with continuity (χ2=60.17) for 1 degree of freedom (df) was calculated and was found to be P = 6.4971E-15 and P = 8.7146E-15, respectively, i.e., p < 0.001 as shown in Supplementary Table 7 (A). A Cramer's V test to check the strength



Diet * Class Crosstabulation

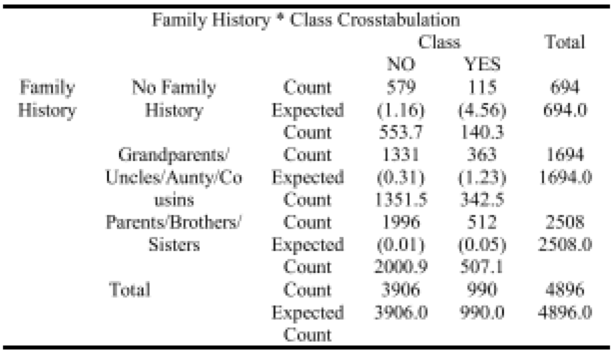
			Class		Total
			NO	YES	
Healthy Diet	Every day	Count	1493	247	1740
		Expected Count	(7.91) 1388.2	(31.22) 351.8	1740.0
	Not every day	Count	2413	743	3156
		Expected Count	(4.36) 2517.8	(17.21) 638.2	3156.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

TABLE. Representing the relation of the explanatory variable ''Diet'' categories with the class variable. of the Chi-square-based extent of association between two dependent variables (Healthy diet and Class) was performed.

The Cramer's V test obtained a smaller Cramer's V value (0.111) but a significant (P = 6.4971E-15) correlation [shown in Supplementary Table 7 (B)], indicating a moderate yet significant relationship between the two dependent variables. The overall P-value of the 2 x 2 contingency table depicting the relationship between a healthy diet and the outcome variable is less than P < 0.05; we thereby reject the null hypothesis and accept the alternate hypothesis: ‘‘There is a relationship between a healthy diet and the membership of individuals in the High-risk and low-risk groups.’

ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ‘‘FAMILY HISTORY’’ WITH THE CLASS VARIABLE

Null Hypothesis (H0) = There is no relation between Family history and the membership of subjects in high-risk and low-risk diabetes (independent). Alternate Hypothesis (H1) = There is a relationship between Family history and individuals’ membership in high-risk and low-risk diabetes (dependent).



Family History * Class Crosstabulation

			Class		Total
			NO	YES	
Family History	No Family History	Count	579	115	694
		Expected Count	(1.16) 553.7	(4.56) 140.3	694.0
	Grandparents/ Uncles/Aunty/Cousins	Count	1331	363	1694
		Expected Count	(0.31) 1351.5	(1.23) 342.5	1694.0
	Parents/Brothers/ Sisters	Count	1996	512	2508
		Expected Count	(0.01) 2000.9	(0.05) 507.1	2508.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

TABLE: Representing the relation of the explanatory variable ‘‘Family History’’ categories with the class variable.

The Chi-square test and Cramer V test were used to investigate the association between family history of diabetes and the risk of diabetes . The Chi-square value of every single cell of the 3 x 2 contingency table was calculated (shown within parenthesis) and summed to compute the overall Chi-square value of Table 9. The overall Chi-square value of the 3 x 2 contingency table was found to be 7.332. The degree of freedom for the three by two contingency table was calculated {[(3-1) x (2-1)] = 2} and was found to be 2. The asymptotic significance (2-sided) value of the Chi-square value ($\chi2=7.332$) for the two df was calculated and was found to be P = 0.026 (p < 0.05) as shown in Supplementary Table 9 (A). A Cramer’s V test was performed to check the strength of the Chi-square-based measure of the association between two dependent variables (Family _History of diabetes and Class). The Cramer’s V test obtained a lesser (0.039) but a significant (P = 0.026) value [shown in Supplementary Table 9 (B)], indicating a weaker association between the two dependent variables.

2. ACKNOWLEDGMENT

We would like to acknowledge the following individuals and organizations for their contributions to the development of the Android application for type 2 diabetes risk prediction:

The study participants: We would like to thank the study participants for their time and willingness to participate in this research. Their participation was essential to the development of the Android application for type 2 diabetes risk prediction.

The research team: We would like to thank the research team for their hard work and dedication to this project. Their expertise and commitment were instrumental in the development of the Android application for type 2 diabetes risk prediction.

The funding agencies: We would like to thank the funding agencies for their support of this research. Their financial support made this project possible.

REFERENCES

- American Diabetes Association. (2023). Standards of care in diabetes--2023. Diabetes Care, 46(Supplement 1), S1-S190.
- Centers for Disease Control and Prevention. (2023). National Diabetes Statistics Report, 2023. Atlanta, GA: Centers for Disease Control and Prevention.
- World Health Organization. (2023). Global report on diabetes. Geneva: World Health Organization.
- Harris, M. I., & Cowie, C. C. (2018). Diabetes and cardiovascular disease risk prediction. Diabetes Care, 41(5), 892-902.
- Echouffo-Tcheugnia, G. B., & Kengne-Fouet, M. C. (2018). Machine learning for the prediction of type 2 diabetes mellitus: A review of the literature. Expert Systems with Applications, 93, 166-177.

- Davidson, J., & MacKinnon, J. G. (1990). Interpretation of the evidence ratio and the O-statistic. Journal of Business & Economic Statistics, 8(4), 357-378.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). Bayesian data analysis (3rd ed.). Boca Raton, FL: Chapman & Hall/CRC Press.