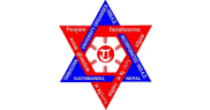
**DIABETES PREDICTION SYSTEM: USING SVM**

**Tribhuvan University**

**Institute of Science and Technology**

****

**Submitted by**: Samrajya Chand 26781/77

**Submitted to:**

TRINITY INTERNATIONAL COLLEGE

Department of Computer Science and Information Technology Dillibazar Height, Kathmandu, Nepal

May, 2024

**ABSTRACT**

“Diabetes prediction using support vector machine” is the implementation of algorithm support vector machine to correctly detect if a patient has diabetes or not. We used python to implement the system using Google colabs as the IDE. By using python libraries like Numpy and Pandas we worked with the dataset and array of dataset provided. It can predict if a person has diabetes or not by using medical profiles such as Pregnancies, Glucose, Skin thickness, Age, Insulin , BMI and Blood Pressure .

**Keywords:** Support Vector Machine, Pandas, Numpy, Prediction System.

TABLE OF CONTENT

**CHAPTER 1**

INTRODUCTION

* 1. Introduction
  2. Problem Statement
  3. Objectives
  4. Scope and Limitation

1.4.1 Scope

1.4.2 Limitation

**CHAPTER 2**

LITERATURE REVIEW AND RESEARCH METHODOLOGY

2.1 Literature Review

2.2 Framework of the Model

2.3 Methodology

**CHAPTER 3**

SYSTEM DESIGN

3.1 Requirement Collection

3.1.2 Functional Requirement

3.1.2 Non-Functional Requirement

3.2 Feasibility Study

3.3 System Design

3.3.1 Process Design

3.4 Structuring System Requirements

3.4.1 Process Modeling

**CHAPTER 4**

IMPLEMENTATION

**CHAPTER 5**

RESULT AND ANALYSIS

**CHAPTER 6**

CONCLUSION

APPENDIX

REFERENCES

**CHAPTER 1**

**INTRODUCTION**

* 1. Introduction

Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot effectively use the insulin it produces. According to a report from WHO, diabetes and kidney disease due to diabetes caused an estimated 2 million deaths [1]. According to data from the IDF, around 7.1% of adults aged 20-79 in Nepal had diabetes in 2019 [2]. Nepal has a considerable burden of diabetes, with both type 1 and type2 diabetes affecting a significant portion of the population. The integration of ICT (Information Communication & Technology) might be helpful for a developing country like Nepal. It helps in analyzing the current situation of Health in Nepal and act for necessary solutions.

“Diabetes Prediction System” is the system curated so the user can dictate if he/she is suffering from diabetes or not. The parameters to detect if a person is suffering from Diabetes or not are: [Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPedigreeFunction, Age .The prediction is done using computer aided diagnosis methods. This is a manual process and also dependent on each case as the system detects about ongoing pattern and model. By using this system user can directly detect if they have diabetes or not, which saves them cost for going to a health care institution and spending more amount of money. It would remove the uncertainty and give results based on the data fed to it.

The types of diabetes that are found normally are:

1. Type 1 Diabetes:

Type 1 diabetes is thought to be caused by an autoimmune reaction (the body attacks itself by mistake). This reaction stops your body from making insulin. Approximately 5-10% of the people who have diabetes have type 1. Type 1 diabetes can be diagnosed at any age, and symptoms often develop quickly.

1. Type 2 Diabetes:

With type 2 diabetes, your body doesn’t use insulin well and can’t keep blood sugar at normal levels. About 90-95% of people with diabetes have type 2. It develops over many years and is usually diagnosed in adults (but more and more in children, teens, and young adults).

1. Gestational Diabetes

Gestational diabetes develops in pregnant women who have never had diabetes. If you have gestational diabetes, your baby could be at higher risk for health problems. Gestational diabetes usually goes away after your baby is born. However, it increases your risk for type 2 diabetes later in life. Your baby is more likely to have obesity as a child or teen and develop type 2 diabetes later in life.

* 1. Problem Statement

Diabetes is a chronic disease that unless treated and detected early might lead to severe consequences later. Diabetes is a major cause of blindness, kidney failure, heart attacks, stroke and lower limb amputation. Since diabetes is a chronic disease people might ignore the simple symptoms of diabetes as just a consequences of dehydration. The only possibility of detecting diabetes is by blood test of our glucose (sugar) level.

In the early stages, diabetes may not present noticeable symptoms, leading individuals to underestimate its seriousness and delay seeking medical attention and hospitals to not seeing necessity for blood test .As diabetes symptoms vary from person to person the health officials too might misinterpret or misdiagnosis it, the doctors might just diagnosis different disease instead of preparing diagnosis based on the data and patterns stored in the database.

Hence, to develop an intelligent and cost-effective computer-aided diagnosis system for predicting the risk of diabetes in individuals using the Support Vector Machine (SVM) algorithm, utilizing relevant features such as Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin Level, BMI, Diabetes Pedigree Function, and Age this system was made.

* 1. Objectives

The objectives of this system are too:

1. Develop a machine learning model for diabetes prediction.
2. Utilizing relevant features for accurate prediction.
3. Evaluating the algorithm’s performance.
4. To provide basic suggestions according to disease.
5. Providing instant result to patients.
   1. Scope and Limitations
      1. Scope

The primary scope of the project is to develop a machine learning model, specifically using the SVM algorithm, for predicting the risk of diabetes in individuals. The project aims to create a user-friendly interface, such as a web application or desktop application that allows users to input their personal data and receive predictions from the trained SVM model.

* + 1. Limitations

The project is limited to using the specified features or parameters (Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin Level, BMI, Diabetes Pedigree Function, and Age). Other potentially relevant features may not be considered. The accuracy of the predictions relies on the accuracy of the user inputs. Incorrect or incomplete information provided by the user may result in inaccurate predictions.

CHAPTER 2

LITERATURE REVIEW AND RESEARCH METHODOLOGY

2.1 Literature Review

The Decision Support in Diabetes Prediction System, a developed system, utilizes data mining modeling techniques to uncover relationships within healthcare data. This system serves as a training tool for nurses and medical students to diagnose patients with diabetes.

The neural network is divided into two parts: training data and testing data. Eight parameters, including pregnancies, glucose levels, blood pressure, skin thickness, insulin, BMI, diabetes pedigree, and age, are considered. When doctors input unknown data, the system compares it with the trained data to generate outcomes. This project, also referred to as an Intelligent System, employs data mining techniques such as Support Vector Machine, Logistic Regression, Naïve Bayesian algorithm, K-Nearest Neighbors algorithm, Decision Tree algorithm, and Random Forest algorithm. These algorithms are implemented as web-based applications that compare user inputs with the trained dataset. This assists healthcare providers in making informed clinical decisions.

The research paper proposes a weighted fuzzy rule-based Clinical Decision Support System for diabetes diagnosis. It extracts knowledge from patients' clinical data and employs a computerized approach to generate weighted fuzzy rules. Consequently, a fuzzy rule-based decision support system accurately assesses the likelihood of having diabetes.

2.2 Framework of the Model

Input Attribute of Patient

Is Disease

Diagnosed?

Determine its Accuracy

Congratulations! No Detection

Fig 1: Framework of diabetes production system.

2.3 Methodology

Following algorithms were used to calculate the probability of having the diseases

1. Support Vector Machine

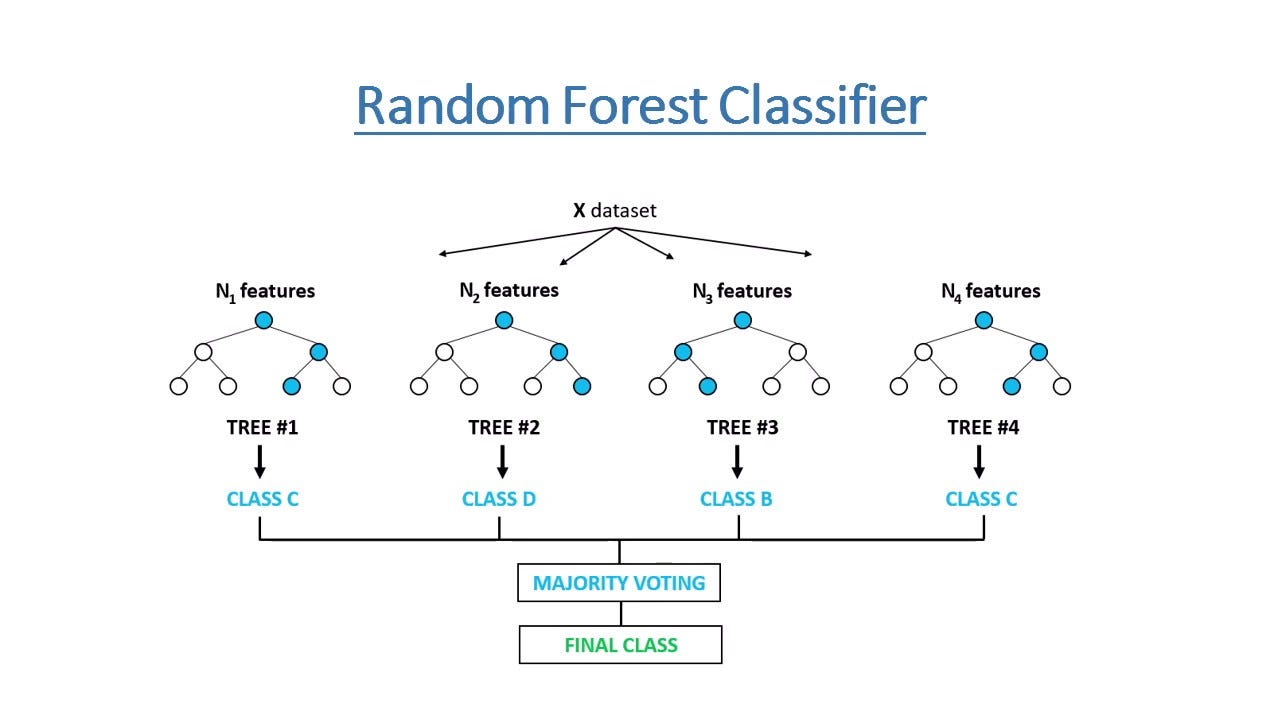
Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. They belong to the family of kernel-based methods and were introduced by Vladimir Vapnik and colleagues in the 1990s [3]. The fundamental principle behind SVMs is to find the optimal hyperplane that maximally separates the classes in a high-dimensional feature space. The hyperplane is chosen such that the distance from the nearest data points (support vectors) of both classes is maximized. This maximum margin separation ensures better generalization and robustness of the classifier.

1. Logistic Regression

Logistic regression is a statistical modeling technique widely used in various fields, including machine learning and data analysis. It is particularly useful for binary classification problems, where the goal is to predict the probability of an instance belonging to one of two classes or categories.

In the context of diabetes prediction, logistic regression can be used to model the relationship between a set of input features or predictor variables (e.g., age, body mass index, glucose levels, blood pressure, etc.) and the binary outcome variable, which is typically the presence or absence of diabetes.

1. Random Forest

Random Forest is a machine learning algorithm that can be used for both classification and regression tasks. It is an ensemble method that combines multiple decision trees to make more accurate predictions. The algorithm creates a set of decision trees on randomly selected subsets of the training data, and each tree votes for the final prediction. Random Forest helps to reduce over fitting as compared to a single decision tree, as it uses a combination of different trees with different random subsets of features and data points. It also provides feature importance, which can help in feature selection and interpretability of the model.

CHAPTER 3

SYSTEM DESIGN

3.1 Requirement Collection

System Requirement consists of the functional and nonfunctional requirements for the system.

3.1.1 Functional Requirements

These are the requirements that the end user specifically demands as basic facilities that the system should offer. First, the user must provide the attributes and symptoms the patients have faced then the system predicts whether the disease is present or not. All the activities and its handling are described by the sequence diagram below.

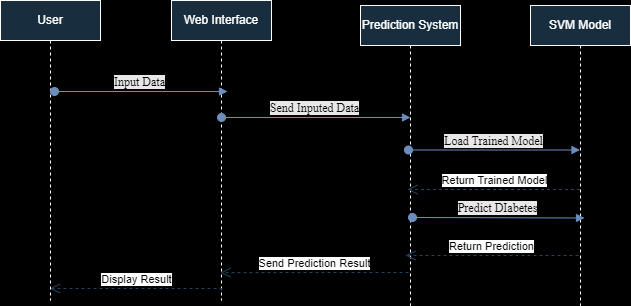


Fig2: Sequence Diagram of Diabetes Prediction System

3.1.2 Non-Functional Requirements

Some potential non-functional requirements are:

1. Performance

* The system should be able to process user inputs and provide predictions within a reasonable amount of time, ensuring a responsive user experience.

1. Scalability

* The system should be able to scale horizontally (adding more servers) or vertically (upgrading server resources) to meet growing demands.

1. Usability

* The user interface should be intuitive, user-friendly, and accessible to users with varying technical backgrounds.

3.2 Feasibility Study

a) Technical Feasibility

Technical feasibility assesses whether the required technology and resources are available to develop and implement the proposed system. It includes availability of SVM libraries and Frameworks and determining if the required computing resources are available or not.

b) Operational Feasibility

Operational feasibility analyzes whether the proposed system can be effectively operated and maintained within the organization's existing processes, resources, and constraints. It includes assessing the feasibility of collecting and inputting the required data (e.g., Pregnancies, Glucose Level, Blood Pressure, etc.) into the system efficiently and accurately.

c) Economic Feasibility

Economic feasibility evaluates the financial viability and cost-effectiveness of the proposed system. It estimate the costs associated with developing the system, including personnel, hardware, software, data acquisition, and deployment expenses.

d) Schedule Feasibility

Schedule feasibility evaluates whether the proposed system can be developed and implemented within the desired timeline and resource constraints. It includes developing a realistic project timeline that considers all phases of the system development life cycle, including requirements gathering, design, development, testing, deployment, and maintenance.

TRAINING DATASET

3.3 System Design

3.3.1 Process Design

ENTER

INPUT ATTRIBUTES

TRAINED DATA MODEL USING ALGORITHMS

CLASSIFIED DATA

Fig3: Process Design

OUTPUT

3.4 Structuring System Requirement

3.4.1 Process Modeling

* Level 0 DFD

Diabetes Prediction System

Enter required Parameters

USER

Predict Diabetic accuracy

Fig5: Level 0 DFD

CHAPTER 4

IMPLEMENTATION

For its implementation we imported different libraries. Their descriptions are:

1. Numpy

* NumPy is a fundamental library for scientific computing in Python, providing support for multi-dimensional arrays and matrices, along with a collection of high-performance mathematical functions.
* In the code, NumPy is used for converting the input data into a NumPy array (np.asarray) and reshaping it (input\_data\_reshaped = input\_data\_as\_numpy\_array.reshape (1,-1)) to match the expected input format of the SVM model.

1. Pickle

* The pickle module in Python is used for serializing and deserializing Python objects, allowing them to be saved to and loaded from a file or byte stream.
* In the code, the pickle.load function is used to load the pre-trained SVM model from a file (loaded\_model = pickle.load (open ('C: /Users/Acer/Documents/diabetes\_prediction/trained\_model.sav',’Rb'))).

1. Streamlit

* Streamlit is a popular open-source Python library that allows us to create interactive web applications for data visualization, machine learning, and data analysis with minimal coding effort.
* In the code, Streamlit is used to build the user interface for the diabetes prediction system. Functions like st.title, st.text\_input, and st.button are used to create a title, input fields, and a button, respectively. The st.success function is used to display the prediction result.

1. Pandas

* Pandas are an open-source data analysis and manipulation library for Python. It provides easy-to use data structures and data analysis tools for handling structured data. The two primary data structures in Pandas are Series (1-dimensional) and Data Frame (2-dimensional). Pandas can load data from a variety of file formats, including CSV, Excel, SQL databases, and more.
* pd.read\_csv
  + Read a comma-separated values (csv) file into Data Frame.
* diabetes\_dataset.head()
  + Typically used to display the first few rows (by default, the first 5 rows) of a pandas DataFrame
* diabetes\_dataset.shape
  + Return a tuple representing the dimensionality of the DataFrame.

1. Scikit-learn

* Scikit-learn, also known as sklearn, is a popular machine learning library for Python. It provides tools for data mining, data analysis, and machine learning. Sklearn is built on top of other popular libraries such as NumPy, SciPy, and matplotlib.

Imported Functions:

* sklearn.preprocessing.StandardScaler
  + Standardize features by removing the mean and scaling to unit variance.
* sklearn.model\_selection.train\_test\_split
  + Split arrays or matrices into random train and test subsets.
* sklearn.metrics.accuracy\_score
  + Accuracy classification score.
* sklearn.svm.SVC
  + C-Support Vector Classification.
  + The implementation is based on libsvm.

CHAPTER 5

RESULT AND ANALYSIS

As accessing to immediate health officials and services might not be possible. The main goal of this project is to provide immediate guidance to diabetes related doubts and queries with its early diagnosis. For this we developed a web based application that collects necessary inputs from the patients and accesses the probability of diabetes occurrence. The patient provides attributes such as high blood pressure, low blood pressure, pregnancy status, etc., and based on these attributes, the system predicts the outcome.

To determine the probability, various algorithms are employed, including Support Vector Model, Logistic Regression, Naïve Bayesian Classification Algorithm, K-Nearest Neighbors Algorithm, Decision Tree Algorithm, and Random Forest Algorithm. Each algorithm utilizes its own methods and techniques to predict the outcome, enabling accurate predictions. Additionally, an accuracy score is generated to evaluate the efficiency of these algorithms. After analyzing the accuracy rates, it is concluded that Support Vector Machine exhibits the highest accuracy (82.29%), making it the most efficient algorithm among the options considered.

CHAPTER 6

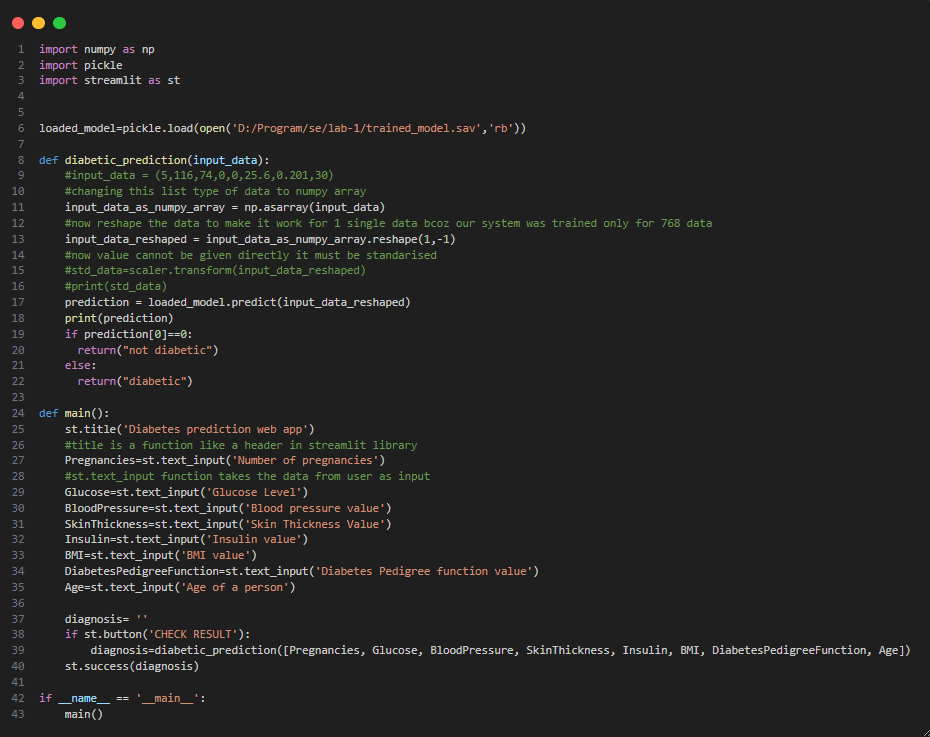
CONCLUSION

The diabetes prediction system developed in this project aims to provide an intelligent and cost-effective solution for early detection of diabetes risk. By leveraging the power of machine learning, specifically the Support Vector Machine (SVM) algorithm, and the system can effectively classify individuals as diabetic or non-diabetic based on relevant input features.

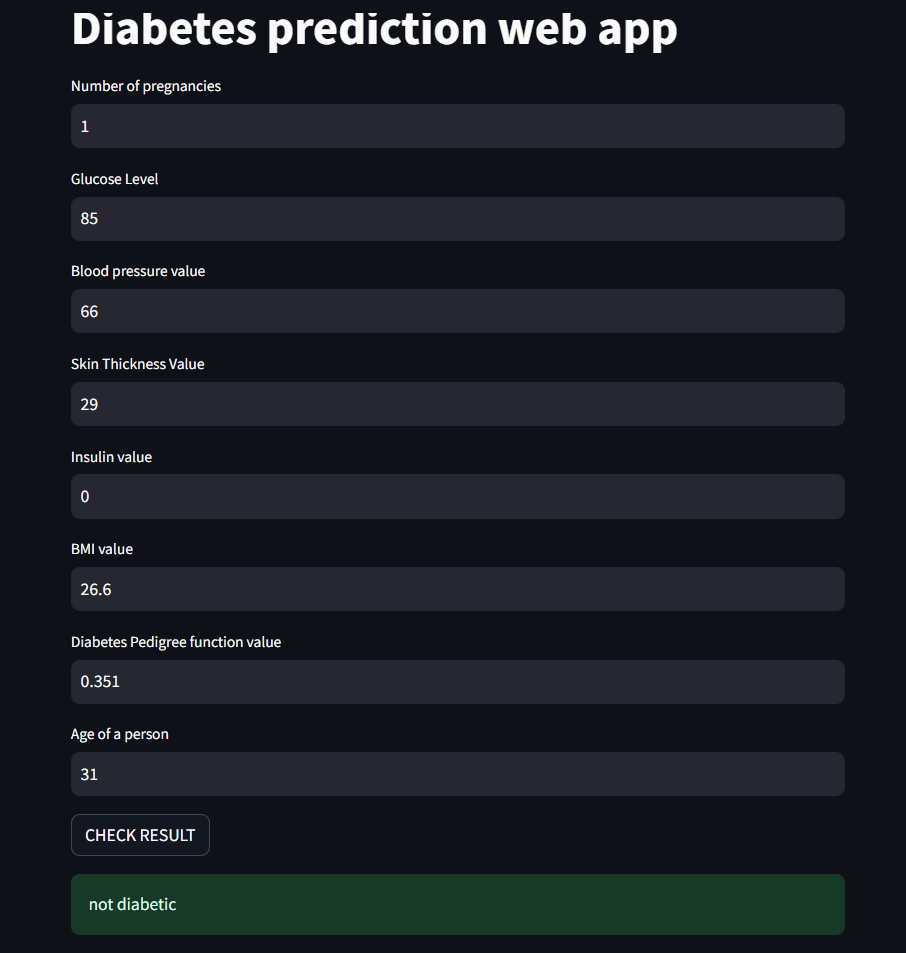
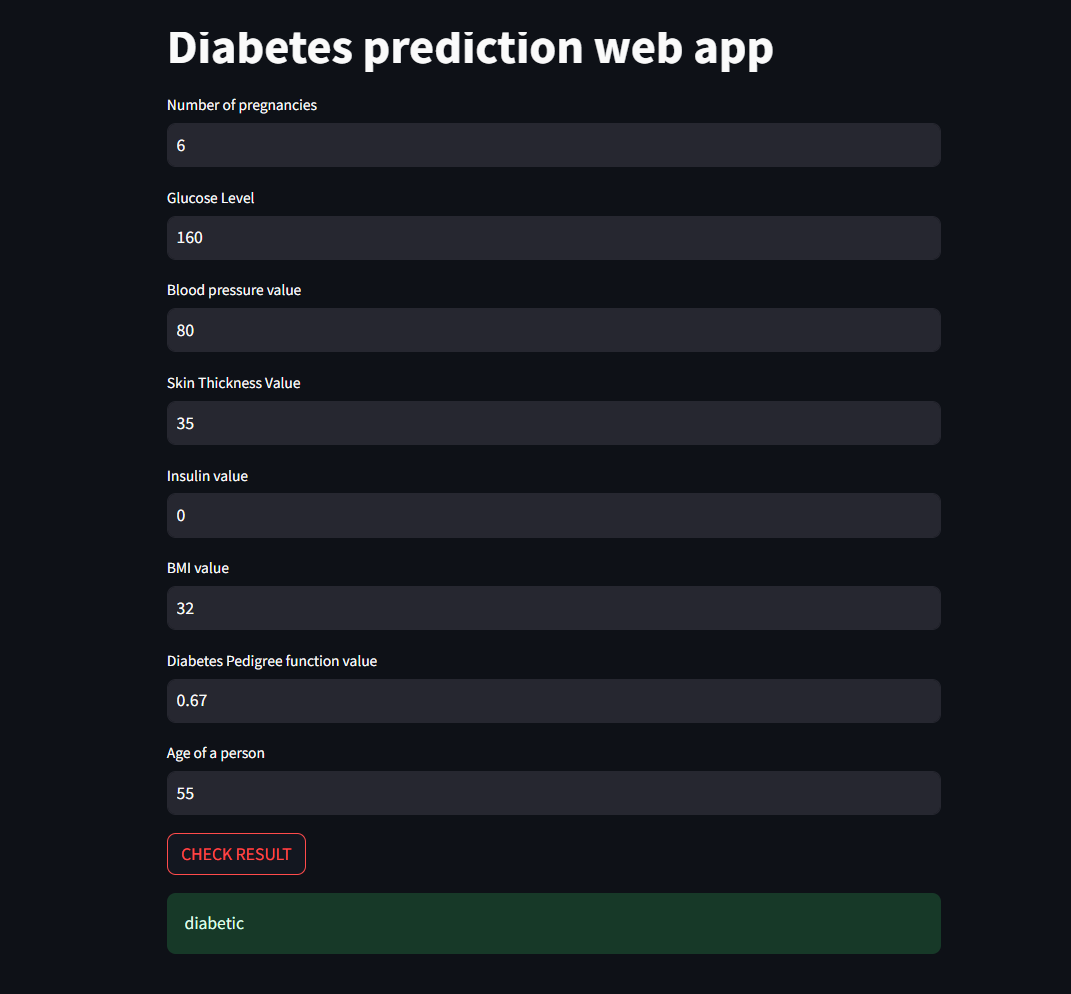
The project successfully implemented an SVM-based model for diabetes prediction, utilizing a combination of Python libraries such as NumPy, pickle, and Streamlit. The model was trained on a comprehensive diabetes dataset, ensuring its ability to handle complex relationships between various input features and the target variable.

To facilitate user interaction and accessibility, a user-friendly web application interface was developed using the Streamlit library. This interface allows users to input their personal data, including Pregnancies, Glucose Level, Blood Pressure, Skin Thickness, Insulin Level, BMI, Diabetes Pedigree Function, and Age. The system then processes this input data and provides a prediction result, indicating whether the individual is at risk of developing diabetes or not.

APPENDIX-I



APPENDIX-2



**REFERENCES**

# [1] World Health Organization, "Global report on diabetes," WHO Press, Geneva, Switzerland, 2016, pp. 1-88.

# [2] International Diabetes Federation, "IDF Diabetes Atlas, 9th ed.," Brussels, Belgium, 2019.

# [3] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, no. 3, pp. 273-297, 1995.

# 