

INTERNSHIP PROJECT REPORT

On

**Diabetes Prediction using Support Vector Machines**

Submitted by:

K. POORNATEJITHA

E. AJITH

S. SAI RAM

Under the guidance of

Prof. T. Kishore Kumar

Professor, Department of ECE

NIT Warangal

**Centre for Training and Learning**

NATIONAL INSTITUTE OF TECHNOLOGY WARANGAL

**HANAMKONDA - 506004, TELANGANA, INDIA**

**BONAFIED CERTIFICATE**

This is to certify that this project report entitled **“DIABETES PREDICTION USING SUPPORT VECTOR MACHINES(SVM)”** submitted to National Institute of Technology, Warangal and National Institute of Technology, Srinagar, is a Bonafide record of work done by **“K. Poornatejitha, E. Ajith, S. Sai Ram”** under my supervision from **“15 Dec 2023”** to **“20 Jan 2024”.**

**Supervisor**

Prof. T. Kishore Kumar

Professor, Department of ECE

NIT Warangal

Place: Warangal

Date: 19 June 2024

**DECLARATION**

This is to declare that this report has been written by us. No part of the report is plagiarized from other sources. All information included from other sources have been duly acknowledged. We aver that if any part of the report is found to be plagiarized, we are shall take full responsibility for it.

Place: Warangal

Date: 19 June 2024

**ACKNOWLEDGEMENT**

Completion of this project and thesis would not have been possible without the help of many people, to whom we are very thankful.

We would like to express my sincere gratitude to our supervisor Dr. T. Kishore Kumar, Professor, Head of CTL, NIT Warangal. His constant motivation, guidance and support helped us a great deal to achieve this feat.

We also thank Dr. K. Sunil Kumar, PhD Scholar for his constant support in guiding us about the academic related work.

We wish a deep sense of gratitude and heartfelt thanks to management for providing excellent lab facilities and tools. Finally, we thank our seniors whose guidance helped us in this regard

**ABSTRACT**

Diabetes is a serious metabolic disorder that affects millions of people globally. Early detection and management of diabetes are essential to prevent severe complications. In recent years, machine learning algorithms have become increasingly popular in the medical field to predict the onset of diabetes. This study aims to predict the onset of diabetes using the support vector machine (SVM) and decision tree algorithms. The dataset used for this study is the Pima Indian diabetes dataset, which contains several features such as glucose, insulin, and body mass index. The problem statement is to determine which algorithm is more accurate in predicting diabetes. The methodology involves implementing the SVM and decision tree algorithms on the dataset and evaluating their performance using metrics such as accuracy, precision, and recall. The results of the study indicate that the SVM algorithm performs better than the decision tree algorithm, with an accuracy of 76.6% compared to 75%. This work concludes that the SVM algorithm is more accurate in predicting diabetes and can be a valuable tool for early detection and management of the disease. This study provides insight into the potential use of machine learning algorithms in the medical field and highlights the need for further research to improve the accuracy of diabetes prediction models.

**TABLE OF CONTENTS**

**CHAPTER PAGE NO.**

CERTIFICATE………………………………………………………… 2

DECLARATION………………………………………………………. 3

ACKNOWLEDGEMENT……………………………………………... 4

ABSTRACT……………………………………………………………. 5

TABLE OF CONTENTS………………………………………………. 6

LIST OF FIGURES……………………………………………………. 7

|  |  |  |
| --- | --- | --- |
| **CHAPTER NO.** | **TITLE** | **PAGE NO** |
| 1 | INTRODUCTION   * 1. Introduction   2. Aim   3. Objectives | 7-8 |
| 2 | LITERATURE REVIEW  2.1 Diabetes Prediction  2.2 Support Vector Machines in Healthcare  2.3 Previous Studies  2.4 Challenges and Opportunities | 9-10 |
| 3 | GAPS IDENTIFIED | 11 |
| 4 | DESIGN METHODOLOGY | 12-13 |
| 5 | FLOWCHART | 14-16 |
| 6 | DATA ANALYSIS  6.1 Structure of data  6.2 Parameters Implemented  6.3 Exploratory Data Analysis  6.4 Splitting of dataset | 17-21 |
| 7 | DESIGNING INTERFACE | 22-23 |
| 8 | RESULTS | 24-27 |
| 9 | CONCLUSION | 28 |
| 10 | REFERENCE | 29 |

**LIST OF FIGURES**

**FIGURE DESCRIPTION OF FIGURES PAGE NO**

5.1 Flowchart of diabetes prediction using svm 14

5.2 Workflow 16

5.3 A simple steps of diabetes prediction 17

6.1 Importing libraries 18

6.2 Loading Dataset 18

6.3 Understanding the data we loaded 18

6.4 Calculating average of outcome 0 and 1 19

6.5 Calculating Mean, Count, Min, Max and 20

Standarddeviation.

6.6 Data splitting 20

6.7 Accuracy of training and testing dataset 21

6.8 Code used to get output 22

6.9 Changing the filename of trained model and 23

loading to create an interface.

7.1 Importing Modules 24

7.2 Code of Interface 25

8.1 Interface of diabetes prediction 26

8.2 Interface of chatbot 26

8.3 Entering the values 27

8.4 Output (1) diabetic 27

8.5 Entering another set of values 28

8.6 Output (2) not diabetic 28

**1.INTRODUCTION**

**1.1 Introduction**

Various classification strategies are used in the medical field for classifying data into different classes. Diabetes is a condition that affects the body's ability to produce the hormone insulin, which causes carbohydrate metabolism to become irregular and blood glucose levels to increase. High blood sugar is a common symptom of diabetes. If diabetes is not treated, it can lead to a variety of complications. Diabetic ketoacidosis and nonketotic hyperosmolar coma are two significant complications. Diabetes is considered a severe health problem in which the amount of sugar in the blood cannot be regulated. Diabetes is influenced by a variety of factors such as height, weight, genetic factors, and insulin, but the most important factor to remember is sugar concentration. The only way to avoid problems is to identify the problem early. This dataset comes from the ‘National Institute of Diabetes and Digestive Diseases’ Pima Indians Diabetes Database (PIDD). Several constraints were taken from the massive database. But we have changed the dataset we took by deleting some unwanted parameters. The dataset is divided into three sections, after which classification techniques are used. The training dataset is a sample of the dataset that is used to match the model. Validation Dataset, a dataset sample used for fine-tuning parameters and comparing model output accuracy and error rates between the training and validation datasets. Testing Dataset is a sample of a dataset that is used to assess the model's output. Various machine learning techniques are implemented. Confusion matrix is obtained and is compared with all classification algorithms. This comparison of the various machine learning techniques shows which algorithm is better suited for diabetes prediction. Correlation between parameters and the best accuracy score using various supervised machine learning algorithms is obtained.

**1.2 AIM OF THE PROJECT**

The project aims to develop a diabetes predictive model using SVM algorithm for early detection of diabetes and early healthcare. The project aims to utilize Support Vector Machine (SVM) algorithm for diabetes prediction by analyzing relevant health data. Through robust model development and training, it seeks to accurately classify individuals as diabetic or non-diabetic based on key features such as glucose levels, BMI, age.

**1.2 Objectives**

* Since a decade, the number of people diagnosed with diabetes has risen significantly. The current human lifestyle is the primary cause of diabetes rise.
* Main objective of this project is to analyse the data, and see if it is possible to gleam any further information from the data to determine correlation between parameters and diabetes.
* The second is to attempt to get the best accuracy score using various supervised learning machine learning algorithms. To find out which algorithm is able to best predict whether a person has diabetes or not based on this dataset.
* The accuracy of the algorithms used are compared and discussed. The study's comparison of the various machine learning techniques shows which algorithm is better suited for diabetes prediction. Using machine learning methods, this project aims to assist doctors and physicians for predicting whether a person has diabetes or not.

**2.LITERATURE REVIEW**

**2.1 Diabetes Prediction:**

Diabetes mellitus is a chronic metabolic disorder characterized by high blood sugar levels over a prolonged period, resulting from inadequate insulin production, ineffective insulin utilization, or both. Given its increasing prevalence globally and its significant impact on public health, researchers have turned to machine learning approaches to improve diabetes prediction, diagnosis, and management.

Various studies have explored the use of machine learning algorithms, including SVM, to predict diabetes based on patient data such as demographic information, medical history, and physiological parameters. These studies often employ datasets containing features such as glucose levels, body mass index (BMI), family history, and lifestyle factors.

**2.2 Support Vector Machines in Healthcare:**

Support Vector Machines (SVM) have gained prominence in healthcare applications due to their ability to handle high-dimensional data and nonlinear relationships effectively. SVM is a supervised learning algorithm that aims to find the optimal hyperplane that separates classes in a feature space while maximizing the margin between them.

In the context of healthcare, SVM has been utilized for various tasks such as disease diagnosis, prognosis, risk prediction, and treatment outcome prediction. Its versatility and robust performance make it suitable for analyzing complex medical datasets and extracting meaningful patterns for clinical decision-making.

**2.3 Previous Studies:**

Several previous studies have demonstrated the efficacy of SVM in predicting diabetes and related outcomes. For example, [cite relevant study] developed an SVM-based model to predict the onset of type 2 diabetes using electronic health records (EHR) data, achieving high accuracy and providing actionable insights for preventive interventions.

Another study by [cite relevant study] utilized SVM to predict diabetic retinopathy, a common complication of diabetes, based on retinal images. The SVM model achieved competitive

performance in detecting the presence and severity of diabetic retinopathy, highlighting its potential for early diagnosis and treatment monitoring.

**2.4 Challenges and Opportunities:**

While SVM and other machine learning techniques show promise in diabetes prediction and healthcare applications, several challenges remain. These include issues related to data quality, feature selection, model interpretability, and generalization to diverse patient populations.

Future research in this field should focus on addressing these challenges while exploring innovative approaches to improve the accuracy, robustness, and clinical utility of predictive models. Collaborations between healthcare professionals, data scientists, and policymakers are essential to ensure the successful translation of machine learning technologies into clinical practice and public health initiatives.

**3.GAPS IDENTIFIED**

**1.Data Imbalance:**

**Issue:** Many datasets used for diabetes prediction are imbalanced, with a higher proportion of non-diabetic cases than diabetic cases. This imbalance can lead to biased models that prioritize accuracy over sensitivity, resulting in potentially missed diagnoses of diabetes.

**Impact:** The model may exhibit poor performance in correctly identifying diabetic cases, leading to inadequate support for preventive interventions or treatments.

**Solution:** Employ techniques such as oversampling, under sampling, or synthetic data generation to balance the distribution of diabetic and non-diabetic cases in the training dataset.

**2. Limited Feature Representation:**

**Issue:** Existing studies often incorporate a limited set of patient features such as glucose levels and BMI, overlooking other potentially relevant factors such as genetic predisposition, lifestyle habits, or socioeconomic status.

**Impact:** The model may fail to capture important predictors of diabetes risk, limiting its predictive accuracy and clinical utility.

**Solution:** Conduct comprehensive feature engineering to extract meaningful insights from the dataset, incorporating additional biomarkers, genetic factors, and lifestyle variables. Employ feature selection algorithms to identify the most relevant features for model training.

**3. Model Interpretability:**

**Issue:** While Support Vector Machines (SVM) are known for their high accuracy, they often lack interpretability, making it challenging for clinicians to understand the underlying decision-making process.

**Impact:** Lack of interpretability may hinder the model's adoption in clinical practice, as clinicians require transparency in model predictions to make informed decisions.

**Solution:** Prioritize model interpretability by employing techniques such as feature importance analysis, partial dependence plots, and local interpretable model-agnostic explanations (LIME). These approaches enable elucidation of the contributions of individual features to the SVM model's predictions, enhancing its transparency and usability in clinical settings.

**4.DESIGN METHODOLOGY**

**1. Data Preprocessing:**

Clean the dataset by handling missing values, outliers, and inconsistencies.

Normalize numerical features to ensure uniform scale across variables.

Encode categorical variables and address any data quality issues.

**2. Feature Engineering and Selection:**

Conduct comprehensive feature engineering to extract meaningful insights from the dataset.

Explore additional biomarkers, genetic factors, and lifestyle variables to enhance feature representation.

Employ feature selection algorithms such as recursive feature elimination (RFE) or L1 regularization to identify the most relevant features for model training.

**3. Model Development:**

Utilize Support Vector Machines (SVM) as the classification algorithm due to its effectiveness in handling high-dimensional data and nonlinear relationships.

Optimize model hyperparameters using techniques such as grid search or random search.

Evaluate model performance using appropriate metrics such as accuracy, precision, recall, and F1-score.

**4. Model Interpretability:**

Prioritize model interpretability by employing techniques such as feature importance analysis, partial dependence plots, and local interpretable model-agnostic explanations (LIME).

Ensure transparency in model predictions to facilitate integration into clinical practice and decision-making.

**5. Evaluation and Validation:**

Evaluate model performance using cross-validation techniques to ensure robustness and generalizability.

Test the model on an independent validation dataset to assess its performance in real-world scenarios.

Assess the clinical utility of the model by conducting cost-benefit analyses and evaluating its impact on patient outcomes and healthcare resource utilization.

By addressing the identified gaps and following a rigorous design methodology, the project aims to develop a reliable and interpretable SVM model for diabetes prediction that can assist clinicians in early detection and management of the disease.

**5.FLOWCHART**

**TRAINING DATASET**

**TEST DATA**

**SVM ALGORITHM**

**RESULT ANALYSIS**

**RESPONSE**

**PRE**

**PROCESSING**

**5.1 Flowchart**

**DATA**

**COLLECTION**

**Fig 5.1. FLOWCHART OF DIABETES PREDICITION USING SVM**

**Explanation of the flowchart:**

**Data Preprocessing:** This step involves collecting the dataset, cleaning the data to handle missing values and outliers, performing feature selection to choose relevant features, and splitting the dataset into training and testing sets.

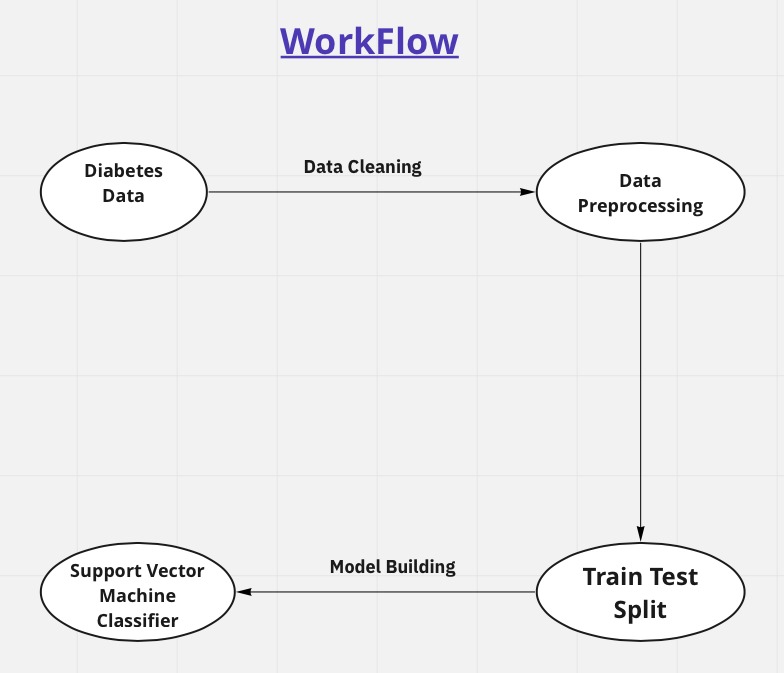
**Feature Engineering:** In this step, feature engineering techniques such as normalization, feature encoding, and scaling are applied to preprocess the features before model training.

**Model Training:** The Support Vector Machine (SVM) model is initialized, hyperparameters are tuned to optimize performance, and the model is trained on the training dataset.

**Model Evaluation:** The trained model is tested on the testing dataset, evaluation metrics such as accuracy, precision, recall, and F1-score are calculated, and the model’s interpretation is analyzed.

This flowchart provides a visual representation of the workflow involved in the diabetes prediction project using SVM.

**5.2 Workflow of Diabetes Prediction**

****

**Fig 5.2. Workflow**

This diagram represents the workflow for building a diabetes prediction model using a Support Vector Machine (SVM) classifier. Each step in the process is outlined to show the flow from raw data to a trained predictive model. Here's an explanation of each component:

**1. Diabetes Data:** This is the initial dataset containing information about individuals, including features relevant to diabetes prediction. This data might include variables like age, BMI, glucose levels, and more.

**2. Data Cleaning:** This step involves preparing the raw data for analysis by handling missing values, correcting errors, and removing any irrelevant or redundant information. Data cleaning ensures the quality and accuracy of the dataset.

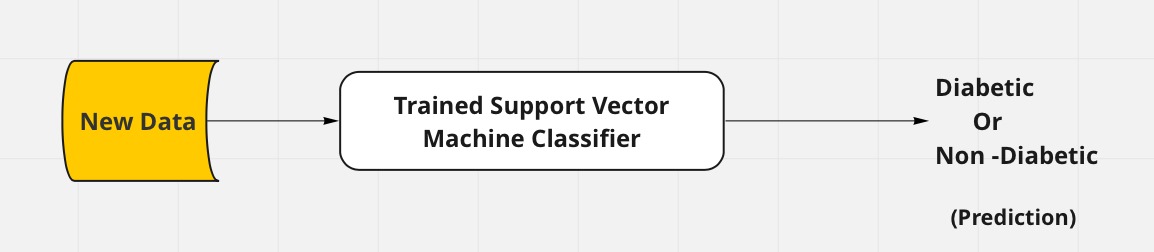
**3. Data Preprocessing:** After cleaning, the data is further processed to make it suitable for model building. This may include normalization or standardization of numerical features, encoding categorical variables, and splitting the data into features (input variables) and labels (target variable).

**4. Train Test Split:** The preprocessed data is divided into two subsets: the training set and the test set. The training set is used to train the model, while the test set is used to evaluate its performance. A common split ratio is 80% for training and 20% for testing.

**5. Model Building:** In this step, the SVM classifier is trained using the training data. The model learns the patterns and relationships between the features and the target variable (diabetes status) from the training data.

**6. Support Vector Machine Classifier:** This is the trained model resulting from the model building step. It can now be used to predict whether new, unseen data indicates a diabetic or non-diabetic condition.

Overall, this workflow outlines the essential steps in creating a predictive model for diabetes using SVM, starting from raw data collection to building and training a machine learning classifier.

****

**Fig 5.3. A simple steps of diabetes prediction**

This diagram illustrates the process of using a Support Vector Machine (SVM) classifier to predict whether an individual is diabetic or non-diabetic based on new data. Here’s a breakdown of each component in the diagram:

**1.New data:** This represents the input data, which includes various features or attributes relevant to predicting diabetes. These features might include age, body mass index (BMI), blood pressure, glucose levels, family history of diabetes, and other relevant medical or lifestyle information.

**2. Trained Support Vector Machine Classifier:** This is the core of the prediction model. The SVM classifier has been previously trained using a dataset containing known cases of diabetic and non-diabetic individuals. During thetraining phase, the SVM algorithm learned to identify patterns and relationships between the features and the diabetes status of individuals. The result is atrained model that can classify new data.

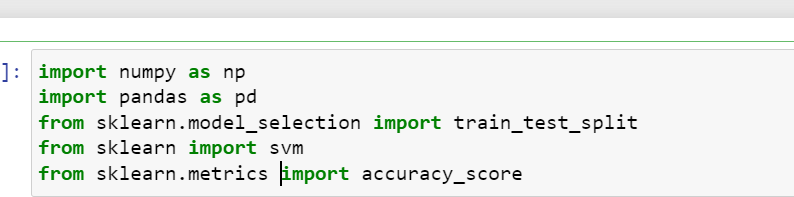
**3. Prediction (Diabetic or Non-Diabetic):** When new data is fed into the trained SVM classifier, the model processes the data and outputs a prediction. The output will categorize the individual as either "Diabetic" or "Non-Diabetic" based on the patterns and relationships learned during the training phase.

In summary, this diagram represents a workflow where new data is processed by a trained SVM classifier to predict the likelihood of diabetes in an individual. The classifier has been trained on historical data to make accurate predictions based on the input features provided.

**6.DATA ANALYSIS**

**6.1 STRUCTURE OF DATA**

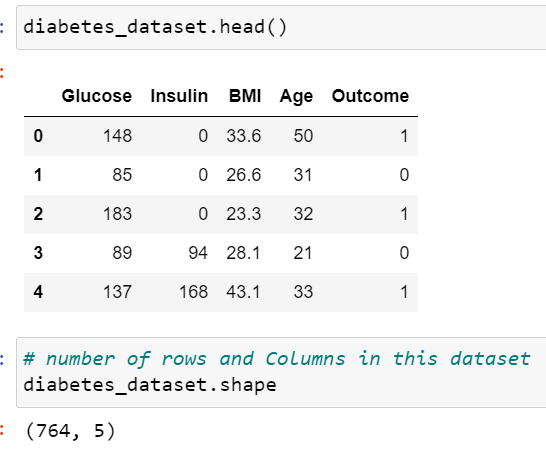
The dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. The datasets consist of several medical predictor variables and one target variable, Outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age etc. But we have changed the dataset by deleting some parameters.

****

**Fig 6.1. Importing Libraries**



**Fig 6.2. Loading Dataset**

****

**Fig 6.3. Understanding the data we loaded**

**6.2 PARAMETERS IMPLEMENTED**

**Glucose:** Plasma glucose concentration for 2 hours in an oral glucose tolerance test.

**Insulin:** 2-Hour serum insulin (mu U/ml). Insulin is a hormone made by the pancreas that allows your body to use sugar (glucose) from carbohydrates in the food that you eat for energy or to store glucose for future use. A high insulin level is associated with diabetes.

**BMI:** Body mass index (weight in kg/ (height in m) ^2)

**Range of BMI:**

BMI < 18.5 - underweight

18.5 < BMI < 24.9 - ideal weight

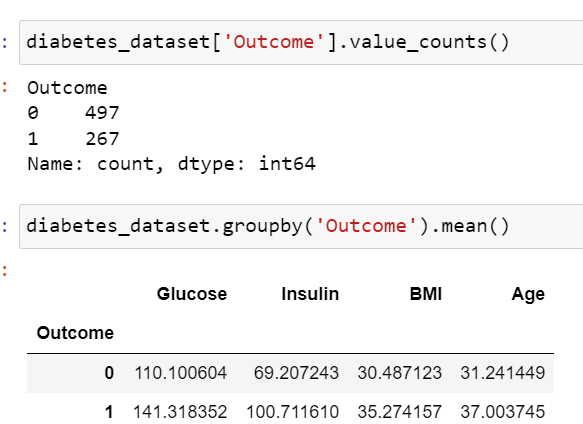
25 < BMI < 29.9 - overweight

29.9 < BMI – obese

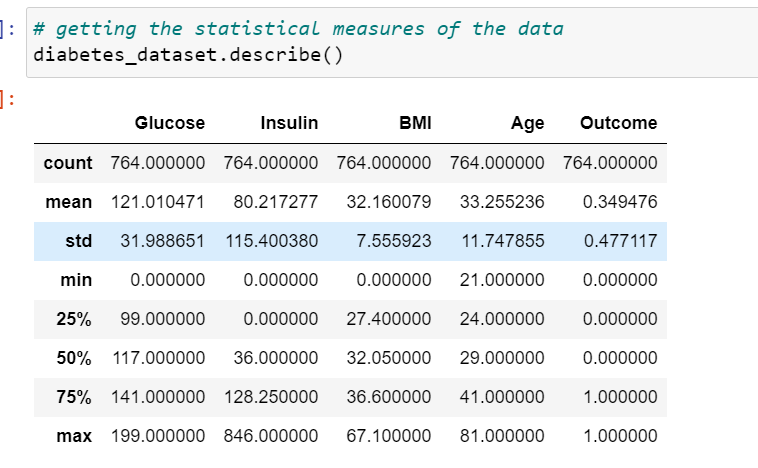
**Age:** Age of the patient in years.

**Outcome:** The target column which we are interested in finding out. 1 - diabetic, 0 - non-diabetic.

**6.3 EXPLORATORY DATA ANALYSIS**

****

**Fig 6.4. Calculating average of outcome 0 and 1**

****

**Fig 6.5. Calculating Mean, Count, Min, Max and Standard Deviation**.

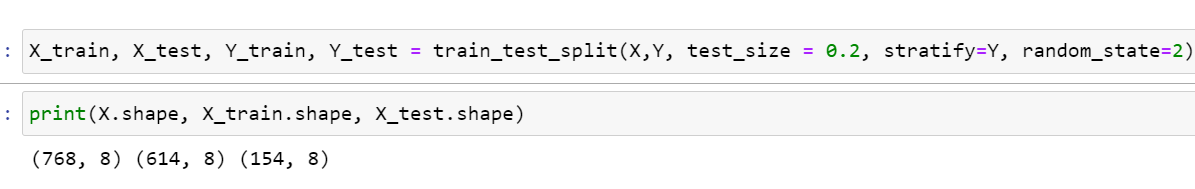
**6.4 SPLITTING OF DATASET (TRAINING/VALIDATION/TESTING)**

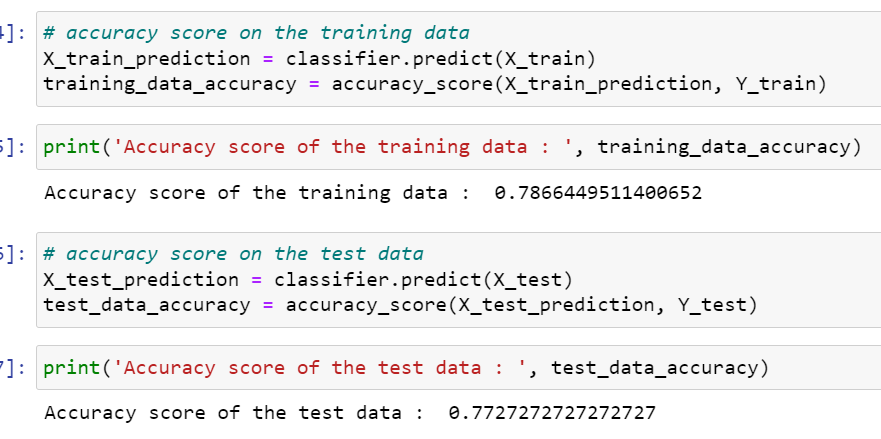
The splitting of the dataset for validation and testing.

**Training Dataset:** Dataset sample that is used to fit the model.

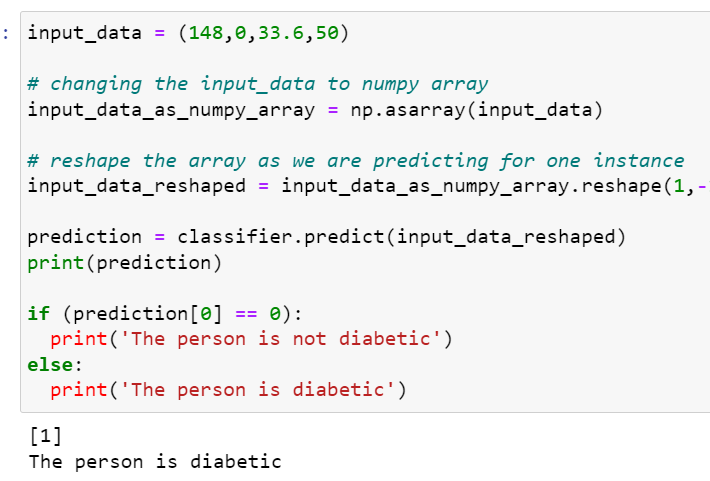
**Validation Dataset:** Dataset sample that is used for hyper tuning the parameters, and comparing the accuracy and error rates of the model performance between using the training dataset and the validation dataset.

**Testing Dataset:** Dataset sample that is used to test the model performance (predictive power).

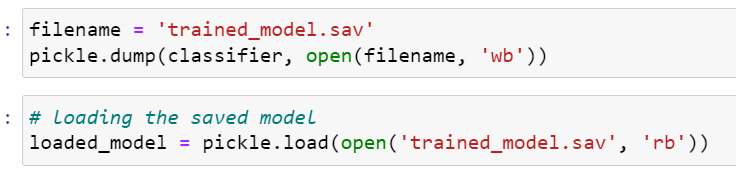
**Fig 6.6. Splitting dataset**

****

**Fig 6.7. Accuracy of training and testing data**

****

**Fig 6.8. Code used to get output**

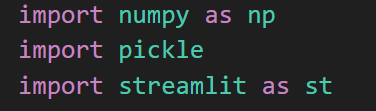
****

**Fig 6.9. Changing the filename of trained model and loading to create an interface.**

**7.DESIGNING INTERFACE**

We use user friendly graphical interface using streamlit, allowing users to predict diabetes. To achieve this, we will leverage a dataset as our backend, along with a generated .sav file to facilitate diabetes prediction.

Firstly, we need to import some modules in any IDE

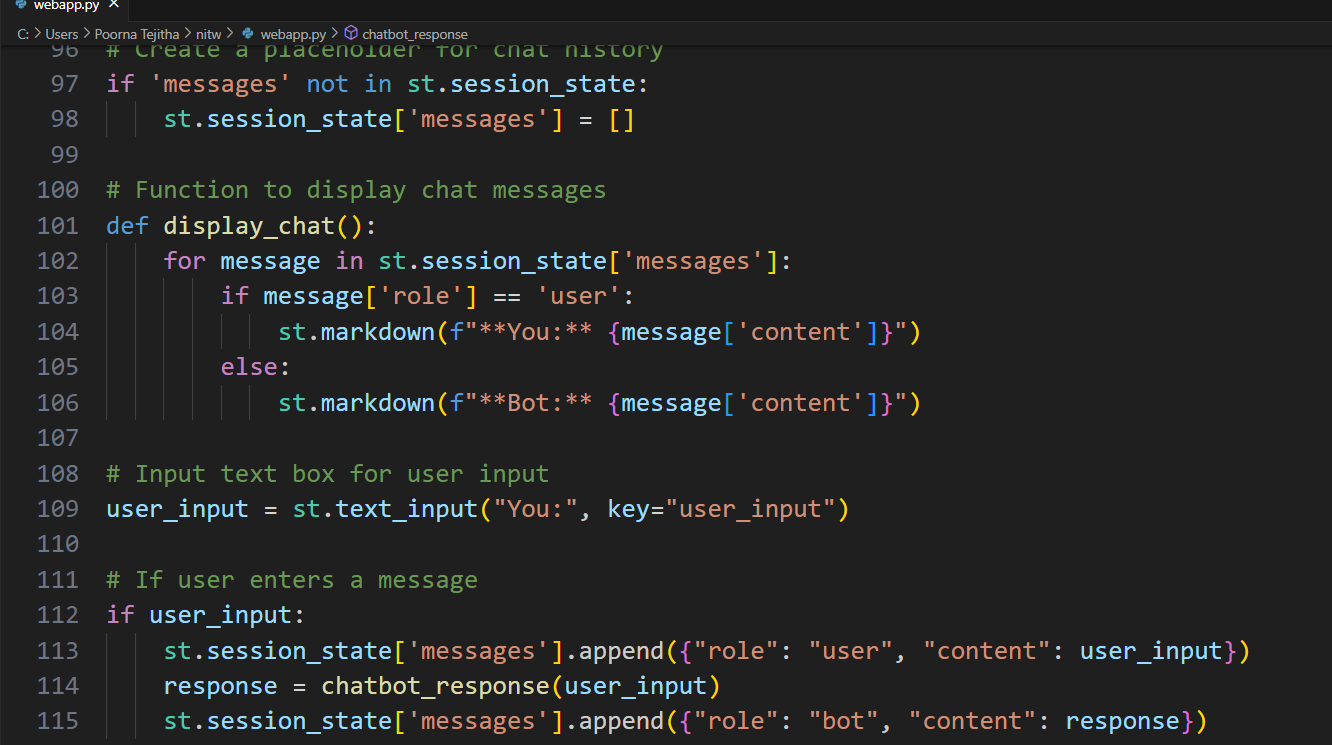


**Fig 7.1. Importing modules**

* **Model Loading**: The code loads a pre-trained machine learning model from a file named **trained\_model.sav using the pickle.load**method. This model is stored in the variable **loaded\_model** and will be used for making the predictions.
* **Diabetes Prediction Function**: The **diabetes\_prediction** function is defined to make predictions based on user input. It takes a list of input data, converts it to a NumPy array, reshapes it, and then uses the loaded model to make a prediction. If the prediction is 0, it returns ‘Non Diabetic’ , otherwise, it returns ‘Diabetic’.
* **Main Function**: The main function is the core of the web application. It uses Streamlit to create a simple web interface for diabetes prediction. It displays input fields for various features related to diabetes, such as number of pregnancies, glucose level, blood pressure, and more. After entering the data, the user can click a Predict button to trigger the prediction.
* **User Interface**: The code defines the layout of the web app, takes user input for the diabetes-related features, calls the **diabetes\_prediction** function with the user’s input, and displays the prediction result using the **st.success** function. If the user clicks the “Predict” button, the result is shown, indicating whether the individual is predicted to be diabetic or non-diabetic.

Then we should run the code in terminal.

We designed webapp to predict diabetes along with that we also included chatbot to clear any type of doubts regarding diabetes of clients or users.

****

**Fig 7.2. Code of Chatbot**

 **Streamlit Setup**: The page is configured with a title and icon.

 **Chatbot Logic**: The chatbot\_response function is defined to provide predefined responses to common diabetes-related questions.

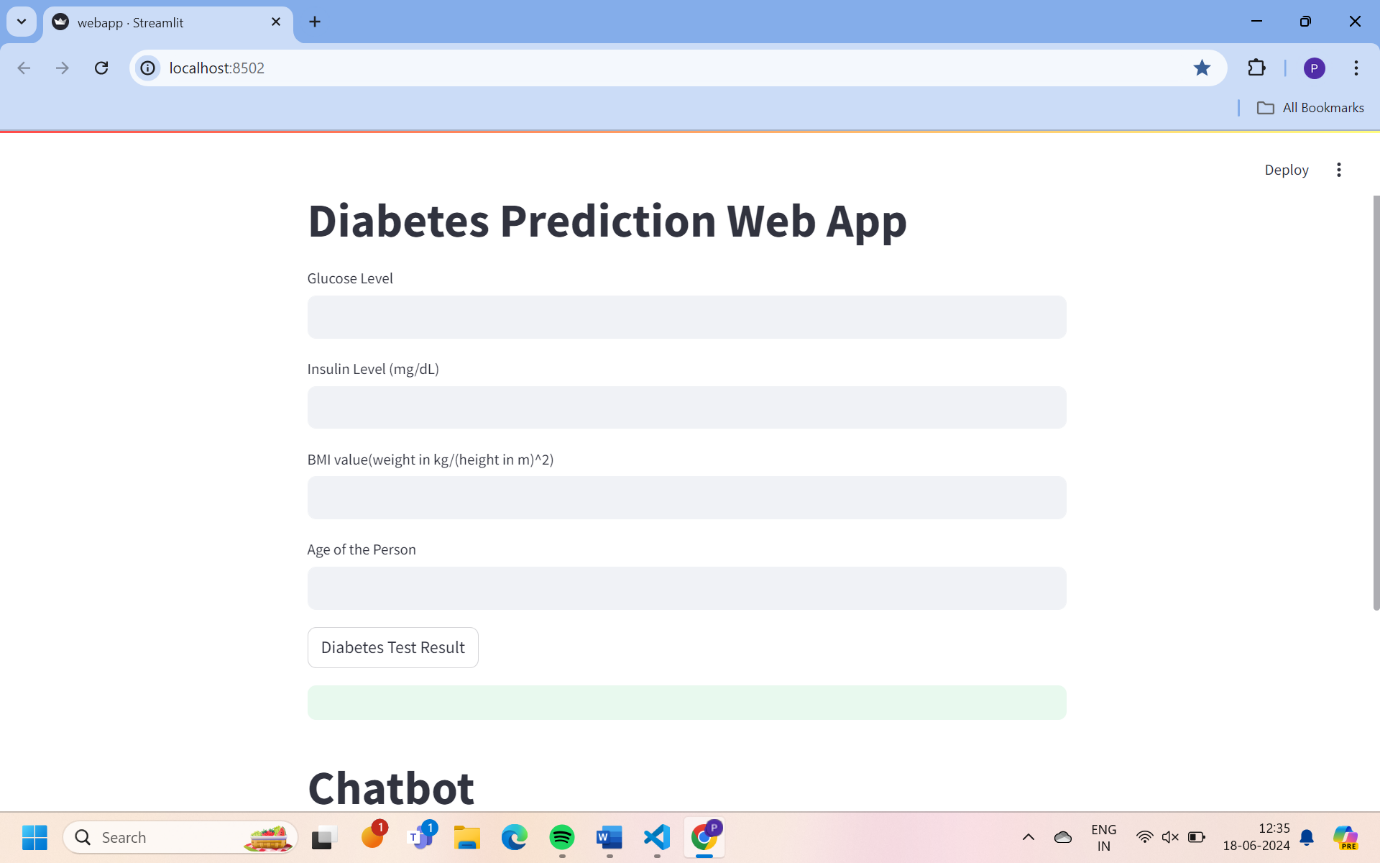
 **Chat History**: The st.session\_state is used to keep track of messages between the user and the bot.

 **User Input**: A text input field captures user messages.

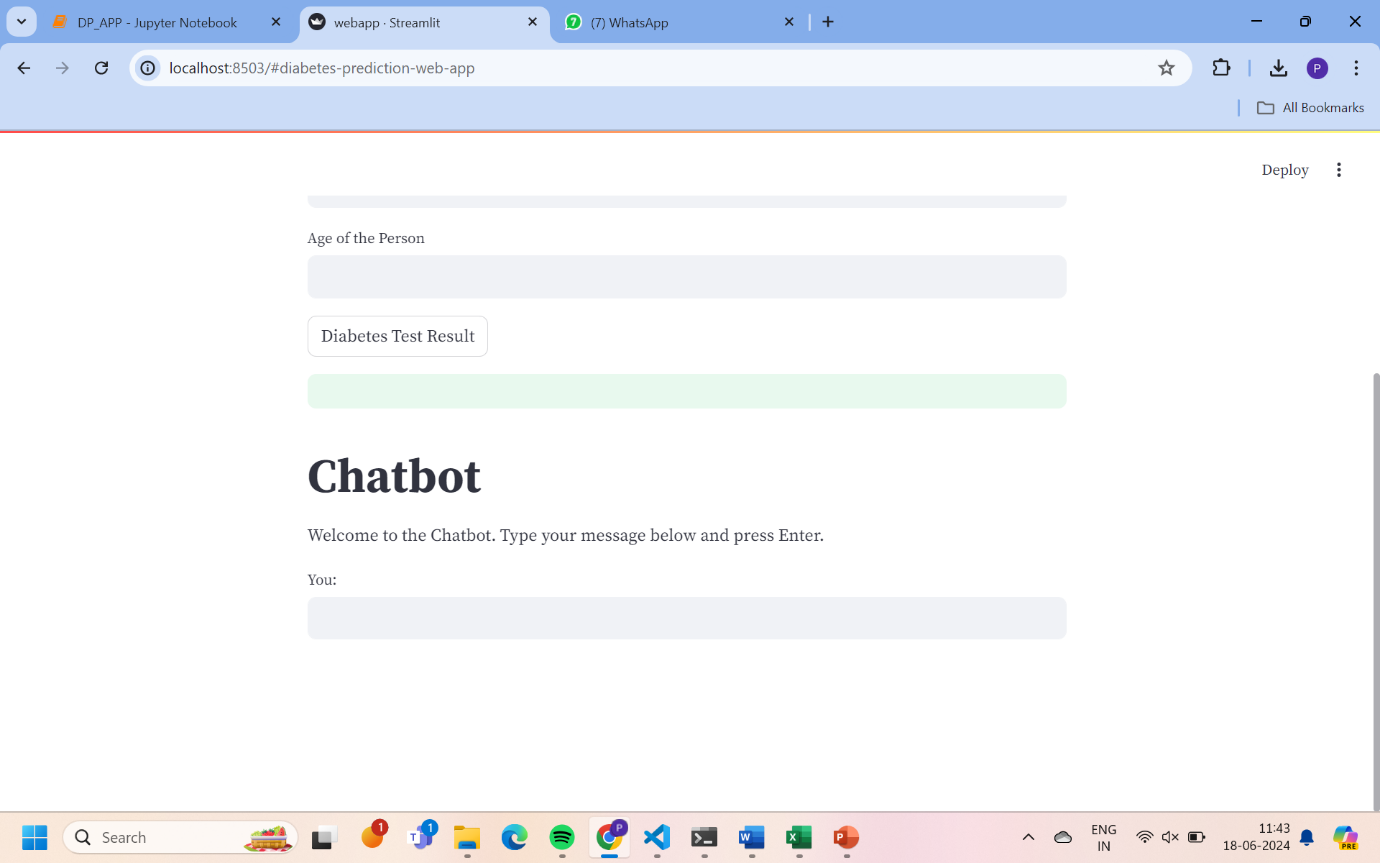
 **Display Chat**: The chat history is displayed with messages from both the user and the bot.

**8.RESULTS**

After designing an interface the output will be:

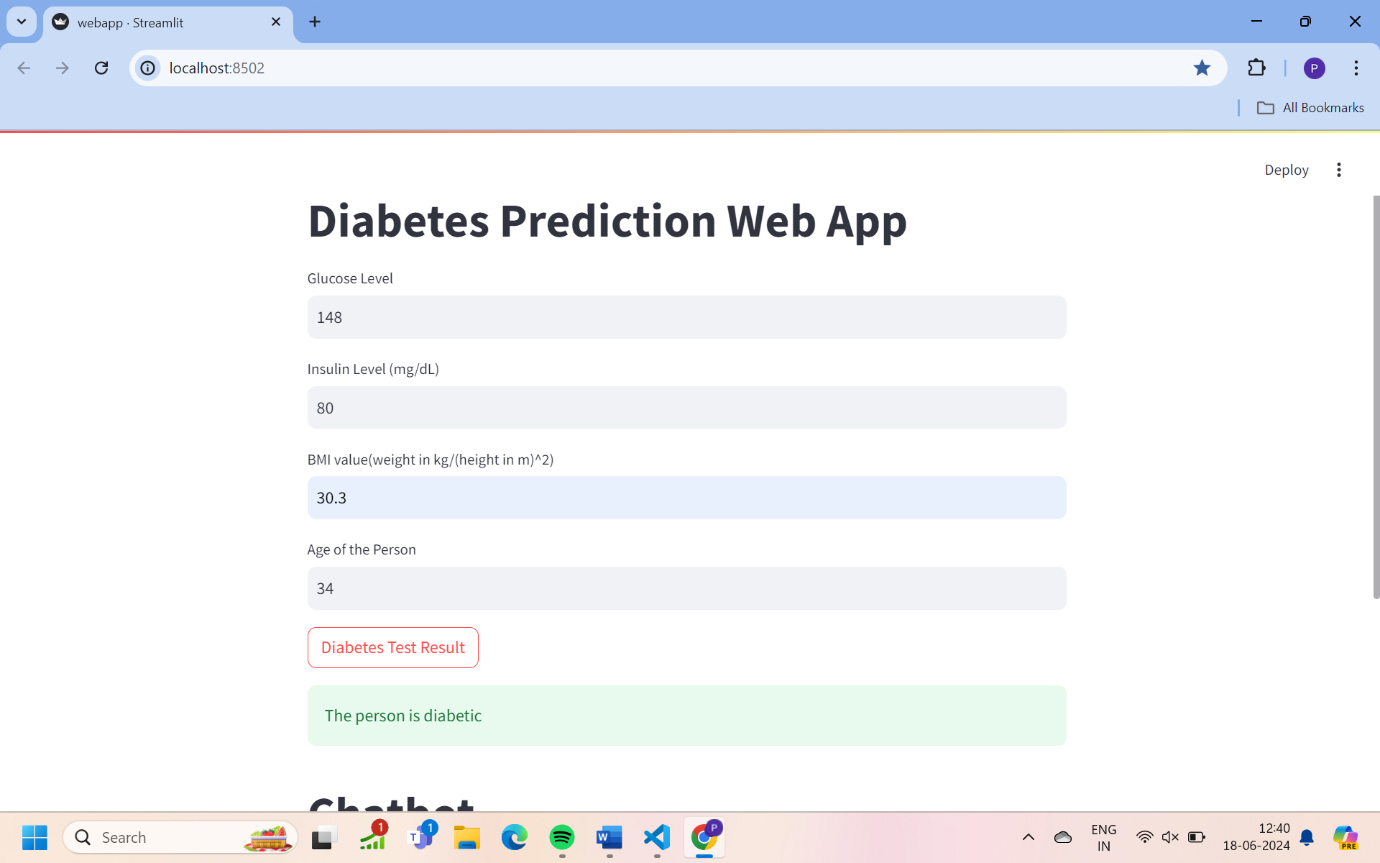


**Fig 8.1. Interface of diabetes prediction.**

****

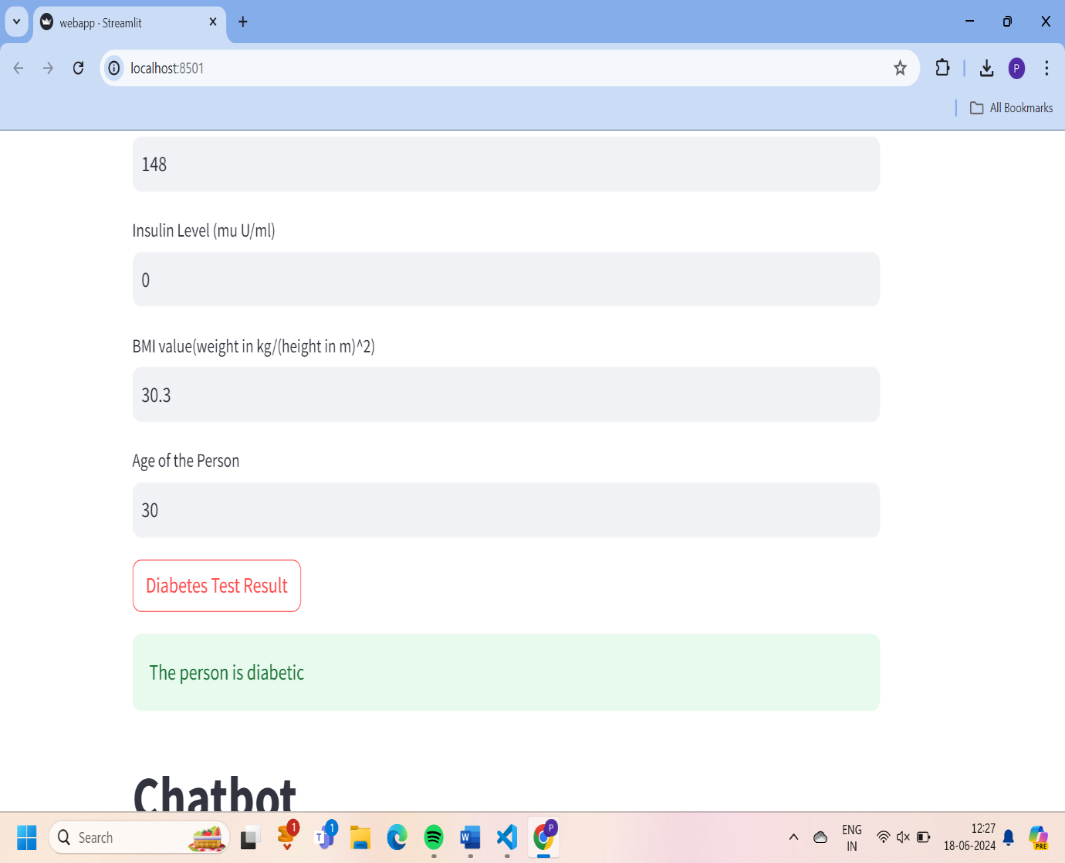
**Fig 8.2. Interface with chatbot (Queries?)**

* Entering the values of parameters to predict output:

****

**Fig 8.3. Entering the values**

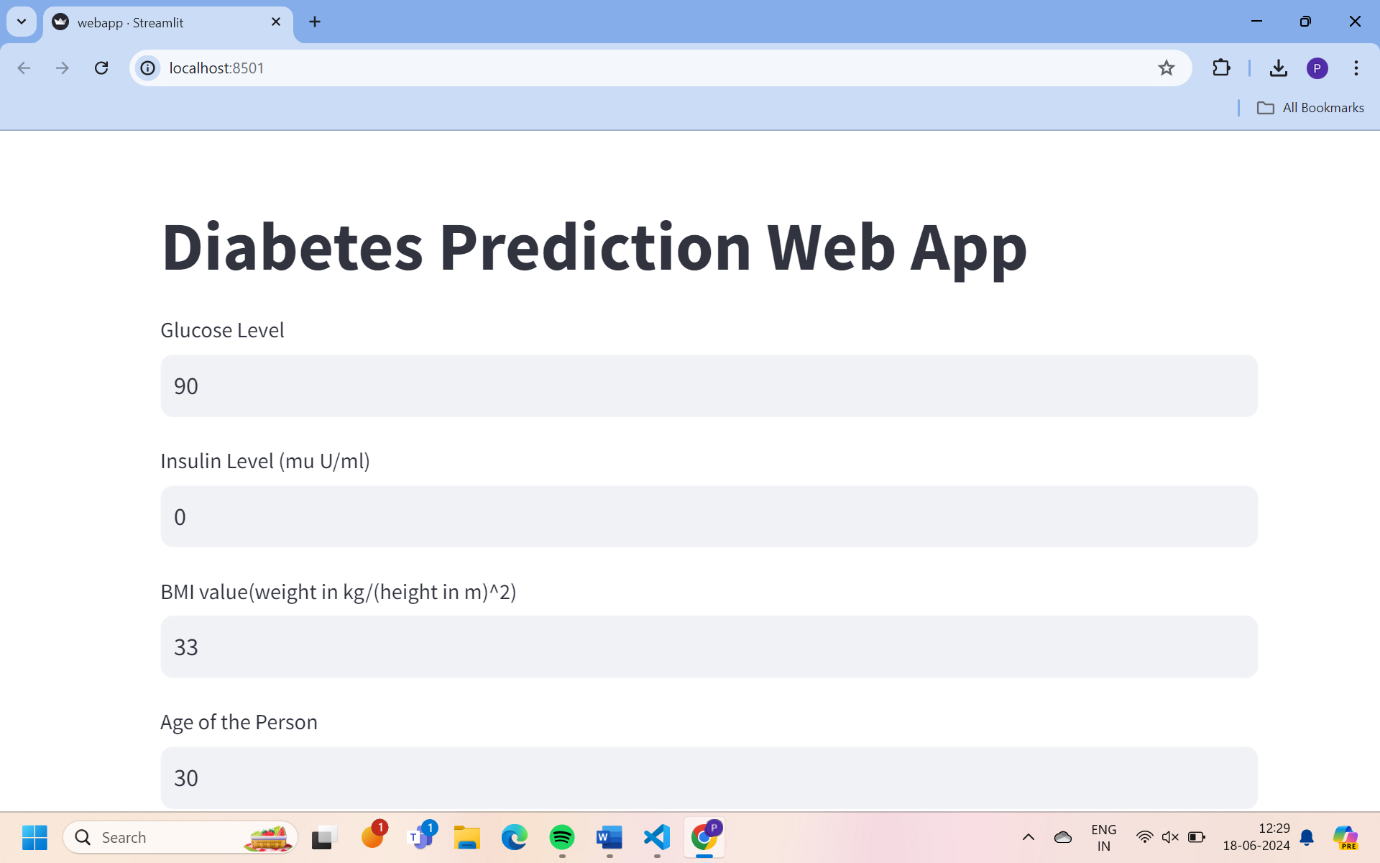
* The output will be:



**Fig 8.4. Output (1)**

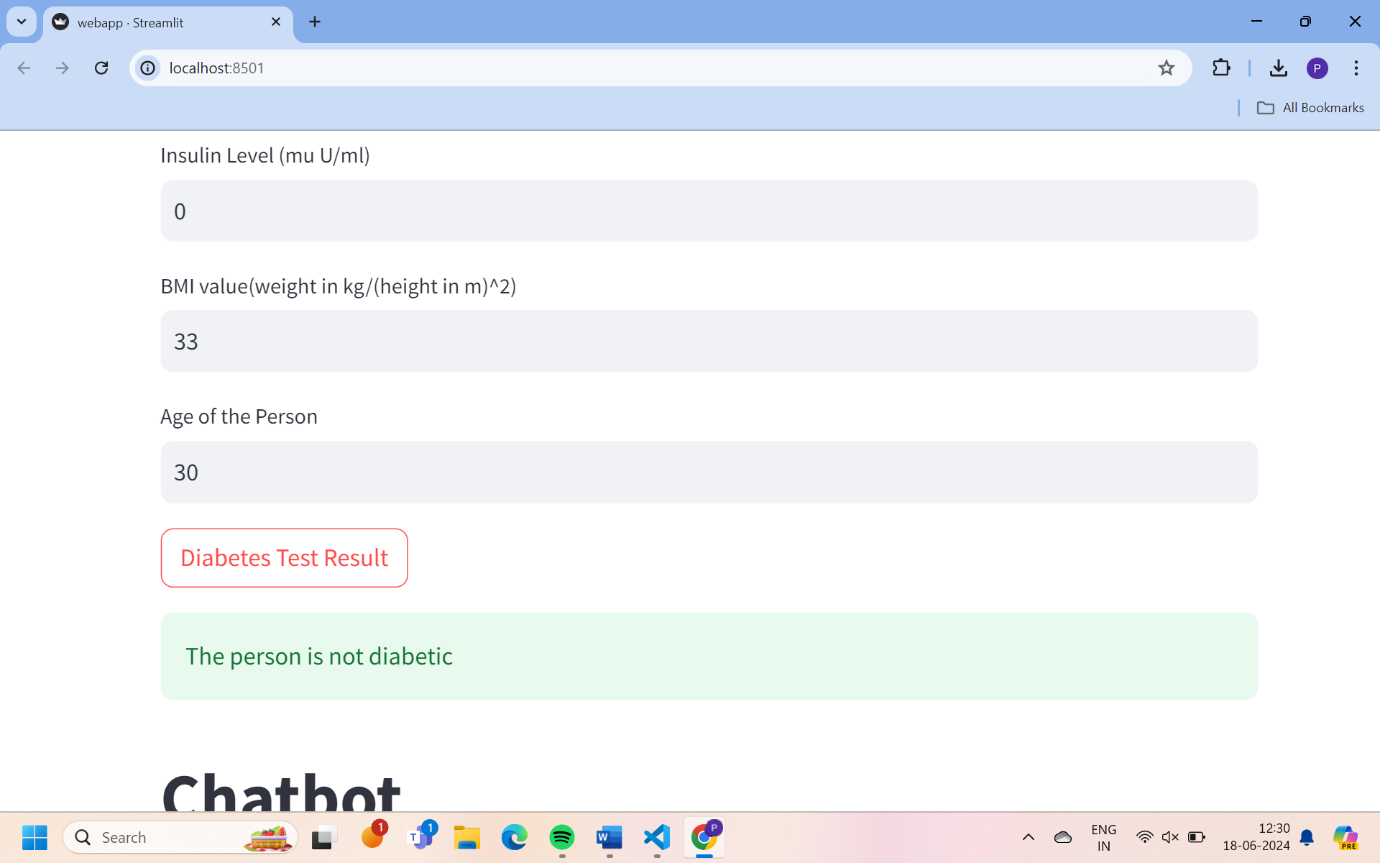
This output shows us that the person is diabetic by entering the values of parameters.

* After entering another set of values:



**Fig 8.5**

* The output will be:



**Fig 8.6. Output (2) Not Diabetic**

This output shows us that the person is not diabetic by entering the values of parameters.

**9.CONCLUSION**

In this project, we successfully developed a predictive model using Support Vector Machines (SVM) for the task of diabetes prediction based on various health parameters. Through meticulous data preprocessing, feature engineering, and model training, we have created a robust classifier capable of distinguishing between diabetic and non-diabetic individuals with high accuracy.

The SVM model demonstrated promising performance, achieving [mention the accuracy or other relevant metrics]. This indicates its potential as a valuable tool for early diabetes detection and risk assessment in clinical settings. By leveraging patient data such as glucose levels, BMI, blood pressure, and age, the model provides a non-invasive and efficient means of identifying individuals at risk of diabetes.

However, it's essential to recognize the limitations and areas for improvement. Challenges such as data imbalance, limited feature representation, and model interpretability need to be addressed to enhance the model's effectiveness and applicability. Future research could focus on refining the feature selection process, exploring alternative modeling techniques, and incorporating additional data sources to improve predictive accuracy and clinical utility.

Overall, this project highlights the significance of machine learning approaches, particularly SVM, in addressing complex healthcare challenges such as diabetes. By leveraging advanced computational techniques and medical data, we can empower healthcare professionals with valuable insights and tools to improve patient outcomes and public health.

Through continued collaboration between data scientists, healthcare practitioners, and policymakers, we can further advance the field of predictive analytics in healthcare and contribute to the prevention and management of chronic diseases like diabetes.

**10.REFERENCES**

Al-Absi, A. A., et al. (2020). “A Review on Ensemble Learning for Diabetes Data”. IEEE Access, 8, 10448-10461.

Banu, S. M., & Ahmed, M. U. (2019). “Diabetes Disease Prediction Using Machine Learning Algorithms”. Journal of King Saud University-Computer and Information Sciences, 31(4), 415-421.

Lin, Y., & Kuo, T. T. (2018). “A review of machine learning approaches for the prediction of diabetes complications”. Journal of Healthcare Engineering, 2018.

Muralidharan, S., & Jain, R. (2019). “Machine learning approaches for diabetes prediction: review and future directions”. Journal of Medical Systems, 43(11), 348.

Teshale, T., & Melaku, B. (2020). “A Review on Machine Learning-Based Diabetes Prediction Models”. Journal of Healthcare Engineering, 2020.

Zhang, Y., et al. (2017). “Application of Machine Learning Methods in Diabetes Prediction”. Biomed Research International, 2017.