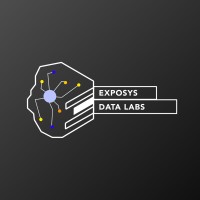
**EXPOSYS DATA LABS**



Internship report on

**DIABETES PREDICTION SYSTEM USING DATA SCIENCE**

A Dissertation work submitted in partial fulfilment of the requirement for the award of the degree of

Internshi

A Dissertation work submitted in partial fulfillment of the requirement for the award of the degree of

**Internship**

Name: Shivangi Singh Mentor Name:

College Name: GN Group of Institute Mr. Vishnu Vardhan

Duration of Internship:

1 Month [11-02-24] – [15-03-24]

**ABSTRACT**

With increasing population no. of cases for diabetes are also increasing. Diabetes mellitus, a chronic metabolic disorder, poses a significant health challenge worldwide. Early detection and timely intervention are crucial for managing diabetes and preventing associated complications.

To overcome this problem the techniques of machine learning can easily be utilized in order to extract useful pieces of information of the specified data in healthcare communities, biomedical fields etc.

This project demonstrates the diabetes prediction system developed using Machine Learning algorithms such as Random forest, Naïve based classifier, KNearest neighbors. The aim of developing classifier system using machine learning algo is to immediately help to solve health related issues by predicting the diabetes at early stage.

**INDEX**

| **S. No.** | **CONTENTS** | | | **PAGE NO.** |
| --- | --- | --- | --- | --- |
| **1** | Abstract | | | **3** |
| **2** | List of Figures | | | **7** |
| **3** | Chapter - 1 | | |  |
|  | Introduction | | | **5** |
|  | **1.1** | | Data Science | **5** |
|  | **1.2** | | Machine Learning | **6** |
| **4** | Chapter – 2 | | |  |
|  | Existing Methods | | | **7** |
|  | **2.1** | | Issues in existing methods | **7** |
| **5** | Chapter – 3 | | |  |
|  | Proposed Method | | | **8** |
|  | **3.1** | | KNearest Neighbors | **8** |
|  | **3.2** | | Algorithm | **9** |
|  | **3.3** | | Important features of KNN | **10** |
| **6** | Chapter – 4 | | |  |
|  | Methodology | | | **11** |
|  | **4.1** | | Prediction model | **12** |
|  | **4.2** | | Data Preprocessing | **12** |
|  | **4.3** | | Training and Testing | **12** |
|  | **4.4** | | Data model |  |
| **7** | Chapter – 5 | | |  |
|  | Implementation | | | **13** |
|  | **5.1** | Problem Statement | | **13** |
|  | **5.2** | Requirements | | **13** |
|  | **5.3** | Implementation steps | | **14** |
|  | **5.4** | Snapshots | | **17** |
| **8** | Conclusion | | | **19** |
| **9** | References | | | **20** |

**LIST OF FIGURES**

| **S. No.** | **FIGURE No.** | **HEADING** | **PAGE No.** |
| --- | --- | --- | --- |
| 1 | Fig. – 3.1 | KNN | 9 |
| 2 | Fig. – 3.2 | Algorithm | 10 |
| 3 | Fig. – 4.1 | Prediction model | 11 |
| 4 | Fig. – 5.1 | Importing Libraries | 14 |
| 5 | Fig. – 5.2 | Loading data | 15 |
| 6 | Fig. – 5.3 | Shape of dataset | 15 |
| 7 | Fig. – 5.4 | Checking missing values | 16 |
| 8 | Fig. – 5.5 | Train Test split | 16 |
| 9 | Fig. – 5.6 | Home Page | 17 |
| 10 | Fig. – 5.7 | Login Page | 17 |
| 11 | Fig. – 5.8 | Prediction Result | 18 |

**CHAPTER 1**

**INTRODUCTION**

**1.1 Data Science**

Data science is a multidisciplinary field that combines various techniques and methods to extract insights and knowledge from structured and unstructured data. It encompasses a wide range of tools and approaches, including statistics, machine learning, data mining, and data visualization, to analyze complex datasets and make data-driven decisions.

**Key components of data science include:**

**Data Collection:** Gathering data from various sources, including databases, sensors, social media, and the internet.

**Data Cleaning and Preprocessing:** Removing inconsistencies, errors, and irrelevant information from the data to ensure accuracy and reliability.

**Exploratory Data Analysis (EDA):** Exploring and visualizing the data to understand its structure, identify patterns, and gain insights into relationships between variables.

**Statistical Analysis:** Applying statistical techniques to analyze data and test hypotheses, such as regression analysis, hypothesis testing, and analysis of variance (ANOVA).

**Machine Learning:** Building predictive models and algorithms that can learn from data and make predictions or decisions without being explicitly programmed. This includes supervised learning, unsupervised learning, and reinforcement learning techniques.

**Data Visualization:** Creating visual representations of data, such as charts, graphs, and dashboards, to communicate insights effectively and facilitate understanding.

**Model Evaluation and Validation:** Assessing the performance of predictive models and validating their accuracy and reliability using techniques such as cross-validation and model evaluation metrics.

**1.2 Machine Learning**

Machine learning is a subset of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data without being explicitly programmed. The primary goal of machine learning is to enable computers to automatically learn and improve from experience, without human intervention, by identifying patterns, trends, and relationships within data.

**Key concepts and techniques in machine learning include**:

**Supervised Learning:** In supervised learning, the algorithm is trained on labeled data, where each input data point is associated with a corresponding target label or output. The algorithm learns to map input features to the correct output by minimizing the difference between predicted and actual values. Common supervised learning algorithms include linear regression, logistic regression, decision trees, support vector machines, and neural networks.

**Unsupervised Learning:** In unsupervised learning, the algorithm is trained on unlabeled data, and its goal is to find patterns, clusters, or hidden structures within the data. Unsupervised learning techniques include clustering algorithms such as K-means clustering and hierarchical clustering, as well as dimensionality reduction techniques like principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE).

**Reinforcement Learning:** Reinforcement learning involves training agents to make sequential decisions in an environment to maximize cumulative rewards. The agent learns through trial and error by interacting with the environment and receiving feedback in the form of rewards or penalties. Reinforcement learning algorithms include Q-learning, deep Q-networks (DQN), and policy gradient methods.

**Feature Engineering:** Feature engineering involves selecting, transforming, and creating new features from raw data to improve the performance of machine learning models. It plays a crucial role in preprocessing and preparing data for modeling tasks.

**Model Evaluation and Validation:** Evaluating the performance of machine learning models is essential to ensure their effectiveness and reliability. Common evaluation metrics include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (AUC-ROC).

**CHAPTER 2**

**EXISTING METHOD**

Existing many research handled for diabetes detection. Data mining approach like clustering, classifications were studied in existing system.

Diabetes prediction using algorithms such as bound algo and linear regression was proposed.

**2.1. Issues in Diabetes Prediction Systems:**

**Data quality and analysis:** One of the primary challenges in diabetes prediction is the quality and availability of data. Missing values, inaccuracies, and inconsistencies in data can hinder the development of accurate predictive models.

**Imbalanced datasets:** Imbalanced datasets where the number of individuals with diabetes is significantly lower than those without diabetes can pose challenges for model training and evaluation, leading to biased predictions.

**Feature selection interpretability:** Selecting relevant features and interpreting the output of complex machine learning models can be challenging, particularly in healthcare settings where interpretability and explainability are critical for clinical decision-making.

**Chapter 3**

**PROPOSED METHOD**

**3.1 KNearest Neighbors**  
The k-Nearest Neighbors (k-NN) algorithm is a simple yet effective supervised machine learning technique used for classification and regression tasks. In classification tasks, it is particularly useful for identifying the class of a new data point based on the majority class of its nearest neighbors in the feature space.

**Working of KNN Algorithm**

**1. Training Phase:** During the training phase, the algorithm simply stores all available data points along with their corresponding class labels.

**2. Prediction Phase:**

* Given a new data point, the algorithm calculates its distance to all other data points in the training set. The most commonly used distance metric is Euclidean distance, although other metrics like Manhattan or cosine distance can also be used.
* It then selects the k nearest data points (neighbors) to the new data point based on the calculated distances.
* For classification, the algorithm assigns the class label that is most frequent among the k nearest neighbors to the new data point. In regression tasks, it computes the average (or weighted average) of the target values of the k nearest neighbors and assigns it as the predicted value for the new data point.

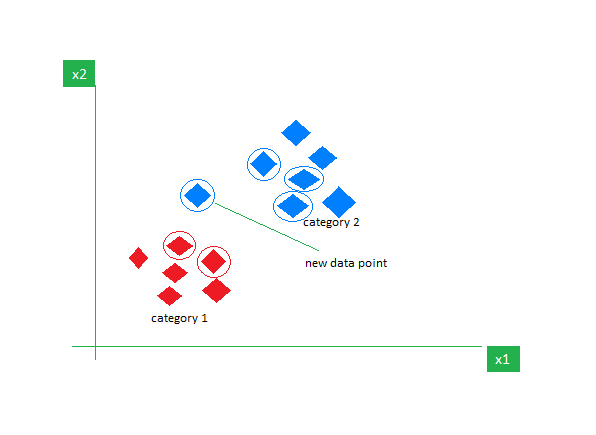


Fig.3.1 KNN

**3.2. Algorithm**

**Step-1:** Select the number K of the neighbors.

**Step-2:** Calculate the Euclidean distance of **K number of neighbors.**

**Step-3:** Take the K nearest neighbors as per the calculated Euclidean distance.

**Step-4:** Among these k neighbors, count the number of the data points in each category.

**Step-5:** Assign the new data points to that category for which the number of the neighbor is maximum.

**Step-6:** Our model is ready.



Fig.3.2 Algorithm

**3.3 Important Features of KNN**

1. Simple and intuitive:k-NN is conceptually simple and easy to understand. It relies on the principle that similar instances are likely to belong to the same class or have similar target values.

2. Non-parametric: k-NN is a non-parametric method, meaning it makes no assumptions about the underlying data distribution. This flexibility allows it to capture complex patterns in the data without imposing specific constraints.

3. Instance based learning: NN is an instance-based learning algorithm, which means it memorizes the entire training dataset. During prediction, it compares new instances to the training data to make decisions, rather than learning explicit models.

**CHAPTER 4**

**METHODOLOGY**

We start with importing pandas and Numpy and load a .csv dataset using pandas. To see first five rows we use diabetes\_data.head () function. We also use matplotlib and sklearn.

**4.1. Prediction Model**

Diabetes dataset

Data Preprocessing

Test Set

Training Set

Ml Model

Fig.4.1 Prediction model

**4.2 Data preprocessing**

Gather a dataset containing features such as age, BMI, blood pressure, glucose levels, family history of diabetes, physical activity level, and dietary habits.

Preprocess the data by handling missing values, scaling the features, and encoding categorical variables if necessary.

**4.3 Training and Testing**

We are using Machine learning model algorithm called KNearest Neighbors. We train the model using train datasets. The Trained model is saved as result.py.

The splitting into trained and test is occurred. The average splitting is 80% training data and 20% testing data. Our model accuracy is 80% apprx.

**4.4 Data Model**

The trained model is saved in result.pkl file which is used to predict the output of the user’s input.

We use flask to deploy our html page to ml model. Web page contains a form with some input based on that it will predict the output.

**CHAPTER 5**

**IMPLEMENTATION**

**5.1 Problem Statement**

**The goal is to predict**

1. Prepare the data-set using several methods to train the model.

2. Build a model which can give high accuracy of predicting the disease.

**5.2. Requirements**

* Software

1. VS CODE
2. Python (3.8.5)

* Website:

Diabetes dataset from kaggle (<https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>)

* Libraries

1. Pandas (3.8.5)
2. Matplotlib (3.7.5)
3. Scikit Learn (1.3.2)
4. Pickle
5. Os

**5.3. Implementation Steps**

Step By Step implementation of project:

**Step1:** Import all the required libraries such as Pandas , Matplotlib etc.

For example:

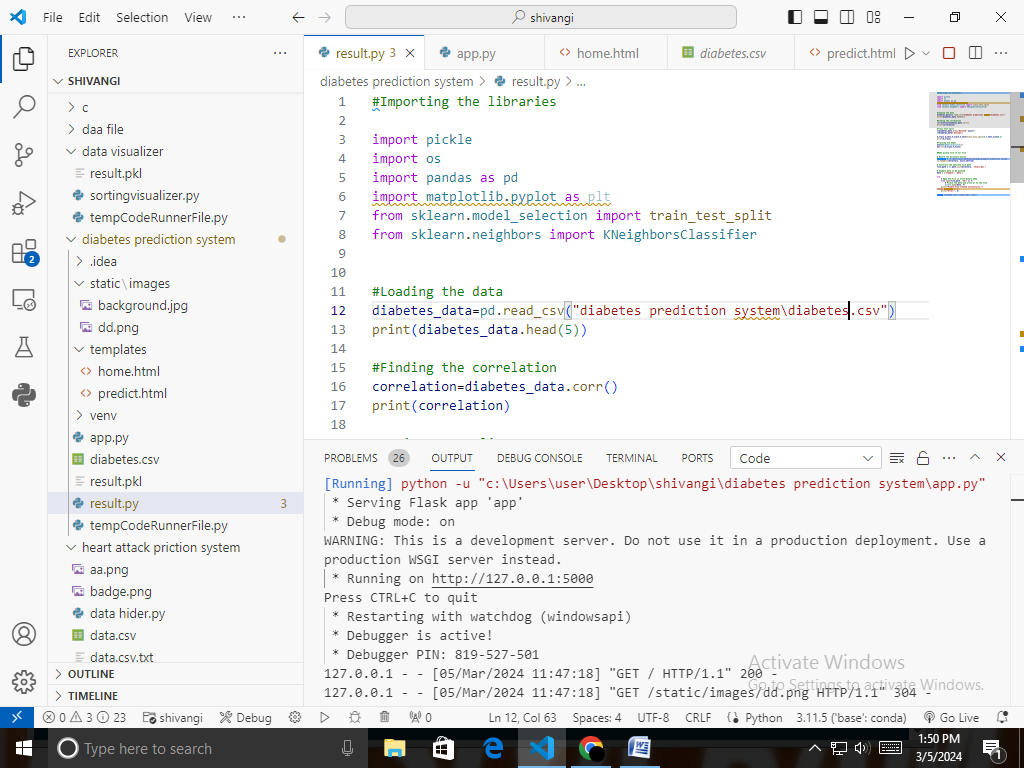


Fig.5.1 Importing Libraries

**Step2:** Read the Dataset and stored it in variable.

In the dataset there are various variable

Pregnancies: This variable represents the number of times a woman has been pregnant.

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

Blood Pressure: This variable represents the diastolic blood pressure (mm Hg) of the individual.

Skin Thickness: This variable indicates the thickness of the triceps skin fold in millimeters.

Insulin: This variable represents the 2-hour serum insulin level measured in micro units per milliliter (mu U/ml) during a glucose tolerance test.

BMI: Body Mass Index (BMI) is calculated as the individual's weight in kilograms divided by the square of their height in meters.

Diabetes Pedigree Function: This variable is a function that represents the likelihood of diabetes based on family history. It combines information about the diabetes history of relatives and the age of onset to provide a numerical value indicating genetic predisposition to diabetes.

Age: Age is the age of the individual in years.

Outcome: This variable is a binary indicator (0 or 1) representing the presence or absence of diabetes. A value of 1 typically indicates that the individual has been diagnosed with diabetes, while a value of 0 indicates no diagnosis.

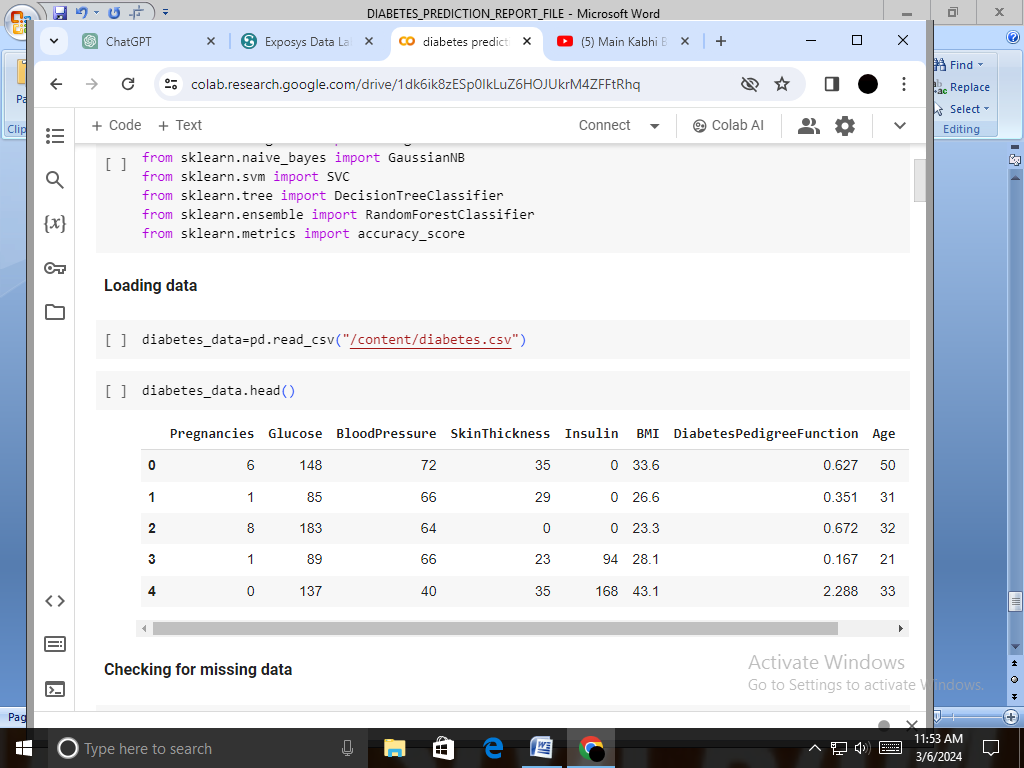


Fig5.2 Loading data

**Step3:** Next step is to check the shape of the dataset or data frame.

For example:

Shape of the dataset: (768, 9)

This means, in dataset there are 768 rows and 9 columns.

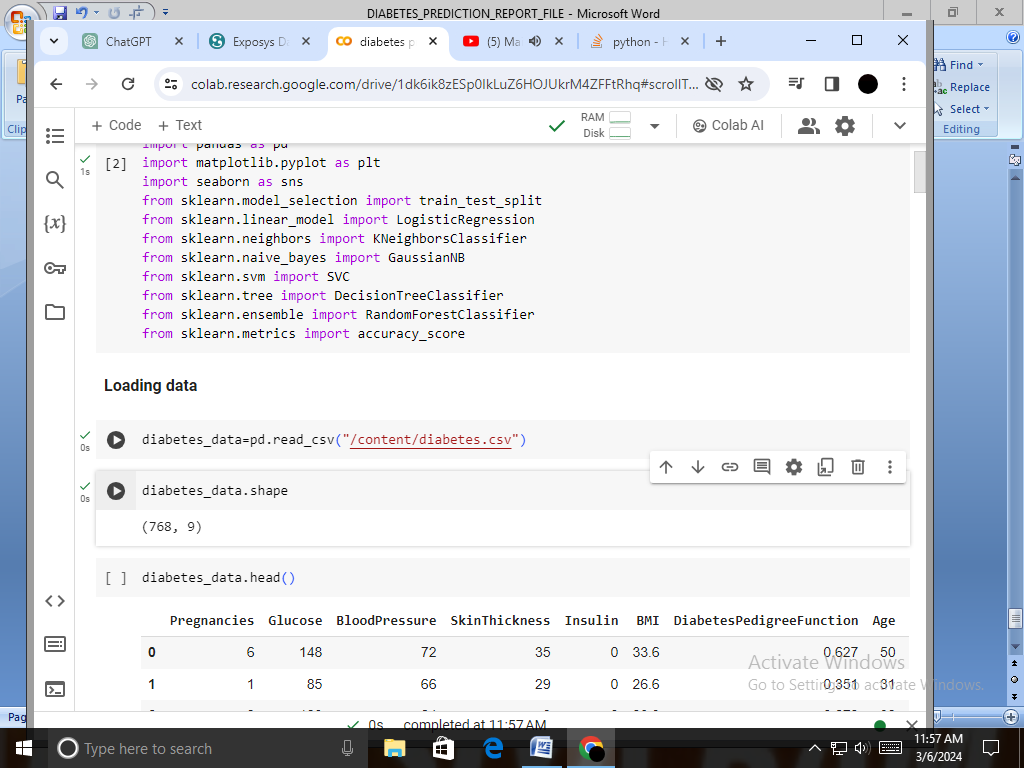


Fig5.3 Shape of dataset

**Step4:** Now next step is to check if there are missing values in our dataset or not. Here there is no missing values hence we can move further.

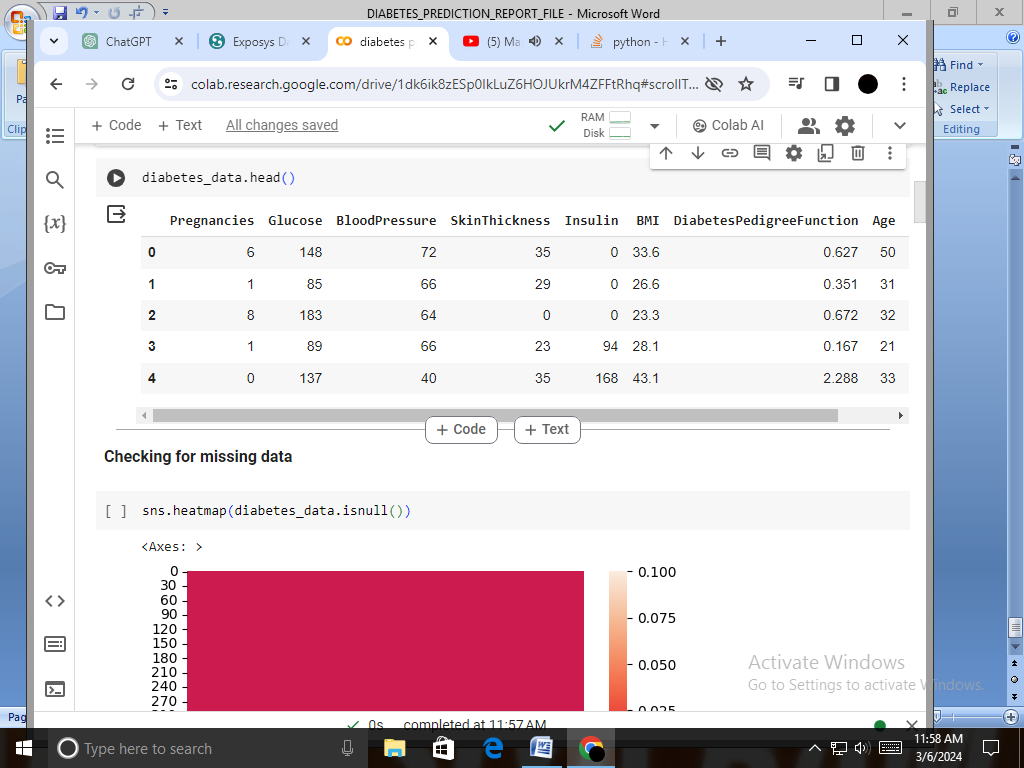


Fig5.4 checking missing value

Step5: Now the next step is to train data and do train test split.

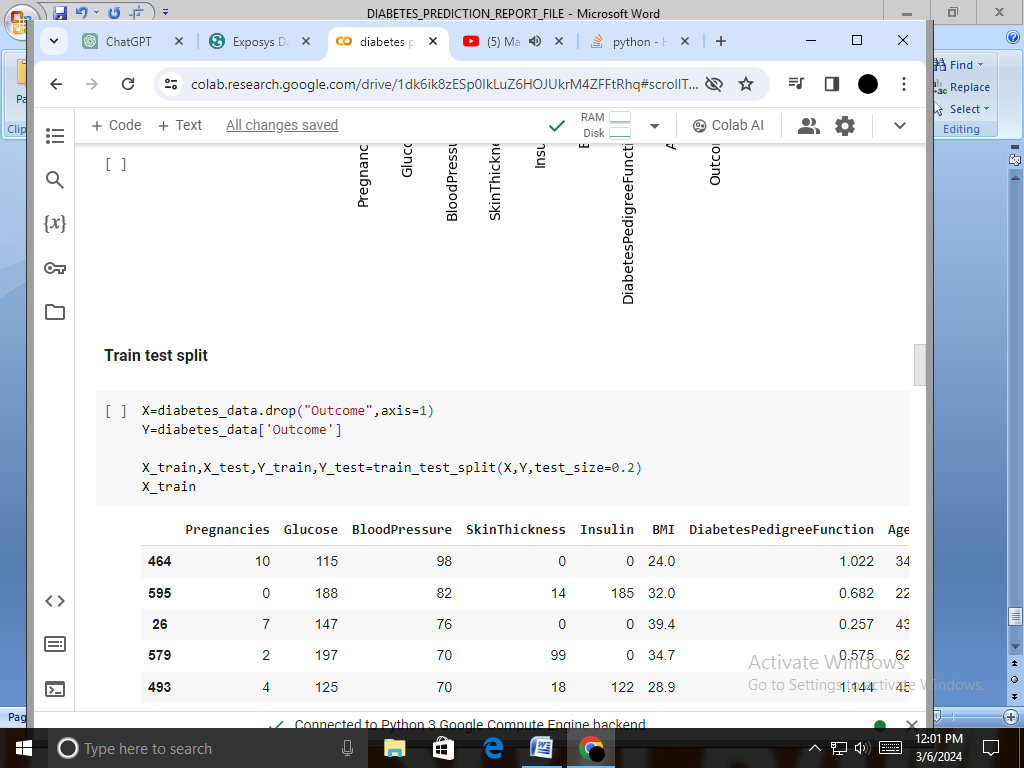


Fig5.5 Train Test split

Step6: Next step is select model. Here we are using KNN, Naïve bayes, linear regression, Decision Tree, SVM, Random forest.

Step7: After model building check for highest accuracy model.

Step8: After checking for highest accuracy model Make a model using it and check whether it is working or not.



Fig5.6 Home Page

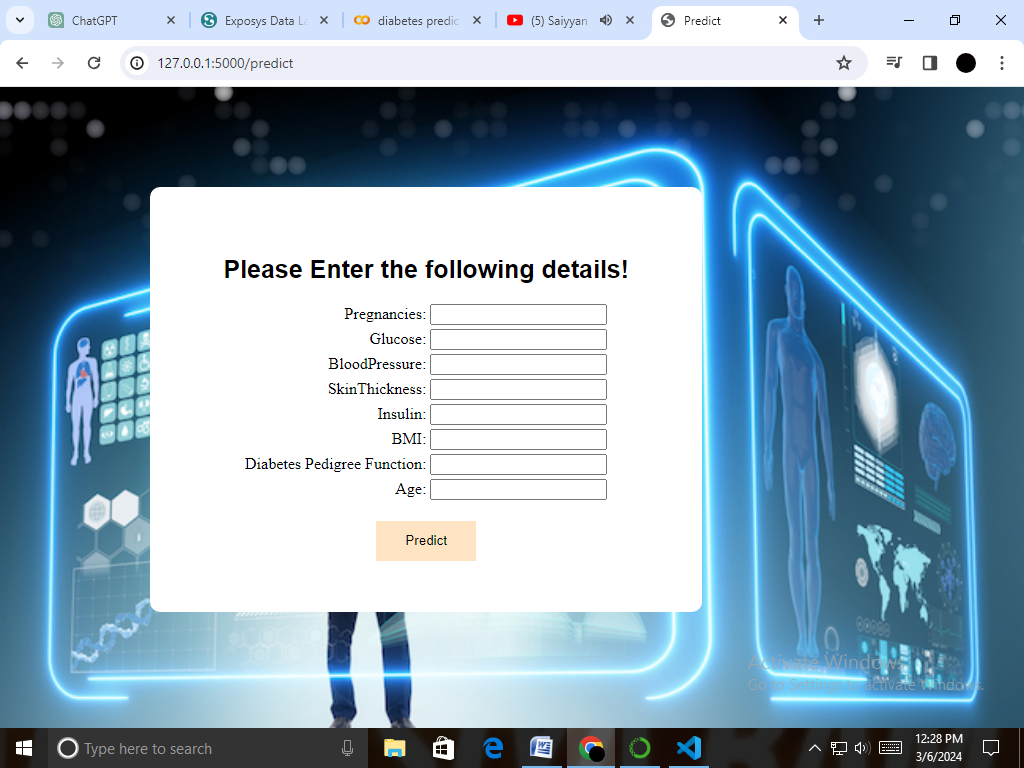


Fig5.6 Login Page

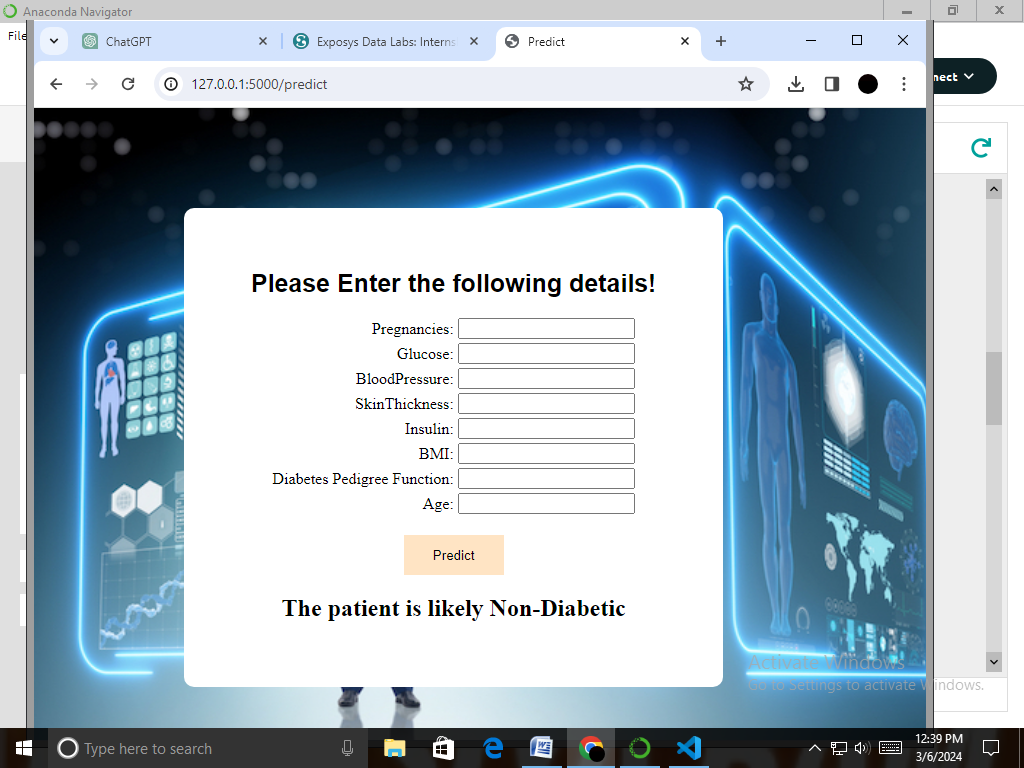


Fig5.6 Prediction Result

**Conclusion**

In this project we performed all functions from Data collecting to model deployment using Flask. In this different libraries and classification are used. On the basis of accuracy we use KNearest Neighbors (KNN) algorithm in our project. This system allows users to input their diabetes-related data through a web interface, and the trained KNN model predicts whether they are likely to have diabetes or not. The integration with Flask enables easy deployment and interaction with the model through a web browser.

In future improvements we

1. Add more sophisticated machine learning algorithms for comparison.
2. Enhance the user interface with more features or better styling.
3. Deploy the application to a production server for public access

**REFRENCES**

* <https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset>)
* <https://en.wikipedia.org/wiki/Data_science>
* <https://en.wikipedia.org/wiki/Flask_(web_framework)>
* Research on diabetes Prediction model (<https://www.niddk.nih.gov/about-niddk/research-areas/diabetes>)