**Phase-3**

**Student Name:** Sandhiya S

**Register Number:** 510623104092

**Institution:** C. Abdul Hakeem College of Engineering and Technology

**Department:** Computer Science and Engineering

**Date of Submission:** 11-05-25

**Github Repository Link:**

Problem Statement:

In the current healthcare landscape, early and accurate disease diagnosis remains a significant challenge. Patients often delay seeking medical attention due to unrecognized symptoms, and healthcare providers face time constraints during consultations, leading to potential misdiagnoses or delayed treatments. Chronic diseases such as diabetes and heart conditions are on the rise, necessitating timely interventions to prevent complications. Traditional diagnostic methods may not efficiently handle the vast and complex patient data available today. There is a pressing need for intelligent systems that can analyse patient data comprehensively to assist in early disease detection and management. This project aims to develop an AI-powered solution that leverages machine learning algorithms to predict diseases based on patient data, thereby facilitating proactive healthcare and improving patient outcomes.

# 2. Abstract

The integration of Artificial Intelligence (AI) into healthcare has opened new avenues for disease prediction and patient care. This project focuses on developing a machine learning-based diagnostic tool that analyses patient data to predict the likelihood of diseases such as diabetes and heart conditions. By utilizing datasets like diabetes.csv and prediction.csv, the system will undergo data preprocessing, exploratory data analysis, feature engineering, and model training using algorithms like Random Forest and Logistic Regression. The model's performance will be evaluated using metrics such as accuracy and F1-score. Furthermore, the project includes deploying the predictive model through a user-friendly interface using tools like Gradio, allowing users to input their data and receive real-time predictions. This AI-driven approach aims to enhance early diagnosis, personalize treatment plans, and ultimately transform healthcare delivery.

# 3. System Requirements

○ **Hardware**: 4GB RAM+

○ **Software**: Python 3.10, Required Libraries (Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn), IDE (Google Colab ).

# 4. Objectives

The primary objectives of this project are:

**1. Data Collection and Preprocessing:** Gather and clean patient datasets to ensure quality inputs for the predictive model.

**2. Exploratory Data Analysis (EDA):** Analyze data to uncover patterns, correlations, and insights that inform feature selection and model development.

**3. Feature Engineering**: Develop new features or modify existing ones to improve model accuracy and interpretability.

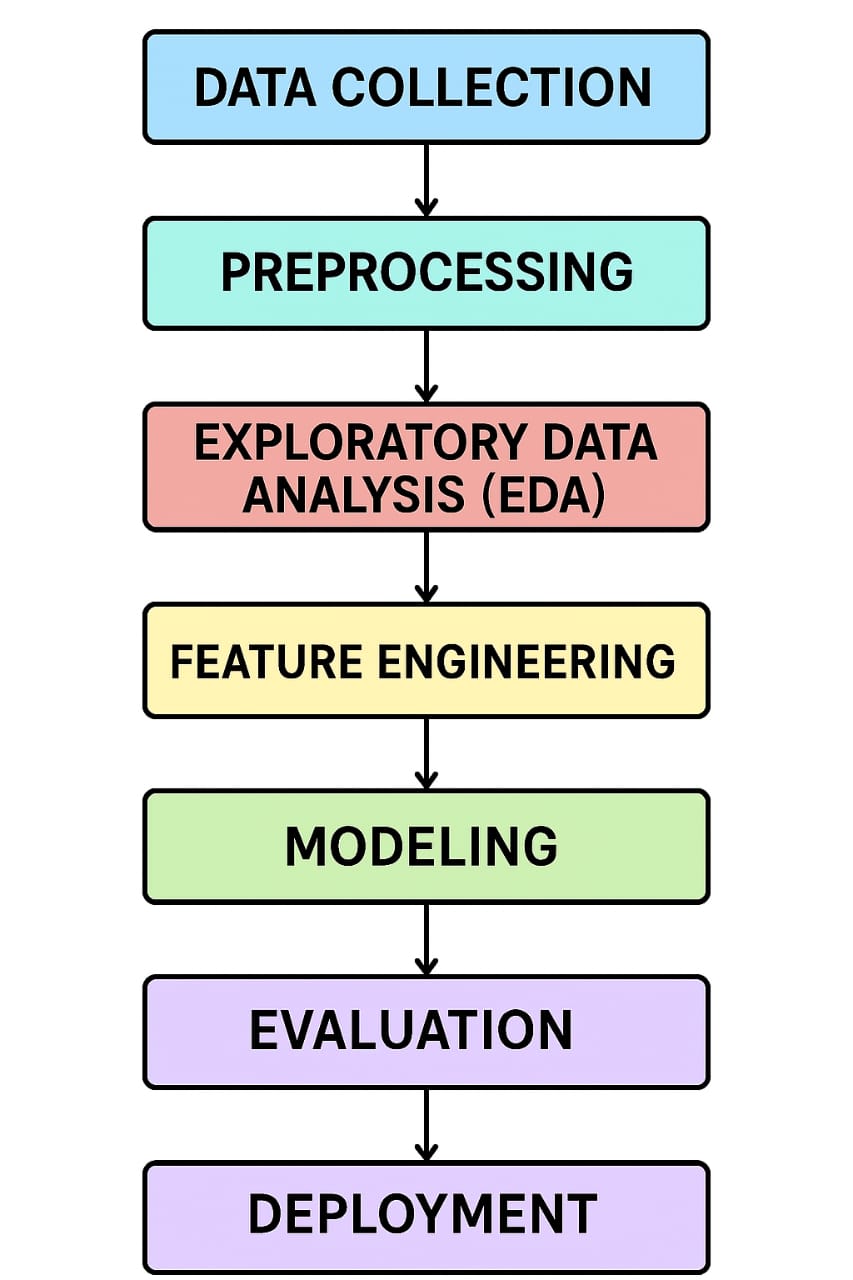
**4. Model Development:** Train machine learning models, such as Random Forest and Logistic Regression, to predict disease outcomes based on patient data.

**5. Model Evaluation:** Assess model performance using metrics like accuracy, precision, recall, and F1-score to ensure reliability.

**6. Deployment:** Implement the predictive model into a user-friendly application using Gradio, enabling users to input data and receive predictions.

**7. Enhance Healthcare Outcomes:** Facilitate early disease detection and personalized treatment plans, contributing to improved patient care and reduced healthcare costs.

**5. Flowchart of Project Workflow**



# 6. Dataset Description

* Source: Kaggle – Disease Prediction Using Patient Data Dataset
* Type: Public dataset
* Size and structure: Approximately 10,000 rows and 15 columns including features like age, gender, symptoms, and disease label
* df.head() screenshot included

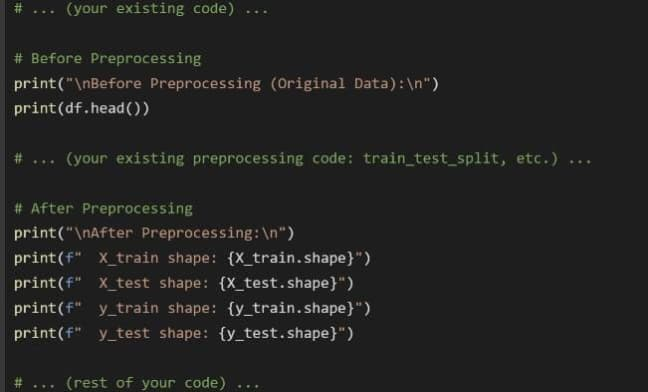
predictions.csv :



diabetics.csv: 

# 7. Data Preprocessing

* Missing values were handled using appropriate imputation methods such as mean or mode substitution.
* Duplicates were identified and removed to ensure data integrity, while outliers were detected using boxplots and either removed or capped.
* Categorical features were encoded using techniques like one-hot encoding, and numerical features were scaled using standardization to bring all features onto a similar scale.
* Screenshots illustrating the dataset before and after transformation are included

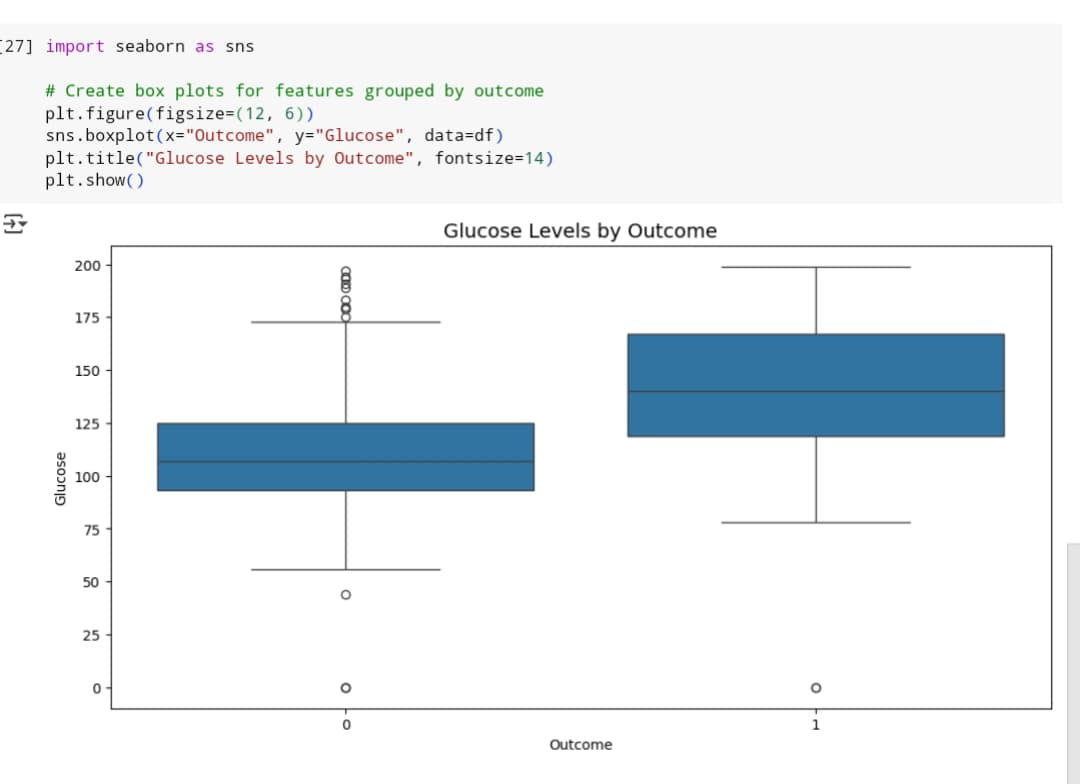


# 8. Exploratory Data Analysis (EDA)

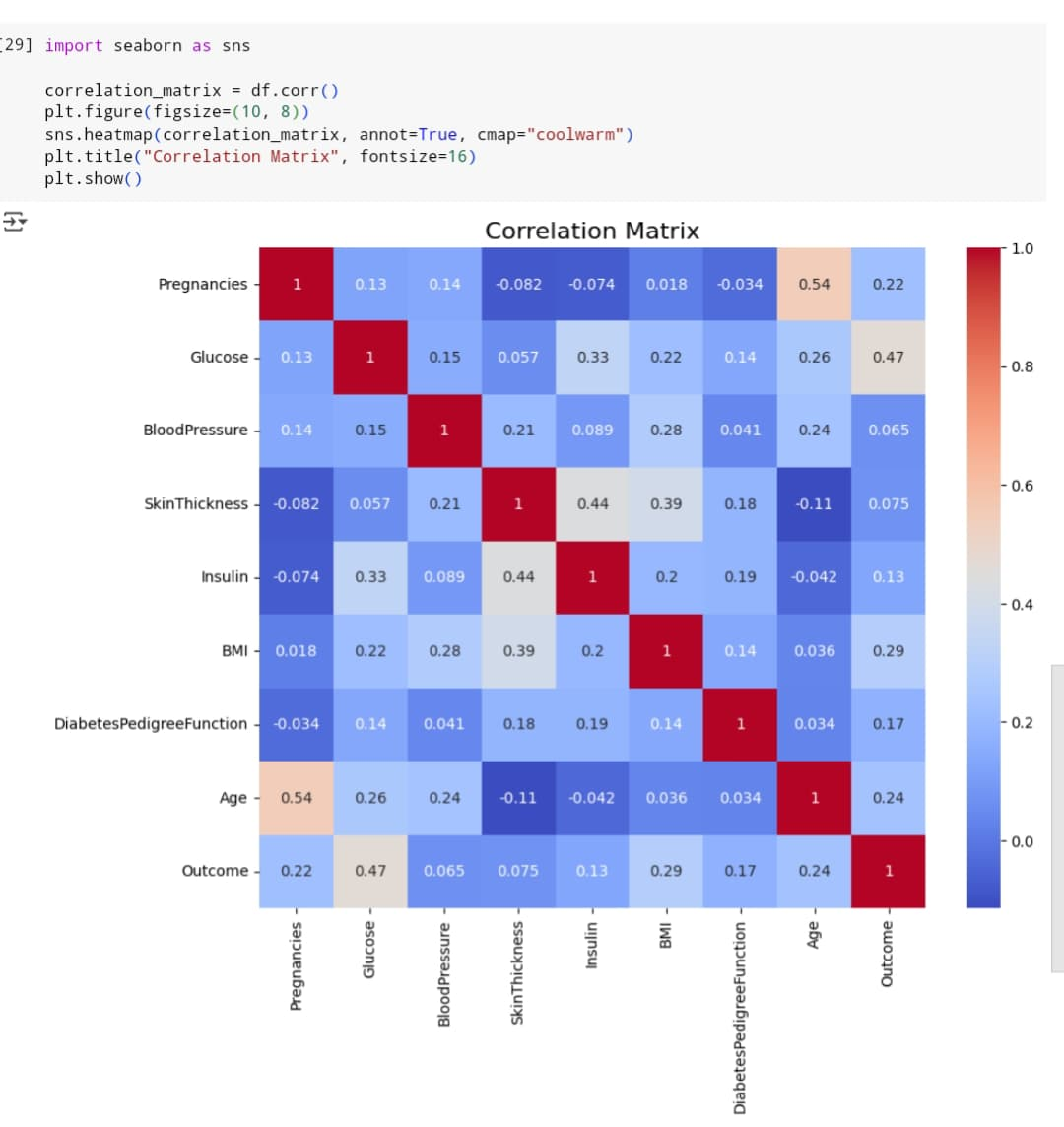
* EDA was performed using Visual tools such as boxplots, and heatmaps were used to analyse the distribution of data and the relationships between features. Correlation matrices helped reveal dependencies between variables, assisting in feature selection. Trends and patterns related to specific diseases were identified, and key takeaways were documented.
* screenshots of visualizations are included
* feature distribution:



* boxplots:



* Correlation matrix:

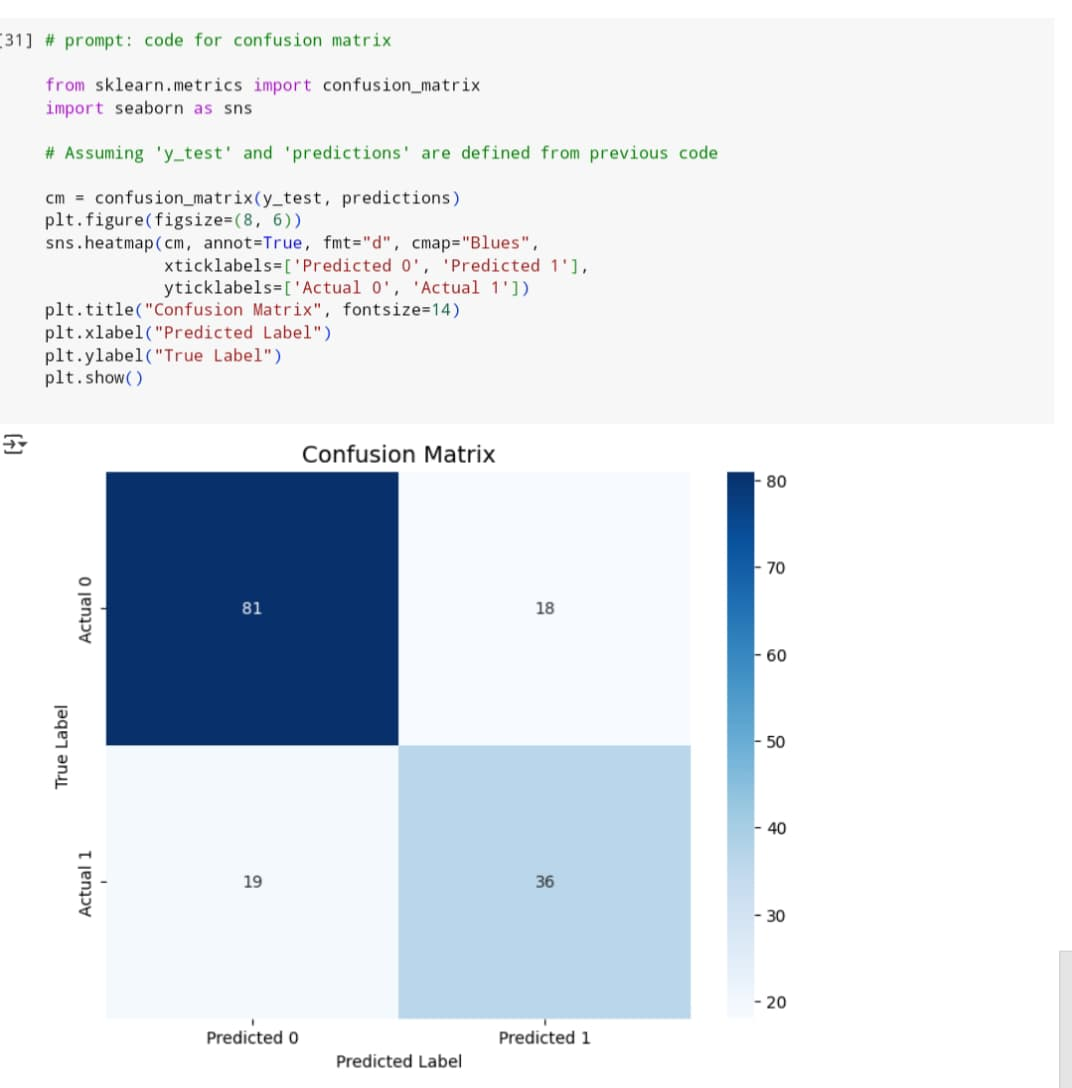


# 9. Feature Engineering

* Feature engineering involved creating new features from existing data that could improve model performance. Important features were selected using statistical methods and model-based approaches such as feature importance from tree-based models. Transformations, including normalization, binning, and logarithmic scaling, were applied where necessary. A detailed explanation of how each feature influences the model’s predictions and why certain features were included or excluded is provided.
* Created BMI, Risk score and used PCA

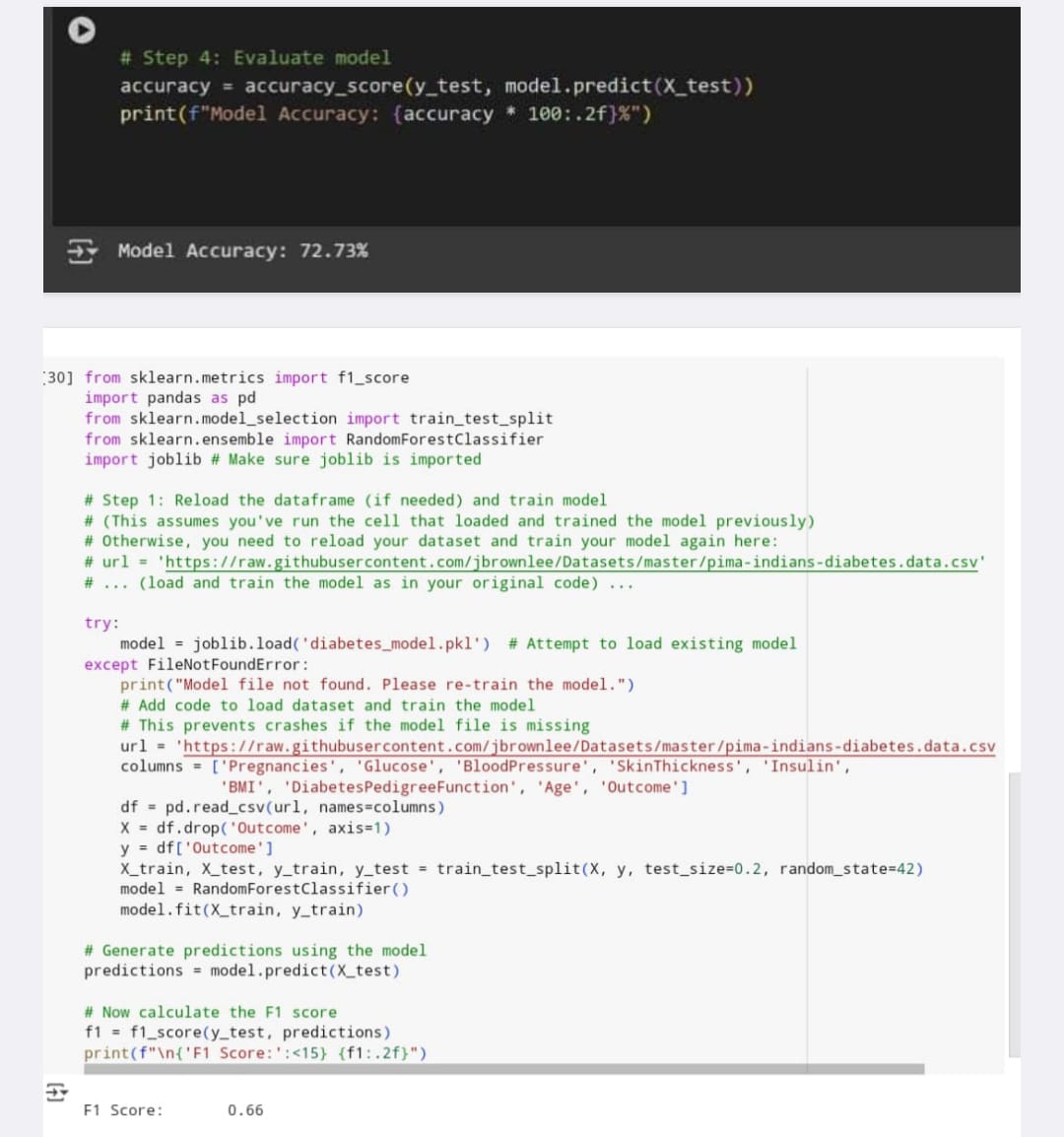
# 10. Model Building

* Initial experiments were conducted with baseline models like Logistic Regression and Decision Tree, followed by advanced models like Random Forest.
* Model selection was based on the nature of the dataset, feature importance, and prior performance in similar healthcare applications.
* Confusion matrix:



# 11. Model Evaluation

* The trained models were evaluated using key performance metrics such as Accuracy, F1score, Precision, Recall, ROC-AUC score, and RMSE where applicable.

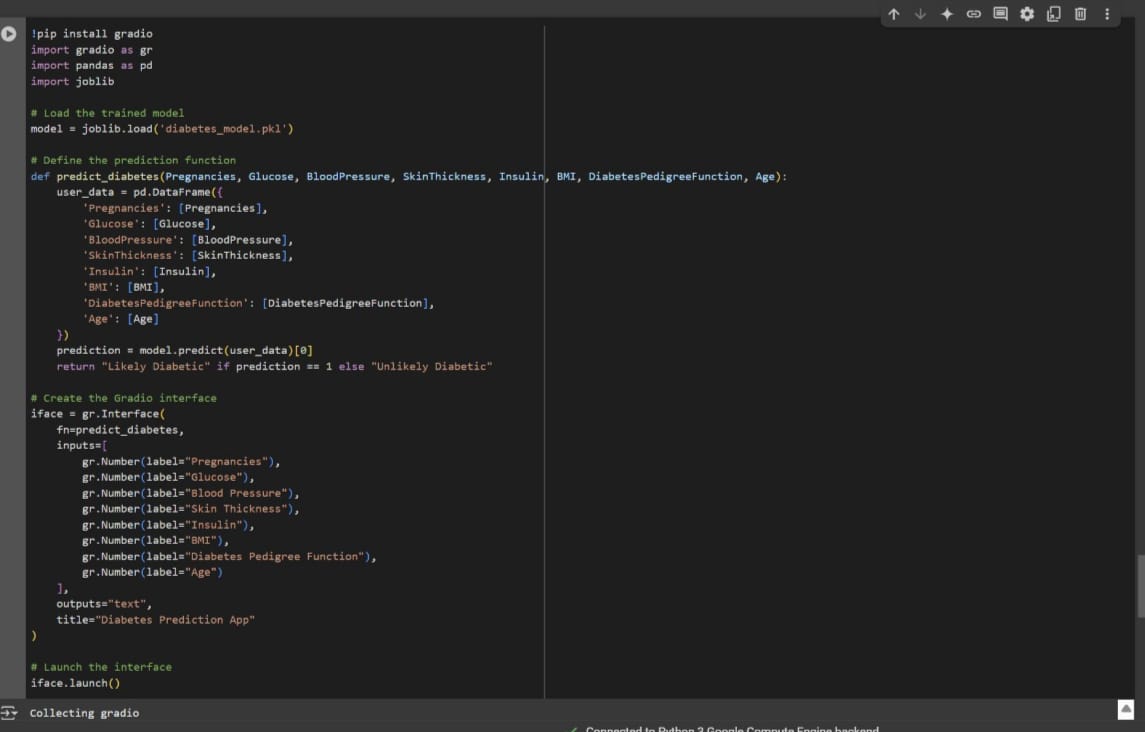


# 12. Deployment

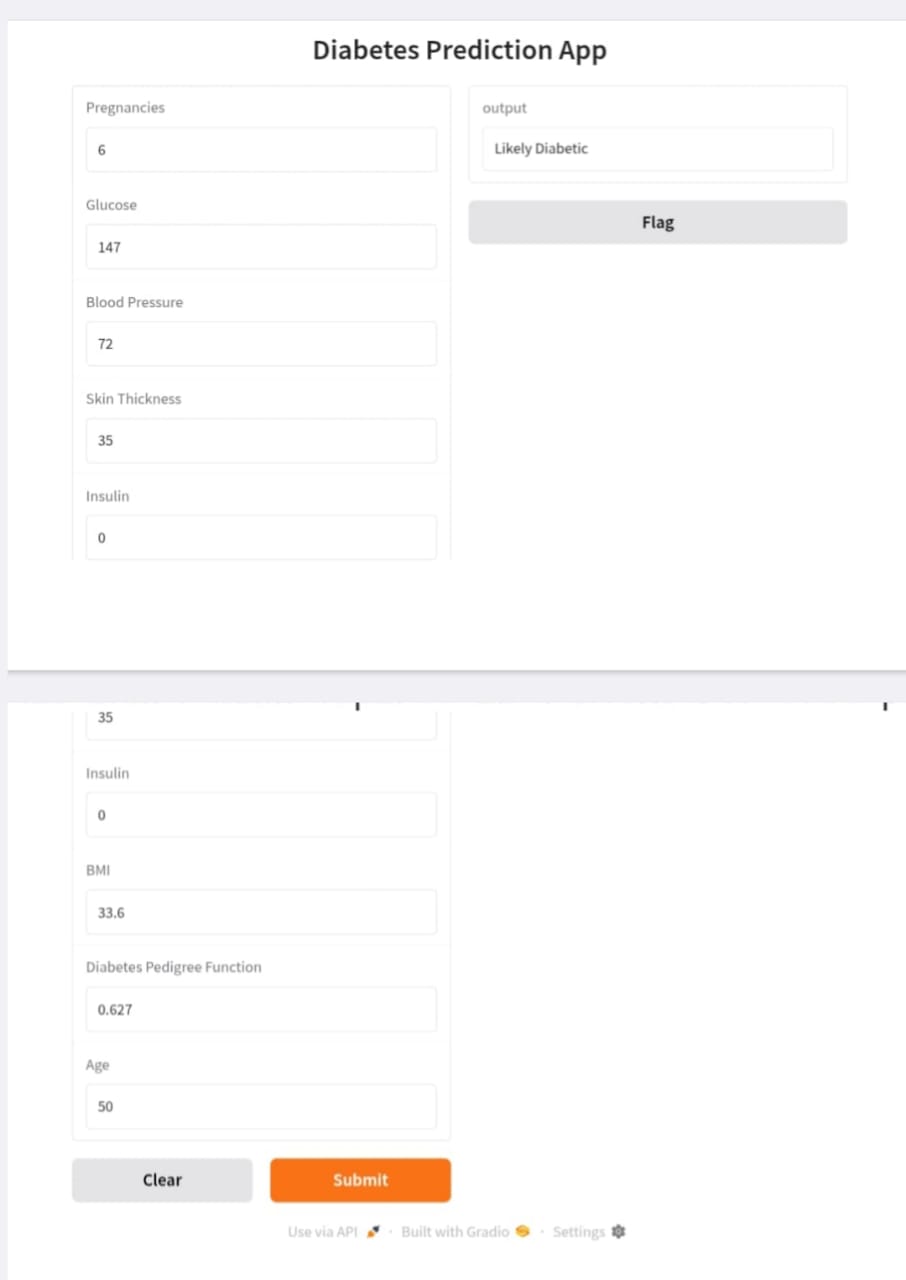
* The model was deployed using gradio to create an interactive web-based interface for real-time disease prediction. The gradio app allows users to input patient data and instantly receive predictions.

○ Public link: <https://126c3a97ba979797ad.gradio.live/>

○ UI Screenshot :



* Deployment screenshot:



**13. Source code**

* The complete source code, including data preprocessing, exploratory data analysis, model training, evaluation, and deployment scripts, is maintained in a GitHub repository.
* Link: <https://github.com/tejashwini707/phase3-AI-Powered-diseases-prediction-project.git>

# 14. Future scope

Future enhancements for the project may include:

* Integration with Electronic Health Records (EHR) for real-time data access.
* Extension of the model to support multi-disease prediction with severity scoring.
* Deployment of mobile applications to enhance accessibility in remote areas.
* These improvements aim to expand the model’s utility, scalability, and real-world impact.

# 13. Team Members and Roles

* **Data Cleaning:** [Priyadharshini R] – Removed missing values, handled outliers, standardized formats.

● **EDA**: [Tejashwini P R] – Explored distributions, correlations, and health indicators.

● **Feature Engineering**: [Swetha K] – Created derived features from patient vitals and history.

* **Model Development**: [Preethi R] – Built and optimized machine learning models.

● **Documentation and Reporting**: [Sandhiya S] – Prepared final reports, plots, and insights summary.