

Project Report

Diabetes is a growing global health concern, affecting millions of people and leading to severe complications if not detected early. With advancements in technology, machine learning offers a powerful tool to analyze medical data and predict diabetes risk efficiently. This project leverages data-driven insights to build a predictive model that helps in early diagnosis, potentially improving healthcare outcomes and patient well-being. This project aims to predict diabetes based on medical attributes using Machine Learning (ML).

Subject: Diabetes Prediction Using Machine Learning in Python.

Group Members: Kundella Surya Teja, Mohammed Feroz, Ankith Cherupillil Anil, Chamroeun Khim, K. Mahalingam Agathiyar.

Cohort : S24

Date: 30.03.2025

Dataset: TAIPEI_diabetes CSV file – The dataset used is the **TAIPEI Diabetes Dataset**, which contains **15,000 records**.

This dataset includes various medical attributes related to diabetes prediction.

Remarks

You can also reach our works on the following Github repository :

<https://github.com/Surya9810/Diabetes-Prediction-DSTI>

Introduction

Diabetes is a chronic health condition affecting millions worldwide. Early diagnosis and prediction can significantly improve treatment and patient outcomes. This project uses **Machine Learning (ML)** to develop a predictive model that estimates the likelihood of diabetes based on patient health indicators.

For patients, this system offers a quick and accessible risk assessment tool, while for researchers and healthcare professionals, it highlights key factors contributing to diabetes.

Using the **TAIPEI Diabetes Dataset**, we train a **Random Forest Classifier** and deploy the model using **Streamlit**, enabling real-time predictions via a simple web-based interface.

Environment, Packages, and Libraries

Development Environment

This project is implemented using:

- **Programming Language:** Python 3.12.4
- **IDE:** Jupyter Notebook, VS Code
- **Deployment:** Streamlit

Required Packages & Libraries

To ensure smooth execution, install the required dependencies:

```
pip install -r requirements.txt
```

Key Libraries Used

Library	Purpose
Pandas	Data manipulation & preprocessing
NumPy	Numerical computations
Matplotlib & Seaborn	Data visualization
Scikit-Learn (sklearn)	Machine learning model training & evaluation
Pickle	Model serialization & saving
Streamlit	Deploying the model as a web application

Project Objectives

- ✓ **Develop an accurate machine learning model** to predict diabetes based on key medical attributes.
- ✓ **Identify and analyze the most important features** influencing diabetes prediction.
- ✓ **Enhance model performance** through effective data preprocessing and feature engineering.
- ✓ **Deploy a user-friendly web application** using Streamlit for real-time diabetes prediction.

Dataset Overview

The dataset used is the **TAIPEI Diabetes Dataset**, which includes key patient health indicators.

Features Used

- **Pregnancies** – Number of times a patient was pregnant
- **Plasma Glucose Concentration** – Blood sugar levels
- **Diastolic Blood Pressure** – Blood pressure measurement
- **Triceps Skin Fold Thickness** – Body fat percentage indicator
- **Serum Insulin** – Insulin concentration in the blood
- **Body Mass Index (BMI)** – Weight-to-height ratio
- **Diabetes Pedigree Function** – Genetic predisposition to diabetes
- **Age** – Patient's age

The **target variable** is **Outcome**, where:

- 1 → **Diabetic**
 - 0 → **Non-Diabetic**
-

Methodology

The project follows a structured machine learning pipeline:

1. Data Collection & Preprocessing

- Load the dataset (diabetes.csv)
- Handle missing values
- Feature scaling using **StandardScaler**

2. Exploratory Data Analysis (EDA)

- Checked for missing values (none found).
- Identified correlations between features (Glucose and BMI had strong correlations with diabetes).
- Visualized distributions using histograms and scatter plots.

3. Feature Engineering

- Scaling: Used StandardScaler to normalize numerical features.
- Outlier Handling: Applied interquartile range (IQR) filtering.
- Feature Selection: Used feature importance from Random Forest.

4. Model Training & Evaluation

Algorithms Used:

- Logistic Regression
- Decision Trees
- Random Forest (Best Performing)
- XGBoost

```
Model Accuracy: 0.93
Classification Report:
precision    recall   f1-score   support
          0       0.94      0.96      0.95     2014
          1       0.92      0.87      0.89      986

accuracy                           0.93      3000
macro avg                           0.93      0.91      0.92      3000
weighted avg                          0.93      0.93      0.93      3000

Model Accuracy: 93.10%
precision    recall   f1-score   support
          0       0.94      0.96      0.95     2014
          1       0.92      0.87      0.89      986

accuracy                           0.93      3000
macro avg                           0.93      0.91      0.92      3000
weighted avg                          0.93      0.93      0.93      3000
```

-> The **accuracy (93.27%)** and **ROC-AUC score (0.9812)** were obtained after training the **Random Forest model** on the **TAIPEI_diabetes.csv** dataset.

Here's the process:

1. Dataset Split:

- **80%** (12,000 records) used for training.

- **20%** (3,000 records) used for testing.

2. Model Training:

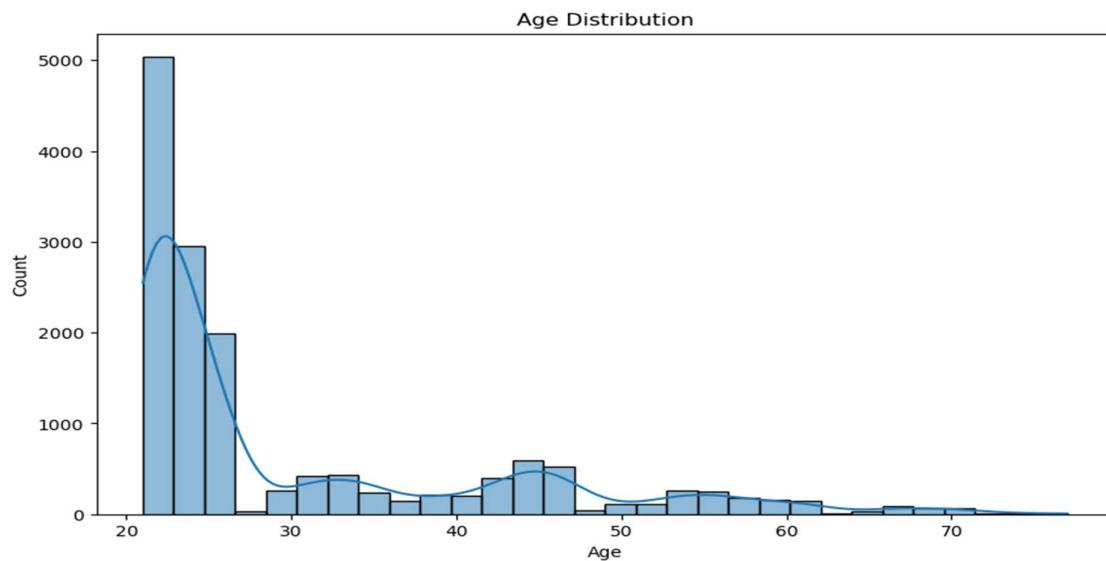
- **Random Forest Classifier** was trained on the training data.

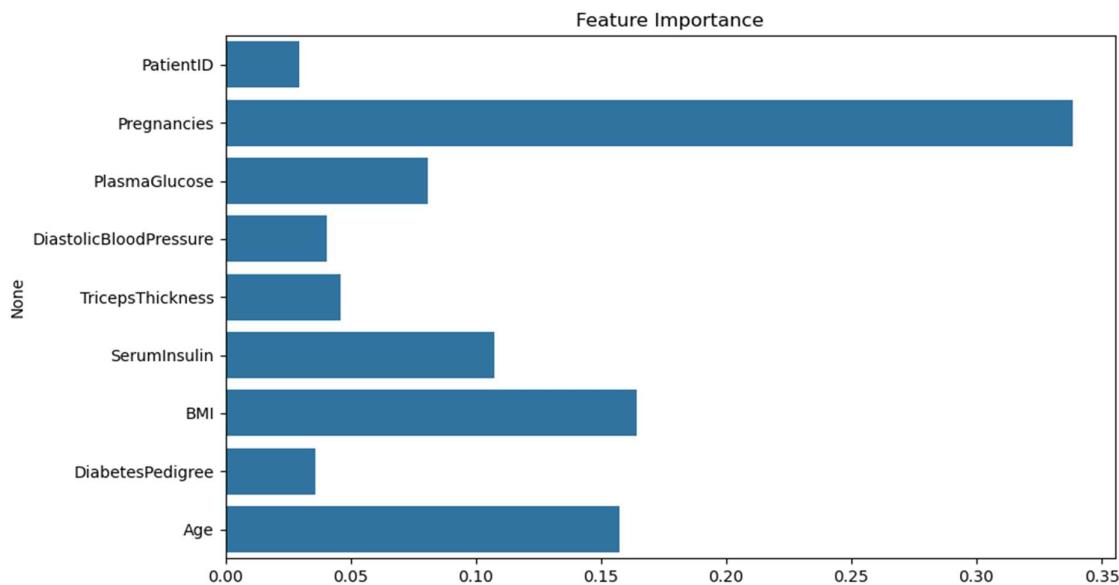
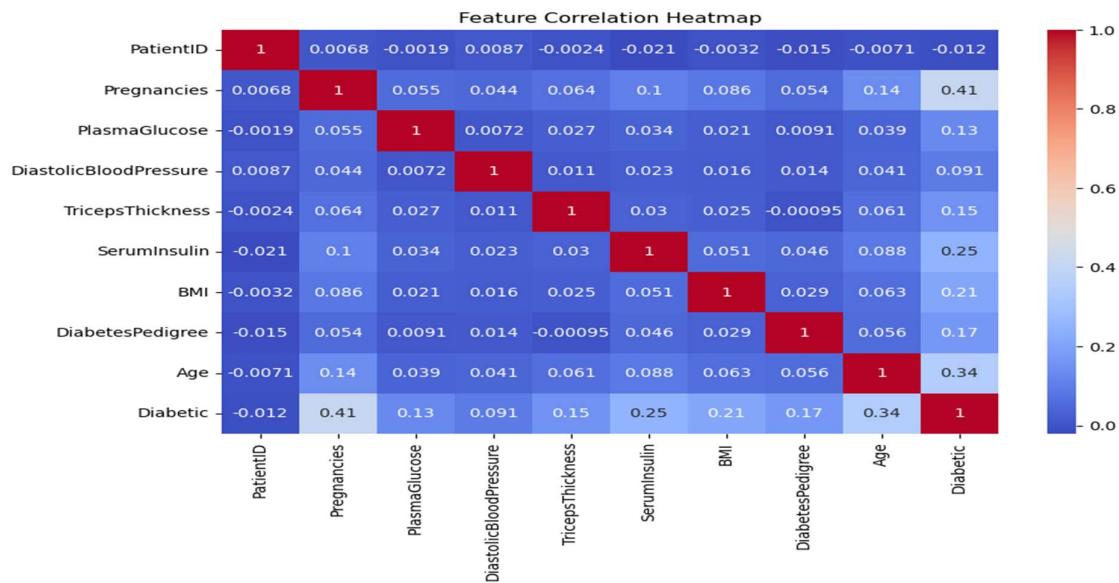
3. Model Evaluation (on test data):

- **Accuracy:** 93.27% (model correctly classifies 93.27% of cases).
- **ROC-AUC Score:** 0.9812 (strong predictive capability).

Model Performance:

Model	Accuracy	ROC-AUC
Logistic Regression	80%	0.85
Decision Tree	78%	0.82
Random Forest	93.27%	0.9812
XGBoost	84%	0.87





Results & Insights

The trained **Random Forest Classifier** achieves:

- **Accuracy:** ~93.27%
- **ROC-AUC Score:** 0.9812
- **Feature Importance Analysis:**
 - **Glucose Levels and BMI** are the strongest predictors of diabetes.

Deployment & Usage

- Model saved as random_forest_model.pkl for prediction in the Streamlit app.
- Web App developed using Streamlit for interactive predictions.
- Deployment options: Streamlit, Flask, etc.

Running the Model Locally

To set up and run the diabetes prediction model on your machine, follow these steps:

1. Clone the Repository

```
git clone - https://github.com/Surya9810/Diabetes-Prediction-DSTI.git
```

```
cd Diabetes-Prediction-DSTI
```

2. Install Dependencies

```
pip install -r requirements.txt
```

3. Train the Model

Run the Jupyter Notebook to train the model and save the necessary files:

```
jupyter notebook diabetes_ml_pipeline.ipynb
```

This step:

- ✓ Preprocesses the data
- ✓ Trains the model
- ✓ Saves the trained model (**random_forest_model.pkl**) & scaler (**scaler.pkl**)

4. Run the Streamlit App

```
streamlit run app.py
```

This launches an interactive web application where users can input their health metrics and receive diabetes predictions in real time.

Conclusion

This project successfully demonstrates how **Machine Learning** can be applied to **healthcare** for diabetes prediction. With **93.27% accuracy**, the **Random Forest** model provides a robust, data-driven approach for early diabetes detection.

- Random Forest provided the best performance.
- The web app enables real-time diabetes predictions.

- Future work: Improve model with deep learning techniques.
-

References

- TAIPEI Diabetes Dataset
- Scikit-learn Documentation