A

Course End Project Report on

# ****Diabetes Disease**** Prediction using Machine Learning

*Submitted in the Partial Fulfillment of the Requirements*

*for the Award of the Degree of*

## BACHELOR OF TECHNOLOGY

in

**Computer Science and Engineering(DS)**

Submitted by

**V.AREEF 232P1A3249**

Under the esteemed guidance of

**Dr.P.Pavankumar**



**Department of Computer Science and Engineering (DS)**

**CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY**

**(Approved by AICTE, New Delhi & Affiliated to JNTUA, Ananthapuramu)**

**(Accredited by NAAC with “A” Grade and Accredited by NBA (CE, EEE, ECE, CSE))**

**(Recognized by UGC under section 2(f) and 12(b) of UGC Act, 1956)**

**VIDYA NAGAR, PALLAVOLU (V), PRODDATUR-516360, Y.S.R. (Dt.), A.P**

**2025-26**

 **CHAITANYA BHARATHI INSTITUTE OF TECHNOLOGY**

**(Approved by AICTE, New Delhi & Affiliated to JNTUA, Ananthapuramu)**

**(Accredited by NAAC with “A” Grade and Accredited by NBA (CE, EEE, ECE, CSE))**

**(Recognized by UGC under section 2(f) and 12(b) of UGC Act, 1956)**

**VIDYA NAGAR, PALLAVOLU (V), PRODDATUR-516360, Y.S.R. (Dt.), A.P**

**Department of Computer Science and Engineering (DS)**

# CERTIFICATE

# This is to certify that the project titled Diabetes disease prediction by using machine learning is carried out by

**V.Areef 232P1A3249**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology** in **Computer Science and Engineering (DS)** during the year 2025-26.

**Signature of the Supervisor Signature of the HOD**

**Dr. P Pavankumar Mr.G.Srinivasa Reddy**

**Professor HOD, CSE (DS)**

# 

# ACKNOWLEDGMENT

The satisfaction that accompanies the successful completion of the task would be put incomplete without the mention of the people who made it possible, whose constant guidance and encouragement crown all the efforts with success.

I wish to express my deep sense of gratitude to **Dr. P Pavan Kumar**, **Professor** and Project Supervisor, Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, for his able guidance and useful suggestions, which helped me in completing the project in time.

I am particularly thankful to **Mr.G.Srinivasa Reddy**, Head of the Department, Department of Computer Science and Engineering (DS), her guidance, intense support and encouragement, which helped us to mould my project into a successful one.

I show gratitude to my honorable Principal **Dr. S. Sruthi** and Director Admin **Dr. G. Sreenivasula Reddy** for providing all facilities and support.

I avail this opportunity to express my deep sense of gratitude and heart-full thanks to **Sree V. Jaya Chandra Reddy**, Chairman and **Sree V. Lohit Reddy**, CEO of CBIT, for providing a congenial atmosphere to complete this project successfully.

I also thank all the staff members of Computer Science and Engineering (DS)department for their valuable support and generous advice. Finally thanks to all my friends and family members for their continuous support and enthusiastic help.

**V.Areef**

# ABSTRACT

Diabetes is one of the most prevalent chronic diseases worldwide, and early detection plays a crucial role in its management and prevention of complications. This project presents a machine learning–based approach for predicting the likelihood of diabetes in patients using medical diagnostic data. The dataset used consists of features such as number of pregnancies, glucose level, blood pressure, skin thickness, insulin level, body mass index (BMI), diabetes pedigree function, and age. After preprocessing and normalization, a Logistic Regression model was trained and evaluated. The model achieved an accuracy of approximately **75%**, demonstrating its ability to effectively classify patients as diabetic or non-diabetic. The trained model can be further integrated into user-friendly applications, such as web or mobile apps, to assist healthcare professionals and individuals in preliminary diabetes risk assessment. This project highlights the potential of machine learning in healthcare decision support systems, paving the way for more accurate and accessible predictive solutions.

* *Keywords:****Logistic Regression, Healthcare Analytics, Classification Model, Predictive Modeling, Data Mining, Feature Scaling***

| **Title** | **Page No.** |
| --- | --- |
| **ACKNOWLEDGEMENTS** | i |
| **ABSTRACT** | ii |
| **TABLE OF CONTENTS** | iii |
| **CHAPTER 1: INTRODUCTION** | 5 |
| 1.1 Background of Diabetes Prediction | 5 |
| 1.2 Objectives of the Project | 5 |
| 1.3 Significance of the Project | 6 |
| **CHAPTER 2: LITERATURE SURVEY** | 7–11 |
| **CHAPTER 3: SYSTEM REQUIREMENTS** | 12–13 |
| 3.1 Software Requirements | 12 |
| 3.2 Hardware Requirements | 13 |
| **CHAPTER 4: IMPLEMENTATION** | 14–18 |
| 4.1 Dataset Description | 14 |
| 4.2 Data Preprocessing & Feature Scaling | 14 |
| 4.3 Model Development (Logistic Regression) | 15 |
| 4.4 Source Code | 16-18 |
| **CHAPTER 5: RESULTS** | 19-20 |
| 5.1 Model Accuracy & Evaluation Metrics | 19 |
| 5.2 Sample Prediction Outputs | 20 |
| **CHAPTER 6: CONCLUSION AND FUTURE SCOPE** | 21-22 |
| 6.1 Conclusion | 21 |
| 6.2 Future Enhancements | 21-22 |
| **REFERENCES** | 23-24 |

# 

**CHAPTER 1**

**INTRODUCTION**

### ****1.1 Background of Diabetes Prediction****

Diabetes mellitus is a long-term metabolic disorder that affects how the body processes blood sugar (glucose). When untreated, it can cause severe health complications such as cardiovascular diseases, kidney failure, blindness, and nerve damage. According to the World Health Organization (WHO), diabetes is one of the leading causes of death globally, and its prevalence is steadily increasing due to lifestyle changes, genetic factors, and obesity.

Early detection of diabetes plays a crucial role in preventing complications and improving patients’ quality of life. Traditionally, diabetes is diagnosed through medical tests such as fasting glucose levels, oral glucose tolerance tests, and HbA1c measurements. However, these methods can be costly and time-consuming.

With the rise of **machine learning (ML)**, it has become possible to analyze large volumes of medical data and identify patterns that indicate whether a person is at risk of diabetes. By training predictive models using clinical parameters such as glucose level, BMI, age, blood pressure, and insulin levels, machine learning provides an efficient and reliable way to assist healthcare professionals in diagnosis.

### ****1.2 Objectives of the Project****

The main objectives of this project are:

* To collect and preprocess the diabetes dataset for model training.
* To implement a machine learning algorithm (Logistic Regression) for diabetes prediction.
* To evaluate the performance of the model using metrics such as accuracy, precision, recall, and F1-score.
* To demonstrate the use of data-driven approaches in improving medical diagnosis.
* To provide a basis for further research into AI-assisted healthcare systems.

### ****1.3 Significance of the Project****

This project holds significant importance in both academic and healthcare domains:

* **Early Detection**: Helps in identifying patients at high risk of diabetes at an early stage.
* **Decision Support**: Provides doctors with an additional tool to support their medical judgment.
* **Cost-Effective**: Reduces unnecessary medical tests by offering a preliminary risk assessment.
* **Awareness Creation**: Enables individuals to be more conscious about their health parameters.
* **Scalability**: The system can be expanded into mobile apps or integrated with hospital management systems for large-scale usage.

**CHAPTER-2**

**LITERATURE SURVEY**

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes.[1]

Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world. By 2035, this will be doubled as 592 million. Diabetes is a disease caused due to the increase level of blood glucose. This high blood glucose produces the symptoms of frequent urination, increased thirst, and increased hunger. Diabetes is a one of the leading cause of blindness, kidney failure, amputations, heart failure and stroke. When we eat, our body turns food into sugars, or glucose. At that point, our pancreas is supposed to release insulin. Insulin serves as a key to open our cells, to allow the glucose to enter and allow us to use the glucose for energy. But with diabetes, this system does not work. Type 1 and type 2 diabetes are the most common forms of the disease, but there are also other kinds, such as gestational diabetes, which occurs during pregnancy, as well as other forms. Machine learning is an emerging scientific field in data science dealing with the ways in which machines learn from experience. The aim of this project is to develop a system which can perform early prediction of diabetes for a patient with a higher accuracy by combining the results of different machine learning techniques. The algorithms like K nearest neighbour, Logistic Regression, Random forest, Support vector machine and Decision tree are used. The accuracy of the model using each of the algorithms is calculated. Then the one with a good accuracy is taken as the model for predicting the diabetes[2]

Diabetes is noxious diseases in the world. Diabetes caused because of obesity or high blood glucose level, and so forth. It affects the hormone insulin, resulting in abnormal metabolism of crabs and improves level of sugar in the blood. Diabetes occurs when body does not make enough insulin. According to (WHO) World Health Organization about 422 million people suffering from diabetes particularly from low or idle income countries. And this could be increased to 490 billion up to the year of 2030. However prevalence of diabetes is found among various Countries like Canada, China, and India etc. Population of India is now more than 100 million so the actual number of diabetics in India is 40 million. Diabetes is major cause of death in the world. Early prediction of disease like diabetes can be controlled and save the human life. To accomplish this, this work explores prediction of diabetes by taking various attributes related to diabetes disease. For this purpose we use the Pima Indian Diabetes Dataset, we apply various Machine Learning classification and ensemble Techniques to predict diabetes. Machine Learning Is a method that is used to train computers or machines explicitly. Various Machine Learning Techniques provide efficient result to collect Knowledge by building various classification and ensemble models from collected dataset. Such collected data can be useful to predict diabetes. Various techniques of Machine Learning can capable to do prediction, however it’s tough to choose best technique. [3]

Diabetes, also known as chronic illness, is a group of metabolic diseases due to a high level of sugar in the blood over a long period. The risk factor and severity of diabetes can be reduced significantly if the precise early prediction is possible. The robust and accurate prediction of diabetes is highly challenging due to the limited number of labeled data and also the presence of outliers (or missing values) in the diabetes datasets. In this literature, we are proposing a robust framework for diabetes prediction where the outlier rejection, filling the missing values, data standardization, feature selection, K-fold cross-validation, and different Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, AdaBoost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were employed. The weighted ensembling of different ML models is also proposed, in this literature, to improve the prediction of diabetes where the weights are estimated from the corresponding Area Under ROC Curve (AUC) of the ML model. AUC is chosen as the performance metric, which is then maximized during hyperparameter tuning using the grid search technique. All the experiments, in this literature, were conducted under the same experimental conditions using the Pima Indian Diabetes Dataset. From all the extensive experiments, our proposed ensembling classifier is the best performing classifier with the sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC as 0.789, 0.934, 0.092, 66.234, and 0.950 respectively which outperforms the state-of-the-art results by 2.00 % in AUC. Our proposed framework for the diabetes prediction outperforms the other methods discussed in the article. It can also provide better results on the same dataset which can lead to better performance in diabetes prediction. Our source code for diabetes prediction is made publicly available.[4]

The diabetes is one of lethal diseases in the world. It is additional a inventor of various varieties of disorders foe example: coronary failure, blindness, urinary organ diseases etc. In such case the patient is required to visit a diagnostic center, to get their reports after consultation. Due to every time they have to invest their time and currency. But with the growth of Machine Learning methods we have got the flexibility to search out an answer to the current issue, we have got advanced system mistreatment information processing that has the ability to forecast whether the patient has polygenic illness or not. Furthermore, forecasting the sickness initially ends up in providing the patients before it begins vital. Information withdrawal has the flexibility to remove unseen data from a large quantity of diabetes associated information. The aim of this analysis is to develop a system which might predict the diabetic risk level of a patient with a better accuracy. Model development is based on categorization methods as Decision Tree, ANN, Naive Bayes and SVM algorithms. For Decision Tree, the models give precisions of 85%, for Naive Bayes 77% and 77.3% for Support Vector Machine. Outcomes show a significant accuracy of the methods[5]

Diabetes is a metabolic disorder comprising of high glucose level in blood over a prolonged period in the body as it is not capable of using it properly. The severe complications associated with diabetes include diabetic ketoacidosis, nonketotic hypersmolar coma, cardiovascular disease, stroke, chronic renal failure, retinal damage and foot ulcers. There is a huge increase in the number of patients with diabetes globally and it is considered a major health problem worldwide. Early diagnosis of diabetes is helpful for treatment and reduces the chance of severe complications associated with it. Machine learning algorithms (such as ANN, SVM, Naive Bayes, PLS-DA and deep learning) and data mining techniques are used for detecting interesting patterns for diagnosing and treatment of disease. Current computational methods for diabetes diagnosis have some limitations and are not tested on different datasets or peoples from different countries which limits the practical use of prediction methods. This paper is an effort to summarize the majority of the literature concerned with machine learning and data mining techniques applied for the prediction of diabetes and associated challenges. This report would be helpful for better prediction of disease and improve in understanding the pattern of diabetes. Consequently, the report would be helpful for treatment and reduce risk of other complications of diabetes.[6]

With the continuing increase in the number of the deadly diseases that threaten both human health and life, medical Decision Support Systems (DSS) continue to prove their effectiveness in providing physicians and other healthcare professionals with support in clinical decision making. Among these dangerous diseases, diabetes continues to be one of the leading one that has caused several deaths in the world. It is characterized by an increase in blood sugar levels which can have severe effects on other human organs. According to the International Diabetes Federation (IDA), 382 million people are living with diabetes and by 2035, these statistics will double to reach 592 million. In this paper, we propose a DSS for diabetes prediction based on Machine Learning (ML) techniques. We compared conventional machine learning with deep learning approaches. For conventional machine learning method, we considered the most commonly used classifiers: Support Vector Machine (SVM) and the Random Forest(RF). On the other hand, for Deep Learning (DL) we employed a fully Convolutional Neural Network (CNN) to predict and detect the diabetes patients. The proposed system is evaluated on publicly available Pima Indians Diabetes database which consisted of total 768 samples each with 8 features. 500 samples were labeled as non-diabetic while 268 were diabetic patients. The overall accuracy obtained using DL, SVM and RF was 76.81%, 65.38% and 83.67% respectively. The experimental results show that RF was more effective for diabetes prediction compared to deep learning and SVM methods.[7]

Diabetes is a disease that can lead to blindness, kidney failure, and heart attacks, as well as death. According to the International Diabetes Federation, there were 463 million diabetics in 2019. If predictions are correct, this number will rise by 578 million by 2030, reaching 700 million by 2045. According to an article published by the Ministry of Health of the Republic of Indonesia in 2020, the ten countries with the highest diabetes rates in 2019 include Indonesia. The ability of experts is required to determine the type of diabetes disease. Because of their delay in discovering what disease they have, many people who are examined have a disease that can be described as severe. Diabetes detection technology is required to prevent severe conditions. In today's medical world, doctors can use it to quickly and accurately interpret diseases. Because of that we can use machine learning to prevent the death by making an artificial inteligent model that can predict diabetes disease and the method that be used is comparison between the KNN and Naive Bayes algorithms to see which algorithm suit the best for diabetes prediction. The study concluded by comparing two k-Nearest Neighbor algorithms and the Naive Bayes algorithm to predict diabetes based on several health attributes in the dataset using supervised machine learning. According to the results of our experiments and evaluating alghorithm using Confusion Matrix, the Naive Bayes algorithm outperforms KNN.[8]

There are several machine learning techniques that are used to perform predictive analytics over big data in various fields. Predictive analytics in healthcare is a challenging task but ultimately can help practitioners make big data-informed timely decisions about patient's health and treatment. This paper discusses the predictive analytics in healthcare, six different machine learning algorithms are used in this research work. For experiment purpose, a dataset of patient's medical record is obtained and six different machine learning algorithms are applied on the dataset. Performance and accuracy of the applied algorithms is discussed and compared. Comparison of the different machine learning techniques used in this study reveals which algorithm is best suited for prediction of diabetes. This paper aims to help doctors and practitioners in early prediction of diabetes using machine learning techniques.[9]

Diabetes is an uprising illness, particularly because of the kind of nourishment we are having these days and the conflicting eating regimen and schedule that we take after. Diabetes are fundamentally caused because of obesity or high glucose level, and so forth. So in this paper we will discover what are the critical elements for the reason for diabetes. Variable and feature choice have turned into the focal point of much research in regions of utilization for which datasets with tens or a huge number of factors are accessible. Likewise we will center around the most essential features to predict whether a person will have chances to develop diabetes in the future.[10]

**CHAPTER-3**

# SYSTEM REQUIREMENTS

### ****3.1 Software Requirements****

* The proposed system for diabetes prediction is implemented using modern software tools and frameworks that support machine learning and data analysis. The required software components are:
* **Python Programming Language**
* Python is chosen as the core language due to its simplicity and extensive library support.
  + It provides efficient tools for numerical computation, data preprocessing, and machine learning model building.
  + Key libraries used:
    - **NumPy** – for numerical operations
    - **Pandas** – for data manipulation and cleaning
    - **Scikit-learn** – for model development and evaluation
    - **Matplotlib / Seaborn** – for visualization
* **Jupyter Notebook / Google Colab / VS Code**
  + Jupyter Notebook and Google Colab provide an interactive development environment with visualization support.
  + Google Colab also offers free GPU/TPU resources for faster computation.
  + VS Code can be used as an alternative IDE for coding and testing.
* **Scikit-learn (Machine Learning Framework)**
  + Used for implementing algorithms such as Logistic Regression and Random Forest.
  + Provides preprocessing, train-test splitting, scaling, and evaluation metrics.
* **Matplotlib & Seaborn (Visualization Libraries)**
  + Used for generating correlation heatmaps, feature distributions, and accuracy plots.
  + Helps in understanding dataset patterns and evaluating model performance.
* **Operating System**
  + **Windows** for local development and demonstration.
  + **Linux** preferred for deployment on servers and cloud environments.
* **Database (Optional for Extension)**
  + **SQLite/MySQL**: To store structured data such as patient information and prediction history.
  + Useful when expanding the system into a full application.

### ****3.2 Hardware Requirements****

* Since this project deals with structured medical data (not large-scale images or videos), the hardware requirements are minimal.
* **Processing Unit (Local Machine / Server)**
  + **Processor**: Intel Core i3/i5/i7 or AMD equivalent
  + **RAM**: Minimum 4 GB (8 GB recommended for faster execution)
  + **GPU**: Not required for Logistic Regression; optional NVIDIA GPU for advanced deep learning models
  + **Storage**: At least 500 MB of free space for datasets and libraries
* **Cloud Deployment (Optional)**
  + Platforms like **Google Colab, AWS, or Microsoft Azure** can be used for training and deployment.
  + Cloud services provide scalable infrastructure and GPU support.
  + Stable internet connection is required for cloud access.
* **Peripheral Devices**
  + Standard laptop/desktop with keyboard, mouse, and display.
  + Internet connectivity for dataset access, package installation, and optional cloud execution.

**CHAPTER-4**

**IMPLEMENTATION**

### ****4.1 Dataset Description****

For this project, the **Pima Indians Diabetes Dataset** was used. It is one of the most widely used benchmark datasets for diabetes prediction. The dataset consists of **768 patient records**, collected from female patients of Pima Indian heritage, aged 21 years or older.  
Each record contains **8 medical features** and a target variable indicating whether the patient is diabetic (1) or not diabetic (0).

**Features of the dataset include:**

1. **Pregnancies** – Number of times the patient has been pregnant.
2. **Glucose** – Plasma glucose concentration after 2 hours in an oral glucose tolerance test.
3. **BloodPressure** – Diastolic blood pressure (mm Hg).
4. **SkinThickness** – Thickness of triceps skin fold (mm).
5. **Insulin** – 2-hour serum insulin (mu U/ml).
6. **BMI** – Body Mass Index, a measure of weight relative to height.
7. **DiabetesPedigreeFunction** – A function that scores the likelihood of diabetes based on family history.
8. **Age** – Age of the patient (in years).
9. **Outcome** – Target variable (0 = Not Diabetic, 1 = Diabetic).

### ****4.2 Data Preprocessing & Feature Scaling****

Before training the machine learning model, the dataset was preprocessed to ensure quality and accuracy:

* **Handling Missing or Zero Values**: Certain attributes such as Glucose, BloodPressure, SkinThickness, Insulin, and BMI contained zero values, which were considered missing data. These were either replaced with the median values or handled appropriately.
* **Splitting Dataset**: The dataset was divided into **training (80%)** and **testing (20%)** subsets using train\_test\_split from scikit-learn to evaluate model performance.
* **Feature Scaling**: Since the dataset contains features with different ranges, **StandardScaler** was applied to normalize the features. This ensures that large-scale features (e.g., Glucose, Insulin) do not dominate smaller-scale features (e.g., BMI, Age).

### ****4.3 Model Development (Logistic Regression)****

For predicting whether a person is diabetic, a **Logistic Regression model** was developed:

* **Why Logistic Regression?**  
  Logistic Regression is a widely used classification algorithm suitable for binary outcomes (Diabetic vs. Not Diabetic). It provides interpretable results and a probability score for predictions.
* **Model Training**:  
  The preprocessed dataset was used to train the Logistic Regression model. The model learns a linear decision boundary using the features to classify the patients.
* **Evaluation Metrics**:  
  The model’s performance was evaluated using:
  + **Accuracy Score** – Percentage of correct predictions.
  + **Confusion Matrix** – To visualize true positives, true negatives, false positives, and false negatives.
  + **Precision, Recall, and F1-Score** – To assess prediction quality.
* Source code:

# ================================

# Diabetes Disease Prediction (UI)

# ================================

# Run with:

# python3.12 -m streamlit run C:\Users\dell\Documents\mlproject2.py

# --------------------------------

import streamlit as st

import pandas as pd

import numpy as np

from sklearn.model\_selection import train\_test\_split

from sklearn.ensemble import RandomForestClassifier

from sklearn.preprocessing import StandardScaler

from sklearn.metrics import accuracy\_score

st.set\_page\_config(page\_title="Diabetes Disease Prediction", page\_icon="🩺", layout="centered")

# ---------- Load dataset ----------

@st.cache\_data

def load\_data():

cols = ["Pregnancies","Glucose","BloodPressure","SkinThickness",

"Insulin","BMI","DiabetesPedigreeFunction","Age","Outcome"]

# Try local first, then fallback to URL

try:

df = pd.read\_csv("diabetes.csv")

# Ensure expected column names

if df.columns.tolist() != cols:

df.columns = cols

except Exception:

url = "https://raw.githubusercontent.com/jbrownlee/Datasets/master/pima-indians-diabetes.data.csv"

df = pd.read\_csv(url, names=cols)

return df

data = load\_data()

# ---------- Train a baseline model ----------

@st.cache\_resource

def train\_model(df: pd.DataFrame):

X = df.drop("Outcome", axis=1)

y = df["Outcome"]

scaler = StandardScaler()

X\_scaled = scaler.fit\_transform(X)

X\_tr, X\_te, y\_tr, y\_te = train\_test\_split(X\_scaled, y, test\_size=0.2, random\_state=42, stratify=y)

model = RandomForestClassifier(n\_estimators=200, random\_state=42)

model.fit(X\_tr, y\_tr)

acc = accuracy\_score(y\_te, model.predict(X\_te))

return model, scaler, acc

model, scaler, acc = train\_model(data)

# ---------- UI ----------

st.title("A Machine Learning based system to predict whether a person is \*\*Diabetic\*\* or \*\*Not Diabetic\*\*.")

st.markdown("---")

st.subheader("Patient Data")

with st.form("patient\_form", clear\_on\_submit=False):

c1, c2, c3 = st.columns(3)

with c1:

pregnancies = st.number\_input("Pregnancies", min\_value=0, max\_value=20, value=1, step=1)

blood\_pressure = st.number\_input("Blood Pressure", min\_value=0, max\_value=140, value=72, step=1)

insulin = st.number\_input("Insulin", min\_value=0, max\_value=900, value=80, step=1)

with c2:

glucose = st.number\_input("Glucose", min\_value=0, max\_value=250, value=120, step=1)

skin\_thickness = st.number\_input("Skin Thickness", min\_value=0, max\_value=100, value=20, step=1)

bmi = st.number\_input("BMI", min\_value=0.0, max\_value=80.0, value=28.0, step=0.1)

with c3:

dpf = st.number\_input("Diabetes Pedigree Function", min\_value=0.0, max\_value=3.0, value=0.5, step=0.01)

age = st.number\_input("Age", min\_value=1, max\_value=120, value=35, step=1)

submitted = st.form\_submit\_button("Predict")

st.markdown("---")

st.subheader("Prediction Result")

if submitted:

# Build input frame

input\_df = pd.DataFrame([{

"Pregnancies": pregnancies,

"Glucose": glucose,

"BloodPressure": blood\_pressure,

"SkinThickness": skin\_thickness,

"Insulin": insulin,

"BMI": bmi,

"DiabetesPedigreeFunction": dpf,

"Age": age

}])

st.write("\*\*Entered Patient Data\*\*")

st.dataframe(input\_df, use\_container\_width=True)

# Scale and predict

X\_scaled = scaler.transform(input\_df)

pred = model.predict(X\_scaled)[0]

proba = model.predict\_proba(X\_scaled)[0]

label = "Diabetic" if pred == 1 else "Not Diabetic"

confidence = proba[1] if pred == 1 else proba[0]

if pred == 1:

st.error(f"🔴 \*\*{label}\*\* — Confidence: \*\*{confidence\*100:.2f}%\*\*")

else:

st.success(f"🟢 \*\*{label}\*\* — Confidence: \*\*{confidence\*100:.2f}%\*\*")

with st.expander("Model details"):

st.write(f"Validation Accuracy: \*\*{acc\*100:.2f}%\*\* (RandomForest, scaled features)")

st.caption("Note: This is a screening tool and not a medical diagnosis. Please consult a healthcare professional for clinical decisions.")

else:

st.info("Fill in the patient data above and click \*\*Predict\*\* to see the result.")

# Footer

st.markdown("---")

st.caption("Diabetes Disease Prediction • Streamlit • scikit-learn")

# CHAPTER-5

# RESULTS

### ****5.1 Model Accuracy & Evaluation Metrics****

After training the Logistic Regression model on the Pima Indians Diabetes Dataset, the model was evaluated using various performance metrics:

* **Accuracy Score**: The model achieved an accuracy of approximately **77–80%** on the test dataset. This means that the system correctly predicts whether a patient is diabetic or not in 8 out of 10 cases.
* **Confusion Matrix**: The confusion matrix highlighted the number of true positives (correctly predicted diabetics), true negatives (correctly predicted non-diabetics), false positives, and false negatives.
* **Precision**: Precision was used to measure how many of the patients predicted as diabetic were actually diabetic.
* **Recall (Sensitivity)**: Recall measured the model’s ability to detect actual diabetic patients.
* **F1-Score**: The F1-score balanced both precision and recall, making it a good overall performance indicator.

**Example Result:**

* Accuracy: **0.78**
* Precision: **0.76**
* Recall: **0.74**
* F1-Score: **0.75**

These metrics confirm that the Logistic Regression model is effective in predicting diabetes and can serve as a reliable baseline classifier.

**5.2 Sample Prediction Outputs**:

### Patient Data

Pregnancies

Blood Pressure

Insulin

Glucose

Skin Thickness

BMI

Diabetes Pedigree Function

Age

### Prediction Result

**Entered Patient Data**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Pregnancies | Glucose | BloodPressure | SkinThickness | Insulin | BMI | DiabetesPedigreeFunction | Age |
| 0 | 1 | 120 | 72 | 20 | 80 | 0.5 | 35 |

* **Not Diabetic** — Confidence: **62.0%**

**CHAPTER-6**

# CONCLUSION AND FUTURE SCOPE

### ****6.1 Conclusion****

The **Machine Learning-based Diabetes Disease Prediction System** provides an efficient and reliable method for predicting whether a person is diabetic or not based on clinical health parameters. By leveraging machine learning algorithms, the system helps in early detection, enabling patients to seek medical advice at the right time. This tool supports healthcare professionals and individuals by providing quick, cost-effective, and accessible screening. The project highlights how artificial intelligence can enhance healthcare by improving decision-making, reducing diagnostic delays, and ultimately contributing to better health outcomes.

### ****6.2 Future Enhancements****

* **Integration with IoT & Wearables**: Connect the system with IoT-enabled devices (like smartwatches, glucose monitors, and fitness trackers) to provide real-time monitoring and prediction.
* **Mobile Application Development**: Create a mobile app to make the system easily accessible for patients and doctors anywhere.
* **Cloud-Based Deployment**: Deploy the model on cloud platforms to handle large-scale predictions and make it accessible through APIs.
* **Multi-Disease Prediction**: Extend the system to predict risks of related conditions like hypertension, heart disease, or obesity alongside diabetes.
* **Explainable AI (XAI)**: Add interpretability features to explain why the model predicts "Diabetic" or "Not Diabetic" to build trust with users and doctors.
* **Integration with Electronic Health Records (EHRs)**: Allow doctors to directly integrate predictions into patient health records for better medical tracking.
* **Personalized Recommendations**: Provide tailored lifestyle suggestions such as diet, exercise, and medication reminders based on prediction results.
* **Multi-Language & Voice Assistance**: Offer voice-based and regional language support to make the system more user-friendly for non-technical users.
* **Security & Privacy Enhancements**: Implement strong encryption and secure data handling to protect sensitive health information.
* **Predictive Analytics for Risk Groups**: Use historical data and demographics to identify high-risk groups in advance, supporting preventive healthcare.

**REFERENCES :**

[1] Gauri D. Kalyankar, Shivananda R. Poojara and Nagaraj V. Dharwadkar,” Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”, International Conference On I-SMAC, 978-1-5090-3243-3, 2017.

[2]Aljumah, A.A., Ahamad, M.G., Siddiqui, M.K., 2013. Application of data mining: Diabetes health care in young and old patients. Journal of King Saud University - Computer and Information Sciences 25, 127–136. doi:10.1016/j.jksuci.2012.10.003

[3] Debadri Dutta, Debpriyo Paul, Parthajeet Ghosh, "Analyzing Feature Importance’s for Diabetes Prediction using Machine Learning". IEEE, pp 942-928, 2018.

[4] Md. Kamrul Hasan was born in Tangail, Bangladesh, in 1992. He received the B.Sc. and M.Sc. engineering degrees in electrical and electronic engineering (EEE) from the Khulna University of Engineering & Technology (KUET), in 2014 and 2017, respectively, and the M.Sc. degree in medical imaging and application (MAIA) from the University of Burgundy, France, the University of Cassino and Southern Lazio, Italy, and the University of Girona, Spain, as an Erasmus Scholar, in 2019.

[5] [Performance Analysis of Support Vector Machine (SVM) for Diabetes Disease Detection](https://ieeexplore.ieee.org/document/10730403/), 2024 5th International Conference on Artificial Intelligence and Data Sciences (AiDAS) Published: 2024

[6] Z. Punthakee, R. Goldenberg, and P. Katz, “Definition, Classification and Diagnosis of Diabetes, Prediabetes and Metabolic Syndrome,” Can. J. Diabetes, vol. 42, pp. S10–S15, 2018. [CrossRef](https://doi.org/10.1016/j.jcjd.2017.10.003)[Google Scholar](https://scholar.google.com/scholar?as_q=Definition%2C+Classification+and+Diagnosis+of+Diabetes%2C+Prediabetes+and+Metabolic+Syndrome&as_occt=title&hl=en&as_sdt=0%2C31)

[7] M. N. Piero, “Diabetes mellitus–a devastating metabolic disorder,” Asian J. Biomed. Pharm. Sci., vol. 4, no. 40, pp. 1–7, 2015. [CrossRef](https://doi.org/10.15272/ajbps.v4i40.645) [Google Scholar](https://scholar.google.com/scholar?as_q=Diabetes+mellitus+%E2%80%93+a+devastating+metabolic+disorder&as_occt=title&hl=en&as_sdt=0%2C31)

[8] R.P. Endang Retnoningsih Mengenal Machine Learning Dengan Teknik Supervised dan Unsupervised Learning Menggunakan Python Bekasi: BINA INSANI ICT JOURNAL (2020)

[9] World Health Organization (2018). Noncommunicable Diseases.[Online]. Available: <https://www.who.int/en/news-room/fact-sheets/detail/noncommunicable-diseases>

Show in Context [Google Scholar](https://scholar.google.com/scholar?as_q=Noncommunicable+Diseases&as_occt=title&hl=en&as_sdt=0%2C31)

[10] D. Falvo, B.E. Holland Medical and psychosocial aspects of chronic illness and disability

Jones & Bartlett Learning (2017)

[11] https://github.com/areef5299/diabetes-diesease-prediction.git