Diabetes Prediction using Machine Learning

1. Introduction

Diabetes is a serious chronic disease that occurs when the body cannot produce or properly use insulin.
According to the World Health Organization (WHO), diabetes is one of the fastest-growing health challenges, and early detection is very important for treatment.

In this project, we build a **machine learning model** that can predict whether a patient is diabetic or not, based on certain medical parameters.
We use the **Pima Indians Diabetes Dataset**, which is a widely used dataset in machine learning research.

2. Problem Statement

The main question this project answers is:
"Given the medical details of a patient, can we predict if the person is diabetic?"

This is a **binary classification problem** where:

- **0 = Non-Diabetic**
- **1 = Diabetic**

3. Objectives

- To perform **data analysis and visualization** on the diabetes dataset.
- To preprocess the dataset and prepare it for training.
- To build and train a **Support Vector Machine (SVM)** model.
- To evaluate the model using accuracy, confusion matrix, and charts.
- To create a system that can take new patient input and predict diabetes.

4. Dataset Description

The dataset has **768 records** and **9 columns** (8 input features + 1 target).

Feature	**Description**
Pregnancies	Number of times pregnant
Glucose	Plasma glucose concentration (mg/dL)
BloodPressure	Diastolic blood pressure (mm Hg)
SkinThickness	Triceps skin fold thickness (mm)
Insulin	2-hour serum insulin (mu U/ml)
BMI	Body Mass Index (weight/height^2)
DiabetesPedigreeFunction	Likelihood of diabetes based on family history
Age	Patient age (years)
Outcome	Target (0 = Non-Diabetic, 1 = Diabetic)

---

5. Methodology

The steps followed in this project are:

Step 1: Data Collection and Exploration

- Loaded the dataset using **Pandas**.
- Checked dataset shape, missing values, and summary statistics.
- Visualized data distribution using **Seaborn and Matplotlib**.

Step 2: Data Preprocessing

- Divided dataset into **features (X)** and **target (Y)**.
- Used **StandardScaler** to normalize data (important for SVM).
- Split dataset into **training set (80%)** and **testing set (20%)** using stratified sampling.

Step 3: Model Training

- Used **Support Vector Machine (SVM)** with linear kernel.
- Trained the model using training data.

Step 4: Model Evaluation

- Calculated **training accuracy** and **testing accuracy**.
- Generated **confusion matrix** to analyze classification performance.
- Plotted accuracy comparison chart.

Step 5: Prediction System

- Created a function to input patient details (Glucose, BMI, Age, etc.).
- Preprocessed the input and used the trained model to predict diabetes.

---

6. Results and Analysis

- **Training Accuracy:** ~ 0.79 (79%)
- **Testing Accuracy:** ~ 0.77 (77%)

This shows the model generalizes well and is **not overfitting**.

🔷 **Key Insights from EDA:**

- Diabetic patients generally have **higher glucose levels**.
- BMI and Age also strongly influence diabetes.
- Skin thickness and insulin levels show weaker correlations.

🔷 **Confusion Matrix (Example):**

	Predicted Non-Diabetic	**Predicted Diabetic**
Actual Non-Diabetic	80	20
Actual Diabetic	15	39

- The confusion matrix shows most patients are correctly classified.

---

7. Visualizations

1. **Class Distribution Chart** → Shows number of diabetic vs non-diabetic patients.

2. **Correlation Heatmap** → Shows which features are most related to diabetes (Glucose, BMI, Age).
3. **Boxplot of Glucose vs Outcome** → Diabetic patients have higher glucose values.
4. **Confusion Matrix Heatmap** → Visual evaluation of correct and incorrect predictions.
5. **Training vs Testing Accuracy Chart** → Comparison of model performance.

---

8. Conclusion

- The **SVM model** achieved **77–79% accuracy** on the diabetes dataset.
- Glucose, BMI, and Age are strong predictors of diabetes.
- The model can successfully classify patients into **diabetic** or **non-diabetic** categories.

Future Work:

- Use more advanced models (Random Forest, Logistic Regression, Deep Learning).
- Perform **hyperparameter tuning** for SVM.
- Deploy the model as a **web application** (Flask or Streamlit).
- Collect more real-world data for higher accuracy.

---

9. Technologies Used

- **Python**
- **NumPy, Pandas** → Data handling
- **Matplotlib, Seaborn** → Visualization
- **Scikit-learn** → ML algorithms, preprocessing, metrics

---

10. References

- Pima Indians Diabetes Dataset (Kaggle / UCI Repository)
- Scikit-learn Documentation
- Data Science & ML Tutorials