**Project Report**

**Diabetes Prediction Using Machine Learning**

**Submitted by: DOOR DARSHITA PUROHIT**  
**Class:** MCA  
**University:** Jaypee University, Anoopshahr

**1. Introduction**

Diabetes is a rapidly growing global health concern that can lead to severe complications if not detected early. Predictive modeling using machine learning offers a powerful tool to assist healthcare professionals in identifying potential diabetic patients through automated analysis of clinical data. This project aims to develop a reliable diabetes prediction model using commonly available patient features.

**2. Objective**

To build a machine learning model capable of accurately predicting whether a person is diabetic or non-diabetic based on medical and physiological attributes. This can aid in early diagnosis and better preventive care strategies.

**3. Dataset Description**

* **Source**: Publicly available diabetes dataset (Pima Indians Diabetes Database).
* **Number of Records**: 768 samples.
* **Features Used**:
  + **Pregnancies** – Number of times pregnant
  + **Glucose** – Plasma glucose concentration
  + **Blood Pressure** – Diastolic blood pressure (mm Hg)
  + **Skin Thickness** – Triceps skin fold thickness (mm)
  + **Insulin** – 2-Hour serum insulin (mu U/ml)
  + **BMI** – Body Mass Index
  + **Diabetes Pedigree Function** – Genetic relationship impact
  + **Age** – Age in years
  + **Outcome** – Class variable (0 = Non-diabetic, 1 = Diabetic)

**4. Tools & Technologies**

* **Platform**: Google Colab (Cloud-based Python environment)
* **Language**: Python
* **Libraries**:
  + pandas and numpy for data manipulation
  + matplotlib and seaborn for data visualization
  + scikit-learn for machine learning models and evaluation

**5. Data Preprocessing**

* Checked for missing and zero values in critical columns.
* Replaced invalid zeros with mean/median where necessary.
* Standardized the data using StandardScaler to ensure uniformity in feature scaling.
* Split the data into training and testing sets (80:20 ratio).

**6. Model Development**

Several machine learning algorithms were trained and tested:

* **Logistic Regression**
* **K-Nearest Neighbors (KNN)**
* **Decision Tree Classifier**
* **Random Forest Classifier**
* **Support Vector Machine (SVM)**

Each model was evaluated for performance using test data. Hyperparameter tuning was also applied to improve results.

**7. Evaluation Metrics**

The following metrics were used to assess model performance:

* **Accuracy Score**
* **Confusion Matrix**
* **Precision, Recall, and F1-Score**
* **ROC-AUC Curve**

Visual tools such as heatmaps and classification reports were used to interpret the results effectively.

**8. Results**

* The **Random Forest Classifier** yielded the highest accuracy of **[Insert your actual accuracy here, e.g., 82.4%]**.
* The model demonstrated a balanced trade-off between precision and recall, making it robust for real-world application.

**9. Conclusion**

This project successfully implemented a machine learning approach to predict diabetes using real-world medical data. The models, particularly ensemble techniques like Random Forest, showed promising accuracy and reliability. With proper deployment and continuous data feeding, such a system can act as an early diagnostic support tool in the healthcare sector.

**10. Future Scope**

* Implement deep learning (e.g., neural networks) for improved prediction.
* Integrate real-time prediction into mobile or web-based health apps.
* Extend the model with more diverse datasets and patient demographics.
* Address data imbalance with oversampling techniques like SMOTE.