Diabetes Clinical Prediction Project Documentation

1. Project Overview

This project focuses on the development of a Diabetes Clinical Prediction Web Application using a trained machine learning model. The goal is to assist clinicians and healthcare practitioners in assessing the risk of diabetes in patients based on clinical, lifestyle, and laboratory data. The model predicts the likelihood of diabetes and provides interpretability through a user-friendly Streamlit interface.

2. Objective

The main objective of this project is to provide a reliable, data-driven tool that allows medical professionals to input patient details and receive a probabilistic prediction of diabetes risk. The system is designed to enhance preventive healthcare and support clinical decision-making.

3. Data and Feature Selection

The dataset used for model training includes various lifestyle, physiological, and medical history parameters. After preprocessing and feature selection, the following features were chosen as the most relevant predictors of diabetes risk:

- 1. age – Age of the patient in years.
- 2. alcohol_consumption_per_week – Number of alcoholic drinks consumed per week; excessive intake raises metabolic risks.
- 3. physical_activity_minutes_per_week – Weekly minutes of physical activity; higher values reduce diabetes risk.
- 4. sleep_hours_per_day – Average sleep duration per day; too little or too much sleep affects metabolism.
- 5. screen_time_hours_per_day – Average daily screen exposure; prolonged sedentary screen time increases risk.
- 6. family_history_diabetes – Encoded as 1 if the patient has a family history of diabetes; 0 otherwise.
- 7. hypertension_history – Encoded as 1 if the patient has a history of hypertension (high blood pressure); 0 otherwise.
- 8. cardiovascular_history – Encoded as 1 if the patient has a cardiovascular disease history; 0 otherwise.
- 9. bmi – Body Mass Index (weight/height^2). BMI > 25 indicates overweight; > 30 obesity.
- 10. systolic_bp – Systolic blood pressure (mmHg).
- 11. diastolic_bp – Diastolic blood pressure (mmHg).
- 12. cholesterol_total – Total serum cholesterol level (mg/dL).
- 13. glucose_fasting – Fasting plasma glucose concentration (mg/dL).
- 14. glucose_postprandial – Two-hour postprandial glucose level (mg/dL).

- 15. insulin_level – Fasting insulin concentration (µU/mL).
- 16. Smoking_Status_Encoded – Encoded smoking status: 1 for smokers, 0 for non-smokers.

4. Encoded Columns Explanation

Some categorical variables were encoded into numeric form to make them suitable for machine learning model training. Below is a description of each encoded column:

- family_history_diabetes: 1 = Patient has a family history of diabetes, 0 = No family history.

- hypertension_history: 1 = History of hypertension, 0 = No hypertension.

- cardiovascular_history: 1 = History of cardiovascular disease, 0 = No such history.

- Smoking_Status_Encoded: 1 = Smoker, 0 = Non-smoker.

5. Model Development

The model was developed using Python’s scikit-learn library. Various supervised learning algorithms were evaluated, and the model achieving the best balance of accuracy, sensitivity, and specificity was selected. The trained model was serialized and saved in a '.joblib' file for deployment in the Streamlit web application.

6. Application Interface

The web application was built using Streamlit, providing a clean, aesthetic, and intuitive user interface. The application allows clinicians to input patient data, adjust the decision threshold slider, and view the resulting risk probability along with an interpretation ('Diabetes likely' or 'Diabetes unlikely'). Additional sections include feature explanations, model information, and an about page.

7. Decision Threshold Explanation

The threshold slider enables clinicians to adjust the model's sensitivity and specificity. Lower thresholds increase sensitivity (detecting more at-risk patients) but may increase false positives, while higher thresholds increase specificity but may miss early-stage cases. The default threshold is set at 0.5, representing a balanced decision point.

8. Clinical Use Disclaimer

This predictive model is intended as a clinical decision support tool. It should not replace professional medical judgment or diagnostic testing. Clinicians are advised to interpret predictions within the context of each patient’s complete medical history and laboratory findings.