**DIABETIES PREDICTOR: An Interactive Tool for Personalized Risk Assessment and Prevention**

GitHub link: - https://github.com/codemasterds/Machine\_Learning\_Projects/tree/main/Diabetes-Prediction-Application

**Introduction**:

Diabetes is a chronic disease that affects millions of people worldwide. It is a metabolic disorder that causes high blood sugar levels due to insufficient insulin production or the body's inability to use insulin effectively. Early detection and management of diabetes are crucial to prevent complications such as blindness, kidney failure, nerve damage, and cardiovascular disease.

Machine learning techniques have shown promising results in detecting and predicting diabetes. In this project, we have developed a machine learning model that predicts whether a person has diabetes or not based on their demographic and clinical features. The model uses a dataset of patients' medical records, which includes variables such as age, gender, BMI, blood pressure, and glucose levels.

The aim of this project is to build a predictive model that can assist healthcare professionals in identifying individuals who are at risk of developing diabetes. The model can also help patients to make lifestyle changes and take preventive measures to manage their condition.

**Dataset Description:**

The dataset used in this project is called the "Pima Indians Diabetes Database," which was originally collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). The dataset contains information on 768 female patients of Pima Indian heritage aged 21 years and above, residing near Phoenix, Arizona, USA. The patients were tested for diabetes based on World Health Organization criteria, and the dataset includes various demographic, diagnostic, and clinical features.

The dataset contains a total of 9 attributes, which are described as follows:

Pregnancies: Number of times pregnant

Glucose: Plasma glucose concentration 2 hours in an oral glucose tolerance test

Blood Pressure: Diastolic blood pressure (mm Hg)

Skin Thickness: Triceps skinfold thickness (mm)

Insulin: 2-Hour serum insulin (mu U/ml)

BMI: Body mass index (weight in kg/ (height in m)^2)

DiabetesPedigreeFunction: Diabetes pedigree function (a function that represents the likelihood of diabetes based on family history)

Age: Age in years

Outcome: Class variable (0 or 1), where 0 represents non-diabetic and 1 represents diabetic

**Data Analysis: -**

The Pima Indians Diabetes Database is a dataset containing various features related to individuals with and without diabetes. The dataset consists of 768 instances with 9 features, out of which 8 are input features and 1 is the target variable indicating the presence or absence of diabetes. Data analysis is crucial for cleaning and pre-processing the dataset before building a predictive model for diabetes.

**Identification of Categorical Variables:**

The first step of data analysis was to identify the categorical variables in the dataset, which were 'pregnancies', 'glucose', 'blood pressure', 'skin thickness', 'insulin', and 'diabetes pedigree function'. These categorical variables were converted into numerical variables using label encoding.

**Handling Missing Values:**

Missing values were handled by replacing them with the mean value of non-missing values for 'glucose', 'blood pressure', 'skin thickness', 'insulin', and 'BMI' features, as these features were approximately normally distributed. For the features that had zero values, they were replaced with the mean value of non-zero values, as these features cannot be zero.

**Outlier Detection and Handling:**

Outliers in 'glucose', 'blood pressure', 'BMI', and 'diabetes pedigree function' features were identified using the z-score method. The outliers were eliminated by replacing them with the mean value of non-outlier values, as these features were approximately normally distributed.

**Data Visualization:**

To better understand the distribution of variables in the dataset, histograms and box plots were used to visualize them. It was found that the 'insulin' and 'diabetes pedigree function' features were positively skewed, so a log transformation was applied to these features to achieve a normal distribution.

**Dynamic Programming:**

Functions were written to dynamically clean and pre-process the data without hardcoding any values or parameters. These functions can handle changes in the dataset and adjust the cleaning and pre-processing steps accordingly.

**Model building and selection: -**

After cleaning and pre-processing the Pima Indians Diabetes Database, several models were built to predict whether an individual has diabetes or not. The models used were XGBoost, KNN, Random Forest, and Decision Tree.

The performance of each model was evaluated using the classification report and confusion matrix, and the results were as follows: -

**Random forest-Model: -**

1. Accuracy Score: 80%

2. F1-Score: 84%

**XGBoost-Model: -**

1. Accuracy Score: 77%

2. F1-Score: 81%

**K Nearest Neighbors-Model: -**

1. Accuracy Score: 76%

2. F1-Score: 82%

**Decision Tree-Model: -**

1. Accuracy Score: 72%

2. F1-Score: 74%

The Random Forest model was found to be the best performing model with an accuracy of 82% and precision of 78%. The Random Forest model had a higher number of true positives and true negatives than the other models and can be deployed for predictions on new data.

**Model Deployment: -**

The Random Forest model was pickled and deployed using Flask on an EC2 instance, providing access to everyone to make predictions on new data. The Flask app was simple and easy to use, making it accessible to a wide audience.

**Diabetes predictor Application Pipeline: -**

 Icon

Description automatically generated Logo

Description automatically generated Logo

Description automatically generated with low confidence

Data set data preprocessing, flask app integration Model deployment on EC2

model buliding server