Diabetes Prediction Dataset

The **Diabetes prediction dataset** is a collection of medical and demographic data from patients, along with their diabetes status (positive or negative). The data includes features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. This dataset can be used to build machine learning models to predict diabetes in patients based on their medical history and demographic information. This can be useful for healthcare professionals in identifying patients who may be at risk of developing diabetes and in developing personalized treatment plans. Additionally, the dataset can be used by researchers to explore the relationships between various medical and demographic factors and the likelihood of developing diabetes.

About this file

The **diabetes_prediction_dataset.csv file** contains medical and demographic data of patients along with their diabetes status, whether positive or negative. It consists of various features such as age, gender, body mass index (BMI), hypertension, heart disease, smoking history, HbA1c level, and blood glucose level. The Dataset can be utilized to construct machine learning models that can predict the likelihood of diabetes in patients based on their medical history and demographic details.

- **Biological Sex (Gender)**: Gender refers to the biological sex of the individual, which can have an impact on their susceptibility to diabetes. This column indicates the gender of the individual, typically denoted as male or female.
- **Age**: Age is an important factor as diabetes is more commonly diagnosed in older adults. Age ranges from 0 to 80 in our dataset.
- **Hypertension**: Hypertension is a medical condition in which the blood pressure in the arteries is persistently elevated. This column indicates whether the individual has hypertension, with values of 0 indicating absence and 1 indicating presence.
- **Heart Disease**: Heart disease is another medical condition that is associated with an increased risk of developing diabetes. This column indicates whether the individual has heart disease, with values of 0 indicating absence and 1 indicating presence.
- **Smoking History**: Smoking history is also considered a risk factor for diabetes and can exacerbate the complications associated. This column indicates the individual's smoking history, typically denoted as non-smoker or smoker.
- **BMI (Body Mass Index)**: BMI is a measure of body fat based on weight and height. Higher BMI values are linked to a higher risk of diabetes.
- **HbA1c (Haemoglobin A1c) Level**: HbA1c level is a measure of a person's average blood sugar level over the past 2-3 months. Higher levels indicate poorer blood sugar control and are associated with a higher risk of diabetes.
- **Blood Glucose Level**: Blood glucose level refers to the amount of glucose in the bloodstream at a given time. High blood glucose levels are a key indicator of diabetes risk.
- **Diabetes**: Diabetes is the target variable being predicted, with values of 1 indicating the presence of diabetes and 0 indicating the absence. This column indicates whether the individual has been diagnosed with diabetes.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a crucial step in understanding the dataset and uncovering patterns, relationships, and anomalies within the data. Here's a suggested EDA for the provided columns:

Summary Statistics:

Calculate basic statistics such as mean, median, standard deviation, minimum, and maximum for numerical columns (Age, BMI, HbA1c level, Blood glucose level).

For categorical columns (Gender, Hypertension, Heart disease, Smoking history, Diabetes), calculate frequency counts and percentages for each category.

Data Visualization:

Histograms and Boxplots: Visualize the distribution of numerical variables (Age, BMI, HbA1c level, Blood glucose level) using histograms and boxplots to identify outliers and understand the spread of the data.

Bar Charts: Plot bar charts for categorical variables (Gender, Hypertension, Heart disease, Smoking history, Diabetes) to visualize the distribution of categories.

Scatter Plots: Explore relationships between numerical variables (e.g., Age vs. BMI) using scatter plots to identify potential correlations or patterns.

Heatmap: Create a heatmap to visualize the correlation matrix between numerical variables. This helps identify correlated features and potential multicollinearity issues.

Feature Relationships:

Explore relationships between different features and the target variable (Diabetes) using visualizations such as stacked bar charts or grouped boxplots to understand how each feature may be related to the presence or absence of diabetes.

Group Analysis:

Perform group analysis to compare statistics or distributions between different groups (e.g., gender, presence of hypertension, presence of heart disease) and their impact on diabetes risk.

Link to Dataset: **https://drive.google.com/file/d/1INcXfo2qK3qQeiewOccs5sEJ3HluyVGh/view?usp=sharing**