Diabetes Dataset

1. Dataset Overview

Diabetes is one of the diseases with no proper treatment and with proceeding stages it has no cure. Thereby, making it a cause of slow death among many age groups. Therefore having a chance to detect diabetes in its early stages with useful information of a patient with blood profiles can provide some extra years to the subject under observation. The following dataset that is being used for analysis here is collected from kaggle[1]. The dataset possesses 9 columns and 768 rows. The 9 column represents the features of the dataset and the output classes are represented as **yes** and **no** values.
The 9 features of the dataset are as follows:

1. Pregnancies
2. Glucose
3. Blood Pressure
4. Skin Thickness
5. Insulin
6. BMI
7. Diabetes Pedigree Function
8. Age
9. Outcome

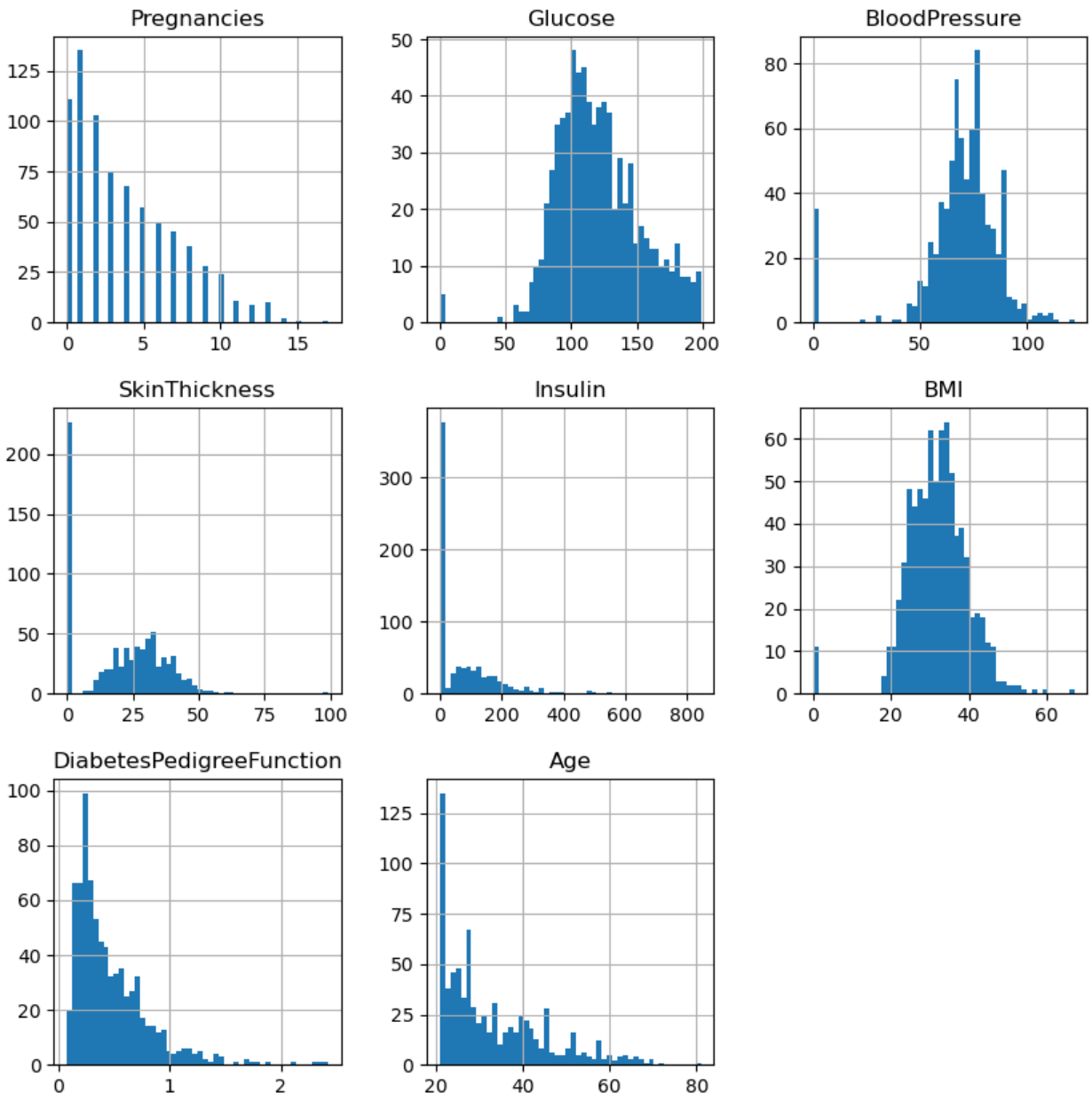
The ninth feature of the dataset is output class which are:

1. Yes represented by 1
2. No represented by 0

The “yes” class signifies all the patients who have diabetes whereas the “no” class signifies that there is no trace of diabetes. The dataset is only about the females suffering from diabetes especially how is the pregnancies are affecting women's health.
The dataset comes under the class of binary classification tasks. The dataset is currently maintained by kaggle and the best possible accuracy achieved on this dataset is 100% using a deep neural network with 4 dense layers[2].

2. Dataset analysis and visualization using pandas and matplotlib

For the current project the dataset is analyzed with the pandas library of python. Python provides an easy and efficient way to analyze data via pandas. The read csv command has been used to study the dataset in the form of a data frame. A data frame is generally a tabular representation of the data with every entry in the data frame is indexed and each column is labeled[3]. Similar commands have been used to analyze data statistically. After observing the dataset it is observed that there are 768 rows in total with 9 columns, wherein the ninth column represents the output class. Out of all the 768 entries there are 384 entries belonging to each output class with no missing values. After data analysis the data is visualized using matplotlib library. The visualization of the data helps in seeing how the data looks like. With the help of matplotlib library we can plot each column with numerical entries to see how the data is varying from one entry to another[4].



References

[1] https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/code

[2] https://www.kaggle.com/code/miladaghalari/diabetes

[3] https://pandas.pydata.org/docs/reference/api/pandas.read_csv.html

[4] https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.hist.html