**Clinical Information System 473:**

**DIABETICS PROJECT:**

**Diabetes-Prediction:**

Data mining project to detect if a person is diabetic using logistic regression in R.

**Dataset Description:**

* All patients here are females at least 21 years old of Pima Indian heritage.
* Number of Instances: 768
* Number of Attributes: eight plus class

**Attributes:**

1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/ (height in m) ^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1)

Class Distribution: (class value 1 is interpreted as "evaluated positive for diabetes")

|  |  |
| --- | --- |
| **Class Value** | **Number of instances** |
| 0 | 500 |
| 1 | 268 |
|  |  |

* Mean age of the population is 33 years.
* Average pregnancy is 3.84 ~ 4 times.
* Plasma glucose concentration 2 hours in an oral glucose tolerance test is 120

Here In Data mining project, I have imported data in csv format, which is already present in my system database, next to that I must check dimension of data such that I have import every row of data base. After that I have allocated column name to data.

Then I checked for null values in dataset, as this dataset does not have any null values. Hence, we are good to go with data analysis. For this data set I have no null values.

Then carefully analyze summary of data which gives us most information like mean age, mean glucose, etc., also with this summary we can analyze that if it contains any outliers. Outliers are determined by observing the difference between mean and median for that column.

Now we have to set seed so that we will get same results every time when we conduct same results.