**EXPOSYS DATA LABS**

**DATA SCIENCE**

**INTERNSHIP PROJECT REPORT**

**NAME : NIKHIL SONI**

**MOB. : 9792658562**

**ROLL NO. : 190013135054**

**(UNIVERSITY OF LUCKNOW)**

**Project Report on**

**Diabetes Prediction**

**Problem Statement:** Diabetes is one of the deadliest diseases in the world. It is not only a disease but also creator of different kinds of diseases like heart attack, blindness etc. The normal identifying process is that patients need to visit a diagnostic centre, consult their doctor, and sit tight for a day or more to get their reports. So, the objective of this project is to identify whether the patient has diabetes or not based on diagnostic measurements.

**Dataset Used:** The dataset used has been obtained from UCI Machine Learning Repository having 769 records of Female Patients exclusively.

There are 8 independent variables:

1. Pregnancies: No. of times pregnant

2. Glucose: Plasma Glucose Concentration a 2 hour in an oral glucose tolerance test (mg/dl)

3. Blood Pressure: Diastolic Blood Pressure(mmHg) If Diastolic B.P > 90 means High B.P (High Probability of Diabetes) Diastolic B.P < 60 means low B.P (Less Probability of Diabetes)

4. Skin Thickness: Triceps Skin Fold Thickness (mm) – A value used to estimate body fat.

Normal Triceps Skinfold Thickness in women is 23mm. Higher thickness leads to obesity and chances of diabetes increases.

5. Insulin: 2-Hour Serum Insulin (mu U/ml)

6. BMI: Body Mass Index (weight in kg/ height in m2 ) Body Mass Index of 18.5 to 25 is within the normal range BMI between 25 and 30 then it falls within the overweight range. A BMI of 30 or over falls within the obese range.

7. Diabetes Pedigree Function: It provides information about diabetes history in relatives and genetic relationship of those relatives with patients.

Higher Pedigree Function means patient is more likely to have diabetes.

8. Age (years)

9. Outcome: Class Variable (0 or 1) where ‘0’ denotes patient is not having diabetes and ‘1’ denotes patient having diabete.

**Proposed Methodology:** In the conducted research the purpose is to classify the data available into diabetic or non-diabetic using the supervised learning algorithms. The dataset will be divided into training and testing sets. In order to achieve more accuracy we must train more data. Than we will to a comparative analysis on the results achieved from the algorithms for early detection of diabetes. The models like Support vector machine, logistic regression proves to be most useful in detection of diabetes in a patient. The

center objective of our model is to achieve a better accuracy and overall improvement in early diagnosis of diabetes.

**Algorithms Used:** As we have to classify the data into patients having diabetes or not, the best method which can be used is Classification and Regression Tree Algorithm (CART), because in this, the dataset is divided into training and testing data. Further we can easily classify and predict the outcome using nodes and internodes.

**Software Package Used**: Python-Scikit Learn, NumPy, SciPy, Matplotlib.

**Advantage of this project:** The rules derived will be helpful for doctors to identify patients suffering from diabetes. Further predicting the disease early leads to treating the patient before it becomes critical.

**Conclusion:** A fast and accurate diabetes prediction system is proposed in this paper. The proposed system used 768 instances within 8 attributes for each one of PID dataset. The used data is preprocessed in order to remove the unwanted data, and lead to faster processing time.

Moreover, the dividing technique of the dataset into subset, made an optimal classification result. The proposed system focused on the features analysis and classification parts. The propositions of these parts lead to an optimal achievement. The results of experiments illustrated the effects of using the algorithms of the proposed system through achieving a higher classification rate that the other systems.

Proposed Methodology