FCI MINI PROJECT

Submitted By: Apeksha Chavan
UID: 2017130013
BE COMPS
Professor: Preeti Jani Ma’am

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.

Diabetes is a disease that occurs when your blood glucose level, also called as blood sugar, is too high.
Insulin, a hormone made by the pancreas, helps glucose from food get into your cells to be used for energy. Overtime, having too much glucose in your blood can cause health problems.

Diabetes contribute to high blood pressure and is linked with high cholesterol which significantly increases the risk of heart attacks, strokes and other cardio vascular diseases.
The kidneys are another organ that is at particular risk of damage as a result of diabetes and the risk is again
increased by poorly controlled diabetes, high blood pressure and high cholesterol. If diabetes has caused nerve damage,
then this can lead to nausea, constipation or diorrhoea. Diabetes affect on the skin is usually a result of it's affect on the nerves
and circulation which can lead to dry skin; slow healing of cuts, burns and wounds,
fungal and bacterial infection and the loss of feeling in the foot.

Why I chose this Topic:

Due to increase in diabetes cases worldwide and due to the overgrowing negligence of people towards this issue,

the medical industry is in need for predicting diabetes to control and create awareness among the patients to avoid fatal health issues or even death in the years to come. With a proper data set and a trained machine learning model, doctors can predict whether a person is likely to have diabetes or long in the long run which will help them to take proper preliminary actions to avoid the worst.
So, the objective of this project is to identify whether the patient has diabetes or not based on diagnostic measurements.

Dataset Used:

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases”. In particular, the dataset consists of 769 records of **Female Patients** exclusively.

From the domain knowledge, I have analysed and found out the ranges of values and its effects on diabetes for each continuous variable in the dataset. Based upon these ranges we will categorize the continuous variables for implementing the decision tree in the next step. Also, we can utilize these ranges to come up with appropriate null value replacement for each independent variable.

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)

Plasma Glucose Test	Normal	Prediabetes	Diabetes
2 hour post-prandial	Below 7.8 mmol/l Below 140 mg/dl	7.8 to 11.0 mmol/l 140 to 199 mg/dl	11.1 mmol/l or more 200 mg/dl or more

A) 2-hour value between 140 and 200 mg/dL (7.8 and 11.1 mmol/L) is called impaired glucose tolerance. This is called "pre-diabetes." It means you are at increased risk of developing diabetes over time. A glucose level of 200 mg/dL (11.1 mmol/L) or higher is used to diagnose diabetes.

3. Blood Pressure: Diastolic Blood Pressure(mmHg)
If Diastolic B.P > 75 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)

	Normal Insulin Level
2 Hours After Glucose	16-166 mIU/L

Values above this range can be alarming.

6. BMI: Body Mass Index (weight in kg/ height in m^2) 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. Obesity: It provides information about those patients prone to over-weight. Higher weight/ obesity means patient is more likely to have diabetes.

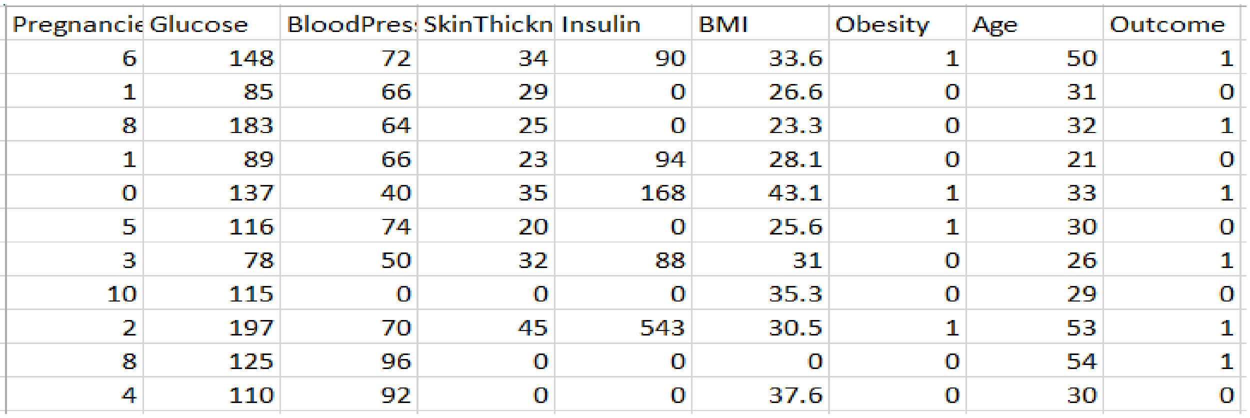
8. Age (years)

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

The **dependent variable** is whether the patient is having diabetes or not.

Data Cleaning will take place as data has got lot of missing values. Handling missing values can be done either by replacing null values with mode or mean or replacing the null value with a random variable.

Sample Data



Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	Obesity	Age	Outcome
6	148	72	34	90	33.6	1	50	1
1	85	66	29	0	26.6	0	31	0
8	183	64	25	0	23.3	0	32	1
1	89	66	23	94	28.1	0	21	0
0	137	40	35	168	43.1	1	33	1
5	116	74	20	0	25.6	1	30	0
3	78	50	32	88	31	0	26	1
10	115	0	0	0	35.3	0	29	0
2	197	70	45	543	30.5	1	53	1
8	125	96	0	0	0	0	54	1
4	110	92	0	0	37.6	0	30	0

Algorithms Used: As we have to classify the data into patients having diabetes or not, the best method which can be used is Random Forest Classifer, 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.

Results:

Glucose level, BMI, pregnancies and Age have significant influence on the m odel, specially glucose level and BMI.
higher blood pressure is correlated with a person being highly diabetic.

Although age was more correlated than BMI to the output variables (as we saw during data exploration), the model relies more on BMI.

The performance metrics used in the evaluation are:
Accuracy Score: proportion of correct predictions out of the whole dataset.

In this project, I have used Random Forest classifier as it gives the highest accuracy out of all
algorithms, the Random Forest Classifer model has achieved an accuracy score of 99% in test data,
i.e. out of all diabetic patients, 99% of them will be correctly classified using medical diagnostic measurements.