INTRODUCTION:-

In this study most favourable and common 21 variables are correlated to predict and anticipate diabetes. The data referred for the prediction is collected from Kaggle since it is authentic and realistic data obtained from CDC. Several groups are created based whether they are carriers of diabetes or not and their habits. This study design is typically selected to have a dual benefit. One would be to get actual scenario for individuals and second to attain real base data for my educational purpose. This study will prove my technical strengths which can be explored to assessor. The disease (diabetes) is selected for this study because this disease is increasing rapidly across human age. The study will definitely help to individual as well as medicinal practitioner to avoid most favourable variables so that this disease can be prevented well in advance.

Diabetes is a disease that occurs when your blood glucose, also called blood sugar, is too high. Glucose is your body’s main source of energy. Your body can make glucose, but glucose also comes from the food you eat. Insulin is a hormone made by the pancreas that helps glucose get into your cells to be used for energy. If you have diabetes, your body doesn’t make enough—or any—insulin, or doesn’t use insulin properly. Glucose then stays in your blood and doesn’t reach your cells.

Diabetes raises the risk for damage to the eyes, kidneys, nerves, and heart. Diabetes is also linked to some types of cancer. Taking steps to prevent or manage diabetes may lower your risk of developing diabetes health problems. **If you have a mother, father, sister, or brother with diabetes, you are more likely to get diabetes yourself. You are also more likely to have prediabetes.**

The most common types of diabetes are type 1, type 2, and gestational diabetes.

Chronic diabetes conditions include type 1 diabetes and type 2 diabetes. Potentially reversible diabetes conditions include prediabetes and gestational diabetes.

Type 1 diabetes:-If you have type 1 diabetes, your body makes little or no insulin. Your immune system attacks and destroys the cells in your pancreas that make insulin. [Type 1 diabetes](https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes/type-1-diabetes) occurs when your immune system, the body’s system for fighting infection, attacks and destroys the [insulin](https://www.niddk.nih.gov/Dictionary/I/insulin)-producing beta cells of the [pancreas](https://www.niddk.nih.gov/Dictionary/P/pancreas). Scientists think type 1 diabetes is caused by [genes](https://www.niddk.nih.gov/Dictionary/G/gene) and environmental factors, such as viruses, that might trigger the disease. Type 1 diabetes is usually diagnosed in children and young adults, although it can appear at any age. People with type 1 diabetes need to take insulin every day to stay alive.

### Type 2 diabetes:- If you have type 2 diabetes, the cells in your body don’t use insulin properly. The pancreas may be making insulin but is not making enough insulin to keep your blood glucose level in the normal range. Type 2 diabetes is the most common type of diabetes. You are more likely to develop type 2 diabetes if you have risk factors, such as overweight or obesity, and a family history of the disease. You can develop type 2 diabetes at any age, even during childhood. You can help delay or prevent type 2 diabetes by knowing the risk factors and taking steps toward a healthier lifestyle, such as losing weight or preventing weight gain.

### Gestational diabetes:- Gestational diabetes is a type of diabetes that develops during pregnancy. Most of the time, this type of diabetes goes away after the baby is born. However, if you’ve had gestational diabetes, you have a higher chance of developing type 2 diabetes later in life. Sometimes diabetes diagnosed during pregnancy is type 2 diabetes.

Some of the symptoms of type 1 diabetes and type 2 diabetes are: Feeling more thirsty than usual, urinating often, losing weight without trying, feeling tired and weak, having blurry vision, having slow-healing sores.

Glucose — a sugar — is a source of energy for the cells that make up muscles and other tissues. Glucose comes from two major sources: food and the liver. Sugar is absorbed into the bloodstream, where it enters cells with the help of insulin. The liver stores and makes glucose. When glucose levels are low, such as when you haven't eaten in a while, the liver breaks down stored glycogen into glucose. This keeps your glucose level within a typical range.

The exact cause of most types of diabetes is unknown. In all cases, sugar builds up in the bloodstream. Race or ethnicity also may raise your risk of developing type 2 diabetes. Sometimes family members of people with type 1 diabetes are tested for the presence of diabetes immune system cells.

No cure for diabetes currently exists, but the disease can go into remission. When diabetes goes into remission, it means that the body does not show any signs of diabetes, although the disease is technically still present. Doctors have not come to a final consensus on what exactly constitutes remission, but they all include A1C levels below 6.5% as a significant factor. A1C levels indicate a person’s blood sugar levels over 3 months.

India has around 101 million people living with diabetes and another 136 million people in pre-diabetes stages, found a recently published study by the Madras Diabetes Research Foundation and Indian Council of Medical Research.

The Diabetes Health Indicators Dataset contains healthcare statistics and lifestyle survey information about people in general along with their diagnosis of diabetes. The features consist of demographic information (eg. sex, age), presence of comorbid conditions (eg. high blood pressure, high cholesterol), lifestyle factors (eg. smoker, alcohol consumption) and answers to survey questions for each patient. The target variable for classification is whether a patient has diabetes or not.

**Objective:** The objective of this project is to build a model that best classify whether the parameters contribute or trigger diabetes.

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Type | Description | Values |
| Diabetes\_binary | Binary | Diagnosed with diabetes or not | 0= no diabetes  1= diabetes or pre diabetes. |
| HighBP | Binary | Diagnosed with high BP or not | 0= no high BP  1=high BP |
| HighChol | Binary | Diagnosed with cholesterol or not | 0= no high cholesterol  1= high cholesterol |
| CholCheck | Binary | Did a Cholesterol check in last 5 years | 0= not checked  1= checked |
| BMI | Continuous | Body Mass Index | Body Mass Index |
| Smoker | Binary | Smoked at least 100 cigarettes in your entire life | 0= no  1= yes |
| Stroke | Binary | Ever had a stroke | 0= no  1=yes |
| HeartDiseaseorAttack | Binary | Coronary heart disease or Myocardial Infarction. | 0= no  1= yes |
| PhysActivity | Binary | Physical Activity in last 30 days | 0= no  1= yes |
| Fruits | Binary | Consumption of at least 1 fruit a day | 0= no  1= yes |
| Veggies | Binary | Consumption of vegetable 1 or more than once per day | 0= no  1= yes |
| HvyAlcoholConsump | Binary | Heavy drinkers (14>for Men and 7> for Women) | 0= no  1= yes |
| AnyHealthcare | Binary | Any kind of health care coverage. | 0= no  1= yes |
| NoDocbcCost | Binary | Was there any time when you wanted to see a doctor but couldn’t because of cost | 0= no  1= yes |
| Genhlth | Ordinal | What would you say is your general health (scale 1-5) | 1= excellent  2=very good  3= good  4= fair  5= poor |
| MentHlth | Continuous | How many days during the past 30 days was your mental health not good (this includes depression, anxiety, stress etc.) | 1-30 days |
| PhysHlth | Continuous | How many days during the past 30 days was your physical health not good (this includes illness, injury etc.) | 1-30 days |
| DiffWalk | Binary | Do you have any difficulty climbing or walking on stairs. | 0= no  1= yes |
| Sex | Binary |  | 0= female  1= male |
| Age | Ordinal | 13 level age categories | 1= 18-24  9= 60-64  13= 80 or older |
| Education | Ordinal | Education level (scale 1-6) | 1= never attended school.  6= graduation of 4 years or more. |
| Income | Ordinal | Income scale (scale 1-8) | 1=less than $10,000  5=less than $35,000  8=less than $75,000 or more |

Interpretations:-

In python I developed certain plots. Plots are easier to understand and give an overview of the data. I have developed bar graphs of all the binary data. This includes High BP, High Cholesterol, Chol Check, Smoker, Stroke, Heart Disease or Attack, Physical Activity, Fruits, Veggies, Heavy Alcohol Consumption, Any Health care, No Doctor because of cost, Difficulty Walking, Sex.

When these graphs are plotted you can see that in the graphs with high BP, high cholesterol tend to have high count of diabetic individuals then non-diabetic. This means we assume that these features might be significantly contributing or triggering diabetes.

But the other features like cholesterol check, smoker, stroke, heart disease or attack, physical activity, fruit, veggies, heavy alcohol consumption, health care, no doctor because of cost, difficulty walking, sex don’t show a great difference between the diabetic and non-diabetic individuals. This means we can assume that they don’t contribute or trigger diabetes. We will future perform models and make sure that this is the case. We drop those which prove to be insignificant and develop a model based on the features that are significant.

For all the features which were ordinal type data I developed histograms. This includes general health, age, education and income. You can see that as the general health goes on deteriorating the number of diabetic individuals go high. In the case of age as the age advances the diabetic individuals goes on increasing. In case of education and income the diabetic and non-diabetics are more or less same in al the classes. So, we may assume that they don’t influence diabetes. Although we will confirm this in the later models.

For all the features which are continuous type data I have developed histograms and box plots. We can see that as the BMI increases the number of diabetic individuals increase. Where as in physical health and mental health in both cases we can say the diabetic are more compared to the non-diabetic individuals.

Finally, through a heat map you can see the correlation matrix. It shows the correlation of the diabetic individuals with all the features. This heat map makes it easier to figure out which features may be significant in tigering diabetes. We can see that high BP, high cholesterol, BMI, heart disease or attack, general health, difficulty walking, age show higher correlation than some other groups. We will confirm our predictions in the further models to get accurate results.

You can check the python codes for cleaning the data, balancing the data, and plotting the graphs [here.](../../Users/samiksha/PycharmProjects/pythonProjectIMP/exportToHTML/main.py.html)

You can check the R codes for Binary logistic regression [here.](R%20for%20diabetes.R)