PHASE 2:
Project=[Diabetes Prediction system]

NAME- MD ALTAMASH
YEAR:- 3RD
DEPT:- COMPUTER SCIENCE & ENGINEERING(C.S.E)

TABLE OF CONTENTS:-

1.1:- INTRODUCTION:

Diabetes mellitus is characterized by abnormally high levels of sugar (glucose) in the blood.

When the amount of glucose in the blood increases, e.g., after a meal, it triggers the release of the hormone insulin from the pancreas. Insulin stimulates muscle and fat cells to remove glucose from the blood and stimulates the liver to metabolize glucose, causing the blood sugar level to decrease to normal levels.

1.2:- DATA PREPARATION:

As a Data Scientist, the most tedious task which we encounter is the acquiring and the preparation of a data set. Even though there is an abundance of data in this era, it is still hard to find a suitable data set that suits the problem you are trying to tackle. If there aren't any suitable data sets to be found, you might have to create your own.

2.:- DATA EXPLORATION:

When encountered with a data set, first we should analyze and "**get to know**" the data set. This step is necessary to familiarize with the data, to gain some understanding of the potential features and to see if data cleaning is needed.

First, we will import the necessary libraries and import our data set to the Jupyter notebook. We can observe the mentioned columns in the data set.

```
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as snsdiabetes = pd.read_csv('datasets/diabetes.csv')
diabetes.columns
```

3.:- Data Cleaning

The next phase of the machine learning work flow is data cleaning. Considered to be one of the crucial steps of the workflow, because it can make or break the model. There is a saying in machine learning **"Better data beats fancier algorithms"**, which suggests better data gives you better resulting models.

There are several factors to consider in the data cleaning process.

1. Duplicate or irrelevant observations.
2. Bad labeling of data, same category occurring multiple times.
3. Missing or null data points.
4. Unexpected outliers.

4.:-Missing or Null Data points

We can find any missing or null data points of the data set (if there is any) using the following pandas function.

```
diabetes.isnull().sum()
diabetes.isna().sum()
```

5.:- Unexpected Outliers

When analyzing the histogram we can identify that there are some outliers in some columns. We will further analyze those outliers and determine what we can do about them.

6.:- DATA PREPROCESSING:

n this paper, propose an approach on dataset preprocessing, which is applied to diabetes prediction. The preprocessing approach consists of the following process: missing value process, imbalanced data process, feature importance process, and data augmentation process.