Diabetes Data Set

**Problem statement**:  
Analyse the dataset and predict whether person will have Diabetes or not by developing a machine learning model.

It contains information about 768 women.

Minimum age 21 years

**have one target variable:**

**Outcome**:**1** (if having diabetes) and **0** (if not)

**We have 8 explanatory variables(all numeric) which are:**

1. **Pregnancies** : Number of times pregnant.
2. **Glucose**:♣ Oral glucose tolerance test — OGTT (two hour plasma glucose concentration after 75g anhydrous glucose in mg/dl)
3. **BloodPressure**: Blood Pressure (Diastolic Blood Pressure in mmHg)
4. **SkinThickness** : Triceps skin fold thickness (in mm)
5. **Insulin**: 2 h serum insulin in mu U/ml
6. **BMI** : Body Mass Index in kg/m2
7. **DiabetesPedigreeFunction** : Function that represents how likely they are to get the disease by extrapolating from their ancestor’s history.
8. **Age** : Age in years.

**Problem Statement**

1. Check the data for missing values, and replaced the missing values with appropriate measures. (Give Reason)
2. Identify the columns which are useful to you.
3. Get rid of unwanted columns of data.
4. Remove duplicates values from the columns
5. Convert any special or unwanted characters into numerical or meaningful data.
6. find the **summary** of the dataset, **histogram**, to see the distribution of each variable, **boxplots**, to check for outliers, **line plots, count plots, bar plots.**
7. Check is their strong correlation between the features.
8. Plotted**pairplots,** to see relationship between each pair of variables, **jointplots**, to see univariate and bivariate relationship simultaneously.
9. Make different model and check the accuracy wither other models and find the best classification model.

DATA NAM: “diadetes.csv”