**CAPSTONE HEATHCARE PROJECT**

**Context:**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

**Approach:**

Following pointers will be helpful to structure your findings.

1. Perform descriptive analysis. It is very important to understand the variables and corresponding values. We need to think through - Can minimum value of below listed columns be zero (0)? On these columns, a value of zero does not make sense and thus indicates missing value.

* Glucose
* BloodPressure
* SkinThickness
* Insulin
* BMI

How will you treat these values?

1. Visually explore these variables, you may need to look for the distribution of these variables using histograms. Treat the missing values accordingly.
2. We observe integer as well as float data-type of variables in this dataset. Create a count (frequency) plot describing the data types and the count of variables.

**Week 1 Results are as follows –**

* There are 768 rows and 9 columns in this dataset
* The dataset have nine attributes in which there are eight independent variables (Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunction,Age) and one dependent variable (Outcome).
* BMI and DiabetesPedigreeFunction are a float data type and rest of the variables are integer data type.
* The Variables have a lot of zero values which can be represented as missing values,
* The missing values '0' is replaced by mean to explore the dataset.
* The Outcome variable shows that there are 500 non-diabetic people and 268 diabetic people.
* It means that 65.1% people are diabetic and 34.9% people are non-diabetic.
* The parameters Glucose, BloodPressure, BMI are normally distributed.
* Pregnancies,Insulin,Age,DiabetesPedigreeFunction are rightly skewed.
* BloodPressure,SkinThickness,Insulin,BMI have outliers.

*The code for the results are here -*































