**Logistic regression** is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

# The project

This project predict whether a person has diabetes or not using Machine Learning, based on information about the patient such as blood pressure, body mass index (BMI), age, etc.

## Overview

The data was collected and made available by “National Institute of Diabetes and Digestive and Kidney Diseases” (I downloaded this data through [kaggle.com](http://kaggle.com/)) as part of the Pima Indians Diabetes Database Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here belong to the Pima Indian heritage (subgroup of Native Americans), and are females of ages 21 and above.

I used Python and some of its popular data science related packages. First of all, I import pandas to read the data from a CSV file and manipulate it for further use. l also use numpy to convert out data into a format suitable to feed our classification model. Used seaborn and matplotlib for visualizations, then imported Logistic Regression algorithm from sklearn. This algorithm will help build a classification model. Lastly, I will used joblib available in sklearn to save our model for future use.

# Types of Questions Logistic Regression Can Answer

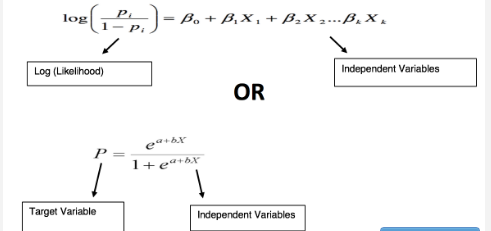
How does the probability of getting lung cancer (yes vs. no) change for every additional pound a person is overweight and for every pack of cigarettes smoked per day?

Do body weight, calorie intake, fat intake, and age have an influence on the probability of having a heart attack (yes vs. no)?

In this project I was able to predict using independent variables such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function and Age to determine if a patient is diabetic or not.  
The dependent variable should be dichotomous in nature (binary)

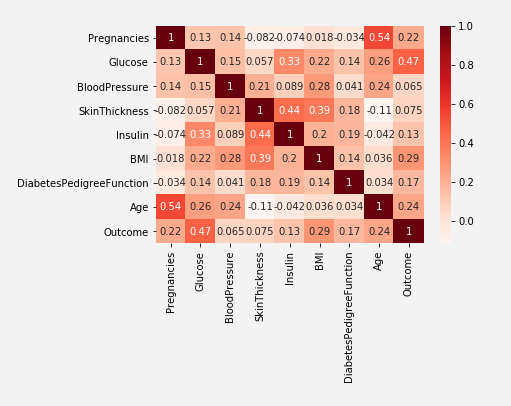
Mathematically, logistic regression estimates a multiple linear regression function defined as:

**Logistic** **Regression Equation**



# Data Exploration

By finding correlation of every pair of features (and the outcome variable), and visualize the correlations using a heatmap. Brighter colors indicate more correlation. Aside the colours I included the values for each pair correlation. From the table below , glucose levels, age, BMI and number of pregnancies all have significant correlation with the outcome variable, this can be seen from their values and colours on the heatmap. Also the correlation between pairs of features, like blood pressure, skin thickness, insulin have light colours and values almost at zero (0) which implies that they have low significance



# Dataset Preparation

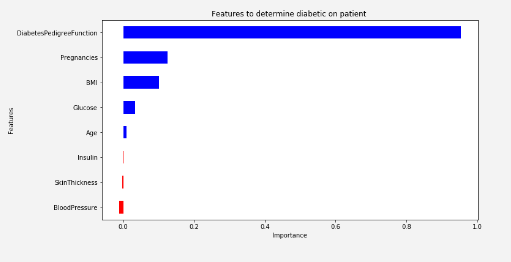
This stage is spit the dataset into 3 parts, Train, Test and check

The data set consists of record of 767 patients in total. To train our model I used 650 records, 100 records for testing, and the last 17 records to cross check our model.

# Training and Evaluating Machine Learning Model

I used a machine learning model called logistic regression\_. Since the model is readily available in sklearn. First, I created an instance called data\_check and then use the fit function to train the model. Next, I used my test data to find out accuracy of the model, have gave 78.0%

I also decided to interprete the model by ploting a horizintal bar chart to show which features really influenced diabetic and those that didn’t.



From the Image above like the heat map, it shows that Insulin, skin thickness and blood pressure least influence the chances of a patient having Daibetic. But pregnancies, glucose, BMI, DiabetesPedigreeFunction, age, outcome have higher influence on having diabetics

# Making Predictions with the model

I used the unused data in check\_data and predicted the outcome of the first ten records, i was able to obtain amazing result from the model.  
Patient test Positive from data collected: 1

Patient test Negative from data collected: 0

Patient test Negative from data collected: 0

Patient test Positive from data collected: 1

Patient test Positive from data collected: 1

Patient test Negative from data collected: 0

Patient test Negative from data collected: 0

Patient test Negative from data collected: 0

Patient test Negative from data collected: 0

Patient test Positive from data collected: 1

Patient test Negative from data collected: 0