What is classification

In terms of data science , any task that has the capability to distinguish the data into different classes is called a classification Method.

What is the need

Human brain tends to classify anything depending upon various conditions.
For example –
- Will you pass or fail in an examination?
    - You judge this basing upon various factors like your preparation level , depth in understanding, the lectures attended, the assignments done against the course etc.
- Which movie to watch today ?
    - You decide this basing on various factors like the storyline, the action , the drama, the actors , the directors , the cinematography, the hype against the movie , ImDB rating etc.

Though you won't know consciously but you are in a way classifying a situation or taking decision after you have done the classification .

Why do you need an algorithm?

Can you quantify your decision !
Can you process many number of factors !

One of the methods for classification is logistic regression. This is a very widely used method which basing on certain factors tells you what is the chance that a case falls into one category.
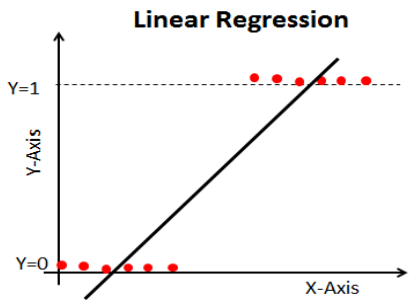
The power of logistic regression can be extended to multiclass classification but for the matter of simplicity we will keep it to binary classification(which means classifying into 2 classes)

How does it function !

Logistic regression is a special case of linear regression. For that we have to know what is **linear regression !**
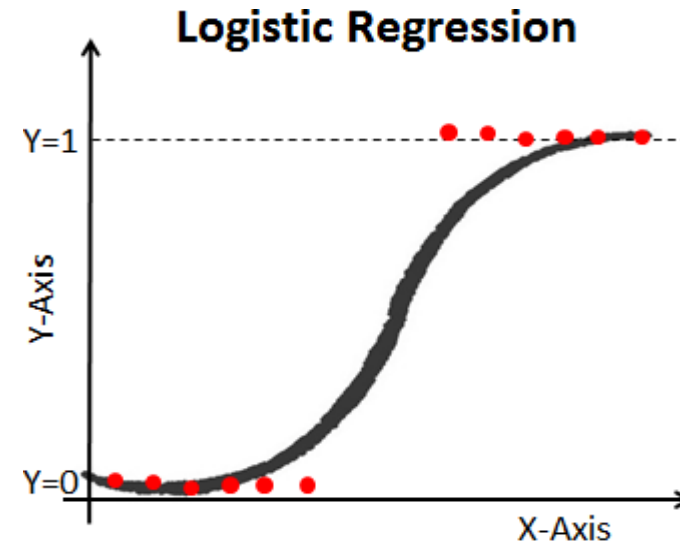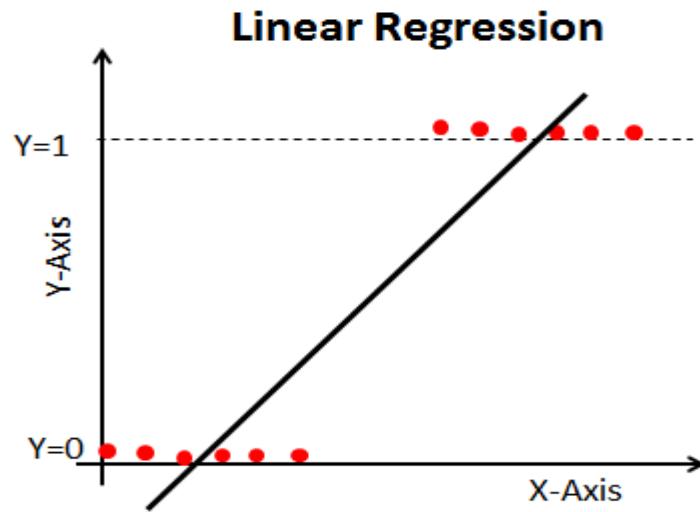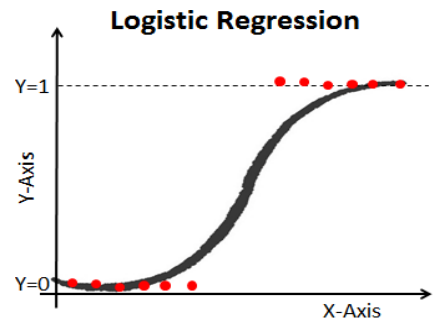

CONTD..

Linear Regression

Analogous to the equation of a straight line(Linear regression)

$$y = \beta0 + \beta1X1 + \beta2X2 + \ldots + \beta nXn$$

Sigmoid function

$$p = 1/1 + e^{-y}$$

Apply sigmoid on linear and convert into logistic

$$p = 1/1 + e^{-(\beta0 + \beta1X1 + \beta2X2 \ldots \beta nXn)}$$

Logistic Regression

$$\log_b \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Let's convert it into plain English :

P is the probability of happening. Here as we are talking about binomial classification, the algorithm will tell you 'yes' or 'no'.

Can anybody tell me what does that equation mean !

What does increase in betas imply !

What will very high value of x tend to move the equation !

https://www.kaggle.com/uciml/pima-indians-diabetes-database

This is a dataset which says depending on certain factors , if a person has diabetes or not.

Let's see how this works.

Now I want to plant this thought in your head. That you are seeing some factors and how they impact.

Can these factors act among themselves and form another factor that probably you can't see .

Human eyes might not find the answer to this.

Let me know what's your thought.