Logistic Regression

1. This is a statistical method to predict a class variable
2. We use a transformation technique to convert a continuous value to probabilities using a logit function (hence the name Logistic)
    a. We have seen the linear regression equation right? Logistic regression equation also looks similar but there is a slight variation in interpretation
    b. Logistic regression equation is $S= m_1X_1 + m_2X_2 + \ldots\ldots + m_nX_n + C$ (just like in case of linear regression) where X's are independent variables, m's are coefficients and C is a constant we need to find which gives S.
    c. S is called log of odds.
        i. Odds= probability of success divided by probability of failure $\frac{P}{1-p}$
        ii. $S=\log(\frac{P}{1-p})$. This is known as log odds
        iii. From simplification of the above equation, we get $p=\frac{1}{1+e^{-S}}$ where $S = m_1X_1 + m_2X_2 + \ldots\ldots + m_nX_n + C$
3. We fix a threshold probability for success, let us say 0.5 for example. Once we get the probability of success for each data point, we check if the obtained probability is greater than 0.5. If yes we classify it to one class or else classify it to other class
4. The decision surface for the logistic regression is a straight line or a plane. On either side of this plane we have data points that belong to different classes
5. The predictions might go wrong as well. These are errors.
    a. Just like in case of linear regression where we use least square method to get the coefficients such that the error in minimum, here we use a method called **maximum likelihood**
    b. As in the case a linear regression, we have a cost function for logistic regression. The best separator line is the one that can generalize well on the unseen data for label classification. The cost function is defined as
    $$-\sum[y_i ln\hat{y}_i + (1-y_i)ln(1-\hat{y}_i)]$$
        i. For a binary classification y_i is either 1 or 0 (success or failure) and p is probability of success
        ii. Our task is to minimize this error. The method we use is gradient descent
    c. A confusion matrix is a structure that gives error report between the actual and the predicted classes
        i. Accuracy- Sum of diagonal elements/ total sum of matrix
        ii. Recall- Out of total actual positives how many are predicted positive
        iii. Precision- Out of total predicted positives how many are actual positive
        iv. F1- Is a combination of both Recall and Precision
        v. Of the two classes of the target variable, the one class that is most important for us will be termed a positive class
        vi. We can also fix the threshold based on ROC curves
            1. A plot between True positive rate (TPR) and false positive rates (FPR) for all probabilities for success from 0 to 1 and at each value TPR and FPR is computed and plotted
            2. The threshold value at which there is a high TPR and reasonable FPR is considered a threshold probability value for the model
            3. This is also used to compare multiple models. The model that has highest area under the curve (AUC value is chosen to be the best model)