# REPORT-FOML

## ASSIGNMENT-4 ES19BTECH11003

5.a) cross-entropy error function:-(t*log(p)+(1-t)*log(1-p))

5.b)
Learning rate=0.1
updated logistic regression model
*Initial values*
θ0=-1
θ1=1.5
θ2=0.5
*After iteration 1:*
θ0=-1.00316626
θ1= 1.50535086
θ2= 0.50196867
Loss=0.5569500797547652

5.c)
At the convergence of gradient descent
Loss after convergence:0.01423
precision:0.8
Recall:0.6666666

6)
My 1st method which has resulted in the best RMSE value is **Random Forest Regression**
RMSE value for test set for 2500000 train samples:3.20202(taken from Kaggle)

My 2nd method which has resulted in the best RMSE value is ***XGBoost Regression***

RMSE value for test set for 2500000 train samples:3.38860(taken from Kaggle)

*Reasons for best RMSE value with* ***Random Forest Regression.***

- Random Forest Regression is a supervised learning algorithm that uses predictions from multiple machine learning algorithms to make a more accurate prediction than a single model.

- It reduces the overfitting problems in decision trees and also **reduces the variance** and therefore **improves the accuracy**.

- Random Forest works well with both **categorical and continuous variables** which is very much an advantage for the given dataset.

- Random Forest can automatically **handle missing values** and is also very less impacted by noise.

- The Random Forest algorithm is very **stable**. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.

*Reasons for better RMSE value with* ***XGBoost Regression***

- XGBoost is an efficient implementation of gradient boosting that can be used for regression predictive modeling.

- It uses the power of parallel processing which is very much useful for our dataset which consists of a large number of rows of data.

- It supports regularization and is designed to handle missing data with its in-build features.

- XGBoost make splits up to the **max_depth** specified and then start pruning the tree backward and remove splits beyond which there is no positive gain.

Even after trying various Regression methods which are Decision tree regression, gradient boost regression, and a few other regression techniques by observing the RMSE values, we can clearly see that as the dataset is furthermore divided from 8 to 14 columns and as the newly added columns being exact integers within a limited range so there will be high importance value for these attributes when compared with latitudes, longitudes which helped in better performance for XGboost and RandomForest Regression.