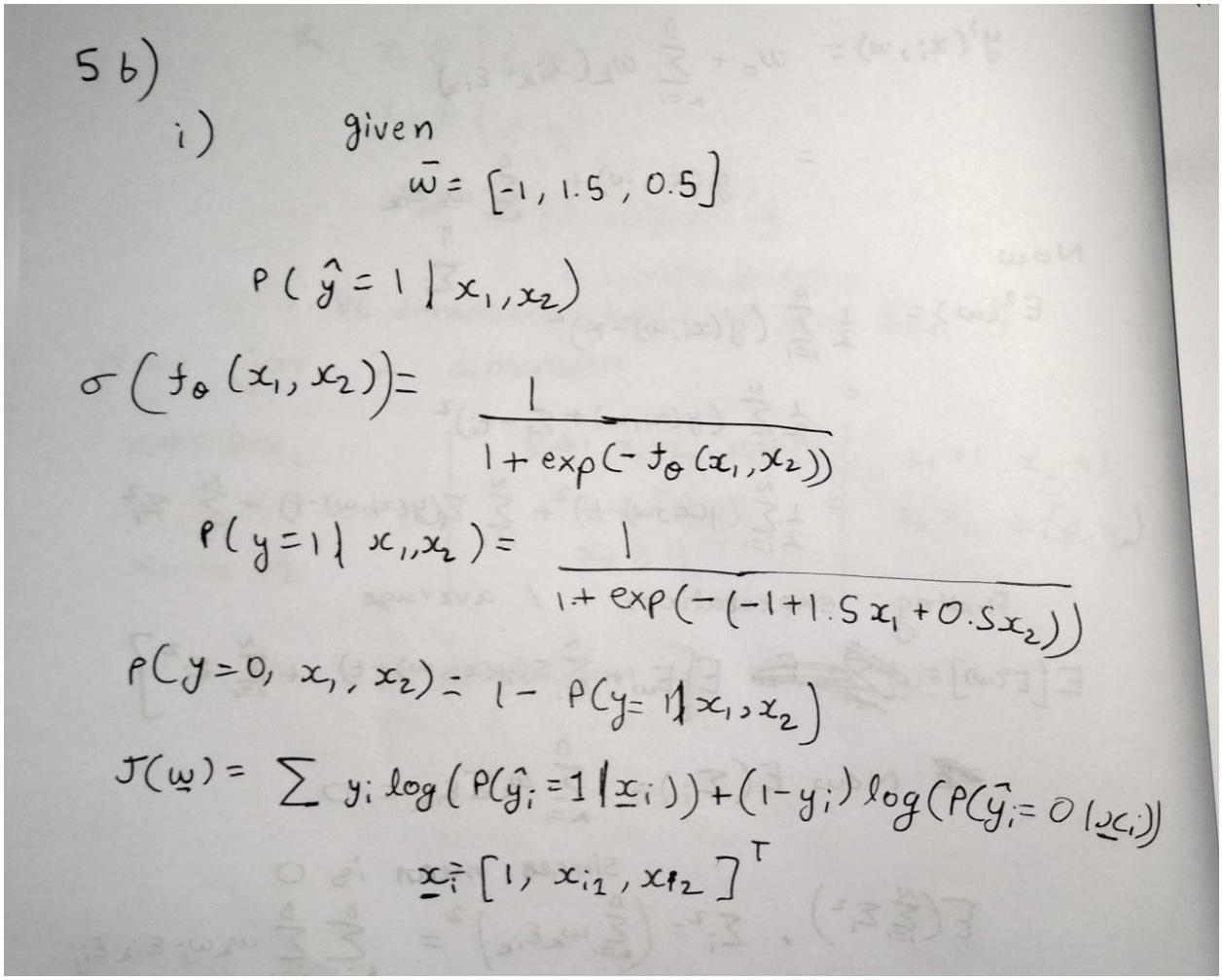
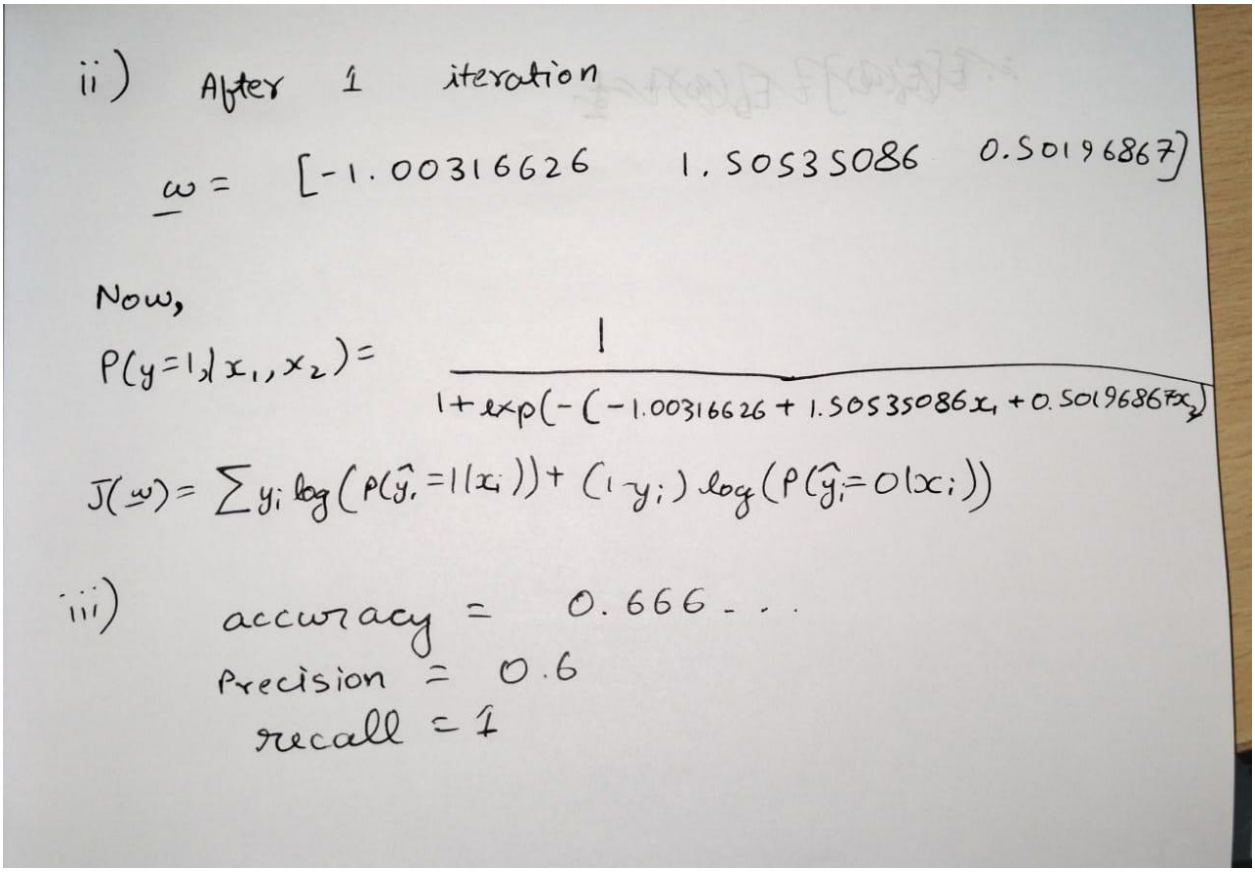
Assignment-4

AI20BTECH11006

Question-5)





Question-6)

Best Accuracies

1) 3.63367
2) 3.92116

Method-1

```
KNeighborsRegressor
```

I used KNeighborsRegressor with n_neighbors = 7, and 30 million data samples for training. I used only 4 columns all being latitude and longitude ones, since more features reduces accuracy possibly because of the curse of dimensionality. I used large number of data points because knn regressor is very fast.

Method-2

```
GradientBoostingRegressor(n_estimators=500, max_depth=6)
```

I used Gradient Boosting Regressor because it is a standard go to ensemble method, I trained it only for 100,000 data points, it gives an accuracy of 3.92116, I used more columns for this Regressor because unlike knn, more information is actually better for for Gradient Boosting Regressor.

Why these models performed better

Knn regressor performed well because it is using 30 million data points for a comparable number of data points it would score far worse. Besides longitude and latitude are rather important factors and would naturally have a higher weight, normalising longitude to be of the same range as latitude didn't change accuracy by much.
Gradient Boosting Regressor performed very slightly better than xgboost, but the difference is rather low and can be closed by hypertuning xgboost. I tried using randomForestRegressor but my kernel died while running it, it was slow compared to other listed models.