Assignment 3 Report

Devansh Agarwal 200316
Utkarsh Aditya 201066
Harshit Singhai 2000434
Antariksh Choudhary 200159
Abhinav Zade 200019
Siddi Akshay 200985

Answers to the Report questions

1.

	Ozone O_3	NO_2	Training Time	Prediction Time
LinearSVR ()	6.83	6.47	2.4s	10.9ms
Logistic Regression(500)	6.41	6.18	$\sim$1min	64ms
Linear Regression()	5.58	6.53	0.2s	7ms

MAE for various Linear Models

We can see from the above table that the method which gives us the best results in terms of MAE (Mean Absolute error) and time taken for prediction is the the **Linear Regression** model.
The loss function used is the Least squares loss and regularization used is l2. This is essentially the same as the **Ridge regression** so we have not mentioned the Ridge regression model explicitly
The MAE we get in this case is 5.63 for O_3 and 6.54 for NO_2

	Ozone O_3	NO_2	Training Time	Prediction Time
LinearSVR ()	6.24	6.48	1.9s	$\sim$10ms
Logistic Regression(max_iter=500)	6.65	6.075	45s	$\sim$80ms
Linear Regression()	5.63	6.54	0.2s	6.17ms

For filling the first table we have used all the 20000 data point as training data but in the second table we have split and shuffled our training data as 15000 data points as train and 5000 data points as test

2. For training and testing of the Non linear models we have split the train.csv dataset into 2 parts , 15000 test points for training and 5000 test points for testing along with shuffling of data points.
If we don't follow the above procedure the model overfits to training data and gives us a negative R2 score value.
Splitting and shuffling the data points as above improves the performance of our models and gives us accurate results.

	Ozone O_3	NO_2	Prediction Time
ElasticNet ()	5.38	6.49	$\sim$0.16s
KNeighborsRegressor	3.13	2.59	$\sim$0.29s
DecisionTreeRegressor	0.052	0.053	$\sim$0.01s
MLPRegressor	5.19	5.46	$\sim$0.09
RandomForestRegressor	1.37	1.05	$\sim$0.79

MAE for various Non Linear Models at train data

For Non linear models, we are getting the best result by Decision Tree when we are training on 20k data points and testing on train set itself. However, when we split the test set into train (15k data points) and test (5k data points), the results change significantly.

	Ozone O_3	NO_2
ElasticNet ()	5.5	6.45
KNeighborsRegressor	3.95	3.43
DecisionTreeRegressor	5.18	3.64
MLPRegressor	5.41	5.71
RandomForestRegressor	3.66	2.87

MAE for various Non Linear Models at train test data split

The Random Forest regression model gives us very good results in terms of MAE but the model size becomes very large (about 320MB) as compared to kNN (about 3MB) so it is not practical to use this method. So, our final approach is kNN approach because it performs well on all paramters (time, size, MAE).

	Ozone O_3	NO_2
KNN (K=5)	3.88	3.32
KNN (K=10)	3.98	3.39
KNN (K=3)	4.05	3.37

MAE for various KNN Models with hyperparameter tuning

The best results which we get for our given data set combining both linear and non linear models comes from the **KNN (K-Nearest Neighbour)** model.
Tuning our K value gives us the best performance at K=5.