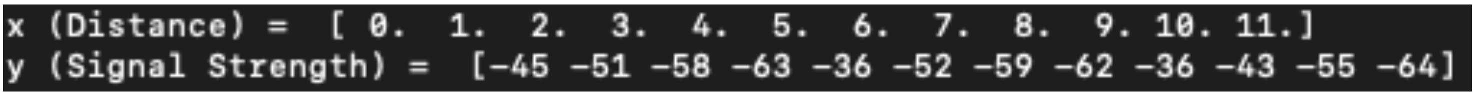
SML Assignment-5

-Harkishan Singh (2017233)

Gaussian Process Regression(GPR)

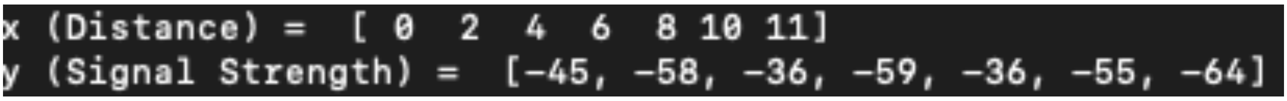
For Q2, we had to implement GPR. I have used “sklearn.gaussian_process” library for that. The data is provided in the question itself.

Original data given to us in the question :



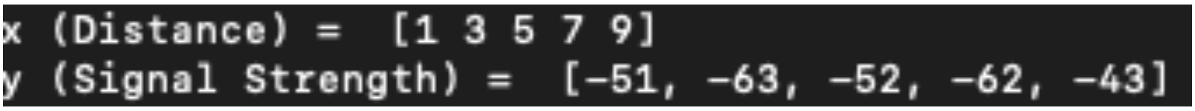
```
x (Distance) =  [ 0.  1.  2.  3.  4.  5.  6.  7.  8.  9. 10. 11.]
y (Signal Strength) =  [-45 -51 -58 -63 -36 -52 -59 -62 -36 -43 -55 -64]
```

Train data :



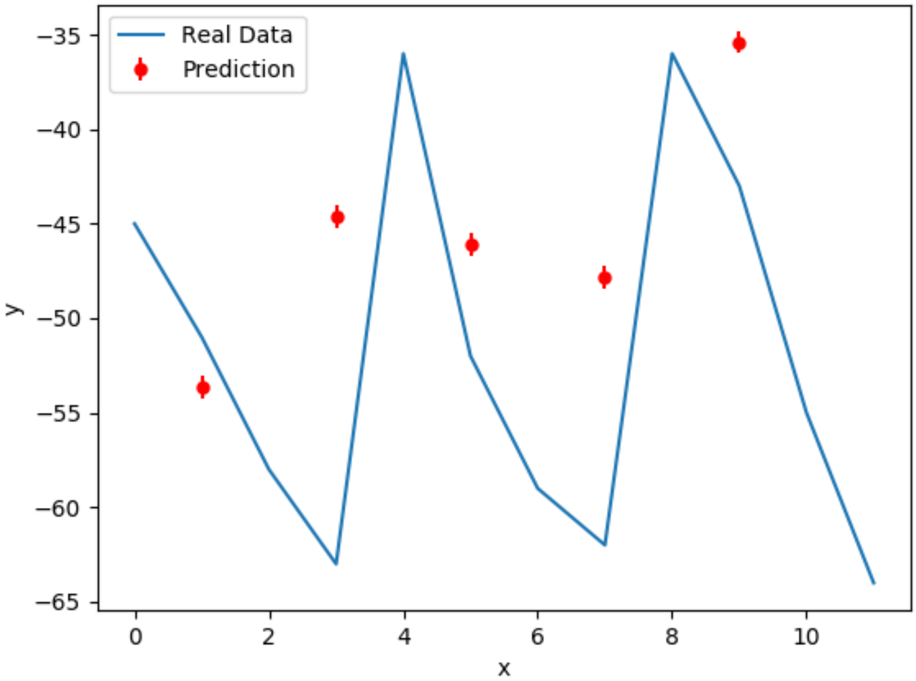
```
x (Distance) =  [ 0  2  4  6  8 10 11]
y (Signal Strength) =  [-45, -58, -36, -59, -36, -55, -64]
```

Test data :



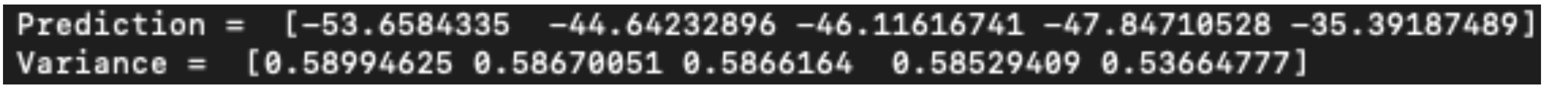
```
x (Distance) =  [1 3 5 7 9]
y (Signal Strength) =  [-51, -63, -52, -62, -43]
```

The graph I am getting by initiating the gaussian process (with default learning parameters) are :



Observations/Note : The red spots are the predicted value of signal strength of the testing data and the line around the red spots corresponds to the variance.

Here are numbers used to built the graph :



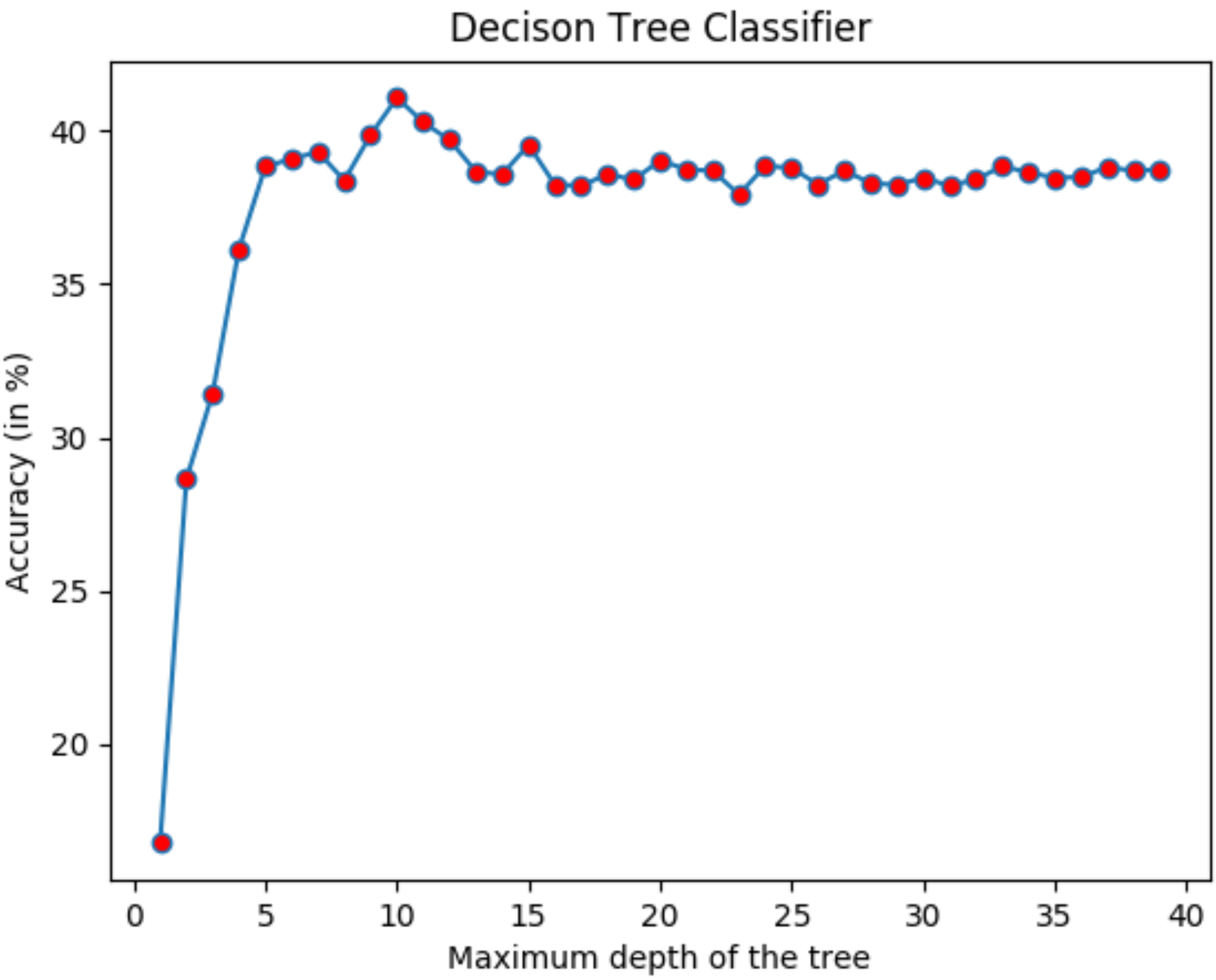
```
Prediction =  [-53.6584335  -44.64232896 -46.11616741 -47.84710528 -35.39187489]
Variance =  [0.58994625 0.58670051 0.5866164  0.58529409 0.53664777]
```

Data Set

That data we have contains PM2.5 quantity of time period : 2010-2014. I have divided the data in two file : “train_data.csv” which contains the training data. Training data is of years 2010 and 2012. Similarly, I have created “test_data.csv” file which contains data of years 2011 and 2013. 2014 data is not used.

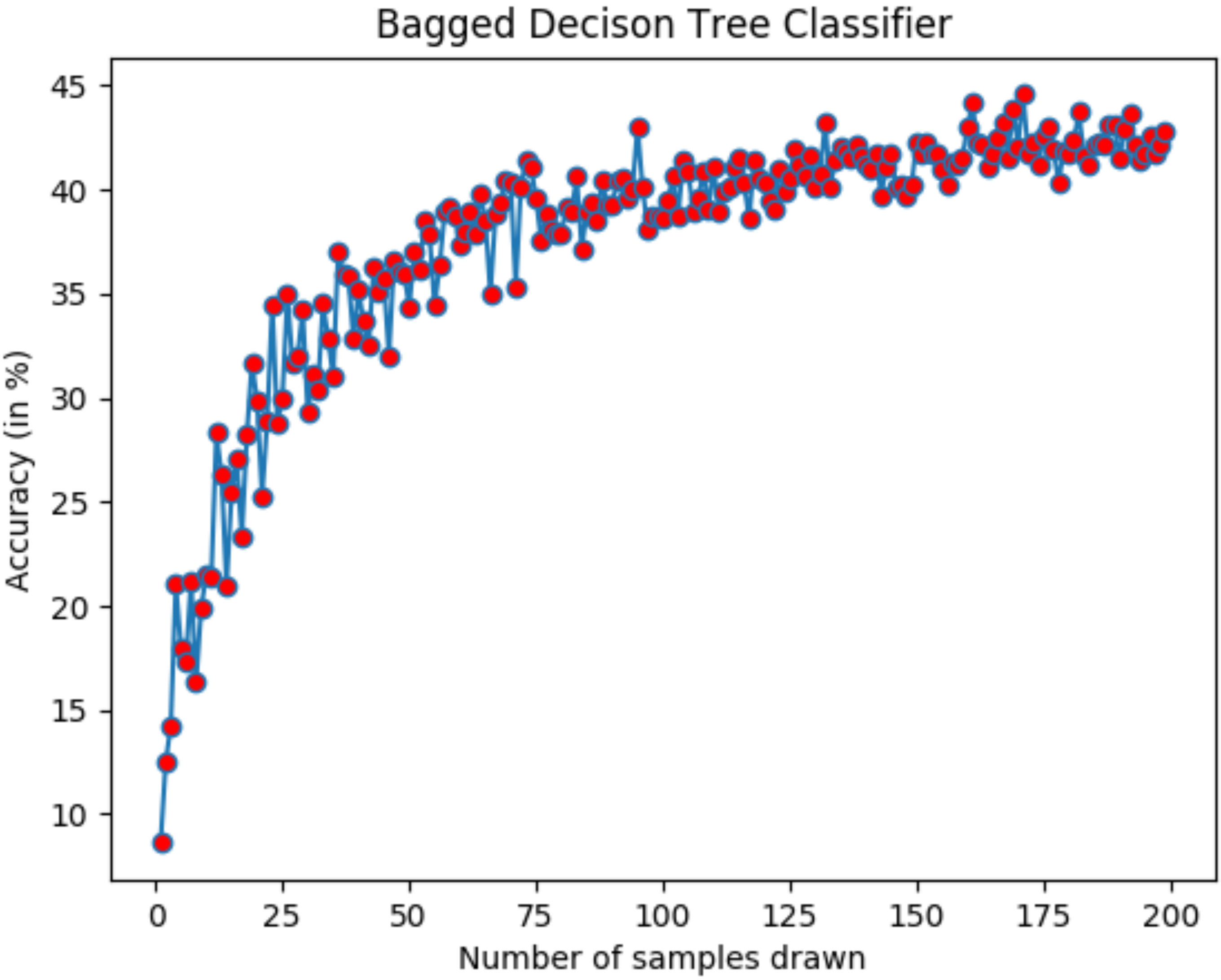
Classifier

In this part, we have to predict the month using the other information available.
Analysis of the prediction by the decision tree classifier of various depth is given below :



We can see that after the maximum depth of tree has reached 15 - 17, we can't see any marginal improvement in the accuracy. Also, a point t note here is that the accuracy have never jumps over 45%, which is bad.

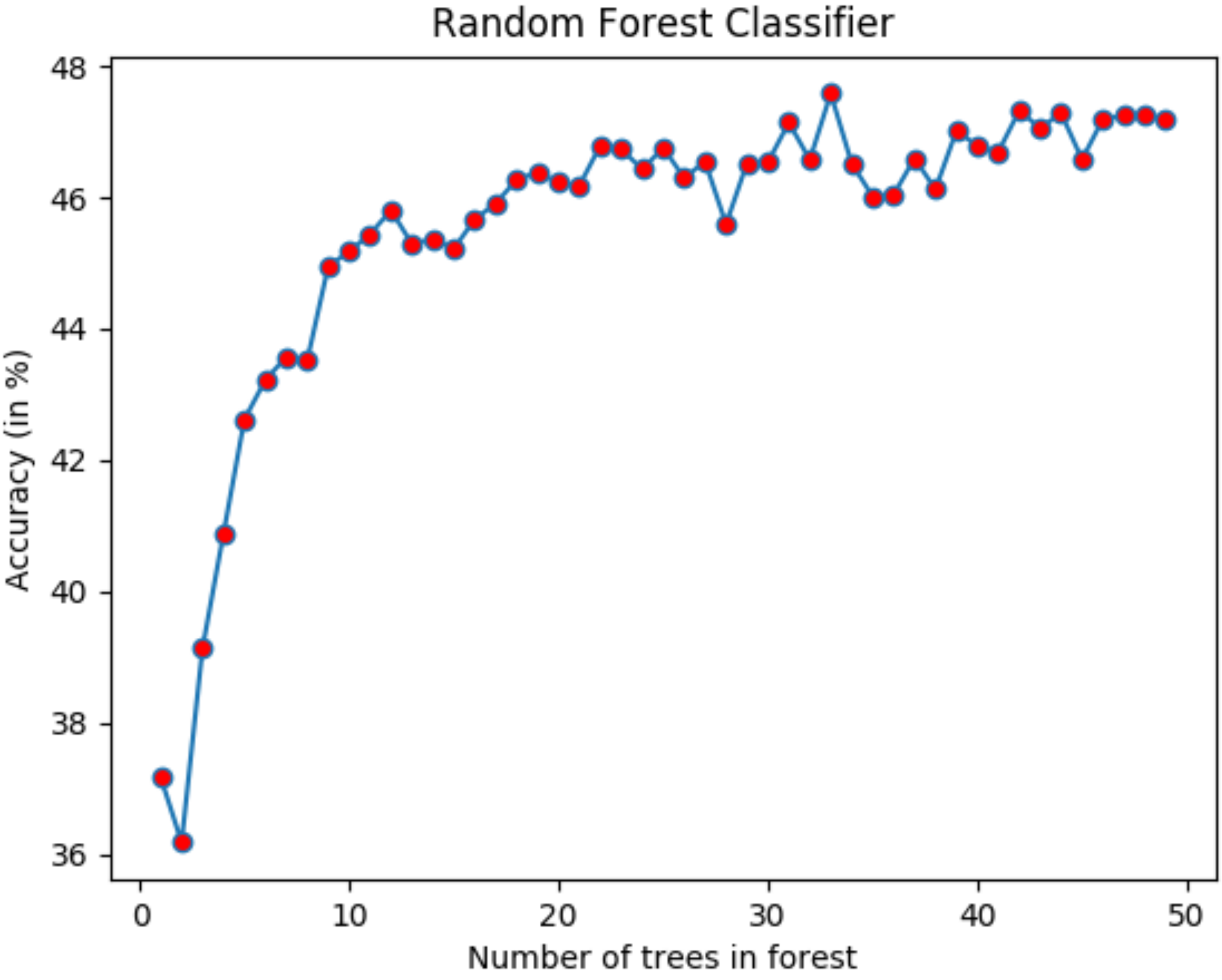
Bagged Decision Tree Classifier :



X - axis represents the number of samples drawn. This clearly a better option that normal decision tree.

Next page ==>

Random Forest Classifier :



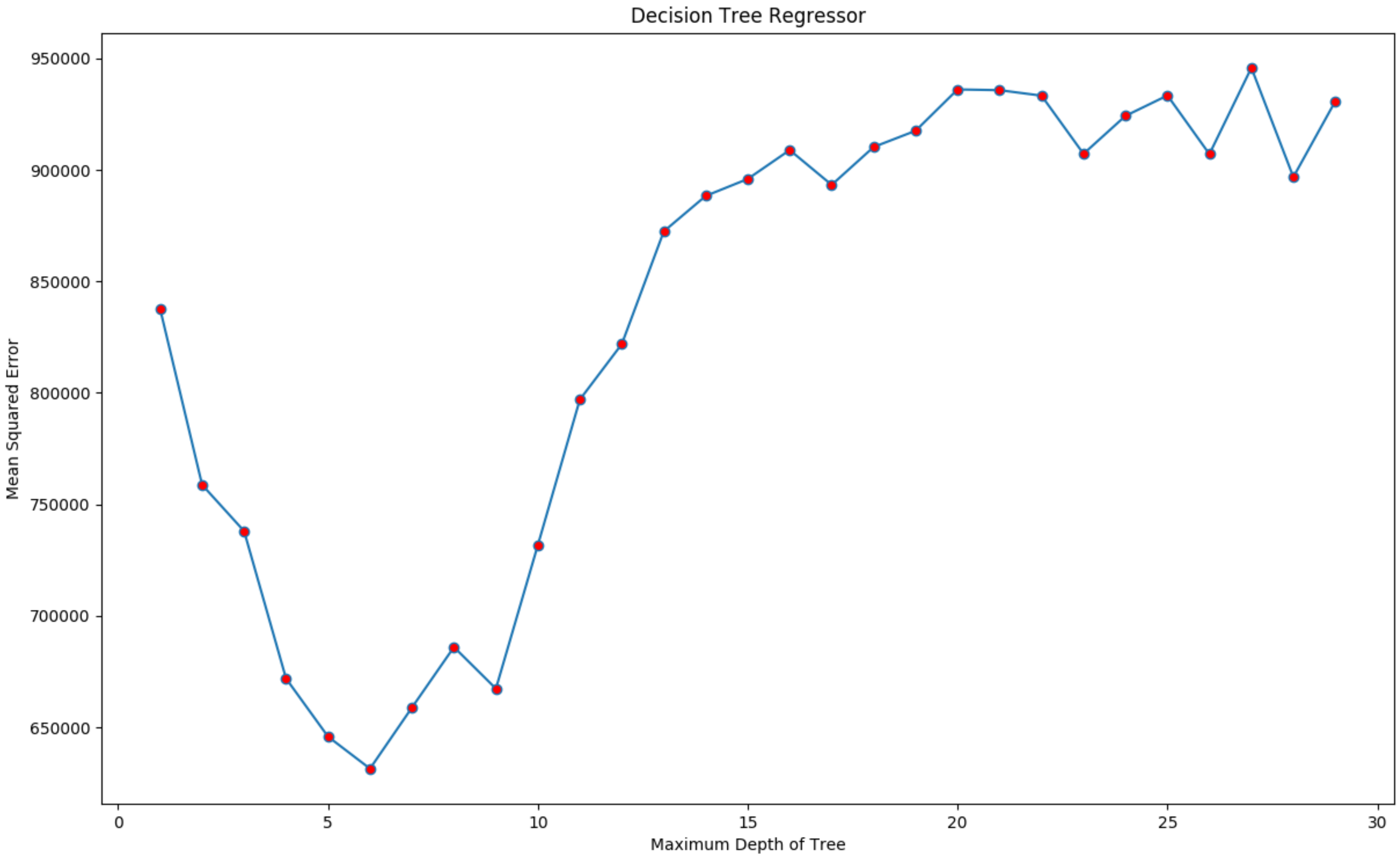
It seems like the random forest could grow more above the accuracy but the algorithm was taking long time to converge so I had to stop the algorithm for smaller numbers.

Conclusion : Random forest works best for classification task with highest accuracy.

Regression

Here, we have to predict the value of PM2.5 using regressive methods of tree.

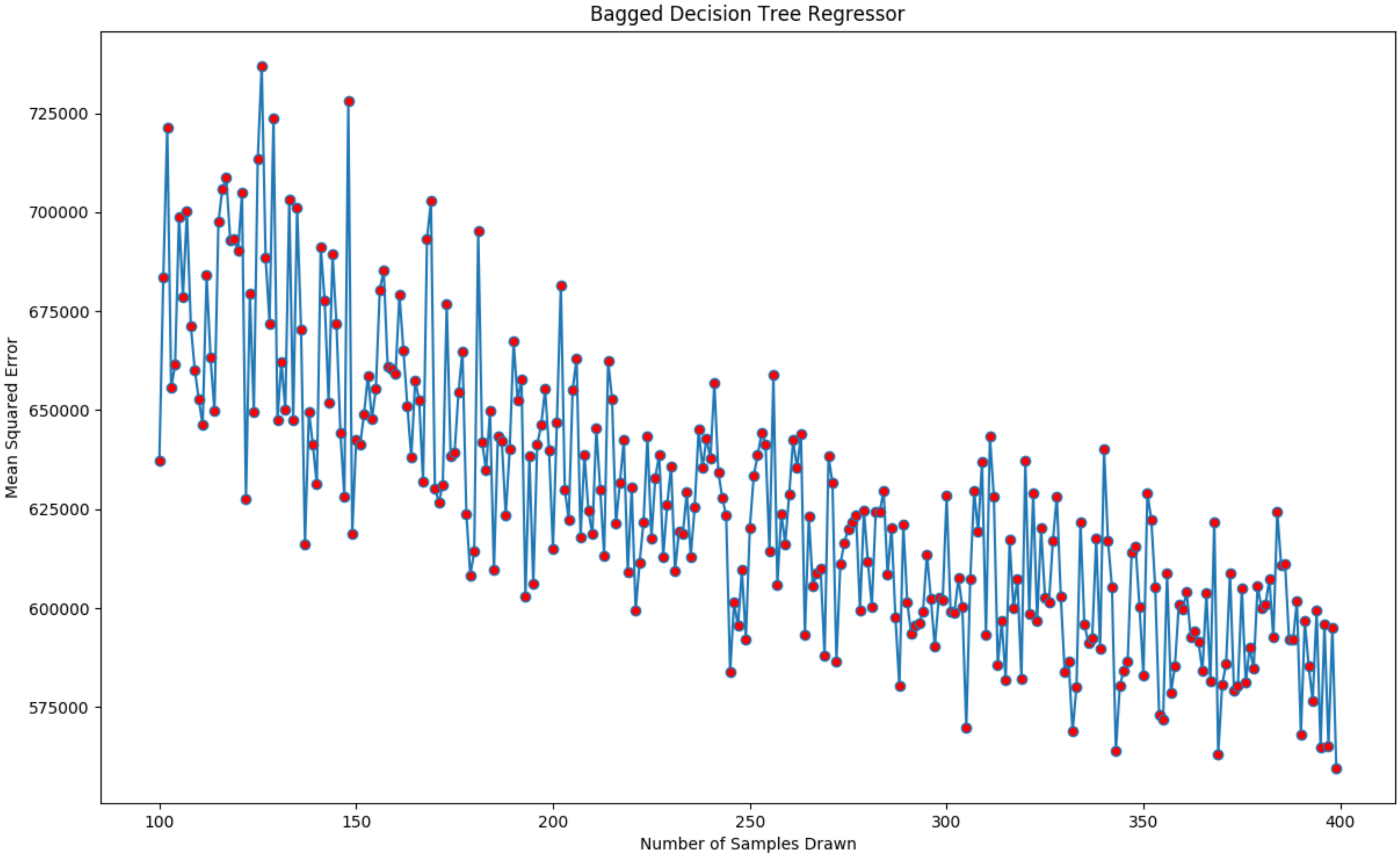
Decision Tree Regression :



X- axis is the depth of tree.

Observation : MSE first decreases for some intervals and the it increases. It means that as we increase the depth of tree it leads to **overfitting**.
So, depth of tree = 6 leads to best results and the MSE at depth of tree = 6 is 631493.6535570803, which is very high error in itself.

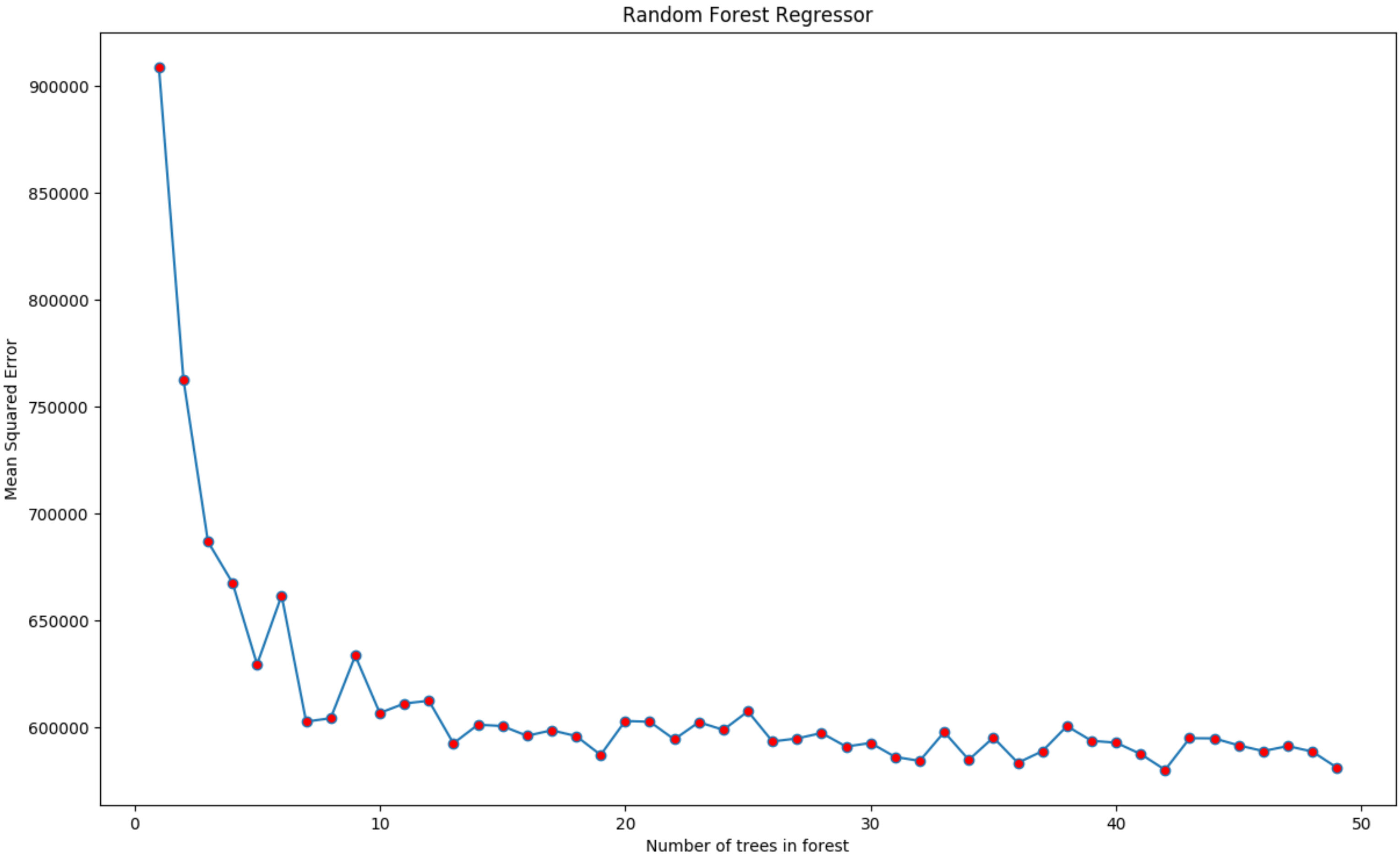
Bagged Decision Tree Regressor :



We can notice that MSE is decreasing with increasing number of samples drawn. After a time the graph does not goes down. Also, if we compare it with decision tree regressor, its minimum error (bagged) is less than that of normal decision tree (Its less than 575K).

Next page==>

Random Forest Regressor :



Observation : MSE drops marginally and after that it settles down and do not show much fluctuations.

Conclusion : Bagged Decision Tree Regressor has performed best for regression task with lowest MSE.