Assignment 5

Dibyendu Roy Chaudhuri
MT19034

Classification

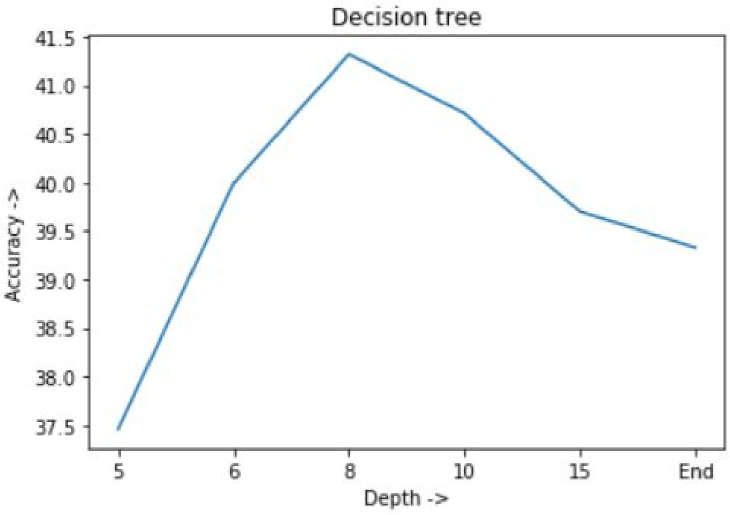
Assumption: NA values of pm2.5 is filled by the average of rest elements (98.61)

Target Field: Month

Classifier: Decision Tree

Sklearn default decision tree: 38.89

Iteration	Depth	Accuracy(%)
1	5	37.46
2	6	39.99
3	8	41.32
4	10	40.71
5	15	39.70
6	End	39.33

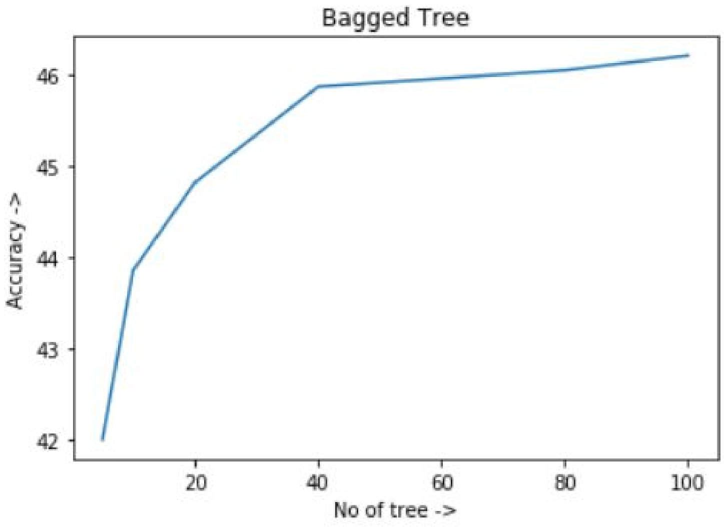


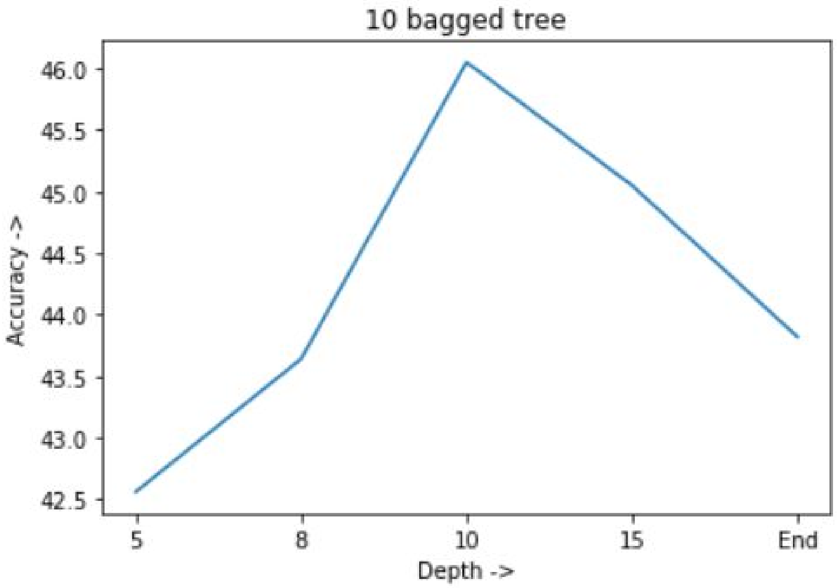
We can see the accuracy of the decision tree is increasing up to depth 8 and later it started to overfit and accuracy fell.

Classifier: Bagged Tree

Sklearn default bagged tree: 44.69

Iteration	No tree	Depth	Data size(%)	Accuracy(%)
1	5	5	70	40.07
2	5	8	70	42.01
3	5	10	70	43.86
4	10	5	70	42.56
5	10	8	70	43.64
6	10	10	70	46.05
7	10	15	70	45.05
8	40	To the end	70	44.82
9	80	To the end	50	45.87
10	100	To the end	50	46.21





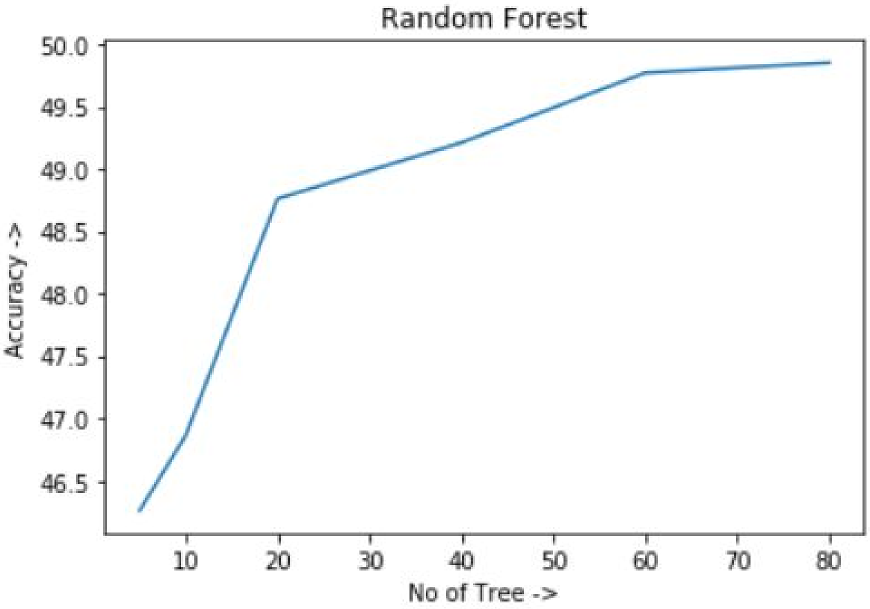
Accuracy of the model increases with increasing numbers of bagged trees and slowly get into saturation. If we reduce training size then up to 50% mark, accuracy is increasing. The depth effect is similar to a decision tree.

Classifier: Random Forest

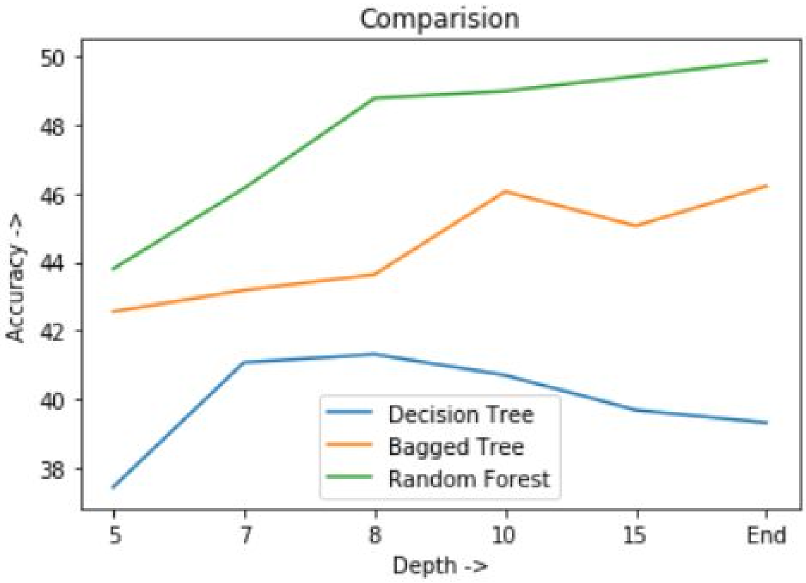
Sklearn default Random forest: 48.32

Iteration	No tree	Depth	Dataset size	No of features	Accuracy
1	5	8	70	4	46.26
2	5	10	70	4	46.15
3	10	8	70	4	46.86
4	10	8	70	6	45.73
5	20	8	70	4	48.76
6	20	8	70	6	46.73
7	20	15	70	4	49.40
8	40	To the end	70	4	49.21

9	60	To the end	70	4	49.77
10	80	To the end	70	4	49.85



Accuracy of the model increases with increasing numbers of bagged trees and slowly get into saturation. If we reduce training size then up to 50% mark, accuracy is increasing. The depth effect is similar to a decision tree. With no feature close to sqrt(n), we got maximum accuracy.



Regression

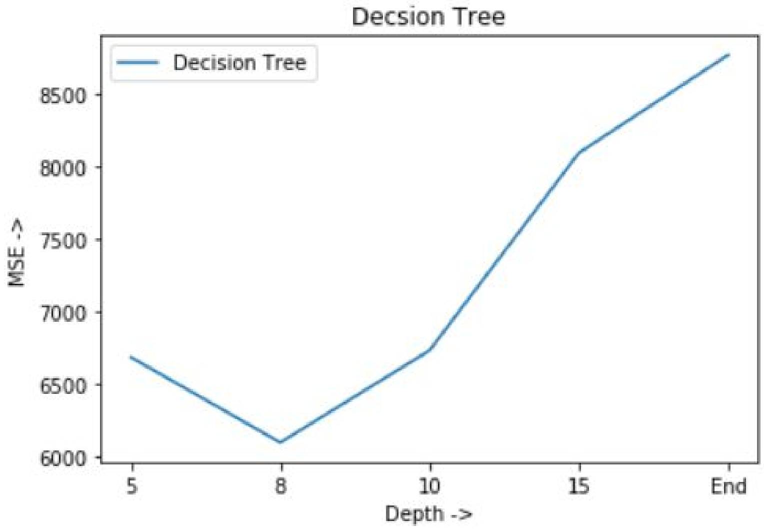
Assumption: NA values of pm2.5 is filled by the average of rest elements (98.61)

Target Field: PM2.5

Classifier: Decision Tree

Sklearn default decision tree MSE: 8681.79

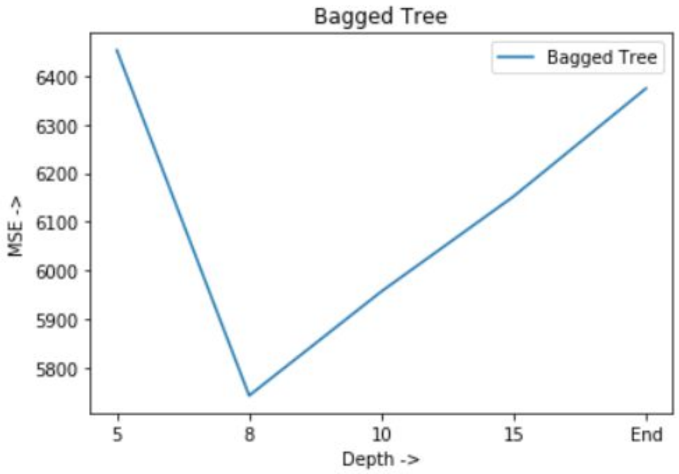
Iteration	Depth	MSE	MAE	SD
1	5	6684.60	56.81	58.79
2	8	6098.76	53.31	57.06
3	10	6731.87	53.57	62.14
4	15	8090.99	57.86	68.86
5	End	8764.42	61.56	70.52

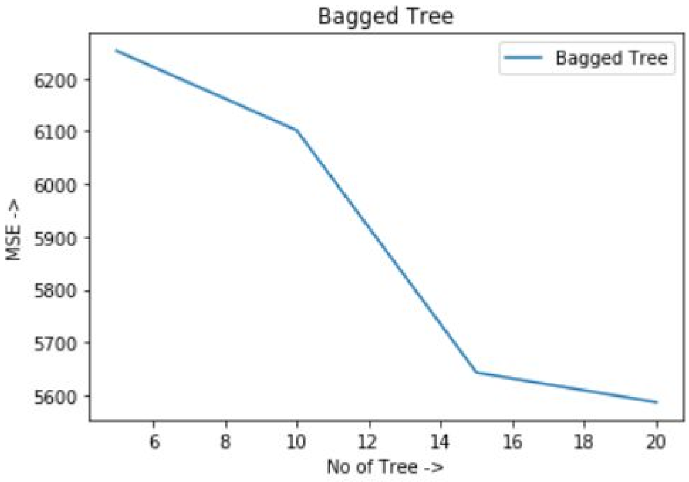


We can see the MSE of the decision tree is decreasing up to depth 8 and later it started to overfit and MSE increased.

Classifier: Bagged Tree

Iteration	No tree	Depth	Dataset	MSE	MAE	SD
1	10	5	70	6452.20	55.98	57.59
2	10	8	70	5743.31	51.88	55.25
3	20	8	70	5587.16	51.16	54.48
4	10	10	70	5957.03	54.82	54.32
5	10	10	50	5690.44	51.83	54.80
6	10	15	70	6152.32	55.32	55.59
7	10	End	70	6374	55.78	57.11

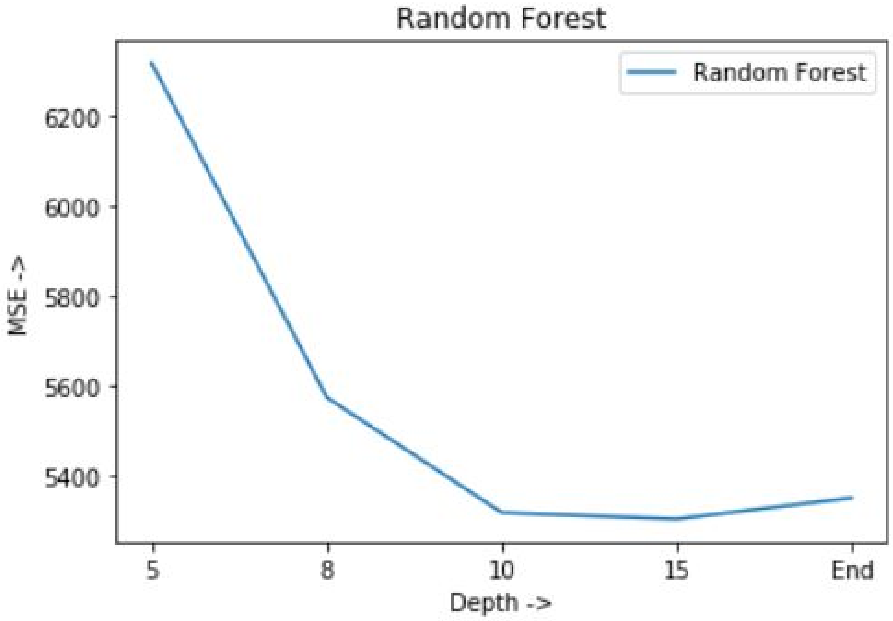


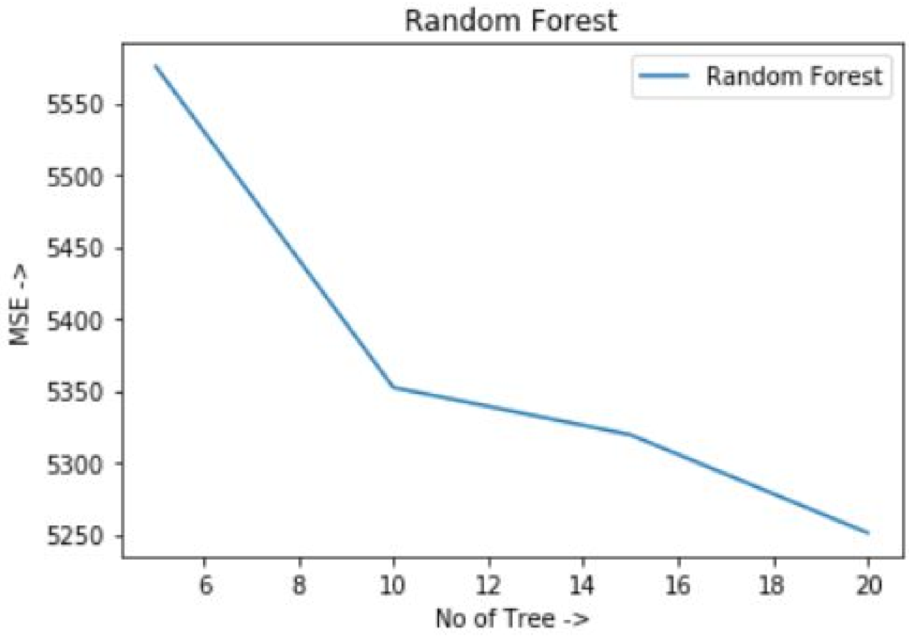


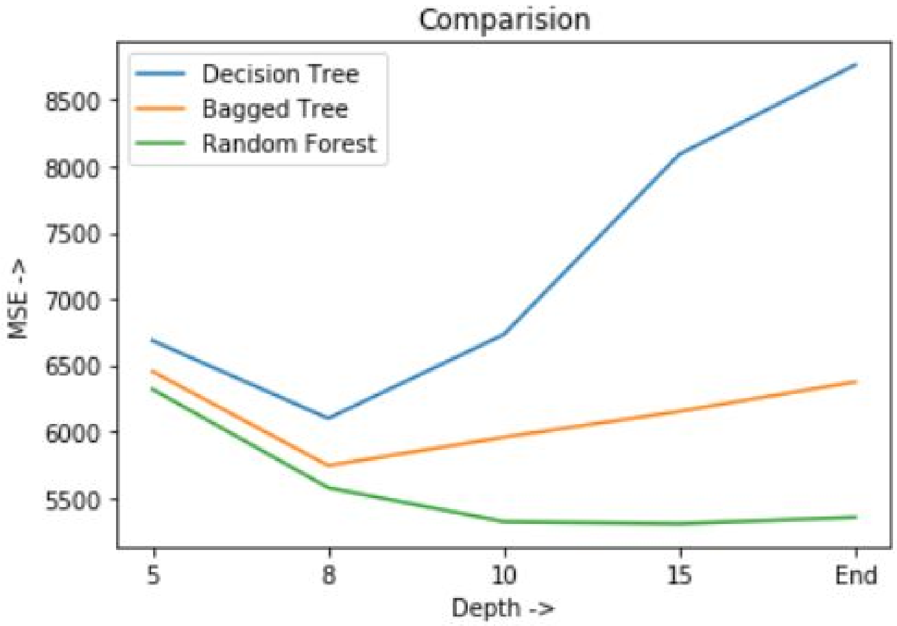
MSE of the model decreases with increasing numbers of bagged trees and slowly get into saturation. If we reduce training size then up to 50% mark, MSE is decreasing. The depth effect is similar to a decision tree.

Classifier: Random Forest

Iteration	No tree	Depth	Dataset	Feature	MSE	MAE	SD
1	10	5	70	4	6317.81	56.47	55.93
2	10	8	70	4	5575.73	51.88	53.69
3	10	10	70	4	5319.76	50.00	53.09
4	20	10	70	4	5234.79	47.25	52.14
5	10	15	70	4	5305.69	48.65	54.20
6	10	15	70	6	5450.41	50.67	54.40
7	10	End	70	4	5352.18	49.06	54.26







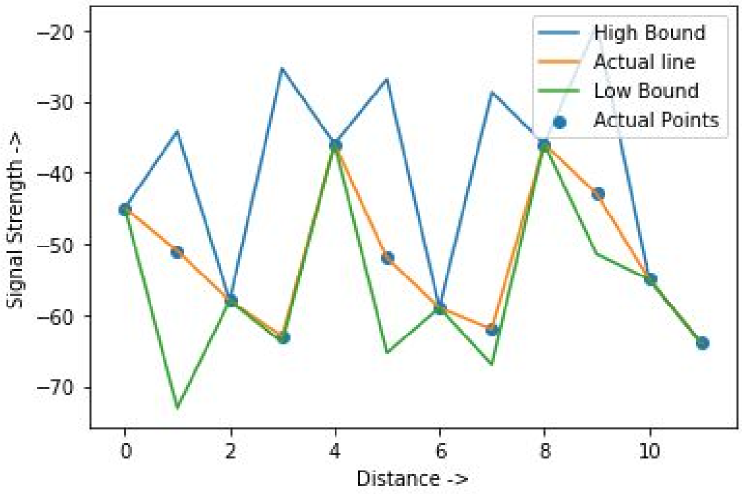
MSE of the model decreases with increasing numbers of bagged trees and slowly get into saturation. If we reduce training size then up to 50% mark, MSE is decreasing. The depth effect is similar to a decision tree. With no feature close to sqrt(n), we got the minimum MSE.

Gaussian Process Regression

Var(f)=111.67 L=1 (Code is done from scratch)

Predicted point: [2,4,6,8,10] Train point: [1,3,5,7,9,11,12]

Predicted means-
[-53.65843350378308, -44.64232895978484, -46.11616741047627, -47.84710528763023, -35.39187489436043]
Predicted variance-
[38.866452019186426, 38.43996100253564, 38.42894020968639, 38.25588744613346, 32.16093467010228]



Here between lower and, upper bound showed confidence region. We can see the predicted dataset are successfully fitted between the confidence region. Here actual point show target points and actual line show ideal function line