Qn 1:

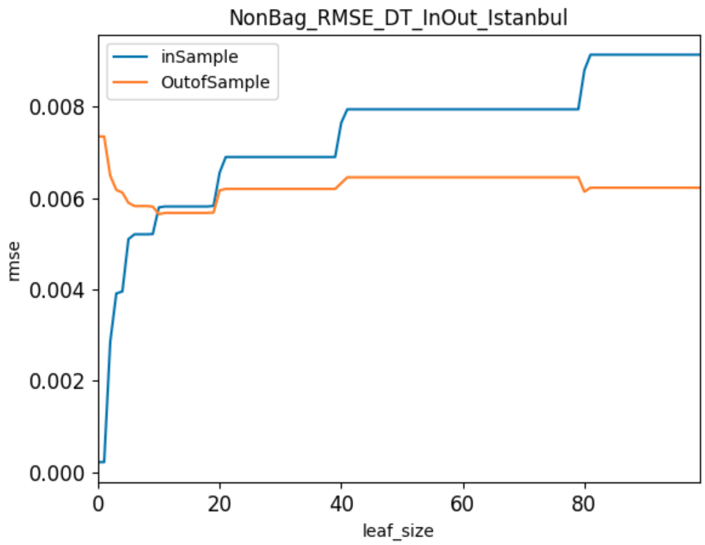

Figure 1. NonBag_RSME_DT

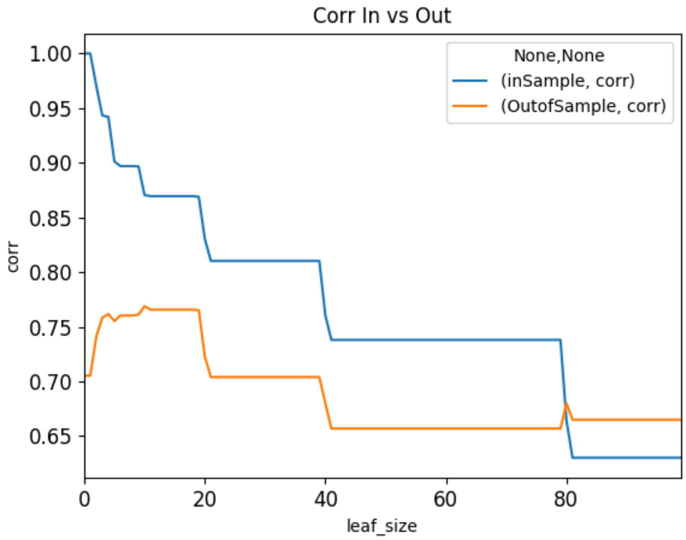

Figure 2. NonBag_Corr_DT

Theoretically, In-Sample RMSE should be lowest (which mean very predictive model) when leaf_size = 1 because it perfectly fit the data. As the leaf_size increases, In-Sample RMSE starts to increase (the model is worsen) because it won't perfectly fit the data anymore. On the other hand, Out-of-Sample RMSE should be very high when leaf_size = 1 (not very good model) since that model is supposed to over-fit the in-sample data. However, as leaf_size increase, the model become more general and start to fit Out-of-Sample data better. This leads to the increase in Out-of-Sample RMSE. With the DTLearner I have built, using Istanbul.csv, it seems that when leaf_size is around 22, I start to see the over-fitting of the model. It means that, increasing the leaf_size will increase In-Sample RMSE but not decrease the Out-of-Sample RMSE. An ideal model would have the Out-of-Sample RMSE keep decreasing.

I can also use the Corr vs leaf_size graph to observe the similar result. In_Sample_Corr started out very high (suggesting a very strong (over-fit) model). It worsen as the leaf_size increases since the model become too generic. With Out-of-Sample Correlation, I see that the corr started out much lower (suggesting not a strong model). Its correlation slowly increases up to around leaf_size= 20. Then the model worsens (evident through the decreasing correlation).

Qn 2:

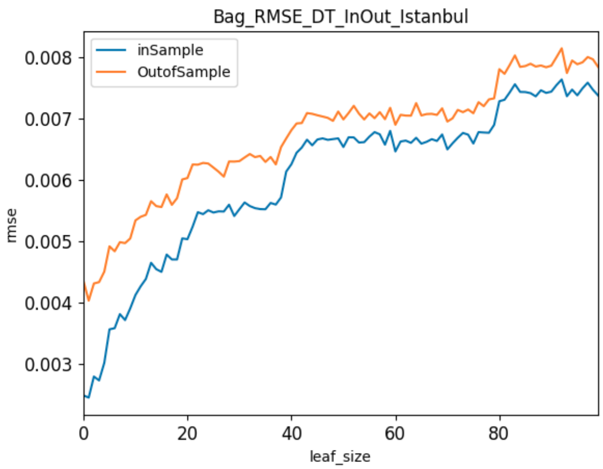

Figure 3. Bag_RSME_DT

Quick glance at Figure 3 shows that there is no clear over-fitting region where over-fitting occurs as I do not really see an area where In-Sample RMSE diverge from Out -of-Sample RMSE. To better understand the result, I will compare the result of the BagLearner model (Figure 3) and Non-Bagged DTLearner model (Figure 1).

Compared to the Non-bag DTLearner model, the BagLearner model show consistent RMSE trend. It is expected that with small leaf_size the In-Sample RMSE of BagLearner started out very low (indicating that it trains in-sample data well) and then increases as the leaf_size. Such trend is very similar to the In-Sample RMSE of the Non-Bagged DTLearner model. However, I can see that the BagLearner model is more general than the Non-Bagged DTLearner: the best In-Sample RMSE of BagLearner model is slightly higher than that Non-Bagged DTLearner model while the worse In-Sample RMSE of BagLearner is smaller than that of the Non-Bagged DTLearner. Overall, the In-Sample RMSE trend of the BagLearner and NonBagLearner are not very different with the BagLearner seems to make a less generic model and thus perform better

However, what truly separates the BagLearner and Non-Bagged DTLearner is the performance of the Out-of-Sample RMSE. A comparison of figure 1 and 3 clearly shows that the BagLearner outperforms Non-Bagged DTLearner. With low leaf_size, the Out-of-Sample RMSE of the BagLearner model is also pretty small compared to that of the Non-Bagged DTLearner. However, as the model becomes more general (leaf_size increases), the Out-of-Sample RMSE of BagLearner model continues to follow the trend of the In-Sample RMSE. There is no region where Out-of-Sample RMSE improves while In-Sample RMSE degenerates.

Qn 3:

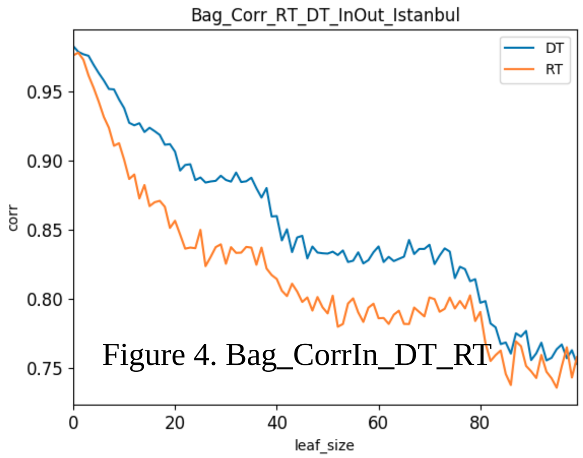

Figure 4. Bag_CorrIn_DT_RT

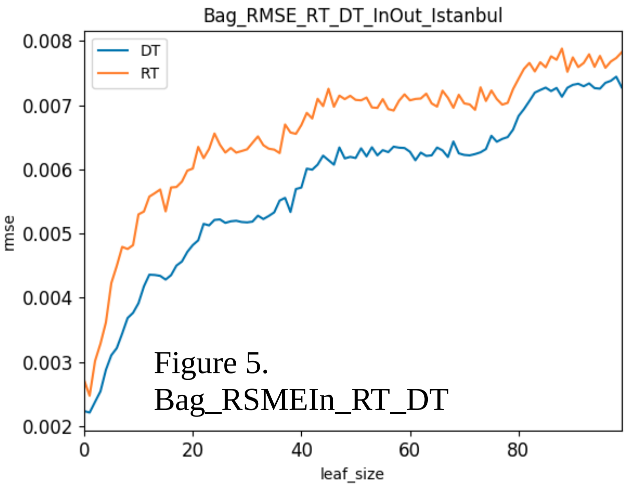

Figure 5.
Bag_RSMEIn_RT_DT

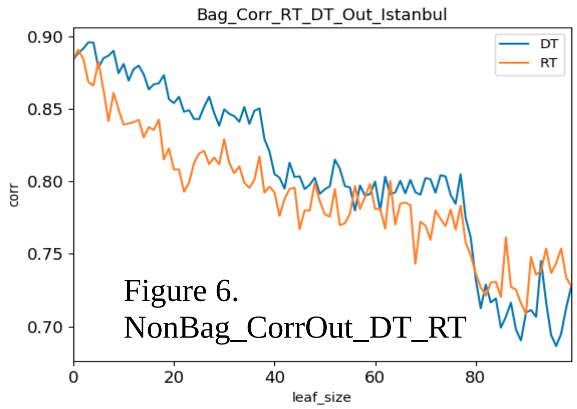

Figure 6.
NonBag_CorrOut_DT_RT

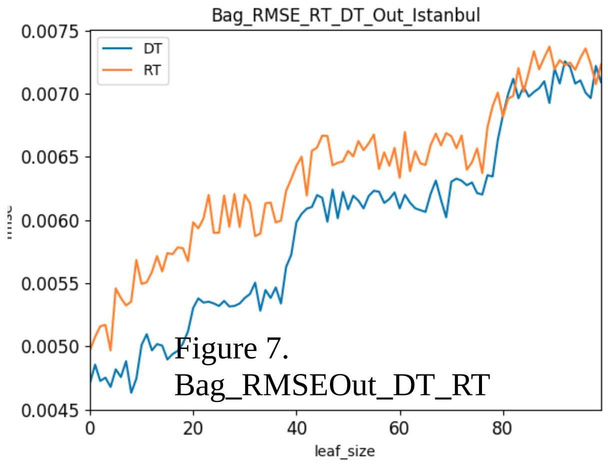

Figure 7.
Bag_RMSEOut_DT_RT

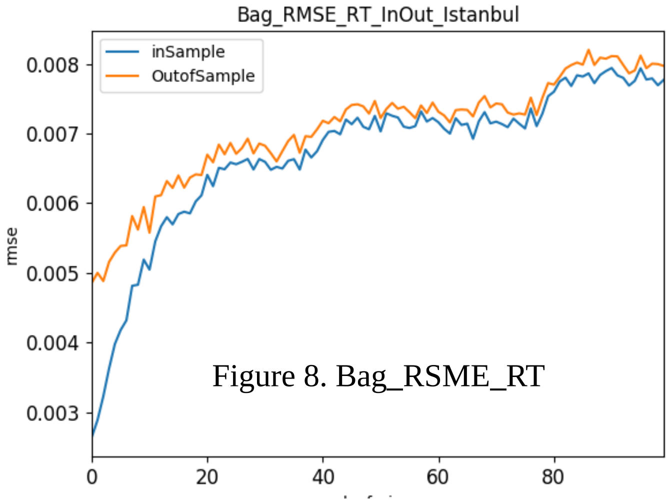

Figure 8. Bag_RSME_RT

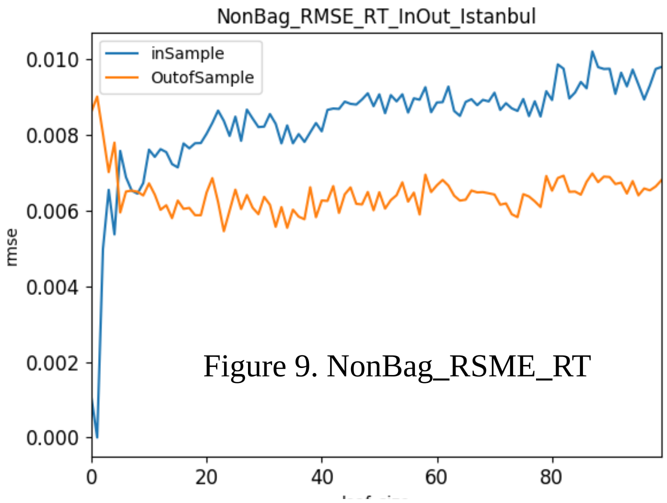

Figure 9. NonBag_RSME_RT

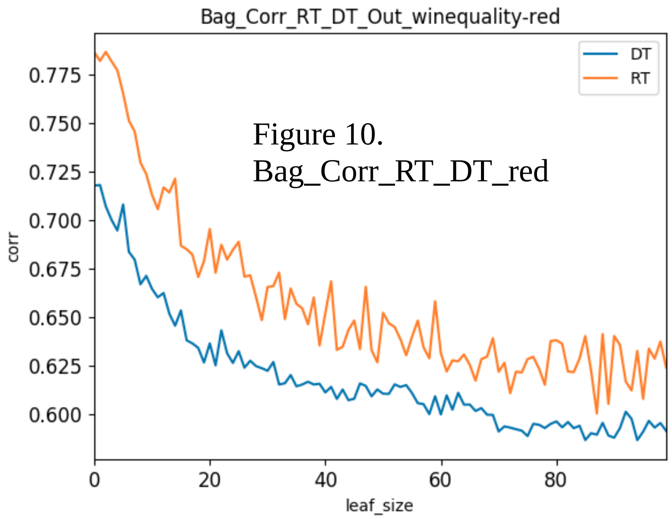

Figure 10.
Bag_Corr_RT_DT_red

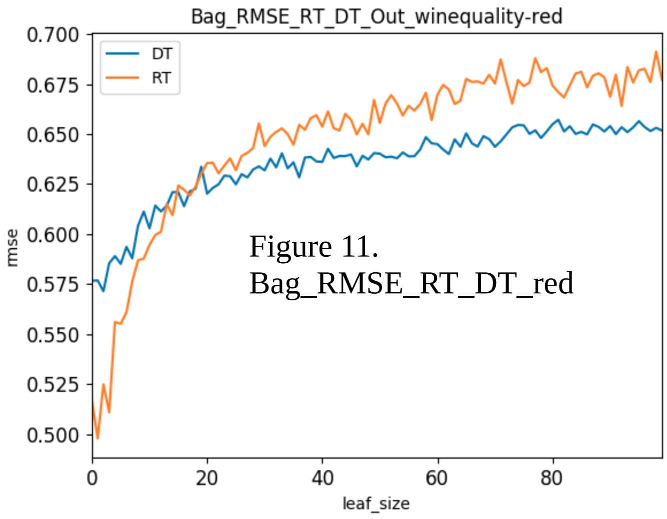

Figure 11.
Bag_RMSE_RT_DT_red

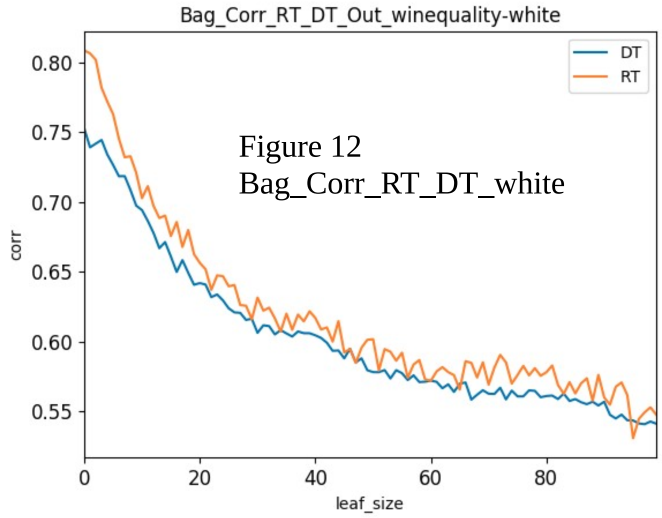

Figure 12
Bag_Corr_RT_DT_white

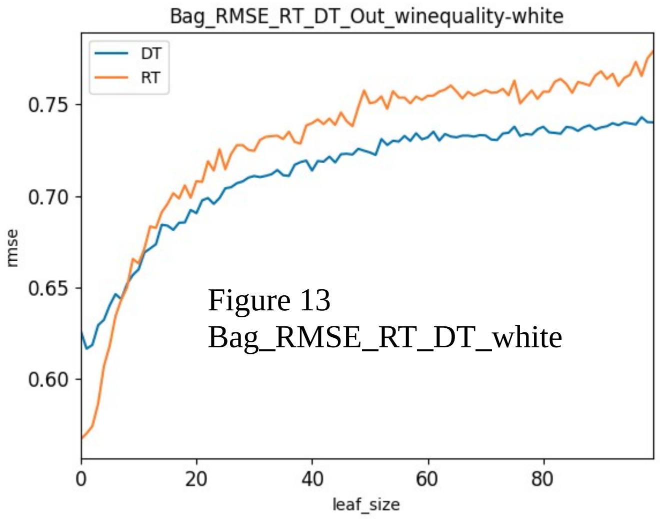

Figure 13
Bag_RMSE_RT_DT_white

Firstly, since I already have discussed the results of BagLearner and NonBagLearner from DTLearner, it is fruitful to compare the result between those of DTLearner and RTLearner. Upon comparing Figure 1, 3 to Figure 8, 9, I first observe that the RTLearner did not eliminate over-fitting for NonBagLearner. A close comparison between figure 1 and 3 shows that, there are not much different in the trend as well as the range of RMSE values between BagLearners of RT Learner and DTLearner.

However, as I dived deeper to understand the RMSE and Corr values between RT Learner and DTLearner, I notice from Figure 4 and 5 that, the In-Sample RMSE of DTLearner is smaller than that of the RTLeanner. Similarly, when I compare Figure 6 and 7, I also notice that the Out-of-Sample RMSE of DTLearner is smaller than that of RTLearner. These evidence suggests that the BagDTLearner seems to perform better than the BagRTLearner.

When I compare the Correlation values between DTLearner and RTLearner, I also come to similar conclusion as I found that Correlation values of BagDTLearner is better (higher) than that of BagRTLearner. This again points to the fact that DTLearner seems to be better than RTLearner.

To further confirm out suspicion that DTLearner is better than RTLearner, I test similar values on other datasets. Firstly, I look at the winequality-red data. To our surprise, the result was not as straightforward. In Figure 10., the correlation value of bagRTLearner is higher than that bagDTLearner which suggests that BagRTLearner seems to be a better model. However, the RMSE values graph in Figure 11 shows that there are no real winner between the two models.

Similar, when I swap the winequality-white data into the model, I do not see a clear winner between BagRTLearner and BagDTLearner as BagRTLearner has better correlation values while the RMSE value are much murky.

These results seems to suggest that the performance of RTLearner depends on the specific type of data and randomization while the performance of DTLearner is more consistent.