IA4 report
Name: Yijie Ren

General Introduction:

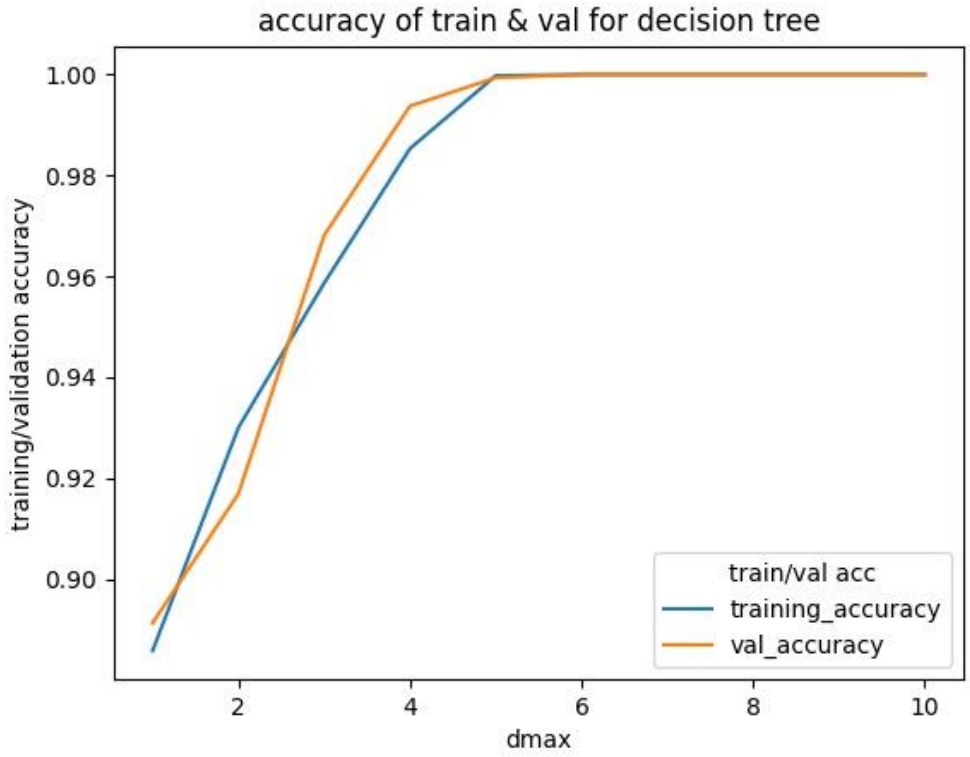
This report illustrates the results of decision tree, random forest and Adaboost based on single decision tree. It is surprising for me that the decision tree itself is so powerful, since the model is quite simple. The same happens to random forest and Adaboost, which reminds me of Occam's Razor. The simplest is the best.

Part 1:

(a)

The root feature is "odor=n", and the other splits right beneath the root are "bruises?=f" and "spore-print-color=r". The information gain of 3 features are 0.5351, 0.3952 and 0.1029, respectively.

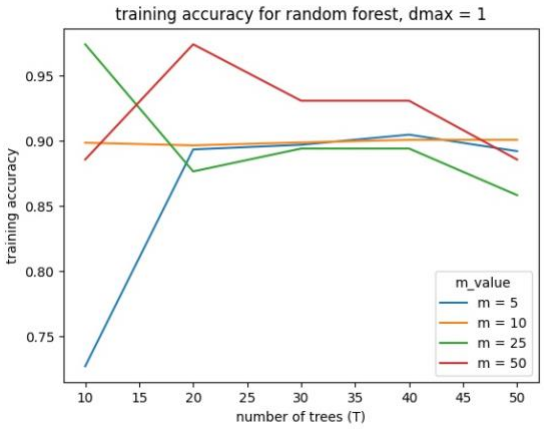
(b)

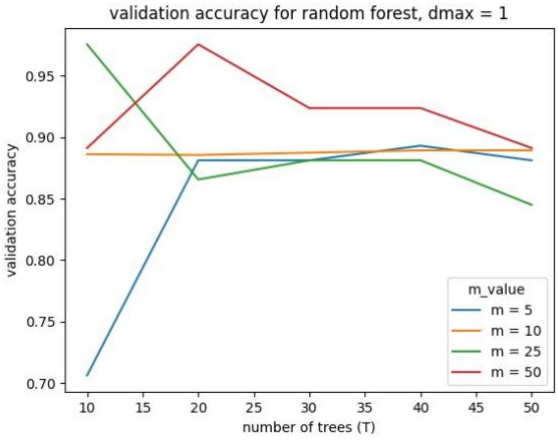


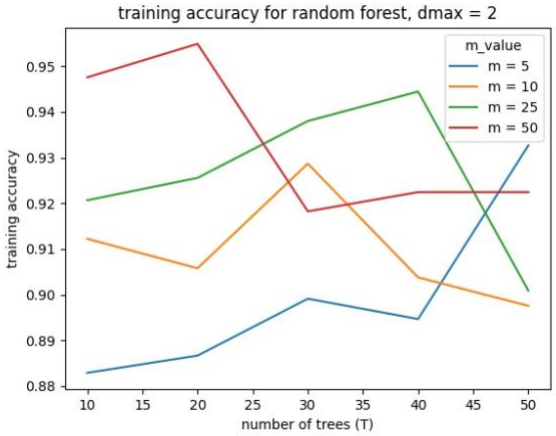
When depth = 5, the training accuracy reaches to 100%. I would say there is no overfitting, because both training and validation accuracies are increasing as dmax increases, which indicates that the decision tree is well-trained.

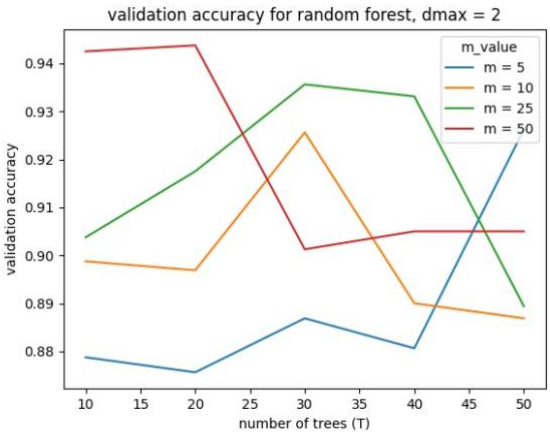
Part 2:

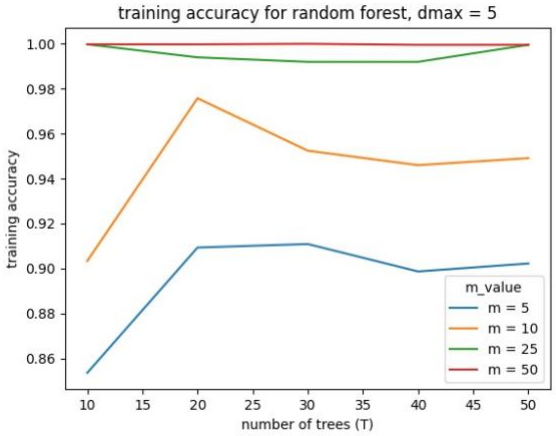
(a)

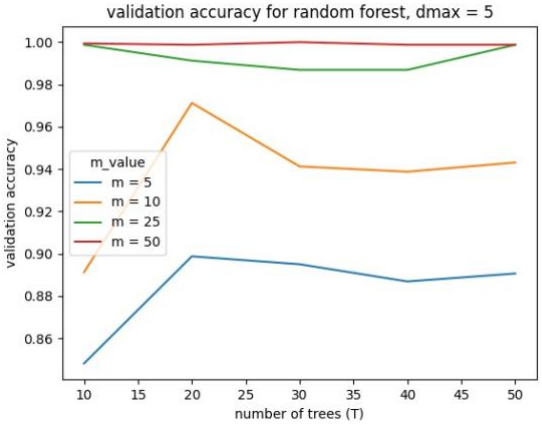










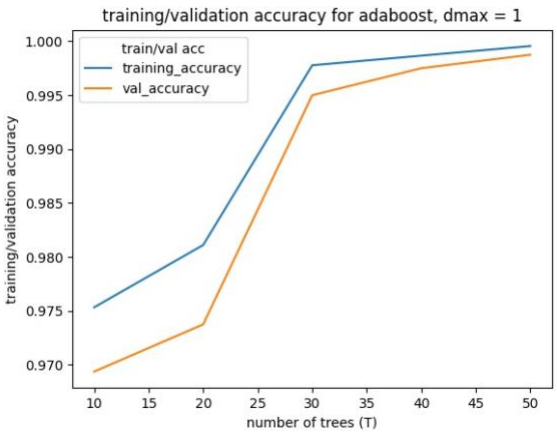


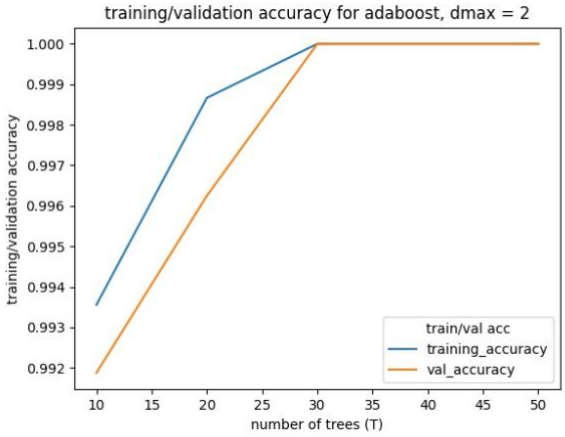
Comparing the training accuracy curves and the validation accuracy curves, I do not think the model is overfitting, since each pair of training/validation curves almost follow the same "curve patterns". However, after comparing to the training/validation accuracies of dmax=5, I think the models are underfitting, for both dmax =1 and dmax =2 – the accuracies can be larger. Also, when dmax=5, the curves of m=5 and m=10 are also underfitting, since both training and validation accuracies are not stable, they are still fluctuating. If the trees in random forest could be more, I assume the accuracies can also increase.

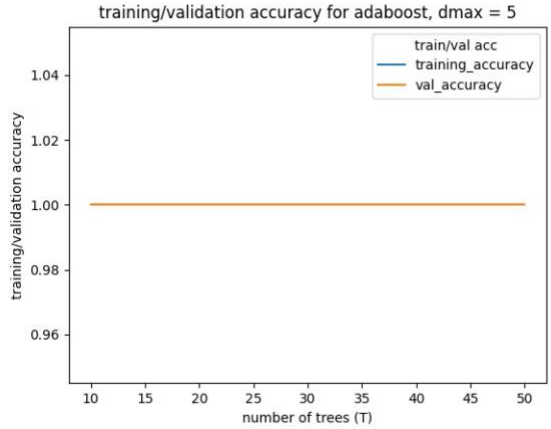
(b)
When dmax=1, the dominating factor of performance loss should be both the bias and variance, caused by both shallow depth of the trees and small number of the trees in random forest. Because the model is not fully trained, both the bias and variance are very large.
When dmax=2, the dominating factor of performance loss should be the variance, caused by the small number of feature selections, such as m=5. Because the results of the models are changing a lot when the training data (mainly the different feature selections) is changing slightly.
When dmax=5, the performance loss is mainly because of the bias, caused by small feature selections. The depth is almost deep enough for this dataset, so as we observed from the figure of dmax=5, the more features selected for each tree in the random forest, the more accurate results we can obtain. Thus, the bias might exist when selected features=5 (m=5), the model cannot achieve the best performance.
As for alternative configurations, I will try to let dmax=6, m =50, T=50. Although from figure in Part 2 (a), the accuracy has already reach to 100% when dmax=5, if there are more data coming in, the higher the dmax is, the performance of larger dataset can be better.

Bonus Part:

(a)







From the 3 figures above, I think all models are not overfitting, since the "curve patterns" are almost the same for both training and validation accuracies. However, since both training and validation accuracies can reach to 100% when the hyperparameters are configured well, the models with accuracies less than 100% can be considered underfitting, which include the configurations of (dmax=2, T=10), (dmax=2, T=20) and all models with dmax=1.

The reason of underfitting might be the models are not fully trained. The number of trees in each ensemble is not enough to construct a model with high accuracy. As illustrated in the figures, when the number of trees increases, no matter how small the dmax value is, the training/validation accuracies are always increasing.

(b)
When dmax=1, the dominating factor of performance loss should be both the bias and variance, caused by the small number of trees in the ensemble. Because the model is not fully trained, both the bias and variance are very large.
When dmax=2, the dominating factor of performance loss should be the bias, still caused by the small number of trees in the ensemble. However, even if the number of trees is only 10 (T=10), the model can still achieve 99.2% accuracy on validation dataset, I don't think the bias here is that much to consider.
When dmax=5, there is no performance loss here, since the model is completely trained well.
As for alternative configurations, I will try to let dmax=3 or 4, along with T=10 or 20. Although the model performs perfectly when dmax=5, the training time is too long. If we could get the same results with less dmax and less T, the model will be better.