Q1. S.T bootstrapping each sample contains about 2/3 unique examples

Say, 'N' observations in total

- we randomly draw 'N' observations from the dataset w replacement
- $\therefore$ Prob for each select $= 1/N$

Prob of not selecting $\Rightarrow (1 - 1/N)$

Prob of 1 obs NEVER getting selected for all 'N' $\Big\} \Rightarrow \therefore N \times \left(1 - \frac{1}{N}\right)$

$\because$ Draws are independent

$$P(\text{unique}) = 1 - \left(1 - \frac{1}{N}\right)^N = 1 - \left(\frac{N-1}{N}\right)^N$$

Sub w N=3 for simplicity $\Rightarrow$ $1 - \left(\frac{2}{3}\right)^3 = 1 - \frac{8}{27} = \frac{19}{27} \approx 0.70$

$\therefore$ 70% $\Rightarrow$ unique $\rightarrow \approx \frac{2}{3}$

Q2) Simple majority among K learned hyp.
Each hyp has error ε
↳ maybe not independent of each other

Is the error of the ensemble never worse than ε?

It can be worse than ε.

Proof:

Say we have 5 cases, and 3 hypothesis

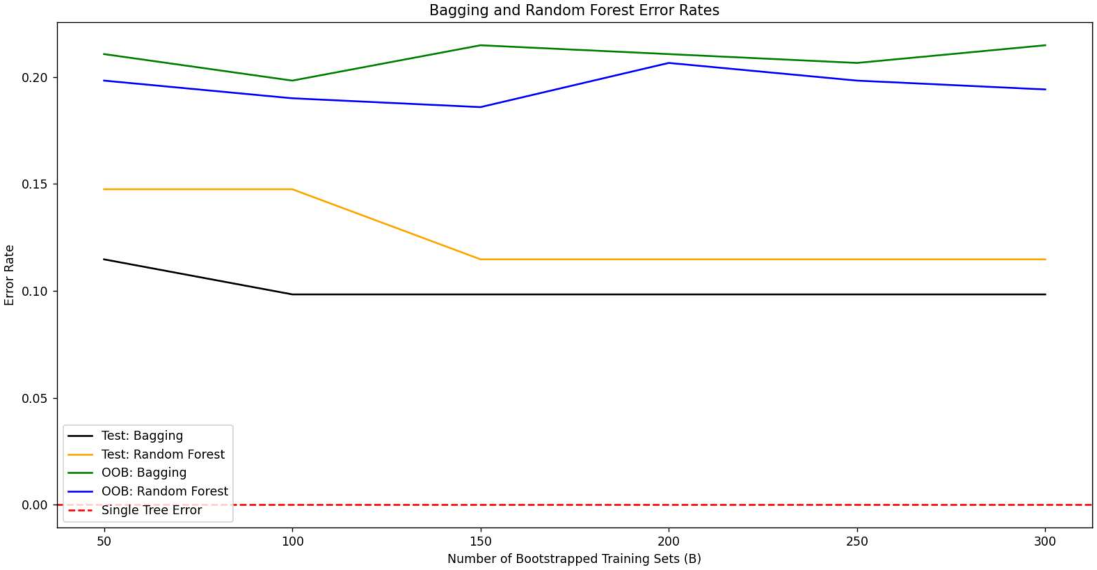
Errors are:

i) Hypothesis 1 fails for cases 1, 3, 4 } $\varepsilon = 3/5$
ii) Hypothesis 2 fails for cases 1, 2 5
iii) Hypothesis 3 fails for cases 2, 3, 4

∴ The simple majority decides 1, 2, 3, 4 → fails erroneously which means
Overall error $= 4/5 = \varepsilon'$

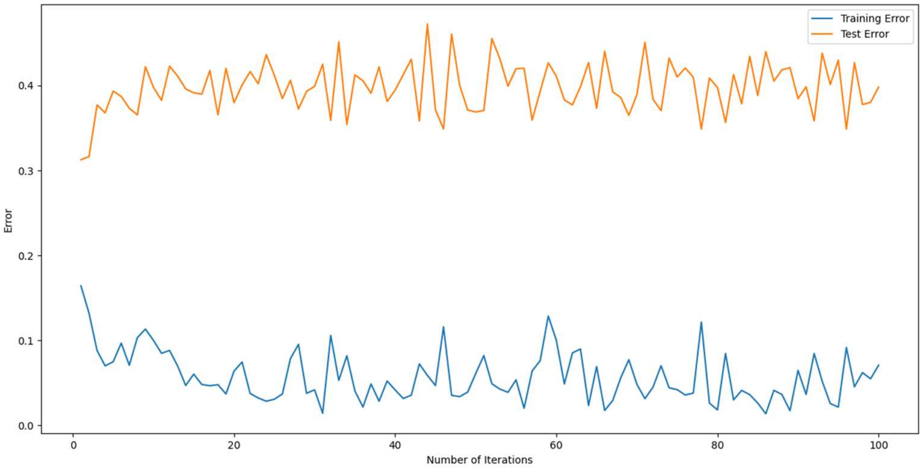
$$\boxed{\varepsilon < \varepsilon'}$$

Question 3:



Question 4:

b) Plot the training error as well as test error and discuss its behavior.

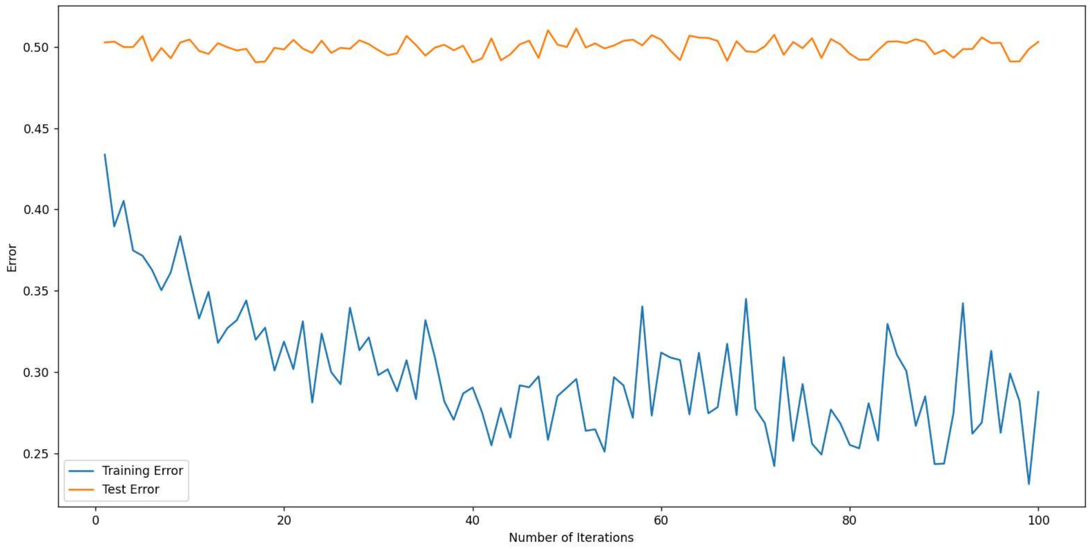


As expected, training error is lower than test error and although there are irregularities throughout, there is a reduction in both errors.

c) Investigating the number of iterations needed for the test error to start rising

It starts rising from the first iteration, at least in my code.

d) Repeat the AdaBoost experiments with new dataset



The reduction in training error is much more evident and thus the model gets better with more iterations.