Group_A8 Lab 1 Block 2 - ENSEMBLE METHODS AND MIXTURE MODELS

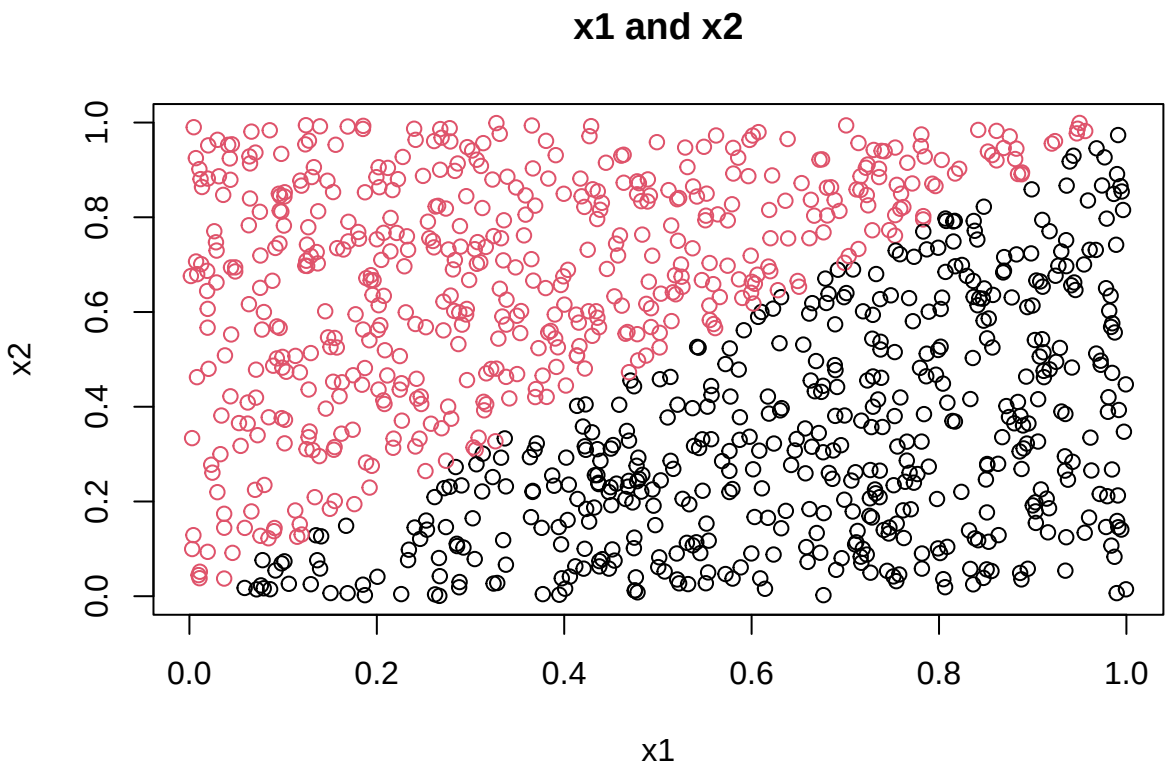
Ayesha Gamage-(ayega981)/Muditha Cherangani(mudch175)

2023-12-06

1. ENSEMBLE METHODS

Part a.

Here used 1000 training data sets of size 100 and Report results for when the random forest has 1, 10 and 100

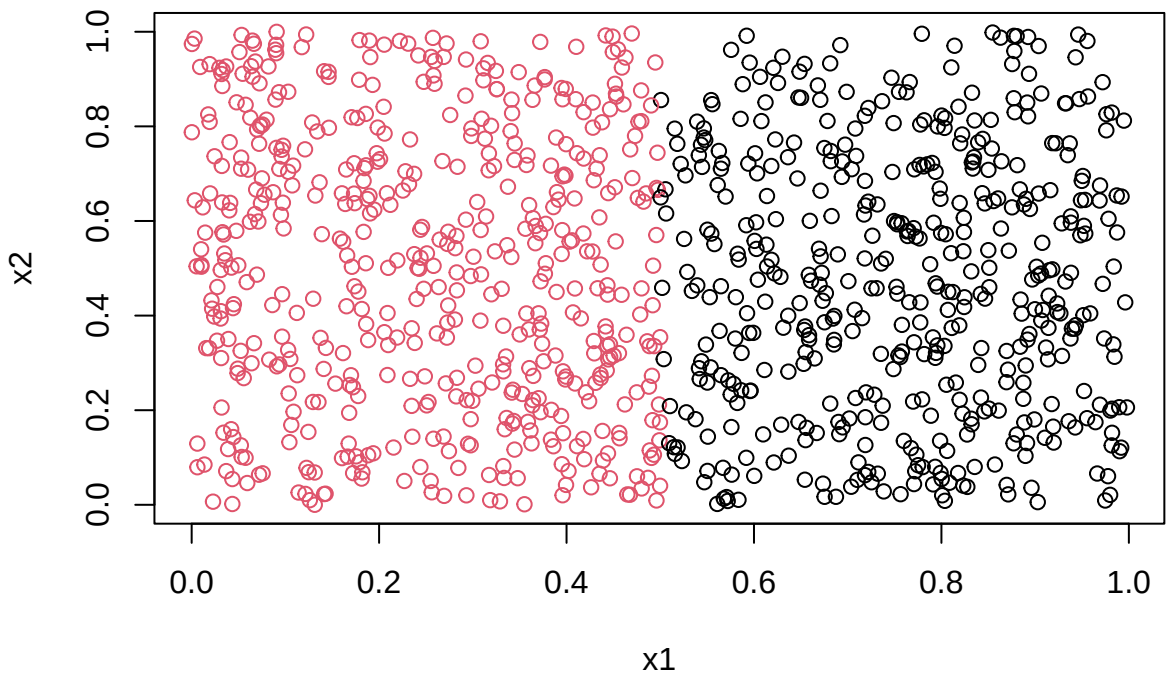


trees.

```
## Misclassification_errors
## Mean error when tree 1 > 0.209062
## Mean error when tree 1 > 0.128525
## Mean error when tree 1> 0.102176
```

Misclassification_errors part a

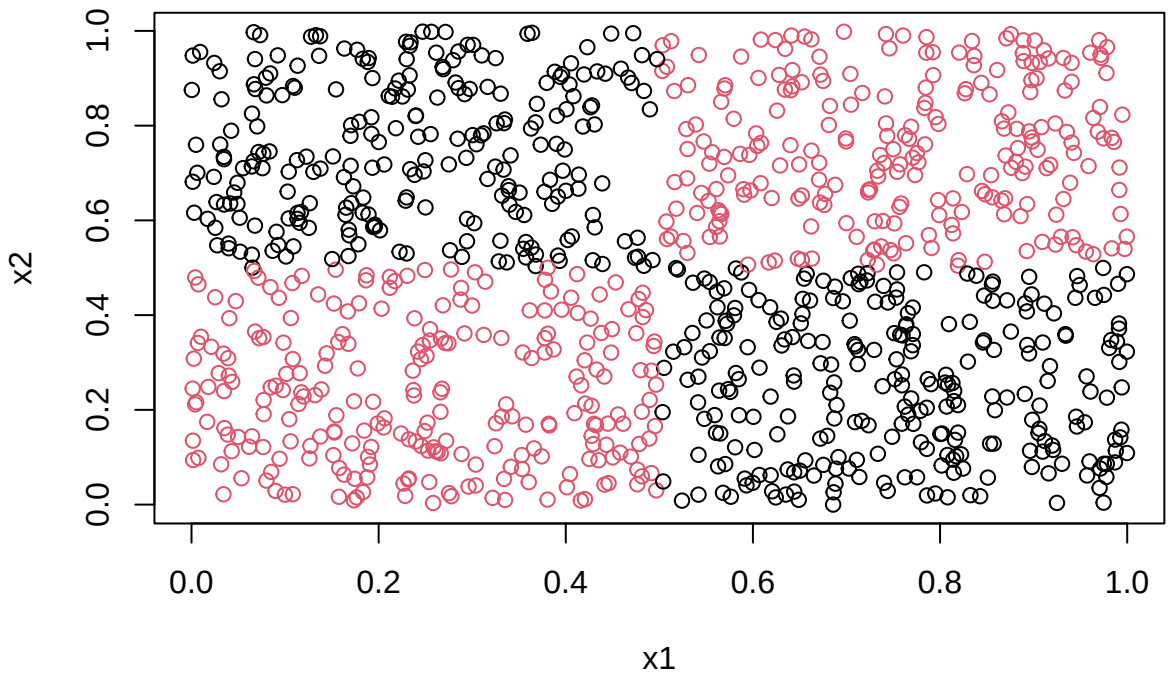
part b.




```
## Misclassification_errors
## Mean error when tree 1 > 0.095316
## Mean error when tree 1 > 0.015269
## Mean error when tree 1> 0.006106
```

Misclassification_errors part b

Part c.




```
## Misclassification_errors
## Mean error when tree 1 > 0.257463
## Mean error when tree 1 > 0.142149
## Mean error when tree 1> 0.091343
```

Misclassification_errors part c

What happens with the mean error rate when the number of trees in the random forest grows? Why?

As the number of trees in a random forest grows, the mean error rate tends to decrease. Misclassification_errors plot illustrate clearly that, when number of trees is 1 error is higher than number of trees are 10 and 100. Because Random forest used bagging to reduce variance. Following formula used to compute variances,

$$Var|\frac{1}{B}\sum_{b=1}^{B} z_b| = \frac{1-\rho}{B}\rho^2 + \rho\sigma^2$$

According to this formula, we can reduce variances by increase number of trees in the forest.Adding more trees may strike a balance between bias and variance.

The third dataset represents a slightly more complicated classification problem than the first one. Still, you should get better performance for it when using sufficient trees in the random forest. Explain why you get better performance.

When compere third error graph with other error graph, it also perform as simple data sets.One reason is when B increase the variance is decrease.But the complexity of data can have a significant impact on the variance and σ will increase with complexity. Since σ increase Var() increased.With the high variances,lead to over fitting. But here we used Random forest and it is combination of decision tree , bagging and decorelation.

$$Random forest = decision tree + bagging + decorelation$$

The base models' predictions can be seen as random variables.Since data set more complicated,in the bagging, reduces the variance of the base model's predictions without increasing the bias. As well as, by aggregating multiple models, to strike a balance between bias and variance, providing improved generalization performance on complex datasets. There for complex data set also have better performance.