Assignment 2:

Q1(a)

The Decision Tree made is extended by creating a function called random forests that estimates a given number of estimators.

Q1(c)

The parallel implementation of the tree yields similar accuracy compared to the serially implemented tree. However, slight performance benefits are observed in certain implementations.

Q1(d)

The accuracy in this case obtained is given as **94%** with **20** estimators and 2 features per decision tree in case of Random Forests.

The same 2 features on a single tree yield an accuracy of about **90%** with the same number of features.

Q1(e)

On 5 cross-validation evaluation the performance on all the estimators 5 estimators(though this is subject to change) yields the best performance on the testing set of data.

Q3.
(b)

On adding false columns to the dataset we observe that the accuracy of the individual iteration falls. However, the overall accuracy of classification is not affected drastically.

Q4.

On carrying out 100 rounds of bagging, the individual random variation in the data due to the individual estimators cancels leaving a plot close to a straight line. This fit is less prone to variance as compared to the individual polynomial fits.

Q5.

(a)The random numbers in the data are generated using the current time in python.

The random number is generated after a number of iterations that are a function of the current time in milliseconds. The random number is taken after applying a pseudo random number sequence for multiple iterations.

(b)

On invoking the function about 1000 times we observe that the generated is close to uniform with each number occurring at least once.

However, certain numbers are more probable than others which depicts that this generation method is not truly random.