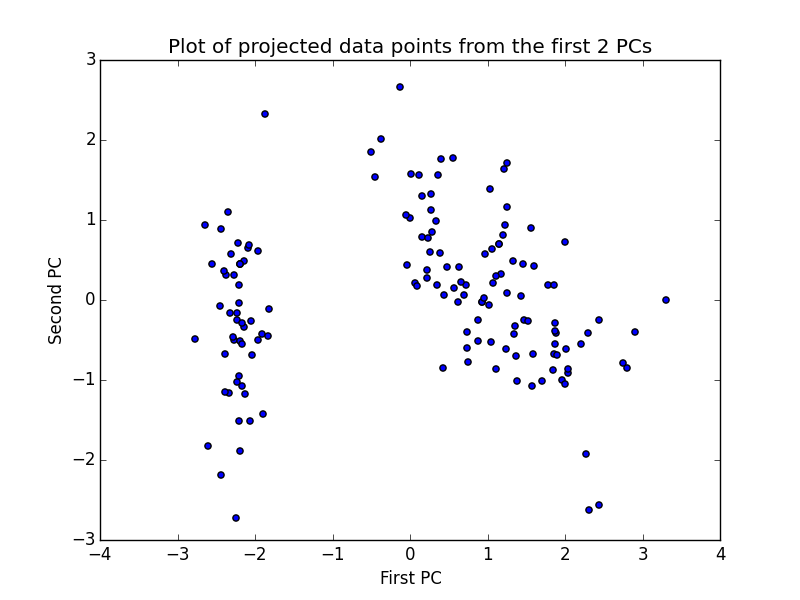
Assignment 4

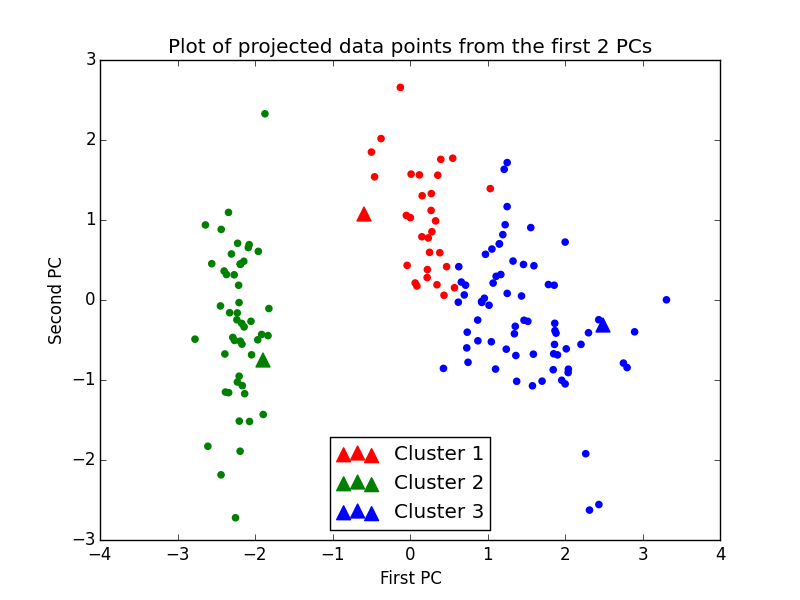
# Exercise 1

Normalising data to zero mean and variance one does not influence the training process and the classification accuracy of a CART or Random Forest. This is because only the range of one feature is used to define the split at each node, so the features are not compared to each other and scaling the features has no effect on the result.

# Exercise 2

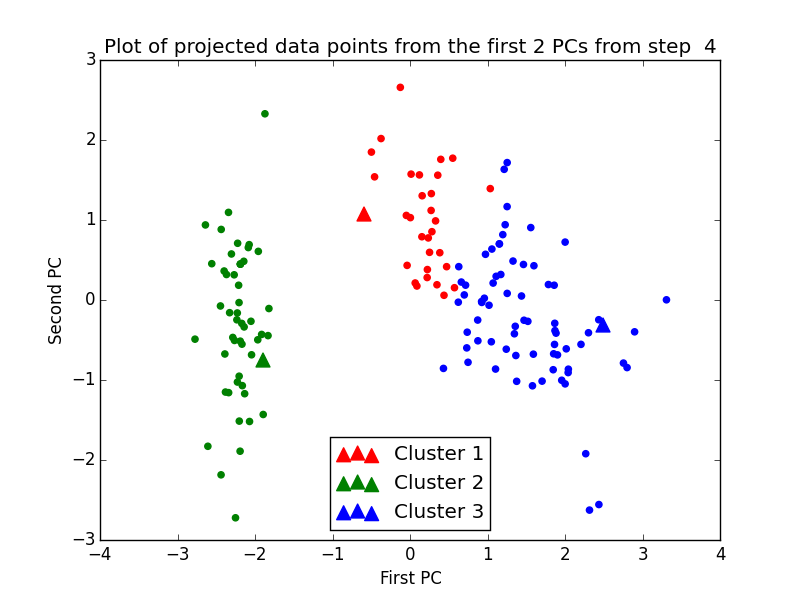
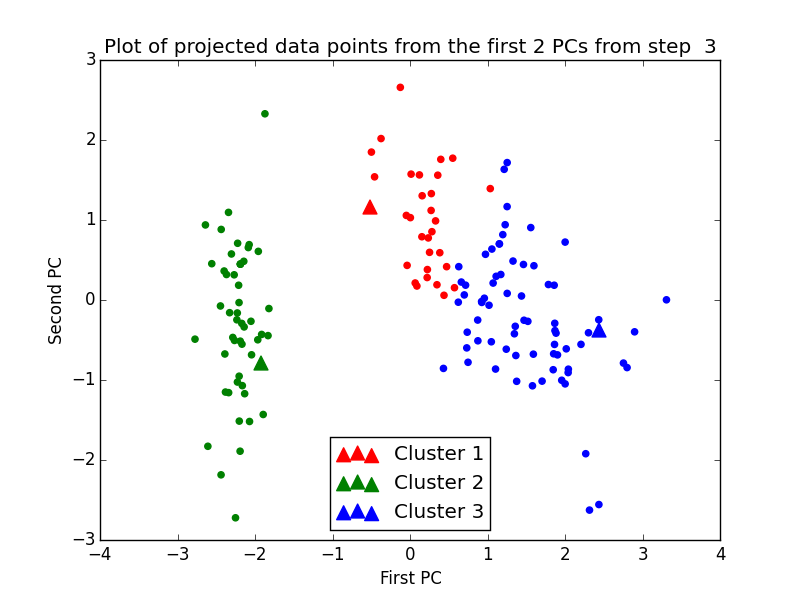
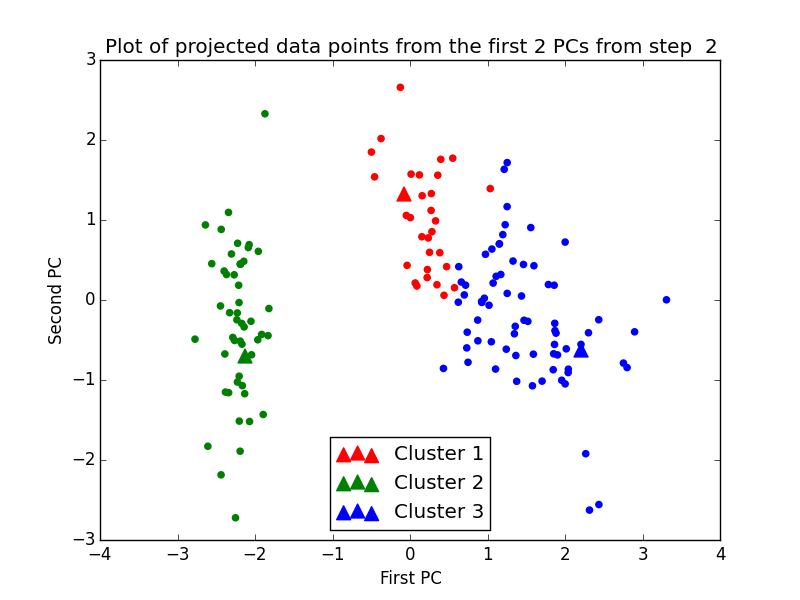
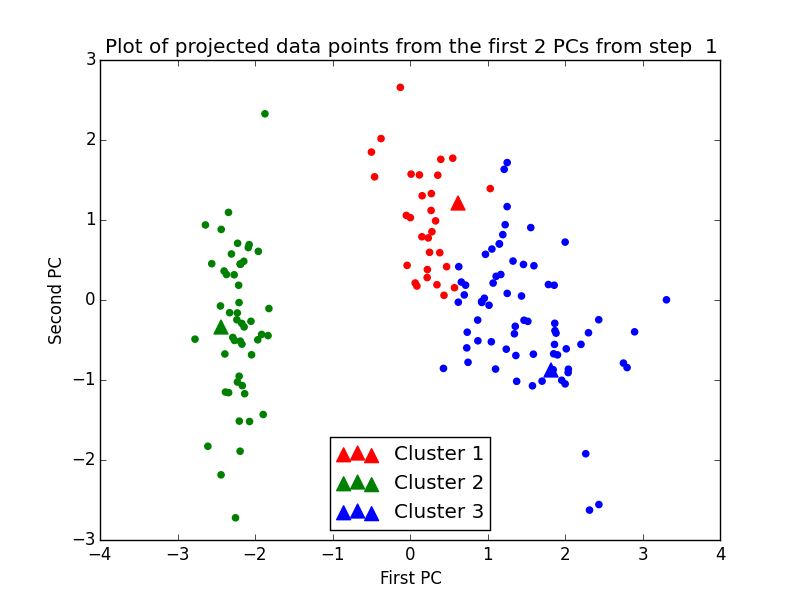
c) 

The plot seems to show 2 distinct clusters.

d) 

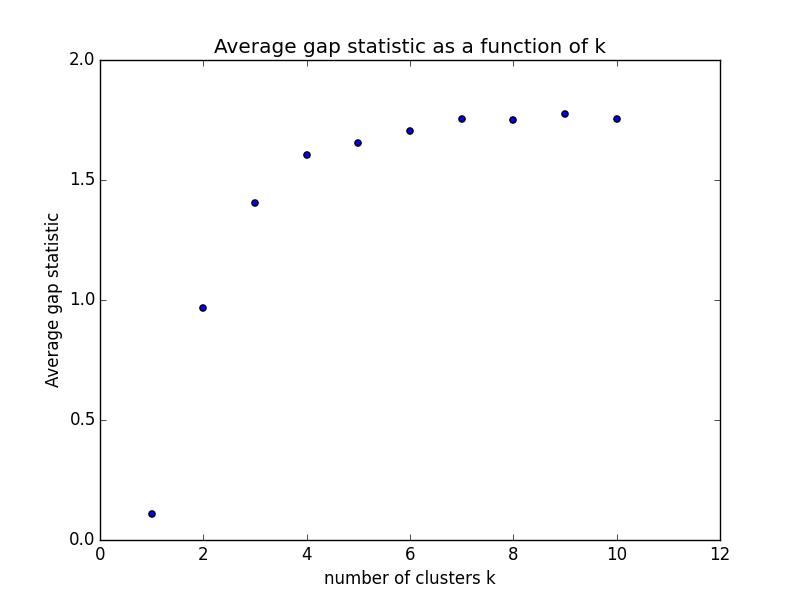
There is a slight overlap between data points belonging to cluster 1 and cluster 3, but there are three distinct clusters.

# Exercise 3



Initially, the centres of the clusters are not in the correct position. As the new centres are calculated in the following steps, the centres eventually move into the correct position.

# Exercise 4

b) 

The value of k with the best performance is 9 because as it can be shown from the plot above that it gives the maximum gap statistic, which gives the most reasonable choice for number of clusters.