Assignment 1

STK 802

Connor McDonald
u16040725

Behavioural Analytics

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science
University of Pretoria
Date: 14 October 2021

1 k-NN Mode Seeking

1.1 k-NN Density Estimate

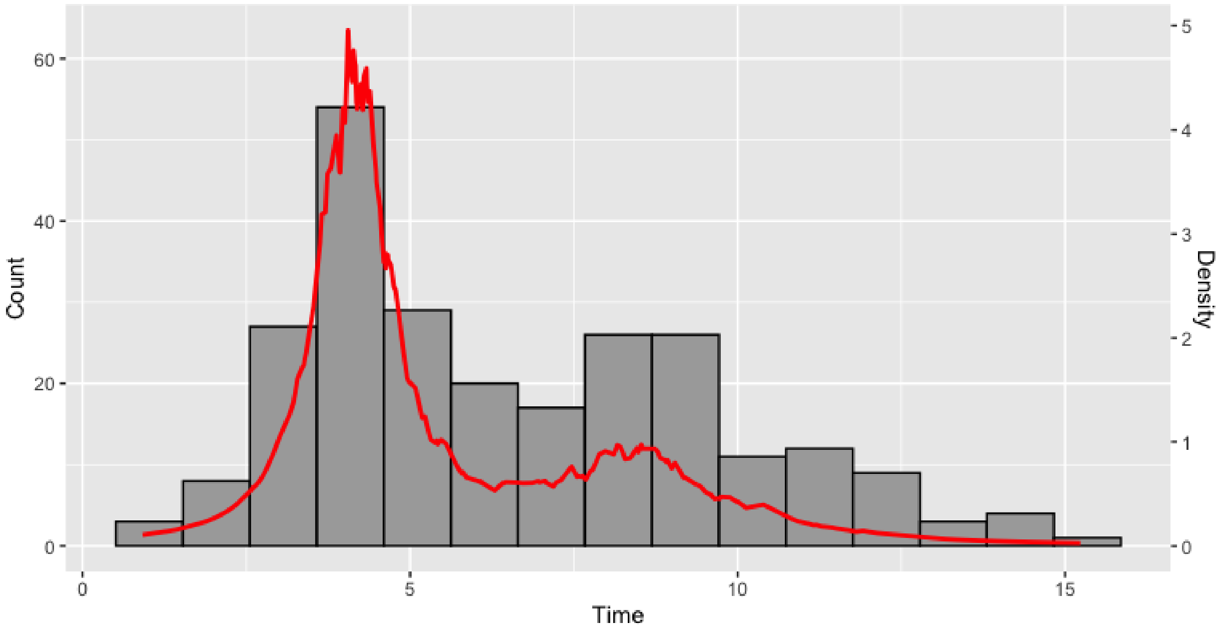
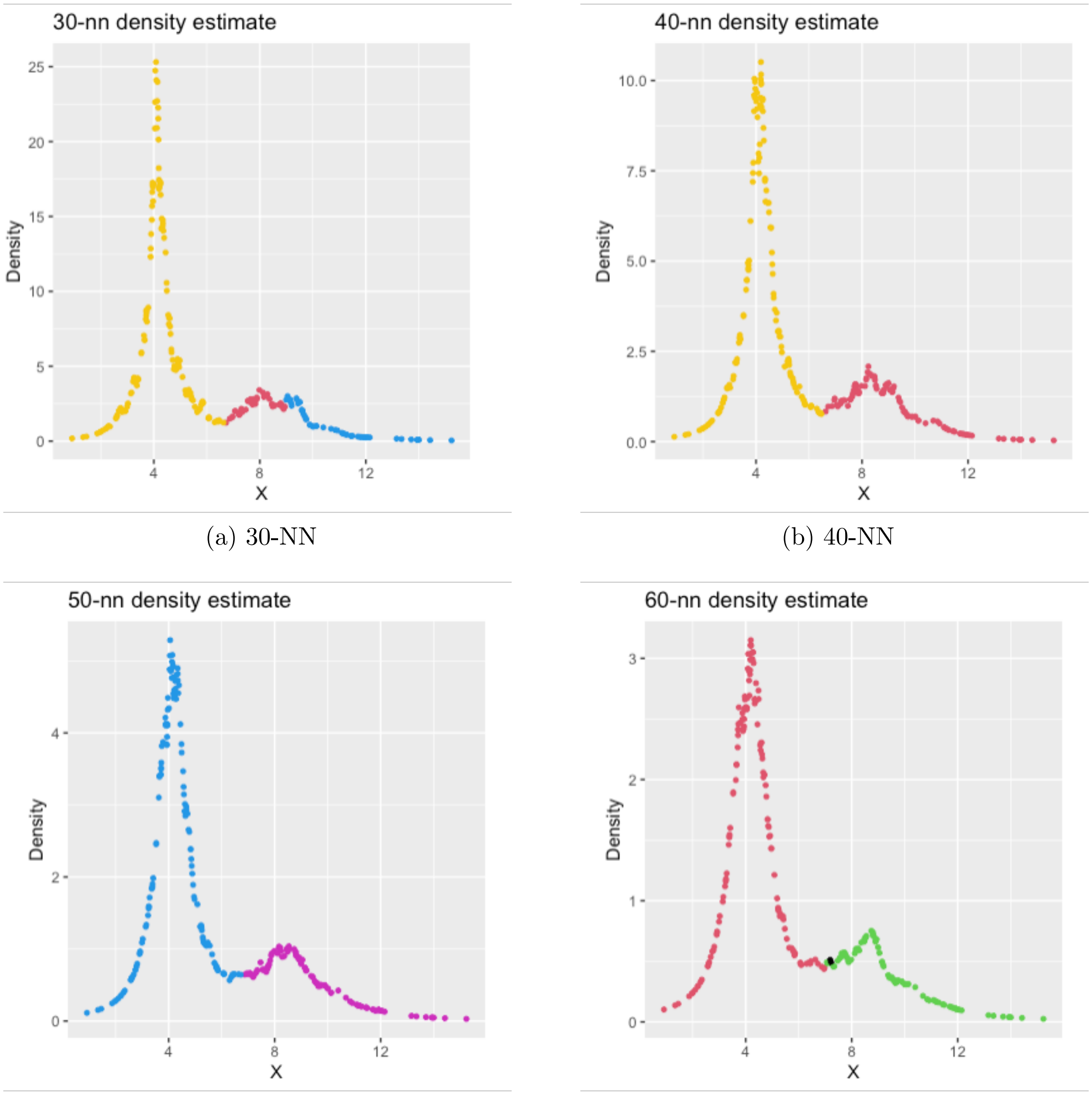


Figure 1

1.2 Cluster Solution

Various values of k were investigated in attempt to find the optimal clustering solution, it was found that the optimal number of clusters was two, which was later confirmed by the dendrogram in figure 3. Interestingly, the number of clusters found typically has an inverse relationship, however it was observed that when going from k = 50 to k = 60, an additional cluster was identified with a much smaller cluster with a proportion of only 1.2% (Figure 2d), this becomes more relevant when using mixture modelling to cluster the data. Finally a k-value of 40 was selected as the optimal value for k as it created two clusters whilst also achieving a balance between smoothness and number of modes found.

The plots of the cluster solution with various k values is shown on the following page.



(a) 30-NN

(b) 40-NN

(c) 50-NN

(d) 60-NN

Figure 2

1.3 Mean, Variance and Proportions

Table 1: Cluster Statistics for k = 30

Cluster Index	**Index Value**	**Mean**	**Variance**	**Proportion**
170	7.981523	7.955511	1.4295300	0.208
95	4.0810298	4.204749	1.4295300	0.564
124	9.045791	10.963116	2.4867021	0.228

Table 2: Cluster Statistics for k = 40

Cluster Index	**Index Value**	**Mean**	**Variance**	**Proportion**
178	8.239222	9.501913	3.791600	0.440
135	4.185595	4.187453	1.397328	0.560

Table 3: Cluster Statistics for k = 50

Cluster Index	**Index Value**	**Mean**	**Variance**	**Proportion**
6	8.5315159	9.554271	3.710242	0.432
68	4.0549596	4.222482	1.464046	0.568

Table 4: Cluster Statistics for k = 60

Cluster Index	**Index Value**	**Mean**	**Variance**	**Proportion**
99	8.5315159	9.700390	3.5413327183	0.408
210	4.190760	4.278543	1.5833833063	0.580
1	8.5520962	4.278543	0.0003587644	0.012

1.4 Co-presence Ensemble Method

When integrating the co-presence ensemble method the parameters in Table 5 were used for the algorithm. The dendrogram in figure 3 was then generated, and clearly indicates the longest lifeline with two clusters which corresponds with what was observed in figure 2. Furthermore, there is evidence of a third cluster which is barely noticeable at the bottom right of the dendrogram, when using a smaller number of clustering trials between 25 and 100, the third cluster definitely became more prominent.

Table 5: Ensemble Parameters

Parameter Symbol	**Parameter Description**	**Parameter Value**
P	proportion of original data to be randomly sampled	0.8
M	number of clustering trials	500
K	range of k-values to be randomly selected and used in k-NN clustering	k ϵ [10:80]

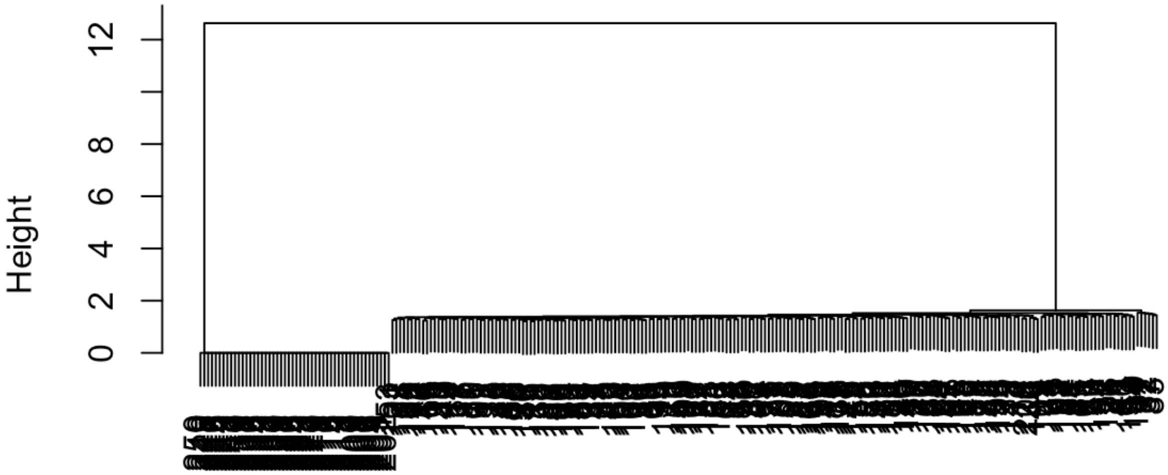


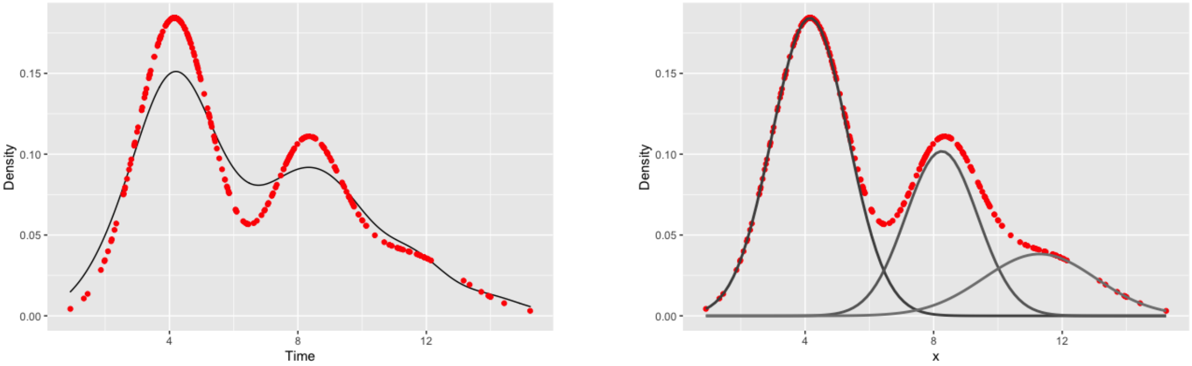
Figure 3: Dendrogram

2 Mixture Modelling

2.1 Estimated Mixture Model and Parameters

The estimated mixture model is shown in red in figure 4, figure 4a shows the mixture model overlayed on a kernel density estimate of the data, and figure 4b shows the individual distributions making up each component of the mixture model. The parameters for each of these distributions can be seen in Table 6. The final mixture model overlayed on a histogram of the data can be seen in figure 5.

Note: To reproduce these results a random seed of "2021" must be used in the mixtools library



(a)

(b)

Figure 4

Table 6: Mixture Modelling Parameters

Parameter	**Left Distribution**	**Middle Distribution**	**Right Distribution**
λ	0.546	0.286	0.168
μ	4.15	8.25	11.30
σ	1.18	1.12	1.75

2.2 Mixture Model Overlayed On Histogram

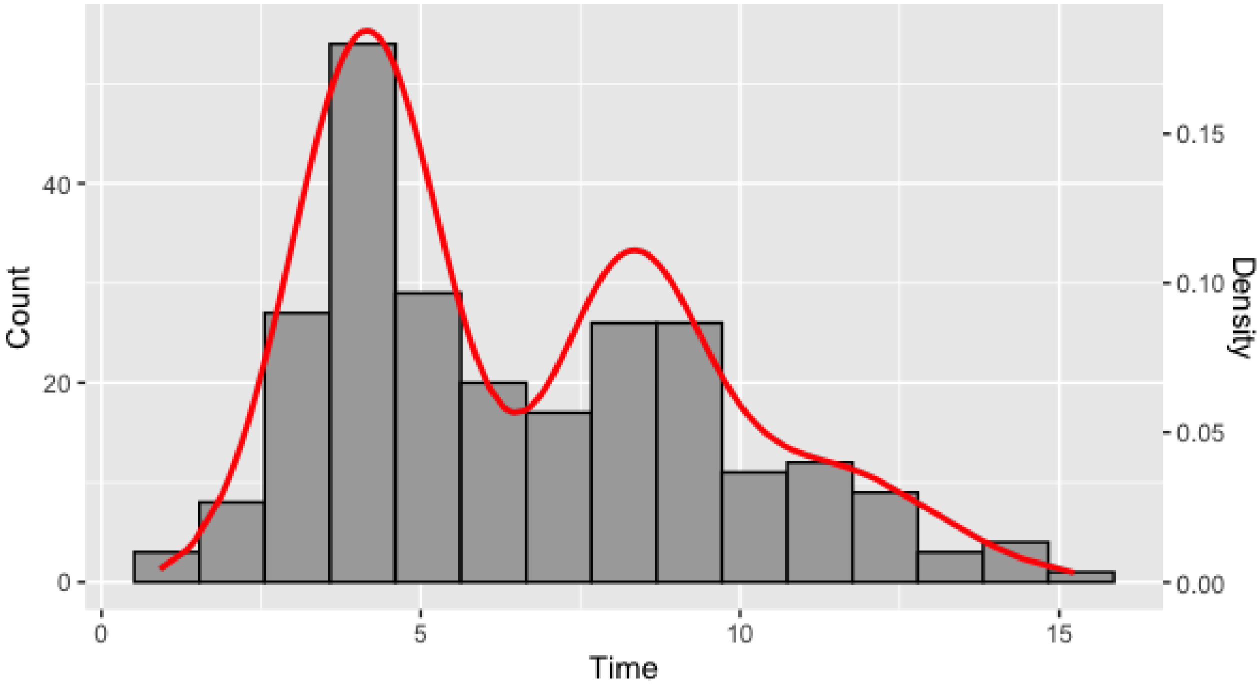


Figure 5

3 Results

3.1 Accuracy

Table 7: Model Classification Accuracy

Model	**Accuracy**
k-NN Mode Seeking	73.2%
Mixture Modelling	77.6%

3.2 Degree of correspondence

The metric used to find the degree of correspondence between the two solutions uses M iterations where a random subset of indexes is selected from the original dataset of proportion P, these indexes are then compared on their k-NN and mixture modelling classification. The number of matching classifications is divided by the sample size to get the percentage of matching classifications for the sample and this value is stored in a list. This is repeated M times and an average of all iterations is finally computed to get the degree of correspondence.

For this dataset, an M value of 10000 was used and a P value of 0.8, meaning that 10000 samples were generated of size $0.8 \times 250 = 200$. An matching percentage was calculated for each sample and the average of all sample matching percentages was used as the final figure, which happened to be 86.4%. Essentially this means that the classifications between the two models are the same 86.4% of the time.