2.3.

ALL	Case	Control	Unknown
Cluster1	8.71%	50.31%	15.28%
Cluster2	81.35%	42.44%	72.27%
Cluster3	9.94%	7.25%	12.46%
	100%	100%	100%

Table1. All features cluster K-means

Filtered	Case	Control	Unknown
Cluster1	0.00%	0.00%	62.67%
Cluster2	100.00%	100.00%	30.87%
Cluster3	0.00%	0.00%	6.46%
	100%	100%	100%

Table2. Filtered features cluster K-means

2.4

All	Case	Control	Unknown
Cluster1	9.22%	49.37%	24.38%
Cluster2	14.96%	39.14%	28.91%
Cluster3	75.82%	11.50%	46.71%
	100%	100%	100%

All features GMM

Filtered	Case	Control	Unknown
Cluster1	0.00%	97.78%	0.00%
Cluster2	19.57%	2.22%	15.90%
Cluster3	80.43%	0.00%	84.10%
	100%	100%	100%

Filtered Feature GMM

2.5

a.

For a cluster center c_{t+1} at t+1 iteration, $c_{t+1} = \frac{c_t n_t \alpha + B_t m_t}{\alpha n_t + m_t}$ $n_{t+1} = n_t + m_t$

where n_t is number of elements in cluster centered at c_t

, B_t is center of current batch, m_t is number of elements in the batch and α is a weight.

As shown in the equation, when c_t is the current center of a cluster, allocate a new data points to the nearest cluster and update the cluster center to c_{t+1}.

Due to α constant, we tend to give the most recent batch more weight in

comparison to the past data points. This is how the **forgetfulness** works in the streaming K-means.

Pros : this can be utilized in the streaming data, instead of static batch, which can reflect the data points in the real time, dynamically.
Cons: it might encounter unexpected data points, which current algorithm cannot handle, making the overall yield incorrect or unrepresentative.

b.

All	Case	Control	Unknown
Cluster1	12.91%	32.28%	36.56%
Cluster2	29.10%	58.97%	36.56%
Cluster3	57.99%	8.76%	26.87%
	100%	100%	100%

All features Streaming K-means

Filtered	Case	Control	Unknown
Cluster1	7.68%	0.00%	4.51%
Cluster2	25.10%	100.00%	68.10%
Cluster3	67.21%	0.00%	27.38%
	100%	100%	100%

Filtered features Streaming K-means

2.6

a. The purity score improved largely after the filtration of the features for all three models. The purity score is higher in K-means when used the all features, whereas GMM being higher when used the filtered features. In fact, GMM showed the highest purity when the features were filtered, out of all models, both filtered and non-filtered.

b.

k	K-means All features	K-means Filtered features	GMM All features	GMM Filtered features
2	0.47831	0.56847	0.47831	0.65022
5	0.47831	0.58779	0.58779	0.85685
10	0.70309	0.89238	0.54637	0.69783
15	0.66133	0.89169	0.62852	0.89479