2.3 b)

Table 1: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	82.2746%	41.7722%	64.3424%
Cluster 2	9.0164%	11.4979%	22.3922%
Cluster 3	8.7090%	46.7299%	13.2654%
	100%	100%	100%

Table 2: Clustering with 3 centers using filtered features

Percentage Cluster	Case	Control	Unknown
Cluster 1	91.5984%	100%	97.8462%
Cluster 2	3.0738%	0%	1.0256%
Cluster 3	5.3278%	0%	1.1282%
	100%	100%	100%

2.4 b)

Table 1: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	67.0082%	36.0759%	53.1746%
Cluster 2	16.4959%	1.6878%	16.7234%
Cluster 3	16.4959%	62.2363%	30.1020%
	100%	100%	100%

Table 2: Clustering with 3 centers using filtered features

Percentage Cluster	Case	Control	Unknown
Cluster 1	19.5697%	2.2152%	84.2051%
Cluster 2	80.4303%	0%	15.7949%
Cluster 3	0%	97.7848%	0%
	100%	100%	100%

2.5

a) Data comes in batches in streaming k-means algorithm and each batch of data is assigned to its nearest center, compute new cluster centers, updates weights by time unit and then each cluster is updated using the following formula:

$$c_{t+1} = \frac{c_t n_t \alpha + x_t m_t}{n_t \alpha + m_t}$$

$$n_{t+1} = n_t + m_t$$

ct = previous center of the for the cluster
nt = number of points assigned to the cluster
xt = new cluster center from the current batch
mt = number of points added to the cluster in the current batch
a = decay factor or forgetfulness factor

The forgetfulness factor α is used to ignore the past. When α=1 all data will be used from the beginning and when α=0 only the most recent data will be used. Recent data batches are given more weight.

Pros: Best for real time data or live stream data or new data because the cluster gets estimated dynamically. Does not depend on previous or historical data.
Cons: Since the data is real time, it can have errors in it which may cause the algorithm to yield incorrect results.

c)

Table 1: Clustering with 3 centers using all features

Percentage Cluster	Case	Control	Unknown
Cluster 1	11.9877%	31.8565%	31.2358%
Cluster 2	48.2582%	51.5823%	36.8481%
Cluster 3	39.7541%	16.5612%	31.9161%
	100%	100%	100%

Table 2: Clustering with 3 centers using filtered features

Percentage Cluster	Case	Control	Unknown
Cluster 1	7.6844%	0%	4.5128%
Cluster 2	25.1025%	100%	68.1026%
Cluster 3	67.2131%	0%	27.3846%
	100%	100%	100%

2.6

a) In the K-means algorithm, most of the of the distribution is allocated in the first cluster as compared to the rest. The reason for this behavior could be because of bias due to an imbalance in label distribution. In the GMM algorithm, data is distributed fairly randomly although in Table 2, there were a few clusters which did not have any distribution at all. The purity for Kmeans decreased from all features to filtered features but the purity increased for GMM, and Streaming KMeans.

b)

Table 3: Purity values for different number of clusters

k	k-Means All features	k-Means All Filtered features	GMM All Features	GMM Filtered features

2	0.51708	0.35426	0.51708	0.38565
5	0.58379	0.68334	0.50217	0.69231
10	0.60710	0.89548	0.61876	0.88893
15	0.68411	0.89790	0.60412	0.89100

Based on the above results purity values increases as the number of k increases for both K-Means and GMM.