Solution for Part II

Question 2) PCA has the following shortcomings
a) PCA is limited to linearity and therefore if a non-linear model produces a better solution then the latter method is more preferred.
b) PCA needs the components to be perpendicular, though in some cases, that may not be the best solution.
c) PCA assumes that columns with low variance are not useful, which might not be true in prediction setups (especially classification problem with class imbalance)

Question 3) K - means Clustering

K-Means algorithm is the process of dividing the N data points into K groups or clusters. Here the steps of the algorithm are:

1. Start by choosing K random points the initial cluster centres.
2. Assign each data point to their nearest cluster centre. The most common way of measuring the distance between the points is the Euclidean distance.
3. For each cluster, compute the new cluster centre which will be the mean of all cluster members.
4. Now re-assign all the data points to the different clusters by using the new cluster centres.
5. Keep iterating through step 3 & 4 until there are no further changes possible.

Hierarchical Clustering
Hierarchical Clustering proceeds a bit differently from the K-means Clustering method in the following ways
Given a set of N items to be clustered, the steps in the hierarchical clustering are:
1. Calculate the NxN distance (similarity) matrix, which calculates the distance of each data point from the other.
2. Start by assigning each item to its own cluster, so that if you have N items, you now have N clusters, each containing just one item.
3. Find the closest (most similar) pair of clusters and merge them into a single cluster, so that now you have one less cluster.
4. Compute distances (similarities) between the new cluster and each of the old clusters.
5. Repeat steps 3 and 4 until all items are clustered into a single cluster of size N.

Generally, for large datasets, it is preferred to used K-means clustering whereas for smaller ones we use Hierarchical Clustering. The reason for this is that Hierarchical Clustering is computationally expensive. In each iteration, it runs on every cluster that has been formed previously and store them in the memory as well. Therefore, it uses a lot of RAM and if one has limited memory bandwidth, then it becomes a problem to get good clusters.

K- means, on the other hand, runs iteratively each time without any burden on the memory, since only the new K centroids need to be stored(given that K is available to use from the start). Since it would only compute new distances in each step(by assigning each point to its nearest cluster) and not store them in the memory for further comparisons, K- means is much faster to use.