Final assignment: Supplemental Material

Akiva Bruno Melka
332629393

1 Scaling

Apart from removing the features corresponding to the DM-SNR curves, we also performed scaling of the remaining data to ensure that all would contributr equally in the clustering. However, as seen in Fig. 1, this does not provide a better representation of the data set.

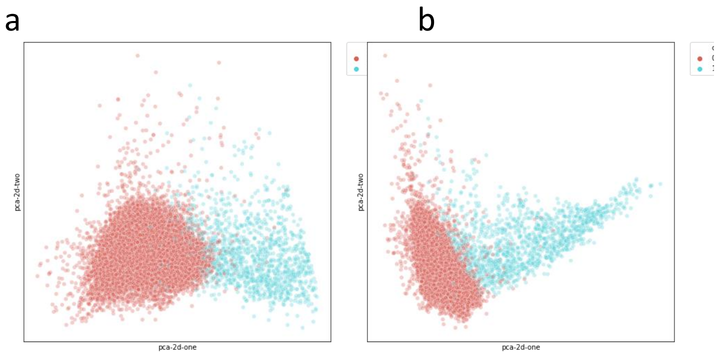

Figure 1: 2-dimension representation of the data set using PCA. Plot (a) corresponds to the data set without scaling and Plot (b) with scaling.

2 t-SNE representation

We use the t-SNE algorithm to represent the data set in 2-diemnsion. The PCA representation is more relevant (Fig. 2).

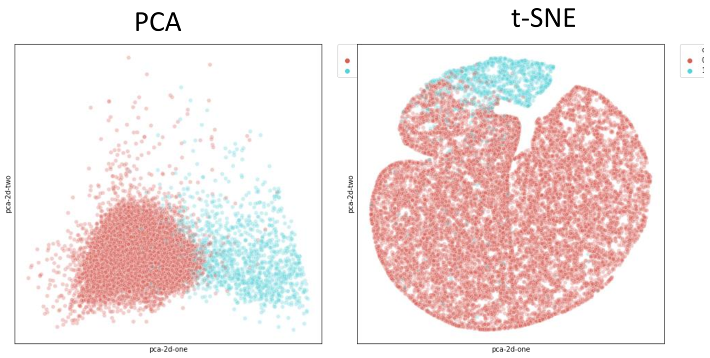

Figure 2: 2-dimension representation of the data set using PCA on plot (a) and t-SNE on plot (b).

3 Dimension reduction

Apart from using the PCA algorithm for representation, we also implemented dimension reduction to improve the clustering. As observed on Fig. 3, the 2-dimension representation remains the same.

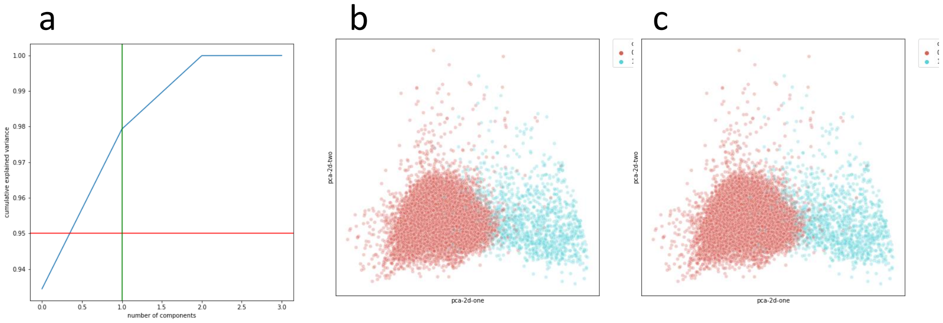

Figure 3: Dimension reduction using PCA. Plot (a) corresponds to the variance profile with respect to the number of components. Plot (b) is the representation before reduction and plot (c) after. Those two are similar.

4 Clustering with 3 clusters.

Even though the original classes were binary (1 for pulsar and 0 for non-pulsar), we implemented the clustering for "regular" algorithms with three clusters (see Fig. 4). Although the silhouette score decreased significantly from 0.57 on average to 0.39, the Mutual Information increased from 0.11 on average to 0.145.

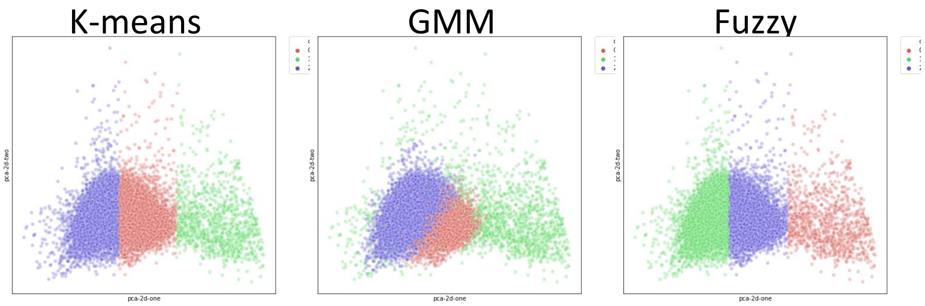

Figure 4: 2-dimension representation of the data set using PCA after clustering using 3 clusters.