Applied Machine Learning:

Goal: Implement clustering algorithms + apply feature dimensionality algorithms

Breast cancer dataset:

About the problem –

Breast cancer typically forms in the cells of the breast & is the most common cancer diagnosed in women in the United States. What I wasn't aware of is that breast cancer can occur in men as well. The dataset being used contains variables which were recorded during observation.

The signs & symptoms may include –

1. A breast lump or thickening that feels different from the surrounding tissue
2. Changes in size, shape or appearance of a breast
3. Changes to the skin over the breast, such as dimpling
4. A newly inverted nipple
5. Peeling, scaling, crusting or flaking of the pigmented area of skin surrounding nipple
6. Redness or the pitting of the skin

Doctors estimate that about 5 to 10 percent of breast cancer cases are linked to gene mutations passed through generations of a family. While these signs can be observed at home, there are also alternative proactive steps which can be taken to minimize the chances of breast cancer –

1. Breast cancer screening
2. Performing breast self-exam for breast cancer awareness
3. Drinking alcohol in moderation
    i. No more than one drink a day
4. Exercise for most days of a week
    i. At least 30 minutes of exercise
5. Limit postmenopausal hormone therapy
6. Maintain a healthy weight
7. Maintain a healthy diet

Reference –

https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470

The dataset in use, has variables which were recorded for patients upon their arrival at a Wisconsin clinic. To predict a malignant or benign case of breast cancer, we can look at the following information provided by the variables:

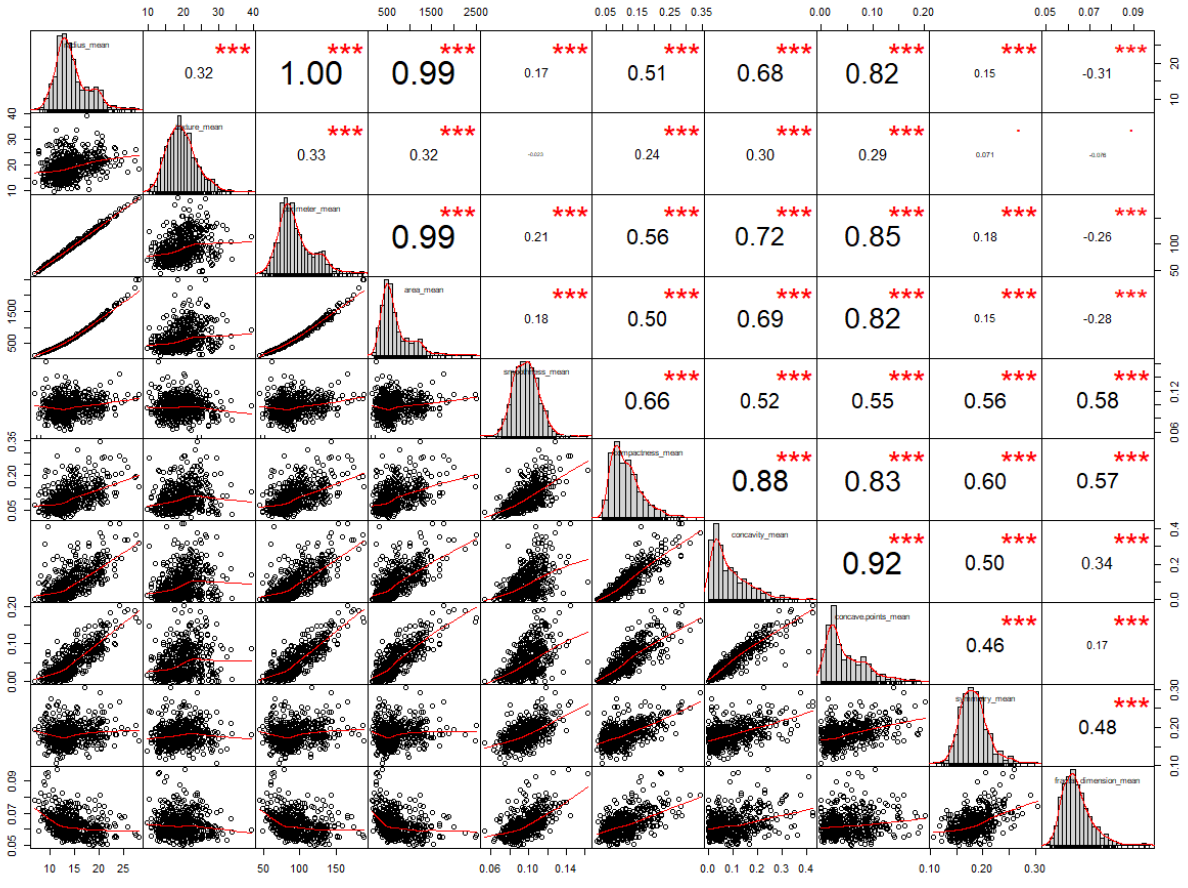
1. Breast lump or thickening –
    a. texture_mean
    b. smoothness_mean
    c. concavity_mean
    d. concave.points_mean
2. Changes in size/shape:
    a. radius_mean
    b. area_mean
    c. perimeter_mean
3. Inversion of nipple:
    a. symmetry_mean
4. Growth of tumor:
    a. fractal_dimension_mean

Recorded cases:

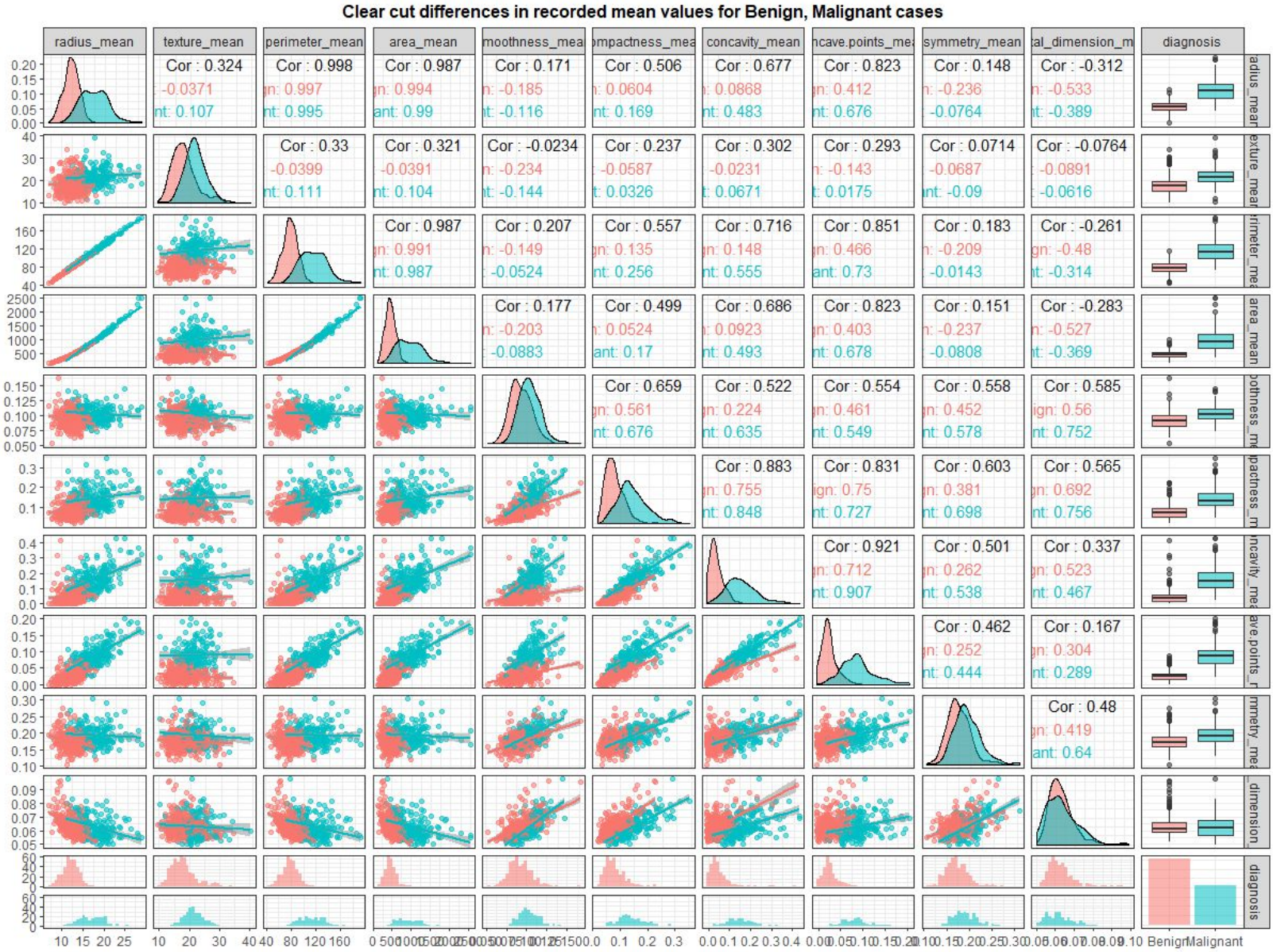
357 - Benign 212 - Malignant

Since these were values were recorded, these variables were also provided with their standard errors & the worst value recorded for each case. There was a high degree of correlation observed in the dataset.

Between mean values



There are only 2 clusters in this dataset – Benign cases, Malignant cases. I wanted to check if there were features that displayed clear differences in measurement before proceeding with clustering models + feature dimensionality algorithms on neural networks. I plotted a correlation chart with the features & the diagnosis recorded. Clearly, there are differences in mean values observed:



Due to the number of features, I decided to run PCA to see the number of principal components enough to capture variance.

6 Principal components were enough to capture 87% of the variance.

The variables most correlated with the PCA dimensions helpful to capture variance were –

Dimension 1 –

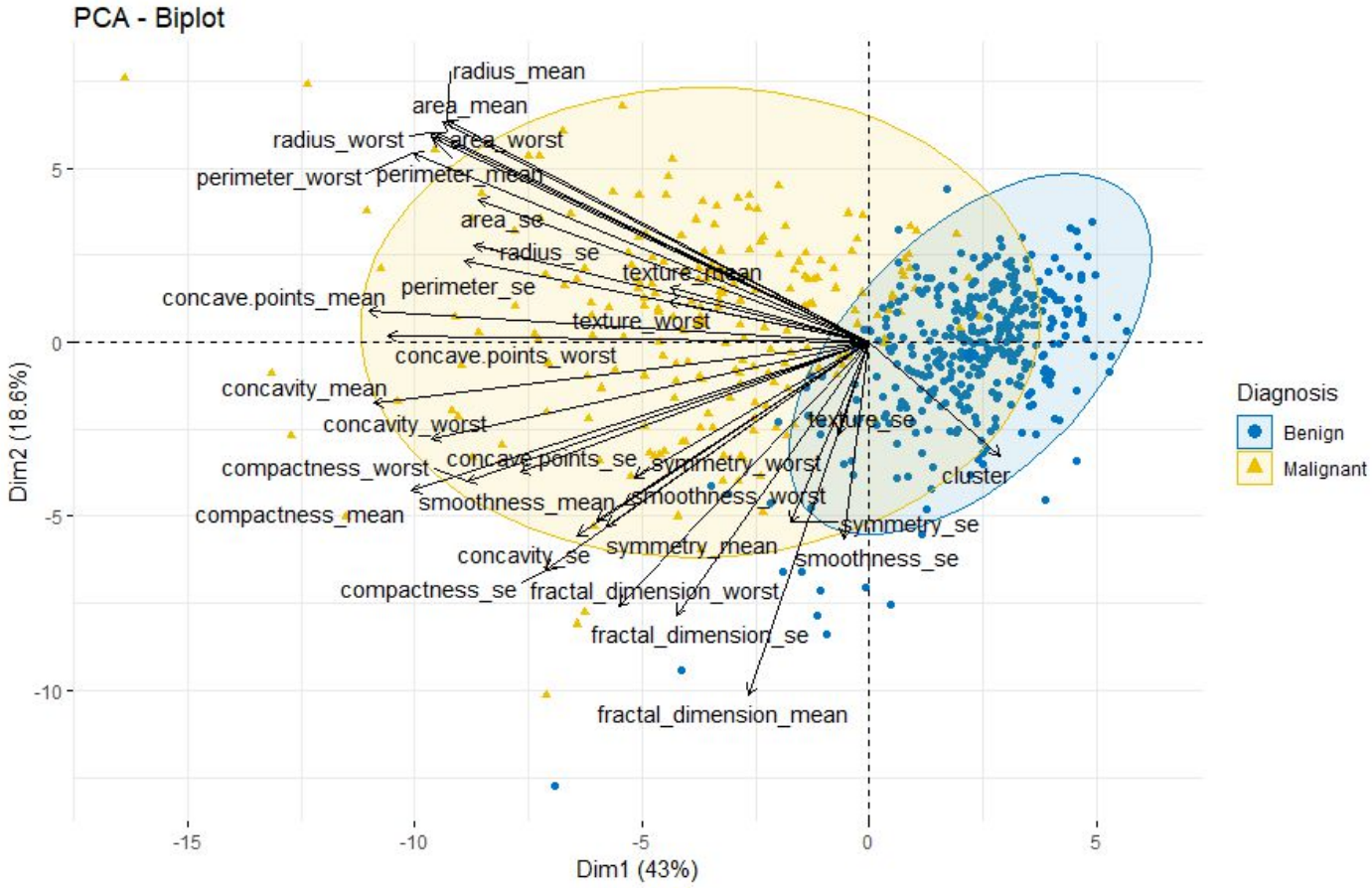
Radius mean – Perimeter mean – Area mean – Compactness mean – Concavity mean - Concave points mean – Radius standard error – Perimeter standard error – Area standard error – Radius worst – Perimeter worst – Area worst – Compactness worst – Concavity worst – Concave points worst

Dimension 2 –

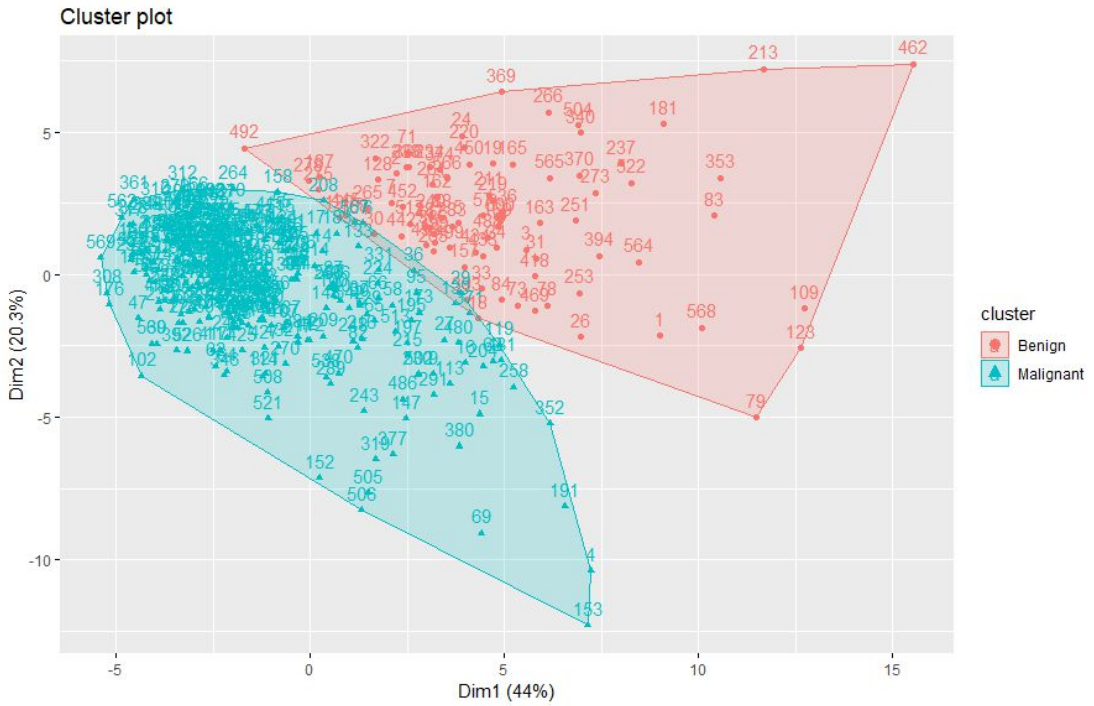
Fractal dimension mean

Dimension 4 –

Texture mean – Texture worst



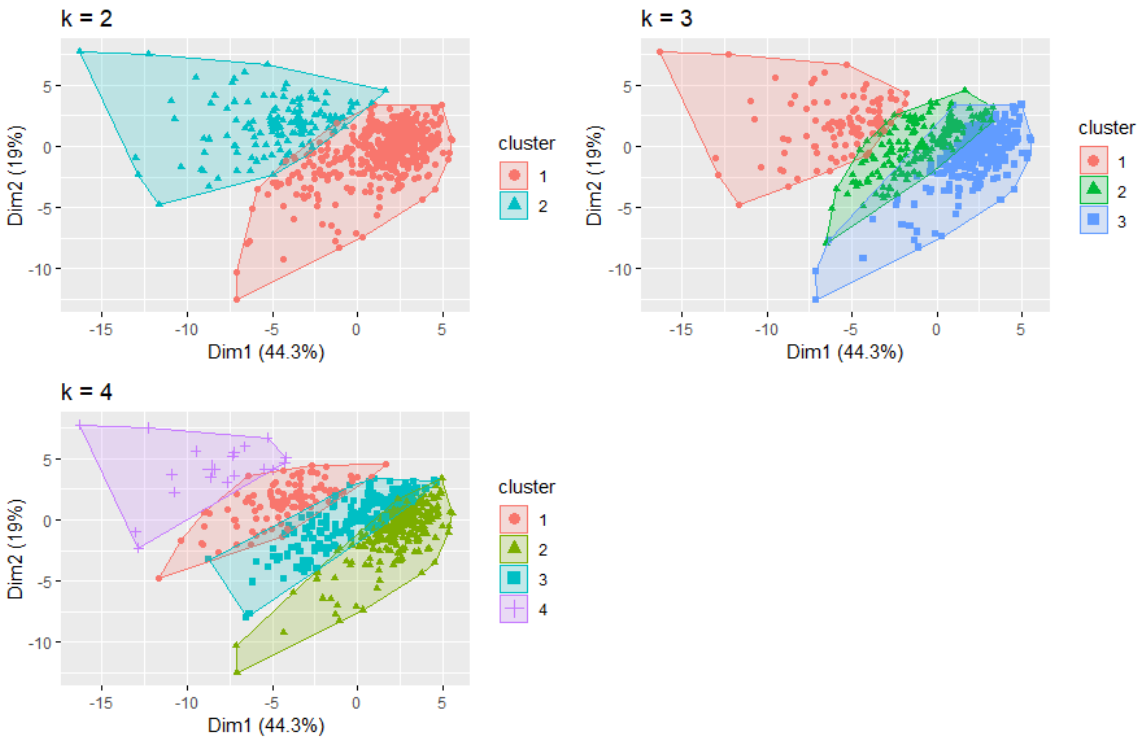
I tried to implement K – means using a predict function I found online & with these 2 clusters. For reasons I couldn't figure out, the accuracy was simply 18.3% in the confusion matrix – there is a coding error I couldn't fix where Benign had moved into Malignant in the table. But there were 2 clear clusters:



After this, I implemented the random projections algorithm. I needed 3 functions which I found online –

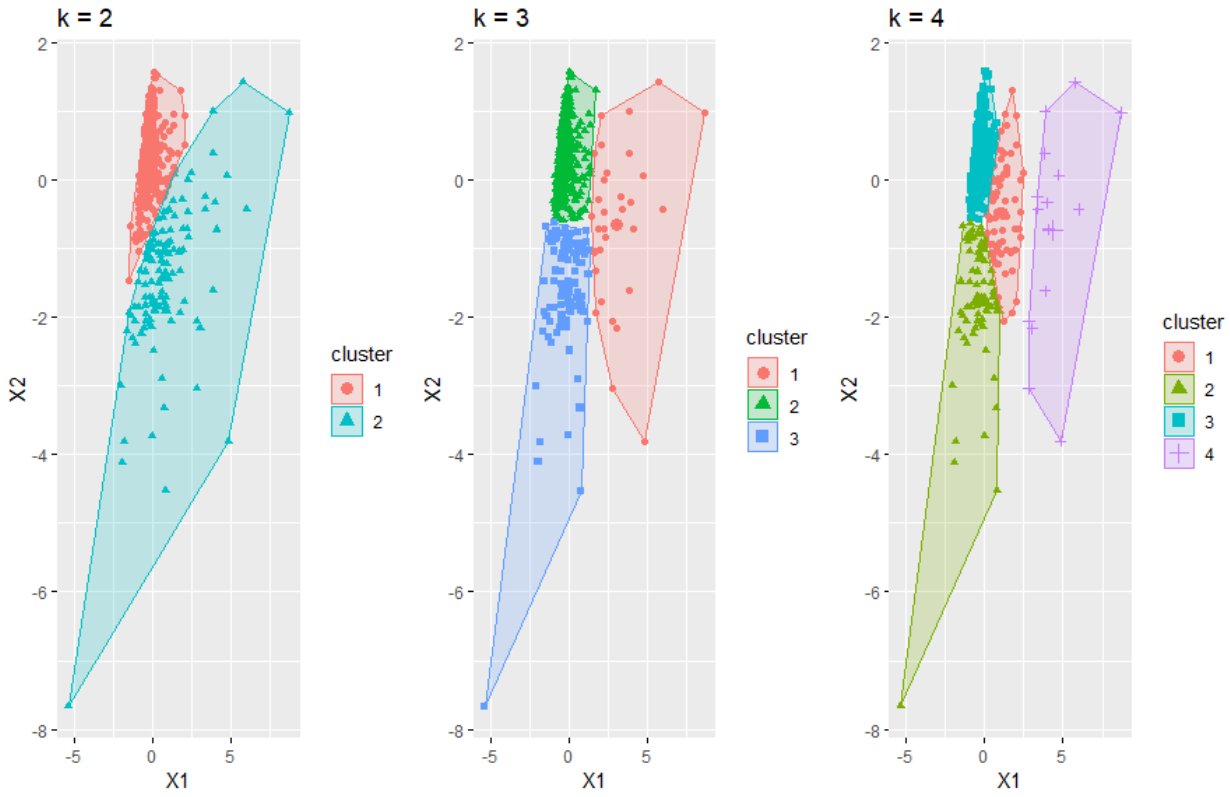
1. Johnson Lindenstrauss Min Dim –
    a. On reading the user guide, this function helps us estimate conservatively the minimal size of the random subspace to guarantee a bounded distortion
2. Random matrix generation & random projection

A within – sum – of squares plot revealed an elbow at 2. I ran the k-means function for 2,3 & 4 clusters. Clearly this dataset has 2 clusters. There is an overlap in observations beyond 2 clusters & this isn't appropriate for this dataset.



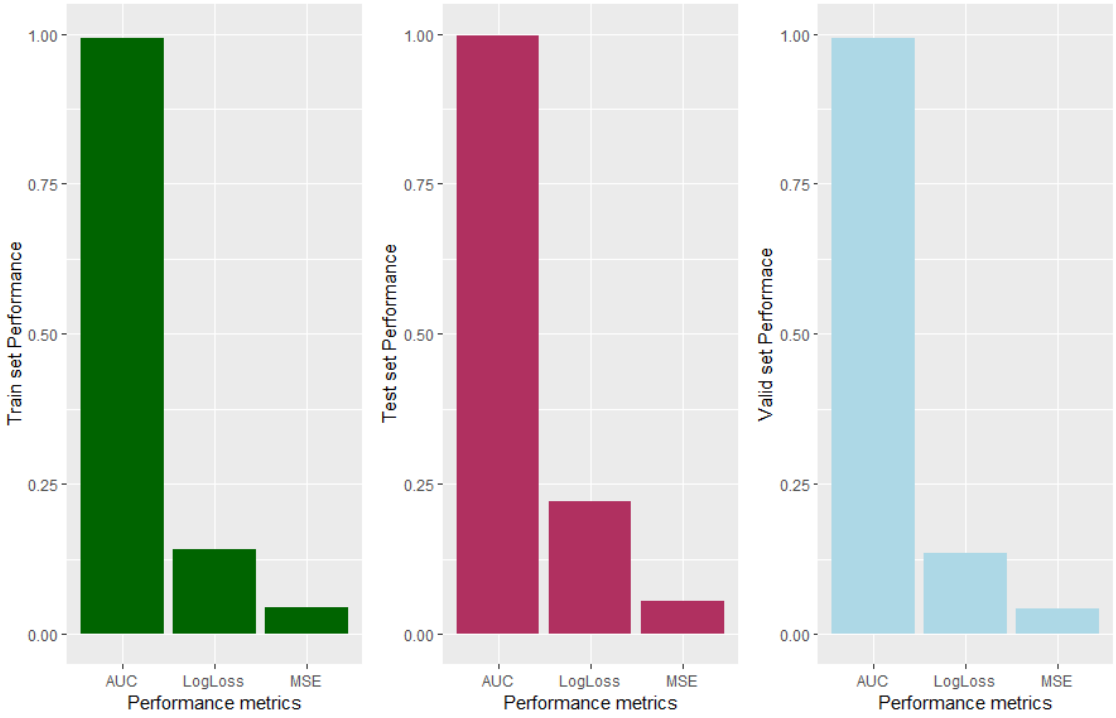
On running Boruta variable importance method to reduce the number of features for my dataset & then perform K-means, I observed overlapping observations between clusters. As mentioned before, this wasn't appropriate.

Then I performed dimension reduction using ICA. The observations were now more spread out, but 2 clusters were clear with a boundary separation.



Next, I ran expectation maximization algorithm. It suggested an optimal value of 8 clusters.

Next, I implemented neural networks on PCA & obtained AUC of 1.00.



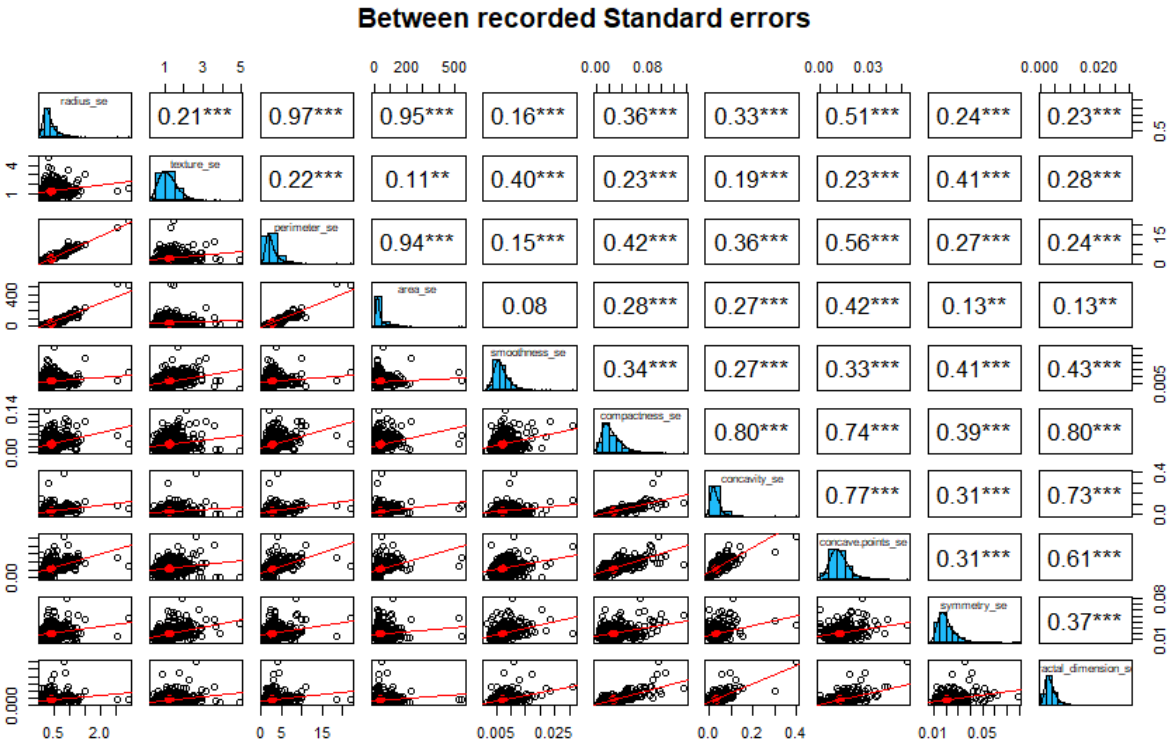
Neural networks implementation, observed AUC -

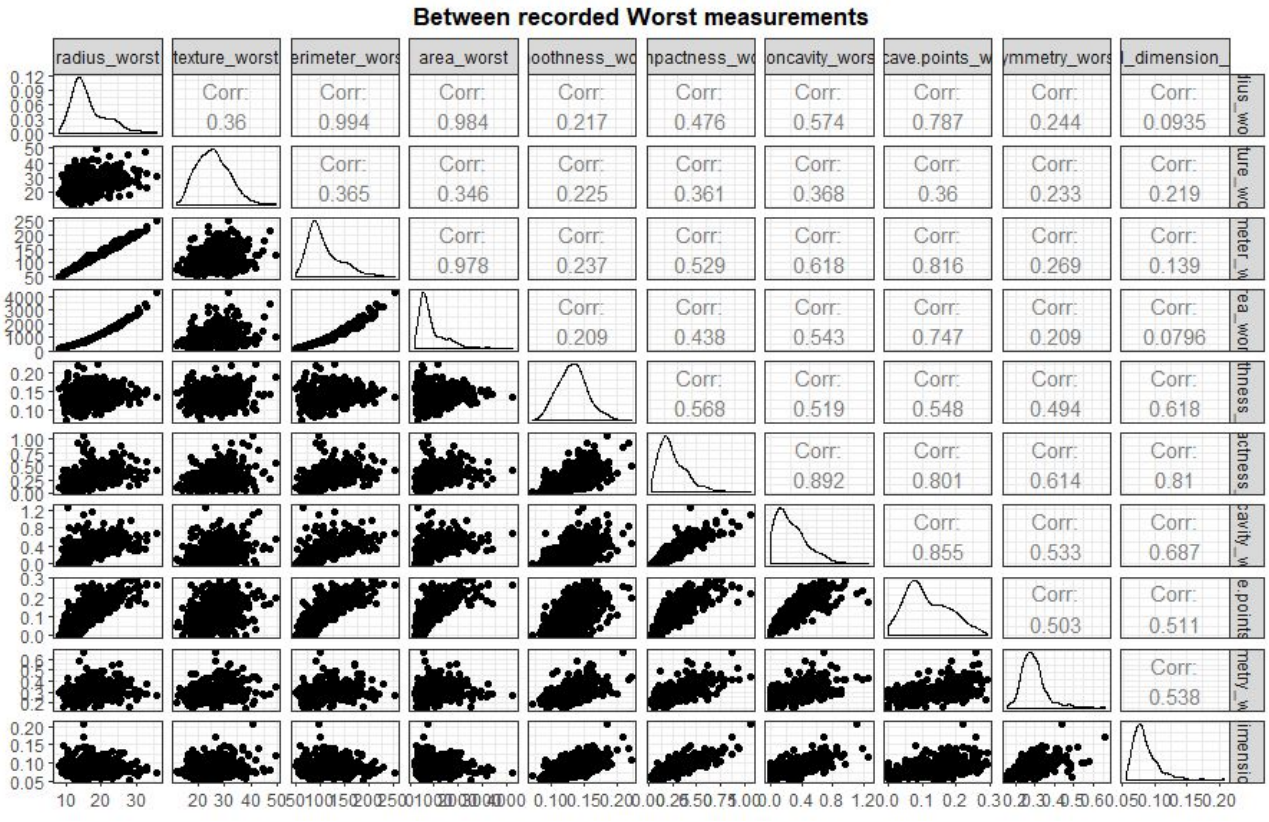
Algorithm	Training	Validation	Test
PCA	0.9928	0.9967	0.9928
ICA	0.9782	0.9924	0.9801
Random Projection	0.9863	0.9978	0.9871
K-means clustering	0.8348	0.9090	0.8341
Expectation Maximization	0.50	0.50	0.50

I implemented the algorithms as h2o objects, which was the easiest method to deploy on neural networks. In fairness, I wanted to run algorithms separately & write code separately for both datasets

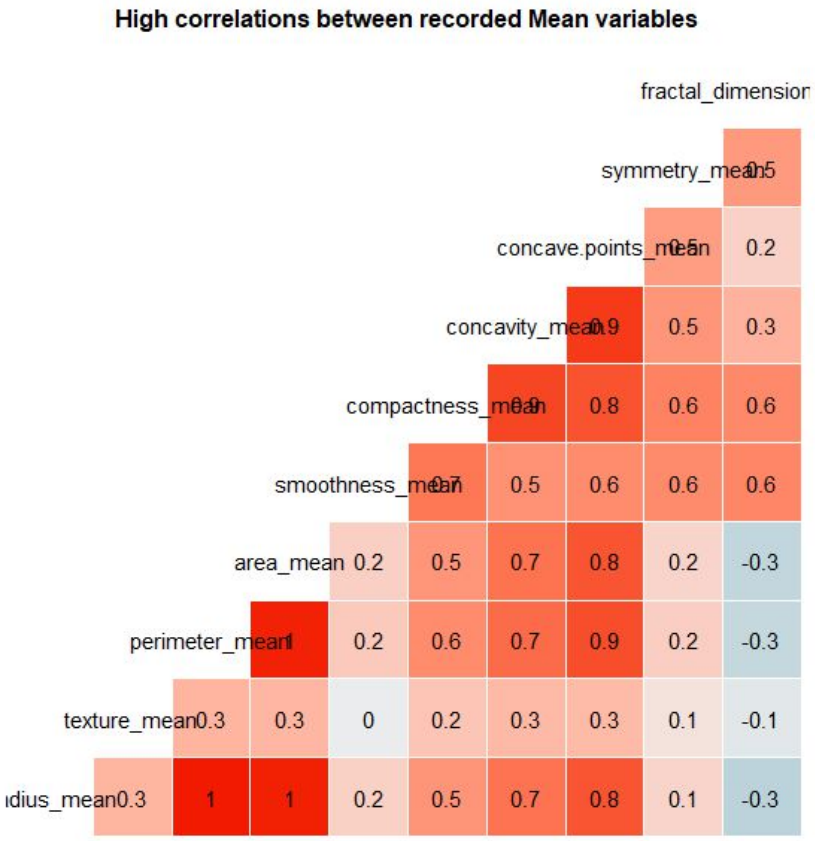
Further plots.

Wisconsin Breast Cancer dataset -

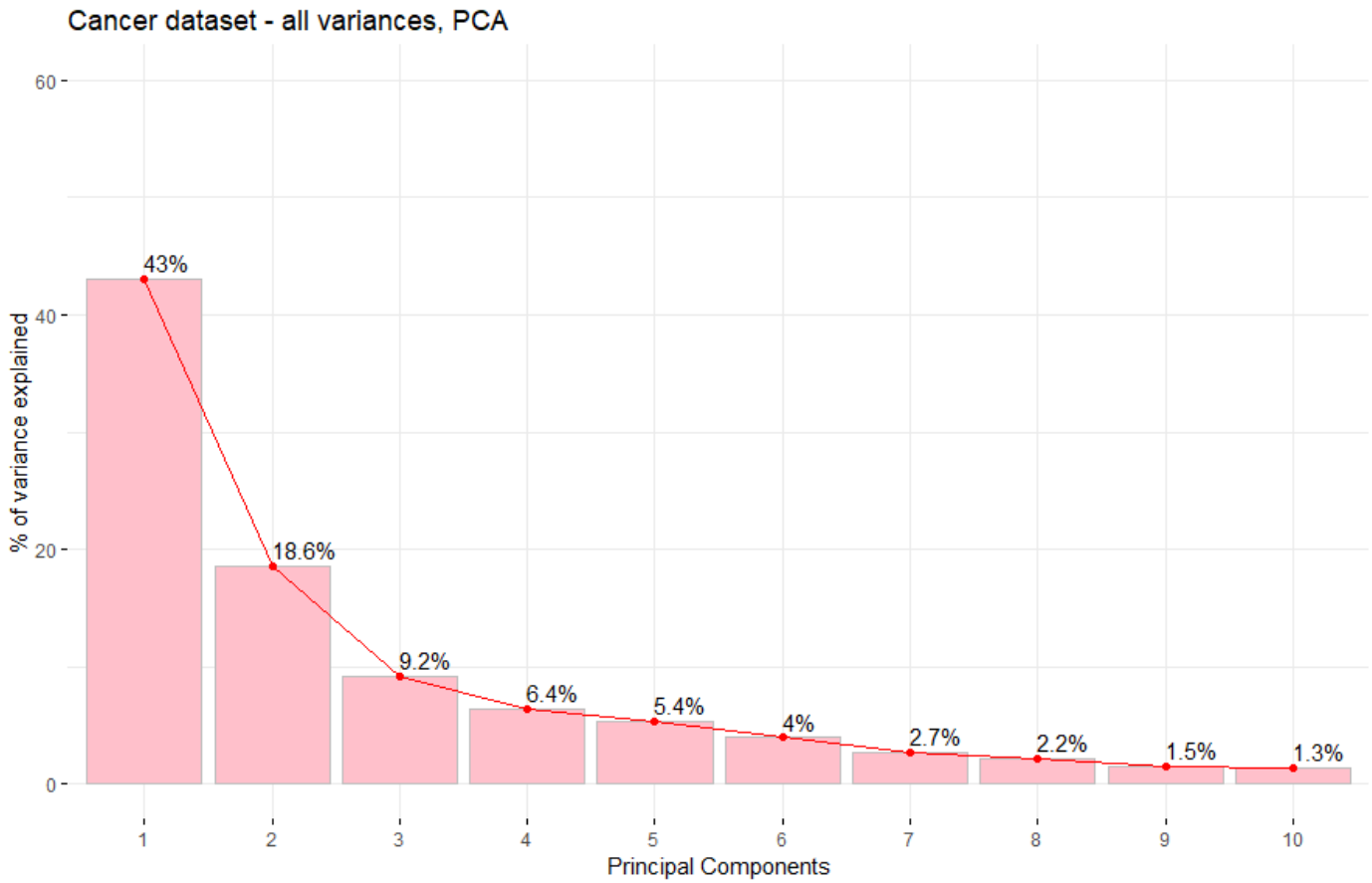




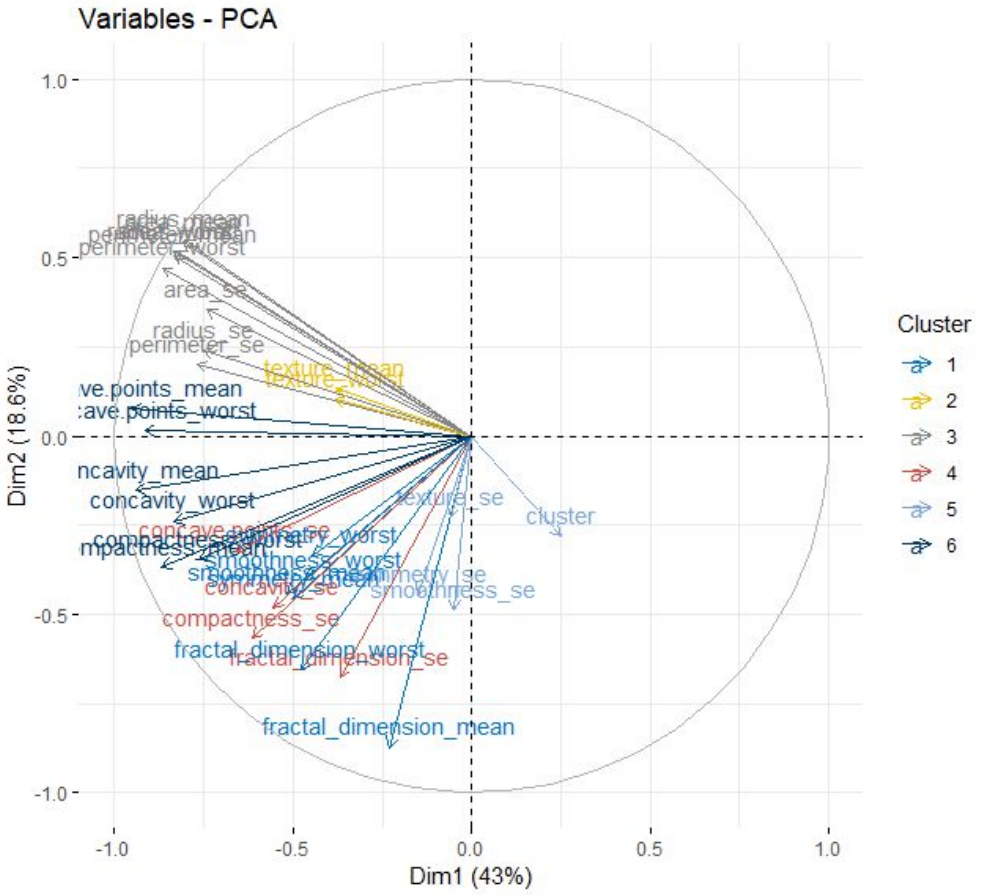
Presence of high correlation within the dataset:



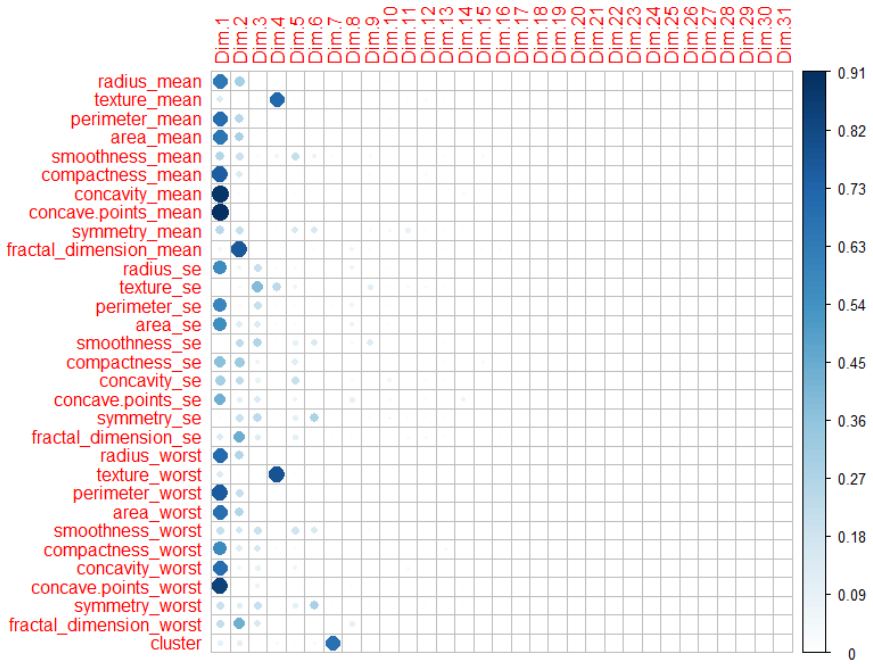
PCA to observe the number of principal components needed –



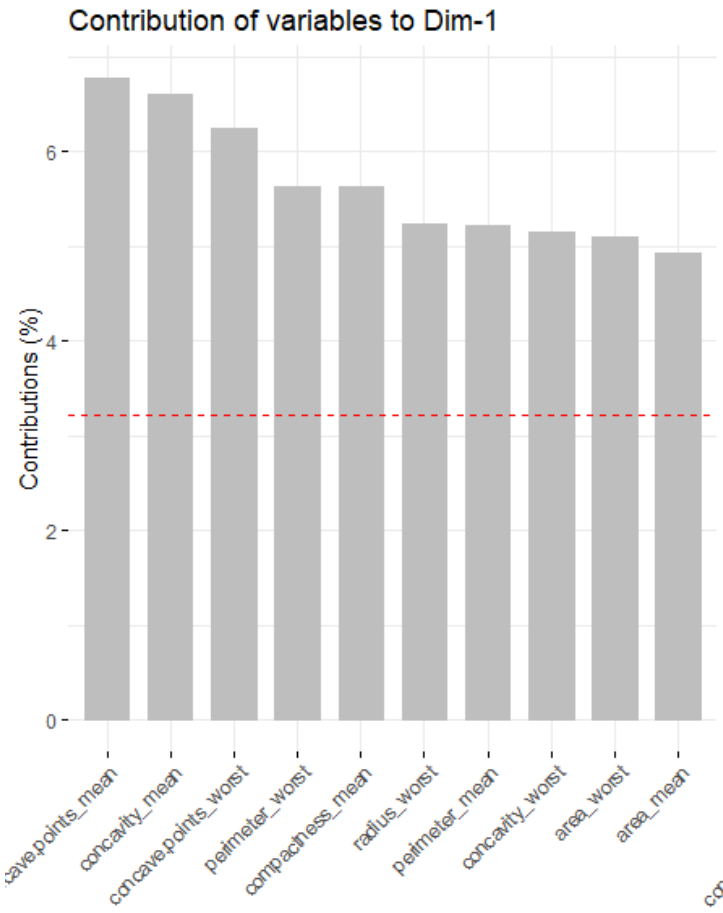
The biplot here clearly shows 2 distinct directions for the clusters –

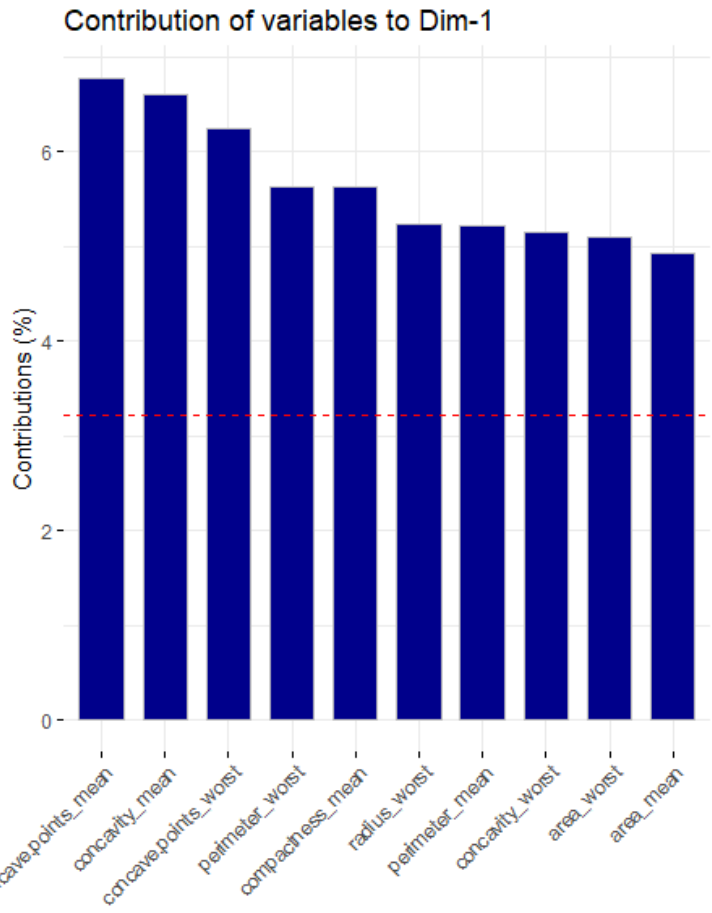


Cos2 for the variables observed in PCA –

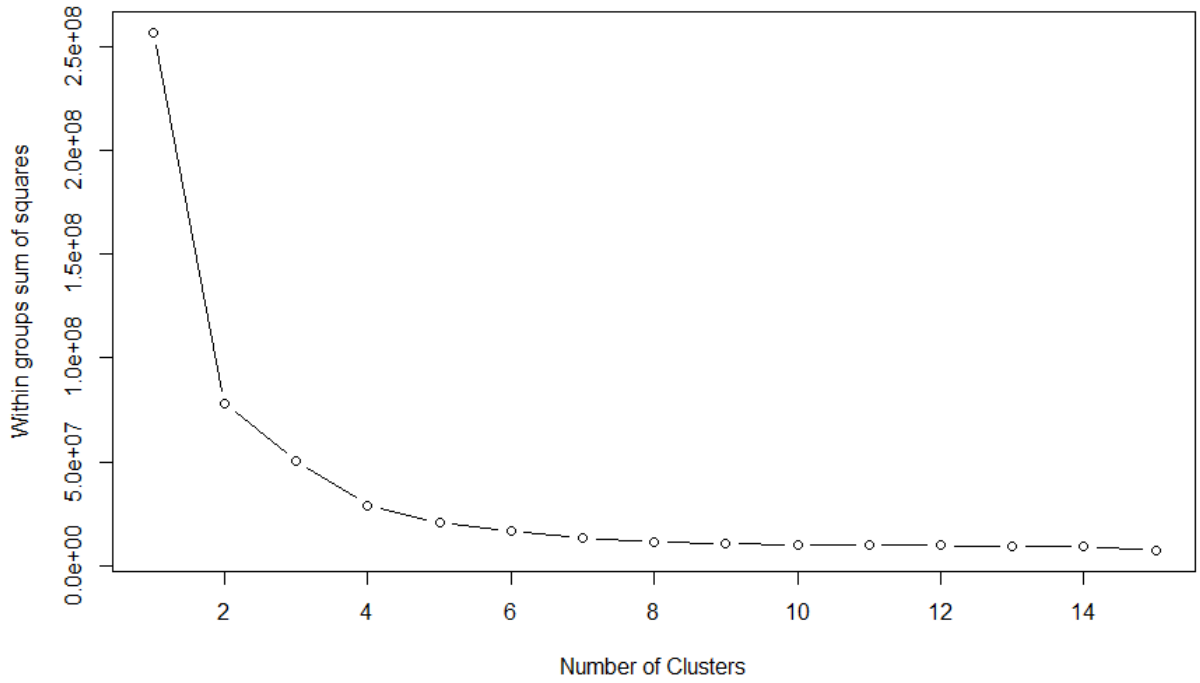


Here was the variable contribution in both the dimensions. I'm guessing that the same variables important in both dimensions means that there are clear cut differences for both these clusters.

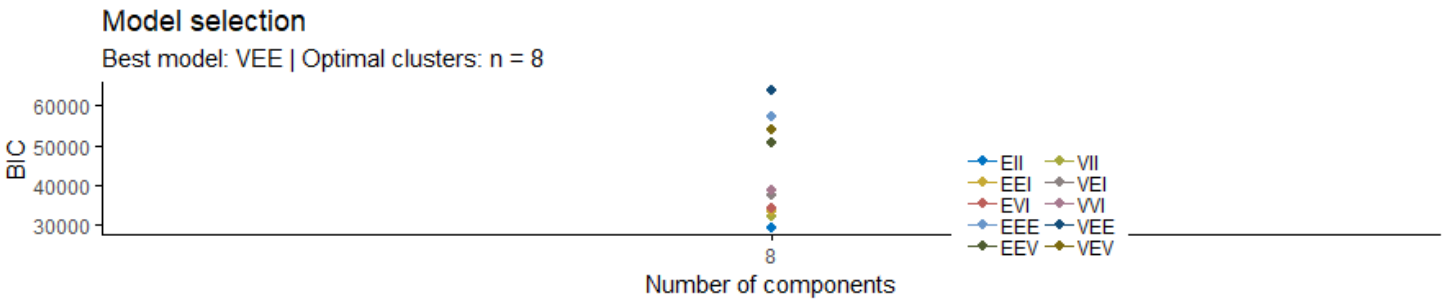


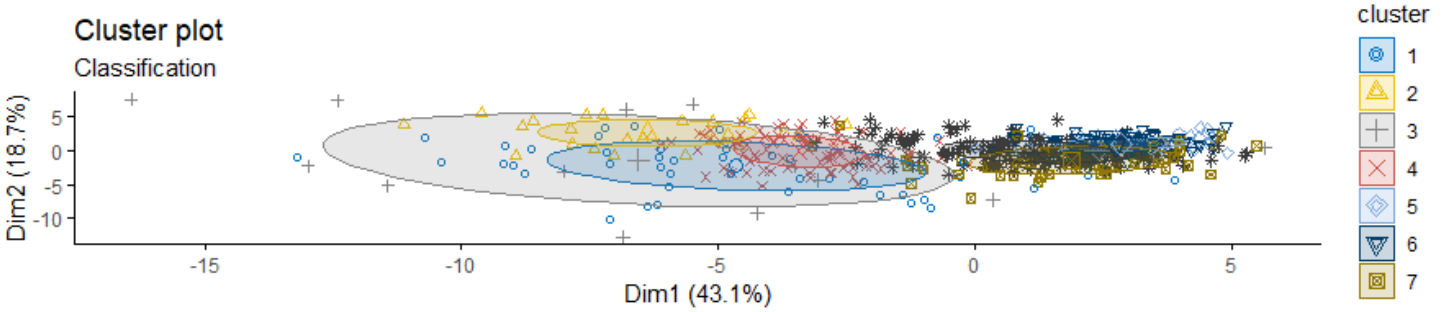


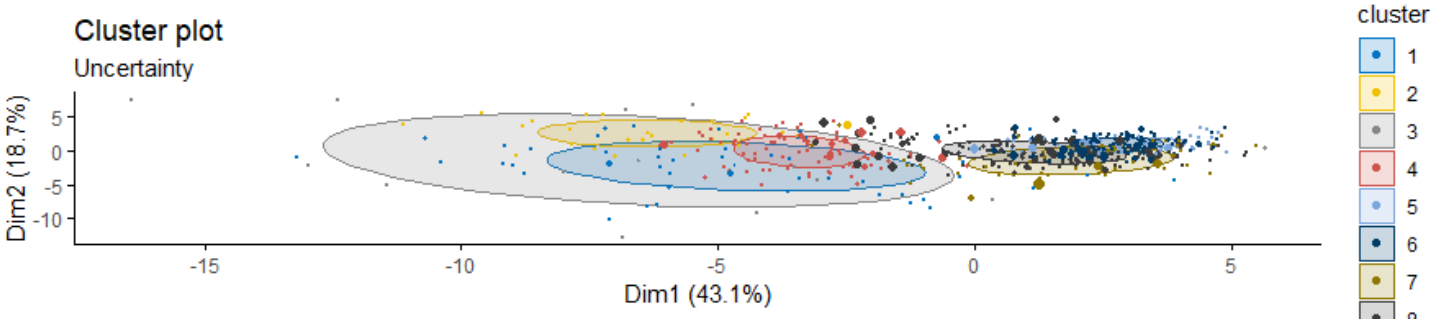
Elbow plot revealed decreasing within groups sum of squares beyond 2 clusters:



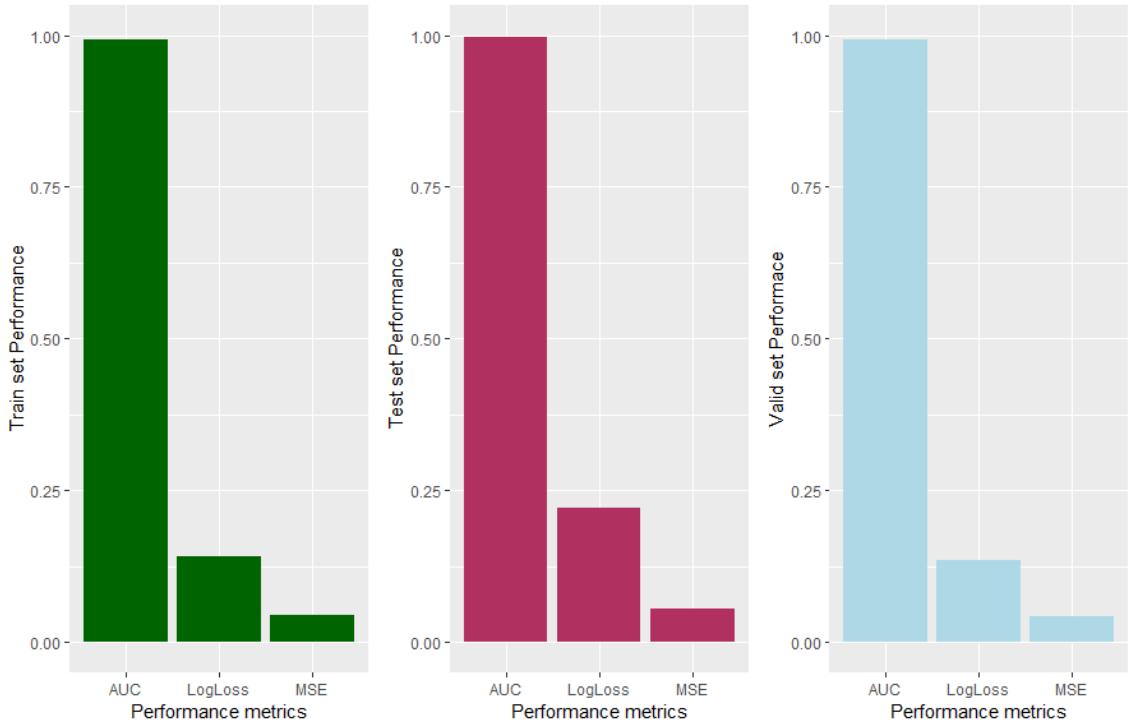
I wasn't successful implementing expectation maximization & it revealed 8 clusters.







PCA on Neural Network –



ICA on Neural Network –