Applied Machine Learning – Assignment 4

Goal: Implement clustering algorithms + apply feature dimensionality algorithms

Secondary school student performance:

It wasn't as clear cut to observe the number of clusters in this dataset. I initially assumed there would be groups of students which could be simply split upon their grades, addresses & school. This wasn't the case.

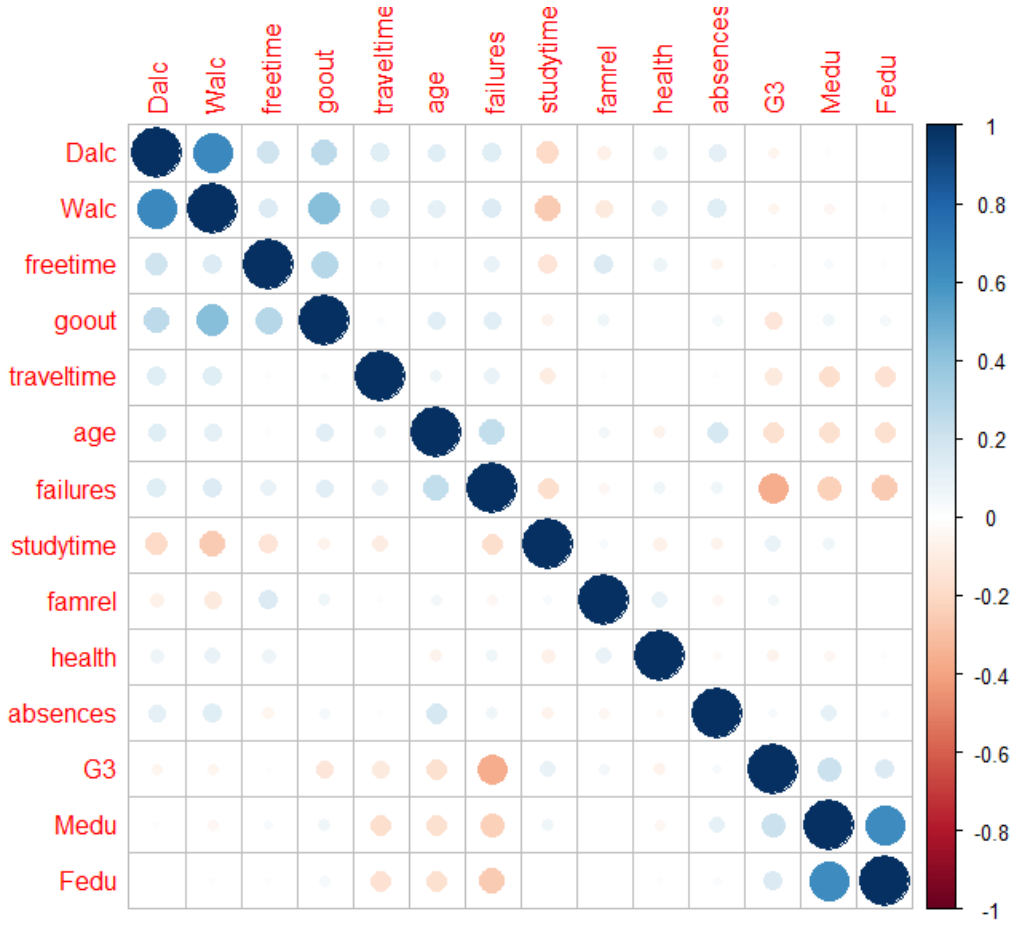
349 – Gabriel Pereira

46 – Mousinho da Silveira

Correlation plot –

I used the following variable selection methods –

1. Boruta
    a. Important variables – Fedu, failures, schoolsup & higher
2. Recursive partitioning



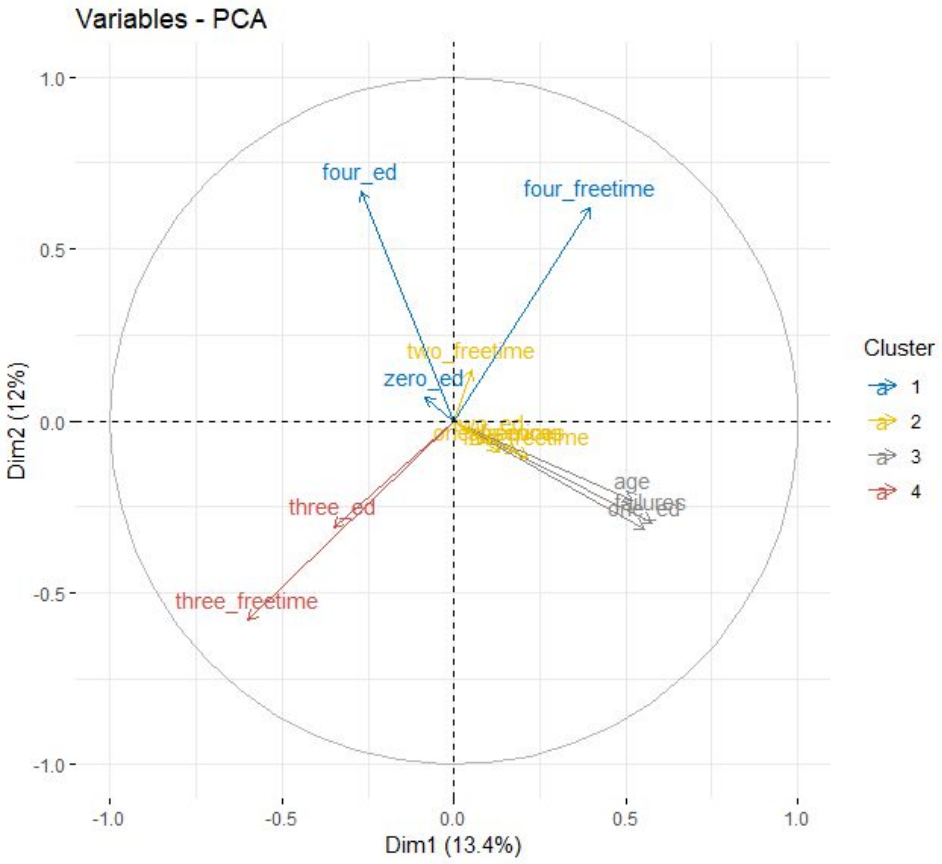
    a. Important variables – only failures
3. Recursive random forest
    a. Important variables – failures, absences, age, Fedu

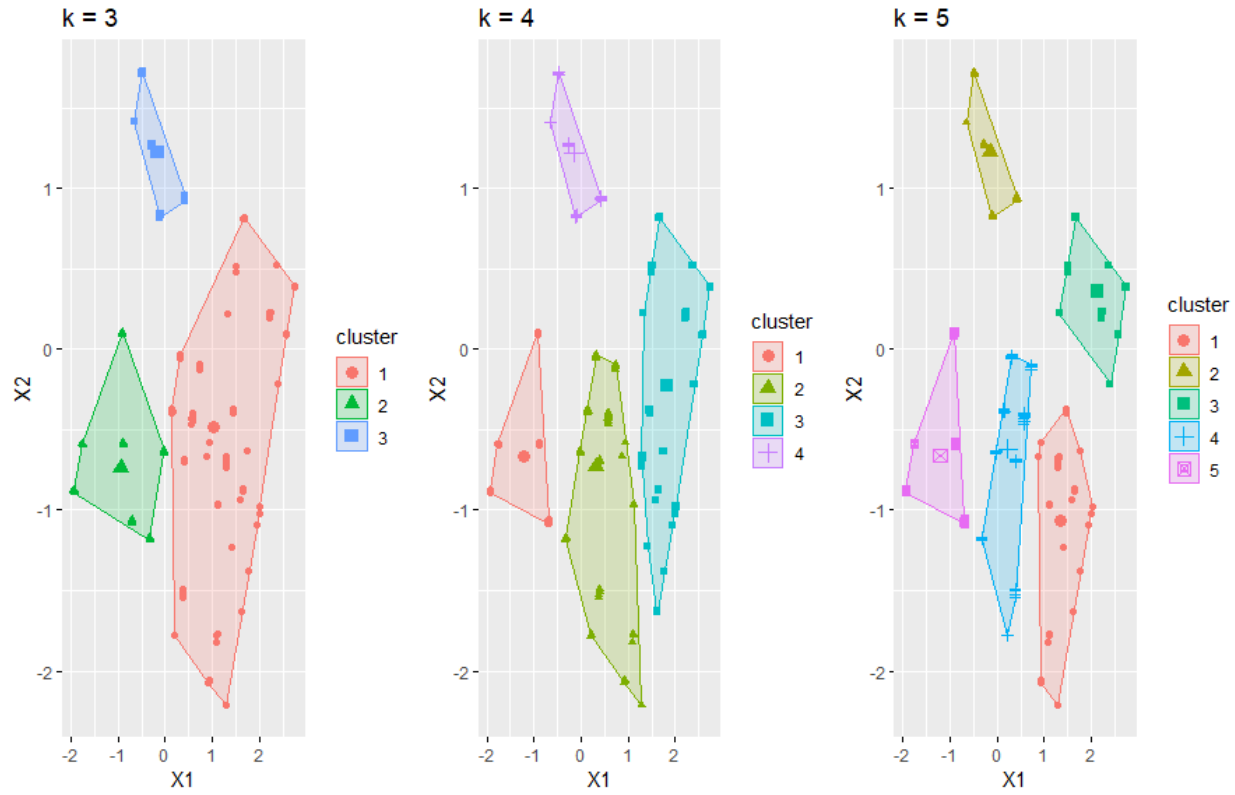
Based on this & wanting to use numeric variables out of the 33 variables, I used freetime, Fedu. Age, failures, absences to predict student's failure or passing.

For this dataset, 8 principal components gave a variance of 89%.

There were 4 clusters in the dataset based on the PCA output & these were the directions output for the variables -

Next, I implemented dimension reduction using ICA. I could observe 3 clear clusters based on this output & it seemed like I was overfitting with 4 clusters.

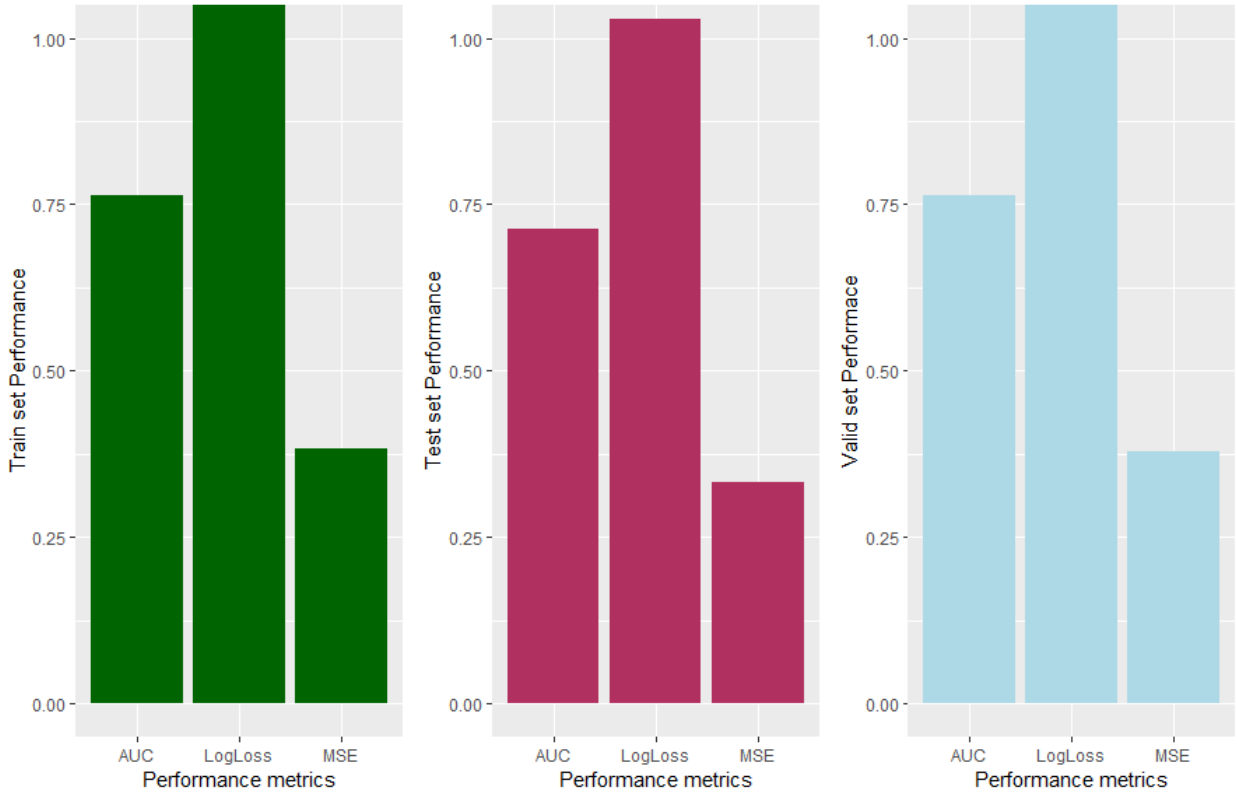




Expectation Maximization algorithm gave an output of optimal clusters of 8. I wasn't sure of this being a correct representation of the data.

Running Neural networks:

I couldn't obtain an accuracy greater than 60% in the previous assignments. On running PCA output on neural network, I managed to get an accuracy of 75%.



Neural networks implementation, observed AUC -

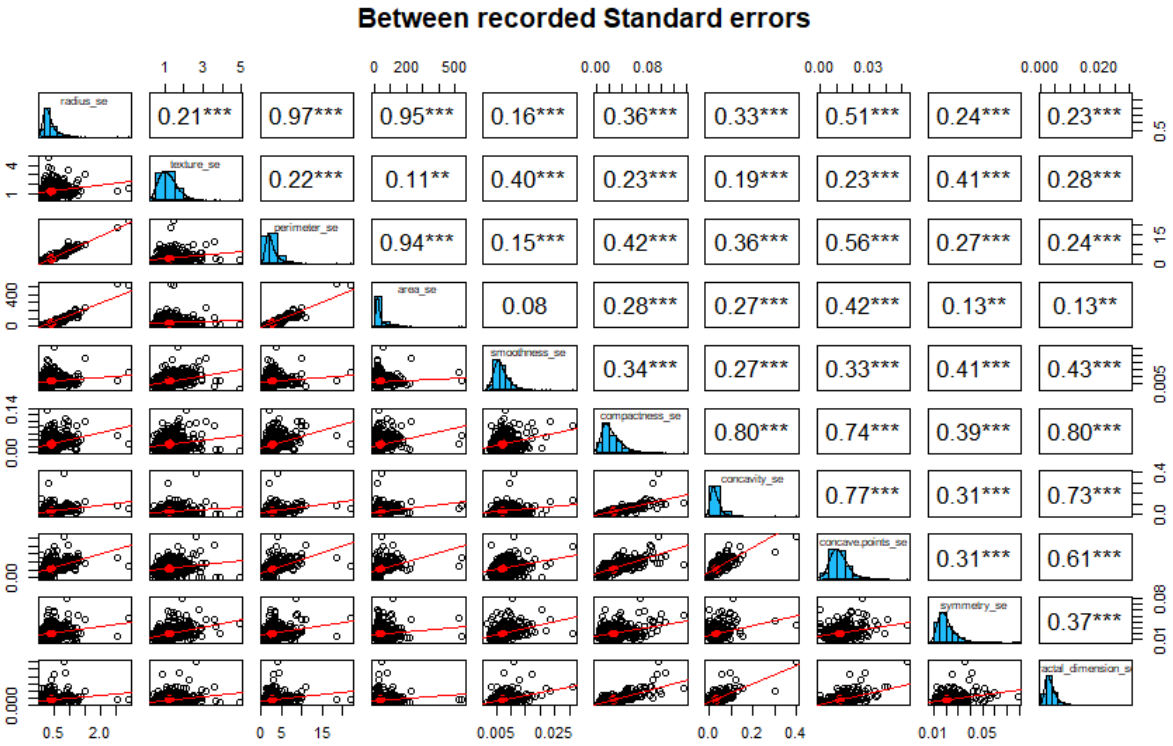
Algorithm	Training	Validation	Test
PCA	0.7627	0.7132	0.7627
ICA	0.5926	0.6370	0.5755
Random Projection	0.6105	0.7105	0.6883
K-means clustering	0.5939	0.6703	0.5939
Expectation Maximization	0.50	0.50	0.50

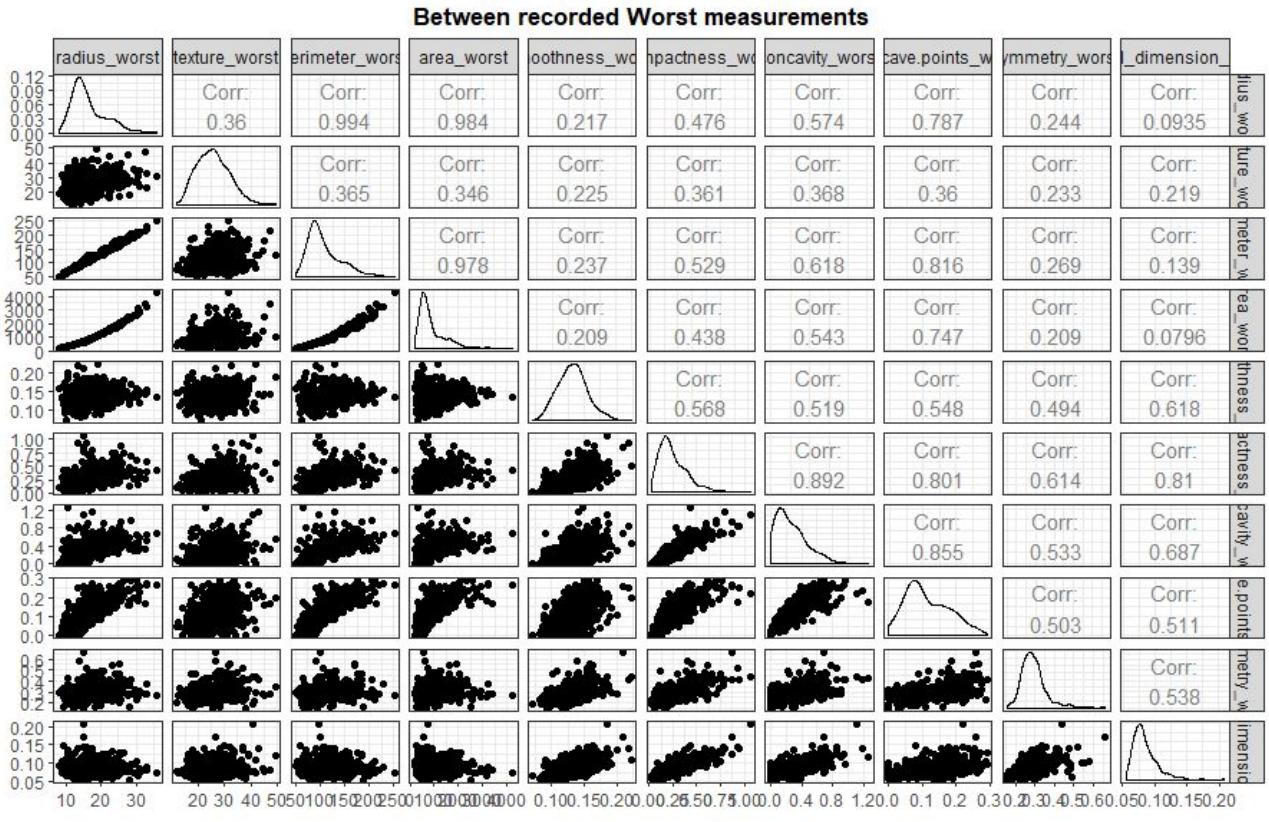
--

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

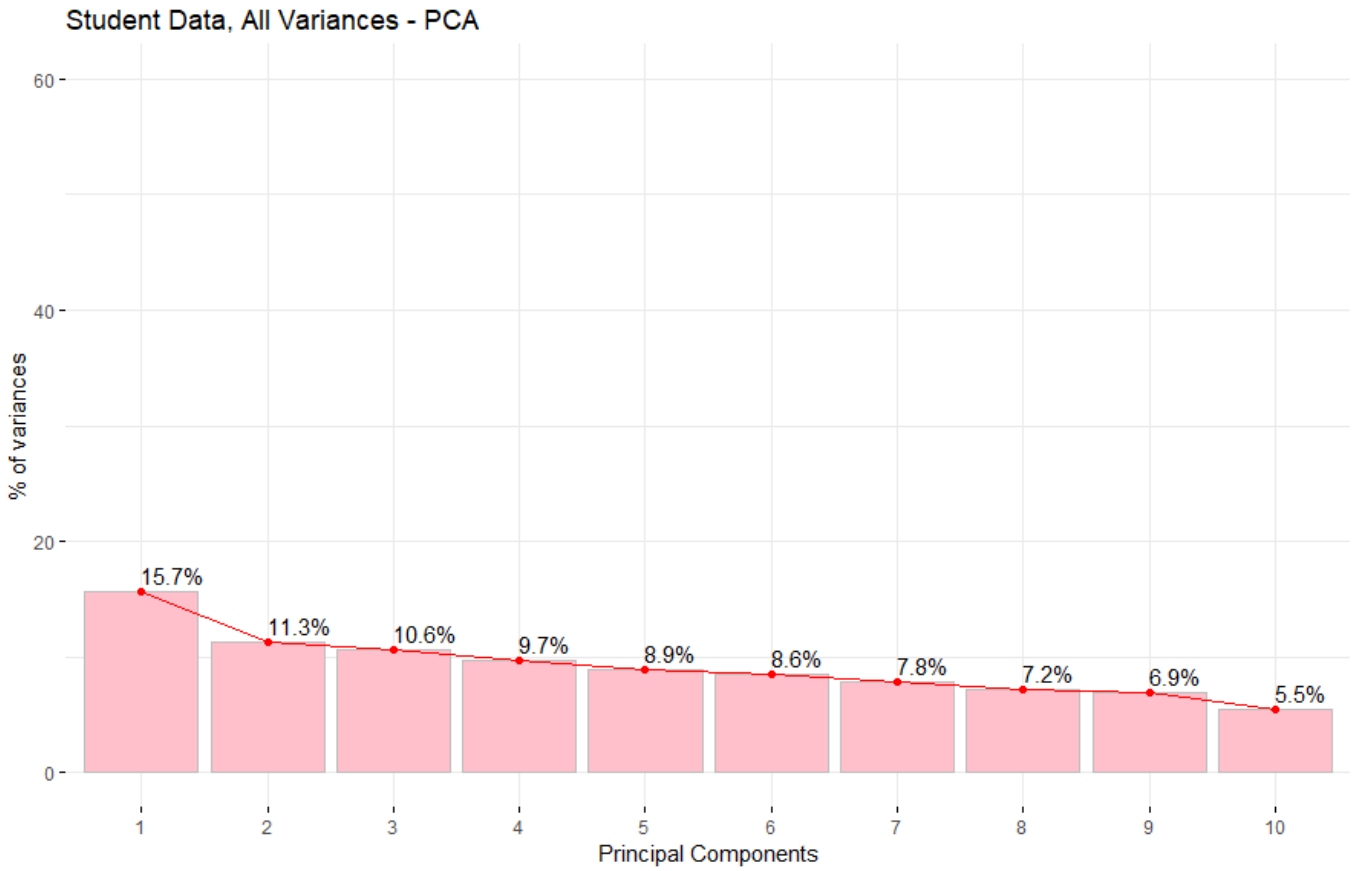
Further plots.

Student dataset –

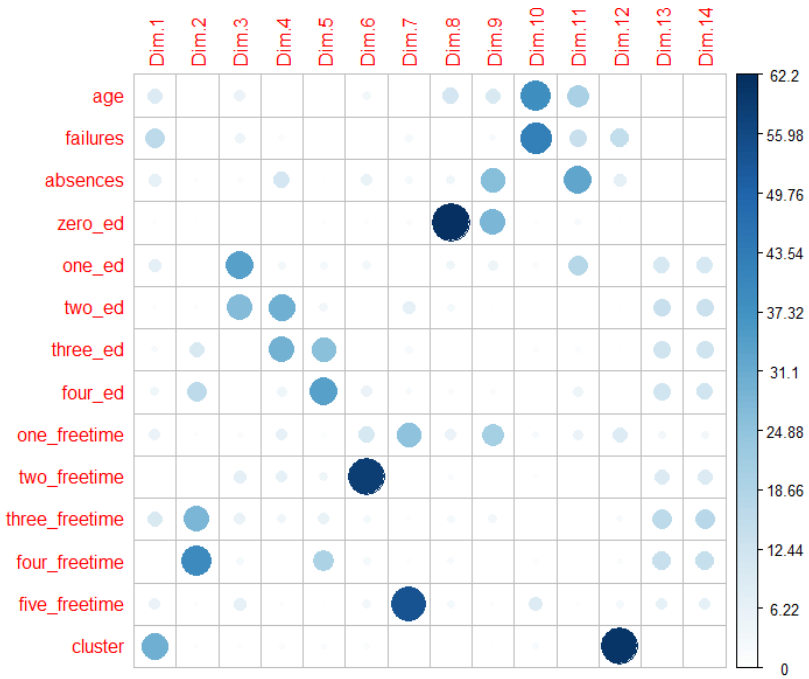




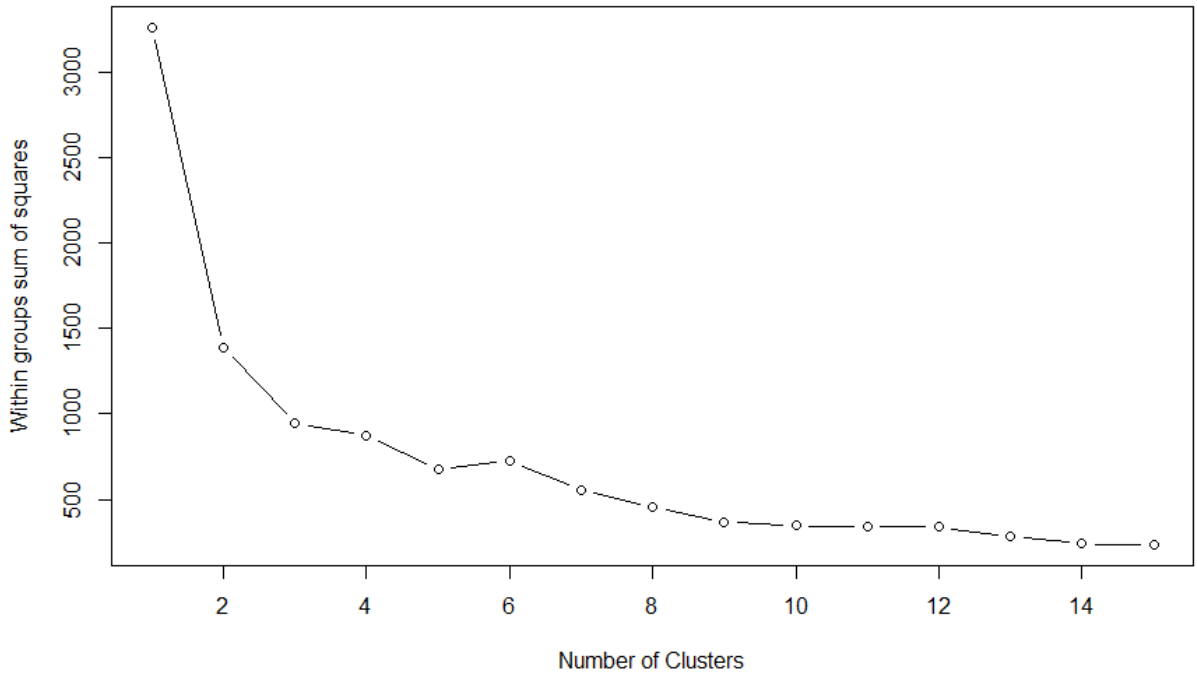
Principal components contribution:



Contributions to PCA
by variables:

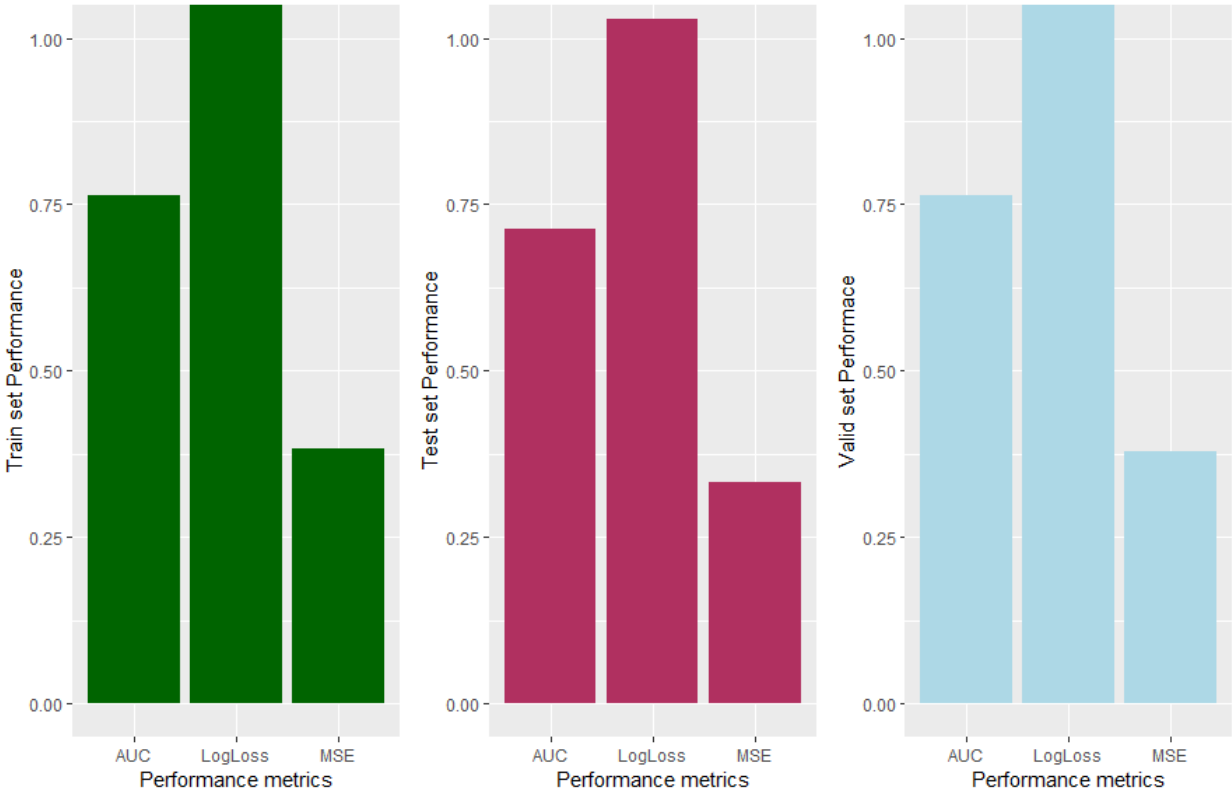


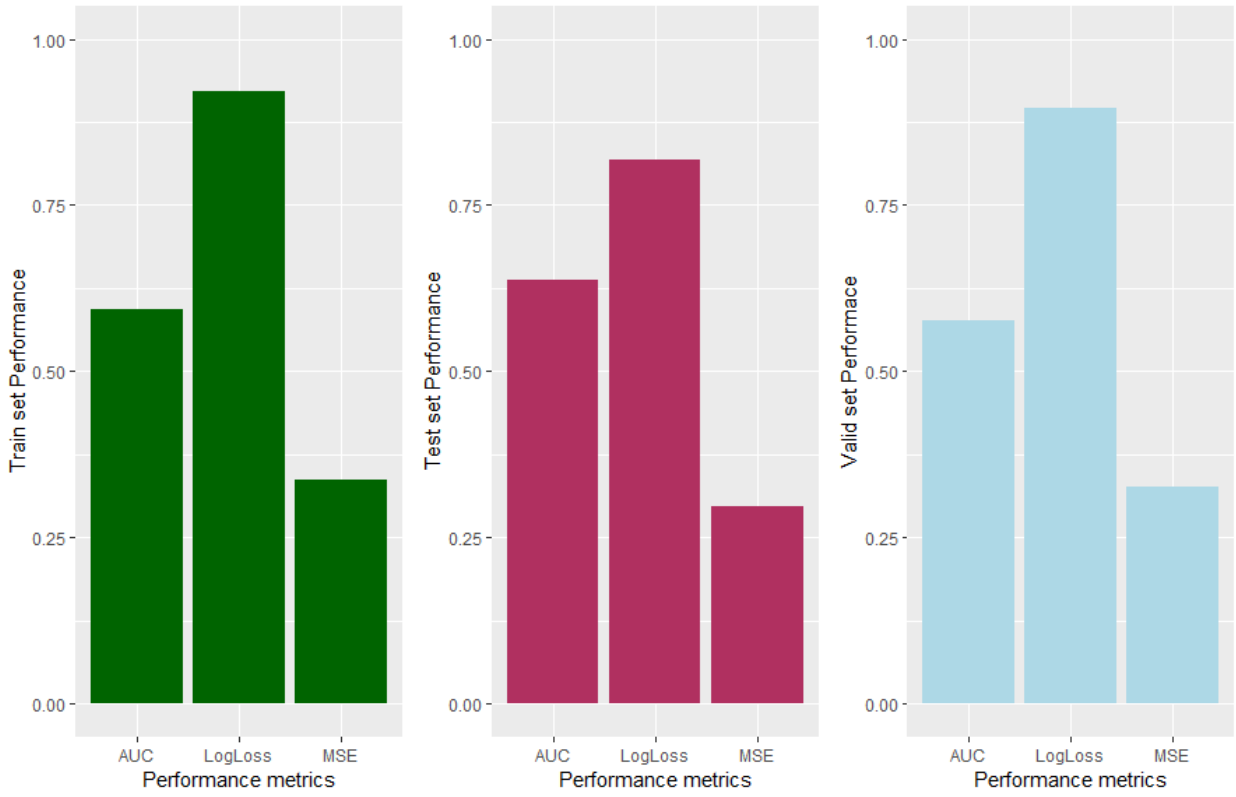
The elbow plot shows
nothing significant after 4 clusters –



PCA on Neural Network –

ICA on Neural Network –





Random Projections on Neural Network -