REPORT – CLASSIFICATION CHALLENGE OF ALZHEIMER'S DISEASE USING MRIs AND GENE EXPRESSION DATA

Name: **Anwai Archit**

Date of Submission**: 4th June 2021**

Subject: **Statistical Learning and Data Mining (Medical Imaging and Applications)**

With the spiking demand of early diagnosis of the people with Alzheimer's Disease (AD), the need of understanding the macro stages of AD before the Dementia comes into role-play is a worldwide concern. The multivariate study for deficits is our primary task in this Final Project.

Pre-Processing the Dataset:

- *Importing the Libraries, Training & Test Data*: The open-ended importing of requisite libraries to make our functions work and the *six* training & test csv files for the binary classifications.
- *Removing the Patient IDs from Training & Test Data:* Subject_id (Patient Identification) is one of non-contributing features to the binary classification.
- *Binary Encoding of the Training Labels:* Converting the Labels to zeros and ones.
- *Storing the Patient ID & Labels of Test Data for Prediction Analysis:* Subject_id will be the indicator of our predictions & labels of each training dataset to derive the formula for fitting the classifiers with the best-selected features.

Applying **Feature Engineering** on the Training Dataset(s):

- *Checking for Normality:* Studying the Shapiro-Wilk Normality Test, assuming normality for p > 0.05 & ggqqplot and ggdensity plots of the training dataset features to better understand the distribution.
- *Checking for Collinearity:* The function *vifcor* identifies variables with collinearity problem from the input variables & we remove the unwanted collinear features from the training dataset.
- *Checking for Correlation:* The function *cor* identifies correlation within the input variables in the correlation matrix & we remove the highly correlated features from the training dataset.
- *LASSO Regression for Feature Selection:* LASSO reduces the coefficients of unwanted features to zero in the process of L1 Regularization & shrinking of coefficients, removing the unneeded variables altogether. A high positive or low negative implies more importance that is variable. The respective plots show the relation of AUC & logs of lambda for included predictors & the convergent trails for using the most significant variables. We identify the best lambda value & study the importance of features.
- *PCA:* The function *ggbiplot* aids the study of principal components. Based on our experimentation, the classes are inseparable for the macro stages; hence, we choose to avoid its usage. We simply visualise the best components for our understanding.
- *Handling the Class Imbalance (**Note: For ADCN & ADMCI only**):* Using ROSE to deal with binary classification problems of imbalanced classes.
- *Cross Validation:* 10 Iterations of 5-fold Cross Validation.
- *Classification Models:* Support Vector Machines, K-Nearest Neighbours, Logistic Regression, Linear Discriminant Analysis - using the four aforementioned classifiers on the best features of

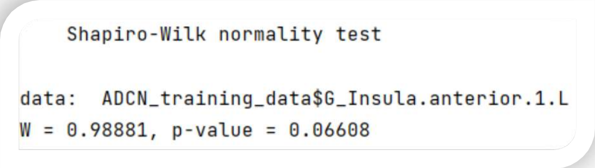
the training dataset, fitting the model and making predictions on the test datasets to find the binary macro stage classification.

- *Evaluation Metrics:* MCC and AUC for each classification.
- *Labelling the Macro Stages from Binary to Original:* Individual classes (AD/MCI/CN).
- *Saving the Best Features and Best Classification Model*

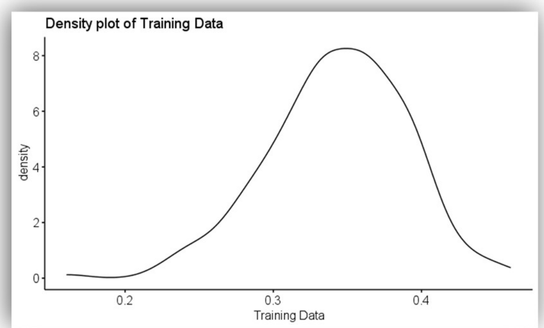
1. **<u>Alzheimer's Disease – Cognitively Normal (ADCN) Binary Patients Classification</u>**

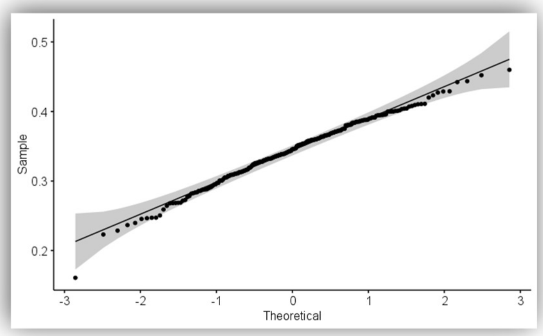
   - Requisites (**Datasets**): ADCNtrain.csv (for experimentations and model building) & ADCNtest.csv (for predictions).

- *Normality Test, Quantile-Quantile and Density Plots on ADCN Training Features:*

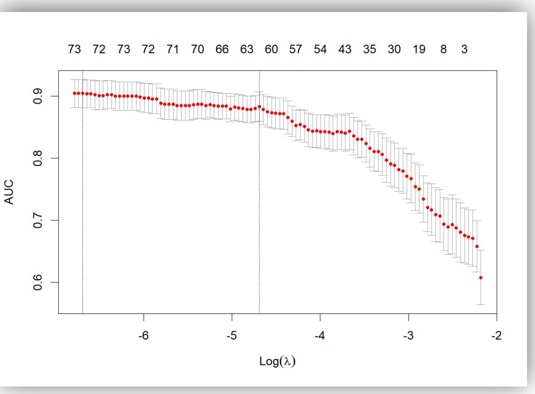


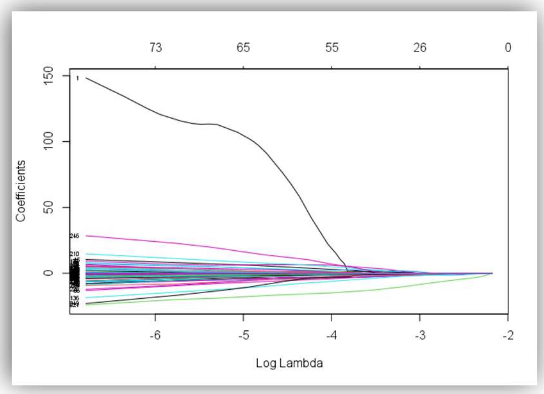
```
	Shapiro-Wilk normality test

data:  ADCN_training_data$G_Insula.anterior.1.L
W = 0.98881, p-value = 0.06608
```





- *Collinearity Test: "33 variables from the 566 input variables have collinearity problem."*
- *Correlation Test:* With a cutoff of correlation 0.8 and above, we have exempted them from the feature set.
- *LASSO Regression for Feature Selection:* Best Lambda Value -"*Min Lambda: 0.001240231*"

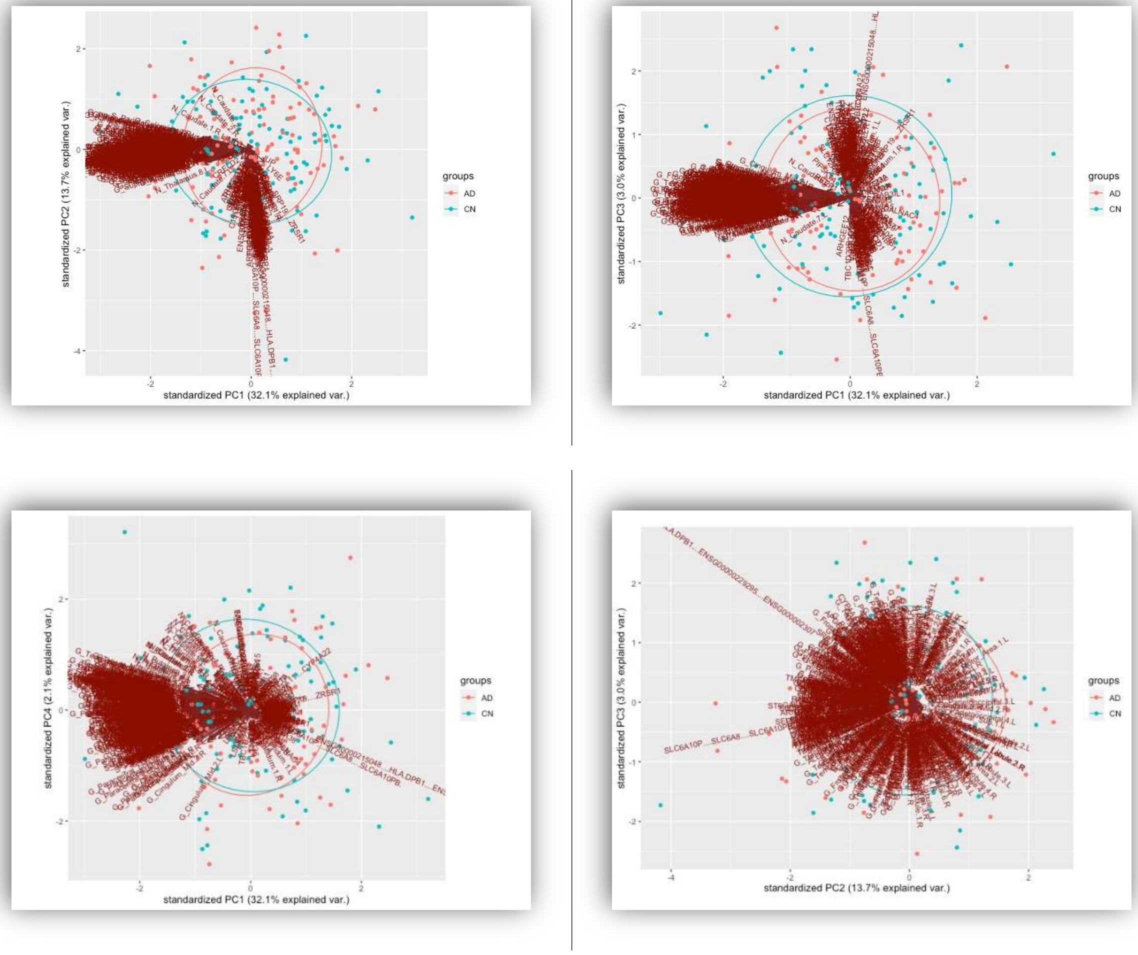




- *PCA:* "Figure-1"
- *Evaluation Metrics on Training Dataset:*

Classifiers	MCC	AUC	Accuracy
SVM	0.8449852	0.9725	0.9210453

KNN	0.6943566	0.9325	0.7730527
LR	**0.9181485**	**0.9752941**	0.9021647
LDA	0.7619488	0.9525	0.9154672

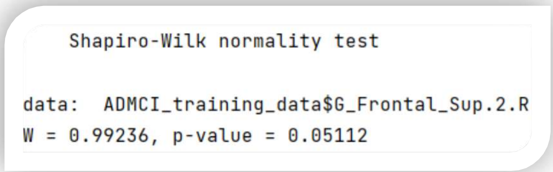


.

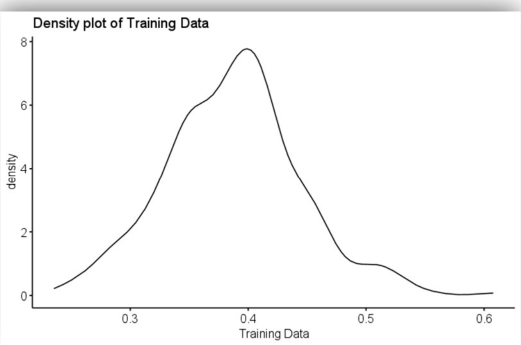
Figure - 1: PCA for ADCN

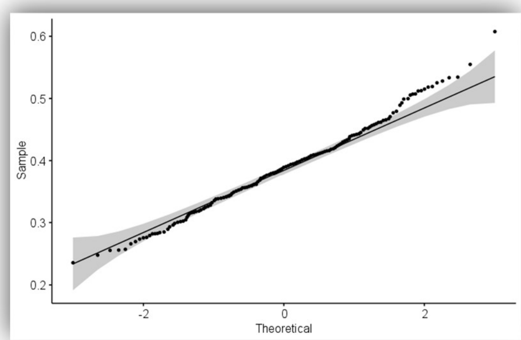
2. **Alzheimer's Disease – Mild Cognitive Impairment (ADMCI) Binary Patients Classification**
   - Requisites (**Datasets**): ADMCItrain.csv (for experimentations and model building) & ADMCItest.csv (for predictions).

❖ *Normality Test, Quantile-Quantile and Density Plots on ADCN Training Features:*

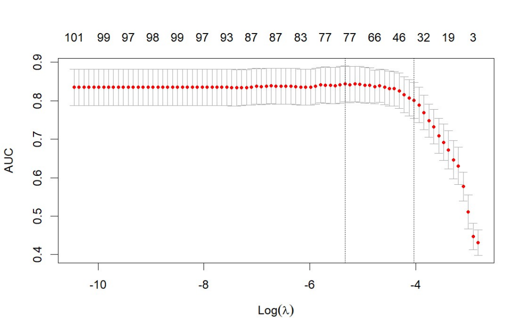


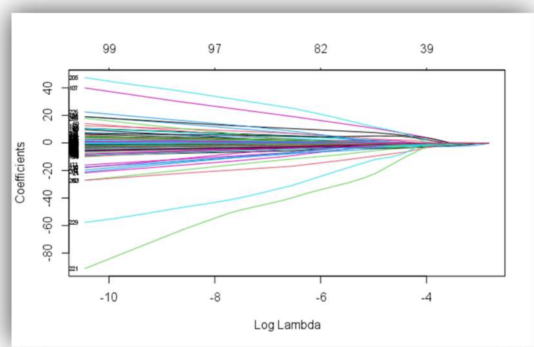
```
    Shapiro-Wilk normality test

data:  ADMCI_training_data$G_Frontal_Sup.2.R
W = 0.99236, p-value = 0.05112
```



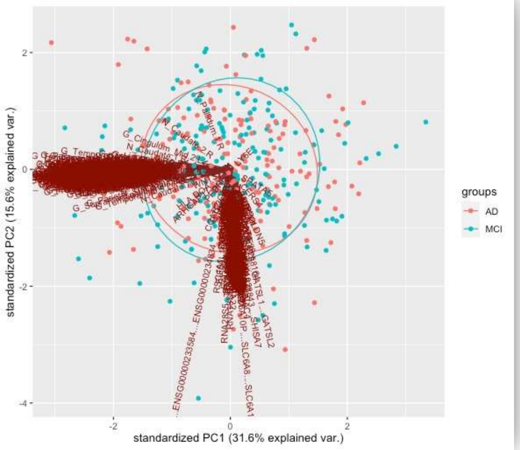


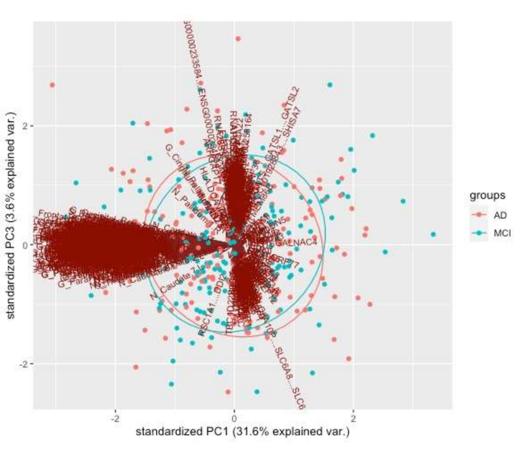
- *Collinearity Test: "40 variables from the 596 input variables have collinearity problem."*
- *Correlation Test:* With a cutoff of correlation 0.8 and above, we have exempted them from the feature set.
- *LASSO Regression for Feature Selection:* Best Lambda Value -"*Min Lambda: 0.004820303*"

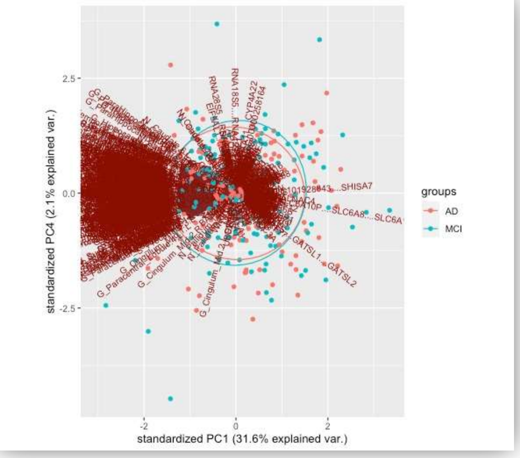


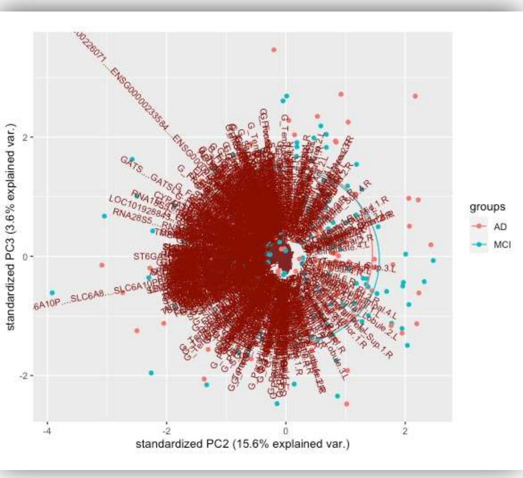


- *PCA:*









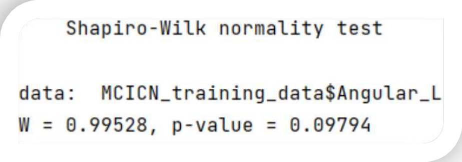
- *Evaluation Metrics on Training Dataset:*

Classifiers	MCC	AUC	Accuracy
SVM	0.6710383	0.9502924	0.9055757
KNN	0.3576531	0.8069315	0.6683730
LR	**0.7617239**	**0.9692982**	0.895261
LDA	0.7165009	0.9605263	0.9008294

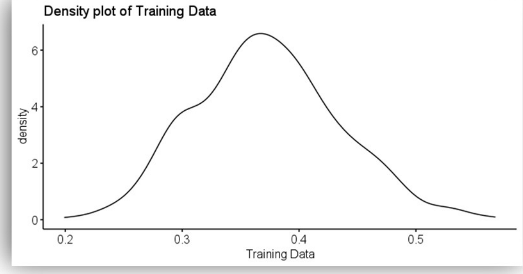
3. Mild Cognitive Impairment – Cognitively Normal (MCICN) Binary Patients Classification

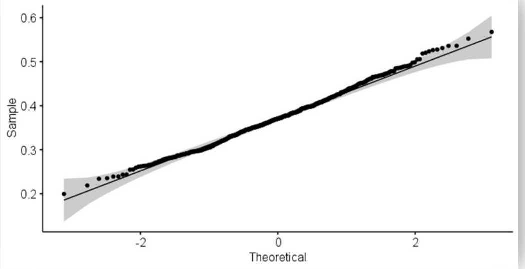
- Requisites (**Datasets**): MCICNtrain.csv (for experimentations and model building) & MCICNtest.csv (for predictions).

- *Normality Test, Quantile-Quantile and Density Plots on ADCN Training Features:*

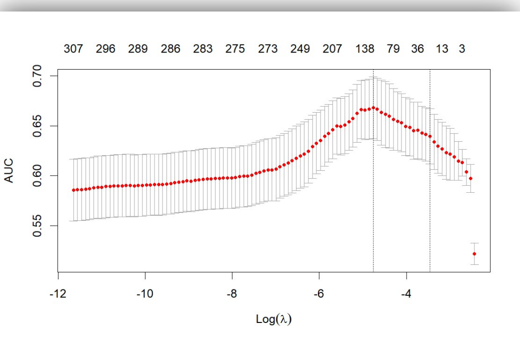


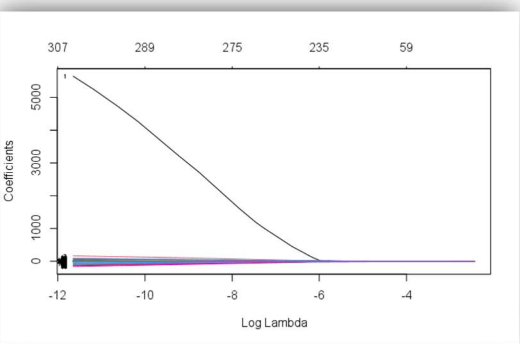
```
        Shapiro-Wilk normality test

data:  MCICN_training_data$Angular_L
W = 0.99528, p-value = 0.09794
```



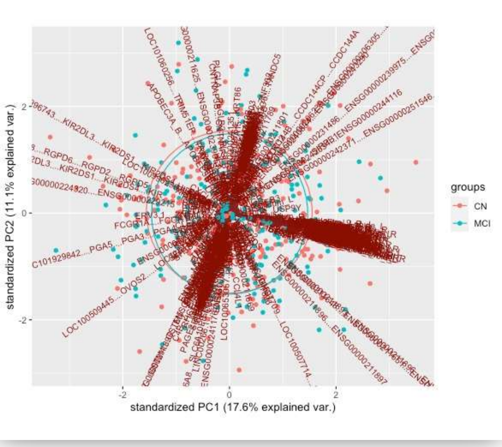


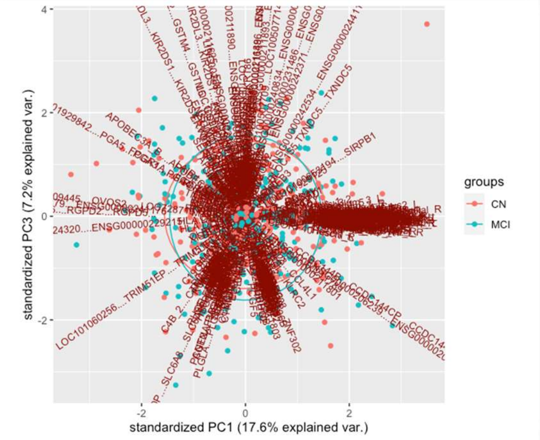
- *Collinearity Test: "58 variables from the 421 input variables have collinearity problem."*
- *Correlation Test:* With a cutoff of correlation 0.9 and above, we have exempted them from the feature set.
- *LASSO Regression for Feature Selection:* Best Lambda Value -"*Min Lambda: 0.008552971*"

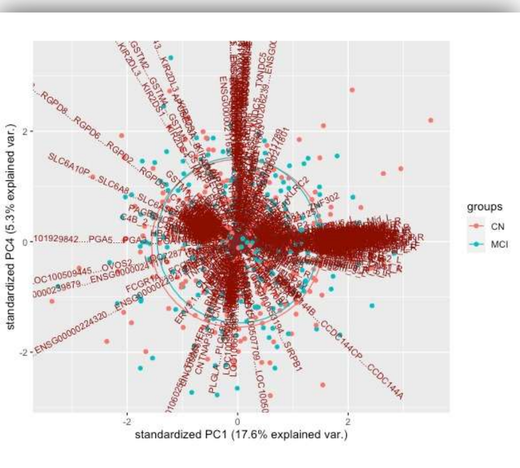


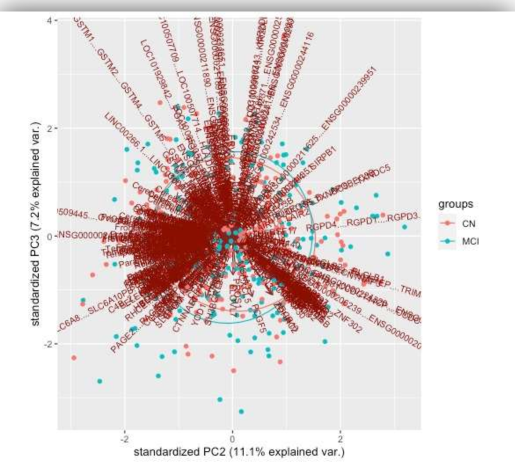


- *PCA:*









- *Evaluation Metrics on the Training Dataset:*

Classifiers	MCC	AUC	Accuracy
SVM	**0.7723091**	**0.8813743**	0.7310262
KNN	0.3320034	0.6388304	0.5986952
LR	0.7568452	0.8755263	0.7231974
LDA	0.7370809	0.8661404	0.7501886