STATISTICAL LEARNING AND LARGE DATA

Module I - Filippo Salmaso

Module II – Pietro Carlotti, Laura Pittalis, Filippo Salmaso

THE DATASET

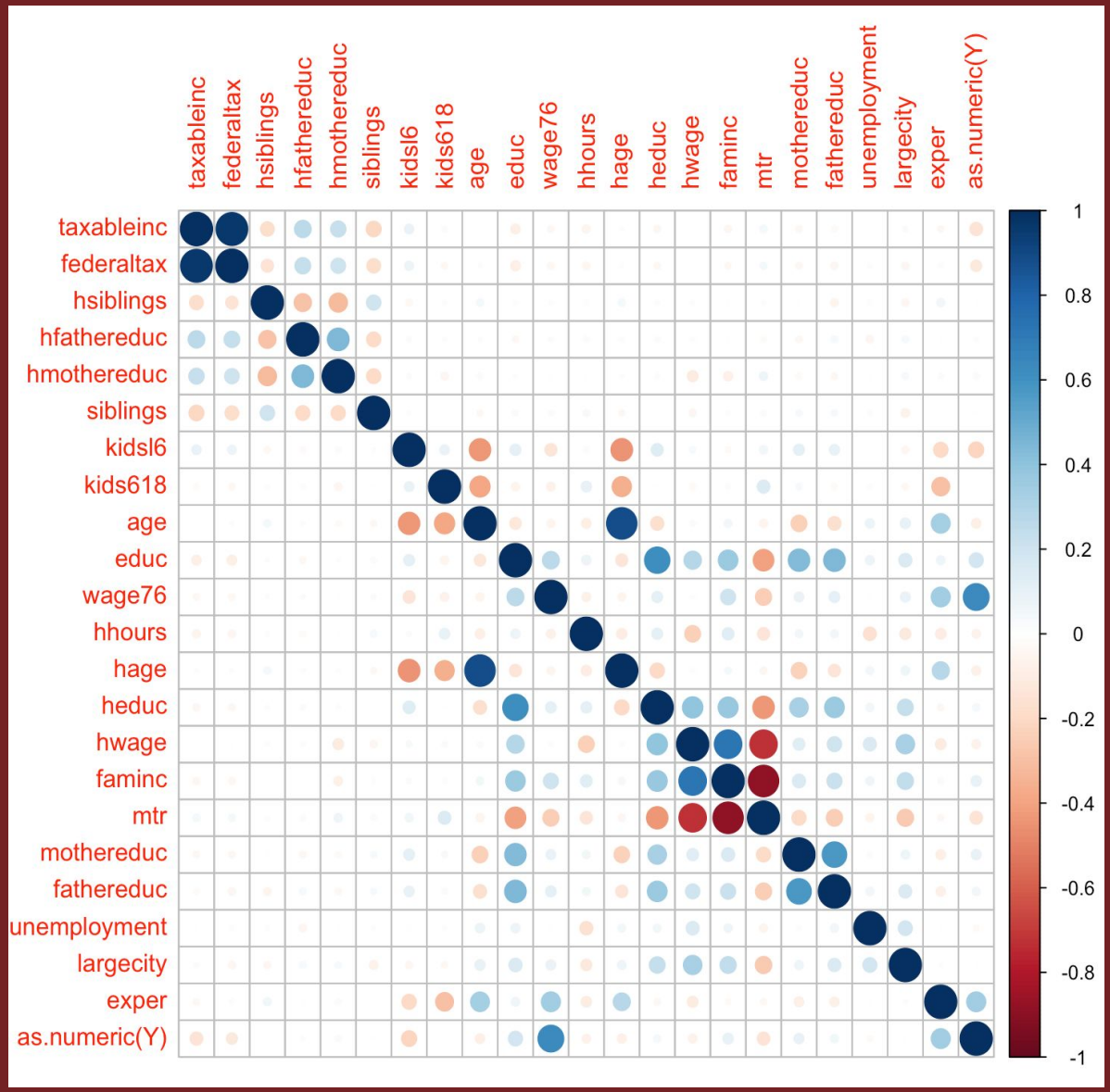
- Mroz Data on Female Labour Supply
- Source: Mroz (1987)
- It is a 1976 Panel Study of Income Dynamics
- 753 rows, 23 variables

THE RESEARCH QUESTION

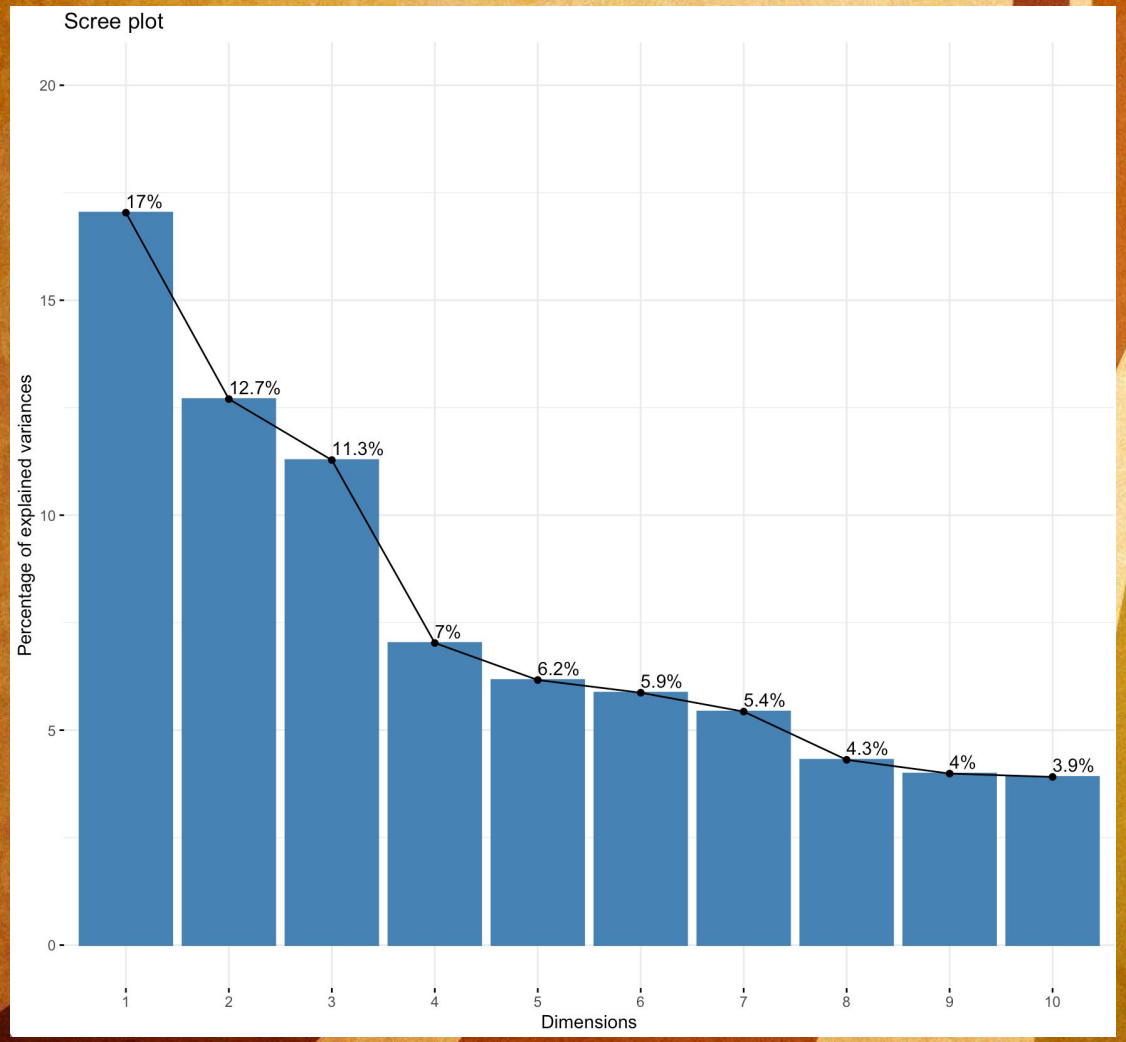
- Unsupervised Analysis: deriving the main sources of variability in the data, trying to group observations and testing the accuracy
- Supervised Analysis: assessing the determinants of women's access in the U.S. labour market in the mid-70s (Y=lfp)

DATA PREPROCESSING

- Correction of the skewness of the variables --> Square root
- Standardisation to [0, 1]
- Normalisation through mean and sd
- A Corrplot is never a bad thing!

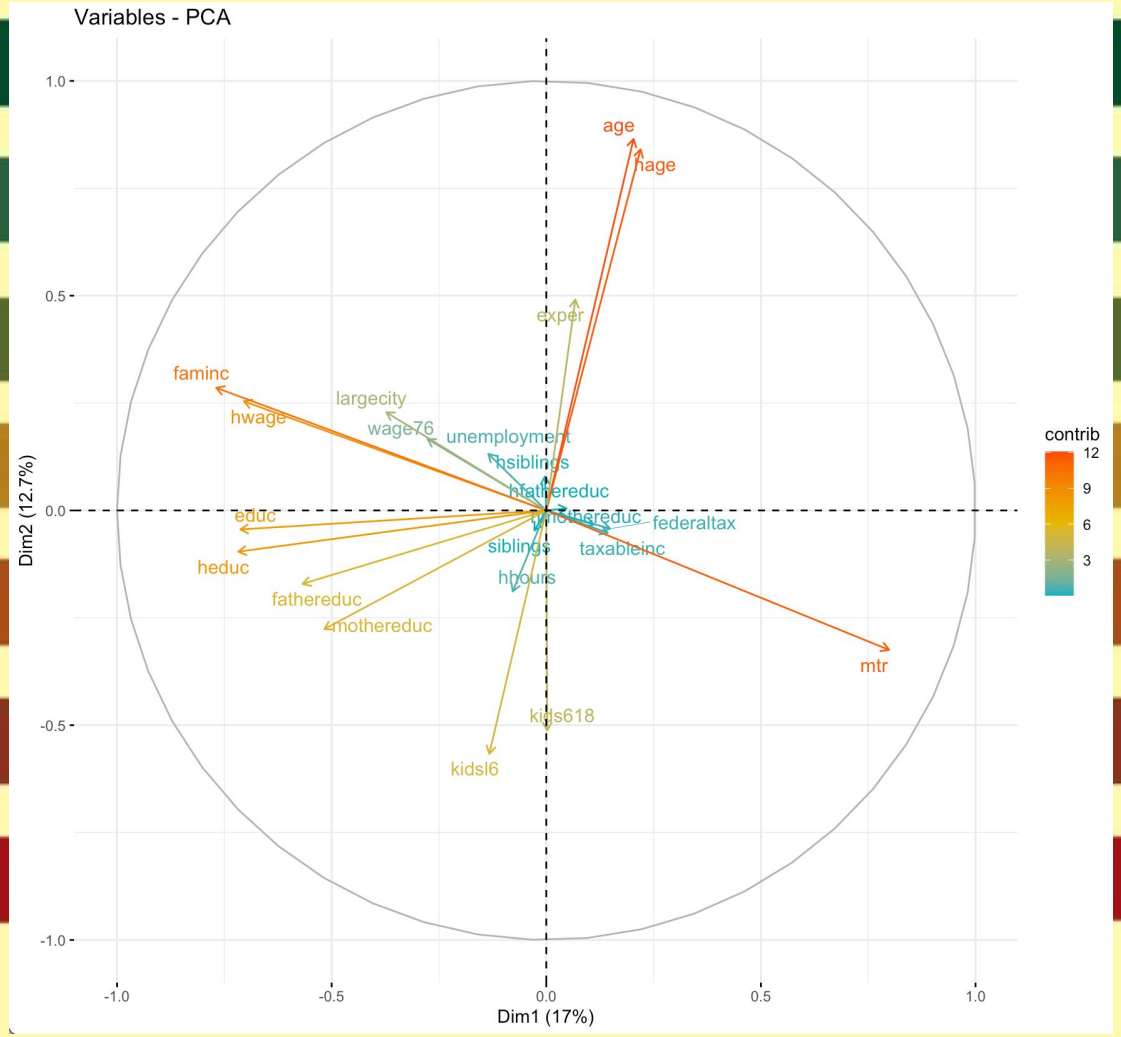


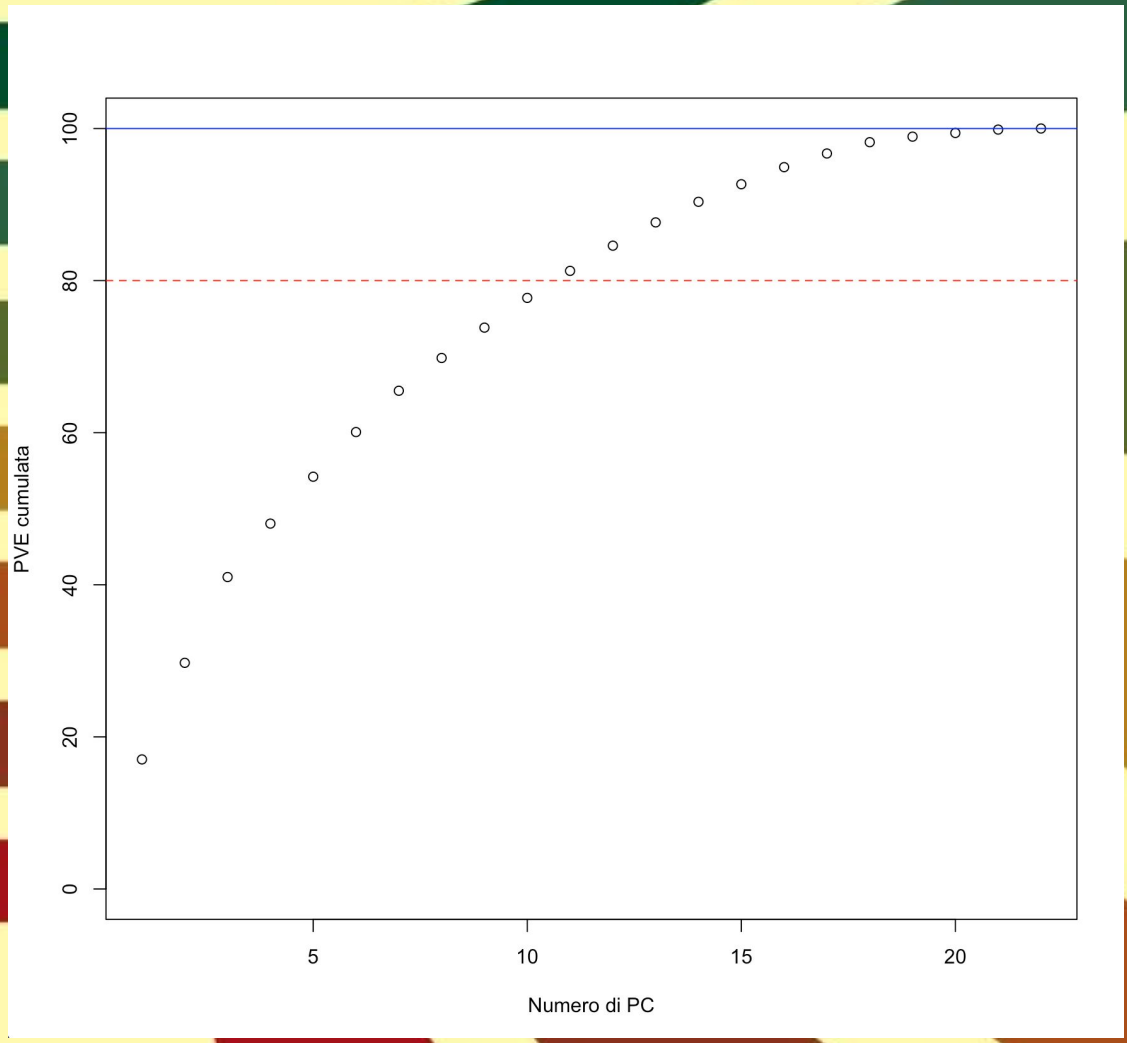
PRINCIPAL COMPONENT ANALYSIS



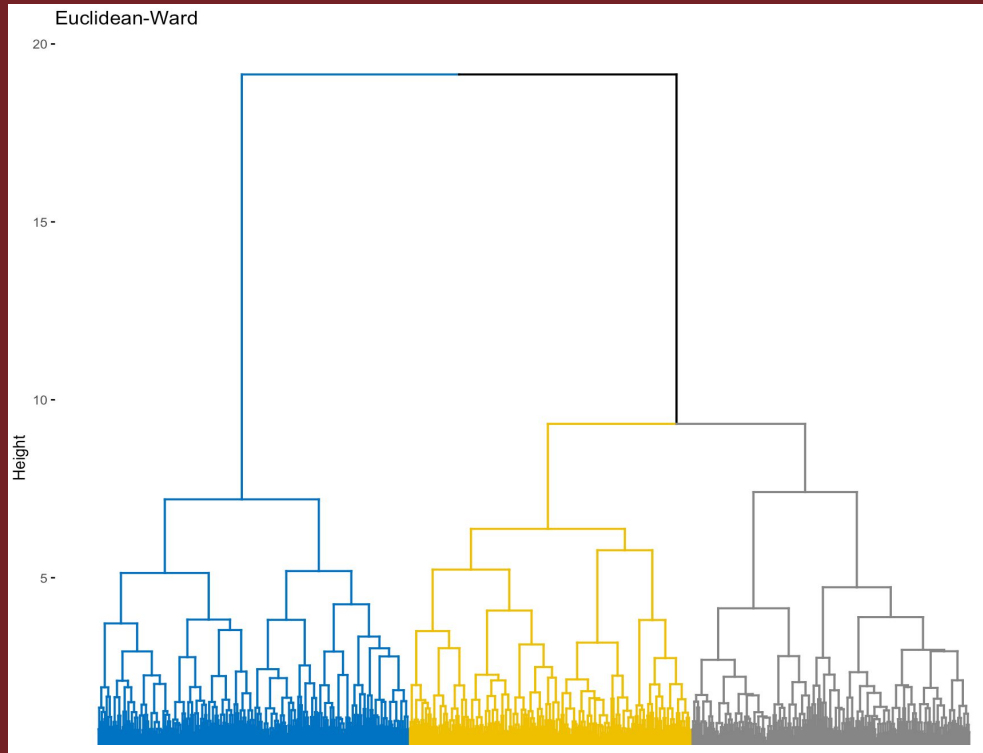
	eigenvalue	variance.percent	cumulative.variance.percent
Dim.1	3.74804735	17.0365789	17.03658
Dim.2	2.79423939	12.7010881	29.73767
Dim.3	2.48197112	11.2816869	41.01935
Dim.4	1.54603477	7.0274308	48.04678
Dim.5	1.35651205	6.1659639	54.21275
Dim.6	1.29165945	5.8711793	60.08393
Dim.7	1.19497374	5.4316988	65.51563
Dim.8	0.94821722	4.3100783	69.82570
Dim.9	0.87775412	3.9897915	73.81550
Dim.10	0.86056709	3.9116686	77.72716
Dim.11	0.78155209	3.5525095	81.27967
Dim.12	0.72890453	3.3132024	84.59288
Dim.13	0.67333647	3.0606203	87.65350
Dim.14	0.59616976	2.7098625	90.36336
Dim.15	0.50837165	2.3107802	92.67414
Dim.16	0.49398866	2.2454030	94.91954
Dim.17	0.39524152	1.7965524	96.71610
Dim.18	0.32647181	1.4839628	98.20006
Dim.19	0.16100986	0.7318630	98.93192
Dim.20	0.10618837	0.4826744	99.41460
Dim.21	0.09805915	0.4457234	99.86032
Dim.22	0.03072985	0.1396811	100.00000

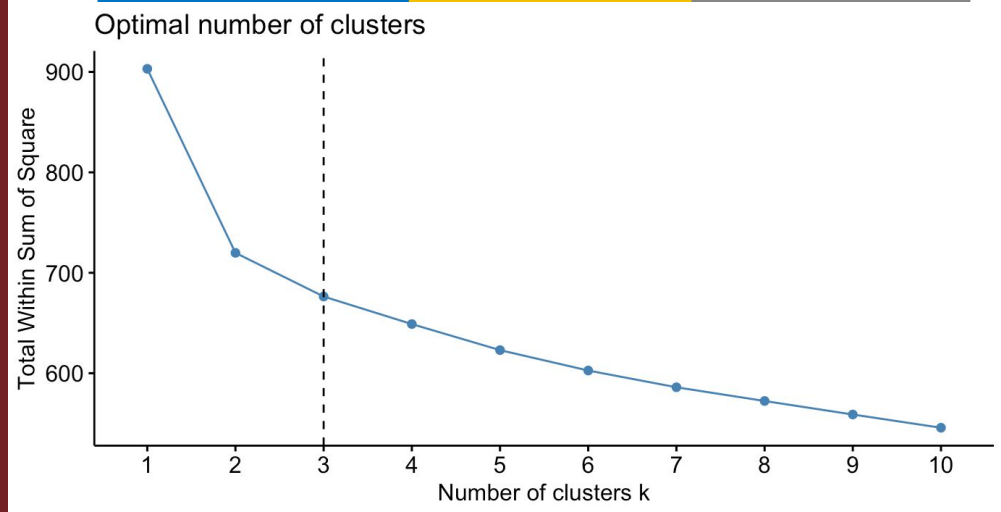
BIPLOT AND CUMULATIVE PVE

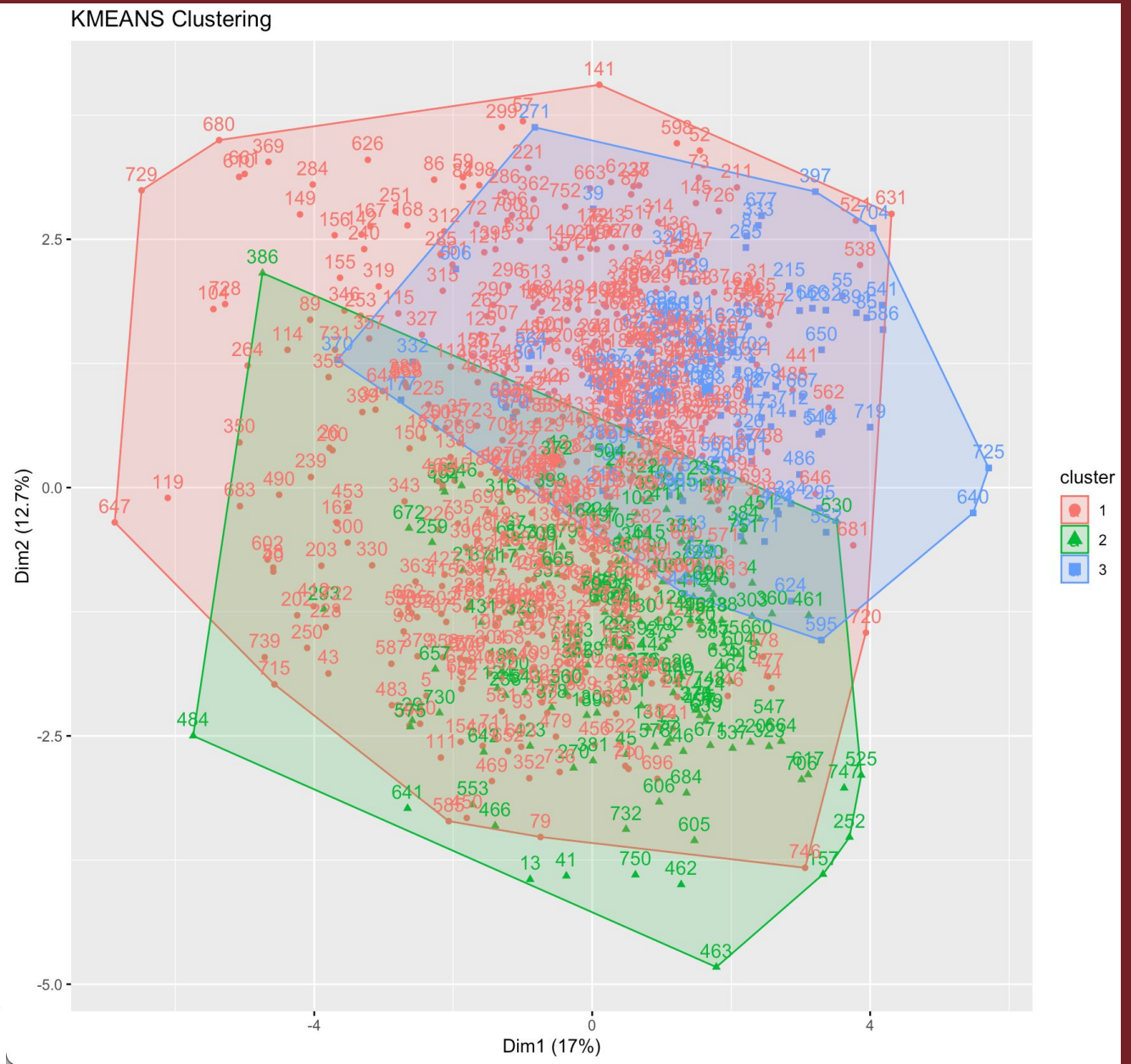




CLUSTERING: AVERAGE SILHOUETTE WIDTH 0.12 (WARD HIERARCHICAL CLUSTERING)







CLASSIFICATION (CONFUSION MATRICES)

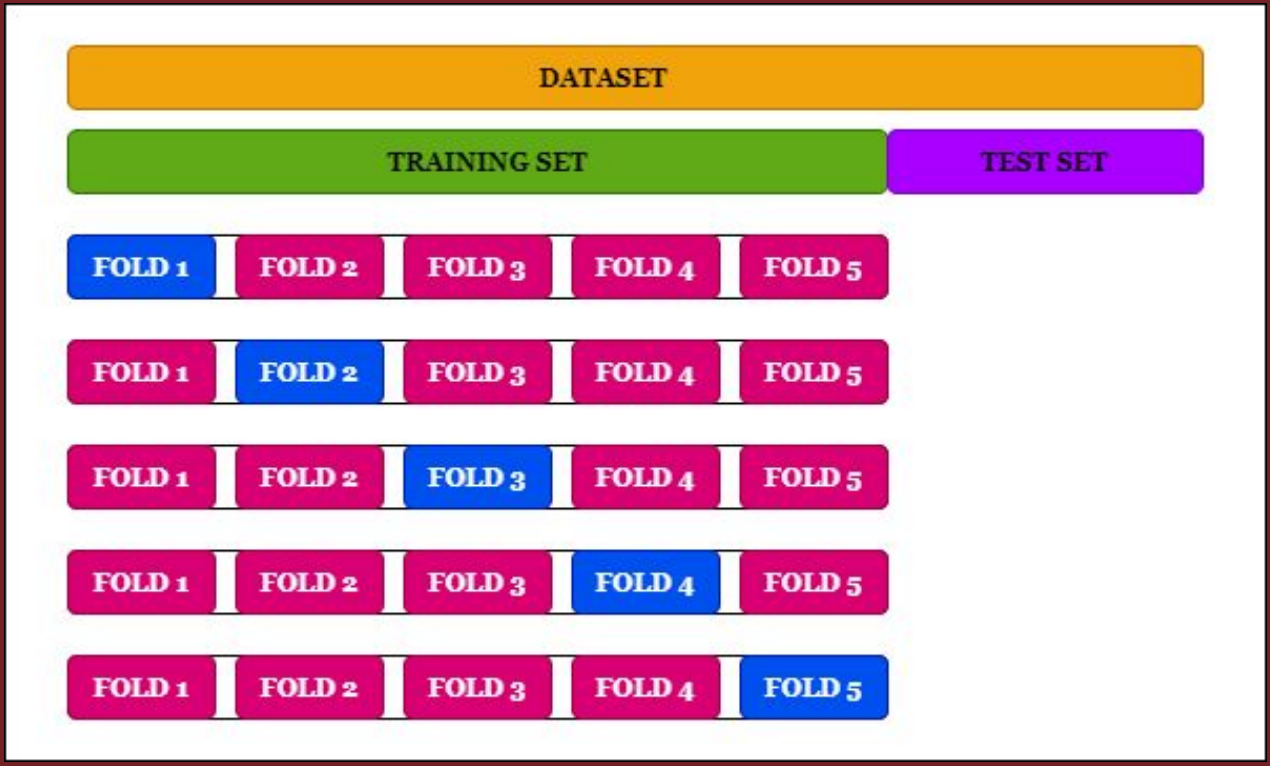
Complete Logit: AIC 372.32
Stepwise Logit: AIC down to 357.79

```
> Logit_confusion_matrix$table
          Reference
Prediction  0  1
         0 70 18
         1 11 89
> Logit_stepwise_confusion_matrix$table
          Reference
Prediction  0  1
         0 70 19
         1 11 88
> SVM_confusion_matrix$table
          Reference
Prediction  0  1
         0 71 21
         1 10 86
```

```
> LDA_confusion_matrix$table
          Reference
Prediction  0  1
         0 72 20
         1  9 87
> QDA_confusion_matrix$table
          Reference
Prediction  0  1
         0 69 14
         1 12 93
> KNN_confusion_matrix$table
          Reference
Prediction  0  1
         0 69 22
         1 12 85
```

```
> PC_logit_confusion_matrix
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  76   5
         1   5 102

> Naive_Bayes_confusion_matrix$table
          Reference
Prediction  0  1
         0 72 21
         1  9 86
```

CROSS VALIDATION



Training set: 703 observations (approximately 93%)

Test set: 50 observations

Number of folds: 10

Iterations: 20

EVALUATION METRIC

ACCURACY:

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

PRECISION AND RECALL

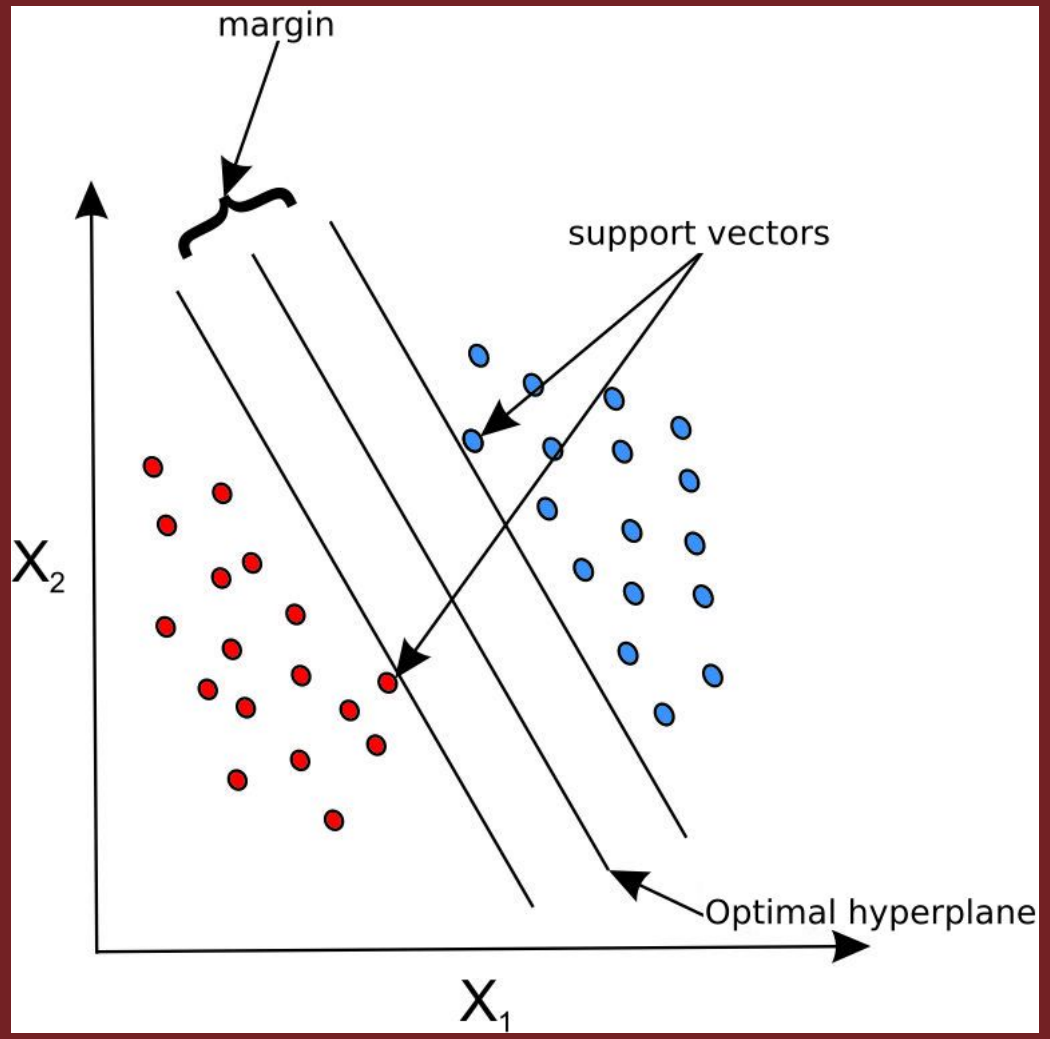
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

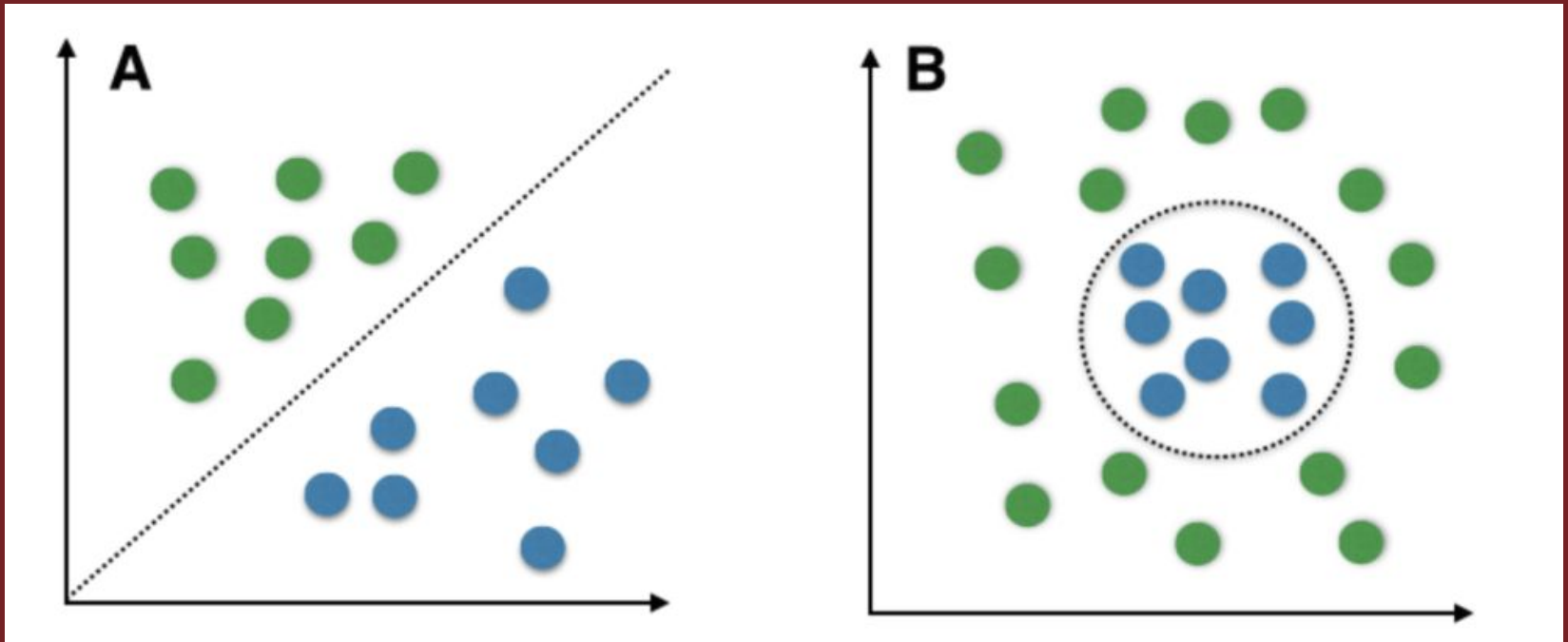
CLASSIFIERS

- Logit
- Stepwise feature selection logit
- LDA
- QDA
- KNN
- PC logit
- SVM
- Naive Bayes
- Decision tree
- Random forest

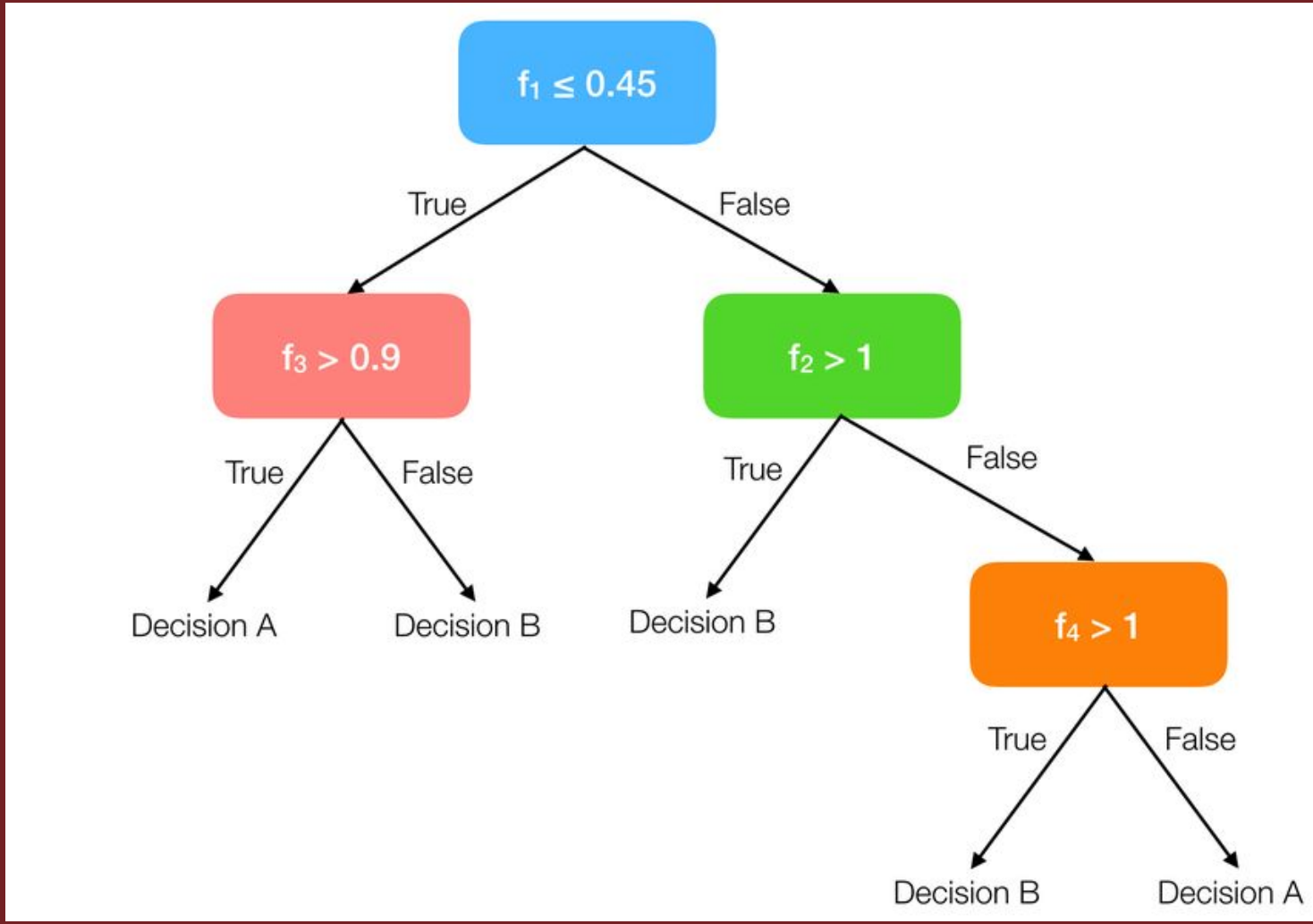
SVM CLASSIFIER



SVM CLASSIFIER



DECISION TREE CLASSIFIER



DECISION TREE CLASSIFIER

Impurity Criterion

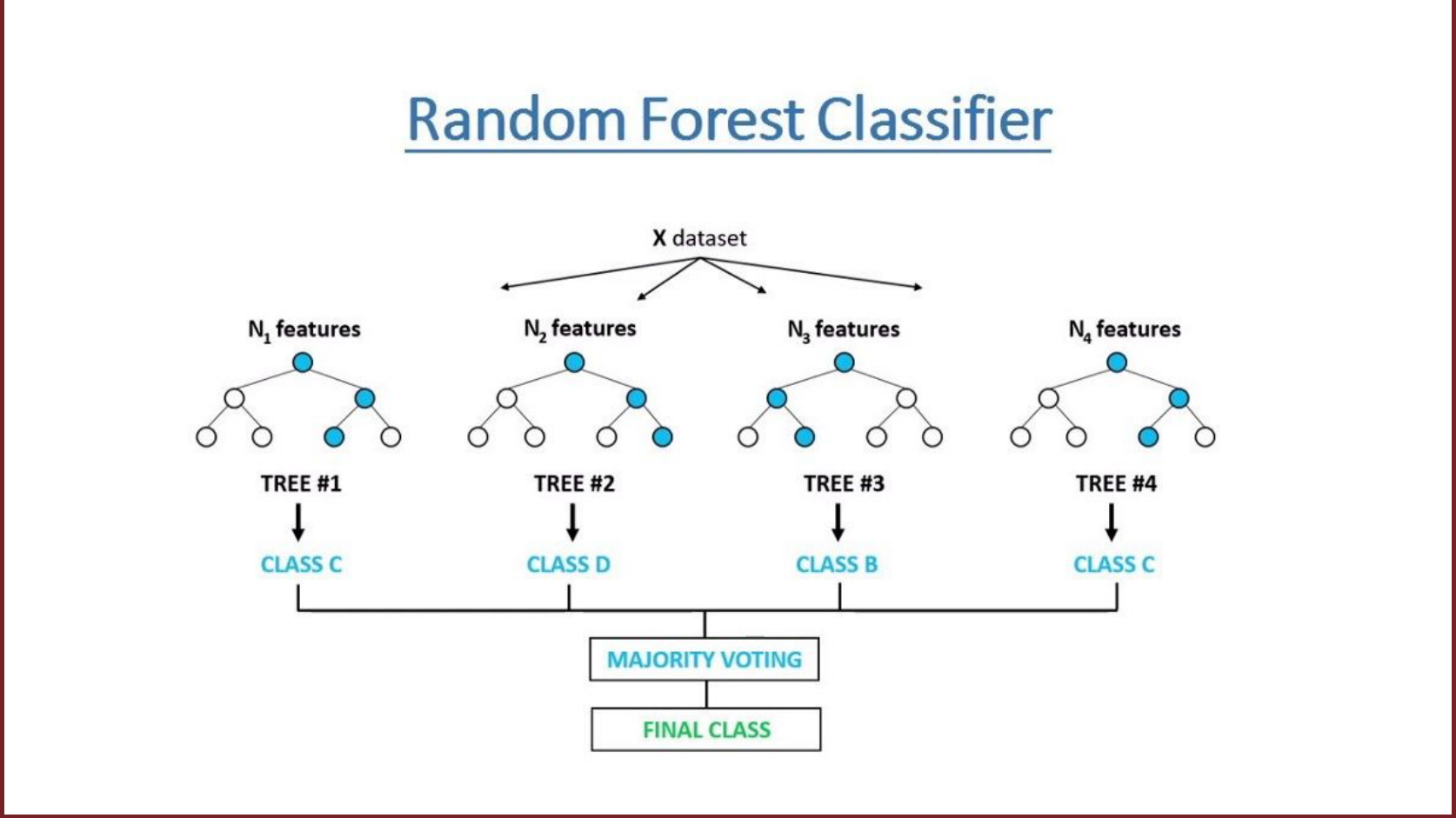
Gini Index

$$I_G = 1 - \sum_{j=1}^{c} p_j^2$$

Entropy

$$I_H = -\sum_{j=1}^{c} p_j log_2(p_j)$$

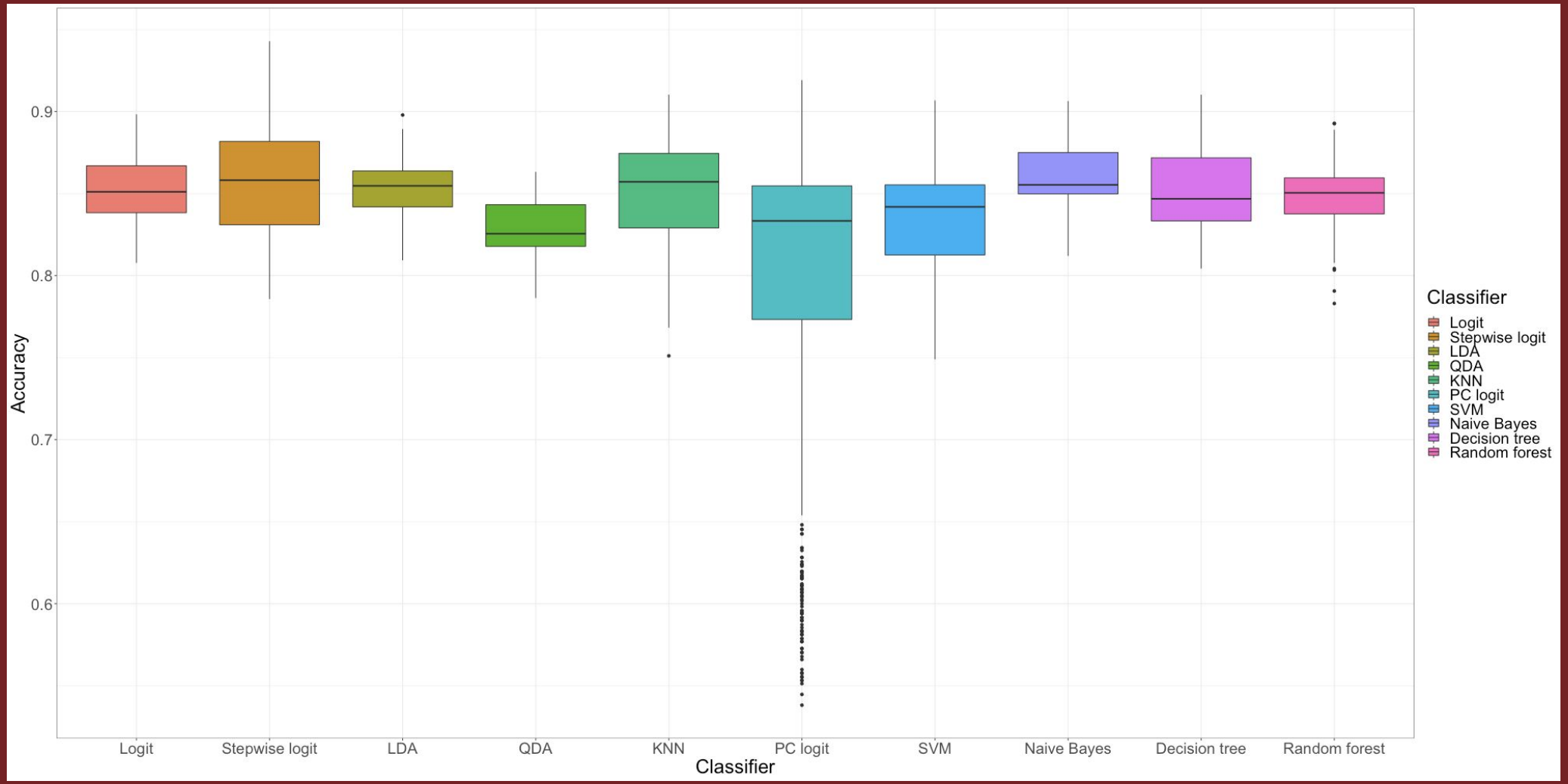
RANDOM FOREST CLASSIFIER



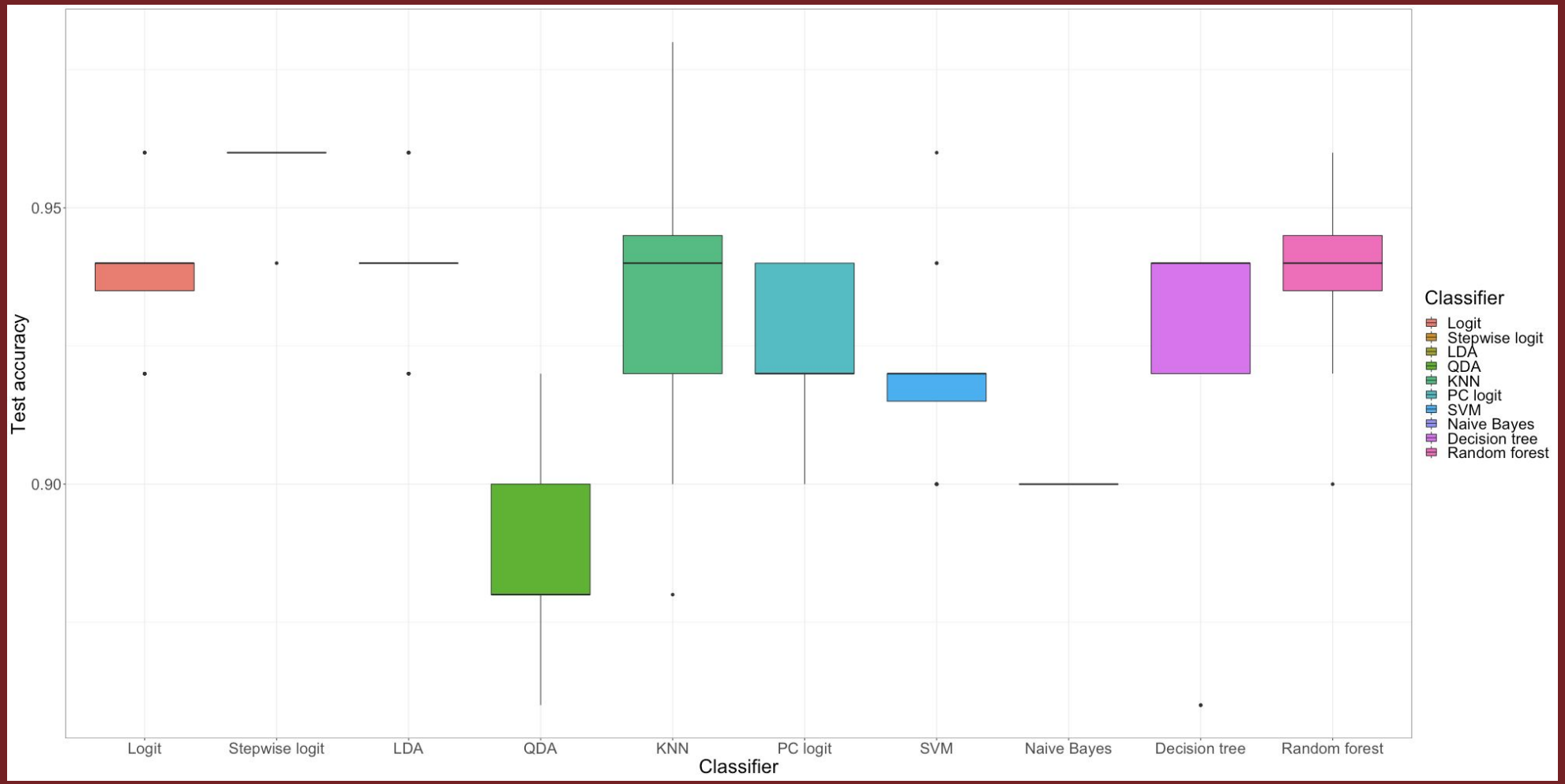
TUNING PARAMETERS

- PC logit: number of principal components
- SVM: kernel function
- KNN: number of neighbors
- Decision tree: maximal depth
- Random forest: number of random variables at each node

CROSS VALIDATION ACCURACIES



TEST SET ACCURACIES



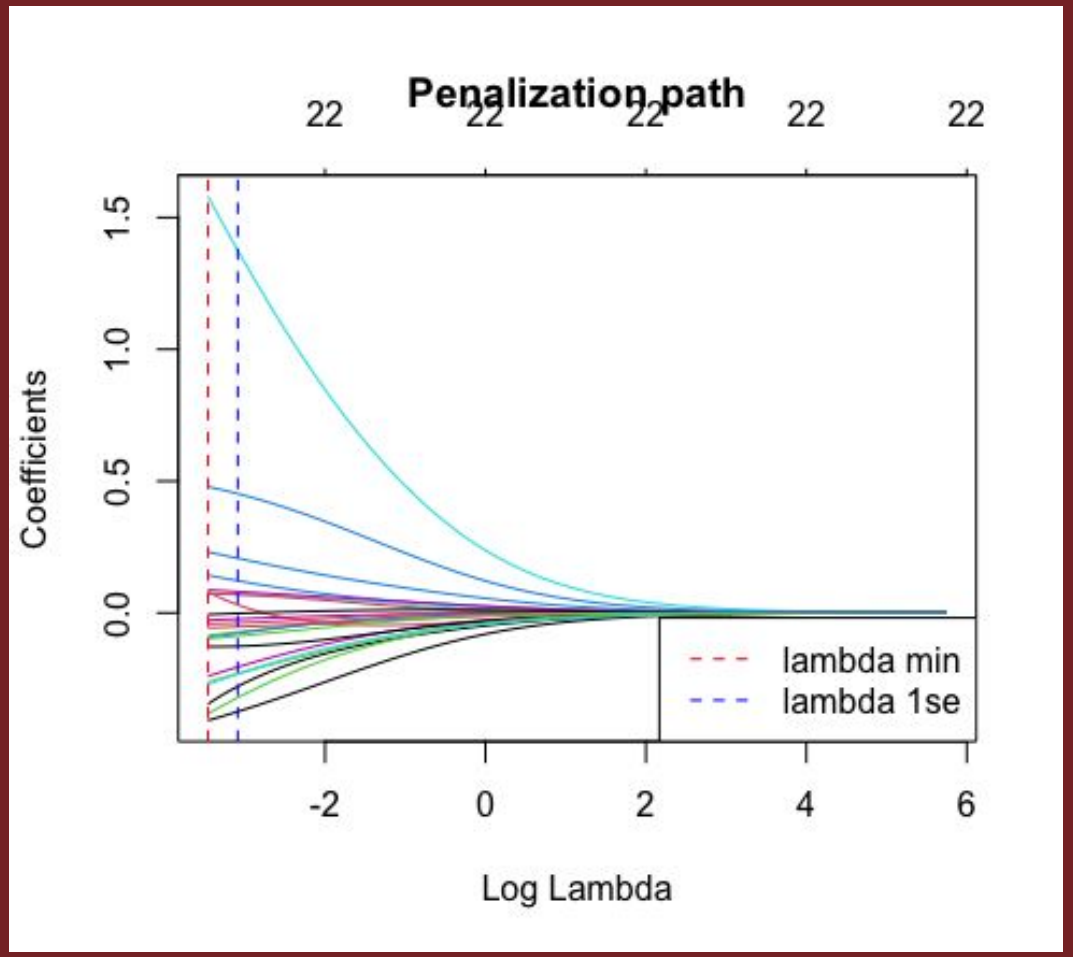
Is there any multicollinearity issue among our predictors?

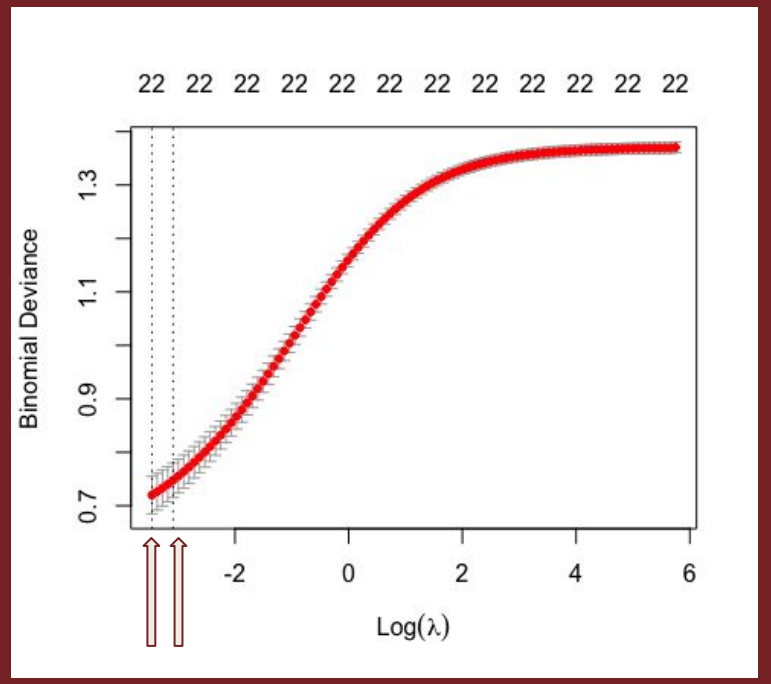
$$VIF_j = \frac{1}{1 - R^2_{X_j|otherX_s}}$$

Table 1: VIF

taxableinc	18.484
federaltax	17.806
hsiblings	1.278
hfathereduc	1.266
hmothereduc	1.341
siblings	1.132
kidsl6	1.471
kids618	1.522
age	5.486
educ	2.025
wage76	1.125
hhours	1.743
hage	5.082
heduc	1.995
hwage	3.875
faminc	5.507
mtr	6.687
mothereduc	1.664
fathereduc	1.585
unemployment	1.107
largecity	1.198
exper	1.300

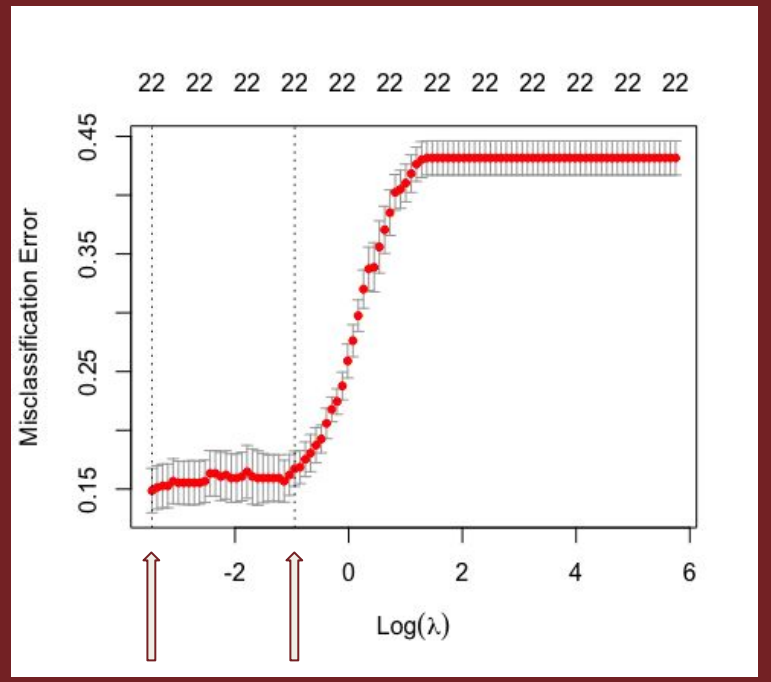
Ridge : lambda fit





MINIMUM lambda:
0.031

WITHIN 1 SE lambda:
0.045



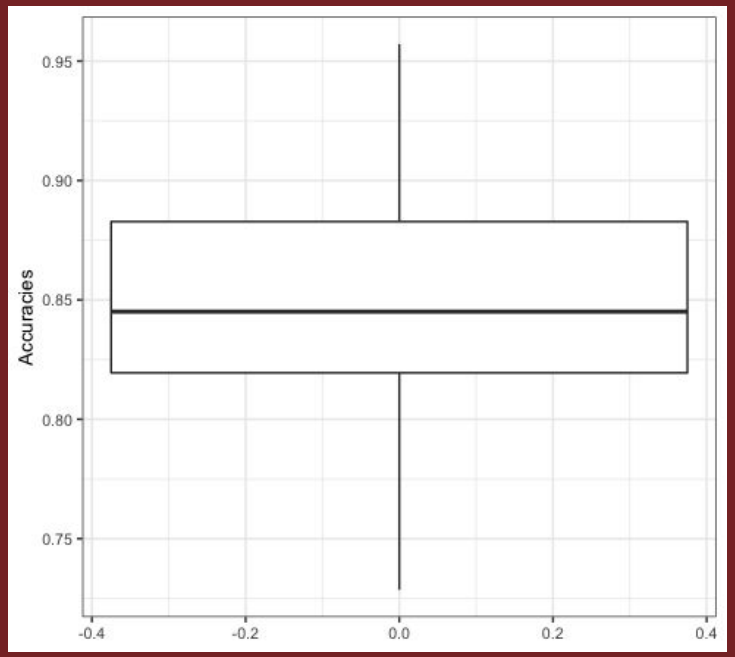
MINIMUM lambda:
0.031

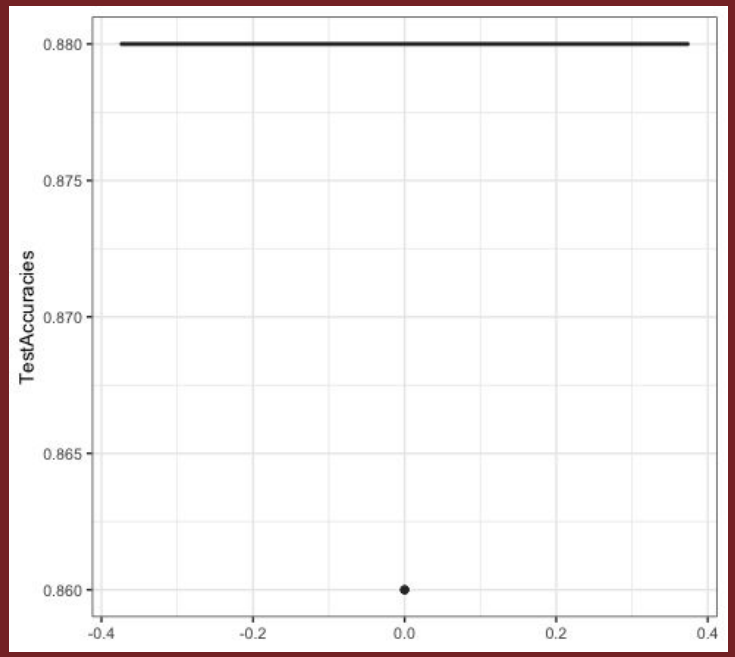
WITHIN 1 SE lambda:
0.38

Ridge: performance

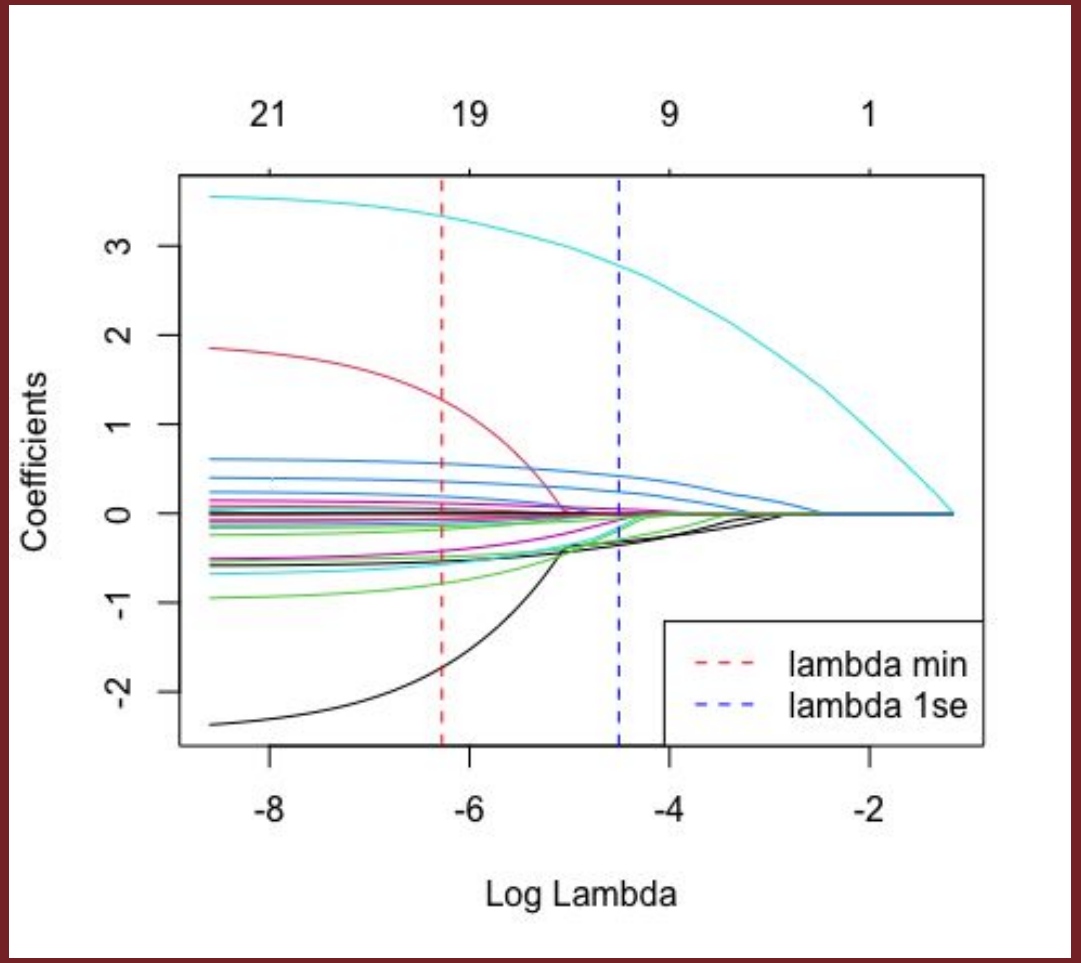
(Intercept)	0.72862180
taxableinc	-0.32890420
federaltax	0.09439418
hsiblings	-0.03954700
hfathereduc	-0.11409241
hmothereduc	0.10674335
siblings	-0.02000602
kidsl6	-0.41434821
kids618	0.10376023
age	-0.30504744
educ	0.19011423
wage76	1.57958476
hhours	-0.30812579
hage	-0.11067974
heduc	0.00207064
hwage	-0.37027908
faminc	0.09517143
mtr	-0.12652498
mothereduc	0.10630340
fathereduc	0.01351770
unemployment	-0.06521322
largecity	-0.08110368
exper	0.45722393

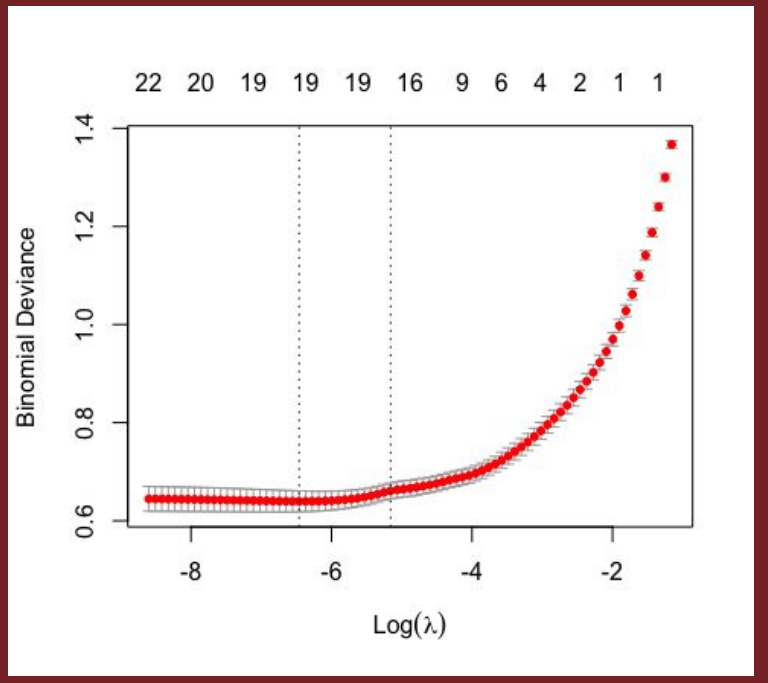
```
          Reference
Prediction  0  1
         0 71 22
         1 10 85

               Accuracy : 0.8298
                 95% CI : (0.7683, 0.8806)
    No Information Rate : 0.5691
    P-Value [Acc > NIR] : 2.426e-14

                  Kappa : 0.6591
```





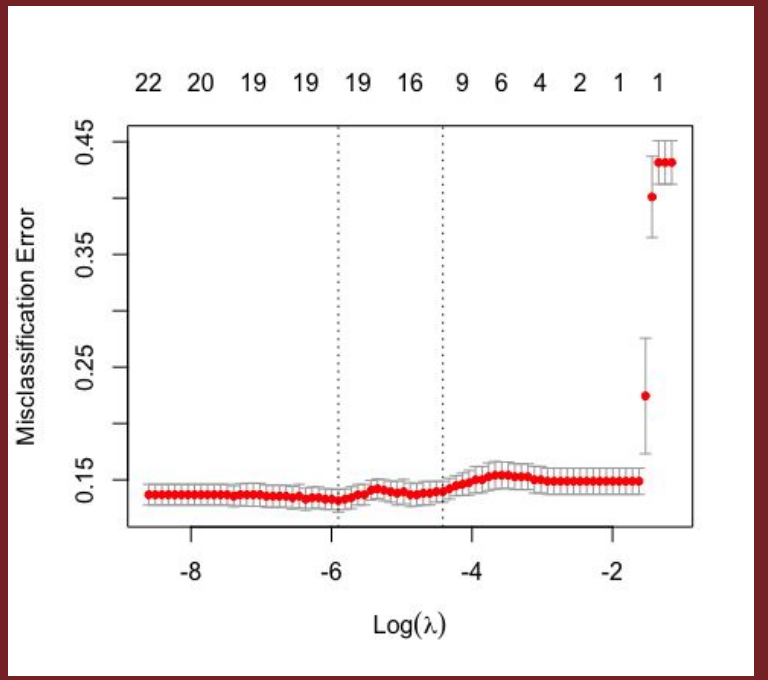
LASSO: lambda fit





MINIMUM lambda: 0.0018

WITHIN 1 SE lambda: 0.011



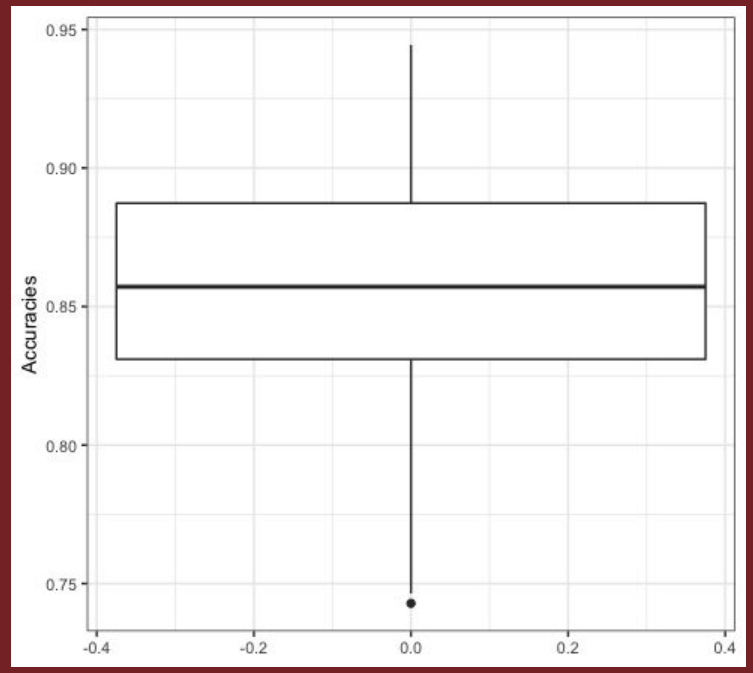
MINIMUM lambda: 0.0027

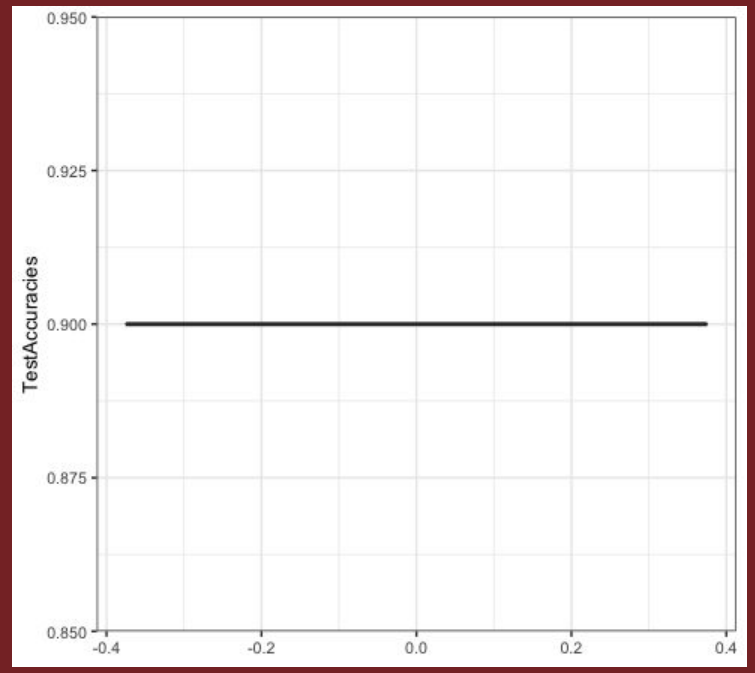
WITHIN 1 SE lambda: 0.012

LASSO: performance

(Intercept)	1.79713152
taxableinc	-1.41707851
federaltax	1.00243934
hsiblings	-0.12781629
hfathereduc	-0.19922906
hmothereduc	0.10118867
siblings	-0.04872464
kidsl6	-0.57419284
kids618	0.04908124
age	-0.55986586
educ	0.32703043
wage76	3.47930598
hhours	-0.49125279
hage	.
heduc	0.05868827
hwage	-0.69387248
faminc	0.22298429
mtr	-0.21052734
mothereduc	0.18764200
fathereduc	.
unemployment	-0.08452289
largecity	-0.15189722
exper	0.52782743

```
          Reference
Prediction  0  1
         0 71 22
         1 10 85

               Accuracy : 0.8298
                 95% CI : (0.7683, 0.8806)
    No Information Rate : 0.5691
    P-Value [Acc > NIR] : 2.426e-14

                  Kappa : 0.6591
```

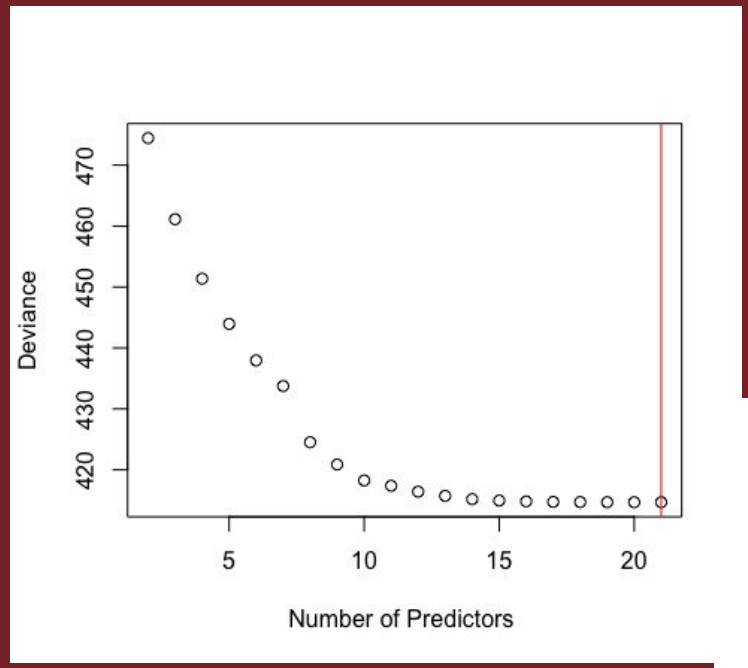


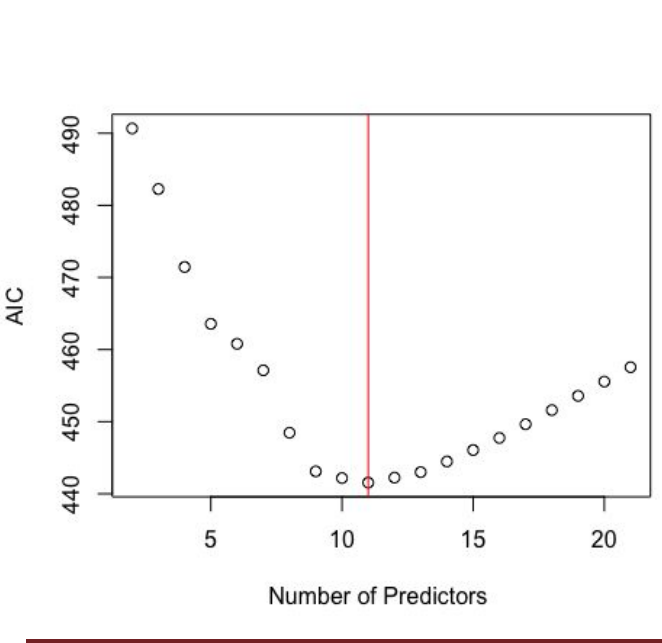


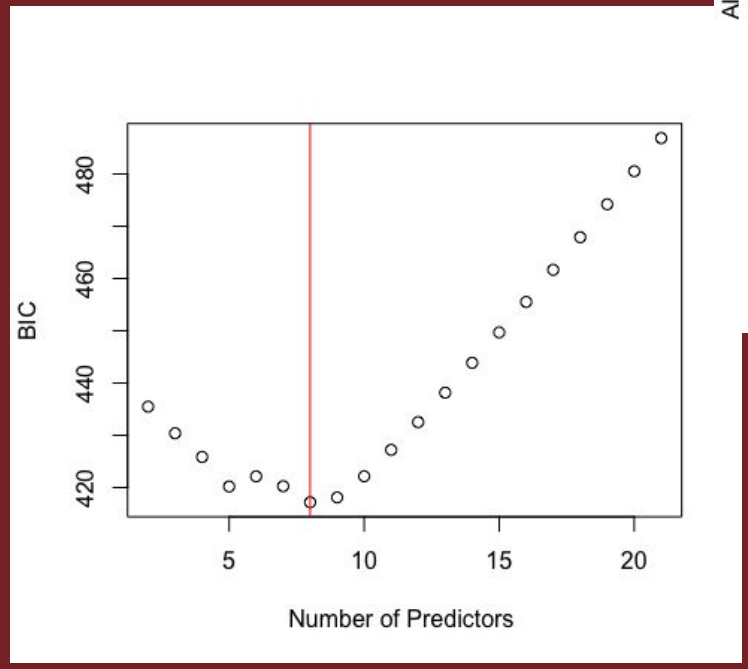
Best Subset Selection with Logit

$$AIC = -2\log L(\hat{\vartheta}_{ML}) + 2d$$

$$BIC = -2\log L(\hat{\vartheta}_{ML}) + \log(n)d$$







FUTURE PLANS

- Testing new classifiers (neural links, PC lasso, adding interactions…)
- Supervised dimension reduction
- More cross validation on tuning parameters

REFERENCES

- Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. Econometrica, 55(4), 765–799. https://doi.org/10.2307/1911029
- Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). An introduction to statistical learning: with applications in R. Spinger.
- Durgesh, K. S., & Lekha, B. (2010). Data classification using support vector machine. *Journal of theoretical and applied information technology*, *12*(1), 1-7.
- Swain, P. H., & Hauska, H. (1977). The decision tree classifier: Design and potential. *IEEE Transactions on Geoscience Electronics*, *15*(3), 142-147.
- Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, *2*(3), 18-22.

THANK YOU FOR YOUR ATTENTION!