TÉCNICO LISBOA

Predictability of Human Development Index

Multivariate Analysis

Professor Maria do Rosário De Oliveira Silva

Pedro Sousa, 42022
André Luís, 98638
Carlos Sequeira, 87638
Sara Cruz, 79410
Pedro Dias, 39953

TÉCNICO LISBOA

Outline

- Introduction.
- Data set preparation.
- Statistical Analysis.
- Principal Component Analysis.
- Clustering.
- Classification.
- Conclusion.

TÉCNICO LISBOA

Introduction

- **Objective:** explore the World Bank data to find additional explanations predictive of the progress of HDI in the 21st century.

HDI Rank (2018)	Country	1990	1991	...	2018
170	Afghanistan	0.298	0.304	...	0.496
69	Albania	0.644	0.625	...	0.791
82	Algeria	0.578	0.582	...	0.759
36	Andorra	na	na	...	0.857
149	Angola	na	na	...	0.574

Table 1: Table: sample from HDI.csv file. A 212x31 matrix.

Country Name	Country Code	Indicator Name	1960	...	2018
Arab World	ARB	Access to cl...	na	...	na
Arab World	ARB	Adjusted sav...	na	...	5.084
Arab World	ARB	Adolescent f...	134.8	...	46.01
Arab World	ARB	Age dependen...	88.06	...	61.17
Arab World	ARB	Arable land ...	na	...	na

Table 2: Sample from WDIData.csv file. A 9504x66 matrix.

Data set preparation

HDI = f(WDI)

year

country

WDI

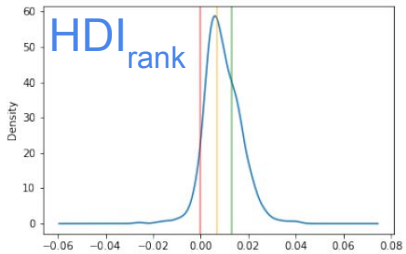
year

country

WDI_{Δ}, WDI, HDI_{Δ}, HDI_{rank}

country.year

2002
2003 } WDI_{Δ}
2004 WDI_{2004}
2005 } HDI_{Δ}
2006

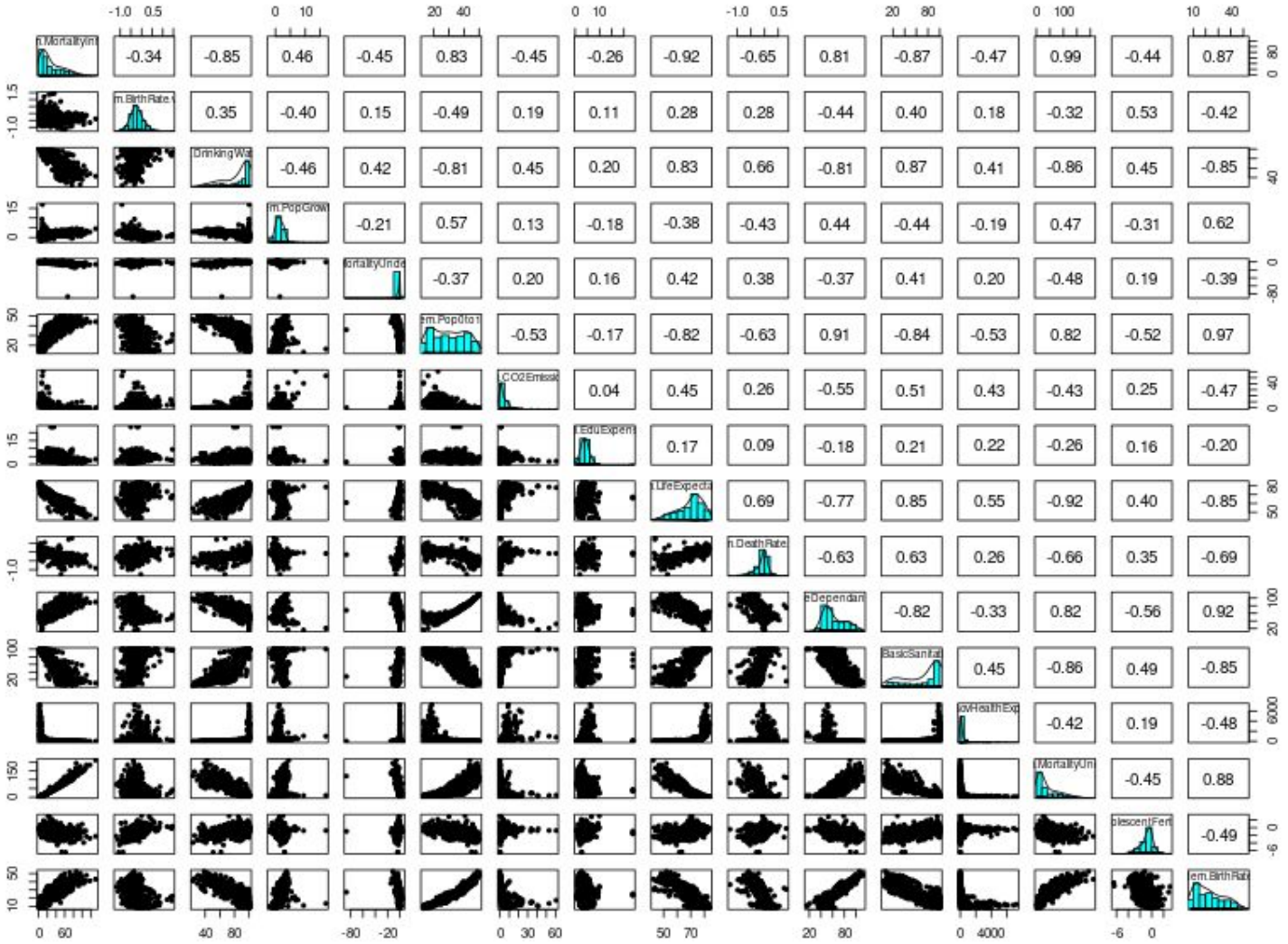


Data set feature reduction

Voting Poll: Take the role of decisor and get a (still large) subset of features to assess as predictors:

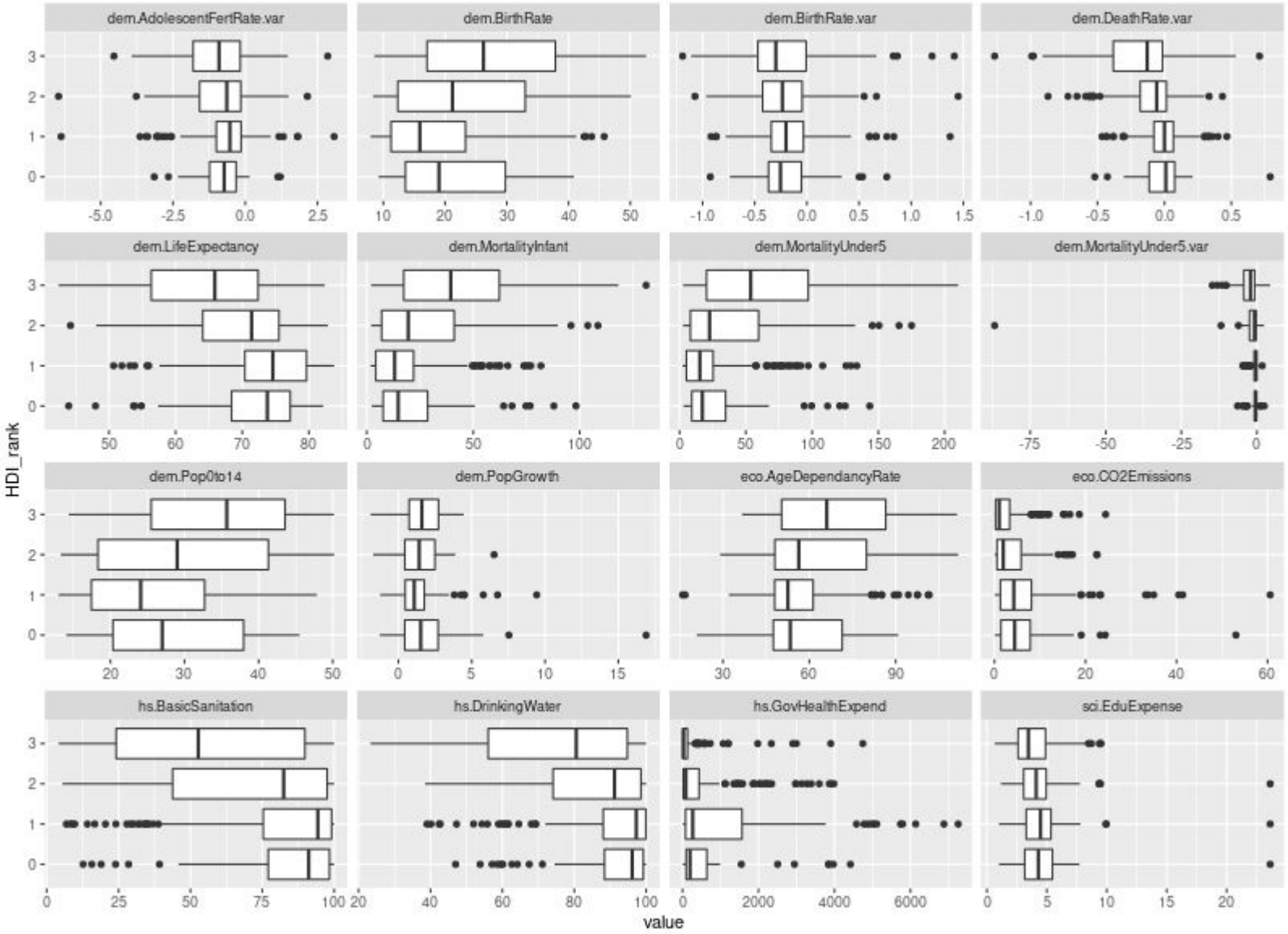
Applying mRMR feature reduction allowed us to further reduce the feature set to 16 variables.

mRMR 16 feature pairs.panels



TÉCNICO LISBOA

Boxplots by outcome



Principal Component Analysis

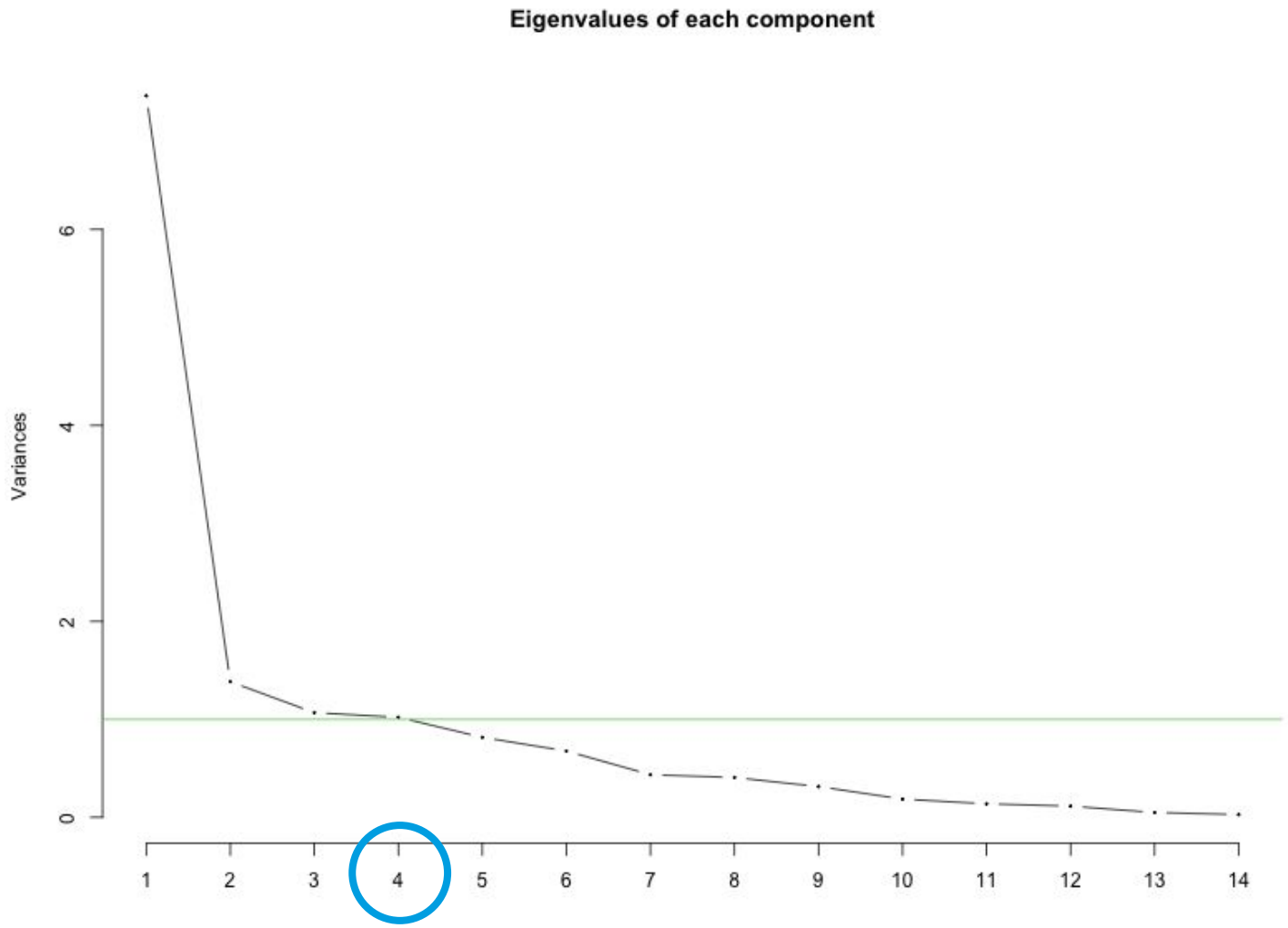
```
Call:
PcaClassic(x = x_train, scale = TRUE)
Importance of components:
                          PC1     PC2     PC3     PC4     PC5     PC6     PC7     PC8     PC9     PC10
Standard deviation     2.7004 1.15461 1.04757 1.01291 0.89882 0.82673 0.68065 0.64488 0.58849 0.44083
Proportion of Variance 0.5209 0.09522 0.07839 0.07329 0.05771 0.04882 0.03309 0.02971 0.02474 0.01388
Cumulative Proportion  0.5209 0.61611 0.69450 0.76778 0.82549 0.87431 0.90740 0.93710 0.96184 0.97572
                          PC11    PC12    PC13    PC14
Standard deviation     0.38236 0.34945 0.24689 0.10302
Proportion of Variance 0.01044 0.00872 0.00435 0.00076
Cumulative Proportion  0.98617 0.99489 0.99924 1.00000
```

1. Choose k such that $\lambda_i \geq \bar{\lambda}$, for $i = 1, \ldots, k$ → $k = 4$
2. Choose k such that

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq 0.8$$ → $k = 5$

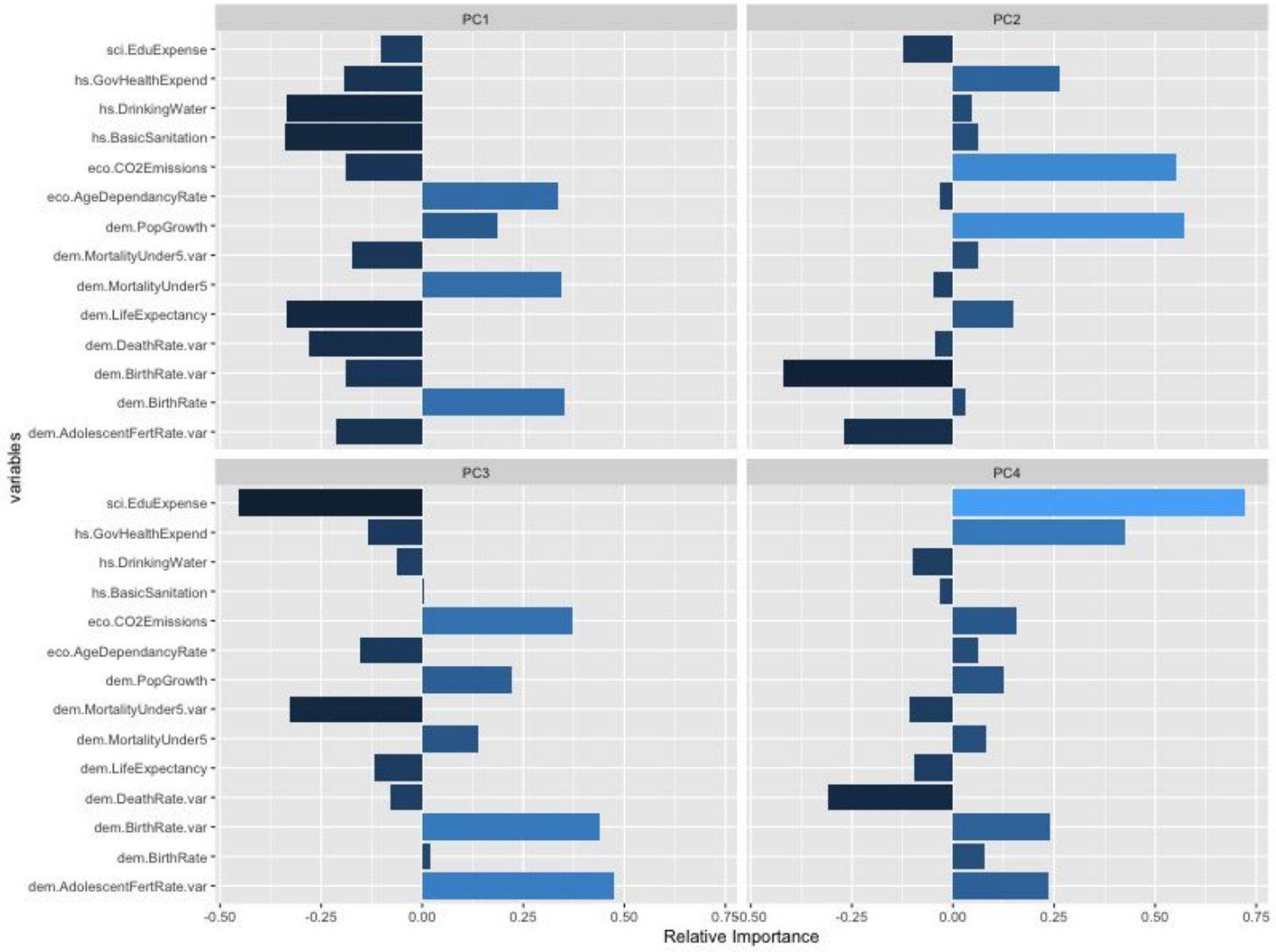
TÉCNICO LISBOA

Principal Component Analysis



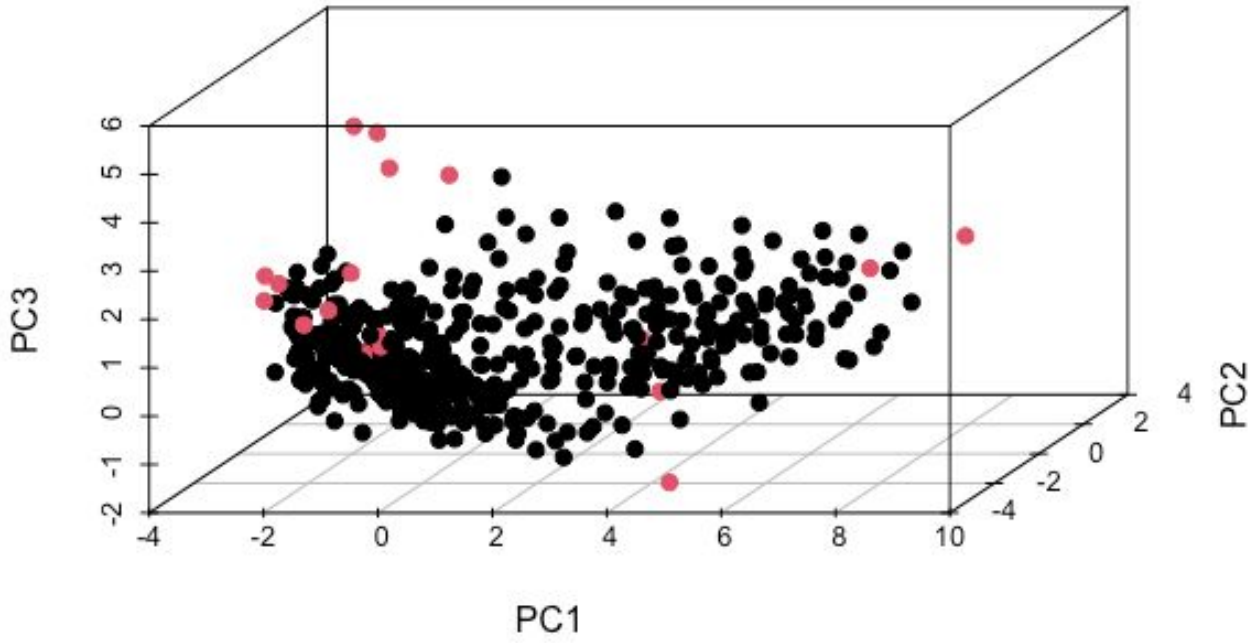
TÉCNICO LISBOA

Principal Component Analysis



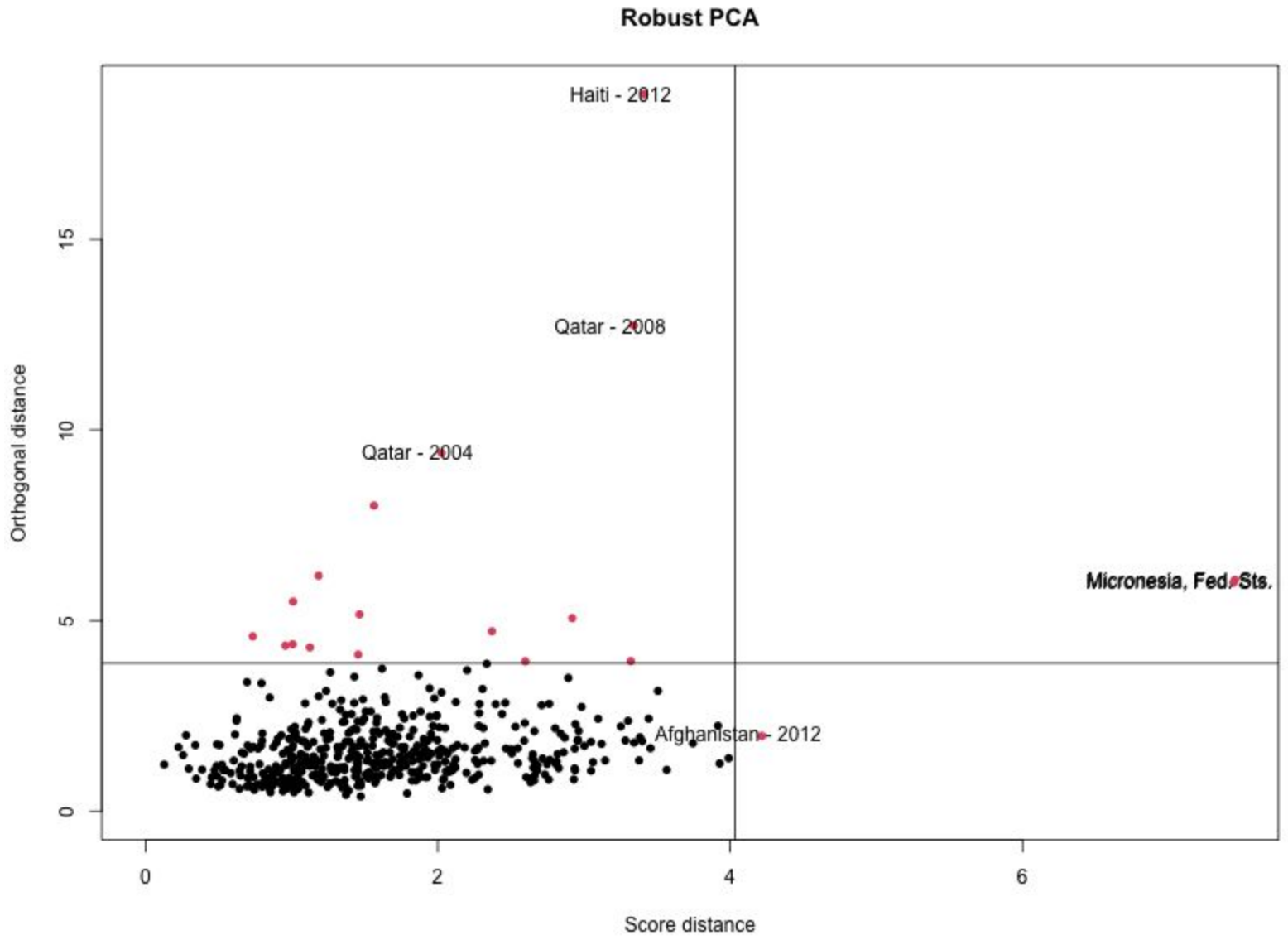
Principal Component Analysis

```
Call:
PcaHubert(x = x_train, scale = TRUE, crit.pca.distances = 0.999)
Importance of components:
                          PC1   PC2    PC3
Standard deviation      2.570 1.082 0.9064
Proportion of Variance  0.768 0.136 0.0955
Cumulative Proportion   0.768 0.904 1.0000
```



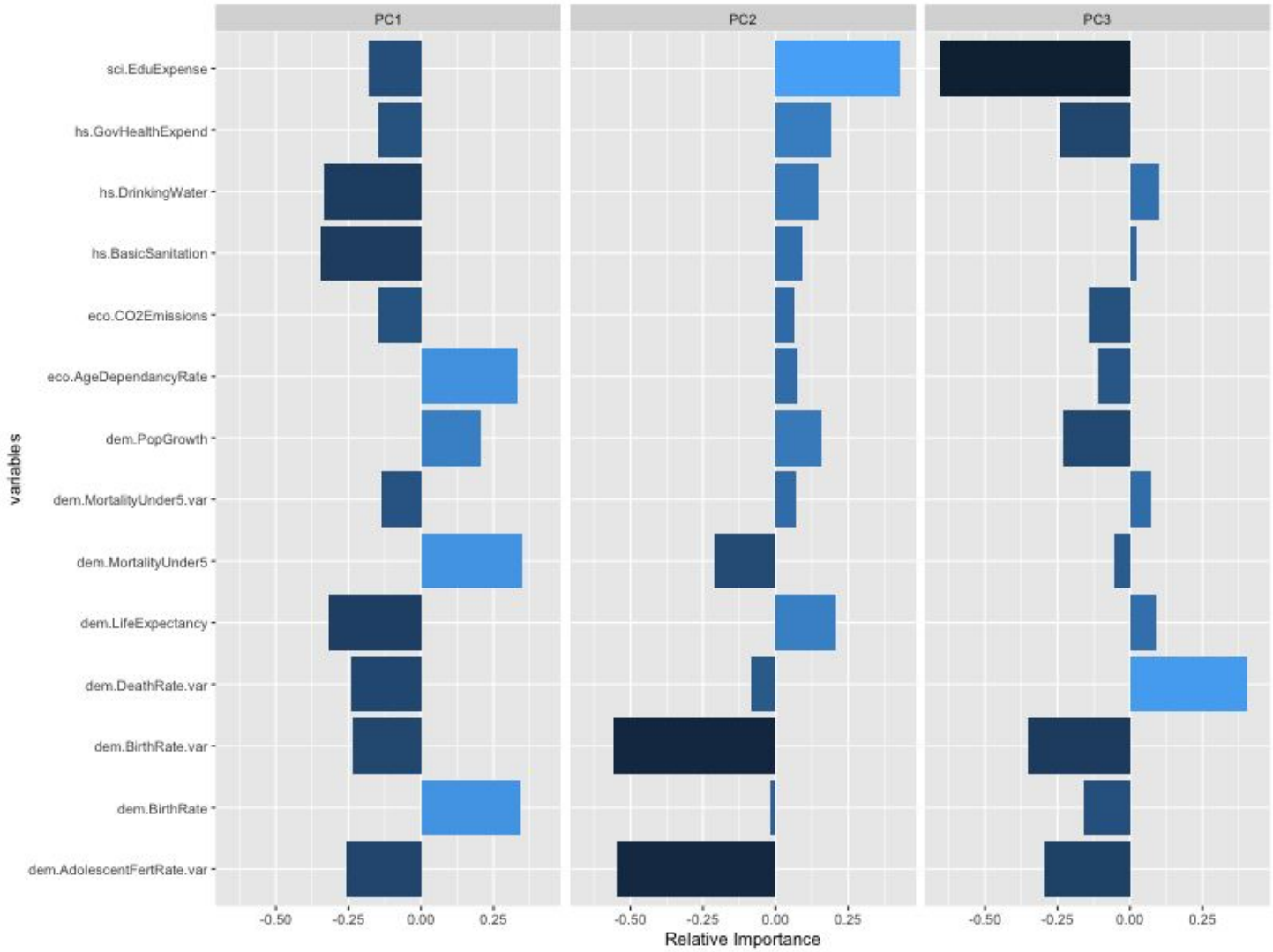
TÉCNICO LISBOA

Principal Component Analysis

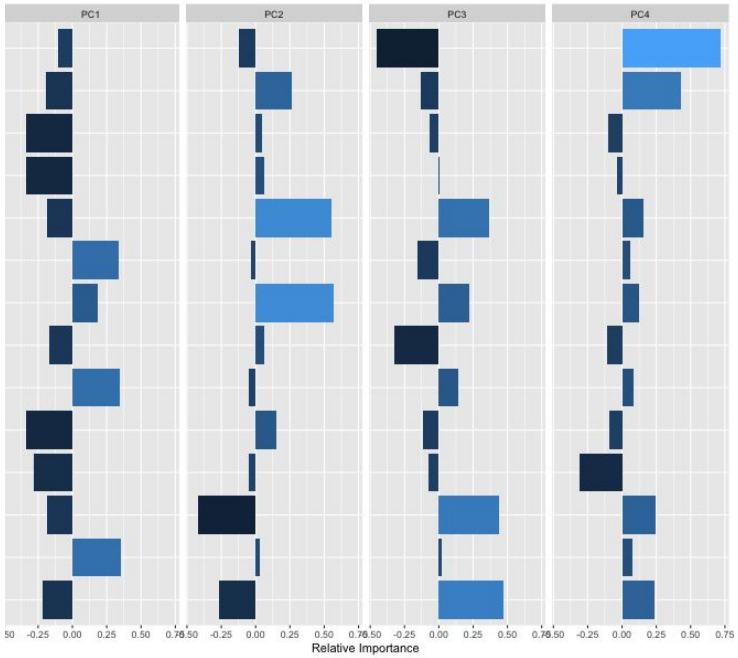


TÉCNICO LISBOA

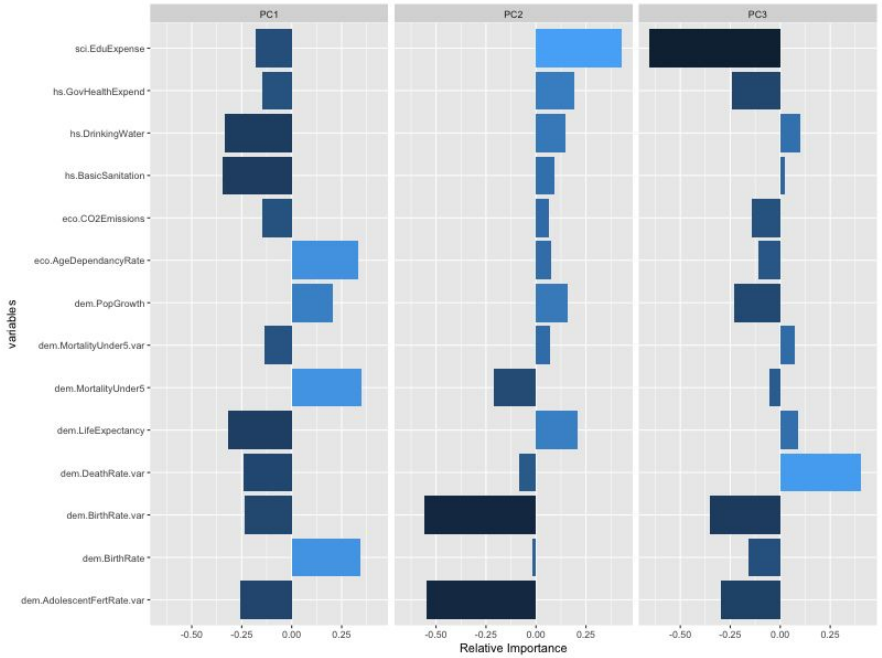
Principal Component Analysis



Principal Component Analysis



Loadings Classical



Loadings Robust

TÉCNICO LISBOA

Classification

	rMRM		rMRM+ROBPCA		rMRM+PCA	
	acc	b.acc	acc	b.acc	acc	b.acc
Random Forest	45.1	58.8	**37.4**	**54.7**	44.0	58.2
Naïve Bayes	43.0	58.6	43.1	57.1	39.7	55.5
LDA	44.5	59.2	43.3	57.2	42.5	56.9
QDA	44.5	59.6	44.9	58.2	40.9	56.3
KNN	43.3	57.5	41.1	56.1	42.3	56.8

Classification

- Predicting HDI_rank label with Random Forest

	Real			
Pred	Negative	Low	Medium	High
Negative	0	1	0	1
Low	4	21	10	5
Medium	1	10	16	14
High	4	7	11	17

Table 6: Confusion Matrix.

	Metric		
	Precision	Recall	F1-Score
Negative	0	0	-
Low	0.525	0.538	0.532
Medium	0.390	0.432	0.410
High	0.435	0.459	0.447

Table 7: Metrics for each class.

	Score
Accuracy	44.3
Balanced Accuracy	57.9

Table 8: Class prediction scores (Random Forest).

Classification

- Predicting clustering outcome

	Classes			
Clusters	Negative	Low	Medium	High
-	0	0	0	0
Three	1	12	5	3
One	6	20	18	15
Two	2	7	14	19

Table 9: Confusion Matrix combining clusters and classes.

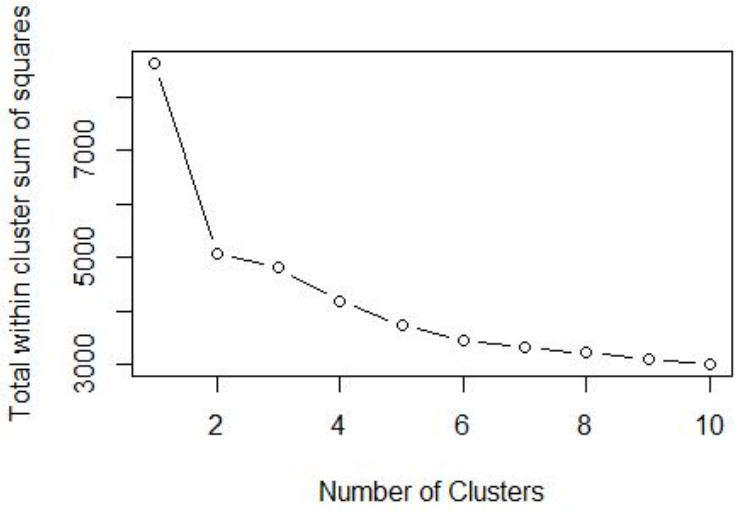
	Score
Accuracy	40.2
Balanced Accuracy	55.6

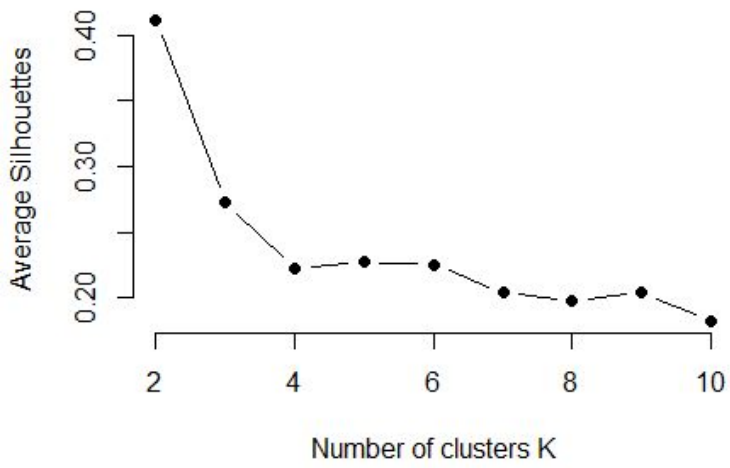
Table 10: Use of clusters for prediction.

Clustering

mRMR data set

Criteria to select k:



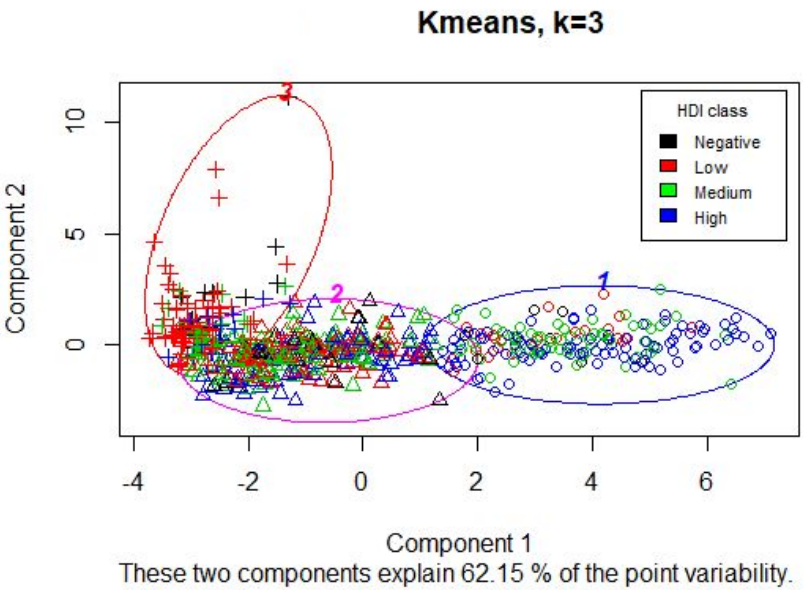


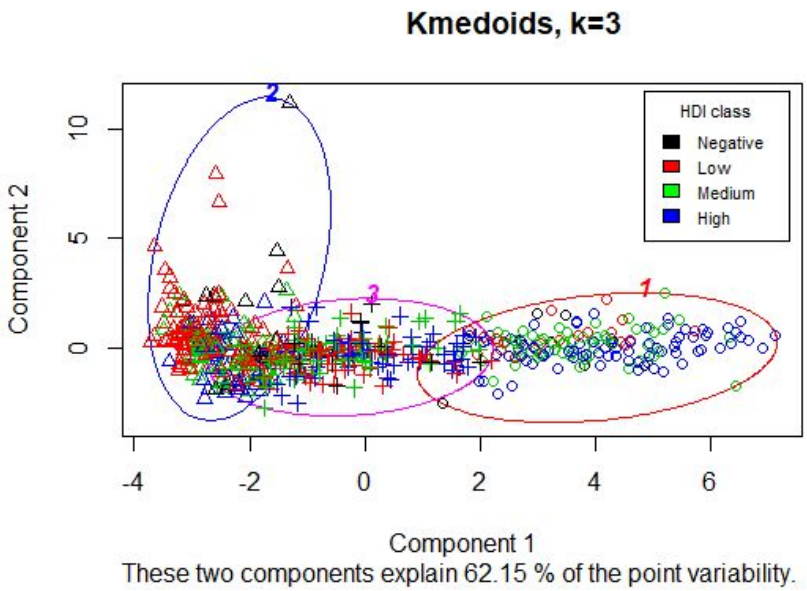
$$BdT = \frac{\sum SSE - \sum_i^k WCSS_i}{\sum SSE}$$

$$BdT_{k=2} = 0.413$$
$$BdT_{k=3} = 0.4885$$

Clustering

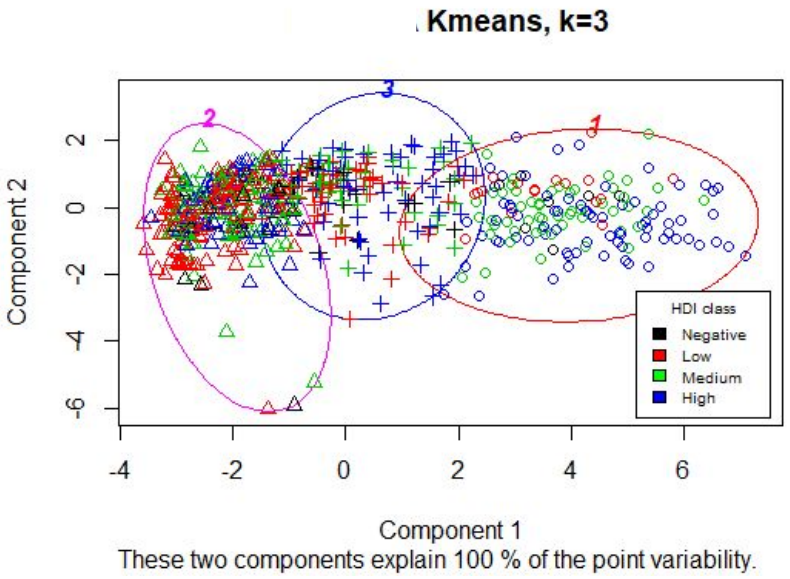
mRMR data set

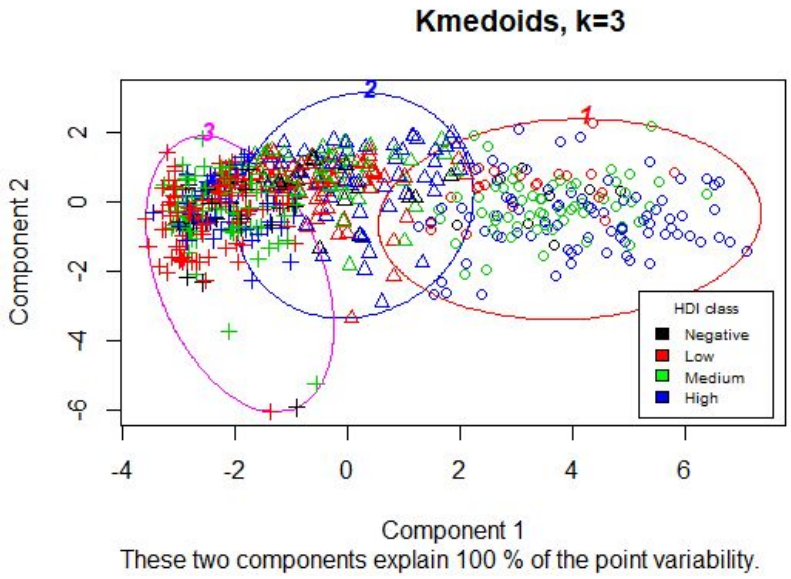




Clustering

Robust Principal Components





TÉCNICO LISBOA

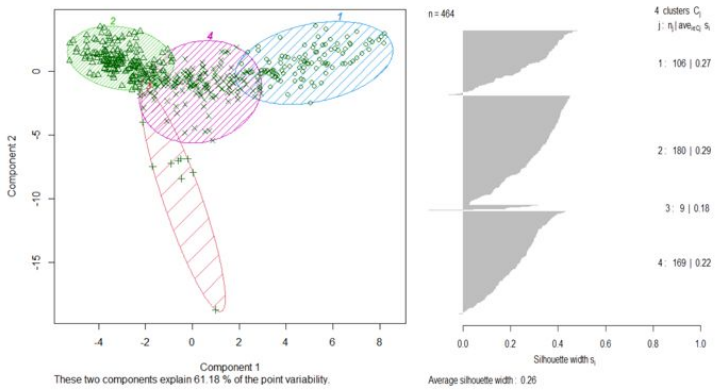
Conclusion

- There is **low predictability** of the HDI progress using WDI variables as predictors.
- The results suggest that this problem would be better tackled as a **regression problem**.
- Split the data into a **different set of classes using different criteria** and try to learn if this divisions do represent some discernible separation among the observations.

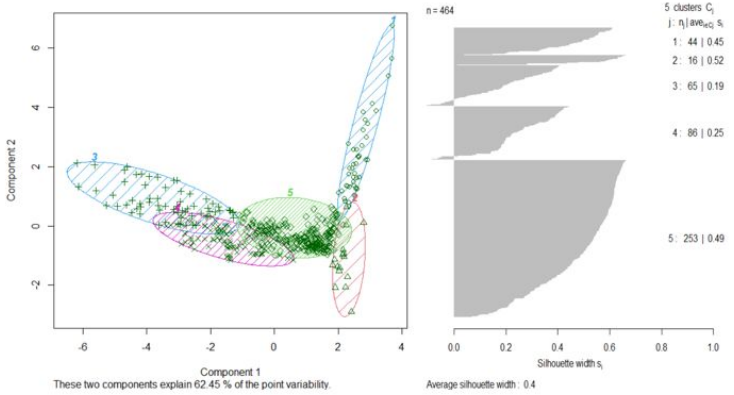
Further work

- **Data reduction through clustering**
  Clustering theme data sets using K-means

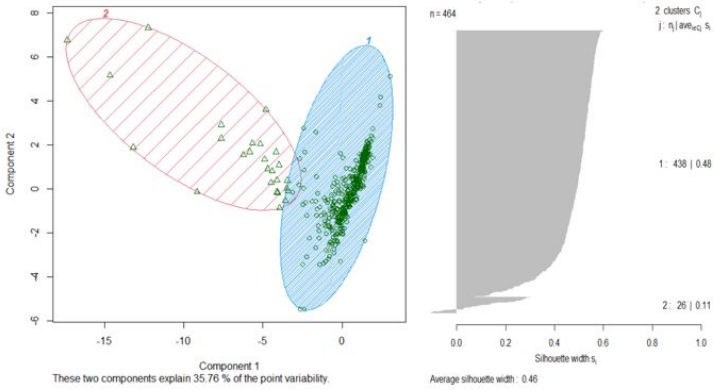
demographic k=4



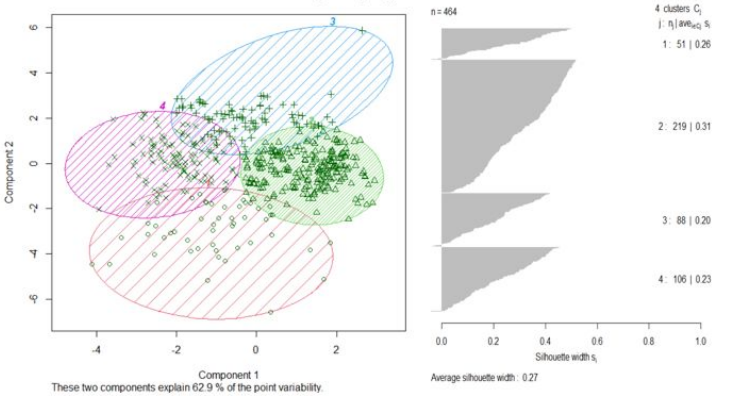
health and sanitation k=5



economics k=2



geographic k=4



TÉCNICO LISBOA

Thank you!