Predictability of Human Development Index

Pedro Sousa* Andre Luis† Carlos Sequeira‡ Sara Cruz§ Pedro Dias¶

Abstract

The United Nations Development Programme (UNDP) provides a widely accepted ranking of countries according to their human development status [1]. The annual score of each country is called Human Development Index (HDI) and it combines four development indicators: life expectancy for health, expected years of schooling, mean of years of schooling for education and Gross National Income (GNI) per capita for standard of living. Explaining and predicting the progress of such index can be used to evaluate the quality of public policies as well as to assess what better explains inequalities and certain status in human development. Factors such as available natural resources, accumulated capital, social norms, fiscal policies, re-distributive policies, political institutions quality, all should play a role in the HDI outcome, but often not in the direction intuition would point to.

1 Introduction

The Human Development Report 2019 [2] is subtitled: *Human development beyond income, beyond averages, beyond today: Inequalities in human development in the 21st century.* The report includes an in-depth exploration of the HDI progress in the 21st century up to 2018 for almost all countries in the world, and, in chapter 7, it concludes that inequalities in basic capabilities are falling, however, in some areas perceived as essential beyond 2020, inequalities in human development are growing. It also concludes that inequalities in the distribution of opportunities between men and women have improved. Then the report discusses in detail some of the spotlights derived from the data analysis, as example, it discusses how to addresses constraints in social choice, or how to balance productivity and equity with environmental sustainability.

Our project objective will be to explore the World Bank data [3] to find additional explanations predictive of the progress of HDI in the 21st century which could contribute to the discussion held in chapter 7 of [2].

Because we want to explain the progress of the Human Development Index, our *label* or *dependent* variable, will be a two year difference between HDI for every given country. In detail, such difference will be defined as the difference between the country HDI at a given year and the HDI two years later as follows:

$$HDI_{\triangle} = HDI_{y+2} - HDI_y$$

As assumption, we considered the year to year changes less informative because it is often the case that some of the statistics used to derive the HDI are not reported every year, creating an additional measurement lag between HDI and the real changes it assesses. Thus, instead, we take a two year change to have a more robust label, one that is also more sensitive to sustainable changes as opposed to one-off variations

*Instituto Superior Técnico – 42022 - pedrodesousa@tecnico.ulisboa.pt
†Instituto Superior Técnico – 98638 – andre.t.luis@tecnico.ulisboa.pt
‡Instituto Superior Técnico – 87638 – carlos.r.sequeira@tecnico.ulisboa.pt
§Instituto Superior Técnico – 79410 – sara.cruz@tecnico.ulisboa.pt
¶Instituto Superior Técnico – 39953 – pedroruivodias@tecnico.ulisboa.pt

(*e.g.*: natural catastrophe, formula revisions, election year book cooking). Moreover, because the quality of political institutions can be questioned in many countries, the quality of the statistics and reporting provided is also often questionable, thus, focusing on the variation makes the overall data analysis less biased by inter country variability in the quality of statistics or political institutions.

Our source for the HDI data was the UNDP Human Development Data Center [4]. A sample from the HDI original data set is represented by table 1

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.

The World Development Indicators, our *independent* or *explanatory variables* or *features*, were also pre-processed to include variations over 2 years, and the reasoning is the same as for the HDI, just that the difference is in this case taken comparing given year value with the value reported two years before as follows:

$$WDI_{\triangle} = WDI_y - WDI_{y-2}$$

The hypothesis is that an increasing or decreasing trend can contribute as predictor to the label variable. Taking all into account, a 4 year period is contributing to any given sample and thus, to maximize sample independence, we are not overlapping periods in the process of producing the samples out of the WDI and HDI data sets, instead, we will build one sample for each country every four years for a period that starts in 1998 and ends in 2018.

Our source for the WDI data was the World Bank [5]. A sample from the WDI original data set can be found in table 2.

2 Data set preparation

This section explains the main assumptions and the rational underlying the decisions made for the variable and object selection task that preceded the statistical analysis.

Both data sets, WDIData.csv and HDI.csv, have several country aggregates derived from countries already represented in the data set (dependent samples). For that reason we discarded aggregates from both data sets. Aggregates such as "East Asia and the Pacific", "Europe and Central Asia", "Latin America and the Caribbean", "South Asia", "Sub-Saharan Africa" or "Least Developed Countries".

Both data sets need also to be unpivoted from wide to long format before concatenating all data along the rows. Finally, all variables can become represented as columns by pivoting the concatenated rows. The result is a conventional tabular data set still quite populated with empty values.

To avoid empty values and to focus on a comprehensible set of explanatory variables, we selected 36 World Development Indicators (WDI) excluding those metrics that have more than 20 empty values and excluding variants of the same metric differing in the unit or quantity used as reference. Empty values in most cases simply mean that such metric is only available for few countries or for more recent dates.

The data set preparation includes the calculation of $WDI_{\triangle}$ as follows:

$$WDI_{\triangle} = WDI_y - WDI_{y-2}, \forall y \in \{2000, 2004, 2008, 2012, 2016\}$$

Likewise, we also calculate the HDI_{Δ} as follows:

$$HDI_{\Delta} = HDI_{y+2} - HDI_{y}, \forall y \in \{2000, 2004, 2008, 2012, 2016\}$$

And from it, we derive a classification label that splits the HDI_{Δ} into four fairly balanced bins, in detail: $HDI_{\Delta} \leq 0$; $0 < HDI_{\Delta} \leq 0.007$; $0.007 < HDI_{\Delta} \leq 0.013$ and $HDI_{\Delta} > 0.013$. The rational for this 4 bins is the following: lowest bin follows the intuition that a less than 0 HDI_{Δ} is a significant group that should have its own category. However, once most samples do register a higher than 0 value, we broke the higher than 0 portion in 3 groups using percentiles so that it results in fairly balanced classes to avoid majority class bias. As result, we got the 0.007 and 0.013 thresholds (see figure 1).

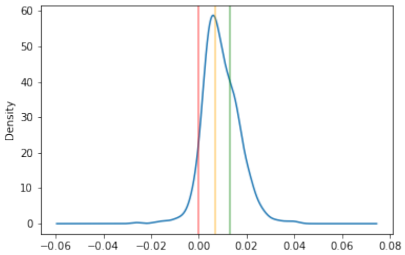


Figure 1: HDI_{Δ} distribution split into 4 classes

HDI.Delta.cat	Label	Number of samples
0	negative	36
1	low	200
2	medium	190
3	high	190

Table 3: Number of samples per class.

The outcome from the data preparation is the file represented by table 4. Moreover, we also renamed all the variables making the names both comprehensible and short enough to ease presentation in charts and tables.

Country.Year	eco.CleanCook	geo.UrbanPop	...	HDI.Delta	HDI.Delta.cat
Afghanistan - 2008	1.949	23.113	...	0.028	3
Afghanistan - 2012	4.818	23.948	...	0.009	2
Afghanistan - 2016	7.630	24.803	...	0.005	1
Albania - 2004	1.261	44.575	...	0.017	3
Albania - 2008	16.169	48.902	...	0.016	3

Table 4: Sample from the prepared data (WDI.csv file). A 616x74 matrix.

The label column *HDI.Delta.cat* has four levels, from 0 to 3, corresponding to the following labels: *negative*, *low*, *medium* and *high*, qualifying the progress of the HDI along the two years that followed the predictors measurement. It should be pointed out that the result from the proposed class split makes the dataset labels unbalanced given that *negative* labels are very under-represented compared to the other three classes as demonstrated in table 3.

As referred previously, the original indicator set, from [5], is extensive, we started with 187 potential explanatory variables, being many of them highly correlated and easily reducible using just common sense. Taking the role of project owner, the group elements took a pool on which features could be dropped before any statistical analysis. Such voting established 36 variables resulting in the set of variables listed in appendix A. All except three of the variables in appendix A include a variant concerning the difference between current year and the 2 years before (WDI_{Δ}). The exceptions are marked with "*" and the reason for such exception is the fact that the variable represents already a first order time difference.

3 Statistical Analysis

For all variables within each group, a preliminary analysis was performed, which included plotting the histogram and infer to some probable distribution from it, compute the correlation between variables and also boxplots of each variable by outcome.

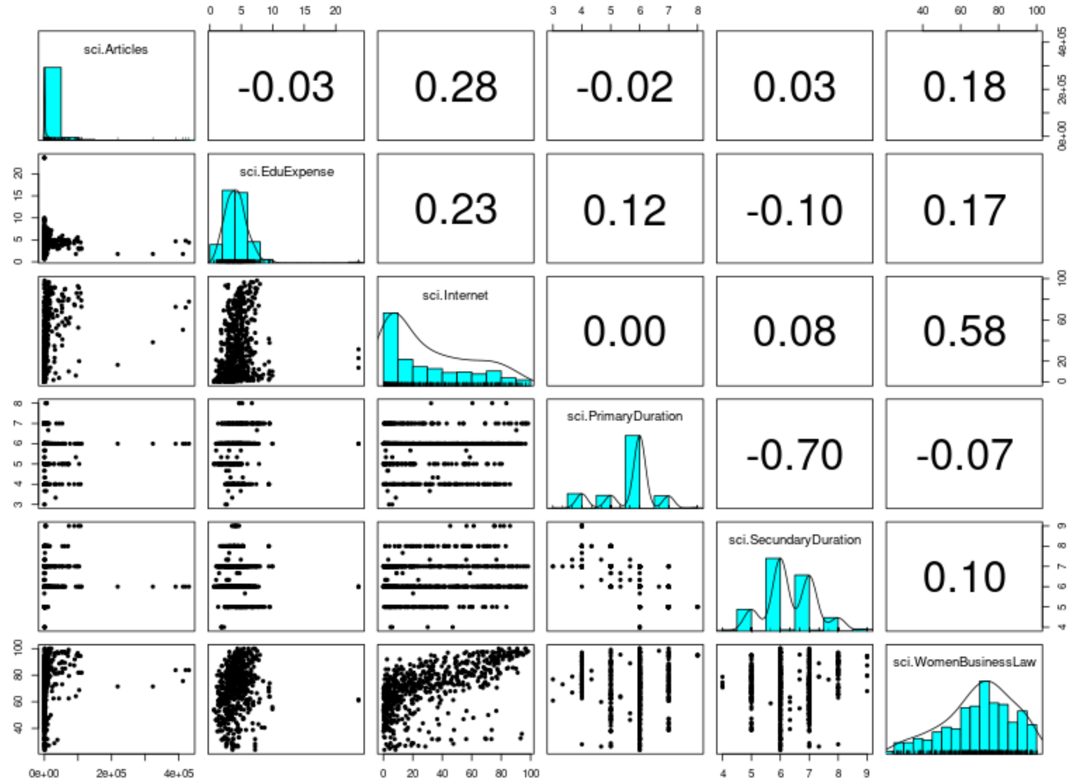


Figure 2: Pairs panels, box plots by outcome and Q-Q plots for the science and education variables.

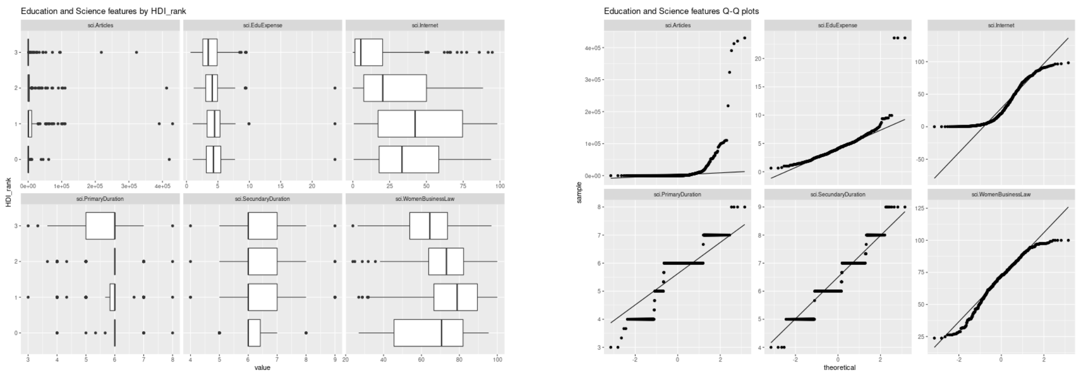


Figure 3: Box plots by outcome and Q-Q plots for the science and education variables.

For instance, from such analysis we could infer that *sci.WomenBusinessLaw* follows a distribution that is fairly normal, given the good fit of data to the quantiles of a normal distribution, as shown in the respective

Q-Q plot. Also one can observe that it is positively correlated with *sci.Internet.*

This approach, although proving to give some insight from the variables lacks the ability to identify correlations (and therefore linear interactions) between variables from distinct defined groups (and also from their variations grouped in separate sets as well). As it is very hard to assess all combinations of the considered variables in sensible amounts of time and effort, and present such an analysis in a comprehensible way.

To ease this task, we adopted an automated feature reduction strategy [1], namely the *mRMR* feature selection proposed by [6] and implemented in the R package `mRMRe`.

mRMR stands for "minimum Redundancy, Maximum Relevanc" and tries to choose a given j number of features such that, of all possible j-set of features, the one that maximizes the dependency on the target feature. This is done in an incremental procedure by calculating the information gain in order to estimate the former dependency. Using this method it is difficult to determine which is the correct number of variables will give the best result, and the number of 16 variables was chosen as a large enough number of variables which are still presentable and interpretable.

The reduced feature set that was selected was the following: *dem.MortalityInfant*, *dem.BirthRate.var*, *hs.DrinkingWater*, *dem.PopGrowth*, *dem.MortalityUnder5.var*, *dem.Pop0to14*, *eco.CO2Emissions*, *sci.EduExpense*, *dem.LifeExpectancy*, *dem.DeathRate.var*, *eco.AgeDependancyRate*, *hs.BasicSanitation*, *hs.GovHealthExpend*, *dem.MortalityUnder5*, *dem.AdolescentFertRate.var* and *dem.BirthRate.*

The `pairs.panels` plot of the mRMR 16 features subset is presented on figure 14 in appendix C.

Looking to the Pearson correlations between variable pairs (refer to figure 15 from appendix C), there are two which stand out as highly correlated: *dem.BirthRate* and *dem.Pop0to14* with a correlation of 0.97; and dem.MortalityInfant with *dem.MortalityUnder5*, having a correlation of 0.99. Also the correlation profile with the remaining features seems to be similar within each pair, and therefore one of the features in each of the highly correlated pairs can be dropped due to redundancy. As a result *dem.Pop0to14* and *dem.MortalityInfant* were taken out of the feature selection, keeping the remaining representative features.

From the heatmap, shown in C figure 15, given that the correlation is, for almost any variable, fairly significant (with absolute value not close to zero), we can deduce that dependence among variables exists and it is not negligible, the only exception being *sci.EduExpense*, which is less correlated with every other feature. We can observe also that:

- *dem.BirthRate* is positively correlated with, *eco.AgeDependancyRate* and *dem.MortalityUnder5*;
- *dem.AdolescentFertRate.var* is mildly inversely correlated with *eco.AgeDependancyRate*;
- *dem.MortalityUnder5* is strongly inversely correlated with *hs.BasicSanitation* and *hs.DrinkingWater*;
- *sci.EduExpense* is not as strongly correlated as the previous features, yet it mildly correlates inversely with *dem.MortalityUnder5*;
- *eco.CO2Emissions* are inversely correlated with *dem.BirthRate*;
- *hs.GovHealthExpend* positively correlates with *dem.LifeExpectancy.*

Also worth mentioning is the existence of some specific extreme values in the some features, which can be observed in figure 14, which can be found in appendix C:

- in *dem.MortalityUnder5.var* an occurrence stands out as an extreme, taking the value of -86.80 for Haiti in 2012, which indicates that in 2010 the mortality under 5 was abnormally high. This is most likely due to the earthquake which took place in that here and leveled Port-au Prince city to the ground, killing many people as direct consequence, but also indirectly by compromising the sanitation conditions in the affected areas for some time, which left small children susceptible to illness, disease and death;

[1] As graceful suggestion of our professor.

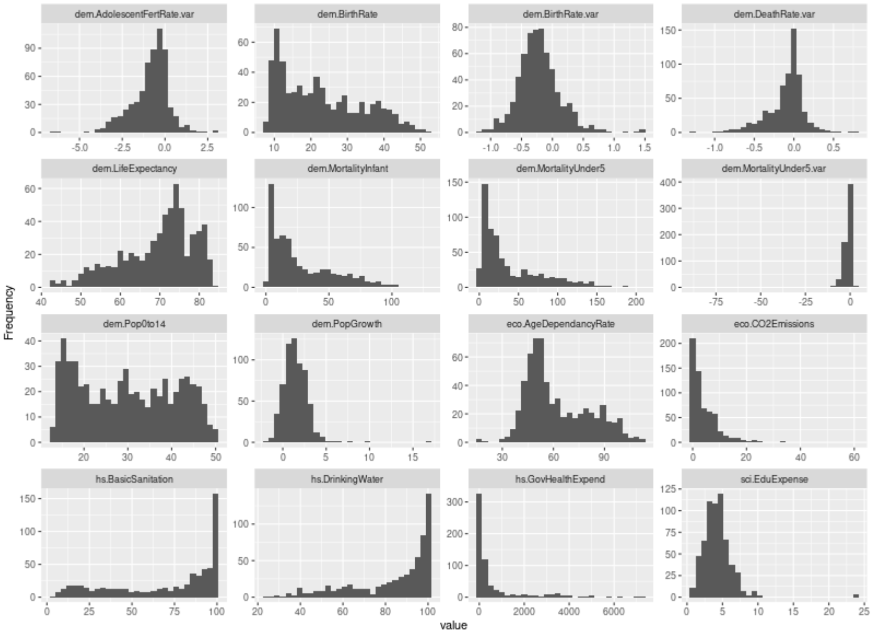


Figure 4: mRMR 16 Histograms

- in *sci.EduExpense* we can identify three values which are also very separate from all other observations, having the value of 23, 6% of the GNI. These values belong to the Micronesia Federation on the years regarding 2008, 2012 and 2016, and can be due to the government financing the expenses of students who might need to go abroad to follow and conclude their higher academic studies.

Histograms in figure 4 and Q-Q plots in figure 5 were obtained for all features selected by *mRMR*. Firstly one can identify that the redundant and excluded features do have a identical distribution to the ones that remained in the selected 14 feature set. Secondly, while some features like *dem.AdolescentFertRate.var* and *sci.EduExpense* do tend to follow some normal distribution to most extent of the observed instances, *hs.BasicSanitation* and *hs.GovHealthExpend* among others do not, and this can have implications in the following tasks and should be taken into account.

Finally box plots by outcome were also obtained in figure 6, and although there are features to which each group mean is different for each class, the variation of mean is not monotonous and the there is a extensive overlap among classes within each feature, without any exception. This most likely indicates that the class outcome is not very discernible by the chosen (and available) predictor variables, and so the classifier that is to be implemented will most likely struggle to have a decent performance in identifying the correct class of any occurrence in such indiscernible population.

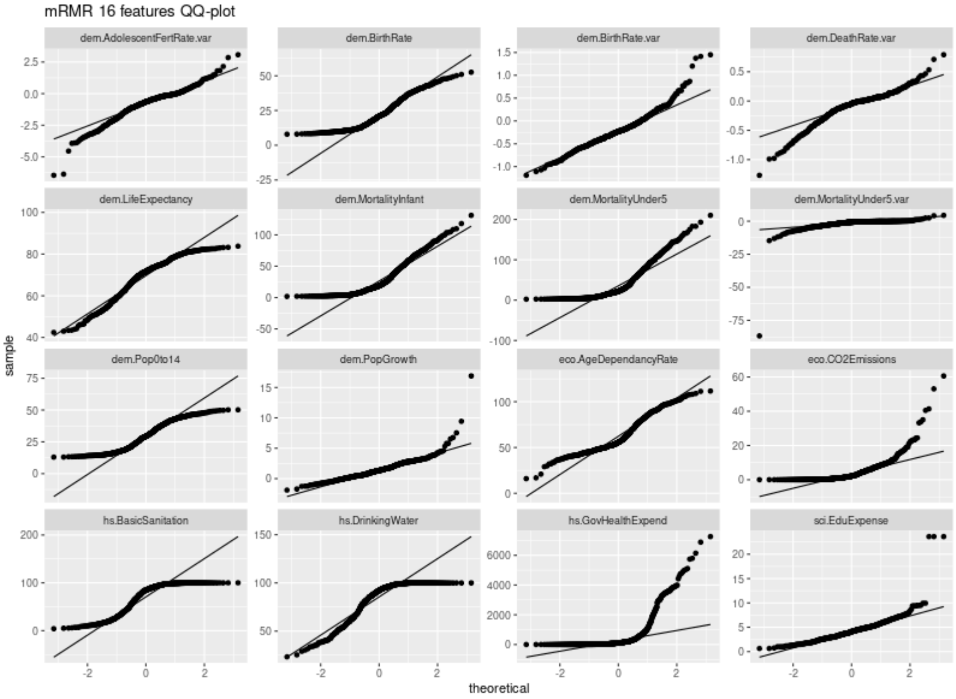


Figure 5: Q-Q plots of mRMR reduced dataset

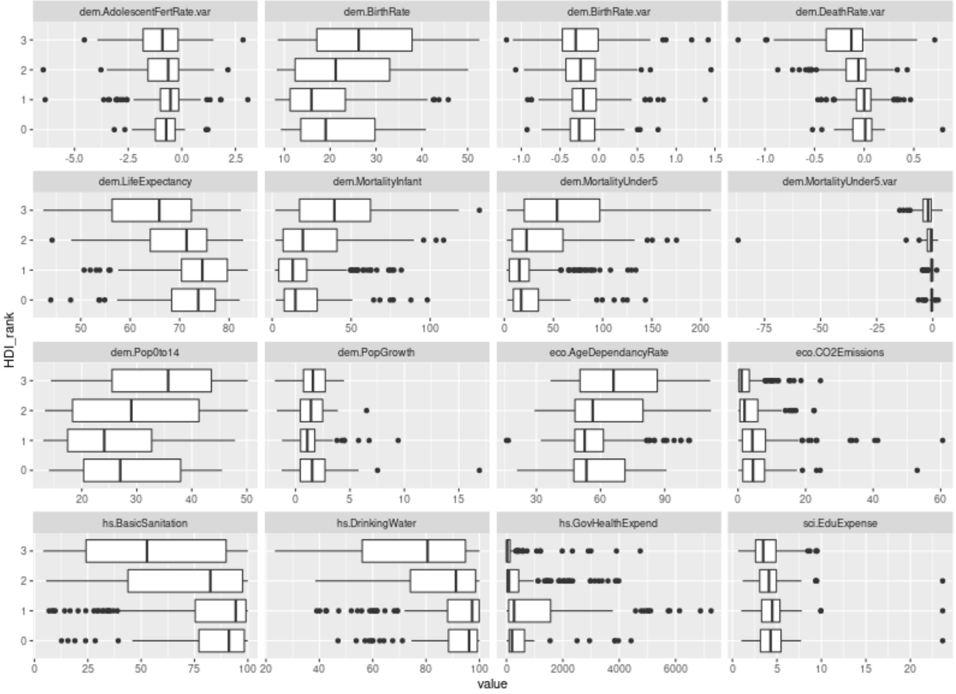


Figure 6: Box plots by *HDI_rank* of mRMR reduced dataset

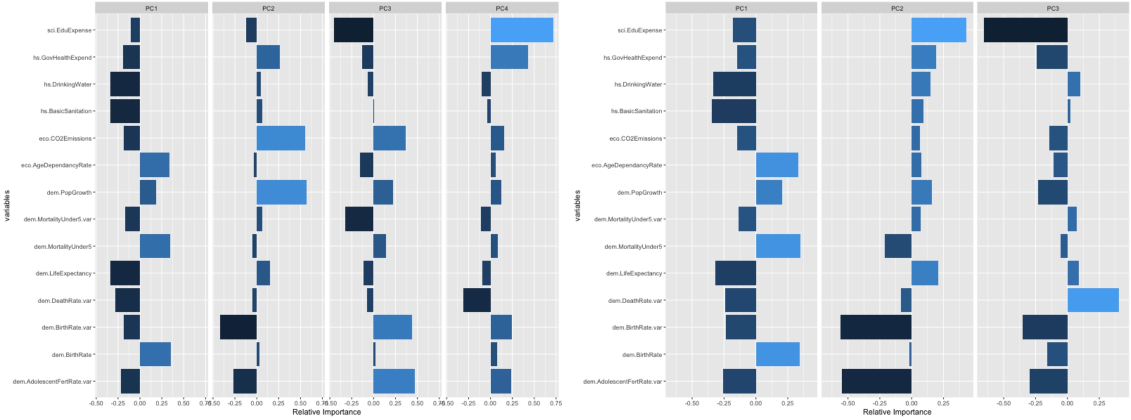
3.1 Principal Component Analysis

Even though the previous analysis allowed a reduction from roughly 70 to 14 explanatory variables, this number could still be intractable for some further study and interpretation. Thus, we proceed by analyzing the principal components, in order to find the linear combinations within the set of variables that better explain the variability of the data, and provide a smaller and more practicable set of uncorrelated variables to continue our work, as described in [7] and [8].

Prior to the Principal Component Analysis, and because the output from PCA is later used in a classification task, we split the data between *train* (80%) and *test* (20%) using the R package `caret` to ensure all labels are represented in both sets as they are in the complete data set. The principal components were learned from the training set and then applied to the test set after the analysis and before the clustering exercise, which takes the complete data as input.

Since there are different scales among data, we use the standardized version of the PCA, so that it is able to better reveal relationships between variables and be more informative. To perform the classical PCA, we used the available method implemented in the R package `rrcov`. Upon the event of choosing the number of components to retain, we resorted to two criteria and chose the minimum k among those. The first was to find k such that $\lambda_i \geq \bar{\lambda}$, for $i = 1, \dots, k$, *i.e.*, find the principal components whose variance is above the mean of all variance, which, since we are dealing with standardized data, is 1. The second was to find k such that $\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{p} \lambda_j} \geq 0.8$, meaning that 80% of the variance is explained by the first k principal components. The method that yielded the lowest value of k was the first one and hence we proceeded with 4 principal components.

With this analysis we are also able to assess the relative importance of each variable at each principal component by observing the magnitude of the respective loadings. As seen in figure 7a, the coefficients of the first eigenvalue vary between -0.35 and 0.35, and, since the relative importance of a variable is given by the magnitude of its corresponding coefficient, rather than its sign, we can assess that there are several variables contributing to the variance explained by the first component (52.6%) which are *hs.DrinkingWater*, *hs.BasicSanitation*, *eco.AgeDependencyRate*, *dem.MortatilityUnder5*, *dem.LifeExpectancy*, and *dem.BirthRate*.



(a) Classical PCA (selected principal components).

(b) Robust PCA.

Figure 7: Relative importance of each variable given the loadings of the PC's.

Not only is PCA scale sensitive, as its results might be severely biased by the presence of outliers, as pointed out by Hubert, Rousseeuw, and Vanden Branden in [9]. The authors propose a robust method for principal component analysis (ROBPCA) which is also implemented in the same `rrcov` package. We apply the method to the data in order to not only assess the presence of outliers, but also to proceed the work with a sounder combination of variables.

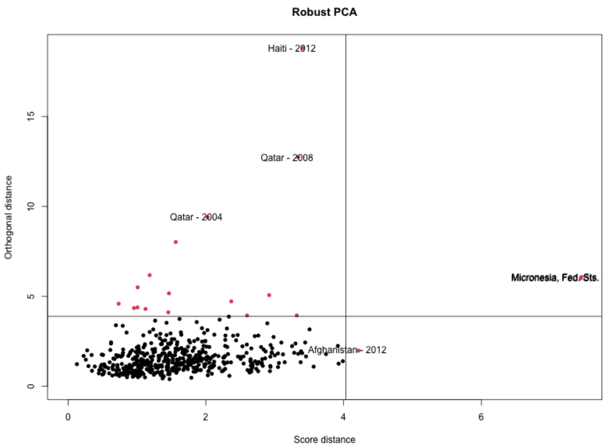


Figure 8: Outlier detection with robust PCA.

ROBPCA classifies the data points into four types: the regular observations, the good leverage points (points which are close to the PCA space, but far from the regular observations), the orthogonal outliers (points whose orthogonal distance to the PCA is large, but lie among the other points when projected on the PCA space), and the bad leverage points, which have both large orthogonal distance and score distance (distance on the PCA space). As we can observe in figure 8, there are two almost overlapped points of the bad leverage type corresponding to the observations of the Micronesia Federation, which have already been identified on the preliminary analysis, some borderline orthogonal outliers, which appear in red above the horizontal line, and some prominent as the observation Haiti-2012, an the two observations of Qatar. Finally it is possible to observe a good leverage point – the Afghanistan-2012 observation, whose orthogonal distance is small, but yet lies beyond the threshold of the accepted distance on the PCA space.

Despite the presence of outliers, the relative importance of each variable on the first principal component does not differ too much from the classical PCA. But we can observe some differences on the two remaing components. A notable aspect of the differences between both methods is the fact that the robust PCA is able to explain 100% of the total variance while yielding only 3 components.

4 Clustering

In this section we will look for natural clusters in the data set that resulted from applying mRMR feature selection and after the Principal Component analysis. We will explore the natural clusters using K-means [10] and K-medoids [11]. With K-medoids we can take advantage from its robustness in dealing with outlier samples and from its flexibility to take distance metrics other than Euclidean.

4.1 K-means clustering of mRMR data set

The choice regarding the K-means number of clusters will be taken inspecting the outcome from three distinct methods: the Elbow method [12], the average silhouette coefficient [13], and finally validated using the ratio between the between sum of squares and the total sum of squares (BdT) calculated as follows: $BdT = \frac{\sum SSE - \sum_i^k WCSS_i}{\sum SSE}$ (from [14]). The Elbow method, refer to figure 10, suggests that k=2 is an appropriate number of clusters. The silhouette coefficient suggests also that k=2 or k=3 provides good separation between the clusters (> 0.25). Finally, to decide between k=2 and k=3, we compare result from the BdT index for each k: once for $BdT_{k=2} = 0.413$ and $BdT_{k=3} = 0.4885$, $BdT_{k=3} > BdT_{k=2}$, so we finally go for $k = 3$. The final clusters are represented in figure 11a as a 2d projection based on the first two

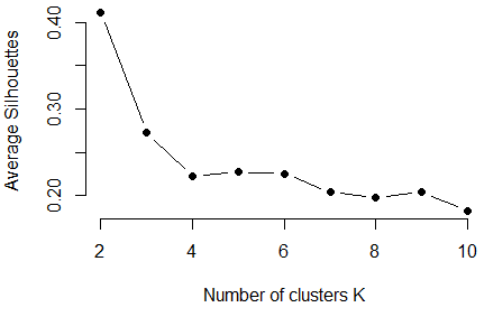


Figure 9: Average silhouette coefficient vs k

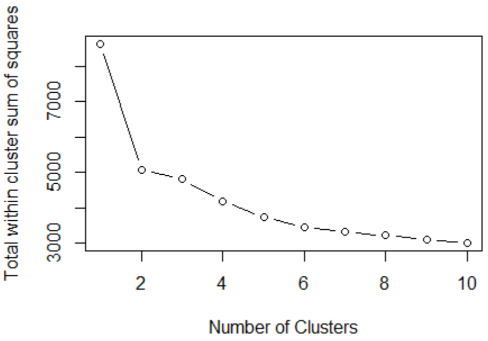
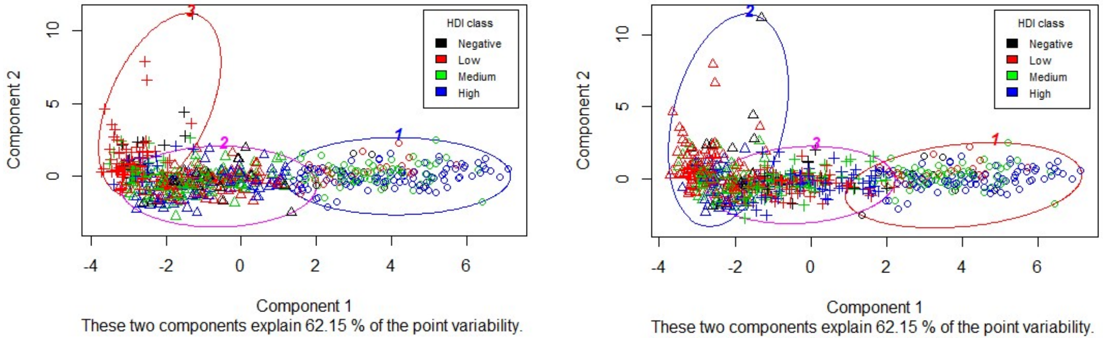


Figure 10: Elbow method

principal components. It is a reliable representation of the clusters separation since it accounts for 62% of the variance, giving a good intuition on the final result.

4.2 K-medoids clustering of mRMR data set

Following the same criteria to choose the number of clusters, we performed K-medoids on the mRMR feature selection data set, and represent the final clusters (k=3) as a 2d projection in figure 11b. The use of K-medoids is convenient when the number of variables tends to be quite small increasing the chances of having outlier samples caused by single variable extreme values, making necessary a method less sensitive to outliers. Additionally, we tested Manhattan distance because and it outperformed Euclidean distance when comparing cluster separability.



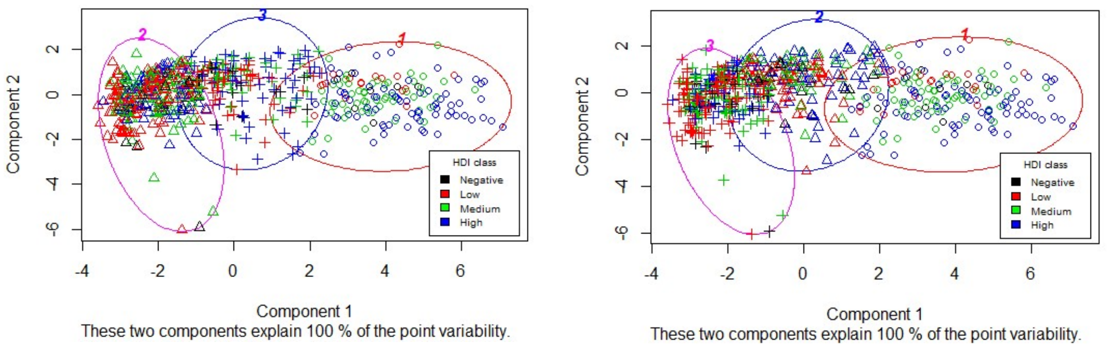
(a) 2d projection of K-means clusters for k=3

(b) 2d projection of K-medoids clusters for k=3

Figure 11: K-means and K-medoids clustering on the mRMR feature selection data set.

4.3 K-means and K-medoids clustering on the Robust Principal Components

In order to work on a smaller feature space, we performed clustering on the classical and robust PCA results presented above, taking advantage of the interpretation of each principal component to characterize our cluster solution. The criteria to choose k was the same as above and the result is k = 3. Specifically on the Robust PCA, the first 2 robust principal components were used and produced the results presented in figures 12a and 12b.



(a) 2d projection of K-means clusters for k=3

(b) 2d projection of K-medoids clusters for k=3

Figure 12: K-means and k-medoids on the robust principal components.

5 Classification

To solve the classification problem: the prediction of the factorized HDI_Δ, we will test several commonly used classifiers, namely Random Forest, Naïve Bayes, Linear Discriminant Analysis, Quadratic Discriminant Analysis and K-Nearest Neighbors. To test the classifiers we are going to take the train data set and do 5-Fold Cross-Validation while doing grid search to find the best hyper-parameters for tuning each classifier. Both these procedures are streamlined in the `caret` R package and all the tested classifiers are also supported by `caret`. As a relevant remark, we are not repeating the PCA for every validation fold, we assume that there can be some over-fit in the validation but it shall not affect significantly the model choice.

	rMRM label		rMRM+ROBPCA label		rMRM+PCA label		rMRM cluster	
	acc	b.acc	acc	b.acc	acc	b.acc	acc	b.acc
Random Forest	45.1	58.8	37.4	54.7	44.0	58.2	94.9	95.7
Naïve Bayes	43.0	58.6	43.1	57.1	39.7	55.5	91.8	94.8
LDA	44.5	59.2	43.3	57.2	42.5	56.9	95.1	95.2
QDA	44.5	59.6	44.9	58.2	40.9	56.3	93.1	95.6
KNN	43.3	57.5	41.1	56.1	42.3	56.8	94.5	94.2

Table 5: Validation scores for each classifier and for each dimensionality reduction technique.

Table 5 summarizes the best scores obtained for each classifier and the outcome is that, to predict the HDI progress, the best choice is the use of Random Forest [15] before applying any PCA.

5.1 Classification predicting *HDI_rank* label

To provide performance metrics for the final model predicting the HDI progress, we made estimates using the test set and summarized the result in tables 6, 7 and 6.

	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).

5.2 Classification predicting clustering outcome

We decided to present the results from the Random Forest classifier in the clustering prediction. These results are presented in tables 9 , 10 and 11. As we can observe in the confusion matrix, only two predictions were wrong resulting in 95.2% accuracy and 96.1% balanced accuracy.

	Real		
Pred	One	Two	Three
One	54	0	0
Two	2	42	0
Three	3	0	21

Table 9: Confusion Matrix.

	Metric		
	Precision	Recall	F1-Score
One	1.0	0.915	0.956
Two	0.955	1.0	0.977
Three	0.875	1.0	0.933

Table 10: Metrics for each cluster.

	Score
Accuracy	95.9
Balanced Accuracy	97.7

Table 11: Cluster prediction scores (Random Forest).

The performance obtained when predicting the cluster in the test set excels in all metrics, being almost perfect in every class. Although this outcome is somehow predictable. the problem lies in the interpretation of each obtained cluster: how can we assign meaning to the partitioning obtained? The overlap of the original classes with the cluster subject to the new classification presented in the clustering representation in figures 12 shows that in every cluster one can find at least some few representatives of the original classes. Also along the PCA's axis used to represent the clusters, there seems to be no clear tendency that can separate the original classes as intended by the objective of classifying the *HDI_rank*. In the end, the success of a classification exercise implies that the clustering is also able to split the data set into some properly separated clusters, but what indeed separates each of the obtained cluster is of little use for the classification exercise.

The clustering provides a fairly clear data separation, and if the objective is to find some attribute that distinguishes between elements of some set, then clustering might aid into such a task, providing also meaning to the obtained partitions, which tends to be a challenging task. On the other hand, classification can be challenging if the explanatory variables available are in fact irrelevant to the classes that are the target of the prediction. Yet, regardless of unimpressive performance of the model obtained using the variables, it is always possible to make some interpretation of the outcome, worst case scenario, we can conclude that the used data might be irrelevant given the objective.

6 Conclusion

From the results we can conclude that there is predictability of the HDI progress using WDI variables as predictors. If that was not the case we would expect a balanced accuracy close to 50%, and an accuracy close to 33% given that we have four classes and one of them is very under represented in the data set. Instead, we observed an accuracy of 44% and balanced accuracy of 58% (results from table).

We addressed the HDI progress prediction as a classification problem but the results suggest that it could be better tackled as a regression problem. This conclusion also derives from the strategy used to split the classes, in fact we are taking a continuous quantity and splitting it in ranges without using any natural boundary found in the distribution, as there was none. As expected, there is no natural clear split between classes and the cluster analysis also supports this conclusion. So, as further work, it would be interesting to tackle the problem with regression methods.

Also as further work, if persisting in addressing the HDI progress prediction as a classification problem, we suggest to explore the *Data reduction through clustering* technique from [16] and demonstrated at superficial level in appendix B.

Another hypothesis might be to try to split the data into a different set of classes using different criteria, in order to maximize the chances of having a useful classifier, and try to learn if this divisions do represent some discernible separation among the observations.

What we can imply from this analysis is that, although the indicators used as predictors are available, they either are unrelated with the outcome or the time lapse considered is not relevant to the outcome: it might be that two year difference is too short for some features and too long for others, and while in some cases the effect of one effect might already have passed in the 2-year time considered, for another feature the effect is not yet reflected in the outcome. A deeper analysis, considering several periods for variation (or even a time series analysis) might improve the results.

References

[1] Wikipedia, “Human development index — wikipedia, the free encyclopedia,” 2020, [Online; accessed 30-November-2020]. [Online]. Available: https://en.wikipedia.org/wiki/Human_Development_Index

[2] U. N. D. Programme, *Human Development Report 2019*, 2019. [Online]. Available: https://www.un-ilibrary.org/content/publication/838f78fd-en

[3] W. Bank, “DataBank World Development Indicators.” [Online]. Available: https://databank.worldbank.org/reports.aspx?source=world-development-indicators#

[4] U. N. D. Programme, “Human development data center - human development reports,” 2020, [Online; accessed 30-November-2020]. [Online]. Available: http://hdr.undp.org/en/data

[5] W. Bank, “World development indicators - databank,” 2020, [Online; accessed 30-November-2020]. [Online]. Available: https://databank.worldbank.org/source/world-development-indicators

[6] Hanchuan Peng, Fuhui Long, and C. Ding, “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[7] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning.* Springer, 2013, vol. 112.

[8] K. Mardia, J. Kent, and J. Bibby, *Multivariate Analysis.* Academic Press, 1979.

[9] M. Hubert, P. J. Rousseeuw, and K. Vanden Branden, “Robpca: a new approach to robust principal component analysis,” *Technometrics*, vol. 47, no. 1, pp. 64–79, 2005.

[10] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” in *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, L. M. L. Cam and J. Neyman, Eds., vol. 1. University of California Press, 1967, pp. 281–297.

[11] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis.* John Wiley, 1990.

[12] Wikipedia, “Elbow method (clustering) — wikipedia, the free encyclopedia,” 2020, [Online; accessed 30-November-2020]. [Online]. Available: https://en.wikipedia.org/wiki/Elbow_method_(clustering)

[13] P. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53–65, Nov. 1987. [Online]. Available: http://svn.donarmstrong.com/don/trunk/projects/research/papers_to_read/statistics/silhouettes_a_graphical_aid_to_the_interpretation_and_validation_of_cluster_analysis_rousseeuw_j_comp_app_math_20_53_1987.pdf

[14] L. Mouselimis, “Clusterr,” May 2020. [Online]. Available: https://cran.r-project.org/web/packages/ClusterR/vignettes/the_clusterR_package.html

[15] L. Breiman, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324

[16] R. Evans, B. Pfahringer, and G. Holmes, “Clustering for classification,” 08 2011, pp. 1 – 8.

A Variables list

- demographic:
  - dem.AdolescentFertRate: Adolescent fertility rate expressed in number of births per 1000 women ages 15-19;
  - dem.BirthRate: Birth rate, crude (per 1000 people);
  - dem.DeathRate: Death rate, crude (per 1000 people);
  - dem.FemalePop: Population, female (% of total population);
  - dem.FertilityRate: Fertility rate, total (births per woman);
  - dem.LifeExpectancy: Life expectancy at birth, total (years);
  - dem.MortalityInfant: Mortality rate, infant (per 1000 live births);
  - dem.MortalityUnder5: Mortality rate, under-5 (per 1000 live births);
  - dem.over64: Population ages 65 and above (% of total population);
  - dem.Pop0to14: Population ages 0-14 (% of total population);
  - dem.Pop15to64: Population ages 15-64 (% of total population);
  - dem.PopGrowth: Population growth (annual %) *,
- economics:
  - eco.AgeDependancyRate: Age dependency ratio (% of working-age population);
  - eco.CleanCook: Access to clean fuels and technologies for cooking (% of population);
  - eco.CO2Emissions: CO2 emissions (metric tons per capita);
  - eco.Exports: Merchandise exports (current US$);
  - eco.FoodProdIdx: Food production index (2004-2006 = 100);
  - eco.GDP: GDP (current US$);
  - eco.Imports: Merchandise imports (current US$);
  - eco.Inflation: Inflation, GDP deflator (annual %);
  - eco.MerchTrade: Merchandise trade (% of GDP),
- geographic:
  - geo.ArableLand: Arable land (% of land area);
  - geo.RuralPop: Rural population (% of total population);
  - geo.RuralPopGrowth: Rural population growth (annual %) *;
  - geo.UrbanPop: Urban population (% of total population);
  - geo.UrbanPopGrowth: Urban population growth (annual %) *,
- health and sanitation:
  - hs.BasicSanitation: People using at least basic sanitation services (% of population);
  - hs.DrinkingWater: People using at least basic drinking water services (% of population);
  - hs.GovHealthExpend: Domestic general government health expenditure per capita (current US$);
  - hs.OpenDefecation: People practicing open defecation (% of population),
- education and science:
  - sci.Articles: Scientific and technical journal articles;

- sci.EduExpense: Adjusted savings: education expenditure (% of GNI);
- sci.Internet: Individuals using the Internet (% of population);
- sci.PrimaryDuration: Primary education, duration (years);
- sci.SecundaryDuration: Secondary education, duration (years);
- sci.WomenBusinessLaw: Women Business and the Law Index Score (scale 1-100).

B Data reduction through clustering

In this appendix we demonstrate the use of the technique *data reduction through clustering* as proposed in [16]. Without exploring the proposed technique or the results in much detail, we will demonstrate that this different approach is a viable alternative compared to the use of mRMR plus PCA.

B.1 Clustering theme data sets using K-means

The choice regarding the number of clusters will be taken inspecting visually the 2d projection of the first two principal components and using the average silhouette coefficient, we want it to be higher than 0.25 on all clusters. Figure 13 summarizes the decision and information used to select k for each data set (*i.e.* one k for each theme).

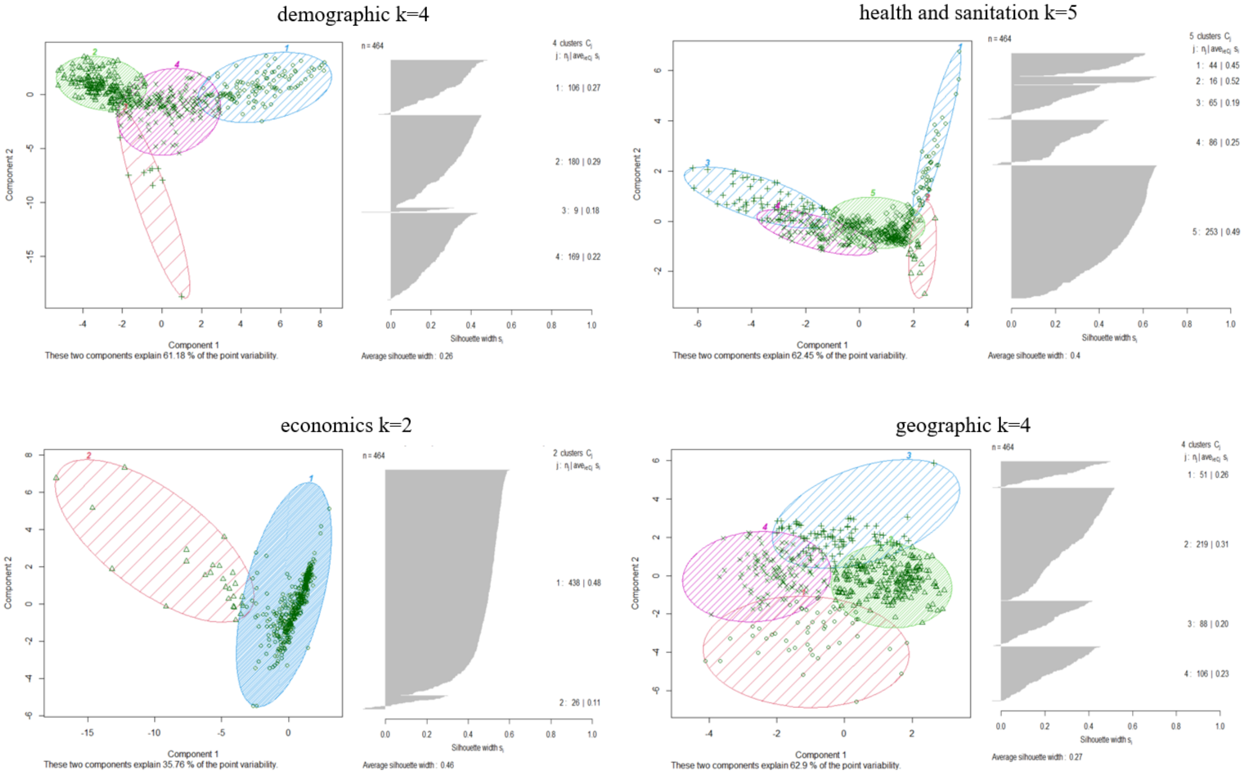


Figure 13: Outcome of K-medoids for each theme.

B.2 Classification using clustering outcome

The classification technique chosen for this task was Random Forests [15] because it can take decision trees as weak learners and these handle, by design, categorical variables such as the output from the clustering data reduction. Using Random Forests with decision trees as weak learners, we do not need to convert our variables to binary, *i.e.*: to create dummies. We however need to use some strategy in order to handle the minority class (labeled *negative*) so that it becomes represented in the prediction and the balanced accuracy score is not so affected. To handle imbalanced classes there are several options, we chose to use class weights because undersampling is not an option given the already small size of the dataset, and oversampling is also not a good solution when all the variables have distributions that cannot be easily modeled.

	Real			
Pred	Negative	Low	Medium	High
Negative	1	0	2	0
Low	5	26	14	9
Medium	6	16	22	20
High	0	6	8	17

Table 12: Confusion Matrix for Random Forest classification after data reduction through clustering.

	Score
Accuracy	0.434
Balanced Accuracy	0.583

Table 13: Random Forest after data reduction through clustering scores.

B.3 Result analysis

The clustering applied for each of the themes, *demographic*, *economics*, *geographic*, *health and sanitation* and *education and science*, reduced the data set to only 4 columns, one for each theme corresponding to the cluster identification, but excluding *education and science* because the average silhouette failed to meet the established criterion (> 0.25). Taking this very reduced data set composed only of categorical columns, as input to our classifier, we were able to reach scores comparable to those achieved in section 5 that used as input data sets reduced via mRMR and PCA. As conclusion, the approach presented in this appendix can be explored further as a viable alternative for the data reduction task. Moreover, this approach makes it possible to estimates the importance of each theme to the HDI progress prediction, using the feature importance derived from the Random Forest R implementation. The importance of each features would be derived from the obtained theme clusters using for instance the PCA loadings also used to visualize the clusters.

C Oversized figures

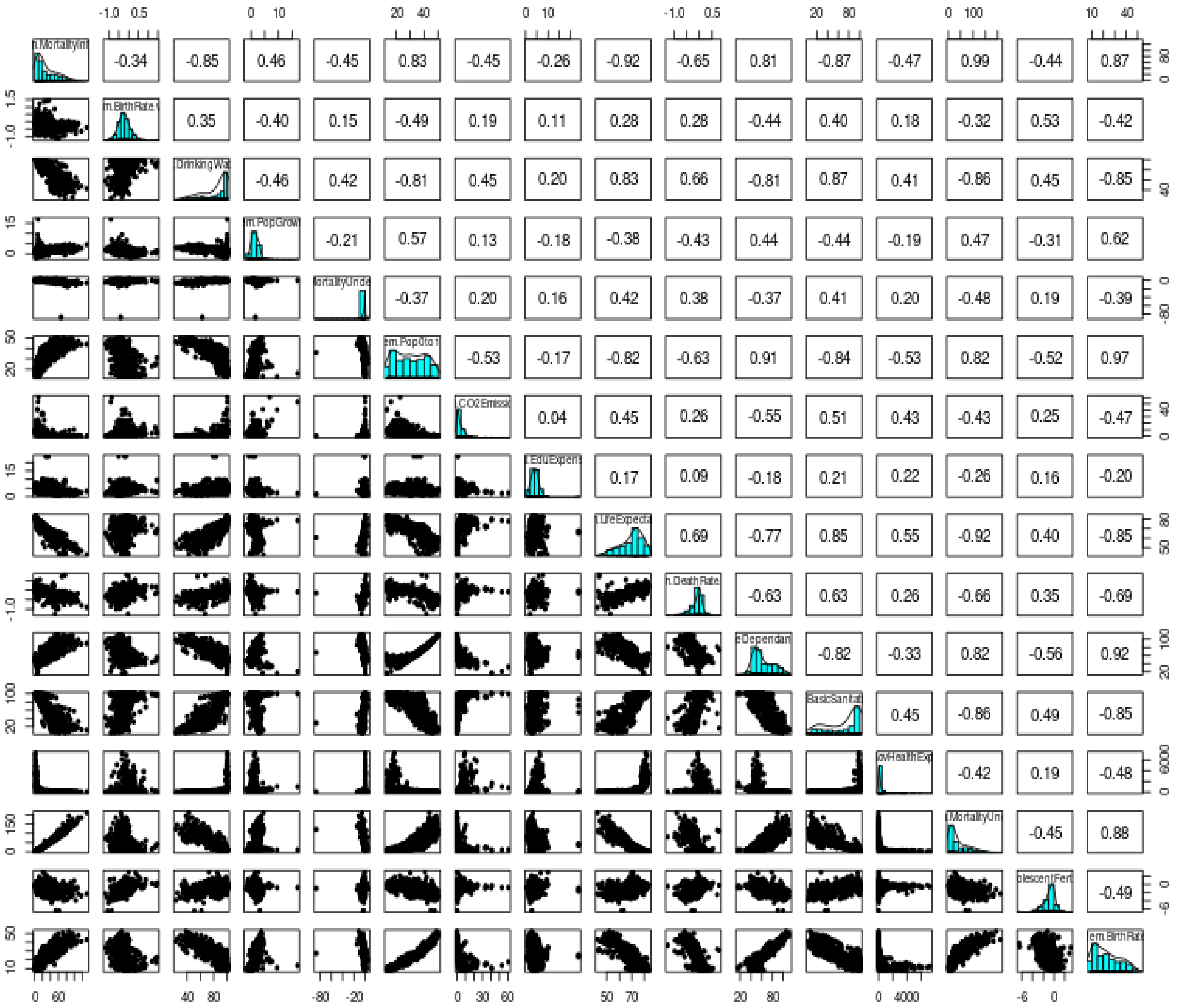


Figure 14: Pairs panel plots of mRMR reduced dataset

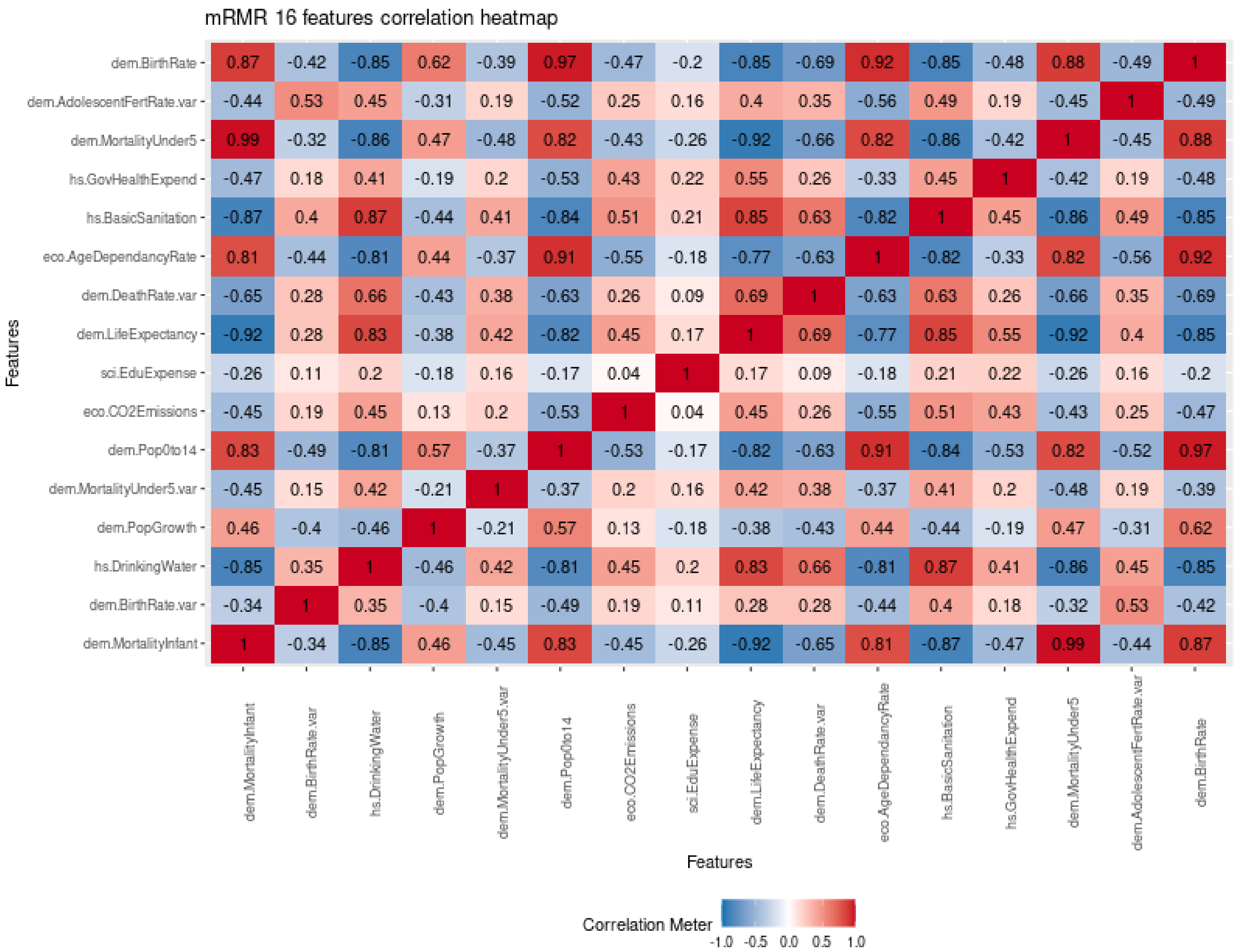


Figure 15: mRMR 16 Correlation Heatmap