A Knowledge Discovery from Data Process to Assess the Reliability of the Human Development Index

Hassenet Slimani

A Knowledge Discovery from Data Process to Assess the Reliability of the Human Development Index

Hassenet Slimani
Higher Institute of Computer Science, University of Jendouba, Tunisia

ABSTRACT

The United Nations Development Program (UNDP) defines a set of indicators about the level of human development in countries, collects data and issues reports annually. This paper targets assessing the ability of the the UNDP Human Development Index (HDI) to reflect a factual image of human development in a country. Assessment is based on a reproducible process of knowledge discovery from data (KDD). Main results include: (i) Disparities among the sample dataset countries are most visible for the distribution of the gross national income per capita (GNI-pC) dimension; (ii) The formula used to calculate the HDI weakens the contribution of the GNI-pC, so some countries are catching up the same HDI of other countries having higher GNI-pC only due to a higher expected indicator (like expected years of schooling). This is not in favor of the HDI which is supposed to reflect factual development level while an expected measure can not reflect facts; and (iii) The clustering found in the data mining step appears factually realistic relatively to the HDI data including the high disparities in GNI_pC.

Keywords: Correlation analysis, principal component analysis, data pre-processing, data clustering, data post-processing, data scaling, United Nations Development Program, life expectancy, expected years of schooling, mean years of schooling, gross national income per capita

INTRODUCTION

"Human development" is an umbrella term that gathers the dreams of wellbeing for individuals, communities and nations. People work on their development in order to achieve the life they value. In an attractive article about "human development", the website "Measure of America" (Measure of America team, 2021) defines human development as "the process of enlarging people's freedoms and opportunities and improving their well-being"; "It is about the real freedom ordinary people have to decide who to be, what to do, and how to live". The same website emphasizes how the concept of "capabilities" is a central concept for the human development approach since capabilities mean what people can do and what they can become; capabilities are the equipment, one has, to pursue that wanted life of value. These capabilities basically include good health, access to knowledge, and a decent material standard of living.
The UNDP calculates, for each country, a "Human Development Index (HDI)" based on these three basic capabilities a.k.a dimensions of the HDI. UNDP reports show that HDI is a summary measure of achievements in three key dimensions of the human development: (i) a long and healthy life evaluated through the index of life expectancy at birth (Life_Expectancy), (ii) access to knowledge evaluated

through the education index based on expected years of schooling (E_Y_S) and mean years of schooling (M_Y_S), and (iii) a decent standard of living evaluated through the gross national income per capita (GNI_pC) index (UNDP, 2018; UNDP, 2019) .

Since it has been adopted by the United Nations, the HDI has been target of large research work that can be two-folded: (i) the studies that just use HDI to study, evaluate or predict development performance for some countries (Boutayeb & Serghini, 2006; Santos et al., 2017; El Katat et al., 2019) and (ii) the other studies that question the relevance of the HDI as a measure that would reflect factual information about the real level of human development in a country (Blanchflower & Oswald, 2005; Bagolin & Comim, 2008; Porter et al., 2014;Verma, 2017; Lepeley, 2017).

In the same perspective of evaluating HDI ability to convey the real level of nations' development, this paper revisits the United Nations HDI through, a different approach, a knowledge discovery from data process where data about HDI and its components, for some countries, are analyzed in several stages. While the followed process and its outputs remain reproducible for any set of countries, this paper uses a sample dataset of human development indicators in the Arab countries for the year 2018.

The knowledge discovery process allows extracting knowledge from data (Han & Kamber, 2006). It is, mainly, composed of three consecutive stages briefly introduced as follows:

1. Data pre-processing where data is prepared for the step of mining through data cleaning, summarization, integration, transformation and reduction.
2. Data mining where specific tasks, like classification, clustering, mining of patterns and associations, prediction, etc., are applied in order to extract data patterns (classes, clusters, patterns and associations, etc.).
3. Data post-processing where the discovered patterns are evaluated to filter out pertinent knowledge that could be visualized.

Throughout the KDD process stages, this paper attempts gaining insight into:(i) the links between HDI values and underlying data; (ii) any interesting details hidden in the fact that a country x has an equal or a higher HDI than a country y?; (iii) how HDI dimensions rank in contributing to the disparities among countries? And how well this ranking is reflected on HDI values?; and (iv) to which extent a clustering algorithm, based on the HDI data, would reflect the facts of living in the dataset countries?.

The contributions are as follows:

1. The use of a KDD process to evaluate HDI reliability; this comes in coherence with the data era, we are witnessing, where knowledge is discovered from data.
2. Confirming that disparities among the sample dataset countries are most visible for the distribution of the GNI-pC dimension. But contribution of the GNI_pC is weakened through the formula used by the UNDP to calculate the HDI .
3. Showing that some countries are catching up the same HDI of other countries having noticeably higher GNI_pC only due to a higher expected indicator (like expected years of schooling). This is not in favor of the HDI which is supposed to reflect the factual level of development while an expected measure can't reflect facts.
4. Showing that a hierarchical clustering of the dataset countries is able to give a factually realistic output that takes into account all HDI data disparities including the high disparities related to the GNI_pC.
5. Confirming the need for improving HDI capture of a country's efforts in human development and the need for assessment methods that focus on quality more than quantitative aspects.

In the two first steps of data pre-processing and data mining, most of the work is done using the programming language of statistical computing and graphics R.
The remainder of the current paper studies related works, reports work in the three steps of the KDD process then ends up by a conclusion and perspectives.

RELATED WORKS

HDI has been tackled by many research works. Some works use it to evaluate status of human development for some countries or to predict it. One work sample focuses on forecasting the human development index and life expectancy for Latin American countries for the period of 2015-2020 using data mining techniques (Santos et al., 2017). Another example studies the position of one country (Lebanon) among middle east countries regarding financial development indicators; the authors use data mining techniques like k-means clustering and k-Nearest Neighbors classification (El Katat et al., 2019). A third example focuses on the relationship between health indicators(Life expectancy at birth, Maternal Mortality, Infant mortality, etc.) and human development in the Arab region. Data analysis using principal components analysis is used to compare the achievements of the Arab countries in terms of the considered health indicators (Boutayeb & Serghini, 2006).
Other research works question the HDI itself as a measure of human development. One work sample questions the HDI assessment that brings Australia as the third in the world (Blanchflower & Oswald, 2005). It includes a review of work on economics of happiness, implications for policymakers and an analysis of new data on approximately 50,000 randomly sampled individuals from 35 nations. The data is about well-being questions that covered the levels of satisfaction with one's life in general, with family life and with main job. The analysis of the gathered data ended up at authentic facts of low performance for Australia in several happiness indicators. The authors emphasized their purpose was not to reject HDI methods, but rather to argue that much remains to be understood in this area of well being assessment among nations.
Another work attempts answering the question " To what extent is the HDI a successful alternative to the GDP?" (Bagolin & Comim, 2008). It acknowledges that HDI represents indeed an advancement over solely-income centered indicators (like GDP) not only in terms of the characterization of the multidimensional nature of development but also in terms of its refined theoretical basis. Its research method consists in an analysis of the HDI's evolution since its creation, looking at the contributions and criticisms put forward and in an investigation of the correlation between high HDI and people's real capabilities and/or opportunities. Its ultimate finding is that despite HDI flexible evolution, the index is still unable to reply to the majority of the criticisms that it has received. For instance, in the HDI formula, education represents a third of the index weight and higher education has the same weight as fundamental education; could higher education be considered a basic capability? Also income, which represents all standard of living aspects, goes through a diminishing returns to scale in the HDI; why the same does not apply to education?
Some recent works focus on alternatives to the UN HDI like the Gross National Happiness Index (GNHI) and the Social Progress Index (SPI). The GNHI is meant to convey more fully the breadth and texture of peoples' lives than the standard welfare measure of GDP per capita, and the HDI (Verma, 2017;Lepeley, 2017). GNHI values collective happiness as the goal of governance. It emphasizes harmony with nature and traditional values as expressed in the nine domains of happiness and four pillars of GNHI. The four pillars of GNHI are:(i) sustainable and equitable socio-economic development;(ii) environmental conservation;(iii) preservation and promotion of culture; and (iv) good governance. The nine domains of GNHI are psychological well-being, health, time use, education, cultural diversity and resilience, good governance, community vitality, ecological diversity and resilience, and living standards. The index is calculated based on indicators representing these pillars and domains.

As for the SPI, it measures the capacity of a society to (i) meet the basic human needs of its citizens, (ii) establish the building blocks that allow citizens and communities to enhance and sustain the quality of their lives, and (iii) create the conditions for all individuals to reach their full potential (Porter et al., 2014).
The work in the current paper continues questioning the relevance of the UNDP HDI in mirroring factual human development level among nations and that based on a KDD process that will be presented in the next section.

KNOWLEDGE DISCOVERY FROM HDI DATA

In this section, the KDD process is run for the human development data of the Arab countries for the year 2018 shown in Table 1. These data are gathered from the UNDP data center at this web link: http://hdr.undp.org/en/data. Work related to the three steps of the KDD process is reported in the next subsections.

Table 1. HD indicators of The Arab countries for the year 2018.

N°	Country Name	Life_Expectancy	M_Y_S	E_Y_S	GNI_pC	Composite HDI
1	Egypt	71.8	7.3	15.3	10744	0.700
2	Algeria	76.7	8	14.7	13639	0.759
3	Iraq	70.5	7.3	11.1	13200	0.689
4	Sudan	65.1	3.7	7.7	3962	0.413
5	Morocco	76.5	5.5	13.1	7480	0.676
6	Saudi Arabia	75	9.7	17	49338	0.857
7	Yemen	66.1	3.2	8.7	1433	0.463
8	Syria	71.8	5.1	8.8	2725	0.549
9	Tunisia	76.5	7.2	15.1	10677	0.739
10	United Arab Emirates	77.8	11	13.6	66912	0.866
11	Jordan	74.4	10.5	11.9	8268	0.723
12	Libya	72.7	7.6	12.8	11685	0.708
13	Palestine	73.9	9.1	12.8	5314	0.690
14	Lebanon	78.9	8.7	11.3	11136	0.730
15	Oman	77.6	9.7	14.7	37039	0.834
16	Kuwait	75.4	7.3	13.8	71164	0.808
17	Mauritania	64.7	4.6	8.5	3746	0.527
18	Qatar	80.1	9.7	12.2	110489	0.848
19	Bahrain	77.2	9.4	15.3	40399	0.838

Data Pre-Processing

The pre-processing applied on the sample dataset consists in its cleaning, summarization, integration, transformation and reduction.

Data cleaning

Data cleaning, in general, attempts to fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data (Han & Kamber, 2006). In the considered sample dataset, too many

missing values have been noticed with three countries (Djibouti, Comoros and Somalia). Not taking into account these three countries in this study has been the option made. So, the study covers nineteen Arab countries out of twenty two.

Descriptive data summarization

Descriptive data summarization helps in studying the general characteristics of the data and identifying the presence of noise or outliers, which is useful for successful data cleaning and data integration. Table 2 shows a descriptive data summary for the considered dataset; A reading through this summary components follows.

Measures of the central tendency like mean, median and midrange allow for determining the center of the data. The most common and effective numeric measure of the "center" of a variable distribution is the (arithmetic) mean. The mean measure is sensitive to noise and outlier data points because a small number of such data can substantially influence the mean value (Han & Kamber, 2006). This is visible in the considered dataset for the variable GNI_pC where Qatar appears as an outlier with a very high value of GNI_pc; the GNI_pC has a mean value which is far from the median value, so, GNI_pC is noticeably skewed. Similar to the median, the midrange (the average of the largest and smallest values for a variable), confirms the skewness of the distribution of this variable.

Table 2. Univariate statistics

	Life_Expectancy	**M_Y_S**	**E_Y_S**	**GNI_pC**	**Composite HDI**
Min	64.70	3.20	7.70	1433.00	0.41
Max	80.10	11.00	17.00	110489.00	0.87
Range	15.40	7.80	9.30	109056.00	0.45
Mean	73.83	7.61	12.55	25228.95	0.71
Median	75.00	7.60	12.80	11136.00	0.72
Midrange	72.40	7.10	12.35	55961.00	0.64
Skew	-0.88	-0.50	-0.42	1.69	-0.87
Variance	19.63	4.99	6.67	849266347.73	0.02
Standard deviation	4.43	2.23	2.58	29142.17	0.13

The skewness coefficient (skew in Table 2) confirms this skewness as the GNI-pc variable has the highest absolute value among all the considered variables. In fact, skewness is a measure of the symmetry in a distribution. A symmetrical dataset, like a normal distribution, has a skewness equal to zero. Skewness coefficient essentially measures the relative size of the two tails and we have the following rule:

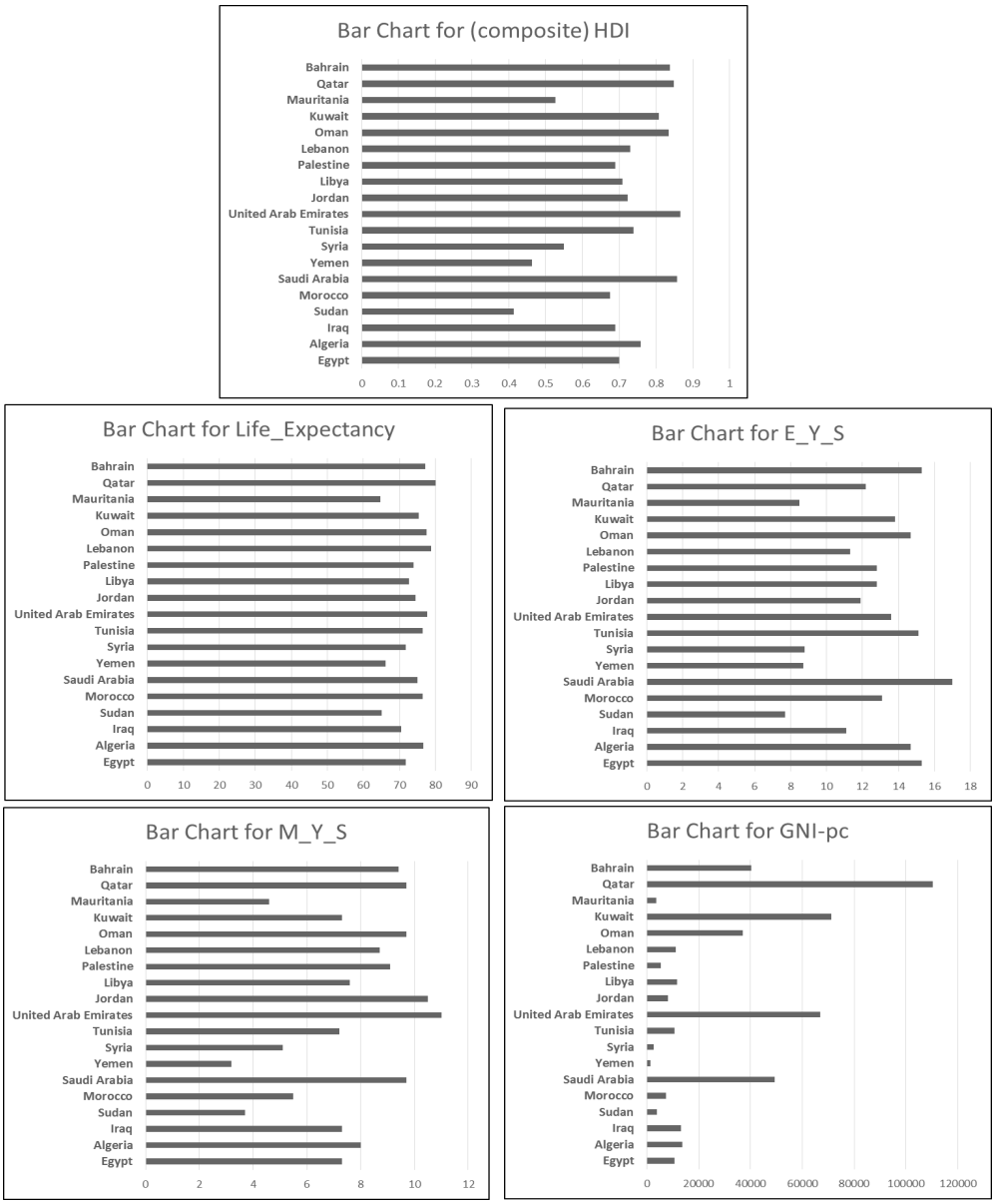
- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical
- If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data are moderately skewed
- If the skewness is less than -1 or greater than 1, the data are highly skewed

So, for the case of the GNI-pC, the skew coefficient is greater than one, the distribution is positively skewed i.e. outliers are located on the higher range of the data values and are pulling the mean in the positive direction.

Measures of data dispersion describe how much data varies (spread out) around a central value. Two distinct data samples may have the same mean or median, but completely different levels of variability, or vice versa. In table 2, measures of data dispersion (the variance and the standard deviation) confirm a high variation of the GNI-pc among the dataset countries. Nevertheless, that high variation of the GNI-pc does not show up at the level of the composite HDI which has a very low standard deviation (near zero value: 0.13). Having this strong variation of the GNI-pC hidden at the level of the HDI is just because of the way the HDI is calculated as it will be explained later.

There are many types of graphs for the display of data distributions and summaries that can be used for the visual inspection of the data. For the running case, bar charts and scatter plots have been chosen; they are illustrated in Figure 1 and Figure 2.

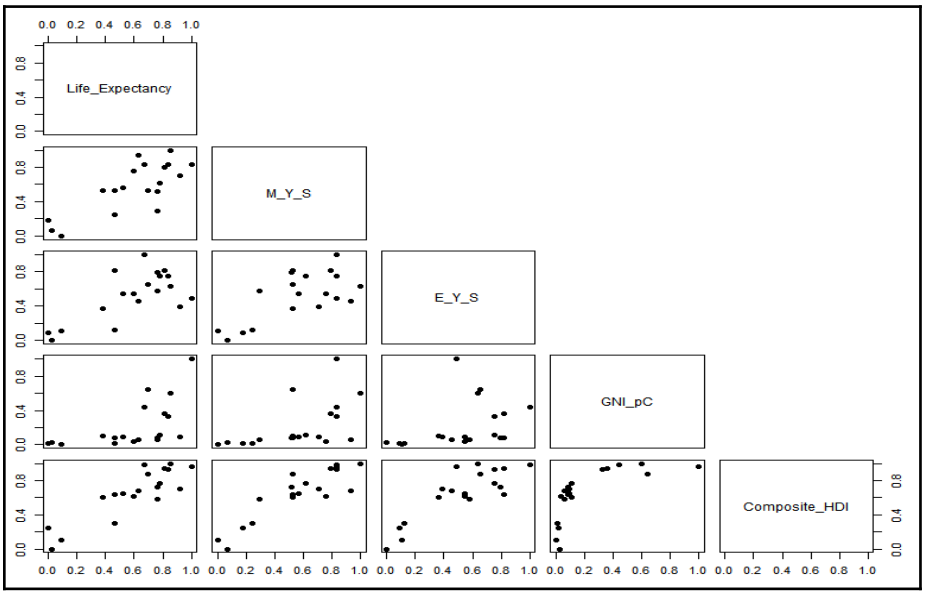
Figure 1. Bar charts of the sample dataset variables



Through the bar charts of Figure 1, it is visible that disparities among the Arab countries are most visible for the GNI-pC distribution. More importantly, even with low or average GNI-pC, some countries are catching up the same HDI as other countries with better GNI-pC. For example, Bahrain and Qatar are having nearly the same HDI when Bahrain's GNI-pC is, at least, a third of Qatar's GNI-pC. Bahrain is achieving such competitiveness only by a better expected indicator (E_Y_S) than Qatar Life_Expectancy. Similarly, Tunisia is having a HDI of 0.739, very close to the HDI of Lebanon (0.73) just by surpassing Lebanon in the E_Y_S by about four expected years of schooling. This is not in favor of the HDI credibility in reflecting the factual level of development for a country since an expected measure can't reflect facts.

A scatter plot is one of the most effective graphical methods for determining if there appears to be a relationship, a pattern, or a trend between two numerical attributes (Han & Kamber, 2006). Given n attributes, a scatter-plot matrix is a symmetric matrix consisting of an n × n scatter plots grid that provides a visualization of each attribute with every other attribute.

Figure 2. Scatter plots of the dataset variables.



In scatter plot matrix of Figure 2, there is no observed correlation between the plotted variables except between the composite HDI and its components. As explained in the technical notes of the human development report for 2018 (and 2019, too), there is almost linear shape of evolution of the HDI as a function of life expectancy, mean years of schooling and expected years of schooling. But, "as each additional dollar of income has a smaller effect on expanding capabilities in the dimension of a decent standard of living , the transformation function from income to capabilities is likely to be concave and the natural logarithm is used for income" ; this is reflected by the concave shape of the evolution of the HDI as a function of the GNI_pC (Kahneman & Deaton, 2010; UNDP, 2018; UNDP, 2019).

Data integration

Data integration, in general, merges data from multiple sources into a coherent data store, such as a data warehouse. The dataset used in this paper is coming from only one source which is the human development reports by the UNDP and data files consist in one excel file. Data integration, also, includes

detecting redundancy in data. Some redundancies can be detected by correlation analysis. Given two attributes, such analysis can measure how strongly one attribute implies the other based on the available data (Han & Kamber, 2006). Correlation is a bivariate analysis that measures the strengths of a linear association between two variables. For numerical attributes, correlation between two attributes, A and B, can be evaluated by computing the correlation coefficient (Pearson's product moment coefficient). Given a set of observations described with a set of attributes/variables, a correlation matrix is a symmetric matrix showing correlation coefficients between pairs of variables. Figure 3 shows the correlation matrix of the considered dataset variables.

Figure 3. Correlation matrix

	Life_Expectancy	M_Y_S	E_Y_S	GNI_pC	Composite_HDI
Life_Expectancy	1.00	0.78	0.69	0.57	0.87
M_Y_S	0.78	1.00	0.68	0.54	0.89
E_Y_S	0.69	0.68	1.00	0.38	0.84
GNI_pC	0.57	0.54	0.38	1.00	0.69
Composite_HDI	0.87	0.89	0.84	0.69	1.00

>

Figure 3 shows a strong correlation between the composite HDI and each one of its three components (M_Y_S, E_Y_S and Life_Expectancy). On the contrary, it shows an average correlation between the composite HDI and the GNI_pC component. This is understandable when taking into account the formulas linking the composite HDI to its components (UNDP,2018; UNDP, 2019). In fact, the composite HDI is calculated as the geometric mean between life expectancy index, education index and gross national income per capita index. Both life expectancy index and education index are linear functions of their actual values since , for a given country, we have:

- $\text{Life expectancy index} = \dfrac{\text{Life expectancy} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$ (1)

- $\text{Education index} = \dfrac{\text{Expected years of schooling index} + \text{Mean years of schooling index}}{2}$ (2)

where:

- $\text{Expected years of schooling index} = \dfrac{\text{Expected years of schooling} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$ (3)

- $\text{Mean years of schooling index} = \dfrac{\text{Mean years of schooling} - \text{minimum value}}{\text{maximum value} - \text{minimum value}}$ (4)

Differently from life expectancy index and education index, to calculate the GNI_pC index, the natural logarithm is used as follows:

- $\text{GNI\_pC index} = \dfrac{\ln(\text{actual GNI\_pC}) - \ln(\text{minimum GNI\_pC})}{\ln(\text{maximum GNI\_pC}) - \ln(\text{minimum GNI\_pC})}$ (5)

where: *ln* is the natural logarithm.

This way limits the contribution of the GNI_pC to the HDI and results in a weak correlation between both variables.

Data Transformation

In the process of KDD, data often needs to undergo certain transformations before the mining process itself. Options for transforming data include smoothing, aggregation, generalization of the data, normalization and attribute construction (Han & Kamber, 2006).

The transformation needed for the current case study is normalization, or more specifically, data scaling which means changing the range of the data values. The range is often set at [0, 1] for numerical values. By such a transformation, the shape of the distribution doesn't change. Furthermore, bringing all variables to the same scale is needed by many algorithms like the principal component analysis which is used as data reduction tool in next subsection of this paper. In fact, data scaling helps in preventing attributes with initially large ranges from out-weighing attributes with initially smaller ranges. Such a case of range variability appears clearly in the considered dataset where GNI_pc has larger values than the remaining indicators.

Methods for data scaling/normalization include min-max normalization and z-score normalization. Min-max normalization performs a linear transformation on the original data. It maps values of an attribute, A, from the range *[minA, maxA]* to corresponding values in a new range *[new_minA, new_maxA]* following this formula:

$$newX = \left[\frac{X - minA}{maxA - minA}\right] * \left[\text{new\_maxA} - \text{new\_minA}\right] + \text{new\_minA},$$

where *minA* and *maxA* are the minimum and maximum values of the attribute A respectively. When, the new range is [0, 1], the formula simply becomes: $newX = \frac{X - minA}{maxA - minA}$

In z-score normalization (or zero-mean normalization), the values for an attribute, A, are normalized based on the mean and standard deviation of A. A value, v, of A is normalized to v' such that $v' = \frac{v - \mu}{\sigma}$ where μ and σ are the mean and standard deviation, respectively, of attribute A. It is clear that the z-score normalization maps the original distribution to a distribution with a zero mean and a unit variance.

Min-max normalization is sensitive to outliers; it may be dominated by outliers like for the attribute GNI_pc in the sample dataset. The z-score normalization is more helpful in this case of an attribute with outliers.

Scatter plots for the dataset variables after min-max normalization and zero-mean normalization are shown by Figure 4 and Figure 5, respectively. When comparing both plots with that of the original data plot in Figure 2, we see that they are similar; this gives a visual confirmation that relationships between variables are kept by min-max normalization and zero-mean normalization.

Figure 4.Scatter plots matrix of the min-max normalized data

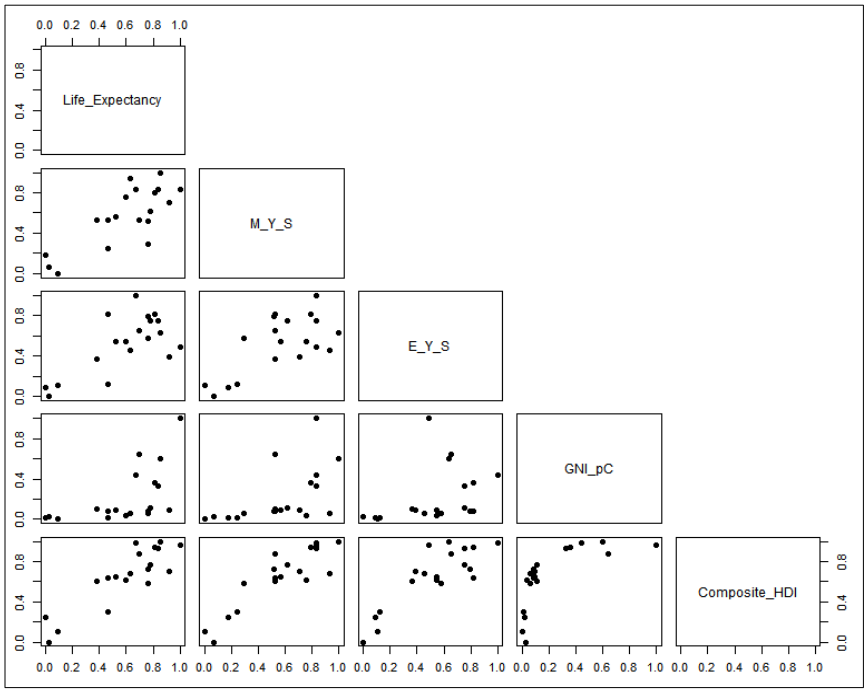
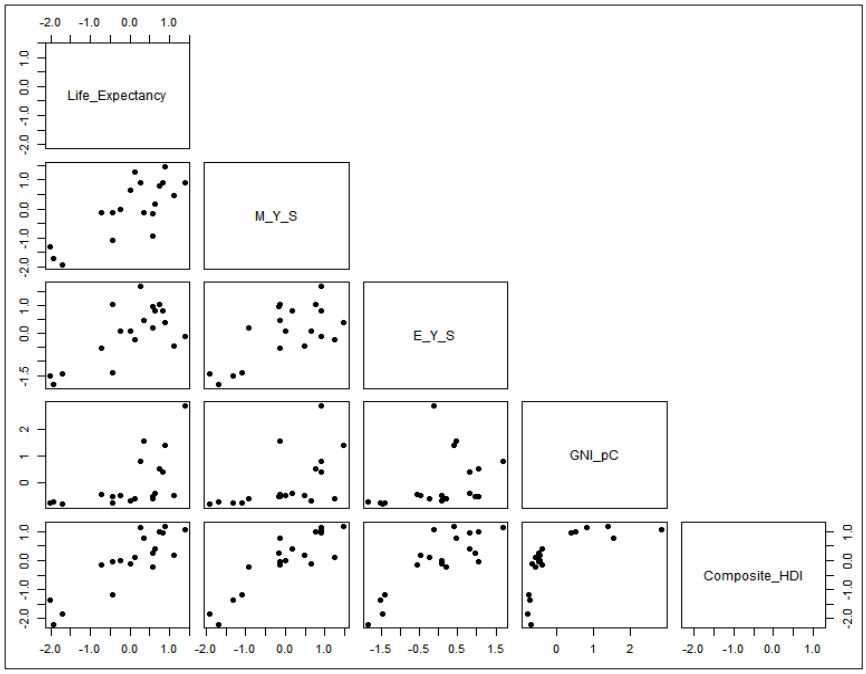


Figure 5. Scatter plots matrix of the z-score -normalized data



Data Reduction

Data reduction techniques allow for the selection of a pertinent subset of the original dataset variables as "a reduced representation of the data set" from which the same (or almost the same) analytical results can be extracted by mining processes. The principal component analysis (PCA) is a data reduction technique that identifies important relationships in a dataset through quantifying the importance of these relationships so the most important relationships can be kept. In the considered case, applying a PCA would help in digging more inside the question "what are the most representative variables among the four dimensions of the HDI". So, PCA is applied on the first four variables only i.e. Life_Expectancy, M_Y_S, E_Y_S and GNI_pC. In what follows, some selected outputs of the PCA are discussed (1-3):

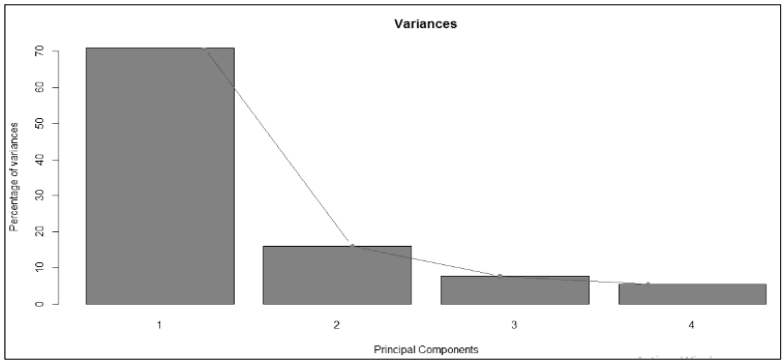
1. Proportions of variance retained by principal components (Figure 6 and Figure 7).

Figure 6. Proportions of variance retained by the four principal components

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	2.84	70.89	70.89
comp 2	0.64	16.04	86.93
comp 3	0.30	7.62	94.55
comp 4	0.22	5.45	100.00

Eigenvalues correspond to the variance explained by each principal component. The first eigenvalue is about 4.5 times the second value whose value is about 0.64 << 1. It is reported in literature related to the PCA that a principal component with an eigenvalue > 1 indicates that the principal component accounts for more variance than accounted by one of the original variables in standardized data. This is commonly used as a cutoff point to determine the number of principal components to retain. Figures 6 and 7 show that most of the variance (about 70.9%) is carried by the first principal component which is the only one component whose eigenvalue value is greater than 1. About 87% of the variance contained in the data is retained by the first two principal components (Jolliffe, 2002).

Figure 7. Graph of the variance percentages associated with the principal components(Scree plot)



2. Contributions of the variables to the principal components (Figure 8)

Figure 8. Percentages of variables' contributions to the principal components

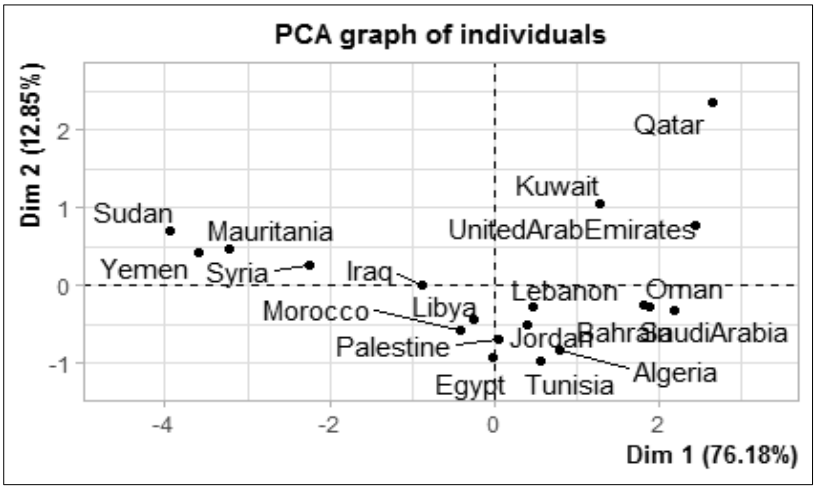
	Dim.1	Dim.2	Dim.3	Dim.4
Life_Expectancy	29.40	0.60	11.51	58.49
M_Y_S	28.69	1.20	29.97	40.15
E_Y_S	23.81	27.69	47.64	0.86
GNI_pC	18.10	70.52	10.88	0.50

The contributions of the four HDI components to the variance explained by the first principal component (Dim.1 in Figure 8) are relatively close ranging from about 18% to 29%; so almost all the four dimensions are closely contributing to the variance explained by the first principal component. Contributions to the variance explained by the second principal component (Dim.2) are mainly carried by the GNI_pC. So, in addition to its considerable contribution to the variance explained by the first principal component, GNI_pC is carrying most of the variance explained by the second principal component. Therefore, GNI_pC is a strong parameter of variance among the Arab countries. However, GNI_pC participation in the HDI is weakened through the formula currently used by the UNDP.

3. Graph of individuals in principal components coordinate system

Figure 9 shows the coordinates of individuals relatively to the two first principal components. In accordance with the definition of the PCA, Figure 9 shows an important variance of the individuals (the set of countries) through the first principal component (Dim1) and slightly less variance through the second principal component (Dim2). Also, clusters of the countries already show up. Thus, we move to the next step of our KDD process: Data mining through cluster analysis.

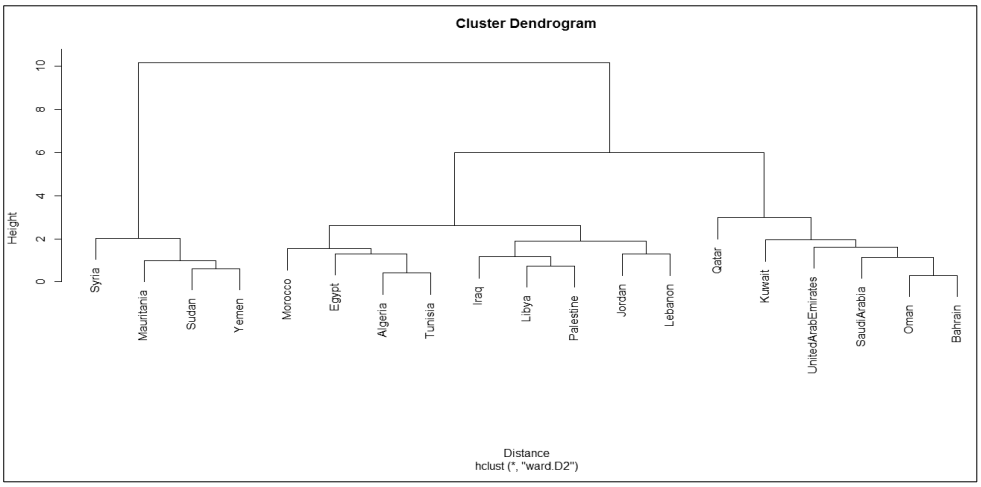
Figure 9. Graph of individuals in principal components coordinate system.



Data mining through cluster analysis

At this level of the KDD process, data mining step is started up; a cluster analysis task is chosen. An agglomerative hierarchical clustering method (Han & Kamber, 2006) is applied taking into account all considered variables .i.e both the composite HDI and its four components (Life_Expectancy, M_Y_S, E_Y_S and GNI_pC). The output of the clustering method is a hierarchical tree represented by the dendrogram of Figure 10. In the dendogram, each leaf corresponds to one country and, as we move up the tree, countries that are similar to each other are combined into branches, which are themselves fused at a higher height. The height of the fusion, provided on the vertical axis, indicates the dissimilarity (the distance) between two objects/clusters. The higher the height of the fusion, the less similar the objects are.

Figure 10. Dendogram of an agglomerative hierarchical clustering of the dataset individuals



The key to interpreting a dendrogram is to focus on the height at which any two objects are joined together. The dendrogram shows that the biggest difference between clusters is between the cluster of *{Syria, Mauritania, Sudan and Yemen}* versus that of the remaining Arab countries. Then, comes the difference between the cluster of *{Morocco, Egypt, Algeria, Tunisia, Iraq, Libya, Palestine, Jordan, Lebanon}* versus the cluster of *{Qatar, Kuwait, United Arab Emirates, Saudi Arabia, Oman, Bahrain}*. An attempt, to find out what brings the items of a cluster together, here follows:

- The countries of the cluster *{Syria, Mauritania, Sudan and Yemen}*, witness wars, economic bans or deep economic difficulties. In this cluster, the HDI is strictly inferior to 0.6 and the GNI_pC ranges between 1433$ and 3962$
- The cluster of *{Morocco, Egypt, Algeria, Tunisia, Iraq, Libya, Palestine, Jordan, Lebanon}* brings together countries that have HDI greater than 0.6 and strictly inferior to 0.8. These countries have a GNI_pC that ranges between 5314$ and 13639$.
- The cluster of *{Oman, Bahrain, SaudiArabia, UnitedArabEmirates, Kuwait, Qatar}* brings together the countries that have a HDI greater than 0.8 and a GNI_pC ranging between 37039 and 110489; These are the (Arabian) Gulf countries missing Iraq which has known wars and instability since years ago.

Overall this clustering is highly in accordance with the original data reflecting differences in all the considered variables including the highly skewed variable GNI_pC. As previously noticed in this paper, when looking to single values of the HDI, two countries may have similar HDI only due to an expected

indicator (life expectancy or expected years of schooling) like the two pairs previously mentioned {Tunisia, Lebanon} and {Bahrain, Qatar}. This ambiguity is cleared with the hierarchical clustering since in the dendogram of Figure 10, Tunisia is in a slightly lower sub-cluster than Lebanon, and Bahrain is in a noticeably lower sub-cluster than Qatar. Definitely, data mining through cluster analysis gives knowledge that reflects the factual level of human-development in a country better than the summary values of the HDI.

DATA POST-PROCESSING (DISCUSSION OF RESULTS)

In this post-processing step, the goal is to gather the knowledge discovered throughout the steps of the KDD process, to evaluate this knowledge and interpret it in order to answer the questions raised at the beginning of this research work. Combining findings made throughout the two first steps of the KDD process (.i.e pre-processing and mining), main conclusions could include:

- Descriptive data summarization showed that disparities among the sample dataset countries are most visible for the GNI-pC distribution. Also, in response to the question "what variables participate most in the variance of the considered countries?", the PCA analysis showed that all the four HDI dimensions are closely contributing to the variance explained by the first principal component. But, contributions to the variance explained by the second principal component are mainly carried by the GNI_pC.
- Despite being a strong parameter of variance among the sample dataset countries, GNI_pC sees its contribution to the HDI weakened through the formula currently used by the UNDP.
- Some countries are catching up the same HDI as other countries with better GNI-pC just due to a higher expected indicator. This is not in favor of the HDI credibility in reflecting the factual level of development for a country as an expected measure (like E_Y_S) can't reflect facts about the level of human development in a country.
- The step of data mining using an agglomerative hierarchical clustering resulted in clusters that faithfully reflect all the variables including the high variation of the GNI_pC. Also, the clustering could separate pairs of countries that have very close values of the HDI into separate clusters taking into account the real distributions of the variables without any interference like weakening the contribution of some variables as done with the HDI. The clustering could show that the cluster of countries that have the highest HDI are those that have the highest GNI_pC compared to other countries. Also, by reflecting the underlying data without interference, the clustering could reflect the facts of living within the Arab region countries. In fact, it is known that the GCC countries have higher and better standards of living and that people, in many other countries with competitive HDI struggle in earning their living.
- In addition to the choice made of weakening the contribution of the GNI_pC to the HDI, the HDI is a summary for four components which are calculated based on mean values and, very often, they hide inequalities among regions of the same country. Also, a question may arise, here, about these components: Are life expectancy, mean years of schooling, expected years of schooling and gross national income per capita enough to evaluate the development of a society ? Certainly, in essence, quality of life and quality of schooling do matter more than numbers of years as many people live a long but uncomfortable life and many people spend long years in schools without truly learning or earning the expected level of knowledge and skills. Also, a successful investment in an acceptable amount of the income to lead a life of value is a strong asset for personal and nations development.

CONCLUSION

Throughout this paper, the UNDP HDI is revisited through a KDD process questioning the relevance of the HDI as a measure that would reflect factual information about the real level of human development in a country. Human development data of the Arab countries for the year 2018 is used as a sample dataset. The adopted KDD process shows that, overall, disparities among the dataset countries are most visible for the GNI-pC distribution, then for the three other dimensions of the HDI (M_Y_S, E_Y_S, Life_Expectancy) and for the HDI itself. PCA analysis shows GNI_pC as a strong parameter of variance among the dataset countries. However, GNI_pC participation in the HDI is weakened through the formula currently used by the UNDP. Also, even with low or average GNI-pC, some countries are catching up the same HDI as other countries with better GNI-pC just due to a better expected indicator solely as observed in some cases. Obviously, this limits the ability of the HDI in reflecting facts of the human development level. Correlation analysis shows average correlations between the GNI_pC and the remaining components of the HDI. This remains understandable as people with high GNI_pC may not be interested in long schooling or may not live for long. Also, a person may study for long and get high degrees without getting an adequate job or earning a good living.

The step of data mining using an agglomerative hierarchical clustering resulted in a factually realistic clustering of the dataset countries; a clustering that faithfully reflects all variables including the high variation of the GNI_pC. In fact, the clustering could separate pairs of countries that have very close values of the HDI into separate clusters taking into account the real distributions of the variables without any interference like weakening the contribution of some variables as done with the HDI.

Overall, the KDD process, as run in this paper, shows data as a strong asset when valued through reliable knowledge extraction process. The KDD process allowed composing an in-depth view of the human development data and allowed for the extraction of reliable knowledge from the data without any alteration. The KDD process is highly informative than the variable HDI taken solely. The HDI definitely appears as a quantitative summary for some aspects of human development and it needs to be improved for better capture of factual level of human development. Also, there is need for quality-focused assessment tools to assess quality of life, quality of schooling and quality of management of a country financial resources allowing a satisfying level of the GNI_pC; this would be more insightful in guiding nations on the way of their developments.

REFERENCES

Bagolin, I.P., & Comim, F. (2008). Human Development Index (HDI) and its family of indexes: an evolving critical review. Revista de Economia, 34(2), 7-28.

Blanchflower, D.G., & Oswald, A.J. (2005). Happiness And The Human Development Index: The Paradox Of Australia. Australian Economic Review, 38(3), 307-318.

Boutayeb, A., & Serghini, M. (2006). Health indicators and human development in the Arab region. International Journal of Health Geographics. https://doi.org/10.1186/1476-072X-5-61 .

El Katat, S., Kalakech, A., Kalakech, M., & Denis, H. (2019). Financial Development Indicators: A Comparative Study between Lebanon and Middle East Countries Based on Data Mining Techniques. The International Arab Journal of Information Technology (IAJIT), 16(special issue),499-505.

Han, J., & Kamber M. (2006). Data Mining: Concepts and Techniques (2nd ed.). Elsevier.

Jolliffe, I.T. (2002). Principal Component Analysis (Second Edition). Springer.

Kahneman, D., & Deaton, A. (2010). High income improves evaluation of life but not emotional well-being. Proceedings of the National Academy of Sciences, 107(38), 16489-16493.

Lepeley, M.-T. (2017). Bhutan's Gross National Happiness: An Approach to Human Centred Sustainable Development. South Asian Journal of Human Resources Management, 4(2), 174–184.

Measure of America team.(2021). Measure of America website. https://measureofamerica.org/human-development/ (accessed January 2021)

Porter, M. E., Stern, S., & Green, M. (2014). Social progress index 2014:Methodological Report. Social Progress Imperative. https://www.socialprogress.org/static/82a7f907b051ef1cd040c580ac5f497d/2014-social-progress-index-methodology.pdf (accessed August 2021)

Santos, C.B., Pedroso, B., Guimaraes, A.M., Carvalho, D.R., & Pilatti, L.A. (2017). Forecasting of Human Development Index of Latin American Countries Through Data Mining Techniques. IEEE Latin America Transactions, 15, 1747-1753.

UNDP. (2018). Human Development Report for 2018 Technical Notes. http://hdr.undp.org/sites/default/files/hdr2018_technical_notes.pdf

UNDP. (2019). Human Development Report for 2019 Technical Notes. http://hdr.undp.org/sites/default/files/hdr2019_technical_notes.pdf

Verma R. (2017). Gross National Happiness: meaning, measure and degrowth in a living development alternative. Journal of Political Ecology, 24(1). p.476-490.