A Comparative Analysis of Countries' Performance According to SDG Indicators based on Machine Learning

Guilherme Souza[1], **Julian Santos**[1], **Gabriel SantClair**[1], **Janaina Gomide**[1], **Luan Santos**[1]

[1]Polytechnic Institute – Federal Univeristy of Rio de Janeiro (UFRJ)
Postal Code 15.064 – 27.930-560 – Macaé – RJ – Brazil

{gui.eng.santos, gsantclair, janainagomide}@gmail.com,
julianmartins.engenharia@outlook.com, luan.santos@pep.ufrj.br

Abstract. *The Sustainable Development Goals (SDGs) are part of a global effort to reduce the impacts of climate change, promoting social justice and economic growth. The United Nations provides a database with hundreds of indicators to track the SDGs since 2016 for a total of 302 regions. This work aims to assess which countries are in a similar situation regarding sustainable development. Principal Component Analysis was used to reduce the dimension of the dataset and k-means algorithm was used to cluster countries according to their SDGs indicators. For the years of 2016, 2017 and 2018 were obtained 11, 13 and 11 groups, respectively. This paper also analyses clusters changes throughout the years.*

1. Introduction

The Sustainable Development Goals (SDGs) came into effect on January 1^{st} of 2016 as part of Agenda 2030. There are 17 Goals, they contemplate the environmental, social, economic and institutional area - which concerns how to put these objectives into practice. SDGs replaced the 8 Millennium Development Goals (MDGs) which were supposed to be achieved by 2015. The foundations for the SDGs were built at the Rio+20 conference, that was held in Rio de Janeiro, and were adopted by the United Nations (UN) Summit on Sustainable Development in 2015 [Santos and Santos 2017]. Most of United Nations member states, 193 countries, signed the treaty committing to achieve the targets proposed, as the UN Secretary-General Mr. Ban Ki-moon[in China 2015] underscored:

> "[Agenda 2030] It is a roadmap to ending global poverty, building a life of dignity for all and leaving no one behind. It is also a clarion call to work in partnership and intensify efforts to share prosperity, empower people's livelihoods, ensure peace and heal our planet for the benefit of this and future generations [...]."

Each SDG has a number of targets measured by indicators in order to quantify performance of the related parties and serving as a parameter to meet the established goals. According to UK Parliament [Booth 2010], targets are clear expressions of public policies. They describe what must be done by those responsible and let citizens know the priorities of governments. Besides, [Davis et al. 2012] characterizes indicators as "distance governance technologies that help in the more efficient allocation of resources".

There are a total 302 regions in the UN database with 231 indicators - considering all 17 SDGs - that track the Agenda 2030 and provide an open database available to the international community. There are hundred thousands instances considering all the available data since 2016. Therefore, it is challenging to analyze all information in order to monitor and compare countries' performance in relation to the Agenda 2030.

On the other hand, according to [Bishop 2006], Machine Learning (ML) aims to develop methods and algorithms to learn from data and forecast based on data. Those methods have been successfully used to describe the behavior of large datasets. Some works have analyzed SDG indicators using ML such as in [Ferreira et al. 2020] where the authors present a survey of some of these papers. Others authors, as [Kijewska and Bluszcz 2016] and [Jabbari et al. 2019], used the k-mean algorithm to assess the emission levels of Greenhouse Gases and to classify countries as "Developed" or "Developing", respectively.

The current economic scenario may undermine the path to the SDGs, thus structuring all available information can be of fundamental importance to determine the most effective strategy in order to accomplish the Agenda 2030. This work aims to assess which countries are in a similar situation regarding sustainable development. In order to achieve this goal, an unsupervised learning algorithm was applied to group countries according to their SDGs indicators. After that the results are evaluated analyzing the clusters in a spatial and temporal perspectives. All goals were considered to study the performance of countries through the years of 2016, 2017 and 2018. The results of the clustering algorithm were considered to visualize similarities and differences on the clusters through the years.

The following sections present the related works in Section 2; Section 3 outlines the methodology applied, describing and developing the algorithm and the database used; Section 4 presents results' analysis and discussions and Section 5 contains the conclusions and future works.

2. Related Works

Several works have already analysed SDGs considering different approaches and focusing on different objects of study. Some studies considered only one country, such as [Shahbaz et al. 2021], who focused their analysis on India. Their work used nonlinear auto regressive distributed lag approach to assess whether per capita income, energy use, commercial openness and oil price influenced in CO_2 emissions. Similarly, [Raszkowski and Bartniczak 2019] concentrated their study in Poland. The purpose of the work was to determine the status of implementation of SDGs in the country. The authors used the data for the years 2010 to 2016 and dynamic analysis methods to estimate the magnitude of indicators changes through the respective time period [Raszkowski and Bartniczak 2019].

The work presented in [Lim et al. 2016], on the other hand, assessed the performance of 188 countries considering 33 health-related indicators. They rescaled the indicators - considering as zero the worst value and as 100 the best value in the years assessed - and also used statistical methods to systematically compile data. Spline regression was applied to assess the relationship between socio-demographic indicators and health-related SDG indicators.

The authors in [Ferreira et al. 2020] made an in-depth review of ML algorithms already applied for monitoring SDGs performance of countries, presenting the contributions of articles that applied different types of learning algorithms for satellite monitoring. In [Kijewska and Bluszcz 2016], the authors used the K-Means algorithm to study the emission levels of Greenhouse Gases (GHG) from European Union countries. The study considered 28 countries and their attributes were considered as the emission levels of the following gases: carbon dioxide (CO_2), methane (CH_4), nitrogen oxides (NO_x) and nitrous oxide (N_2O). Information from *Eurostat*[1] served as a database. [Jabbari et al. 2019] proposed the use of ML algorithms to group countries in the categories: "Developed" and "Developing". For this, SDGs indicators were used after undergoing a pretreatment and a dimensional reduction. The clustering method adopted was K-Means, with the coefficient *Average Silhouette* used to find the optimal number of groups.

The work proposed in this article differs from the others aforementioned since it considers all SDGs, 251 countries, through the years of 2016, 2017 and 2018. The Principal Component Analysis (PCA) method was used to reduce the dimension of the dataset before using the K-Means method to group countries according to their performance in the years 2016, 2017 and 2018.

3. Methodology

The proposed methodology are illustrated in Figure 1, the following subsections detail each step.

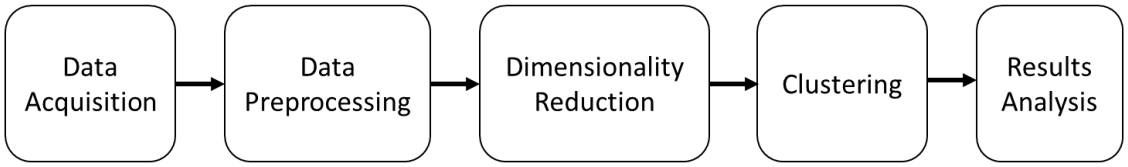

Figure 1. Steps of the proposed methodology.

3.1. Data Acquisition

The dataset used in this article was provided by the UN[2]. The raw data is composed of 302 regions that do not necessarily correspond to a country (e.g.: America, Middle East, Europe, Western Africa, among others). Each instance of the database corresponds to an indicator from a respective region - considering a specific year. There are over 290k lines in the raw dataset - considering 302 regions, 3 years and 231 indicators. Data is structured in 47 columns representing specifications of each measurement (e.g.: country, year, region, information source, etc.). Among all features from the raw dataset - referring to different aspects of each instance - only one refers to the measurement of the corresponding indicator. The SDGs are as follows:

- **SGD 1:** No Poverty
- **SGD 2:** Zero Hunger
- **SGD 3:** Good Health and Well-being

[1]Eurostat: `https://ec.europa.eu/eurostat`
[2]UN: `https://unstats.un.org/sdgs/indicators/database`

- **SGD 4:** Quality Education
- **SDG 5:** Gender Equality
- **SDG 6:** Clean Water and Sanitation
- **SDG 7:** Affordable and Clean Energy
- **SDG 8:** Decent Work and Economic Growth
- **SDG 9:** Industry, Innovation and Infrastructure
- **SGD 10:** Reduced Inequalities
- **SDG 11:** Sustainable Cities and Communities
- **SDG 12:** Responsible Consumption and Production
- **SDG 13:** Climate Action
- **SDG 14:** Life Bellow Water
- **SDG 15:** Life on Land
- **SDG 16:** Peace, Justice and Strong Institutions
- **SGD 17:** Partnership for the Goals

3.2. Data Preprocessing

The data preprocessing is used to prepare the dataset for ML and the steps are shown in Figure 2.

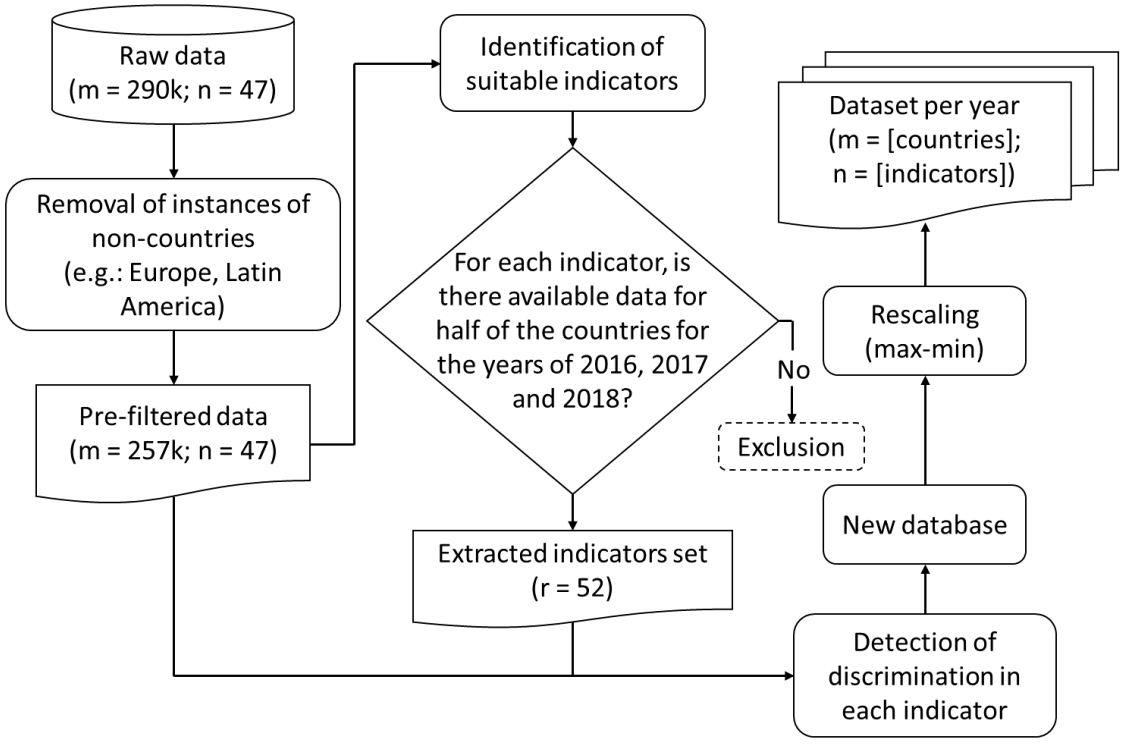

Figure 2. Data preprocessing, in which *n* and *m* represents the number of attributes and instances, respectively.

Instances that did not refer to countries in the 302 regions were disregarded. In addition, indicators that did not have values for at least half of the 250 countries present in the database were removed - regarding the years 2016, 2017 and 2018 - considering that when more than 50% of the data is missing results should only be used for hypothesis

generating [Madley-Dowd et al. 2019]. From this process, 52 indicators remained, out of the 231 that were made available. For the purpose of detecting the relevant indicators, a list was created to serve as a filter to identify the indicators to be considered.

Indicators had different degrees of detail. Some had information by age, location (rural or urban), type of product, etc. In order to benefit from this information, each variation of the indicators had been treated as an independent indicator, after the filters aforementioned. For example, the 2.1.2 indicator was divided into 12 indicators (separation by gender, age, severe or moderate food insecurity, among others). After the data preparation process, the missing information was filled with the average value of the respective indicator.

The features were adjusted using *max-min* scale so that every attribute varied in the same range, since some algorithms converge faster when scaled [Géron 2019]. In the *max-min* scale values vary between 0 and 1, and to transform the dataset the following equation is applied:

$$x_{max-min} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

in which x_{min} is the smallest value for the respective attribute and x_{max} is the largest value. At the end of this process there were 250 countries (rows) and 1146 indicators (columns) in the dataset - when considering ramifications from the previous 52, after the aforementioned filters.

3.3. Dimensionality Reduction

After the previous mentioned data treatment, the dataset presented over 1000 features for only 250 instances - see Table 1. According to [Géron 2019], in high dimensional datasets instances tends to be far apart from each other and, therefore, the results of some ML algorithms are much less reliable.

Table 1. Instances and features in each step of the methodology.

	Instances	Features
Raw Data	290.000	47
Pre-filtered data	257.000	47
Before PCA	250	1146
After PCA	250	4

In order to achieve better results using a reasonable amount of data, while facilitating the later interpretation of the outcome, the Principal Component Analysis (PCA) algorithm was applied. Its main objective is reducing the number of characteristics of the dataset while preserving the maximum amount of variance, according to a predefined criteria [Bishop 2006]. After the data treatment all indicators were considered, when applying the PCA optimization.

3.4. Clustering

In order to achieve the groups of similar countries per year the k-Means clustering algorithm was applied. K-Means is one of the most famous clustering algorithms, which is based on the concept of centroids. In this algorithm, the number of groups (k) is predetermined, and the results depends directly of this parameter [Bishop 2006]. The number of clusters must be large enough to represent the specific characteristics of the dataset, while small enough not to mischaracterize the cluster and avoid overfitting [Kijewska and Bluszcz 2016]. However, it should be noted that there is no single criterion for estimating the value of k.

This article adopted the *Silhouette coefficient* proposed by [Rousseeuw 1987] as a metric capable of analyzing the performance of clustering tasks. The advantages of this method are: *(i)* the possibility of building a graphical visualizations of multidimensional datasets; and *(ii)* the fact that this metric does not depend on the clustering algorithm or external data. Thus, it is possible to compare ML algorithms, with different parameters, and choose which best represents the data. Defined as $(b-a)/\max(a,b)$ with a representing the average distance between each example (instance) and the centroid of its cluster and b representing the average distance between the centroid of a cluster and each example from another group - the closest. The *Silhouette coefficient* varies between -1 and 1. Having that -1 means the clusters are poorly defined - close to each other - and 1 means the clusters are well defined - well delineated.

Although the *Silhouette Coefficient* was used as a quantitative comparison parameter for the results, as can be observed in Figure 3, the greater the number of components considered from the PCA the lesser *Silhouette Coefficient* would be obtained.

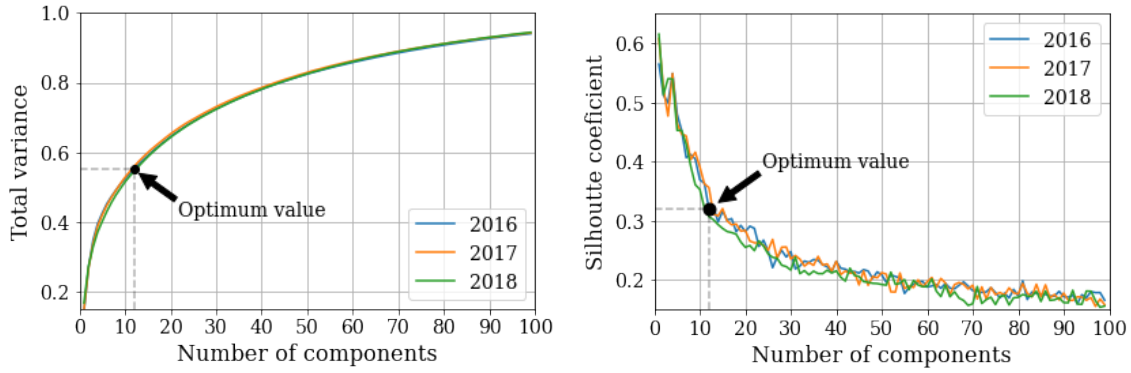

Figure 3. (a) Variance learning curve, (b) Annual clusters' distribution.

In order to overcome this barrier, the cluster's optimal number of groups (k) as well as the optimal number of components (t), to be considered from the PCA, were obtained through a proposed loss function, see Equation 2. The number of components varied from 1 to 100, being selected the one with the smaller associated loss. The objective of this method is to find the best set of hyperparamenters that maximizes the variance ratio (μ) and the cluster's *Silhouette Coefficient* (S) simultaneously.

$$L = \left(\frac{\mu(k,t) - \mu_{max}}{\mu_{max}} \right)^2 + \left(\frac{S(k,t) - S_{max}}{S_{max}} \right)^2 \quad (2)$$

On Table 2 we can observe the number of clusters (k), the number components - considered in the PCA algorithm - and the variance ratio (μ) associated with the respective amount of components obtained for the years of 2016, 2017 and 2018.

Table 2. Hyperparameters, variance and Silhouette Coefficient - per year.

Year	k	N. of Components	μ	Silhouette Coefficient
2016	11	4	0.3943	0.5486
2017	13	4	0.3875	0.5487
2018	11	4	0.3707	0.5394

Once the training parameters were selected, the results were compiled and, for the present work, we focused on the analysis of which indicators had the biggest impact on the components of the PCA - and therefore would better summarize the groups - and on the general behavior of the groups throughout the years.

4. Results and Discussions

In this section, the results of countries clustering, in relation to their SDGs indicators and throughout the years, are presented.

4.1. Indicators Considered

As mentioned before, the PCA algorithm was applied considering all available data for the years of 2016, 2017 and 2018 - after treating the database. There were 1146 attributes in the final database before applying PCA, considering the subdivisions of each indicator (e.g.: per sex, per age, per region, etc.). In order to interpret the results the first four components of the PCA had to be translated into indicators of the SDGs. Figure 4 displays the accumulated percentage of influence of each SDGs in relation to PCA's components.

As can be observed, the components remained representing almost the same goals throughout the years, the most significant change occurred on the second and fourth components on the last year. It matters to say that:

- The first and fourth components represented, almost entirely, the SDGs 10 and 16 - **Reduced Inequalities** and **Peace, Justice and Strong Institutions**, respectively;
- The second and third components represented, almost entirely, SDGs 3 and 4 - **Good Health and Well-being** and **Quality Education**, respectively;
- SGDs 2, 9, 10, 15 and 17 were also significantly considered on some components of the PCA.

Table 3 presents the main indicators considered in each SDG. Only indicators that contributed with more than 10% of the respective PCA component are being reported. These indicators are described as follows:

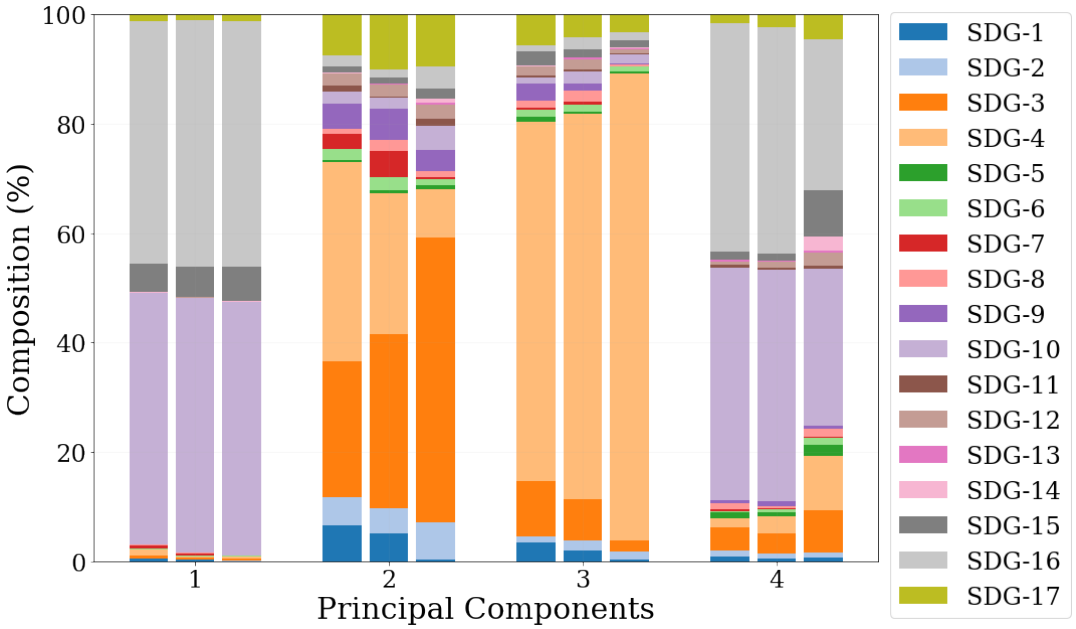

Figure 4. Considered SDGs for the respective component of PCA

Table 3. Indicators and their influence - per PCA component.

PCA Component	Indicator	Influence [%]
1	10.6.1	44, 76
	16.8.1	44, 76
2	3.d.1	25, 37
	3.2.1	16, 63
3	4.1.2	80, 80
4	10.6.1	23, 65
	16.8.1	23, 65

- **SGD 3: Good Health and Well-being**
  - **Indicator 3.2.1:** Under-5 mortality rate.
  - **Indicator 3.d.1:** International Health Regulations (IHR) capacity and health emergency preparedness.
- **SGD 4: Quality Education**
  - **Indicator 4.1.2:** Completion rate (primary education, lower secondary education, upper secondary education).
- **SGD 10: Reduced Inequalities**
  - **Indicator 10.6.1:** Proportion of members and voting rights of developing countries in international organizations;
- **SGD 16: Peace, Justice and Strong Institutions**
  - **Indicator 16.8.1:** Proportion of members and voting rights of developing countries in international organizations.

As there was a considerable amount of missing information in UN data base, these were the most relevant indicators - to distinguish instances and acknowledge patterns - considered in the clustering algorithm. Therefore, these indicators are assumed to summarize countries SDG status, in relation to each other.

4.2. Clusters Analysis

In Figure 5 it is possible to observe the groups obtained after k-Means. Each point represents a country and the axis are the first two components of the PCA. For the years of 2016 and 2018 the number of clusters that minimized the Equation 2 where 11, meanwhile for the year of 2017 the number of clusters obtained was 13.

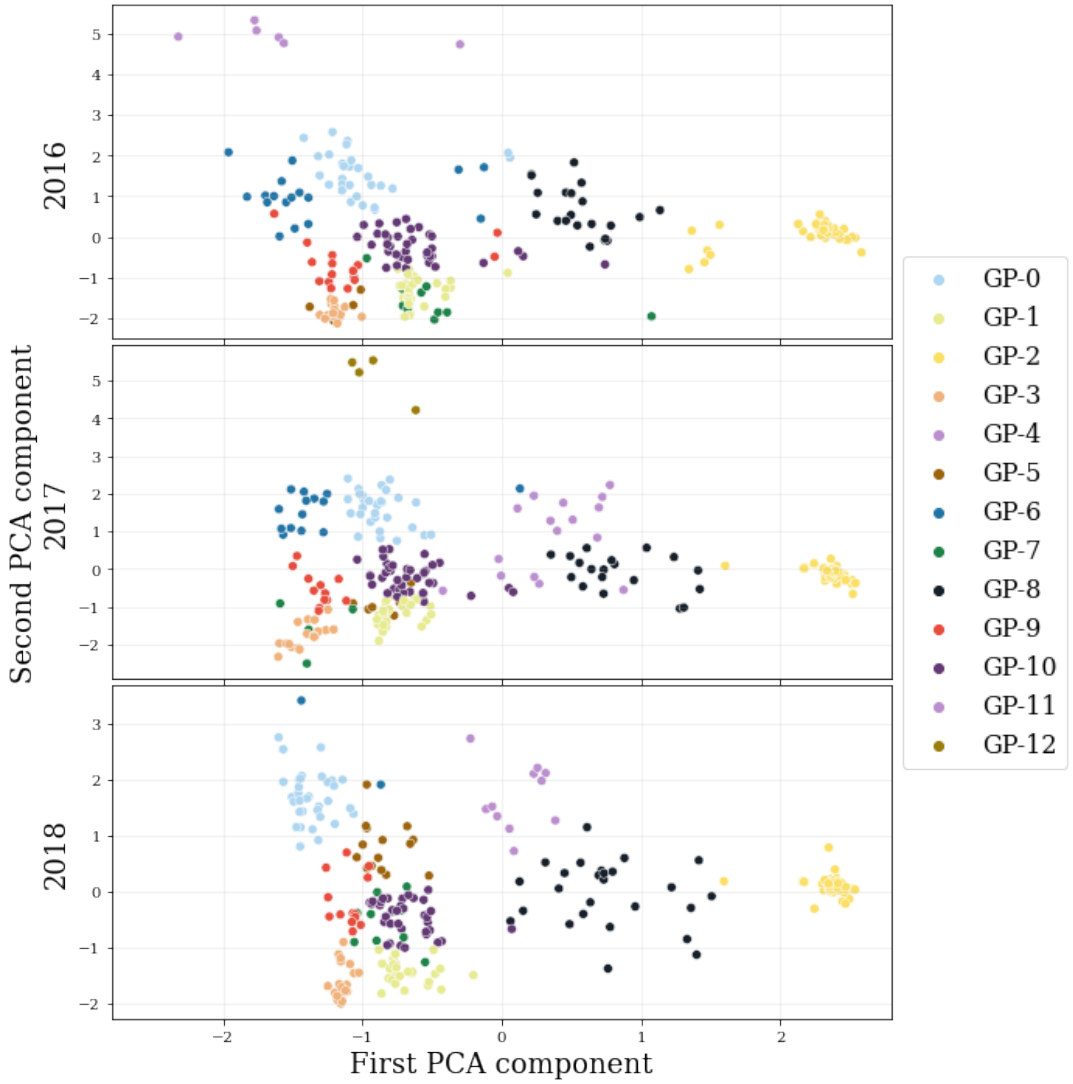

Figure 5. Clusters obtained for the years of: 2016, 2017 and 2018.

The largest group obtained by the algorithm had 61 countries, in the year of 2016. For the year of 2018 the largest group had 55 instances, which included countries such as: Scotland, Puerto Rico, Sudan, Greenland, French Guiana, and other insular countries. In the year of 2016 the smallest group (GP-4) had 6 countries, including Brazil and South Africa, and for 2018 there was a cluster (GP-6) with only 2 countries - Pakistan and Benin.

Table 4. Centroid position of each cluster for the year of 2018.

Cluster	Principal Component			
	1	2	3	4
GP-0	-1,3644	1,6779	-0,1812	-0,5156
GP-1	-0,6904	-1,4298	0,2557	0,2306
GP-2	2,3528	0,1056	0,0500	-0,1709
GP-3	-1,1229	-1,5919	0,6466	-1,3388
GP-4	0,1075	1,6880	-0,1623	-0,6753
GP-5	-0,8517	0,7934	-0,2138	0,9782
GP-6	-1,1529	2,6619	6,7145	1,9210
GP-7	-0,8311	-0,5538	-1,9913	0,1865
GP-8	0,7576	-0,0513	-0,0764	0,2965
GP-9	-1,0967	-0,1794	0,0754	-0,3271
GP-10	-0,6832	-0,4638	-0,1789	0,8554

All developed countries were assigned to GP-3 or GP-1, at the bottom left of the Figure 5 (2018). The centroid of these groups presented similar values for the second principal component meanwhile contrasting from the others (Table 4), and these groups had virtually no reduction of instances. And, as can be observed in Figure 5 (2018), GP-2 is isolated from the other clusters - highest first component, Table 4. This group consists mostly of insular countries and so it is believed that, due to its specific geographical conditions, such countries remained in the same, and isolated, cluster throughout the years.

Furthermore, it is important to understand how countries are performing in the SGDs over the years, in order to know which committed parties are obtaining tangible results. Figure 6 presents the relative motion of the analysed countries, in relation to the clusters, throughout the years.

From Figure 6 it is possible to notice that some clusters suffered considerable changes through the years, such as groups 4, 5, 6, 7 and 9. GP-5 presented the biggest increase in number of instances, in 2016 it had 4 countries and by the year of 2018 it reached 15. It is also remarkable that GP-6, in 2018, had only two countries - Pakistan and Benin - presenting the highest value in the third principal component (strongly influenced by indicator $4.1.2$). Some countries suffered changes in their classifications every year, such as: Russia, Kazakhstan, Thailand, Saudi Arabia, Iraq, Chile, among others.

Analyzing Figure 5, considering the three years, in combination with Figure 6 a relative movement of GP-4 can be observed, towards the bottom of the Figure 5. This groups consists of African countries and the most dissonant feature is related to the first component of the PCA - which is related to reduced inequalities, peace, justice and strong institutions. Therefore, those countries are having tangible improvements when related to these areas, considering developed countries as a reference (GP-1 and GP-3). Although other African countries where assigned to group 0, which remained stable throughout the years.

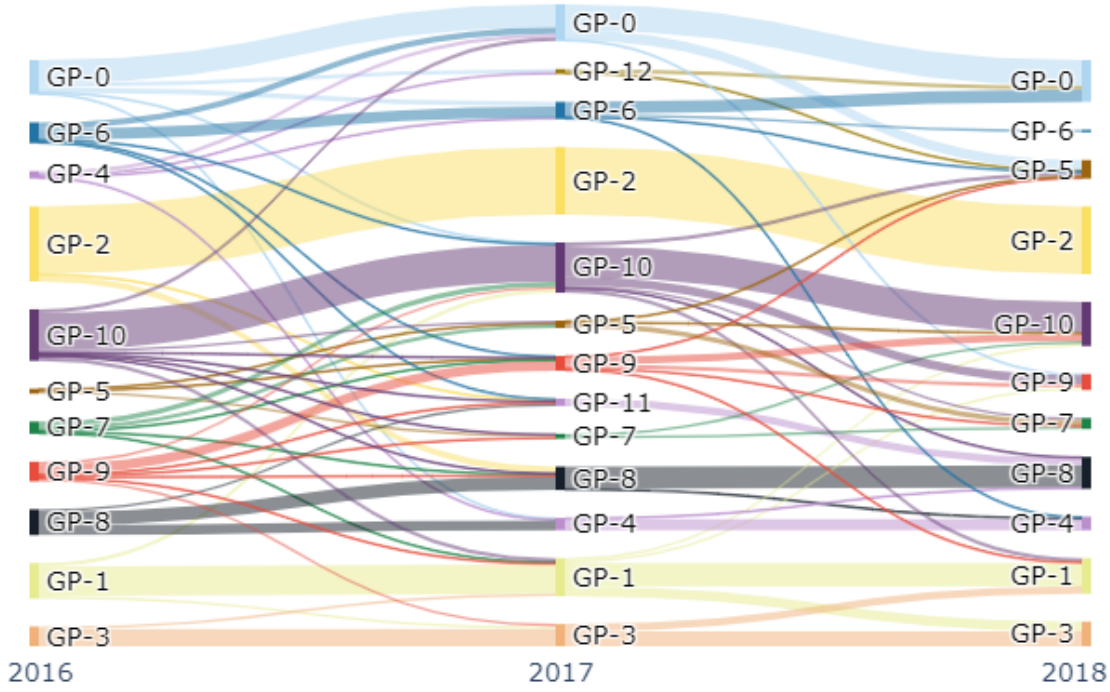

Figure 6. Clusters evolution through years

5. Conclusion and Future Works

Sustainable development is becoming more and more relevant nowadays, as it can be observed in the current international political scenario. UN provides a database - per year - of countries and regions performance in all 17 SDGs, according to established indicators from each. There are over 290k instances in UN database, hence establishing relations among instances can be a challenge. The present article used a clustering algorithm to compare countries according to their performance on the SDGs.

The results allowed a systematic analysis and interpretation of the countries. These analyses were possible due to the proposed methodology that perform the data preprocessing, dimension reduction and clustering of the countries through the years. To determine the number of clusters and the number of features to be considered a loss function was proposed , combining the *Silhouette Coefficient* with the total variance obtained by the dimensionality reduction algorithm.

SDGs 3, 4, 10 and 16 had the biggest weight when reducing the data dimensionality - considering under-5 mortality rate, educational level and representation in international organizations. The algorithm divided all 250 countries in 11 groups for the years of 2016 and 2018. Most insular countries were assigned to the same group, forming the largest groups throughout the years (55 instances), while being the most distinct group. Developed countries were assigned to two groups.

As future work, it is planned to continue the analysis of the classifications obtained by the algorithm. However, it was observed a considerable amount of missing information for several countries in the UN database and, due to the initial treatment of the data, some indicators were greatly influenced by the geographical properties of the countries (continental and insular countries), that may have biased the results of this study. It is

noteworthy that the presented algorithm considered information from all SDGs for three years, providing a wide range of possible analysis, considering a specific year or over the years.

References

Bishop, C. M. (2006). Pattern recognition and Machine Learning. Springer.

Booth, L. (2010). Targets as a policy tool. https://www.parliament.uk/globalassets/documents/commons/lib/research/key_issues/key-issues-targets-as-a-policy-tool.pdf. Accessed: 2021-08-03.

Davis, K. E., Kingsbury, B., and Merry, S. E. (2012). Indicators as a technology of global governance. Law & Society Review, 46(1):71–104.

Ferreira, B., Iten, M., and Silva, R. (2020). Monitoring sustainable development by means of earth observation data and machine learning: a review. Environ Sci Eur, 32(120).

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems. O'Reilly Media.

in China, U. N. (2015). Sustainable development goals officially adopted by 193 countries. http://www.un.org.cn/info/6/620.html. Accessed: 2021-07-13.

Jabbari, M., Motlagh, M. S., Ashrafi, K., and Abdoli, G. (2019). Differentiating countries based on the sustainable development proximities using the sdg indicators. Environment, Development and Sustainability, pages 1–19.

Kijewska, A. and Bluszcz, A. (2016). Research of varying levels of greenhouse gas emissions in european countries using the k-means method. Atmospheric Pollution Research, 7(5):935–944.

Lim, S. S., Allen, K., Bhutta, Z. A., Dandona, L., Forouzanfar, M. H., Fullman, N., Gething, P. W., Goldberg, E. M., Hay, S. I., Holmberg, M., et al. (2016). Measuring the health-related sustainable development goals in 188 countries: a baseline analysis from the global burden of disease study 2015. The Lancet, 388(10053):1813–1850.

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of clinical epidemiology, 110:63–73.

Raszkowski, A. and Bartniczak, B. (2019). On the road to sustainability: Implementation of the 2030 agenda sustainable development goals (sdg) in poland. Sustainability, 11(2).

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation an validation of cluster analysis. Journal of Computational and Applied Mathematics, pages 53–65.

Santos, L. and Santos, T. (2017). Os ods e seus indicadores: novas classes gramaticais, uma mesma morfologia. Pontes, 13:13–16.

Shahbaz, M., Sharma, R., Sinha, A., and Jiao, Z. (2021). Analyzing nonlinear impact of economic growth drivers on co2 emissions: Designing an sdg framework for india. Energy Policy, 148:111965.