# Multivariate Analysis Report

Armando Sobrinho Assembleia 89614,
Carolina Francisco Constantino 89618,
João Afonso Fernandes Batista 89629,
Ruotong Yang 89638

**Abstract**

Humanity's purpose has been a vivid discussion matter for centuries. Happiness is the answer many find and is thought to be one of the noblest aspirations people can have. By implementing unsupervised learning clustering methods, such as Single, Average and Complete Linkage, Ward's Method, K-Means and K-Medoids, it is shown that the countries can be split in two groups, which is coherent with the idea of separating the world in two fractions : happy and unhappy countries. After applying different external indexes to the clusters and evaluating the presence of noise in the data, it is possible to conclude that the K-Means is the most reliable. In this paper, it is also studied which features have the biggest influence on the happiness of the citizens of a given country. By implementing supervised learning method to study the data, it was possible to conclude that happier countries have high values of GPD, Life expectancy and Year of Schooling, whereas unhappy countries have high values of Homicide Rate and number of Refugees. By having 13 features about a country, it is possible to determine if the citizens of that country are happy or unhappy with about 80% certainty. Knowing which factor have the bigger impact on the happiness of people is very important to governmental institutions when making decisions that would impact the lives of many. This paper concludes that the pillars to a happy society are a strong economy, health, education and a sense of safety.

*Keywords:* Happiness, Principal Component Analyses, Cluster Analyses, Supervised Learning, Decision Tree, Neural Network

## 1. Introduction

What influences and defines the happiness of individuals? That is a question with a very interesting and hard to reach answer. Philosophers and thinkers have tried for millennia to answer it and have not found a satisfactory answer. In a society where the purpose of life is "to live the most happy life possible", this question has never been more relevant. Today, data is generated at an unprecedented rate, so it is possible to try to do what thinkers failed. To answer: "What makes someone happy?".

For 154 countries, features like `GDP per Capita`, `Suicide Rate` or `Unemployment` are studied in an attempt to understand which features have higher importance on the happiness of the citizens of such countries, based on the "World Happiness Report" [1]. We also want to study how many clusters the data tends to naturally have. Finally, the development of a classifier that, given features about a country, would classify that country on a scale of happiness.

In Section 3, we will make a preliminary analysis of our dataset by imputing missing values, analysing outliers, and performing dimensional reduction techniques.

After that, in Section 4, we will do unsupervised cluster analysis applying Single Linkage, Complete Linkage, Average Linkage, Ward's Method, K-Means and K-Medoids to the different datasets found in the previous Section. In order to find the best number of cluster and best dataset, we will apply internal and external indexes to the clusters found. Later on, we study the robustness of the best methods, by analysing our dataset with added noise.

In supervised learning, Section 5, to create classifying methods to our dataset, four methods will be applied: Discriminant analyses with linear and quadratic rules, Decision Trees and Neural Networks. Further, the same classifying methods will be applied to the partition clusters obtained.

Finally, in the last Section 6 we will set our conclusions and potential future works.

## 2. Dataset Description

The World Happiness Report, from Kaggle [1], ranks 154 countries by their happiness levels, by asking people how would they rate their happiness.

From the **Human Development Data Center** [2], we chose several indicators from different dimensions of 2015, which are sourced from international data agencies, to create our dataset.

First of all, the dataset is composed by 154 observations of 21 variables. The variables studied for the different countries are:

`Country` [1]: name of the countries
`Region` [1]: indicates where the country belongs to. It is divided it to 10 zones: 'AustraliaandNewZealand', 'CentralandEasternEurope', 'EasternAsia', 'LatinAmericaandCaribbean', 'MiddleEastandNorthernAfrica', 'NorthAmerica', 'SoutheasternAsia', 'SouthernAsia', 'Sub-SaharanAfrica' and 'WesternEurope'.
`Happiness Score` [1]: a metric measured by asking people how would they rate their country happiness on a scale of 0 to 10, where 10 is the happiest.
`Class`: we consider two classes of countries based on the happiness score, Figure 1. The class 1 contains 78 observations which are classified as unhappy and the class 2 has 76 observations which are classified as happy.
`Mean years of schooling` [3]: number of years that a child is expected to study.
`Carbon Dioxide Emissions per GDP` [4]: emissions stemming from the burning of fossil fuels, gas flaring and cement manufacture, and emitted by forest biomass through reduction of forest, expressed in kilograms.
`Change florest area` [5]: the percentage under forest cover
`Gender Inequality Index` [5]: reflects the inequality between women and men in reproductive health, empowerment and the labour market.
`Child malnutrition` [6]: the percentage of children, under age 5, who are more than two standard deviation below the median height for age.
`Current health expenditure` [6]: the spending on healthcare goods and services, expressed as a percentage of GDP.
`Homicide rate` [7]: of unlawful deaths inflicted upon a person with the intent to cause death or serious injury, expressed per 100,000 people.
`Net migration rate` [8]: the ratio of the difference between the number of in-migrants and out-migrants from a country to the average population, expressed per 1,000 people.
`Unemployment` [9]: the percentage of the labour force population that is not employed but is available for work.
`Vulnerable employment` [9]: contributing family workers and own-account workers as a percentage of total employment.
`Suicide Rate` [10]: the number of deaths from purposely self-inflicted injuries, expressed per 100,000 people.
`Contraceptive prevalence` [11]: the percentage of women (married or in-union) in reproductive age currently using any contraceptive method.
`GDP per capita` Gross Domestic Product in a particular period divided by the total population in the same period.
`Life expectancy ate birth` [8]: the number of years a newborn is expected to live.
`Median age` [8] (years): age that divides the population distribution into two equal parts—that is, 50 percent of the population is above that age and 50 percent is below it.
`Refugees by country of origin` [12]: number of people (thousands) who have fled their country of origin because of a well founded fear of persecution due to their race, religion, nationality, political opinion or membership in a particular social group and who cannot or do not want to return to their country of origin.
`Share parliament by women` [13]: the proportion of seats held by women in the national parliament expressed as a percentage of total seats.

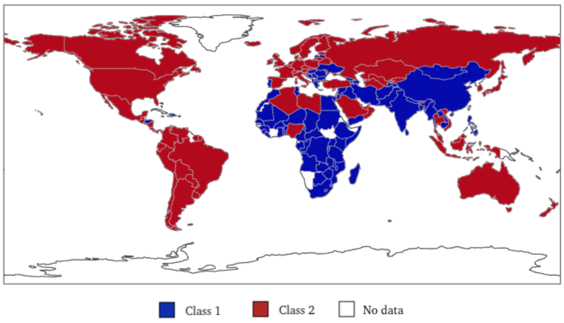

**Figure 1: World map with the original dataset class of happiness.**

## 3. Pre-process

Firstly, we used the command `as.factor` in R to define the variables `Country`, `Region` and `Class` as factors.

### *3.1. Imputation : dealing with missing values*

We noticed that there were some missing values for each variable, except for `Net migration rate`, `Unemployment`, `Vulnerable employment`, `Life expectancy at birth` and `Median age`.

To avoid extras difficulties, since the majority of the procedures that were intended to be used imply dealing with a dataset free of empty cells, we implemented a particular method of imputation called **Predictive Mean Matching** [14]. This algorithm allows us to predict the missing values in the dataset and, as a consequence, it grants the possibility of applying traditional techniques when analysing the data. This procedure was implemented using `mice` library available in `R` [15].

Due to different range between each variable, specially `GDP per capita` that has much larger values when considering the others, we decided to standardize all of our observations with the exception of `Happiness Score`.

### *3.2. Outliers' Analysis*

By plotting the boxplot of the standardized data, Figure 2, it can be notice that only a few variables have noticeable outliers. After a quick analysis, we realised that 76 of our countries are outliers in at least one dimension. However, the majority of them are outliers in just one or two dimensions and the maximum number of dimensions of an outlier is 4, corresponding to Lesotho (country 94). As the dataset has in total 17 variables, we decided to not delete any outlier.

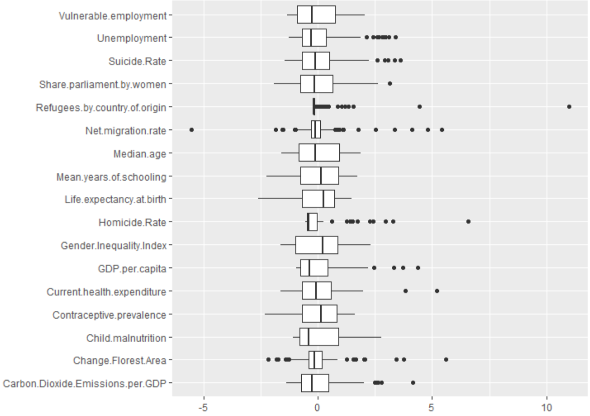

**Figure 2: Boxplot of the dataset with standardization.**

### *3.3. Data Reduction*

In this report, we will be considering four different transformations of the original dataset.

`Standardized` : The first one is based on the standardization of the original data .

`PCA` : The second transformation is the result of a principal component analysis preserving a total variability of 80%. This new dataset allowed a significant data reduction equivalent to ten variables.

`Corr` : The third modification was obtained by removing high correlated variables in order to guarantee a correlation matrix with absolute values below 0.8. In this particular case, by observing the correlation plot, Figure 3, we decided to eliminate the following variables :`Gender Inequality`, `Child Malnutrition`, `Vulnerable Employment` and `Median Age`. Consequently, this transformation achieved a data reduction of four .

`Corr + PCA` : The forth alteration is a mixture of what we have done in the second and third transformation. Firstly, we applied the same procedure used in the third modification, eliminating exactly the same variables. After that, we proceeded as in the second transformation. This procedure allowed a data reduction of ten variables .

Further on, it will be important to give some interpretation to some of the principal components related to this last transformation. We highlight the following principal components:

- **PC1** : gives the contrast between this two sets of features - {`Homicide Rate`, `Refugees by country of origin`} and {`Mean years of schooling`, `Life expectancy at birth`, `GDP per capita`}
- **PC3** : gives the contrast between {`Refugees by country of origin`} and {`Suicide Rate`}
- **PC5** : gives the contrast between {`Suicide Rate`} and {`Homicide Rate`, `Unemployment`}.

High positive values of one of this PC represent high values of the features present in the first set and low values of features in the second set. Negative values represent the opposite.

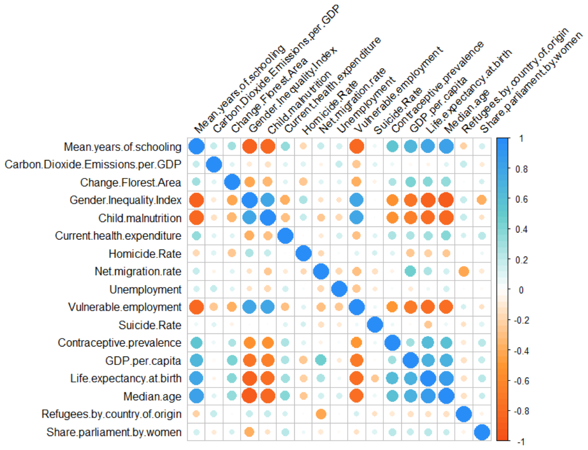

**Figure 3: Correlation matrix of the 17 features under study.**

## 4. Unsupervised Learning

### *4.1. Clustering: Internal and External indexes*

In this section, we will use the following clustering methods:

- Hierarchical Clustering
  - Single Linkage
  - Complete Linkage
  - Average Linkage
  - Ward's Method
- Partitioning Method
  - K-Means
  - K-Medoids

In order to choose the best number of clusters, we decided to consider five indexes [16]: Calinski-Harabasz (CH), Dunn (D), Gamma (G), Ratkowsky-Lance (RL), Wemmert-Gancarski (WG).

Using the function `intCriteria` in library `clusterCrit` available in R [16], we obtained the best number of clusters for each transformation of the original dataset, taking into consideration all the indexes and clustering methods previously mentioned. We established the range of the best number of clusters to be between 2 and 30. Both Table 1a and Table 1b illustrate these results.

Analysing Table 1, we can conclude that the indexes suggest partitioning the data into two clusters. This result is clear when interpreting Table 1a. However, the Table 1b proposes the idea of creating a much higher number of clusters. Since the data has only 154 countries, its partitioning into a large number of clusters (21 or more) may be inappropriate for further analysis and interpretation. Choosing two clusters based on the previous results is also coherent with the original number of classes of the data.
For these reasons, from now on we will be considering the partition of the data into two clusters.

We are now interested about finding the best method and transformed dataset, which provide a reliable classification considering the original classes of the dataset. To study that, we applied different criteria such as overall accuracy, recalls, precisions and overall F1-score. These results are displayed in Table 2.

We can observe that the performances of the linkage clustering methods are identically low. This can be justified by the fact that each one of this three clustering methods will provide two clusters, where one of those clusters has only one element.

The Ward's Method has a better performance comparing with the previous methods, especially for dataset with `Corr + PCA`.

When analysing K-Means and K-Medoids, we can infer that they have overall better result than hierarchical clustering methods. As a matter of fact, these partitioning methods and the Ward's Method are the ones where further study is worth.

We may also highlight that the `Corr + PCA` dataset provides a relevant amount of data reduction and also demonstrates high values of external indexes, regardless of which one of the final three methods are under study. For this reason, we will continuing to analyse Ward's Method, K-Means and K-Medoids jointly with `Corr + PCA` dataset, whilst trying to choose what is the best partition of the data.

| | Single Linkage | | | | | Complete Linkage | | | | | Average Linkage | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CH | D | G | RL | WG | CH | D | G | RL | WG | CH | D | G | RL | WG |
| Standardized | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 7 | 2 | 2 | 7 | 2 |
| PCA | 2 | 2 | 2 | 2 | 2 | 4 | 2 | 2 | 4 | 2 | 3 | 2 | 2 | 3 | 2 |
| Corr | 2 | 2 | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 2 | 28 | 2 | 2 | 10 | 2 |
| Corr + PCA | 2 | 2 | 2 | 2 | 2 | 6 | 2 | 2 | 6 | 2 | 27 | 2 | 2 | 3 | 2 |

**(a)**

| | Ward's Method | | | | | K-Means | | | | | K-Medoids | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CH | D | G | RL | WG | CH | D | G | RL | WG | CH | D | G | RL | WG |
| Standardized | 2 | 27 | 30 | 2 | 2 | 2 | 28 | 29 | 2 | 2 | 2 | 21 | 23 | 2 | 2 |
| PCA | 2 | 29 | 30 | 5 | 2 | 2 | 28 | 29 | 5 | 2 | 2 | 22 | 30 | 7 | 2 |
| Corr | 2 | 24 | 30 | 3 | 30 | 2 | 28 | 30 | 2 | 30 | 2 | 28 | 30 | 3 | 30 |
| Corr + PCA | 2 | 28 | 29 | 4 | 30 | 2 | 29 | 30 | 4 | 30 | 2 | 25 | 30 | 7 | 30 |

**(b)**

**Table 1: Best number of clusters based on internal indexes regarding different datasets and clustering methods.** The internal indexes CH, D, G, RL and WG stand for Calinski-Harabasz, Dunn, Gamma, Ratkowsky-Lance and Wemmert-Gancarski, respectively. It was considered the range of the best number of clusters to be between 2 and 30.

| | Single Linkage | | | | | | Complete Linkage | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 |
| Standardized | 0.500 | 0.987 | 0.000 | 0.503 | 0.000 | 0.333 | 0.500 | 0.987 | 0.000 | 0.503 | 0.000 | 0.333 |
| PCA | | | | | | | | | | | | |
| Corr | | | | | | | | | | | | |
| Corr + PCA | | | | | | | | | | | | |

**(a)**

| | Average Linkage | | | | | | Ward's Method | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 |
| Standardized | 0.500 | 0.987 | 0.000 | 0.503 | 0.000 | 0.333 | 0.734 | 0.564 | 0.908 | 0.670 | 0.863 | 0.749 |
| PCA | | | | | | | 0.727 | 0.513 | 0.947 | 0.655 | 0.909 | 0.752 |
| Corr | | | | | | | 0.721 | 0.500 | 0.947 | 0.649 | 0.907 | 0.746 |
| Corr + PCA | | | | | | | 0.734 | 0.487 | 0.987 | 0.652 | 0.974 | 0.769 |

**(b)**

| | K-means | | | | | | K-medoids | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 | ACC | Pr(0) | Pr(1) | Re(0) | Re(1) | F1 |
| Standardized | 0.786 | 0.667 | 0.908 | 0.726 | 0.881 | 0.795 | 0.766 | 0.603 | 0.934 | 0.696 | 0.904 | 0.783 |
| PCA | 0.786 | 0.667 | 0.908 | 0.881 | 0.726 | 0.783 | 0.779 | 0.628 | 0.934 | 0.710 | 0.907 | 0.793 |
| Corr | 0.766 | 0.615 | 0.921 | 0.700 | 0.889 | 0.780 | 0.740 | 0.641 | 0.842 | 0.696 | 0.806 | 0.746 |
| Corr + PCA | 0.766 | 0.628 | 0.908 | 0.704 | 0.875 | 0.778 | 0.792 | 0.667 | 0.921 | 0.729 | 0.897 | 0.803 |

**(c)**

**Table 2: External indexes regarding different datasets and clustering methods.** The external indexes ACC, Pr(0), Pr(1), Re(0), Re(1) and F1 stand for overall accuracy, negative predicted value, positive predicted value, specificity, sensitivity and overall F1-score, respectively.

### *4.2. Noise*

By previous results, we conclude that the best performing methods are K-medoids, Ward's Method and K-means, which had very similar results. However, we need a robust method of clustering, because the data under study is based on real-world observations, so it is most likely to have some noise attached.

For this reason, to find the best method, in terms of its robustness, we decided to add noise to our dataset. We, then, faced an issue: which noise should be considered? To work around this problem, we added a range of noise following a normal distribution with mean 0 and variance between 0 and 0.2, with step size of 0.001 to the dataset `Corr + PCA`. We considered this new dataset as `Corr + PCA + Noise`.

Then, we applied the three different methods of clustering to both of dataset `Corr + PCA` and `Corr + PCA + Noise`. With the resulting partitions, we calculated the agreement between them, for each method. Due to the randomness of our methods and noise, we executed this relative index procedure 20 times for each different noise, displaying in Figure 4 the mean results of multiple iterations. As it can be seen, the best method dealing with noise is K-Means, which has better agreement values for the noise added, than both of the other methods.

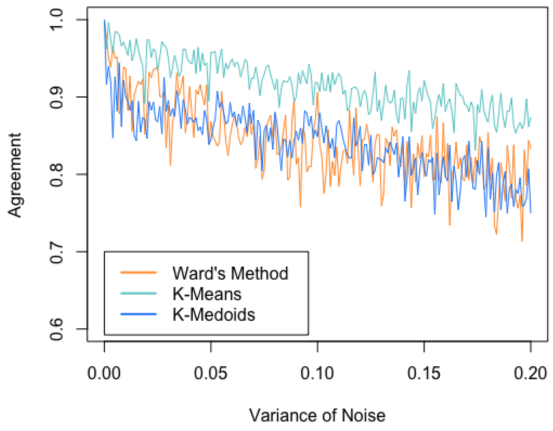

**Figure 4: Agreement between the clusters of `Corr + PCA` dataset and `Corr + PCA + Noise` by** Ward's Method, K-means and K-medoids, considering different variance's noise.

We finally conclude that the best partition cluster is the one obtained by applying K-Means to the dataset `Cor + PCA`. This choice grants a high data reduction and a reliable and robust clustering. The clusters can be seen in Figure 5.

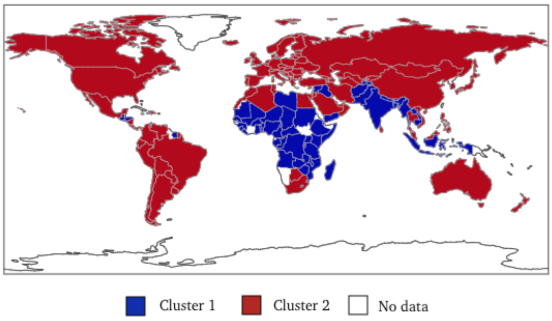

**Figure 5: World map with clusters of happiness found by applying K-Means to dataset `Cor + PCA`.**

## 5. Supervised Learning

In this section, we used the dataset `Corr + PCA` to create classifying methods using supervised learning. Four methods were applied and studied: Discriminant analyses with linear and quadratic rules, Decision Trees and Neural Networks.

Our dataset has 154 countries, which is not a large amount of data. Taking this into consideration, we chose a leave-one-out approach, since that is the method that uses the larger amount of data to train. However, training Decisions Trees and Neural Networks need a lot of computational power and a leave-one-out approach would take too long to compute. So we decided to use 90% of the dataset to train and 10% to test. This will lead to a random result that depends on the splitting done to the dataset. As such, we average the results over 20 simulations. The results are shown in the Table 3.

The results are satisfying with an Overall F1-Score ranging from 0.757 to 0.825. The method with the best performance is, without a doubt, the Linear Classifier.

We believe this is due to the low size of the sample for the amount of measures our dataset has. The more measures, the higher the number of parameter the Quadratic Classifier has to estimate. Therefore, the Quadratic Classifier needs more data to estimate them properly, while the linear classifier does not suffer from this issue.

| | Linear Classifier | Quadratic Classifier | Decision Tree | Neural Network |
|---|---|---|---|---|
| Overall Accuracy | 0.825 | 0.773 | 0.761 | 0.781 |
| Sensitivity | 0.855 | 0.816 | 0.745 | 0.774 |
| Specificity | 0.795 | 0.731 | 0.775 | 0.788 |
| Positive Predicted Value | 0.802 | 0.747 | 0.768 | 0.787 |
| Negative Predicted Value | 0.849 | 0.803 | 0.780 | 0.788 |
| Overall F1-Score | 0.825 | 0.772 | 0.757 | 0.779 |

**Table 3: External indexes for Discriminant analyses applied to the original class of the dataset.**

The classifying based on Decision Trees (done with the library `rpart`) is the method that would result in more misclassifications. Nevertheless, this method has a great quality of showing us what are the most important features used to differentiate the countries. The most important feature selected by the algorithm is the PC1, that can be seen in Figure 6, which confirms the validity of our Principal Component Analyses. More concretely, this shows that the features that have the higher influence on the happiness of people are `GDP per Capita`, `Life Expectancy at Birth`, `Mean Years of Schooling`, `Refugees by country of origin` and `Homicide Rate`.

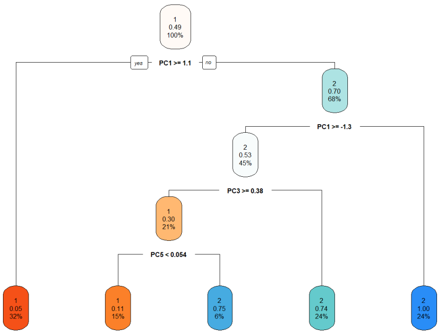

**Figure 6: Decision Tree when applied to original classes.**

Finally, using the library `neuralnet`, we decided to train a Neural Network [17]. The efficiency of neural networks rise tremendously if the data are min-max normalized. So we changed the minimum of each feature to 0, the maximum to 1, and every other value proportionally. Here is an example of a Neural Network with 3 middle layers, Figure 7.

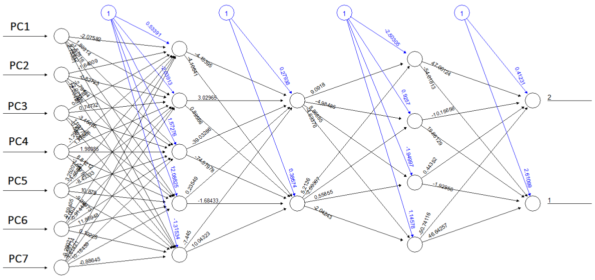

**Figure 7: Neural Network when applied to original classes.**

The neural network has a better F1-score than both the decision tree and the quadratic classifier. However, it is still worse than the linear classifier. Neural Network is a technology that is on the rise since the amount of data generated today is very large. Neural Network needs a big training set to learn and as a result this method is the one that suffer most from the small dimension of our dataset.

### *5.1. Supervised Learning on clusters*

We were also asked to perform the same analyses as in the previous section, but now on the classes from the clusters obtained in Section 4.2. We expect the performance of this analyses to be overall better than on the previous one. Our reasoning is based on the fact that the algorithms, that produce the clusters, analyse the data presented to them, find patterns and group the data according to those patterns. Classification algorithms have better performances on data that has patterns. So the algorithms we will apply to the cluster partitions will utilize those same patterns the cluster algorithms introduced to the data to separate the data more efficiently. The results are shown in the Table 4.

|  | Linear Classifier | Quadratic Classifier | Decision Tree | Neural Network |
|---|---|---|---|---|
| Overall Accuracy | 0.941 | 0.929 | 0.964 | 0.958 |
| Sensitivity | 0.875 | 0.857 | 0.935 | 0.942 |
| Specificity | 0.979 | 0.969 | 0.981 | 0.968 |
| Positive Predicted Value | 0.961 | 0.941 | 0.966 | 0.941 |
| Negative Predicted Value | 0.932 | 0.922 | 0.964 | 0.968 |
| Overall F1-Score | 0.936 | 0.921 | 0.959 | 0.953 |

**Table 4: External indexes for Discriminant analyses applied to the cluster partitions.**

We can see that the methods achieve higher levels of performance. The Linear Classifier is still better than the Quadratic Classifier. However, the Decision Tree and the Neural Network outshine the other methods with an Overall F1-Score higher than 0.95. As suspected, the classification on the partitions produced by the clustering algorithms is far more accurate.

It is important to note that the Decision Trees produced in this analyses, most of the times, only use one rule/question to make their decision, always based on the first Principal Component, as can be seen in Figure 8.

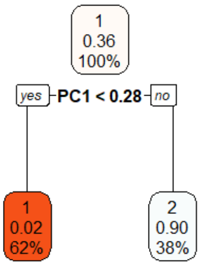

**Figure 8: Decision Tree when applied to the clustering partitions**.

Clustering is a good method to find patterns in the data that are not easy to identify. However, the interpretation of the partitions produced by the algorithms is of extreme importance, because they will produce patterns even if they are not present in the data.

Clustering is also useful if the original data might have classification errors. For example, if Switzerland, a country with very high happiness score, is incorrectly classified as unhappy, then the classification algorithms will most likely classified it as happy, since it has a very high `GDP per capita` and low `Homicide Rate`, diminishing the performance of the classifier. However, cluster analyses would unveil this problem and insert this country in the cluster that contains, mostly, happy countries. Applying classification algorithms on the partitions made by the clusters would provide virtuous results as Switzerland would be, in fact, classified as a happy country.

We believe applying the classification methods to the partitions made by the clustering analysis would make sense if the data has strong patterns and if the chance of the original classification on the dataset being wrong is considerable. However, the cluster analysis might artificially impose patterns, hiding some important information in the process.

Even though the analysis on the original classes has a lower performance, we recommend classifying based on those because of the nature of our problem. Firstly, the data was gathered from trustworthy governmental institutions and thus, we believe the chance of misclassifications is low. Furthermore, choosing the clustering partitions strategy would be to accept that the happiness of the people living in certain countries can be fully explained by the measures we are analysing. And although we conclude that the measures certainly have a very big impact, they alone might not be enough to explain everything. Happiness is a very subjective matter that can be influenced by a lot of different factors, which makes it real hard to measure.

## 6. Conclusions

### *6.1. Conclusion*

To sum up, we saw that partitioning the dataset into two clusters is an appropriate idea. After that, we concluded that the Ward's Method, K-Means and K-Medoids provide similar results, regarding external indexes. This analysis also pointed out that the dataset with `Corr + PCA` was the one which presented the best balance between data reduction and high values of external indexes. At this stage we were still indecisive about the best method to choose among those three. We finally concluded that the K-Means jointly with the `Corr+PCA` transformed dataset is the best combination of strategies, regarding resistance to noisy data.

Classification on the partitions made by the clusters produced marvelous results with Overall F1-Scores over 0.95. However, we decided, due to the nature of our problem, to classify using the original classes. In this classification, the Linear Classifier had the best performance with an Overall F1-Score of 0.825. The performance of the other classifiers, including a Neural Network, suffered due to the sample size of 154 countries, which is not enough to train efficiently these kinds of methods. In the end, a classification that is correct, on average, for every 4 out of 5 countries is a very satisfactory result, having in mind the nature of our problem.

From the 17 initial features, and after rejecting 4 of those due to high correlation with another feature, only 7 dimensions remained after the dimensional reduction analysis was done. It is thought that happiness is influence by a lot of things but we concluded that the most important factor is wealth and health. Education and Security are also not far behind in importance. Without achieving satisfactory results on these areas, a country will hardly have happy citizens.

The human mind is very complex and an emotion like happiness cannot be easily understood, let alone quantified. This presents a challenge to this analyses, but it is also what gives it its merit.

### *6.2. Conversation with researcher*

To discover the features that are more likely to influence the happiness of the citizens of a given country, we decided to apply different learning methods to our datasets.

Firstly, our analysis shows that the best number of clusters is two. This means that the natural way to divide the countries is into 2 classes: happy countries and unhappy countries.
As can be seen in Figure 5, the unhappy countries are concentrated in Sub-Saharan Africa and Southern Asia, whereas the other regions,that are being considered, can be associated to happy countries.

Secondly, the variables `Gender Inequality`, `Child Malnutrition`, `Vulnerable Employment` and `Median Age` are not necessary to be studied since they have a very high correlation with other features and would not provide much information.

By analyzing the dataset, we realized that the variables `GDP per Capita` and `Life Expectancy at Birth` have the biggest influence on the classification of the level of happiness, `Mean Years of Schooling`, `Refugees by country of origin` and `Homicide Rate` are also important. Hence, to have happy citizens, a country needs a prosperous economy, a strong public health, a sense of security and a solid education.

On the other hand, our result of classification has an accuracy, in the best case, of 0.825, when using a Linear Classifier. This is quite satisfactory, but might suffer from two factors: the complex nature of the problem and the low sample size of our dataset. The first problem might be overcome by adding more significant features that might be useful to explain happiness. The second might be very difficult to pass over since the number of countries in the world is very limited. A possible solution could be to change the granularity of the analysis to compare regions or cities instead of whole countries.

Therefore, we propose as future work, doing the same analyses but with other, more varied, measures. For example, most followed religion, if drugs are legalized or not, or the most common climate. A lot of conclusion and relations can be made between features and the happiness of people.

Another possibility is changing the granularity, going from analysing countries to a analysing smaller regions like states, or even cities or villages. An expected conclusion is that richer regions/cities will be happier but many more factor will have an impact in the matter.

Also, it would be interesting to study the happiness over several years, to see how the world has changed in the previous years, and analyze the evolution of the contributions to the happiness of a country in each period of time.

## References

[1] Kaggle. World Hapiness Report. `https://www.kaggle.com/unsdsn/world-happiness`, .

[2] Human Development Data Center. UNITED NATIONS DEVELOPMENT PROGRAMME Human Development Reports. `http://hdr.undp.org/en/data`, .

[3] UNESCO Institute for Statistics (2020). Barro and Lee (2018), ICF Macro Demographic and Health Surveys, UNICEF Multiple Indicator Cluster Surveys and OECD (2019b), .

[4] United Nations Statistics Division (2020a). Global SDG Indicators Database (accessed 21 July 2020). https://unstats.un.org/sdgs/indicators/database/., .

[5] HDRO. World Development Indicators database. Washington, DC.(Accessed 22 July 2020.). http://data.worldbank.org., .

[6] World Bank (2020a). World Development Indicators database. Washington, DC.(Accessed 22 July 2020). http://data.worldbank.org, .

[7] UNODC (United Nations Office on Drugs and Crime) (2020. DATAUNODC.(Accessed 21 July 2020). https://dataunodc.un.org, .

[8] UNDESA (2019a). World Population Prospects: The 2019 Revision. Rev 1. New York. (accessed 30 April 2020). https://population.un.org/wpp/., .

[9] ILO (2020). ILOSTAT database (accessed 21 July 2020). https://ilostat.ilo.org/data/., .

[10] World Health Organization. Global Health Observatory. (Accessed 21 July 2020). www.who.int/gho/, .

[11] UNDESA (2020). World Contraceptive Use 2020. New York. (accessed 21 July 2020). https://www.un.org/en/development/desa/population/publications/dataset/contraception/wcu2020.asp., .

[12] UNHCR. Office of the United Nations High Commissioner for Refugees. www.unhcr.org/, .

[13] IPU. Parline database: Monthly ranking of women in national parliaments. https://data.ipu.org/women-ranking, .

[14] Gerko Vink, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. Predictive mean matching imputation of semicontinuous variables. *Statistica Neerlandica*, 68(1):61–90, 2014. doi: https://doi.org/10.1111/stan.12023.

[15] Stef Buuren and Catharina Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45, December 2011. doi: 10.18637/jss.v045.i03.

[16] Bernard Desgraupes. Clustering Indices, Package clusterCrit for R . https://cran.r-project.org/web/packages/clusterCrit/vignettes/clusterCrit.pdf, .

[17] Conceicao Amado, Claudia Nunes, and Alberto Sardinha. *Análise Estatística de Dados Financeiros.* Sociedade Portuguesa de Estatística, 2019.