Università degli Studi di Milano

APPLICATION OF STATISTICAL LEARNING TECHNIQUES FOR COUNTRIES SOCIO-ECONOMIC DEVELOPMENT ANALYSIS

November 17, 2021

Angelina Khatiwada

MSc, Data Science and Economics

Contents

ABSTRACT

We study the differences in social, economic and environmental development of all the countries worldwide. We analyzed the United Nations countries dataset containing key development indicators such as GDP, life expectancy, levels of employment, education, and healthcare among others. The main purpose is to study countries socio-economic level using both unsupervised and supervised learning techniques. Principal Component Analysis, K-Means and Hierarchical Clustering methods are applied to explore hidden patterns in data and group the most similar countries. Moreover, all the countries are classified into four categories based on their development level using Random Forest, Multinomial Logistic Regression, K-NN and Neural Network models. Finally, we discuss the challenges of small dataset analysis such as high-dimensional feature space, class imbalance, overfitting, outliers and the techniques to overcome them.

1 INTRODUCTION

The process of a country economic, social, political and environmental development in the context of global competition depends on a number of indicators. The report form Boston Consulting Group 'Comparing Socioeconomic Development Across Nations' highlights at least 10 dimensions of socioeconomic development such as income, employment, income equality, economic stability, health, education, governance, the environment, infrastructure, and civil society [3]. These dimensions stand out as the most significant in differentiating leading countries in terms of both current levels of development and recent progress. The wealthiest countries with the highest current-level development scores are European nations such as Switzerland and Norway as well as Australia, New Zealand, Canada and the U.S.

According to the report produced by the United Nations Department of Economic and Social Affairs (UN DESA), all countries are classified into one of three broad categories: developed economies, economies in transition and developing economies [4]. The composition of these groupings is intended to reflect basic economic country conditions. Several countries (in particular the economies in transition) have characteristics that could place them in more than one category. Europeon Union countries, Major Developed Economies (G7), Norway, Iceland, Switzerland, Australia and New Zealand are categorized as Developed economies. South-Eastern Europe and CIS countries are considered as Economies in Transition. Geographical regions for developing economies are the following: Africa, East Asia, South Asia, Western Asia, and Latin America and the Caribbean. Within 'developing' category, the subgroup of the least developed countries (LDCs) is determined based on the per capita GNI, a human assets index and an economic vulnerability index.

For our research purposes, we analysed the data extracted from the UNData dataset containing key statistical indicators for 229 countries [2]. We also added the UN country development level categorization for classification purposes. We manually grouped all the countries in four classes, allocating the least developed economies in a separate class.

2 DATASET EXPLORATORY ANALYSIS

2.1 DATASET DESCRIPTION

The countries dataset based on UNData contains key statistical indicators of the countries covering four sections: General Information, Economic Indicators, Social Indicators, Environmental Infrastructure Indicators. The UNData is a web-based data service provided by The United Nations Statistics Division [2]. It makes possible to search and download a variety of statistical resources compiled by the United Nations statistical system and other international agencies. Our dataset contains information about 229 countries presented by 52 indicators that cover a wide range of statistical themes including agriculture, education, energy, environment, gender, health, labour market, manufacturing, population and migration, science and technology, transport and trade. The dataset also contains country label, country region and socio-economic development label (developed, developing, economies in transition. The dataset was uploaded from Kaggle competition and reflects the countries development level in 2017 [1]. Indicators names and their codes are presented in Appendix A, Table 3.

2.2 DATA CLEANING AND NULL VALUES HANDLING

Missing data of different format is set to NA, all the indicators are set as numeric class variables. We checked the percentage of missing values for all the numeric variables. The majority of columns has less that 10% of null values. Columns with the highest rate of missing values are: Net Official Development Assist. received (% of GNI) (100% missing), Health: Physicians (per 1000 pop.) (45.9% missing) and Pop. using improved sanitation facilities (urban/rural, %) (41.9% missing). We decided to remove the columns with less than 50% of present values, i.e. only one column.

Moreover, we analyzed null values in the rows and dropped the rows with more than 50% of NA values. For instance, Saint Pierre and Miquelon, Tokelau, Isle of Man, Monaco, American Samoa were among the countries eliminated. In total, we removed 15 countries, and most of them are dependent territories, Island Nations or the countries with a very small population. Final dataset includes 214 countries and 51 indicators.

After removing rows and columns not containing enough information, we performed an imputation of the remaining missing values. Instead of using simple techniques such as column median, mean, backward or forward fill, we applied MissForest imputation which is based on the Random Forest algorithm. Relevant studies suggest that MissForest outperforms other algorithms such as KNN-Impute, in some cases by over 50% [10]. MissForest is a non parametric imputation method applicable to various variable types. It yield OOB (out of bag) imputation error estimate and provides a high level of control on imputation process.

As a next step, We calculated basic descriptive statistics for the data (Appendix A, Table 3). All the 51 indicators are encoded to facilitate the analysis and further visualization. The dataset contains different social, economic, environmental indicators such as population, GDP per capita, employment in different economic sectors, life expectancy, etc., and all the variables are present in different units, therefore, data scaling is performed. We also checked that none of the variables contains near zero variance in it.

3 UNSUPERVISED LEARNING: PCA AND CLUSTERING

We analyzed the differences in social, economic and environmental development of the countries included in the UNData dataset using different unsupervised techniques. The main objectives were 1) to investigate underlying patterns in the countries dataset 2) to identify significant features and perform dimensionality reduction 3) to study the similarity among the countries 4) to create the groups of homogeneous countries in terms of socio-economic development. PCA was used to perform dimensionality reduction and visualize the results. We also explored K-means and Hierarchical Clustering methods for grouping similar countries based on the indicators given. Clustering results were compared to the original geographical regions grouping and development level classes present in the dataset.

3.1 CORRELATION MATRIX

After cleaning, imputing and standardizing the data, we calculated the correlation matrix between all the 51 variables. One of the important conditions to perform dimensionality reduction is the strong correlation between the variables. If the variables are not correlated, there is a higher number of dimensions that we need to keep to represent the information within the dataset. Figure 1 shows the correlation matrix between all the variables. We can see from the correlation plot that there are at least 3 groups of highly correlated indicators (please, refer to the Appendix A, Table 3 to see the description of the indicators):

- The first group includes I16, I25, I23, I11, I33, I8, I26, I29.
- The second one includes I10, I31, I35, I50, I49, I44, I27, I39, I28, I30, I40, I7, I41, I24, I42 among others. The indicators in this group are negatively correlated with the ones from the first group.
- The third group contains I19, I20, I5, I47, I2, I45, I1.

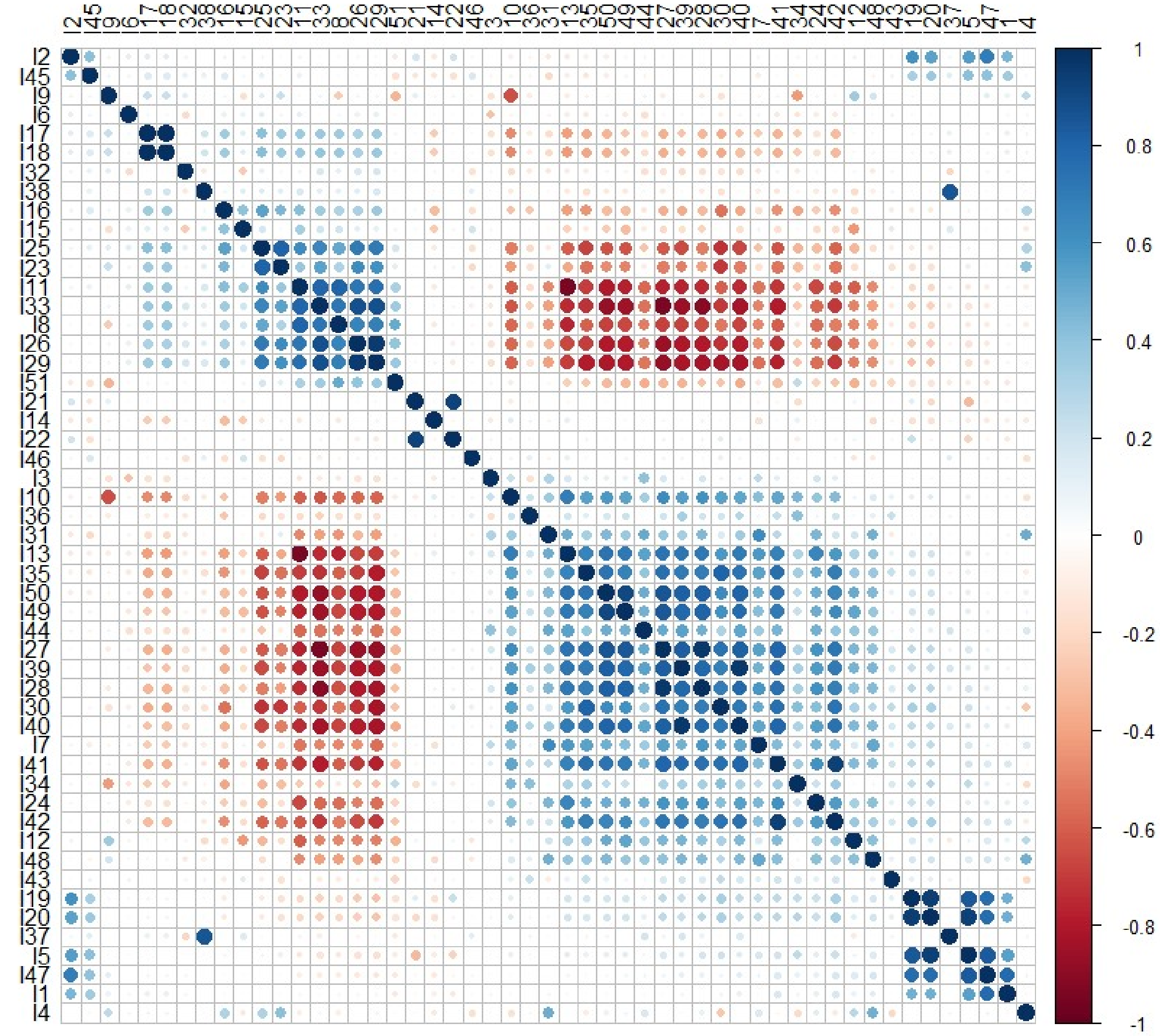


Figure 1: Correlation plot between the indicators: 1 - strong positive correlation, -1 - strong negative

High correlation may be explained by the fact that some of the variables are the linear combinations of the other variables. For example, Economy: Agriculture (% of GVA), Economy: Industry (% of GVA), Economy: Services and other activity (% of GVA) are a linear combination of each other. These variables do not add any information to the dataset and may be replaced by a lower-dimensional representation.

3.2 Principal Component Analysis

Principal component analysis (PCA) is an algorithm that reduces the dimensionality of the data while retaining most of the variation. It accomplishes this reduction by identifying directions, called principal components, along which the variation in the data is maximal [7]. We applied PCA to represent our dataset with a relatively few number of components instead of 51 indicators. We have already standardized the data before the analysis, so the principal components represent normalized eigenvectors of the covariance matrix between the indicators. They are ordered according to proportion of the variation in the data they contain. We visualize the percentage of the variance explained by the first ten components using a scree plot (Figure 2). The first principal component explains 32.8% of the variance in the dataset, the second component explains 9.5% of variance, third – 7%, etc. Eigenvalues, variance percentage explained and the cumulative variance percentage explained for all 51 principal components are shown in Appendix A, Table 4. We can see that 12 principal components have an eigenvalue larger than 1 and explain about 80% of the variance in the data. Therefore, we might take these 12 components to perform a dimensionality reduction preserving the patterns in the data and getting rid of noise. We also computed the correlation plot between the principal components to make sure that the new dimensions are not correlated with each other (Appendix B, Figure 15).

Using only two principal components, we miss around 60% of variance contained in the features, however, for visualization purposes we will stick to the two-dimensional representation. As a next step, we study the relation between original variables and principal components (Figure 3). All the indicators are plotted against two principal components. We also calculated the quality of representation for variables on the factor map and the contributions of the variables to the principal components (are reflected by the colors in the Figure 3). Positively correlated variables point to the same direction, negatively correlated are positioned on the opposite sides of the plot origin. The distance between variables and

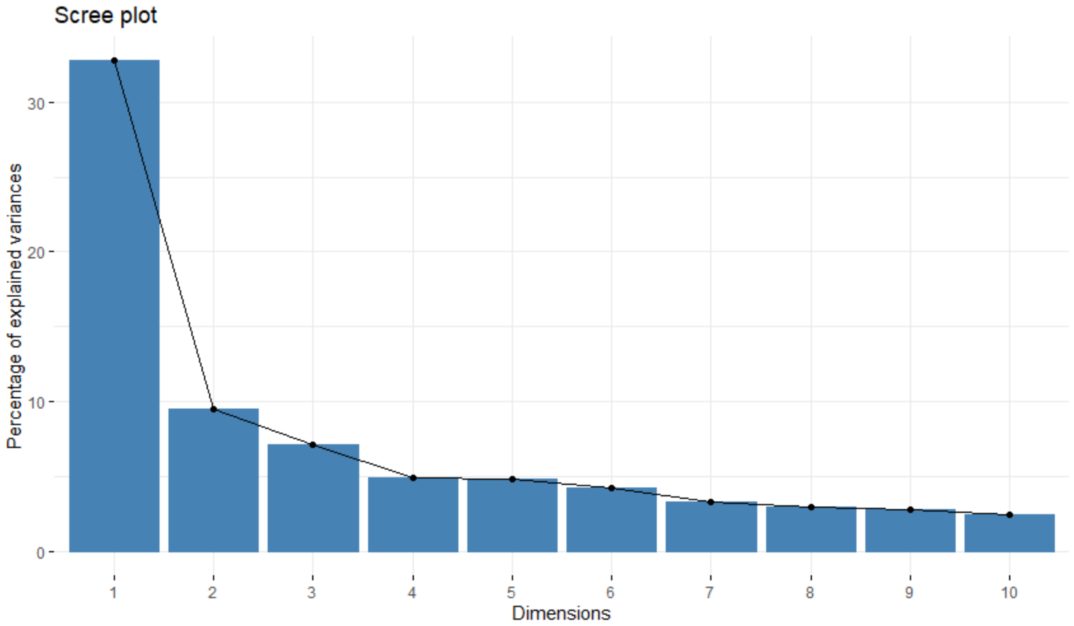


Figure 2: PCA Scree plot: variance percentage explained by the principal components

the origin measures the quality of the variables on the factor map, such that variables that are away from the origin are well represented on the principal component. Variables that are closer to the center of the circle are worse represented by the PCs.

We also calculated PCA loadings - the coefficients of the linear combination of the original variables for each PC. Larger loadings (positive or negative) indicate that a particular variable has a strong relationship with a PC. The loadings for the first 10 components are presented in the Appendix A, Table 5. For instance, the first component is positively correlated with: I8 - Economy: Agriculture (% of GVA); I11 - Employment: Agriculture (% of employed); I26 - Fertility rate, total (live births per woman); I29 - Population age distribution (0-14 years, %); I33 - Infant mortality rate (per 1000 live births). It is negatively correlated with the I49 - Pop. using improved drinking water (urban,%); I30 - Population age distribution (60+ years,%); I27 - Life expectancy at birth (females, years), I13 - Employment: Services (% of employed).

The second component can be explained by the following indicators: I1 - Surface area (km2); I2 - Population in thousands (2017); I5 - GDP: Gross domestic product (million current US$); I19 - International trade: Exports (million US$); I20 - International trade: Imports

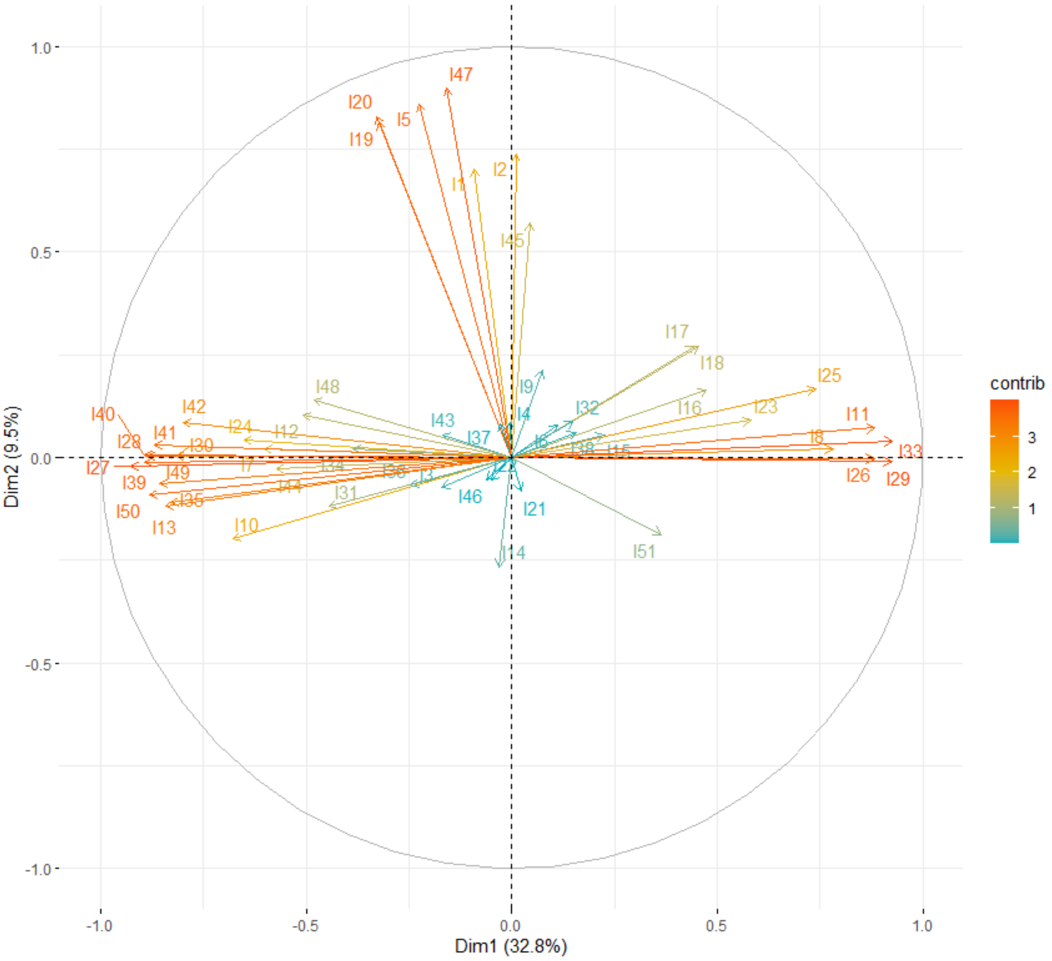


Figure 3: Variables plotted against the PC1 and PC2. Color indicates the contribution of the variables to the PCs.

(million US$); I47 - CO2 emission estimates (million tons/tons per capita).

The third component has a higher positive loadings for I4 - Sex ratio (m per 100 f, 2017); I9 - Economy: Industry (% of GVA); I16 - Labour force participation (male %); I31 - International migrant stock (% of total pop.) that are in the center of correlation plot and are not correlated with PC1 and PC2. This means that the indicators in the countries dataset are quite diverse and should be covered by more than two dimensions. Moreover, PCA results are consistent with the correlation plot analysis: positively correlated variables point in the same direction and explain the same component. For example, indicators from the first group are positively correlated with PC1 and the ones from the second group are negatively correlated with it. For indicator groups description, please see section 3.1.

As a next step, we plot all the countries against the two main components to study similarities and differences between them and determine whether countries form groups (Appendix B, Figure 16). We can notice that the countries with the same socio-economic-environmental profile are grouped together, for example European countries with developed economies such as Belgium, The Netherlands, the UK, Germany, Italy, France have a large negative score for the PC1 and a small score for the PC2, whereas, developing countries such as Chad, Tanzania, Liberia, Madagascar have a high score for the PC1 and are on the opposite side of the plot. Countries with transitive economies (Eastern Europe countries) and some South American countries are in the middle of the plot. Indeed, the analysis of the variables-components relation showed that the indicators relevant to the developed economies are negatively correlated with the PC1. The color indicates the quality of representation of each sample by the principal components.

We also studied the relation between principal components and countries development level (Appendix B, Figure 17). We can see that the 'economies in transition' class is overlapping with the 'developed' and 'developing' when all the observations are plotted in two dimensions. For a more detailed analysis we can sample the region (Europe or Asia, for example) and perform the visualization over the sample (may be considered as a future work).

3.3 K-Means Clustering

Clustering analysis is one of the main unsupervised methods to partition, group the data and find the hidden patterns in it. In this paper we explore the standard k-means and hierarchical clustering and compare their results.

K-means is a numerical, non-deterministic, iterative method. The algorithm consists of two separate phases: 1) selecting k centers randomly with value k (number of clusters) fixed in advance 2) assigning each sample to the nearest center [6]. After all the samples are included in the clusters, the centroids are recalculated again. This iterative process of assigning samples to the clusters continues repeatedly until the within-cluster variances (squared Euclidean distances) are minimized. Several algorithms have been proposed in the literature to determine an optimal number of clusters: i) rule of thumb; ii) elbow method calculating within-cluster sums of squares; iii) Information Criterion approach; iv) silhouette method; v) gap statistic [6]. We explored elbow and silhouette methods to determine the number of clusters before performing k-means clustering on the dataset with 214 countries. The Euclidean distance was considered to determine the distance between the data points.

Elbow method is based on minimising the within cluster sums of squares (inter-cluster variation): we choose a number of clusters such that adding a new cluster does not improve much the total WSS. Figure 4 shows that the optimal number of clusters with an elbow method can be chosen as 4 (the bend in the plot indicates the best number).

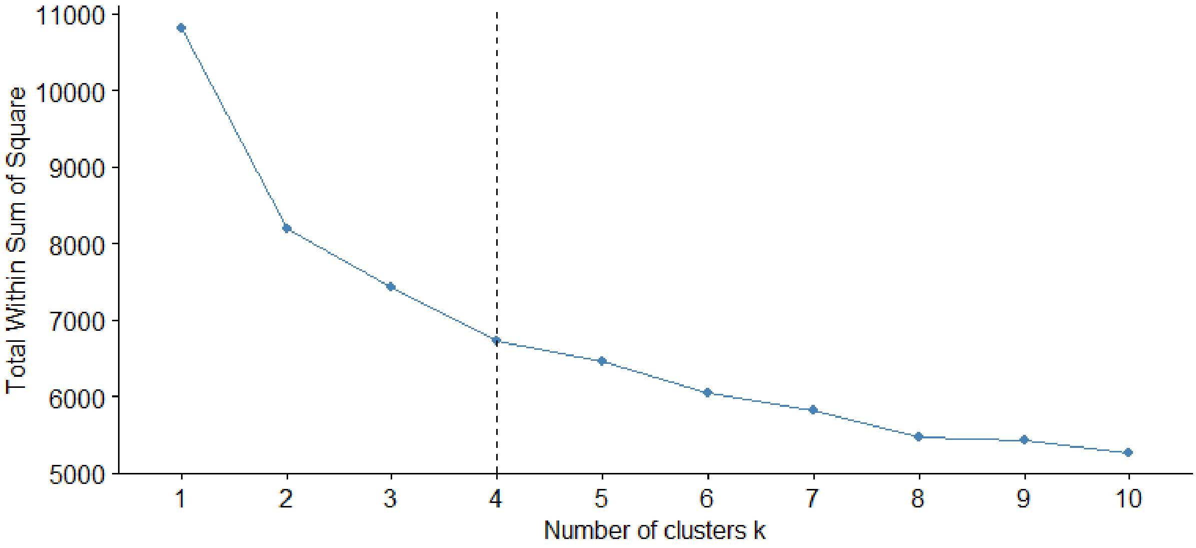


Figure 4: An optimal number of clusters for K-means clustering defined with an elbow method

The average silhouette method measures the quality of the clustering and identifies how well each data point fits into the cluster. The optimal number of clusters k is the one that maximize the average silhouette score. Figure 5 depicts an optimal number of clusters that maximizes the silhouette score is 3.

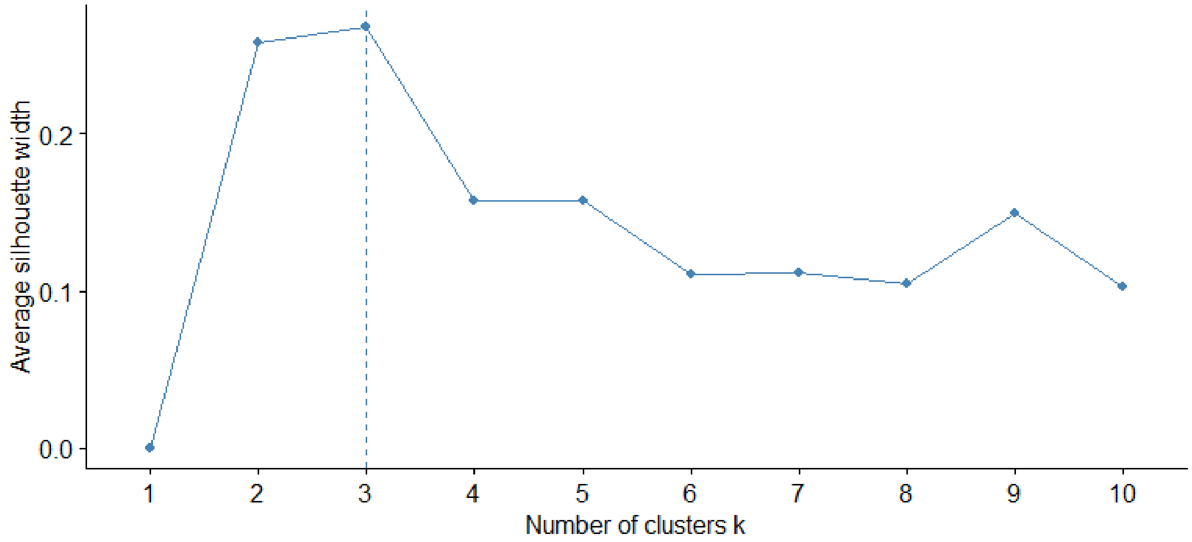


Figure 5: An optimal number of clusters for K-means clustering defined with a silhouette method

We decided to proceed with the four clusters and ran the k-means over the countries dataset. The result is plotted in the two-dimensional space using PCs (Figure 6). We can see that the outcome is quite similar to the one we got when we plotted the countries by the development level over PC1 and PC2 (Appendix B, Figure 17).

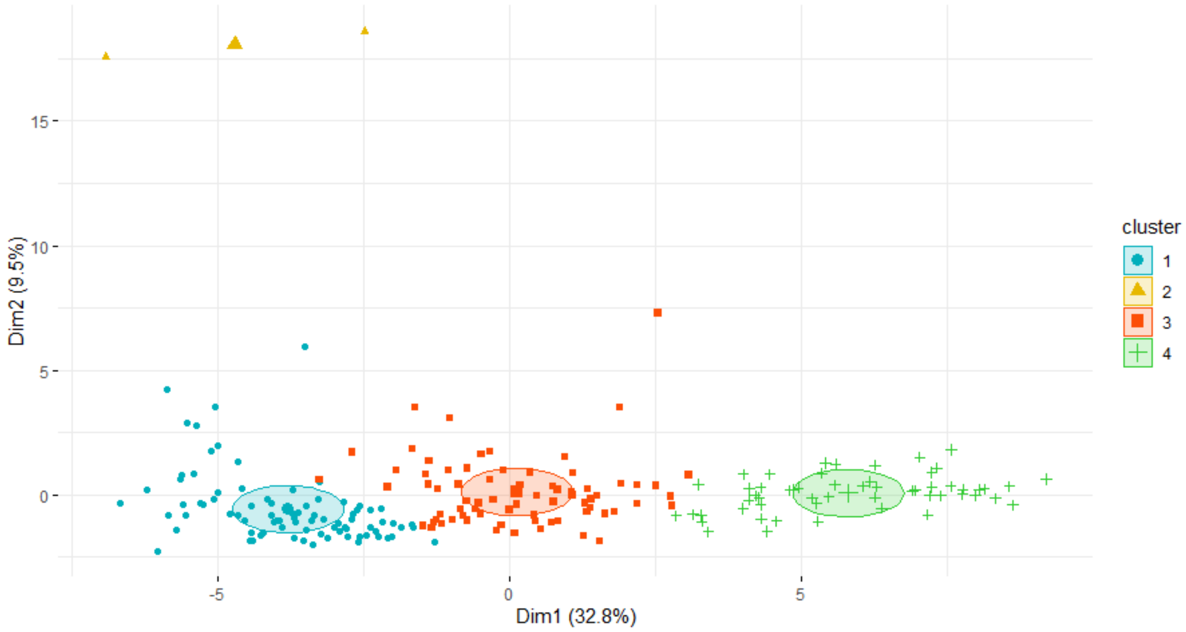


Figure 6: The result of K-means clustering of the countries. K = 4

We studied the association between clusters and countries socio-economic development classes (Figure 7). All the developed countries except for United States of America are in the Cluster 1 and most of the least developed countries are in the Cluster 4.

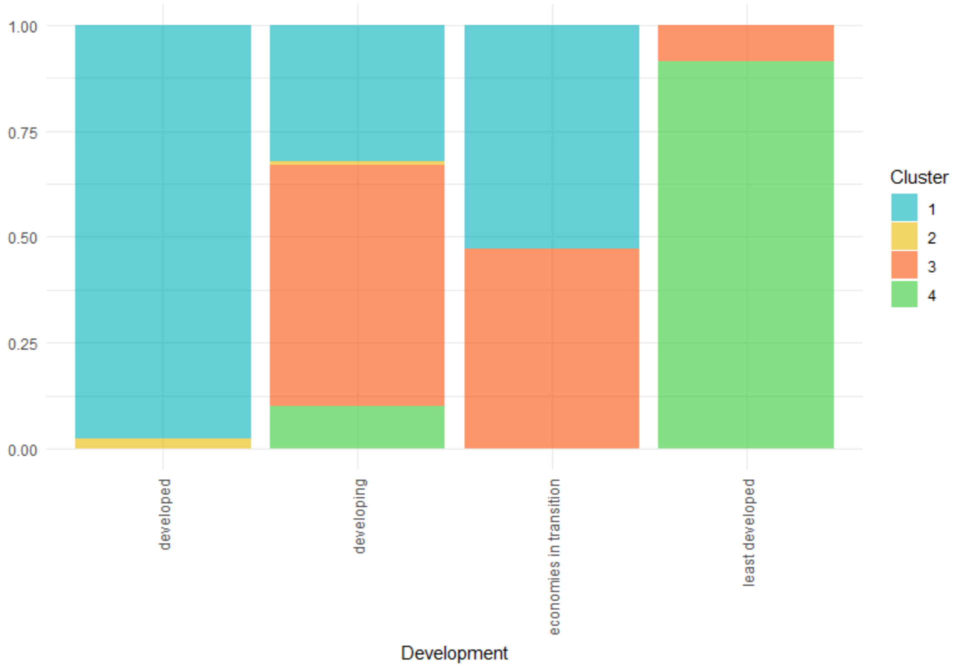


Figure 7: Association between clusters and the countries socio-economic development classes

Subsequently, we studied the relation between the clusters and the country regions and the analysis is shown in Figure 8. Cluster 1 includes all the Eastern Europe, Northern Europe, Western Europe as well as most of the Southern Europe, Caribbean and North American counties. Central America, Central Asia, Micronesia, Northern Africa, Polynesia, South-Eastern Asia, South America, South Asia and Western Asia are included into the Cluster 3. Finally, Cluster 4 is mostly composed by East African, Melanesia, Middle African, South African and West African countries.

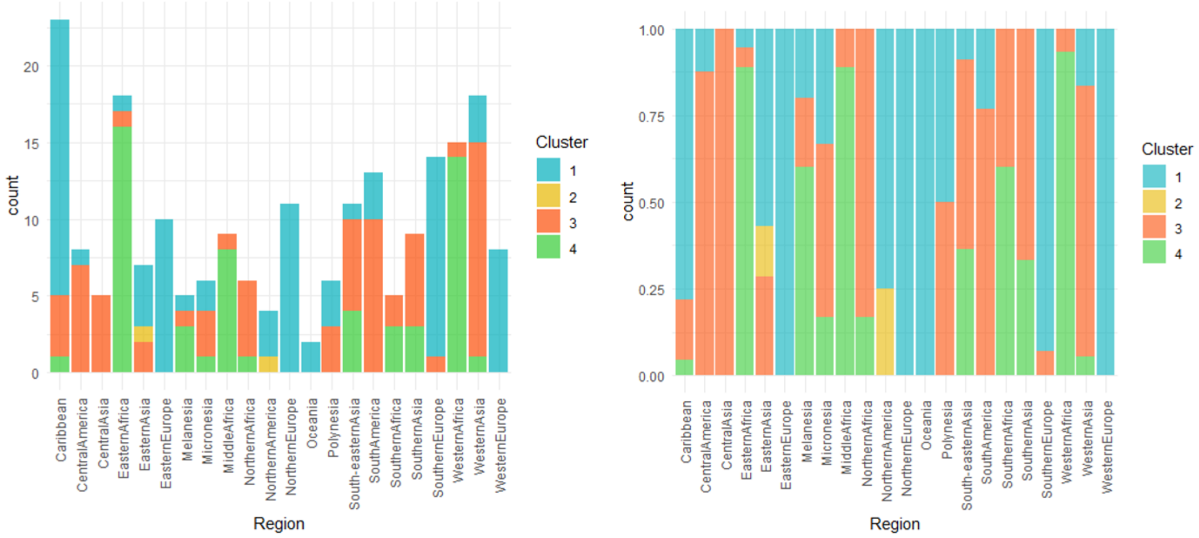


Figure 8: Association between clusters and the countries region

3.4 Hierarchical Clustering

Hierarchical clustering is based on the notion that the data points closer in space are more similar than the ones lying farther away. Hierarchical clustering models follow two approaches: 1) the algorithm starts by classifying all the data points into a separate cluster, then two nearest clusters are merged into the same group and the process repeats until only one cluster is left (agglomerative) 2) all data points are merged into a single cluster and then partitioned as the distance increases (divisive) [6].

The result of hierarchical clustering is shown as a dendrogram in Appendix B, Figure 18. The countries are merged iteratively into clusters based on their closeness measured by Euclidean distance. Ward's linkage method is used to combine the clusters. The point in the dendrogram where two clusters are merged represents the distance between two clusters in the data space. We decided to cut the dendogram at the height 0.30 with the countries combined into four clusters. We can see that The USA and China are combined in one cluster again and that most of the European countries ae also grouped together. For example, Italy is closer to France among all the counties. Then Italy-France cluster is merged with Spain and afterwards, joined with the cluster consisting of Austria, Slovenia and Portugal. The future study may include hierarchical clustering analysis over the regions applying different linkage methods and distances.

The results of the K-means and hierarchical clustering are quite consistent: both of the methods showed that the optimal number of clusters is three or four and depicted the real structure of the data. The main challenges of K-means clustering is that we need to specify number of clusters in advance, the results depend on the position of the initial centroids and the algorithm itself is sensitive to the outliers. To avoid the last one, we might either remove outliers or apply robust algorithms such as k-median or k-medoids clustering. On the contrary, hierarchical clustering does not require the number of the clusters to be set and the results are reproducible, however, it depends on the graphical representation and is not suitable for large datasets [6].

4 OUTLIERS DETECTION

An outlier is a country that is significantly different from others in terms of indicators. Outliers can be univariate, e.g. are found in a distribution of a single feature, or multivariate that are detected in a n-dimensional space. There are several approaches to detect outliers and in our research we considered descriptive statistics method using boxplots, Isolation forest and PCA inversion.

Interquartile range A boxplot helps to visualize a quantitative variable by displaying minimum, median, first, third quartiles and maximum values and shows the potential outliers using the interquartile range criterion. We plotted boxplots over the standartized data to compare all the 51 indicators on the same scale (Figure 9). We can observe that almost all the indicators have outliers. For example, I3 -Population density (per km2, 2017), I5 - GDP: Gross domestic product (million current US$) have more anomaly observations than the others. We cannot get rid of outliers or set their values to maximum or minimum merely based on the boxplot analysis. There are only 214 observations in the dataset and the presence of the ouliers may be explained by the structure of the data: each country is diverse and by removing an outlier we may lose some valuable information. Moreover, IQR is a univariate method returning outliers detected for one column and it is not suited for our dataset with 51 dimensions. Therefore, we also explored multivariate techniques such as Isolation Forest and PCA inversion.

Isolation Forest Isolation Forest is a tree-based algorithm detecting anomalies that are i) the minority consisting of fewer instances and ii) have attribute-values that are very different from those of normal instances [8]. We ran the Isolation Forest algorithm implemented by solitude package. Number of observations to build a tree in the forest was set to 200 and they were chosen with replacement; number of trees - to 500. Each observation received a score from 0 to 1, where 1 is close to being an outlier. We specified a threshold for anomalies at 0.60 which is a 85-percentile of an anomaly score. For instance, China, United States of

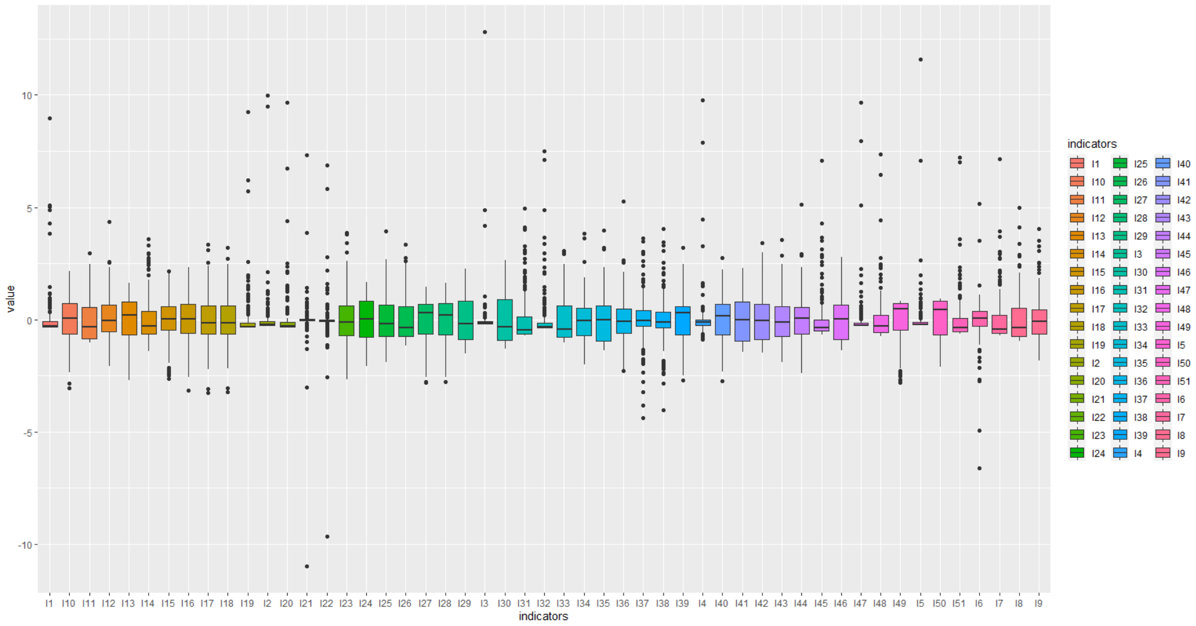


Figure 9: Potential outliers identified with the interquartile range criterion

America, Chad, Russian Federation, Hong Kong SAR are among the outliers identified by Isolation Forest algorithm. These are the countries that were located in the margins of a 2-dimenstional PCA plot (Appendix B, Figure 16). A number of countries that are indicated as outliers highly depends on the threshold specified. In this case, we set up a higher threshold to compare Isolation Forest results with the PCA inversion and then combine the outcomes.

PCA inversion To get consistent results, we decided to perform PCA inversion to find the outliers using reconstruction error. For the analysis we consider 12 components which describe almost 80% of the variance in the data (we assume that remaining 20% is noise) and then reconstruct the original data given these PCs. The reconstructed data is similar, but not the same as the original one. We calculated the Mean Squared Error between the original data and the reconstructed data and the observations with the highest error were considered as anomalies. However, one of the major concern of this method is that since PCA relies on the sample covariance matrix, the variance explained by the principal components can be distorted by the outliers. However, it is still a good method to find the multivariate outliers and is commonly used in anomaly detection [11]. We set the threshold as a 85-percentile of the reconstruction error score. The top outliers are Russian Federation, Colombia, India.

Finally, we combined the results of Isolation Forest and PCA inversion: 14 countries (0.07 of the dataset) were identified by both of the algorithms as anomalies (Table 1) and we considered them for the future analysis.

Country	**Region**	**Development**	**Iforest anomaly score**	**PCA reconstruction loss**
Russian Federation	EasternEurope	economies in transition	0.6182	1.2975
Colombia	SouthAmerica	developing	0.6050	1.0553
China, Macao SAR	EasternAsia	developing	0.6171	0.7633
Angola	MiddleAfrica	least developed	0.6005	0.5935
Syrian Arab Republic	WesternAsia	developing	0.6112	0.5903
Timor-Leste	South-easternAsia	least developed	0.6103	0.5197
United Arab Emirates	WesternAsia	developing	0.6061	0.4524
Bermuda	NorthernAmerica	developing	0.6063	0.4169
Cuba	Caribbean	developing	0.6069	0.4000
Qatar	WesternAsia	developing	0.6188	0.3921
Equatorial Guinea	MiddleAfrica	developing	0.6119	0.3727
Nigeria	WesternAfrica	developing	0.6041	0.3275
Greece	SouthernEurope	developed	0.6267	0.3211
Madagascar	EasternAfrica	least developed	0.6068	0.3072

Table 1: Outliers identified with both Isolation Forest and PCA inversion methods

5 SUPERVISED LEARNING: CLASSIFICATION

The objective of the supervised learning in this project is to classify all the countries in 4 categories: developed, economies in transition, developing and the least developed countries (based on the United Nations classification). The labels were manually extracted from The World Economic Situation and Prospects report produced by the United Nations Department of Economic and Social Affairs, and added to the dataset. The classes are imbalanced and have the following distribution: developed (41 countries), developing (109), economies in transition (17), least developed countries (47). We implemented Random Forest, Multinomial Logistic Regression, K-NN and a simple ANN classification models and studied their performance on the training and test sets.

Classification models with a low bias and a high variance can not generalize while models with a high bias and a low variance are oversimplified and underfit the data. We are dealing with a very small dataset in this project and in fact these types of datasets are very common in a real world setting. Classifiers trained on a small number of observations might overfit and produce inaccurate results. However, there are various techniques to avoid overfitting and improving the prediction power:

1. Choose a simple model with a smaller number of parameters (logistic regression with regularization, tree-based models with a limited maximum depth).

2. Remove outliers from data.

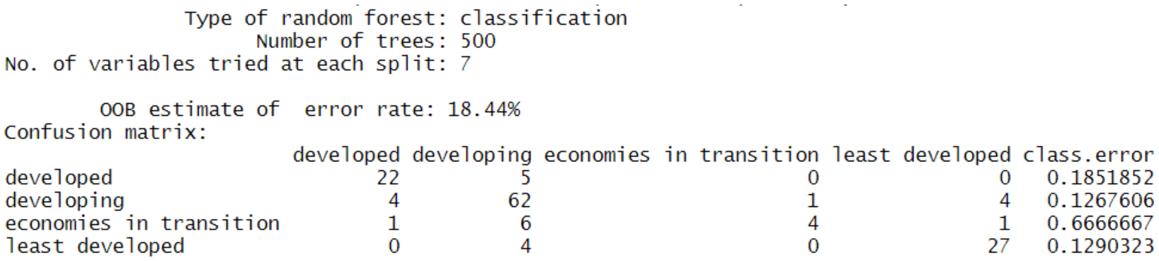
3. Select only relevant features. Overfitting might occur when there are few observations and a large number of features.

4. Extend the dataset if it is imbalanced by augmenting the dataset with synthetic samples or adding more information from other sources.

5.1 Random Forest

Random Forest classifier belongs to the Ensemble models and is used to improve the predictive performance of Decision Trees. Instead of training a single complex model that tends to overfit, Random Forest generates several models (trees) on different bootstrapped samples from training data and then reduces the variance in the trees by averaging them. The idea is to create many trees in the way correlation between them is as small as possible. A random subset of predictors is sampled each time during the training split [9].

As a first step, we run the Random Forest Classifier over the dataset containing 214 countries and 51 standardized indicators. To evaluate the model we split the dataset into the training set (p = 0.65) and the testing set (p = 0.35) and set the random seed for results reproducibility. With the default parameters (number of trees = 500 and the mtry = sqrt(p) = 7), we get the Out of Bag error (OOB) equal to 18.44%. The number of trees should not be set to a very small number, to ensure that every observation is predicted at least few times. Figure 10 shows the performance of the classifier on the training set. We can see that 'developed', 'developing' and 'least developed' classes have a pretty low error, whereas countries with economies in transition tend to fall into 'developing' category. This might be explained by the fact that we have less observations for this class and also by the nature of the data: countries with economies in transition have features of both developed and developing countries.



```
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 7

        OOB estimate of  error rate: 18.44%
Confusion matrix:
                        developed developing economies in transition least developed class.error
developed                      22          5                       0               0   0.1851852
developing                      4         62                       1               4   0.1267606
economies in transition         1          6                       4               1   0.6666667
least developed                 0          4                       0              27   0.1290323
```

Figure 10: Random Forest Classifier setting and performance on the training set

Figure 11 shows how the OOB error (black) and errors for each class change if we increase the number of trees: 500 trees proves to be a good number to run the classifier.

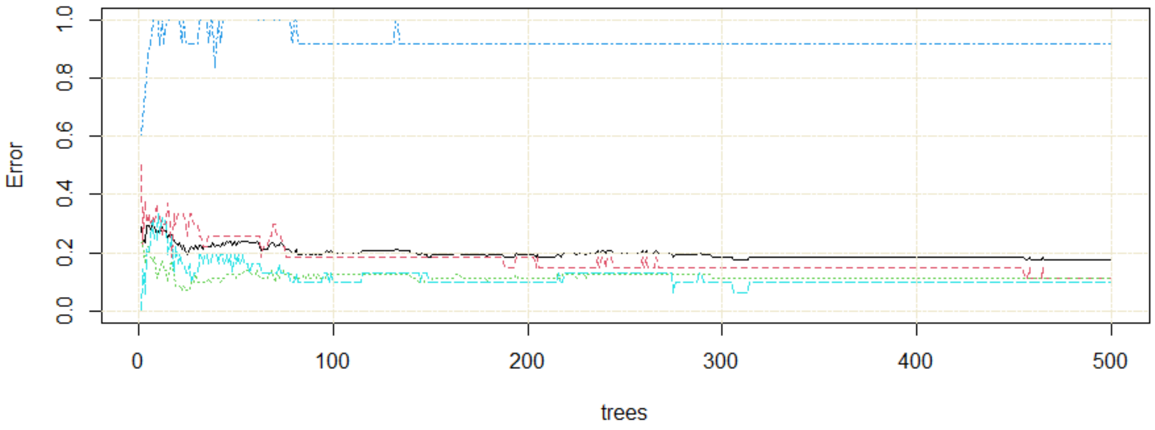
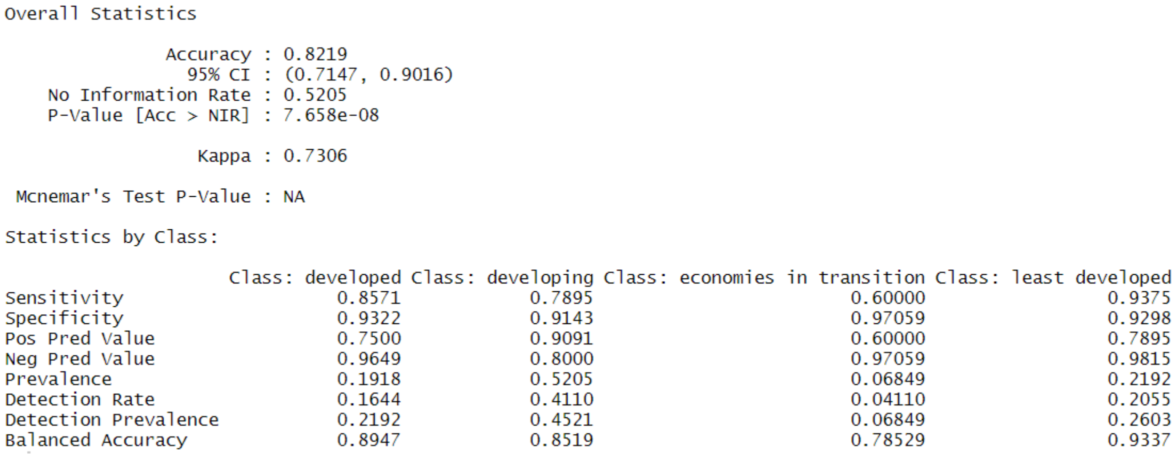


Figure 11: OOB error (black) and errors for each class against the number of trees

The random forest model achieves accuracy of 0.82 on a testing set, with the sensitivity and specificity for different classes shown in Figure 12. However, the result depends on the random nature of the train-test split and is might not be representative as the size of the test set is very small.



```
Overall Statistics

               Accuracy : 0.8219
                 95% CI : (0.7147, 0.9016)
    No Information Rate : 0.5205
    P-Value [Acc > NIR] : 7.658e-08

                  Kappa : 0.7306

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: developed Class: developing Class: economies in transition Class: least developed
Sensitivity                    0.8571            0.7895                        0.60000                 0.9375
Specificity                    0.9322            0.9143                        0.97059                 0.9298
Pos Pred Value                 0.7500            0.9091                        0.60000                 0.7895
Neg Pred Value                 0.9649            0.8000                        0.97059                 0.9815
Prevalence                     0.1918            0.5205                        0.06849                 0.2192
Detection Rate                 0.1644            0.4110                        0.04110                 0.2055
Detection Prevalence           0.2192            0.4521                        0.06849                 0.2603
Balanced Accuracy              0.8947            0.8519                        0.78529                 0.9337
```

Figure 12: Random Forest Classifier performance on the testing set

We also explored how much each indicator contributed to the model using feature importance plot (Figure 13). The most important variables are I35 - Health: Physicians (per 1000 pop.), I7 - GDP per capita (current US$), I48 - Energy production, primary (Petajoules), I33 - Infant mortality rate (per 1000 live births). MeanDecreaseAccuracy gives an estimate of the loss in prediction when the indicator is omitted from the training set. MeanDecreaseGini measures how important that feature is to split the data correctly. Random Forest model is not affected by multicollinearity in variables in terms of prediction. However, feature importance score might be influenced: the overall importance of correlated variables is reduced when they are included together in the split [9].

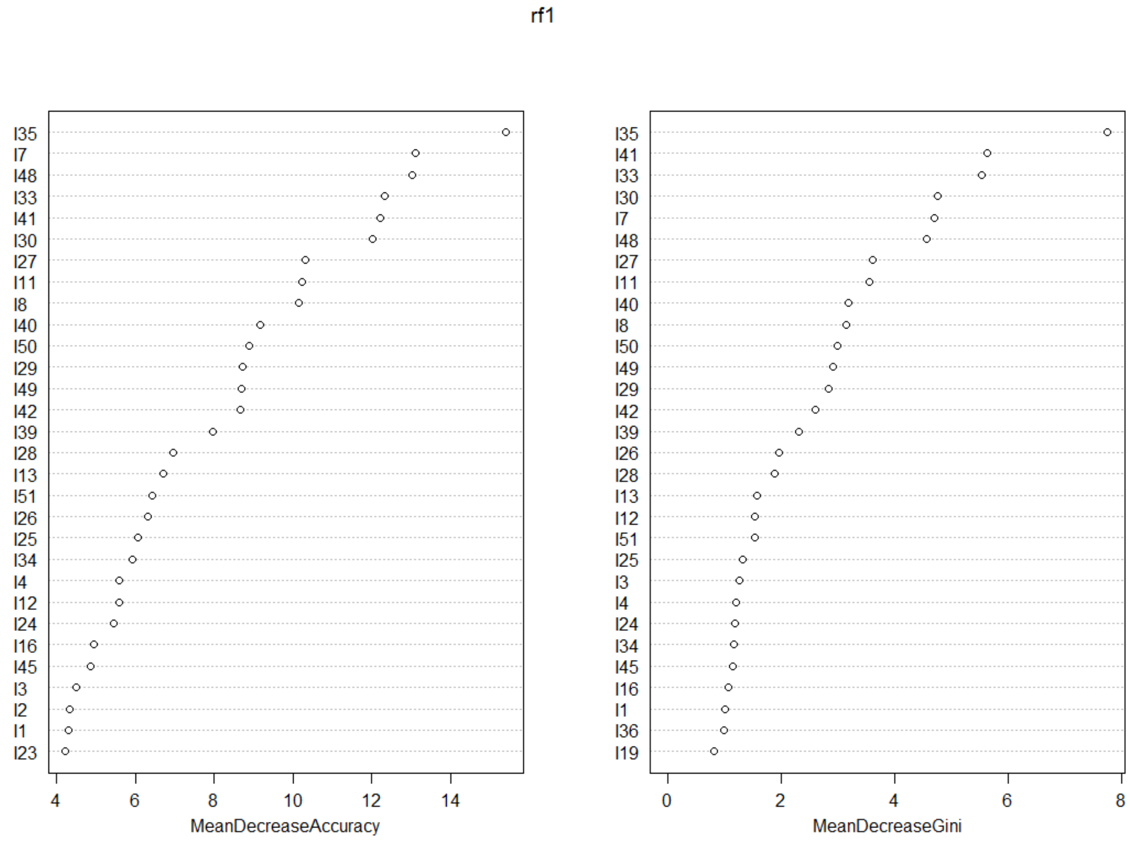


Figure 13: Importance of the indicators in Random Forest model with the original parameters

To eliminate the effect of multicollinearity, we run the Random Forest model with 1) all the indicators and 2) 12 principal components on the entire set of data (214 countries). The classifier trained using all the indicators resulted in 16.82% OOB estimate of error rate, whereas the model with the principal components resulted in 23.36% OBB error. The first model identified I30, I48, I7, I33, I35 as the most important features (the same we got while training the model on the part of the dataset, however, the order has changed a bit). Mean-while, the result of the second model is presented in Figure 14. We can see that the first principal component is more important for the model than the others.

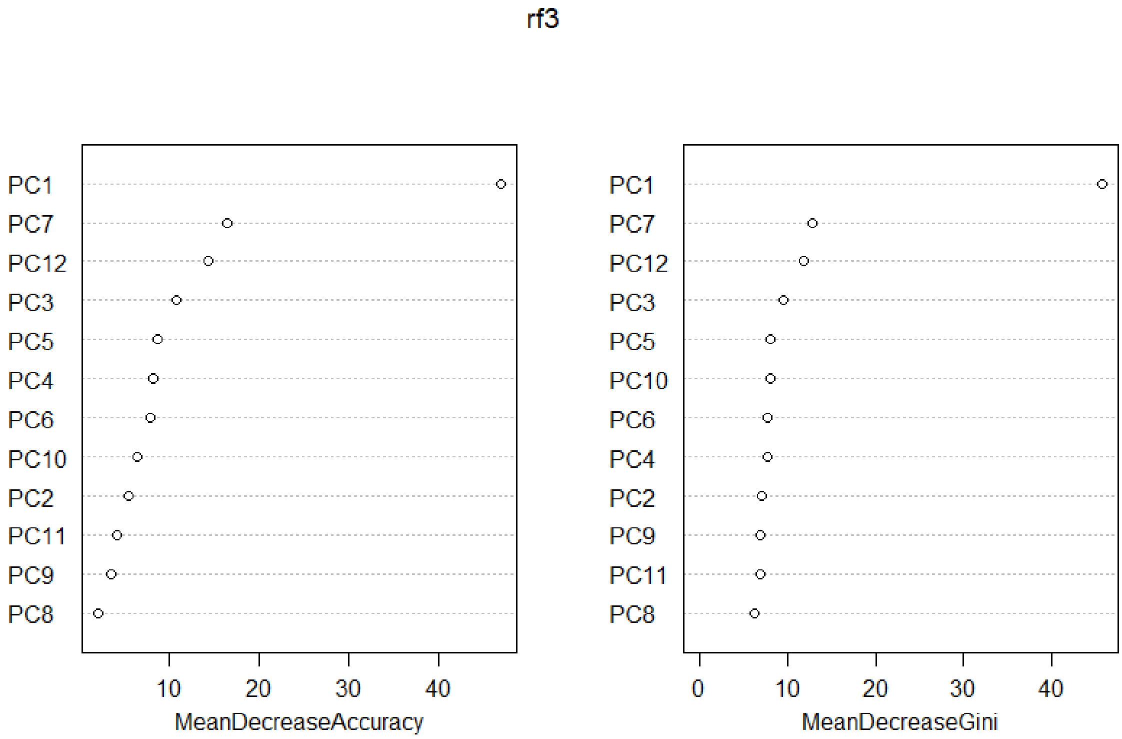


Figure 14: Importance of the PCs in the Random Forest model with 12 principal components

We also ran the model on the dataset removing the outliers detected in Section 4. OOB error was 17.5% and the important features did not change, i.e. that the model performance and interpretability did not improve.

5.2 Comparing Random Forest with other models

Finally, we compared performance of the Random Forest with the other multiclass classification models such as Multinomial Logistic Regression, K-Nearest Neighbors algorithm and a simple Neural Network. The evaluation is done using both training and test sets with all the indicators and observations included (Table 2). We used the accuracy score (training set OOB score for Random Forest) to compare the results.

Model	Train set accuracy	Test set accuracy
Random Forest	0.82	0.82
Multinomial Logistic Regression	1.00	0.64
K-NN (k = 8)	0.78	0.81
Neural Network	0.96	0.66

Table 2: Comparing different models performance on the train and test sets

1. **Multinomial Logistic Regression** is used to predict the categorical target variable with more than 2 classes. The log odds of the outcomes are modeled as a linear combination of the features. We ran the model with the default parameters. We can see that the classifier is overfitting and some regularization should be applied to overcome this problem. Moreover, we need to get rid of multicollinearity and outliers to improve performance and data interpretability.

2. **K-NN** is a non-parametric algorithm that assigns a new data point to the closest class, given the k number of the neighbors. It works similarly to the K-means clustering and the distance between two points in our case is measured by Euclidean Distance. We took k = 8, however, we can apply different techniques such as elbow method to find the better number of neighbours. The method shows high accuracy in both training and test sets and does not overfit. However, this classifier is also affected by the imbalanced classes, it is outlier sensitive and lacks interpretability in terms of features.

3. **Neural network**: we used a neuralnet package to run a simple neural network [5]. Softmax activation function was applied to normalize the output of a network to a probability distribution over predicted classes.Two hidden layers with ten and four neurons (number of outcome variables) were used. The algorithm type is set to resilient backpropagation with weight backtracking. As an output we got the probability of each neuron then predicted the class with the highest probability value. We can see that NN overfits as well and this problem might be solved by optimizing the hyperparam-

eters and extending both training and test sets as NN requires much more data than traditional statistical learning algorithms.

We can conclude that Random Forest outperforms other models in terms of prediction accuracy on the test set. Although the algorithm has some limitations, it is one of the best classifiers for dealing with multiclass classification problem in a small dataset. Random Forest handles imbalanced data, it is a low-biased and moderate-variance model (overfitting problem is addressed by averaging the results over the trees). Moreover, it is based on a bootstrap resampling and training decision trees on the samples, which helps to solve the problem of a data scarcity in some way.

6 CONCLUSION

We analyzed the United Nations countries dataset containing various socio-economic indicators. We applied different unsupervised and supervised learning techniques to study countries similarity and also to classify them according to the development level.

We can conclude than the results achieved by K-means and Hierarchical Clustering are consistent and show that there are three or four natural clusters within the dataset. This reflects the real structure of the data with the four categories provided by the UN: developed, developing, the least developed countries and economies in transition. However, both K-means and Hierarchical Clustering have some drawbacks, for instance, K-means is sensitive to the outliers. To avoid the effect of outliers on the grouping, K-medoids or K-median clustering methods are proposed as a part of future exploration. Moreover, studying the counties by geographic regions might be valuable to understand how the nations are different within one region and discover the reasons behind it.

As for classification, Random Forest model outperformed other models in terms of interpretability and prediction accuracy on the test set. This algorithm is suitable for dealing with a classification problem on a small dataset.

Finally, we discussed some of the statistical learning analysis challenges such as class imbalance, overfitting, presence of outliers, data scarcity and various techniques that may help to overcome them. One of the possible extensions of the project is augmenting the dataset by adding more observations and indicators to it.

Bibliography

[1] Kaggle: UN countries dataset 2017. `https://www.kaggle.com/sudalairajkumar/undata-country-profiles`. [Online; Accessed: 2021-11-15].

[2] UNdata, international statistical database. `http://data.un.org/Host.aspx?Content=About`. [Online; Accessed: 2021-11-15].

[3] Boston Consulting Group. Comparing Socioeconomic Development Across Nations. `https://www.bcg.com/publications/2012/public-sector-globalization-comparing-socioeconomic-development`, 2012. Online; Accessed: 2021-11-15.

[4] UN ESCAP, UN ECA, UN ECE, UN ESCWA, UN ECLAC, et al. World economic situation and prospects 2017. 2017.

[5] Stefan Fritsch, Frauke Guenther, and Maintainer Frauke Guenther. Package 'neuralnet'. *Training of Neural Networks*, 2019.

[6] Paulene Govender and Venkataraman Sivakumar. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 11(1):40–56, 2020.

[7] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[8] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 6(1):1–39, 2012.

[9] Victor Francisco Rodriguez-Galiano, Bardan Ghimire, John Rogan, Mario Chica-Olmo, and Juan Pedro Rigol-Sanchez. An assessment of the effectiveness of a random forest classifier for land-cover classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 67:93–104, 2012.

[10] Daniel J Stekhoven. missforest: Nonparametric missing value imputation using random forest. *Astrophysics Source Code Library*, pages ascl–1505, 2015.

[11] Bo Zong, Qi Song, Martin Renqiang Min, Wei Cheng, Cristian Lumezanu, Daeki Cho, and Haifeng Chen. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In *International conference on learning representations*, 2018.

APPENDIX A

Indicators	**I**	**Mean**	**Variance**
Surface area (km2)	I1	631353.6	1843929.43
Population in thousands (2017)	I2	35211.15	137899.13
Population density (per km2, 2017)	I3	334.07	1602.89
Sex ratio (m per 100 f, 2017)	I4	101.49	20.45
GDP: Gross domestic product (million current US$)	I5	349854.26	1529954.92
GDP growth rate (annual %, const. 2005 prices)	I6	2.48	4.64
GDP per capita (current US$)	I7	15287.26	21548.02
Economy: Agriculture (% of GVA)	I8	11.21	11.99
Economy: Industry (% of GVA)	I9	27.62	12.91
Economy: Services and other activity (% of GVA)	I10	61.23	15.29
Employment: Agriculture (% of employed)	I11	23.6	22.83
Employment: Industry (% of employed)	I12	18.73	8.09
Employment: Services (% of employed)	I13	57.28	18.78
Unemployment (% of labour force)	I14	9.01	6.27
Labour force participation (female %)	I15	52.76	15.3
Labour force participation (male %)	I16	73.44	8.57
Agricultural production index (2004-2006=100)	I17	115.99	24.97
Food production index (2004-2006=100)	I18	116.77	25.57
International trade: Exports (million US$)	I19	74650.75	221944.67
International trade: Imports (million US$)	I20	75726.27	225075.67
International trade: Balance (million US$)	I21	-736.33	72679.84
Balance of payments, current account (million US$)	I22	1026.72	48140.32
Population growth rate (average annual %)	I23	1.34	1.36
Urban population (% of total population)	I24	58.96	24.4
Urban population growth rate (average annual %)	I25	1.93	1.67
Fertility rate, total (live births per woman)	I26	2.81	1.39
Life expectancy at birth (females, years)	I27	74.1	8.36
Life expectancy at birth (males, years)	I28	69.13	7.67
Population age distribution (0-14 years, %)	I29	27.15	10.27
Population age distribution (60+ years, %)	I30	12.67	7.83
International migrant stock (% of total pop.)	I31	11.22	15.66
Refugees and others of concern to UNHCR (in thousands)	I32	317.9	907.89
Infant mortality rate (per 1000 live births	I33	24.51	22.76
Health: Total expenditure (% of GDP)	I34	6.86	2.66
Health: Physicians (per 1000 pop.)	I35	1.93	1.4
Education: Government expenditure (% of GDP)	I36	4.59	1.5
Education: Primary gross enrol. ratio (f per 100 pop.)	I37	102.42	12.88
Education: Primary gross enrol. ratio (m per 100 pop.)	I38	104.89	12.87
Education: Secondary gross enrol. ratio (f per 100 pop.)	I39	84.4	28.87
Education: Secondary gross enrol. ratio (m per 100 pop.)	I40	84.06	26.13
Education: Tertiary gross enrol. ratio (f per 100 pop.)	I41	43.4	30.06
Education: Tertiary gross enrol. ratio (m per 100 pop.)	I42	34.78	22.85
Seats held by women in national parliaments %	I43	21.17	11.27
Mobile-cellular subscriptions (per 100 inhabitants)	I44	108.31	42.36
Individuals using the Internet (per 100 inhabitants)	I45	210.5	303.86
Threatened species (number)	I46	32.55	23.88
CO2 emission estimates (million tons/tons per capita)	I47	2699.5	10250.39
Energy production, primary (Petajoules)	I48	87.9	117.88
Pop. using improved drinking water (urban, %)	I49	81.22	23.09
Pop. using improved drinking water, rural %)	I50	70.07	31.81
Pop. using improved sanitation facilities (urban/rural, %)	I51	5.25	8.2

Table 3: Descriptive statistics and indicators encoding

PC	Eigenvalue	Variance percentage	Cumulative Variance Percentage
Dim.1	16.7087	32.7622	32.7622
Dim.2	4.8317	9.474	42.2362
Dim.3	3.6102	7.0788	49.315
Dim.4	2.5058	4.9133	54.2283
Dim.5	2.4781	4.8591	59.0874
Dim.6	2.1483	4.2124	63.2998
Dim.7	1.6692	3.2728	66.5726
Dim.8	1.5158	2.9721	69.5448
Dim.9	1.4004	2.7459	72.2907
Dim.10	1.2441	2.4395	74.7302
Dim.11	1.0936	2.1444	76.8746
Dim.12	1.0816	2.1208	78.9954
Dim.13	0.9431	1.8492	80.8446
Dim.14	0.9016	1.7679	82.6125
Dim.15	0.8327	1.6327	84.2452
Dim.16	0.7536	1.4777	85.7229
Dim.17	0.6152	1.2063	86.9291
Dim.18	0.6036	1.1834	88.1126
Dim.19	0.5271	1.0335	89.1461
Dim.20	0.5171	1.014	90.16
Dim.21	0.4926	0.966	91.126
Dim.22	0.4661	0.9139	92.04
Dim.23	0.4118	0.8075	92.8475
Dim.24	0.3804	0.7459	93.5934
Dim.25	0.3516	0.6893	94.2827
Dim.26	0.3369	0.6606	94.9433
Dim.27	0.2983	0.5849	95.5281
Dim.28	0.2903	0.5692	96.0973
Dim.29	0.2529	0.4959	96.5932
Dim.30	0.2306	0.4522	97.0454
Dim.31	0.2111	0.4138	97.4592
Dim.32	0.1991	0.3904	97.8496
Dim.33	0.1749	0.343	98.1926
Dim.34	0.1491	0.2923	98.4848
Dim.35	0.1416	0.2776	98.7624
Dim.36	0.1157	0.2269	98.9894
Dim.37	0.1022	0.2005	99.1898
Dim.38	0.0931	0.1826	99.3725
Dim.39	0.0771	0.1512	99.5237
Dim.40	0.0493	0.0967	99.6204
Dim.41	0.0473	0.0927	99.7131
Dim.42	0.0397	0.0779	99.791
Dim.43	0.0363	0.0713	99.8622
Dim.44	0.0236	0.0463	99.9085
Dim.45	0.0157	0.0307	99.9392
Dim.46	0.0137	0.0269	99.9661
Dim.47	0.0101	0.0198	99.9859
Dim.48	0.0038	0.0075	99.9933
Dim.49	0.0029	0.0058	99.9991
Dim.50	0.0004	0.0009	99.9999
Dim.51	0	0.0001	100

Table 4: PCA eigenvectors, variance explained, cumulative variance percentage explained by the principal components

	PC1	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
I1	-0.02246	0.31962	-0.04491	0.01509	-0.05063	0.02835	-0.02606	0.05109	-0.06364	0.03336
I2	0.00252	0.33635	-0.03586	0.11126	0.02007	-0.1771	-0.12761	-0.01324	0.06334	-0.15152
I3	-0.04185	-0.03369	0.05764	-0.2317	-0.07005	-0.22241	-0.10594	0.43503	-0.17889	0.04983
I4	-0.0008	0.03919	0.37311	-0.12033	-0.11989	-0.00782	-0.0592	-0.30011	0.03642	-0.12313
I5	-0.05519	0.39153	-0.11801	-0.10993	-0.11448	0.05699	-0.02103	-0.02873	0.02876	0.00932
I6	0.02744	0.03559	0.00169	0.02721	0.10028	-0.00178	0.12136	-0.36795	0.46941	0.02394
I7	-0.14752	0.00909	0.1034	-0.25648	-0.03321	-0.13199	0.12485	-0.02114	-0.04489	0.19816
I8	0.19136	0.00944	-0.12026	-0.04878	0.04182	-0.04444	0.09793	0.00411	-0.16247	-0.2046
I9	0.0184	0.0966	0.28621	0.31223	-0.00231	0.13074	-0.04051	-0.01052	0.01733	0.32984
I10	-0.16634	-0.09018	-0.14784	-0.22621	-0.03176	-0.07721	-0.04294	0.00918	0.11269	-0.11741
I11	0.21626	0.03255	-0.08933	-0.03041	0.10114	-0.04254	0.07516	0.03867	-0.0003	-0.01332
I12	-0.12442	0.04688	0.2063	0.27129	-0.10838	0.08323	0.01909	-0.03772	-0.05716	-0.13309
I13	-0.20593	-0.0536	0.0244	-0.07715	-0.07476	0.01542	-0.09016	-0.02271	0.02319	0.06597
I14	-0.00764	-0.12154	-0.1251	0.13099	-0.16567	0.12141	-0.34664	-0.14401	0.03114	0.10722
I15	0.05518	0.02533	-0.05484	-0.27582	0.28599	-0.16166	0.24498	0.09353	0.18501	0.18615
I16	0.11602	0.07431	0.21965	-0.1507	0.05675	-0.09422	0.1668	0.11185	0.21603	-0.12908
I17	0.10862	0.12348	0.18343	0.06838	0.22378	0.15899	0.26544	0.06422	-0.23731	-0.22107
I18	0.11134	0.12299	0.17958	0.06969	0.21869	0.15959	0.26884	0.06359	-0.24498	-0.22422
I19	-0.07936	0.37112	-0.0781	-0.00491	-0.00865	-0.19826	-0.05843	-0.08631	-0.04746	0.05375
I20	-0.08032	0.37776	-0.10563	-0.0898	-0.07643	-0.03344	-0.01785	-0.04127	-0.01086	0.05879
I21	0.00533	-0.03761	0.08837	0.26419	0.20923	-0.50235	-0.12276	-0.13576	-0.11191	-0.01418
I22	-0.01224	-0.02449	0.05602	0.23372	0.20929	-0.53984	-0.10302	-0.12692	-0.11567	-0.00234
I23	0.14259	0.04096	0.27172	-0.16568	-0.10508	0.01747	-0.07184	-0.13005	-0.08942	0.1102
I24	-0.15952	0.0195	0.12192	-0.06414	-0.05756	0.02015	-0.00929	0.01635	-0.15394	0.11378
I25	0.18098	0.07596	0.17707	-0.13438	-0.03303	-0.03319	-0.01851	-0.08394	-0.03471	0.07095
I26	0.21585	-0.0013	-0.01482	-0.0526	-0.08242	0.01486	-0.03012	-0.0948	-0.11201	0.16733
I27	-0.22681	-0.00997	0.0223	-0.01114	0.0585	-0.00215	0.03325	0.02206	0.04686	-0.1101
I28	-0.2172	0.00183	0.04748	-0.0617	0.04296	-0.02173	0.03158	-0.03849	0.02687	-0.11232
I29	0.22627	-0.00475	-0.02407	-0.0206	-0.05441	0.05721	-0.06311	-0.05892	-0.05852	0.09501
I30	-0.19965	0.00183	-0.15803	0.05749	0.08581	-0.06388	0.14796	0.06548	-0.04286	0.0808
I31	-0.10878	-0.05384	0.27207	-0.29401	-0.11826	-0.11585	-0.0735	-0.06116	-0.00293	0.00997
I32	0.03649	0.03975	-0.00492	0.20647	-0.25456	0.03683	-0.05697	0.19683	-0.15347	0.06051
I33	0.22674	0.01815	-0.03521	-0.01166	-0.0526	-0.01553	0.01235	0.00809	-0.10568	0.12514
I34	-0.09435	0.00949	-0.27219	-0.17694	-0.04769	0.04091	0.06923	-0.23907	-0.33035	-0.07346
I35	-0.20342	-0.04984	-0.0404	0.03685	0.0056	-0.04612	0.12876	0.05358	-0.05723	0.09173
I36	-0.06018	-0.03022	-0.12642	-0.00902	0.09732	0.15608	0.07107	-0.42639	-0.19351	0.07578
I37	-0.00725	0.03785	0.04195	-0.16255	0.42433	0.17361	-0.40837	-0.01618	-0.08397	-0.0361
I38	0.03822	0.02819	0.0384	-0.14651	0.39246	0.17955	-0.40069	0.0013	-0.13904	0.01847
I39	-0.21837	-0.00482	0.03131	0.02145	0.1207	0.08228	-0.00432	-0.08305	-0.01625	-0.04866
I40	-0.21853	0.00295	0.02774	0.02921	0.11588	0.06312	0.01319	-0.04134	-0.03248	-0.0187
I41	-0.2125	0.01425	-0.03724	0.03894	0.01165	0.05367	0.07047	0.0317	-0.10998	0.04928
I42	-0.19567	0.03875	-0.06672	0.07395	0.02011	0.02528	0.05597	0.05529	-0.16735	0.06263
I43	-0.04126	0.02559	-0.08062	0.01254	0.21165	0.03275	0.21722	-0.0404	-0.114	0.51578
I44	-0.14027	-0.01285	0.20528	-0.10398	0.03387	-0.0004	-0.17206	0.20763	-0.05409	0.08221
I45	0.01054	0.25945	-0.05084	-0.00021	0.08173	0.13147	-0.04657	0.14528	0.21297	-0.05952
I46	-0.0151	-0.02476	-0.12866	-0.0179	0.21318	0.05653	-0.0893	0.0023	0.20533	0.01829
I47	-0.0386	0.4088	-0.03124	-0.01053	-0.08514	-0.04116	-0.08876	-0.03173	0.00345	-0.02013
I48	-0.11774	0.0636	0.26347	-0.10799	-0.0355	0.00989	0.17311	-0.19514	-0.06454	0.10626
I49	-0.20959	-0.02884	0.06294	0.0727	0.00277	0.06868	0.00617	0.00616	0.02501	-0.19297
I50	-0.21587	-0.04162	0.06739	0.01952	0.00005	0.04385	0.00815	-0.00489	0.03878	-0.1404
I51	0.08903	-0.0852	-0.1616	-0.21026	-0.13763	-0.13874	-0.00045	-0.19618	-0.2236	-0.26411

Table 5: PCA loadings as the coefficients of the linear combination of the original variables for 10 principal components

Appendix B

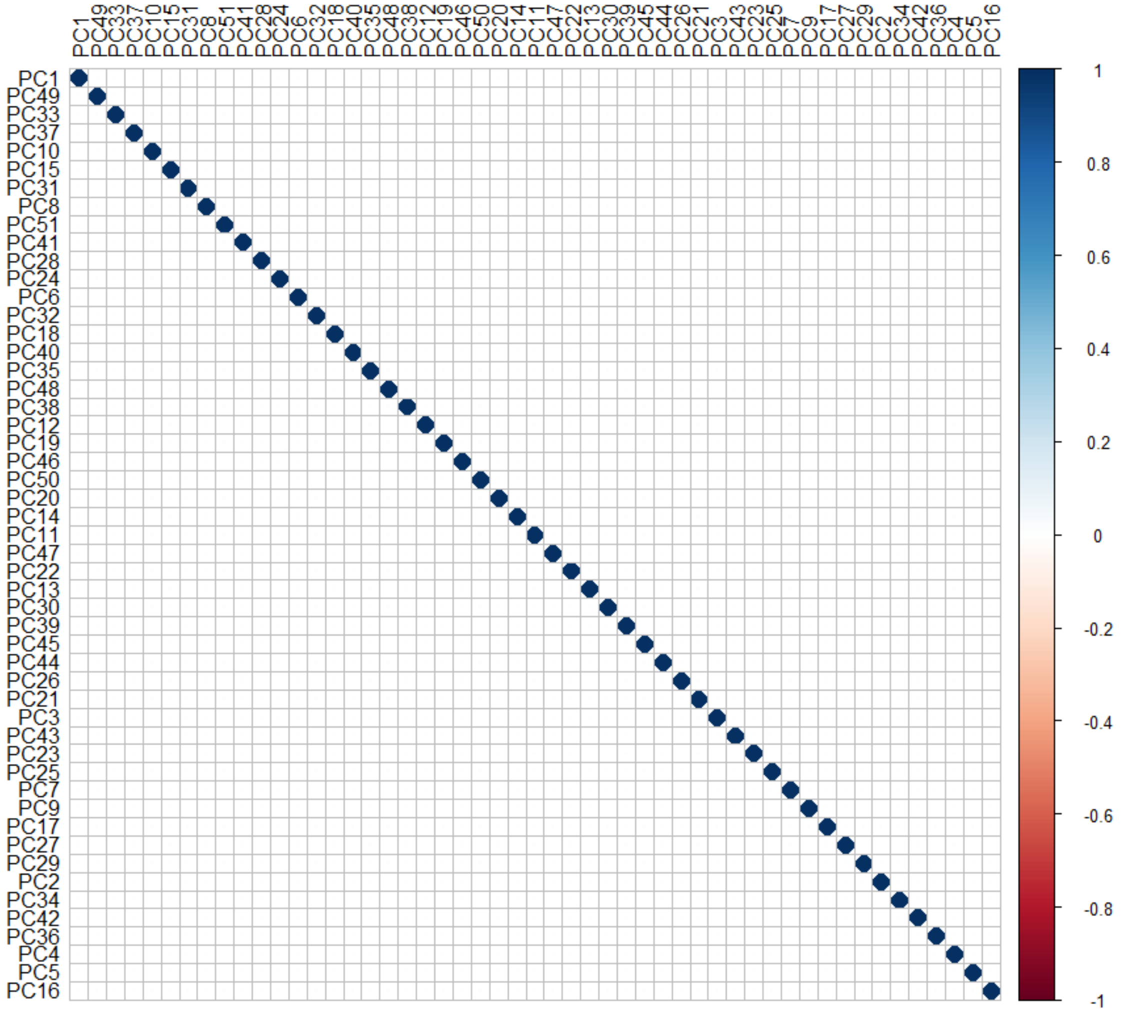


Figure 15: Correlation plot between the principal componets: 1 - strong positive correlation, -1 - strong negative correlation

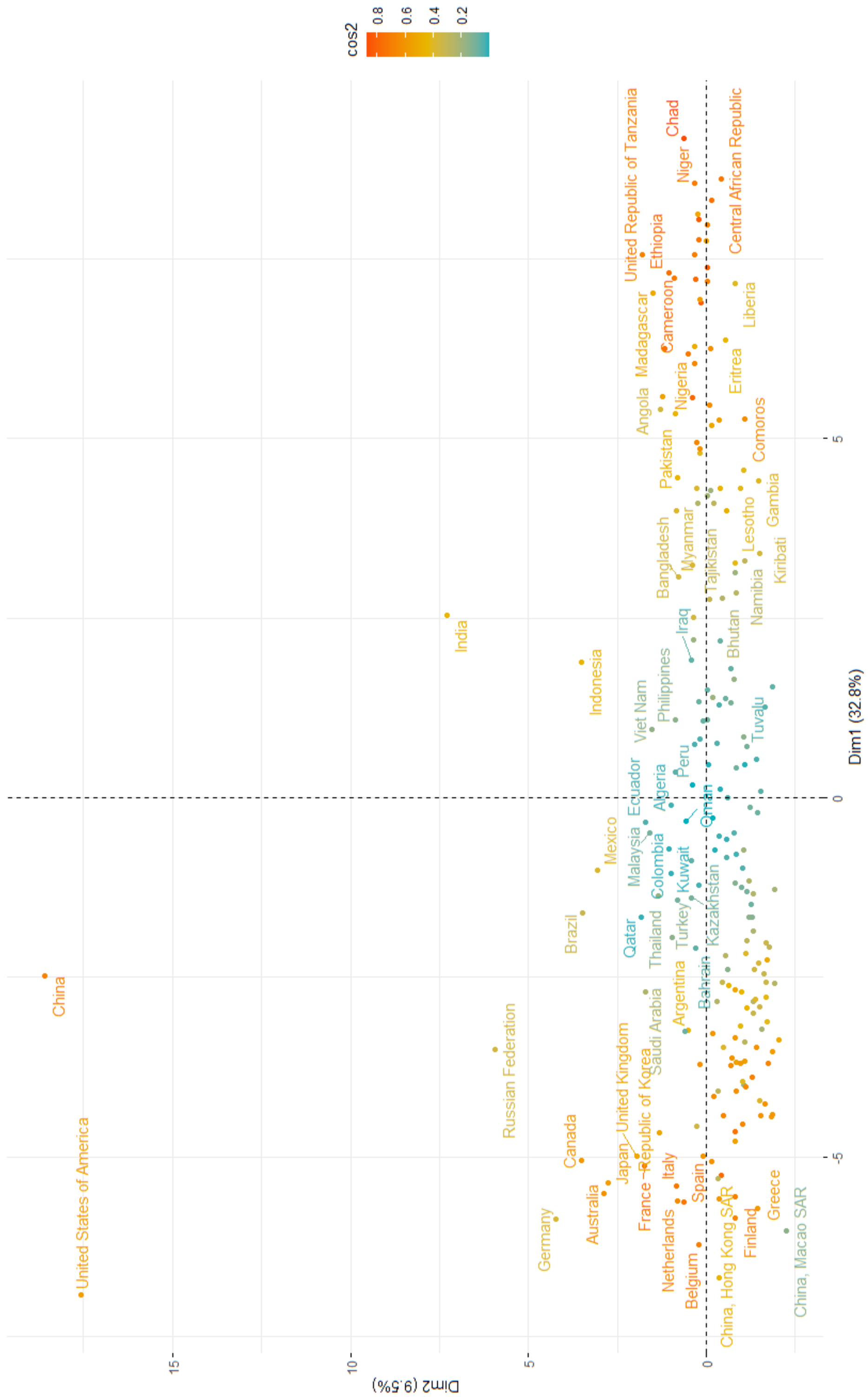


Figure 16: Countries plotted against PC1 and PC2. Color indicates the quality of representation of each sample by the PCs.

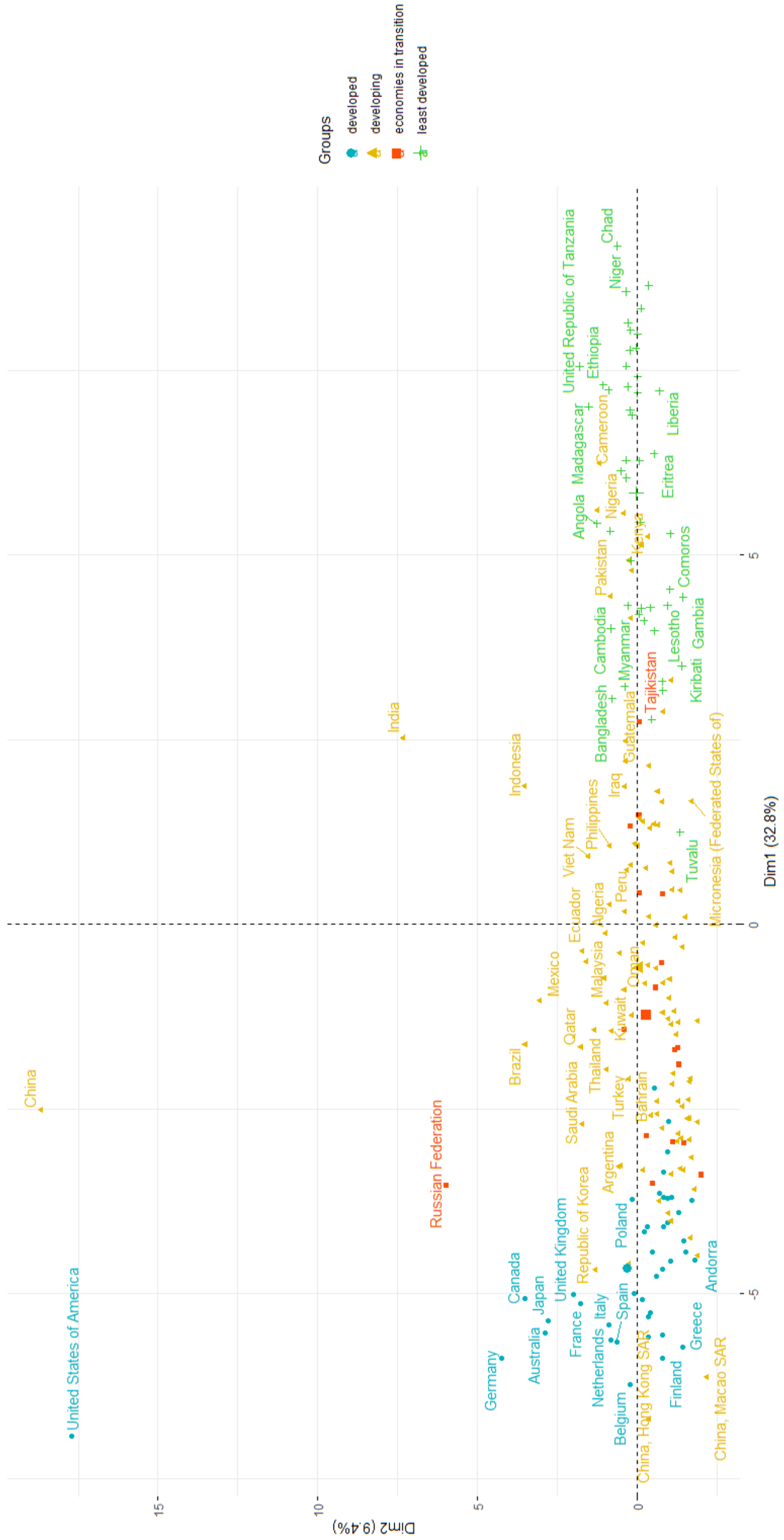


Figure 17: Countries plotted against PC1 and PC2. Color indicates the development level of the country.

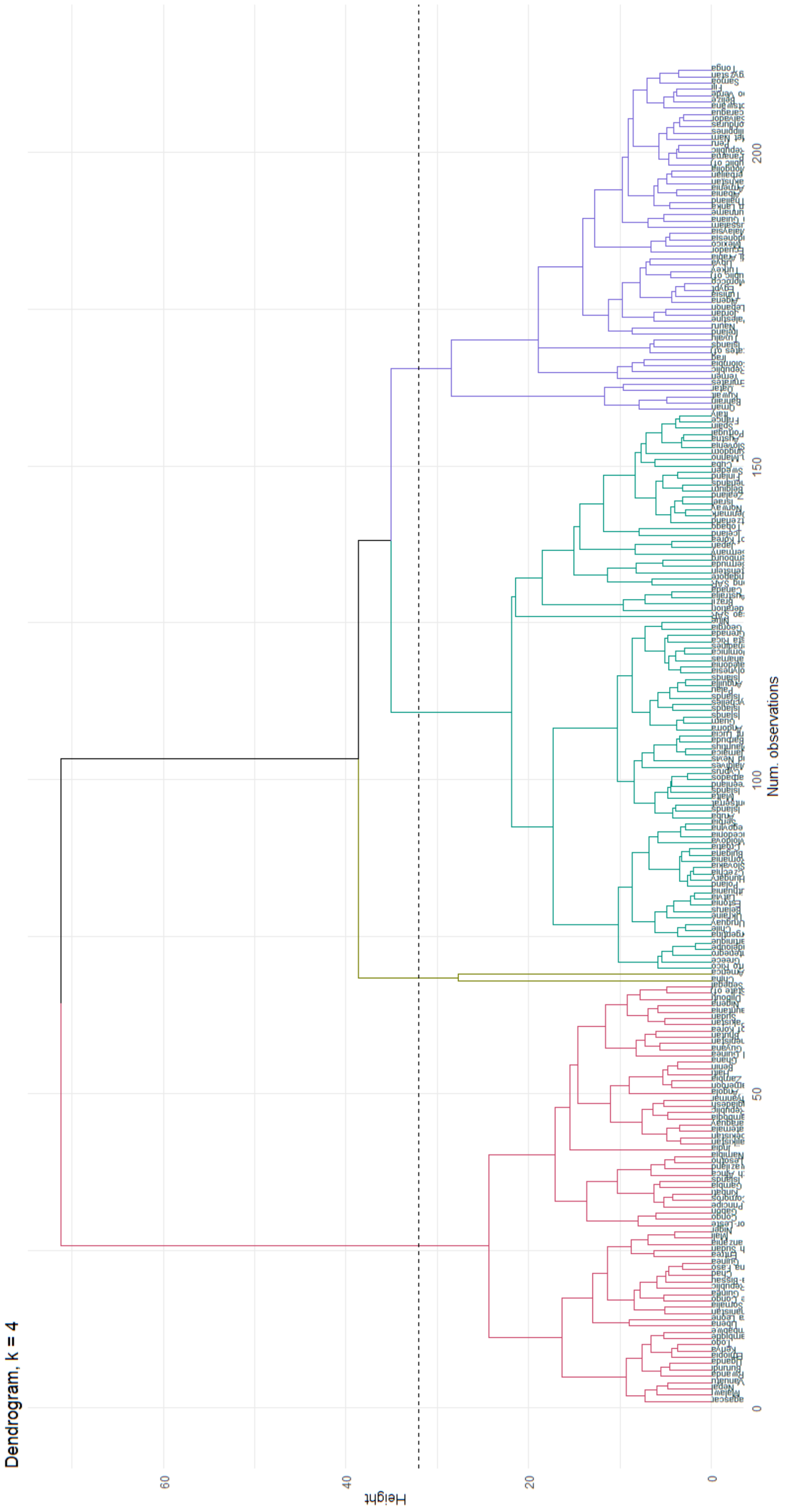


Figure 18: Dendrogram of Hierarchical Clustering with all the countries split into 4 clusters. Ward linkage applied