ECONSTOR
Make Your Publications Visible.

A Service of

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Chakraborty, Aditya; Tsokos, Chris P.

Article

A real data-driven clustering approach for countries based on happiness score

Amfiteatru Economic Journal

Provided in Cooperation with:
The Bucharest University of Economic Studies

Suggested Citation: Chakraborty, Aditya; Tsokos, Chris P. (2021) : A real data-driven clustering approach for countries based on happiness score, Amfiteatru Economic Journal, ISSN 2247-9104, The Bucharest University of Economic Studies, Bucharest, Vol. 23, Iss. Special Issue No. 15, pp. 1031-1045, https://doi.org/10.24818/EA/2021/S15/1031

This Version is available at:
https://hdl.handle.net/10419/281616

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

https://creativecommons.org/licenses/by/4.0/

WWW.ECONSTOR.EU

Mitglied der Leibniz-Gemeinschaft

A REAL DATA-DRIVEN CLUSTERING APPROACH FOR COUNTRIES BASED ON HAPPINESS SCORE

Aditya Chakraborty[1] and Chris P Tsokos[2*]
[1)2)] *University of South Florida, Tampa, FL, USA*

Please cite this article as:
Chakraborty, A. and Tsokos, C.P., 2021. A Real Data-Driven Clustering Approach for Countries Based on Happiness Score. *Amfiteatru Economic*, 23(Special Issue No. 15), pp. 1031-1045.

DOI: 10.24818/EA/2021/S15/1031

Article History
Received: 9 July 2021
Revised: 22 August 2021
Accepted: 28 September 2021

Abstract
In machine learning and data science literature, clustering is the task of dividing the observations (data points) into several categories in such a way that data points falling into one group are being dissimilar than the data points falling to the other groups such that the variation within a group is minimized and the variation between the groups is maximized. It falls under the class of unsupervised learning techniques. It is primarily a tool to classify individuals on the basis of similarity and dissimilarity between them. Our present study utilizes the world happiness data of 156 countries collected by the Gallup World Poll. Our study proposes a useful clustering approach with a very high degree of accuracy to classify different countries of the world based on several economic and social indicators. The most appropriate clustering algorithm has been selected based on different statistical methods. We also proceed to rank the top ten countries in each of the three clusters according to their happiness score. The three leading countries in terms of happiness from cluster 1 (medium happiness), cluster 2 (high happiness), and cluster 3 (low happiness) are Oman, Denmark, and Guyana, respectively, followed by United Arab Emirates, Finland, and Pakistan. Finally, we use four popular machine learning classification algorithms to validate our cluster-based algorithm and obtained very consistent results with high accuracy.

Keywords: Clustering Algorithms, Subjective Well Being (SWB), Stability Measures, Machine Learning Classification Algorithms, Economic Indicators

JEL Classification: C00, C02, C19, C40, C49, C65, Y91, Y10

[*] Corresponding author, **Chris P Tsokos** – e-mail: ctsokos@usf.edu

Authors' ORCID:
Aditya Chakraborty: orcid.org/0000-0002-1476-113X
Chris P. Tsokos: orcid.org/0000-0002-5470-7410

Introduction

Describing happiness is subjective as it depends on individual perspectives. However, feeling good is associated with happiness; ancient philosophers like Aristotle believed that the ultimate aim of human life was eudaimonia which is often translated as "happiness", but more likely means "human flourishing" or "a good life". For many years, social researchers are specifically concerned with self-reports of subjective well-being (SWB), which may be a simple answer to the question "Is a particular person little happy, medium happy, or very happy?" A plethora of people from many countries are asked the same types of questions over many years, and as reviewed in Frey and Stutzer (2002), researchers have begun to use data to tackle a variety of relevant questions relating to happiness or well-being. Economist Richard Easterlin (1974), using happiness data, reported first that increase in personal income over time is not the only driving factor in an increasing level of happiness (Di Tella and MacCulloch, 2006). Being happy is not only associated with personal well-being but also with productivity at a large scale. Throughout history, it has been seen as the ultimate end of temporal existence. Economists performed evidence-based analysis to show that happiness and productivity go hand in hand, increasing productivity by approximately 12% (Oswald et al., 2015). Sabatini and Medicine (2014) analyzed a representative sample of 817 from Italy and found a strong correlation between happiness and perceived good health based on a probit regression model after controlling for several relevant socio-economic phenomena. Researchers (Moeinaddini et al., 2020) proposed a new score to measure personal happiness by identifying the contributing factors based on the Oxford Happiness Questionnaire (OHQ) and used Structural Equation Modeling (SEM) to study the relationship of the individual OHQ items in explaining personal happiness in Skudai, Johor, Malaysia. Recently, social researchers are prone to use sophisticated machine learning techniques in applied sciences (Chakraborty and Tsokos, 2021a) to answer specific types of questions regarding the happiness of individuals and the socio-economic conditions of a country as a whole (Chakraborty and Tsokos, 2021b,c). They developed a data-driven analytical model with high performance to predict the happiness of the developed countries based on different social and economic indicators. Howell and Howell (2008) has shown that there is a positive correlation with the subjective well-being (SWB) economic status of a country. Yarkoni and Westfall (2017) reviewed some of the basic concepts and methods of statistical machine learning and provided instances where these concepts have been implemented to perform important applied psychological research that focuses on predictive research questions. They also recommended that an increased focus on prediction, rather than descriptive explanation, might lead us to a greater understanding of the unknown parameter of interest. Chaipornkaew and Prexawanprasut (2019) developed a machine learning prediction model for human happiness using four popular methods, namely KNN, Multi-Layer Perceptron, Naïve Bayes, and Decision Tree. Their proposed model suggested that the Decision Tree with Random Over-sampler technique is the best prediction algorithm for the analyzed data.

The main goals of our study are the following:

- Clustering is a great tool to investigate any unknown pattern in the data. However, it is necessary to verify the quality of the data set, that is, to check if the data is clusterable or not. We plan to check if there is any random pattern in the data before performing clustering.

- We want to develop an appropriate clustering algorithm by implementing different algorithms to choose from, which will classify similar observations (countries) with a high level of accuracy based on the indicators.

- After the selection of the accurate clustering algorithm, we proceed to perform an analysis of individual clusters to choose the indicators which are most influential.

- We then rank the top ten countries based on the happiness score in each cluster. It would offer us an indication of which countries fall into the low happiness, middle happiness, and the highest happiness categories.

- For the validation purpose of our clustering, we plan to perform four very popular machine learning classification algorithms.

- Finally, for an overall comparison, we plan to create a *global cluster map* to show the position of clusters in the map. It provides a general idea about happiness and also other socio-economic conditions related to the happiness of individual countries. Then, we proceed to compare the cluster map with Figure no. 1, where we create a *world map* based on happiness scores.

In the following Figure no. 1, light pink indicates the countries with the highest happiness scores. These countries include the United States of America, western parts of Europe, Canada, and Australia. On the contrary, red indicates the countries with the least happiness score. We see that the countries with the least happiness scores are some countries in Africa and Asia. We will rank these countries in each cluster based on our clustering algorithm in Section 5.

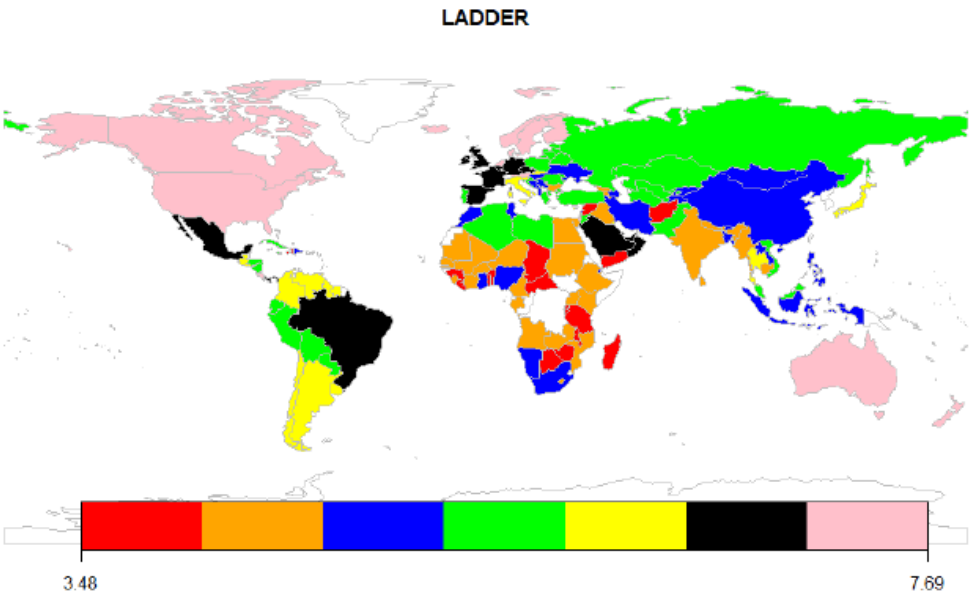

Figure no. 1. World map showing the Happiness Scores (LADDER) of different countries

1. The data

Our study data has been obtained from the World Happiness Report 2019 website (Helliwell et al., 2019), where they used the Gallup Poll to get the answers to specific questions relating to the socio-economic status and how happy the citizens perceive themselves to be. The data has been collected for a total of 156 countries from 2005 to 2018. However, in our study, we only considered the data of developed countries (sorted based on the human development index [HDI]) in the world. Individuals were asked specific questions, and as a result of their response as a whole, a score was produced, which is termed the national average. In our data, the average scores of the developed countries from 2005 to 2018 were tabulated. We have the Cantril life ladder/ladder (an imaginary ladder, with steps numbered from 0 at the bottom to 10 at the top representing the worst possible condition of life and the best possible condition of life of an individual, respectively) as our response and eleven economic indicators in our study. The eleven attributable variables (indicators) that the data was collected on are LOG_GDP (X1) (Log GDP), SOC_SUPPORT (X2) (Social Support), LIFE_EXPECT(X3)(Life Expectancy), FREEDOM (X4), Generosity (X5), PER_CORR (X6) (Perception of Corruption), POS_AFFECT (X7)(Positive Affect), NEG_AFFECT (X8) (Negative Affect), CONF_GOV (X9) (Confidence in Government), DEM_QUALITY(X10) (Democratic quality), and, DEL_QUALITY (X11) (Delivery quality). The description of the indicators can be found in (Chakraborty and Tsokos, 2021b).

2. Research methodology

2.1 Investigating clustering pattern

Before performing any kind of cluster analysis, it is primitive to check if the data we are trying to analyze is clusterable. Analysis of non-clusterable data might produce misleading results if we proceed with clustering of the data.

2.1.1 The Visual Assessment of cluster Tendency (VAT)

The VAT technique (Bezdek and Hathaway, 2002) is a good practice to express the data pictorially for getting a visual representation for the clustering assessment. The technique of deciding whether clusters are present as a step prior to actual clustering is called the assessing of clustering tendency. For the visual evaluation of clustering tendency, dissimilarity matrix between observations has been constructed. In the following Figure no. 2, the color level is proportional to the value of the dissimilarity between observations: pure red if $dist(\chi_i, \chi_j) = 0$ and pure blue if $dist(\chi_i, \chi_j) = 1$. The dissimilarity matrix image in Figure no. 2 confirms that there is a clustering pattern in the study data.

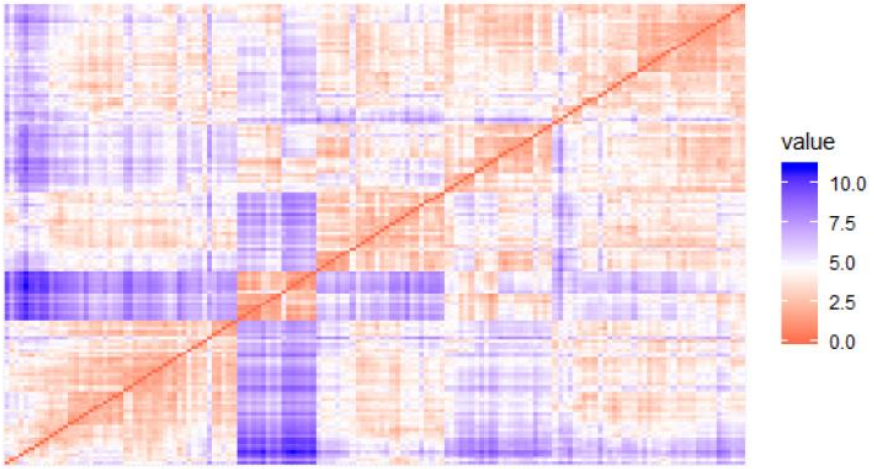

Figure no. 2. The clustering tendency of happiness data.

Notes: Red: High Similarity, Blue: Low Similarity

3. Selecting the best clustering algorithm

Selecting the best clustering algorithm can be a tough call for a research scientist. One of the rudimentary challenges of clustering is how to evaluate results without any supplementary knowledge beforehand. A usual approach for the evaluation of clustering results is to use validity indexes. Clustering validity approaches can use two criteria (Rendón et al., 2011); External criteria (evaluate the result with respect to a pre-specified structure), internal criteria (evaluate the result with respect to information intrinsic to the data alone). Hence, different types of indexes are used to solve different types of problems, and index selection depends on the kind of available information. A usual approach is to use *stability measures*. It is a special sort of internal measure criteria (Zambon et al., 2016), which assesses the uniformity in a clustering mechanism by comparing it with the clusters obtained after each column is removed, one at a time. There are four measures that fall into the stability measure. In all of these cases, the average is taken over all the deleted columns, and all measures must be minimized. The four measures are Average Proportion of Non-overlap (APN), Average Distance (AD), Average Distance between Means (ADM), and Figure of Merit (FOM). For a detailed explanations of these measures, see Della Mea et al., 2006; Jackson, 2008; Borg et al., 2012; Kermani, 2017. For the selection of the most accurate clustering algorithm, we compared among three popular clustering algorithms, namely k-means, hierarchical, and PAM, and we selected the stability as our measure of choosing the most appropriate clustering algorithm as it gives a robust result throughout the four above mentioned stability measures. The following Table no. 1 illustrates that k-means have been chosen by the four stability measures as an optimal algorithm.

Table no. 1. Comparison four stability measures to select the appropriate clustering algorithm

Measure	Score	Algorithm
APN	0.63	k-means
AD	3.326	k-means
ADM	0.3	k-means
FOM	0.765	k-means

4. Optimal number of clusters

Deciding the optimal number of clusters for a set of data is a basic challenge in clustering techniques, such as k-means, k-medoids (PAM), and hierarchical clustering, which requires the user to select the appropriate number of clusters k beforhand. The method of selecting the number of clusters is somehow subjective and also dependent upon using the techniques for computing similarities and the parameters used for partitioning. However, are almost thirty methods (indices, see (Charrad et al., 2014)) to decide the optimum number of clusters; the most popular methods include the elbow method, silhouette methods (Rousseeuw and mathematics, 1987), Hartigan (Telgarsky and Vattani, 2010), and Gap Statistic method (Tibshirani et al., 2001; Kassambara, 2017). (Figures no. 3 and no. 4)

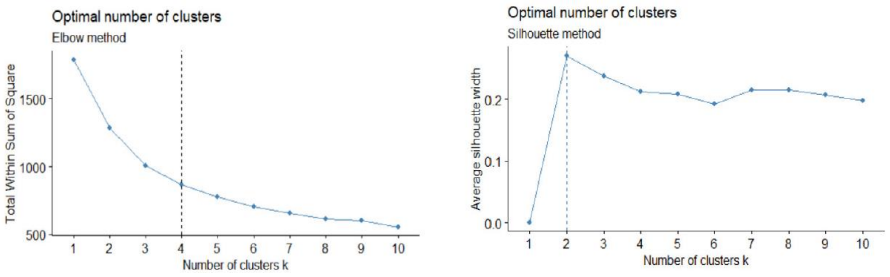

Figure no. 3. Elbow Method (left) and Average silhouette plot (Right) showing the optimum number of clusters

The elbow plot and the average silhouette plot propose the appropriate cluster numbers as *four* and *two*, respectively.

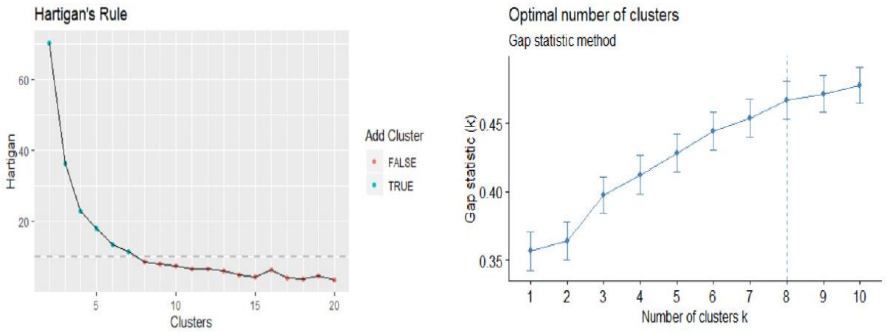

Figure no. 4. Hartigan's Plot (left) and Gap Statistic Plot (Right) showing the optimum number of clusters

The Hartigan's Plot and the Gap Statistic Plot propose the appropriate cluster numbers as *five* and *eight*, respectively.

4.1 The ultimate cluster choice based on clustering validity indices

Selecting the desired number of clusters for a set of data is a basic challenge in clustering and selecting several clustering algorithms lead to different clusters of data. There are several Clustering validity indices in the literature that combine information regarding intra-cluster compactness and inter-cluster isolation, as well as other factors, such as geometric or statistical properties of the data, the number of data objects, and dissimilarity or similarity measurements.Some of those indices are "dunn" , "kl", "ch", "hartigan", "ccc", etc. By selecting the standard Euclidean distance measure, we found the following result based on several Clustering validity indices. (Table no. 2)

Table no. 2. Comparison among several clustering validity indices based on majority votes

Result showing the outputs based on Clustering validity indices
5 proposed 2 as the best number of clusters
12 proposed 3 as the best number of clusters
2 proposed 6 as the best number of clusters
4 proposed 10 as the best number of clusters

5. K-means clustering algorithm

One of the primary objectives of our study is to investigate which countries fall into similar groups (clusters) based on the relevant socio-economic indicators. After we have selected the number of clusters and the appropriate clustering algorithm, we are in a position to perform the k-means clustering with clusters 3 (that is, group the data into three clusters/subgroups). We see that our clustering resulted in 3 clusters with sizes 55, 69, and 26, respectively. The following figure projects the multi-dimensional data into three clusters that are formed as a result of k-means clustering. (Figure no. 5)

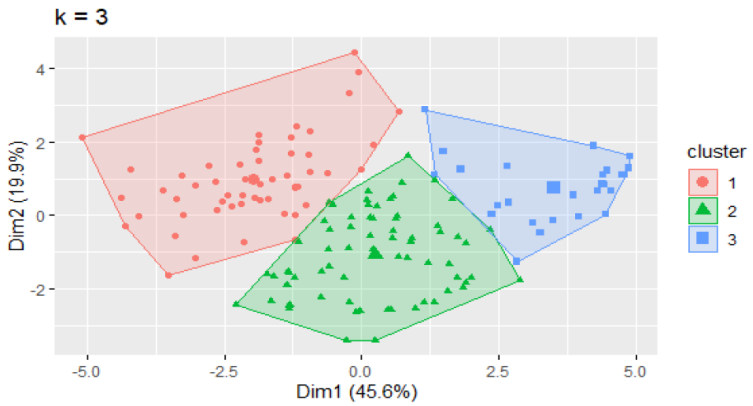

Figure no. 5. The three clusters obtained by k-means clustering

We now proceed to perform some exploratory analysis to fetch some information regarding the behavior of the indicators in each cluster.

From the following Figure no. 6, we notice that countries falling within cluster 2 happens to be happiest among the three, followed by cluster 1 and cluster 3. Most probably, cluster 2 contains the most developed and democratic countries of the world, and cluster 3 contains most of the underdeveloped and developing and less democratic countries. It seems Cluster 2 contains most of the developing countries, followed by Cluster 1 and Cluster 3.

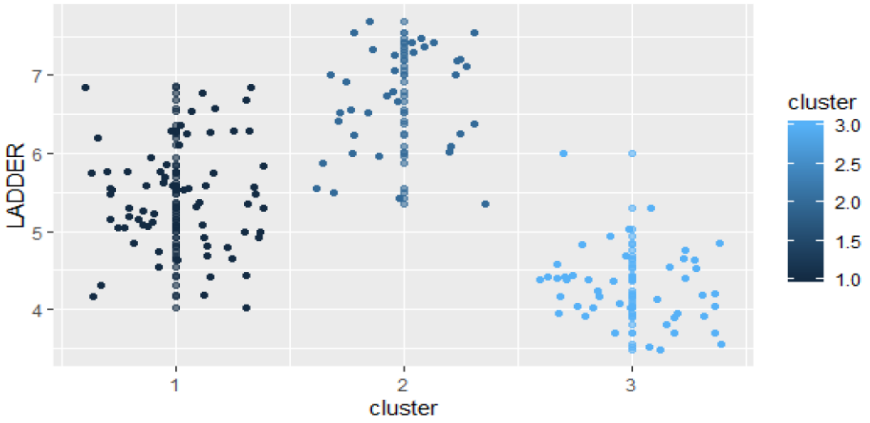

Figure no. 6. The distribution of happiness throughout three clusters

After implementing the k-means algorithm, we have computed the cluster means for all eleven indicators. We can plot the cluster means for all eleven indicators throughout the three clusters to compare how each attributable variable (indicators) behaves in these three clusters. As Figure no. 7 illustrates, we notice that each average values of the indicators in cluster 2 are higher than that of cluster 1 and cluster 3 except NEG_AFFECT (X_8), PER_COR (X_6), and CONF_GOV (X_9). On the other hand, every average numerical measure of indicators in cluster 3 is worse than that of cluster 2 and cluster 1 except Generosity (X_5) and CONF_GOV (X_9). By studying the graphs, we see an almost opposite pattern between cluster 2 and cluster 3. For example, we see the variable SOC_SUPPORT (X_2) in cluster 3 has been placed into a completely opposite position when compared to cluster 2. By visualizing the pattern, we might tell that most developed countries are placed into cluster 2, and most underdeveloped countries are classified into cluster 3. Cluster 1 behaves pretty averagely when compared with cluster 2 and cluster 3. However, those countries classified into cluster 1 have the lowest average Generosity (X_5) values which need to be further analyzed. It is important to observe that the average value for CONF_GOV (X_9) in cluster 3 is higher than that of cluster 1 and cluster 2. Also, it seems that there must be some correlation between cluster 1 and cluster 2 as they show a parallel trend for most of the indicators.

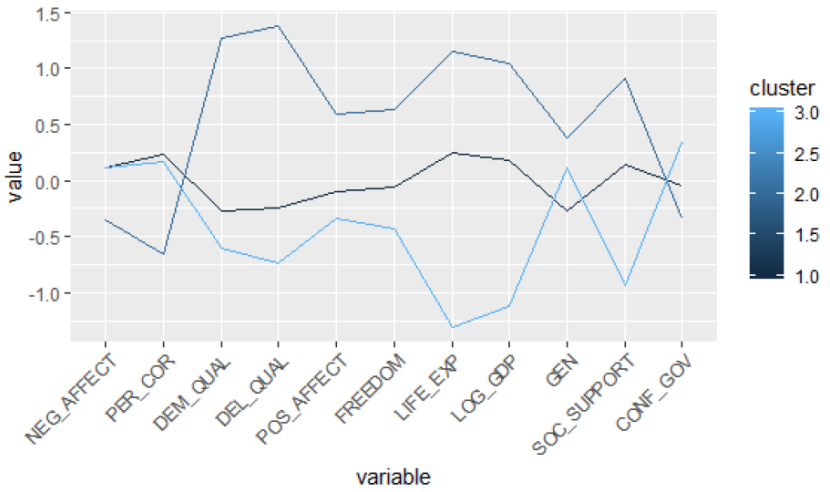

Figure no. 7. The distribution of indicators throughout three clusters

For a better understanding, we can visualize the variability of three clusters with respect to each indicator. From the following Figure no. 8, we can obtain some very interesting facts about each indicators factor in each cluster. We see that countries belonging to cluster 3 has especially significantly low measures for indicators LIFE_EXP (Life Expectancy), SOC_SUPPORT (Social Support), and LOG_GDP (logarithm of GDP). Since we got an idea that underdeveloped countries belong to cluster 3, they can expect to have low GDP and life expectancy. One striking fact that these countries also have very low scores for social support (X_2) which means that on an average the citizens of those countries does not feel to get support from their relative, friends or Governments when they are in trouble. One interesting fact to notice that, however, there is not much difference among the Generosity (X_5) within the three clusters; some countries in cluster 3 outperforms some countries in cluster 2 when it comes to Generosity (X_5). Also, the countries belonging to cluster 1 and cluster 3 happen to have almost the same average measures of perception of corruption (X_6).

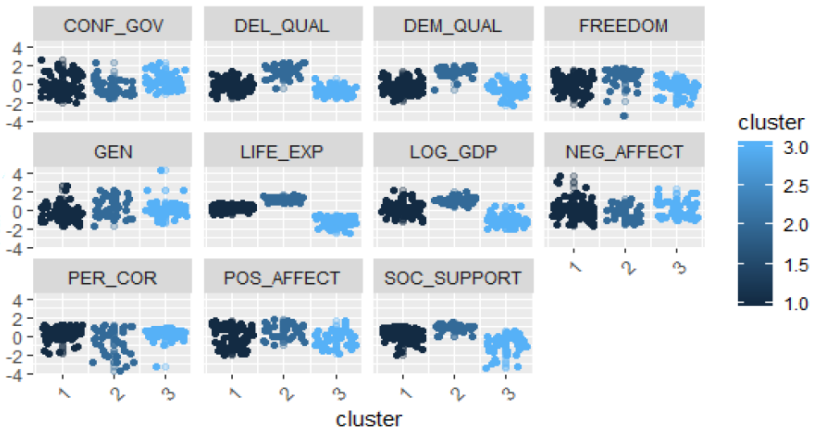

Figure no. 8. The distribution of each indicators individually for three clusters

One of the most important consequences of the clustering aspect is that we can rank the different countries in the world in each cluster based on the happiness score. Since there is a positive correlation between happiness and democracy, by knowing the name of the happiest countries in the cluster, we might be able to guess their socio-economic status. Table no. 3 below shows the top 10 countries in each cluster based on happiness score.

Table no. 3. Ranking of countries in different clusters based on happiness score

Rank	Cluster 2	Cluster 1	Cluster 3
1	Denmark	Oman	Guyana
2	Finland	UAE	Pakistan
3	Norway	Mexico	Nigeria
4	Netherlands	Brazil	Laos
5	Canada	Qatar	South Africa
6	Iceland	Saudi Arabia	Djibouti
7	Sweden	Argentina	Ghana
8	New Zealand	Kuwait	Namibia
9	Australia	Colombia	Mozambique
10	Austria	Trinidad and Tobago	Zambia

6. Validation of clustering algorithm

After clustering the observations into three clusters, the next important thing is to assess the validity of the clustering algorithm. To assess the performance of our k-means clustering algorithm, we have used four machine learning classification algorithm to see how well the data has been clustered with a high degree of accuracy. We chose Decision Tree RPART (Quinlan, 1986; Therneau et al., 2015; Jauhari and Supianto, 2019), Decision Tree C5.0 (Pandya and Pandya, 2015), Random Forest (Breiman, 2001; Lantz, 2019), and Extreme Gradient Boosted Tree (XGBTREE) (Friedman and analysis, 2002; Natekin and Knoll, 2013; Chen et al., 2015) classification algorithms for checking cluster validity. The reason behind selecting these two classification algorithms is that they have no distributional assumptions and are also very popular supervised algorithms that happen to work well with a large amount of data. If we get high classification accuracy by these two methods, we can conclude that the k-means clustering method we have implemented here is one of the good if not the best clustering algorithm for the data we have. After we implement the four popular machine learning classification algorithms to judge the performance of our k-means clustering algorithm, it is important to check and evaluate the performance of the classification algorithms in terms of evaluation matrices. We will access the quality of the classification algorithms based on the following two evaluation matrices.

6.1 Accuracy

Accuracy is one of the most widely used metrics for evaluating classification models. Conventionally, multi-class accuracy is defined as the average number of correct predictions as follows.

$$Accuracy = \frac{1}{N}\sum_{k=1}^{|G|}\sum_{x:g(x)=k} I(g(x) = \hat{g}(x)) \quad (1)$$

Where G is the number of classes, $g(x)$ and $\hat{g}(x)$ are the classifier and estimated value of the classifier respectively . $I(\cdot)$ is an indicator function which takes the value 1 if the classes match and 0 otherwise.

6.2 Cohen's Kappa

Cohen's Kappa or Kappa statistic is a very useful metric in machine learning when we deal with a multi-class classification problem. Basically, it suggests how better the desired classifier performs over the performance of a random classifier that simply makes arbitrary guesses according to the frequency of each class. Always Cohen's kappa (Vieira et al., 2010) is less than or equal to 1 and values zero or less, implies that the classifier is not useful. It is defined as follows.

$$\kappa = \frac{(p_0 - p_e)}{(1 - p_e)} = 1 - \frac{(1 - p_0)}{(1 - p_e)} \quad (2)$$

where p_0 is the observed accuracy, the number of instances that were classified correctly and p_e is the expected accuracy, the accuracy that any random classifier would be expected to achieve based on the confusion matrix. It is very interesting to note that we got very high accuracy and Kappa (κ) values using the four machine learning classification algorithm implying the great performance of our k-means algorithm.

The following Table no. 4 illustrates the values of accuracy and Kappa(κ) generated by the four classification algorithm.

Table no. 4. Comparing the performance of different classification methods

Method	Accuracy	Kappa
Rpart	96.6	94.7
C5.0	97.3	95.9
Random Forest	97.8	96.5
XGBTREE	99.7	99.1

From the above Table no. 4, we see that the classification is pretty good and consistent with respect to accuracy and kappa measure based on the four popular classification algorithms.

The following Figure no. 9 shows a visual comparison of the four different classification methods based on accuracy and kappa. It shows that the extreme gradient boosted tree performs the best, followed by random forest, decision tree C5.0, and decision tree RPART. For the first three algorithms in the figure, we see the median accuracy and kappa measures are pretty indistinguishable.

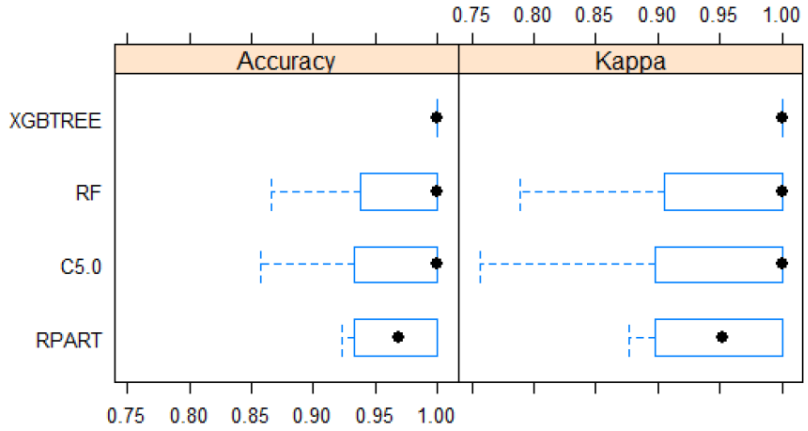

Figure no. 9. Visual representation of four classification methods in terms of accuracy and kappa

Conclusions

By implementing an appropriate clustering algorithm to our happiness data, we have successfully accomplished all the goals introduced in the introduction.

We have shown statistically and visually that there is a meaningful clustering pattern in our happiness data.

We have implemented different methodologies to select an optimal number of clusters as three and the most appropriate clustering algorithm as k-means.

We have compared the happiness scores for different clusters and have done exploratory data analysis to understand which indicators contribute the most to each cluster.

We have ranked the top ten countries in each of the three clusters according to their happiness score. The three leading countries in terms of happiness from cluster 1, cluster 2, and cluster 3 are Oman, Denmark, and Guyana, respectively, followed by Finland, United Arab Emirates, and Pakistan.

For the validation purpose of our clustering algorithm, we have selected four popular machine learning classification algorithms to compare with. We got outstanding classification accuracy, which was also pretty consistent throughout the four methods implying that our cluster has been instrumental.

The following Figure no. 10 shows that our clustering has been very useful if we compare it with Figure no. 1 in section 1. As we have guessed earlier, the happiest countries are those which fall into cluster 2 (green), followed by cluster 1 (red) and cluster 3 (blue).

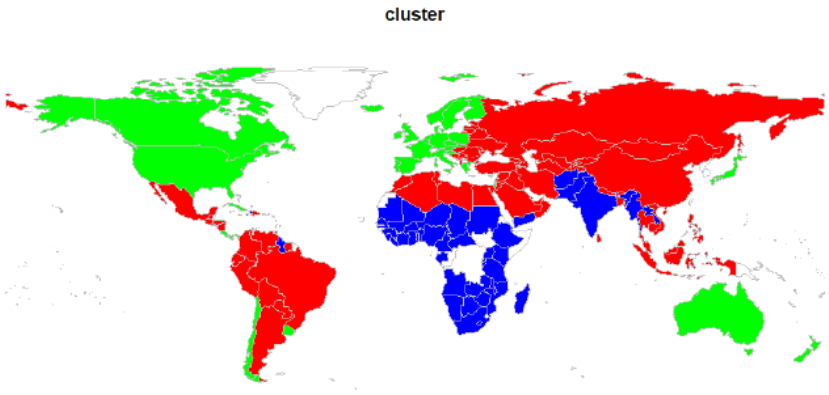

Figure no. 10. The distribution of three clusters in world map

The way we have ranked that countries in different clusters by happiness score, one can rank the countries based on all indicators. This will provide a tremendous amount of information about the economic condition of individual countries, and also, at the same time, those countries with a low score would be able to understand which indicators they are supposed to be working on. The study's utility encourages economists and other social scientists to pay closer attention to subjective well-being as a causative force. Furthermore, because individual happiness in an organization has a favorable impact on production, the most influential metrics can be employed for companys' promotion plans. They may be useful for managers and human resources professionals for making decision making and strategic planning. One of the essential parts of our research is the ability to visualize subjective well-being (SWB) and the factors influencing it from a global perspective. One might also try different types of clustering algorithms such as PAM (Partition Around Medoids), Hierarchical clustering, etc., and can also evaluate their accuracy by using different classification algorithms. It would also be interesting to investigate the performance of dimension reduction techniques as PCA (Principal Component Analysis), PLS (Partial Least Square), and Factor Analysis techniques to use the components as potential indicators to predict happiness score in the future.

References

Bezdek, J.C. and Hathaway, R.J., 2002, May. VAT: A tool for visual assessment of (cluster) tendency. In: S.n., *The 2002 International Joint Conference on Neural Networks.* Honolulu, HI, USA, 12-17 May 2002. S.l: IEEE.

Borg, M., Badr, I. and Royle, G.J., 2012. The use of a figure-of-merit (FOM) for optimisation in digital mammography: a literature review. *Radiation protection dosimetry*, 151(1), pp.81-88.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5-32.

Chaipornkaew, P. and Prexawanprasut, T., 2019. A Prediction Model for Human Happiness Using Machine Learning Techniques. In: s.n., *2019 5th International Conference on Science in Information Technology*. Yogyakarta, Indonesia, 23-24 October 2019. S.l: IEEE.

Chakraborty, A. and Tsokos, C.P., 2021a. A Real Data-Driven Analytical Model to Predict Happiness. *Sch J Phys Math Stat*, 8(3), pp.45-61.

Chakraborty, A. and Tsokos, C.P., 2021b. Parametric and Non-Parametric Survival Analysis of Patients with Acute Myeloid Leukemia (AML). *Open Journal of Applied Sciences*, 11(01), p.126.

Chakraborty, A. and Tsokos, C., 2021c. Survival Analysis for Pancreatic Cancer Patients using Cox-Proportional Hazard (CPH) Model. *Global Journal Of Medical Research*, [e-journal] 21(3-F). doi:10.34257/GJMRFVOL21IS3PG29

Charrad, M., Ghazzali, N., Boiteau, V. and Niknafs, A., 2014. NbClust: an R package for determining the relevant number of clusters in a data set. *Journal of statistical software*, 61(1), pp.1-36.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y. and Cho, H., 2015. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4), pp.1-4.

Della Mea, V., Demartini, G., Di Gaspero, L. and Mizzaro, S., 2006. Measuring retrieval effectiveness with average distance measure (ADM). *Information Wissenschaft und Praxis*, 57(8), pp.433-443.

Di Tella, R., and MacCulloch, R., 2006. Some uses of happiness data in economics. *Journal of economic perspectives*, 20(1), pp.25-46.

Friedman, J.H., 2002. Stochastic gradient boosting. *Computational statistics & data analysis*, 38(4), pp.367-378.

Hastie, T., Tibshirani, R. and Walther, G., 2001. Estimating the number of data clusters via the Gap statistic. *J Roy Stat Soc B*, 63, pp.411-423.

Helliwell, J., Layard, R. and Sachs, J., 2019. *World Happiness Report 2019*. [online] New York: Sustainable Development Solutions Network. Available at: <https://worldhappiness.report/ed/2019/#read> [Accessed 8 September 2021].

Howell, R.T. and Howell, C.J., 2008. The relation of economic status to subjective well-being in developing countries: a meta-analysis. *Psychological bulletin*, 134(4), p.536.

Jackson, M.O., 2008. Average Distance, Diameter, and Clustering in Social Networks with Homophily. In: C. Papadimitriou and S. Zhang eds., 2008. *Internet and Network Economics. WINE 2008. Lecture Notes in Computer Science*, vol 5385. Berlin, Heidelberg: Springer. https://doi.org/10.1007/978-3-540-92185-1_3.

Jauhari, F. and Supianto, A.A., 2019. Building student's performance decision tree classifier using boosting algorithm. *Indones. J. Electr. Eng. Comput. Sci*, 14(3), pp.1298-1304.

Kassambara, A., 2017. *Practical guide to cluster analysis in R: Unsupervised machine learning*. Vol. 1. S.l: Sthda.

Kermani, F., 2017. Validation of clustering methods for medical data sets. *ActaHealthMedica*, 2(1), pp.116-116.

Lantz, B., 2019. *Machine learning with R: expert techniques for predictive modeling*. S.l: Packt publishing ltd.

Moeinaddini, M., Asadi-Shekari, Z., Aghaabbasi, M., Saadi, I., Shah, M.Z. and Cools, M., 2020. Proposing a new score to measure personal happiness by identifying the contributing factors. *Measurement*, 151, p.107115.

Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, p.21.

Oswald, A.J., Proto, E. and Sgroi, D., 2015. Happiness and productivity. *Journal of Labor Economics*, 33(4), pp.789-822.

Pandya, R. and Pandya, J., 2015. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. *International Journal of Computer Applications*, 117(16), pp.18-21.

Pang, S.L. and Gong, J.Z., 2009. C5. 0 classification algorithm and application on individual credit evaluation of banks. *Systems Engineering-Theory & Practice*, 29(12), pp.94-104.

Quinlan, J.R., 1986. Induction of decision trees. *Machine Learning*, 1(1), pp.81-106.

Rendón, E., Abundez, I., Arizmendi, A. and Quiroz, E.M., 2011. Internal versus external cluster validation indexes. *International Journal of Computers and Communications*, 5(1), pp.27-34.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, pp.53-65.

Sabatini, F., 2014. The relationship between happiness and health: evidence from Italy. *Social Science & Medicine*, 114, pp.178-187.

Telgarsky, M. and Vattani, A., 2010, March. Hartigan's method: k-means clustering without voronoi. [online] Available at: <http://proceedings.mlr.press/v9/ telgarsky10a/ telgarsky10a.pdf> [Accessed 8 September 2021].

Therneau, T., Atkinson, B., Ripley, B. and Ripley, M.B., 2015. *Package 'rpart'*. [online] Available at: <cran.ma.ic.ac.uk/web/packages/rpart/rpart.pdf> [Accessed 20 April 2016].

Vieira, S.M., Kaymak, U. and Sousa, J.M., 2010. Cohen's kappa coefficient as a performance measure for feature selection. In: S.n., *International Conference on Fuzzy Systems*. Barcelona, Spain, 18-23 July 2010. s.l: IEEE.

Yarkoni, T. and Westfall, J., 2017. Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), pp.1100-1122.

Zambon, G., Benocci, R. and Brambilla, G., 2016. Statistical road classification applied to stratified spatial sampling of road traffic noise in urban areas. *International Journal of Environmental Research*, 10(3), pp.411-420.