A Comparative Analysis of Machine Learning Techniques for Exploring Country Clustering Based on Life Expectancy

Dennis Lim Kam Ho  
*Computer Science Department School of Computer Science*  
Bina Nusantara UniversityJakarta, Indonesia  
dennis.kam@binus.ac.id

Marcellino Bonamutial  
*Computer Science Department School of Computer Science*Bina Nusantara UniversityJakarta, Indonesia  
marcellino.bonamutial@binus.ac.id

Emmanuel Brandon Hamdi  
*Computer Science Department School of Computer Science*  
Bina Nusantara UniversityJakarta, Indonesia  
emmanuel.hamdi@binus.ac.id

Karli Eka Setiawan  
*Computer Science Department, School*

*of Computer Science*  
Bina Nusantara University

Jakarta, Indonesia  
karli.setiawan@binus.ac.id

Muhammad Fikri Hasani  
*Computer Science Department, School*

*of Computer Science*  
Bina Nusantara University

Jakarta, Indonesia  
karli.setiawan@binus.ac.id

***Abstract*—** **This study employs data mining techniques and clustering algorithms to analyze life expectancy factors and discover influential elements. Using World Health Organization data, including life expectancy figures for various countries and years, the study examines mental and physical health conditions, disease rates, and accidents. Three clustering models—DBSCAN, k-means, and fuzzy c-means—are developed and evaluated using metrics such as Silhouette score, Davis Bouldin Index (DBI), and Calinski-Harabasz index. The results favor the k-means model, with a Silhouette coefficient of 0.7812, a Calinski-Harabasz score of 541.2260, and a Davies-Bouldin score of 0.3297. Compared to previous studies, this k-means model excels, highlighting its effectiveness in analyzing life expectancy factors. The findings contribute to global efforts in improving life expectancy and offer insights for policymakers, UNDP, and other stakeholders. Identifying key factors can guide targeted interventions, resource allocation, and policy development to enhance healthcare infrastructure and address poverty in underprivileged countries. This research emphasizes data-driven decision-making and inspires collaborative efforts for a healthier and more equitable world.**

*Keywords—Clustering, Artificial Intelligence, Life Expectancy, Country, Model Analysis*

# Introduction

Currently, technology is increasingly advanced and developing rapidly. Developments occur in various fields including in the field of artificial intelligence. With the development of the field of artificial intelligence, many advantages such as analysis of statistical data to serve a better understanding of trends in real life are possible in the science and mathematics field [1].

Other fields are also developing, one of which is the health sector. Artificial intelligence and the health sector has been used in the health sector in diagnosis, prediction, Imaging, Treatment, and even in drug production, all with a proven higher degree of accuracy and a lower human intervention with many diseases that can be diagnosed and treated earlier with the help artificial intelligence prediction that shows promising result in the field of medicine that with the development of the health sector [2]. Life expectancy can increase because disease and health are one of the most significant factors that can influence the life expectancy of a person [3]. By acknowledging these data hopefully, the policymakers of each country can contribute to increase their quality of life and reassure their health facilities to support communal health [4].

By knowing these factors related to life expectancy, data mining techniques with clustering can be carried out with various appropriate machine learning algorithms [5]. The existence of interrelated factors that affect life expectancy can be the basis for using datasets that describe the conditions of each country as material for analysis [6]. This study will compare several suitable algorithms such as k-means, fuzzy c-means clustering [1], and DBSCAN [7], and look for appropriate parameter values to produce good performance.

This research is aimed to potentially contribute to global efforts in improving life expectancy that can give valuable insight for the policy makers of each nation, international organization that is United Nations Development Program (UNDP), and other stakeholders that can help guide targeted interventions, resource allocation, and policy development aimed at eradicating poverty, enhancing healthcare infrastructure, and promoting investment in underprivileged countries. This can identify the key factors that can influence life expectancy and understanding their relationships. Once understood, this study can serve as a foundation for data-driven decision-making and inspire collaborative efforts to create a healthier, more equitable world.

# Literature Review

From the literature, life expectancy can be used to compare social categories within countries or regions [8]. Various socioeconomic factors influence a country's life expectancy [9], which has been increasing as shown by some studies [10] and [11]. However, health factors such as an unhealthy lifestyle [12, 13], pollution [14], treatable diseases [15], social needs [16], social determinants [17], and mortality rates [18] can negatively impact life expectancy. More factors including cancer, COVID-19 [19], mental psychology [16, 17], and other determinants that indirectly affect health. As a result, the rate of increase in life expectancy may slow down or even decrease. Additionally, factors such as economic inequality, unemployment rates, low education levels, and population growth have a negative impact on life expectancy [20-21].

These factors can be used in data mining through clustering [22], an effective pattern-finding algorithm [23]. Clustering has been used in previous studies to cluster healthy life expectancy factors (HLE) to better forecast HLE [24], and to find the relation between gender equality and how it affects health indicators [25]. Proposed methods involve grouping countries using models based on previous research, such as k-means clustering and fuzzy c-means [1]. DBSCAN is also appropriate for this case [26] and will be compared to the models. Spectral clustering showed the best results in a previous study [27], but each algorithm has its use cases and can be improved with modifications such as by adding a statistical based method [28]. These models can be implemented on different platforms, such as the RapidMiner Tool [4] or the scikit-learn library in Python [7]. Life expectancy is useful as a parameter for classifying countries and running the model [29].

The outcomes of each model are distinct, with DBSCAN being unable to detect anomalies in datasets that have the same trend or seasonality, making it appropriate for our research data since it does not involve time series data [30]. Besides DBSCAN, the k-means model is also a promising option as it has a fast execution time and model performance comparable to other models like fuzzy c-means [6]. The fuzzy c-means model can also be utilized in this study [31] to compare and evaluate parameters that can enhance model performance and surpass other models proposed in this paper.

# Research Methodology

The research process begins with the selection of the dataset that will be used for analysis by the model developed in the next stage. After that, the development of three models is carried out, each using a different approach and algorithms to process the data with the same goal which is to classify existing countries. Following that, the analysis of the performance results of each model is conducted using three different algorithms, aiming to identify the model with the best performance as the outcome of this research. The Fig. 1. below shows the illustration of the workflow of this research.

A diagram of data processing

Description automatically generated with medium confidence

1. Clustering analysis workflow

The models that compared in this research are DBSCAN, k-means, and Fuzzy C-Means as the three clustering models lies in their unique algorithms and strengths. DBSCAN's density-based approach is well-suited for identifying clusters with varying shapes and densities, while k-means excels in partitioning data into well-defined, non-overlapping clusters when the number of clusters is known. Fuzzy C-Means, on the other hand, offers a soft clustering method allowing data points to belong to multiple clusters with varying degrees of membership, making it suitable for datasets with uncertain or overlapping cluster boundaries. The comparison of three algorithms provides a comprehensive approach to clustering tasks, catering to diverse data distributions and structures.

The research commences with data collection from Kaggle, followed by data preprocessing and modelling on the Google Colaboratory platform with Python programming. The initial step involves downloading the dataset in CSV format and importing it into Google Colab, where data manipulation is executed in dataframe format involving the pandas library. Specifically, data for the year 2019 is extracted, while columns predominantly containing null (['BMI']) or zero values (['Trichomoniasis', 'Genital herpes', 'Trichuriasis', 'Food-bourne trematodes', 'Depressive disorders', 'Autism and Asperger syndrome', 'Childhood behavioural disorders', 'Non-migraine headache', 'Infertility']) are eliminated.

To address missing values, mean imputation is applied, taking into account the dataset's contextual relevance. Subsequently, data is separated and extracted exclusively for the male gender category, given the adoption of unsupervised learning models that obviate the need for country labels.

The dataset is standardized using the standard scaler from the scikit-learn library, followed by the application of Principal Component Analysis (PCA) to extract two principal components that represent the dataset. This dimensionality reduction enables the creation of an interpretable two-dimensional plot, enhancing the visualization's comprehensibility.

The study then progresses to modeling, with the pre-processed dataset input into each model: DBSCAN, k-means, and fuzzy c-means. To optimize hyperparameters, the silhouette score is harnessed, guiding the selection of the epsilon value for DBSCAN, the k value for k-means, and the c value for fuzzy c-means within specified ranges (0-1 for DBSCAN, 2-20 for k-means and fuzzy c-means). The elbow method, based on the WCSS (Within Cluster Sum of Squares) score to validates the chosen parameter values. Subsequently, clustering result plots are generated for each model.

The final stage entails conducting comprehensive metric analyses, including the Calinski-Harabasz index and the Davies-Bouldin index, for each model. These metrics form the foundation for the rigorous comparison of model quality, enabling informed insights into the efficacy of clustering models for life expectancy prediction.

## Dataset

This research is based on World Life Expectancy taken from Global Health Observatory (GHO) which is a public health data repository established by the World Health Organization (WHO). The dataset depicts the life expectancy figures for each country in the years 2000, 2010, 2015, and 2019, with each country having two different rows indicating the gender differences and its features. The dataset used also has a total of 147 columns describing each country's life expectancy-related information such as mental health conditions, physical health conditions, accident rates, and many types of disease rate in each country.

The data mainly consist of integer and floats, whereas the country is object type since it is a name of a country in the dataset. Since we are clustering the life expectancy, using Explanatory Data Analysis on the dataset using correlation heatmap to find the relation of the column containing life expectancy. The data shown that the 'life expectancy' column is shown positive linear relationship with 'Schooling' and 'Income Composition of Resources', while 'Population' has near to no linear relationship with 'life expectancy'. 'Adult Mortality' and 'HIV/AIDS' has strong negative linear relationship with 'Life expectancy'.

The data must be preprocessed first to prepare it for the clustering model fitting stage. The first step is to extract the data for the latest year, which is 2019, to create a more accurate model reflecting the current conditions. Then, columns that are not representative, characterized by having many zero or empty values, are removed to reduce dimensionality and avoid the curse of dimensionality that can lower our model performance [32]. Afterward, columns with empty or null values are handled by imputing them with the mean value of the respective column (mean imputation) since this method is suitable for the data type [33]. Further method to improve the previous research which aims to reduce the dataset dimension used in this research is PCA (Principal Component Analysis) method to reduce dataset into only two columns based on the components so that it can be deployed to human-friendly graph.

Once the data is separated based on gender for separate data analysis, as both have unique data distribution characteristics [34], the labels in the dataset are separated first because clustering models are trained using unlabeled data as part of unsupervised learning. The existing labels are stored so that we can know the countries can be categorized based on the clustering results. Subsequently, the data is normalized using the standard scaler to improve model performance through a comparable data range [35].

## Models

The proposed models for this study are DBSCAN, k-means, and fuzzy c-means. The proposed model then set to a default parameter and fit to the proposed model with the provided dataset. This study then will compare the result of the models to find the best option for this case.

1. DBSCAN as Density-Based Spatial Clustering of Applications with Noise is a machine learning algorithm that utilizes two inputs: the radius (ε) and the minimum number of points (minPts). It classifies points based on their proximity within the radius neighborhood. When a core point is identified, DBSCAN forms a cluster by including all points within the specified radius ε from that point. If any of the newly added points are also core points, the cluster expands to include their epsilon neighborhoods, and this expansion process continues until no unexpanded core points remain in the cluster. Points that cannot be assigned to any cluster are considered noise. Therefore, the DBSCAN algorithm is introduced as a clustering model [28].
2. K-means is a popular clustering algorithm used to partition data into k distinct clusters [1]. It operates iteratively by assigning data points to the nearest cluster centroid and then recalculating the centroids based on the mean of the points assigned to each cluster. This process continues until convergence, resulting in clusters that minimize the sum of squared distances from data points to their respective centroids. k-means is computationally efficient and widely applied in various fields for data segmentation and pattern recognition tasks. However, it requires specifying the number of clusters (k) in advance and may be sensitive to the initial centroid positions.
3. Fuzzy c-means is a machine learning algorithm that determines the data clustering using the fuzzy theory, the method itself pointing to determine the data to be a member. The method itself is very easy to implement and robust against outliers, hence fuzzy c-means model algorithm is proposed [1]. In FCM, each data point is assigned a membership value indicating its degree of belongingness to each cluster. These membership values are represented as fuzzy numbers between 0 and 1. The algorithm iteratively updates the membership values and cluster centroids until convergence is reached. The update process involves recalculating the centroids based on the weighted average of data points using their membership values. The weights reflect the degree of influence of each data point on the centroid calculation.

## Performance Metrics

To test our model research, we used three different types of cluster evaluation which are the Silhouette score, Davis Bouldin Index (DBI), and Calinski-Harabasz index methods to evaluate the model results.

1. Silhouette score is a cluster evaluation method that uses the silhouette coefficient of every sample which is used to find the performance of the clustering algorithm and the recommended number of clusters with a positive silhouette score that is closer to 1 means the prediction data is in the right cluster, a score that is near 0 means that the prediction data probably belongs to another cluster, with a negative score that is closer to -1 means the prediction data is wrong, the silhouette score can be computed by taking the means out of every silhouette coefficient from the dataset using the equation With variable a being the average intra-cluster distance and variable b being the average nearest-cluster distance. [36].

|  |  |
| --- | --- |
|  | (1) |

1. Davis Bouldin Index is one of the methods for evaluating the clustering model quality by seeing the similarity of each data in a cluster and how different it is compared to other data in a different cluster or to finding the distance between data points in the same cluster and the distance between each cluster with a Davis Bouldin Index score that is closer the 0 means the clustering is optimal. [37], the Davis Bouldin Index can be calculated with the following equation.

|  |  |  |
| --- | --- | --- |
|  | (2) |  |

1. The equation states that the distance between two cluster centroids, ci and cj, is calculated based on the average distances of all elements in cluster i and j to their respective centroids. The values 𝑖, 𝜎𝑖, and 𝜎j represent the average distances, while d(ci, cj) represents the distance between the two centroids. The equation is applicable when there are multiple clusters, and the parameter n denotes the total number of clusters. This equation is referenced from source [38].
2. Calinski-Harabasz index is another method for determining a cluster quality by calculating the variance of the sum of the square of the distance of each point of data compared to the cluster centroid, the higher the index score means the higher the separation and the quality of the cluster with each data inside each cluster are similar and different compared to each cluster, the Calinski-Harabasz index can be calculated using the equation below where k is the number of the cluster from n observation, BCSM as the separation between each cluster, WCSM as the compactness of data inside the cluster. [39]

|  |  |
| --- | --- |
|  | (3) |
|  |  |

# Results and Discussion

After the research process has been done to each of the different models with Davies Bouldin Index, Silhouette Coefficient, and Calinski-Harabasz Index as the metrics to evaluate the performance of the model, the result is as follows.

## DBSCAN

After the model was fine-tuned, the epsilon value of 0.9 was obtained from the highest silhouette score as shown as on Fig. 2. and a min\_samples values of 4 is used.

A picture containing line, diagram, plot, slope

Description automatically generated

1. DBSCAN Silhouette score per epsilon value.

A picture containing screenshot, text, diagram, line

Description automatically generated

1. DBSCAN Clustering Result.

From Fig. 3. Above there are countries that have a high level of similarity that is shown by the overlapping dots, the model also has a silhouette coefficient score of 0.7327, a Davies-Bouldin Index score of 1.881, and a Calinski-Harabasz score of 64.2857.

## k-means

The optimal k value for this model was determined after using the elbow method and the silhouette coefficient which resulted in the following result as shown in Fig. 4, and Fig 5. That shows a value of 6 is the most optimal k value.

A picture containing text, line, plot, diagram

Description automatically generated

1. k-means Elbow Method using WCSS result.

A picture containing line, diagram, plot, text

Description automatically generated

1. k-means Silhouette score per k values.

A picture containing screenshot, text, diagram, line

Description automatically generated

1. k-means clustering result.

The result of the k-means clustering model with 6 k values is as shown as on Fig. 6. And the model also has a silhouette coefficient score of 0.7812, a Davies-Bouldin Index score of 0.3297, and a Calinski-Harabasz score of 541.2260.

## Fuzzy c-means

The C value of 3 was obtained using the result from the silhouette score as shown in Fig. 7.

A picture containing text, diagram, line, plot

Description automatically generated

1. fuzzy c-means Silhouette score per C values.

A picture containing screenshot, text, diagram, line

Description automatically generated

1. fuzzy c-means clustering result.

The Fig. 8. Shows the clustering result using the fuzzy c-means clustering, the model also resulted in a silhouette score of 0.7931, a Davies-Bouldin Index score of 0.8133, and a Calinski-Harabasz index score of 256.7199,

1. Clustering Model Result Comparison

|  |  |  |  |
| --- | --- | --- | --- |
| **Name** | **Silhouette score** | **Calinski-Harabasz score** | **Davies-Bouldin score** |
| DBSCAN | 0.7327 | 64.2857 | 1.8881 |
| k-means | 0.7812 | 541.226 | 0.3297 |
| fuzzy c-means | 0.7931 | 256.7199 | 0.8133 |

The table above shows the result of the clustering model in a table form and was compared to one another to find the most optimal model in this type of research.

1. Calinski-Harabasz score comparation.
2. Davis-Bouldin and Silhouette score comparation.

## Discussion

The Calinski-Harabasz score and Davies-Bouldin score are better alternatives to the Silhouette score for cluster evaluation. Unlike the Silhouette score, they consider both compactness and separation of clusters. The Calinski-Harabasz score provides a global measure of clustering quality, considering the dispersion of data and providing a more robust evaluation. On the other hand, the Davies-Bouldin score considers cluster density and overlap, offering a more comprehensive assessment of cluster dispersion. These metrics are valuable for comparing different clustering models and gaining deeper insights into their performance in various data scenarios.

The result from Table I. or Fig.9. shows that the k-means is the best performing model with a Silhouette coefficient metric of 0.7812, Calinski-Harabasz score of 541.2260, and a Davies-Bouldin score of 0.3297, the result compared to previous study done on [8] with a k-means and fuzzy c-means, using silhouette coefficient metrics resulting 0.726 and 0.896 respectively and a Calinski-Harabasz score of 354.423 for the k-means and 354.423 for the fuzzy c-means, indicated that this research has a better score in k-means while having a lower score in fuzzy c-means.

The first metric is Silhouette score which ranges from 0 to 1, with a higher score indicating better model performance. According to the Silhouette score, the models can be ranked from best to worst as follows: Fuzzy C-means, K-means, and DBSCAN. The Calinski-Harabasz score ranges from 0 to positive infinity, and a larger score indicates better model performance. Contrary to the Silhouette score, the ranking changes when using the Calinski-Harabasz score. According to this metric, the models are ranked as follows: K-means performs the best, followed by Fuzzy C-means, and DBSCAN takes the last position. The Davies-Bouldin score also ranges from 0 to positive infinity, but here, a smaller score indicates better model performance. The ranking based on the Davies-Bouldin score is different from the previous two metrics. According to this evaluation, the models are ranked as follows: K-means is in the first position, Fuzzy C-means comes next, and DBSCAN is placed last.

The result also is supported with previous study [40] which use DBSCAN and k-means clustering to find the relationship with national epidemic strategy response to COVID-19 and shows that the k-means clustering model is the most optimal model.

# Conclusions

In this research we explored the relationship between the features from the dataset by the Global Health Observatory (GHO) using clustering model, the clustering model that was chosen are three popular clustering model, which is DBSCAN, fuzzy c-means, and k-means.

The result from the study showed that the k-means algorithm has the most favorable results in terms of performance which is supported by the evaluating the clustering results using three different evaluation metrics, which are the silhouette metric, with a score of 0.781, and the Calinski-Harabasz index, which a result of 541.226, overall the k-means algorithm is shown to be a suitable algorithm for this task with a high level of separation and compactness in the cluster result that indicates an effectiveness in indicating the underlying pattern and structure of the data.

##### References

1. L. Wulandari and B. Olga Yogantara, “Algorithm Analysis of k-means and fuzzy c-means for Clustering Countries Based on Economy and Health,” vol. 15, no. 2, pp. 1979–276, 2022, doi: 10.30998/faktorexacta.v15i2.12106.
2. N. Shahira Pisal, S. Abdul-Rahman, M. Hanafiah, and S. I. Kamarudin, “PREDICTION OF LIFE EXPECTANCY FOR ASIAN POPULATION USING MACHINE LEARNING ALGORITHMS,” Malaysian Journal of Computing, vol. 7, no. 2, pp. 1150–1161, 2022, doi: 10.24191/mjoc.v7i2.18218.
3. P. Liu, L. Lu, J. Zhang, T. Huo, S. Liu, and Z. Ye, “Application of Artificial Intelligence in Medicine: An Overview,” Current Medical Science, vol. 41, Dec. 2021, doi: 10.1007/s11596-021-2474-3.
4. J. Hutagalung, N. L. W. S. R. Ginantra, G. W. Bhawika, W. G. S. Parwita, A. Wanto, and P. D. Panjaitan, “COVID-19 Cases and Deaths in Southeast Asia Clustering using k-means Algorithm,” in Journal of Physics: Conference Series, IOP Publishing Ltd, Feb. 2021. doi: 10.1088/1742-6596/1783/1/012027.
5. R. Cerqueti and V. Ficcadenti, “Combining rank-size and k-means for clustering countries over the COVID-19 new deaths per million,” Chaos Solitons Fractals, vol. 158, May 2022, doi: 10.1016/j.chaos.2022.111975.
6. A. A. B. Wirayuda and M. F. Chan, “A Systematic Review of Sociodemographic, Macroeconomic, and Health Resources Factors on Life Expectancy,” Asia-Pacific Journal of Public Health, vol. 33, no. 4. SAGE Publications Inc., pp. 335–356, May 01, 2021. doi: 10.1177/1010539520983671.
7. P. Jain, M. S. Bajpai, and R. Pamula, “A Modified DBSCAN Algorithm for Anomaly Detection in Time-series Data with Seasonality,” International Arab Journal of Information Technology, vol. 19, no. 1, pp. 23–28, Jan. 2022, doi: 10.34028/iajit/19/1/3.
8. S. A. Paramita, C. Yamazaki, and H. Koyama, “Determinants of life expectancy and clustering of provinces to improve life expectancy: An ecological study in Indonesia,” BMC Public Health, vol. 20, no. 1, pp. 1–8, 2020, doi: 10.1186/s12889-020-8408-3.
9. M. R. Mahmoudi, D. Baleanu, Z. Mansor, B. A. Tuan, and K. H. Pho, “fuzzy clustering method to compare the spread rate of Covid-19 in the high risks countries,” Chaos Solitons Fractals, vol. 140, Nov. 2020, doi: 10.1016/j.chaos.2020.110230.
10. K. J. Foreman et al., “Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016–40 for 195 countries and territories,” The Lancet, vol. 392, no. 10159, pp. 2052–2090, Nov. 2018, doi: 10.1016/S0140-6736(18)31694-5.
11. H. Wang et al., “Global age-sex-specific fertility, mortality, healthy life expectancy (HALE), and population estimates in 204 countries and territories, 1950–2019: a comprehensive demographic analysis for the Global Burden of Disease Study 2019,” The Lancet, vol. 396, no. 10258, pp. 1160–1203, Oct. 2020, doi: 10.1016/S0140-6736(20)30977-6.
12. Y. Li et al., “Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study,” BMJ, vol. 368, p. l6669, Jan. 2020, doi: 10.1136/bmj.l6669.
13. Y. V. Chudasama et al., “Healthy lifestyle and life expectancy in people with multimorbidity in the UK Biobank: A longitudinal cohort study,” PLOS Medicine, vol. 17, no. 9, p. e1003332, Sep. 2020, doi: 10.1371/journal.pmed.1003332.
14. J. Lelieveld, A. Pozzer, U. Pöschl, M. Fnais, A. Haines, and T. Münzel, “Loss of life expectancy from air pollution compared to other risk factors: a worldwide perspective,” Cardiovascular Research, vol. 116, no. 11, pp. 1910–1917, Sep. 2020, doi: 10.1093/cvr/cvaa025.
15. J. E. Bennett, J. Pearson-Stuttard, V. Kontis, S. Capewell, I. Wolfe, and M. Ezzati, “Contributions of diseases and injuries to widening life expectancy inequalities in England from 2001 to 2016: a population-based analysis of vital registration data,” The Lancet Public Health, vol. 3, no. 12, pp. e586–e597, Dec. 2018, doi: 10.1016/S2468-2667(18)30214-7.
16. S. K. Dini and A. Fauzan, “Clustering Provinces in Indonesia based on Community Welfare Indicators,” EKSAKTA J. Sci. Data Anal., vol. 1, no. 1, pp. 56–63, 2020, doi: 10.20885/eksakta.vol1.iss1.art9.
17. N. K. Schiltz, K. Chagin, and A. R. Sehgal, “Clustering of Social Determinants of Health Among Patients,” J. Prim. Care Community Heal., vol. 13, 2022, doi: 10.1177/21501319221113543.
18. X. Gu et al., “Interactions between life expectancy and the incidence and mortality rates of cancer in China: A population-based cluster analysis,” Cancer Commun., vol. 38, no. 1, pp. 1–15, 2018, doi: 10.1186/s40880-018-0308-x.
19. S. A. Rizvi, M. Umair, and M. A. Cheema, “Clustering of countries for COVID-19 cases based on disease prevalence, health systems and environmental indicators,” Chaos, Solitons and Fractals, vol. 151, p. 111240, 2021, doi: 10.1016/j.chaos.2021.111240.
20. P. Atorkey and C. Owiredua, “Clustering of multiple health risk behaviours and association with socio-demographic characteristics and psychological distress among adolescents in Ghana: A latent class analysis,” SSM - Popul. Heal., vol. 13, p. 100707, 2021, doi: 10.1016/j.ssmph.2020.100707.
21. S. Wang and Z. Ren, “Spatial variations and macroeconomic determinants of life expectancy and mortality rate in China: a county-level study based on spatial analysis models,” Int. J. Public Health, vol. 64, no. 5, pp. 773–783, 2019, doi: 10.1007/s00038-019-01251-y.
22. Z. Rustam and S. Hartini, “Classification of Breast Cancer using Fast fuzzy Clustering based on Kernel,” IOP Conf. Ser. Mater. Sci. Eng., vol. 546, no. 5, 2019, doi: 10.1088/1757-899X/546/5/052067.
23. A. Leogrande, “Munich Personal RePEc Archive Healthy Life Expectancy at Birth in the Italian Regions,” no. 112860, 2022.
24. S. Levantesi, A. Nigri, G. Piscopo, and A. Spelta, “Multi-country clustering-based forecasting of healthy life expectancy,” Qual Quant, Jan. 2023, doi: 10.1007/s11135-022-01611-6.
25. A. Heinz, C. Catunda, C. van Duin, T. Torsheim, and H. Willems, “Patterns of Health-Related Gender Inequalities—A Cluster Analysis of 45 Countries,” Journal of Adolescent Health, vol. 66, no. 6, Supplement, pp. S29–S39, Jun. 2020, doi: 10.1016/j.jadohealth.2020.02.011.
26. U. Rangana Jayathilaka, “International Market Segmentation Of Countries Using Unsupervised Machine Learning Algorithms,” 2021.
27. M. Z. Rodriguez et al., “Clustering algorithms: A comparative approach,” PLOS ONE, vol. 14, no. 1, p. e0210236, Jan. 2019, doi: 10.1371/journal.pone.0210236.
28. Y. Xie and S. Shekhar, “Significant DBSCAN towards Statistically Robust Clustering,” in Proceedings of the 16th International Symposium on Spatial and Temporal Databases, in SSTD ’19. New York, NY, USA: Association for Computing Machinery, Aug. 2019, pp. 31–40. doi: 10.1145/3340964.3340968.
29. K. Zhou and S. Yang, “Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering,” Pattern Analysis and Applications, vol. 23, no. 1, pp. 455–466, Feb. 2020, doi: 10.1007/s10044-019-00783-6.
30. M. Amiruzzaman, R. Rahman, M. R. Islam, and R. M. Nor, “Evaluation of DBSCAN algorithm on different programming languages: An exploratory study,” in 2021 5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2021, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICEEICT53905.2021.9667925.
31. J. R. Valverde et al., “Determinants of inequalities in life expectancy: an international comparative study of eight risk factors,” 2019. [Online]. Available: [www.thelancet.com/](http://www.thelancet.com/)
32. M. Mittal, L. M. Goyal, D. J. Hemanth, and J. K. Sethi, “Clustering approaches for high-dimensional databases: A review,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 9, no. 3. Wiley-Blackwell, May 01, 2019. doi: 10.1002/widm.1300.
33. P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, “Missing Data in Clinical Research: A Tutorial on Multiple Imputation,” Canadian Journal of Cardiology, vol. 37, no. 9. Elsevier Inc., pp. 1322–1331, Sep. 01, 2021. doi: 10.1016/j.cjca.2020.11.010.
34. G. E. Hodes and C. N. Epperson, “Sex Differences in Vulnerability and Resilience to Stress Across the Life Span,” Biological Psychiatry, vol. 86, no. 6. Elsevier USA, pp. 421–432, Sep. 15, 2019. doi: 10.1016/j.biopsych.2019.04.028.
35. S. Fan et al., “Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data,” Anal Chem, vol. 91, no. 5, pp. 3590–3596, Mar. 2019, doi: 10.1021/acs.analchem.8b05592.
36. K. R. Shahapure and C. Nicholas, “Cluster Quality Analysis Using Silhouette score,” in 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA), Oct. 2020, pp. 747–748. doi: 10.1109/DSAA49011.2020.00096.
37. S. Surono and R. D. A. Putri, “Optimization of fuzzy c-means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analysis,” Int. J. fuzzy Syst., vol. 23, no. 1, pp. 139–144, Feb. 2021, doi: 10.1007/s40815-020-00997-5.
38. G. Liu, “A Optimization of fuzzy c-means Clustering Algorithm with Combination of Minkowski and Chebyshev Distance Using Principal Component Analys estimation.” arXiv, Sep. 29, 2022. doi: 10.48550/arXiv.2207.01294.
39. A. Rachwał et al., “Determining the Quality of a Dataset in Clustering Terms,” Applied Sciences, vol. 13, no. 5, Art. no. 5, Jan. 2023, doi: 10.3390/app13052942.
40. Y. Shuai, C. Jiang, X. Su, C. Yuan and X. Huang, "A Hybrid Clustering Model for Analyzing COVID-19 National Prevention and Control Strategy," 2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE), Beijing, China, 2020, pp. 68-71, doi: 10.1109/ICCSSE50399.2020.9171941.