Khatulistiwa: Jurnal Pendidikan dan Sosial Humaniora
Vol.3, No.1 Maret 2023
e-ISSN: 2962-4037; p-ISSN: 2962-4452, Hal 142-150

Komparasi Performa Fuzzy C-Means dan Random Forest (Studi Kasus: Indeks Modal Sosial Indonesia)

Pardomuan Robinson Sihombing
BPS-Statistics Indonesia

Ade Marsinta Arsani
BPS-Statistics Indonesia

Wisnu Pratiko
BPS-Statistics Indonesia

Sri Murtiningsih
BPS-Statistics Indonesia

Jl. Dr. Sutomo 6-8 Jakarta 10710 Indonesia
Email: robinson@bps.go.id

Abstract*. This study aims to test the performance of the Fuzzy C-Means Cluster method with Random Forest Clustering. The data used is the dimension data for the Social Capital Index in 34 Provinces in Indonesia in 2021. The social capital index consists of three dimensions, namely the dimensions of Trust, Social Participation and Tolerance. Data sourced from the BPS- Statistics Indonesia. The optimal number of clusters suggested by using the Elbow method technique is 3 clusters. By paying attention to the largest Silouhette and R square values, the random forest method is better than Fuzzi C-Means. The same thing is seen based on the criteria of smaller AIC and BIC values, the random forest model is better than fuzzy c-means. Cluster 3 is the cluster with the best dimension values where all values are positive or above the average. On the other hand, cluster 1 is the province with the worst dimension values because all dimension values are negative, below the average data. Comprehensive and targeted policies are needed so that the social capital index in Indonesia can be evenly distributed and increases every year.*

Keywords*: Fuzzy C-Means, Clusters, Random Forest Social Capital, Silouhette*

Abstrak. Penelitian ini bertujuan menguji performa metode Fuzzy C-Means Klaster dengan Random Forest Clustering. Data yang digunakan data dimensi Indeks Modal Sosial di 34 Provinsi di Indonesia tahun 2021. Indeks mdoal social terdiri atas tiga dimensi yaitu dimensi Rasa Percaya, Partisipasi Sosial dan Toleransi. Data bersumber dari Badan Pusat Statistik (BPS). Banyaknya klaster optimum yang disarankan dengan menggunakan teknik metode Elbow adalah sebanyak 3 klaster. Dengan memperhatikan nilai Silouhette dan R square terbesar metode random forest lebih baiknya daripada Fuzzi C-Means. Hal senada jika dilihat berdasarkan kriteria nilai AIC dan BIC yang lebih kecil model random forest lebih baik daripada fuzzy c-means. Klaster 3 merupakan klaster dengan nilai dimensi terbaik dimana nilainya semuanya di positif atau di atas rata-rata. Di sisi lain klaster 1 merupakan provinsi dengan nilai dimensi terburuk karena semua

Received Januari 04, 2023; Revised Januari 09, 2023; Januari 15, 2023
*Pardomuan Robinson Sihombing, robinson@bps.go.id

nilai dimensinya negative, di bawah rata-rata data. Dibutuhkan kebijakan yang komprehensif dan tepat sasaran sehingga indeks modal social di Indonesia dapat merata dan meningkat setiap tahunnya.

Kata kunci: 3 Fuzzy C-Means, Klaster, Modal Sosial Random Forest, Silouhette

LATAR BELAKANG

Modal sosial merupakan suatu variable yang berhubungan dengan kehidupan suatu masyrakat. Fukuyama (1995) mendefenisikan modal sosial sebagai serangkaian nilai atau norma-norma informal yang dimiliki bersama diantara para anggota suatu kelompok masyarakat yang saling terkait, yang didasarkan pada nilai kepercayaan, norma, dan jaringan sosial. Modal sosial ini merupakan suatu kapabilitas yang muncul dari kepercayaan di dalam sebuah masyarakat secara umum. Dalam indeks modal sosial terdapat tiga dimensi yaitu Rasa Percaya, Partisipasi Sosial dan Toleransi.

Pada tahun 2005, Badan Pusat Statistik (BPS) menginisiasi penyusunan pengukuran modal sosial di Indonesia. Studi modal sosial yang dilaksanakan BPS pada tahun 2005 dan 2006 bertujuan untuk mengkaji kelayakan variabel yang secara teoritis merupakan determinan modal sosial. Studi terus berlanjut pada tahun 2007, 2009, 2014, 2017 dan publikasi terbaru tahun 2022.

Untuk terus meningkatkan modal sosial, pemerintah melakukan berbagai usaha dan kebijakan. Tentu saja usaha dan kebijakan yang dilakukan harus tetap sasaran dengan melihat masing-masing kondisi suatu wilayah. Oleh karena ini perlu dilakukan pengelompokan wilayah berdasarkan nilai dimensi sosial masing-masing dimensi sehingga kebijakan yang dilakukan sesuai dengan kebutuhan masing-masing wilayah.

Salah satu analisis dalam statistik yang digunakan untuk mengelompokkan subjek adalah analisis klaster (gerombol/ cluster). Analisis klaster termasuk bagian dalam analisis multivariat (Rencher & Christensen, 2012). Pada umumnya analisis klaster dapat dibagi dua yaitu klaster hirarki dan klaster non hirarki (Johnson & Dean, 2008). Klaster hirarki melakukan pengelompokan data berdasarkan kemipiran jarak maupun korelasi antar variabel. Sedangkan dalam klaster non hirarki peneliti sudah menetukan terlebih dahulu jumlah klaster yang diinginkan.

Klaster nonhirarki secara klasik pada umumnya menggunkan metode k-means dan k-median. Pengembangan selanjutnya berbasis *machine learning* seperti fuzzy c-means

dan random forest. Perbandingan metode machine learning dan metode klasik sudah dilakukan beberapa peneliti. Misalnya Syarif et al (2018) membandingan Algoritme K-Means dengan Algoritme Fuzzy C Means (FCM) dalam Clustering Moda Transportasi Berbasis GPS. Dari hasil pengujian, disimpulkan bahwa algoritma FCM lebih unggul dibandingkan K-Means. Alzubaidi et al. (2018) menerapkan metode random forest pada clusterimng data. Hasilnya dapat meningkatkan performa clustering. Yu et al.(2021) membandingan Algoritme K-Means dengan Randoms Forest (RFC) dalam megevaluasi kandungan gas reservoir metana batubara. Dari hasil pengujian, disimpulkan bahwa algoritma RFC lebih unggul dibandingkan K-Means. Sokoutia et al. (2015) membandingkan performa random forest dan fuzzy c-means. Hasil penelitian menunjukkan bahwa performa RFC lebih baik dari FCM.

Berdasarkan permasalahan di atas dimana masih terdapat gap hasil penelitian antara metode klaster. Peneliti tertarik menguji performa random forest dan Fuzzy C-Means dalam mengelompokkan provinsi-provinsi di Indonesia berdasarkan data 3 dimensi modal sosial tahun 2021.

METODE PENELITIAN

Data yang digunakan dalam penelitian ini berasal dari publikasi Badan Pusat Statistik (BPS, 2022). Di mana data indeks modal sosial menggunakan tiga dimensi yaitu dimensi Rasa Percaya, Partisipasi Sosial dan Toleransi. Karena semua variabel memiliki satuan yang sama maka tidak perlu dilakukan tranformasi data baik menggunakan logaritma maupun nilai standar data (*z score*).

Fuzzy C-means (FCM)

Fuzzy C-Mean merupakan pengembangan dari KMeans dengan menggabungkan prinsip *fuzzy* dengan metode K-Means. Perbedaannya dimaan data yang di-*cluster* menggunakan FCM akan menjadi aggota dari setiap *cluster* yang ada. Ikatan data dengan *cluster* ditentukan oleh nilai keanggotannya yang berada pada rentang 0 hingga 1. Pada model FCM, tingkat keberadaan data dalam suatu kelas atau cluster ditentukan oleh derajat keanggotaannya (Mas'udin et al., 2018)

Random Forest

Model random forest merupakan pengembangan dari decision tree (pohon keputusan). Setiap pohon memiliki banyak node yang diatur secara hierarkis, dimana informasi ditransfer dari arah atas ke bawah (Alzubaidi et al., 2018). Random Forest memiliki keuntungan karena tidak sensitif terhadap data; tidak ada masalah overfitting dan dapat mengurutkan variabel yang berkontribusi pada prediksi.

Elbow Method

Elbow method merupakan metoda yang sering dipakai untuk menentukan jumlah klaster yang akan digunakan pada k-means clustering dengan melihat persentase hasil perbandingan antara jumlah cluster yang akan membentuk siku pada suatu titik (Madhulatha, 2012). Pada umumnya hasil persentase yang berbeda dari setiap nilai cluster dapat ditunjukan dengan menggunakan grafik sebagai sumber informasinya.

Kriteria Pemilihan Model

Pada penelitian ini pemilihan model didasarkan pada kriteria Silhouette (Struyf et al., 1997), dimana dipilih model yang nilai terbesar. Adapun formulanya yang digunakan:

$$s(i) = \frac{b(i)-a(i)}{max\{a(i)-b(i)\}} \quad (1)$$

dengan: $a(i) := \frac{\sum_{j \epsilon A, j \neq 1} d(i,j)}{|A|^{-1}}$ *(2)*

$$d(i) := \frac{\sum_{j \epsilon C,} d(i,j)}{|C|} \quad (3)$$

$$b(i) := \min_{C \neq A} d(i,j) \quad (4)$$

di mana,

A = banyaknya data di klaster A

d(i)= jarak

b(i) =nilai minimum dari jarak rata-rata data ke-*i* dengan semua data di klaster berbeda.

C = banyaknya data di klaster C

Selain itu digunakan kriteria error meliputi AIC (Akaike, 1974) dan BIC (Gideon Schwarz, 1978) dan koefisien determinasi (R^2). Model terbaik adalah model yang

memiliki nilai AIC dan BIC terkecil (Widarjono, 2007) dan koefisien determinasi terbesar (Gujarati, 2004). Adapun formula yang digunakan adalah:

$$AIC = -2\,L(\hat{\theta}) + 2p \tag{5}$$

$$BIC = -2\,L(\hat{\theta}) + p\ln(n) \tag{6}$$

$$R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \tag{7}$$

dengan $L(\hat{\theta})$ adalah nilai likelihood, dan p adalah jumlah parameter yang akan diestimasi termasuk konstanta, n adalah jumlah sampel. Nilai $\hat{Y}$ adalah nilai prediksi variabel dependen dari model, dan Y adalah nilai observasi variabel dependen.

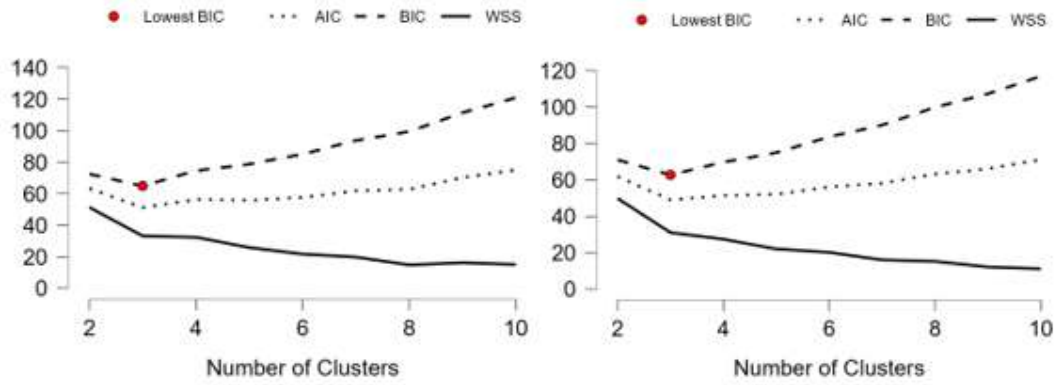
HASIL DAN PEMBAHASAN

Sebelum lebih lanjut membahas pengelompokan provinsi dilakukan analisis deskriptif pada variabel penelitian. Pada Tabel 1 secara rata-rata nilai Dimensi Rasa Percaya sebesar 74,276 poin, dengan nilai terendah sebesar 67.970 poin pada Provinsi Papua, dengan nilai tertinggi sebesar 80,740 poin pada Provinsi Sulawesi Utara. Secara rata-rata nilai Dimensi Partisipasi Sosial sebesar 75,369 poin, dengan nilai terendah sebesar 69,960 poin pada Provinsi banten, dengan nilai tertinggi sebesar 80,260 poin pada Provinsi Sulawesi Utara. Secara rata-rata nilai Dimensi Toleransi sebesar 59,964 poin, dengan nilai terendah sebesar 46,950 poin pada Provinsi Aceh, dengan nilai tertinggi sebesar 77,060 poin pada Provinsi Kalimantan Utara. Variasi nilai dimensi terbesar pada Dimensi Toleransi dengan nilai standar deviasi sebesar 5,791.

Tabel 1. Statistik Deskriptif Variabel Penelitian

Variabel	Rasa Percaya	Partisipasi Sosial	Toleransi
Rata-rata	74,276	75,369	59,964
Standart Deviasi	3,240	2,649	5,791
Minimum	67,970	69,960	46,950
Maksimum	80,740	80,260	77,060

Pada gambar 1 dapat dilihat banyaknya klaster optimum yang disarankan dengan metode Elbow dengan kriteria BIC terkecil. Dengan menggunakan metode Elbow Plot terpilih sebanyak 3 klaster baik pada metode Fuzzy Means maupun Random Forest.



Gambar 1. Pemilihan Banyaknya Klaster Pada K-Means dan Fuzzy Means dengan Plot Elbow

Pada Tabel 2 terlihat banyaknya anggota tiap klaster. Pada klaster fuzzy c-means terdapat 3 Klaster yang masing-masing anggotara terdiri atas 11, 13 dan 10 provinsi. Sedangkan pada random forest cluster terdapat 3 klaster yang anggotanya masing-masing terdiri atas 14, 15 ddan 5 provinsi.

Tabel 2. Banyaknya Anggota Klaster

Klaster	1	2	3
Fuzzy c-means	11	13	10
Random Forest	14	15	5

Selanjutnya dilakukan pemilihan model terbaik dengan membandingkan kriteria dari masing-masing metode yaitu indeks Silhouette, kriteria error (AIC dan BIC) dan koefisien determinaasi R^2. Jika dilihat dari nilai indeks Silhouette maka model cluster random forest memiliki nilai yang lebih besar yaitu 0,39. Sedangkan jika dinilai dari nilai AIC dan BIC maka metode random forest memiliki nilai yang lebih kecil yaitu sebesar 49,08 dan 62,82 sedangkan fuzzy c-mean sebesar 95,48 dan 109,22. Di sisi lain jika dilihat dari R^2 maka metode random forest memliki nilai yang lebih besar daripada fuzzy c-mean yaitu sebesar 0.686. Sehingga dapat dikatakan model random forest lebih baik daripada model Fuzzy C-Means dalam mengelompokkan provinsi di Indonesia

berdasarkan nilai dimensi indeks modal sosial karena nilai AIC dan BIC yang lebih kecil dan indeks Silhouettes R^2 yang lebih besar.

Tabel 3. Kriteria Pemilihan Model Terbaik

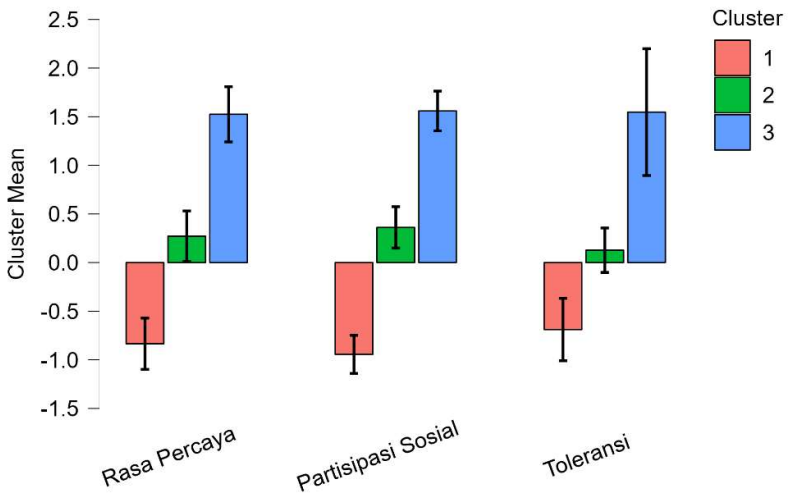
Klaster	R^2	AIC	BIC	Silhouette
Fuzzy c-means	0,032	95,48	109,22	0,07
Random Forest	0,686	49,08	62,82	0,39

Pada Tabel 4 terlihat nilai rata-rata masing-masing variabel per klaster. Nilai positif menunjukkan bahwa nilai variabel kelompok dalam klaster itu di bawah rata-rata keseluruhan data, sedangkan nilai negatif menunjukkan bahwa nilai variabel kelompok itu di atas rata-rata keseluruhan data. Klaster 1 memiliki nilai negatif untuk ketiga dimensi. Anggota pada klaster ini adalah Aceh, Sumatera Utara, Sumatera Barat, Riau, Bengkulu, Kep. Bangka Belitung, DKI Jakarta, Jawa Barat, Jawa Timur, Banten, Nusa Tenggara Barat, Kalimantan Selatan, Sulawesi Selatan, Papua. Hal ini mengindikasikan bahwa provinsi ini masih memiliki modal sosial yang sangat rendah. Di sisi lain Klaster 3 memiliki nilai positif untuk ketiga dimensi. Anggota pada klaster ini adalah Nusa Tenggara Timur, Kalimantan Utara, Sulawesi Utara, Sulawesi Tengah, Maluku. Hal ini mengindikasikan bahwa provinsi-provinsi ini sudah sangat baik dari dari semua aspek indeks modal sosial.

Tabel 4. Klaster Means Masing-Masing Variabel

Klaster Mean	Rasa Percaya	Partisipasi Sosial	Toleransi
Cluster 1	-0.835	-0.944	-0.689
Cluster 2	0.271	0.361	0.127
Cluster 3	1.525	1.559	1.547

Klaster 2 juga memiliki nilai yang positif pada semua dimensi tetapi nilainya di bawah klister 1. Anggota klaster ini termasuk provinsi Jambi, Sumatera Selatan, Lampung, Kepulauan Riau, Jawa Tengah, DI Yogyakarta, Bali, Kalimantan Barat, Kalimantan Tengah, Kalimantan Timur, Sulawesi Tenggara, Gorontalo, Sulawesi Barat, Maluku Utara, Papua Barat.



Gambar 2. Cluster Means

Selanjutnya didefeniskan urutan dimensi terpengting yang membentuk klaster. Pada Tabel 5 bahwa dimensi terpenting adalah partisipasi sosial dengan nilai ineks sebesar 11,033. Sedangkan dimensi terkecil adalah toleransi dengan gini indek 10,286.

Tabel 5. Feature Importance

Dimensi	Mean decrease in Gini Index
Partisipasi Sosial	11,033
Rasa Percaya	11,011
Toleransi	10,286

KESIMPULAN DAN SARAN

Banyaknya klaster optimum yang disarakankan dengan menggunakan teknik metode Elbow dengan kriteria BIC adalah sebanyak 3 klaster baik pada metode random forest maupun Fuzzy C-Means. Dengan memperhatikan nilai Silouhette dan R square terbesar metode random foerst lebih baik dari Fuzzi C-Means karena nilainya lebih besar. Hal senada jika dilihat berdasarkan kriteria nilai AIC dan BIC yang lebih kecil model random forest lebih baik dari fuzzy time series.

Dari ketiga klaster yang terbentuk klaster 1 memiliki nilai means negatif yang artinya nilai semua dimensi anggota klaster di bawah nilai rata-rata seluruh data. Sedangkan klaster 2 dan 3 memiliki nilai means positif artinya nilai semua dimensi anggota klaster di atas nilai rata-rata seluruh data.

Untuk penelitian selanjutnya dapat menambahkan variabel potensial lainnya seperti kemiskinan, faktor lingkungan hidup, factor budaya dan lainnya. Dari sisi metode dapat menambah metode k-means klsster, k-median klsster dan k-harmonik klaster..

DAFTAR REFERENSI

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alzubaidi, L., Arkah, Z. M., & Hasan, R. I. (2018). Using random forest algorithm for clustering. *Journal of Engineering and Applied Sciences*, *13*(21), 9189–9193. https://doi.org/10.3923/jeasci.2018.9189.9193

BPS. (2022). *Statistik Modal Sosial 2021*.

Fukuyama, F. (1995). *Trust: The Social Virtues and the Creation of Prosperity*. Free Press.

Gideon Schwarz. (1978). Estimating The Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Gujarati, D. (2004). *Basic Econometrics BY Gujarati* (pp. 1–1002). McGraw-Hill Inc.

Madhulatha, T. . (2012). An Overview On Clustering Methods. *IOSR Journal Engineering*, *2*(4), 719–725.

Mas'udin, P. E., Farida, A., & Mustafa, L. D. (2018). ClusteringData RemunerasiDosen Untuk Penilaian Kinerja Menggunakan Fuzzy c-Means. *Jurnal Resti*, 288–294.

Sokoutia, B., Rezvanb, F., & Dastmalchi, S. (2015). Applying random forest and subtractive fuzzy c-means clustering techniques for developing a novel G protein-coupled receptors discrimination method using pseudo amino acid composition. *Molecular BioSystems*, *3*, 2–11. https://doi.org/10.1039/b000000x

Struyf, A., Hubert, M., & P. J. Rousseeuw. (1997). Clustering in an Object-Oriented Environment. *Journal of Statictical Software*, *1*(4), 1–30.

Widarjono, A. (2007). *Ekonometrika: Teori dan Aplikasi untuk Ekonomi dan Bisnis*. Ekonosia Fakultas Ekonomi Universitas Islam Indonesia.

Yu, J., Zhu, L., Qin, R., Zhang, Z., Li, L., & Huang, T. (2021). Combining k-means clustering and random forest to evaluate the gas content of coalbed bed methane reservoirs. *Geofluids*, *2021*. https://doi.org/10.1155/2021/932156