Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

Implementasi Big Data Analytics Untuk Deteksi Pola Penyalahgunaan Bantuan Sosial Menggunakan Metode Clustering

R. Mahdalena Simanjorang[1*], Agustina Simangunsong[2], Dini Auliah[3], Rabbiatul Adawiyah[4]
[1,2,3,4]Teknologi Informasi, STMIK Pelita Nusantara, Medan, Indonesia
Email: [1] lenasinaga30@gmail.com, [2] agustinasimangunsong93@gmail.com , [3] diniauliah@gmail.com , [4]rabbiatuladawiyah14@icloud.com
Email Penulis Korespondensi: [1] lenasinaga30@gmail.com

Info Artikel

Kata Kunci:
Big Data Analytics
Clustering
Bantuan Sosial
Deteksi Pola
DTKS

Abstrak

Kajian ini bertujuan untuk menerapkan pendekatan Big Data Analytics guna mengidentifikasi pola penyimpangan dalam penyaluran bantuan sosial dengan memanfaatkan algoritma K-Means Clustering terhadap 10.000 data penerima dari Data Terpadu Kesejahteraan Sosial (DTKS). Proses analisis dilakukan melalui kerangka Knowledge Discovery in Databases (KDD) yang mencakup tahapan kurasi data, pembersihan data, transformasi data, pengelompokan, serta evaluasi hasil menggunakan bantuan perangkat Python dan RapidMiner. Temuan penelitian menunjukkan bahwa algoritma K-Means Clustering mampu mempartisi penerima bantuan ke dalam tiga kelompok berbeda. Kelompok pertama yang terdiri dari 4.570 data (45,7%) merupakan penerima yang sah dengan ciri pendapatan terbatas dan kepemilikan aset rendah. Kelompok kedua mencakup 3.200 data (32%) yang berpotensi mengalami kesalahan data dengan kondisi ekonomi menengah. Sementara itu, kelompok ketiga sebanyak 2.060 data (20,6%) terdeteksi sebagai kelompok dengan indikasi penyalahgunaan, yang dicirikan oleh pendapatan rata-rata mencapai Rp6.500.000 serta kepemilikan aset yang tinggi. Proses validasi dengan Silhouette Score memperoleh nilai 0,68, sementara Davies-Bouldin Index menghasilkan angka 0,62, yang keduanya menandakan bahwa struktur cluster yang terbentuk tergolong memadai. Penelitian ini memberikan sumbangsih bagi pengembangan Big Data Analytics dalam sistem pengambilan keputusan di sektor kesejahteraan sosial, terutama dalam upaya mendeteksi potensi penyalahgunaan program bantuan.

Keywords:
Big Data Analytics
Clustering
Social Assistance
Pattern Detection
DTKS

Abstract

This study aims to apply a Big Data Analytics approach to identify patterns of irregularities in social assistance distribution by utilizing the K-Means Clustering algorithm on 10,000 recipient records from the Integrated Social Welfare Data (DTKS). The analytical process was conducted through the Knowledge Discovery in Databases (KDD) framework, encompassing data curation, data cleaning, data transformation, clustering, and result evaluation using Python and RapidMiner tools. The findings reveal that the K-Means Clustering algorithm successfully partitioned social assistance recipients into three distinct groups. The first group, consisting of 4,570 records (45.7%), represents eligible recipients characterized by limited income and low asset ownership. The second group comprises 3,200 records (32%) with potential data errors and middle economic characteristics. Meanwhile, the third group of 2,060 records (20.6%) was identified as having indications of misuse, marked by an average income of Rp6,500,000 and high asset ownership. Validation using the Silhouette Score yielded a value of 0.68, while the Davies-Bouldin Index produced 0.62, both indicating that the cluster structures formed are reasonably adequate. This research contributes to the development of Big Data Analytics in decision support systems within the social welfare sector, particularly in detecting potential misuse of assistance programs.

1. PENDAHULUAN

Big Big Data Analytics telah menjadi fondasi penting dalam transformasi digital di berbagai sektor, termasuk dalam pengelolaan program bantuan sosial [1]. Seiring dengan meningkatnya volume data penerima bantuan yang dihasilkan dari berbagai sumber seperti Data Terpadu Kesejahteraan Sosial (DTKS), sistem pengelolaan data tradisional tidak lagi

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

mampu menangani kompleksitas dan ukuran data yang terus berkembang [2]. Big Data Analytics menawarkan kemampuan untuk memproses, menganalisis, dan mengekstrak informasi berharga dari kumpulan data berukuran besar yang sebelumnya sulit dikelola dengan metode konvensional [3]. Tanpa dukungan Big Data Analytics, upaya memastikan bantuan sosial tepat sasaran dan dikelola secara akuntabel akan sulit terwujud di tengah kompleksitas data yang ada.

Di tengah berbagai upaya pengurangan kemiskinan, bantuan sosial hadir sebagai instrumen utama pemerintah untuk meningkatkan kesejahteraan masyarakat[4]. Beragam bantuan yang disalurkan mulai dari uang tunai, sembako, subsidi pendidikan dan kesehatan, hingga pelatihan keterampilan menjadikan program ini sebagai penyangga utama bagi jutaan keluarga yang terbatas secara ekonomi [5]. Namun, pelaksanaan bantuan sosial sering kali dihambat oleh berbagai tantangan administratif dan logistik [6]. Volume penerima yang sangat besar, alokasi sumber daya yang terbatas, serta kebijakan yang kompleks menjadi beberapa masalah yang sering dihadapi dalam penyaluran bantuan.

Di tingkat implementasi, masalah paling mendasar adalah buruknya kualitas data penerima tidak akurat dan sudah kadaluwarsa yang berujung pada kesalahan penyaluran bantuan [7]. Terbatasnya akses terhadap teknologi pemutakhiran data penduduk semakin memperparah kondisi tersebut, sehingga bantuan tidak menjangkau kelompok masyarakat yang paling membutuhkan[8]. Penyaluran bantuan yang masih bergantung pada rekomendasi perangkat desa setempat juga dinilai kurang efisien dan berpotensi menimbulkan subjektivitas dalam penentuan penerima manfaat [9].

Penelitian terdahulu telah membuktikan efektivitas pendekatan berbasis Big Data dalam mendukung ketepatan penyaluran bantuan sosial. Amaliyah dkk. [10] mengelompokkan data kepala keluarga dan menemukan bahwa sejumlah keluarga termasuk dalam kategori prioritas namun belum tercatat sebagai penerima bantuan. Sari Ufriani dkk. [11] mengelompokkan data masyarakat dan mengidentifikasi warga yang layak menerima prioritas namun tidak terdaftar. Fitriani dkk. [12] menerapkan pendekatan clustering pada Program Keluarga Harapan (PKH) dan menghasilkan klaster dengan kualitas yang cukup baik. Penelitian lain oleh Daulay dan Wandri [6] menggunakan kombinasi metode clustering dan klasifikasi untuk seleksi penerima beasiswa dan berhasil mencapai akurasi yang tinggi. Sompa dan Ishak [13] juga menerapkan pendekatan serupa untuk mengelompokkan calon penerima beasiswa berdasarkan tingkat ekonomi. Bengnga dan Ishak [14]menggabungkan XGBoost untuk seleksi atribut dengan K-Means dan berhasil mengidentifikasi klaster dengan nilai DBI yang baik.

Meskipun penelitian-penelitian tersebut telah menunjukkan hasil yang positif, sebagian besar masih menggunakan data dalam skala terbatas dan belum memanfaatkan pendekatan Big Data Analytics secara optimal [15]. Selain itu, penelitian sebelumnya lebih fokus pada pengelompokan penerima bantuan tanpa melakukan analisis mendalam terhadap pola penyalahgunaan yang mungkin terjadi dalam skala big data [1]. Hal ini menciptakan kesenjangan antara kebutuhan pengelolaan data bantuan sosial yang semakin besar dengan kemampuan teknologi yang diterapkan.

Berdasarkan uraian di atas, terdapat gap analysis antara penelitian-penelitian sebelumnya dengan kebutuhan saat ini. Penelitian terdahulu belum memanfaatkan pendekatan Big Data Analytics secara optimal pada dataset yang besar, serta belum melakukan analisis pola penyalahgunaan secara komprehensif. Penelitian ini menawarkan nilai kebaruan melalui implementasi Big Data Analytics yang dipadukan dengan algoritma K-Means Clustering pada dataset yang relatif lebih besar, serta analisis komprehensif terhadap pola penyimpangan bantuan sosial yang ditelaah dari sudut pandang karakteristik sosial dan ekonomi para penerima dalam skala big data.

Penelitian ini memiliki dua sasaran utama. Pertama, mengimplementasikan Big Data Analytics melalui algoritma K-Means Clustering untuk menelusuri pola-pola penyimpangan dalam penyaluran bantuan sosial yang terekam dalam Data Terpadu Kesejahteraan Sosial (DTKS), sekaligus menguji tingkat ketepatan metode tersebut. Kedua, mendorong perbaikan efisiensi dan percepatan alur distribusi dana bantuan sosial dengan mengoptimalkan pemanfaatan teknologi Big Data Analytics.

2. METODOLOGI PENELITIAN

Penelitian ini mengadopsi pendekatan Big Data Analytics dengan mengusung metode deskriptif kuantitatif sebagai kerangka analisisnya. Proses analisis data pada penelitian ini mengandalkan algoritma K-Means Clustering dengan berpedoman pada pendekatan Knowledge Discovery in Database (KDD), sebuah framework yang sering digunakan dalam Big Data Analytics.

2.1 Pengumpulan dan persiapan Data

Sumber data utama dalam penelitian ini adalah 10.000 rekaman penerima bantuan sosial yang berasal dari Data Terpadu Kesejahteraan Sosial (DTKS). Data terdiri dari berbagai atribut seperti NIK, nama, kecamatan, status rumah, luas bangunan, jenis lantai, sumber air, kepemilikan aset, jumlah tanggungan, status pekerjaan, pendapatan per bulan, desil,

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

program diterima, dan status verifikasi. Knowledge Discovery in Database (KDD) menjadi acuan utama dalam proses pengolahan data big data pada penelitian ini.

Tabel 1. Sampel Data Penerima Bantuan Sosial

No	Nama Kepala Keluarga	Usia	Status Perkawinan	Pekerjaan	Pendapatan (Rp)	Tanggungan	Status Verifikasi
1	Entin Kartini	66	Cerai Mati	Mengurus Rumah Tangga	500.000	2	Sah
2	Ujang Ahmad Solihin	40	Kawin Belum Tercatat	Belum/Tidak Bekerja	0	2	Bermasalah
3	Nani Rochani	75	Cerai Mati	Mengurus Rumah Tangga	500.000	3	Sah
4	Jojo Warjo Permana	65	Kawin Belum Tercatat	Buruh Harian Lepas	1.000.000	2	Sah
5	Ade Sutardi	58	Kawin Belum Tercatat	Buruh Harian Lepas	1.000.000	4	Bermasalah

Sumber: Hasil Peneliti, 2025

2.2 Pra-pemrosesan Data Big Data

Tahap seleksi data dilakukan untuk menghilangkan data yang tidak diperlukan dalam analisis, sehingga diperoleh atribut yang relevan berupa data numerik dan kategorikal . Tahap pembersihan data bertugas menyingkirkan rekaman ganda berdasarkan NIK sekaligus menangani nilai-nilai yang tidak terisi (*missing values*). Dari total 10.000 data awal, proses ini menyisakan 9.830 data yang dinyatakan layak untuk diproses. Setelah itu, data ditransformasi ke dalam bentuk numerik dengan melakukan *encoding* terhadap variabel kategorikal dan normalisasi *Min-Max Scaling*—sebuah langkah standar yang lazim diterapkan dalam analisis Big Data.

Tabel 2. Fitur yang Digunakan dalam Big Data Analytics

No	Jenis Fitur	Nama Fitur	Keterangan
1	Numerik	Luas Bangunan	Luas bangunan tempat tinggal (m^2)
2	Numerik	Kepemilikan Aset	Skor kepemilikan aset (0-10)
3	Numerik	Jumlah Tanggungan	Jumlah anggota keluarga
4	Numerik	Pendapatan Bulan	Pendapatan per bulan (Rp)

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

No	Jenis Fitur	Nama Fitur	Keterangan
5	Numerik	Desil	Tingkat kesejahteraan (1-10)
6	Kategorikal	Status Rumah	Milik/Sewa/Kontrak
7	Kategorikal	Jenis Lantai	Tanah/Semen/Keramik/Marmer
8	Kategorikal	Sumber Air	Tidak Layak/Layak
9	Kategorikal	Status Pekerjaan KK	Pekerjaan kepala keluarga
10	Kategorikal	Kecamatan	Wilayah tempat tinggal
11	Kategorikal	Program Diterima	Jenis program bantuan

Sumber: Hasil Peneliti, 2025

2.3 Reduksi Dimensi dengan PCA pada Big Data

Principal Component Analysis (PCA) diterapkan untuk mereduksi dimensi data, menyusutkan 11 fitur menjadi 8 komponen utama yang mampu menerangkan 91,2% dari total varians yang ada. Hal ini dilakukan untuk mengurangi kompleksitas komputasi tanpa kehilangan informasi penting dalam proses clustering pada big data, seperti yang juga diterapkan dalam penelitian big data clustering oleh Sreedhar dkk. [7] dan Zerhari dkk. [13].

2.4 K-Means Clustering untuk Analisis Big Data

K-Means merupakan algoritma clustering yang mengelompokkan data ke dalam kelompok-kelompok yang saling terpisah. Penerapannya melalui enam tahapan: (1) menentukan jumlah cluster yang paling sesuai dengan metode Elbow dan Silhouette Score; (2) menginisialisasi centroid awal secara acak; (3) mengukur jarak masing-masing data ke setiap centroid dengan Euclidean Distance; (4) menempatkan data ke cluster berdasarkan jarak paling pendek; (5) menyesuaikan posisi centroid sesuai rata-rata anggota cluster; serta (6) mengulangi proses dari tahap 3 hingga 5 sampai tidak ditemukan lagi perubahan anggota cluster.

Untuk mengukur jarak antara data dan centroid, penelitian ini menggunakan formula Euclidean Distance yang disajikan sebagai berikut:

$$\mathrm{d}(x_i, c_k) = \sqrt{\sum_{j=1}^{m} (x_{ij} - c_{kj})^2} \qquad (1)$$

Keterangan:
a) $d(xi, ck)$= besaran jarak yang dihitung dari data ke-i menuju centroid cluster ke-k
b) xij = nilai spesifik data ke-I pada fitur ke-J
c) ckj = nilai centroid cluster ke-k pada fitur ke-j
d) m = jumlah fitur yang digunakan

Setelah semua data dialokasikan ke cluster terdekat, centroid baru dihitung menggunakan rumus:

$$c_k = \frac{1}{N_k} = \sum_{i=1}^{N_k} \mathcal{X}_i^{(k)} \qquad (2)$$

Keterangan:
a) c_k : centroid hasil pembaruan untuk cluster ke $-\ k$
b) N_k : total data yang menjadi anggota cluster ke- k
c) $\mathcal{X}_i^{(k)}$: data ke- i yang termasuk dalam cluster ke- k

Proses ini diulang secara iteratif hingga tidak terjadi perubahan anggota cluster atau mencapai konvergensi. Pendekatan ini sejalan dengan penelitian Jin [11]tentang analisis big data perilaku siswa dan penelitian Haripriya dkk. [1] tentang big data clustering pada data kesehatan.

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

2.5 Validasi Clustering pada Big Data

Penelitian ini mengevaluasi kualitas hasil clustering dengan menerapkan dua metode validasi yang umum digunakan dalam lingkungan Big Data Analytics. Silhouette Score bertugas mengukur kemiripan suatu titik data dengan kelompoknya sendiri apabila dibandingkan dengan kelompok lain. Di sisi lain, Davies-Bouldin Index bertugas menghitung rasio antara jarak internal cluster dan jarak antar cluster semakin kecil angka yang dihasilkan, semakin baik kualitas pengelompokan yang dicapai.

Perhitungan Silhouette Score untuk setiap data:

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \qquad (3)$$

Keterangan:

a) $S(i)$ = skor silhouette untuk data ke-i
b) $a(i)$ = jarak rata-rata ke-i dengan anggota lain dalam cluster yang sama
c) $b(i)$ = jarak rata-rata ke-I dengan anggota dari cluster terdekat lainnya

Silhouette Score memiliki rentang nilai dari -1 hingga 1. Jika nilainya mendekati 1, hal ini menunjukkan bahwa kualitas clustering yang dihasilkan tergolong sangat baik.

Perhitungan Davies-Bouldin Index:

$$DBI = \frac{1}{\mathcal{K}} \sum_{i=1}^{k} \max_{i \neq j} \left(\frac{\sigma i + \sigma j}{d(c_i, c_j)} \right) \qquad (4)$$

Keterangan:

a) DBI = Davies-Bouldin Index
b) k = total cluster yang dibentuk dalam proses clustering
c) σi = nilai rata-rata jarak setiap data dalam cluster i*i* ke titik pusat (centroid) cluster tersebut
d) $d(ci, cj)$ = jarak antar centroid dari cluster *i* dan cluster *j*

Nilai DBI yang rendah—terutama yang mendekati 0—menandakan bahwa kualitas cluster yang terbentuk tergolong baik. Hal ini disebabkan oleh kecilnya jarak antar anggota dalam satu cluster dan besarnya jarak yang memisahkan antar cluster.

Pendekatan validasi ini juga digunakan dalam penelitian ElHaj dan Alshamsi tentang big data clustering pada data hidrologi.

3. HASIL DAN PEMBAHASAN

3.1 Hasil Eksplorasi Big Data

Hasil analisis Big Data pada data DTKS menunjukkan bahwa data awal berjumlah 10.000 data dengan 15 variabel. Setelah proses pembersihan data dari duplikat dan nilai hilang, diperoleh 9.830 data siap olah . Berdasarkan analisis statistik deskriptif big data, rata-rata pendapatan penerima bantuan adalah Rp2.850.000 dengan standar deviasi Rp1.950.000. Rata-rata kepemilikan aset berada pada skor 3,8 dari skala 0-10. Distribusi status verifikasi menunjukkan bahwa 60% data berstatus "Sah", 25% berstatus "Bermasalah", dan 15% berstatus "Tidak Berhak" .

3.2 Optimalisasi Jumlah Cluster dalam Analisis Big Data

Penetapan jumlah cluster yang ideal pada data big data dilakukan dengan memanfaatkan metode Elbow dan Silhouette Coefficient. Dari hasil analisis Sum of Squared Errors (SSE) untuk K bernilai 1 hingga 10, terlihat penurunan yang sangat tajam sampai dengan K=3, sebelum akhirnya penurunannya mulai melambat. Hal ini mengisyaratkan bahwa K=3 adalah jumlah cluster yang paling sesuai. Keyakinan ini diperkuat oleh capaian Silhouette Score tertinggi pada K=3 yakni 0,68 dan Davies-Bouldin Index terendah pada angka yang sama yaitu 0,62.

Tabel 3. Metrik Evaluasi Big Data Clustering untuk Berbagai Nilai K

Jumlah Cluster (K)	SSE	Silhouette Score	Davies-Bouldin Index
2	18.450,23	0,62	0,78
3	12.340,67	0,68	0,62

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

Jumlah Cluster (K)	SSE	Silhouette Score	Davies-Bouldin Index
4	9.870,45	0,65	0,71
5	8.120,34	0,61	0,79

Sumber: Hasil Peneliti, 2025

3.3 Hasil Clustering K-Means pada Big Data

Berdasarkan K=3, algoritma K-Means berhasil mengelompokkan 9.830 data penerima bantuan ke dalam tiga cluster :

Tabel 4. Distribusi Cluster Penerima Bansos pada Big Data

Cluster	Jumlah Data	Persentase	Interpretasi
Cluster 0	4.570	45,7%	Penerima Sah
Cluster 1	3.200	32,0%	Potensi Kekeliruan Data
Cluster 2	2.060	20,6%	Potensi Penyalahgunaan

Sumber: Hasil Peneliti, 2025

3.4 Profiling dan Interpretasi Cluster pada Big Data

Analisis centroid atau pusat cluster dilakukan untuk memahami karakteristik setiap kelompok penerima bantuan dalam skala big data .

Tabel 5. Profil Tiga Cluster Penerima Bansos pada Big Data

Variabel	Cluster 0 (Sah)	Cluster 1 (Kekeliruan)	Cluster 2 (Penyalahgunaan)
Jumlah Data	4.570 (45,7%)	3.200 (32,0%)	2.060 (20,6%)
Pendapatan Rata-rata	Rp1.200.000	Rp3.800.000	Rp6.500.000
Kepemilikan Aset	1,2 (Rendah)	2,8 (Sedang)	4,5 (Tinggi)
Luas Bangunan	30 m²	50 m²	90 m²
Jumlah Tanggungan	4 orang	3 orang	2 orang
Rata-rata Desil	3,2	5,8	8,5

Sumber: Hasil Peneliti, 2025

Cluster 0 merupakan kelompok Penerima Sah dengan karakteristik pendapatan rendah (rata-rata Rp1.200.000), kepemilikan aset rendah (skor 1,2), dan kondisi tempat tinggal yang kurang layak . Kelompok ini merupakan target utama program bantuan sosial dan mencakup 45,7% dari total data big data. Komposisi status verifikasi pada cluster ini didominasi oleh status "Sah" sebesar 85%.

Cluster 1 merupakan kelompok dengan Potensi Kekeliruan Data yang memiliki pendapatan sedang (rata-rata Rp3.800.000) dan kepemilikan aset sedang (skor 2,8) . Kelompok ini mencakup 32% dari total data dan memerlukan verifikasi ulang. Komposisi status verifikasi pada cluster ini terdiri dari 50% "Sah", 35% "Bermasalah", dan 15% "Tidak Berhak".

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

Cluster 2 merupakan kelompok dengan Potensi Penyalahgunaan tertinggi. Kelompok ini memiliki pendapatan tinggi (rata-rata Rp6.500.000), kepemilikan aset tinggi (skor 4,5), kondisi tempat tinggal yang sangat layak (luas rata-rata 90 m²), dan jumlah tanggungan yang rendah (rata-rata 2 orang) . Kelompok ini mencakup 20,6% dari total data. Komposisi status verifikasi pada cluster ini menunjukkan bahwa 80% data berstatus "Tidak Berhak".

3.5 Validasi dan Evaluasi Big Data Clustering

Hasil validasi internal menunjukkan Silhouette Score sebesar 0,68, yang menandakan struktur cluster yang terbentuk tergolong baik dan kompak. Adapun Davies-Bouldin Index mencapai 0,62, yang mengindikasikan bahwa jarak antar cluster terpisah dengan cukup baik. Sementara itu, validasi eksternal dilakukan melalui pembandingan antara label cluster yang dihasilkan dengan status verifikasi yang ada pada data. Hasil klasifikasi menunjukkan akurasi prediksi sebesar 73,5%.

Tabel 6. Confusion Matrix Hasil Big Data Clustering

Status Aktual	Prediksi Sah	Prediksi Bermasalah	Prediksi Tidak Berhak
Sah	3.890	450	230
Bermasalah	320	2.100	280
Tidak Berhak	150	310	2.100

Sumber: Hasil Peneliti, 2025

3.6 Analisis Pola Penyalahgunaan pada Big Data

Berdasarkan hasil Big Data Analytics, teridentifikasi lima pola utama penyalahgunaan bantuan sosial :

1. Pendapatan Tinggi Tetap Menerima Bansos: Sebanyak 2.060 penerima (20,6%) memiliki pendapatan rata-rata Rp6.500.000 namun masih terdaftar sebagai penerima bantuan sosial .
2. Kepemilikan Aset Tidak Sesuai: Kelompok potensi penyalahgunaan memiliki skor kepemilikan aset rata-rata 4,5 dari skala 10 .
3. Kondisi Tempat Tinggal Layak: Rata-rata luas bangunan pada kelompok potensi penyalahgunaan adalah 90 m² .
4. Jumlah Tanggungan Rendah: Rata-rata jumlah tanggungan pada kelompok potensi penyalahgunaan hanya 2 orang .
5. Desil Kesejahteraan Tidak Sesuai: Rata-rata desil pada kelompok potensi penyalahgunaan adalah 8,5 .

3.7 Pembahasan

Hasil penelitian ini menegaskan bahwa penggunaan Big Data Analytics melalui algoritma K-Means Clustering tergolong efektif dalam mengungkap pola-pola penyimpangan bantuan sosial yang berlangsung pada skala big data.Cluster 2 yang teridentifikasi sebagai kelompok potensi penyalahgunaan memiliki karakteristik yang sangat kontras dengan Cluster 0 sebagai kelompok penerima sah . Hal ini sesuai dengan penelitian Daulay dan Wandri yang menunjukkan bahwa kombinasi metode clustering dan klasifikasi dapat meningkatkan akurasi dalam seleksi penerima bantuan pada skala big data.

Keberhasilan Big Data Analytics dalam mengidentifikasi kelompok potensi penyalahgunaan juga didukung oleh hasil validasi eksternal yang menunjukkan akurasi prediksi sebesar 73,5% . Penelitian oleh Sreedhar dkk. juga menunjukkan bahwa penerapan Big Data Analytics pada platform seperti Hadoop dapat meningkatkan efisiensi clustering pada dataset yang sangat besar. Penelitian ini menyuguhkan implikasi praktis yang penting bagi para pemangku kepentingan yang mengelola program bantuan sosial.

4. KESIMPULAN

Berdasarkan implementasi Big Data Analytics pada penelitian ini, terbukti bahwa algoritma K-Means Clustering efektif dalam mendeteksi pola-pola penyimpangan penyaluran bantuan sosial.Dari 9.830 data yang dianalisis, terbentuk tiga cluster dengan karakteristik yang berbeda. Cluster 0 (Penerima Sah) berjumlah 4.570 data (45,7%) dengan pendapatan rendah dan aset minim. Cluster 1 (Potensi Kekeliruan Data) berjumlah 3.200 data (32%) dengan karakteristik ekonomi menengah. Cluster 2 (Potensi Penyalahgunaan) berjumlah 2.060 data (20,6%) dengan pendapatan tinggi dan kepemilikan aset tinggi. Validasi Big Data Analytics menggunakan Silhouette Score menghasilkan nilai 0,68 dan Davies-Bouldin

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**

Jurnal Kolaborasi
Sains dan Ilmu Terapan
Volume 5 Nomor 1 Juli Desember 2026 | Page -1-8
ISSN 2964- 5581 (media online)
https://utilityprojectsolution.org/ejournal/index.php/JuKSIT
Penerbit : Utiliti Project Solution

Index sebesar 0,62 yang mengindikasikan cluster terbentuk dengan baik. Akurasi prediksi sebesar 73,5% menunjukkan bahwa model Big Data Analytics cukup handal dalam mengelompokkan penerima bantuan.

Penelitian ini memberikan kontribusi praktis bagi pemerintah dalam meningkatkan akurasi dan efisiensi penyaluran bantuan sosial melalui Big Data Analytics. Dengan menerapkan teknologi Big Data Analytics, pihak terkait dapat mengidentifikasi kelompok masyarakat yang benar-benar membutuhkan serta mendeteksi potensi penyalahgunaan, sehingga bantuan sosial dapat tepat sasaran. Penelitian selanjutnya disarankan untuk menggunakan variabel tambahan seperti riwayat penerimaan bantuan sebelumnya dan data geospasial untuk meningkatkan akurasi Big Data Analytics, serta membandingkan hasil K-Means dengan metode clustering lain seperti DBSCAN atau Hierarchical Clustering pada skala big data yang lebih besar.

UCAPAN TERIMAKASIH

Penghargaan setinggi-tingginya saya sampaikan kepada STMIK Pelita Nusantara Medan atas segala bentuk dukungan, baik sarana maupun prasarana, yang telah memungkinkan penelitian ini berjalan lancar dan mencapai hasil yang diharapkan.

REFERENCES

[1] R. Haripriya, N. Khare, M. Pandey, and S. Biswas, “Decentralized big data mining : federated learning for clustering youth tobacco use in India,” J. Big Data, 2024, doi: 10.1186/s40537-024-01042-0.
[2] S. Daulay and R. Wandri, “Integrating K - Means Clustering and K - Nearest Neighbor Classification for Effective Scholarship Recipient Selection,” vol. 14, pp. 235–248, 2025.
[3] F. H. Basuki et al., “BIG DATA ANALYTICS IN PUBLIC SECTOR AUDIT : TRANSFORMING RISK ASSESSMENT AND FRAUD DETECTION,” vol. 2, no. 3, pp. 986–1002, 2025.
[4] A. S. Shirkhorshidi, S. Aghabozorgi, T. Y. Wah, and T. Herawan, “Big Data Clustering : A Review”.
[5] O. Kurasova, V. Marcinkevi˘, V. Medvedev, and A. Rape˘, “Strategies for Big Data Clustering,” 2014, doi: 10.1109/ICTAI.2014.115.
[6] K. Elhaj and D. Alshamsi, “GeoTemporal clustering for aquifer delineation : a big data approach to synchronizing and analyzing variable - length groundwater time series,” J. Big Data, 2025, doi: 10.1186/s40537-025-01060-6.
[7] C. Sreedhar, N. Kasiviswanath, and P. C. Reddy, “Clustering large datasets using K - means modified inter and intra clustering (KM - I2C) in Hadoop,” J. Big Data, 2017, doi: 10.1186/s40537-017-0087-2.
[8] D. Novita, “PENERAPAN K-MEANS UNTUK PRIORITAS PENERIMA BANTUAN,” pp. 105–114, 2025.
[9] M. I. Zuhendra and R. Hidayat, “Penerapan Data Mining Untuk Klasterisasi Tingkat Kemiskinan Berdasarkan Data Terpadu Kesejahteraan Sosial (DTKS),” vol. 7, pp. 32–42, 2024.
[10] S. Rainaya and S. Dewi, “Penerapan Data Mining Penyaluran Dana Bantuan Sosial Menggunakan Metode K-Means Clustering di Desa Rancaekek Kulon,” vol. 4, no. 3, pp. 433–444, 2025.
[11] J. Jin, “Student behavior patterns in vocational education big data based on clustering algorithm,” 2025.
[12] M. Astriani, M. Hamu, and A. C. Talakua, “Optimalisasi Manajemen Bantuan Sosial Dengan Implementasi Data Mining Menggunakan Algoritma K-Means Clustering,” vol. 1, no. 1, pp. 7–12, 2024.
[13] B. Zerhari, A. A. Lahcen, and S. Mouline, “Big Data Clustering : Algorithms and Challenges”.
[14] J. Informatika, D. Rekayasa, K. Jakakom, S. Amaliyah, and S. R. Agustini, “Penerapan Data Mining Untuk Menentukan Kelompok Prioritas Penerima Bantuan PKH Menggunakan Metode Clustering K-Means Pada Desa Kuala Dendang Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM),” vol. 3, no. April, pp. 453–458, 2023.
[15] D. I. Desa and T. Ciamis, “PENERAPAN ALGORITMA K-MEANS CLUSTERING UNTUK IDENTIFIKASI KELAYAKAN PENERIMA BANTUAN PROGRAM KELUARGA HARAPAN (PKH),” vol. 7, no. 6, pp. 3363–3369, 2023.

Submitted: **17/06/2026**; Accepted: **29/06/2026**; Published:**08/07/2026**