

DATA MINING UNTUK KLASIFIKASI STATUS GIZI DESA DI KABUPATEN MALAKA MENGGUNAKAN METODE *K-NEAREST NEIGHBOR*

Brigita Yulia Lestari Fahik¹, Bertha S. Djahi², Nelci D. Rumlaklak³
^{1,2,3} Jurusan Ilmu Komputer, Fakultas Sains dan Teknik, Universitas Nusa Cendana

INTISARI

Klasifikasi status desa menurut jumlah penderita gizi buruk sangatlah penting dalam mengantisipasi kasus gizi buruk di sebuah daerah, khususnya bagi daerah-daerah yang ada di Kabupaten Malaka. Kasus gizi buruk yang tercatat cukup banyak di Kabupaten Malaka menuntut pemerintah Kabupaten Malaka agar segera mengantisipasi masalah tersebut. Untuk mengatasinya, pada penelitian ini digunakan metode *k-Nearest Neighbor* untuk mengklasifikasi status desa-desa yang ada di Kabupaten Malaka berdasarkan nilai tingkat balita dengan berat badan dibawah garis merah kedalam tiga kelas target yaitu rendah, sedang, dan tinggi. Sebelum dilakukan proses klasifikasi, terlebih dahulu dilakukan proses klastering menggunakan metode *K-Means* agar seluruh data dapat terbagi dalam kelas yang telah ditentukan. Data yang digunakan dalam penelitian ini sebanyak 174 data yang diambil dari tahun 2013-2015. Hasil akhir, setelah dilakukan validasi data hasil *clustering* didapat kemiripan dengan data asli sebesar 98,25%, dan hasil pengujian sistem sebesar 93,10%. Penentuan nilai *k* terbaik dengan data uji sebanyak 34 buah dan data latih sebesar 140 buah yaitu pada $k=7$ dengan rata-rata presentase kemiripan sebesar 95.53%.

Kata Kunci: Klasifikasi, Klastering, gizi buruk, TBGM, *k-Nearest Neighbor*, *K-Means*, *purity measure*.

ABSTRACT

Classification of village status according to the number of malnourished patients is very important in anticipating malnutrition cases in a region, especially for the areas in the district of Malaka. Cases of malnutrition recorded quite a lot in the District of Malaka demanded the district government of Malaka to immediately anticipate the problem. To overcome this problem, we used k-Nearest Neighbor method to classify the status of villages in Malaka District based on the level of under-five children under the red line into three target classes: low, medium, and high. Prior to the classification process, clustering process is done using K-Means method so that all data can be divided into classes that have been determined. The data used in this study as many as 174 data taken from the year 2013-2015. The final result, after validation of clustering data obtained resemblance to the original data of 98.25%, and the results of system testing of 93.10%. Determination of the best value of k with the test data of 34 pieces and the training data of 140 pieces is at $k = 7$ with the average percentage of similarity of 95.53%.

Keywords: *Classification, clustering, malnutrition, TBGM, k-Nearest Neighbor, K-Means, purity measure*

I. PENDAHULUAN

Gizi buruk merupakan salah satu masalah utama dalam tatanan kehidupan dunia. Di negara Indonesia sendiri terdapat empat provinsi yang tidak pernah absen tiap tahun dalam kasus gizi buruk, salah satunya adalah provinsi Nusa Tenggara Timur (NTT). Kabupaten Malaka merupakan salah satu kabupaten di NTT dengan jumlah balita yang bermasalah dengan gizi tergolong cukup banyak. Pada tahun 2014, Badan pusat statistik menyatakan bahwa dari 15.619 balita yang ditimbang di kabupaten Malaka, 498 balita diantaranya dinyatakan memiliki berat badan dibawah garis merah [5].

Dalam menyelesaikan masalah ini digunakan sebuah metode klasifikasi untuk membangun sebuah sistem pembantu pengambil keputusan berdasarkan solusi pada kasus sebelumnya. Metode klasifikasi *k-Nearest Neighbor* terkenal dengan kelebihanannya yaitu tangguh terhadap *training* data yang memiliki banyak *noise* serta *efektif* apabila memiliki *training* data yang sangat besar. Selain *k-Nearest Neighbor*, metode tambahan yang digunakan dalam penelitian ini adalah metode *K-Means clustering*. Metode ini sangat cocok untuk mengelompokkan *dataset* yang belum terbagi kedalam kelas *label*.

II. MATERI DAN METODE

2.1 Data Penelitian

Data yang digunakan dalam penelitian ini adalah data jumlah kasus balita penderita gizi buruk pada 58 desa dari 5 Kecamatan di Kabupaten Malaka, tahun 2013 sampai 2015. Rekam data yang digunakan adalah jumlah balita dengan berat badan dibawah garis merah (TBGM) dari sebuah desa dari bulan januari sampai bulan desember selama satu tahun.

2.2 Data Mining

Data mining merupakan salah satu rangkaian dari proses pencarian pengetahuan pada *database* (*Knowledge Discovery in Database/KDD*). *KDD* adalah keseluruhan proses untuk mencari dan mengidentifikasi pola (*pattern*) dalam data, dimana pola yang ditemukan bersifat sah, baru, dapat bermanfaat dan dapat dimengerti.

2.3 Algoritma *k-Nearest Neighbor*

Algoritma *k-Nearest Neighbor* (*KNN*) adalah sebuah metode untuk melakukan penggolongan terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut [23]. *K-Nearest Neighbor* dilakukan dengan mencari kelompok objek dalam data *training* yang paling dekat (mirip) dengan objek pada data baru atau data *testing*.

Rumus yang digunakan untuk menghitung jarak antar objek yaitu *Euclidean Distance* seperti pada persamaan:

$$d(x_i, y_i) = \sqrt{(x_i - y_i)^2} \dots\dots\dots(1)$$

Keterangan:

x_i = record 1, objek ke-i
 y_i = record ke 2, objek ke i
 $d(x_i, y_i)$ = nilai jarak antara x_i dan y_i

Berikut tiga kelompok klasifikasi yang digunakan untuk mengklasifikasi status desa [20] seperti pada tabel 1.

Tabel 1. Klasifikasi nilai Tingkat Balita dengan Berat Badan dibawah Garis Merah

No	Rentang (%)	Label
1.	TBGM < 3	Rendah
2.	3 => TBGM <6	Sedang
3.	TBGM => 6	Tinggi

2.4 Algoritma *K-Means Clustering*

Metode *clustering* mengacu pada pengelompokkan data, observasi atau kasus berdasar kemiripan objek yang diteliti. Sebuah cluster adalah suatu kumpulan data yang mirip dengan lainnya atau ketidakmiripan data pada kelompok lain [1]. Metode *K-Means* adalah teknik penyekatan (*partition*) yang membagi atau memisahkan objek yang berbeda, dan mengumpulkan objek yang sama ke dalam jumlah *K* daerah yang telah ditentukan sebelumnya dan setiap objek

harus masuk dalam kelompok tersebut. Metode ini dimulai dengan pembentukan prototipe *cluster* di awal kemudian secara iteratif prototipe *cluster* ini diperbaiki hingga tidak terjadi perubahan yang signifikan pada prototipe *cluster*. Rumus yang digunakan untuk mengukur jarak yaitu - *Euclidean Distance* pada Persamaan (1). Berikut algoritma *K-Means clustering*:

1. Menentukan banyak *K cluster* yang ingin dibentuk.
2. Membangkitkan nilai random untuk pusat awal *cluster* sebanyak *K cluster*.
3. Menghitung jarak setiap data terhadap masing-masing *centroid* menggunakan rumus *euclidean distance* seperti pada Persamaan (1).
4. Mengelompokkan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
5. Memperbaharui nilai *centroid* berdasarkan dari rata-rata *cluster* dengan persamaan:

$$C_k = \frac{1}{nk} \sum d_i \dots \dots \dots (2)$$

Keterangan:

C_k = *cluster*

nk = jumlah data dalam *cluster*

d_i = jumlah tiap objek yang masuk dalam masing-masing *cluster*.

6. Ulangi langkah 2 sampai langkah ke 5 hingga anggota tiap *cluster* tidak ada yang berpindah.

Jika langkah 6 telah terpenuhi, maka nilai rata-rata pusat *cluster* pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

2.5 Tahap Perancangan Sistem Klasifikasi

Sistem Klasifikasi status desa dimulai dengan admin memasukan data jumlah balita BGM (Bawah Garis Merah) dan total balita yang ditimbang tiap bulannya dari sebuah desa. Selanjutnya sistem akan melakukan pengelompokkan awal *dataset* menggunakan perhitungan *K-Means*. Data hasil *clustering* inilah yang seterusnya akan digunakan oleh sistem sebagai *dataset* pada proses klasifikasi. Kerangka kerja sistem klasifikasi status desa dapat dilihat pada gambar 1.

III. HASIL DAN PEMBAHASAN

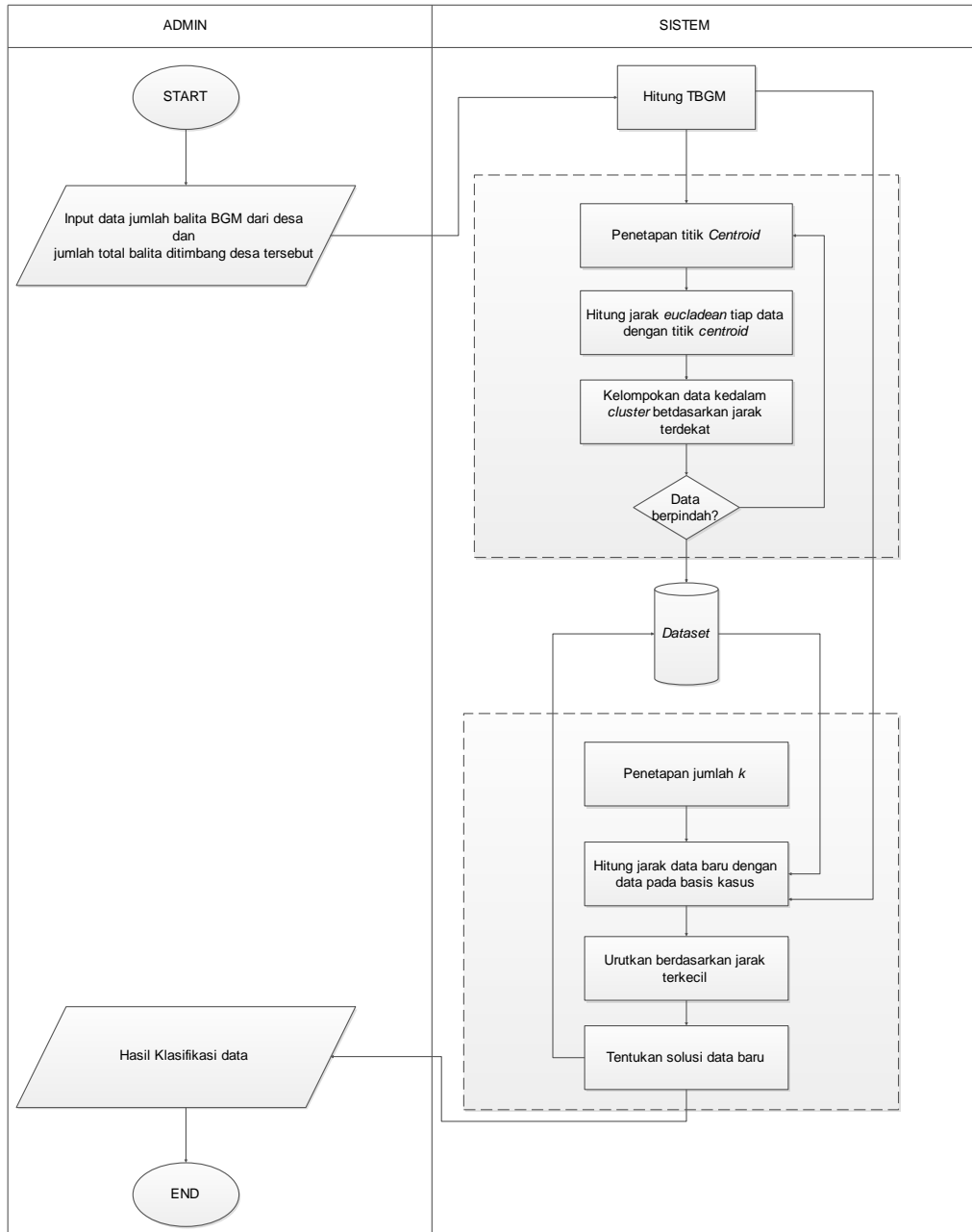
3.1 Hasil Pengujian Sistem Klasifikasi

Sistem klasifikasi status desa telah diberikan pengujian dengan 3 macam cara uji. Cara pertama yaitu, dilakukan perbandingan antara *dataset* hasil *clustering* dengan data asli. Pengujian dilakukan sebanyak 5 kali percobaan, dengan memilih secara acak kemungkinan-kemungkinan titik *centroid* awal yang berbeda. Dari hasil pengujian ini diperoleh presentase kemiripan tertinggi dengan data asli terdapat pada percobaan ke-2 nilai *purity* yang dihasilkan yaitu sebesar 98,83%. Diagram perbandingan titik *centroid* awal dapat dilihat pada gambar 2.

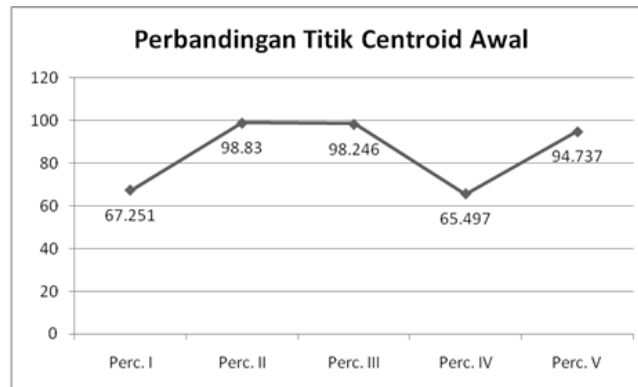
Pengujian ke-2 adalah penentuan nilai *k* terbaik untuk proses klasifikasi. Pada bagian ini, pengujian dilakukan dengan membagi *dataset* menjadi dua bagian, yaitu 140 *dataset* sebagai data latih dan 34 *dataset* sebagai data uji. Pengujian dilakukan sebanyak 5 kali dengan melakukan proses klasifikasi terhadap data uji. Setelah membagi *dataset* menjadi dua bagian, akan dilakukan proses perhitungan klasifikasi terhadap 34 data uji. Proses pengujian dilakukan sampai 5 kali pembagian data uji dengan jumlah nilai *k* dari 1 sampai 10. Hasil pengujian menunjukkan nilai rata-rata presentase kecocokan tertinggi ada pada *k* ke-7 dengan presentase kecocokan sebesar 95,53%. Maka *k*=7 akan digunakan sebagai *k* terbaik untuk melakukan pengujian terhadap sistem klasifikasi status desa berdasarkan jumlah balita BGM. Diagram perbandingan nilai *k* dapat dilihat pada gambar 3.

Pengujian ke-3, seluruh data akan diklasifikasi kembali dengan menggunakan nilai *k*=7. Presentase kinerja yang dihasilkan oleh sistem klasifikasi status desa yaitu sebesar 93,103% dengan hasil klasifikasi data yaitu, 73 data kedalam kelas rendah dengan 70 data diklasifikasi

sesuai dan 3 data tidak sesuai, 39 data kedalam kelas sedang dengan 37 data sesuai dan 2 data tidak sesuai, dan 62 data kedalam kelas tinggi dengan 55 data sesuai dan 7 data tidak sesuai.



Gambar 1. Kerangka kerja sistem klasifikasi status desa



Gambar 2. Diagram perbandingan titik centroid awal



Gambar 3. Diagram perbandingan nilai k

3.2 Pembahasan

Dari hasil pengujian pertama dapat diketahui bahwa Setelah melakukan lima kali percobaan, diperoleh presentase kemiripan tertinggi dengan data asli terdapat pada percobaan ke-2 yaitu sebesar 98,83%, dengan 72 data dikelompokkan kedalam kelas rendah, 47 data kedalam kelas sedang, dan 55 data kedalam kelas tinggi. Berdasarkan pengujian yang dilakukan, maka pemilihan titik *centroid* dan hasil *clustering* data dari percobaan ke-2 digunakan sebagai *dataset* untuk melakukan pengujian pada sistem klasifikasi.

Sedangkan pada pengujian ke-2 dapat diketahui bahwa hasil klasifikasi yang menunjukkan nilai rata-rata presentaste kecocokan tertinggi ada pada k ke-7 dengan presentase kecocokan sebesar 95,53%. Maka $k=7$ akan digunakan sebagai nilai k untuk melakukan pengujian terhadap sistem klasifikasi status desa berdasarkan jumlah balita BGM.

Berdasarkan hasil pengujian klasifikasi data dengan menggunakan $k=7$, diperoleh hasil klasifikasi data yaitu, 73 data kedalam kelas rendah dengan 70 data diklasifikasi sesuai dan 3 data tidak sesuai, 39 data kedalam kelas sedang dengan 37 data sesuai dan 2 data tidak sesuai, dan 62 data kedalam kelas tinggi dengan 55 data sesuai dan 7 data tidak sesuai.

IV. PENUTUP

4.1 Kesimpulan

Berdasarkan implementasi dan pengujian sistem klasifikasi status desa yang telah dilakukan sebelumnya, maka dapat disimpulkan bahwa:

1. Metode *K-Means clustering* mampu mengelompokkan desa berdasarkan nilai TBGM dengan nilai *purity* hasil *clustering* sebesar 98,83%.

2. Metode *k-NN* mampu mengklasifikasi status desa dengan baik dengan presentase sistem sebesar 93.10%.
3. Nilai *k* terbaik didapat pada $k=7$ dengan rata-rata nilai akurasi yang dihasilkan setelah lima kali pembagian data uji adalah sebesar 95,53%.

4.2 Saran

Perlu adanya penelitian lebih lanjut mengenai klasifikasi status desa berdasarkan jumlah balita BGM di kabupaten Malaka, serta penelitian lebih lanjut mengenai pelatihan data menggunakan metode *K-Means*, seperti menambah parameter ataupun jumlah *dataset* agar sistem dapat belajar dengan baik dalam menemukan pola data.

DAFTAR PUSTAKA

- [1] Agusta, Y., 2007, *K-Means-Penerapan, Permasalahan dan Metode Terkait*, Jurnal Sistem dan Informatika Vol.3, pp : 47-60, Denpasar, Bali.
- [2] Anonim, 2013, *Pembentukan Kabupaten Malaka di Provinsi Nusa Tenggara Timur*.
- [3] Anonim, 2015, *Buku Putih Sanitasi Kabupaten Malaka*.
- [4] Azis, A., 2015, *Penderita Gizi Buruk di NTT Hampir 2000 balita*, <https://m.tempo.co/read/news/2015/06/27/173678797/hampir-2-000-balita-di-ntt-menderita-gizi-buruk>, diakses 21 Mei 2016.
- [5] BPS, 2014, *Status Gizi Balita Menurut Kabupaten/Kota di Provinsi Nusa Tenggara Timur*, <http://ntt.bps.go.id/index.php/linkTabelStatis/356>, diakses 23 Mei 2016.
- [6] Fadilah, I., 2013, *Definisi Gizi Buruk Pada Anak, Gejala, Penyebab, dan Cara Penanganannya*, <http://www.idmedis.com/2014/11/definisi-gizi-buruk-pada-anak-gejala.html>, diakses 4 Juni 2016.
- [7] Fathansyah, 1999, *Basis Data, Informatika Bandung*, Bandung.
- [8] Fuada, N., Muljati, S., Hidayat, T., 2012, *Penentuan Daerah Rawan Gizi Berdasarkan Analisis Spasial, Media Litbang Kesehatan*.
- [9] Gizinet, 2010, *Kasus Gizi Buruk: Empat Provinsi Tak Pernah Absen*, <http://gizi.depkes.go.id/kasus-gizi-buruk-empat-provinsi-tak-pernah-absen>, diakses 21 Mei 2016.
- [10] Hermaduanty, N., Kusumadewi, S., 2008, *Sistem Pendukung Keputusan Berbasis SMS Untuk Menentukan Status Gizi dengan metode *k-Nearest Neighbor**, Teknik Informatika Fakultas Teknologi Industri, Universitas Islam Indonesia, Yogyakarta.
- [11] Huang, Z., (1998), *Extension to the K-means Algorithm, for Clustering Large Datasets with Categorical Values*, *Data Mining and Knowledge Discovery*, vol.2, hal.283-304
- [12] Imandoust, S., B., Bolandraftar, M., 2013, *Application of K-Nearest Neighbor (KNN) Approach for Predicting Economic Events: Theoretical Background*, Vol. 3.605-610, Payame Noor University, Iran.
- [13] Irdiansyah, E., 2012, *Penerapan Data Mining Pada Penjualan Produk Minuman di PT. Pepsi Cola Indobeverages Menggunakan Metode Clustering*, Universitas Komputer Indonesia.

- [14] Krisandi, N., Prihandono, B., 2013, Algoritma *k-Nearest Neighbor* dalam klasifikasi data hasil produksi kelapa sawit pada PT.Minamas kecamatan Parindu, Pontianak.
- [15] Larose D, T., 2006, *Data Mining Methods and Models*, Jhon Wiley & Sons, Inc. Hoboken New Jersey.
- [16] Oetomo, B., S., D., 2002, *Perencanaan & Pembangunan Sistem Informasi*, Andi, Yogyakarta.
- [17] Permana, B., Ukar, K., 2011, *Praktis Menguasai Microsoft Access 2007/2010*, PT Elex Media Komputindo, Jakarta.
- [18] Rohmawati, W., N., Defianti, S., Jajuli, M., 2015, Implementasi Algoritma *K-Means* dalam pengklasteran mahasiswa pelamar beasiswa, Teknik Informatika Fakultas Ilmu Komputer, Universitas Singaperbangsa, Karawang.
- [19] Soekirman, 2000, Ilmu Gizi dan Aplikasinya, Departemen Pendidikan Nasional, Jakarta.
- [20] Swastina, L., Lareno, B., 2014, Model Penentuan Potensi Status Gizi Bermasalah Menggunakan *Decision Tree*, Teknik Informatika, STMIK Indonesia Banjarmasin, Banjarmasin.
- [21] Supriasa, 2001, *Penilaian Status Gizi*, EGC, Jakarta.
- [22] Tiaratuti A., S., Sudaryanto S., 2015, Implementasi Metode *K-Nearest Neighbor* Dalam Peramalan Penjualan Mobil Pada PT Bengawan Abadi, Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Semarang.
- [23] Widiarsana, O., Putra, N.W., Budiayasa, P.G.I., Bismantara, A.N.I., Mahajaya, S., N., 2011, Data Mining: Metode Clasification K-Nearest Neighbor (KNN), Teknologi Informasi, Universitas Udayana, Bali.
- [24] Yuswanto, S., 2010, Boom..! Visual Studio .Net 2010 Meledak, Cerdas Pustaka Publisher, Jakarta.