KLUSTERISASI GENDER MENGGUNAKAN ALGORITMA K-MEANS

GENDER CLUSTERING WITH K-MEANS ALGORITHM

Wenefrida Tulit Ina[1)], Molina O.Odja[2)], Stephanie Pella[3)], Frans Likadja[4)]
[1)]Prodi Teknik Elektro Universitas Nusa Cendana
Jl. Adi Sucipto Penfui Kupang, NTT
[2,3,4)]Prodi Teknik Elektro Universitas Nusa Cendana
Jl. Adi Sucipto Penfui Kupang NTT
[1)]e-mail: wenefrida_ina@staf.undana.ac.id

ABSTRAK

Penelitian di bidang Data Mining sangat penting bagi pengembangan pengetahuan bidang mechine learning, dimana salah satu bagiannya adalah pengelompokan/klusterisasi terhadap suatu data set untuk memperoleh informasi penting dari data set tersebut.

Penelitian ini bertujuan menghasilkan pola pengelompokan/kluster dari data set gender berdasarkan data set nama-nama manusia yang diambil dari data set pada UCI_ML (UCI Machine Learning) tahun 2020 yang berjumlah 147.268 data. Data set ini akan dikelompokkan menjadi 2 kluster gender yaitu Male dan Female menggunakan algoritma K-Means.

Hasil penelitian ini menunjukan bahwa dari keseluruhan data yang digunakan, terdapat 61% (89840) Female dan 39% (57428) Male.

Kata Kunci : Klasterisasi, Gender, Algoritma K Means

ABSTRACT

Research in the field of Data Mining is very important for the knowledge development in the field of machine learning, where one of the parts is clustering a data set to obtain important information from a data set.

This study aims to produce a pattern of clusters from the gender data set based on the data set of human names taken from the data set in UCI_ML (UCI Machine Learning) in 2020 which amounted to 147,268 data. This data set is grouped into 2 gender clusters, namely Male and Female using the K-Means algorithm.

The results of this study indicate that from the overall data used, there are 61% (89840) Female and 39% (57428) Male.

Keywords: Clustering, Gender, K Means Algorithm

PENDAHULUAN

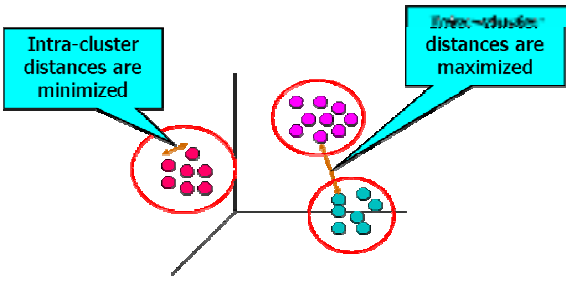
Data mining merupakan suatu bidang ilmu yang berhubungan dengan pengolahan data, analisa data, dan machine learning dengan tujuan menemukan suatu informasi baru dari sekumpulan data yang sangat banyak, dimana data-data tersebut lebih banyak merupakan data buangan atau pun data yang dianggap tak berguna. Data-data tersebut dapat dikelompokan/dikluster berdasarkan ciri-ciri tertentu sehingga dapat membentuk suatu pola yang mana pola tersebut akan menjadi pola prediksi untuk data-data yang baru.

Pada penelitian ini, data yang akan dikluster/dikelompokan adalah data nama-nama manusia yang tersimpan dalan data set gender pada UCI_ML tahun 2020. Dari data nama-nama ini akan diklusterisasi berdasarkan gender yaitu Female (Wanita) dan Male (Pria). Data set gender dari UCI-ML tahun 2020 berjumlah 147.268 data nama manusia, akan diolah menggunakan algoritma k-means untuk mendapatkan pengelompakan seperti yang sudah dijabarkan.

Tujuan penelitiian ini yakni mendapatkan kluster gender Female dan Male dari data set nama manusia yang digunakan. Untuk mencapai tujuan ini, penulis menggunakan algooritma K-Means pada aplikasi data mining WEKA dalam mengolah data ini.

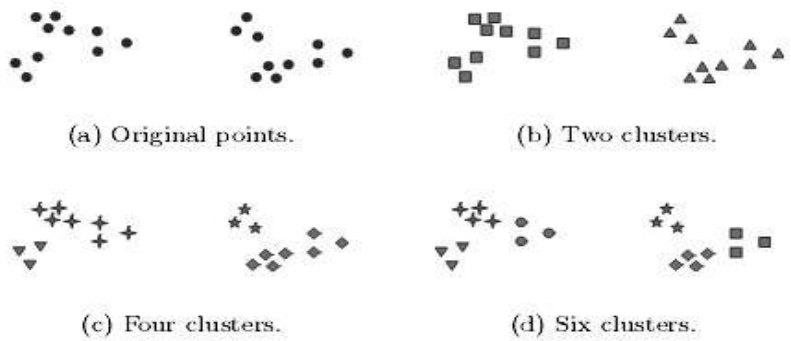
KLUSTERISASI

Klusterisasi (*clustering*) merupakan pengelompokan obyek-obyek data hanya berdasarkan pada informasi yang terdapat pada data tersebut, dan menjelaskan obyek dan relasinya (Tan, 2006).Tujuan analisis *cluster* agar obyek-obyek di dalam kelompok adalah mirip (memiliki karakteristik yang sama) satu dengan lainnya, dan berbeda (atau tidak berhubungan) dengan obyek dalam kelompok lainnya. Semakin besar tingkat kemiripan/homogenitas di dalam satu kelompok dan semakin besar tingkat perbedaan diantara kelompok, maka semakin baik (atau lebih berbeda) kluster tersebut. Gambar 1. merupakan ilustrasi prinsip *clustering.*



Gambar 1. Prinsip *Clustering*

Gambar 2. menunjukkan dua puluh titik dengan tiga cara membagi titik-titik tersebut dalam *cluster*, yang merupakan ilustrasi bagaimana definisi *cluster* tidak presisi dan definisi terbaik tergantung dari kondisi data serta hasil yang diinginkan.



Gambar 2. Beberapa cara menentukan Kluster dalam kelompok data.

Algoritma K-Means

K-Means merupakan metode analisis kelompok (klusterisasi) yang mengarah pada partisi dari N obyek pengamatan ke dalam K kelompok (kluster), dimana setiap obyek pengamatan dimiliki oleh sebuah kelompok dengan rata-rata (Mean) terdekat. Algoritma k-means mempartisi data ke dalam kelompok sehingga data yang memiliki karakteristik sama dimasukan ke dalam satu kelompok yang sama, dan data yang memiliki karakteristik berbeda dikelompokkan ke dalam kelompok yang lain.

Proses algoritma k-means sebagai berikut: (Prasetyo, 2012)

1. Tentukan jumlah kluster/kelompok yang diinginkan sesuai jenis data
2. Alokasikan data ke dalam kelompok secara acak
3. Hitung pusat kelompok (sentroid/rata-rata) dari data yang ada di masing-masing kelompok
4. Alokasikan masing-masing data ke sentroid/rata-rata terdekat.
5. Apabila ada data yang berpindah kelompok, atau apabila ada perubahan nilai sentroid di atas nilai ambang yang telah ditentukan, atau apabila perubahan nilai pada fungsi obyektif yang digunakan masih di atas nilai ambang yang ditentukan maka kembali ke langkah no.3.

Lokasi sentroid setiap kelompok yang diambil dari rata-rata (Mean) semua nilai data pada setiap fiturnya harus dihitung kembali.

Untuk menghitung sentroid fitur ke-i digunakan persamaan :

$$C_i = \frac{1}{M}\sum_{j=1}^{M} X_j$$

Dimana : C : Sentroid

M : Jumlah data dalam sebuah kelompok

Untuk menghitung jarak data ke pusat data (sentroid) digunakan persamaan Euclidean Distance, sebagai berikut :

$$D(X_2, X_1) = \|X_2 - X_1\|_2 = \sqrt{\sum_{j=1}^{p} |X_{2j} - X_{1j}|^2}$$

METODOLOGI PENELITIAN

Prosedur Penelitian

Adapun tahapan penelitian yang dilaksanakan sbb:

1. Pengumpulan Data

Pengumpulan data merupakan tahap awal penelitian. Data data yang dibutuhkan adalah dataset gender yang ada pada UCI_ML tahun 2020 berjumlah 147.268

2. Praprocessing Data

   Pada tahap ini dilakukan praprocessing data dimana data-data dari dataset UCI_ML di transformasikan ke dalam format data sesuai yang dibutuhkan dalam proses klusterisasi.

3. Pengolahan Data menggunakan Aplikasi WEKA dengan algoritma K-Means.

   Proses utama terjadi pada tahapan ini, yakni data-data gender yang sudah ditransformasi, diolah dengan algoritma k-means pada WEKA untuk mendapatkan klusterisasi gender yang diinginkan.

4. Pengujian Pola dengan Data Uji

   Pada tahapan ini, pola gender yang dihasilkan diuji menggunakan data uji.

5. Evaluasi Hasil Pengujian

   Proses terakhir adalah mengevaluasi hasil pengujian, apakah pola yg dihasilkan pada tahap 3 sesuai dengan hasil pengujian menggunakan data uji.

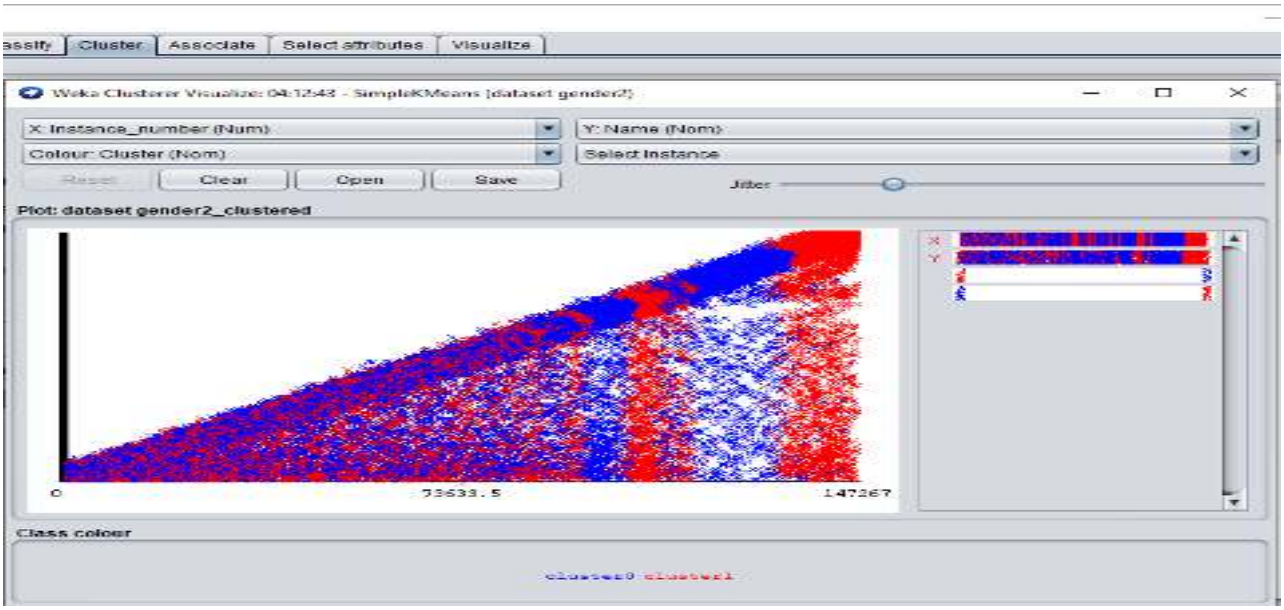
HASIL DAN PEMBAHASAN

Data nama manusia yang digunakan dalam penelitian berjumlah 147.268 yang diambil dari Data set UCI-ML tahun 2020.

Pada praprosesing data, diketahui bahwa ada sejumlah data *missing* yaitu sebanyak 254 data, seperti yang ditunjukkan pada gambar 3.

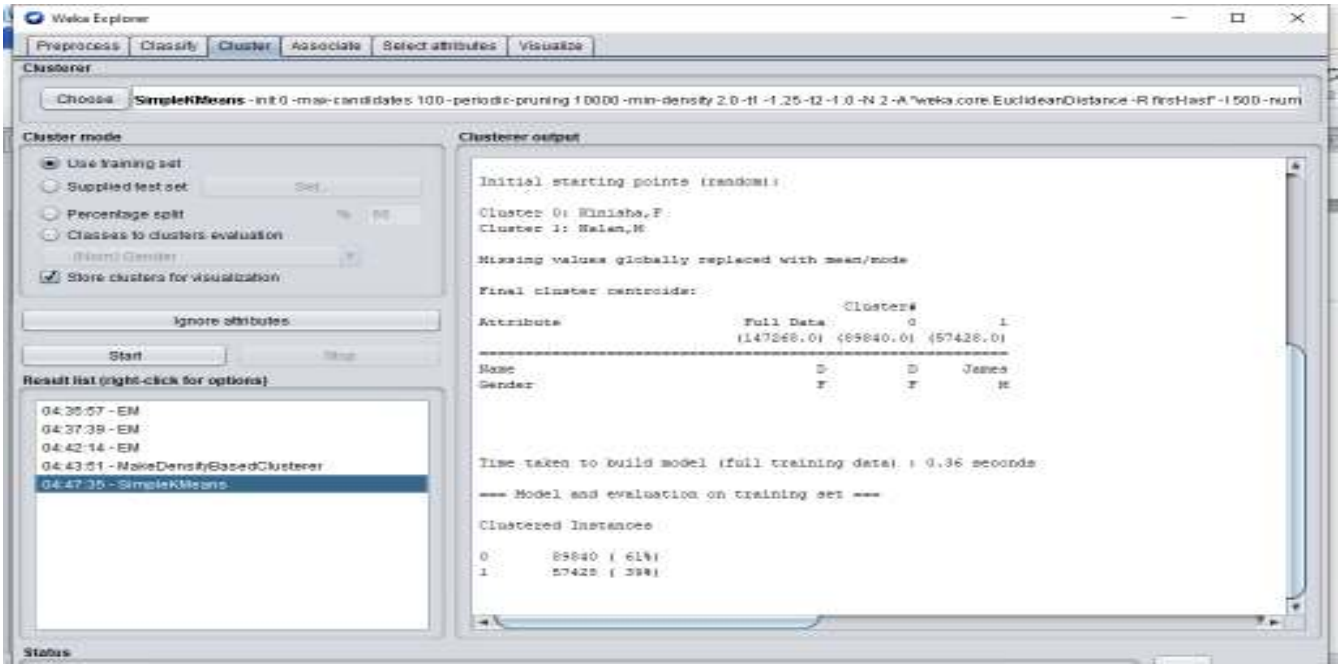
Gambar 3. Praproses data dengan WEKA

Kondisi keseluruhan data sebelu dilakukan klusterisasi dapat diluhat pada gambar 4.



Gambar 4. Kondisi persebaran data sebelum diproses.

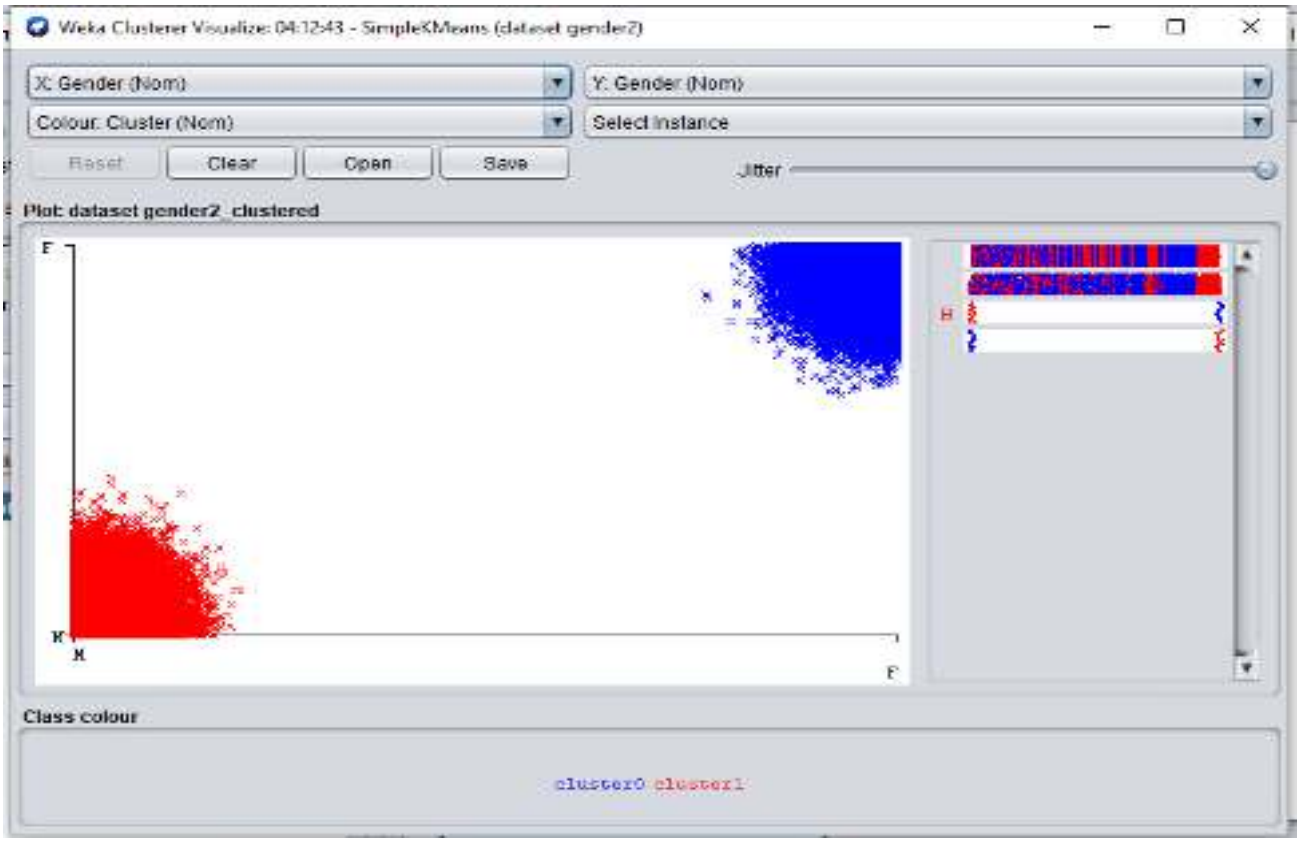
Setelah praproses ini, maka kembali dilakukan pengecekan data, kemudian dilanjutkan pada proses pengolahan data yaitu menggunakan Algoritma K-Means yang ada pada WEKA, seperti yang ditunjukan pada gambar 5.



Gambar 4. Hasil Klusterisasi menggunakan Algoritma K-Means pada WEKA.

Dari hasil ini, diketahui bahwa kluster 0 menunjukkan kelompok Female, titik sentroidnya pada “Kinisha” dengan jumlah anggota kelompok Female sebanyak 89.840 (61%) sedangkan kluster 1 menunjukkan kelompok Male, titik sentroidnya pada “Nalan”, dengan jumlah anggota kelompok Male sebanyak 57.428 (39%).

Klusterisasi data secara keseluruhan diperlihatkan pada gambar 6.



Gambar 6. Bentuk kluster yang dihasilkan Algoritma K-Means pada WEKA.

Warna merah menunjukkan kluster 1 (Male), sedangkan warna biru menunjukkan kelompok kluster 0 (Female).

KESIMPULAN

Berdasarkan hasil pengolahan data yang dilakukan oleh penulis, maka diambil kesimpulan bahwa algoritma K-Means dapat menghasilkan klusterisasi gender dengan baik sesuai dengan karakteristik data gender yang digunakan oleh penulis..

DAFTAR PUSTAKA

Anita F. Febrianti, Antonito H. Cabral, Gangga Anuraga. 2018. *K-Means Clustering dengan Metode Elbow Untuk Pengelompokkan Kabupaten dan Kota di Jawa Timur Berdasarkan Indikator Kemiskinan*. SNHRP-1/2018.

Asroni, Ronald Adrian. 2015. *Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan WEKA (Studi Kasus pada Jurusan Teknik Informatika UMM malang)*.. Jurnal Ilmiah SEMESTA TEKNIKA. Vol. 18. No.1. 76-82.

Baginda Harahap. 2018. *Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus pada UD.Toko Bangunan YD Indarung)*. Ready Star-2. ISSN (Cetak): 2620-6048. ISSN (online) :2686-6641

Budi Santosa, 2007. *Data Mining Terapan dengan MATLAB*. Penerbit Graha Ilmu

Eko Prasetyo, 2012. *Data Mining Konsep dan Aplikasi MATLAB*. Penerbit Andi

Muhammad Bakri. 2017. *Penerapan Data Mining Untuk Clustering Kualitas Batu Bara Dalam Proses Pembakaran di PLTU Sebalang Menggunakan Metode K-Means*. Jurnal TEKNOINFO Vol.11 No.1.2017. 1-4

Tan, P. et al. 2006. *Introduction to data Mining*. Boston: Pearson Education