KOMPARASI ALGORITMA CLUSTERING DENGAN DATASET PENYEBARAN COVID-19 DI INDONESIA PERIODE MARET-MEI 2020

Trientje Marlein Tamtelahitu[1]
[1]Program Studi Informatika, Universitas Kristen Indonesia Maluku
Email: [1]trienmarlein77@gmail.com

Masuk: 15 Juli 2020, Revisi masuk: 27 Juli 2020, Diterima: 29 Juli 2020

ABSTRACT

In data mining, there is a predictive model, namely predicting the value of different sample data sets, and testing into three types such as classification, regression and time series. While descriptive models allow us to determine patterns in sample data and divide them into groups, summaries and association rules. Report on the results of experiments on algorithms that are quite widely used in the field of machine learning. This experiment aims to measure performance on commonly used datasets in machine learning studies. The main performance factor to be compared in this experiment is the level of accuracy of the independent experiments on the dataset used.

This research uses clustering algorithm method to compare various clustering algorithms using Weka Tools to find out which algorithm will be more convenient for users to do clustering algorithm using the Covid-10 distribution map dataset in Indonesia from March-May 2020. K-means taking the points closest to the center whereas Farthest-First picks the furthest points. Farthest-First can complete the clustering process but with a lower quality than K-Means. And other experiments, on the method of Making Based on Clusterd Density and EM (Expectation-Maximization) prove the same accuracy. The EM grouping method proves low (less than 50%) of the results comparing the Clusterd Based Densitity Making Method, with a percentage reaching 74%.

Keywords*: Clustering algorithm, Data mining, Weka tools.*

INTISARI

Dalam data mining, ada model prediktif, yakni memprediksi nilai dari set data sampel yang berbeda, dan diklasifikasikan menjadi tiga jenis seperti klasifikasi, regresi dan deret waktu. Sedangkan model deskriptif memungkinkan kita untuk menentukan pola dalam data sampel dan dibagi lagi menjadi pengelompokan, peringkasan dan aturan asosiasi. Laporan hasil eksperimen pada algoritma yang cukup banyak digunakan dalam bidang *machine learning*. Eksperimen ini bertujuan untuk mengukur performansi algoritma pada dataset yang secara umum digunakan pada penelitian-penelitian *machine learning*. Faktor performansi utama yang ingin diperbandingkan pada eksperimen ini adalah tingkat akurasi dari algoritma independen terhadap dataset yang digunakan.

Penelitian ini menggunakan metode algoritma *clustering* untuk membandingkan berbagai algoritma *clustering* dengan menggunakan *Weka Tools* untuk mengetahui algoritma mana yang akan lebih nyaman bagi pengguna untuk melakukan algoritma *clustering* dengan menggunakan dataset peta penyebaran Covid-10 di Indonesia periode Maret-Mei 2020. Hasil eksperimen menggambarkan bahwa *K-Means* mengambil titik-titik yang paling dekat terhadap pusat sedangkan *Farthest-First* mengambil titik-titik terjauh. *Farthest-First* dapat menyelesaikan proses klasteringnya namun dengan kualitas yang lebih rendah dari K-Means. Dan eksperimen lainnya, pada metode *Make Densitity Based Clusterd* dan EM (Expectation-Maximization) menunjukkan akurasi yang sama baik. Metode pengelompokan EM menunjukkan akurasi rendah (kurang dari 50%) dari hasil dibandingkan Metode *Make Densitity Based Clusterd*, dengan presentase 74%.

Kata-kata kunci: *Algoritma clustering, Data mining, Weka tools.*

PENDAHULUAN

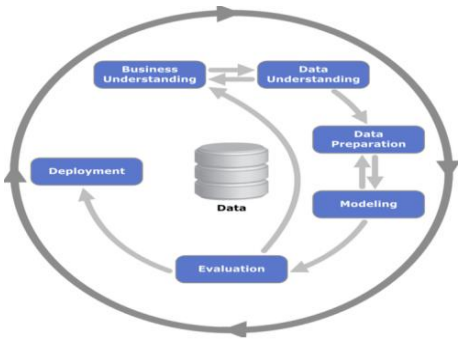
Clustering adalah salah satu model deskriptif untuk mengelompokkan satu set objek ke dalam kelompok-kelompok tertentu sesuai dengan hubungannya. *Clustering* adalah teknik yang digunakan dalam banyak

bidang seperti analisis gambar, pengenalan pola, analisis data statistik, dan sebagainya. *Clustering* adalah metode pembagian data menjadi kelompok-kelompok benda yang sama. Setiap *cluster* terdiri dari berbagai objek yang sama dan berbeda dibandingkan dengan objek kelompok lain (Chauhan, dkk., 2010). *WEKA Tools* digunakan untuk membandingkan berbagai algoritma pengelompokan. Ini digunakan karena menyediakan *interface* yang lebih baik bagi pengguna daripada dibandingkan dengan alat penambangan data lainnya. Dalam tulisan ini, ada perbandingan algoritma pengelompokan berbasis partisi dan non-partisi. Alasan pemilihan *WEKA Tools*, karena tidak rumit dan mudah tanpa harus memiliki pengetahuan mendalam tentang teknik *data mining*.

Mengacu pada hasil-hasil penelitian sebelumnya tentang penerapan metode *clustering*, yaitu penerapan metode *k-means* untuk *clustering* mahasiswa berdasarkan nilai akademik dengan Weka Interface (Asroni, 2015) analisis algoritma *K-Medoids clustering* dalam pengelompokan penyebaran Covid-19 di Indonesia (Sindi, dkk., 2020), dan *coal trade data clustering using K-Means* dengan studi kasus pada PT. Global Bangkit Utama (Rahman dan Wiranto, 2017), dalam makalah ini disajikan berbagai teknik pengelompokan dan perbandingannya menggunakan WEKA.

Data Mining

Data Mining merupakan bagian dari proses *Knowledge Discovery in Database* (KDD) (Han san Kamber, 2006). Proses KDD diilustrasikan pada Gambar 1.



Gambar 1. Tahapan Proses KDD

Data mining juga dapat diartikan secara luas berdasarkan kemampuannya yaitu proses menemukan *interesting knowledge* dari sejumlah data yang besar di *database*, data *warehouse*, atau tempat penyimpanan lainnya. *Data mining* dapat digunakan pada beberapa kasus yang meliputi ekonomi, bisnis, intelektual yang dapat dikategorikan menjadi 6 bagian *task* yaitu *classification, estimation, prediction, affinitygrouping, clustering, description* dan *profiling* (Berry dan Linoff, 2004).

Clustering

Clustering merupakan suatu proses pengelompokkan suatu *record*, observasi, atau pengelompokan kelas yang memiliki kesamaan objek. Perbedaan *clustering* dengan klasifikasi yaitu tidak adanya variabel target dalam melakukan suatu pengelompokan pada proses *clustering*. *Clustering* sering dilakukan sebagai langkah awal dalam proses *data mining* saat melakukkan suatu metode analisis (Sindi, dkk., 2020).

Farthest First

Algoritma farthest first menggunakan pemilihan secara acak untuk menentukan centroid dalam setiap pembentukan cluster. Untuk setiap perhitungan dilakukan dengan membandingkan setiap jarak antar kejadian dan mencari jarak yang terdekat dengan centroid. Pemilihan untuk cluster centroid selanjutnya menggunakan jarak yang terjauh dari cluster centroid yang aktif. Proses ini akan terus diulang sampai jumlah cluster yang terbentuk lebih dari batas yang telah ditetapkan (Sharma, dkk., 2012).

K-Means

K-Means merupakan algoritma clustering yang berulang-ulang. Algoritma K-Means dimulai dengan pemilihan secara acak K untuk cluster centroid (nilai K umumnya ditetapkan dahulu). Setiap kejadian membentuk sebuah cluster kemudian dicari sebagai center kemudian jika jumlah anggota cluster sama dengan nilai K, maka *cluster* tersebut ditutup.Selanjutnya setiap kejadian yang telah terbentuk centroid akan diproses ulang. Proses ini akan diulang sampai cluster centroid menjadi stabil (Jain dan Dubes, 1988).

EM (Expectation-Maximization)

Pada algoritma EM setiap cluster sama dengan distibution probability (kemungkinan penyebaran) dan untuk setiap kejadian data digunakan parameter nilai estimate pada

setiap distribution. Algoritma pencarian yang digunakan adalah maximum *likehood*. Algoritma ini menguraikan parameter dari distribution dengan cara melakukan secara berulang-ulang untuk memperkirakan nilai *expected* dari parameter dengan hipotesis yang digunakan. Hipotesis tersebut dihitung ulang dengan expected values. EM terdiri dari dua tahap, yaitu *estimation* dan maximization. Pada tahap *estimation* dilakukan perhitungan expected values dari parameter menggunakan hipotesis. Pada Tahap maximization dihitung nilai hipotesis maximum likehood dengan mengasumsikan parameter sama dengan expected value dari tahap estimation. Kedua tahap tersebut dilakukan berulang-ulang sampai hipotesa dari converge (terpusat) mencapai nilai yang stationer (Witten, dkk., 2002).

Make Density Based *Cluster*

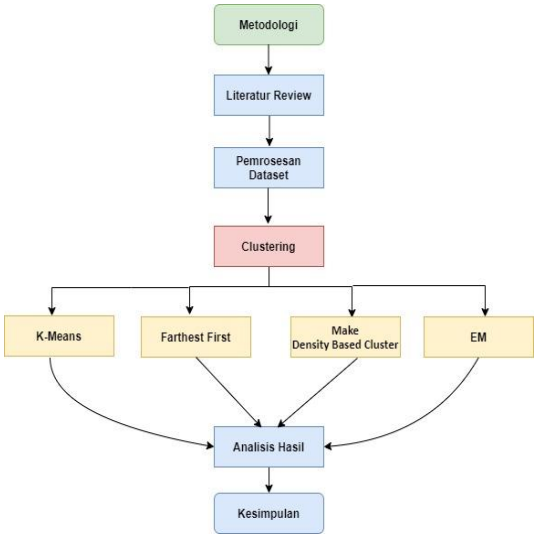
Algoritma ini didukung dalam analisis menggunakan WEKA. Dalam algoritma ini proses menemukan kembali *cluster* dilakukan dengan bentuk yang acak (*arbitary*). Pertumbuhan setiap *region* dengan kepadatan yang cukup dari setiap *cluster* mengikuti jaringan (rantai) dari setiap objek yang terhubung dengan *region*. Dalam model ini menghasilkan setiap *estimate* anggota di setiap *cluster* (Pelleg dna Moore, 2015).

WEKA

WEKA (*Waikato Environment for Knowledge Analysis*) merupakan aplikasi *data mining* yang bersifat *open source* berbasis Java. WEKA pertama kali dikembangkan oleh Universitas Waikato Selandia Baru sebelum menjadi bagian di Pentanho. Weka terdiri dari koleksi algoritma *machine learning* yang dapat digunakan untuk melakukan generalisasi atau formulasi dari sekumpulan data (https://waikato.github .io/weka-wiki/documentation/, 21 Juni 2020) Karena Weka ditulis dalam bahasa pemrograman Java, maka Weka juga didukung oleh GUI yang sangat baik dan *user friendly*, dapat mengolah berbagai file data seperti **.csv* dan *.*arff* serta memiliki fitur utama seperti *data preprocessing tools, learning algorithms* dan berbagai metode evaluasi (Eibe, 2011).

METODE

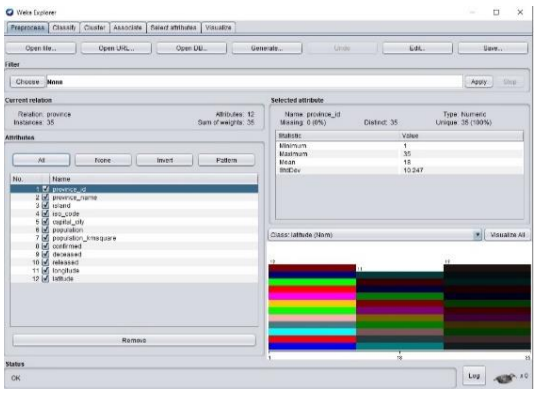
Penelitian ini dilakukan dalam empat tahapan (Gambar 2). Pada tahap pertama diuraikan beberapa algoritma *clustering* yang akan dikomparasi. Pada tahap kedua dijelaskan dasar-dasar perbandingan algoritma *clustering*. Pada tahap ketiga dibahas tentang perbandingan efektifitas antar metode *clustering*. Terakhir, tahap keempat adalah penyimpulan hasil penelitian.



Gambar 2. Alur Penelitian

Dataset

Untuk melakukan *clustering* dengan *raw data* diperlukan *dataset* untuk diuji coba. Dalam penelitian ini dataset diambil dari repositori *kaggle* dengan *dataset penyebaran* covid-19 di Indonesia, yang terdiri dari 9 *attribute* dan 35 *instances* (Gambar 3). Semua *dataset* tidak memiliki label dan akan dilakukan *clustering* menggunakan algoritma yang tersedia di WEKA. Koleksi teks terdiri dari atribut nama dokumen (https://www.kaggle.com/ardisra gen/indonesia-coronavirus-cases, 16 Juni 2020).



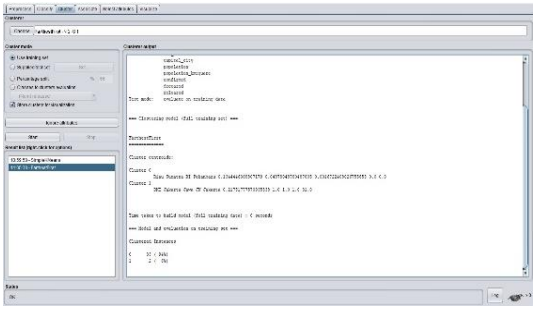
Gambar 3. Persiapan Dataset

ISSN: 1979-8415
E-ISSN: 2714-8025

PEMBAHASAN

Simulasi Algoritma Farthest First

Simulasi dilakukan pada algoritma *Farthest First*, yaitu menentukan *centroid* dalam setiap pembentukan *cluster* dilakukan pemilihan secara acak (Gambar 4).



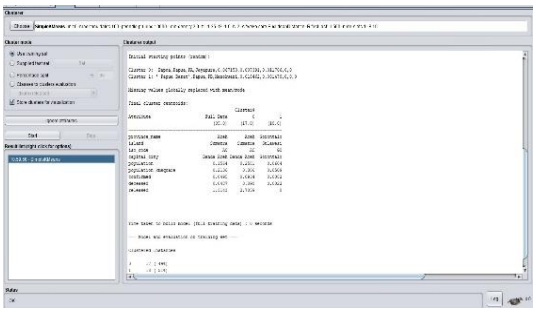
Gambar 4. Simulasi Algoritma Farthest First

Hasil simulasi algoritma Farthest First sebagai berikut:

```
=== Run information ===
Scheme: weka.clusterers.FarthestFirst
Relation:province_unsupervised.attribute.Normaliz
e-S1.0-T0.0
Instances: 35
Attributes: 9
  province_name
  island
  iso_code
  capital_city
  population
  population_kmsquare
  confirmed
  deceased
  released
Test mode: evaluate on training data
Clustering model (full training set)
FarthestFirst
==============
Cluster centroids:
Cluster 0
Riau Sumatra RI Pekanbaru 0.1344446368807179
0.04878048780487805 0.0016722408026755853
0.0 0.0
Cluster 1
DKI Jakarta Jawa JK Jakarta
0.21751777570935338 1.0 1.0 1.0 31.0
Time taken (full training data): 0 seconds
Model and evaluation on training set
Clustered Instances:
0 33 (94%)
1 2 (6%)
```

Simulasi Algoritma K-Means

Simulasi algoritma K-Means adalah mengklasifikasikan objek data yang diberikan ke kelompok *k* yang berbeda melalui metode iteratif yang cenderung konvergen ke minimum lokal (Gambar 5). Jadi hasil dari cluster yang dihasilkan adalah padat dan independen satu sama lain (Huang, 1998).



Gambar 5. Simulasi Algoritma K-Means

Hasil simulasi algoritma K-Means sebagai berikut:

```
=== Run information ===
Scheme: weka.clusterers.SimpleKMeans
Relation:province_unsupervised.attribute.Normaliz
e-S1.0-T0.0
 Instances: 35
 Attributes: 9
  province_name
  island
  iso_code
  capital_city
  population
  population_kmsquare
  confirmed
  deceased
  released
Test mode: evaluate on training data
Clustering model (full training set)
kMeans
======
Number of iterations: 4
Within cluster sum of squared errors:
124.64930333408826
Initial starting points (random):
Cluster 0:
'Papua,Papua,PA,Jayapura,0.067153,0.007391,0.01
1706,0,0
Cluster1:
'PapuaBarat',Papua,PB,Manokwari,0.018462,0.001
478,0,0,0
Final cluster centroids:
```

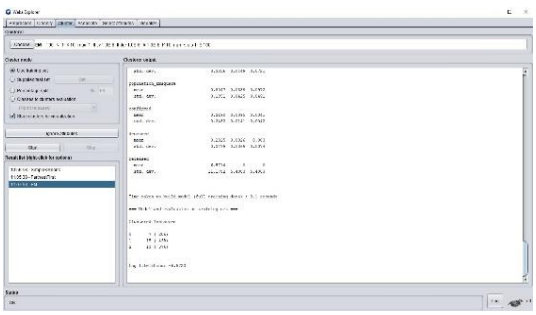
		Cluster#	
Attribute	Full Data	0	1
	(35.0)	(17.0)	(18.0)
province_name	Aceh	Aceh	Gorontalo
island	Sumatra	Sumatra	Sulawesi
iso_code	AC	AC	GO
capital_city	Banda Aceh	Banda Aceh	Gorontalo
population	0.1564	0.2581	0.0604
population_kmsquare	0.2136	0.386	0.0509
confirmed	0.0495	0.0934	0.0082
deceased	0.0487	0.098	0.0022
released	1.3143	2.7059	0

ISSN: 1979-8415
E-ISSN: 2714-8025

Time taken (full training data) : 0 seconds
Model and evaluation on training set
Clustered Instances:
0 17 (49%)
1 18 (51%)

Simulasi Algoritma EM (Expectation-Maximization)

Fungsi algoritma EM (*Expectation-Maximization*) (Gambar 6) adalah menemukan nilai estimasi *Maximum Likelihood* dari parameter dalam sebuah model probabilistik (manning, 2009).



Gambar 6. Simulasi Algoritma Expectation-Maximization

Hasil simulasi algoritma EM sebagai berikut:
=== Run information ===
Scheme:
weka.clusterers.EM -I 100 -N -1 -X 10 -max -1 -ll-cv 1.0E-6 -ll-iter 1.0E-6 -M 1.0E-6 -K 10 -num-slots 1 -S 100
Relation:province_weka.filters.unsupervised.attribute.
Normalize-S1.0-T0.0
Instances: 35
Attributes: 9
province_name
island
iso_code
capital_city
population
population_kmsquare
confirmed
deceased
released
Test mode: evaluate on training data
Clustering model (full training set)
EM
==
Number of clusters selected by cross validation: 3
Number of iterations performed: 0

Attribute	Cluster 0 (0.2)	1 (0.43)	2 (0.37)
province_name			
Aceh	1	1	2
Bali	2	1	1
Banten	2	1	1
Bengkulu	1	1	2
DI Yogyakarta	2	1	1
DKI Jakarta	2	1	1
Gorontalo	1	2	1
Jambi	1	1	2
Jawa Barat	2	1	1
Jawa Tengah	2	1	1
Jawa Timur	2	1	1
Kalimantan Barat	1	2	1
Kalimantan Selatan	1	1	2
Kalimantan Tengah	1	2	1
Kalimantan Timur	1	2	1
Kalimantan Utara	1	2	1
Kepulauan Bangka Belitung	1	1	2
Kepulauan Riau	1	1	2
Lampung	1	1	2
Maluku	1	2	1
Maluku Utara	1	2	1
Nusa Tenggara Barat	1	1	2
Nusa Tenggara Timur	1	1	2
Papua	1	2	1
Papua Barat	1	2	1
Riau	1	1	2
Sulawesi Barat	1	2	1
Sulawesi Selatan	1	2	1
Sulawesi Tengah	1	2	1
Sulawesi Tenggara	1	2	1
Sulawesi Utara	1	2	1
Sumatra Barat	1	1	2
Sumatra Selatan	1	1	2
Sumatra Utara	1	1	2
Unknown	1	2	1
[total]	42	50	48

Attribute	Cluster 0 (0.2)	1 (0.43)	2 (0.37)
island			
Sumatra	1	1	11
Kepulauan Nusa Tenggara	2	1	3
Jawa	7	1	1
Sulawesi	1	7	1
Kalimantan	1	5	2
Kepulauan Maluku	1	3	1
Papua	1	3	1

Attribute	Cluster 0 (0.2)	1 (0.43)	2 (0.37)
capital_city			
Banda Aceh	1	1	2
Denpasar	2	1	1
Serang	2	1	1
Bengkulu	1	1	2
Yogyakarta	2	1	1
Jakarta	2	1	1
Gorontalo	1	2	1
Jambi	1	1	2
Bandung	2	1	1
Semarang	2	1	1
Surabaya	2	1	1
Pontianak	1	2	1
Banjarmasin	1	1	2
Palangka Raya	1	2	1
Samarinda	1	2	1
Tanjung Selor	1	2	1
Pangkalpinang	1	1	2
Tanjungpinang	1	1	2
Bandar Lampung	1	1	2
Ambon	1	2	1
Sofifi	1	2	1
Mataram	1	1	2
Kupang	1	1	2
Jayapura	1	2	1
Manokwari	1	2	1
Pekanbaru	1	1	2
Mamuju	1	2	1
Makassar	1	2	1
Palu	1	2	1
Kendari	1	2	1
Manado	1	2	1
Padang	1	1	2
Palembang	1	1	2
Medan	1	1	2
Unknown	1	2	1
[total]	42	50	48

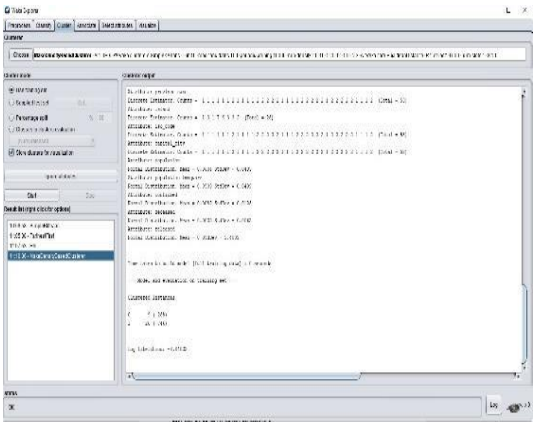
ISSN: 1979-8415
E-ISSN: 2714-8025

Attribute	Cluster 0 (0.2)	1 (0.43)	2 (0.37)
population			
mean	0.4579	0.0526	0.1139
std. dev.	0.3856	0.0449	0.0721
population_kmsquare			
mean	0.8047	0.0388	0.0972
std. dev.	0.1851	0.0425	0.0491
confirmed			
mean	0.2198	0.0095	0.0041
std. dev.	0.3482	0.0141	0.0042
deceased			
mean	0.2325	0.0026	0.003
std. dev.	0.3479	0.0069	0.0074

Time taken (full training data):
0.1 seconds
Model and evaluation on training set
Clustered Instances:
0 7 (20%)
1 15 (43%)
2 13 (37%)
Log likelihood: -6.5722

Simulasi Algoritma Make Density Based Cluster

Algoritma *Make Density Based Cluster* (Gambar 7) mengklasifikasikan objek data yang digunakan untuk mencari clusters yang betuknya berubah-ubah (*arbitary shape*) yang dimodelkan berupa daerah yang pada (*dense regions*) pada ruang data yang dipisahkan oleh daerah yang jarang (*sparse regions*) (Han, dkk., 2012).



Gambar 7. Simulasi Algoritma Make Density Based Cluster

Hasil simulasi algoritma Make Density Based Cluster sebagai berikut:
== Run information ==
Scheme:
weka.clusterers.MakeDensityBasedClusterer-M1.0E-6-W
Relation:Province.weka.filters.unsupervised.attribute.Normalize-S1.0-T0.0
Instances: 35
Attributes: 9
province_name
island
iso_code
capital_city
population
population_kmsquare
confirmed
deceased
released
Test mode: evaluate on training data
Clustering model (full training set)
MakeDensityBasedClusterer:
Wrapped clusterer:
kMeans
======
Number of iterations: 4
Within cluster sum of squared errors:
124.64930333408826

Initial starting points (random):
Cluster0:
Papua,Papua,PA,Jayapura,0.067153,0.007391,0.011706,0,0
Cluster1:
PapuaBarat',Papua,PB,Manokwari,0.018462,0.001478,0,0,0
Final cluster centroids:

Attribute	Full Data (35.0)	Cluster# 0 (17.0)	1 (18.0)
province_name	Aceh	Aceh	Gorontalo
island	Sumatra	Sumatra	Sulawesi
iso_code	AC	AC	GO
capital_city	Banda Aceh	Banda Aceh	Gorontalo
population	0.1564	0.2581	0.0604
population_kmsquare	0.2136	0.386	0.0509
confirmed	0.0495	0.0934	0.0082
deceased	0.0487	0.098	0.0022
released	1.3143	2.7059	0

Fitted estimators (with ML estimates of variance):
Cluster: 0 Prior probability: 0.4865

Time taken (full training data): 0 seconds
Model and evaluation on training set

Clustered Instances

0 9 (26%)
1 26 (74%)
Log likelihood: -8.84022

Tabel 1 menunjukkan ringkasan hasil analisis dari empat algoritma yang dikomparasi.

Tabel 1. Perbandingan hasil

Algoritma	Jumlah Cluster	Cluster Instance	Jumlah Iterasi	Time Taken
Farthest First	2	33 (94%) 2 (6%)	0	0 second
K-Means	2	17 (49% 18 (51%)	4	0 second
EM	3	7 (20%) 15 (43%) 13 (37%)	0	0,1 second
MDBC	2	9 (26%) 26 (74%)	4	0 second

KESIMPULAN

Dari hasil analisa WEKA, pada *K-Means*, 2 *cluster* dengan persentase *instance cluster* 17 (49%), 18 (51%), dan Farthest-First 2 *cluster* dengan persentase instance cluster 33 (94%), 2 (6%). K-Means mengambil titik-titik yang paling dekat terhadap pusat, sedangkan Farthest-First mengambil titik-titik terjauh. Farthest-First dapat menyelesaikan proses klasteringnya namun dengan kualitas yang lebih rendah dari K-Means. Pada EM (*Expectation-Maximization*), 3 cluster dengan persentase *instance cluster* 7 (20%), 15 (43%), dan 13 (37%). Pada *Make Density Based Clustered*, 2 cluster dengan persentase *instance cluster* 9 (26%), 26 (74%). Kedua metode pengelompokan yang diuji menunjukkan akurasi yang sama baik. Metode pengelompokan EM menunjukkan akurasi rendah (kurang dari 50%) dibandingkan Metode *Make Densitity Based Clusterd* yaitu presentase 74%.

DAFTAR PUSTAKA

Chauhan R, Kaur H, and Alam M A, Data Clustering Method for Discovering Clusters in Spatial Cancer Databases, *International Journal of Computer Applications*, 10(6), November 2010.

Asroni, A.R., 2015, Penerapan Metode K-Means Untuk Clustering Mahasiswa Berdasarkan Nilai Akademik dengan Weka Interface Studi Kasus Pada Jurusan Teknik Informatika UMM Magelang, *Jurnal Ilmiah Semesta Teknika*, 18(1): 76-82, Mei 2015.

Sindi, S., Ningse, W.R.O., Sihombing, I.A., Zer, F.I.R.H., Hartama, D., 2020, Analisis Algoritma K-Medoids Clustring Dalam Pengelompokan Penyebaran Covid-19 di Indonesia, *Jurnal Teknologi Informasi,* 4(1): 166-173.

Rahman, A.T., 2017, Coal Trade Data Clustering Using K-Means (Case Study PT. Global Bangkit Utama), *ITSMART: Jurnal Teknologi dan Informasi*, 6(1): 24-31.

Han, J. and Kamber, M., 2006. *Data Mining: Concepts and Techniques*, 2nd edition, San Francisco: Elsevier. Inc.

Berry, M.J. and Linoff, G.S., 2004, *Data Mining Techniques For Marketing, Sales, & Customer Relationship Management*, 2nd edition, Indiana: Wiley Publishing, Inc.

Sharma, N., Bajpai, A., and Litoriya, R., 2012, Comparison the various clustering algorithms of Weka Tools, *International Journal of Emerging Technology and Advanced Engineering*, 2(5), May 2012.

Jain K., and Dubes, R.C., 1988, Algorithms for Clustering Data, New Jersey: Prentice-Hall, Inc.

Witten, I.H., Frank, E., Trigg, T., Hall, M., 2002, Holmes,G., and Cunningham, S.J., 2002, *WEKA: Practical Machine Learning Tools and Techniques with Java Implementations*.

Pelleg D. and Moore, A., 2002, X-means: Extending K-means with Efficient Estimation of the Number of Clusters, *CEUR Workshop Proc*., 1542, January 2002, pp. 33-36.

https://waikato.github.io/weka-wiki/documentation/, 21 Juni 2020.

Eibe, F., 2011, *Machine Learning with WEKA,* Department of Computer Science, University of Waikato, New Zealand.

https://www.kaggle.com/ardisragen/indonesia-coronavirus-cases, 16 Juni 2020.

Huang, Z., 1988, *Extensions to The K-Mens Algorithm for Clustering Large Data Sets*

with Categorical Values, *Data Mining and Knowledge Discovery*, 2:283-304.

Manning, C.D., 2009, *An Introduction to Information Retrieval*, Cambridge University Press Cambridge, England.

BIODATA PENULIS

Trientje Marlein Tamtelahitu, S.Kom, M.Kom., lahir di Ambon tanggal 11 Oktober 1977, menyelesaikan pendidikan S1 bidang ilmu Teknik Informatika dari Institut Sains dan Teknologi Palapa tahun 1999, dan S2 bidang ilmu Sistem Informasi dari Universitas Diponegoro tahun 2011. Saat ini tercatat sebagai Dosen Tetap di Universitas Kristen Indonesia Maluku dengan jabatan akademik Asisten Ahli pada bidang minat Informatika.