Algoritma *ClusterMix K-Prototypes* Untuk Menangkap Karakteristik Pasien Berdasarkan Variabel Penciri Mortalitas Pasien Dengan Gagal Jantung

Raditya Novidianto[1*] dan Kartika Fithriasari[2]
[1,2] Departemen Statistika, Fakultas Sains dan Analitika Data, Institut Teknologi Sepuluh Nopember
*Corresponding author: radit@bps.go.id

Received: 18 January 2021
Accepted: 27 March 2021
Published: 31 March 2021

ABSTRAK – ***Cardiovascular Disease* (CVD) atau penyakit kardiovaskular adalah salah satu penyebab utama kematian cukup besar di seluruh dunia yang berujung pada kejadian gagal jantung. Organiasasi kesehatan WHO menyebutkan jumlah orang yang meninggal karena penyakit kardiovaskuler akibat gagal jantung setiap tahun memiliki rata-rata 17,9 juta kematian setiap tahunnya, yaitu sekitar 31 persen dari total kematian secara global. Pendeteksian faktor mortalitas pasien gagal jantung perlu dibentuk segmentasi yang berguna untuk memperkecil peluang terjadinya kematian akibat gagal jantung. Salah satunya dengan menggunakan variabel penciri mortalitas akibat gagal jantung dengan cara menerapkan algoritma *k-prototypes*. Hasil penggerombolan terbentuk 2 kluster yang dianggap optimal berdasarkan nilai koefisien *silhouette* tertinggi yaitu sebesar 0.5777. Hasil penelitian dilakukan segementasi pasien dengan variabel penciri mortalitas pasien gagal jantung yang menunjukan bahwa kluster 1 merupakan gerombol pasien yang memiliki resiko rendah terhadap peluang mortalitas akibat gagal jantung dan kluster 2 merupakan gerombol pasien dengan karaktistik pasien dengan resiko yang tinggi terhadap peluang mortalitas akibat gagal jantung. Segementasi tersebut didasari dari nilai rata-rata setiap variabel penciri dari faktor mortalitas gagal jantung pada setiap kluster yang dibandingkan dengan kondisi normal pada variabel *serum creatine, ejection fraction*, usia, *serum sodium*, tekanan darah, anemia, *creatinine phosphokinase*, *plateles*, merokok, jenis kelamin dan diabetes.**

Keywords– **Penyakit kardiovakuler, ClusterMix, Algoritma K-Prototype, Koefisien *Silhouette*.**

I. PENDAHULUAN

Cardiovascular Disease (CVD) atau penyakit kardiovaskuler adalah salah satu penyebab utama kematian dan kecacatan cukup besar di seluruh dunia karena adanya gangguan pada jantung dan pembuluh darah termasuk penyakit jantung koroner, stroke, gagal jantung dan jenis patologi lainnya. Sehingga persepsi antara dokter dan pasien diperlukan untuk melakukan lebih banyak perhatian agar tingkat kesembuhan penyakit ini dapat diatasi dengan baik [1]. Penyakit kardiovaskuler adalah penyakit mematikan karena data dari Organisasi Kesehatan Dunia (WHO) menyebutkan bahwa orang meninggal karena penyakit kardivaskuler setiap tahun memiliki rata-rata 17,9 juta per tahun, yaitu sekitar 31 persen dari total kematian secara global [2]. Penyakit kardiovaskuler membuat otot jantung bekerja lebih cepat sehingga menyebabkan gagal jantung. Keadaan organ jantung pada manusia secara usia perlahan akan semakin melemah dan semakin lama akan semakin berat untuk memompa darah dengan sebagaimana mestinya. Saat kondisi jantung melemah, terdapat zat tertentu yang akan dilepaskan didalam darah. Salah satunya dikarenakan pasien memiliki riwayat penyakit bawaan seperti anemia, penyakit diabetes, tekanan darah, penyakit lainnya dan faktor lainnya sehingga zat tertentu tersebut memiliki efek racun didalam darah sehingga dapat menyebabkan kondisi gagal jantung [3].

Gagal jantung merupakan suatu keaadan dimana otot-otot dinding jantung mulai mengendur, membesar dan mulai membatasi memompa darah ke jantung [4]. Penyebab terjadinya gagal jantung dapat karena *ejection fraction* yaitu proporsi darah yang dipompa keluar dari jantung selama satu kontraksi dengan nilai persentase berkisar antara 50% dan 75%. Beberapa penyabab terjadinya kondisi gagal jantung yaitu dapat dikarena berkurangnya *ejection fraction* (HFrEF), biasanya dikenal sebagai gagal jantung karena disfungsi *systolik* ventrikel kiri atau gagal jantung *systolik*, ditandai dengan *ejection fraction* yang lebih kecil dari 40%. Selanjutnya gagal jantung dengan *ejection fraction* yang stabil (HFpEF), biasanya disebut juga gagal jantung *diastolik* atau gagal jantung dengan *ejection fraction* normal. Dalam hal ini, ventrikel kiri berkontraksi normal selama *systole*, tetapi ventrikel kaku dan gagal rileks secara normal selama *diastole*, sehingga mengganggu pengisian [5].

Jantung merupakan organ vital yang paling penting karena fungsinya berhubungan dengan peluang seseorang untuk hidup. Menganalisis kelangsungan hidup pasien gagal jantung menjadi prioritas bagi dokter yang bertujuan untuk memperbaiki kondisi kesehatan pasien, tetapi sampai saat ini tindakan penyembuhan secara klinis pasien gagal jantung cenderung masih tergolong minim karena karakteristik pasien gagal jantung sangat susah dideteksi [6].

Catatan kesehatan pasien atau biasa disebut *Electronic Health Records* (EHR) merupakan catatan rekamedis yang digunakan sebagai sumber informasi mengenai karakteristik dari pasien gagal jantung sehingga dapat diketahui atau didalami peran karakteristik demografi dan variabel lainnya baik secara langsung maupun tak langsung dalam sebuah praktik klinis kesembuhan pasien gagal jantung [7]. Sebuah studi mempelajari pola umum kelangsungan hidup yang

menunjukkan intensitas mortalitas yang tinggi pasien gagal jantung pada hari-hari awal dan kemudian meningkat secara bertahap hingga akhir penelitian [8].

Sehingga faktor mortalitas pasien gagal jantung dapat modelkan dengan mempertimbangkan usia, *ejection fraction*, *serum creatine*, *serum sodium*, anemia, *plateles*, *creatinine phosphokinase*, tekanan darah, jenis kelamin, diabetes, dan status merokok berpotensi berkontribusi pada kematian [4]. Peran besar pada variabel faktor mortalitas pasien gagal jantung tergambar melalui sebuah algoritma pada *machine learning* sehingga diperoleh *importance variabel* (Variabel Penting) penentu kejadian mortalitas gagal jantung terurut yaitu *serum creatine, ejection fraction,* usia, *serum sodium*, tekanan darah, anemia, *creatinine phosphokinase*, *plateles*, merokok, jenis kelamin dan diabetes [9]. Penelitian lanjutan adalah menentukan sebuah segemen dari faktor mortalitas dari pasien gagal jantung untuk melihat karakter dari pasien gagal jantung berdasarkan tingkat *similarity* atau kesamaan dengan asumsi yang muncul dalam satu kluster merupakan pasien yang memiliki karakteristik yang cenderung homogen dalam kluster dan haterogen antar kluster [10].

Saat pengelompokkan *dataset* dengan unit observasi yang besar maka metode yang sering digunakan merupakan metode *k-means* atau *k-modes* sehingga dapat terebentuk segmentasi dari karakteristik setiap kluster [11]. Adapun kelamahan dalam menggunakan metodologi tersebut yiatu jumlah kluster yang perlu ditentukan terlebih dahulu sebelum algortima tersebut diterapkan dan metode *k-means* hanya bisa digunakan pada data kontinyu serta *k-modes* hanya bisa digunakan pada data kategorik [12]. Dalam dunia nyata tipe data sangatlah luas bahkan cenderung pada tipe data yang bersifat campuran, sehingga terdapat algoritma modifikasi *k-means* dan *k-modes* untuk mengintegerasikan algoritma tersebut maka dibangun algoritma *clusterMix K-Prototypes* [13].

Berdasarkan latar belakang di atas, peneliti tertarik untuk membahas mengenai analisis klaster data campuran atau biasa yang disebut *ClusterMix* dengan algoritma data campuran dengan menggunakan *k-prototypes* pada proses pengelompokkan pasien CVD yang mengalami gagal jantung. Validitas pengukuran kemiripan pada penelitian ini berdasarkan koefisien *silhoutte* untuk mendapatakan *k* atau jumlah kelompok yang optimum. Tujuan dari penelitian ini adalah memperoleh hasil pengelompokkan yang optimal pada proses pengelompokkan pasien gagal jantung dan membentuk segementasi pasien berdasarkan kemiripan variabel untuk kepentingan penanganan pasien gagal jantung lebih lanjut. Pentingnya informasi tentang hasil kluster ini dapat membantu pada tenaga medis dalam mengambil tindakan berdasarkan segmentasi yang terbentuk pada pasien gagal jantung sehingga kejadian gagal jantung dapat diminimalisir. Bagi dunia *machine learning* hal ini merupakan media pembelajaran dengan metode baru yaitu *ClusterMix* dalam membentuk segmentasi dengan data campuran.

II. TINJAUAN PUSTAKA

A. Statistik Deskriptif

Statististik deskriptif merupakan salah satu metode dasar yang digunakan untuk menggambarkan suatu keadaan tertentu dengan cara mengumpulkan, mengolah hingga mendesiminasikan hasil dari pengumpulan data [14]. Pada penelitian ini akan digunakan metode deskriptif seperti rata-rata, standard devaiasi, median dan modus untuk mendeskripsikan variabel penciri pasien yang mengalami gagal jantung. Statistik deskriptif akan disajikan dalam bentuk tabel, gambar, *heatmap*, grafik maupun *boxplot*.

B. Analisis Kluster

Analisis kluster merupakan salah satu teknik peubah ganda yang bertujuan untuk mengelompokkan sejumlah objek berdasarkan kemiripan karakteristik yang dimilikinya. Objek yang terkelompok dalam satu kluster memiliki tingkat kemiripan yang tinggi dan objek antar kluster memiliki tingkat kemiripan yang rendah [15]. Langkah secara umum dilakukan dalam analisis kluster yaitu menentukan ukuran kemiripan, metode penggerombolan, melakukan penggerombolan dan yang terakhir yaitu interpretasi hasil penggerombolan [16].

C. Koefisien Silhoutte

Koefisien *silhouette* merupakan metode yang sering digunakan dalam analisis kluster untuk menentukan jumlah *k* (jumlah kluster) yang tepat dalam proses *clustering* [17]. Koefisien *silhoutte* dapat juga digunakan untuk mengukur kualitas kluster yang telah terbentuk [18]. Pengukuran koefisien *silhouette* dirumuskan sebagai berikut [19]

$$S_i = \frac{b_i - a_i}{\max(a_i b_i)} \qquad (1)$$

Dimana,

a_i : Jarak rata-rata antara objek *i* dengan seluruh objek yang berada dalam satu kluster yang sama

b_i : Jarak rata-rata antara objek *i* dengan seluruh objek yang berada pada kluster terdekat

D. Ukuran Kemiripan (Similarity)

Ukuran kemiripan digunakan pada analisis kluster dengan menggunakan jarak antar objek dan jarak antar kluster. Kerena penelitian ini menggunakan data campuran maka jarak yang digunakan sebagai berikut.

2.1 Jarak Eucledean

Jarak *euclidean* digunakan untuk mengukur jarak antar objek dengan data bertipe numerik. Salah satu penggunaan yaitu pada algoritma *k-means*. Jarak *euclidean* antar objek ke-*i* dan objek ke-*j* dengan *p* variabel adalah sebagai berikut [20].

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{1/2} \quad (2)$$

Dengan :

d_{ij} : jarak *euclidean* antara objek ke-*i* dengan objek ke-*j*
x_{ik} : nilai objek ke-*i* pada peubah ke-*k*
x_{jk} : nilai objek ke-*j* pada peubah ke-*k*
p : banyaknya peubah yang diamati

2.2 Jarak Tipe Data Kategorik

Algoritma *k-modes* digunakan untuk menggerombolkan semua data yang bertipe kategorik. Ukuran jarak yang digunakan algoritma *k-modes* adalah sebagai berikut [21].

$$d_2(X,Y) = \sum_{j=1}^{p} (x_j - y_j)^2 + \gamma \sum_{j=p+1}^{m} \delta(x_j, y_j) \quad (3)$$

$d_2(X,Y)$ = Ukuran jarak antara objek X dan Y (data campuran)
$\sum_{j=1}^{p}(x_j - y_j)^2$ = Ukuran jarak untuk data yang bertipe numerik
$\sum_{j=p+1}^{m} \delta(x_j, y_j$ = Ukuran jarak untuk data yang bertipe kategorik
γ = Parameter penimbang

E. Algortima K-Means

Algoritma *k-means* merupakan metode penggerombolan tidak berhirarki yang menentukan penggerombolan setiap objek berdasarkan nilai rataan terdekat. Langkah-langkah algoritma *k-means* yaitu membagi objek dalam *k* inisial kluster, melakukan penggerombolan objek kedalam kluster yang memiliki nilai rata-rata terdekat, menghitung kembali nilai rataan untuk kluster yang menerima objek baru maupun kehilangan objek dan pengulangan sampai tidak ada lagi perpindahan objek [22].

F. Algoritma K-Modes

Algoritma *k-modes* menggunakan ukuran jarak untuk data kategorik. Tahapan algoritma *k-modes* yaitu menentukan *k* inisial modus untuk setiap kluster, mengalokasikan objek ke kluster berdasarkan modus terdekat, uji kembali jarak objek terhadap modus terakhir, alokasi kembali jika ada objek yang dekat dengan kluster lain dan ulangi sampai tidak ada objek yang berganti kluster [21]

G. Algoritma K-Prototypes

Algoritme *k-prototypes* menggunakan ukuran jarak campuran yang dicirikan oleh γ. Variabel γ merupakan parameter penimbang yang digunakan untuk menyeimbangkan proporsi dua fungsi jarak untuk data bertipe numerik dan kategorik. Algoritma *k-prototypes* dapat dilakukan dengan tahapan sebagai berikut [23]

1. Menentukan banyaknya kluster (k) yang akan dibentuk. Batas minimum ukuran k adalah $\sqrt{n}$ kluster, sedangkan batas maksimum k adalah atau $n/2$ dimana n merupakan banyaknya amatan.
2. Menentukan k inisial *prototypes* yaitu $Z_1, Z_2, \ldots, Z_k$ sebagai pusat kluster di masing-masing kluster.
3. Melakukan perhitungan jarak semua observasi pada dataset terhadap inisial kluster awal. Ukuran jarak yang digunakan adalah ukuran jarak campuran.
4. Mengalokasikan semua observasi ke dalam kluster yang memiliki jarak *prototype* terdekat dengan objek yang diukur.
5. Melakukan perhitungan titik pusat kluster yang baru setelah semua objek dialokasikan,
6. Merealokasikan semua data observasi pada dataset terhadap *prototype* yang baru.

Apabila titik pusat kluster tidak berubah atau sudah konvergen maka proses algoritma berhenti. Tetapi apabila pusat masih berubah-ubah secara signifikan maka proses kembali ke tahap 2 sampai 5 hingga iterasi maksimum tercapai atau sudah tidak ada lagi perpindahan objek.

III. METODOLOGI

A. *Sumber Data*

Data yang digunakan dalam penelitian ini merupakan catatan *EHR* (Rekam Medis) 299 pasien gagal jantung yang dikumpulkan *Faisalabad Institute of Cardiology* dan *Allied Hospital* di Faisalabad (Punjab, Pakistan), selama bulan April – Desember 2015 [4] [24]. Pasien terdiri dari 105 wanita dan 194 pria, dan usia berkisar antara 40 dan 95 tahun. Kumpulan data tersebut merupakan data yang berisikan 12 variabel dari hasil *EHR* yang dilaporkan seperti informasi klinis, tubuh, dan gaya hidup. Data tersebut dijelaskan secara singkat beberapa variabel yang berbentuk biner yaitu anemia, tekanan darah, diabetes, jenis kelamin, dan kebiasaan merokok.

B. *Variabel Penelitian*

Dalam penelitian ini terdapat 12 variabel yang berasal dari variabel penciri faktor yang mempengaruhi tingkat mortalitas pasien gagal jantung berdasarkan penelitian terdahulu yang telah melakukan ketepatan klasifikasi dan akurasi berdasarkan *attribute variabel importance* tertinggi yang didapatkan dari algoritma *random forest* [4]. Proses segementasi untuk memisahkan karakteristik pasien gagal jantung menggunakan variabel penciri mortalitasnya sebagai variabel *input* untuk dilakukan pengelompokkan objek berdasarkan *similarity* sebagai berikut

Tabel 1. Variabel Penelitian

Kode	Variabel	Definisi	Satuan	Batas	Jenis Data
X1	Usia	Usia Pasien (tahun)	Years	[40,..., 95]	Numerik
X2	Anemia	Memiliki 2 nilai (Ya/Tidak)	*Boolean*	0, 1	Kategorik
X3	Tekanan Darah	Tipe Data memiliki 2 nilai	*Boolean*	0, 1	Kategorik
X4	*Creatinine phosphokinase* (CPK)	Kadar kreatin dalam darah (mcg/L)	mcg/L	[23,..., 7861]	Numerik
X5	Diabetes	Memiliki 2 nilai (Ya/Tidak)	*Boolean*	0, 1	Kategorik
X6	*Ejection fraction*	Persen darah meninggalkan jantung saat kontraksi (%)	Percentage	[14,..., 80]	Numerik
X7	Jenis Kelamin	Memiliki 2 nilai (Ya/Tidak)	Binary	0, 1	Kategorik
X8	*Platelets*	Kadar Trombosit dalam darah kiloplateles/mL	kiloplatelets/mL	[25.01,..., 850.00]	Numerik
X9	*Serum creatinine*	Tingkat Kreatin dalam darah (mg/dL)	mg/dL	[0.50,..., 9.40]	Numerik
X10	*Serum sodium*	Tingkat sodium dalam darah mEq/L	mEq/L	[114,..., 148]	Numerik
X11	Merokok	Memiliki 2 nilai (Ya/Tidak)	*Boolean*	0, 1	Kategorik
X12	Waktu (Periode Tindakan)	Periode Waktu tindakan (Hari)	Days	[4,...,285]	Numerik

C. *Langkah Analisis*

Langkah analisis yang digunakan dalam penelitian ini sebagai berikut

1. Melakukan tahap *pre-processing* yaitu dengan eksplorasi data dengan tahapan *cleaning data* yang meliputi memeriksa terhadap data hilang atau *missing data*, kemudian mengeksplorasi data sehingga dapat diketahui gambaran deskriptif mengenai karakteristik pasien gagal jantung serta yang terakhir melakukan proses *standardization* atau pembakuan untuk data numerik.
2. Menentukan banyaknya kluster atau *k* dengan menghitung koefisien *silhouette*.
3. Melakukan penggerombolan pasien gagal jantung berdasarkan variabel penciri menggunkan algoritma *k-prototypes* dengan ukuran jarak campuran.
4. Melakukan interpretasi terhadap hasil kluster optimal yang didapatkan.

IV. ANALISIS DAN PEMBAHASAN

Berikut merupakan analisis pembahasan mengenai variabel penciri dari pasien gagal jantung disebuah Rumah Sakit di Pakistan dapat tergambar melalu sebaran data berdasarkan statistik deskriptif dan analisis lebih dalam dengan menggunakan algoritma *k-prototypes.*

A. *Statistik Deskriptif*

Dalam penelitian ini dilakukan analisis secara deskriptif mengenai variabel penciri dari pesien penderita gagal jantung yang tergambar dari objek yang diteliti. Sebelum melakukan analisis lebih dalam maka dilakukan pengecekkan terhadap data yang digunakan dan disusun secara deskriptif dengan menggunakan tabel sebagai berikut

Tabel 2. Deskriptif Variabel Numerik

Variabel	Jumlah Obeservasi	Nilai Minimal	Nilai Maximum	Rata-rata	Standard Deviasi
X1	299	40	95	60,83	11,8
X4	299	23	7861	581,84	970,2
X6	299	14	80	38,08	11,8
X8	299	25.100	850.000	263.358	97.804,2
X9	299	1	9	1,39	1,03
X10	299	113	148	136,63	4,4
X12	299	4	285	130,26	77,6

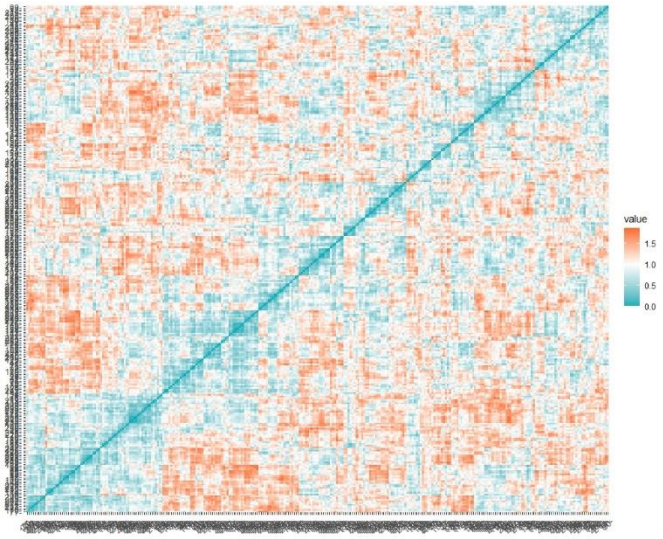
Tabel 2 menunjukkan gambaran dari variabel numerik pada penelitian ini yaitu usia pasien, *CPK* , *ejection fraction*, *plateles, serum creatinine, serum sodium* dan waktu. Pada tabel menunjukkan bahwa semua nilai observasi tidak terdapat data yang *missing value* sehingga dapat dilakukan observasi lebih lanjut. Hasil ekplorasi data terdapat observasi yang dianggap dianggap *outlier* melalui gambar pada *boxplot* yaitu pada variabel *CPK*, *platelets*, *serum creatinine* dan *serum sodium*. Karena data pasien merupakan data riil lapangan dan tidak mudah didapatkan sehingga observasi yang dianggap *outlier* tersebut tetap dimasukkan kedalam penelitian. Pada variabel yang bertipe kategorik ditunjukkan pada tabel berikut

Tabel 3. Deskriptif Variabel Kategorik

Variabel	Jumlah Obeservasi	Nilai Minimal	Nilai Maximum	Persen Kategori "0"	Persen Kategori "1"
X2	299	0	1	56,90	43,1
X3	299	0	1	58,20	41,8
X5	299	0	1	64,90	35,1
X7	299	0	1	35,10	64,9
X11	299	0	1	67,90	32,1

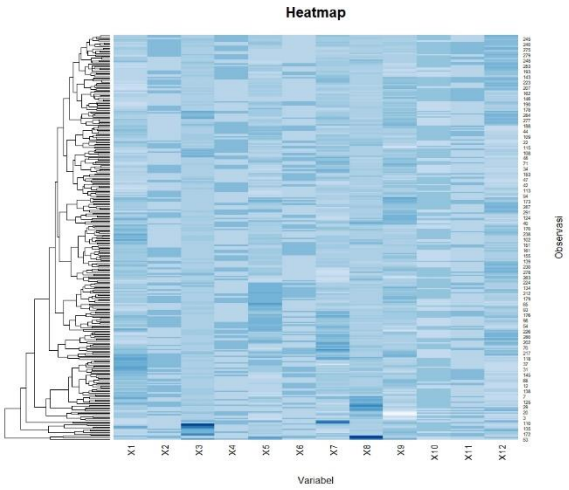
Tabel diatas menunjukkan hasil ekplorasi data kategorik pada penelitian ini sebanyak 5 variabel. Hasil ekplorasi menujukkan bahwa tidak terdapat *missing value* pada variabel anemia, tekanan darah, diabetes, jenis kelamin dan kebiasaan merokok. Variabel anemia (X2), tekanan darah (X3), diabetes (X5) dan merokok (X11) merupakan variabel dengan besarnya persentasi berkode 0 lebih besar dari pada yang berkode 1, sedangkan untuk variabel jenis kelamin (X7) yaitu sebagian besar jenis kelamin menujukkan bahwa pasien laki-laki lebih banyak dibandingkan pasien yang berjenis kelamin perempuan. Pada sebagian besar pasien gagal jantung memiliki minimal satu penyakit bawaan seperti contoh diabetes, tekanan darah atau anemia, namun juga terdapat pasien yang tidak memiliki penyakit bawaan.

Dari 299 objek yang diteliti maka akan dilakukan analisis berdasarkan visual mengenai pola hubungan antar objek dengan menggunakan *heatmap*, sehingga bisa diketahui pola hubungan anatar observasi berdasarkan kekuatan variabel penciri setiap pasien sehingga peneliti bisa melihat pola kluster lebih awal sebelum masuk kedalam algoritma *k-prototypes*. Data yang divisualisasikan merupakan data yang telah dilakukan *standardization* pada data numerik yang bertujuan untuk menghilangkan perbedaan satuan pada variabel numerik sehingga tidak terlihat dominan. Proses pengelompokan yaitu mendeteksi observasi melalui *similarity* dengan menggunakan matriks jarak sebagai berikut



Gambar 1. *Heatmap* matriks jarak antar observasi

Proses penglompokan dalam data yang berukukan *n x n* dengan jumlah *n* adalah 299 dapat diterlihat dari tingkat kesamaannya dengan menggunakan matriks jarak dengan menggunakan perhitungan jarak yaitu jarak *euclidean*. Kemudian matriks jarak tersebut digambarkan secara visual dengan menggunakan *heatmap* seperti gambar 1 menunjukkan proses penggerombolan dengan menggunakan *heatmap* dengan warna yang bergradasi. Dalam *heatmap* tersebut dapat terlihat objek mana saja yang memiliki kesamaan yang kuat antar observasi yang satu dengan observasi yang lainnya. setelah mengindetifikasi dilanjutkan menggunakan *heatmap* pengelompokkan berdasarkan kesamaan variabelnya yang ditunjukkan secara berhiraki sebagai berikut

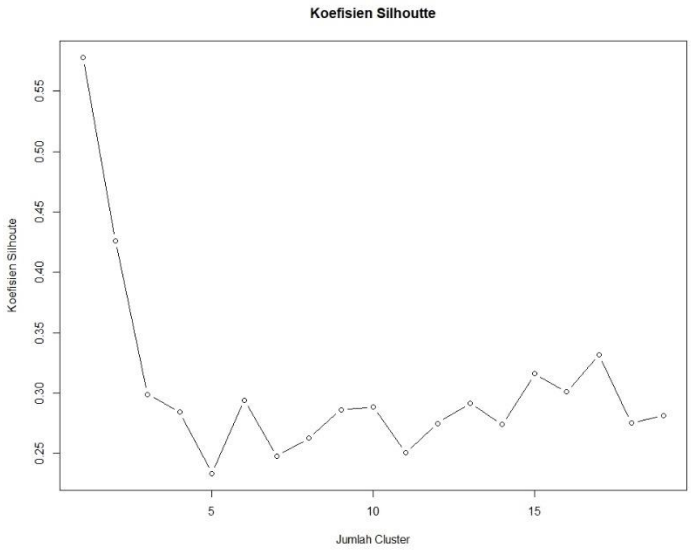


Gambar 2. *Heatmap* pengelompokkan observasi secara hirarki

Gambar 2 menunjukkan bahwa Setalah mendapatkan matriks jarak kemudian dilakukan algoritma pengelompokkan berdasarkan tingkat kesamaannya secara hirarki. Pada gambar 2 terlihat proses dari *n* obeservasi dilakukan pengelompokkan berbentuk dendogram pada sumbu sebelah kiri yang didasarkan pada nilai kesamaan dari variabel yang diteliti yaitu variabel penciri mortalitas pasien gagal jantung. Pada gambar belum ada proses penentuan jumlah *k* secara optimum, namun hanya pengelompokkan berdasar kepada *similarity* dari matriks jarak.

B. Algoritma ClusterMix K-Prototypes

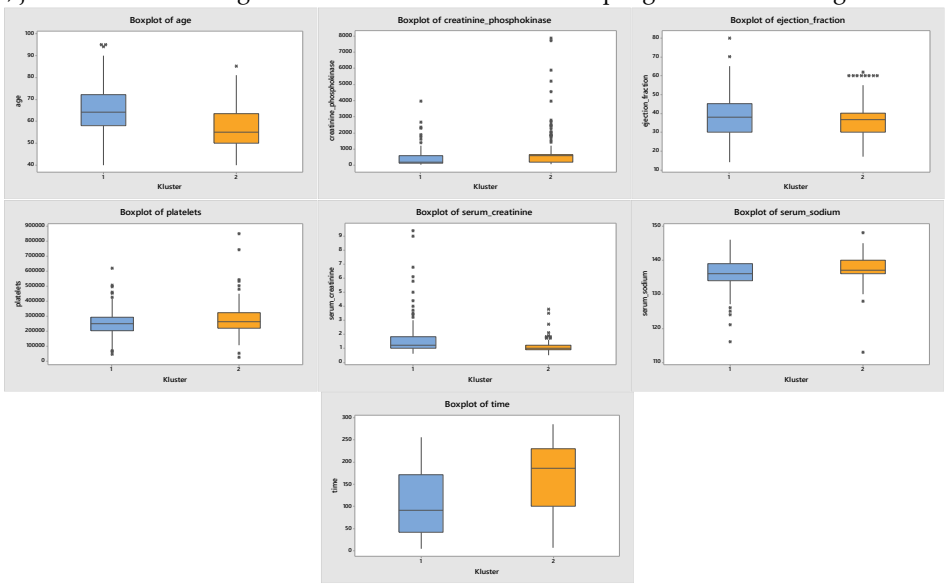
Algortima *k-prototypes* merupakan salah satu teknik dalam melakukan penggerombolan nonhirarki pada suatu objek berdasarkan variabel penciri yang telah tentukan, sehingga perlu penentuan jumlah kluster diawal sebelum proses algoritma dijalankan. Penentuan jumlah kluster yang berbeda akan menghasilkan kesimpulan atau deskripsi kluster yang berbeda. Penentuan jumlah kluster yang optimal merupakan suatu segmentasi yang benar-benar bermakna dan menggambarkan kondisi yang sebenarnya. Penetuan banyaknya kluster diperoleh dari cara mengevaluasi perhitungan koefisien *silhouette* pada setiap kluster yang terbentuk sehingga dapat diperoleh hasil penggerombolan yang homogen dalam satu kluster dan heterogen antar kluster. Namun dalam beberapa penelitian penentuan jumlah kluster juga dapat dilakukan secara subjektif oleh peneliti tergantung tujuan dari peneliti untuk menggambarkan suatu observasi dari objek yang di teliti. Semakin besar nilai koefisien *silhouette* maka kluster tersebut merupakan jumlah kluster yang optimal dengan anggapan semakin homogen kluster yang terbentuk maka semakin tinggi tingkat korelasi objek yang ada didalam sehingga nilai koefisien *silhouette* juga akan semakin tinggi. Berikut perhitungan koefisisen *silhouette* pada dengan batasan perhitungan minimal kluster yang terbentuk yaitu 2 dan maksimal 20 kluster yang terbentuk sebagai berikut



Gambar 3. Koefisien *Silhoutte*

Pemilihan jumlah kluster dilakukan secara bertahap dimulai dari jumlah kluster sebanyak 2 hingga 20 kluster. Terjadi fluktusi nilai koefisien *silhouette* pada setiap tahap jumlah kluster atau nilai *k*. Gambar 3 menunjukkan adanya penurunan nilai koefisien *silhouette* dari k=2 dengan nilai koefisien 0,5777 hingga k=5 dengan nilai koefisien 0,23361 , kemudian mengalami peningkatan pada k=6 hingga k=7, kemudian cenderung turun dari k=8 hingga k=11 dan seterusnya sesuai dengan gambar 3. Koefisien *silhouette* terbesar dihasilkan pada k=2, sehingga ditetapkan sebagai jumlah kluster yang

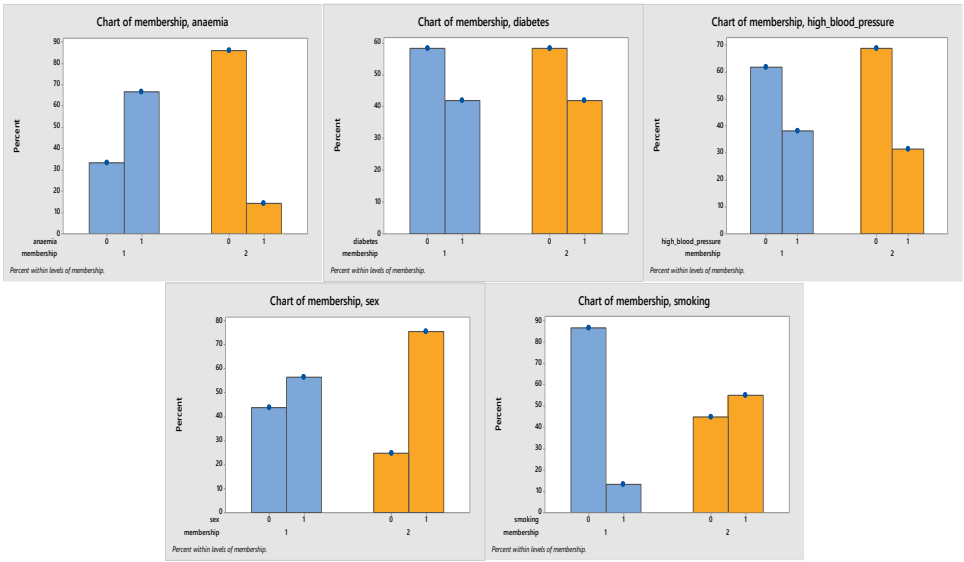
optimal. Pembentukan kluster pada algoritma *k-prototypes* ditentukan juga koefisien penimbang (γ) sesuai dengan persamaan 3. Berdasarkan hasil pengolahan, diperoleh koefisien penimbang (γ) yang sama pada setiap tahapan jumlah kluster dari k=2 sampai k=20 yaitu sebesar 2,1509. besarnya nilai koefisien penimbang (γ) ditentukan oleh banyaknya jumlah objek, jumlah variabel kategorik dan variabel numerik. Hasil pengolahan data sebagai berikut



Gambar 4. *Boxplot* k=2 untuk variabel numerik

Gambar 4 menunjukkan hasil pengolahan sesuai dengan k=2 dengan koefisien penimbang 2,1509 digambarkan dari *boxplot.* Terbentuk 2 kluster dengan jumlah anggota kluster 1 yaitu sebanyak 81 observasi dan anggota kluster 2 yaitu sebanyak 218 observasi. Visualisasi terlihat distribusi dalam sebuah *boxplot* menunjukkan bahwa perbedaannya cukup kecil persebarannya untuk variabel numerik, yang cukup terlihat perbedaannya yaitu pada variabel waktu, *serum sodium, ejection fraction* dan usia. Gambar 4 menunjukkan bahwa *boxplot* setelah terbentuk kluster memiliki distribusi yang cukup beragam dan observasi yang *outlier* didalamnya.

Variabel usia pasien menunjukkan bahwa rata-rata kluster 1 sebesar 64,71 tahun lebih besar dari kluster 2 yang memiliki rata-rata usia pasien sebesaar 56,060 tahun. Kadar enzim dalam darah atau CPK pada pasien pada kluster 1 memiliki rata-rata 387 mcg/L sedangkan pada kluster 2 memiliki rata-rata 822 mcg/L. Rata-rata persentasi darah meninggalkan jantung pada satu kali kontraksi atau disebut *ejection fraction* pada pasien kluster 1 sebesar 38,67 persen lebih tinggi dibanding kluster 2 sebesar 37,36 persen. Variabel *platelets* atau merupakan kadar trombosit dalam darah pada kluster 1 memiliki rata-rata kadar trombosit sebesar 252.448 kiloplatelets/mL lebih kecil dibandingkan pada kluster 2 yaitu sebesar 276.792 kiloplatelets/mL. Kadar *cratine* dalam dalah pada kluster 1 memiliki rata-rata sebesar 1,612 mg/dL lebih besar dari kluster 2 yang memiliki rata-rata sebesar 1,1255 mg/dL. Level *sodium* dalam darah pada pasien dengan kluster 1 memiliki rata-rata sebesar 136,01 mEq/L lebih rendah dibandingkan dengan pasien pada kluster 2 yang meiliki level *sodium* dalam darah sebesar 137,39 mEq/L. Kemudian yang paling akhir yaitu waktu tindakan pada pasien gagal jantung pada kluster 1 menunjukkan lamanya tindakan pada pasien gagal jantung memiliki rata-rata sebesar 103 hari lebih singkat dibanding pasien kluster 2 yang memiliki rata lamanya tindakan terhadap pasien gagal jantung sebesar 163 hari.



Gambar 5. *Barchart* variabel kategorik

Gambar 5 menujukkan distribusi observasi terhadap kluster yang telah dibentuk. Pada variabel anemia menunjukkan bahwa kluster 1 sebagian besar pasien memiliki penyakit anemia sedangkan pada kluster 2 sebagian besar tidak memiliki penyakit anemia. Variabel diabetes dan variabel tekanan darah menunjukkan bahwa kluster 1 dan kluster 2 memiliki pola yang sama bahwa sebagian besar pasien tidak memiliki penyakit bawaan diabetes dan tekanan darah. Variabel jenis kelamin kluster 1 dan kluster 2 sebagian besar pasien berjenis kelamin laki-laki. Variabel kebiasaan merokok kluster 1 sebagian besar pasien tidak memiliki kebiasaan merokok dan pada kluster 2 sebagian besar memiliki kebiasan merokok.

Tabel 4. Perbandingan variabel untuk semua pasien dengan kluster

Variabel	Satuan	Seluruh Pasien	kluster 1	Kluster 2
Usia	Tahun	60,83	64,711	56,06
Anaemia	*Boolean*	0	1	0
High blood pressure	*Boolean*	0	0	0
Creatinine Phosphokinase (CPK)	mcg/L	581,84	387	822
Diabetes	*Boolean*	0	0	0
Ejection fraction	Persen	38,08	38,67	37,366
Jenis Kelamin	*Binary*	1	1	1
Platelets	kiloplatelets/mL	263.358,03	252.448	276.792
Serum creatinine	mg/dL	1,39	1,612	1,1255
Serum sodium	mEq/L	136,63	136,01	137,39
Merokok	*Boolean*	0	0	1
Waktu	Hari	130,26	103,01	163,82

Usia pada kluster 1 menunjukkan rata-rata sebesar 64,71 tahun dan pada kluster 2 sebesar 56,06 tahun, menunjukkan kluster 1 usia pasien lebih tua dibandingkan kluster 2. Usia muda lebih sering terkena gagal jantung akibat gaya hidup, pola hidup, keturunan hingga riwayat penyakit, sehingga pada pembentukan kluster cenderung variabel usia menunjukkan sebagai petunjuk mengenai gaya hidup atau pola hidup pasien gagal jantung. Sebagian besar kluster 1 dan 2 didominasi pasien dengan tidak memiliki penyakit bawaan seperti tekanan darah dan diabtes, sedangkan anemia pada kluster 1 didominasi pasien yang memiliki penyakit bawaan anemia, sedangkan pada kluster 2 didominasi pasien yang tidak memiliki penyakit bawaan anemia. Pada kondisi normal kadar CPK atau *Creatinine Phosphokinase* dalam darah memiliki angka sebesar 20 - 200 mcg/L, pada kluster 1 dominan kondisi CPK pasien mendekati normal dengan rata-rata sebesar 387 mcg/L namun pada kluster 2 cenderung di dominasi pasien dengan kadar CPK yang cenderung menjauhi kondisi normal dengan rata-rata sebesar 833 mcg/L. *Ejection fraction* memiliki batas normal sekitar 50-75 persen untuk orang dewasa, semakin kebawah maka kondisi kemampuan jantung memompa darah semakin tidak efektif. Pada kluster 1 menunjukkan secara dominan memiliki nilai *ejection fraction* lebih besar dari pada kluster 2 yaitu sebesar 38,08 persen dan 37,36 persen, sehingga keduanya menunjukkan bahwa kondisi jantung yang tidak normal dalam memompa darah ke seluruh tubuh namun kondisi kluster 1 cenderung lebih baik dibandingkan kluster 2.

Platelets merupakan kadar trombosit dalam darah, dalam keadaan normal manusia memiliki kadar trombosit sebesar 150.000-400.000 kiloplatelets/mL. kadar trombosit tinggi pada kondisi jantung menyebabkan beragam hal seperti penggumpalan darah yang dapat menyebabkan pecahnya pembuluh darah. Pada kluster 1 memiliki jumlah trombosit lebih kecil dibandingkan kluster 2 yaitu sebesar 252.448 kiloplatelets/mL dan 276.792 kiloplatelets/mL, hal tersebut menyebabkan kluster 1 memiliki kondisi yang cukup baik dibandingkan dengan kluster 2. Kadar kreatin atau *serum creatine* dalam tubuh manusia secara normal pada laki-laki sebesar 0,6 – 1,2 mg/dL dan pada perempuan 0,5 – 1,1 mg/dL. Kadar keratin yang tinggi biasanya terjadi pada orang yang memiliki gagal ginjal. kluster 1 memiliki kadar kratin yang lebih tinggi dari pada kluster 2 yaitu 1,612 mg/dL dan 1,1255 mg/dL sehinnga kondisi kluster 2 labih baik dibandingkan kluster 1. Sedangkan kadar natrium dalam darah atau serum *sodium* dalam kondisi normal memiliki kadar *natrium* sebesar 135-145 mEq/L, sehingga tingginya kadar *natrium* berpotensi untuk seseorang mudah mengalami *hiponatrium*. kluster 1 memiliki rata-rata kadar *natrium* lebih kecil dibandingkan kluster 2 yaitu sebesar 136,01 mEq/L dan 137,39 mEq/L, hal tersebut menyebabkan kluster 1 memiliki kondisi kadar *natrium* lebih baik dari pada kluster 2. Waktu perawatan pasien yang semakin lama menunjukkan bahwa semakin kompleks kondisi suatu pasien gagal jantung tersebut. Tergambar pada kluster 1 yang memiliki rata-rata waktu rawat lebih singkat dibandingkan rata-rata waktu rawat pasien pada kluster 2. Pada kebiasan merokok dimiliki sebagian besar pada kluster 2 dibandingkan kluster 1, hal tersebut menunjukkan bahwa kebiasaan yang bisa mengganggu kondisi jantung terdapat pada kluster 2 dibandingkan kluster 1. Berdasarkan keterangan diatas disimpulkan bahwa secara umum kluster 1 merupakan karakteristik pasien dengan resiko rendah sedangkan kluster 2 merupakan karakteristik pasien dengan resiko tinggi. Secara keseluruhan dapat dilakukan segmentasi pada pasien dengan variabel penciri mortalitas pada pasien gagal jantung sebagai berikut

Tabel 3. Segmentasi variabel penciri mortalitas pasien gagal jantung

Kluster	Karakteristik Pasien Gagal Jantung
1	Rata-rata usia pasien penyakit jantung pada rumah sakit berusia 64 sampai 65 tahun Sebagian besar pasien penyakit jantung memiliki penyakit bawaan anemia Sebagian besar pasien penyakti jantung tidak memiliki tekanan darah Rata-rata kadar kreatin pasien penyakti jantung dalam darah sebesar 387 mcg/L Sebagian pasien penyakit jantung tidak memiliki penyakit bawaan diabetes Rata-rata persentase darah yang keluar dari jantung pada saat kontraksi sebesar 38-39 persen Sebagian besar pasien penyakit jantung memiliki jenis kelamin laki-laki Rata-rata kadar trombosit darah pasien penyakit jantung sebesar 252448 kiloplatelets/mL Rata-rata level kreatin dalam darah pasien penyakit jantung sebesar 1,612 mg/dL Rata-rata level sodium dalam darah pasien panyakit jantung sebesar 136,63 mEq/L Sebagian besar pasien penyakit jantung tidak memiliki kebiasaan merokok Rata-rata lama waktu tindakan pasien penyakit jantung selama 130-131 hari
2	Rata-rata usia pasien penyakit jantung pada rumah sakit berusia 56 sampai 57 tahun Sebagian besar pasien penyakit jantung tidak memiliki penyakit bawaan anemia Sebagian besar pasien penyakti jantung tidak memiliki tekanan darah Rata-rata kadar kreatin pasien penyakti jantung dalam darah sebesar 822 mcg/L Sebagian pasien penyakit jantung tidak memiliki penyakit bawaan diabetes Rata-rata persentase darah yang keluar dari jantung pada saat kontraksi sebesar 37-38 persen Sebagian besar pasien penyakit jantung memiliki jenis kelamin laki-laki Rata-rata kadar trombosit darah pasien penyakit jantung sebesar 276.792 kiloplatelets/mL Rata-rata level kreatin dalam darah pasien penyakit jantung sebesar 1,1255 mg/dL Rata-rata level sodium dalam darah pasien panyakit jantung sebesar 137,39 mEq/L Sebagian besar pasien penyakit jantung memiliki kebiasaan merokok Rata-rata lama waktu tindakan pasien penyakit jantung selama 163-164 hari

V. KESIMPULAN DAN SARAN

Hasil penggerombolan observasi dengan menggunakan algorima *k-prototypes* dari beberapa percobaan menunjukkan jumlah kluster yang terbentuk yaitu sebanyak 2 kluster. Penentuan kluster yang optimum yaitu dengan menggunakan koefisien *silhouette* sebesar 0,5777 yang digunakan sebagai bahan evaluasi keragaman didalam kluster. Pemilihan kluster optimum didasari nilai koefisien *silhouette* paling besar diatanra koefisien *silhouette* yang dihasilkan pada jumlah kluster yang lain. Hasil penelitian dilakukan segementasi pasien dengan variabel penciri mortalitas pasien gagal jantung yang menunjukan bahwa kluster 1 merupakan gerombol pasien yang memiliki resiko rendah terhadap peluang mortalitas akibat gagal jantung dan kluster 2 merupakan gerombol pasien dengan karaktistik pasien gagal jantung dengan resiko yang tinggi terhadap peluang mortalitas akibat gagal jantung. Segementasi tersebut didasari dari nilai rata-rata setiap variabel penciri dari faktor mortalitas gagal jantung pada setiap kluster yang dibandingkan dengan kondisi normal melalui variabel *serum creatine*, *ejection fraction*, usia, *serum sodium*, tekanan darah, anemia, *creatinine phosphokinase*, *plateles*, merokok, jenis kelamin dan diabetes.

Pada penelitian ini terdapat saran mengenai adanya perbedaan karakteristik antara kedua kluster sehingga perlunya analasis lebih dalam mengenai tindakan medis yang merujuk kepada karakteristik kedua kluster tersebut sehingga dapat menurunkan tingat mortalitas akibat adanya gagal jantung. Penelitian ini hanya menggunakan analaisis kluster sehingga perlunya pengukuran lain dengan menggunakan variabel mortalitas untuk menunjukkan *variabael importance* pada karakteristik mortalitas pasien gagal jantung, sehingga variabel tersebut menjadi penentu dalam mengidentifikasi pasien yang beresiko untuk berpeluang terjadinya gagal jantung. Dalam penelitian ini metode yang digunakan merupakan metode *unsupervised*, sehingga perlu perbandingan dengan metode yang bersifat *supervised*.

REFERENSI

[1] J. Barallobre-Barreiro, Y.-L. Chung, and M. Mayr, "Proteomics and metabolomics for mechanistic insights and biomarker discovery in cardiovascular disease," *Rev. Española Cardiol. (English Ed.*, vol. 66, no. 8, pp. 657–661, 2013.

[2] World Health Organization, "WHO." https://www.who.int/cardiovascular_diseases/world-heart-day/en/ (accessed Jan. 07, 2020).

[3] A. B. I. NATIONAL HEART, LUNG, "No Title." https://www.nhlbi.nih.gov/health-topics/heart-failure (accessed Jan. 08, 2020).

[4] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS One*, vol. 12, no. 7, p. e0181001, 2017.

[5] F. Meng *et al.*, "Machine learning for prediction of sudden cardiac death in heart failure patients with low left ventricular ejection fraction: study protocol for a retroprospective multicentre registry in China," *BMJ Open*, vol. 9, no. 5, p. e023724,

2019.

[6] T. A. Buchan *et al.*, "Physician prediction versus model predicted prognosis in ambulatory patients with heart failure," *J. Hear. Lung Transplant.*, vol. 38, no. 4, p. S381, 2019.

[7] B. Chapman, A. D. DeVore, R. J. Mentz, and M. Metra, "Clinical profiles in acute heart failure: an urgent need for a new approach," *ESC Hear. Fail.*, vol. 6, no. 3, pp. 464–474, 2019.

[8] L. Chiodo, M. Casula, E. Tragni, A. Baragetti, D. Norata, and A. L. Catapano, "Profilo cardiometabolico in una coorte lombarda: lo studio PLIC. Cardio-metabolic profile in a cohort from Lombardy region: the PLIC study," *G. Ital. di Farm. e Farm.*, vol. 9, no. 2, pp. 35–53, 2017.

[9] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 16, 2020.

[10] Y. Al-Kofahi, W. Lassoued, W. Lee, and B. Roysam, "Improved automatic detection and segmentation of cell nuclei in histopathology images," *IEEE Trans. Biomed. Eng.*, vol. 57, no. 4, pp. 841–852, 2009.

[11] P. Arora and S. Varshney, "Analysis of k-means and k-medoids algorithm for big data," *Procedia Comput. Sci.*, vol. 78, pp. 507–512, 2016.

[12] T. S. Madhulatha, "Comparison between k-means and k-medoids clustering algorithms," in *International Conference on Advances in Computing and Information Technology*, 2011, pp. 472–481.

[13] R. Madhuri, M. R. Murty, J. V. R. Murthy, P. P. Reddy, and S. C. Satapathy, "Cluster analysis on different data sets using K-modes and K-prototype algorithms," in *ICT and Critical Infrastructure: Proceedings of the 48th Annual Convention of Computer Society of India-Vol II*, 2014, pp. 137–144.

[14] J. Supranto, "Statistik Deskriptif." Jakarta: Airlangga, 1988.

[15] A. A. Mattjik, I. Sumertajaya, G. N. A. Wibawa, and A. F. Hadi, "Sidik peubah ganda dengan menggunakan SAS." 2011.

[16] S. Sharma and S. Sharma, "Applied multivariate techniques," 1996.

[17] S. G. Rao and A. Govardhan, "Performance validation of the modified k-means clustering algorithm clusters data," *Int. J. Sci. Eng. Res.*, vol. 6, no. 10, pp. 726–730, 2015.

[18] Z. Ansari, M. F. Azeem, W. Ahmed, and A. V. Babu, "Quantitative evaluation of performance and validity indices for clustering the web navigational sessions," *arXiv Prepr. arXiv1507.03340*, 2015.

[19] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.

[20] N. J. Salkind, *Encyclopedia of measurement and statistics*. SAGE publications, 2006.

[21] Z. Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values," *Data Min. Knowl. Discov.*, vol. 2, no. 3, pp. 283–304, 1998.

[22] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, vol. 5, no. 8. Prentice hall Upper Saddle River, NJ, 2002.

[23] G. Gan, C. Ma, and W. Jianhong, "Center-based clustering algorithms," *Data Clust. Theory, Algorithms Appl.*, 2007.

[24] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PloS one*, 2017. https://plos.figshare.com/articles/dataset/Survival_analysis_of_heart_failure_patients_A_case_study/5227684 (accessed Jan. 08, 2020).