BERKALA SAINSTEK 2022 10(3): 124-132
doi: 10.19184/bst.v10i3.22259
https://jurnal.unej.ac.id/index.php/BST/article/view/22259

RESEARCH ARTICLE

Classification of Cardiovascular Disease Gene Data Using Discriminant Analysis and Support Vector Machine (SVM)

(Pengklasifikasian Data Gen Penyakit Kardiovaskuler Menggunakan Analisis Diskriminan dan Support Vector Machine (SVM))

Rizky Prayogo, Dian Anggraeni*, Alfian Futuhul Hadi
Jurusan Matematika, Fakultas MIPA, Universitas Jember,
Jl. Kalimantan No.37, Jember, Indonesia

ABSTRACT

Cardiovascular disease is a disease caused by impaired function of the heart and blood vessels. This disease is caused by many factors, one of which is genetics, while the causes are age, gender, and family history. In this study, classification of 62 individuals with normal response and cardiovascular disease was carried out. Discriminant Analysis (AD) is a method that classifies data into two or more groups based on several variables where data that has been entered into one group will not be included in another group. The Support Vector Machine (SVM) performs classification by building an N-dimensional hyperplane that optimally separates data into two categories in the input space. Furthermore, AD and SVM will be compared to get which method has the best accuracy, after that it will be added to clustering using k-means and k-means kernels to improve the accuracy of each method. The results of this study are AD and SVM have accuracy values of 83.33% and 91.66%, for AD and SVM which are subjected to k-means have accuracy values of 91.66 % and 91.66 %, and for AD and SVM subjected to k-means kernel has an accuracy value of 100 % and 100 %.

Penyakit kardiovaskuler adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah. Penyakit ini disebabkan oleh banyak faktor salah satunya genetika, adapun penyebabnya adalah umur, jenis kelamin, dan riwayat keluarga. Dalam penelitian ini dilakukan klasifikasi terhadap 62 individu dengan respon normal dan terjangkit kardiovaskuler. Analisis diskriminan (AD) metode yang mengklasifikasikan data kedalam dua atau beberapa kelompok berdasarkan beberapa variabel dimana data yang sudah masuk kedalam satu kelompok tidak akan masuk kedalam kelompok yang lain. *Support Vector Machine* (SVM) melakukan klasifikasi dengan membangun sebuah hyperplane N-dimensi yang optimal memisahkan data menjadi dua kategori pada input space. Selanjutnya akan dibandingkan AD dan SVM untuk didapatkan mana metode yang mempunyai akurasi terbaik, setelah itu akan ditambahi diklastering menggunakan k-means dan kerel k-means untuk memperbaiki hasil akurasi dari masing-masing metode. Didapatkan hasil dari penelitian ini adalah AD dan SVM mempunyai nilai akurasi sebesar 83,33% dan 91,66%, untuk AD dan SVM yang dikenai k-means mempunyai nilai akurasi sebesar 91,66 % dan 91,66 %, dan untuk AD dan SVM yang dikenai kernel k-means mempunyai nilai akurasi sebesar 100 % dan 100 %.

Keywords: Kardiovaskuler, Analisis diskriminan, *Support Vector Machine* (SVM), K-Means, Kernel K-Means.

*) Corresponding author:
Dian Anggraeni
E-mail: dian_a.fmipa@unej.ac.id

PENDAHULUAN

Penyakit kardiovaskuler adalah penyakit yang disebabkan gangguan fungsi jantung dan pembuluh darah. *World Health Organization* (WHO) melaporkan penyakit kardiovaskuler adalah penyakit kematian nomer satu di dunia. Ada banyak macam penyakit kardiovaskuler, tetapi yang paling umum dan paling terkenal adalah penyakit jantung koroner, stroke dan gagal jantung [1]. Penyakit kardiovaskuler adalah penyakit yang disebabkan oleh penyumbatan aliran darah yang membawa sari-sari makanan dan oksigen menuju otot jantung [2]. Kebanyakan faktor-faktor yang mempengaruhi penyakit kardiovaskular yaitu merokok, kurang aktifitas gerak, obesitas dan tekanan darah tinggi atau hipertensi [3]. Selain faktor kebiasaan hidup yang bisa mempengaruhi penyakit kardiovaskular ada juga faktor keturunan seperti halnya jenis kelamin, umur dan riwayat keluarga (genetika) [1].

Penggunaan sistem klasifikasi dalam mengdiagnosis penyakit kardiovaskular meningkat,

Dalam bidang ilmu statistika kita bisa mengklasifikasikan genetika penderita penyakit kardiovaskuler untuk mengetahui apakah orang tersebut dengan gen yang terkandung dalam tubuhnya mempunyai resiko penyakit kardiovaskuler. pengklasifikasian penyakit kardiovaskuler dalam disiplin ilmu satatistika antara lain Analisis Diskriminan (AD) dan *Support Vector Machine (SVM).*

Analisis diskriminan merupakan metode statistika untuk mengelompokkan atau mengklasifikasikan sejumlah obyek ke dalam dua atau beberapa kelompok berdasarkan beberapa variabel. Analisis diskriminan mengasumsikan data berdistribusi normal multivariat dan matriks kovarian yang homogen. Sedangkan SVM menggunakan cara pengklasifikasian dengan menggunakan Hyperplane N-dimensi yang bisa memisahkan data menjadi dua kategori yang berbeda, secara sederhana konsep SVM adalah sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas. *Pattern* yang merupakan anggota buah kelas : +1 dan -1 dan berbagi *alternative* garis pemisah (*Discrimination Boundaries*). *Margin* adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing kelas. Pattern yang paling dekat ini disebut sebagai *support vector* [4].

Penelitian dilakukan Suwardika [5] tentang metode SVM dibandingkan dengan Analisis Diskriminan pada data *Echocardiogram* dengan akurasi 88 % untuk SVM dan 96% untuk analisis diskriminan. Penelitian yang dilakukan Huang dkk. membandingkin model *logistics regression*, *multivariate discriminan analysis*, *neural network* dan SVM, pada analisis peringkat kredit dari hasil penelitian tersebut model yang mempunyai *accuracy* ketepatan lebih tinggi adalah dengan model NN dan SVM yaitu 80 % [6].

Selain itu penelitian ini mencari metode mana dan akurat untuk klasifikasi data yang berukuran besar. Secara garis besar penilitian ini akan dilakukan dengan menggunakan 2 metode berbeda yakni analisis diskriminan dan SVM. Serta data akan dibagi menjadi dua yakni untuk data *training* dan data *testing*, dan untuk mendapatkan hasil terbaik akan menggunakan K-Means dan Kernel K-Means.

METODE PENELITIAN

Kardiovaskuler

Kardiovaskuler adalah penyakit yang pada umumnya menyerang jantung dan sesekali menyerang pembuluh darah, kardiovaskular banyak macamnya yang sering kita dengar adalah jantung koroner, stroke dan gagal jantung. Faktor yang sering dijumpai ada 2 (dua) faktor yang mempengruhi penyakit kardiovaskuler yang pertama karena keturunan/bawaan (*congenital heart diseases*) dan yang kedua karena didapatkan sebab pola hidup (*acquired heart diseases*) menurut *survey* terbaru penyakit kardiovaskuler adalah penyakit yang paling banyak menyebabkan kematian diseluruh dunia. . Pada tahun 2005 ada sekitar 17,5 juga orang mengidap penyakit ini dan 30 % meninggal pada tahun yang sama [7].

Faktor-faktor yang mempengaruhi penyakit kardiovaskuler secara garis besar dibedakan menjadi 2 (dua) yang pertama dikenal dengan faktor resiko yang kedua dikenal dengan faktor prognosis. Faktor yang dapat memperberat atau mempengaruhi perjalanan penyakit kardiovaskular yang telah ada disebut sebagai faktor prognosis, istilah faktor risiko digunakan untuk menggambarkan faktor-faktor yang dapat mempermudah serta faktor-faktor yang dapat memperberat atau mempengaruhi perjalanan timbulnya penyakit kardiovaskular, namun secara prakteknya kedua faktor ini sering disandingkan karena faktor resiko sudah mencakup faktor prognosis.

Faktor risiko kardiovaskuler adalah faktor-faktor yang memudahkan timbul dan memberat dari penyakit ini sendiri. Secara umum, faktor risiko ini dibedakan atas faktor risiko yang tidak dapat diubah (seperti umur, jenis kelamin, ras dan riwayat keluarga menderita kelainan kardiovaskuler) dan faktor risiko yang dapat diubah (seperti kebiasaan merokok, diet, aktivitas fisik yang kurang, kegemukan, tekanan darah tinggi, penyakit diabetes dan sebagainya) [8].

Klasifikasi

Klasifikasi adalah aspek yang harus ada pada *data mining*. Teknik klasifikasi telah banyak digunakan dalam berbagai permasalahan dalam sebuah penelitian.Klasifikasi merupakan suatu metode pengelompokan data yang akan mempelajari *data training* dengan menggunakan algoritma pengklasifikasian, adapun beberapa algoritma klasifikasi antara lain *Bayesian classification, K-Nearst Neighbor, Decision Tree Induction, Case-Based Reasoning, Genetic Algorithms, Support Vector Mechine* (SVM) dan *Linear Discriminant Analysis* [9].

Analisis Diskriminan (AD)

Analisis diskriminan (AD) merupakan metode statistika untuk mengelompokkan atau mengklasifikasikan sejumlah obyek ke dalam dua atau beberapa kelompok berdasarkan beberapa variabel. Setiap obyek yang diklasifikasikan akan menjadi anggota dari salah satu kelompok dan tidak ada obyek yang menjadi anggota lebih dari satu kelompok [10]. Analisis diskriminan mengasumsikan data berdistribusi normal multivariat dan matriks kovarian yang homogen. Uji asumsi distribusi normal multivariat dapat dilakukan Chi-Square plot sedangkan asumsi homogenitas matriks kovarian diuji dengan uji Box's M [11]. Terdapat asumsi yang harus dipenuhi dalam analisis diskriminan yaitu bahwa data harus memenuhi asumsi berdistribusi normal multivariate dan matrik varian kovarian antara kategori terjangkit dan kategori normal homogen. Dengan meggunakan hipotesis.

Hipotesis:

H_0 : Data berdistribusi normal multivariate.

H_1 : Data tidak berdistrbusi normal multivariate.

$\alpha = 0.05$

Uji homogenan matrik varian dan kovarian antara kategori 0 dan 1 dengan menggunkan Box-M dengan hepotesis sebagai berikut.

Hipotesis:

$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_k$ dimana k=observasi sampai ke-k

$H_1 : \Sigma_i \neq \Sigma_j$ untuk $i \neq j$ dimana $i, j = 1,2, \ldots, n$

$\alpha = 0.05$

***Support Vector Machine* (SVM)**

Suatu *Support Vector Machine* (SVM) melakukan klasifikasi dengan membangun sebuah hyperplane N-dimensi yang optimal memisahkan data menjadi dua kategori. Konsep SVM dapat dijelaskan secara sederhana sebagai usaha mencari *hyperplane* terbaik yang berfungsi sebagai pemisah dua buah kelas pada *input space*. *Pattern* yang merupakan anggota dari dua buah kelas : +1 dan -1 dan berbagi alternative garis pemisah (*discrimination boundaries*). Margin adalah jarak antara *hyperplane* tersebut dengan *pattern* terdekat dari masing-masing kelas. *Pattern* yang paling dekat ini disebut sebagai *support vector* [4].

Linear Support Vector Machine

Setiap data yang ada pada SVM dinotasikan sebagai $x_i \varepsilon R^p, i = 1,2, \ldots, n$ dimana n adalah banyaknya data. Kelas positif dinotasikan dengan +1,dan kelas negatif dinotasikan dengan -1. Dengan demikian ,tiap data dan dari masing-masing kelas dinotasikan dengan : $y_i \varepsilon \{0,1\}$. Diasumsikan bahwa kedua kelas tersebut dapat dipisahkan secara baik oleh masing-masing *hyperplane* di *D-dimensional feature space*. *Hyperplane* tersebut didefinisikan sebaga berkut:

$$\vec{w}.\vec{x_i} + b = 0 \quad (1)$$

Data $\vec{x_i}$ yang tergolong ke dalam kelas -1 (sampel negatif) adalah yang memenuhi pertidaksamaan berikut:

$$\vec{w}.\vec{x_i} + b \geq -1 \quad (2)$$

Adapun data $\vec{x_i}$ yang tergolong ke dalam kelas 1 (sampel positif) adalah yang memenuhi pertidaksamaan berikut :

$$\vec{w}.\vec{x_i} + b \geq 1 \quad (3)$$

Margin terbesar bisa dihitung dengan memaksimalkan jarak antara hyperplane dan pattern terdekat. Jarak ini dirumuskan sebagai $1/||\vec{w}||$ ($||\vec{w}||$ adalah norm dari vector w). Selanjutnya,masalah ini diformulasikan ke dalam *Quadratic Programming*(QP) problem, dengan meminimalkan invers persamaan berikut:

$$||\vec{w}||^2 = w^T \quad (4)$$

$$y_i(w.x_i + b) - 1 \geq 0, \forall_i \quad (5)$$

Optimasi ini dapat diselesaikan dengan menggunakan terknik komputasi salah satunya lagrange multipliers :

$$L(\vec{w}, b, a) = \frac{1}{2}||\vec{w}||^2 - \sum_{i=2}^{n} \propto_i y_i(\vec{w}.\vec{x}_i + b - 1) \quad (6)$$

α_i adalah Lagrange multiplier yang berkorespondensi dengan x_i. Nilai α_i adalah nol atau positif. Untuk menyelesaikan masalah tersebut. Pertama-tama meminimakan L terhadap w,dan memaksimalkan L terhadap α_i. Dengan memodifikasi persamaan (9), memaksimalkan masalah diatas dapat direpresentasikan dalam α_i.

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x}_i \vec{x}_j \quad (7)$$

$$\alpha_i \geq 0, \sum_{i=1}^{n} \alpha_i y_i = 0 \quad (8)$$

Solusi dari permasalahan di atas menghasilkan banyak nilai α_i nol. Data yang berkorespondensi dengan α_i yang tak nol,merupakan *support vectors*,yaitu

data yang memiliki jarak terdekat dengan *hyperplane* [13].

Non Linear Support Vector Mechine

SVM merupakan salah satu varian dari *linear machine* sehingga banyak dapat dipakai untuk menyelesaikan masalah yang sifatnya *linear separable*. Untuk dapat dipakai dalam kasus *non-linear*,pertama-tama data yang berada pada ruang vector awal ($\{x_i \epsilon \Re^D\}$) berdimensi D,harus dipetakan ke ruang vector baru yang berdimensi lebih tinggi ($\{x_i' \epsilon \Re^Q\}$). Fungsi pemetaan tersebut dinotasikan sebagai $\Phi(x)$. Pemetaan ini bertujuan untuk merepresentasikan data kedalam format yang *linaer separable* pada ruang vector yang baru.

$$\Phi : \Re^d \rightarrow \Re^q, d < q \quad (9)\ (15)$$

Proses optimimisasi pada fase ini memerlukan perhitungan *dot product* dua buah variabel pada ruang vektor baru. Dot product kedua buah vektor (x_i) dan (x_j) dinotasikan sebagai $\Phi(x_i).\Phi(x_j)$. nilai *dot product* kedua buah vektor ini dapat dihitung secara tak langsung,yaitu tanpa mengetahui fungsi transformasi Φ. komponen kedua buah vetor tersbut diruang vektor asal sebagai berikut.

$$K(x_i, x_j) = \Phi(x_i).\Phi(x_j) \quad (10)$$

Berbagai jenis fungsi dapat dipakai sebagai kernel k,sebagaimana tercantum pada Tabel 1 [12].

Tabel 1. Fungsi kernel dalam SVM

Nama kernel	Definisi
Linear	$K(\overrightarrow{x_i}.\overrightarrow{x_j}) = (\overrightarrow{x_i}.\overrightarrow{x_j})$
Polynomial	$K(\overrightarrow{x_i}.\overrightarrow{x_j}) = (\overrightarrow{x_i}.\overrightarrow{x_j} + 1)^p$
Gaussian RBF	$K(\overrightarrow{x_i}.\overrightarrow{x_j}) = \exp(-\frac{\|\| \overrightarrow{x_i}.\overrightarrow{x_j}\|\|^2}{2\sigma^2})$
Sigmoid	$K(\overrightarrow{x_i}.\overrightarrow{x_j}) = \tanh(\alpha x_i.x_j + \beta)$

Tabel 2. *Confusion matrix*

Confusion Matrix		Nilai Prediksi	
		TRUE	FALSE
Nilai Riil	TRUE	TP (True positive)	FP (False positive)
	FALSE	FN (False negative)	TN (True negative)

Klastering K-Means dan Kernel K-Means

Klastering secara umum digunakan untuk melihat kesamaan antar objek dan mengelompokkannya ke dalam beberapa bagian. Peran klastering dalam ilmu pengetahuan menjadi semakin pesat berkembang seiring dengan kebutuhannya yang semakin meningkat, K-means dan Kernel K-Means merupakan salah satu algoritma populer dalam pembentukan klastering. Selain karena kemudahan, algoritma ini juga mampu memberikan hasil yang efektif.

K-Means adalah algoritma klastering untuk menemukan kelompok dari objek yang *non overlapping* [14]. K-Means juga dianggap sebagai algoritma yang efektif untuk mengelompokkan suatu data [15]. K-means adalah algoritma klastering dalam bidang data mining yang hanya mampu menangkap fitur kelas yang linear . Hal ini digunakan untuk klastering analisis, dan memiliki efisiensi tinggi pada partisi data terutama dalam dataset besar [16].

Kernel k-means bekerja dengan mengubah data dari *initial space* ke dalam *featured space* dan k-means dijalankan menggunakan data *featured space* tersebut. Kernel K-Means algoritma klastering yang mampu menangkap fitur pembagian kelas yang *non linear* dari data yang diinisialisasi, tidak hanya itu Kernel K-means juga mampu mengekspresikan hasil kompleksnya pemetaan *nonlinear* [17].

Kurva *Receiver Operating Characteristic* (ROC)

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horizontal dan *true positive* sebagai garis vertikal [18]. *Confusion matrix* adalah alat ukur berbentuk matriks yang digunakan untuk mendapatkan jumlah ketepatan klasifikasi terhadap kelas dengan algoritma yang dipakai. Berikut disajikan bentuk confusion matrix pada Tabel 2.

Pada Tabel 2 nilai TP (*True Positive*) dan TN (*True Negative*) menunjukan tingkat ketepatan klasifikasi. Berikut formulasi untuk menghitung akurasi, *specitifity*, *sensitivity* pada pembentukan model klasifikasi ditunjukkan pada persamaan (11), persamaan (12), persamaan (13).

$$Akurasi = \frac{TP + TN}{TP + FP + TN + FN} \quad (11)$$

$$Specitifity = \frac{TN}{FP + TN} \quad (12)$$

$$Sensitivity = \frac{TP}{FN + TP} \quad (13)$$

Menunjukkan proporsi klasifikasi kelas *negative* (*false positive rate*), dalam kasus ini termasuk dalam kategori normal. *Sensitivity* menunjukkan klasifikasi kelas *positive* (*true positive rate*). ROC memiliki tingkat nilai diagnose/klasifikasi yaitu sebagai berikut: [19]

Tabel 3. Tingkat akurasi kurva ROC

Akurasi	Tingkat klasifikasi
0,90-1,00	*excellent classification*
0,80-0,90	*good classification*
0,70-0,80	*fair classification*
0,60-0,70	*poor classification*
0,50-0,60	*Failure*

Model Development

Data yang digunakan oleh penulis terdiri dari Variabel dependen dan independen, variabel dependen yaitu 62 individu yang diklasifikasi kedalam kelas ya (terjangkit kardiovaskuler) didefinisikan dengan *positive*, dan tidak (normal) didefinisikan dengan *negative*. Sedangkan variabel independen berupa 2000 variabel gen penyakit kardiovaskuler pada setiap individu. Tahapan selanjutnya adalah menganalisis, proses analisis sebagai berikut.

1. Mengambil data awal pengamatan gen penyakit kardiovaskuler
2. Data dibagi menjadi 2 dengan perbandingan 80 % untuk *training* dan 20 % untuk *testing* dengan proporsi sama setiap kelasnya.
3. Data diklasifikasi menggunakan AD dan SVM tanpa klastering setelah itu dicari kurva ROC dan AUC
4. Data diklastering menggunakan K-Means setelah itu data klasifikasi dengan AD dan SVM setelah itu dicari kurva ROC dan AUC
5. Data diklastering menggunakan Kernel K-Means setelah itu data klasifikasi dengan AD dan SVM setelah itu dicari kurva ROC dan AUC
6. Didapatkan hasil akurasi metode terbaik.

HASIL DAN PEMBAHASAN

Data microarray yang digunakan yaitu dalam bentuk 62 individu dengan masing-masing individu jumlahnya genetikanya sebanyak 2000 variabel. Sebanyak 22 individu tergolong ke dalam kelas negative (normal), dan 40 individu termasuk kelas positive (terjangkit kardiovaskuler), ekspresi gen yang tersedia akan dibagi menjadi 80 % untuk training dan 20 % untuk testing. Setelah itu akan dibandingkan AD, AD K-Means, AD Kernel K-Means, SVM, SVM K-Means, dan SVM Kernel K-Means.

Klasifikasi AD dan SVM Tanpa Klastering

Klasifikasi AD tanpa kastering akan membagi data training dan testing dengan jumlah data training sebanyak 50 data dan jumlah data testing sebanyak 12 data. dalam tahap AD dan SVM, Jumlah pembagian data dapat dilihat pada tabel 3.

Untuk hasil dari AD dan SVM data testing bisa dilihat pada Tabel 4.

Tabel 4. Confsusion Matrix Pengujian Data Testing AD dan SVM Tanpa Klastering

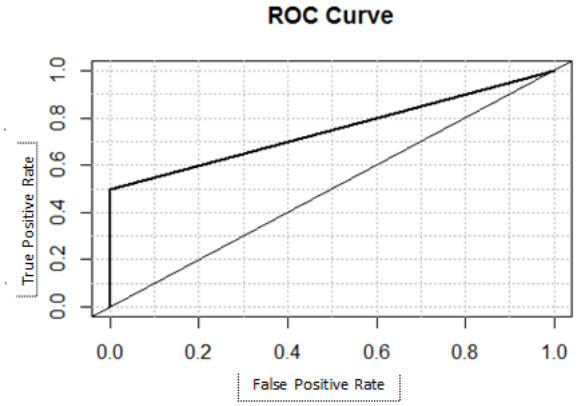
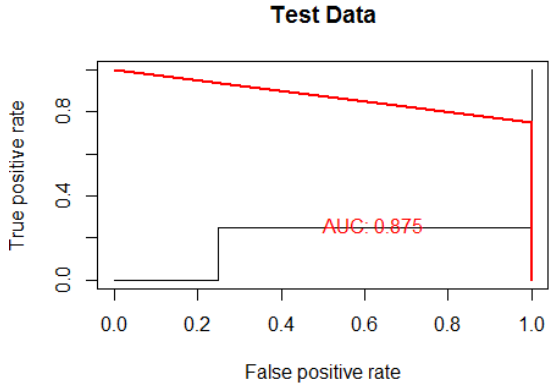
Confusion Matrix	AD	SVM	AD	SVM
	Normal		Kardiovaskuler	
Normal	8	8	0	0
Kardiovaskuler	2	1	2	3

Jumlah data testing yang akan diuji pada AD adalah sebanyak 12 data dengan perbandingan 8 data normal dan 2 data terjangkit kardiovaskuler, terjadi kesalahan prediksi data normal sebanyak 2 orang yang masuk kedalam kelas terjangkit kardiovaskuler. Didapatkan hasil akurasi dari data testing yaitu sebesar 83.33 %, Hasil akurasi prediksi proses testing menghasilkan tingkat akurasi 83,3 %, performa specitifity sebesar 100 %, sedangkan performa sensitivity sebesar 80 %.

Untuk hasil dari SVM Jumlah data testing yang akan diuji adalah sebanyak 12 data dengan perbandingan 8 data normal dan 3 terjangkit kardiovaskuler, pada kelas normal terjadi kesalahan prediksi data yang terbaca dalam kelas terjangkit

kardiovaskuler sebanyak 1 orang didapatkan hasil akurasi prediksi data testing sebesar 91,66 %. Hasil akurasi prediksi proses testing menghasilkan tigkat akurasi 91,6 %, performa specitifity sebesar 100 %, sedangkan performa sensitivity sebesar 88,8 %. Untuk kurva ROC seperti Tabel 5.

Tabel 5. Perjumpaan MEP di Resort Kucur TNAP

Kurva ROC dan AUC	Interpretasi
 Gambar 1. Kurva ROC data *testing* AD tanpa klastering	Hasil evaluasi dengan kurva ROC terhadap data *training* pada gambar didapatkan nilai akurasi sebesar 0,83, maka menunjukan bahwa termasuk *good classification*.
 Gambar 2. Kurva ROC data *testing* SVM tanpa klastering	Hasil evaluasi dengan kurva ROC terhadap data *testing* pada gambar didapatkan nilai akurasi sebesar 0,916, dan diketahui nilai AUC sebesar 0,875 maka menunjukan bahwa termasuk *excellent classification*.

Klasifikasi AD dan SVM Menggunakan K-Means

Klasifikasi AD dan SVM menggunakan K-Means dengan jumlah data *training* sebanyak 50 data dan jumlah data *testing* sebanyak 12 data. Kemudian data yang sudah dipartisi dibentuk menjadi data *frame* untuk digunakan dalam tahap AD dan SVM menggunakan K-Means. Jumlah pembagian data dapat dilihat pada tabel 6 berikut.

Untuk hasil dari AD dan SVM dengan K-Means data *testing* bisa dilihat pada tabel 6. Berikut.

Tabel 6. *Confsusion Matrix* Pengujian Data *Testing* AD dan SVM dengan K-Means

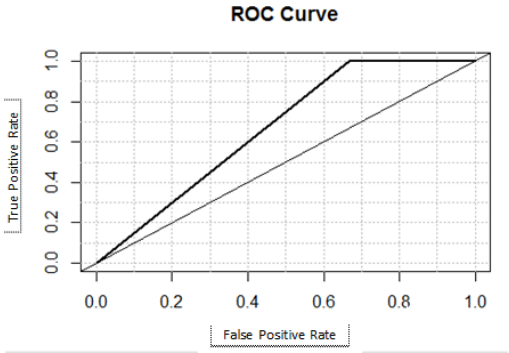
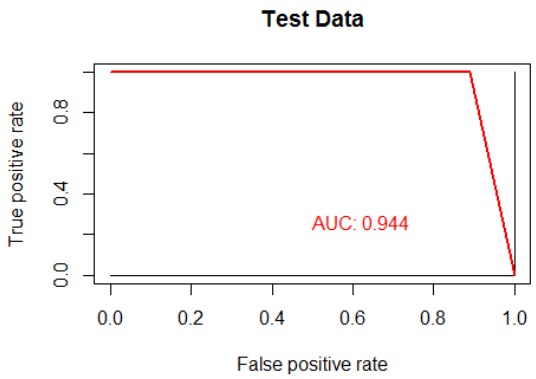
Confusion Matrix	AD	SVM	AD	SVM
	Normal		Kardiovaskuler	
Normal	8	8	1	1
Kardiovaskuler	0	0	3	3

Jumlah data *testing* AD K-Means yang akan diuji adalah sebanyak 12 data dengan perbandingan 8 data normal dan 3 terjangkit kardiovaskuler, pada kelas terjangkit kardiovaskuler terjadi kesalahan prediksi data yang terbaca dalam kelas terjangkit normal sebanyak 1 orang. Didapatka hasil akurasi data *testing* sebesar 91,66 %. Hasil akurasi prediksi proses *testing* menghasilkan tigkat akurasi 91,6 %, performa *specitifity* sebesar 75 %, sedangkan performa *sensitivity* sebesar 100 %.

Jumlah data *testing* SVM K-Means yang akan diuji adalah sebanyak 12 data dengan perbandingan 8 data normal dan 3 terjangkit kardiovaskuler, pada kelas terjangkit kardiovaskuler terjadi kesalahan prediksi data yang terbaca dalam kelas normal sebanyak 1 orang didapatkan hasil akurasi prediksi data *testing* sebesar 91,66 %. Hasil akurasi prediksi proses *testing*

menghasilkan tigkat akurasi 91,6 %, performa *specitifity* sebesar 75 %, sedangkan performa *sensitivity* sebesar 100 %. Untuk kurva ROC seperti Tabel 7.

Tabel 7. Kurva ROC dan AUC AD dan SVM dengan K-Means

Kurva ROC dan AUC	Interpretasi
 Gambar 3. Kurva ROC data *testing* AD menggunakan k-means	Hasil evaluasi dengan kurva ROC terhadap data *testing* pada gambar didapatkan nilai akurasi sebesar 0,916, maka menunjukan bahwa termasuk *excellent classification.*
 Gambar 4. Kurva ROC data *testing* SVM menggunakan K-Means	Hasil evaluasi dengan kurva ROC terhadap data *testing* pada gambar didapatkan nilai akurasi sebesar 1 pada program, dan diketahui nilai AUC sebesar 0,944 maka menunjukan bahwa termasuk *excellent classification.*

Klasifikasi AD dan SVM menggunakan Kernel K-Means

Klasifikasi AD dan SVM menggunakan Kernel K-Means dengan menggunakanjumlah data *training* sebanyak 50 data dan jumlah data *testing* sebanyak 12 data. Kemudian data yang sudah dipartisi dibentuk menjadi data *frame* untuk digunakan dalam tahap AD dan SVM menggunakan Kernel K-Means, Jumlah pembagian data dapat dilihat pada Tabel 8.

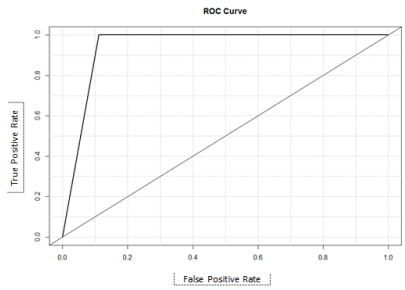
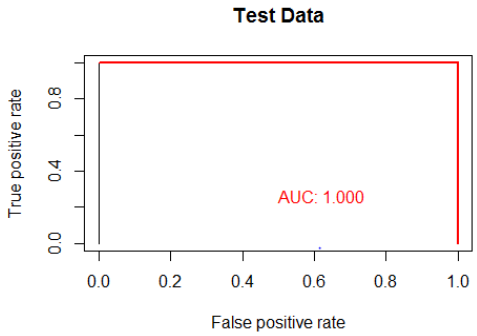
Tabel 8. *Confsusion Matrix* Pengujian Data *Testing* AD dan SVM dengan Kernel K-Means

Confusion Matrix	AD	SVM	AD	SVM
	Normal		Kardiovaskuler	
Normal	3	3	0	0
Kardiovaskuler	0	0	9	9

Jumlah data *testing* AD Kernel K-Means yang akan diuji adalah sebanyak 12 data dengan perbandingan 3 data normal dan 9 terjangkit kardiovaskuler, didapatka hasil akurasi prediksi data *testing* sebesar 100 %. Hasil perhitungan di atas menghasilkan akurasi prediksi proses *testing* menghasilkan tigkat akurasi 100 %, untuk performa *specitifity* dan performa *sensitivity* sama-sama sebesar 100 %. Untuk kurva ROC seperti gambar 6. dibawah.

Jumlah data *testing* SVM Kernel K-Mean yang akan diuji adalah sebanyak 12 data dengan perbandingan 3 data normal dan 9 terjangkit kardiovaskular, didapatka hasil akurasi prediksi data *training* sebesar 100 %. Hasil akurasi prediksi proses *testing* menghasilkan tigkat akurasi 100 %, performa *specitifity* sebesar 100 %, sedangkan performa *sensitivity* sebesar 100 %. Untuk kurva ROC seperti Tabel 9.

Tabel 9. Kurva ROC dan AUC AD dan SVM dengan Kernel K-Means

Kurva ROC dan AUC	Interpretasi
 Gambar 5. Kurva ROC data *testing* AD menggunakan Kernel K-Means	Hasil evaluasi dengan kurva ROC terhadap data *training* pada gambar didapatkan nilai akurasi sebesar 1, maka menunjukan bahwa termasuk *excellent classification.*
 Gambar 6. Kurva ROC data *testing* SVM menggunakan Kernel K-Means	Hasil evaluasi dengan kurva ROC terhadap data *testing* pada gambar didapatkan nilai akurasi sebesar 1, dan diketahui nilai AUC sebesar 1 maka menunjukan bahwa termasuk *excellent classification.*

Hasil Klasifikasi Terbaik

Hasil akurasi terbaik didapatkan dari membandingkan antara AD, AD K-Means, ADK Kernel K-Means, SVM, SVM K-Means, SVM Kernel K-Means. Pencarian hasil klasifikasi terbaik ini dilihat dari tingkat akurasi terbaik. Hasil akurasi terbaik dari masing-masing metode diatas dapat dilihat pada Tabel 10.

Tabel 10. Hasil akurasi terbaik masing-masing metode

Jenis Data	Tanpa klastering		Dengan klastering			
			K Means		Kernel K-Means	
	AD	SVM	AD	SVM	AD	SVM
Training	98%	100%	100%	100%	100%	100%
Testing	83,33%	91,66%	91,66%	91,66%	100%	100%

Dari tabel diatas bisa diketahui bahwa untuk metode terbaik ditunjukan oleh AD Kernel K-Means dan SVM Kernel K-Means dengan tingkat akurasi masing-masing 100 %.

KESIMPULAN

Percobaan klasifikasi data gen penyakit kardiovaskular dilakukan dengan Menggunakan 6 metode yaitu AD, AD K-Means, AD Kernel K-Means, SVM, SVM K-Means, dan SVM Kernel K-Means. Metode terbaik ditunjukan dengan tingkat akurasi tertinggi pada data *training* dan *testing,* untuk hasil dari masing-masing metode adalah sebagai berikut :

1. Hasil dari AD tanpa klastering untuk akurasi dari data *training* 98 % dan *testing* 83,33 %, untuk akurasi AD K-Means adalah untuk data *training* 100% dan *testing* 100 %, sedangkan akurasi dari AD Kernel K-Means data *training* dan *testing* sama-sama mempunyai akurasi 100 %.
2. Hasil dari SVM tanpa klastering untuk akurasi dari data *training* 100 % dan *testing* 91,66 %, untuk akurasi SVM K-Means adalah untuk data *training* 100 % dan *testing* 91,66 %, sedangkan akurasi dari

SVM Kernel K-Means dari dari data *training* dan *testing* sama-sama sebesar 100 %.

3. Hasil perbandingan metode terbaik adalah metode yang menggunakan Kernel K-Means baik pada AD maupun SVM dengan akurasi data *training* dan *testing* sebesar 100 %.

DAFTAR PUSTAKA

[1] Kementerian Kesehatan RI. 2014. Situasi Kesehatan Jantung. Jakarta: Infodatin.

[2] Kurniadi, Helmanu. Stop Gejala Penakit Jantung Koroner. Yogyakarta: Familia, 2013.

[3] World Health Organization, World Health Statistic 2013. Geneva: WHO Press, 2013

[4] N. Cristianini and J. S. Taylor, an Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge Press University, 2000.

[5] G. Suwardika, "Pengklasifikasian pada data echocardiogram dengan menggunakan support vector machine dan analisis diskriminan," *International Journal of Natural Science and Engineering*, vol.1, no. 1 pp. 1-7, 2017.

[6] Z. Huang, H. Chen, C. J. Hsu, W. H. Chen, and S. Wu, "Credit analysis with support vector machines and neural network: a market comparative study, decision support system 37. *Elsevier*, pp. 543-558, 2003

[7] L. H. Lindholm and S. Mendhis, "Prevention of cardiovascular in developing countries," *The Lancet*, vol. 370, no. 9589, pp. 720-722, 2007.

[8] J. Mackay and G. A. Mensah. The atlas of heart disease and stroke, Geneva WHO, 2004, pp. 30-49.

[9] Khan, aurngzeb, B baharudin, L. H. Lee, and K. Khan. Review of Machine Learning: Algorithms and Applications, New York : CRC press, 2010.

[10] R. A. Johnson and D. W. Wichern, Applied Multivariate Statistical Analysis. NJ: PrenticeHall, 1992.

[11] W. R. Dillon and M. Goldstein, Multivariate Analysis Methods and Application. NY: John Wiley and Sons, Inc, 1984.

[12] A. S. Nugroho, A.B. Witarto and D. Handoko, Support vector machines: Teori Aplikasinya Dalam Bioinformatika. Kuliah Umum Ilmu Komputer.com, 2003

[13] E. Prasetyo, Data mining Konsep dan Aplikasi Menggunakan MATLAB, Yogyakarata: Andi, 2012

[14] X. Wu, B. Wu, J. Sun, S. Qiu, and X. Li, A hybrid fuzzy K-harmonic means clustering algorithm, Aplied mathematical model, 2015, pp. 3398- 3409

[15] D. T. Larose, Discovering Knowledge in Data: An Introduction to Data Mining, New Jersey: Wiley Interscience, 2005.

[16] Y. Yao, Y. Liu, Y. Yu, H. Xu, W. Lv, Z. Li, and X. Chen, "K-SVM: An effective SVM algorithm based on k-means clustering," *Journal of Computers*, pp. 2632-2639, 2013.

[17] K. Aprianto, "Optimasi kernel k-means dan pengelompokan kabupaten/kota berdasarkan indeks pembangunan manusia Indonesia," *J Math and Its Appl*, vol. 15, no. 1, pp. 1-15, 2018.

[18] C. Vercellis, Business Intelligence. United Kingdom: John Wiley and Sons, 2009.

[19] F. Gorunescu, Data Mining Concepts, Models and Techniques. Berlin: Springer, 2011.