Gede Suwardika. (2017). Pengelompokan dan Klasifikasi pada Data Hepatitis dengan Menggunakan Support Vector Machine (SVM), Classification and Regression Tree (Cart) dan Regresi Logistik Biner. *Journal of Education Research and Evaluation*. Vol.1 (3) pp. 183-191.

Pengelompokan Dan Klasifikasi Pada Data Hepatitis Dengan Menggunakan Support Vector Machine (SVM), Classification And Regression Tree (Cart) Dan Regresi Logistik Biner

Gede Suwardika[1, *]

[1] Universitas Terbuka, UPBJJ-UT Denpasar

Abstrak

Hepatitis adalah peradangan pada hati karena toxin, seperti kimia atauobat ataupun agen penyebab infeksi. Hepatitis yang berlangsung kurang dari 6 bulan disebut "hepatitis akut", hepatitis yang berlangsung lebih dari 6 bulan disebut "hepatitis kronis".Hepatitis biasanya terjadi karena virus, terutama salah satu dari kelima virus hepatitis, yaitu A, B, C, D atau E. Hepatitis juga bisa terjadi karena infeksi virus lainnya, seperti mononukleosis infeksiosa, demam kuning dan infeksi sitomegalovirus. Penyebab hepatitis non-virus yang utama adalah alkohol dan obat-obatan.Dalam penelitian ini dilakukan tes terhadap 155 pasien dengan respon meninggal atau hidup. Untuk itu penerapan Data Mining akan dilakukan pada kasus diatas, memanfaatkan salah satu teknik yaitu Data Classification, sejumlah data testing yang tersedia akan di analisis serta dibandingkan dengan data training untuk dilakukan prediksi meninggal atau hidup.Hasil ketepatan klasifikasi antara data training dengan data testing dengan analisis regresi logistik adalah 79,4% sedangkan dengan menggunakan SVM diperoleh sebesar 80%. Pengelompokan dengan menggunakan K-Means dan Kernel K-Means menghasilkan ketepatan pengelompokan yang berbeda. Ini menunjukkan bahwa data hepatitis memiliki pengelompokan yang baik. Kemudian hasil pengelompokan pada Kernel K-Means dibandingkan dengan data aktual yang diklasifikasikan dengan menggunakan regresi logistik, SVM dan CART dimana dihasilkan bahwa data hasil dari Kernel K-Means memiliki ketepatan klasifikasi yang lebih baik dibandingkan dengan hasil klasifikasi pada data aktual.

Keywords:
Regresi Logistik Biner, SVM, Kernel K-Means, K-Means, CART.

Pendahuluan

Penyakit Hepatitis adalah penyakit yang disebabkan oleh beberapa jenis virus yangmenyerang dan menyebabkan peradangan serta merusak sel-sel organ hatimanusia. Hepatitis diketegorikan dalam beberapa golongan, diantaranya hepatitis A,B,C,D,E,F dan G. Di Indonesia penderita penyakit Hepatitis umumnya cenderung lebih banyak mengalami golongan hepatitis B dan hepatitis C, namun disini kita akan membahas pada hidup atau matai pada penyakit Hepatitis, data diambil langsung dari website http://archive.ics.uci.edu/ml/datasets.html (UCI Machine Learning Repository)

Masalah klasifikasi banyak dijumpai dalam kehidupan sehari-hari seperti dalam penentuan diterima atau tidaknya pengajuan kredit dalam bidang perbankan, hingga diagnosis suatu penyakit di bidang kedokteran. Klasifikasi merupakan salah satu bentuk peramalan yang memiliki nilai keluaran diskrit, dan bertujuan untuk menemukan suatu fungsi keputusan f(x)yang secara akurat memprediksi kelas dari data (Santosa, 2007). Pola data dipelajari dengan pendekatan supervised learning untuk memprediksi data berikutnya yang memiliki kemiripan. Dalam pendekatan ini, label keluaran telah dikelompokkan, sehingga fungsi pemisah antara label satu dengan lainnya dapat dicari dengan mempelajari data kelas-kelas yang telah ada untuk mengklasifikasi data baru. Data yang digunakan untuk melatih fungsi disebut datatraining, sedangkan data untuk menguji model disebut data testing.

Dalam data mining dan machine learning telah dikembangkan berbagai metode klasifikasi2 seperti analisis diskriminan (linear discriminantanalysis), decision tree, Artificial Neural Networks, hingga Support Vector Machines (SVM). Pada beberapa penelitian, metode SVM telah terbukti mampu melakukan klasifikasi dengan baik untuk berbagai kasus. Menurut Frie, et. al., pencarian hyperplane dengan menggunakan program kuadratik SVM memiliki kelemahan yakni proses komputasi yang berat, berakibat pada waktu komputasiyang panjang.

Data Mining merupakan salah satu solusi yang dapat diterapkan untuk permasalahan data diatas. Data Mining itu sendiri adalah serangkaian proses yang dilakukan pada sejumlah data besar untuk diolah

* Corresponding author.
E-mail Addresses: isuwardika@ecampus.ut.ac.id (Gede Suwardika).

dan dihasilkan informasi yang lebih berguna, disiplin ilmu ini mengkaji berbagai metode yang umum digunakan untuk melakukan pengolahan data tersebut, salah satu metode pengolahan dalam prosesnya adalah klasifikasi data.

Klasifikasi data biasa digunakan pada sejumlah data yang telah di ketahui data induknya, untuk kemudian dijadikan data training/data model yang hasilnya akan menjadi keputusan prediksi dari sejumlah data yang serupa namun belum lengkap pada salah satu atributnya. Support Vector Machines (SVM) adalah sistem learning yang menggunakan sebuah ruang hipotesis fungsi linier dalam ruang fitur berdimensi tinggi, dilatih dengan menggunakan sebuah algoritma pembelajar dari teori optimasi yang mengimplementasikan sebuah bias learnig yang diturunkan dari teori learning statistika. Strategi learning yang diperkenalkan oleh Vapnik dan timnya merupakan sebuah metode yang powerful dalam beberapa tahun sejak diperkenalkan dan telah melebihi sistem yang lain dalam berbagai aplikasi.

Konsep dasar SVM adalah: (1) Class Separation, pada dasarnya, SVM mencari bidang hyperplane yang memisahkan secara optimal antara dua kelas dengan memaksimalkan margin antara titik terdekat kelas tersebut. Pada Gambar 1, terlihat bahwa titik yang berada pada batas dinamakan support vectors, dan bagian tengah margin merupakan bidang hyperplane yang memisahkan secara optimal. (2) Overlapping Classes, titik-titik data pada sisi "salah" dari diskriminan margin diturunkan untuk mengurangi pengaruhnya (soft margin). (3) Non Linearity, ketika tidak dapat ditemukan pemisah berbentuk linier, titik-titik data biasanya diproyeksikan ke dalam ruang dimensi yang lebih tinggi dimana titik-titik data secara efektif akan menjadi pemisah linier (proyeksi ini direaliasikan melalui teknik kernel). Dan (4) Problem Solution, semua tugas tersebut dapat diformulasikan sebagai permasalahan optimasi quadratik yang dapat diselesaikan dengan teknik yang diketahui.

.

Metode Penelitian

Data yang digunakan Hipatitis yang berasal dari UCI Machine Learning Repository. Variabel-variabel yang digunakan adalah: Class: DIE, LIVE, AGE: 10, 20, 30, 40, 50, 60, 70, 80, SEX: male, female, STEROID: no, yes, ANTIVIRALS: no, yes, FATIGUE: no, yes, MALAISE: no, yes, ANOREXIA: no, yes, LIVER BIG: no, yes, LIVER FIRM: no, yes, SPLEEN PALPABLE: no, yes, SPIDERS: no, yes, ASCITES: no, yes, VARICES: no, yes, BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00, ALK PHOSPHATE: 33, 80, 120, 160, 200, 250, SGOT: 13, 100, 200, 300, 400, 500, ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0, PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90, HISTOLOGY: no, yes.

Langkah-langkah yang dilakukan dalam penelitian ini yaitu: (1) engevaluasi data hepatitis apakah terdapat missing value. Kemudian setelah diketahui missing value yang besar pada variable, tetapi pada data ini variabel-variabel yang diketahui missing value tidak dihapus,tetapi tetap digunakan dengan cara dengan cara mengganti data pada variabel-variabel yang missing value tersebut menggunakan nilai mean. (2) Mengklasifikasikan data dengan menggunakan analisis regresi logistik biner. (3) Mengklasifikasikan data menggunakan SVM dengan bantuan software Matlab setelah sebelumnya membagi data menjadi 116 data training dan 39 data sebagai testing. (4) Pembandingan ketepatan klasifikasi antara analisis regresi logistik dan SVM. (5) Mengklasifikasikan dengan menggunakan K-Means dan Kernel K-Means kemudian menentukan hasil prediksi terbaik yang mendekati data aktual. Dan (6) Data aktual diklasifikasikan menggunakan analisis regresi logistik, SVM dan CART kemudian membandingkan hasilnya. Hasil prediksi terbaik pada langkah 7 diklasifikasikan menggunakan analisis regresi logistik, SVM dan CART kemudian membandingkan hasilnya.

Analisis dan Pembahasan

Preprocessing Data Missing Value

Missing value adalah informasi yang tidak tersedia untuk sebuah objek (kasus).Missing value terjadi karena informasi untuk sesuatu tentang objek tidak diberikan,sulit dicari, atau memang informasi tersebut tidak ada.Missing value pada dasarnya tidak bermasalah bagi keseluruhan data, apalagi jikajumlahnya hanya sedikit, misal hanya 1 % dari seluruh data. Namun jika persentasedata yang hilang tersebut cukup besar, maka perlu dilakukan pengujian apakah datayang mengandung banyak missing tersebut masih layak diproses lebih lanjut ataukahtidak. Cara lain dalam penanganan missing value yaitu: menghilangkan/membuang kasus atau objek yang mengandung missing value dan menghapus variabel (kolom) yang mengandung missing value.

Pada tahap ini akan dilakukan pengevaluasian terhadap banyaknya missing value. Variabel X19(PROTIME) memiliki banyak missing value yaitu sebesar 43,2%, karena dari 20 variabel yang digunakan ada 15 variabel yang missing value, tetapi disini tidak ada yang dihilangkan, data missing value

tetap digunakan dengan cara mengganti data pada variabel-variabel yang missing value tersebut menggunakan nilai mean.

Tabel 1. Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes[a]	
				Count	Percent	Low	High
v2	155	41.2000	12.56588	0	.0	0	1
v15	149	1.4275	1.21215	6	3.9	0	17
v16	126	105.3254	51.50811	29	18.7	0	5
v17	151	85.8940	89.65089	4	2.6	0	13
v18	139	3.8173	.65152	16	10.3	1	1
v19	88	61.8523	22.87524	67	43.2	0	0
v1	155			0	.0		
v3	155			0	.0		
v4	154			1	.6		
v5	155			0	.0		
v6	154			1	.6		
v7	154			1	.6		
v8	154			1	.6		
v9	145			10	6.5		
v10	144			11	7.1		
v11	150			5	3.2		
v12	150			5	3.2		
v13	150			5	3.2		
v14	150			5	3.2		
v20	155			0	.0		

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

Pada beberapa variabel dan observasi yang memiliki missing value dengan persentase kecil, maka kekosongan nilai dapat diisi dengan mean yang diperoleh dari masing-masing variable seperti pada Tabel dibawah ini :

Tabel. Statistics

		v4	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16	v17	v18	v19
N	Valid	154	154	154	154	145	144	150	150	150	150	149	126	151	139	88
	Missing	1	1	1	1	10	11	5	5	5	5	6	29	4	16	67
Mean		1.5065	1.3506	1.6039	1.7922	1.8276	1.5833	1.8000	1.6600	1.8667	1.8800	1.4275	105.3254	85.8940	3.8173	61.8523

Tabel. Statistika Deskriptif Masing-masing Variabel

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
Class	155	1,00	2,00	1,7935	,40607	,165
Age	155	20,00	78,00	41,6065	12,47100	155,526
Sex	155	1,00	2,00	1,1032	,30524	,093
Steroid	155	1,00	2,00	1,5065	,49996	,250
Antivirals	155	1,00	2,00	1,8452	,36292	,132
Fatigue	155	1,00	2,00	1,3506	,47717	,228
Malaise	155	1,00	2,00	1,6039	,48909	,239
Anorexia	155	1,00	2,00	1,7922	,40573	,165
LiverBig	155	1,00	2,00	1,8276	,36654	,134
LiverFirm	155	1,00	2,00	1,5833	,47673	,227
SpleenPalpable	155	1,00	2,00	1,8000	,39477	,156
Spider	155	1,00	2,00	1,6600	,46752	,219
Ascites	155	1,00	2,00	1,8667	,33549	,113
Varices	155	1,00	2,00	1,8800	,32071	,103
Balirubin	155	,30	8,00	1,4275	1,18830	1,412
AlkPhosphate	155	26,00	295,00	105,3254	46,40558	2153,478
Sgot	155	14,00	648,00	85,8940	88,47893	7828,521
Albumin	155	2,10	6,40	3,8173	,61675	,380

Protime	155	,00	100,00	61,8523	17,19353	295,617
Histology	155	1,00	2,00	1,4516	,49927	,249
Valid N (listwise)	155					

Tabel. Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	155	100,0
	Missing Cases	0	,0
	Total	155	100,0
Unselected Cases		0	,0
Total		155	100,0

a. If weight is in effect, see classification table for the total number of cases.

Dependent Variable Encoding

Original Value	Internal Value
DIE	0
LIVE	1

Block 0: Beginning Block

Tabel. Classification Table[a,b]

Observed			Predicted		
			Class		Percentage Correct
			DIE	LIVE	
Step 0	Class	DIE	0	32	,0
		LIVE	0	123	100,0
	Overall Percentage				79,4

a. Constant is included in the model.
b. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	1,346	,198	46,037	1	,000	3,844

Variables not in the Equation

			Score	df	Sig.
Step 0	Variables	Age	6,491	1	,011
		Sex	4,642	1	,031
		Steroid	2,808	1	,094
		Antivirals	2,627	1	,105
		Fatigue	14,800	1	,000
		Malaise	17,663	1	,000
		Anorexia	2,703	1	,100
		LiverBig	,808	1	,369
		LiverFirm	,534	1	,465
		SpleenPalpable	8,556	1	,003
		Spider	23,816	1	,000
		Ascites	34,282	1	,000
		Varices	20,423	1	,000
		Balirubin	31,453	1	,000
		AlkPhosphate	3,082	1	,079
		Sgot	,885	1	,347

	Albumin	33,634	1	,000
	Protime	14,631	1	,000
	Histology	17,693	1	,000
Overall Statistics		73,670	19	,000

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	85,246	19	,000
	Block	85,246	19	,000
	Model	85,246	19	,000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	72,611[a]	,423	,662

a. Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Classification Table[a]

Observed			Predicted		
			Class		Percentage Correct
			DIE	LIVE	
Step 1	Class	DIE	23	9	71,9
		LIVE	7	116	94,3
	Overall Percentage				89,7

a. The cut value is ,500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	Age	-,056	,030	3,386	1	,066	,946
	Sex	21,244	8055,172	,000	1	,998	1682796751,786
	Steroid	1,223	,834	2,150	1	,143	3,398
	Antivirals	-,065	1,158	,003	1	,955	,937
	Fatigue	,776	1,081	,516	1	,473	2,173
	Malaise	,539	,914	,347	1	,556	1,714
	Anorexia	-2,196	1,012	4,703	1	,030	,111
	LiverBig	-1,056	1,093	,933	1	,334	,348
	LiverFirm	-,813	,901	,814	1	,367	,444
	SpleenPalpable	,031	,898	,001	1	,972	1,032
	Spider	2,413	,854	7,974	1	,005	11,164
	Ascites	1,284	1,108	1,344	1	,246	3,612
	Varices	,908	1,004	,818	1	,366	2,480
	Balirubin	-,745	,340	4,794	1	,029	,475
	AlkPhosphate	-,002	,007	,057	1	,812	,998
	Sgot	-,001	,004	,051	1	,821	,999
	Albumin	,986	,746	1,746	1	,186	2,679
	Protime	,022	,024	,874	1	,350	1,023
	Histology	,458	,804	,325	1	,569	1,581
	Constant	-25,656	8055,174	,000	1	,997	,000

a. Variable(s) entered on step 1: Age, Sex, Steroid, Antivirals, Fatigue, Malaise, Anorexia, LiverBig, LiverFirm, SpleenPalpable, Spider, Ascites, Varices, Balirubin, AlkPhosphate, Sgot, Albumin, Protime, Histology.

Interpretasi Hasil :

- Tabel **dependen variable enconding** menunjukkan variabel Class diberi kode 1= LIVE dan 2 = DIE
- Output block : 0 beginning block
  1. Classification table menunjukkan tabel 2x2 dengan kolom berupa predicted values dari variabel dependen dan baris berupa nilai data aktual yang diamati. Untuk model yang sempurna, semua cases akan terletak pada diagonal tabel dan overall percentage akan bernilai 100%. Jika model regresi logistic mempunyai variance sama, maka nilai persen (%) padakedua baris hampir sama. Overall percentage yang memprediksi model dengan benar mempunyai nilai cukup baik sebesar $\frac{132}{155}$x 100% - 79,4 %
  2. Tabel **variables in the equation** yang hanya berisi **constant** memberikan nilai b_0 = 1,346 atau exp(1,346)= $e^{1,346}$ = 3,844. Karena responden yang mempunyai penyakit hepatitis dengan keadaan hidup (LIVE) ada 132 dan dengan keadaan mati (DIE) rendah ada 32, maka odd ratio = $\frac{32}{132}$ = 0,2424.
  3. Uji **wald** pada tabel **variables in the equation** digunakan untuk menguji apakah masing-masing koefisien regresi logistik signifikan. Uji **wald** sama dengan kuadrat dari rasio koefisien regresi logistic B dan standar error S.E . dalam kasus ini uji wald :
     = $[\frac{B}{S.E}]^2$ = $[\frac{1,346}{0,198}]^2$ = 46.21252933. *P*-value = 0,000 lebih kecil dari α = 0,05.
     Maka kesimpulannya **constant** dari model regresi logistic ini signifikan.
- Pada output block 1 : method enter
  1. Tabel **omnibus test of model coefficients** memberikan nilai chi-square goodness-of-fit test sebesar 85,246 dengan derajat kebebasan = 19,*P*-value=0,000 lebih kecil dari α = 0,05.. sehingga hasil uji ini sngat signifikan, chi-square goodness-of-fit test disini digunakan untuk menguji hipotesis :
     H_0 : memasukkan variabel independen ke dalam model tidak akan menambah kemampuan predeksi model regresi logistik
  2. Tabel **model summary** memberikan nilai statistic **-2 loglikehood = 72,611** . semakin kecil nilai **-2 loglikehood** semakin baik.
  3. Koefisien **cox & snall R square** pada tabel **model summary** dapat diinterpretasikan sama seperti koefisien determinasi R^2 pada regresi berganda. Tetapi karena nilai **cox & snall R square** biasanya lebih kecil dari 1 maka sukar untuk di interpretasikan dan jangan digunakan
  4. Koefisien **nagelkerkeR square** pada tabel **model summary** merupakan modifikasi dari koefisiensi **cox & snall R square** agar nilai maksimumnya bias mencapai satu dan mempunyai kisaran nilai antara 0 dan 1, sama seperti koefisien determinasi R^2 pada regresi linear nerganda. Nilai koefisien **nagelkerkeR square** umumnya lebih besar dari koefisien **cox & snall R square** tapi cenderung lebih kecil dibandingkan dengan nilai koefisien R^2 pada regrei linear berganda. Dalam contoh ini koefisien **nagelkerkeR square = 0,662.**
  5. Hasil perhitungan koefisien dari model regresi logistik biner ini terlihat pada tabel **variables in the equation** sebagai berikut :
     In $\left(\frac{\pi}{1-\pi}\right)$ = -25,656 – 0,056 Age + 21,244 Sex + 1,223 Steroid – 0,065 Antivirals + 0,776 Fatigue + 0,539 Malaise – 2,196 Anorexia – 1,056 LiverBig -0,813 LiverFirm + 0,031 SpleenPalpable + 2,413 Spider + 1,284 Ascites + 0,908 Varices – 0,745 Balirubin – 0,002 AlkPhosphate – 0,001 Sgot + 0,986 Albumin + 0,022 Protime + 0,458 Histology Atau $\frac{\pi}{1-\pi}$ =exp (-25,656 – 0,056 Age + 21,244 Sex + 1,223 Steroid – 0,065 Antivirals + 0,776 Fatigue + 0,539 Malaise – 2,196 Anorexia – 1,056 LiverBig -0,813 LiverFirm + 0,031 SpleenPalpable + 2,413 Spider + 1,284 Ascites + 0,908 Varices – 0,745 Balirubin – 0,002 AlkPhosphate – 0,001 Sgot + 0,986 Albumin + 0,022 Protime + 0,458 Histology)
  6. Kolom **Exp(B)** merupakan *odds ratio* yang diprediksi oleh model, misalnya :
     a. Untuk koefisien variabel **Age** :
        exp (-0,056) = $e^{-0,056}$ = 0,946
     b. untuk koefisien variabel **Sex :**
        exp (21,244) = $e^{21,244}$ = 1682796752
     c. Untuk **constant : (**exp -25,656) = $e^{-25,656}$ = 7,20677183E-12 ~ 0,000
  7. U**ji wald** manguji masing-masing koefisien regresi logistic, misalnya :

a. Untuk koefisien variabel **Age:**
   = $\left(\frac{B}{S.E}\right)^2 = \left(\frac{-0,056}{0,030}\right)^2$ = 3,386. *P*-value = 0,066 lebih kecil dari α = 0,05, maka koefisien regresi untuk variabel **Age**tidak signifikan.
b. Untuk koefisien variabel **Anorexia :**
   $\left(\frac{-2,196}{1,012}\right)^2$ = 4,703. *P*-value = 0,030 lebih kecil dari α = 0,05, maka koefisien regresi untuk variabel **Anorexia** signifikan.
c. Untuk **Constant** :
   $\left(\frac{-25,656}{8055,174}\right)^2$ = 0,000. *P*-value = 0,997 lebih besar dari α = 0,05, maka koefisien regresi untuk variabel **constant** tidak signifikan, artinya dari variabel-variabel prediktor, tidak semua variabel mempengaruhi LIVE dan DIE pada penyakit hepatitis.

Support Vector Machine (SVM)

SVM adalah suatu teknik yang relatif baru untuk melakukan prediksi, baik dalam kasus klasifikasi maupun regresi. Dalam hal ini data yang ingin diklasifikasikan adalah data echocardiogram. Dalam pengklasifikasian SVM ini ingin diketahui variabel y prediksi berdasarkan data training dan data testing, sehingga nantinya diketahui ketepatan y prediksi terhadap variabel y yang sebenarnya. Dengan bantuan program Matlab, didapatkan hasil seperti tabel berikut ini:

Tabel. Hasil Prediksi Y Dan Y Testing

Data Training	Data Testing	Hasil Prediksi		Ketepatan Klasifikasi
		Sesuai	Tidak sesuai	
78	77	43	24	69%
116	39	25	14	64%
145	10	8	2	80%

Jadi ketepatan klasifikasi dengan SVM antara y prediksi dengan y testing yang terbaik adalah sebesar 80% dengan training sebanyak 145dan 10 sebagai testing.

K-Means dan Kernel K-Means

Pengelompokan pada data hepatitis dengan menggunakan K-Means menghasilkan pengelompokan seperti pada tabel dibawah ini :

Tabel. Hasil Pengelompokan Menggunakan K-Means

Kelas	TotalAktual	Hasil Pengelompokan
-1	32	16
1	123	139

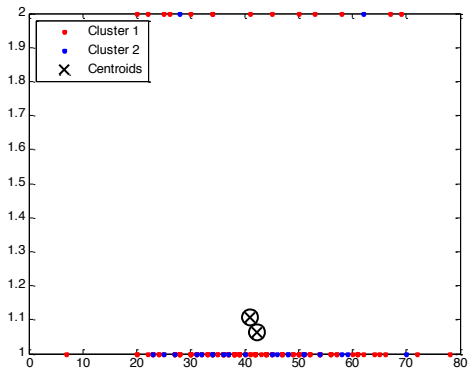
Dari hasil pengelompokan menggunakan K-Means, dapat dilihat bahwa sebanyak 16 data dikelompokkan dalam kelas pertama, sedangkan sisanya sebanyak 139 data dikelompokkan dalam kelas kedua.

Hasil yang didapatkan dari K-Means akan dibandingkan dengan kernel K-Means yang kemudian dibandingkan hasilnya untuk mendapatkan hasil pengelompokan terbaik yang mendekati data aktual. Dengan menggunakan kernel 'rbf'(Radial Basis Function) dan 'poly' menghasilkan error terkecil didapatkan hasil klasifikasi seperti pada Tabel dibawah ini :

Tabel. Hasil Pengelompokan Menggunakan Kernel K-Means

Kelas	Total	Hasil Pengelompokan
-1	32	22
1	123	133

Dapat dilihat bahwa dihasilkan pengelompokan yang sama persis. Gambar dibawah ini menunjukkan plot hasil pengelompokan dengan menggunakan metode kernel. Pada plot kernel dibandingkan dengan hasil kernel dengan error yang besar sebagai pembanding.



Gamba. Plot Hasil Pengelompokan Dengan K-Means

Klasifikasi Data Aktual dan Hasil K-Means Menggunakan Analisis Regresi Logistik, SVM dan CART

Pengklasifikasian pada data aktual berikut ini tanpa membagi data menjadi data testing dan data training dilakukan untuk mengetahui bagaimana hasil ketepatan klasifikasi dengan menggunakan analisis regresi logistik, SVM dan CART. Hasil pengklasifikasian dibandingkan dengan klasifikasi dengan menggunakan data hasil pengelompokan dengan menggunakan Kernel K-Means. Tabel dibawah ini, merupakan perbandingan hasil dengan menggunakan ketiga data dengan regresi logistik, SVM dan CART.

Tabel. PerbandinganKetepatan Klasifikasi

Y	Metode		
	Regresi Logistik	SVM	CART
Data Aktual	79,4%	79,4%	83,2%
Hasil K-Means	89,7%	90,3%	100%
Hasil Kernel K-Means	85,8%	98,7%	98,7%

Pengklasifikasian antara ketiga data dengan ketiga metode menghasilkan ketepatan klasifikasi dengan nilai yang besar. Pada data aktual, ketepatan klasifikasi terbesar adalah dengan menggunakan metode CART yaitu sebesar 83,2%, sedangkan pada data hasil K-Means dihasil ketepatan sebesar 100% dengan menggunakan CART dan data hasil Kernel K-Means ketepatan sebesar 98,7% dengan menggunakan metode SVM dan CART.

Berdasarkan tabel diatas dapat dilihat bahwa ketepatan klasifikasi ketiga data menggunakan ketiga metode menghasilkan ketepatan yang lebih besar dengan menggunakan data hasil dari Kernel K-Means. Hal ini dikarenakan data tersebut sudah merupakan hasil dari pengelompokan dengan metode K-Means, sehingga pengklasifikasian-nya akan lebih baik dibandingkan dengan data aktual.

Kesimpulan

Kesimpulan yang dapat dibuat berdasarkan hasil klasifikasi yang telah dilakukan adalah: Pada metode pengklasifikasian diperoleh bahwa hasil klasifikasi CART dan Pengklasifikasian dengan menggunakan K-Means dan Kernel K-Means menghasilkan ketepatan klasifikasi yang berbeda, yang mana dari kedua metode tersebut Kernel K-Means menunjukkan bahwa data hepatitis memiliki pengelompokan yang baik.

Daftar Pustakan

Agresti, A. 2007. *An Introduction to Categorical Data Analysis Second Edition*. USA : A John Wiley & Sons, Inc.

D.C. Montgomery. 1991. *Design and Analysis of Experiments*, Third Edition. John Wiley & Sons.

Gunn S. R. 1998. *Support Vector Machines for Classification and Regression*. TechnicalReport. University of Southampton.

Johnson, R.A. and Winchern, D.W. 2007. *Applied Multivariate Statistical Analys*is. USA: Pearson Education International

Lim, T.S. 1997. *Contraceptive Method Choice*. http://archive.ics.uci.edu, diakses 6 April 2012.

Nancy, JA. 1999. *Contraception : Present and Future*. Medical Journal of Indonesia. Vol.8: No. 1 .

Ratna M., Susilaningrum D. 2006. *Buku Ajar Analisys Multivariat FMIPA-ITS Surabaya.*

Ruslan, Mohammad. 2000.*Pengelompokan Wilayah di Jawa Timur Berdasarkan Komponen Penyusun Indeks Kemiskinan Manusia Sesudah dan Sebelum Krisis Ekonomi.*

Santosa, B. 2007. *Data Mining : Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta : Graha Ilmu.

Sobirin. 2006. *Mengenal Lebih Dalam Aneka Alat Kontrasepsi,*http://www.kafka.web.id/forum/kesehatan1, diakses 6 April 2012.

Santoso, Budi.2007. *Data Mining Terapan dengan Matlab*. Yogyakarta : Graha Ilmu

Santoso, Budi.2007. *Teori & Aplikasi Data Mining*. Yogyakarta : Graha Ilmu

Trihendredi, Cornelius., 2005. *Step by step SPSS 13 Analisis Data Statistik*. Yogyakarta : ANDI

Http://archive.ics.uci.edu/ml/datasets.html (UCI Machine Learning Repository)