RANDOM FOREST DAN *MULTIVARIATE ADAPTIVE REGRESSION SPLINE* (MARS) *BINARY RESPONSE* UNTUK KLASIFIKASI PENDERITA HIV/AIDS DI SURABAYA

[1]Nidhomuddin, [2]Bambang Widjanarko Otok
[1,2]Jurusan Statistika,Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Sepuluh November Surabaya
Alamat e-mail : mnidhomulmuluk@gmail.com

ABSTRAK

Human Immunodeficiency Virus (HIV) merupakan salah satu virus yang menurunkan sistem kekebalan tubuh. *Acquired immunodeficiency syndrome* (AIDS) merupakan penyakit menular yang disebabkan infeksi HIV. Pada tahun 2010, Jawa Timur menempati posisi kedua sedangkan tahun 2011 posisi keempat untuk kasus HIV/AIDS di Indonesia. Meskipun peringkatnya menurun namun jumlah kasusnya mengalami peningkatan yaitu 235 kasus (6,6%) dari tahun 2010. Hubungan variabel respon dan variabel prediktor terkadang tidak diketahui bentuk fungsi regresinya, karena itu menggunakan pendekatan regresi nonparametrik. Penelitian ini memiliki variabel prediktor yang berjumlah banyak sehingga menggunakan metode *Multivariate Adaptive Regrression Spline* (MARS). Untuk meningkatkan akurasi klasifikasi menggunakan metode *resampling* yakni *random forest* (RF) serta kombinasi antara metode MARS dan RF yang disebut RF MARS. Penelitian ini bertujuan untuk mendapat model terbaik dengan metode MARS berdasarkan nilai GCV minimum serta variabel-variabel yang berpengaruh terhadap HIV/AIDS di Surabaya dan mendapatkan tingkat akurasi klasifikasi penderita HIV/AIDS dengan metode MARS, RF, dan RF MARS.. Model MARS terbaik diperoleh saat kombinasi BF = 26, MI =3, dan MO = 0. Nilai GCV sebesar 0,1687. Dari 13 variabel prediktor yang dianalisis, 5 variabel masuk ke dalam model MARS terbaik yakni variabel usia, pekerjaan, pernah ditahan kasus NAPZA, status nikah, dan selalu pakai jarum steril. Akurasi klasifikasi status HIV/AIDS di Surabaya menggunakan metode MARS sebesar 80,28%. Pada metode RF diperoleh klasifikasi terbaik sebesar 97,80%. Pada RF MARS diperoleh Akurasi klasifikasi terbaik sebesar 91,00%.

Kata Kunci **:** *Multivariate Adaptive Regrression Spline, random forest,* HIV/AIDS

PENDAHULUAN

Human Immunodeficiency Virus (HIV) merupakan salah satu jenis virus yang menurunkan sistem kekebalan tubuh, sehingga orang yang terkena virus ini menjadi rentan terhadap beragam infeksi [11]. Berdasarkan Dinas Kesehatan Jawa Timur tahun 2012, *Acquired immunodeficiency syndrome* (AIDS) merupakan penyakit menular yang disebabkan oleh infeksi HIV yang menyerang sistem kekebalan tubuh. HIV/AIDS menjadi masalah kesehatan masyarakat yang penting di seluruh dunia pada awal abad ke 21. Mobilitas internasional dari individu yang terinfeksi memungkinkan terjadinya penyebaran HIV/AIDS secara global.

Jawa Timur merupakan salah satu provinsi yang berpotensi dalam menyumbang tingginya jumlah kasus AIDS di Indonesia. Berdasarkan Analisa Penyusunan Kineja Makro Ekonomi dan

Sosial Jawa Timur, diestimasi bahwa populasi rawan tertular HIV di Jawa Timur diperkirakan mencapai 0.05 persen penduduk. Pada tahun 2008 kasus kumulatif AIDS melonjak dua kali lipat dari kasus tahun 2007. Untuk menekan laju pertumbuhan jumlah kasus AIDS dapat dilakukan dengan cara mengetahui faktor-faktor yang berhubungan dengan AIDS dan berpotensi dalam meningkatkan jumlah kasus AIDS [7].

Berdasarkan Ditjen PPM dan PL Depkes RI tahun 2011, pada tahun 2010 Jawa Timur berada pada posisi kedua sedangkan pada tahun 2011 pada posisi keempat untuk kasus HIV/AIDS di Indonesia. Meskipun menunjukkan penurunan peringkat namun jumlah kasusnya tetap mengalami peningkatan yaitu 235 kasus (6,6%) dari tahun 2010. Surabaya memiliki jumlah kasus HIV/AIDS terbesar di Jawa Timur, hal ini dikarenakan adanya tempat lokalisasi yang berada di Dolly, sehingga banyak wanita pekerja seks yang rentan terhadap penyakit HIV/AIDS.

Penelitian sebelumnya tentang HIV/AIDS telah dilakukan oleh [10] yang meneliti Prevalensi dan Faktor Resiko HIV pada *Generalized Epidemic* di Tanah Papua Menggunakan Metode Regresi Logistik dengan Stratifikasi. [5] membahas preventif atau pencegahannya dengan metode *indepth interview*.

Keterbatasan informasi, bentuk fungsi, dan tidak jelasnya pola hubungan antara variabel respon dengan prediktor merupakan pertimbangan sehingga digunakan pendekatan regresi nonparametrik. Friedman mengenalkan metode regresi nonparametrik untuk kasus multivariate yang variabel bebasnya lebih dari dua dan dinamakan dengan metode MARS [4]. Dalam metode MARS terdapat MARS respon kontinu dan MARS respon kategorik. Pada MARS respon kategorik menggunakan Boostrap dalam MARS [8], sedangkan untuk MARS respon kontinu pemodelan MARS pada nilai ujian masuk terhadap Ipk [2].

Tingkat akurasi dari suatu metode klasifikasi dapat ditingkatkan dengan tujuan memberikan hasil klasifikasi yang lebih baik dan menurunkan tingkat kesalahan klasifikasi maka dilakukan metode *resampling* dalam penyusunan modelnya untuk menurunkan tingkat kesalahan klasifikasi. Bagging (*bootstrap aggregating*) dan Boosting merupakan metode ensemble yang relatif baru namun telah menjadi populer. Salah satu metode ensemble yang terbaru ialah *random forest* yang dikembangkan dari proses Bagging.

Random forest pertama kali dikenalkan oleh Breiman pada Tahun 2001. Dalam penelitiannya menunjukkan kelebihan *random forest* antara lain dapat menghasilkan error yang lebih rendah, memberikan hasil yang bagus dalam klasifikasi, dapat mengatasi data *training* dalam jumlah sangat besar secara efisien, dan metode yang efektif untuk mengestimasi *missing data* [1]. Penelitian sebelumnya tentang *random forest* dilakukan oleh [9] melakukan penelitian tentang *web caching* dengan membandingkan akurasi klasifikasinya menggunakan metode CART, MARS, *random forest* dan *Tree Net.* Penelitian tentang penerapan metode *random forest* dalam *driver analysis* [3]. Penelitian metode *ensemble* pada klasifikasi kemiskinan di Kabupaten Jombang dan diperoleh bahwa *random forest* memberikan akurasi klasifikasi yang terbaik [6].

METODE PENELITIAN

Sumber Data dan Variabel Penelitian

Data yang digunakan dalam penelitian ini adalah data sekunder berupa data kasus penderita HIV/AIDS di Kota Surabaya yang didapatkan dari skripsi S1 ITS Surabaya yang disusun

oleh Romaiza Millah Hanifa pada tahun 2013.

Banyaknya data yang digunakan pada penelitian ini sebanyak 218 sampel yang terdiri dari klien dengan status HIV/AIDS negatif dan klien dengan status HIV/AIDS positif. Jumlah masing-masing status dapat dilihat pada Tabel 1 di bawah ini.

Tabel 1. Jumlah dan Persentase Status HIV/AIDS

Status HIV/AIDS	Jumlah	Persentase
Negatif	170	78,0%
Positif	48	22,0%
Total	218	100,0%

Variabel respon (*Y*) dan Variabel-variabel prediktor adalah sebagai berikut

Tabel 2. Variabel penelitian

Variabel	**Nama Variabel**	**Kategori**
Y	Status Hiv	1 = Negatif 2 = Positif
Variabel Identitas Klien		
x_1	Jenis Kelamin	1 = Laki-Laki 2 = Perempuan
x_2	Usia	-
x_3	Pendidikan	1 = SMP 2 = SMA 3 = S1 4 = Tidak Bersekolah
x_4	Pekerjaan	1 = beresiko 2 = tidak beresiko
Pola Perilaku		
x_5	Status Nikah	1 = Kawin 2 = Cerai 3 = Tidak Kawin 4 = Tidak Open
x_6	Pasangan Tetap	1 = Ada, laki-laki 2 = Ada, Perempuan 3 = Tidak Ada
x_7	Pasangan Tidak Tetap	1 = Ada, laki-laki 2 = Ada, Perempuan 3 = Ada, laki-laki dan perempuan 4 = Tidak Ada
x_8	Selalu Pakai Kondom	1 = Ya 2 = Tidak
Riwayat Penggunaan Jarum Suntik		
x_9	Zat Yang Disuntikkan	1 = Putau 2 = Buphre 3 = Anti Depresan 4 = Putau dan Buphre 5 = Putau, Buphre, Metadhone, dan Anti Depresan
x_{10}	Selalu Pakai Jarum Steril	1 = Ya 2 = Tidak
x_{11}	Selalu Pakai Jarum Untuk Sendiri	1 = Ya 2 = Tidak
x_{12}	Pernah Ditahan Terkait Kasus Napza	1 = Ya 2 = Tidak
x_{13}	Pernah Ditahan Terkait Kasus Lain	1 = Ya 2 = Tidak

Langkah Penelitian

Langkah-langkah dalam penelitian ini adalah sebagai berikut:

Untuk mendapatkan pemodelan dengan pendekatan MARS respons biner adalah sebagai berikut

1. Mendeskriptifkan variabel respons dan variabel prediktor dalam pembentukan model.
2. Mendapatkan model MARS terbaik dengan trial dan error dengan tahapan sebagai berikut:
   a. Menentukan maksimum *basis function* (BF) = 26, 39, dan 52.
   b. Menentukan maksimum interaksi (MI) = 1, 2, dan 3.
   c. Menentukan minimal jumlah pengamatan setiap knots (MO) = 0, 1, 2, dan 3
3. Mendapatkan model terbaik dengan nilai GCV yang paling minimum.

4. Mendapatkan variabel yang masuk dalam model terbaik berdasarkan langkah ke 3.
5. Menentukan akurasi ketepatan klasifikasi.

Menentukan ketepatan klasifikasi dengan metode *random forest*

1. Menentukan *m* jumlah variabel prediktor yang diambil secara acak dan *k* pohon yang akan dibentuk untuk digunakan dalam klasifikasi *random forest*. Nilai *k* yang disarankan untuk digunakan pada metode bagging juga dicobakan yakni *k* = 50. Umumnya *k* = 50 sudah memberikan hasil yang memuaskan untuk masalah klasifikasi (Breiman, 1996). Sementara itu $k \geq 100$ cenderung menghasilkan tingkat misklasifikasi yang rendah (Sutton, 2005). Nilai *m* dan *k* yang dicobakan adalah:

$$m = \begin{cases} m_1 = \frac{1}{2}\left|\sqrt{p}\right| = 2 \\ m_2 = \left|\sqrt{p}\right| = 4 \\ m_3 = 2\left|\sqrt{p}\right| = 8 \end{cases}$$

$$k \begin{cases} k_1 = 25 \\ k_2 = 50 \\ k_3 = 100 \\ k_4 = 500 \\ k_5 = 1000 \end{cases}$$

2. Mengambil n sampel dengan teknik resampling dengan pengembalian sehingga diperoleh dataset baru D*
3. Membentuk *tree model* dari dataset D* dengan kombinasi *m* variabel prediktor yang diambil secara acak dan *k* buah ukuran pohon.
4. Melakukan voting mayoritas untuk setiap kali pohon.
5. Menentukan akurasi ketepatan klasifikasi.

Menentukan ketepatan klasifikasi dengan metode *random forest* MARS

1. Mendapatkan variabel-variabel yang masuk dalam model MARS terbaik pada bagian A langkah ke 4.
2. Mendapatkan p^* yaitu banyaknya variabel predictor pada model MARS terbaik
3. Menentukan *m* jumlah variabel prediktor yang diambil secara acak dan *k* pohon yang akan dibentuk untuk digunakan dalam klasifikasi *random forest*. Nilai *m* dan *k* yang dicobakan adalah:

$$m = \begin{cases} m_1 = \frac{1}{2}\left|\sqrt{p^*}\right| \\ m_2 = \left|\sqrt{p^*}\right| \\ m_3 = 2\left|\sqrt{p^*}\right| \end{cases}$$

$$k \begin{cases} k_1 = 25 \\ k_2 = 50 \\ k_3 = 100 \\ k_4 = 500 \\ k_5 = 1000 \end{cases}$$

4. Mengambil n sampel dengan teknik resampling dengan pengembalian sehingga diperoleh dataset baru D*
5. Membentuk *tree model* dari dataset D* dengan kombinasi *m* variabel prediktor yang diambil secara acak dan *k* buah ukuran pohon.
6. Melakukan voting mayoritas untuk setiap pohon.
7. Menentukan akurasi ketepatan klasifikasi.

HASIL PENELITIAN

A. Pemodelan Status HIV/AIDS Menggunakan MARS

Pemodelan status HIV/AIDS menggunakan pendekatan MARS dengan *Trial and error* yang dilakukan merujuk

dari Friedman (1991) dengan mengkombinasikan banyaknya *basis function* (BF), *maximum interaction* (MI) dan *minimum number of observation* (MO). banyaknya BF yang digunakan dalam pengolahan ini adalah 2 sampai dengan 4 kali banyaknya variabel prediktor yang diduga berpengaruh terhadap variabel respon. MI yang digunakan adalah 1,2 atau 3. Minimum observasi (MO) antar knot yang digunakan adalah 0, 1, 2 atau 3.

Tahap pembentukan model dilakukan dengan mengkombinasikan nilai-nilai BF, MI, dan MO yang telah ditentukan. Pemilihan model terbaik dilihat dari nilai GCV terkecil, namun bila GCV bernilai sama maka dilihat pada model yang memiliki ketapatan klasifikasi terbesar.

Tabel 4. *Trial And Error* Penentuan Model Terbaik MARS Status HIV/AIDS

Kombinasi			GCV	MSE	R^2	Keakuratan Klasifikasi (%)
BF	MI	MO				
26	1	0	0,1733	0,000	0,000	22,02
26	1	1	0,1714	0,156	0,112	66,06
26	1	2	0,1713	0,156	0,113	66,06
26	1	3	0,1711	0,156	0,114	66,06
26	2	0	0,1701	0,164	0,052	78,44
26	2	1	0,1724	0,161	0,073	79,82
26	2	2	0,1724	0,161	0,073	79,82
26	2	3	0,1733	0,000	0,000	22,02
26	3	0	0,1693	0,158	0,090	78,44
26	**3**	**1**	**0,1687***	**0,149**	**0,152**	**80,28***
26	3	2	0,1689	0,163	0,059	79,40
26	3	3	0,1687	0,163	0,060	79,40
39	1	0	0,1733	0,000	0,000	22,02
39	1	1	0,1733	0,000	0,000	22,02
39	1	2	0,1716	0,156	0,113	66,06
39	1	3	0,1715	0,156	0,114	66,06
39	2	0	0,1703	0,164	0,052	78,44
39	2	1	0,1727	0,161	0,073	79,82
39	2	2	0,1727	0,161	0,073	79,82
39	2	3	0,1728	0,167	0,038	29,82
39	3	0	0,1703	0,164	0,052	78,44
39	3	1	0,1690	0,163	0,052	79,40
39	3	2	0,1690	0,163	0,059	79,40
39	3	3	0,1688	0,163	0,060	79,40
52	1	0	0,1733	0,000	0,000	22,02
52	1	1	0,1728	0,161	0,082	73,85
52	1	2	0,1728	0,161	0,082	73,85
52	1	3	0,1728	0,161	0,082	73,85
52	2	0	0,1703	0,164	0,052	78,44
52	2	1	0,1724	0,161	0,073	79,82
52	2	2	0,1724	0,161	0,073	79,82
52	2	3	0,1729	0,167	0,038	29,82
52	3	0	0,1696	0,158	0,090	78,44
52	3	1	0,1691	0,163	0,058	79,40
52	3	2	0,1691	0,163	0,059	79,40
52	3	3	0,1687	0,163	0,060	79,40

Berdasarkan kriteria pemilihan model terbaik MARS maka yang terpilih adalah dengan model BF: 26, MI: 3 dan MO : 1 dengan bentuk model :

$\hat{f}(x)$= -0.134 + 0.110 * BF1 + 0.019 * BF2 + 0.025 * BF7 + 0.366 * BF11

BF1 = max(0, X2 - 48.000);
BF2 = max(0, 48.000 - X2);
BF5 = (X12 = 1) * BF2;
BF7 = (X4 = 1) * BF5;
BF10 = (X10 = 2);
BF11 = (X5 = 3) * BF10;

Pada Tabel 5 dapat dilihat variabel-variabel yang berpengaruh signifikan pada model.

Tabel 5. Variabel-Variabel Yang Mempengaruhi Pengurangan Nilai GCV Status HIV/AIDS

No	Variabel	Tingkat kepentingan (%)	-GCV
1	Usia (X2)	100,000	0,174
2	Pekerjaan (X4)	99,126	0,174
3	Pernah ditahan Kasus NAPZA (X12)	99,126	0,174
4	Status Nikah (X5)	54,906	0,170
5	Selalu Pakai Jarum Steril (X10)	54,906	0,170
6	Jenis Kelamin (X1)	0,000	0,169
7	Pendidikan (X3)	0,000	0,169
8	Pasangan Tetap (X6)	0,000	0,169
9	Pasangan Tidak Tetap (X7)	0,000	0,169
10	Selalu Pakai Kondom (X8)	0,000	0,169
11	Zat yang Disuntikkan (X9)	0,000	0,169
12	Selalu Pakai Jarum Sendiri (X11)	0,000	0,169
13	Pernah Ditahan Terkait Kasus Selain NAPZA (X13)	0,000	0,169

Pada Tabel 5 di atas dapat terlihat bahwa variabel usia adalah variabel

terpenting pada model MARS dengan tingkat kepentingan 100%. Kemudian diikuti berturut-turut oleh Pekerjaan, Pernah ditahan Kasus NAPZA, Status Nikah, Selalu Pakai Jarum Steril dengan besar kontribusi pada model adalah sebesar 99,126%, 99,126%, 54,906%, dan 54,906%. 8 variabel memiliki tingkat kepentingan 0,000% yang berarrti variabel-variabel tersebut tidak masuk dalam model karena sudah terwakili oleh variabel-variabel yang masuk model MARS.

Nilai minus GCV menunjukkan bahwa apabila variabel usia (X2) dimasukkan dalam model, maka nilai GCV akan berkurang sebesar 0,174. Apabila variabel pekerjaan (X4) dimasukkan dalam model, maka nilai GCV akan berkurang sebesar 0,174. Apabila variabel pernah ditahan kasus NAPZA (X12) dimasukkan dalam model, maka nilai GCV akan berkurang sebesar 0,174. Begitu juga status nikah (X5) dan selalu pakai jarum steril (X10) nilai GCV akan berkurang 0,170 dan 0,170. Kemudian variabel X1, X3, X6, X7, X8, X9, X11, dan X13 apabila dimasukkan dalam model maka nilai GCV akan berkurang masing-masing sebesar 0,169.

B. Akurasi Klasifikasi Status HIV/AIDS dengan Metode MARS

Akurasi klasifikasi status HIV/AIDS yakni status negatif dan status positif berdasarkan model MARS dihitung dengan menggunakan nilai ketepatan klasifikasi dapat dilihat pada Tabel 6

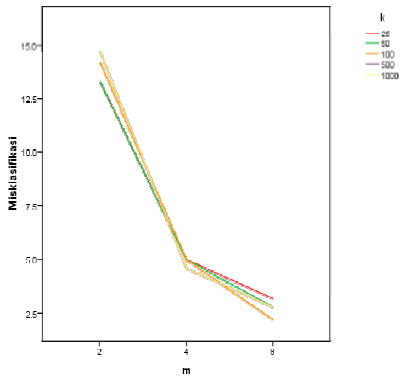
Tabel 6. Tabel Hasil Klasifikasi Status HIV/AIDS dengan Metode MARS

Kelas Aktual	Prediksi Kelas	
	Negatif	Positif
Negatif	160	10
Positif	33	15
Keakuratan Klasifikasi Total (%)		80,28
APER (100% - 80,28 %)		19,72
Sensitivity		94,12
Specificity		31,25

Total keakuratan klasifikasi sebesar 80,28% dan nilai APER (tingkat kesalahan klasifikasi)sebesar 19,72%.

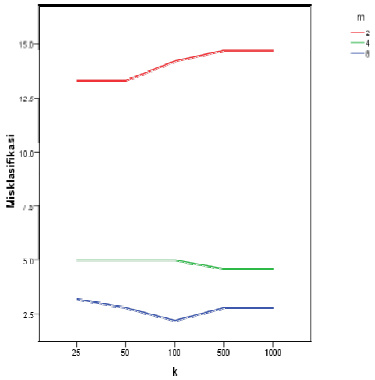
C. Akurasi Klasifikasi Status HIV/AIDS dengan Metode *Random Forest*

Akurasi prediksi *random forest* dapat diukur dari tingkat misklasifikasinya. Ukuran contoh peubah penjelas (*m*) dan ukuran *random forest* (*k*) menentukan stabil dan tingginya akurasi klasifikasi (Dewi, dkk. 2011).



Gambar 1. Tingkat Misklasifikasi *Random Forest* Berukuran *k* pada Beberapa Peubah Penjelas *m*

Pada gambar 1 menunjukkan perubahan nilai *m* menyebabkan tingkat misklasifikasi menjadi semakin turun. Tingkat misklasifikasi terendah selalu dicapai saat $m = 2\sqrt{p} = 8$. Hal tersebut menunjukkan bahwa *m* = 8 adalah *m* optimal.



Gambar 2. Tingkat Misklasifikasi *Random Forest* Peubah Penjelas *m* pada ukuran *k*

Pada gambar 2 menunjukkan perubahan misklasifikasi akibat berubahnya nilai *k*. terlihat bahwa perubahan nilai *k* berbeda-beda pada setiap pengambilan *m*. pada saat *m* = 2, semakin besar nilai *k* maka semakin besar pula tingkat misklasifikasi yang terjadi. Pada saat *m* = 4, semakin besar nilai *k* maka semakin kecil tingkat misklasifikasi. Pada saat *m* = 8, ketika ukuran *k* antara 25 sampai 100 nilai tingkat misklasifikasinya turun, kemudian saat *k* = 500 tingkat misklasifikasinya meningkat tetapi tidak begitu signifikan. Pada gambar terlihat pula tingkat misklasifikasi terendah terjadi pada saat *k* = 100. Dengan demikian dapat dikatakan bahwa akurasi *random forest* akan mencapai optimal saat *m* = 8 dan konvergen saat menggunakan 100 pohon dengan tingkat akurasi klasifikasi sebesar 97,8%.

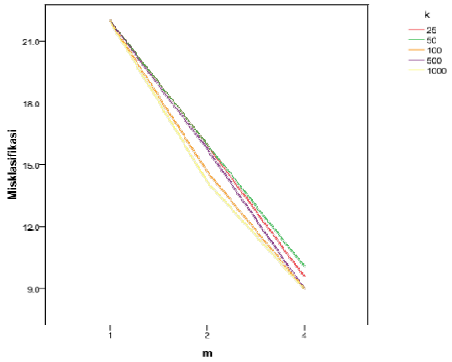
Tabel 7. Tabel Hasil Klasifikasi Status HIV/AIDS dengan Metode *Random Forest*

Kelas Aktual	Prediksi Kelas	
	Negatif	Positif
Negatif	170	0
Positif	5	43
Keakuratan Klasifikasi Total (%)		97,8
Aper (100% - 97,8 %)		2,2
Sensitivity		100
Specificity		95,55

Pada Tabel 7 di atas dapat diperoleh bahwa akurasi klasifikasi sebesar 97,8% dan kesalahan klasifikasi sebesar 2,8%.

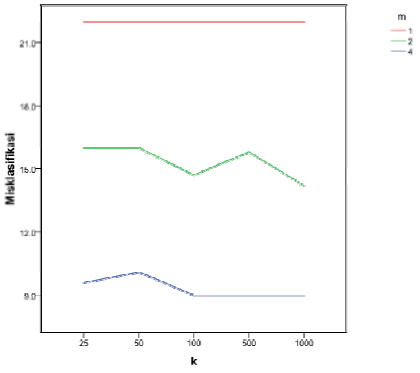
D. Akurasi Klasifikasi Status HIV/AIDS dengan Metode *Random Forest* MARS

Analisis klasifikasi dengan metode MARS menghasilkan model terbaik dan variabel-variabel yang masuk pada model. Berdasarkan tabel 4 diperoleh variabel yang masuk dalam model MARS terbaik yakni usia, pekerjaan, status nikah, selalu memakai jarum steril, dan pernah ditahan terkait NAPZA. Berdasarkan variabel tersebut kemudian akan dianalisis akurasi klasifikasinya menggunakan metode *random forest*. Metode gabungan ini yang disebut dengan *random forest* MARS. Langkah selanjutnya melakukan simulasi pada nilai *m* dan *k* yang telah ditentukan pada metodologi penelitian.



Gambar 3. Tingkat Misklasifikasi *Random Forest* MARS Berukuran *k* pada Beberapa Peubah Penjelas *m*

Pada gambar 3 terdapat tiga jenis peubah penjelas *m*, yakni 1, 2, dan 4 sesuai dengan metodologi pada bab 3. Gambar tersebut menunjukkan perubahan nilai *m* menyebabkan tingkat misklasifikasi menjadi semakin turun. Tingkat misklasifikasi terendah selalu dicapai saat $m = 2\sqrt{p} = 4$. Hal tersebut menunjukkan bahwa *m* = 4 adalah *m* optimal.



Gambar 4. Tingkat Misklasifikasi *Random Forest* MARS Peubah Penjelas *m* pada ukuran *k*

Gambar 4 menunjukkan perubahan misklasifikasi akibat berubahnya nilai *k*. terlihat bahwa perubahan nilai *k* berbeda-beda pada setiap pengambilan *m*. pada *m* = 1, tingkat misklasifikasi selalu sama pada setiap penambahan *k*. Pada saat *m* = 2, ketika ukuran *k* antara 25 sampai 100 nilai tingkat misklasifikasinya turun, kemudian saat *k* = 500 tingkat misklasifikasinya meningkat tetapi tidak begitu signifikan kemudian turun lagi saat *k* = 1000. Pada saat *m* = 4, semakin besar nilai *k* maka semakin kecil tingkat misklasifikasi dan terjadi konvergen saat *k* = 100. Dengan demikian dapat dikatakan bahwa akurasi *random forest* akan mencapai optimal saat *m* = 4 dan konvergen saat menggunakan 100 pohon dengan tingkat akurasi klasifikasi sebesar 91,0%..

Tabel 8. Tabel Hasil Klasifikasi Status HIV/AIDS dengan Metode *Random Forest* MARS

Kelas Aktual	Prediksi Kelas	
	Negatif	Positif
Negatif	168	2
Positif	18	30
Keakuratan Klasifikasi Total (%)		91,0
APER (100% - 91,0 %)		9,0
Sensitivity		98,82
Specificity		62,50

Pada Tabel 8 di atas dapat diperoleh bahwa akurasi klasifikasi sebesar 91% dan kesalahan klasifikasi sebesar 9%.

E. Perbandingan Akurasi Klasifikasi Status HIV/AIDS pada Metode MARS, *Random Forest,* dan *Random Forest* MARS

Kinerja metode klasifikasi diukur dari akurasi klasifikasi. Setelah melakukan analisis pada masing-masing metode diperoleh akurasi klasifikasinya pada tabel 9 berikut.

Tabe**l 9.** Tabel Akurasi Klasifikasi Status HIV/AIDS dengan Metode MARS, *Random Forest*, *Random Forest* MARS

Metode	Keakuratan Klasifikasi (%)
MARS	80,28
RF	97,80
RF MARS	91,00

Pada Tabel 9 di atas terlihat bahwa metode *random forest* memiliki akurasi klasifikasi tertinggi yakni sebesar 97,8%. Sehingga dapat disimpulkan untuk analisis klasifikasi status HIV/AIDS di Surabaya lebih baik menggunakan metode *random forest*.

KESIMPULAN

Model terbaik status HIV/AIDS di Surabaya memuat 5 variabel yang signifikan, variabel yang memiliki kepentingan paling tinggi untuk status HIV/AIDS adalah usia kemudian diikuti oleh Pekerjaan, Pernah ditahan Kasus NAPZA, Status Nikah, dan Selalu Pakai Jarum Steril. Tingkat keakurasian klasifikasi status HIV/AIDS di Surabaya menggunakan metode MARS menghasilkan akurasi sebesar 80,28%. Akurasi klasifikasi menggunakan metode *random forest* menghasilkan akurasi sebesar 97,80%. Akurasi klasifikasi menggunakan metode *random forest* MARS menghasilkan akurasi sebesar 91,00%. klasifikasi metode *random forest* lebih baik dibandingkan metode MARS dan *random forest* MARS.

DAFTAR PUSTAKA

[1] Breiman, L., 2001, *Random Forest.* Machine learning, 45(1):5-32. Kluwer Academic Publisher. Belanda.

[2] Budiantara, I.N., Suryadi, F., Otok, B.W., Guritno, S., 2006, Pemodelan

B-Spline dan MARS Pada Nilai Ujian Masu kterhadap IPK Mahasiswa Jurusan Disain Komunikasi Visual *UK*. Petra Surabaya; *Jurnal Teknik Industri*, Vol 8 No. 1,Universitas Petra.

[3] Dewi, N.K., Syafitri, U.D., Mulyadi, S.Y., 2011, Penerapan Metode Random Forest dalam Driver Analysis. *Forum Statistika dan Komputasi* 16(1):35-43.

[4] Friedman, J.H., 1991, Multivariate Adaptive Regression Spline (With Discussion), *The Annals of Statistics*, Vol. 19, hal. 1-141.

[5] Haryanto., Islami, I., Soemartono., Zauhar, S., 2010, *Implementasi Kebijakan Pencegahan dan Penanggulangan HIV/AIDS dan Infeksi Menular Seksual (IMS) di Kabupaten Jayapura.*

[6] Muttaqin, M.J. dan Bambang, W.O, 2013, *Metode Ensemble pada CART untuk Perbaikan Klasifikasi Kemiskinan*. Seminar Nasional Pascasarjan XI, Agustus 2013, Pascasarjana, ITS.

[7] Oktarina, Hanafi, F., dan Budisuari, M.A., 2009, Hubungan antara Karakteristik Responden, Keadaan Wilayah dengan Pengetahuan, Sikap terhadap HIV/AIDS pada Masyarakat Indonesia. *Buletin Penelitian Sistem Kesehatan 2009;* 24 : 362-36.

[8] Otok, B.W., Guritno, S., Subanar, Haryatmi, S. (2006), Bootstrap dalam MARS untuk Klasifikasi Perbankan. *Inferensi Jurnal Statistik*, Volume 2, N0. 1, Januari 2006. FMIPA ITS Surabaya..

[9] Sulaiman, S., Shamsuddin, S.M., Abraham, A. (2011), *Intelligent Web Caching Using Adaptive Regression Trees, Splines, Random Forests and Tree Net*. IEEE, 108-114.

[10] Susilo, B., 2009, *Prevalensi dan Faktor Resiko HIV pada Generalized Epidemic di Tanah Papua Menggunakan Regresi Logistik dengan Stratifikasi (Studi Kasus Surveilands Terpadu HIV-Perilaku (STHP) 2006).* Surabaya : Program Pasca Sarjana, Institul Teknologi Sepuluh Nopember.

[11] WHO, 2007, *Technical Working Group for The Development of an HIV/AIDS Diagnostic Support Toolkit*: p.2.