

PERBANDINGAN METODE REGRESI LOGISTIK BINER DAN *MULTIVARIATE ADAPTIVE REGRESSION SPLINE* (MARS) PADA PEMINATAN JURUSAN SMA (Studi Kasus SMA Negeri 2 Semarang)

Ratih Binadari¹, Yuciana Wilandari², Suparti³

¹Mahasiswa Jurusan Statistika FSM Universitas Diponegoro

^{2,3}Staff Pengajar Jurusan Statistika FSM Universitas Diponegoro

(ratihbinadari@gmail.com, yuciana.wilandari@gmail.com, supartisudargo@yahoo.co.id)

ABSTRACT

Major specialization at High School is aimed to gives opened opportunity for students to choose subject that are interest and develop their potential in accordance with the abilities, interests, talents, and personality. Major specialization at High school is influenced by some factors. To detect those factors, used biner logistic regression method and Multivariate Adaptive Regression Spline (MARS). Biner Logistic Regression is method that describes relationship between dependent variable and some independent variable, with independent variable has been coded 1 as representing the presence of the characteristic, and 0 as representing the absence of the characteristic. MARS is multivariate nonparametric regression method that development of Recursive Partitioning Regression (RPR) method and Spline method for high dimensional data that produces accurate prediction and continuous models on knots. Both of the methods are compared to know the best method used in research. From the result of analysis using biner logistic regression method and MARS, concluded that major specialization has been influenced by mathematic score, science score and relationship between students and friends. From proportion test, concluded classification that formed by regression logistic is as good as by MARS.

Keywords : Major specialization at High School, Biner Logistic Regression, Mutlivariate Adaptive Regression Spline (MARS), Clasification

1. PENDAHULUAN

Pendidikan adalah usaha sadar dan terencana untuk mewujudkan suasana belajar dan proses pembelajaran agar peserta didik secara aktif mengembangkan potensi dirinya untuk memiliki kekuatan spiritual keagamaan, pengendalian diri, kepribadian, kecerdasan, akhlak mulia, serta keterampilan yang diperlukan dirinya, masyarakat, bangsa dan negara (Undang-Undang Republik Indonesia Nomor 20 Tahun 2003, Pasal 1 Ayat 1). Pada jenjang Sekolah Menengah Atas, peserta didik akan mengikuti peminatan jurusan. Peminatan jurusan dilakukan untuk memberikan peluang kepada peserta didik untuk memilih mata pelajaran yang diminati, mendalami materi mata pelajaran dan mengembangkan berbagai potensi yang dimilikinya secara fleksibel. Ada beberapa faktor yang mempengaruhi peserta didik dalam peminatan jurusan. Menurut Syah (2009), faktor yang mempengaruhi antara lain faktor internal dan faktor eksternal. Dalam ilmu statistika, banyak metode yang dapat digunakan untuk mengetahui pengaruh variabel prediktor terhadap variabel responnya yang bersifat kategori. Beberapa metode tersebut adalah metode regresi logistik biner dan *Multivariate Adaptive Regression Spline* (MARS).

2. TINJAUAN PUSTAKA

2.1. Penjurusan di SMA

Pada Undang-Undang Sisdiknas 2003 tentang tujuan pendidikan menengah, terdapat 2 arahan yaitu mempersiapkan siswa ke jenjang perguruan tinggi dengan adanya penjurusan di SMA dan untuk terjun ke masyarakat (bekerja) dengan adanya sekolah kejuruan (SMK). Berdasarkan Peraturan Pemerintah mengenai pendidikan, penjurusan SMA dilakukan ketika siswa masuk SMA atau ketika siswa duduk di bangku kelas X. Sudiarto

(2013) menyebutkan bahwa tujuan diadakannya peminatan jurusan adalah untuk memberikan probabilitas yang lebih terbuka kepada peserta didik untuk memilih mata pelajaran yang diminati, mendalami materi mata pelajaran dan mengembangkan berbagai potensi yang dimilikinya secara fleksibel sesuai dengan kemampuan dasar umum (kecerdasan), bakat, minat dan karakteristik kepribadian tanpa dibatasi dengan sekat-sekat penjurusan yang terlalu kaku. Seperti yang sudah diatur dalam Undang-Undang, minat jurusan yang ada di SMA adalah ilmu alam, ilmu sosial dan ilmu bahasa.

Dalam penentuan minat jurusan, ada beberapa faktor yang mempengaruhi siswa. Menurut Syah (2009), faktor tersebut meliputi:

- a. Faktor Internal
Faktor Internal merupakan faktor yang berasal dari diri siswa itu sendiri, yang meliputi aspek fisiologis (bersifat jasmaniah) dan psikologis (bersifat rohaniah).
- b. Faktor Eksternal
Faktor eksternal merupakan faktor yang berasal dari lingkungan luar, yang meliputi lingkungan sosial seperti teman dan guru serta lingkungan nonsosial seperti keadaan sekolah, cuaca, dan waktu belajar dan lainnya.

2.2. Regresi Logistik Biner

2.2.1. Model Regresi Logistik Biner

Analisis regresi adalah suatu metode statistika yang digunakan untuk menganalisis dan menggambarkan hubungan antara variabel respon dengan beberapa variabel prediktor. Metode ini digunakan untuk menganalisis data yang variabel responnya berupa data kuantitatif (Montgomery dan Peck, 1992). Metode regresi biasa tidak dapat digunakan jika data variabel responnya berupa data kualitatif/kategori karena variabel responnya bersifat non numeris, keragaman respon yang tidak konstan serta ketidaknormalan sebaran dari variabel respon. Salah satu cara untuk mengatasi masalah ini adalah dengan menggunakan metode regresi logistik. Menurut Hosmer dan Lemeshow (2000), metode regresi logistik adalah suatu metode analisis statistika yang menjelaskan hubungan antara variabel respon yang berskala kategori yang memiliki dua kategori (biner) atau lebih dengan satu atau lebih variabel prediktor. Variabel respon biner adalah variabel respon yang hanya bernilai 1 untuk keberadaan suatu karakteristik dan 0 untuk ketidakberadaan karakteristik tersebut. Model regresi logistik adalah sebagai berikut:

$$\pi(x_i) = \frac{e^{\hat{g}(x_i)}}{1 + e^{\hat{g}(x_i)}}$$

dengan $\hat{g}(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}$

2.2.2. Estimasi Parameter

Menurut Hosmer dan Lemeshow (2000), metode estimasi yang digunakan untuk menaksir parameter pada model regresi logistik adalah metode *Maximum Likelihood Estimation* (MLE). Pada dasarnya, metode maksimum likelihood menghasilkan nilai parameter yang memaksimalkan probabilitas data observasi. Karena berdistribusi binomial dan observasi diasumsikan saling bebas, fungsi likelihoodnya adalah sebagai berikut:

$$l(\beta) = \prod_{i=1}^n \{\pi(x_i)\}^{y_i} \{1 - \pi(x_i)\}^{1-y_i} \quad (1)$$

Untuk mengestimasi nilai β , maka Persamaan (1) harus dimaksimalkan. Fungsi log likelihoodnya yaitu:

$$L(\beta) = \ln \left(\prod_{i=1}^n \{\pi(x_i)\}^{y_i} \{1 - \pi(x_i)\}^{1-y_i} \right) = \sum_{i=1}^n [y_i e^{\hat{g}(x_i)} - \ln (1 + e^{\hat{g}(x_i)})]$$

Untuk mendapatkan nilai β yang memaksimalkan nilai $L(\beta)$, maka $L(\beta)$ diturunkan terhadap $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ dengan hasil dari persamaan adalah nol. Jika ditulis dalam bentuk matriks, persamaan turunan pertamanya adalah:

$$\mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}(\mathbf{x}_i))$$

dengan:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}, \text{ dan } \mathbf{Y} - \boldsymbol{\pi}(\mathbf{x}_i) = \begin{bmatrix} y_1 - \pi(x_i) \\ y_2 - \pi(x_i) \\ \dots \\ y_n - \pi(x_i) \end{bmatrix}$$

Sedangkan bentuk matriks turunan keduanya adalah:

$$\mathbf{X}'\mathbf{V}\mathbf{X}$$

dengan:

$$\mathbf{V} = \begin{bmatrix} \pi(x_i)[1 - \pi(x_i)] & 0 & \dots & 0 \\ 0 & \pi(x_i)[1 - \pi(x_i)] & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \pi(x_i)[1 - \pi(x_i)] \end{bmatrix}$$

Karena estimasi parameternya tidak dapat dihitung langsung dari turunan pertama, maka digunakan metode numerik. Menurut Hastie, *et al.* (2009), untuk mendapatkan estimasi parameter, digunakan metode Newton Raphson dengan langkah sebagai berikut:

- Menentukan nilai taksiran awal untuk $\hat{\boldsymbol{\beta}} = \mathbf{0}$
- Menghitung $\mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}(\mathbf{x}_i))$ dan invers dari $\mathbf{X}'\mathbf{V}\mathbf{X}$
- Menghitung taksiran baru untuk setiap $(d + 1)$ dengan rumus:
$$\hat{\boldsymbol{\beta}}^{(d+1)} = \hat{\boldsymbol{\beta}}^{(d)} + \{\mathbf{X}'\mathbf{V}\mathbf{X}\}^{-1}\{\mathbf{X}'(\mathbf{Y} - \boldsymbol{\pi}(\mathbf{x}_i))\}$$
- Proses iterasi berhenti jika didapat hasil yang konvergen, $\hat{\boldsymbol{\beta}}^{(d+1)} \cong \hat{\boldsymbol{\beta}}^{(d)}$

Setelah mendapatkan estimasi parameter, langkah selanjutnya adalah menguji signifikansi parameter baik secara bersama-sama ataupun masing-masing.

2.2.3. Uji Signifikansi Parameter

Menurut Hosmer dan Lemeshow (2000), uji signifikansi parameter yang digunakan adalah uji Rasio Likelihood dan uji Wald.

- Uji Rasio Likelihood

Uji Rasio Likelihood adalah uji signifikansi parameter secara keseluruhan atau bersama-sama.

Hipotesis

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0$$

$$H_1 : \text{minimal ada satu } \beta_j \neq 0, \text{ dengan } j = 1, 2, \dots, p.$$

$$\text{Statistik Uji : } G = -2 \ln \left(\frac{\text{likelihood tanpa variabel bebas}}{\text{likelihood dengan variabel bebas}} \right)$$

$$\text{Kriteria Uji : } H_0 \text{ ditolak jika } G > \chi^2_{(\alpha; p)}$$

- Uji Wald

Uji wald digunakan untuk mengetahui apakah masing-masing variabel prediktornya memiliki pengaruh terhadap model atau tidak.

Hipotesis

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0, \text{ untuk } j = 1, 2, \dots, p.$$

$$\text{Statistik Uji : } W = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right)^2$$

$$\text{Kriteria Uji : } H_0 \text{ ditolak jika } W > \chi^2_{(\alpha; 1)}$$

2.3. Multivariate Adaptive Regression Spline (MARS)

Multivariate Adaptive Regression Spline (MARS) adalah salah satu regresi nonparametrik yang mengkombinasi regresi *spline* dan *Recursive Partitioning regression* (RPR) yang pertama kali diperkenalkan oleh Friedman tahun 1991. Menurut Friedman (1991), regresi *spline* menghasilkan persamaan bentuk parametrik polinomial tersegmen (terbagi dalam beberapa *region*), sedangkan RPR merupakan salah satu pendekatan komputasi yang digunakan untuk data berdimensi tinggi. Itu sebabnya, metode MARS dapat digunakan pada data yang berdimensi tinggi yaitu data yang memiliki variabel prediktor $3 \leq p \leq 20$ dan memiliki ukuran sampel yang berukuran $50 \leq n \leq 2000$.

Menurut Nash dan Bradford (2001), beberapa hal yang perlu diperhatikan dalam menggunakan metode MARS antara lain *knot* dan fungsi basis. *Knot* adalah sebuah titik yang memisahkan akhir suatu wilayah data dengan awal suatu wilayah data yang lain. *Knot* pada MARS dipilih menggunakan *forward stepwise* dan *backward stepwise*. Penempatan *knot* tergantung pada penentuan banyaknya amatan antar *knot*. Banyaknya amatan/observasi pada masing-masing *knot* disebut sebagai *Minimum Observation* (MO). MO yang digunakan adalah 0, 1, 2, dan 3. *Basis Function* (BF) adalah suatu fungsi yang dipisahkan oleh titik-titik *knot* yang menjelaskan hubungan antara variabel prediktor dan variabel respon. Metode MARS membentuk fungsi basis dengan prosedur *forward stepwise* dan *backward*. Friedman (1991) menyatakan bahwa jumlah BF adalah 2 sampai dengan 4 kali jumlah variabel prediktor, sedangkan jumlah *Maximum Interaction* (MI) adalah 1, 2 dan 3 dengan pertimbangan jika $MI > 3$ akan menghasilkan model yang kompleks dan interpretasinya akan hampir sama.

Menurut Friedman (1991), model untuk metode MARS adalah sebagai berikut:

$$\hat{f}(x) = \hat{a}_0 + \sum_{m=1}^M \hat{a}_m \prod_{k=1}^{K_m} [S_{km}(x_{j(k,m)} - t_{km})]_+$$

dengan :

\hat{a}_0 = koefisien konstanta fungsi basis

\hat{a}_m = koefisien dari fungsi basis ke-m

M = maksimum fungsi basis

K_m = derajat interaksi pada fungsi basis ke-m

S_{km} = tanda + atau - untuk interaksi ke-k, fungsi basis ke-m

$x_{j(k,m)}$ = variabel prediktor ke j, interaksi ke-k dan fungsi basis ke-m

t_{km} = nilai *knot* dari variabel prediktor $x_{j(k,m)}$

Menurut Hastie, *et al.* (2009), untuk model MARS dengan variabel kontinu, estimasi modelnya menggunakan OLS. Sedangkan model MARS dengan variabel respon biner, estimasi modelnya menggunakan metode *Maximum Likelihood Estimation* (MLE). Sama seperti metode regresi logistik, untuk mengestimasi nilai \hat{a} , dicari turunan pertama dan kedua dari fungsi log likelihoodnya kemudian dilanjutkan dengan iterasi.

2.4. Model Terbaik

Menurut Friedman (1991), model terbaik pada MARS ditentukan berdasarkan kriteria *Generalized Cross Validation* (GCV) minimum yang didefinisikan sebagai berikut:

$$GCV(M) = \frac{\frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}_M(x_i)]^2}{\left[1 - \frac{\tilde{C}(M)}{n}\right]^2}$$

dengan:

n = banyak pengamatan

M = jumlah fungsi basis

$$\tilde{C}(M) = C(M) + d.M, \text{ nilai } d \text{ terbaik berada pada interval } 2 \leq d \leq 4$$

$$C(M) = \text{Trace} [\mathbf{B} (\mathbf{B}' \mathbf{B})^{-1} \mathbf{B}'] + 1$$

2.5. Klasifikasi

Menurut Hosmer dan Lemeshow (2000), langkah awal klasifikasi dari variabel respon biner adalah menentukan titik potong. Variabel respon yang memiliki dua kategori (biner) dapat digunakan titik potong sebesar 0,50 dengan ketentuan jika $\pi(x) \geq 0,50$, maka hasil prediksi adalah 1 dan jika $\pi(x) < 0,50$, maka hasil prediksinya adalah 0. Klasifikasi pada pendekatan analisis regresi logistik dan MARS dapat menggunakan model probabilitas $\pi(x)$, yaitu:

$$\pi(x_i) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \dots + \hat{\beta}_p x_{pi}}} \text{ untuk regresi logistik biner}$$

$$\pi(x_i) = \frac{e^{\hat{a}_0 + \sum_{m=1}^M \hat{a}_m \prod_{k=1}^{K_m} [S_{km}(x_{p(k,m)} - t_{km})]_+}}}{1 + e^{\hat{a}_0 + \sum_{m=1}^M \hat{a}_m \prod_{k=1}^{K_m} [S_{km}(x_{p(k,m)} - t_{km})]_+}} \text{ untuk MARS}$$

2.6. Evaluasi Ketepatan Klasifikasi

Pengukuran klasifikasi dilakukan dengan matriks konfusi (*confussion matrix*), seperti berikut:

Tabel 1. Matriks Konfusi Klasifikasi Dua Kelas

Hasil Observasi	Prediksi	
	y_0	y_1
y_0	f_{00}	f_{01}
y_1	f_{10}	f_{11}

dengan:

f_{00} = jumlah objek pengamatan dari y_0 dan diklasifikasikan sebagai y_0

f_{01} = jumlah objek pengamatan dari y_0 dan diklasifikasikan sebagai y_1

f_{10} = jumlah objek pengamatan dari y_1 dan diklasifikasikan sebagai y_0

f_{11} = jumlah objek pengamatan dari y_1 dan diklasifikasikan sebagai y_1

Untuk mengevaluasi ketepatan klasifikasi, digunakan uji beda dua proporsi. Proporsi masing-masing metode didapatkan dengan cara menghitung nilai akurasi. Keakuratan suatu hasil prediksi digunakan untuk mengetahui besarnya proporsi data yang diklasifikasikan secara benar. Menurut Prasetyo (2012), keakuratan hasil prediksi dihitung dengan rumus:

$$PR = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}}$$

Selanjutnya dilakukan uji beda dua proporsi. Menurut Sugiarto (2000), langkah uji beda dua proporsi yaitu:

Hipotesis

H_0 : $PR_1 = PR_2$ (tidak ada perbedaan signifikan dari kedua metode)

H_1 : $PR_1 \neq PR_2$ (ada perbedaan signifikan dari kedua metode)

Taraf Signifikansi: $\alpha = 5\%$

Statistik Uji:

$$Z_{hitung} = \frac{PR_1 - PR_2}{\sqrt{(PR_{gab}(1 - PR_{gab}))x(\frac{1}{n_1} + \frac{1}{n_2})}}$$

dengan :

PR_1 = Proporsi metode regresi logistik biner

PR_2 = Proporsi metode MARS

PR_{gab} = Proporsi gabungan yaitu $\frac{n_1PR_1+n_2PR_2}{n_1+n_2}$

n_1 = ukuran sampel pada metode regresi logistik biner

n_2 = ukuran sampel pada metode MARS

Kriteria Uji : H_0 ditolak jika $Z_{hitung} > Z_{\alpha/2}$ atau $Z_{hitung} < -Z_{\alpha/2}$

Jika H_0 ditolak, maka terdapat perbedaan yang signifikan antara sistem klasifikasi metode regresi logistik biner dengan metode MARS. Sistem klasifikasi terbaik adalah sistem klasifikasi yang mempunyai nilai akurasi paling tinggi.

3. METODE PENELITIAN

3.1. Jenis dan Sumber Data

Data yang digunakan dalam penelitian ini adalah data primer yang diperoleh dengan membagikan kuisioner pada siswa SMA Negeri 2 Semarang.

3.2. Variabel Penelitian

Variabel yang digunakan dalam penelitian ini adalah:

Tabel 2. Variabel Penelitian

Variabel	Definisi
Variabel Respon	
Minat Jurusan (Y)	1 = IPA 2 = IPS
Variabel Prediktor	
Nilai UN Matematika (X_1)	
Nilai UN IPA (X_2)	
Nilai UN Bahasa Inggris (X_3)	
Hubungan Siswa dengan Teman (X_4)	1 = Rendah 2 = Sedang 3 = Tinggi
Hubungan Siswa dengan Guru (X_5)	1 = Rendah 2 = Sedang 3 = Tinggi
Hubungan Siswa dengan Keluarga (X_6)	1 = Rendah 2 = Sedang 3 = Tinggi
Motivasi dari Diri Sendiri (X_7)	1 = Rendah 2 = Sedang 3 = Tinggi

3.3. Langkah Analisis

Langkah analisis penelitian ini adalah:

1. Pengambilan data dilanjutkan dengan pengkodean data agar menjadi data siap olah
2. Menganalisis data dengan metode regresi logistik biner
3. Menganalisis data dengan metode MARS
4. Membandingkan hasil klasifikasi kedua metode

4. HASIL DAN PEMBAHASAN

4.1. Model awal metode Regresi Logistik Biner

Model awal yang terbentuk adalah:

$$\pi(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}}$$

dengan: $\hat{g}(x) = -15,335 + 1,045(X_1) + 1,269(X_2) - 0,472(X_3) + 1,277(X_{4_1}) + 1,061(X_{4_2}) - 0,748(X_{6_1}) - 0,484(X_{6_2}) + 1,895(X_{7_1}) + 0,335(X_{7_2})$

Selanjutnya adalah melakukan uji signifikansi parameter baik secara bersama-sama ataupun masing-masing variabel prediktor.

1. Uji Rasio Likelihood

Hipotesis

$H_0 : \beta_1 = \beta_2 = \dots = \beta_9 = 0$ (secara bersama-sama variabel prediktor tidak mempengaruhi model)

$H_1 : \text{minimal ada satu } \beta_j \neq 0, \text{ dengan } j = 1, 2, \dots, 9.$ (secara bersama-sama variabel prediktor mempengaruhi model)

Taraf Signifikansi : $\alpha = 5\%$

Statistik Uji : $G = -2 \ln \left(\frac{\text{likelihood tanpa variabel bebas}}{\text{likelihood dengan variabel bebas}} \right) = 89,658$

Kriteria Uji: H_0 ditolak jika $G > \chi^2_{(0,05;9)}$

Keputusan : Karena $(G = 89,658) > (\chi^2_{(0,05;9)} = 16,919)$ maka H_0 ditolak

Kesimpulan: jadi pada taraf signifikansi 5%, disimpulkan bahwa secara bersama-sama variabel prediktor mempengaruhi model.

2. Uji Wald

Hipotesis

$H_0 : \beta_j = 0$ (variabel ke-j tidak mempengaruhi model)

$H_1 : \beta_j \neq 0, \text{ untuk } j = 1, 2, \dots, 9.$ (variabel ke-j mempengaruhi model)

Taraf Signifikansi : $\alpha = 5\%$

Statistik Uji: $W = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right)^2$

Nilai Wald untuk masing-masing variabel adalah sebagai berikut:

Tabel 3. Nilai Wald untuk Setiap Parameter

Variabel Prediktor (X_j)	Wald (W)	$\chi^2_{(0,05;1)}$	Keputusan
Nilai UN Matematika (X_1)	16,191	3,841	H_0 ditolak
Nilai UN IPA (X_2)	28,024	3,841	H_0 ditolak
Nilai UN Bahasa Inggris (X_3)	2,416	3,841	H_0 diterima
Hubungan Siswa dengan Teman (X_4)			
Rendah (X_{4_1})	8,925	3,841	H_0 ditolak
Sedang (X_{4_2})	4,797	3,841	H_0 ditolak
Hubungan Siswa dengan Keluarga (X_6)			
Rendah (X_{6_1})	0,528	3,841	H_0 diterima
Sedang (X_{6_2})	1,306	3,841	H_0 diterima
Motivasi dari Diri Sendiri (X_7)			
Rendah (X_{7_1})	3,307	3,841	H_0 diterima
Sedang (X_{7_2})	0,170	3,841	H_0 diterima

Kriteria Uji: H_0 ditolak jika $W > \chi^2_{(0,05;1)}$

Kesimpulan : jadi dengan melihat Tabel 3, pada taraf signifikansi 5% disimpulkan bahwa variabel X_1, X_2 dan X_4 mempengaruhi model sedangkan variabel X_3, X_6 dan X_7 tidak mempengaruhi model.

Selanjutnya dilakukan pembentukan model akhir menggunakan variabel yang berpengaruh pada model. Model akhir yang didapat adalah:

$$\pi(x) = \frac{e^{\hat{g}(x_i)}}{1 + e^{\hat{g}(x_i)}}$$

dengan $\hat{g}(x_i) = -18,019 + 0,883(X_1) + 1,295(X_2) + 1,111(X_{4,1}) + 0,950(X_{4,2})$

4.2. Klasifikasi Regresi Logistik

Dengan menghitung probabilitas masing-masing observasi, maka didapatkan hasil klasifikasi untuk metode regresi logistik biner yaitu:

Tabel 4. Hasil Klasifikasi Metode Regresi Logistik Biner

Hasil Observasi	Hasil Prediksi	
	Kelas 0 (jurusan IPS)	Kelas 1 (jurusan IPA)
Kelas 0 (jurusan IPS)	21	38
Kelas 1 (jurusan IPA)	16	238
Proporsi	$\frac{21 + 238}{21 + 38 + 16 + 238} = 0,82748$	

Pada Tabel 4, terlihat bahwa ada 21 siswa jurusan IPS yang diprediksi masuk jurusan IPS, 38 siswa jurusan IPS yang diprediksi masuk jurusan IPA, 16 siswa jurusan IPA yang diprediksi masuk jurusan IPS dan 238 siswa jurusan IPA yang diprediksi masuk jurusan IPA.

4.3. Model MARS

Untuk mendapatkan model terbaik, perlu dilakukan *trial and error* terhadap kombinasi Basis Function (BF), *Maximum Interaction* (MI) dan *Minimum Observation* (MO). Hasil kombinasi BF, MI dan MO yang menghasilkan nilai GCV terkecil adalah kombinasi BF = 18, MI = 3 dan MO = 1. dengan nilai GCV yang dihasilkan adalah:

$$GCV(M) = \frac{\frac{1}{n} \sum_{i=1}^n [y_i - \hat{f}_M(x_i)]^2}{\left[1 - \frac{\tilde{C}(M)}{n}\right]^2} = \frac{\frac{1}{313} [33,609]}{\left[1 - \frac{13,571}{313}\right]^2} = 0,11733$$

dengan model yang terbentuk adalah:

$$\hat{f}(x) = 0,136 + 0,237*BF1 - 0,290*BF6 - 0,233*BF10 + 0,241*BF17$$

dengan:

$$BF1 = \max\{0; X_2 - 7,50\}$$

$$BF3 = \max\{0; X_1 - 5,25\}$$

$$BF4 = \{X_4 = 3\} * BF3$$

$$BF5 = \{X_4 = 1 \text{ OR } X_4 = 2\} * BF3$$

$$BF6 = \max\{0; X_2 - 7,50\} * BF5$$

$$BF10 = \max\{0; X_2 - 7,00\} * BF4$$

$$BF17 = \max\{0; X_2 - 6,75\} * BF3$$

Dari model terbaik yang dihasilkan, dapat disimpulkan bahwa variabel-variabel yang mempengaruhi model adalah nilai UN Matematika (X_1), nilai UN IPA (X_2) dan Hubungan dengan Teman (X_4).

4.4. Klasifikasi MARS

Dengan model probabilitas yang telah terbentuk, maka didapatkan hasil klasifikasi untuk metode MARS yaitu:

Tabel 5. Hasil Klasifikasi Metode MARS

Hasil Observasi	Hasil Prediksi	
	Kelas 0 (jurusan IPS)	Kelas 1 (jurusan IPA)
Kelas 0 (jurusan IPS)	30	29
Kelas 1 (jurusan IPA)	23	231
Proporsi	$\frac{30 + 231}{30 + 29 + 23 + 231} = 0,83387$	

Pada Tabel 5, terlihat bahwa pada hasil klasifikasi peminatan jurusan SMA menggunakan metode MARS terdapat 30 siswa jurusan IPS yang diprediksi masuk jurusan IPS, 29 siswa jurusan IPS yang diprediksi masuk jurusan IPA, 23 siswa jurusan IPA yang diprediksi masuk jurusan IPS dan 231 siswa jurusan IPA yang diprediksi masuk jurusan IPA.

4.5. Evaluasi Klasifikasi

Untuk mendapatkan metode terbaik, maka dilakukan evaluasi ketepatan klasifikasi dengan melakukan uji beda dua proporsi.

Hipotesis:

$H_0 : PR_1 = PR_2$ (tidak ada perbedaan signifikan dari kedua metode)

$H_1 : PR_1 \neq PR_2$ (ada perbedaan signifikan dari kedua metode)

Taraf Signifikansi: $\alpha = 5\%$

Statistik Uji:

$$PR_{gab} = \frac{313 \times 0,82748 + 313 \times 0,83387}{313 + 313} = \frac{520}{626} = 0,83067$$

$$Z_{hitung} = \frac{0,82748 - 0,83387}{\sqrt{(0,83067(1 - 0,83067)) \times \left(\frac{1}{313} + \frac{1}{313}\right)}} = -0,03690$$

Kriteria Uji: H_0 ditolak jika $Z_{hitung} < -Z_{\alpha/2}$ atau $Z_{hitung} > Z_{\alpha/2}$

Keputusan: karena $Z_{hitung} = -0,03690 > -Z_{\alpha/2} = -1,960$ maka H_0 diterima.

Kesimpulan: jadi pada taraf signifikansi 5% disimpulkan bahwa tidak ada perbedaan antara klasifikasi yang dihasilkan metode regresi logistik dan metode MARS pada peminatan jurusan SMA. Dengan kata lain, kedua metode tersebut menghasilkan sistem klasifikasi pada peminatan jurusan SMA yang sama baiknya.

5. KESIMPULAN

Dari analisis yang telah dilakukan, maka didapatkan kesimpulan yaitu baik menggunakan metode regresi logistik biner maupun metode MARS, faktor yang diduga

mempengaruhi peminatan jurusan SMA adalah nilai UN Matematika, nilai UN IPA dan hubungan siswa dengan teman. Sistem klasifikasi peminatan jurusan SMA yang dihasilkan metode regresi logistik biner dan metode MARS sama baiknya.

DAFTAR PUSTAKA

- Anonim. 2012. *Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 – Sisdiknas*. [online]. Tersedia: <http://www.unpad.ac.id/wp-content/uploads/2012/10/UU20-2003-Sisdiknas.pdf> (diakses 5 Januari 2015).
- Friedman, J. H. 1991. Multivariate Adaptive Regression Splines. *The Annals of Statistics*, Vol. 19.
- Hastie, T., Tibshirani, R. And Friedman, J. H. 2009. *The Elements of Statistical Learning: Data Mining, Inference and Prediction Second Edition*. New York: Springer Science Business Media.
- Hosmer, D. W. and Lemeshow S. 2000. *Applied Logistic Regression*. United States of American: Sons Inc.
- Montgomery, D.C. and Peck, E.A. 1992. *Introduction to Linear Regression Analysis, 2nd Edition*. John Wiley & Sons, New York.
- Nash, M. S. and Bradford, D. F. 2001. *Parametric and Non Parametric Logistic Regression for Prediction of Precense/ Absence of an Amphibian*. [online]. Tersedia: <http://www.epa.gov/esd/land-sci/pdf/0081eb02.pdf> (diakses 12 Maret 2015)
- Prasetyo, E. 2012. *Data Mining Konsep dan Aplikasi Menggunakan MATLAB*. Yogyakarta: Penerbit ANDI Yogyakarta
- Sudiarto, A. 2013. *Pedoman Arah Peminatan Draf 2*. [online]. Tersedia: <http://www.akursudianto.com/wp-content/uploads/2013/06/pedoman-arrah-peminatan-draf-2.pdf> (diakses 20 Desember 2014).
- Sugiarto, D.S. 2000. *Metode Statistika*. Jakarta: Gramedia Pustaka Utama.
- Syah, M. 2009. *Psikologi Pendidikan dengan Pendekatan Baru*. Bandung: Remaja Rosdakarya.