BAB II

TINJAUAN PUSTAKA

2.1 Tinjauan Non Statistik

2.1.1 Penjurusan SMA

Pada Undang-Undang Sisdiknas 2003 tentang tujuan pendidikan menengah, terdapat dua 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 mengeni pendiidkan, penjurusan SMA dilakukan ketika siswa masuk SMA atau ketika siswa duduk dibangku 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 kemalmpuan 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 Teknik Pengambilan Sampel

Sampel merupakan bagian populasi penelitian yang digunakan untuk memperkirakan hasil suatu penelitian sedangkan teknik sampling adalah bagian dari metodologi statistika yang berkaitan dengan cara-cara pegambilan sampel.

Teknik sampling adalah cara untuk menentukan sampel yang jumlahnya sesuai dengan ukuran sampel yang akan dijadikan sumber data sebenarnya, dengan memperhatikan sifat-sifatdan penyebaran populasi agar diperoleh sampel yang representatif (Margono 2004)

Tujuan dilakukannya pengambilan sampel adalah karena populasi terlalu banyak atau jangkauan terlalu luas sehingga tidak memungkinkan dilakukan

pengambilan data pada seluruh populasi, keterbatasan tenaga,waktu dan biaya, adanya asumsi bahwa seluruh populasi seragam sehingga bisa diwakili oleh sampel.

Pada penelitian ini peneliti menguakan teknik pengambilan sampling *Purposive Sampling*. *Purposive Sampling* adalah teknik sampling yang sering digunakan. Metode ini menggunakan kriteria yang telah dipilih oleh peneliti dalam memilih sampel. Kriteria pemilihan sampel terbagi menjadi kriteria inklusi dan kriteria eksklusi.

Kriteria inklusi merupakan kriteria sampel yang diinginkan peneliti berdasarkan tujuan penelitian. Sedangkan kriteria eksklusi merupakan kriteria khusus yang menyebabkan calon responden yang memenuhi kriteria inklusi harus dikeluarkan dari kelompok penelitian.

2.3 Regresi Logistik Biner

2.3.1 Model Regresi Logistik Biner

Analisis regresi logistik biner digunakan untuk menjelaskan hubungan antara variabel respon yang berupa data dikotomik/biner dengan variabel bebas yang berupa data berskala interval dan atau kategorik.

Regresi Logistik biner (*logistic regression*) sebenarnya sama dengan regresi berganda, hanya variabel terkaitnya merupakan variabel dummy (0 dan 1). Seperti pada penelitian ini variabel terkaitnya adalah 0 jika memilih penjurusan IPA dan 1 jika memilih penjurusan IPS. Tidak seperti *regresi logistik* biasa, *regresi logistik*

biner tidal mengansumsikan hubungan antara variabel independen dan dependen secara linier.

Model yang digunakan pada *regresi logistik* biner adalah :

$$\log(P/1-p) = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta kX \qquad (2.1)$$

Dimana p adalah kemungkinan bahwa Y = 1, dan X1, X2, X3 adalah variabel independen, dan b adalah koefisien regresi.

2.3.2. 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:

$$l(\beta) = \prod_{i=1}^{n} \{\pi(\chi_i)\}^{yi} \{1-\pi(\chi_i)\}^{1-yi} \qquad (2.2)$$

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

$$L(\beta) = \ln\left(\prod_{i=1}^{n} \{\pi(x_i)\}^{yi} \{1-\pi(x_i)\}^{1-yi}\right) = \sum_{i=1}^{n} \left[y_i e^{\hat{g}(x_i)} - \ln(1+e^{\hat{g}(x_i)})\right] \qquad (2.3)$$

Untuk mendapatkan nilai yang memaksimalkan nilai β, maka diturunkan terhadap β_0, β_1, β_2, ..., β_p dengan hasil dari persamaan adalah nol. Jika ditulis dalam bentuk matriks, persamaan turunan pertamanya adalah:

dengan: $X'\left(Y - \pi\left(x_{i)}\right)\right)$

$$\mathrm{X}=\begin{bmatrix} 1 & x_{11} & \cdots & x_{p1} \\ 1 & x_{12} & \cdots & x_{p2} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{1n} & \cdots & x_{pn} \end{bmatrix}, \text{ dan } Y-\pi(x_i)=\begin{bmatrix} y_1 & - & \pi(x_i) \\ y_2 & - & \pi(x_i) \\ & & \cdots \\ y_n & - & \pi(x_i) \end{bmatrix} \qquad (2.4)$$

Sedangkan untuk matriks turunan keduanya adalah :

$$V=\begin{bmatrix} \pi(x_i)[1-\pi(x_i)] & 0 & \cdots & 0 \\ 0 & \pi(x_i)[1-\pi(x_i)] & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \pi(x_i)[1-\pi(x_i)] \end{bmatrix} \qquad (2.5)$$

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:

a. Menentukan nilai taksiran awal untuk $\hat{\beta}=0$

b. Menghitung $X'(Y-\pi(x_i))$ dan invers dari $X'VX$

c. Menghitung taksiran baru untuk setiap (d +1) dengan rumus:

$$\hat{\beta}^{(d+1)}=\hat{\beta}^{(d)}+\{X'VX\}^{-1}\{X'(Y-\pi(x_i))\} \qquad (2.6)$$

d. Proses iterasi berhenti jika didapat hasil yang konvergen, $\hat{\beta}^{(d+1)}\cong\hat{\beta}^{d}$

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

2.3.3 Uji Signifikansi Parameter

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

a. Uji Rasio Likelihood

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

Hipotesis :

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

H_1 : minimal ada satu $\beta_j \neq 0$ dengan j =1, 2, …, p.

Statistik Uji : G = -2ln

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

b. Uji Wald

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

Hipotesis

$H_0 : \beta_j \neq 0$

$H_1 : \beta_j \neq 0$, untuk j = 1, 2, ..., p.

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

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

2.4 *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 $50 \leq n \leq 2000$. Variabel respon yang diolah pada metode MARS pun dapat berbentuk kontinu atau biner (Kriner, 2007).

Menurut Nash dan Bradford (2001), beberapa hal yang perlu diperhatikan dalam menggunakan metode MARS antara lain *knot* dan fungs 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:

$$\hat{f}(x) = \hat{a}_0 + \sum_{m=1}^{M} \hat{a}_m \prod_{k=1}^{K_m} \left[S_{km} \left(x_{j(k,m)} - t_{km} \right) \right]_+ \qquad (2.7)$$

dengan :

$\hat{a}_0$ = koefisien konstanta fungsi basis

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

M = maksimum fungsi basis

Km = derajat interaksi pada fungsi basis ke-m

Skm = 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

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, dicari turunan pertama dan kedua dari fungsi log likelihoodnya kemudian dilanjutkan dengan iterasi.

2.5 Model Terbaik

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

$$GCV(M) = \frac{\frac{1}{n}\sum_{i=1}^{n}\left[yi - \hat{f}_M(x_i)\right]^2}{\left[1 - \frac{\tilde{C}(M)}{n}\right]^2} \qquad (2.8)$$

dengan:

n = Banyaknya pengamatan

M = Jumlah fungsi basis

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

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

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 hasiln 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:

$$\pi(x_i) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_p x_{pi}}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_{1i}+\ldots+\hat{\beta}_p x_{pi}}} \quad \text{Untuk regresi logistik Biner} \qquad (2.9)$$

$$\pi(x_i) = \frac{e^{\hat{a}_0+\sum_{m=1}^{M}\hat{a}_m\prod_{k=1}^{K_m}\left[{}^{s}km\left({}^{x}p(k,m)-{}^{t}km\right)\right]_+}}{1+e^{\hat{a}_0+\sum_{m=1}^{M}\hat{a}_m\prod_{k=1}^{k_m}\left[{}^{s}km\left({}^{x}p(k,m)-{}^{t}km\right)\right]_+}} \quad \text{Untuk MARS} \qquad (2.10)$$

2.7 Evaluasi Ketepatan Klasifikasi

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

Tabel 2.1. Matrik 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 akurasinya. Keakurasian suatu hasil prediksi digunakan untuk mengetahui besarnya proporsi data yang diklasifikasikan secara benar. Menurut Prasetyo (2012), keakurasian hasil prediksi dihitung dengan rumus:

$$PR = \frac{f_{00} + f_{11}}{f_{00} + f_{01} + f_{10} + f_{11}} \qquad (2.11)$$

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: α = 5%

Statitik Uji : $$Z_{hitung} = \frac{PR_1 - PR_2}{\sqrt{\left(PR_{gab}\left(1 - PR_{gab}\right) x \left(\frac{1}{n_1} + \frac{1}{n_2} \right) \right)}} \quad (2.11)$$

dengan :

PR_1 = Proporsi metode regresi logistik biner

PR_2 = Proporsi metode MARS

PR_{gab} = Proporsi gabungan yaitu $\frac{n_1 PR_1 + n_2 PR_2}{n_1 + n_2}$

n1 = ukuran sampel pada metode regresi logistik biner

n2 = 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.

2.8 *Apparent Error Rate* (APER) dan *Total Accuracy Rate* (TAR)

Menurut Prasetyo (2012) , sebuah sistem yang melakukan klasifikasi diharapkan dapat melakukan klasifikasi semua *data set* dengan benar, tetapi tidak dapat dipungkiri bahwa kinerja suatu sistem tidak bisa 100% benar sehingga sebuah sistem klasifikasi juga harus diukur kinerjanya. Pada umumnya, pengukuran klasifikasi dilakukan dengan matriks konfusi *(confusion matrix).* Jika y_0 dan y_1 merupakan subjek pengklasifikasian, maka bentuk matriks konfusi dapat dilihat pada Tabel 3.2

Tabel 2.2 Hasil Klasifikasi Metode MARS

Hasil Observasi	Hasil Prediksi	
	Kelas 0	Kelas 1
Kelas 0	f_{00}	f_{01}
Kelas 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

Setiap sel fij dalam matriks menyatakan jumlah data dari kelas i yang hasil prediksinya masuk ke kelas j, f_{11} adalah jumlah data dalam kelas 1 yang secara benar dipetakan ke kelas 1, f_{10} adalah jumlah data dalam kelas 1 yang dipetakan secara salah ke kelas 0, f_{01} adalah jumlah data dalam kelas 0 yang dipetakan secara salah ke kelas 1, f_{00} adalah jumlah data dalam kelas 0 yang dipetakan secara benar ke kelas 0 (Prasetyo, 2012).

Menurut Johnson dan Wichern (1992), *Apparent Error Rate* (APER) adalah prosedur evaluasi yang digunakan untuk melihat kesalahan klasifikasi yang dilakukan oleh suatu fungsi klasifikasi. Nilai APER menunjukkan nilai proporsi sampel yang salah diklasifikasikan pada fungsi klasifikasi.

Total Accuracy Rate (TAR) digunakan untuk menghitung ketepatan klasifikasi pada hasil pengelompokan. Nilai TAR dapat menyatakan representasi proporsi sampel yang tepat diklasifikasi-kan.