Model Hibrida Untuk Penjurusan Siswa SMA

Purwanto[1], Sutini Dharma Oetomo[2], Ricardus Anggi Pramunendar[3]
Magister Teknik Informatika, Universitas Dian Nuswantoro Semarang
[1]Email : mypoenk@gmail.com; [2]Email : tintin@karangturi.sch.id; [3]Email : ricardus.anggi@yahoo.com

ABSTRAK

Penjurusan siswa di SMA merupakan rutinitas penting setiap tahun. Pada umumnya terdapat 2 jurusan utama di setiap sekolah, yaitu jurusan IPA dan IPS. Pemilihan jurusan ini sangat penting karena berkaitan dengan jurusan fakultas yang dapat dipilih oleh siswa pada jenjang pendidikan selanjutnya. Oleh sebab itu, diperlukan model yang cocok dari variabel-variabel yang mempengaruhi penjurusan tersebut. Dalam penelitian ini penulis mengusulkan model hibrida untuk penjurusan siswa di SMA. Model hibrida ini menggabungkan metode Logistic Regression dengan Support Vector Machine (SVM). SVM ini merupakan metode yang lebih handal dibandingkan metode-metode analisis lainnya.Hasil yang didapat dari penelitian menunjukkan model hibrida Logistic Regression dengan SVM memiliki tingkat akurasi yang lebih tinggi dibandingkan dengan metode SVM biasa.

Kata kunci: Support Vector Machine, Logistic Regression.

1. PENDAHULUAN

Logistic Regression (LR) merupakan salah satu metode analisis regresi yang digunakan untuk memprediksi hasil pengelompokan variabel dependent (terikat) berdasarkan satu atau lebih variabel independent (bebas)[1]. Metode LR ini dapat digunakan untuk melihat hubungan antar variabel independent dengan variabel dependent atau bahkan hubungan antar variabel independent itu sendiri. Relasi antar variabel dapat dilihat menggunakan teori Hafner yang menyatakan bahwa relasi antar variabel dapat diukur antara -1 hingga +1 dan dinyatakan dengan variabel *r* (Correlation Coeffisient). Angka 0 menunjukkan tidak adanya relasi antar variabel tersebut. Angka +1 dan -1 menunjukkan relasi yang tinggi antar variabel. Angka +1 menunjukkan relasi yang positif dan -1 menunjukkan relasi yang negatif[2]. Tabel hubungan antar variabel menuruf Hafner adalah sebagai berikut:

Tabel 1. Tabel nilai correlation coeffisient

Correation Coeffisient:	Hubungan antar variabel:
0.00 – 0.25	kecil atau tidak ada
0.30 – 0.45	Cukup
0.50 – 0.75	cukup hingga baik
0.80 – 1.00	kuat hingga sempurna

Variabel R^2 (R Square) digunakan untuk melihat apakah model sudah cocok atau belum. Pada model non-linearitas, ada dua macam metode yang digunakan, yaitu Cox and Snell R Square dan Nagelkerke R Square. Nagelkerke R Square merupakan perbaikan dari Cox and Snell R Square[3]. Nilai yang dihasilkan oleh variabel R^2 ini menunjukkan persentase kecocokan model.
Metode analisis lain yang digunakan dalam penelitian ini adalah Support Vector Machine (SVM). Metode ini pertama kali dikenalkan oleh Vapnikdan digunakan dalam pemecahan masalah *pattern recognition* dan estimasi fungsi non linier[4].Metode SVM ini bertujuan meminimalkan *upper bound* kesalahan generalisasi dengan memaksimalkan margin antara *hyperplane*(fungsi pemisah) dan data[5]. Metode Support Vector Machine ini banyak digunakan dalam penelitian-penelitian lain, seperti text classifier untuk artikel-artikel berbahasa Arab yang diteliti oleh Abdel Wadood Mohammad Abdel Wadood Mesleh[6]; text mining oleh Neelima Guduru[7].

2.DATA SET

Dataset yang digunakan dalam penelitian ini diambil dari pusat data siswa SMA Karangturi kelas 10 periode tahun ajaran 2012 – 2013. Keseluruhan jumlah data siswa yang digunakan adalah 150 data dengan jumlah variabel 6 yang terdiri dari 5 variabel independent dan 1 variabel dependent. Keenam variabel yang digunakan tersebut dapat dijabarkan sebagai berikut:

Tabel 2. Tabel Variabel yang digunakan dalam penelitian

Nama Variabel	**Keterangan**	**Pengukuran**
Kepercayaan diri	rasa percaya diri dalam melakukan segala sesuatu	5-Baik sekali, 4-Baik, 3-Cukup, 2-Kurang, 1-Sangat kurang
Kemandirian	kemandirian siswa dalam belajar	5-Baik sekali, 4-Baik, 3-Cukup, 2-Kurang, 1-Sangat kurang
Motivasi berprestasi	keinginan untuk maju dan berkembang	5-Baik sekali, 4-Baik, 3-Cukup, 2-Kurang, 1-Sangat kurang
Kemampuan berhitung	kemampuan siswa dalam perhitungan dan logika matematika	5-Baik sekali, 4-Baik, 3-Cukup, 2-Kurang, 1-Sangat kurang
Intelligence quotient (IQ)	hasil test IQ	Numerik
Jurusan	jurusan IPA (1) dan IPS (0)	1-IPA, 0-IPS

Contoh data set siswa yang digunakan:

Tabel 3. Data set Siswa SMA Karangturi kelas 10

No.	Kepercayaan Diri	Kemandirian	Motivasi Berprestasi	Kemampuan Berhitung	Minat Membaca	IQ	Jurusan
1	2	2	2	1	2	98	0
2	2	2	4	3	3	123	1
3	2	2	3	3	3	86	0
4	3	3	2	4	2	139	1
5	3	3	2	4	2	117	1
6	3	3	3	2	3	113	0

3. DASAR TEORI

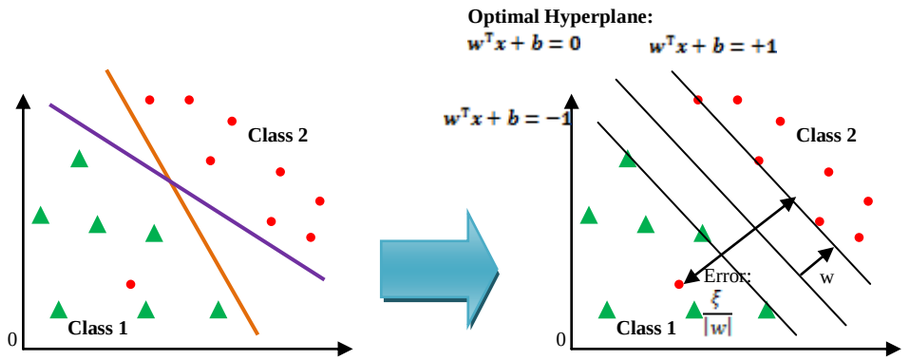
Logistic Regression (LR) merupakan model matematis yang digunakan untuk memprediksi variabel dependent dari satu atau beberapa variabel independent baik yang bersifat kategorial maupun kontinyu dan mendapatkan model yang cocok. Model yang didapat tersebut menggambarkan hubungan antara variabel dependent dan independent.

Secara matematis, LR dapat digambarkan sebagai berikut:

$$ln\left(\frac{p}{1-p}\right) = a + b_1x_1 + b_2x_2 + \cdots + b_ix_i \qquad (1)$$

$ln\left(\frac{p}{1-p}\right)$ ini menunjukkan logaritma natural dari probabilitas suatu peristiwa untuk terjadi dan probabilitas suatu peristiwa untuk tidak terjadi.Variabel **a** menggambarkan konstanta (intersep), variabel **b** merupakan koefisien regresi variabel prediktor (slope), variabel **x** adalah variabel independent yang pengaruhnya diteliti dan variabel **p** adalah probabilitas terjadinya "peristiwa" dari variabel dependent.

Metode SVM dikembangkan untuk memecahkan masalah data mining seperti klasifikasi, regresi dan fitur seleksi. Dalam bidang klasifikasi, metode SVM pertama kali diajukan oleh Vapnik tahun 1995. Metode SVM ini menentukan pemisahan optimal*hyperplane* yang mengklasifikasikan titik data ke dalam kategori yang berbeda.



Gambar 1. Mencari fungsi pemisah (hyperplane) secara linier

Dimana *w* merupakan bobot vektor dan *b* adalah bias.

Model SVM dapat diformulakan sebagai berikut:

Minimalkan

$\frac{1}{2}\|w\|^2$ (2)

Subject to

$y_i(\omega.x_i + \beta) \geq 1, \forall_i$

dimana ω, β adalah konstanta yang akan dicari

x_i : input dan y_i : output

$\|w\|^2 = \sqrt{(w,w)}$ dan merupakan 2-norm dari *w*

4. METODE PENELITIAN

Data set siswa yang digunakan dalam penelitian ini diuji potensial covariate-nya menggunakan Logistic Regression terlebih dahulu untuk mendapatkan variabel independent yang berpengaruh secara signifikan terhadap variabel dependent. Uji potential covariate Wald ini menghasilkan nilai sig. (p-value). Jika p-value yang dihasilkan lebih besar dari 0.25, maka variabel independent yang diuji tidak dapat menjadi kandidat model.

Setelah didapat variabel kandidat model, maka pengujian selanjutnya adalah uji kolinearitas antar variabel independent. Untuk menghasilkan model yang cocok maka di antara variabel independent yang telah menjadi kandidat model tidak boleh terjadi kolinearitas. Dengan menggunakan uji kolinearitas Spearman Rho akan didapat Correlation Coefficient antar variabel independent. Nilai Correlation Coefficient yang dibawah 0.8 menunjukkan bahwa diantara variabel independent tidak terjadi kolinearitas.
Langkah selanjutnya kandidat model yang sudah fit tersebut diuji menggunakan metode SVM. Data set siswa dibagi menjadi 2 macam, yaitu: 90% data training dengan 10% data test dan 80% data training dengan 20% data test. Pengujian menggunakan nilai Coefficient 0.1, 0.5 dan 0.9. Nilai gamma yang digunakan adalah 1.0 dengan kernel Radial.

5. HASIL DAN PEMBAHASAN

Hasil dari uji potential covariate terhadap variabel-variabel independent adalah sebagai berikut:

Tabel 4. Hasil uji potential covariate

nilai p-value	Jurusan
Kepercayaan diri	0.430
Kemandirian	0.159
Motivasi berprestasi	0.029
Kemampuan berhitung	0.785
IQ	0.000

Uji potential covariate Wald menghasilkan nilai signifikan hubungan variabel independent dengan variabel depndent. Nilai signifikan (p-value) ini memiliki nilai ambang 0.25. Semakin melebihi ambang nilai p-value suatu variabel menunjukkan hubungan variabel independent tersebut tidak signifikan terhadap variabel dependent[11]. Pada tabel 4diatas nilai p-value variabel Kepercayaan diri dan Kemampuan berhitung > 0.25 sehingga kedua variabel tersebut tidak dapat menjadi kandidat model.
Uji kolinearitas Spearman Rho menghasilkan nilai correlation coefficient antar variabel. Nilai correlation coefficient yang baik adalah kurang dari 0.8 karena nilai koefisien yang diatas 0.8 menunjukkan korelasi yang sangat kuat[12]. Uji kolinearitas ini dapat dilihat sebagai berikut:

Tabel 5. Hasil uji colinearitas

Correlations				
Variabel	**Kemandirian**	**Motivasi berprestasi**	**Minat membaca**	**IQ**
Kemandirian	1.000	0.042	-0.024	0.130
Motivasi berprestasi	0.042	1.000	0.052	0.051
Minat membaca	-0.024	0.52	1.000	-0.161
IQ	0.130	0.051	-0.161	1.000

Tabel 5. Menunjukkan hasil uji kolinieritasantar variabel independen. Dari hasil test, nilai Correlation Coefficient yang didapat tidak > 0.8, maka dapat disimpulkan tidak ada collinearitas antar variabel.
Pada tahap akhir pemilihan kandidat model, dilakukan uji potential covariate untuk kandidat model terpilih. Uji ini digunakan untuk melihat nilai signifikan tertinggi. Variabel dengan nilai signifikan tertinggi akan dikeluarkan dari kandidat model.
Tabel 6.Dibawah menunjukan hasil analisa LR multivariate dengan semua potensial covariate. Nilai p-value (sig.) untuk variabel Motivasi berprestasi paling tinggi diantara variabel independen lainnya, sehingga variabel ini dikeluarkan dari kandidat model.

Tabel 6. Tabel hasil analisa LR multivariate

	p-value
Kemandirian	0.193
Motivasi berprestasi	0.892
Minat membaca	0.116
IQ	0.000

Setelah didapat model yang *fit*, kemudian mulai diuji menggunakan metode SVM. Data set siswa yang digunakan dibagi menjadi 2 macam, yaitu: 90% data training dengan data testing sebesar 10% dan 80% data training dengan 20% data testing. Kernel yang digunakan adalah kernel Radial dengan nilaiCoefficientbervariasi 0.1; 0.5 dan 0.9. Nilai gamma = 1. Hasil pengujian dapat dilihat pada tabel berikut:

Tabel 7. Perbandingan nilai akurasi model dengan metode LR, SVM dan LR-SVM tanpa pembagian dataset

Kernel	*Coeffisient*	*Akurasi LR*	*Akurasi SVM*	*Akurasi LR-SVM*
Radial	*0.1*	*61.33%*	*52.00%*	*77.33%*
	0.5	*70.00%*	*70.67%*	*80.00%*
	0.9	*70.67%*	*74.00%*	*78.67%*

Pengujian berikutnya dilakukan menggunakan dataset yang dibagi menjadi data training sebesar 90% dan data testing sebesar 10%.

Tabel 8. Perbandingan nilai akurasi model dengan metode LR, SVM dan LR-SVM (90% training – 10% testing)

Kernel	*Coeffisient*	*Akurasi LR*	*Akurasi SVM*	*Akurasi LR-SVM*
Radial	*0.1*	*42.00%*	*40.00%*	*73.33%*
	0.5	*53.33%*	*54.00%*	*74.00%*
	0.9	*54.00%*	*59.33%*	*90.67%*

Pengujian berikutnya dilakukan menggunakan dataset yang dibagi menjadi data training sebesar 80% dan data testing sebesar 20%.

Tabel 9. Perbandingan nilai akurasi model dengan metode LR, SVM dan LR-SVM (80% training – 10% testing)

Kernel	*Coeffisient*	*Akurasi LR*	*Akurasi SVM*	*Akurasi LR-SVM*
Radial	*0.1*	*41.33%*	*40.00%*	*66.67%*
	0.5	*53.33%*	*53.67%*	*73.33%*
	0.9	*53.33%*	*55.00%*	*74.00%*

Dari hasil pengujian diatas didapat kecenderungan metode SVM yang lebih baik dibanding metode LR untuk nilai coeffisient diatas 0.1. Sedangkan model Hibrida LR-SVM yang diusulkan nampak lebih baik dibandingkan kedua metode lainnya. Namun pada saat data training diturunkan menjadi 80% keakurasian ketiga metode menurun. Hal ini terjadi karena dengan berkurangnya jumlah data training akan mempengaruhi nilai akurasi metode-metode lainnya.

6. KESIMPULAN

Pada penelitian ini penulis menguji penggunaan metode Logistic Regression, Support Vector Machine dan model hibrida LR-SVM. Dari pengujian diatas, didapatkan persentase nilai akurasi model dimana persentase model hibrida nampak paling bagus dibandingkan dengan metode lainnya.

Daftar Pustaka

[1] Karl L. Wuensch, “Binary Logistic Regression with PASW/SPSS”, 2011.
[2] Hafner, “Linear Regression and Correlation”, Bab 7, hal. 212 – 213.
[3] H.I. Mbachu, E.C. Nduka, M.E. Nja, “Designing a Pseudo R-Squared Goodness-of-Fit Measure in Generalized Linear Models”, Journal of Mathematics Research, Vol. 4, No. 2, April 2012.
[4] Moh. Yamin Darsyah, “Menakar Tingkat Akurasi Support Vector Machine Study Kasus Kanker Payudara”, Universitas Muhammadiyah Semarang: Statistika, Vol. 1, No. 1, Mei 2013, halaman 15 – 20.
[5] Yuh-Jye Lee, Yi-Ren Yeh, Hsing-Kuo Pao, “An Introduction to Support Vector Machines”, National Taiwan University of Science and Technology.
[6] Abdel Wadood Mohammad Abdel Wadood Mesleh, “Support Vector Machine Text Classifier for Arabic Articles: Ant Colony Optimization-Based Feature Subset Selection”, Arab Academy for Banking and Financial Sciences, 2008.
[7] Neelima Guduru, “Text Mining with Support Vector Machines and Non-Negative Matrix Factorization Algorithms”, University of Rhode Island, 2006.
[8] Zoran Bursac, C Heath Gauss, David Keith Williams and David W Hosmer, “Purposeful selection of variables in logistic regression”, Biology and Medicine 2008, 3:17, DOI 10.1186/1751-0473-3-17.
[9] Dr. Rick Yount, “Research Design and Statistical Analysis in Christian Ministry”, Bab 22: Correlation Coefficient, halaman 22-1, Edisi ke 4, 2006.