

KLASIFIKASI TUBERKULOSIS DENGAN PENDEKATAN METODE SUPPORTS VECTOR MACHINE (SVM)

Moh. Yamin Darsyah

Program Studi Statistika Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Muhammadiyah Semarang
Alamat e-mail : yamindarsyah@unimus.ac.id

ABSTRAK

Tuberkulosis (TBC atau TB) adalah suatu penyakit infeksi yang disebabkan oleh basil tahan asam disingkat BTA dengan nama lengkap bakteri *Mycobacterium Tuberculosis*. TBC menjadi salah satu penyebab kematian terbesar di Indonesia sehingga perlu penanganan khusus dalam pencegahan TBC. Data Kesehatan di Kota Semarang menunjukkan TBC menjadi salah satu kontributor terbesar dalam kejadian luar biasa di Provinsi Jawa Tengah. Beberapa faktor yang mempengaruhi penyakit TBC antara lain faktor lingkungan, jenis pekerjaan, status sosial, kebiasaan merokok dan minuman keras. Dalam menganalisis klasifikasi kasus pasien TBC terinfeksi atau tidak terinfeksi maka digunakan pendekatan Supports Vector Machine (SVM) dan Regresi logistik. Hasil menunjukkan SVM mampu mengukur ketepatan klasifikasi dengan akurasi lebih tinggi. Hasil akurasi yang diperoleh SVM dengan fungsi kernel RBF sebesar 98%.

Kata Kunci : SVM, RBF, TBC

PENDAHULUAN

Tuberkulosis (TBC atau TB) adalah suatu penyakit infeksi yang disebabkan oleh basil tahan asam disingkat BTA dengan nama lengkap bakteri *Mycobacterium Tuberculosis*. Bakteri ini merupakan bakteri basil yang sangat kuat sehingga memerlukan waktu lama untuk mengobatinya. Bakteri ini lebih sering menginfeksi organ paru-paru dibandingkan bagian lain dari tubuh manusia, sehingga selama ini kasus tuberkulosis yang sering terjadi di Indonesia adalah kasus tuberkulosis paru [4].

Penyakit tuberkulosis biasanya menular melalui udara yang tercemar dengan bakteri *Mycobacterium Tuberculosis* yang dilepaskan pada saat penderita batuk, sedangkan pada anak-anak sumber infeksi umumnya berasal dari penderita tuberkulosis dewasa. Basil

tuberkulosis terhisap melalui saluran pernapasan masuk ke dalam paru-paru, kemudian basil masuk lagi ke saluran limfe paru-paru dan menyebar ke berbagai organ tubuh melalui aliran darah. Selain manusia, satwa juga dapat terinfeksi dan menularkan penyakit tuberkulosis kepada manusia melalui kotorannya [4].

Tuberkulosis (TBC) merupakan masalah kesehatan, baik dari sisi angka kematian (*mortalitas*), angka kejadian penyakit (*morbiditas*), maupun diagnosis dan terapinya. Pada tahun 2004, *WHO Global Surveillance* memperkirakan di Indonesia terdapat 583.000 penderita tuberkulosis baru tiap tahun dengan 262.000 BTA positif atau insidens rate kira-kira 130 tiap 100.000 penduduk dan kematian akibat tuberkulosis diperkirakan menimpa 140.000 penduduk tiap tahun. Hasil Survey Kesehatan Rumah Tangga (SKRT) tahun 2010 menunjukkan bahwa tuberkulosis

merupakan penyebab kematian nomor tiga setelah penyakit kardiovaskuler dan penyakit saluran pernapasan pada semua golongan usia dan nomor satu dari golongan infeksi (Depkes, 2010). Kasus tuberkulosis terutama terjadi pada usia produktif kerja, yaitu kelompok umur 15 sampai 49 tahun yang berdampak pada kualitas SDM sehingga bisa mengganggu perekonomian keluarga, masyarakat dan negara [4].

Dengan adanya peningkatan kasus penularan infeksi tuberkulosis paru yang telah dilaporkan saat ini maka perlu adanya kajian teoritis terkait tentang penentuan variabel-variabel yang berpengaruh terhadap angka kejadian tuberkulosis paru, hal ini dimaksudkan agar jumlah penderita tuberkulosis paru di Indonesia dapat diminimalkan. Data Dinas Kesehatan Kota Semarang tahun 2010 menunjukkan bahwa Semarang sebagai kota penyumbang terbanyak angka kejadian tuberkulosis paru (TB Paru) di Jawa Tengah. Oleh sebab itu, pada penelitian ini diharapkan dapat diketahui model yang mewakili variabel-variabel yang mempengaruhi angka kejadian penyakit TB Paru di Kota Semarang.

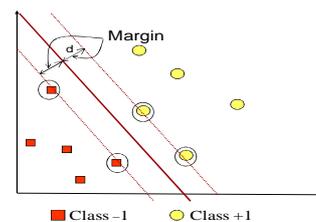
Penelitian sebelumnya tentang variabel-variabel yang diduga mempengaruhi tuberkulosis paru meliputi faktor umur, jenis kelamin, tingkat pendidikan, pekerjaan, kebiasaan merokok, kepadatan hunian kamar, ventilasi, kondisi rumah, kelembapan udara, status gizi, keadaan sosial ekonomi, dan perilaku (Prabu, 2008).

Dari penjelasan tersebut dapat dikatakan bahwa banyak faktor yang mempengaruhi angka kejadian penyakit TB Paru sehingga perlu dilakukan identifikasi faktor-faktor yang paling berpengaruh agar nantinya dapat digunakan untuk perencanaan pencegahan dan penanganan penyakit TB Paru sehingga angka kejadian penyakit ini dapat diminimalisir. Untuk

mendeskripsikan karakteristik anggota rumah tangga (ART) yang terinfeksi TB Paru dan melihat faktor yang paling berpengaruh pada suatu kelompok maka digunakan metode klasifikasi yang tepat berdasarkan faktor-faktor yang mempengaruhinya.

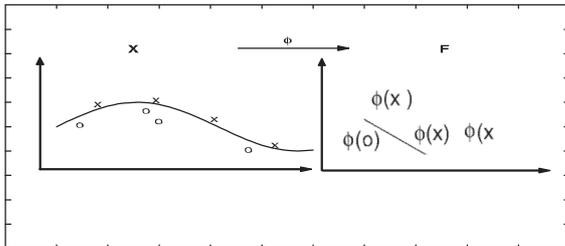
Support Vector Machine (SVM) salah satu metode *learning machine* yang bekerja atas prinsip *Structural Risk Minimization* (SRM) dengan tujuan menemukan *hyperplane* terbaik yang memisahkan dua buah *class* pada *input space* [8]. Selain itu SVM juga bertujuan untuk meminimalkan batas atas dari general eror. Keuntungan lain menggunakan SVM adalah metode ini dapat dianalisis secara teoritis menggunakan konsep teori pembelajaran komputasi.

Prinsip dasar SVM adalah *linier classifier*, kemudian dikembangkan untuk dapat bekerja pada kasus *non linier* dengan memasukkan konsep kernel pada ruang kerja berdimensi tinggi [8]. Misal diberikan himpunan $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ dimana data yang tersedia dinotasikan sebagai $\mathbf{x}_i \in R^n, i = 1, 2, \dots, m$. Gambar 2.1 memperlihatkan beberapa *pattern* yang merupakan anggota dari dua buah *class* : +1 dan -1. *Pattern* yang tergabung pada class -1 disimbolkan dengan kotak warna merah, sedangkan *pattern* pada class +1, disimbolkan dengan lingkaran warna kuning. Problem klasifikasi dapat diterjemahkan dengan usaha menemukan garis (*hyperplane*) yang memisahkan antara kedua kelompok tersebut.



Gambar 1. Optimal Hyperplane SVM

Pada umumnya, dalam permasalahan nyata, jarang ditemukan data *linear separable*. Sehingga fungsi Kernel digunakan dalam Support Vector Machine untuk mengatasi data non-linier. Dengan memasukkan fungsi Kernel, maka problem data non-linier menjadi linier dalam space baru seperti tampak pada ilustrasi berikut.



Gambar 2. Ilustrasi Problem Non-Linier menjadi Linier dengan Kernel SVM

Secara matematis, beberapa fungsi Kernel dijelaskan sebagai berikut:

1. Kernel Linier: $x^T x$,
2. Kernel Polynomial: $(x^T x_i + 1)^p$,
3. Kernel Radial Basis Function (RBF): $\exp\left(-\frac{1}{2\sigma^2} \|x - x_i\|^2\right)$,
4. Kernel Tangent hyperbolic: $\tanh(\beta x^T x_i + \beta_1)$, dimana $\beta, \beta_1 \in \mathbb{R}$

Regresi logistik adalah metode statistika yang mempelajari tentang pola hubungan secara matematis antara satu variabel respon (y) yang bersifat nominal atau ordinal dengan satu atau lebih variabel prediktor (x). Variabel respon pada regresi logistik merupakan variabel biner atau dikotomis. Menurut jenis skala dan variabel respon yang digunakan regresi logistik dibagi menjadi 3 macam, yaitu regresi logistik biner, multinomial dan ordinal.

Analisis regresi logistik biner adalah metode regresi yang bertujuan untuk mengetahui pengaruh antara variabel dependen (y) dan variabel independen (x) dimana variabel y menghasilkan 2 kategori yaitu 0 dan 1 (Hosmer dan Lemeshow, 1989). Sehingga variabel y

mengikuti distribusi Bernoulli dengan fungsi probabilitasnya sebagai berikut.

$$f(y) = \pi^y (1 - \pi)^{1-y} ; y = 0, 1 \quad (1)$$

Dimana jika $y = 0$ maka $f(y) = 1 - \pi$ dan jika $y = 1$ maka $f(y) = \pi$. Fungsi regresi logistiknya dapat dituliskan sebagai berikut.

$$f(z) = \frac{1}{1+e^{-z}} \text{ ekuivalen } f(z) = \frac{e^z}{1+e^z} \quad (2)$$

Dengan $z = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$

METODE PENELITIAN

Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari data Riset Kesehatan Dasar (Riskesdas) dan Survei Sosial Ekonomi Nasional (Susenas) tahun 2010 se-Indonesia yang bersumber dari Puslitbang Sistem dan Kebijakan Kesehatan, Badan Litbangkes Departemen Kesehatan Republik Indonesia.

Variabel yang digunakan dalam penelitian ini terdiri atas variabel dependen dan variabel independen, sebagai berikut :

1. Seorang anggota rumah tangga (ART) dikatakan terinfeksi penyakit tuberkulosis paru jika anggota rumah tangga (ART) tersebut pernah dinyatakan positif terinfeksi penyakit tuberkulosis paru dalam rentang satu bulan terakhir sebelum pelaksanaan survei (September-November 2010) dan sudah dikonfirmasi melalui pemeriksaan darah oleh tenaga ahli kesehatan (dokter/perawat/bidan). Ini merupakan variabel dependen (Y) yang berskala biner dengan ketentuan kode 1 bagi anggota rumah tangga (ART) yang terinfeksi penyakit tuberkulosis paru dan kode 2 bagi anggota rumah tangga (ART) yang tidak terinfeksi penyakit tuberkulosis paru.
2. Variabel independen (X) yang digunakan dalam penelitian ini

merupakan faktor-faktor yang berpengaruh terhadap angka kejadian penyakit tuberkulosis paru di Provinsi Papua Barat, baik dari segi lingkungan maupun perilaku masyarakat. Faktor-faktor yang digunakan dalam penelitian ini antara lain:

a. Pendidikan (X_1)

Variabel ini menunjukkan status pendidikan tertinggi yang dimiliki ART. Variabel ini berskala ordinal dengan enam kategori yaitu kode 1 jika tidak pernah bersekolah, kode 2 jika tidak tamat SD, kode 3 jika tamat SD, kode 4 untuk pendidikan SLTP, kode 5 untuk pendidikan SLTA, dan kode 6 untuk perguruan tinggi.

b. Pekerjaan (X_2)

Variabel ini menunjukkan jenis pekerjaan utama ART. Variabel ini memiliki skala nominal dengan delapan kategori yaitu kode 1 jika tidak bekerja, kode 2 untuk buruh (pekerja yang mendapat upah dalam mengolah pekerjaan orang lain, seperti buruh tani, buruh bangunan, buruh angkat angkut, buruh pekerja), kode 3 untuk nelayan, kode 4 untuk petani, kode 5 untuk wiraswasta, kode 6 untuk pegawai swasta, kode 7 untuk PNS (bekerja di pemerintahan sebagai pegawai negeri sipil), dan kode 8 untuk lainnya.

c. Status Sosial Ekonomi (X_3)

Variabel ini berskala nominal yang menunjukkan status ekonomi ART dengan dua kategori yaitu kode 1 yang menunjukkan status miskin dan kode 2 untuk status tidak miskin.

d. Kebiasaan Merokok (X_4)

Variabel ini menunjukkan kebiasaan ART dalam mengkonsumsi produk tembakau dalam satu bulan terakhir (bukan

hanya berupa rokok tetapi juga termasuk cerutu, cangklong, rokok linting, tembakau yang dikunyah). Variabel ini memiliki skala ordinal dengan empat kategori yaitu kode 1 jika merokok dengan frekuensi setiap hari, kode 2 jika merokok dengan frekuensi kadang-kadang, kode 3 jika tidak pernah merokok tetapi sebelumnya pernah merokok, dan kode 4 jika tidak pernah merokok sama sekali.

e. Kebiasaan Konsumsi Alkohol (X_5)

Variabel ini menunjukkan kebiasaan ART dalam mengkonsumsi minuman yang mengandung alkohol dalam 12 bulan terakhir (minuman alkohol bermerk : contohnya bir, whiskey, vodka, anggur/wine, dll serta minuman alkohol tradisional : contohnya tuak, poteng, sopi). Variabel ini memiliki skala nominal dengan dua kategori yaitu kode 1 jika mengkonsumsi alkohol dalam 12 bulan terakhir, dan kode 2 jika tidak mengkonsumsi alkohol dalam 12 bulan terakhir.

HASIL PENELITIAN

Tabel 1. Pengujian Parameter Serentak

	<i>Chi-square</i>	<i>df</i>	<i>Sig.</i>
Step	113.789	5	.000
Block	113.789	5	.000
Model	113.789	5	.000

Berdasarkan Tabel 1 menunjukkan bahwa pengujian parameter secara serentak membuktikan bahwa seluruh variabel independen berpengaruh terhadap penderita penyakit tuberkulosis paru dengan nilai yang sangat signifikan (<5%).

Berikut pengujian variabel secara parsial ditunjukkan Tabel 2. Bahwa Variabel Jenis pekerjaan (X_2), Kebiasaan merokok (X_4), Kebiasaan konsumsi alkohol (X_5) merupakan variabel yang sangat berpengaruh secara signifikan terhadap penderita penyakit tuberkulosis paru sedangkan variabel tingkat pendidikan dan status sosial ekonomi tidak berpengaruh secara signifikan (>5%).

Tabel 2. Pengujian Parameter Parsial

Variabel x	Sig.
1	.056
2	.043
3	.051
4	.039
5	.043

Tingkat akurasi dalam kasus pengklasifikasian sangat diperlukan karena mengindikasikan tepatnya suatu kategori penyakit itu terinfeksi atau tidak pada pasien sehingga untuk mengetahui tingkat akurasi dalam kasus tuberkulosis paru perlu di lakukan untuk mendeteksi akurasi kebenarannya.

Tabel 3. Pengujian Parameter Parsial

Tingkat akurasi	SVM (RBF)	Regresi logistik
Y	.981	.713

Dalam Tabel 3 menunjukkan tingkat akurasi begitu tinggi sebesar 98% pada metode klasifikasi dengan menggunakan SVM dengan kernel RBF, apabila ketepatan akurasi dalam klasifikasi tinggi maka tindakan penanganan secara medis akan tepat dilakukan.

KESIMPULAN

Variabel-variabel Jenis pekerjaan (X_2), Kebiasaan merokok (X_4), Kebiasaan konsumsi alkohol (X_5) merupakan variabel yang sangat berpengaruh secara signifikan terhadap

penderita penyakit TB paru. Tingkat akurasi dalam klasifikasi dengan metode SVM diperoleh sebesar 98%. Untuk penelitian selanjutnya bisa menggunakan metode klasifikasi lainnya agar tingkat akurasi diketahui untuk penanganan medis yang tepat.

DAFTAR PUSTAKA

- [1] Agresti, A. 1990. *Categorical Data Analysis*. New York: John Wiley and Sons.
- [2] Breiman, L., Friedman, J., Olshen, R., dan Stone, C. 1993. *Classification and Regression Trees*. New York - London: Chapman Hall.
- [3] Darsyah, M.Y. 2013. *Menakar Tingkat Akurasi Support Vector Machine Pada Kasus Kanker Payudara*. Jurnal Statistika UNIMUS Vol.1 Nomor.1. Semarang.
- [4] Indriani, D., Adiningsih, S., Mahmudiono, T. (2005). *Faktor resiko yang mem-pengaruhi kejadian TB paru pada anak jalanan dengan studi kasus di Yayasan Insani Surabaya*. Jurnal FKM UA. Surabaya
- [5] Johnson, R.A. and Wichern D.W. 2002. *Applied Multivariate Statistical Analysis. 5th Ed.* New Jersey USA: Prentice-Hall Ic.
- [6] Le, C. T. 1998. *Applied Categorical Data Analysis*. USA: John Wiley and Sons.
- [7] Lewis dan Roger J., 2000. *an Introduction to Classification and Regression Trees (CART) Analysis*. Presented at the 2000.
- [8] Santosa, B. 2007. *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu.