MODEL DIAGNOSIS TUBERKULOSIS MENGGUNAKAN *k-NEAREST NEIGHBOR* BERBASIS SELEKSI ATRIBUT

Ratih Sari Wardani[1)], Purwanto[2)]
[1] Kesehatan Masyarakat, Universitas Muhammadiyah Semarang
email: ratihsw@gmail.com
[2] Ilmu Komputer, Universitas Dian Nuswantoro
email: purwanto@dsn.dinus.ac.id

Abstract

The objective of this paper is to obtain a diagnosis model of Tuberculosis (TB) using k-Nearest Neighbor based on feature selection. Data is collected from BKPM Semarang, Central Java. The data consist of characteristics, anamnesis, physical examination, laboratory test results, radiological examination, duration of cough and sputum color. The results indicate that the k-Nearest Neighbor based on backward elimination model improvements as high as 78.66% % compared to individual models.

Keywords: *k-Nearest Neighbor, backward ellimination , Tuberkulosis, diagnosis, pengambilan keputusan*

1. PENDAHULUAN

Tuberkulosis (TB) sampai saat ini masih menjadi permasalahan kesehatan di Indonesia. Tuberkulosis (TB) merupakan penyakit menular yang dapat menyerang Paru yang disebabkan oleh Mycobacterium tuberculosis. World Health Organization (WHO) melaporkan terdapat 9 juta kasus baru (*incident rate*)

TB pada tahun 2013. angka kecenderungan terus meningkat dari 6,6 juta kasus pada tahun 1990, 8,3 juta kasus pada tahun 2000 dan 9,24 juta kasus pada tahun 2006, 9,3 juta kasus pada tahun 2007, 9,4 juta kasus pada tahun 2008, 9,4 juta kasus pada tahun 2009, dan 8,7 juta kasus pada tahun 2011. (WHO, 2009a; WHO, 2009b; WHO, 2010; WHO,2014). Diantara 9 juta kasus baru TB pada tahun 2013, terdapat sekitar 1,1 juta (13%) kasus HIV positif (WHO, 2014).

TB memerlukan penegakan diagnosis yang akurat. Ketidak akuratan diagnosis menyebabkan pengobatan yang diberikan pada pasien menjadi tidak tepat. Kasus resistensi terhadap obat TB (MDR TB) yang sampai sekarang semakin meningkat menjadi 60.000 kasus pada tahun 2011 (WHO, 2012). *Joint External Monitoring Mission* (JEMM) tahun 2007 melaporkan sekitar 10 hingga 30% pasien TB di rumah sakit DOTS tidak melaksanakan proses diagnosis dan sputum secara lengkap (Depkes & Stop TB partnership, 2007). Probandari dkk (2010) juga telah melaporkan hasil penelitian pada 61 rumah sakit di Jawa, sekitar 13-53% pasien TB paru dewasa tidak mendapatkan tata kelola diagnosis sesuai standar.

Model-model yang dapat dipergunakan untuk klasifikasi dan regresi telah sukses diaplikasikan di berbagai bidang diantaranya dengan regresi logistik, naïve bayes, neural network, dan Decision Tree dan lain-lain (Whitten dk, 2011; Santosa, 2007). Belum banyak penelitian tentang prediksi TB. Penelitian Santos, et al. (2004) telah menggunakan model neural network untuk diagnosis TB paru tetapi akurasinya belum begitu memuaskan. Peneliti lain Permanasari dkk. (2009) telah memprediksi insiden TB menggunakan univariate (satu variabel) data. Purwanto dkk. (2010) memenggunakan model hybrid pada data morbidity tuberculosis.

Hasil penelitian yang dilakukan oleh Wardani & Purwanto (2014), diperoleh model diagnosis TB didasarkan pada sistem penatalaksanaan TB sesuai dengan strategi DOTS menggunakan variabel karakteristik

pasien, anamnesis, pemeriksaan fisik, pemeriksaan laboratorium dan pemeriksaan radiologi. Hasil eksperimen menunjukkan bahwa model regresi logistik diperoleh akurasi 50%, sedangkan jika menggunakan model regresi logistik berbasis backward elimination menghasilkan akurasi sebesar 72,22%.

Berdasarkan hal di atas, perlu dikembangkan model prediksi untuk diagnosis TB dengan menggali variabel lain yang meliputi data yang diperoleh dari anamnesis, hasil pemeriksaan fisik, hasil pemeriksaan laboratorium, dan data pemeriksaan radiologi, termasuk lama batuk dan warna dahak.

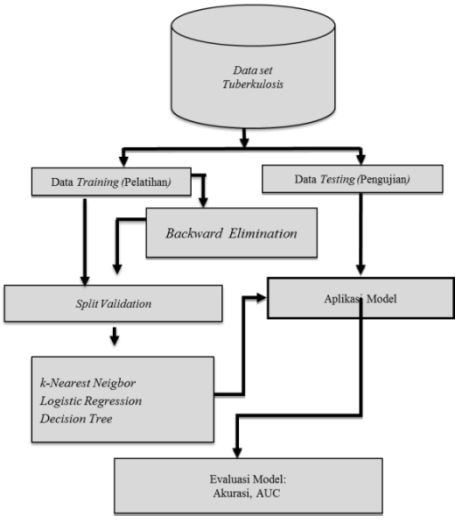
2. METODE PENELITIAN

Metode Pengumpulan Data

Data TB dalam penelitian ini diperoleh dari BKPM Semarang, yaitu data TB Paru dewasa tahun 2013 s/d 2014. Data primer diperoleh dengan wawancara tenaga medis dan pengelola TB untuk verifikasi dan validasi data. Data sekunder diperoleh dari data rekam medis terdiri dari 15 variabel pada tahun I yaitu data anamnesis (jenis kelamin, umur, batuk (ya/tidak), berdahak (ya/tidak), darah (ya/tidak), sesak nafas (ya/tidak), berat badan (menurun/ tetap/ naik), panas badan (ya/tidak), nafsu makan(ya/tidak), dan batuk anggota keluarga (ya/tidak), data pemeriksaan fisik (denyut nadi dan tekanan darah) dan data pemeriksaan radiologi (thorax) dan data diagnosis TB, sedangkan variabel pada tahun ke II ditambahkan 2 variabel yaitu lama batuk dan warna dahak.

Desain eksperimen

Desain eksperimen dalam penelitian ini digambarkan seperti terlihat di Gambar 1.



Gambar 1. Diagram Alir Eksperimen

Eksperimen dengan Model Klasifikasi

Model klasifikasi yang digunakan dalam penelitian ini meliputi *logistic regression*, *k-NN*, dan *Decision Tree*. Peneliti akan menggunakan model ini secara individual dan dengan menkombinasi model-model klasifikasi ini dengan teknik *backward elimination dan forward selection*.

Evaluasi Hasil Eksperimen

Untuk mengukur kinerja dari model klasifikasi, kinerja hasil experimen dilakukan dengan menggunakan teknik *confusion matrix*, yaitu nilai akurasinya dan AUC. Rumus-rumus yang digunakan untuk menghitung nilai *accuracy*, *precision*, dan *recall* adalah sebagai berikut (Gorunescu, 2011)

		PREDICTED CLASS	
		Class= Yes/Positive	Class= No/ Negative
OBSERVED CLASS	Class= Yes/Positive	a (TP=True Positive)	b (FN=False Negative)
	Class= No/ Negative	c (FP=False Positive)	d (TN=True Negative)

$$Accuracy = \left(\frac{a+d}{a+b+c+d}\right) = \frac{TP+TN}{TP+TN+FP+FN}$$

Alur penelitian

Pengolahan awal data

Pengolahan awal data dilakukan untuk pembersihan data TB, meliputi menghilangkan semua nilai *missing value* dan *outlier data* sebelum dilakukan analisis data lebih lanjut. Selanjutnya dilakukan pembagian data untuk digunakan dalam proses pembelajaran *(training)* dan pengujian *(testing)*.

Model Prediksi

Pada bagian ini, model yang diusulkan dipergunakan untuk diagnosis TB. Pengaturan dan pemilihan nilai dari parameter-parameter dan arsitektur melalui eksperimen-ekperimen dilakukan untuk memperoleh konfigurasi model yang sesuai, sehingga diperoleh akurasi diagnosis TB yang baik.

Seleksi Variabel

Metode ini digunakan untuk memilih variabel yang sesuai dalam pengembangan model. Pada penelitian ini teknik seleksi variabel yang dipakai adalah *backward ellimination*.

Model yang Digunakan

Selanjutnya dilakukan evaluasi model. Dalam penelitian ini model yang dipergunakan adalah regresi logistik, *k-Nearest Neighbor*, dan *Decision Tree.* Penelitian ini juga mengkombinasi model logistik, *k-Nearest Neighbor*, dan *Decision Tree* dengan metode *backward elimination.*

Evaluasi Model

Evaluasi pada model yang telah dikembangkan dilaksanakan untuk memberikan justifikasi dari model yang diusulkan.

3. HASIL DAN PEMBAHASAN

Pengumpulan data

Data set pada penelitian ini dikumpulkan dari data rekam medis pasien TB di BKPM wilayah Semarang tahun 2014 sebesar 461 pasien. Selanjutnya data tersebut dilakukan pengolahan awal data dengan tujuan untuk menghilangkan *missing* data maupun *outlier* data sehingga data siap untuk digunakan sebagai dasar pembangunan model prediksi. Setelah dilakukan *pre-processing* data, terdapat 5 buah data yang missing value/outlier data. Sehingga data yang digunakan dalam mengembangkan model prediksi sebanyak 456 buah. Variabel yang digunakan untuk prediksi terdiri dari karakteristik pasien, anamnesis, pemeriksaan tanda vital dan pemeriksaan radiologi sesuai dengan standar penatalaksanaan TB yang ditetapkan serta lama batuk dan warna dahak.

Hasil Analisis

Pada penelitian ini telah dilakukan analisis data dengan menggunakan tiga buah model yaitu regresi logistik, *k-Nearest Neighbor* dan *Decision Tree* (C4.5). Kinerja model diukur dengan nilai akurasi dan nilai AUC. Analisis dilakukan dalam dua tahap yaitu tahap pertama melakukan analisis dengan individual model dan tahap kedua dengan menggabungkan ketiga model tersebut dengan metode seleksi variabel *Backward elimination*. Peneliti melakukan perbandingan hasil kinerja semua model individu dan kombinasi metode seleksi variabel dengan ketiga model. Disamping itu, peneliti juga membandingkan dengan metode seleksi lain yaitu *forward selection*. Berikut adalah hasil-hasil yang diperoleh dari eksperimen-ekperimen yang telah dilakukan.

Individual Model

Pada model individu, peneliti menggunakan model regresi logistik, *k-Nearest Neighbor*, dan *Decision Tree.* Pada eksperimen pengembangan model, data yang telah dikumpulkan dibagi menjadi dua bagian, yaitu 90% untuk training dan 10 % untuk testing. Berikut ini adalah hasil akurasi dan nilai AUC dari ketiga model tersebut:

Model *k-Nearest Neighbour*

Pada model KNN, peneliti melakukan eksperimen dengan nilai *k* pada KNN bervariasi dari 1 sampai dengan 19. Hal ini dilakukan untuk menentukan nilai *k* yang terbaik. Berikut adalah hasil akurasi dan nilai AUC yang diperoleh.

Tabel 1. Kinerja model menggunakan K-Nearest Neighbour

k	*accuracy*	*precision*	*recall*	AUC
1	52,17%	62,50%	46,67%	0,500
3	52,17%	56,25%	50,00%	0,493
5	45,64%	43,75%	46,67%	0,442
7	54,35%	43,75%	60,00%	0,394
9	50,00%	37,50%	56,67%	0,492
11	**54,35%**	**43,75%**	**60,00%**	**0,534**
13	47,83%	43,75%	50,00%	0,533
15	45,65%	37,50%	50,00%	0,495
17	52,17%	43,75%	56,67%	0,488
19	52,27%	50,00%	53,35%	0,526

Kinerja terbaik diperoleh pada k =11, dengan akurasi sebesar 54,35%, presisi 43,75%, recall 60,00% dan AUC 0,534

Model *Decision Tree (C4.5)*

Model *Decision Tree (C4.5)* diimplementasikan untuk diagnosis tuberculosis. Peneliti mencari model C4.5 terbaik dengan menggunakan bermacam-macam criterion, yaitu gain ratio, information gain dan gini index. Berikut adalah hasil akurasi dan nilai AUC yang diperoleh.

Tabel 2. Kinerja model menggunakan Decision Tree

Kriteria	accuracy	precision	recall	AUC
Gain Rasio	34,78%	100%	0,00%	0,500
Information Gain	24,78%	100%	0,00%	0,500
Gini index	50,00%	56,25%	46,67 %	0,573

Kinerja terbaik diperoleh pada kriteria Gini Index, dengan akurasi sebesar 50,00%, presisi 56,25%, recall 46,67% dan AUC 0,573

Model Regresi Logistik

Model *Regresi Logistik* juga diterapkan untuk diagnosis tuberculosis, hasil kinerja model *accuracy* 34,78%, *precision* 100%, *recall* 0,00% dan AUC : 0,00

Adapun bobot-bobot dari regresi logistik adalah:

Bias/ konstanta: -0.049
w[jk = 1.0] = 0.049
w[jk = 0.0] = -0.039
w[batuk = 0.0] = -0.005
w[batuk = 1.0] = 0.129
w[Berdahak = 0.0] = -0.050
w[Berdahak = 1.0] = 0.180
w[warna = 2.0] = -0.160
w[warna = 3.0] = -0.058
w[warna = 1.0] = 0.229
w[warna = 0.0] = 0.017
w[darah = 1.0] = 0.000
w[darah = 0.0] = -0.002
w[sesak = 1.0] = -0.308
w[sesak = 0.0] = 0.170
w[panas = 1.0] = 0.028
w[panas = 0.0] = -0.026
w[naf_makan = 0.0] = 0.104
w[naf_makan = 1.0] = -0.078
w[bb = 1.0] = 0.028
w[bb = 0.0] = -0.026
w[batuk_angg = 1.0] = -0.020
w[batuk_angg = 0.0] = 0.079
w[thorax = 1.0] = 0.006
w[thorax = 2.0] = -0.045
w[thorax = 0.0] = -0.049
w[umur] = -0.033
w[lama batuk(minggu)] = -0.116
w[denyut_nadi] = -0.337
w[sistole] = 0.282
w[diastole] = 0.098

Model Klasifikasi berbasis *Bacward Ellimination*

Untuk meningkatkan kinerja model, peneliti mengembangkan model klasifikasi berbasis Backward Ellimination untuk prediksi diagnosis tuberculosis. Peneliti menggabungkan metode seleksi variabel/ atribut yaitu Backward Ellimination dengan tiga model klasifikasi kNN, regresi logistic dan decision tree (C4.5).

Model *k-Nearest Neighbor* berbasis *Bacward Ellimination*

Model KNN berbasis *Bacward Ellimination* merupakan model pengembangan kombinasi metode *Bacward Ellimination* untuk menyeleksi variabel yang sesuai dengan metode klasifikasi KNN. Peneliti melakukan eksperimen dengan nilai *k* pada KNN berbeda-beda dari 1 sampai dengan 19 untuk menentukan nilai *k* yang terbaik. Berikut adalah hasil akurasi dan nilai AUC yang diperoleh.

Tabel 3. Kinerja model menggunakan K-Nearest Neighbour berbasis Backward Ellimination

k	*accuracy*	*precision*	*recall*	AUC
1	65,22%	62,50%	68,18%	0,500
3	73,91%	72,00%	76,19%	0,676
5	**78,26%**	**72,73%**	**83,33%**	**0,807**
7	71,74%	68,18%	75,00%	0,795
9	67,39%	68,18%	66,67%	0,759
11	73,91%	72,73%	75,00%	0,802
13	67,39%	72,73%	62,50%	0,762
15	69,57%	68,18%	70,83%	0,789
17	69,57%	63,64%	75,00%	0,821
19	73,91%	81,82%	66,67%	0,830

Dari hasil-hasil tersebut di atas, maka nilai k pada K-NN terbaik adalah k=5. Hal ini ditunjukkan dari nilai akurasi tertinggi sebesar 78.26 %. Sedangkan atribut/ variabel yang terseleksi terlihat dalam tabel berikut:

Tabel 4. Variabel dan bobot menggunakan KNN+Backward Ellimination

Atribut/ Variabel	Bobot
Jk	1.0
Umur	1.0
Batuk	1.0
lama batuk(minggu)	1.0
Berdahak	1.0
Warna	1.0
Darah	1.0
Sesak	1.0
Panas	1.0
naf_makan	1.0
Bb	1.0
batuk_angg	1.0
denyut_nadi	1.0
Systole	0.0
diastole	1.0
Thorax	1.0

Model Decision Tree berbasis *Bacward Ellimination*

Pengembangan model *Decision Tree* berbasis backward elimination diterapkan dengan berbagai criterion, yaitu gain ratio, information gain dan gini index. Hasil kinerja akurasi dan nilai AUC yang diperoleh disajikan berikut ini. Decision Tree berbasis *Backward Ellimination* menggunakan gain ratio dan information gain menghasilkan nilai kinerja yang sama. Hasil kinerja dari model ini adalah sebagai berikut:

Tabel 5. Kinerja model menggunakan Decision Tree berbasis Backward Ellimination

Kriteria	accuracy	precision	Recall	AUC
GR IG	52,17%	0,00%	100%	0,500
Gini index	65,22%	58,82%	68,97%	0,533

Ket: GR=gain Rasio
IG=information gain

Dari hasil kinerja model Decision Tree berbasis Bakcward Ellimination di atas, terlihat bahwa *criterion* dengan menggunakan gini index memberikan hasil terbaik yaitu nilai accuracy sebesar 65.22% dan nilai AUC sebesar 0.533.

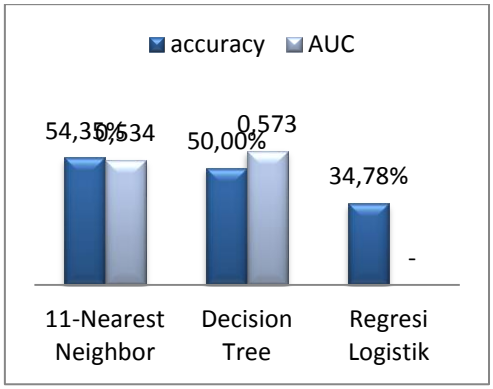
Model Regresi Logistik berbasis *Bacward Ellimination*

Hasil kinerja model Regresi Logistik berbasis Bacward Ellimination untuk prediksi diagnosis tuberculosis, hasil kinerja model *accuracy* 60,87, *precision* 100%, *recall* 0,00% dan AUC : 0,00

Perbandingan Model untuk Prediksi Diagnosis Tuberculosis

Perbandingan Model untuk Prediksi Diagnosis Tuberculosis Individu

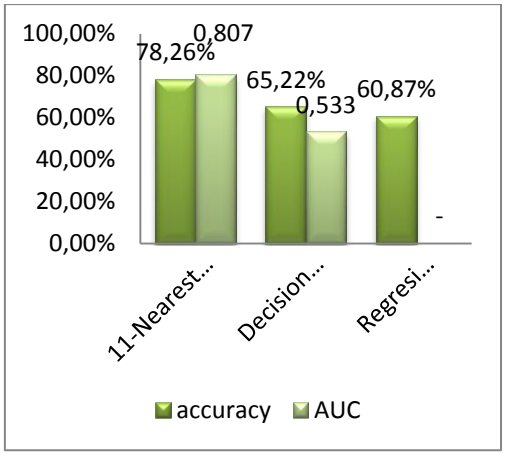
Kinerja model-model klasifikasi regresi logistik, KNN dan Decision Tree untuk prediksi diagnosis tuberculosis telah didapatkan dengan bervariasi. Untuk model klasifikasi individu yang terbaik adalah KNN dengan nilai accuracy 54.35% dan AUC 0.534. Sementara untuk model regresi logistik, kinerja akurasinya 34.78% dan AUC=0.0 Serta untuk model decision tree akurasinya 50.00% dan AUC sebesar 0.573. Nilai AUC decision tree lebih tinggi daripada nilai AUC pada model KNN, tetapi jika diklasifikasikan kedua nilai AUC tersebut menurut Gurunescu (2011) masih dalam dalam kelompok yang sama.
Hal ini menunjukkan bahwa model KNN memberikan hasil yang terbaik untuk model individu (gambar 2).



Gambar 2. Perbandingan kinerja model individu

Perbandingan Model untuk Diagnosis Tuberculosis berbasis backward ellimination

Perbandingan kinerja model berbasis backward ellimination ditunjukkan pada gambar 3



Gambar 3. Perbandingan Kinerja Model Berbasis *Backward Ellimination*

Hasilnya diperoleh k-NN berbasis *backward ellimination* memberikan hasil kinerja yang terbaik yaitu akurasi sebesar 78.66% dan AUC sebesar 0.806 (*good classification*). Sementara regresi logistik berbasis backward elimination (Akurasi: 60.87%; AUC: 0.0) dan decision tree berbasis backward elimination (akurasi: 65.22%; AUC: 0.533). KNN berbasis backward ellimination juga memberikan hasil kinerja yang terbaik jika dibandingkan dengan KNN berbasis forward selection. Model KNN (k=5) berbasis forward selection memberikan kinerja akurasi 67.39% dan AUC: 0.603.

4. KESIMPULAN

Hasil eksperimen menunjukkan bahwa model k-NN berbasis *backward elimination* memberikan kinerja yang terbaik jika dibandingkan dengan model-model lain. Model k-NN berbasis *backward elimination* memberikan nilai akurasi sebesar 78.66% dan nilai AUC sebesar 0.806 yang mengindikasikan bahwa model adalah *good classification*.

5. REFERENSI

Depkes & Stop TB Partnership.(2007). Report of the joint External TB Monitoring Mission Indonesia (16-27 April 2007).Jakarta: Depkes

Depkes.(2011). Pedoman Nasional Penanggulangan Tuberkulosis. Jakarta: Depkes

Dinkes Prop. Jateng (2012). Potret BKPM 2012. Semarang : BKPM

Gorunescu, F. 2011. Data Mining: Concepts, Models and Techniques, Intelligent Systems. Reference Library, 12, 319–330

Permanasari, A. E., Rambli, D. R. A., Dominic, D.D., 2009. A Comparative Study of Unvariate Forecasting Methods for Predicting Tuberculosis Incidence on Human, Proceeding of 2009 Student Conference on Research and Development (SCORed 2009), Malaysia, pp 188-191

Probandari, A., Utarini, A. dan Karin, A.H., (2010), Missed Opportunity for Standardized Diagnosis and Treatment among Adult Tuberculosis Patient in hospital Involved in Public Private Mix for Directly Observed Treatment Short Cource Strategy in Indonesia : a cross sectional study. BMC Health Service Research.

Purwanto, Eswaran, C., dan Logeswaran, R. (2010). An Adaptive Hybrid Algorithm for Time Series Prediction in

Healthcare. Proceedings of the International Conference on Computational Intelligence, Modelling and Simulation (CIMSiM 2010), Bali, Indonesia (IEEE), 21-26. ISBN: 978-1-4244-8652-6

Santosa, B.(2007).Data Mining, Teknik Pemanfaatan Data untuk Keperluan Bisnis.Yogyakarta: Graha Ilmu

Wardani, R. S., Purwanto.(2014). Model Pengambilan Keputusan dalam Prediksi Kasus Tuberkulosis Menggunakan Regresi Logistik Berbasis Backward Elimination, Prosiding Seminar Nasional Hasil-hasil Penelitian dan Pengabdian, ISSN: 978.979.704

Witten, I. H, Eibe, F dan Mark, H.A.(2011). Data Mining: Practical Machine Learning Tools and Techniques 3rd Edition, Elsevier

WHO, (2009a). Global Tuberculosis Control : A Short Update to the 2009 Report, Geneva, Switzerland: WHO Press.

WHO, (2009b). Global Tuberculosis Control : Epidemiology, Strategy, Financing : WHO report 2009, Switzerland: WHO Press.

WHO, (2010). Global Tuberculosis Control : WHO Repots 2010, Switzerland: WHO Press.

WHO, (2014). Global Tuberculosis Control : WHO Repots 2014, Switzerland: WHO Press.

WHO, (2012).Global Tuberculosis Control Report. Swizerland: WHO Press.