Jurnal Riset Komputer (JURIKOM), Vol. 5 No. 5, Oktober 2018
ISSN 2407-389X (Media Cetak)
Hal: 498-502

Aplikasi Diagnosa Penyakit Tuberculosis Menggunakan Algoritma Naive Bayes

Amrin, Hafdiarsya Saiyar

Universitas Bina Sarana Informatika Jakarta, Indonesia
Email: amrin.ain@bsi.ac.id, hafdiarsya.hyr@bsi.ac.id

Abstrak

Sangat penting bagi dokter untuk melakukan diagnosa secara dini penyakit *tuberculosis* agar dapat mengurangi penularan penyakit tersebut kepada masyarakat luas. Pada penelitian ini, penulis akan menerapkan metode klasifikasi data mining, yaitu metode *Naïve Bayes* untuk mendiagnosa penyakit *tuberculosis*. Berdasarkan hasil pengukuran performa dari model tersebut dengan menggunakan metode pengujian *Cross Validation*, *Confusion Matrix* dan Kurva ROC, diketahui bahwa metode *Naïve Bayes* dengan tingkat akurasi sebesar 94,18% dan nilai area under the curva (AUC) sebesar 0,977. Hal ini menunjukkan model yang dihasilkan termasuk katagori klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

Kata Kunci: Metode N*aive Bayes*, C*onfusion Matrix*, Kurva ROC

Abstract

It is important for doctors to make an early diagnosis of tuberculosis in order to reduce the transmission of the disease to the wider community. In this study, the authors will apply methods of data mining classification, Naïve Bayes to diagnose tuberculosis disease. Based on the performance measurement results of the models using Cross Validation, Confusion Matrix and ROC Curve methods, it is known that Naïve Bayes method with accuracy of 94.18% and under the curva (AUC) value of 0.97. This shows that the models that are produced including the category of classification is very good because it has an AUC value between 0.90-1.00.

Keywords: Naive Bayes, Confusion Matrix, ROC Curva

1. PENDAHULUAN

Tuberculosis yang disingkat TBC atau TB adalah penyakit menular yang disebabkan oleh bakteri *Mycobacterium Tuberculosis* yang ditularkan melalui udara (*dropet nuclei*) saat seorang pasien TBC batuk dan percikan ludah yang mengandung bakteri tersebut terhirup oleh orang lain saat bernapas (Widoyono, 2011). Penyakit TB (Orhan dan Tanrikulu, 2010) adalah penyakit menular yang disebabkan oleh bakteri yang disebut *Mycobacterium tuberculosis* dan merupakan penyebab kematian paling tinggi yang terjadi pada usia produktif 15-50 tahun, kelompok ekonomi lemah, dan berpendidikan rendah. Penyakit ini dapat menular sehingga perlu penanganan yang intensif, setidaknya diperlukan pengobatan minimal 6 bulan secara rutin dan terus menerus. Sedangkan Indonesia menempati peringkat ke-2 di dunia setelah india dengan pasien TBC terbanyak dan diperkirakan ada 1.020.000 kasus TB di Indonesi (Kemenkes, 2018).

Penularan *tuberkulosi*s (TBC) sangat cepat melalui udara. Bagi penderita diharapkan selalu melakukan pemeriksaan dan pengobatan sampai tuntas. TBC ditularkan melalui udara. Percikan ludah atau dahak yang dikeluarkan menjadi media penularan yang sangat cepat di dunia ini. Penularan TBC melalui udara akan sangat rentan terjadi di ruang publik. Dari berbagai penelitian akan ada puluhan ribu kuman yang keluar dari batuk dan bersin. Oleh karenanya diharapkan masyarakat untuk menggunakan masker di tempat-tempat umum dan senantiasa berperilaku hidup bersih dan sehat (Kemenkes, 2018).

Klasifikasi data penyakit TB pada medis merupakan tugas penting dalam memprediksi penyakit, bahkan dapat membantu dokter dalam mengambil keputusan diagnosis penyakit tersebut (Fine, 2012), dengan demikian sangat penting melakukan diagnosa secara dini agar dapat mengurangi penularan TB kepada masyarakat luas. Pada penelitian ini, penulis akan menerapkan metode klasifikasi data mining, yaitu Naïve Bayes untuk mendiagnosa penyakit *Tuberculosis*. Data yang penulis gunakan adalah data pasien puskesmas bojonggede yang terdiagnosa *tuberculosis*.

Menurut Han dan Kamber dalam (Amrin, 2016) *Data mining* adalah rangkaian proses untuk menggali nilai tambah berupa informasi yang belum terekplorasi dari sebuah basis data, melakukan ekplorasi dengan cara-cara tertentu untuk memanipulasi data menjadi informasi yang lebih berharga dengan cara mengektraksi dan mengenali pola penting dari basis data. Menurut Daryl Pregibons dalam (Gorunescu, 2011) "*Data mining* adalah perpaduan dari ilmu statistik, kecerdasan buatan, dan penelitian bidang *database*". Nama *data mining* berasal dari kemiripan antara pencarian informasi yang bernilai dari *database* yang besar dengan menambang sebuah gunung untuk sesuatu yang bernilai (Sumathi, 2006). Keduanya memerlukan penyaringan melalui sejumlah besar material, atau menyelidiki dengan cerdas untuk mencari keberadaan sesuatu yang disebut bernilai tadi.

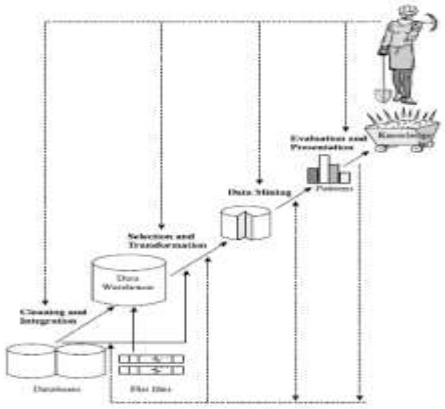
Data Mining merupakan teknologi baru yang sangat berguna untuk membantu perusahaan-perusahaan menemukan informasi yang sangat penting dari gudang data mereka. Beberapa aplikasi data mining fokus

pada prediksi, mereka meramalkan apa yang akan terjadi dalam situasi baru dari data yang menggambarkan apa yang terjadi di masa lalu (Witten, 2011).

2. TEORITIS

2.1 Tahapan Proses Data Mining

Data mining sering disebut juga Knowledge Discovery in Database atau disingkat menjadi KDD, adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar (Santosa, 2007). Gambar tahapan pembuatan aplikasi data mining ditunjukan pada gambar 1 berikut ini:



Gambar 1. Tahapan Proses KDD
Sumber: Han & Kamber (2006)

Gambar 1 menunjukkan langkah dalam proses *data mining.* Proses dalam tahap *data mining* terdiri dari tiga langkah utama, yaitu (Sogala, 2006):

1. *Data Preparation*
   Pada langkah ini, data dipilih, dibersihkan, dan dilakukan *preprocessed* mengikuti pedoman dan *knowledge* dari ahli domain yang menangkap dan mengintegrasikan data internal dan eksternal ke dalam tinjauan organisasi secara menyeluruh.
2. Algoritma *data mining*
   Penggunaan algoritma *data mining* dilakukan pada langkah ini untuk menggali data yang terintegrasi untuk memudahkan identifikasi informasi bernilai.
3. Fase analisa data
   Keluaran dari data mining dievaluasi untuk melihat apakah *knowledge* domain ditemukan dalam bentuk *rule* yang telah diekstrak dari jaringan.

2.2 Naive Bayes

Klasifikasi Bayes (Kusrini, 2009) adalah pengklasifikasian statistik yang dapat digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Klasifikasi Bayes juga dikenal dengan *Naïve Bayes*.

$$P(x|y) = \frac{P(y|x)\,P(x)}{P(y)} \qquad (1)$$

keterangan :
y = data dengan kelas yang belum diketahui
x = hipotesis data y merupakan suatu kelas spesifik
P(x | y) = probabilitas hipotesis x berdasar kondisi y (*posteriori probability*)
P(x) = probabilitas hipotesis x (*prior probability*)
P(y | x) = probabilitas y berdasarkan kondisi pada hipotesis x
P(y) = probabilitas dari y

2.3 Evaluasi dan Validasi Model

Untuk mengukur akurasi model maka dilakukan evaluasi dan validasi menggunakan teknik:

1. *Confusion matrix*

Confusion Matrix adalah alat (*tools*) visualisasi yang biasa digunakan pada supervised learning. Tiap kolom pada matriks adalah contoh kelas prediksi, sedangkan tiap baris mewakili kejadian di kelas yang sebenarnya (Gorunescu, 2011). Confusion matrix berisi informasi aktual (actual) dan prediksi predicted) pada sisitem klasifikasi.

2. Kurva ROC (*Reciever Operating Characteristic*)

Kurva ROC menunjukkan akurasi dan membandingkan klasifikasi secara visual. ROC mengekspresikan *confusion matrix*. ROC adalah grafik dua dimensi dengan *false positives* sebagai garis horisontal dan *true positives* sebagai garis vertikal (Vecellis, 2009). *The area under curve* (AUC) dihitung untuk mengukur perbedaan performansi metode yang digunakan. AUC dihitung menggunakan rumus: (Liao, 2007)

$$\theta^r = \frac{1}{mn}\sum_{j=1}^{n}\sum_{i=1}^{m}\psi\,(xi^r, xj^r) \qquad (2)$$

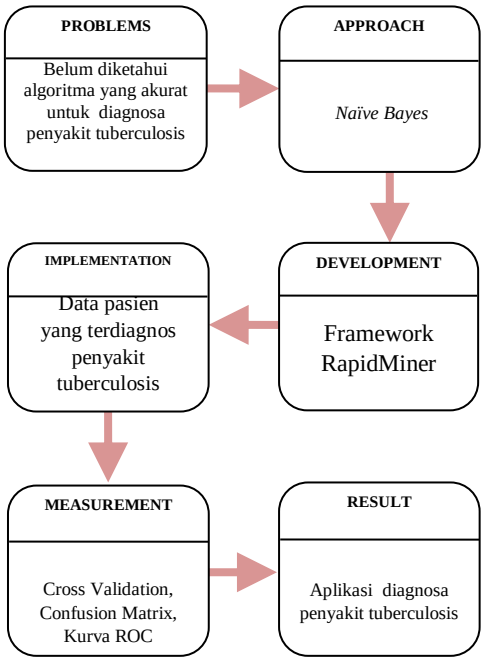
Dimana

$$\boldsymbol{\psi}(X,Y) = \begin{cases} 1 & Y < X \\ \frac{1}{2} & Y = X \\ 0 & Y > X \end{cases}$$

Performance keakurasian AUC dapat diklasifikasikan menjadi lima kelompok yaitu (Gorunescu, 2011):
0.90 – 1.00 = *Exellent Clasification*
0.80 – 0.90 = *Good Clasification*
0.70 – 0.80 = *Fair Clasification*
0.60 – 0.70 = *Poor Clasification*
0.50– 0.60 = *Failure*

3. ANALISA DAN PEMBAHASAN

Penelitian ini terdiri dari beberapa tahap seperti terlihat pada kerangka pemikiran Gambar 2 Permasalahan (*problem*) pada penelitian ini adalah Belum diketahui algoritma yang akurat untuk diagnosa penyakit tuberculosis. Untuk itu dibuat *approach* (model) *Naive Bayes* untuk memecahkan permasalahan kemudian dilakukan pengujian terhadap kinerja dari ketiga metode tersebut. Pengujian menggunakan metode *Cross Validation*, *Confusion Matrix* dan kurva ROC. Untuk mengembangkan aplikasi (*development*) berdasarkan model yang dibuat, digunakan Rapid Miner.



Gambar 2. Kerangka Pemikiran Pemecahan Masalah

Pada penelitian ini data yang digunakan sebanyak 136 data pasien tuberculosis (TBC) baik yang positif maupun yang negatif. Variabel input pada penelitian ini terdiri dari enam variabel, yaitu: 1. Keringat pada malam hari tanpa aktivitas fisik, 2. Berat badan turun, 3. Nafsu makan berkurang 4. Mudah lelah dan lemah, 5. Demam, 6. Batuk berdahak lebih dari tiga minggu disertai batuk darah, Sedangkan variabel output adalah variabel penyakit TBC. Perangkat lunak yang digunakan untuk menganalisa adalah RapidMiner *versi 5.3*.

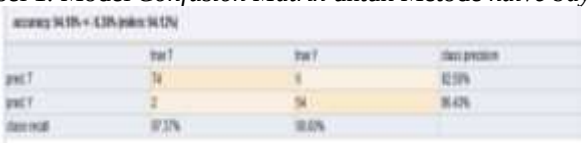
3.1 Pengujian Model

Model yang telah dibentuk diuji tingkat akurasinya dengan memasukan data uji yang berasal dari data *training*. Karena data yang didapat dalam penelitian ini setelah proses *preprocessing* hanya 136 data maka digunakan metode *cross validation* ,*Confusion Matrix*, dan Kurva ROC untuk menguji tingkat akurasi. Untuk nilai akurasi metode *naïve bayes* sebesar 94.18%.

1. Confusion Matrix

Tabel 1 adalah *confusion matrix* untuk metode *naïve bayes*. diketahui dari 136 data, 74 diklasifikasikan tidak (negatif) sesuai dengan prediksi yang dilakukan dengan metode *naïve bayes*, lalu 6 data diprediksi tidak (nagatif) tetapi ternyata ya (positif), 54 data *class* ya (positif) diprediksi sesuai, dan 2 data diprediksi ya (positif) ternyata tidak (negatif).

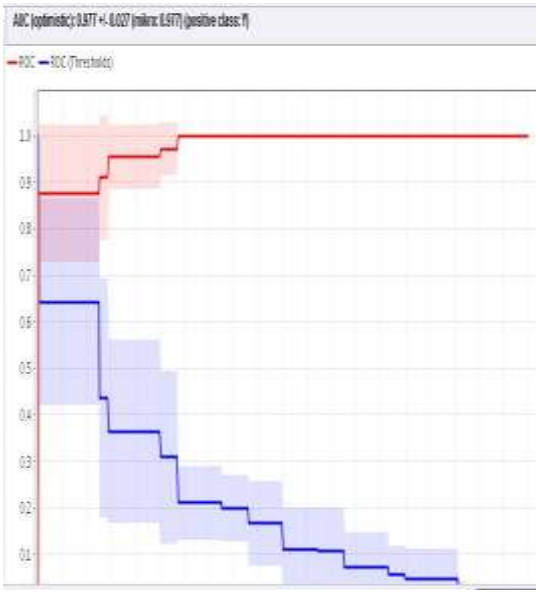
Tabel 1. Model *Confusion Matrix* untuk Metode *naïve bayes*



Sumber: Hasil Pengolahan Menggunakan RapidMiner 5.3 (2018)

2. Kurva ROC

Hasil perhitungan divisualisasikan dengan kurva ROC. Kurva ROC pada gambar 3 mengekspresikan *confusion matrix*. Garis horizontal adalah *false positives* dan garis vertikal *true positives*.



Gambar 3 Kurva ROC dengan Metode *Naïve Bayes*
Sumber: Hasil Pengolahan Menggunakan RapidMiner 5.3 (2018)

Dari gambar di atas terlihat bahwa nilai *area under curve* (AUC) metode naïve bayes adalah 0.977. Untuk klasifikasi *data mining*, nilai AUC dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

a. 0.90-1.00 = klasifikasi sangat baik
b. 0.80-0.90 = klasifikasi baik
c. 0.70-0.80 = klasifikasi cukup
d. 0.60-0.70 = klasifikasi buruk

e. 0.50-0.60 = klasifikasi salah

Berdasarkan pengelompokkan di atas maka dapat disimpukan bahwa metode *naïve bayes* termasuk klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

3.2 Rancangan Aplikasi Data Mining

Berdasarkan metode tersebut kemudian dirancang dan dibuatlah aplikasi diagnosa penyakit *tuberculosis* menggunakan algoritma data mining, dalam hal ini *naïve bayes*, maka implementasi hasil rancangan aplikasi seperti terlihat pada gambar 4 berikut ini:

Gambar 4. GUI Sistem Prediksi Diagnosa Penyakit TBC

4. KESIMPULAN

Kesimpulan yang dapat diambil berdasarkan penelitian ini adalah bahwa performa model *naïve bayes* memberikan tingkat akurasi kebenaran sebesar 94,18% dengan nilai area under the curve (AUC) sebesar 0,977. Hal ini menunjukkan bahwa model tersebut termasuk katagori klasifikasi sangat baik karena memiliki nilai AUC antara 0.90-1.00.

REFERENCES

Amrin, A. (2016). Data Mining Dengan Regresi Linier Berganda Untuk Peramalan Tingkat Inflasi. *Jurnal Techno Nusa Mandiri*, *XIII*(1), 74–79. Retrieved from http://ejournal.nusamandiri.ac.id/ejurnal/index.php/techno/article/view/268

Fine, J. (2012). *An Overview Of Statistical Methods in Diagnostic Medicine.* Chapel Hill.

Gorunescu, F. (2011). *Data Mining: Concepts, Models, and Techniques.* Verlag Berlin Heidelberg: Springer.

Han, J., & Kamber, M. (2006). *Data Mining Concept and Tehniques.* San Fransisco: Morgan Kauffman.

Kusrini, &. E. (2009). *Algoritma Data Mining* . Yogyakarta: Andi Publishing.

Liao. (2007). *Recent Advances in Data Mining of Enterprise Data: Algorithms and Application.* Singapore: World Scientific Publishing.

Orhan, E., Temurtas, F., & Tanrıkulu, A. Ç. (2010). *Tuberculosis Disease Diagnosis Using Artificial Neural Networks*. Springer, 299-302.

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

Sogala, S. S. (2006). *Comparing the Efficacy of the Decision Trees with Logistic Regression for Credit Risk Analysis.* India.

Sumathi, &. S. (2006). *Introduction to Data Mining and its Applications.* Berlin Heidelberg New York: Springer.

Vercellis, C. (2009). *Business Intelligent: Data Mining and Optimization for Decision Making. Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.* Southern Gate, Chichester, West Sussex: John Willey & Sons, Ltd.: Vercellis, Carlo (2009). Business Intelligent: Data Mining and OpJohn Willey & Sons, Ltd.

Widoyono. (2011). *Penyakit Tropis Epidemiologi, Penularan, Pencegahan dan Pemberantasan.* Jakarta: Erlangga.

Witten, I. H. (2011). *Data Mining: Practical Machine Learning and Tools.* Burlington: Morgan Kaufmann Publisher.