

ALGORITMA KLASIFIKASI NAIVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK PREDIKSI PENYAKIT TUBERCULOSIS (TB)

Erika Mutiara

Universitas Bina Sarana Informatika
Jl. Keramat Raya No. 18 Ringroad Barat, Cengkareng, Jakarta Barat
e-mail: erika.emb@bsi.ac.id

ABSTRAK

Penyakit Tuberculosis (TB) merupakan penyakit infeksi paru yang yang berpotensi serius, paling menular dan berbahaya. Menurut World Health Organization (WHO) penyakit tuberculosis masuk kedalam salah satu dari sepuluh penyebab kematian diseluruh dunia. Lebih dari 10 juta orang terkena infeksi TB dan sekitar 60% terjadi dinegara berkembang seperti Indonesia. Sudah banyak penelitian yang membahas tentang prediksi penyakit tuberculosis. Pada penelitian kali ini dilakukan optimasi metode Naive Bayes dengan metode Particle Swarm Optimization sebagai seleksi atribut untuk meningkatkan akurasi prediksi yang diaplikasikan terhadap data pasien yang dinyatakan Positif Tuberculosis dan Negatif Tuberculosis. Setelah dilakukan pengujian dengan dua model yaitu metode naive bayes dan optimasi naive bayes menggunakan particle swarm optimization sehingga hasil yang didapat dengan menggunakan Naive Bayes diperoleh accucary sebesar 92,69% dengan nilai AUC 0,992 sedangkan pengujian dengan menggunakan optimasi Naive Bayes dengan Particle Swarm Optimization didapatkan nilai accucary sebesar 98,76% dengan nilai AUC 0,999 dengan tingkat diagnosa excellent classification karena hasil AUC-nya antara 0,90-1,00. Sehingga kedua metode tersebut memiliki perbedaan tingkat akurasi sebesar 6,07% dan perbedaan nilai AUC sebesar 0,007.

Keywords: Tuberculosis, Naive Bayes, Seleksi Atribut, Particle Swarm Optimization

ABSTRACT

Tuberculosis (TB) is a potentially serious, most contagious and dangerous lung infection. According to the World Health Organization (WHO) tuberculosis disease is one of the ten leading causes of death worldwide. More than 10 million people exposed to TB infected and around 60% occur in a country growing as Indonesia. There have been many studies that discuss the prediction of tuberculosis. In this study the optimization of the Naive Bayes method was carried out with the Particle Swarm Optimization method as attribute selection to improve predictions accuracy applied to patient data that was declared positive tuberculosis and negative tuberculosis. After testing with two models namely Naive Bayes Method and Naive Bayes Optimization using Particle Swarm Optimization, the results obtained by using the Naive Bayes obtained accucary of 92.69% with the value of AUC 0.992 while testing using Naive Bayes Optimization with Particle Swarm Optimization obtained an accucary value of 98.76% with the value of AUC 0.999 with excellent classification diagnosis because the AUC results are between 0.90-1.00. So that the two methods have difference levels of accuracy that is equal to 6.07% and the difference in AUC value of 0.007.

Keywords: Tuberculosis, Naive Bayes, Selection Attributes, Particle Swarm Optimization

1. Pendahuluan

Pencemaran udara merupakan salah satu peristiwa dimana dapat menyebabkan gangguan serius pada manusia. Pencemaran udara bisa disebabkan oleh asap kendaraan bermotor, asap industri pabrik, limbah pertanian, asap rokok, dan lainnya. Dampak dari pencemaran udara salah satunya dapat membahayakan kesehatan kita hingga menyebabkan kematian. Salah satu yang bisa kita lihat adalah kebiasaan merokok ditambah lagi pencemaran udara yang sangat tinggi dari polusi, itu semua dapat menimbulkan berbagai penyakit paru-paru, diantara penyakit paru yang paling menular dan berbahaya adalah Tuberculosis (TB) (Setyo & Sudrajat 2017). Tuberculosis adalah penyakit menular yang disebabkan oleh *bacillus* yang disebut *mycobacterium tuberculosis*, TB merupakan salah satu dari sepuluh penyebab kematian diseluruh dunia meskipun penyakit ini dapat disembuhkan dan dapat dicegah. Menurut *World Health Organization (WHO)* lebih dari 10 juta orang terkena infeksi TB sekitar 60% terjadi di enam negara termasuk India, Indonesia, Cina, Nigeria, Pakistan dan Afrika Selatan (Zulvia, Kuo & Roflin, 2017).

Tahun 2013 berdasarkan laporan WHO memperkirakan sudah terdapat 8,6 juta kasus TB dan pada tahun 2012 sudah diperkirakan terdapat 450.000 orang yang menderita TB dan 170.000 diantaranya meninggal dunia (Kementrian Kesehatan RI, 2018) dan disetiap tahunnya ada lebih dari 8 juta kasus orang yang tertular TB, dengan sekitar 2 juta orang pertahun meninggal akibat penyakit ini (Ekata, et al, 2016).

Banyak sebab yang menjadikan TB paru sebagai ancaman global dan sangat serius, mengingat bahwa apabila seorang penderita dengan alasan apapun tidak pergi berobat maka dalam waktu kurang lebih setahun akan menularkan ke 10 sampai 15 orang lain yang sehat. Penularannya pun sangat mudah sekali dapat melalui interaksi komunikasi antara penderita TB dengan lawan bicara bahkan dengan bersin pun mudah sekali menularkan ke orang lain yang sehat (Abdullah, 2017).

Dengan klasifikasi data penyakit TB pada medis dapat membantu dokter dalam mengambil keputusan diagnosis penyakit, sangatlah penting melakukan diagnosis secara dini agar dapat mengurangi penularan TB dan tingkat kematian kepada masyarakat luas (Saputra, 2014). Selain itu, deteksi dini

terhadap penyakit TB memungkinkan penderita penyakit ini dapat disembuhkan lebih cepat. Banyak peneliti yang sudah melakukan prediksi penyakit TB dengan metode klasifikasi data mining, namun belum diketahui metode apa saja yang paling akurat dalam memprediksi penyakit TB.

Sudah banyak peneliti yang melakukan penelitian untuk memprediksi penyakit TB seperti penelitian yang dilakukan oleh Rusdah et al (2015) tentang diagnosa awal tuberkulosis paru menggunakan metode *ensemble*, penelitian selanjutnya dilakukan oleh Ekata et al (2016) tentang diagnosa tuberkulosis paru menggunakan *fuzzy inference system* dan penelitian selanjutnya yang dilakukan oleh Yuangli et al (2017) tentang klasifikasi otomatis tuberkulosis paru dan sarcoidosis dengan membandingkan metode *random forest* dan algoritma klasifikasi data mining yang lainnya.

Metode Naive Bayes merupakan metode klasifikasi data mining yang dapat memecahkan masalah dari metode-metode sebelumnya hal ini merupakan alasan mengapa naive bayes merupakan salah satu metode yang tepat dipakai untuk memecahkan masalah dalam keterbatasan sampel data yang ada (Dewi, 2016). Tetapi ada kelemahan pada metode naive bayes yaitu lamanya waktu dan tingkat akurasi prediksi yang belum maksimal. Masalah ini tentunya menjadi perhatian banyak peneliti untuk dapat memperbaiki metode naive bayes agar kinerja waktu dan akurasi prediksi menjadi lebih singkat tetapi kinerja akurasi tetap baik (Rosandy, 2016). Kelemahan ini dapat dihilangkan dengan melakukan optimasi seperti yang dilakukan oleh Buani (2016) menggunakan algoritma genetika. Dalam penelitian ini, metode yang digunakan untuk memilih variabel yang relevan adalah algoritma yang dilandasi oleh kebiasaan alam yang disebut *Particle Swarm Optimization (PSO)*.

Berdasarkan uraian diatas untuk menangani kelemahan yang masih ada maka akan diterapkan metode *Naive Bayes* berbasis *feature selection Particle Swarm Optimization (PSO)* untuk mengetahui seberapa akurat *feature selection PSO* dalam meningkatkan akurasi *Naive Bayes* dalam memprediksi penyakit tuberkulosis.

2. Tinjauan Pustaka

2.1. Tuberculosis (TB)

Kata "*Tuberculosis*" diciptakan oleh Johann Lukas Schonle pada tahun 1839 dari bahasa latin "*tuberculum*" yang berarti benjolan kecil, pembengkakan atau jerawat (Ekata, et al, 2016). Tuberculosis (TB) disebabkan oleh suatu penyakit infeksi kuman menular yang mempunyai nama yang sama yaitu bakteri *Mycobacterium Tuberculosis*, kuman atau bekteri ini dapat menyerang berbagai organ terutama paru-paru, penyakit tuberculosis ini bila tidak diobati atau pengobatannya tidak tuntas atau tidak selesai maka dapat menimbulkan komplikasi berbahaya hingga kematian. Gejala utama yang sering terlihat sebagai tanda-tanda terkena penyakit TB antara lain (Kementerian Kesehatan RI, 2018):

1. Batuk berdahak selama dua minggu atau lebih
2. Dahak bercampur darah
3. Batuk darah
4. Sesak nafas
5. Badan lemas
6. Malaise
7. Nafsu makan menurun
8. BB (Berat Badan) menurun
9. Berkeringat saat malam hari tanpa kegiatan fisik
10. Demam meriang lebih dari satu bulan

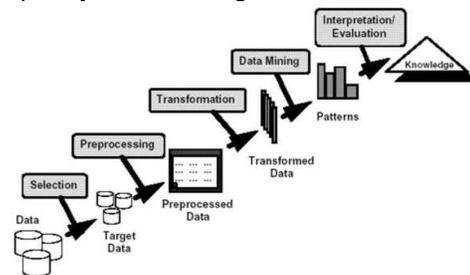
Adapun penularan bakteri *Mycobacterium Tuberculosis* yaitu dengan cara batuk penderita TB mengeluarkan dahak yang mengandung basil-basil virus TB paru dan virus yang mengandung dan bertebaran diudara sehingga terhirup orang sehat dan masuk ke paru-paru kemudian mengakibatkan penyakit TB paru dengan berbagai tahapan. TB pun dapat menular orang yang tinggal dirumah padat, kurang sinar matahari dan siklus buruk atau lembab karena bakteri *Mycobacterium Tuberculosis* akan dapat menetap lama dan berkembangbiak tapi jikalau banyak udara dan sinar matahari sirkulasi dan ventilasi baik maka bakteri tidak akan bertahan lama sekitar 1 sampai 2 jam (Ibrahim, 2017).

2.2. Data Mining

Data mining merupakan proses menelusuri pengetahuan terbaru, pola dan tren yang dipilih dari jumlah data yang besar dan disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan teknik matematika (Ardiyansyah, Rahayuningsih dan Maulana, 2018). *Data mining* kemudian dikenal dengan nama *Knowledge-discovery in Database (KDD)* adalah kegiatan yang

meliputi pengumpulan, pemakaian data historis untuk memecahkan pola atau hubungan keteraturan dalam set data yang berukuran besar. Keluaran dari data *mining* ini bisa dipakai untuk memperbaiki pengambilan keputusan dimasa depan berdasarkan informasi yang diperoleh dari data masa lalu (Setyo & Sudradjat, 2017)

Proses KDD secara garis besar terdiri dari dapat dijelaskan sebagai berikut:



Sumber: (Metisen dan Sari, 2015)

Gambar 2.1. Tahapan Proses Knowledge Discovery in Database

Secara garis besar gambar diatas, menjelaskan langkah-langkah utama dalam proses KDD yaitu sebagai berikut (Metisen dan Sari, 2015):

1. *Data Selection*, tahap ini melakukan pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi, data hasil seleksi akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.
2. *Pre-processing/Cleaning*, operasi dasar seperti penghapusan noise. Proses ini membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan data contohnya seperti kesalahan cetak. Data bisa diperkaya dengan data atau informasi eksternal yang relevan.
3. *Transformation*, pada tahap ini proses integrasi pada data yang telah dipilih, sehingga data sesuai untuk proses data mining.
4. *Data mining*, pada tahap ini menentukan tipe data mining yang akan digunakan, seperti klasifikasi, clustering atau regresi dan lain-lain tergantung pada proses dan tujuan KDD secara keseluruhan.
5. *Interpretation/Evaluation*, penerjemah pola-pola yang dihasilkan dari data mining. Pola informasi yang akan dihasilkan perlu ditampilkan dalam bentuk yang mudah dimengerti.

2.3. Algoritma Klasifikasi Data Mining

Algoritma Klasifikasi Data Mining merupakan suatu pengelompokan data untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas label atau target yang bertujuan untuk memprediksi kelas dari objek yang label kelasnya tidak diketahui. Performa algoritma data mining pada banyak kasus tergantung kepada kualitas dataset yang digunakan, karena data training yang berkualitas rendah dapat menyebabkan klasifikasi yang lemah juga (Ardiyansyah, Rahayuningsih dan Maulana, 2018).

Dalam klasifikasi data mining proses klasifikasi didasarkan pada empat komponen mendasar (Gorunescu, 2011) yaitu:

1. Kelas

Variabel dependen yang merupakan variabel kategoris yang mewakili 'label' yang terdapat pada objek. Contoh kelas-kelas tersebut adalah resiko penyakit hepatitis, resiko kredit, loyalitas pelanggan, jenis gempa, jenis bintang.

2. *Predictor*

Variabel independen dari suatu model yang diwakili oleh karakteristik atau atribut dari data yang akan diklasifikasikan dan berdasarkan klasifikasi yang akan dibuat. Contoh predictor misalnya status perkawinan, tekanan darah, merokok, konsumsi alkohol, frekuensi pembelian, musim, arah angin dan kecepatan.

3. *Training Dataset*

Kumpulan data yang berisi nilai dari kedua komponen diatas yang digunakan untuk melatih model untuk mengenali kelas yang sesuai berdasarkan predictor yang tersedia. Contoh dari dataset tersebut misalnya kelompok pasien pada serangan jantung.

4. *Testing Dataset*

Data yang berisi data baru yang akan diklasifikasi oleh model yang akan dibuat dan akurasi klasifikasi dapat dievaluasi.

2.4. Naive Bayes

Naive Bayes dikemukakan oleh ilmuwan Inggris *Thomas Bayes*. *Naive Bayes* memprediksi peluang dimasa depan dengan berdasarkan pengalaman dimasa sebelumnya. *Naive Bayes* juga dinilai berpotensi baik dalam mengklasifikasikan dokumen dibandingkan dengan metode klasifikasi lainnya dalam hal akurasi (Wahyuningsih dan Patima, 2018). *Naive Bayes* merupakan metode yang tidak memiliki aturan, *Naive Bayes* menggunakan cabang matematika yang dikenal dengan teori probabilitas untuk mencari peluang

terbesar dari beberapa kemungkinan klasifikasi, dengan melihat frekuensi tiap klasifikasi pada data training. *Klasifikasi Naive Bayes* merupakan pengklasifikasian statistik yang digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Klasifikasi bayesian didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree* dan *neural network* (Dewi, 2016).

Bayes rule digunakan untuk menghitung *probabilistik* suatu *class*. Algoritma *Naive Bayes* memberikan suatu cara dengan mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula baru yang digunakan untuk menghitung peluang dari tiap kemungkinan yang akan terjadi. Bentuk umum dari *teorema bayes* seperti dibawah ini:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \quad (2.1)$$

Dimana:

X : Data dengan kelas yang belum diketahui

H : Hipotesis data X merupakan suatu kelas spesifik

P(H|X) : *Probabilitas* hipotesis H berdasar kondisi X (*posterior probability*)

P(H) : Probabilitas hipotesis H (*prior probability*)

P(X|H) : *Probabilitas* X berdasar kondisi pada hipotesis H

P(X) : *Probabilitas* dari X

Naive Bayes adalah penyederhanaan metode *Bayes*. *Teorema Bayes* setelah disederhanakan menjadi:

$$P(H|X) = P(X|H)P(X) \quad (2.2)$$

Pada *Bayes rule* diterapkan untuk menghitung *posterior* dan probabilitas dari data sebelumnya. Dalam analisis *Bayesian*, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi yaitu informasi *prior* dan informasi *posterior* untuk menghasilkan *probabilitas* menggunakan aturan *Bayes*.

2.5. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) sering digunakan dalam sebuah penelitian, karena *Particle Swarm Optimization* memiliki kesamaan sifat dengan *Genetic Algorithm (GA)*. Keuntungan dari PSO adalah mudah diterapkan dan ada beberapa parameter

untuk menyesuaikan-nya (Widiastuti, Santosa dan Supriyanto, 2014). *Particle Swarm Optimization* merupakan teknik optimasi stokastik berbasis populasi yang diusulkan pada tahun 1995 oleh Kennedy dan Eberhart. Pengembangan PSO didasarkan pada metafora interaksi sosial dan komunikasi dari pergerakan kawanan burung atau ikan (*bird flocking* atau *fish schooling*) (Delicate et al., 2014).

Hu et al (2003) *Particle Swarm Optimization* menggunakan teknik perhitungan evolusioner (Wajhillah, 2014):

1. *Particle Swarm Optimization* diinisialisasi dengan sekumpulan solusi acak.
2. *Particle Swarm Optimization* mencari solusi yang optimum dengan memperbarui generasi.
3. Perkembangan populasi berdasarkan pada generasi sebelumnya.

Pada algoritma *Particle Swarm Optimization*, pencarian solusi dilakukan oleh suatu populasi yang terdiri dari beberapa partikel. Ketika Populasi dibangkitkan secara *random* atau acak dengan batasan permasalahan yang dihadapi. Setiap partikel merepresentasikan partikel atau solusi dari permasalahan yang sedang dihadapi. Partikel tersebut mencari solusi yang optimal dengan melintasi ruang pencarian dengan cara partikel terkait melakukan penyesuaian terhadap posisi yang terbaik dari setiap partikel tersebut atau disebut *local best* dan posisi partikel terbaik dari seluruh kawanan atau disebut *global best* selama melintasi ruang pencarian. Jadi penyebaran informasi terjadi dalam partikel itu sendiri dan antara suatu partikel dengan partikel terbaik dari seluruh kawanan selama proses pencarian solusi. Setelah itu dilakukan proses pencarian untuk mencari posisi terbaik setiap partikel dalam jumlah iterasi tertentu sampai didapatkannya posisi relatif yang tetap (*steady*) atau mencapai batas *iterasi* yang telah ditetapkan. Pada setiap *iterasi* (t), setiap solusi yang direpresentasikan oleh partikel i , dievaluasi performanya dengan cara memasukkan solusi tersebut kedalam nilai *fitness function*.

Setiap partikel memerlukan titik pada suatu dimensi ruang tertentu kemudian terdapat 2 faktor yang memberikan karakter terhadap status partikel pada ruang pencarian yaitu pada posisi partikel (X) dan pada kecepatan partikel (Y). Formulasi matematika yang menggambarkan posisi dan kecepatan partikel suatu ruang dimensi tertentu sebagai berikut:

$$X_i(t) = X_{i1}(t), X_{i2}(t), \dots, X_{in}(t) \quad (2.3)$$

$$V_i(t) = V_{i1}(t), V_{i2}(t), \dots, V_{in}(t) \quad (2.4)$$

Persamaan diatas (2.4) digunakan untuk menggambarkan kecepatan partikel yang baru berdasarkan kecepatan pada kecepatan sebelumnya, jarak antara posisi saat ini dengan posisi partikel terbaik atau *local best*, dan jarak antara posisi saat ini dengan posisi terbaik dalam kawanan atau *global best*. Kemudian partikel terbang menuju posisi yang baru berdasarkan persamaan (2.5).

$$V_i(t) = v_{i1}(t-1) + c_1 r_1 (X_i^L - X_i(t-1)) + c_2 r_2 (X_i^G - X_i(t-1)) \quad (2.5)$$

$$X_i(t) = v_i(t) + X_i(t-1) \quad (2.6)$$

Dimana:

$V_i(t)$: kecepatan partikel ke- i pada *iterasi* ke- i

$X_i(t)$: posisi partikel saat ini pada partikel ke- i pada *iterasi* ke- i

t : *iterasi*

X_i^L : *local best* dari partikel ke- i

X_i^G : *global best* dari seluruh kawanan

$c_1 c_2$: konstanta akselerasi atau *learning rate*

$r_1 r_2$: bilangan *random* atau acak yang bernilai antara 0 sampai dengan 1

2.6. Validasi (*K-Fold Cross Validation*)

K-Fold Cross validation merupakan suatu metode evaluasi dimana pada metode ini data yang digunakan dalam jumlah yang sama untuk *training* dan tepat satu kali untuk *testing* (Nugroho, Nhita dan Trantoro, 2016). Dengan membagi data secara acak kedalam k bagian dan masing-masing bagian akan dilakukan proses klasifikasi, secara umum pengujian nilai k dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi (Ramanda, 2015). *Cross validation* yang paling sering digunakan adalah *10-fold cross validation*. Prinsip dari *10-fold cross validation* adalah 9:1 dimana 9 bagian menjadi *data training* dan 1 bagian menjadi *data testing*, sehingga 10 bagian tersebut dapat berkesempatan menjadi data testing (Buani, 2016).

2.7. Evaluasi *Confusion Matrix* dan *Receiver Operating Characteristic (ROC) Curve*

Confusion Matrix adalah suatu metode untuk melakukan evaluasi dengan menggunakan tabel *matrix* yang digunakan pada konsep *data mining* untuk melakukan perhitungan akurasi (Rosandy, 2016). Evaluasi dengan menggunakan fungsi *confusion matrix* akan menghasilkan nilai *accuracy*, *precision*, dan *recall* (Buani, 2016). Ham & Kamber (2006) Nilai *accuracy* adalah persentase dari jumlah *record* data yang diklasifikasikan secara baik dan benar dengan menggunakan sebuah algoritma dan dapat membuat klasifikasi setelah dilakukan pengujian hasil klasifikasi tersebut. Nilai *precision* atau juga dikenal dengan nama *confidence value* merupakan proporsi dari jumlah kasus yang diprediksi mendapatkan hasil positif dimana nilainya juga akan positif pada data sebenarnya. Powers (2011) nilai dari *Recall* atau *sensivity value* merupakan proporsi dari jumlah kasus yang bernilai positif yang sebenarnya dan diprediksi positif secara benar (Buani, 2016).

Tabel 2.1 Model *Confusion matrix*

| Aktual | Classified as | |
|--------|---------------------|---------------------|
| | + | - |
| + | True Positives (A) | False Negatives (B) |
| - | False Positives (C) | True Negatives (D) |

Sumber: (Rosandy, 2016, p. 57)

Model *confusion matrix* pada tabel 2.2 dapat dijelaskan sebagai berikut, dimana:

True Positives (A) merupakan jumlah *record positive* yang diklasifikasikan sebagai *positive*.

False Negatives (B) merupakan jumlah *record positive* yang diklasifikasikan sebagai *negative*.

False Postitives (C) merupakan jumlah *record negative* yang diklasifikasikan sebagai *positive*.

True Negatives (D) merupakan jumlah *record negative* yang diklasifikasikan *negative*, lalu masukkan data uji.

Setelah data uji dimasukkan kedalam *confusion matrix*, maka dimehitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *accuracy*, *sensitivity* untuk mengukur proporsi 'positif' yang benar yang didiagnosa dengan benar, *spesificity* untuk mengukur proporsi 'negatif' yang benar yang didiagnosa dengan benar, PPV (*Positive Predictive Value*) adalah proporsi kasus dengan hasil tes 'positif' yang didiagnosa dengan benar, dan NPV (*Negative Predictive Value*) adalah proporsi kasus dengan hasil tes 'negatif' yang didiagnosa dengan benar.

Dapat dihitung menggunakan rumus (Gorunescu, 2011):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2.8)$$

$$Spesificity = \frac{TN}{TN+F} \quad (2.9)$$

$$PPV = \frac{TP}{TP+FP} \quad (2.10)$$

$$NPV = \frac{TN}{TN+FN} \quad (2.11)$$

Keterangan:

TP : jumlah *true positives*

TN : jumlah *true negatives*

FP : jumlah *false positives*

FN : jumlah *false negatives*

Sedangkan fungsi *ROC Curve* adalah untuk memperlihatkan akurasi dan membandingkan klasifikasi secara visual. *ROC* mengekspresikan *Confusion Matrix*. *ROC* merupakan grafik dua dimensi dengan garis *horizontal* sebagai *false positive* dan garis *vertical* sebagai *true positive*. Secara teknis kurva *ROC* juga dikenal sebagai grafik *ROC*, dua dimensi grafik dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X (Buani, 2016).

Gorunescu (2011) hasil perhitungan dapat divisualisasikan dengan *ROC Curve* atau *AUC (Area Under Curve)*. Berikut tingkat nilai diagnosa dari *ROC Curve* (Buani, 2016) yaitu:

- Akurasi bernilai 0.90-1.00 sama dengan *Excellent Classification*
- Akurasi bernilai 0.80-0.90 sama dengan *Good Classification*
- Akurasi bernilai 0.70-0.80 sama dengan *Fair Classification*
- Akurasi bernilai 0.60-0.70 sama dengan *Poor Classification*
- Akurasi bernilai 0.50-0.60 sama dengan *Failure*

3. Metode Penelitian

3.1. Jenis Penelitian

Jenis penelitian pada penelitian ini adalah jenis penelitian eksperimen, yaitu penelitian yang dilakukan dengan cara menguji kebenaran sebuah hipotesis dengan statistik yang melibatkan penyelidikan beberapa variabel dengan menggunakan tes tertentu dan menghubungkan dengan

masalah penelitian. Tujuan dari penelitian ini untuk melakukan komparasi dan mengevaluasi metode *Naive Bayes* sebagai algoritma terpilih dan metode *Naive Bayes* Berbasis *Particle Swarm Optimization (PSO)* untuk mengetahui algoritma yang memiliki keakuratan lebih tinggi dalam memprediksi penyakit tuberculosis. Penelitian ini menggunakan data primer yang diperoleh dari hasil riset di salah satu Rumah Sakit yang ada di wilayah Kabupaten Sukabumi.

3.2. Metode Pemilihan Populasi dan Sampel

Populasi dalam penelitian ini yaitu 346 data pasien tuberculosis tahun 2015-2018 yang berasal dari data internal salah satu Rumah Sakit yang ada di Kabupaten Sukabumi. Sedangkan Sampel dari penelitian ini adalah data rekam medik penderita penyakit TB Positif dan TB Negatif, data tersebut bersifat intern yang belum dipublikasikan oleh pihak RS.

3.3. Metode Pengumpulan Data

Dalam penelitian ini metode pengumpulan data untuk mendapatkan sumber data yang digunakan adalah metode pengumpulan data primer. Data utama diperoleh dari data rekam medik RS pasien yang positif TB dan pasien yang negatif TB, sedangkan data pendukung lainnya didapat dari buku, jurnal dan publikasi lainnya.

3.4. Instrumen Penelitian

Instrumen penelitian ini sebagai berikut:

1. Penelitian ini menggunakan data primer berupa data pasien yang menderita positif TB dan negatif TB yang akan digunakan sebagai instrumen guna memperoleh data dalam proses mendiagnosa penyakit TB.
2. Data disajikan dalam bentuk tabulasi model dan variabel masing-masing sebanyak 346 pasien yang terdiri dari pasien TB sebanyak 230 dan pasien bukan TB sebanyak 116 yang diambil tahun 2015 sd 2018.
3. Perangkat lunak yang digunakan untuk menganalisis adalah Rapidminer versi 8.2.

3.5. Metode Analisis dan Pengujian Data

Dalam penelitian ini pengujian data menggunakan metode eksperimen dengan model *Cross Industry Standard Process for Data Mining (CRISP-DM)* yang terdiri dari enam tahap (Listriani, Setyaningrum dan Eka, 2016), yaitu:

1. Tahap *Business Understanding*

Berdasarkan dari hasil laporan RS tahun 2015 - 2018 yang menderita penyakit TB itu sekitar 346 jiwa orang, untuk mengurangi jumlah resiko penyakit TB Positif harus dilakukan pemeriksaan terhadap pasien agar dapat diketahui penyakit secara dini. Banyak peneliti yang telah dilakukan untuk memprediksi penyakit TB, pada penelitian tersebut diantaranya penelitian Ekata et al (2016) dengan diperoleh akurasi sebesar 92%, namun demikian akurasi tersebut masih memiliki kemungkinan untuk ditingkatkan dengan penerapan optimasi metode *Naive Bayes* berbasis *Particle Swarm Optimization* dalam memprediksi penyakit TB.

2. Tahap Data *Understanding*

Data yang digunakan adalah data primer, yang didapat dari data rekam medik RS, didalam data tersebut dapat diketahui status pasien yang positif TB dan negatif TB yang terdiri dari 9 atribut *predictor* dan 1 atribut hasil, dengan jumlah 346 record yang terdiri dari 230 record (66,5%) positif TB dan 116 record (33,5%) negatif TB. Atribut-atribut yang menjadi parameter terlihat pada Tabel 3.1.

Tabel 3.1. Atribut dan Nilai Kategori Penyakit TB

| Atribut | Kategori |
|--------------------|---------------------|
| Lama Batuk | > 3 Bulan |
| | 1 Bulan -< 3 Bulan |
| | 1 Minggu -< 1 Bulan |
| | Tidak Batuk |
| Jenis Batuk | Batuk Berdahak |
| | Batuk Berdarah |
| | Batuk Kering |
| | Tidak Batuk |
| Sesak Nafas | < 1 Minggu |
| | > 1 Bulan |
| | 1 Minggu - 1 Bulan |
| | Tidak Sesak |
| Sakit Dinding Dada | Sakit Dada Terus |
| | Sakit Setiap Batuk |
| | Tidak Sakit Dada |
| Lemas | > 2 Minggu |
| | 1-2 Minggu |
| | Tidak Lemas |
| | Ya |

| | |
|--------------------|--------------------|
| Hilang Nafsu Makan | Tidak |
| BB Menurun | Ya |
| | Tidak |
| Keringat Malam | Ya |
| | Tidak |
| Demam | Agak Demam |
| | Demam Panas Sekali |
| | Kadang-Kadang |
| | Tidak Demam |
| Hasil | Negatif |
| | Positif |

3. Tahap Data Preparation

Jumlah data yang diperoleh pada penelitian ini sebanyak 346 *record*, baik pasien TB dan pasien bukan TB, akan tetapi data tersebut masih mengandung inkonsisten data atau anomali. Untuk mendapatkan data yang berkualitas dilakukan teknik *preprocessing* data yang didapat dari data internal RS sebanyak 346 *record* direduksi dengan menghilangkan duplikasi menjadi 258 *record*, 165 *record* Positif TB dan 93 *record* Negatif TB.

4. Tahap Modelling

Tahap ini dilakukan untuk menerapkan teknik yang tepat guna mendapatkan hasil yang optimal dalam memprediksi penyakit TB. Pada penelitian ini model yang digunakan yaitu metode *Naive Bayes* dan metode *Naive Bayes* berbasis *Particle Swarm Optimization (PSO)* yang kemudian menghasilkan sejumlah aturan.

5. Tahap Evaluation

Tahap ini melakukan pengujian terhadap model-model untuk mendapatkan informasi model yang akurat. Tahap Evaluasi dan tahap validasi menggunakan *Confusion Matrix* dan *ROC Curve*.

6. Tahap Deployment

Pada tahap ini diterapkan model yang memiliki akurasi tinggi atau yang paling baik pada RS yang relevan untuk mendiagnosa atau memprediksi penyakit TB dengan menggunakan data baru.

4. Hasil dan Pembahasan

4.1. Hasil Penelitian

Tujuan penelitian ini yaitu dengan menerapkan *Particle Swarm Optimization (PSO)* untuk mengeliminasi atribut input pada

metode *Naive Bayes*, untuk meningkatkan tingkat akurasi prediksi penyakit tuberculosus. Data dianalisis dengan melakukan dua perbandingan yaitu menggunakan metode *Naive Bayes* dan metode *Naive Bayes* berbasis *Particle Swarm Optimization*.

4.1.1. Eksperimen dan Pengujian Model Naive Bayes

Pembuatan model *naive bayes* dilakukan pada *dataset* yang terdiri dari 10 atribut yang merupakan atribut dari diagnosa penyakit TB dan *class* yang merupakan hasil akhir prediksi. Pada saat membuat model *naive bayes* terlebih dahulu kita harus mencari nilai *probabilitas hipotesis* untuk masing-masing kelas $P(H)$ (Probabilitas Hipotesis). Hipotesis yang ada yaitu pasien positif TB dan pasien negatif TB. Data *training* yang digunakan dengan total data yaitu 258 dengan 165 pasien yang positif TB dan 93 pasien yang negatif TB, perhitungan probabilitas seperti dibawah ini:

$$P(\text{Positif TB}) = 165:258 = 0,6395349$$

$$P(\text{Negatif TB}) = 93:258 = 0,3604651$$

Setelah probabilitas untuk tiap hipotesis diketahui langkah selanjutnya adalah menghitung probabilitas kondisi tertentu (probabilitas X) berdasarkan probabilitas tiap hipotesis (probabilitas H) atau dinamakan *probabilitas prior* dengan menggunakan *naive bayes*.

Perhitungan diatas dapat dibuatkan model *Naive Bayes* dengan *Framework Rapidminer versi 8.2*. Sehingga didapatkan hasil seperti gambar 4.1 dibawah ini:

SimpleDistribution

Distribution model for label attribute HASIL

Class Negatif (0.360)
9 distributions

Class Positif (0.640)
9 distributions

Gambar 4.1 Hasil Perhitungan Probabilitas dengan RapidMiner

4.1.2. Hasil Pengujian dengan Metode Naive Bayes

Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan nilai *accuracy* dan nilai *Area Under Curve (AUC)*.

a. Evaluasi Model dengan Confusion Matrix

Model *confusion matrix* akan membentuk *matrix* yang terdiri dari *tupel positif* atau *true positif* dan *tupel negatif* atau *true negatif*, kemudian masukkan data *testing* yang sudah disiapkan terlebih dahulu kedalam *confusion matrix* sehingga didapatkan hasil pada tabel dibawah ini:

Tabel 4.1 Confusion Matrix Metode Naive Bayes

| accuracy: 92.69% +/- 7.59% (micro average: 92.64%) | | | |
|--|--------------|--------------|-----------------|
| | true Negatif | true Positif | class precision |
| pred. Negatif | 77 | 3 | 96.25% |
| pred. Positif | 16 | 162 | 91.01% |
| class recall | 82.80% | 98.18% | |

Sumber: Data Hasil Olahan dengan Rapid Miner

Berdasarkan tabel diatas data testing terdapat rincian jumlah True Positif (TP) 162, False Negatif (FN) 16, False Positif (FP) 3, dan True Negatif (TN) 77. Dari data tersebut maka dapat dihitung nilai accuracy, sensitivity, spesificity, PPV dan NPV. Data hasil olahan dapat dilihat pada tabel dibawah ini:

Tabel 4.2 Nilai Accuracy , Sensitivity, Spesificity, PPV dan NPV Naive Bayes

| | Nilai |
|-------------|--------|
| Accuracy | 0,9264 |
| Sensitivity | 0,9101 |
| Spesificity | 0,9625 |
| PPV | 0,9818 |
| NPV | 0,8280 |

Secara manual data tersebut dihitung dengan menggunakan persamaan (2.7), (2.8), (2.9), (2.10) dan (2.11), sebagai berikut:

$$Accuracy = \frac{TP + T}{TP + TN + FP + F} = \frac{162 + 77}{162 + 77 + 3 + 1} = 0,9264$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{162}{162 + 1} = 0,9101$$

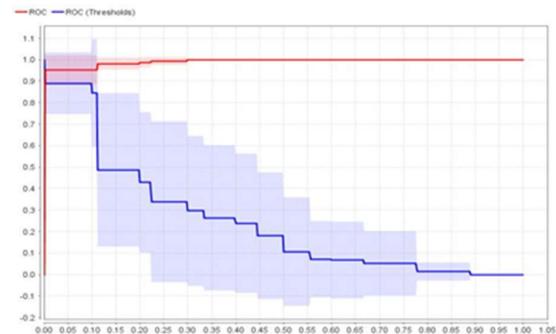
$$Spesificity = \frac{TN}{TN + FP} = \frac{77}{77 + 3} = 0,9625$$

$$PPV = \frac{TP}{TP + FP} = \frac{162}{162 + 3} = 0,9818s$$

$$NPV = \frac{TN}{TN + F} = \frac{77}{77 + 16} = 0,8280$$

b. Evaluasi dengan ROC Curve

Hasil pengujian terhadap data testing untuk metode naive bayes terdapat nilai ROC diketahui pada gambar dibawah ini:



Gambar 4.2 Nilai AUC Metode Naive Bayes dalam grafik ROC

Berdasarkan nilai AUC sebesar 0,992 yang ditunjukkan gambar diatas maka akurasi memiliki tingkat *Excellent Classification*.

4.1.3. Eksperimen dan Pengujian Model Naive Bayes berbasis PSO

Seleksi fitur yang digunakan pada penelitian ini yaitu *particle swarm optimization*. Dengan PSO data yang akan diolah akan diberikan bobot untuk membantu meningkatkan hasil perhitungan, pemberian bobot ini diberikan secara acak dengan menentukan nilai minimum dan maksimum bobot.

Indikator untuk PSO yaitu *population*, *inertia*, dan *global best weight* ikut diuji untuk mendapatkan akurasi yang maksimal. Populasi yang digunakan adalah 5 populasi. Dalam penyesuaian indikator PSO *maximum number of generation* 30, nilai *inertia* 1,0, nilai *global best* 100,0 dan *local best* 50,0. Jumlah partikel yang diambil 5. Nilai C1 dan C2 masing-masing 1 karena partikel berada dalam putaran pertama.

Perhitungan pada keseluruhan data dilakukan dengan rapid miner. Model yang terbentuk dengan metode naive bayes sendiri akan dibandingkan dengan model yang terbentuk setelah proses pembobotan atribut. Nilai bobot atribut akan diberikan kepada oleh setiap partikel PSO. Model yang terbentuk akan diuji sehingga dapat diketahui akurasi model yang terbentuk, setelah itu partikel PSO akan bergerak kearah partikel terbaik agar akurasi model yang dihasilkan dapat lebih tinggi dari sebelumnya. Hasil dari proses optimasi fitur oleh PSO menggunakan rapid miner sebagai berikut:

| attribute | weight |
|--------------------|--------|
| LAMA BATUK | 1 |
| JENIS BATUK | 1 |
| SESAK NAFAS | 0 |
| SAKIT DINDING DADA | 0 |
| LEMAS | 1 |
| HILANG NAFSU MAKAN | 1 |
| BB MENURUN | 1 |
| KERINGAT MALAM | 0 |
| DEMAM | 1 |

Gambar 4.3 Hasil Proses Optimize Selection Menggunakan Rapid Miner

Terdapat 9 atribut yang digunakan dan hasil pembobotan atribut yaitu 3 atribut mempunyai bobot 0 atau tidak berpengaruh terhadap akurasi yaitu Sesak Nafas, Sakit Dinding Dada dan Keringat Malam dan 6 atribut mempunyai bobot 1 diantaranya Lama Batuk, Jenis Batuk, Lemas, Hilang Nafsu Makan, BB Menurun dan Demam. Hal ini dapat mempengaruhi peningkatan hasil *accuracy* pada proses klasifikasi sedangkan atribut yang bernilai 0 dapat dihilangkan karena tidak mempunyai pengaruh pada akurasi prediksi penyakit tuberkulosis.

4.1.4. Hasil Pengujian dengan Metode Naive Bayes berbasis PSO

Hasil dari uji coba yang dilakukan yaitu untuk menghasilkan nilai *accuracy* dan nilai *AUC* (Area Under Curve) menggunakan metode *naive bayes* berbasis PSO sebagai berikut:

a. Evaluasi Model dengan Confusion Matrix

Model *confusion matrix* akan membentuk *matrix* yang terdiri dari *tupel positif* atau *true positif* dan *tupel negatif* atau *true negatif*, kemudian masukan data *testing* yang sudah disiapkan terlebih dahulu kedalam *confusion matrix* sehingga didapatkan hasil pada tabel dibawah ini:

Tabel 4.3 Confusion Matrix Metode Naive Bayes berbasis PSO

| accuracy: 98.72% | | | |
|------------------|--------------|--------------|-----------------|
| | true Negatif | true Positif | class precision |
| pred. Negatif | 27 | 0 | 100.00% |
| pred. Positif | 1 | 50 | 98.04% |
| class recall | 96.43% | 100.00% | |

Sumber: Data Hasil Olahan dengan Rapid Miner

Berdasarkan tabel diatas data testing terdapat rincian jumlah True Positif (TP) 50, False Negatif (FN) 1, False Positif (FP) 0, dan True Negatif (TN) 27. Dari data tersebut maka akan dihitung nilai *accuracy*, *sensitivity*, *specificity*, *PPV* dan *NPV*. Data hasil olahan dapat dilihat pada tabel 4.

Tabel 4.4 Nilai Accuracy, Sensivity, Specificity, PPV dan NPV Naive Bayes berbasis PSO

| | Nilai |
|--------------------|--------|
| <i>Accuracy</i> | 0,9872 |
| <i>Sensitivity</i> | 0,9804 |
| <i>Specificity</i> | 1 |
| <i>PPV</i> | 1 |
| <i>NPV</i> | 0,9643 |

Secara manual data tersebut dihitung menggunakan persamaan (2.7), (2.8), (2.9), (2.10) dan (2.11) sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{50 + 27}{50 + 27 + 0 + 1} = 0,9872$$

$$Sensitivity = \frac{TP}{TP + FN} = \frac{50}{50 + 1} = 0,9804$$

$$Specificity = \frac{TN}{TN + FP} = \frac{27}{27 + 0} = 1$$

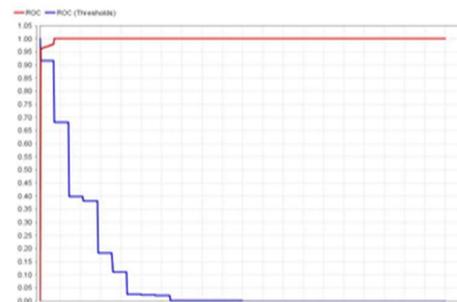
$$PPV = \frac{TP}{TP + FP} = \frac{50}{50 + 0} = 1$$

$$NPV = \frac{TN}{TN + FN} = \frac{27}{27 + 1} = 0,9643$$

Dari nilai-nilai tersebut diketahui bahwa nilai *Negative Predictive Value* memiliki nilai yang paling rendah diantara nilai pada pengujian lainnya.

b. Evaluasi dengan ROC Curve

Hasil pengujian terhadap data testing untuk metode Naive Bayes berbasis PSO terdapat nilai ROC seperti gambar dibawah ini:



Gambar 4.4 Nilai AUC Naive Bayes berbasis PSO dalam Grafik ROC

Berdasarkan nilai AUC sebesar 0,999 yang ditunjukkan pada gambar diatas maka akurasi memiliki tingkat *Excellent Classification*.

4.2. Pembahasan

Hasil pengujian menyimpulkan bahwa atribut yang mempengaruhi pengujian adalah 7 atribut yang terdiri dari 6 atribut *predictor* dan 1 atribut *class*. Hasil pengujian model *Naive Bayes* tanpa seleksi atribut dibandingkan dengan model *Naive Bayes* dengan seleksi atribut menggunakan *PSO* dapat dilihat pada tabel 4.7.

Tabel 4.5 Pengujian Metode Naive Bayes dan Naive Bayes berbasis PSO

| | Accucary | AUC |
|--------------------------|----------|-------|
| Naive Bayes | 92,69 % | 0,992 |
| Naive Bayes berbasis PSO | 98,72% | 0,999 |

Dari hasil pengujian diatas, dengan dilakukan evaluasi baik secara *confusion matrix* maupun *ROC Curve* terbukti bahwa pengujian yang dilakukan oleh metode *Naive Bayes* berbasis *Particle Swarm Optimization (PSO)* memiliki nilai akurasi yang lebih tinggi dibandingkan hanya dengan menggunakan satu metode *Naive Bayes*. Nilai akurasi untuk metode *naive bayes* sebesar 92,69% dengan nilai AUC 0,992 dan nilai akurasi model *naive bayes* dengan seleksi atribut menggunakan *PSO* sebesar 98,72% dan AUC 0,999. Berdasarkan nilai tersebut diperoleh selisih akurasi sebesar 6,03% dan selisih AUC sebesar 0,007.

5. Kesimpulan

Dari pembahasan-pembahasan diatas maka dapat ditarik kesimpulan bahwa:

1. Penelitian dengan menerapkan *particle swarm optimization* untuk pemilihan atribut pada metode *naive bayes* dilakukan agar dapat meningkatkan akurasi prediksi dalam mendiagnosa penyakit tuberculosis.
2. Hasil penelitian untuk nilai akurasi metode *naive bayes* senilai 92,69%, sedangkan untuk nilai akurasi *naive bayes* berbasis *PSO* sebesar 98,76% sehingga tampak selisih nilai akurasi yaitu sebesar 6,07%. Sedangkan evaluasi menggunakan *ROC Curve* diperoleh hasil untuk metode *naive bayes* bernilai 0,992 dengan tingkat diagnosa *excellent classification* dan metode *naive bayes* berbasis *PSO*

bernilai 0,999 dengan tingkat diagnosa *excellent classification*, didapatkan selisih nilai AUC sebesar 0,007.

3. Dari 10 atribut yang terdapat pada dataset pasien penyakit TB kemudian selanjutnya dilakukan seleksi atribut hanya menjadi 7 atribut yang digunakan dalam menentukan prediksi penyakit TB, atribut-atribut tersebut yaitu jenis batuk, sesak nafas, lemas, hilang nafsu makan, berat badan menurun, demam dan satu atribut prediksi.

Sehingga dapat disimpulkan bahwa penerapan teknik optimasi *particle swarm optimization* mampu menyeleksi atribut pada metode *naive bayes*, sehingga menghasilkan tingkat akurasi diagnosa penyakit tuberculosis yang lebih baik dibandingkan dengan menggunakan metode *individual naive bayes*. Dengan demikian metode *naive bayes* berbasis *PSO* merupakan metode yang sangat baik dalam memprediksi secara dini diagnosa penyakit TB. Berdasarkan proses pengujian dan kesimpulan yang telah dilakukan, agar penelitian ini bisa ditingkatkan berikut adalah saran-saran yang diusulkan yaitu:

1. Penelitian ini diharapkan dapat digunakan pada instansi dinas kesehatan khususnya pada perusahaan riset dilakukan untuk lebih meningkatkan efisiensi dan kecepatan dalam pengambilan suatu keputusan pasien.
2. Menambahkan jumlah data yang lebih besar lagi dan atribut yang lebih banyak sehingga hasil pengukuran yang akan didapatkan lebih baik lagi.
3. Menggunakan metode optimasi lainnya seperti *Genetik Algorithm*, *Bagging* dan metode optimasi lainnya.
4. Melakukan pengembangan dengan menggunakan metode *feature selection* yang lain seperti *forward selection*, *genetic algorithm* dan metode *feature selection* lainnya untuk menyeleksi atribut yang berpengaruh kuat, sehingga atribut yang dipakai hanya sedikit namun tidak mengurangi akurasi dari algoritma yang digunakan.
5. Penelitian ini dapat dikembangkan lagi dengan membandingkan algoritma data mining lainnya misal *supprot vector machine* dengan mengoptimalkan parameter menggunakan *genetic algorithm* atau dengan menggunakan pengujian model yang sama untuk dataset public sebagai data sekunder dan data hasil riset sebagai data primer.

6. Diterapkan pada perangkat lunak sesuai perkembangan IPTEK,s seperti menggunakan mobile computing.

6. DAFTAR PUSTAKA

- Abdullah, R. (2017, Juli-Desember). Studi Karakteristik Penderita TB Paru Aktif ditinjau dari Lesi Foto Thorax di RS Dr Wahidin Sudirohusodo Makasar Pada Periode Januari - Desember 2016. *JF FKI K UINAM*, Vol.II, 10-21.
- Ardiyansyah, Rahayuningsih, P. A., & Maulana, R. (2018, Juni). Analisis Perbandingan Algoritma Klasifikasi Data Mining Untuk Dataset Blogger Dengan Rapid Miner. *Jurnal Khatulistiwa Informatika*, VOL. VI(p-ISSN: 2339-1928 & e-ISSN: 2579-633X), 20-28.
- Buani, D. P. (2016). Optimasi Algoritma Naïve Bayes dengan Menggunakan Algoritma Genetika untuk Prediksi Kesuburan (Fertility). *Jurnal Evolusi* , Volume 4 Nomor 1(ISSN : 2338 – 8161), 54-63.
- Delice, Y., Aydogan , E. K., Ozcan, U., & Ilkay, M. S. (2014, September). A modified particle swarm optimization algorithm to mixed-model two-sided assembly line balancing. *J Intell Manuf*(DOI 10.1007/s10845-014-0959-7), 1-14.
- Dewi, A. (2016, Maret 60-66). Komparasi 5 Metode Algoritma Klasifikasi Data Mining Pada Prediksi Keberhasilan Pemasaran Produk Layanan Perbankan. *Jurnal Techno Nusa Mandiri*, Vol. XIII, No. 1(ISSN 1978-2136).
- [1] Ekata, Tyagi, P. K., Gupta, N. K., & Gupta, S. (2016, November). Diagnosis of Pulmonary Tuberculosis using Fuzzy Inference System. *IEEE Second International Innovative Applications of Computational Intelligence on Power, Energy and Controls with their Impact on Humanity (CIPECH)*(DOI: 10.1109/CIPECH.2016.7918726), 3-7.
- Gorunescu, F. (2011). *Data Mining: Concepts, Model and Techniques* (Vol. Vol. 12, DOI 10.1007/978-3-642-19721-5). Romania.
- Ibrahim, I. (2017, Maret). Faktor Yang Mempengaruhi Kejadian TB Paru di Wilayah Kota Tidore. *Global Health Science*, Volume 2(Issue 1 , ISSN 2503-5088), 34-40.
- Kementrian Kesehatan RI. (2018). *InfoDatin Pusat Data dan Informasi Tuberkulosis*. Jakarta.
- Listriani, D., Setyaningrum, A. H., & Eka, F. (2016, Oktober). Penerapan Metode Asosiasi Menggunakan Metode Apriori Pada Aplikasi Analisa Pola Belanja Konsumen (Studi Kasus Toko Buku Gramedia Bintaro). *Jurnal Teknik Informatika*, VOL. 9 NO. 2(ISSN 1979-9160), 120-127.
- Metisen, B. M., & Sari, H. L. (2015, September). Analisis Clustering Menggunakan Metode K-Means Dalam Pengelompokan Penjualan Produk Pada Swalayan Fadhila. *Jurnal Media Infotama*, Vol. 11 No.2(ISSN 1858 – 2680), 110-118.
- Nugroho, D., Nhita, F., & Trantoro, D. (Agustus 2016). Prediksi Penyakit Menggunakan Genetic Algorithm (GA) dan Naive Bayes Untuk Data Berdimensi Tinggi. *e-Proceeding of Engineering*, Vol.3, No.2 , hal. 3889-3899.
- Ramanda, K. (2015, Agustus). Penerapan Particle Swarm Optimization Sebagai Seleksi Fitur Prediksi Kelahiran Prematur Pada Algoritma Neural Network. *Jurnal Teknik Komputer AMIK BSI*, VOL. I NO. 2 (ISSN. 2442-2436), 178-183.
- Rosandy, T. (2016, Mei). Perbandingan Metode Naive Bayes Classifier dengan Metode Decison Tree (C4.5) Untuk Menganalisa Kelancaran Pembiayaan (Study Kasus: KSPPS/BMT Al-Fadhila). *Jurnal TIM Darmajaya*, Vol. 02 No. 01(ISSN: 2442-5567 | E-ISSN: 2443-289X), 52-62.
- Rusdah, Winarko, E., & Wardoyo, R. (2015, November). Preliminary Diagnosis of Pulmonary Tuberculosis Using Ensemble Method. *IEEE International Conference on Data and Software Engineering (ICoDSE)*(DOI: 10.1109/ICoDSE.2015.7436993), 175-180.

-
- Saputra, R. A. (2014). Komparasi Algoritma Klasifikasi Data Mining Untuk Memprediksi Penyakit Tuberculosis (TB): Studi Kasus Puskesmas Karawang Sukabumi. Seminar Nasional Inovasi dan Tren (SNIT) , (hal. 1-8). Sukabumi.
- Setyo, J. S., & Sudrajat, A. (2017, September). Penerapan Metode C4.5 Terhadap Penyakit Tuberculosis Paru. *Jurnal Kajian Ilmiah*, Volume 17, No. 3(ISSN 1410-9794, EISSN 2597-792X), 111-118.
- Wahyuningsih, D., & Patima, E. (2018, Februari). Penerapan Naive Bayes Untuk Penerimaan Beasiswa. *Jurnal Telematika*, Vol. 11 No. 1(ISSN : 1979 – 925X & e-ISSN : 2442 - 4528), 135-147.
- Wajhillah, R. (2014, September). Optimasi Algoritma Klasifikasi C4.5 Berbasis Particle Swarm Optimization Untuk Prediksi Penyakit Jantung. *SWABUMI*, VOL I, No. 1(ISSN 2355-990X), 1-12.
- Widiastuti, N. A., Santosa, S., & Supriyanto, C. (2014, Februari). Algoritma Klasifikasi Data Mining Naive Bayes Berbasis Particle Swarm Optimization Untuk Deteksi Penyakit Jantung. *Jurnal Pseudocode*, Volume 1, Nomor 1(ISSN 2355 – 5920), 11-14.
- Yuanli., Hong, & Fei. (2017, October). Automatic Classification of Pulmonary Tuberculosis and Sarcoidosis based on Random Forest. *IEEE International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI 2017)*(DOI: 10.1109/CISP-BMEI.2017.8302280), 1-5.
- Zulvia, F. E., Kuo, R., & Roflin, E. (2017, July). An Initial Screening Method for Tuberculosis Diseases Using a Multi-objective Gradient Evolution-based Support Vector Machine and C5.0 Decision Tree. *IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*(DOI: 10.1109/COMPSAC.2017.57), 204-209.