 ISBN: 978-602-60280-1-3

Perbandingan Kinerja Algoritme CART dan *Naïve Bayesian* Untuk Mendiagnosa Penyakit Diabetes Mellitus

Pungkas Subarkah[1], Irfan Santiko[2], Tri Astuti[3]
[1]*Informatics Technopreneurship, Universitas Amikom Yogyakarta, Indonesia*
[2]*Sistem Informasi, STMIK Amikom Purwokerto, Indonesia*
[3]*Teknik Informatika, STMIK Amikom Purwokerto, Indonesia*
[1]*subarkah18.pungkas@gmail.com,* [2]*irfan.santiko@amikompurwokerto.ac.id,* [3]*tri_astuti@amikompurwokerto.ac.id*

Abstrak

Penyakit diabetes mellitus merupakan salah satu penyakit yang mengancam kesehatan yang serius, dapat mengakibatkan kematian dan *Word Health Organazitation (WHO)* memperkirakan setiap 10 detik ada satu orang pasien diabetes yang meninggal karena penyakit ini. Metode yang digunakan dalam penelitian ini yaitu identifikasi masalah, pengumpulan data, tahap *pre-processing*, metode klasifikasi, validasi dan evaluasi serta penarikan kesimpulan. Algoritma yang digunakan dalam penelitian ini adalah CART dan *Naïve Bayes* dengan menggunakan *dataset* diambil dari repository *database* UCI Indian Pima yang terdiri dari data klinis pasien yang terdeteksi positif dan negatif penyakit diabetes mellitus. Adapun metode validasi dan evaluasi yang digunakan yaitu *10-cross validation* dan *confusion matrix* untuk penilaian *precision, recall dan F-measure*. Hasil perhitungan yang telah dilakukan, didapatkan hasil akurasi pada algoritma CART sebesar 76.9337% dengan *precision* 0.764%, *recall* 0.769% dan F-Measure 0.765%. Sedangkan *dataset* diabetes yang diuji dengan algoritma *Naïve Bayes* mendapatkan nilai akurasi sebesar 73.7569% dengan *precision* 0.732% , *recall* 0.738% , dan F- Measure 0.734%. Dari hasil tersebut dapat disimpulkan bahwa untuk mendiagnosa penyakit diabetes mellitus disarankan menggunakan algoritma CART.

Kata Kunci : Kinerja, Diagnosis, Algoritma

Abstract

Diabetes Mellitus is a disease that treatern serious health, can lead to death and Organization World Health Organization (WHO) estimates that every 10 seconds one person died because of this diseases. The methods used in this research were problem identification, data collection, pre-processing phase, method of classification, validation and evalution as well as the withdrawal conclusion. The algorithms used in this study were CART and Naive Bayes using a dataset retrieved from tthe database repository UCI Pima Indians which consists of clinical data of patients wete detected positive and negative diabetes mellitus. The validation and evaluation methods used were 10-cross validation and confusion for matrix assesment of precision, recall and F-Measure. The results of calculations have been done. The result accuracy on CART algorihtm was 76.9337% with 0764% precision, 0769% recall and 0765% F-Measure. Diabetes datasets in the test with the Naive Bayes algorithm to get the value of 73.7569% accuracy with precision 0732%, 0738% recall, and 0734% F-Measure. From these results it can be concluded that in order to diagnose diabetes, it is advised to use an algorithm mellitus CART.

Keyword : Performance, Diagnosis, Algorithm

I. PENDAHULUAN

Penyakit diabetes mellitus merupakan salah satu penyakit yang mengancam kesehatan yang serius dan dapat mengakibatkan kematian baik di Indonesia maupun di Dunia. Menurut survei yang dilakukan WHO tahun 2005, Indonesia sebagai Negara *lower-middle income* menempati urutan ke 4 dengan jumlah penderita diabetes mellitus terbesar di dunia setelah India, China, dan Amerika Serikat [1]. Berdasarkan Profil Kesahatan Indonesia tahun 2008, diabetes mellitus merupakan penyebab peringkat enam untuk semua umur di Indonesia dengan proporsi kematian 5,7 % dibawah stroke, TB, Hipertensi, cedera dan perinatal.

Terdapat dua type diabetes mellitus, yaitu diabetes tipe 1 yang merupakan diabetes yang tergantung pada insulin, dimana pankreas menghasilkan sedikit insulin atau sama sekali tidak menghasilkan insulin. Sedangkan pada diabetes

mellitus tipe 2, pankreas tetap menghasilkan insulin tetapi kadang kadarnya lebih tinggi dari norma kejadian ini akan menyebabkan tubuh membentuk kekebalan terhadap efeknya, sehingga kekurangan insulin relative. Tanda dan gejala umum biasanya dialami penderita diabetes mellitus [2] :

a. Rasa haus berlebihan (*Polidipsi*)
b. Sering kencing (*Poliuri*) terutama pada malam hari
c. Sering merasa lapar (*Poliphagi*)
d. Berat badan turun dengan cepat
e. Kesemutan pada tangan dan kaki
f. Penglihatan jadi kabur
g. Luka sulit sembuh
h. Impotensi
i. Penyakit kulit akibat jamur lipatan kulit
j. Pada ibu – ibu sering melahirkan bayi besar dengan berat badan >= 4 kg.

Hal ini diperkuat oleh Badan Kesehatan Dunia, *World Health Organizatation* memperkirakan , setiap 10 detik ada satu orang pasien diabetes yang meninggal karena penyakit itu dan memperkirakan bahwa 177 juta penduduk mengidap penyakit diabetes mellitus. Banyak algoritma telah digunakan dalam mendiagnosa penyakit diabetes mellitus diantaranya dengan menggunakan algoritma Optimasi Koloni Semut , dan algoritma C4.5 berbasis *Particle Swarm Optimatization*. Peneliti lain menggunakan algoritma *Decision Tree* J48 dan ID3 dan hasil akurasi dari algoritma *Decision Tree* J48 lebih unggul dibandingkan dengan algoritma ID3.

Penelitian yang dilakukan oleh I Putu Dody Lesmana menggunakan algoritma *Decision Tree* J48 dan ID3 dalam pengklasifikasian diagnosis penyakit diabetes mellitus, dan membandingkan dua algoritma dengan beberapa parameter *Pregnant, Plasma-glucose, Diastolic Blood-Pressure, Trisepts Skin Fold Thickness, Insulin, Body Mass Index, Diabetes Pedigree Function,Age, Class Variable*. Dari perbandingan kedua algoritma tersebut algoritma J48 memiliki tingkat akurasi yang lebih baik sebesar 74.72% dibandingkan dengan algoritma ID3 dengan akurasi 72.64%.

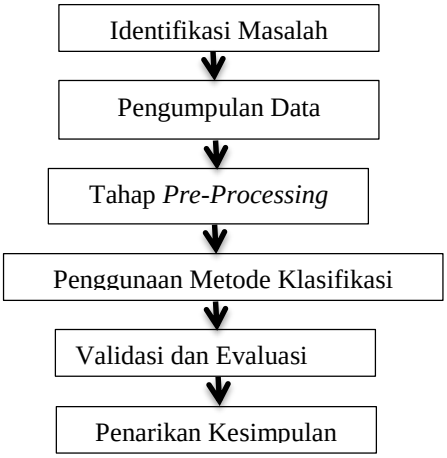
Penelitian yang dilakukan oleh Farid Nurhidayat, penentuan besar akurasi metode klasifikasi menggunakan algoritma C4.5 berbasis particle swarm optimatization pada prediksi penyakit diabetes mellitus. Adapun tujuan dari penelitian ini ialah untuk mendapatkan *rule* dalam memprediksi penyakit diabetes mellitus. Hasil dari penelitian ini dapat disimpulkan bahwa algoritma C4.5 berbasis *particle swarm optimization* memiliki akurasi dan nilai AUCR lebih tinggi daripada algoritma C4.5 dengan selisih nilai *accuracy* 3.28% dan nilai AUC 0.12%.

Penelitian yang dilakukan oleh Triajianto, Purwananto, Soelaiman, implementasi sistem klasifikasi fuzzy berbasis koloni semut untuk mendiagnosa penyakit diabetes. Tujuan penelitian ini ialah untuk memprediksi penyakit diabetes dengan sistem pakar. Metode yang digunakan adalah klasifikasi fuzzy dengan performa terbaik yang dihasilkan oleh model adalah akurasi sebesar 78.55%.

Dari kajian pustaka yang telah dilakukan terhadap beberapa paper utama, belum ada penelitian yang menggunakan algoritma CART perbandingan dengan algoritma *Naïve Bayes* yang digunakan untuk mendiagnosa penyakit diabetes mellitus. Kedua algoritma ini diuji dengan *software* Weka.

II. METODE PENELITIAN

Adapun konsep penelitian bisa dilihat pada gambar 1 dibawah ini:



Gambar. 1. Model Konsep penelitian

Keterangan :

A. Identifikasi Masalah

Proses mengidentifikasi masalah dalam penelitian dengan upaya mengetahui permasalahan yang berkaitan dengan poin – poin mendiagnosa penyakit diabetes mellitus.

B. Pengumpulan Data

Pengumpulan data yang dilakukan oleh penulis dengan mengambil *dataset* UCI repository Indian Pima

C. Tahap *Pre-Processing*

Pada tahap ini peneliti melakukan proses seleksi data dengan tujuan mendapatkan data yang valid.

D. Penggunaan Metode Klasifikasi

Dalam penelitian ini penulis menggunakan metode algoritma *CART dan Naïve Bayes.*

E. Validasi dan Evaluasi

Dalam tahap ini dilakukan validasi dan pengukuran keakuratan hasil yang dicapai oleh algoritma menggunakan aplikasi Weka.

F. Penarikan Kesimpulan

Dalam tahap ini adalah menyimpulkan hasil yang diperoleh dari algoritma CART dan *Naive Bayes* yang memberikan akurasi terbaik untuk mendiagnosa penyakit diabetes mellitus

III. HASIL DAN PEMBAHASAN

Berdasarkan metodologi diatas, maka dapat dilihat beberapa hasil sebagai berikut:

A. Identifikasi Masalah

Dalam penelitian ini penulis melakukan studi literature yang terkait dengan penelitian dan mendapatkan algoritma yang sesuai untuk penelitian ini yaitu CART dan *Naive Bayes*.

B. Pengumpulan Data

Data yang digunakan yaitu mengambil dari repository Indian Pima Diabetes yang terdiri dari 768 data klinis yang semuanya berasal dari kelamin wanita dengan umur sekurang-kurangnya 21 tahun.

C. Tahap Pre-processing

Setelah melakukan analisis terhadap dataset Indian Pima , diketahui bahwa tidak semua atribut memilki nilai yang lengkap, dimana kelengkapan nilai atribut sangat mempengaruhi klasifikasi. Atribut yang memiliki jumlah data tidak lengkap yaitu pregnant sebanyak 111, atribut *glucose* sebanyak 5, atribut DBP sebanyak 35, atribut TSFT sebanyak 227, atribut INS sebanyak 374 dan atribut BMI sebanyak 11. Sedangkan pada atribut *age* dan *class* memiliki nilai yang lengkap. Untuk menangani *missing value* dilakukan :

a. Nilai nol pada atribut pregnant dapat diasumsikan bahwa nilai tersebut menyatakan pasien belum pernah melahirkan, sehingga hal ini dimungkinkan sesuai dengan kondisi yang sebenarnya.
b. Data dengan nilai nol pada atribut glucose, DBP dan BMI dapat dihilangkan karena jumlahnya tidak terlalu banyak sehingga tidak begitu mempengaruhi hasil klasifikasi.
c. Karena atribut TSFT dan INS memiliki jumlah nilai yang tidak ada sangat besar, maka kedua atribut ini tidak dapat dihilangkan dan tidak dapat dipakai dalam pengklasifikasian. Oleh karena itu, dalam penelitian ini atribut TSFT dan INS tidak digunakan.

Setelah proses penanganan nilai yang tidak lengkap (*misssing value*) dilakukan dengan aturan diatas, maka didapatkan 724 data (249 *class* positif dan 475 *class* negatif) dari 768 data asli dan siap diolah lebih lanjut dengan atribut *pregnant, glucose*, DBP, BMI, *Age* dan *class*.

D. Proses Metode Klasifikasi

Dari hasil perhitungan dan uji coba menggunaka aplikasi weka dengan algoritma CART menghasilkan akurasi sebesar 76.9337%. Nilai akurasi tersebut didapatkan dari hasil perhitungan dari *precision*, *recall*, dan *F-Measure*. Hasil perhitungan nilai akurasi berdasarkan *confusion matrix* disajikan pada tabel 1 sebagai berikut :

Tabel 1. Nilai akurasi berdasarkan *confusion matrix*

Class	*Precision*	*Recall*	*F-Measure*
Tested_negative	0.804	0.857	0.83
Tested_positive	0.688	0.602	0.642
Weighted Avg	0.764	0.769	0.765

Sedangkan apabila pengklasifikasian menggunakan algoritma *Naive Bayes* menggunakan aplikasi weka menghasilkan *precision, recall dan F-Measure*. Hasil perhitungan nilai akurasi berdasarkan *confusion matrix* disajikan pada tabel 2 sebagai berikut :

Tabel 2. Nilai akurasi berdasarkan *confusion matrix*

Class	*Precision*	*Recall*	*F-Measure*
Tested_negative	0.804	0.857	0.83
Tested_positive	0.688	0.602	0.642
Weighted Avg	0.764	0.769	0.765

Berikut perbedaan hasil akurasi yang diperoleh menggunakan algoritma CART dan *Naive Bayes* pada tabel 3.

Tabel 3. Perbandingan Hasil akurasi CART dan *Naive Bayes*

Algoritma	Hasil Akurasi	*Precision*	*Recall*	*F-Measure*	Waktu
CART	76.9337 %	0.764	0.769	0.765	0.16 *second*
Naive Bayes	73.7569 %	0.732	0.738	0.734	0.04 *second*

Perbedaan akurasi yang diperoleh dengan menggunakan algoritma CART dan *Naive Bayes* sebesar 3.1768%. Waktu yang digunakan saat *running dataset* pada aplikasi Weka juga berbeda antara algoritma CART dan *Naive Bayes* yaitu selisih 0.12 *second.* Dalam algoritma CART terdapat suatu perbedaan yaitu secara rekursif membagi *record* pada data latihan ke dalam subset-subset yang memiliki nilai atribut target (kelas) yang sama, hal ini menyebabkan waktu kompilasi sistem menjadi lama.

E. Validasi dan Evaluasi

Tabel 4 dan 5 merupakan tabel hasil *confusion matrix* dari pengujian *dataset* menggunakan algoritma CART dan *Naive Bayes* dengna *10-fold cross validation*.

Tabel 4. *Confusion matrix* CART

	Positif Diabetes	Negative Diabetes
Positif Diabetes	407	68
Negative Diabetes	99	150
724	506	218

Tabel 5. *Confusion matrix Naive Bayes*

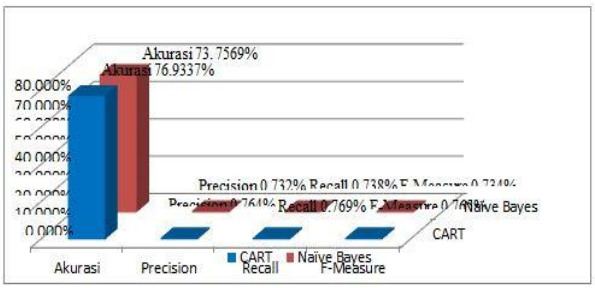
	Positif Diabetes	Negative Diabetes
Positif Diabetes	394	81
Negative Diabetes	109	140
724	503	221

Dari tabel 4 terlihat bahwa jumlah data hasil bentukan *rule* yang terkena penyakit diabetes mellitus yang sama dengan data testing yang juga terkena diabetes sebanyak 407. Kemudian, jumlah data hasil bentukan *rule* yang tidak terkena penyakit diabetes mellitus dengan data testing yang terkena diabetes sebanyak 68. Selanjutnya, jumlah data hasil bentukan *rule* yang terkena diabetes dan data testing yang tidak terkena diabetes sebanyak 99. Terakhir, jumlah data hasil bentukan *rule* yang tidak terkena diabetes yang sama dengan data testing yang juga tidak terkena diabetes sebanyak 150.

Sedangkan dari tabel 5 6 terlihat bahwa jumlah data hasil bentukan *rule* yang terkena penyakit diabetes mellitus yang sama dengan data testing yang juga terkena diabetes sebanyak 394. Kemudian, jumlah data hasil bentukan *rule* yang tidak terkena penyakit diabetes mellitus dengan data testing yang terkena diabetes sebanyak 81. Selanjutnya, jumlah data hasil bentukan *rule* yang terkena diabetes dan data testing yang tidak terkena diabetes sebanyak 109. Terakhir, jumlah data hasil bentukan *rule* yang tidak terkena diabetes yang sama dengan data testing yang juga tidak terkena diabetes sebanyak 140.

F. Penarikan Kesimpulan

Dari perhitungan yang telah dilakukan pada kedua algoritma, didapatkan hasil akurasi dari masing – masing algoritma yaitu 76.9337% dengan nilai precision 0.764%, Recall 0.769% dan F-Measure 0.765% pada algoritma CART dan 73.7569% pada algoritma Naïve Bayes dengan nilai precision 0.732%, Recall 0.738% dan F-Measure 0.734%. Berikut merupakan grafik hasil akurasi dari perhitungan pada algoritma CART dan Naïve Bayes pada gambar 2.



Gambar 2. Perbandingan Hasil Akurasi CART dan *Naive Bayes*

Dengan demikian setelah melihat hasl perhitungan diatas, untuk menentukan diagnosa penyakit mellitus lebih baik menggunakan algoritma CART karena tingkat akurasinya lebih tinggi dibandingkan dengan algoritma *naïve bayes*.

IV. KESIMPULAN

Dari hasil penelitian dilakukan, maka dapat ditarik beberapa kesimpulan sebagai berikut :

- Algoritma CART memiliki tingkat akurasi lebih baik daripada algoritma *Naive Bayes* dengan selisih akurasi 3.1768%.
- Dalam penerapan algoritma CART mempunyai sifat rekursif yaitu membagi *record* pada data latihan ke dalam subset-subset yang memiliki nilai atribut target (kelas) yang sama.
- Saran untuk penelitian selanjutnya dilakukannya penanganan nilai yang hilang (*missing value*) pada setiap atribut.
- Penulis berharap untuk penelitian selanjutnya menggunakan algoritma yang lain dengan melihat tingkat akurasi yang lebih tinggi.

DAFTAR PUSTAKA

[1] Bramer, M. 2007. Principles Of Data Mining. London : Springer.
[2] Depkes,R.I.2009. Profil Kesehatan Indonesia. Jakarta : Depkes RI
[3] Diabetes Care. 2004. *Global prevalence of diabetes: estimates for the year 2000 and projections for 2030*
[4] Gorunescu,F.2011. *Data mining Concepts, Models and Techniques*. Verlan Berlin Heidelberg: Spinger.
[5] Han,J., & Kamber,M.2006. *Data Mining Concepts And Techniques*. Verlag Berlin Heidelberg : Spinger
[6] Indri Rahmayuni.2014.Perbandingan Performansi Algoritma C4.5 dan CART dalam Klasifikasi Data Nilai Mahasiswa Prodi Teknik Komputer Politeknik Negeri Padang.Jurnal TEKNOIF Vol.2 No. 1 April 2014.
[7] Jayalskshmi, T., Santhakumaran, A., " *Impact of Prepocessing for diagnosis of diabetes mellitus using artificial neural network*," Machine Learning and Computing (ICMLC),2010 Second International Conference on, vol., no., pp.109-112,9-11 Feb.2010.
[8] Kemenkes RI.2014.Situasi dan Analisis Diabetes. Jakarta : Kemenkes RI
[9] Komalasari, Wieta B. 2007. Metode Pohon Regresi Untuk Eksplorasi Data Dengan Peubah Yang Banyak Dan Kompleks. Jurnal Infomatika Pertanian Volume 16 No. 1, Juli 2007.
[10] Kundari, Eska Sarti.2015.Perbandingan Kinerja Metode Naïve Bayes dan C4.5 dalam Pengklasifikasian Penyakit Diabetes Mellitus di Rumah Sakit Kumala Siwi Kudus. SKRIPSI. Universitas Dian Nuswantoro.
[11] Kusrini, & Lutfhi,E. T. 2009.Algoritma Data Mining.Yogyakarta:Andi Offset.
[12]. Larose, D. T., 2005. *Discovering Knowledge In Data : An Introduction To Data Mining*. New Jersey : Wiley-nterscience.
[13] Lesmana,I Putu Dody.2012. Perbandingan Kinerja Decisiion Tree J48 dan ID3 dalam Pengklasifikasian Diagnosis Penyakit Diabetes Mellitus. Jurnal Teknologi dan Informatika Vol. 2 No. 2 Mei 2012.
[14] Nurhidayat,Farid.2013. Penentuan Besar Akurasi Metode Klasifikasi Menggunakan Algoritma C4.5 Berbasis Particle Swarm Optimatization Pada Prediksi Penyakit Diabetes. Tugas Akhir. Fakultas Ilmu Komputer Universitas Dian Nuswantoro Semarang.

[15] Nuriyah. 2013. Perbandingan Metode chi-square automstic interaction detection (chaid) dan classification and regression tree (cart) Dalam Menentukan Klasifikasi Alumni UIN Sunan Kalijaga Berdasarkan Masa Studi. SKRIPSI. Fakultas Sains dan Teknologi Universitas Islam Negeri Sunan Kalijaga.

[16] Patil, B.M., Joshi,R.C., Toshniwal,D.2010. Assosiation rule for classification of type 2 diabetic patients. Machine Learning And Computing (ICMLC),pp.330-334

[17] Patil, B.M., Joshi,R.C., Toshniwal,D.2010. Assosiation rule for classification of type 2 diabetic patients. *Machine Learning And Computing (ICMLC*),pp.330-334

[18] Pima Indians Diabetes Dataset, UCI Machine Learning Repository , diambil dari http://archive.ics.edu/ml/datasets/Pima+Indians+Diabetes. Diakses 29 Agustus 2016

[19] Timofeev, Roman.2004.*Classification and Regression Trees (CART) Theory and Aplications*.Humboldt University :,Berlin