YAYASAN AKRAB PEKANBARU
Jurnal AKRAB JUARA
Volume 4 Nomor 2 Edisi Mei 2019 (95-104)

9

PERBANDINGAN ALGORITMA CART DAN *NAÏVE BAYESIAN* PADA KASUS DIAGNOSIS PENYAKIT DIABETES

Hellik Hermawan, Irfan Santiko
Program Studi Sistem Informasi, STMIK AMIKOM Purwokerto
(Naskah diterima: 1 Maret 2019, disetujui: 20 April 2019)

Abstract

Diabetes mellitus is a disease that threatens serious health, can cause death and the World Health Organization (WHO) estimates that every 10 seconds there is one diabetes patient who dies of this disease. This makes researchers and practitioners focus their attention on detecting / diagnosing diabetes mellitus and preventing it because this disease can cause complications. The method used in this research is problem identification, data collection, pre-processing stage, classification method, validation and evaluation and conclusion drawing. The algorithm used in this study is CART and Naïve Bayes by using a dataset taken from the UCI Indian Pima database repository which consists of clinical data of patients who detected positive and negative diabetes mellitus. The validation and evaluation methods used are 10-cross validation and confusion. Matrix for precision, recall and F-Measure. The results of calculations that have been done, the results of the accuracy of the CART algorithm are 76.9337% with precision 0.764%, recall 0.769%, and F-Measure 0.765%. While the diabetes dataset tested by the Naïve Bayes algorithm gets an accuracy value of 73.7569% with precision 0.732%, recall 0.738%, and F-Measure 0.734%. From these results it can be concluded that to diagnose diabetes mellitus it is recommended to use the CART algorithm.
***Keywords**: Performance, Diagnosis, Algorithm*

Abstrak

Penyakit diabetes mellitus merupakan salah satu penyakit yang mengancam kesehatan yang serius, dapat mengakibatkan kematian dan *World Health Organizatation* (WHO) memperkirakan setiap 10 detik ada satu orang pasien diabetes yang meninggal karena penyakit ini.Hal ini menjadikan para peneliti dan praktisi memusatkan perhatiannya untuk mendeteksi atau mendiagnosis penyakit diabetes mellitus dan mencegahnya karena penyakit ini bisa menimbulkan komplikasi. 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 di uji 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 mendiagnosis penyakit diabetes mellitus disarankan menggunakan algoritma CART.
Kata Kunci: Kinerja, Diagnosis, Algoritma.

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 *World Health Organization* (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 (Depkes RI,2009). Berdasarkan profil kesehatan 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. Hal ini diperkuat oleh WHO (2003), penderita penyakit Diabetes Mellitus angkanya mencapai 194 juta jiwa atau 5,1 persen dari penduduk dunia usia dewasa dan pada tahun 2025 diperkirakan meningkat menjadi 333 juta jiwa. Khususnya, di Indonesia, penderita Diabetes Mellitus semakin meningkat. Pada tahun 2000, penderita Diabetes Mellitus telah mencapai 8,4 juta jiwa dan diperkirakan bahwa prevalensi penderita Diabetes Mellitus tahun 2030 di Indonesia mencapai 21,3 juta orang (Diabetes Care,2004).

II. KAJIAN TEORI

Beberapa penelitian yang terkait diagnosis penyakit diabetes mellitus yaitu oleh Triajianto, dkk (2013) yang berjudul Implementasi Sistem Klasifikasi Fuzzy Berbasis Optimasi Koloni Semut untuk Mendiagnosa Penyakit Diabetes. Hasil didapatkan tingkat akurasi algoritma dari algoritma optimasi koloni semut sebesar 78,55%. Dalam penelitian Kundari (2015) menggunakana algoritma *Naïve Bayes* dan C.4.5 menghasilkan nilai akurasi pada algoritma *Naïve bayes* sebesar 74,32% dan C.4.5 sebesar 85,13%. Hal serupa dilakukan oleh Lesmana (2012) namun hanya menggunakan algoritma J48 dan mendapatkan tingkat akurasi sebesar 74,72%.

Dari mengkaji penelitian yang telah dilakukan, oleh penulis merasa perlu. melakukan penelitian sejenis dengan menggunakan algoritma CART dan *Naïve Bayes.* Algoritma CART merupakan algoritma yang nilai akurasinya tinggi terlihat pada penelitian terkait (Gorunescu, 2011) dan algoritma *Naïve Bayes* merupakan mempunyai algoritma klasifi-kasi Bayesian memiliki kemampuan klasifikasi serupa dengan *decision tree, neural network* (Kusrini & E.T, 2009).

III. METODE PENELITIAN

Metode pengumpulan data yang digunakan dalam penelitian ini adalah studi pustaka dengan mengambil data sekunder.

1. Studi Pustaka

   Studi pustaka merupakan metode pengum-pulan data dengan tujuan untuk mengum-pulkan informasi berupa sumber tertulis (buku, majalah ilmiah, arsip, dokumen resmi dan karya ilmiah), gambar maupun dokumen elektronik yang mendukung dalam proses penulisan (Sangadji & Sopiah, 2010).

2. Data Sekunder

Data sekunder adalah sumber data peneli-tian yang diperoleh peneliti secara tidak langsung melalui media perantara (diperoleh, dicatat, atau telah diteliti oleh pihak lain). Data sekunder biasanya umumnya berupa bukti, catatan atau laporan dan historis yang telah tersusun dalam arsip (data dokumenter) yang telah dipublikasikan dan tidak dipublikasikan (Sangadji & Sopiah, 2010). Data ini digunakan karena sumber data diambil dari repository *UCI* Indian Pima Diabetes.

Alur penelitian yang digunakan yaitu:

1. Identifikasi Masalah

Proses identifikasi masalah dilakukan sebagai upaya mengetahui permasalahan serta metode yang sesuai untuk penelitian ini.

2. Pengumpulan Data

Dalam penelitian ini, data sekunder yang digunakan diambil dari repository *database* UCI Indian Pima Diabetes yang terdiri dari 768 data klinis.

3. Tahap *Pre-processing*

Tahap ini dilakukan untuk mendapat-kan data bersih dan siap untuk digunakan. Tahap *pre-processing* data meliputi identi-fikasi dan pemilihan atribut (*attribute identification and selection*), penanganan nilai atribut yang tidak lengkap (*handling missing values*), dan proses diskritisasi nilai.

4. Penggunaan Metode Klasifikasi

Langkah-langkah yang dilakukan dalam algoritma CART yaitu:

a. *Dataset* Indian Pima diklasifikasikan menggunakan algoritma CART (Simple Cart) dalam aplikasi Weka.
b. Kemudian dilakukan proses pelatihan dan pengujian dengan menggunakan metode *10-cross validation*.
c. Didapatkan hasil *classifier* dan menghasilkan *confusion matrix*.
d. Dapat melihat hasil *rule* pohon keputusan hasil output algoritma CART.
e. Dari hasil *confusion matrix* dapat dihitung nilai *precision, recall, F-Measure* dengan menjabarkan *confu-sion matrix* menjadi *of confusion*.

Sedangkan langkah – langkah yang dilakukan dalam algoritma *Naïve Bayes* adalah

a. *Dataset* Indian Pima diklasifikasikan menggunakan algoritma *Naïve Bayes* (Naive Bayes) dalam aplikasi Weka.
b. Kemudian dilakukan proses pelatihan dan pengujian dengan menggunakan metode *10-cross validation*.
c. Didapatkan hasil *classifier* dan menghasilkan *confusion matrix*.
d. Dari hasil *confusion matrix* dapat dihitung nilai *precision, recall, F-Measure* dengan menjabarkan *confu-sion matrix* menjadi *of confusion*.

5. Validasi dan Evaluasi

Dalam tahap ini dilakukan validasi dan pengukuran keakuratan hasil yang dicapai oleh model menggunakan teknik yang terdapat dalam aplikasi weka yaitu *confusion matrix* dan *cross-validation*.

6. Penarikan Kesimpulan

Tahapan selanjutnya yaitu menyim-pulkan hasil yang diperoleh dari penelitian. Algoritma CART atau *Naïve Bayes* yang memberikan hasil akurasi terbaik untuk mendiagnosis penyakit diabetes mellitus berdasarkan nilai *precision, recall, F-Measure* dari masing-masing algoritma.

1. Tahap *Pre-processing*

Setelah melakukan analisis terhadap *dataset* Indian Pima, diketahui bahwa tidak semua atribut memilki nilai yang lengkap, dimana kelengkapan nilai atribut sangat sebanyak 374 dan atribut BMI sebanyak 11. Sedangkan pada atribut *age* dan *class* 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 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 (*missing 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, DPF, *Age* dan *class*.

Tetapi, terlebih dahulu dilakukan proses diskritisasi atribut. Tujuannya untuk mempermudah pengelompokkan nilai ber-dasarkan kriteria yang telah ditetapkan. Hal ini juga bertujuan untuk menyederhanakan nilai permasalahan dan meningkatkan aku-rasi dalam proses pembelajaran (Lesmana, 2012). Atribut *glucose* dibagi menjadi tiga, yaitu *low, medium* dan *high*. Atribut DBP dibagi menjadi tiga, yaitu *normal*, *normal–to-high* dan *high* (Patil, dkk, 2010). Sedang-kan atribut BMI dikelompokkan menjadi empat, yaitu *low*, normal, *obese*, dan *severely-obese* (Patil,dkk,2010). Atribut DPF terbagi menjadi dua kelompok yaitu *low* dan *high*. Atribut class dibagi menjadi dua kelompok, yaitu positif diabetes dan negatif diabetes.

2. Proses Metode Klasifikasi

Dari hasil perhitungan dan uji coba menggunakan aplikasi weka dengan algoritma CART menghasilkan nilai akura-si sebesar 76.9337%. Nilai akurasi tersebut didapatkan dari hasil perhitungan dari *precision, recall*, dan *F-measure*. Hasil perhitungan nilai akurasi

4.2 *confusion matrix* disajikan pada table:

Tabel 4.2 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

Sedangkan apabila pengklasifikasian menggunakan algoritma *Naïve Bayes* menggunakan aplikasi weka menghasilkan nilai akurasi sebesar 73.7569%. Nilai akurasi didapatkan dari hasil perhitungan *precision, recall* dan *F-Measure*. Hasil perhitungan nilai akurasi bersadarlan *confusion matrix* disajikan pada tabel 4.3 sebagai berikut:

Tabel 4.3 nilai akurasi berdasarkan *confusion matrix*

Class	*Precision*	*Recall*	*F-Measure*
Tested_negative	0.783	0.829	0.806
Tested_positive	0.633	0.562	0.596
Weighted Avg	0.732	0.738	0.734

Tabel 4.4 Perbandingan Hasil Akurasi CART dan *Naïve Bayes*

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

Perbedaan akurasi yang diperoleh dengan menggunakan algoritma CART dan *Naïve Bayes* sebesar 3.1768 %. Waktu yang digunakan pada saat *running dataset* pada aplikasi Weka juga berbeda antara algoritma CART dan *Naïve 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.

3. Validasi dan Evaluasi

Tabel 4.5 dan 4.6 merupakan tabel hasil *confusion matrix* dari pengujian *dataset*

menggunakan algoritma CART dan *Naïve Bayes* dengan *10-fold cross validation.*

Tabel 4.5 *Confusion matrix* CART

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

Tabel 4.6 *Confusion matrix Naïve Bayes*

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

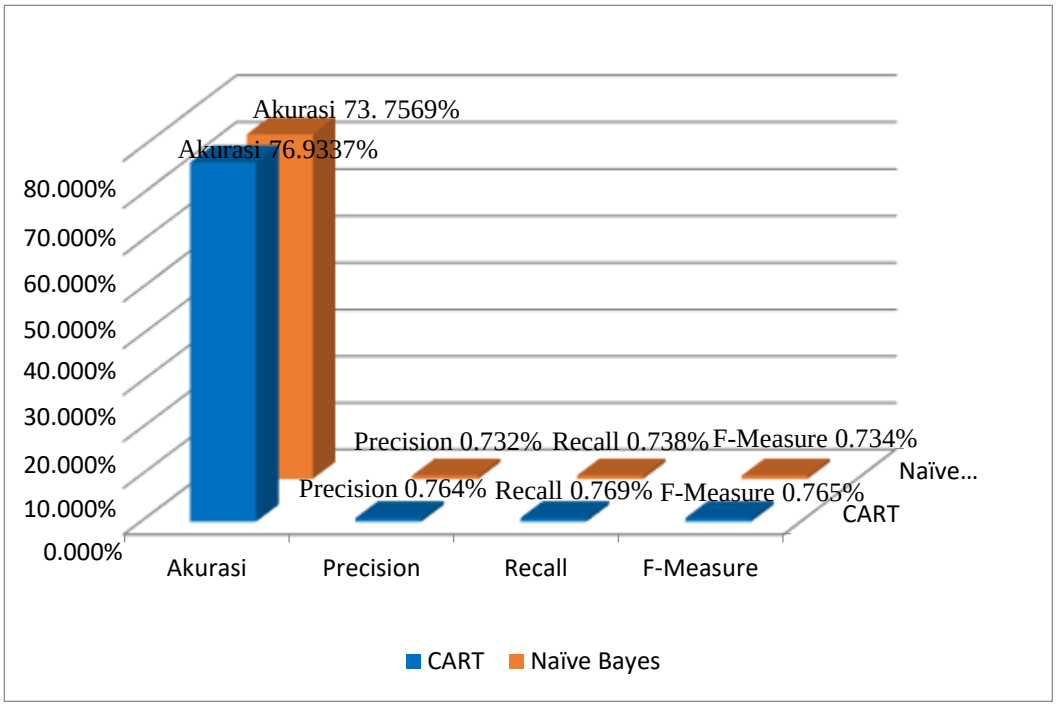
Dari tabel 4.5 terlihat bahwa jumlah data hasil bentukan *rule* yang terkena penyakit diabetes mellitus yang sama dengan data *testing* yang juga terkena diabetes sebanyak 407. Kemudian, jumlah data hasil bentukan *rule* yang tidak terkena penyakit diabetes mellitus dengan data *testing* yang terkena 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 4.6 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.

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

V. KESIMPULAN

Setelah melalui tahapan *pre-processing* tersebut dan dilakukan perhitungan dengan menggunakan dua algoritma yaitu CART dan *Naïve Bayes*, serta sudah dilakukan tahap evaluasi dengan *confusion matrix* didapatkan hasil akurasi sebesar 76.9337% dengan *precision* 0.764%, recall 0.769% dan F-Measure 0.765% pada algoritma CART dan sedangkan pada hasil akurasi *Naïve Bayes* yaitu 73.7569% dengan nilai *precision* 0.732%, recall 0.738% dan F-Measure 0.734%. Selanjutnya hasil *confusion matrix* dari dua algoritma yaitu algoritma CART dan *Naïve Bayes* dapat dilihat dengan grafik hasil akurasi dari perhitungan dua algortima tersebut pada gambar 5.1.



Gambar 5.1 Perbandingan hasil akurasi algoritma CART dan *Naïve Bayes*

DAFTAR PUSTAKA

Bramer, M. 2007. *Principles Of Data Mining*. London: Springer

Depkes,R.I. 2009. Profil Kesehatan Indonesia. Jakarta: Depkes RI

Diabetes Care. 2004. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030.

Gorunescu,F. 2011. *Data mining Concepts, Models and Techniques*. Verlan Berlin Heidelberg: Spinger

Han,J., & Kamber,M. 2006. *Data Mining Concepts And Techniques*. Verlag Berlin Heidelberg: Spinger

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.*

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.

Kemenkes RI.2014.Situasi dan Analisis Diabetes. Jakarta: Kemenkes RI

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*.

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.

Kusrini, & Lutfhi,E. T. 2009. *Algoritma Data Mining*.Yogyakarta: Andi Offset.

Larose, D. T., 2005. *Discovering Knowledge In Data: An Introduction To Data Mining*. New Jersey: Wiley-nterscience.

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.

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.

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.

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

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

Prasetyo, Eko. 2012. *Data Mining Konsep dan Aplikasi Menggunakan Matlab*. Yogyakarta: Penerbit Andi

RISKESDAS, Badan Penelitian dan Pengembangan Kesehatan Kementrian Kesehatan RI. 2013.

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

Sangadji.E.M dan Sopiah. 2010. *Metodologi Penelitian*. Yogyakarta : Andi

Sulianta,F , dan Juju,D. 2010. *Data Mining Meramalkan Bisnis Perusahaan*. Jakarta: Elex Media Komputindo.

Susanto, S., dan Suryadi, D.. 2010. *Pengantar Data Mining*. Yogyakarta: Andi

Triajianto, J., Purwananto, Y., dan Soelaiman, R. 2013. Implementasi Sistem Klasifikasi Fuzzy Berbasis Optimasi Koloni Semut untuk Diagnosa Penyakit Diabetes. *Jurnal Teknik Pomits Vol. 2, No. 1 (2013) ISSN : 2337 – 3539 (2301-9271 Print).*

Trisnawati, Setyorogo. 2013. Faktor Risiko Kejadian Diabetes Melitus Tipe II Di Puskesmas Kecamatan Cengkareng Jakarta Barat Tahun 2012. *Jurnal Ilmiah Kesehatan, 5 (1); Januari 2013*

WEKA, Machine Learning Group at University of Waikato,diambil dari http://www.cs.waikato.ac.nz/ml/weka/ . Diakses 29 Agustus 2016.

Written, I. H., Frank, E., & Hall, M.A. 2011. *Data Mining Practical Machine*. Burlington: Elsevie.