PERBANDINGAN KINERJA ALGORITME C4.5 DAN NAÏVE BAYES MENGKLASIFIKASI PENYAKIT DIABETES

Hendra Marcos[1], Hengky Setiawan Utomo[2]

[1,2]Program Studi Teknik Informatika, STMIK Amikom Purwokerto
Jl. Pol. Soemarto Depan SPN Watumas Purwokerto
Telp. (0281) 623321 Fax. (0281) 623196
e-mail :[1]hendra.marcos@amikompurwokerto.ac.id

ABSTRACT

Diabetes or can be called with diabetes or blood sugar disease is a disease that is hard to cure but can be controlled blood sugar levels. This causes people with diabetes is increasing every year. This study aims to determine which algorithm that has the best classification accuracy, so that it can be used to assist in classifying whether a person has diabetes or not. The data used is the Pima Indians Diabetes dataset obtained from the UCI machine learning. Processing of data mining is divided into two stages, namely stage of data preprocessing and feature selection. Results of the research that has been done, C4.5 algorithm has an accuracy of 73.82% and increased to 74.87%, subsequent to the selection of attributes. While naïve Bayes has an accuracy rate of 76.30% and increased to 77.47%. The end result of this research is naïve bayes algorithm is better than C4.5 algorithms because it has a better accuracy rate

Keywords—*C4.5, Naïve Bayes, Diabetes*

ABSTRAK

Diabetes atau dapat disebut dengan kencing manis atau penyakit gula darah merupakan penyakit yang sukar disembuhkan namun kadar gula darah dapat dikontrol. Hal ini menyebabkan penderita penyakit diabetes semakin meningkat setiap tahunnya. Penelitian ini bertujuan untuk mengetahui algoritme mana yang memiliki tingkat akurasi klasifikasi paling baik, sehingga bisa digunakan untuk membantu dalam mengklasifikasi apakah seseorang terkena penyakit diabetes atau tidak. Data yang digunakan adalah *Pima Indians Diabetesdataset* yang diperoleh dari UCI *machine learning*. Pengolahan *data mining* dibagi menjadi dua tahap, yaitu tahap *preprocessingdata* dan seleksi fitur. Hasil penelitian yang telah dilakukan, algoritme C4.5 memiliki akurasi sebesar 73.82% dan meningkat menjadi 74,87% setelah dilakukannya seleksi atribut. Sedangkan *naïve bayes* memiliki tingkat akurasi sebesar 76,30% dan meningkat menjadi 77,47%. Hasil akhir dari penelitian ini adalah algoritme *naïve bayes* lebih baik dari pada algoritme C4.5 karena memiliki tingkat akurasi yang lebih baik.

Kata Kunci—Perbandingan kinerja algoritme, C4.5, Naïve Bayes, Diabetes

I. PENDAHULUAN

Diabetes atau dapat disebut dengan penyakit gula darah adalah salah satu jenispenyakit kronis yang mempunyai tanda awal berupa meningkatnya kadar gula darah akibat adanya gangguan sistem

metabolisme didalam tubuh. Diabetes sukar untuk disembuhkan namun kadar gula darah dapat dikontrol. Menurut laporan WHO [1][2], Indonesia menempatiurutan ke-empat terbesar dari jumlah penderita diabetes mellitus (DM) dengan prevalensi 8,6% dari total penduduk dan WHO memprediksi kenaikan jumlah penyandang DM di Indonesia dari 8,4 juta pada tahun 2000 menjadi sekitar 21,3 juta pada tahun 2030.

Sedangkan *International Diabetes Federation* (IDF) pada tahun 2009 memprediksi kenaikan jumlah penyandang DM dari 7 juta pada tahun 2009 menjadi 12 juta pada tahun 2030[3]. Dari laporan tersebut menunjukkan peningkatan jumlah penyandang DM sebanyak 2-3 kali lipat pada tahun 2030 [4]. Seperti halnya WHO dan IDF, dari hasil riset yang telah dilakukan oleh Riset Kesehatan Dasar (Riskesdas) pada tahun 2007 dan 2013 proporsi DM meningkat hampir dua kali lipat[5].

Dari banyaknya jumlah penduduk yang terkena penyakit diabetes yang semakin tahun makin meningkat maka diperlukan sebuah diagnosis komputer dengan menggunakan algoritme tertentu untuk mengklasifikasi apakah seseorang terkena penyakit diabetes ataupun tidak berdasarkan data-data pasien dengan teknik *data mining*[6].

Beberapa penelitian yang telah dilakukanmenggunakan *machine learning*dari dataset yang sama yaitu UCI *machine learning,* diantaranya menggunakan *extreme machine learning* [7], *bayesian*[8] menunjukkan hasil yang tidak transparan dan relatif hasil akurasi masih rendah 72,3%. Metode *artificial neural network* juga telah digunakan pada penelitian [9], yang menggunakan *principal component analysis* (PCA) untuk seleksi fitur/atribut nya dengan hasil akurasi sekitar 66,8% s.d 68%. Maka dari itu pada penelitian ini diusulkan teknik algoritme C.45 dan *naive bayes* dan membandingkan kinerja kedua algoritme dalam mengklasifikasi dengan evaluasi *accuracy, precision dan recall*.

II. METODE PENELITIAN

2.1 Pengukuran Kinerja

Pengukuran kinerja algoritme menggunakan bantuan perangkat lunak *data mining* yaitu menggunakan aplikasi weka. Algoritme yang diuji yaitu C4.5 dengan *Naïve Bayes*. Pengukuran kinerja dilihat dari *confusion matrix* dengan mencari nilai *precision, recall* dan nilai *accuracy*.

a. *Precision* merupakan tingkat ketepatan antara informasi yang diminta oleh pengguna dengan jawaban yang diberikan oleh sistem.

b. *recall* adalah tingkat keberhasilan sistem dalam menemukan kembali sebuah informasi.
c. *accuracy* didefinisikan sebagai tingkat kedekatan antara nilai prediksi dengan nilai aktual.

Secara umum *precision, recall* dan *accuracy* dapat dirumuskan seperti pada tabel 1 berikut:

Tabel 1 Rumus *precision*, *recall* dan *accuracy*

	Classified Positive	***Classified Negative***
Actual Positive	TP (*True Positive*)	FN (*False Nagative*)
Actual Negative	FP (*False Positive*)	TN (*True Negative*)

a. $Precission = \frac{TP}{TP+FP} \times 100\%$

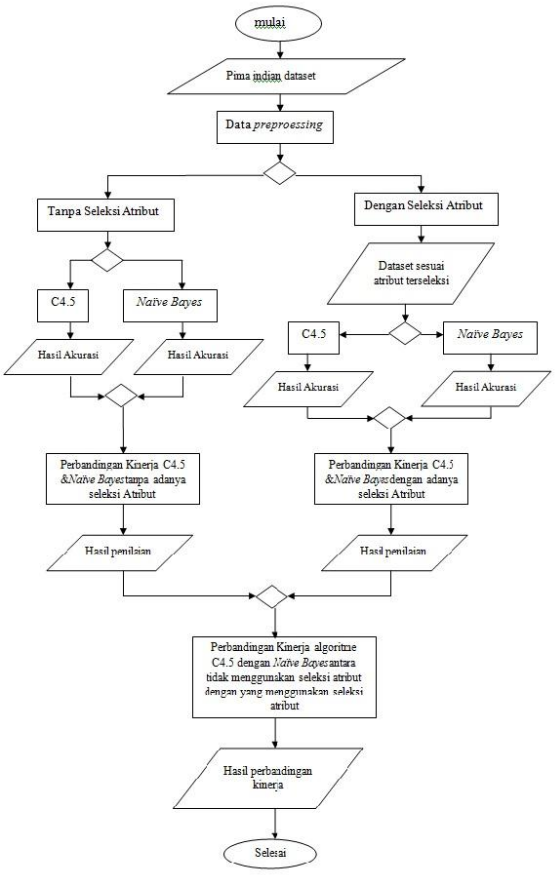
b. $Recall = \frac{TP}{TP+FN} \times 100\%$

c. $Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \times 100\%$

Pengukuran kinerja algoritme akan dilakukan dengan dua cara, yaitu untuk pertama akan membandingkan kinerja kedua algoritme tanpa adanya seleksi atribut dan yang kedua dengan adanya seleksi atribut, yaitu dengan menggunakan *correlation-based feature selection* (CFS) pada aplikasi WEKA. Metode pengujian sistem menggunakan *cross validation* 10, baik *dataset* yang belum terseleksi atribut maupun yang sudah dilakukan seleksi atribut.

2.2. *Dataset*

Dataset yang digunakan yaitu berupa data sekunder *pima indians diabetes*diambil dari *repository* UCI *machinelearning*. Langkah-langkah yang dilakukanpada penelitian ini, dapat dilihat pada gambar 1



Gambar 1 Diagram alur penelitian

2.3. Uji Hipotesis

Setelah didapatkan hasil perbandingan kinerja langkah selanjutnya yaitu dengan melakukan uji hipotesis. Pengujian uji hipotesis akan dibagi menjadi dua langkah :

a. Pengujian hipotesis pertama

Pengujian hipotesis pertama dilakukan dengan membandingkan rata-rata nilai kinerja algoritme C4.5 dengan *naïve bayes.*

b. Pengujian hipotesis kedua

Pengujian hipotesis kedua dilakukan dengan uji t data berpasangan (*paired sample t test*). Penghitungan dilakukan melalui prosedur *paired sample t test* dengan *software* SPSS *for windows* versi 22(Sugiyono, 2012).

III. HASIL DAN PEMBAHASAN

3.1.*Dataset*

Dataset terdiri dari 8 atribut dan 768 *instance* yang semuanya berasal dari jenis kelamin wanita dengan umur sekurang-kurangnya 21 tahun[10].

Tabel 2. Atribut dataset diabetes Pima Indians

Atribut	Deskripsi	Satuan	Tipe Data
Number of times pregnant	Banyaknya kehamilan	-	*Numerik*
Plasma glucose concentration	Kadar glukosa dua jam setelah makan	mg/dL	*Numerik*
Diastolic blood pressure	Tekanan darah	mm Hg	*Numerik*
Triceps skin fold Thickness	Ketebalan kulit	mm	*Numerik*
Insulin	Insulin	mu U/ml	*Numerik*
Body mass index	Berat tubuh	Kg/m^2	*Numerik*
Diabetes pedigree Function	Riwayat diabetes dalam keluarga	-	*Numerik*
Age	Umur	Years	*Numerik*
Class Variable	Positif diabetes (1) dan negative diabetes(0)	-	*Nominal*

3.2.*Preprocessing data*

Preprocessing data yang dilakukan yaitu dengan menormalisasi atribut *class* dengan nilai 0 menjadi *tested_negative* sedangkan dengan nilai 1 menjadi *tested_positive*, sehingga ketika sudah pembuatan *file* arff maka dapat dibuka menggunakan aplikasi weka dan data dapat diproses.

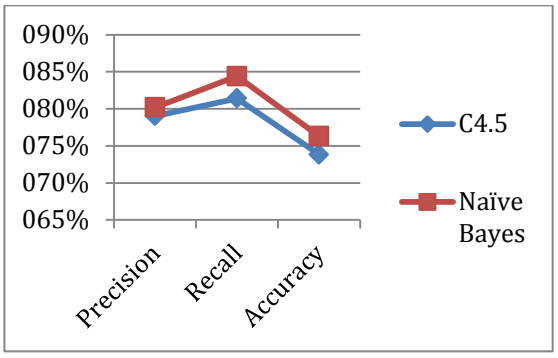
Perbandingan kinerja algoritme sebelum proses seleksi atribut

Pada langkah ini perbandingan yang dilakukan yaitu dengan menghitung nilai *recall*, *precision* dan *accuracy* tanpa adanya seleksi atribut pada dataset yang digunakan. Dari perhitungan yang dilakukan mendapatkan hasil seperti pada tabel 3 berikut :

Tabel 3. Hasil perhitungan tanpa seleksi atribut

Algoritme	*Precision*	*Recall*	*Accuracy*
C4.5	79.03%	81.40%	73.82%
Naïve Bayes	80.20%	84.40%	76.30%

Pada tabel 3 menunjukkan bahwa nilai *precision,recall* dan *accuracy*yang tertinggi algoritme*naïve bayes*. Hal ini membuktikan bahwa kinerja algoritme *naïve bayes* lebih unggul jika dibandingkan dengan kinerja algoritme C4.5. Untuk lebih jelasnya dapat dilihat pada gambar 2 berikut.



Gambar 2. Hasil perbandingan kinerja algoritme

Perbandingan kinerja algoritme sesudah proses seleksi atribut

Pengukuran kenerja selanjutnya yaitu dengan meyeleksi jumlah atribut dari *dataset* yang ada dengan menggunakan fitur "*Select attributes*" pada aplikasi weka. Seleksi atribut yang digunakan yaitu *Correlation-based Featur Selection* (CFS). Dari *dataset* yang ada dengan jumlah data 768 dan memiliki 9 atribut akan diseleksi fitur untuk mencari fitur atribut yang mempunyai korelasi antar fitur rendah namun mempunyai korelasi yang tinggi terhadap kelas.

```
Selected attributes: 2,6,7,8 : 4
                     plas
                     mass
                     pedi
                     age
```

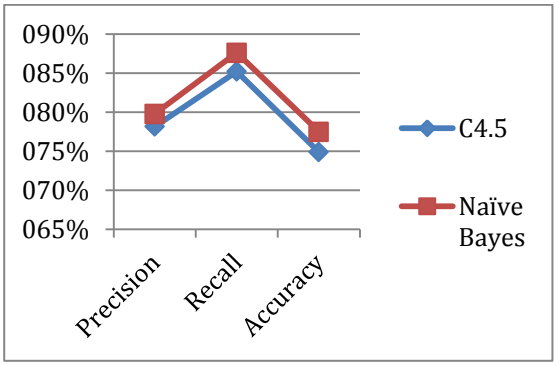
Gambar 3. Hasil seleksi atribut

Ketika *dataset* sudah diubah sesuai atribut yang terseleksi maka langah selanjutnya ialah mengukur kinerja kedua algoritme dengan mencari nilai *precision*, *recall* dan *accuracy* dari *confussion matrix*yang diperoleh dari pemrosesan *dataset* pada aplikasi weka. Proses pengukuran, *dataset* yang digunakan yaitu yang sudah diubah sesuai atribut yang terseleksi.Perhitungan yang dilakukan mendapatkan hasil seperti pada tabel 4 berikut :

Tabel 4. Hasil perhitungan setelah seleksi atribut

Algoritme	*Precision*	*Recall*	*Accuracy*
C4.5	78.16%	85.20%	74.87%
Naïve Bayes	79.78%	87.60%	77.47%

Dari data yang ada pada tabel 4menunjukan bahwa nilai *precision*,*recall* dan *accuracy* yang tertinggi diperoleh oleh Naïve Bayes. Hal ini membuktikan bahwa kinerja algoritme Naïve Bayes lebih unggul jika dibandingkan dengan kinerja algoritme C4.5. Untuk lebih jelasnya dapat dilihat pada gambar 4 berikut.



Gambar 4. Hasil perbandingan kinerja algoritme setelah seleksi atribut

Uji t

Digunakan dalam penelitian ini untuk menguji perbedaan antara algoritme C4.5 dan *naïve bayes*. Data yang digunakan untuk pengujian adalah data

kinerja algoritme sesudah seleksi atribut.Uji t yang digunakan adalah uji t data berpasangan (*paired sample t test*). *Software* yang digunakan untuk menghitung nilai t adalah *SPSS for windows* versi 22.

Tabel 5. Statistik deskriptif data penelitian

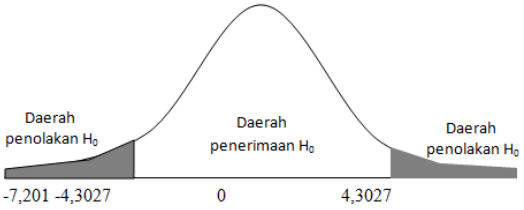
Algoritme	Rata-rata	N
C4.5	,7943	3
Naïve Bayes	,8163	3

Data tersebut selanjutnya diuji dengan uji t. Hasil uji t dapat dilihat pada tabel 6.Berdasarkan tabel Tabel 6 dapat diketahui rata-rata selisih sebesar -0,02200. Hal ini menunjukkan rata-rata nilai algoritme C.4.5 lebih kecil dibandingkan dengan rata-rata nilai Algoritme *naïve bayes* dengan dengan rata-rata selisih sebesat 2,2 persen. Uji t digunakan untuk menguji signifikansi selisih kinerja algoritme C4.5 dan algoritme *naïve bayes*.

Tabel 6. Statistik deskriptif data penelitian

Pasangah Algoritme	Paired Differences Mean	t	df	Sig. (2-tailed)
C4.5 - Naïve Bayes	-,02200	-7,201	2	,019

Nilai uji t sebesar -7,201 dengan signifikansi 0,019. Nilai signifikansi tersebut lebih kecil dari α= 0,05, sehingga H_0 ditolak dan H_1 diterima. Hal ini berarti ada perbedaan antara nilai kinerja Algoritme C4.5 dan algoritme *naïve bayes*. Perbedaan tersebut menunjukkan nilai kinerja algoritme *naïve bayes* lebih besar dibandingkan dengan kinerja algoritme C4.5 dengan selisih yang signifikan pada tingkat keyakinan 95%. Kurva uji t tersebut dapat dilihat pada gambar 5.



Gambar 5. Daerah penerimaan dan penolakan H_0 uji t

Pengujian hipotesis

a. Pengujian hipotesis pertama

Hipotesis pertama menyatakan algoritme *naïve bayes* memiliki tingkat akurasi ketepatan lebih baik dibandingkan dengan algoritme C4.5 dalam menentukan potensi terjadinya penyakit diabetes. Berdasarkan hasil analisis data dapat diketahui rata-rata nilai kinerja algoritme C4.5 sebesar 79,43%, sedangkan nilai rata-rata kinerja algoritme *naïve bayes* sebesar 81,63%. Berdasarkan nilai rata-rata tersebut dapat diketahui nilai rata-rata kinerja algoritme *naïve bayes* lebih tinggi dibandingkan dengan nilai

kinerja algoritme C4.5, sehingga dapat disimpulkan algoritme *naïve bayes* memiliki tingkat akurasi ketepatan lebih baik dibandingkan dengan algoritme C4.5 dalam menentukan potensi terjadinya penyakit diabetes, sehingga hipotesis pertama diterima.

b. Pengujian hipotesis kedua

Hipotesis kedua menyatakan terdapat perbedaan nilai kinerja antara algoritme C4.5 dengan *naïve bayes.* Berdasarkan hasil pengujian dengan uji t data berpasangan diperoleh kesimpulan ada perbedaan antara nilai kinerja Algoritme C4.5 dan algoritme *naïve bayes*. Perbedaan tersebut menunjukkan nilai kinerja algoritme *naïve bayes* lebih besar dibandingkan dengan kinerja algoritme C4.5 dengan selisih yang signifikan pada tingkat keyakinan 95%. Berdasarkan hasil uji t ini maka dapat disimpulkan terdapat perbedaan nilai kinerja antara algoritme C4.5 dengan *naïve bayes,* sehingga hipotesis kedua dapat diterima..

IV. SIMPULAN

Dari pengukuran kinerja kedua algoritme yang telah dilakukan, algoritme*naïve bayes* memiliki kinerja (*precision*, *recall* dan *accuracy*) lebih baik jika dibandingkan dengan algoritme C4.5. Ketika dilakukan seleksi atribut terjadi penurunan kinerja pada *precision* untuk masing masing algoritme, namun pada kinerja *recall* dan *accuracy* dengan dilakukannya seleksi atribut mengalami peningkatan. Hasil akurasi yang diperoleh dari perhitungan yang telah dilakukan, algoritme C4.5 memiliki akurasi sebesar 73.82% dan meningkat menjadi 74.87% setelah dilakukannya seleksi atribut. Sedangkan *naïve bayes* memiliki tingkat akurasi sebesar 76.3% dan meningkat menjadi 77.47%. Dari semua perhitungan yang telah dilakukan, dapat disimpulkan algoritme *naïve bayes* memiliki tingkat akurasi ketepatan lebih baik dibandingkan dengan algoritme C4.5 dalam mengklasifiksi risiko terjadinya penyakit diabetes. Hasil analisis statistik antara nilai kinerja algoritme C4.5 dan *naïve bayes* dapat disimpulkan terdapat perbedaan nilai kinerja antara algoritme C4.5 dengan *naïve bayes*.

PENELITIAN LANJUTAN

Dari penelitian yang dilakukan, peneliti menggunakan metode pengujian sistem *cross validation*, dan metode seleksi fitur *correlation-based feature selection* (CFS), sehingga untuk peneliti lain bisa menggunakan metode pengujian sistem dan metode seleksi fitur yang lain.

DAFTAR PUSTAKA

[1] S. Kumari dan A. Singh, "A Data Mining Approach for the Diagnosis of Diabetes Mellitus," *IEEE* , no. 978-1-4673-4603-0, pp. 373-375, 2012.

[2] WHO, "WHO," 2016. [Online]. Available: http://goo.gl/a6BToB. [Diakses 20 Januari 2016].

[3] I. D. Federation, "International Diabetes Federation," 2016. [Online]. Available: http://www.idf.org/who-we-are. [Diakses 2 Mei 2016].

[4] PERKENI, "Perkumpulan Endokrinologi Indonesia," 2011. [Online]. Available: http://perkeni.freeservers.com/kons_dm.html. [Diakses 23 Februari 2016].

[5] "Riset Kesehatan Dasar," 2016. [Online]. Available: http://goo.gl/mX0hCm. [Diakses 17 Mei 2016].

[6] E. Prasetyo, Data Mining Konsep dan Aplikasi Menggunakan Matlab, Yogyakarta: Andi Publisher, 2012.

[7] J. J. Pangaribuan dan Suharjito, "Diagnosis of Diabetes Mellitus Using Extreme Learning Machine," dalam *International Conference on Information Technology Systems and Innovation (ICITSI)*, 2014.

[8] Y. Guo, G. Bai dan Y. Hu, "Using Bayes Network for Prediction of Type-2 Diabetes," *International Conference for Internet Technology and Secured Transactions IEEE,* 2012.

[9] T. Jayalakshmi dan Dr.A.Santhakumaran, "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Network," dalam *International Conference on Data Storage and Data Engineering*, 2010.

[10] "UCI repository dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes. [Diakses 4 April 2016].