Analisis Hasil Implementasi Algoritma C4.5 dan *Naïve Bayes Clasifier* Dalam Menentukan Lokasi Prioritas Penyuluhan Program Keluarga Berencana

Febrina Sari[1], David Saro[2]
[1)] Program Studi Informatika, Sekolah Tinggi Teknologi Dumai
Jl. Utama Karya Bukit Batrem II
Email: febri_ghaniya@yahoo.co.id

ABSTRAK

Pemerintah memberikan serangkaian usaha untuk menekan laju pertumbuhan penduduk agar tidak terjadi ledakan penduduk yang lebih besar. Salah satu cara yang dilakukan oleh pemerintah adalah dengan menggalakkan Program Keluarga Berencana (KB). Program Keluarga Berencana yang dicanangkan oleh pemerintah untuk menekan angka kelahiran yang tinggi ini belum sepenuhnya terlaksana dengan baik karena lokasi penyuluhan program KB yang belum tepat sasaran. Oleh karena itu diperlukan suatu sistem yang dapat membantu Badan Kependudukan dan Keluarga Berencana Nasional (BKKBN) dalam menentukan lokasi Prioritas Penyuluhan Program Keluarga Berencana agar penyuluhan tepat sasaran. Metode yang digunakan dalam penelitian ini adalah Algoritma C4.5 dan *Naïve Bayes Clasifier* yang mana keduanya merupakan metode pada teknik klasifikasi data mining. Hasil proses klasifikasi dalam menentukan lokasi prioritas penyuluhan program keluarga berencana menunjukkan bahwa Algoritma C4.5 memiliki tingkat akurasi 2% lebih baik dibandingkan dengan metode *Naïve Bayes*.

Kata kunci: *Data Mining*, Algoritma C4.5, *Naïve Bayes*, Keluarga Berencana.

ABSTRACT

The government provides a series of efforts to curb the rate of population growth in order to avoid a larger population explosion. One way that the government is doing is by promoting Family Planning Program (KB). The family planning program launched by the government to reduce the high birth rate has not been fully implemented as the location of extension program that has not been well targeted. Therefore needed a system that can help National Population and Family Planning Agency (BKKBN) in determining location of Priority of Counseling Program Family Planning in order to counseling on target. The method used in this research is Algorithm C4.5 and Naïve Bayes Clasifier which are both methods of data mining classification techniques. The result of the classification process in determining the priority location of the family planning extension program shows that Algorithm C4.5 has 2% accuracy level better than Naïve Bayes method.

Keywords: *Data Mining, Algoritma C4.5, Naïve Bayes, Family Planning.*

Pendahuluan

Program keluarga berencana yang dicanangkan oleh pemerintah untuk membudayakan norma keluarga kecil bahagia dan sejahtera serta menekan angka kelahiran yang tinggi ini belum sepenuhnya terlaksana dengan baik karena penyuluhan program keluarga berencana diadakan di beberapa daerah yang tingkat kelahirannya rendah, jika penetapan lokasi penyuluhan yang selalu tidak tepat sasaran dikawatirkan program ini tidak dapat mencapai tujuan. Oleh karena itu penetapan lokasi prioritas penyuluhan program keluarga berencana yang tepat sangatlah penting sehingga diperlukan suatu sistem dan metode *data mining* yang dapat menyelesaikan permasalahan tersebut dan mengolah *database* kelahiran bayi yang ada di Dinas Penduduk dan Catatan Sipil yang hasil akhirnya dapat digunakan BKKBN dalam menentukan lokasi prioritas Penyuluhan Program KB.

Data mining adalah suatu proses menemukan hubungan yang berarti, pola dan kecenderungan dengan memeriksa sekumpulan besar data yang tersimpan dalam penyimpanan dengan menggunakan teknik pengenalan pola seperti teknik statistik dan matematika. *Data mining* merupakan gabungan dari beberapa disiplin ilmu yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistik, database, dan visualisasi untuk penanganan permasalaan pengambilan informasi dari *database* yang besar (Larose, 2005).

Secara singkat bahwa Algoritma C4.5 dan *Naive Bayes Clasifier* merupakan metode klasifikasi pada *text mining*. Klasifikasi adalah proses menemukan kumpulan pola atau fungsi-fungsi yang mendeskripsikan dan memisahkan kelas data satu dengan lainnya, untuk dapat digunakan dalam memprediksi data yang belum memiliki kelas data tertentu (Han, 2006). Klasifikasi pertama kali diterapkan pada bidang tanaman yang mengklasifikasikan suatu spesies tertentu, seperti yang dilakukan oleh Carolus von Linne (atau dikenal dengan nama Carolus Linneaeus) yang pertama kali mengklasifikasikan spesies berdasarkan karakteristik fisik (Mardi, 2014).

Penelitian mengenai komperasi metode dalam *data mining* telah banyak dilakukan sebelumnya dengan jumlah data dan atribut yang berbeda-beda. Salah satu penelitian yang dilakukan oleh Defiyanti. Menggunakan metode C4.5 dan ID3 dalam mengklasifikasi spam mail dengan jumlah data dan atribut yang bervariasi. Dari penelitian ini didapat sebuah hasil bahwa nilai akurasi tertinggi yang diperoleh oleh algoritma C4.5 adalah sebesar 72,38% dengan jumlah atribut sebesar 52, sedangkan untuk algoritma ID3 memperoleh nilai akurasi tertinggi sebesar 73,20% pada jumlah atribut sebesar 58.

Penelitian berikutnya dilakukan oleh Iskandar. Perbandingan Akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan *Naive Bayes Clasifier*. Dalam penelitian ini disimpulkan bahwa Algoritma C4.5 memiliki tingkat akurasi yang lebih baik 3% dibandingkan dengan metode *Naive Bayes*. Meskipun demikian kedua metode memiliki nilai presisi dan *recall* yang tidak jauh berbeda, hal ini dikarenakan, kedua metode menggunakan jumlah fitur/atribut yang sama.

Dalam penelitian ini penulis juga akan mencoba melakukan perbandingan terhadap hasil yang diperoleh Algoritma C4.5 dan *Naive Bayes Clasifier* dalam menganalisa lokasi prioritas penyuluhan program KB, dan menggunakan *Tools* WEKA. WEKA adalah sebuah paket *tools machine learning* praktis. WEKA merupakan singkatan dari "*Waikato Environment for Knowledge Analysis*" yang dibuat di Universitas Waikato New Zealand untuk penelitian, pendidikan dan berbagai aplikasi. WEKA mampu menyelesaikan masalah-masalah data mining di dunia nyata, khususnya klasifikasi yang mendasari pendekatan *machine learning*. Perangkat lunak ini ditulis dalam hirarki *class java* dengan metode berorientasi objek dan dapat berjalan hampir di semua platfrom (Bouckaert, 2008).

Implementasi Algoritma C4.5 dan *Naïve Bayes Clasifier*

Secara umum algoritma C4.5 dalam membangun pohon keputusan langkah-langkahnya adalah sebagai berikut (Bramer, 2007).

a. Pilih atribut sebagai akar
b. Buat cabang untuk tiap-tiap nilai
c. Bagi kasus dalam cabang
d. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama.

Untuk memilih atribut sebagai akar, didasarkan pada nilai *Gain* tertinggi dari atribut-atribut yang ada. Untuk menghitung *Gain* digunakan rumus seperti tertera dalam persamaan (1).

$$Gain(S, A) = Entropy(S) - \sum_{i=1}^{n} \frac{|Si|}{|S|} * Entropy(Si) \quad (1)$$

Keterangan:

S	: himpunan kasus
A	: atribut
n	: jumlah partisi atribut A
$\|S_i\|$	: jumlah kasus pada partisi ke-i
$\|S\|$	: jumlah kasus dalam S

Sementara itu perhitungan nilai *entropy* dapat dirumuskan dalam persamaan (2) berikut ini.

$$Entropy(S) = \sum_{i=1}^{n} - pi * Log_2 pi \quad (2)$$

Keterangan:

S	: himpunan kasus
A	: fitur
n	: jumlah partisi S
pi	: proporsi dari Si terhadap S

Proses pengulangan pada metode decision tree ini akan berhenti apabila:

1. Semua data telah terbagi rata
2. Tidak ada lagi atribut yang bisa dibagi lagi
3. Tidak ada data record dalam cabang yang kosong

Naïve Bayes Clasifier (NBC) merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan inggris Thomas bayes, yaitu memprediksi peluang dimasa depan berdasarkan pengalaman dimasa sebelumnya sehingga dikenal sebagai teorema bayes. Teorema tersebut diasumsikan memiliki atribut saling bebas (independen). Dasar teori yang digunakan adalah teorema bayes yang ditunjukkan oleh persaman (3).

P(a | b) = (p(b | a) * p(a)) / p(b)

Keterangan: peluang a sebagai b, diperoleh dari peluang b saat a, peluang a dan peluang b.

Metode Penelitian

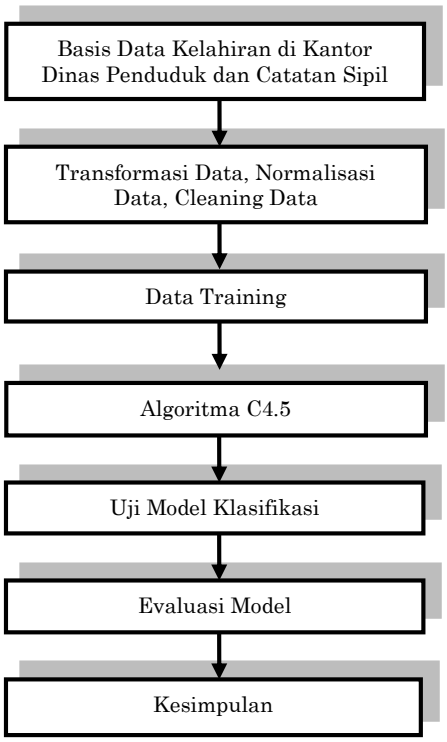
Metodologi penelitian merupakan urutan-urutan yang dilakukan dalam sebuah penelitian. Metodologi penelitian ini bertujuan agar penelitian bisa lebih terkonsep dan terstruktur, sehingga setiap tahapan akan dapat dilihat pencapaiannya sesuai dengan tujuan yang diharapkan terhadap penelitian tersebut.

Sampel dan Sumber Data Penelitian

Sampel dalam penelitian ini adalah data kelahiran bayi pada setiap Kelurahan yang ada di Kecamatan Dumai Timur Tahun 2016. Karena kesuksesan program KB disuatu daerah ditentukan dari tinggi rendahnya angka kelahiran bayi di daerah tersebut. Data kelahiran bayi di setiap Kelurahan yang ada di Kecamatan Dumai Timur bersumber dari *database* Aplikasi Sistem Informasi Administrasi Kependudukan Kota Dumai.

Kerangka Kerja Penelitian

Kerangka kerja ini merupakan langkah-langkah yang akan dilakukan dalam penyelesaian masalah yang akan dibahas, sedangkan metode yang digunakan dalam penelitian ini bertujuan untuk memperlihatkan bagaimana sebuah model klasifikasi *data mining* bias memberikan solusi untuk mengklasifikasikan lokasi prioritas penyuluhan program keluarga berencana berdasarkan atribut yang ada. Adapun metode dan kerangka kerja ini ditunjukkan pada Gambar 1.



Gambar 1 Kerangka kerja penelitian

Berdasarkan kerangka kerja pada Gambar 1 maka masing-masing langkahnya dapat diuraikan seperti berikut ini:

1. Transformasi Data
   Basis data yang diperoleh masih berupa data yang mengandung banyak atribut yang tidak diperlukan sehingga perlu dilakukan transformasi data dengan membuang sebagian atribut yang tidak memiliki kaitan dengan topik penelitian.
2. Normalisasi Data
   Proses normalisasi data yang dimaksud adalah mengubah jenis skala pengukuran yang semula berbentuk numerikal menjadi nominal.
3. *Cleaning* Data
   Proses Pembersihan data yang tidak relevan termasuk data *missing* dalam atribut (Yuhefizar, 2013).
4. Training Data
   Proses pelatihan data diambil dari sebagian data yang terdapat pada basis data kelahiran bayi. Besarnya proporsi data yang dilakukan pengujian adalah 70% untuk *training*, sedangkan sisanya digunakan untuk uji coba model.
5. Uji Model Klasifikasi
   Proses uji model dilakukan setelah proses training data selesai dilakukan, jumlah data yang dilakukan uji model sebesar 30% dari basis data kelahiran bayi.
6. Evaluasi Model

Evaluasi model dilakukan dengan melihat tingkat akurasi metode melalui *confusion matrix* dan tabel akurasi serta presisi untuk model yang digunakan.

Hasil dan Pembahasan

Hasil klasifikasi akan dihadirkan dalam bentuk *Confusion Matrix*. *Predict Class* dan *Actual Class*. Pengujian model akan menggunakan Model *Confusion Matrix* 2x2 yang ditunjukkan pada Tabel 1.

Tabel 1 Model *confusion matrix* 2x2

		Predict Class	
		Class A	**Class B**
Actual Class	**Class A**	AA	AB
	Class B	BA	BB

Hasil pengujian ditunjukkan pada Gambar 2 *Confusion Matrix* Algoritma Klasifikasi C4.5 yang diperoleh merupakan evaluasi dari kinerja model klasifikasi, dan bukti terjadinya hasil proses pada klasifikasi tersebut yang telah tersedia didalamnya.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 911   4 |   a = PRIORITAS
   1  18 |   b = TIDAK PRIORITAS
```

Gambar 2 *Confuision matrix* metode C4.5

Selain akurasi dan *Confusion Matrix*, sebuah model klasifikasi bisa dilihat dari nilai *recall* dan presisinya. Presisi merupakan probabilitas bahwa sebuah item yang terpilih adalah relevan. Sedangkan *recall* adalah rasio dari item yang relevan yang dipilih terhadap total jumlah item yang relevan. Hasil *recall* dan presisi memiliki nilai antara 0-1. Semakin tinggi nilainya, maka semakin baik.

Berdasarkan informasi diatas, kemudian akan dilakukan proses perhitungan nilai rata-rata persentasi akurasi keberhasilan dengan mengunakan persamaan (3) dan *error rate* pada *confusion matrix* data *training* dengan menggunakan persamaan (4) berikut ini.

$$\text{Akurasi} = \frac{banyaknya\ prediksi\ yang\ benar}{total\ banyaknya\ prediksi} \quad (3)$$

$$\text{Akurasi} = \frac{911 + 18}{911 + 4 + 1 + 18} = \frac{929}{934} = 0{,}995$$

Maka Nilai Persentase Akurasi adalah

= 0,995 x 100%

= 99,5%

$$Error\ Rate = \frac{banyaknya\ prediksi\ yang\ salah}{total\ banyaknya\ prediksi} \quad (4)$$

$$Error\ Rate = \frac{4+1}{911+4+1+18} = \frac{5}{934} = 0{,}005$$

Maka Nilai Persentase *Error Rate* adalah

= 0,005 x 100%

= 0,5%

Nilai akurasi serta error rate data training dengan menggunakan algoritma C4.5 memiliki nilai akurasi 99,5% dan *Error Rate* sebesar 0,5%.

Berikutnya Gambar 3, merupakan *Confusion Matrix* Hasil pengujian dari metode *Naive Bayes Clasifier*.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 915   0 |   a = PRIORITAS
  19   0 |   b = TIDAK PRIORITAS
```

Gambar 3 *Confusion matrix* metode *naive bayes clasifier*

Berdasarkan informasi diatas, kemudian akan dilakukan proses perhitungan nilai rata-rata persentasi akurasi keberhasilan dan *error rate* pada *confusion matrix* data training untuk metode *Naïve Bayes Clasifier*.

$$Akurasi = \frac{banyaknya\ prediksi\ yang\ benar}{total\ banyaknya\ prediksi}$$

$$Akurasi = \frac{915+0}{915+0+19+0} = \frac{915}{934} = 0{,}98$$

Maka Nilai Persentase Akurasi adalah

= 0,98 x 100%

= 98%

$$Error\ Rate = \frac{banyaknya\ prediksi\ yang\ salah}{total\ banyaknya\ prediksi}$$

$$Error\ Rate = \frac{0+19}{915+0+19+0} = \frac{19}{934} = 0{,}020$$

Maka Nilai Persentase *Error Rate* adalah

= 0,020 x 100%

= 0,2%

Simpulan

Dari nilai akurasi serta *error rate* data *training* dengan menggunakan Algoritma C4.5 memiliki nilai akurasi lebih dari 90%, yakni dengan nilai 99,5% sedangkan nilai akurasi yang diperoleh oleh *Naïve Bayes Clasifier* adalah 97%. Hal ini menunjukkan bahwa Algoritma C4.5 lebih unggul 2% dari *Naïve Bayes Clasifier* sehingga Algoritma C4.5 lebih baik digunakan pada data set kelahiran bayi yang ada di Dinas Penduduk dan Catatan Sipil, dalam menentukan lokasi prioritas penyuluhan program keluarga berencana di Kecamatan Dumai Timur.

Daftar Pustaka

Larose, D. T. (2005). *Discovering Knowledge In Data.Mining An Introduction To Data Mining, Wiley Interscience.*

Han, Jiawei dan Kember, Micheline. (2006). *Data Mining*: *Concepts and Techniques Second Edition*. Morgan Kaufmann Publisher.

Mardi, Yuli. (2014). *Analisa Rekam Medis untuk Menentukan Penyakit Terbanyak berdasarkan International Classification Of Disease (ICD) menggunakan Decision Tree C4.5 (Studi Kasus: RSU. CBMC Dumai).* UPI YPTK Dumai.

Defiyanti, S. *Perbandingan Kinerja Algoritma ID3 dan C4.5 dalam klasifikasi spam-mail. Universitas Gunadarma*. Jakarta.

Iskandar, D. Suprapto, Yoyon K. (2013). *Perbandingan Akurasi Klasifikasi Tingkat Kemiskinan antara Algoritma C4.5 dan Naive Bayes Clasifier*. Java Journal of Electrical and Engineering, Vol. 11, No. 1.

Bouckaert, Remco R.; Frank, Eibe, dkk. (2008). WEKA Manual For Version 3-6-0. New Zealand: University of Waikato.

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

Yuhefizar, Budi Santoso, I Ketut Eddy P, Yoyon K Suprapto. (2013). *Combination of Cluster Method for Segmentation of Web Visitors*. Jurnal TELKOMNIKA Vol 11, No 1, Maret 2013.