Perbandingan akurasi klasifikasi tingkat kemiskinan antara algoritma C4.5 dan Naïve Bayes Clasifier

Derick Iskandar[1], Yoyon K. Suprapto[2]
Jurusan Teknik Elektro
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
derick14@mhs.ee.its.ac.id[1], yoyonsuprapto@ee.its.ac.id[2]

***Abstract*— Kemiskinan merupakan salah satu masalah yang dialami oleh beberapa Negara berkembang, termasuk indonesia. Banyak cara yang dilakukan untuk menanggulangi kemiskinan, diantaranya dengan program bantuan sosial untuk rakyat miskin. Bentuk bantuan sosial yang diberikan oleh pemerintah disesuaikan dengan tingkat kemiskinan yang ada disuatu wilayah sehingga pemberian bantuan sosial tersebut tidak salah sasaran. Pada penelitian kali ini kami menggunakan BDT (Basis Data Terpadu) yang dikeluarkan oleh TNP2K dalam menentukan klasifikasi tingkat kemiskinan. Adapun metode yang digunakan dalam penelitian ini adalah Naïve Bayes Clasifier (NBC) dan Algoritma C4.5 yang keduanya merupakan metode pada teknik klasifikasi data mining. Pegujian akan dilakukan dengan menggunakan 14 atribut. Hasil dari proses klasifikasi diperoleh bahwa metode C4.5 memiliki tingkat akurasi 3% lebih baik jika dibandingkan dengan metode Naïve Bayes.**

***Keywords*— C4.5, Naïve Bayes, Data mining, kemiskinan**

I. PENDAHULUAN

Indonesia merupakan salah satu Negara berkembang di asia khususnya asia tenggara. Salah satu masalah yang sering dihadapi oleh Negara berkembang adalah kemiskinan. Berdasarkan data yang dikeluarkan oleh Badan Pusat Statistik, angka kemiskinan di Indonesia pada tahun 1999 mencapai 47.97 juta jiwa. Pada tahun 2011 jumlah penduduk miskin menjadi 30.02 juta jiwa. Badan Pusat Statistik melakukan pendataan kependudukan khususnya masalah kemiskinan setiap 3 tahun sekali. Proses pendataan dilakukan dengan cara door to door langsung menuju rumah tangga sasaran.

Bagi pemerintah Indonesia masalah kemiskinan merupakan masalah lama yang belum dan sulit untuk diselesaikan [8]. Pemerintah sendiri telah melakukan beberapa upaya dalam melakukan pengentasan kemiskinan diantaranya melalui program bantuan sosial diantaranya Bantuan langsung tunai (BLT), Program Keluarga Harapan (PKH) dll. Salah satu kesulitan yang terkadang dihadapi oleh pemerintah adalah proses pembagian bantuan sosial yang tidak merata dan tepat sasaran. Kendala tersebut bisa terjadi karena faktor teknis maupun nonteknis. Basis Data Terpadu yang dikeluarkan oleh TNP2K merupakan salah satu dasar yang digunakan dalam penentuan klasifikasi tingkat kemiskinan tersebut. Di dalamnya terdapat setidaknya 14 atribut yang menentukan tingkat kemiskinan diantaranya tingkat pendidikan, jenis lapangan usaha, kedudukan dalam pekerjaan, status tempat tinggal, jenis atap, jenis dinding, jenis lantai, air minum yang digunakan, penerangan yang digunakan, bahan bakar memasak, fasilitas buang air besar, fasilitas tempat pembuangan akhir, kepesertaan KB, serta jumlah anggota rumah tangga dalam satu rumah.

Salah satu pemodelan yang bisa digunakan untuk klasifikasi tingkat kemiskinan tersebut adalah dengan menggunakan data mining. Dalam penelitian kali ini penulis mencoba membandingkan model klasifikasi yang dibentuk oleh teknik data mining antara algoritma decision tree C4.5 dan Naïve Bayes Clasifier (NBC). Metode ini di uji cobakan dengan menggunakan jumlah data sebanyak 15256 data set dengan 14 atribut pendukung. Hasil penelitian diperoleh bahwa metode decision tree memiliki tingkat akurasi lebih baik 3% dibandingkan dengan metode naïve bayes.

II. PENELITIAN SEBELUMNYA

Penelitian mengenai perbandingan/komparasi metode dalam data mining telah banyak dilakukan sebelumnya dengan jumlah data dan atribut yang berbeda-beda. Salah satu penelitian yang dilakukan oleh Hastuti [4] mengenai prediksi mahasiswa nonaktif dengan menggunakan metode klasifikasi data mining. Penelitian dilakukan dengan jumlah data sebanyak 3.861 data set dengan jumlah atribut sebanyak 21 item. Hasilnya di dapatkan bahwa nilai akurasi tertinggi diperoleh dengan menggunakan algoritma decision tree C4.5 sebesar 95,29%.

Penelitian berikutnya dilakukan oleh defiyanti [3]. Menggunakan metode C4.5 dan ID3 dalam mengklasifikasi spam mail dengan jumlah atribut dan jumlah data yang bervariasi. Dari penelitian ini didapatkan sebuah hasil bahwa nilai akurasi tertinggi yang diperoleh oleh algoritma C4.5 yakni sebesar 72,38% dengan jumlah atribut sebesar 52, sedangkan untuk algoritma ID3 memperoleh nilai akurasi tertinggi sebesar 73,20% pada jumlah atribut sebesar 58.

III. TINJAUAN PUSTAKA

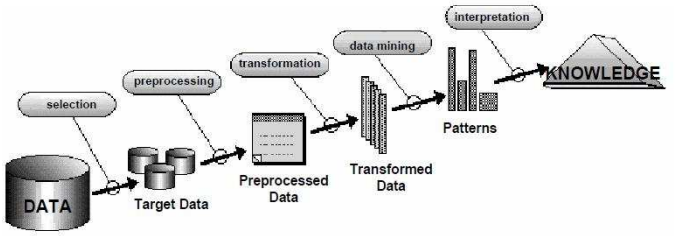
A. Basis Data Terpadu

Merupakan sistem data elektronik yang memuat informasi sosial, ekonomi, dan demografi dari sekitar 24,5 juta rumah tangga atau 96 juta individu dengan status kesejahteraan terendah di Indonesia. Sumber utama Basis Data Terpadu adalah hasil kegiatan Pendataan Program Perlindungan Sosial yang dilaksanakan oleh Badan Pusat Statistik (BPS) pada bulan Juli - Desember 2011 (PPLS 2011). Basis Data Terpadu digunakan untuk memperbaiki kualitas penetapan sasaran program-program perlindungan sosial. Basis Data Terpadu membantu perencanaan program, memperbaiki penggunaan

anggaran dan sumber daya program perlindungan sosial. Dengan menggunakan data dari Basis Data Terpadu, jumlah dan sasaran penerima manfaat program dapat dianalisis sejak awal perencanaan program. Hal ini akan membantu mengurangi kesalahan dalam penetapan sasaran program perlindungan sosial.

B. Data Mining

Data mining merupakan sebuah teknik untuk menggali informasi tersembunyi untuk memperoleh manfaat lebih dari data yang ada [8]. Menurut witten data mining bisa diartikan serangkaian proses mendapatkan pengetahuan atau pola dari kumpulan data. Diantara tujuan data mining adalah untuk melakukan klasifikasi, klasterisasi, menemukan pola asosiasi hingga melakukan peramalan (predicting). Istilah data mining lebih populer dengan sebutan KDD (Knowledge Discovery from Database). Proses KDD dapat dilihat pada gambar 1 mulai dari pemilihan atribut data hingga terciptalah sebuah pengetahuan.



Gambar 1. KDD process

C. Klasifikasi

Merupakan suatu teknik dengan melihat pada kelakuan dan atribut dari kelompok yang telah didefinisikan.Teknik ini dapat memberikan klasifikasi pada data baru dengan memanipulasi datayang ada yang telah diklasifikasi dan dengan menggunakan hasilnya untuk memberikan sejumlahaturan. Aturan-aturan tersebut digunakan pada data-data baru untuk diklasifikasi. Teknik inimenggunkan supervised induction yang memanfaatkan kumpulan pengujian dari data set yang terklasifikasi.

D. Algoritma C4.5

Algoritma C4.5 merupakan salah satu algoritma decision tree yang paling efektif untuk melakukan klasifikasi [2]. Algoritma C4.5 merupakan pengembangan dari algoritma ID3 (*Iterative Dichotomiser*) yang ditemukan oleh J Ross Quinlan. Pohon keputusan ini dibangun dengan cara membagi data secara rekursif hingga tiap bagian terdiri dari data yang berasal dari kelas yang sama. Bentuk pemecahan (*split)* yang digunakan untuk membagi data tergantung dari jenis atribut yang digunakan dalam *split*. Secara umum algoritma decision tree memiliki tahapan sebagai berikut :

1. Pilih atribut sebagai akar

   Memilih atribut sebagai akar bisa dihitung dengan melihat nilai gain dari masing-masing atribut. Nilai gain tertinggi nantinya akan menjadi akar yang pertama. Bentuk persamaan untuk mendapatkan nilai gain bisa dilihat pada Persamaan *(1).*

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

Dimana :
S : Himpunan Kasus
A : Atribut
n : jumlah partisi atribut A
|Si| : jumlah kasus pada partisi ke –i
|S| : jumlah kasus dalam S
Sedangkan nilai entropy sendiri bisa diperoleh dari persamaan *(2).*

$$Entropy\ (S) = \sum_{i=0}^{n} -pi * log_2\ pi \qquad \textbf{(2)}$$

Dimana:
S : Himpunan kasus
n : jumlah partisi S
Pi : Proporsi Si terhadap S

2. Ulangi metode di atas hingga semua data terbagi.
3. Proses pengulangan pada metode decision tree ini akan berhenti jika :
   - Semua data telah terbagi rata
   - Tidak ada lagi atribut yang bisa di bagi lagi
   - Tidak ada data record dalam cabang yang kosong

Pada algoritma C4.5 terdapat pembeda dari ID3 yakni adanya RasioGain yang berfungi sebagai pemecah atribut. Persamaan *(3)* memperlihatkan bagaimana RasioGain diperoleh.

$$RasioGain\ (s,j) = \frac{Gain\ (s,j)}{SplitInfo\ (s,j)} \qquad \textbf{(3)}$$

Dimana :
S : Himpunan kasus
j : fitur ke- j

Sedangkan untuk split info dapat diperoleh dengan rumus pada Persamaan *(4).*

$$SplitInfo\ (s,j) = -\ \sum_{i=1}^{k} p(Vi \mid s) log_2\ p(Vi \mid s) \qquad \textbf{(4)}$$

Dimana :
k : jumlah pemecahan

E. Naïve Bayes Clasifier (NBC)

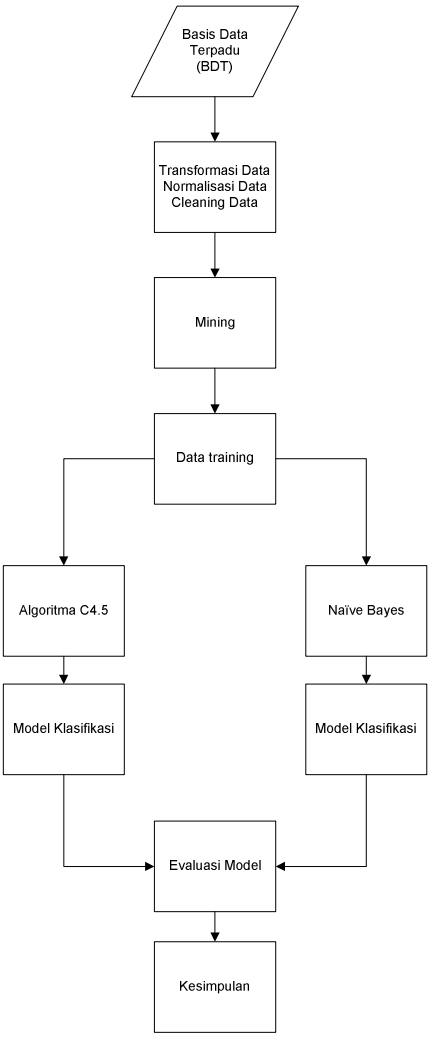
Merupakan pengklasifikasian dengan metode probabilitas dan statistik yang dikemukakan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi peluang di masa depan berdasarkan pengalaman di masa sebelumnya sehingga dikenal sebagai teorema Bayes. Teorema tersebut di asumsikan memiliki atribut saling bebas (independen).
Dasar teori yang digunakan dalam melakukan klasifikasi ini adalah teorema bayes yang ditunjukkan oleh persamaan *(5).*

$$P\ (A|B) = (p\ (B|A) * p(A)) / p(B) \qquad \textbf{(5)}$$

Peluang A sebagai B, diperoleh dari peluang B saat A, peluang A dan peluang B.

IV. METODE PENELITIAN

Metode yang digunakan dalam penelitian ini adalah bertujuan untuk memperlihatkan bagaimana sebuah model klasifikasi data mining bisa memberikan solusi untuk mengklasifikasikan tingkat kemiskinan berdasarkan atribut yang ada. Tahapan penelitian bisa dilihat pada gambar 2.



Gambar 2. Alur penelitian

1. Transformasi Data

Basis Data Terpadu 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 yakni mengubah jenis skala pengukuran yang semula numeric menjadi nominal.

3. Cleaning Data

Proses membersihkan data yang tidak relevan termasuk data missing dalam atribut [10]. Jumlah atribut setelah dilakukan cleaning ditunjukkan oleh tabel 1.

Tabel 1. Jumlah data setelah proses Cleaning

Jumlah Data awal	Jumlah data setelah *cleaning*
15.256 data set	13.928 data set

4. Training Data

Proses pelatihan data diambil dari sebagian data yang terdapat pada BDT. Besarnya proporsi data yang dilakukan pengujian adalah 60% untuk training, sedangkan sisanya digunakan untuk uji coba model.

5. Uji Model

Proses uji model dilakukan setelah proses training data selesai dilakukan. Jumlah data yang dilakukan uji model sebesar 40 % dari BDT.

6. Evaluasi Model

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

V. HASIL DAN PEMBAHASAN

Sebelum data dilakukan training, maka dipecah menjadi 2 bagian:

1. Data latih
2. Data uji

Keduanya dibagi menurut proporsi jenis klasifikasi yang telah terbentuk, masing-masing 60% data latih dan 40% data uji. Pembagian proporsi data sesuai dengan tabel 2.

Tabel 2. Pembagian proporsi data training dan testing

Jenis klasifikasi	Jumlah data training	Jumlah data testing	Total
Hampir miskin	2269 data set	1513 data set	3782 data set
Miskin	2293 data set	1528 data set	3821 data set
Sangat miskin	3795 data set	2530 data set	6325 data set

A. Pengujian Model

Hasil klasifikasi akan di hadirkan dalam betuk *confusion matrix*. Tabel ini terdiri dari *predict class* dan *actual class*. Model *confusion matrix* 3x3 ditunjukkan pada tabel 3.

Tabel 3. Model *confusion matrix*

		Predict Class		
		Class A	Class B	Class C
Actual Class	Class A	AA	AB	AC
	Class B	BA	BB	BC
	Class C	CA	CB	CC

Nilai akurasi model diperoleh dari persamaan *(6)*, jumlah data yang tepat diklasifikasikan dibagi dengan total data.

$$Akurasi = \frac{AA+BB+CC}{AA+AB+AC+BA+BB+BC+CA+CB+CC} \quad \textbf{(6)}$$

Dengan bantuan tools WEKA, maka di dapatkan tabel *confusion matrix* untuk metode C4.5 seperti yang ditunjukkan oleh gambar 3. Sedangkan untuk metode naïve bayes bisa dilihat pada gambar 4.

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 2121  110  318 |    a = sangat miskin
  175  977  370 |    b = hampir miskin
  519  496  485 |    c = miskin
```

Gambar 3. *Confusion matrix* metode C4.5

```
=== Confusion Matrix ===

    a    b    c   <-- classified as
 2159  181  209 |    a = sangat miskin
  307  910  305 |    b = hampir miskin
  671  481  348 |    c = miskin
```

Gambar 4. *Confusion matrix* metode naïve bayes

Tabel 4 menunjukkan perbandingan hasil akurasi 2 model diatas. Nilai akurasi pada metode C4.5 3% lebih baik jika dibandingkan dengan naïve bayes.

Tabel 4. Hasil akurasi

Metode	Akurasi
C4.5	64%
Naïve Bayes	61%

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. Nilai presisi ditunjukkan pada persamaan *(7)*.

$$Presisi\ \boldsymbol{i} = \frac{Ai}{Ai+Bi+Ci} \quad \textbf{(7)}$$

Sedangkan *recall* adalah rasio dari item yang relevan yang dipilih terhadap total jumlah item yang relevan. Nilai recall dapat diperoleh dari persamaan *(8)*.

$$Recall\ \boldsymbol{i} = \frac{iA}{iA+iB+iC} \quad \textbf{(8)}$$

Hasil presisi dan *recall* yang diperoleh dari model klasifikasi diatas ditunjukkan oleh tabel 5. Hasil recall dan presisi memiliki nilai antara 0-1. Semakin tinggi nilainya, maka semakin baik.

Tabel 5. Nilai presisi dan *recall*

Jenis klasifikasi	C4.5		Naïve Bayes	
	Presisi	Recall	Presisi	Recall
Sangat miskin	0.753	0.832	0.688	0.847
Miskin	0.413	0.323	0.404	0.232
Hampir miskin	0.617	0.642	0.579	0.598

Jika dilihat dari tabel diatas maka secara umum metode C4.5 memiliki nilai presisi dan recall yang lebih baik daripada metode naïve bayes.

VI. KESIMPULAN DAN SARAN

Berdasarkan hasil komparasi antara algoritma C4.5 dan Naïve Bayes untuk mengklasifikasikan tingkat kemiskinan dengan 14 atribut dan jumlah data yang telah di *cleaning* sebesar 13.928 data set dapat disimpulkan bahwa algoritma C4.5 memilliki tingkat akurasi yang lebih baik 3% dibandingkan dengan metode naïve bayes yang bernilai 63%. Meskipun demikian dilihat dari nilai presisi dan recall untuk masing-masing metode hanya memiliki selisih yang tidak jauh berbeda. Hal ini menunjukkan bahwa untuk jumlah fitur/atribut yang sama akan menghasilkan nilai akurasi yang tidak jauh berbeda.

Saran untuk penelitian berikutnya adalah bisa dilakukan optimasi pada tahap pemilihan atribut sehingga kompleksitas atribut dapat berkurang. Dengan demikian diharapkan nilai akurasi dan presisi akan meningkat.

REFERENCES

[1] Aradea,Satriyo A., Ariyan, Z., Yuliana,A. 2011. Penerapan Decision Tree untuk penentuan pola data Penerimaan Mahasiswa Baru. Jurnal Penelitian Sitrotika Vol 7 No 1. Universitas Diponegoro, Semarang.

[2] Chauhan, H and Chauhan, A. 2013. Implementation of decision tree algorithmC4.5. International Journal of Scientific and Research Publication Vol 3 issue 10, October 2013.

[3] Defiyanti, S. Perbandingan kinerja Algoritma ID3 dan C4.5 dalam klasifikasi spam-mail. Universitas Gunadarma. Jakarta.

[4] Hastuti, K .2012. Analisis komparasi Algoritma Klasifikasi Data Mining untuk prediksi mahasiswa non aktif. Seminar Nasional Teknologi Informasi dan Komunikasi Terapan 2012. Universitas Dian Nuswantoro, Semarang.

[5] Ian H. Witten, frank Eibe, and Mark A. Hall, Data Mining: Practical Machine Learning Tools and Techniques, 3rd ed., Asma Stephan and Burlington, Eds. United States of America: Morgan Kaufmann, 2011.

[6] Prasetyo, Eko.2014.Data mining mengolah data menjadi informasi menggunakan matlab. Yogyakarta : penerbit andi.

[7] http://bdt.tnp2k.go.id/ diakses pada 2 oktober 2015.

[8] Sutaat. 2006. Hasil-hasil Penelitian Tahun 2006 Puslitbang Kesejahteraan Sosial. Pusat Penelitian dan Pengembangan Kesejahteraan Sosial. Badan Pendidikan dan Kesejahteraan Sosial, Departemen Sosial Republik Indonesia : Jakarta.

[9] Tan, dkk, 2006. Introduction to Data Mining. Pearson Education, Inc.

[10] Yuhefizar, Budi Santosa, 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.