Perbandingan Kinerja Algoritma Naive Bayes dan C4.5 Untuk Klasifikasi Harga Pangan

Nurani
Department of Computer Engineering,
Academy of Information and
Computer Management Rizky Makassar
nurani.nannii@gmail.com

Afif
Department of Computer Engineering,
Academy of Information and
Computer Management Rizky Makassar
afif.sudrahsyah@gmail.com

Prediksi harga pangan merupakan proses analisis yang diperlukan oleh sebuah sistem pada distribusi penjualan, metode yang bisa digunakan untuk prediksi salah satu adalah teknik data mining. Data mining didefinisikan sebagai suatu proses untuk mencari pola dari sekumpulan data yang terdapat di dalam database untuk kemudian dianalisis sehingga menghasilkan suatu informasi. Algoritma data mining yang digunakan adalah Naive Bayes dan C.45 dengan pengujian Precision, Recall serta Accuracy untuk setiap data training dan data testing yang telah diberikan. Berdasarkan hasil pengujian, semakin banyaknya data training yang digunakan, maka nilai precision, recall dan accuracy akan semakin meningkat. Selain itu, hasil klasifikasi pada algoritma Naive Bayes dan C.45 tidak dapat memberikan nilai yang absolut atau mutlak. Dengan menggunakan alat bantu WEKA (Waikato Environment for Knowledge Analaysis) Hasil perbandingan menunjukkan bahwa metode C45 memiliki tingkat akurasi tertinggi 65% dibandingkan algoritma naive bayes dan nearest neighbour yaitu mencapai 60%.

Keywords - Data Mining; Naïve Bayes; C.45; Precision; Recall, accuracy.

I. PENDAHULUAN

Data mining adalah proses yang memanfaatkan suatu metode untuk memperoleh pola dari suatu data, sedangkan menurut Mirkin Data mining didefinisikan sebagai suatu proses untuk mencari pola dari sekumpulan data yang terdapat di dalam database untuk kemudian dianalisis sehingga menghasilkan suatu informasi tertentu untuk dimanfaatkan pada proses selanjutnya. Salah satu pendekatan yang dapat digunakan untuk menganalisis sekumpulan data adalah klasifikasi.

Klasifikasi merupakan salah satu teknik data mining yang digunakan untuk membangun suatu model dari sampel data yang belum terklasifikasi untuk digunakan mengklasifikasi sampel data baru ke dalam kelas-kelas yang sejenis. Klasifikasi termasuk ke dalam supervised learning karena menggunakan sekumpulan data untuk dianalisis terlebih dahulu, kemudian pola dari hasil analisis tersebut digunakan untuk pengklasifikasian data uji. Proses klasifikasi data terdiri dari pembelajaran dan klasifikasi. Pada pembelajaran data training dianalisis menggunakan algoritma klasifikasi, selanjutnya pada klasifikasi digunakan data testing untuk memastikan tingkat akurasi dari rule klasifikasi yang digunakan. Teknik klasifikasi dibagi menjadi lima kategori berdasarkan perbedaan konsep matematika, yaitu berbasis statistik, berbasis jarak, berbasis pohon keputusan, berbasis jaringan syaraf, dan berbasis rule. Ada banyak algoritma dari masing-masing kategori tersebut, namun yang populer dan sering digunakan diantaranya yaitu naive bayes & C 45

II. TEORI DASAR

Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai *database* besar. Istilah *data mining* memiliki hakikat sebagai disiplin ilmu yang tujuan utamanya adalah untuk menemukan, menggali, atau menambang pengetahuan dari data atau informasi yang kita miliki. *Data mining*, sering juga disebut sebagai *Knowledge Discovery in Database* (KDD).

KDD adalah kegiatan yang meliputi pengumpulan, pemakaian data, historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar .

1. Metode Pelatihan

Secara garis besar metode pelatihan yang digunakan dalam teknik-teknik *data mining* dibedakan ke dalam dua pendekatan, yaitu :

- *Unsupervised learning*, metode ini dierapkan tanpa adanya latihan *(training)* dan tanpa ada guru *(teacher)*. Guru di sini adalah label dari data.
- *Supervised learning*, yaitu metode belajar dengan adanya latihan dan pelatih. Dalam pendekatan ini, untuk menemukan fungsi keputusan, fungsi pemisah atau fungsi regresi, digunakan beberapa contoh data yang mempunyai output atau label selama proses *training*.

2. Pengelompokan *Data Mining*

Ada beberapa teknik yang dimiliki *data mining* berdasarkan tugas yang bisa dilakukan, yaitu :

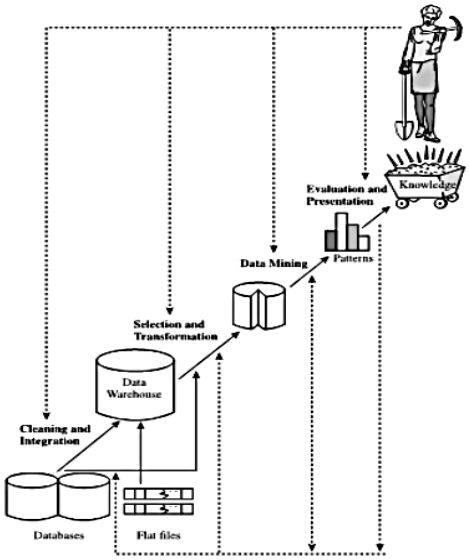
- Deskripsi
  Para peneliti biasanya mencoba menemukan cara untuk mendeskripsikan pola dan trend yang tersembunyi dalam data.
- Estimasi

Estimasi mirip dengan klasifikasi, kecuali variabel tujuan yang lebih kearah numerik dari pada kategori.

- Prediksi
  Prediksi memiliki kemiripan dengan estimasi dan klasifikasi. Hanya saja, prediksi hasilnya menunjukkan sesuatu yang belum terjadi (mungkin terjadi di masa depan).
- Klasifikasi
  Dalam klasifikasi variabel, tujuan bersifat kategorik. Misalnya, kita akan mengklasifikasikan pendapatan dalam tiga kelas, yaitu pendapatan tinggi, pendapatan sedang, dan pendapatan rendah
- Clustering
  Clustering lebih ke arah pengelompokan record, pengamatan, atau kasus dalam kelas yang memiliki kemiripan.
- Asosiasi
  Mengidentifikasi hubungan antara berbagai peristiwa yang terjadi pada satu waktu.

3. Tahap-tahap *Data Mining*

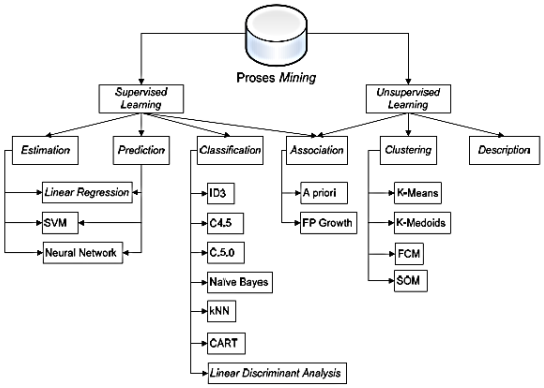
Sebagai suatu rangkaian proses, *data mining* dapat dibagi menjadi beberapa tahap prose yang diilustrasikan pada Gambar 1. Tahap-tahap tersebut bersifat interaktif, pemakai terlibat langsung atau dengan perantaraan *knowledge base*.



Gambar 1 Tahap-tahap Data Mining

Tahap-tahap *data mining* adalah sebagai berikut:

- Pembersihan data *(data cleaning)*
  Pembersihan data merupakan proses menghilang-kan *noise* dan data yang tidak konsisten atau data tidak relevan.
- Integrasi data *(data integration)*
  Integrasi data merupakan penggabungan data dari berbagai *database* ke dalam satu *database* baru.
- Seleksi data *(data selection)*
  Data yang ada pada database sering kali tidak semuanya dipakai, oleh karena itu hanya data yang sesuai untuk dianalisis yang akan diambil dari *database*.
- Transformasi data *(data transformation)*
  Data diubah atau digabung ke dalam format yang sesuai untuk diproses dalam *data mining*.
- Proses *mining*
  Merupakan suatu proses utama saat metode diterapkan untuk menemukan pengetahuan berharga dan tersembunyi dari data. Beberapa metode yang dapat digunakan berdasarkan pengelompokan *data mining* dapat dilihat pada Gambar 2



Gambar 2 data mining

- Evaluasi pola *(pattern evaluation)*
  Untuk mengidentifikasi pola-pola menarik ke dalam *knowledge based* yang ditemukan.
- Presentasi pengetahuan *(knowledge presentation)*
  Merupakan visualisasi dan penyajian pengetahuan mengenai metode yang digunakan untuk memperoleh pengetahuan yang diperoleh pengguna.

Menurut Prasetyo, klasifikasi merupakan suatu Pekerjaan menilai objek data untuk memasukkannya ke dalam kelas tertentu dari sejumlah kelas yang tersedia. Dalam klasifikasi terdapat dua proses yang dilakukan yaitu dengan membangun model untuk disimpan sebagai memori dan menggunakan model tersebut untuk melakukan pengenalan atau klasifikasi atau prediksi pada suatu data lain supaya diketahui di kelas mana objek data tersebut dimasukkan berdasarkan model yang telah disimpan dalam memori.

Sistem dalam klasifikasi diharapkan mampu melakukan klasifikasi semua set data dengan benar, namun tidak dapat dipungkiri bahwa kesalahan akan terjadi dalam proses pengklasifikasian tersebut sehingga perlunya dilakukan pengukuran kinerja dari sistem klasifikasi tersebut. Umumnya, pengukuran kinerja klasifikasi dilakukan dengan matriks konfusi *(confusion matrix)*. Matriks konfusi merupakan tabel pencatat hasil kerja klasifikasi. Contoh dari matriks

konfusi untuk dua kelas (biner) dapat dilihat pada Tabel 1.

Tabel 1. Matriks Konfusi untuk Klasifikasi Dua Kelas

f_{ij}		Kelas hasil prediksi (j)	
		Kelas = 1	Kelas = 2
Kelas asli (i)	Kelas = 1	f_{11}	f_{12}
	Kelas = 2	f_{21}	f_{22}

Jadhav et all (2016) menyatakan bahwa Naïve Bayes Classifier adalah suatu model independen yang membahas mengenai klasifikasi sederhana berdasarkan teorema Bayes. Naïve Bayes merupakan suatu algoritma yang dapat mengklasifikasikan suatu variable tertentu dengan menggunakan metode probabilitas dan statistik. Secara garis besar algoritma Naïve Bayes dapat dijelaskan seperti persamaan (1)

$$P(R|S) = \frac{P(R)P(S|R)}{\mathrm{P(S)}}$$

Keterangan:

R : Data yang belum diketahui kelasnya
S : Hipotesis pada data R yang merupakan class khusus
P(R|S) : Nilai probabilitas pada hipotesis R yang berdasarkan kondisi S
P(R) : Nilai probabilitas pada hipotesis R
P(S|R) : Nilai probabilitas S yang berdasarkan dengan kondisi hipotesis R
P(S) : Nilai probabilitas S

Dengan menggunakan persamaan diatas, data yang telah diperoleh dapat diproses dengan algoritma Naive Bayes untuk penilaian data yang akan diklasifikasikan.

Purushottam, et al (2016) menyatakan algoritma C4.5 merupakan algoritma yang dipergunakan dalam membentuk *decision tree* (pengambilan keputusan). Algoritma C4.5 adalah
salah satu algoritma dalam induksi *decision tree* yaitu ID3 (*Iterative Dichotomiser* 3) yang dikembangan oleh J. Ross Quinlan. Dalam prosedur algoritma ID3, input berupa sampel *training*, label *training* dan atribut. Algoritma C4.5 ini merupakan pengembangan dari ID3. Ide dasar dari algoritma ini adalah pembuatan pohon keputusan berdasarkan pemilihan atribut yang memiliki prioritas tertinggi atau dapat disebut memiliki nilai *gain* tertinggi berdasarkan nilai *entropy* atribut tersebut sebagai poros atribut klasifikasi. Kemudian secara rekursif cabang-cabang pohon diperluas sehingga seluruh pohon terbentuk. Terdapat empat langkah dalam proses pembuatan pohon keputusan pada algoritma C4.5, yaitu:

a. Memilih atribut sebagai akar.
b. Membuat cabang untuk masing-masing nilai.
c. Membagi setiap kasus dalam cabang.
d. Mengulangi proses dalam setiap cabang sehingga semua kasus dalam cabang memiliki kelas yang sama.

Kemudian dilakukan perhitungan untuk mencari nilai *entropy* dan *gain*. Berikut ini rumus untuk mencari nilai *entropy* dan *gain*.

$$Entropy(S) = \sum_{j=1}^{k} - pj \log_2 pj$$

persamaan yang digunakan dalam perhitungan *entropy* untuk menentukan *heterogenity* dari sebuah kumpulan *data sample* (Amin et al, 2015). Berikut keterangannya :

: Himpunan kasus
: Jumlah partisi S
: Jumlah kasus pada partisi ke-j

$$Gain(A) = Entropy(S) - \sum_{i=1}^{k} \frac{|S_i|}{|S|} \times Entropy(S_i)$$

rumus yang digunakan dalam perhitungan *gain* setelah melakukanperhitungan *entropy*. Berikut keterangannya :

: Atribut dari dataset
: Jumlah partisi S
: Himpunan kasus

Dengan mengetahui rumus–rumus diatas, data yang telah diperoleh dapat dimasukkan dan diproses dengan algoritma C4.5 untuk proses pembuatan *decision tree*

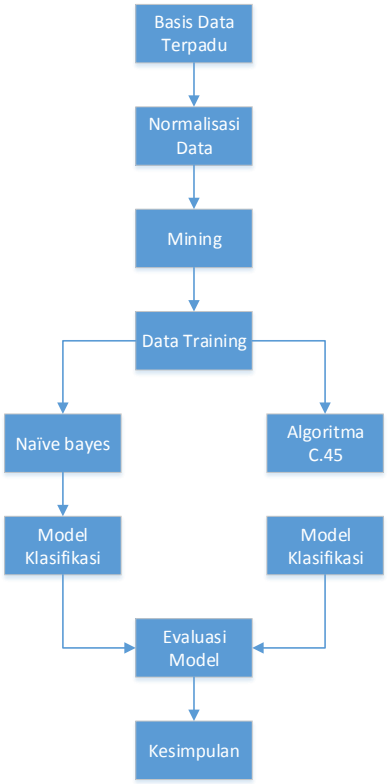
WEKA adalah sebuah alat yang digunakan untuk membandingkan beberapa algoritma machine learning yang bisa diaplikasikan untuk permasalahan data mining. WEKA dikembangkan oleh University of Wakaito, New Zealand yang bersifat open source.

Penelitian yang dilakukan oleh Seongwook Youn dan Dennis Mcleod (2006) menggunakan WEKA sebagai alat bantu untuk membandingkan kinerja tiga algoritma yaitu Neural Network, Support Vektor Mechine (SVM),Naïve Bayesian dan C4.5. Keempat algoritma tersebut digunakan dalam kasus yang sama yaitu mengklasifikasikan email menjadi spam atau non-spam.

Beberapa kelebihan yang dimiliki WEKA antara lain mudah digunakan, berbasis GUI (Graphical Interface User) dan bisa digunakan untuk mengintegrasikan metode baru yang dibuat sendiri dengan beberapa ketentuan.

III. METODOLOGI 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 .



Gambar 3 Alur Perbandingan

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. Jumlah atribut setelah dilakukan cleaning.
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.

IV. HASIL DAN PEMBAHASAN

Sebelum data dilakukan training, maka dipecah menjadi 2
bagian:

1. Data training
2. Data testing

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.

jenis klasifikasi	jumlah data training	jumlah data testing	total
Naik	2125	1733	3858
Normal	2323	1543	3866
Turun	2437	1439	3876

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.

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

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

$$Akurasi = \frac{AA + BB + CC}{AA + AB + AB + BA + BB + BC + CA + CB + CC}$$

Dengan bantuan tools WEKA, maka di dapatkan tabel confusion matrix untuk metode C4.5 dan metode naïve bayes bisa pada tabel berikut menunjukkan perbandingan hasil akurasi modeldiatas. Nilai akurasi pada metode C4.5 5% lebih baik jika dibandingkan dengan naïve bayes

Tabel 3 Hasil Akurasi

No	Metode	Akurasi
1	Naïve Bayes	60 %
2	C4.5	65 %

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 dan dibutuhkan. Nilai presisi ditunjukkan pada persamaan

$$Presisi\ /\ Recall/Accuracy = \frac{A}{A+B+C}$$

Sedangkan recall adalah rasio dari item yang relevan yang dipilih terhadap total jumlah item yang relevan. dan accuracy adalah persentase dari total data harga yang benar diidentifikasi. Hasil presisi, *recall dan accuracy* yang diperoleh dari model klasifikasi diatas ditunjukkan oleh tabel 5. Hasilnya memiliki nilai antara 0-1. Semakin tinggi nilainya, maka semakin baik.

Tabel.4 Nilai presisi, recall dan accuracy

jenis klasifikasi	Naïve Bayes			C 4.5		
	Presisi	Recall	Accuracy	Presisi	Recall	Accuracy
Naik	0.575	0.417	0.332	0.791	0.841	0.523
Normal	0.331	0.157	0.312	0.423	0.312	0.498
Turun	0.468	0.618	0.398	0.621	0.671	0.578

Dari tabel diatas dari perbandingan metode, dengan parameter presisi, recall dan accuracy dapat dilihat algoritma C 4.5 lebih ungggul dari pada naïve bayes.

V. KESIMPULAN

Berdasarkan hasil komparasi antara algoritma Naïve Bayes dan algoritma C4.5 untuk mengklasifikasikan harga pangan dengan 50 atribut. Algoritma C4.5 memilliki tingkat akurasi yang lebih baik 5% dibandingkan dengan metode naïve bayes yang bernilai 65%. Meskipun demikian dilihat dari nilai presisi, recall dan accuracy 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.

DAFTAR PUSTAKA

[1] J. Iawe. Han, M. Kamber, and J. Pei, 2012, Data Mining Concept and Techniques

[2] B. Mirkin, 2011, "Data Analysis, Mathematical Statistics, Machine Learning, Data Mining: Similarities and Differences," 2011 Int. Conf. Adv. Comput. Sci. Inf. Syst., Vol. 2, pp. 1–8.

[3] M. Ramageri, 2010, "Data Mining Techniques and Applications," Indian J. Comput. Sci. Eng., Vol. 1, No. 4, pp. 301–305.

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

[5] 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.

[6] M. Ramageri, 2010, "Data Mining Techniques and Applications," Indian J. Comput. Sci. Eng., Vol. 1, No. 4, pp. 301–305.

[7] Larose, Daniel T. 2005. Discovering Knowledge In Data: An Introduction to Data Mining. New Jersey: John Wiley and Sons Inc.

[8] Berry, M.W. dan M. Browne. 2006. Lecture Notes in Data Mining. Singapore: World Scientific Publishing Co. Pte. Ltd.

[9] Han, J and Kamber, M. 2006. Data Mining Concepts and Techniques, second edition. California: Morgan Kaufman.

[10] I. A. Mubarok, N. A. Wesiani, and A. Rusdiansyah, "Pengembangan Prototype Knowledge Management System berbasis Case Based Reasoning bagi Peningkatan Aksesibilitas UMKM Dalam Permodalan Usaha," pp. 1–6.