

VALIDITAS SUATU ALAMAT MENGGUNAKAN DECISION TREE DENGAN ALGORITMA ID3

Fanuel Nugroho S⁽¹⁾, Harianto Kristanto⁽²⁾, Yetli Oslan⁽³⁾

Intisari:

Kita mengetahui bahwa bidang ilmu dan teknologi informasi sangat dibutuhkan karena tidak dapat diragukan lagi mempunyai suatu pengaruh yang besar dalam kehidupan saat ini. Untuk suatu perusahaan yang secara khusus bergerak di bidang jasa, permasalahan yang dihadapi adalah ketika pihak perusahaan mengirimkan barang dengan tujuan alamat yang diperoleh dari pihak pelanggan, seringkali pihak pelanggan tidak memberikan alamat pengiriman secara tepat sehingga pihak perusahaan harus memprediksi alamat tersebut secara manual. Dalam melakukan hal ini, membutuhkan waktu yang tidak sedikit sehingga mengurangi kualitas pelayanan terhadap pelanggan. Dari permasalahan ini, diperlukan suatu solusi yang dapat membantu perusahaan dalam melakukan prediksi suatu alamat yang benar.

Salah satu solusi memprediksi validitas alamat adalah dengan membuat suatu sistem yang dianggap mampu melakukan prediksi suatu alamat secara tepat. Metode decision tree dengan algoritma ID3 merupakan salah satu metode dari data mining yang digunakan untuk mengklasifikasikan data sampel ke dalam kelas-kelas tertentu. Berdasarkan kemampuan dari metode ini, kemudian dilakukan penelitian dalam bentuk skripsi untuk menganalisis keefektifitasan metode ini dalam melakukan prediksi alamat menggunakan kelas yang terbentuk dari metode ini. Bentuk penelitian yang dilakukan adalah dengan melihat tingkat kebenaran yang dihasilkan oleh metode ini dalam melakukan validitas prediksi suatu alamat pada sekumpulan data uji yang diteliti.

Hasil akhir dari skripsi adalah penelitian terhadap hasil prediksi suatu alamat pada sejumlah data pengujian dengan persentase tingkat kebenaran, kesalahan, dan data yang tidak bisa diprediksi untuk mengetahui keefektifitasan metode decision tree dengan algoritma ID3 dalam melakukan prediksi suatu alamat.

PENDAHULUAN

Suatu perusahaan yang memiliki jasa pengiriman barang yang bergerak di dalam kota mempunyai beberapa masalah. Salah satu analisis yang bisa dilakukan adalah pengecekan kebenaran alamat karena sering kali pelanggan tidak memberikan alamat pengiriman secara tepat sehingga pihak perusahaan harus memprediksi alamat tersebut secara manual.

Penelitian ini dilakukan dengan menerapkan metode *Decision Tree* menggunakan

algoritma ID3, serta untuk mengetahui kemampuan metode ini dalam menangani pengelompokan data-data alamat, dan juga untuk mengetahui kemampuan metode ini dalam memberikan hasil keputusan alamat yang tepat.

DECISION TREE

Salah satu metode *data mining* yang umum digunakan adalah *decision tree* yang merupakan sebuah *flow chart* yang seperti struktur pohon, dimana setiap titik pohon merupakan atribut

⁽¹⁾ Fanuel Nugroho S, Alumni Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana

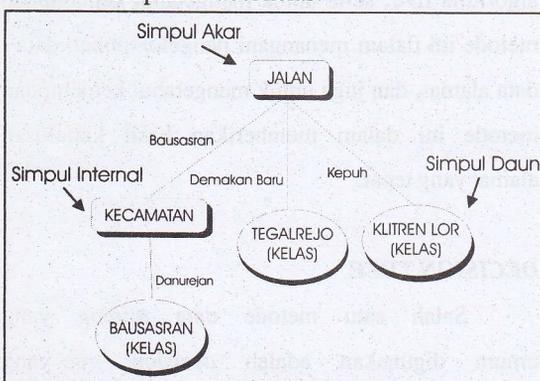
⁽²⁾ Harianto Kristanto, Dosen Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana

⁽³⁾ Yetli Oslan, Dosen Teknik Informatika, Fakultas Teknik, Universitas Kristen Duta Wacana

yang telah diuji. Bagian awal dari pohon keputusan ini adalah titik akar (*root*), sedangkan setiap cabang dari *decision tree* merupakan pembagian berdasarkan hasil uji, dan titik akhir (*leaf*) merupakan pembagian kelas yang dihasilkan. Pada umumnya proses dari sistem *decision tree* adalah mengadopsi strategi pencarian *top-down* untuk solusi ruang pencariannya. Pada proses mengklasifikasikan sampel yang tidak diketahui, nilai atribut akan diuji pada *decision tree* dengan cara melacak jalur dari titik akar sampai titik akhir, kemudian akan diprediksikan kelas yang ditempati sampel baru tersebut.

Decision tree mempunyai 3 tipe simpul yaitu:

- ⊙ simpul akar dimana tidak memiliki cabang yang masuk dan memiliki cabang lebih dari satu, terkadang tidak memiliki cabang sama sekali.
- ⊙ simpul internal dimana hanya memiliki 1 cabang yang masuk, dan memiliki lebih dari 1 cabang yang keluar.
- ⊙ simpul daun atau simpul akhir dimana hanya memiliki 1 cabang yang masuk, dan tidak memiliki cabang sama sekali dan menandai bahwa simpul tersebut merupakan label kelas.



Gambar 1. Decision Tree

Tahap awal dilakukan pengujian simpul akar, jika pada pengujian simpul akar menghasilkan sesuatu maka proses pengujian juga dilakukan pada setiap cabang berdasarkan hasil dari pengujian. Hal ini berlaku juga untuk simpul internal dimana suatu

kondisi pengujian baru akan diterapkan pada simpul daun.

METODE *DECISION TREE CLASSIFICATION* DENGAN ALGORITMA ID3

Algoritma ID3 juga dikenal dengan nama C4.5 diperkenalkan dan dikembangkan pertama kali oleh Quinlan. Algoritma ID3 membentuk pohon keputusan dengan cara pembagian dan menguasai sampel secara rekursif dari atas ke bawah. Algoritma ID3 dimulai dengan semua data yang ada sebagai akar dari pohon keputusan. Sebuah atribut yang dipilih akan menjadi pembagi dari sampel tersebut. Untuk setiap atribut dari cabang yang telah dibentuk, semua sampel yang memiliki nilai yang sama dengan atribut cabang akan masuk dalam anggotanya dan dinamakan anak cabang. Berikut ini merupakan algoritma dari *Decision Tree* :

Masukan : Data Sampel, daftar dari atribut tes

1. Buat *Node N*.
2. Jika sampel pada kelas yang sama, maka *N* akan menjadi *leaf node* dan diberi nama dengan nama kelas kemudian proses berhenti. Jika tidak, maka lanjutkan proses ke langkah 3.
3. Untuk setiap atribut tes, lakukan perhitungan *information gain* masing-masing atribut tes.
4. Pilih atribut tes, atribut dengan *information gain* terbesar, sebagai atribut *split* yang akan digunakan untuk membagi data sampel
5. Beri nama pada *node N* dengan atribut *split*.
6. Untuk setiap nilai, a_i , dari atribut *split*. Lakukan langkah 6 sampai dengan langkah 8.
7. Buat sampel baru, s_i , dengan kondisi atribut *split* sama dengan a_i .
8. Lakukan pembuatan *node* menggunakan Prosedur *Decision Tree* dengan s_i sebagai data sampel

baru, dan daftar atribut dari sampel tersebut.

Tujuan dari pengukuran nilai *information gain* adalah untuk memilih atribut yang akan dijadikan cabang pada pembentukan pohon keputusan. Atribut yang memiliki nilai *information gain* yang tertinggi akan dipilih menjadi atribut uji untuk dijadikan cabang pohon. Berikut ini rumus untuk menghitung informasi yang diperlukan mengelompokkan sampel adalah:

$$I(S_1, S_2, \dots, S_m) = - \sum_{i=1}^m \frac{S_i}{|S|} \log_2 \frac{S_i}{|S|}$$

Berikut ini rumus untuk menghitung informasi masing-masing atribut adalah :

$$E(A) = \sum_{j=1}^v \frac{S_{1j} + \dots + S_{mj}}{S} I(S_{1j}, S_2, \dots, S_{mj})$$

dimana $I(S_{1j}, \dots, S_{mj})$ dapat dihitung dengan rumus:

$$I(S_{1j}, S_2, \dots, S_{mj}) = - \sum_{i=1}^m \frac{S_{ij}}{|S_j|} \log_2 \frac{S_{ij}}{|S_j|}$$

Sesudah mendapatkan nilai informasi mengelompokkan sampel dan nilai informasi pada masing-masing atribut, maka nilai *information gain* dapat dihitung dengan rumus:

$$Gain(A) = I(S_1, S_2, \dots, S_m) - E(A)$$

Informasi gain dari setiap atribut akan dihitung, dan atribut yang mendapatkan nilai *gain* yang terbesar akan dipilih menjadi *test attribute* untuk dijadikan cabang dari pohon keputusan yang akan dibentuk. Nilai dengan *gain* yang tertinggi akan dijadikan batas untuk pembagian kelompoknya, dengan rumus $Z \leq Y$ dan $Y > Z$.

IMPLEMENTASI SISTEM

Pada bagian ini, akan dilakukan penginputan untuk melakukan prediksi alamat dengan contoh data masukan yaitu “bausasran DN III gondomanan telp. (0274) 543232” dan dapat

dilihat pada gambar berikut ini :

Gambar 2. Form Prediksi Alamat

Form cleaning parsing akan muncul ketika tombol “PROSES” ditekan serta masukan alamat membutuhkan pembersihan data untuk menyesuaikan pengelompokan data masukan alamat dengan *data warehouse* agar proses analisis dengan metode *decision tree* bisa dilakukan. *Form cleaning parsing* terdiri dari 3 tahap yang dijelaskan dengan gambar-gambar berikut :

Gambar 3. Form Cleaning Parsing Tahap 1

Form Cleaning Parsing Tahap 1 ini ditampilkan ketika sistem mendeteksi data masukan alamat mempunyai lebih dari satu lokasi untuk setiap bagian kelompok yang mengharuskan pengguna untuk memilih salah satu lokasi dari setiap bagian kelompok. Pada contoh data yang dimasukkan, sistem mendeteksi dua lokasi pada bagian kelompok

jalan yaitu 'bausasran' dan 'gondomanan'. Dalam hal ini, jalan 'bausasran' dipilih sebagai kelompok jalan yang benar. Setelah pemilihan selesai, pengguna menekan tombol "Lanjut" untuk melanjutkan ke proses berikutnya.

| Nama | Atribut Tambahan | Keterangan Hasil |
|------|------------------|------------------|
| DN | DN III | |
| | | |
| | | |

Gambar 4. Form Cleanning Parsing Tahap 2

Form *Cleanning Parsing Tahap 2* ini ditampilkan ketika sistem mendeteksi ada bagian dari data masukan alamat yang tidak mempunyai bagian kelompok. Jika bagian data tersebut merupakan atribut tambahan, maka pengguna harus memilih bagian tersebut. Jika tidak, maka bagian data tersebut akan dihilangkan dari data masukan alamat. Pada contoh data yang dimasukkan, sistem mendeteksi kata 'III' sebagai atribut yang tidak mempunyai bagian dalam kelompok. Dalam hal ini, angka 'III' dipilih sebagai atribut tambahan dari kata 'DN'. Setelah pemilihan selesai, pengguna menekan tombol "Lanjut" untuk melanjutkan ke proses berikutnya.

| Sebelum | Setelah |
|---------------|--------------|
| (0274) 543232 | (0274)543232 |
| | |
| | |

Gambar 5. Form Cleanning Parsing Tahap 3

Form *Cleanning Parsing Tahap 3* ini ditampilkan ketika sistem melakukan pembersihan penulisan yang sesuai dengan kriteria penulisan

sistem yang ada pada data masukan alamat. Pada contoh data yang dimasukkan, sistem mendeteksi kata '(0274) 543232' sebagai kata yang penulisannya dianggap tidak tepat sehingga sistem memberikan penulisan yang menurut kriteria sistem benar yaitu '(0274)543232'. Setelah selesai, pengguna menekan tombol "Lanjut" untuk melanjutkan proses.

Gambar 6. Form Konfirmasi Cleanning Parsing

Form Konfirmasi *Cleanning Parsing* ini ditampilkan ketika sistem selesai melakukan proses *Cleanning Parsing* dan menampilkan hasil proses. Jika pengguna ingin menekan tombol "Selesai" maka proses *Cleanning Parsing* akan berhenti lalu sistem akan menampilkan hasil pengelompokan pada form *Prediksi alamat*. Tampilan setelah proses pengelompokan dilakukan dapat dilihat pada gambar berikut:

Gambar 7. Form Prediksi Alamat setelah proses *Cleanning Parsing*

Proses selanjutnya setelah pengelompokan bagian alamat dilakukan, pengguna dapat melakukan analisis untuk mendapatkan hasil prediksi alamat dari metode *decision tree* dengan menekan tombol “ANALISIS” yang berada di bagian tengah sebelah kanan pada *form*. Proses analisis ini dapat menampilkan suatu *tree* jika *decision tree* terbentuk serta menampilkan prediksi alamat. Hasil prediksi dari masukan data dapat dilihat pada gambar berikut ini :

VALIDITAS INPUT SUATU ALAMAT
MENGGUNAKAN DECISION TREE DENGAN ALGORITMA ID3

Alamat: Bausasran DN III gondomansi telp. (0274)543232

TUJUAN: SIMPAN

TIPE ALAMAT : JALAN
NAMA : BAUSASRAN
NO TELP : (0274)543232

KELURAHAN : BAUSASRAN
KODE POS :
KECAMATAH : DANUREJAN
KABUPATEH :

HIM_JALAH (ROOT)
+ ABUBAKAR ALI BAUSASRAN
- BAUSASRAN (KELAS)
+ COKRODIRJAN (KELAS)

PREDIKSI ALAMAT :
JALAN BAUSASRAN DN III,
KELURAHAN BAUSASRAN, KECAMATAH DANUREJAN,
KABUPATEN YOGYAKARTA, KODEPOS 55211, TELP. (0274)543232

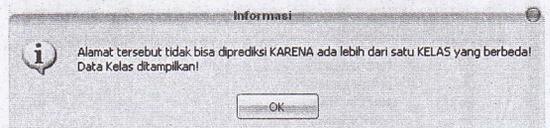
Gambar 8. Form Prediksi Alamat setelah proses ANALISA

Dari contoh kasus data masukan tersebut akan diperoleh suatu hasil prediksi alamat sebagai berikut :

JALAN BAUSASRAN DN III,
KELURAHAN BAUSASRAN,
KECAMATAN DANUREJAN,
KABUPATEN YOGYAKARTA,
KODEPOS 55211, TELP. (0274)543232

Suatu masukan alamat dengan contoh data masukan adalah “JLN. WAHIDIN, KEL. KOTABARU” akan menghasilkan suatu *decision tree* dengan nilai simpul internal terdiri dari dua kelas. Sistem menangani hal tersebut dengan pemberian suatu pesan yang memberikan informasi terhadap pengguna bahwa masukan alamat tersebut

tidak bisa diprediksi serta informasi tentang terbentuknya *decision tree* yang pada simpul internalnya terdapat nilai yang terdiri lebih dari satu kelas. Sistem kemudian menampilkan setiap kelas tersebut. Berikut ini merupakan pesan yang ditampilkan oleh sistem serta contoh kelas yang ditampilkan ketika nilai simpul internal pada *decision tree* terdiri lebih dari satu kelas :



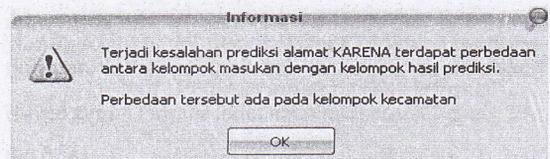
Gambar 9. Pemberian informasi untuk nilai yang lebih dari satu kelas

| Jalan | Kode Pos | Kecamatan | Kabupaten | Kelas |
|-------------|----------|--------------|------------|-------------|
| DR. WAHIDIN | 55222 | GONDOKUSUMAN | YOGYAKARTA | KLITREN LOR |
| DR. WAHIDIN | 55224 | GONDOKUSUMAN | YOGYAKARTA | KOTABARU |

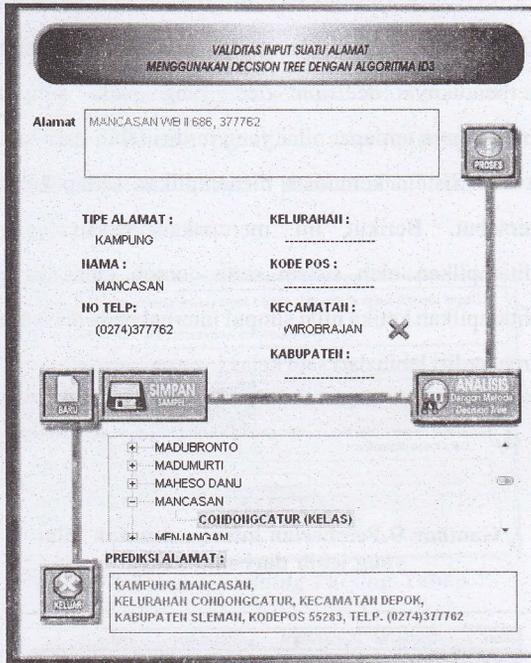
OK

Gambar 10. Informasi data kelas pada nilai yang lebih dari satu kelas

Suatu masukan alamat dengan contoh “MANCASAN WB II 686, 377762” akan menghasilkan suatu hasil prediksi alamat yang salah. Sistem menangani hal tersebut dengan pemberian suatu pesan yang memberikan informasi bahwa terjadi perbedaan hasil pada kelompok masukan dan kelompok hasil prediksi serta memberikan tanda “X” pada bagian kelompok masukan yang tidak sesuai dengan kelompok hasil prediksi. Berikut ini merupakan pesan yang ditampilkan oleh sistem dan tampilan pada *form* prediksi alamat :



Gambar 11. Pesan untuk perbedaan hasil prediksi dengan masukan alamat



Gambar 12. Pemberian tanda "X" pada kelompok masukan alamat yang salah

ANALISIS SISTEM

Pada bagian analisis sistem ini, dilakukan pengujian dengan memasukan data alamat secara acak sebanyak 114 data. Secara keseluruhan, sistem ini dianggap mampu untuk memenuhi kebutuhan pengguna dalam melakukan prediksi alamat.

Berikut ini merupakan pengujian yang dilakukan terhadap sistem untuk menguji hasil prediksi data sampel terhadap kebenaran suatu alamat.

Jumlah data uji sampel alamat = 114 data.

◆ Kebenaran sistem dalam melakukan prediksi

Secara keseluruhan, jumlah data dengan kebenaran prediksi sebanyak 100 data, tetapi tingkat kebenaran tersebut terdiri dari 2 kategori yaitu terbentuk *decision tree* dan tidak terbentuk *decision tree* yang dikarenakan pada tabel sampel hanya terdiri dari 1 data. Berikut ini merupakan jumlah hasil prediksi dengan kategori tersebut :

- Terbentuk *decision tree* : 32 data sampel.
- Tidak terbentuk *decision tree* : 68 data

sampel.

◆ Kesalahan sistem dalam melakukan prediksi

Secara keseluruhan, jumlah data dengan kesalahan prediksi sebanyak 4 data dengan terbentuk *decision tree*.

◆ Data yang tidak bisa diprediksi

Sistem tidak bisa melakukan prediksi pada 10 data sampel yang dimasukkan.

◆ Persentase

- Kebenaran prediksi

persentase secara keseluruhan :

$$\frac{100}{114} \times 100\% = 87,72\%$$

Persentase dengan kategori :

- Terbentuk *decision tree*

persentase :

$$\frac{32}{114} \times 100\% = 28,07\%$$

- Tidak terbentuk *decision tree*

persentase :

$$\frac{68}{114} \times 100\% = 59,65\%$$

- Tidak bisa diprediksi

persentase :

$$\frac{10}{114} \times 100\% = 8,77\%$$

- Kesalahan prediksi, dengan terbentuk *decision tree*

persentase secara keseluruhan :

$$\frac{4}{114} \times 100\% = 3,51\%$$

KEKURANGAN SISTEM

A. Metode *decision tree* tidak dapat melakukan prediksi terhadap suatu kelompok jalan yang mempunyai 2 kelurahan dengan kelurahan

sebagai parameter kelas. Hal ini menyebabkan prediksi terhadap kelompok jalan tersebut tidak terlalu akurat, karena kelompok jalan tersebut terdiri dari dua kelas yang berbeda.

Contoh suatu jalan yang terdiri dari 2 kelas :
Jalan A.M. Sangaji -> kelurahan Cokrodiningratan dan kelurahan Karangwaru
Jalan Sultan Agung -> kelurahan Purwokinanti dan kelurahan Gunungketur

B. Metode *decision tree* tidak dapat melakukan prediksi terhadap suatu nama jalan yang sama tetapi berbeda lokasi. Hal ini menyebabkan prediksi terhadap kelompok jalan tersebut menjadi tidak akurat.

C. Metode *decision tree* tidak dapat melakukan prediksi terhadap suatu jalan yang tidak terdapat dalam tabel sampel karena *decision tree* terbentuk berdasarkan data yang ada dalam tabel sampel. Berdasarkan hal itu, suatu alamat yang tidak terdapat dalam tabel sampel menyebabkan tidak terbentuknya *decision tree* sehingga data tersebut tidak bisa diprediksi.

KESIMPULAN

Berdasarkan perancangan dan implementasi program, maka dapat diperoleh kesimpulan sebagai berikut :

◆ Berdasarkan hasil pengujian, dapat dilihat metode *decision tree* jika diimplementasikan dalam sistem ini untuk memprediksi suatu alamat mempunyai persentase tingkat kebenaran secara keseluruhan sebesar 87,72 %, tingkat kesalahan secara keseluruhan 3,51 %, dan data yang tidak bisa

diprediksi sebesar 8,77 %

◆ Hasil prediksi dari metode *decision tree* menjadi tidak akurat karena tidak mampu menangani suatu data alamat yang terdiri dari dua kelas yang berbeda atau nama yang sama tetapi lokasi yang berbeda.

◆ Metode *decision tree* tidak mampu melakukan prediksi pada data alamat yang tidak terdapat pada data sampel

DAFTAR PUSTAKA

- Oetomo, Budi Sutedjo Dharma, S.Kom.,MM., *Perencanaan & Pembangunan Sistem Informasi*, Penerbit ANDI Offset, Yogyakarta, 2002
- Kristanto, Harianto, Ir., *Konsep Dan Perancangan Database*, Penerbit ANDI Offset, Yogyakarta, 1994
- Han, Jiawei dan Micheline Kamber, *Data Mining : Concepts and Techniques*, Morgan Kaufmann Publisher,. San Francisco, 2001
- Kantardzic, Mehmed, *Data Mining : Concepts, Models, Methods, and Algorithms*, John Willey & Sons, New Jersey, 2003
- Turban, Efraim, *Decision Support Systems and Intelligent Systems*, Fifth Edition, Prentice-Hall Upper Saddle River, New Jersey, 1998
- Munir, Rinaldi, Ir. dan Ir. Leoni Lidya, *Algoritma dan Pemrograman*, CV. Informatika, Bandung, 1997