Volume 4 Nomor 1, April 2021
e-ISSN : 2621-234X

Penggunaan Metode C4.5 Untuk Kasus Klasifikasi

Albert Yohan[1], Hendy Leonardo[2], Christnatalis[3]

Fakultas Teknologi dan Ilmu Komputer, Program Studi Teknik Informatika, Indonesia[1], Fakultas Teknologi dan Ilmu Komputer, Program Studi Teknik Informatika, Indonesia[2], Fakultas Teknologi dan Ilmu Komputer, Program Studi Teknik Informatika, Indonesia

Email: albertyohan90@gmail.com[1], hendileo27@gmail.com[2], christ.natalis@gmail.com[3]

Abstrak

Pada zaman sekarang terdapat banyak bidang pekerjaan yang menggunakan data baik perusahaan maupun toko. Data yang dimiliki tentunya memiliki jumlah yang banyak sehingga kurang dipedulikan dan dianggap tidak penting padahal data tersebut dapat digunakan sebagai informasi untuk mengembangkan usaha mereka. Data mining merupakan metode yang cocok untuk mengolah data data tersebut. Data mining bisa digunakan untuk klasifikasi, prediksi, asosiasi dan regresi. Pada data mining terdapat banyak metode yang dapat digunakan untuk mengolah sekumpulan data yang banyak seperti algoritma C4.5, naive bayes, K-means dan masih banyak lagi. Algoritma C.45 merupakan salah satu metode untuk melakukan klasifikasi. Algoritma C4.5 merupakan algoritma modifikasi dari algoritma ID3. Ada beberapa modifikasi yang dilakukan pada algoritma C.45 seperti bisa menangani atribut kontinu serta missing value.Tujuan dari penulisan artikel berikut yaitu untuk mengetahui kasus yang pernah digunakan peneliti terdahulu dengan metode C4.5 dan hasil akurasi yang didapat terhadap kasus yang dibahas Hasil dari penelitian yang dilakukan peneliti terdahulu dengan metode C4.5 untuk kasus klasifikasi mendapat nilai akurasi yang cukup baik sehingga dapat disimpulkan bahwa metode C4.5 cocok digunakan pada kasus klasifikasi

Keywords *– Kasus Klasifikasi, Data Mining, Algoritma C4.5*

1. Latar Belakang

Data mining merupakan metode untuk memperoleh informasi berharga dari sekumpulan data, dimana proses dilakukan dengan menerapkan berbagai ilmu lain seperti matematika, statistika dan pengenalan pola. Data Mining dapat digunakan dan diterapkan dalam berbagai bidang, seperti bidang kedokteran, pemasaran, pendidikan dan bidang lainnya yang menggunakan data[1]. Data mining dapat digunakan untuk melakukan proses klasifikasi, prediksi, perkiraan dan memperoleh informasi yang bermanfaat dari sekumpulan data berukuran besar. Metode C4.5 merupakan suatu algoritma yang cocok untuk melakukan klasifikasi. Hasil dari klasifikasi dengan metode C4.5 yaitu sebuah pohon keputusan serta rules dari atribut yang dipakai[2].

Metode C4.5 adalah metode klasifikasi yang cukup terkenal untuk mengklasifikasikan data baik data yang bertipe diskrit maupun numerik. Pohon keputusan yang terbentuk dari klasifikasi dengan metode C4.5 dapat menjadi informasi baru dari atribut yang digunakan seperti bisa menemukan hubungan antar atribut ataupun *prediction*[3]. Algoritma C4.5 merupakan algoritma yang dimodifikasi dari dari algoritma ID3. Input dari algoritma ID3 berupa label training, sampel training dan atribut. Beberapa modifikasi yang dilakukan pada metode C4.5 seperti bisa mengatasi missing value, kontinuitas data, dan pruning [4]. Kelebihan utama dari algoritma C4.5 adalah mampu menghasilkan model berupa aturan yang mudah diinterpretasikan ataupun tree, tingkat akurasi yang cukup bagus, mampu memproses atribut dengan tipe numerik dan diskrit. Model yang dihasilkan dari proses "belajar" terhadap data pelatihan pada algoritma C4.5 adalah berupa sebuah pohon keputusan. Pohon keputusan inilah yang akan digunakan untuk melakukan klasifikasi, prediksi ataupun untuk hal lainnya[5].

Untuk mengetahui penerapan metode C4.5 dalam menyelesaikan berbagai kasus klasifikasi dalam kehidupan sehari-hari, maka disusun sebuah artikel penelitian dengan judul "**Penggunaan Metode C4.5 untuk Kasus Klasifikasi**".

2. Metode

Algoritma C4.5 merupakan modifikasi dari algoritma ID3. Ada beberapa modifikasi yang dilakukan pada algoritma C4.5 seperti bisa menangani atribut kontinu serta missing value.

Berikut beberapa kelebihan dari metode C4.5 :

1. Menghasilkan pohon keputusan yang mudah dimengerti.
2. Menghasilkan *rules* yang dapat dijadikan informasi untuk mengklasifikasikan ataupun memprediksi sesuatu
3. Metode C4.5 mampu mengatasi missing value, kontinuitas data dan prunning

Hasil dari klasifikasi dengan metode C4.5 yaitu:

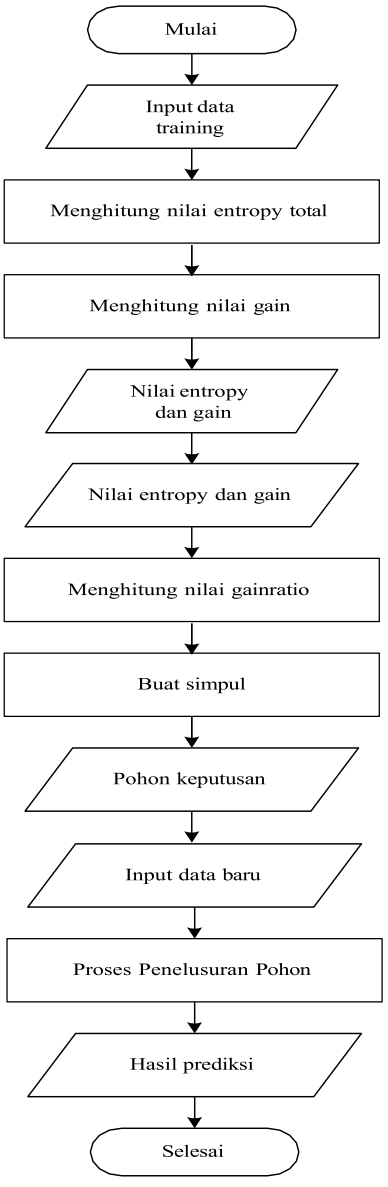
1. Pohon keputusan.
   Pohon keputusan dapat dibentuk dari atribut atribut yang digunakan dari sebuah data. Pada pohon keputusan ini akan didapat root node dimana atribut yang berada di root node merupakan atribut yang paling berpengaruh pada suatu data
2. Sejumlah aturan (*rules).*
   Rules dapat dibentuk dari pohon keputusan yang dibentuk dengan metode C4.5. *Rules* yang terbentuk dapat digunakan untuk mengklasifikasikan ataupun memprediksi sesuatu. Hasil dari *rules* yang terbentuk berupa *if* dan *then*

Decision tree adalah pohon keputusan yang terbentuk dari atribut yang digunakan pada suatu data. Secara umum decision tree terbagi menjadi 3 bagian yaitu simpul root yang merupakan titik awal suatu pohon keputusan, simpul branch yang merupakan percabangan dari akar yang membentuk rangkaian tindakan yang tersedia saat membuat keputusan, dan simpul leaf yang merupakan keputusan akhir untuk suatu decision tree. Decision tree dapat menghasilkan aturan (rules) dari atribut atribut yang digunakan

Tahapan untuk mengolah data dengan algoritma C4.5 sebagai berikut:

1. Mempersiapkan data yang sudah diklasifikasikan terlebih dahulu
2. Lalu tentukan atribut akarnya. Atribut akar dipilih berdasarkan atribut dengan nilai gain yang paling tinggi. Namun sebelum menghitung nilai gain, lakukan perhitungan terhadap nilai entrophy dari masing masing atribut terlebih dahulu.
   Untuk mendapatkan nilai entrophy dapat digunakan rumus:
   Entrophy(S) = $\sum_{j=1}^{k} -pj * log2 * pj$
   Dimana :
   S = banyaknya kasus
   Pj = probabilitas jumlah kasus (ya/tidak)/total kasus
3. Setelah entrophy dari masing masing atribut telah didapat maka lakukan perhitungan terhadap nilai gain.
   Untuk mendapatkan nilai gain dapat digunakan rumus
   Gain = entrophy(S) - $\sum_{j=1}^{k} \frac{|Sj|}{|S|} * entrophy(S)$
   Dimana :
   S = banyaknya kasus
   Sj = jumlah kasus pada partisi j
   S = jumlah seluruh sampel data
4. Lakukan langkah kedua kembali hingga semua atribut memiliki kelas seperti kelas ya dan kelas tidak

Proses kerja dari algoritma C4.5 dapat digambarkan dalam bentuk flowchart seperti terlihat pada gambar 1.



Gambar 1. *Flowchart* dari Algoritma C4.5

3. Hasil dan Pembahasan

3.1 Hasil

Berdasarkan jurnal yang telah dikumpulkan, penulis akan menguraikan ulasan dari beberapa jurnal tersebut sebagai berikut : (1) "Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing" yang ditulis oleh Eka Fitriani, Riska Aryanti, Atang Saepudin, Dian Ardiansyah pada tahun 2020. Penelitian ini menggunakan metode C4.5 untuk klasifikasi apakah seseorang lolos atau tidak dalam penempatan tenaga marketing. Akurasi yang didapat pada penelitian ini sebesar 91,10%. (2) "Klasifikasi Faktor Faktor Penyebab Penyakit Diabetes Melitus di Rumah Sakit UNHAS Menggunakan Algoritma C4.5" yang ditulis oleh Dewi Rahma Ente, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin, Andreza pada tahun 2020. Penelitian ini menggunakan metode C.45 untuk mengetahui faktor yang mempengaruhi seseorang terkena penyakit diabetes mellitus. Rata rata akurasi yang didapat pada penelitian ini sebesar 98,5%.

Dalam jurnal lain dengan judul (3) "Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Menggunakan Algoritma C4.5" yang ditulis oleh Putri Ayu Mardhiyah, Riki Ruli A Siregar, Pritasari Paluningsih pada tahun 2020. Penelitian ini menggunakan metode C4.5 untuk prediksi seorang nasabah macet atau tidak dalam melakukan pembayaran kredit. Akurasi yang didapat pada penelitian ini sebesar 70,93%. (4) "Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika" yang ditulis oleh Tri Novika, Poningsih, Harly Okprana, Agus Perdana Windarto, Hasudungan Siahaan pada tahun 2021. Penelitian ini menggunakan metode C4.5 untuk klasifikasi konsep pemahaman siswa pada pelajaran matematika. Akurasi yang didapat pada penelitian ini sebesar 96%. (5) "Analisa Klasifikasi C4.5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi" yang ditulis oleh Khairunnissa Fanny Irnanda, Dedy Hartama, Agus Perdana Windarto pada tahun 2021. Penelitian ini menggunakan metode C4.5 untuk klasifikasi faktor penyebab menurunnya prestasi belajar mahasiswa pada masa pandemi. Akurasi yang didapat pada penelitian ini sebesar 97,5%. (6) "Penerapan Algoritma Decision Tree C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit Santa Maria Pemalang" yang ditulis oleh Sigit Abdillah pada tahun 2015. Penelitian ini menggunakan metode C4.5 untuk klasifikasi pasien yang terkena penyakit stroke. Akurasi yang didapat dengan 130 data training sebesar 82,31% dan pada 26 data testing sebesar 76,92%

Berikut ini adalah tabel pengukuran dari enam jurnal yang telah penulis kumpulkan tentang penggunaan metode C4.5 untuk kasus klasifikasi. Jurnal tersebut digunakan sebagai data dalam penelitian ini. Tabel pengukuran dari enam jurnal yang telah penulis review dapat dilihat pada tabel 1.

Volume 4 Nomor 1, April 2021
e-ISSN : 2541-2019

Tabel 1. Pengukuran Perbandingan

Judul	Nama peneliti		Pengukuran			
		Kasus	Dataset	Atribut	Akurasi	Hasil
Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing	Eka Fitriani, Riska Aryanti, Atang Saepudin, Dian Ardiansyah	Metode C.45 digunakan untuk klasifikasi seseorang lolos atau tidak dalam penempatan tenaga marketing	Dataset yang digunakan yaitu data 359 calon marketing dari perusahaan *outsourcing*	Atribut yang digunakan sebanyak 8 atribut yaitu jenis kelamin, tinggi badan, berat badan, usia, jarak dari tempat tinggal ke tempat kerja, pengalaman, tes kapabilitas, targeting 3 bulan	Akurasi yang didapat pada penelitian ini sebesar 91,10%	Pada pohon keputusan yang terbentuk atribut targeting 3 bulan menjadi atribut yang paling berpengaruh terhadap seseorang lolos atau tidak dalam penempatan tenaga marketing
Klasifikasi Faktor Faktor Penyebab Penyakit Diabetes Melitus di Rumah Sakit UNHAS Menggunakan Algoritma C4.5	Dewi Rahma Ente, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin, Andreza	Metode C4.5 digunakan untuk mengetahui faktor yang mempengaruhi seseorang terkena penyakit diabetes melitus	Dataset yang digunakan pada penelitian ini yaitu 127 data pasien dari rumah sakit pendidikan UNHAS	Atribut yang digunakan sebanyak 9 atribut yaitu jenis kelamin, usia, berat badan, tinggi badan, GDP, HDL, LDL, kolestrol total, trigleserida	Rata rata akurasi yang didapat pada penelitian ini sebesar 98,5%	Pada pohon keputusan yang terbentuk atribut GDP (glukosa darah puasa) menjadi atribut yang paling berpengaruh terhadap seseorang terkena penyakit diabetes melitus
Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Menggunakan Algoritma C4.5	Putri Ayu Mardhiyah, Riki Ruli A Siregar, Pritasari Paluningsih	Metode C4.5 digunakan untuk prediksi seorang nasabah macet atau tidak dalam melakukan pembayaran kredit	Dataset yang digunakan pada penelitian ini diambil dari UCI *Machine Learning*	Atribut yang digunakan sebanyak 6 atribut yaitu jumlah kredit, status, usia, bulan-1, bulan-2, bulan-3	Akurasi yang didapat pada penelitian ini sebesar 70,93%	Pada pohon keputusan yang terbentuk atribut bulan-3 menjadi atribut yang paling berpengaruh terhadap seseorang macet atau tidak dalam melakukan pembayaran kredit

Judul	Nama peneliti		Pengukuran			
		Kasus	Dataset	Atribut	Akurasi	Hasil
Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika	Tri Novika, Poningsih, Harly Okprana, Agus Perdana Windarto, Hasudungan Siahaan	Metode C.45 digunakan untuk klasifikasi konsep pemahaman siswa pada pelajaran matematika	Dataset yang digunakan pada penelitian ini yaitu hasil kuesioner dari siswa kelas VIII T/A 2019/2020	Atribut yang digunakan sebanyak 6 atribut yaitu minat siswa, cara belajar siswa, motivasi siswa, cara mengajar guru, media pembelajaran, sarana dan prasarana	Akurasi yang didapat pada penelitian ini sebesar 96%	Pada pohon keputusan yang terbentuk motivasi siswa menjadi atribut yang paling berpengaruh terhadap seoramg siswa memahami pelajaran matematika
Analisa Klasifikasi C4.5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi	Khairunnissa Fanny Irnanda, Dedy Hartama, Agus Perdana Windarto	Metode C4.5 digunakan untuk klasifikasi faktor penyebab menurunnya prestasi belajar mahasiswa pada masa pandemi	Dataset yang digunakan pada penelitian ini yaitu dari wawancara dan angket yang diberikan kepada mahasiswa semester 7 STIKOM Tunas Bangsa	Atribut yang digunakan sebanyak 5 atribut yaitu cara belajar, waktu belajar, pemahaman materi, pemberian tugas, lingkungan	Akurasi yang didapat pada penelitian ini sebesar 97,5%	Pada pohon keputusan yang terbentuk pemahaman materi menjadi atribut yang paling berpengaruh terhadap menurunnya prestasi belajar seorang mahasiswa
Penerapan Algoritma Decision Tree C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit Santa Maria Pemalang	Sigit Abdillah	Metode C.45 digunakan untuk klasifikasi pasien yang terkena penyakit stroke	Dataset yang digunakan pada penelitian ini yaitu data pasien dari rumah sakit santa maria pemalang sebanyak 156 data	Atribut yang digunakan hanyalah atribut yang paling mempengaruhi keakuratan metode C4.5 untuk klasifikasi penyakit stroke sebanyak 3 atribut yaitu umur, hipertensi, diabetes	Akurasi yang didapat dengan 130 data training sebesar 82,31% dan pada 26 data testing sebesar 76,92%	Pada pohon keputusan yang terbentuk diabetes menjadi atribut yang paling berpengaruh terhadap seseorang terkena penyakit stroke

3.2 Pembahasan

Dari beberapa jurnal yang telah penulis baca dan *review*, penulis menemukan satu jurnal yang membahas tentang penggunaan metode C4.5 untuk kasus klasifikasi dengan jelas dan terperinci. Jurnal tersebut memiliki kelebihan dari jurnal lain yang penulis baca karena peneliti pada jurnal tersebut tidak menggunakan semua atribut yang ada namun melakukan seleksi atribut terlebih dahulu dengan diskusi dengan pakar bidang kesehatan pada rumah sakit yang menjadi objek penelitiannya untuk mengetahui atribut mana saja yang paling mempengaruhi keakuratan terhadap objek yang diteliti

Judul : "Penerapan Algoritma Decision Tree C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit Santa Maria Pemalang"

Peneliti: Sigit Abdillah

Tabel 2. Pengukuran

Kasus yang dibahas	Metode C.45 digunakan untuk klasifikasi pasien yang terkena penyakit stroke
Atribut yang digunakan	Dari 17 atribut yang didapat, peneliti hanya menggunakan 3 atribut yang paling mempengaruhi keakuratan dalam pengaruh penyakit stroke yaitu umur, hipertensi, dan diabetes
Akurasi	Akurasi yang didapat dengan 130 data training sebesar 82,31% dan pada 26 data testing sebesar 76,92%
Jumlah dataset	Dataset yang digunakan pada penelitian ini yaitu data pasien dari rumah sakit santa maria pemalang sebanyak 156 data
Hasil	Pada pohon keputusan yang terbentuk diabetes menjadi atribut yang paling berpengaruh terhadap seseorang terkena penyakit stroke

4. Kesimpulan dan Saran

4.1 Kesimpulan

Kesimpulan yang dapat diambil dari artikel tersebut sebagai berikut :

1. Metode C4.5 cocok digunakan untuk kasus klasifikasi dan menghasilkan nilai akurasi yang cukup baik.
2. Pemilihan atribut yang digunakan sangat berpengaruh terhadap rules yang dihasilkan.
3. Semakin banyak data yang digunakan tidak menjamin akan menghasilkan nilai akurasi yang tinggi.

4.2 Saran

Saran yang dapat diberikan untuk artikel tersebut sebagai berikut :

1. Perlunya data yang kompleks agar mendapat nilai akurasi yang tinggi.
2. Perlunya melakukan seleksi data untuk memilih atribut mana saja yang paling berpengaruh terhadap penelitian yang dilakukan agar menghasilkan pohon keputusan yang akurat.

Referensi

[1] E. Fitriani, R. Aryanti, A. Saepudin, and D. Ardiansyah, "Penerapan Algoritma C4.5 Untuk Klasifikasi Penempatan Tenaga Marketing," *Paradig. - J. Komput. dan Inform.*, vol. 22, no. 1, pp. 72–78, 2020, doi: 10.31294/p.v22i1.6898.

[2] D. R. Ente, S. A. Thamrin, S. Arifin, H. Kuswanto, and A. Andreza, "Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5," *Indones. J. Stat. Its Appl.*, vol. 4, no. 1, pp. 80–88, 2020, doi: 10.29244/ijsa.v4i1.330.

[3] M. P. A. S. R. R. A. P. Pritasari, "Klasifikasi Untuk Memprediksi Pembayaran Kartu Kredit Macet Jurnal Teknologia," *J. Teknol.*, vol. 3, no. 1, pp. 91–101, 2020.

[4] T. Novika, H. Okprana, A. P. Windarto, and H. Siahaan, "Penerapan Data Mining Klasifikasi Tingkat Pemahaman Siswa Pada Pelajaran Matematika," vol. 5, pp. 9–17, 2021, doi: 10.30865/mib.v5i1.2498.

[5] K. F. Irnanda, D. Hartama, and A. P. Windarto, "Analisa Klasifikasi C4 . 5 Terhadap Faktor Penyebab Menurunnya Prestasi Belajar Mahasiswa Pada Masa Pandemi," *J. Media Inform. Budidarma*, vol. 5, no. 1, pp. 327–331, 2021, doi: 10.30865/mib.v5i1.2763.

[6] S. Abdillah, "Penerapan Algoritma Decision Tree C4.5 Untuk Diagnosa Penyakit Stroke Dengan Klasifikasi Data Mining Pada Rumah Sakit Santra Maria Pemalang," *J. Tek. Inform.*, pp. 1–12, 2011.

[7] A. Asroni, B. Masajeng Respati, and S. Riyadi, "Penerapan Algoritma C4.5 untuk Klasifikasi Jenis Pekerjaan Alumni di Universitas Muhammadiyah Yogyakarta," *Semesta Tek.*, vol. 21, no. 2, pp. 158–165, 2018, doi: 10.18196/st.212222.