PRISMA 3 (2020): 72-76
PRISMA, PROSIDING SEMINAR NASIONAL MATEMATIKA
https://journal.unnes.ac.id/sju/index.php/prisma/
ISSN 2613-9189

# Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C5.0 untuk Atribut Kontinu dan Diskrit

Afif Zakiy Abdullah[a,*], Dewi Retno Sari Saputro[b], Bowo Winarno[c]

[a, b, c] *Program Studi Matematika FMIPA Universitas Sebelas Maret, Jl. Ir Sutami No.36 A, Surakarta 57126, Indonesia*

** Alamat Surel: abdullahazakiy@student.uns.ac.id*

Abstrak

Data mining merupakan proses semi automatis yang menggunakan teknik statistik, matematika, keceradan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan dalam *database.* Klasikasi dengan pohon keputusan merupakan salah satu teknik data mining. Pohon keputusan adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap *node* internal menandakan suatu tes pada atribut dan setiap cabang merepresentasikan hasil tes, dan *node* daun merepresentasikan kelas-kelas. Alur pada pohon keputusan ditelusuri dari *node* akar ke *node* daun yang memegang prediksi kelas. Pohon keputusan terdiri dari *node* yang membentuk pohon yang berakar, semua *node* memiliki satu masukan. *Node* yang keluar disebut *node* tes. *Node* yang lain disebut *node* keputusan atau sering disebut *node* daun. Setiap *node* internal membagi dua atau lebih sub-ruang sesuai dengan kategori atribut dan akan dipartisi sesuai dengan nilai kategori kasus. Kasus-kasus tersebut membentuk pohon keputusan, yang menghasilkan *problem solving*. Dalam menentukan pohon keputusan ada beberapa algoritme salah satunya adalah algoritme C5.0. Pohon keputusan berbasis algoritme C5.0 merupakan penyempurnaan dari algoritme ID3 dan C4.5. Algortime C5.0 dapat menangani atribut kontinu dan diskrit yang tidak dapat ditangani oleh algoritme ID3. Pada penelitian dilakukan kajian klasifikasi dengan pohon keputusan berbasis algoritme C5.0.

Kata kunci:

data mining, pohon keputusan, klasifikasi, algoritme C5.0



## 1. Pendahuluan

Data mining merupakan proses semi automatis yang menggunakan teknik statistik, matematika, keceradan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan dalam *database*. Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu deskripsi, estimasi, prediksi, klasifikasi, pengklasteran, dan asosiasi. Klasfikasi dengan pohon keputusan merupakan salah satu teknik data mining. Pohon keputusan adalah struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap *node* internal menandakan suatu tes pada atribut dan setiap cabang merepresentasikan hasil tes, dan *node* daun merepresentasikan kelas-kelas. Dalam menentukan pohon keputusan ada beberapa algoritme salah satunya adalah algoritme C5.0.

Pohon keputusan berbasis algoritme C5.0 merupakan penyempurnaan dari algoritme ID3 dan C4.5. Algoritme ID3, C4.5 dan C5.0 merupakan algoritme yang ditemukan oleh J. R. Quinlan, dimana C4.5 merupakan penyempurnaan dari ID3. Beberapa penelitian sebelumnya memberikan gambaran bahwa, C5.0 merupakan algoritme dalam klasifikasi yang menggunakan waktu lebih singkat dibandingkan algoritme yang lain, memori yang digunakan sedikit, dan memberikan fasilitas berupa pemilihan fitur, *cross validation*, pengurangan *error pruning* serta kompleksitas model (Galathiya *et al.*, 2012) dan (Pandya *et al.*, 2015), C5.0 memiliki akurasi yang tinggi pada penerapan kasus lalu lintas jaringan komputer dengan tujuh aplikasi yang berbeda (Bujlow *et al*., 2012). C5.0 lebih unggul dari C4.5 pada besar pohon keputusan, lamanya proses, dan besar eror yang dihasilkan dengan penerapannya pada data

*To cite this article:*
Abdullah, A.Z., Saputro, D.R.S, & Winarno, B. (2020). Klasifikasi dengan Pohon Keputusan Berbasis Algoritme C5.0 untuk Atribut Kontinu dan Diskrit. *PRISMA, Prosiding Seminar Nasional Matematika 3*, 72-76

kanker Tiroid (Upadhayay *et al.*, 2013), C5.0 diusulkan untuk menggunakan *rough set theory-information entropy- discernible matrix discretization* (*RSIEDM*) yang lebih rasional, lebih akurat untuk atribut diskrit secara kontinu, dan pohon keputusan yang dihasilkan lebih akurat, didasarkan pada penelitian terhadap statistik becana kilat (Hou *et al.*, 2014), penggunaan C5.0 dan *one-class* SVM sukses mendeteksi penyalahgunaan dan anomali sistem deteksi intrusi (Rani & Xavier, 2015), C5.0 dapat mengklasifikasikan atau memprediksi data hasil dengan baik dan akurasi sebesar 70.4% (Septiandi, 2016), penggunaan C5.0 sebagai pembuat kelas-kelas untuk diterapkan pada pemantauan perubahan dinamis perkebunan karet di China (Sun *et al*., 2017), penggunaan algoritme C5.0 dapat menghasilkan model dengan akurasi tinggi dalam menentukan faktor yang dominan mempengaruhi permintaan kartu perdana internet studi kasus Vidha Ponsel (Hutabarat, 2018). Pada penelitian dilakukan kajian klasifikasi dengan algoritme C5.0 berbasis pohon keputusan.

## 2. Pembahasan

Data mining adalah suatu proses semi automatis (belum sepenuhnya automatis) dalam menganalisis data dalam jumlah yang besar untuk memperoleh suatu pengetahuan, selain itu ada juga yang mengatakan bahwa Data mining merupakan metode yang dapat digunakan untuk memperoleh suatu informasi sebagai pertimbangan dalam mengambil keputusan. Data mining *is predicted to be* "*one of the most revolutionary developments of the next decade*". Data mining *is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognotion technologies as well as statistical and mathematical techniques* (Larose, 2005). Data mining merupakan proses automatis terhadap data yang sangat besar yang sudah ada untuk mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat (Kusrini & Luthfi, 2009). "Menurut Turban dkk, data mining adalah proses yang menggunakan teknik statistik, matematika, keceradan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan dalam *database* besar" (Kusrini & Luthfi, 2009). Dari definisi-definisi yang telah disampaikan, data mining dapat dituliskan sebagai proses semi automatis yang menggunakan teknik statistik, matematika, keceradan buatan dan *machine learning* untuk mengekstraksi dan mengidentifikasi informasi pengetahuan yang potensial dan bermanfaat yang tersimpan dalam *database*.

Data mining dibagi menjadi beberapa kelompok berdasarkan tugas yang dapat dilakukan, yaitu deskripsi, estimasi, prediksi, klasifikasi, pengklusteran, dan asosiasi (Larose, 2005). Klasifikasi sebagai salah satu teknik data mining merupakan proses menentukan model atau fungsi yang membedakan konsep atau kelas data, dengan tujuan dapat memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Menurut Kashyap dan Chauhan (2016), klasifikasi merupakan metode untuk memberikan atau membuat kelas data berdasarkan kelas kategorikal yang telah ditentukan, Ada beberapa algoritme yang digunakan untuk proses klasifikasi, yaitu pohon keputusan, *Naïve Biased Classification*, *Generalized Linear Models* (GLM), *Super Vector Machine* dan lain-lain.

### *2.1. Pohon keputusan (Decision Tree)*

Pohon keputusan merupakan struktur *flowchart* yang menyerupai *tree* (pohon), dimana setiap *node* internal menandakan suatu tes pada atribut dan setiap cabang merepresentasikan hasil tes, dan *node* daun merepresentasikan kelas-kelas. Proses klasifikasi untuk memeroleh pohon keputusan dapat dilakukan dengan beberapa algoritme, seperti ID3, C4.5, C5.0, dan lainnya.

Algoritme *generate decision tree* sebagai berikut (Han *et al.*, 2012):

- partisi data, D, data percobaan yang sudah ditentukan label kelasnya,
- *attribute_list*, himpunan yang terdiri dari kandidat atribut, dan
- *attribute_selection_method*, prosedur untuk menentukan kriteria pemotongan yang berisi *tuple* data terbaik ke kelas masing-masing.

Menurut Han *et al.* (2012) algoritme klasifikasi pohon keputusan diuraikan sebagai berikut:

- membuat *node* N;
- jika semua *tupel* di D memiliki kelas yang sama misal C, Maka *node* N sebagai *leaf node* (*node* daun) dan diberi label dengan kelas C;

- jika *attribute_list* kosong, Maka jadikan *node* N sebagai *node* daun dan diberi label sama dengan nilai kelas terbanyak pada sampel;
- menerapkan *attribute_selection_method* (D, *attribute_list*) untuk memperoleh *splitting_creterion* terbaik;
- memberi label *node* N dengan *splitting_creterion*;
- jika atribut bernilai diskrit dan diperbolehkan untuk dipisah (*multiway splits*), maka *attribute_list* ← *attribute_list - splitting_attribute*;
- untuk setiap nilai j dari atribut percobaan yang diketahui;
  - Buat $D_j$ menjadi kumpulan data *tuple* D untuk memenuhi hasil j,
  - Jika $D_j$ kosong, maka Tambahkan *node* daun yang diberi label sama dengan nilai kelas yang terbanyak pada D ke *node* N,
  - Selainnya, tambah cabang baru di bawah dengan memanggil fungsi *generate_decision_tree* ($D_j$, *attribute_list*) ke *node* N.
- kembali ke N.

*2.2. Algoritme C5.0*

Algoritme C5.0 merupakan pengembangan atau perbaikan dari algoritme C4.5 dimana pengembangan dilakukan pada penggunaan memori, pemisahan atribut, dan percepatan (Rulequest, 2019). Menurut Quinlan (1993), algoritme C4.5 merupakan pengembangan atau perbaikan dari algoritme ID3, perbaikan yang dilakukan seperti dapat menangani atribut diskrit atau kontinu, dapat menangani atribut yang kosong atau *missing value*, dapat meringkas atau memangkas pohon keputusan, sementara ID3 ditemukan oleh Quinlan (1986). Menurut Hou *et al*. (2014), dengan mengasumsikan bahwa *S* adalah himpunan sampel percobaan *X* membatasi *n* atribut yang membagi *S* menjadi *n* himpunan bagian ($S_1, S, S_3, \cdots, S_n$), jumlahan sampel *S* adalah $|S|$, jumlahan sampel yang termuat pada $C_i$ dimana *I = 1, 2, ..., n* adalah $freq(C_i, S)$, peluang setiap kelas yang termuat pada $C_i$ adalah $\frac{freq(C_i,S)}{|S|}$ sehingga informasi yang disampaikan adalah $-\log_2\left(\frac{freq(C_i,S)}{|S|}\right)$ bits, dan untuk konten informasi kasar yang diperlukan untuk mengidentifikasi samua sampel pada *S* adalah *info*(*S*) atau dapat dituliskan pada persamaan (2.1)

$$info(S) = -\sum_{i=1}^{n} \frac{freq(C_i, S)}{|S|} \log_2\left(\frac{freq(C_i, S)}{|S|}\right) \tag{2.1}$$

setelah membagi *S* menjadi *n* himpunan bagian, informasi entropi dapat dihitung. Nilai ekspektasi *S* ditulis sebagai

$$info_X(S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} info(S_i) \tag{2.2}$$

untuk mengukur *info*(*S*) pada partisi *X* berdasarkan verifikasi atribut, *gain*(*X*) yang terlihat pada persamaan (2.3) dapat digunakan. Atribut dengan info gain yang besar dapat diselesaikan dengan cara memartisi dengan rumus yang ditulis sebagai

$$gain(X) = info(S) - info_X(S) \tag{2.3}$$

untuk menghitung kriteria *gain_ratio*, hitung *split info* dengan menganalogikan dengan persamaan (2.1) diperoleh

$$Split\_info(X) = -\sum_{i=1}^{n} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

didapat

$$Gain\_ratio(X) = \frac{gain(X)}{Split\_info(X)}.$$

Menurut Revathy dan Lawrance (2017) klasifikasi dengan pohon keputusan berbasis algoritme C5.0 diuraikan sebagai berikut:

*Input:*

- partisi data, S, satu set tupel pelatihan dan label kelas yang terkait,
- *attribute_list*, himpunan kandidat atribut, dan

- *attribute_selection_method*, prosedur untuk menentukan kriteria pemotongan yang berisi *tuple* data terbaik ke kelas masing-masing. Kriteria yang dimaksud terdiri dari *splitting_attribute* dan *split-point* atau *splitting subset*.

*Output:* pohon keputusan C5.0.

Metode:

- membuat *node* N;
- jika semua *tupel* di *S* memiliki kelas yang sama misal $S_i$ $(i = 1,2, \dots, n)$, maka *node* N sebagai *leaf node* (*node* daun) dan diberi label dengan kelas $S_i$;
- jika *attribute_list* kosong, maka jadikan *node* N sebagai *node* daun dan diberi label sama dengan nilai kelas terbanyak pada sampel;
- menerapkan *attribute_selection_method* (*S*, *attribute_list*) untuk memperoleh *splitting_creterion* terbaik;
- memberi label *node* N dengan *splitting_creterion*;
- jika atribut bernilai diskrit dan diperbolehkan untuk dipisah (*multiway splits*), maka *attribute_list* ← *attribute_list* - *splitting_attribute*
- untuk setiap nilai *j* dari atribut percobaan yang diketahui;
  - Buat $S_j$ menjadi kumpulan data *tuple S* untuk memenuhi hasil *j*,
  - Jika $S_j$ kosong, maka Tambahkan *node* daun yang diberi label sama dengan nilai kelas yang terbanyak pada *S* ke *node* N,
  - Selainnya, tambahkan *node* ke pohon keputusan hasil dari penghitungan C5.0 ($S_j$, *attribute_list*) ke *node* N.
- kembali ke N.

### *2.3. Atribut algoritme C5.0*

See5/C5.0 adalah alat penambangan data yang canggih untuk menemukan pola yang menggambarkan kategori, menggabungkannya ke dalam klasifikasi, dan menggunakannya untuk membuat prediksi (Rulequest, 2019).

Secara umum, terdapat dua subkelompok atribut penting yang dapat diolah oleh C5.0, yaitu

1. Nilai dari atribut yang didefinisikan secara eksplisit, ia terdapat pada data dalam salah satu dari beberapa bentuk.
   a) atribut diskrit memiliki nilai yang diambil dari sekumpulan nilai nominal,
   b) atribut kontinu memiliki nilai numerik,
   c) atribut tanggal menyimpan tanggal kalender,
   d) atribut waktu memegang waktu jam,
   e) atribut cap waktu memegang tanggal dan waktu, dan
   f) label atribut hanya berfungsi untuk mengidentifikasi kasus tertentu.
2. Nilai atribut yang didefinisikan secara implisit ditentukan oleh rumus.

Pada atribut diskrit memiliki cabang terpisah untuk setiap nilai yang ada dalam data, sementara atribut kontinu dibuat menjadi kontinu sepotong-sepotong atau dibuat menjadi beberapa interval tertentu (Rulequest, 2019).

## 3. Simpulan

Berdasarkan pembahasan yang telah dilakukan, dapat disimpulkan bahwa klasifikasi dengan pohon keputusan berbasis algoritme C5.0 merupakan penyempurnaan dari algoritme ID3 dan C4.5. Algortime C5.0 dapat menangani atribut kontinu dan diskrit yang tidak dapat ditangani oleh algoritme ID3.

## Daftar Pustaka

Bujlow, T., Riaz, T., & Pedersen, J. M. (2012). A Method for Classification of Network Traffic Based on C5.0 Machine Learning Algorithm. In *Proceeding of ICNC'12: 2012 International Conference on*

*Computing, Networking and Comminocations (ICNC): Workshop on Computing, Networking and Communications* (pp. 244-248). Hawaii, USA.

Galathiya, A. S., Ganatrand, A. P., & Bhensdadia, C. K. (2012). Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. *International Journal of Computer Science and Information Technologies*, 3(2), 3427-3431.

Han, J., Kamber, M., & Pei, J. (2012). Data Mining: Concepts and Techniques Third Edition. Morgan Kaufmann. Massachusetts (US).

Hou, S., Hou, R., Shi, X., Wang, J., & Yuan, C. (2014). Research on C5.0 Algorithm Improvement and the Test in Lightning Disaster Statistics. *International Journal of Control and Automation*, 7(1), 181-190.

Hutabarat, C. (2018). Penerapan Data Mining untuk Memprediksi Permintaan Produk Kartu Perdana Internet Menggunakan Algoritma C5.0 (Studi Kasus: Vidha Ponsel). Jurnal Pelita Informatika, 17(2), 168-173.

Kashyap, G., & Chauhan, E. (2016). Parametric Comparisons of Classification Techniques in Data Mining Applications. *International Journal of Engineering Development and Research*, 4(2), 1117-1123.

Kusrini & Luthfi, E. T. (2009). Algoritma Data Mining. ANDI. Yogyakarta.

Larose, D. T. (2005). Discovery Knowledge in Data: an Introduction to Data Mining. John Willey & Sons, Inc. Hoboken, New Jersey.

Pandya, R., & Pandya, J. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications*, 117(16), 18-21.

Quinlan, J. R. (1986). Induction of Decision Trees. Kluwer *Academic Publishers*, Boston.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*. San Mateo, California.

Rani, M. S., & Xavier, S. B. (2015). A Hybrid Intrusion Detection System Based on C5.0 Decision Tree and One-Class SVM. *International Journal of Current Engineering and Technology*, 5(3), 2001-2007.

Revathy, R., & Lawrance, R. (2017). Comparative Analysis of C4.5 and C5.0 Algorithms on Crop Pest Data. *International Journal of Innovative Research in Computer and Communication Engineering*, 5(1), 50-58.

Rulequest. (2019). Data Mining Tools See5 and C5.0. (*online*). (https://www.rulequest.com/see5-info.html, diakses pada 31 Mei 2019).

Septiandi, R. (2016). Analisis Data Keterlambatan Bahan Baku Berdasarkan Pendekatan Data Warehouse dan Pohon Keputusan C5.0. (*Master's Thesis*). Institut Pertanian Bogor, Bogor.

Sun, Z., Leinenkugel, P., Guo, H., Huang, C., & Kuenzer, C. (2017). Extracting Distribution and Expansion of Rubber Plantations from Landsat Imagery Using The C5.0 Decision Tree Method. *Journal of Applied Remote Sensing*, 11(2), 026011-1–026011-21.

Upadhayay, A., Shukla, S., & Kumar, S. (2013). *International Journal of Computer Science and Communication Networks*, 3(1), 64-68.