FORMAT: Jurnal Ilmiah Teknik Informatika
Akreditasi Sinta 5 RISTEKBRIN No. 164/E/KPT/2021; E-ISSN 2722-7162 | P-ISSN 2089-5615

Vol. 13, No. 1,
January 2024,
1-8. 2024

STUDI TENTANG ALGORITMA C5.0 DALAM MEMPREDIKSI KEPATUHAN NASABAH DALAM MEMBAYAR PAJAK PERTAMBAHAN NILAI

Moch. Anjas Aprihartha[1,]; M.Husniyadi [2]; Taufik Nur Alam [3]

[1,2,3] Fakultas Ilmu Komputer, Universitas Dian Nuswantoro, Jl Imam Bonjol No. 207, Semarang Jawa Tengah

[1] anjas.aprihartha@dsn.dinus.ac.id, [2] 1182022000012@mhs.dinus.ac.id, [3] 118202300068@mhs.dinus.ac.id

Kata kunci:
C5.0, data mining, decision tree, tax, predict

Abstract

Data mining is the process of extracting valuable patterns, information, and knowledge from large data sets. Data mining has an important role in identifying and minimizing risks in various lives. One of the algorithms of data mining is the decision tree type C5.0. The C5.0 algorithm is an algorithm used to solve classification problems. The C5.0 method can be applied in various sectors such as the taxation sector. Paying taxes is an obligation by an individual or entity paid to the State. The value added tax is the highest contributory tax because it is collected several times to companies. Factors that affect customer compliance in paying value added tax are income, entity form, and reporting status. This study aims to predict public compliance in paying value added tax using the C5.0 method. This research aims to produce a classification model that can be a potential solution for dealing with prediction problems in customer compliance with paying taxes. The results of the study obtained an average accuracy of the C5.0 model of 66,5%. Based on this accuracy value, the model can be categorized as still weak in predicting the status of value added tax payments.

Pendahuluan

A. Latar Belakang

Penambangan data (*data mining*) merupakan proses penggalian pola, informasi, dan pengetahuan berharga dari kumpulan data besar. Dalam proses penggalian pola tersebut maka perlu menggunakan alat dan teknik tertentu untuk menganalisis data dari sumber data yang besar [1]. Teknik yang mendukung dan paling banyak digunakan dalam data mining adalah *machine learning*. Dalam pempelajari tentang *machine learning*, memahami algoritma pembelajaran mesin adalah hal yang sangat penting. Salah satu algoritma yang banyak digunakan adalah *supervised learning*, yang biasanya digunakan untuk prediksi atau klasifikasi [2]. Salah satu algoritma dalam *supervised learning* adalah algoritma *decision tree* tipe C5.0.

D*ecision tree* tipe C5.0 merupakan algoritma yang digunakan untuk mengatasi kelemahan C4.5 dan algoritma ID3 dan mengurangi ukuran pohon keputusan dengan memangkas bagian pohon [3]. Jika dalam proses pembentukan pohon, algoritma C4.5 melibatkan perhitungan *entropy* dan *information gain*, namun pada algoritma C5.0 dilakukan

perhitungan tambahan yaitu *gain ratio. Gain ratio* merupakan perbandingan antara nilai *entropy* dan nilai *information gain.* Hal ini menjadikan algoritma C5.0 memiliki beberapa keunggulan yaitu memberikan hasil yang akurat, membutuhkan sedikit memori, dibutuhkan sedikit waktu untuk membangun model, dan dapat menanggani data katagorik, kontinu, dan data yang hilang. Metode C5.0 dapat diterapkan dalam berbagai sektor seperti sektor perpajakan.

Menurut Undang-Undang Republik Indonesia Nomor 28 tahun 2007 tentang perubahan ketiga atas Undang-Undang No 6 Tahun 1983 mengenai Ketentuan Umum dan Tata Cara Perpajakan, pajak adalah konstribusi wajib kepada Negara yang terutang oleh orang pribadi atau badan yang bersifat memaksa berdasarkan Undang-Undang dengan tidak mendapatkan imbalan secara langsung dan digunakan untuk keperluan negara bagi sebesar-besarnya kemakmuran rakyat. Menurut Hakim, *et. al* [4], pajak pertambahan nilai (PPN) dinyatakan sebagai pajak penyumbang tertinggi karena dipungut beberapa kali kepada perusahaan. Ini dikarenakan bertambahnya faktor-faktor produksi pada setiap jalur perusahaan dalam menyiapkan, menghasilkan, menyalurkan, memperdagangkan barang atau pemberian layanan jasa kepada konsumen. Berikut dipaparkan penelitan terdahulu terkait penerapan *data mining* pada sektor perpajakan.

Penelitian oleh Ripai, *et. al* [5], menggunakan algoritma regresi logistik biner dalam mengetahui kepatuhan masyarakat dalam membayar pajak daerah. Hasil penelitan menunjukkan tingkat akurasi model sebesar 93,97%. Penelitian oleh Shaumi, *et. al* [6], menggunakan algoritma *naive bayes* dalam mengklasifikasikan ketepatan waktu sesorang membayar pajak. Pada jumlah wajib pajak sebanyak 1.647 wajib pajak diperoleh akurasi sebesar 99,33%. Pada penelitian Ramaulidyah dan Goejantoro [7], peneliti menggunakan algoritma naive bayes dan k-nearest neigbour (KNN) dalam klasifikasi status pembayaran pajak nilai di KPP Pratama Samarinda ULU. Hasil pengukuran diperoleh metode naive bayes memberikan performa yang lebih baik dibandingkan KNN.

Berdasarkan paparan yang telah dijelaskan, penelitian ini akan menggali lebih jauh potensi penerapan metode C5.0 menggunakan data dari Ramaulidyah dan Goejantoro [7]. Penelitian ini diharapkan dapat menjadi solusi potensial untuk menanggani masalah prediksi kepatuhan nasabah dalam pembayaran pajak.

Metode penelitian

A. Data dan Variabel Penelitian

Data yang diaplikasikan pada penelitian ini adalah data sekunder yang diperoleh dalam penelitian Ramaulidyah dan Goejantoro [7]. Sampel penelitian sebanyak 205 data sampel yang berasal dari laporan wajib pajak badan KPP Pratama Samarinda ULU. Data yang diberikan instansi tersebut meliputi status pembayaran pajak, pendapatan, bentuk badan, dan status pelaporan. Keterangan tiap variabel dapat dilihat pada Tabel 1.

Tabel 1. Variable Penelitian

Variabel	Kategori
Status Pembayaran Pajak (Y)	1) Patuh 2) Tidak Patuh
Pendapatan (X_1)	1) Kurang dari 100 juta 2) 100 juta – 500 juta 3) 500 juta – 1 miliar 4) 1 miliar – 10 miliar 5) Lebih dari 10 miliar
Bentuk Badan Usaha (X_2)	1) BUMN/BUMD 2) CV 3) Koperasi 4) PT 5) Lainnya
Status Pelaporan (X_3)	1) Tepat Waktu

	2) Tidak Tepat Waktu

B. Decision Tree Tipe C5.0

Decision tree merupakan metode klasifikasi nonparametrik yang sangat sering digunakan dalam berbagai studi kasus. Model *decision tree* pada dasarnya adalah sebuah pohon yang melibatkan serangkaian keputusan node, di antaranya node akar, node internal, dan node daun. Alur model diawali node akar lalu turun ke node internal hingga berakhir di node daun, dengan solusi untuk pertanyaan terkait berada node daun yang telah diberi label. Perpecahan dibuat di setiap node dengan membuat keputusan biner yang memisahkan satu kelas atau beberapa kelas dari dataset global.

Misalkan C_i adalah kategori dalam himpunan S dengan $i = 1, 2, \cdots, m$. Probabilitas p_i merupakan perbandingan antara total amatan pada kelas C_i di S terhadap keseluruhan amatan di S. *Entropy* S dapat dihitung sebagai berikut.

$$Entropy\ (S) = -\sum_{i=1}^{m} p_i \log_2 p_i \quad (1)$$

Algoritma C5.0 adalah sejenis algoritma yang menghitung pemisahan terbaik berdasarkan *information gain*. *Information gain* merupakan ukuran berbasis probabilitas yang digunakan untuk menghitung tingkat pengurangan ketidakpastian. Umumnya pohon keputusan dibentuk dengan menghitung pemisahan dengan *information gain* terbesar hingga diperoleh solusi terbaik [8].

Misalkan untuk setiap variabel independen A, pertimbangkan himpunan $V(A)$ yang berisi kemungkinan nilai di A. Untuk setiap $v \in V(A)$, misalkan S_v adalah himpunan semua elemen S yang mempunyai nilai v untuk variabel A. Perolehan informasi dari A terhadap S dinyatakan sebagai berikut.

$$Information\ Gain(S, A) = Entropy\ (S) - \sum_{v \in V(A)} \frac{|S_v|}{|S|} \times Entropy\ (S_v) \quad (2)$$

Kemudian menghitung *gain ratio* menggunakan persamaan (1) dan (2).

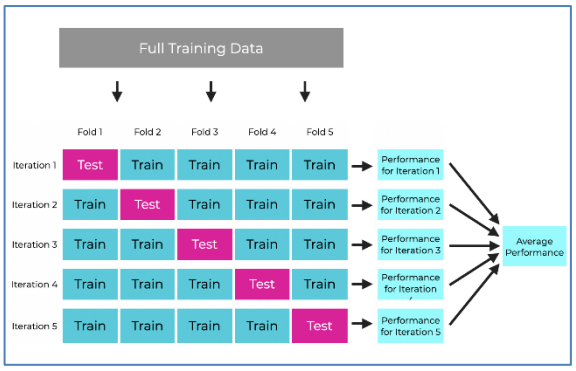
$$Gain\ Ratio = \frac{Information\ Gain(S,A)}{\sum_{v \in V(A)} Entropy\ (S_v)} \quad (3)$$

C. k-Fold Cross Validation

Metode *k-fold cross validation* merupakan teknik yang digunakan untuk menyelidiki keakuratan verifikasi model. Dataset dibagi menjadi n lipatan, di antaranya satu lipatan digunakan untuk validasi, dan $n - 1$ lipatan adalah data pelatihan. Ketika setiap lipatan dipilih sebagai set validasi, n akurasi model diperoleh dengan mengulangi langkah yang sama. Akhirnya, rata-rata akurasi yang diperoleh ini, yang disebut sebagai akurasi validasi silang, dianggap sebagai akurasi model. Dibandingkan dengan pendekatan tradisional, metode ini sangat membantu untuk mengatasi masalah *overfitting* dan meningkatkan kemampuan generalisasi model *decision tree*. Garre *et al.* [9] menerapkan pendekatan validasi silang 10 kali lipat dalam meminimalkan *overfitting* untuk beberapa algoritma pembelajaran mesin termasuk *decision tree*, *random forest*, *gradien boosting*, regresi lasso, regresi ridge, *elastic net*, dan *spline*. Menurut Nguyen *et al.* [10], validasi silang k-*fold* berhasil meminimalkan masalah *overfitting*.

Langkah-langkah dalam menggunakan teknik *k-fold cross validation* sebagai berikut.

1. Menentukan jumlah lipatan (*k*).
2. Membagi data berdasarkan k kelompok
3. Melakukan pengulangan sebanyak k kali. Pada tiap pengulangan, dipilih satu kelompok sebagai data *testing* dan k-1 kelompok sebagai data *training.*
4. Melakukan pelatihan model dengan data *training.*
5. Menguji performa model
6. Ulangi proses 3 sampai 5 untuk setiap kelompok.
7. Hitung rata-rata kinerja setiap model.



Gambar 1. *k-fold Cross Validation*

D. Uji Performa Model

Confusion matrix merupakan tabel yang memberikan hasil diuji dengan data *testing*. Metode ini paling intuitif untuk melihat kinerja klasifikasi [11]. Akurasi merupakan persentase dari total keseluruhan amatan yang terprediksi atau terklasifikasi secara tepat [12].

Tabel 2. *Confusion Matrix*

			Kelas Prediksi	
			Ya	Tidak
	Kelas Observasi	Ya	*a*	*b*
		Tidak	*c*	*d*

$$\text{Akurasi} = \frac{a+d}{a+b+c+d} \times 100\% \qquad (4)$$

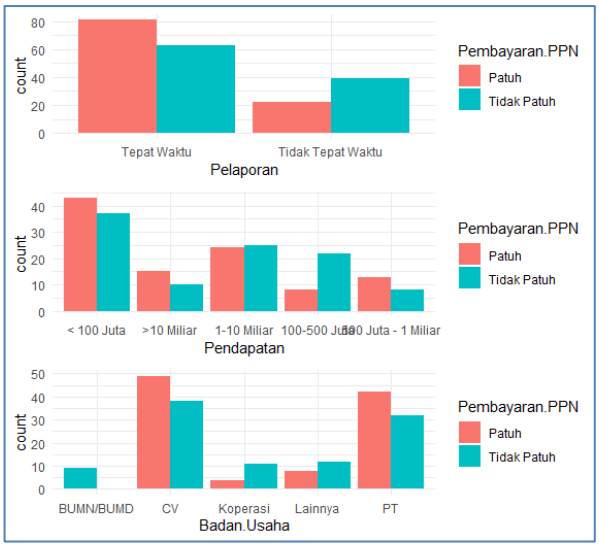
Interpretasi kategori nilai akurasi sebagai berikut [13].

- 50% - 59,99% = sangat lemah
- 60% - 69,99% = lemah
- 70% - 79,99% = sedang
- 80% - 89,99% = kuat
- 90% - 100% = sangat kuat

Hasil dan Pembahasan

A. Eksplorasi Data

Pada Gambar 2, grafik yang memperlihatkan variabel Status Pelaporan, terdapat 81 nasabah yang melaporkan surat pemberitahuan tahunan (SPT) masa PPN dalam jangka waktu yang ditentukan patuh membayar PPN. Sedangkan 22 nasabah yang melaporkan SPT masa PPN dalam jangka waktu yang ditentukan tidak patuh membayar PPN. Pada grafik yang menampilkan variabel Pendapatan, terdapat 43 nasabah yang memiliki pendapatan pertahun kurang dari 100 juta patuh membayar PPN dan terdapat 8 nasabah yang memiliki pendapatan pertahun diantara 100 juta hingga 500 juta patuh membayar pajak. Sebaliknya terdapat 8 nasabah yang memiliki pendapatan pertahun diantara 500 juta hingga 1 miliar tidak patuh membayar pajak. Pada grafik yang menampilkan variabel Bentuk Badan Usaha, terdapat 49 nasabah berbentuk CV patuh membayar pajak tetapi 38 lainnya tidak patuh membayar pajak. Sementara itu terdapat 4 nasabah berbentuk koperasi patuh dalam membayar pajak.



Gambar 2. Asosiasi Variabel Independen Terhadap Variabel Dependen

B. Evaluasi Performa Model C5.0

Sebelum melakukan pemodelan, langkah pertama adalah menetapkan sebanyak 10 lipatan dalam membangun model. Kemudian proses pembentukan model C5.0 hingga uji model dengan data *testing* dibantu oleh *software* R 4.3.2. menggunakan *packages C50*. Hasil uji model prediksi pada lipatan ke-10 disajikan dalam tabel *confusion matrix*.

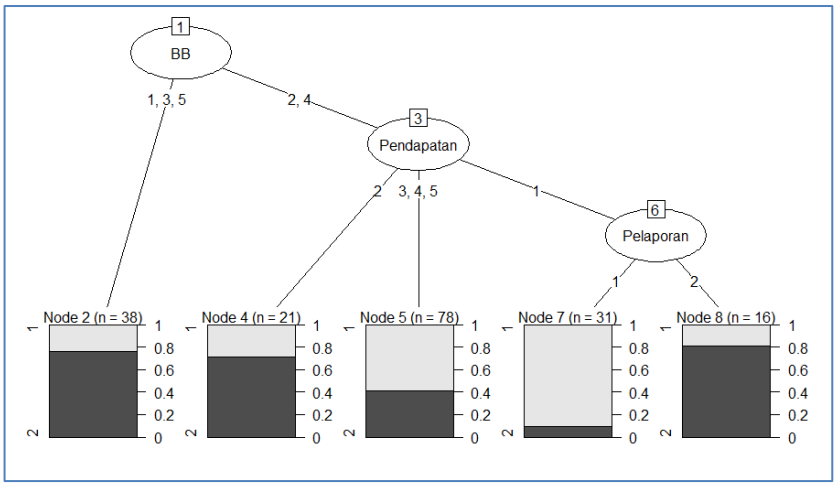
Tabel 3. *Confusion Matrix* Model *C5.0*

		Kelas Prediksi	
		Patuh	Tidak Patuh
Kelas Observasi	Patuh	6	3
	Tidak Patuh	5	7

Berdasarkan Tabel 3 dijelaskan bahwa kelas patuh yang tepat terprediksi sebagai kelas patuh sebanyak 6 amatan dan kelas tidak patuh terprediksi tepat tidak patuh sebanyak 7 amatan. Sementara itu, ada 8 amatan yang terdeteksi misklasifikasi. Kemudian menghitung performa model dengan menghitung akurasi menggunakan persamaan (10).

$$Akurasi = \frac{6+7}{6+7+5+3} \times 100\%$$
$$= \frac{13}{21} \times 100\%$$
$$= 61{,}9\%$$

Diperoleh akurasi yang dihasilkan dari uji model C5.0 terhadap data *testing* sebesar 61,9%. Artinya model berhasil memprediksi dengan tepat status nasabah yang membayar PPN secara keseluruhan sebesar 61,9% sisanya 38,1% terjadi misklasifikasi. Model C5.0 pada lipatan ke-10 ditunjukkan pada Gambar 3.



Gambar 3. Model C5.0

Pada Gambar 5 diperlihatkan simpul akar yang memisahkan data berdasarkan bentuk badan usaha (BB). Data pada kategori BUMN/BUMD, koperasi, dan lainnya masuk pada node 2. Pada node 2 sebagian besar data di kelas 2 yaitu kelas tidak patuh.
Data dengan kategori PT dan CV dipisahkan lagi berdasarkan pendapatan. Data dengan pendapatan 100-500 juta masuk pada node 4. Pada node 4 sebagaian besar data berada di kelas tidak patuh. Data dengan pendapatan > 500 juta masuk pada node 5. Pada node 5, sebanyak 60% data berada pada kelas patuh.
Data dengan kategori pendapatan < 100 juta dipisahkan lagi berdasarkan status pelaporan. Data dengan status pelaporan tepat waktu masuk pada node 7. Pada node 7 sebagian besar data berada pada kelas patuh. Sementara itu, data dengan status pelaporan tidak tepat waktu berada pada node 8. Pada node 8, sebanyak 80% data masuk pada kelas tidak patuh.

Tabel 4. 10-*fold Cross Validation Model* C5.0

Lipatan ke-*i*	**Akurasi**
1	57,1%
2	80%
3	61,9%
4	85%
5	47,6%
6	65%
7	60%
8	76,2%
9	70%
10	61,9%
Rata-rata	**66,5%**

Pada Tabel 4 memperlihatkan model untuk setiap lipatan menghasilkan akurasi yang berbeda-beda. Akurasi tertinggi pada lipatan ke-4 sebesar 85% sedangkan akurasi terendah pada lipatan ke-5 sebesar 47,6%. Rata-rata untuk setiap lipatan diperoleh akurasi sebesar 66,5%. Hasil ini menunjukkan bahwa model masih lemah dalam memprediksi dengan tepat status nasabah yang membayar PPN.

Kesimpulan

Berdasarkan hasil dan pembahasan maka dapat disimpulkan bahwa metode klasifikasi C5.0 dengan 10-*fold cross validation* menghasilkan rata-rata akurasi sebesar 66,5%. Berdasarkan akurasi yang dihasilkan maka model dapat dikategorikan masih lemah dalam memprediksi status pembayaran pajak PPN. Namun perlu diketahui bahwa hasil penelitian ini masih membutuhkan variabel independen lebih banyak lagi untuk meningkatkan akurasi model. Selain itu diharapkan penelitian selanjutnya dapat menggunakan metodel klasifikasi lain seperti *random forest*, regresi logistik, *support vector machine*, dll.

Referensi

[1] M. A. Aprihartha, J. Prasetya, and S. I. Fallo, "Implementasi CART-Real Adaboost dalam Memprediksi Minat Pelanggan Membeli Sepatu," *Jurnal EurekaMatika*, vol. 12, no. 1, pp. 35-46, May 2024.

[2] M. A. Aprihartha, F. Astutik and N. Sulistianingsih, "Comparison of Naïve Bayes, CART, dan CART Adaboost Methods in Predicting Tire Product Sales," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 3, pp. 596-605, June 2024.

[3] S. Rajeswari and K. Suthendran, "C5.0: Advanced Decision Tree (ADT) classification model for agricultural data analysis on cloud," *Computers and Electronics in Agriculture*, vol. 156, pp. 530–539, Jan. 2019.

[4] L. Hakim, S. Sabil, A. S. Lestiningsih, and D. P. Widodo, "Pengaruh Pemungutan Pajak Pertambahan Nilai Terhadap Laporan Keuangan," Jurnal SIKAP (Sistem Informasi, Keuangan, Auditing Dan Perpajakan), vol. 4, no. 1, pp. 1-11, 2019.

[5] M. Ripai, U. Hayati, W. Widyawati, and H. Susana, "Klasifikasi Surat Pemberitahuan Pajak Daerah Menggunakan Metode Regresi Logistik Biner Untuk Mengetahui Patuh Dan Tidak Patuh Dalam Pembayaran Pajak Daerah," KOPERTIP: Scientific Journal of Informatics Management and Computer, vol. 6, no. 1, pp. 27-33, 2022.

[6] A. R. Shaumi, M. F. Ali, and M. T. A. M. Simbolon, "Penerapan Data Mining menggunakan Metode Teknik Classification untuk Melihat Potensi Kepatuhan Wajib Pajak Bumi dan Bangunan," JUKI: Jurnal Komputer dan Informatika, vol. 4, no. 2, pp. 171-182, 2022.

[7] F. N. Rahmaulidyah, M. N. Hayati, and R. Goejantoro, "Perbandingan Metode Klasifikasi Naive Bayes Dan K-Nearest Neighbor Pada Data Status Pembayaran Pajak Pertambahan Nilai di Kantor Pelayanan Pajak Pratama Samarinda Ulu," Eksponensial, vol. 12, no. 2, pp. 161-164, 2021.

[8] Z. Guo, Y. Shi, F. Huang, X. Fan, and J. Huang, "Landslide susceptibility zonation method based on C5.0 decision tree and K-means cluster algorithms to improve the efficiency of risk management," Geoscience Frontiers, vol. 12, no. 6, p. 101249, Nov. 2021.

[9] A. Garre, M. C. Ruiz, and E. Hontoria, "Application of Machine Learning to support production planning of a food industry in the context of waste generation under uncertainty," Operations Research Perspectives, vol. 7, p. 100147, 2020.

[10] X. C. Nguyen et al., "Development of machine learning - based models to forecast solid waste generation in residential areas: A case study from Vietnam," Resources, Conservation and Recycling, vol. 167, p. 105381, Apr. 2021.

[11] J. Prasetya, S. I. Fallo, and M. A. Aprihartha, "Stacking

Machine Learning Model for Predict Hotel Booking Cancellations," *Jurnal Matematika, Statistika dan Komputasi*, vol. 20, no. 3, pp. 525-537, 2024.

[12] F. Gorunescu, Data Mining: Concepts, models and techniques, vol. 12. Springer Science & Business Media, 2011.

[13] T. T. Maskoen and A. Masthura, "Nilai Area Under Curve dan Akurasi Neutrophil Gelatinase Associated Lipocalin untuk Diagnosis Acute Kidney Injury pada Pasien Politrauma di Instalasi Gawat Darurat RSUP dr. Hasan Sadikin Bandung," Maj Anest dan Crit Care, vol. 35, no. 3, pp. 158–164, 2017.

Moch. Anjas Aprihartha
Dosen universitas Dian Nuswantoro,
yang berfokus pada matematika.

M. Husniyadi
Mahasiswa dari Universitas
Dian Nuswantoro, Fakultas Ilmu Komputer.

Taufik Nur Alam
Mahasiswa dari Universitas
Dian Nuswantoro, Fakultas Ilmu Komputer.