KLASIFIKASI PEMILIHAN PRODUK PERBANKAN DENGAN ALGORITMA *DECISION TREE* ID3 DAN C4.5

Putri Anggita Kristianti, Singgih Jatmiko
Perangkat Lunak Sistem Informasi, Universitas Gunadarma
Jl. Kenari I, RT.4/RW.5, Kenari, Kec. Senen, Jakarta Pusat
Anggita314@gmail.com

ABSTRAK

Seiring meningkatnya kebutuhan nasabah, bank berusaha menarik nasabah dengan menawarkan produk yang paling sesuai. Produk merupakan faktor dalam mempengaruhi keputusan nasabah dalam memilih layanan yang tepat. Penelitian ini bertujuan untuk melakukan klasifikasi produk bank dengan membandingkan pengukuran akurasi, presisi dan *recall* algoritma *decision tree* ID3 dan C4.5 menggunakan *tools* RapidMiner. Dalam penelitian ini *dataset* tersedia dengan jumlah 16.848 terdiri dari 13 atribut, 12 atribut prediksi antara lain *age, gender, religion,* ntb, *household, education, salary, realiation, potential, response_status,* status_nasabah, jenis_produk dan 1 atribut tujuan yaitu nama_produk. Atribut jenis_produk merupakan *node* pertama dan yang paling berpengaruh. Penelitian ini menggunakan metode CRISP-DM yang terdiri dari tahap *business understanding, data understanding, data preparation, modelling* dan *evaluation*. Hasil penelitian menunjukan algoritma ID3 memiliki rata-rata akurasi 79,09%, presisi 77,56%, dan *recall* 79,07%. Sementara itu, untuk algoritma C4.5, memiliki rata-rata akurasi 80,03%, presisi 78,26%, dan *recall* 80,03%. Algoritma C4.5 unggul dalam nilai akurasi, presisi, dan *recall* daripada algoritma ID3, menurut hasil implementasi secara keseluruhan. Dengan demikian, dapat dikatakan bahwa algoritma C4.5 adalah algoritma yang paling efektif untuk klasifikasi pemilihan produk perbankan.

Kata kunci : *Klasifikasi, Bank, Akurasi, ID3, C4.5*

1. PENDAHULUAN

Setiap perusahaan menggunakan aktivitas pemasaran sebagai standar dalam proses penyampaian produk kepada konsumen dan mencapai tujuan target penjualan produk sebaik mungkin. Terdapat berbagai faktor yang berkontribusi pada peningkatan penjualan; ini termasuk faktor internal yang berkaitan dengan pemasaran perusahaan, seperti promosi, kualitas layanan dan faktor eksternal yang berkaitan dengan keputusan konsumen untuk menentukan produk tertentu. Sebelum melakukan pembelian, konsumen menggunakan proses pengambilan keputusan untuk menentukan masalah mereka, meneliti produk dan mengevaluasi seberapa efektif setiap pilihan dapat mengatasi masalah mereka. Pemasaran oleh sales adalah salah satu aktivitas penjualan dalam pemasaran yang memiliki bagian yang sangat dibutuhkan, karena merupakan salah satu elemen kunci dalam menjalankan bisnis yang sukses. Dengan melakukan tindakan-tindakan seperti mencari pelanggan, mencari prospek, menjelaskan produk atau layanan, bernegosiasi, dan menutup penjualan, perusahaan dapat mencapai tujuan penjualan mereka. Pemasaran oleh sales di industri perbankan memiliki tanggung jawab untuk mencari nasabah yang ingin menghimpun dananya sebanyak mungkin. Karena tujuan perbankan adalah untuk mencapai tujuan dan target yang telah ditetapkan. Pemasaran sales menjadi salah satu elemen yang memiliki potensi tinggi untuk mencapai target penjualan perbankan [1].

Seiring perkembangan teknologi dan meningkatnya kebutuhan konsumen, masing-masing bank berusaha menarik nasabah dengan memberikan pelayanan yang lebih baik untuk setiap nasabahnya. Saat memilih bank, nasabah tidak hanya mempertimbangkan suku bunga tetapi juga hal-hal seperti layanan dan produk yang ditawarkan. Saat ini bank dihadapkan pada tantangan untuk menawarkan berbagai produk yang tidak hanya memenuhi kebutuhan dasar nasabah, tetapi juga menawarkan nilai tambah yang lebih tinggi. Selain itu, persaingan yang semakin ketat di industri perbankan mendorong bank untuk lebih fokus dalam pengembangan produk yang tepat sasaran. Dengan adanya klasifikasi produk, bank dapat lebih mudah mengidentifikasi celah di pasar, menyesuaikan layanan dengan kebutuhan pasar yang terus berkembang dan mengembangkan produk baru. Penerapan *data mining* di industri perbankan menjadi salah satu kunci dalam melakukan identifikasi dan keefektifan dalam mendapatkan, mengelola, dan menganalisa data nasabah dan produk pada kegiatan perbankan. Penelitian ini menjadi relevan dan penting untuk membantu bank dalam mengoptimalkan strategi mereka, meningkatkan rekomendasi produk dan mengoptimalkan penawaran produk.

Adapun penelitian terdahulu telah menunjukan efektivitas algoritma ID3 dan C4.5 dalam berbagai bidang, penelitian dalam memprediksi kelulusan mahasiswa menghasilkan akurasi tertinggi algoritma C4.5 yaitu 81,88% [2].

Sementara algoritma ID3 dalam penelitian faktor kepuasan konsumen memiliki akurasi lebih tinggi yaitu 98,50% [3]. Penelitian lain membahas prediksi mahasiswa yang putus kuliah algoritma C4.5 lebih unggul dari ID3 dalam hal akurasi dan presisi [4]. Begitu juga dalam seleksi asisten laboratorium, C4.5

mencatat akurasi 96,67% menggunakan model *k-fold* [5].

Sementara itu, pada kasus klasifikasi tersangka narkoba, C4.5 tetap lebih baik dari ID3 meskipun kalah dari algoritma CHAID [6].

Saat ini belum ada kajian secara spesifik menerapkan algoritma ini dalam pemilihan produk perbankan. Oleh karena itu, penelitian ini mengisi celah tersebut dengan menerapkan algoritma *decision tree* ID3 dan C4.5 untuk membantu dalam proses pemilihan produk perbankan.

Penelitian ini dilakukan perbandingan antara dua algoritma yaitu *decision tree* ID3 dan C4.5, untuk menganalisis klasifikasi produk perbankan dengan membandingkan nilai akurasi, presisi dan *recall* yang didapatkan dari kedua algoritma. Perbedaan utama C4.5 dari ID3 yaitu pemilihan atribut yang dilakukan dengan menggunakan *gain ratio* dan hasil pohon keputusan C4.5 dipangkas setelah dibentuk. Pemilihan algoritma *decision tree* ID3 dan C4.5 dalam penelitian ini didasarkan pada beberapa pertimbangan. Algoritma ID3 dipilih karena dapat memberikan pohon keputusan yang sederhana dan mudah dipahami, sementara C4.5 merupakan pengembangan dari ID3 yang memperbaiki kelemahan terkait dengan kemampuan untuk menghitung atribut numerik dan data kosong. Pada tahap akhir penelitian ini dilakukan evaluasi kinerja dari kedua algoritma dalam memilih produk perbankan yang sesuai.

2. TINJAUAN PUSTAKA

Penelitian terdahulu dilakukan dalam bidang pendidikan untuk memprediksi ketepatan waktu lulus mahasiswa, menunjukan bahwa prediksi kelulusan mahasiswa dengan hasil nilai akurasi tertinggi dihasilkan oleh algoritma C4.5 yaitu 81,88% pada data pengujian 30%. Hasil implementasi untuk komposisi data sebesar 90%, 80%, 70%, 30%, 20%, dan 10% menunjukkan bahwa nilai akurasi algoritma C4.5 rata-rata lebih tinggi daripada algoritma ID3 [2].

Penelitian kedua berfokus pada faktor kepuasan konsumen, penelitian yang berfokus pada faktor-faktor yang memengaruhi kepuasan pelanggan, elemen-elemen yang memiliki pengaruh paling besar terhadap kepuasan pelanggan dikelompokkan menggunakan pendekatan klasifikasi menggunakan algoritma C4.5 dan ID3. Selain itu, pendekatan *cross validation* digunakan untuk membandingkan algoritma C4.5 dan ID3. Tingkat akurasi, presisi, dan *recall* dari pengujian algoritma ID3 lebih tinggi, masing-masing sebesar 98,50%, 98,77%, dan 99,38% [3].

Penelitian ketiga membahas prediksi mahasiswa yang putus kuliah menggunakan metode pohon keputusan C4.5 dan ID3 yang berguna untuk membantu kampus dalam mengantisipasi mahasiswa putus kuliah. Temuan studi menghasilkan 18 aturan untuk algoritma ID3 dan 8 aturan untuk algoritma C4.5. Akurasi, presisi dan *recall* rata-rata algoritma ID3 selama pengujian masing-masing adalah 95,17%, 94,7%, dan 96,18%. Sementara itu, akurasi, presisi dan *recall* rata-rata algoritma C4.5 selama pengujian masing-masing adalah 96,45%, 96,90%, dan 95,38%. Penelitian ini membuktikan bahwa penggunaan algoritma C4.5 lebih baik dalam memprediksi mahasiswa yang putus kuliah di STEKOM [4].

Penelitian keempat meneliti tentang penerimaan kandidat asisten laboratorium fisika dasar. Setelah membandingkan akurasi, presisi dan *recall* dari kedua algoritma, maka dapat disimpulkan bahwa algoritma C4.5 lebih akurat dibandingkan algoritma ID3. Dengan nilai variasi 2, 4, 6, 8, dan 10 *fold*. Maka digunakan model *K-Fold cross validation* untuk menghasilkan nilai pengujian. Pada *k-fold* = 2, nilai akurasi, algoritma C4.5 adalah 96,67% [5].

Berdasarkan penelitian lain dalam menentukan tersangka kasus narkoba di Jawa Barat berdasarkan hasil uji performa menggunakan 10-*fold cross validation*, algoritma CHAID memiliki nilai akurasi tertinggi yaitu 73.89% sedangkan C4.5 akurasinya 72.44%, lalu nilai akurasi paling rendah didapatkan oleh ID3 sebesar 70.14%. Hasil klasifikasi dengan algoritma CHAID sebagai pemodelan yang terpilih, dan algoritma C4.5 lebih baik dari pada ID3 [6].

3. METODE PENELITIAN

Metode penelitian ini didasarkan pada pendekatan klasifikasi. Dalam *data mining*, klasifikasi adalah teknik dalam yang mengelompokkan data menurut kriteria tertentu ke dalam kelas atau kategori. Tujuan klasifikasi yaitu untuk membangun model atau aturan yang dapat digunakan untuk memprediksi kelas atau kategori data yang tidak diketahui. Sedangkan *data mining* adalah teknik pengekstrasian informasi dari data yang jumlahnya sangat besar sehingga dari data tersebut dapat diperoleh suatu informasi yang bermanfaat bagi pihak yang membutuhkan [7].

Alur proses utama yang digunakan dalam penelitian ini untuk mendukung proses pembuatan model *data mining* adalah CRISP-DM (*Cross Industry Standard Process for Data Mining*), yang berfungsi sebagai landasan metodologi. Ada lima langkah dalam metode CRISP-DM yaitu:

3.1. *Business Understanding*

Memahami tujuan dan kebutuhan dari sudut pandang bisnis adalah tahap pertama dalam *business understanding* [8].

Pada tahap ini terdapat empat fase yaitu *determine business objective, asses situation, determine data mining goals* dan *produce project plan* [9]:

a. *Determine Business Objective*, Dengan penerapan algoritma *decision tree* ID3 dan C4.5, bank dapat mengoptimalkan proses klasifikasi pemilihan produk bank secara otomatis, dapat meningkatkan dan mengukur akurasi klasifikasi produk bank, mempercepat proses pengambilan keputusan terkait produk perbankan dan meningkatkan

tingkat keberhasilan penjualan dan kepuasan nasabah.

b. *Asses Situation*, saat ini *sales* bank masih kesulitan dalam mengidentifikasi kebutuhan nasabah secara akurat, diketahui dari tahun 2020 hingga 2024 sebanyak 77 nasabah menolak dan 206 nasabah memiliki tingkat minat rendah dalam penawaran produk dari jumlah total data 16.848 nasabah. Proses ini masih banyak bergantung pada interaksi *manual* sales, sehingga memiliki beberapa kendala yaitu potensi ketidakakuratan dalam pemilihan produk, proses pengambilan keputusan lambat sehingga nasabah mungkin ditawarkan produk yang kurang relevan dan tidak ada sistem otomatis untuk meningkatkan peluang penjualan dalam menawarkan produk yang paling sesuai. Saat ini proses pemilihan produk dan penawaran produk kepada nasabah melibatkan sales, nasabah dan sistem bank.

c. *Determine Data Mining Goals*, tujuan utama dari *data mining* dalam penelitian ini adalah untuk membangun model klasifikasi yang dapat merekomendasikan produk, melakukan analisis klasifikasi produk bank dengan membandingkan akurasi, presisi dan *recall* algoritma *decision tree* ID3 dan C4.5, mengevaluasi kinerja dari model.

d. *Produce Project Plan*
Untuk mencapai tujuan tersebut, studi ini dilakukan melalui tahapan pengumpulan data historis nasabah termasuk transaksi demograsi dan produk yang diambil, tahap kedua *preprocessing data* mencakup perbersihan *dataset* sehingga fitur yang dilakukan uji coba hanya yang relevan untuk penelitian, tahap ketiga pembangunan model dengan algoritma ID3 dan C4.5, tahap terakhir dilakukan evaluasi model dengan *confussion matrix* untuk mengetahui tingkat akurasi, presisi dan *recall* yang terdapat dalam masing-masing pengujian algoritma serta membandingkan nilai terbaik dalam pengujian *dataset*.

3.2. *Data Understanding*

Dataset yang digunakan pada penelitian ini bersumber dari *database* dengan jumlah *dataset* yaitu 16.848 dengan rentang waktu 22 April 2020 hingga 17 Oktober 2024. Atribut *dataset* yang memiliki 12 atribut prediksi antara lain *age, gender, religion*, ntb, *household, education, salary, realiation, potential, response_status*, status_nasabah, jenis_produk dan 1 atribut tujuan yaitu nama_produk.

Tabel 1. Atribut *dataset*

Atribut	Nilai
Age	< 84
Gender	L dan P
Religion	Islam, Kristen, Hindu, Budha, Konghucu
NTB	Yes dan No
Household	Milik Pribadi, Sewa, Milik Keluarga, Rumah Dinas
Education	SD, SMP, SMA, D1, D2, D3, D4, S1, S2, S3
Salary	< Rp 37.000.000.000
Realiation	< Rp. 7.000.000.000
Potential	< Rp. 600.000.000.000
Response_status	Nasabah Tertarik, Nasabah Bisa di Call/Visit Kembali, Tidak Diangkat, Nasabah Menolak, Tidak dirumah, Nomor Tidak Aktif, Terhubung Tapi Salah Sambung, Salah Alamat dan lain-lain
Status_nasabah	*Hot, warm, cold*
Jenis_produk	*Funding, Lending*
Nama_produk	Tabungan, KPR / KPA, Kredit Ringan, Kredit Agunan Rumah dan Deposito Perorangan

Berikut penjelasan atribut pada tabel 1:

a. *Age*, umur nasabah.
b. *Gender*, jenis kelamin nasabah.
c. *Religion*, agama.
d. Ntb, *new to bank* merupakan nasabah baru yang belum memiliki akun *bank.*
e. *Household*, kepemilikan tempat tinggal.
f. *Education*, tingkat pendidikan nasabah.
g. *Salary*, pendapatan terakhir.
h. *Realiation*, jumlah uang saat ini dalam bentuk produk bank.
i. *Potential*, kemampuan nasabah untuk meningkatkan jumlah uang simpanan.
j. *Response_status*, tingkat respon terkait interaksi sales dengan nasabah.
k. *Status_nasabah*, tingkat ketertarikan atau kesiapan.
l. *Jenis_produk*, Kategori produk. Jika *funding* adalah aktivitas pengumpulan dana, jika *lending* merupakan pinjaman dana.
m. *Nama_produk*, Produk yang ditawarkan.

3.3. *Data Preparation*

Pada tahap ini mencakup proses pembersihan pada *dataset* sehingga fitur yang dilakukan uji coba hanya yang relevan untuk penelitian, sehingga dapat dijadikan masukan dalam tahap pemodelan (*modeling*).

Gambar 1. *Data preparation*

Pada gambar 1 merupakan tahap *data preparation* yang terdiri dari:

a. *Data cleaning*, memeriksa *dataset* apakah terdapat data yang tidak relevan, menghapus atribut yang tidak digunakan, menghapus data duplikat serta *missing value*.
b. *Data selection* dan *data transformation*, pemilihan atribut prediksi dan atribut target dan dilakukan proses *data transformation* untuk mengubah data ke format yang sesuai seperti melakukan normalisasi data kategorikal.

c. Diskretisasi data merupakan proses pengelompokkan atau pengkategorian atribut bertipe numerik menjadi kategori.

Tabel 2. Diskretisasi data

Atribut	*Range*	Nilai
Age	1	< 31 tahun
	2	31 – 40 tahun
	3	> 40 tahun
Salary	1	< Rp. 4.901.550
	2	Rp. 4.901.550 – Rp. 9.001.000
	3	> Rp. 9.001.000
Realiation	1	Rp. 0
	2	Rp. 0 - Rp. 55.227.561
Realiation	3	> Rp. 55.227.561
Potential	1	< Rp. 100.000.000
	2	Rp. 100.000.000 – Rp. 390.140.000
	3	> Rp. 390.140.000

Hasil dari diskretisasi terlihat pada tabel 2. Diskretisasi data dilakukan pada atribut *age, salary*, *realiation* dan *potential*.

d. *Split data*, hasil dari *data preprocessing* didapatkan data yang sudah diolah sebanyak 16.332. *Split data* menggunakan *hold-out cross validation*. Lima skenario berbeda akan digunakan untuk membagi data dalam *hold-out cross validation*.

3.4. *Modelling*

Dalam proses *modelling* digunakan bantuan *software* RapidMiner. *Software* ini memiliki banyak operator yang sudah dibangun, dan mencakup semua tahap proses pengolahan data, seperti pembersihan data, pemilihan fitur, dan pemodelan [10].

Tahap pemodelan menggunakan algoritma *decision tree* ID3 dan C4.5. *decision tree* adalah metode klasifikasi yang digunakan untuk membantu menentukan pengambilan keputusan dalam pemecahan masalah. Dilakukan perhitungan untuk menentukan akar dari pohon keputusan. Langkah-langkah dalam menentukan akar pohon keputusan menggunakan algoritma *decision tree* adalah sebagai berikut:

a. Menghitung *entropy* untuk seluruh data. *Entropy* digunakan untuk mengukur ketidakpastian dari kumpulan data.

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

Pada persamaan 1 n adalah jumlah kategori (jenis_produk), p_i adalah probabilitas dari setiap kategori i. Dan log_2 adalah logaritma basis.

b. Menghitung *information gain* untuk setiap atribut. *information gain* menentukan berapa banyak ketidakpastian yang berkurang setelah membagi data berdasarkan atribut tertentu.

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

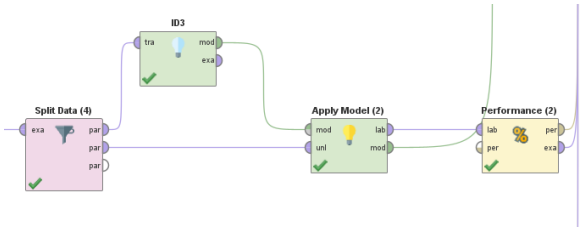
c. Menghitung *split information*. *split information* adalah pengukuran tambahan yang mempertimbangkan jumlah partisi yang dihasilkan oleh atribut.

$$SplitInfo(A) = -\sum_{v \in Values(A)} \frac{|S_v|}{|S|} log_2 \frac{|S_v|}{|S|} \quad (3)$$

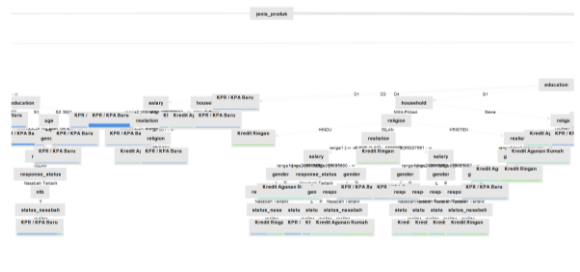
d. Menghitung *gain ratio*. *Gain ratio* adalah rasio antara *information gain* dan *split information*.

$$Gain\ Ratio(A) = \frac{Gain(S,A)}{SplitInfo(A)} \quad (4)$$

Algoritma ID3 dikenal juga sebagai *Iterative Dichotomiser 3* adalah teknik untuk membuat pohon keputusan. Dari sekumpulan fakta, algoritma ID3 dapat memberikan pohon keputusan yang mudah dipahami. Dengan memilih atribut yang akan diperiksa yang membentuk dasar pohon keputusan (akar), proses konstruksi diselesaikan dari atas ke bawah. Pendekatan ID3 digunakan untuk membuat keputusan model. Untuk mengidentifikasi simpul akar, langkah pertama adalah menghitung perolehan *Information Gain* setiap atribut.



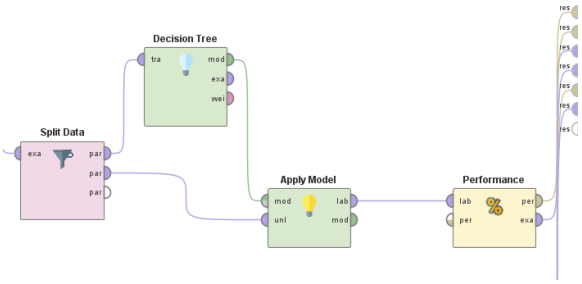
Gambar 2. *Modelling* ID3 RapidMiner



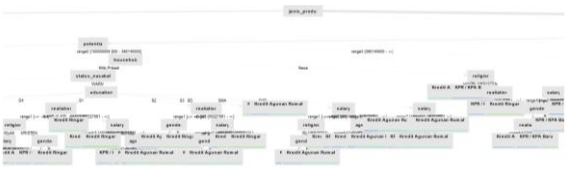
Gambar 3. Model *decision tree* ID3

Pada gambar 3 terlihat akar atau *node* awal dari pembentukan model algoritma ID3 adalah atribut jenis_produk dengan nilai 0,840 sebagai nilai tertinggi *information gain*. Nilai tersebut didapat dari hasil perhitungan manual menggunakan *data testing* 10% dari 16.332 data yaitu sebanyak 1.634 data dengan bantuan Microsoft Excel.

Pengembangan dari algoritma ID3 adalah algoritma C4.5. Karena hasilnya lebih akurat dan memiliki kemampuan untuk menghitung atribut numerik dan data kosong. Prosedur pembentukan pohon keputusan C4.5 mirip dengan proses algoritma ID3. Tentukan nilai *entropy* terlebih dahulu, kemudian nilai *gain* untuk menentukan akar. Untuk menghitung *gain ratio* perlu diketahui adanya suatu istilah yang disebut *Split Information*. Setelah nilai *Gain* dan *Split Information* diketahui, selanjutnya menghitung *Gain Ratio.*



Gambar 4. *Modelling* C4.5 RapidMiner



Gambar 5. Model *decision tree* C4.5

Gambar 5 menunjukkan akar pembentukan model algoritma C4.5 adalah atribut jenis_produk dengan nilai 1 sebagai nilai tertinggi *gain ratio*. Nilai tersebut didapat dari hasil perhitungan manual menggunakan *data testing* 10% dari 16.332 yaitu sebanyak 1.634 data dengan bantuan Microsoft Excel.

3.5. *Evaluation*

Suatu prosedur diperlukan untuk mengevaluasi bagaimana model bekerja setelah dilatih. Kemudian algoritma *decision tree* ID3 dan C4.5 dibandingkan untuk menentukan mana yang memiliki akurasi, presisi dan *recall* terbaik dalam pengujian *dataset.* Hasilnya akan dievaluasi secara sederhana dengan *confussion matrix*. *Confusion matri*x berguna untuk mendapatkan nilai akurasi, presisi, dan *recall*. Nilai *Confusion matrix* umumnya ditunjukkan dalam satuan persen [11].

Confussion matrix berisi metrik sebagai berikut:

a. TP adalah *True Positive*, yaitu jumlah data positif yang terklasifikasi dengan benar oleh sistem.
b. TN adalah *True Negative*, yaitu jumlah data negatif yang terklasifikasi dengan benar oleh sistem.
c. FN adalah *False Negative*, yaitu jumlah data negatif namun terklasifikasi salah oleh sistem.
d. FP adalah *False Positive*, yaitu jumlah data positif namun terklasifikasi salah oleh sistem.

Persamaan yang digunakan untuk menghitung metrik evaluasi adalah:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \quad (6)$$

$$Precision = \frac{TP}{TP+FP}\ x\ 100\% \quad (7)$$

$$Recall = \frac{TP}{TP+FN}\ x\ 100\% \quad (8)$$

Jumlah label pada data ada lima kelas produk perbankan yaitu Tabungan, KPR / KPA, Kredit Ringan, Kredit Agunan Rumah dan Deposito Perorangan, sehingga *confussion* matrix yang digunakan pada penelitian ini adalah *confussion matrix multi-class. Multi-class* data merupakan kondisi dimana *dataset* memiliki lebih dari dua label atau kelas (*multi-class*). Persamaan yang digunakan untuk *confussion matrix multi-class*:

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i+TN_i}{TP_i+TN_i+FP_i+FN_i}\ x\ 100\% \quad (9)$$

$$Weighted\ Precision = \frac{\sum_{i=1}^{N} Precision_i\ x\ Number_i}{\sum_{i=1}^{N} Number_i}\ x\ 100\%\ [12] \quad (10)$$

$$Weighted\ Recall = \frac{\sum_{i=1}^{N} Recall_i\ x\ Number_i}{\sum_{i=1}^{N} Number_i}\ x\ 100\%\ [12] \quad (11)$$

4. HASIL DAN PEMBAHASAN

Tahap ini membahas hasil model algoritma ID3 dan C4.5 hingga mendapatkan nilai akurasi, presisi dan nilai *recall*. Pada tahap terakhir algoritma akan dibandingkan dan dievaluasi untuk menentukan algoritma terbaik.

4.1. *Hold-out Cross Validation*

Salah satu Teknik yang sering digunakan untuk evaluasi model adalah *cross validation*. Jenis yang paling sederhana dari *cross validation* adalah *hold-out validation*. Adapun metode ini membagi *dataset* secara acak ke dalam dua kategori *dataset* yaitu *dataset training* yang digunakan untuk membuat model sementara data *testing* digunakan untuk menguji kinerja model yang dibentuk oleh data *training* [13].

Pada *hold-out cross validation* dilakukan *split data* kedalam 5 skenario dalam pembagian data yaitu skenario 1 yaitu proporsi data *training*:*testing* 90:10, skenario 2 yaitu proporsi data 80:20, skenario 3 proporsi data 70:30, skenario 4 proporsi data 60:40, skenario 5 proporsi data 50:50. Hasil *hold-out cross validation* akan dihitung dan dibandingkan nilai akurasi, presisi dan *recall* terbaik dari setiap algoritma.

4.2. Hasil Akurasi, Presisi, *Recall* ID3 dan C4.5

Jumlah klasifikasi yang benar dibagi dengan total jumlah record klasifikasi menunjukkan tingkat akurasi secara keseluruhan. Pada persamaan 9 merupakan perhitungan akurasi pada *confussion matrix multi-class.*

Tabel 3. Perbandingan nilai akurasi model

Skenario	Proporsi		Akurasi	
	Train	*Test*	ID3	C4.5
1	90%	10%	79,07%	80,78%
2	80%	20%	79,60%	79,45%
3	70%	30%	78,94%	79,88%
4	60%	40%	79,26%	79,87%
5	50%	50%	78,56%	80,19%
Rata-Rata Akurasi			79,09%	80,03%

Pada tabel 3 dapat dilihat bahwa pembagian proporsi data mempengaruhi akurasi pada setiap algoritma, algoritma C4.5 memiliki akurasi lebih tinggi pada proporsi 90:10, 70:30, 60:40 dan 50:50 dibandingkan algoritma ID3 dimana akurasi tertinggi pada algoritma C4.5 adalah proporsi 90:10 sebanyak

80,78% dan algoritma ID3 unggul hanya pada proporsi 80:20 sebanyak 79,60% dibandingkan C4.5. Adapun Rata-rata akurasi algoritma ID3 adalah 79,09% dan rata-rata akurasi algoritma C4.5 adalah 80,03%.

Proporsi prediksi positif yang benar-benar relevan dari semua prediksi yang dinyatakan positif oleh model disebut *precision*. Model klasifikasi memiliki lima kelas, maka *precision* akan dihitung terpisah untuk masing-masing kelas. Oleh karena itu, hasil evaluasi akan menampilkan beberapa nilai *precision* (satu untuk setiap kelas), serta ukuran performa dalam klasifikasi *multi-class* yang memperhitungkan proporsi jumlah sampel di setiap kelas (*Weighted Precision*). Pada persamaan 10 merupakan perhitungan presisi *confussion matrix multi-class.*

Tabel 4. Perbandingan nilai presisi model

Skenario	Proporsi		Presisi	
	Train	*Test*	ID3	C4.5
1	90%	10%	77,03%	79,15%
2	80%	20%	78,01%	77,37%
3	70%	30%	77,71%	77,47%
4	60%	40%	77,94%	77,84%
5	50%	50%	77,10%	78,64%
Rata-Rata Presisi			77,56%	78,26%

Pada tabel 4 dapat dilihat bahwa pembagian proporsi data mempengaruhi presisi pada setiap algoritma, algoritma ID3 memiliki presisi lebih tinggi pada proporsi 80:20, 70:30 dan 60:40 dibandingkan algoritma C4.5 dimana presisi tertinggi pada algoritma ID3 adalah proporsi 80:20 sebanyak 78,01% dan algoritma C4.5 unggul hanya pada proporsi 90:10 dan 50:50 dengan presisi tertinggi yaitu 79,15%. Adapun Rata-rata presisi algoritma ID3 adalah 77,56% dan rata-rata presisi algoritma C4.5 adalah 78,26%.

Recall adalah beberapa persen data kategori positif yang terklasifikasi dengan benar oleh sistem sedangkan *weighted recall* memperhitungkan bobot dari setiap kelas dalam sebuah *dataset* yang tidak seimbang. *Recall* untuk setiap kelas dalam data klasifikasi dapat diperoleh pada persamaan 11.

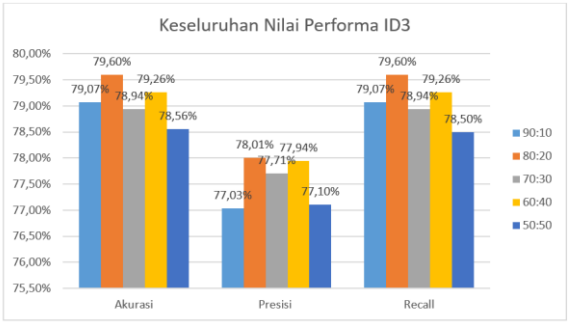
Tabel 5. Perbandingan nilai *recall* model

Skenario	Proporsi		*Recall*	
	Train	*Test*	ID3	C4.5
1	90%	10%	79,07%	80,78%
2	80%	20%	79,60%	79,45%
3	70%	30%	78,94%	79,88%
4	60%	40%	79,26%	79,87%
5	50%	50%	78,5%	80,19%
Rata-Rata *Recall*			79,07%	80,03%

Pada tabel 5 dapat dilihat bahwa pembagian proporsi data mempengaruhi *recall* pada setiap algoritma, algoritma C4.5 memiliki *recall* lebih tinggi pada proporsi 90:10, 70:30, 60:40 dan 50:50 dibandingkan algoritma ID3 dimana akurasi tertinggi pada algoritma C4.5 adalah proporsi 90:10 sebanyak 80,78% dan algoritma ID3 unggul hanya pada proporsi 80:20 dengan nilai *recall* tertinggi yaitu 79,60%. Adapun Rata-rata *recall* algoritma ID3 adalah 79,07% dan rata-rata *recall* algoritma C4.5 adalah 80,03%.

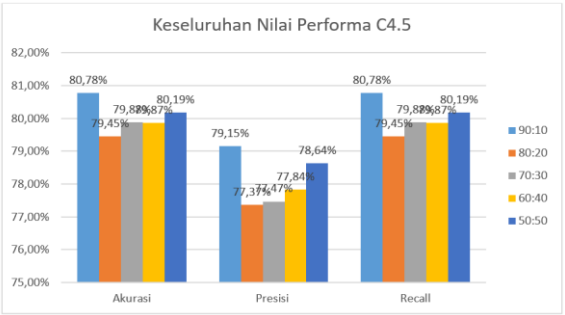
4.3. Evaluasi Model

Hasil yang didapatkan dari *output confussion matrix* diolah menjadi nilai performa untuk masing-masing algoritma. Nilai performa ini berupa akurasi, presisi dan *recall*. Secara keseluruhan nilai performa algoritma ID3 dapat dilihat pada gambar 6 dan keseluruhan nilai performa algoritma C4.5 dapat dilihat pada gambar 7.



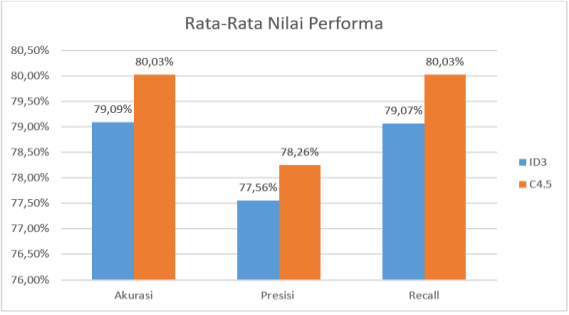
Gambar 6. Grafik keseluruhan nilai performa ID3

Gambar 6 menunjukkan bahwa data pengujian yang dilakukan menggunakan algoritma ID3 menunjukkan rata-rata akurasi sebesar 79,09%, presisi sebesar 77,56% dan *recall* sebesar 79,07%.



Gambar 7. Grafik keseluruhan nilai performa C4.5

Gambar 7 menunjukkan bahwa data pengujian yang dilakukan menggunakan algoritma C4.5 menunjukkan rata-rata akurasi sebesar 80,03%, presisi sebesar 78,26% dan *recall* sebesar 80,03%.



Gambar 8. Rata-rata nilai performa

Pada gambar 8 menunjukan rata-rata nilai performa algoritma C4.5 lebih besar dibandingkan algoritma ID3. Nilai *recall* yang lebih tinggi dari presisi maka model berhasil menangkap sebagian besar nasabah yang benar-benar memilih produk bank (*true positive* tinggi), model akan menjangkau sebanyak mungkin nasabah yang mungkin tertarik.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian dan pengujian terhadap *dataset* dengan jumlah data 16.848 terdiri dari 13 atribut, 12 atribut prediksi *age*, *gender*, *religion*, ntb, *household*, *education*, *salary*, *realiation*, *potential*, *response_status*, status_nasabah, jenis_produk dan 1 atribut tujuan yaitu nama_produk. Klasifikasi ini terdapat 5 kelas produk klasifikasi yaitu Tabungan, KPR/KPA, Kredit Ringan, Kredit Agunan Rumah dan Deposito. Atribut jenis_produk merupakan *node* pertama dan yang paling berpengaruh. Pengujian algoritma ID3 menunjukkan rata-rata akurasi 79,09%, presisi 77,56%, dan *recall* 79,07%. Sementara itu, untuk algoritma C4.5, diperoleh rata-rata akurasi 80,03%, presisi 78,26%, dan *recall* 80,03%. %. Algoritma C4.5 unggul dalam nilai akurasi, presisi, dan *recall* daripada algoritma ID3, menurut hasil implementasi secara keseluruhan. Dengan demikian, dapat dikatakan bahwa algoritma C4.5 adalah algoritma yang paling efektif untuk klasifikasi pemilihan produk perbankan. Berdasarkan kesimpulan, penelitian selanjutnya disarankan menggunakan *dataset* yang lebih besar untuk meningkatkan generalisasi model dan mengombinasikan algoritma *decision tree* dengan teknik lain seperti *ensemble learning* (*random forest* atau *boosting*) untuk meningkatkan akurasi klasifikasi, serta mengembangkan sistem berbasis *machine learning* dengan implementasi model yang dihasilkan ke dalam sistem.

DAFTAR PUSTAKA

[1] Z. Lailatul Ulfa, D. Herliandis Shodiqin, and M. Syafi, "STRATEGI MARKETING SALES FORCE (SF) DALAM MEMASARKAN PRODUK PEMBIAYAAN PENSIUN BERKAH DI BANK SYARIAH INDONESIA KCP BANYUWANGI ROGOJAMPI 1," *Jurnal Ekonomi Syariah*, vol. 6, 2024.

[2] T. Faizah and A. Jananto, "Perbandingan Algoritma C4.5 dan ID3 Untuk Prediksi Ketepatan Waktu Lulus Mahasiswa," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 8, pp. 980–990, 2021.

[3] M. Farhan Arief, I. H. Wahyono, and N. L. Chusna, "ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN ID3 UNTUK FAKTOR KEPUASAN KONSUMEN WARUNG ICHA STEAK & SEAFOOD," vol. 2, pp. 28–40, 2022.

[4] L. Rajendra Haidar, E. Sediyono, A. Iriani, and J. O. Notohamidjojo Blotongan Sidorejo, "ANALISA PREDIKSI MAHASISWA DROP OUT MENGGUNAKAN METODE DECISION TREE DENGAN ALGORITMA ID3 dan C4.5," *TRANSFORMATIKA*, vol. 17, no. 2, pp. 97–106, 2020.

[5] W. Supriyatin and Y. Rianto, "Comparative Analysis Accuracy ID3 Algorithm and C4.5 Algorithm in Selection of Candidates Basic Physics Laboratory Assistant," *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, vol. 21, no. 1, pp. 1–14, Jan. 2024, doi: 10.33751/komputasi.v21i1.9198.

[6] A. Permana, A. Hananto, S. B. Nugroho, and A. Wibowo, "Komparasi Performa Algoritma ID3, C4.5, CHAID Dalam Profiling Tersangka Kasus Narkoba Di Jawa Barat," 2021.

[7] D. P. Utomo and M. Mesran, "Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 4, no. 2, pp. 437–444, Apr. 2020, doi: 10.30865/MIB.V4I2.2080.

[8] N. Wayan Wardani *et al.*, "PENERAPAN DATA MINING UNTUK KLASIFIKASI PENJUALAN BARANG TERLARIS MENGGUNAKAN METODE DECISION TREE C4.5," 2022.

[9] K. Suhada, A. Elanda, and A. Aziz, "Dirgamaya Jurnal Manajemen dan Sistem Informasi Klasifikasi Predikat Tingkat Kelulusan Mahasiswa Program Studi Teknik Informatika dengan Menggunakan Algoritma C4.5 (Studi Kasus: STMIK Rosma Karawang)," 2021.

[10] M. R. Nahjan, N. Heryana, and A. Voutama, "IMPLEMENTASI RAPIDMINER DENGAN METODE CLUSTERING K-MEANS UNTUK ANALISA PENJUALAN PADA TOKO OJ CELL," 2023.

[11] W. I. Rahayu, C. Prianto, and E. A. Novia, "PERBANDINGAN ALGORITMA K-MEANS DAN NAÏVE BAYES UNTUK MEMPREDIKSI PRIORITAS PEMBAYARAN TAGIHAN RUMAH SAKIT BERDASARKAN TINGKAT KEPENTINGAN PADA PT. PERTAMINA (PERSERO)," *Jurnal Teknik Informatika*, vol. 13, no. 2, 2021.

[12] L. Fu, P. Liang, X. Li, and C. Yang, "A machine learning based ensemble method for automatic multiclass classification of decisions," in *ACM International Conference Proceeding Series*, Association for Computing Machinery, Jun. 2021, pp. 40–49. doi: 10.1145/3463274.3463325.

[13]M. R. Dewi, "KLASIFIKASI AKSES INTERNET OLEH ANAK-ANAK DAN REMAJA DEWASA DI JAWA TIMUR MENGGUNAKAN SUPPORT VECTOR MACHINE," *J. Ris. & Ap. Mat*, vol. 4, no. 1, pp. 17–27, 2020