JSiI | Jurnal Sistem Informasi | Vol. 9 | No. 1 | Maret 2022 | 17-26

p-ISSN: 2406-7768
e-ISSN: 2581-2181

OPTIMASI ALGORITMA C4.5 MENGGUNAKAN METODE *FORWARD SELECTION* DAN *STRATIFIED SAMPLING* UNTUK PREDIKSI KELAYAKAN KREDIT

Ibnu Ubaedi[1], Yan Mitha Djaksana[2]

[1,2] Program Studi Teknik Informatika Fakultas Teknik Universitas Pamulang
Jl. Surya Kencana No.1 Pamulang Barat, Pamulang - Tangerang Selatan, Banten

[1]ubaediibnu@gmail.com
[2]dosen01994@unpam.ac.id

Abstrak

Kredit merupakan dana yang diberikan oleh bank kepada pihak lain berdasarkan perjanjian pinjam-meminjam, yang mewajibkan peminjam melunasi pinjamannya setelah jangka waktu tertentu. Sebelum memberi kredit, bank perlu menganalisis kemampuan nasabah dalam melunasi utangnya untuk mengurangi resiko kredit. Analisis kredit yang dilakukan oleh bank terkadang tidak akurat, sehingga menimbulkan kredit macet. Masalah ini akan diselesaikan dengan memanfaatkan teknologi data mining, yaitu membuat pohon keputusan menggunakan algoritma C4.5 untuk memprediksi kelayakan kredit. Tetapi, algoritma C4.5 memiliki masalah penurunan akurasi ketika dihadapkan dengan atribut dan jumlah data yang besar. Masalah ini akan diselesaikan dengan metode *Feature Selection* dan *Stratified Sampling* yang terbukti dapat menyelesaikan masalah atribut dan jumlah data yang besar. Hasil penelitian menunjukkan bahwa pohon keputusan yang dihasilkan dari algoritma C4.5 memiliki akurasi cukup baik sebesar 79,11% dan metode *Feature Selection* dan *Stratified Sampling* berhasil meningkatkan akurasi algoritma C4.5 sebesar 9,2% dalam memprediksi kelayakan kredit.

Kata kunci - Optimasi Algoritma C4.5, *Feature Selection*, *Stratified Sampling*, Kelayakan Kredit, Atribut dan Jumlah Data yang Besar

I. PENDAHULUAN

Kredit merupakan dana yang diberikan oleh bank kepada pihak lain berdasarkan perjanjian pinjam-meminjam, yang mewajibkan peminjam melunasi pinjamannya setelah jangka waktu tertentu, dengan memberi bunga sebagai imbalannya.

Sebelum memberi kredit, bank perlu menganalisis kemampuan nasabah dalam melunasi pinjaman beserta dengan bunganya. Hal ini dilakukan untuk mengurangi resiko kerugian yang disebabkan ketidak-mampuan nasabah dalam melunasi utangnya[1]. Analisis kredit umumnya dilakukan dengan prinsip 5C, antara lain *character* (kepribadian), *capacity* (kapasitas), *capital* (modal), *collateral* (jaminan), dan *condition of Economy* (keadaan perekonomian)[2]. Akan tetapi, cara tersebut memiliki masalah pada waktu proses yang lama[3] dan hasil yang kurang akurat terbukti dengan masih banyak kredit yang bermasalah (macet)[4].

Masalah tersebut dapat diatasi dengan memanfaatkan teknologi *data mining*. *Data Mining* merupakan disiplin ilmu yang mempelajari metode untuk menyarikan pengetahuan atau menemukan pola dari suatu data yang besar[5]. Pola yang dihasilkan dari pembelajaran algoritma *data mining* akan digunakan untuk memprediksi layak atau tidak nasabah menerima kredit.

Ada beberapa algoritma data mining yang dapat digunakan untuk memprediksi kelayakan kredit, seperti *K-Nearest Neighbor*[6], *Naive Bayes*[7], dan C4.5[8][9][10][11]. *K-Nearest Neighbor* memiliki kelebihan pada model pembelajaran yang tidak mengasumsikan apa-apa terhadap distribusi data (non parametrik), akan tetapi memiliki kelemahan rentan terhadap atribut yang non-informatif[12]. *Naïve Bayes* memiliki kelebihan pada model yang mudah dibuat dan proses perhitungan yang cepat, akan tetapi memiliki kelemahan hanya mengandalkan satu probabilitas (peluang) saja saat mengukur akurasi[13].

C4.5 dapat mengatasi kerentanan atribut non-informatif dan pengukuran akurasi yang mengandalkan satu probabilitas saja dengan membuat model pohon keputusan. Akan tetapi, pohon keputusan memiliki kelemahan ketika dihadapkan dengan atribut dan jumlah data yang besar[14].

Berdasarkan penelitian[13][15] metode *Feature Selection* dapat mengatasi besarnya jumlah atribut, dan penelitian[16][17][18][19][20] metode *Sampling* berhasil mengatasi besarnya jumlah data. Pada penelitian ini, metode

p-ISSN: 2406-7768
e-ISSN: 2581-2181

Feature Selection dan *Stratified Sampling* akan diterapkan pada algoritma C4.5 untuk memprediksi kelayakan kredit.

Berdasarkan latar belakang masalah yang telah diuraikan diatas, dapat dibuat rumusan masalah sebagai berikut Seberapa efektif algoritma C4.5 dalam memprediksi kelayakan kredit? Bagaimana peningkatan akurasi algoritma C4.5 apabila metode *Feature Selection* dan *Stratified Sampling* diterapkan dalam memprediksi kelayakan kredit?

Tujuan penelitian ini membuat pola berupa pohon keputusan menggunakan algoritma C4.5 untuk memprediksi kelayakan kredit. Selain itu, penelitian ini juga bertujuan untuk mengoptimasi algoritma C4.5 menggunakan metode *Feature Selection* dan *Stratified Sampling* untuk memprediksi kelayakan Kredit.

II. METODOLOGI PENELITIAN

2.1 *Data Mining*

Data Mining merupakan disiplin ilmu yang mempelajari metode untuk menyarikan pengetahuan atau menemukan pola dari suatu data yang besar[5]. Proses menyarikan atau ekstraksi data jadi pengetahuan dimulai dari data yang tidak memiliki arti, kemudian diolah menjadi informasi berupa rangkuman dan statistik data, terakhir metode/algoritma data mining diterapkan pada informasi untuk menghasilkan pengetahuan berupa pola, rumus, aturan atau model.

Data mining memiliki fungsi mencari pengetahuan yang bermanfaat dari sekumpulan data yang banyak. Menurut Larose[21] terdapat 5 peran utama *data mining*, antara lain:

1. Estimasi
   Digunakan untuk memperikirakan nilai yang belum diketahui, target nilai bersifat numerik dari atribut numerik. Contohnya estimasi waktu pengiriman *pizza*.
2. *Forecasting*
   Sama dengan estimasi, bedanya terdapat penambahan atribut *time series*. Contohnya *forecasting* harga saham.
3. Klasifikasi
   Digunakan untuk memprediksi kejadian dimasa depan. Target prediksi bersifat nominal dari atribut nominal atau numerik. Contohnya klasifikasi tingkat kelulusan mahasiswa tepat waktu.
4. Klastering
   Digunakan untuk mengelompokkan objek yang serupa dalam satu klaster, tetapi antar klaster mempunyai karakter yang berbeda. Klatering tidak memiliki target, pengelompokan dibuat dari atribut yang bersifat numerik. Contohnya klaster jenis pelanggan.
5. Asosiasi
   Digunakan untuk mencari hubungan yang ada pada nilai atribut dari sekumpulan data yang sering muncul secara bersamaan. Contohnya asosiasi pembelian barang di supermarket.

2.2 Klasifikasi *Data Mining*

Klasifikasi pada *data mining* merupakan proses membagi objek tertentu menjadi beberapa kategori sesuai dengan sifatnya masing-masing[22]. Hasil dari klasifikasi digunakan untuk memprediksi nilai yang belum diketahui dari variabel yang ada. Terdapat 4 komponen dasar dalam proses klasifikasi *data mining*, antara lain:

1. *Class*
   Merupakan variabel yang dipengaruhi oleh variabel lain atau biasa disebut dengan variabel dependen. Variabel ini bersifat kategorikal yang mewakili *"label"* yang terkandung dalam data. Contohnya status kredit nasabah: macet atau lancar.
2. *Predictors*
   Merupakan variabel yang mempengaruhi variabel lain atau biasa disebut dengan variabel independen. Variabel diwakilkan oleh karakteristik (atribut) data. Contohnya usia, jumlah pinjaman, jangka waktu, dan lain sebagainya.
3. *Training Dataset*
   Merupakan kumpulan data yang memuat nilai dari *class* dan *predictors* yang digunakan untuk menentukan *class* yang cocok berdasarkan *predictors*. Contohnya kumpulan data risiko kredit.
4. *Testing Dataset*
   Merupakan data baru yang digunakan untuk menguji model yang telah dibuat dan hasil akurasinya akan dievaluasi.

2.3 Algoritma C4.5

Algoritma C4.5 merupakan algoritma yang membuat pohon keputusan, dikembangkan oleh J. Ross Quinlain dari algoritma ID3 yang sudah dia temukan sebelumnya. Pohon keputusan merupakan model klasifikasi yang dibuat menggunakan struktur pohon. Caranya, data pada tabel diubah menjadi model pohon, kemudian model pohon diubah menjadi aturan lalu disederhanakan. Pohon keputusan digunakan untuk menguji data dan menemukan hubungan tersembunyi antara variabel dependen dengan variabel independen[5].

Pohon keputusan begitu populer karena aturan klasfikasi yang sederhana dan mudah dimengerti. Selain itu, proses pembelajarannya relatif lebih cepat bila dibandingan dengan metode klasifikasi lainnya.

Ada beberapa tahapan untuk membuat pohon keputusan dalam algoritma C4.5, antara lain[21]:

1. Siapkan *data training*
   Data training digunakan sebagai objek dalam membuat sebuah pohon keputusan pada algoritma C4.5.
2. Pilih atribut sebagai akar

p-ISSN: 2406-7768
e-ISSN: 2581-2181

Atribut akar dipilih berdasarkan nilai *gain ratio* tertinggi dari atribut-atribut yang ada. Untuk mendapatkan nilai *gain ratio*, dilakukan dengan beberapa tahapan sebagai berikut:

1. Hitung *Entropy* Akar

(1)

$$\text{Entropy}(S)=\sum_{i=1}^{n} - pi^{*} \log_2 p\, i$$

Keterangan:
S = himpunan kasus
n = jumlah partisi S
pi = proporsi dari Si terhadap S

2. Hitung *Gain* Akar

(2)

$$\text{Gain}(S,A)=\text{Entropy}(S)-\sum_{i=1}^{n} \frac{|S_i|}{|S|} * \text{Entropy}(S_i)$$

Keterangan:
S = himpunan kasus
A = atribut
n = jumlah partisi atribut A
| Si | = jumlah kasus pada partisi ke-i
| S | = jumlah kasus dalam S

3. Hitung *Split Info* Akar

(3)

$$\text{SplitInfo}_A(D)=-\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|}\right)$$

Keterangan:
Dj = jumlah kasus pada partisi ke-i
D = jumlah kasus dalam D

4. Hitung *Gain Ratio* Akar

(4)

$$\text{GainRatio}(A)=\frac{\text{Gain}(A)}{\text{SplitInfo}(A)}$$

3. Buat cabang untuk tiap-tiap nilai
   Setelah mendapat atribut akar, buat cabang dari atribut akar tersebut.
4. Ulangi proses untuk setiap cabang sampai semua kasus pada cabang memiliki kelas yang sama

2.4 *Forward Selection*

Data set bisa berisi ratusan atribut, yang diantaranya mungkin tidak relevan dengan pola yang ingin dicari. Misalnya ingin mengklasifikasi status kredit nasabah yang kreditnya macet atau lancar, atribut seperti nomor telepon mungkin tidak relevan bila dibandingkan dengan atribut usia dan riwayat kredit. Mempertahankan atribut yang tidak relevan dapat memperlambat proses pencarian pola dan membuat algoritma data mining bingung sehingga menghasilkan pola yang buruk.

Untuk mengatasi atribut yang tidak relevan, dapat dilakukan dengan menguranginya (menghapusnya)[5]. Pilih atribut yang terbaik dengan kontribusi akurasi yang tinggi dan hapus atribut yang tidak relevan atau bisa disebut dengan *Feature Selection*. Tujuan *Feature Selection* adalah menemukan kumpulan atribut yang minim dari sekumpulan atribut yang berjumlah banyak. Manfaat *Feature Selection* adalah membantu membuat pola yang mudah dipahami.

Forward Selection merupakan salah satu *Feature Selection*. Prosedur dimulai dengan himpunan kosong dari atribut yang ingin dikurangi. Kemudian, atribut diuji satu-persatu dan dipilih atribut terbaik dengan dampak ke akurasi paling tinggi. Lalu, lakukan iterasi pengujian berikutnya terus menerus dan berhenti sampai atribut yang diuji tidak memberikan dampak akurasi yang signifikan. Penjelasannya sebagai berikut[5]:
Atribut asli: {A01, A02, A03, A04, A05, A06}
Tahap pengurangan atribut:

1. {A01}
2. {A01, A04}
3. Pengurangan atribut {A01, A04, A06}

2.5 *Stratified Sampling*

Pengambilan sampel dapat digunakan untuk mereduksi data. Populasi data yang besar dapat diwakili oleh sampel acak yang ukurannya jauh lebih kecil. Ada beberapa teknik pengambilan sampel, salah satunya *Stratified Sampling*.

Stratified Sampling merupakan teknik pengambilan sampel secara acak dengan memperhatikan tingkatan (strata) pada populasi. *Dataset* dibagi menjadi beberapa bagian yang terpisah (strata), kemudian sampel diambil secara acak berdasarkan strata yang sudah dibuat[5]. Adapun tahapan *Stratified Sampling* sebagai berikut:

1. Pertama, populasi N dibagi menjadi beberapa sub-sub populasi yang masing-masing sub-sub populasi tersebut terdiri dari elemen $N1$, $N2$, $N3$, ..., NL.
2. Kemudian, diantara sub-sub populasi, tidak boleh ada tumpang tindih, sehingga $N1 + N2 + N3 + \cdots + NL = N$.
3. Terakhir, ambil sampel secara acak dari masing-masing sub populasi dengan alokasi sampel yang proporsional.

Sebelum pengambilan sampel, menentukan ukuran sampel merupakan hal yang penting. Sampel yang diambil harus mencerminkan populasi. Ada beberapa cara dalam menentukan ukuran sampel. Salah satunya yang paling banyak dan umum menggunakan teori slovin dijelaskan dengan rumus sebagai berikut:

(5)

$$n=\frac{N}{1+Ne^2}$$

p-ISSN: 2406-7768
e-ISSN: 2581-2181

Keterangan:
n = besar sampel
N = ukuran populasi atau jumlah elemen dalam populasi
e = nilai presisi atau tingkat signifikansi yang telah ditentukan. Umumnya dalam penelitian tingkat signifikansi ditentukan diantara 90%, 95%, dan 99%.

2.6 Evaluasi dan Validasi Hasil

1. *Cross Validation*

   Cross Validation merupakan proses pembagian (pemisahan) data secara acak menjadi beberapa *subset* dengan ukuran yang sama[5]. *Subset* digunakan untuk *data testing* dan sisanya untuk *data training*. Dengan demikian, setiap data memiliki kesempatan yang sama untuk jadi *data training* dan *testing*. *Cross Validation* digunakan untuk menghindari *data testing* yang tumpang tindih. *10-Fold Cross-Validation* akan mengulang pengujian sebanyak 10 kali dan hasil pengukurannya merupakan nilai rata-rata dari 10 kali pengujian tersebut.

2. *Confusion Matrix*

   Confusion Matrix merupakan alat untuk menganalisis kemampuan pengklasifikasi dalam mengidentifikasi atribut dari kelas yang berbeda[5]. *Confusion Matrix* memberikan penilaian benar atau salah berdasarkan kinerja klasifikasi objeknya[22]. *Confusion Matrix* pada tabel I merupakan matrik 2 dimensi yang digunakan untuk membandingkan hasil prediksi dengan kenyataan.

Tabel 1. *Confusion matrix*

Actual class				
Predict class		Yes	No	Total
	Yes	TP	FP	P’
	No	FN	TN	N’
	Total	P	N	P + N

Keterangan:
1. *Predict class yes, actual class yes*: jumlah data yang diprediksi *yes* dan kenyataannya *yes* (TP).
2. *Predict class no, actual class no*: jumlah data yang diprediksi *no* dan kenyataannya *no* (TN).
3. *Predict class yes, actual class no*: jumlah data yang diprediksi *yes* dan kenyataannya *no* (FP).
4. *Predict class no, actual class yes*: jumlah data yang diprediksi *no* dan kenyataannya *yes* (FN).

Berikut adalah persamaan model *Confusion Matrix*:

1. Akurasi

   Merupakan persentase tupel set pengujian yang diklasifikasikan dengan benar oleh pengklasifikasi.

(6)

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

2. Sensitivitas

   Merupakan proporsi tupel positif yang diidentifikasi dengan benar (tingkat positif sebenarnya).

(7)

$$\text{Sensitivitas} = \frac{TP}{TP+FN}$$

3. Spesifisitas

   Merupakan proporsi tupel negatif yang diidentifikasi dengan benar (tingkat negatif yang sebenarnya).

(8)

$$\text{Spesifisitas} = \frac{TN}{TN+FP}$$

4. PPV

   Merupakan proporsi kasus dengan hasil tes positif yang didiagnosis dengan benar.

(9)

$$\text{PPV} = \frac{TP}{TP+FP}$$

5. NPV

   Merupakan mengukur proporsi 'negatif' benar yang diidentifikasi dengan benar.

(9)

$$\text{NPV} = \frac{TN}{TN+FN}$$

3. *Area Under ROC Curve*

   Kurva ROC (*Receiver Operating Characteristic*) merupakan grafik antara sensitivitas sumbu Y dan 1-spesifisitas sumbu X. Kurva ROC ini tampaknya menggambarkan korespondensi antara sumbu Y atau sensitivitas dengan sumbu X atau spesifisitas. Kurva ROC banyak digunakan dalam pelatihan dan *data mining.*

   Nilai Kurva ROC dapat dijadikan bahan evaluasi sehingga kita dapat membandingkan beberapa algoritma. Pada kurva ROC, klasifikasi merupakan metode visualisasi, pengorganisasian dan pemilihan klasifikasi berdasarkan kinerja algoritma[22].

   Menurut akurasi nilai AUC (*Area Under ROC Curve*) dalam klasifikasi *data mining* dibagi menjadi lima kelompok[22], yaitu:
   1. 0.90 - 1.00 = klasifikasi yang sangat baik
   2. 0.80 - 0.90 = klasifikasi yang baik
   3. 0.70 - 0.80 = klasifikasi yang cukup
   4. 0.60 - 0.70 = klasifikasi yang buruk
   5. 0.50 - 0.60 = klasifikasi yang salah

2.7 *UCI Machine Learning Repository*

UCI Machine Learning Repository merupakan kumpulan *database*, teori *domain*, dan generator data

p-ISSN: 2406-7768
e-ISSN: 2581-2181

yang digunakan oleh komunitas *machine learning* untuk menganalisis secara empiris algoritma *machine learning*. Data dibuat sebagai arsip ftp pada tahun 1987 oleh David Aha dan mahasiswa pascasarjana lain di Universitas California Irvine.

Sejak saat itu, data telah digunakan secara luas oleh siswa, pengajar, dan peneliti di seluruh dunia sebagai sumber utama kumpulan data *machine learning*. Data telah dikutip lebih dari 1000 kali, yang menjadikannya salah satu dari 100 "makalah" teratas yang paling banyak dikutip dalam semua ilmu komputer.

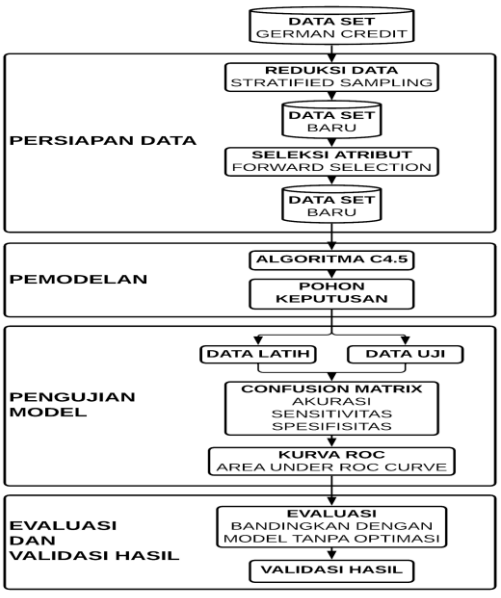
2.8 RapidMiner

RapidMiner merupakan platform perangkat lunak *data science* yang menyediakan lingkungan untuk pembelajaran mesin (*machine learning*), pembelajaran mendalam (*deep learning*), penambangan teks (*text mining*), dan analisis prediktif (*predictive analytics*). Aplikasi ini digunakan untuk bisnis, penelitian, pendidikan, pelatihan, pembuatan *prototipe*, dan pengembangan aplikasi. RapidMiner mendukung semua tahapan pembelajaran mesin termasuk persiapan data, visualisasi hasil, validasi dan pengoptimalan.

Di bidang akademik, RapidMiner digunakan oleh mahasiswa, dosen atau peneliti yang berpengalaman untuk memodelkan sistem berbasis kecerdasan buatan dalam sistem informasi dan teknik informatika untuk memodelkan sistem berbasis kecerdasan buatan (optimasi, pengenalan pola gambar/teks/grafik, prediksi).

2.9 Kerangka Pemikiran

Penelitian ini dilakukan berdasarkan Gambar 1 sebagai berikut:



Gambar 1. Kerangka Pemikiran

III. HASIL DAN PEMBAHASAN

3.1 Desain Penelitian

Metode penelitian yang digunakan adalah metode penelitian eksperimen dan studi pustaka, yang dibagi menjadi beberapa tahapan sebagai berikut:

1. Pengumpulan Data
   Tahapan pertama yang dilakukan pada penelitian ini adalah pengumpulan data. Data dikumpulkan dari UCI *Machine Learning Repository* bersifat terbuka dan banyak digunakan peneliti data mining.
2. Pengolahan Awal Data
   Tahapan berikutnya adalah pengolahan awal data. Pada penelitian ini pengolahan awal data dilakukan dengan mengurangi data (reduksi data) menggunakan *Stratified Sampling*.
3. Eksperimen dan Pengujian Model
   Tahapan inti pada penelitian ini adalah eksperimen dan pengujian model. Sebelum membuat pemodelan, peneliti menseleksi atribut yang ada pada *dataset*, memilih atribut apa saja yang paling berpengaruh pada akurasi, dan menghilangkan atribut yang tidak berpengaruh. Kemudian, baru membangun pohon keputusan berdasarkan eksperimen yang telah dilakukan sebelumnya dan menguji seberapa akurat model yang telah dihasilkan.
4. Evaluasi dan Validasi Hasil
   Tahapan terakhir pada penelitian ini adalah evaluasi dan validasi hasil. Evaluasi dilakukan dengan melihat seberapa besar pengaruh eksperimen yang telah dilakukan dengan membandingkan model tanpa eksperimen. Baru kemudian model dapat diterapkan.

3.2 Pengumpulan Data

Penelitian ini menggunakan *dataset german credit data* yang bersifat publik (terbuka) diperoleh dari situs UCI *Machine Learning Repository* https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data).

Dataset ini memiliki 1000 *record* terdiri dari 700 *record* dengan status kredit lancar dan 300 *record* dengan status kredit macet. Kemudian, dataset ini memiliki 21 atribut, yang terbagi menjadi 20 atribut biasa (7 atribut numerik dan 13 atribut nominal) dan 1 atribut target yang bersifat nominal. Karakteristik dataset ini dijelaskan pada tabel II sebagai berikut:

Tabel 2.I *German credit data*

No	Nama atribut	Tipe data	Deskripsi (Isi atribut)
A1	Status	*Polyno*	A11 0 DM

p-ISSN: 2406-7768
e-ISSN: 2581-2181

	rekening koran	*minal*	A12 Kurang dari 200 DM A13 Lebih/sama dengan 200 DM A14 Tidak ada
A2	Durasi (bulan)	*Integer*	Berisi angka
A3	Riwayat kredit	*Polyno minal*	A30 Tidak mengambil kredit A31 Semua kredit telah lunas A32 Kredit yang ada telah dibayar A33 Pembayaran tertunda A34 Pembayaran bermasalah
A4	Tujuan kredit	*Polyno minal*	A40 Mobil baru A41 Mobil bekas A42 Peralatan A43 Furnitur A44 Tv A45 Peralatan rumah tangga A46 Perbaikan A47 Kursus A48 Bisnis
A5	Jumlah kredit	*Integer*	Berisi angka
A6	Tabungan	*Polyno minal*	A61 Kurang dari 100 DM A62 100 – 499 DM A63 500 – 999 DM A64 Lebih/sama dengan 100 DM A65 Tidak ada
A7	Lama bekerja	*Polyno minal*	A71 Tidak bekerja A72 Kurang dari setahun A73 1 – 3 tahun A74 4 – 6 tahun A75 Lebih/sama dengan 7 tahun
A8	Tingkat angsuran dalam persentase pendapatan yang dapat disisihkan	*Integer*	Berisi angka
A9	Jenis kelamin dan status	*Polyno minal*	A91 Pria bercerai A92 Wanita menikah/bercerai A93 Pria lajang A94 Pria menikah
A10	Debitur lain/penjamin	*Polyno minal*	A101 Tidak ada A102 Pemohon bersama A103 Penjamin
A11	Lama tinggal (tahun)	*Integer*	Berisi angka
A12	Aset/kekayaan	*Polyno minal*	A121 Real estate A122 Asuransi jiwa A123 Mobil atau lain-lain A124 Tidak ada
A13	Usia (tahun)	*Integer*	Berisi angka
A14	Cicilan lain	*Polyno minal*	A141 Bank A142 Toko A143 Tidak ada
A15	Tempat tinggal	*Polyno minal*	A151 Sewa A152 Pemilik A153 Gratis
A16	Jumlah kredit pada bank ini	*Integer*	Berisi angka
A17	Pekerjaan	*Polyno minal*	A171 Tidak terampil (penduduk) A172 Tidak terampil (non penduduk) A173 Pegawai terampil A174 Pengusaha
A18	Jumlah tanggungan	*Integer*	Berisi angka
A19	Telepon	*Binominal*	A191 Tidak ada A192 Terdaftar
A20	Pekerja asing	*Binominal*	A201 Ya A202 Tidak
A21	Status kredit	*Binominal*	1 Lancar 2 Macet

3.3 Pengolahan Data Awal

Setelah dikumpulkan, data diolah terlebih dahulu sebelum dibuat pemodelan. Pengolahan awal data dilakukan dengan menggunakan *Stratified Sampling*. Sampel data diambil secara acak dari populasi data dengan memperhatikan tingkat distribusi data untuk membentuk sebuah *dataset* baru.

Ukuran sampel ditentukan menggunakan rumus slovin dengan taraf signifikansi sebesar 90% dengan perhitungan sebagai berikut:

$$\text{sampel} = \frac{1000}{1+1000*0{,}1^2} = 91$$

Dari 91 sampel yang akan diambil, alokasikan secara proporsional agar sampel yang diambil

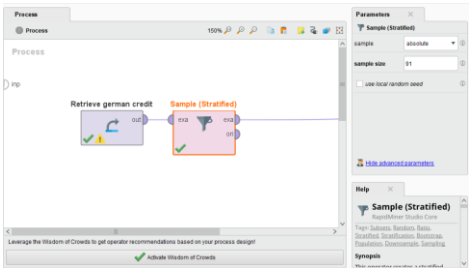
p-ISSN: 2406-7768
e-ISSN: 2581-2181

mencerminkan populasi. Perhitungan alokasi sampel sebagai berikut:

$$\text{sampel (kredit lancar)} = \frac{700}{1000} * 91 = 64$$

$$\text{sampel (kredit macet)} = \frac{300}{1000} * 91 = 27$$

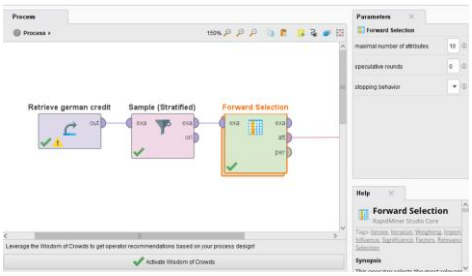
Untuk mempermudah proses *Stratified Sampling* digunakan aplikasi RapidMiner dengan tampilan pada Gambar 2 sebagai berikut:



Gambar 2. *Stratified sampling*

3.4 Eksperimen dan Pengujian Model

Setelah mendapatkan *dataset* baru yang ukurannya lebih kecil (*sampling*), eksperimen dilakukan dengan memilih atribut terbaik dan menghapus atribut yang tidak memiliki kontribusi ke akurasi. Untuk memilih (menyeleksi) atribut terbaik menggunakan metode *Forward Selection*. Setiap atribut diuji satu-persatu dengan cara dibangun model, kemudian model diuji untuk melihat seberapa besar akurasi yang dihasilkan. Atribut dengan akurasi yang paling besar dipilih, kemudian lakukan pengujian kembali dengan atribut-atribut yang masih ada. Proses dilakukan terus-menerus dan berhenti jika atribut yang diuji sudah tidak memberikan peningkatan akurasi yang signifikan. Untuk mempermudah proses penilaian atribut, digunakan aplikasi RapidMiner dengan tampilan pada Gambar 3 sebagai berikut:



Gambar 3. *Forward selection*

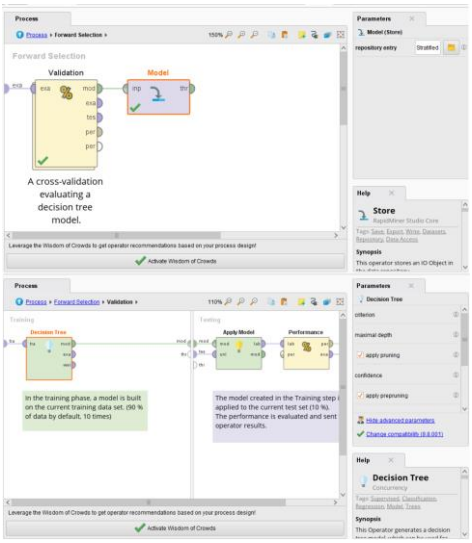
Proses pada gambar 3 menghasilkan penilaian atribut yang dijelaskan pada Tabel III sebagai berikut:

Tabel 3. Penilaian atribut

No	Nama Atribut	Bobot
A18	Jumlah tanggungan	1
A16	Jumlah kredit pada bank ini	1
A15	Tempat tinggal	1
A12	Aset/kekayaan	1
A3	Riwayat kredit	1
A20	Pekerja asing	0
A19	Telepon	0
A17	Pekerjaan	0
A14	Cicilan lain	0
A13	Usia (tahun)	0
A11	Lama tinggal (tahun)	0
A10	Debitur lain/penjamin	0
A9	Jenis kelamin dan status	0
A8	Tingkat angsuran dalam persentase pendapatan yang dapat disisihkan	0
A7	Lama bekerja	0
A6	Tabungan	0
A5	Jumlah kredit	0
A4	Tujuan kredit	0
A2	Durasi (bulan)	0
A1	Status rekening koran	0

Berdasarkan Tabel III atribut yang memiliki bobot 1 merupakan atribut terbaik dan dipilih untuk dibuat model. Sementara atribut dengan bobot 0 merupakan atribut yang tidak memiliki kontribusi ke akurasi dan akan dihapus sehingga menghasilkan *dataset* (baru) yang jumlah atributnya lebih sedikit.

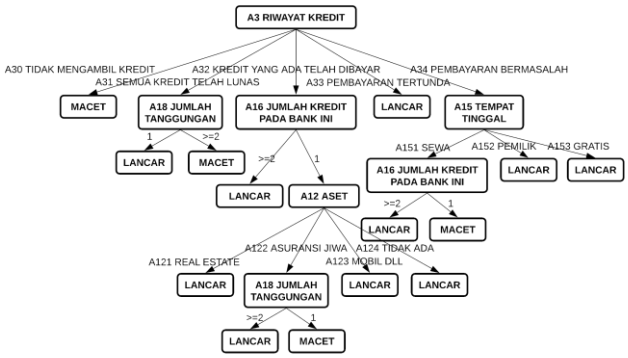
Setelah eksperimen dilakukan, sekarang masuk ke tahap pemodelan dengan membuat pohon keputusan menggunakan aplikasi RapidMiner dengan tampilan pada Gambar 4 sebagai berikut:



Gambar 4. *Modeling*

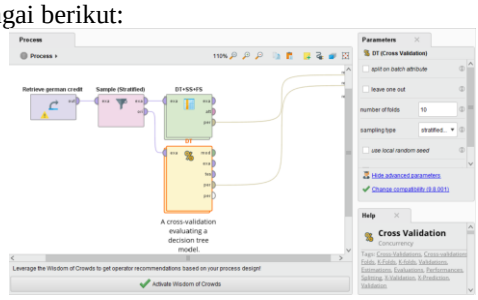
p-ISSN: 2406-7768
e-ISSN: 2581-2181

Proses pada Gambar 4 menghasilkan pola yang dijelaskan pada Gambar 5 sebagai berikut:



Gambar 5. Pohon keputusan

Berdasarkan Gambar 5, pohon keputusan perlu dilakukan pengujian dan untuk mengetahui seberapa besar nilai akurasi, sensitifitas, spesifisitas, dan *Area Under ROC Curve (AUC)* yang dihasilkan. Pengujian dilakukan menggunakan aplikasi RapidMiner dengan *10 Cross Validation* dengan tampilan pada Gambar 6 sebagai berikut:



Gambar 6. Pengujian model

3.5 Hasil Eksperimen dan Pengujian Model

Eksperimen dilakukan menggunakan laptop Lenovo Yoga Slim 7 dengan prosesor Intel generasi ke-10 Core i5-1035G1 CPU @ 1.00 GHz, memori (RAM) 8.00 GB, sistem operasi Windows 10 64-bit dan aplikasi RapidMiner Studio Educational versi 9.8.001 pada Gambar 6 menghasilkan klasifikasi (prediksi) status kredit pada Tabel IV sebagai berikut:

Tabel 4. *Confusion matrix*

	Kenyataan Lancar	**Kenyataan Macet**	**Total**
Prediksi Lancar	60	15	75
Prediksi Macet	4	12	16
Total	64	27	91

3.6 Evaluasi dan Validasi Hasil

Berdasarkan hasil klasifikasi pada Tabel IV dilakukan pengukuran dengan persamaan sebagai berikut:

$$\text{Akurasi} = \frac{60+12}{60+12+15+4} \times 100\% = 79{,}11\%$$

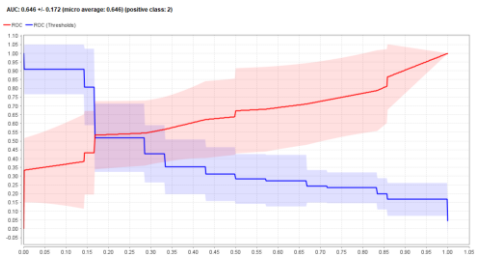
$$\text{Sensitivitas} = \frac{60}{60+4} \times 100\% = 93{,}75\%$$

$$\text{Spesifisitas} = \frac{12}{12+15} \times 100\% = 44{,}44\%$$

$$\text{PPV} = \frac{60}{60+15} \times 100\% = 80{,}00\%$$

$$\text{NPV} = \frac{12}{12+4} \times 100\% = 75{,}00\%$$

Dari hasil pengukuran, dibuat kurva ROC untuk mengetahui nilai *Area Under ROC Curve* (AUC) dengan aplikasi RapidMiner menghasilkan tampilan pada Gambar 7 sebagai berikut:



Gambar 7. AUC

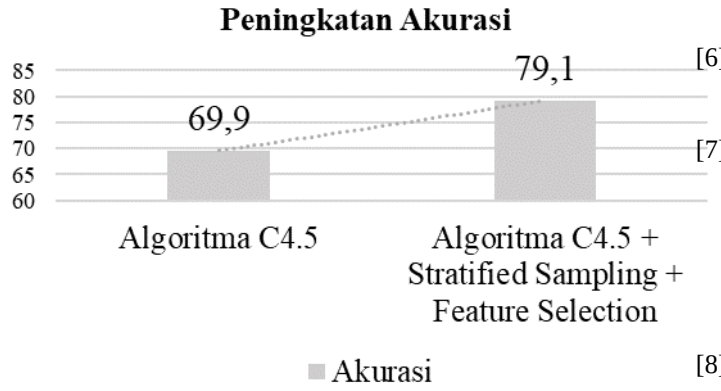
Berdasarkan kurva ROC pada Gambar 7, nilai AUC yang dihasilkan sebesar 0,646. Kemudian dibuat pula perhitungan algoritma C4.5 tanpa optimasi untuk mengetahui seberapa besar peningkatan yang dihasilkan. Hasil pengukurannya dijelaskan pada Tabel V sebagai berikut:

Tabel 5. Hasil pengukuran

Model	**Akurasi**	**Sensitivitas**	**Spesifisitas**	**PPV**	**NPV**	**AUC**
Algoritma C4.5	69,90%	88,00%	27,67%	73,95%	49,70%	0,705
Algoritma C4.5 + *Stratified Sampling* + *Forward Selection*	79,11%	93,75%	44,44%	80,00%	75,00%	0,646

Berdasarkan hasil pengukuran pada Tabel V, akurasi algoritma C4.5 + *Stratified Sampling* + *Forward Selection* dibandingkan dengan algoritma C4.5 tanpa optimasi yang dijelaskan pada Gambar 8 sebagai berikut:

p-ISSN: 2406-7768
e-ISSN: 2581-2181



Gambar 8. Hasil peningkatan akurasi

Berdasarkan Gambar 8 hasil pengujian model algoritma C4.5 + *Stratified Sampling* + *Feature Selection* mengalami peningkatan akurasi sebesar 9,2% saat dibandingkan dengan model algoritma C4.5 tanpa optimasi dalam memprediksi kelayakan kredit.

IV. KESIMPULAN

Berdasarkan hasil penelitian, dapat ditarik kesimpulan yang menjawab rumusan masalah sebagai berikut:

1. Algoritma C4.5 terbukti efektif dalam memprediksi kelayakan kredit dengan tingkat akurasi sebesar 79,11%.
2. Metode *Forward Selection* dan *Stratified Sampling* terbukti berhasil meningkatkan akurasi algoritma C4.5 sebesar 9,2% dalam memprediksi kelayakan kredit.

REFERENSI

[1] R. A. Saraswati, "Peranan Analisis Laporan Keuangan, Penilaian Prinsip 5C Calon Debitur Dan Pengawasan Kredit Terhadap Efektivitas Pemberian Kredit Pada Pd Bpr Bank Pasar Kabupaten Temanggung," *Nominal, Barom. Ris. Akunt. dan Manaj.*, vol. 1, no. 1, 2012, doi: 10.21831/nominal.v1i1.994.

[2] N. Wahyuni, "Penerapan Prinsip 5C Dalam Pemberian Kredit Sebagai Perlindungan Bank," *Lex J. Kaji. Huk. Keadilan*, vol. 1, no. 1, 2017, doi: 10.25139/lex.v1i1.236.

[3] N. Eprianti, "Penerapan Prinsip 5C Terhadap Tingkat Non Performing Financing (Npf)," *Amwaluna J. Ekon. dan Keuang. Syariah*, vol. 3, no. 2, 2019, doi: 10.29313/amwaluna.v3i2.4645.

[4] OJK, "Laporan Profil Industri Perbankan Triwulan III 2020," *Otoritas Jasa Keuang.*, 2020.

[5] J. Han, M. Kamber, and J. Pei, "Third Edition : Data Mining Concepts and Techniques," *J. Chem. Inf. Model.*, vol. 53, no. 9, pp. 1689–1699, 2012, [Online]. Available: http://library.books24x7.com/toc.aspx?bkid=44712.

[6] H. Harafani and H. A. Al-kautsar, "Meningkatkan Kinerja K-Nn Untuk Klasifikasi Kanker," vol. 18, no. 1, 2021.

[7] D. Riana, D. Riana, P. Pembiayaan, and N. Koperasi, "Algoritma Naïve Bayes , Decision Tree , dan SVM untuk Klasifikasi Persetujuan Pembiayaan Nasabah Koperasi Syariah Naïve Bayes , Decision Tree , and SVM Algorithm for Classification of Sharia," vol. 7, no. July, pp. 77–82, 2019, doi: 10.14710/jtsiskom.7.2.2019.77-82.

[8] A. Rifai, R. Aulianita, and D. Mining, "Komparasi Algoritma Klasifikasi C4.5 dan Naïve Bayes Berbasis Particle Swarm Optimization Untuk Penentuan Resiko Kredit," vol. 10, no. 2, 2018.

[9] M. Algoritma, C. Untuk, N. Iriadi, and N. Nuraeni, "Kajian Penerapan Metode Klasifikasi Data Kelayakan Kredit Pada Bank," pp. 132–137.

[10] S. Atribut, P. Algoritma, M. Genetik, and A. Dan, "Bagging Untuk Analisa Kelayakan," vol. 04, no. 02, pp. 174–183, 2017.

[11] C. Menggunakan and B. Pso, "Kelayakan Kredit Bank," vol. II, no. 1, pp. 26–30, 2019.

[12] S. Harlina, "Data Mining Pada Penentuan Kelayakan Kredit Menggunakan Algoritma K-Nn Berbasis Forward Selection Data Mining on Credit Feasibility Determination Using K-Nn Algorithm Based on Forward Selection," *CCIT J.*, vol. 11, no. 2, pp. 236–244, 2018, doi: 10.33050/ccit.v11i2.591.

[13] M. Hasan, "Prediksi Tingkat Kelancaran Pembayaran Kredit Bank Menggunakan Algoritma Naïve Bayes Berbasis Forward Selection," *Ilk. J. Ilm.*, vol. 9, no. 3, pp. 317–324, 2017, doi: 10.33096/ilkom.v9i3.163.317-324.

[14] A. Nurzahputra and M. A. Muslim, "Peningkatan Akurasi Pada Algoritma C4.5 Menggunakan Adaboost Untuk Meminimalkan Resiko Kredit," *Pros. SNATIF*, no. 1, pp. 243–247, 2017, [Online]. Available: https://media.neliti.com/media/publications/173704-ID-none.pdf.

[15] E. Algoritma, C. Dan, F. Feature, P. Gede, S. Cipta, and G. S. Mahendra, "Selection Untuk Menentukan Debitur Baik Dan Debitur Bermasalah Pada Produk Kredit Tanpa Agunan (Kta)," vol. 9, no. 1, pp. 39–46, 2020.

[16] T. Mardiana, "Optimasi Naïve Bayes Dengan Particle Swarm Optimization Dan Stratified Untuk Prediksi Kredit Macet Pada Koperasi," *J. Ris. Inform.*, vol. 1, no. 1, pp. 43–50, 2018, doi: 10.34288/jri.v1i1.13.

[17] A. Syukron and A. Subekti, "Penerapan Metode Random Over-Under Sampling dan Random Forest Untuk Klasifikasi Penilaian Kredit," *J. Inform.*, vol. 5, no. 2, pp. 175–185, 2018, doi: 10.31311/ji.v5i2.4158.

p-ISSN: 2406-7768
e-ISSN: 2581-2181

[18] K. Algoritma, L. D. A. Dan, and Y. Ramdhani, "Naïve Bayes Dengan Optimasi Fitur Untuk Klasifikasi," vol. II, no. 2, pp. 434–441, 2015.

[19] Y. N. Dewi and F. A. Sariasih, "Metode Sample Bootstrapping Untuk Meningkatkan Performa," vol. 12, no. 1, 2019.

[20] A. A. Agustian *et al.*, "Data Mining Optimization Using Sample Bootstrapping and Particle Swarm Optimization in the Credit Approval Classification," vol. 2, no. 1, pp. 18–27, 2019.

[21] D. T. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. New Jersey: John Wiley & Sons, 2005.

[22] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. Verlag Berlin Heidelberg: Springer, 2011.