JUTIS (Jurnal Teknik Informatika Unis)
Vol. 12 No. 1 Bulan November 2024 **p-ISSN: 2252-5351**
Program Studi Teknik Informatika – Fakultas Teknik **e-ISSN: 2656-0860**
Universitas Islam Syekh-Yusuf Tangerang

Prediksi Kelancaran Nasabah Terhadap Pembayaran Angsuran Pinjaman Dengan Menggunakan Metode Algoritma Klasifikasi C4.5

Gikaldi Arbiyan Septuri [1*], Puspita Nurul Sabrina[2], Herdi Ashaury[3]
[1, 2, 3] Informatika, Sains dan Informatika, Universitas Jenderal Achmad Yani,
Cimahi Indonesia, Kode Pos 40513

gikaldiarbiyans18@if.unjani.ac.id[1*] , puspita.sabrina@lecture.unjani.ac.id[2] ,
herdi.ashaury@lecture.unjani.ac.id[3]

Abstrak

Pinjaman adalah salah satu kegiatan terpenting lembaga keuangan. Masalah yang umum terjadi di lembaga keuangan adalah banyak nasabah yang kesulitan membayar cicilan, sehingga perusahaan berpotensi mengalami kerugian. Masalah ini diselesaikan dengan teknik data mining., yaitu menggunakan algoritma C4.5 untuk memprediksi kelancaran nasabah pembayaran terhadap angsuran. Tetapi, Data yang tidak seimbang merupakan tantangan bagi kinerja algoritma C4.5. Keadaan dimana model hanya mampu memprediksi kelas mayoritas. Akibatnya model C4.5 cenderung mengenali data minoritas sebagai data mayoritas. Masalah ini akan diselesaikan dengan metode penyeimbangan data Random Oversampling. Hasil dari penelitian ini mengungkapkan bahwa dengan hanya metode algoritma C4.5 menunjukkan akurasi sebesar 0.88 namun hanya memprediksi kelas mayoritas, Sedangkan penerapan Random Oversampling terbukti algoritma C4.5 berhasil mengenali distribusi kedua kelas dengan akurasi sebesar 0.79 dalam memprediksi kelancaran nasabah dalam membayar angsuran.

Kata kunci: *Angsuran, Algoritma C4.5, Nasabah, Prediksi kelancaran pembayaran, Random Oversampling.*

A. Pendahuluan

Lembaga keuangan adalah suatu usaha yang memiliki aset berupa instrumen keuangan atau tagihan, bukan instrumen non-keuangan. Lembaga-lembaga ini fokus pada peminjaman dan penyaluran uang dalam bentuk bank(Nugraha et al., 2020).Selain itu, lembaga keuangan menawarkan beragam produk layanan keuangan, seperti tabungan, pinjaman, asuransi, skema pensiun, metode pembayaran, transfer dana, dan pinjaman. Lembaga keuangan sering kali menghadapi berbagai tantangan yang terus mempengaruhi operasional mereka. Dari sudut pandang perkreditan, permasalahan yang sama dan penyebabnya terkadang muncul. Penyebab utama permasalahan tersebut bukan pada lemahnya sistem dan regulasi Bank Indonesia, namun pada kualitas sumber daya manusia yang mengelola pinjaman pada lembaga keuangan tersebut.

Sesuai Surat Keputusan Bersama Ulama Perindustrian Uang dan Bursa No.1169/KMK.01/1991 tanggal 21 November 1991 tentang Kegiatan Menyewa, Menyewa adalah suatu organisasi yang mendukung gerakan melalui pemberian modal barang

dagangan untuk diikutsertakan dalam organisasi dalam jangka waktu tertentu. . Dukungan ini bergantung pada angsuran sesekali dan disertai dengan kemungkinan bagi organisasi untuk membeli barang modal atau memperluas jangka waktu sewa mengingat harga yang disepakati. Seperti organisasi keuangan lainnya, persewaan juga menghadapi masalah kredit yang berbeda. Permasalahan yang sering muncul disebabkan oleh pembeli, misalnya pembeli yang dianggap memenuhi syarat namun secara finansial sudah lewat jatuh tempo porsinya, hingga sepeda motornya harus direview oleh pihak organisasi, bahkan ada pula yang dicopot bersama sepeda motor yang dikreditkan. Masalah ini sebagian besar disebabkan oleh kesalahan pemeriksaan pejabat pencatatan (ahli kredit). Petugas akun sering kali lalai menyelesaikan pemeriksaan intermiten dan pengawasan kredit yang memuaskan setelah pelanggan mendapatkan kantor, baik manajemen langsung maupun regulasi (Gumelar et al., 2021).

Pada lembaga keuangan di daerah bandung, data pinjaman uang dengan agunan kendaraan bermotor dapat diperoleh dengan atribut User, Kecamatan, Kelurahan, Tenor, Nomor Cicilan, Sisa Cicilan, Sisa Angsuran, Tenggat Pembayaran, Angsuran, Jumlah Pinjamann Yang Belum Dibayarkan, Bucket Akhir,Kode Aset, Nomor Mesin, Nomor Rangka, Klasifikasi, Masalah, pekerjaan dan Status Bayar.

Dari data yang dikumpulkan, terdapat ketidakseimbangan jumlah data pada atribut Status Bayar yang persentasenya cenderung condong ke nasabah yang sudah membayar cicilan sebesar 88%, sedangkan jumlah data yang belum membayar adalah sebesar 12%, dan pada atribut ini memiliki tipe data yang berbeda dimana data yang sudah membayar dituliskan dengan tipe data numerik dengan jumlah cicilan yang harus dibayarkan (contohnya 2684000) sementara yang belum membayar dituliskan dengan tipe data kategorikal (belum bayar). Karena itu dapat mempengaruhi kelancaran pada proses penelitian, maka perlu dilakukan data balancing dan juga transformasi data untuk meningkatkan akurasi pada penilitian, dimana penelitian ini akan berfokus dalam memprediksi kelancaran terhadap nasabah dalam pembayaran angsuran pinjaman.

Beberapa penelitian terdahulu yang terkait dengan tema penelitian ini diantaranya : "Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm" oleh Wahyu Nugraha, Muhammad Sony Maulana, Agung Sasongko (Nugraha et al., 2020). Penelitian ini bertujuan untuk mengatasi masalah ketidakseimbangan kelas dalam dataset pelatihan dengan menggunakan strategi undersampling. Penelitian ini bertujuan untuk meningkatkan kinerja algoritma C4.5. Hasilnya algoritma C4.5 tanpa menggunakan metode tambahan untuk mengatasi dataset yang tidak seimbang memiliki akurasi yang tinggi tidak bisa mengenali distribusi kelas minoritas dan mengabaikannya, tetapi saat metode untuk menangani dataset tidak seimbang akurasi dari algoritma C4.5 cenderung menurun, namun algoritma C4.5 dapat mengenali distribusi kelas mayoritas maupun minoritas (Nugraha et al., 2020). "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance" oleh Gagah Gumelar, Norlaila , Quratul Ain, Riza Marsuciati, Silvi Agustanti Bambang, Andi Sunyoto, M. Syukri Mustafa (Gumelar et al., 2021). Penelitian ini bertujuan untuk menyeimbangkan dataset yang tidak seimbang dan kemudian mengevaluasi kinerja

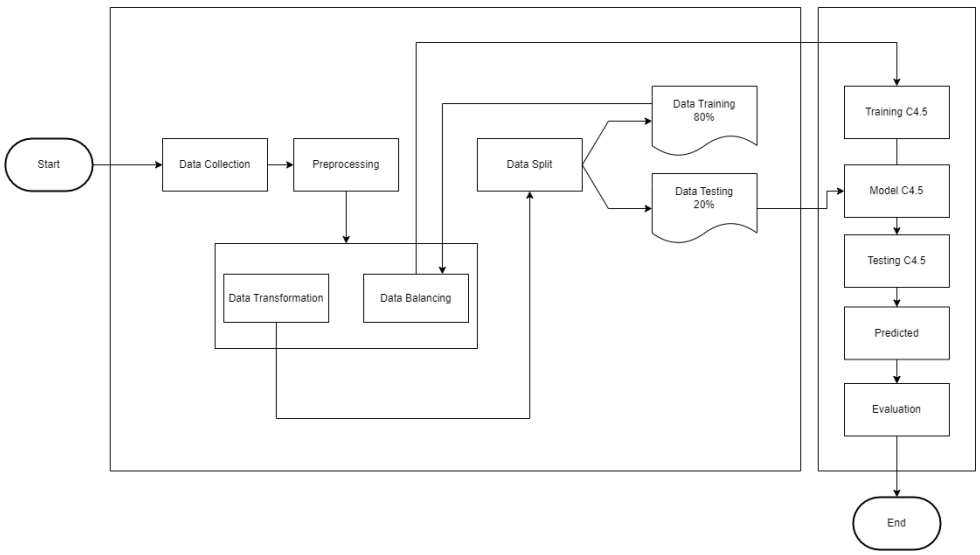
berbagai algoritma klasifikasi (C4.5, Naïve Bayes, KNN, dan SVM) sebelum dan sesudah penggunaan oversampling. Hasil pada penelitian adalah bahwa implementasi resampling smote yang dikombinasikan dengan algoritma klasifikasi dapat meningkatkan akurasi pada algoritma NB sebesar 24%, algoritma KNN sebesar 1%, dan algoritma DT sebesar 2 % (Gumelar et al., 2021). "Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results" oleh Roweida Mohammed, Jumanah Rawashdeh, dan Malak Abdullah(Mohammed et al., 2020). Tujuan dari penelitian ini untuk mengeksplorasi dan membandingkan kinerja dua teknik resampling yang umum digunakan, yaitu oversampling dan undersampling, dalam mengatasi masalah ketidakseimbangan data pada tugas klasifikasi. Hasil dari penelitian ini dapat disimpulkan bahwa teknik oversampling menunjukkan kinerja yang lebih baik dibandingkan dengan teknik undersampling dalam mengatasi masalah ketidakseimbangan kelas pada dataset Santander Customer Transaction Prediction dari Kaggle. Penelitian ini menunjukkan bahwa oversampling memberikan skor yang lebih tinggi dalam berbagai metrik evaluasi (seperti recall, precision, dan accuracy) dibandingkan dengan undersampling ketika diterapkan pada berbagai model klasifikasi machine learning (Mohammed et al., 2020).

Dari hasil kedua penelitian di atas, pernelitian terdahulu terdapat kemiripan topik penelitian dengan studi kasus yang berbeda, dan pada penelitian-penelitian terdahulu dapat diambil kesimpulan bahwa algoritma klasifikasi rentan terhadap dataset dengan distribusi kelas yang tidak seimbang, khususnya metode prediksi menggunakan C4.5 pada penelitian ini.

B. Metode

Perancangan Umum Tahap Penelitian

Pada subbab ini, penulis membahas metode desain yang digunakan dan langkah-langkah yang terlibat dalam penelitian ini. Di bawah ini adalah gambar prosedur penelitian berikut :



Gambar 1. Perancangan Umum Tahap Penelitian

Perolehan Data

Perolehan data nasabah ini diambil dari salah satu perbankan, dengan 18 atribut diantaranya User, Kecamatan, Kelurahan, Tenor, Installment No, Sisa Cicilan, Sisa Angsuran, Tenggat Pembayaran, Angsuran, Jumlah Outstanding Balance, Bucket Akhir,Kode Aset, Nomor Mesin, Nomor Rangka, Klasifikasi, Masalah, Pekerjaan dan Status Bayar. Data akan dilakukan pre-processing. Attribut data yang diperoleh dapat dilihat pada tabel 1.

Tabel 1
Atribut Data

No	Atribut	Deskripsi
1	User	ID nasabah
2	Kecamatan	Alamat kecamatan nasabah
3	Kelurahan	Alamat kelurahan nasabah
4	Tenor	Jangka waktu pinjaman
5	Installment No	Nomor cicilan
6	Sisa Cicilan	Jumlah sisa cicilan yang harus dibayarkan
7	Sisa Angsuran	Sisa angsuran yang dimiliki nasabah
8	Tenggat Pembayaran	Tenggat hari pembayaran
9	Angsuran	Jumlah uang yang harus dibayarkan tiap cicilan
10	Outstanding Balance	Jumlah uang yang belum dibayarkan
11	Bucket Akhir	Pengelompokan jumlah hari keterlambatan pembayaran
12	Kode Aset	Kode aset kendaraan
13	Nomor Mesin	Nomor mesin kendaraan
14	Nomor Rangka	Nomor rangka kendaraan
15	Klasifikasi	Pengelompokan ketersediaan nasabah dan aset nasabah
16	Masalah	Kendala nasabah terhadap angsuran
17	Pekerjaan	Pekerjaan nasabah
18	Status Bayar	Status nasabah terhadap pembayaran

Tahapan Pre-processing

Tahap pre-processing dilakukan untuk memahami isi dataset yang diperoleh. Dalam tahap ini akan di seleksi atribut yang sesuai dengan kebutuhan, mengubah data agar bisa diolah oleh model machine learning dengan cara menangani data yang tidak lengkap, data yang kosong, data yang format tulisannya tidak konsisten, data yang tipe data nya tidak sesuai, menambahkan atribut baru yang sesuai dengan kebutuhan.

Data Selection

Dalam tahapan data selection ini, seleksi data yang dilakukan adalah penghapusan atribut – atribut yang tidak digunakan dalam analisis seperti user, kecamatan, kelurahan, tenor, sisa cicilan, tenggat pembayaran, angsuran, outstanding balance, bucket akhir, kode asset, nomor mesin, nomor rangka, klasifikasi, masalah, pekerjaan, status bayar, kode kecamatan, kode klasifikasi, kode kategori, tahun, dan user encoded menggunakan teknik Recursive Feature Elimination. Atribut – atribut tersebut tidak memiliki korelasi yang kuat dengan atribut lainnya. Sehingga atribut tersebut tidak diperlukan pada penelitian ini.

Sehingga atribut yang digunakan pada penelitian ini diantaranya adalah kode_klasifikasi, bayar_kode, angsuran_kategori, OSBalance_kategori, Tenor_Kategori dan tanggal_kategori. Dari jumlah 2301 data record dan 16 atribut yang terdapat dalam dataset didapatkan hanya 6 atribut dan 2301 record. Hasil dari selection dapat dilihat pada tabel Gambar 2.

Kode_Klasifikasi	bayar_kode	Angsuran_kategori	OSBalance_kategori	Tenor_kategori	tanggal_kategori
1	0	3	3	2	3
1	1	1	2	2	2
1	1	1	2	2	1
1	0	2	3	3	3
1	1	3	3	2	1
1	0	3	3	3	2
1	0	3	3	2	1
1	0	3	3	3	3
1	0	2	3	3	3
1	0	1	2	2	3
1	0	3	3	3	3
1	1	3	3	3	3
1	0	3	3	2	2
1	1	1	3	3	3
1	0	3	3	2	3
1	1	1	2	2	1
1	0	1	2	2	3
1	0	3	3	2	1
1	1	3	3	2	3

Gambar 2. Atribut Transformation

Data Transformation

Transformasi data dilakukan untuk merubah data menjadi kategorik, diantaranya dilakukan pada beberapa atribut. Atribut yang ditranformasi antara lain kecamatan, klasifikasi, pekerjaan, bayar, bucket akhir, tahun, bulan, tanggal, user menggunakan teknik Label Encoding, dan Ordinal Encoding untuk mengubah data kategorikal menjadi numerik, dan teknik Interquartile untuk mengukur sebaran data dengan cara menghitung rentang antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Hasil transformasi dapat dilihat pada tabel dibawah, pembagian menjadi tiga kelompok untuk mengurutkan nilai dengan mengelompokan kategori Rendah, Sedang, Tinggi untuk menghilangkan outlier pada dataset.

Tabel 2
Transformasi Atribut Klasifikasi

Klasifikasi => Kode Klasifikasi	
Sebelum Transformasi	**Sesudah Transformasi**
Nasabah Ada Unit Ada	1
Nasabah Ada Unit Tidak Ada	2
Nasabah Tidak Ada Unit Ada	3
Nasabah dan Unit Tidak Ada	4

Tabel 3
Transformasi Atribut OSBalance

OSBalance => OSBalance Kategori		Keterangan
Sebelum Transformasi	**Sesudah Transformasi**	
0 - 13571999	1	Rendah
13572000 - 27143999	2	Sedang
27144000 - 40716000	3	Tinggi

Train/Test Split

Pada proses ini digunakan metode yang dapat mengevaluasi performa model machine learning. Metode evaluasi model ini membagi dataset menjadi dua bagian yakni bagian yang digunakan untuk training data dan untuk testing data dengan proporsi tertentu. Train data digunakan untuk fit model machine learning, sedangkan test data digunakan untuk mengevaluasi hasil fit model tersebut. Dataset yang telah diperoleh sebelumnya dibagi menjadi 2 data latih sebanyak 80% dan data uji sebanyak 20%.

Handling Data Imbalance

Pada tahap ini, data minoritas akan diperlakukan dengan teknik Random Oversampling, yang merupakan metode oversampling untuk menyeimbangkan distribusi dataset dengan cara menduplikasi data minoritas secara acak sehingga data minoritas hingga jumlahnya seimbang dengan data mayoritas (Wongvorachan et al., 2023).

Hyperparameter

Hyperparameter adalah variabel yang tidak dihitung secara otomatis oleh model selama proses pelatihan, tetapi harus ditentukan sebelumnya oleh pengguna. Dalam domain machine learning dan deep learning, hyperparameter memainkan peran kunci dalam menentukan cara model belajar dan beroperasi. Menemukan pengaturan hyperparameter yang optimal adalah kunci untuk mencapai kinerja model yang terbaik. Proses ini sering melibatkan eksperimen dan teknik validasi silang untuk menemukan kombinasi hyperparameter yang optimal untuk data yang belum pernah terlihat sebelumnya (Nasrullah et al., 2024)

Prediksi dengan metode C4.5

Pada langkah ini merupakan proses perhitungan C4.5. Dengan menentukan atribut yang akan digunakan untuk melakukan proses perhitungan dengan data yang digunakan adalah data yang melewati tahap preprocessing yaitu 2301 dataset.

Tabel 4
Penerapan C4.5

Kriteria		Bayar	Belum	Jumlah	Entropy	Gain	Split Info	Gain Ratio
Total		2029	272	2301	0,524192			
Kode Klasifikasi	1	1953	269	2222	0,532406	0,003445	0,244457	0,014093969
	2	63	1	64	0,116115			
	3	3	0	3	0			
	4	10	2	12	0,650022			
Obstacle Kategori	1	495	80	575	0,581957	0,001054	1,499783	0,0007026
	2	508	67	575	0,519279			
	3	1026	125	1151	0,495682			
Sisa angsuran kategori	1	444	86	530	0,639699	0,003778	1,411969	0,002675776
	2	409	47	456	0,478653			
	3	1176	139	1315	0,486819			
Tenor kategori	1	1	0	1	0	0,000275	0,950602	0,000289093
	2	743	93	836	0,503657			
	3	1285	179	1464	0,535845			
Tanggal Kategori	1	510	63	573	0,499758	0,000159	1,484593	0,000107213
	2	478	66	544	0,533153			
	3	1041	143	1184	0,53159			

Pada iterasi ke-1 diperoleh gain ratio terbesar adalah 0,002675776 yaitu kriteria "Sisa Angsuran Kategori". Maka Sisa Angsuran Kategori merupakan akar pohon keputusan (root). Dilakukan proses perhitungan untuk node berikutnya sampai proses perhitungan berakhir dengan setiap cabang pohon keputusan sudah memiliki kelas yang sama. Pohon keputusan yang akan menjadi aturan-aturan keputusan (rule) (Situmeang et al., 2022).

Evaluasi Confusion Matrix logistic regression

Serangkaian teknik yang digunakan untuk mengukur kinerja model Logistic Regression dalam memprediksi kategori biner atau mengklasifikasikan data ke dalam dua kelas salah satunya mengunkan Confusion Matrix adalah salah satu cara untuk mengukur kinerja model klasifikasi. Confusion matrix adalah tabel yang digunakan untuk menganalisis kinerja model berdasarkan jumlah data yang diklasifikasikan dengan benar dan salah.

C. Hasil dan Pembahasan

Implementasi

Tahap Implementasi merupakan langkah krusial di mana algoritma yang telah dirancang diimplementasikan dalam sistem, berdasarkan analisis mendalam dan perancangan yang telah disusun dengan pendekatan fungsional. Pada fase ini, sistem dikembangkan dengan menerjemahkan rancangan detail menjadi kode Python. Proses ini melibatkan transformasi spesifikasi teknis dan rancangan mendetail ke dalam kode yang dapat dieksekusi oleh komputer, memastikan bahwa semua fitur dan

p-ISSN: 2252-5351
e-ISSN: 2656-0860

fungsionalitas yang direncanakan dapat beroperasi sesuai dengan tujuan yang ditetapkan.

Hasil Implementasi

Proses menerapkan teknik dan algoritma data mining yang telah dirancang untuk menganalisis data dan mengungkap pola, tren, atau pengetahuan yang terkait dengan pertanyaan penelitian yang sedang diteliti berikut adalah hasil implementasi.

Hasil Implementasi Label Encoder, dan Ordinal Encoding

Dari data yang didapat sebelumnya harus ditransformasi kedalam tipe data numerik sebelum dilatih ke dalam metode C4.5.

le_Kecama	le_Klasifik	kerjaan_k	ayar_kod	de_kategc	tahun	bulan	tanggal
24	1	0	0	3	2023	7	25
12	1	0	1	3	2023	11	12
22	1	0	1	3	2023	2	1
70	1	0	0	2	2023	8	18
109	1	0	1	3	2023	6	2
27	1	0	0	1	2023	3	9
47	1	0	0	3	2023	10	7
60	1	0	0	3	2023	4	26
97	1	0	0	3	2023	1	24
27	1	0	0	3	2023	5	24
54	1	0	0	3	2023	3	30
62	1	0	1	3	2023	8	28
46	1	0	0	1	2023	10	14
7	1	0	1	3	2023	12	26
62	1	0	0	3	2023	5	16
91	1	0	1	3	2023	9	8
16	1	0	0	3	2023	2	21
22	1	0	0	3	2023	4	8
49	1	0	1	1	2023	7	18
83	1	0	1	2	2023	7	2

Gambar 3. Hasil Implementasi Label Encoder

Hasil Implementasi Interquartile Range

Dari data yang didapat sebelumnya harus dikelompokan menjadi 3 kelompok mennggunakann teknik Interquartile Range dengan menentukan Q1, Mean, dan Q3 pada atribut yang akan dikelompokkan.

Angsuran_kategori	OSBalance_kategori	InstallmentNo_kategori	sisa angsuran_kategor	Tenor_kategori	bulan_kategori	tanggal_kategori
3	3	1	3	2	1	3
1	2	1	3	2	1	2
1	2	1	3	2	1	1
2	3	1	3	3	1	3
3	3	1	3	2	1	1
3	3	1	3	3	1	2
3	3	1	3	2	1	1
3	3	1	3	3	1	3
2	3	1	3	3	1	3
1	2	1	3	2	1	3
3	3	1	3	3	1	3
3	3	1	3	3	1	3
3	3	1	3	2	1	2
1	3	1	3	3	1	3
3	3	1	3	2	1	3
1	2	1	3	2	1	1
1	2	1	3	2	1	3
3	3	1	3	2	1	1
3	3	1	3	2	1	3

Gambar 4. Hasil Implementasi Interquartile Range

Hasil Implementasi Data Selection

Dari data yang didapat sebelumnya, data tersebut harus di Selection untuk mengurangi kompleksitas dan volume data yang ada sehingga fokus hanya pada data yang penting. Sehingga atribut – atribut yang digunakan untuk analisis prediksi adalah kode_klasifikasi, bayar_kode, angsuran_kategori, OSBalance_kategori, Tenor_Kategori dan tanggal_kategori menggunakan teknik Recursive Feature Elimination.

```
#tidak usah
data_A=['Kode_Klasifikasi','pekerjaan_kode','kode_kategori','Angsuran_kategori','OSBalance_kategori','InstallmentNo_kategori','sisa angsu
#data_x=['InstallmentNo','angsuran','OSBalance','Kode_Klasifikasi','pekerjaan_kode','bulan','tanggal'] #rekomedasi +2
#data_A=['InstallmentNo','angsuran','OSBalance','Kode_Klasifikasi','pekerjaan_kode','tanggal']
data_B=['bayar_kode']
A = df[data_A]

B = df[data_B]

A_train, A_test, B_train, B_test = train_test_split(A, B, test_size = 0.2)
UT_FRE =tree.DecisionTreeClassifier(max_depth=10, min_samples_split=10, criterion="entropy",random_state=42,max_leaf_nodes=100)
UT_FRE.fit(A_train, B_train)
rfe = fs.RFE(UT_FRE)
rfe.fit(A,B)
print(f'Support = {rfe.support_}')
print(f'Ranking = {rfe.ranking_}')
#sebelum ['Kode_Klasifikasi1','pekerjaan_kode','kode_kategori','Angsuran_kategori','OSBalance_kategori1','InstallmentNo_kategori','sisa a
#sesudah ['Kode_Klasifikasi','OSBalance_kategori','sisa angsuran_kategori','Tenor_kategori','tanggal_kategori']
```

```
Support = [ True False False False  True False  True  True  True False]
Ranking = [1 4 2 3 1 5 1 1 1 6]
```

Gambar 5. Hasil Implementasi Data Selection

Hasil Implementasi Train/Test Split

Proses membagi data latih(train) dan data uji(test) pada mesin pembelajaran dengan pembagian 80% untuk data latih(test), dan 20% untuk data latih(test).

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 42, stratify = y)
print(f'jumlah dari dimensi X_train, {len(X_train)}')
print(f'jumlah dari dimensi y_train, {len(y_train)}')
print(f'jumlah dari dimensi X_test, {len(X_test)}')
print(f'jumlah dari dimensi y_test, {len(y_test)}')
```

```
jumlah dari dimensi X_train, 1840
jumlah dari dimensi y_train, 1840
jumlah dari dimensi X_test, 461
jumlah dari dimensi y_test, 461
```

Gambar 6. Hasil Implementasi Train /Test Split

Hasil Implementasi Handling Data Imbalance

Dari data yang didapat sebelumnya point 3.2.4 dikarenakan terjadi ketidakseimbangan kelas pada data latih dimana kelas mayoritas (0) memiliki data record = 1622. Dan kelas minoritas (1) memiliki data record = 218, dengan menggunakan teknik Random Oversamplig label mayoritas (0), dan minoritas (1) memiliki jumlah data record yang sama.

```
[27]: #RandomOverSampling
      ros = RandomOverSampler(sampling_strategy=1.0, random_state=42) # String
      x_train_ros, y_train_ros = ros.fit_resample(X_train, y_train)

[28]: y_train_ros.value_counts()

[28]: bayar_kode
      0            1622
      1            1622
      Name: count, dtype: int64
```

Gambar 7. Hasil Implementasi Handling Data Imbalance

Hasil Implementasi Hyperparameter

Teknik pengoptimalisasi mesin yang digunakan pada penelitian ini adalah Hyperparameter Grid yang bertugas untuk mengoptimalkan pengaturan atau konfigurasi dari algoritma model statistik untuk menemukan pola terbaik, pada konfigurasi terbaik yang dihasilkan menggunakan teknik ini adalah :

'classifier__criterion': 'entropy', 'classifier__max_depth': 10, 'classifier__max_leaf_nodes': 15, 'classifier__min_samples_split': 2

Hasil Implementasi C4.5

Metode yang digunakan untuk prediksi menggunakann C4.5 dengan menggunakan dataset yang sebelumnya sudah melalui pre-processing data untuk menentukan nilai prediksi dengan konfigurasi terbaik seperti berikut:

(criterion="entropy", max_depth=10 min_samples_split= 2, random_state=42,max_leaf_nodes=15)

Hasil Evaluasi Confusion Matrix

Berikut adalah hasil Confusion Matrix dari metode C4.5

```
Confusion matrix

 [[353  54]
 [ 44  10]]

True Positives(TP) =  353

True Negatives(TN) =  10

False Positives(FP) =  54

False Negatives(FN) =  44
```

Gambar 8. Hasil Evaluasi Confusion Matrix

Dari informasi tersebut dapat dijelaskan dengan perhitungan :

1. Precision (Presisi): Persentase data positif yang diklasifikasikan dengan benar dari semua data yang diklasifikasikan sebagai positif. Rumus: TP / (TP + FP),

p-ISSN: 2252-5351
e-ISSN: 2656-0860

pada Confusion matrix yang telah diberikan maka akan mendapatkan hasil sebagai berikut

Precision = 353 / (353 + 54) = 0.8673

2. Accuracy (Akurasi): Persentase data yang diklasifikasikan dengan benar dari keseluruhan data. Rumus: (TP + TN) / (TP + TN + FP + FN) Confusion matrix yang telah diberikan maka akan mendapatkan hasil sebagai berikut

   Accuracy (353 + 10) / (353 + 10 + 54 + 44) = 0.7874

3. Recall :seberapa baik model dapat mengidentifikasi semua data positif yang sebenarnya. Rumus: Recall = TP / (TP + FN), Confusion matrix yang telah diberikan maka akan mendapatkan hasil sebagi berikut

   Recall = 353 / (353+ 44) = 0.8891

4. F1-Score: Menggabungkan precision dan recall dalam satu metrik. Rumus: 2 * (Precision * Recall) / (Precision + Recall), Confusion matrix yang telah diberikan maka akan mendapatkan hasil sebagi berikut

   F1-Score = 2 * (Precision * Recall) / (Precision + Recall)

   F1-Score = 2 * (0.7813 * 0.8983) / (0.7813 + 0.8983) = 0.8781

Evaluasi Hasil

Evaluasi hasil dari perbandingan metode-metode yang digunakan pada saat pre-processing data, seperti Imbalance Data Handling, Feature Selection, dan Hyperparameter, untuk mengetahui hasil akurasi dan f1-score terbaik. Berikut adalah tabel perbandingan akurasi dari prediksi menggunakan C4.5 dengan menggunakan metode-metode balancing data, dan hyperparameter tuning :

Tabel 5
Perbandingan akurasi dari prediksi

NO	Metode Tambahan	Label	Precision	Recall	F1-score	Accuracy
1	None	0	0.88	1.00	0.94	0.88
		1	0.00	0.00	0.00	
2	RFE, SMOTE	0	0.89	0.83	0.86	0.76
		1	0.16	0.24	0.19	
3	RFE, ROS	0	0.89	0.87	0.88	0.79
		1	0.16	0.19	0.17	
4	RFE, RUS	0	0.89	0.78	0.83	0.72
		1	0.14	0.28	0.19	

Berdasarkan Tabel 4 dapat disimpulkan bahwa hasil dari prediksi terbaik adalah menggunakan algoritma C4.5 dengan metode Feature Selection untuk memilih dan menghapus atribut yang kurang relevan sehingga atribut yang digunakan adalah

"kode_klasifikasi", "bayar_kode", "angsuran_kategori", "OSBalance_kategori", "Tenor_Kategori" dan "tanggal_kategori", dengan teknik untuk mengangani imbalance data menggunakan Random Oversampling untuk menangani dataset yang tidak seimbang sehingga menghasilkan data latih yang seimbang dengan label 0 (sudah membayar) sebanyak 1622 data, dan label 1 (belum bayar) sebanyak 1622 , yang menghasilkan akurasi sebesar 0.79 dengan F1 Score label 0 sebesar 0.88, dan label 1 sebesar 0.17.

D. Kesimpulan

Berdasarkan hasil evaluasi tersebut didapatkan metode yang paling optimal untuk penelitian ini adalah menggunakan metode Random Oversampling, Feature Selection, yang menghasilkan akurasi sebesar 0.79, dengan F1-score untuk label 0 sebesar 0.88, dan label 1 sebesar 0.17.

Dari pernyataan di atas dapat disimpulkan bahwa metode Random Oversampling untuk menangani data imbalance pada data nasabah memberikan hasil terbaik sehingga algoritma C4.5 dapat mengenali distribusi kelas mayoritas dan mayoritas pada prediksi kelancaran pembayaraan dengan akurasi meningkat menjadi 0.79, dengan rasio data latih 80% dan data uji 20%.

Daftar Pustaka

Gumelar, G., Ain, Q., Marsuciati, R., Agustanti Bambang, S., Sunyoto, A., & Syukri Mustafa, M. (2021). Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance. *SISFOTEK : Sistem Informasi Dan Teknologi*, 250–255.

Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. *2020 11th International Conference on Information and Communication Systems, ICICS 2020*, 243–248. https://doi.org/10.1109/ICICS49469.2020.239556

Nasrullah, M. F., Saedudin, R. R., & Hamami, F. (2024). Perbandingan Akurasi Algoritma C4.5 Dan. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *9*(2), 628–638.

Nugraha, W., Maulana, M. S., & Sasongko, A. (2020). Clustering Based Undersampling for Handling Class Imbalance in C4.5 Classification Algorithm. *Journal of Physics: Conference Series*, *1641*(1). https://doi.org/10.1088/1742-6596/1641/1/012014

Situmeang, N., Jaya, I. K., & Yohanna, M. (2022). Penerapan Metode Decision Tree C4.5 Dalam Memprediksi Kelancaran Pembayaran Kredit Di Bpr Wahana Bersama Kpum. *METHOMIKA Jurnal Manajemen Informatika Dan Komputerisasi Akuntansi*, *6*(6), 215–220. https://doi.org/10.46880/jmika.vol6no2.pp215-220

Wongvorachan, T., He, S., & Bulut, O. (2023). A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information (Switzerland)*, *14*(1). https://doi.org/10.3390/info14010054