IMPLEMENTASI METODE NAIVE BAYES UNTUK KLASIFIKASI GOLONGAN PEMBAYARAN SYAHRIYAH DI PONDOK PESANTREN DARUSSAADAH

Aprillian Putra Pratama[1)], Eka Yuniar[2)], Weda Adistianaya Dewa[3)]
Program Studi S1-Sistem Informasi, STMIK PPKIA Pradnya Paramita[1)]
email: aprillian_20510001@stimata.ac.id
Program Studi S1-Sistem Informasi, STMIK PPKIA Pradnya Paramita[2)]
email: eka@stimata.ac.id
Program Studi S1-Sistem Informasi, STMIK PPKIA Pradnya Paramita[3)]
email: weda@stimata.ac.id

Abstract

Currently, Pondok Pesantren Darussa'adah continues to utilize a traditional approach in managing its operational systems, particularly in the administration of syahriyah (monthly tuition) payments. The payment system at this pesantren implements a cross-subsidy mechanism, where students from more economically advantaged backgrounds help subsidize those from less privileged families. However, the classification process for syahriyah payment tiers remains manual and is based on a form that evaluates Monthly Income, Number of Dependents, and Outstanding Loans. This form is submitted alongside the new student registration form. Such a conventional method is considered inefficient and prone to human error in the classification process.

Therefore, this study aims to classify syahriyah payment categories using the Naive Bayes method. A dataset comprising 1,000 student records from Pondok Pesantren Darussa'adah was utilized, with 70% allocated for training and 30% for testing. The study evaluates and compares the classification performance of four different models employing the same Naive Bayes technique. The best-performing model achieved an overall accuracy of 96%, with precision and recall also at 96%. Class-wise accuracy results indicate 100% accuracy for the "Capable" category, 96% for the "Middle" category, and 93% for the "Underprivileged" category.

Keywords: *Classification, Payment, Data Mining, Naive Bayes Classifier (NBC), Synthetic Minority Over-Sampling Technique (SMOTE)*

1. PENDAHULUAN

Pondok Pesantren Darussa'adah Gubugklakah merupakan salah satu lembaga pendidikan islam yang terletak di Desa Gubugklakah, Kecamatan Poncokusumo, Kabupaten Malang. Pondok Pesantren ini didirikan pada tahun 1991 oleh Ust. Nur Hasanuddin. Saat ini terdapat beberapa instansi pendidikan di bawah naungan Pondok Pesantren Darussa'adah, antara lain; MTs Darussa'adah, SMP-BP Darussa'adah, MA Darussa'adah, SMK Darussa'adah, dan Madrasah Diniyah Darussa'adah. Terhitung sejak Juni 2023, Pondok Pesantren ini memiliki santri sebanyak 1.200 orang, serta 100 orang pengajar dengan 70% pengajar lulusan Yaman.

Saat ini, Pondok Pesantren Darussa'adah masih menggunakan sistem tradisional dalam operasional manajemen lembaga secara keseluruhan, salah satunya adalah bidang pembayaran *syahriyah* (bulanan). Sistem pembayaran *syahriyah* di Pondok Pesantren Darussa'adah Gubugklakah dapat dikatakan menganut sistem subsidi silang. Santri dengan kemampuan perekonomian yang baik memberikan subsidi kepada santri dengan kemampuan perekonomian yang kurang. Namun, sistem pengkalsifikasian pembayaran *syahriyah* di Pondok Pesantren Darussa'adah Gubugklakah masih dilakukan menggunakan cara konvensional, dengan meninjau formulir golongan pembayaran yang dibuktikan dengan Penghasilan Bulanan, Jumlah Tanggungan, Total Aset, Kredit/Cicilan, kemudian dikumpulkan bersama dengan formulir pendaftaran santri baru. Hal ini tentu sangat tidak efektif dikarenakan klasifikasi di tiap golongan *syahriyah* tidak memiliki variabel

untuk menentukan golongan pembayaran secara spesifik.

Selama ini Pondok Pesantren Darussa'adah Gubugklakah menentukan tiga golongan pembayaran *syahriyah* yang berbeda, yaitu golongan I (Rp. 300.000) untuk wali santri dengan perekonomian rendah, golongan II (Rp. 400.000) untuk wali santri dengan perekonomian menengah, dan golongan III (Rp. 500.000) untuk wali santri dengan perekonomian atas. Namun proses pengklasifikasian golongan pembayaran *syahriyah* secara konvensional cenderung tidak efektif dan hasil pengklasifikasian kurang akurat karena memungkinkan terjadinya kesalahan.

Untuk itu, penelitian ini bertujuan untuk mengoptimalkan pengklasifikasian golongan pembayaran syahriyah di Pondok Pesantren Darussa'adah menggunakan metode Naïve Bayes. Dengan mengklasifikasikasikan santri menjadi Mampu, Menengah, dan Tidak Mampu, diharapkan dapat meningkatkan keakuratan dan meminimalisir terjadinya kesalahan pada proses pengklasifikasian yang sebelumnya dilakukan secara tradisional.

2. KAJIAN LITERATUR

2.1. Data Mining

Data mining merupakan proses penggalian informasi tersembunyi dari sejumlah besar data untuk menemukan pola, hubungan, atau pengetahuan baru yang dapat digunakan dalam pengambilan keputusan. Menurut Utomo dan Mesran (2020), data mining adalah analisis terhadap kumpulan data untuk menemukan hubungan yang tidak terduga dan menyederhanakan data menjadi bentuk yang lebih mudah dipahami serta bermanfaat.

2.1.1. Tahapan Data Mining

Pada proses data mining, terdapat beberapa tahap penting, antara lain;

a. Pembersihan data *(data cleaning)*: adalah proses menghilangkan nilai yang hilang, *noise*, dan data yang tidak konsisten.
b. Integrasi data *(data integration)*: merupakan tahap penggabungan data yang bertujuan untuk memindahkan seluruh data yang telah dibersihkan ke dalam satu tabel.
c. Transformasi data *(data tranformation)*: adalah proses yang dilakukan untuk mengkonversi data ke format atau label yang diinginkan.
d. Proses mining: merupakan proses utama penerapan metode untuk menemukan informasi yang tersembunyi dalam data.
e. Evaluasi pola: langkah ini merupakan pengidentifikasian pola-pola menarik yang ditemukan. Evaluasi ini bertujuan untuk memeriksa apakah temuan yang dihasilkan sesuai dengan harapan atau tidak.
f. Presentasi pengetahuan: merupakan proses visualisasi dan representasi informasi dari penelitian.

2.2. Klasifikasi

Klasifikasi merupakan proses analisis data yang digunakan untuk menentukan label kelas dari suatu sampel data yang akan diklasifikasikan (Srirahayu & Pribadie, 2023).

Proses klasifikasi tidak hanya bertujuan untuk prediksi, tetapi juga untuk memberikan dasar bagi pengambilan keputusan yang lebih informasional. Dengan memahami pola-pola yang ada dalam data, model klasifikasi dapat digunakan untuk segmentasi dan pengelompokan data, identifikasi keterkaitan yang mungkin tidak terlihat secara langsung, dan optimasi proses bisnis dengan memberikan wawasan tentang faktor-faktor yang paling berpengaruh. Selain itu, klasifikasi membantu dalam meramalkan perilaku masa depan dan menyaring informasi yang relevan, memungkinkan fokus pada aspek-aspek khusus dari data.

Melalui tujuannya untuk meningkatkan akurasi dan efisiensi pengambilan keputusan, klasifikasi menjadi alat yang sangat penting dalam mendukung analisis data dan pengambilan keputusan di berbagai bidang, seperti pendidikan, bisnis, kesehatan, dan pemasaran.

2.3. Algoritma Naïve Bayes

Metode *Naive Bayes* adalah algoritma klasifikasi yang berdasarkan pada Teorema Bayes. Metode ini bekerja melalui suatu pendekatan klasifikasi yang dimanfaatkan untuk memproyeksikan kelas suatu entitas berdasarkan serangkaian atribut yang akan dianalisis (Suarpurningsih, Utami, & Estiyanti, 2022).

Dengan mengacu pada Teorema Bayes, metode ini mengasumsikan independensi antar atribut, di mana setiap atribut dianggap tidak

memiliki ketergantungan yang signifikan terhadap nilai-nilai yang diberikan pada variabel kelas. Pendekatan ini memerlukan data latihan yang relatif kecil untuk menentukan estimasi parameter yang diperlukan dalam proses klasifikasi (Fauziah & Dana, 2023).

Teorema bayes memiliki bentuk umum sebagai berikut:

$$P(H|X) = (P(X \mid H)P(H))/(P(X))$$

Keterangan :

X = Data dengan kelas yang belum diketahui

H = Hipotesis data X

P(H|X) = Probabilitas hipotesis H berdasarkan kondisi X

P(H) = Probabilitas hipotesis H

P(X|H) = Probabilitas X berdasarkan kondisi H

P(X) = Probabilitas dari X

Klasifikasi menggunakan metode naïve bayes melalui tiga proses klasifikasi yaitu fase pelabelan, fase *Pre-Processing*, dan fase *Processing*.

2.3.1. *Pre-Processing*

Proses *Pre-Processing* Data merupakan serangkaian langkah yang bertujuan untuk melakukan pembersihan data dengan maksud untuk menstandardisasi bentuk kata dan mengurangi volume kata (Yusra, Olivita, & Fitriani , 2016). *Pre-Processing* memiliki beberapa tahap, namun, sesuai data yang telah tersedia, penelitian ini hanya membutuhkan *Encoding*, Pelabelan dan *Synthetic Minority Over-Sampling Technique (SMOTE)* pada fase *Pre-Processing*.

a. Encoding: adalah proses mengonversi data mentah atau data dalam format yang sulit diolah menjadi format yang lebih cocok untuk analisis atau pemodelan.
b. Pelabelan: merupakan tahapan pemberian tag atau label pada data yang belum diolah, sehingga dapat diklasifikasikan ke dalam kelas tertentu. Dalam penelitian ini pelabelan yang di gunakan yaitu Mampu, Menengah, dan Tidak Mampu.
c. *Synthetic Minority Over-Sampling Technique (SMOTE):* adalah metode yang digunakan untuk melakukan *over-sampling* pada kelas minoritas dengan menghasilkan data sintetis.

2.3.2. *Processing*

Pada fase ini data akan dibagi menjadi dua bagian *(splitting data)*, yaitu data pelatihan *(training data)* dan data pengujian *(testing data)*.

a. *Splitting data*: pembagian data menjadi dua subset, yaitu data pelatihan *(training)* dan data pengujian *(testing)*.
b. *Training data*: adalah data yang digunakan untuk melatih model atau algoritma pembelajaran mesin *(machine learning)* agar dapat mengenali pola, relasi, atau struktur tertentu di dalam data.
c. *Testing data*: merupakan data yang digunakan untuk menguji dan mengevaluasi kinerja model yang telah dibangun menggunakan *training data*.

2.4. Jupyter Notebook

Jupyter merupakan singkatan yang terbentuk dari tiga bahasa pemrograman, yaitu Julia (Ju), Python (Py), dan R. Jupyter adalah suatu aplikasi web yang tersedia secara gratis dan paling umum digunakan oleh para ilmuwan data. Aplikasi ini dirancang untuk membuat dan berbagi dokumen yang menggabungkan kode pemrograman, hasil perhitungan, visualisasi, dan teks. Keberadaan ketiga bahasa pemrograman pada Jupyter memiliki peranan penting dalam lingkungan kerja seorang ilmuwan data. Penggunaan Jupyter sebagai platform analisis data dengan bahasa pemrograman Python dalam penelitian ini merupakan pilihan yang tepat. Jupyter memiliki keunggulan dalam kemampuannya menjalankan kode secara per baris, sehingga memudahkan identifikasi dan perbaikan kesalahan dengan hanya memodifikasi satu baris kode yang dianggap bermasalah (Oktavian & Budi, 2020).

3. METODE PENELITIAN

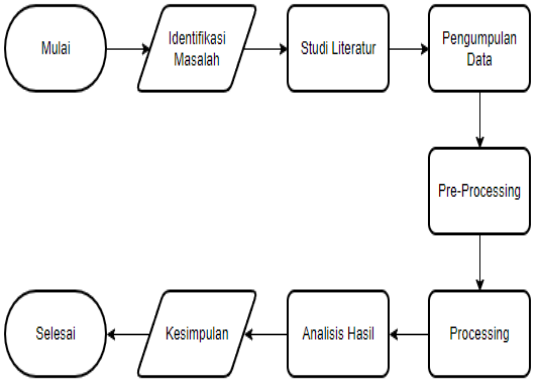
Penelitian ini bersifat kuantitatif menggunakan metode Naive Bayes. Penelitian ini juga akan membandingkan hasil akurasi dari beberapa pengklasifikasian. Pengklasifikasian yang dimaksud dapat dilihat pada Tabel berikut.

Tabel 1 Model pengklasifikasian yang akan dilakukan

Model	Encoding	SMOTE
I	Ya	Ya
II	Ya	Tidak
III	Tidak	Ya
IV	Tidak	Tidak

Tabel diatas menunjukkan bahwa pada penelitian ini akan dilakukan empat model pengklasifikasian yang berbeda. Model pertama akan menerapkan *Encoding* dan *SMOTE*, model kedua akan menerapkan *Encoding* dan tanpa menggunakan *SMOTE*, model ketiga tidak menerapkan *Encoding* dan menerapkan *SMOTE*, dan model keempat tidak menerapkan *Encoding* dan *SMOTE*.

Penelitian ini akan dilakukan dengan beberapa tahapan sebagai berikut.



Gambar 1 Tahapan Penelitian

a. Identifikasi Masalah

Berdasarkan hasil observasi yang telah dilakukan di Pondok pesantren Darussa'adah Gubugklakah, diketahui bahwa permasalahan yang terjadi adalah belum optimalnya pengklasifikasian golongan pembayaran *syahriyah* yang masih dilakukan dengan cara tradisional. Hal ini menyebabkan seringnya terjadi kesalahan bahkan kecurangan dalam proses pengklasifikasian pembayaran *syahriyah*.

b. Studi Literatur

Penelitian ini menggunakan metode pengumpulan data dengan cara memperoleh data dari pihak manajemen keuangan pesantren, membaca, mencatat, dan mengolah bahan penelitian dari jurnal dan buku yang telah digunakan atau dibaca.

c. Pengumpulan Data

Penelitian ini menggunakan data santri yang didapat dari Pondok Pesantren Darussa'adah. Data yang didapat berupa Nama Santri, Nama Wali Santri, Penghasilan Bulanan, Jumlah Tanggungan, dan Kredit/Cicilan. Penelitian ini menggunakan kolom Penghasilan Bulanan, Jumlah Tanggungan, dan Kredit/Cicilan sebagai acuan dalam pemberian label.

d. Pre-Processing

Pada penelitian ini, tahap *Pre-Processing* akan dilakukan *encoding*, pelabelan dan *Synthetic Minority Over-Sampling Technique (SMOTE)*.

1. Encoding

Pada penelitian ini *encoding* akan dilakukan pada pengklasifikasian Model I dan II. Kolom Penghasilan 0 sampai 2.000.000 akan diberi kode 1, penghasilan diatas 2.000.000 dan sampai 4.000.000 akan diberi kode 2, dan penghasilan diatas 4.000.000 akan diberi label 3. Untuk kolom Jumlah Tanggungan lebih besar sama dengan 5 akan diberi kode 1, jumlah tanggungan 3 sampai 4 akan diberi kode 2, dan jumlah tangungan dibawah sama dengan 3 akan diberi kode 3. Kemudian kolom Kredit/Cicilan diatas 3.000.000 akan diberi kode 1, kredit dan utang diatas 1.000.000 sampai 3.000.000 akan diberi label 2, dan kredit dan utang 0 sampai 1.000.000 akan diberi label 3.

Tabel 2 Data sebelum Encoding

Sebelum Encoding		
Penghasilan	**Jumlah Tanggungan**	**Kredit/Cicilan**
2300000	4	100000
3000000	3	0
7000000	4	2500000
1600000	4	0
5000000	4	200000
7000000	3	500000
2900000	6	0
3200000	4	0
8000000	3	2000000
7000000	4	2500000
5000000	4	200000

Sebelum Encoding		
Penghasilan	**Jumlah Tanggungan**	**Kredit/Cicilan**
2300000	4	100000

Data pada tabel diatas akan diubah menjadi angka atau numerik.

Tabel 3 Data setelah Encoding

Setelah Encoding		
Penghasilan	**Jumlah Tanggungan**	**Kredit/Cicilan**
2	2	1
2	3	1
3	2	2
1	2	1
3	2	1
3	3	1
2	1	1
1	3	1
3	3	2
3	2	2
3	2	1
2	2	1

Pengklasifikasian Model III dan IV tidak menerapkan *Encoding* dalam fase *Pre-Processing*. Data asli sebelum *Encoding* akan langsung digunakan untuk pelabelan.

2. Pelabelan

Data yang telah diperoleh selanjutnya akan diberi label. Label yang diberikan yaitu Tidak Mampu, Menengah, dan Mampu.

Pengklasifikasian akan menggunakan ketentuan khusus dari pihak pesantren dalam pemberian label. Berikut adalah rumus ketentuan pelabelan dari pihak pesantren.

$$P - (JT \times 500.000) - KC$$

Keterangan:

P : Penghasilan

JT : Jumlah Tanggungan

KC : Kredit Cicilan

Jika hasil pengurangan diatas 2.000.000 maka akan diberi label Mampu. Jika hasil pengurangan 500.000 sampai dengan 2.000.000 maka akan diberi label Menengah. Jika hasil pengurangan dibawah 500.000 maka akan diberi label Tidak Mampu.

Tabel 4 Pelabelan data Model I dan II

Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label
2	2	1	Tidak Mampu
2	3	1	Menengah
3	2	2	Mampu
1	2	1	Tidak Mampu
3	2	1	Menengah
3	3	1	Mampu
2	1	1	Tidak Mampu
1	3	1	Tidak Mampu
3	3	2	Mampu
3	2	2	Mampu
3	2	1	Menengah
2	2	1	Tidak Mampu

Pelabelan pada pengklasifikasian model III dan IV memiliki perbedaan dengan model I dan II, karena model III dan IV tidak menerapkan *Encoding* pada fase *Pre-Processing*.

3. *Synthetic Minority Over-Sampling Technique (SMOTE)*

Fungsi dari *SMOTE* pada penelitian ini adalah untuk menyeimbangkan ketimpangan label data. Pada penelitian ini *SMOTE* akan diterapkan pada pengklasifikasian model I dan III.

Tabel 5 Penerapan SMOTE pada Model I

Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label
2	2	1	Tidak Mampu
2	3	1	Menengah
3	2	2	Mampu
1	2	1	Tidak Mampu
3	2	1	Menengah

Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label
3	3	1	Mampu
2	1	1	Tidak Mampu
1	3	1	Tidak Mampu
3	3	2	Mampu
3	2	2	Mampu
3	2	1	Menengah
2	2	1	Tidak Mampu
2	3	1	Menengah
3	2	1	Menengah
3	3	2	Mampu

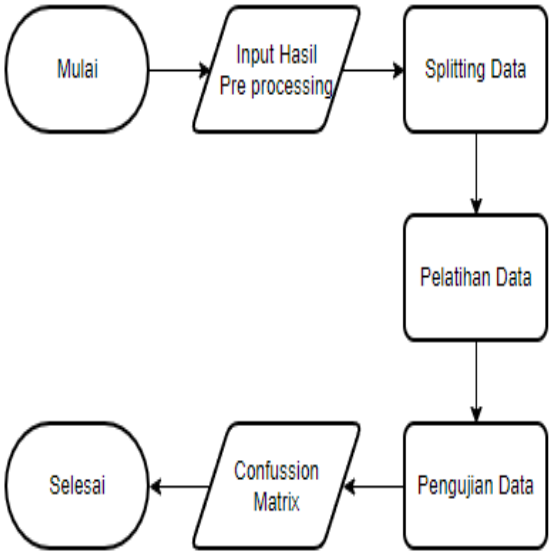
Tabel 6 Penerapan SMOTE pada Model III

Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label
2300000	4	100000	Tidak Mampu
3000000	3	0	Menengah
7000000	4	2500000	Mampu
1600000	4	0	Tidak Mampu
5000000	4	200000	Menengah
7000000	3	500000	Mampu
2900000	6	0	Tidak Mampu
3200000	4	0	Tidak Mampu
8000000	3	2000000	Mampu
7000000	4	2500000	Mampu
5000000	4	200000	Menengah
2300000	4	100000	Tidak Mampu
3000000	3	0	Menengah
5000000	4	200000	Menengah
8000000	3	2000000	Mampu

Hasil *SMOTE* pada kedua tabel diatas ditunjukkan dengan kolom berwarna kuning. Label yang jumlah sebelumnya timpang menjadi setara setelah dilakukan *SMOTE*.

e. Processing

Pada tahap *Processing*, hasil *Pre-Processing* dijalankan melalui *splitting data*, pelatihan data, dan pengujian data untuk mendapatkan *Confusion matrix* dan akurasi dari algoritma yang digunakan. Berikut adalah tahap Processing.



Gambar 2 Tahap Processing

1. Input Data

Data yang telah melalui fase *pre-processing* akan di input kedalam sistem untuk selanjutnya akan dilakukan *splitting*.

2. *Splitting*

Splitting merupakan proses dimana *dataset* akan dibagi menjadi dua, menghasilkan 70% data latih dan 30% data uji. Splitting data pada penelitian ini akan diterapkan pada masing-masing model. Model I dan III yang berjumlah 15 dataset setelah penerapan SMOTE menghasilkan 11 data latih dan 4 data uji. Model II dan IV tetap berjumlah 12 dataset, menghasilkan 9 data latih dan 3 data uji.

3. Pelatihan Data

Sebelum melakukan pengujian data, sistem akan dilatih menggunakan data latih terlebih dahulu dengan tujuan untuk mendapatkan hasil yang terbaik. Langkah awal pengimplementasian perhitungan algoritma *Naive Bayes* adalah menghitung probabilitas dari setiap label dan atribut. Adapun cara menghitung probabilitas label yaitu dengan membagi jumlah kelas yang muncul dengan seluruh total kelas.

Tabel 7 Probabilitas Label Data Latih Model I dan III

Probabilitas Label		
Mampu	4	0,36
Menengah	3	0,27
Tidak Mampu	4	0,36

Tabel 8 Probabillitas Label Data Latih Model II dan IV

Probabilitas Label		
Mampu	3	0,33
Menengah	2	0,22
Tidak Mampu	4	0,44

Setelah probabilitas dari label sudah ditemukan, maka selanjutnya menghitung probabilitas atribut pada masing-masing model.

Tabel 9 Probabilitas atribut Model I

Penghasilan					
1		2		3	
0	0	0	0	4	1
0	0	1	0,33	2	0,67
2	0,5	2	0,5	0	0
Jumlah Tanggungan					
1		2		3	
0	0	2	0,5	2	0,5
0	0	2	0,67	1	0,33
1	0,25	2	0,5	1	0,25
Kredit/Cicilan					
1		2		3	
1	0,25	3	0,75	0	0
3	1	0	0	0	0
4	1	0	0	0	0

Tabel 10 probabilitas atribut Model II

Penghasilan					
1		2		3	
0	0	0	0	3	1
0	0	1	0,5	1	0,5
2	0,5	2	0,5	0	0
Jumlah Tanggungan					
1		2		3	
0	0	1	0,33	2	0,67
0	0	1	0,5	1	0,5
1	0,25	2	0,5	1	0,25
Kredit/Cicilan					
1		2		3	
1	0,33	2	0.67	0	0
2	1	0	0	0	0
4	1	0	0	0	0

Tabel 11 probabilitas atribut Model III

Penghasilan					
0 - 2jt		>2jt - 4jt		>4jt	
0	0	0	0	0	0
0	0	1	0	0	1
1	0,25	3	1	0,25	3
Jumlah Tanggungan					
1 - 2		3 - 4		>5	
0	0	4	0	0	4
0	0	3	0	0	3
0	0	3	0	0	3
Kredit/Cicilan					
0 - 1jt		>1jt - 2jt		>2jt	
1	0,25	1	1	0,25	1
3	1	0	3	1	0
4	1	0	4	1	0

Tabel 12 Probabilitas atribut Model IV

Penghasilan					
0 - 2jt		>2jt - 4jt		>4jt	
0	0	0	0	3	1
0	0	1	0,5	1	0,5
1	0,25	3	0,75	0	0
Jumlah Tanggungan					
1 - 2		3 - 4		>5	
0	0	3	1	0	0
0	0	2	1	0	0
0	0	3	0,75	1	0,25

Kredit/Cicilan					
0 - 1jt		>1jt - 2jt		>2jt	
1	0,33	1	0,33	1	0,33
2	1	0	0	0	0
4	1	0	0	0	0

4. Pengujian Data

Setelah probabilitas diketahui, maka dilakukan perhitungan pada masing-masing model. Perhitungan data uji dilakukan dengan cara mengalikan probabilitas label dengan probabilitas atribut.

Tabel 13 Hasil Penghitungan

Hasil Penghitungan Data Uji Model I				
Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label Aktual	Label Prediksi
2	2	1	Tidak Mampu	Tidak Mampu
2	3	1	Menengah	Tidak Mampu
3	2	1	Menengah	Menengah
3	3	2	Mampu	Mampu
Hasil Penghitungan Data Uji Model II				
Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label Aktual	Label Prediksi
3	2	2	Mampu	Mampu
3	2	1	Menengah	Menengah
2	2	1	Tidak Mampu	Tidak Mampu
Hasil Penghitungan Data Uji Model III				
Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label Aktual	Label Prediksi
2300000	4	100000	Tidak Mampu	Tidak Mampu
3000000	3	0	Menengah	Tidak Mampu
5000000	4	200000	Menengah	Menengah
8000000	3	2000000	Mampu	Mampu
Hasil Penghitungan Data Uji Model IV				
Penghasilan	Jumlah Tanggungan	Kredit/Cicilan	Label Aktual	Label Prediksi
7000000	4	2500000	Mampu	Mampu
5000000	4	200000	Menengah	Menengah
2300000	4	100000	Tidak Mampu	Tidak Mampu

Penghitungan prediksi algoritma *Naïve Bayes* dari seluruh model menunjukkan dua data uji pada Model I dan II terdapat masing-masing satu data dengan kolom berwarna kuning yang tidak terprediksi dengan akurat.

5. *Confusion Matrix*

Setelah data berhasil di klasifikasi menggunakan metode *Naive Bayes* selanjutnya hasil dari prediksi akan di lakukan pengujian pada masing-masing model untuk mengetahui seberapa tepat hasil prediksi dengan menggunakan *Confusion Matrix*.

Tabel 14 Confusion Matrix Model I

Aktual	Prediksi		
	Tidak Mampu	Menengah	Mampu
Tidak Mampu	1	0	0
Menengah	1	1	0
Mampu	0	0	1

Tabel 15 Confusion Matrix Model II

Aktual	Prediksi		
	Tidak Mampu	Menengah	Mampu
Tidak Mampu	1	0	0
Menengah	0	1	0
Mampu	0	0	1

Tabel 16 Confusion Matrix Model III

Aktual	Prediksi		
	Tidak Mampu	Menengah	Mampu
Tidak Mampu	1	0	0
Menengah	1	1	0
Mampu	0	0	1

Tabel 17 Confusion Matrix Model IV

Aktual	Prediksi		
	Tidak Mampu	Menengah	Mampu
Tidak Mampu	1	0	0
Menengah	0	1	0
Mampu	0	0	1

Setelah hasil prediksi sudah diuji menggunakan *Confusion Matrix* kemudian akan dihitung nilai *Accuracy*, *Precision*, dan *Recall*.

a) Akurasi *(Accuracy)*

Rumus akurasi adalah sebagai berikut:

$$\frac{TP + TN}{(TP + TN + FP + FN)}$$

Implementasi hasil *Confusion Matrix* dari rumus diatas pada masing-masing model adalah Model I $\frac{3}{4} = 0.75 = 75\%$, Model II $\frac{3}{3} = 1 = 100\%$, Model III $\frac{3}{4} = 0.75 = 75\%$, Model IV $\frac{3}{3} = 1 = 100\%$

b) Presisi *(Precision)*

Rumus Presisi adalah sebagai berikut:

$$\frac{TP}{(TP + FP)}$$

Implementasi hasil *Confusion Matrix* dari rumus diatas pada masing-masing model adalah Model I $\frac{0,5+1+1}{3} = 0,83 = 83\%$, Model II $\frac{1+1+1}{3} = 1 = 100\%$, Model III $\frac{0,5+1+1}{3} = 0,83 = 83\%$, Model IV $\frac{1+1+1}{3} = 1 = 100\%$

c) Recall

Rumus Recall adalah sebagai berikut:

$$\frac{TP}{(TP + FN)}$$

Implementasi hasil *Confusion Matrix* dari rumus diatas pada masing-masing model adalah Model I $\frac{1+0.5+1}{3} = 0,83 = 83\%$, Model II $\frac{1+1+1}{3} = 1 = 100\%$, Model III $\frac{1+0.5+1}{3} = 0,83 = 83\%$, Model IV $\frac{1+1+1}{3} = 1 = 100\%$

Model II dan IV mampu memprediksi data uji dengan akurat, sementara Model I dan III menghasilkan satu kesalahan prediksi, di mana label aktual 'Menengah' terprediksi sebagai 'Tidak Mampu'. Kesalahan ini umum terjadi, dan akan menghasilkan *Confusion Matrix* untuk mengevaluasi akurasi, presisi, dan recall pada hasil prediksi.

Tabel 18 Accuracy, Precision, dan Recall

Hasil	Pengujian			
	Model I	Model II	Model III	Model IV
Akurasi	75%	100%	75%	100%
Presisi	83%	100%	83%	100%
Recall	83%	100%	83%	100%

4. HASIL DAN PEMBAHASAN

Pada bagian ini menjelaskan tentang pengujian dan hasil pengklasifikasian menggunkan algoritma Naïve Bayes untuk menglasifikasi data pembayaran syahriyah Pondok Pesantren Darussa'adah.

a. Pre-Processing

1) Encoding dilakukan pada pengujian Model I dan II dengan memberikan kode berupa angka 1, 2, dan 3 pada masing-masing kolom.

new_data_kode

	Kode Penghasilan	Kode Tanggungan	Kode Kredit dan Utang
0	2	3	1
1	3	2	1
2	2	3	1
3	3	3	2
4	2	3	1
...	...	...	...
495	2	3	1
496	2	3	1
497	2	3	1
498	2	3	1
499	2	3	1

500 rows × 3 columns

Gambar 3 Hasil Encoding Model I dan II

2) Pelabelan dilakukan pada semua model. Label yang diberikan yaitu Tidak Mampu, Menengah, dan Mampu.

data_baru

	Kode Penghasilan	Kode Tanggungan	Kode Kredit dan Utang	Label
0	2	3	1	Menengah
1	3	2	1	Menengah
2	2	3	1	Menengah
3	3	3	2	Mampu
4	2	3	1	Menengah
...	...	...	...	...
495	2	3	1	Menengah
496	2	3	1	Menengah
497	2	3	1	Menengah
498	2	3	1	Menengah
499	2	3	1	Menengah

500 rows × 4 columns

Gambar 4 Hasil dari Pelabelan

3) *Synthetic Minority Over-Sampling Technique (SMOTE)* akan diterapkan pada pengujian Model I dan III. Dari data yang ada, jumlah data pada Model I dan III memiliki ketimpangan yang cukup signifikan. Data aktual pada Model I dengan label Tidak Mampu berjumlah 210 data, label Menengah 315 data, dan label Mampu 475 data. Sedangkan unutk Model III, data dengan label Tidak Mampu berjumlah 168 data, label Menengah 255 data, dan label Mampu 577 data. Oleh karena itu *SMOTE* diterapkan untuk menyeimbangkan data.

```
Data setelah SMOTE:
Label
Menengah      475
Mampu         475
Tidak Mampu   475
Name: count, dtype: int64
      Kode Penghasilan  Kode Tanggungan  Kode Kredit dan Utang        Label
0                    2                3                      1    Menengah
1                    3                2                      1    Menengah
2                    2                3                      1    Menengah
3                    3                3                      2       Mampu
4                    2                3                      1    Menengah
...                ...              ...                    ...          ...
1420                 2                2                      1  Tidak Mampu
1421                 2                2                      1  Tidak Mampu
1422                 2                2                      1  Tidak Mampu
1423                 2                2                      1  Tidak Mampu
1424                 2                2                      1  Tidak Mampu

[1425 rows x 4 columns]
```

Gambar 5 Hasil SMOTE Model I

b. Processing

1) *Splitting data* bertujuan untuk membagi data menjadi data latih dan data uji. Pada penelitian ini, *splitting data* akan dilakukan pada masing-masing model, data akan dibagi menjadi 70% untuk data latih dan 30% data uji.

2) Pelatihan Data dilakukan pada data training masing-masing model menggunakan algoritma Naïve Bayes. Langkah ini bertujuan untuk melatih mesin untuk mempelajari dataset.

3) Pengujian Data diterapkan pada data testing untuk mengetahui keakuratan prediksi mesin.

```
# Melakukan prediksi pada data pengujian
predicted = clf.predict(X_test.drop('Label', axis=1))
```

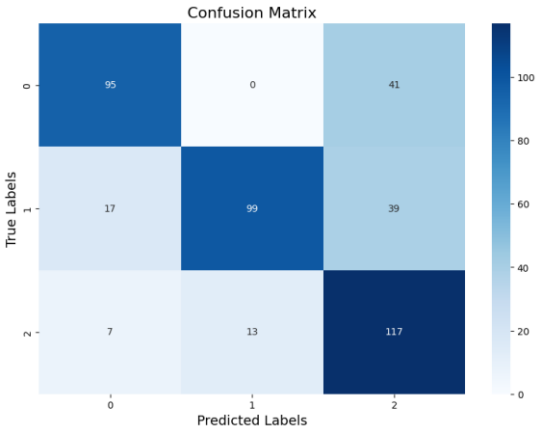
Gambar 6 Sourcecode Algoritma Naive Bayes

Hasil dari implementasi algoritma *Naive Bayes* pada data *testing* masing-masing model akan dibandingkan untuk mengetahui performa model terbaik.

Tabel 19 Hasil Performa tiap Model

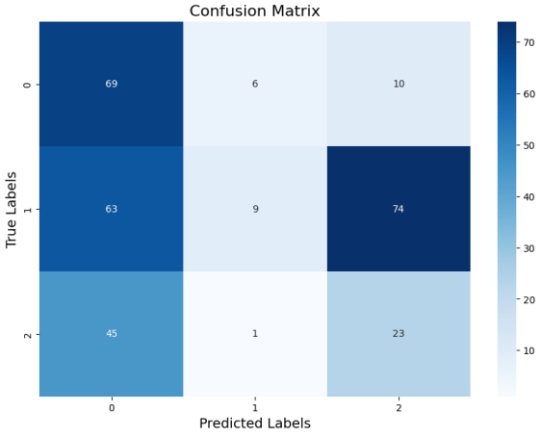
Hasil	Pengujian			
	Model I	Model II	Model III	Model IV
Akurasi	73%	34%	96%	95%
Presisi	76%	39%	96%	94%
Recall	73%	40%	96%	97%

4) *Confussion Matrix* dihasilkan dari implementasi algoritma *Naive Bayes*. Pada penelitian ini menghasilkan empat *Confussion Matrix* dari masing-masing empat model.



Gambar 7 Confusion Matrix Model I

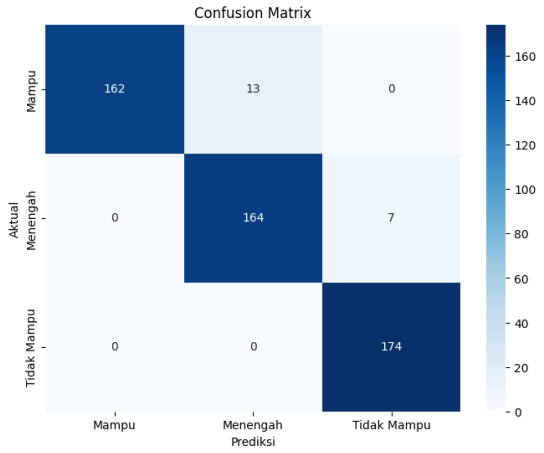
Confusion Matrix Model I dapat diketahui tepat dan tidak tepatnya prediksi model dalam mengklasifikasi data. Pada label Mampu terdapat 95 data yang terprediksi dengan benar, 0 data terprediksi pada label Menengah, dan 41 data terprediksi pada label Tidak Mampu. Pada label Menengah terdapat 99 data yang terprediksi dengan benar, 17 data terprediksi pada label Mampu, dan 39 data terprediksi pada label tidak Mampu. Pada label Tidak Mampu terdapat 117 data yang terprediksi dengan benar, 7 data yang terprediksi pada label Mampu, dan 13 data yang terprediksi dalam label Menengah.



Gambar 8 Confusion Matrix Model II

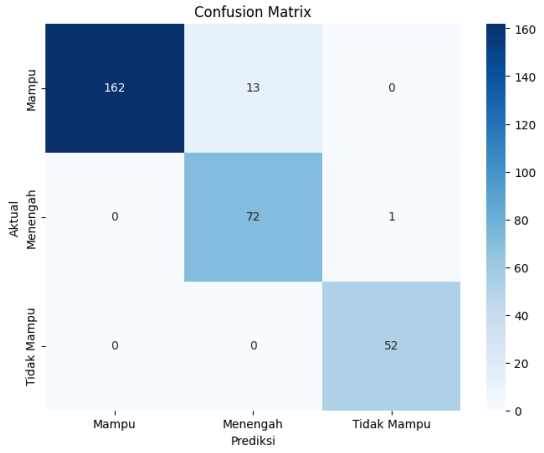
Confusion Matrix Model II menunjukkan bahwa pada label Mampu terdapat 69 data yang terprediksi dengan benar, 6 data terprediksi pada label Menengah, dan 10 data terprediksi pada label Tidak Mampu. Pada label Menengah terdapat 9 data yang terprediksi dengan benar, 63 data terprediksi pada label Mampu, dan 74 data terprediksi pada label tidak Mampu. Pada label Tidak Mampu terdapat 23 data yang terprediksi dengan benar, 45 data yang terprediksi pada label Mampu, dan 1 data yang terprediksi

dalam label Menengah.



Gambar 9 Confusion Matrix Model III

Confusion Matrix Model III menunjukkan bahwa pada label Mampu terdapat 162 data yang terprediksi dengan benar, 13 data terprediksi pada label Menengah, dan 0 data terprediksi pada label Tidak Mampu. Pada label Menengah terdapat 164 data yang terprediksi dengan benar, 0 data terprediksi pada label Mampu, dan 7 data terprediksi pada label tidak Mampu. Pada label Tidak Mampu terdapat 174 data yang terprediksi dengan benar, 0 data yang terprediksi pada label Mampu, dan 0 data yang terprediksi dalam label Menengah.



Gambar 10 Confusion Matrix Model IV

Confusion Matrix Model IV menunjukkan bahwa pada label Mampu terdapat 162 data yang terprediksi dengan benar, 13 data terprediksi pada label Menengah, dan 0 data terprediksi pada label Tidak Mampu. Pada label Menengah terdapat 72 data yang terprediksi dengan benar, 0 data terprediksi pada label Mampu, dan 1 data terprediksi pada label tidak Mampu. Pada label Tidak Mampu terdapat 52 data yang terprediksi dengan benar, 0 data yang terprediksi pada label Mampu, dan 0 data yang terprediksi dalam label Menengah.

5. KESIMPULAN

Penelitian ini melibatkan empat model yang berbeda untuk membandingkan performa model terbaik. Penelitian ini menggunakan metode *Naive Bayes* dengan data *training* sebesar 70% dari *dataset* dan data *testing* sebesar 30% dari *dataset*. Pengujian yang telah dilakukan pada keempat model menghasilkan akurasi, presisi, dan recall yang berbeda, dan Model III adalah yang terbaik dengan akurasi sebesar 96%, presisi sebesar 96%, dan *recall* sebesar 96%.

6. REFERENSI

[1] Suarpurningsih, N.K.A., et al. (2022). Klasifikasi Kelayakan Kredit dengan Naive Bayes.
[2] Marlina, D., & Bakri, M. (2021). Tahapan Data Mining dalam Prediksi Nasabah.
[3] Hudayat, M., et al. (2023). Perbandingan Naive Bayes dan KNN.
[4] Arhami, M., & Nasir, M. (2020). Data Mining: Algoritma dan Implementasi.