VOL. 2 NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

JOURNAL OF COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE (JCSAI)

PENERAPAN NAÏVE BAYES DENGAN BALANCING DATA SMOTE UNTUK PENINGKATAN PREDIKSI NASABAH DEPOSITO BANK

Melan DumasarI Gultom[1], Rani Dwi Destianti[1], Gina Noviamanti[3].

Program Studi Manajemen Informatika[123]
POLITEKNIK GEOLOGI DAN PERTAMBANGAN AGP
melan_dprri@yahoo.com[1], ranidesna3112@gmail.com[1], ginanoviamanti02@gmail.com[3].

(*) Corresponding Author : ranidesna3112@gmail.com[1]

Abstract—*This study aims to address the challenges of classifying potential customers for opening bank deposits, where manual approaches are often time-consuming, inefficient, and inaccurate in capturing complex patterns in customer data. As a solution, this research employs the Naïve Bayes algorithm implemented through the Knowledge Discovery in Databases (KDD) process. The data used is sourced from the Bank Marketing dataset, comprising 41,188 entries and 17 attributes, including demographic information, campaign history, and target outcomes. The research process involves data selection, preprocessing, data transformation, model training, and performance evaluation.The results demonstrate that the Naïve Bayes algorithm can achieve an accuracy of 78.10%, with a precision of 72.42% and recall of 75.76% for customers who open deposits. The SMOTE balancing technique proved effective in enhancing the model's sensitivity to minority classes. Confusion matrix analysis revealed 238 false positives and 200 false negatives, indicating areas for further improvement in model development.This research has significant implications for the banking sector, particularly in optimizing data-driven marketing strategies. The developed model enables banks to allocate marketing resources more efficiently, accurately identify potential customers, and improve operational effectiveness. Future research is suggested to explore other machine learning algorithms, broaden data sources, develop new features, and test the model in real-world environments to enhance generalization and classification performance.*

Keywords: *Deposit Classification, Naïve Bayes Algorithm, Knowledge Discovery in Databases, Machine Learning, Data-Driven Marketing Strategy*

Abstrak—Permasalahan Klasifikasi nasabah potensial yang membuka deposito bank. Tantangan ini muncul karena pendekatan manual yang sering memakan waktu, kurang efisien, dan tidak akurat dalam menangkap pola kompleks dalam data nasabah. Metode Sebagai solusi, penelitian ini menggunakan algoritma Naïve Bayes yang diterapkan melalui proses Knowledge Discovery in Databases (KDD). Data yang digunakan berasal dari dataset Bank Marketing dengan 41.188 entri dan 17 atribut, mencakup informasi demografi, riwayat kampanye pemasaran, dan hasil target. Proses penelitian melibatkan tahapan seleksi data, praproses data, transformasi data, pelatihan model, dan evaluasi performa. Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes mampu mencapai akurasi sebesar 78.10%, dengan presisi sebesar 72.42% dan recall sebesar 75.76% untuk kelas nasabah yang membuka deposito. Teknik balancing dataset menggunakan SMOTE terbukti meningkatkan sensitivitas model terhadap kelas minoritas. Analisis confusion matrix mengungkapkan adanya 238 prediksi positif palsu dan 200 prediksi negatif palsu, yang menunjukkan area untuk perbaikan lebih lanjut dalam pengembangan model. Tujuan Penelitian ini memberikan implikasi penting bagi sektor perbankan, terutama dalam mengoptimalkan strategi pemasaran berbasis data. Dengan model yang dihasilkan, bank dapat mengalokasikan sumber daya pemasaran lebih efisien, mengidentifikasi nasabah potensial secara akurat, dan meningkatkan efektivitas operasional.

Kata Kunci : Klasifikasi Deposito, Algoritma Naïve Bayes, Knowledge Discovery in Databases, Machine Learning, Strategi Pemasaran Berbasis Data.

INTRODUCTION

Sektor perbankan merupakan salah satu pilar penting dalam pembangunan ekonomi modern. Perbankan tidak hanya menyediakan layanan dasar berupa penyimpanan dan transfer dana, tetapi juga memainkan peran strategis dalam menggerakkan roda perekonomian melalui penyaluran kredit dan penyediaan produk investasi. Salah satu produk investasi yang cukup diminati masyarakat adalah deposito bank. Produk ini menawarkan keuntungan

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

berupa tingkat bunga tetap dengan risiko relatif rendah, sehingga menjadi pilihan utama bagi nasabah yang ingin menabung dengan aman sekaligus memperoleh imbal hasil yang stabil [6].

Pertumbuhan dana deposito di Indonesia menunjukkan tren positif. Menurut laporan Bank Indonesia (2023), peningkatan minat masyarakat terhadap deposito mencerminkan adanya kepercayaan publik terhadap stabilitas sektor perbankan. Namun, dalam praktiknya, bank menghadapi tantangan untuk mengidentifikasi nasabah potensial yang berpeluang besar membuka deposito. Proses identifikasi ini sangat krusial bagi keberhasilan strategi pemasaran bank, karena memungkinkan lembaga keuangan untuk mengalokasikan sumber daya secara lebih efisien, meningkatkan efektivitas kampanye, serta memperkuat daya saing di pasar jasa keuangan [6].

Metode yang selama ini digunakan oleh banyak bank dalam menentukan nasabah potensial masih bersifat konvensional. Analisis manual berdasarkan data historis nasabah sering kali tidak mampu menangkap pola kompleks yang tersembunyi dalam jumlah data yang besar. Proses ini bukan hanya memakan waktu, tetapi juga berisiko menghasilkan bias dalam pengambilan keputusan. Hal ini sejalan dengan penelitian Luthfi et al. (2020), yang menyebutkan bahwa pendekatan tradisional sering gagal mendeteksi hubungan non-linear antaratribut nasabah, sehingga tingkat akurasinya rendah [6].

Seiring perkembangan teknologi, machine learning hadir sebagai solusi untuk mengatasi keterbatasan analisis manual. Salah satu algoritma yang banyak digunakan dalam klasifikasi data adalah Naïve Bayes. Algoritma ini bekerja berdasarkan Teorema Bayes dengan asumsi independensi antar fitur. Meskipun sederhana, Naïve Bayes terbukti mampu memberikan hasil klasifikasi yang cukup kompetitif dibandingkan algoritma lainnya, terutama pada dataset yang berskala besar dan memiliki variabel kategorikal [6]. Riyadi dan Setiawan (2020) menunjukkan bahwa penggunaan kombinasi algoritma C4.5 dan Naïve Bayes berhasil meningkatkan akurasi klasifikasi nasabah deposito hingga 85% .

Namun, penerapan algoritma Naïve Bayes juga menghadapi tantangan, khususnya ketika berhadapan dengan dataset yang tidak seimbang (imbalanced dataset). Dalam kasus klasifikasi nasabah deposito, jumlah nasabah yang benar-benar membuka deposito biasanya lebih sedikit dibandingkan dengan yang tidak membuka deposito. Ketidakseimbangan ini dapat menyebabkan bias pada model, di mana algoritma lebih cenderung memprediksi kelas mayoritas, sehingga menurunkan sensitivitas terhadap kelas minoritas yang justru lebih penting bagi bank. Sebagai contoh, model dapat menghasilkan akurasi tinggi secara keseluruhan, tetapi gagal mendeteksi sebagian besar nasabah yang benar-benar berpotensi membuka deposito.

Untuk mengatasi masalah tersebut, penelitian ini mengintegrasikan algoritma Naïve Bayes dengan metode Synthetic Minority Over-Sampling Technique (SMOTE). SMOTE merupakan salah satu teknik *data balancing* yang bekerja dengan cara menghasilkan sampel sintetis pada kelas minoritas berdasarkan interpolasi antar data yang ada. Dengan menyeimbangkan distribusi kelas, model klasifikasi dapat belajar dengan lebih baik sehingga meningkatkan kemampuan deteksi terhadap nasabah yang benar-benar potensial [6].

Pendekatan ini selaras dengan proses Knowledge Discovery in Databases (KDD) yang digunakan dalam penelitian, yaitu mencakup tahapan seleksi data, preprocessing, transformasi, penerapan algoritma, dan evaluasi model. KDD memberikan kerangka kerja sistematis untuk mengolah data berskala besar menjadi pengetahuan yang bermakna. Dalam penelitian ini, KDD digunakan untuk memastikan data nasabah diproses secara optimal sebelum diterapkan ke dalam model Naïve Bayes yang dilengkapi balancing SMOTE .

Selain itu, keunggulan lain dari penggunaan Naïve Bayes adalah efisiensi komputasi yang tinggi. Algoritma ini mampu mengolah dataset berukuran besar dengan kecepatan yang baik, sehingga cocok digunakan dalam dunia perbankan yang memiliki data dinamis dan terus bertambah setiap hari. Dengan mengombinasikan kecepatan Naïve Bayes dan keefektifan SMOTE dalam menangani dataset tidak seimbang, penelitian ini diharapkan dapat menghasilkan model klasifikasi yang lebih akurat dan aplikatif untuk mendukung strategi pemasaran bank.

Manfaat praktis dari penelitian ini sangat besar bagi industri perbankan. Pertama, model yang dikembangkan dapat membantu bank dalam menyusun strategi pemasaran berbasis data, sehingga kampanye dapat difokuskan pada segmen nasabah yang benar-benar potensial. Kedua, model ini dapat meningkatkan efisiensi operasional dengan mengurangi pemborosan sumber daya pada nasabah yang tidak tertarik dengan produk deposito. Ketiga, dengan meningkatnya tingkat konversi nasabah potensial menjadi nasabah deposito, bank dapat memperkuat posisi keuangannya sekaligus memberikan kontribusi pada pertumbuhan ekonomi nasional.

Lebih jauh, penelitian ini juga memiliki kontribusi teoretis, yaitu memperluas penerapan algoritma Naïve Bayes dalam domain perbankan dengan pendekatan balancing data. Meskipun

algoritma ini telah banyak digunakan di berbagai bidang, penerapannya pada kasus klasifikasi deposito bank dengan integrasi SMOTE masih relatif jarang dilakukan. Dengan demikian, penelitian ini memberikan novelty berupa pendekatan hibrid antara Naïve Bayes dan SMOTE yang diharapkan dapat menjadi rujukan untuk penelitian serupa di masa depan.

MATERIALS AND METHODS

A. Dataset

Dataset yang digunakan dalam penelitian ini berasal dari data nasabah bank yang diperoleh melalui sumber sekunder, sebagaimana tercantum dalam laporan penelitian

Dataset tersebut berisi atribut-atribut yang relevan dengan keputusan nasabah untuk membuka deposito, antara lain umur, pekerjaan, status perkawinan, pendidikan, saldo rekening, frekuensi transaksi, serta status kepemilikan pinjaman. Variabel target berupa kelas biner, yaitu ya (nasabah membuka deposito) dan tidak (nasabah tidak membuka deposito) Distribusi data pada variabel target menunjukkan ketidakseimbangan kelas, di mana jumlah nasabah yang tidak membuka deposito jauh lebih besar dibandingkan dengan nasabah yang membuka deposito. Kondisi ini berpotensi menimbulkan bias pada model klasifikasi, sehingga diperlukan teknik balancing data sebelum penerapan algoritma

B. Teknik Pengumpulan Data

Data yang digunakan bersifat data sekunder dan telah tersedia dalam bentuk terstruktur (file CSV). Pengumpulan dilakukan dengan mengunduh dataset kemudian diproses menggunakan perangkat lunak Python. Beberapa pustaka yang digunakan dalam analisis adalah pandas untuk pengolahan data, scikit-learn untuk pembangunan model *machine learning*, serta imbalanced-learn untuk penerapan teknik SMOTE

C. Preprocessing Data

Sebelum digunakan pada model, dataset diproses melalui tahapan preprocessing berikut:

1. Data Cleaning**:** Menghapus data duplikat serta menangani *missing values*. Untuk atribut numerik digunakan metode imputasi rata-rata (*mean imputation*), sedangkan untuk atribut kategorikal digunakan modus (*mode imputation*).
2. Encoding**:** Mengubah variabel kategorikal seperti pekerjaan, status perkawinan, dan pendidikan menjadi representasi numerik dengan metode *label encoding*.
3. Normalisasi**:** Variabel numerik seperti saldo rekening dan frekuensi transaksi dinormalisasi dengan teknik Min-Max Normalization agar berada pada skala 0–1, sehingga memperkecil bias dalam perhitungan probabilitas Naïve Bayes.

D. Synthetic Minority Over-Sampling Technique (SMOTE)

Untuk mengatasi permasalahan distribusi data yang tidak seimbang, penelitian ini menerapkan SMOTE (Synthetic Minority Over-Sampling Technique)**.** SMOTE bekerja dengan cara menghasilkan data sintetis pada kelas minoritas dengan melakukan interpolasi antar titik data terdekat [1].

Secara umum, mekanisme SMOTE adalah:

1. Menentukan jumlah data sintetis yang dibutuhkan.
2. Memilih tetangga terdekat dari setiap data minoritas.
3. Menghasilkan sampel baru dengan interpolasi linier antara data asli dan tetangga terdekat tersebut.

Dengan penerapan SMOTE, distribusi data target menjadi lebih seimbang, sehingga meningkatkan sensitivitas model Naïve Bayes terhadap kelas minoritas.

E. Algoritma Naïve Bayes

Algoritma Naïve Bayes digunakan sebagai metode utama klasifikasi. Algoritma ini bekerja berdasarkan Teorema Bayes, dengan asumsi bahwa setiap atribut bersifat independen terhadap atribut lainnya.

Persamaan dasar probabilitas bersyarat adalah:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

dengan:

1. P(C|X)P(C|X)P(C|X): probabilitas kelas CCC diberikan fitur XXX,
2. P(X|C)P(X|C)P(X|C): probabilitas fitur XXX muncul pada kelas CCC,
3. P(C)P(C)P(C): probabilitas awal kelas CCC,
4. P(X)P(X)P(X): probabilitas fitur XXX.

Model Naïve Bayes menghitung probabilitas masing-masing kelas untuk data uji, kemudian memilih kelas dengan nilai probabilitas tertinggi sebagai hasil prediksi [2].

F. Evaluasi Model

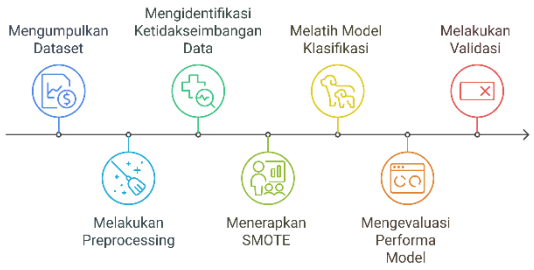
Kinerja model dievaluasi menggunakan metrik berikut [3]:

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

1. Accuracy (Akurasi)**:** proporsi prediksi yang benar terhadap total data.
2. Precision (Presisi)**:** ketepatan prediksi positif yang benar dibandingkan dengan seluruh prediksi positif.
3. Recall (Sensitivitas)**:** kemampuan model dalam mendeteksi seluruh kasus positif.
4. F1-score: rata-rata harmonis presisi dan recall.

Untuk memvalidasi hasil, digunakan metode 10-Fold Cross Validation [4], di mana dataset dibagi menjadi sepuluh bagian, sembilan digunakan untuk pelatihan dan satu untuk pengujian secara bergantian. Metode ini memastikan model tidak hanya efektif pada data latih, tetapi juga mampu digeneralisasikan pada data baru.

G. Alur Penelitian



Gambar 1. Alur Penelitian

Proses analisis data nasabah bank secara sistematis untuk membangun model prediksi nasabah potensial deposito. Tahap pertama adalah mengumpulkan dataset, yaitu memperoleh data nasabah dari sumber sekunder yang berisi atribut demografis dan finansial, seperti usia, status pekerjaan, tingkat pendidikan, saldo, serta riwayat pinjaman. Data ini menjadi dasar penting dalam membangun model klasifikasi. Selanjutnya dilakukan preprocessing, yang mencakup pembersihan data dari nilai yang hilang (*missing values*), penghapusan duplikasi, pengkodean data kategorikal menjadi numerik, serta normalisasi agar seluruh variabel berada pada skala yang eragam.

Tahap berikutnya adalah mengidentifikasi ketidakseimbangan data. Dalam kasus deposito bank, sering kali jumlah nasabah yang tidak membuka deposito jauh lebih besar dibandingkan dengan yang membuka deposito, sehingga data menjadi tidak seimbang. Untuk mengatasi hal tersebut diterapkan metode SMOTE (Synthetic Minority Over-Sampling Technique). Metode ini bekerja dengan menambahkan data sintetis pada kelas minoritas melalui interpolasi, sehingga distribusi kelas menjadi lebih seimbang dan model dapat belajar secara lebih adil.

Setelah data seimbang, proses berlanjut ke tahap melatih model klasifikasi, dalam hal ini menggunakan algoritma Naïve Bayes. Algoritma ini dipilih karena efisien, mudah diimplementasikan, dan mampu bekerja baik dengan data berukuran besar. Setelah model terbentuk, dilakukan evaluasi performa model menggunakan metrik standar, seperti akurasi, presisi, recall, dan F1-score, untuk mengukur sejauh mana model dapat mengklasifikasikan nasabah dengan benar. Tahap terakhir adalah validasi model, yang bertujuan memastikan keandalan hasil klasifikasi dan kemampuan generalisasi model terhadap data baru. Validasi dilakukan dengan teknik *k-fold cross validation*, sehingga hasil evaluasi lebih robust dan tidak bias terhadap data tertentu.

RESULTS AND DISCUSSION

Data Selection

Tahap pertama adalah Seleksi Data, yang bertujuan untuk memilih atribut-atribut relevan dari dataset Bank Marketing yang tersedia di platform Kaggle. Dataset ini memiliki 17 atribut, yang terdiri dari: Atribut Demografi dan Sosial:

Age (usia nasabah)Job (pekerjaan)Marital (status perkawinan)Education (tingkat pendidikan)Atribut Ekonomi:Default (status kredit macet)Housing (kepemilikan rumah)Loan (kepemilikan pinjaman pribadi)Atribut Kampanye

Kontak: Contact (jenis kontak seperti telepon seluler atau telepon rumah)Month (bulan kampanye terakhir dilakukan)Day_of_week (hari dalam seminggu ketika kontak terakhir dilakukan)Duration (durasi kontak terakhir dalam detik)Atribut Riwayat

Kampanye:Campaign (jumlah kontak yang dilakukan dalam kampanye saat ini)Pdays (jumlah hari sejak nasabah terakhir dihubungi pada kampanye sebelumnya)Previous (jumlah kontak yang dilakukan sebelum kampanye ini)Poutcome (hasil kampanye sebelumnya)Atribut

Tambahan:Emp.var.rate (tingkat variabel pekerjaan)Cons.price.idx (indeks harga konsumen)Cons.conf.idx (indeks kepercayaan konsumen)Euribor3m (suku bunga Euribor 3 bulan)Nr.employed (jumlah karyawan yang dipekerjakan)

Target Variable: y (apakah nasabah membuka deposito atau tidak, dengan nilai “yes” atau “no”).

Seleksi data dilakukan dengan mempertimbangkan relevansi atribut terhadap tujuan klasifikasi, yaitu memprediksi nasabah potensial yang membuka deposito. Atribut seperti Age, Job, Education, Contact, dan y dipastikan digunakan karena memiliki pengaruh langsung terhadap keputusan nasabah. Atribut dengan variabilitas rendah atau yang kurang relevan, seperti Day_of_week, dapat diabaikan jika terbukti tidak signifikan terhadap model. Setelah seleksi, data difilter untuk menghilangkan nilai yang hilang (missing values) dan memastikan setiap atribut memiliki format yang konsisten untuk tahap berikutnya dalam proses KDD.

Tabel 1 Data Kaggle

age	job	balance	housing	deposit
32	admin.	678	yes	yes
31	management	141	yes	yes
41	selfemployed	94	yes	yes
22	admin.	897	yes	yes
31	unemployed	1214	no	yes
69	retired	324	no	yes
30	selfemployed	2666	no	yes
49	bluecollar	10613	no	yes
29	management	697	no	yes
43	bluecollar	953	yes	yes

Preprocesing

Tahap kedua adalah Pra-proses Data, yang bertujuan untuk membersihkan dan mempersiapkan data agar siap digunakan dalam analisis. Pada tahap ini, dilakukan beberapa langkah penting. Langkah pertama adalah pembersihan data, di mana nilai-nilai yang hilang (missing values) pada atribut-atribut penting seperti Age, Job, dan Education diidentifikasi dan diatasi. Nilai kosong digantikan dengan median untuk atribut numerik atau modus untuk atribut kategori guna menjaga integritas data tanpa menghilangkan informasi penting. Data duplikat juga dihapus untuk memastikan bahwa setiap baris mewakili nasabah yang unik.

Selanjutnya, dilakukan normalisasi data untuk atribut numerik seperti Age, Duration, dan Campaign menggunakan metode Min-Max Scaling. Proses ini memastikan bahwa nilai-nilai berada dalam rentang [0,1], sehingga menghindari dominasi atribut dengan skala besar terhadap model. Untuk atribut kategori seperti Job, Marital, Education, dan Contact, diterapkan teknik one-hot encoding. Transformasi ini mengubah kategori menjadi representasi numerik, misalnya nilai "married," "single," dan "divorced" diubah menjadi kolom biner terpisah. Atribut target y (yes/no) juga dikodekan menjadi nilai 1 untuk "yes" dan 0 untuk "no."

Selain itu, atribut yang tidak relevan atau memiliki variabilitas rendah, seperti Day_of_week, dihapus jika terbukti tidak memberikan kontribusi signifikan terhadap prediksi. Atribut seperti Duration dapat dihilangkan jika model harus mencerminkan kondisi sebelum kampanye dilakukan, karena durasi hanya diketahui setelah kontak dilakukan. Untuk mengatasi ketidakseimbangan data pada atribut target (y), dilakukan balancing dataset menggunakan metode seperti SMOTE (Synthetic Minority Oversampling Technique) atau undersampling pada kelas mayoritas. Hal ini memastikan bahwa model tidak bias terhadap kelas dengan frekuensi lebih tinggi.

Setelah semua langkah selesai, dilakukan pemeriksaan konsistensi untuk memastikan tidak ada kesalahan format atau nilai yang tidak sesuai. Dataset yang telah diproses kemudian siap digunakan untuk tahap transformasi lebih lanjut atau langsung diterapkan dalam model Naïve Bayes.

Tabel 2 Data Preprocessing

age	job	housing	deposit
32	admin.	yes	yes
31	management	yes	yes
41	selfemployed	yes	yes
22	admin.	yes	yes
31	unemployed	no	yes
69	retired	no	yes
30	selfemployed	no	yes
49	bluecollar	no	yes
29	management	no	yes
43	bluecollar	yes	yes

Transformation

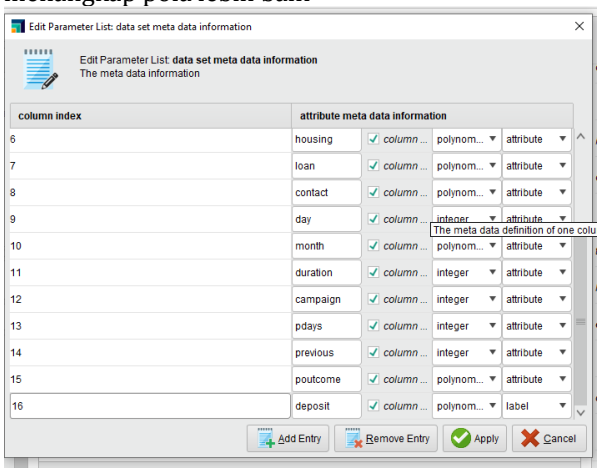
Tahap berikutnya adalah Transformasi Data, yang bertujuan untuk mengubah data mentah menjadi format yang sesuai untuk analisis menggunakan algoritma Naïve Bayes. Transformasi ini dimulai dengan encoding variabel kategori seperti Job, Marital, dan Education menjadi representasi numerik menggunakan teknik one-hot encoding. Teknik ini memastikan bahwa semua atribut kategori diubah menjadi kolom biner, sehingga algoritma dapat memprosesnya secara optimal.

Selanjutnya, dilakukan pembuatan atribut baru berdasarkan kombinasi atau agregasi data yang ada. Sebagai contoh, atribut seperti Campaign dan Previous dapat digunakan untuk membuat fitur baru yang mencerminkan total interaksi selama

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

kampanye sebelumnya. Selain itu, atribut numerik seperti Duration dapat dikategorikan ke dalam rentang tertentu untuk membantu mengidentifikasi pola spesifik dalam data.

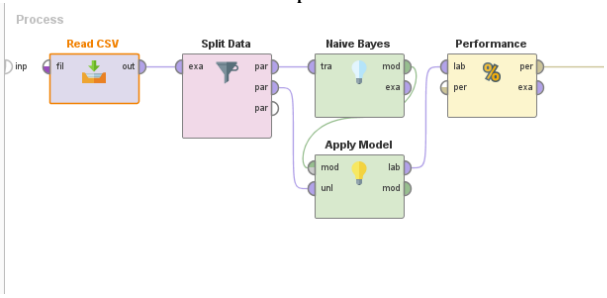
Pada tahap ini, transformasi logaritmik atau normalisasi lanjutan juga diterapkan pada atribut numerik dengan distribusi skewed, seperti Age atau Balance, untuk meningkatkan kinerja model. Transformasi ini membantu menyamakan distribusi data sehingga algoritma dapat menangkap pola lebih baik



Gambar 2 Transformation

Desain Algoritma

Setelah semua transformasi selesai, dataset diperiksa ulang untuk memastikan konsistensi data. Proses ini melibatkan validasi bahwa semua atribut berada dalam format numerik, tidak ada data yang hilang, dan struktur dataset sesuai dengan kebutuhan algoritma. Dataset yang telah melalui proses transformasi kemudian siap digunakan dalam tahap penerapan algoritma Naïve Bayes.Tahap berikutnya adalah Penerapan dan Pengujian Algoritma Naïve Bayes, di mana model klasifikasi dibangun menggunakan algoritma tersebut. Data latih yang telah diproses digunakan untuk melatih model, sehingga model dapat mempelajari pola-pola penting dalam data. Proses ini menghasilkan model yang siap untuk digunakan dalam klasifikasi nasabah potensial.



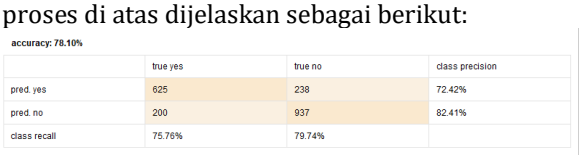
Gambar 3 Algoritma Naïve Bayes

Berdasarkan gambar 3 diatas menggambarkan proses alur analisis dengan menggunakan algoritma Naïve Bayes untuk klasifikasi, yang dijelaskan dalam langkah-langkah berikut:

1. Read CSV: Pada tahap ini, data diimpor dari file CSV yang telah dipersiapkan sebelumnya. Data ini mencakup informasi nasabah, termasuk atribut demografi, riwayat kampanye, dan hasil target (apakah membuka deposito atau tidak). Proses ini merupakan awal dari pipeline analisis.
2. Split Data: Setelah data diimpor, langkah ini membagi dataset menjadi dua bagian utama: data latih (training set) dan data uji (testing set). Biasanya, pembagian dilakukan dengan rasio tertentu, seperti 80:20 atau 70:30, untuk memastikan bahwa model dilatih pada data yang cukup dan diuji pada data yang belum pernah dilihat.
3. Naïve Bayes : Pada tahap ini, algoritma Naïve Bayes digunakan untuk membangun model klasifikasi berdasarkan data latih. Algoritma ini menghitung probabilitas untuk setiap kelas berdasarkan atribut-atribut data dan menentukan kelas yang memiliki probabilitas tertinggi sebagai prediksi.
4. Apply Model : Model yang telah dilatih diterapkan pada data uji untuk menghasilkan prediksi. Tahap ini menguji seberapa baik model dapat mengklasifikasikan data baru yang tidak pernah digunakan selama proses pelatihan.
5. Performance : Tahap terakhir adalah evaluasi kinerja model. Berbagai metrik, seperti akurasi, presisi, recall, dan F1-score, digunakan untuk menilai efektivitas model. Hasil evaluasi ini membantu menentukan apakah model memenuhi tujuan klasifikasi dan seberapa baik prediksi yang dihasilkan.

Evaluasi

Tahapan evaluasi dalam penelitian menggunakan algoritma Naïve Bayes sesuai dengan proses di atas dijelaskan sebagai berikut:



accuracy: 78.10%

	true yes	true no	class precision
pred. yes	625	238	72.42%
pred. no	200	937	82.41%
class recall	75.76%	79.74%	

Gambar 4 Hasil Matrik

Berdasarkan tabel hasil evaluasi model dengan algoritma Naïve Bayes, berikut adalah uraian hasil evaluasi secara rinci:

Akurasi Model : Akurasi model sebesar **78.10%**, yang menunjukkan bahwa model Naïve Bayes mampu mengklasifikasikan data dengan benar sebanyak 78.10% dari total data uji.

Confusion Matrix :

a) **True Yes** (625): Prediksi "yes" (nasabah membuka deposito) yang benar sesuai dengan data aktual.

b) **True No** (937): Prediksi "no" (nasabah tidak membuka deposito) yang benar sesuai dengan data aktual.
c) **False Positives (238)**: Prediksi "yes" yang salah (model memprediksi nasabah akan membuka deposito, tetapi kenyataannya tidak).
d) **False Negatives (200)**: Prediksi "no" yang salah (model memprediksi nasabah tidak akan membuka deposito, tetapi kenyataannya membuka).

Presisi (Precision) :

a) Presisi untuk kelas **"yes"**: **72.42%**, artinya dari semua prediksi "yes", hanya 72.42% yang benar-benar nasabah membuka deposito.
b) Presisi untuk kelas **"no"**: **82.41%**, artinya dari semua prediksi "no", 82.41% benar-benar nasabah tidak membuka deposito.

Recall (Sensitivity):

a) Recall untuk kelas **"yes"**: **75.76%**, artinya model mampu menangkap 75.76% dari total nasabah yang benar-benar membuka deposito.
b) Recall untuk kelas **"no"**: **79.74%**, artinya model mampu menangkap 79.74% dari total nasabah yang benar-benar tidak membuka deposito.

Analisis Hasil:

a) Model menunjukkan performa yang cukup baik, tetapi presisi untuk kelas "yes" lebih rendah dibandingkan kelas "no". Hal ini mungkin disebabkan oleh distribusi data yang tidak seimbang, di mana lebih banyak nasabah yang tidak membuka deposito dibandingkan yang membuka.
b) Recall yang lebih tinggi untuk kedua kelas menunjukkan bahwa model cukup sensitif dalam mendeteksi masing-masing kelas.

Pembahasan

Hasil penelitian menunjukkan bahwa algoritma Naïve Bayes dapat diterapkan secara efektif untuk meningkatkan model klasifikasi deposito bank. Berdasarkan evaluasi yang dilakukan, model berhasil mencapai akurasi sebesar 78.10%. Hal ini mencerminkan bahwa dari seluruh data uji yang digunakan, hampir 80% prediksi yang dibuat oleh model adalah benar. Presisi untuk kelas "yes" sebesar 72.42% menunjukkan bahwa model cukup baik dalam mengidentifikasi nasabah yang berpotensi membuka deposito, meskipun masih terdapat kesalahan pada prediksi positif palsu (false positives).

Salah satu kekuatan model ini adalah recall untuk kelas "yes" sebesar 75.76%, yang menunjukkan kemampuan model dalam menangkap sebagian besar nasabah yang benar-benar membuka deposito. Sementara itu, presisi untuk kelas "no" sebesar 82.41% dan recall sebesar 79.74% menunjukkan performa yang konsisten dalam mengidentifikasi nasabah yang tidak membuka deposito. Namun, perbedaan antara presisi dan recall pada kedua kelas mengindikasikan bahwa model cenderung lebih konservatif dalam memprediksi nasabah membuka deposito dibandingkan dengan memprediksi nasabah yang tidak membuka deposito.

Analisis confusion matrix menunjukkan bahwa terdapat 238 prediksi positif palsu (false positives), di mana model memprediksi nasabah akan membuka deposito, tetapi kenyataannya tidak. Selain itu, terdapat 200 prediksi negatif palsu (false negatives), di mana model memprediksi nasabah tidak akan membuka deposito, tetapi kenyataannya membuka. Pola kesalahan ini dapat dianalisis lebih lanjut untuk mengidentifikasi faktor-faktor yang menyebabkan ketidakakuratan prediksi.

Hasil ini memberikan implikasi penting bagi sektor perbankan. Dengan model yang dikembangkan, bank dapat mengidentifikasi nasabah potensial untuk membuka deposito secara lebih akurat, sehingga meningkatkan efisiensi strategi pemasaran. Model ini juga dapat membantu bank dalam mengalokasikan sumber daya pemasaran secara lebih efektif, misalnya dengan fokus pada nasabah yang memiliki probabilitas tinggi untuk membuka deposito.

Selain itu, hasil penelitian ini menunjukkan bahwa balancing dataset menggunakan SMOTE dapat membantu meningkatkan performa model. Dengan memastikan bahwa kelas minoritas (nasabah yang membuka deposito) memiliki representasi yang cukup dalam data latih, model dapat lebih sensitif terhadap pola-pola yang terkait dengan kelas tersebut. Hal ini menjadi bukti bahwa teknik praproses data memiliki peran penting dalam meningkatkan kualitas hasil analisis.

Namun, penelitian ini juga memiliki beberapa keterbatasan. Salah satunya adalah fokus pada satu algoritma, yaitu Naïve Bayes. Meskipun algoritma ini memiliki kelebihan dalam hal kesederhanaan dan efisiensi komputasi, eksplorasi algoritma lain seperti Random Forest atau Gradient Boosting dapat memberikan perbandingan yang lebih komprehensif. Selain itu, meskipun dataset yang digunakan memiliki 41.188 entri, penambahan data dari berbagai bank atau sumber lain dapat meningkatkan generalisasi model.

Hasil penelitian ini memberikan kontribusi signifikan terhadap pengembangan strategi pemasaran berbasis data di sektor perbankan. Dengan memanfaatkan algoritma machine learning seperti Naïve Bayes, bank dapat membuat keputusan yang lebih informatif dan strategis. Ke depan, penelitian lebih lanjut dapat dilakukan dengan mengeksplorasi atribut tambahan atau

VOL. 2. NO. 1 JANUARY 2025
P-ISSN: XXXX-XXXX | E-ISSN: XXXX-XXXX
DOI: XX.XXXXX/jcsai.vXXX.XXXX

menggabungkan algoritma yang berbeda untuk meningkatkan akurasi dan efisiensi model.

KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa algoritma Naïve Bayes dapat digunakan secara efektif untuk meningkatkan model klasifikasi deposito bank. Dengan akurasi model sebesar 78.10% dan performa yang cukup baik dalam presisi serta recall, algoritma ini membuktikan kemampuannya dalam mengidentifikasi nasabah potensial yang membuka deposito. Proses praproses data, seperti balancing dataset menggunakan SMOTE dan normalisasi atribut, terbukti berperan penting dalam meningkatkan performa model.

Implikasi dari penelitian ini mencakup kemampuan bank untuk mengoptimalkan strategi pemasaran berbasis data, mengalokasikan sumber daya secara lebih efisien, dan meningkatkan efisiensi operasional. Namun, penelitian ini juga menunjukkan bahwa eksplorasi lebih lanjut pada algoritma lain atau integrasi berbagai algoritma dapat memberikan hasil yang lebih optimal. Penelitian masa depan diharapkan dapat mengatasi keterbatasan ini dengan mempertimbangkan variasi dataset yang lebih luas dan algoritma pembelajaran mesin yang lebih kompleks.

REFERENCE

[1] Agung, A., Wahyu, G., Erlangga, S., Gede, I., Gunadi, A., Made, I., & Sunarya, G. (2020). *Kombinasi Oversampling dan Undersampling dalam Menangani Class Imbalanced dan Overlapping pada Klasifikasi Data Bank Marketing*. https://s.id/jurnalresistor

[2]Arya Pramudya, R., Mutiarachim, A., & Setya Sunarka, P. (2024). Klasifikasi Pola Pembelian Kendaraan Bermotor Untuk Merancang Strategi Promosi Terarah Menggunakan Algoritma Logistic Regression. *Indo-Fintech Intellectuals: Journal of Economics and Business*, *4*(5), 2549–2557. https://doi.org/10.54373/ifijeb.v4i5.212

[3] Astofa, A., & Sutono, E. (2024). *Implementasi Algoritma Nave Bayes Untuk Memprediksi Kelayakan Kredit Nasabah*. https://journal.mediapublikasi.id/index.php/logic

[4] Aulia, R., & Fadli, M. (2023). Perbandingan Algoritme C5.0 dan Random Forest menggunakan Data Bank Marketing. In *JUTIF: Jurnal Teknologi Informasi* (Vol. 1, Issue 1).

[5] Dwi, E., Aini, N., Khasanah, R. A., Ristyawan, A., Diniati, E., Nusantara, U., & Kediri, P. (2024). Penggunaan Data Mining untuk Prediksi tingkat Obesitas di Meksiko Menggunakan Metode Random Forest. In *Agustus* (Vol. 8). Online.

[6] Dwi Handayani. (2020). Analisis Prediksi Nasabah yang Berpotensi Membuka Deposito pada Bank Umum di Bekasi Menggunakan Algoritma C4.5 dan Naive Bayes. *Jurnal Ilmiah Komputasi*, *19*(3). https://doi.org/10.32409/jikstik.19.3.66

[7] Fikri Eina, M., & Herry Chrisnanto, Y. (2024). KLASIFIKASI TELEMARKETING MENGGUNAKAN NAÏVE BAYES CLASSIFICATION DAN WRAPPER SEQUENTIAL FEATURE SELECTION TELEMARKETING CLASSIFICATION USING NAÏVE BAYES CLASSIFICATION AND WRAPPER SEQUENTIAL FEATURE SELECTION. *Journal of Information Technology and Computer Science (INTECOMS)*, *7*(4).

[8] Homepage, J., Kenia, S., Loka, P., & Marsal, A. (2023). *MALCOM: Indonesian Journal of Machine Learning and Computer Science Comparison Algorithm of K-Nearest Neighbor and Naïve Bayes Classifier for Classifying Nutritional Status in Toddlers Perbandingan Algoritma K-Nearest Neighbor dan Naïve Bayes Classifier Untuk Klasifikasi Status Gizi Pada Balita*. *3*, 8–14.

[9] Koklu, N., & Sulak, S. A. (2024). Using Artificial Intelligence Techniques for the Analysis of Obesity Status According to the Individuals' Social and Physical Activities. *Sinop Üniversitesi Fen Bilimleri Dergisi*, *9*(1), 217–239. https://doi.org/10.33484/sinopfbd.1445215

[10] Leonard, B., Panggabean, E., Klasifikasi, :, Potensial, N., Aziz, F., Komputer, J. I., Mipa, F., & Pancasakti, U. (2023). *Klasifikasi Nasabah Potensial menggunakan Algoritma Ensemble Least Square Support Vector Machine dengan AdaBoost*. *8*(3).