ISSN:2302-8149
e-ISSN:2540-9719

Optimasi Algoritma Naive Bayes dengan Kombinasi SMOTETomek untuk Imbalance Class Fraud Detection

Optimization of the Naive Bayes Algorithm with SMOTETomek Combination for Imbalance Class Fraud Detection

[1]Arief Tri Arsanto, [2]Arif Faizin, [3]Moch. Lutfi*, [4]Zulfatun Nikmatus Saadah
[1,2,3,4]Teknik informatika, Universitas Yudharta Pasuruan, Indonesia
*email: moch.lutfi@yudharta.ac.id

(***received*:** 28 October 2024, ***revised*:** 3 November 2024, ***accepted*:** 6 November 2024)

Abstrak

Penggunaan kartu kredit di jaman modern semakin meningkat oleh karna itu perlu pencegahan dengan Teknologi yang telah digunakan seperti *Address Verification Systems* (AVS), *Card Verification Method* (CVM) dan Nomor Identifikasi Pribadi (PIN). Analisa dataset perlu dilakukan untuk mengalisa histori transaksi yang sudah perna dilakukan, Dalam dataset fraud detection terlihat adanya attribut yang mengalami imbalance data. Imbalance class dalam dataset merupakan masalah signifikan dalam machine learning yang dapat memepengaruhi kinerja model secara keseluruhan. Dalam satu kelas jumlah sampel mayoritas memiliki jumlah lebih banyak dari jumlah kelas minoritas. pada penelitian ini digunakan pendekatan *oversampling* dengan menggunakan kombinasi *smote* dan *tomek-link*. Fokus pada penelitian ini adalah klasifikasi card fraud deteksi pada dataset yang tidak seimbang atau dengan kelas imbalanced dilakukan dengan memanfaatkan metode Naive Bayes sebagai algoritma klasifikasi. Selain itu, kombinasi teknik *resampling* juga diterapkan untuk mengatasi kelas yang tidak seimbang dalam dataset ini melalui pendekatan *SMOTETomek*. *SMOTETomek* adalah metode yang mengurangi jumlah sampel dengan memperhatikan dua data yang berdekatan dari kelas minoritas dan mayoritas. Sedangkan dari permasalahan diatas hasil kinerja naïve bayes yang mengalami masalah dengan data imblance pada penelitian ini diusulkan metode *resampling* dengan harapan bisa memperbaiki kinerja dari algoritma naïve bayes dan pada hasil curva ROC AUC bahwa metode *SMOTETomek* dapat memperbaiki performance algoritma naïve bayes, Semakin tinggi skor ROC-AUC, semakin baik kinerja model dalam hal kemampuannya untuk membedakan antara dua kelas, namun pada hasil akurasi tidak mengalami perubahan yang signifikan.

Kata kunci: fraud detection, sampling, naive bayes

Abstract

The use of credit cards in the modern era is increasing. Therefore, it is necessary to prevent it with the use of technology such as address verification systems (AVS), card verification methods (CVM), and personal identification Numbers (PIN). Dataset analysis needs to be carried out to analyze the history of transactions that have been carried out. In the fraud detection dataset, it can be seen that there are attributes that cause data imbalance. Class imbalance in a dataset is a significant problem in machine learning that can affect overall model performance. The number of majority samples is more significant in one class than the number of minority classes. This research used an oversampling approach using a combination of smote and tomek-link. The focus of this research is card fraud classification. Detection of imbalanced datasets or imbalanced classes is carried out using the Naive Bayes method as a classification algorithm. In addition, a combination of resampling techniques is also applied to overcome imbalanced classes in this dataset through the SMOTETomek approach. SMOTETomek is a method that reduces the number of samples by considering two adjacent data from the minority and majority classes. Meanwhile, from the problems above, the results of the performance of Naïve Bayes, which experienced issues with data imbalance in this study, a resampling method was proposed in the hope of improving the performance of the Naïve Bayes algorithm and in the results of the AUC ROC curve, the SMOTETomek method could improve the performance of the Naïve Bayes algorithm. The

 ISSN:2302-8149
e-ISSN:2540-9719

higher the ROC score. -AUC, the better the model performance in terms of its ability to differentiate between two classes, but the accuracy results do not experience a significant change.
***Keywords**: fraud detection, sampling, naive Bayes*

1 Pendahuluan

Penipuan sudah ada sejak manusia dilahirkan dan dapat terjadi dalam berbagai bentuk kejahatan yang tidak terbatas. Penipuan merupakan tindakan disengaja yang dilakukan oleh individu maupun kelompok untuk memperoleh keuntungan finansial yang signifikan. Selain itu perkembangan teknologi juga dapat meningkan kasus penipuan misalnya melalui e-commerce[1] dan informasi pada kartu kredit. Penipu serig memanfaatkan kelemahan dari sistem keamanan atau kurangnya kewaspadaan pihak korban untuk melancarkan aksinya. Dampak dari penipuan tidak hanya merugikan finansial tetapi juga dapat merusak reputasi korban baik individu maupun perusahaan.

Penggunaan kartu kredit di jaman modern semakin meningkat oleh karna itu perlu pencegahan dengan Teknologi yang telah digunakan seperti *Address Verification Systems* (AVS), *Card Verification Method* (CVM) dan Nomor Identifikasi Pribadi (PIN). Analisa dataset perlu dilakukan untuk mengalisa histori transaksi yang sudah perna dilakukan, akan tetapi dalam analisa dataset perlu adanya metode data mining [2]untuk mengekstrak data menjadi pengetahuan. Dalam dataset *fraud detection*[3] terlihat adanya *attribut* yang mengalami *imbalance* data[1][4][5]. Penelitian yang dilakukan oleh [6] yang mengusulkan *customized Bayesian Network Classifier* untuk *fraud detection*, penelitian dilakukan perbaikan algortima yaitu dengan melakukan customisasi pada metode BNC dengan mengusulkan *Hyper-Heuristic Evolutionary Algorithm* (HHEA).

Imbalance class dalam dataset merupakan masalah signifikan dalam *machine learning* yang dapat memepengaruhi kinerja model secara keseluruhan, dalam satu kelas jumlah sampel mayoritas memiliki jumlah lebih banyak dari jumlah kelas minoritas. Maka model klasifikasi cenderung ke arah kelas mayoritas dan menghasilkan kinerja yang buruk dalam mengidentifikasi kelas minoritas. Masalah ini dapat menyebabkan kinerja yang buruk dalam memprediksi maupun klasifikasi class minoritas dan menyebabkan penurunan akurasi secara keseluruhan.

Pada penelitian yang diusulkan oleh Xu dkk[7] fraud detection menggunakan *deep boosting decision trees*, pendekatan ini menggabungkan teknik *gradient boosting* dan *neural network* kombinasi dari metode tradisional dan *deep learning* menghasilkan model struktur *node decision tree* pada *neural network* yang lebih efektif dan hasil kombinasi dapat meningkatkan kinerja algoritma serta dapat menanggani dataset yang tidak seimbang *(imbalance class)*. Pada penelitian yang diusulkan oleh Zhu dkk[8] mengatasi masalah klasifikasi imbalance data menggunakan extreem learning machine, penerapanya dilakukan pembobotan parameter. Optimasi *Weighted Extreme Learning Machine* diusulkan comparasi mutasi probabilitas dengan metode *PSO, GA* dan *self learning deadline*.

Maka diperlukan teknik penanganan *Imbalance class* untuk menangani distribusi data yang tidak seimbang antara *class* mayoritas dan class minoritas agar data menjadi seimbang[9]. Penelitian yang dilakukan oleh SAkila dkk[10] untuk mengatasi data *imbalance class* pada data *fraud detection* diusulkan metode bayesian dan teknik ensemble, gabungan dari kedua metode ini di namakan *Risk Induced Bayesian Inference Bagging (RIBIB).* Bagging diharapkan mampu meningkat kinerja bayesian untuk melakukan probabilitas prediksi resiko transaksi. Ada beberapa teknik yang dapat digunakan untuk menangani *imbalance data* yaitu teknik *oversampling* dan *undersampling* [11][12].

Penelitian yang dilakukan oleh Lutfi dkk[12] untuk menangani dataset yang tidak seimbang (*imbalance class*) dapat diseimbangkan menggunakan metode seperti *undersampling*, *oversampling*, dan *SMOTE*. Metode undersampling bekerja dengan cara mengurangi jumlah sampel dari class mayoritas sehingga lebih seimbang dengan class minoritas. Meskipun efektif dalam beberapa kasus, undersampling dapat menyebabkan hilangnya beberapa informasi dari class mayoritas. Di sisi lain, metode oversampling meningkatkan jumlah sampel dari class minoritas dengan cara menduplikasi data yang ada.

Namun pada penelitian ini digunakan pendekatan *oversampling* dengan menggunakan kombinasi *smote* dan *tomek-link*[13]. Fokus pada penelitian ini adalah klasifikasi *card fraud detection* terhadap dataset yang tidak seimbang atau kelas yang imbalanced. Algoritma klasifikasi yang digunakan adalah *Naive Bayes*. Pemilihan *Naive Bayes* didasarkan pada kesesuaian algoritma ini dengan atribut yang

terdapat dalam dataset, serta tingkat akurasinya yang tinggi berdasarkan penelitian sebelumnya oleh Menzies [14], Lessman [15] dan juga Putri & Friyendie [16]. Selain menerapkan algoritma *Naive Bayes* sebagai pengklasifikasi, juga diterapkan teknik resampling untuk mengatasi kelas yang tidak seimbang dalam dataset dengan pendekatan *SMOTETomek*. *SMOTETomek* adalah metode yang mengurangi sampel dengan mempertimbangkan dua data yang berdekatan antara kelas minoritas dan kelas mayoritas.

Penelitian ini akan fokus pada penanganan *imbalance data* pada dataset *fraud detection* untuk mengklasifikasi kemungkinan pemalsuan kartu kredit dengan menggunakan teknik *SMOTETomek.* Dalam penelitian ini diusulkan metode resampling *SMOTETomek* diharapkan bisa menyelesaikan permasalahan ketidak seimbangan *class*, sedangkan algoritma klasifikasi *Naïve Bayes* diusulkan untuk deteksi kejahatan atau pemalsuan data traksansi bank.

2 Tinjauan Literatur

Menangani data tidak seimbang pada dataset *fraud detection* ada dua cara yaitu dengan teknik *RUS* dan *ROS*, pada penelitian ini digunakan teknik *ROS* salah satunya dengan menggunakan metode *SMOTETomek* untuk menanggani data tidak seimbang dan model naïve bayes digunakan sebagai klasifikasi, ada beberapa peneliti yang sudah menggunakan metode tersebut salah satunya yang dilakukan oleh J. F. Díez-Pastor [17] diusulkan pendekatan baru untuk mengklasifikasikan dua kelas dataset yang tidak seimbang yang disebut *Random Balance*. Setiap anggota *ensemble Random Balance* dilatih dengan data latih dan ditambah dengan contoh buatan yang diperoleh dengan menggunakan *SMOTE*. Pembaruan ini dalam pendekatan ini yaitu dengan adanya proporsi kelas untuk setiap anggota ansambel dipilih secara acak.

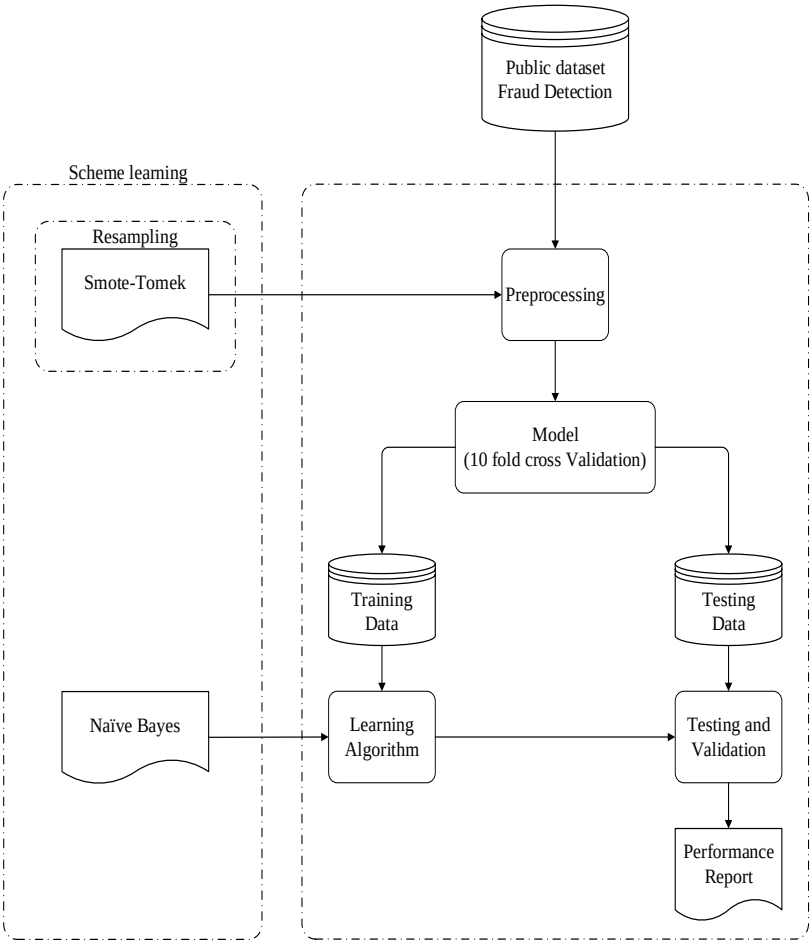
Penelitian yang dilakukan oleh Xu dkk[18] dilakukan analisa secara menyeluruh untuk klasifikasi data yang tidak seimbang dalam diagnosis medis, dalam penelitian ini mengusulkan algoritma pengambilan sampel hibrida yang disebut *RFMSE*, yang menggabungkan orientasi Misklasifikasi Teknik pengambilan sampel berlebih minoritas sintetis (*M-SMOTE*) dan tetangga terdekat yang diedit (*ENN*) berdasarkan *Random Forest* (*RF*). *M-SMOTE* digunakan untuk meningkatkan jumlah sampel pada kelas minoritas, sedangkan tingkat *over-sampling M-SMOTE* adalah tingkat kesalahan klasifikasi dari RF. Kemudian *ENN* digunakan untuk menghilangkan noise dari sampel mayoritas.

Dan penelitian yang dilakukan oleh Jiang dkk [19] mengusulkan *CSAWNB* untuk meningkatkan klasifikasi Naïve Bayes dengan memperkenalkan atribut bobot spesifik kelas yang dioptimalkan. Pendekatan ini membuktikan bahwa dengan sedikit modifikasi pada model dasar, meningkatkan performa yang signifikan.

3 Metode Penelitian

Pada penelitian ini diusulkan metode Resampling pada dataset *card fraud detection* Untuk menangani masalah data yang tidak seimbang. Teknik resampling yang digunakan adalah *SMOTETomek* dan algoritma naive bayes sebagai metode klasifikasi. Proses validasi menggunakan metode *10-fold cross validation*, dan hasil pengukuran dilakukan dengan *confusion matrix*. Berikut kerangka kerja metode yang diusulkan ditunjukkan Gambar 1.

 ISSN:2302-8149
e-ISSN:2540-9719

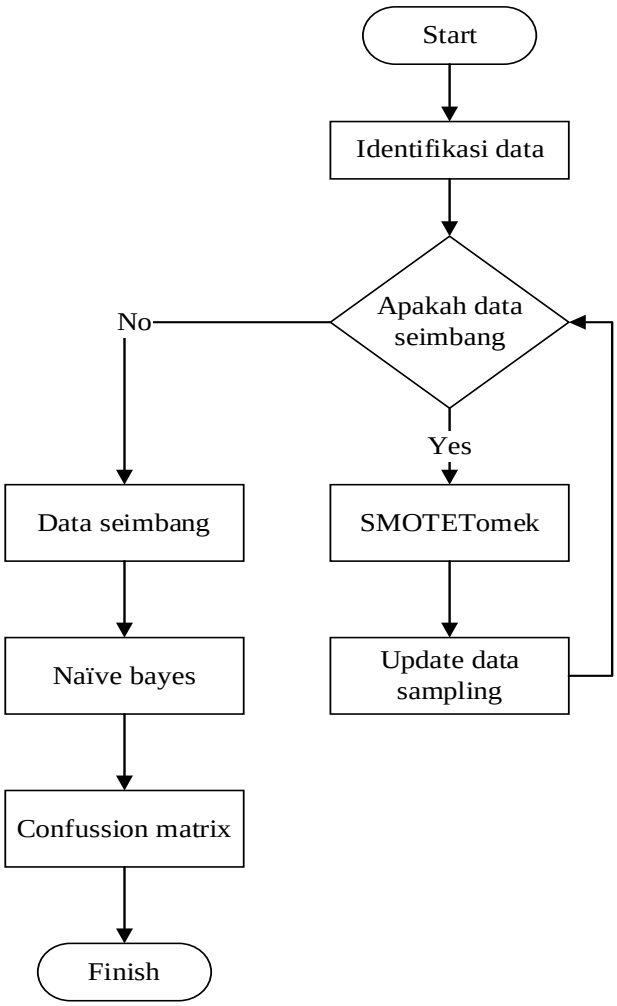


Gambar 1. Kerangka kerja metode yang diusulkan

Pada metode yang diusulkan, dataset yang diinput dibagi menjadi dua bagian utama, yaitu data latih (training data) dan data uji (testing data). Pembagian ini bertujuan untuk memastikan bahwa model yang dikembangkan dapat diuji secara efektif dan objektif. Sebagian dari dataset dialokasikan sebagai data uji, yang digunakan untuk mengukur kinerja model setelah dilatih. Sisanya digunakan sebagai data latih, yang akan digunakan dalam proses pembelajaran model. Kemudian dilakukan praproses dengan menggunakan *SMOTETomek* untuk mendapatkan hasil terbaik.

Proses praproses ini sangat penting untuk mendapatkan hasil terbaik dalam klasifikasi. Dengan menggunakan kombinasi *SMOTETomek*, dataset yang dihasilkan menjadi lebih seimbang dan bersih, yang pada gilirannya meningkatkan akurasi dan kinerja model klasifikasi yang akan dibangun. Model yang dihasilkan dari data yang telah diproses diharapkan bisa memberikan dampak prediksi yang signifikan atau lebih akurat. Berikut ini flowchart metode yang diusulkan ditunjukan Gambar 2.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 2. *flowchart* metode yang diusulkan

Algoritma Naive Bayes digunakan untuk meningkatkan hasil pada proses klasifikasi dan prediksi dataset dalam penelitian ini. Metode yang diusulkan ini bertujuan untuk menghasilkan data probabilitas terdekat sehingga dapat menunjukkan kinerja yang lebih baik, diukur dengan metrik seperti f-measure. Berikut adalah tahapan proses klasifikasi dengan Naive Bayes dalam penelitian ini:

1. Memulai dengan mengidentifikasi data.
   Proses klasifikasi dimulai dengan mengidentifikasi dataset yang akan digunakan. Identifikasi ini melibatkan pengecekan kelengkapan data, format data, dan distribusi kelas dalam dataset. Tujuan dari tahap ini adalah memastikan bahwa data yang akan diproses memenuhi syarat dan bebas dari kesalahan atau kekurangan.
2. Menyeimbangkan dataset dengan metode resampling dan melakukan update data.
   Setelah identifikasi data, langkah selanjutnya adalah memeriksa keseimbangan data. Jika data tidak seimbang, diterapkan metode *resampling* menggunakan teknik *SMOTETomek*. SMOTE (*Synthetic Minority Over-sampling Technique*) digunakan untuk meningkatkan jumlah data kelas sampel dari kelas yang minoritas dengan cara membuat contoh-contoh baru secara *sintesis*. Setelah itu, Tomek Links digunakan untuk membersihkan data yang berlebihan dan menghapus pasangan data yang berdekatan tetapi memiliki label yang berbeda, sehingga menghasilkan dataset yang lebih seimbang dan berkualitas.
3. *Update* data setelah *resampling*.
   Setelah menerapkan *SMOTETomek*, data diperbarui untuk mencerminkan hasil *resampling*. Data yang telah diperbarui ini kemudian digunakan untuk proses pembelajaran model. Dengan data yang lebih seimbang, model yang dihasilkan diharapkan memiliki kinerja yang lebih baik dalam mendeteksi kecurangan.
4. Melakukan proses *learning* dengan algoritma naive bayes dengan menghitung jumlah *Mean*

 ISSN:2302-8149
e-ISSN:2540-9719

dan standar deviasi setiap parameter.

5. Melakukan validasi dengan menggunakan *confusion matrix*.
   Confusion matrix akan membantu dalam mengevaluasi kinerja model dengan menghitung metrik-metrik seperti akurasi, *presisi*, *recall*, dan *f-measure*.

3.1 *SMOTETomek*

Adalah teknik *resampling* yang bertujuan untuk mengatasi data yang mengandung noise dan masalah ketidakseimbangan kelas. metode ini menggunakan teknik apabila tetangga terdekat ditemukan memiliki kelas yang berbeda dari data yang dianalisa maka data kelas mayoritas dihapus. Sebagai contoh, jika terdapat dua sampel data x dan z dari kelas yang berbeda, dengan d(x, z) disebut *Tomek link*. jika tidak ada sampel z* yang membuat d(x, z*) < d(x, z) [20]atau d(z, x*) < d(x, z). Jika dua sampel membentuk *Tomek-Links*, salah satu atau keduanya dianggap sebagai data *noise* atau berada di batas klasifikasi.

3.2 *Teorema Naive Bayes*

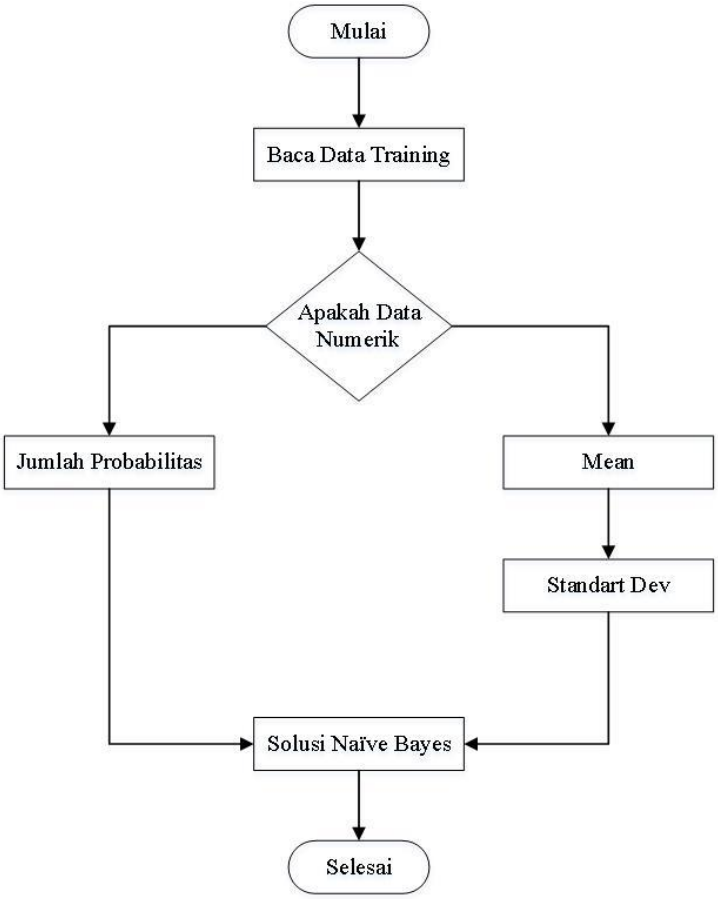
Merupakan algoritma klasifikasi sederhana yang menghitung probabilitas dan statistic yang diemukakan oleh ilmuwan Inggris Thomas Bayes[21]. Berikut persamaan teorema Naive Bayes :

$$P(C|F) = \frac{P(C).P(F|C)}{P(F)} \quad (1)$$

Dari rumus persamaan 1 tersebut dapat dipaparkan peluang masuknya objek dengan atribut tertentu dalam kelas C (*posterior*) adalah peluang munculnya kelas C (seringkali disebut prior), dikali dengan peluang kemunculan atribut objek pada kelas C (disebut *likelihood*), dibagi dengan peluang kemunculan atribut objek secara global (disebut evidence). Sehingga diperoleh rumus persamaan 2 sebagai berikut :

$$Posterior = \frac{Prior.likelihood}{evidence} \quad (2)$$

Berikut ini alur proses klasifikasi *Naive Bayes* :



Gambar 3. Alur proses klasifikasi naive bayes

 ISSN:2302-8149
e-ISSN:2540-9719

Pada Gambar 3 dijelaskan sebelum dilakukan pengolahan data tahapan proses naive bayes harus dilakukan terlebih dahulu untuk mengambarkan proses pengolahan data dengan tujuan untuk memperjelas pembuatan program yang akan digunakan untuk mengukur kinerja yang diusulkan. Berikut langkah naive bayes:

1. Mulai siapkan data
2. Hitung *mean* dan *standev*
3. Hitung probabilitas *prior*
4. Hitung *likelihood* setiap fitur
5. Hitung probabilitas *posterior*
6. Klasifikasi

3.3 *Matrix Confusion*

Kinerja pengukuran model klasifikasi sering kali dilakukan dengan menggunakan matriks kebingungan *(confusion matrix)*, yang terdiri dari empat istilah utama yang menggambarkan hasil dari sistem pengelompokan :

1. *True Negative (TN)*
   True Negative adalah jumlah instance negatif dalam dataset yang berhasil diidentifikasi dengan benar oleh model sebagai negatif. Artinya, model tidak salah dalam mengklasifikasikan data ini sebagai negatif. Misalnya, dalam konteks deteksi penipuan kartu kredit, TN adalah jumlah transaksi yang benar-benar sah (non-penipuan) dan model mengklasifikasikannya dengan tepat sebagai transaksi sah.
2. *False Positive (FP)*
   False Positive adalah jumlah instance negatif yang salah diidentifikasi oleh model sebagai positif. Ini berarti model mengklasifikasikan data yang sebenarnya tidak mencurigakan sebagai data penipuan. Dalam contoh deteksi penipuan, FP adalah transaksi sah yang salah diidentifikasi oleh model sebagai transaksi penipuan. Ini adalah kesalahan yang dapat menyebabkan ketidaknyamanan bagi pengguna yang sah.
3. *True Positive (TP)*
   True Positive adalah jumlah instance positif dalam dataset yang berhasil diidentifikasi dengan benar oleh model sebagai positif. Dengan kata lain, ini adalah kasus di mana model dengan benar mendeteksi data yang sebenarnya mencurigakan sebagai data penipuan. Dalam deteksi penipuan kartu kredit, TP adalah jumlah transaksi penipuan yang benar-benar diidentifikasi sebagai penipuan oleh model.
4. *False Negative (FN)*
   False Negative adalah jumlah instance positif yang salah diidentifikasi oleh model sebagai negatif. Ini berarti model gagal mendeteksi data yang sebenarnya mencurigakan, mengklasifikasikannya sebagai data sah. Dalam deteksi penipuan, FN adalah transaksi penipuan yang salah diidentifikasi oleh model sebagai transaksi sah. Ini adalah kesalahan yang berpotensi berbahaya karena penipuan yang sebenarnya tidak terdeteksi dan dapat merugikan pengguna dan institusi keuangan.

Tabel 1 Confusion matrix

Kelas	Terklasifikasi Positif	Terklasifikasi Negatif
Positif	TP (*True Positif*)	FN (*False Negatif*)
Negatif	FP (*False Positif*)	TN (*True Negatif*)

Adapun rumus dari tabel diatas:

1. $Precision = \left(\frac{TP}{TP+FP}\right) \times 100\%$ (3)
2. $Recall = \left(\frac{TP}{TP+FN}\right) \times 100\%$ (4)
3. $Accuracy = \left(\frac{TP+TN}{TP+FP+FN+FP}\right) \times 100\%$ (5)

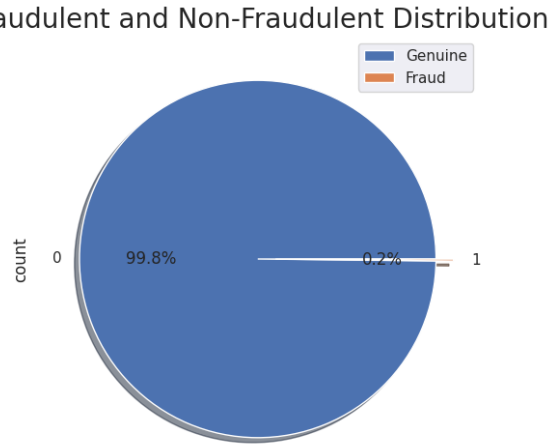
Dengan memahami keempat istilah dalam confusion matrix ini, kita dapat mengevaluasi kinerja model secara lebih mendalam. Model yang ideal adalah model yang memiliki jumlah *TP* dan *TN*

 ISSN:2302-8149
e-ISSN:2540-9719

yang tinggi, serta jumlah *FP* dan *FN* yang rendah. Penggunaan confusion matrix memungkinkan kita untuk melihat secara spesifik di mana model membuat kesalahan dan di mana ia berkinerja dengan baik, sehingga kita dapat mengambil langkah-langkah yang diperlukan untuk meningkatkan akurasi dan keandalannya. Evaluasi melalui confusion matrix juga membantu dalam menentukan trade-off antara sensitivitas *(recall)* dan spesifisitas *(precision)*, yang sangat penting dalam berbagai aplikasi seperti deteksi penipuan[22].

4 Hasil dan Pembahasan

Pada penelitian ini, kami menganalisis distribusi data transaksi untuk mendeteksi penipuan kartu kredit. Gambar 4 menunjukkan distribusi transaksi penipuan dan non-penipuan dalam dataset yang digunakan.



Gambar 4. Distribusi transaksi penipuan dan non-penipuan

Dalam penelitian ini, analisis distribusi data transaksi dilakukan untuk mendeteksi penipuan kartu kredit. Dari distribusi transaksi yang diilustrasikan, terlihat ketidakseimbangan yang sangat mencolok antara transaksi penipuan dan non-penipuan, di mana transaksi asli (*genuine*) mendominasi dengan 99,8%, sedangkan hanya 0,2% transaksi yang teridentifikasi sebagai penipuan. Ketidakseimbangan ini menjadi tantangan utama dalam pengembangan model klasifikasi, karena model memiliki kecenderungan untuk lebih mudah mengenali pola kelas mayoritas, yakni transaksi asli, dan mengabaikan pola kelas minoritas, yakni transaksi penipuan.

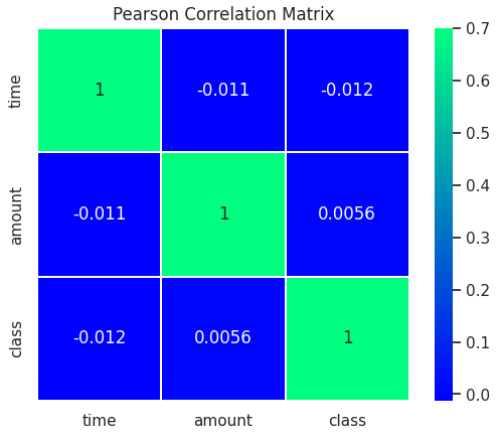
Ketika dihadapkan dengan data tidak seimbang seperti ini, metode klasifikasi sering kali memberikan performa yang bias terhadap kelas mayoritas. Oleh karena itu, solusi yang tepat dalam menangani ketidakseimbangan ini menjadi krusial. Pendekatan seperti resampling, baik itu oversampling maupun undersampling, dapat menjadi alternatif untuk memperbaiki distribusi kelas sebelum data dilatih pada model prediksi. Penelitian ini juga menganalisis distribusi waktu untuk kedua jenis transaksi, yang memberikan wawasan tambahan mengenai perbedaan pola antara transaksi asli dan penipuan. Dengan demikian, penelitian ini tidak hanya mengidentifikasi tantangan yang ada, tetapi juga memberikan dasar untuk pengembangan metode yang lebih efektif dalam mendeteksi anomali transaksi dalam data yang tidak seimbang, distribusi transaksi untuk kedua kategori tersebut, yang ditunjukkan pada Gambar 5.

ISSN:2302-8149
e-ISSN:2540-9719

Gambar 5. Distribusi waktu transaksi penipuan (kiri) dan transaksi non-penipuan (kanan)

Dari Gambar 5, terlihat bahwa transaksi penipuan *(Fraud)* terjadi pada berbagai waktu dengan pola yang tidak teratur, sedangkan transaksi asli *(Geneuine)* menunjukkan pola yang lebih teratur dengan puncak pada waktu-waktu tertentu. Hal ini dapat memberikan petunjuk tambahan dalam deteksi penipuan, dimana pola waktu transaksi dapat menjadi indikator penting.

Terakhir, kami menghitung korelasi antara berbagai fitur dalam dataset menggunakan matriks korelasi Pearson. Gambar 6 menunjukkan matriks korelasi Pearson untuk fitur *Time*, *Amount*, dan *Class.*

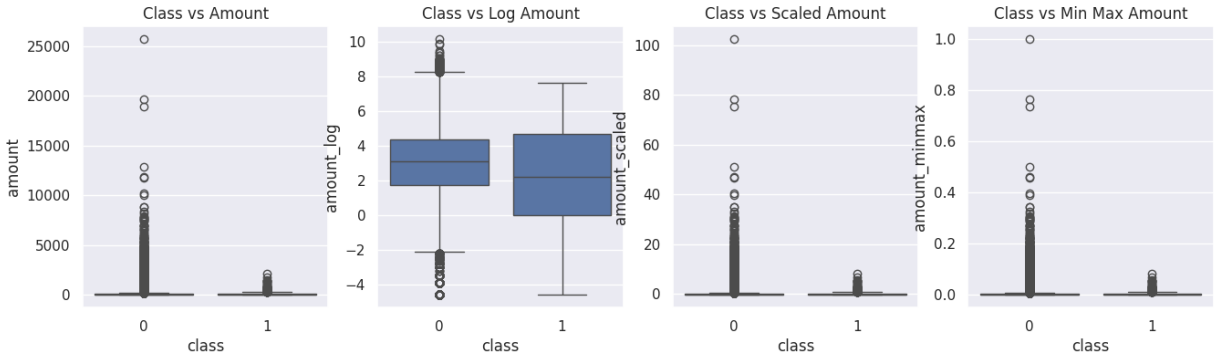


Gambar 6. Matriks korelasi pearson antara fitur *time*, *amount*, dan *class*

Dari Gambar 6, dapat dilihat bahwa tidak ada korelasi yang signifikan antara fitur-fitur tersebut. Nilai korelasi yang mendekati nol menunjukkan bahwa fitur-fitur tersebut tidak memiliki hubungan linear yang kuat satu sama lain, sehingga fitur-fitur ini dapat memberikan informasi yang berbeda untuk model deteksi penipuan.

Untuk memahami lebih lanjut perbedaan antara transaksi penipuan dan non-penipuan berdasarkan nilai transaksi, kami juga melakukan beberapa teknik rekayasa fitur dan visualisasinya dalam Gambar 7.

 ISSN:2302-8149
e-ISSN:2540-9719



Gambar 7. Rekayasa fitur untuk visualisasi nilai yang lebih baik. *class vs amount, class vs log amount, class vs scaled amount, class vs min max amount*

Terlihat bahwa penggunaan teknik rekayasa fitur seperti *Log Transformation* dan *Scaling* dapat membantu dalam visualisasi dan pemahaman data. Misalnya, visualisasi *Log Amo*unt menunjukkan perbedaan distribusi yang lebih jelas antara transaksi penipuan dan non-penipuan.

Gambar 8. Hasil ROC AUC metode naïve bayes

Gambar 9. Hasil ROC AUC metode NB+SMOTE tomek

Dari hasil Gambar 8 diatas menunjukkan bahwa metode naïve bayes kurang efektif apabila menangani data imbalance. Ketika class minoritas memiliki jumlah sampel yang jauh lebih sedikit daripada class mayoritas maka model klasifikasi cenderung ke arah kelas mayoritas dan menghasilkan kinerja yang buruk dalam mengidentifikasi kelas minoritas. Masalah ini dapat menyebabkan kinerja yang buruk dalam memprediksi maupun klasifikasi class minoritas dan menyebabkan penurunan akurasi secara keseluruhan. Sedangkan pada Gambar 9 diusulkan metode *SmoteTomek* untuk menanggani class imbalance pada naive bayes, hasil metode pengujian yang dilakukan dapat dilihat pada gambar 9 bahwa metode *SmoteTomek* sangat efektif digunakan pada dataset *imbalance*.

Tabel 2. Perbandingan hasil kinerja algoritma

	NB	NB+SMOTETomek
Accuracy	0.97	0.97
AUC	0.90	**0.92**
Precision	0.05	0.06
Recall	0.84	0.86
F1	0.10	0.11

Tabel 2 membandingan kinerja antara algoritma Naïve Bayes (NB) dan kombinasi Naïve Bayes dengan teknik *resampling SMOTETomek*. Dari tabel tersebut, dapat diambil beberapa kesimpulan terkait kinerja kedua model dalam menangani ketidakseimbangan data transaksi penipuan kartu kredit yaitu:

1. Akurasi (*Accuracy*): Kedua model menunjukkan hasil akurasi yang sama, yaitu 0,97. Ini menunjukkan bahwa meskipun metode *SMOTETomek* diterapkan, hasil prediksi benar secara keseluruhan (benar untuk kedua kelas) tidak berubah secara signifikan.

2. AUC (*Area Under Curve*): AUC dari model Naïve Bayes murni adalah 0,90. Sedangkan dengan penambahan *SMOTETomek*, AUC meningkat menjadi 0,92. Hal ini menunjukkan peningkatan kemampuan model untuk membedakan antara kelas penipuan dan non-penipuan setelah mengatasi ketidakseimbangan data.
3. Precision: Naïve Bayes tanpa *SMOTETomek* menghasilkan *precision* yang rendah, hanya 0,05 sementara dengan *SMOTETomek* naik menjadi 0,06. *Precision* yang rendah mengindikasikan bahwa proporsi prediksi kelas penipuan yang benar sangat kecil, menandakan banyaknya *false positive*.
4. Recall: *Recall* meningkat dari 0,84 menjadi 0,86 setelah penerapan *SMOTETomek*, yang menunjukkan bahwa model dengan *SMOTETomek* lebih baik dalam menemukan transaksi penipuan.
5. F1-Score: Naïve Bayes awalnya memiliki F1 score 0,10, dan setelah menggunakan *SMOTETomek*, nilainya sedikit naik menjadi 0,11. *F1-Score* rendah menegaskan bahwa kinerja keseluruhan dalam mendeteksi kelas minoritas (penipuan) masih jauh dari optimal, meskipun ada peningkatan.

Dengan demikian, penelitian ini menunjukkan bahwa *resampling* dapat memperbaiki performa model dalam mendeteksi kelas minoritas, meskipun ada batasan dalam hal *precision* yang rendah. Hal ini mengindikasikan bahwa, untuk mencapai kinerja yang lebih baik, mungkin diperlukan eksplorasi lebih lanjut terhadap metode *imbalance* yang lebih kompleks atau kombinasi dengan algoritma yang lebih baik.

5 Kesimpulan

Imbalance class merupakan masalah pada dataset dalam bidang *machine learning* yang dapat mempengaruhi kinerja dari model klasifikasi yang diusulkan, pada penelitian ini diusulkan pendekatan *oversampling* dengan menggunakan kombinasi *smote* dan *tomek-link* atau disebut *SmoteTomek* yang digunakan untuk mengatasi masalah data *imbalance* pada dataset *fraud detection* sebelum dilakukan proses training pada model klasifikasi naive bayes. Hasil penelitian yang telah dilakukan dengan model yang diusulkan yaitu teknik *resampling* (*SMOTETomek*) dan naive bayes bahwa metode ini sangat baik hasil performance pada kurva AUC namun pada hasil akurasi tidak ada perubahan yang signifikan, hal ini menunjukan bawah untuk mengatasi masalah imbalance data sangat baik digunakan dengan model *SMOTETomek* namun tidak baik apabila digunakan untuk meningkatkan akurasi pada dataset fraud detection

Referensi

[1] B. Lebichot, Y.-A. Le Borgne, L. He-Guelton, F. Oblé, and G. Bontempi, “Deep-Learning Domain Adaptation Techniques for Credit Cards Fraud Detection,” no. January, pp. 78–88, 2020, doi: 10.1007/978-3-030-16841-4_8.

[2] A. D. Pozzolo, “Adaptive Machine Learning for Credit Card Fraud Detection Declaration of Authorship,” *Dr. - Univ. Libr. Bruxelles*, no. December, p. 199, 2015, [Online]. Available: https://www.ulb.ac.be/di/map/adalpozz/pdf/Dalpozzolo2015PhD.pdf%0Ahttp://www.ulb.ac.be/di/map/adalpozz/

[3] N. Ofek, L. Rokach, R. Stern, and A. Shabtai, “Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem,” *Neurocomputing*, vol. 243, pp. 88–102, 2017, doi: 10.1016/j.neucom.2017.03.011.

[4] H. Huang, B. Liu, X. Xue, J. Cao, and X. Chen, “Imbalanced credit card fraud detection data: A solution based on hybrid neural network and clustering-based undersampling technique,” *Appl. Soft Comput.*, vol. 154, p. 111368, 2024, doi: 10.1016/J.ASOC.2024.111368.

[5] G. Tong and J. Shen, “Financial transaction fraud detector based on imbalance learning and graph neural network,” *Appl. Soft Comput.*, vol. 149, p. 110984, Dec. 2023, doi:

 ISSN:2302-8149
e-ISSN:2540-9719

10.1016/J.ASOC.2023.110984.

[6] A. G. C. de Sá, A. C. M. Pereira, and G. L. Pappa, "A customized classification algorithm for credit card fraud detection," *Eng. Appl. Artif. Intell.*, vol. 72, no. March, pp. 21–29, 2018, doi: 10.1016/j.engappai.2018.03.011.

[7] B. Xu, Y. Wang, X. Liao, and K. Wang, "Efficient fraud detection using deep boosting decision trees," *Decis. Support Syst.*, vol. 175, no. 28, p. 114037, 2023, doi: 10.1016/j.dss.2023.114037.

[8] H. Zhu, G. Liu, M. Zhou, Y. Xie, A. Abusorrah, and Q. Kang, "Optimizing Weighted Extreme Learning Machines for imbalanced classification and application to credit card fraud detection," *Neurocomputing*, vol. 407, pp. 50–62, 2020, doi: 10.1016/j.neucom.2020.04.078.

[9] M. A. Islam, M. A. Uddin, S. Aryal, and G. Stea, "An ensemble learning approach for anomaly detection in credit card data with imbalanced and overlapped classes," *J. Inf. Secur. Appl.*, vol. 78, no. October, p. 103618, 2023, doi: 10.1016/j.jisa.2023.103618.

[10] S. Akila and U. Srinivasulu Reddy, "Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection," *J. Comput. Sci.*, vol. 27, pp. 247–254, 2018, doi: 10.1016/j.jocs.2018.06.009.

[11] S. B. Belhaouari, A. Islam, K. Kassoul, A. Al-Fuqaha, and A. Bouzerdoum, "Oversampling techniques for imbalanced data in regression," *Expert Syst. Appl.*, vol. 252, no. PB, p. 124118, 2024, doi: 10.1016/j.eswa.2024.124118.

[12] M. Lutfi, A. T. Arsanto, M. F. Amrulloh, and U. Kulsum, "Penanganan Data Tidak Seimbang Menggunakan Hybrid Method Resampling Pada Algoritma Naive Bayes Untuk Software Defect Prediction," *Informatics J.*, vol. 8, no. 2, 2023.

[13] E. P. Kondy, S. Siswanto, and N. Ilyas, "Data Balancing Approach Using Combine Sampling on Sentiment Analysis With K - Nearest Neighbor," *Sist. J. Sist. Inf.*, vol. 13, pp. 1836–1851, 2024.

[14] T. Menzies, Z. Milton, B. Turhan, B. Cukic, Y. Jiang, and A. Bener, "Defect prediction from static code features: Current results, limitations, new approaches," *Autom. Softw. Eng.*, vol. 17, no. 4, pp. 375–407, 2010, doi: 10.1007/s10515-010-0069-5.

[15] S. Lessmann, B. Baesens, C. Mues, and S. Pietsch, "Benchmarking classification models for software defect prediction: A proposed framework and novel findings," *IEEE Trans. Softw. Eng.*, vol. 34, no. 4, pp. 485–496, 2008, doi: 10.1109/TSE.2008.35.

[16] S. A. Putri, "Prediksi Cacat Software Dengan Teknik Sampel Dan Seleksi Fitur Pada Bayesian Network," *J. Kaji. Ilm.*, vol. 19, no. 1, p. 17, 2019, doi: 10.31599/jki.v19i1.314.

[17] J. F. Díez-Pastor, J. J. Rodríguez, C. García-Osorio, and L. I. Kuncheva, "Random Balance: Ensembles of variable priors classifiers for imbalanced data," *Knowledge-Based Syst.*, vol. 85, no. May, pp. 96–111, 2015, doi: 10.1016/j.knosys.2015.04.022.

[18] Z. Xu, D. Shen, T. Nie, and Y. Kou, "A hybrid sampling algorithm combining M-SMOTE and ENN based on Random forest for medical imbalanced data," *J. Biomed. Inform.*, vol. 107, no. June, p. 103465, 2020, doi: 10.1016/j.jbi.2020.103465.

[19] L. Jiang, L. Zhang, L. Yu, and D. Wang, "Class-specific attribute weighted naive Bayes," *Pattern Recognit.*, vol. 88, pp. 321–330, 2019, doi: 10.1016/j.patcog.2018.11.032.

[20] T. T. H. Le, Y. Shin, M. Kim, and H. Kim, "Towards unbalanced multiclass intrusion detection

 ISSN:2302-8149
e-ISSN:2540-9719

with hybrid sampling methods and ensemble classification," *Appl. Soft Comput.*, vol. 157, no. February, p. 111517, 2024, doi: 10.1016/j.asoc.2024.111517.

[21] C. Cassidy, "Parameter tuning Naïve Bayes for automatic patent classification," *World Pat. Inf.*, vol. 61, no. March, p. 101968, 2020, doi: 10.1016/j.wpi.2020.101968.

[22] M. A. Latief, L. R. Nabila, W. Miftakhurrahman, S. Ma, H. Tantyoko, and M. A. Latief, "Handling Imbalance Data Using Hybrid Sampling SMOTE-ENN in Lung Cancer Classification Corresponding Author :," vol. 3, no. 1, pp. 11–18, 2024, doi: 10.30812/IJECSA.v3i1.3758.