Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

Optimalisasi Akurasi Algoritma Naïve Bayes Dengan Metode Syntetic Minority Oversampling Technique (Smote) Pada Data Numerik

Hizbul Izzi[1*], Arief Setyanto[2], Anggit Dwi Hartanto[3]
[1,2,3] Informatics, Universitas Amikom Yogyakarta Indonesia
*hizbul_izzi@students.amikom.ac.id

Abstrak

Pada penelitian ini akan mengklasifikasikan data numerik yaitu data loan yang diambil dari Kaggle. Data yang digunakan berjumlah 9578 dataset yang meliputi kelas data dengan peminjam dapat menyelesaikan kredit sebanyak 8045 record dan peminjaman yang tidak dapat menyelesaikan kredit sebanyak 1533 record. Dari jumlah data tersebut terdapat ketidakseimbangan kelas sehingga perlu dilakukan penyeimbangan agar mendapatkan hasil klasifikasi yang lebih akurat. Tujuan dari penelitian ini adalah meningkatkan akurasi algoritma Naïve Bayes dalam mengklasifikasikan data numerik. Penipuan dalam transaksi keuangan adalah contoh kasus data tidak seimbang, di mana jumlah transaksi yang sah jauh lebih banyak dibandingkan yang merupakan penipuan. Optimalisasi akurasi pada kelas minoritas (penipuan) sangat penting untuk menghindari kerugian. Metode yang digunakan untuk meningkatkan akurasi algoritma yaitu Syntetic Minority Oversampling Tecnique (SMOTE) dengan cara meng-over sampling minoritas dataset. Selain itu juga menggunakan metoe K-Fold Cross Validation untuk mengevaluasi performa dari proses algoritma yang digunakan. Praproses data dilakukan untuk membersihkan data dari nilai-nilai yang hilang dan tidak valid serta menormalisasi data supaya semua fitur berada dalam skala yang sama dan sesuai untuk analisis klasifikasi. Berdasarkan hasil analisis yang dilakukan, sebelum penerapan SMOTE kemampuan model mengenali kelas minoritas sebesar 16.1%, sedangkan setelah penerapan SMOTE kemampuan model mengenali kelas minoritas menjadi 48.8%. selain itu juga sebelum penerapan SMOTE model mampu memprediksi kelas minoritas dengan benar sebanyak 10 kasus sedangkan setelah penerapan SMOTE, model mampu memprediksi kelas minoritas dengan benar sebanyak 102 kasus. Sehingga dapat disimpulkan bahwa teknik SMOTE mampu meningkatkan kemampuan model.

Kata kunci : Data Numerik, Klasifikasi, K-Fold Cross Validation, Naïve Bayes, SMOTE

Abstract

This research will classify numerical data, namely loan data taken from Kaggle. The data used amounted to 9578 datasets which included data classes with borrowers able to complete credit as many as 8045 records and loans that could not complete credit as many as 1533 records. From the amount of data there is an imbalance of classes so it is necessary to do balancing in order to get more accurate classification results. The purpose of this research is to improve the accuracy of the Naïve Bayes algorithm in classifying numerical data. Fraud in financial transactions is an example of a case of imbalanced data, where the number of legitimate transactions is much greater than those that are fraudulent. Optimizing accuracy in minority (fraud) classes is very important to avoid losses. The method used to improve the accuracy of the algorithm is the Synthetic Minority Oversampling Technique (SMOTE) by over sampling the minority of the dataset. In addition, it also uses the K-Fold Cross Validation method to evaluate the performance of the algorithm process used. Data preprocessing is done to clean the data from missing and invalid values and normalize the data so that all features are on the same scale and suitable for classification analysis. Based on the results of the analysis conducted, before the application of SMOTE the model's ability to recognize minority classes was 16.1%, while after the application of SMOTE the model's ability to recognize minority classes became 48.8%. besides that, before the application of SMOTE the model was able to predict the minority class correctly in 10 cases while after the application of SMOTE, the model was able to predict the minority class correctly in 102 cases. So it can be concluded that the SMOTE technique is able to improve the ability of the model..

Keywords : *Classification, K-Fold Cross Validation, Naive Bayes, SMOTE.*

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

1. Pendahuluan

Klasifikasi adalah tugas dasar dari analisis data yang berfungsi memberikan label kelas untuk kasus yang dijelaskan oleh satu set atribut. Sehingga tujuan dari klasifikasi adalah kebenaran dalam memprediksi sebuah nilai. Beberapa algoritma yang dapat digunakan untuk klasifikasi adalah Naïve Bayes, Decision tree, Artificial Neural Network, dan lain sebagainya[1]. Hasil klasifikasi akan sangat berpengaruh jika terdapat data yang tidak seimbang. Ketidakseimbangan data adalah kondisi dimana jumlah sampel di satu kelas jauh lebih banyak dibandingkan dengan kelas lainnya. Untuk mengatasi masalah ketidakseimbangan ini, berbagai teknik telah dikembangkan salah satunya adalah *Synthetic Minority Oversampling Technique* (SMOTE)[2].

Dalam era perkembangan teknologi yang sangat cepat, analisis data numerik menjadi sangat penting dalam berbagai bidang, seperti bisnis, medis, dan ilmu pengetahuan. Data yang digunakan pada penelitian ini adalah data *loan* yang diperoleh dari Kaggle. Penelitian ini menggunakan 9578 dataset yang meliputi kelas peminjam dapat menyelesaikan kredit sebanyak 8045 record dan peminjam yang tidak dapat menyelesaikan kredit sebanyak 1533 record.

Sehingga masalah masalah dalam penelitian ini adalah adanya ketidakseimbangan kelas antara kelas data dengan peminjam yang dapat menyelesaikan kredit dan yang tidak dapat menyelesaikan kredit. Sehingga perulu dilakukan penyeimbangan data untuk memperoleh hasil klasifikasi yang optimal.

Metode *oversampling* seperti SMOTE digunakan untuk menangani ketidakseimbangan kelas dan meningkatkan kinerja model klasifikasi pada data yang akan digunakan. Terdapat metode lain yang digunakan untuk menyeimbangkan data seperti, *undersampling* dan beberapa algoritma lain seperti *Adaboost* dan *Bagging* dapat digunakan sebagai alternatif. Metode SMOTE menjadi salah satu metode yang efektif dalam mengatasi ketidakseimbangan kelas pada model klasifikasi. Karena SMOTE dapat menghasilkan data sintetis yang mirip dengan data minoritas, sehingga data sintetis tersebut dapat digunakan sebagai data latih yang lebih representatif. Dengan demikian, model klasifikasi dapat lebih efektif dalam mengklasifikasi data minoritas[3].

Pada penelitian ini peneliti juga menambahkan penggunaan metode *K-Fold Cross Validation. K-Fold Cross Validation* berguna untuk mengevaluasi performa dari algoritma yang digunakan

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

2. Tinjauan Pustaka

2.1. Penelitian Terkait

- Penelitian yang dilakukan oleh Shujuan Wang, Yuntao Dai dan Jihong Shen dan Jingxue dengan judul *Research on expansion and classification of imbalanced data based on SMOTE algorithm*. Penelitian tersebut membahasa mengenai perluasan dan klasifikasi data tidak seimbang berdasarkan algoritma SMOTE yang fokus pada mengatasi tantangan yang ditimbulkan oleh dataset tidak seimbang dalam pembelajaran mesin. Hasil eksperimen menunjukkan bahwa algoritma SMOTE yang ditingkatkan secara efektif mengatasi ketidakseimbangan kelas dan meningkatkan akurasi klasifikasi. Temuan ini menyoroti pentingnya pra-pemrosesan data dalam mengembangkan model pembelajaran mesin yang kuat untuk data tidak seimbang[4].
- Penelitian oleh Hairani, Anthony Anggrawan dan Dadang Priyanto dengan judul *Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link.* Penelitian tersebut bertujuan untuk menerapkan metode *Smote-Tomek Link* dan *Random Forest* dalam klasifikasi diabetes. Data yang digunakan adalah data diabetes yang diperoleh dari Kaggle sebanyak 768 data dengan delapan atribut input dan 1 output atribut sebagai kelas. Berdasarkan pengujian yang dilakukan dengan cara membagi data menggunakan validasi silang 10 kali lipat, algoritma *Random Forest* dengan *Smote-Tomek Link* mendapatkan akurasi, sensitivitas, presisi, dan skor F1 yang lebih baik dibandingkan dengan *Random Forest* dengan Smote. Algoritma Random Forest dengan Smote-Tomeklink memiliki 86,4% akurasi, sensitivitas 88,2%, presisi 82,3%, dan skor F1 85,1%. Dengan demikian, penggunaan Smote-Tomek Link dapat meningkatkan kinerja metode hutan acak berdasarkan akurasi, sensitivitas, presisi, dan skor F1[5].
- Penelitian yang dilakukan oleh Aliyah Kurniasih dan Khaira Isyara dengan judul penggunaan metode SMOTE pada naïve bayes gaussian untuk klasifikasi mahasiswa *drop out*. Penelitian ini bertujuan untuk mengklasifikasikan kasus mahasiswa yang akan di drop out atau tidak menggunakan metode Naïve Bayes Gaussian dengan teknik oversampling SMOTE untuk mengatasi imbalance class. hasil dari penelitian ini yaitu Sebelum oversampling, model klasifikasi memiliki akurasi sebesar 84%, sedangkan setelah oversampling, akurasi meningkat menjadi 86%. Hal ini menunjukkan bahwa terdapat peningkatan

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

besarnya nilai akurasi setelah penerapan oversampling menggunakan SMOTE sebesar 2%. Dapat disimpulkan bahwa penggunaan oversampling dengan SMOTE berhasil meningkatkan kemampuan model dalam mengklasifikasikan dengan benar jumlah tuple dalam data uji [6].

- Penelitian yang dilakukan oleh FaidhlulRahman dan Mustikasari dengan judul Optimalisasi Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Binning Dan *Synthetic Minority Oversampling Technique* (SMOTE). Penelitian ini bertujuan untuk mengoptimalkan prediksi kelulusan mahasiswa tepat waktu menggunakan metode Binning untuk mengelompokkan variabel ke dalam kategori diskrit dan Synthetic Minority Oversampling Technique (SMOTE) untuk mengatasi ketidakseimbangan kelas pada dataset. Hasil dari penelitian ini, pertama yaitu peningkatan performa dengan Teknik Binning dan SMOTE, implementasi teknik Binning dan Synthetic Minority Oversampling Technique (SMOTE) pada model Random Forest dan Decision Tree secara signifikan meningkatkan kinerja prediksi kelulusan mahasiswa tepat waktu, kedua yaitu efektivitas model Random Forest dan Decision Tree di mana hasil menunjukkan bahwa model Random Forest memiliki performa yang baik dalam mengklasifikasikan kelas mayoritas. Namun, ketika digabungkan dengan teknik SMOTE, model ini mampu meningkatkan kemampuan dalam mengklasifikasikan kelas minoritas dengan tingkat Recall yang lebih baik[7].
- Penelitian oleh Widya Amalia Putri Riswandha dengan judul Evaluasi Performa *Synthetic Minority Oversampling Technique* (Smote) Untuk Mengatasi Klasifikasi Data Tidak Seimbang Pada Metode K-Nearest Neighbor (KNN) Dan Support Vector Machine (SVM). Penelitian ini berfokus pada pengembangan metode oversampling baru untuk mengatasi masalah data tidak seimbang dalam diagnosis kesalahan mekanis. Algoritma MeanRadius-SMOTE yang diusulkan menggunakan radius rata-rata untuk meningkatkan kualitas data sintetis yang dihasilkan. Penulis mengevaluasi kinerja metode ini melalui berbagai eksperimen dan membandingkannya dengan teknik oversampling yang ada. Hasil eksperimen menunjukkan bahwa MeanRadius-SMOTE secara signifikan meningkatkan akurasi diagnosis kesalahan dibandingkan dengan metode oversampling tradisional. Temuan ini menunjukkan potensi besar dari pendekatan baru ini dalam meningkatkan keandalan diagnosis kesalahan mekanis pada data tidak seimbang[2].

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

2.2. Landasan Teori

[1] Klasifikasi

Klasifikasi (*Classification*) merupakan proses untuk menemukan sekumpulan model yang menjelaskan dan membedakan kelas-kelas data, sehingga model tersebut dapat digunakan untuk memprediksi nilai suatu kelas yang belum diketahui pada sebuah objek. Untuk mendapatkan model, kita harus melakukan analisis terhadap data latih (training set)[8]. Klasifikasi merupakan proses yang terdiri dari dua tahap, yaitu tahap pembelajaran dan tahap pengklasifikasian. Pada tahap pembelajaran, sebuah algoritma klasifikasi akan membangun sebuah model klasifikasi dengan cara menganalisis training data [9].

[2] Algoritma Naive Bayes

Naïve bayes adalah sebuah algoritma supervised learning berdasarkan teorema Bayes yang digunakan untuk memecahkan masalah klasifikasi dengan mengikuti pendekatan probabilistik. Naïve bayes dikemukakan oleh ilmuwan inggris Thomas Bayes, yaitu memprediksi peluang dimasa depan berdasarkan pengalaman sebelumnya sehingga dikenal sebagai Teorema Bayes [10]. Naive Bayes berpotensi baik untuk mengklasifikasikan data karena kesederhanaannya. Naive Bayes dirancang untuk dipergunakan dengan asumsi bahwa antar satu kelas dengan kelas yang lain tidak saling bergantung (independen). Pada klasifikasi Naive Bayes, proses pembelajaran lebih ditekankan pada mengestimasikan probabilitas. Keuntungan dari pendekatan Naive Bayes adalah pengklasifikasian akan mendapatkan nilai error yang lebih kecil ketika data set berjumlah besar [11]. Klasifikasi Naive Bayes terbukti memiliki akurasi dan kecepatan yang tinggi saat dipublikasikan kedalam basis data dengan jumlah yang besar.

[3] Imbalance Class

Ketidakseimbangan kelas (*imbalance class*) mengacu pada kondisi di mana jumlah sampel antara kelas-kelas yang berbeda dalam dataset tidak seimbang atau tidak proporsional. Ini berarti ada satu atau beberapa kelas yang memiliki jumlah sampel yang jauh lebih sedikit atau jauh lebih banyak daripada kelas lainnya. Masalah ketidakseimbangan kelas dapat mempengaruhi kinerja model pembelajaran mesin. Beberapa dampak yang mungkin terjadi adalah model dapat cenderung memprediksi secara dominan kelas mayoritas karena penyebaran yang tidak seimbang. Ini mengakibatkan kinerja model yang buruk dalam mengidentifikasi sampel dari kelas minoritas [12].

[4] SMOTE

SMOTE (*Synthetic Minority Over-sampling Technique*) ialah salah satu metode yang banyak digunakan untuk mengatasi ketidakseimbangan kelas dalam dataset. Teknik ini berkerja dengan cara mengambil sampel data baru. Jumlah data sampel yang diambil menyesuaikan dengan jumlah data minoritas[13]. Metode SMOTE terbagi menjadi beberapa langkah yakni de-noising, oversampling dan juga penyaringan [14]. Jika ketidakseimbangan kelas tersebut tidak ditangani atau diabaikan maka dapat menyebabkan model memiliki bias yang sangat signifikan terhadap kelas mayoritas. Hal ini dapat mengakibatkan model yang tidak peka terhadap kasus-kasus dalam kelas minoritas yang mungkin memiliki nilai prediktif yang penting. Akurasi model dapat sangat tinggi karena

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

dominasi kelas mayoritas, tetapi hasil prediksi pada kelas minoritas akan menjadi sangat rendah. SMOTE (*Synthetic Minority Over-sampling Technique* bertujuan untuk meningkatkan jumlah sampel dalam kelas minoritas dengan menciptakan sampel sintetis berdasarkan data yang ada [2].

[5] K-Fold Cross Validation

K-Fold Cross Validation merupakan teknik yang digunakan untuk menguji ataupun melatih data yang akan digunakan[15]. *Cross Validation* merupakan Teknik yang dilakukan untuk dapat menilai atau dapat memvalidasi tingkat keakuratan dari sebuah model berdasarkan dataset tertentu yang mana model tersebut digunakan untuk melakukan prediksi atau klasifikasi. Dalam metode K-Fold Cross Validation, data dibagi menjadi K bagian setelah dilakukan proses pembobotan term sebelumnya. Percobaan akan dilakukan sebanyak K kali dengan menggunakan satu bagian sebagai data uji[16]. K-fold Cross Validation dapat memberikan solusi terhadap masalah akurasi yang berbeda saat menggunakan set tes yang berbeda pada waktu evaluasi kinerja model. Dengan menggunakan K-fold Cross Validation data akan dibagi kedalam K bagian / fold dan dari setiap fold yang ada akan digunakan sebagai set pengujian[17]

3. Metode Penelitian

Pengerjaan penelitian ini menggunakan *Software* R Studio, berdasarkan implementasi yang digambarkan sebagai berikut:

3.1. Pengumpulan Data

Data yang digunakan merupakan data sekunder yakni data yang diperoleh dari website Kaggle yakni data dari lendingclub.com sejak tahun 2007–2010. LendingClub.com adalah platform pinjaman *peer-to-peer* yang menghubungkan peminjam yang membutuhkan dana dengan investor yang memiliki dana untuk diinvestasikan. Dalam data loan terdapat 9578 dataset yang meliputi kelas data dengan peminjam dapat menyelesaikan kredit sebanyak 8045 record dan peminjam yang tidak dapat menyelesaikan kredit sebanyak 1533 record yang dimana menunjukkan adanya ketidakseimbangan kelas.

3.2. Pengolahan Data

1. Praprosessing Data

Tahap ini mencakup pembersihan dan normalisasi data. Pembersihan data dilakukan unruk membersihkan data dari nilai-nilai yang hilang atau tidak valid. Sedangkan normalisasi data digunakan untuk memastikan bahwa semua fitur berada dalam skala yang sama dan skala yang memang sesuai untuk analisis klasifikasi

2. Pembagian Dataset

Pada tahap ini data dibagi menjadi data latih dan data uji. Proporsi pembagian ini adalah 70% untuk data latih dan 30% untuk data uji. Pembagian data dilakukan secara acak untuk memastikan representasi yang adil dari setiap kelas dalam data latih dan data uji.

3.3. Teknik Analisis Data

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

Teknik analisis data yang digunakan dalam penelitian merupakan algoritma naïve bayes dalam melakukan proses klasifikasi dan dengan menggunakan metode SMOTE untuk membantu mengoptimalkan akurasi algoritma naïve bayes dikarenakan data yang digunakana merupakan data yang tidak seimbang.

1. Penerapan SMOTE

Menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) pada data latih untuk menyeimbangkan jumlah sampel antara kelas mayoritas dan minoritas. SMOTE menghasilkan sampel sintetis dari kelas minoritas dengan melakukan interpolasi antara sampel-sampel minoritas yang ada. Ini dilakukan untuk mengatasi masalah ketidakseimbangan kelas yang dapat mempengaruhi performa model.

2. Klasifikasi Model

Pada tahap ini Algoritma yang digunakan dalam pengklasifikasian yaitu Naive Bayes. Model Naive Bayes dilatih menggunakan data latih yang belum dan telah dioversample dengan SMOTE.

3. K-Fold Cross Validation

Melakukan K-Fold Cross Validation untuk mengevaluasi model lebih lanjut. Untuk memastikan model yang dilatih tidak overfitting dan memberikan gambaran yang lebih akurat tentang performa model pada data yang tidak terlihat.

3.4. Evaluasi

Evaluasi dilakukan dengan tujuan mengukur kinerja model klasifikasi yang telah dilakuakan. Menggunakan data uji untuk memprediksi dan mengevaluasi performa model. Hitung dan bandingkan akurasi model sebelum dan sesudah penerapan SMOTE

4. Hasil dan Pembahasan

4.1. Hasil Penerapan Teknik SMOTE

Teknik SMOTE digunakan untuk menyeimbangkan data yang tidak seimbang sehingga hasil akurasi dari klasifikasi yang diperoleh lebih optimal.

Tabel 1. Perbandingan kelas sebelum SMOTE

Kelas 0	Kelas 1
0.8399457	0.1600543

Jumlah data yang digunakan yakni sebanyak 9578, dengan data training yang digunakan sebanyak 6706 dan data testing sebanyak 2872. Komposisi perbandingan data training dan testing sebesar 0.7 dan 0.3. Kelas mayoritas dan kelas minoritas akan menggunakan kode untuk memudahkan dalam analisis. Kelas mayoritas atau kelas untuk peminjam yang dapat menyelesaikan kredit berkode 0, sedangkan untuk kelas minoritas atau kelas untuk peminjam yang tidak dapat menyelesaikan kredit berkode 1.

Berdasarkan ketidakseimbangan data yang ada, maka akan dilakukan pengolahan data untuk

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

mengatasi masalah ketidakseimbangan kelas dalam dataset. Metode ini bekerja dengan cara menghasilkan data sintetik dari kelas minoritas sehingga menghasilkan data yang seimbang antara kelas mayoritas dengan kelas minoritas.

Tabel 2. Perbandingan kelas setelah SMOTE

Kelas 0	Kelas 1
0.512377	0.487623

Teknik SMOTE yang digunakan yakni *package* ROSE (*Random Over Sampling Examples)* yang dimana teknik tersebut digunakan untuk meresampling data dengan menghasilkan sampel sintetis berdasarkan distribusi kelas dan dapat membantu meningkatkan performa model. Setelah melakukan proses penyeimbangan data dengan package ROSE, terlihat bahwa proporsi antara variabel 0 dan 1 hanya berbeda 0.2%, hal tersebut menunjukkan bahwa data yang akan digunakan telah seimbang.

4.2. Klasifikasi Model

1. Naive Bayes Tanpa Teknik SMOTE

Tabel 3. Kemampuan model sebelum SMOTE

Kelas 0	Kelas 1
0.8398449	0.1601551

Berdasarkan tabel diatas, diperoleh informasi bahwa kemampuan model untuk masing-masing kelas yakni kelas 0 sebesar 0.839 (sekitar 83.9%), kemudian untuk kelas 1 sebesar 0.161 (sekitar 16.1%).

2. Naive Bayes dengan Teknik SMOTE

Tabel 4. Keamampuan model setelah SMOTE

Kelas 0	Kelas 1
0.512377	0.487623

Berdasarkan tabel diatas, diperoleh informasi bahwa kemampuan model untuk masing-masing kelas setelah penerapan teknik SMOTE, yakni untuk kelas 0 sebesar 0.512 (sekitar 51.2%), kemudian untuk kelas 1 sebesar 0.488 (sekitar 48.8%). Hal tersebut menunjukkan bahwa data antara kelas 1 dan 0 telah seimbang.

4.3. Penerapan K-Fold Cross Validation

1. Sebelum penerapan SMOTE

Tabel 5. Akurasi sebelum SMOTE

FALSE	TRUE
0.8015198	0.8379037

Dari output diatas, diperoleh akurasi untuk FALSE sebesar 0.8015 (sekitar 80.15%), begitu juga untuk akurasi TRUE diperoleh sebesar 0.8379 (sekitar 83.79%), dengan kemampuan model untuk mengenali kelas minoritas sebesar 16.1%.

2. Setelah penerapan SMOTE

Tabel 6. Akurasi setelah SMOTE

FALSE	TRUE
0.6368839	0.6859439

Dari output diatas, diperoleh akurasi untuk FALSE sebesar 0.6368 (sekitar 63.68%) begitu juga untuk akurasi TRUE diperoleh sebesar

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

0.6859 (sekitar 68.59%) dengan kemampuan model untuk mengenali kelas minoritas sebesar 48.8%.

4.4. Evaluasi

1. Sebelum penerapan SMOTE

a. Confusion matrix: berdasarkan nilai dari *confusion matrix* diperoleh bahwa model hanya mampu memprediksi kelas 0 sebanyak 2395 kasus, dan untuk yang bukan kelas 0 sebanyak 449 kasus sedangkan sebanyak 18 kasus diprediksi sebagai kelas 1 yang tidak benar dan 10 kasus sebagai kelas 1 yang benar.

b. Akurasi 83.74% yang berarti bahwa model memiliki nilai akurasi yang baik namun memiliki kekurangan karena tidak mampu memprediksi kelas 0 dengan benar

c. Recall (sensitivity) 0.99% yang berarti bahwa model sangat baik dalam memprediksi sebagian besar dari kasus positif. Hanya 1% dari kasus positif yang terlewat (false negatives).

d. Presisi 0.842, berarti bahwa dari semua prediksi yang dibuat oleh model sebagai positif yakni 84.2% di antaranya benar-benar positif.

e. F1-score 0.91 (91%): berarti bahwa model memiliki keseimbangan yang baik antara precision dan recall. Ini menunjukkan bahwa model dapat diandalkan dalam memprediksi kelas positif, dengan 91% dari prediksi positif yang benar dan kemampuan yang tinggi untuk menangkap kasus positif

2. Setelah penerapan SMOTE

a. Confusion matrix: berdasarkan nilai dari *confusion matrix* diperoleh bahwa model mampu memprediksi kelas 0 dengan benar sebanyak 2138 kasus, sedangkan untuk bukan 0 sebanyak 357 kasus, sebanyak 275 kasus kelas 0 yang salah diprediksi sebagai kelas 1, dan sebanyak 102 kasus kelas 1 diprediksi sebagai kelas 1.

b. Akurasi 77.99% berarti bahwa model benar dalam memprediksi kelas (baik positif maupun negatif) sekitar 78% dari semua prediksi yang dibuat, model tersebut menunjukkan bahwa hampir 78% dari semua contoh dalam dataset (baik yang positif maupun negatif) diprediksi dengan benar.

c. *Recall* (*sensitivity*) 88.60% yang berarti bahwa proporsi kelas 0 yang terdeteksi dengan benar sebesar 88.60%

d. Presisi 85.69%, yang menunjukkan bahwa sekitar 85.69% prediksi kelas 0 adalah benar

e. F1-score 87.0% menandakan bahwa model berfungsi dengan baik, terutama di situasi di mana keseimbangan antara menangkap kasus positif dan

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

meminimalkan kesalahan prediksi sangat penting.

4.5. Pembahasan

Data yang tidak seimbang merujuk pada kondisi dimana distribusi kelas dalam dataset tidak sebanding. Kondisi tersebut dapat memunculkan beberapa masalah diantaranya model cenderung memprediksi kelas mayoritas. Salah satu metode yang dapat digunakan untuk menangani masalah ketidakseimbangan adalah dengan menggunakan salah satu metode *oversampling* yakni SMOTE. Cara kerja teknik SMOTE yakni dengan membuat sampel sintetis baru berdasarkan kombinasi linier antara data yang ada pada kelas minoritas dengan kelas mayoritas. Berdasarkan hasil analisis, data yang awalnya *imbalance* yakni 0.84 : 0.16 setelah penerapan teknik SMOTE menjadi *blance* yakni 0.51 : 0.49. Setelah data seimbang barulah kemudian melakukan tahapan klasifikasi model yang dimana model klasifikasi yang digunakan yaknik algoritma Naïve Bayes. Berdasarkan hasil analisis, diperoleh bahwa probabilitas atau peluang awal untuk masing-masing kelas setelah penerapan teknik SMOTE lebih seimbang dibandingkan dengan sebelum penerapan teknnik SMOTE.

Penerapan teknik *K Fold Cross Validation* pada data yang belum seimbang menghasilkan nilai kappa yang rendah sehingga mengindikasikan model tidak bekerja secara konsisten. Sedangkan pada data yang seimbang, diperoleh nilai kappa yang lebih tinggi. Namun, untuk mengetahui apakah model termasuk ke dalam kategori baik atau tidak, dapat menggunakan beberapa metric evaluasi lainnya seperti nilai *recall, presisi,* dan F1-score. Pada data yang telah seimbang diperoleh informasi bahwa model lebih mampu memprediksi kelas minoritas dengan lebih baik yang dilihat melalui nilai dari *confusion matrix.* Kemudian model juga mampu menyeimbangkan antara menangkap kelas positif dan negatif.

5. Kesimpulan

Algoritma naïve bayes mampu mengklasifikasikan data yang tidak seimbang. Dalam penelitian ini kelas 0 dan kelas 1 memiliki proporsi sebersar 0.839% dan 0.161%, namun setelah diterapkan metode SMOTE, proporsi kelas 0 dan 1 menjadi seimbang dengan nilai 0.489 untuk kelas 1 dan 0.510 untuk kelas 0. Akurasi yang diperoleh setelah menerapkan teknik SMOTE lebih rendah dibandingkan dengan sebelum menerapkan teknik SMOTE. Namun pada metric evaluasi lainnya seperti *confusion matrix* didapatan hasil sebelum penerapan SMOTE kemampuan model memprediksi kelas minoritas dengan benar sebanyak 10 kasus, sedangkan setelah penerapan SMOTE kemampuan model memprediksi kelas minoritas dengan benar sebanyak 102 kasus. Penelitian ini

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

dapat dijadikan landasan bagi peneliti selanjutnya mengenai data yang tidak seimbang dan teknik penyeimbangan data menggunakan SMOTE. Penelitian ini juga dapat dikembangkan menggunakan algoritma lainnya selain dari algoritma Naïve Bayes dan dapat menggunakan teknik penyeimbangan data lainnya

[6] Daftar Pustaka

[1] Pratiwi D, Awangga RM, Setyawan MYH. Seleksi Calon Kelulusan Tepat Waktu Mahasiswa Teknik Informatika Menggunakan Metode Naive Bayes. Kreatif; 2020.

[2] Riswandha WAP. Evaluasi Performa Synthetic Minority Oversampling Technique (SMOTE) Untuk Mengatasi Klasifikasi Data Tidak Seimbang Pada Metode K-Nearest Neighbor (KNN) Dan Support Vector Machine (SVM). 2023;

[3] Nursyahfitri R, Rozikin C, Adam RI. Penerapan Metode SMOTE dalam Klasifikasi Daerah Rawan Banjir di Karawang Menggunakan Algoritma Naive Bayes. J Sist dan Teknol Inf. 2022;10(4):339.

[4] Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. Sci Rep [Internet]. 2021;11(1):1–11. Available from: https://doi.org/10.1038/s41598-021-03430-5

[5] Hairani H, Anggrawan A, Priyanto D. Improvement Performance of the Random Forest Method on Unbalanced Diabetes Data Classification Using Smote-Tomek Link. Int J Informatics Vis. 2023;7(1):258–64.

[6] Kurniasih A, Isyara K. Penggunaan Metode SMOTE pada Naïve Bayes Gaussian untuk Klasifikasi Mahasiswa Drop Out. Semin Nas Mhs Ilmu Komput dan Apl. 2023;616–23.

[7] Rahman F, Negeri Alauddin Makassar I, Alauddin Makassar N. Optimalisasi Prediksi Kelulusan Mahasiswa Tepat Waktu Menggunakan Binning Dan Synthetic Minority Oversampling Technique (Smote). J Artif Intell Data Sci. 2024;4(1):29–35.

[8] Pambudi L. … Untuk Menganalisis Kepuasan Peserta Program Indonesia Bisa Baca Quran Menggunakan Algoritma Decision Tree (C4. 5) Berbasis …. J Teknorama (Informatika dan …. 2023;1(1):14–20.

[9] Heliyanti Susana. Penerapan Model Klasifikasi Metode Naive Bayes Terhadap Penggunaan Akses Internet. J Ris Sist Inf dan Teknol Inf. 2022;4(1):1–8.

[10] Sobri A, Satrianansyah S, Noverendi BA. Implementasi Sistem Pakar Diagnosis Penyakit Pada Ibu Hamil Menggunakan Metode Naïve Bayes. J Inf Syst Res. 2023;4(4):1245–52.

[11] Arsa D, Weni I, Fahreza A. Analisis Sentimen Opini Publik Terhadap Pariwisara di Masa Pandemi COVID-19 Menggunakan Metode Naïve Bayes. J Telemat. 2022;17(1):49–54.

[12] Kurniadi D, Nuraeni F, Firmansyah M, Komputer JI, Korespondensi P. Klasifikasi Masyarakat Penerima Bantuan Langsung Tunai Dana Desa Menggunakan Naive Bayes dan SMOTE. J Teknol Inf dan Ilmu Komput. 2022;10(2):309–20.

[13] Pulungan MP, Purnomo A, Kurniasih A. Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Kepribadian MBTI Menggunakan Naive Bayes Classifier. J Teknol Inf dan Ilmu Komput. 2023;10(7):1493–502.

[14] Andreyestha A, Azizah QN. Analisa Sentimen Kicauan Twitter Tokopedia Dengan Optimalisasi Data Tidak Seimbang Menggunakan Algoritma SMOTE. Infotek J Inform dan Teknol. 2022;5(1):108–16.

Infotek : Jurnal Informatika dan Teknologi
Vol. 8 No. 1, Januari 2025
Hal. 217-227

e-ISSN 2614-8773

DOI : 10.29408/jit.v8i1.28340 Link : https://dx.doi.org/10.29408/jit.v8i1.28340

[15] Alwanda AY, Utami E, Yaqin A. Analisis Klasifikasi Konsentrasi Mahasiswa Menggunakan Algoritma K-Nearest Neighbor. Infotek J Inform dan Teknol. 2024;7(2).

[16] Fathoni FM, Putra CA, Nurlaili AL. Klasifikasi Penyakit Daun Anggur Menggunakan Metode K-Nearest Neighbor Berdasarkan Gray Level Co-Occurrence Matrix. Biner J Ilm Inform dan Komput. 2024;3(1):8–15.

[17] Duan F, Zhang S, Yan Y, Cai Z. An Oversampling Method of Unbalanced Data for Mechanical Fault Diagnosis Based on MeanRadius-SMOTE. Sensors. 2022;22(14).