KLIK: Kajian Ilmiah Informatika dan Komputer
ISSN 2723-3898 (Media Online)
Vol 4, No 3, Desember 2023, Hal 1551-1561
DOI 10.30865/klik.v4i3.1486
https://djournals.com/klik

Perbandingan Algoritma Naïve Bayes dan K-Nearest Neighbor Pada Imbalace Class Dataset Penyakit Diabetes

Muhammad Rousydi Hunafa*, Arief Hermawan

Fakultas Sains & Teknologi, Program Studi Informatika, Universitas Teknologi Yogyakarta, Yogyakarta, Indonesia
Email: [1,*]muhammad.5200411483@student.uty.ac.id, [2]ariefdb@uty.ac.id
Email Penulis Korespondensi: muhammad.5200411483@student.uty.ac.id

Abstrak−Penyakit diabetes menjadi perhatian global karena dampaknya yang signifikan terhadap kesehatan masyarakat. Pengelolaan penyakit ini menjadi fokus utama untuk mencegah komplikasi serius. Kemajuan teknologi, khususnya model machine learning, telah membuka peluang baru dalam identifikasi penyakit diabetes. Studi ini membandingkan kinerja algoritma klasifikasi Naive Bayes dan K-Nearest Neighbor (KNN) pada dataset diabetes yang tidak seimbang. Tujuan utamanya adalah mengevaluasi performa algoritma-algoritma ini dalam memprediksi penyakit diabetes dengan mempertimbangkan ketidakseimbangan kelas. Metode klasifikasi diterapkan pada dataset yang telah dikumpulkan sebelumnya. Hasil penelitian menunjukkan bahwa Naive Bayes dengan teknik SMOTE menunjukkan performa terbaik dengan akurasi 71.66%, diikuti oleh Naive Bayes tanpa SMOTE (76.03%), dan KNN dengan SMOTE (80.47%). Meskipun KNN tanpa SMOTE memiliki akurasi tertinggi (83.02%), Naive Bayes dengan SMOTE menunjukkan keseimbangan yang lebih baik antara akurasi, presisi, dan recall. Penggunaan teknik SMOTE meningkatkan performa Naive Bayes dengan peningkatan presisi dan recall, menunjukkan kemampuannya dalam mengatasi ketidakseimbangan kelas pada dataset diabetes. Studi ini memberikan wawasan tentang pemilihan algoritma terbaik dan teknik penanganan ketidakseimbangan kelas yang efektif dalam memprediksi penyakit diabetes pada dataset yang tidak seimbang.

Kata Kunci: Diabetes; K-NN; Naïve Bayes; Penyakit; Perbandingan

Abstract−Diabetes is a global health concern due to its significant impact on public health. Managing this disease is crucial to prevent serious complications. Technological advancements, particularly in machine learning models, have opened new avenues in diabetes identification. This study compares the performance of the Naive Bayes and K-Nearest Neighbor (KNN) classification algorithms on an imbalanced diabetes dataset. The primary aim is to evaluate these algorithms' performance in predicting diabetes while considering class imbalance. Classification methods were applied to previously collected datasets. The research findings demonstrate that Naive Bayes with the SMOTE technique exhibited the best performance with an accuracy of 71.66%, followed by Naive Bayes without SMOTE (76.03%), and KNN with SMOTE (80.47%). Although KNN without SMOTE showed the highest accuracy (83.02%), Naive Bayes with SMOTE showcased a better balance between accuracy, precision, and recall. The utilization of the SMOTE technique improved Naive Bayes' performance by enhancing precision and recall, indicating its capability to address class imbalance in the diabetes dataset. This study offers insights into selecting the best algorithms and effective techniques for handling class imbalance to predict diabetes on imbalanced datasets.

Keywords: Diabetes; K-NN; Naïve Bayes; Diesease; Comparison

1. PENDAHULUAN

Diabetes, menurut International Diabetes Federation (IDF), merupakan kondisi kronis yang ditandai dengan tingginya kadar gula darah dalam tubuh [1]. Penyakit ini telah menjadi masalah kesehatan global yang memengaruhi jutaan orang di seluruh dunia. Tingginya kadar gula darah pada diabetes, jika tidak terkontrol dengan baik, dapat menyebabkan berbagai komplikasi serius yang memengaruhi organ tubuh, termasuk gangguan pada mata, ginjal, saraf, serta meningkatkan risiko serangan jantung dan stroke [2]. Pengelolaan diabetes untuk mencegah komplikasi yang berpotensi mengancam jiwa memerlukan pengendalian kadar gula darah yang baik. Terapi untuk diabetes tidak hanya terfokus pada pengaturan diet dan olahraga, tetapi juga melibatkan penggunaan obat-obatan seperti insulin, metformin, sulfonilurea, dan inhibitor SGLT2 (Sodium-Glucose Cotransporter-2) untuk membantu mengontrol kondisi ini. Kesadaran akan risiko komplikasi yang terkait dengan diabetes, bersama dengan peran penting obat-obatan dalam pengelolaannya, memperkuat perlunya pemahaman yang mendalam serta pendekatan holistik dalam menangani kondisi ini.

Kemajuan teknologi yang terus berkembang telah mengubah cara aparat kesehatan memandang identifikasi suatu penyakit. Penggunaan model machine learning menjadi salah satu hasil dari perkembangan teknologi ini yang memiliki dampak luas, termasuk dalam bidang kesehatan. Penerapan model machine learning dapat digunakan untuk mengklasifikasikan penyakit diabetes pada pasien [3]. Dengan memanfaatkan data pasien yang telah terkumpul sebelumnya, proses pengklasifikasian penyakit diabetes menjadi lebih efisien dan mengurangi ketergantungan pada sumber daya manusia.

Banyak model prediksi machine learning yang umumnya digunakan dalam pendeteksian penyakit diabetes merupakan teknik klasifikasi [4]. Model klasifikasi berfungsi dengan mencari pola dalam dataset yang berhubungan dengan penyakit diabetes, yang sering tersedia secara luas dalam berbagai repositori data terbuka [5]. Model klasifikasi ini bertujuan untuk memprediksi apakah seseorang menderita diabetes atau tidak. Penelitian yang dilakukan oleh [6] ini mengevaluasi penggunaan metode KNN yang dimodifikasi untuk mengklasifikasikan pasien-pasien dengan diabetes. Mereka memanfaatkan dataset pasien diabetes terbuka yang terdiri dari 768 pasien untuk mengembangkan model mereka. Hasilnya menunjukkan bahwa model yang dibangun memiliki tingkat akurasi sebesar 89% dalam klasifikasi pasien-pasien diabetes. Selain itu, penelitian ini juga menghasilkan sebuah paket perangkat lunak dalam bahasa pemrograman Python yang didasarkan pada model tersebut untuk membantu dalam diagnosis penyakit diabetes.

Penelitian [7] ini mengevaluasi beberapa teknik data mining untuk mengklasifikasikan diabetes. Mereka menggunakan dataset yang diperoleh dari UCI machine learning depository yang terdiri dari 520 instansi, masing-masing memiliki 17 atribut. Tujuh algoritma klasifikasi yang berbeda, termasuk *Bayes Network, Naïve Bayes*, J48, *Random Tree*, *Random Forest*, K-NN, dan SVM, dipelajari pada dataset ini. Hasil yang diperoleh menunjukkan bahwa algoritma K-NN memiliki akurasi tertinggi sebesar 98.07% dan dianggap sebagai metode terbaik untuk mengidentifikasi dan mengklasifikasikan penyakit diabetes pada dataset yang diteliti.. Penelitian lain yang dilakukan oleh [8] menunjuk Penelitian ini mengevaluasi prediksi diabetes dengan menggunakan rekam medis elektronik dari pasien-pasien diabetes. Dua algoritma klasifikasi data mining, *Naive Bayes* dan *Support Vector Machine*, dibandingkan untuk analisis prediktif. Hasil analisis menunjukkan akurasi antara 75% hingga 78%, memperlihatkan potensi untuk penggunaan kedua algoritma dalam memprediksi diabetes berdasarkan rekam medis elektronik.

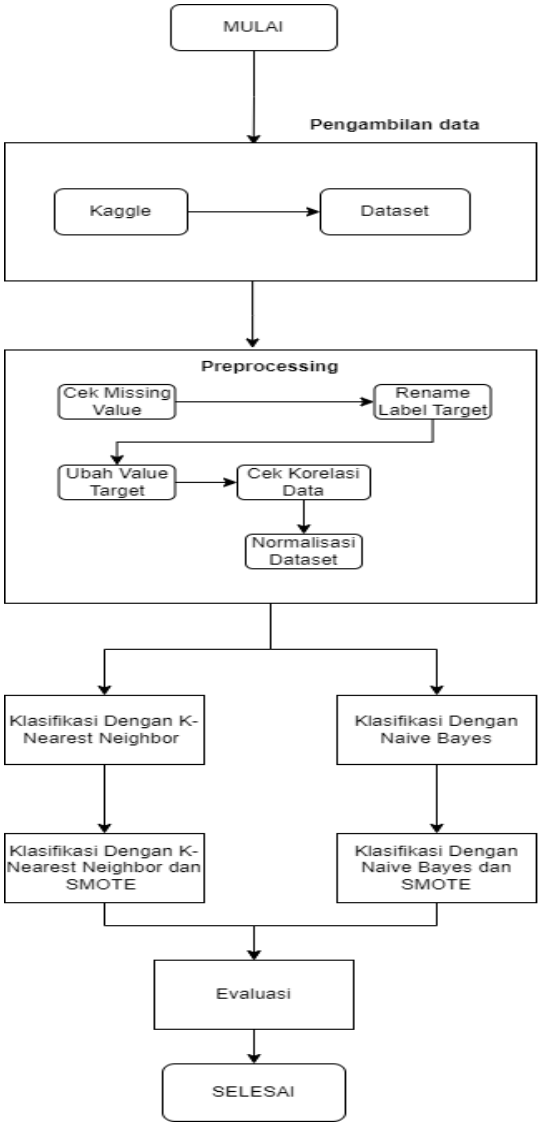
Penelitian [9] ini mengevaluasi kinerja algoritma *K-Nearest Neighbor* (KNN) dalam klasifikasi diabetes menggunakan dataset yang relevan. Penelitian ini memperhatikan pengaruh perubahan nilai K pada akurasi, presisi, recall, dan F-Measure dari metode KNN. Hasil penelitian menunjukkan bahwa nilai K tertentu mempengaruhi performa KNN dalam mengklasifikasikan dataset penderita diabetes. Sementara penelitian [10] ini mengevaluasi kinerja dua algoritma klasifikasi, *Naïve Bayes* dan *K-Nearest Neighbors* (KNN), dalam mengklasifikasikan penyakit diabetes melitus. Metode eksperimen menggunakan lima pembagian data dan mengukur akurasi, recall, dan presisi dari kedua algoritma. Hasil penelitian menunjukkan bahwa algoritma *Naïve Bayes* memiliki tingkat akurasi lebih tinggi, mencapai 80%, sementara KNN memiliki akurasi tertinggi sebesar 75%. Meskipun demikian, algoritma KNN memberikan nilai recall yang lebih tinggi (0.92), sementara presisi tertinggi dihasilkan oleh algoritma *Naïve Bayes* (0.86). Penelitian ini memberikan wawasan tentang keefektifan kedua algoritma dalam mengklasifikasikan diabetes melitus, dengan memberikan informasi yang dapat menjadi referensi dan panduan bagi penelitian lanjutan di bidang ini. Terakhir Penelitian [11] ini membandingkan kinerja beberapa pengklasifikasi pohon pembelajaran mesin seperti *Random Forest*, C4.5, *Random Tree*, REPTree, dan *Logistic Model Tree* (LMT) dalam memprediksi Diabetes Mellitus. Evaluasi dilakukan berdasarkan akurasi dan *True Positive Rate* (TPR). Hasil analisis menunjukkan bahwa dalam memprediksi diabetes mellitus, pengklasifikasi pohon pembelajaran mesin *Logistic Model Tree* (LMT) mencapai akurasi lebih tinggi sebesar 79.31%, *True Positive Ra*te (TPR) sebesar 0.739, dan waktu eksekusi 1.09 detik yang lebih baik dibandingkan dengan pengklasifikasi lain yang diteliti.

Penelitian ini bertujuan untuk membandingkan kinerja dua algoritma klasifikasi, yaitu *Naive Bayes* dan *K-Nearest Neighbor* (KNN), pada dataset penyakit diabetes yang tidak seimbang. Fokus utama penelitian adalah mengevaluasi performa algoritma-algoritma ini dalam memprediksi penyakit diabetes menggunakan teknik klasifikasi pada dataset yang memiliki ketidakseimbangan kelas. Tujuan utama dari perbandingan ini adalah untuk menentukan algoritma yang paling efektif dalam mengatasi ketidakseimbangan kelas pada dataset diabetes sambil mempertimbangkan faktor-faktor seperti akurasi, presisi, dan recall. Selain itu, penelitian ini juga ingin mengidentifikasi apakah penggunaan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) dalam menangani ketidakseimbangan kelas dapat meningkatkan kinerja algoritma-algoritma klasifikasi yang diuji, khususnya *Naive Bayes* dan KNN [12]. Dengan demikian, tujuan utama penelitian adalah untuk memberikan wawasan tentang pemilihan algoritma terbaik dan teknik penanganan ketidakseimbangan kelas yang paling efektif dalam memprediksi penyakit diabetes pada dataset yang tidak seimbang.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini mengalami berbagai proses tahapan. Pertama adalah tahapan pengambilan data yang sumbernya berasal dari Kaggle. Setelah data didapatkan, terjadi tahap prepocessing yang dimana data akan diolah kembali pada tahapan ini. Setelah data diprepocessing, data akan diklasifikasikan dengan model *Naïve Bayes* dan K-NN (*K-Nearest Neighbor*) pada tahapan ini juga akan diterapkan teknik SMOTE (Synthetic Minority Oversampling Technique). Setelah klasifikasi dilakukan maka akan dilanjutkan dengan evaluasi yang dimana tahapan ini ditandai dengan *Confusion Matrix*. Proses tahapan-tahapan ini ditunjukkan pada gambar 1 dibawah.



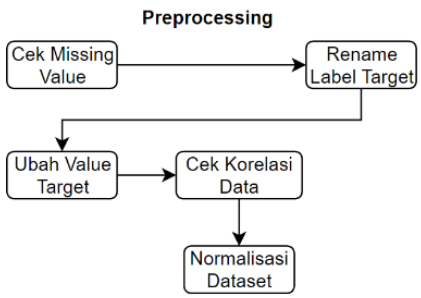
Gambar 1. Tahapan Penelitian

2.2 Pengambilan Data

Penelitian ini mendapatkan data dari Kaggle yang bersifat open dataset. Proses pengambilan data dilakukan dengan cara pengunduhan langsung dari Kaggle. Dataset inilah yang akan digunakan dalam penelititian ini yang dimana klasifikasi yang menggunakan algoritma K-NN dan *Naïve Bayes* yang dibantu teknik SMOTE.

2.3 Preprocessing Dataset

Pada tahapan ini dilakukan berbagai proses pengolahan data agar data dapat digunakan untuk proses klasifikasi [13]. Tahapan pertama yang dilakukan adalah pengecekan missing value. Selanjutnya dilakukan rename label target, lalu perubahan value target. Selanjutnya dilakukan pengecekan korelasi data untuk memilih atribut yang memiliki bobot tinggi. Setelah dilakukan pengecekan korelasi, dilanjutkan dengan normalisasi data agar menjadi format yang lebih konsisten. Proses dari tahapan preprocessing dapat dilihat pada gambar 2 dibawah.



Gambar 1. Preprocessing Dataset

2.3.1 Cek Missing Value

Langkah pertama yang dilakukan dalam tahap preprocessing adalah pengecekan mising value. Tahap ini memastikan bahwa dalam dataset tidak ditemukan data yang kosong.

2.3.2 Rename Label Target

Pada tahap ini label target akan diganti nama. Tujuan dari penggantian nama label target adalah untuk memudahkan dalam proses modeling.

2.3.3 Ubah Value Target

Pada tahap ini value dari target terkadang tidak sesuai untuk hasil dari penelitian, pada tahapan ini dilakukan perubahan isi dari target. Hal ini dilakukan untuk menghindari hasil yang tidak sesuai.

2.3.4 Cek Korelasi Data

Tahap pengecekan korelasi data dilakukan untuk menemukan kolom mana saja yang memiliki nilai korelasi tertinggi dalam dataset. Tahapan ini bertujuan untuk memilih kolom yang berbobot dan menghindari kolom yang memiliki nilai korelasi yang rendah karena dapat menyebabkan penurunan nilai akurasi [14].

2.3.5 Normalisasi Dataset

Tahapan normalisasi data dilakukan untuk menghasilkan value data yang batasannya nilainya telah tertentukan. Normalisasi dilakukan untuk memudahkan dalam melakukan analisis dalam modeling karena rentan nilai yang telah ditentukan [15].

2.4 K-Nearest Neighbor

Algoritma *K-Nearest Neighbor* (KNN) adalah model dalam supervised learning di mana data baru diklasifikasikan berdasarkan jaraknya ke sejumlah tetangga terdekat yang ditentukan oleh parameter K. KNN memanfaatkan mayoritas kategori dari K tetangga terdekat untuk mengklasifikasikan data baru. Prinsip kerjanya adalah dengan mencari tetangga terdekat dari data baru berdasarkan perhitungan jarak, sering kali menggunakan Euclidean Distance, untuk menentukan kesamaan atribut antara data baru dengan sampel latihan. Tujuan utama dari algoritma ini adalah mengelompokkan objek baru ke dalam kategori yang sesuai berdasarkan atribut yang dimiliki, dengan memanfaatkan informasi dari tetangga terdekat yang dihitung menggunakan metode perhitungan jarak seperti *Euclidean Distance* yang dipresentasikan pada persamaan 1 berikut [16].

$$E(x,y) = \sqrt{\sum_i^n 0\ (xi - yi)^2} \quad (1)$$

2.5 Naïve Bayes

Naive Bayes adalah salah satu metode klasifikasi dalam machine learning yang berbasis pada teorema probabilitas Bayes. Metode ini mengasumsikan independensi antara setiap fitur dalam dataset, meskipun demikian, mesin pembelajaran *Naive Bayes* tetap mampu memberikan prediksi yang kuat dengan menghitung probabilitas setiap fitur terhadap kelas tertentu. Dengan menggunakan prinsip probabilitas, *Naive Bayes* dapat melakukan klasifikasi dengan menghitung probabilitas setiap kelas berdasarkan informasi yang ada, dan memilih kelas dengan probabilitas tertinggi sebagai prediksi akhir. Metode ini sering digunakan dalam klasifikasi teks, analisis sentimen, serta aplikasi di mana asumsi independensi fitur bisa bekerja secara efektif [17]. Persamaan teorema Bayes dan interpretasinya ditunjukkan pada Persamaan (2) berikut [18].

$$P(X) = P(H)P(H)P(X) \quad (2)$$

2.6 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) adalah teknik yang digunakan dalam pemrosesan data untuk menangani ketidakseimbangan kelas dalam dataset. Metode ini bekerja dengan membuat sampel sintetis dari kelas minoritas dengan cara menghasilkan contoh-contoh baru yang serupa tetapi tidak identik dengan data yang ada [19]. SMOTE membantu meningkatkan representasi kelas minoritas dengan cara mengisi celah antara data yang ada, sehingga memperkuat kinerja model dalam memprediksi kelas minoritas dengan lebih baik tanpa mempengaruhi kelas mayoritas. Teknik ini sering digunakan dalam situasi di mana terdapat ketimpangan signifikan antara jumlah sampel antar kelas dalam dataset. Proses algoritma pada SMOTE melibatkan perhitungan selisih vektor fitur antara data pada kelas minoritas dan tetangga terdekatnya, lalu mengalikannya dengan nilai acak antara 0 hingga 1. Hasil kalkulasi ini kemudian ditambahkan kembali ke vektor fitur aslinya untuk menghasilkan data baru yang bersifat sintetis [20].

$$X_{bew} = X_i + (\widehat{X}_i - X_i) \times \delta \quad (3)$$

2.7 Evaluasi

Dalam Tahapan evaluasi menggunakan *Confusion Matrix* sebagai pengukur kinerja algortima *Naïve Bayes* dan *K-Nearest Neighbor*.*Confusion Matrix* adalah tabel probabilitas yang merupakan alat statistik untuk analisis kombinasi. *Confusion Matriks* didefinisikan untuk ukuran kualitas pada data spasial [21]. Dapat dilihat *Confusion Matriks* dengan empat nilai prediksi dan aktual pada tabel 1 dan rumus pada persamaan 3.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Tabel 1. Confusion Matrix

TP *(True Positive)*	FP *(False Positive)* *Type I Error*
FN *(False Negative)* *Type II Error*	TN *(True Negative)*

3. HASIL DAN PEMBAHASAN

3.1 Pengambilan Dataset

Penelitian ini menggunakan open dataset yang berasal dari Kaggel. Proses ini dilakukan dengan mendownload data secara langsung pada situs Kaggle. Situs dataset terlihat pada gambar 3.

Gambar 2. Situs Dataset

3.2 Preprocessing Dataset

3.2.1 Cek Missing Value

Pada tahapan ini menggunakan fitur dari *library DataFrame* untuk pengecekan missing valuenya.

Cek Missing Value:

	0
Diabetes_012	0
HighBP	0
HighChol	0
CholCheck	0
BMI	0
Smoker	0
Stroke	0
HeartDiseaseorAttack	0
PhysActivity	0
Fruits	0

Gambar 3. Missing Value

3.2.2 Rename Label Target

Label terget yang sebelumnya “Diabetes_012” diubah menjadi “Diabetes” untuk memudahkan dalam analisis dan modeling.

Tabel 2. Label sebelum rename

Diabetes_012	HighBP	HighChol	…
0	1	1	…
0	0	0	…
0	1	1	…

Tabel 3 setelah rename

Diabetes	HighBP	HighChol	…
0	1	1	…
0	0	0	…
0	1	1	…

3.2.3 Ubah Value Target

Valeu pada kolom target "Diabetes" memiliki nilai 0,1,dan 2, tetapi pada penelitian ini hasil yang dibutuhkan hanya terkena penyakit (1) dan tidak terkena penyakit (0). Oleh karena itu nilai 1 dan 2 akan dikelompokkan ke terkena penyakit (1), pengelompokkannya sendiri berdasarkan pada ketentuan yang ada pada laman situs dataset di Kaggle.

Tabel 4 Value sebelum diubah

Diabetes_	HighBP	HighChol	…
0	0	1	…
2	0	1	…
1	1	0	…

Tabel 5 Value setelah diubah

Diabetes_	HighBP	HighChol	…
0	0	1	…
2	0	1	…
1	1	0	…

3.2.4 Cek Korelasi Data

Pengecekan bobot korelasi dataset dilakukan dengan *library pandas* untuk memilih kolom yang berbobot dan menghindari kolom yang memiliki nilai korelasi yang rendah.

Tabel 6 Bobot korelasi

	Diabetes
Diabetes	1
GenHlth	0.3008
HighBP	0.2703
BMI	0.2239

3.2.5 Normalisasi Dataset

Proses normalisasi dilakukan dengan menggunakan *library sklearn* dengan *MinMaxScaler* sehingga nilai yang akan didapatkan berada direntan 0 dingga 1.

Tabel 7 Dataset sebelum normalisasi

Diabetes	HighBP	HighChol	BMI	…
0	1	1	40	…
0	0	0	25	…
0	1	1	28	…

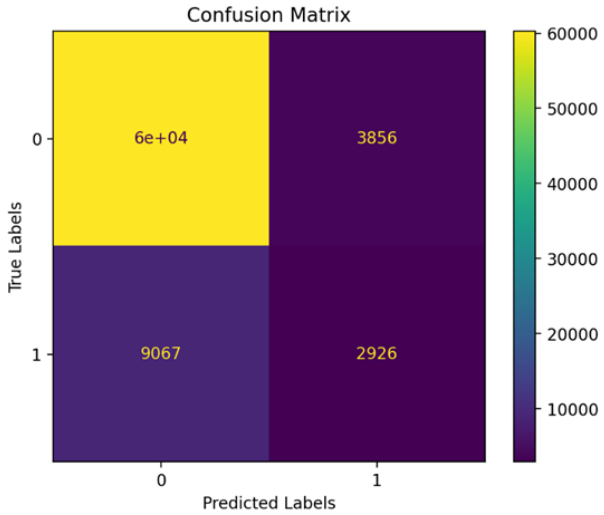
Tabel 8 Dataset setelah normalisasi

Diabetes	HighBP	HighChol	BMI	…
0	1	1	0.3256	…
0	0	0	0.1512	…
0	1	1	0.186	…

3.3 Analisis K-Nearest Neighbor

3.3.1 K-Nearest Neighbor Tanpa SMOTE

Pada analisis yang melibatkan sebuah *Confusion Matrix*, terdapat informasi yang penting untuk mengevaluasi performa suatu model. Dalam kasus ini, terdapat 60.000 data positif yang berhasil diprediksi dengan benar, sedangkan 2926 data negatif diprediksi dengan tepat. Namun, terdapat 3856 data negatif yang salah diprediksi sebagai positif, serta 9064 data positif yang keliru diprediksi sebagai negatif. Hasil evaluasi model juga menunjukkan bahwa presisi sebesar 43.14%, mengindikasikan bahwa dari semua data yang diprediksi positif, sekitar 43.14% di antaranya benar-benar positif. Sementara itu, recall sebesar 24.4% menunjukkan bahwa dari semua data yang sebenarnya positif, hanya sekitar 24.4% yang berhasil diprediksi dengan benar oleh model. F1 score, yang menggabungkan antara presisi dan recall, tercatat sebesar 31.17%, yang menunjukkan seberapa baik model dapat mengkompromikan presisi dan recall. Meskipun akurasi keseluruhan mencapai 83.02%, perlu dipertimbangkan bahwa akurasi tidak selalu mencerminkan kinerja yang baik jika data memiliki ketidakseimbangan kelas.



Gambar 4 Confusion Matrix KNN tanpa SMOTE

Implementasi metode *K-Nearest Neighbor* (KNN) tanpa SMOTE pada analisis confusion matrix ini mengindikasikan bahwa model KNN digunakan untuk klasifikasi data dengan tetap mempertahankan ketidakseimbangan kelas aslinya. KNN adalah salah satu metode dalam machine learning yang digunakan untuk klasifikasi berdasarkan keterdekatannya dengan tetangga terdekat. Dalam konteks ini, KNN digunakan untuk memprediksi kelas dari setiap data baru berdasarkan sejumlah K (jumlah tetangga terdekat) terdekat dalam dataset.

Namun, tanpa penggunaan teknik SMOTE (*Synthetic Minority Over-sampling Technique*), model KNN cenderung tidak menangani ketidakseimbangan kelas. SMOTE adalah teknik oversampling yang umumnya digunakan untuk menangani dataset yang tidak seimbang dengan menciptakan sampel sintetis dari kelas minoritas sehingga menciptakan keseimbangan antara kelas mayoritas dan minoritas. Dalam kasus ini, penggunaan KNN tanpa SMOTE menghasilkan performa yang kurang optimal dalam mengatasi ketidakseimbangan kelas, yang tercermin dari presisi yang rendah (43.14%) dan recall yang rendah (24.4%).`

Meskipun akurasi keseluruhan model KNN mencapai 83.02%, hal ini menunjukkan seberapa baik model mampu memprediksi secara tepat dari seluruh data yang ada, tetapi tidak memberikan gambaran yang lengkap dalam kasus ketidakseimbangan kelas. Oleh karena itu, dalam situasi ketidakseimbangan kelas seperti pada analisis confusion matrix tersebut, penggunaan KNN tanpa SMOTE dapat menghasilkan hasil yang kurang memuaskan dalam hal presisi dan recall, yang merupakan indikator penting dalam penanganan ketidakseimbangan kelas pada dataset.

3.3.2 K-Nearest Neighbor Dengan SMOTE

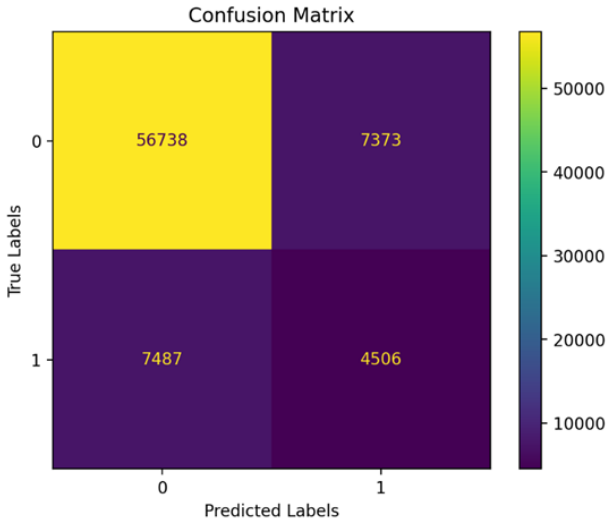
Implementasi metode K-Nearest Neighbor (KNN) dengan SMOTE (Synthetic Minority Over-sampling Technique) pada analisis confusion matrix ini menunjukkan bahwa model KNN diterapkan dengan teknik SMOTE untuk menangani ketidakseimbangan kelas pada dataset. SMOTE adalah salah satu teknik oversampling yang umum digunakan dalam machine learning untuk menangani dataset yang memiliki ketidakseimbangan kelas dengan membuat sampel sintetis dari kelas minoritas.

Dalam konteks ini, penggunaan KNN dengan SMOTE bertujuan untuk meningkatkan kinerja model dalam mengklasifikasikan kelas minoritas (positif) dengan membuat sampel sintetis yang serupa dengan data minoritas yang ada, sehingga menciptakan keseimbangan antara kelas mayoritas dan minoritas. Hasilnya, jumlah data positif yang diprediksi dengan tepat meningkat secara signifikan (dari 45.506 menjadi total 56.738) setelah penerapan SMOTE.

Dalam evaluasi hasil *Confusion Matrix* untuk model *K-Nearest Neighbors* (KNN) dengan metode SMOTE (*Synthetic Minority Over-sampling Technique*), terdapat beberapa informasi penting yang dapat dianalisis. Dari total 56.738 data positif, model berhasil memprediksi dengan tepat sebanyak 45.506 data negatif, sementara jumlah data negatif yang diprediksi secara akurat mencapai 4.506. Namun, terdapat 7.373 data negatif yang salah diprediksi sebagai positif, dan 7.487 data positif yang keliru diprediksi sebagai negatif.

Hasil evaluasi parameter menunjukkan nilai presisi sebesar 37.93%. Hal ini menggambarkan bahwa dari semua data yang diprediksi positif oleh model KNN dengan SMOTE, sekitar 37.93% di antaranya adalah benar-benar positif. Sementara itu, nilai recall sebesar 37.57% menunjukkan bahwa dari keseluruhan data yang sebenarnya positif, model berhasil mengidentifikasi sekitar 37.57% di antaranya dengan tepat. F1 score, yang menggabungkan presisi dan recall, mencapai 37.75%, menunjukkan seberapa baik model mampu mengkompromikan keseimbangan antara presisi dan recall.

Meskipun akurasi keseluruhan mencapai 80.47%, penting untuk dicatat bahwa keseimbangan kelas (*imbalance*) dapat mempengaruhi interpretasi performa model. Dalam hal ini, hasil yang relatif seimbang antara presisi dan recall menunjukkan bahwa model KNN dengan SMOTE dapat memprediksi kelas positif dan negatif dengan proporsi yang lebih serupa. Namun, evaluasi lebih lanjut mungkin diperlukan untuk memperbaiki jumlah kesalahan prediksi kelas positif dan negatif agar lebih mendekati nilai yang diharapkan.



Gambar 5 Confusion Matrix KNN dengan SMOTE

3.4 Analisis Naïve Bayes

3.4.1 Naïve Bayes Tanpa SMOTE

Dari hasil confusion matrix untuk model Naive Bayes tanpa menggunakan metode SMOTE, terlihat bahwa dari total 50.956 data positif, model berhasil memprediksi dengan tepat sebanyak 6.905 data negatif. Selain itu, jumlah data negatif yang diprediksi secara akurat mencapai 6.905, namun terdapat 13.155 data negatif yang salah diprediksi sebagai positif. Di sisi lain, sebanyak 5.088 data positif keliru diprediksi sebagai negatif.

Parameter evaluasi model menunjukkan nilai presisi sebesar 34.42%. Hal ini mengindikasikan bahwa dari semua data yang diprediksi positif oleh model Naive Bayes, sekitar 34.42% di antaranya adalah benar-benar positif. Sementara nilai recall sebesar 57.58% menunjukkan bahwa dari total data yang sebenarnya positif, model berhasil mengidentifikasi sekitar 57.58% di antaranya dengan tepat. F1 score, yang merupakan gabungan dari presisi dan recall, mencapai 43.08%, yang menunjukkan seberapa baik model dapat mengkompromikan keseimbangan antara presisi dan recall.

Meskipun akurasi keseluruhan mencapai 76.03%, penting untuk dicatat bahwa model ini memiliki recall yang relatif lebih tinggi daripada presisinya. Artinya, model cenderung lebih baik dalam mengidentifikasi data yang sebenarnya positif daripada memastikan bahwa prediksi positifnya benar-benar akurat. Evaluasi lebih lanjut mungkin diperlukan untuk memperbaiki jumlah kesalahan prediksi, terutama dalam mengklasifikasikan data negatif yang salah diprediksi sebagai positif.

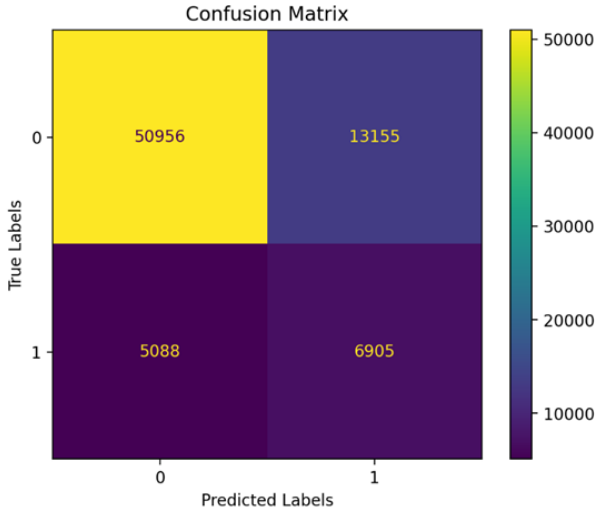
Implementasi metode Naïve Bayes tanpa menggunakan SMOTE (Synthetic Minority Over-sampling Technique) menunjukkan bahwa model Naïve Bayes digunakan untuk melakukan klasifikasi pada dataset tanpa penanganan ketidakseimbangan kelas. Naïve Bayes adalah salah satu metode klasifikasi yang sering digunakan dalam machine learning untuk memprediksi probabilitas dari suatu kelas dengan asumsi independensi antara fitur-fitur yang ada.

Dalam konteks ini, penggunaan Naïve Bayes tanpa SMOTE berarti bahwa model Naïve Bayes digunakan secara langsung pada dataset yang tidak seimbang, tanpa melakukan proses oversampling atau undersampling pada kelas minoritas atau mayoritas. Naïve Bayes melakukan prediksi berdasarkan probabilitas yang dihitung dari fitur-fitur yang ada dalam dataset, dengan asumsi bahwa fitur-fitur tersebut saling independen.

Hasil dari confusion matrix menunjukkan bahwa model Naïve Bayes tanpa SMOTE memiliki akurasi keseluruhan sebesar 76.03%, yang mengindikasikan seberapa baik model dapat memprediksi secara tepat keseluruhan dari seluruh data yang ada. Namun, presisi (34.42%) yang rendah menunjukkan bahwa dari semua data yang diprediksi positif oleh model, hanya sekitar 34.42% di antaranya adalah benar-benar positif. Sementara itu, recall (57.58%) yang lebih tinggi menunjukkan bahwa dari keseluruhan data yang sebenarnya positif, model berhasil mengidentifikasi sekitar 57.58% di antaranya dengan tepat.

Namun demikian, terdapat banyak data negatif yang salah diprediksi sebagai positif (false positives), serta beberapa data positif yang salah diprediksi sebagai negatif (false negatives). Hal ini menandakan bahwa model memiliki kecenderungan untuk lebih sensitif dalam mengidentifikasi data yang sebenarnya positif (recall yang tinggi), namun kurang spesifik dalam memastikan bahwa prediksi positifnya benar-benar akurat (presisi yang rendah).

Dalam konteks penggunaan Naïve Bayes tanpa SMOTE, model lebih fokus pada pola dan hubungan fitur-fitur dalam data, namun tidak mengatasi secara langsung masalah ketidakseimbangan kelas. Evaluasi lebih lanjut atau penanganan khusus terhadap ketidakseimbangan kelas mungkin diperlukan untuk meningkatkan performa model dalam mengklasifikasikan kelas minoritas atau mayoritas secara lebih tepat.



Gambar 6 Confusion Matrix Naïve Bayes tanpa SMOTE

3.4.2 Naïve Bayes Dengan SMOTE

Dari hasil confusion matrix yang diberikan, terlihat bahwa dalam pengujian tersebut, terdapat beberapa aspek penting yang menggambarkan performa model klasifikasi. Dari total 45.825 data positif, model berhasil memprediksi dengan tepat sebanyak 8.714 data negatif, sedangkan jumlah data negatif yang diprediksi secara akurat mencapai 8.714. Namun, terdapat 18.256 data negatif yang salah diprediksi sebagai positif, sementara 3.279 data positif keliru diprediksi sebagai negatif.

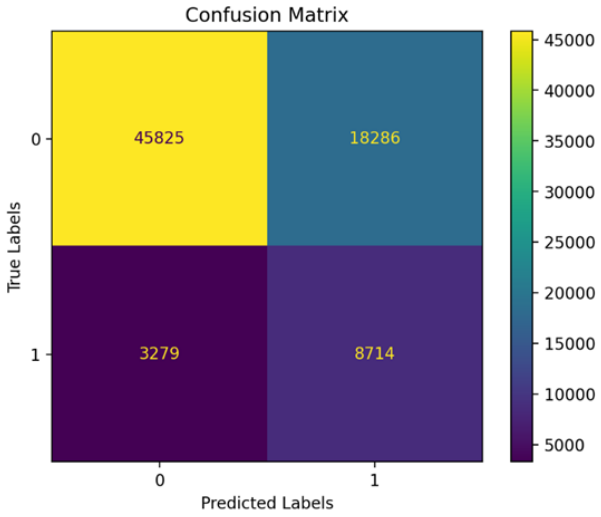
Parameter evaluasi model menunjukkan nilai presisi sebesar 32.26%. Hal ini mengindikasikan bahwa dari semua data yang diprediksi positif oleh model, sekitar 32.26% di antaranya adalah benar-benar positif. Sementara itu, nilai recall sebesar 72.72% menunjukkan bahwa dari total data yang sebenarnya positif, model berhasil mengidentifikasi sekitar 72.72% di antaranya dengan tepat. F1 score, yang merupakan harmonisasi dari presisi dan recall, mencapai 44.69%, menggambarkan seberapa baik model mampu mengkompromikan antara presisi dan recall.

Implementasi metode Naïve Bayes dengan SMOTE (Synthetic Minority Over-sampling Technique) mengacu pada penggunaan model klasifikasi Naïve Bayes yang telah diperkuat dengan teknik oversampling SMOTE untuk menangani ketidakseimbangan kelas pada dataset. Naïve Bayes adalah metode klasifikasi probabilitas yang didasarkan pada asumsi independensi antara fitur-fitur dalam dataset. SMOTE adalah teknik yang digunakan dalam machine learning untuk menangani ketidakseimbangan kelas dengan membuat sampel sintetis dari kelas minoritas agar seimbang dengan kelas mayoritas.

Dalam konteks ini, penggunaan Naïve Bayes dengan SMOTE bertujuan untuk meningkatkan performa model dalam mengklasifikasikan kelas minoritas (positif) dengan cara menciptakan sampel sintetis dari kelas minoritas. Dengan demikian, model Naïve Bayes diperkuat dengan data sintetis yang mirip dengan kelas minoritas yang ada, sehingga membantu model untuk belajar dan membuat prediksi yang lebih akurat terhadap kelas minoritas.

Hasil dari confusion matrix menunjukkan bahwa penggunaan Naïve Bayes dengan SMOTE menghasilkan akurasi keseluruhan sebesar 72.72%, yang mengindikasikan seberapa baik model dapat memprediksi dengan tepat keseluruhan dari seluruh data yang ada. Namun, presisi (32.26%) yang rendah menggambarkan bahwa dari semua data yang diprediksi positif oleh model, hanya sekitar 32.26% di antaranya adalah benar-benar positif. Sementara itu, recall (72.72%) yang lebih tinggi menunjukkan bahwa dari total data yang sebenarnya positif, model berhasil mengidentifikasi sekitar 72.72% di antaranya dengan tepat.

Penerapan SMOTE dalam Naïve Bayes membantu model dalam meningkatkan kemampuannya dalam mengenali kelas minoritas (positif) dengan memperkuat sampel dari kelas tersebut. Namun, terdapat juga sejumlah data negatif yang salah diprediksi sebagai positif dan beberapa data positif yang keliru diprediksi sebagai negatif, menandakan bahwa masih ada ruang untuk perbaikan dalam meningkatkan presisi dan spesifisitas model dalam mengklasifikasikan kelas. Evaluasi lebih lanjut dan fine-tuning mungkin diperlukan untuk mencapai keseimbangan yang lebih baik antara presisi dan recall serta meningkatkan performa model dalam mengatasi ketidakseimbangan kelas pada dataset.



Gambar 7 Confusion Matrix Naïve Bayes dengan SMOTE

3.5 Analisis Perbandingan Metode

Hasil analisis perbandingan model KNN dan Naïve Bayes menunjukkan bahwa dalam hal akurasi dan presisi, model K-NN tanpa SMOTE memberikan nilai tertinggi. Sementara itu, model *Naïve Bayes* dengan SMOTE memberikan nilai tertinggi dalam hal recall dan F1 Score. Recall mengukur kemampuan model untuk mengidentifikasi sebanyak mungkin instans positif. F1 score adalah ukuran gabungan antara presisi dan recall, yang memberikan gambaran tentang keseimbangan antara keduanya. Dalam konteks ini, model *Naïve Bayes* dengan SMOTE mampu mengenali instans positif dengan baik, meskipun akurasi dan presisinya mungkin sedikit lebih rendah dibandingkan dengan model K-NN tanpa SMOTE. Hal ini menunjukkan bahwa model naïve bayes dengan SMOTE memiliki kemampuan yang baik dalam mengurangi jumlah *false negative* (kelas positif yang salah diklasifikasikan sebagai negatif).

Tabel 9 Perbandingan Model

Model	Accuracy	Precision	Recall	F1
K-NN	**83.02%**	**43.14%**	24.4%	31.17%
K-NN SMOTE	80.47%	37.93%	37.57%	37.75%
Naïve Bayes	76.03%	34.42%	57.58%	43.08%
Naïve Bayes SMOTE	71.66%	32.27%	**72.66%**	**44.7%**

4. KESIMPULAN

Berdasarkan penelitian ini, dilakukan perbandingan antara algoritma Naive Bayes dan K-Nearest Neighbor (KNN) pada dataset penyakit diabetes yang tidak seimbang. Hasil menunjukkan bahwa Naive Bayes dengan teknik SMOTE memberikan performa terbaik dalam hal akurasi dengan nilai sebesar 71.66%, diikuti oleh Naive Bayes tanpa SMOTE dengan nilai akurasi sebesar 76.03%. KNN dengan SMOTE juga memberikan hasil yang baik dengan akurasi sebesar 80.47%, sementara KNN tanpa SMOTE memiliki akurasi yang paling tinggi, yaitu 83.02%. Hal ini menunjukkan bahwa KNN, terutama tanpa menggunakan SMOTE, memiliki kemampuan yang lebih baik dalam memprediksi kelas pada dataset diabetes yang tidak seimbang. Selain itu, penggunaan teknik SMOTE juga menjadi faktor penting dalam penelitian ini. Hasil menunjukkan bahwa penggunaan SMOTE dapat memberikan peningkatan dalam kinerja algoritma, terutama pada Naive Bayes. Meskipun Naive Bayes dengan SMOTE memiliki akurasi yang sedikit lebih rendah dibandingkan Naive Bayes tanpa SMOTE, penggunaan SMOTE dapat meningkatkan presisi (32.27% menjadi 44.7%) dan recall (57.58% menjadi 72.66%). Hal ini menunjukkan bahwa SMOTE dapat membantu dalam mengatasi masalah ketidakseimbangan kelas pada dataset diabetes. Dalam penelitian ini, berdasarkan faktor akurasi dan penggunaan SMOTE, kesimpulan yang dapat diambil adalah KNN tanpa SMOTE memberikan akurasi tertinggi, tetapi Naive Bayes dengan SMOTE memberikan keseimbangan yang lebih baik antara akurasi, presisi, dan recall. Oleh karena itu, penggunaan Naive Bayes dengan teknik SMOTE mungkin lebih disarankan untuk memprediksi kelas pada dataset diabetes yang tidak seimbang.

REFERENCES

[1] International Diabetes Federation, “About Diabetes.” Diakses: 26 November 2023. [Daring]. Tersedia pada: https://idf.org/about-diabetes/what-is-diabetes/

[2] Kementerian Kesehatan RI, Tetap Produktif, Cegah, dan Atasi Diabetes Melitus. Jakarta: Pusat Data dan Informasi Kementerian Kesehatan RI, 2020.

[3] M. D. Purbolaksono, M. Irvan Tantowi, A. Imam Hidayat, dan A. Adiwijaya, "Perbandingan Support Vector Machine dan Modified Balanced Random Forest dalam Deteksi Pasien Penyakit Diabetes," Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi), vol. 5, no. 2, hlm. 393–399, Apr 2021, doi: 10.29207/resti.v5i2.3008.

[4] A. Ridwan, "Penerapan Algoritma Naïve Bayes Untuk Klasifikasi Penyakit Diabetes Mellitus," Jurnal Sistem Komputer & Kecerdasan Buatan, vol. 4, no. 1, hlm. 15–21, 2020, doi: https://doi.org/10.47970/siskom-kb.v4i1.169.

[5] M. M. F. Islam, R. Ferdousi, S. Rahman, dan H. Bushra, "Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques," dalam Advances in Intelligent Systems and Computing, vol. 992, 2020, hlm. 113–125. doi: 10.1007/978-981-13-8798-2_12.

[6] V. Lopatka, I. Meniailov, dan K. Bazilevych, "Classification and Prediction of Diabetes Disease Using Modified k-neighbors Method," dalam 2021 IEEE 12th International Conference on Electronics and Information Technologies (ELIT), 2021, hlm. 46–50. doi: 10.1109/ELIT53502.2021.9501151.

[7] K. Alpan dan G. S. İlgi, "Classification of Diabetes Dataset with Data Mining Techniques by Using WEKA Approach," dalam 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), 2020, hlm. 1–7. doi: 10.1109/ISMSIT50672.2020.9254720.

[8] R. S. Raj, D. S. Sanjay, M. Kusuma, dan S. Sampath, "Comparison of Support Vector Machine and Naïve Bayes Classifiers for Predicting Diabetes," dalam 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019, hlm. 41–45. doi: 10.1109/ICATIECE45860.2019.9063792.

[9] A. M. Argina, "Penerapan Metode Klasifikasi K-Nearest Neigbor pada Dataset Penderita Penyakit Diabetes," Indonesian Journal of Data and Science, vol. 1, no. 2, hlm. 29–33, Jul 2020, doi: 10.33096/ijodas.v1i2.11.

[10] N. Marito Putry dan B. Nurina Sari, "KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELITUS," Jurnal Sains dan Manajemen, vol. 10, no. 1, 2022, doi: https://doi.org/10.31294/evolusi.v10i1.12514.

[11] D. Vigneswari, N. K. Kumar, V. G. Raj, A. Gugan, dan S. R. Vikash, "Machine Learning Tree Classifiers in Predicting Diabetes Mellitus," dalam 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), 2019, hlm. 84–87. doi: 10.1109/ICACCS.2019.8728388.

[12] D. Elreedy dan A. F. Atiya, "A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," Inf Sci (N Y), vol. 505, hlm. 32–64, Des 2019, doi: 10.1016/J.INS.2019.07.070.

[13] S.-A. N. Alexandropoulos, S. B. Kotsiantis, dan M. N. Vrahatis, "Data preprocessing in predictive data mining," Knowl Eng Rev, vol. 34, hlm. e1, 2019, doi: DOI: 10.1017/S026988891800036X.

[14] J. Peng dkk., "DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python," dalam Proceedings of the 2021 International Conference on Management of Data, dalam SIGMOD '21. New York, NY, USA: Association for Computing Machinery, 2021, hlm. 2271–2280. doi: 10.1145/3448016.3457330.

[15] D. Singh dan B. Singh, "Investigating the impact of data normalization on classification performance," Appl Soft Comput, vol. 97, hlm. 105524, Des 2020, doi: 10.1016/J.ASOC.2019.105524.

[16] A. Nikmatul Kasanah, U. Pujianto, T. Elektro, F. Teknik, dan U. Negeri Malang, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," JURNAL RESTI (Rekayasa Sist.Teknol. Inf.), vol. 1, no. 3, hlm. 196–201, 2019, doi: https://doi.org/10.29207/resti.v3i2.945.

[17] N. G. Ramadhan dan A. Khoirunnisa, "Klasifikasi Data Malaria Menggunakan Metode Support Vector Machine," JURNAL MEDIA INFORMATIKA BUDIDARMA, vol. 5, no. 4, hlm. 1580, Okt 2021, doi: 10.30865/mib.v5i4.3347.

[18] M. A. Maricar dan Dian Pramana, "Perbandingan Akurasi Naïve Bayes dan K-Nearest Neighbor pada Klasifikasi untuk Meramalkan Status Pekerjaan Alumni ITB STIKOM Bali," Jurnal Sistem dan Informatika (JSI), vol. 14, no. 1, hlm. 16–22, Nov 2019, doi: 10.30864/jsi.v14i1.233.

[19] J. Li, Q. Zhu, Q. Wu, dan Z. Fan, "A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors," Inf Sci (N Y), vol. 565, hlm. 438–455, Jul 2021, doi: 10.1016/J.INS.2021.03.041.

[20] E. Sutoyo dan M. Asri Fadlurrahman, "Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network," JEPIN (Jurnal Edukasi dan Penelitian Informatika) , vol. 6, no. 3, hlm. 379–385, 2020, doi: https://dx.doi.org/10.26418/jp.v6i3.42896.

[21] J. Xu, Y. Zhang, dan D. Miao, "Three-way confusion matrix for classification: A measure driven view," Inf Sci (N Y), vol. 507, hlm. 772–794, Jan 2020, doi: 10.1016/J.INS.2019.06.064.