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

# Implementasi Algoritma *Synthetic Minority Over-Sampling Technique* untuk Menangani Ketidakseimbangan Kelas pada Dataset Klasifikasi

**[1] Mulia Sulistiyono*, [2] Yoga Pristyanto, [3] Sumarni Adi, [4] Gagah Gumelar**

[1,4] Informatika, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta
[2,3] Sistem Informasi, Fakultas Ilmu Komputer, Universitas Amikom Yogyakarta

Jl. Padjajaran, Ring Road Utara, Kel. Condongcatur, Kec. Depok, Kab. Sleman, Prop. Daerah Istimewa Yogyakarta Indonesia
*e-mail: muliasulistiyono@amikom.ac.id

(***received*:** 23 Februari 2021, ***revised*:** 19 Maret 2021, ***accepted*:** 6 Mei 2021)

## Abstrak

Pada penelitian ini dilakukan penangganan ketidakseimbangan kelas terhadap kelas minoritas menggunakan teknik resampling yaitu oversampling. Algoritma oversampling yang digunakan adalah Synthetic Minority Over-sampling Technique (SMOTE). Hasil dari penelitian ini dibandingkan dengan hasil klasifikasi tanpa resampling. Uji evaluasi yang digunakan ialah akurasi, Geometric Mean (g-mean), dan Confussion Matrix (CM). Penanganan distribusi kelas yang tidak seimbang pada dataset menggunakan algoritma SMOTE dapat meningkatkan nilai akurasi maupun g-mean pada algoritma Naïve Bayes, SVM, KNN dan Decision Tree. Hal tersebut menunjukkan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap pra-pemrosesan data memberikan pengaruh terhadap nilai akurasi maupun g-mean algoritma Naïve Bayes, SVM, KNN dan Decision Tree. Pada scenario percobaan yang telah dilakukan algoritma Naïve Bayes memiliki akurasi paling baik 96,43 %, SVM dengan 99,02 %, KNN dengan 97,29 % dan Decision Tree dengan nilai 97,29 % pada dataset ecoli 15,8 setelah dilakukan SMOTE dengan 10 fold cross validation. Sedangkan memiliki nilai G-mean paling baik 96,42 % untuk algoritma Naïve Bayes, SVM dengan 99,37 %, KNN dengan 99,53 % dan Decision Tree dengan nilai 96,29 % pada dataset ecoli 15,8 setelah dilakukan SMOTE dengan 10 fold cross validation.

**Kata Kunci** : Data Mining, Klasifikasi, Imbalance Ratio (IR), Oversampling, Synthetic Minority Over-sampling Technique (SMOTE)

## *Abstract*

*In this research, the subscriber of class imbalance to the minority class was carried out using a resampling technique, namely oversampling. The oversampling algorithm used is Synthetic Minority Over-sampling Technique (SMOTE). The results of this study were compared with the results of the classification without resampling. The evaluation tests used are accuracy, Geometric Mean (g-mean), and Confusion Matrix (CM). Handling the unbalanced class distribution on the dataset using the SMOTE algorithm can increase the accuracy and g-mean values of the Naïve Bayes, SVM, KNN and Decision Tree algorithms. This shows that the handling process of the unbalanced class distribution at the pre-processing stage has an effect on the accuracy and g-mean values of the Naïve Bayes, SVM, KNN and Decision Tree algorithms. In the experimental scenario that has been carried out the Naïve Bayes algorithm has the best accuracy of 96.43%, SVM with 99.02%, KNN with 97.29% and Decision Tree with a value of 97.29% on the ecoli dataset of 15.8 after SMOTE with 10 fold cross validation. Meanwhile, it has the best G-mean value of 96.42% for the Naïve Bayes algorithm, SVM with 99.37%, KNN with 99.53% and Decision Tree with a value of 96.29% in the ecoli dataset of 15.8 after SMOTE with 10 fold cross validation.*

***Keywords:*** *Data Mining, Classification, Imbalance Ratio (IR), Oversampling, Synthetic Minority Over-sampling Technique (SMOTE)*

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

# 1 Pendahuluan

Data Mining atau sering dikenal dengan istilah Knowledge Discovery in Database (KDD) merupakan serangkaian kegiatan yang menggabungkan berbagai cabang ilmu pengetahuan menjadi satu terdiri atas sistem basis data, statistika, *machine learning*, *visualization*, dan informasi pengetahuan untuk menganalisis sebuah set data yang besar guna mendapatkan pola atau karakteristik data yang bermanfaat. Data Mining telah berhasil menyumbang manfaat dalam berbagai bidang ilmu pengetahuan seperti: bisnis, bioinformatika, genetika, kedokteran, pendidikan [1]. Pembelajaran Mesin (*Machine Learning*) merupakan bidang studi ilmiah yang mampu memberikan kemampuan melakukan tugas tertentu tanpa diberikan instruksi secara eksplisit, dengan mengandalkan pola dan inferensi. Kemampuan belajar menjadi dominan ditentukan oleh algoritma yang dapat dicapai baik menggunakan kaidah, pendekatan statistik dan pendekatan fisiologis. Algoritma pembelajaran membangun model matematika berdasarkan data sampel atau lebih dikenal dengan istilah data *training*, untuk membuat prediksi atau keputusan [2]. Aplikasi pembelajaran mesin sangat beraneka ragam, seperti penyaringan spam email, *computer vision*, dan sebagainya, Dimana program yang semacam ini akan sulit dikembangan dengan pemrograman eksplisit. *Supervised Learning* merupakan bagian dari *Machine Learning* untuk jenis pembelajaran mesin memetakan data yang disertai dengan anotasi manusia biasa disebut dengan istilah label. Label merupakan target yang diinginkan dari data latih [3]. Kemudian algoritma mempelajari dari data latih biasanya berupa *array* atau vektor, kadang kadang disebut dengan vektor fitur dan menghasilkan sebuah model statistik yang mampu memetakan masukan yang baru menjadi keluaran yang tepat [4]. Bisa diasumsikan bahwa *supervised learning* belajar dari sebuah contoh. Salah satu metode supervised learning yang populer ialah klasifikasi.

Klasifikasi merupakan teknik pengalian informasi dari data. Klasifikasi adalah metode untuk menyusun data secara sistematis atau menurut aturan atau kaidah atau kaidah yang telah ditetapkan. Sebuah kaidah atau aturan diperoleh dari pembelajaran sebuah himpunan data. Namun seringkali para peneliti tidak memperhatikan keseimbangan distribusi kelas yang memicu terjadi misclassification sebagai contoh dalam dunia medis, sedikit pasien dideteksi menderita kanker ganas, daripada pasien menderita kanker jinak karena dominasi data kanker jinak, hal ini berpotensi kesulitan mendapatkan hasil klasifikasi yang kuat [5][6]. Selain keberadaan ketidakseimbangan kelas seringkali dijumpai dataset yang memiliki dimensi yang tinggi, hal ini ditandai dengan banyaknya jumlah fitur, tentunya akan memiliki pengaruh terhadap proses penerapan teknik data mining itu sendiri baik klasifikasi, klasterisasi maupun prediksi, permasalahan yang sering timbul adalah kinerja algoritma klasifikasi baik dari sisi waktu komputasi maupun dari sisi akurasinya sebagai akibat beberapa attribute feature yang tidak memiliki relevansi dengan attribute class [7]. Salah satu metode klasifikasi adalah supervised classification dimana batasan kelas ditentukan dari awal [8]. Akurasi didalam supervised classification dapat dikontrol dengan memperhatikan kualitas data yang digunakan untuk training terhadap algoritma. Permasalahan terhadap data mining klasifikasi adalah keadaan dataset yang tidak seimbang untuk data training terhadap algoritma, agar menghasilkan akurasi yang optimal. Ketidakseimbangan dataset adalah keadaan dimana distribusi kelas didalam dataset tidak seimbang.

Sebuah kelas dikatakan tidak seimbang apabila ada suatu kelas yang memiliki data yang lebih banyak dibandingkan dengan kelas lainnya [9]. Kelompok kelas dengan jumlah data yang banyak disebut dengan kelas mayoritas, sedangkan kelompok kelas dengan jumlah yang sedikit disebut dengan kelas minoritas. Perbandingan antara kelas minoritas dengan kelas mayoritas disebut dengan Imbalance Ratio (IR) atau rasio ketidakseimbangan. Semakin besar perbedaan antara kelas minoritas dengan kelas mayoritas maka nilai dari Imbalance Ratio (IR) atau rasio ketidakseimbangan semakin besar. Ketidakseimbangan dataset pada data mining adalah masalah yang serius. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas [10]. Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan. Apabila algoritma klasifikasi di implementasikan langsung terhadap dataset yang imbalance maka akan mengalami penurunan performa [9].

Pada penelitian ini, peneliti akan melakukan penangganan ketidakseimbangan kelas terhadap kelas minoritas menggunakan teknik resampling yaitu oversampling. Teknik oversampling dipilih karena tidak menggurangi dataset akan tetapi menambah dataset yang kurang pada kelas minoritas.

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

Algoritma oversampling yang digunakan adalah Synthetic Minority Over-sampling Technique (SMOTE), algoritma ini dipilih dari beberapa algoritma resampling karena SMOTE menghasilkan akurasi yang baik dan efektif dalam menangani kelas yang tidak seimbang karena mengurangi overfitting [9].

Hal ini bertujuan untuk menyeimbangkan kelas pada dataset sehingga dapat meningkatkan kinerja dari algoritma klasifikasi. Hasil dari penelitian ini akan dibandingkan dengan hasil klasifikasi tanpa resampling. Uji evaluasi yang digunakan ialah akurasi, Geometric Mean (g-mean), dan Confussion Matrix (CM) [10]. Data pengujian yang digunakan adalah data public yang peneliti dapatkan dari situs KEELS data mining yang menyediakan dataset dengan angka Imbalance Rasio (IR) yang berbeda – beda. Dataset dari KEELS ini digunakan untuk menguji dan membandingkan usulan peneliti dalam menangani masalah ketidakseimbangan kelas pada sebuah dataset klasifikasi.

## 2 Tinjauan Literatur

Dalam melakukan penelitian ini peneliti merujuk kepada beberapa penelitian ilmiah yang telah dilakukan sebelumnya sebagai bahan acuan pada penerapan algoritma Synthetic Minority Over-sampling Technique (SMOTE) untuk mengangani ketidakseimbangan kelas pada dataset klasifikasi. Berikut ini beberapa penelitian yang sudah dilakukan sebelumnya mengenai Synthetic Minority Over-sampling Technique (SMOTE) untuk mengatasi ketidakseimbangan data.

Rimbun Siringoringo (2018) dengan judul penelitiannya "Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote Dan K-Nearest Neighbor" dalam penelitian ini dilakukan perbandingan antara klasifikasi dengan algoritma k-NN pada dataset Credit Card Fraud yang tidak seimbang secara langsung tanpa implementasi Synthetic Minority Over-sampling Technique (SMOTE), dan implementasi Synthetic Minority Over-sampling Technique (SMOTE) sebelum dilakukan klasifikasi terhadap dataset Credit Card Fraud yang tidak seimbang menggunakan algoritma k-NN. Dataset yang digunakan memiliki Imbalance Ratio (IR) sebesar 3,521 dan memiliki 25 atribut data. Pengujian hasil penelitian dilakukan dengan skema 10-fold cross validation, g-mean, f-measure, dan confusion matrix. Hasil penerapan algoritma Synthetic Minority Over-sampling Technique (SMOTE) dapat meningkatkan rata-rata G-Mean dari 53,4% ke 81,0% dan rata-rata F-Measure dari 38,7 ke 81,8% [10].

Hairani, Noor Akhmad Setiawan, Teguh Bharata Adji (2016) dengan judul penelitiannya "Metode Klasifikasi Data Mining Dan Teknik Sampling Smote Menangani Class Imbalance Untuk Segmentasi Customer Pada Industri Perbankan" dari penelitian yang sudah dilakukan penanganan imbalanced class atau ketidak seimbangan kelas menggunakan Synthetic Minority Over-sampling Technique (SMOTE) kemudian dilakukan klasifikasi menggunakan beberapa algoritma klasifikasi yaitu J48+, SVM, dan Naïve Bayes. Hasil penanganan dataset yang tidak seimbang dibandingkan dengan dataset yang tidak dilakukan penganganan. Dataset yang digunakan memiliki 17 atribut data. Pengujian hasil penelitian dilakukan dengan 30 Cross Validation dan Confusion Matrix (CM). hasil penerapan algoritma Synthetic Minority Over-sampling Technique (SMOTE) metode J48+ Synthetic Minority Over-sampling Technique (SMOTE) memiliki tingkat akurasi dan sensitivity paling tinggi yaitu sebesar 0,93% dan 0,93%. Sedangkan metode SVM memiliki nilai specificity yang paling tinggi sebesar 0.99% dan metode Naive Bayes memiliki waktu komputasi yang paling cepat dibandingkan ketiga metode lainnya sebesar 0.38 seconds. Kombinasi, metode J48 + Synthetic Minority Over-sampling Technique (SMOTE) mampu menangani class imbalance dataset Bank Direct Marketing pada industri perbankan dibandingkan dengan metode SVM dan Naive Bayes [11].

Mustaqim Mustaqim, Budi Warsito, Bayu Surarso (2019) dengan judul penelitiannya "Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan" dari penelitian yang sudah dilakukan terhadap penanganan imbalance class atau ketidakseimbangan kelas pada dataset pemakaian alat kontrasepsi implan dengan 300 data, terdiri dari 285 data mayoritas (tidak hamil) dan 15 data minoritas (hamil). Dari 300 data dibagi menjadi dua bagian, 270 digunakan untuk data training dan 30 data uji akurasi. Dari 270 data yang digunakan untuk data training, terdapat 13 data training pada kelas minoritas dan 257 data training pada kelas mayoritas. Data training pada kelas minoritas diduplikasi sebanyak data pada kelas mayoritas. Dataset terdiri dari 257 data pada kelas mayoritas, 13 data pada kelas minoritas asli, dan 244 data pada kelas

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

minoritas buatan sehingga jumlah data training menjadi 514. Algoritma klasifikasi yang digunakan NN, DT, SVM, dan LR. Pengujian hasil penelitian dilakukan dengan confusion matrix dan 10-fold cross validation. Penerapan metode kombinasi SMOTE dan NN Backpropagation untuk prediksi pemakaian alat kontrasepsi implan menghasilkan akurasi prediksi 96,1%. Implementasi kombinasi SMOTE dan NN Backpropagation mampu mengatasi imbalance class dengan akurasi 96,1% [9].

Ah-Pine dalam penelitian yang berjudul "*A study of synthetic oversampling for twitter imbalanced sentiment analysis*", melakukan penelitian terhadap ketidakseimbangan kelas menggunakan 3 *dataset binary class* menggunakan teknik oversampling SMOTE, Borderline-SMOTE dan Adaptive Synthetic (ADASYN), dan klasifikasi menggunakan CART dan Logistic Regression. Penelitian ini menunjukkan bahwa ketiga metode mampu mengatasi masalah *imbalance dataset*, dengan meningkatnya kemampuan mendeteksi kelas minoritas dan peningkatan nilai g – mean [12].

N. A. Verdikha dkk, dalam penelitian "Komparasi Metode Undersampling Untuk Klasifikasi Teks Ujaran Kebencian" dilakukan penanganan terhadap ketidak seimbangan kelas menggunakan metode oversampling SMOTE, Borderline - SMOTE versi 1 dan 2, SMOTE-SVM, ADASYN dataset yang digunakan bersumber dari data tweet (multi class) Hasil dari penelitian menunjukkan bahwa kelima metode oversampling mampu meningkatkan kinerja klasifikasi ditandai dengan meningkatnya nilai g – mean pada saat evaluasi [13]. Perbandingan lebih lengkap mengenai perbandingan penelitian bisa dilihat di tabel 1.

**Tabel 1 Matrik Literatur Review dan Posisi Penelitian**

| No | Judul | Peneliti | Pokok Penelitian | Perbandingan |
|---|---|---|---|---|
| **1.** | Klasifikasi Data Tidak Seimbang Menggunakan Algoritma Smote dan K-Nearest Neighbor | Rimbun Siringoringo (2018) | Menguji implementasi algoritma SMOTE untuk mengatasi data tidak seimbang dengan Imbalance ratio pada dataset credit card fraud klasifikasi menggunakan algoritma KNN . | Penelitian yang dilakukan menggunakan lebih dari 1 algoritma klasifikasi, yaitu : C45, Naïve Bayes, K-NN, dan SVM. selain itu dataset yang digunakan menggunakan level imbalance ratio yang berbeda-beda untuk mengetahui perbedaan akurasi pada setiap level imbalance ratio. |
| **2.** | Metode Klasifikasi Data Mining dan Teknik Sampling Smote Menangani Class Imbalance untuk Segmentasi Customer pada Industri Perbankan | Akhmad Setiawan, Teguh Bharata Adji (2016) | Menguji implementasi algoritma SMOTE untuk mengatasi data tidak seimbang pada Segmentasi Customer pada Industri Perbankan klasifikasi menggunakan algoritma J48+, SVM, dan Naïve Bayes. | Penelitian yang dilakukan menggunakan algoritma klasifikasi, yaitu : C45, Naïve Bayes, K-NN, dan SVM. selain itu dataset yang digunakan menggunakan level imbalance ratio yang berbeda-beda untuk mengetahui perbedaan akurasi pada setiap level imbalance ratio. |
| **3.** | Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan | Mustaqim Mustaqim, Budi Warsito, Bayu Surarso (2019) | Menguji implementasi algoritma SMOTE untuk mengatasi data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan menggunakan algoritma klasifikasi algoritma NN, DT, SVM dan LR. | Penelitian yang akan dilakukan menggunakan algoritma klasifikasi, yaitu : C45, Naïve Bayes, K-NN, dan SVM. selain itu dataset yang digunakan menggunakan level imbalance ratio yang berbeda-beda untuk mengetahui perbedaan akurasi pada setiap level imbalance ratio. |

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

| | | | | |
|---|---|---|---|---|
| **4.** | A study of synthetic oversampling for twitter imbalanced sentiment analysis | J. Ah-Pine and E. P. S. Morales (2016) | Membahas perbandingan kemampuan handling imbalance dataset dengan metode oversampling SMOTE, Borderline SMOTE, dan ADASYN | Penelitian yang akan dilakukan akan dilakukan transformasi data, dan fitur seleksi sebelum memasuki tahap klasifikasi |
| **5.** | Komparasi Metode Undersampling Untuk Klasifikasi Teks Ujaran Kebencian | N. A. Verdikha, T. B. Adji, and A. E. Permanasari (2017) | Menguji perbandingan hasil resampling dengan menggunakan metode, SMOTE, Borderline SMOTE V1, Borderline SMOTE V2, SMOTE-SVM, dan ADASYN yang di klasifikasi oleh metode SVM, dan dilakukan evaluasi dengan skor matrik berdasarkan hasil G-Means | Pada paper ini hanya di ujikan dalam satu dataset, dan tidak ditemukan adaya proses tranformasi data dimana metode ADASYN membutuhkan distance untuk sintesis data baru, dalam penelitian ini dilibatkan proses transormasi data dan seleksi fitur serta hasil yang ditampilkan memiliki variasi dataset dengan rasio berbeda. |

Pada penelitian ini setelah implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE) terhadap dataset yang tidak seimbang selanjutnya akan dilihat akurasi model sebelum dilakukan implementasi dan sesudah dilakukan implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE) menggunakan algoritma klasifikasi. Algoritma klasiikasi yang digunakan untuk penggujian adalah C45, Naïve Bayes, KNN dan SVM. Kelas minoritas dataset imbalance akan menjadi input dari sistem, sebelum masuk kedalam system kelas minoritas harus diubah dulu ke data numerik. kemudian sistem akan memproses kelas minoritas dengan Synthetic Minority Over-sampling Technique (SMOTE).

Untuk mengukur untuk mengukur kinerja dari model yang dihasilkan salah satunya menggunakan Confusion Matrix. *confusion matrix* yang juga dikenal *dengan error matrix*, *error matrix* adalah tata letak tabel spesifik yang memungkinkan visualisasi dari sebuah algoritma [14]. Presisi atau confidence adalah proporsi kasus yang diprediksi positif yang juga positif benar pada data sebenarnya. Recall atau sensitivity adalah proporsi kasus positif yang sebenarnya diprediksi positif secara benar [15].

**Tabel 2 Model Confusion Matrix**

| **Predicted Class** | **Actual Class** | |
|---|---|---|
| | **+** | **-** |
| + | True Positives(TP) | False Positives (FP) |
| - | False Negatives (FN) | True Negatives (TN) |

Perhitungan akurasi dengan tabel Confusion Matrix adalah seperti terdapat pada persamaan (1) sebagai berikut.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

Presisi didefinisikan sebagai rasio item relevan yang dipilih terhadap semua item yang terpilih. Presisi dapat diartikan sebagai kecocokan antara permintaan informasi dengan jawaban terhadap permintaan tersebut. Rumus presisi seperti terdapat pada persamaan (2) adalah:

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

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

Recall didefinisikan sebagai rasio dari item relevan yang dipilih terhadap total jumlah item relevan yang tersedia. Recall dihitung dengan rumus seperti terdapat pada persamaan (3):

$$Recall\ atau\ sensitivity = \frac{TP}{TP+FN} \quad (3)$$

Presisi dan Recall dapat diberi nilai dalam bentuk angka dengan menggunakan perhitungan persentase (1-100%) atau dengan menggunakan bilangan antara 0-1. Sistem rekomendasi akan dianggap baik jika nilai presisi dan recallnya tinggi.

$$Specificity = \frac{TN}{TN+FP} \quad (4)$$

$$G - Mean = \sqrt{Sensitivity \times Specificity} \quad (5)$$

G-Mean adalah Indeks yang digunakan untuk mengukur keseluruhan performa dari klasifikasi atau overall classification performance seperti terdapat pada persamaan (4) dan persamaan (5) [14].

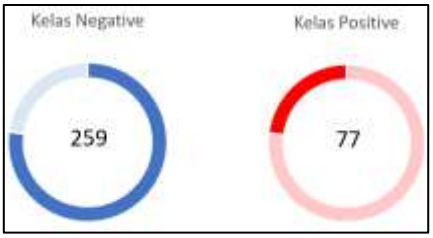
Hasil akhir dari sistem adalah data replikasi dari data minoritas. Sebelum diklasifikasikan data kelas minoritas + data replikasi, hasil smote digabung dengan data mayoritas sehingga membentuk data yang balance. Dalam penelitian ini menggunakan bahasa pemrograman python dan jupyter notebook untuk visualisasinya. Sistem dibangun dengan bahasa pemrograman python, dan visualisasi menggunakan jupyter notebook. Sistem hanya digunakan untuk implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE).

## 3 Metode Penelitian

### 3.1 Analisa Data

Data yang akan dianalisa ada 2 yaitu data imbalance yang belum dilakukan preprocessing menggunakan algoritma Synthetic Minority Over-sampling Technique (SMOTE) dan dengan data yang sudah dilakukan preprocessing dengan algoritma Synthetic Minority Over-sampling Technique (SMOTE). Dataset akan dibagi menjadi 2 bagian yaitu data untuk training dan data untuk testing. Besar data untuk training 80 % sedangkan data untuk testing sebesar 20%.

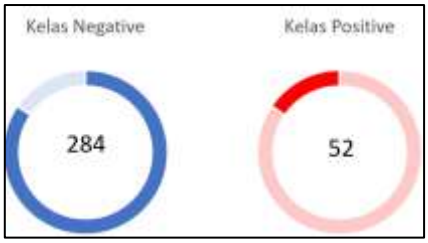
Jumlah dataset yang digunakan sebanyak 4 yaitu dataset ecoli dengan IR 3,3 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 259 instance sedangkan jumlah kelas minoritas 77 instance, ecoli dengan IR 5,4 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 284 instance sedangkan jumlah kelas minoritas 52 instance, ecoli dengan IR 8,6 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 301 instance sedangkan jumlah kelas minoritas 35 instance, dan ecoli dengan IR 15, 4 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 316 instance sedangkan jumlah kelas minoritas 20 instance. seperti ilustrasi pada gambar 1.




**Gambar 1 ilustrasi dataset ecoli dengan IR 3,3**

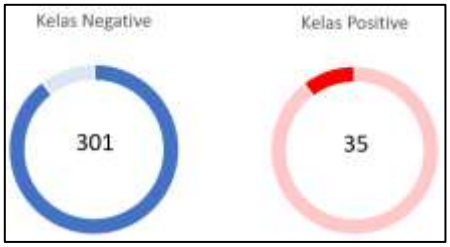
Ecoli dengan IR 5,4 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 284 instance sedangkan jumlah kelas minoritas 52 instance seperti pada ilustrasi gambar 2.

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



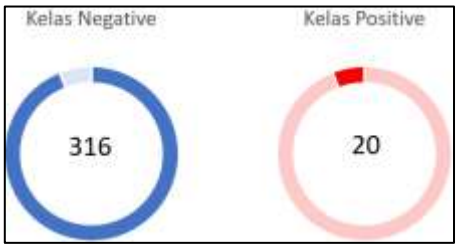

**Gambar 2 ilustrasi dataset ecoli dengan IR 5,4**

Ecoli dengan IR 8,6 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 301 instance sedangkan jumlah kelas minoritas 35 instance seperti pada ilustrasi gambar 3.




**Gambar 3 ilustrasi dataset ecoli dengan IR 8,6**

Ecoli dengan IR 15,4 memiliki jumlah 336 Instance dengan jumlah kelas mayoritas 316 instance sedangkan jumlah kelas minoritas 20 instance seperti ilustrasi pada gambar 4.




**Gambar 4 ilustrasi dataset ecoli dengan IR 15,4**

### 3.2 Proses Sistem

Dataset imbalance akan menjadi input dari sistem, kemudian sistem akan memproses data tidak seimbang dengan Synthetic Minority Over-sampling Technique (SMOTE). Hasil akhir dari sistem adalah dataset yang seimbang kemudian dilakukan klasifikasi dengan algoritma klasifikasi Naïve Bayes ,SVM, K-NN, dan dan Decision Tree. Gambaran proses sistem dan alur data yang dibuat seperti terdapat pada gambar 5.


**Gambar 5 Gambaran proses sistem**

### 3.3 Alur Penelitian

Secara lebih jelas alur penelitian yang dilakukan dalam penelitian ini seperti terdapat pada gambar 7 .

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


**Gambar 7 Alur penelitian yang dilakukan**

### 3.4 Proses SMOTE

Pada saat implementasi algoritma SMOTE, tahap ini akan menambahkan jumlah instance kelas minoritas agar seimbang terhadap kelas mayoritasnya.Peningkatan jumlah instance pada kelas minoritas, mengacu kepada jumlah presentase SMOTE serta kelas terdekat dari nilai k nearest neighbor agar menghasilkan data sintesis sesuai dengan yang diharapkan. Presentase dari SMOTE merupakan jumlah presentase dari kelas minoritas awal yang akan dibangkitkan menjadi instance sintesis. Secara umum fungsi dari algoritma SMOTE ditulis dengan SMOTE (X, N, k) , dimana X merupakan data minoritas, N merupakan presentase jumlah instance yang akan diciptakan, dan k merupakan jumlah instance terdekat dari instance yang dicari dengan rumus jarak euclidean distance dapat dihitung dengan persamaan (6).

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} + \cdots + (X_n - Y_n)^2 \quad (6)$$

Misalkan atribut dengan data 2, maka jarak euclidean distance seperti dalam persamaan (7) yaitu :

$$dist = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2} \quad (7)$$

Setelah dilakukan perhitungan jarak terdekat dari instance menggunakan Euclidean distance, setelah itu membuat data replikasi dari instance terdekat seperti dalam persamaan (8).

$$X_{syn} = X_i + (X_{knn} - X_i) \times \sigma \quad (8)$$

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

Dimana Xsyn merupakan data sintesis hasil replikasi, Xi merupakan data ke-i pada kelas minoritas, Xknn merupakan data kelas minoritas yang memiliki jarak terdekat dengan data Xi, Sedangkan σ merupakan bilangan random antara 0 dan 1.

**Tabel 3 Contoh dataset sebelum dilakukan proses SMOTE**

| No | Atribut 1 | Atribut 2 | Kelas |
|---|---|---|---|
| 1. | 1 | 2 | 1 |
| 2. | 2 | 3 | 1 |
| 3. | 4 | 3 | 1 |
| 4. | 6 | 2 | 2 |
| 5. | 6 | 4 | 2 |
| 6. | 5 | 4 | 2 |
| 7. | 4 | 4 | 2 |
| 8. | 5 | 6 | 2 |
| 9. | 6 | 3 | 2 |
| 10. | 4 | 5 | 2 |
| 11. | 6 | 7 | 2 |
| 12. | 5 | 3 | 2 |

Pada tabel 3 diketahui bahwa kelas minoritas memiliki 3 instance, sedangkan 9 instance berada pada kelas mayoritas. Berikut tahapan untuk mereplikasi data jumlah instance kelas minoritas (kelas 1). Langkah pertama, setiap instance kelas minoritas (kelas 1) dihitung jaraknya dengan menggunakan euclidean distance, untuk mendapatkan instance tetangga terdekat Xknn dengan nilai k = 3 untuk direplikasi sebagai berikut :

Data ke-1 dari kelas minoritas, dengan setiap data pada kelas minoritas :

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(1-1)^2 + (2-2)^2} = \sqrt{0}$$

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(1-2)^2 + (2-3)^2} = \sqrt{2}$$

$$d\left(\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(1-4)^2 + (2-3)^2} = \sqrt{10}$$

Jarak dari setiap data diurutkan dari yang terkecil = $\sqrt{0}$ , $\sqrt{2}$ , $\sqrt{10}$. Data ke-2 dari kelas minoritas, dengan setiap data pada kelas minoritas :

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(2-1)^2 + (3-2)^2} = \sqrt{2}$$

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(2-2)^2 + (3-3)^2} = \sqrt{0}$$

$$d\left(\begin{bmatrix}2\\3\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(2-4)^2 + (3-3)^2} = \sqrt{4}$$

Jarak dari setiap data diurutkan dari yang terkecil = $\sqrt{0}$ , $\sqrt{2}$ , $\sqrt{4}$. Data ke-3 dari kelas minoritas, dengan setiap data pada kelas minoritas :

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}1\\2\end{bmatrix}\right) = \sqrt{(4-1)^2 + (3-2)^2} = \sqrt{10}$$

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}2\\3\end{bmatrix}\right) = \sqrt{(4-2)^2 + (3-3)^2} = \sqrt{4}$$

$$d\left(\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}4\\3\end{bmatrix}\right) = \sqrt{(4-4)^2 + (3-3)^2} = \sqrt{0}$$

Jarak dari setiap data diurutkan dari yang terkecil = $\sqrt{0}$ , $\sqrt{4}$ , $\sqrt{10}$. Jumlah data kelas minoritas (kelas 1) yang semula berjumlah 3 maka untuk memperoleh keseimbangan kelas, data pada kelas minoritas harus direplikasi sebanyak dua kali (200%) N dari SMOTE = 200.

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

Langkah kedua adalah membangkitkan data sintesis, menggunakan persamaan (3). Berikut perhitungan data sintesis pada kelas minoritas (Kelas 1), dengan replikasi N = 200 Xknn yang digunakan adalah acak data dari banyak N.

Data ke-1 :

$$X_{syn} = [1,2] + ([1,2] - [1,2]) \times 0,12 = [1,2]$$
$$X_{syn} = [1,2] + ([2,3] - [1,2]) \times 0,26 = [2,26\ ,3,26]$$

Data ke-2 :

$$X_{syn} = [2,3] + ([4,3] - [2,3]) \times 0,31 = [2,62\ ,3]$$
$$X_{syn} = [2,3] + ([1,2] - [2,3]) \times 0,21 = [1,79\ ,2,79]$$

Data ke-3 :

$$X_{syn} = [4,3] + ([1,2] - [4,3]) \times 0,34 = [2,98\ ,2,66]$$
$$X_{syn} = [4,3] + ([2,3] - [4,3]) \times 0,17 = [3,66\ ,3]$$

Tabel distribusi sebelum dilakukan proses SMOTE dapat dilihat pada tabel 3, perbandingan komposisi data sebelum dan sesudah dilakukan dapat dilihat pada tabel 4. Sedangkan contoh data setelah dilakukan proses SMOTE, dapat dilihat pada tabel 5.

**Tabel 4 Contoh komposisi dataset sebelum dan setelah dilakukan proses SMOTE**

| Kelas | Data Awal | Replikasi | Data Baru |
|---|---|---|---|
| Kelas Mayoritas | 9 | - | - |
| Kelas Minoritas | 3 | 200 % | 6 |

**Tabel 5 Contoh dataset setelah dilakukan proses SMOTE**

| No | Atribut 1 | Atribut 2 | Kelas |
|---|---|---|---|
| 1. | 1 | 2 | 1 |
| 2. | 2 | 3 | 1 |
| 3. | 4 | 3 | 1 |
| 4. | 6 | 2 | 2 |
| 5. | 6 | 4 | 2 |
| 6. | 5 | 4 | 2 |
| 7. | 4 | 4 | 2 |
| 8. | 5 | 6 | 2 |
| 9. | 6 | 3 | 2 |
| 10. | 4 | 5 | 2 |
| 11. | 6 | 7 | 2 |
| 12. | 5 | 3 | 2 |
| 13.* | 1 | 2 | 1 |
| 14. * | 2.26 | 3.26 | 1 |
| 15. * | 2.62 | 3 | 1 |
| 16. * | 1.79 | 2.79 | 1 |
| 17. * | 2.98 | 2.66 | 1 |

*) data sintesis baru

### 3.5 Pengujian Akurasi

Hasil dari output sistem digunakan untuk uji akurasi penerapan algoritma Synthetic Minority Over-sampling Technique (SMOTE), dengan membandingkan dataset yang imbalance dengan dataset yang sudah dilakukan sampling menggunakan algoritma Synthetic Minority Over-sampling Technique (SMOTE) masing – masing data diklasifikasikan dengan algoritma klasifikasi kemudian hasil dari klasifikasi diukur akurasinya dengan confusion matrix. Sehingga didapat akurasi antara klasifikasi data sebelum diimplementasikannya algoritma Synthetic Minority Over-sampling Technique (SMOTE) dengan data yang sudah

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

diimplementasikan algoritma Synthetic Minority Over-sampling Technique (SMOTE). Hasil akurasi dari masing – masing data tersebut kemudian dibandingkan untuk mengetahui perbedaan sebelum dan sesudah implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE).

# 4 Hasil dan Pembahasan

## 4.1 Pra Pemrosesan Data

Proses implementasi SMOTE pada setiap dataset ecoli dengan IR 3,3, IR 5,4, IR 8,6 dan IR 15,4 ialah sebagai berikut. Jumlah awal kelas minoritas untuk IR 3,3 sebanyak 77 Instance, untuk IR 5,4 sebanyak 52 Instance, untuk IR 8,6 sebanyak 35 Instance dan 20 Instance untuk IR 15,4 maka setiap instance dapat memiliki 5 tetangga terdekat. Oleh karena itu instance tersebut dapat melihat tetangganya dengan faktor k=3, dengan N%=200 untuk IR 3,3 5,4, faktor k=5, dengan N%=400 untuk IR 5,4, faktor k=7, dengan N%=700 untuk IR 8,6, dan faktor k=15, dengan N%=1400 untuk IR 15,4 maka jumlah instance hasil dari proses SMOTE data replikasi yang dihasilkan sebanyak 154 instance, 208 instance, 245 instance dan 280 instance. Sehingga hasil akhir oversampling menggunakan SMOTE adalah jumlah instance awal ditambahkan dengan jumlah instance hasil repilkasi. Berikut Tabel 6 merupakan ilustrasi proses SMOTE pada setiap dataset ecoli.

**Tabel 6 Contoh dataset setelah dilakukan proses SMOTE**

| Dataset | Kondisi Awal Kelas Minoritas | SMOTE | | Kondisi Akhir Kelas Minoritas |
|---|---|---|---|---|
| | | K | N% | |
| Ecoli IR 3,3 | 77 Instance | 3 | 200 % | 231 Instance |
| Dari 77 Instance dinaikan menjadi (→) 231 Instance | | | | |
| Ecoli IR 5,4 | 52 Instance | 5 | 400 % | 260 Instance |
| Dari 52 Instance dinaikan menjadi (→) 260 Instance | | | | |
| Ecoli IR 8,6 | 35 Instance | 7 | 700 % | 280 Instance |
| Dari 35 Instance dinaikan menjadi (→) 280 Instance | | | | |
| Ecoli IR 15,4 | 20 Instance | 15 | 1400 % | 300 Instance |
| Dari 20 Instance dinaikan menjadi (→) 300 Instance | | | | |

Setelah dilakukan proses oversampling dengan SMOTE pada setiap dataset, jumlah instance kelas minoritas maupun menjadi relatif seimbang [3]. Apabila selisih antara kelas minoritas dengan kelas mayoritas tidak lebih dari 35% atau perbandingan kelas mayoritasnnya kurang dari dua kali ditambah satu dari kelas minoritasnya, sehingga dapat diasumsikan distribusi kelasnya relatif seimbang.

## 4.2 Implementasi Algoritma Klasifikasi

Untuk melihat perbedaan antara dataset sebelum dilakukan prapemrosesan teknik sampling SMOTE dengan dataset yang sudah dilakukan langkah prapemrosesan dengan teknik sampling SMOTE maka diimplementasikan algoritma klasifikasi Naïve Bayes, SVM, K-NN, dan Decision Tree pada dataset sebelum dan sesudah dilakukan langkah prapemrosesan dengan teknik sampling SMOTE.

a. Pengujian akurasi algoritma klasifikasi Naïve Bayes, SVM, K-NN, dan dan Decision Tree pada setiap dataset ecoli sebelum dan setelah dilakukan prapemrosesan SMOTE.
   Perhitungan akurasi dilakukan menggunakan persamaan (1) dengan mengambil data dari actual classification dan prediction classification confusion matrix klasifikasi naïve bayes pada dataset ecoli dengan IR 3,3 seperti terdapat pada tabel 7 sebelum dilakukan prapemrosesan SMOTE seperti .

   **Tabel 7 Confusion matrix klasifikasi Naïve Bayes dengan Cross-validation fold 10 pada Dataset ecoli IR 3,3**

| Actual Classification | Prediction Classification | |
|---|---|---|
| | Positive (kelas Negative) | Negative ( kelas Positive) |
| Positive ( kelas Negative) | 229 | 30 |
| Negative ( kelas Positive) | 20 | 57 |

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

$$Akurasi = \frac{TP + TN}{TP + TN + FP + FN}$$

$$= \frac{229+57}{229+57+30+20} \times 100\%$$
$$= \frac{286}{336} \times 100\% = \mathbf{85,12}\ \%$$

Hasil lebih lengkap untuk Hasil pengujian akurasi algoritma klasifikasi Naïve Bayes sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli seperti terdapat pada tabel 8, 9, 10, 11 .

**Tabel 8 Hasil pengujian akurasi algoritma klasifikasi Naïve Bayes sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | Naïve Bayes | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 85,12 % | 85,53 % | 85,51 % | 86,77 % |
| 2. | Ecoli IR 5,4 | 91,67 % | 90,88 % | 89,70 % | 90,81 % |
| 3. | Ecoli IR 8,6 | 85.71 % | 86.35 % | 89,84 % | 90,15 % |
| 4. | Ecoli IR 15,4 | 95,83 % | 93,85 % | 96,43 % | 96,42 % |

**Tabel 9 Hasil pengujian akurasi algoritma klasifikasi SVM sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | SVM | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 91,67 % | 90,68 % | 91,02 % | 91,24 % |
| 2. | Ecoli IR 5,4 | 96,13 % | 95,70 % | 95,95 % | 95,98 % |
| 3. | Ecoli IR 8,6 | 92,86 % | 92,79 % | 93,80 % | 93,48 % |
| 4. | Ecoli IR 15,4 | 98,81 % | 98,97 % | 99,02 % | 98,81 % |

**Tabel 10 Hasil pengujian akurasi algoritma klasifikasi K-NN sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | K-NN | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 90,47 % | 90,5 % | 92,04 % | 92,20 % |
| 2. | Ecoli IR 5,4 | 96,13 % | 95.68 % | 90,07 % | 94,81 % |
| 3. | Ecoli IR 8,6 | 92,86 % | 92,35 % | 94,15 % | 93,35 % |
| 4. | Ecoli IR 15,4 | 99,10 % | 98,94 % | 99,02 % | 98,93 % |

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

**Tabel 11 Hasil pengujian akurasi algoritma klasifikasi Decision Tree sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | Decision Tree | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 89,28 % | 93,42 % | 92,04 % | 93 % |
| 2. | Ecoli IR 5,4 | 94,05 % | 93 % | 92,46 % | 92,11 % |
| 3. | Ecoli IR 8,6 | 91,67 % | 91,38 % | 92,60 % | 91,26 % |
| 4. | Ecoli IR 15,4 | 96.43 % | 97,29 % | 96,27 % | 96,06 % |

b. Pengujian G-Mean algoritma klasifikasi Naïve Bayes, SVM, K-NN, dan dan Decision Tree pada setiap dataset ecoli sebelum dan setelah dilakukan prapemrosesan SMOTE

Untuk perhitungan g-mean dilakukan setelah sebelumnya menghitung presisi, recall atau sensitivity serta specificity seperti terdapat pada persamaan (2)(3)(4). Contoh perhitungan seperti terdapat di bawah ini

$$Presisi = \frac{TP}{TP + FP}$$
$$= \frac{229}{229+30} \times 100\%$$
$$= \frac{229}{259} \times 100\% = \mathbf{88,42}\%$$
$$Recall\ atau\ sensitivity = \frac{TP}{TP + FN}$$
$$= \frac{229}{29+20} \times 100\%$$
$$= \frac{229}{249} \times 100\% = \mathbf{91,97}\ \%$$
$$Specificity = \frac{TN}{TN + FP}$$
$$= \frac{57}{57+20} \times 100\% = \mathbf{88,42}\ \%$$
$$= \frac{57}{77} \times 100\% = \mathbf{74,02}\ \%$$
$$G - Mean = \sqrt{Sensitivity\ \times\ Specificity}$$
$$= \sqrt{91,97\ \times 74,02}\ = \sqrt{6807,62}\ = \mathbf{82,51}\ \%$$

Hasil lebih lengkap untuk Hasil pengujian G-Mean algoritma klasifikasi Naïve Bayes sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli seperti terdapat pada tabel 12, 13, 14, 15 .

**Tabel 12 Hasil G-Mean algoritma klasifikasi Naïve Bayes sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | Naïve Bayes | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 82,51 % | 78,54 % | 86,11 % | 86,69 % |
| 2. | Ecoli IR 5,4 | 80,85 % | 80,89 % | 91,14 % | 90,77 % |
| 3. | Ecoli IR 8,6 | 61,45 % | 62,92 % | 89,98 % | 90,24 % |
| 4. | Ecoli IR 15,4 | 76,92 % | 96,40 % | 96,42 % | 96,40 % |

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

**Tabel 13 Hasil G-Mean algoritma klasifikasi SVM sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | SVM | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 89,69 % | 88,32 % | 91,17 % | 91,31 % |
| 2. | Ecoli IR 5,4 | 92,77 % | 92,27 % | 95,98 % | 96,04 % |
| 3. | Ecoli IR 8,6 | 79,88 % | 79,93 % | 93,87 % | 93,54 % |
| 4. | Ecoli IR 15,4 | 99,37 % | 99,17 % | 98,99 % | 98,80 % |

**Tabel 14 Hasil G-Mean algoritma klasifikasi K-NN sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | K-NN | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 87,41 % | 87,73 % | 92,03 % | 92,20 % |
| 2. | Ecoli IR 5,4 | 91,61 % | 91,02 % | 95,55 % | 94,76 % |
| 3. | Ecoli IR 8,6 | 79,88 % | 78,58 % | 94,35 % | 93,58 % |
| 4. | Ecoli IR 15,4 | 99,53 % | 98,08 % | 99,01 % | 98,92 % |

**Tabel 15 Hasil G-Mean algoritma klasifikasi Decision Tree sebelum dan setelah dilakukan prapemrosesan SMOTE dengan Cross-validation fold 10 dan 80% data training pada setiap dataset ecoli**

| No | Dataset | Decision Tree | | | |
|---|---|---|---|---|---|
| | | Sebelum SMOTE | | Setelah SMOTE | |
| | | 10-fold | 80 % | 10-fold | 80 % |
| 1. | Ecoli IR 3,3 | 84.43 % | 83,96 % | 91,98 % | 92,42 % |
| 2. | Ecoli IR 5,4 | 91,43 % | 89,20 % | 92,43 % | 92,07 % |
| 3. | Ecoli IR 8,6 | 75,89 % | 74,77 % | 92,57 % | 91,26 % |
| 4. | Ecoli IR 15,4 | 84,04 % | 89,95 % | 96,29 % | 96,05 % |

### 4.3 Evaluasi

Penanganan distribusi kelas yang tidak seimbang pada dataset menggunakan algoritme SMOTE dapat meningkatkan nilai akurasi maupun g-mean pada algoritma Decision Tree. Hal tersebut menunjukkan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap pra-pemrosesan data memberikan pengaruh terhadap nilai akurasi maupun g-mean algoritme Decision Tree. Pada scenario percobaan yang telah dilakukan algoritma SVM memiliki akurasi paling baik 97,29 % pada dataset ecoli 15,8 sebelum dilakukan SMOTE dengan 80% data training, dan memiliki nilai G-mean paling baik 96,29 % pada dataset ecoli 15,8 setelah dilakukan SMOTE dengan 10 fold cross validation.

Nilai akurasi dan nilai G-mean algoritma Naïve Bayes konsisten dengan performanya pada setiap level imbalance ratio, sebelum implementasi SMOTE memiliki performa yang tidak baik, sedangkan setelah diimplementasikan SMOTE algoritma Naïve Bayes memiliki peningkatan akurasi yang konsisten. Sehingga dapat ditarik kesimpulan bahwa kombinasi SMOTE + Naïve Bayes paling efektif digunakan pada dataset imbalance dengan level yang berbeda-beda pada skema 10 fold cross validation maupun 80% data testing yang diujikan sebanyak 50 kali.

Implementasi dari algoritma SMOTE membantu dalam pengklasifikasian terhadap data kelas minoritas yang menjadi masalah pada latarbelakang penelitian ini, hal tersebut dapat dilihat pada tabel confusion matrix bagian nilai FN atau false negative dimana jika dibandingkan dengan tabel confusion matrix sebelum dan sesudah implementasi SMOTE terdapat penurunan dari nilai

FN atau false negative. Selain itu dapat dilihat pada perbandingan nilai g-mean sebelum implementasi SMOTE dan setelah implementasi SMOTE.

## 5 Kesimpulan

Penanganan distribusi kelas yang tidak seimbang pada dataset menggunakan algoritma SMOTE dapat meningkatkan nilai akurasi maupun g-mean pada algoritma Naïve Bayes, SVM, KNN dan Decision Tree. Hal tersebut menunjukkan bahwa proses penanganan terhadap distribusi kelas yang tidak seimbang pada tahap pra-pemrosesan data memberikan pengaruh terhadap nilai akurasi maupun g-mean algoritma Naïve Bayes, SVM, KNN dan Decision Tree. Pada scenario percobaan yang telah dilakukan algoritma Naïve Bayes memiliki akurasi paling baik 96,43 %, 99,02 % untuk SVM, sedangkan KNN dan Decision Tree memiliki akurasi paling baik 97,29 % pada dataset ecoli 15,8 setelah dilakukan SMOTE dengan 10 fold cross validation, dan memiliki nilai G-mean paling baik 96,42 % untuk Naïve Bayes, 99,37 % untuk SVM, 99,53 % untuk KNN serta 96,29 % untuk Decision Tree pada dataset ecoli 15,8 setelah dilakukan SMOTE dengan 10 fold cross validation.

## Ucapan Terima Kasih

Ucapan terima kasih ditujukan kepada Lembaga Penelitian Universitas AMIKOM Yogyakarta yang telah memberikan pendanaan dengan skema Internal Pemula untuk kegiatan penelitian ini.

## Referensi

[1] Han, Jiawei. Data Mining: Concepts and Techniques, Third Edition. 3rd ed. Waltham, Mass.: Morgan Kaufmann Publishers, 2012.
[2] C. M. Bishop, "Bishop - Pattern Recognition and Machine Learning - Springer 2006," *Antimicrob. Agents Chemother.*, 2014, doi: 10.1128/AAC.03728-14.
[3] S. Russell and P. Norvig, *Artificial Intelligence A Modern Approach Third Edition*. 2010.
[4] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning (Adaptive Computation and Machine Learning series)*. 2012.
[5] Y. Sun, A. K. C. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," Int. J. Pattern Recognit. Artif. Intell., 2009, doi: 10.1142/S0218001409007326.
[6] M. Bach, A. Werner, J. Żywiec, and W. Pluskiewicz, "The study of under- and over-sampling methods' utility in analysis of highly imbalanced data on osteoporosis," Inf. Sci. (Ny)., 2017, doi: 10.1016/j.ins.2016.09.038.
[7] Y. Pristyanto, S. Adi, and A. Sunyoto, "The effect of feature selection on classification algorithms in credit approval," 2019 Int. Conf. Inf. Commun. Technol. ICOIACT 2019, pp. 451–456, 2019, doi: 10.1109/ICOIACT46704.2019.8938523.
[8] Septiani, I. P. A. Citra, and A. S. A. Nugraha, "JURNAL GEOGRAFI Perbandingan Metode Supervised Classification dan Unsupervised Classification terhadap Penutup Lahan di Kabupaten Buleleng," vol. 16, no. 196, pp. 90–96, 2019, doi: 10.15294/jg.v16i2.19777.
[9] M. Mustaqim, B. Warsito, and B. Surarso, "Kombinasi Synthetic Minority Oversampling Technique (SMOTE) dan Neural Network Backpropagation untuk menangani data tidak seimbang pada prediksi pemakaian alat kontrasepsi implan," Regist. J. Ilm. Teknol. Sist. Inf., vol. 5, no. 2, p. 128, 2019, doi: 10.26594/register.v5i2.1705.
[10] A. Smote and D. A. N. Neighbor, "Klasifikasi Data Tidak Seimbang Menggunakan," vol. 3, no. 1, pp. 44–49.
[11] Hairani, N. A. Setiawan, and T. B. Adji, "Metode Klasifikasi Data Mining dan Teknik Sampling Smote ... (Hairani dkk.)," Semin. Nas. Sains dan Teknol., pp. 168–172, 2016.
[12] J. Ah-Pine and E. P. S. Morales, "A study of synthetic oversampling for twitter imbalanced sentiment analysis," CEUR Workshop Proc., vol. 1646, pp. 17–24, 2016.
[13] N. A. Verdikha, T. B. Adji, and A. E. Permanasari, "Komparasi Metode Oversampling Untuk Klasifikasi Teks Ujaran Kebencian," Semant. 2017 Komparasi, pp. 195–202, 2018.
[14] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation Measures for Models Assessment over Imbalanced Data Sets," J. Inf. Eng. Appl., vol. 3, no. 10, pp. 27–38, 2013, [Online]. Available: http://www.iiste.org/Journals/index.php/JIEA/article/view/7633.
[15] A. Ilham, "Komparasi Algoritma Kasifikasi dengan Pendekatan Level Data Untuk Menangani Data Kelas Tidak Seimbang," J. Ilm. Ilmu Komput.,vol.3, no.1, pp.1–6,2017,doi: 10.35329/jiik.v3i1.60