# Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung

Akhmad Syukron
Program Studi Sistem Informasi Akuntansi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
akhmad.khy@bsi.ac.id

Sardiarinto
Program Studi Sistem Informasi Akuntansi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
sardiarinto.sdo@bsi.ac.id

Eko Saputro
Program Studi Sistem Informasi
Universitas Bina Sarana Informatika
Jakarta, Indonesia
eko.eto@bsi.ac.id

Pudji Widodo
Program Studi Teknik Komputer
Universitas Bina Sarana Informatika
Jakarta, Indonesia
pudji.piw@bsi.ac.id

***Abstract*—Classification is a process carried out to find a model with the aim of estimating the class of an object whose class is unknown. One of the problems encountered in classification is about class imbalances (class imbalances) in which a dataset contains a number of classes whose data are uneven so that it has an adverse impact on the classification results. How to overcome class imbalance in classification by applying the SMOTE (Synthetic Minority Over-sampling Technique) method. Testing data on heart failure by applying the SMOTE method can improve the performance accuracy of several classification algorithms. The performance results obtained show that the SMOTE Random Forest classifier model has a higher accuracy value compared to other models. The accuracy value is 0.881 or 88.1% and the AUC value is 0.947 or 94.7%. From this it can be concluded that the algorithm with the best performance is SMOTE Random Forest.**

**Keywords—Classification, Class Imbalance, Smote, Heart Failure**

***Abstrak*— Klasifikasi adalah suatu proses yang dilakukan untuk menemukan sebuah model dengan tujuan untuk memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui. Salah satu permasalahan yang dihadapi pada klasifikasi adalah tentang ketidakseimbangan kelas (*imbalance Class*) yang mana suatu dataset terdapat jumlah jumlah kelas yang datanya tidak merata Sehingga memberikan dampak yang tidak baik pada hasil klasifikasi. Cara mengatasi ketidakseimbangan kelas pada klasifikasi dengan menerapkan metode SMOTE (*Synthetic Minority Over-sampling Technique*). Pengujian data penyakit gagal jantung dengan penerapan metode metode SMOTE dapat meningkatkan permforma akurasi dari beberapa algoritma klasifikasi. Hasil kinerja yang didapatkan menunjukkan bahwa dengan model pengklasifikasi SMOTE Random Forest punya nilai akurasi yang lebih tinggi dibandingkan dengan model yang lain. Untuk nilai accuracy sebesar 0,881 atau 88,1% dan nilai AUCnya sebesar 0.947 atau 94,7%. Dari sini dapat disimpulkan bahwa algoritma dengan performa terbaik yaitu SMOTE Random Forest.**

***Keywords*— *Klasifikasi, Ketidakseimbangan Kelas, Smote, Gagal Jantung***

## PENDAHULUAN

Klasifikasi adalah suatu proses yang dilakukan untuk menemukan sebuah model yang menjelaskan serta membedakan sebuah konsep atau kelas data dengan tujuan untuk memperkirakan kelas dari suatu objek yang kelasnya tidak diketahui[6]. Salah satu permasalahan yang dihadapi pada klasifikasi adalah tentang ketidakseimbangan kelas (*imbalance Class*) yang mana suatu dataset terdapat jumlah kelas yang datanya memiliki perbedaan yang signifikan antara kelas minor dengan kelas mayor[3]. Sehingga dapat mengakibatkan klasifikasi yang dihasilkan tidak sesuai, karena kelas minoritas sering salah diklasisifikasi sebagai kelas mayoritas. [13]. Pada penelitian sebelumnya, yang dilakukan oleh Laila, Hikmah dan Megasari tentang penerapan teknik oversampling, underampling dan SMOTE untuk mengatasi ketidakseimbangan kelas dengan menggunakan dua metode klasifikasi SVM serta Random forest untuk klasifikasi bagi penerima bidikmisi seprovinsi Jawa Timur pada Tahun 2017[12]. Berdasarkan permasalahan yang ada pada ketidakseimbangan kelas, maka pada penelitian ini akan membahas tentang cara mengatasi ketidakseimbangan kelas pada klasifikasi dengan menerapkan metode SMOTE pada algoritma klasifikasi. Penerapan SMOTE pada algoritma klasifikasi dapat mengurangi terjadinya overfitting yang merupakan salah satu kelemahan dari teknik oversampling. Tujuan dari penelitian ini adalah untuk mengatasi masalah ketidakseimbangan kelas dan untuk meningkatkan nilai akurasi pada algoritma klasisifikasi pada prediksi penyakit gagal jantung dengan membandingkan 8 algoritma klasifikasi yang terdiri dari algoritma C45, SVM, Naïve Bayes, K-NN, Neural Network, Random Forest, Adaboost, Bagging untuk pemodelan dan pengujian unuk menentukan algoritma klasifikasi terbaik pada prediksi gagal jantung.

a) Gagal Jantung

Gagal jantung adalah kondisi dimana jantung tidak mampu memompa darah sesuai dengan kebutuhan jaringan. Gagal jantung kronis adalah masalah utama di seluruh dunia. Gagal jantung kronis dapat menyerang siapa saja, tua atau muda, namun lebih sering menyerang orang tua [8]. Gagal

DOI: https://doi.org/10/25047/jtit.v10i1.312 

jantung merupakan penyakit yang mengakibatkan mortalitas dan morbiditas yang tinggi[13]. Penyakit kardiovaskular merupakan sekelompok penyakit gangguan jantung dan pembuluh darah yang merupakan penyakit jantung koroner, jantung rematik, serebrofaskular, dan gangguan jantung lainnya[5].

b) Klasifikasi

Klasifikasi adalah teknik yang digunakan untuk menemukan pola untuk menjelaskan atau membedakan konsep atau kelas data. Tujuannya adalah untuk dapat memperkirakan kelas dari suatu objek yang pengenalnya tidak diketahui[6]. Klasifikasi merupakan teknik penambangan data yang menetapkan kelas ke kumpulan data untuk memungkinkan prediksi dan analisis yang lebih akurat[1]. Definisi klasifikasi oleh para ahli data mining sangat bervariasi. Ini pada dasarnya melibatkan mendapatkan model atau pola dengan memeriksa proses dan menggunakan algoritma klasifikasi yang ada[2].

c) Ketidakseimbangan Kelas(Imbalance Class)

Ketidakseimbangan kelas adalah jumlah data yang tidak seimbang antar kelas. Adanya ketidakseimbangan kelas dapat membuat metode klasifikasi cenderung mengklasifikasikan kelas yang tidak sesuai/kelas mayoritas terhadap kelas yang benar/kelas minoritas[7]. Beberapa metode yang diguanakan untuk mengatasi ketidakseimbangan kelas yaitu undersampling dan oversampling. Undersampling merupakan teknik resampling untuk menyeimbangkan dataset dengan mengurangi data pada kelas mayor. Sedangkan Oversampling dalah tenik resampling untuk menyeimbangkan dataset dengan meningkatkan ukuran pada kelas minor. Salah satu teknik oversampling adalah teknik SMOTE[3].

d) SMOTE(Synthetic Minority Over-sampling Technique)

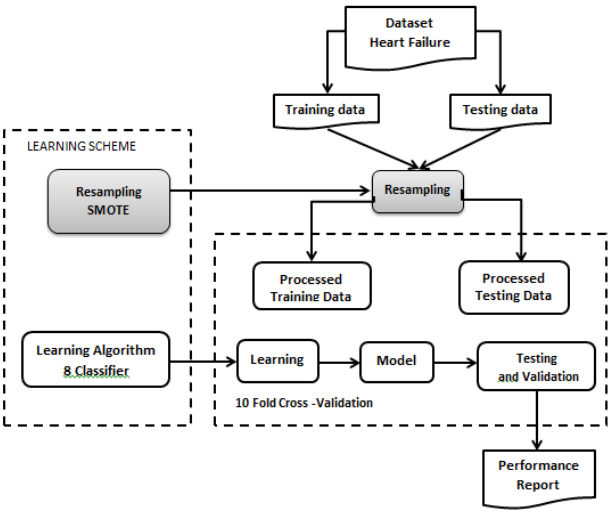
Synthetic Minority Oversampling Technique (SMOTE) algoritma preprocessing yang paling banyak digunakan dalam permasalahan data yang tidak seimbang[14]. SMOTE bekerja dengan memodifikasi kumpulan data yang tidak seimbang dengan membuat kumpulan data sintetik baru dari kelas minoritas dengan tujuan meningkatkan efisiensi metode klasifikasi[4].

e) Random Forest

Random forest merupakan metode pembelajaran ansambel yang digunakan untuk klasifikasi yang bekerja dengan membangun beberapa pohon keputusan dan menghasilkan keluaran berupa prediksi mode kelas atau rata-rata pohon individu[14]. Random Forest diartikan sebagai prinsip umum dari sekumpulan pohon keputusan acak[10]. Random forest didefinisikan sebagai kombinasi dari pohon keputusan. Jumlah pohon keputusan mempengaruhi keakuratan struktur acak secara keseluruhan[9].

## METODE PENELITIAN

Untuk memudahkan dalam melaksanakan penelitaian, maka dibuatlah alur penelitian mulai dari pengumpulan data, preprocessing, pemodelan, pengujian, serta evaluasi. Untuk tahapan penelitian yang akan dilakukan dapat dilihat pada gambar 1.




Gambar 1. Metode Usulan

Dari gambar 1 dapat dilihat alur penelitian yang akan dilakukan , dimulai dari pengumpulan dataset yang diambil dari data sekunder yang ada di website kaggle.com. Data yang digunakan adalah dataset Heart Failure dengan jumlah atribut 12 dengan atribut kelas, dan jumlah data 299 instance. Kemudian dilakukan proses resampling dengan menggunakan metode SMOTE (*Synthetic Minority Oversampling Technique*) untuk mengatasi ketidakseimbangan kelas pada dataset Heart Failure. Dari dataset baru yang telah dilakukan resampling, kemudian dilakukan pemodelan menggunakan 8 algoritma klasifikasi, yaitu Naive Bayes, C45, SVM, Random Forest, Neural Network, KNN, Adabost dan Bagging. Dengan evaluasi pengujian menggunakan 10 *Fold Crosvalidation* untuk menghitung nilai akurasi, *F-Measure* serta *Area under Roc Curve* (AUC).

## HASIL DAN PEMBAHASAN

Pada bagian ini akan membahas tentang hasil eksperimen yang sudah dilakukan menggunakan aplikasi Weka 3.8.1 dengan membuat pemodelan tanpa menggunakan teknik resampling dan yang menggunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) dengan menggunakan 8 algoritma klasifikasi untuk pemodelan dan pengujian. Kemudian hasil pengujian yang telah dilakukan akan dibandingkan untuk menentukan algoritma dengan nilai akurasi terbaik.

### *A. Dataset*

Pada penelitian ini menggunakan data sekunder yang diambil dari website kaggle.com yaitu dataset Heart Failure Prediction. Yang terdiri dari 12 Atribute utama(age, anaemia, creatinine phosphokinase, Diabetes, ejection fraction, high_blood_pressure, platelets, serum creatinine, serum sodium, sex, Smoking, time) dan 1 Atribut kelas binner (death_event) dan jumlah instance 299 yang terdiri dari 96 kelas yes, dan 203 instance kelas no.

### *B. Klasifikasi Random Forest*

Pada tahapan ini dilakukan experiment untuk membuat pemodelan dan pengujian terhadap algoritma random forest pada prediksi gagal jantung tanpa menggunakan proses resampling. Hasil Pengujian akan dibandingkan dengan 7 algoritma klasifikasi yang lain untuk menentukan model yang

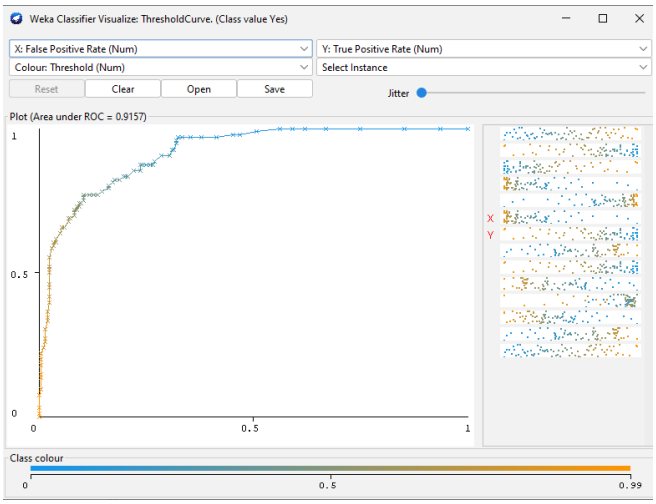

DOI: https://doi.org/10.25047/jtit.v10i1.312 

menghasilkan nilai akurasi yang paling baik. Untuk hasil uji klasifikasi tanpa resampling tersaji pada tabel 1.

Tabel 1. Perbandingan Kinerja Akurasi Pemodelan Tanpa Resampling

| No | *Algoritm* | *Preci-ssion* | *Recall* | *Accuracy* | *F-Measure* | *AUC* |
|---|---|---|---|---|---|---|
| 1 | C45 | 0,803 | 0,806 | 0,806 | 0,804 | 0,750 |
| 2 | SVM | 0,833 | 0,836 | 0,836 | 0,831 | 0,789 |
| 3 | Naïve Bayes | 0,770 | 0,776 | 0,776 | 0,761 | 0,852 |
| 4 | Random Forest | 0,847 | 0,849 | 0,849 | 0,848 | 0,916 |
| 5 | Neural Network | 0,738 | 0,742 | 0,742 | 0,740 | 0,792 |
| 6 | KNN | 0,689 | 0,709 | 0,709 | 0,663 | 0,655 |
| 7 | Adabost | 0,833 | 0,836 | 0,836 | 0,833 | 0,876 |
| 8 | Bagging | 0,833 | 0,836 | 0,831 | 0,836 | 0,889 |

Berdasarkan hasil yang diperoleh dari tabel 1, memperlihatkan bahwa nilai akurasi terendah pada algoritma K-NN dengan nilai akurasi 0,709 atau 70,9% dan nilai AUC sebesar 0,655 atau 65,5%. Sedangkan Nilai akurasi tertinggi yaitu pada algoritma random forest denga nilai akurasi sebesar 0,849 atau 84,9% dan nilai AUC sebesar 0,916 atau 91,6%. Maka dapat disimpulkan bahwa algoritma random forest memiliki tingkat akurasi yang paling tinggi yaitu sebesar, 0,849 dan nilai F-Measure:0,848, dan AUC: 0,916 untuk memprediksi gagal jantung. Untuk gambar Area Under Roc Curve klasifikasi dengan Algoritma Random forest dapat di lihat pada Gambar 2.




Gambar 2. Area Under Roc Curve Random Forest

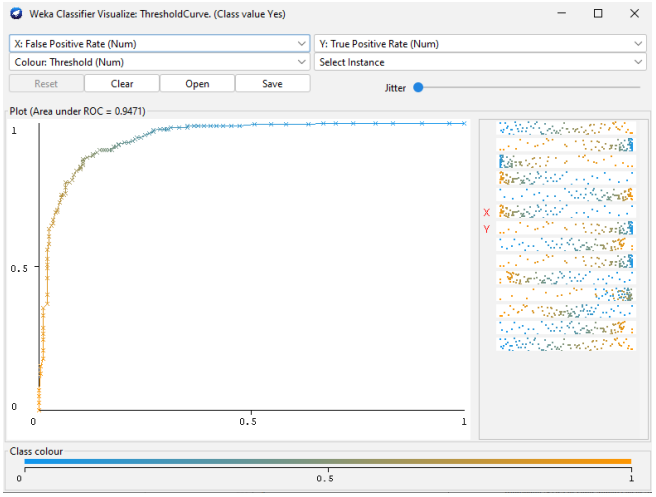
*C. Klasifikasi Smote dengan Random Forest*

Eksperimen dilakukan dengan penerapan metode SMOTE pada algoritma klasifikasi random forest. Dimana dataset yang sebelunya tidak seimbang, dengan penerapan smote maka data akan menjadi lebih seimbang dengan melakukan oversampling pada data minoritas. Jumlah instace pada dataset heart failure prediction berubah menjadi 395 instance, dengan jumlah kelas yes pada atribut death_event menjadi 192 dan kelas no 203. Dari hasil pemodelan dan pengujian pada algoritma klasifikasi dapat dirangkum dalam tabel 2.

Tabel 2. Perbandingan Kinerja Akurasi Pemodelan Dengan Resampling

| No | *Algoritm* | *Prec-ission* | *Recall* | *Accuracy* | *F-Measure* | *AUC* |
|---|---|---|---|---|---|---|
| 1 | Smote+C45 | 0,821 | 0,820 | 0,820 | 0,820 | 0,842 |
| 2 | Smote+SVM | 0,823 | 0,823 | 0,823 | 0,823 | 0,823 |
| 3 | Smote+Naïve Bayes | 0,789 | 0,782 | 0,780 | 0,780 | 0,868 |
| 4 | Smote+Random Forest | 0,881 | 0,881 | 0,881 | 0,881 | 0,947 |
| 5 | Smote+Neural Network | 0,752 | 0,752 | 0,752 | 0,752 | 0,841 |
| 6 | Smote+KNN | 0,740 | 0,739 | 0,739 | 0,739 | 0,804 |
| 7 | Smote+Adabost | 0,832 | 0,830 | 0,830 | 0,830 | 0,915 |
| 8 | Smote+Bagging | 0,861 | 0,861 | 0,861 | 0,861 | 0,918 |

Berdasarkan tabel 2, maka dapat disimpulkan bahwa penerapan metode SMOTE (Synthetic Minority Over-sampling Technique) dapat meningkatkan permforma akurasi dari 6 algoritma klasifikasi. Penerapan SMOTE pada algoritma C45 mengalami peningkatan akurasi sebesar 0,014. Penerapan SMOTE pada algoritma naïve Bayes mengalami peningkatan akurasi sebesar 0,004. Penerapan SMOTE pada Algoritma Neural Network mengalami peningkatan sebesar 0,010. Sedangkan pada algoritma K-NN mengalami peningkatan akurasi sebesar 0,030. Penerapan SMOTE pada algoritma Bagging mengelami peningkatan akurasi sebesar 0,030. Pada algoritma Random Forest mengalami peningkatan akurasi sebesar 0,032. Sedangkan pada algoritma SVM dan Adaboost mengalami penurunan akurasi setelah penerapan SMOTE sebesar 0,013 dan 0,006. Dari 8 algotitma klasifikasi yang digunakan, peningkatan akurasi tertinggi pada algoritma random forest. Berikut adalah gambar dari Area Under Roc Curve klasifikasi dengan Algoritma Random forest dengan penerapan SMOTE.



*Gambar 3. AUC SMOTE Random Forest*

## KESIMPULAN

Dengan menerapkan metode SMOTE ini dapat meningkatkan permforma akurasi dari beberapa algoritma klasifikasi kecuali metode SVM dan adabost. Maka disimpulkan bahwa metode SMOTE dapat meningkatkan performa akurasi secara efektif pada klasifikasi tidak seimbang untuk prediksi penyakit gagal jantung. Hasil kinerja yang diperoleh menunjukan bahwa model pengklasifikasi SMOTE Random Forest memiliki nilai accuracy yang lebih tinggi dibandingkan dengan beberapa model lainnya dengan nilai accuracy sebesar 0,881 atau 88,1% yang dan nilai AUC sebesar 0.947 atau 94,7%. Maka dapat disimpulkan bahwa Algoritma yang memiliki performa terbaik adalah adalah SMOTE Random Forest. Maka dapat disimpulkan bahwa Algoritma tersebut dapat digunakan untuk menyelesaikan masalah pada klasifikasi penentuan penyakit gagal jantung.

## PENGHARGAAN

Ucapan terima kasih kepada Universitas Bina Sarana Informatika, serta rekan-rekan dosen yang telah memberikan kontribusinya dalam penelitian ini. Sehingga Penelitian ini dapat terlaksana dan diselesaikan dengan baik.

## REFERENSI

[1]Apriliah, W., Kurniawan, I., Baydhowi, M., & Haryati, T. (2021). Prediksi Kemungkinan Diabetes pada Tahap Awal Menggunakan Algoritma Klasifikasi Random Forest. *Jurnal Sistem Informasi*, *10*(1), 163–171. http://sistemasi.ftik.unisi.ac.id

[2] Arhami, M., & Nasir, M. (2020). *Data Mining - Algoritma dan Implementasi*. Andi.

[3]Astuti, Dwi, F., & Lenti, Nova, F. (2021). Implementasi SMOTE Untuk mengatasi Imbalance Class Pada Klasifikasi Car Evolution Menggunakan K-NN. *Jurnal Jupiter*, *13*(1), 89–98.

[4]Cahyaningtyas, C., Nataliani, Y., & Widiasari, I. R. (2021). Analisis Sentimen Pada Rating Aplikasi Shopee Menggunakan Metode Decision Tree Berbasis SMOTE. *Aiti*, *18*(2), 173–184. https://doi.org/10.24246/aiti.v18i2.173-184

[5]Falbanyo, Alvian, R. (2022). *Ilmu Keperawatan Komunitas*. Nasya Expanding Management.

[6]Fitriani, Dwi, R., Yasin, H., & Tarno. (2021). Penanganan Klasifikasi Kelas Data Tidak Seimbang Dengan Random Over Sampling Pada Naive Bayes. *Jurnal Gausian*, *10*, 11–20.

[7]Hairani, H., Suweleh, A. S., & Susilowaty, D. (2020). Penanganan Ketidak Seimbangan Kelas Menggunakan Pendekatan Level Data. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *20*(1), 109–116. https://doi.org/10.30812/matrik.v20i1.846

[8]Lumi, A. P., Joseph, V. F. F., & Polii, N. C. I. (2021). Rehabilitasi Jantung pada Pasien Gagal Jantung Kronik. *Jurnal Biomedik:JBM*, *13*(3), 309. https://doi.org/10.35790/jbm.v13i3.33448

[9]Mu'Alim, F., & Hiday, R. (2022). Implementasi Metode Random Forest Untuk Penjurusan Siswa di Madrasah Aliyah Negeri Sintang. *Jupiter*, *14*(1), 116–125. https://www.neliti.com/publications/441871/implementasi-metode-random-forest-untuk-penjurusan-siswa-di-madrasah-aliyah-nege#cite

[10]Normah, Rifai, B., Vambudi, S., & Maulana, R. (2022). Analisa Sentimen Perkembangan Vtuber Dengan Metode Support Vector Machine Berbasis SMOTE. *Jurnal Teknik Komputer AMIK BSI*, *8*(2), 174–180. https://doi.org/10.31294/jtk.v4i2

[11]Nursita, H., & Pratiwi, A. (2020). Peningkatan Kualitas Hidup Pada Pasien Gagal Jantung: A Narrative Review Article. *Jurnal Berita Ilmu Keperawatan*, *13*(1), 11. https://doi.org/10.23917/bik.v13i1.11916

[12]Qadrini, L., Hikmah, H., & Megasari, M. (2022). Oversampling, Undersampling, Smote SVM dan Random Forest pada Klasifikasi Penerima Bidikmisi Sejawa Timur Tahun 2017. *Journal of Computer System and Informatics (JoSYC)*, *3*(4), 386–391. https://doi.org/10.47065/josyc.v3i4.2154

[13]R, & Siringoringo. (2018). Klasifikasi Data Tidak Seimbang Menggunakan Algoritma SMOTE dan KNearest Neighbor. *Jurnal ISD*, *3*(1).

[14]Ramadhan, N. G., & Adhinata, F. D. (2021). Teknik Smote Dan Gini Score Dalam Klasifikasi Kanker Payudara. *RADIAL : Jurnal Peradaban Sains, Rekayasa Dan Teknologi*, *9*(2), 125–134. https://doi.org/10.37971/radial.v9i2.229



DOI: https://doi.org/10.25047/jtit.v10i1.312 