IMPLEMENTASI TEKNIK *OVERSAMPLING* SMOTE-NC PADA ALGORITMA *K-NEAREST NEIGHBOR*

WINA PRIANI
H1091211005

SKRIPSI

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

IMPLEMENTASI TEKNIK *OVERSAMPLING* SMOTE-NC PADA ALGORITMA *K-NEAREST NEIGHBOR*

WINA PRIANI

H1091211005

SKRIPSI

Sebagai salah satu syarat untuk memperoleh gelar
Sarjana Statistika pada Program Studi Statistika

PROGRAM STUDI STATISTIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS TANJUNGPURA
PONTIANAK
2025

LEMBAR PENGESAHAN

Judul Tugas Akhir	:	Implementasi Teknik *Oversampling* SMOTE-NC pada Algoritma *K-Nearest Neighbor*
Nama Mahasiswa	:	Wina Priani
NIM	:	H1091211005
Jurusan/Program Studi	:	Matematika/Statistika
Tanggal Lulus	:	19 Mei 2025
SK Pembimbing	:	No. 2585/UN22.8/TD.06/2024/Tanggal 2 September 2024
SK Penguji	:	No. 808/UN22.8/TD.06/2025/Tanggal 10 Maret 2025

Dosen Pembimbing

Pembimbing I	Pembimbing II
Dr. Evy Sulistianingsih, M.Sc.	Nurfitri Imro'ah, M.Si.
NIP 198502172008122006	NIP 198907182019032021

Dosen Penguji

Ketua Penguji	Anggota Penguji
Shantika Martha, M.Si.	Hendra Perdana, M.Sc.
198403082008122003	198810102019031020

Pimpinan Sidang (merangkap anggota penguji)	Sekretaris Sidang (merangkap anggota penguji)
Dr. Evy Sulistianingsih, M.Sc.	Nurfitri Imro'ah, M.Si.
NIP 198502172008122006	NIP 198907182019032021

Mengesahkan
Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Tanjungpura

Prof. Dr. Gusrizal, M.Si.
NIP 197108022000031001

PERNYATAAN INTEGRITAS AKADEMIK

Saya yang bertanda tangan di bawah ini:

Nama : Wina Priani

NIM : H1091211005

Program Studi/ Jurusan : Statistika/ Matematika

Fakultas : Matematika dan Ilmu Pengetahuan Alam

dengan ini menyatakan bahwa dokumen ilmiah Tugas Akhir yang disajikan ini tidak mengandung unsur pelanggaran integritas akademik sesuai Peraturan Menteri Pendidikan, Kebudayaan, Riset, Dan Teknologi Republik Indonesia Nomor 39 Tahun 2021. Apabila di kemudian hari dokumen ilmiah Tugas Akhir ini mengandung unsur pelanggaran integritas akademik sesuai ketentuan perundangan tersebut, maka saya bersedia menerima sanksi akademik dan/atau sanksi hukum yang berlaku.

Demikian pernyatan ini untuk digunakan sebagaimana mestinya.

Pontianak, 19 Mei 2025

Wina Priani
H1091211005

Implementasi Teknik *Oversampling* SMOTE-NC Pada Algoritma *K-Nearest Neighbor*

Abstrak

Synthetic Minority Oversampling Technique-Nominal Continuous (SMOTE-NC) adalah metode *oversampling* yang digunakan untuk menangani ketidakseimbangan kelas pada data dengan kombinasi fitur numerik dan kategorikal. Teknik ini membuat data sintetis pada kelas minoritas dengan mempertimbangkan kedua jenis fitur tersebut agar distribusi data lebih seimbang. Algoritma *K-Nearest Neighbor* (K-NN) adalah algoritma klasifikasi yang bekerja dengan cara mencari sejumlah tetangga terdekat (berdasarkan jarak) dari data yang akan diprediksi dan menentukan kelasnya berdasarkan mayoritas kelas dari tetangga tersebut. Algoritma ini sederhana dan efektif untuk klasifikasi data. Penelitian ini membahas penerapan teknik *oversampling* SMOTE-NC pada algoritma K-NN untuk mengatasi ketidakseimbangan kelas pada data pasien gagal jantung. Data yang digunakan adalah *Heart Failure Clinical Record* dari Kaggle, yang mencakup 299 pasien dengan 11 atribut independen dan 1 atribut dependen. Setelah proses *pre-processing*, data dibagi menjadi data latih (70%) dan data uji (30%). SMOTE-NC diterapkan untuk meningkatkan jumlah data kelas minoritas (pasien meninggal) menjadi seimbang dengan kelas mayoritas (pasien selamat). Algoritma K-NN digunakan untuk klasifikasi dengan berbagai nilai parameter *K*. Evaluasi dilakukan menggunakan *confusion matrix*, dan difokuskan pada nilai sensitivitas sebagai ukuran kinerja model. Hasil penelitian menunjukkan bahwa nilai sensitivitas tertinggi yang diperoleh adalah sebesar 72,41%. Hal ini menunjukkan bahwa penerapan SMOTE-NC pada algoritma K-NN cukup efektif dalam meningkatkan kemampuan model dalam mendeteksi pasien yang meninggal akibat gagal jantung.

Kata Kunci: SMOTE-NC, K-NN, Sensitivitas, Klasifikasi.

Implementation Of SMOTE-NC Oversampling Technique in the K-Nearest Neighbor Algorithm

Abstract

Synthetic Minority Oversampling Technique–Nominal Continuous (SMOTE-NC) is an oversampling method used to handle class imbalance in datasets that contain a combination of numerical and categorical features. This technique generates synthetic data for the minority class by considering both types of features to create a more balanced data distribution. The K-Nearest Neighbor (K-NN) algorithm is a classification algorithm that works by identifying a number of nearest neighbors (based on distance) to the data point being predicted and assigning its class based on the majority class among those neighbors. This algorithm is simple and effective for data classification. This study discusses the application of the SMOTE-NC oversampling technique to the K-NN algorithm to address class imbalance in heart failure patient data. The dataset used is the Heart Failure Clinical Record from Kaggle, consisting of 299 patients with 11 independent attributes and 1 dependent attribute. After pre-processing, the data was divided into training data (70%) and testing data (30%). SMOTE-NC was applied to increase the number of minority class data (deceased patients) to be balanced with the majority class (surviving patients). The K-NN algorithm was used for classification with various K parameter values. Evaluation was conducted using a confusion matrix, focusing on sensitivity as the primary performance measure. The results showed that the highest sensitivity achieved was 72.41%. This indicates that the application of SMOTE-NC to the K-NN algorithm is reasonably effective in improving the model's ability to detect patients who died from heart failure.

Keywords: SMOTE-NC, K-NN, Sensitivity, Classification.

PRAKATA

Puji dan syukur penulis panjatkan kehadirat Tuhan Yang Maha Esa karena atas segala berkat, rahmat, dan hidayah-Nya, sehingga penulis dapat menyelesaikan skripsi. Skripsi yang berjudul **"Implementasi Teknik *Oversampling* SMOTE-NC pada Algoritma *K-Nearest Neighbor*"** merupakan salah satu syarat untuk memperoleh gelar Sarjana Statistika pada Program Studi Statistika di Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Tanjungpura, Pontianak.

Penulisan skripsi ini tidak terlepas dari dukungan dan bimbingan dari beberapa pihak, oleh karena itu penulis ingin mengucapkan terima kasih yang sebesar-besarnya kepada:

1. Kedua orang tua saya, adik-adik saya, dan keluarga saya yang sangat saya cintai yang telah memberikan dorongan, nasehat, serta dukungan moril dan materil.
2. Ibu Dr. Evy Sulistianingsih, M.Sc. selaku Dosen Pembimbing Pertama yang telah memberikan arahan dan motivasi untuk menyelesaikan penulisan skripsi ini.
3. Ibu Nurfitri Imro'ah, M.Si., selaku Dosen Pembimbing Kedua yang juga telah memberikan arahan serta motivasi untuk menyelesaikan penulisan skripsi ini.
4. Ibu Shantika Martha, M.Si., selaku Dosen Penguji Pertama yang telah memberikan saran dan masukan untuk perbaikan skripsi ini sehingga menjadi lebih baik.
5. Bapak Hendra Perdana, M.Sc., selaku Dosen Penguji Kedua yang telah memberikan saran dan masukan untuk perbaikan skripsi ini sehingga menjadi lebih baik.
6. Teman-teman Statistika Angkatan 2021 yang telah membersamai ketika perkuliahan terutama kepada Diva Rahma Kamila, Fadhela Trifaiza, Hana Salsabila, Mira Asmara dan Sy. Farini Nurhaliza yang telah memberikan semangat dalam penulisan skripsi ini.
7. Semua orang yang saya kenal yang memberikan dukungan, nasehat, motivasi, dan doa yang tidak dapat saya sebutkan satu persatu.

Penulis berharap dapat memperoleh saran dan masukan untuk kesempurnaan skripsi ini. Semoga skripsi ini dapat bermanfaat bagi semua pihak yang membutuhkannya. Akhir kata penulis mengucapkan terima kasih kepada semua pihak yang telah berperan dari awal hingga akhir dalam penulisan skripsi ini.

Pontianak, 19 Mei 2025

Wina Priani

DAFTAR ISI

LEMBAR PENGESAHAN ..... **ii**
PERNYATAAN INTEGRITAS AKADEMIK ..... **iii**
Abstrak ..... **iv**
Abstract ..... **v**
PRAKATA ..... **vi**
DAFTAR ISI ..... **viii**
DAFTAR GAMBAR ..... **x**
DAFTAR TABEL ..... **xi**
DAFTAR LAMPIRAN ..... **xii**
DAFTAR SIMBOL ..... **xiii**
BAB I PENDAHULUAN **1**
1.1 Latar Belakang 1
1.2 Rumusan Masalah 4
1.3 Tujuan Penelitian 4
1.4 Batasan Masalah 4
1.5 Tinjauan Pustaka 4
1.6 Metodologi Penelitian 6
BAB II LANDASAN TEORI **9**
2.1 *Data Mining* 9
2.2 *Imbalanced* Data 10
2.3 Gagal Jantung 11
BAB III *SYNTHETIC MINORITY OVERSAMPLING TECHNIQUE-NOMINAL CONTINOUS* DAN ALGORITMA *K-NEAREST NEIGHBOR* .. **13**
3.1 *Synthetic Minority Oversampling Technique-Nominal Continous* (SMOTE-NC) 13
3.2 Algoritma *K-Nearest Neighbor* 15
3.3 Klasifikasi 16
3.4 Evaluasi Kinerja Klasifikasi 17
BAB IV HASIL DAN PEMBAHASAN **19**

4.1 Data Penelitian.....19
4.2 Statistika Deskriptif.....20
4.3 Pembagian Data Latih dan Data Uji.....22
4.4 Penerapan *Synthetic Minority Oversampling Technique-Nominal Continous* (SMOTE-NC).....23
4.5 Penerapan Algoritma *K-Nearest Neighbor*.....25
BAB V PENUTUP..... 29
5.1. Kesimpulan.....30
5.2. Saran.....30
DAFTAR PUSTAKA..... 31
LAMPIRAN.....35

DAFTAR GAMBAR

Gambar 1. 1 *Flowchart* Implementasi SMOTE-NC pada Algoritma K-NN 8
Gambar 4. 1 Diagram Lingkaran Status Hidup Pasien..... 20

DAFTAR TABEL

Tabel 3. 1 *Confusion Matrix*..... 17
Tabel 3. 2 Performa Kinerja Klasifikasi..... 18
Tabel 4. 1 Atribut Penelitian 19
Tabel 4. 2 Statistik Deskriptif Atribut Independen Numerik 21
Tabel 4. 3 Statistik Deskriptif Atribut Independen Kategorikal..... 21
Tabel 4. 4 Pembagian Data Latih dan Data Uji..... 22
Tabel 4. 5 Tiga Tetangga Terdekat untuk Data Minoritas Pertama 24
Tabel 4. 6 Normalisasi Data Latih SMOTE-NC 26
Tabel 4. 7 Normalisasi Data Uji..... 26
Tabel 4. 8 *Confusion Matrix* Hasil Klasifikasi Data Uji untuk K=5 27
Tabel 4. 9 Akurasi, Spesifisitas, dan Sensitivitas Parameter K-NN..... 28
Tabel 4.10 Perbandingain Sensitivitas Sebelum dan Sesudah SMOTE-NC..... 29

DAFTAR LAMPIRAN

Lampiran I Data Latih Hasil SMOTE-NC 35
Lampiran II Hasil Klasifikasi Data Uji 36

DAFTAR SIMBOL

σ	: Standar deviasi
N	: Banyaknya sampel pada data latih
x	: Nilai dari data
x_i	: Data latih ke-i, dengan $i = 1,2, \dots, p$
x_g	: Data latih SMOTE-NC ke- g, dengan $g = 1,2, \dots, h$
μ	: Nilai rata-rata
$x_{i,j}$	: Data latih ke-i atribut ke-j, dengan $i = 1,2, \dots, p$ dan $j = 1,2, \dots, n$
$x_{g,j}$	: Data latih SMOTE-NC ke- g, dengan $g = 1,2, \dots, h$ dan $j = 1,2, \dots, n$
$y_{i,j}$	: Data uji ke-i atribut ke-j, dengan $i = 1,2, \dots, s$ dan $j = 1,2, \dots, n$
n	: Banyaknya atribut independen
$D(x_i, x_i)$	: Jarak antar data latih, dengan $i = 1,2, \dots, w$
$D(x_g, y_i)$	: Jarak antara data latih SMOTE-NC ke-g dengan data uji ke-i, dengan $i = 1,2, \dots, v$
x_{mnr_i}	: Data minor ke-i, dengan $i = 1,2, \dots, t$
x_{knn}	: Tetangga terdekat
δ	: Bilangan acak antara 0 dan 1
x_{new_k}	: Data sintetis ke-k hasil dari SMOTE, dengan $k = 1,2, \dots, u$
x_{norm}	: Data hasil normalisasi
$\min(x)$	: Nilai minimum dari kumpulan data
$\max(x)$	: Nilai maksimum dari kumpulan data

BAB I
PENDAHULUAN

1.1 Latar Belakang

Data mining adalah proses menggunakan teknik statistik untuk mengidentifikasi pola dan pengetahuan menarik dari kumpulan data yang berukuran besar (Muslim, Prasetiyo, dan Mawarni, 2019). Salah satu fungsi yang paling sering digunakan dalam *data mining* adalah klasifikasi. Klasifikasi umumnya digunakan untuk membuat keputusan berdasarkan wawasan baru yang diperoleh dari pengolahan data sebelumnya melalui perhitungan algoritma tertentu (Indrayanti, Sugianti, dan Karomi, 2017). Salah satu algoritma yang sering diterapkan dalam proses klasifikasi pada *data mining* adalah Algoritma *K-Nearest Neighbor*. Algoritma K-NN adalah metode yang digunakan untuk mengklasifikasikan objek dengan merujuk pada kelas data latih yang memiliki jarak terdekat (Baharuddin, Azis, dan Hasanuddin, 2019). Algoritma K-NN dinyatakan efektif untuk data besar dan yang mengandung *noise*, serta menunjukkan kinerja yang baik dalam hasil klasifikasi (Putry dan Sari, 2022).

Hasil klasifikasi dapat dipengaruhi oleh ketidakseimbangan kelas, di mana satu kelas dalam data memiliki jumlah kasus yang jauh lebih banyak dibandingkan kelas lainnya. Data dengan ketidakseimbangan kelas cenderung mengklasifikasikan data baru berdasarkan kelas mayoritas dan mengabaikan kelas minoritas, yang mengakibatkan kinerja klasifikasi menjadi kurang optimal (Tanti, Sirait, dan Andri, 2021). Masalah tersebut dapat diatasi dengan menyeimbangkan data untuk meningkatkan akurasi, yakni dengan menghasilkan sampel sintetis pada kelas minoritas menggunakan teknik *oversampling Synthetic Minority Oversampling Technique* (SMOTE) sebelum proses klasifikasi dilakukan.

SMOTE adalah sebuah teknik yang umum digunakan untuk mengatasi ketidakseimbangan kelas dalam data (Siringoringo, 2018). Metode ini menciptakan sampel tambahan dari kelas yang minoritas dengan melakukan pengambilan sampel ulang, sehingga dapat membuat dataset menjadi seimbang. Sedangkan K-NN

adalah metode klasifikasi dengan mencari jarak terdekat antara data yang akan dievaluasi dengan kelas tetangga terdekatnya dalam data pelatihan (Kustiyahningsih & Syafa'ah, 2015). Penggunaan SMOTE dalam proses klasifikasi data dengan algoritma seperti K-NN dapat diterapkan di berbagai bidang, termasuk dalam sektor kesehatan. Saat ini, banyak catatan dalam bidang kesehatan mengenai hasil pemeriksaan pasien untuk berbagai penyakit yang dapat dimanfaatkan untuk mencari informasi penting, termasuk data mengenai penyakit gagal jantung.

Penyakit gagal jantung termasuk dalam kategori penyakit kardiovaskular. Gagal jantung adalah kondisi di mana jantung tidak mampu memompa cukup darah untuk memenuhi kebutuhan tubuh (Pratama et al., 2022). Penyakit jantung merupakan salah satu penyakit yang cukup berbahaya ketika menyerang seseorang. Penyebab utamanya sering kali berasal dari gaya hidup yang kurang sehat, seperti konsumsi makanan tinggi kolesterol, kebiasaan mengonsumsi alkohol, penggunaan tembakau, pola makan yang tidak teratur, serta faktor-faktor lain yang dapat memicu risiko penyakit tersebut (Zulhaq Jasman et al., 2022). Penyakit jantung sering dialami oleh pria, dengan perbandingan sekitar satu hingga tiga orang berisiko mengalami penyakit ini sebelum usia 60 tahun. Sementara itu, pada perempuan, perbandingannya sekitar satu dari sepuluh yang berpotensi menderita penyakit jantung. Skala yang cukup besar terkait dengan penyakit jantung membuatnya menjadi salah satu penyakit yang menghasilkan jumlah data pasien yang signifikan (Putra & Rini, 2019). Banyak orang baru menyadari bahwa mereka mengidap penyakit jantung ketika sudah terlambat, karena gejala awal sering kali tidak tampak jelas, sehingga penyakit ini baru terdeteksi setelah muncul komplikasi. Akibatnya, penyakit jantung tergolong sebagai Penyakit Tidak Menular (PTM) yang serius dan semakin menyebar ke berbagai negara.

Gagal jantung menyebabkan sekitar 17,9 juta kematian setiap tahun di seluruh dunia dan memiliki prevalensi yang lebih tinggi di Asia (Nugraha, 2021). Menurut *World Health Organization* (WHO), prevalensi penyakit gagal jantung di Amerika Serikat pada tahun 2013 mencapai sekitar 550.000 kasus per tahun, sementara *American Heart Association* (AHA) melaporkan bahwa sekitar 375.000 orang meninggal setiap tahun akibat penyakit gagal jantung di negara tersebut. Menurut

data dari Kementerian Kesehatan Republik Indonesia, penyakit sistem sirkulasi merupakan penyebab utama kematian di Indonesia. Data menunjukkan bahwa penyakit jantung koroner, yang termasuk dalam kategori penyakit sistem sirkulasi, menjadi penyebab utama kematian di Indonesia, dengan persentase sebesar 26,4% dari seluruh kematian, dan profil Kesehatan Indonesia pada tahun 2023 menyebutkan bahwa penyakit jantung merupakan penyebab kematian kedua tertinggi di Indonesia.

Ketidakstabilan kadar gula darah dapat berdampak serius pada organ tubuh, meningkatkan risiko penyakit kardiovaskular, merusak ginjal, serta menyebabkan berbagai komplikasi lainnya. Diagnosis dini dan penanganan yang efektif pada individu yang menderita penyakit jantung memiliki peran yang sangat penting (Artanti et al., 2024). Maka dari itu, penting untuk mengembangkan metode klasifikasi yang akurat dan efektif dalam proses diagnosis penyakit kardiovaskular (Mohi Uddin et al., 2023). Data pasien yang telah terkumpul sebelumnya dapat dimanfaatkan dengan mengolahnya menggunakan teknik klasifikasi dalam *data mining*, sehingga menghasilkan informasi yang dapat mendukung keputusan dalam mengidentifikasi penyakit gagal jantung. Jumlah data rekam medis yang cenderung besar sering kali menghadapi masalah ketidakseimbangan kelas, sehingga teknik oversampling SMOTE diperlukan untuk meningkatkan keakuratan algoritma klasifikasi (Marlisa et al., 2024).

Oleh karena itu, dilakukan penelitian untuk mengatasi masalah ketidakseimbangan kelas dalam kasus penentuan status gagal jantung dengan menerapkan teknik *oversampling* SMOTE-NC, sehingga dapat dilihat bagaimana kinerja dari algoritma klasifikasi K-NN. Penelitian yang dilakukan diharapkan dapat menghasilkan klasifikasi yang akurat, sehingga dapat berfungsi sebagai sistem pendukung keputusan dalam mengidentifikasi penyakit gagal jantung, khususnya dalam sektor kesehatan.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, rumusan masalah dalam penelitian ini adalah mengenai penerapan dan mengevaluasi teknik SMOTE dalam Algoritma K-NN pada studi kasus klasifikasi status penyakit gagal jantung.

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah yang telah dirumuskan, tujuan yang ingin dicapai dari penelitian ini adalah:

1. Menerapkan teknik SMOTE-NC dalam Algoritma K-NN untuk studi kasus klasifikasi status penyakit gagal jantung.
2. Mengevaluasi performa model berdasarkan sensitivitas sebagai ukuran keberhasilan dalam mendeteksi pasien yang meninggal akibat gagal jantung.

1.4 Batasan Masalah

Batasan masalah dalam penelitian ini yaitu sebagai berikut:

1. Data yang digunakan dalam penelitian ini merupakan data *heart failure clinical record dataset* yang diperoleh dari laman internet Kaggle dengan alamat website: confu
2. Atribut dependen yang digunakan yaitu status gagal ginjal, sedangkan atribut independennya yaitu usia, penurunan sel darah merah, kreatinin fosfokinase, diabetes, fraksi ejeksi, tekanan darah tinggi, trombosit, kreatinin serum, natrium serum, jenis kelamin, dan status perokok.
3. Nilai ketetanggaan yang digunakan pada Algoritma K-NN yaitu K = 3, 5, 7, 9,11,15.

1.5 Tinjauan Pustaka

Penggunaan SMOTE dalam konteks klasifikasi telah menghasilkan peningkatan kinerja dibandingkan dengan metode klasifikasi yang sudah ada sebelumnya. Beberapa penelitian yang membahas tentang pengklasifikasian oversampling SMOTE pada algoritma K-NN sudah dilakukan. Penelitian yang

dilakukan oleh Anis Nikmatul Hasanah, dkk., pada tahun 2017 menggunakan pendekatan oversampling SMOTE pada algoritma K-NN untuk mengatasi masalah *imbalance class* dalam klasifikasi objektivitas berita online. Dalam penelitian tersebut menerapkann variasi nilai *K* tetangga, yakni 1, 3, 5, 7, dan 9, menunjukkan bahwa penggunaan SMOTE dapat meningkatkan akurasi algoritma K-NN pada *K*=1 dan *K*=3 dengan peningkatan rata- rata sebesar 3,36. Namun, saat nilai *K* adalah 5, 7, dan 9, terdapat penurunan rata-rata akurasi sebesar 6,67 dengan nilai *K* adalah nilai tetangga terdekatnya.

Penelitian yang dilakukan oleh Rimbun Siringoringo pada tahun 2018 menerapkan *oversampling* SMOTE untuk menyelesaikan masalah ketidak seimbangan kelas pada dataset *Credit Card Fraud*. Penelitian tersebut menerapkan nilai *K* dalam rentang 1, 2, 3, 5, 7, dan 9, dimana didapatkan rata-rata performa *G-Mean* sebesar 81,0% untuk performa kombinasi dari SMOTE dan K-NN, sedangkan untuk performa K-NN saja adalah 53,4%. Sementara itu, rata-rata performa *F-Measure* untuk skema SMOTE+K-NN adalah 81,8%, sedangkan untuk skema K-NN saja adalah 38,7%. Dari hasil eksperimen tersebut, dapat disimpulkan bahwa penerapan SMOTE dan K-NN efektif dalam menangani ketidakseimbangan kelas pada *dataset Credit Card Fraud* dengan menghasilkan nilai *G-Mean* dan *F-Measure* yang lebih tinggi daripada K-NN saja. Hal ini menunjukkan bahwa metode SMOTE efektif dalam meningkatkan kinerja klasifikasi pada data yang tidak seimbang.

Penelitian yang dilakukan oleh Sultan Maula Chamzah pada tahun 2022 menerapkan *oversampling* SMOTE pada *data text* menggunakan K-NN. Penelitian tersebut menggunakan 5000 data *review* Tokopedia yang terdiri dari 3975 data negatif dan 1025 data positif. Dari 5000 data dibagi menjadi dua bagian, 70% data latih dan 30% data uji. Penerapan metode SMOTE pada K-NN menghasilkan tingkat akurasi yang lebih tinggi, mencapai 90%, dibandingkan dengan menggunakan K-NN saja yang memiliki tingkat akurasi sebesar 82%.

Penelitian menggunakan Algoritma K-NN dalam mengklasifikasikan penyakit kardiovaskular dilakukan oleh Vera, Faisal, dan Fachrul tahun 2024. Data

yang digunakan merupakan data rekam medis 299 pasien gagal jantung di Punjab, Pakistan. Atribut yang digunakan dalam penelitian tersebut yaitu sebanyak 11 atribut yakni usia, anemia, tekanan darah tinggi, kadar CPK dalam darah, diabetes, fraksi ejeksi, trombosit, kreatinin serum, natrium serum, status perokok, dan status diagnosa pasien dengan kelas pasien selamat sebanyak 203 dan pasien meninggal 96 orang. Berdasarkan hasil penelitian dilakukan split data yaitu dengan rasio 60:40, yang mana data latih 60% dan data uji 40%. Diperoleh hasil akurasi 91% serta nilai presisi 90%, *recall* 93%, dan F1-*score* 92%. Selain didapatkan hasil akurasi, presisi, *recall*, dan F1-*score* didapatkan juga hasil AUC sebesar 0,92 atau 92%. Hasil tersebut sangat baik karena nilai AUC mendekati nilai 1.

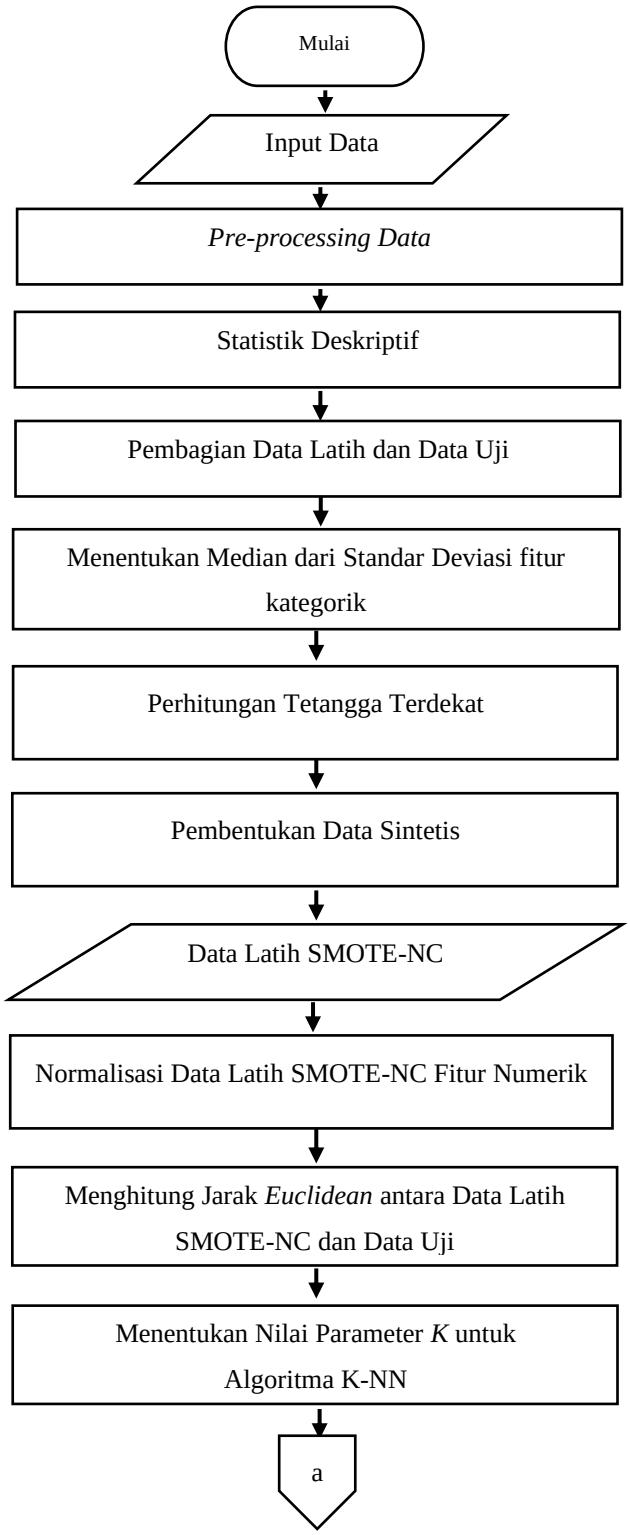
1.6 Metodologi Penelitian

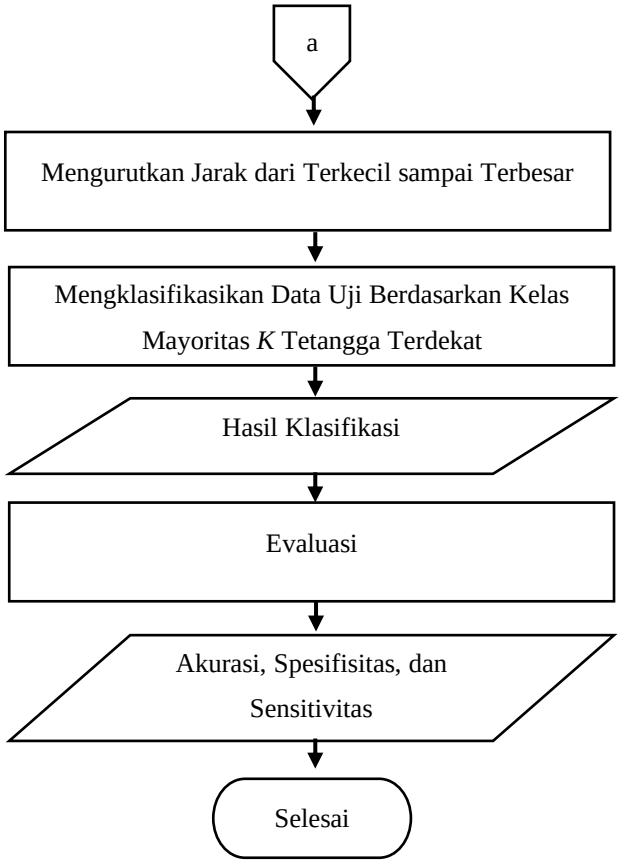
Langkah awal pada penelitian ini yaitu melakukan *pre-processing* data pada data *input* dengan memilih atribut berdasarkan faktor-faktor yang mempengaruhi penyakit gagal jantung dan menghapus data yang tidak memiliki nilai (*missing values*), sehingga terbentuk dataset baru yang hanya mengandung data yang diperlukan. Data yang telah disiapkan diproses dengan membuat statistik deskriptif untuk memahami gambaran umum data yang akan digunakan. Setelah itu, dataset dibagi menjadi data pelatihan dan data pengujian.

Tahap selanjutnya yaitu mengatasi permasalahan ketidakseimbangan pada data dengan menerapkan teknik SMOTE-NC pada data latih dengan menggunakan nilai ketetanggaan *K* sebanyak 3. Proses selanjutnya yaitu membuat sampel sintetis pada kelas minoritas yang kemudian dibangkitkan. Data sintetis yang dihasilkan ditambahkan ke data latih awal, yang kemudian menghasilkan data latih baru yang disebut data latih SMOTE-NC.

Langkah selanjutnya dilanjutkan dengan melakukan normalisasi pada data latih SMOTE-NC dan data uji kemudian menghitung jarak antara data latih SMOTE-NC dengan data uji, kemudian mengurutkan jarak yang diperolah dari jarak terkecil sampai terbesar, lalu menentukan nilai parameter *K* untuk algoritma K-NN dan melakukan klasifikasi pada data uji dengan mempertimbangkan kelas mayoritas dari *K* tetangga terdekat sebagai acuan. Setelah itu, dilakukan evaluasi

dan perbandingan terhadap kinerja pengklasifikasian, dan kemudian didapatkan hasil akurasi, spesifisitas dan sensitivitas dari klasifikasti tersebut.





Gambar 1. 1 *Flowchart* Implementasi SMOTE-NC pada Algoritma K-NN