

**KLASIFIKASI PENYAKIT KARDIOVASKULAR MENGGUNAKAN  
ALGORITMA *K-NEAREST NEIGHBORS* (KNN)**

**SKRIPSI**

**Oleh:  
VERA ARTANTI  
NIM. 200605110039**



**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM  
MALANG  
2024**

**KLASIFIKASI PENYAKIT KARDIOVASKULAR MENGGUNAKAN  
ALGORITMA *K-NEAREST NEIGHBORS* (KNN)**

**SKRIPSI**

Diajukan Kepada:  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Maulana Malik Ibrahim Malang  
Untuk memenuhi Salah Satu Persyaratan Dalam  
Memperoleh Gelar Sarjana Komputer (S.Kom)

Oleh:  
**VERA ARTANTI**  
**NIM. 200605110039**

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS SAINS DAN TEKNOLOGI  
UNIVERSITAS ISLAM NEGERI MAULANA MALIK IBRAHIM  
MALANG  
2024**

## HALAMAN PERSETUJUAN

### KLASIFIKASI PENYAKIT KARDIOVASKULAR MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBORS* (KNN)

#### SKRIPSI

Oleh:  
**VERA ARTANTI**  
NIM. 200605110039

Telah Diperiksa dan Disetujui untuk Diuji  
Tanggal: 30 April 2024

Pembimbing I



Dr. Muhammad Faisal, M.T  
NIP. 19740510 200501 1 007

Pembimbing II



Dr. Fachrul Kurniawan, M.MT, IPM  
NIP. 19771020 200912 1 001

Mengetahui,  
Ketua Program Studi Teknik Informatika  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Maulana Malik Ibrahim Malang



Dr. Fachrul Kurniawan, M.MT, IPM  
NIP. 19771020 200912 1 001

## HALAMAN PENGESAHAN

### KLASIFIKASI PENYAKIT KARDIOVASKULAR MENGGUNAKAN ALGORITMA *K-NEAREST NEIGHBORS* (KNN)

#### SKRIPSI

Oleh:  
**VERA ARTANTI**  
NIM. 200605110039

Telah Dipertahankan di Depan Dewan Penguji Skripsi  
dan Dinyatakan Diterima Sebagai Salah Satu Persyaratan  
Untuk Memperoleh Gelar Sarjana Komputer (S.Kom)  
Tanggal: 14 Mei 2024

#### Susunan Dewan Penguji

Ketua Penguji : Hani Nurhayati, M.T  
NIP. 19780625 200801 2 006

Anggota Penguji I : Johan Ericka Wahyu P, M.Kom  
NIP. 19831213 201903 1 004

Anggota Penguji II : Dr. Muhammad Faisal, M.T  
NIP. 19740510 200501 1 007

Anggota Penguji III : Dr. Fachrul Kurniawan, M.MT, IPM  
NIP. 19771020 200912 1 001

(  )  
(  )  
(  )  
(  )

Mengetahui dan Mengesahkan,  
Ketua Program Studi Teknik Informatika  
Fakultas Sains dan Teknologi  
Universitas Islam Negeri Maulana Malik Ibrahim Malang



  
Fachrul Kurniawan, M.MT, IPM  
NIP. 19771020 200912 1 001

## PERNYATAAN KEASLIAN TULISAN

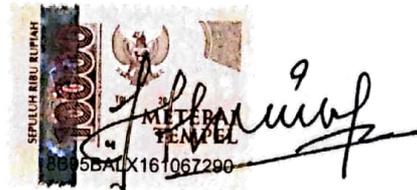
Saya yang bertanda tangan di bawah ini:

Nama : Vera Artanti  
NIM : 200605110039  
Fakultas / Program Studi : Sains dan Teknologi / Teknik Informatika  
Judul Skripsi : Klasifikasi Penyakit Kardiovaskular Menggunakan Algoritma *K-Nearest Neighbors* (KNN)

Menyatakan dengan sebenarnya bahwa Skripsi yang saya tulis ini benar-benar merupakan hasil karya saya sendiri, bukan merupakan pengambil alihan data, tulisan, atau pikiran orang lain yang saya akui sebagai hasil tulisan atau pikiran saya sendiri, kecuali dengan mencantumkan sumber cuplikan pada daftar pustaka.

Apabila dikemudian hari terbukti atau dapat dibuktikan skripsi ini merupakan hasil jiplakan, maka saya bersedia menerima sanksi atas perbuatan tersebut.

Malang, 30 April 2024  
Yang membuat pernyataan,



Vera Artanti  
NIM. 200605110039

## HALAMAN MOTTO

“Orang lain tidak akan paham *struggle* dan masa sulit kita, yang mereka ingin tau hanya *success stories*. Berjuanglah untuk diri sendiri, tunjukkan bahwa kita bisa melewati semua rintangan. Setiap kali ingin menyerah, ingatlah sudah sejauh apa kita melangkah. Kelak di masa depan kita akan bangga dengan apa yang kita perjuangkan hari ini.”

لَا يُكَلِّفُ اللَّهُ نَفْسًا إِلَّا وُسْعَهَا

“Allah SWT tidak membebani seseorang melainkan sesuai dengan kesanggupannya”

(Q.S Al-Baqarah:286)

## HALAMAN PERSEMBAHAN

Alhamdulillah Wasyukurillah segala Puji Syukur atas kehadiran Allah SWT atas limpahan Rahmat, Taufiq, hidayah, dan inayah-Nya kepada penulis beserta keluarga dan saudara lainnya sehingga penulis dapat menyelesaikan tugas akhir ini. Dengan rasa hormat dan terimakasih, penulis mempersembahkan skripsi tugas akhir ini kepada:

1. Pintu surgaku yaitu Ibu tersayang Ngatmi. Beribu-ribu terimakasih penulis ucapkan atas limpahan kasih sayang yang tak terhingga yang telah mengasuh, merawat, mendidik dan membesarkan penulis, serta iringan doa yang tulus tiada hentinya untuk keberhasilan penulis.
2. Penguat langkahku yaitu Ayah Danial Salam. Terimakasih penulis ucapkan atas kehadiran ayah dalam kehidupan penulis. Kehadiran ayah membuat penulis mendapatkan kembali kasih sayang seorang ayah, terimakasih ayah atas semangat bekerja yang tak pernah kenal lelah demi pendidikan penulis.
3. Adikku tersayang Saktika Ihza Maulana, terimakasih atas semangat dan selalu menghibur penulis.
4. (.almh) Nenek tersayang Ginah. Terimakasih telah merawat penulis dari kecil, nenek yang memberikan semangat agar penulis bisa berpendidikan setinggi mungkin, sekarang penulis telah berhasil menyelesaikan S1.
5. Bapak Dr. Muhammad Faisal, M.T dan Bapak Dr. Fachrul Kurniawan, M.MT selaku Dosen Pembimbing Penulis yang yang sabar serta tulus dalam memberikan bimbingan dan arahan kepada penulis dalam penyelesaian tugas akhir.
6. Diri saya sendiri. Vera Artanti. Apresiasi sebesar-besarnya karena telah menyelesaikan apa yang telah dimulai. Terimakasih karena terus berusaha dan tidak menyerah, serta menikmati setiap proses. Terimakasih sudah bertahan, berbahagialah dimanapun berada, jangan dengarkan orang lain yang memandang sebelah mata, karena mereka hanya ingin melihat kesuksesan kita bukan proses yang kita lalui.

## KATA PENGANTAR

*Assalamualaikum Wr. Wb*

Puji dan syukur penulis panjatkan kehadiran Tuhan Yang Maha Esa, Allah subhanahu wa ta'ala yang telah memberikan Taufik dan Hidayah-Nya kepada penulis sehingga dapat menyelesaikan skripsi yang berjudul “Klasifikasi Penyakit Kardiovaskular Menggunakan Algoritma *K-Nearest Neighbors* (KNN)” dengan baik.

Banyak pihak yang terlibat dalam penulisan skripsi ini yang telah memberikan dukungan baik moril maupun materil. Untuk itu dalam kesempatan kali ini penulis ingin mengucapkan banyak terimakasih kepada:

1. Prof. Dr. H. M. Zainuddin, MA, selaku Rektor Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
2. Prof. Dr. Sri Harini, M.Si, selaku Dekan Fakultas Sains dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang beserta jajarannya.
3. Dr. Fachrul Kurniawan ST., M.MT, IPM selaku Ketua Program Studi Teknik Informatika Universitas Islam Negeri Maulana Malik Ibrahim Malang serta pembimbing II.
4. Dr. Muhammad Faisal, M.T selaku Dosen Pembimbing I yang telah dengan sabar memberikan arahan baik dalam penulisan hingga program yang dibuat dalam menyelesaikan skripsi ini.
5. Dr. Ir. M. Amin Hariyadi, M.T selaku Dosen Wali yang telah memberikan arahan dalam proses perkuliahan.
6. Segenap Dosen, Laboran dan jajaran pada Program Studi Teknik Informatika yang telah memberikan bimbingan dan bantuan selama studi.
7. Nia Faricha, S.Si selaku Admin Program Studi Teknik Informatika yang dengan sabar membantu, memberikan arahan, informasi terkait perkuliahan.
8. Kedua orangtua penulis Ibu Ngatmi dan Ayah Danial Salam yang selalu memberi dukungan dan perhatian serta selalu memberikan yang terbaik untuk kelancaran putrinya dalam pendidikan.

9. Adik penulis Saktika Ihza Maulana yang selalu memberi semangat dan menghibur.
10. Sahabat penulis “*enter new subject*” yang beranggotakan Zidan, Rizka, Bima terimakasih atas segala bantuan dan semangat yang diberikan serta semua kenangan dari awal perkuliahan yang masih *online*, bertemu hanya lewat *zoom meeting* sampai akhirnya bisa bertemu secara langsung dan saling membantu dalam perkuliahan, sampai terselesaikannya skripsi ini. *See u on top guys!*
11. Teman kos alex Icha, Sasa, Mimin, Lala, Kakwid penghuni kos alex terimakasih sudah membantu, mendukung serta menghibur penulis di masa-masa hilang harapan. selalu sigap disaat penulis membutuhkan pertolongan, dan banyak hal lain. *Sending u a warm hug! And See u on top guys!*
12. Seluruh keluarga besar Saudara Teknik Informatika UIN Malang terkhusus Angkatan 2020 “Integer”. serta teman-teman yang dekat dengan penulis tidak bisa disebutkan satu persatu, terimakasih telah memberikan *support*, motivasi dan bantuannya kepada penulis.
13. Seluruh pihak yang telah terlibat secara langsung maupun tidak langsung dalam proses penyusunan skripsi sejauh ini.
14. *Last but not least*, untuk “Vera Artanti” terimakasih sudah semangat, tidak pernah berhenti mencoba. Tetaplah berjuang walaupun sering jatuh, jangan pernah puas pencapaian saat ini terus kejar cita-cita dan jangan pernah sombong atas apa yang telah Allah titipkan di dunia ini. Kamu keren dan hebat Vera! *You’re trying your best and that’s great, just keep going!*

Penulis menyadari dalam penulisan skripsi ini tidak luput dari kesalahan yang jauh dari kata sempurna. Oleh karena itu, penulis mengharapkan kritik dan saran yang membangun sehingga skripsi ini dapat lebih dikembangkan.

Malang, 30 April 2024

Penulis

## DAFTAR ISI

<b>HALAMAN PENGAJUAN</b> .....	<b>ii</b>
<b>HALAMAN PERSETUJUAN</b> .....	<b>iii</b>
<b>HALAMAN PENGESAHAN</b> .....	<b>iv</b>
<b>PERNYATAAN KEASLIAN TULISAN</b> .....	<b>v</b>
<b>HALAMAN MOTTO</b> .....	<b>vi</b>
<b>HALAMAN PERSEMBAHAN</b> .....	<b>vii</b>
<b>KATA PENGANTAR</b> .....	<b>viii</b>
<b>DAFTAR ISI</b> .....	<b>x</b>
<b>DAFTAR GAMBAR</b> .....	<b>xii</b>
<b>DAFTAR TABEL</b> .....	<b>xiii</b>
<b>ABSTRAK</b> .....	<b>xiv</b>
<b>ABSTRACT</b> .....	<b>xv</b>
<b>مستخلص البحث</b> .....	<b>xvi</b>
<b>BAB I PENDAHULUAN</b> .....	<b>1</b>
1.1 Latar Belakang .....	1
1.2 Rumusan Masalah .....	6
1.3 Batasan Masalah .....	6
1.4 Tujuan Penelitian .....	6
1.5 Manfaat Penelitian .....	6
<b>BAB II STUDI PUSTAKA</b> .....	<b>7</b>
2.1 Penelitian Terkait .....	7
2.2 Penyakit Kardiovaskular .....	11
2.3 Klasifikasi .....	13
2.4 <i>K-Nearest Neighbors</i> (KNN) .....	14
2.5 <i>GridSearchCV</i> .....	16
2.6 <i>Synthetic Minority Oversampling Technique</i> (SMOTE) .....	17
2.7 <i>Confusion Matrix</i> .....	18
2.8 <i>Receiver Operating Characteristic</i> (ROC) .....	20
<b>BAB III METODOLOGI PENELITIAN</b> .....	<b>22</b>
3.1 Desain Sistem .....	22
3.2 Pengumpulan Data .....	23
3.3 <i>Preprocessing</i> .....	25
3.3.1 Seleksi Fitur .....	25
3.3.2 <i>Synthetic Minority Oversampling Technique</i> (SMOTE) .....	26
3.3.3 Normalisasi Data .....	27
3.4 <i>Split Data</i> .....	29
3.5 Implementasi <i>K-Nearest Neighbor</i> (KNN) .....	29
3.5.1 Contoh Implementasi <i>K-Nearest Neighbor</i> (KNN) .....	31
3.6 Proses <i>Hyperparameter</i> .....	33
3.7 Skenario Uji Coba .....	34
<b>BAB IV UJI COBA DAN PEMBAHASAN</b> .....	<b>36</b>
4.1 Hasil Uji Coba .....	36
4.1.1 Pengujian Model A .....	36

4.1.1.1 Tidak Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	36
4.1.1.2 Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	39
4.1.2 Pengujian Model B.....	41
4.1.2.1 Tidak Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	41
4.1.2.2 Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	43
4.1.3 Pengujian Model C.....	45
4.1.3.1 Tidak Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	46
4.1.3.2 Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	48
4.2 Pembahasan.....	50
4.3 Integrasi Islam.....	58
4.3.1 Muamalah Mu'Allah.....	58
4.3.2 Muamalah Mu'Annas.....	59
<b>BAB V KESIMPULAN DAN SARAN.....</b>	<b>62</b>
5.1 Kesimpulan.....	62
5.2 Saran.....	63
<b>DAFTAR PUSTAKA</b>	
<b>LAMPIRAN</b>	

## DAFTAR GAMBAR

Gambar 2. 1 <i>Receiver Operator Characteristic</i> .....	21
Gambar 2. 2 <i>Area Under Curve</i> .....	21
Gambar 3. 1 Desain Sistem .....	22
Gambar 3. 2 Hasil Perhitungan <i>Fitur Importance</i> .....	26
Gambar 3. 3 <i>Flowchart K-Nearest Neighbor</i> .....	29
Gambar 4. 1 Curva ROC Model A tanpa seleksi fitur dan <i>SMOTE</i> .....	38
Gambar 4. 2 Curva ROC Model A menggunakan seleksi fitur dan <i>SMOTE</i> .....	40
Gambar 4. 3 Curva ROC Model B tanpa seleksi fitur dan <i>SMOTE</i> .....	43
Gambar 4. 4 Curva ROC Model B menggunakan seleksi fitur dan <i>SMOTE</i> .....	45
Gambar 4. 5 Curva ROC Model C tanpa seleksi fitur dan <i>SMOTE</i> .....	47
Gambar 4. 6 Curva ROC Model C menggunakan seleksi fitur dan <i>SMOTE</i> .....	49
Gambar 4. 7 <i>Imbalance Data</i> .....	50
Gambar 4. 8 <i>Balance Data</i> .....	51
Gambar 4. 9 Tingkat Kepentingan ( <i>Importance</i> ) Fitur .....	52
Gambar 4. 10 Hasil Pengujian Tanpa Seleksi Fitur dan <i>SMOTE</i> .....	53
Gambar 4. 11 Hasil Pengujian Menggunakan Seleksi Fitur dan <i>SMOTE</i> .....	54
Gambar 4. 12 Hasil Perbandingan Dua Pengujian .....	55

## DAFTAR TABEL

Tabel 2. 1 Perbandingan Penelitian.....	9
Tabel 2. 2 <i>Confusion matrix</i> .....	18
Tabel 3. 1 Deskripsi Data.....	23
Tabel 3. 2 Contoh sampel <i>dataset</i> .....	24
Tabel 3. 3 Contoh <i>dataset</i> sebelum normalisasi.....	28
Tabel 3. 4 Contoh <i>dataset</i> setelah normalisasi.....	28
Tabel 3. 5 Perhitungan <i>Euclidean Distance</i> .....	31
Tabel 3. 6 Hasil perhitungan <i>Euclidean Distance</i> .....	33
Tabel 3. 7 Parameter .....	34
Tabel 3. 8 Skenario pembagian data .....	34
Tabel 4. 1 Hasil rata-rata skor pengujian tanpa seleksi fitur dan <i>SMOTE</i> .....	37
Tabel 4. 2 <i>Confusion matrix</i> tanpa seleksi fitur dan <i>SMOTE</i> .....	37
Tabel 4. 3 Nilai <i>accuracy, precision, recall, f1-score</i> .....	38
Tabel 4. 4 Hasil rata-rata skor pengujian dengan seleksi fitur dan <i>SMOTE</i> .....	39
Tabel 4. 5 <i>Confusion matrix</i> menggunakan seleksi fitur dan <i>SMOTE</i> .....	39
Tabel 4. 6 Nilai <i>accuracy, precision, recall, f1-score</i> .....	40
Tabel 4. 7 Hasil rata-rata skor pengujian tanpa seleksi fitur dan <i>SMOTE</i> .....	41
Tabel 4. 8 <i>Confusion matrix</i> tanpa seleksi fitur dan <i>SMOTE</i> .....	42
Tabel 4. 9 Nilai <i>accuracy, precision, recall, f1-score</i> .....	42
Tabel 4. 10 Hasil rata-rata skor pengujian dengan seleksi fitur dan <i>SMOTE</i> .....	44
Tabel 4. 11 <i>Confusion matrix</i> dengan seleksi fitur dan <i>SMOTE</i> .....	44
Tabel 4. 12 Nilai <i>accuracy, precision, recall, f1-score</i> .....	45
Tabel 4. 13 Hasil rata-rata skor pengujian tanpa seleksi fitur dan <i>SMOTE</i> .....	46
Tabel 4. 14 <i>Confusion matrix</i> tanpa seleksi fitur dan <i>SMOTE</i> .....	46
Tabel 4. 15 Nilai <i>accuracy, precision, recall, f1-score</i> .....	47
Tabel 4. 16 Hasil rata-rata skor pengujian dengan seleksi fitur dan <i>SMOTE</i> .....	48
Tabel 4. 17 <i>Confusion matrix</i> dengan seleksi fitur dan <i>SMOTE</i> .....	48
Tabel 4. 18 Nilai <i>accuracy, precision, recall, f1-score</i> .....	49
Tabel 4. 19 Fitur yang digunakan .....	52
Tabel 4. 20 Hasil Pengujian tanpa seleksi fitur dan <i>SMOTE</i> .....	53
Tabel 4. 21 Hasil Pengujian menggunakan seleksi fitur dan <i>SMOTE</i> .....	54
Tabel 4. 22 Hasil <i>Precision, Recall, F1-score</i> setelah <i>Preprocessing</i> .....	55
Tabel 4. 23 Hasil Perbandingan Akurasi.....	56

## ABSTRAK

Artanti, Vera. 2024. *Klasifikasi Penyakit Kardiovaskular Menggunakan Algoritma K-Nearest Neighbors (KNN)*. Skripsi. Program Studi Teknik Informatika, Fakultas Sains dan Teknologi. Universitas Islam Negeri Maulana Malik Ibrahim Malang. Pembimbing: (I) Dr. Muhammad Faisal, M.T. (II) Dr. Fachrul Kurniawan, M.MT.

**Kata Kunci:** Klasifikasi, Penyakit Kardiovaskular, *K-Nearest Neighbors*

Penyakit kardiovaskular merupakan salah satu tantangan serius dalam bidang kesehatan global. Menurut data penyakit kardiovaskular menduduki peringkat pertama sebagai penyebab kematian di seluruh dunia. Pada tahun 2019 diperkirakan sekitar 17,9 juta kematian akibat penyakit kardiovaskular, atau sekitar 32% dari total kematian global. Salah satu kejadian umum dari penyakit ini adalah gagal jantung, yang memiliki dampak serius pada kualitas hidup dan harapan hidup individu. Dalam penelitian ini digunakan metode *K-Nearest Neighbors* (KNN) untuk mengklasifikasikan penyakit kardiovaskular. Tujuan penelitian ini adalah mengetahui kinerja *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular. Fitur yang digunakan antara lain *time*, *serum creatine*, *ejection fraction*, *age*, *serum sodium*, *creatinine phosphokinase*, *platelets*. Data yang digunakan pada penelitian ini adalah data “*Heart Failure Prediction*” yang diambil dari *BMC Medical Informatics and Decision Making* berdasarkan rekam medis 299 pasien gagal jantung yang kemudian dilakukan *Preprocessing* data yang meliputi seleksi fitur, *SMOTE*, normalisasi data, serta split data. Pada penelitian ini *split* data terbagi menjadi 3 model, yaitu model A dengan perbandingan 80% data pelatihan: 20% data pengujian, model B dengan perbandingan 70% data pelatihan: 30% data pengujian, model C dengan perbandingan 60% data pelatihan: 40% data pengujian. Selanjutnya pada masing-masing pembagian model akan dilakukan perbandingan antara pengujian tidak menggunakan seleksi fitur dan tidak dilakukan proses *SMOTE* dengan pengujian menggunakan seleksi fitur dan dilakukan proses *SMOTE*. Nilai akurasi terbaik pada model C rasio data 60% data pelatihan dan 40% data pengujian yang telah dilakukan seleksi fitur dan proses *SMOTE* dengan nilai akurasi 90% dan nilai *AUC* 94% serta didapatkan nilai presisi 89%, *recall* 90%, dan *f1-score* 90%.

## ABSTRACT

Artanti, Vera. 2024. *Classification of Cardiovascular Diseases Using the K-Nearest Neighbors (KNN) Algorithm*. Department of Informatics Engineering Faculty of Science and Technology Islamic State University of Maulana Malik Ibrahim Malang. Supervisors: (I) Dr. Muhammad Faisal, M.T. (II) Dr. Fachrul Kurniawan, M.MT.

**Keywords:** Classification, Cardiovascular Disease, K-Nearest Neighbors

Cardiovascular disease is one of the most serious challenges in global health. According to data, cardiovascular disease is the number one cause of death worldwide. In 2019, an estimated 17.9 million deaths from cardiovascular disease are expected, or about 32% of total global deaths. One of the common occurrences of these diseases is heart failure, which has a serious impact on the quality of life and life expectancy of individuals. In this study, the K-Nearest Neighbors (KNN) method was used to classify cardiovascular diseases. The aim of this study is to evaluate the effectiveness of K-Nearest Neighbors (KNN) well through the calculation of accuracy, recall, precision, and f1-score in the classification of cardiovascular diseases. The features used include time, serum creatine, ejection fraction, age, serum sodium, creatine phosphokinase, platelets. The data used in this study is the "Heart Failure Prediction" data taken from BMC Medical Informatics and Decision Making based on the medical records of 299 heart failure patients which are then subjected to data Preprocessing which includes feature selection, SMOTE, data normalization, and data split. In this study, the split data is divided into 3 models, namely model A with a ratio of 80% training data: 20% testing data, model B with a ratio of 70% training data: 30% testing data, model C with a ratio of 60% training data: 40% testing data. Furthermore, in each division of the model, a comparison will be made between testing not using feature selection and not doing the SMOTE process with testing using feature selection and doing the SMOTE process. The best accuracy value is in model C with a data ratio of 60% training data and 40% test data that has been carried out feature selection and SMOTE process with an accuracy value of 90% and an AUC value of 94% and obtained a precision value of 89%, recall 90%, and f1-score 90%.

## مستخلص البحث

أرنتي، فير. ٢٠٢٤. تصنيف الأمراض القلبية الوعائية باستخدام *K-Nearest Neighbours (KNN)*. البحث الجامعي. قسم هندسة المعلوماتية، كلية العلوم والتكنولوجيا، الجامعة الإسلامية الحكومية مولانا مالك إبراهيم مالانج. المشرف: (١) الدكتور محمد فيصل الماجستير (٢) الدكتور فاهرول كرنويان الماجستير

الكلمات الرئيسية: التصنيف، *K-Nearest Neighbours*، الأمراض القلبية الوعائية

تعد أمراض القلب والأوعية الدموية أحد أخطر التحديات في مجال الصحة العالمية. ووفقاً للبيانات، فإن أمراض القلب والأوعية الدموية هي السبب الأول للوفاة في جميع أنحاء العالم. في عام 2019، من المتوقع حدوث ما يقدر بنحو 17.9 مليون حالة وفاة بسبب أمراض القلب والأوعية الدموية، أو حوالي 32% من إجمالي الوفيات العالمية. أحد الأمراض الشائعة لهذه الأمراض هو قصور القلب، والذي له تأثير خطير على جودة الحياة ومتوسط العمر المتوقع للأفراد. في هذه الدراسة، تم استخدام طريقة *K-Nearest Neighbours (KNN)* لتصنيف أمراض القلب والأوعية الدموية. والغرض من هذه الدراسة هو تحديد أداء طريقة *K-Nearest Neighbours (KNN)* بشكل جيد من خلال حساب الدقة والاستدعاء والدقة ودرجة  $f1$  في تصنيف أمراض القلب والأوعية الدموية. تشمل السمات المستخدمة الوقت والكرياتين في المصل والجزء المقذوف والعمر والصوديوم في المصل وفوسفوكيناز الكرياتين والصفائح الدموية. والبيانات المستخدمة في هذه الدراسة هي بيانات "*Heart Failure Prediction*" المأخوذة من *BMC* للمعلوماتية الطبية واتخاذ القرار بناءً على السجلات الطبية لـ 299 مريضاً بقصور القلب، ثم أجريت المعالجة المسبقة للبيانات التي تتضمن اختيار السمات و *SMOTE* وتطبيع البيانات وتقسيم البيانات. في هذه الدراسة، تم تقسيم البيانات المقسمة إلى 3 نماذج، وهي النموذج A بنسبة 80% بيانات تدريب: 20% بيانات اختبار، والنموذج B بنسبة 70% بيانات تدريب: 30% بيانات اختبار، والنموذج C بنسبة 60% بيانات تدريب: 40% بيانات اختبار. علاوة على ذلك، في كل قسم من أقسام النماذج، سيتم إجراء مقارنة بين الاختبار الذي لا يستخدم اختيار الملامح وعدم إجراء عملية *SMOTE* مع الاختبار الذي يستخدم اختبار الملامح وإجراء عملية *SMOTE*. أفضل قيمة دقة هي في النموذج (ج) بنسبة بيانات 60% بيانات تدريب و 40% بيانات اختبار والتي تم إجراء عملية اختيار الملامح وعملية *SMOTE* بقيمة دقة 90% وقيمة *AUC* 94% وحصل على قيمة دقة 89% واسترجاع 90% ودرجة  $f1$  90%.

# **BAB I**

## **PENDAHULUAN**

### **1.1 Latar Belakang**

Penyakit kardiovaskular yang juga dikenal sebagai penyakit gagal jantung dan pembuluh darah adalah tantangan serius dalam bidang kesehatan global. Menurut data penyakit ini menduduki peringkat pertama sebagai penyakit kematian di dunia. Tahun 2019 diperkirakan 17,9 juta kematian akibat penyakit kardiovaskular atau sekitar 32% dari total kematian global (*World Health Organization, 2021*). Salah satu kejadian umum dari penyakit ini adalah gagal jantung, yang memiliki dampak serius pada kualitas hidup dan harapan hidup individu. Lebih dari tiga per empat atau sekitar 80% kematian akibat penyakit kardiovaskular terjadi di negara-negara berpenghasilan rendah hingga menengah, termasuk Indonesia (*WHO, 2016; American Heart Association & American Stroke Association, 2015*).

Tahun 2014, Kementerian Kesehatan Republik Indonesia (Kemenkes RI) menggambarkan penyakit kardiovaskular sebagai kondisi medis yang melibatkan gangguan pada jantung dan sistem pembuluh darah, termasuk penyakit jantung koroner, gagal jantung, hipertensi, dan *stroke*. Data dari tahun 2013 menunjukkan bahwa prevalensi penyakit kardiovaskular di Indonesia adalah sekitar 1,5% untuk penyakit jantung koroner, 12,1% untuk *stroke*, 9,5% untuk hipertensi, dan 0,3% untuk gagal jantung. (Kemenkes RI, 2014).

Meskipun penyakit kardiovaskular dapat bersifat kompleks dan beragam, banyak faktor risiko penyakit kardiovaskular diantaranya semakin bertambahnya usia seseorang, maka semakin tinggi risiko terkena penyakit kardiovaskular. Faktor lain yaitu memiliki riwayat penyakit kardiovaskular dalam keluarga, diabetes, tekanan darah tinggi, obesitas (kegemukan), Pola hidup tidak sehat, dan Stres. (Kemenkes RI, 2021). Faktor-faktor lain seperti penggunaan tembakau, pola makan kurang sehat, kurangnya aktivitas fisik, serta konsumsi alkohol berlebihan telah terbukti berkontribusi pada risiko terjadinya penyakit ini (Putri, 2021).

Strategi Kementerian Kesehatan (Kemenkes) untuk mencegah penyakit kardiovaskular atau penyakit jantung meliputi beberapa langkah. Pertama, menyebarkan informasi dan edukasi sebanyak mungkin kepada masyarakat melalui berbagai media. Kedua, melakukan deteksi dini dengan mengimbau masyarakat untuk mengukur tekanan darah, gula darah, indeks massa tubuh, atau lingkar perut. Ketiga, memberikan perlindungan khusus, seperti menerapkan kawasan tanpa rokok. Terakhir, pengobatan yang dilakukan ketika seseorang sudah terdiagnosis penyakit jantung. (Kemenkes RI, 2019).

Oleh karena itu, tindakan pencegahan yang efektif untuk mengatasi faktor risiko ini menjadi penting dalam mengurangi angka kejadian penyakit kardiovaskular. Salah satu metode yang efektif untuk mengatasi penyakit kardiovaskular adalah dengan mengembangkan strategi pencegahan yang mencakup seluruh populasi. Hal ini mencakup upaya untuk mengedukasi masyarakat tentang dampak gaya hidup yang sehat, mempromosikan perubahan perilaku positif, dan meningkatkan kesadaran akan pentingnya pencegahan

(Nugraha, 2021). Namun, untuk melaksanakan strategi ini dengan efektif, diperlukan pemahaman yang lebih mendalam tentang faktor-faktor risiko dan cara mengidentifikasi individu yang berisiko tinggi.

Adapun dalam aspek penelitian ini terintegrasi keilmuan dalam ayat Al-Qur'an dalam surat Al-Hujurat ayat 13.

يَا أَيُّهَا النَّاسُ إِنَّا خَلَقْنَاكُمْ مِنْ ذَكَرٍ وَأُنْثَىٰ وَجَعَلْنَاكُمْ شُعُوبًا وَقَبَائِلَ لِتَعَارَفُوا ۗ إِنَّ أَكْرَمَكُمْ عِنْدَ اللَّهِ أَتْقَىٰكُمْ ۗ إِنَّ اللَّهَ عَلِيمٌ  
حَبِيرٌ (١٣)

*“Wahai manusia, sesungguhnya Kami telah menciptakan kamu dari seorang laki-laki dan perempuan. Kemudian, Kami menjadikan kamu berbangsa-bangsa dan bersuku-suku agar kamu saling mengenal. Sesungguhnya yang paling mulia di antara kamu di sisi Allah adalah orang yang paling bertakwa. Sesungguhnya Allah Maha Mengetahui lagi Mahateliti.” (QS. Al-Hujurat:13)*

Tafsir Ibnu Katsir mengajarkan pentingnya menghormati perbedaan agar kita dapat saling mengenal. Perbedaan yang paling mendasar adalah antara laki-laki dan perempuan, yang kemudian berkembang menjadi berbagai bangsa. Semua manusia adalah ciptaan Allah, dan yang paling mulia di hadapan-Nya adalah mereka yang bertakwa. Manusia terbaik menurut konsep ini adalah yang paling baik dalam membaca Al-Qur'an, paling bertakwa, aktif dalam menegakkan kebaikan dan mencegah kemungkaran, serta rajin menjalin silaturahmi. Ibnu Katsir menekankan bahwa perbedaan ras, suku, dan bangsa bukan untuk menjadi dasar perpecahan, melainkan untuk saling mengenal dan saling melengkapi. Kemuliaan di sisi Allah SWT menurut Ibnu Katsir, tidak ditentukan oleh ras, suku, atau bangsa, melainkan oleh ketakwaan dan ketaatan kepada Allah SWT (Shofwan & Munib, 2023).

Surat Al-Hujurat ayat 13 menekankan pentingnya persaudaraan dan toleransi antara berbagai bangsa dan suku. Ini bisa dikaitkan dengan klasifikasi dalam konteks saling mengenal dan memahami satu sama lain. Al-Qur'an mengajarkan bahwa perbedaan-perbedaan ini bukan dimaksudkan untuk menimbulkan permusuhan, tetapi untuk memungkinkan manusia saling berbagi pengetahuan dan pemahaman. Klasifikasi antar-bangsa dan suku tidak dianggap sebagai dasar untuk merendahkan atau mendiskriminasi satu sama lain tetapi sebagai titik awal untuk membangun hubungan yang baik dan saling menghormati dengan takwa kepada Allah sebagai standar keutamaan. Dalam konteks kesehatan memahami dan mengklasifikasikan penyakit kardiovaskular dengan benar merupakan langkah awal yang penting dalam penanganan dan pencegahan penyakit tersebut.

Penggunaan teknologi dan analisis data menjadi semakin relevan. Klasifikasi penyakit kardiovaskular berdasarkan faktor-faktor risiko tertentu dapat membantu dalam mengidentifikasi kelompok individu yang lebih rentan terhadap penyakit ini. Salah satu pendekatan yang dapat digunakan adalah algoritma *K-Nearest Neighbors* (KNN). Metode *K-Nearest Neighbors* (KNN) merupakan teknik klasifikasi yang menggunakan pola-pola data dalam dataset untuk menentukan kategori atau kelas dari sampel yang belum diketahui (Dhany, 2021). Pengklasifikasi *K-Nearest Neighbor* (KNN) telah membuktikan fleksibilitas yang tinggi dalam permasalahan klasifikasi yang kompleks, memungkinkan penanganan data yang bervariasi dan *heterogen* dengan hasil yang akurat. Hal ini menjadikannya pilihan yang populer dalam bidang medis untuk membantu diagnosa dan pengambilan keputusan klinis (Shaban *et al.*, 2020).

Penelitian dilakukan oleh (Samosir *et al.*, 2021) berhasil menerapkan metode *K-Nearest Neighbors* (KNN) untuk mengklasifikasikan pasien dengan penyakit jantung, mencapai akurasi sebesar 0,686 Area Under the Curve (AUC), 0,645 *Classification Accuracy* (CA), 0,641 *F1 Score*, 0,642 *Precision*, dan 0,645 *Recall*. Dataset yang digunakan berasal dari *Cleveland Clinic Foundation*, yang telah diadopsi oleh *Hungarian Institute of Cardiology* di *Budapest*.

Penelitian sekarang akan mengimplementasikan metode *K-Nearest Neighbors* (KNN) pada klasifikasi penyakit kardiovaskular. Peneliti akan melakukan skenario pengujian yaitu dengan tidak menggunakan seleksi fitur dan tidak melakukan *Synthetic Minority Oversampling Technique* (SMOTE) serta pengujian yang menggunakan seleksi fitur dan melakukan *Synthetic Minority Oversampling Technique* (SMOTE). Skenario pengujian tersebut akan dibandingkan hasilnya mana yang akurasinya lebih baik. Pengujian juga akan dilakukan dengan menggunakan tiga model pembagian *split* data, hal itu juga akan mempengaruhi hasil dari pengujian. Setelah dilakukan semua pengujian akan di bandingkan hasil yang lebih baik dan lebih akurat. Melalui penggunaan data rekam medis dan informasi tentang faktor risiko kardiovaskular, algoritma KNN diharapkan dapat membantu dalam klasifikasi risiko pasien penyakit kardiovaskular. Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk menerapkan algoritma KNN pada *dataset* untuk klasifikasi penyakit kardiovaskular. Diharapkan bahwa hasil penelitian ini dapat memberikan kontribusi dalam diagnosis pasien dengan penyakit kardiovaskular.

## 1.2 Rumusan Masalah

Berdasarkan pada latar belakang yang dijelaskan, rumusan masalah yang diambil adalah bagaimana mengetahui model *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular?

## 1.3 Batasan Masalah

1. Data diambil dari *BMC Medical Informatics and Decision Making* yang diambil oleh *Davide Chicco* dan *Giuseppe Jurman* berdasarkan rekam medis 299 pasien gagal jantung di Punjab, Pakistan (Chicco & Jurman, 2020).
2. Range nilai parameter (k) yang digunakan dalam pemodelan adalah 2 sampai 11 dan menggunakan *GridSearchCV* untuk mencari parameter terbaik.
3. Menggunakan 7 fitur.
4. Bahasa pemrograman menggunakan *Python*.

## 1.4 Tujuan Penelitian

Mengetahui kinerja *K-Nearest Neighbors* (KNN) dengan baik melalui perhitungan *accuracy*, *recall*, *precision*, dan *f1-score* pada klasifikasi penyakit kardiovaskular.

## 1.5 Manfaat Penelitian

1. Meningkatkan keakuratan diagnosis penyakit kardiovaskular melalui penggunaan model *K-Nearest Neighbors* (KNN).
2. Diharapkan dapat menjadi sumber referensi yang berguna bagi penelitian selanjutnya.

## BAB II

### STUDI PUSTAKA

#### 2.1 Penelitian Terkait

Penelitian yang dilakukan oleh (Samosir *et al.*, 2021) membandingkan tiga metode klasifikasi, yaitu *Random Forest*, *Naïve Bayes*, dan *K-Nearest Neighbor*, dalam mengklasifikasikan data penyakit jantung. Penelitian ini mengevaluasi kinerja ketiga metode tersebut dalam hal akurasi, *recall*, dan *precision*. Meskipun dilakukan dengan *dataset* berukuran kecil, studi ini dapat diperluas melalui pengujian menggunakan dataset yang lebih besar. Dataset yang digunakan diambil dari *Cleveland Clinic Foundation* yang telah diadopsi oleh *Hungarian Institute of Cardiology* di *Budapest*.

Penelitian (Fiqriansyah *et al.*, 2022) melakukan penelitian yang menerapkan Algoritma *Naïve Bayes* untuk mengidentifikasi pasien dengan penyakit gagal jantung. Penelitian ini menyoroti potensi penyakit jantung untuk menyerang khususnya pada pria dengan usia di bawah 60 tahun. Penelitian ini memusatkan perhatian pada penyelidikan menggunakan algoritma *Naïve Bayes*.

Penelitian (Subarkah *et al.*, 2022) tentang *Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification*. Penelitian tersebut menggunakan *dataset Heart Failure Prediction*. Dalam penelitiannya ini mengklasifikasikan pasien berdasarkan data dengan menggunakan algoritma *naïve bayes*.

Penelitian (Munandar & Munir, 2022) membahas implementasi *K-Nearest Neighbor* untuk menciptakan *prototipe* sistem pakar yang dapat mengidentifikasi dini penyakit jantung. Penelitian ini mencatat pencapaian dalam pembuatan *prototipe* sistem pakar berbasis web yang mampu mengklasifikasikan penyakit kardiovaskular, yang diharapkan dapat meningkatkan layanan kesehatan khususnya dalam penanganan pasien penyakit jantung. Namun, meskipun demikian, akurasi yang diperoleh masih rendah.

Penelitian (Shaban *et al.*, 2020) tentang “*A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier*”. Pada penelitian ini *K-Nearest Neighbor* (KNN) telah membuktikan fleksibilitas yang tinggi dalam permasalahan klasifikasi yang kompleks.

Penelitian (Kiran *et al.*, 2023) tentang “*A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for cardiovascular disease detection and classification*”. Penelitian ini bertujuan untuk menemukan kerangka klasifikasi pembelajaran mesin dengan presisi diagnostik tertinggi.

Penelitian (Putri, 2021) tentang “*Analisis Performa Metode K- Nearest Neighbor* (KNN) dan *Crossvalidation* pada Data Penyakit *Cardiovascular*”. Tujuan dari penelitian ini adalah mengukur performa (akurasi, presisi, *recall* dan *f-measure*) metode KNN dan *crossvalidation* pada *dataset cardiovascular*.

Penelitian (Argina, 2020) tentang “*Penerapan Metode Klasifikasi K-Nearest Neighbor* pada *Dataset* Penderita Penyakit Diabetes”. Tujuan penelitiannya yaitu mengukur akurasi, presisi, *recall*, dan *f-measure*.

Tabel 2. 1 Perbandingan Penelitian

No	Peneliti	Objek	Metode	Hasil	Perbandingan Penelitian ini
1.	(Samosir <i>et al.</i> , 2021)	Klasifikasi Data Penyakit Jantung	<i>Random Forest</i> , <i>Naïve Bayes</i> dan <i>K-Nearest Neighbor</i>	<i>Naïve Bayes</i> memiliki rerata hasil akurasi sebesar 0,91 <i>AUC</i> , 0,84 <i>CA</i> , 0,84 <i>F1</i> , 0,839 <i>Precision</i> dan 0,84 <i>Recall</i> . <i>Random Forest</i> sebesar 0,884 <i>AUC</i> , 0,788 <i>CA</i> , 0,787 <i>F1</i> , 0,788 <i>Precision</i> , 0,788 <i>Recall</i> <i>KNN</i> memiliki akurasi sebesar 0,686 <i>AUC</i> , 0,645 <i>CA</i> , 0,641 <i>F1</i> , 0,642 <i>Precision</i> , 0,645 <i>Recall</i>	a) Menggunakan metode <i>K-Nearest Neighbor</i> b) (Samosir <i>et al.</i> , 2021) Pengambilan <i>dataset</i> berasal dari lokasi <i>Cleveland Clinic Foundation</i> , Sedangkan untuk Peneliti sekarang mengambil data di “ <i>BMC Medical Informatics and Decision Making</i> ” (2020)
2.	(Fqriansyah <i>et al.</i> , 2022)	Mengetahui Pasien Penyakit Gagal Jantung	<i>Naïve Bayes</i>	Akurasi 80%	a) Menggunakan metode <i>K-Nearest Neighbor</i> b) (Fqriansyah <i>et al.</i> , 2022) Data berasal dari Data Riset Kesehatan Dasar 2013 yang dikeluarkan oleh Badan Litbangkes Kementerian Kesehatan RI dan Data Penduduk Sasaran dari Pusdatin Kementerian Kesehatan RI. Namun, peneliti saat ini menggunakan data yang diambil dari “ <i>BMC Medical Informatics and Decision Making</i> ” (2020)
3.	(Subarkah <i>et al.</i> , 2022)	“Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classificatio”	<i>Naïve Bayes</i>	Akurasi 80.6%	a) Menggunakan metode <i>K-Nearest Neighbor</i> . b) Menggunakan <i>GridSearchCV</i> untuk <i>Hyperparameter</i>

No	Peneliti	Objek	Metode	Hasil	Perbandingan Penelitian ini
4.	(Munandar & Munir, 2022)	“Prototype Sistem Pakar Identifikasi Dini Penyakit Jantung”	<i>K-Nearest Neighbor</i>	Tingkat akurasi sebesar 70.04%	a) Mengambil data di <i>BMC Medical Informatics and Decision Making</i> (2020), dan menggunakan 7 fitur b) Menggunakan <i>GridSearchCV</i> untuk <i>Hyperparameter</i>
5.	(Shaban <i>et al.</i> , 2020)	“A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection”	<i>KNN classifier</i>	Nilai presisi, <i>recall</i> , akurasi, dan error masing-masing mencapai 0,72, 0,71, 0,93, dan 0,07.	a) Objek Penyakit kardiovaskular b) Mengambil data di <i>BMC Medical Informatics and Decision Making</i> (2020)
6.	(Kiran <i>et al.</i> , 2023)	<i>Optimizer for cardiovascular disease detection and classification</i>	<i>Decision Tree</i>	Akurasi ringkasan model sebesar 97,89%, sensitivitas rata-rata sebesar 97,89%, presisi rata-rata 97,86%, dan rata-rata model dan skor F1 sebesar 97,43%.	a) Metode menggunakan <i>K-Nearest Neighbor</i> . b) Mengambil data di “ <i>BMC Medical Informatics and Decision Making</i> ” (2020) c) Menggunakan <i>GridSearchCV</i> untuk <i>Hyperparameter</i>
7.	(Putri, 2021)	Penyakit <i>Cardiovascular</i>	<i>K-Nearest Neighbor (KNN)</i>	Tingkat akurasi mencapai 87%, dengan presisi dan <i>recall</i> masing-masing sebesar 87% dan 97%, serta nilai <i>f-measure</i> sebesar 92%.	a) Mengambil data di <i>BMC Medical Informatics and Decision Making</i> (2020) b) Menggunakan <i>GridsearchCV</i> untuk menentukan <i>hyperparameter</i>
8.	(Argina, 2020)	Klasifikasi Penderita Penyakit Diabetes	<i>K-Nearest Neighbor</i>	Nilai akurasi sebesar 39%	a) Mengambil data di <i>BMC Medical Informatics and Decision Making</i> (2020) b) Menggunakan <i>GridsearchCV</i> untuk menentukan <i>hyperparameter</i>

Tabel 2.1 diatas menjelaskan penelitian sebelumnya yang telah menggunakan beberapa metodenya adalah *K-Nearest Neighbor*, *Naïve Bayes*, *Decision Tree* digunakan untuk mengklasifikasi penyakit, dan beberapa penelitian diatas objeknya yaitu tentang penyakit kardiovaskular. Hasil menunjukkan bahwa dapat memberikan akurasi yang baik dalam pengklasifikasian penyakit. Penelitian saat ini akan mengklasifikasikan pasien yang menderita penyakit kardiovaskular. Metode yang digunakan adalah *K-Nearest Neighbor* untuk klasifikasi, dan hasilnya akan dievaluasi dengan menggunakan *Receiver Operating Characteristic (ROC)* untuk menilai kinerja sistem klasifikasi ini. Penelitian ini akan menghadirkan tiga pembaruan. Pertama, penggunaan *dataset* dari *BMC Medical Informatics and Decision Making (2020)* yang berjudul “*Heart Failure Prediction*” yang open akses. Kedua, menggunakan 7 fitur untuk data inputannya. Terakhir, menggunakan metode *K-Nearest Neighbor* untuk mengklasifikasikan pasien penderita kardiovaskular agar memudahkan dokter dalam mendiagnosanya.

## **2.2 Penyakit Kardiovaskular**

Penyakit kardiovaskular merupakan jenis penyakit yang melibatkan gangguan pada jantung dan pembuluh darah, dan saat ini menjadi salah satu tantangan kesehatan global yang mendalam. Penyakit kardiovaskular meliputi sejumlah kondisi, seperti gagal jantung, hipertensi (tekanan darah tinggi), *stroke*, dan penyakit pembuluh darah. Berdasarkan data terkini, penyakit kardiovaskular merupakan penyebab utama kematian di seluruh dunia. Pada tahun 2019, penyakit kardiovaskular menyebabkan sekitar 18,6 juta kematian, yang setara dengan sekitar 32% dari total kematian global. (*World Health Organization, 2021*).

Faktor risiko utama yang berhubungan dengan munculnya penyakit kardiovaskular meliputi tekanan darah tinggi, kadar kolesterol tinggi, diabetes, obesitas, kurangnya aktivitas fisik, serta pola makan tidak sehat yang kaya akan lemak dan garam. (*American Heart Association*, 2021). Selain itu, faktor-faktor genetik dan usia juga memainkan peran dalam risiko mengembangkan Penyakit Kardiovaskular. Pola makan yang tidak seimbang, yang seringkali mencakup konsumsi makanan cepat saji dan tinggi gula tambahan, juga dapat berkontribusi pada risiko penyakit kardiovaskular (Kemenkes RI, 2021).

Strategi Kementerian Kesehatan (Kemenkes) untuk pencegahan penyakit kardiovaskular atau penyakit jantung meliputi beberapa langkah. Pertama, menyebarkan informasi edukatif kepada masyarakat melalui berbagai media sebanyak mungkin. Kedua, mendorong deteksi dini dengan mengajak masyarakat untuk secara rutin memeriksa tekanan darah, gula darah, indeks massa tubuh, atau lingkar perut. Ketiga, memperkenalkan langkah-langkah perlindungan khusus, seperti mengimplementasikan kawasan tanpa rokok. Terakhir, memberikan perhatian pada pengobatan, yang merupakan langkah terakhir setelah seseorang didiagnosis dengan penyakit jantung (Kemenkes RI, 2019).

Diagnosis dan manajemen Penyakit Kardiovaskular melibatkan berbagai tes laboratorium untuk memantau tekanan darah, kadar kolesterol, dan gula darah. Penggunaan teknologi medis seperti *elektrokardiogram* (EKG) dan pencitraan jantung (seperti *angiografi*) dapat membantu dalam diagnosis penyakit jantung spesifik (*American College of Cardiology*, 2021). Selain itu, perubahan gaya hidup

seperti perbaikan diet, olahraga teratur, dan pengelolaan stres adalah komponen penting dari pengobatan dan pencegahan penyakit kardiovaskular.

Upaya pencegahan juga melibatkan penggunaan obat-obatan seperti antihipertensi dan statin untuk mengatur tekanan darah dan kolesterol. Kampanye kesehatan masyarakat dan edukasi adalah langkah penting untuk meningkatkan kesadaran masyarakat tentang faktor risiko penyakit kardiovaskular dan mendorong perilaku sehat. Penelitian terus berlanjut untuk memahami lebih dalam tentang penyakit kardiovaskular dan mengembangkan terapi yang lebih efektif (Putri, 2021).

### **2.3 Klasifikasi**

Menurut definisi Menurut definisi dalam KBBI (Kamus Besar Bahasa Indonesia), klasifikasi adalah tindakan mengatur suatu objek dalam kelompok atau kategori berdasarkan prinsip atau kriteria yang telah disetujui. Klasifikasi adalah langkah untuk membagi sekelompok objek menjadi kelompok-kelompok berdasarkan atribut atau karakteristik yang dimiliki oleh masing-masing objek (Mourya *et al.*, 2018). Tujuan klasifikasi ini untuk mengelompokkan objek-objek yang memiliki kesamaan sifat dan memisahkan mereka dari objek-objek yang memiliki perbedaan. Dalam proses klasifikasi, objek-objek tersebut dikategorikan berdasarkan ciri khususnya, sehingga objek yang termasuk dalam satu kategori akan memiliki kesamaan, sementara objek dari kategori yang berbeda memiliki perbedaan. Proses klasifikasi ini berguna untuk mengorganisasi data secara terstruktur, memungkinkan penggunaannya dalam berbagai analisis data lebih

lanjut, termasuk prediksi, pemodelan, pengambilan keputusan, dan lainnya (Mantecipour *et al.*, 2018).

Metode klasifikasi melibatkan penggunaan berbagai algoritma dan teknik, termasuk pembelajaran mesin, statistik, dan pengolahan sinyal. Algoritma klasifikasi ini dipergunakan untuk mengategorikan data ke dalam berbagai kelas berdasarkan atribut-atribut atau fitur-fitur spesifik yang terdapat dalam data tersebut. Hasilnya adalah model klasifikasi yang dapat digunakan untuk mengklasifikasi kelas atau kategori yang sesuai untuk data baru yang belum dikenal (Nugraha, 2021). Klasifikasi memiliki beragam aplikasi di berbagai sektor, termasuk di ilmu komputer, ilmu kesehatan, bisnis, dan bidang lainnya. Di dalam ilmu komputer, klasifikasi digunakan untuk pengenalan pola, pemrosesan bahasa alami, serta pengenalan wajah. Di bidang kesehatan, klasifikasi dimanfaatkan untuk mendiagnosis penyakit berdasarkan gejala yang dialami oleh pasien. (Annisa, 2019).

Evaluasi klasifikasi merupakan tahap krusial dalam memverifikasi kinerja model klasifikasi. Ini melibatkan penilaian seberapa baik model dapat mengklasifikasikan data dengan tepat. Dalam proses tersebut, berbagai metrik evaluasi seperti akurasi, presisi, *recall*, dan *F1-score* digunakan untuk mengukur performa model. (Widodo Romy Budhi, 2022).

#### **2.4 *K-Nearest Neighbors* (KNN)**

Algoritma *K-Nearest Neighbors* (KNN) merupakan algoritma klasifikasi yang populer dalam bidang *machine learning*. Konsep dasar *K-Nearest Neighbors* (KNN) adalah sederhana, yaitu mencari "tetangga terdekat" dari suatu data

berdasarkan kesamaan fitur atau atribut (Baharuddin *et al.*, 2019). Cara kerja *K-Nearest Neighbors* (KNN) melibatkan perbandingan data yang akan diklasifikasikan dengan data pelatihan yang sudah ada dalam *dataset* (Cunningham & Delany, 2021). Data akan diklasifikasikan ke dalam kelas mayoritas dari tetangga terdekatnya (Widodo Romy Budhi, 2022).

Salah satu keunggulan dari algoritma *K-Nearest Neighbors* (KNN) adalah kesederhanaannya yang memudahkan pemahaman dan implementasi. KNN dapat diterapkan baik untuk masalah klasifikasi maupun regresi, walaupun lebih umum digunakan dalam konteks klasifikasi. (Sahar, 2020). Selain itu, *K-Nearest Neighbors* (KNN) tidak memerlukan asumsi tentang distribusi data, sehingga cocok untuk berbagai jenis data. Namun, ada juga beberapa kelemahan *K-Nearest Neighbors* (KNN), seperti sensitif terhadap *outlier* dan komputasi jika *dataset* besar. Selain itu, pemilihan parameter jumlah tetangga terdekat (nilai  $k$ ) yang akan digunakan juga bisa menjadi tantangan (Cunningham & Delany, 2021).

*K-Nearest Neighbors* (KNN) telah diterapkan dalam berbagai kasus penggunaan seperti kesehatan atau medis (Novitasari *et al.*, 2019).. Salah satu contohnya adalah dalam sistem diagnosis medis, dimana *K-Nearest Neighbors* (KNN) dapat digunakan untuk membantu mengklasifikasikan penyakit berdasarkan data medis pasien seperti riwayat penyakit, hasil tes, dan parameter klinis lainnya (Shaban *et al.*, 2020).

Langkah-langkah dalam menerapkan algoritma KNN menurut (Daqiqil, 2021) dalam bukunya yang berjudul “*Machine Learning* Teori, Studi Kasus dan Implementasi Menggunakan *Python*” adalah sebagai berikut:

1. Menentukan nilai  $k$ , yaitu jumlah tetangga terdekat yang akan dipertimbangkan.
2. Menghitung jarak antara titik baru dengan seluruh data latih yang ada.
3. Memilih  $k$  titik terdekat dengan titik data baru.
4. Menghitung jumlah titik terdekat berdasarkan kategori atau kelas untuk kasus klasifikasi.

Perhitungan jarak sangat penting untuk keberhasilan algoritma ini. Jarak menunjukkan tingkat kesamaan antara data uji dan data latih. Semakin besar jaraknya, semakin besar perbedaannya, dan sebaliknya. Jarak yang dapat digunakan yaitu *Euclidean Distance*. *Euclidean* yaitu jarak yang paling umum digunakan pada data numerik. *Euclidean distance* atau jarak *Euclidean* adalah metrik jarak ukur antara dua vektor dengan menghitung akar kuadrat dari jumlah selisih kuadrat antara keduanya (Zhang *et al.*, 2017). Untuk jarak *euclidean* didefinisikan sebagai berikut:

$$d(x, y) = \sqrt{(\sum_{i=1}^n (x_i - y_i)^2)} \quad (2.1)$$

Keterangan:

$(x_i)$  = data *training*

$(y_i)$  = data *testing*

$(i)$  = variabel data

$d$  = jarak

## 2.5 *GridSearchCV*

*GridSearchCV* merupakan metode seleksi optimal yang bertujuan untuk menemukan kombinasi *hyperparameter* terbaik guna meningkatkan kinerja model. Pendekatan ini menggabungkan berbagai kombinasi *hyperparameter* untuk

menghitung rata-rata performa model pada setiap kombinasi tersebut. Proses ini memungkinkan pemilihan *hyperparameter* yang paling optimal untuk diterapkan pada model. GridSearchCV adalah bagian dari pustaka scikit-learn yang melakukan validasi silang untuk beberapa model sekaligus dan secara otomatis menentukan hyperparameter masing-masing model secara sistematis. (Maisat *et al.*, 2023).

## 2.6 Synthetic Minority Oversampling Technique (SMOTE)

*Synthetic Minority Oversampling Technique (SMOTE)* adalah sebuah teknik yang digunakan untuk menyeimbangkan distribusi jumlah sampel data pada kelas minoritas dengan cara membuat sampel sintetis hingga jumlahnya seimbang dengan jumlah sampel pada kelas mayoritas. Penggunaan metode *SMOTE* dapat meningkatkan risiko overfitting karena duplikasi data pada kelas minoritas dapat menyebabkan adanya data pelatihan yang sama. *SMOTE* beroperasi dengan menciptakan sampel sintetis baru untuk data minoritas, dimana pembuatan sampel sintetis didasarkan pada prinsip k-nearest neighbor (tetangga terdekat). Implementasi *SMOTE* ini tidak mengubah jumlah kasus mayoritas. *Instans* baru bukan hanya salinan kasus tetapi sebagai gantinya, algoritma mengambil sampel ruang fitur untuk setiap kelas target dan tetangga terdekatnya (Nikmatul Kasanah *et al.*, 2021).

Langkah-langkah dalam menerapkan *SMOTE* adalah:

1. Menghitung jarak antara data pada kelas minoritas dan menetapkan jumlah tetangga terdekat ( $k$ ) yang akan digunakan menggunakan jarak Euclidean. Rumus untuk mencari jarak euclidean seperti pada rumus (2.1).
2. Kemudian menciptakan data sintesis dengan persamaan berikut:

$$x_{syn} = x_i (x_{knn} - x_i) \times \sigma \quad (2.2)$$

Keterangan:

$x_{syn}$  = data sintesis yang akan dibuat

$x_i$  = data yang akan dijadikan referensi untuk pembuatan data sintetis

$x_{knn}$  = data yang paling dekat jaraknya dari data referensi

$\sigma$  = nilai acak antara 0 dan 1

## 2.7 Confusion Matrix

*Confusion matrix* merupakan teknik yang digunakan untuk menghitung performa model. Dengan menggunakan confusion matrix, kita dapat menghasilkan nilai akurasi, presisi, dan *recall* sebagai evaluasi. Dilihat pada tabel 2.2 *Confusion matrix model*.

Tabel 2. 2 *Confusion matrix*

<i>Actual Class</i>		<i>Predicted Class</i>	
		<i>Positive (P)</i>	<i>Negative (N)</i>
<i>Positive (P)</i>		<i>True Positive (TP)</i>	<i>False Negative (FN)</i>
<i>Negative (N)</i>		<i>FP (False Positive)</i>	<i>TN (True Negative)</i>

TP (*True Positif*) adalah metrik yang mengukur jumlah data positif yang berhasil diklasifikasikan dengan benar oleh sistem. Sebaliknya, FP (*False Positif*) adalah ukuran yang menghitung jumlah data positif yang salah diklasifikasikan oleh sistem. FN (*False Negative*) adalah parameter yang menghitung jumlah data negatif yang salah diklasifikasikan oleh sistem. TN (*True Negative*) adalah ukuran yang menghitung jumlah data negatif yang berhasil diklasifikasikan dengan benar oleh sistem. *Accuracy*, *precision*, *recall*, dan *f1-score* merupakan evaluasi yang akan dihitung berdasarkan nilai-nilai dalam *confusion matrix*. Persamaan dalam menghitung *accuracy*, *precision*, *recall*, dan *f1-score* akan digunakan untuk mengevaluasi kinerja sistem.

Akurasi (*Accuracy*) adalah parameter yang mengukur sejauh mana model mampu melakukan klasifikasi secara benar. Akurasi mengindikasikan sejauh mana model mampu mengklasifikasi dengan tepat baik untuk hasil positif maupun negatif dalam keseluruhan *dataset* yang ada. Akurasi dihitung sebagai rasio antara jumlah klasifikasi yang benar (*True Positive* dan *True Negative*) dengan total jumlah *instance* yang diuji (Baharuddin *et al.*, 2019). Nilai akurasi dapat dihitung menggunakan rumus (2.3).

$$Akurasi = \frac{TP+TN}{TP+FP+FN+TN} \quad (2.3)$$

Keterangan:

- TP = *True Positive* (jumlah sampel positif yang terklasifikasi dengan benar)
- FN = *False Negative* (jumlah sampel positif yang terklasifikasi salah)
- FP = *False Positive* (jumlah sampel yang terklasifikasi salah sebagai positif)
- TN = *True Negative* (jumlah sampel yang terklasifikasi benar sebagai negatif)

Presisi (*Precision*) ukuran untuk mengevaluasi sejauh mana hasil klasifikasi model sesuai dengan data yang diminta. Secara lebih rinci, presisi mengukur tingkat keakuratan klasifikasi positif, yaitu sejauh mana hasil yang diklasifikasi sebagai positif yang sesungguhnya juga benar-benar positif. Ini dihitung dengan membandingkan jumlah klasifikasi positif yang tepat dengan jumlah total hasil klasifikasi positif. Presisi membantu untuk memahami seberapa banyak dari data yang telah diklasifikasi sebagai positif yang memang benar-benar positif dalam kenyataannya. Nilai *precision* dapat dihitung menggunakan rumus (2.4).

$$Precision = \frac{TP}{TP+FP} \quad (2.4)$$

Keterangan:

- TP = *True Positive* (jumlah sampel positif yang terklasifikasi dengan benar)
- FP = *False Positive* (jumlah sampel yang terklasifikasi salah sebagai positif)

*Recall* adalah metrik yang mengukur kemampuan model dalam menemukan informasi yang benar. Ini dihitung dengan membandingkan jumlah klasifikasi positif yang benar dengan total data yang sebenarnya positif. Nilai *recall* dapat dihitung menggunakan rumus (2.5).

$$Recall = \frac{TP}{TP+FN} \quad (2.5)$$

Keterangan:

TP = *True Positive* (jumlah sampel positif yang terklasifikasi dengan benar)

FN = *False Negative* (jumlah sampel positif yang terklasifikasi salah)

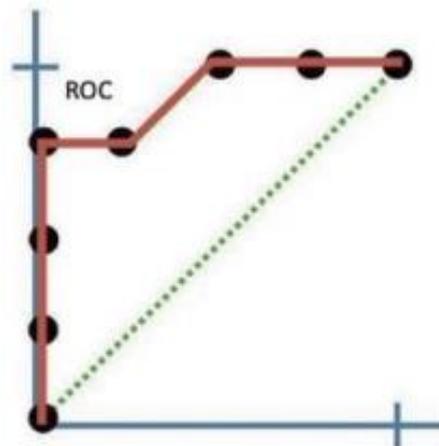
*F1-score* adalah metrik yang menggabungkan rata-rata *precision* dan *recall* secara terbobot. Nilai *f1-score* dapat dihitung menggunakan rumus (2.6).

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (2.6)$$

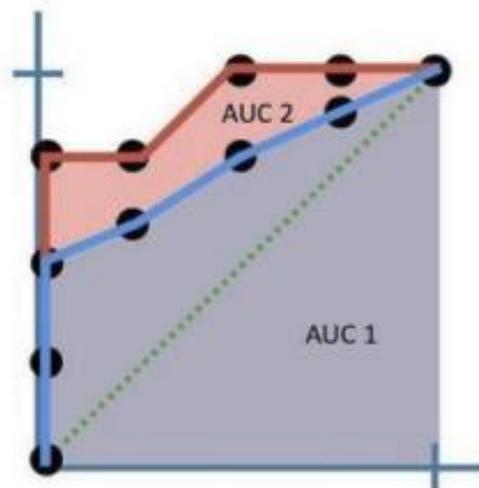
## 2.8 Receiver Operating Characteristic (ROC)

*Receiver Operating Characteristic (ROC)* merupakan sebuah kurva yang digunakan untuk mengevaluasi kinerja model klasifikasi, terutama dalam bidang pengenalan pola atau diagnostik medis. (Ahnan *et al.*, 2023). Selain *confusion matrix*, kurva *receiver operating characteristic (ROC)* digunakan untuk mengevaluasi akurasi model machine learning. *Confusion matrix* memberikan performa algoritma klasifikasi dalam bentuk angka, tetapi untuk representasi visual yang lebih jelas *ROC* atau kurva *precision-recall* (Narabel & Budi, 2020). *ROC* didapatkan dari perhitungan nilai *false positive rate* dan *true positive rate* yang dihasilkan dari *confusion matrix*. Untuk membandingkan nilai kinerja masing-

masing algoritma dapat dilakukan dengan membandingkan luas *AUC* (*Area Under Curve*) seperti pada gambar dibawah ini adalah kurva *ROC* dan *AUC* untuk lebih jelasnya (Kristiawan & Widjaja, 2021).



Gambar 2. 1 Receiver Operator Characteristic  
Sumber: Kristiawan & Widjaja, 2021



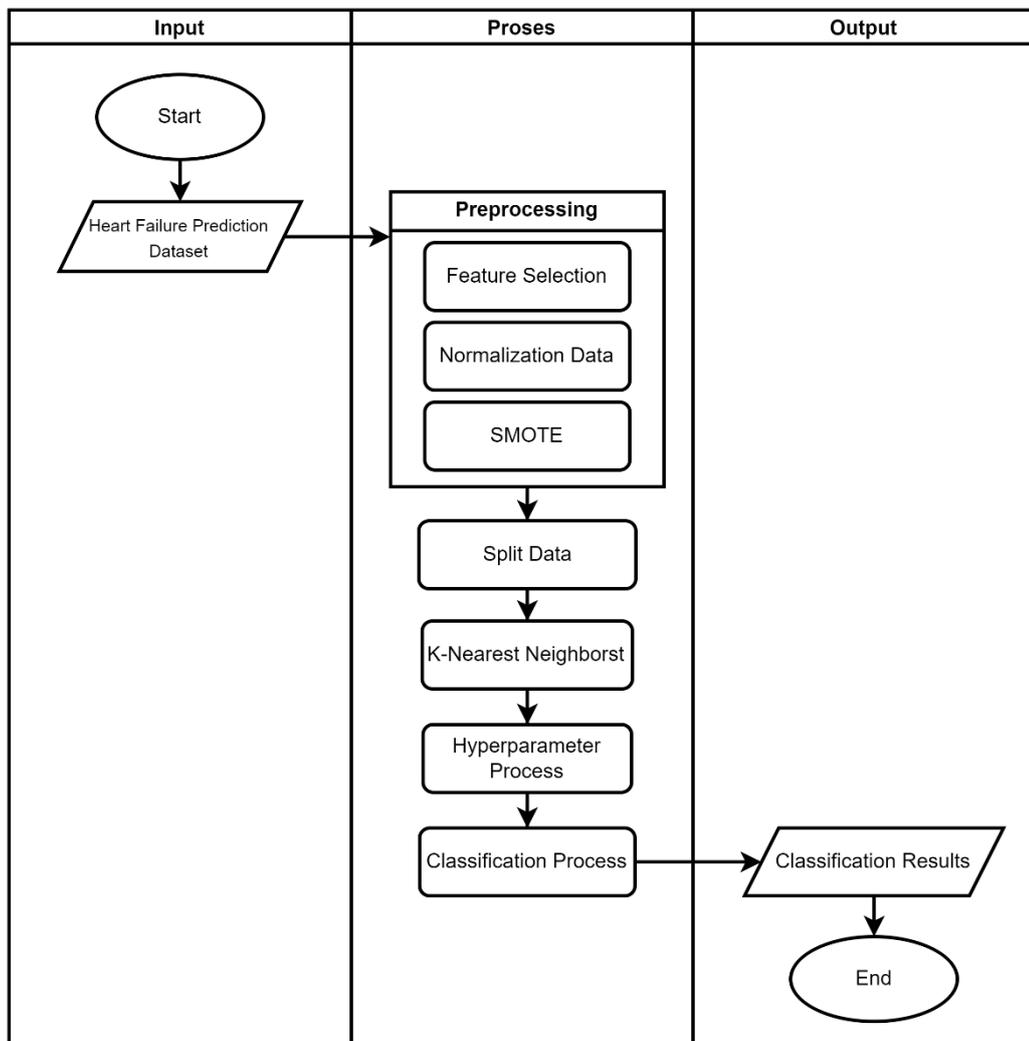
Gambar 2. 2 Area Under Curve  
Sumber: Kristiawan & Widjaja, 2021

### BAB III

## METODOLOGI PENELITIAN

### 3.1 Desain Sistem

Desain sistem merupakan tahapan yang akan digunakan oleh peneliti untuk melakukan penelitian. Proses klasifikasi penyakit kardiovaskular pada penelitian ini memerlukan data, *Preprocessing*, implementasi *K-Nearest Neighbors* (KNN), dan evaluasi. Gambar 3.1 menunjukkan perancangan sistem.



Gambar 3. 1 Desain Sistem

### 3.2 Pengumpulan Data

Langkah ini peneliti akan melakukan pengumpulan data. Data pada penelitian ini yaitu *dataset* “*Heart Failure Prediction*” yang merupakan *open access* data web dari “*BMC Medical Informatics and Decision Making*” (Chicco & Jurman, 2020). *Dataset* ini dikumpulkan oleh *Davide Chicco* dan *Giuseppe Jurman* menggunakan rekam medis dari 299 pasien yang mengalami gagal jantung di Punjab, Pakistan. Data ini dikumpulkan di *Institut Kardiologi Faisalabad* bekerja sama dengan Rumah Sakit *Faisalabad* selama periode April hingga Desember 2015. Penjelasan dari fitur-fitur pada *dataset* pada tabel 3.1. Contoh sampel data pelatihan dan data pengujian dapat dilihat pada tabel 3.2.

Tabel 3. 1 Deskripsi Data

Fitur	Penjelasan	Pengukuran	Jangkauan
<i>Age</i>	Usia Pasien	Tahun	[40,..., 95]
<i>Anaemia</i>	Penurunan sel darah merah	Boolean	0, 1
<i>High blood pressure</i>	Tekanan darah tinggi (Jika pasien menderita hipertensi)	Boolean	0, 1
<i>Creatinine phosphokinase</i>	Tingkat enzim dalam darah	mcg/L	[23,..., 7861]
<i>Diabetes</i>	Jika pasien menderita diabetes	Boolean	0, 1
<i>Ejection fraction</i>	Persentase darah yang keluar dari jantung pada setiap kontraksi	Percentage	[14,..., 80]
<i>Sex</i>	Jenis kelamin pasien	Binary	0, 1
<i>Platelets</i>	Trombosit di dalam darah	kiloplatelets/mL	[25.01,..., 850.00]
<i>Serum creatinine</i>	Tingkat kreatinin dalam darah	mg/dL	[0.50,..., 9.40]
<i>Serum sodium</i>	Tingkat sodium dalam darah	mEq/L	[114,..., 148]
<i>Smoking</i>	Jika pasien merokok	Boolean	0, 1
<i>Time</i>	Periode tindak lanjut	Days	[4,...,285]
(target) <i>death event</i>	Jika pasien meninggal selama masa tindak lanjut	Boolean	0, 1

Pada tabel ini adalah contoh sampel *dataset* untuk data pelatihan dan data pengujian

Tabel 3. 2 Contoh sampel *dataset*

<b>Contoh Data Latih</b>												
<i>Age</i>	<i>Anaemia</i>	<i>High blood pressure</i>	<i>Creatinine phosphokinase</i>	<i>Diabetes</i>	<i>Ejection fraction</i>	<i>Sex</i>	<i>Platelets</i>	<i>Serum creatinine</i>	<i>Serum sodium</i>	<i>Smoking</i>	<i>Time</i>	<i>(target) death event</i>
75	0	1	582	0	20	1	265000.00	1.9	130	0	4	1
55	0	0	7861	0	38	1	263358.03	1.1	136	0	6	1
65	0	0	146	0	20	1	162000.00	1.3	129	1	7	1
50	1	0	111	0	20	1	210000.00	1.9	137	0	7	1
65	1	0	160	1	20	0	327000.00	2.7	116	0	8	1
90	1	1	47	0	40	1	204000.00	2.1	132	1	8	1
75	1	0	246	0	15	1	127000.00	1.2	137	0	10	1
60	1	0	315	1	60	1	454000.00	1.1	131	1	10	1
65	0	0	157	0	65	0	263358.03	1.5	138	0	10	1
80	1	1	123	0	35	1	388000.00	9.4	133	1	10	1
<b>Contoh Data Uji</b>												
<i>Age</i>	<i>Anaemia</i>	<i>High blood pressure</i>	<i>Creatinine phosphokinase</i>	<i>Diabetes</i>	<i>Ejection fraction</i>	<i>Sex</i>	<i>Platelets</i>	<i>Serum creatinine</i>	<i>Serum sodium</i>	<i>Smoking</i>	<i>Time</i>	<i>(target) death event</i>
75	1	1	81	0	38	1	368000.00	4.0	131	1	10	1

### 3.3 Preprocessing

*Preprocessing* adalah langkah pertama dalam proses pengolahan data sebelum memasuki tahap utama. Tahapan ini data disiapkan untuk analisis dengan tujuan memaksimalkan hasilnya. Penting untuk memberikan perhatian khusus pada tahap *Preprocessing* karena teknik yang diterapkan di sini akan memengaruhi cara model mempelajari data yang diberikan. *Preprocessing* yang dilakukan oleh peneliti ada tiga yaitu seleksi fitur, *SMOTE*, dan normalisasi data.

#### 3.3.1 Seleksi Fitur

Seleksi fitur merupakan tahap krusial dalam proses klasifikasi. Tujuannya adalah untuk memilih fitur-fitur yang signifikan dari sekumpulan fitur yang ada dalam dataset. Metode yang digunakan untuk seleksi fitur adalah *Feature Importance*. Perhitungan *Feature Importance* dengan rumus sebagai berikut:

$$ni_j = w_j C_j - w_{left(j)} C_{left(j)} - w_{right(j)} C_{right(j)} \quad (3.1)$$

Keterangan:

$ni$	= <i>feature importance</i>
$w_j$	= metrik fitur
$C_j$	= jumlah <i>instance</i>
$w_{left(j)}$	= metrik simpul anak kiri
$C_{left(j)}$	= jumlah <i>instance</i> anak kiri
$w_{right(j)}$	= metrik simpul anak kanan
$C_{right(j)}$	= jumlah <i>instance</i> anak kanan

Feature Importance:	
	Importance
time	35.04
serum_creatinine	16.76
ejection_fraction	11.47
age	8.76
serum_sodium	7.39
creatinine_phosphokinase	7.31
platelets	6.85
smoking	1.42
sex	1.40
anaemia	1.31
diabetes	1.29
high_blood_pressure	1.00

Gambar 3. 2 Hasil Perhitungan *Fiture Importance*

Dari gambar 3.2 adalah hasil perhitungan *fiture importance* dari setiap fitur yang ada pada *dataset*. Hasil tersebut menunjukkan urutan nilai *importance* paling tinggi sampai yang paling rendah. 12 fitur tersebut akan digunakan 7 fitur yang memiliki nilai tertinggi untuk digunakan dalam proses selanjutnya.

### 3.3.2 *Synthetic Minority Oversampling Technique (SMOTE)*

Tahap ini yaitu melakukan *Synthetic Minority Oversampling Technique (SMOTE)* digunakan untuk menyeimbangkan data. Data yang digunakan dalam penelitian ini memiliki ketidakseimbangan (*imbalance*) sehingga diperlukan penggunaan teknik *SMOTE* untuk menyeimbangkan distribusi kelasnya. Persebaran kelas pada *dataset "Heart Failure Prediction"* tidak seimbang targetnya. Kelas 0 (yang tidak meninggal) memiliki 203 data, sedangkan kelas 1 (yang meninggal) memiliki 96 data. Data tersebut tidak seimbang, karena itu akan menggunakan teknik *SMOTE* untuk menangani ketidakseimbangan data.

### 3.3.3 Normalisasi Data

Sebelum digunakan, data perlu menjalani proses normalisasi. Normalisasi data merupakan proses mengubah data sehingga semua nilai dalam dataset memiliki skala yang seragam. Tujuannya yaitu untuk menghindari masalah akibat perbedaan skala antar variabel dan untuk memastikan bahwa setiap variabel memiliki pengaruh yang seimbang dalam analisis statistik. Peneliti akan menggunakan *min-max scaling* yang menskalakan data kisaran 0 (nilai minimum) hingga 1 (nilai maksimum). Penskalaan diperlukan untuk *K-Nearest Neighbors* (KNN) yang sensitif terhadap perbedaan skala antar variabel ketika melakukan klasifikasi. Rumus normalisasi *Min-Max* seperti ditunjukkan dalam persamaan (3.1).

Berikut rumus persamaan *Min-Max Scaling*.

$$\text{Normalisasi} = \frac{x - \min}{\max - \min} \quad (3.2)$$

Normalisasi min-max bertujuan untuk mengubah data dengan skala yang bervariasi menjadi skala dari 0-1. Berikut ini perbandingan dari data sebelum dilakukan normalisasi (Tabel 3.3) , (Tabel 3.4) setelah dilakukan normalisasi.

Tabel 3. 3 Contoh *dataset* sebelum normalisasi

<i>Age</i>	<i>Anaemia</i>	<i>High blood pressure</i>	<i>Creatinine phosphokinase</i>	<i>Diabetes</i>	<i>Ejection fraction</i>	<i>Sex</i>	<i>Platelets</i>	<i>Serum creatinine</i>	<i>Serum sodium</i>	<i>Smoking</i>	<i>Time</i>
75	0	1	582	0	20	1	265000.00	1.9	130	0	4
55	0	0	7861	0	38	1	263358.03	1.1	136	0	6
65	0	0	146	0	20	1	162000.00	1.3	129	1	7
50	1	0	111	0	20	1	210000.00	1.9	137	0	7
65	1	0	160	1	20	0	327000.00	2.7	116	0	8

Tabel 3. 4 Contoh *dataset* setelah normalisasi

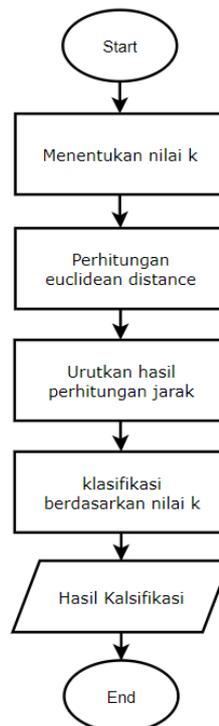
<i>Age</i>	<i>Anaemia</i>	<i>High blood pressure</i>	<i>Creatinine phosphokinase</i>	<i>Diabetes</i>	<i>Ejection fraction</i>	<i>Sex</i>	<i>Platelets</i>	<i>Serum creatinine</i>	<i>Serum sodium</i>	<i>Smoking</i>	<i>Time</i>
0.703598	0.0	1.0	0.635752	0.0	0.165986	1.0	0.435767	0.737676	0.336595	0.0	0.000000
0.341690	0.0	0.0	1.000000	0.0	0.512213	1.0	0.433664	0.489581	0.514993	0.0	0.029502
0.532920	0.0	0.0	0.390982	0.0	0.165986	1.0	0.290625	0.575762	0.310354	1.0	0.042589
0.236341	1.0	0.0	0.337651	0.0	0.165986	1.0	0.362001	0.737676	0.548482	0.0	0.042589
0.532920	1.0	0.0	0.408434	1.0	0.165986	0.0	0.511545	0.845279	0.044543	0.0	0.054846

### 3.4 Split Data

Setelah normalisasi data, langkah berikutnya adalah membagi data menjadi data pelatihan dan data pengujian untuk digunakan dalam proses klasifikasi. Terdapat tiga model yang akan digunakan dengan rasio pembagian yang berbeda. Model A menggunakan 80% data pelatihan dan 20% data pengujian, model B menggunakan 70% data pelatihan dan 30% data pengujian, dan model C menggunakan 60% data pelatihan dan 40% data pengujian.

### 3.5 Implementasi *K-Nearest Neighbor* (KNN)

Tahapan ini akan mengimplementasikan metode *K-Nearest Neighbor* (KNN) untuk melakukan klasifikasi pada penyakit kardiovaskular. Adapun *flowchart* dari implementasi metode *K-Nearest Neighbor* pada penelitian ini adalah sebagaimana gambar 3.3.



Gambar 3. 3 *Flowchart K-Nearest Neighbor*

Dari gambar 3.3 langkah-langkah implementasi KNN dapat dijelaskan sebagai berikut:

1. Menetapkan nilai  $k$  yang terbaik.
2. Setelah menentukan nilai  $k$  yang terbaik langkah selanjutnya yaitu klasifikasi menggunakan model K-NN. Pada tahapan ini klasifikasi dilakukan dengan menghitung jarak antar data menggunakan metode *Euclidean Distance*.
3. Menyusun hasil perhitungan langkah kedua secara berurutan dari yang terkecil hingga terbesar.
4. Melakukan klasifikasi *nearest neighbor* berdasarkan nilai  $k$ .
5. Menggunakan model *K-Nearest Neighbor* untuk melakukan klasifikasi.

### 3.5.1 Contoh Implementasi *K-Nearest Neighbor* (KNN)

Adapun contoh implementasi *K-Nearest Neighbor* pada penelitian ini dengan menggunakan rumus *Euclidean Distance* dengan *dataset* yang ada pada tabel 3.2 seperti berikut:

Tabel 3. 5 Perhitungan *Euclidean Distance*

No	Perhitungan <i>Euclidean distance</i>
1.	$\sqrt{(75 - 75)^2 + (0 - 1)^2 + (1 - 1)^2 + (582 - 81)^2 + (0 - 0)^2 + (20 - 38)^2 + (1 - 1)^2 + (265000.00 - 368000.00)^2 + (1.9 - 4.0)^2 + (130 - 131)^2 + (0 - 1)^2 + (4 - 10)^2 + (1 - 1)^2}$
2.	$\sqrt{(55 - 75)^2 + (0 - 1)^2 + (0 - 1)^2 + (7861 - 81)^2 + (0 - 0)^2 + (38 - 38)^2 + (1 - 1)^2 + (263358.03 - 368000.00)^2 + (1.1 - 4.0)^2 + (136 - 131)^2 + (0 - 1)^2 + (6 - 10)^2 + (1 - 1)^2}$
3.	$\sqrt{(65 - 75)^2 + (0 - 1)^2 + (0 - 1)^2 + (146 - 81)^2 + (0 - 0)^2 + (20 - 38)^2 + (1 - 1)^2 + (162000.00 - 368000.00)^2 + (1.3 - 4.0)^2 + (129 - 131)^2 + (1 - 1)^2 + (7 - 10)^2 + (1 - 1)^2}$
4.	$\sqrt{(50 - 75)^2 + (1 - 1)^2 + (0 - 1)^2 + (111 - 81)^2 + (0 - 0)^2 + (20 - 38)^2 + (1 - 1)^2 + (210000.00 - 368000.00)^2 + (1.9 - 4.0)^2 + (137 - 131)^2 + (0 - 1)^2 + (7 - 10)^2 + (1 - 1)^2}$
5.	$\sqrt{(65 - 75)^2 + (1 - 1)^2 + (0 - 1)^2 + (160 - 81)^2 + (1 - 0)^2 + (20 - 38)^2 + (0 - 1)^2 + (327000.00 - 368000.00)^2 + (2.7 - 4.0)^2 + (116 - 131)^2 + (0 - 1)^2 + (8 - 10)^2 + (1 - 1)^2}$
6.	$\sqrt{(90 - 75)^2 + (1 - 1)^2 + (1 - 1)^2 + (47 - 81)^2 + (0 - 0)^2 + (40 - 38)^2 + (1 - 1)^2 + (204000.00 - 368000.00)^2 + (2.1 - 4.0)^2 + (132 - 131)^2 + (1 - 1)^2 + (8 - 10)^2 + (1 - 1)^2}$

7.	$\sqrt{(75 - 75)^2 + (1 - 1)^2 + (0 - 1)^2 + (246 - 81)^2 + (0 - 0)^2 + (15 - 38)^2 + (1 - 1)^2 + (127000.00 - 368000.00)^2 + (1.2 - 4.0)^2 + (137 - 131)^2 + (0 - 1)^2 + (10 - 10)^2 + (1 - 1)^2}$
8.	$\sqrt{(60 - 75)^2 + (1 - 1)^2 + (0 - 1)^2 + (315 - 81)^2 + (1 - 0)^2 + (60 - 38)^2 + (1 - 1)^2 + (454000.00 - 368000.00)^2 + (1.1 - 4.0)^2 + (131 - 131)^2 + (1 - 1)^2 + (10 - 10)^2 + (1 - 1)^2}$
9.	$\sqrt{(65 - 75)^2 + (0 - 1)^2 + (0 - 1)^2 + (157 - 81)^2 + (0 - 0)^2 + (65 - 38)^2 + (0 - 1)^2 + (263358.03 - 368000.00)^2 + (1.5 - 4.0)^2 + (138 - 131)^2 + (0 - 1)^2 + (10 - 10)^2 + (1 - 1)^2}$
10.	$\sqrt{(80 - 75)^2 + (1 - 1)^2 + (1 - 1)^2 + (123 - 81)^2 + (0 - 0)^2 + (35 - 38)^2 + (1 - 1)^2 + (388000.00 - 368000.00)^2 + (9.4 - 4.0)^2 + (133 - 131)^2 + (1 - 1)^2 + (10 - 10)^2 + (1 - 1)^2}$

Tabel 3. 6 Hasil perhitungan *Euclidean Distance*

No	<i>Euclidean distance</i>	Rangking
1.	103011.52	4
2.	104635.22	6
3.	206262.78	9
4.	158078.56	7
5.	41003.15	2
6.	163967.41	8
7.	240944.21	10
8.	86000.00	3
9.	104635.07	5
10	20000.05	1

Tabel perhitungan jarak pada tabel 3.5 dan 3.6 diatas menggunakan rumus perhitungan jarak *Euclidean Distance*:

$$d(x, y) = \sqrt{(\sum_{i=1}^n (x_i - y_i)^2)} \quad (3.3)$$

Dimana:

$(x_i)$  : data *training*                       $(i)$  : variabel data  
 $(y_i)$  : data *testing*                      d : jarak

### 3.6 Proses *Hyperparameter*

Proses *hyperparameter* yaitu untuk mencari parameter optimal dengan menggunakan *GridSearchCV*. *GridSearchCV* adalah teknik yang digunakan untuk mencari *hyperparameter* terbaik guna membentuk model yang optimal sehingga menghasilkan skor akurasi tertinggi. *GridSearchCV* dirancang untuk menentukan *hyperparameter* optimal yang akan menghasilkan performa terbaik untuk model *K-Nearest Neighbors*. Berikut parameter penting yang digunakan peneliti:

Tabel 3. 7 Parameter

Parameter	Keterangan
$n\_neighbors=list(range(2,12))$	Nilai jarak k yang dicari mulai dari 2-11
$weight\_options = [“uniform”, “distance”]$	Berfungsi mengontrol bagaimana tetangga diberi "bobot" saat melakukan klasifikasi

Sebelumnya peneliti menentukan parameter apa saja yang akan dioptimalkan. Peneliti menggunakan parameter yang sesuai dengan tabel 3.7. *GridSearchCV* digunakan untuk menentukan *hyperparameter* optimal yang akan menghasilkan performa terbaik untuk model *K-Nearest Neighbors*.

### 3.7 Skenario Uji Coba

Penelitian ini melakukan penerapan tiga model data latih dan data uji. Pembagian dilakukan dengan model A menggunakan rasio 80% data pelatihan dan 20% data pengujian, model B menggunakan rasio 70% data pelatihan dan 30% data pengujian, serta model C menggunakan rasio 60% data pelatihan dan 40% data pengujian. Kemudian, pada setiap pembagian model tersebut, akan dilakukan penyetelan hyperparameter menggunakan *GridSearchCV* untuk menemukan nilai k atau jarak yang optimal pada model *K-Nearest Neighbor*. Pengujian model ini terdapat dua skenario pengujian yaitu:

1. Pengujian tidak menggunakan seleksi fitur dan *SMOTE*,
2. Pengujian menggunakan seleksi fitur dan *SMOTE*.

Tabel 3. 8 Skenario pembagian data

Model	Persentase Data Latih	Persentase data Uji
A	80%	20%
B	70%	30%
C	60%	40%

Langkah terakhir dalam penelitian ini adalah mengevaluasi kinerja penggunaan K-Nearest Neighbor dalam klasifikasi penyakit kardiovaskular. Tahapan ini peneliti menggunakan *confusion matrix* untuk menunjukkan berapa banyak klasifikasi model *K-Nearest Neighbor* yang akurat dan salah pada setiap kelas. Dari evaluasi *confusion matrix*, berbagai parameter kinerja model seperti akurasi, *recall*, presisi, dan *f1-score* juga dapat diidentifikasi. Selain menggunakan *confusion matrix*, peneliti juga akan mengevaluasi dengan menggunakan *curva receiver operating characteristic (ROC)*. Kurva ROC dibentuk dari perhitungan *nilai false positive rate* dan *true positive rate* yang dihasilkan dari *confusion matrix* untuk mendapatkan nilai *Area Under the Curve (AUC)*.

## BAB IV

### UJI COBA DAN PEMBAHASAN

#### 4.1 Hasil Uji Coba

Bagian ini menguraikan hasil analisis dari pengujian sistem berdasarkan skenario pengujian yang dijelaskan dalam sub-bab 3.7 yang terdapat tiga model yaitu model A data latih 80%, data uji 20%, model B data latih 70% data uji 30%, model C data latih 60% dan data uji 40%. Hasil pengujian ini untuk mengetahui performa dari metode *K-Nearest Neighbor* dalam mengklasifikasi penyakit kardiovaskular pada *dataset Heart Failure Prediction* yang telah terbagi menjadi tiga model pengujian yang akan diketahui akurasi. Penelitian ini melakukan pengujian menggunakan seleksi fitur, menerapkan *SMOTE* dan tidak menggunakan seleksi fitur, tidak menetapkan *SMOTE*. Hasil pengujiannya ini akan dibandingkan hasilnya.

##### 4.1.1 Pengujian Model A

Model A perbandingan yang digunakan antara data pelatihan dengan data pengujian sebesar 80:20. Pengujian model ini terdapat dua skenario pengujian yaitu tidak menggunakan seleksi fitur dan *SMOTE*, serta pengujian dengan menggunakan seleksi fitur dan *SMOTE*.

##### 4.1.1.1 Tidak Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian pertama ini yaitu tidak menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian dilakukan

dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai  $k$  atau  $n\_neighbor$  terbaik.

Tabel 4. 1 Hasil rata-rata skor pengujian tanpa seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.197323	{'n_neighbors': 2}
0.349961	{'n_neighbors': 3}
0.186111	{'n_neighbors': 4}
0.255758	{'n_neighbors': 5}
0.158016	{'n_neighbors': 6}
0.275339	{'n_neighbors': 7}
0.181919	{'n_neighbors': 8}
0.328446	{'n_neighbors': 9}
0.219658	{'n_neighbors': 10}
0.280668	{'n_neighbors': 11}

Sesuai tabel 4.1 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 3$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.349961.

Tabel 4. 2 *Confusion matrix* tanpa seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
5	39	2	14

Tabel 4.2 menunjukkan bahwa model *K-Nearest Neighbor* tanpa dilakukan seleksi fitur, dan *SMOTE* terdapat 5 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 39 sampel yang berhasil diklasifikasi negatif dengan benar, ada 2 sampel yang salah diklasifikasi sebagai positif, dan terdapat 14 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.2 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{5+39}{5+39+2+14} = 0.73$$

$$Precision = \frac{5}{5+2} = 0.71$$

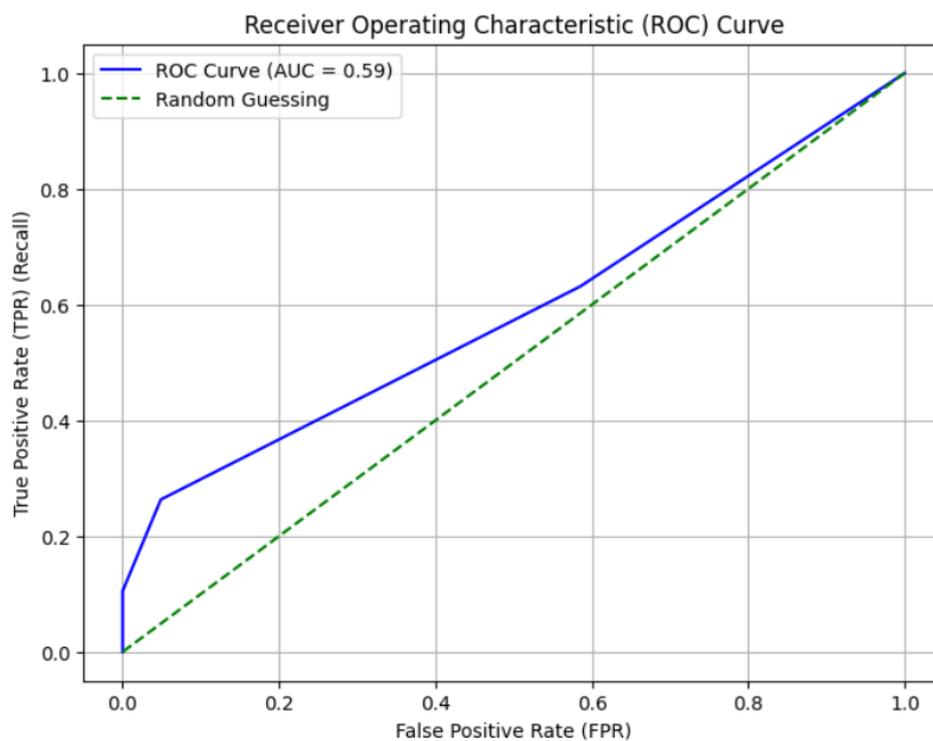
$$Recall = \frac{5}{5+14} = 0.26$$

$$F1 - Score = \frac{2 \times 0.71 \times 0.26}{0.71 + 0.26} = 0.38$$

Tabel 4. 3 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0.73	0.71	0.26	0.38

Pengujian tanpa seleksi fitur dan *SMOTE* pada model A ini didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.1 dibawah.

Gambar 4. 1 *Curva ROC* Model A tanpa seleksi fitur dan *SMOTE*

Gambar 4.1 menunjukkan hasil *AUC* 0.59. *AUC* di atas menunjukkan bahwa model memiliki kemampuan yang signifikan dalam membedakan antara kelas

positif dan negatif, tetapi belum mencapai kinerja yang sangat baik karena jauh dari angka 1.

#### 4.1.1.2 Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian kedua ini yaitu menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai  $k$  atau  $n\_neighbor$  terbaik.

Tabel 4. 4 Hasil rata-rata skor pengujian dengan seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.817314	{'n_neighbors': 2}
0.850399	{'n_neighbors': 3}
0.832756	{'n_neighbors': 4}
0.866960	{'n_neighbors': 5}
0.845947	{'n_neighbors': 6}
0.879312	{'n_neighbors': 7}
0.858611	{'n_neighbors': 8}
0.877553	{'n_neighbors': 9}
0.875600	{'n_neighbors': 10}
0.873691	{'n_neighbors': 11}

Sesuai tabel 4.4 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 7$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.879312.

Tabel 4. 5 *Confusion matrix* menggunakan seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
38	34	7	3

Tabel 4.5 menunjukkan bahwa model *K-Nearest Neighbor* terdapat 38 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 34 sampel yang berhasil

diklasifikasi negatif dengan benar, ada 7 sampel yang salah diklasifikasi sebagai positif, dan terdapat 3 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.5 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{38+34}{38+34+7+3} = 0.88$$

$$Precision = \frac{38}{38+7} = 0.84$$

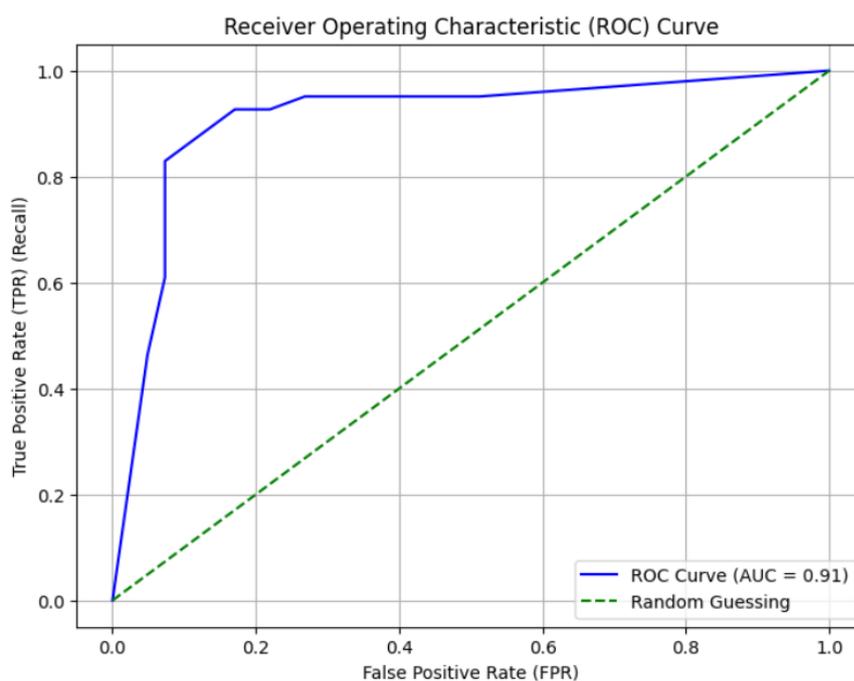
$$Recall = \frac{38}{38+3} = 0.93$$

$$F1 - Score = \frac{2 \times 0.84 \times 0.93}{0.84 + 0.93} = 0.88$$

Tabel 4. 6 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0.88	0.84	0.93	0.88

Pengujian menggunakan seleksi fitur dan *SMOTE* pada model A didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.2 dibawah.



Gambar 4. 2 *Curva ROC* Model A menggunakan seleksi fitur dan *SMOTE*

Gambar 4.2 menunjukkan hasil *AUC* sebesar 0.91. Nilai *AUC* yang mendekati 1 menunjukkan bahwa model memiliki kemampuan yang sangat tinggi dalam membedakan antara kelas positif dan negatif. Hasil di atas memiliki tingkat *True Positive Rate* (TPR) yang tinggi.

#### 4.1.2 Pengujian Model B

Model B perbandingan yang digunakan antara data latih dengan data uji sebesar 70:30. Pengujian model ini terdapat dua skenario pengujian yaitu tidak menggunakan seleksi fitur dan *SMOTE*, serta pengujian dengan menggunakan seleksi fitur dan *SMOTE*.

##### 4.1.2.1 Tidak Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian pertama ini yaitu tidak menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai *k* atau *n\_neighbor* terbaik.

Tabel 4. 7 Hasil rata-rata skor pengujian tanpa seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.274444	{'n_neighbors': 2}
0.419747	{'n_neighbors': 3}
0.189444	{'n_neighbors': 4}
0.365354	{'n_neighbors': 5}
0.221032	{'n_neighbors': 6}
0.401111	{'n_neighbors': 7}
0.230556	{'n_neighbors': 8}
0.414177	{'n_neighbors': 9}
0.266667	{'n_neighbors': 10}
0.366854	{'n_neighbors': 11}

Sesuai tabel 4.7 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 3$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.419747.

Tabel 4. 8 *Confusion matrix* tanpa seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
7	52	9	22

Tabel 4.8 menunjukkan bahwa model *K-Nearest Neighbor* tanpa dilakukan seleksi fitur, dan *SMOTE* terdapat 7 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 52 sampel yang berhasil diklasifikasi negatif dengan benar, ada 9 sampel yang salah diklasifikasi sebagai positif, dan terdapat 22 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.8 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{7+52}{7+52+9+22} = 0.66$$

$$Precision = \frac{7}{7+9} = 0.44$$

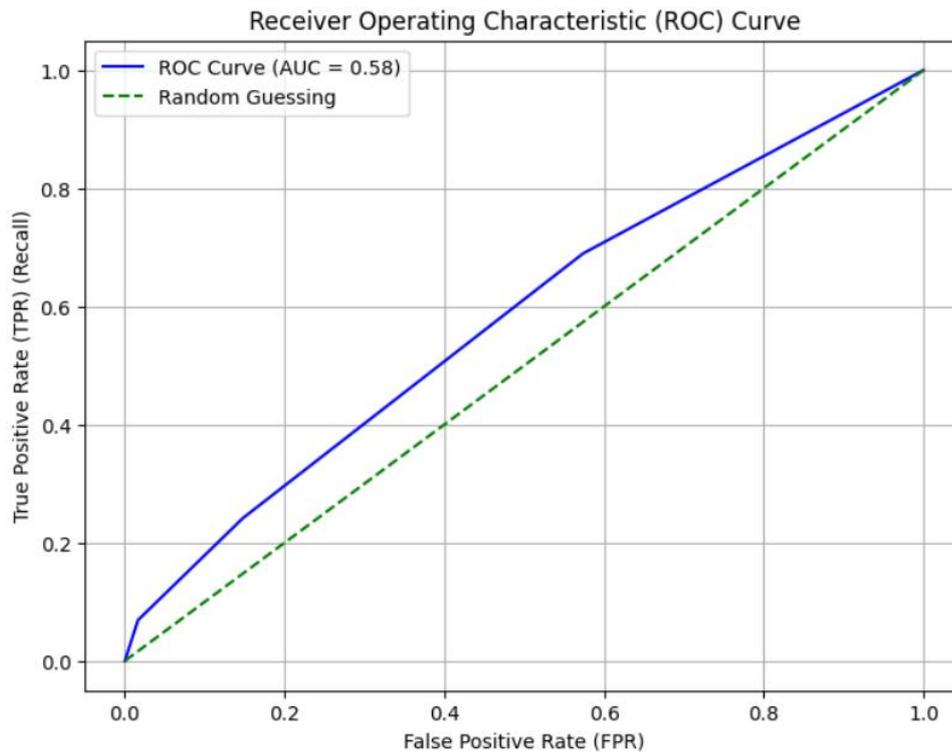
$$Recall = \frac{7}{7+22} = 0.24$$

$$F1 - Score = \frac{2 \times 0.44 \times 0.24}{0.44 + 0.24} = 0.31$$

Tabel 4. 9 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<b><i>Accuracy</i></b>	<b><i>Precision</i></b>	<b><i>Recall</i></b>	<b><i>F1-score</i></b>
0.66	0.44	0.24	0.31

Pengujian tanpa seleksi fitur dan *SMOTE* pada model B ini didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.3 dibawah.



Gambar 4.3 *Curva ROC Model B tanpa seleksi fitur dan SMOTE*

Gambar 4.3 menunjukkan *AUC* sebesar 0.58. *AUC* yang berada di atas menunjukkan bahwa model memiliki kemampuan yang cukup baik dalam membedakan antara kelas positif dan negatif, tetapi belum mencapai kinerja yang sangat baik karena jauh dari angka 1.

#### 4.1.2.2 Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian kedua ini yaitu menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai  $k$  atau  $n\_neighbor$  terbaik.

Tabel 4. 10 Hasil rata-rata skor pengujian dengan seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.809635	{'n_neighbors': 2}
0.863294	{'n_neighbors': 3}
0.842915	{'n_neighbors': 4}
0.867453	{'n_neighbors': 5}
0.839974	{'n_neighbors': 6}
0.869057	{'n_neighbors': 7}
0.868080	{'n_neighbors': 8}
0.866001	{'n_neighbors': 9}
0.881649	{'n_neighbors': 10}
0.881839	{'n_neighbors': 11}

Tabel 4.10 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 11$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.881839.

Tabel 4. 11 *Confusion matrix* dengan seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
56	53	8	5

Tabel 4.11 menunjukkan bahwa model *K-Nearest Neighbor* terdapat 56 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 53 sampel yang berhasil diklasifikasi negatif dengan benar, ada 8 sampel yang salah diklasifikasi sebagai positif, dan terdapat 5 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.11 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*. Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{56+53}{56+53+8+5} = 0.89$$

$$Precision = \frac{56}{56+8} = 0.88$$

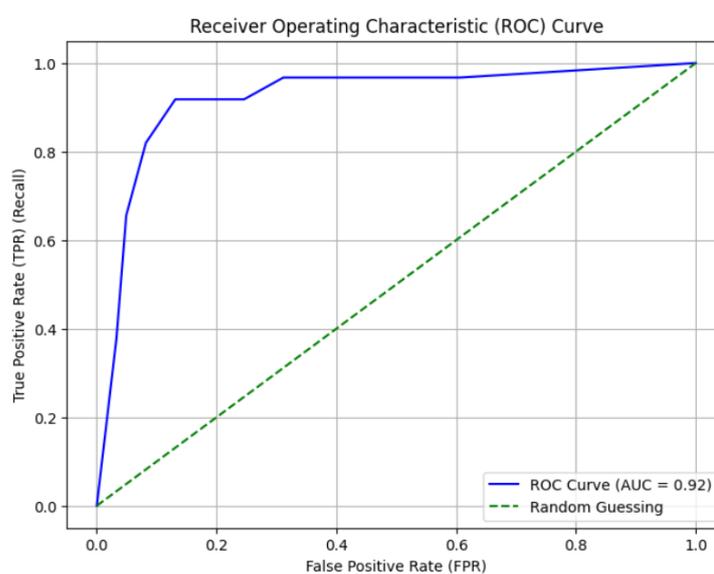
$$Recall = \frac{56}{56+5} = 0.92$$

$$F1 - Score = \frac{2 \times 0.88 \times 0.92}{0.88 + 0.92} = 0.90$$

Tabel 4. 12 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0.89	0.88	0.92	0.90

Pengujian menggunakan seleksi fitur dan *SMOTE* pada model B didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.4 dibawah.

Gambar 4. 4 *Curva ROC* Model B menggunakan seleksi fitur dan *SMOTE*

Gambar 4.4 menunjukkan hasil *AUC* sebesar 0.92. Nilai *AUC* yang mendekati 1 menunjukkan bahwa model memiliki kemampuan yang sangat tinggi dalam membedakan antara kelas positif dan negatif. Hasil di atas memiliki tingkat *True Positive Rate* (TPR) yang tinggi.

#### 4.1.3 Pengujian Model C

Model C perbandingan yang digunakan antara data latih dengan data uji sebesar 60:40. Pengujian model ini terdapat dua skenario pengujian yaitu tidak

menggunakan seleksi fitur dan *SMOTE*, serta pengujian dengan menggunakan seleksi fitur dan *SMOTE*.

#### 4.1.3.1 Tidak Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian pertama ini yaitu tidak menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai  $k$  atau  $n\_neighbor$  terbaik.

Tabel 4. 13 Hasil rata-rata skor pengujian tanpa seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.249683	{'n_neighbors': 2}
0.414870	{'n_neighbors': 3}
0.286349	{'n_neighbors': 4}
0.373889	{'n_neighbors': 5}
0.322857	{'n_neighbors': 6}
0.405397	{'n_neighbors': 7}
0.294841	{'n_neighbors': 8}
0.430317	{'n_neighbors': 9}
0.301190	{'n_neighbors': 10}
0.378175	{'n_neighbors': 11}

Sesuai tabel 4.13 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 9$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.430317.

Tabel 4. 14 *Confusion matrix* tanpa seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
10	75	6	29

Tabel 4.14 menunjukkan bahwa model *K-Nearest Neighbor* tanpa dilakukan seleksi fitur, dan *SMOTE* terdapat 10 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 75 sampel yang berhasil diklasifikasi negatif dengan benar,

ada 6 sampel yang salah diklasifikasi sebagai positif, dan terdapat 29 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.14 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*.

Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{10+75}{10+75+6+29} = 0.71$$

$$Precision = \frac{10}{10+6} = 0.62$$

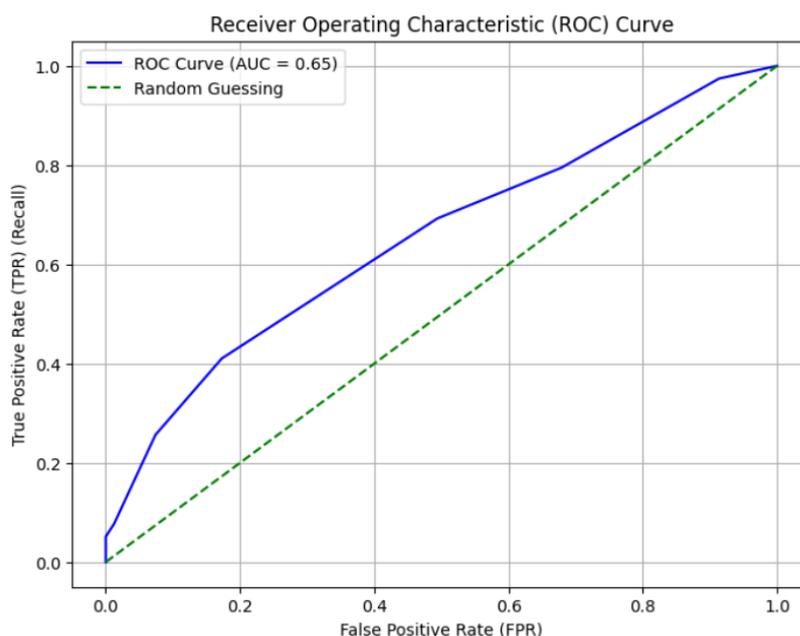
$$Recall = \frac{10}{10+29} = 0.26$$

$$F1 - Score = \frac{2 \times 0.62 \times 0.26}{0.62 + 0.26} = 0.36$$

Tabel 4. 15 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0.71	0.62	0.26	0.36

Pengujian tanpa seleksi fitur dan *SMOTE* pada model C ini didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.5 dibawah.



Gambar 4. 5 *Curva ROC* Model C tanpa seleksi fitur dan *SMOTE*

Gambar 4.5 menunjukkan hasil *AUC* 0.65. *AUC* di atas menunjukkan model memiliki kemampuan yang cukup baik untuk membedakan antara kelas positif dan negatif, tetapi belum mencapai kinerja yang sangat baik karena jauh dari angka 1.

#### 4.1.3.2 Menggunakan Seleksi Fitur dan *SMOTE*

Skenario pengujian kedua ini yaitu menggunakan seleksi fitur dan *SMOTE* dengan menggunakan penyetelan *hyperparameter*. Pengujian ini dilakukan dengan menggunakan *GridSearchCV* untuk mencari konfigurasi parameter nilai  $k$  atau  $n\_neighbor$  terbaik.

Tabel 4. 16 Hasil rata-rata skor pengujian dengan seleksi fitur dan *SMOTE*

<i>mean_test_score</i>	<i>params</i>
0.762976	{'n_neighbors': 2}
0.869284	{'n_neighbors': 3}
0.839076	{'n_neighbors': 4}
0.875669	{'n_neighbors': 5}
0.823318	{'n_neighbors': 6}
0.859506	{'n_neighbors': 7}
0.824845	{'n_neighbors': 8}
0.855720	{'n_neighbors': 9}
0.836012	{'n_neighbors': 10}
0.844570	{'n_neighbors': 11}

Tabel 4.16 menunjukkan bahwa nilai  $k$  (nilai  $n\_neighbor$ ) terbaik yang digunakan pada model *K-Nearest Neighbor* adalah  $k = 5$  dengan skor rata-rata data pengujian untuk setiap kombinasi *hyperparameter* adalah = 0.875669.

Tabel 4. 17 *Confusion matrix* dengan seleksi fitur dan *SMOTE*

<b>TP</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>
73	73	9	8

Tabel 4.17 menunjukkan bahwa model *K-Nearest Neighbor* terdapat 73 sampel yang berhasil diklasifikasi positif dengan benar, terdapat 73 sampel yang berhasil

diklasifikasi negatif dengan benar, ada 9 sampel yang salah diklasifikasi sebagai positif, dan terdapat 8 sampel yang salah diklasifikasi sebagai negatif.

Tabel 4.17 dapat diketahui nilai dari *accuracy*, *precision*, *recall* dan *f1-score*.

Perhitungan dari nilai-nilai tersebut adalah sebagai berikut.

$$Accuracy = \frac{73+73}{73+73+9+8} = 0.90$$

$$Precision = \frac{73}{73+9} = 0.89$$

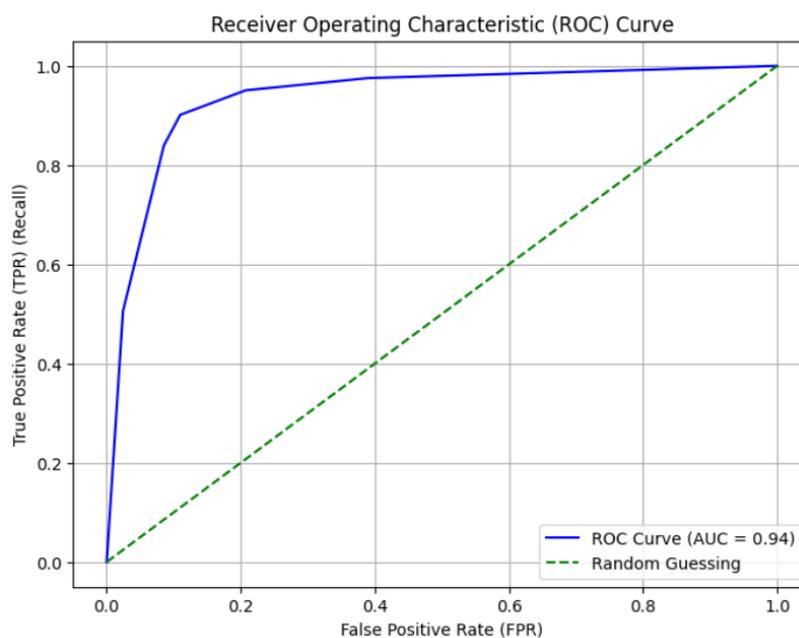
$$Recall = \frac{73}{73+8} = 0.90$$

$$F1 - Score = \frac{2 \times 0.89 \times 0.90}{0.89 + 0.90} = 0.90$$

Tabel 4. 18 Nilai *accuracy*, *precision*, *recall*, *f1-score*

<i>Accuracy</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
0.90	0.89	0.90	0.90

Pengujian menggunakan seleksi fitur dan *SMOTE* pada model C didapatkan hasil visualisasi kurva *ROC* seperti gambar 4.6 dibawah.

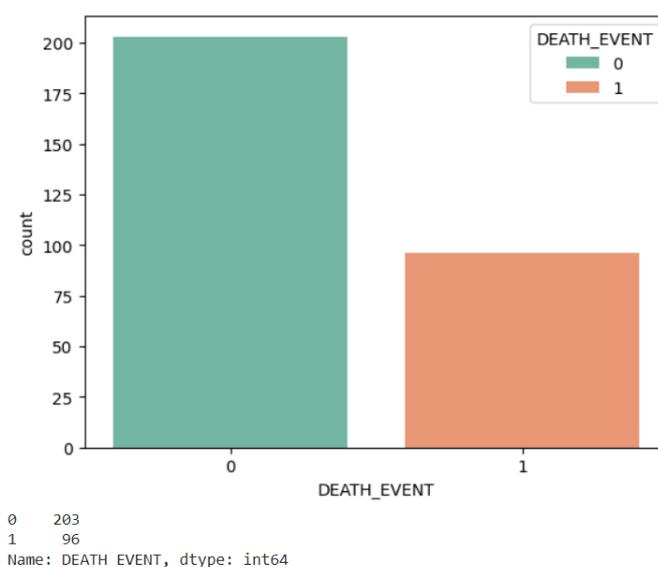


Gambar 4. 6 Curva *ROC* Model C menggunakan seleksi fitur dan *SMOTE*

Gambar 4.6 menunjukkan hasil  $AUC$  0.94.  $AUC$  di atas mendekati 1 menunjukkan bahwa model memiliki kemampuan yang sangat baik untuk membedakan antara kelas positif dan negatif. Hasil di atas memiliki tingkat *True Positive Rate* (TPR) yang tinggi.

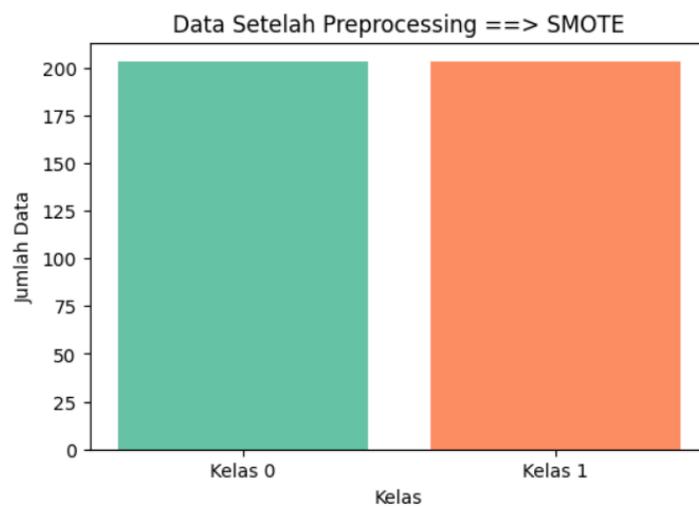
## 4.2 Pembahasan

Sub-bab ini merupakan penjelasan dari hasil uji coba untuk menentukan skenario yang paling optimal dalam melatih model *K-Nearest Neighbor* (KNN). *Dataset* yang digunakan untuk penelitian ini berasal dari *dataset* “*Heart Failure Prediction*” yang merupakan *open access* dari “*BMC Medical Informatics and Decision Making*” (Chicco & Jurman, 2020). *Dataset* ini dikumpulkan oleh Davide Chicco dan Giuseppe Jurman berdasarkan rekam medis 299 pasien gagal jantung di Punjab, Pakistan. Jumlah keseluruhan fitur yang ada pada *dataset* ini berjumlah 12 fitur. Namun, dalam penelitian ini hanya menggunakan 7 fitur yang tingkat kepentingannya tinggi.



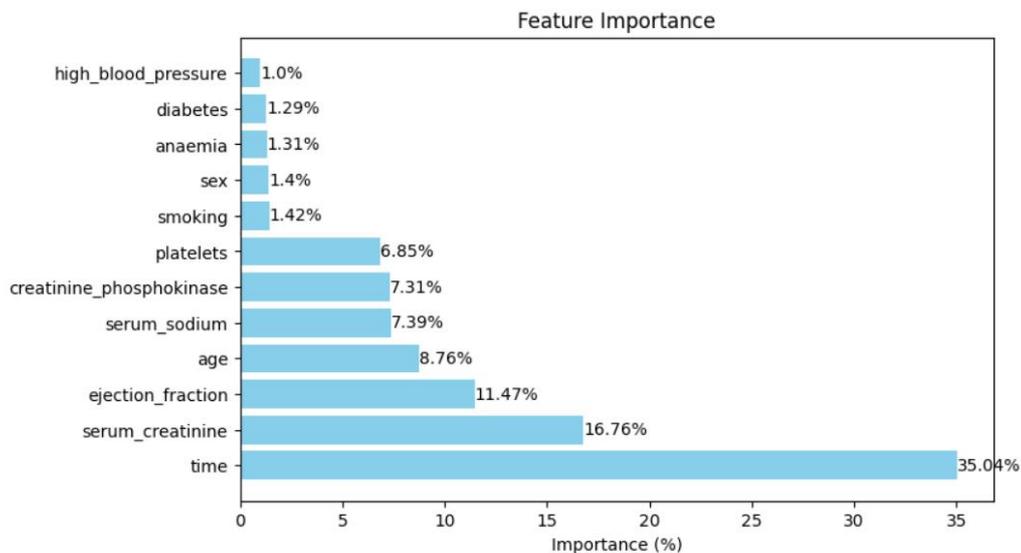
Gambar 4. 7 *Imbalance Data*

Gambar 4.7 menunjukkan bahwa kelas target pada *dataset “Heart Failure Prediction”* imbalance (data tidak seimbang). Kelas 0 (yang tidak meninggal) memiliki 203 data, sedangkan kelas 1 (yang meninggal) memiliki 96 data. Data tersebut tidak seimbang, oleh karena itu akan melakukan *Synthetic Minority Oversampling Technique (SMOTE)* untuk menangani ketidakseimbangan data.



Gambar 4. 8 *Balance Data*

Gambar 4.8 menunjukkan bahwa kelas target pada *dataset “Heart Failure Prediction”* sudah *balance* (seimbang). Data tersebut sudah melewati *Preprocessing* yaitu dengan dilakukannya *Synthetic Minority Oversampling Technique (SMOTE)*. Dapat dilihat hasilnya bahwa data tersebut sudah *balance*. Setelah dilakukannya proses *SMOTE* maka akan dilakukan seleksi fitur untuk memilih fitur yang memiliki tingkat kepentingan tinggi.



Gambar 4. 9 Tingkat Kepentingan (*Importance*) Fitur

Gambar 4.9 tercermin dalam plot yang dihasilkan, dimana setiap fitur memiliki tingkat kepentingan yang ditunjukkan *importance* setiap fitur. Setelah menganalisis tingkat kepentingan fitur, peneliti dapat memahami kontribusi relatif dari masing-masing fitur terhadap klasifikasi hasil yang bisa digunakan untuk melihat faktor-faktor yang mempengaruhi hasil klasifikasi. Fitur yang digunakan oleh peneliti yaitu di tabel 4.19.

Tabel 4. 19 Fitur yang digunakan

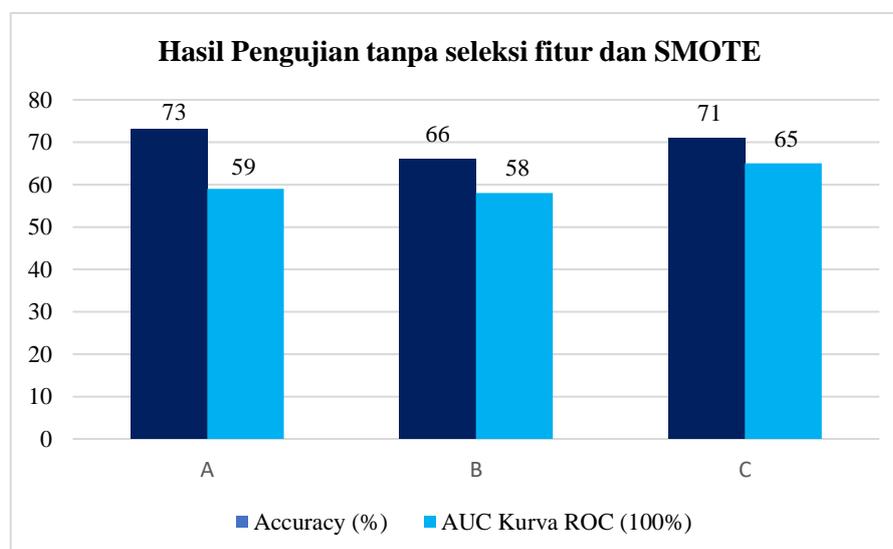
<i>Feature</i>	<i>Importance (%)</i>
<i>time</i>	35.05%
<i>serum creatinine</i>	16.76%
<i>ejection fraction</i>	11.47%
<i>age</i>	8.76%
<i>serum sodium</i>	7.39%
<i>creatinine phosphokinase</i>	7.31%
<i>platelets</i>	6.85%

Setelah proses seleksi fitur, langkah selanjutnya adalah membagi data pelatihan dan data pengujian menjadi 3 model yang sudah ditentukan. Model A perbandingan

80:20, model B dengan perbandingan 70:30, dan model C dengan perbandingan 60:40. Langkah selanjutnya adalah penyetelan *hyperparameter* dengan model *K-Nearest Neighbor* menggunakan penyetelan *hyperparameter* dengan *GridSearchCV* dan melakukan perhitungan jarak *Euclidean distance* pada pengujian skenario tidak ada seleksi fitur, *SMOTE* dan menggunakan seleksi fitur, *SMOTE* dari tiga model pembagian data. Berikut tabel hasil *accuracy* dan nilai *AUC* pada kurva *ROC* dari skenario tiap model.

Tabel 4. 20 Hasil Pengujian tanpa seleksi fitur dan *SMOTE*

Model	Rasio Data	<i>Hyperparameter</i>	<i>Accuracy</i>	<i>AUC Kurva ROC</i>
A	80:20	3	0.73	0.59
B	70:30	3	0.66	0.58
C	60:40	9	0.71	0.65



Gambar 4. 10 Hasil Pengujian Tanpa Seleksi Fitur dan *SMOTE*

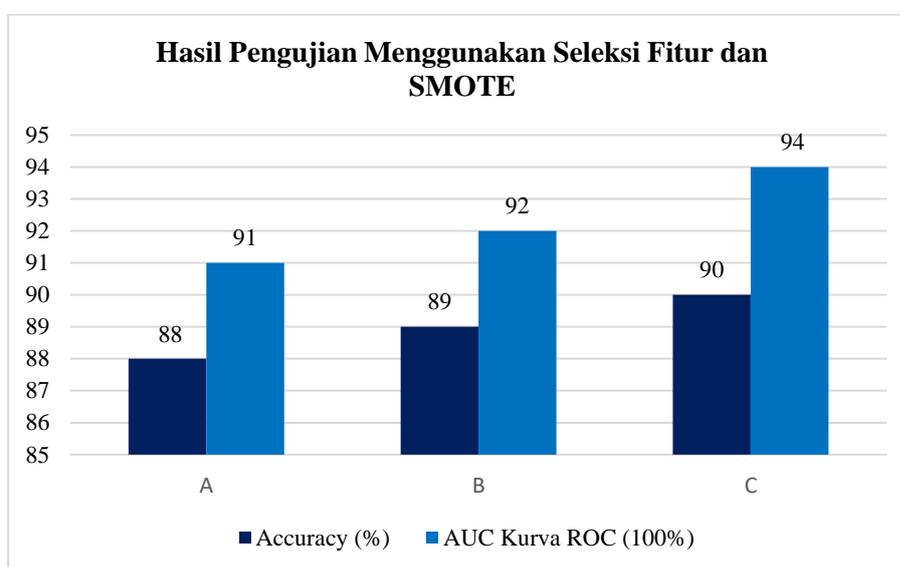
Tabel 4.20 dan Gambar 4.10 menunjukkan hasil pengujian tanpa seleksi fitur dan *SMOTE*. Hasil pengujian tersebut menunjukkan bahwa akurasi yang didapatkan kurang baik. Model A didapatkan akurasi 0.73 dan nilai *AUC* pada

kurva *ROC* 0.59. Model B didapatkan akurasi 0.66 dan nilai *AUC* kurva *ROC* 0.58.

Model C didapatkan akurasi 0.71 dan nilai *AUC* kurva *ROC* 0.65.

Tabel 4. 21 Hasil Pengujian menggunakan seleksi fitur dan *SMOTE*

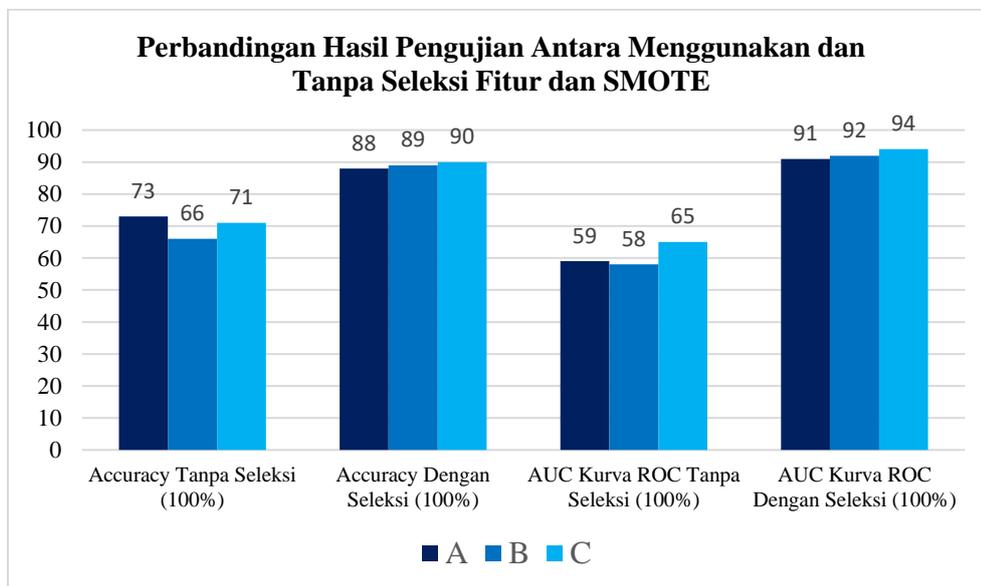
Model	Rasio Data	Hyperparameter	Accuracy	AUC Kurva ROC
A	80:20	7	0.88	0.91
B	70:30	11	0.89	0.92
C	60:40	5	0.90	0.94



Gambar 4. 11 Hasil Pengujian Menggunakan Seleksi Fitur dan *SMOTE*

Tabel 4.21 dan Gambar 4.11 menunjukkan hasil pengujian menggunakan seleksi fitur dan *SMOTE* bahwa algoritma *K-Nearest Neighbor* memberikan performa yang baik dalam melakukan klasifikasi penyakit kardiovaskular. Hasil akurasi terbaik diperoleh dari model pengujian C dengan rasio data 60:40 dan *hyperparameter GridsearchCV* 5 mendapatkan hasil akurasi 0.90. Selain nilai akurasi didapatkan hasil nilai *AUC* pada model pengujian C yaitu 0.94. Hasil tersebut dapat dikatakan sangat baik karena *AUC* model pengujian C mendekati 1 yang menunjukkan bahwa model memiliki kemampuan yang sangat baik untuk

membedakan antara kelas positif dan negatif. Hasil pengujian dengan seleksi fitur dan *SMOTE* diatas memiliki tingkat *True Positive Rate* (TPR) yang tinggi.



Gambar 4. 12 Hasil Perbandingan Dua Pengujian

Gambar 4.12 menunjukkan perbandingan hasil kedua skenario pengujian yaitu tidak menggunakan seleksi fitur dan *SMOTE* dan dengan menggunakan seleksi fitur dan menggunakan *SMOTE*. Hasilnya menunjukkan bahwa hasil lebih bagus menggunakan seleksi fitur dan *SMOTE* karena telah melewati *Preprocessing*. Dapat dilihat perbandingan model A, B dan C pada gambar diatas bahwa model C yang menggunakan seleksi fitur dan menggunakan *SMOTE* mendapatkan hasil yang tinggi dan lebih baik daripada model yang lain. Model C mendapatkan hasil akurasi 90% dan nilai *AUC* 94% setelah melewati *Preprocessing*.

Tabel 4. 22 Hasil *Precision*, *Recall*, *F1-score* setelah *Preprocessing*

Model	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>
A	0.84	0.93	0.88
B	0.88	0.92	0.90
C	0.89	0.90	0.90

Selain nilai akurasi, berdasarkan tabel 4.22 diatas yang telah melewati *Preprocessing* dapat disimpulkan bahwa Model C menunjukkan kinerja terbaik secara keseluruhan karena memiliki nilai presisi tertinggi, serta nilai *recall* dan *f1-score* yang tinggi menunjukkan keseimbangan yang baik antara presisi dan *recall*, sedangkan model A *recall* tinggi tetapi presisi dan *f1-score* tidak seimbang. Begitupun model B *recall*-nya tinggi tetapi tidak seimbang dengan presisinya.

Berikut ini adalah hasil perbandingan akurasi antara peneliti saat ini dan peneliti sebelumnya.

Tabel 4. 23 Hasil Perbandingan Akurasi

Sitasi	Metode	Akurasi Terbaik
(Fiqriansyah <i>et al.</i> , 2022)	<i>Naïve Bayes</i>	80%
(Munandar & Munir, 2022)	<i>K-Nearest Neighbor</i>	70.04%
(Putri, 2021)	<i>K-Nearest Neighbor</i>	87%
(Subarkah <i>et al.</i> , 2022)	<i>Naïve Bayes</i>	80.6%
(Chicco & Jurman, 2020).	<i>K-Nearest Neighbor, Random forests, Decision tree, Naïve bayes</i>	<i>Random forests</i> Akurasi 74% dan AUC 80%
<b>Peneliti</b>	<b><i>K-Nearest Neighbor</i></b>	<b>90% dan AUC 94%</b>

Tabel 4.23 adalah hasil dari penelitian terdahulu dan penelitian sekarang oleh peneliti. Hasil dari “Penerapan Algoritma Naïve Bayes Untuk Mengetahui Pasien Penyakit Gagal Jantung” (Fiqriansyah *et al.*, 2022) menunjukkan bahwa algoritma *naïve bayes* didapatkan hasil akurasi 80% yang menggunakan *dataset (heart disease)* dengan menggunakan 6 fitur. Hasil dari “Implementasi *K-Nearest Neighbor* Untuk *Prototype* Sistem Pakar Identifikasi Dini Penyakit Jantung” (Munandar & Munir, 2022) menunjukkan tingkat akurasi *k-nearest neighbor* sebesar 70.04% yang menggunakan *dataset* catatan medis pasien pada salah satu rumah sakit pemerintah di Provinsi Banten. Data target diperoleh sebanyak 240

record, terdiri atas 15 variabel dan 8 jenis penyakit jantung. Hasil dari “Analisis Performa Metode *K- Nearest Neighbor* (KNN) dan *Crossvalidation* pada Data Penyakit *Cardiovascular*” (Putri, 2021) menunjukkan tingkat akurasi terbaik yaitu sebesar 87%, presisi 87%, *recall* 97%, dan *f-measure* 92%, pada  $K=3$  yang menggunakan *dataset* dari UCI machine learning repository. Hasil dari “*Comparison of correlated algorithm accuracy Naive Bayes Classifier and Naive Bayes Classifier for heart failure classification*” menunjukkan tingkat akurasi 80.6% yang menggunakan *dataset Heart Failure Prediction* dengan 299 data. Hasil dari (Chicco & Jurman, 2020) menunjukkan bahwa pada jurnal ini mengklasifikasi pasien berdasarkan *serum creatinine* dan *ejection fraction*, yang menggunakan beberapa metode dan evaluasi *ROC*. *Dataset* yang digunakan adalah *Heart Failure Prediction* dengan 299 data. Dari penelitian ini dihasilkan *Random forests* memiliki akurasi 74% dan *AUC* 80%.

Peneliti mendapatkan hasil yaitu akurasi 90% dan *AUC* dengan hasil 94%. Hasil tersebut didapatkan dari beberapa proses dan evaluasi. Peneliti menggunakan *GridsearchCV* untuk mendapatkan *Hyperparameter* yang baik dan menggunakan seleksi fitur serta *SMOTE*. Peneliti merujuk pada penelitian (Subarkah *et al.*, 2022) karena pada penelitiannya tersebut menggunakan *dataset* yang sama, tetapi algoritma yang digunakan berbeda sedangkan pada penelitian (Chicco & Jurman, 2020) menggunakan *dataset* yang sama, evaluasi yang sama dan menguji beberapa metode, tetapi peneliti sekarang berbeda di fitur yang digunakan untuk pengujian.

Menurut (Subarkah *et al.*, 2022) 90%-100% merupakan kategori klasifikasi sangat baik, 80%-90% merupakan kategori klasifikasi baik, 70%-80% merupakan

kategori klasifikasi cukup baik, 60%-70% merupakan kategori klasifikasi kurang baik, dan 50%-60% merupakan kategori klasifikasi gagal (Gorunescu, 2009). Berdasarkan pernyataan tersebut maka hasil penelitian ini termasuk kategori klasifikasi sangat baik karena menghasilkan nilai akurasi 90% dan nilai *AUC* 94%.

### 4.3 Integrasi Islam

Dengan mendasarkan penelitian ini pada prinsip-prinsip agama melalui integrasi islam yaitu berdasarkan al-quran dan hadist, diharapkan mampu menekankan pentingnya kesehatan dan kesadaran akan peran Allah SWT dan Manusia. Berikut adalah integrasi yang menjelaskan Muamalah Mu'Allah dan Muamalah Mu'annas.

#### 4.3.1 Muamalah Mu'Allah

إِنَّ اللَّهَ تَعَالَى أَنْزَلَ الدَّاءَ وَالذَّوَاءَ وَجَعَلَ لِكُلِّ دَاءٍ دَوَاءً فَتَدَاوَوْا وَلَا تَدَاوَوْا بِالْحَرَامِ

*“Sesungguhnya Allah menurunkan penyakit dan obatnya dan menjadikan bagi setiap penyakit ada obatnya. Maka berobatlah kalian, dan jangan kalian berobat dengan yang haram” (HR. Dawud)*

Menurut Imam Nawawi, hadis ini menunjukkan bahwa Allah SWT memiliki kuasa penuh atas segala sesuatu, termasuk penyakit dan obatnya. Allah SWT menciptakan penyakit sebagai ujian bagi manusia, dan Dia pula yang menciptakan obatnya. Oleh karena itu, manusia wajib bersyukur kepada Allah SWT atas kesehatan yang diberikan dan berikhtiar untuk mencari kesembuhan ketika sakit.

Hadits tersebut menunjukkan Nabi menganjurkan kepada umatnya untuk segera berobat ketika sakit. Nabi tak melarang seseorang yang sakit untuk berobat.

Karena berobat merupakan suatu ikhtiar untuk segera sembuh dari penyakit. *Rasulullah Shallallahu 'Alaihi Wa Sallam* dalam sabdanya menjelaskan bahwasannya semua penyakit pasti diciptakan pula obatnya.

حَدَّثَنَا هَارُونُ بْنُ مَعْرُوفٍ وَأَبُو الطَّاهِرِ وَأَحْمَدُ بْنُ عِيسَى قَالُوا حَدَّثَنَا ابْنُ وَهْبٍ أَخْبَرَنِي عَمْرُو وَهُوَ ابْنُ الْحَارِثِ عَنْ عَبْدِ رَبِّهِ بْنِ سَعِيدٍ عَنْ أَبِي الزُّبَيْرِ عَنْ جَابِرٍ عَنْ رَسُولِ اللَّهِ صَلَّى اللَّهُ عَلَيْهِ وَسَلَّمَ أَنَّهُ قَالَ لِكُلِّ دَاءٍ دَوَاءٌ فَإِذَا أُصِيبَ وَجَلَّ عَرَّ دَوَاءُ الدَّاءِ بَرًّا بِإِذْنِ اللَّهِ

*“Telah menceritakan kepada kami Harun bin Ma'ruf dan Abu Ath Thahir serta Ahmad bin 'Isa mereka berkata; Telah menceritakan kepada kami Ibnu Wahb; Telah mengabarkan kepadaku 'Amru, yaitu Ibnu al-Harits dari 'Abdu Rabbih bin Sa'id dari Abu Az Zubair dari Jabir dari Rasulullah shallallahu 'alaihi wasallam, beliau bersabda: "Setiap penyakit ada obatnya. Apabila ditemukan obat yang tepat untuk suatu penyakit, akan sembuhlah penyakit itu dengan izin Allah 'azza wajalla." (HR Muslim Nomor 4084).*

Menurut Ibnu Qayyim al-Jauziyyah dalam kitabnya yang berjudul *AthThibban-Nabawi* hadis di atas mengandung pengabsahan terhadap adanya sebab musahab. Menurutnya setiap penyakit pasti ada lawannya, yaitu pengobatan yang menjadi kebalikan dari penyakitnya, karena segala sesuatu yang diciptakan Allah pasti ada lawannya (Hafil, 2020). Adanya sistem klasifikasi ini merupakan suatu bentuk pencegahan dini dimana merupakan suatu bentuk langkah pengobatan awal sehingga tidak terjadinya komplikasi penyakit yang lebih parah dan tidak terjadinya kematian.

#### 4.3.2 Muamalah Mu'Annas

إِنَّمَا الْمُؤْمِنُونَ إِخْوَةٌ فَأَصْلِحُوا بَيْنَ أَخَوَيْكُمْ ۗ وَاتَّقُوا اللَّهَ لَعَلَّكُمْ تُرْحَمُونَ

*“Sesungguhnya orang-orang mukmin itu bersaudara, karena itu damaikanlah kedua saudaramu (yang bertikai) dan bertakwalah kepada Allah agar kamu dirahmati.” (QS. Al-Hujurat:10)*

Tafsir Ibnu Katsir pada ayat ini menjelaskan bahwa orang-orang beriman itu bersaudara seakan-akan mereka satu keluarga dalam hal iman dan Islam. Mereka harus saling tolong-menolong, saling mengasihi, dan saling membela. Oleh karena itu, Allah SWT memerintahkan kepada mereka untuk mendamaikan dua orang yang berselisih di antara mereka. Hendaklah mereka berusaha dengan penuh kesabaran dan kebijaksanaan untuk menyelesaikan perselisihan itu dengan cara yang adil dan damai. Allah SWT memerintahkan mereka untuk bertakwa kepada Allah SWT dalam segala perbuatan mereka, termasuk dalam mendamaikan dua orang yang berselisih. Dengan demikian, mereka akan mendapatkan rahmat dari Allah SWT, yaitu pahala yang besar dan kebahagiaan yang sempurna.

Sebagai ilustrasi dari sikap saling membantu dalam kebaikan dan ketakwaan, Rasulullah Shallallahu ‘alaihi wa sallam bersabda:

انصُرْ أَهْلَكَ ظَالِمًا أَوْ مَظْلُومًا قَالُوا يَا رَسُولَ اللَّهِ هَذَا نَنْصُرُهُ مَظْلُومًا فَكَيْفَ نَنْصُرُهُ ظَالِمًا قَالَ تَأْخُذُ فَوْقَ يَدَيْهِ

*“Bantulah saudaramu, baik dalam keadaan sedang berbuat zhalim atau sedang teraniaya. Ada yang bertanya: “Wahai Rasulullah, kami akan menolong orang yang teraniaya. Bagaimana menolong orang yang sedang berbuat zhalim?” Beliau menjawab: “Dengan menghalanginya melakukan kezhaliman. Itulah bentuk bantuanmu kepadanya.” [HR. al-Bukhâri]*

Penelitian ini berpotensi meningkatkan akurasi diagnosis dan hasil pengobatan, yang pada akhirnya berkontribusi pada kesejahteraan individu dan masyarakat secara keseluruhan. Hubungan keduanya terletak pada semangat kolaborasi dan

pengejaran tujuan bersama. Penelitian ini menunjukkan contoh nyata upaya kolaborasi para ilmuwan dalam mengatasi masalah kesehatan yang mendesak. Keduanya menunjukkan pentingnya bekerja sama untuk kemajuan umat manusia. Al-Qur'an berfungsi sebagai pengingat tentang pertimbangan etis dan moral yang harus memandu penelitian ilmiah. Kemajuan teknologi dan kedokteran memberikan manfaat luar biasa, namun harus diimbangi dengan rasa tanggung jawab dan belas kasih kepada mereka yang akan mendapat manfaat dari inovasi tersebut.

## BAB V

### KESIMPULAN DAN SARAN

#### 5.1 Kesimpulan

Berdasarkan rangkaian skenario uji coba dalam penelitian ini, terdapat tiga model yang digunakan yaitu Model A dengan pembagian data latih 80% dan data uji 20%, Model B dengan pembagian data latih 70% dan data uji 30%, serta Model C dengan pembagian data latih 60% dan data uji 40%. Selain itu, terdapat dua skenario uji: satu tanpa menggunakan SMOTE dan seleksi fitur, dan satu lagi dengan penggunaan SMOTE dan seleksi fitur.

Hasil pengujian tanpa menggunakan *SMOTE* dan seleksi fitur yaitu model A mendapatkan hasil akurasi 73% dan nilai *AUC* 59%, model B mendapatkan hasil akurasi 66% dan nilai *AUC* 58%, model C mendapatkan hasil akurasi 71% dan nilai *AUC* 65%, sedangkan hasil pengujian menggunakan *SMOTE* dan seleksi fitur yaitu model A mendapatkan hasil akurasi 88% dan nilai *AUC* 91%, model B mendapatkan hasil akurasi 89% dan nilai *AUC* 92%, model C mendapatkan hasil akurasi 90% dan nilai *AUC* 94%.

Kesimpulan dari kedua pengujian yaitu yang menggunakan *SMOTE* dan seleksi fitur mendapatkan hasil yang terbaik, yaitu pada model C mendapatkan hasil akurasi 90 % dan nilai *AUC* 94%. Selain mendapatkan hasil akurasi dan nilai *AUC*, didapatkan juga nilai *precision*, *recall*, dan *f1-score*, dapat dilihat pada tabel (4.22). Tabel tersebut menunjukkan bahwa Model C memiliki kinerja terbaik secara

keseluruhan karena memiliki nilai presisi 89%, nilai *recall* 90% dan *f1-score* 90% menunjukkan keseimbangan yang baik antara presisi, *recall* dan *f1-score*.

Kesimpulan penelitian yang dilakukan dengan menggunakan metode klasifikasi *K-Nearest Neighbor* menghasilkan nilai terbaik pada pengujian model C dengan perbandingan yaitu 60% data latih 40% data uji yaitu nilai akurasi 90%, nilai *AUC* 94%, nilai presisi 89%, nilai *recall* 90% dan *f1-score* 90% yang baik untuk klasifikasi penyakit kardiovaskular.

## 5.2 Saran

Peneliti menyadari terdapat beberapa kekurangan di dalam penelitian ini dan perlunya kritik dan saran yang membangun untuk meningkatkan penelitian selanjutnya. Perbaikan yang disarankan meliputi:

1. Dapat menggunakan metode klasifikasi yang berbeda sehingga dapat membandingkan nilai akurasi yang didapatkan agar mendapatkan metode klasifikasi yang terbaik.
2. Dapat menggunakan metode perhitungan jarak yang berbeda pada *K-Nearest Neighbor* seperti metode *Manhattan distance*, *Minkowski distance*, dan *Chebyshev distance*.
3. Dapat menggunakan proses normalisasi data yang berbeda seperti teknik *Z-Score* dan *Decimal scaling*.
4. Dapat menggunakan proses *hyperparameter* yang berbeda selain *GridSearchCV* untuk *K-Nearest Neighbor*.

## DAFTAR PUSTAKA

- Ahnan, M., Farida, I. N., & Helilintar, R. (2023). Sistem Pendukung Keputusan Kinerja Karyawan Terbaik Dengan Menggunakan Kombinasi Metode TOPSIS Dan Metode ROC. *JSITIK*, 2(1).
- Annisa, R. (2019). Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung. *Jurnal Teknik Informatika Kaputama (JTIK)*, 3(1).
- Argina, A. M. (2020). *Indonesian Journal of Data and Science Penerapan Metode Klasifikasi K-Nearest Neighbor pada Dataset Penderita Penyakit Diabetes*. 1(2), 29–33.
- Baharuddin, M. M., Azis, H., & Hasanuddin, T. (2019). Analisis performa metode k-nearest neighbor untuk identifikasi jenis kaca. *ILKOM Jurnal Ilmiah*, 11(3), 269–274. <https://doi.org/10.33096/ilkom.v11i3.489.269-274>
- Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC Medical Informatics and Decision Making*, 20(1). <https://doi.org/10.1186/s12911-020-1023-5>
- Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. <https://doi.org/10.1145/3459665>
- Cunningham, P., & Delany, S. J. (2021). K-Nearest Neighbour Classifiers-A Tutorial. In *ACM Computing Surveys* (Vol. 54, Issue 6). Association for Computing Machinery. <https://doi.org/10.1145/3459665>
- Dhany, H. W. (2021). *Seminar Nasional Informatika (SENATIKA) Prosiding SENATIKA 2021 Performa Algoritma K-Nearest Neighbour dalam Memprediksi Penyakit Jantung*. <https://www.kaggle.com/>
- Fiqriansyah, R., Akbar, F., Casanova Andiko, V., Gibran Ahmad, K., Rasywir, E., Meisak, D., Pratama, Y., & Feranika, A. (2022). *Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM) Penerapan Algoritma Naïve Bayes Untuk Mengetahui Pasien Penyakit Gagal Jantung*. <https://m5.gs/bE42Nm>
- Gorunescu. (2009). *Computational intelligenc: collaboration, fusion and emergence*. Springer.
- Kiran, S., Reddy, G. R., Girija, S. P., Venkatramulu, S., Dorthi, K., & Chandra Shekhar Rao, V. (2023). A Gradient Boosted Decision Tree with Binary Spotted Hyena Optimizer for cardiovascular disease detection and

classification. *Healthcare Analytics*, 3.  
<https://doi.org/10.1016/j.health.2023.100173>

Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel. *Jurnal Teknik Informatika Dan sisteminformasi*, 7.  
<https://doi.org/10.28932/jutisi.v7i1.3182>

Mahidinha, R. C. A., & Masithoh, S. Al. (2022). Nilai Moral Kisah Nabi Ayub dalam Al-Quran (Studi Tafsir Tematik Wahbah Zuhaili dalam Kitab Al-Munir). *Jurnal Keislaman, Pemikiran Islam Dan Living Qur'an*, 1.

Maisat, Z., Darmawan, E., Fauzan Dianta, A., & Korespondensi (2023). Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM. *Online) Teknologi: Jurnal Ilmiah Sistem Informasi*, 13(1), 8–15.  
<https://doi.org/10.26594/teknologi.v13i1.3098>

Manteqipour, M., Hadigheh, A. G., Mahmoudvand, R., & Safari, A. (2018). Grouping Objects to Homogeneous Classes Satisfying Requisite Mass. In *Journal of AI and Data Mining* (Vol. 6, Issue 1).

Mourya, S., Srinivas, P. V. S., & Seetha, M. (2018). Multi-Attributes Web Objects Classification based on Class-Attribute Relation Patterns Learning Approach. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 9, Issue 12).  
[www.ijacsa.thesai.org](http://www.ijacsa.thesai.org)

Munandar, T. A., & Munir, A. Q. (2022). Implementasi K-Nearest Neighbor Untuk Prototype Sistem Pakar Identifikasi Dini Penyakit Jantung K-Nearest Neighbor for Prototype Expert System for Early Identification of Heart Disease. *Jurnal Teknik Informatika*.

Narabel, J., & Budi, S. (2020). Deteksi Dini Status Keanggotaan Industri Kebugaran Menggunakan Pendekatan Supervised Learning. *Jurnal Teknik Informatika Dan Sistem Informasi*, 6(2).  
<https://doi.org/10.28932/jutisi.v6i2.2675>

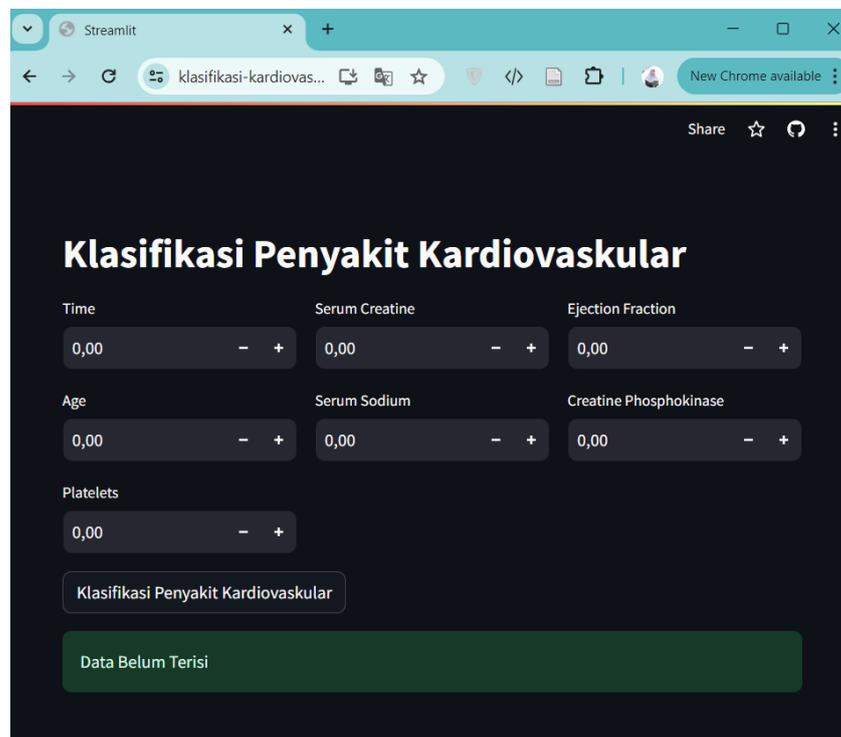
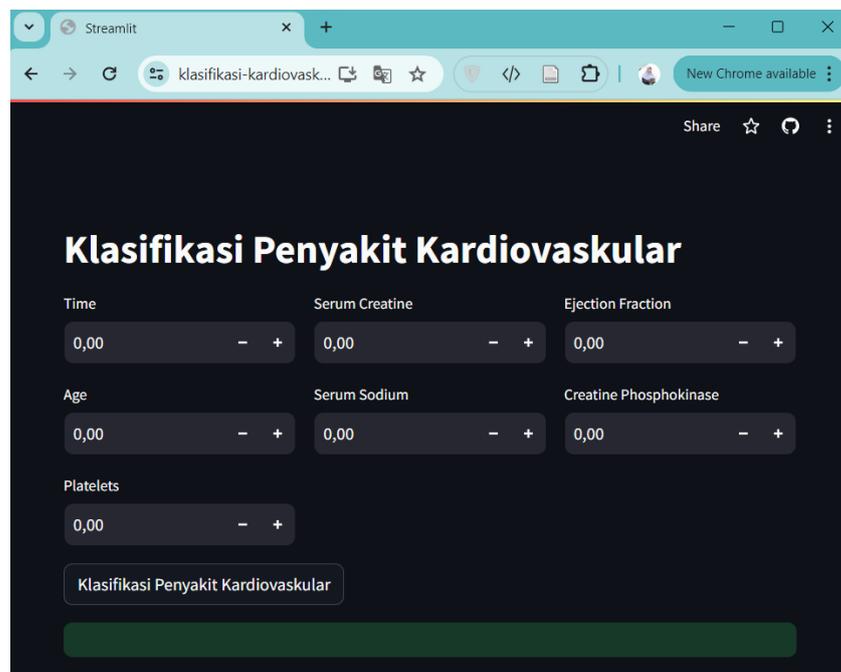
Nikmatul Kasanah, A., Pujiyanto, U., Elektro, T., Teknik, F., & Negeri Malang, U. (2021). Penerapan Teknik *SMOTE* untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN. *JURNAL RESTI (Rekayasa Sistem Dan Teknologi Informasi)*, 1(3), 196–201.

Novitasari, H. B., Hadiano, N., Sfenrianto, Rahmawati, A., Prasetyo, R., Miharja, J., & Gata, W. (2019). K-nearest neighbor analysis to predict the accuracy of product delivery using administration of raw material model in the cosmetic industry (PT Cedefindo). *Journal of Physics: Conference Series*, 1367(1). <https://doi.org/10.1088/1742-6596/1367/1/012008>

- Nugraha, W. (2021). *Prediksi penyakit jantung cardiovascular menggunakan model algoritma klasifikasi*. <https://www.kaggle.com/andrewmvd/heart->
- Putri, I. P. (2021). Analisis Performa Metode K-Nearest Neighbor (KNN) dan Crossvalidation pada Data Penyakit Cardiovascular. *Indonesian Journal of Data and Science (IJODAS)*, 2(1), 21–28.
- Sahar. (2020). Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Classifier pada Data Set Penyakit Jantung. *Indonesian Journal of Data and Science (IJODAS)*, 1(3), 79–86.
- Samosir, A., Hasibuan, M., Justino, W. E., & Hariyono, T. (2021). *Komparasi Algoritma Random Forest, Naïve Bayes dan K-Nearest Neighbor Dalam klasifikasi Data Penyakit Jantung*.
- Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2020). A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier. *Knowledge-Based Systems*, 205. <https://doi.org/10.1016/j.knosys.2020.106270>
- Shofwan, I., & Munib, A. (2023). Pendidikan Karakter Sosial Qur’ani: Studi Tafsir Surat Al-Hujurat Ayat 11-13. *Jurnal Intelektual: Jurnal Pendidikan Dan Studi Keislaman*, 13(1), 72–84. <https://doi.org/10.33367/ji.v13i1.3500>
- Subarkah, P., Damayanti, W. R., & Permana, R. A. (2022). Comparison of Correlated Algorithm Accuracy Naive Bayes Classifier and Naive Bayes Classifier for Classification of heart failure. *ILKOM Jurnal Ilmiah*, 14(2), 120–125. <https://doi.org/10.33096/ilkom.v14i2.1148.120-125>
- Syamil, A. (2023). Makna Hadits Sampaikan Dariku Walau Satu Ayat, Ballighu ’Anni walau Ayah. 9 May 2023. <https://www.risalahislam.com/2023/05/makna-hadits-sampaikan-dariku-walau.html>
- TafsirQ. (n.d.). Tafsir Surat Al-Hujurat ayat 10-13. Retrieved March 1, 2024, from <https://tafsirq.com/5-al-maidah/ayat-2#tafsir-quraish-shihab>
- Widodo Romy Budhi. (2022). 1.1.\_Romy\_MONOGRAF\_Machine\_Learning. *Media Nusa Creative*.
- Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for kNN Classification. *ACM Transactions on Intelligent Systems and Technology*, 8(3). <https://doi.org/10.1145/2990508>

# LAMPIRAN

## Lampiran 1: Hasil Deploy



Streamlit x +

klasifikasi-kardiovas... ☆ New Chrome available

Share ☆ ↻ ⋮

## Klasifikasi Penyakit Kardiovaskular

Time	Serum Creatine	Ejection Fraction
0,97 - +	0,49 - +	0,51 - +
Age	Serum Sodium	Creatine Phosphokinase
0,48 - +	0,77 - +	0,27 - +
Platelets		
0,62 - +		

Klasifikasi Penyakit Kardiovaskular

Pasien Tidak Meninggal

Streamlit x +

klasifikasi-kardiovas... ☆ New Chrome available

Share ☆ ↻ ⋮

## Klasifikasi Penyakit Kardiovaskular

Time	Serum Creatine	Ejection Fraction
0,03 - +	0,49 - +	0,51 - +
Age	Serum Sodium	Creatine Phosphokinase
0,34 - +	0,51 - +	1,00 - +
Platelets		
0,43 - +		

Klasifikasi Penyakit Kardiovaskular

Pasien Meninggal