**Pengembangan Model Prediksi Penyakit Jantung dengan Metode SMOTE-IPF dan SVM**

**SKRIPSI**

**Disusun untuk memenuhi sebagian persyaratan**

**mencapai derajat Sarjana**

****

**Disusun Oleh:**

Arddian Darmawan

2200018157

**PROGRAM STUDI S1 INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS AHMAD DAHLAN**

**2025**

# HALAMAN JUDUL

**JUDUL SKRIPSI (DITULIS SECARA SINGKAT DAN JELAS MAKSIMAL 20 KATA)**

**SKRIPSI**



**Disusun Oleh:**

ARDDIAN DARMAWAN

2200018157

**PROGRAM STUDI S1 INFORMATIKA**

**FAKULTAS TEKNOLOGI INDUSTRI**

**UNIVERSITAS AHMAD DAHLAN**

**2025**

# LEMBAR PERSETUJUAN PEMBIMBING

**SKRIPSI**

**JUDUL SKRIPSI (DITULIS SECARA SINGKAT DAN JELAS MAKSIMAL 20 KATA)**

Dipersiapkan dan disusun oleh:

**ARDDIAN DARMAWAN**

**2200018157**

**Program Studi S1 Informatika**

**Fakultas Teknologi Industri**

**Universitas Ahmad Dahlan**

**Telah disetujui oleh:**

**Pembimbing**

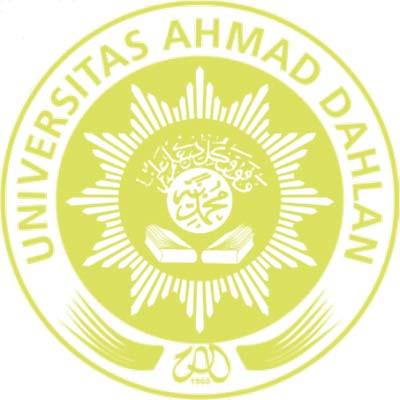
**Nama Pembimbing dengan Gelar**

**NIP/NIPM.**

# LEMBAR PENGESAHAN

**SKRIPSI**

**JUDUL SKRIPSI (DITULIS SECARA SINGKAT DAN JELAS MAKSIMAL 20 KATA)**

Dipersiapkan dan disusun oleh:

**NAMA MAHASISWA**

**NIM**

**Telah dipertahankan di depan Dewan Penguji**

**pada tanggal Hari Bulan Tahun**

**dan dinyatakan telah memenuhi syarat**

**Susunan Dewan Penguji**

**Ketua : Nama Ketua Penguji dengan Gelar …………………………………**

**Penguji 1 : Nama Penguji 1 dengan Gelar …………………………………**

**Penguji 2 : Nama Penguji 2 dengan Gelar …………………………………**

**Yogyakarta, Hari Bulan Tahun**

**Dekan Fakultas Teknologi Industri**

**Universitas Ahmad Dahlan**

**Nama Dekan FTI dengan Gelar**

**NIP/NIPM.**

# LEMBAR PERNYATAAN KEASLIAN

**SURAT PERNYATAAN**

Yang bertanda tangan di bawah ini:

Nama : ………………………………………………………………………………………………

NIM : ………………………………………………………………………………………………

Prodi : Informatika

Judul TA/Skripsi : ………………………………………………………………………………………………

………………………………………………………………………………………………

………………………………………………………………………………………………

Dengan ini saya menyatakan bahwa Laporan Tugas Akhir ini tidak terdapat karya yang pernah diajukan untuk memperoleh gelar Ahli Madya/Kesarjanaan di suatu Perguruan Tinggi, dan sepanjang pengetahuan saya juga tidak terdapat karya atau pendapat yang pernah ditulis atau diterbitkan oleh orang lain, kecuali yang secara tertulis diacu dalam naskah ini dan disebutkan dalam daftar pustaka.

Yogyakarta, Hari Bulan Tahun

|  |  |
| --- | --- |
| Mengetahui,  Dosen Pembimbing  Nama Dosen Pembimbing dengan Gelar  NIP/NIPM. | Yang menyatakan,  Nama Mahasiswa  NIM |

# KATA PENGANTAR

Tuliskan kata pengantar disini.

# DAFTAR ISI

Contents

[HALAMAN JUDUL 1](#_Toc208346278)

[LEMBAR PERSETUJUAN PEMBIMBING 2](#_Toc208346279)

[LEMBAR PENGESAHAN 3](#_Toc208346280)

[LEMBAR PERNYATAAN KEASLIAN 4](#_Toc208346281)

[KATA PENGANTAR 5](#_Toc208346282)

[DAFTAR ISI 6](#_Toc208346283)

[DAFTAR GAMBAR 8](#_Toc208346284)

[DAFTAR TABEL (Jika Ada) 9](#_Toc208346285)

[DAFTAR KODE PROGRAM (Jika Ada) 10](#_Toc208346286)

[DAFTAR LAMPIRAN 11](#_Toc208346287)

[DAFTAR SINGKATAN DAN ARTI LAMBANG (Jika Diperlukan) 12](#_Toc208346288)

[ABSTRAK 13](#_Toc208346289)

[BAB I. Pendahuluan 1](#_Toc208346290)

[1.1. Latar Belakang Masalah 1](#_Toc208346291)

[1.2. Batasan Masalah Penelitian 3](#_Toc208346292)

[1.3. Rumusan Masalah 5](#_Toc208346293)

[1.4. Tujuan Penelitian 6](#_Toc208346294)

[1.5. Manfaat Penelitian 6](#_Toc208346295)

[BAB II. Tinjauan Pustaka 8](#_Toc208346296)

[2.1. Kajian Penelitian Terdahulu 8](#_Toc208346297)

[2.2. Landasan Teori 17](#_Toc208346298)

[2.2.1. Machine Learning 17](#_Toc208346299)

[2.2.2. Support Vector Machine (SVM) 19](#_Toc208346300)

[2.2.3. Ketidakseimbangan Data (*Class* *Imbalance*) 21](#_Toc208346301)

[2.2.4. Metode Penanganan Data Tidak Seimbang 22](#_Toc208346302)

[2.2.5. Synthetic Minority Oversampling Technique (SMOTE) 22](#_Toc208346303)

[2.2.6. SMOTE-IPF (SMOTE with Iterative Partitioning Filter) 23](#_Toc208346304)

[2.2.7. Metrik Evaluasi Kinerja Model 25](#_Toc208346305)

[BAB III. METODOLOGI PENELITIAN 28](#_Toc208346306)

[3.1. Pengumpulan Data 28](#_Toc208346307)

[3.2. *Software* dan *Hardware* 29](#_Toc208346308)

[3.2.1. Software 29](#_Toc208346309)

[3.2.2. Hardware 29](#_Toc208346310)

[3.3. Tahapan Penelitian 31](#_Toc208346311)

[3.3.1. Explorasi Data 31](#_Toc208346312)

[3.3.2. Analisis Masalah Data Imbalance 32](#_Toc208346313)

[3.3.3. Analisis Model 33](#_Toc208346314)

[3.4. Evaluasi Model 34](#_Toc208346315)

[BAB IV. HASIL DAN PEMBAHASAN 37](#_Toc208346316)

[4.1. Contoh Sub Bab 37](#_Toc208346317)

[4.1.1. Contoh Sub Sub Bab 37](#_Toc208346318)

[4.2. Contoh Sub Bab 38](#_Toc208346319)

[BAB V. KESIMPULAN DAN SARAN 39](#_Toc208346320)

[5.1. Kesimpulan 39](#_Toc208346321)

[5.2. Saran 39](#_Toc208346322)

[DAFTAR PUSTAKA 40](#_Toc208346323)

[LAMPIRAN 45](#_Toc208346324)

[Lampiran 1. Bukti Pernyataan Diterima (ACCEPTED) 45](#_Toc208346325)

[Lampiran 2. Bukti Terakreditasi atau Terindeks SCOPUS 45](#_Toc208346326)

[Lampiran 3. Bukti Bayar (untuk jurnal yang berbayar) 45](#_Toc208346327)

# DAFTAR GAMBAR

# DAFTAR TABEL (Jika Ada)

[Tabel 2.1. Perbandingan beberapa penelitian terdahulu 4](#_heading=h.yhwi9cxo87ff)

[Tabel 2.2. Perbandingan beberapa penelitian terdahulu mengenai klasifikasi 5](#_heading=h.2wm8bkse6d6a)

# DAFTAR KODE PROGRAM (Jika Ada)

[Kode Program 4.1. Menampilkan tulisan “Hello World!” 9](#_heading=h.poodahthp3ux)

# DAFTAR LAMPIRAN

[LAMPIRAN 13](#_heading=h.sqtc21c2vku3)

[Lampiran 1. Bukti Pernyataan Diterima (ACCEPTED) 13](#_heading=h.9boewg7hi8be)

[Lampiran 2. Bukti Terakreditasi atau Terindeks SCOPUS 13](#_heading=h.gaoo8c3xwrgj)

[Lampiran 3. Bukti Bayar (untuk jurnal yang berbayar) 13](#_heading=h.dnxve5ypnnnu)

# DAFTAR SINGKATAN DAN ARTI LAMBANG (Jika Diperlukan)

FTI : Fakultas Teknologi Industri

UAD : Universitas Ahmad Dahlan

dst

# ABSTRAK

Penulisan abstrak sebagai berikut: 1) Menggunakan jarak antar baris 1 spasi, 2) Jenis Huruf ; Calibri dengan ukuran 11, 3) Terdiri dari Tiga alinea/paragraph.

Abstrak berisi uraian singkat tetapi lengkap yang terdiri dari 250-300 kata dan memberikan gambaran menyeluruh tentang isi skripsi. Abstrak terdiri dari 3 alinia, dengan cakupan: alinia 1 berisi tentang masalah, tujuan dan manfaat penelitian. Alinia 2 berisi tentang metode penelitian yang mengacu pada bab 3. Alinia 3 berisi tentang temuan-temuan sebagai simpulan dari pengujian yang dilalukan yang didukung dengan data-data kuantitatif hasil pengujian.

Pada akhir penulisan abstrak disertai dengan kata kunci 3 sampai 5 kata kunci diurutkan berdasarkan abjad. Kata kunci merupakan kata yang menjadi fokus dalam penelitian yang biasanya diambil dari judul Skripsi. Tiap kata kunci dipisahkan dengan tanda baca titik koma (;).

Kata kunci: kata kunci 1; kata kunci 2; kata kunci 3; kata kunci 4; kata kunci 5

# Pendahuluan

## Latar Belakang Masalah

Penyakit jantung merupakan salah satu penyebab utama kematian di dunia. Berdasarkan data dari *World Health Organization* (WHO), penyakit kardiovaskular, termasuk penyakit jantung, menyebabkan sekitar 17,9 juta kematian setiap tahun[1], menjadikannya penyebab kematian terbesar di dunia. Di Indonesia sendiri, data dari Kementerian Kesehatan RI menunjukkan bahwa penyakit jantung merupakan salah satu penyebab kematian tertinggi, dengan angka prevalensi yang terus meningkat setiap tahunnya[2]. Oleh karena itu, deteksi dini terhadap risiko penyakit jantung menjadi sangat penting untuk menekan angka kematian dan meningkatkan kualitas hidup pasien.

Secara medis, penyakit jantung dapat dideteksi melalui berbagai metode, seperti *elektrokardiogram* (EKG/ECG), *ekokardiografi*, tes darah (*biomarker* seperti *troponin* dan *kolestrol*), tes *treadmill* (*stress* *test*), hingga pencitraan seperti CT scan dan MRI jantung. Namun, metode-metode ini sering kali membutuhkan biaya tinggi, waktu yang lama, serta tenaga medis yang profesional untuk menginterpretasi hasilnya. Oleh karena itu, dibutuhkan suatu mekanisme yang lebih cepat dan lebih murah dalam melakukan prediksi penyakit jantung. Pendekatan berbasis *Machine Learning* menjadi alternatif yang dapat membantu analisis faktor resiko penyakit jantung secara cepat dan efisien.

Salah satu dataset yang banyak di gunakan dalam penelitian prediksi penyakit jantung adalah *Framingham Heart Study* Dataset[3]. Dataset ini berisi berbagai variabel kesehatan yang dapat digunakan untuk memprediksi risiko seseorang terkena penyakit jantung dalam 10 tahun mendatang. Namun, permasalahan utama dalam dataset ini adalah adanya ketidakseimbangan kelas (*imbalanced data*), dimana jumlah individu yang mengalami penyakit jantung jauh lebih sedikit dibandingkan individu yang tidak mengalami penyakit tersebut. Dari total data yang tersedia, terdapat 644 individu yang mengalami penyakit jantung dan 3.596 individu yang tidak mengalaminya. Ketidakseimbangan ini menyebabkan model klasifikasi cenderung lebih akurat dalam memprediksi kelompok mayoritas (tidak sakit jantung), tetapi memiliki performa rendah dalam mendeteksi kelompok minoritas (sakit jantung).

Untuk mengatasi permasalahan ketidakseimbangan *dataset*, metode *Synthetic Minority Over-sampling Technique* (SMOTE) telah banyak digunakan. Metode SMOTE merupakan teknik pemodelan ulang data dengan cara membuat sampel *sintetis* dari kelas minoritas melalui *interpolasi* antara contoh-contoh yang ada[4]. Dengan demikian, SMOTE membantu meningkatkan performa model klasifikasi dengan menciptakan *dataset* yang lebih seimbang, tanpa hanya melakukan duplikasi data, sehingga mengurangi risiko terjad inya *overfitting[5]*.

Penelitian ini akan menerapkan teknik SMOTE *Improved Preserving Filtering* (SMOTE-IPF). Teknik ini tidak hanya menambahkan sampel sintetis ke kelas minoritas, tetapi juga melakukan *filtering* untuk menghindari penambahan data yang berpotensi menjadi *noise*[6], sehingga dapat menghasilkan data yang lebih berkuliatas dan seimbang. Setelah proses penyeimbangan data, model *Support Vector Machine (*SVM*)* akan digunakan sebagai metode klasifikasi untuk memprediksi risiko penyakit jantung berdasarkan fitur yang tersedia dalam dataset[7].

SVM telah terbukti memiliki performa yang sangat baik dalam klasifikasi kasus-kasus di bidang kesehatan[8]. Metode ini telah banyak diterapkan dalam berbagai studi, seperti pada klasifikasi diagnosis kanker payudara menggunakan *Wisconsin Breast Cancer Dataset[9]*, prediksi penyakit diabetes dengan memanfaatkan *Pima Indian Diabetes Dataset[10]*, selain itu, SVM juga digunakan untuk mendeteksi *aritmia jantung* melalui sinyal *electrocardiogram* (ECG)[11], dan dalam memprediksi penyakit Stroke[12]. Dengan berbagai penerapan tersebut, SVM menunjukkan kapabilitas yang kuat dalam menangani data kesehatan yang kompleks dan bervariasi. Oleh karena itu, pada penelitian ini SVM akan digunakan untuk memprediksi risiko penyakit jantung pada Framingham Heart Study Dataset.

## Batasan Masalah Penelitian

Penelitian ini difokuskan pada permasalahan ketidakseimbangan data (*imbalanced data*) dalam proses prediksi risiko penyakit jantung menggunakan Framingham Heart Study Dataset. Untuk menghindari ruang lingkup yang terlalu luas dan agar penelitian dapat dilakukan secara fokus dan mendalam, maka ditetapkan batasan masalah sebagai berikut:

1. Penelitian ini hanya membahas permasalahan ketidakseimbangan data yang terdapat pada Framingham Heart Study Dataset dan dampaknya terhadap performa model klasifikasi.
2. Fokus penelitian diarahkan pada upaya penanganan imbalanced data menggunakan metode SMOTE-IPFsebagai solusi untuk meningkatkan akurasi klasifikasi risiko penyakit jantung.
3. Model klasifikasi yang digunakan dalam penelitian ini dibatasi pada SVM*.*
4. Penelitian ini dilakukan menggunakan data sekunder dari Framingham Heart Study yang telah tersedia secara publik, tanpa melibatkan data primer atau pengambilan data baru.
5. Wilayah kajian bersifat umum/global sesuai cakupan data Framingham, dan tidak membatasi pada suatu negara atau wilayah tertentu.
6. Penelitian dilakukan berdasarkan data yang tersedia dalam dataset tanpa mempertimbangkan perkembangan data terbaru setelah dataset dirilis.

## Rumusan Masalah

Permasalahan dalam penelitian ini berfokus pada ketidakseimbangan data (imbalanced data) dalam Framingham Heart Study Dataset yang dapat mempengaruhi akurasi dalam prediksi risiko penyakit jantung. Berdasarkan hal tersebut, maka rumusan masalah dalam penelitian ini adalah:

1. Bagaimana cara mengatasi permasalahan ketidakseimbangan data pada Framingham Heart Study Dataset agar hasil klasifikasi lebih optimal
2. Bagaimana penerapan metode SMOTE-IPF dalam menangani data tidak seimbang pada dataset penyakit jantung
3. Bagaimana perfoma algoritma SVM dalam mengklasifikasikan risiko penyakit jantung setelah dilakukan penyeimbangan data dengan metode SMOTE-IPF

## Tujuan Penelitian

Penelitian ini bertujuan untuk memberikan solusi terhadap permasalahan ketidakseimbangan data dalam proses prediksi risiko penyakit jantung menggunakan pendekatan keilmuan di bidang Informatika, khususnya dalam data mining dan machine learning. Adapun tujuan dari penelitian ini adalah:

1. Mengatasi permasalahan ketidakseimbangan data pada Framingham Heart Study Dataset dengan menerapkan teknik SMOTE-IPF
2. Menerapkan metode SVM untuk mengklasifikasikan risiko penyakit jantung berdasarkan data yang telah diseimbangkan
3. Mengevaluasi performa klasifikasi SVM setelah dilakukan penyeimbangan data, dengan menggunakan *metrik evaluasi* seperti *akurasi, precision, recall, dan F1-score*.

## Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat sebagai berikut:

1. Bagi akademisi dan peneliti:

Penelitian ini diharapkan dapat berkontribusi sebagai referensi dalam pengembangan metode penanganan ketidakseimbangan data pada kasus klasifikasi, khususnya dengan penerapan metode SMOTE-IPF dan *algoritma* SVM. Hal ini berkontribusi terhadap peningkatan pengetahuan di bidang *data mining* dan *klasifikasi prediktif.*

1. Bagi praktisi kesehatan dan *data analyst* di bidang medis:

Hasil penelitian ini dapat memberikan gambaran mengenai pentingnya kualitas data dalam proses prediksi risiko penyakit jantung serta menunjukkan efektivitas penggunaan teknik penyeimbangan data untuk meningkatkan akurasi sistem prediksi berbasis *machine learning*.

1. Bagi pengembang sistem prediksi medis:

Penelitian ini dapat menjadi acuan dalam pengembangan sistem pendukung keputusan (*decision support system*) di bidang kesehatan, khususnya dalam membangun model prediksi penyakit jantung yang lebih akurat dan adil dalam menangani data dengan distribusi tidak seimbang.

# Tinjauan Pustaka



## Kajian Penelitian Terdahulu

Untuk menangani masalah data tidak seimbang, berbagai varian SMOTE telah dikembangkan. penelitian [13] secara spesifik bertujuan untuk membandingkan pengaruh dua metode *oversampling*, yaitu SMOTE-IPF dan SMOTE Borderline, terhadap performa klasifikasi. Penelitian tersebut menguji kedua metode pada sembilan dataset tidak seimbang yang berbeda, termasuk dataset Abalone 19. Metode klasifikasi yang digunakan adalah SVM dan *Naive Bayes* sebagai *base classifier*. Hasil eksperimen menunjukkan bahwa proses *oversampling* secara konsisten berhasil meningkatkan performa model. Sebuah temuan dari penelitian tersebut adalah metode SMOTE Borderline menunjukkan hasil yang lebih unggul dibandingkan SMOTE-IPF pada kasus dataset Abalone 19.

Sebuah studi lain[14] bertujuan untuk meningkatkan prediksi kelangsungan hidup pasien penyakit jantung menggunakan SMOTE. Penelitian ini menggunakan dataset *Heart Failure Clinical Records* dan menguji delapan model klasifikasi yang berbeda, termasuk *Decision Tree* (DT) dan *Random Forest* (RF). Hasil studi menunjukkan bahwa kombinasi *Random Forest* dengan SMOTE memberikan performa terbaik dengan akurasi mencapai 95,12%.

Efektivitas SMOTE juga dibuktikan dalam sebuah penelitian yang berfokus pada klasifikasi penyakit gagal jantung [15]. Penelitian tersebut menerapkan metode SMOTE untuk meningkatkan performa dari beberapa *algoritma klasifikasi*. Dataset yang digunakan adalah data penyakit gagal jantung. Hasil penelitiannya menunjukkan bahwa model dengan kinerja terbaik adalah *Random Forest* yang dipadukan dengan SMOTE, yang berhasil mencapai nilai akurasi sebesar 88,1% dan AUC (*Area Under the Curve*) sebesar 94,7%.

Dukungan lebih lanjut terhadap efektivitas penanganan data tidak seimbang datang dari penelitian [5]. Penelitian tersebut secara spesifik mengevaluasi dampak SMOTE pada pengklasifikasi *Random Forest* untuk prediksi penyakit jantung. Keunggulan dari studi ini adalah penggunaan beragam metrik evaluasi yang cocok untuk data tidak seimbang, seperti *Sensitivity*, *G-Mean*, dan *Youden's Index*. Hasilnya membuktikan bahwa SMOTE tidak hanya mengurangi *overfitting*, tetapi juga meningkatkan kinerja model secara signifikan di semua indikator, dengan peningkatan skor *Accuracy* sebesar 3,45% dan *Sensitivity* sebesar 7,1%.

Selain pendekatan di tingkat data, masalah ketidakseimbangan kelas juga ditangani di tingkat *algoritma*. Salah satu penelitian[16] mengusulkan pendekatan *cost-sensitive learning* untuk meningkatkan deteksi penyakit jantung. Kerangka kerja yang diusulkan terdiri dari *Supervised Infinite Feature Selection* (Inf-FSs) untuk seleksi fitur, dan *Improved Weighted Random Forest* (IWRF) sebagai model klasifikasi. Metode IWRF bekerja dengan memberikan bobot penalti yang lebih tinggi pada kesalahan klasifikasi kelas minoritas. Dengan menggunakan dataset Statlog dan *Heart Disease Clinical Records*, hasil penelitian menunjukkan bahwa model IWRF yang dioptimalkan mampu mengungguli berbagai model *machine learning* lainnya, termasuk dalam perbandingan langsung dengan SMOTE-RF.

Di sisi lain, terdapat studi kritis yang menginvestigasi validitas dari teknik *oversampling* itu sendiri [17]. Penelitian tersebut tidak mengusulkan metode klasifikasi baru, melainkan sebuah sistem untuk memvalidasi apakah sampel *sintetis* yang dihasilkan oleh metode *oversampling* benar-benar merepresentasikan kelas *minoritas*. Dengan menggunakan tiga dataset medis, termasuk Framingham, studi tersebut menemukan bahwa semua teknik *oversampling* yang diuji—termasuk SMOTE dan SMOTE-IPF—cenderung menghasilkan sampel-sampel keliru yang lebih dekat dengan karakteristik kelas mayoritas. Berdasarkan temuan ini, disimpulkan bahwa *oversampling* adalah praktik yang berisiko dan harus dipertimbangkan dengan sangat hati-hati karena dapat menyebabkan diagnosis yang salah.

Namun Penelitian [18] melakukan evaluasi terhadap beberapa prediksi terbaik pada kasus penyakit jantung. Penelitian tersebut menggunakan dataset penyakit jantung dari *Centers for Disease Control and Prevention* (CDC) yang berisi lebih dari 300.000 data. Untuk menangani masalah ketidakseimbangan kelas, penelitian ini menerapkan dua teknik *oversampling*, yaitu SMOTE dan ADASYN. Model klasifikasi yang diuji mencakup *Random Forest*, C45, *Support Vector Machine* (SVM), dan *Logistic Regression*. Hasil penelitian menunjukkan bahwa kombinasi Random Forest denganSMOTE mencapai kinerja terbaik dengan akurasi 94,54%. Namun, sebuah temuan penting dari studi ini adalah penerapan SMOTE dan ADASYN justru menurunkan performa akurasi pada model SVM secara signifikan.

Namun pada penelitian[19] *oversampling* dan seleksi fitur terhadap kinerja model SVM menunjukan performa yang baik. Penelitian tersebut menggunakan dataset stroke dari Kaggle yang terdiri dari 5.110 data dan 12 fitur. Metode yang diuji meliputi empat teknik *oversampling* (SMOTE, Borderline-SMOTE, SMOTE-ENN, dan SMOTE-IPF) , dua metode seleksi fitur (*Information Gain* dan *Chi-Square*) , serta tiga jenis *kernel* SVM (Linear, RBF, dan Polinomial). Hasil eksperimen menunjukkan bahwa model SVM tanpa *oversampling* gagal mendeteksi kelas minoritas (*recall* 0%). Setelah penerapan *oversampling*, kombinasi Borderline-SMOTE dengan *kernel* RBF mencapai performa tertinggi, dengan akurasi 96,86% dan *F1-score* 96,79%. Selain itu, kombinasi SMOTE-IPF dengan kernel RBF juga menunjukkan hasil yang sangat kompetitif, dengan perolehan akurasi 96,50% dan F1-score 96,41%.

penelitian[20] yang memperkenalkan SMOTE-IPF. Penelitian tersebut bertujuan untuk mengatasi keterbatasan signifikan dari metode SMOTE standar, yang performanya dapat menurun pada dataset yang mengandung *noise* dan *borderline examples*. Metode yang diusulkan adalah sebuah ekstensi dari SMOTE yang mengintegrasikan sebuah filter *noise* berbasis *ensamble*, yaitu IPF, yang dijalankan setelah proses *oversampling* untuk membersihkan data. Untuk memvalidasi keunggulan metodenya, penelitian tersebut melakukan studi eksperimental yang komprehensif pada berbagai jenis dataset (sintetis, dunia nyata, dan dunia nyata dengan tambahan *noise*) dengan membandingkan SMOTE-IPF dengan SMOTE standar dan varian-varian lainnya seperti SMOTE-ENN dan *Borderline*-SMOTE. Hasil penelitian yang dievaluasi menggunakan metrik AUC dan uji statistik secara konsisten menunjukkan bahwa SMOTE-IPF memiliki performa yang lebih unggul dibandingkan metode-metode pembandingnya di semua skenario pengujian, terutama pada dataset yang paling kompleks.

Tabel 2.1. Perbandingan beberapa penelitian terdahulu mengenai klasifikasi

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Peneliti (Sitasi) \*** | **Dataset** | **Variabel** | **Metode** | **Hasil\*** |
| **Ayu Dkk**[13] | Ecoli  Abalone19  Glass-6  Wine  Liver disorder  Thyroid-2  Vehicle1  Pima-ssl10  Page-block | Semua variable dari tiap tiap dataset | SVM, Navie bayes | Oversampling selalu meningkatkan performa. **SMOTE Borderline** memberikan peningkatan tertinggi pada dataset Abalone 19 untuk SVM (akurasi naik dari 64.5% ke 84.5%). |
| **Priyadarshinee dan Panda** [14] | Heart Failure Clinical Records |  | SMOTE-IPF | Terklasifikasi menjadi 2 kelompok dengan akurasi mencapai 85% |
| **Akhmad Syukron**[15] | Heart Failure  Prediction. | age, anaemia,  creatinine phosphokinase, Diabetes, ejection fraction,  high\_blood\_pressure, platelets, serum creatinine, serum  sodium, sex, Smoking, time | SMOTE dan Random Forrest,C45,  SVM,  Naïve Bayes, Neural Network, KNN, Ada Bost, Bagging | Hasil kinerja yang  didapat menunjukkan dengan model pengklasifikasi  SMOTE Random Forest punya nilai akurasi yang lebih tinggi  dibandingkan dengan model yang lain. Untuk nilai accuracy  sebesar 0,881 atau 88,1% dan nilai AUCnya sebesar 0.947 atau  94,7%. Dari sini dapat disimpulkan bahwa algoritma dengan  performa terbaik yaitu SMOTE Random Forest. |
| **Erlin, Erlin Dkk**[5] | UCI Machine Learning Repository (299 data) |  | **Oversampling:** SMOTE (standar)  **Klasifikasi:** Random Forest. | SMOTE meningkatkan semua metrik, contoh: **Akurasi +3.45%**, **Sensitivity +7.1%**. Nilai akurasi akhir **88,1%**. |
| **Abdellatif dkk**[16] | Improving the Heart Disease Detection and Patients’ Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest | Statlog dan Heart Disease Clinical Records | Seleksi Fitur: Inf-FS (Supervised Infinite Feature Selection)  Klasifikasi & Penanganan Imbalance: IWRF (Improved Weighted Random Forest), yaitu memberikan bobot/penalti lebih tinggi pada kelas minoritas. Optimasi: Bayesian Optimization (BO) untuk tuning hyperparameter IWRF | Model Inf-FS-BO-IWRF yang diusulkan terbukti lebih unggul dibandingkan model lain seperti SVM, KNN, dan RF standar. Dalam perbandingan langsung, IWRF juga menunjukkan hasil yang lebih baik daripada SMOTE-RF |
| **Hassanat dkk.**[17] | Dataset Diabetes,  Dataset Framingham,  Dataset Theoracic Surgery | 1. Diabetes (8 fitur) Framingham (15 fitur) Thoracic Surgery (16 fitur) | Bukan metode klasifikasi, melainkan sebuah sistem validasi untuk menguji apakah sampel sintetis yang dihasilkan oleh berbagai metode oversampling (termasuk SMOTE & SMOTE-IPF) benar-benar milik kelas minoritas. | Menunjukkan bahwa semua metode oversampling yang diuji menghasilkan sampel yang keliru (lebih mirip ke kelas mayoritas). Pada dataset Framingham, SMOTE-IPF memiliki *error rate* 12,4%. Mereka menyimpulkan bahwa oversampling adalah praktik yang berisiko untuk data medis. |
| **Masruriyah dkk.**[18] | Data pasien Penyakit Jantung dengan total objek lebih dari 300.000 data dengan delapan belas  variabel dan satu kelas target. S | BMI  Smoking  AlcoholDrinking  Stroke  PhysicalHealth  MentalHealth  DiffWalking  Sex  AgeCategory  Race  Diabetic  PhysicalActivity  GenHealth  SleepTime  Asthma  KidneyDisease  SkinCancer  HeartDisease (Kelas Target) | Metode yang di gunakan adalah  SMOTE (Synthetic Minority Oversampling Technique)  ADASYN (Adaptive Synthetic Sampling)  Random Forest (RF)  C45 (sebagai jenis *Decision Tree*)  Support Vector Machine (SVM)  Logistic Regression (LG)  K-Fold Cross Validation | *oversampling* SMOTE dengan algoritma *Random Forest* menunjukkan kinerja terbaik dalam prediksi penyakit jantung. Melalui validasi K-Fold 10, model gabungan ini mampu mencapai tingkat akurasi hingga  94,54% dengan nilai kurva ROC 98,4%. Angka tersebut merupakan peningkatan yang signifikan dari performa awal  *Random Forest* sebagai model tunggal yang memiliki akurasi 90,71%. Selain itu, performa SMOTE +  *Random Forest* ini juga terbukti lebih baik daripada kombinasi yang menggunakan ADASYN |
| **Martini dkk.**[19] | Dataset stroke dari Kaggle (5.110 data, 12 fitur). | • Id • Gender • Age • Hypertension • Heart disease • Ever\_married • Work\_type • Residence\_type • Avg\_glucose\_level • Bmi • Smoking status | **Penanganan Imbalance:** • SMOTE  • Borderline-SMOTE  • SMOTE-ENN  • SMOTE-IPF | kombinasi Borderline-SMOTE dengan *kernel* RBF mencapai performa tertinggi, dengan akurasi 96,86% dan *F1-score* 96,79%. Selain itu, kombinasi SMOTE-IPF dengan kernel RBF juga menunjukkan hasil yang sangat kompetitif, dengan perolehan akurasi 96,50% dan F1-score 96,41%. |
| **Sáez dkk.**[20] | • Sintetis: 30 dataset dengan berbagai level noise dan bentuk batas kelas (non-linear).  • Dunia Nyata: 9 dataset dari repositori KEEL (termasuk Cleveland, Pima, Ecoli).  • Dunia Nyata dengan Noise Tambahan: 9 dataset dunia nyata yang ditambahkan class noise & attribute noise. | Bervariasi (karena menggunakan banyak dataset), tidak fokus pada set fitur spesifik | • Metode Usulan: SMOTE-IPF, yaitu menerapkan SMOTE terlebih dahulu, kemudian diikuti dengan pemfilteran menggunakan Iterative-Partitioning Filter (IPF) untuk membersihkan *noise*. • Metode Pembanding: SMOTE standar, SMOTE-ENN, SMOTE-TL, Borderline-SMOTE, dan Safe-Level-SMOTE. • Klasifikasi: Menggunakan C4.5 sebagai *classifier* utama untuk melaporkan hasil, dan juga diuji dengan *classifier* lain termasuk SVM. • Evaluasi: AUC (*Area Under the Curve*) dan uji statistik (Wilcoxon, Friedman). | *•* ***SMOTE-IPF secara statistik terbukti lebih unggul*** *dibandingkan SMOTE standar dan semua varian SMOTE pembanding lainnya di semua skenario pengujian (sintetis, dunia nyata, dan dunia nyata dengan noise). • Keunggulan SMOTE-IPF paling signifikan pada dataset yang paling kompleks (memiliki batas non-linear dan tingkat noise yang tinggi). • Kesimpulan serupa mengenai keunggulan SMOTE-IPF juga ditemukan saat menggunakan classifier lain, termasuk* ***SVM****.* |

\*Wajib ada

## Landasan Teori

### Machine Learning

*Machine learning* adalah cabang dari *Artificial Intelegent* yang memungkinkan sistem komputer untuk belajar dari data dan pengalaman tanpa harus diprogram secara eksplisit. Alih-alih mengikuti instruksi yang kaku, algoritma *machine learning* membangun model matematika berdasarkan data sampel untuk membuat prediksi atau keputusan.

Machine learning terbagi menjadi 3 jenis yang pertama adalah *supervised learning*, model belajar dari data yang sudah diberi label. Artinya, setiap data masukan *(input)* sudah memiliki pasangan keluaran (output) yang benar. Tujuannya adalah agar model dapat mempelajari pola hubungan antara *input* dan *output*, sehingga nantinya bisa memprediksi *output* untuk data baru yang belum pernah dilihat.

Lalu ada *unsupervised learning* Berbeda dengan *supervised learning*, *unsupervised learning* menggunakan data yang tidak berlabel. Model tidak diberi tahu "jawaban" yang benar. Sebaliknya, tujuannya adalah untuk menemukan pola, struktur, atau hubungan tersembunyi dalam data secara mandiri.

Lalu ada *Reinforcement learning,* *Reinforcement learning* adalah jenis pembelajaran di mana sebuah model belajar untuk membuat keputusan dengan cara berinteraksi dengan lingkungannya. Model akan menerima imbalan (*reward*) untuk tindakan yang benar dan hukuman (*punishment*) untuk tindakan yang salah. Tujuannya adalah memaksimalkan total imbalan dari waktu ke waktu.

Tahapan *machine* *Learning* Di bagi menjadi 2 tahap, yang pertama adalah tahap pelatihan Pada tahapan ini model *machine learning* "belajar" dari data. Prosesnya adalah dengan memberikan data training kepada *algoritma*. *Algoritma* akan menganalisis data yang sudah berisi jawaban atau pola yang benar, untuk menemukan hubungan, pola, dan parameter yang paling sesuai.Tujuan dari tahap ini adalah untuk "*mengkalibrasi*" model sehingga ia dapat membuat prediksi atau keputusan yang akurat. Hasil dari tahap pelatihan adalah sebuah model yang terlatih.

Klasifikasi merupakan salah satu tugas dalam *supervised machine learning* yang bertujuan untuk membangun sebuah model yang dapat memprediksi label kelas yang bersifat *diskrit* (*kategorikal*) untuk suatu objek berdasarkan serangkaian fitur atau atribut yang dimilikinya[21]. Tujuan utama dari model klasifikasi adalah untuk belajar dari pola yang ada pada dataset latih, di mana setiap sampel telah memiliki label kelas yang benar, untuk dapat menggeneralisasi dan memprediksi label pada data baru yang belum pernah dilihat sebelumnya[21]. Dalam bidang medis, klasifikasi digunakan secara luas untuk membantu proses diagnosis, misalnya untuk menentukan apakah seorang pasien menderita penyakit jantung atau tidak berdasarkan data klinisnya.[22]

### Support Vector Machine (SVM)

SVM adalah algoritma klasifikasi yang kuat dan populer karena kemampuannya menangani data yang kompleks dan berdimensi tinggi. Prinsip dasar SVM adalah menemukan sebuah bidang pemisah atau *hyperplane* optimal yang dapat membedakan antar kelas data dalam ruang fitur[22].

*Hyperplane* yang dianggap optimal adalah yang memiliki *margin* atau jarak fungsional terbesar terhadap titik data terdekat dari setiap kelas. Titik-titik data yang berada tepat di batas margin ini disebut sebagai *support vectors*, karena posisi merekalah yang secara krusial "mendukung" dan menentukan orientasi dari *hyperplane*. Untuk kasus di mana data tidak dapat dipisahkan secara linear, SVM memanfaatkan sebuah mekanisme cerdas yang disebut *kernel trick*. Fungsi kernel (misalnya, Linear, Polinomial, atau Radial Basis Function/RBF) memetakan data ke ruang berdimensi lebih tinggi di mana pemisahan linear menjadi mungkin, tanpa perlu melakukan perhitungan di ruang dimensi tinggi tersebut secara eksplisit.[8]

Fungsi keputusan untuk SVM dalam memprediksi kelas dari sebuah data baru (x) dapat direpresentasikan dalam rumus berikut.

Prediksi=sign(i=1 (1)

Keterangan:

Prediksi: Hasil akhir prediksi kelas (biasanya bernilai +1 atau -1).

sign: Fungsi Signum, yang menentukan kelas berdasarkan tanda (positif atau negatif) dari hasil perhitungan di dalam kurung.

x: Vektor fitur dari data baru yang akan diklasifikasikan.

xᵢ: Vektor fitur dari *support vector* ke-i.

yᵢ: Label kelas dari *support vector* ke-i (bernilai +1 atau -1).

αᵢ (alpha): Bobot atau pengali Lagrange yang menunjukkan tingkat kepentingan dari setiap *support vector* xi​.

K(xᵢ, x): Fungsi Kernel, yang menghitung kemiripan antara data baru x dan *support vector* xi​.

b: Bias atau *intercept*, yang menyesuaikan posisi *hyperplane*.

m: Jumlah total dari *support vectors*.

### Ketidakseimbangan Data (*Class* *Imbalance*)

Salah satu tantangan persisten dalam aplikasi *machine learning* di dunia nyata adalah masalah ketidakseimbangan kelas (*class imbalance*). Masalah ini terjadi ketika distribusi sampel dalam dataset tidak merata antar kelas. Seringkali, kelas yang dominan disebut sebagai kelas mayoritas, sementara kelas dengan jumlah sampel yang jauh lebih sedikit disebut sebagai kelas minoritas. Dalam dataset medis, kelas minoritas seringkali menjadi kelas yang paling penting untuk dideteksi, misalnya kelas pasien yang terdiagnosis positif menderita penyakit.

Dampak utama dari ketidakseimbangan data adalah model klasifikasi yang dilatih akan cenderung bias terhadap kelas mayoritas. Model dapat mencapai nilai akurasi yang tinggi hanya dengan memprediksi mayoritas kelas, namun dengan performa yang sangat buruk dalam mengidentifikasi kelas minoritas. Hal ini sangat berisiko dalam aplikasi medis karena kesalahan dalam mengklasifikasikan kelas pasien (minoritas) memiliki dampak yang jauh lebih serius daripada kesalahan dalam mengklasifikasikan kelas non-pasien (mayoritas).

### Metode Penanganan Data Tidak Seimbang

Secara umum, solusi untuk mengatasi masalah ketidakseimbangan data dapat dikategorikan menjadi tiga pendekatan utama: pendekatan di level data (*data-level*), pendekatan di level algoritma (*algorithm-level*), dan pendekatan hibrida. Penelitian ini berfokus pada pendekatan di level data, yang secara langsung memodifikasi dataset untuk menciptakan distribusi kelas yang lebih seimbang sebelum data tersebut dilatih oleh model. Pendekatan ini terbagi lagi menjadi dua teknik utama:

1. Undersampling: Mengurangi jumlah sampel pada kelas mayoritas.
2. Oversampling: Memperbanyak jumlah sampel pada kelas minoritas.

### Synthetic Minority Oversampling Technique (SMOTE)

SMOTE adalah salah satu metode *oversampling* yang paling populer dan sering digunakan untuk mengatasi masalah ketidakseimbangan kelas. Berbeda dengan metode *random oversampling* yang hanya menduplikasi data, SMOTE bekerja dengan menciptakan sampel "sintetis" baru di ruang fitur. Proses ini membantu model untuk menciptakan region keputusan yang lebih besar dan lebih jelas untuk kelas minoritas.

Cara kerja SMOTE secara konseptual adalah sebagai berikut.

1. Pilih sebuah sampel dari kelas minoritas secara acak.
2. Identifikasi *k*-tetangga terdekat (*k-nearest neighbors*) dari sampel tersebut yang juga berasal dari kelas minoritas.
3. Pilih salah satu dari tetangga tersebut secara acak.
4. Buat sampel sintetis baru pada titik manapun di sepanjang garis lurus yang menghubungkan sampel asli dengan tetangga yang telah dipilih.

Meskipun efektif, SMOTE memiliki keterbatasan. Metode ini dapat menciptakan *noise* jika sampel minoritas yang dipilih sebagai basis pembuatan sampel baru merupakan sebuah *outlier* atau berada di area yang tumpang tindih dengan kelas mayoritas. Beberapa studi kritis bahkan menunjukkan bahwa ada risiko di mana sampel sintetis yang dibuat oleh SMOTE sebenarnya lebih mirip dengan karakteristik kelas mayoritas, sebuah "jebakan" (*jeopardy*) yang dapat menyesatkan proses *training*, terutama pada data medis.

### SMOTE-IPF (SMOTE with Iterative Partitioning Filter)

SMOTE-IPF adalah salah satu varian dari teknik SMOTE yang menggabungkan proses *oversampling* dengan *Iterative Proportional Fitting* (IPF). Metode ini bertujuan untuk mengatasi ketidakseimbangan kelas secara iteratif dengan cara menghapus sampel kelas mayoritas yang mempengaruhi distribusi data dan pada saat yang sama menghasilkan sampel sintetis untuk kelas minoritas [20]

Proses kerja dari SMOTE-IPF melibatkan beberapa tahapan utama sebagai berikut.

1. Penentuan KNN (*K-Nearest Neighbors*): Langkah pertama adalah memanfaatkan algoritma KNN untuk mengidentifikasi tetangga terdekat bagi setiap sampel di dalam kelas minoritas [20]
2. Seleksi Sampel Minoritas: Dilakukan seleksi terhadap sampel-sampel spesifik dari kelas minoritas yang akan dijadikan basis untuk proses pembuatan data sintetis[20]
3. Proses *Iteratif*: Sebuah proses iteratif dijalankan untuk setiap sampel minoritas yang telah dipilih, di mana dalam setiap iterasinya sampel tersebut digunakan untuk menciptakan data sintetis baru[20].
4. Pembaruan Distribusi: Terakhir, distribusi kelas sintetis diperbarui berdasarkan distribusi aktual dari kelas minoritas dan mayoritas. Tahapan ini bertujuan untuk memastikan bahwa sampel baru yang dihasilkan tidak hanya terkonsentrasi di area yang padat, tetapi juga dapat mencerminkan distribusi asli dari kelas minoritas [20].

Dengan memisahkan proses pembersihan *noise* dari proses pembuatan sampel sintetis, SMOTE-IPF bertujuan untuk memastikan bahwa SMOTE hanya bekerja pada data minoritas yang representatif, sehingga mengurangi risiko pembuatan sampel sintetis di area yang salah dan meningkatkan kualitas data latih secara keseluruhan.

### Metrik Evaluasi Kinerja Model

Untuk mengevaluasi performa model klasifikasi secara akurat, terutama pada kasus data tidak seimbang, digunakan serangkaian metrik yang dihitung dari *confusion matrix*. *Confusion matrix* membandingkan hasil prediksi model dengan kelas aktual dan terdiri dari empat komponen: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN).

True Positive (TP) menunjukkan jumlah data yang diprediksi sebagai kelas positif oleh model dan ternyata memang benar merupakan kelas positif. Misalnya, dalam kasus deteksi penyakit, TP terjadi ketika model memprediksi seorang pasien menderita penyakit dan pada kenyataannya pasien tersebut memang menderita penyakit tersebut.

True Negative (TN) adalah jumlah data yang diprediksi sebagai kelas negatif dan memang benar merupakan kelas negatif. Sebagai contoh, model memprediksi seorang pasien tidak menderita penyakit, dan kenyataannya pasien tersebut memang sehat.

False Positive (FP) terjadi ketika model memprediksi kelas positif, padahal sebenarnya data tersebut termasuk dalam kelas negatif. Contohnya adalah ketika model memprediksi seseorang menderita penyakit, tetapi kenyataannya orang tersebut sehat. Kesalahan jenis ini dikenal juga sebagai *Type I Error*.

Sebaliknya, False Negative (FN) terjadi ketika model memprediksi kelas negatif, padahal data sebenarnya termasuk dalam kelas positif. Misalnya, model memprediksi seorang pasien tidak menderita penyakit, padahal pasien tersebut sebenarnya sakit. Kesalahan ini dikenal sebagai *Type II Error*.

Metrik-metrik yang digunakan dalam penelitian ini antara lain.

1. Accuracy, Mengukur proporsi total prediksi yang benar dari keseluruhan data.

​ (2.2)

1. Precision, Mengukur proporsi prediksi positif yang benar dari total prediksi positif yang dibuat oleh model.

​(2.3)

1. Recall (Sensitivity) Mengukur proporsi prediksi positif yang benar dari total sampel positif yang sebenarnya.

​(2.3)

1. Specificity Mengukur proporsi prediksi negatif yang benar dari total sampel negatif yang sebenarnya.

​(2.5)

1. F1-Score, Rata-rata harmonik dari *Precision* dan *Recall*, memberikan skor tunggal yang menyeimbangkan kedua metrik tersebut.

​(2.6)

1. AUC (Area Under the Curve), Metrik ini mengukur keseluruhan kemampuan sebuah model untuk membedakan antara kelas positif dan negatif di semua ambang batas klasifikasi. Nilai AUC yang mendekati 1 menunjukkan performa pemisahan yang sangat baik.

# METODOLOGI PENELITIAN



## Pengumpulan Data

Penelitian ini menggunakan data sekunder yang diperoleh dari dataset *Framingham Heart Study*, sebuah dataset yang umum digunakan dalam penelitian prediksi penyakit jantung. Dataset ini diunduh dari situs *Kaggle*, yang merupakan platform berbagi dataset dan kompetisi ilmu data.

Dataset Framingham ini berisi informasi medis dan demografis dari sejumlah individu, termasuk variabel-variabel seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, riwayat merokok, riwayat diabetes, dan status penyakit jantung. Total terdapat 15 atribut (fitur) dan lebih dari 4.000 baris data.

Teknik pengumpulan data dilakukan melalui pengunduhan langsung dari sumber publik, sehingga tidak melibatkan pengukuran langsung, wawancara, ataupun survei. Pemilihan dataset ini didasarkan pada pertimbangan kesesuaiannya dengan topik penelitian, yaitu prediksi penyakit jantung, serta karena dataset ini banyak digunakan sebagai standar dalam studi klasifikasi kesehatan.

## *Software* dan *Hardware*

### Software

Penelitian ini menggunakan perangkat lunak berbasis cloud, yaitu Google Colaboratory (Colab), untuk seluruh proses eksplorasi data, praproses, implementasi algoritma, dan evaluasi model. Google Colab menyediakan lingkungan *notebook* berbasis Python yang mendukung eksekusi kode langsung di browser tanpa perlu instalasi lokal, serta mendukung pustaka *machine learning* populer seperti.

1. *NumPy* dan *Pandas* untuk manipulasi data
2. *Matplotlib* dan *Seaborn* untuk visualisasi
3. *Scikit*-*learn* untuk pemodelan dan evaluasi klasifikasi menggunakan SVM
4. *Imbalanced*-*learn* untuk penerapan teknik oversampling SMOTE dan SMOTE-IPF

Penggunaan *Google* *Colab* dipilih karena kemudahan akses, integrasi dengan Google Drive, serta ketersediaan sumber daya komputasi GPU/TPU secara gratis.

### Hardware

Eksekusi program dalam penelitian ini dilakukan melalui platform cloud Google Colab, sehingga spesifikasi perangkat keras lokal tidak menjadi faktor utama dalam proses pelatihan maupun eksekusi model. Namun, perangkat keras lokal tetap berperan penting dalam aktivitas pendukung seperti akses ke platform, penulisan laporan, serta persiapan dan eksplorasi data awal.

Untuk keperluan tersebut, peneliti menggunakan sebuah laptop HP Pavilion X360 yang dibekali dengan prosesor Intel Core i3-1215U (generasi ke-12). Laptop ini dilengkapi dengan RAM sebesar 8 GB dan GPU terintegrasi Intel UHD Graphics, serta menjalankan sistem operasi Windows 10 (64-bit). Akses ke platform Google Colab dilakukan melalui peramban berbasis Chromium seperti Google Chrome atau Brave, yang kompatibel dan mendukung fitur-fitur modern berbasis web.

Sementara itu, Google Colab sebagai runtime environment menyediakan spesifikasi perangkat keras virtual yang cukup mumpuni untuk keperluan komputasi ilmiah. Umumnya, runtime default menggunakan CPU virtual 2-core Intel Xeon, dengan RAM berkisar antara 12 hingga 25 GB, tergantung pada kondisi server dan pilihan runtime. Jika peneliti mengaktifkan opsi GPU, maka sistem dapat menyediakan GPU seperti NVIDIA Tesla K80 atau NVIDIA T4, yang sangat membantu dalam mempercepat proses pelatihan model. Untuk penyimpanan, Google Colab terintegrasi langsung dengan Google Drive, sehingga data dan hasil eksperimen dapat disimpan dan diakses secara fleksibel melalui cloud.

## Tahapan Penelitian

### Explorasi Data

Tahap ini merupakan proses awal yang sangat penting dalam penelitian, yaitu pemahaman terhadap dataset Framingham Heart Study yang menjadi dasar analisis. Tujuannya adalah untuk mengenali struktur, kualitas, dan karakteristik data sebelum masuk ke tahap pemodelan.

Langkah pertama yang dilakukan adalah pemeriksaan struktur data, yang mencakup identifikasi jumlah total data, jumlah fitur atau atribut yang tersedia, serta pengelompokan jenis atribut apakah bersifat numerik atau kategorik. Langkah ini memberikan gambaran awal mengenai kompleksitas dan tipe data yang akan dianalisis.

Selanjutnya, dilakukan proses pembersihan data (data cleaning) untuk memastikan bahwa data yang digunakan dalam pemodelan bebas dari gangguan yang dapat menurunkan akurasi. Hal ini mencakup penanganan terhadap data yang hilang (missing values), baik dengan cara menghapus baris yang tidak lengkap atau mengisi nilai yang hilang dengan metode tertentu, seperti rata-rata atau median. Selain itu, format data juga disesuaikan agar kompatibel dengan algoritma yang akan digunakan.

Tahap berikutnya adalah analisis distribusi kelas pada kolom target, yaitu TenYearCHD, yang menunjukkan apakah seseorang mengalami penyakit jantung dalam 10 tahun ke depan. Dari analisis ini diketahui bahwa data cenderung tidak seimbang, di mana jumlah kasus non-penyakit jantung jauh lebih banyak dibandingkan dengan kasus penyakit jantung, yang dapat memengaruhi performa model klasifikasi.

Untuk memahami karakteristik setiap atribut numerik, dilakukan analisis statistik deskriptif. Analisis ini meliputi perhitungan nilai rata-rata, standar deviasi, nilai minimum, dan maksimum, yang membantu dalam mengenali distribusi serta variabilitas setiap fitur dalam dataset.

Terakhir, dilakukan visualisasi data untuk mendukung pemahaman yang lebih mendalam. Berbagai teknik visual seperti histogram, boxplot, dan heatmap digunakan untuk melihat pola distribusi data, hubungan antar variabel, serta mendeteksi adanya outlier atau nilai pencilan yang ekstrem. Visualisasi ini juga membantu dalam mengidentifikasi fitur-fitur yang memiliki korelasi tinggi dengan target, yang nantinya berguna dalam pemilihan fitur untuk pemodelan.

### Analisis Masalah Data Imbalance

Setelah diketahui bahwa distribusi kelas target tidak seimbang, dilakukan analisis dampak imbalance terhadap performa model. Model cenderung bias terhadap kelas mayoritas sehingga diperlukan teknik oversampling. Oleh karena itu, dilakukan eksperimen menggunakan dua pendekatan.

1. Menggunakan SMOTE tanpa IPF
2. Menggunakan SMOTE-IPF

Tujuannya adalah untuk mengetahui seberapa besar peningkatan performa model SVM ketika diterapkan teknik oversampling tersebut.

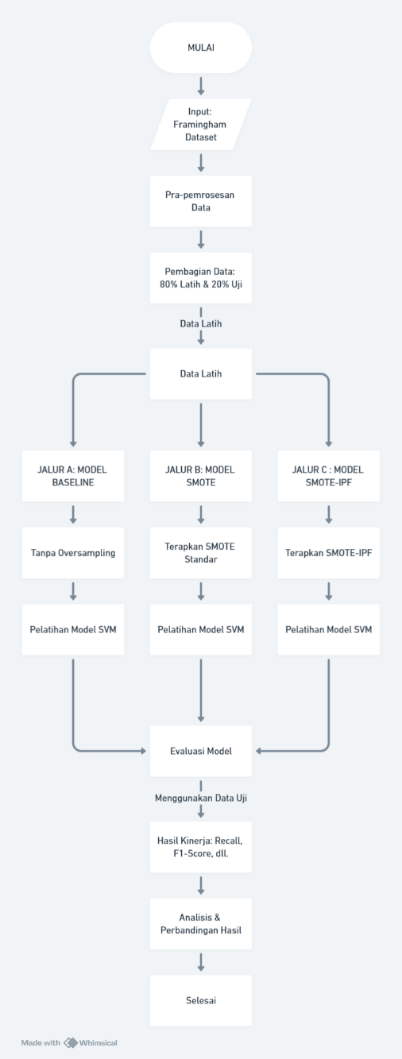
### Analisis Model

Model yang digunakan dalam penelitian ini adalah Support Vector Machine (SVM). Alasan pemilihan SVM adalah karena algoritma ini bekerja dengan baik pada data berdimensi tinggi dan efektif pada kasus klasifikasi biner seperti prediksi penyakit jantung.

Untuk menguji efektivitas teknik oversampling, model SVM dilatih dan dievaluasi dengan menggunakan data.

1. Tanpa *oversampling*
2. Dengan SMOTE
3. Dengan SMOTE-IPF

Evaluasi dilakukan dengan confusion matrix dan metrik seperti accuracy, precision, recall, dan F1-score untuk mengetahui dampak dari masing-masing pendekatan terhadap performa klasifikasi.



Gambar 3.1

## Evaluasi Model

Metode Evaluasi yang digunakan dalam penelitian ini bertujuan untuk mengevaluasi performa model klasifikasi yang telah dibangun dengan tiga pendekatan data.

1. SVM Tanpa teknik oversampling (baseline)
2. SVM Dengan teknik SMOTE
3. SVM Dengan teknik SMOTE-IPF

Evaluasi model dalam penelitian ini dilakukan dengan membandingkan hasil klasifikasi menggunakan sejumlah metrik evaluasi yang umum digunakan dalam analisis data klasifikasi, khususnya ketika menangani dataset yang tidak seimbang.

Metrik pertama yang digunakan adalah *accuracy*, yaitu ukuran proporsi prediksi yang dilakukan model dengan benar terhadap seluruh data. Meskipun sering digunakan, metrik ini bisa menyesatkan bila terjadi ketidakseimbangan kelas, karena model bisa menghasilkan akurasi tinggi hanya dengan memprediksi mayoritas kelas.

Untuk mengatasi keterbatasan tersebut, digunakan pula metrik *precision* dan *recall*. *Precision* mengukur sejauh mana prediksi positif dari model benar-benar merupakan kasus positif, sehingga metrik ini penting ketika kesalahan positif palsu (*false positive*) harus diminimalkan. Sementara itu, *recall* mengukur kemampuan model dalam mendeteksi seluruh kasus positif yang sebenarnya, sangat penting bila kesalahan negatif palsu (*false negative*) berdampak besar.

Sebagai pelengkap, digunakan juga *F1*-*score*, yaitu rata-rata harmonik dari *precision* dan *recall*. Metrik ini sangat berguna dalam situasi data tidak seimbang karena memberikan keseimbangan antara keduanya.

Untuk memberikan gambaran visual dan numerik yang lebih rinci, juga digunakan confusion matrix, yang menampilkan jumlah prediksi benar dan salah untuk tiap kelas. Matriks ini membantu dalam mengidentifikasi pola kesalahan model, misalnya jika model sering salah dalam mengklasifikasikan kelas minoritas.

Agar hasil evaluasi lebih stabil dan tidak bias terhadap pembagian data tertentu, digunakan teknik *Stratified K-Fold Cross Validation*. Teknik ini membagi dataset menjadi K bagian (*fold*) secara acak, namun tetap mempertahankan distribusi kelas yang proporsional di setiap *fold*. Model dilatih dan diuji sebanyak K kali, dengan setiap *fold* bergantian menjadi data uji, lalu hasil evaluasi dirata-rata untuk mendapatkan performa yang lebih representatif.

# HASIL DAN PEMBAHASAN



## Contoh Sub Bab

Hasil dan pembahasan merupakan bagian terpenting penting dari skripsi karena memuat semua temuan ilmiah yang diperoleh sebagai data hasil penelitian. Pembahasan dari hasil penelitian yang diperoleh dapat disajikan dalam bentuk uraian teoritik, baik secara kualitatif maupun kuantitatif. Dalam pembahasan ini juga dapat dilakukan perbandingan hasil penelitian yang dilakukan dengan hasil penelitian yang dilakukan oleh para peneliti terdahulu. Secara ilmiah hasil penelitian dapat berupa temuan baru, perbaikan, penegasan, atau penolakan interpretasi dari hasil peneliti sebelumnya. Hasil penelitian disajikan secara lengkap dan jelas dalam bentuk uraian yang dilengkapi dengan gambar/grafik, tabel, kurva, foto, atau dalam bentuk lain sesuai.

### Contoh Sub Sub Bab

Gambar 4.1 merupakan contoh penulisan kode program dalam naskah.

1. Penulisan Kode Program ditulis dalam frame dan diberi penomoran baris
2. Jenis huruf Courier New 10 pt, dengan 1 spasi.
3. Penomoran Kode Program sama dengan penomoran gambar. Gunakan style Gambar Judul.

|  |
| --- |
| 1. // Your First C++ Program 2. #include <iostream> 3. int main() { 4. std::cout << "Hello World!"; 5. return 0; 6. } |

Kode Program 4.1. Menampilkan tulisan “Hello World!”

Keterangan: angka 4 menunjukkan bab dan angka 1 menunjukkan urutan Kode Program dalam setiap bab.

## Contoh Sub Bab

Pembahasan dari hasil penelitian yang diperoleh dapat disajikan dalam bentuk uraian teoritik, baik secara kualitatif maupun kuantitatif. Dalam pembahasan ini juga dapat dilakukan perbandingan hasil penelitian yang dilakukan dengan hasil penelitian yang dilakukan oleh para peneliti terdahulu. Secara ilmiah hasil penelitian dapat berupa temuan baru, perbaikan, penegasan, atau penolakan interpretasi dari hasil peneliti sebelumnya.

# KESIMPULAN DAN SARAN



## Kesimpulan

Kesimpulan memuat secara singkat dan jelas tentang hasil penelitian yang diperoleh sesuai dengan tujuan penelitian. Kesimpulan merupakan rangkuman hasil yang dicapai dan merupakan jawaban rumusan masalah.

## Saran

Saran berisi hal-hal atau masalah yang perlu disampaikan untuk penelitian lebih lanjut. Bagian ini menguraikan saran‐saran yang perlu diperhatikan berdasarkan keterbatasan yang ditemukan dan asumsi yang dibuat, termasuk saran untuk pengembangan lebih lanjut.

# DAFTAR PUSTAKA

[1] World Health Organization, “Cardiovascular diseases,” https://www.who.int/health-topics/cardiovascular-diseases.

[2] Kementrian Kesehatan, “Penyakit Jantung Penyebab Utama Kematian, Kemenkes Perkuat Layanan Primer,” https://sehatnegeriku.kemkes.go.id/baca/rilis-media/20220929/0541166/penyakit-jantung-penyebab-utama-kematian-kemenkes-perkuat-layanan-primer/.

[3] T. R. Dawber, G. F. Meadors, and F. E. Moore, “Epidemiological Approaches to Heart Disease: The Framingham Study\*.”

[4] K. Ujaran, K. Ridwan, E. Heni Hermaliani, M. Ernawati, and C. Author, “Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada,” 2024. [Online]. Available: http://jurnal.bsi.ac.id/index.php/co-science

[5] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, “Dampak SMOTE terhadap Kinerja Random Forest Classifier berdasarkan Data Tidak seimbang,” *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

[6] A. Syukron, E. Saputro, and P. Widodo, “Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” 2023. [Online]. Available: https://doi.org/10/25047/jtit.v10i1.312

[7] L. N. Farida and S. Bahri, “Klasifikasi Gagal Jantung menggunakan Metode SVM (Support Vector Machine),” *Komputika : Jurnal Sistem Komputer*, vol. 13, no. 2, pp. 149–156, Oct. 2024, doi: 10.34010/komputika.v13i2.11330.

[8] M. G. Pradana, P. H. Saputro, and D. P. Wijaya, “KOMPARASI METODE SUPPORT VECTOR MACHINE DAN NAÏVE BAYES DALAM KLASIFIKASI PELUANG PENYAKIT SERANGAN JANTUNG,” *Indonesian Journal of Business Intelligence (IJUBI)*, vol. 5, no. 2, p. 87, Dec. 2022, doi: 10.21927/ijubi.v5i2.2659.

[9] J. KUSUMA, B. H. HAYADI, W. WANAYUMINI, and R. ROSNELLY, “Komparasi Metode Multi Layer Perceptron (MLP) dan Support Vector Machine (SVM) untuk Klasifikasi Kanker Payudara,” *MIND Journal*, vol. 7, no. 1, pp. 51–60, Jun. 2022, doi: 10.26760/mindjournal.v7i1.51-60.

[10] H. Hairani, K. E. Saputro, and S. Fadli, “K-means-SMOTE for handling class imbalance in the classification of diabetes with C4.5, SVM, and naive Bayes,” *Jurnal Teknologi dan Sistem Komputer*, vol. 8, no. 2, pp. 89–93, Apr. 2020, doi: 10.14710/jtsiskom.8.2.2020.89-93.

[11] B. Penelitian *et al.*, “KLASIFIKASI PENYAKIT JANTUNG MENGGUNAKAN METODE SUPPORT VECTOR MACHINE BERDASARKAN PERBANDINGAN ALGORITMA PEMBACAAN WAKTU DENGAN TEKSTUR SINYAL SEBAGAI METODE EKSTRAKSI SINYAL EKG,” vol. 02, no. 01, pp. 7–11, 2020.

[12] Alexsander, Ahmad Nazri, Rio Agus Panbudi, and Junadhi, “Implementasi Algoritma SVM dalam Memprediksi Penyakit Stroke,” *JOURNAL ZETROEM*, vol. 6, no. 2, pp. 1–5, Nov. 2024, doi: 10.36526/ztr.v6i2.3676.

[13] G. A. Pradipta and Putu Desiana Wulaning Ayu, “Kombinasi Inisial Filtering Oversampling dengan Metode Ensemble Classifier pada Klasifikasi Data Imbalanced,” *Jurnal Sistem dan Informatika (JSI)*, vol. 17, no. 2, pp. 137–145, May 2023, doi: 10.30864/jsi.v17i2.591.

[14] S. Priyadarshinee and M. Panda, “CARDIAC DISEASE PREDICTION USING SMOTE AND MACHINE LEARNING CLASSIFIERS,” *Journal of Pharmaceutical Negative Results ¦*, vol. 13, p. 2022, doi: 10.47750/pnr.2022.13.S08.108.

[15] A. Syukron, E. Saputro, and P. Widodo, “Penerapan Metode Smote Untuk Mengatasi Ketidakseimbangan Kelas Pada Prediksi Gagal Jantung,” 2023. [Online]. Available: https://doi.org/10/25047/jtit.v10i1.312

[16] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, “Improving the Heart Disease Detection and Patients’ Survival Using Supervised Infinite Feature Selection and Improved Weighted Random Forest,” *IEEE Access*, vol. 10, pp. 67363–67372, 2022, doi: 10.1109/ACCESS.2022.3185129.

[17] A. Hassanat *et al.*, “The Jeopardy of Learning from Over-Sampled Class-Imbalanced Medical Datasets,” in *2023 IEEE Symposium on Computers and Communications (ISCC)*, IEEE, Jul. 2023, pp. 1–7. doi: 10.1109/ISCC58397.2023.10218211.

[18] P. Studi Informatika and U. Buana Perjuangan Karawangsa, “MIND (Multimedia Artificial Intelligent Networking Database Evaluasi Algoritma Pembelajaran Terbimbing terhadap Dataset Penyakit Jantung yang telah Dilakukan Oversampling ANIS FITRI NUR MASRURIYAH, HILDA YULIA NOVITA, CICI EMILIA SUKMAWATI, SITI NOVIANTI NURAINI ARIF, ANGGA RAMDA RAMADHAN,” *Journal MIND Journal | ISSN*, vol. 8, no. 2, pp. 242–253, 2023, doi: 10.26760/mindjournal.v8i2.242-253.

[19] Luh Ayu Martini, G. A. Pradipta, and R. R. Huizen, “Analysis of the Impact of Data Oversampling on the Support Vector Machine Method for Stroke Disease Classification,” *Journal of Electronics, Electromedical Engineering, and Medical Informatics*, vol. 7, no. 2, pp. 404–421, Mar. 2025, doi: 10.35882/jeeemi.v7i2.698.

[20] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Inf Sci (N Y)*, vol. 291, pp. 184–203, Jan. 2015, doi: 10.1016/j.ins.2014.08.051.

[21] N. Ranti, M. 1\*, and K. H. Hanif, “Klasifikasi Penyakit Kanker Payudara Menggunakan Perbandingan Algoritma Machine Learning,” vol. 3, no. 1, pp. 1–6, 2022, [Online]. Available: http://creativecommons.org/licences/by/4.0/

[22] G. F. Fahrudin, S. Suroso, and S. Soim, “Pengembangan Model Support Vector Machine untuk Meningkatkan Akurasi Klasifikasi Diagnosis Penyakit Jantung,” *Jurnal Teknologi Sistem Informasi dan Aplikasi*, vol. 7, no. 3, pp. 1418–1428, Jul. 2024, doi: 10.32493/jtsi.v7i3.42254.

Daftar Pustaka disajikan pada halaman tersendiri dengan judul ‘DAFTAR PUSTAKA’, diketik dengan huruf kapital dan diletakkan pada sisi kiri halaman. Daftar Pustaka memuat semua pustaka yang digunakan dalam penelitian. penulisan daftar pustaka mengacu pada sistem angka IEEE dan diurutkan sesuai dengan sitasi yang dilakukan. Berikut merupakan contoh penulisan daftar pustaka menggunakan format IEEE. Gunakan style Daftar Acuan.

[1] G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955. (references)

[2] J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[3] I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963, pp. 271–350.

[4] K. Elissa, “Title of paper if known,” unpublished.

[5] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.

[6] Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science, 1989.

# LAMPIRAN

Lampiran berisi dokumen-dokumen yang digunakan sebagai pendukung penelitian, seperti dokumen pengujian. Berikut contoh penulisan lampiran untuk Skripsi Berbasis Publikasi.

## Lampiran 1. Bukti Pernyataan Diterima (ACCEPTED)

## Lampiran 2. Bukti Terakreditasi atau Terindeks SCOPUS

## Lampiran 3. Bukti Bayar (untuk jurnal yang berbayar)