BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Relevan

Penelitian yang relevan ini dilakukan untuk membandingkan penelitian sebelumnya dari jurnal dan paper yang sesuai sebagai referensi topik penelitian untuk memperkuat posisi penelitian ini sendiri. Berikut adalah beberapa contoh penelitian terdahulu tentang topik ini.

Penelitian Pertama yang diteliti oleh Aziz, Ishak, dan Abasa pada tahun 2024 yang berjudul *"Klasifikasi Tingkat Depresi Menggunakan Metode Support Vector Machine"* membahas penerapan algoritma SVM untuk mengklasifikasikan tingkat depresi berdasarkan data hasil kuesioner responden. Dalam penelitian ini, penulis melakukan tahapan pengumpulan data, preprocessing (termasuk normalisasi dan transformasi), serta pembagian data menjadi data latih dan data uji menggunakan metode *k-fold cross validation*. Model SVM diuji dengan beberapa jenis *kernel* seperti linear, polynomial, dan RBF, dengan evaluasi performa menggunakan akurasi, presisi, recall, dan F1-Score. Hasil penelitian menunjukkan bahwa SVM memiliki kemampuan klasifikasi yang baik terhadap data kesehatan mental sehingga pendekatan ini berpotensi diadaptasi untuk kasus klasifikasi penyakit jantung berdasarkan faktor klinis dan gaya hidup (Aziz1 et al., 2024).

Penelitian ketiga yang dipimpin oleh Betül Akalın, Ülkü Veranyurt, dan Ozan Veranyurt pada tahun 2020 denganjudul "Classification of Individuals at Risk of Heart Disease Using Machine Learning" ini membahas penggunaan algoritma pembelajaran mesin, seperti Gaussian Bayes, K-Nearest Neighbor (KNN), dan Random Forest, untuk mengklasifikasikan risiko seseorang mengalami penyakit jantung. Penelitian ini menggunakan dataset dari University of Cleveland yang berisi data dari 303 pasien. Dalam penelitian tersebut, para peneliti membandingkan hasil

dari algoritma dengan dan tanpa proses scaling fitur, serta melakukan pengecekan validasi silang sebanyak 50 kali untuk setiap algoritma. Hasil menunjukkan bahwa ketika data telah diubah skala, algoritma Random Forest memberikan akurasi tertinggi yaitu 82,50%, disusul oleh Gaussian Bayes dengan 80,52% dan KNN sebesar 80,52%. Namun, ketika data tidak dikasal, akurasi KNN turun hingga 65,28%. Penelitian ini menunjukkan bahwa penggunaan teknik pra-pengolahan data seperti scaling dapat meningkatkan akurasi dalam mengklasifikasikan risiko penyakit jantung, serta membuktikan bahwa pembelajaran mesin memiliki potensi besar dalam mendeteksi penyakit jantung secara dini(AKALIN et al., 2020).

Penelitian ketiga yang diteliti oleh Ammar Farisi, Ahmad Homaidi, dan Hermanto pada tahun 2025 dengan judul "Prediksi Penyakit Diabetes Menggunakan Algoritma Support Vector Machine (SVM)" membahas cara menggunakan algoritma SVM untuk mengklasifikasikan risiko seseorang terkena diabetes berdasarkan dataset Pima Indians Diabetes yang berisi 768 data pasien dari Kaggle. Penelitian ini menggunakan pendekatan Knowledge Discovery in Databases (KDD) yang mencakup beberapa tahap, seperti pra-pemrosesan data, normalisasi, pemilihan fitur, pemodelan, evaluasi, dan pembuatan sistem berbasis web dengan framework Streamlit. Model SVM yang digunakan memiliki kernel linear dan parameter C = 1.0, serta mampu memberikan hasil yang baik dengan akurasi 82,4%, presisi 79,6%, recall 76,1%, dan F1-score 77,8%. Hasil ini lebih baik dibandingkan algoritma Decision Tree yang memiliki akurasi 76,8% dan Naive Bayes dengan akurasi 74,5%. Penelitian ini menunjukkan bahwa algoritma SVM mampu memberikan prediksi yang stabil dan cepat pada data medis yang kompleks, serta bisa digunakan dalam aplikasi web untuk mendeteksi penyakit diabetes secara mandiri (Farisi & Homaidi, 2025).

Penelitian keempat yang dilakukan oleh Gusti Ayu Putu Febriyanti dan Anna Baita (2025) berjudul "Comparison of Support Vector Machine and Decision Tree Algorithm Performance with Undersampling Approach in Predicting Heart Disease Based on Lifestyle" membahas penggunaan algoritma Support Vector Machine (SVM) dan Decision Tree (DT) untuk memprediksi penyakit jantung berdasarkan faktor gaya hidup. Teknik undersampling digunakan untuk mengatasi ketidakseimbangan data. Dataset yang digunakan berasal dari Kaggle dengan nama Heart_2020_cleaned dan terdiri dari 319.795 data pasien. Hasil penelitian menunjukkan bahwa model SVM tanpa menggunakan undersampling memiliki akurasi tertinggi yaitu 92%, sedangkan ketika menggunakan undersampling, akurasi menurun menjadi 76%. Teknik undersampling membantu menyeimbangkan data dan meningkatkan kemampuan model dalam mengenali kelas yang jumlahnya sedikit, meskipun mengurangi akurasi secara keseluruhan. Dari hasil tersebut, disimpulkan bahwa model SVM lebih unggul dalam kinerja prediksi pada data yang tidak menggunakan teknik undersampling (Ayu et al., 2025).

Penelitian kelima yang dilakukan oleh Siskawati Rahayu dan Yuni Yamasari (2024) berjudul "Klasifikasi Penyakit Stroke dengan Metode Support Vector Machine (SVM)" membahas penerapan algoritma SVM dengan empat jenis kernel, yaitu linear, polynomial, radial basis function (RBF), dan sigmoid, dalam mengklasifikasikan penyakit stroke. Penelitian ini menggunakan dataset publik Stroke Prediction Dataset dari Kaggle yang berisi 5110 data dengan 12 atribut. Proses penelitian terdiri dari beberapa tahap, yaitu data cleaning, label encoding, dan membagi data menjadi beberapa rasio untuk menguji kemampuan model. Hasil penelitian menunjukkan bahwa kernel polynomial memberikan hasil terbaik dengan akurasi 78,86%, presisi 73,98%, dan recall 56,75% pada rasio data 80:20 dibandingkan dengan kernel lainnya. Kesimpulan dari penelitian ini menyatakan bahwa SVM dengan kernel polynomial memiliki kinerja terbaik dalam mengklasifikasikan penyakit stroke dan bisa menjadi dasar untuk pengembangan

metode serupa dalam kasus klasifikasi penyakit jantung berbasis faktor klinis dan gaya hidup (Rahayu & Yamasari, 2024).

Penelitian keenam yang diteliti oleh Hilda Apriyani dan Kurniati tahun 2020 memiliki judul “Perbandingan Metode Naïve Bayes dan Support Vector Machine dalam Klasifikasi Penyakit Diabetes Melitus”. Penelitian ini membahas perbandingan antara metode Support Vector Machine (SVM) dan Naïve Bayes dalam mengklasifikasikan penyakit diabetes melitus menggunakan 613 rekam medis dengan 9 atribut. Mereka menggunakan alat WEKA dan metode 10-fold cross-validation untuk mengevaluasi hasil. Hasil menunjukkan bahwa SVM dengan kernel polynomial memberikan akurasi tertinggi sebesar 96,27% dengan error 3,73%, yang lebih baik daripada Naïve Bayes yang hanya mencapai 92,07%. Sementara itu, SVM dengan kernel RBF mencapai akurasi sebesar 80,89%. Temuan ini menunjukkan bahwa SVM, khususnya dengan kernel polynomial, memiliki keunggulan dalam klasifikasi klinis berdasarkan data rekam medis. Relevansi penelitian ini bagi skripsi Anda adalah penggunaan SVM sebagai model utama dalam klasifikasi risiko serangan jantung berdasarkan faktor klinis dan gaya hidup. Penelitian ini juga menekankan pentingnya pemilihan jenis kernel dan penggunaan validasi bertingkat untuk memastikan model dapat berfungsi baik di berbagai situasi (Apriyani, 2020) .

Penelitian ke tujuh yang dilakukan oleh Fatwa Abdusyukur pada tahun 2023 berjudul "Penerapan Algoritma Support Vector Machine (SVM) untuk Klasifikasi Pencemaran Nama Baik di Media Sosial Twitter" membahas penggunaan algoritma SVM untuk mendeteksi dan mengklasifikasikan tweet yang berisi pencemaran nama baik berdasarkan teks. Penelitian ini menggunakan metode CRISP-DM dengan tahapan business understanding, data preparation, modeling, evaluation, dan deployment terhadap 6.000 data tweet, yang terbagi secara merata antara kategori pencemaran nama baik dan bukan. Model dilatih menggunakan teknik k-fold cross validation sebanyak 10 kali dan dievaluasi dengan confusion matrix untuk mengukur

tingkat akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa model SVM mencapai akurasi tertinggi sebesar 87,7% dengan rata-rata akurasi 92% pada data latih dan 85% pada data uji, serta dikategorikan sebagai model yang cocok, mampu mengenali data secara baik dan efektif dalam mengklasifikasikan teks berbasis opini publik di media sosial (Abdusyukur, 2023).

Penelitian oleh Mardewi, Nuru Yarkuran, Sofyan, dan Firman Aziz (2023) berjudul "Klasifikasi Kategori Obat Menggunakan Algoritma Support Vector Machine" membahas penerapan metode SVM dengan tiga jenis kernel, yaitu linear, polynomial, dan RBF, untuk mengklasifikasikan kategori obat. Penelitian ini menggunakan 200 data dengan lima variabel prediktor, yaitu usia, jenis kelamin, tekanan darah, kolesterol, dan perbandingan natrium terhadap kalium (Na-to-K). Penelitian melalui tahapan preprocessing dan k-fold cross-validation. Hasil menunjukkan bahwa kernel linear dan polynomial memiliki akurasi terbaik sebesar 95,0%, sedikit lebih tinggi dibandingkan kernel RBF yang mencapai 94,5%. Temuan ini menunjukkan pentingnya memilih jenis kernel yang tepat dalam algoritma SVM. Selain itu, hasil penelitian membuktikan bahwa SVM masih efektif digunakan pada dataset medis kecil dengan fitur klinis yang sederhana. Temuan ini relevan bagi skripsi Anda karena dapat mendukung strategi pemilihan kernel dan metode validasi dalam mengklasifikasikan risiko serangan jantung berdasarkan faktor klinis dan gaya hidup (Kreatindo Manokwari et al., 2023).

Penelitian kesembilan yang dilakukan oleh Danis Rifa Nurqotimah, Ahsanun Naseh Khudori, dan Risqy Siwi Pradini pada tahun 2024 berjudul “Implementasi Algoritma Support Vector Machine (SVM) untuk Klasifikasi Penyakit Stroke” membahas penerapan algoritma SVM dalam mengklasifikasikan penyakit stroke. Penelitian ini menggunakan dataset medis yang diambil dari situs data.world.com, terdiri dari 40.910 data dengan 10 atribut utama, seperti usia, hipertensi, penyakit jantung, indeks massa tubuh (BMI), dan status perokok. Peneliti menggunakan tool

Orange untuk proses pra-pemrosesan data, normalisasi, pembagian data, maupun evaluasi model melalui metode cross-validation dan random sampling dengan berbagai jenis kernel. Hasil penelitian menunjukkan bahwa kernel RBF dengan metode cross-validation memberikan hasil terbaik, dengan akurasi sebesar 94,8%, presisi 95,0%, recall 94,8%, dan F1-score sebesar 92,3%. Dari hasil tersebut, penelitian menyimpulkan bahwa algoritma SVM, terutama dengan kernel RBF, memiliki performa yang sangat baik dalam klasifikasi penyakit stroke dan dapat digunakan sebagai alat bantu dalam diagnosis medis berbasis pembelajaran mesin (Danis Rifa Nurqotimah et al., 2024).

Penelitian kesepuluh yang teliti oleh Anita dkk. (2025) berjudul "Klasifikasi Faktor Risiko Penyakit Jantung Menggunakan Machine Learning" membahas cara mengelompokkan faktor-faktor risiko penyakit jantung di kalangan populasi Indonesia dengan menggunakan berbagai jenis algoritma machine learning. Data yang digunakan berasal dari Kaggle dan terdiri dari 920 rekor pasien Indonesia yang memiliki berbagai atribut klinis dan gaya hidup yang berkaitan dengan risiko penyakit jantung. Penelitian ini menunjukkan bahwa model machine learning mampu mengenali faktor-faktor risiko penyakit jantung secara efektif dalam konteks data Indonesia, sehingga memiliki potensi untuk membantu dalam mendeteksi penyakit jantung lebih awal dan mendukung pengambilan keputusan dalam bidang klinis. Hasil penelitian ini relevan dengan penelitian tersebut, terutama dalam membandingkan kemampuan algoritma SVM dengan metode lainnya dalam menggabungkan fitur klinis dan gaya hidup (Anita et al., n.d.)

Tabel 2. 1 Penelitian Yang relevan

No	Penulis	Tahun	Hasil Penelitian	Keterkaitan	Gap Penelitian
1	Aziz et al.	2024	SVM klasifikasi depresi, hasil akurasi tinggi dengan kernel linear, polynomial, dan RBF.	Menunjukkan efektivitas SVM dan tahapan evaluasi model.	**Domain non-jantung, tanpa faktor klinis.**
2	Akalın et al.	2020	Random Forest terbaik (82,5%) setelah scaling; KNN & Bayes menurun tanpa scaling.	Menegaskan pentingnya preprocessing dalam klasifikasi jantung.	**Tidak mengunakan SVM**
3	Farisi et al.	2025	SVM (linear) akurasi 82,4%, lebih baik dari DT & NB.	SVM efektif untuk prediksi penyakit medis.	**Domain diabetes, bukan jantung.**
4	Febriyanti & Baita	2025	SVM akurasi 92% (tanpa undersampling); turun 76% (dengan undersampling).	Relevan: faktor gaya hidup & SVM pada data jantung.	**Tidak ada uji dataset eksternal.**
5	Rahayu & Yamasari	2024	Kernel polynomial terbaik (78,86%) untuk klasifikasi stroke.	Memvalidasi pemilihan kernel SVM pada data medis.	**Bukan penyakit jantung.**
6	Apriyani & Kurniati	2020	SVM polynomial akurasi tertinggi 96,27%.	Bukti kekuatan SVM dalam data klinis.	**Tidak menggunakan metode RBF**

7	Abdusyukur	2023	SVM akurasi 87,7% klasifikasi teks.	Menunjukkan stabilitas SVM pada data besar.	**Domain non-medis.**
8	Mardewi et al.	2023	SVM linear/polynomial akurasi 95% pada data obat.	Menunjukkan pengaruh kernel SVM.	**Dataset kecil & sederhana.**
9	Nurqotimah et al.	2024	SVM RBF akurasi 94,8% pada data stroke besar.	Bukti kuat SVM efektif pada data medis besar.	**Fokus stroke, bukan jantung.**
10	Anita et al.	2025	ML identifikasi faktor risiko jantung (920 pasien Indonesia).	Konteks populasi Indonesia sangat relevan.	**Tidak jelaskan algoritma & metrik detail.**

Berdasarkan tabel perbandingan hasil penelitian di atas, dapat disimpulkan bahwa sebagian besar penelitian sebelumnya menunjukkan algoritma Support Vector Machine (SVM) memiliki kemampuan yang bagus dalam mengklasifikasi data medis, baik untuk penyakit jantung, stroke, maupun diabetes. Hasil penelitian menunjukkan bahwa pemilihan jenis kernel, teknik pengolahan data, serta metode pengujian seperti cross-validation sangat memengaruhi tingkat akurasi model. Selain itu, penggunaan SVM juga terbukti efektif dalam menangani data yang memiliki banyak variabel dan memiliki kemampuan untuk beradaptasi dengan baik, terutama ketika digabungkan dengan teknik normalisasi dan penyesuaian parameter.

Meskipun demikian, dari seluruh penelitian yang dianalisis, masih ada beberapa hal yang belum tercakup dalam konteks penelitian Anda, yaitu penerapan SVM khusus untuk mengevaluasi risiko serangan jantung berdasarkan faktor klinis dan gaya hidup masyarakat Indonesia. Mayoritas penelitian sebelumnya lebih

menitikberatkan pada penyakit lain seperti diabetes atau stroke, atau menggunakan data dari berbagai negara, sehingga belum sepenuhnya mewakili ciri khas populasi lokal. Oleh karena itu, penelitian Anda memiliki potensi untuk memberikan kontribusi baru dalam hal prediksi penyakit jantung berbasis data lokal dengan mempertimbangkan variabel klinis dan gaya hidup yang relevan.

2.2 Landasan Teori

Landasan teori berisi penjelasan mengenai konsep, teori, dan penelitian terdahulu yang menjadi dasar dalam penyusunan penelitian ini. Teori-teori yang disajikan memberikan pemahaman terhadap variabel, metode, serta pendekatan yang digunakan dalam proses penelitian, sehingga dapat memperkuat argumentasi ilmiah penelitian ini.

2.2.1 Serangan Jantung

Serangan Jantung merupakan penyakit yang mematikan dimana penyakit ini terjadi ketika gangguan jantung serius ketika otot jantung tidak mendapatkan aliran darah. penyakit jantung koroner (PJK) merupakan bentuk manifestasi utama dari penyakit kardiovaskular yang disebabkan oleh penyumbatan aliran darah pada arteri koroner akibat aterosklerosis.

Serangan jantung atau infark miokard adalah cedera akut pada otot jantung akibat iskemia (kekurangan aliran darah/oksigen) yang berlangsung cukup lama, sehingga menimbulkan kematian sel otot jantung. Secara klinis, diagnosis ditegakkan bila terdapat kenaikan/penurunan biomarker jantung (terutama troponin) dengan setidaknya satu nilai di atas batas rujukan persentil ke-99, disertai bukti iskemia seperti gejala khas, perubahan EKG, temuan imaging, atau bukti trombus koroner. Penyebab tersering adalah oklusi arteri koroner karena ruptur/erosi plak aterosklerotik yang memicu pembentukan trombus. Definisi ini mengikuti *Fourth Universal Definition of Myocardial Infarction* dan dirujuk dalam pedoman ESC terbaru untuk sindrom koroner akut.

faktor-faktor penyebab/risiko serangan jantung meliputi yang tidak dapat dimodifikasi (usia makin tua, laki-laki, riwayat keluarga/keturunan) dan yang dapat dimodifikasi seperti hipertensi, dislipidemia (LDL tinggi/non-HDL tinggi), diabetes mellitus, merokok, konsumsi alkohol berlebih, obesitas/lingkar pinggang berlebih, kurang aktivitas fisik, serta pola makan tidak sehat (tinggi garam, lemak trans, rendah buah-sayur); juga berkontribusi stres/psikososial, polusi udara, penyakit ginjal kronik, dan kondisi inflamasi tertentu. Bukti global dan pedoman klinis terbaru menempatkan hipertensi sebagai faktor risiko utama dan menegaskan peran kuat perilaku (tobacco, diet tak sehat, inaktivitas, alkohol) dan faktor metabolik (tekanan darah, kolesterol, glukosa) dalam memicu aterosklerosis koroner yang berujung oklusi trombotik. Dalam konteks Indonesia, indikator negara menunjukkan tingginya penggunaan tembakau pada laki-laki dan profil non-HDL pada perempuan yang relatif tinggi, sehingga pengendalian faktor-faktor ini menjadi prioritas. Rangkuman ini disarikan dari WHO/WHF (observatorium & fact sheet) dan pedoman ESC/consensus definisi MI.

2.2.2 Machine Learning

Machine learning adalah bidang dalam kecerdasan buatan yang berfokus pada pengembangan sistem yang mampu belajar dari data untuk mengenali pola dan membuat keputusan tanpa perlu pemrograman eksplisit. Dalam Algoritma *machine learning* membantu komputer dalam mempelajari pola dan tren dari data yang tersedia, sehingga mesin dapat digunakan untuk mengambil keputusan secara lebih akurat (Yessy Asri et al., 2024). Beberapa dekade yang lalu, Machine Learning menjadi salah satu alat yang sangat umum untuk digunakan pada setiap pekerjaan yang membutuhkan ekstraksi informasi dari suatu kumpulan data dengan ukuran yang besar(Haganta Depari et al., n.d.). Pada umumnya, proses pembelajaran pada machine learning terdiri dari dua tahapan utama, yaitu pelatihan (training) dan pengujian (testing). Kedua tahap ini dilakukan dengan cara membagi dataset menjadi data latih (train data) dan data uji (test data). Namun, dalam beberapa kasus,

pembagian data juga dapat ditambahkan tahap validasi (validation) untuk menyesuaikan kebutuhan analisis dan memperoleh hasil model yang lebih optimal sesuai tujuan penelitian(Yessy Asri, Dr. Dra. Dwina Kuswardani, et al., 2025).

Dalam dilakukan pemodelan machine learning, menjadi pendekatan populer untuk menganalisis data medis karena kemampuannya mengidentifikasi pola dari kumpulan data besar. Penelitian ini bertujuan untuk melakukan model klasifikasi penyakit jantung menggunakan Support Vector Machine. Sehingga Evaluasi machine learning dilakukan eksperimental untuk menganalisis hasil percobaan, dengan harapan dapat meningkatkan akurasi identifikasi pasien yang berisiko terkena penyakit jantung (Anita et al., n.d.).

Selain itu, Machine Learning terkait erat dengan konsep data mining yang bertujuan mencari pola dan pengetahuan tersembunyi dari data yang besar. Proses belajarnya terdiri dari beberapa tahap seperti persiapan data, melatih model, menguji model, serta mengevaluasi hasilnya dengan metrik seperti akurasi, presisi, dan recall. Dalam bidang kesehatan, Machine Learning digunakan untuk menganalisis data secara prediktif, sehingga bisa mendeteksi penyakit lebih cepat, lebih efisien, dan didasarkan pada bukti ilmiah.

2.2.3 Klasifikasi

Algoritma klasifikasi atau klasterisasi adalah metode dalam pembelajaran mesin yang digunakan untuk mengelompokkan data ke dalam kategori tertentu. Contohnya, menentukan apakah sebuah email adalah spam atau bukan spam, atau memprediksi apakah seorang pelanggan akan melakukan pembelian atau tidak. Klasifikasi merupakan tugas *supervised learning* untuk memetakan fitur (variabel input) menjadi label kelas (mis. “berisiko” vs “tidak berisiko”). Model belajar dari data berlabel sehingga bisa memprediksi kelas pada data baru. Masalah klasifikasi umum: biner (2 kelas), multikelas (>2 kelas), dan multilabel (satu contoh bisa memiliki banyak label sekaligus). Metode klasifikasi mencakup berbagai algoritma seperti Naive Bayes, Decision Tree, Support Vector Machine (SVM), Random

Forest, dan K-Nearest Neighbor (KNN). Setiap metode memiliki karakteristik yang berbeda; SVM misalnya, mencari *hyperplane* dengan margin maksimum; kuat pada data berdimensi tinggi; fleksibel lewat kernel (linear/polynomial/RBF).

Alur klasifikasi mencakup empat tahap utama, yaitu

- memahami data,
- mempersiapkan data,
- membuat model, dan
- mengevaluasi model.

Proses persiapan data meliputi mengisi nilai yang hilang menggunakan median atau modus, mendeteksi nilai-nilai yang tidak normal, mengubah data kategori menjadi bentuk kode yang bisa dipahami oleh model, serta menyesuaikan skala nilai numerik agar model seperti SVM atau kNN tidak bias karena fitur yang berskala besar. Untuk meningkatkan kemampuan model dalam menggeneralisasi, bisa dilakukan pemilihan fitur terbaik dengan metode filter seperti ANOVA atau Mutual Information, metode wrapper seperti RFE, maupun reduksi dimensi seperti PCA apabila terdapat banyak fitur yang saling berkorelasi. Semua langkah ini sebaiknya diatur dalam satu pipeline agar tidak terjadi kebocoran data antara data latih dan data uji. Masalah penting dalam data kesehatan adalah ketidakseimbangan jumlah kelas (kasus positif sedikit). Mengandalkan akurasi saja bisa mengelirukan karena model bisa dominan pada kelas mayoritas namun gagal mengenali pasien yang berisiko.

Oleh karena itu, evaluasi lebih menekankan pada recall dan sensitivitas untuk mengurangi false negative, F1-score, balanced accuracy, ROC-AUC, serta PR-AUC ketika tingkat ketidakseimbangan tinggi. Beberapa strategi untuk menangani ketidakseimbangan kelas mencakup penimbangan kelas, oversampling (contoh SMOTE), undersampling, atau kombinasi dengan ensemble. Semua strategi ini dievaluasi menggunakan Stratified k-fold cross-validation agar distribusi kelas tetap seimbang di setiap lipatan. Tahap tuning dan validasi sangat memengaruhi hasil

akhir. Pemilihan model didasarkan pada ukuran data, jumlah fitur, tingkat non-linearitas, kebutuhan interpretasi, serta dampak kesalahan.

2.2.4 Algoritma Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan metode klasifikasi yang bekerja dengan mencari pemisah optimal yang memiliki jarak maksimum terhadap data dari masing-masing kelas, di mana data yang berada paling dekat dengan pemisah tersebut disebut support vectors . Metode ini dapat digunakan untuk klasifikasi biner maupun multikelas dengan memanfaatkan fungsi kernel (Jatnika et al., 2024). Konsep utamanya adalah mencari hyperplane terbaik yang dapat memisahkan dua kelas data secara optimal. SVM bekerja dengan cara memaksimalkan jarak atau margin antara data dari dua kelas yang berbeda, sehingga menghasilkan batas keputusan yang lebih stabil. Ketika data tidak bisa dipisahkan secara linear, SVM menggunakan fungsi kernel untuk memindahkan data ke ruang berdimensi lebih tinggi, sehingga data bisa dipisahkan secara linear di ruang tersebut. Beberapa jenis fungsi kernel yang sering digunakan adalah linear, polynomial, RBF (radial basis function), dan sigmoid (Cortes et al., 1995).

Secara teknis, garis atau bidang yang berfungsi sebagai pemisah antar kelas data disebut *hyperplane*. Pada metode Support Vector Machine (SVM), tujuan utamanya adalah menemukan *hyperplane* yang paling optimal, yaitu pemisah yang mampu membedakan kelas-kelas data dengan jarak (margin) terbesar. Margin ini penting karena semakin besar jaraknya, semakin baik kemampuan model dalam melakukan generalisasi terhadap data baru.

Data-data yang posisinya paling dekat dengan *hyperplane* dinamakan *support vectors*. Titik-titik inilah yang sangat berpengaruh dalam menentukan letak dan arah *hyperplane*, karena perubahan pada *support vectors* dapat mengubah hasil pemisahan secara keseluruhan.

Apabila data dapat dipisahkan secara linear, maka SVM hanya perlu mencari garis lurus (untuk data dua dimensi) atau bidang datar (untuk data berdimensi lebih

tinggi) yang paling optimal. Namun, dalam banyak kasus nyata, data tidak dapat dipisahkan secara linear. Untuk mengatasi hal tersebut, SVM menggunakan metode yang disebut *Kernel Trick*, yaitu teknik yang memetakan data ke ruang berdimensi lebih tinggi. Dengan cara ini, data yang sebelumnya sulit dipisahkan menjadi lebih mudah untuk dipisahkan menggunakan *hyperplane* linear di ruang baru tersebut(Yessy Asri, Dwina Kuswardani, et al., 2025).

Keunggulan lain dari SVM adalah kemampuannya bekerja efektif pada dataset dengan jumlah kecil tapi berdimensi tinggi, serta lebih tahan terhadap overfitting dibandingkan algoritma lain. Implementasi SVM dalam penelitian ini dilakukan dalam beberapa tahap, seperti penskalaan fitur, pembagian data menjadi data latih dan uji, pelatihan model menggunakan kernel RBF, serta evaluasi performa melalui confusion matrix. Parameter C dan γ diatur secara optimal menggunakan metode grid search untuk mendapatkan keseimbangan antara kompleksitas model dan kemampuan generalisasi terhadap data baru. Dalam bidang kesehatan, SVM terkenal efektif dalam menganalisis data berdimensi tinggi dan kompleks, seperti data rekam medis, citra medis, atau hasil laboratorium. Dalam konteks klasifikasi risiko serangan jantung, SVM digunakan untuk menganalisis hubungan antara variabel klinis seperti tekanan darah, kolesterol, glukosa, dan detak jantung dengan variabel gaya hidup seperti merokok, aktivitas fisik, dan pola makan. Pendekatan ini diharapkan memberikan prediksi yang akurat dan stabil, serta mendukung upaya deteksi dini penyakit jantung berbasis data lokal Indonesia (Akalin et al., 2020).

Menurut Burges proses klasifikasi dibedakan menjadi dua jenis berdasarkan kernel yang digunakan, yaitu SVM Linear dan SVM Non-Linear(Rahayu & Yamasari, 2024). berikut tabel rumus dan penjelasan tersebut:

Tabel 2. 2 Tabel Rumus Kernel SVM

No	Fungsi Kernel	Rumus
1.	Linier	$\boldsymbol{K(x, x_k) = x_k^T x}$
2.	Polinomial	$\boldsymbol{K(x, x_k) = (x_k^T x + 1)}$
3.	Gaussian (Radial Basis Function / RBF)	$\boldsymbol{K(x, x_k) = \exp\left(\frac{-\| x_i - x_j \|^2}{2\sigma^2}\right)}$
4.	Sigmoid	$\boldsymbol{K(x, x_k) = k(x_k^T x) + 0}$

Sumber: Diadaptasi dari buku referensi *Optimalisasi Analisis Sentimen Dengan Spelling Corrector*

1. SVM Linear

Support Vector Machine (SVM) linear adalah pendekatan klasifikasi yang berupaya menemukan sebuah *hyperplane*—garis pada 2D, bidang pada 3D, atau batas keputusan pada dimensi lebih tinggi—yang memisahkan dua kelompok data dengan margin terbesar. Jika pola data memang dapat dipisahkan secara lurus (linearly separable), model ini biasanya menghasilkan batas keputusan yang sederhana, stabil, dan interpretatif. Dalam kerangka ini, kita tidak melakukan transformasi fitur yang rumit; model langsung bekerja pada ruang asli menggunakan kernel linier. Fungsi kernelnya berbentuk $K(x, x_i) = x \cdot x_i$, yaitu hasil kali titik (dot product) antara vektor fitur sampel uji dan sampel pelatihan. Sifatnya yang hemat parameter membuat SVM linier unggul di data berdimensi besar namun dengan struktur pemisahan yang cukup jelas, serta relatif mudah ditata melalui parameter regularisasi (misalnya C) untuk menyeimbangkan *margin* besar dan kesalahan klasifikasi.

2. SVM Non-linear

Ketika data tidak bisa dipisahkan hanya dengan garis lurus di ruang aslinya, SVM non-linier menggunakan "trik kernel" untuk memetakan data ke ruang fitur berdimensi lebih tinggi, tempat di mana pemisahan linier menjadi mungkin. Alih-alih menghitung pemetaan eksplisit $\phi(x)$, SVM menghitung kemiripan antar titik melalui fungsi kernel $K(x, x_i) = \langle \phi(x), \phi(x_i) \rangle$. Dengan demikian, kita memperoleh

fleksibilitas bentuk batas keputusan tanpa biaya komputasi pemetaan eksplisit. Beberapa kernel non-linier yang umum dipakai antara lain:

a. Kernel polynomial.

Kernel ini memproyeksikan data ke ruang dengan derajat lebih tinggi untuk menangkap hubungan non-linier antar fitur. Bentuk umumnya $K(x, x_i) = (x \cdot x_i + r)^d$, di mana d adalah derajat polinom dan r adalah *bias* (sering ditulis c atau γ_0). Dengan memilih d yang sesuai, model dapat membentuk *hyperplane* pada ruang terangkat yang memisahkan kelas dengan lebih efektif, terutama ketika interaksi antar fitur penting untuk akurasi.

b. Kernel Gaussian RBF.

Kernel Radial Basis Function (RBF) sangat populer untuk data yang batas kelasnya melengkung atau berkelok. Bentuknya $K(x, x_i) = \exp(-\gamma \| x - x_i \|^2)$, dengan γ mengatur "jangkauan pengaruh" tiap titik pelatihan. Ruang fitur tersirat dari RBF berdimensi tak hingga, sehingga secara teoritis mampu merepresentasikan beragam pola non-linier. Dengan pemilihan γ dan C yang tepat, SVM RBF dapat menghasilkan solusi pemisahan yang efektif sekaligus unik dalam ruang Hilbert reproduktifnya, sehingga sering menjadi *baseline* kuat untuk banyak dataset non-linier.

c. Kernel Sigmoid.

Kernel sigmoid berakar pada fungsi aktivasi jaringan saraf, dengan bentuk umum $K(x, x_i) = \tanh(\alpha\, x \cdot x_i + c)$. Secara intuitif, ia meniru perilaku lapisan tersembunyi tunggal pada perceptron berfungsi aktivasi *tanh*. Meski demikian, kernel ini cenderung sensitif terhadap pilihan parameter α dan c, dan tidak selalu memenuhi sifat positif-definit untuk semua kombinasi parameter serta distribusi data. Akibatnya, kestabilan dan kinerjanya dapat kurang konsisten dibandingkan RBF atau polinomial, sehingga penggunaannya biasanya memerlukan *tuning* yang lebih hati-hati.

Intinya, SVM linier cocok ketika pola pemisahan hampir lurus dan kita menginginkan model sederhana yang kuat terhadap *overfitting*. Sebaliknya, SVM non-linier—melalui pilihan kernel seperti polinomial, RBF, atau sigmoid—memberikan fleksibilitas untuk menangkap struktur batas keputusan yang kompleks pada data dunia nyata. Pemilihan kernel dan hiperparameter (misalnya C, γ, d, α, c) menjadi kunci untuk menyeimbangkan kemampuan representasi, generalisasi, dan stabilitas solusi yang dihasilkan.

2.2.5 Kernel Radial Basis Function

Kernel RBF (Radial Basis Function) adalah salah satu jenis kernel yang sering dipakai dalam machine learning, khususnya pada metode Support Vector Machine (SVM), untuk mengukur seberapa mirip satu data dengan data lainnya(Arifin et al., 2023). Kernel ini bekerja dengan cara memetakan data dari ruang berdimensi rendah ke ruang berdimensi lebih tinggi, supaya data yang awalnya sulit dipisahkan secara garis lurus bisa menjadi lebih mudah dipisahkan.

Kernel RBF menilai kemiripan berdasarkan jarak antar data. Jika dua data posisinya berdekatan, maka nilai kemiripannya akan besar, sedangkan jika jaraknya jauh, maka kemiripannya akan kecil. Karena kemampuannya menangani hubungan yang tidak linear, kernel RBF sangat cocok digunakan pada berbagai permasalahan seperti klasifikasi dan regresi. Inilah yang membuat kernel RBF menjadi salah satu kernel yang paling sering digunakan.

Rumus kernel RBF dapat dituliskan sebagai berikut:

$$K(x, x_k) = \exp\left(\frac{-\| x_i - x_j \|^2}{2\sigma^2}\right)$$

Pada persamaan tersebut:

- x dan x’ merupakan dua buah vektor fitur yang ingin dibandingkan tingkat kemiripannya
- γ (gamma) adalah parameter yang mengatur seberapa besar pengaruh satu titik data terhadap titik data lainnya.

Secara konsep, kernel RBF berfungsi untuk memetakan data yang awalnya bersifat non-linear di ruang asli ke dalam ruang fitur berdimensi lebih tinggi, sehingga data tersebut menjadi lebih mudah dipisahkan secara linear. Dengan kata lain, RBF membantu algoritma SVM dalam menemukan pola pemisah yang tidak bisa dilihat secara langsung di ruang data awal.

Nilai dari Nilai dari $K(x, x')$ akan mendekati 1 jika dua data x dan x' sangat dekat atau mirip, sangat dekat atau mirip, dan akan mendekati 0 jika kedua data tersebut sangat berbeda atau berjauhan. Mekanisme ini sangat membantu SVM dalam menentukan support vector serta margin terbaik untuk memisahkan kelas-kelas data. Parameter γ memiliki peran penting dalam mengatur kompleksitas model:

- Jika γ bernilai kecil, maka pengaruh tiap titik data akan menyebar lebih luas. Akibatnya, model yang dihasilkan menjadi lebih halus dan cenderung lebih sederhana, namun bisa kurang mampu menangkap detail pola data.
- Jika γ bernilai besar, maka pengaruh setiap titik data hanya berlaku di area yang sangat sempit. Hal ini membuat model menjadi lebih kompleks dan sensitif terhadap data training, sehingga berpotensi mengalami overfitting.

Oleh karena itu, pemilihan nilai γ yang tepat sangat penting agar model SVM dengan kernel RBF dapat memberikan hasil yang optimal, yaitu mampu mengenali pola data dengan baik tanpa terlalu kaku maupun terlalu bebas.

2.2.6 Synthetic Minority Over-sampling Technique (SMOTE)

SMOTE (*Synthetic Minority Over-Sampling Technique*) merupakan salah satu teknik over-sampling yang digunakan untuk mengatasi masalah ketidakseimbangan data *(imbalanced data*), yaitu kondisi ketika jumlah data pada

kelas minoritas jauh lebih sedikit dibandingkan kelas mayoritas. Dalam kondisi seperti ini, model machine learning cenderung lebih memihak kelas mayoritas sehingga kurang optimal dalam mengenali kelas minoritas, padahal kelas minoritas sering kali justru merupakan kelas yang paling penting untuk dideteksi.

SMOTE juga memiliki keterbatasan. Salah satu kelemahan utamanya adalah potensi terjadinya over-generalisasi, karena SMOTE membangkitkan data baru tanpa mempertimbangkan secara langsung keberadaan data kelas mayoritas di sekitar data minoritas. Akibatnya, data sintetis yang dihasilkan bisa terlalu mendekati wilayah kelas mayoritas dan menyebabkan tumpang tindih antar kelas (*overlapping*), yang berpotensi menurunkan kinerja klasifikasi jika tidak dikontrol dengan baik(Chawla et al., 2002).

Dalam konteks penelitian penulis tentang klasifikasi risiko serangan jantung di Indonesia, penggunaan SMOTE menjadi sangat relevan. Umumnya, data pasien dengan risiko tinggi serangan jantung jumlahnya lebih sedikit dibandingkan pasien dengan risiko rendah atau normal. Kondisi ini menyebabkan model klasifikasi cenderung lebih akurat dalam memprediksi pasien berisiko rendah, tetapi kurang sensitif dalam mendeteksi pasien berisiko tinggi, padahal kelompok inilah yang paling krusial dalam dunia medis.

Dengan menerapkan SMOTE, jumlah data pasien berisiko tinggi dapat ditingkatkan secara sintetis sehingga lebih seimbang dengan data pasien berisiko rendah. Hal ini membantu model machine learning, seperti SVM atau algoritma klasifikasi lainnya, untuk belajar lebih baik dalam mengenali pola pasien berisiko tinggi serangan jantung. Dampaknya, performa model dalam mendeteksi risiko serangan jantung menjadi lebih baik, khususnya dalam meningkatkan kemampuan model mengenali pasien yang benar-benar berisiko, sehingga dapat mendukung upaya pencegahan dan penanganan dini penyakit jantung di Indonesia.

2.2.7 Confusion Matrix

Confusion matrix merupakan tabel 2×2 untuk tugas klasifikasi biner yang merangkum perbandingan prediksi model dengan label sebenarnya. Tabel ini memisahkan keluaran ke dalam empat kategori: *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). Pada proses evaluasi, metrik yang lazim digunakan mencakup akurasi, presisi, dan recall, dengan presisi serta recall berperan menilai ketepatan sistem dalam mengenali kelas yang relevan. Presisi sendiri menggambarkan tingkat kecocokan antara data yang dinyatakan positif oleh model dengan informasi yang benar-benar diperlukan (proporsi prediksi positif yang benar). (Rahayu & Yamasari, 2024).

- Presisi

  Presisi—sering disebut *Positive Predictive Value* (PPV)—menunjukkan ketepatan prediksi positif sebuah model: dari seluruh sampel yang diprediksi positif, berapa persen yang benar-benar positif. Ukuran ini menilai seberapa "bersih" hasil positif model dari kesalahan. Secara matematis, presisi didefinisikan sebagai:

  $$Precision = TP/(TP + FP) \quad (2.1)$$

- Akurasi

  Akurasi menunjukkan berapa besar prediksi yang benar dibandingkan semua prediksi yang dibuat oleh model. Dalam klasifikasi dua kategori, "benar" mencakup jumlah prediksi yang benar positif (TP) dan benar negatif (TN). Nilai akurasi berada antara 0 hingga 1 (atau 0 hingga 100 persen). Akurasi didefinisikan sebagai seberapa jauh nilai yang didapatkan dan nilai yang sebenarnya berbeda (Rahayu & Yamasari, 2024).

  $$Accuracy = (TP + TN)/(TP + TN + FP + FN) \quad (2.2)$$

- Recall

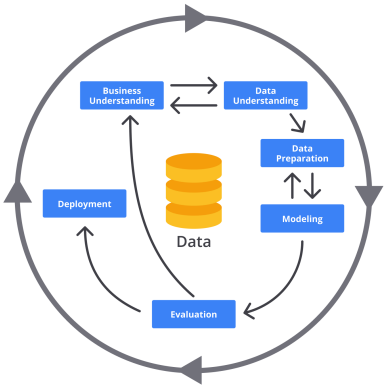
  Recall mengukur seberapa baik model mampu mendeteksi semua kasus yang sebenarnya positif. Dengan kata lain, dari semua pasien yang sebenarnya berisiko, berapa persen yang berhasil terdeteksi oleh model sebagai berisiko. Karena fokusnya pada hal-hal yang tidak terlewat, recall sangat penting dalam konteks medis. Nilainya berkisar antara 0 hingga 1 (atau 0 hingga 100 persen); semakin tinggi, semakin sedikit kesalahan negatif (FN). Tingkat keberhasilan sistem dalam menemukan Kembali data disebut recall (Rahayu & Yamasari, 2024).

$$\text{Recall} = \frac{TP}{TP + FN} \quad 2.3$$

2.2.8 CRISP-DM

CRISP-DM (Cross-Industry Standard Process for Data Mining) merupakan sebuah standar proses yang dikembangkan sejak tahun 1996 untuk membantu pelaksanaan analisis data secara terstruktur (Ruli Siregar et al., 2017). Metode ini digunakan sebagai panduan dalam mengolah dan menganalisis data pada berbagai bidang industri maupun penelitian, sehingga dapat membantu dalam menemukan solusi terhadap permasalahan bisnis atau kajian ilmiah yang sedang diteliti. Untuk menjalankan proyek data mining secara terstruktur dan berulang melalui enam tahapan utama, yaitu: pemahaman bisnis, pemahaman data, persiapan data, pembangunan model, evaluasi, dan penerapan model. CRISP-DM (Cross Industry Standard Process for Data Mining) merupakan kerangka proses yang secara luas digunakan dalam penelitian dan praktik data mining. Model ini dikembangkan agar bersifat independen terhadap industri dan perangkat yang digunakan, menyediakan alur kerja yang sistematis dari pemahaman bisnis hingga deployment (Wirth & Hipp, n.d.).

Menurut Schröer et al. (2021), CRISP-DM tetap menjadi standar de facto dalam proyek data mining, namun banyak studi menunjukkan bahwa fase deployment seringkali kurang diperhatikan atau dilaporkan secara terperinci (Schröer et al., 2021). Makalah adaptasi metodologi data mining juga menunjukkan bahwa CRISP-DM sering digunakan “as-is” atau dengan modifikasi minor tergantung kebutuhan domain, terutama untuk menangani volume data besar (big data) dan integrasi ke sistem operasional organisasi (Plotnikova et al., 2020). Sementara itu, penelitian terbaru juga menekankan bahwa CRISP-DM perlu dikembangkan agar lebih fleksibel dalam konteks data science modern, seperti penambahan fasa atau integrasi dengan kerangka agile(Shimaoka et al., 2024).



Gambar 2. 1 CRISP-DM

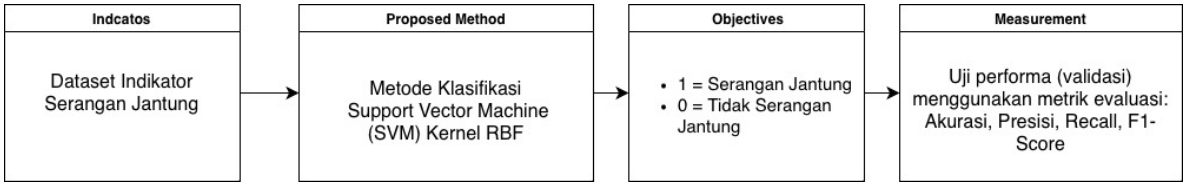
Sumber : (Dicoding.com)

Dalam penelitian mengklasifikasikan risiko serangan jantung dengan menggunakan SVM, metode CRISP-DM membantu membagi proses kerja menjadi beberapa tahap yang logis dan terstruktur. Tahapan tersebut meliputi:

- Business Understanding, yaitu menentukan tujuan kesehatan masyarakat serta target dalam memprediksi risiko serangan jantung,

- Data Understanding, yaitu memahami karakteristik data klinis dan pola hidup pasien,
- Data Preparation, yaitu membersihkan data, menormalkan data, mengisi nilai yang hilang, serta mengekstrak fitur yang relevan,
- Modeling, yaitu proses melatih model SVM dengan memilih jenis kernel dan parameter yang tepat,
- Evaluation, yaitu mengukur tingkat akurasi, recall, F1-score, validitas model.

2.3 Kerangka Pemikiran



Gambar 2. 2 Kerangka Pemikiran

Gambar 2.2 yang menjelaskan alur kerangka pemikiran dalam penelitian ini. Kerangka pemikiran tersebut menggambarkan hubungan antar komponen penelitian, sehingga dapat memperlihatkan bagaimana setiap bagian saling berkaitan dan mendukung proses penelitian secara keseluruhan. Adapun penjelasan dari masing-masing komponen dalam kerangka pemikiran tersebut adalah sebagai berikut:

1. **Indicator**

   Tahap pertama dalam penelitian ini adalah penggunaan dataset yang berisi berbagai indikator kesehatan yang berkaitan dengan risiko serangan jantung. Indikator tersebut meliputi faktor-faktor seperti usia, tekanan darah, kadar kolesterol, riwayat diabetes, serta faktor gaya hidup lainnya. Dataset ini menjadi dasar utama dalam proses pelatihan dan pengujian model klasifikasi.

2. **Proposed Method (Metode yang Diusulkan)**

   Metode yang digunakan dalam penelitian ini adalah metode klasifikasi Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF).

Pemilihan SVM kernel RBF didasarkan pada kemampuannya dalam menangani data dengan pola hubungan yang tidak linear. Dalam konteks data kesehatan, hubungan antar variabel sering kali kompleks, sehingga diperlukan metode yang mampu membentuk batas pemisah yang fleksibel dan optimal.

3. **Objectives (Tujuan Klasifikasi)**

   Tujuan dari penelitian ini adalah melakukan klasifikasi terhadap risiko serangan jantung dengan dua kategori, yaitu:

   - 1 = Serangan Jantung (berisiko)
   - 0 = Tidak Serangan Jantung (tidak berisiko)

   Model akan mempelajari pola dari data pelatihan untuk dapat membedakan kedua kelas tersebut secara akurat.

4. **Measurement (Pengukuran Performa Model)**

   Tahap terakhir adalah evaluasi atau pengukuran performa model. Setelah model dilatih dan diuji, hasil klasifikasi dievaluasi menggunakan beberapa metrik, yaitu akurasi, presisi, recall, dan F1-score. Metrik-metrik ini digunakan untuk mengetahui seberapa baik model dalam mengklasifikasikan risiko serangan jantung, terutama dalam mendeteksi pasien yang benar-benar berisiko

   .