BAB II
TINJAUAN PUSTAKA

2.1 Kajian Pustaka Relevan

Penelitian yang relevan dilakukan untuk membandingkan penelitian yang sudah pernah dilakukan dan berkaitan dengan topik klasifikasi hipertensi serta penerapan metode *Fuzzy C-Means* dan *Support Vector Machine* yang digunakan dalam penelitian ini. Tinjauan ini membantu memperkuat posisi penelitian yang sedang dilakukan, sekaligus menunjukkan celah penelitian yang belum terjawab oleh studi sebelumnya.

Pada penelitian yang dilakukan oleh Wiharto dan Suryani dengan judul *The Comparison of Clustering Algorithms K-Means and Fuzzy C-Means for Segmentation Retinal Blood Vessels*, dibahas perbandingan kinerja algoritma K-Means dan *Fuzzy C-Means* pada tugas segmentasi pembuluh darah retina menggunakan dataset DRIVE dan STARE. Hasil penelitian tersebut menunjukkan bahwa *Fuzzy C-Means* memberikan performa yang secara signifikan lebih baik dibanding K-Means berdasarkan metrik Sensitivity, Specificity, Accuracy, dan AUC, sehingga menguatkan alasan pemilihan FCM sebagai metode *clustering* pada data medis yang memiliki batas kelas tidak tegas.

Zainul Arifin dalam penelitiannya yang berjudul Penerapan Algoritma Support Vector Machine Berbasis Kernel Radial Basis Function dalam Klasifikasi Sel Kanker menggunakan algoritma Support Vector Machine (SVM) dengan kernel Radial Basis Function (RBF) untuk mengklasifikasikan sel kanker berdasarkan dataset Breast Cancer Wisconsin. Dataset tersebut terdiri atas 699 sampel sel dengan 10 fitur morfologis yang dikategorikan sebagai jinak atau ganas. Setelah melalui pra-pemrosesan dan optimasi parameter menggunakan grid search, hasil pengujian menunjukkan bahwa SVM-RBF mampu mencapai akurasi sebesar 96%, disertai nilai precision, recall, dan F1-score yang tinggi. Penelitian ini mengindikasikan bahwa SVM dengan kernel RBF merupakan metode yang efektif dan andal sebagai pendukung diagnosis kanker berbasis pembelajaran mesin.

Jatnika dalam penelitiannya yang berjudul "A Comparative Study on Data Collection Methods: Investigating Optimal Datasets for Data Mining Analysis" membandingkan kedua metode pengumpulan data, yaitu kuesioner dan web mining, dalam konteks analisis data mining dengan menggunakan algoritma Support Vector Machine (SVM) dan Naive Bayes Classifier. Hasil analisis menunjukkan bahwa penggunaan data berdasarkan

kuesioner memberikan akurasi SVM sebesar 82,39% dengan nilai AUC sebesar 0,911, yang lebih baik dibandingkan data yang diperoleh dari web mining. Temuan ini menunjukkan bahwa sifat dan tingkat kualitas dataset sangat memengaruhi hasil kerja model SVM saat melakukan proses klasifikasi.

Nandan N dalam penelitiannya berjudul Machine Learning Approach for Parkinson's Disease Detection: A Comparative Study of SVM Kernels on DaTSCAN Data membandingkan kinerja sejumlah kernel Support Vector Machine (SVM) untuk mendeteksi penyakit Parkinson menggunakan data DaTSCAN berbasis nilai Striatal Binding Ratio. Hasil analisis menunjukkan bahwa SVM dengan kernel Radial Basis Function (RBF) mencapai akurasi tertinggi sebesar 98,12%, melebihi kernel Polynomial dan Sigmoid serta algoritma lain seperti Random Forest, Logistic Regression, K-Nearest Neighbors (KNN), dan Convolutional Neural Network (CNN). Temuan ini menegaskan keunggulan SVM-RBF dalam menangani pola data nonlinier pada deteksi dini Parkinson.

Utami melakukan penelitian berjudul *Data Clustering of Confirmed COVID-19 Patients Using Fuzzy C-Means* yang memanfaatkan FCM untuk mengelompokkan pasien COVID-19 berdasarkan tingkat keparahan gejala. Penelitian ini menunjukkan bahwa FCM mampu membentuk dua cluster utama dengan akurasi sekitar 80% dan menggambarkan perbedaan pola gejala dan kelompok umur, sehingga menguatkan penggunaan FCM untuk mengelompokkan kasus penyakit berdasarkan karakteristik klinis. Penerapan *Fuzzy C-Means* untuk mengelompokkan pasien berdasarkan gejala ini sangat relevan dengan teori *clustering* berbasis himpunan fuzzy dan menunjukkan bahwa FCM mampu mengungkap pola tingkat keparahan penyakit, sebagaimana penelitian ini memanfaatkannya untuk memetakan kelompok risiko hipertensi.

Rahmanesta dalam penelitian *Implementation of Fuzzy C-Means Algorithm for Clustering Provinces in Indonesia Based on Micro and Small Industry Ratio in Village Areas* menggunakan FCM untuk mengelompokkan provinsi di Indonesia berdasarkan rasio industri mikro dan kecil di wilayah desa. Hasil penelitian menunjukkan terbentuknya empat cluster dengan nilai *Silhouette Coefficient* 0,6406 yang menandakan kualitas pengelompokan yang baik, dan memberi bukti bahwa FCM efektif digunakan untuk analisis spasial berbasis wilayah. Penggunaan FCM untuk mengelompokkan provinsi berdasarkan indikator ekonomi ini sangat relevan dengan konsep analisis spasial dan *soft clustering*, serta menjadi rujukan bahwa FCM dapat digunakan untuk pengelompokan

wilayah, yang sejalan dengan tujuan penelitian ini dalam memetakan wilayah risiko hipertensi.

Malika Dakhola Jannah dalam penelitiannya yang berjudul Klasifikasi Penyakit Hipertensi Menggunakan *Support Vector Machine* dan *Naive Bayes* ini membandingkan algoritma *Support Vector Machine (SVM)* dan *Naive Bayes* untuk klasifikasi status hipertensi berdasarkan data rekam medis pasien. Dataset terdiri dari 1.898 sampel dari Puskesmas di Karawang, yang telah melalui pra-pemrosesan seperti normalisasi dan pembagian data latih-uji. Hasil evaluasi menunjukkan SVM mencapai akurasi 95,71% dengan F1-score 89,00%, sementara Naive Bayes memperoleh akurasi 93,37% dengan F1-score 84,68%, sehingga SVM lebih unggul dalam keseimbangan presisi dan recall. Penelitian ini memberikan wawasan tentang efektivitas algoritma klasifikasi dalam mendukung deteksi dini hipertensi di layanan kesehatan primer, menegaskan pentingnya pemilihan model yang sesuai untuk aplikasi medis berbasis pembelajaran mesin.

Bangun dalam *Support Vector Machine for Classifying Heart Failure, Hypertension, and Normal Heart Condition* mengembangkan model SVM untuk mengklasifikasikan tiga kondisi kardiovaskular, yaitu gagal jantung, hipertensi, dan kondisi jantung normal. Penelitian ini menunjukkan bahwa SVM mampu memberikan nilai akurasi, presisi, dan recall yang tinggi secara konsisten pada data klinis, sehingga mendukung penggunaan SVM dalam sistem pendukung keputusan medis multikelas.

Fita Sheila Gomiasti dalam penelitiannya yang berjudul *Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine* menggunakan algoritma Support Vector Machine (SVM) dengan penyetelan hyperparameter untuk mengklasifikasikan kanker paru-paru berdasarkan data Lung Cancer Survey. Setelah proses pra-pemrosesan dan penyeimbangan data, SVM dengan kernel RBF yang dioptimasi menggunakan Random Grid Search memberikan hasil kinerja terbaik, yaitu akurasi sebesar 99% dan F1-score 99%, serta menunjukkan performa yang lebih baik dibandingkan SVM tanpa tuning maupun algoritma klasifikasi lainnya. Temuan ini menegaskan bahwa optimasi parameter pada SVM efektif untuk meningkatkan akurasi klasifikasi pada aplikasi medis berbasis pembelajaran mesin.

Penelitian Nisa berjudul *SVM Optimization for Autism Spectrum Disorder Classification: A Comparison of PCA, PSO, and Grid Search* memang tidak spesifik membahas hipertensi, namun relevan dari sisi optimasi SVM. Studi ini menunjukkan

bahwa pendekatan *Grid Search* tanpa PCA mampu menghasilkan akurasi hingga 98,2% dengan AUC 0,997, sehingga memberikan referensi teknis untuk strategi optimasi SVM yang juga dapat diadaptasi dalam konteks klasifikasi hipertensi.

able 2.1Kajian Pustaka Relevan

No	Judul	Metode	Hasil	Perbandingan
1	*The Comparison of Clustering Algorithms K-Means and Fuzzy C Means for Segmentation Retinal Blood Vessels* (Wiharto & Suryani, 2020)	Preprocessing (green channel, CLAHE, median filter), segmentasi (clustering dan thresholding menggunakan mean & median), serta analisis performa menggunakan uji statistik dengan tingkat kepercayaan 95%.	*Fuzzy C-Means* menunjukkan performa segmentasi yang secara signifikan lebih baik dibanding K-Means (p-value < 0,05) pada dataset DRIVE dan STARE, dengan AUC berada di kisaran 70–80%.	Penelitian ini menerapkan metode FCM pada data kasus hipertensi untuk menganalisis pengelompokan wilayah, sementara penelitian sebelumnya menggunakan FCM pada data citra medis.
2	Penerapan *Algoritma Support Vector Machine* Berbasis Kernel *Radial Basis Function* dalam Klasifikasi Sel Kanker (Arifin., 2023)	*Support Vector Machine*	Model SVM-RBF mampu mengklasifikasikan sel kanker dengan akurasi 96%, serta menunjukkan nilai precision, recall, dan F1-score yang tinggi.	Penelitian sebelumnya menggunakan metode SVM-RBF pada dataset publik yang memiliki banyak fitur morfologis untuk melakukan klasifikasi sel kanker. Sementara itu, penelitian ini menerapkan SVM-RBF pada data klinis dari layanan

No	Judul	Metode	Hasil	Perbandingan
				kesehatan primer dengan variabel tekanan darah dan usia untuk melakukan klasifikasi tingkat hipertensi.
3	*A Comparative Study on Data Collection Methods: Investigating Optimal Datasets for Data Mining Analysis (Jatnika., 2024)*	SVM dan Naive Bayes	SVM pada data berbasis kuesioner menghasilkan akurasi 82,39% dengan AUC 0,911, lebih tinggi dibandingkan data web mining..	Penelitian tersebut membandingkan metode pengumpulan data untuk analisis klasifikasi menggunakan SVM dan Naive Bayes. Sementara itu, penelitian ini berfokus pada klasifikasi tingkat hipertensi menggunakan SVM pada data klinis nyata di fasilitas kesehatan primer.
4	*Machine Learning Approach for Parkinson's Disease Detection: A Comparative*	*SVM (RBF, Polynomial, Sigmoid)*	SVM dengan kernel RBF memberikan performa terbaik dengan akurasi 98,12%, mengungguli	Penelitian sebelumnya mengulas beberapa kernel SVM yang digunakan untuk mendeteksi penyakit

No	Judul	Metode	Hasil	Perbandingan
	Study of SVM Kernels on DaTSCAN Data (Nandan, 2025)		kernel lain serta Random Forest, Logistic Regression, KNN, dan CNN.	Parkinson. Penelitian ini menggunakan kernel RBF untuk mengklasifikasikan tingkat hipertensi sesuai dengan karakteristik data yang digunakan.
5	*Data Clustering of Confirmed COVID-19 Patients Using Fuzzy C-Means (Utami, 2023)*	*Fuzzy C-Means* Clustering	Model clustering FCM berhasil membedakan dua cluster utama (terkonfirmasi & tidak terkonfirmasi COVID-19) dengan akurasi 80%. Cluster 1 didominasi oleh pasien dengan gejala di atas level 5, usia 21–40 tahun, sementara cluster 0 gejalanya lebih ringan dan rata-rata usia 18–43 tahun.	Penelitian tersebut menggunakan FCM untuk mengelompokkan pasien COVID-19 berdasarkan tingkat gejala. Sementara itu, penelitian ini menerapkan FCM untuk mengelompokkan wilayah berdasarkan tingkat risiko hipertensi.
6	*Implementation of Fuzzy C-Means Algorithm for Clustering Provinces in*	*Fuzzy C-Means*	*Fuzzy C-Means* menghasilkan 4 cluster provinsi dengan nilai *Silhouette*	Penelitian tersebut menerapkan FCM pada data indikator ekonomi untuk pengelompokan

No	Judul	Metode	Hasil	Perbandingan
	Indonesia Based on Micro and Small Industry Ratio in Village Areas (Rahmanesta, 2024)		*Coefficient* 0,6406, yang menunjukkan hasil pengelompokan baik dan terstruktur.	wilayah. Sedangkan penelitian ini menggunakan FCM pada data kesehatan untuk mengelompokkan wilayah berdasarkan distribusi kasus hipertensi.
7	Klasifikasi Penyakit Hipertensi Menggunakan Support Vector Machine dan Naive Bayes (Jannah, 2025)	*Support Vector Machine (SVM) dan Naive Bayes*	SVM menunjukkan performa lebih unggul dengan akurasi 95,71% dan F1-score 89,00% dibandingkan Naive Bayes, sehingga lebih seimbang dalam presisi dan recall untuk klasifikasi hipertensi.	Penelitian tersebut menggunakan 20 variabel klinis dan menerapkan klasifikasi biner (hipertensi dan tidak hipertensi). Sementara itu, penelitian ini menggunakan variabel utama tekanan darah dan usia, serta menerapkan klasifikasi lima tingkat hipertensi.
8	*Support Vector Machine for Classifying Heart Failure, Hypertension,*	*Support Vector Machine*	Model SVM berhasil mengklasifikasikan tiga kondisi kardiovaskular	Penelitian tersebut mengklasifikasikan tiga kondisi kardiovaskular menggunakan

No	Judul	Metode	Hasil	Perbandingan
	and Normal Heart Condition (Bangun, 2025)		dengan tingkat akurasi, presisi, dan recall yang tinggi serta menunjukkan kinerja yang stabil dalam validasi silang.	SVM. Sedangkan penelitian ini secara khusus mengklasifikasikan tingkat hipertensi dalam lima kategori tingkat keparahan.
9	*Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine (Gomiasti, 2024)*	SVM (kernel RBF) dengan hyperparameter tuning (Random Grid Search)	SVM yang dioptimasi menghasilkan akurasi 99% dan F1-score 99%, serta mengungguli SVM tanpa tuning dan algoritma klasifikasi lain, menunjukkan pentingnya optimasi parameter.	Penelitian tersebut berfokus pada optimasi hyperparameter SVM untuk meningkatkan akurasi klasifikasi kanker paru-paru. Sementara itu, penelitian ini menerapkan SVM pada data layanan kesehatan tingkat pertama dengan penanganan ketidakseimbangan kelas menggunakan SMOTE.
10	SVM Optimization for Autism Spectrum	*Support Vector Machine*, Particle	SVM Grid Search (tanpa PCA) unggul: akurasi	Penelitian tersebut membandingkan metode optimasi

No	Judul	Metode	Hasil	Perbandingan
	Disorder Classification: A Comparison of PCA, PSO, and Grid Search (Nisa et al., 2025)	Swarm Optimization, Grid Search	98,2%, AUC 0,997, efisiensi tinggi	parameter SVM seperti PCA, PSO, dan Grid Search. Sedangkan penelitian ini berfokus pada penerapan SVM untuk klasifikasi tingkat hipertensi dengan 5 kelas.
11	Penerapan Metode Fuzzy C-Means dan Support Vector Machine Untuk Klasifikasi Tingkat Hipertensi Pasien Puskesmas Pangkalan Balai	*Fuzzy C-Means, Support Vector Machine*		Penelitian sebelumnya sebagian besar menggunakan klasifikasi biner atau tiga kelas, sedangkan penelitian ini menerapkan klasifikasi lima tingkat hipertensi sehingga menghasilkan pembagian tingkat risiko yang lebih detail.

2.2 Landasan Teori

2.2.1 Hipertensi

Tekanan darah tinggi adalah kondisi medis yang terjadi ketika tekanan darah seseorang terus-menerus berada di atas tingkat normal. Menurut klasifikasi dari World Health Organization (WHO) dan Kementerian Kesehatan Republik Indonesia

(Kemenkes RI), seseorang dikatakan mengalami hipertensi jika tekanan darah sistoliknya mencapai atau melebihi 140 mmHg dan/atau tekanan darah diastoliknya mencapai atau melebihi 90 mmHg (WHO, 2024). Kondisi ini merupakan salah satu faktor risiko utama penyakit kardiovaskular seperti stroke, gagal jantung, dan penyakit ginjal kronis (Bangun et al., 2025a).

Penyebab hipertensi bisa dibagi menjadi dua jenis utama. Yang pertama adalah hipertensi primer, yang tidak memiliki penyebab spesifik, tetapi berkaitan dengan gaya hidup, makanan yang dikonsumsi, serta faktor keturunan. Jenis kedua adalah hipertensi sekunder, yang disebabkan oleh kondisi medis lain seperti gangguan pada ginjal atau masalah pada kelenjar endokrin.

Dalam penelitian ini, data pasien penderita hipertensi yang diperoleh dari Puskesmas Pangkalan Balai dimanfaatkan sebagai sumber utama dalam membangun sistem klasifikasi yang bertujuan untuk mengidentifikasi tingkat hipertensi secara komputasional melalui penerapan metode data mining.

2.2.2 Data Mining

Data mining adalah proses untuk mencari dan menemukan pola yang tersembunyi dalam kumpulan data besar. Proses ini menggunakan berbagai metode seperti statistik, matematika, dan teknik kecerdasan buatan. Data mining merupakan bagian dari Knowledge Discovery in Databases (KDD), yang mencakup beberapa tahap, seperti memilih data, mempersiapkan data (preprocessing), mengubah data, menggunakan algoritma untuk menemukan pola atau model, serta mengevaluasi dan memahami hasil yang diperoleh.

Tahapan umum data mining meliputi:

a. Data Selection – memilih data relevan dari sumber data yang tersedia.
b. Preprocessing – membersihkan data dari noise, missing value, dan inkonsistensi.
c. Transformation – mengubah data ke dalam format yang sesuai untuk analisis.
d. Data Mining Process – menerapkan algoritma tertentu untuk menemukan pola atau model.
e. Evaluation & Interpretation – menilai performa model dan menginterpretasikan hasilnya.

Dalam penelitian ini, data mining digunakan untuk mengklasifikasi tingkat hipertensi pasien dengan mengombinasikan metode *Fuzzy C-Means* untuk proses pengelompokan awal dan *Support Vector Machine* untuk proses klasifikasi akhir.

2.2.3 Klasifikasi

Klasifikasi merupakan salah satu teknik utama dalam *supervised learning* pada data mining, di mana model dilatih menggunakan data yang telah berlabel untuk mempelajari hubungan antara atribut input dan kelas output(Setiawan et al., n.d.). Setelah model selesai dilatih, sistem dapat mengklasifikasikan data baru ke dalam kelas tertentu berdasarkan pola yang telah dipelajari dari data sebelumnya(Pratiwi et al., 2025).

Dalam bidang kesehatan, metode klasifikasi sering digunakan untuk membantu proses diagnosis penyakit dan penentuan tingkat keparahan suatu kondisi berdasarkan variabel medis seperti usia, jenis kelamin, tekanan darah sistolik, dan tekanan darah diastolik. Beberapa algoritma klasifikasi yang umum digunakan antara lain *Decision Tree*, *K-Nearest Neighbor (KNN)*, *Naïve Bayes*, serta *Support Vector Machine (SVM)(Setiawan et al., n.d.)*.

Pada penelitian ini, metode klasifikasi digunakan untuk menentukan tingkat hipertensi pasien berdasarkan atribut usia, tekanan darah sistolik, dan tekanan darah diastolik dengan menggunakan algoritma Support Vector Machine (SVM). Proses klasifikasi dilakukan secara terpisah dari proses clustering. Sementara itu, metode Fuzzy C-Means (FCM) digunakan untuk mengelompokkan wilayah kelurahan berdasarkan distribusi kasus hipertensi. Dengan pemisahan pendekatan tersebut, analisis dapat dilakukan secara lebih terfokus sesuai karakteristik masing-masing metode.

2.2.4 *Clustering*

Clustering dalam data mining merupakan suatu metode analisis yang bertujuan mengelompokkan sekumpulan objek ke dalam beberapa cluster berdasarkan kedekatan karakteristiknya tanpa membutuhkan informasi kelas terlebih dahulu. Teknik ini digunakan untuk mempartisi data ke dalam kelompok-kelompok yang seragam di dalam cluster namun berbeda satu sama lain antarcluster, sehingga pola dan struktur tersembunyi di dalam data dapat dikenali dengan lebih mudah. Dalam bidang kesehatan masyarakat, clustering sering dimanfaatkan untuk membagi wilayah atau kelompok

pasien menurut indikator kejadian suatu penyakit sehingga terbentuk kelompok berisiko rendah, sedang, dan tinggi yang dapat dijadikan acuan penentuan prioritas program penanggulangan. Di antara berbagai metode clustering, *Fuzzy C-Means* (FCM) merupakan algoritma yang populer karena bekerja dengan konsep himpunan fuzzy, di mana setiap objek memiliki derajat keanggotaan terhadap seluruh cluster, sehingga lebih sesuai untuk menggambarkan ketidakpastian(Purba et al., 2025).

2.2.5 *Fuzzy C-Means (FCM*)

Fuzzy C-Means (FCM) adalah algoritma *clustering* berbasis teori himpunan fuzzy yang mengelompokkan data ke dalam sejumlah klaster dengan memberikan derajat keanggotaan pada setiap data untuk tiap klaster. Dalam FCM, satu objek tidak langsung dipandang sebagai anggota mutlak satu klaster, melainkan memiliki tingkat keanggotaan bernilai antara 0 sampai 1 pada beberapa klaster, sehingga struktur klaster yang terbentuk mampu merepresentasikan ketidakpastian dan tumpang tindih antar kelompok pada data dunia nyata(Nugraha et al., 2023).

Secara konsep, langkah utama algoritma FCM:

1. Menentukan jumlah klaster C, pangkat fuzzy m>1, nilai maksimum iterasi, dan batas error konvergensi.
2. Menginisialisasi matriks keanggotaan awal u_{ij} secara acak (derajat keanggotaan data *i* terhadap klaster *j*).
3. Menghitung pusat klaster (centroid) berbobot fuzzy:

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{2.1}$$

   Keterangan:
   a. c_j: pusat (centroid) dari cluster ke-j
   b. x_i: data ke-i
   c. u_{ij}: derajat keanggotaan data ke-i terhadap cluster ke-j
   d. m: *fuzziness coefficient* (biasanya $m > 1$)
   e. N: jumlah total data
   f. i: indeks data
   g. j: indeks cluster
4. Menghitung ulang derajat keanggotaan:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C} \left(\frac{|x_i - c_j|}{|x_i - c_k|} \right)^{\frac{2}{m-1}}} \quad (2,2)$$

Keterangan:

a. : derajat keanggotaan data ke-i pada cluster ke-j
b. x_i: data ke-i
c. c_j: pusat cluster ke-j
d. c_k: pusat cluster ke-k
e. C: jumlah cluster
f. m: *fuzziness coefficient*
g. $\| x_i - c_j \|$: jarak antara data ke-i dan pusat cluster ke-j (umumnya jarak Euclidean)
h. k: indeks cluster pembanding

5. Menghitung fungsi objektif utama FCM:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} ||x_i - c_j||^2 \quad (2,3)$$

Keterangan:

a. J_m: fungsi objektif FCM
b. u_{ij}: derajat keanggotaan data ke-i pada cluster ke-j
c. m: *fuzziness coefficient*
d. x_i: data ke-i
e. c_j: pusat cluster ke-j
f. N: jumlah total data
g. C: jumlah cluster

6. Mengecek konvergensi: jika perubahan J_m atau matriks keanggotaan sudah sangat kecil, iterasi dihentikan; jika belum, kembali ke langkah 3.

2.2.6 *Support Vector Machine (SVM)*

Support Vector Machine (SVM) adalah algoritma pembelajaran yang didasarkan pada pengawasan (supervised learning), yang digunakan untuk tugas

klasifikasi maupun regresi(Jannah et al., 2025). Pokok utama dari SVM adalah mencari sebuah garis pemisah (hyperplane) yang bisa memisahkan data dari berbagai kelas dengan jarak pemisah (margin) yang terbesar. Konsep ini pertama kali dikembangkan oleh Vapnik pada tahun 1995, dan sejak itu menjadi salah satu metode yang sangat efektif dalam pengenalan pola dan klasifikasi data yang bersifat non-linear.

1) Bentuk linear SVM

Secara matematis, bentuk umum *hyperplane* pada SVM dapat dituliskan sebagai:

$$f(x) = w^T x + b \tag{2.4}$$

dengan:

a. w= vektor bobot,

b. x= data input,

c. b= bias.

Hyperplane optimal diperoleh dengan meminimalkan norma bobot agar margin maksimum, yang dirumuskan sebagai:

$$\min_{w,b} \quad \frac{1}{2} \| w \|^2 \tag{2.5}$$

dengan kendala:

$$y_i(w^T x_i + b) \geq 1, i = 1,2, \dots, n \tag{2.6}$$

di mana:

a. x_i: data ke-i,

b. y_i: label kelas (misal −1 atau +1),

c. n: jumlah data.

Formulasi ini menunjukkan bahwa SVM mencari batas pemisah dengan margin terbesar sekaligus menjaga agar data terklasifikasi dengan benar.

2) Bentuk dengan Kernel (Untuk Data Non-Linear)

Jika data tidak dapat dipisahkan secara linear, SVM menggunakan **fungsi kernel** untuk memetakan data ke ruang berdimensi lebih tinggi tanpa menghitung transformasi secara eksplisit (*kernel trick*). Dalam bentuk dual, fungsi keputusan SVM dinyatakan sebagai:

$$f(x) = \sum_{i=1}^{n} \alpha_i \, y_i K(x_i, x) + b \tag{2.7}$$

dengan:

a. α_i: koefisien Lagrange,

b. y_i: label kelas dari support vector,

c. x_i: support vector,

d. $K(x_i, x)$: fungsi kernel,

e. b: bias.

Melalui pendekatan ini, SVM dapat membentuk batas pemisah yang lebih fleksibel sesuai karakteristik data.

2.2.7 *Radial Basis Function (RBF).*

Dalam banyak kasus, data tidak dapat dipisahkan secara linier pada data awal. Oleh karena itu, SVM memanfaatkan fungsi kernel untuk memetakan data ke ruang berdimensi lebih tinggi tanpa harus menghitung transformasi secara eksplisit. Salah satu kernel yang paling umum digunakan adalah Radial Basis Function (RBF).

Kernel RBF memiliki kemampuan yang baik dalam menangani pola data non-linier karena bersifat lokal dan sensitif terhadap jarak antar data. Fungsi kernel RBF didefinisikan sebagai:

$$K(x_i, x_j) = \exp\left(-\gamma \parallel x_i - x_j \parallel^2\right) \tag{2,8}$$

dengan:

$K(x_i, x_j)$: nilai kemiripan antara dua data

x_i, x_j: vektor data

γ: parameter kernel yang mengatur tingkat sensitivitas jarak

$\parallel x_i - x_j \parallel^2$: jarak kuadrat antar data

Nilai parameter γsangat berpengaruh terhadap performa model. Nilai γyang terlalu besar dapat menyebabkan *overfitting* karena model menjadi terlalu sensitif terhadap data latih, sedangkan nilai yang terlalu kecil dapat menyebabkan *underfitting* karena model kurang mampu menangkap pola data.

Berdasarkan hasil penelitian pada jurnal rujukan, penggunaan kernel RBF pada SVM memberikan kemampuan pembelajaran yang tinggi serta akurasi prediksi yang baik pada data yang memiliki hubungan non-linier. Hal ini menjadikan kernel RBF sebagai pilihan yang umum dalam pemodelan klasifikasi di berbagai bidang penelitian(Gao et al., 2022).

2.2.8 *Synthetic Minority Over-Sampling Technique* (SMOTE)

Synthetic Minority Over-Sampling Technique (SMOTE) merupakan salah satu metode *data preprocessing* yang digunakan untuk menangani permasalahan ketidakseimbangan kelas *(class imbalance)* pada dataset. Ketidakseimbangan kelas terjadi ketika jumlah data pada satu kelas jauh lebih sedikit dibandingkan kelas lainnya, sehingga model pembelajaran mesin cenderung lebih bias terhadap kelas mayoritas dan menghasilkan performa prediksi yang kurang optimal.

SMOTE bekerja dengan cara menambahkan data sintetis pada kelas minoritas, bukan dengan menggandakan data yang sudah ada. Data sintetis tersebut dibentuk berdasarkan kedekatan antar data minoritas menggunakan perhitungan jarak, umumnya jarak Euclidean. Dengan pendekatan ini, data baru yang dihasilkan tetap memiliki karakteristik yang mirip dengan data asli, namun tidak identik, sehingga mampu mengurangi risiko *overfitting*.

Secara konseptual, pembentukan data sintetis pada SMOTE dilakukan dengan rumus:

$$x_{baru} = x_i + \lambda(x_{nn} - x_i) \tag{2.9}$$

Keterangan:

x_i: data minoritas asli

x_{nn}: data minoritas tetangga terdekat (*nearest neighbor*)

λ: bilangan acak antara 0 dan 1

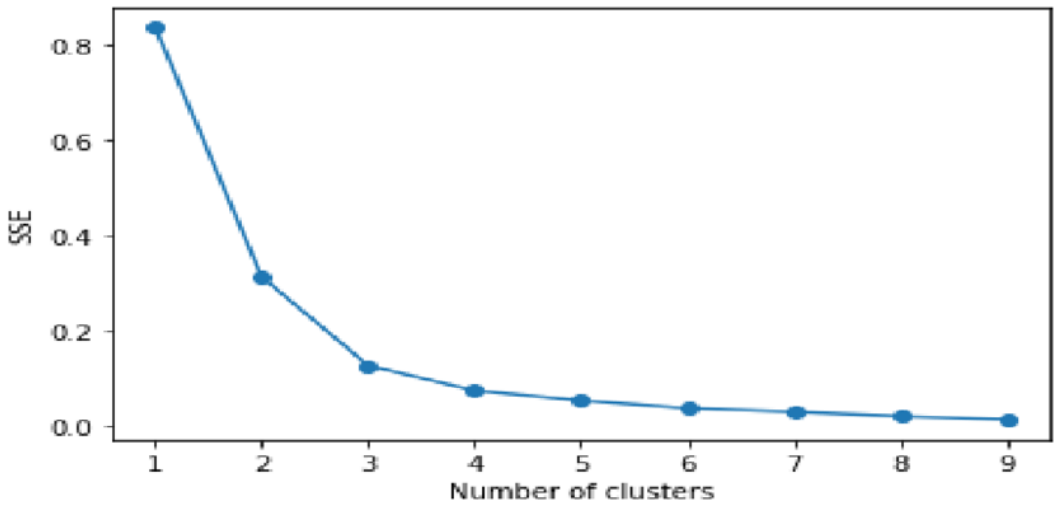
x_{baru}: data sintetis yang dihasilkan

Melalui persamaan tersebut, SMOTE membentuk titik data baru di antara dua data minoritas yang berdekatan, sehingga distribusi data menjadi lebih merata. Teknik ini banyak diterapkan pada bidang kesehatan dan biomedis karena sering dijumpai kondisi dataset yang tidak seimbang, seperti jumlah pasien sehat yang jauh lebih banyak dibandingkan pasien dengan kondisi tertentu.

Berdasarkan penelitian pada jurnal rujukan, penerapan SMOTE mampu meningkatkan performa beberapa algoritma klasifikasi, khususnya pada model berbasis pohon (*tree-based classifiers*), dengan cara memperbaiki distribusi data latih sebelum proses pelatihan model dilakukan(Ishaq et al., 2021).

2.2.9 *Elbow Method*

Dalam analisis clustering, Elbow method digunakan untuk menentukan jumlah cluster k yang paling sesuai dengan cara melihat perubahan nilai Sum of Squared Error (SSE) untuk beberapa kandidat nilai k. Pada setiap nilai k, algoritma clustering dijalankan lalu dihitung nilai SSE, kemudian hubungan antara k dan SSE digambarkan dalam bentuk grafik untuk mengidentifikasi titik jumlah cluster yang memberikan penurunan SSE paling signifikan.



Gambar 2.1 Elbow Method

Pada algoritma *Fuzzy C-Means*, SSE yang digunakan dalam Elbow method dinyatakan oleh fungsi objektif berikut:

$$SSE_{fuzzy} = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}|x_i - c_j|^2 \qquad (2.10)$$

Keterangan:

1. SSE_{fuzzy} : nilai Sum of Squared Error pada algoritma *Fuzzy C-Means*.
2. N : jumlah data.
3. C : jumlah cluster.
4. x_i : data ke-i.
5. c_j : pusat (centroid) cluster ke-j.
6. u_{ij} : derajat keanggotaan data x_i terhadap cluster ke-j.
7. m : parameter fuzzifier pada *Fuzzy C-Means*.

Dalam penelitian ini, Elbow method diterapkan dengan cara menjalankan algoritma *Fuzzy C-Means* untuk beberapa nilai k (misalnya $k = 2,3,4,5,6$), kemudian untuk setiap k dihitung nilai SSE_{fuzzy} sesuai rumus di atas. Nilai SSE tersebut disimpan dan diplot terhadap jumlah cluster sehingga membentuk kurva penurunan SSE; titik "siku" (elbow) ditentukan pada posisi di mana penurunan SSE mulai melambat ketika nilai k ditambah, dan titik ini dipilih sebagai jumlah cluster yang dianggap optimal (Nugraha et al., 2023).

2.2.10 Davies-Bouldin Index (DBI)

Davies–Bouldin Index (DBI) merupakan indeks validitas internal yang digunakan untuk mengevaluasi kualitas cluster berdasarkan nilai separation dan cohesion(Rochman et al., 2022). Indeks ini menilai seberapa kompak data di dalam setiap cluster dan seberapa besar pemisahan antarcluster, sehingga dapat digunakan untuk membandingkan beberapa skenario jumlah cluster dan memilih konfigurasi terbaik. Nilai DBI yang lebih kecil menunjukkan bahwa cluster yang terbentuk semakin baik karena sebaran data di dalam cluster relatif kecil dan jarak antar centroid cluster cukup besar.

Secara matematis, DBI didefinisikan sebagai

$$DBI = \frac{1}{k}\sum_{i=1}^{k} R_i \tag{2.11}$$

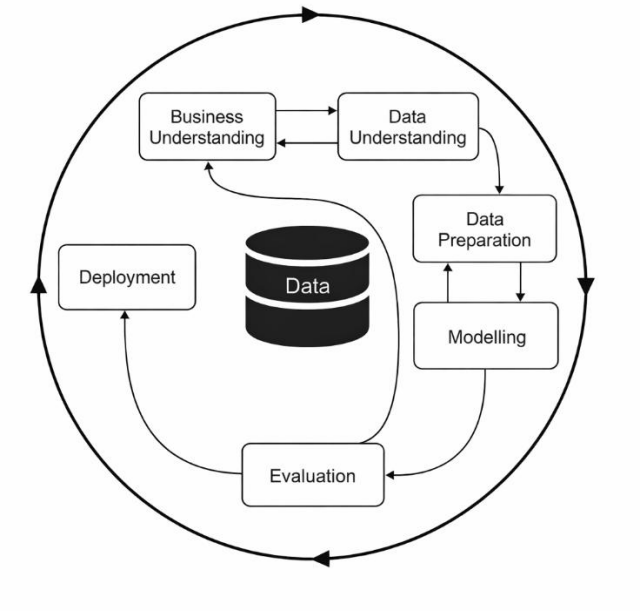
Sebagai

$$R_i = \max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \quad (2.12)$$

di mana k adalah jumlah cluster, S_i menyatakan dispersi dalam cluster i (misalnya rata-rata jarak Euclidean antara titik-titik dalam cluster i terhadap centroid c_i), dan M_{ij} adalah jarak antara centroid cluster i dan centroid cluster j. Nilai R_i merepresentasikan rasio kemiripan terburuk antara cluster i dan cluster lain, sehingga DBI merupakan rata-rata dari rasio kemiripan terburuk seluruh cluster.

2.2.11 Cross Industry Standard Process for Data Mining (CRISP DM)

CRISP-DM (Cross-Industry Standard Process for Data Mining) adalah metodologi standar yang banyak digunakan untuk merancang dan melaksanakan proyek data mining maupun data science. Metodologi ini menyediakan kerangka kerja umum yang sistematis sehingga proses analisis data dapat berjalan terstruktur, mulai dari tahap perumusan masalah sampai pemanfaatan model di lingkungan operasional(Schröer et al., 2021), yang akan di gambarkan sebagai berikut :



Gambar 2 .2 CRISP-DM

Sumber: (Rianti et al., n.d.)

1. *Business Understanding*

   Tahap ini fokus memahami konteks dan tujuan bisnis secara jelas sebelum menyentuh data. Tujuannya supaya rumusan masalah data mining selaras dengan

kebutuhan organisasi, lengkap dengan batasan, asumsi, dan rencana proyek yang akan dijalankan.

2. *Data understanding*

   Pada tahap ini data yang relevan dikumpulkan, dideskripsikan, dan dieksplorasi secara awal. Aktivitas umum meliputi pengecekan struktur data, distribusi nilai, pola awal, serta identifikasi masalah kualitas data seperti missing value atau outlier.

3. *Data preparation*

   Tahap ini menyiapkan dataset akhir yang siap dipakai untuk pemodelan dan biasanya memakan porsi waktu paling besar. Kegiatannya mencakup seleksi variabel, pembersihan data, transformasi, penggabungan beberapa sumber data, sampai pembentukan format data yang sesuai untuk algoritma yang akan digunakan.

4. *Modeling*

   Pada tahap modeling dipilih teknik analisis atau algoritma yang sesuai (misalnya klasifikasi, regresi, clustering), lalu model dibangun dan disetel parameternya. Beberapa model biasanya dibandingkan untuk melihat mana yang memiliki performa terbaik terhadap tujuan yang telah ditetapkan.

5. *Evaluation*

   Tahap ini mengevaluasi model secara lebih menyeluruh, bukan hanya dari metrik teknis, tetapi juga kesesuaiannya dengan tujuan bisnis. Di sini diputuskan apakah model sudah layak digunakan, perlu perbaikan, atau harus kembali ke tahap sebelumnya untuk penyesuaian.

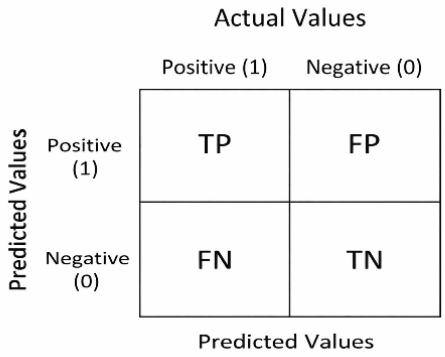
6. *Deployment*

   Tahap deployment menerapkan model atau insight ke lingkungan nyata, misalnya dalam bentuk laporan, dashboard, integrasi ke sistem operasional, atau layanan prediktif. Selain implementasi awal, tahap ini juga mencakup rencana monitoring dan pemeliharaan model karena data dan kondisi bisnis dapat berubah dari waktu ke waktu.

2.2.12 Confusion Matrix

Confusion matrix merupakan salah satu metode evaluasi yang digunakan untuk menilai kinerja model klasifikasi. Matriks ini menyajikan perbandingan antara kelas aktual dan kelas hasil prediksi model dalam bentuk tabel dua dimensi(Sathyanarayanan,

2024a). Melalui confusion matrix, peneliti dapat mengetahui jumlah prediksi yang benar maupun salah untuk setiap kelas secara terperinci.



Gambar 2 .3 Confusion Matrix

Sumber: (Hakim et al., 2025)

Secara umum, confusion matrix terdiri atas empat komponen utama, yaitu true positive (TP), true negative (TN), false positive (FP), dan false negative (FN). True positive menunjukkan jumlah data yang diprediksi positif dan memang benar positif, sedangkan true negative menunjukkan jumlah data yang diprediksi negatif dan memang benar negatif. False positive dan false negative merepresentasikan kesalahan prediksi model.

Berdasarkan komponen-komponen tersebut, berbagai metrik evaluasi dapat dihitung, seperti akurasi, presisi, dan recall. Akurasi menggambarkan proporsi prediksi yang benar terhadap keseluruhan data, presisi menunjukkan tingkat ketepatan model dalam memprediksi kelas positif, sedangkan recall menggambarkan kemampuan model dalam mendeteksi seluruh data yang benar-benar termasuk dalam kelas positif.

a. Accuracy

Accuracy merupakan metrik evaluasi yang mengukur proporsi prediksi yang benar dibandingkan dengan seluruh jumlah data. Accuracy dihitung dengan menjumlahkan True Positive dan True Negative, kemudian dibagi dengan total seluruh prediksi(Sathyanarayanan, 2024).

Rumus accuracy dinyatakan sebagai berikut:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.13}$$

Meskipun accuracy mudah dipahami dan sering digunakan, metrik ini memiliki keterbatasan pada dataset yang tidak seimbang (*imbalanced dataset*). Dalam

kondisi di mana salah satu kelas jauh lebih dominan, model dapat memperoleh nilai accuracy yang tinggi hanya dengan memprediksi kelas mayoritas, meskipun gagal mengidentifikasi kelas minoritas secara akurat. Fenomena ini dikenal sebagai *accuracy paradox*

b. Precision

Precision mengukur ketepatan prediksi pada kelas positif, yaitu proporsi data yang diprediksi positif dan benar-benar merupakan kelas positif.

Secara matematis, precision dirumuskan sebagai:

$$Precision = \frac{TP}{TP + FP} \tag{2.14}$$

Precision menjadi penting dalam situasi di mana kesalahan prediksi positif (false positive) memiliki konsekuensi besar. Precision yang tinggi menunjukkan bahwa model memiliki tingkat kesalahan false positive yang rendah.

Namun, precision tidak mempertimbangkan jumlah false negative, sehingga belum cukup untuk menggambarkan kemampuan model dalam menangkap seluruh kasus positif(Hinojosa Lee et al., 2024).

c. *Recall*

Recall, yang juga disebut sebagai *sensitivity* atau *true positive rate*, merupakan ukuran kemampuan model dalam mendeteksi seluruh data positif yang sebenarnya ada.

Rumus recall adalah sebagai berikut:

$$Recall = \frac{TP}{TP + FN} \tag{2.15}$$

Recall penting dalam kondisi di mana kesalahan tidak mendeteksi kasus positif (false negative) memiliki dampak serius. Nilai recall yang tinggi menunjukkan bahwa model mampu mengidentifikasi sebagian besar kasus positif yang tersedia dalam data.

Dalam penelitian evaluasi model pada dataset tidak seimbang, recall sering digunakan bersama precision untuk memberikan gambaran performa yang lebih adil terhadap kelas minoritas(Sujon et al., 2025).

d. F1-Score

F1-score merupakan metrik yang menggabungkan precision dan recall dalam satu ukuran tunggal dengan menggunakan rata-rata harmonis. F1-score

memberikan keseimbangan antara ketepatan prediksi positif (precision) dan kemampuan mendeteksi seluruh kasus positif (recall).

Rumus F1-score dinyatakan sebagai:

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall} \tag{2.16}$$

Dalam konteks dataset tidak seimbang, F1-score dianggap lebih representatif dibandingkan accuracy karena mempertimbangkan keseimbangan antara precision dan recall. Penelitian dalam *Computers & Geosciences* juga menyatakan bahwa kombinasi metrik seperti accuracy, precision, recall, dan F1-score memberikan evaluasi yang lebih informatif terhadap model, khususnya pada data yang memiliki distribusi kelas tidak merata(Conciatori et al., 2024).

2.2.13 Flowchart

Flowchart merupakan representasi grafis yang digunakan untuk menggambarkan urutan langkah atau alur suatu proses secara sistematis. Flowchart menyajikan tahapan proses dalam bentuk simbol-simbol standar yang saling terhubung oleh garis berarah, sehingga hubungan antar langkah dapat dipahami secara jelas dan terstruktur. Diagram ini membantu memvisualisasikan proses yang kompleks menjadi lebih sederhana dan logis.

Dalam perancangan sistem dan pengembangan perangkat lunak, flowchart digunakan untuk memodelkan logika proses sebelum sistem diimplementasikan. Flowchart mempermudah identifikasi urutan kegiatan, percabangan keputusan, serta kemungkinan keluaran dari setiap tahapan proses. Dengan demikian, flowchart berfungsi sebagai alat bantu analisis dan dokumentasi sistem (Zhang et al., 2023).

Selain sebagai alat dokumentasi, flowchart juga berperan dalam meningkatkan pemahaman terhadap struktur dan hubungan antar tahapan proses. Visualisasi berbasis flowchart memungkinkan proses yang kompleks ditampilkan secara lebih intuitif dan sistematis, sehingga memudahkan analisis dan evaluasi sistem.

Dalam konteks penelitian, flowchart sering digunakan untuk menggambarkan tahapan metodologi penelitian mulai dari pengumpulan data, pengolahan data, penerapan metode, hingga evaluasi hasil. Penyajian dalam bentuk flowchart membantu pembaca memahami alur penelitian secara menyeluruh dan terstruktur.

2.3 Kerangka Pemikiran

Gambar 2 .4 Kerangka Pemikiran

Kerangka pemikiran pada penelitian ini digunakan untuk memberikan gambaran alur penelitian yang dimulai dari tahap pengumpulan data hingga diperolehnya hasil analisis sesuai dengan tujuan penelitian. Kerangka ini disusun secara sistematis agar setiap tahapan penelitian saling berkaitan dan mendukung proses pengambilan keputusan.

Berikut penjelasan dari masing-masing komponen dalam kerangka pemikiran penelitian ini:

1. *Input*

   Input dalam penelitian ini berupa data rekam hipertensi tahun 2025 yang diperoleh dari fasilitas pelayanan kesehatan tingkat pertama, yaitu Puskesmas Pangkalan Balai. Data tersebut memuat informasi status hipertensi pasien yang kemudian dikelompokkan berdasarkan wilayah desa atau kelurahan. Data ini menjadi dasar utama dalam proses analisis lebih lanjut, baik untuk klasifikasi maupun clustering.

2. *Method*

   Metode yang digunakan dalam penelitian ini meliputi beberapa pendekatan, yaitu:

   a. *Support Vector Machine* (SVM)

      Metode SVM digunakan untuk melakukan klasifikasi status hipertensi pasien, sehingga dapat diketahui pola klasifikasi berdasarkan data yang tersedia.

   b. *Fuzzy C-Means* (FCM)

Metode FCM digunakan untuk melakukan clustering wilayah desa atau kelurahan berdasarkan jumlah kasus hipertensi. Metode ini dipilih karena mampu menangani data yang bersifat tidak tegas (fuzzy), di mana suatu wilayah dapat memiliki tingkat keanggotaan pada lebih dari satu cluster.

c. CRISP-DM

CRISP-DM digunakan sebagai kerangka kerja penelitian data mining yang mengatur seluruh tahapan penelitian, mulai dari pemahaman data, pengolahan data, pemodelan, evaluasi, hingga penyajian hasil.

d. Elbow Method

Elbow Method digunakan untuk membantu menentukan jumlah cluster yang optimal dengan melihat perubahan nilai fungsi objektif terhadap jumlah cluster yang dibentuk.

3. *Objectives*

Tujuan dari penelitian ini adalah untuk mengklasifikasikan status hipertensi pasien menggunakan metode SVM, mengelompokkan wilayah desa atau kelurahan berdasarkan tingkat risiko hipertensi menggunakan metode *Fuzzy C-Means*, dan Mengidentifikasi wilayah dengan tingkat kasus hipertensi tertinggi sebagai bahan pertimbangan dalam pengambilan keputusan di bidang kesehatan masyarakat.

4. Pengujian

Tahap pengujian dilakukan untuk mengevaluasi kinerja metode yang digunakan, yaitu:

a. Confusion Matrix digunakan untuk mengevaluasi hasil klasifikasi SVM.

b. Davies-Bouldin Index (DBI) digunakan untuk menilai kualitas hasil clustering *Fuzzy C-Means*, di mana nilai DBI yang lebih kecil menunjukkan kualitas cluster yang lebih baik.