BAB 2
LANDASAN TEORI

Pada penelitian ini digunakan beberapa teori dan metode yaitu Penyakit Jantung, Machine Learning, Supervised Learning, Standard Scaler, SMOTE, Support Vector Machine (SVM) dan Confusion Matrix.

2.1 Penyakit Jantung

Penyakit jantung koroner (PJK) berawal dari proses aterosklerosis, yaitu penumpukan lemak (kolesterol) di dinding arteri yang menyebabkan penyempitan pembuluh darah. Lemak dalam makanan diubah menjadi kolesterol dan trigliserida, lalu dibawa oleh lipoprotein seperti LDL (lemak jahat) ke seluruh tubuh. Kolesterol LDL yang berlebih dan teroksidasi akan menempel di dinding arteri dan memicu reaksi peradangan. Sel-sel kekebalan tubuh seperti monosit dan makrofag akan memakan LDL ini dan membentuk sel busa, yang menumpuk menjadi kerak lemak (*fatty streaks*) dan berkembang menjadi plak. Plak ini dapat pecah dan memicu pembentukan gumpalan darah (*trombus*), yang bisa menyumbat aliran darah ke jantung dan menyebabkan gejala seperti nyeri dada (*angina*) hingga serangan jantung (*infark miokard*). Intinya, penyakit jantung terjadi karena penyumbatan pembuluh darah oleh lemak yang menyebabkan aliran darah ke jantung terganggu, dan bisa berujung pada kondisi berbahaya atau kematian [23].

2.2 Machine Learning

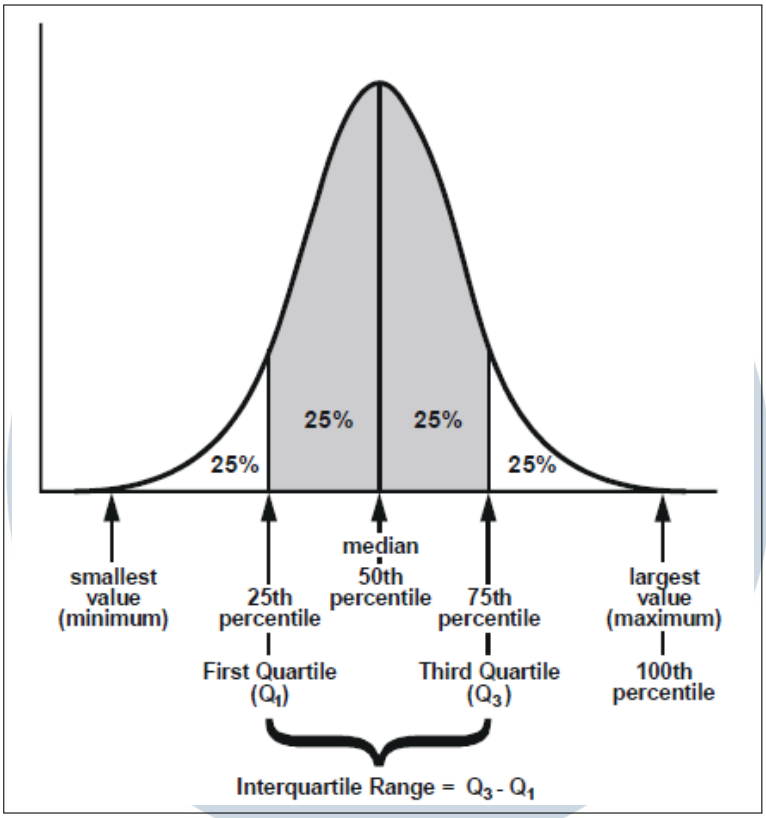
Machine learning (ML) adalah cabang dari kecerdasan buatan (AI) yang semakin digunakan dalam banyak bidang. Secara sederhana, *machine learning* memungkinkan komputer memahami data dan mengambil keputusan atau melakukan klasifikasi terhadap suatu tugas, baik dengan bantuan manusia maupun secara otomatis tanpa campur tangan langsung [24]. Dalam *machine learning*, pengalaman diperoleh dari berbagai informasi yang sudah tersedia dan digunakan sebagai dasar untuk menyelesaikan masalah di masa depan. Pembelajaran mesin sendiri terdiri dari beberapa metode, antara lain *Supervised Learning*, *Unsupervised Learning*, dan *Reinforcement Learning* [25]. Penelitian ini menggunakan metode *Supervised Learning*.

2.3 Supervised Learning

Supervised learning adalah metode machine learning di mana algoritma dilatih menggunakan data yang sudah memiliki label, yaitu input dan output yang telah ditentukan. Algoritma mempelajari hubungan antara input dan output tersebut untuk membuat prediksi terhadap data baru. Metode ini umumnya digunakan untuk klasifikasi dan regresi, seperti memprediksi nilai atau mengkategorikan data. Contoh algoritma yang digunakan dalam supervised learning antara lain K-Nearest Neighbors, Naïve Bayes, dan Decision Tree [26]. Berbeda dengan Unsupervised learning yang merupakan metode machine learning yang mencoba mengelompokkan data tanpa label atau target hasil. Data yang digunakan bersifat mentah dan tidak terstruktur, tanpa kriteria atau umpan balik yang jelas [27].*Supervised Learning* ini mirip seperti manusia belajar dari pengalaman masa lalu untuk meningkatkan kemampuan dalam menjalani tugas sehari-hari. Bedanya, komputer tidak memiliki pengalaman, jadi machine learning belajar dari data yang dikumpulkan di masa lalu sebagai representasi dari pengalaman tersebut [28].

2.4 *Interquartile Range (IQR)*

Interquartile Range (IQR) merupakan teknik yang digunakan untuk mendeteksi outlier pada data yang bersifat kontinu. IQR dihitung sebagai selisih antara kuartil ketiga (Q3) dan kuartil pertama (Q1). IQR juga dikenal dengan istilah midspread atau middle 50%, dan secara teknis disebut sebagai H-spread. Metode ini membagi data menjadi empat bagian yang sama besar berdasarkan urutan nilainya, dengan titik-titik pembaginya dikenal sebagai kuartil pertama (Q1), kuartil kedua (Q2), dan kuartil ketiga (Q3). IQR juga dapat divisualisasikan secara sederhana melalui diagram boxplot, yang memberikan gambaran tentang penyebaran data dan posisi nilai-nilai ekstrem [29]. Gambar 2.1. adalah visualisasi sederhana interquartile range



Gambar 2.1. Visualisasi Interquartile Range

Penggunaan metode IQR memiliki keunggulan karena bersifat non-parametrik, sehingga tidak bergantung pada asumsi distribusi data normal. Hal ini menjadikan IQR sangat efektif dalam mengidentifikasi nilai-nilai ekstrem yang dapat mengganggu analisis, terutama pada dataset yang memiliki sebaran data tidak simetris atau mengandung noise. Selain itu, metode ini cukup sederhana secara komputasional namun tetap memberikan hasil yang akurat dalam menyaring data yang menyimpang secara signifikan. Oleh karena itu, metode IQR menjadi salah satu pendekatan yang direkomendasikan dalam tahap pembersihan data untuk meningkatkan kualitas dan keandalan model pembelajaran mesin.

2.5 Standard Scaler

Banyak algoritma machine learning bekerja lebih baik ketika variabel input numerik diskalakan ke dalam rentang standar. Hal ini termasuk algoritma yang menggunakan penjumlahan berbobot dari input, seperti regresi linier, dan algoritma yang menggunakan ukuran jarak, seperti k-nearest neighbours dan SVM. Dua teknik paling populer untuk menskalakan data numerik sebelum pemodelan adalah

normalisasi dan standardisasi. Normalisasi menskalakan setiap variabel input secara terpisah ke dalam rentang 0–1, yaitu rentang nilai floating point di mana kita memiliki presisi paling tinggi. Standardisasi menskalakan setiap variabel input secara terpisah dengan cara mengurangkan nilai rata-rata (disebut centering) dan membaginya dengan standar deviasi, sehingga distribusi data bergeser memiliki rata-rata nol dan standar deviasi satu. Ada beberapa macam fitur scaling diantara lain Standard Scaler, The min-max scaler, Robust Scaler, Normalizer.

Dalam proyek ini digunakan teknik StandardScaler sebagai metode normalisasi fitur, karena model utama yang digunakan adalah Support Vector Machine (SVC) dengan kernel RBF, yang sensitif terhadap perbedaan skala antar fitur. StandardScaler efektif dalam membuat distribusi data memiliki rata-rata nol dan deviasi standar satu, yang sangat membantu dalam menjaga margin optimal pada pemisahan kelas. Selain itu, dibandingkan metode lain seperti MinMax atau Normalizer, StandardScaler memberikan performa yang lebih stabil pada data dengan distribusi yang relatif normal dan tidak terlalu terpengaruh oleh skew ringan.

Standard Scaler, atau yang lebih umum dikenal sebagai teknik standardisasi, mentransformasi dataset agar memiliki nilai rata-rata nol dan standar deviasi satu. Nilai yang telah ditransformasi, atau yang sering disebut sebagai z-score.

Dengan melakukan standardisasi, semua fitur akan berada dalam skala yang sama, sehingga model machine learning yang sensitif terhadap skala data (seperti SVM, KNN, Logistic Regression, dll.) dapat bekerja lebih optimal.

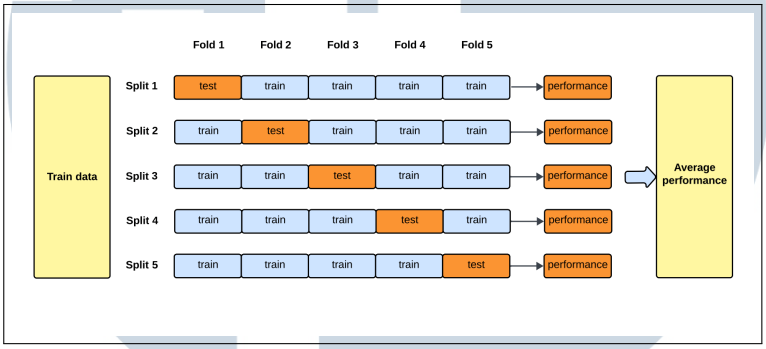
2.6 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE merupakan pendekatan over-sampling di mana kelas minoritas diperbanyak dengan membuat contoh “sintetik” alih-alih menyalin data secara langsung (over-sampling dengan pengulangan). Di mana data pelatihan tambahan dibuat dengan memodifikasi data nyata, misalnya melalui rotasi atau distorsi. Namun dalam SMOTE, contoh sintetik dihasilkan dengan cara yang lebih umum, yaitu dengan bekerja di ruang fitur (feature space), bukan di ruang data mentah (data space).Kelas minoritas di-over-sample dengan cara mengambil setiap sampel dari kelas minoritas, lalu membuat contoh sintetik di sepanjang garis yang menghubungkan sampel tersebut dengan salah satu tetangganya yang paling dekat dalam kelas yang sama. Banyaknya tetangga yang digunakan tergantung pada tingkat over-sampling yang diinginkan.

Dengan membuat data sintetik di sepanjang garis antar titik dalam

kelas minoritas, SMOTE secara tidak langsung memperbesar wilayah keputusan (decision boundary) kelas minoritas. Hal ini memungkinkan model pembelajaran mesin yang dilatih di atasnya menjadi lebih sensitif dan seimbang dalam memprediksi kedua kelas, khususnya saat menghadapi data yang sangat tidak seimbang [30].

2.7 Cross Validation



Gambar 2.2. Visualisasi cross validation

Visualisai Cross Validation dapat dilihat pada Gambar 2.2. Cross-validation merupakan metode resampling yang umum digunakan untuk mengevaluasi kemampuan generalisasi model prediktif dan mencegah overfitting. Teknik ini bekerja dengan membagi data menjadi beberapa subset, di mana model dilatih pada sebagian data dan diuji pada sisanya. Dalam pembelajaran mesin, cross-validation termasuk metode Monte Carlo yang sering digunakan dalam supervised learning untuk memastikan model tidak hanya akurat pada data latih, tetapi juga mampu memprediksi data baru. Pendekatan ini memungkinkan evaluasi performa model secara efisien tanpa memerlukan data uji eksternal, sehingga membantu peneliti menilai kualitas model sebelum proses validasi lanjutan dilakukan. Pada penelitian ini teknik cross validation yang digunakan adalah K-fold cross-validation.

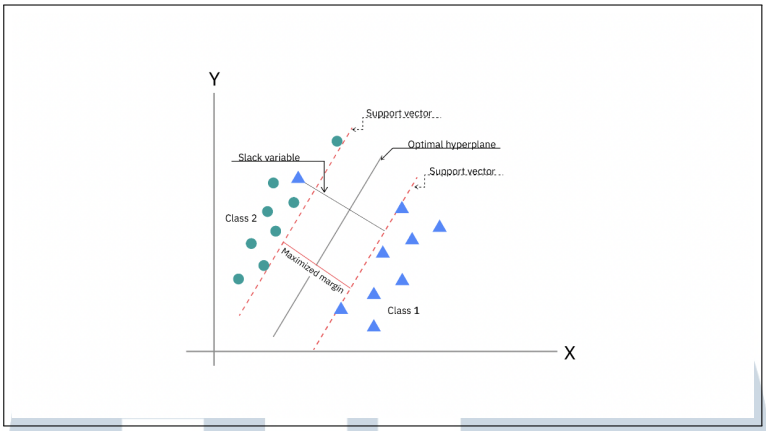
K-fold cross-validation adalah metode evaluasi yang umum digunakan dalam pembelajaran mesin untuk mengukur kemampuan generalisasi model terhadap data baru. Teknik ini membagi data menjadi beberapa subset yang tidak tumpang tindih, di mana setiap subset secara bergiliran digunakan sebagai data validasi, sementara sisanya digunakan untuk pelatihan. Hasil evaluasi dari seluruh putaran dirata-ratakan untuk memperoleh gambaran performa model secara keseluruhan.

Metode ini efisien karena memanfaatkan seluruh data yang tersedia dan menghasilkan estimasi performa yang lebih stabil. Dalam praktiknya, k-fold cross-validation sering dipadukan dengan stratifikasi untuk menjaga proporsi kelas tetap seimbang di setiap subset, terutama pada dataset yang tidak seimbang. Pengulangan k-fold dengan kombinasi subset berbeda dapat dilakukan untuk mengurangi variansi, namun peningkatan kestabilannya relatif kecil. Oleh karena itu, pendekatan ini tetap menjadi pilihan yang disarankan dalam evaluasi model prediktif [31].

2.8 Support Vector Machine (SVM)

Support Vector Machine (SVM) adalah teknik klasifikasi dalam machine learning yang digunakan untuk menganalisis data dan menemukan pola dalam tugas klasifikasi maupun regresi. SVM biasanya digunakan ketika data dibagi menjadi dua kelas (two-class problem) untuk membantu memisahkan dan mengidentifikasi pola antar kelas tersebut [32].Kekuatan SVM terletak pada kemampuannya dalam mengenali pola klasifikasi data dengan akurasi yang seimbang dan hasil yang konsisten. Meskipun terkadang digunakan untuk regresi, SVM lebih dikenal sebagai alat yang sangat populer untuk tugas klasifikasi. Fleksibilitasnya yang tinggi membuatnya banyak digunakan dalam berbagai bidang ilmu data [21], termasuk dalam penelitian memprediksi penyakit jantung.

Dalam konsep SVM terdapat 3 aspek penting, diantaranya adalah Hyperplane yaitu garis (untuk 2D), bidang (untuk 3D), atau dimensi lebih tinggi yang membagi ruang fitur menjadi dua bagian, yang kedua ada Support Vector yaitu titik-titik data terdekat dengan hyperplane. Titik-titik ini sangat penting karena menentukan posisi hyperplane, yang terakhir adalah Margin yaitu jarak antara hyperplane dan support vector terdekat. SVM berusaha memaksimalkan margin agar pemisahan antar kelas seoptimal mungkin. Fungsi keputusan pada SVM secara sederhana adalah sebuah "hyperplane" atau bidang pemisah optimal yang digunakan untuk membedakan (atau mengklasifikasikan) data ke dalam kelas yang berbeda berdasarkan pola informasi atau fitur dari data tersebut. Hyperplane ini kemudian digunakan untuk memprediksi label yang paling mungkin dari data baru yang belum pernah dilihat sebelumnya [21] [33]. Visualisasi sederhana hyperplane dapat dilihat pada Gambar 2.3.



Gambar 2.3. Visualisasi Sederhana Hyperplane

Algoritma Support Vector Machine (SVM) banyak digunakan dalam machine learning karena kemampuannya dalam menangani masalah klasifikasi, baik yang bersifat linier maupun nonlinier. Namun, ketika data tidak bisa dipisahkan dengan garis lurus (secara linier), SVM menggunakan sesuatu yang disebut fungsi kernel. Fungsi ini mengubah data ke dalam ruang dimensi yang lebih tinggi sehingga data bisa dipisahkan secara linier di ruang tersebut. Proses ini dikenal sebagai ”kernel trick” [34].

Fungsi kernel digunakan dalam SVM sebagai cara matematis untuk mengubah data berdimensi rendah menjadi bentuk berdimensi lebih tinggi, sehingga data yang awalnya sulit dipisahkan (non-linear) menjadi lebih mudah untuk dipisahkan secara linear. Dengan kata lain, kernel membantu proses klasifikasi yang lebih kompleks tanpa perlu mengolah data secara eksplisit di dimensi yang lebih tinggi [34].

Fungsi basis radial adalah salah satu fungsi kernel yang paling populer dan digunakan pada penelitian ini. Fungsi ini menambahkan semacam “tonjolan” (bump) di sekitar setiap titik data. ”Bump” di sini bukan objek fisik, tapi representasi dari pengaruh lokal suatu titik data terhadap sekelilingnya di ruang fitur seperti tonjolan kecil yang memperkuat sinyal di sekitar titik tersebut.

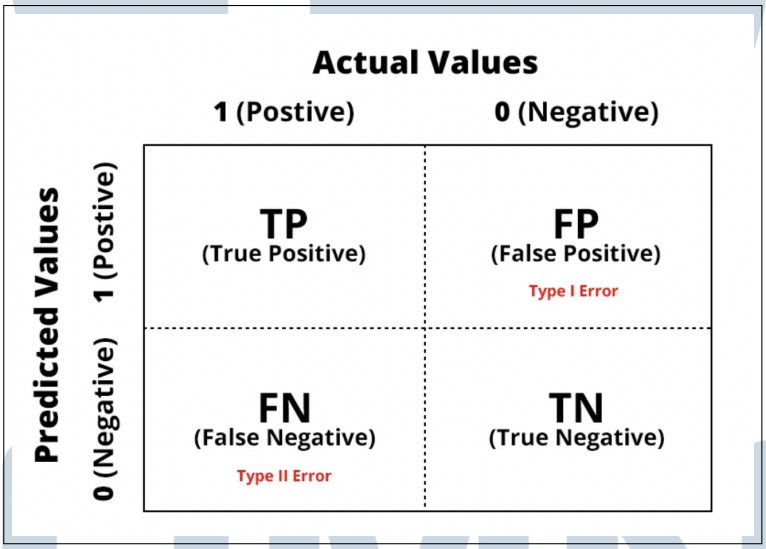
Fungsi kernel Radial Basis Function (RBF) digunakan dalam penelitian ini karena mampu menangani data yang tidak dapat dipisahkan secara linear dengan memetakan data ke ruang berdimensi lebih tinggi. RBF menambahkan “tonjolan” di sekitar tiap titik data, yang merepresentasikan pengaruh lokal titik tersebut terhadap sekelilingnya, sehingga model dapat membentuk batas keputusan yang fleksibel. Dalam penelitian ini, kernel RBF digunakan untuk menguji dua hyperparameter penting, yaitu C dan gamma. Parameter C mengatur toleransi terhadap kesalahan

klasifikasi, sementara gamma menentukan seberapa jauh pengaruh satu titik data dalam membentuk model. Kombinasi keduanya diuji untuk menemukan konfigurasi terbaik, dan hasilnya kernel RBF memberikan performa klasifikasi paling optimal dibanding kernel lainnya dalam mendeteksi penyakit jantung.

Pemilihan jenis kernel yang tepat sangat bergantung pada karakteristik data yang digunakan serta jenis masalah yang ingin diselesaikan.

2.9 Confusion Matrix

Kinerja sebuah model dapat di evaluasi menggunakan nilai yang didapatkan oleh confusion matrix yang dapat dilihat pada Gambar 2.4.



Gambar 2.4. Confusion Matrix

Terdapat 4 istilah sebagai representasi hasil proses klasifikasi pada confusion matrix. Keempat istilah tersebut adalah True Positive (TP) adalah data positif diprediksi benar contohnya Pasien sakit kanker dan diprediksi sakit, True Negative (TN) adalah Data negatif diprediksi benar contohnya Pasien sehat dan diprediksi sehat, False Positive (FP) adalah data negatif diprediksi positif contohnya Pasien sehat diprediksi sakit dan False Negative (FN) adalah Data positif diprediksi negatif contohnya pasien sakit diprediksi sehat.

Confusion matrix digunakan untuk menghitung berbagai *performance metrics* untuk mengukur kinerja model yang telah dibuat. Pada bagian ini mari dipahami beberapa performance metrics populer yang umum dan sering digunakan: accuracy, precission, dan recall.

Rumus 2.1 menunjukkan cara perhitungan *Accuracy*.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1}$$

Rumus 2.2 menunjukkan cara perhitungan *Precision*.

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

Rumus 2.3 menunjukkan cara perhitungan *Recall*.

$$Recall = \frac{TP}{TP + FN} \tag{2.3}$$

Rumus 2.4 menunjukkan cara perhitungan *F1-Score*.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.4}$$

Dari hasil confusion matrix tersebut dapat dihasilkan beberapa evaluation mectric yaitu Accuracy, Precision, Recall dan F1-Score. Accuracy adalah proporsi dari prediksi yang benar. Ini mengukur seberapa baik sebuah model klasifikasi memprediksi suatu kondisi. Precision adalah adalah ukuran yang menunjukkan seberapa akurat model dalam memprediksi data positif. Precision dihitung dengan membandingkan jumlah prediksi positif yang benar terhadap seluruh prediksi positif yang dibuat oleh model. Recall menggambarkan keberhasilan model dalam menemukan kembali sebuah informasi. Maka, recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. F1-Score adalah metrik yang mempertimbangkan Presisi dan Sensitivitas secara bersamaan. Nilai ini merupakan rata-rata harmonik antara presisi dan recall [35, 36].