BAB II
TINJAUAN PUSTAKA

2.1 Studi Literatur

Studi literatur ini dilakukan dengan tujuan untuk menemukan referensi dari berbagai penelitian sebelumnya, dimana akan dijadikan acuan dalam menentukan tahap-tahap sistematis dalam penyusunan penelitian ini. Beberapa penelitian sebelumnya adalah sebagai berikut :

***Tabel 2. 1** Studi Literature*

No	Penulis	Judul	Algoritma dan akurasi	Dataset
1	Elias Dritsas, dkk (2022)	*Cardiovascular Disease Risk Prediction with Supervised Machine Learnin Techniques* [5]	• Naïve Bayes (NB) : 59,59 % • Support Vector Machine (SVM) : 70.61 % • Random Forest (RF) : 70.86 % • Logistic Regression (LR) : 72.06 %	70.000 data *"Cardiovascular Disease Dataset"*
2	I Ketut Adian Jayaditya, (2023)	Implementasi Random Forest pada Klasifikasi Penyakit Kardiovaskular dengan Hyperparameter Tuning Grid Search [15]	Random Forest (FR) • RF Tanpa Grid Search CV : 69.65 % • RF dengan Grid Search CV : 73.06 %	70.000 data *"Cardiovascular Disease Dataset"*
3	Muhamad Ichsan	Peningkatan Kinerja Akurasi Prediksi	Logistic Regression (LR)	768 data *"Pima Indians*

	Gunawan, dkk (2020)	Penyakit Diabetes Mellitus Menggunakan Metode Grid Seacrh pada Algoritma Logistic Regression [16]	• LR Tanpa Grid Search CV : 72.22 % • LR dengan Grid Search CV : 83.33 %	*Diabetes Database"*

Tabel 2.1 di atas menjelaskan tentang beberapa penelitian yang pernah dilakukan sebelumnya tentang penyakit kardiovaskular. Penelitian pertama dilakukan oleh Elias Dritsas, dkk pada tahun 2022 [5]. Penelitian ini menggunakan data dari *Kaggle* dengan jumlah 70.000 data (*Cardiovascular Disease Dataset)*. Beberapa algoritma yang digunakan dan hasil akurasinya antara lain Naïve Bayes (59.59 %), Support Vector Machine (70.61 %), Random Forest (70.86 %), Logistic Regression (72.06 %). Dalam penelitian ini belum terdapat adanya sebuah model prediktif yang disempurnakan guna adanya peningkatan akurasi. Terdapat peluang untuk meningkatkan hasil dari penelitian ini agar dapat menambah pemahaman terkait klasifikasi/prediksi lebih awal terkait dengan penyakit kardiovaskular sehingga adanya hasil pencegahan yang lebih baik.

Penelitian sebelumnya juga tentang penyakit kardiovaskular juga dilakukan oleh I Ketut Adian Jayaditya, dkk pada tahun 2023 [15]. Data yang digunakan sejumlah 70.000 (*Cardiovascular Disease Dataset)* yang didapatkan dari situs *Kaggle.* Algoritma yang digunakan adalah Random Forest dengan akurasi 69.65% tanpa *grid search* dan 73.06% menggunakan *grid search.* Hal ini menunjukkan bahwa *grid search* mampu mengoptimalkan akurasi sehingga dapat mengklasifikasikan penyakit kardiovaskular secara lebih tepat.

Terdapat penelitian lain yang menunjukkan tentang penggunaan metode *logistic regression* dan *grid search*. Penelitian itu dilakukan oleh Muhamad Ichsan Gunawan, dkk pada tahun 2020 [16]. Dataset yang digunakan dalam penelitian tersebut berjumlah 768 data yang diambil dari *"Pima Indians Diabetes Database".*

Algoritma yang digunakan dalam penelitian ini adalah *logistic regression* dengan hasil akurasi sebesar 72.22% tanpa *grid search* dan 83.33% menggunakan *grid search.* Penelitian tersebut berkontribusi dengan meningkatkan akurasi model menggunakan metodologi sistematis yaitu *grid search.* Kontribusi ini penting dilakukan untuk memajukan pemahaman dan penerapan *machine learning* dalam perawatan kesehatan.

Berdasarkan penelitian yang telah dilakukan sebelumnya, maka penelitian ini bertujuan untuk mengembangkan algoritma *Logistic Regression* dengan optimasi *grid search CV* yang dapat mengklasifikasi penyakit kardiovaskular secara akurat. Algoritma dan optimasi yang dikembangkan diharapkan dapat membantu dalam klasifikasi, prediksi serta pencegahan penyakit kardiovaskular sehingga dapat memberikan hasil yang baik dalam perawatan kesehatan.

2.2 Penyakit Kardiovaskular

Berbagai kondisi seperti penyakit jantung koroner, penyakit serebrovaskular, penyakit arteri perifer, penyakit jantung rematik, dan penyakit jantung bawaan termasuk dalam kategori penyakit kardiovaskular [1]. Dengan sekitar 19,7 juta kematian pada tahun 2019, penyakit ini masih menjadi penyebab kematian nomor satu di seluruh dunia [17]. Ada beberapa faktor yang bisa menyebabkan terjadi nya penyakit ini. Faktor seperti riwayat penyakit keluarga, usia dan jenis kelamin merupakan faktor risiko yang tidak bisa diubah. Sedangkan tingkat kolestrol, diabetes, kurangnya aktivitas fisik, pengonsumsian rokok dan alkohol adalah faktor-faktor risiko yang bisa diubah.

2.3 Machine Learning

Machine learning adalah cabang dari kecerdasan buatan (*artificial intelligent)* dan ilmu komputer (*computer science)* yang berfokus pada pengembangan algoritma dan model statistik yang memungkinkan komputer untuk melakukan tugas - tugas tertentu dan membuat keputusan berdasarkan data yang tersedia. *Machine learning* menggunakan berbagai metode komputasi yang dirancang berdasarkan data dan pengalaman sebelumnya [19]. Algoritma dalam

machine learning bertujuan untuk mempelajari model atau kumpulan data sehingga dapat memprediksi label data dengan tepat. Algoritma *machine learning* secara umum dibagi menjadi 4 tipe [19], akan di jelaskan sebagai berikut :

2.3.1 *Supervised Learning*

Supervised learning merupakan tugas dari *machine learning* untuk memetakan input-output menggunakan data latih yang sudah berlabel. Tugas umum dari *supervised learning* adalah "klasifikasi" dan "regresi" [19].

2.3.2 Unsupervised Learning

Unsupervised Learning menganalisis kumpulan data yang tidak berlabel. Banyak digunakan untuk "clustering" dan "anomaly detection" [19].

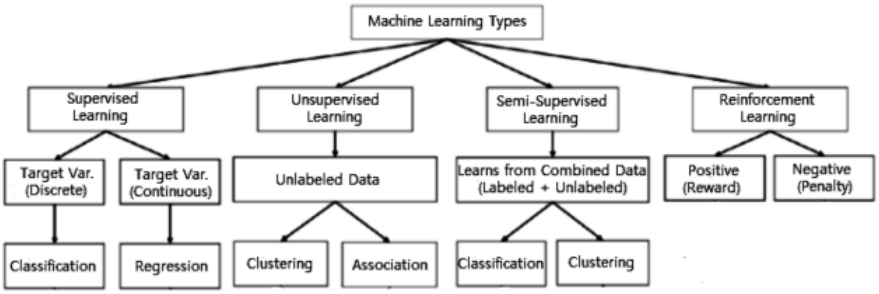
2.3.3 Semi-supervised

Semi-supervised merupakan gabungan dari *supervised learning* dan *unsupervised learning* dimana fungsinya untuk menganalisis kumpulan data gabungan yakni data berlabel dan tidak berlabel. Banyak digunakan untuk klasifikasi teks [19].

2.3.4 Reinforcement

Reinforcement adalah sejenis algoritma yang lebih ampuh untuk melatih model AI *(Artificial Intelligent)* yang dapat membentu mengoptimalkan efisiensi operasional yang canggih seperti sistem robotika [19].

Penjelasan lebih lanjut tentang tipe – tipe dari *machine learning* akan ditampilkan dalam gambar 2.1 sebagai berikut :



Gambar 2. 1 *4 Tipe Machine Learning [19]*

2.4 Klasifikasi

Klasifikasi merupakan proses menemukan serangkaian model atau fungsi yang mampu mengidentifikasi dan membedakan berbagai konsep atau kelas data. Tujuannya adalah agar model tersebut dapat digunakan untuk memprediksi kelas dari suatu data yang belum memiliki label kelas. Klasifikasi termasuk dalam kategori *supervised machine learning.* Proses klasifikasi terdiri dari dua tahap, yaitu tahap pelatihan *(training)* dan tahap prediksi (klasifikasi) [7]. Klasifikasi dibagi menjadi 3 tipe [19] yang akan dijelaskan sebagai berikut :

2.4.1 Klasifikasi Biner

Klasifikasi biner mengacu pada tugas klasifikasi yang ha nya memiliki 2 label kelas, seperti "benar atau salah", "ya atau tidak", "menderita atau tidak menderita" [19].

2.4.2 Klasifikasi Multiclasss

Klasifikasi Multiclass mengacu pada tugas klasifikasi yang memiliki lebih dari 2 label kelas. Contoh dari klasifikasi ini antara lain ketika akan mengklasifikasikan jenis serangan jaringan yang masuk ke dalam 4 kelas, seperti DoS *(Denial of Service Attack)*, U2R *(User to Root Attack)*, R2L *(Root To Local Attack)*, dan *Probbing Attack* [19].

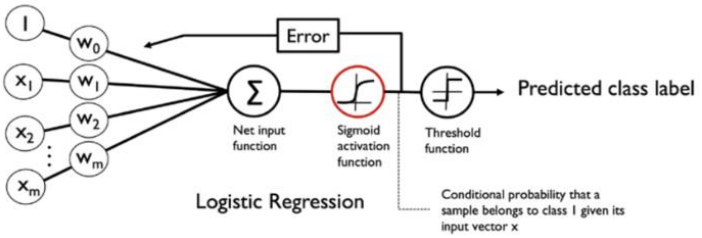
2.4.3 Klasifikasi Multi-Label

Pada tipe klasifikasi multi-label mengacu pada data yang memiliki lebih dari 2 kelas, dimana setiap kelas bisa memiliki lebih dari 1 label yang terkait. Contoh dari klasifikasi ini adalah ketika akan mengklasifikasikan berita dari *Google* yang dapat disajikan dalam beberapa kategori seperti ("nama kota", "teknologi", "berita terkini") [19].

2.5 Logistic Regression

Logistic regression adalah sebuah algoritma yang digunakan untuk menganalisis data yang bersifat kategoris, biasanya datanya bersifat biner seperti 0 atau 1, ya atau tidak, benar atau salah [20]. Dalam *machine learning, logistic regression* merupakan model statistik berbasis probabilitas umum yang digunakan

untuk memecahkan masalah klasifikasi dengan menentukan hubungan antara varibel bebas dan variabel terikat [19]. *Logistic Regression* banyak digunakan di berbagai bidang termasuk kedokteran, ilmu sosial, dan *machine learning* untuk tugas seperti memprediksi hasil penyakit, perilaku pelanggan, dan masalah klasifikasi biner. Arsitektur dari *logistic regression* akan ditampilkan dalam gambar 2.2 sebagai berikut :



***Gambar 2. 2** Arsitektur Model Logistic Regression [21]*

2.6 Hyperparameter Tuning

Hyperparameter tuning berperan penting dalam mengoptimalkan kinerja dan meningkatkan performa algoritma *machine learning*. Tujuan dari penggunaan *hyperparameter tuning* adalah untuk mencegah overfitting, mengurangi waktu pelatihan dan mencapai skor maksimum pada hasil evaluasi model [22]. Terdapat beberapa jenis *hyperparameter tuning*, diantaranya adalah *GridSearch CV, RandomSearch CV, bayesian optimization,* dan *evolutionary optimization* [23].

2.7 Grid Search CV

Grid search adalah metode yang digunakan untuk menentukan kombinasi terbaik antara model dan *hyperparameter* dengan cara yang sistematis. Prosesnya melibatkan pengujian dari setiap kombinasi *hyperparameter* yang telah ditentukan sebelumnya, di ikuti dengan validasi pada setiap kombinasi tersebut [12]. *Grid search* sering dikombinasikan dengan metode *cross-validation* (CV) untuk memperkirakan keakuratan kinerja model pada data yang terbatas [24]. Metode ini secara otomatis membantu dalam menentukan *hyperparameter* terbaik dalam sebuah prediksi. Cara kerja *grid search* adalah sebagai berikut [25] :

1. Menginisialisasi semua parameter yang ada
2. Mengkombinasikan semua nilai parameter

3. Melakukan pelatihan dengan metode *machine learning*
4. Mengevaluasi klasifikasi yang dihasilkan dengan data uji
5. Menyimpan hasil klasifikasi terbaik dan kombinasi nilai parameter terbaik

2.8 Evaluasi Model

Evaluasi model merupakan proses penting yang digunakan untuk penilaian kinerja dan akurasi model dalam melakukan klasifikasi atau prediksi pada suatu data. Salah satu jenis evaluasi model adalah *confusion matrix. Confusion matrix* merupakan sebuah tabel yang menunjukan jumlah prediksi yang benar dan salah untuk setiap kelas. Tabel ini memberikan gambaran yang lebih jelas mengenai kinerja model dengan memisahkan antara prediksi yang benar dan salah untuk masing-masing kelas [14]. Sebagai contoh *confusion matrix* akan ditampilkan pada tabel berikut:

***Tabel 2. 2** Confusion Matrix*

	Keadaan Data Sebenarnya		
		TRUE	FALSE
Hasil Prediksi	TRUE	TP (*True Positive)*	FP *(False Positive)*
	FALSE	FN *(False Negative)*	TN (True Negative)

Komponen utama *confusion matrix* akan dijelaskan sebagai berikut :

1. True Positive (TP) : Jumlah sampel bernilai true yang hasil prediksinya positif.
2. True Negative (TN) : Jumlah sampel bernilai true yang hasil prediksinya negatif.
3. False Positive (FP) : Jumlah sampel bernilai false dan hasil prediksinya positif.
4. False Negative (FN) : Jumlah sampel bernilai false yang hasil prediksinya negatif.

Dalam *confusion matrix* terdapat beberapa nilai yang dapat dihasilkan dan ditampilkan dalam bentuk *classification report* yang menampilkan nilai akurasi, presisi, recall, dan F1-Score. Rumus untuk mendapatkan nilai tersebut bisa dilihat pada persamaan (1) sampai (4) sebagai berikut [14] :

2.8.1 Akurasi

Akurasi menggambarkan tingkat ketepatan sistem dalam mengklasifikasikan data secara benar. Nilai akurasi diperoleh dari perbandingan antara jumlah data yang terklasfikasi benar dengan total keseluruhan data.

$$Akurasi = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

2.8.2 Presisi

Presisi adalah nilai dari berapa banyak data dengan kategori terklasifikasi benar dibandingkan dengan total keseluruhan data kategori terklasifikasi benar.

$$Presisi = \frac{TP}{TP+FP} \quad (2)$$

2.8.3 Recall

Recall dilakukan untuk mengetahui perbandingan jumlah data kategori terklasifikasi benar oleh sistem dengan jumlah data kategori terklasifikasi benar dan salah.

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

2.8.4 F1-Score

F1-Score merupakan penggabungan dari presisi dan recall.

$$F1-SCORE = \frac{2\ x\ Presisi\ x\ Recall}{Presisi+Recall} \quad (4)$$