BAB II
TINJAUAN PUSTAKA

2.1. Penyakit Jantung

Penyakit jantung merupakan gangguan yang terjadi pada sistem pembuluh darah besar sehingga menyebabkan jantung dan peredaran darah tidak berfungsi sebagaimana mestinya. Penyakit yang berhubungan dengan organ jantung dan pembuluh darah antara lain: gagal jantung, jantung koroner, dan jantung rematik [6]. Penyakit jantung koroner (PJK) adalah penyakit jantung dan pembuluh darah yang disebabkan karena penyempitan arteri koroner. Penyempitan pembuluh darah terjadi karena proses aterosklerosis atau spasme atau kombinasi keduanya. Aterosklerosis yang terjadi karena timbunan kolesterol dan jaringan ikat pada dinding pembuluh darah secara perlahan-lahan, hal ini sering ditandai dengan keluhan nyeri pada dada. Pada waktu jantung harus bekerja lebih keras terjadi ketidakseimbangan antara kebutuhan dan asupan oksigen, hal inilah yang menyebabkan nyeri dada. Kalau pembuluh darah tersumbat sama sekali, pemasokan darah ke jantung akan terhenti dan kejadian inilah yang disebut dengan serangan jantung. Penyakit jantung sering disebut “suddent death”. Seseorang kemungkinan mengalami serangan jantung karena *iskemia miokard* atau kekurangan oksigen pada otot jantung atau sering disebut dengan nyeri dada.

Beberapa faktor yang bisa menimbulkan penyakit jantung antara lain:

1. bertambahnya usia,
2. gaya hidup
3. stres
4. kurangnya waktu istirahat
5. kurangnya berolah raga
6. merokok
7. obesitas
8. dislipidemia
9. permasalahan dalam diagnosa klinis penyakit jantung

2.2. Data Mining

Data mining merupakan proses menelusuri pengetahuan terbaru, pola dan tren yang dipilih dari jumlah data yang besar dan disimpan dalam repositori atau tempat penyimpanan dengan menggunakan teknik pengenalan pola serta statistik dan teknik matematika [10]. Data mining kemudian dikenal dengan nama *Knowladge-discovey in Database* (KDD) adalah kegiatan yang meliputi pengumpulan, pemakaian data historis untuk memecahkan pola atau hubungan keteraturan dalam set data yang berukuran besar. Keluaran dari data mining ini bisa dipakai untuk memperbaiki pengambilan keputusan dimasa depan berdasarkan informasi yang diperoleh dari data masa lalu [11]. Proses KDD secara garis besar terdiri dari dapat dijelaskan sebagai berikut:

1. *Data Selection*, tahap ini melakukan pemilihan data dari sekumpulan data operasional perlu dilakukan sebelum tahap penggalian informasi, data hasil seleksi akan digunakan untuk proses data mining, disimpan dalam suatu berkas terpisah dari basis data operasional.
2. *Pre-processing/Cleaning*, operasi dasar seperti penghapusan noise. Proses ini membuang duplikasi data, memeriksa data yang inkonsisten dan memperbaiki kesalahan data contohnya seperti kesalahan cetak. Data bisa diperkaya dengan data atau informasi ekternal yang relavan.
3. *Transformation*, pada tahap ini proses integrasi pada data yang telah dipilih, sehingga data sesuai untuk proses data mining.
4. *Data mining*, pada tahap ini menentukan tipe data mining yang akan digunakan, seperti klasifikasi, clustering atau regresi dan lain-lain tergantung pada proses dan tujuan KDD secara keseluruhan.
5. *Interpretation/Evaluation*, penerjemah pola-pola yang dihasilkan dari data mining. Pola informasi yang akan dihasilkan perlu ditampilkan dalam bentuk yang mudah dimengerti.

2.3. Algoritma Klasifikasi *Data Mining*

Algoritma Klasifikasi Data Mining merupakan suatu pengelompakan data untuk memprediksi nilai dari sekelompok atribut dalam menggambarkan dan membedakan kelas label atau target yang bertujuan untuk memprediksi kelas dari

objek yang label kelasnya tidak diketahui. Performa algoritma data mining pada banyak kasus tergantung kepada kualitas dataset yang digunakan, karena data training yang berkualitas rendah dapat menyebabkan klasifikasi yang lemah juga [10]. Dalam klasifikasi data mining proses klasifikasi didasarkan pada empat komponen mendasar yaitu [12]:

1. *Kelas*

   Variabel dependen yang merupakan variabel kategoris yang mewakili 'label' yang terdapat pada objek. Contoh kelaskelas tersebut adalah resiko penyakit hepatitis, resiko kredit, loyalitas pelanggan, jenis gempa, jenis bintang.

2. *Predictor*

   Variabel independen dari suatu model yang diwakili oleh karakteristik atau atribut dari data yang akan diklasifikasikan dan berdasarkan klasifikasi yang akan dibuat. Contoh predictor misalnya status perkawinan, tekanan darah, merokok, konsumsi alkohol, frekuensi pembelian, musim, arah angin dan kecepatan.

3. *Training Dataset*

   Kumpulan data yang berisi nilai dari kedua komponen diatas yang digunakan untuk melatih model untuk mengenali kelas yang sesuai berdasarkan predictor yang tersedia. Contoh dari dataset tersebut misalnya kelompok pasien pada serangan jantung.

4. *Testing Dataset*

   Data yang berisi data baru yang akan diklasifikasi oleh model yang akan dibuat dan akurasi klasifikasi dapat dievaluasi.

2.4. Naïve Bayes

Naive Bayes dikemukakan oleh ilmuan Inggris *Thomas Bayes*. *Naive Bayes* memprediksi peluang dimasa depan dengan berdasarkan pengalaman dimasa sebelumnya. *Naive Bayes* juga dinilai berpotensi baik dalam mengklasifikasikan dokumen dibandingkan dengan metode klasifikasi lainnya dalam hal akurasi (Wahyuningsih dan Patima, 2018). *Naive bayes* merupakan metode yang tidak memiliki aturan, *Naive Bayes* menggunakan cabang matematika yang dikenal dengan teori probabilistik untuk mencari peluang terbesar dari beberapa

kemungkinan klasifikasi, dengan melihat frekuensi tiap klasifikasi pada data training. Klasifikasi *Naive Bayes* merupakan pengklasifikasian statistik yang digunakan untuk memprediksi probabilitas keanggotaan suatu kelas. Klasifikasi bayesian didasarkan pada teorema bayes yang memiliki kemampuan klasifikasi serupa dengan *decision tree dan neural network* [13].

Bayes rule digunakan untuk menghitung *probabilistik* suatu *class*. *Algoritma Naive Bayes* memberikan suatu cara dengan mengkombinasikan peluang terdahulu dengan syarat kemungkinan menjadi sebuah formula baru yang digunakan untuk menghitung peluang dari tiap kemungkinan yang akan terjadi. Bentuk umum dari *teorema bayes* seperti dibawah ini:

$$P(H|X) = \frac{P(X|H)P(H)}{P(X)} \qquad (2.1)$$

Dimana:

X : Data dengan kelas yang belum diketahui

H : Hipotesis data X merupakan suatu kelas spesifik

P(H|X) : *Probabilitas* hipotesis H berdasar kondisi X *(posterior probability)*

P(H) : *Probabilitas* hipotesis H *(prior probability)*

P(X|H) : *Probabilitas* X berdasar kondisi pada hipotesis H

P(X) : *Probabilitas* dari X

Naive bayes adalah penyederhanaan *metode bayes*. *Teorema bayes* setelah disederhanakan menjadi:

$$P(H|X) = P(X|H)P(X) \qquad (2.2)$$

Pada *bayes rule* diterapkan untuk menghitung *posterior* dan *probabilitas* dari data sebelumnya. Dalam analisis bayesian, klasifikasi akhir dihasilkan dengan menggabungkan kedua sumber informasi yaitu informasi *prior* dan informasi *posterior* untuk menghasilkan probabilitas menggunakan aturan *bayes*.

2.5. Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) sering digunakan dalam sebuah penelitian, karena Particle Swarm Optimization memiliki kesamaan sifat dengan *Genetic Algorithm (GA)*. Keuntungan dari *PSO* adalah mudah diterapkan dan ada beberapa parameter untuk menyesuaikan-nya [6]. *Particle Swarm Optimization* merupakan

teknik optimasi stokastik berbasis populasi yang diusulkan pada yahun 1995 oleh *Kennedy* dan *Eberhart.* Pengembangan *PSO* didasarkan pada metafora interaksi sosial dan komunikasi dari pergerakan kawanan burung atau ikan *(bird flocking atau fish schooling)* [14].

Hu et al (2003) Particle Swarm Optimization menggunakan teknik perhitungan evolusioner [15]:

1. *Particle Swarm Optimization* diinisialisasi dengan sekumpulan solusi acak.
2. *Particle Swarm Optimization* mencari solusi yang optimum dengan memperbarui generasi.
3. Perkembangan polulasi berdasarkan pada generasi sebelumnya.

Pada algoritma *Particle Swarm Optimization*, pencarian solusi dilakukan oleh suatu populasi yang terdiri dari beberapa partikel. Ketika Populasi dibangkitkan secara random atau acak dengan batasan permasalahan yang dihadapi. Setiap partikel merepresentasikan partikel atau solusi dari permasalahan yang sedang dihadapi. Partikel tersebut mencari solusi yang optimal dengan melintasi ruang pencarian dengan cara partikel terkait melakukan penyesuaian terhadap posisi yang terbaik dari setiap partikel tersebut atau disebut *local best* dan posisi partikel terbaik dari seluruh kawanan atau disebut *global best* selama melintasi ruang pencarian. Jadi penyebaran informasi terjadi dalam partikel itu sendiri dan antara suatu partikel dengan partikel terbaik dari seluruh kawanan selama proses pencarian solusi. Setelah itu dilakukan proses pencarian untuk mencari posisi terbaik setiap partikel dalam jumlah iterasi tertentu sampai didapatkannya posisi relatif yang tetap *(steady)* atau mencapai batas iterasi yang telah ditetapkan. Pada setiap *iterasi (t)*, setiap solusi yang direpresentasikan oleh partikel *i*, dievaluasi performanya dengan cara memasukkan solusi tersebut kedalam nilai *fitness function*.

Setiap partikel memerlukan titik pada suatu dimensi ruang tertentu kemudian terdapat 2 faktor yang memberikan karakter terhadap status partikel pada ruang pencarian yaitu pada posisi partikel (X) dan pada kecepatan partikel (Y). Formulasi matematika yang menggambarkan posisi dan kecapatan partikel suatu ruang dimensi tertentu sebagai berikut:

$$X_i(t) = X_{i1}(t), X_{i2}(t), ..., X_{in}(t) \quad (2.3)$$

$$V_i(t) = V_{i1}(t), V_{i2}(t), ..., V_{in}(t) \quad (2.4)$$

Persamaan diatas (2.4) digunakan untuk menggambarkan kecepatan partikel yang baru berdasarkan kecepatan pada kecepatan sebelumnya, jarak antara posisi saat ini dengan posisi partikel terbaik atau *local best*, dan jarak antara posisi saat ini dengan posisi terbaik dalam kawanan atau *global best*. Kemudian partikel terbang menuju posisi yang baru berdasarkan persamaan (2.5).

$$V_i(t) = v_{i1}(t-1) + c_1 r_1 (X_i^L - X_i(t-1) + c_2 r_2 (X_i^G - X_i(t-1)) \quad (2.5)$$

$$X_{i\,(t)} = v_i(t) + X_i(t-1) \quad (2.6)$$

Dimana :

$V_i(t)$: kecepatan partikel ke-*i* pada *iterasi* ke-*i*

$X_{i\,(t)}$: posisi partikel saat ini pada partikel ke-*i* pada *iterasi* ke-*i*

t : *iterasi*

X_i^L : *local best* dari partikel ke-*i*

X_i^G : *global best* dari seluruh kawanan

$c_1 c_2$: konstanta akselerasi atau *learning rate*

$r_1 r_2$: bilangan *random* atau acak yang bernilai antara 0 sampai dengan 1

2.6. Validasi *(K-Fold Cross Validation)*

K-Fold Cross validation merupakan suatu metode evaluasi dimana pada metode ini data yang digunakan dalam jumlah yang sama untuk *training* dan tepat satu kali untuk *testing* [16]. Dengan mambagi data secara acak kedalam *k* bagian dan masing-masing bagian akan dilakukan proses klasifikasi, secara umum pengujian nilai *k* dilakukan sebanyak 10 kali untuk memperkirakan akurasi estimasi[17]. *Cross validation* yang paling sering digunakan adalah *10-fold cross validation*. Prinsip dari *10-fold cross validation* adalah 9;1 dimana 9 bagian menjadi data *training* dan 1 bagian menjadi data *testing*, sehingga 10 bagian tersebut dapat berkesempatan menjadi data *testing* [18].

Cara kerja *k-folds cross validation*, yaitu total data dibagi menjadi n bagian, *iterasi* atau *fold* ke 1, yaitu bagian ke 1 menjadi *testing*, bagian sisanya menjadi data *training*, kemudian hitung akurasi menggunakan persamaan berikut :

$$Accuracy = \frac{jumah\ klasfikasi\ benar}{jumlah\ data\ uji} x\ 100\%$$

Keterangan :

Jumlah klasifikasi benar : jumlah prediksi klasifikasi yang tepat

Jumlah data uji : jumlah dataset yang digunakan untuk *testing*

2.7. Klasifikasi

Klasifikasi merupakan metode untuk menentukan sebuah anggota kedalam suatu kelas tertentu yang telah ditentukan sebelumnya. Anggota tersebut dimasukan kedalam kelas tertentu berdasarkan persamaan karakter dari data terserbut. Teknik klasifikasi banyak digunakan dalam penerapan sistem klasifikasi untuk kasus tertentu. Dalam klasifikasi pembagian dataset dibagi menjadi 2 yaitu data latih dan data uji, dimana semua dataset akan dibagi kedalam dua cluster untuk dilakukan pelatihan dan pengujian terhadap data tersebut. Pembagian data tersebut juga akan menentukan hasil akurasi dari penerapan suatu metode dalam Teknik klasifikasi tersebut [19].

2.8. Evaluasi Confusion Matrix dan *Receiver Operating Characteristic (ROC) Curve*

Confusion Matrix adalah suatu metode untuk melakukan evaluasi dengan menggunakan tabel *matrix* yang digunakan pada konsep *data mining* untuk melakukan perhitungan akurasi [20]. Evaluasi dengan menggunakan fungsi *confusion matrix* akan menghasilkan nilai *accuracy, precision,* dan *recall*[18]. Nilai *accuracy* adalah *persentase* dari jumlah *record* data yang diklasifikasikan secara baik dan benar dengan menggunakan sebuah algoritma dan dapat membuat klasifikasi setelah dilakukan pengujian hasil klasifikasi tersebut. Nilai *precision* atau juga dikenal dengan nama *confidence value* merupakan proporsi dari jumlah kasus yang diprediksi mendapatkan hasil positif dimana nilainya juga akan positif pada data sebenarnya. Nilai dari *Recall* atau *sensivity value* merupakan proporsi

dari jumlah kasus yang bernilai positif yang sebenarnya dan diprediksi positif secara benar [18].

Tabel 2. 1 Model *Confusion Matrix*

Aktual	Classified as	
	+	-
+	True Positives (A)	False Negatives (B)
-	False Positives (C)	True Negatives (D)

Sumber: (Rosandy, 2016, p. 57)

Model confusion matrix pada tabel 2.1 dapat dijelaskan sebagai berikut, dimana:

True Positives (A) merupakan jumlah *record positive* yang diklasifikasikan sebagai *positive*.

False Negatives (B) merupakan jumlah *record positive* yang diklasifikasikan sebagai *negative*.

False Postifives (C) merupakan jumlah *record negative* yang diklasifikasikan sebagai *positive*.

True Negatives (D) merupakan jumlah *record negative* yang diklasifikasikan *negative*, lalu masukkan data uji.

Setelah data uji dimasukkan kedalam *confusion matrix*, maka dimehitung nilai-nilai yang telah dimasukkan tersebut untuk dihitung jumlah *accuracy*, *sensitivity* untuk mengukur proporsi 'positif' yang benar yang didiagnosa dengan benar, spesificity untuk mengukur proporsi 'negatif' yang benar yang didiagnosa dengan benar, PPV *(Positive Predictive Value)* adalah proporsi kasus dengan hasil tes 'positif' yang didiagnosa dengan benar, dan NPV *(Negative Predictive Value)* adalah proporsi kasus dengan hasil tes 'negatif' yang didiagnosa dengan benar.

Dapat dihitung menggunakan rumus [12]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2.7)$$

$$Sensitivity = \frac{TP}{TP+FN} \quad (2.8)$$

$$Spesificity = \frac{TN}{TN+F} \quad (2.9)$$

$$PPV = \frac{TP}{TP+FP} \quad (2.10)$$

$$NPV = \frac{TN}{TN+FN} \quad (2.11)$$

Keterangan:

TP : jumlah *true positives*

TN : jumlah *true negatives*

FP : jumlah *false positives*

FN : jumlah *false negatives*

Sedangkan fungsi *ROC Curve* adalah untuk memperlihatkan akurasi dan membandingkan klasifikasi secara *visual*. ROC mengekspresikan *Confusion Matrix*. ROC merupakan grafik dua dimensi dengan garis *horizontal* sebagai *false positive* dan garis *vertical* sebagai *true positive*. Secara teknis kurva ROC juga dikenal sebagai grafik ROC, dua dimensi grafik dimana tingkat TP diplot pada sumbu Y dan tingkat FP diplot pada sumbu X [18].

Hasil perhitungan dapat divisualisasikan dengan ROC Curve atau AUC *(Area Under Curve)*[12]. Berikut tingkat nilai diagnosa dari *ROC Curve* yaitu [8]:

a. Akurasi bernilai 0.90-1.00 sama dengan *Excellent Classification*

b. Akurasi bernilai 0.80-0.90 sama dengan *Good Classification*

c. Akurasi bernilai 0.70-0.80 sama dengan *Fair Classification*

d. Akurasi bernilai 0.60-0.70 sama dengan *Poor Classification*

e. Akurasi bernilai 0.50-0.60 sama dengan *Failure*

2.9. Hasil Penelitian yang Relevan

Berikut adalah beberapa hasil penelitian yang menjadi referensi dan memberikan banyak masukan kepada penulis :

Tabel 2. 2 Hasil Penelitian yang Relevan

No	Judul	Nama & Tahun	Jumlah & Atribut	*Jumlah Data (Record)*	*Algoritma*	*Selection Feature/ Optimasi*	*Preprocesing*	*Validation*	*Open Source*	Akurasi
1	*A Clinical support system for Prediction of Heart Disease using Machine Learning Techniques*	[5]	*14 Atribut, age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, num*	*303 Records*	*Naïve Bayes*	-	*Transformation data*	*cross validation*	*Heart disease UCI repository*	Hasil akurasi = 82.17%

2	*Algoritma Klasifikasi data mining naïve bayes berbasis Particle Swarm Optimization untuk deteksi penyakit jantung*	[6]	-	*300 Records*	*Naïve Bayes*	*PSO*	-	-	*medical check-up laboratorium*	Data dibagi 75% untuk data *training* dan 25% untuk data *testing* dengan hasil akurasi sebesar = 92.86%
3	*Perancangan Sistem Klasifikasi Penyakit Jantung Mengunakan Naïve Bayes*	[7]	*15 Atribut, age, trestbps, chol, sex, cpfbsthalach, exang, oldpeak, slope, c, thal, restecg, num*	*303 Records*	*Naïve Bayes*	-	-	*5-FOLD Cross Validation*	*Heart disease UCI repository*	Data kedalam 5 *subset* yaitu : 60 *record* = 88,62%, 120 *record* = 89,04%, [8]180 *record* = 91,48%, 240 *record* = 91,89%, 303 *record* = 92,02%
4	*Implementasi Algoritma Naïve Bayes Untuk Klasifikasi Penderita Penyakit Jantung Di Indonesia Menggunakan Rapid Miner*	[8]	*7 Atribut, usia, jk, tipe sakit dada, tekanan darah, kkolesterol, gula darah, detak jantung maksiamal*	*500 Records*	*Naïve Bayes*	-	*Data Selection, Data Cleaning, dan Data Transformation*	*split validation*	*www.kaggle.com/ & archive.isuuci.edu/*	Data dibagi 80% *data training* dan 20% *data testing* dengan hasil akurasi sebesar = 70.00%
5	*Analisis Komparasi Algoritma Klasifikasi Data Mining Untuk Prediksi Penderita Penyakit Jantung*	[21]	*8 Atribut , atribut Tidak disebutkan*	-	*Decision Tree, Naïve Bayes, k- Nearest Neighbour, Random Forest, dan Decison Stump*		-	*cross validation*	*Heart disease UCI repository*	*Decision Tree*n = 60.08%, *Naïve Bayes* = 60.77%, *k-Nearest Neighbour* = 78.95%, *Random Forest* = 80.38%, *dan Decison*

										Stump = 78.95%.
6	*Penerapan Algoritma Klasifikasi Nearest Neighbor (K-NN) Untuk Mendeteksi Penyakit Jantung*	[22]	*13 Atribut, Age, Sex, Chest Pain Type, Tresting Blood Pressure, Serum cholestoral dalam mg/dl, Fasting blood sugar>120 mg/dl, Resting electrocardiographic result, The Slope of the peak exercise ST segment, Excercise Induced Angina, Old Peak, CA (Number of Major Vessels), Maximum Heart Rate, Achieved (Thalac), Thal*	*110 Records*	*Klasifikasi Nearest Neighbor (K-NN)*	-	-	-	*University of California Irvine Machine learning data repository*	*KKN k = 9* data dibagi *100 records (training data)* dan *10 records (testing data).* dengan hasil akurasi = 70%.

Dari review beberapa jurnal baik jurnal internasional maupun nasional pada tabel diatas ada beberapa belum dilakukan validasi model untuk mengevaluasi kinerja model atau algoritma. Untuk itu peneliti ingin mengembangkan model teknik validasi menggunakan *cross validation* dengan fitur *Particle Swarm Optimization* yang dikombinasikan pada algoritma *naïve bayes* untuk memperoleh hasil akurasi yang maksimal, kemudian setelah dilakukan pengukuran dengan *cross validation* dan fitur *Particle Swarm Optimization*, maka algoritma yang memiliki tingkat akurasi tinggi yang akan digunakan untuk memprediksi gejala penyakit jantung.