BAB II
LANDASAN TEORI

2.1 Penelitian Terkait

Beberapa penelitian terkait dengan menggunakan teknik Data Mining dengan penerapan Algoritma *Decision Tree* C.45 diantaranya:

Tabel 2. 1 Review Jurnal

No	Judul, Penulis, Tahun	Dataset	Metode	Hasil	Kekurangan	Kelebihan
1	Perbandingan Klasifikasi Metode *Naive* Bayes dan Metode *Decision Tree* Algoritma (J48) pada Pasien Penderita Penyakit *Stroke* di RSUD Abdul Wahab Sjahranie Samarinda, Irene Lishania 2019	Data set pada penelitian ini diambil dari RSUD Abdul Wahab Sjahranie Samarinda berupa data Kuantitatif	Menggunakan Metode *Naive* Bayes dan Metode *Decision Tree*	Dari 12 atribut yang terdapat dalam dataset yaitu Jenis, kelamin, tekanan darah, diabetes melitus, dislipidemia, kadar asam urat, penyakit jantung, status stroke. Penelitian ini menggunakan perbandingan Algoritma *Naive* Bayes dan *Decision Tree*.	1. Tidak ada grafik 2. Dalam penelitian, ada terlalu banyak konten teoretis dalam diskusi 3. Hasil akurasi masil belum optimal yaitu *Naive Bayes* sebesar 81,25 dan *Decission Tree* 87,25	- Jurnal menggunakan bahasa yang mudah dipahami - Penelitian ini memberikan informasi hasil akurasi prediksi dengan perbandingan algoritma dan dapat mengetahui penggunaan algoritma terbaik atau tertinggi akurasinya.

2	PENERAPAN ALGORITMA DECISION TREE C4.5 UNTUK DIAGNOSA PENYAKIT STROKE DENGAN KLASIFIKASI DATA MINING PADA RUMAH SAKIT SANTA MARIA PEMALANG, Sigit Abdillah 2011	Data set pada penelitian ini diambil dari Rumah Sakit Santa Maria Pemalang berupa data Kuantitatif	DECISION TREE C4.5	Pada pemodelan algoritma C4.5 dengan menggunakan pembobotan atribut juga menghasilkan akurasi sebesar 76,92% .	1. Tidak menyajikan hasil per metode secara detail 2. Tidak ada grafik 3. Hasil akurasi kurang optimal.	1. Jurnal menggunakan data kuantitatif yang mudah dipahami 2. Jurnal menggunakan bahasa yang mudah dipahami
3	Klasifikasi Penurunan Fungsi Kognitif Pasien Stroke Menggunakan Metode Klasifikasi *Random Forest*, Muhammad Shidqi Fadlilah, 2019	Data set pada penelitian ini berupa data Kuantitatif	*Random Forest*	Hasil penelitian dengan algoritma *random forest* maka ditetapkan bahwa akurasi Hasil rata-rata akurasi yang didapatkan dari semua percobaan adalah 53,094%	1. Jumlah dataset sedikit 2. Hasil akurasi kurang optimal. 3. Tidak dicantumkan atribut yang terlibat pada dataset	1. Jurnal menggunakan data kuantitatif yang mudah dipahami 2. Jurnal menggunakan bahasa yang mudah dipahami
4	Penerapan Metode Adaboost Untuk Mengoptimasi Prediksi Penyakit Stroke Dengan Algoritma Naïve Bayes, Agus Byna 2020	Data set pada penelitian ini diambil dari situs web www.kaggle.com adapun variabel dari data tersebut adalah	Metode Adaboost Dengan Algoritma Naïve Bayes	Pada penelitian ini didapat hasil pengujian, dengan dilakukan evaluasi baik secara confusion matrix beberapa algoritma yang diujikan terlihat hasil yang memiliki akurasi paling tinggi adalah optimasi adaboost dengan	1. Tidak dijelaskan mendetail mengenai perhituangannya 2. Pembahasan yang di berikan singkat	1. Dataset yang di dipergunakan cukup baik 2. Dari penelitian ini kita bisa menyimpulkan perbandingan algoritma murni naive bayes dan di tambahkan

		jenis kelamin, umur, hipertensi, serangan jantung, status pernikahan, pekerjaan, tempat tinggal, rata-rata gula darah, BMI, status merokok		naïve bayes akurasi sebesar 0,981 ini lebih besar dari algoritma naïve bayes yang belum di optimasi ada selisih perbedaan sebelum di optimasi dan sesudahnya. Selisihnya adalah 0,005		metode adaboost
5	KOMPARASI PENERAPAN METODE BAGGING DAN ADABOOST PADA ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT STROKE, NUR DIANA SAPUTRI 2021	Dataset yang digunakan adalah *Dataset Stroke Disease* dari situs web Kaggle 2021 version 1 dengan atribut id, gender, age, hypertension, heart disease, ever married, work_type, residence, avg_glucose_level, BMI, Smoking_Status	Metode bagging Dan adaboost pada algoritma C4.5	Berdasarkan hasil perbandingan anlisa, algoritma C4.5 menghasilkan nilai akurasi sebesar 92,87%, nilai presisi 27 %, spesifisitas 98%, sensivisitas 13%, dan f1score sebesar 18%. Sedangkan akurasi dari algoritma C4.5 setelah menerapkan metode bagging adalah 95,02%, nilai presisi 50%, spesifisitas 100%, sensivisitas 3%, dan f1score sebesar 5% serta akurasi dari algoritma C4.5 setelah menerapkan	Untuk menambah akurasi algoritma, akan lebih baik apabila dioptimasi dengan menambahkan Bagging Dan Adaboost Pada Algoritma C4.5	Teori dan model analisis yang digunakan sudah sesuai. Bahasa yang digunakan penulis mudah untuk memahami maksud dan tujuan pembaca. Analisisnya sangat detail dan mudah dipahami. detail penulis dalam memberikan hasil yang diperoleh dalam melakukan penelitiannya

		dan stroke		metode adaboost adalah sebesar 94,63% dengan nilai presisi 46%, spesitifitas 100%, sensivisitas 8%, dan f1score sebesar 13%.		
6	KOMPARASI ALGORITMA C4.5 BERBASIS PSO DAN GA UNTUK DIAGNOSA PENYAKIT STROKE, Ramdhan Saepul Rohman 2020	dataset repository yang dapat diperoleh melalui alamat web https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data sedangkan data pendukung didapatkan dari buku, jurnal dan publikasi lainnya	Algoritma C4.5 Berbasis PSO	Hasil dari pengujian diatas menjelaskan bahwa pengujian C4.5 menjadi lebih baik ketika menggunakan optimasi *Particle Swarm Optimization* dan *Genetik Algorithm.* Terbukti dengan menggunakan C4.5 berbasis PSO maka akurasi yang diperoleh sebesar 91,63% dibanding dengan C4.5 saja yang hanya mendapatkan 89,93%. Namun meskipun demikian tingkat akurasi yang didapatkan jauh lebih tinggi menggunakan GA dibanding dengan menggunakan PSO dengan tingkat akurasi sebesar 92,02%	Untuk menambah akurasi algoritma, akan lebih baik apabila dioptimasi dengan menambahkan GA Pada Algoritma C4.5	1. Dataset yang di dipergunakan cukup baik 2. Dari penelitian ini kita bisa menyimpulkan perbandingan algoritma murni C4.5 dan di tambahkan metode PSO dan GA

Dari hasil review beberapa jurnal diatas baik jurnal nasional dan internasional maka dapat disimpulkan untuk mendapatkan akurasi yang baik maka data yang dibutuhkan pada penelitian yaitu minimal 500 record data, dan metode yang digunakan untuk beberapa jurnal yang sudah di review tingkat akurasinya cenderung lebih tinggi menggunakan algoritma *Decision Tree* C4.5 dengan mencapai akurasi 92.%. hasil perbandingan akurasi dapat dilihat dibawah ini:

Tabel 2. 2 Hasil Akurasi Penelitian Sebelumnya

No	Jurnal	Algoritma	Akurasi
1	Irene Lishania 2019	*Naive Bayes* *Decission Tree* J48	81,25 % 87,25 %
2	Sigit Abdillah 2011	C4.5	76,92% .
3	Muhammad Shidqi Fadlilah**,** 2019	*Random Forest*	53,094%
4	Agus Byna 2020	Naïve Bayes	80%
5	NUR DIANA SAPUTRI 2021	C4.5	92,87%,
6	Ramdhan Saepul Rohman 2020	C4.5	89,93%.

2.2 Landasan Teori

2.2.1 Penyakit Stroke

a. Pengertian Penyakit Stroke

Stroke adalah serangan otak yang timbul secara mendadak dimana terjadi gangguan fungsi otak sebagian atau menyeluruh sebagai akibat dari gangguan aliran darah oleh karena sumbatan atau pecahnya pembuluh darah tertentu di otak, sehingga menyebabkan sel-sel otak kekurangan darah, oksigen atau zat-zat makanan dan akhirnya dapat terjadi kematian sel-sel tersebut dalam waktu relatif singkat [8].

Stroke merupakan gangguan fungsi saraf pusat yang berkembang sangat cepat baik menit maupun jam dengan perburukan ringan sampai berat kemudian menetap atau bahkan membaik secara cepat atau perlahan-lahan tergantung tingkat keparahan stroke dan cepat serta tepatnya intervensi pengobatan. Karena setiap bagian otak memiliki fungsi-fungsi tertentu, maka gejala dan tanda stroke pada setiap individu sangat bervariasi, tergantung pembuluh darah mana yang terkena dan bagian otak mana yang terganggu [8].

b. Faktor Risiko Penyakit Stroke

Faktor risiko stroke adalah kondisi atau penyakit atau kelainan yang terdapat pada seseorang yang memiliki potensi untuk memudahkan orang tersebut mengalami serangan stroke pada suatu saat [8]. Jika seseorang terdapat faktor-faktor risiko untuk terjadinya serangan stroke disebut sebagai stroke prone profile [3]. Terdapat dua macam faktor risiko penyakit stroke, yaitu faktor risiko yang dapat diubah atau dikendalikan dan faktor risiko yang tidak dapat diubah atau dikendalikan. Faktor risiko yang tidak dapat diubah atau dikendalikan [9] meliputi:

1. Usia

Stroke sering terjadi pada orang yang telah lanjut usia (tua). Setiap penambahan 10 tahun setelah usia 55 tahun, terdapat peningkatan risiko penyakit stroke sebanyak dua kali lipat.

2. Jenis Kelamin

Stroke lebih mungkin pada pria dibandingkan pada wanita. Namun, lebih dari separuh kematian stroke total yang terjadi pada wanita. Penggunaan pil KB dan kehamilan meningkatkan risiko stroke bagi perempuan.

3. Ras

Kematian akibat penyakit stroke lebih banyak terjadi pada orang Afrika-Amerika daripada orang kulit putih. Hal ini dikarenakan mereka mempunyai risiko lebih tinggi menderita tekanan darah tinggi, diabetes, dan obesitas. Sedangkan faktor risiko yang dapat diubah atau dikendalikan (Valensia, 2015) meliputi:

4. Tekanan Darah

Tekanan darah adalah tekanan yang terjadi pada pembuluh darah arteri saat darah dipompa oleh jantung untuk dialirkan ke seluruh tubuh. Tekanan darah sistolik adalah tekanan darah yang terjadi pada saat otot jantung berkontrakasi. Sedangkan tekanan darah diastolik adalah tekanan darah yang terjadi pada saat otot jantung beristirahat atau tidak sedang berkontraksi

5. Kadar Gula Darah

Gula darah adalah bahan bakar tubuh yang dibutuhkan untuk kerja otak, sistem saraf, dan jaringan tubuh yang lain. Gula darah yang terdapat di dalam tubuh dihasilkan oleh makanan yang mengandung karbohidrat,

protein, dan lemak. Rata-rata, kadar gula darah normal adalah sebagai berikut:

a. Gula darah 8 jam sebelum makan atau setelah bangun pagi (70-110 mg/dl).

b. Gula darah 2 jam setelah makan (100-150 mg/dl).

c. Gula darah acak (70-125 mg/dl).

6. Kadar Kolesterol Total

Kolesterol total merupakan kadar keseluruhan kolesterol yang beredar dalam tubuh manusia. Kolesterol adalah lipid amfipatik dan merupakan kompenen struktural esensial pada membran plasma. Senyawa kolesterol total ini disintesis dibanyak jaringan dari asetil-KoA dan merupakan prekusor utama semua steroid lain di dalam tubuh termasuk kortikosteroid, hormone seks, asam empedu, dan vitamin D.

7. Low Density Lipoprotein (LDL)

Kolesterol LDL disebut sebagai kolesterol jahat disebabkan peranannya membawa kolesterol total ke banyak jaringan di dalam tubuh. Sehingga memberikan peluang terjadinya penumpukan kolesterol di berbagai jaringan tubuh, termasuk diantaranya dalam pembuluh darah.

8. Asam Urat

Penyakit asam urat adalah penyakit yang timbul akibat kadar asam urat darah yang berlebihan. Yang menyebabkan kadar asam urat darah

berlebihan adalah produksi asam urat di dalam tubuh lebih banyak dari pembuangannya. Organ yang bisa terserang adalah sendi, otot, jaringan di sekitar sendi, telinga, kelopak mata, jantung, ginjal, dan lain-lain.

9. Blood Urea Nitrogen (BUN)

Blood Urea Nitrogen (BUN) dapat didefinisikan sebagai jumlah nitrogen urea yang hadir dalam darah. Urea adalah produk limbah yang dibentuk dalam tubuh selama proses pemecahan protein. Selama metabolisme protein, protein diubah menjadi asam amino yang juga menghasilkan amonia. Urea tidak lain adalah substansi yang dibentuk oleh beberapa molekul amonia. Metabolisme protein berlangsung dalam hati dan dengan demikian urea juga diproduksi oleh hati. Selanjutnya, urea ditransfer ke ginjal melalui aliran darah dan dikeluarkan dari tubuh dalam bentuk urin. Dengan demikian, setiap disfungsi ginjal akan menyebabkan kadar tinggi atau rendah BUN dalam darah.

10. Kreatinin (creatinine)

Kreatinin (creatinine) adalah produk penguraian dari kreatin fosfat dalam metabolisme otot dan dihasilkan dari kreatin (creatine). Kreatinin pada dasarnya merupakan limbah kimia yang selanjutnya diangkut ke ginjal melalui aliran darah untuk dikeluarkan melalui urin. Kadar kreatinin dapat diukur dalam urin serta darah. Tingkat kreatinin dalam darah umumnya tetap normal. Dengan demikian, ginjal yang berfungsi normal juga akan menunjukkan tingkat normal kreatinin dalam darah. Tapi ketika ginjal tidak berfungsi dengan baik, jumlah kreatinin dalam darah akan meningkat.

2.2.2 Data Mining

Data mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terkait dari berbagai database besar. (Turban, 2005), Proses Data mining menerapkan teknik statistik, machine learning, matematika, dan kecerdasan buatan dalam mengekstraksi dan identifikasi pengetahuan dari berbagai basis data besar. Data mining dapat pula didefenisikan rangkaian proses penggalian nilai tambah dari suatu perkumpulan data yakni nilai informasi yang belum diketahui cara manual.

Definisi umum dari data mining itu sendiri adalah proses pencarian polapola yang tersembunyi (hidden patern) berupa pengetahuan (knowledge) yang tidak diketahui sebelumnya dari suatu sekumpulan data yang mana data tersebut dapat berada di dalam database, data werehouse, atau media penyimpanan informasi yang lain. Hal penting yang terkait di dalam data mining adalah:

1. Data mining merupakan suatu proses otomatis terhadap data yang sudah ada.

2. Data yang akan diproses berupa data yang sangat besar.

3. Tujuan data mining adalah mendapatkan hubungan atau pola yang mungkin memberikan indikasi yang bermanfaat [10].

Data mining memilki beberapa teknik yang terkenal dan sering digunakan oleh peneliti, diantaranya seperti clustering, classification, association, dan

beberapa perkembangan teknik sesuai dengan perubahan kecendrungan data pada saat ini.

Pola yang disajikan mudah dipahami berlaku untuk data yang akan diprediksi dengan derajat kepastian tertentu, penggalian datanya dengan memiliki beberapa nama alternatif meskipun eksaknya berbeda seperti KDD (*Knowledge Discovery in Database*) , analisis pola, arkeologi data, pemanenan informasi, intelegensia bisnis. Data mining dikelompokan menjadi beberapa kelompok yaitu:

1. Deskripsi Menggambarkan pola yang terdapat dalam data yang memberikan kemungkinan penjelasan untuk suatu pola atau kecendrungan.

2. Estimasi Estimasi hampir sama dengan klasifikasi, kecuali variabel target estimasi lebih ke arah numerik dari pada ke arah kategori. Dibangun dengan record lengkap nilai dari variabel target sebagai nilai prediksi. Selanjutnya, pada peninjauan berikutnya estimasi nilai dari variabel target dibuat berdasarkan nilai variabel prediksi.

3. Prediksi Prediksi nilai dari hasil akan ada di masa mendatang .

4. Klasifikasi Dalam klasifikasi, terdapat target variabel kategori.

5. Pengklusteran Pengklusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan.

2.2.3 Klasifikasi

Klasifikasi adalah urutan yang sangat penting dalam data komunitas pertambangan. Klasifikasi adalah salah satu prediksi teknik data mining yang membuat prediksi tentang data nilai menggunakan hasil yang diketahui yang ditemukan dari kumpulan data yang berbeda. Masalah akurasi dari banyak algoritma klasifikasi adalah diketahui mengalami penurunan informasi saat dihadapi dengan data yang tidak seimbang, misalnya ketika distribusi sampel lintas kelas sangat miring [11]. Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, yang, misalnya, dapat dipartisi menjaditiga kelas atau kategori: berpenghasilan tinggi, menengah pendapatan, dan pendapatan rendah. Model data mining memeriksa satu set besar catatan, masing-masing catatan yang berisi informasi tentangvariabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnisdan penelitian meliputi :[12].

a. Menentukan apakah transaksi kartu kredit tertentu adalah penipuan

b. Menempatkan siswa baru pada jalur tertentu yang berkaitan dengan kebutuhankhusus

c. Menilai apakah aplikasi hipotek adalah risiko kredit yang baik atau buruk

d. Mendiagnosis apakah ada penyakit tertentu

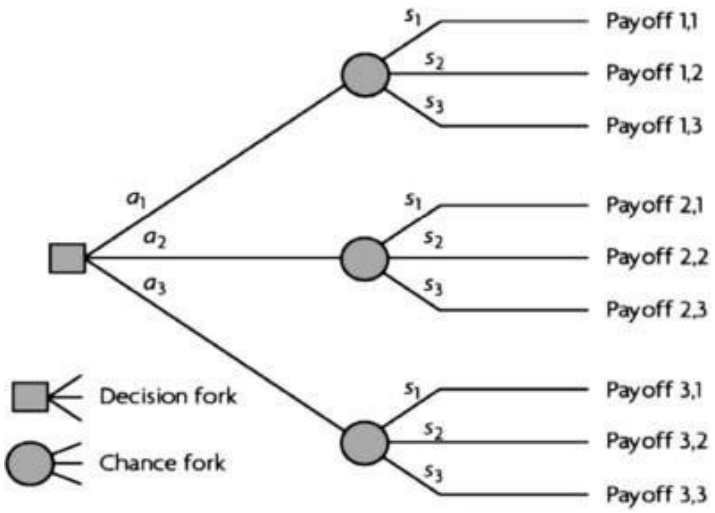
e. Menentukan apakah surat wasiat ditulis oleh almarhum yang sebenarnya, ataucurangoleh orang lain

f. Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkankemungkinan ancaman teroris

Klasifikasi yang dilakukan secara manualadalahklasifikasi yang dilakukan oleh manusiatanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, SupportVector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan [13].

2.2.4 *Decision Tree*

Decision Tree adalah struktur flowchart yang menyerupai Tree (pohon), dimana setiap simpul internal menandakan suatu tes pada atribut, setiap cabang merepresentasikan hasil tes,dan simpul daun merepresentasikan kelas atau distribusi kelas. Alur pada *Decision Tree* di telusuri dari simpul akar ke simpul daun yang memegang prediksi [14].



Gambar 2.1 Bentuk *Decision Tree* Secara umum

Decision tree memiliki training sample berupa sekumpulan data yang nantinya akan digunakan untuk membangun sebuah tree yang telah diuji kebenarannya. Secara umum *Decision Tree* adalah untuk membangun pohon keputusan sebagai berikut :

a. Pilih atribut sebagai akar

b. Buat cabang untuk setiap nilai

c. Bagi kasus dalam cabang

d. d. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang yang memiliki kelas yang sama. Rumus menghitung nilai entropy menggunakan persamaan:

$$\text{Entropy (S)} \sum_{i=1}^{n} -pi\ log2\ pi \qquad (2.1)$$

Keterangan :

S = himpunan kasus

n = jumlah partisi atribut A

Pi = proporsi Si terhadap S

|Si| = jumlah kasus pada partisi ke i

|S| = jumlah kasus dalam S

A= atribut Rumus untuk mencari nilai gain :

$$\text{Gaint (S,A)} \sum_{f-1} \frac{|si|}{|s|} \text{Entropy (Si)} \qquad (2.2)$$

2.2.5 *Split Validation*

Split Validation adalah teknik validasi yang membagi data menjadi dua bagian secara acak, sebagian sebagai data training dan sebagian lainnya sebagai data testing. Dengan menggunakan *Split* Data akan dilakukan percobaan training berdasarkan *split ratio* yang telah ditentukan sebelumnya, untuk kemudian sisa dari *split ratio* data *training* akan dianggap sebagai data *testing*. Data *training* adalah data yang akan dipakai dalam melakukan pembelajaran sedangkan data *testing* adalah data yang belum pernah dipakai sebagai pembelajaran dan akan berfungsi sebagai data pengujian kebenaran atau keakurasian hasil pembelajaran [7].

2.2.6 Seleksi Fitur

Seleksi fitur merupakan proses yang melibatkan subset dari kumpulan fitur yang menghasilkan keluaran seperti keseluruhan kumpulan fitur. Seleksi fitur biasanya digunakan untuk memilih fitur yang optimal, mereduksi dimensi, meningkatkan akurasi algoritma klasifier, dan menghapus fitur yang tidak relevan [15]. Tujuan utama dari seleksi fitur adalah untuk mengurangi jumlah fitur yang digunakan dalam klasifikasi dengan tetap menjaga akurasi klasifikasi yang dapat diterima. Pemilihan fitur dapat berdampak besar pada keefektifan algoritma klasifikasi yang dihasilkan, dalam beberapa kasus, sebagai hasil dari pemilihan fitur, akurasi klasifikasi yang akan datang dapat ditingkatkan [6].

Manfaat melakukan pemilihan fitur sebelum memodelkan data Anda adalah sebagai berikut:

a. Mengurangi Overfitting: Data yang lebih sedikit berarti lebih sedikit kesempatan untuk membuat keputusan berdasarkan noise.
b. Meningkatkan Akurasi: Data yang kurang menyesatkan berarti akurasi pemodelan meningkat.
c. Mengurangi Kompleksitas: lebih sedikit titik data mengurangi kompleksitas algoritme dan membuatnya lebih mudah dipahami.
d. Pelatihan Lebih Cepat: Ini memungkinkan algoritme pembelajaran mesin untuk berlatih lebih cepat. 18
e. Pada sistem ini digunakan dua proses seleksi fitur yaitu proses seleksi fitur sekuensial/forward dan proses seleksi fitur mundur. Pada sistem ini digunakan dua proses seleksi fitur yaitu proses seleksi fitur sekuensial/forward dan proses seleksi fitur mundur.

2.2.7 *Particle Swarm Optimization* (PSO)

Particle swarm optimization adalah salah satu optimasi yang dapat digunakan untuk pengambilan keputusan. PSO adalah teknik optimasi dengan cara menghitung terus menerus calon solusi dengan menggunakan suatu acuan kualitas. PSO mengoptimasi permasalahan dengan cara menggerakkan partikel atau calon solusi di dalam permasalahan menggunakan fungsi tertentu untuk posisi dan kecepatan dari partikel. Pergerakan pertikel dipengaruhi oleh solusi terbaik dari partikel tersebut, dan solusi terbaik secara umum yang didapatkan dari partikel lain. Sekumpulan partikel ini dinamakan *swarm*, *swarm* ini akan bergerak menuju solusi terbaik.

vn+1 = vn + $c1rand() * (pbest,n - CurrentPositionn) + c2rand2() * (gbest,n - CurrentPositionn)$

Particle Swarm Optimization (PSO) adalah teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter. Dalam *Particle Swarm Optimization* (PSO), terdapat beberapa teknik untuk optimasi antara lain meningkatkan bobot atribut dari semua atribut atau variabel yang digunakan, memilih atribut (*attribute selection*), dan seleksi fitur. Particle swarm optimization adalah suatu algoritma yang banyak terinspirasi dari perilaku sosial hewan seperti burung, lebah dan ikan. Seekor hewan dalam algoritma PSO akan dianggap sebagai partikel. Partikel ini akan dipengaruhi oleh kecerdasan dari individu hewan itu sendiri dan dan kecerdasan dari partikel lain dalam satu kelompok. Apabila satu partikel menemukan jalan yang tepat dan terpendek menuju ke suatu sumber makanan, maka yang terjadi adalah partikel-partikel lain tersebut akan mengikuti partikel yang telah menemukan jalan yang tepat dan terpendek tadi [16].

Secara garis besar prosedur PSO dapat dilakukan dalam beberapa langkah.

1. Inisialisasi kecepatan awal bernilai 0 untuk semua partikel seperti pada Persamaan 13.

   (Vi,j(t)=0)

   Vi,j merupakan kecepatan, j adalah letak partikel dan i adalah letak individu dan t adalah iterasi.

2. Inisialisasi posisi awal partikel dengan batasan sesuai range [xmin,max]. proses inisialisasi posisi terdapat pada Persamaan 14.

   $x(t)=x\min+r(x\max-x\min)$

   X merupakan posisi partikel dan r adalah nilai random

3. Inisialisasi Pbest dan Gbest awal dimana pada iterasi ke 0 nilai Pbest sama dengan posisi awal sesuai dengan Persamaan 15 dan Gbest merupakan Pbest dengan nilai fitness terbaik.

   (Pbesti,j(t)=xi,j(t))

   Pbest merupakan personal best pada individu ke-i dan partikel ke-j. Xij merupakan posisi partikel

4. Update kecepatan dilakukan untuk menentukan arah perpindahan posisi partikel yang ada di populasi. Kecepatan dihitung sesuai Persamaan 16. Terdapat Batasan untuk kecepatan yang digunakan yaitu berdasarkan nilai maksimum dan minimum posisi partikel untuk menentukan batas kecepatan maksimum dan minimum yang dipengaruhi oleh interval (k) yang sebaiknya dilakukan pada proses inisialisasi. Proses update dilakukan seperti pada Persamaan 7 dan 8.

   $t + 1\ t\ t\ t\ t\ t$

   $$\mathbf{v}\ i, j = w.\ vi, j + c\mathbf{1}r\mathbf{1}(Pbesi, j - xi, j) + c\mathbf{2}r\mathbf{2}(Gbesg, j - xi, j) \quad (2.3)$$

   $$vj\ max = k(xj,max-xj,\min)\ k\ \epsilon(0,1] \quad (2.4)$$

$if\ vij(t+1) > vmaxj\ then\ vij(t+1) = vj\ max\ max$

$if\ vij(t+1) < -v\ j\ then\ vij(t+1)\ \ max\ = -v\ j$

Nilai c1 dan c2 adalah koefisien akselerasi, nilai r1 dan r2 adalah partikel random, nilai w adalah bobot inertia.

5. Update posisi dilakukan untuk menentukan posisi terbaru dari setiap partikel berdasarkan hasil update kecepatan sebelumnya. Setelah didapatkan nilai kecepatan maka dilanjutan dengan perhitungan sigmoid dari kecepatan tersebut sesuai dengan Persamaan 9 Kemudian hasil signoid yang telah didapat akan diproses lebih lanjut pada Persamaan 10 sehingga didapatkan posisi terbaru Setelah itu menentukan hasil fitness terbaru yang tentunya juga akan mendapat nilai Pbest terbaru.

$sig(vi, j) = 2 + e1 - v\ t\ , j = 1,2, . . d$

$t\ t+1\ \ if\ rand[0,1] > sig(vi, j)\ then\ x\ i, j = 0$

$t\ t+1\ \ if\ rand[0,1] < sig(vi, j)\ then\ x\ i, j = 1$

$j = 1,2, \ldots . d$

6. Update Pbest, yaitu dengan membandingkan nilai fitness dari Pbest pada iterasi sebelumnya dengan fitness dari update Posisi. Nilai yang terbaik akan menjadi Pbest yang baru pada iterasi selanjutnya.

$k = 1 + decimal(s1)\ x\ a - 1$ (21)

$2n1 - 1$

2.2.8 Akurasi

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal,akurasiadalah sebagian kecil dari prediksi model kami yang benar. Secara formal, akurasi memiliki definisi sebagai berikut :

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}}$$

Untuk klasifikasi biner, akurasi juga dapat dihitung dalam hal positif dan negatif sebagai berikut :

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN}$$

Dimana

TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.2.9 *Precision*

Precision menggambarkan tingkat keakuratan antar data yang diminta dengan hasil prediksi yang diberikan oleh model, maka *precision* merupakan rasio prediksi benar positif dibandingkan dengan hasil yang diprediksi positif. Dari semua kelasi positif yang telah diprediksi dengan benar, berapa banyak data yang benar-benar positif [17]. Adapun rumus untuk mencari *precision* sebagai berikut:

$$precision = \frac{1}{2} \times \left(\frac{negatif - F.N}{negatif - F.N + F.P} + \frac{positif - F.P}{positif - F.P + F.N} \right) \times 100\%$$

2.2.10 *Recall*

Recall menggambarkan keberhasilan model dalam menentukan Kembali sebuah informasi. Maka *recall* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif [17]. Adapun rumus untuk mencari *recall* sebagai berikut:

$$recall = \frac{1}{2} \times \left(\frac{negatif - F.N}{negatif} + \frac{positif - F.P}{positif} \right) \times 100\%$$

2.2.11 *Confusion Matrix*

Matriks konfigurasi adalah tabel yang terdiri dari jumlah baris data uji yang diprediksi benar dan salah dengan model klasifikasi yang digunakan. Tabel *Confusion Matrix* diperlukan untuk memilih kinerja terbaik dari sebuah model klasifikasi [18].

2.2.12 Kurva ROC dan AUC

Dalam *Machine Learning*, pengukuran kinerja adalah tugas penting. Jadi dalam masalah klasifikasi, kita dapat mengandalkan Kurva AUC - ROC. Ketika kita

perlu memeriksa atau memvisualisasikan kinerja masalah klasifikasi multi-kelas, kita menggunakan kurva AUC (*Area Under The Curve*) ROC (*Receiver Operating Characteristics*). Ini adalah salah satu metrik evaluasi terpenting untuk memeriksa kinerja model klasifikasi apa pun. Itu juga ditulis sebagai AUROC (Area Di Bawah Karakteristik Operasi Penerima) [19].

Area di Bawah Kurva (AUC) adalah ukuran dari kesesuaian metode. AUC mewakili nilai sensitivitas dan spesifisitas dengan nilai batas 0 sampai 1 . Selanjutnya, hasil evaluasi ini dikategorikan berdasarkan nilai yang diperoleh dari setiap pengukuran [20]. Gorunescu [21] mengkategorikan hasil klasifikasi berdasarkan pada nilai AUC sebagai berikut:

• 0,90 – 1,00 = klasifikasi sangat baik;

• 0,80 – 0,90 = klasifikasi baik;

• 0,70 – 0,80 = klasifikasi wajar;

• 0,60 – 0,70 = klasifikasi buruk;

• 0,50 – 0,60 = gagal.