BAB II
TINJAUAN PUSTAKA

2.1. Penelitian Terkait

Penelitian sebelumnya yang menjadi latar belakang penelitian ini dijabarkan pada tabel di bawah ini:

No	Judul, Penulis, Tahun	Dataset	Metode	Hasil
1	Implementasi Data Mining Untuk Prediksi Penyakit Diabetes Dengan Algoritma C4.5 Sanni Ucha Putri, Eka Irawan, Fitri Rizky 2021 [1]	- Usia (Tahun) - Tekanan Darah (NmHg) - Denyut Nadi (x/menit) - Berat Badan (kg) - Kadar Gula Darah (mg/dl) - Variabel	Menggunakan Metode Decision Tree dengan Algoritma C4.5	Dari perhitungan terdapat delapan rules yang dapat dijadikan sebagai referensi dalam memprediksi penyakit diabetes. Adapun aturan rules berupa lima rules keputusan positif dan tiga rules keputusan negatif dapat dijabarkan melalui narasi sebagai berikut : a) Jika Kadar Gula Darah (mg/dl) = Normal dan Tekanan Darah (NmHg) = Rendah, maka Hasil = Positif b) Jika Kadar Gula Darah (mg/dl) = Normal dan Tekanan Darah (NmHg) = Tinggi, maka Hasil = Negatif

				c) Jika Kadar Gula Darah (mg/dl) = Rendah, maka Hasil = Negatif d) Jika Kadar Gula Darah (mg/dl) = Tinggi dan Berat Badan (kg)= Avarage dan Tekanan Darah (NmHg) = Normal, maka Hasil = Negatif e) Jika Kadar Gula Darah (mg/dl) = Tinggi dan Berat Badan (kg)= Avarage dan Tekanan Darah (NmHg) = Rendah, maka Hasil = Positif (Positif=1dan Negatif=1) f) Jika Kadar Gula Darah (mg/dl) = Tinggi dan Berat Badan (kg) = Avarage dan Tekanan Darah (NmHg) = Tinggi, maka Hasil = Positif (Positif=13 dan Negatif=1) g) Jika Kadar Gula Darah (mg/dl) = Tinggi dan Berat Badan (kg)= Over Weight, maka Hasil = Positif (Positif =18 dan Negatif=0)

				h) Jika Kadar Gula Darah (mg/dl) = Tinggi dan Berat Badan (kg)= Under Weight, maka Hasil = Negatif
2	Klasifikasi Faktor-Faktor Penyebab Penyakit Diabetes Melitus Di Rumah Sakit Unhas Menggunakan Algoritma C4.5 Dewi Rahma Ente, Sri Astuti Thamrin, Hedi Kuswanto, Samsul Arifin, Andreza 2020 [2]	- Jenis Kelamin (JK) - Usia - Berat Badan (BB) - Tinggi Badan (TB) - Glukosa Darah Puasa (GDP) - Kolesterol HDL - Kolesterol LDL - Kolesterol Total (Kol.Tot) - Trigliserida (Tg) - Status DM (Hasil)	Menggunakan Metode Decision Tree dengan Algoritma C4.5	Adapun aturan atau rule yang terbentuk yaitu jika GDP>126 mg/dl maka seseorang positif berpeluang mengidap DM (Diabetes Melitus). Jika GDP<126 mg/dl dan kolesterol LDL>110 mg/dl maka seseorang positif berpeluang mengidap DM. Jika kolesterol LDL<110 mg/dl dan GDP<101 maka seseorang negatif berpeluang DM. Jika kolesterol LDL< 110 mg/dl, GDP berada antara 101-126 mg/dl dan usia< 66 tahun maka peluang seseorang akan menderita penyakit DM itu kecil. Jika kolesterol LDL<110 mg/dl, GDP berada antara 101-126 mg/dl dengan usia>66 tahun maka

				seseorang berpeluang besar akan menderita penyakit DM. Berdasarkan dari rule yang didapatkan dengan algoritma C4.5 maka terdapat empat atribut yang dapat digunakan untuk mengidentifikasi faktor-faktor yang substansial mempengaruhi seseorang menderita penyakit DM yaitu kolesterol GDP, LDL, usia, dan berat badan. Pengukuran akurasi data latih dan data uji dari algoritma C4.5 dengan validasi silang lipat 10 setelah proses seleksi atribut dapat ketahui nilai akurasi. Nilai akurasi memiliki rentang antara 50% sampai dengan 100% dengan tingkat akurasi rata-rata prediksi yaitu 98,5%. Ini berarti model yang didapatkan sangat baik dengan tingkat akurasi sangat tinggi.

				Faktor-faktor yang mempengaruhi status DM secara substansial adalah glukosa darah puasa (GDP), kolesterol LDL, usia dan berat badan. Dengan mengetahui faktor-faktor yang mempengaruhi status DM penderita maka komplikasi serius akibat DM ini dapat dicegah sedini mungkin
3	Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5 Fida Maisa Hana 2020 [3]	- Umur - Jenis Kelamin - Polyuria - Polydipsia - Suddenweight loss - Weakness - Polyphagia - Genital thrush - Visual blurring - Itching - Irritability - Delayed healing - Partial paresis	Menggunakan Metode Decision Tree dengan Algoritma C4.5	Dari 16 atribut yang terdapat dalam dataset diabetes yaitu umur, Alopecia, Gender, Polyuria, Polydipsia, sudden weight loss, weakness, Polyphagia, Genital thrush, Irritability, delayed healing, partial paresis, Itching, visual blurring, muscle stiffness, dan Obesitas dapat dijadikan sebagai data untuk klasifikasi penderita penyakit diabates. Penelitian ini menggunakan Algoritma C4.5 untuk pengklasifikasian

		- Muscle stiffness - Alopecia - Obesitas - Kelas		seseorang terkena penyakit diabetes atau tidak. Dari 520 data dibagi menjadi 416 sebagai data training dan 104 sebagai data testing. Dari hasil Pengujian menghasilkan akurasi yang cukup besar yaitu 97,12 % Precision sebesar 93,02% %, dan Recall sebesar 100,00% Adapun Kurva ROC (Reciver Operating Characteristic) menunjukkan algoritma C4.5 memiliki nilai AUC sebesar 0.994 yang artinya Excellent Classification, ini menunjukkan bahwa menggunakan Algoritma C4.5 untuk klasifikasi penderita penyakit diabetes menghasilkan akurasi yang tinggi.
4	Optimasi Parameter K Pada Algoritma K-Nearest Neighbour Untuk Klasifikasi Penyakit	- Pregnant - Plasma-Glucose - Diastolic Blood-Pressure	Menggunakan Metode K-Nearest Neighbour	Berdasarkan hasil penelitian yang dilakukan terhadap dataset penyakit diabetes. Algoritma klasifikasi KNN dengan nilai k=13 menempati peringkat akurasi

	Diabetes Mellitus Indrayanti, Devi Sugianti, M. Adib Al Karomi 2017 [4]	- Tricepts Skin Fold Thickness - Insulin - Body Mass Index - Diabetes Pedigree Function - Age - Class Variable		tertinggi yaitu 75,14%. Hasil penelitian menunjukkan bahwa nilai K=13 merupakan nilai k yang paling optimal dengan tingkat akurasi sebesar 75,14%. Hasil prosentase ini didapatkan dari proses percobaan sebanyak 10 kali dengan data yang diacak (10 fold cross validation) kemudian hasil klasifikasi yang muncul dibandingkan dengan data sebenarnya. Perhitungan ini menggunakan confussion matrix untuk menentukan prosentase data yang sesuai dengan kenyataan dibandingkan jumlah keseluruhan data yang ada. Nilai k yang digunakan dalama penelitian ini seluruhnya adalah nilai ganjil dikarenakan label atau hasil akhir dari klasifikasi hanya memiliki 2 kemungkinan yaitu positif dan negatif.

				Penentuan hasil klasifikasi untuk nilai k lebih dari 1 digunakan metode hasil terbanyak atau mayoritas hasil klasifikasi
5	Penerapan Data Mining untuk Identifikasi Penyakit Diabetes Melitus dengan Menggunakan Metode Klasifikasi Faizal Aris, Benyamin 2019 [5]	- Berat Badan - Jenis Kelamin - Tekanan Darah - Kadar Gula Darah - Tipe Penyakit Diabetes	Menggunakan Metode Decision Tree dengan Algoritma C4.5	R1: IF glu sewaktu Normal AND TD Tinggi AND Usia >55 AND BB <=50 THEN B R2: IF glu sewaktu Normal AND TD Tinggi AND Usia >55 AND BB 51-60 THEN A R3: IF glu sewaktu Normal AND TD Tinggi AND Usia <=55 THEN B Klasifikasi data penderita diabetes dengan teknik data mining klasifikasi yang menggunakan algoritma C.45 menghasilkan rule yang dapat digunakan untuk prediksi penyakit diabetes. Hasil dari pohon keputusan berisi tentang analisa terhadap factor-faktor seseorang berpotensi terkena penyakit diabetes dengan melihat kepada atribut-atribut seperti jenis kelamin (jenis kelamin hanya

				sebagai atribut bantuan yang tidak dapat dijadikan perhitungan dalam memprediksi riwayat penyakit diabetes), berat badan, tekanan darah, dan kadar gula darah dan variabel riwayat penyakit turunan dan tidak turunan.
6	Perbandingan Algoritma Klasifikasi Data Mining Model C4.5 dan Naive Bayes Untuk Prediksi Penyakit Diabetes Fatmawati 2016 [6]	- Jumlah Hamil - Konsentrasi Glukosa - Tekanan Darah - Lipatan Kulit - Serum Insulin - IMB - Riwayat Diabetes - Umur - Hasil	Perbandingan C4.5 dan Naive Bayes	Hasil perbandingan antara C4.5 dan Naive Bayes diukur tingkat akurasinya menggunakan pengujian Confusion Matrix dan Kurva ROC. Berdasarkan hasil pengukuraan tingkat akurasi kedua algoritma tersebut, diketahui bahwa nilai akurasi C4.5 adalah 73.30% dan nilai AUC adalah 0.733, sedangkan nilai akurasi Naive Bayes 75.13% dan nilai AUC adalah 0.810 dapat disimpulkan bahwa dengan menggunakan model Naive Bayes lebih tinggi tingkat akurasinya, dengan peningkatan akurasi sebesar 1.83% dan

				peningkatan nilai AUC sebesar 0.077 sedangkan hasil pengujian dari prediksi diabetes hasilnya termasuk Good Clasification.

Tabel 2.1 Review Jurnal

Dari hasil review beberapa jurnal nasional tersebut, dapat saya tarik kesimpulan agar dataset dapat memiliki akurasi yang baik maka data yang dibutuhkan semakin banyak semakin baik, dengan jumlah data diatas 500 fitur, dan metode yang digunakan untuk beberapa jurnal tersebut tingkat akurasi cenderung lebih tinggi dengan menggunakan Decision Tree. Hasil dapat dilihat pada tabel berikut:

No Jurnal	Metode Terbaik Beserta Tingkat Akurasinya
1	Metode Decision Tree dengan Algoritma C4.5 (90,00%)
2	Metode Decision Tree dengan Algoritma C4.5 (97,50%)
3	Metode Decision Tree dengan Algoritma C4.5 (97,12%)
4	Metode K-Nearest Neighbour (75,14%)
5	Metode Decision Tree dengan Algoritma C4.5 (tidak dijelaskan tingkat akurasi)
6	Perbandingan Metode Decision Tree (75,13%) dan Naive Bayes (73,30%)

Tabel 2.2 Hasil Akurasi Review Jurnal

2.2. Diabetes

Diabetes yang juga dikenal di Indonesia dengan istilah penyakit kencing manis adalah sekelompok gangguan metabolisme yang ditandai dengan kadar gula darah yang tinggi selama periode waktu yang lama [7]. Gejala umum yaitu sering buang air kecil, haus meningkat, dan nafsu makan meningkat. Jika tidak diobati, diabetes dapat menyebabkan banyak komplikasi. Komplikasi akut dapat mencakup ketoasidosis diabetik, keadaan hiperglikemik hiperosmolar, atau kematian. Komplikasi jangka panjang yang serius yaitu penyakit kardiovaskular, stroke, penyakit ginjal kronis, borok kaki, kerusakan saraf, kerusakan mata, dan gangguan kognitif [7].

Glukosa adalah karbohidrat alamiah yang digunakan tubuh sebagai sumber energi. Kadar glukosa pada darah dikendalikan oleh beberapa hormon [8]. Insulin adalah hormon yang dibuat oleh pankreas. Ketika kita makan, pankreas membuat insulin untuk mengirimkan pesan pada sel-sel lainnya di tubuh. Insulin ini memerintahkan sel-sel untuk mengambil glukosa dari darah. Glukosa digunakan oleh sel-sel untuk pembuatan energi. Glukosa yang berlebih disimpan dalam sel-sel sebagai glikogen. Pada saat kadar gula darah mencapai tingkat rendah tertentu, sel-sel memecah glikogen menjadi glukosa untuk menciptakan energi.

Diabetes disebabkan oleh pankreas yang tidak memproduksi cukup insulin, atau sel-sel tubuh tidak merespons dengan baik terhadap insulin yang diproduksi [9]. Terdapat tiga jenis utama diabetes yaitu:

1. Diabetes melitus tipe 1 disebabkan karena pankreas gagal untuk memproduksi insulin yang cukup karena kehilangan sel beta. Jenis ini sebelumnya disebut sebagai "diabetes mellitus tergantung insulin" (IDDM) atau "diabetes remaja". Hilangnya sel beta disebabkan oleh respons autoimun. Penyebab respons autoimun ini tidak diketahui.
2. Diabetes melitus tipe 2 dimulai dengan resistensi insulin, suatu kondisi yang mana sel-sel gagal merespons insulin dengan baik. Seiring perkembangan penyakit, kekurangan insulin juga dapat terjadi. Bentuk ini sebelumnya disebut sebagai "diabetes mellitus non-dependen insulin" (NIDDM) atau "diabetes onset dewasa". Penyebab paling umum yaitu kombinasi dari berat badan berlebihan dan kurang olahraga.
3. Diabetes gestasional adalah bentuk utama ketiga, dan terjadi ketika wanita hamil sebelumnya tanpa riwayat diabetes, lalu mengalami kadar gula darah tinggi saat hamil.

Diabetes melitus tipe 1 harus dikelola dengan suntikan insulin. Pencegahan dan pengobatan diabetes tipe 2 melibatkan menjaga pola makan yang sehat, olahraga fisik secara teratur, berat badan normal, dan menghindari

penggunaan tembakau. Diabetes tipe 2 dapat diobati dengan Antidiabetik oral seperti sensitizer insulin dengan atau tanpa insulin. Kontrol tekanan darah dan menjaga perawatan kaki dan mata secara baik merupakan langkah penting bagi penderita penyakit ini. Insulin dan beberapa obat oral dapat menyebabkan gula darah rendah. Operasi penurunan berat badan pada orang-orang yang mengalami obesitas kadang-kadang merupakan upaya yang efektif pada mereka yang menderita diabetes tipe 2. Diabetes gestasional biasanya sembuh setelah kelahiran bayi

2.3. Data Mining

Data mining adalah suatu proses penggalian terhadap data berukuran besar yang belum diketahui sebelumnya [10]. Data mining juga didefinisikan sebagai bagian dari proses penggalian pengetahuan dalam database yang sering disebut dengan istilah Knowledge Discovery in Database (KDD) [11]. KDD menyajikan data mining yang terstruktur dengan baik dan proses standar, berhubungan erat dengan manajer, pengambil keputusan, dan mereka terlibat dalam menyebarkan hasil. Pertumbuhan luar biasa yang sedangberlangsung di bidang data mining dan Knowledge Discovery telah didorong oleh penemuan dari berbagai hal [11]:

a. Pertumbuhan eksplosif dalam pengumpulan data, seperti yang dicontohkan oleh pemindaian di Supermarket
b. Penyimpanan data di gudang data, sehingga seluruh perusahaan memiliki akses kedatabase terkini yang andal
c. Ketersediaan peningkatan akses ke data dari navigasi Web dan intranet
d. Tekanan persaingan untuk meningkatkan pangsa pasar dalam ekonomi global
e. Pengembangan suite perangkat lunak penambangan data komersial yang siap pakai
f. Pertumbuhan luar biasa dalam daya komputasi dan kapasitas penyimpanan

Sebagai bagian dari proses yang ada di dalam KDD, maka data mining di dahului denganproses pemilihan data, pembersihan data, pre- processing, dan transformasi data. Ada tiga tahappenting dalam KDD adalah sebagai berikut [11] :

a. Data preprocessing, Proses ini bertujuan untuk mentransformasikan data input ke dalam format yang sesuai untuk kemudian dianalisa. Dalam tahap ini dilakukan proses penggabungan data dari berbagai sumber, pembersihan data untuk menghilangkan noise data dan data ganda, serta memilih atribut data yang diperlukan bagi proses data mining.
b. Data mining, Proses ini bertujuan untuk medapatkan pola-pola dan informasi yang tersembunyi di dalam basis data. Ada beberapa teknik yang dapat digunakan dalam data mining untuk mendapatkan pola-pola dan informasi tersembunyi, yaitu classification, neural network, decision tree, genetic algorithm, clustering, OLAP (Online Analitycal Processing), dan association rules.
c. Postprocessing, Proses ini bertujuan untuk memastikan hanya hasil yang valid dan berguna yang dapat digunakan oleh pihak yang berkepentingan. Contoh dari proses ini adalah proses visualisasi, yaitu proses untuk menganalisa dan mengeksplorasi datadan hasil dari proses data mining dari berbagai sudut pandang

2.4. 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 [12]. Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, misalnya, dapat dipartisi menjadi tiga 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, atau curang oleh orang lain
f. Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkankemungkinan ancaman teroris

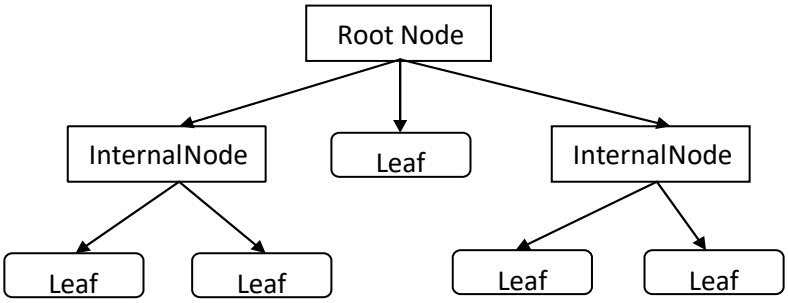
Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, SupportVector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan [12].

2.5. Decision Tree

Decision tree adalah flow-chart seperti struktur tree, dimana tiap internal node menunjukkan sebuah test pada sebuah atribut, tiap cabang menunjukkan hasil dari test, dan leaf node menunjukkan class-class atau class distribution. Selain karena pembangunannya relatif cepat, hasil dari model yang dibangun mudah untuk dipahami [13]. Pada decision tree terdapat 3 jenis node, yaitu:

a. Root Node, merupakan node paling atas, pada node ini tidak ada input dan bisa tidak mempunyai output atau mempunyai output lebih dari satu.

b. Internal Node, merupakan node percabangan, pada node ini hanya terdapat satu input dan mempunyai output minimal dua
c. Leaf node atau terminal node, merupakan node akhir, pada node ini hanya terdapat satu input dan tidak mempunyai output.

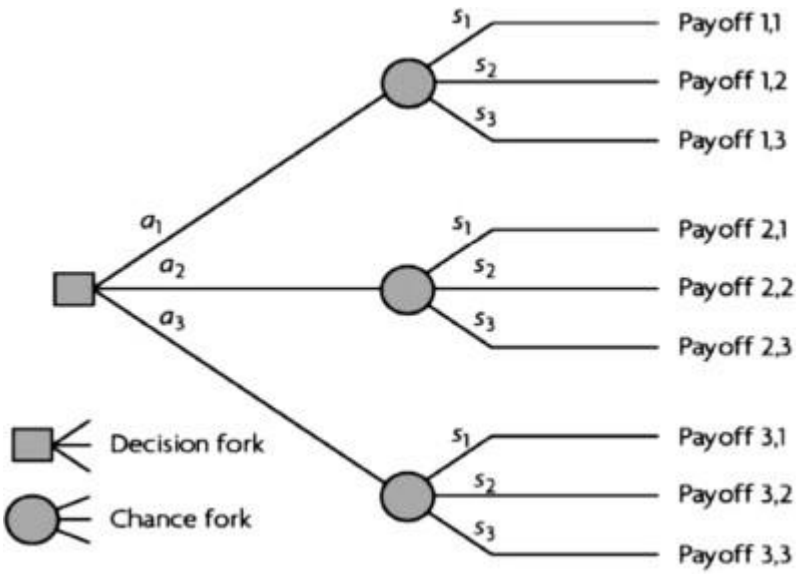


Gambar 2.1 Model Decision Tree

Pohon keputusan merupakan metode klasifikasi dan prediksi yang sangat kuat dan terkenal. Metode pohon keputusan mengubah fakta yang sangat besar menjadi pohon keputusan yang merepresentasikan aturan. Aturan dapat dengan mudah dipahami dengan bahasa alami. Selain itu dapat diekspresikan dalam bentuk bahasa basis data seperti Structure Query Language untuk mencari record pada kategori tertentu [9].

Decision Tree juga berguna untuk mengeksplorasi data, menemukan hubungan tersembunyi antara sejumlah calon variabel input dengan variabel target. Sebuah pohon keputusan adalah sebuah struktur yang dapat digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan record yang lebih kecil dengan menerapkan serangkaian aturan keputusan, dengan masing-masing rangkaian pembagian, anggota himpunan hasil menjadi mirip satu dengan yang lain.

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 [11].



Gambar 2.2 Bentuk Decision Tree Secara umum

Decision tree memiliki training sample berupa sekumpulan data yang nantinya akandigunakan 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. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang yang memiliki kelas yang sama. Rumus menghitung nilai entropy menggunakan persamaan (1) dan (2) :

Entropy (S) $\sum -pi\ log2\ pi\ n\ i = 1$ (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 :

Gaint (S,A) $\sum n\ f - 1\ \left|si\right|\ \left|s\right|\ Entropy\ (Si)$ (2)

2.6. 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 Validation 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 [12].

2.7. 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 [14]. 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 [14].

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 algoritma dan membuatnya lebih mudah dipahami.
d. Pelatihan Lebih Cepat: Ini memungkinkan algoritma pembelajaran mesin untuk berlatih lebih cepat.

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.

2.8. 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 [15].

$$v_{n+1} = v_n + c_1 rand() * (p_{best,n} - CurrentPosition_n) + c_2 rand2() * (g_{best,n} - CurrentPosition_n)$$

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 [8]. 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 (3).

   $(V_{i,j(t)=}0)$ (3)

   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 $[x_{min,max}]$. proses inisialisasi posisi terdapat pada Persamaan (4).

   $x(t)=x\min+r(x\max-x\min)$ (4)

   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 (5) dan Gbest merupakan Pbest dengan nilai fitness terbaik.

   $(Pbest_{i,j(t)}=x_{i,j(t)})$ (5)

   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 (6). 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 (6) dan (7).

$t+1$ t t t t t

$$\text{v } i,j = w.\, vi,j + c1r1(Pbesi,j - xi,j) + c2r2(Gbesg,j - xi,j) \quad (6)$$

$vj\ max = k(\underline{\qquad\qquad} xj,max - xj,\min)\ k \in (0,1]$

(17) 2

$$if\ v_{ij}(t+1) > vmaxj \quad then\ v_{ij}(t+1) = vj^{\,max} \quad (7)$$

max

$if\ v_{ij}(t+1) < -v \quad j \quad then\ v_{ij}(t+1)$

max

$= -vj$

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 (8) Kemudian hasil signoid yang telah didapat akan diproses lebih lanjut pada Persamaan (9) sehingga didapatkan posisi terbaru Setelah itu menentukan hasil fitness terbaru yang tentunya juga akan mendapat nilai Pbest terbaru.

t

$$sig(vi,j) = \frac{2}{+e^{1-v}}\, t\ ,\ j = 1,2,\ldots d \quad (8)$$

i,j

t $t+1$

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

t $t+1$

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

$j = 1,2,\ldots . d$ (9)

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 seperti persamaan (10).

$$k = 1 + decimal(s1\ )\ x^{a-1} \qquad (10)$$
$$2n1-1$$

2.9. Akurasi

Akurasi adalah salah satu metrik untuk mengevaluasi model klasifikasi. Secara informal, akurasi adalah sebagian kecil dari prediksi model kami yang benar. Secara formal, akurasi memiliki definisi sebagaimana ditunjukan pada Persamaan (11) dan (12) (Romadhon & Kurniawan, 2021):

$$\text{Akurasi} = \frac{\text{Number of Correect Prediction}}{\text{Total Number of Prediction}} \qquad (11)$$

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

$$\text{Akurasi} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (12)$$

Dimana

TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.10. Confusion Matrix

Confusion Matriks 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 [12].

2.11. 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) [12].