

BAB II TINJAUAN PUSTAKA

2.1 Penelitian Terkait

Beberapa penelitian terkait dengan menggunakan teknik Data Mining metode klasifikasi untuk memprediksi penyakit kanker payudara diantaranya:

Tabel 2.1 Review Jurnal

No	Judul, penulis, tahun	Dataset	Metode	Hasil
1.	Analisis Algoritma Klasifikasi Neural Network untuk diagnosis penyakit kanker payudara Fitra Septia Nugraha, Muhammad Ja'far Shidiq, Sri Rahayu, Jurnal PILAR Nusa Mandiri Vol. 15, September 2019 (Nugraha et al., 2019)	Terdapat 10 atribut yaitu age, BMI, glucose, insulin, homa, leptin, adiponectin, resistin, MCP.1 dan diagnosis Data pada penelitian ini merupakan data sekunder yang diperoleh dari Machine Learning Repository UCI (Universitas California,	menggunakan metode Neural Network dan tools Rapid Miner 9.0	Dengan menggunakan metode Neural Network dan tools Rapid Miner 9.0 bertujuan untuk memprediksi diagnosis kanker payudara dan kemudian menghasilkan nilai accuracy 71,83%, precision 81,08% dan recall 69,17% dengan AUC sebesar 0,806 yang artinya klasifikasi dinyatakan baik sehingga pasien

No	Judul, penulis, tahun	Dataset	Metode	Hasil
		Invene) dengan alamat web https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Coimbra		dengan parameter yang ada dapat diprediksi mana saja yang merupakan pasien kanker payudara dan mana yang bukan, sehingga pola ini dapat digunakan sebagai tolak ukur diagnosis sehingga dapat dideteksi lebih dini dan diharapkan dapat menekan angka kematian akibat kanker payudara
2.	Optimasi Fitur Menggunakan Backward Elimination Dan Algoritma SVM Untuk Klasifikasi Kanker Payudara Farizul Ma'arif, Toni Arifin JURNAL INFORMATIKA, Vol.4 No.1 April 2017 (Algoritma et al., 2017)	Tahap awal dilakukan identifikasi dataset, diketahui bahwa jumlah dataset WBC (Wisconsin Breast Cancer) sebanyak 699 record	Backward Elimination dan Support Vector Machine (SVM)	Berdasarkan hasil dan pembahasan yang telah dilakukan, dapat disimpulkan bahwa metode Support Vector Machine (SVM) adalah Algoritma yang baik diantara algoritma yang

No	Judul, penulis, tahun	Dataset	Metode	Hasil
		1. Clump Thickness, 2 Uniformity of Cell Size, 3 Uniformity of Cell Shape ,4 Marginal Adhesion, 5 Single Epithelial Cell Size, 6 Bare Nuclei, 7 Bland Chromatin, 8 Normal Nucleoli 9, Mitoses ,10 Class (Benign dan Malignant)		penulis telah uji untuk pengklasifikasian Kanker Payudara menggunakan Dataset WBC (Wisconsin Breast Cancer). Dimana nilai klasifikasi performansi Akurasi dan AUC nya adalah yang tertinggi, sedangkan untuk penggabungan algoritma seleksi fitur Backward Elimination dan Support Vector Machine (SVM) mendapatkan peningkatan Akurasi sebesar 14% sehingga nilai tingkat akurasi akhirnya sebesar 97.14% dan nilai AUC mencapai 0.995.

No	Judul, penulis, tahun	Dataset	Metode	Hasil
3.	Penerapan Algoritma C4.5 untuk Klasifikasi Tingkat Keganasan Kanker Payudara Dwi Ayu Nursela 2019 (Dwi Ayu Nursela (2010). Penerapan Algoritma C4.5 Untuk Klasifikasi Tingkat Keganasan Kanker Payudara, [s.d.]	Menggunakan Machine learning sebagai metode klasifikasi kanker payudara, Dataset berisi 699 sampel yang diambil dari needle aspirates dari jaringan kanker payudara manusia.	Decision Trees	Pada pemodelan algoritma C4.5 dengan menggunakan pembobotan atribut juga menghasilkan akurasi sebesar 98,57%. Akurasi tersebut diperoleh dari kesesuaian antara prediksi klasifikasi dan hasil klasifikasi
4	Predicting Breast Cancer Survivability Using Data Mining Techniques Abdelghani Bellaachia, Erhan Guven 2010 (Bellaachia & Guven, [s.d.]	The EOD field is composed of five fields including the EOD code. These fields (size of tumor, number of positive nodes, number of nodes, and number of primaries)	In this paper, we have investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and	The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other two techniques.

No	Judul, penulis, tahun	Dataset	Metode	Hasil
		<p>contain missing information coded such as '999', '99' or '9' representing the 'unknown' information</p> <p>The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database</p>	the C4.5 algorithms	C4.5 86,7 %
5.	<p>Prediction of benign and malignant breast cancer using data mining techniques</p> <p>Vikas Chaurasia , Saurabh Pal, and BB Tiwari</p>	<p>We used three popular data mining algorithms (Nai"ve Bayes, RBF Network, J48) to develop the prediction models using a</p>	Nai"ve Bayes, RBF Network, J48	<p>The results (based on average accuracy Breast Cancer dataset) indicated that the Nai"ve Bayes is the best predictor with 97.36% accuracy on the holdout sample (this</p>

No	Judul, penulis, tahun	Dataset	Metode	Hasil
	Journal of Algorithms & Computational Technology 2018, Vol. 12(2) 119–126 (Chaurasia et al., 2018)	large dataset (683 breast cancer cases). We also used 10-fold cross-validation methods to measure the unbiased estimate of the three prediction models for performance comparison purposes		prediction accuracy is better than any reported in the literature), RBF Network came out to be the second with 96.77% accuracy, J48 came out third with 93.41% accuracy
6.	Penerapan Algoritma Klasifikasi C4.5 Untuk Diagnosis Penyakit Kanker Payudara Laily Hermawanti, 2012 (Hermawanti, [s.d.])	Dataset penyakit kanker payudara yang digunakan adalah Wisconsin Breast Cancer (WBC) dari UCI Dataset Repository	algoritma C4.5	Setelah dilakukan evaluasi dengan algoritma C4.5, hasil akurasi tinggi. Nilai akurasi untuk algoritma klasifikasi C4.5 senilai 94.56% dan nilai AUC untuk algoritma C4.5 senilai 0.941

No	Judul, penulis, tahun	Dataset	Metode	Hasil
7.	Penerapan Optimasi Berbasis Particple Swarm Optimization (PSO) Algoritma Naïve Bayes Dan K-Nearst Neighbor Sebagai Perbandingan Untuk Mencari Kenerja Terbaik Dalam Mendeteksi Kanker Payudara Taghfirul Azhima Yoga Siswa , Prihandoko 2018 (Azhima & Siswa, 2018)	Dataset penyakit kanker payudara yang digunakan adalah Wisconsin Breast Cancer (WBC) dari UCI Dataset Repository	Naïve Bayes, K-Nearest Neighbor	Terbukti optimasi Particle Swarm Optimization (PSO) dapat meningkatkan kinerja pada akurasi Naïve Bayes sebesar 97.37% menjadi 97.81%, dan akurasi K-Nearest Neighbor dari 95.45% mengalami peningkatan menjadi 96.63%

Dari hasil review beberapa jurnal baik jurnal nasional dan internasional saya menyimpulkan agar data set dapat terakurasi dengan baik maka data yang dibutuhkan peneliti semakin banyak akan semakin baik akurasinya dengan jumlah data diatas 500 fitur , dan metode yang digunakan untuk beberapa jurnal yang saya review tingkat akurasi cenderung lebih tinggi dengan menggunakan Decision Tree dengan akurasi mencapai 98,57 %. Hasil dapat dilihat pada tabel berikut :

Tabel 2.2 Hasil Akurasi Review Jurnal

No Jurnal	Metode Terbaik dan akurasi
1	Neural Network 71,83%
2	Support Vector Machine (SVM) 97,14%
3	SVM 97.59%
4	Decision Tree 98,57 %
5	RBF Network 96.77%
6	algoritma C4.5 94,56%
7	Naïve Bayes 97.81%,

2.2 Kanker Payudara

Kanker payudara adalah keganasan pada sel-sel yang terdapat pada jaringan payudara, bisa berasal dari komponen kelenjarnya maupun komponen selain kelenjar seperti jaringan lemak, pembuluh darah, dan persarafan jaringan payudara. Kanker payudara adalah suatu penyakit dimana terjadi pertumbuhan berlebihan atau perkembangan tidak terkontrol dari sel sel (jaringan) payudara. Kanker payudara adalah kanker yang paling umum terjadi pada wanita (Rasjidi, 2010).

Disebut kanker payudara ketika sejumlah sel di dalam payudara tumbuh dan berkembang dengan tidak terkendali. Kanker payudara merupakan suatu kondisi dimana sel telah kehilangan pengendalian dan mekanisme normalnya, sehingga terjadi pertumbuhan yang tidak normal, cepat dan tidak terkendali yang terjadi pada jaringan payudara. Kanker payudara pada umumnya menyerang pada kaum wanita, tetapi tidak menutup kemungkinan juga dapat menyerang kaum laki-laki, walaupun kemungkinan menyerang kaum laki-laki itu sangat kecil sekali yaitu 1:1000. Kanker payudara adalah salah satu jenis kanker yang juga menjadi penyebab kematian terbesar kaum wanita di dunia, termasuk indonesia (Entan Teram Zettira et al., 2017).

Kanker payudara merupakan keganasan pada jaringan payudara yang dapat berasal dari epitel duktus maupun lobulusnya. Kanker payudara adalah tumor ganas yang terbentuk dari sel-sel payudara yang tumbuh dan berkembang tanpa terkendali sehingga dapat menyebar di antara jaringan atau organ di dekat payudara atau ke bagian tubuh lainnya (profil Kesehatan Indonesia Tahun 2020, [s.d.]). Kanker payudara merupakan kanker ganas pada payudara atau salah satu payudara. Kanker ini adalah suatu penyakit neoplasma ganas yang berasal dari parenchyma (bagian organ yang produktif). Kanker bisa mulai tumbuh di dalam kelenjar susu, saluran susu, jaringan lemak maupun jaringan ikat pada payudara. Kanker payudara disebabkan oleh adanya kerusakan pada materi genetik sel yang kemudian bersentuhan dengan bahan kimia yang mempercepat pembiakan sel yang diperlukan untuk berkembang menjadi sel kanker yang lebih ganas (Abdullah, Rozi. 2012. *Ca Mamae (Kanker Payudara)*, [s.d.]

2.3 Klasifikasi Kanker Payudara

Menurut *National Breast Cancer Foundation*, terdapat 2 tipe yaitu Jinak dan ganas payudara, yaitu:

1. Tumor Jinak Payudara (*Benign Breast Tumors*) Tumor jinak payudara adalah pertumbuhan sel yang tidak normal tetapi tidak menyebar keluar payudara dan tidak mengancam nyawa manusia. Namun, tumor jinak payudara dapat meningkat menjadi kanker sehingga tumor jinak payudara tetap patut diwaspadai. Ketika tumor didiagnosa sebagai tumor jinak, dokter selalu membiarkannya daripada mengangkatnya. Meskipun biasanya tumor ini tidak menyerang sekitar jaringan payudara, kadang tumor ini bisa tumbuh menekan organ dan menyebabkan luka atau masalah lain. Pada keadaan ini, tumor harus segera diangkat.
2. Tumor Ganas Payudara (*Malignant Breast Tumor/Breast Cancer*) Tumor ganas payudara adalah kumpulan sel kanker yang tumbuh dan dapat menyebar pada berbagai bagian tubuh. Tumor ganas berbahaya karena menyerang sekitar jaringan payudara. Ketika tumor dicurigai sebagai tumor ganas, maka dokter akan melakukan biopsi untuk mendiagnosa tumor (Nugraha et al. 2019).

2.4 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 (Editioi Efraim Turban Jay Aronson & Liang, 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 didefinisikan 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.

(Kusrini dan Taufiq., 2009)

Data mining memiliki beberapa teknik yang terkenal dan sering digunakan oleh peneliti, diantaranya seperti clustering, classification, association, dan beberapa perkembangan teknik sesuai dengan perubahan kecenderungan 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 dikelompokkan menjadi beberapa kelompok yaitu :

1. Deskripsi Menggambarkan pola yang terdapat dalam data yang memberikan kemungkinan penjelasan untuk suatu pola atau kecenderungan.
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.
2. Prediksi Prediksi nilai dari hasil akan ada di masa mendatang .
3. Klasifikasi Dalam klasifikasi, terdapat target variabel kategori.
Pengkusteran Pengkusteran merupakan pengelompokan record, pengamatan, atau memperhatikan dan membentuk kelas objek-objek yang memiliki kemiripan (Kusrini dan Taufiq., 2009).

2.5 Klasifikasi

Klasifikasi adalah proses untuk menemukan model atau fungsi yang menjelaskan atau membedakan konsep atau kelas data, dengan tujuan untuk memperkirakan kelas dari suatu objek yang labelnya tidak diketahui. Model itu sendiri dapat berupa aturan jika-maka (if-then), berupa pohon keputusan (decision tree), jaringan saraf tiruan (Sistem Informasi STMIK Nusa Mandiri Jakarta Jl Damai No et al., 2016b).

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 (Misdrum et al., 2021). Dalam klasifikasi, ada variabel kategoris target, seperti braket pendapatan, yang, 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 tentang variabel target serta satu set input atau prediktor variable. Contoh tugas klasifikasi dalam bisnis dan penelitian meliputi :

(Wibawa et al., 2018):

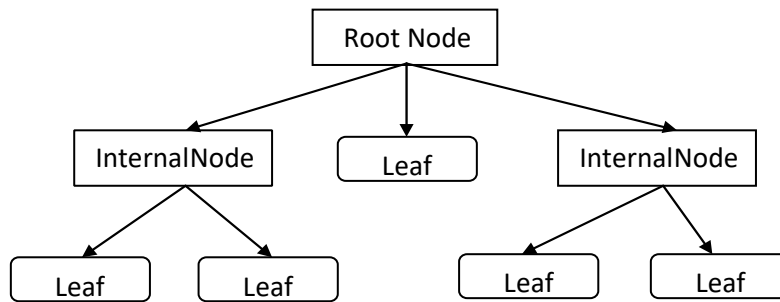
- a. Menentukan apakah transaksi kartu kredit tertentu adalah penipuan
- b. Menempatkan siswa baru pada jalur tertentu yang berkaitan dengan kebutuhan khusus
- c. Menilai apakah aplikasi hipotek adalah risiko kredit yang baik atau buruk
- d. Mendiagnosis apakah ada penyakit tertentu
- e. Menentukan apakah surat wasiat ditulis oleh almarhum yang sebenarnya, atau curang oleh orang lain
- f. Mengidentifikasi apakah perilaku keuangan atau pribadi tertentu menunjukkan kemungkinan ancaman teroris

Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari algoritma cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa algoritma, diantaranya *Naïve Bayes*, *Support Vector Machine*, *Decision Tree*, *Fuzzy* dan Jaringan Saraf Tiruan (Wibawa et al., 2018)

2.6. 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. 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.

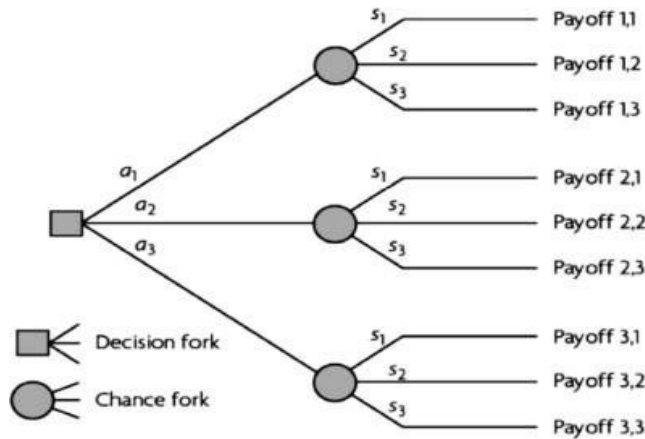


Gambar 2.1 Model Decision Tree

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 (Kusrini dan Emha, 2009).

Pohon keputusan 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 (Aripin, [s.d.]



Gambar 2.2 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. Ulangi proses untuk masing-masing cabang sampai semua kasus pada cabang yang memiliki kelas yang sama. Rumus menghitung nilai entropy menggunakan persamaan 1 dan persamaan 2 seperti rumus berikut :

$$\text{Entropy (S)} \sum_{i=1}^n - p_i \log_2 p_i \quad (1)$$

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 :

$$\text{Gain (S,A)} \sum_{i=1}^n f - 1 \left| \frac{s_i}{s} \right| \text{Entropy (Si)} \quad (2)$$

2.7 Rapid Miner

Rapid Miner adalah perangkat lunak yang dibuat oleh Dr. Markus Hofmann dari Institute of Technology Blanchardstown dan Raif Klinkenberg dari rapid-i.com dengan tampilan GUI (Graphical User Interface) sehingga memudahkan pengguna dalam menggunakan perangkat lunak ini. Perangkat lunak ini bersifat open source dan dibuat dengan menggunakan bahasa java dibawah lisensi GNU Public License dan Rapid Miner dapat dijalankan disistem operasi manapun. Dengan menggunakan Rapid Miner, tidak dibutuhkan kemampuan koding khusus, karena semua fasilitas sudah disediakan. Rapid Miner dikhususkan untuk penggunaan data mining. (Sistem Informasi STMIK Nusa Mandiri Jakarta Jl Damai No et al., 2016b)

2.8 Split Validation

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 (Witten et al., [s.d.]

2.9 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 *classifier*, dan menghapus fitur yang tidak relevan (Syahid Pebriadi & Saubari, 2019). 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 (Prasetio, 2020).

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.
- 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 (Prasetio, 2020)

2.9.1. Particle Swarm Optimization

Particle swarm optimization, disingkat sebagai PSO, didasarkan pada perilaku sebuah kawanan serangga, seperti semut, rayap, lebah atau burung. Algoritma PSO meniru perilaku sosial organisme ini. Perilaku sosial terdiri dari tindakan individu dan pengaruh dari individu-individu lain dalam suatu kelompok. Kata “partikel” menunjukkan individu, misalnya seekor burung dalam kawanan burung. Setiap individu atau partikel berperilaku saling terhubung dengan cara menggunakan kecerdasannya (*intelligence*) sendiri dan juga dipengaruhi perilaku kelompok kolektifnya. Dengan demikian, jika satu partikel atau seekor burung menemukan jalan yang tepat atau pendek menuju ke sumber makanan, sisa kelompok yang lain juga akan dapat segera mengikuti jalan tersebut meskipun lokasi mereka jauh di kelompok tersebut.

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 partikel 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. (Habiburrahman et al., 2017)

$$v_{n+1} = v_n + c_1 \text{rand}() * (p_{best,n} - \text{CurrentPosition}_n) + c_2 \text{rand}2() * (g_{best,n} - \text{CurrentPosition}_n)$$

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 (Mustopa, 2021). 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 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 (Habiburrahman et al., 2017).

Secara garis besar prosedur PSO dapat dilakukan dalam beberapa langkah.

1. Inisialisasi kecepatan awal bernilai 0 untuk semua partikel seperti pada Persamaan 3.

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

$V_{i,j}$ merupakan kecepatan, j adalah letak partikel dan i adalah letak individu dan t adalah iterasi.

2. Inisialisasi posisi awal partikel dengan batasan sesuai range $[x_{\min}, x_{\max}]$. proses inisialisasi posisi terdapat pada Persamaan 4.

$$x(t) = x_{\min} + r(x_{\max} - x_{\min}) \quad (4)$$

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 dan Gbest merupakan Pbest dengan nilai fitness terbaik .

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

Pbest merupakan personal best pada individu ke-i dan partikel ke-j. x_{ij} 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 6 dan 7.

$$v_{i,j}^{t+1} = w \cdot v_{i,j}^t + c_1 r_1 (Pbest_{i,j}^t - x_{i,j}^t) + c_2 r_2 (Gbest_{j}^t - x_{i,j}^t) \quad (6)$$

$$v_{j \max} = k \left(\frac{x_{j,\max} - x_{j,\min}}{x_{j,\max} - x_{j,\min}} \right) \quad k \in (0,1]$$

$$\text{if } v_{ij}(t+1) > v_{maxj} \text{ then } v_{ij}(t+1) = v_{j \max} \quad (7)$$

max

$$\text{if } v_{ij}(t+1) < -v_{j \max} \text{ then } v_{ij}(t+1) = -v_{j \max}$$

max

$$= -v_{j \max}$$

Nilai c_1 dan c_2 adalah koefisien akselerasi, nilai r_1 dan r_2 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 dilanjutkan 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.

$$v_{i,j}^{t+1} = w \cdot v_{i,j}^t + e^{-\alpha t} \cdot r \cdot (p_{best} - x_{i,j}^t) + e^{-\alpha t} \cdot r \cdot (g_{best} - x_{i,j}^t) \quad (8)$$

$$x_{i,j}^{t+1} = \begin{cases} x_{i,j}^t & \text{if } rand[0,1] > sig(v_{i,j}^t) \\ v_{i,j}^{t+1} & \text{if } rand[0,1] < sig(v_{i,j}^t) \end{cases} \quad (9)$$

- 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.

$$k = 1 + decimal(s1) \cdot x^{\alpha-1} \quad (10)$$

Algoritma PSO (*Particle Swarm Optimization*) adalah salah satu algoritma optimasi yang dapat digunakan untuk pengambilan keputusan. Tetapi bisa juga digunakan untuk pencarian jalur. Contoh yang dibahas kali ini adalah mengenai pencarian posisi dengan pengembalian nilai fungsi minimal.

Particle Swarm Optimization adalah teknik optimasi dengan cara menghitung secara terus menerus calon solusi dengan menggunakan suatu acuan kualitas. Algoritma ini mengoptimasi permasalahan dengan cara menggerakkan partikel atau calon solusi di dalam ruang permasalahan menggunakan fungsi tertentu untuk posisi dan kecepatan dari partikel. Pergerakan partikel dipengaruhi oleh solusi terbaik partikel tersebut, dan solusi terbaik secara umum yang didapatkan dari partikel lain. Sekumpulan partikel ini dinamakan swarm, dan pada akhirnya swarm ini akan bergerak menuju kepada solusi terbaik (Habiburrahman et al. 2017).

2.10 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 sebagai berikut (Romadhon & Kurniawan, 2021):

$$\text{Akurasi} = \frac{\text{Number of Correct Prediction}}{\text{Total Number of Prediction}} \quad (11)$$

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

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

Dimana TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

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 (Romadhon & Kurniawan, 2021)

2.12 AUC (*Area Under The Curve*)

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) (Sistem Informasi STMIK Nusa Mandiri Jakarta Jl Damai No et al., 2016a) .

Kurva ROC (*Receiver Operating Characteristics*) banyak digunakan untuk menilai hasil prediksi, kurva ROC adalah teknik untuk memvisualisasikan, mengatur, dan memilih pengklasifikasian berdasarkan kinerja mereka (Gorunescu, 2011). Kurva ROC (*Receiver Operating Characteristics*) adalah tool dua dimensi yang digunakan untuk menilai kinerja klasifikasi yang menggunakan dua class keputusan, masingmasing objek dipetakan ke salah satu elemen dari himpunan pasangan, positif atau negatif. Pada kurva ROC, TP rate diplot pada sumbu Y dan FP rate diplot pada sumbu X. Untuk klasifikasi data mining, nilai AUC (*Area Under The Curve*) dapat dibagi menjadi beberapa kelompok (Gorunescu, 2011).

- 1) 0.90-1.00 = Excellent Classification
- 2) 0.80-0.90 = Good Classification
- 3) 0.70-0.80 = Fair Classification
- 4) 0.60-0.70 = Poor Classification
- 5) 0.50-0.60=Failur (Sistem Informasi STMIK Nusa Mandiri Jakarta Jl Damai No et al., 2016a).