BAB II
LANDASAN TEORI

2.1. Penelitian Terkait

Sebelum melakukan penelitian ini, terlebih dahulu peneliti mengumpulkan dan mengkaji hasil-hasil penelitian terkait yang pernah dilakukan oleh peneliti sebelumnya. Hal tersebut penting untuk mengetahui *state of the art* pada bidang yang akan diteliti. Prediksi nilai akurasi dengan teknik data mining mengguanakan Algoritma Particle Swarm Optimization Dan Juga Algoritma Decision Tree C4.5 telah banyak dilakukan oleh peneliti-peneliti sebelumnya, beberapa penelitian terkait yang menjadi *key paper* pada penelitian ini. Berikut adalah penelitian terkait yang menjadi dasar dalam penelitian ini.

Tabel 2. 1 Review Penelitian Terdahulu

No	Judul Penelitian	Tahun	Author	Metode	Kesimpulan	Saran Penelitian
1	Meningkatkan Kinerja Decision Tree C4.5 Dengan Seleksi Fitur Korelasi Pearson Pada Deteksi Penyakit Diabetes	2022	Mohammad Burhan Hanif, Galet Guntoro Setiaji [10]	Decision Tree C4.5 dan Korelasi Pearson	Dalam penelitian ini dapat terlihat keberhasilan peningkatan nilai akurasi dan nilai AUC dari algoritma C4.5 melalui penerapan metode seleksi fitur korelasi pearson. Peningkatan nilai besaran akurasi dari 95.31% menjadi 96.16% dengan nilai AUC sebesar 0.936 menjadi 0.949	Dalam penelitian ini penggunaan algoritma lain sangat diperlukan untuk penelitian selanjutnya. Pencarian fitur-fitur penting dalam penelitian ini memainkan peran penting. Sehingga optimasi berfokus pada optimalisasi seleksi fitur dari data untuk model algoritma C4.5 dan model lain sangat diperlukan.
2	Penerapan Machine Learning	2021	Agus Surip , Muhama	Decision Tree, PSO	Dalam penelitian ini Algoritma decision tree yang	saran dari peneliti yang dijadikan bahan

	menggunakan algoritma C4.5 berbasis PSO dalam Menganalisa Data Siswa Putus Sekolah		d Aji Pratama , Irfan Ali , Arif Rinaldi Dikanan da , Ade Irma Purnama sari [8]		dioptimasi dengan PSO memiliki akurasi lebih baik dibandingkan dengan decision tree yang tanpa dioptimasi. Hal ini didasarkan pada hasil dari decision tree memiliki class precission prediksi tidak sebesar 87.74 % dan prediksi ya sebesar 98.36 %. Sedangkan hasil dari decision tree yang dioptimasi mengunakan PSO memiliki class precission prediksi tidak sebesar 91.74 % dan prediksi ya class precission sebesar 96.38 %. Dari hasil klasifikasi diketahui nilai akurasi dari model decision tree adalah 90.86 % artinya keakuratan dalam klasifikasi tersebut sudah cukup baik. Sedangkan Dari hasil klasifikasi model decision tree yang dioptimasi mengunakan PSO diketahui nilai akurasi adalah 92.95 %. Artinya keakuratan dalam klasifikasi tersebut menjadi lebih baik dari decision tree yang tanpa optimasi PSO.	renungan penelitian yang akan datang, agar hasil penelitian ini lebih baik lagi, dapat menggunakan jumlah data yang lebih banyak lagi dan melakaukan pengujian dengan dataset yang lebih lengkap atributnya sehingga model yang sudah dihasilkan akan lebih teruji lagi kedepanya; Menggunakan algoritma lain dalam klasifikasi data misalnya ANN atau Naïve Bayes; Untuk mendapatkan nilai akurasi yang lebih baik lagi dari penelitian ini, dapat melakukan optimasi dengan teknik optimasi lainnya
3	Klasifikasi Penyakit Diabetes Mellitus	2022	Ronna Putri Fadhillah , Raisya	C4.5 Fitur Heatmap	Berdasarkan penelitian yang telah dilakukan menegnai faktor-	Saran dalam penelitian ini yang dapat dilakukan dalam

	Berdasarkan Faktor-Faktor Penyebab Diabetes Menggunakan Algoritma C4.5		Rahma, Arni Sepharni, Ratna Mufidah, Betha Nurina Sari, Agung Pangestu [11]		faktor penyakit terjadinya diabetes melitus menggunakan algoritma C4.5 dan seleksi Fitur Heatmap yang dilakukan dengan melihat nilai korelasi fitur tertinggi terhadap variabel outcome sebagai target value. Pembagian data pada penelitian ini menggunakan grid search cross validation, di mana pada proses tersebut data dibagi dengan cross validation 10 yang menghasilkan akurasi sebesar 76%	penelitian selanjutnya adalah penggunaan algoritma lain untuk peningkatan akurasi penelitian dan juga tidak disebutkan dalam penlitian ini akurasi hasil dari klasifikasi dengan c4.5 saja sehingga tidak diketahui apakah terjadi peningkatan akurasi yang signifikan atau tidak.
4	Komparasi Algoritma C4.5 Berbasis Pso Dan Ga Untuk Diagnosa Penyakit Stroke	2020	Ramdhan Saepul Rohman, Rizal Amegia Saputra, Dasya Arif Firmansa [7]	Algoritma C4.5, *Particle Swarm Optimization*, *Genetic Algorithm*	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%	Dalam penelitian ini perbandingan algoritma murni C4.5 dan di tambahkan metode PSO dan GA menghasilkan nilai yang cukup baik namun dapat mencoba menambahkan algoritma lain untuk mendapatkan hasil yang lebih baik.

5	KOMPARASI PENERAPAN METODE BAGGING DAN ADABOOST PADA ALGORITMA C4.5 UNTUK PREDIKSI PENYAKIT STROKE	2020	Nur Diana Saputri, Khalid Khalid, Dwi Rolliawati [12]	Metode bagging Dan adaboost pada algoritma C4.5	Hasil pengujian klasifikasi menggunakan *Confusion Matrix* dan *k-fold cross validation* untuk algoritma C4.5 menghasilkan akurasi sebesar 92.87%. Kemudian hasil akurasi dari algoritma C4.5 dengan metode bagging meningkat menjadi 95.02% dan ketika dikombinasikan dengan metode Adaboost nilai akurasinya juga meningkat menjadi 94.63%. Dari hasil akurasi tersebut dapat disimpulkan bahwa penggabungan algoritma pengklasifikasi tunggal yaitu algoritma C4.5 dengan metode bagging dan Adaboost terbukti dapat meningkatkan performa klasifikasi.	dalam penelitian ini kedua metode tersebut masih kurang signifikan dalam meningkatkan nilai akurasi dari algoritma C4.5. Selain itu, metode bagging dan Adaboost juga tidak dapat meningkatkan nilai sensivisitas dan *f1-score* dari algortima C4.5, karena nilai kelas *false negative* yang dihasilkan masih tinggi. Sehingga dapat dicoba dengan menambahkan seleksi fitur lain untuk meningkatkan nilai sesitivitasnya.
6	A Review On Optimization Techniques Using Data Mining	2019	Anjali Agrawal, Mahesh Parmar [13]	Optimization techniques, genetic algorithm, particle swarm optimization, Ant colony optimization, support vector machine	Dalam hasil penelitian terdapat tabel perbandingan untuk membandingkan keuntungan dan keterbatasan teknik pengoptimalan dan algoritma pada masing – masing algoritma yang dibahas.	Dalam jurnal ini hanya terdapat perbandingan antara satu algoritma dan algoritma lain baik dari kekurangan tiap-tiap algoritma dan juga kelebihan pada masing - masing algoritma.
7	MonkeyPox2022Tweets: The First Public	2022	Nirmala Thakur [14]	Big Data; Data Mining;	karya ini menghadirkan Monkey-	Pada penelitian ini hanyalah crawling untuk pengumpulan

	Twitter Dataset on the 2022 MonkeyPox Outbreak			Data Science	Pox2022Tweets, kumpulan data dengan akses terbuka berisi lebih dari 255.000 Tweet terkait wabah cacar monyet tahun 2022 yang diposting di Twitter sejak kasus pertama wabah ini terdeteksi pada 7 Mei 2022. Kumpulan data tersebut sesuai dengan kebijakan privasi, perjanjian pengembang, dan panduan untuk redistribusi konten Twitter, serta dengan prinsip FAIR (Findability, Accessibility, Interoperability, dan Reusability) untuk manajemen data ilmiah.	dataset tentang wabah Monkeypox
8	A Forecasting Prognosis of the Monkeypox Outbreak Based on a Comprehensive Statistical and Regression Analysis	2022	Farhana Yasmin , and Sami Azam, Md. Mehedi Hassan, Sadika Zaman, Si Thu Aung , Asif Karim [9]	machine learning forecasting prediction	penelitian ini menunjukkan sembilan metode peramalan yang berbeda dan menemukan bahwa pemodelan learning adalah model forecasting yang paling dapat diandalkan dengan membandingkannya dengan kinerja delapan model lainnya.	Studi komprehensif tentang penyakit cacar monyet masih sulit dilakukan karena kurangnya data yang tersedia serta ketidaklengkapan dataset. Saran Kedepannya untuk melanjutkan penelitian ini dengan mengumpulkan dataset gambar yang terkait untuk mendeteksi penyakit monkeypox menggunakan pendekatan deep learning.

Dalam hasil review beberapa jurnal diatas baik jurnal nasional dan internasional Dapat dilihat bahwa hasil perhitungan pada masing – masing algoritma dan metode yang digunakan berbeda sehingga mempengaruhi hasil prediksi yang diperoleh. Dari hasil review pada jurnal dalam penelitian terdahulu penggunaan algoritma *Decision Tree* C4.5 memiliki hasil akurasi yang sangat baik terutama saat di gabungkan dengan mengguanakan algoritma *Particle Swarm Optimization* (PSO). Dapat terlihat nilai akurasinya dengan menggunakan *Decision Tree* C4.5 dan *Particle Swarm Optimization* (PSO) dapat meningkat menjadi 96.95 %. Seperti dalam penelitian [8] tentang, "Penerapan Machine Learning menggunakan algoritma C4.5 berbasis PSO dalam Menganalisa Data Siswa Putus Sekolah". Berikut adalah tabel komparasi dari jurnal pada penelitian tersebut setelah di review.

Tabel 2. 2 Tabel komparasi jurnal peneltian sebelumnya

NO	JURNAL	METODE	AKURASI
1	Meningkatkan Kinerja Decision Tree C4.5 Dengan Seleksi Fitur Korelasi Pearson Pada Deteksi Penyakit Diabetes	Decision Tree C4.5 dan Korelasi Pearson	96.16%
2	Penerapan Machine Learning menggunakan algoritma C4.5 berbasis PSO dalam Menganalisa Data Siswa Putus Sekolah	Decision Tree dan PSO	96.95 %.
3	Klasifikasi Penyakit Diabetes Mellitus Berdasarkan Faktor-Faktor Penyebab Diabetes Menggunakan Algoritma C4.5	C4.5 dan Fitur Heatmap	76 %
4	Komparasi Algoritma C4.5 Berbasis Pso Dan Ga Untuk Diagnosa Penyakit Stroke	C4.5 dan *Particle Swarm Optimization* C4.5 dan *Genetic Algorithm*	95.02% 94.63%
5	Komparasi Penerapan Metode Bagging Dan Adaboost Pada Algoritma C4.5 Untuk Prediksi Penyakit Stroke	C4.5 dengan *bagging* C4.5 dengan *adaboost*	95.02% 94.63%.

Dapat terlihat bahwa dalam penelitian sebelumnya pengguanakan algoritma Decision Tree C4.5 dan PSO memiliki hasil akurasi yang relatif sangat baik. Kemudian seteleh mengkomparasi jurnal pada penelitian sebelumnya peneliti juga

mengkomparasi tiap – tiap algoritma yang menjadi refrensi dalam melakukan penelitian ini dengan mengguanakan tool Rapidminer dan juga satu dataset publik yang didapat melalui situs publik yaitu kaggle.com untuk mengkomparasi tiap tiap algoritma yang menjadi refrensi dalam menentukan algoritma yang akan dipakai oleh penelitian ini. Berikut adalah hasil komparasi dengan Tool Rapidmner tersebut.

Tabel 2. 3 Komparasi dengan Rapidminer

NO	Pengujian Algoritma	Gambar	Akurasi
1	Dengan C4.5	accuracy: 94.91% +/- 0.33% (micro average: 94.91%) true 1 | true 0 | class precision pred. 1 | 8 | 19 | 29.63% pred. 0 | 241 | 4842 | 95.26% class recall | 3.21% | 99.61%	94.15 %
2	C4.5 dan PSO	accuracy: 96.13% +/- 0.06% (micro average: 96.13%) true 1 | true 0 | class precision pred. 1 | 0 | 0 | 0.00% pred. 0 | 249 | 4861 | 95.13% class recall | 0.00% | 100.00%	96.13 %
3	C4.5 dan Adaboost	accuracy: 94.91% +/- 0.33% (micro average: 94.91%) true 1 | true 0 | class precision pred. 1 | 8 | 19 | 29.63% pred. 0 | 241 | 4842 | 95.26% class recall | 3.21% | 99.61%	94.91 %
4	C4.5 dan bagging	accuracy: 95.17% +/- 0.13% (micro average: 95.17%) true 1 | true 0 | class precision pred. 1 | 4 | 2 | 66.67% pred. 0 | 245 | 4859 | 95.20% class recall | 1.61% | 99.96%	95.17 %

2.2. Landasan Teori

2.2.1. Wabah Monkeypox

Cacar adalah salah satu penyakit menular yang harus ditangani dengan serius. Wabah cacar telah terjadi dari masa ke masa, namun saat ini telah diberantas melalui program vaksinasi yang diadakan di seluruh dunia. Kasus cacar terakhir di dunia terjadi pada tahun 1977 di Somalia. Setelah itu, penyakit cacar menjadi mulai berkurang sehingga vaksinasi rutin terhadap penyakit cacar di kalangan masyarakat mulai dihentikan karena dianggap sudah tidak diperlukan pencegahan lagi terhadap penyakit cacar. Cacar monyet (monkeypox) merupakan penyakit infeksi virus yang

disebabkan oleh virus dengan genus orthopoxvirus. Virus cacar monyet ditemukan pada tahun 1958 saat dilakukan isolasi dari lesi vesikuloid pustular di antara monyet tawanan di Kopenhagen. Penyakit cacar monyet sebagian besar terjadi di hutan hujan Afrika bagian tengah dan barat. Orang- orang yang tinggal di sekitar kawasan berhutan mungkin memiliki resiko terpapar yang dapat menyebabkan infeksi subklinis. Namun baru- baru ini, muncul penyakit cacar monyet di Amerika Serikat pada hewan pengerat liar yang diimpor dari Afrika [3].

Virus cacar monyet telah ditemukan pada lesi kulit dan sebagian besar atau semua sekresi dan ekskresi (misalnya, urin, feses, dan eksudat oral, hidung, dan konjungtiva) pada hewan. Rute penularan yang mungkin termasuk inhalasi, inokulasi langsung ke luka di kulit, dan menelan jaringan yang terinfeksi. Anjing padang rumput yang terinfeksi secara eksperimental dapat menyebarkan virus cacar monyet hingga 21 hari setelah inokulasi, dan terdapat bukti yang terbatas menunjukkan bahwa beberapa hewan kecil, seperti tikus dormice dan tikus berkantung raksasa Gambia, mungkin membawa virus ini selama beberapa minggu atau bulan. Manusia dapat terinfeksi melalui gigitan hewan, aerosol selama kontak dekat, atau kontak langsung dengan lesi, darah, atau cairan tubuh. Penularan seksual dicurigai dalam beberapa kasus, ketika ada lesi pada alat kelamin, dan transmisi transplasenta juga telah tercatat. Di Afrika, kasus klinis sering dikaitkan dengan penanganan, penyiapan, dan makan hewan liar. Di A.S., sebagian besar kasus terjadi di antara orang-orang yang memiliki kontak langsung dekat dengan anjing padang rumput; beberapa infeksi tampaknya diperoleh melalui goresan dan gigitan, atau melalui luka terbuka. Virus cacar monyet telah diisolasi dari manusia hingga 18 hari setelah timbulnya ruam, dan keropeng yang pecah selama pemulihan ditemukan mengandung sejumlah besar virus menular. Penularan dari orang ke orang tampaknya tidak mampu mempertahankan virus dalam populasi manusia.[2]

a. Tanda – Tanda dan Gejala Cacar Monyet

Dalam artikel yang diterbitkan oleh [4] Orang yang terinfeksi virus monkeypox akan mulai menunjukan gejala pertmanya setelah 6-16 hari setelah paparan. Masa inkubasi virus cacar monyet bisa berkisar antara 6-13 hari, namun bisa juga terjadi dalam rentang yang lebih panjang, yakni 5-21 hari. Selama tidak memnuculkan

gejala, seseorang tetap bisa menularkan virus cacar monyet kepada orang lain. Dilansir dari WHO, kemunculan gejala cacar monyet terbagi dalam dua periode infeksi, yaitu periode invasi dan periode erupsi kulit, sebagaimana dijelaskan berikut ini:

1. Periode invasi

Periode invasi terjadi dalam 0-5 hari setelah terinfeksi virusnya pertama kali. Saat seseorang berada dalam masa invasi, dirinya akan menunjukan beberapa gejala cacar monyet seperti demam, sakit kepala, limfadenopati (pembengkakan kelenjar getah bening), sakit punggung, nyeri otot, lemas parah (asthenia). Pembengkakan kelenjar getah bening itulah yang menjadi ciri pembeda anara cacar monyet dengan jenis cacar lainya.

2. Periode Erupso Kulit

Periode ini terjadi pada 1-3 hari setelah demam muncul. Gejala utama dalam fase ini adalah munculnya ruam kulit. Ruam kulit pertama kali muncul di wajah dan kemudian menyebar ke seluruh tubuh. Wajah dan telapak tangan serta kaku adalah area yang paling terdampak ruam ini. Kemunculan ruam juga bisa ditemukan pada membran mukosa yang terletak di tenggorokan, area alat, termasuk jaringan mata dan kornea. Ruam terbentuk biasanya diawalai dengan bintik bintik hingga berubah menjadi vesikel atau lenting, yaitu lepuhan kulit yang berisi cairan. Dalam waktu beberapa hari, ruam akan berubah mengering membentuk kerak (keropong) di kulit.

2.2.2. Data Mining

Data mining adalah proses menemukan pola dan tren yang berguna dalam kumpulan data yang besar. Data mining juga seharusnya lebih tepat disebut *knowledge mining from data*. Namun dalam jangka pendek, penambangan pengetahuan mungkin tidak mencerminkan penekanan pada penambangan dari sejumlah besar data. Selain itu, banyak istilah lain yang memiliki arti serupa dengan penambangan data, seperti: *knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology*, and *data dredging* [15]. Data mining dibagi menjadi beberapa kelompok sesuai dengan tugas yang dapat dilakukan, yaitu:

a. Description (Deskripsi)

Model penambangan data harus setransparan mungkin. Artinya, hasil model data mining harus menggambarkan pola yang jelas yang terbuka untuk interpretasi dan interpretasi intuitif. Beberapa metode data mining lebih cocok untuk interpretasi transparan daripada yang lain. Misalnya, pohon keputusan memberikan interpretasi hasil yang intuitif dan ramah manusia. Deskripsi berkualitas tinggi seringkali dapat dicapai melalui analisis data eksplorasi, metode grafis untuk mengeksplorasi data untuk pola dan tren.

b. Estimation (Estimasi)

Dalam estimasi, satu set prediktor numerik dan/atau kategoris digunakan untuk memperkirakan nilai variabel target numerik. Model dibangun menggunakan catatan "penuh", yang memberikan nilai variabel target serta variabel prediktor. Kemudian, untuk pengamatan baru, nilai variabel target diestimasi berdasarkan nilai prediksi.

c. Prediction (Prediksi)

Prediksi mirip dengan klasifikasi dan estimasi, hanya saja hasilnya ada di masa depan. Jika sesuai, setiap metode dan teknik yang digunakan untuk klasifikasi dan estimasi juga dapat digunakan untuk prediksi. Ini termasuk estimasi titik dan estimasi interval kepercayaan, regresi linier sederhana dan korelasi, dan metode statistik tradisional untuk regresi berganda.

d. Classification (Klasifikasi)

Klasifikasi adalah bentuk analisis data yang mengekstrak model yang menggambarkan kategori / kelas data penting dan memiliki variabel target. Jenis model ini disebut klasifikasi, yang dapat memprediksi label kelas kategorikal (diskrit, tidak berurutan). Klasifikasi memiliki banyak aplikasi, termasuk deteksi penipuan, pemasaran bertarget, prediksi kinerja, manufaktur, dan diagnosis medis. Salah satu algoritma untuk klasifikasi adalah menggunakan algoritma Decision

Tree C4.5

e. Clustering (Pengelompokan)

Clustering mengacu pada pengelompokan record, observasi, atau kasus ke dalam kelas objek yang serupa. Cluster adalah kumpulan record yang mirip satu sama lain dan tidak mirip dengan record di cluster lain. Clustering berbeda dari klasifikasi karena tidak ada variabel target untuk clustering. Tugas pengelompokan tidak berusaha untuk mengklasifikasikan, memperkirakan, atau memprediksi nilai variabel target.

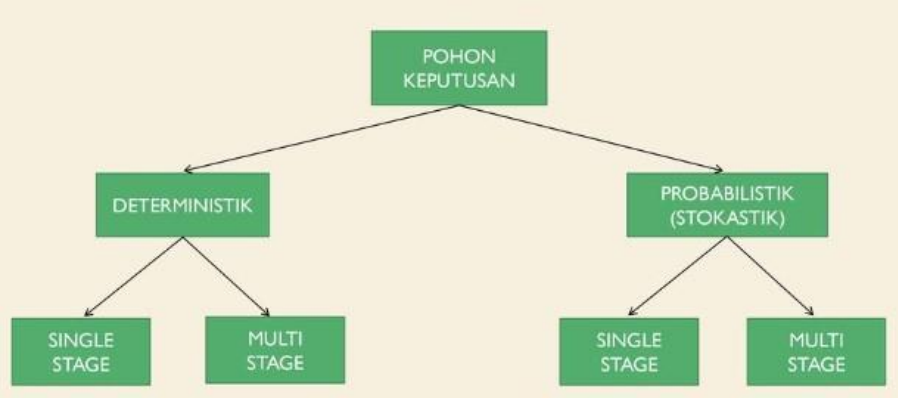
f. Association (Asosiasi)

Tugas asosiatif data mining adalah tugas mencari tahu atribut mana yang “bersama”. Paling umum di dunia bisnis, di mana disebut analisis afinitas atau analisis keranjang belanja, tugas asosiasi bertujuan untuk mengungkap aturan yang mengukur hubungan antara dua atau lebih atribut. Aturan pengaitan berbentuk "ikuti jika pertama", bersama dengan metrik dukungan dan kepercayaan yang terkait dengan aturan.

2.2.3. Algoritma Decision Tree C4.5

Decision Tree adalah struktur flowchart berbentuk pohon, dimana simpul bagian dalam merupakan suatu tes pada atribut kemudian setiap cabang menampilkan hasil tes dan simpul daun menampilkan kelas atau penyebaran kelas. Salah satu algoritma yang dapat digunakan untuk membuat pohon keputusan adalah algoritma C4.5. Algoritma C4.5 merupakan algoritma yang sangat populer digunakan oleh banyak peneliti di dunia. Algoritma C4.5 sebagai versi perbaikan ID3 merupakan sebuah algoritma yang diperkenalkan oleh Quinlan. Akan tetapi kelemahan hanya atribut bertipe kategorikal (nominal atau ordinal) saja yang bias di induksi oleh decision tree, sedangkan untuk menangani atribut bertipe numerik interval atau rasio tidak dapat menggunakan algoritma ID3. Sehingga dapat diketahui kelebihan algoritma C4.5 daripada algoritma ID3 antara lain, dapat menangani atribut dengan tipe numerik, memangkas pohon keputusan, dan menurunkan rule set. Penentuan fitur

atau atribut sebagai pemecah simpul pada algoritma C4.5 menggunakan kriteria [6].



Gambar 2. 1 Bentuk Pohon Keputusan

Hal pertama yang dilakukan dalam algoritma C4.5 adalah menghitung akar pohon. Akar akan diambil dari atribut – atribut yang akan dipilih, dan dengan menghitung nilai gain dari masing-masing atribut maka nilai gain ratio tertinggi akan menjadi akar pertama. Sebelum menghitung nilai gain dari atribut, terlebih dahulu menghitung nilai entropy.

$$Entropy(\mathrm{D}) = \sum_{i=1}^{m} -p_i \log_2(p_i)$$

Ket : D = Himpunan kasus

m = Jumlah partisi S

pi = Proporsi Si terhadap S

$$Gain\ (A)\ =\ Entropy(D)\ -\sum_{i=1}^{n} \frac{|D_i|}{|D|} * Entropy(D_i)$$

Ket : D = Himpunan Kasus

A = Atribut

n = Jumlah partisi atribut A

$|Si|$ = Jumlah sampel pada partisi ke -i

|S| = Jumlah sampel dalam S

Untuk bisa menghitung Gain Ratio diperlukan sebuah term yang baru yaitu Split Information. Yang dimana untuk memilih atribut test untuk simpul menggunakan nilai Gain Ratio tertinggi pada setiap atributnya. Split Information menggunakan normalisasi pada Information gain dengan persamaan.

$$SplitInfo_A\ (D) = -\sum_{i=1}^{v} \frac{|Dj|}{|D|}\ x\ log_2(\frac{|Dj|}{|D|})$$

Pemilihan pada Atribut sebagai Root Node menggunakan perhitungan Gain Ratio. Gain Ratio menggunakan persamaan.

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo_A(D)}$$

2.2.4. Particle Swarm Optimization

Particle Swarm Optimization adalah salah satu algoritma optimasi Swarm Intelligence (SI). Algoritma SI terutama terinspirasi oleh kelompok hewan yang bekerja sama tanpa pemimpin. Selain PSO, masih ada algoritma SI lainnya, seperti Bee Colony, Bat Algorithm, Cuckoo Search, dan algoritma lainnya. Jadi Particle Swarm Optimization atau disingkat PSO. PSO memiliki tiga komponen utama, antara lain: partikel, komponen kognitif dan sosial, dan kecepatan partikel. Setiap partikel mewakili solusi. Pembelajaran partikel terdiri dari dua elemen, kombinasi pengalaman partikel (disebut pembelajaran kognitif) dan pembelajaran dari seluruh populasi (disebut pembelajaran sosial). Pembelajaran kognitif sebagai pBest adalah posisi terbaik yang pernah dicapai partikel, dan pembelajaran sosial sebagai gBest adalah posisi terbaik untuk semua partikel di swarm. pBest dan gBest menghitung kecepatan partikel, dan kecepatan menghitung posisi berikutnya [8]. Tahapan penggunaan algoritma PSO adalah sebagai berikut

a. Initialization: Inisialisasi sekumpulan partikel dengan posisi dan kecepatan acak dalam dimensi D ruang pencarian.

b. Evaluation: Mengevaluasi kebugaran setiap partikel dalam populasi.

c. Update: Hitung kecepatan setiap partikel dengan persamaan, dan pindah ke posisi berikutnya sesuai rumus persamaan.

d. Termination: Jika kondisi terminasi yang ditentukan terpenuhi, hentikan algoritma; jika tidak, lanjutkan ke langkah selanjutnya.

koefisien akselerasi terdapat empat alternatif, yaitu constant acceleration coefficient, self-tuned (ST), self-adaptive acceleration coefficient (SAAC) dan fitness-based acceleration coefficient (FR). Dalam koefisien percepatan constant nilai C_1 dan C_2 selalu sama, yaitu 2. Dalam koefisien self-tuned (ST) nilai C_1 dan C_2 seperti pada persamaan.

2.2.5. Seleksi Fitur

Dalam penelitian [15] dikatakan bahwa banyak masalah seperti klasifikasi, sejumlah besar fitur diperkenalkan untuk menggambarkan konsep target dengan baik. Namun, sejumlah besar fitur dapat menyebabkan masalah, yang merupakan hambatan utama untuk klasifikasi. Pada saat yang sama, tidak semua fitur berguna untuk klasifikasi. Fitur yang tidak relevan dan berlebihan bahkan meningkatkan tingkat kesalahan klasifikasi. Pemilihan fitur dapat mengurangi jumlah fitur dengan menghilangkan fitur yang tidak relevan dan berlebihan, sehingga meningkatkan efisiensi dan/atau meningkatkan kinerja klasifikasi. Di banyak aplikasi, ukuran kumpulan data sangat besar sehingga pembelajaran mungkin tidak berfungsi dengan baik hingga fitur yang tidak diinginkan ini dihapus. Ini membantu untuk lebih memahami konsep di balik masalah klasifikasi dunia nyata. Metode seleksi fitur mencoba untuk memilih subset fitur yang relevan dengan konsep target.

2.2.6. Validasi Hasil

Dalam pembuatan sebuah model klasifikasi diperlukan pengukuran akurasi dari model tersebut. Cara dalam menghitung akurasi dari suatu algoritma dapat menggunakan confusion matrix dan split validation.

a. Confusion Matrix

Confusion matrix adalah salah satu metode yang digunakan untuk menghitung

presisi suatu penelitian atau eksperimen data mining dan sistem pendukung keputusan. Penggunaan Confusion matrix ini berfungsi sebagai ukuran kinerja sistem klasifikasi. Dalam confusion matrix penyajian hasil proses klasifikasi, yaitu True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN). Nilai True Negative (TN) adalah jumlah data negatif yang terdeteksi dengan benar, sedangkan False Positive (FP) adalah data negatif tetapi terdeteksi sebagai data positif. Positif sejati (TP) kemudian adalah data positif yang dikenali dengan benar. False Negative (FN) adalah kebalikan dari True Positive. data positif, tetapi diakui sebagai data negatif.

Tabel 2. 4 Matriks konfusi untuk 2 kelas

f_{ij}		**Kelas hasil prediksi (j)**	
		Kelas = 1	Kelas = 0
Kelas asli (i)	Kelas = 1	f_{11}	f_{10}
	Kelas = 0	f_{01}	f_{00}

Berdasarkan Tabel 2.4 untuk dapat menghitung tingkat akurasi didapatkan melalui penjumlahan data masing-masing kelas yang diprediksi secara benar yaitu (f11 + f00) dibagi dengan jumlah keseluruhan data. Sedangkan untuk menghitung laju error didapatkan melalui penjumlahan data masing-masing kelas yang diprediksi secara salah yaitu (f10 + f01) dibagi dengan jumlah keseluruhan data (Prasetyo, 2014: 257). Perhitungan hasil akurasi dapat dilihat pada persamaan 1. Sedangkan untuk menghitung laju error (kesalahan prediksi) digunakan persamaan dalam rumus berikut

$$Akurasi = \frac{f_{11}+f_{00}}{f_{11}+f_{00}+f_{01}+f_{00}}$$

$$Laju\ error = \frac{f_{10}+f_{01}}{f_{11}+f_{00}+f_{01}+f_{00}}$$

Kinerja dari sebuah algoritma klasifikasi ditentukan dari pengujian model yang dibentuk dengan data uji [6].

b. Split Validation

Split Validation adalah teknik validasi yang secara acak membagi data menjadi dua bagian, satu bagian sebagai data training dan satu bagian sebagai data testing. Dengan menggunakan Split Ratio, eksperimen pelatihan dilakukan sesuai dengan rasio pemisahan yang telah ditentukan, kemudian rasio pemisahan data training yang tersisa diperlakukan sebagai data testing. Data training adalah data yang digunakan untuk pembelajaran, sedangkan data testing adalah data yang belum pernah digunakan untuk pembelajaran dan akan digunakan sebagai data untuk menguji kebenaran atau keakuratan hasil pembelajaran. *Split validation* bagian *Training* digunakan untuk algoritma klasifikasi *Decision Tree* dan pada bagian *Testing* menggunakan operator *Apply Model* untuk mengaplikasikan model pada data testing dan operator *Performance* yang digunakan untuk menampilkan *accuracy.*[16]