BAB II

TINJAUAN PUSTAKA

2.1. Penelitian Terkait

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

Tabel 2.1 . Penelitian Terkait

NO	JUDUL /PENULIS /TAHUN	DATASET	METODE	HASIL
1	Klasifikasi Gejala Penyakit Coronavirus Disease 19 (COVID-19) Menggunakan Machine Learning Suci Anggraini[1] , Muhamad Akbar[2] , Alex Wijaya[3] , Hadi Syaputra[4] , Muhammad Sobri[5] 2021	Dataset yang digunakan peneliti, diambil dari web resmi kaggle.com dari dataset yang ada target (label) yang akan dituju dapat memprediksi gejala dengan tingkatan rendah, sedang dan tinggi	Metode yang digunakan KDD untuk mengklasifikasi pasien terjangkit covid-19 dengan mencari pola data pada dataset. Penelitian ini menggunakan 4 (empat) algoritma yaitu K-Nearest Neighbor (K-NN), Neural Network (NN), Random Forest (RF), dan Naive Bayes dengan bantuan tool rapidminer	Berdasarkan hasil dan pembahasan dari dataset yang telah diuji menggunakan software rapidminer menunjukan nilai absolute count 25,98% untuk tingkat rendah, 54,33% untuk tingkat sedang, dan 19,69% untuk tingkat tinggi. Algoritma Neural Network mempunyai nilai akurasi tertinggi sebesar 73,68%, Random Forest memiliki nilai akurasi sebesar 68,42%, Naive Bayes mendapatkan nilai akurasi sebesar 65,38%, sedangkan Algoritma K-Nearest Neighbor memiliki nilai akurasi paling rendah yaitu 57,89%.

2	Penerapan Data Mining Terhadap Data Covid-19 Menggunakan Algoritma Klasifikasi Rizka Dahlia1 , Nanik Wuryani2 , Sri Hadianti3 , Windu Gata4 , Arina Selawati5 2021	Penelitian in menggunakan data covid 19 didapatkan di repository kaggle.com dengan nama science for covid-19 (ds4c)	Dalam penelitian ini peneliti menggunakan Rapidminer dengan algoritma klasifikasi yaitu Naive Bayes, C4.5, dan Knearest Neighbor dengan melakukan tahapan seleksi, preprocessing, transformasi, data mining, dan data mining. -19 di Korea Selatan. Interpretasi atau evaluasi.	Sebagai hasil dari klasifikasi penelitian ini menggunakan jenis kelamin, usia, kota, kasus infeksi, dan atribut negara bagian menggunakan algoritma c4.5, nave Bayes, dan Rapidminer knn, banyak pasien berpartisipasi dalam kelas terisolasi. .. Dari ketiga algoritma tersebut, algoritma Naive-Bayes mendapatkan akurasi tertinggi dengan hasil 80,79 pada hasil yang terisolasi dan memiliki AUC sebesar 0,881.
3	Machine Learning Algorithms for Predicting SARS-CoV-2 (COVID-19) – A Comparative Analysis L. William Mary, S.Albert Antony Raj	Bagian ini menjelaskan dataset yang digunakan dalam penelitian ini. Untuk memprediksi SARS-CoV-2 diperoleh set data dari survei umum. Dataset memiliki 350 instansi dengan banyak atribut, seperti terinfeksi dan pria dan	Metode data mining dengan menganalisis perbandingan kinerja algoritma klasifikasi SVM, KNN, dan Naïve Bayes.	Dalam penelitian ini tentang prediksi dengan Machine Leraning telah mencapai berbagai tingkat klasifikasi ketepatan. Menggunakan variabel numerik dan kategorikal variabel untuk klasifikasi. Hasil akhirnya adalah Mendukung algoritma Mesin Vektor-85% akurasi, Algoritma K-Nearest Neighbor 80% akurasi, dan Akurasi 66% Naïve Bayes. Diantara ketiganya metode

	2021	wanita yang tidak terinfeksi dari berbagai negara dan kota.		klasifikasi, algoritma SVM menyediaka akurasi tinggi untuk sampel yang disediakan
4	Data Mining untuk Prediksi Status Pasien Covid-19 dengan Pengklasifikasi Naïve Bayes Dewi Yanti Liliana[1] , Hata Maulana[2] , Agus Setiawan[3] 2021	Dataset pasien Covid19 Indonesia diambil dari www.kaggle.com dan diaplikasikan ke RapidMiner.	Penelitian ini menerapkan teknik prediktif di bidang data mining untuk mengklasifikasi kedaruratan pasien. Kami menerapkan pengklasifikasi Naive Bayes untuk membuat model berdasarkan usia dan jenis kelamin yang paling mungkin pulih dari Covid19, dan pasien yang kemungkinan besar akan melanjutkan pengobatan atau meninggal.	Hasil penelitian ini menunjukkan bahwa klasifikasi Naive Bayesian memiliki akurasi yang tinggi yaitu 96,67% dalam mengklasifikasikan kondisi pasien.
5	Prediksi Penyakit Ginjal Kronis Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization Toni Arifin[1] ,	Dalam penelitian ini, digunakan dataset yang didapatkan dari website The UCI Machine Learning Repository.	Menggunakan Algoritma Naive Bayes Classifier Berbasis Particle Swarm Optimization	Dari hasil penelitian naive bayesian classifier. berdasarkan Swarm Optimization memiliki akurasi confusion matrix sebesar 98,75% AUC sebesar 99%. sedangkan Naive Bayes memiliki akurasi matriks konfusi sebesar 97,00% AUC

	Daniel Ariesta[2]			sebesar 99,8%.
6	Optimasi naive Bayes classifier untuk klasifikasi teks pada e-government menggunakan particle swarm optimization Kuncahyo Setyo Nugroho*), Istiadi, Fitri Marisa	Dataset yang digunakan dalam penelitian ini adalah bersumber dari ortal Online (www.sambat.malangkota.go.id).	NBC yang dioptimalkan menggunakan Particle Swarm Optimization (PSO).	Hasil pengujian dengan cross-validation 10x menunjukkan bahwa optimasi Naïve Bayes menggunakan PSO mencapai akurasi 87,44%, hasil ini lebih baik dari kNN yaitu 75 % dan Naïve Bayes 64,38%

Berdasarkan penelitian terkait pada tabel 2.1 diatas dapat disimpulkan penggunaan algoritma Naïve Bayes dalam klasifikasi penyakit covid 19 menghasilkan akurasi yang paling baik yaitu 96,67 %. Selain itu dari penelitian yang menerapkan algoritma naïve bayes bersamaan dengan Particle Swarm Optimization (PSO) mendapatkan akurasi lebih tinggi dibandingkan yang hanya menerapkan algoritma naïve bayes saja tanpa PSO.

2.2. Covid-19

Covid19 berasal dari betacoronavirus (SARSCoV2), yang menyerang saluran pernapasan bagian bawah dan menyebabkan pneumonia pada tubuh manusia. Virus COVID-19 adalah jenis baru dari virus corona. COVID19 dianggap sebagai kerabat dari sindrom pernafasan akut yang parah (SARS) dan coronavirus sindrom pernafasan Timur Tengah (MERS).(Dahlia et al., 2021). Virus Corona

adalah zoonosis, sebagai akibatnya masih ada kemungkinan virus ini asal menurut fauna sampai akhirnya menular ke manusia (Adiba, 2021). Dalam proses penularan virus corona, yaitu melalui droplet yang keluar dari saluran pernapasan manusia saat penderita batuk, bersin, atau berbicara dan dapat ditularkan melalui permukaan benda yang disentuh penderita, kemudian orang lain menyentuh tangan, benda, mata, hidung atau mulut. Ada beberapa gejala umum yang menandakan seseorang terinfeksi virus corona, yaitu nyeri dada, anismia, vertigo, kerusakan jaringan paru-paru, batuk kering, sesak napas, kelelahan kronis, dan mudah lupa (Saenudin et al., 2021)

2.3. Data Mining

Data mining adalah suatu proses penggalian terhadap data yang berukuran besar yang sebelumnya belum diketahui (Adiba, 2021). Data mining juga merupakan bagian dari proses penggalian pengetahuan dalam suatu database yang istilahnya sering disebut dengan Knowledge Discovery in Database (KDD)(Damayanti et al., 2006). 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 sedang berlangsung di bidang data mining dan Knowledge Discovery telah didorong oleh penemuan dari berbagai hal(Daniel, 2005):

a. Pertumbuhan eksplosif dalam pengumpulan data, seperti yang dicontohkan oleh pemindaian di Supermarket
b. Penyimpanan data di gudang data, sehingga seluruh perusahaan memiliki akses ke database 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 KDD, data mining dilakukan sebelum proses seleksi data, pembersihan data, preprocessing, dan transformasi data. Ada tiga langkah penting dalam KDD sebagai berikut(Damayanti et al., 2006):

a. Data preprocessing Proses

   Proses ini dimaksudkan untuk mengubah data masukan ke dalam format yang sesuai untuk analisis lebih lanjut. Pada tahap ini dilakukan proses penggabungan data dari berbagai sumber, pembersihan data untuk menghilangkan data noise dan duplikasi data, dan pemilihan atribut data yang diperlukan untuk penambangan data mining.

b. Data Mining

   Proses ini dimaksudkan untuk mendapatkan pola dan informasi yang tersembunyi dalam database. Beberapa teknik yang dapat digunakan dalam data mining untuk mendapatkan pola dan informasi yang tersembunyi, yaitu klasifikasi, jaringan syaraf tiruan, pohon keputusan, algoritma genetika, clustering, OLAP (Online Analitycal Processing) dan association rules

c. Postprocessing

   Proses ini dimaksudkan untuk memastikan bahwa hanya hasil yang valid dan bermanfaat yang dapat digunakan oleh pihak yang berkepentingan. Contoh dari proses ini adalah visualisasi, yaitu proses menganalisis dan menemukan data dan hasil data mining dari perspektif yang berbeda.

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(Misdram 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:(Daniel, 2005):

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. Tentukan apakah surat wasiat itu benar-benar dibuat oleh almarhum atau oleh penipuan orang lain
f. Menentukan apakah perilaku keuangan atau pribadi tertentu mengindikasi dari kemungkinan ancaman teroris

Klasifikasi yang manusia lakukan tanpa adanya bantuan dari algortima cerdas komputer diartikan sebagai klasifikasi manual. Berbeda dengan klasifikasi yang dilkukan dengan pemberdayaan teknologi, memiliki beberapa algoritma, diantaranya Naïve Bayes, Support Vector Machine, Decission Tree, Fuzzy dan Jaringan Saraf Tiruan (Wibawa et al., 2018).

2.4.1. Algoritma Naïve Bayes

Algoritma Naïve Bayes, teknik yang digunakan adalah percabangan matematis dengan mencari peluang terbesar pada

pengklasifikasi frekuensi setiap kelas data latih dan biasa dikenal dengan teori probabilistic. Rumus untuk menghitung Naive Bayes dijelaskan pada persamaan di bawah (Dahlia et al., 2021):

$$P(x|y) = \frac{P(y|x)xP(x)}{P(y)} \quad (1)$$

$$P(y)=\sum_{n=1}^{n} P(y|x)P(x) \quad (2)$$

Keterangan :

y = data kelas yang belum diketahui

x = hipotesis data y merupakan suatu kelas spesifik

P(x| y) = kemungkinan hipotesis x berdasarkan kondisi y

P(x) = kemungkinan hipotesis x

P(y| x) = kemungkinan y berdasarkan kondisi pada hipotesis x

P(y) = kemungkinan y

Beberapa prinsip teori Bayesian menyatakan bahwa setiap fitur tidak memiliki ketergantungan kuat pada metrik yang tidak terkait dengan keberadaan fitur lain dalam data yang sama. Nave Bayes dalam hipotesis bahwa pengidentifikasi kelas adalah target pemetaan klasifikasi dikaitkan dengan hipotesis korelasi, dan bukti berupa fitur menjadi entri dalam model klasifikasi. Selain itu, Naïve Bayes Classifier memiliki kinerja yang sangat baik dalam beberapa kasus klasifikasi teks (Hakim et al., 2017)

2.5. 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(Pebriadi & Saubari, 2019). Pemilihan fitur merupakan langkah penting dalam proses

klasifikasi karena fitur yang dipilih secara signifikan mempengaruhi keakuratan klasifikasi.Pengklasifikasian dataset yang memiliki banyak fitur memerlukan proses untuk mengurangi fitur yang tidak penting (Nugroho et al., 2018). Seleksi fitur dapat memiliki dampak yang signifikan pada efektivitas algoritma klasifikasi yang dihasilkan, dan dalam beberapa kasus dapat meningkatkan akurasi klasifikasi masa depan sebagai hasil dari seleksi fitur(Prasetio, 2020). Beberapa peneliti telah membandingkan beberapa algoritma klasifikasi dengan algoritma pemilihan fitur untuk hasil terbaik. (Chandani, 2015).

2.5.1. Particle Swarm Optimization (PSO)

Particle swarm Optimization (PSO) adalah teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter. Dalam Particle Swarm Optimization (PSO), terdapat beberapa teknik optimasi, antara lain pembobotan atribut dari semua atribut atau variabel yang digunakan, pemilihan atribut (seleksi atribut), dan daya seleksi atribut(Mustopa, 2021). Particle Swarm Optimization adalah algoritma yang terinspirasi oleh perilaku sosial hewan seperti burung, lebah, dan ikan. Seekor binatang di dalam. Algoritma PSO akan diperlakukan sebagai partikel. Partikel ini akan dipengaruhi oleh kecerdasan hewan itu sendiri dan kecerdasan partikel lain dalam kelompoknya. Jika sebuah partikel menemukan jalur terlurus dan terpendek menuju sumber makanan, maka yang akan terjadi adalah partikel lain akan mengikuti partikel yang telah menemukan jalur lurus dan terpendek sebelumnya (Hakim et al., 2017).

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 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(Hakim 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.

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

   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.

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

   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 sesuai dengan Persamaan 15 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. 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 7 dan 8.

   $$v_{i,j}^{t+1} = w.\, v_{i,j}^{t} + c_1 r_1 (Pbes_{i,j}^{t} - x_{i,j}^{t}) + c_2 r_2 (Gbes_{g,j}^{t} - x_{i,j}^{t}) \quad (6)$$

$$vj^{max} = k\frac{(xj,max - xj,\min)}{2}\ k \in (0,1] \quad (7)$$

$$if\ v_{ij}(t+1) > v_j^{max}\ then\ v_{ij}(t+1) = vj^{max} \quad (8)$$

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

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.

$$sig(v_{i,j}^{t}) = \frac{1}{2+e-v_{i,j}^{t}},\ j = 1,2,\ldots d \quad (9)$$

$$if\ rand[0,1] > sig(v_{i,j}^{t})\ then\ x_{i,j}^{t+1} = 0$$

$$if\ rand[0,1] < sig(v_{i,j}^{t})\ then\ x_{i,j}^{t+1} = 1$$

$$j = 1,2,\ldots . d \quad (10)$$

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.

$$k = 1 + decimal(s1)\ x\frac{a-1}{2n1-1} \quad (11)$$

2.5.2. Genetic Algorithm (GA)/Algoritma Genetika

Algoritma genetika (GA) menyediakan kerangka kerja untuk mempelajari efek dari faktor-faktor yang diilhami secara biologis seperti pemilihan pasangan, reproduksi, mutasi, dan persilangan informasi genetik. Tiga operator digunakan oleh algoritma genetika:

1. Seleksi (Selection)

Operator seleksi mengacu pada metode yang digunakan untuk memilih kromosom yang akan disalin. Fungsi fit mengevaluasi setiap kromosom (kandidat solusi), dan semakin baik kromosom, semakin besar kemungkinan untuk dipilih untuk replikasi.

2. Crossover

Operator crossover melakukan rekombinasi, menghasilkan dua keturunan baru dengan memilih lokus secara acak dan menukar urutan kiri dan kanan lokus itu antara dua kromosom yang dipilih selama seleksi. Misalnya, dalam representasi biner, dua string 11111111 dan 00000000 dapat berpotongan di lokus keenam masing-masing untuk menghasilkan dua substring baru 11111000 dan 00000111.

3. Mutasi (Mutation)

Operator mutasi secara acak mengubah bit atau digit pada posisi tertentu dalam kromosom: namun, biasanya dengan kemungkinan yang sangat kecil. Misalnya, setelah persilangan, substring 11111000 dapat bermutasi untuk posisi kedua pada 10111000. Mutasi tersebut memasukkan informasi baru ke dalam kumpulan gen dan mencegahnya berkumpul terlalu cepat ke arah optimal lokal.

2.6. Cross Validation

Metode validasi silang sering disebut sebagai metode penghapusan, yaitu bertujuan untuk dapat meminimalkan kuadrat dari kesalahan prediksi untuk variabel respons, di mana prediksi respons

didasarkan pada estimator data(Putu & Pratiwi, 2017). Model atau algoritma akan dilatih oleh subset pelatihan dan akan divalidasi oleh subset validasi, kemudian pilihan jenis validasi silang dapat didasarkan pada ukuran kumpulan data Nilai 10Fold dalam validasi silang adalah contohnya dari validasi silang yang ditentukan untuk memilih model terbaik karena cenderung memperkirakan akurasi yang diberikan oleh menjadi lebih baik dalam klasifikasi. Pada nilai 10Fold Cross Validation, data akan dibagi menjadi 10 fold dengan ukuran yang sama, sehingga pada saat mengevaluasi performansi algoritma akan memiliki 10 sub-dataset. (Wong et al., 2019)

2.7. 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). Confusion matrix diartikan matrix 2x2 yang merepresentasikan hasil dari klasifikasi biner pada suatu dataset. Untuk menghitung performa klasifikasi terdapat beberapa rumus umum yang dapat digunakan. Hasil dari nilai akurasi , presisi dan recall bisa ditampilkan berupa persentase (Andika et al., 2019).

2.7.1. Accuracy (Akurasi)

Akurasi adalah salah satu matrix yanf digunakan untuk mengevaluasi model klasifikasi. Secara informal, akurasi merupakan bagian kecil dari prediksi model kami yang benar. Sedangkan secara formal, akurasi memiliki arti sebagai berikut:

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

Akurasi juga dapat dihitung dalam hal negatif dan positif untuk klasifikasi biner sebagai berikut:

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

Dimana TP = True Positif

TN = True Negatif

FP = False Positif

FN = False Negatif

2.7.2. Precision

Precision dalam Confusion Matrix didefinisikan sebagai rasio item terkait yang dipilih dengan semua item yang dipilih. Akurasi adalah kemungkinan bahwa item yang dipilih terkait. Dapat diartikan sebagai kecocokan antara permintaan informasi dan respons terhadap permintaan itu (Swastina, 2013).

2.7.3. Recall

. Recall adalah rasio jumlah dokumen teks terkait yang dikendalikan di antara semua dokumen teks yang relevan dalam suatu koleksi (Andika et al., 2019). Recall adalah probabilitas bahwa item terkait dipilih. Recall dapat dihitung sebagai jumlah rekomendasi relevan yang dipilih oleh pengguna dibagi dengan jumlah semua rekomendasi yang relevan, dipilih dan tidak dipilih (Swastina, 2013)

2.8. 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

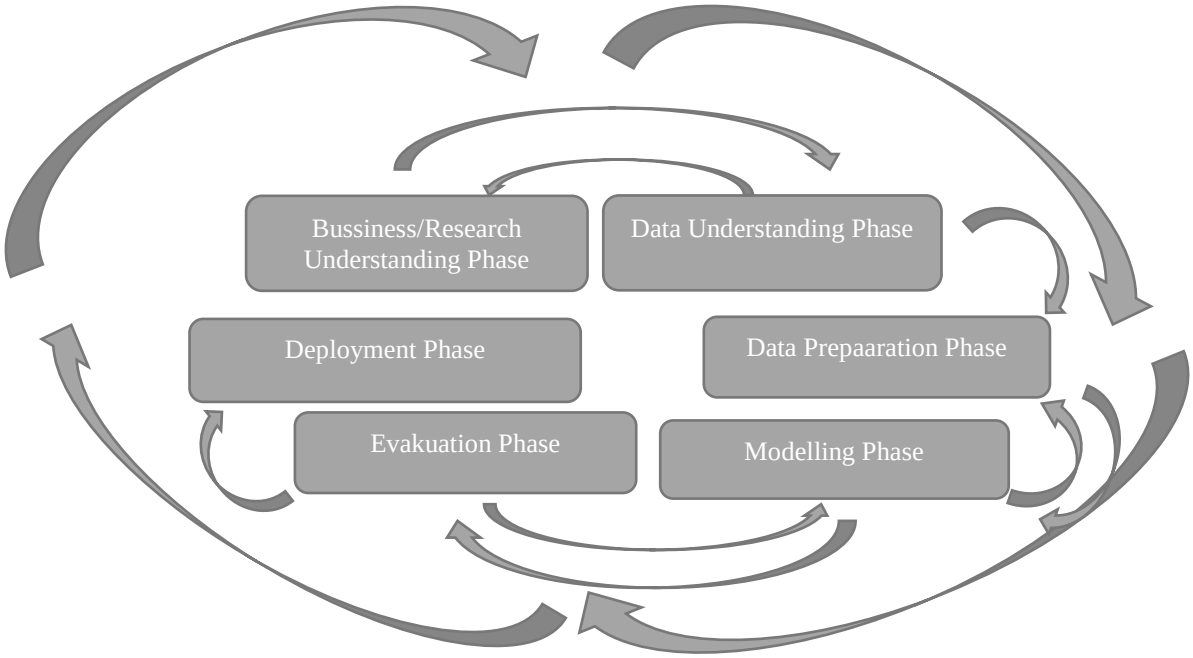
adalah salah satu metrik evaluasi terpenting untuk memeriksa kinerja model klasifikasi apa pun. Itu juga ditulis sebagai AUROC (Area Di Bawah Karakteristik Operasi Penerima) (Sarang Narkhede, 2018). Metode umum untuk menghitung daerah dibawah kurva ROC yaitu Area Under Curve (AUC) dimana bidang yang berada dibawah kurva mempunyai nilai yang selalu berada pada nilai 0,0 dan 1,0. Namun yang menarik untuk dihitung adalah yang mempunyai luas diatas 0,5, semakin tinggi luasnya maka akan semakin baik seperti yang disajikan berikut ini (Hadianto et al., 2019):

- 0.9-1.0 = klasifikasi yang sangat baik (*Excellent Classification*)
- 0.8-0.9 = klasifikasi baik (*Good Classification*)
- 0.7-0.8 = klasifikasi rata-rata (*Fair Classification*)
- 0.6-0.7 = klasifikasi rendah (*Poor Classification)*
- 0.5-0.6 = kegagalan (*Failure Classification)*

2.9. Metode CRISP–DM

CRISP menyediakan proses standar, tidak berpemilik, dan tersedia secara bebas untuk mengintegrasikan penambangan data ke dalam strategi pemecahan masalah keseluruhan perusahaan atau unit penelitian. Menurut CRISPDM, beberapa proyek data mining memiliki siklus hidup yang terdiri dari enam fase, seperti yang diilustrasikan pada Gambar 2.5 . Perhatikan bahwa urutan fase adaptif. Artinya, fase berikutnya dalam urutan sering tergantung pada hasil yang terkait dengan fase sebelumnya. Ketergantungan paling signifikan antara fase ditunjukkan oleh panah. Sebagai contoh, anggaplah kita berada dalam fase pemodelan. Tergantung pada perilaku dan karakteristik model, kita mungkin harus kembali ke fase persiapan data untuk penyempurnaan lebih lanjut

sebelum melanjutkan kemodel fase evaluasi.Sifat iteratif CRISP dilambangkan dengan lingkaran luar pada Gambar 2.1.



Gambar 2.1 CRISP–DM adalah proses adaptif yang berulang.

Seringkali, solusi untuk masalah bisnis atau penelitian tertentu mengarah ke pertanyaan menarik lebih lanjut, yang kemudian dapat diserang menggunakan proses umum yang sama seperti sebelumnya. Pelajaran dari proyek-proyek masa lalu harus selalu dibawa sebagai masukan ke dalam proyek-proyek baru. Berikut ini adalah garis besar dari setiap fase. Meskipun mungkin, masalah yang dihadapi selama fase evaluasi dapat mengirim analis kembali ke salah satu fase sebelumnya untuk perbaikan, untuk kesederhanaan kami hanya menunjukkan loop yang paling umum,kembali ke tahap pemodelan.(Daniel, 2005)

2.10. Rapidminer

Rapidminer adalah software yang digunakan untuk mengolah data. Dengan pemanfaatan prinsip dan algoritma data mining. Rapidminer mengekstrak pola dari kumpulan data besar dengan menggabungkan metode statistik, kecerdasan buatan, dan basis data. Rapidminer memungkinkan pengguna untuk dengan mudah menghitung sejumlah besar data menggunakan operator (Hidayati & Nugroho, 2021). . RapidMiner memiliki Graphical User Interface (GUI) yang sangat efektif untuk desain proses analitis. Ini berisi berbagai Repositori untuk proses, operator, data dan membantu dalam manajemen metadata. Ini membantu dalam perbaikan bug dan deteksi kesalahan. Ini adalah alat visualisasi, mudah digunakan tanpa pengkodean dan merupakan paket lengkap dan serbaguna, berisi ratusan pendekatan yang tersedia untuk integrasi data, pembelajaran mesin, dan simulasi (Ghous, 2020)