

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 DAN PENELITI	DATASET	METODE	HASIL
1	Komparasi Metode Klasifikasi Datang Mining Algoritma C4.5 dan Naïve Bayes Untuk Prediksi Penyakit Hepatitis Wisti Dwi Septiani 2017	Machine Learning Repository UCI (Universitas California Invene) dengan alamat web: http://archive.ics.uci.edu/ml/	Klasifikasi data mining algoritma C.45	Nilai akurasi C.45 adalah 77,29% dan Nilai akurasi Naïve Bayes adalah 83,71 %
2	Menerapkan Teknik Data Mining untuk Mengklasifikasikan Pasien Terduga Infeksi Virus Hepatitis C Reza Safdari , Amir Deghatipour , Marsa Gholamzadeh , Keivan Maghooli 2022	Data yang digunakan berasal dari UCI Machine Learning Repository Website	Menggunakan algoritma SVM, Naive Bayes, Decision Tree, Random Forest, Logistic Regression dan ANN	nilai akurasi C. 45 adalah 96,75% dan akurasi tertinggi Random Forest dengan akurasi 97,29%
3	Penerapan Particle Swarm Optimization (PSO) Untuk Seleksi Atribut Dalam Meningkatkan Akurasi Prediksi Diagnosis Penyakit Hepatitis Dengan Metode Algoritma C4.5 Lis Saumi Ramdhan 2016	Database yang berasal dari http://archive.ics.uci.edu/ml/datasets/Hepatitis sebagai subset dari dataset publik yang digunakan dalam proyek statlog eropa	Klasifikasi data mining algoritma C.45	Akurasi algoritma C4.5 senilai 79,33%, sedangkan untuk nilai akurasi Optimasi algoritma C4.5 menggunakan PSO sebesar

				85,00%
4	Klasifikasi Hepatitis C Virus Menggunakan Algoritma C4.5 Classification Of Hepatitis C Virus Using Algorithm C4.5 Susanto dan Nuri 2022	Data yang digunakan berasal dari UCI Machine Learning Repository Website	Menggunakan algoritma C4.5 dilakukan dengan menerapkan metode lain yaitu Metode Adaboost	Nilai akurasi yang dihasilkan dari Algoritma C4.5 dengan Adaboost sebesar 95,60%
5	Model Data Mining Untuk Merancang Aplikasi Diagnostik Penyakit Inflamasi Hati (Hepatitis) Amrin, Omar Pahlevi 2020	Data yang digunakan berasal dari UCI Machine Learning Repository Website	Menggunakan algoritma Naive Bayes, Decision Tree, K-NN	nilai akurasi C. 45 adalah 70,99% dan nilai akurasi K-KN adalah 67,19% nilai akurasi Naive Bayes adalah 66,14%

2.2 Hepatitis

Hepatitis merupakan suatu penyakit peradangan hati yang umumnya disebabkan oleh virus. Selain itu, hepatitis juga bisa disebabkan oleh alkohol dan penyakit autoimun. Hepatitis virus dapat timbul dari aktivitas yang terkontaminasi virus (misalnya penggunaan jarum suntik, obat suntik, jarum transfusi, jarum tato dan tindik, berhubungan seks dengan penderita hepatitis, atau berinteraksi dengan petugas terkait hepatitis). 5 jenis virus hepatitis yaitu A, B, C, D, kemudian E. Ciri-ciri dari masing-masing jenis ini berbeda-beda, sehingga gejala dan pengobatannya juga berbeda-beda (Handarko, Jefry Latu and Alamsyah., 2015).

Virus hepatitis telah menyebar ke seluruh dunia dan merupakan masalah kesehatan masyarakat global yang utama. Tidak semua kasus hepatitis berkembang, tetapi gejala umum hepatitis termasuk demam, mual hingga muntah, lesu (mendengarkan), mudah memar, dan penyakit kuning (jaundice). Jika tidak diobati, hepatitis dapat berkembang menjadi sirosis (kerusakan hati permanen) dan akhirnya gagal hati. Tes darah adalah cara terbaik untuk memeriksa hepatitis, tetapi biopsi hati, yang menghilangkan sepotong kecil jaringan hati untuk pengujian laboratorium, juga dapat dilakukan. Selain itu, dokter dapat

mendiagnosis hepatitis dengan melakukan pemeriksaan fisik terhadap gejala hepatitis, seperti kulit dan mata menguning. Riwayat kesehatan juga diperlukan untuk mengetahui di mana pasien terpapar virus hepatitis. Hepatitis dapat dicegah dengan menghindari faktor risiko penularan hepatitis dan dengan menerima imunisasi dan vaksinasi (Tinggi *et al.*, 2022).

2.3 Data Mining

Data mining adalah bidang untuk memeriksa database yang sudah ada sebelumnya untuk menggali informasi baru. Bidang ini membentuk dasar Analisa dan digunakan untuk membuat prediksi untuk berbagai bidang seperti pemasaran, keuangan, Medis, cuaca, dan lain-lain. Data mining juga didefinisikan sebagai bagian dari proses penggalian pengetahuan dari database. Hal ini sering disebut sebagai penemuan pengetahuan dalam keputusan database (KDD) dan bertanggung jawab untuk penyebaran hasil. Data mining adalah bidang yang pasti digunakan di bidang Medis (Chatterjee, Al Basir and Takeuchi, 2021). Dalam proses prediksi, persyaratan awal adalah memiliki kumpulan data yang jelas. Data yang kita peroleh perlu dibersihkan terlebih dahulu dengan menggunakan teknik seperti machine learning, statistik, dll. Bidang data mining tidak hanya terbatas pada ritel dan penjualan tetapi juga memiliki banyak aplikasi lain.

Data mining memenuhi tujuan utamanya dengan mengidentifikasi valid, berpotensi berguna, dan mudah dimengerti, korelasi dan pola yang ada dalam data yang ada. Tujuan penambangan data ini dapat dicapai dengan memodelkannya sebagai sifat prediktif atau deskriptif.

1. Regresi- Dalam pemodelan statistik, analisis regresi digunakan untuk mencari hubungan antara variabel dependen dan independen.
2. Clustering- Ini melibatkan pengelompokan item "serupa" bersama-sama dalam bentuk cluster menggunakan algoritma seperti K-means clustering.
3. Klasifikasi- Digunakan untuk memprediksi label kelas kategorikal melalui berbagai model seperti naive Bayes, pohon keputusan.
4. Deteksi Anomali- Teknik data mining juga digunakan untuk mendeteksi pola yang tidak biasa, yang tidak sesuai dengan hasil yang diharapkan yang disebut outlier. Mendukung teknik deteksi Anomali Berbasis Vektor, Teknik Deteksi Anomali Berbasis Clustering dan lain-lain.

5. Peringkasan- Tugas ini membantu memberi pengguna informasi yang lebih baik dan ringkas tentang data dan alat yang paling umum digunakan untuk hal yang sama adalah Excel.

Model prediktif bekerja dengan membuat prediksi tentang nilai data, yang menggunakan hasil yang diketahui yang ditemukan dari kumpulan data yang berbeda. Tugas-tugas yang termasuk dalam model data mining prediktif meliputi klasifikasi, prediksi, regresi dan analisis deret waktu. Model datamining prediktif memprediksi hasil masa depan berdasarkan catatan masa lalu yang ada dalam database atau dengan jawaban yang diketahui. Model deskriptif sebagian besar mengidentifikasi pola atau hubungan dalam kumpulan data. Ini berfungsi untuk mengeksplorasi properti dari data yang diperiksa sebelumnya dan bukan untuk memprediksi properti baru. Model deskriptif mencakup tugas yang harus dilakukan sebagai Pengelompokan, Aturan Asosiasi, Peringkasan, dan Analisis Urutan. Model data mining deskriptif menemukan pola dalam data dan memahami hubungan antara atribut yang diwakili oleh data (Adiba, 2021)

2.4. Klasifikasi

Klasifikasi merupakan tatanan yang sangat penting dalam menambang data komunitas. Klasifikasi adalah teknik penambangan data prediktif yang menggunakan hasil yang diketahui dari kumpulan data yang berbeda untuk membuat prediksi tentang data nilai. Masalah dengan akurasi banyak algoritma klasifikasi adalah bahwa informasi diketahui hilang saat memproses data yang tidak seimbang, misalnya ketika distribusi sampel antar kelas sangat miring (Chatterjee, Al Basir and Takeuchi, 2021). Dalam klasifikasi, Anda memiliki variabel target kategoris, seperti strata pendapatan, yang dapat, misalnya, membagi Anda menjadi tiga kelas atau kategori: pendapatan tinggi, pendapatan menengah, dan pendapatan rendah. Model penambangan data memeriksa kumpulan data besar. Setiap kumpulan data berisi informasi tentang variabel target dan satu set input atau variabel prediktor. Contoh tugas klasifikasi dalam bisnis dan penelitian meliputi (Daniel, 2005):

- a. Menentukan apakah transaksi kartu kredit tertentu adalah penipuan
- b. Penempatan mahasiswa baru pada jalur khusus yang berkaitan dengan kebutuhan khusus
- c. Mengevaluasi apakah aplikasi hipotek menimbulkan risiko kredit

- d. Diagnosis adanya penyakit tertentu
- e. Menentukan apakah wasiat itu ditulis oleh almarhum sendiri atau dipalsukan oleh orang lain
- f. Menentukan Apakah Perilaku Keuangan atau Pribadi Tertentu Mengindikasikan Potensi Ancaman Teroris

Klasifikasi manual adalah klasifikasi yang dilakukan oleh manusia tanpa bantuan algoritma komputer cerdas. Ada beberapa algoritma untuk klasifikasi yang dilakukan menggunakan teknologi ini, antara lain naive Bayes, support vector machine, pohon keputusan, fuzzy, dan jaringan syaraf tiruan (Aji Prasetya Wibawa et al., 2018).

2.5 Decision Tree C.45

Pohon keputusan adalah salah satu jenis algoritma penambangan data yang paling populer untuk klasifikasi dan prediksi. Dtree mengatur catatan dalam struktur pohon yang terdiri dari simpul akar, cabang, dan simpul daun. Node akar berada di bagian atas struktur pohon. Node mewakili atribut, cabang mewakili hasil, lalu daun mewakili keputusan. (Khomsah, no date).

Ada beberapa langkah untuk membangun pohon keputusan menggunakan algoritma C4.5 (Tinggi *et al.*, 2022) yaitu:

- a. Memerlukan pelatihan data, dapat diambil dari data historis yang telah terjadi sebelumnya dan dikelompokkan ke dalam kelas-kelas tertentu.
- b. Tentukan akar pohon dengan menghitung nilai gain tertinggi dari setiap atribut atau nilai indeks entropi terendah. Sebelumnya, nilai indeks entropi dihitung menggunakan rumus:

$$Entropy(i) = \sum_{j=1}^m f(i,j) \cdot \log_2 f(i,j) \quad (1)$$

- c. Nilai gain dengan rumus:

$$gain = - \sum_{i=1}^p \frac{n_i}{n} \cdot IE(i) \quad (2)$$

- d. Untuk menghitung gain ratio perlu diketahui suatu term baru yang disebut Split Information dengan rumus:

$$SplitInformation = - \sum_{t=1}^c \frac{s_t}{s} \log_2 \frac{s_t}{s} \quad (3)$$

- e. Selanjutnya menghitung gain ratio

$$Gainratio(S,A) = \frac{Gain(S,A)}{SplitInformation(S,A)} \quad (4)$$

- f. Ulangi langkah 2 sampai semua record telah terpecah. Proses pemisahan pohon keputusan berakhir ketika:
- 1) Semua tupel dalam catatan simpul m adalah kelas yang sama.
 - 2) Atribut dalam dataset tidak dibagi lagi.
 - 3) Cabang kosong tidak memiliki catatan

2.6 Particle Swarm Optimization (PSO)

Particle Swarm Optimization (PSO) adalah teknik optimasi yang sangat sederhana untuk menerapkan dan memodifikasi beberapa parameter. Particle Swarm Optimization (PSO) memiliki beberapa teknik untuk optimasi, seperti meningkatkan bobot atribut untuk atribut atau variabel yang digunakan, memilih atribut (attribute selection), dan seleksi fitur (Mustopa, 2021). Particle swarm Optimization adalah algoritma yang terinspirasi oleh perilaku sosial hewan seperti burung, lebah, dan ikan. Hewan dalam algoritma PSO dianggap partikel. Partikel ini tunduk pada kecerdasan individu hewan itu sendiri dan kecerdasan partikel lain di dalam kelompok. Jika suatu partikel menemukan jalur terpendek yang benar ke sumber makanan, partikel lain akan mengikuti partikel yang sebelumnya menemukan jalur terpendek yang benar (Hakim, Cholissodin and Widodo, 2017).

Dalam mencari solusi yang optimal, partikel tersebut bergerak pada arah yang terbaik sebelumnya, posisi terbaik secara global. Seperti Ke-i dinyatakan }= dalam ruang dimensi. Posisi terbaik sebelumnya di simpan sebagai sebagai =. Indeks partikel terbaik dari semua di antara kawan group dinyatakan sebagai gbest. Kecepatan partikel dinyatakan sebagai =. Modifikasi kecepatan dan posisi partikel dapat di hitung menggunakan jarak *pbest*, *gbest* seperti ditunjukkan persamaan berikut:

$$v_{i,d} = w * v_{i,d} + c1 * R * (pbest_{i,d} - x_{i,d}) + c2 * R * (gbest_d - x_{i,d}) \quad (5)$$

$$x_{i,d} = x_{i,d} + v_{i,d} \quad (6)$$

Keterangan:

$V_{i, d}$ = Kecepatan partikel ke-i pada iterasi ke-i

w = Faktor bobot inersia

$c1, c2$ = Konstanta akselerasi (learning rate)

R = Bilangan random (0-1)

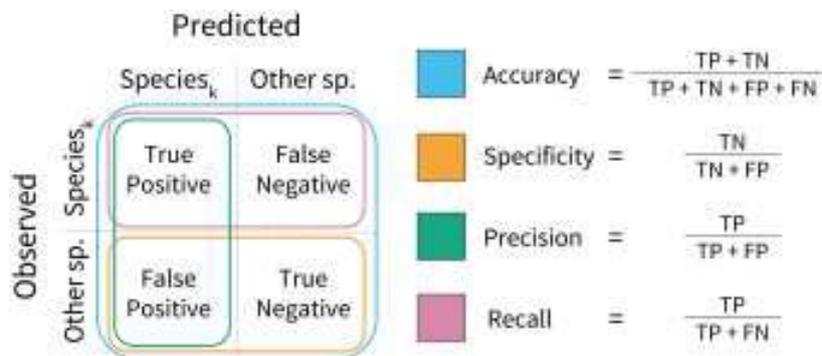
$X_{i, d}$ = Posisi saat ini dari partikel ke-i pada iterasi ke-i

pbesti = Posisi terbaik sebelumnya dari partikel ke-i

gbesti = Partikel terbaik di antara terbaik diantara semua partikel dalam satu kelompok atau populasi.

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 and Kurniawan, 2021). Selama pelatihan model, kinerja dinilai berdasarkan per-sampel menggunakan akurasi model dan skor kerugian log. Akurasi model menghitung proporsi sampel yang diklasifikasikan dengan benar dalam data uji dan nilai akurasi model yang tinggi diinginkan. Log loss menilai apakah probabilitas prediksi dikalibrasi dengan baik, menghukum prediksi yang salah dan tidak pasti. Skor kehilangan log yang rendah menunjukkan bahwa kesalahan klasifikasi terjadi pada tingkat yang mendekati tingkat probabilitas yang diprediksi. Selama pengujian model, kinerja dinilai menggunakan akurasi peringkat-1 dan biaya lintas-entropi (Caprini *et al.*, 2019). Akurasi peringkat-1 dihitung berdasarkan ID spesies mana yang diprediksi dengan probabilitas tertinggi. Skor lintas-entropi mirip dengan fungsi kehilangan log, tetapi diskalakan menggunakan fungsi indikator. Ini dapat ditafsirkan dengan cara yang mirip dengan akurasi dan kehilangan log; akurasi peringkat-1 yang tinggi dan skor entropi silang yang rendah diinginkan.



Gambar 2.1 Metrik Kinerja Model

Metrik pengujian model sekunder dihitung untuk setiap spesies menggunakan data uji. Ini termasuk model spesifisitas, presisi, dan recall. Metrik ini mengungkapkan perilaku model yang skor akurasinya mungkin tidak jelas. Spesifisitas menilai kinerja model pada spesies non-target, menghukum overprediksi spesies target (yaitu, sejumlah besar positif palsu).

Presisi juga menghukum overprediksi, tetapi menilai tingkat overprediksi relatif terhadap tingkat prediksi positif yang benar. Recall menghitung proporsi prediksi positif sejati dengan jumlah total pengamatan positif per spesies. Nilai yang lebih tinggi diinginkan untuk masing-masing. Metrik ini dihitung untuk membantu interpretasi, tetapi tidak digunakan untuk memeringkat kinerja model secara formal. 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:

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}}$$

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

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Dimana TP = Positif Benar, TN = Negatif Benar, FP = Positif Palsu, dan FN = Negatif Palsu.

2.8 Feature Selection

Proses Pemilihan Fitur sangat penting dalam pembelajaran mesin yang sangat memengaruhi kinerja model Anda. Fitur data yang digunakan untuk melatih model atau algoritma pembelajaran mesin memiliki pengaruh besar pada kinerja yang dapat Anda capai. Fitur yang tidak relevan atau hilang dapat berdampak negatif pada kinerja sistem. Pemilihan fitur adalah salah satu langkah pertama dan penting saat melakukan tugas pembelajaran mesin apa pun. Fitur dalam kasus kumpulan data berarti kolom. Ketika kita mendapatkan dataset apa pun, belum tentu setiap kolom (fitur) akan berdampak pada variabel output. Jika kami menambahkan fitur yang tidak relevan ini ke dalam model, itu hanya akan memperburuk model dan dapat mengurangi keakuratan model dan membuat model Anda belajar berdasarkan fitur yang tidak relevan. Hal ini menimbulkan perlunya melakukan seleksi fitur. Ketika datang ke implementasi pemilihan fitur dalam fitur Numerik dan Kategoris harus diperlakukan secara berbeda. Disini kita akan membahas tentang pemilihan fitur Numerik. Oleh karena itu sebelum menerapkan metode berikut, kita perlu memastikan bahwa Data Frame hanya berisi fitur Numerik.

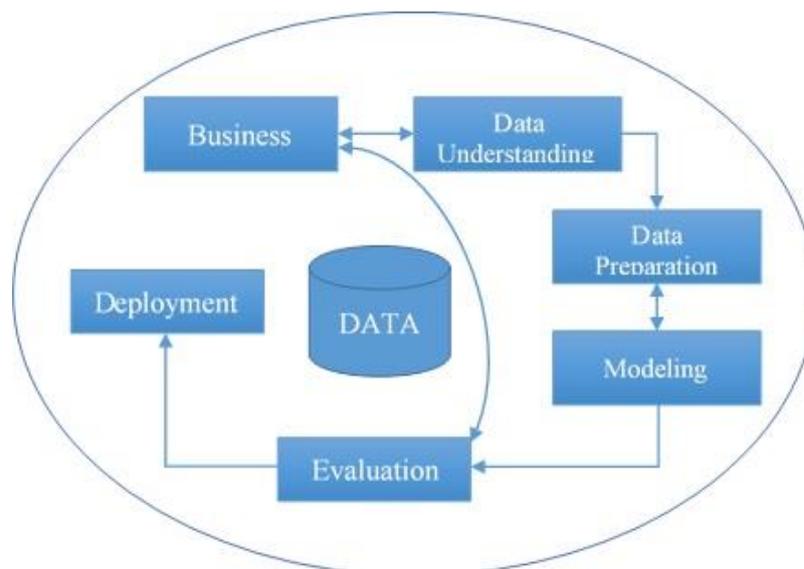
Manfaat melakukan pemilihan fitur sebelum memodelkan data Anda adalah sebagai berikut:

- 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.

Pada sistem ini digunakan dua proses seleksi fitur yaitu proses seleksi fitur sekuensial atau 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 (Bolón-Canedo et al. 2015).

2.9 Metode *Cross Industry Standard Process for Data Mining (CRISP-DM)*

Analisis data menunjukkan bahwa metode CRISP-DM adalah metodologi utama yang digunakan oleh para penambang data.



Gambar. 2.2 Tahapan CRISP-DM

Cross Industry Standard Process for Data Mining (CRISP-DM). CRISPDM merupakan proses dengan enam fase yang secara alami menggambarkan siklus hidup ilmu Data (Mirzakhanyan, 2005), Ini seperti seperangkat pagar pembatas untuk membantu Anda merencanakan, mengatur, dan mengimplementasikan proyek ilmu data (atau pembelajaran mesin). Enam fase tersebut meliputi :

1. Pemahaman Bisnis (Business Understanding)

Fase ini berfokus pada pemahaman tujuan dan kebutuhan proyek. Selain tugas ketiga, tiga tugas lain dalam fase ini adalah aktivitas manajemen proyek dasar yang bersifat universal untuk sebagian besar proyek:

- a. Tentukan tujuan bisnis: Pertama-tama Anda harus “memahami secara menyeluruh, dari perspektif bisnis, apa yang benar-benar ingin dicapai pelanggan.” (CRISP-DM Guide) dan kemudian tentukan kriteria keberhasilan bisnis.
- b. Menilai situasi: Menentukan ketersediaan sumber daya, persyaratan proyek, menilai risiko dan kontinjensi, dan melakukan analisis biayamanfaat.
- c. Tentukan tujuan penambangan data: Selain menentukan tujuan bisnis, Anda juga harus menentukan seperti apa kesuksesan dari perspektif penambangan data teknis.
- d. Menghasilkan rencana proyek: Pilih teknologi dan alat dan tentukan rencana terperinci untuk setiap fase proyek.

Sementara banyak tim terburu-buru melalui fase ini, membangun pemahaman bisnis yang kuat seperti membangun fondasi rumah sangat penting.

2. Pemahaman Data (Data Understanding)

Selanjutnya adalah fase Data Understanding. Menambah dasar Pemahaman Bisnis, ini mendorong fokus untuk mengidentifikasi, mengumpulkan, dan menganalisis kumpulan data yang dapat membantu Anda mencapai tujuan proyek. Fase ini juga memiliki empat tugas:

- a. Kumpulkan data awal: Dapatkan data yang diperlukan dan (jika perlu) masukkan ke dalam alat analisis Anda.
- a. Jelaskan data: Periksa data dan dokumentasikan properti permukaannya seperti format data, jumlah catatan, atau identitas bidang.
- b. Jelajahi data: Gali data lebih dalam. Query, visualisasikan, dan identifikasi hubungan antar data.
- e. Verifikasi kualitas data: Seberapa bersih/kotor datanya?
Dokumentasikan masalah kualitas apa pun.

3. Persiapan data (Data Preparation)

Fase ini, yang sering disebut sebagai “data munging”, menyiapkan kumpulan data akhir untuk pemodelan. Ini memiliki lima tugas:

- a. Pilih data: Tentukan kumpulan data mana yang akan digunakan dan dokumentasikan alasan penyertaan/pengecualian.
- b. Membersihkan data: Seringkali ini adalah tugas terlama. Tanpa itu, Anda mungkin akan menjadi korban sampah-masuk, sampah-keluar. Praktik umum selama tugas ini adalah mengoreksi, mengaitkan, atau menghapus nilai yang salah.
- c. Bangun data: Dapatkan atribut baru yang akan membantu. Misalnya, dapatkan indeks massa tubuh seseorang dari bidang tinggi dan berat badan.
- d. Integrasikan data: Buat kumpulan data baru dengan menggabungkan data dari berbagai sumber.
- e. Format data: Format ulang data seperlunya. Misalnya, Anda dapat mengonversi nilai string yang menyimpan angka menjadi nilai numerik sehingga Anda dapat melakukan operasi matematika.

4. Pemodelan (Modeling)

Di sini Anda mungkin akan membangun dan menilai berbagai model berdasarkan beberapa teknik pemodelan yang berbeda. Fase ini memiliki empat tugas:

- a. Pilih teknik pemodelan: Tentukan algoritma mana yang akan dicoba (misalnya regresi, jaringan saraf).
- b. Hasilkan desain pengujian: Sambil menunggu pendekatan pemodelan, Anda mungkin perlu membagi data menjadi set pelatihan, pengujian, dan validasi.
- c. Model build: Meski terdengar glamor, ini mungkin hanya mengeksekusi beberapa baris kode seperti “`reg = LinearRegression().fit(X, y)`”.
- d. Menilai model: Umumnya, beberapa model bersaing satu sama lain, dan ilmuwan data perlu menginterpretasikan hasil model berdasarkan pengetahuan domain, kriteria keberhasilan yang telah ditentukan sebelumnya, dan desain pengujian.

Meskipun panduan CRISP-DM menyarankan untuk "mengulangi pembuatan model dan penilaian sampai Anda sangat yakin bahwa Anda telah menemukan model terbaik", dalam praktiknya tim harus terus mengulangi sampai mereka menemukan model yang "cukup baik", lanjutkan melalui CRISP -DM siklus hidup, kemudian lebih meningkatkan model di iterasi mendatang.

5. Evaluasi (Evaluation)

Sementara tugas Model Penilaian pada fase Pemodelan berfokus pada penilaian model teknis, fase Evaluasi melihat lebih luas model mana yang paling sesuai dengan bisnis dan apa yang harus dilakukan selanjutnya. Fase ini memiliki tiga tugas:

- a. Evaluasi hasil: Apakah model memenuhi kriteria keberhasilan bisnis?
Yang mana yang harus kami setuju untuk bisnis?
- b. Proses peninjauan: Tinjau pekerjaan yang diselesaikan. Apakah ada yang terlewatkan? Apakah semua langkah dijalankan dengan benar? Ringkas temuan dan perbaiki apa pun jika diperlukan.
- c. Tentukan langkah selanjutnya: Berdasarkan tiga tugas sebelumnya, tentukan apakah akan melanjutkan penerapan, mengulangi lebih lanjut, atau memulai proyek baru.

6. Penyebaran (Deployment)

Sebuah model tidak terlalu berguna kecuali pelanggan dapat mengakses hasilnya. Kompleksitas fase ini sangat bervariasi. Fase terakhir ini memiliki empat tugas:

- a. Merencanakan penyebaran : Kembangkan dan dokumentasikan rencana untuk menerapkan model.
- b. Merencanakan pemantauan dan pemeliharaan: Kembangkan rencana pemantauan dan pemeliharaan yang menyeluruh untuk menghindari masalah selama fase operasional (atau fase pasca proyek) suatu model. Menghasilkan laporan akhir: Tim proyek mendokumentasikan ringkasan proyek yang mungkin mencakup presentasi akhir hasil penambangan data.
- c. Tinjau proyek: Lakukan retrospektif proyek tentang apa yang berjalan dengan baik, apa yang bisa lebih baik, dan bagaimana meningkatkannya di masa depan.

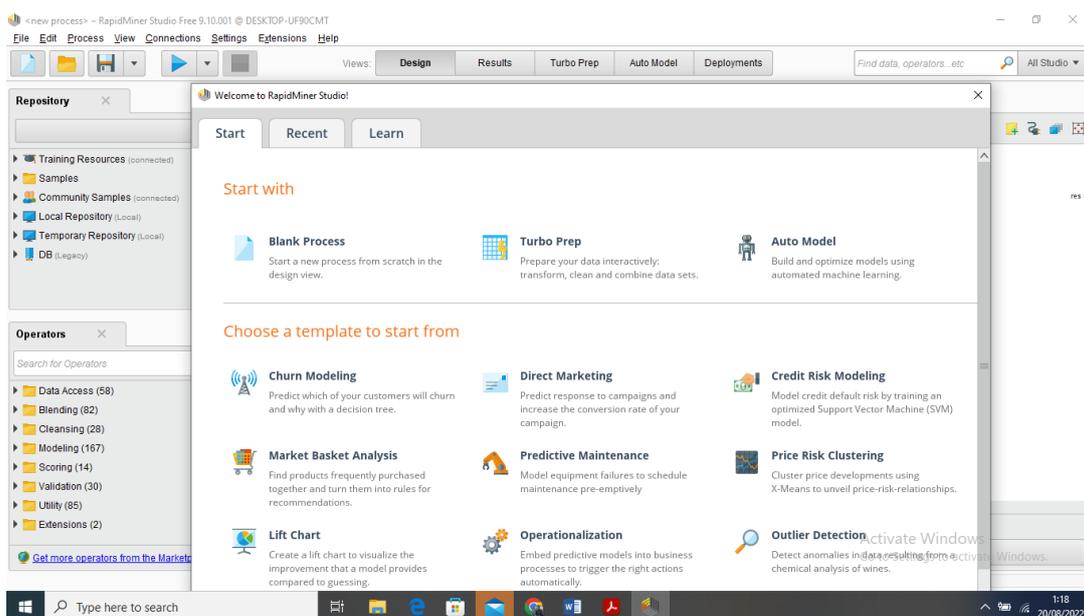
Pekerjaan organisasi Anda mungkin tidak berakhir di situ. Sebagai kerangka kerja proyek, CRISP-DM tidak menguraikan apa yang harus dilakukan setelah proyek (juga dikenal sebagai “operasi”). Tetapi jika model akan diproduksi, pastikan Anda mempertahankan model dalam produksi. Pemantauan konstan dan penyetulan model sesekali sering diperlukan.

2.10 RapidMiner

RapidMiner adalah alat analisis penambangan data yang digunakan untuk menganalisis data dan mendukung berbagai teknik data mining. Digunakan untuk aplikasi industri,

penelitian, pelatihan, pengembangan aplikasi dan pendidikan. Itu mengandung sekitar 100 skema pembelajaran untuk pengelompokan, klasifikasi dan regresi analisis. Ini mendukung sekitar 22 format file seperti .xls, .csv dan sebagainya. Dalam informasi ini dapat diimpor dari berbagai database untuk analisis dan tujuan prediksi (Prasetyo *et al.*, 2021). RapidMiner, sebelumnya dikenal sebagai YALE (Yet Another Learning Environment), dikembangkan mulai tahun 2001 oleh Ralf Klinkenberg, Ingo Mierswa, dan Simon Fischer di Artificial Intelligence Unit of the Technical University of Dortmund. Mulai tahun 2006, perkembangannya dimotori oleh Rapid-I, perusahaan yang didirikan oleh Ingo Mierswa dan Ralf Klinkenberg pada tahun yang sama. Pada tahun 2007, nama perangkat lunak diubah dari YALE menjadi RapidMiner. Pada tahun 2013, perusahaan berganti nama dari Rapid-I menjadi RapidMiner. RapidMiner memiliki beberapa sifat sebagai berikut:

1. Ditulis dengan bahasa pemrograman Java sehingga dapat dijalankan diberbagai sistem operasi;
2. Proses penemuan pengetahuan dimodelkan sebagai operator trees;
3. Representasi XML internal untuk memastikan format standar pertukaran data;
4. Bahasa scripting memungkinkan untuk eksperimen skala besar dan otomatisasi eksperimen;
5. Konsep multi-layer untuk menjamin tampilan data yang efisien dan menjamin penanganan data;
6. Memiliki GUI, command line mode, dan Java API yang dapat dipanggil dari program lain.



Gambar. 2.3 Tools Rapid Miner

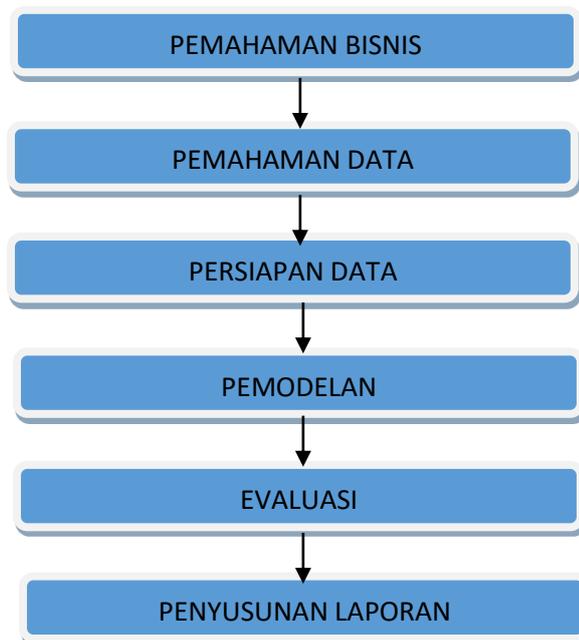
Beberapa fitur RapidMiner antara lain:

1. Banyaknya algoritma data mining, seperti decision tree dan self organization map;
2. Bentuk grafis yang canggih, seperti tumpang tindih diagram histogram, tree chart dan 3D Scatter plots;
3. Banyaknya variasi plugin, seperti text plugin untuk melakukan analisis teks;
4. Menyediakan prosedur data mining dan machine learning termasuk:ETL(extraction, transformation, loading), data preprocessing, visualisasi,modelling dan evaluasi;
5. Proses data mining tersusun atas operator-operator yang nestable, dideskripsikan dengan XML, dan dibuat dengan GUI;
6. Mengintegrasikan proyek data mining Weka dan statistika R.

BAB III METODOLOGI PENELITIAN

3.1 Metode Penelitian

Dalam melakukan analisa dan mencari pola data untuk dijadikan sebuah dataset dalam memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan yang diinginkan maka dibuat alur dalam tahapan penelitian yang akan dilakukan berikut:



Gambar. 3.1 Tahapan Data Mining

Tahapan pada gambar 2 Adalah proses data mining pada penelitian ini :

1. Tahap Pemahaman bisnis juga dapat disebut fase pemahaman penelitian yang meliputi tujuan dan persyaratan proyek dengan jelas dalam hal bisnis atau unit penelitian secara keseluruhan, menterjemahkan tujuan dan batasan ke dalam perumusan definisi masalah data mining, menyiapkan strategi awal untuk mencapai tujuan tersebut.
2. Tahap pemahaman data dengan melakukan pengumpulan data menggunakan analisis data eksplorasi untuk membiasakan diri dengan data dan menemukan wawasan awal, mengevaluasi kualitas data, apabila diinginkan pilih subset menarik yang mungkin berisi pola yang dapat ditindaklanjuti.
3. Tahap persiapan data adalah mempersiapkan data mentah awal kumpulan data akhir yang akan digunakan untuk semua tahapan berikutnya. Pada tahapan ini

memilih kasus dan variabel yang diinginkan, menganalisis yang sesuai untuk analisis. Melakukan transformasi pada variabel tertentu, jika diperlukan, membersihkan data mentah sehingga siap untuk alat pemodelan.

Data tersebut berisi 615 observasi dan 14 atribut laboratorium dan nilai demografis pendonor darah dan pasien Hepatitis C data bersumber dari (<https://www.kaggle.com/datasets/fedesoriano/hepatitis-c-dataset?resource=download&select=HepatitisCdata.csv>). Seluruh atribut pada data kecuali Kategori dan Jenis Kelamin adalah nominal dan numerik. Atribut 1 sd 3 merujuk pada data pasien, 4 sampai dengan 13 adalah hasil cek darah laboratorium. Target atau label data yang digunakan adalah Category atribut yang digunakan untuk perhitungan 12 atribut data.

Tabel 3.1 Atribut-atribut Dataset

No.	Atribut	Tipe	Keterangan
1	Att1	integer	(Pasien ID/No.)
2	Age	integer	(Usia)
3	Sex	polynomial	(Jenis Kelamin)
4	ALB	polynomial	(jumlah albumin)
5	ALP	polynomial	(jumlah alkaline phosphatase)
6	ALT	polynomial	(jumlah alanin transaminase)
7	AST	polynomial	(jumlah aspartat aminotransferase)
8	BIL	polynomial	(jumlah bilirubin)
9	CHE	polynomial	(jumlah kolinesterase)
10	CHOL	polynomial	(jumlah kolesterol)
11	CREA	polynomial	(jumlah kreatin)
12	GGT	polynomial	(jumlah gamma- lutamil transferase)
13	PROT	polynomial	(jumlah protein)

Tabel 3.1 Atribut-atribut Dataset

14	Category (Diagnosa)	polynomial	0=Blood Donor (0=donor darah)
			3=Cirrhosis (penyakit hati stadium akhir)
			2=Fibrosis (jaringan parut fibrotik)
			1=Hepatitis (1=peradangan hati)
			0s=Suspect Blood Donor (0s=dugaan Donor Darah)

Tabel diatas menjelaskan masing-masing atribut dari data pada penelitian ini. Tujuan dari dataset ini adalah untuk menentukan diagnosa penyakit hepatitis C.

4. Tahap pemodelan, Tahapan ini dilakukan dengan memilih dan menerapkan teknik pemodelan yang sesuai. Kalibrasi pengaturan model untuk mengoptimalkan hasil, seringkali, beberapa teknik berbeda dapat digunakan untuk masalah data mining yang sama, apabila diperlukan mengulangi kembali ke tahapan persiapan data untuk membawa bentuk data sesuai dengan persyaratan khusus dari teknik data mining tertentu.



Gambar. 1.2 Tahapan Pemodelan

5. Tahap evaluasi adalah mengevaluasi satu atau lebih model yang disampaikan dalam tahap pemodelan untuk kualitas dan efektivitas sebelum menerapkannya untuk digunakan di lapangan.. Pada tahap ini yang akan dilakukan adalah menentukan model sebenarnya mencapai tujuan yang ditetapkan untuk itu pada tahapan pertama, menetapkan apakah beberapa aspek penting dari masalah bisnis atau penelitian belum diperhitungkan secara memadai, mengambil keputusan tentang

penggunaan hasil data mining menggunakan tools RapidMiner. Dengan tools tersebut akan dilakukan pengujian keakurasian dengan beberapa teknik yaitu confusion matrix

6. Tahap penyusunan laporan adalah mendokumentasikan yang berkaitan dengan penelitian yang dilakukan. Pencatatan hasil penelitian dan diterjemahkan dalam pembahasan untuk dapat disajikan dalam bentuk laporan dan dapat dijadikan literatur.

3.2 Kerangka Pemikiran

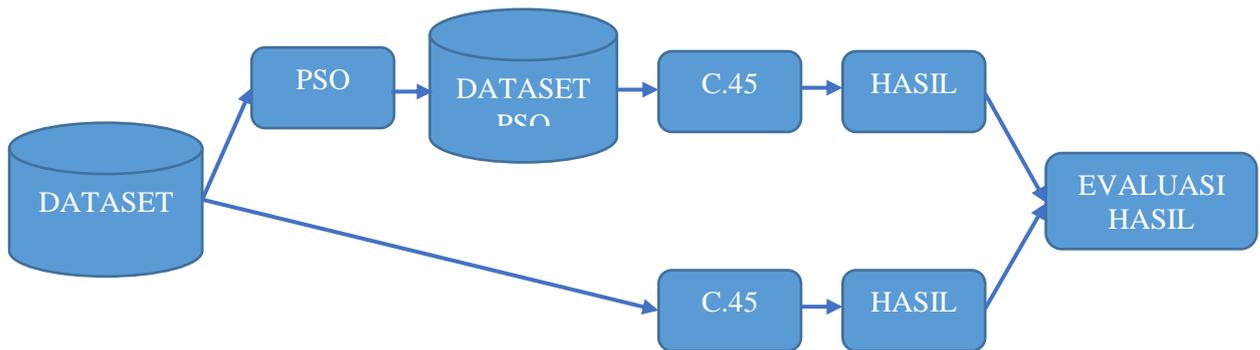
MASALAH		
Hasil yang didapatkan pada percobaan sebelumnya menggunakan algoritma C4.5 lebih rendah dibandingkan Algoritma <i>Particle Swarm Optimization (PSO)</i> maka perlu adanya peningkatan pada algoritma C4.5		
STUDI LITERATUR		
Mencari Referensi-referensi teori yang sesuai dengan permasalahan yang akan dilakukannya penelitian.		
PERCOBAAN		
Data	Metode	Tools
Dataset Hepatitis C	Algoritma C4.5	Rapidminer
	Algoritma C.45+ <i>Particle Swarm Optimization (PSO)</i>	
PENGUKURAN		
Accuracy, Precision, Recall, Performance Vektor		
HASIL		
Teknik optimasi menggunakan algoritma PSO berhasil meningkatkan akurasi dari algoritma C4.5 dalam penentuan hepatitis C.		

Gambar. 3.3 Kerangka Pemikiran

Dalam penelitian ini perlu adanya kerangka pemikiran yang digunakan untuk sebagai landasan serta pedoman agar penelitian ini berjalan sesuai dengan alur yang direncanakan. Permasalahan pada penelitian ini adalah belum adanya metode yang dapat digunakan untuk mengoptimasi data hepatitis C. Metode yang digunakan pada penelitian ini adalah algoritma C4.5 Decision tree dan juga algoritma *Particle Swarm Optimization (PSO)* untuk dilakukan pengujian. Pengujian dari metode yang telah diterapkan menggunakan cara *Confusion Matrix* dan *Accuracy Performance Vector (Performance)*. Untuk *tool* yang digunakan untuk melakukan pengujian metode adalah aplikasi RapidMiner.

3.3 Desain Penelitian

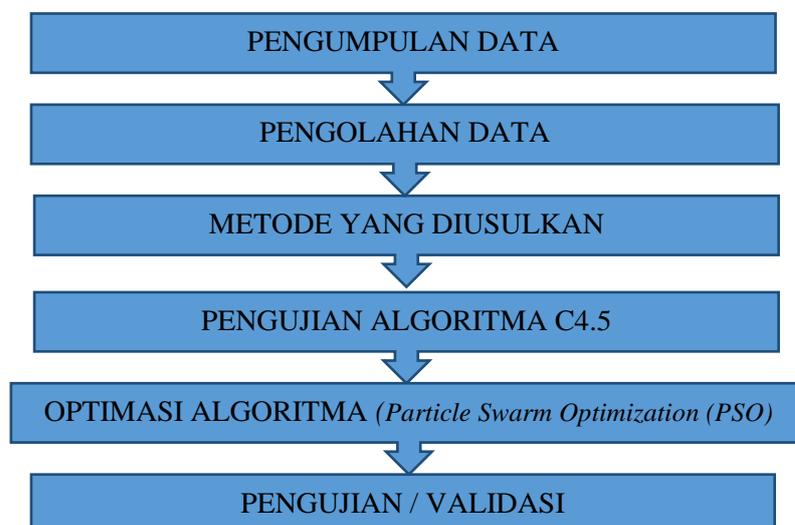
Dalam penelitian ini melalui tahapan dilakukannya uji coba menggunakan metode algoritma C4.5 dan Optimasi *Particle Swarm Optimization (PSO)*. Data dianalisa dengan menggunakan algoritma sesuai dengan metode yang telah di rencanakan, setelah itu membandingkan metode dengan optimasi *Particle Swarm Optimization (PSO)* untuk melihat perbandingan tertinggi tingkat akurasi. Pada tahapan uji coba ini akan dilakukan beberapa langkah atau tahapan pengujian data yaitu seperti berikut:



Gambar. 3.4 Skema Riset

Dari gambar metode yang diusulkan akan mendapat dan mengetahui hasil yang didapatkan oleh algoritma C4.5, kemudian pada hasil tersebut akan ditambahkan algoritma *Particle Swarm Optimization (PSO)* sebagai optimasi yang digunakan untuk meningkatkan hasil dari algoritma C4.5. Pada tahap akhir hasil dari keduanya akan dibandingkan sehingga akan diketahui hasil dari uji coba seberapa efektifnya tingkat optimasi algoritma *Particle Swarm Optimization (PSO)* pada dataset Hepatitis C.

3.4 Tahapan Penelitian



Gambar. 3.5 Tahapan Penelitian

Adapun tahapan penelitian yang akan dilakukan sebagai berikut :

1) Pengumpulan Data

Pada bagian pengumpulan data dijelaskan dari mana data dalam penelitian ini didapatkan, meliputi data kuantitatif dan data kualitatif. Data kuantitatif yang diperoleh dari sumber perusahaan untuk keperluan penelitian, sedangkan data kualitatif berisi tentang data yang dihasilkan dari penelitian.

2) Pengolahan Awal Data

Pengelompokan awal data menjelaskan tentang tahapan awal data mining. Pengolahan awal data meliputi proses input data ke format yang dibutuhkan, dalam pengelompokan dan penentuan atribut data.

3) Metode yang Diusulkan

Metode yang Diusulkan menjelaskan tentang metode yang diusulkan untuk mengoptimasi klasifikasi penyakit hepatitis. Penjelasan ini meliputi pengaturan dan pemilihan dari atribut-atribut yang digunakan sebagai parameter dan arsitektur melalui uji coba.

4) Eksperimen dan Pengujian Metode

Pada bagian Eksperimen dan Pengujian Metode menjelaskan tentang langkah-langkah eksperimen meliputi cara pemilihan arsitektur yang tepat dari model atau metode yang diusulkan sehingga didapatkan hasil yang dapat membuktikan bahwa metode yang digunakan adalah tepat.

5) Evaluasi dan Validasi Hasil

Pada bagian Evaluasi dan Validasi Hasil dijelaskan tentang evaluasi dan validasi hasil penerapan metode pada penelitian yang dilakukan.

3.5 Alat dan Bahan

Penelitian ini menggunakan perangkat keras Laptop Lenovo Thinkpad X260 Intel® Core™ i5 CPU 2.50 Ghz, RAM 8 GB sedangkan perangkat lunak yang digunakan Windows 10 Pro, Aplikasi RapidMiner Studio9.10.001 yang digunakan untuk pengolahan data.