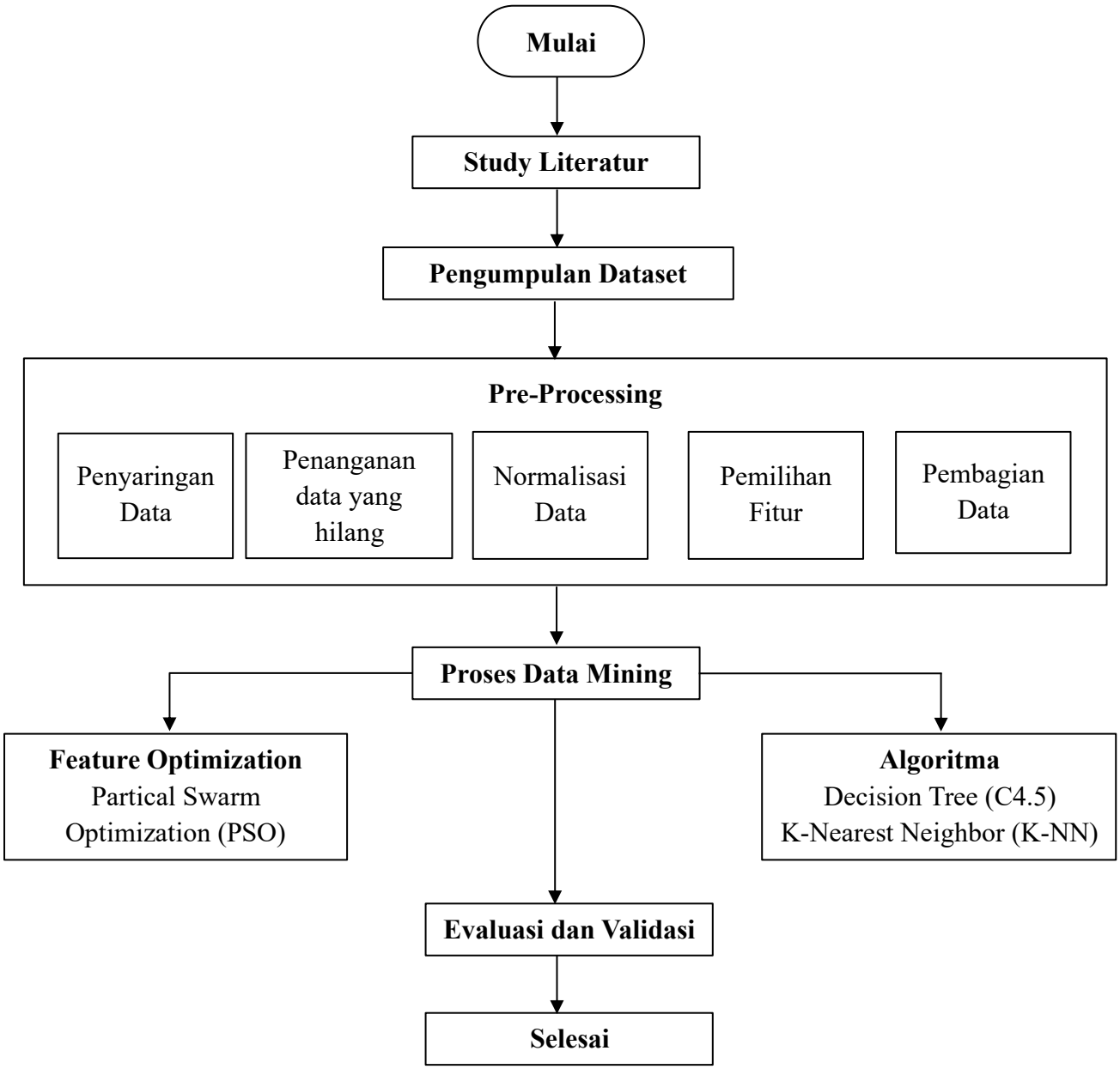
BAB III
METODOLOGI PENELITIAN

Bab ini dimulai dengan memperkenalkan skema penelitian dan alat yang digunakan dalam penelitian ini, yang merupakan alat Machine Learning (ML) open-source yang siap pakai. Selain itu, dijelaskan pula mengenai dataset yang digunakan dalam penelitian, Penelitian ini bertujuan untuk membandingkan performa antara algoritma *K-Nearest Neighbors* (KNN) dan *algoritma C4.5* (C45) dalam klasifikasi kanker payudara menggunakan Rapidminer. Dalam bab ini, akan dibahas mengenai rancangan penelitian, data dan sumber data yang digunakan, teknik pengolahan data, serta analisis data untuk mendapatkan hasil penelitian yang akurat.

Skema Penelitian

Pada bab ini akan membahas langkah-langkah dari proses penelitian yang akan dilaksanakan. Dalam melakukan analisa dan mencari pola data untuk dijadikan sebuah dataset dalam memudahkan penelitian dan dapat berjalan dengan sistematis dan memenuhi tujuan, maka dibuat alur dalam tahapan penelitian yang akan dilakukan sebagai berikut:



Gambar 3.1. Alur Dalam Tahapan Penelitian

Alat dan Bahan

Berikut adalah alat dan bahan yang digunakan dalam penelitian ini:

1. Hardware

Kebutuhan perangkat keras (hardware) yang digunakan : Laptop Infinix INBook_X1, Processor : Intel(R) Core(TM) i3-1005G1 CPU @ 1.20GHz 1.19

GHz. RAM 8,00 GB (7,74 GB usable). System Type : 64-bit operating system, x64-based processor . Harddisk dengan kapasitas 237 GB.

2. Software

Kebutuhan perangkat lunak (software) yang digunakan :

a. Rapid Miner 10.1
b. Microsoft Excel 2019
c. Microsoft Windows 11

Pengumpulan Data

Pengumpulan Dataset.

Data yang digunakan merupakan dataset publik berupa data yang diperoleh dari Kaggle dengan link https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data. yang terdiri dari informasi demografis, kebiasaan, dan catatan medis historis. Beberapa pasien memutuskan untuk tidak menjawab beberapa pertanyaan karena masalah privasi (missing value). Jumlah data adalah 570 data dengan 32 atribut.

Kriteria Inklusi dan Eksklusi

a. Kriteria Inklusi: Menentukan kriteria untuk memilih data pasien yang relevan dengan penelitian ini, misalnya pasien perempuan dengan diagnosis kanker payudara, usia tertentu, dan data yang lengkap.
b. Kriteria Eksklusi: Menentukan kriteria untuk menghapus data yang tidak relevan atau tidak lengkap, seperti data pasien dengan riwayat kanker payudara yang tidak lengkap atau data yang memiliki nilai yang hilang atau tidak valid.

Pre-processing dan Pengolahan Data (Cleaning Data).

Preprocessing dan pengolahan data dilakukan pada Rapidminer dengan melakukan normalisasi, penghapusan fitur yang tidak berguna, dan treatment data yang hilang. Data berjumlah 570 record data dengan 32 atribut.

Penyaringan Data

1. Identifikasi data yang tidak lengkap: Melakukan identifikasi terhadap data yang tidak lengkap, misalnya kolom yang memiliki banyak nilai yang hilang atau data yang tidak relevan.
2. Penghapusan data yang tidak lengkap: Menghapus data yang tidak lengkap dari kumpulan data untuk memastikan konsistensi dan keandalan data yang digunakan dalam analisis.

Penanganan Data yang Hilang

1. Analisis data yang hilang: Menganalisis pola dan karakteristik data yang hilang, seperti apakah data yang hilang secara acak atau memiliki pola tertentu.
2. Teknik pengisian data yang hilang: Menggunakan teknik pengisian data yang hilang, seperti pengisian nilai rata-rata, pengisian dengan nilai median, atau penggunaan metode imputasi data seperti regresi atau K-Nearest Neighbor (KNN) untuk mengisi nilai yang hilang.

Normalisasi Data

1. Pemahaman fitur data: Menganalisis dan memahami fitur-fitur data yang dikumpulkan, misalnya mengidentifikasi apakah ada variabel yang perlu dinormalisasi.
2. Normalisasi data numerik: Melakukan normalisasi data numerik, seperti normalisasi Min-Max atau normalisasi Z-score, untuk memastikan data memiliki skala yang serupa dan tidak mendominasi pengaruh model.

Pemilihan Fitur

1. Analisis fitur: Melakukan analisis statistik atau metode seleksi fitur lainnya untuk mengevaluasi pentingnya setiap fitur dalam mempengaruhi hasil klasifikasi.
2. Pemilihan fitur: Memilih subset fitur yang paling relevan dan memberikan kontribusi signifikan dalam klasifikasi kanker payudara. Metode seperti Information Gain, Chi-Square, atau Recursive Feature Elimination (RFE) dapat digunakan dalam proses pemilihan fitur.

Pembagian Data

1. Data pelatihan dan data uji: Memisahkan data menjadi dua subset, yaitu data pelatihan yang akan digunakan untuk melatih algoritma klasifikasi, dan data uji yang akan digunakan untuk menguji kinerja algoritma.
2. Proporsi pembagian data: Menentukan proporsi pembagian data pelatihan dan data uji, misalnya menggunakan metode pembagian 70:30 atau 80:20, tergantung pada jumlah data yang tersedia dan kompleksitas model yang diinginkan.

Penerapan Data Mining

Penerapan Algoritma K-Nearest Neighbor (K-NN)

Penerapan algoritma K-Nearest Neighbor (KNN) dalam klasifikasi kanker payudara melibatkan beberapa langkah berikut:

1. Persiapan Data:

   Kumpulkan dataset yang berisi informasi tentang pasien, termasuk fitur-fitur yang relevan untuk deteksi kanker payudara seperti usia, ukuran tumor, bentuk tumor, kepadatan, dan sebagainya.

   Lakukan preprocessing data seperti membersihkan data dari missing values atau outlier, dan transformasi data jika diperlukan.
2. Pembagian Data:

   Bagi dataset menjadi dua subset: data training dan data testing.

   Data training digunakan untuk melatih model KNN, sedangkan data testing digunakan untuk menguji kinerja model yang telah dilatih.
3. Normalisasi Fitur:

   Jika fitur-fitur pada dataset memiliki skala yang berbeda, normalisasikan fitur-fitur tersebut ke dalam rentang yang seragam. Hal ini penting untuk memastikan bahwa semua fitur memiliki pengaruh yang seimbang pada perhitungan jarak dalam algoritma KNN.
4. Pemilihan Parameter:

Pilih nilai K yang optimal, yaitu jumlah tetangga terdekat yang akan digunakan dalam pengklasifikasian.

Pemilihan nilai K dapat dilakukan melalui eksperimen dan validasi silang (cross-validation) untuk menentukan performa terbaik.

5. Perhitungan Jarak:

   Hitung jarak antara data testing dengan setiap data training menggunakan metrik jarak seperti Euclidean, Manhattan, atau metrik jarak lainnya.

   Pilih K tetangga terdekat berdasarkan jarak yang terkecil.

6. Voting:

   Berdasarkan K tetangga terdekat, tentukan kelas mayoritas dari tetangga tersebut.

   Gunakan metode voting (misalnya, majority voting) untuk menentukan kelas prediksi untuk data testing.

7. Evaluasi Model:

   Evaluasi kinerja model KNN menggunakan metrik evaluasi seperti akurasi, presisi, recall, atau F1-score.

   Gunakan data testing yang memiliki label kelas yang diketahui untuk membandingkan hasil prediksi dengan nilai sebenarnya dan mengukur kinerja model.

Penerapan Algoritma Decision Tree (C4.5)

Pada bagian ini, akan dilakukan penerapan algoritma Decision Tree C4.5 untuk menklasifikasi kanker payudara. Algoritma C4.5 merupakan salah satu algoritma yang sering digunakan dalam data mining untuk melakukan klasifikasi atau pengelompokan berdasarkan aturan-aturan yang dihasilkan dari pohon keputusan.

Penerapan algoritma C4.5 dalam penelitian ini memberikan manfaat berupa representasi visual dari analisis yang dilakukan menggunakan teknik data mining. Hasil pohon keputusan dari algoritma ini memungkinkan pengguna untuk melihat langkah-langkah yang diambil dalam proses klasifikasi dan prediksi. Keunggulan ini memungkinkan observasi visual terhadap proses prediksi, yang mempermudah interpretasi hasil dan pengambilan Keputusan.

Pohon keputusan telah banyak dimanfaatkan dalam berbagai disiplin ilmu, termasuk dalam bidang kesehatan untuk mendiagnosis penyakit pada pasien, dalam ilmu komputer untuk menganalisis struktur data, dalam psikologi untuk teori pengambilan keputusan, dan dalam berbagai bidang lainnya. Dalam konteks klasifikasi kanker payudara, pohon keputusan berdasarkan algoritma C4.5 dapat membantu dokter atau peneliti dalam melakukan klasifikasi pasien berdasarkan atribut-atribut yang relevan dan menghasilkan prediksi mengenai kemungkinan keberadaan kanker payudara.

Pada pohon keputusan, cabang-cabang yang terbentuk merupakan pertanyaan klasifikasi yang digunakan untuk memisahkan data menjadi kelompok-kelompok yang lebih homogen. Setiap cabang mewakili aturan-aturan yang digunakan untuk mengelompokkan data. Sementara itu, daun-daun pohon keputusan merupakan kelas-kelas atau segmen-segmen yang ditemukan setelah proses klasifikasi selesai.

Rumus yang digunakan dalam algoritma C4.5 untuk menghitung entropy merupakan salah satu kunci utama dalam pembentukan pohon keputusan. Entropy digunakan untuk mengukur tingkat ketidakpastian atau ketidaksempurnaan di dalam data. Dengan menggunakan rumus entropy, algoritma C4.5 akan memilih atribut yang paling informatif untuk dibagi dan membentuk cabang-cabang pada pohon keputusan.

Dengan menerapkan algoritma Decision Tree C 4.5 pada penelitian ini, diharapkan dapat diperoleh model yang efektif untuk menklasifikasi kanker payudara berdasarkan atribut-atribut yang relevan. Pohon keputusan yang dihasilkan akan memberikan gambaran yang jelas dan dapat dimengerti tentang aturan-aturan yang digunakan untuk klasifikasi, sehingga dapat membantu dalam pengambilan keputusan medis yang lebih akurat dan cepat.

Penerapan Particle Swarm Optimization (PSO):

Particle Swarm Optimization (PSO) digunakan untuk mengoptimalkan parameter-parameter yang digunakan dalam algoritma K-NN dan Decision Tree C4.5.

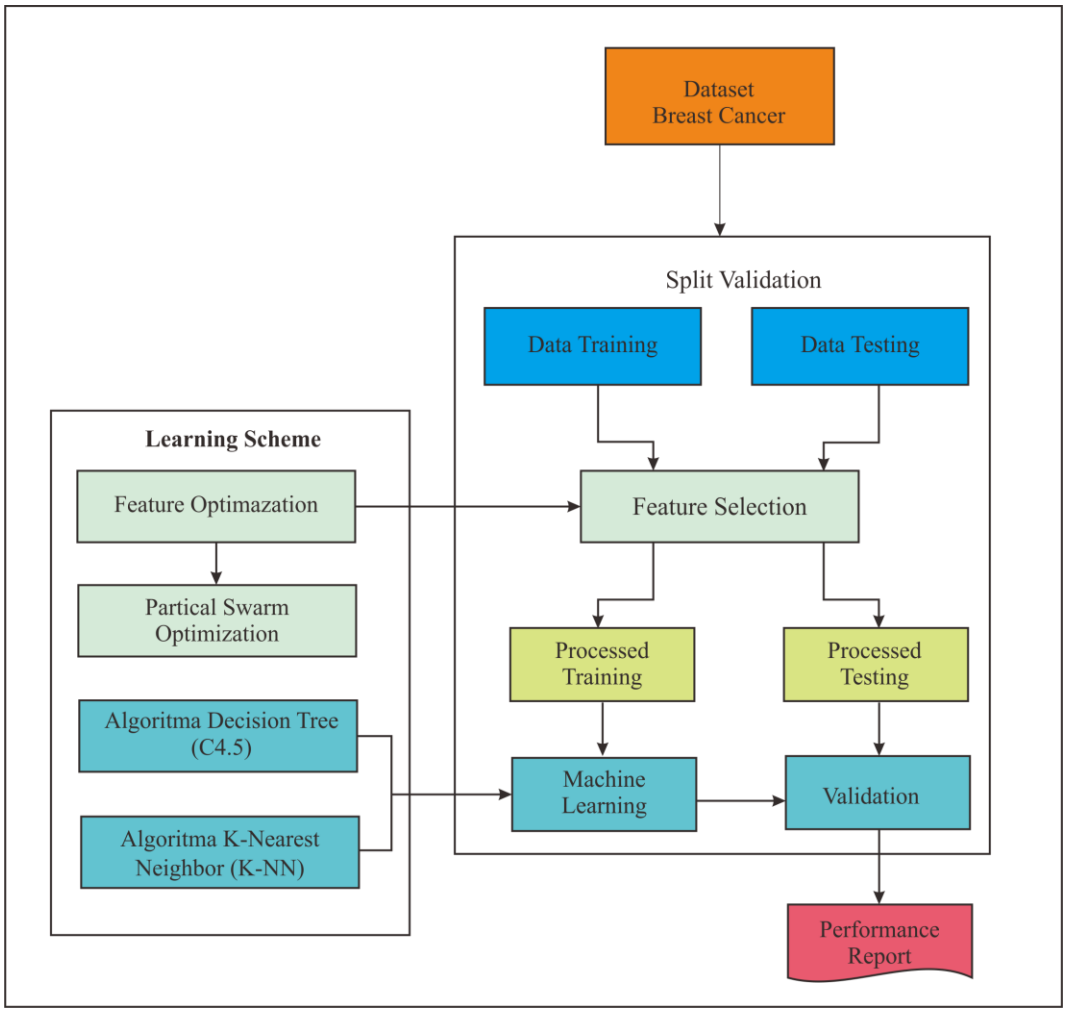
Langkah yang dilakukan adalah menentukan parameter-parameter yang akan dioptimalkan, seperti jumlah tetangga terdekat dalam K-NN dan kriteria splitting dalam C4.5.

Selanjutnya, dilakukan inisialisasi partikel-partikel dalam ruang pencarian.

PSO akan melakukan iterasi untuk mencari solusi terbaik yang menghasilkan akurasi klasifikasi tertinggi.

Solusi terbaik yang dihasilkan akan digunakan untuk mengoptimalkan parameter-parameter dalam kedua algoritma klasifikasi.

Berikut ini adalah proses data Mining yang dilakukan :



Gambar 3.2. Proses Data Mining