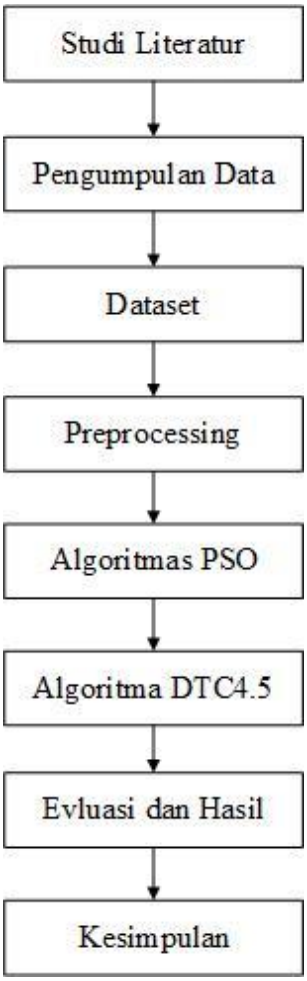
BAB III
METODE PENELITIAN

3.1. Metode 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 yang diinginkan maka dibuat alur penelitian sebagai tahapan dalam penelitian kali ini.



Gambar 3. 1 Alur Penelitian

3.2. Studi Literatur

Tahapan yang pertamakali dilakukan yaitu studi lieratur atau studi awal yang merupakan tahap yang pertama kali dilakukan dalam penelitian. Pada tahap ini, peneliti melakukan observasi atau 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. Dalam studi literatur ini juga dilakukan pencarian terhadap sumber-sumber teori yang relevan dengan topik penelitian dari berbagai sumber seperti buku, penelitian terkait yang ada dalam suatu jurnal, internet, dan lain sebagainya yang mendukung proses penelitian dengan tujuan untuk memperkuat permasalahan.

3.3. Pengumpulan Data

Tahap awal yang dilakukan adalah pengumpulan data. Data yang digunakan dalam penelitian ini adalah data *public* yaitu data tentang wabah monkeypox. pengumpulan data menggunakan analisis data eksplorasi untuk membiasakan diri dengan data dan menemukan wawasan awal dan mengevaluasi kualitas data. Data penelitian yang digunakan diperoleh dari sutus Kaggle.com (https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset)

3.4. Preprocessing

Pada bagian ini merupakan tahapan preprocessing data yang akan disiapkan agar nantinya dapat digunakan dengan Data Transformation dan Split Validation untuk pembagian datanya. Karena kualitas data dapat mempengaruhi akurasi itu sendiri. Dalam data transformation dilakukan tahap pre-processing data yaitu normalisasi data dengan cara MinMax Normalization yang bertujuan untuk membuat beberapa variabel memiliki rentang nilai yang sama, tidak terlalu besar atau terlalu kecil, dengan begitu dapat mempermudah analisis statistik. Sedangkan dalam split validation akan dibagi menjadi data training dan data testing menggunakan perbandingan yang diinginkan. Perbandingan yang digunakan dapat seperti 70:30, 50:50 dan masih banyak perbandingan yang dapat digunakan. Penggunaan perbandingan yang berbeda juga dapat mempengaruhi hasil dari akurasi yang

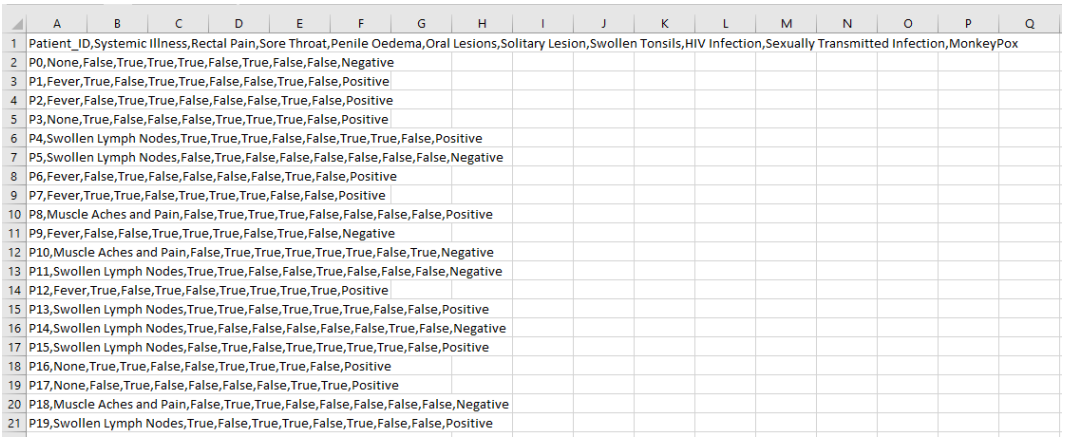
didapatkan nantinya. Dalam preprocessing data akan dilakukan langkah langkah sebagai berikut.

a. Seleksi Data

Proses seleksi data dilakukan untuk memilih data keseluruhan dari dataset yang digunakan. Dataset yang didapatkan berasal dari dataset public yaitu Kaggle.com tentang prediksi Monkeypox. Dari hasil yang di dapatkan akan digunakan data sebanyak 25.000 data pasien yang terinfeksi Monkeypox untuk perhitungan data mining.

b. Integrasi Data

Pada proses ini adalah pemrosesan dari data mentah dalam dataset yang akan diolah sehingga data siap untuk dilakukan preprocessing. Pada data awal yaitu data mentah adalah sebagai berikut.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Patient_ID,Systemic Illness,Rectal Pain,Sore Throat,Penile Oedema,Oral Lesions,Solitary Lesion,Swollen Tonsils,HIV Infection,Sexually Transmitted Infection,MonkeyPox																
2	P0,None,False,True,True,True,False,True,False,False,Negative																
3	P1,Fever,True,False,True,True,False,False,True,False,Positive																
4	P2,Fever,False,True,True,False,False,False,True,False,Positive																
5	P3,None,True,False,False,False,True,True,True,False,Positive																
6	P4,Swollen Lymph Nodes,True,True,True,False,False,True,True,False,Positive																
7	P5,Swollen Lymph Nodes,False,True,False,False,False,False,False,False,Negative																
8	P6,Fever,False,True,False,False,False,False,True,False,Positive																
9	P7,Fever,True,True,False,True,True,True,False,False,Positive																
10	P8,Muscle Aches and Pain,False,True,True,True,False,False,False,False,Positive																
11	P9,Fever,False,False,True,True,True,False,True,False,Negative																
12	P10,Muscle Aches and Pain,False,True,True,True,True,True,False,True,Negative																
13	P11,Swollen Lymph Nodes,True,True,False,False,True,False,False,False,Negative																
14	P12,Fever,True,False,True,False,True,True,True,True,Positive																
15	P13,Swollen Lymph Nodes,True,True,False,True,True,True,False,False,Positive																
16	P14,Swollen Lymph Nodes,True,False,False,False,False,False,True,False,Negative																
17	P15,Swollen Lymph Nodes,False,True,False,True,True,True,True,False,Positive																
18	P16,None,True,True,False,False,True,True,True,False,Positive																
19	P17,None,False,True,False,False,False,False,True,True,Positive																
20	P18,Muscle Aches and Pain,False,True,True,False,False,False,False,False,Negative																
21	P19,Swollen Lymph Nodes,True,False,True,True,False,True,False,False,Positive																

Gambar 3. 2 Dataset sebelum di integrasi

Data yang yang telah dilakukan integrasi adalah pada gambar 3.3 dimana data telah di integrasi dilakukan pemisahan data terhadap setiap atribut yang dipisahkan oleh koma dengan cara pemisahan text ke kolom dimana yang data awalnya hanya dipisahkan oleh koma kemudian data dapat di definisikan untuk dilakukan preprocessing data pada dataset. Berikut adalah dataset yang telah dilakukan integrasi.

	A	B	C	D	E	F	G	H	I	J	K
1	Patient_ID	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	MonkeyPox
2	P0	None	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	Negative
3	P1	Fever	TRUE	FALSE	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	Positive
4	P2	Fever	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	TRUE	FALSE	Positive
5	P3	None	TRUE	FALSE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	Positive
6	P4	Swollen Lymph Nodes	TRUE	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	FALSE	Positive
7	P5	Swollen Lymph Nodes	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	FALSE	Negative
8	P6	Fever	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Positive
9	P7	Fever	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	Positive
10	P8	Muscle Aches and Pain	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	Positive
11	P9	Fever	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	TRUE	FALSE	Negative
12	P10	Muscle Aches and Pain	FALSE	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE	Negative
13	P11	Swollen Lymph Nodes	TRUE	TRUE	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	Negative
14	P12	Fever	TRUE	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	Positive
15	P13	Swollen Lymph Nodes	TRUE	TRUE	FALSE	TRUE	TRUE	TRUE	FALSE	FALSE	Positive
16	P14	Swollen Lymph Nodes	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	TRUE	FALSE	Negative
17	P15	Swollen Lymph Nodes	FALSE	TRUE	FALSE	TRUE	TRUE	TRUE	TRUE	FALSE	Positive
18	P16	None	TRUE	TRUE	FALSE	FALSE	TRUE	TRUE	TRUE	FALSE	Positive
19	P17	None	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE	TRUE	TRUE	Positive
20	P18	Muscle Aches and Pain	FALSE	TRUE	TRUE	FALSE	FALSE	FALSE	FALSE	FALSE	Negative
21	P19	Swollen Lymph Nodes	TRUE	FALSE	TRUE	TRUE	FALSE	TRUE	FALSE	FALSE	Positive

Gambar 3. 3 Data setelah integrasi

c. Missing Data

Missing data dapat diartikan sebagai data atau informasi yang hilang dalam suatu data baik yang hilang ataupun tidak tersedia mengenai subyek penelitian dalam atribut ataupun variable data tersebut. Dalam penelitian ini akan dilakukan pengecekan data hilang pada dataset. Pengecekan dilakukan dengan pengecekan pada kolom dataset mana yang terdapat data yang hilang agar pada proses mining tidak ada data yang missing. Berikut adalah hasil pengecekan data missing pada dataset Monkeypox.

```
Patient_ID                      0
Systemic Illness                0
Rectal Pain                     0
Sore Throat                     0
Penile Oedema                   0
Oral Lesions                    0
Solitary Lesion                 0
Swollen Tonsils                 0
HIV Infection                   0
Sexually Transmitted Infection  0
MonkeyPox                       0
dtype: int64
```

Gambar 3. 4 Pengecekan Missing data pada Dataset

Dari gambar hasil pengecekan missing data yang hilang diatas dapat dilihat bahwa tidak ada data yang missing sehinga data siap untuk dilakukan pengolahan pada tahap yang selanjutnya.

d. Transformasi Data

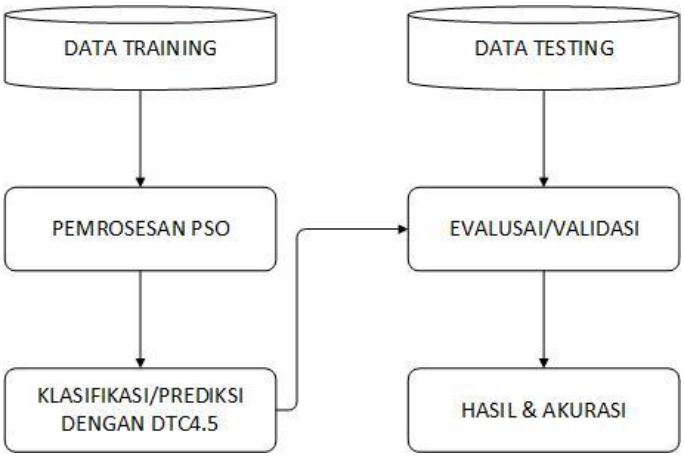
Transformasi data adalah tahapaan yang dilakukan untuk mengubah skala pengukuran data asli menjadi bentuk lain untuk tujuan Analisa data. Dalam penelitian ini akan dilakukan pengelompokan atribut berdasarkan dataset sehingga perubahan data yang dilakukan siap diolah dalam pengolahan yang akan dilakukan dengan bantuan tools rapidminer. Atribut pada dataset berjumlah 11 atribut dimana terdapat 1 atribut yang dijadikan label yaitu atribut monkeypox. Berikut adalah penjelasannya dalam table 3.1.

Tabel 3. 1 Deskripsi Atribut pada Dataset

Nama Atribut	**Deskripsi & Jenis Atribut**	**Kriteria/Keterangan**
Patient_ID	Id pasien	P0-P24999
Systemic Illness	Penyakit Sitemik	- Muscle Aches and Paint - Swollen Lymph Nodes - Fever - None
Rectal Pain	Nyeri Rektum	- True - False
Sore Throat	Sakit Tenggorokan	- True - False
Penile Oedema	Edema Penis	- True - False
Oral Lesions	Lesi Mulut	- True - False
Solitary Lesion	Lesi soliter	- True - False
Swollen Tonsils	Tonsil Bengkak	- True - False
HIV Infection	Infeksi HIV	- True - False
Sexually Transmitted Infection	penyakit menular seksual	- True - False
MonkeyPox	Monkeypox status	- Positive - Negative

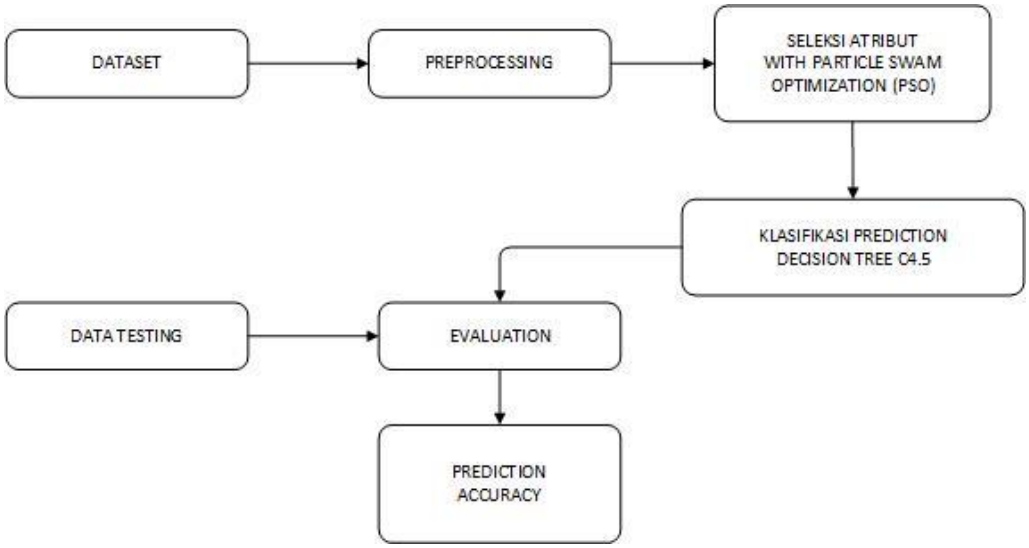
3.5. Proses Mining

Dalam proses mining menggunakan algoritma *Particle Swam Optimization* (PSO) dan Algoritma *Decision Tree* C4.5 untuk meningkatkan akurasi. Dalam tahapan ini juga akan dilakukan perancangan model dari proses mining pada penelitian dan gambaran model yang akan dilakukan dalam pengujian dengan mengkombinasikan algoritma decision tree C4.5 dengan PSO.



Gambar 3. 5 Proses Mining Pada Algoritma PSO dan C4.5

Dalam gambar proses mining diatas setelah dataset dibagi berdasarkan perbandingan maka akan di modeling menggunakan algoritma PSO dan Decision Tree C4.5. diamana pada data training akan dilakukan pemrosesan dengan PSO setelah itu akan diklasifikasi ataupun prediksi menggunakan algoritma DTC4.5 kemudian saat proses evaluasi dan validasi data testing masuk untuk melakukan pemrosesan pada validasi untuk mengetahui hasil dan akurasi. Pada pembentukan pohon keputusan dalam pemrosesan akan ditentukan dari sebuah perhitungan nilai entropy, information gain, split info dan gain ratio. Setelah itu akan dilihat nilai tertinggi dari gain ratio untuk dijadikan sebagai simpul akar dari pohon. Didalam proses pembentukan dilakukan agar seluruh data memiliki kelas dengan cara rekursif sampai decision tree terbentuk. Kemudian pada modeling proses mining dengan PSO dimana pada algoritma ini akan melakukan beberapa tahapan dalam proses pemodelan dimulai dari inisialisasi sampai pada klasifikasi dimana sampai pada kondisi apakah evaluasi nilai dari semua partikel sudah terpenuhi atau belum terpenuhi. Berikut adalah modeling flowchart untuk prediksi penyakit monkeypox.



Gambar 3. 6 Flowchart Modeling PSO dan DTC4.5

Pada Flowchart Modeling PSO dan DTC4.5 diatas dilakukan seleksi atribut. Seleksi atribut ini dilakukan menggunakan algoritma *particle swarm optimization* (PSO). *Dataset* yang telah diseleksi menggunakan PSO kemudian disiapkan untuk proses validasi data untuk dilakukan data uji dimana pada data *training* dilakukan pembelajaran untuk memperoleh pola model kemudian pola tersebut dilakukan pengujian menggunakan data *testing* dengan menggunakan algoritma *decision tree* C4.5 dalam menghitung prediksi akurasi penyakit monkeypox.

3.6. Evaluasi dan Hasil

Pada tahap Evaluasi dan Hasil dilakukan setelah pengolahan data dengan pengujian akurasi, pengujian akurasi pada penelitian ini menggunakan *confusion matrix*. Akurasi akan didapatkan dengan menjumlahkan data yang diprediksi benar, kemudian dibagi dengan keseluruhan data prediksi dan dikali dengan 100%. Pada penelitian ini akan dilakukan beberapa variasi untuk mendapatkan hasil akurasi terbaik yaitu dengan melakukan pengujian algoritma C4.5 akan dilakukan dengan menggunakan perbandingan pembagian data training dan data testing split validation. Kemudian pada pengujian algoritma C4.5 dan seleksi fitur PSO akan dilakukan dengan menggunakan perbandingan pembagian data training dan data testing dengan split validation. Dalam evaluas dan hasil atau dalam validasi hasil

juga digunakan Confusion Matrix dan Cross Validation. Berikut adalah penjelasanya

a. Confusion Matrix

Confusion Matrix digunakan sebagai salah satu alat evaluasi kinerja model. Confusion Matrix adalah sebuah tabel yang digunakan dalam pemrosesan statistik dan machine learning untuk mengukur performa sebuah model klasifikasi. Dalam penelitian ini akan menggunakan Confusion Matrix untuk Mnegukur Performa pada prediksi Wabah Monkeypox. Pada pengukuran performa digunakan nilai-nilai evaluasi matrix sebagai berikut.

1. **Accuracy**: Mengukur sejauh mana model dapat mengklasifikasikan data dengan benar, dinyatakan sebagai (TP + TN) / (TP + TN + FP + FN).
2. **Precision**: Mengukur sejauh mana data yang diklasifikasikan sebagai positif oleh model benar-benar positif, dinyatakan sebagai TP / (TP + FP).
3. **Recall**: (Sensitivity atau True Positive Rate): Mengukur sejauh mana model dapat mendeteksi semua data yang seharusnya positif, dinyatakan sebagai TP / (TP + FN).
4. **ROC AUC**: untuk mengukur kinerja model klasifikasi, terutama dalam konteks klasifikasi biner. ROC AUC mengukur sejauh mana model dapat memisahkan dua kelas yang berbeda dan mengukur kemampuannya dalam mengklasifikasikan data positif dan negatif

b. Cross Validation

Cross Validation (Validasi Silang) adalah teknik yang digunakan dalam pemodelan statistik dan machine learning untuk mengukur kinerja dan keandalan model prediktif. Pada k-fold cross validation akan menggunakan 10 fold validation dengan menggunakan bantuan tool rapid miner

3.7. Alat dan Bahan

Penelitian ini menggunakan perangkat keras Laptop ASUS dengan Processor Intel(R) Core(TM) i3-7020U CPU @ 2.30GHz 1.30 GHz, RAM 4,00 GB,

sedangkan perangkat lunak yang digunakan Microsoft Excel dan RapidMiner Studio untuk pengolahan data dan bahan yang digunakan diambil dari (https://www.kaggle.com/datasets/deepcontractor/monkeypox-dataset-daily-updated).