BAB III
METODE PENELITIAN

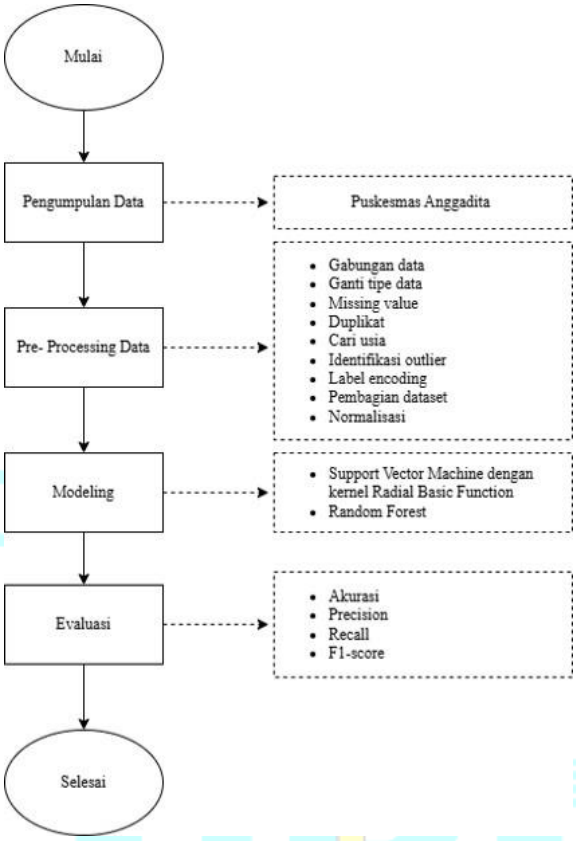
3.1 Objek Penelitian

Objek penelitian ini adalah rekam medis pasien diabetes yang terdiagnosis di setiap tahunnya, data tersebut meliputi, identitas pasien, riwayat penyakit tidak menular pada keluarga/diri sendiri, faktor risiko (merokok, kurang aktifitas fisik, gula, garam, lemak yang berlebihan, kurang makan buah, sayuran dan konsumsi alkohol), pengukuran tekanan darah (sistolik dan diastolik), tinggi badan, berat badan, lingkar perut, Pemeriksaan gula.

Penelitian ini mengumpulkan data dari rekam medis elektronik serta wawancara untuk memperoleh informasi mengenai kebiasaan pola makan dan riwayat keluarga. Proses pengumpulan data dilakukan pada tahun 2023-2024 di Puskesmas Anggadita, Karawang, Jawa Barat. Dengan durasi penelitian selama enam 6 bulan, penelitian ini bertujuan untuk memastikan analisis dan evaluasi model klasifikasi diabetes dilakukan secara optimal. Waktu yang tersedia dimanfaatkan untuk pengolahan data secara mendalam, termasuk tahap pembersihan, transformasi, dan seleksi fitur. Penelitian ini berfokus mengembangkan model klasifikasi yang akurat dalam mendeteksi diabetes, sehingga dapat mendukung pengambilan keputusan medis di Puskesmas Anggadita.

3.2 Prosedur Penelitian

Penelitian ini dirancang dengan prosedur yang terstruktur dan sistematis untuk memastikan setiap langkah dilakukan secara terorganisasi. Prosedur tersebut mencakup beberapa tahap, dimulai dari pengumpulan data hingga evaluasi hasil penelitian. Setiap tahap saling terkait dan dirancang untuk menghasilkan kesimpulan yang valid serta sesuai dengan tujuan penelitian. Secara garis besar, alur tahapan penelitian ini ditunjukkan pada Gambar 1.



Gambar 3. 1 Tahap Penelitian

3.2.1 Pengumpulan Data

Data penelitian diperoleh dari hasil pemeriksaan di Puskesmas Anggadita, dengan total 1001 data pasien. Data ini terbagi ke dalam dua kelas, yaitu ‘Diabetes’ untuk pasien yang terdiagnosis diabetes dan “Tidak” untuk pasien yang tidak terdiagnosis diabetes. Dataset tersebut mencakup berbagai informasi relevan, seperti identitas pasien, riwayat penyakit tidak menular pada keluarga/diri sendiri, faktor risiko (merokok, kurang aktifitas fisik, gula, garam, lemak yang berlebihan, kurang makan buah, sayuran dan konsumsi alkohol).

Data juga mencakup pengukuran tekanan darah (sistolik dan diastolik), tinggi badan, berat badan, lingkar perut, Pemeriksaan gula. Seluruh informasi ini disusun dalam format tabular seperti pada *Excel*, yang bertujuan untuk mempermudah proses analisis dan pengolahan data pada tahap selanjutnya. Adapun dataset dari objek penelitian yang berupa tabel sebagai berikut.

Tabel 3. 1 Dataset Diabetes

No	Nama	Sistol	Diastol	Berat Badan (Kg)	Pemeriksaan Gula	Diagnosis
1	Ruswanti	145	85	62	356	Diabetes
2	Ikarsem	192	113	51	337	Diabetes
3	Pipin	154	81	96	96	Tidak
4	Saturi	127	116	66	112	Tidak
5	Darwil	141	54	41	113	Tidak
…	…	…	…	…	…	…
1001	Sarwono	160	90	66	207	Diabetes

3.2.2 Pre-Processing Data

Data yang tersedia adalah data per tahun, sehingga perlu digabungkan data pertahun tersebut menjadi dataset utuh agar lebih terstruktur. Setelah data berhasil digabungkan, langkah selanjutnya adalah melakukan pemeriksaan dan penyesuaian tipe data pada masing-masing atribut. Misalnya pada atribut tinggi badan dan berat badan, yang awalnya bertipe *object*, diubah menjadi tipe numerik *float* supaya dapat dipakai dalam perhitungan matematis dan pemodelan *machine learning*. Konversi tipe data ini bertujuan untuk memastikan semua atribut numerik memiliki format yang tepat guna mendukung analisis lanjutan.

Langkah berikutnya adalah memeriksa nilai yang hilang (*missing values*) pada dataset. Untuk atribut numerik yang memiliki data kosong, dilakukan proses imputasi dengan menggantinya menggunakan nilai rata-rata (*mean*) dari kolom terkait. Sementara itu, untuk atribut kategorikal, nilai yang kosong diisi menggunakan nilai modus (*mode*), yaitu nilai yang paling sering muncul dalam kolom tersebut. Apabila terdapat nilai kosong pada atribut diagnosis, baris data tersebut akan dihapus guna menjaga kualitas dan ketepatan dalam proses klasifikasi.

Tahap selanjutnya pengecekan data duplikat. Pada tahap ini, dilakukan pemeriksaan seluruh entri data untuk memastikan tidak ada data yang tercatat lebih dari satu kali melalui teknik pendeteksian duplikasi. Data yang terindikasi duplikat kemudian dihapus untuk memastikan bahwa setiap baris mewakili individu yang berbeda, sekaligus menghindari bias dalam hasil analisis model.

Tahapan *pre-processing* dilanjutkan dengan penambahan atribut baru berupa usia pasien. Usia dihitung dengan mengurangi tanggal lahir dari tanggal pemeriksaan, lalu hasilnya dikonversi ke dalam satuan tahun. Nilai usia yang

diperoleh disimpan dalam kolom baru untuk memperkaya dataset, mengingat faktor usia memiliki peran penting sebagai salah satu indikator risiko dalam analisis penyakit diabetes.

Setelah atribut usia berhasil ditambahkan, proses dilanjutkan dengan mendeteksi keberadaan data outlier. Teknik *Z-Score* digunakan untuk mengidentifikasi nilai-nilai ekstrem, di mana data dengan *Z-Score* lebih dari 3 atau kurang dari -3 dikategorikan sebagai outlier. Outlier yang ditemukan kemudian dihapus dari dataset untuk menjaga distribusi data tetap normal dan meningkatkan performa model klasifikasi.

Pada tahap selanjutnya, atribut kategorikal diubah ke dalam format numerik menggunakan metode *Label Encoding*. Variabel seperti jenis kelamin, riwayat penyakit, faktor risiko, rujukan, dan diagnosis dikonversi menjadi nilai angka agar dapat diolah oleh algoritma *machine learning*. Penerapan *Label Encoding* ini bertujuan untuk memudahkan proses pelatihan model yang membutuhkan data input berbentuk numerik.

Setelah semua atribut berhasil dikonversi, dataset kemudian dipisahkan menjadi dua bagian, yaitu 80% untuk data pelatihan (*training*) dan 20% untuk data pengujian (*testing*). Proses pembagian ini dilakukan agar model dapat dilatih menggunakan sebagian besar data dan diuji akurasi nya pada data baru yang belum dikenali sebelumnya, sehingga hasil evaluasi model menjadi lebih objektif dan akurat.

Tahapan akhir yaitu, melakukan normalisasi pada atribut numerik dengan menggunakan metode *Standard Scaler*. Normalisasi ini mengubah data sehingga memiliki rata-rata sebesar 0 dan standar deviasi sebesar 1. Tujuan dari normalisasi adalah untuk menyamakan skala semua fitur, sehingga algoritma *machine learning* dapat berfungsi dengan lebih efektif dan meningkatkan kinerja model secara keseluruhan.

3.2.3 Modeling

Dalam penelitian ini, algoritma *Support Vector Machine* (SVM) dan *Random Forest* diterapkan untuk mengklasifikasikan diabetes. Model SVM diimplementasikan menggunakan kernel *Radial Basis Function* (RBF) dengan nilai

C=1 dan gamma=*'scale'*. Model dilatih menggunakan data latih yang telah dinormalisasi dan dievaluasi kinerjanya pada data uji.

Algoritma 3.1 Membuat Model Klasifikasi dengan SVM

Klasifikasi dengan algoritma SVM
Input : Dataset Training Output : Hasil prediksi model SVM pada data uji 1. Baca dataset 2. Tentukan Kernel yang akan digunakan (RBF) 3. Tentukan nilai *hyperparameter* (C, Gamma) 4. Temukan kombinasi *hyperparameter* terbaik untuk model *training* 5. Melatih model SVM menggunakan data pelatihan untuk mempelajari pola dan membuat *hyperplane* pemisah 6. Hasil prediksi label diabetes pada data uji menggunakan model yang telah dilatih

Pada *Random Forest* juga menggunakan *estimato*r=100 dan *random_state*=42 yang ditentukan secara langsung tanpa melalui proses tuning. Model ini juga dilatih menggunakan data latih yang telah dinormalisasi dan dievaluasi kinerjanya pada data uji.

Algoritma 3.2 Membuat Model Klasifikasi dengan *Random Forest*

Klasifikasi dengan algoritma *Random Forest*
Input : Dataset Training Output : Hasil prediksi model RF pada data uji 1. Baca dataset 2. Membuat objek menggunakan *classifier* dengan *hyperparameter* 3. Melatih model menggunakan data latih 4. Memprediksi label pada data uji 5. Hasil perbandingkan prediksi model dengan label 6. Hasil prediksi label pada data uji menggunakan model yang telah dilatih

3.2.4 Evaluasi

Penelitian ini menggunakan *confusion matrix* dengan evaluasi yang dilakukan untuk mengukur kualitas model klasifikasi menggunakan berbagai matrix tertentu. Akurasi digunakan untuk menggambarkan persentase prediksi yang benar dibandingkan dengan total data yang ada. Presisi mengukur proporsi prediksi positif yang benar dibandingkan dengan seluruh prediksi positif yang dibuat oleh model. pada tahap evaluasi memiliki tujuan untuk mengukur kualitas kemampuan suatu model klasifikasi. Pada tahap ini menggunakan *confusion matrix* guna mencari nilai akurasi, persisi, *recall, dan F1-Score*, hal ini merupakan tahap penting agar dapat menjawab perbandingan model algoritma mana yang memiliki performa tinggi dalam klasifikasi (Ridwan, 2020).

Pada evaluasi yang menggunakan *matrix* terdapat berupa tabel seperti pada tabel berikut, Dimana terdapat record perbandingan hasil klasifikasi data testing atau data uji berdasarkan data *training* atau data latih dengan data sebenarnya (Paramitha et al., 2023).

Tabel 3. 2 *Confussion Matrix*

Kelas Asli	**Prediksi Positif**	**Prediksi Negatif**
Positif	*True Positive* (TP)	*False Negative* (FN)
Negatif	*False Positive* (FP)	*True Negative* (TN)

a. *True positive* adalah jumlah data dengan label positif yang berhasil diklasifikasikan secara benar sebagai positif.
b. *False positive* adalah jumlah data dengan label negatif yang salah diklasifikasikan sebagai positif.
c. *False negative* adalah jumlah data dengan label positif yang salah diklasifikasikan sebagai negatif.
d. *True negative* adalah jumlah data dengan label negatif yang berhasil diklasifikasikan secara benar sebagai negatif.

Keempat metode tersebut memberikan gambaran yang lebih komprehensif tentang kinerja model dalam melakukan klasifikasi sebagai berikut:

Akurasi merupakan tingkat kesesuaian antara hasil prediksi dengan nilai sebenarny (Hidayah & Dodiman, 2024).

$$\text{Akurasi} = \frac{TP + TN}{TP+FP+TN+FN} \text{x}100\% \quad (6)$$

Presisi adalah metode yang digunakan untuk menghitung proporsi data kelas positif yang terprediksi dengan benar dibandingkan dengan seluruh data yang diprediksi sebagai kelas positif (Hidayah & Dodiman, 2024).

$$\text{Presisi} = \frac{TP}{TP+FP} \text{x}100\% \quad (7)$$

Recall adalah metode yang digunakan untuk menghitung persentase data kelas positif yang berhasil terprediksi dengan benar dibandingkan dengan seluruh data kelas positif yang ada (Hidayah & Dodiman, 2024).

$$\textit{Recall} = \frac{TP}{TP+FN} \text{x}100\% \quad (8)$$

F1-score adalah rata-rata harmonik antara presisi dan *recall*. *F1-score* juga dikenal sebagai *F1-measure* (Hidayah & Dodiman, 2024).

$$F1 - score = 2\text{x}\frac{Presisi \text{ x } Recall}{Presisi+Recall} \quad (9)$$