BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek penelitian merupakan aspek yang menjadi pusat kajian dalam sebuah penelitian ilmiah, di mana seluruh proses analisis, perancangan, dan pengujian diarahkan untuk memahami karakteristik dari objek tersebut. Dalam penelitian ini, objek yang diteliti adalah data medis pasien yang digunakan untuk melakukan proses klasifikasi terhadap kemungkinan seseorang menderita diabetes. Dataset yang digunakan bersumber dari *Pima Indians Diabetes Database*, sebuah dataset medis yang telah banyak digunakan secara global untuk pengembangan model pembelajaran mesin di bidang kesehatan.

Dataset ini dikembangkan oleh *National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK)*, lembaga riset medis di bawah naungan pemerintah Amerika Serikat yang secara khusus meneliti berbagai penyakit metabolik dan ginjal. Dataset ini kini tersedia secara terbuka di platform *Kaggle*, sebuah repositori daring yang banyak digunakan oleh penulis, praktisi data, dan akademisi untuk berbagi data serta model analisis berbasis *machine learning*. Dengan ketersediaan dataset yang terbuka, penelitian ini dapat dilakukan secara efisien, terukur, serta memiliki potensi untuk direplikasi oleh penulis lain guna menghasilkan hasil yang konsisten dan dapat diverifikasi.

Dataset *Pima Indians Diabetes* berisi data medis dari 1.200 pasien wanita keturunan suku Pima yang tinggal di wilayah Arizona, Amerika Serikat. Populasi ini dipilih karena memiliki prevalensi diabetes tipe 2 yang tinggi dibandingkan kelompok etnis lainnya. Secara historis, masyarakat Pima memiliki kecenderungan genetik dan gaya hidup tertentu yang meningkatkan risiko diabetes, sehingga menjadikan dataset ini relevan untuk penelitian klasifikasi medis. Data yang terkandung di dalamnya merupakan hasil dari pemeriksaan medis yang terstandarisasi, sehingga dapat digunakan sebagai representasi umum dalam penelitian kesehatan berbasis data.

Data yang terdapat dalam dataset bersifat kuantitatif dan terdiri atas delapan variabel input (fitur independen) serta satu variabel target (dependen). Variabel-variabel tersebut mencakup *Pregnancies* (jumlah kehamilan pasien), *Glucose* (kadar glukosa dalam plasma darah setelah dua jam tes toleransi glukosa), *Blood Pressure* (tekanan darah diastolik), *Skin Thickness* (ketebalan lipatan kulit triceps), *Insulin* (kadar insulin serum), *BMI* (indeks massa tubuh), *Diabetes Pedigree Function* (riwayat keturunan diabetes dalam keluarga), dan *Age* (usia pasien). Sementara itu, variabel target atau *Outcome* bernilai biner dengan kode 1 untuk pasien yang terindikasi menderita diabetes, dan 0 untuk pasien yang tidak menderita diabetes.

Secara umum, dataset ini mencerminkan data medis yang bersifat numerik, sehingga sangat cocok untuk diterapkan dalam model *machine learning* berbasis klasifikasi. Setiap atribut dalam dataset memiliki kontribusi terhadap hasil akhir klasifikasi. Misalnya, kadar glukosa darah dan nilai *BMI* sering dianggap sebagai indikator kuat terhadap risiko diabetes, sementara variabel lain seperti usia dan riwayat keturunan turut memperkuat analisis pola kejadian penyakit tersebut. Keberagaman atribut ini membuat dataset *Pima Indians Diabetes* menjadi objek penelitian yang ideal untuk menguji kinerja algoritma pembelajaran mesin.

Penelitian ini secara khusus memanfaatkan dataset tersebut untuk mengklasifikasikan pasien berdasarkan karakteristik medis menggunakan dua algoritma pembelajaran mesin yang berbeda, yaitu *Artificial Neural Network* (ANN) dan *Decision Tree*. Kedua algoritma ini dipilih karena mewakili dua paradigma pembelajaran yang berbeda namun sama-sama kuat dalam tugas klasifikasi. *Artificial Neural Network* merupakan algoritma berbasis sistem saraf tiruan yang mampu mengenali hubungan non-linear antarvariabel dan beradaptasi terhadap kompleksitas data. Sementara itu, *Decision Tree* merupakan algoritma berbasis aturan (*rule-based algorithm*) yang bekerja dengan membuat struktur keputusan logis dalam bentuk pohon, sehingga hasil klasifikasi dapat dijelaskan dengan mudah dan dipahami oleh pengguna non-teknis, seperti tenaga medis.

Pemilihan kedua algoritma tersebut dilakukan dengan pertimbangan ilmiah bahwa keduanya memiliki karakteristik yang saling melengkapi. *Artificial Neural Network* unggul dalam aspek akurasi dan kemampuan generalisasi terhadap data yang kompleks, namun memiliki kelemahan dalam interpretasi hasil karena bersifat *black box*. Sebaliknya, *Decision Tree* memiliki keunggulan dalam hal interpretabilitas dan kemudahan visualisasi model, namun cenderung lebih rentan terhadap *overfitting* pada dataset yang tidak seimbang. Melalui penelitian komparatif ini, diharapkan dapat diperoleh pemahaman yang lebih mendalam tentang sejauh mana kedua algoritma ini mampu mengklasifikasikan pasien diabetes secara efektif dan efisien [29].

Sumber data yang digunakan dalam penelitian ini bersifat sekunder karena diperoleh dari sumber terbuka yang telah dikumpulkan sebelumnya oleh pihak lain. Meski demikian, data sekunder ini tetap relevan karena berasal dari lembaga resmi dan telah digunakan secara luas dalam berbagai studi internasional. Penggunaan data sekunder memiliki kelebihan dalam efisiensi waktu, biaya, dan sumber daya, namun tetap memerlukan tahap pra-pemrosesan yang ketat. Sebelum digunakan untuk pemodelan, dataset ini harus melalui tahapan *data cleaning* untuk menghapus nilai-nilai yang tidak valid atau kosong (*missing values*), serta normalisasi untuk memastikan semua atribut memiliki skala yang sebanding, khususnya agar model *Artificial Neural Network* dapat berfungsi optimal.

Tahapan penelitian dilakukan berdasarkan metodologi *CRISP-DM (Cross-Industry Standard Process for Data Mining)*, yang terdiri atas enam fase utama yaitu *Problem Understanding, Data Understanding, Data Preparation, Modeling, Evaluation,* dan *Deployment.* Pada tahap awal (*Problem Understanding*), penulis menetapkan tujuan utama penelitian yaitu mengklasifikasikan dan membandingkan performa kedua algoritma terhadap dataset diabetes. Tahap berikutnya (*Data Understanding*) digunakan untuk memahami karakteristik dataset, mengidentifikasi distribusi nilai, serta menentukan atribut yang paling berpengaruh terhadap hasil klasifikasi.

Tahap *Data Preparation* mencakup proses pembersihan, normalisasi, dan pembagian dataset menjadi data latih (*training data*) dan data uji (*testing data*). Setelah itu, tahap *Modeling* dilakukan dengan membangun dua model klasifikasi menggunakan *Python* dalam lingkungan *Jupyter Notebook*. Parameter seperti jumlah neuron, kedalaman pohon, fungsi aktivasi, serta metode optimisasi diuji secara sistematis untuk memperoleh hasil terbaik. Selanjutnya, tahap *Evaluation* digunakan untuk mengukur performa model dengan metrik kuantitatif seperti *accuracy*, *precision*, *recall*, dan *F1-score*.

Secara keseluruhan, penelitian ini bertujuan untuk memberikan kontribusi nyata dalam penerapan algoritma kecerdasan buatan di bidang kesehatan, khususnya dalam analisis klasifikasi penyakit diabetes. Melalui perbandingan kinerja antara *Artificial Neural Network* dan *Decision Tree*, penelitian ini diharapkan mampu menunjukkan algoritma mana yang memiliki performa terbaik dalam mengklasifikasikan pasien berdasarkan data medis yang tersedia.

3.2 Metode Penelitian

Tabel 3.1 Perbandingan Metode Algoritma

Algoritma	Mengapa relevan	Keterkaitan
ANN dan *Decision Tree*	Paling relevan karena penelitian memang bertujuan membandingkan dua algoritma ini dalam satu pipeline CRISP-DM yang sama, untuk melihat *trade-off*, ANN kuat menangkap pola nonlinier, sedangkan Decision Tree kuat pada interpretabilitas dan aturan keputusan.	Selaras dengan fokus CRISP–DM dan replikasi, batasan hanya 2 algoritma, evaluasi pakai akurasi, *precision*, *recall*, *F1*, *confusion matrix*.

Algoritma	Mengapa relevan	Keterkaitan
Logistic Regression	Kurang relevan untuk penelitian ini karena tidak memenuhi dua paradigma yang saling melengkapi seperti yang ditegaskan (nonlinier vs *rule-based*). *Logistic Regression* sangat interpretabel, tapi umumnya berasumsi hubungan lebih linear, sehingga tidak sejalan dengan alasan pemilihan ANN sebagai pembelajar pola nonlinier.	Penelitian sudah menetapkan perbandingan ANN vs Decision Tree sebagai kontribusi dan urgensi (mengisi gap penelitian yang membandingkan keduanya secara konsisten).
SVM	Kurang relevan karena biasanya membutuhkan pemilihan kernel dan pengaturan parameter (*mis. C, gamma*) agar hasil stabil, sementara penelitian ini secara eksplisit tidak mencakup optimasi dan *hyperparameter tuning* tingkat lanjut.	Penelitian menekankan pipeline yang transparan, konsisten, dan replikatif tanpa tuning lanjutan.

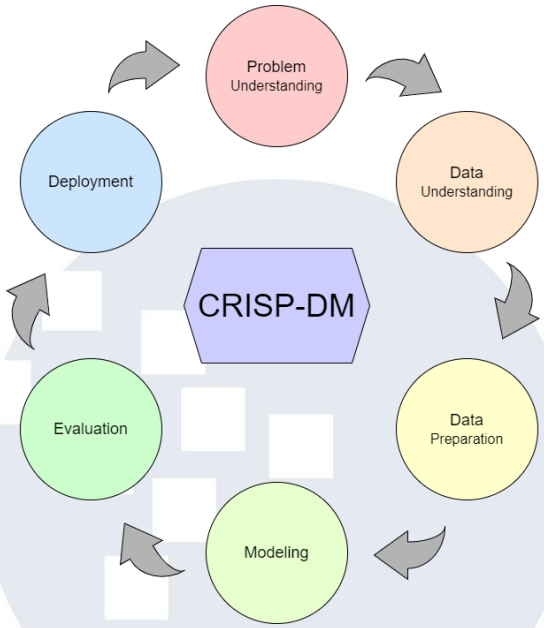
Algoritma	Mengapa relevan	Keterkaitan
K-Nearest Neighbor	Kurang relevan karena k-NN sangat bergantung pada jarak, sehingga sensitif pada skala fitur serta membutuhkan normalisasi ketat, dan tidak menghasilkan aturan keputusan yang mudah diaudit seperti *Decision Tree*. Di penelitian ini, *scaling* memang dibahas sebagai kebutuhan model sensitif skala (ANN), sedangkan *Decision Tree* digunakan karena aturan dan percabangannya mudah ditelusuri.	Rancangan penelitian ini menonjolkan interpretabilitas *Decision Tree* dan kebutuhan *scaling* khusus model sensitif skala (ANN).
Naive Bayes	Kurang relevan karena penelitian ini ingin membandingkan model nonlinier (ANN) vs model *rule-based* yang interpretabel (*Decision Tree*). Naive Bayes biasanya dipilih untuk kesederhanaan asumsi probabilistik, bukan untuk eksplorasi *trade-off* nonlinier vs aturan keputusan yang jadi narasi utama penelitian.	Fokus dan novelty penelitian ini diposisikan pada perbandingan ANN dan *Decision Tree* secara langsung dalam satu kerangka CRISP–DM.

Algoritma	Mengapa relevan	Keterkaitan
Random Forest	Kurang relevan untuk cakupan penelitian ini karena meski sering meningkatkan performa dibanding 1 pohon, *Random Forest* mengurangi keterlacakan aturan keputusan menjadi satu pohon yang mudah dijelaskan (interpretabilitas jadi lebih kompleks). Penelitian ini menekankan *Decision Tree* karena keterbacaan dan aturan keputusan.	Algoritma seperti *Random Forest* disarankan untuk penelitian lanjutan (artinya bukan fokus penelitian saat ini).
Gradient Boosting	Kurang relevan karena biasanya sangat bergantung pada pengaturan parameter dan strategi validasi agar stabil, sedangkan penelitian ini menyatakan tidak melakukan *hyperparameter tuning* tingkat lanjut.	Penelitian ini sendiri menempatkan *XGBoost* sebagai rekomendasi pengembangan, bukan inti kontribusi penelitian.

Penelitian ini menggunakan pendekatan kuantitatif berbasis eksperimen komputasional pada data sekunder untuk membandingkan dua metode klasifikasi *Artificial Neural Network* (ANN) dan *Decision Tree* dalam memprediksi status diabetes. Kerangka kerja mengikuti tahapan baku *data mining* yang terstruktur dan berulang agar proses dapat diaudit serta direplikasi. Dataset yang digunakan adalah turunan *Pima Indians Diabetes* yaitu laman Kaggle yang dipakai merupakan *mirror/hosted copy*, sedangkan pemilik/pengelola awal dataset adalah *National*

Institute of Diabetes and Digestive and Kidney Diseases (*NIDDK*) dan telah lama didistribusikan melalui *UCI Machine Learning Repository*.

3.2.1 Alur Penelitian



Gambar 3.1 Alur Penelitian *CRISP-DM*

Gambar 3.1 merupakan alur penelitian secara menyeluruh sejak perumusan tujuan sampai pelaporan hasil. Diagram ini disusun mengikuti kerangka *CRISP-DM* sehingga tahapan di bagian tengah bersifat iteratif, penulis dapat kembali ke pra-pemrosesan bila evaluasi awal menunjukkan perlunya perbaikan. Dengan membaca alur dari kiri ke kanan, mendapatkan gambaran ringkas tentang bagaimana data sekunder dipersiapkan, model dibangun dan divalidasi, kinerja dibandingkan secara objektif, lalu temuan dirangkum menjadi rekomendasi yang selaras dengan batasan penelitian.

1) Problem Understanding

   Tahap ini menetapkan tujuan penelitian, ruang lingkup, kriteria keberhasilan, serta batasan. Tujuan utamanya yaitu membandingkan ketepatan klasifikasi dua metode (ANN dan *Decision Tree*) untuk prediksi diabetes pada data sekunder. Di sini juga didefinisikan metrik penilaian yang dianggap bermakna (akurasi, *precision*, *recall*, *F1*), asumsi dasar data (populasi Pima, skala variabel), serta risiko yang mungkin memengaruhi

hasil (ketidakseimbangan kelas, nilai nol yang tidak bermakna klinis, dan keterbatasan generalisasi). Luaran tahap ini berupa pernyataan masalah yang jelas, indikator keberhasilan yang terukur, dan rencana kerja ringkas.

2) Data Understanding

Fokusnya memahami karakter data sebelum diolah. Kegiatan meliputi peninjauan struktur dan isi variabel, ringkasan statistik (nilai pusat dan sebaran), pola *missing values*, keberadaan nilai ekstrem, serta komposisi kelas target. Visual yang umum dibuat antara lain histogram, *boxplot*, dan peta *missingness*. Tujuannya adalah memotret "kondisi awal" data agar keputusan pada tahap berikutnya (seperti perlakuan missing atau pencilan) berbasis bukti, bukan asumsi.

3) Data Preparation

Tahap ini menyiapkan data agar layak dianalisis. Kegiatan tipikal yaitu pertama menetapkan aturan kebersihan data contoh, nilai nol pada variabel fisiologis tertentu diperlakukan sebagai missing. Kedua memilih strategi pengisian nilai hilang yang konservatif agar tidak mengubah pola asli data. Ketiga menangani pencilan secara hati-hati (berbasis statistik deskriptif) agar tidak menghapus informasi penting. Keempat menyelaraskan skala antarf fitur bila diperlukan. Terakhir memisahkan data untuk pelatihan dan pengujian secara terstratifikasi agar representasi kelas tetap proporsional. Hasil tahap ini adalah himpunan data "siap model" yang terdokumentasi aturan transformasinya.

4) Modeling

Pada tahap ini dibangun dan disetel dua model klasifikasi (ANN dan *Decision Tree*) menggunakan data yang telah disiapkan. Prinsipnya, tiap model disusun dengan pengaturan yang wajar menurut literatur, lalu dites dengan rancangan validasi yang objektif untuk menilai kestabilan performa. Tujuan utamanya bukan sekadar memperoleh angka tinggi, tetapi memastikan proses penyusunan model transparan, berulang, dan dapat direplikasi pihak lain.

5) Evaluation

Hasil model dievaluasi menggunakan metrik yang relevan (akurasi, *precision*, *recall*, *F1*) dilengkapi *confusion matrix* untuk melihat kesalahan tipe apa yang dominan. Idealnya, disajikan pula rentang ketidakpastian (*confidence interval*) agar mudah memahami ketepatan estimasi, bukan angka titik semata. Karena penelitian ini membandingkan dua model, dilakukan uji perbandingan yang objektif pada data pengujian agar dapat disimpulkan apakah perbedaan performa benar-benar berarti secara statistik dan praktis. Pada tahap ini juga dibahas keterbatasan (misalnya populasi Pima yang spesifik) dan implikasi temuan.

6) Deployment (opsional)

Tahap ini berfokus pada dokumentasi hasil, pembuatan visualisasi performa model, serta interpretasi hasil klasifikasi. Hasil akhir dapat dijadikan acuan bagi penelitian lanjutan atau pengembangan sistem diagnosis berbasis machine learning di bidang kesehatan. Namun hal ini opsional dikarenakan penelitian ini menggunakan data sekunder anonim Pima Indians Diabetes, yang tidak berasal dari lingkungan operasional nyata seperti rumah sakit atau klinik.

Tabel 3.2 Perbandingan Metode

Aspek	CRISP-DM	KDD	SEMMA	TDSP
Fokus Utama	Proses end-to-end	Penemuan pola data	Teknik statistik dan modeling	Proyek machine learning berbasis tim
Tahapan Utama	6 fase: Problem Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, Deployment	Selection, Preprocessing, Transformation, Data Mining, Interpretation	Sample, Explore, Modify, Model, Assess	Business Understanding, Data Acquisition, Modeling, Deployment, Customer Acceptance

Aspek	CRISP-DM	KDD	SEMMA	TDSP
Pendekatan	Iteratif, fleksibel, universal	Lebih fokus pada eksplorasi data dan data mining	Teknis, berorientasi modelling	Agile, kolaboratif, integrasi engineering
Cocok untuk	Semua proyek data science / analitik	Penelitian data mining	Pengguna SAS, analisis statistik	ML production, data engineering
Keterlibatan penelitian	Sangat kuat	Sedang	Lemah	Kuat
Kelengkapan siklus proyek	Paling lengkap sampai deployment	Kurang lengkap pada sisi deployment	Fokus modeling, deployment tidak dibahas	Lengkap, modern, untuk tim
Aksesbilitas	Bebas, open, umum	Konseptual & akademik	Khusus pengguna SAS	Framework Microsoft

Ringkasnya, CRISP-DM memastikan alur penelitian tidak hanya menghasilkan angka performa, tetapi juga jejak proses yang jelas yaitu tujuan yang terukur, data yang dipahami dan dipersiapkan dengan benar, pemodelan yang objektif, evaluasi yang bermakna, dan pelaporan yang dapat dipertanggungjawabkan.

3.2.2 Data Mining

Bagian ini menjelaskan kerangka dan prosedur data mining yang digunakan untuk membangun serta membandingkan dua metode klasifikasi pada data sekunder Pima Indians Diabetes. Bertujuan Penetapan proses analitis yang terstruktur, transparan, dan dapat direplikasi (*reproducible*) agar diperoleh perbandingan yang objektif antara dua model klasifikasi beserta luaran yang bermakna secara akademik maupun praktis.

1) Perumusan Tujuan

   Menetapkan sasaran komparasi, indikator kinerja yang relevan, asumsi, batasan, serta risiko metodologis yang perlu diantisipasi.

2) Data understanding

   Meninjau struktur data, ringkasan statistik setiap variabel, sebaran dan kemiringan, keberadaan missing values dan nilai nol yang tidak bermakna klinis, serta keseimbangan kelas target.

3) Data preparation

   Menetapkan aturan kebersihan data (perlakuan nilai nol tertentu sebagai *missing*), memilih strategi pengisian nilai hilang yang konservatif, menangani pencilan berdasarkan pertimbangan statistik dan substansi, menyelaraskan skala antarf fitur bila diperlukan, serta melakukan pemisahan data pelatihan dan data pengujian secara terstratifikasi untuk menjaga proporsi kelas.

4) Modeling

   Menyusun dua model klasifikasi (ANN dan *Decision Tree*) dengan pengaturan yang wajar menurut praktik literatur, memastikan kesetaraan syarat dalam perbandingan, dan menyiapkan skema penyetelan seperlunya tanpa melampaui batas ruang lingkup penelitian.

5) Validation

   Menerapkan *stratified k-fold cross-validation* pada data latih guna menilai kestabilan kinerja dan meminimalkan overfitting, sekaligus mencatat konfigurasi yang dipilih agar proses dapat diaudit ulang.

6) Evaluation

   Melaporkan kinerja pada data uji menggunakan metrik yang relevan (misalnya akurasi, *precision*, *recall*, F1) disertai *confusion matrix* dan *confidence interval* untuk menekankan ketidakpastian estimasi, bukan sekadar angka titik.

7) Comparison and inference

   Mengujikan perbedaan kinerja kedua model pada data uji agar simpulan komparatif sahih secara statistik dan tidak bergantung pada kebetulan pemisahan data.

8) Reporting

Menyajikan tabel, gambar, dan narasi interpretatif yang menghubungkan temuan dengan konteks data, batasan, serta implikasi praktis–ilmiah diakhiri dengan rekomendasi penggunaan model sesuai kekuatan dan keterbatasannya.

Kualitas data dan integritas proses, Penelitian menjaga konsistensi definisi missing, menerapkan strategi pengisian yang tidak agresif agar pola asli data tetap terpelihara, menangani pencilan secara hati-hati agar informasi penting tidak hilang, serta mendokumentasikan setiap keputusan pra-olah untuk memastikan auditability.

Validasi dan keadilan perbandingan, Proporsi kelas dijaga melalui *stratified spli*t pada data pelatihan dan data penujian yaitu *cross-validation* digunakan untuk menilai kestabilan, dan aturan evaluasi (metrik, prosedur pemilihan, serta pelaporan) diberlakukan sama untuk semua model agar perbandingan objektif.

Evaluasi dan pelaporan hasil, Hasil disajikan dalam bentuk metrik utama per model, *confusion matrix*, serta kurva performa (misalnya ROC/PR) untuk memperlihatkan *trade-off* keputusan. Ringkasan perbandingan menyoroti keunggulan dan kelemahan relatif masing-masing model serta konsekuensi praktisnya terhadap penggunaan di konteks serupa.

Etika, keterbatasan, dan r*eproducibility*, Data yang digunakan bersifat anonim dan dipergunakan untuk tujuan akademik. Keterbatasan generalisasi dicatat mengingat karakter populasi yang spesifik. *Reproducibility* ditegakkan melalui pencatatan langkah, aturan pra-olah, dan skema evaluasi sehingga proses dapat diulang dan diverifikasi pada penelitian lanjutan.

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan *secondary data* berbentuk tabel mengenai diabetes dari populasi Pima yang bersifat anonim. Pengumpulan data dilakukan dengan menelusuri sumber yang kredibel, memverifikasi asal-usul dan deskripsi variabel, mengunduh berkas data beserta penjelasannya, serta mencatat metadata seperti tanggal akses, ukuran berkas, jumlah baris, dan jumlah kolom. Berkas kemudian

disimpan secara terstruktur dengan pemisahan data mentah dan data hasil pra-olah serta dibuat cadangan untuk menjaga keberulangan dan *audit trail*. Kesesuaian nama, urutan, dan tipe variabel diperiksa terhadap dokumentasi sumber yaitu nilai nol yang tidak lazim secara klinis ditandai sebagai *missing values* dan nilai ekstrem diidentifikasi untuk ditangani pada tahap pra-olah. Kriteria inklusi menekankan kelengkapan variabel inti sesuai skema asli, sedangkan baris yang tidak memenuhi kelengkapan minimum dikecualikan. Seluruh proses mengikuti etika akademik, tidak melakukan *re-identification*, dan mematuhi ketentuan penggunaan data. Luaran tahap ini berupa berkas data mentah, ringkasan *data dictionary*, catatan log pengumpulan, dan struktur penyimpanan yang rapi untuk mendukung analisis pada bab berikutnya.

3.4 Variabel Penelitian

3.4.1 Variable Dependen

Variabel dependen dalam penelitian ini adalah *Outcome* (status diabetes), yakni variabel biner yang menunjukkan apakah seorang responden terindikasi diabetes atau tidak. Nilai 1 merepresentasikan kondisi terindikasi diabetes, sedangkan nilai 0 menunjukkan tidak terindikasi. Variabel ini menjadi sasaran prediksi yang ingin dijelaskan oleh kombinasi variabel-variabel penjelas pada data yang digunakan.

3.4.2 Variabel Independen

Variabel independen terdiri dari delapan pengukuran klinis numerik yang umum digunakan untuk memotret kondisi kardiometabolik responden. *Pregnancies* menggambarkan jumlah kehamilan sebagai bagian dari riwayat reproduksi. *Glucose* mencerminkan kadar glukosa plasma dua jam dan merupakan indikator utama kontrol gula darah. *BloodPressure* menunjukkan tekanan darah diastolik yang berkaitan dengan kesehatan kardiometabolik. *SkinThickness* merepresentasikan ketebalan lipatan kulit triseps sebagai proksi lemak subkutan. *Insulin* menunjukkan kadar insulin serum yang terkait dengan regulasi gula darah dan potensi resistensi insulin. *BMI* adalah indeks massa tubuh yang menggambarkan status berat badan relatif. *Diabetes Pedigree*

Function memberikan skor kecenderungan genetik terhadap diabetes. *Age* menunjukkan usia responden. Secara umum, variabel-variabel ini dipilih karena menurut pertimbangan klinis dan literatur, perubahan pada indikator-indikator tersebut berkaitan dengan kemungkinan terjadinya diabetes. Dalam pengolahan data, nilai nol yang tidak realistis secara klinis pada variabel fisiologis diperlakukan sebagai *missing* dan ditangani pada tahap pra-pemrosesan agar hasil analisis tetap andal.

3.5 Teknik Analisis Data

3.5.1 CRISP-DM

Analisis mengikuti kerangka *CRISP–DM* yang berisi enam tahap berulang. Pertama, *problem understanding* yaitu merumuskan tujuan, ruang lingkup, indikator keberhasilan, serta batasan penelitian. Kedua, *data understanding* yaitu menelaah struktur dan karakter data melalui ringkasan statistik, pemeriksaan *missing values*, nilai nol yang tidak bermakna klinis, sebaran variabel, dan keseimbangan kelas. Ketiga, *data preparation* yaitu menetapkan aturan kebersihan data, menandai serta mengisi *missing values* secara konservatif, menangani pencilan dengan hati-hati, menyelaraskan skala fitur bila diperlukan, serta memisahkan data latih dan uji secara terstratifikasi agar proporsi kelas terjaga [30]. Keempat, *modeling* yaitu menyusun dua metode klasifikasi sesuai tujuan penelitian dengan prasyarat evaluasi yang setara sehingga perbandingan konsisten. Kelima, *evaluation* yaitu menilai kinerja pada data uji menggunakan metrik yang relevan (akurasi, *precision*, *recall*, F1) disertai pembacaan *confusion matrix* dan pelaporan ketidakpastian hasil. Keenam, *deployment/reporting* yaitu menyajikan tabel, gambar, dan narasi interpretatif yang menautkan temuan dengan konteks data, keterbatasan, dan implikasi praktis, serta mendokumentasikan langkah agar proses dapat direplikasi.

3.5.2 Phyton

Seluruh pengolahan dilakukan dalam lingkungan kerja *Python* untuk menjamin konsistensi, keterlacakan, dan *reproducibility*. Pengaturan meliputi penetapan versi perangkat lunak, pencatatan *random seed*, serta penataan berkas

yang terstruktur data mentah dan data hasil pra-olah disimpan terpisah, sementara keluaran berupa tabel metrik, *confusion matrix*, dan kurva performa disimpan pada direktori hasil. Alur kerja dimulai dari pembacaan data dan pencatatan metadata, dilanjutkan analisis deskriptif, penerapan *pipeline* pra-pemrosesan sesuai aturan yang telah ditetapkan, penyusunan serta evaluasi model dengan skema validasi yang konsisten, kemudian ekspor hasil beserta *log* proses. Setiap keputusan metodologis mulai dari definisi *missing*, penanganan pencilan, pemisahan data pelatihan dan data pengujian, hingga pemilihan metrik didokumentasikan agar seluruh langkah dapat diulang dan diaudit tanpa perlu menampilkan rincian pemrograman pada bagian ini [31].