BAB III
METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Pada penelitian ini objek yang akan diteliti merupakan penyakit *stroke* yang ditandai oleh penyempitan atau pemblokiran pembuluh darah yang dapat memicu *stroke*. Sebagai kondisi medis kritis, *stroke* terjadi ketika aliran darah ke otak terhenti atau berkurang drastis, mengakibatkan kematian sel otak akibat kekurangan oksigen dan nutrisi. Penyakit ini dapat menyerang siapa saja tanpa memandang usia, jenis kelamin, atau ras, dengan faktor risiko seperti hipertensi, diabetes, merokok, obesitas, dan riwayat keluarga meningkatkan kemungkinan terjadinya *stroke*. Penyakit ini bisa terjadi secara mendadak, menuntut respons medis yang cepat untuk meminimalisir kerusakan dan memaksimalkan pemulihan, menjadikan edukasi dan pencegahan dini sebagai elemen kunci dalam mengurangi insiden *stroke*[36].

Penyakit *stroke* merupakan isu kesehatan global yang berdampak pada jutaan orang di seluruh dunia, tanpa dibatasi oleh geografi, namun dengan variasi prevalensi berdasarkan faktor genetik, lingkungan, dan sosial-ekonomi. Pentingnya penelitian ini terletak pada implikasinya yang signifikan terhadap individu dan sistem kesehatan, *stroke* dapat mengakibatkan kerusakan dari ringan hingga permanen, mempengaruhi kemampuan individu dalam berbicara, berjalan, dan menjalankan aktivitas sehari-hari. Melalui pemahaman mendalam tentang faktor risiko dan mekanisme penyakit, strategi pencegahan dan pengobatan dapat ditargetkan secara lebih efektif. Analisa risiko penyakit *stroke* mencakup identifikasi faktor risiko dan penerapan strategi pencegahan dan intervensi, termasuk adopsi gaya hidup sehat, pengelolaan kondisi medis, serta edukasi

masyarakat mengenai tanda dan gejala *stroke*. Pemanfaatan teknologi diagnostik dan terapeutik terkini juga sangat vital dalam diagnosis dan manajemen *stroke* di lingkungan klinis [37].

3.2 Metode Penelitian

Metode yang digunakan pada penelitian ini adalah kuantitatif. Metode kuantitatif merupakan metode yang menggunakan data numerik atau kuantitatif untuk menjawab pertanyaan penelitian. Metode kuantitatif dapat memuat sampel yang besar sehingga dapat membuat analisis statistik yang kuat serta efektif dalam menguji efektivitas dari algoritma yang ingin diuji [38].

3.2.1 Metode Data Mining

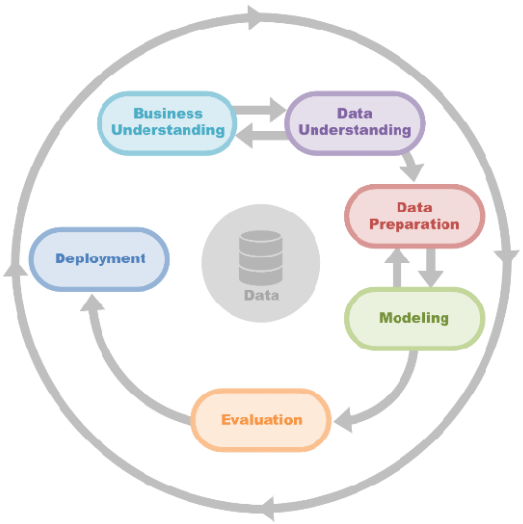
Pada penelitian ini metode penelitian yang akan diterapkan pada penelitian ini di antaranya adalah *Cross-Industri Standard* Process *for Data Mining* (*CRISP-DM*) [39]. Perbandingan metode *CRISP-DM*, *KDD* dan *SEMMA* [40] akan dijelaskan secara lebih lengkap pada tabel 3.1.

Tabel 3. 1 Kelebihan dan batasan *CRISP-DM*

Model Proses	Langkah-langkah	Fokus	Pendekatan Iteratif	Keterbatasan Utama
KDD	*Data Selection and Sampling, Data Processing, Data Transformation, Data Mining, Evaluation*	*Knowledge Generation*	Tidak ada iterasi yang jelas; lebih berorientasi pada tahap data	Kurangnya fokus pada aspek bisnis; Tidak ada fase implementasi atau validasi hasil
SEMMA	*Sample, Explore, Modify, Model, Assess*	*Model Development*	Tidak terdapat fase pemahaman bisnis; Tidak ada definisi eksplisit untuk fase implementasi	Tidak ada bukti yang jelas tentang iterasi antara tugas-tugas
CRISP-DM	*Business Understanding, Data Understanding, Data Preparation, Modeling,*	Penambangan Data	Terdapat iterasi yang diakui; Mendukung fase siklus hidup proyek yang lebih linier	Tidak mempertimbangkan aspek SDM; Kurangnya fase perbaikan yang berkelanjutan

	Evaluation, *Deployment*

Penelitian ini akan menggunakan metode *Cross Industry Standard Process for Data Mining* (*CRISP-DM*) sebagai model alur penelitian dalam melakukan karena menurut perbandingan dinilai cocok untuk klasifikasi menggunakan algoritma berbasis pohon keputusan yaitu *Decision Tree*, *Extra Tree Classifier*, dan *Xgboost* dengan seleksi fitur berbasis algoritma *swarm intelligence* yaitu *ACO* dan *PSO* pada algoritma berbasis pohon keputusan untuk mencari hasil akurasi, presisi, sensitivitas serta waktu pemrosesan untuk mengklasifikasikan penyakit *stroke*. Proses dari *CRISP-DM* [41] dapat dilihat dari Gambar 3.1.



Gambar 3. 1 *Flowchart CRISP-DM*

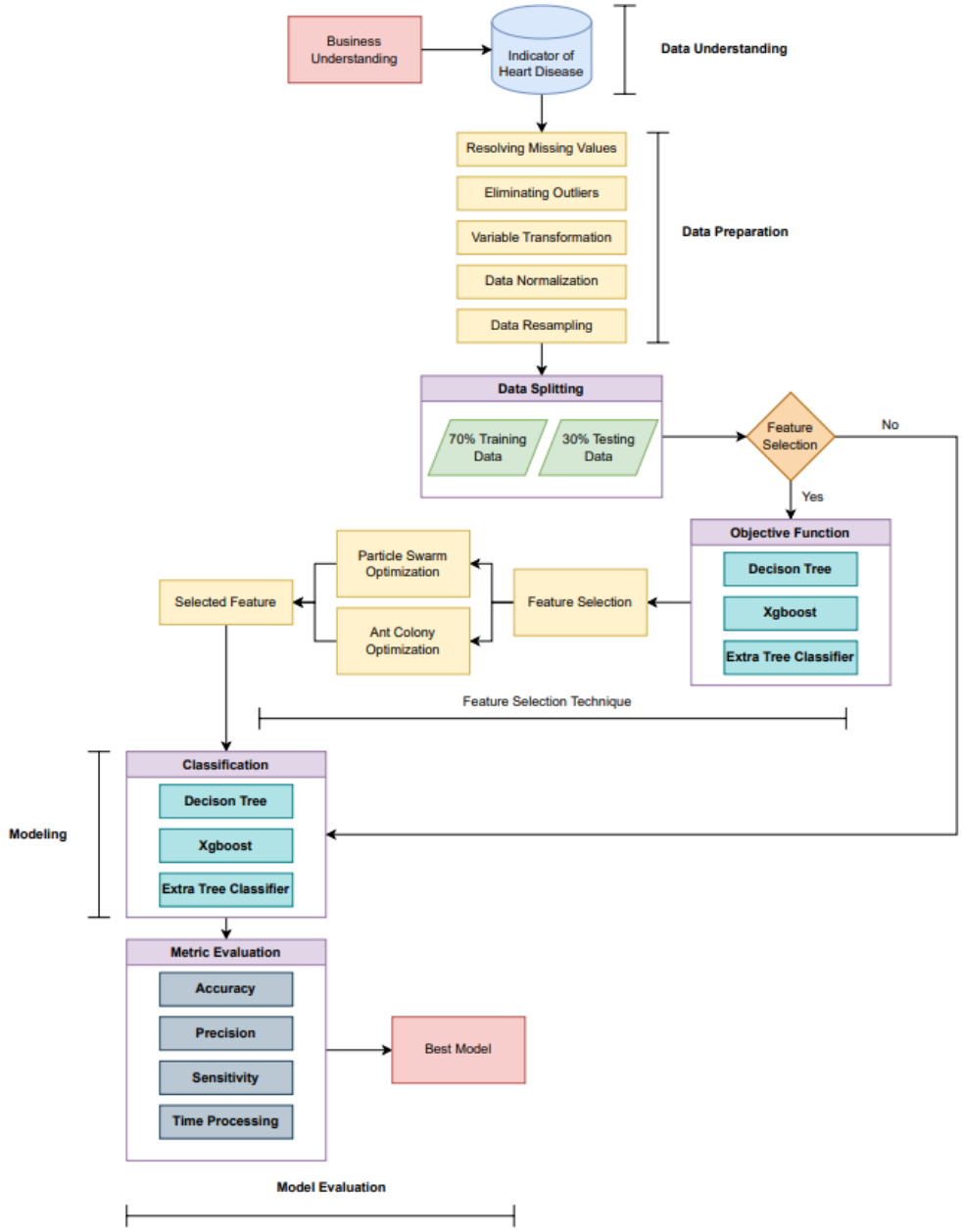
Sumber: Tounsi (2020)

Metode *CRISP-DM* dinilai cocok untuk penelitian ini karena metode *data mining CRISP-DM* memiliki siklus hidup proyek yang linier dengan beberapa aspek sebagai berikut:

1) Tahap *Business Understanding* merupakan tahap awal yang memastikan bahwa tujuan penelitian yang berfokus pada peningkatan akurasi, presisi, sensitivitas, dan efisiensi waktu dalam klasifikasi penyakit stroke sangat relevan. Tahap ini membantu untuk memahami sasaran dan pertanyaan penelitian yang akan dijawab dalam melalui *data mining*.
2) Tahap *Data Understanding* merupakan tahap kedua yang melibatkan eksplorasi data lebih awal untuk mendapatkan wawasan mengenai data yang digunakan, hal ini penting terutama dalam klasifikasi penyakit.
3) Tahap *Data Preparation* merupakan tahap ketiga dan merupakan tahap untuk membersihkan dan memperbaiki data yang sudah dikumpulkan, termasuk penanganan data yang hilang, *outlier*, dan transformasi data jika diperlukan. Tahap ini diperlukan untuk meningkatkan hasil performa yang diinginkan dari algoritma berbasis keputusan yang sensitif terhadap kualitas data.
4) Tahap *Modeling* merupakan tahap keempat dan merupakan tahap yang cocok untuk penelitian ini untuk mendapatkan hasil akurasi, presisi, sensitivitas dan efisiensi waktu dari penggunaan algoritma berbasis pohon keputusan dengan dan tanpa seleksi fitur menggunakan algoritma *PSO* dan *ACO*
5) Tahap Evaluasi merupakan tahap kelima yang merupakan tahap untuk mengetahui seberapa baik hasil *modeling* yang telah dilakukan. Tahap ini cocok karena tujuan penelitian ini adalah membandingkan hasil klasifikasi *stroke* dan mengetahui yang terbaik pada algoritma berbasis pohon keputusan dengan dan tanpa menggunakan algoritma *PSO* dan *ACO*.

3.2.2 Alur Penelitian

Alur Penelitian yang digunakan pada penelitian ini memanfaatkan kerangka kerja *CRISP-DM* untuk mengatur langkah mulai dari *Business Understanding* sampai *Development*. Alur penelitian bisa dilihat pada Gambar 3.2.



Gambar 3. 2 Alur Penelitian

3.2.2.1 Business Understanding

Tahapan *Business understanding* merupakan tahapan awal dari metode *CRISP-DM*. Tahapan *Business understanding* merupakan tahap yang berfokus kepada tujuan dan hal-hal yang diperlukan untuk penelitian yang akan dilakukan. Pengetahuan mengenai hal-hal yang diperlukan dan tujuan akan dibuat menjadi dasar dalam mendefinisikan masalah utama dalam rencana penelitian dan menjadi panduan untuk dapat mencapai tujuan dari penelitian [42]. Pada penelitian kali ini *Business Perspective*-nya adalah melakukan analisa tingkat risiko seseorang untuk terkena *stroke* mengingat tingginya angka kematian yang disebabkan penyakit *stroke* tiap tahunnya. Penelitian ini diharapkan dapat membawa pengetahuan untuk menjaga kesehatan dan menjauhi fitur-fitur yang dapat meningkatkan risiko seseorang untuk terkena penyakit *stroke*.

3.2.2.2 Data Understanding

Tahapan kedua yang akan dilakukan pada metode *CRISP-DM* adalah *Data Understanding*. Tahapan *data understanding* merupakan tahap yang berfokus kepada pengumpulan *dataset* yang akan digunakan dan melakukan eksplorasi terhadap variabel-variabel yang akan digunakan pada *dataset* yang telah dikumpulkan. *Dataset* yang telah dikumpulkan perlu diidentifikasi kualitas dari data dengan menemukan potensi informasi tersembunyi yang ada pada data [42]. Pada penelitian kali ini *dataset* yang digunakan yaitu *indicator of heart disease Dataset* (2022) dengan 40 variabel berbeda.

3.2.2.3 Data Preparation

Tahapan ketiga yang akan dilakukan pada metode *CRISP-DM* adalah tahapan *data preparation* yaitu merupakan tahapan data mentah yang didapat diubah menjadi informasi yang lebih informatif sehingga dapat diproses lebih lanjut lewat proses *data cleansing* dan *data splitting*.

1) *Data Cleansing*

   Data cleansing memiliki tujuan untuk menjaga kualitas dari data yang akan digunakan untuk di analisa. Tahapan yang akan dilakukan pada *data cleansing* adalah menemukan dan memperbaiki *missing value* serta *noisy data*. Proses *data cleansing* akan membuat hasil analisa data akan lebih terpercaya, efektif, dan efisien [43]. Tahap *data cleansing* juga meliputi proses pembuangan *outlier* yang fungsinya untuk meningkatkan akurasi model menjadi lebih akurat karena model dapat lebih baik dalam mencerminkan pola sebenarnya, mengurangi *overfitting*, memperbaiki kualitas data, meningkatkan konsistensi hasil yang diterima dan mempercepat waktu pemrosesan karena menghilangkan bias [44]. Pembuangan *outlier* akan meningkatkan efektivitas dan efisiensi dari kualitas seleksi fitur dan hasil klasifikasi *stroke*.

2) *Data splitting*

   Tahapan *data splitting* yang akan dilakukan adalah membagi *dataset indicator of heart disease* yang telah didapat menjadi dua bagian yaitu *data training* dan *data testing* dengan *range* yang akan disesuaikan dengan kebutuhan dan tujuan dari penelitian. Pada tahap ini juga akan dilakukan pembagian kembali dari *data training* untuk dibagi menjadi *data train* dan *data validation* [45]. Rasio yang digunakan untuk *data splitting*

adalah 70:30, rasio dipilih berdasarkan penelitian [16] yang menggunakan beberapa rasio *data splitting* untuk memprediksi kanker payudara. Dataset yang digunakan adalah *Wisconsin Breast Cancer* dengan empat algoritma klasifikasi yaitu *SVM, Logistic Regression, Decision Forest, dan Neural Network.* Hasilnya algoritma *DF, LR,* dan *NN* memiliki akurasi di atas 98%. Pembagian rasio 70:30 memiliki hasil akurasi yang bagus untuk *Decision Forest* yang merupakan algoritma berbasis pohon keputusan, hal ini sejalan dengan penggunaan algoritma berbasis pohon keputusan yang digunakan pada penelitian ini.

3.2.2.4 Modeling

Pada tahapan *modeling* yang akan dilakukan adalah pemilihan dan pengembangan teknik analisa serta algoritma yang akan dipilih dan digunakan untuk penelitian ini [42]. Pada penelitian ini algoritma yang akan dipakai adalah algoritma berbasis pohon keputusan yaitu *Decision Tree, Xgboost,* dan *Extra tree classifier* dengan dua algoritma *Swarm Intelligence* yaitu *Particle Swarm Optimization,* dan *Ant Colony Optimization*. Pada penelitian ini akan dilakukan *feature selection* yang digunakan dengan tujuan untuk mengurangi kompleksitas algoritma klasifikasi, meningkatkan akurasi klasifikasi dan mengetahui fitur penting yang berpengaruh terhadap tingkat akurasi pada algoritma berbasis pohon keputusan.

a) Membaca *indicator of heart disease* (2022) yang masih mentah.

b) *Data Cleansing,* mempersiapkan data mentah dengan atribut yang nantinya akan digunakan pada saat modeling.

c) Membagi data dengan *data training* sebanyak 70% dan *data testing* sebanyak 30%.

d) Klasifikasi menggunakan algoritma berbasis pohon keputusan yaitu *Decision Tree, Xgboost,* dan *Extra Tree Classifier*.

e) Melakukan seleksi fitur dan mencari fitur terbaik dari *dataset* menggunakan *Particle swarm optimization* dan *Ant Colony Optimization* pada algoritma berbasis pohon keputusan.

f) Hasil Fitur terbaik dari *Particle swarm optimization* dan *Ant Colony Optimization* digunakan untuk melatih dan menguji algoritma berbasis pohon keputusan.

g) Evaluasi performa masing-masing algoritma optimasi pada algoritma berbasis pohon keputusan dengan dan tanpa seleksi fitur.

Penelitian ini menggunakan algoritma berbasis pohon keputusan yaitu *Decision Tree, Extra Tree,* dan *Xgboost*. Algoritma *Decision Tree* memiliki kinerja yang terbukti baik dengan penelitian [6], [7], dan [11] menunjukkan bahwa algoritma *Decision Tree* dapat menghasilkan akurasi yang tinggi dan memiliki kinerja yang baik dalam klasifikasi. Penelitian [6] membandingkan efektivitas *10-cross fold* dan metode *data splitting* pada kasus hujan di Australia menggunakan algoritma *Decision Tree*, hasilnya metode klasifikasi *10-Cross Fold Validation* memberikan akurasi 87.35% dan metode *data splitting* rasio 80:20 memberikan akurasi 85.37%. Penelitian [7] membuktikan bahwa seleksi fitur dengan algoritma *Particle Swarm Optimization (PSO)* pada algoritma *Decision Tree* dapat mempengaruhi efektivitas dan efisiensi waktu, dengan *Decision Tree* akurasi 85.24% dengan waktu pemrosesan

0.11 detik, *Decision Tree* dengan seleksi fitur menggunakan *PSO* akurasi 83.61%, waktu pemrosesan 0.02 detik. Algoritma *Decision Tree* untuk ditingkatkan lebih lanjut dengan seleksi fitur menggunakan *PSO* yang ditunjukkan penelitian [12] menjadi alasan penelitian ini menggunakan algoritma *Decision Tree* karena menunjukkan PSO dapat meningkatkan efektivitas dengan meningkatkan semua aspek kinerja metrik model *Decision Tree* dengan peningkatan akurasi dari 93.50% menjadi 94.56%, presisi dari 94.97% menjadi 96.25%, dan sensitivitas dari 94.62% menjadi 95.01%. Algoritma *Xgboost* memiliki kinerja yang baik dengan kemampuannya menangani *dataset* yang besar, menghindari *overfitting* dan memberikan hasil yang akurat. Penelitian [14] melakukan klasifikasi kanker payudara menggunakan *Xgboost*, hasilnya mendapatkan akurasi sampai 97% dan [15] dengan topik yang sama menunjukkan bahwa algoritma *Xgboost* memiliki akurasi yang tinggi dalam klasifikasi penyakit seperti kanker payudara dengan akurasi mencapai 94.74%. Penelitian [8] menggabungkan metode *CLAHE*, segmentasi *K-means*, dan seleksi fitur menggunakan *PSO* pada *Xgboost*, mencapai akurasi 97%, presisi 98%, dan sensitivitas 97% dan Penelitian [9] menunjukkan bahwa model dengan seleksi fitur menggunakan *ACO* pada *Xgboost* memberikan hasil akurasi tinggi dan efisiensi dalam identifikasi emitor spesifik, dengan akurasi 98.90% pada dataset original dan mendapatkan peningkatan *ACO-Xgboost* dapat meningkatkan akurasi sebesar 0.20% hingga 3.53% dibandingkan dengan algoritma lain. Algoritma *Extra Tree* dipilih karena kinerja baiknya dalam menangani *dataset* yang besar dan kemampuannya dalam menghasilkan akurasi yang baik. Penelitian [10] menunjukkan bahwa *Extra Tree classifier* memberikan kinerja

yang sangat baik dalam mengklasifikasikan gambar medis dengan akurasi 92.13%, presisi 91.85%, dan sensitivitas 90.96%.

Penelitian ini menggunakan algoritma *PSO* dan *ACO* sebagai algoritma optimasi untuk melakukan seleksi fitur terhadap algoritma berbasis pohon keputusan. Algoritma *PSO* telah terbukti baik dapat meningkatkan akurasi dan efisiensi dalam berbagai studi termasuk penelitian [7] yang membuktikan bahwa seleksi fitur menggunakan *PSO* dapat meningkatkan efisiensi waktu dengan hasil akurasi 85.24% dengan waktu pemrosesan 0.11 detik, *Decision Tree* dengan seleksi fitur menggunakan *PSO* akurasi 83.61%, waktu pemrosesan 0.02 detik dan penelitian [12] menunjukkan *PSO* dapat meningkatkan efektivitas dengan meningkatkan semua aspek kinerja metrik model *Decision Tree* dengan peningkatan akurasi, presisi dan sensitivitas dengan hasil peningkatan akurasi dari 93.50% menjadi 94.56%, presisi dari 94.97% menjadi 96.25%, dan sensitivitas dari 94.62% menjadi 95.01%. Hal ini menunjukkan bahwa *PSO* dapat meningkatkan kemampuan model dalam mengidentifikasi kelas yang benar dan lebih akurat. Algoritma *ACO* juga menunjukkan efisiensi dan performa klasifikasi yang tinggi. Penelitian [9] Hasil menunjukkan bahwa *ACO-Xgboost* dapat meningkatkan akurasi sebesar 0.20% hingga 3.53% dibandingkan dengan algoritma lain. Hal ini menunjukkan bahwa model *ACO* pada *Xgboost* dapat meningkatkan akurasi dibandingkan dengan algoritma klasifikasi lain dan tidak hanya meningkatkan akurasi namun menawarkan keseimbangan antara akurasi dan efisiensi pemilihan fitur yang optimal.

3.2.2.5 Evaluation

Pada tahapan *evaluasi* proses yang akan dilakukan adalah peninjauan kembali dari hasil analisa dengan tujuan dan kriteria yang telah ditentukan sebelumnya lewat tahapan *business understanding* [42]. Pada penelitian kali ini evaluasi yang akan dilakukan akan berfokus kepada nilai akurasi, presisi, sensitivitas, dan waktu pemrosesan, serta fitur terpilih. Akurasi merupakan rasio prediksi benar terhadap total prediksi untuk mengukur seberapa sering model membuat prediksi benar. Akurasi merupakan urutan penting pertama karena memberikan gambaran keseluruhan performa model [46]. Sensitivitas adalah rasio prediksi positif yang benar terhadap total jumlah sebenarnya positif untuk mengukur kemampuan model dalam mendeteksi sampel positif sebenarnya. Sensitivitas merupakan urutan penting kedua karena dalam deteksi penyakit sangat penting untuk meminimalkan *false negatives* [47]. Presisi merupakan rasio prediksi positif yang benar terhadap total prediksi positif untuk mengukur seberapa akurat prediksi positif model. Presisi merupakan urutan penting ketiga karena penting ketika biaya kesalahan positif yang tinggi terutama seperti dalam diagnosis penyakit karena kesalahan pengobatan bisa berakibat fatal [48]. Waktu pemrosesan adalah waktu yang diperlukan model untuk melakukan pelatihan dan pengujian. Waktu pemrosesan merupakan urutan keempat karena mengukur efisiensi komputasi dari algoritma sangat penting untuk menjalankan algoritma dalam data jumlah yang besar atau sumber daya komputasi yang terbatas [49].

3.2.2.6 Deployment

Pada tahapan *Deployement* yang merupakan fase terakhir ini, hasil yang didapat dari analisis data akan diterjemahkan ke dalam bentuk rekomendasi yang dapat ditindaklanjuti dengan cara hasil analisis dikomunikasikan dengan efektif proyek analisis dapat sukses [42]. Pada penelitian ini tahapan *deployment* tidak akan dilakukan karena penelitian ini tidak diimplementasikan ke dalam bentuk model yang dijalankan didunia nyata, namun penelitian ini hanya untuk keperluan studi.

3.3 Teknik Pengumpulan Data

3.3.1 Populasi dan Sampel

Teknik pengambilan sampel data akan digunakan pada penelitian kali ini adalah menggunakan skema validasi *train-test*. Penggunaan skema validasi *train-test* akan digunakan pada penelitian ini dengan tujuan untuk mendapatkan dan meningkatkan hasil peninjauan kembali dari model yang dibangun pada penelitian ini. Data *training* yang diambil adalah sebanyak 70% dari total *dataset* akan digunakan untuk melatih model yang dibangun sementara data *testing* yang akan digunakan adalah sebanyak 30% dari total *dataset* digunakan untuk melakukan *training* pada model [50].

3.4 Variabel Penelitian

3.4.1 Variabel Dependen

Pada penelitian kali ini variabel dependen yang dipilih adalah *stroke*. Variabel *Hadstroke* dipilih karena variabel ini menunjukkan hasil diagnosa penyakit *stroke* dengan indikator *true or false* yang dihasilkan dari penggunaan variabel yang lain pada *dataset* yang digunakan.

3.4.2 Variabel Independen

Variabel independen penelitian ini adalah *State, Sex, GeneralHealth, Physical Health, Days, Mental Health Days, Last Checkup Time, Physical Activities, Sleep Hours, Removed Teeth, Had Heart Attack, Had Angina, Had Stroke, Had Asthma, Had Skin Cancer, Had COPD, Had Depressive Disorder, Had Kidney Disease, Had Arthritis, Had Diabetes, Deaf Or Hard Of Hearing, Blind Or Vision Difficulty, Difficulty Concentrating, Difficulty Walking, Difficulty Dressing, Bathing, Difficulty Errands, Smoker Status, E Cigarette Usage, Chest Scan, Race Ethnicity Category, Age Category, Height In Meters, Weight In Kilograms, BMI, Alcohol Drinkers, HIV Testing, Flu Vax Last 12, Pneumo Vax Ever, Tetanus Last 10T dap, High Risk Last Year, Covid Pos.* Setiap variabel independen yang digunakan memiliki definisi dan indikator variabel tersebut. Pada Tabel 3.2 akan ditunjukkan definisi variabel independen dan indikator yang digunakan untuk variabel tersebut serta urutan nomor variabel pada *array*.

Tabel 3. 2 Definisi variabel Independen

No	Variabel	Definisi	Indikator
0	*State*	Wilayah geografis tempat responden tinggal	*Washington, New York, Other*
1	*Sex*	Jenis kelamin responden	*Female, Male*
2	*GeneralHealth*	Penilaian umum kesehatan responden	*Very good, Good, Other*
3	*Physical Health Days*	Jumlah hari dengan kesehatan fisik buruk	*Label Range 0.00 - 30.00, Count Varies*
4	*Mental Health Days*	Jumlah hari dengan kesehatan mental buruk	*Label Range 0.00 - 30.00, Count Varies*
5	*Last Checkup Time*	Waktu terakhir melakukan pemeriksaan kesehatan	*Within past year, Past 2 years, Other*
6	*Physical Activities*	Melakukan aktivitas fisik secara teratur	*True, False*
7	*Sleep Hours*	Jumlah jam tidur rata-rata per hari	*Range 1.00 – 24.00 hours, Count Varies*

8	*Removed Teeth*	Jumlah gigi yang telah dicabut	*None of Them, 1 to 5, Other*
9	*Had Heart Attack*	Riwayat serangan jantung	*True, False*
10	*Had Angina*	Riwayat angina (nyeri dada)	*True, False*
11	*Had Asthma*	Riwayat asma	*True, False*
12	*Had Skin Cancer*	Riwayat kanker kulit	*True, False*
13	*Had COPD*	Riwayat Penyakit Paru Obstruktif Kronis (*COPD*)	*True, False*
14	*Had Depressive Disorder*	Riwayat gangguan depresi	*True, False*
15	*Had Kidney Disease*	Riwayat penyakit ginjal	*True, False*
16	*Had Arthritis*	Riwayat arthritis	*True, False*
17	*Had Diabetes*	Riwayat diabetes	*No, Yes*
18	*Deaf Or Hard of Hearing*	Kesusahan mendengar atau tuli	*True, False*
19	*Blind Or Vision Difficulty*	Kesusahan melihat atau buta	*True, False*
20	*Difficulty Concentrating*	Kesulitan berkonsentrasi	*True, False*
21	*Difficulty Walking*	Kesulitan berjalan	*True, False*
22	*Difficulty Dressing, Bathing*	Kesulitan berpakaian atau mandi	*True, False*
23	*Difficulty Errands*	Kesulitan melakukan tugas atau urusan sehari-hari	*True, False*
24	*Smoker Status*	Status perokok	*Never smoked, Former smoker, Other*
25	*E Cigarette Usage*	Penggunaan rokok elektronik	*Never used, Not at All, Other*
26	*Chest Scan*	Pemeriksaan dada menggunakan scan	*True, False*
27	*Race Ethnicity Category*	Kategori etnis atau ras	*White only, Non-Hispanic, Hispanic, Other*
28	*Age Category*	Kategori usia responden	*Age 65 to 69, Age 60 to 64, Other*
29	*Height In Meters*	Tinggi badan dalam meter	*Label Range 0.91 - 2.41 meters, Count Varies*
30	*Weight In Kilograms*	Berat badan dalam kilogram	*Label Range 28.1-293 kg, Count Varies*

31	*BMI*	Indeks Massa Tubuh	*Label Range 12 – 97.7, Count Varies*
32	*Alcohol Drinkers*	Konsumsi alkohol	*True, False*
33	*HIV Testing*	Pernah melakukan tes HIV	*True, False*
34	*Flu Vax Last 12*	Vaksinasi flu dalam 12 bulan terakhir	*True, False*
35	*Pneumo Vax Ever*	Pernah menerima vaksin pneumokokus	*True, False*
36	*Tetanus Last 10Tdap*	Vaksinasi tetanus dalam 10 tahun terakhir termasuk Tdap	*Yes, No, Other*
37	*High Risk Last Year*	Tinggi risiko kesehatan dalam tahun terakhir	*True, False*
38	*Covid Pos*	Positif COVID-19	*Yes, No, Other*

3.5 Teknik Analisis Data

Tujuan dari dilakukan penelitian adalah menemukan dan mengkomparasikan hasil akurasi, presisi, sensitivitas dan waktu pemrosesan dari algoritma berbasis pohon keputusan yaitu *Decision Tree*, *Xgboost*, dan *Extra Tree* dengan seleksi fitur menggunakan algoritma *Swarm Intelligence* yaitu *PSO, dan ACO* dalam mengklasifikasi fitur yang meningkatkan risiko seseorang untuk terkena *stroke*. Analisa pada penelitian kali ini akan menggunakan bahasa pemrograman *python* yang diakses lewat *platform* dari *jupyter notebook*..