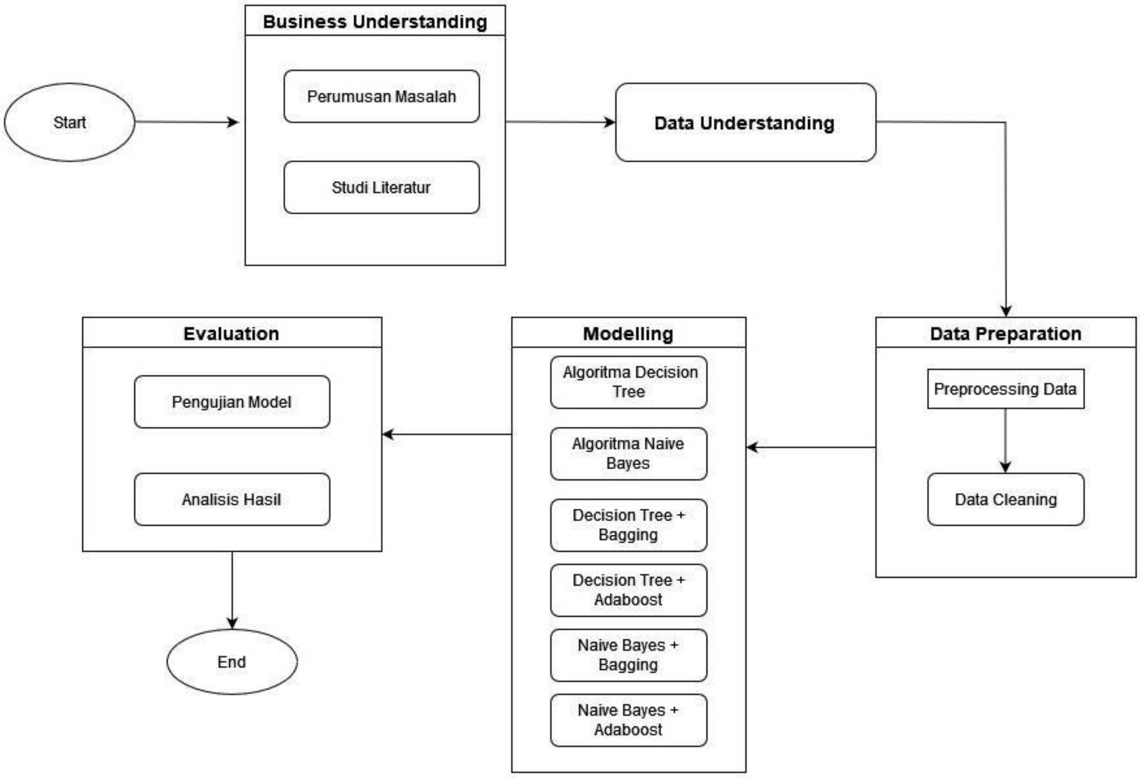
BAB III
METODOLOGI PENELITIAN

3.1 Tahapan Penelitian

Penelitian ini merupakan jenis penelitian eksperimen. Dimana data yang dikumpulkan akan dilakukan pengujian menggunakan model algoritma klasifikasi serta penerapan teknik ensemble. pada metode penelitian ini menggunakan model CRISP-DM yang terdiri dari 6 tahapan. Tahap pertama adalah *Business Understanding* yang didalamnya terdapat proses yaitu perumusan masalah dan studi literatur untuk menambah pemahaman terkait proses bisnis yang akan dilakukan. Tahap kedua adalah *Data Understanding*/pemahaman data yang mencakup pengumpulan data yang relevan untuk di analisis. Tahap ketiga adalah *Data Preparation* atau tahap pengolahan data. Pada tahap ini akan dilakukan *preprocessing data* yaitu pembersihan data yang noise, tidak konsisten atau *missing value*. Tahap keempat adalah *Modelling* atau tahap pemodelan menggunakan algoritma klasifikasi decision tree dan naïve bayes dengan penerapan teknik *ensemble* yaitu bagging dan adaboost. Tahap kelima adalah *Evaluation* yang didalamnya terdapat proses pengujian model yang terdiri dari testing data per-skenario dan perhitungan *confusion matrix* serta analisis hasil. Tahap keenam adalah *Deployment* atau implementasi dari model yang telah dikembangkan dan siap untuk digunakan secara luas. Tetapi, pada penelitian ini hanya dilakukan sampai tahap *evaluation*. Ilustrasi pada Gambar 3.1 menggambarkan tahapan penelitian yang dilakukan.



Gambar 3.1 Tahapan Penelitian

3.1.1 *Business Understanding*

Business Understanding merupakan langkah awal dan kritis dalam proses data mining dan analisis data yang melibatkan pemahaman mendalam tentang masalah bisnis yang sedang dihadapi, tujuan bisnis yang ingin dicapai, serta cara bagaimana solusi analitik dapat memberikan nilai tambah bagi organisasi. Pada penelitian ini *Business Understanding* terdiri dari 2 tahapan yaitu perumusan masalah dan studi literatur terkait prediksi penyakit stroke menggunakan algoritma klasifikasi dan teknik ensemble.

3.1.1.1 Perumusan Masalah

Tahap ini merupakan langkah awal sebelum memulai penelitian, yakni tahap perumusan masalah. Pada tahap ini, pemahaman terhadap masalah penelitian diuraikan berdasarkan latar belakang penelitian tentang prediksi penyakit stroke. Metode yang digunakan dalam penelitian ini adalah algoritma Decision Tree dan Naïve Bayes, dengan penerapan teknik *ensemble* seperti teknik Bagging dan Adaboost untuk memprediksi penyakit stroke. Dengan menerapkan metode ini, diharapkan dapat memastikan apakah seorang pasien mengalami stroke atau tidak dengan hasil prediksi yang akurat.

3.1.1.2 Studi Literatur

Pada tahap ini, dilakukan proses sistematis untuk meninjau dan menganalisis literatur, publikasi, artikel ilmiah, buku, dan sumber informasi lain yang relevan dengan topik penelitian atau masalah yang sedang diteliti. Tujuan dari studi literatur adalah untuk memperoleh pemahaman yang mendalam tentang pengetahuan yang telah ada terkait dengan topik tertentu, mengevaluasi hasil penelitian sebelumnya, mengidentifikasi kesenjangan pengetahuan, dan membangun landasan teoritis yang kokoh untuk penelitian yang sedang dilakukan. Pencarian literatur dilakukan dari berbagai sumber yang relevan dengan penelitian, yang kemudian akan dijadikan sebagai referensi. Proses studi literatur ini dimaksudkan untuk memahami penelitian terdahulu yang terkait dengan stroke dan metode penelitian yang relevan.

3.1.2 *Data Understanding*

Data Understanding adalah tahap dalam proses analisis data yang berfokus pada pemahaman mendalam terhadap data yang akan digunakan dalam analisis. Tujuan dari tahap ini adalah untuk mengumpulkan informasi tentang data, memahami karakteristiknya, mengidentifikasi potensi masalah, dan mempersiapkan data untuk tahap analisis yang lebih lanjut. Pada penelitian ini *data understanding* merupakan tahap pemahaman data penyakit stroke yang diambil dari situs kaggle. *Dataset stroke prediction* merupakan dataset yang bersumber dari *Kaggle* berjumlah 5110 data mentah. Dataset ini terdiri dari sebelas atribut prediktor dan satu atribut target yang merupakan stroke. Atribut stroke mencakup data pasien yang mengalami stroke dan yang tidak. Persentase pasien yang tidak mengalami stroke mencapai 95%, sementara yang mengalami stroke hanya 5%. Dari penjelasan tersebut, dataset ini dapat dikategorikan sebagai dataset yang tidak seimbang karena proporsi pasien yang tidak mengalami stroke lebih besar daripada yang mengalami stroke. Tabel 3.1 di bawah ini menjelaskan masing-masing atribut dari *dataset stroke prediction*.

Tabel 3.1 Penjelasan Atribut *Dataset Stroke Prediction*

Atribut	Deskripsi	Tipe Data
id	Pengenal unik	Integer
gender	Jenis kelamin	Polynominal
age	Umur	Integer
hypertension	Riwayat penyakit hipertensi. (jika 0= Tidak, 1=Ya)	Integer
heart_disease	Riwayat penyakit jantung. (0 = Tidak, 1=Ya)	Integer
ever_married	Pasien pernah menikah atau tidak. (Yes or No)	Polynominal
work_type	Jenis Pekerjaan. (*children, Govt_job, Private, Self-employed, Never_worked*)	Polynominal
Residence_type	Jenis Tempat Tinggal (*Urban*=Perkotaan, *Rural*= Pedesaan)	Polynominal
avg_glucose_level	Rata-rata kadar glukosa/gula darah	Integer
bmi	Indeks massa tubuh	Real
smoking_status	Status perokok. Ada 3 kategori	Polynominal

	(*formerly smoked*=sebelumnya merokok, *smokes*=merokok, *never smoked*= tidak pernah merokok, *Unknown*=tidak diketahui riwayatnya	
stroke	Atribut *predictor* atau penentu, terdiri dari 2 kategori yaitu 0 dan 1, (jika 0 = tidak stroke, 1 = stroke)	Integer

3.1.3 *Data Preparation*

Pada tahap ini data diolah, dibersihkan, dan dipersiapkan untuk analisis lebih lanjut. Langkah-langkah yang umum dilakukan di tahap ini meliputi penghapusan data yang tidak relevan atau tidak lengkap, penanganan nilai yang hilang, normalisasi data, dan transformasi variabel untuk mempersiapkan data yang sesuai untuk pemodelan. Dalam penelitian ini, bagian dari *data preparation* melibatkan proses pengolahan data yang mencakup pembersihan data/*data cleaning*.

3.1.3.1 Pengolahan Data

Tahap pengolahan data merupakan elemen krusial dalam proses analisis data yang melibatkan serangkaian langkah untuk membersihkan, mengatur, dan menyiapkan data mentah agar siap digunakan untuk analisis lebih lanjut. Pada tahap ini, dilakukan proses preprocessing data untuk mengubah data mentah menjadi bentuk yang dapat dengan mudah diproses menggunakan teknik data mining. Terdapat beberapa atribut yang memiliki data tidak konsisten, seperti "N/A". Oleh karena itu, dilakukan pembersihan data atau data cleaning untuk menghapus entri yang tidak sesuai dan menggantinya dengan nilai rata-rata dari atribut yang memiliki nilai

"N/A". Dalam dataset prediksi stroke, terdapat 201 entri yang memiliki nilai "N/A" pada atribut BMI.

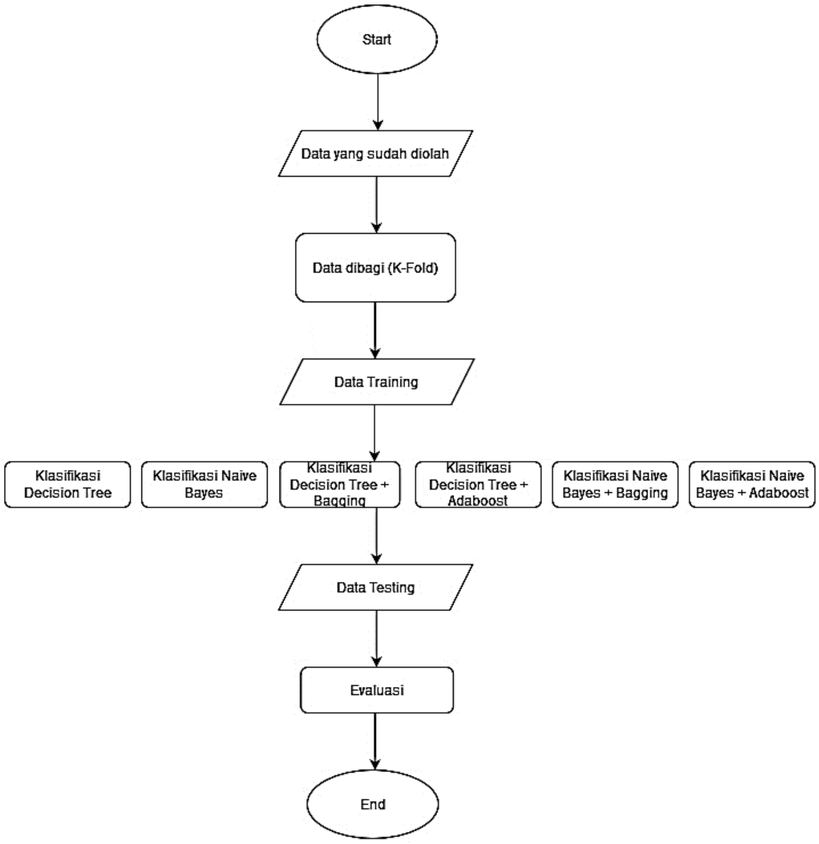
3.1.4 *Modelling* (Pemodelan)

Setelah melewati tahap *data preparation* atau pengolahan data, langkah selanjutnya adalah tahap *modelling/*pelatihan model. Pada tahap keempat ini, teknik pemodelan yang dilakukan menggunakan algoritma Decision Tree dan Naïve Bayes, dengan menerapkan teknik *ensemble* yaitu bagging dan adaboost. Proses ini melibatkan pembagian klasifikasi ke dalam enam skenario yang berbeda. Tabel 3.2 di bawah ini menunjukkan proses skenario klasifikasi.

Tabel 3.2 Skenario Klasifikasi

Decision Tree	Naïve Bayes	Decision Tree + Bagging	Decision Tree + Adaboost	Naïve Bayes + Bagging	Naïve Bayes + Adaboost
Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decision tree saja tanpa menerapkan teknik *Bagging* atau *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes saja tanpa menerapkan teknik *Bagging* atau *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decisiontree dan teknik bagging	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma decisiontree dan teknik *Adaboost*	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes dan teknik bagging	Pada skenario ini akan dilakukan klasifikasi menggunakan algoritma naïve bayes dan *Adaboost*

Berdasarkan skenario sebelumnya, langkah – langkah dalam proses klasifikasi dapat digambarkan ke dalam *flowchart* seperti Gambar 3.2 dibawah ini.



Gambar 3.2 *Flowchart* Klasifikasi Algoritma

Berikut adalah penjelasan tentang *flowchart* klasifikasi Algoritma yang terdapat pada Gambar 3.2 di atas:

1. Data yang telah melalui tahap *preprocessing* dibagi menjadi dua bagian, yaitu data training dan data testing, menggunakan *k-fold cross validation* untuk mendapatkan hasil pengujian yang optimal.
2. Setelah proses *preprocessing data*, dilakukan proses klasifikasi menggunakan beberapa algoritma, termasuk Decision Tree, Naïve Bayes, Decision Tree + Bagging, Decision Tree + Adaboost, Naïve Bayes + Bagging, dan Naïve Bayes + Adaboost.
3. Pengujian klasifikasi dilakukan menggunakan data testing.
4. Evaluasi dilakukan pada tahap pengujian untuk menguji proses klasifikasi dengan melakukan perhitungan untuk mendapatkan nilai akurasi, spesifisitas, sensitivitas, dan presisi.

5. Terdapat enam proses dalam proses klasifikasi, yakni klasifikasi menggunakan Decision Tree, klasifikasi menggunakan Naïve Bayes, Decision Tree + Bagging, Decision Tree + Adaboost, Naïve Bayes + Bagging, dan Naïve Bayes + Adaboost.

3.1.5 *Evaluation* (Evaluasi)

Tahap evaluasi (*evaluation*) dalam proses CRISP-DM adalah tahap yang penting untuk mengukur kinerja model yang telah dibangun selama tahap pemodelan. Tujuannya adalah untuk menilai seberapa baik model dapat memenuhi tujuan bisnis yang telah ditetapkan dan seberapa andal model tersebut dalam melakukan prediksi atau analisis. Pada penelitian ini, evaluasi meliputi tahap pengujian model serta analisis hasil yang diperoleh dari pengujian tersebut.

3.1.5.1 Pengujian Model

Tahap ini terdiri dari dua proses yaitu testing data per-skenario dan perhitungan *confusion matrix*. Pada testing data per-skenario melibatkan pengujian model pada enam skenario klasifikasi dengan menggunakan data yang telah dilabeli. Sebelum melakukan pengujian model, penting untuk melakukan pengolahan data terlebih dahulu agar hasil pengujiannya optimal. Proses pengolahan data dilakukan melalui tahap *preprocessing*. Setelahnya, data akan diteruskan ke model atau algoritma klasifikasi untuk mengevaluasi performa model tersebut. Pengujian model dilakukan menggunakan metode *confusion matrix*. Sebelum memulai pengujian, data dipisahkan menjadi data latih dan data uji. Pengujian dilakukan dengan menggunakan *10 fold cross-validation*, di mana 10 sampel data diambil secara acak untuk mengevaluasi model. Dari berbagai sampel data tersebut, model yang memberikan hasil optimal dipilih dengan mencari rata-rata dari 10 sampel tersebut untuk mendapatkan nilai akurasi, presisi, spesifisitas, dan sensitivitas.

Setelah dilakukan testing data per-skenario selanjutnya menghitung *confusion matrix* secara manual menggunakan rumus matematis untuk mendapatkan nilai akurasi, presisi, spesifisitas, dan sensitivitas. Setelah mendapatkan hasil, maka bandingkan hasil perhitungan secara manual dengan hasil pengujian pada sistem apakah sama atau tidak. Jika sama, berarti pengujian model pada sistem berhasil/valid, begitupun sebaliknya.

3.1.5.2 Analisis Hasil

Tahap analisis hasil bertujuan untuk membandingkan hasil dari setiap skenario yang telah melewati tahap pengujian sebelumnya. Enam skenario ini mencakup pengujian algoritma decision tree, algoritma naïve bayes, algoritma decision tree + bagging, algoritma decision tree + adaboost, algoritma naïve bayes + bagging, dan algoritma naïve bayes + adaboost. Pengujian pada keenam skenario tersebut menghasilkan nilai presisi, spesifisitas, sensitivitas, dan akurasi. Dengan membandingkan nilai-nilai tersebut, kita dapat menentukan metode yang memiliki dampak signifikan pada peningkatan performa klasifikasi. Selain itu, tahap analisis juga melibatkan perbandingan hasil penelitian ini dengan penelitian lain yang menggunakan metode serupa.

3.1.6 Jadwal Penelitian

Kegiatan	Maret	April	Mei	Juni	Juli	Agustus	September	Oktober	November	Desember
Identifikasi Masalah & Pembuatan Bab 1										
Studi Pustaka & Pembuatan Bab 2										
Penentuan Metode Penelitian, Pengumpulan Data & Pembuatan Bab 3										
Eksperimen dan Hasil, Penyusunan Bab 4,										
Penyusunan Bab 5										