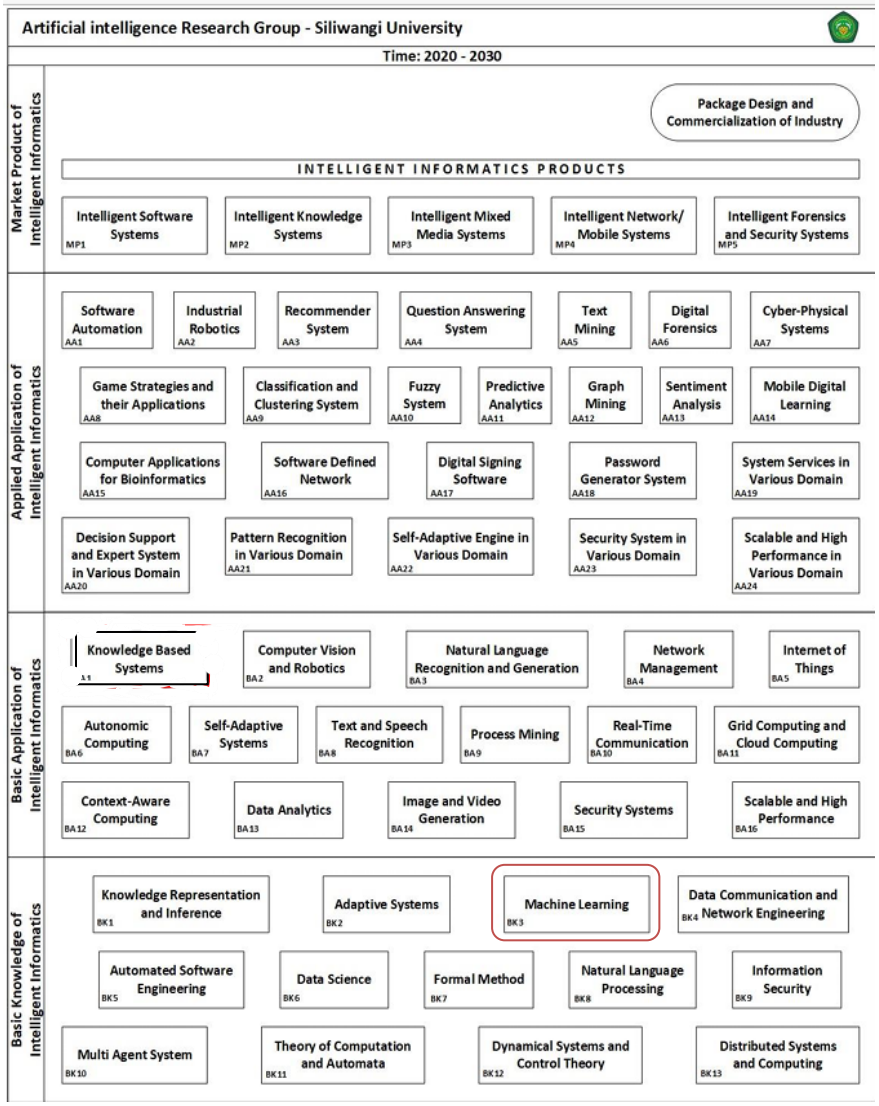
BAB III
METODOLOGI PENELITIAN

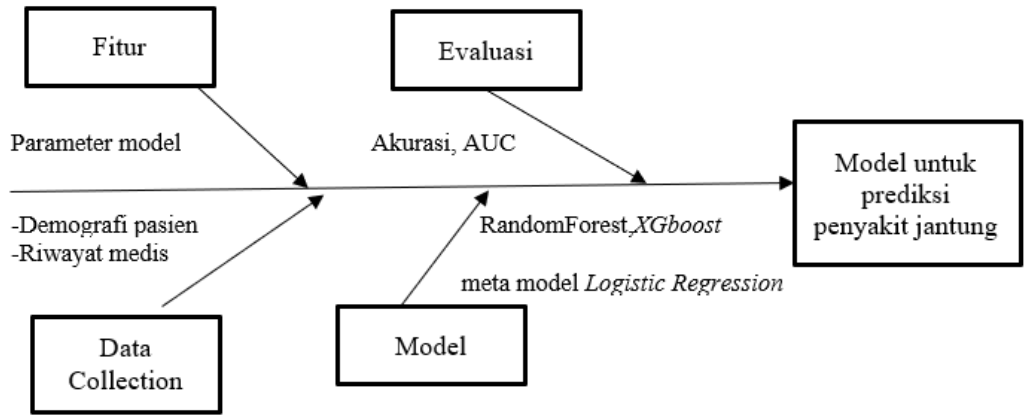
3.1 Roadmap Penelitian

Secara umum, rencana ini selaras dengan roadmap *Artificial Intelligence* di Universitas Siliwangi. Topik yang diangkat berfokus pada yang dipilih yaitu *predictive analysis* pada ranah *Applied Aplication of intelligent informatics*. Roadmap penelitian dapat dilihat pada gambar 3.1 Roadmap penelitian (AIS Universitas Siliwangi, 2020).



Gambar 3. 1 Roadmap Penelitian (AIS Universitas Siliwangi, 2020).

Gambar 3.1 pemilihan topik *Machine learning* pada ranah *Basic Knowledge of Intelligent Informatics* dan *Predictive Analytics* pada ranah *Applied Applications of Intelligent Informatics* dalam penelitian ini bertujuan untuk membangun sebuah pemodelan yang akan menjadi dasar pengetahuan dalam sistem prediksi penyakit jantung. Penelitian ini berfokus pada pengembangan model prediktif dengan pendekatan metode *stacking ensemble XGBoost*, *Random Forest*, dan Meta-Model *Logistic Regression*. Pemodelan ini diharapkan dapat menjadi dasar pengetahuan dalam pengembangan sistem pendukung keputusan medis, yang dapat membantu dalam mendeteksi risiko penyakit jantung secara lebih akurat dan efisien. Gambar 3.2 merupakan fishbone penelitian sebagai bagian dari roadmap.

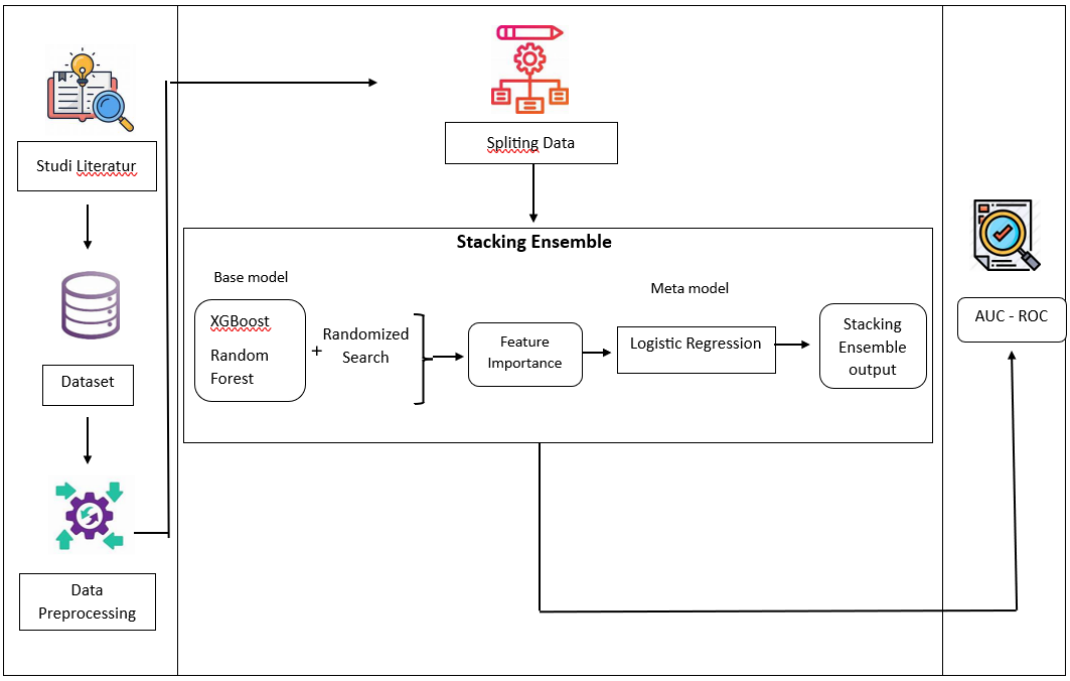


Gambar 3.2 Fishbone Penelitian.

Gambar 3.2 fishbone mempresentasikan *Basic Knowledge of Intelligent Informatics* dan *Applied Applications of Intelligent Informatics*. Penelitian dirancang untuk pemodelan prediksi penyakit jantung. Pengambilan data yang digunakan memiliki informasi demografi pasien, riwayat medis. Kemudian , pemilihan fitur dari parameter untuk model dibutuhkan dalam pemodelan. Selanjutnya, hasil dari pemodelan menggunakan *machine learning* di evaluasi untuk mengetahui performa model dalam memprediksi penyakit jantung.

3.2 Metode Peneitian

Metode penelitian yang digunakan dalam prediksi penyakit jantung pada penelitian ini adalah metode *stacking ensemble*, yang menggabungkan beberapa model pembelajaran mesin untuk meningkatkan akurasi prediksi. Model yang digunakan dalam pendekatan ini terdiri dari *XGBoost, Random Forest*, dan *Logistic Regression* sebagai meta-model. Selain itu, dilakukan optimasi *hyperparameter* menggunakan *Randomized search* untuk mendapatkan kombinasi parameter terbaik yang dapat meningkatkan kinerja model. Proses pemodelan ini melibatkan beberapa tahapan penting, mulai dari *preprocessing* data, penerapan *K-Fold Cross Validation*, training model dasar (base models), optimasi *hyperparameter*, *feature importance* yang mempengaruhi prediksi hingga pembentukan meta-model. Penerapan metode *stacking ensemble* ini bertujuan untuk menggabungkan keunggulan masing-masing model dasar, sehingga dapat memberikan prediksi yang lebih akurat dan andal dalam mendeteksi penyakit jantung. Gambar 3.3 menggambarkan alur metodologi penelitian, di mana *stacking ensemble* diterapkan sebagai strategi utama dalam meningkatkan performa prediksi penyakit jantung.



Gambar 3.3 Tahapan Penelitian.

Gambar 3.3 menunjukkan tahapan utama dalam proses penelitian yang dilakukan untuk membangun model prediksi penyakit jantung. Penelitian ini diawali dengan studi literatur, yaitu mengkaji berbagai jurnal dan penelitian terdahulu terkait model *XGBoost, Random Forest*, optimasi *hyperparameter Randomized search*, serta metode *stacking ensemble*. Setelah itu pengumpulan dataset, yang berisi data demografi pasien serta riwayat medis yang akan digunakan sebagai variabel dalam proses prediksi. Dataset ini kemudian melalui tahap *preprocessing* yang bertujuan untuk membersihkan, mengolah, dan menyesuaikan data agar dapat dianalisis dengan lebih baik serta sesuai dengan kebutuhan penelitian (Wang et al., 2021).

Setelah *preprocessing* selesai, dilakukan integrasi base model menggunakan *XGBoost* dan *Random Forest* secara individu untuk mengevaluasi performa masing-masing model. Tujuan dari tahap ini adalah untuk memahami bagaimana setiap model bekerja dalam memprediksi penyakit jantung dan memperoleh metrik evaluasi awal sebelum diterapkan ke dalam metode *ensemble*. Kemudian, dilakukan optimasi *hyperparameter* menggunakan *Randomized search*, yang bertujuan untuk meningkatkan performa model dengan menemukan kombinasi parameter terbaik (Ansyari et al., 2023) (Saputra et al., 2022).

Sebelum masuk ke pemodelan *stacking*, terlebih dahulu mengetahui fitur-fitur yang mempengaruhi prediksi pada model *XGBoost* dan *Random Forest*. Tahapan berikutnya adalah penggabungan kedua model (*XGBoost dan Random Forest*) dalam pemodelan *stacking ensemble*, yang digabungkan dengan meta-model *Logistic Regression*. Metode *stacking* ini digunakan untuk meningkatkan akurasi prediksi dengan menggabungkan kekuatan dari masing-masing base model agar menghasilkan prediksi yang lebih optimal (Fonda et al., 2024) serta meta model *Logistic Regression* dalam memprediksi penyakit jantung. Pada tahap terakhir, dilakukan evaluasi performa model, di mana hasil prediksi dibandingkan dengan nilai sebenarnya untuk mengukur efektivitas pendekatan yang telah digunakan. Metriks evaluasi yang digunakan dalam penelitian ini adalah akurasi dan AUC - ROC.

3.3 Studi Literatur

Tahap pertama dalam penelitian ini adalah studi literatur, yang dilakukan untuk memperoleh pemahaman mendalam terkait metode, model, dan pendekatan yang relevan dengan topik penelitian. Studi literatur dilakukan terhadap berbagai sumber ilmiah, seperti jurnal, prosiding, dan artikel akademik yang diterbitkan dalam lima tahun terakhir, khususnya yang membahas penerapan *machine learning* dalam prediksi penyakit jantung.

Fokus utama studi literatur adalah pada model *XGBoost* dan *Random Forest*, sebagai model dasar dalam *ensemble*, serta pendekatan *stacking ensemble* dan teknik optimasi *Randomized search*. Dengan melakukan studi literatur, peneliti dapat mengidentifikasi kelebihan, kekurangan, serta hasil dari penelitian sebelumnya untuk dijadikan dasar dalam merancang metode yang digunakan dalam penelitian ini.

3.4 Dataset

Penelitian ini menggunakan data sekunder yang diperoleh dari situs *UCI Repository*, yang dapat diakses melalui tautan https://archive.ics.uci.edu/dataset/45/heart+disease Dataset ini berkaitan dengan permasalahan penyakit jantung dan terdiri dari 303 sampel data. Dataset yang digunakan memiliki 14 atribut, di mana 13 atribut pertama merupakan indikator klinis pasien yang digunakan sebagai variabel prediktor, sedangkan atribut ke-14 berfungsi sebagai target klasifikasi. Atribut-atribut dalam dataset ini meliputi *Age, Sex, Cp, Trestbps, Chol, Fbs (gula darah puasa), Restecg* (hasil elektrokardiografi), *Thalach, Exang, Oldpeak, Slope , Ca, dan Thal*, Target. Berdasarkan penelitian sebelumnya(Ansyari et al., 2023) tabel 3.1 menyajikan deskripsi lebih lanjut mengenai atribut dalam dataset yang digunakan.

Tabel 3. 1 Description of Research Data Attributes

No	Attribute	Description	Category
1	Age	Age in years	Numeric
2	Sex	Sex (01=male; 02=female)	Binary
3	Cp	Chest pain type	Nominal
4	Trestbps	Resting blood pressure	Numeric
5	Chol	Serum cholesterol in mg/dl	Numeric
6	Fbs	Fasting blood sugar> 120 mg/dl 01= true; 02=false	Binary
7	Restecg	Resting electrocardiographic results	Nominal
8	Thalach	Maximum heart rate achieved	Numeric
9	Exang	Exercise-induced angina	Binary
10	Oldpeak	ST depression induced by exercise relative to rest	Numeric
11	Slope	The slope of peak exercise ST segment	Nominal
12	Ca	Number of major vessels (0-3) colored by fluoroscopy	Nominal
13	Thal	3=normal; 6=fixed defect; 7=reversable defect	Nominal
14	Target	Diagnosis of heart disease	Binary

3.5 Data Preprocessing

Tahap ini merupakan langkah awal dalam mempersiapkan data sebelum digunakan dalam pemodelan. Data yang digunakan dalam penelitian ini berasal dari *UCI Heart Disease Dataset*, yang terdiri dari 303 data dengan 14 atribut. Proses *preprocessing* dilakukan agar data lebih siap untuk dianalisis dan digunakan dalam model *machine learning* secara optimal (Wang et al., 2021).

3.5.1. Menangani *Missing Values*

Setelah dataset dimuat, dilakukan pemeriksaan terhadap kemungkinan adanya *missing values* dalam data. Jika terdapat nilai yang hilang, dilakukan penanganan dengan menggantinya menggunakan metode tertentu, seperti mean, median, atau modus, tergantung pada jenis data dan distribusinya. Apabila jumlah nilai yang hilang dalam suatu atribut terlalu besar dan tidak memungkinkan untuk diisi ulang, maka atribut tersebut dapat dipertimbangkan untuk dihapus agar tidak

mengganggu proses analisis lebih lanjut.

3.5.2. *Feature Scalling*

Langkah selanjutnya adalah melakukan *feature scaling* guna menyamakan skala pada seluruh atribut dalam dataset. Hal ini penting untuk mencegah dominasi atribut tertentu yang memiliki nilai lebih besar dibandingkan atribut lainnya. Dalam penelitian ini, digunakan metode *Standardization* (Z-score normalization) yang akan mengubah distribusi data sehingga memiliki mean = 0 dan standar deviasi = 1. Metode ini dipilih karena sesuai untuk model yang sensitif terhadap skala data.

3.6 Spliting Data

Setelah melalui proses *preprocessing*, data kemudian dibagi menjadi dua bagian utama melalui proses *splitting data*, yaitu data *training* dan data testing. Pembagian ini bertujuan agar model *machine learning* dapat dilatih dengan sebagian data, dan diuji performanya pada data yang belum pernah dilihat sebelumnya, sehingga hasil prediksi lebih objektif dan tidak bias. Dalam penelitian ini digunakan metode *K-Fold Cross Validation*, yaitu teknik pembagian data yang lebih stabil dan dapat mengurangi overfitting. Teknik ini bekerja dengan membagi seluruh dataset menjadi *k bagian* (lipatan), kemudian model dilatih sebanyak *k* kali dengan memutar posisi bagian yang digunakan sebagai data uji dan data latih.

3.7 *Stacking Ensemble*

Pada tahap ini, proses pengembangan model prediksi penyakit jantung dilakukan dengan menerapkan metode *stacking ensemble*. Model ini mengombinasikan dua model sebagai base model (*XGBoost dan Random Forest*) yang hasilnya kemudian diproses lebih lanjut oleh meta model (*Logistic Regression*). Selain itu, dilakukan optimasi *hyperparameter* menggunakan *Randomized search* untuk meningkatkan kinerja model. Berikut adalah langkah-langkah yang akan dilakukan.

3.7.1. Base Model

Pada tahap ini, dua model *machine learning* akan diterapkan sebagai base model, yaitu *XGBoost* dan *Random Forest*. *XGBoost* model berbasis pohon keputusan yang terkenal dengan performanya yang tinggi dalam menangani data tabular serta kemampuannya menangani fitur dengan skala berbeda (Aditya et al., 2024). Sedangkan, *Random Forest* model berbasis *ensemble* yang menggabungkan beberapa pohon keputusan untuk meningkatkan akurasi prediksi dan mengurangi risiko *overfitting* dimana memiliki performa yang sangat baik pada data latih tetapi gagal dalam memprediksi data baru (data uji/test data**)** (Zhou et al., 2020). Kedua model ini akan dilatih menggunakan teknik 10-*Fold* Cross Validation, yang membagi data menjadi 10 bagian (*Fold*s), di mana setiap *Fold* akan digunakan secara bergantian sebagai data uji sementara sisanya digunakan sebagai data latih. Teknik ini membantu mengurangi kemungkinan bias dan memastikan hasil yang lebih stabil.

3.7.2 Optimasi *Hyperparameter*

Proses selanjutnya adalah melakukan *hyperparameter tuning* menggunakan metode *Randomized search* setelah base model selesai dibangun. Metode ini secara acak mencari kombinasi parameter terbaik untuk meningkatkan kinerja model. Beberapa parameter yang akan dioptimasi adalah *XGBoost* dengan *Learning rate, max depth, n_estimators, subsample*. Sedangkan *Random Forest* dengan *Number of trees (n_estimators), max depth, min_samples_split*. Tujuan dari proses ini adalah untuk menemukan kombinasi parameter terbaik sehingga model dapat bekerja lebih optimal dalam memprediksi penyakit jantung (Agus Dendi Rachmatsyah, 2024).

3.7.3 *Feature importance*

Analisis *feature importance* pada model *Random Forest* dan *XGBoost* dilakukan setelah proses optimasi *hyperparameter* selesai. Langkah ini bertujuan untuk mengidentifikasi atribut klinis yang paling dominan dalam mempengaruhi klasifikasi pasien. *Feature importance* pada *Random Forest* dihitung berdasarkan

rata-rata pengurangan impurity (mean decrease in Gini), sedangkan pada *XGBoost* menggunakan skor *gain* dari setiap fitur dalam proses pemisahan data.

3.7.4 Meta Model (*stacking ensemble*)

Setelah base model selesai dioptimasi, dilakukan kombinasi model menggunakan metode *stacking ensemble*. Proses *stacking* dilakukan dengan cara output prediksi dari model *XGBoost* dan *Random Forest* digabungkan sebagai fitur baru untuk tahap selanjutnya. Fitur baru tersebut digunakan untuk melatih meta model, yaitu *Logistic Regression*, yang berfungsi sebagai pengambil keputusan akhir dalam klasifikasi penyakit jantung. Metode ini bekerja dengan membangun dua tingkat model klasifikasi, yaitu level 0 (base model) dan level 1 (meta model). Pada tahap awal, dataset digunakan sebagai input untuk beberapa model klasifikasi dasar yang berbeda, yang bekerja secara paralel dan menghasilkan prediksi masing-masing. Prediksi dari setiap model dasar kemudian dikombinasikan untuk membentuk dataset baru. Dataset baru ini, yang terdiri dari output base model, kemudian diberikan ke meta model untuk mempelajari pola hubungan antara output model dasar, sehingga menghasilkan prediksi akhir yang lebih akurat (Ganesan et al., 2022).

3.8 Evaluasi Model

Setelah Metode *stacking* terbentuk, dilakukan evaluasi menggunakan akurasi dan AUC - ROC untuk mengukur performa model dalam membedakan pasien dengan dan tanpa penyakit jantung. Model dengan skor evaluasi terbaik akan dipilih sebagai model final yang digunakan untuk prediksi. AUC-ROC untuk mengukur kemampuan model dalam membedakan antara kelas positif dan negatif berdasarkan area di bawah kurva ROC (Rahim et al., 2022).