BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Objek pada penelitian ini adalah data medis pasien yang bersumber dari dataset MIMIC-IV (Medical Information Mart for Intensive Care). Dataset ini berisi kumpulan data klinis identifikasi dari pasien yang dirawat di *Intensive Care Unit* (ICU) di Beth Israel Deaconess Medical Center antara tahun 2008 dan 2022 [15]. Penelitian ini bertujuan untuk mengembangkan dan membandingkan model klasifikasi *machine learning* yang dapat memprediksi risiko gagal jantung pada pasien-pasien tersebut.

Secara spesifik, dataset yang digunakan bukanlah kumpulan pasien yang seluruhnya sudah teridentifikasi berisiko gagal jantung. Sebaliknya, dataset ini mencakup populasi pasien ICU yang lebih luas. Status "gagal jantung" atau "tidak gagal jantung" untuk setiap pasien tidak tersedia secara langsung, melainkan ditentukan melalui serangkaian proses persiapan data yang sistematis. Proses ini menjadi kunci untuk menciptakan variabel target (*ground truth*) yang akan digunakan untuk melatih model.

Penentuan apakah seorang pasien diklasifikasikan memiliki riwayat gagal jantung dilakukan dengan menganalisis data diagnosisnya. Prosesnya adalah sebagai berikut:

1. Pertama, dilakukan pencarian pada tabel kamus diagnosis (d_icd_diagnoses) menggunakan kata kunci "Heart Failure" untuk mengidentifikasi semua kode diagnosis (ICD code) yang relevan dengan kondisi gagal jantung.

2. Selanjutnya, daftar kode ICD ini digunakan untuk memfilter tabel riwayat diagnosis semua pasien (diagnoses_icd). Pasien yang memiliki setidaknya

satu catatan diagnosis yang cocok dengan kode gagal jantung tersebut akan diidentifikasi.

3. Terakhir, sebuah kolom target biner bernama ***heart_failure*** dibuat pada dataset pasien utama. Pasien yang teridentifikasi memiliki diagnosis gagal jantung diberi label 1, sedangkan pasien lainnya diberi label 0.

Proses pengolahan data ini melibatkan transformasi data dari jumlah baris yang sangat besar menjadi dataset akhir yang siap untuk dimodelkan. Alur reduksi data tersebut adalah sebagai berikut:

1. ***Data Awal (Raw Data)*:** Penelitian dimulai dengan beberapa tabel besar dari MIMIC-IV, di antaranya adalah patients (222.077 baris), diagnoses_icd (6.364.488 baris), chartevents (sampel 87.401 baris), dan labevents (sampel 101.436 baris).

2. ***Filtering Diagnosis*:** Setelah proses filtering diagnosis untuk "Heart Failure", teridentifikasi sebanyak **19.543 pasien unik** yang memiliki riwayat kondisi tersebut.

3. **Penggabungan Data Klinis*:*** Data demografis pasien digabungkan dengan data klinis yang relevan dari *chartevents* (tanda vital) dan *labevents* (hasil tes laboratorium). Setelah digabungkan dan dibersihkan dari nilai yang tidak relevan, dataset ini memiliki **10.973 baris** data observasi.

4. **Agregasi dan *Pivoting*:** Data observasi tersebut kemudian diagregasi untuk setiap pasien, di mana nilai rata-rata dari setiap parameter klinis dihitung. Setelah proses *pivoting* (mengubah format data dari *long* ke *wide*), dataset akhir yang digunakan untuk pemodelan terdiri dari **186 baris**, di mana setiap baris merepresentasikan satu pasien unik dengan fitur-fitur klinisnya.

Dataset akhir inilah yang kemudian digunakan untuk membandingkan kinerja tiga algoritma klasifikasi *Random Forest*, *Support Vector Machine* (SVM), dan *Logistic Regression* dalam upaya menemukan model prediksi gagal jantung yang paling akurat dan andal untuk lingkungan ICU.

3.2 Metode Penelitian

Pemilihan kerangka kerja CRISP-DM (Cross-Industry Standard Process for Data Mining) untuk penelitian ini sangat tepat karena beberapa alasan utama yang didukung oleh praktik standar dalam ilmu data:

1. **Standar Industri dan Akademik**: CRISP-DM adalah metodologi yang paling banyak digunakan dan diakui secara luas, baik di industri maupun di lingkungan akademik, untuk proyek analisis data dan *machine learning*. Menggunakannya memberikan struktur yang logis dan kredibilitas pada metodologi penelitian Anda [80].
2. **Berfokus pada Tujuan (*Business Understanding*)**: Tahapan pertama dalam CRISP-DM adalah *Business Understanding*. Ini memaksa peneliti untuk secara jelas mendefinisikan masalah dan tujuan dari sudut pandang praktis—dalam kasus ini, pentingnya prediksi gagal jantung untuk membantu tenaga medis. Hal ini memastikan bahwa solusi teknis yang dikembangkan benar-benar relevan dan menjawab kebutuhan nyata [81].
3. **Proses yang Terstruktur dan Iteratif**: CRISP-DM membagi proyek yang kompleks menjadi enam tahapan yang jelas dan sistematis (seperti yang dijabarkan pada Gambar 2.1 dan Bab IV). Sifatnya yang siklis (iteratif) memungkinkan peneliti untuk kembali ke tahap sebelumnya jika ditemukan wawasan baru. Misalnya, hasil dari tahap *Evaluation* bisa menginspirasi perbaikan pada tahap *Data Preparation*. Ini mencerminkan sifat asli dari proyek data mining yang jarang sekali berjalan linear [51].
4. **Fleksibel dan Independen dari Teknologi**: Sesuai dengan namanya (*Cross-Industry*), CRISP-DM tidak terikat pada satu industri, masalah, atau teknologi tertentu. Fleksibilitas ini membuatnya sangat cocok untuk diterapkan di berbagai bidang, termasuk bidang medis yang menjadi fokus penelitian ini [51].

Berikut adalah tabel Perbandingan Kerangka Kerja Data Mining:

Tabel 2. 2 Perbandingan Kerangka Kerja Data Mining

Kriteria	CRISP-DM	KDD (Knowledge Discovery in Databases)	SEMMA (Sample, Explore, Modify, Model, Assess)
Fokus Utama	Proses pemecahan masalah bisnis/penelitian secara menyeluruh [51], [81].	Proses teknis untuk mengekstraksi pola/pengetahuan dari data [81], [82].	Alur kerja metodologis yang berorientasi pada langkah-langkah teknis seorang analis data [82].
Tahapan	1. Business Understanding 2. Data Understanding 3. Data Preparation 4. Modeling 5. Evaluation 6. Deployment [51]	1. Selection 2. Pre-processing 3. Transformation 4. Data Mining 5. Interpretation/Evaluation [81]	1. Sample 2. Explore 3. Modify 4. Model 5. Assess [82].
Sifat Proses	Siklis dan iteratif, memungkinkan kembali ke tahap sebelumnya [51].	Umumnya dianggap lebih linear dan berurutan [81].	Siklis dan iteratif, mirip dengan CRISP-DM [82].
Kelebihan	-Sangat terstruktur dan komprehensif -Memprioritaskan tujuan bisnis/penelitian [82].	-Menjadi dasar bagi banyak metodologi lain [81] -Fokus yang kuat pada proses transformasi dan penemuan pola teknis [82].	Sederhana dan mudah diikuti [82].

Tabel 2.2 di atas menyoroti perbedaan fundamental antara tiga kerangka kerja utama dalam *data mining*. Meskipun KDD dan SEMMA menawarkan alur kerja teknis yang kuat untuk seorang analis, keduanya kurang memberikan penekanan pada tahap awal pemahaman masalah bisnis atau tujuan penelitian, yang merupakan fondasi penting dalam sebuah proyek.

Oleh karena itu, **CRISP-DM** dipilih sebagai metodologi dalam penelitian ini karena sifatnya yang menyeluruh (*end-to-end*). Kerangka kerja ini tidak hanya memandu proses teknis mulai dari persiapan data hingga evaluasi, tetapi juga memastikan bahwa seluruh proses analisis tetap selaras dengan tujuan utama penelitian, yaitu mengembangkan model prediksi gagal jantung yang efektif untuk kebutuhan klinis. Pendekatan ini menjadikan hasil penelitian lebih terstruktur, relevan, dan dapat dipertanggungjawabkan.

Berikut ini adalah penjelasan mengenai tahapan – tahapan CRISP – DM yang diimplementasikan pada penelitian ini :

1. *Business Understanding*

   Dalam penelitian ini, langkah pertama yang dilakukan adalah memahami tujuan utama, yaitu mengembangkan model prediksi yang dapat membantu tenaga medis dalam mendeteksi risiko gagal jantung pada pasien ICU. Kondisi ini sangat serius karena bisa menyebabkan komplikasi lain jika tidak ditangani sejak dini. Oleh karena itu, penelitian ini berfokus pada pemanfaatan machine learning untuk memberikan perkiraan yang lebih akurat mengenai pasien yang berisiko tinggi.

2. *Data Understanding*

   Langkah awal yang dilakukan adalah memahami dataset yang digunakan, yaitu MIMIC-IV. Dataset ini berisi berbagai informasi medis pasien yang dirawat di ICU, seperti tekanan darah, kadar kreatinin, dan faktor klinis lainnya. Sebelum masuk ke tahap pemodelan, analisis awal dilakukan untuk melihat apakah ada pola tertentu dalam data. Misalnya, apakah ada hubungan antara potasium dengan kemungkinan gagal jantung? Selain itu, diperiksa juga apakah ada data yang hilang (*missing values*) atau nilai yang tidak wajar (*outliers*) yang dapat mempengaruhi hasil analisis.

3. *Data Preparation*

   Setelah memahami struktur dataset, langkah berikutnya adalah menyiapkan data agar siap digunakan dalam pemodelan. Beberapa tahapan yang dilakukan meliputi:

   1. Membersihkan data dengan mengatasi nilai yang hilang dan menghapus atau mengganti nilai pencilan jika diperlukan.
   2. Mengubah data kategori menjadi numerik menggunakan teknik *encoding*, sehingga bisa dipahami oleh model machine learning.

3. Melakukan normalisasi atau standarisasi, agar skala setiap fitur seimbang dan tidak memengaruhi proses pembelajaran model.
4. Membagi dataset menjadi data latih (80%) dan data uji (20%), supaya model dapat diuji pada data yang belum pernah dilihat sebelumnya.
5. Menggunakan SMOTE untuk menangani data *imbalance*

4. Data Modelling

   Dalam penelitian ini, tiga algoritma yang digunakan adalah *Logistic Regression*, *Random Forest*, dan *Support Vector Machine* (SVM).

   1. *Random Forest* bekerja dengan membangun banyak *decision tree* dan mengambil hasil prediksi yang paling sering muncul. Metode ini membantu mengurangi kemungkinan model terlalu menyesuaikan diri dengan data latih (*overfitting*).
   2. SVM (*Support Vector Machine*) berfungsi dengan menemukan garis pemisah terbaik (*hyperplane*) untuk memisahkan data menjadi dua kategori. Algoritma ini sangat efektif dalam menangani data yang kompleks dan berdimensi tinggi.
   3. *Logistic regression* digunakan untuk klasifikasi biner dengan cara memodelkan probabilitas suatu sampel termasuk dalam salah satu dari dua kelas.

5. Evaluation

   Setelah model selesai dibuat, langkah terakhir adalah mengevaluasi performanya. Beberapa metrik yang digunakan antara lain:

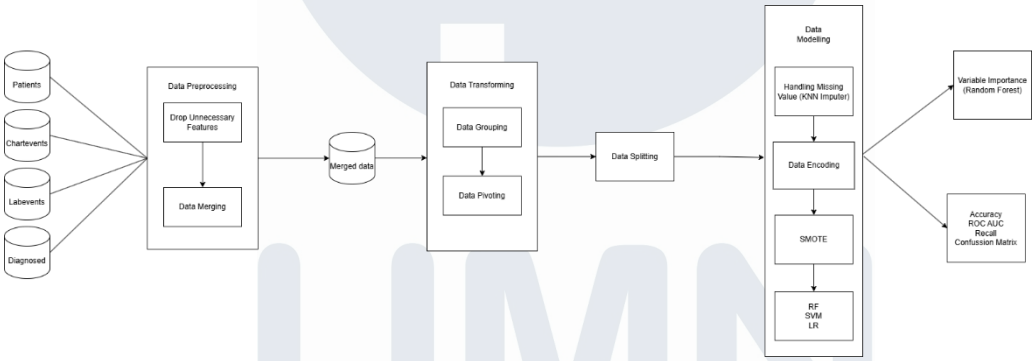
   1. Akurasi
   2. ROC-AUC
   3. Recall
   4. Confussion Matrix

Hasil evaluasi ini menentukan apakah model yang dikembangkan sudah cukup baik atau masih perlu diperbaiki

3.3 Teknik Pengumpulan Data

Penelitian ini menggunakan data "*MIMIC-IV-ECG-Ext-ICD: Diagnostic labels for MIMIC-IV-ECG*" yang didapatkan melalui situs *physionet*. Untuk mendapatkan data tersebut harus melalui beberapa tahap terlebih dahulu. Tahap pertama harus melakukan registrasi di situs physionet tersebut. Setelah itu, tes agar mendapatkan sertifikat "*COLLABORATIVE INSTITUTIONAL TRAINING INITIATIVE* (CITI PROGRAM)". Terdapat dua modul yang berisikan soal pilihan ganda untuk menyelesaikan kelas tersebut. Setelah dua modul berhasil diselesaikan, sertifikat dapat dicetak atau disimpan ke dalam penyimpanan gawai. Setelah mendapatkan sertifikat tersebut, kembali lagi ke halaman *physionet* untuk memasukan sertifikat tersebut. Dan tahap terakhir menunggu hingga semua diverifikasi oleh situs dan data dapat diperoleh.

3.4 Teknik Analisis Data



Gambar 3. 1 Research Freamework

Gambar 3.1 menjabarkan bagaimana proses penelitian ini berlangsung. Penelitian ini dilakukan melalui beberapa tahapan yang terstruktur agar hasil analisis lebih akurat dan dapat diandalkan. Tahap pertama dimulai dari pengambilan data yang bersumber dari database MIMIC-IV. Data yang digunakan mencakup beberapa tabel penting seperti *diagnosed_icd, chartevents, patients,* dan *labevents*.

Setelah data berhasil dikumpulkan, dilakukan proses *pre-processing*, yang melibatkan penghapusan fitur-fitur yang tidak relevan dan penggabungan data dari keempat tabel menjadi satu dataset utuh untuk memudahkan analisis.

Tahap berikutnya adalah *data transforming*. Langkah awal dalam tahap ini adalah pengelompokan data (*data grouping*) dan pemutaran data (*data pivoting*) agar setiap baris merepresentasikan satu pasien unik dengan fitur-fitur klinisnya. Setelah dataset final terbentuk, dilakukan pembagian data *(Data Splitting)* menjadi 80% data latih dan 20% data uji. Pembagian ini bertujuan agar model dapat diuji pada data yang belum pernah dilihat sebelumnya.

Pada *data modelling*, sebuah *pipeline* terstruktur diterapkan. Proses ini dimulai dengan *preprocessing* pada data latih, yang mencakup penanganan nilai yang hilang melalui imputasi dengan *knn-imputer* dan mengubah data kategorikal menjadi numerik melalui *encoding*. Melakukan *preprocessing* setelah pembagian data sangat penting untuk mencegah kebocoran data (*data leakage*). Selanjutnya, untuk mengatasi masalah kelas yang tidak seimbang, teknik *SMOTE* diterapkan hanya pada data latih yang telah di-preprocess.

Data latih yang sudah bersih dan seimbang inilah yang kemudian digunakan untuk melatih tiga model: *Random Forest, Logistic Regression,* dan *Support Vector Machine (SVM).*

Lalu masuk ketahap terakhir, yaitu Evaluation. Kinerja dari setiap model yang telah dilatih diukur menggunakan data uji. Evaluasi dilakukan dengan menganalisis *Confusion Matrix* untuk memahami tipe kesalahan prediksi, serta menghitung metrik kuantitatif utama seperti **Akurasi**, **ROC-AUC**, dan ***Recall* (Sensitivitas)**. Setelah itu dilakukan analisis **tingkat kepentingan fitur (*feature importance*)** dengan fokus pada model ***Random Forest***, karena kemampuannya untuk memberikan skor kepentingan pada setiap fitur. Setelah model dilatih, nilai kepentingan dari setiap fitur diekstrak dan diurutkan untuk menentukan kontributor utama dalam prediksi. Hasil analisis ini memberikan wawasan mendalam mengenai variabel-variabel kunci yang dapat menjadi fokus perhatian bagi tenaga medis.

Dengan kombinasi metrik ini, peneliti dapat membandingkan performa ketiga model secara menyeluruh untuk menentukan model mana yang paling efektif dalam memprediksi risiko gagal jantung.

3.4.1 Variabel Penelitian

Dalam penelitian ini, variabel dibagi menjadi dua kategori utama: variabel dependen (target yang diprediksi) dan variabel independen (fitur-fitur yang digunakan untuk prediksi).

1. **Variabel Dependen (Y)**

   Variabel dependen dalam penelitian ini adalah ***heart_failure*** (gagal jantung). Variabel ini bersifat kategorikal biner yang merepresentasikan kondisi pasien, dengan rincian sebagai berikut:

   1. **Nilai 1**: Menandakan pasien terdiagnosis menderita gagal jantung.
   2. **Nilai 0**: Menandakan pasien tidak terdiagnosis menderita gagal jantung.

   Status gagal jantung ini tidak tersedia secara langsung, melainkan diturunkan selama tahap persiapan data. Label ini ditetapkan dengan cara mengidentifikasi pasien yang memiliki catatan diagnosis dengan kode ICD (*International Classification of Diseases*) yang berhubungan dengan "*Heart Failure*" pada dataset MIMIC-IV.

2. **Variabel Independen (X)**

   Variabel independen adalah fitur-fitur klinis dan demografis yang digunakan oleh model untuk memprediksi variabel dependen. Fitur-fitur ini dipilih berdasarkan relevansinya dengan kondisi kardiovaskular dan ketersediaannya dalam dataset MIMIC-IV. Berikut adalah rincian variabel independen yang digunakan:

   1. **Data Demografis:**

1. *anchor_age*: Usia pasien, merupakan faktor risiko umum yang signifikan dalam banyak penyakit, termasuk kondisi kardiovaskular.
2. *gender*: Jenis kelamin pasien.

2. **Tanda-tanda Vital (*Vital Signs*):**
    1. *heart_rate* (Detak Jantung): Denyut jantung yang tinggi merupakan salah satu faktor risiko utama mortalitas pada pasien gagal jantung. Detak jantung secara umum mencerminkan kondisi kesehatan jantung seseorang [83].
    2. *systolic_bp* (Tekanan Darah Sistolik) dan *diastolic_bp* (Tekanan Darah Diastolik): Tekanan darah sistolik yang rendah diketahui sebagai faktor risiko pada pasien gagal jantung di ICU. Kedua komponen tekanan darah ini memberikan gambaran penting mengenai kondisi sistem peredaran darah [84].
    3. *respiratory* (Laju Pernapasan): Kesulitan bernapas adalah gejala umum gagal jantung, sering kali disebabkan oleh penumpukan cairan di paru-paru (edema paru). Perubahan frekuensi pernapasan dapat menjadi indikator penting adanya gangguan kesehatan [85].
    4. *O2_saturation* (Saturasi Oksigen/SpO_2): Menunjukkan kemampuan sistem pernapasan dalam menyalurkan oksigen ke jaringan tubuh. Pemantauan saturasi oksigen sangat krusial, terutama pada pasien dengan gangguan pernapasan atau pasien di ICU [86].

3. **Hasil Laboratorium (Lab Results):**
    1. *potassium* (Kalium): Mineral ini sangat penting untuk fungsi kontraksi otot, termasuk otot jantung, serta menjaga sinyal listrik pada saraf. Kadar potasium yang tidak normal, baik

terlalu tinggi (*hiperkalemia*) maupun terlalu rendah (*hipokalemia*), dapat menyebabkan gangguan irama jantung [85].

2. *sodium* (Natrium): Berperan vital dalam menjaga keseimbangan cairan tubuh dan regulasi tekanan darah. Kelebihan asupan natrium dapat meningkatkan risiko hipertensi dan penyakit jantung [85], [86].
3. *creatinine* (Kreatinin): Kadar kreatinin adalah indikator penting untuk menilai fungsi ginjal. Adanya gangguan fungsi ginjal merupakan salah satu faktor risiko utama yang memperburuk kondisi pasien gagal jantung [85], [86].
4. BNP (*Brain Natriuretic Peptide*): Merupakan biomarker utama yang dilepaskan jantung sebagai respons terhadap stres atau tekanan, yang umum terjadi pada kondisi gagal jantung. Kadar BNP yang tinggi digunakan secara luas untuk membantu diagnosis dan menilai tingkat keparahan gagal jantung [86].