

BAB 1

PENDAHULUAN

1.1. Latar Belakang Masalah

Penyakit jantung merupakan salah satu penyebab kematian tertinggi baik secara global maupun nasional, dan menjadi ancaman serius terhadap kualitas hidup masyarakat. Berdasarkan laporan dari *World Health Organization (WHO)*, diperkirakan penyakit *kardiovaskular* menyumbang sekitar 17,9 juta kematian setiap tahunnya di seluruh dunia [1]. Di Indonesia sendiri, kasus kematian akibat penyakit jantung diperkirakan mencapai lebih dari 295.000 jiwa per tahun, menunjukkan beban kesehatan masyarakat yang sangat besar [2]. Situasi ini menggarisbawahi urgensi perlunya upaya deteksi dini sebagai langkah penting dalam pencegahan dan mitigasi risiko yang lebih besar di masa depan.

Dalam upaya mendukung proses diagnosis yang lebih cepat dan akurat, pendekatan berbasis teknologi seperti *data mining* mulai banyak dimanfaatkan dalam dunia medis [3]. *Data mining* sendiri merupakan teknik untuk mengekstraksi pola tersembunyi dan informasi bermakna dari kumpulan data berukuran besar. Aplikasinya dalam bidang kesehatan mencakup berbagai kebutuhan, mulai dari prediksi penyakit, analisis risiko, hingga klasifikasi dan pengambilan keputusan klinis berbasis data pasien.

Namun demikian, salah satu tantangan utama dalam implementasi *data mining* terletak pada kualitas data yang digunakan. Data yang tidak lengkap, tidak konsisten, atau mengandung *noise* dapat menurunkan performa algoritma klasifikasi meskipun model yang digunakan sudah tergolong canggih [4]. Oleh karena itu, proses *preprocessing* data menjadi tahap yang sangat penting untuk menjamin kesiapan data sebelum digunakan dalam pemodelan [5].

Salah satu pendekatan metodologis yang paling banyak digunakan dalam proses *data mining* adalah *CRISP-DM* (*Cross Industry Standard Process for Data mining*), yang menawarkan alur kerja sistematis dan iteratif. *CRISP-DM* terdiri dari

enam tahapan utama: *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment* [6].

Meskipun metodologi ini bersifat komprehensif, banyak penelitian sebelumnya yang hanya menerapkan sebagian tahapan *preprocessing* tanpa dokumentasi yang memadai. Kondisi inilah yang mendorong pentingnya penelitian yang secara sistematis menyusun dan mengevaluasi skema *preprocessing* secara lengkap dan terstruktur. Penelitian ini akan difokuskan pada tahapan *data preparation* dalam kerangka *CRISP-DM*, serta hanya menggunakan algoritma *K-Nearest Neighbors (KNN)* sebagai model klasifikasi tunggal yang digunakan agar fokus pada analisis dampak dari kombinasi tahapan *preprocessing* yang berbeda terhadap kualitas klasifikasi.

1.2. Rumusan Masalah

Penelitian ini dilandasi oleh permasalahan utama terkait kualitas data dalam proses klasifikasi penyakit jantung. Rumusan masalah yang diajukan adalah:

1. Bagaimana menyusun skema *preprocessing* data yang optimal dalam kerangka pendekatan *CRISP-DM*?
2. Sejauh mana pengaruh kombinasi tahapan *preprocessing* data terhadap performa klasifikasi penyakit jantung?
3. Skema *preprocessing* mana yang memberikan performa terbaik dalam meningkatkan kualitas klasifikasi berdasarkan metrik evaluasi seperti *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *AUC*?

1.3. Batasan Penelitian

Penelitian ini dibatasi pada penggunaan algoritma *K-Nearest Neighbors (KNN)* sebagai satu – satunya model klasifikasi untuk mengevaluasi dampak berbagai skema *preprocessing* dalam tahap *data preparation* berdasarkan pendekatan *CRISP-DM*, tanpa membandingkan berbagai model klasifikasi lainnya. Seluruh proses eksperimen dan evaluasi dilakukan menggunakan *RapidMiner* sebagai platform utama, sementara proses *Encoding* data kategorikal dilakukan secara terpisah di *Google Colab* menggunakan Python. Penelitian ini tidak membahas implementasi sistem prediksi nyata atau pengujian dengan data klinis,

melainkan terbatas pada evaluasi eksperimental berbasis data sekunder. Selain itu, ruang lingkup penelitian hanya mencakup kombinasi teknik *preprocessing* seperti imputasi, *Normalisasi*, seleksi fitur, dan *balancing (SMOTE)*, tanpa mengevaluasi variasi parameter dalam tiap teknik secara mendalam. Dengan batasan ini, penelitian diharapkan tetap dapat memberikan gambaran sistematis mengenai peran *preprocessing* terhadap performa klasifikasi penyakit jantung, serta menjadi dasar untuk pengembangan lebih lanjut.

1.4. Tujuan Penelitian

Penelitian ini bertujuan untuk mengoptimalkan *preprocessing data* secara terstruktur menggunakan pendekatan *CRISP-DM* guna meningkatkan kualitas klasifikasi penyakit jantung. Optimalisasi dilakukan melalui penyusunan dan pengujian beberapa skema *preprocessing*, serta evaluasi pengaruh masing-masing skema terhadap performa klasifikasi menggunakan algoritma *K-Nearest Neighbors (KNN)*. Sehingga dapat menentukan skema *preprocessing* mana yang paling optimal dalam meningkatkan kualitas klasifikasi

1.5. Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan kontribusi teoritis dalam memperkuat pemahaman akan pentingnya kualitas data dalam siklus *data mining* serta kontribusi praktis berupa panduan *preprocessing data* yang komprehensif. Sementara itu, dari sisi praktis, hasil penelitian ini menghasilkan panduan eksplisit dalam menyusun skema *preprocessing data* yang dapat direplikasi oleh peneliti maupun praktisi, terutama dalam pengembangan sistem deteksi penyakit berbasis data. Skema yang terbukti paling optimal juga dapat menjadi acuan dalam membangun model klasifikasi yang lebih andal untuk keperluan diagnosis awal penyakit jantung, sehingga mendukung pengambilan keputusan klinis yang lebih tepat dan dini.