BAB 1
PENDAHULUAN

1.1 Latar Belakang Masalah

Penyakit jantung adalah masalah kesehatan masyarakat yang paling signifikan terjadi di dunia. Penyakit kardiovaskular (CVD), yang meliputi jantung koroner, serangan jantung, dan gagal jantung, menjadi penyebab utama kematian global dengan sekitar 19,8 juta jiwa meninggal setiap tahun, atau sekitar sepertiga dari seluruh kematian dunia. Mayoritas dari angka kematian ini (sekitar 85 persen) disebabkan oleh serangan jantung dan *stroke*, dengan sebagian besar terjadi di negara-negara yang memiliki penghasilan rendah hingga menengah [1].

Di Indonesia, beban penyakit jantung juga sangat tinggi dan terus meningkat. Data menunjukkan bahwa negara ini mengalami ratusan ribu kematian akibat penyakit kardiovaskular setiap tahun; misalnya pada tahun 2021 tercatat 765.660 kematian akibat CVD. Faktor risiko yang memperburuk beban penyakit jantung di tingkat nasional adalah pola makan tidak sehat, kurangnya aktivitas fisik, merokok, diabetes, dan hipertensi [2].

Diagnosis dini dan akurat terhadap penyakit jantung menjadi aspek krusial untuk menurunkan angka morbiditas dan mortalitas. Deteksi yang terlambat atau tidak tepat sering kali mengakibatkan komplikasi serius atau kematian, sehingga diperlukan sistem pendukung keputusan dengan basis data yang dapat mempermudah tenaga medis dalam tahap diagnosis awal. *Machine learning* menawarkan solusi potensial dalam membantu klasifikasi dan prediksi penyakit jantung dari data rekam medis yang kompleks [3].

Dalam bidang *machine learning*, algoritma *Logistic Regression* (LR) dan *Random Forest* (RF) merupakan dua metode yang sering diterapkan dalam tugas klasifikasi biner, termasuk dalam prediksi penyakit jantung. Metode *Logistic Regression* dikenal karena kemampuannya dalam memberikan interpretasi model yang baik serta asumsi hubungan linier antar variabel. Sementara itu, *Random Forest* memiliki keunggulan dalam memodelkan hubungan non-linear serta menangani kompleksitas fitur melalui pendekatan ensembel yang menghasilkan kinerja prediksi yang tinggi. Sejumlah studi terdahulu menunjukkan bahwa kedua algoritma tersebut mampu menghasilkan performa yang bersaing dalam mengklasifikasikan penyakit jantung berdasarkan data klinis pasien [4].

Namun demikian, sebagian besar dataset di bidang medis, termasuk dataset penyakit jantung, kerap menghadapi permasalahan ketidakseimbangan kelas (*class imbalance*), yaitu kondisi ketika jumlah data pada kelas negatif lebih besar dibandingkan dengan kelas positif (pasien yang terdiagnosis penyakit). Ketimpangan distribusi data ini berpotensi menyebabkan model prediksi menjadi lebih condong ke kelas mayoritas, sehingga kemampuan model dalam mengidentifikasi kelas minoritas yang justru krusial, seperti pasien dengan risiko tinggi menjadi berkurang [5].

Untuk mengatasi permasalahan tersebut, berbagai teknik *oversampling* telah dikembangkan, salah satunya adalah SMOTE (*Synthetic Minority Oversampling Technique*) yang banyak digunakan karena terbukti mampu meningkatkan kinerja model. SMOTE menghasilkan data sintetis pada kelas minoritas dengan memanfaatkan kedekatan antar sampel melalui metode tetangga terdekat, sehingga distribusi data menjadi lebih seimbang. Dengan kondisi ini, model dapat mempelajari pola dari kedua kelas secara lebih optimal. Sejumlah penelitian yang mengombinasikan SMOTE dengan *Random Forest* pada dataset penyakit jantung melaporkan adanya peningkatan akurasi, bahkan mencapai lebih dari 80 persen dalam beberapa kasus [6].

Meskipun SMOTE efektif, varian yang lebih canggih seperti *Borderline*-SMOTE dan *K-Means*-SMOTE dapat menghasilkan sampel sintetis yang lebih representatif dan mengurangi risiko *overfitting* serta *noise*. *Borderline*-SMOTE fokus pada titik yang berada di dekat batas keputusan antara kelas mayoritas–minoritas, sedangkan *K-Means*-SMOTE memanfaatkan *clustering* untuk mengeksplorasi struktur lokal dalam data sebelum membuat sampel sintetis, yang mampu meningkatkan kualitas *balancing data*. Pendekatan ini diyakini dapat memperbaiki kinerja model lebih signifikan dibandingkan SMOTE standar [7].

Penggabungan antara algoritma klasik (*Logistic Regression*) dan ensembel (*Random Forest*) dengan teknik *oversampling* lanjutan seperti *Borderline*-SMOTE dan *K-Means*-SMOTE memberikan arah penelitian yang relevan untuk meningkatkan kinerja prediktif klasifikasi penyakit jantung. Hal ini tidak hanya berkontribusi dalam aspek metodologi *machine learning*, tetapi juga membuka peluang aplikasi nyata dalam sistem informasi kesehatan yang mampu memberikan dukungan keputusan klinis secara lebih akurat dan andal [8].

Dengan latar tersebut, penelitian ini mengeksplorasi pemodelan klasifikasi penyakit jantung menggunakan *Logistic Regression* dan *Random Forest*, serta mengevaluasi dampak teknik *Borderline*-SMOTE dan *K-Means*-SMOTE dalam

menangani ketidakseimbangan data. Tujuannya adalah menghasilkan model klasifikasi yang tidak hanya memiliki metrik evaluasi tinggi (akurasi, *recall*, *F1-score*), tetapi juga lebih sensitif dalam mendeteksi kelas minoritas (kasus penyakit jantung), sehingga layak digunakan sebagai referensi teknologi pendukung diagnosis dini.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan, penelitian ini mengidentifikasi beberapa permasalahan utama yang perlu dikaji, yaitu:

1. Bagaimana performa dari algoritma *Logistic Regression* dan *Random Forest* dalam melakukan klasifikasi penyakit jantung?
2. Bagaimana pengaruh penerapan teknik *Borderline SMOTE* dan *K-Means SMOTE* terhadap performa klasifikasi penyakit jantung?
3. Bagaimana perbandingan kinerja algoritma *Logistic Regression* dan *Random Forest* dengan teknik *Borderline SMOTE* dan *K-Means SMOTE* dalam melakukan klasifikasi penyakit jantung berdasarkan metrik evaluasi *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *Area Under the Curve (AUC)*?

1.3 Batasan Permasalahan

Penelitian ini dibatasi pada sejumlah aspek berikut:

1. Penelitian ini memiliki ruang lingkup yang terbatas pada penerapan dua algoritma *machine learning*, yaitu *Logistic Regression* dan *Random Forest*, dalam pemodelan klasifikasi penyakit jantung. Kajian yang dilakukan difokuskan pada evaluasi kemampuan kedua metode tersebut dalam menangani permasalahan klasifikasi biner, yakni membedakan antara kondisi adanya penyakit jantung dan tidak adanya penyakit jantung.
2. Data yang digunakan dalam penelitian ini merupakan data sekunder berupa dataset *Indicators of Heart Disease* yang diperoleh dari sumber publik pada *Kaggle*, dengan total sekitar 319.795 baris dan 18 kolom. Variabel yang dianalisis mencakup informasi demografis dan kondisi kesehatan, seperti usia, jenis kelamin, *Body Mass Index* (BMI), kebiasaan merokok maupun

aktivitas fisik, pola durasi tidur, serta riwayat penyakit penyerta seperti diabetes dan *stroke*.

3. Permasalahan ketidakseimbangan kelas pada dataset ditangani secara khusus menggunakan dua teknik *oversampling*, yaitu *Borderline SMOTE* dan *K-Means SMOTE*.
4. Perbandingan kinerja model dilakukan berdasarkan metrik evaluasi klasifikasi, yaitu *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *Area Under the Curve* (AUC).

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk mencapai beberapa hal, di antaranya:

1. Mengembangkan performa model *machine learning* berbasis algoritma *Logistic Regression* dan *Random Forest* dalam melakukan klasifikasi penyakit jantung berdasarkan data klinis pasien.
2. Menganalisis pengaruh penerapan teknik penanganan ketidakseimbangan data, yaitu *Borderline SMOTE* dan *K-Means SMOTE*, terhadap peningkatan performa model klasifikasi penyakit jantung.
3. Membandingkan dan mengevaluasi kinerja algoritma *Logistic Regression* dan *Random Forest* yang dikombinasikan dengan teknik *Borderline SMOTE* dan *K-Means SMOTE* berdasarkan metrik evaluasi *Accuracy*, *Precision*, *Recall*, *F1-score*, dan *Area Under the Curve* (AUC), sehingga dapat diketahui kombinasi model dan teknik oversampling yang memberikan hasil paling optimal.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat menghasilkan berbagai manfaat, baik dalam konteks akademik maupun praktis, antara lain:

1. Memberikan kontribusi berupa referensi serta wawasan ilmiah dalam bidang *machine learning*, khususnya terkait implementasi algoritma *Logistic Regression* dan *Random Forest* yang dikombinasikan dengan teknik *oversampling Borderline SMOTE* dan *K-Means SMOTE* pada permasalahan klasifikasi penyakit jantung.

2. Menjadi rujukan dalam menentukan kombinasi model klasifikasi dan teknik *oversampling* yang optimal untuk meningkatkan ketepatan deteksi penyakit jantung. Selain itu, hasil penelitian ini juga dapat dimanfaatkan sebagai dasar dalam pengembangan sistem pendukung keputusan di sektor kesehatan, terutama dalam membantu identifikasi dini penyakit jantung berbasis data klinis pasien.

3. Menjadi referensi bagi penelitian selanjutnya yang berfokus pada eksplorasi penerapan teknologi *machine learning* dalam klasifikasi penyakit jantung maupun penyakit lainnya, dengan memanfaatkan model algoritma yang lebih kompleks, seperti *deep learning*, serta berbagai variasi teknik penanganan ketidakseimbangan data untuk menghasilkan model yang lebih optimal.

1.6 Sistematika Penulisan

Sistematika penulisan laporan adalah sebagai berikut:

- Bab 1 PENDAHULUAN
  Bab ini memaparkan latar belakang penelitian, rumusan masalah, batasan penelitian, tujuan penelitian, manfaat penelitian, serta sistematika penulisan. Bagian ini bertujuan untuk memberikan gambaran umum mengenai urgensi penelitian sekaligus menjelaskan kerangka pemikiran yang digunakan sebagai landasan dalam menjawab permasalahan yang telah dirumuskan.

- Bab 2 LANDASAN TEORI
  Bab ini menyajikan kajian teori yang berkaitan dengan topik penelitian, meliputi tinjauan terhadap penelitian terdahulu, faktor-faktor yang memengaruhi terjadinya penyakit jantung, konsep dasar *machine learning*, metode penyeimbangan data seperti *Borderline SMOTE* dan *K-Means SMOTE*, algoritma yang digunakan, serta metrik yang digunakan dalam evaluasi model.

- Bab 3 METODOLOGI PENELITIAN
  Bab ini menjelaskan metode yang digunakan dalam penelitian, mencakup desain penelitian, proses pengumpulan dan pemilihan dataset, teknik *oversampling* yang diterapkan, serta tahapan analisis yang dilakukan. Selain itu, bab ini juga memaparkan metrik evaluasi yang digunakan untuk

mengukur kinerja algoritma secara komprehensif sesuai dengan tujuan penelitian.

- Bab 4 HASIL DAN DISKUSI
  Bab ini menguraikan hasil penerapan metode pada model klasifikasi yang telah dikembangkan, mencakup analisis data, proses pelatihan dan pengujian model, evaluasi kinerja model, serta pembahasan terhadap hasil yang diperoleh dalam penelitian.
- Bab 5 KESIMPULAN DAN SARAN
  Bab ini memaparkan kesimpulan yang dihasilkan dari penelitian serta rekomendasi yang dapat dijadikan acuan bagi pengembangan penelitian selanjutnya maupun penerapan di lingkungan industri.