

BAB I

PENDAHULUAN

1.1 Latar Belakang

Penyakit Kardiovaskular (PKV) adalah penyakit yang menyerang jantung dan pembuluh darah. Penyakit ini mengakibatkan berhentinya detak jantung manusia sehingga PKV menjadi salah satu penyakit mematikan di dunia, termasuk di Indonesia (Ciumärnean dkk, 2022). Faktor risiko utama yang menyebabkan PKV meliputi gaya hidup yang tidak sehat seperti pola makan yang buruk, kurangnya aktivitas fisik, merokok, dan konsumsi alkohol. Selain itu, terdapat juga faktor fisiologis yang berisiko PKV seperti hipertensi, diabetes, obesitas, dan kadar kolesterol tinggi (Rahim dkk., 2021). Menurut data dari *World Health Organization* (WHO) dalam setahun diperkirakan sebanyak 17,9 juta orang meninggal akibat PKV. Dari jumlah tersebut, 85% disebabkan oleh serangan jantung dan stroke (WHO, 2021). Sementara itu di Indonesia, menurut (Badan Kebijakan Pembangunan Kesehatan (BKPK) Kemenkes, 2023) tercatat sebanyak 877.531 kasus penyakit jantung berdasarkan diagnosis dokter dan pada Provinsi Bengkulu sudah tercatat 6.571 kasus.

Terdapat beberapa upaya pencegahan yang telah dilakukan di bidang medis untuk mengatasi kematian akibat penyakit jantung, seperti tindakan pencegahan berbasis teknologi melalui identifikasi penyakit jantung dengan mengembangkan metode prediksi risiko untuk membantu tenaga medis dalam mengembangkan metode yang efektif guna mendeteksi dini risiko PKV (Maharani dkk., 2024). Namun, di Provinsi Bengkulu masih belum menerapkan hal tersebut dikarenakan oleh banyak faktor dan keterbatasan data menjadikan hal tersebut sebagai tantangan untuk membuat teknologi prediktif secara tepat.

Keterbatasan ketersediaan data rekam medis mengenai PKV di Indonesia merupakan salah satu permasalahan dalam pengembangan sistem prediksi risiko PKV. Hal ini dikarenakan pengaksesan terhadap data pasien di rumah sakit yang cukup terbatas karena alasan privasi data pasien dan prosedur administratif yang kompleks. Berdasarkan kondisi tersebut, penelitian ini bertujuan untuk membangun model prediksi PKV menggunakan dataset gabungan antara dataset lokal Indonesia sebagai data private yang diambil langsung dari RSUD M. Yunus Kota Bengkulu dengan dataset publik sebagai data public. Pemilihan pengambilan data lokal Indonesia dipenelitian ini didasari oleh RSUD M. Yunus Kota Bengkulu merupakan rumah sakit rujukan utama penyakit jantung di Provinsi Bengkulu, sehingga data dari RSUD M. Yunus dapat mencerminkan kondisi masyarakat

yang terdiagnosis PKV di Provinsi Bengkulu. Selain itu, hingga saat ini belum ada penelitian serupa di RSUD M. Yunus yang berfokus pada prediksi PKV. Penelitian ini menggunakan *Shapley Additive Explanations* (SHAP) untuk menganalisis atribut mana yang paling signifikan dalam pemeriksaan PKV. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi nyata melalui dataset lokal.

Pengembangan model prediksi risiko PKV berbasis algoritma *machine learning* menjadi sangat penting dalam upaya pencegahan dan intervensi dini (Dritsas dkk., 2022). *Machine learning* (ML) adalah cabang kecerdasan buatan yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa diprogram secara eksplisit (Balakrishnan dkk, 2021). Dalam bidang kesehatan, ML digunakan untuk berbagai penelitian seperti klasifikasi dan prediksi untuk menganalisis data medis, mengenali pola, dan memberikan prediksi yang akurat mengenai berbagai kondisi kesehatan (Dalal dkk., 2023). Salah satu algoritma yang sering digunakan dalam ML adalah *Random Forest* (Pal dan Parija, 2021).

Algoritma Random Forest (RF) memiliki keunggulan dalam menangani data yang kompleks, beragam, dan memiliki mekanisme dengan cara mengambil sampel secara acak dan menggabungkan hasil dari beberapa pohon keputusan, RF mampu menghasilkan prediksi yang lebih akurat dan mengurangi risiko *overfitting* dibandingkan dengan metode lainnya, seperti *decision tree*, *naïve bayes*, *logistic regression*, dan *support vector machine* (Pal dan Parija, 2021). Dalam bidang kesehatan, RF telah digunakan untuk mendukung diagnosis penyakit, tingkat keparahan penyakit, serta personalisasi pengobatan berdasarkan data pasien (Dalal dkk., 2023).

Penelitian terkait yang membahas mengenai prediksi PKV menggunakan data publik yang digunakan pada penelitian ini juga dilakukan oleh (Jaouja Maiga dkk, 2019) dengan membandingkan beberapa algoritma ML seperti *Random Forest*, *Naïve Bayes*, *K-Nearest Neighbors* (KNN), dan *Logistic Regression* untuk memprediksi risiko PKV berdasarkan data rekam medis. Data yang digunakan berasal dari data publik (Svetlana Ulianova, 2019) yang berjumlah 70.000 data pasien dengan 12 atribut. Pengolahan data dilakukan dengan pendekatan CRISP-DM, mencakup pembersihan data (menghapus nilai tak wajar), pembuatan atribut indeks massa tubuh (BMI), dan normalisasi menggunakan *MinMaxScaler*. Evaluasi dilakukan menggunakan akurasi, ROC curve, *confusion matrix*, dan *classification report*, baik dengan metode *10 fold cross-validation* maupun *train-test split* (80:20). Hasilnya menunjukkan bahwa algoritma RF memberikan performa terbaik dengan akurasi 73% dan sensitivitas 80%. Namun, diperlukan pembahasan lebih lanjut

mengenai optimasi RF guna meningkatkan performa model. Model RF dapat dioptimasi dengan cara melakukan *tuning hyperparameter* untuk mencari parameter yang tepat seperti jumlah pohon dalam hutan (*n_estimators*), kedalaman maksimum pohon (*max_depth*), dan jumlah atribut yang dipilih pada setiap percabangan (*max_features*) serta *feature selection* dan *balancing* data untuk meningkatkan performa RF (Salwa Alexita dkk,2025).

Perbedaan penelitian yang akan dilakukan dengan penelitian terdahulu adalah di bagian fokus penelitian dan dataset yang digunakan. Penelitian sebelumnya (Jaouja Maiga dkk, 2019) berfokus menganalisis keefektifan berbagai algoritma ML dalam memprediksi PKV menggunakan dataset publik. Sedangkan penelitian ini menggunakan data lokal rekam medis PKV yang diambil langsung di RSUD M.Yunus Kota Bengkulu yang digabungkan dengan dataset publik untuk meningkatkan kualitas dan performa model prediksi PKV dari dataset lokal yang terbatas. Serta, menganalisis pengaruh perbedaan jumlah atribut dari data private dan public menggunakan 3 Eksperimen penelitian. Oleh karena itu, penelitian ini dapat membuktikan apakah Eksperimen yang dilakukan bisa menghasilkan hasil model yang baik dan meningkatkan kualitas model dengan melibatkan distribusi lain.

Penelitian ini diharapkan mampu berkontribusi untuk mengatasi permasalahan jumlah data penelitian yang sedikit dengan menggabungkan data dari distribusi lain untuk meningkatkan performa model prediksi PKV menggunakan 3 Eksperimen penelitian (pengurangan jumlah atribut, membandingkan performa data serta menggabungkan data dari 2 sumber berbeda) untuk menjawab permasalahan tersebut. Hal ini dilakukan agar evaluasi performa model tetap relevan dan dapat diterapkan secara praktis dalam lingkungan medis. Selain itu, temuan terkait keterbatasan ketersediaan atribut pada data private juga dapat menjadi masukan bagi pengambil kebijakan, khususnya dalam perbaikan sistem pencatatan rekam medis yang lebih terstandar dan komprehensif untuk mendukung penerapan teknologi prediktif di sektor kesehatan.

1.2 Tujuan Penelitian

Tujuan yang ingin dicapai pada penelitian ini adalah sebagai berikut:

- a. Membangun model prediksi risiko Penyakit Kardiovaskular menggunakan parameter gaya hidup (merokok, alkohol, aktivitas fisik) dan parameter fisiologis (usia, jenis kelamin, tekanan darah, kadar kolesterol, gula darah, berat dan tinggi badan).

- b. Mengidentifikasi atribut yang paling signifikan untuk mendeteksi Penyakit Kardiovaskular menggunakan metode *Shapley Additive Explanations* (SHAP)
- c. Meningkatkan kualitas dan performa model yang memiliki keterbatasan jumlah data (dataset sedikit) dengan menambahkan data dari distribusi lain (dataset publik).

1.3 Manfaat Penelitian

Manfaat yang ingin diberikan pada penelitian ini adalah

- a. Mengidentifikasi faktor utama risiko penyakit kardiovaskular melalui model prediksi berbasis parameter gaya hidup dan fisiologis yang relevan dengan kondisi masyarakat.
- b. Menghasilkan Eksperimen yang efektif untuk meningkatkan performa model pada dataset private yang berjumlah sedikit dengan memanfaatkan dataset publik untuk membuat dataset lebih bervariasi (kombinasi data multisumber).



SEKOLAH PASCASARJANA