BAB II

LANDASAN TEORI

2.1 Analisis Prediksi

Analisis prediksi merupakan proses analisis data yang bertujuan memperkirakan suatu kondisi berdasarkan pola yang terdapat pada data historis (Sari et al., 2023). Dalam data mining, prediksi digunakan untuk membantu pengambilan keputusan dengan memanfaatkan hubungan antara variabel masukan dan variabel target. Analisis prediksi tidak hanya melihat data sebagai catatan masa lalu, tetapi juga mengolah data tersebut menjadi dasar untuk memperkirakan kemungkinan kejadian pada data baru.

Pada penelitian ini, analisis prediksi digunakan untuk memperkirakan risiko penyakit jantung pada pasien. Data pasien berperan sebagai variabel masukan, sedangkan status penyakit jantung berperan sebagai variabel target. Variabel masukan meliputi usia, jenis kelamin, tipe nyeri dada, tekanan darah istirahat, kadar kolesterol, gula darah puasa, hasil elektrokardiografi, detak jantung maksimum, angina akibat aktivitas, oldpeak, slope, ca, dan thal. Variabel target menunjukkan dua kelas, yaitu pasien tidak memiliki indikasi penyakit jantung dan pasien memiliki indikasi penyakit jantung.

Analisis prediksi dalam penelitian ini termasuk prediksi berbasis klasifikasi karena keluaran model berbentuk kelas, bukan angka kontinu. Model tidak menghitung nilai risiko dalam bentuk skor numerik, melainkan menentukan apakah pasien masuk ke kelas 0 atau kelas 1. Kelas 0 menunjukkan tidak ada indikasi penyakit jantung, sedangkan kelas 1 menunjukkan ada indikasi penyakit

jantung. Oleh karena itu, teknik yang digunakan harus mampu mempelajari pola klasifikasi dari data latih dan menerapkannya pada data uji.

Proses analisis prediksi diawali dengan menyiapkan data yang relevan. Data yang digunakan harus melalui tahap preprocessing agar model dapat bekerja secara optimal. Tahapan ini mencakup pemeriksaan missing value, pemeriksaan data ganda, transformasi target, pemeriksaan tipe data, dan pembagian data menjadi data latih serta data uji. Preprocessing penting karena kualitas data sangat memengaruhi kualitas prediksi. Data yang tidak bersih dapat menyebabkan model membaca pola yang salah.

Setelah data siap, model dibangun menggunakan algoritma *Random Forest Classifier*. Algoritma ini mempelajari pola hubungan antara atribut pasien dan target penyakit jantung. Data latih digunakan untuk melatih model, sedangkan data uji digunakan untuk mengukur kemampuan model dalam memprediksi data baru. Pemisahan data latih dan data uji diperlukan agar evaluasi model tidak hanya menggambarkan kemampuan menghafal data, tetapi juga kemampuan generalisasi terhadap data yang belum pernah dipelajari.

Secara umum, hubungan dalam analisis prediksi dapat dituliskan sebagai berikut:

$$Y = f(X)$$

$$X = \{x1, x2, x3, ..., xn\}$$

$$Y = \{0, 1\}$$

Keterangan:

Y adalah hasil prediksi risiko penyakit jantung.

f(X) adalah fungsi model yang dibentuk oleh algoritma *Random Forest Classifier*.

X adalah kumpulan atribut pasien yang digunakan sebagai variabel masukan.

0 menunjukkan pasien tidak memiliki indikasi penyakit jantung.

1 menunjukkan pasien memiliki indikasi penyakit jantung.

Dalam konteks klasifikasi biner, model akan menghasilkan prediksi berdasarkan kelas yang paling sesuai dengan pola data. Jika atribut pasien lebih dekat dengan pola pasien yang memiliki indikasi penyakit jantung, maka model akan memprediksi kelas 1. Sebaliknya, jika atribut pasien lebih dekat dengan pola pasien yang tidak memiliki indikasi penyakit jantung, maka model akan memprediksi kelas 0.

Hasil analisis prediksi perlu dievaluasi menggunakan metrik yang sesuai. Evaluasi dilakukan untuk mengetahui seberapa baik model dalam membedakan pasien berisiko dan tidak berisiko. Metrik evaluasi yang dapat digunakan meliputi accuracy, precision, recall, dan F1-score. Accuracy menunjukkan tingkat ketepatan model secara keseluruhan. Precision menunjukkan ketepatan model ketika memprediksi pasien berisiko. Recall menunjukkan kemampuan model dalam menemukan pasien yang benar-benar berisiko. F1-score menunjukkan keseimbangan antara precision dan recall.

Pada penelitian penyakit jantung, recall menjadi metrik yang penting karena kesalahan dalam mengenali pasien berisiko dapat berdampak serius. Jika model gagal mendeteksi pasien yang sebenarnya memiliki indikasi penyakit jantung, maka pasien tersebut dapat terlambat memperoleh pemeriksaan lanjutan. Oleh karena itu, evaluasi model tidak cukup hanya melihat accuracy, tetapi juga harus memperhatikan precision, recall, dan F1-score.

Analisis prediksi juga dapat diperkuat melalui feature importance. Feature importance digunakan untuk melihat atribut mana yang paling berkontribusi terhadap hasil prediksi. Dalam penelitian ini, feature importance dapat membantu menjelaskan faktor yang paling berpengaruh terhadap prediksi risiko penyakit jantung. Misalnya, atribut seperti chest pain type, thalach, oldpeak, ca, dan thal dapat dianalisis untuk mengetahui kontribusinya terhadap hasil klasifikasi.

Dengan demikian, analisis prediksi dalam penelitian ini tidak hanya bertujuan menghasilkan kelas prediksi, tetapi juga memberikan pemahaman mengenai pola data pasien. Hasil prediksi dapat digunakan sebagai pendukung analisis awal dalam deteksi risiko penyakit jantung. Namun, hasil tersebut tidak dapat dijadikan pengganti diagnosis dokter karena model hanya bekerja berdasarkan pola data yang tersedia. Keputusan medis tetap memerlukan pemeriksaan klinis dan pertimbangan tenaga kesehatan profesional.

2.2 Penyakit Jantung

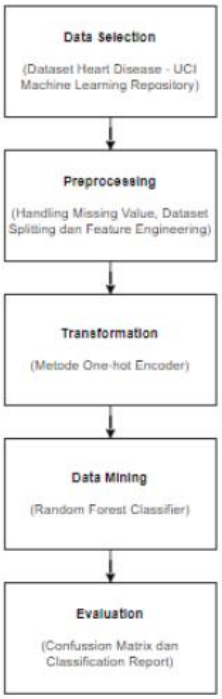
Penyakit jantung adalah kondisi ketika kinerja jantung tidak optimal, sehingga kemampuannya untuk memompa darah dan mengalirkan oksigen ke seluruh tubuh dapat terhambat (Abd Mizwar A. Rahim et al., 2023). Dalam konteks kesehatan masyarakat, penyakit jantung termasuk bagian dari penyakit kardiovaskular. Organisasi Kesehatan Dunia menyatakan bahwa penyakit kardiovaskular menjadi penyebab kematian utama secara global. Pada tahun 2022, penyakit kardiovaskular menyebabkan sekitar 19,8 juta kematian dan mewakili sekitar 32 persen dari seluruh kematian global (Yulianto, 2023). Fakta tersebut

menunjukkan bahwa prediksi risiko penyakit jantung memiliki nilai penting karena dapat mendukung upaya deteksi dini dan pencegahan komplikasi.

Pada penelitian ini, penyakit jantung diposisikan sebagai kelas target yang perlu diprediksi berdasarkan data klinis pasien. Prediksi yang dimaksud bukan pengganti diagnosis dokter, melainkan hasil analisis komputasional yang dapat membantu proses penyaringan awal. Model data mining dapat membaca pola dari data historis, lalu menghasilkan prediksi apakah seorang pasien masuk kategori berisiko atau tidak berisiko berdasarkan atribut yang tersedia.

2.2 *Knowledge Discovery in Database*

Knowledge Discovery In Database (KDD) adalah metode yang digunakan untuk mencari pengetahuan atau informasi yang belum diketahui dari sebuah database (Hidayat dkk., 2024). KDD mencakup tahapan yang saling terkait:



Gambar 2. 1 Knowledge Discovery In Database (KDD)
Sumber: (Alfajr & Defiyanti, 2024)

Proses dalam KDD adalah proses yang terdiri dari rangkaian proses iteratif sebagai berikut :

1. *Data Selection*

   Tahap *data selection* dilakukan dengan memilih data yang relevan dengan tujuan penelitian. Penelitian ini menggunakan *Heart Disease Dataset* yang berasal dari UCI *Machine Learning Repository*. Data yang dipilih terdiri atas 13 atribut prediktor, seperti usia, jenis kelamin, tipe nyeri dada, tekanan darah, kadar kolesterol, gula darah, dan detak jantung maksimum, serta satu atribut target yang menunjukkan ada atau tidaknya indikasi penyakit jantung.

2. *Preprocessing*

   Tahap *preprocessing* bertujuan meningkatkan kualitas data sebelum digunakan dalam pemodelan. Proses ini mencakup pemeriksaan *missing value*, penghapusan data ganda, pemeriksaan tipe data, penanganan nilai yang tidak konsisten, serta pemisahan data menjadi data latih dan data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk mengukur kemampuan model dalam memprediksi data baru.

3. *Transformation*

   Tahap *transformation* dilakukan dengan mengubah data ke dalam format yang sesuai dengan kebutuhan algoritma. Atribut kategorik dapat dikonversi menjadi bentuk numerik menggunakan teknik pengodean, seperti *one-hot encoding*. Variabel target juga ditransformasikan menjadi dua kelas. Nilai 0 menunjukkan tidak terdapat indikasi penyakit jantung, sedangkan nilai 1, 2, 3, dan 4 digabungkan menjadi kelas 1 yang menunjukkan adanya indikasi penyakit jantung.

4. *Data Mining*

   Tahap *data mining* merupakan proses penerapan algoritma untuk menemukan pola dari data. Penelitian ini menggunakan algoritma *Random Forest Classifier*. Algoritma tersebut membangun sejumlah pohon keputusan berdasarkan sampel data dan fitur yang dipilih secara acak. Setiap pohon menghasilkan prediksi, kemudian kelas akhir ditentukan melalui mekanisme *majority voting* atau suara terbanyak.

5. *Evaluation*

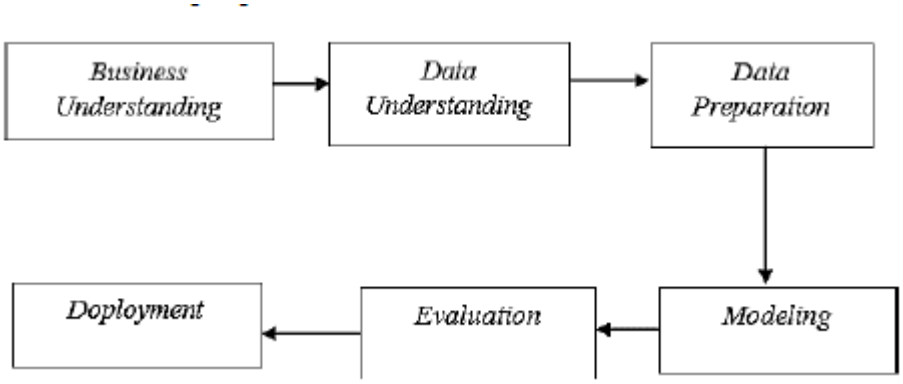
   Tahap *evaluation* bertujuan menilai kualitas pola dan performa model yang telah dibangun. Evaluasi dilakukan menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*. Nilai *recall* pada kelas pasien berisiko perlu mendapat perhatian karena menunjukkan kemampuan model dalam mendeteksi pasien yang memiliki indikasi penyakit jantung. Hasil evaluasi juga digunakan untuk menentukan apakah proses pemodelan telah memadai atau perlu dilakukan perbaikan pada tahap pemilihan data, *preprocessing*, transformasi, atau pengaturan parameter model.

2.2 *Data Mining*

Data mining merupakan suatu proses meringkas suatu pengetahuan menggunakan algoritma untuk mendeteksi pola spesifik, kecenderungan dalam data, dan aturan mekanis (Sari et al., 2023). Data mining bertujuan mengubah data mentah menjadi informasi yang memiliki nilai analitis. Dalam bidang kesehatan, data mining dapat digunakan untuk prediksi penyakit, segmentasi pasien, deteksi anomali, identifikasi faktor risiko, dan pendukung keputusan klinis.

Data mining memiliki beberapa fungsi utama, antara lain klasifikasi, estimasi, prediksi, clustering, asosiasi, dan deteksi outlier. Penelitian ini menggunakan fungsi klasifikasi karena target yang diprediksi berupa kelas biner, yaitu pasien tidak berisiko penyakit jantung dan pasien berisiko penyakit jantung. Klasifikasi bekerja dengan mempelajari hubungan antara atribut masukan dan label target pada data latih, kemudian menggunakan pola tersebut untuk memprediksi label pada data baru.

Pada penelitian ini, data mining digunakan untuk menganalisis prediksi risiko penyakit jantung. Data yang digunakan berbentuk data tabular, yaitu Heart Disease Dataset yang memiliki atribut prediktor dan atribut target. Atribut prediktor berisi karakteristik pasien, sedangkan atribut target menunjukkan ada atau tidaknya indikasi penyakit jantung. Melalui proses data mining, pola hubungan antara atribut pasien dan kelas target dipelajari oleh algoritma *Random Forest Classifier*.



Gambar 2. 2 Proses Data Mining
Sumber: (Munggaran & Nurmalasari, 2025)

Gambar 2.2 menunjukkan bahwa data mining diawali dari data mentah, kemudian dilanjutkan dengan preprocessing, penerapan algoritma, evaluasi, dan interpretasi hasil. Tahapan tersebut menghasilkan pola atau model yang dapat digunakan sebagai dasar pengetahuan. Dalam penelitian ini, pola yang dicari

adalah pola yang dapat membedakan pasien tidak berisiko dan pasien berisiko penyakit jantung.

Gambar tersebut menunjukkan alur metodologi *CRISP-DM (Cross Industry Standard Process for Data Mining)*, yaitu kerangka kerja standar yang sering dipakai dalam penelitian data mining dan machine learning. Alurnya bersifat iteratif, artinya setiap tahap bisa kembali ke tahap sebelumnya jika hasil belum sesuai.

Penjelasan tiap tahapnya sebagai berikut.

1. Business Understanding adalah tahap awal untuk memahami tujuan utama penelitian dari sisi masalah nyata. Pada konteks penelitian penyakit jantung, tahap ini berarti mendefinisikan kebutuhan seperti: bagaimana membantu deteksi dini risiko pasien dan apa dampak klinis yang ingin dicapai. Output dari tahap ini adalah rumusan masalah dan tujuan analitis yang jelas, bukan model.
2. Data Understanding adalah tahap untuk memahami karakteristik data yang tersedia. Peneliti mengecek struktur dataset, jumlah record, jenis variabel, distribusi kelas, serta potensi masalah seperti missing value atau data tidak seimbang. Pada penelitian Anda, ini mencakup pemahaman dataset penyakit jantung dari UCI.
3. Data Preparation adalah tahap paling krusial karena di sini data dibersihkan dan dibentuk agar siap digunakan model. Prosesnya meliputi pembersihan data, transformasi variabel, encoding data kategorik, normalisasi (jika diperlukan), serta pembagian data latih dan data uji. Output tahap ini adalah dataset final yang siap masuk ke algoritma *Random Forest Classifier*.

4. Modeling adalah tahap pembangunan model machine learning. Pada tahap ini algoritma diterapkan ke data latih untuk mempelajari pola hubungan antara fitur pasien (usia, kolesterol, tekanan darah, dll.) dengan label penyakit jantung. Di penelitian Anda, model yang digunakan adalah *Random Forest Classifier* dengan sejumlah pohon keputusan yang digabungkan melalui voting.
5. Evaluation adalah tahap untuk mengukur performa model yang telah dibangun. Evaluasi dilakukan menggunakan metrik seperti accuracy, precision, recall, F1-score, serta confusion matrix. Fokus penting pada kasus medis biasanya ada pada recall, karena kesalahan mendeteksi pasien berisiko (false negative) memiliki dampak yang serius.
6. Deployment adalah tahap akhir untuk menerapkan model ke lingkungan nyata. Dalam konteks penelitian akademik, deployment bisa berupa sistem sederhana, dashboard, atau sekadar simulasi penggunaan model untuk prediksi data baru. Tujuannya agar model tidak hanya berhenti di eksperimen, tetapi bisa digunakan sebagai alat bantu pengambilan keputusan.

Secara keseluruhan, flowchart ini menunjukkan bahwa penelitian data mining bukan hanya soal membangun algoritma, tetapi sebuah siklus lengkap dari pemahaman masalah sampai penerapan hasil model dalam konteks nyata.

Secara umum, tahapan data mining dalam penelitian ini terdiri atas beberapa proses. Pertama, pengumpulan data dilakukan dengan mengambil dataset penyakit jantung dari sumber data sekunder. Kedua, pemahaman data dilakukan dengan memeriksa jumlah data, tipe atribut, rentang nilai, dan distribusi kelas target. Ketiga, preprocessing dilakukan untuk menangani data kosong, data ganda, nilai tidak konsisten, dan format data yang belum sesuai. Keempat,

transformasi dilakukan dengan mengubah target menjadi kelas biner. Kelima, pembagian data dilakukan untuk memisahkan data latih dan data uji. Keenam, pemodelan dilakukan dengan menerapkan algoritma *Random Forest Classifier*. Ketujuh, evaluasi dilakukan dengan confusion matrix, accuracy, precision, recall, dan F1-score. Kedelapan, interpretasi hasil dilakukan dengan membaca feature importance untuk mengetahui atribut yang paling berpengaruh terhadap prediksi.

Data mining memiliki beberapa fungsi utama. Fungsi tersebut meliputi klasifikasi, estimasi, prediksi, clustering, asosiasi, dan deteksi outlier. Setiap fungsi memiliki tujuan yang berbeda. Oleh karena itu, pemilihan fungsi data mining harus disesuaikan dengan bentuk data, tujuan penelitian, dan jenis keluaran yang ingin diperoleh.

1. Klasifikasi

Klasifikasi adalah fungsi data mining yang digunakan untuk menempatkan data ke dalam kelas atau kategori tertentu. Kelas tersebut sudah diketahui sejak awal melalui label pada data latih. Algoritma mempelajari hubungan antara atribut masukan dan label target, kemudian menggunakan pola tersebut untuk menentukan kelas pada data baru. Dalam penelitian ini, klasifikasi digunakan karena target penelitian berbentuk dua kelas, yaitu kelas 0 untuk pasien tidak memiliki indikasi penyakit jantung dan kelas 1 untuk pasien memiliki indikasi penyakit jantung. Contohnya, model mempelajari atribut seperti usia, tekanan darah, kadar kolesterol, detak jantung maksimum, chest pain type, oldpeak, ca, dan thal untuk menentukan apakah seorang pasien masuk kelas berisiko atau tidak berisiko. Algoritma yang digunakan untuk klasifikasi *Random Forest Classifier*.

2. Estimasi

Estimasi adalah fungsi data mining yang digunakan untuk memperkirakan nilai numerik atau nilai kontinu. Berbeda dari klasifikasi yang menghasilkan kelas, estimasi menghasilkan angka. Contohnya adalah memperkirakan kadar kolesterol pasien, memperkirakan tekanan darah, memperkirakan nilai risiko dalam bentuk skor, atau memperkirakan lama rawat pasien berdasarkan atribut klinis tertentu. Estimasi dapat digunakan ketika peneliti membutuhkan keluaran berupa nilai kuantitatif. Pada penelitian ini, fungsi estimasi tidak menjadi fokus utama karena keluaran yang diinginkan bukan angka kontinu, melainkan kelas risiko penyakit jantung.

3. Prediksi

Prediksi adalah fungsi data mining yang digunakan untuk memperkirakan kondisi, nilai, atau kejadian pada masa sekarang maupun masa mendatang berdasarkan pola historis. Prediksi dapat berbentuk prediksi kelas maupun prediksi nilai numerik. Dalam konteks penelitian ini, prediksi digunakan untuk memperkirakan risiko penyakit jantung berdasarkan data pasien. Model dibangun dari data historis yang sudah memiliki label, lalu digunakan untuk memprediksi kelas pada data uji. Kualitas prediksi sangat bergantung pada kualitas data, kecukupan atribut, keseimbangan kelas target, dan kemampuan algoritma dalam membaca pola. Karena penelitian ini menyangkut kesehatan, hasil prediksi harus dipahami sebagai dukungan analisis awal, bukan keputusan klinis final.

4. Clustering

Clustering adalah fungsi data mining yang digunakan untuk mengelompokkan data berdasarkan kemiripan karakteristik tanpa menggunakan

label kelas. Fungsi ini termasuk unsupervised learning karena algoritma tidak diberi target yang benar sejak awal. Dalam bidang kesehatan, clustering dapat digunakan untuk mengelompokkan pasien berdasarkan profil usia, tekanan darah, kolesterol, detak jantung, atau pola pemeriksaan lainnya. Misalnya, pasien dapat dikelompokkan menjadi kelompok risiko rendah, sedang, dan tinggi berdasarkan kemiripan pola data. Pada penelitian ini, clustering tidak digunakan sebagai metode utama karena dataset sudah memiliki label target. Namun, clustering tetap relevan sebagai teori pendukung untuk memahami jenis fungsi data mining.

5. Asosiasi

Asosiasi adalah fungsi data mining yang digunakan untuk menemukan hubungan antaritem atau antaratribut dalam data. Hasil asosiasi biasanya berbentuk aturan if-then. Contohnya, jika pasien memiliki kadar kolesterol tinggi dan tekanan darah tinggi, maka pasien cenderung memiliki risiko penyakit jantung lebih besar. Dalam analisis data transaksi, asosiasi sering digunakan untuk melihat pola pembelian barang. Dalam bidang kesehatan, asosiasi dapat membantu melihat kombinasi faktor yang sering muncul bersama pada pasien tertentu. Meskipun demikian, aturan asosiasi tidak otomatis membuktikan hubungan sebab akibat. Aturan tersebut hanya menunjukkan pola keterkaitan yang ditemukan dalam data.

6. Deteksi Outlier

Deteksi outlier adalah fungsi data mining yang digunakan untuk menemukan data yang memiliki pola berbeda secara mencolok dari sebagian besar data lainnya. Outlier dapat berupa nilai yang sangat tinggi, sangat rendah, tidak wajar, atau jarang muncul. Dalam data kesehatan, outlier dapat terjadi

karena kesalahan input, kesalahan pengukuran, atau kondisi klinis yang memang ekstrem. Contohnya adalah nilai tekanan darah yang terlalu rendah, kadar kolesterol bernilai nol, atau detak jantung maksimum yang berada di luar rentang wajar. Deteksi outlier penting dilakukan sebelum pemodelan karena nilai ekstrem dapat memengaruhi pola yang dipelajari model. Namun, outlier tidak boleh langsung dihapus tanpa pemeriksaan, karena pada data medis outlier bisa saja merepresentasikan kondisi pasien yang penting secara klinis.

Berdasarkan fungsi-fungsi tersebut, penelitian ini berfokus pada klasifikasi dan prediksi. Klasifikasi digunakan karena target data sudah memiliki kelas yang jelas, sedangkan prediksi digunakan karena model diarahkan untuk memperkirakan risiko penyakit jantung pada data pasien yang belum dilihat sebelumnya. Estimasi, clustering, asosiasi, dan deteksi outlier tetap dijelaskan sebagai bagian dari konsep dasar data mining agar landasan teori menjadi lebih lengkap. Pemilihan *Random Forest Classifier* sesuai dengan tujuan penelitian karena algoritma tersebut mampu melakukan klasifikasi biner, menangani banyak atribut, dan memberikan informasi feature importance untuk membantu interpretasi hasil.

2.3 Algoritma *Random Forest Classifier*

Algoritma Random Forest merupakan metode pembelajaran mesin yang efektif untuk analisis data kompleks dan pengklasifikasian (Sibarani, 2025). *Random Forest* sebagai metode yang mengombinasikan banyak decision tree dengan teknik bootstrap sampling dan pemilihan fitur secara acak. Kombinasi

ini membuat model lebih stabil dibandingkan decision tree tunggal karena keputusan akhir tidak bergantung pada satu pohon saja.

Random Forest bekerja dengan dua prinsip utama, yaitu bagging dan random feature selection. Bagging digunakan untuk membuat beberapa subset data latih secara acak dengan pengembalian. Random feature selection digunakan untuk memilih sebagian fitur secara acak pada setiap proses pemisahan node. Kombinasi kedua teknik ini membuat Random Forest lebih stabil, lebih akurat, dan lebih tahan terhadap overfitting dibandingkan decision tree tunggal.

Dalam penelitian ini, *Random Forest Classifier* digunakan untuk memprediksi risiko penyakit jantung berdasarkan atribut pasien, seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, gula darah, detak jantung maksimum, tipe nyeri dada, dan atribut medis lainnya. Model akan mempelajari pola hubungan antara atribut pasien dan kelas target, yaitu pasien tidak berisiko penyakit jantung atau pasien berisiko penyakit jantung.

Secara matematis, misalkan terdapat data latih:

$$D = \{(x1,y1),(x2,y2),...,(xn,yn)\}$$

dengan xi sebagai data atribut pasien dan yi sebagai label kelas target. Random Forest membangun sejumlah B pohon keputusan:

$$h1(x), h2(x), ..., hB(x)$$

Setiap pohon menghasilkan prediksi terhadap data baru x. Prediksi akhir ditentukan dengan majority voting:

$$\hat{y} = \text{mode}\{h1(x), h2(x), ..., hB(x)\}$$

Keterangan:

$\hat{y}$ = hasil prediksi akhir

$hB(x)$ = prediksi dari pohon ke-B

mode = kelas yang memperoleh suara terbanyak.

Sebagai contoh, apabila terdapat 100 pohon keputusan dan 72 pohon memprediksi pasien berisiko penyakit jantung, sedangkan 28 pohon memprediksi pasien tidak berisiko, maka hasil akhir model adalah pasien berisiko penyakit jantung.

Pada proses pembentukan pohon keputusan, Random Forest umumnya menggunakan Gini Index untuk menentukan pemisahan node terbaik. Rumus Gini Index adalah:

$$Gini = 1 - \Sigma(pi^2)$$

atau untuk klasifikasi biner:

$$Gini = 1 - (p0^2 + p1^2)$$

Keterangan:

p0 = proporsi kelas tidak berisiko

p1 = proporsi kelas berisiko.

Semakin kecil nilai Gini Index maka semakin baik kualitas pemisahan node.

Alternatif ukuran impurity adalah Entropy dengan rumus:

$$Entropy = -\Sigma\ pi\ \log2(pi)$$

Semakin kecil nilai entropy menunjukkan bahwa data pada node semakin homogen. Keunggulan Random Forest terletak pada kemampuannya menggabungkan banyak decision tree sehingga hasil prediksi lebih stabil, memiliki akurasi tinggi, mampu menangani data berdimensi besar, relatif tahan

terhadap overfitting, serta dapat menghasilkan nilai feature importance untuk mengetahui atribut yang paling berpengaruh terhadap prediksi penyakit jantung.

Keterbatasan Random Forest adalah interpretasi model lebih kompleks dibandingkan decision tree tunggal dan membutuhkan sumber daya komputasi yang lebih besar ketika jumlah pohon meningkat. Namun demikian, algoritma ini sangat sesuai digunakan dalam penelitian prediksi risiko penyakit jantung karena mampu memberikan performa klasifikasi yang tinggi pada data medis.

Dokumentasi scikit-learn menjelaskan bahwa *Random Forest* merupakan meta estimator yang membangun sejumlah decision tree classifier pada sub-sampel dataset, kemudian menggunakan rata-rata atau voting untuk meningkatkan akurasi prediksi dan mengendalikan overfitting. Dalam klasifikasi, setiap pohon memberikan suara terhadap kelas tertentu. Kelas dengan suara terbanyak menjadi hasil prediksi akhir.

2.3.1 Cara Kerja *Random Forest Classifier*

Cara kerja *Random Forest Classifier* dapat dijelaskan melalui beberapa tahapan. Pertama, algoritma mengambil beberapa sampel data latih secara acak dengan teknik bootstrap. Kedua, untuk setiap sampel bootstrap, algoritma membangun sebuah decision tree. Ketiga, pada setiap proses pemisahan node, algoritma hanya mempertimbangkan sebagian fitur secara acak. Keempat, setiap pohon menghasilkan prediksi. Kelima, hasil akhir ditentukan melalui majority voting untuk klasifikasi.

2.3.2 Konsep Bootstrap Aggregating

Bootstrap aggregating atau bagging adalah teknik pembentukan beberapa subset data latih melalui pengambilan sampel acak dengan pengembalian. Artinya,

satu data dapat muncul lebih dari satu kali dalam subset tertentu. Teknik ini membuat setiap decision tree mempelajari variasi data yang berbeda. Ketika hasil beberapa pohon digabungkan, model menjadi lebih stabil dan tidak mudah dipengaruhi oleh satu pola yang terlalu spesifik pada data latih.

2.3.3 Majority Voting

Majority voting adalah mekanisme pengambilan keputusan akhir dalam *Random Forest Classifier*. Misalnya, jika terdapat 100 pohon keputusan dan 70 pohon memprediksi pasien masuk kelas berisiko, sedangkan 30 pohon memprediksi kelas tidak berisiko, maka hasil akhir model adalah kelas berisiko. Mekanisme ini membuat model lebih kuat karena keputusan akhir berasal dari konsensus banyak pohon.

2.3.4 Parameter Penting *Random Forest*

Kinerja *Random Forest* dipengaruhi oleh beberapa parameter. Parameter tersebut perlu ditentukan secara tepat agar model tidak terlalu sederhana dan tidak terlalu kompleks. Pada penelitian ini, parameter dasar dapat digunakan terlebih dahulu, lalu disesuaikan melalui percobaan jika hasil evaluasi belum memadai.

Tabel 2. 1 Parameter Penting *Random Forest Classifier*

No.	Parameter	Fungsi	Rencana Penggunaan
1	n_estimators	Menentukan jumlah pohon dalam hutan	100 pohon sebagai nilai awal
2	criterion	Menentukan ukuran kualitas pemisahan node	Gini index sebagai nilai awal
3	max_depth	Membatasi kedalaman pohon	None atau nilai tertentu jika terjadi overfitting

4	max_features	Menentukan jumlah fitur yang dipertimbangkan pada setiap split	sqrt sebagai pilihan umum untuk klasifikasi
5	random_state	Menjaga hasil eksperimen agar dapat direplikasi	42
6	class_weight	Mengatur bobot kelas jika distribusi kelas tidak seimbang	balanced jika ketimpangan kelas cukup besar

Sumber: Penelitian, 2026

2.3.5 Kelebihan dan Keterbatasan *Random Forest*

Random Forest memiliki beberapa kelebihan. Algoritma ini mampu menangani banyak fitur, bekerja baik pada data numerik dan kategorik yang sudah dikodekan, relatif tahan terhadap overfitting, dan dapat memberikan informasi feature importance. Feature importance membantu peneliti mengetahui atribut mana yang memiliki kontribusi besar terhadap hasil prediksi. Dalam konteks penyakit jantung, informasi tersebut berguna untuk melihat variabel klinis yang paling kuat dalam membedakan kelas target.

Namun, *Random Forest* juga memiliki keterbatasan. Model ini lebih sulit dijelaskan dibandingkan decision tree tunggal karena keputusan akhir berasal dari banyak pohon. Model juga membutuhkan sumber daya komputasi yang lebih besar jika jumlah pohon, jumlah fitur, dan jumlah data meningkat. Selain itu, *Random Forest* tetap bergantung pada kualitas data. Jika data mengandung bias, nilai hilang yang tidak ditangani, atau label yang salah, maka hasil prediksi juga dapat menurun.

2.6 *Preprocessing* Data

Preprocessing data adalah tahap persiapan data sebelum diterapkan ke dalam algoritma (Putri et al., 1996). Tahap ini penting karena kualitas data sangat memengaruhi kualitas model. Data medis dapat memuat nilai kosong, nilai tidak wajar, perbedaan format, atau ketidakseimbangan kelas. Jika masalah tersebut tidak ditangani, model dapat menghasilkan prediksi yang kurang akurat.

2.6.1 Pembersihan Data

Pembersihan data dilakukan dengan memeriksa nilai kosong, data ganda, dan nilai yang berada di luar batas wajar. Pada data penyakit jantung, pemeriksaan dapat dilakukan terhadap atribut tekanan darah, kolesterol, detak jantung maksimum, dan atribut kategorik seperti chest pain type atau thalassemia. Nilai kosong dapat ditangani dengan menghapus baris, mengisi nilai menggunakan median, atau menggunakan pendekatan lain yang sesuai dengan karakteristik data.

2.6.2 Transformasi Data

Transformasi data dilakukan agar data sesuai dengan kebutuhan algoritma. Dataset merupakan kumpulan dari objek dan sifat atau karakteristik dari suatu objek itu sendiri (atribut) (Bahtiar, 2023). Pada dataset Heart Disease, atribut target asli dapat memiliki nilai 0 sampai 4. Dalam eksperimen klasifikasi biner, nilai 0 dipertahankan sebagai tidak ada penyakit jantung, sedangkan nilai 1 sampai 4 diubah menjadi 1 sebagai ada indikasi penyakit jantung. Transformasi ini mengikuti praktik umum pada eksperimen Cleveland Heart Disease yang membedakan absence dan presence of heart disease (UCI *Machine Learning Repository*, 1988).

2.6.3 Normalisasi Data

Normalisasi data bertujuan menyamakan skala nilai antaratribut. Pada *Random Forest*, normalisasi tidak selalu wajib karena decision tree melakukan pemisahan berdasarkan ambang nilai fitur. Meskipun demikian, normalisasi tetap dapat dilakukan jika penelitian membandingkan *Random Forest* dengan algoritma lain yang sensitif terhadap skala, seperti *K-Nearest Neighbor* atau *Support Vector Machine*. Dalam penelitian ini, normalisasi dapat digunakan sebagai tahap opsional untuk menjaga konsistensi alur data.

2.6.4 Pembagian Data

Pembagian data dilakukan untuk memisahkan data latih dan data uji. Data latih digunakan untuk membangun model, sedangkan data uji digunakan untuk menilai kemampuan model terhadap data yang belum dilihat. Pembagian data yang umum digunakan adalah 80 persen untuk data latih dan 20 persen untuk data uji. Jika jumlah data 303 *records*, maka data latih berjumlah sekitar 242 *records* dan data uji berjumlah sekitar 61 *records*.

2.5 *Tools* Pendukung

Tools pendukung digunakan untuk membantu pengolahan data, pemodelan, evaluasi, dan penulisan laporan. *Tools* yang digunakan harus sesuai dengan kebutuhan penelitian sistem informasi yang berbasis data mining.

2.5.1 Perangkat Keras (*Hardware*)

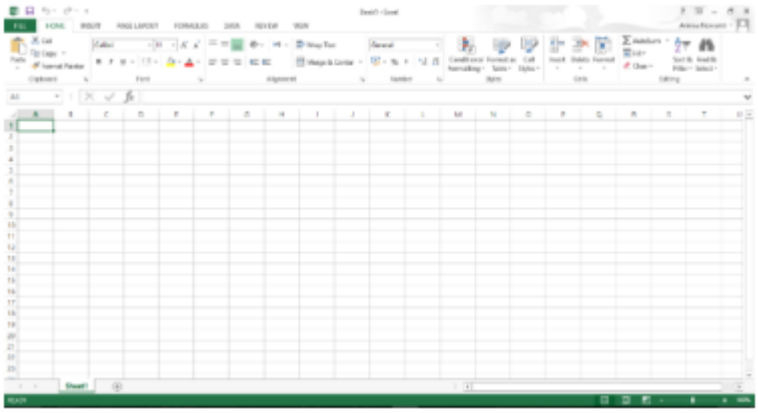
Perangkat keras yang digunakan dalam penelitian ini berupa laptop atau komputer personal. Spesifikasi minimal yang disarankan adalah prosesor Intel Core i3 atau setara, RAM minimal 4 GB, penyimpanan minimal 256 GB, serta

sistem operasi Windows atau Linux. Spesifikasi yang lebih tinggi akan mempercepat proses eksperimen, terutama jika peneliti melakukan percobaan parameter berulang.

2.5.2 Perangkat Lunak (*Software*)

Perangkat lunak yang digunakan dalam penelitian ini terdiri atas aplikasi pengolah data, aplikasi data mining, dan aplikasi pengolah kata. Beberapa perangkat lunak yang digunakan dalam penelitian ini adalah sebagai berikut:

1. Aplikasi Pengolah Angka (*Spreadsheet*) – misalnya Microsoft Excel atau Google Sheets



Gambar 2. 2 Tampilan Awal MS. Excel

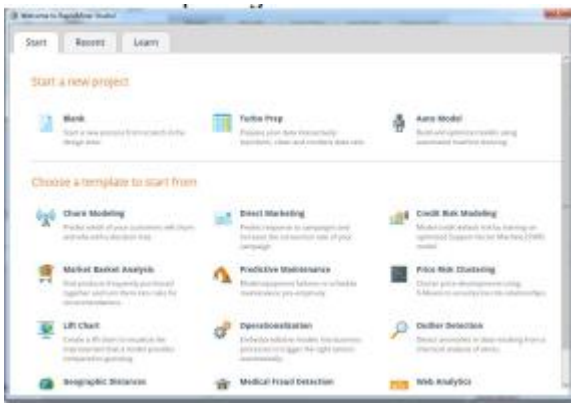
Sumber: (Rianti & Harahap, 2021)

Microsoft excel adalah aplikasi spreadsheet canggih yang bisa digunakan untuk menampilkan data, melakukan pengolahan data, kalkulasi, membuat diagram, laporan, dan semua hal yang berkaitan dengan data yang berupa angka (Rianti & Harahap, 2021). Aplikasi pengolah angka digunakan pada tahap:

2. RapidMiner

Setelah diubah kedalam representasi binerdalam format data tabular, maka

langkah selanjutnya dilakukan eksperimen dan pengujian dengan menggunakan software RapidMiner (Agus Junaidi, 2019).



Gambar 2. 3 Tampilan Awal RapidMiner

Sumber: (Fahmi Julianto et al., 2020)

2.6 Penelitian Terdahulu

Tabel berikut menunjukkan beberapa penelitian terdahulu yang relevan dengan penerapan algoritma Apriori dan *FP-Growth* dalam analisis pola penjualan.

Tabel 2. 2 Penelitian Terdahulu

No.	Penulis & Tahun	Judul	Kontribusi Utama	Hasil Penelitian
1	Ani Dijah Rahajoe, Dandi Azaidane, Brian Akhdan Putra, Mohammad Reza Pahlevy, Bisma Satrio Bimantoro, Fransisco Rivaldi Akash (2026)	Analisis Penyakit Jantung Menggunakan Metode Random Forest	Menggunakan algoritma Random Forest untuk analisis dan klasifikasi penyakit jantung berbasis data medis	Random Forest mampu menghasilkan model klasifikasi dengan tingkat akurasi yang tinggi dan stabil dalam memprediksi penyakit jantung

2	Laura Sari, Annisa Romadloni, Rostika Listyaningrum (2023)	Penerapan Data *Mining* dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma *Random Forest*	Menerapkan Random Forest untuk klasifikasi penyakit berbasis data kesehatan	Random Forest menunjukkan performa baik dalam klasifikasi penyakit dengan tingkat akurasi yang tinggi dibanding metode lain
3	Oktavia Nur Rosita, Niken Rahma Mahmud, Mangara Sitinjak, Yuliana Manobi, Natalia Sibarani (2025)	Klasifikasi Pasien Penyakit Jantung Di Papua Barat Menggunakan Algoritma Random Forest	Mengimplementasikan Random Forest untuk klasifikasi pasien penyakit jantung pada data regional	Model Random Forest mampu mengklasifikasikan pasien dengan baik dan membantu identifikasi risiko penyakit jantung
4	Hidayat, Andi Sunyoto, Hanif Al Fatta (2023)	Klasifikasi Penyakit Jantung Menggunakan Random Forest Clasifier	Menggunakan Random Forest untuk prediksi penyakit jantung berbasis data klinis	Hasil penelitian menunjukkan bahwa Random Forest memiliki performa yang baik dalam klasifikasi data medis
5	Ivana Alhabib, Ahmad Faqih, Fatihanursari Dikananda (2022)	Komparasi Metode *Deep Learning*, *Naïve Bayes* dan *Random Forest* untuk Prediksi Penyakit Jantung	Membandingkan beberapa algoritma machine learning termasuk Random Forest	Random Forest memberikan hasil lebih stabil dibanding Naïve Bayes dan kompetitif terhadap deep learning

6	Nur Halizah Alfajr, Sofi Defiyanti (2024)	Prediksi Penyakit Jantung Menggunakan Metode Random Forest Dan Penerapan Principal Component Analysis (Pca)	Menggabungkan Random Forest dengan PCA untuk reduksi dimensi	PCA membantu meningkatkan efisiensi model, dan Random Forest tetap menunjukkan performa klasifikasi yang tinggi
7	Tiara Sartika (2026)	Prediksi risiko penyakit jantung menggunakan *Random Forest Classifier*	Menerapkan tahapan KDD dan algoritma *Random Forest Classifier* pada *Heart Disease Dataset* yang terdiri atas 303 data, 13 fitur prediktor, dan satu target biner. Penelitian mengevaluasi model menggunakan *confusion matrix*, *accuracy*, *precision*, *recall*, dan *F1-score*, serta menganalisis *feature importance*.	Penelitan masih berlangsung

Sumber: Penelitian, 2026

Berdasarkan penelitian terdahulu, *Random Forest* memiliki dasar teoritis yang kuat dan banyak digunakan dalam klasifikasi penyakit jantung. Penelitian ini berfokus pada penerapan *Random Forest Classifier* untuk memprediksi risiko penyakit jantung dengan data terstruktur. Fokus penelitian bukan hanya melihat akurasi, tetapi juga menjelaskan tahapan *preprocessing*, pembagian data, evaluasi model, dan interpretasi fitur yang berpengaruh.