BAB I
PENDAHULUAN

1.1 Latar Belakang

Pendekatan berbasis data, khususnya algoritma *machine learning*, kini semakin diandalkan untuk mendeteksi dan memprediksi risiko penyakit pada individu (Inna, 2025). Pendekatan ini dinilai efektif karena mampu mengolah data kesehatan yang kompleks dan menemukan pola yang sulit diidentifikasi melalui analisis konvensional (Rahmada & Susanto, 2025). Namun demikian, penerapan *machine learning* pada data medis menghadapi tantangan fundamental, terutama terkait karakteristik data itu sendiri. Data kesehatan biasanya terdiri dari kombinasi variabel numerik dan kategorikal, seperti usia, tekanan darah, status merokok, tingkat aktivitas fisik, maupun riwayat penyakit (Ghosh et al., 2022). Struktur data yang beragam tersebut sering menjadi kendala bagi algoritma klasifikasi tertentu dalam menangkap hubungan antarvariabel secara optimal. Kondisi ini diperparah oleh *missing values*, ketidakkonsistenan data, serta distribusi data yang tidak merata sehingga berdampak pada penurunan kinerja model prediksi (Chen et al., 2023).

Terlepas dari tantangan pada struktur data, permasalahan utama lainnya yang sering muncul dalam data kesehatan adalah ketidakseimbangan kelas (*imbalanced dataset*) di mana pada umumnya, jumlah individu yang tidak menderita penyakit jauh lebih besar dibandingkan jumlah penderita (Salmi et al., 2024). Ketimpangan ini menyebabkan model cenderung bias ke arah kelas mayoritas, yang berakibat pada tingginya angka *false negative* atau kondisi di mana pasien yang sakit diprediksi sehat (Celine et al., 2025). Dengan demikian, proses klasifikasi menjadi sulit karena ketidakseimbangan data dan struktur data yang heterogen. Untuk menghasilkan model yang lebih akurat dan tidak bias terhadap kelas minoritas, diperlukan pendekatan *pre-processing* dan algoritma klasifikasi yang mampu menangani karakteristik data tersebut secara lebih efektif. Merespon tantangan tersebut, pendekatan yang paling ekstensif dirujuk dalam literatur untuk menangani ketidakseimbangan kelas adalah *Synthetic Minority Oversampling Technique* (SMOTE).

Metode SMOTE bekerja dengan membangkitkan sampel sintetis baru melalui interpolasi linier di antara titik data minoritas terdekat (Kovács, 2019). Pendekatan ini diterapkan untuk menghidari duplikasi data secara langsung yang beresiko menyebabkan *overfitting* (Chawla et al., 2002; Fernández et al., 2018). Meskipun SMOTE terbukti efektif pada dataset yang sepenuhnya terdiri dari variabel numerik, metode ini memiliki keterbatasan ketika diterapkan pada data rekam medis yang bersifat heterogen (Sulistiyono et al., 2021). Penerapan perhitungan jarak Euclidean standar yang menjadi basis algoritma SMOTE, sering kali tidak valid pada variabel nominal karena jarak antar kategori tidak dapat dihitung secara linier (Blagus & Lusa, 2013). Akibatnya, pemaksaan metode ini pada data campuran dapat mendistorsi struktur data dan menghasilkan sampel sintetis yang ambigu.

Sebagai upaya awal untuk menangani data campuran tersebut penelitian oleh Chawla et al. (2002) memperkenalkan SMOTE-*Nominal Continuous* (SMOTE-NC) untuk menangani data campuran. Akan tetapi, peneliti tersebut menyebutkan bahwa metode ini masih memiliki kelemahan mendasar karena menghitung jarak antar kategori secara seragam menggunakan median standar deviasi fitur kontinu. Pendekatan seragam ini menyebabkan algoritma gagal menangkap afinitas atau Tingkat keterkaitan spesifik setiap label kategorikal terhadap kelas target.

Merespons keterbatasan tersebut, dikembangkanlah algoritma SMOTE-*Encoded Nominal and Continuous* (SMOTE-ENC), yang dirancang spesifik untuk menjaga integritas fitur pada dataset campuran melalui mekanisme pembangkitan ganda (Mukherjee & Khushi, 2021). Berbeda dengan SMOTE-NC yang hanya mengandalkan modus tetangga untuk variabel nominal, SMOTE-ENC menggunakan *encoding* numerik yang memperhitungkan hubungan kategori dengan kelas minoritas, sembari tetap menjaga interpolasi linier pada data kontinu (Kovács, 2019). Pendekatan ini memastikan bahwa sampel sintetis yang dihasilkan memiliki variabilitas jarak yang lebih realistis dan mampu mempertahankan korelasi antarvariabel heterogen secara lebih akurat dibandingkan SMOTE-NC. Dengan menyeimbangkan distribusi kelas secara proporsional, SMOTE-ENC membantu model untuk memetakan batas keputusan yang lebih presisi dan mengurangi bias terhadap kelas mayoritas (Elreedy & Atiya, 2019).

Keberhasilan penanganan ketidakseimbangan data melalui pendekatan SMOTE-ENC perlu didukung oleh pemilihan algoritma klasifikasi yang mampu mengakomodasi struktur data majemuk tanpa tendensi bias. Dalam konteks ini, algoritma *Quick Unbiased and Efficient Statistical Tree* (QUEST) diajukan sebagai metode pembentukan pohon keputusan yang lebih *robust* dibandingkan pendekatan konvensional. Secara teoritis, pohon klasifikasi yang dibangun melalui algoritma *exhaustive search* seperti yang digunakan pada Classificaton and Regression Tree cenderung memiliki bias dalam memilih variabel yang memiliki banyak titik pemisah (Loh & Shih, 1997). Berbeda pula dengan algoritma *Chi-squared Automatic Interaction Detection* (CHAID) yang cenderung menghasilkan pohon kompleks dengan percabangan ganda (*multi-way splits*), QUEST menawarkan pendekatan pohon biner yang lebih efisien secara komputasi tetapi, tetap *unbiased* dalam seleksi variabel (Lim et al., 2000; Loh, 2014). Hal ini dimungkinkan karena QUEST dirancang untuk memiliki bias yang dapat diabaikan dengan memisahkan proses seleksi variabel dan penentuan titik potong (Loh & Shih, 1997). Mekanisme ini menerapkan uji signifikansi statistik yang ketat, yakni menggunakan uji Chi-square untuk variabel independen bertipe kategorik dan uji ANOVA-F untuk variabel bertipe numerik dalam menentukan prediktor terbaik (Maharani et al., 2017).

Efektivitas QUEST telah dibuktikan dalam berbagai studi komparatif terdahulu yang menunjukkan kinerjanya yang lebih optimal dalam hal akurasi dan efisiensi struktur pohon dibandingkan metode lain. Penelitian oleh Erika et al. (2006) pada dataset dengan variabel campuran menunjukkan bahwa QUEST menghasilkan tingkat kesalahan klasifikasi (*misclassification rate*) yang lebih rendah, yaitu sebesar 0,0014 dibandingkan algoritma CHAID yang mencapai 0,0028. Temuan ini sejalan

dengan studi Faiza et al. (2015) yang mengonfirmasi bahwa meskipun QUEST dan CHAID memiliki tingkat akurasi yang seimbang pada dataset tertentu (87,90%), QUEST lebih unggul dalam menghasilkan struktur pohon biner yang lebih sederhana dan ringkas. Lebih lanjut, Maharani et al. (2017) memperkuat argumen bahwa algoritma QUEST sangat efisien untuk data kompleks karena kemampuannya memproses variabel independen kategorik dan numerik secara simultan tanpa mengorbankan kecepatan komputasi. Dukungan empiris ini menegaskan bahwa QUEST tidak hanya relevan secara teoritis tetapi juga terbukti andal dalam berbagai aplikasi klasifikasi data multidimensi.

Penerapan metode-metode komputasi menjadi sangat krusial ketika diterapkan pada permasalahan kesehatan global seperti penyakit jantung yang hingga kini masih menjadi tantangan kesehatan secara global serta merupakan penyebab kematian utama di berbagai negara, termasuk Indonesia (Tandipanga et al., 2025). World Health Organization (WHO) mencatat bahwa penyakit kardiovaskular atau kelompok penyakit yang memengaruhi jantung dan pembuluh darah, seperti penyakit jantung koroner, stroke, dan hipertensi. Penyakit ini menyebabkan sekitar 19,8 juta kematian pada tahun 2022 atau setara dengan 32% dari total kematian di seluruh dunia (World Health Organization, 2025). Menurut data Riset Kesehatan Dasar (Riskesdas) tren prevalensi penyakit jantung di Indonesia menunjukkan peningkatan signifikan dari 0,50% pada tahun 2013 menjadi 1,5% pada tahun 2018. Lebih spesifik lagi, Survei Kesehatan Indonesia (SKI) 2023 menyoroti Yogyakarta sebagai kota dengan prevalensi tertinggi, mencapai 1,67%. Fakta ini menegaskan bahwa penyakit jantung bukan sekadar isu nasional, melainkan tantangan serius yang memerlukan penanganan komprehensif hingga ke tingkat daerah.

Penyebab penyakit jantung yang bersifat multifaktorial menuntut pemahaman yang komprehensif mengenai faktor-faktor yang memengaruhinya untuk mendukung strategi pencegahan dan penanganan yang efektif (Teshale et al., 2025). Faktor-faktor risiko tradisional penyakit jantung telah dipelajari secara luas. Mengingat sifat penyakit jantung yang multifaktorial, strategi pencegahan yang efektif sangat bergantung pada pemahaman mendalam mengenai faktor risikonya, seperti usia, hipertensi, hiperlipidemia, diabetes, merokok, dan obesitas (Zhou, 2025). Salah satu strategi penting dalam pengendalian penyakit jantung adalah mengidentifikasi dan memahami faktor-faktor penyebabnya (Agraini et al., 2025). Oleh karena itu, diperlukan klasifikasi berdasarkan faktor-faktor yang memengaruhi penyakit jantung. Relevansi integrasi metode SMOTE-ENC dan QUEST dalam klasifikasi penyakit jantung terletak pada urgensi akurasi klinis dan interpretabilitas. Penerapan teknik *oversampling* seperti SMOTE-ENC memegang peranan krusial sebagai tahap pre-processing, khususnya ketika diintegrasikan dengan algoritma pohon keputusan (Sulistiyono et al., 2021). Hal ini memberikan kontribusi langsung terhadap peningkatan sensitivitas model dalam mendeteksi penderita penyakit jantung, di mana kegagalan deteksi (*false negative*) memiliki konsekuensi klinis yang jauh lebih fatal dibandingkan kesalahan prediksi positif (Chicco & Jurman, 2020). Selain itu, relevansi penggunaan QUEST dalam klasifikasi penyakit jantung juga terletak pada efisiensi komputasi dan sifat interpretatif model yang dihasilkan. Dibandingkan

dengan metode *exhaustive search* yang beban komputasinya meningkat secara eksponensial seiring bertambahnya kategori, QUEST terbukti jauh lebih cepat dan efisien, terutama pada dataset berskala besar. Struktur pohon biner yang dihasilkan QUEST mempermudah interpretasi visual bagi tenaga medis, mendukung kebutuhan akan transparansi model prediksi (Lim et al., 2000). Hal ini sangat krusial dalam paradigma *Explainable Artificial Intelligence* di bidang kesehatan, di mana setiap keputusan klinis harus dapat dijelaskan alur logisnya kepada pasien maupun praktisi medis (Tjoa & Guan, 2021). Urgensi penelitian ini terletak pada kebutuhan akan model klasifikasi yang tidak hanya memiliki akurasi tinggi, tetapi juga objektif dalam mengidentifikasi faktor risiko utama penyakit jantung di Yogyakarta. Mengingat adanya ketidakseimbangan proporsi antara penderita dan bukan penderita dalam data SKI 2023, maka implementasi SMOTE-ENC menjadi tahapan krusial untuk memastikan bahwa proses klasifikasi tidak mengalami bias terhadap kelas mayoritas. Dengan menggunakan algoritma QUEST, diharapkan proses klasifikasi menghasilkan *decision tree* yang efisien.

1.2 Batasan Masalah

Batasan masalah pada penelitian ini adalah sebagai berikut:

1. Penyakit jantung yang tergolong pada variabel respon adalah semua kelainan pada jantung termasuk penyakit jantung koroner, gagal jantung (*Decompensatio cordis*), kelainan katup pada jantung, pembengkakan otot jantung, dan berbagai kondisi yang melibatkan penyempitan atau pemblokiran pembuluh darah yang bisa menyebabkan serangan jantung atau strok yang didiagnosis oleh dokter.
2. Pertumbuhan pohon keputusan dibatasi pada kedalaman optimal yang ditentukan saat akurasi model mencapai akurasi maksimum.

1.3 Tujuan dan Manfaat Penelitian

Tujuan penelitian ini adalah sebagai berikut:

1. Menerapkan metode SMOTE-ENC untuk menangani ketidakseimbangan data dan mengukur performa model QUEST dalam mengklasifikasikan kejadian penyakit jantung berdasarkan tingkat akurasi dan kesalahan prediksi (*misclassification rate*).
2. Menentukan faktor risiko utama yang membedakan penderita penyakit jantung dan penduduk sehat di Yogyakarta menggunakan algoritma QUEST

Hasil penelitian diharapkan dapat memberikan manfaat sebagai berikut:

1. Penelitian ini dapat memperkaya literatur terkait penerapan metode SMOTE-ENC pada data kesehatan serta implementasinya pada algoritma QUEST dalam menangani klasifikasi data tidak seimbang.
2. Memberikan wawasan mengenai faktor-faktor risiko yang berkontribusi terhadap kejadian penyakit jantung.

1.4 Landasan Teori

1.4.1 *Machine Learning*

Machine Learning (ML) atau pembelajaran mesin didefinisikan sebagai subbidang dari kecerdasan buatan (*Artificial Intelligence*) yang berfokus pada pengembangan algoritma dan model statistik yang memungkinkan sistem komputer untuk melakukan tugas tertentu tanpa instruksi eksplisit, melainkan dengan mengandalkan pola dan inferensi dari data (Hasan & Abdulazeez, 2024). Berbeda dengan pemrograman konvensional yang bergantung pada aturan-aturan statis yang ditulis secara manual, ML melatih sistem untuk belajar dari data historis (*training data*) guna menghasilkan keputusan atau prediksi yang akurat pada data baru.

Dalam penerapannya, algoritma ML memiliki kemampuan superior dalam menangani data berdimensi tinggi dan kompleksitas non-linear yang sering kali sulit diselesaikan menggunakan metode statistik tradisional. Kemampuan ini membuat ML menjadi instrumen krusial dalam berbagai bidang, termasuk diagnostik medis, di mana algoritma digunakan untuk mengidentifikasi pola tersembunyi dalam rekam medis elektronik guna memprediksi risiko penyakit seperti gangguan kardiovaskular dengan akurasi yang signifikan (Nikhate & Jonnalagedda, 2020).

1.4.2 Data Tidak Seimbang

Data tidak seimbang merupakan tantangan utama dalam *data mining* dan *machine learning*, terutama dalam studi diagnostik medis dan prediksi penyakit, karena dapat merusak keakuratan prediksi dan mengurangi kegunaan alat diagnostik (Salmi et al., 2024). Fenomena ini terjadi ketika distribusi kelas dalam dataset tidak merata, di mana satu kelas yang disebut kelas mayoritas memiliki jumlah sampel yang jauh lebih banyak daripada kelas minoritas. Dalam konteks medis, seperti diagnosis penyakit jantung, pasien sehat (non-penyakit) biasanya melebihi jumlah pasien yang sakit, menjadikan kasus penyakit sebagai kelas minoritas. *Imbalanced Ratio* (IR) untuk mengukur tingkat ketidakseimbangan pada suatu dataset (Celine et al., 2025). IR didefinisikan sebagai rasio antara jumlah data pada kelas mayoritas dan jumlah data pada kelas minoritas. Secara matematis, nilai IR dapat dinyatakan pada Persamaan (1).

$$IR = \frac{N_{mayoritas}}{N_{minoritas}} \tag{1}$$

dengan:

IR : *Imbalanced Ratio*

$N_{mayoritas}$: jumlah entri kelas mayoritas

$N_{minoritas}$: jumlah entri kelas minoritas

Nilai IR = 1 menunjukkan bahwa dataset berada dalam kondisi seimbang sempurna. Dataset dikatakan *imbalanced* apabila IR > 1.5, dan dianggap sangat *imbalanced* apabila IR ≥ 9.

1.4.3 SMOTE - *Encoded Nominal and Continuous*

Synthetic Minority Over-sampling Technique Encoded Nominal and Continuous (SMOTE-ENC), yang diusulkan oleh Mukherjee dan Khushi (2021), adalah pengembangan dari SMOTE untuk menangani dataset yang memuat fitur kategorikal. Berbeda dengan SMOTE standar yang hanya bekerja pada ruang fitur kontinu, SMOTE-ENC mengintegrasikan tahapan transformasi (*encoding*) untuk mengubah fitur nominal menjadi bentuk numerik yang merepresentasikan asosiasi kategori tersebut terhadap kelas target (Mukherjee & Khushi, 2021). Mengingat seluruh variabel prediktor dalam penelitian ini bersifat kategorikal, tahapan SMOTE-ENC dilakukan melalui prosedur berikut:

1. *Encoding* Probabilistik
   Variabel kategorikal tidak memiliki urutan nilai yang dapat dihitung jaraknya secara langsung. Oleh karena itu, setiap kategori c_j diubah menjadi nilai numerik $e(c_j)$ berdasarkan proporsi kemunculannya di kelas minoritas (pasien penyakit jantung) menggunakan Persamaan (2).
   $$e(c_j) = P(minority|c_j) \quad (2)$$
   dengan:
   c_j : kategori ke-j
   $e(c_j)$: nilai numerik hasil encoding atau proporsi kelas minoritas

2. Interpolasi Nilai *Encoding*
   Setelah seluruh data dikonversi menjadi numerik, algoritma menentukan tetangga terdekat $(k - NN)$ menggunakan jarak Euclidean. Selanjutnya, sampel sintetis baru dibangkitkan melalui interpolasi linear pada nilai *encoding* tersebut menggunakan Persamaan (3).
   $$e_{new} = e(x_i) + \lambda(e(x_{nn}) - e(x_i)) \quad (3)$$
   dengan:
   $e(x_i)$: nilai *encoding* dari kategori pada sampel minoritas x_i
   $e(x_{nn})$: nilai *encoding* dari kategori pada sampel tetangga terdekat x_{nn}
   λ : bilangan acak kontinu dari interval $[0,1]$
   e_{new} : nilai hasil interpolasi pada ruang *encoding* yang belum di*decode*

3. *Decoding* (Pemetaan Balik)
   Karena nilai e_{new} yang dihasilkan berupa bilangan kontinu (desimal) yang tidak merujuk pada kategori manapun secara langsung, diperlukan proses *decoding*. Kategori final untuk data sintetis ditentukan dengan memilih kategori asli c_j yang memiliki selisih nilai absolut terkecil terhadap e_{new} menggunakan Persamaan (4).
   $$x_{\text{new}}^{(\text{nom})} = \arg\min_{c_j} \left| e(c_j) - e_{\text{new}} \right| \quad (4)$$
   dengan:
   $x_{new}^{(nom)}$: kategori sintetis untuk variavel nominal
   $e(c_j)$: nilai *encoding* dari kategori c_j

e_{new} : hasil interpolasi numerik dari persamaan 3

Dimana operator $\arg\min$ memilih kategori c_j yang meminimalkan selisih absolut antara nilai *encoding* asli dan hasil interpolasi.

1.4.4 Algoritma Metode *Quick, Unbiased, and Efficient Statistical Tree*

Algoritma *Quick Unbiased Efficient Statistical Tree* (QUEST) merupakan algoritma pembentukan pohon keputusan yang dikembangkan oleh Loh dan Shih (1997). Berbeda dengan metode CART dan CHAID, QUEST dirancang agar proses pemilihan variabel tidak bias terhadap jumlah kategori maupun bentuk skala variabel. Algoritma ini bekerja melalui tiga tahapan utama, yaitu (1) seleksi variabel penyekat, (2) penentuan titik belah, dan (3) pembentukan simpul dengan menggunakan kombinasi linear untuk variabel kategorik. Uraian masing-masing tahap dijelaskan berikut ini.

1.4.4.1 Pemilihan Variabel (*Variable Selection*)

Pada setiap simpul, QUEST menentukan variabel penyekat terbaik menggunakan uji statistik. Karena seluruh variabel prediktor bertipe kategorik, maka digunakan Uji Independensi Chi-Square (χ^2) untuk mengukur keterkaitan antara setiap variabel prediktor dengan variabel respon (kelas target).

1. Pengujian pada Variabel Kategorikal

   Jika variabel X_k bertipe kategorik, QUEST menggunakan uji independensi Chi-Square $(\boldsymbol{\chi}^2)$ pada tabel kontingensi antara kategori variabel dan kelas target. Statistik uji dirumuskan menggunakan Persamaan (5).

$$\chi_k^2 = \sum_{r=1}^{J_t} \sum_{c=1}^{M_t} \frac{(O_{rc} - E_{rc})^2}{E_{rc}} \tag{5}$$

   dengan:

   J_t : jumlah kelas target

   M_t : jumlah kategori variabel

   O_{rc} : frekuensi observasi kategori ke-c dan kelas ke-r

   E_{rc} : frekuensi harapan

   kemudian *p-value* dihitung menggunakan rumus berikut:

$$\hat{\alpha}_2 = Pr\,(\chi^2_{(J_t-1)(M_t-1)} > \chi_k^2) \tag{6}$$

   variabel kategorik dipilih jika *p-value* memenuhi kriteria Bonferroni. Sedangkan frekuensi harapan diperoleh melalui:

$$E_{ij} = \frac{(\text{Total Baris } i) \times (\text{Total Kolom } j)}{\text{Total Keseluruhan } n} \tag{7}$$

2. Pemilihan Variabel Terbaik

   Variabel penyekat akhir (X^*) dipilih dengan membandingkan *p-value* dari seluruh variabel kandidat. Variabel dengan *p-value* terkecil (paling signifikan) akan dipilih untuk membagi simpul:

$$X^* = \arg\min_k \{\hat{\alpha}_1(k), \hat{\alpha}_2(k)\} \quad (8)$$

Variabel X^* kemudian digunakan pada tahap penentuan titik belah.

1.4.4.2 Transformasi Variabel Kategorik (CRIMCOORD)

Untuk menangani variabel kategorik, QUEST tidak membagi kategori secara langsung. Metode ini mentransformasi kategori menjadi nilai numerik diskriminatif menggunakan metode CRIMCOORD (*Classification-Based Categorical Variable Scoring*). Misalkan X adalah variabel prediktor kategorik dengan kategori $b_1, b_2, \dots, b_L$. Selanjutnya transformasikan X menjadi variabel prediktor numerik ξ untuk masing-masing kelas X. Langkah-langkah transformasi variabel prediktor kategorik menjadi variabel prediktor numerik adalah sebagai berikut:

1. Mentrasformasikan masing masing nilai x pada X^* ke vektor *dummy* L dimensi. Setiap kategori c_l dari variabel X direpresentasikan ke dalam bentuk vektor *dummy* v berdimensi L. Vektor ini didefinisikan sebagai $v = (v_1, \dots, v_M)'$, dengan elemen-elemen sebagai berikut:
$$v_l^{(j)} = [v_1, v_2, \dots, v_L]^T$$
$$\text{dengan } v_l = \begin{cases} 1, & jika\ x = b_l \\ 0, & lainnya \end{cases}, l = 1,2, \dots, l$$
2. Menentukan rata-rata untuk X^* menggunakan Persamaan 9 dan 10.
$$\bar{v} = \frac{\sum_l^L f_l v_l^{(j)}}{N_t} \quad (9)$$
$$\bar{v}^{(k)} = \frac{\sum_l^L n_l v_l^{(j)}}{N_{k,t}} \quad (10)$$
Keterangan:
v_l : pengamatan ke-l
$v_l^{(j)}$: pengamatan ke-l pada kelas j
$\bar{v}$: rata-rata untuk semua pengamatan pada simpul t
$\bar{v}^{(k)}$: rata-rata untuk semua pengamatan pada simpul t untuk kelompok ke-k
f_l : jumlah pengamatan pada simpul t untuk vl
n_l : jumlah pengamatan pada simpul t kelompok ke-k untuk vl
N_t : jumlah pengamatan pada simpul t
$N_{k,t}$: jumlah pengamatan pada simpul t kelompok ke-k
3. Menentukan matriks $L \times L$ pada Persamaan 11 dan 12.
$$B = \sum_{k=1}^{K} N_{k,t}(\bar{v}^{(k)} - \bar{v})(\bar{v}^{(k)} - \bar{v})^T \quad (11)$$
$$T = \sum_{k=1}^{K} f_l(v_l - \bar{v})(v_l - \bar{v})^T \quad (12)$$

Keterangan:

$\bar{v}(k)$: rata-rata untuk semua pengamatan pada simpul t untuk kelompok ke-k
$\bar{v}$: rata-rata untuk semua pengamatan pada simpul t
v_l : pengamatan ke-l
f_l : jumlah pengamatan pada simpul t untuk vl

4. Melakukan S*ingular Value Decomposition* (SVD) pada $\boldsymbol{T} = \boldsymbol{QDQ^T}$, Dimana $\boldsymbol{Q}$ adalah matriks orthogonal $L \times L$, $\boldsymbol{D} = diag(d_1, d_2, \dots, d_L)$ dengan $d_1 \geq d_2 \geq \cdots \geq d_L \geq 0$. Meskipun algoritma asli QUEST menggunakan pendekatan SVD untuk menangani matriks singular secara komputasi, pada matriks *scatter* $\boldsymbol{T}$ yang bersifat simetris, hasil SVD secara matematis ekuivalen dengan dekomposisi eigen. Oleh karena itu, dalam perhitungan manual, digunakan pendekatan dekomposisi eigen Ketika matriks ekuivalen.
5. Menentukan $\boldsymbol{D^{-\frac{1}{2}}} = diag(d_1^*, d_2^*, \dots, d_L^*)$,
   dimana $v_l = \begin{cases} d_l^{\frac{1}{2}}, \ jika \ d_L > 0 \\ 0, \ lainnya \end{cases}$
6. Melakukan SVD atau dekomposisi eigen pada $\boldsymbol{D^{-\frac{1}{2}}Q^TBQD^{-\frac{1}{2}}}$, untuk menentukan vektor eigen $\boldsymbol{a}$ yang berhubungan dengan nilai eigen terbesar
7. Menentukan titik koordinat diskriminan dari $\boldsymbol{v}$ yang merupakan proyeksi, yaitu dengan menggunakan Persamaan 13.

$$\xi = \boldsymbol{a^T D^{-\frac{1}{2}} Q^T v} \tag{13}$$

Keterangan:

$\boldsymbol{Q}$: Matriks *orthogonal* yang kolomnya merupakan vektor eigen dari $\boldsymbol{T^TT}$
$\boldsymbol{D}$: Matriks diagonal yang merupakan akar dari nilai eigen $\boldsymbol{T^TT}$
d_l : Elemen matriks diagonal ke-l
d_l^* : Elemen matriks diagonal ke-l setelah dipangkatkan $-\frac{1}{2}$
l : $1,2, \dots, L$

nilai ξ yang dihasilkan kemudian diperlakukan sebagai variabel numerik kontinu. Dengan demikian, titik belah optimal dapat ditentukan menggunakan prosedur QDA (persamaan kuadrat) pada persamaan 14.

1.4.4.3 Penentuan Titik Belah

QUEST menentukan titik belah optimal setelah variabel penyekat terpilih, menggunakan pendekatan *Quadratic Discriminant Analysis* (QDA) untuk kasus pemisahan biner. Setelah variabel kategorik diubah menjadi skor numerik (ξ) melalui proses di atas, titik belah optimal ditentukan menggunakan QDA. Misalkan variabel respon memiliki dua kategori (0 dan 1). Didefinisikan $\bar{x}_0, s_0^2$ sebagai rata-rata dan varians untuk kelas 0, serta $\bar{x}_1, s_1^2$untuk kelas 1. Peluang prior dinotasikan dengan $P(k|t)$. Persamaan diskriminan QDA dibentuk dari persamaan berikut:

$$p(A \mid t) s_A^{-1} \phi\left(\frac{x - \bar{x}_A}{s_A}\right) = p(B \mid t) s_B^{-1} \phi\left(\frac{x - \bar{x}_B}{s_B}\right)$$

dengan:

x	: nilai variabel atau titik belah yang dicari
A, B	: melambangkan dua kelas kategori
$p(A\|t)$	: peluang prior (proporsi data) kelas A
$p(B\|t)$	: peluang prior (proporsi data) kelas B
$s_A\ dan\ s_B$	: standar deviasi dari kelas A dan B
$\bar{x}_A$ dan $\bar{x}_B$	: rata-rata dari kelas A dan B
ϕ	: fungsi densital normal standar
$\frac{x-\bar{x}}{s}$	: z-score

melalui transformasi logaritma, persamaan tersebut dapat disederhanakan menjadi persamaan kuadrat $ax^2 + bx + c = 0$, dengan koefisien sebagai berikut:

$$a = s_0^2 - s_1^2$$
$$b = 2(\bar{x}_0 s_1^2 - \bar{x}_1 s_0^2)$$
$$c = (\bar{x}_1 s_0)^2 - (\bar{x}_0 s_1)^2 + 2s_0^2 s_1^2 \ln\left(\frac{p(0 \mid t) s_1}{p(1 \mid t) s_0}\right)$$

titik belah (d) diperoleh dengan mencari akar-akar persamaan kuadrat tersebut:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{14}$$

Jika terdapat dua akar $(d_1 \text{ dan } d_2)$, maka dipilih nilai d yang paling mendekati rata-rata sampel kelas pertama $(\bar{x}_0)$ dengan ketentuan hasil penyekatan menghasilkan dua simpul yang tidak kosong. Apabila $a = 0$ (varians dianggap homogen), maka titik belah ditentukan menggunakan pendekatan diskriminan linear:

$$d = \frac{\bar{x}_A + \bar{x}_B}{2} - \frac{s_A^2}{\bar{x}_A - \bar{x}_B} \ln\left(\frac{p(A \mid t)}{p(B \mid t)}\right) \tag{15}$$

1.4.5 Matriks Konfusi

Matriks konfusi merupakan instrumen yang digunakan untuk mengevaluasi kinerja model klasifikasi melalui pemetaan perbandingan antara hasil prediksi model terhadap nilai aktual (Purwanto & Widi Nugroho, 2023). Struktur matriks ini dikonstruksi oleh empat komponen utama, yaitu *True Positive* (TP), *False Positive* (FP), *True Negative* (TN), dan *False Negative* (FN) (Mahmoud, 2022). Secara spesifik, Vujović (2021) menjelaskan bahwa komponen TP merepresentasikan jumlah data kelas positif yang berhasil diprediksi secara tepat, sedangkan TN menunjukkan jumlah data kelas negatif yang diklasifikasikan dengan benar. Sebaliknya, kesalahan prediksi diidentifikasi melalui komponen FP, yaitu ketika data kelas negatif diklasifikasikan sebagai positif (*Type I error*), dan FN yang terjadi apabila model salah mengklasifikasikan data kelas positif menjadi negatif (*Type II error*). Dekomposisi nilai-nilai tersebut kemudian menjadi landasan matematis dalam perhitungan berbagai metrik evaluasi kinerja model seperti akurasi, presisi, *recall*, dan *F1-score*.

Tabel 1. Matriks Konfusi

Aktual/Prediksi	Positif	Negatif
Positif	TP	FP
Negatif	FN	TN

Tabel 1 menyajikan hubungan antara prediksi dan nilai aktual dalam klasifikasi yang dikenal sebagai matriks konfusi. Akurasi dapat dihitung berdasarkan komponen matriks konfusi menggunakan persamaan 16.

$$Akurasi = \frac{banyaknya\ prediksi\ yang\ benar}{total\ banyaknya\ prediksi} = \frac{TP + TN}{TP + FP + TN + FN} \quad (16)$$

Keterangan:
TP = *True Positive*
TN = *True Negative*
FP = *False Positive*
FN = *False Negative*

1.4.6 Penyakit Jantung

Jantung memegang peranan vital dalam mendistribusikan oksigen dan nutrisi tubuh (Mechanic et al., 2023). Disfungsi pada mekanisme ini dapat memicu penyakit kardiovaskular (*Cardiovascular Diseases*/CVDs). WHO menyatakan bahwa CVDs merupakan istilah payung bagi sekelompok gangguan yang memengaruhi jantung dan pembuluh darah, mencakup penyakit jantung koroner, penyakit serebrovaskular, penyakit jantung rematik, hingga kondisi akut seperti serangan jantung dan stroke. Secara global, penyakit ini masih menjadi penyebab kematian nomor satu, di mana sebagian besar kasusnya berkaitan erat dengan aterosklerosis suatu kondisi penyempitan arteri akibat penumpukan plak lemak yang menghambat aliran darah vital (Pahwa & Ishwarlal, 2023).

Manajemen penyakit jantung menuntut pendekatan komprehensif yang tidak hanya berfokus pada pengobatan kuratif, tetapi juga strategi preventif melalui modifikasi gaya hidup dan deteksi dini. Tantangan utama dalam penanganan CVDs adalah kompleksitas faktor risiko yang sering kali tidak terdiagnosis hingga mencapai tahap kritis. Dalam konteks ini, identifikasi risiko melibatkan berbagai determinan yang mencakup variabel demografis seperti Jenis Kelamin, Pendidikan, Status Pekerjaan, dan Generasi, serta variabel perilaku kesehatan yang meliputi Frekuensi Kebiasaan Merokok, Frekuensi Mengonsumsi Makanan Berlemak, dan Frekuensi Mengonsumsi Makanan Berpenyedap sebagai faktor yang memengaruhi Penyakit Jantung. Guna mengatasi kompleksitas data tersebut, integrasi teknologi dalam diagnostik medis menjadi semakin krusial. Pemanfaatan *Machine Learning* (ML) terbukti signifikan meningkatkan akurasi prediksi risiko dibandingkan metode statistik konvensional (El-Sofany et al., 2024). Analisis pola data klinis berbasis ML memungkinkan identifikasi dini pasien berisiko tinggi, memfasilitasi intervensi medis yang lebih tepat sasaran untuk menekan angka mortalitas (Deng, 2009)

BAB II
METODE PENELITIAN

2.1 Sumber Data

Penelitian ini menggunakan data sekunder dari Survei Kesehatan Indonesia (SKI) 2023 yang diselenggarakan oleh Kementerian Kesehatan RI. Data mencakup 842 observasi penderita dan non-penderita penyakit jantung di Kota Yogyakarta dengan tujuh variabel prediktor.

2.2 Variabel Penelitian

Penelitian ini menggunakan variabel respon (Y) dan variabel prediktor (X), yang seluruhnya bersifat kategorik. Daftar variabel penelitian disajikan pada Tabel 2.

Tabel 2. Variabel Penelitian

Variabel	Nama Variabel	Definisi Operasional	Kategori
Y	Penyakit Jantung	Kondisi medis yang didiagnosis oleh tenaga kesehatan sebagai penyakit jantung	1: Punya penyakit jantung 2: Tidak punya penyakit jantung
X_1	Jenis Kelamin	Klasifikasi jenis kelamin biologis	1: Laki-laki 2: Perempuan
X_2	Frekuensi Kebiasaan Merokok	Seberapa sering seseorang merokok dalam periode tertentu	1: Setiap hari 2: Tidak setiap hari 3: Tidak pernah
X_3	Pendidikan	Tingkat pendidikan terakhir responden	1: Tidak tamat SMA 2: Tamat SMA 3: Tamat perguruan tinggi
X_4	Status Pekerjaan	Kondisi ketenagakerjaan responden	1: Tidak Bekerja 2: Bekerja
X_5	Generasi	Kelompok usia berdasarkan tahun kelahiran	1: Generasi X 2: Generasi *Baby Boomer*

Variabel	Nama Variabel	Definisi Operasional	Kategori
X_6	Frekuensi Mengonsumsi Makanan Berlemak	Frekuensi konsumsi makanan tinggi lemak jenuh dan kolesterol, seperti daging berlemak, gorengan, makanan bersantan, olahan dengan margarin atau mentega, serta bahan pangan tinggi kolesterol seperti jeroan, telur, dan udang.	1: > 1 kali per hari 2: 1 kali per hari 3: 3-6 kali/minggu 4: 1-2 kali/minggu 5: < 3 kali/bulan 6: Tidak pernah
X_7	Frekuensi Mengonsumsi Makanan Berpenyedap	Frekuensi konsumsi makanan yang mengandung tambahan penyedap rasa atau *Monosodium Glutamate* (MSG), seperti vetsin atau micin, kaldu instan, dan bumbu instan yang mengandung MSG.	1: > 1 kali per hari 2: 1 kali per hari 3: 3-6 kali/minggu 4: 1-2 kali/minggu 5: < 3 kali/bulan 6: Tidak pernah

2.3 Tahapan Analisis Data

Tahapan analisis yang akan dilakukan dalam penelitian ini adalah sebagai berikut:

2.3.1 Tahapan SMOTE-ENC

1. Mengumpulkan data sekunder hasil Survei Kesehatan Indonesia (SKI) tahun 2023 yang mencakup data penderita dan non-penderita penyakit jantung di Yogyakarta.
2. Melakukan penyaringan data (*data cleaning*) untuk memastikan hanya variabel prediktor dan variabel respon yang relevan yang digunakan dalam penelitian. Seluruh variabel bersifat kategorikal dan dipersiapkan untuk proses *encoding* sebelum dilakukan oversampling.
3. Mengidentifikasi ketidakseimbangan kelas (*imbalanced data*) dengan menghitung jumlah sampel pada setiap kelas respon. Ketidakseimbangan ditentukan menggunakan persamaan 1.
4. Memberikan nilai *encoding* untuk setiap atribut kategorikal dengan himpunan kategori $\{c_1, c_2, \dots, c_k\}$, dilakukan proses encoding berbasis asosiasi kategori terhadap kelas minoritas. Nilai encoding untuk kategori c_j diperoleh melalui persamaan 3 yang merepresentasikan probabilitas kemunculan kelas minoritas pada kategori tersebut. Dengan demikian, kategori yang lebih sering muncul dalam kelas minoritas memperoleh nilai *encoding* yang lebih tinggi.

5. Memilih tetangga terdekat dalam kelas minoritas di mana setiap sampel minoritas x_i, dipilih satu tetangga terdekat x_{nn} dari kelas minoritas lain. Pemilihan tetangga dapat dilakukan menggunakan jarak apa pun yang konsisten dalam ruang fitur, karena proses interpolasi dilakukan setelah kategori diubah ke bentuk numerik melalui *encoding*.
6. Pembentukan sampel sintetis pada atribut kategorikal.
7. Sampel sintetis yang dihasilkan ditambahkan ke kelas minoritas hingga diperoleh proporsi kelas yang lebih seimbang.

2.3.2 Tahapan Algoritma QUEST

1. Melakukan seleksi variabel prediktor dengan menguji hubungan antara setiap variabel prediktor dan variabel respon. Mengingat seluruh variabel prediktor dalam penelitian ini bersifat kategorik, maka seleksi variabel dilakukan menggunakan uji independensi Chi-Square.
2. Menentukan variabel penyekat terbaik berdasarkan *p-value* terkecil dari seluruh variabel prediktor yang lolos seleksi. Variabel ini dipilih karena memiliki hubungan paling signifikan dengan variabel respon.
3. Melakukan transformasi CRIMCOORD untuk variabel kategorik, yaitu mengubah kategori menjadi skor numerik diskriminatif yang diturunkan dari vektor eigen matriks antara-kelas dan dalam-kelas. Variabel hasil transformasi kemudian menggunakan prosedur penentuan titik belah yang sama seperti variabel numerik.
4. Menentukan titik belah (*split point*) pada variabel terpilih menggunakan pendekatan *Quadratic Discriminant Analysis (QDA)*. Titik belah diperoleh dari akar persamaan kuadrat yang dibentuk dari perbedaan varians dan rerata antar *superclass*. Jika varians homogen (a = 0), digunakan diskriminan linear sebagai alternatif.
5. Membagi *node* berdasarkan titik belah terbaik apabila *p-value* variabel terpilih lebih kecil atau sama dengan batas signifikansi. Jika tidak signifikan, *node* dinyatakan sebagai *node* terminal.
6. Menghentikan pembentukan *node* baru apabila tidak ada variabel prediktor yang signifikan pada tahap seleksi variabel, atau jumlah pengamatan dalam *node* tidak memenuhi syarat minimum untuk dilakukan pengujian statistik.
7. Membagi data menjadi data pelatihan dan data pengujian, di mana data pelatihan digunakan untuk membangun pohon keputusan QUEST, sedangkan data pengujian digunakan untuk mengukur performa model pada data yang tidak dilibatkan dalam proses pelatihan.
8. Melakukan validasi model menggunakan data pengujian untuk mengevaluasi performa model dalam mengklasifikasikan observasi baru.
9. Menyusun *confusion matrix* dari hasil prediksi data pengujian untuk menghitung metrik evaluasi model, seperti akurasi, presisi, *recall*, dan *F1-score*.