PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS

SKRIPSI

IMMANUEL CLEMENT ONGGO PUTRA

20171000053

TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2024

PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS

SKRIPSI

Diajukan sebagai salah satu syarat untuk kelengkapan gelar kesarjanaan pada

Program Studi Teknik Informatika

Jenjang Pendidikan Strata 1

IMMANUEL CLEMENT ONGGO PUTRA

20171000053

TEKNIK INFORMATIKA

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS BUDDHI DHARMA

TANGERANG

2024

LEMBAR PERSEMBAHAN

"Hati orang berbudi mendapatkan pengetahuan, dan telinganya mencari ilmu."

(Amsal 18:15)

Dengan mengucap puji syukur kepada Tuhan Yang Maha Esa, Skripsi ini kupersembahkan untuk:

1. Bapak Alm. V. Michael Bedjo Irianto dan Ibu Almh. P. Ch. Anneke Halim tercinta yang telah membesarkan aku dan selalu membimbing, mendukung, memotivasi, memberi apa yang terbaik bagiku serta selalu mendoakan aku untuk meraih kesuksesanku.
2. Adik ayahku Ibu Sri Fariani, kakak ibuku Ibu Lenny Halim, dan adikku Christoforus Chrisviancent Onggo Putra yang telah memberikan dukungan semangat serta dorongan yang senantiasa diberikan.
3. Teman-teman kelompok belajar yang selalu berjuang bersama (Mahardika Ardi Manggala, Billy Gozali, Guntur Yoga Pratama, Jonathan Marcellino Pratama).

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERNYATAAN KEASLIAN SKRIPSI

Yang bertanda tangan di bawah ini.

NIM	: 20171000053
Nama	: Immanuel Clement Onggo Putra
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: Database Development

Dengan ini, saya menyatakan bahwa:

1. Skripsi ini adalah asli dan belum pernah diajukan untuk mendapat gelar akademik Sarjana atau kelengkapan studi, baik di Universitas Buddhi Dharma maupun di Perguruan Tinggi lainnya.
2. Skripsi ini saya buat sendiri tanpa bantuan dari pihak lain, kecuali arahan dosen pembimbing.
3. Dalam Skripsi ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dan dicantumkan sebagai acuan dalam naskah dengan disebutkan nama pengarang dan dicantumkan daftar pustaka.
4. Dalam Skripsi ini tidak terdapat pemalsuan (kebohongan), seperti buku, artikel, jurnal, data sekunder, pengolahan data, dan pemalsuan tanda tangan dosen atau Ketua Program Studi Universitas Buddhi Dharma yang dibuktikan dengan keasliannya.
5. Lembar pernyataan ini saya buat dengan sesungguhnya, tanpa paksaan dan apabila dikemudian hari atau pada waktu lainnya terdapat penyimpangan dan ketidak benaran dalam pernyataan ini, saya bersedia menerima sanksi akademik berupa pencabutan gelar akademik yang telah saya peroleh karena Skripsi ini serta sanksi lainnya sesuai dengan peraturan dan norma yang berlaku.

Tangerang, 1 Agustus 2024

Yang membuat pernyataan,

Immanuel Clement Onggo Putra

20171000053

UNIVERSITAS BUDDHI DHARMA

LEMBAR PERSETUJUAN PUBLIKASI KARYA ILMIAH

Yang bertanda tangan di bawah ini.

NIM	: 20171000053
Nama	: Immanuel Clement Onggo Putra
Jenjang Studi	: Strata 1
Program Studi	: Teknik Informatika
Peminatan	: Database Development

Dengan ini menyetujui untuk memberikan ijin kepada pihak Universitas Buddhi Dharma, Hak Bebas Royalti Non – Eksklusif (Non-exclusive Royalty-Free Right) atas karya ilmiah kami yang berjudul: "Penerapan Algoritma C4.5 dalam Data Mining untuk Identifikasi Faktor Risiko Stroke pada Dataset Medis".

Dengan Hak Bebas Royalti Non – Eksklusif ini pihak Universitas Buddhi Dharma berhak menyimpan, mengalih-media atau format-kan, mengelolanya dalam pangkalan data (database), mendistribusikannya, dan menampilkan atau mempublikasikannya di internet atau media lain untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis atau pencipta karya ilmiah tersebut.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Universitas Buddhi Dharma, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini saya buat dengan sebenarnya.

Tangerang, 1 Agustus 2024

Yang membuat pernyataan,

Immanuel Clement Onggo Putra

20171000053

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN PEMBIMBING

PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS

Dibuat Oleh:

NIM	: 20171000053
Nama	: Immanuel Clement Onggo Putra

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Database Development

Tahun Akademik 2023/2024

Disahkan oleh,

Tangerang, 1 Agustus 2024

Pembimbing,

Hartana Wijaya, S.Kom., M.Kom

NIDN: 0412058102

UNIVERSITAS BUDDHI DHARMA

LEMBAR PENGESAHAN SKRIPSI

PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS

Dibuat Oleh:

NIM : 20171000053

Nama : Immanuel Clement Onggo Putra

Telah disetujui untuk dipertahankan di hadapan Tim Penguji Ujian Komprehensif

Program Studi Teknik Informatika

Peminatan Database Development

Tahun Akademik 2023/2024

Disahkan oleh,

Tangerang, 1 Agustus 2024

Dekan,	**Ketua Program Studi,**
Dr. Yakub, M.M., M.Kom	**Hartana Wijaya, S.Kom., M.Kom**
NIDN: 0304056901	**NIDN: 0412058102**

LEMBAR PENGESAHAN TIM PENGUJI

Nama : Immanuel Clement Onggo Putra

NIM : 20171000053

Fakultas : Sains dan Teknologi

Judul Skripsi : Penerapan Algoritma C4.5 dalam *Data Mining* untuk Identifikasi Faktor Risiko *Stroke* pada *Dataset* Medis

Dinyatakan LULUS setelah mempertahankan di depan Tim Penguji pada hari Kamis, 1 Agustus 2024.

Nama Penguji:		Tanda Tangan:
Ketua Sidang	: Riki, M.Kom NIDN: 0431128204	.....
Penguji I	: Dram Renaldi, S.Kom., M.Kom NIDN: 0411019001	.....
Penguji II	: Hartana Wijaya, S.Kom., M.Kom NIDN: 0412058102	.....

Mengetahui,

Dekan Fakultas Sains dan Teknologi

Dr. Yakub, M.M., M.Kom

NIDN: 0304056901

KATA PENGANTAR

Dengan mengucapkan Puji Syukur kepada Tuhan Yang Maha Esa, yang telah memberikan Rahmat dan karunia-Nya kepada penulis sehingga dapat menyusun dan menyelesaikan Skripsi ini dengan judul **"PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS"**. Tujuan utama dari pembuatan Skripsi ini adalah sebagai salah satu syarat kelengkapan dalam menyelesaikan program pendidikan Strata 1 Program Studi Teknik Informatika di Universitas Buddhi Dharma. Dalam penyusunan Skripsi ini penulis banyak menerima bantuan dan dorongan baik moril maupun materiil dari berbagai pihak, maka pada kesempatan ini penulis menyampaikan rasa terima kasih yang sebesar-besarnya kepada:

1. Bapak Dr. Limajatini, S.E., M.M., B.K.P., sebagai Rektor Universitas Buddhi Dharma
2. Bapak Dr. Yakub, S.Kom., M.Kom., M.M., Dekan Fakultas Sains dan Teknologi
3. Bapak Rudy Arijanto, S.Kom., M.Kom., Wakil Dekan Fakultas Sains dan Teknologi
4. Bapak Hartana Wijaya, S.Kom., M.Kom., sebagai Ketua Program Studi Teknik Informatika dan pembimbing yang telah membantu dan memberikan dukungan serta harapan untuk menyelesaikan penulisan Skripsi ini.
5. Orang tua dan keluarga yang selalu memberikan dukungan baik moril dan materiil.
6. Teman-teman yang selalu membantu dan memberikan semangat

Serta semua pihak yang terlalu banyak untuk disebutkan satu-persatu sehingga terwujudnya penulisan ini. Penulis menyadari bahwa penulisan Skripsi ini masih belum sempurna, untuk itu penulis mohon kritik dan saran yang bersifat membangun demi kesempurnaan penulisan di masa yang akan datang.

Akhir kata semoga Skripsi ini dapat berguna bagi penulis khususnya dan bagi para pembaca yang berminat pada umumnya.

Tangerang, 1 Agustus 2024

Penulis

Penerapan Algoritma C4.5 dalam *Data Mining* untuk Identifikasi Faktor Risiko *Stroke* pada *Dataset* Medis
136 halaman + xxi / 53 tabel / 55 gambar / 3 lampiran

ABSTRAK

Stroke adalah masalah kesehatan di seluruh dunia yang terjadi secara signifikan. Berdasarkan data Organisasi Kesehatan Dunia (WHO), insiden *stroke* baru mencapai 13,7 juta kasus per tahun dengan angka kematian sebesar 5,5 juta jiwa. *Data mining* adalah proses dari pengumpulan, pengolahan, dan analisis data untuk memperoleh informasi yang penting. Algoritma C4.5 merupakan algoritma yang paling umum digunakan dalam *data mining* untuk membangun pohon keputusan berdasarkan data yang telah diberi label. Masalah yang ingin diselesaikan dari penelitian ini adalah mengolah data kesehatan untuk mengetahui risiko *stroke* itu sulit dan lama dan masyarakat sulit mendapat informasi tentang *stroke* yang benar dan terpercaya. Untuk teknik pengumpulan data dari *dataset* yang diambil dari situs *Kaggle* dengan *dataset Stroke Prediction* dengan 5110 *record* dan 12 atribut, serta studi pustaka yang mencari informasi-informasi yang terkait dengan penelitian ini untuk membantu penerapan algoritma data mining C4.5 dalam mendeteksi penyebab *stroke*. Pustaka yang digunakan dari berbagai media seperti internet, buku, jurnal, dan media lainnya. Algoritma pohon keputusan di *data mining* dapat diterapkan dalam mendeteksi penyebab seseorang terkena stroke. Setelah dilakukan perhitungan manual dan menggunakan perangkat lunak *RapidMiner*, akurasi hasil prediksi menjadi tolok ukur seberapa efektif algoritma ini dalam mengidentifikasi penyebab *stroke* pada seseorang, dengan menghasilkan model klasifikasi dengan tingkat akurasi 94,89% dan nilai AUC sebesar 0.709. Ada 11 responden yang mengisi kuesioner dengan hasilnya adalah 45,44% memilih sangat setuju, 44,56% memilih setuju, dan 10% yang netral.

Kata Kunci: C4.5, *Database*, *Data Mining*, *Stroke*, Teknik Informatika

Identifying Stroke Risk Factors in Medical Datasets Using the C4.5 Algorithm
136 pages + xxi / 53 tables / 55 images / 3 attachments

ABSTRACT

Stroke is a worldwide health problem that occurs significantly. Based on data from the World Health Organization (WHO), there are 13.7 million new stroke incidents each year, resulting in 5.5 million deaths. Data mining is the process of collecting, processing, and analyzing data to obtain important information. The C4.5 algorithm is the most commonly used in data mining to build decision trees based on labeled data. The problem this study aims to solve is the difficulty and time-consuming nature of processing health data to determine the risk of stroke. The public has difficulty getting correct and reliable information about stroke. For data collection techniques from datasets taken from the Kaggle site with the Stroke Prediction dataset with 5110 records and 12 attributes, as well as literature studies that search for information related to this study to help apply the C4.5 data mining algorithm in detecting the causes of stroke. The libraries are from various media such as the internet, books, journals, and other media. The decision tree algorithm in data mining can be applied to detect the cause of a person having a stroke. After manual calculations and using RapidMiner software, the accuracy of the prediction results is a benchmark for how effective this algorithm is in identifying the cause of stroke in a person, by producing a classification model with an accuracy rate of 94.89% and an AUC value of 0.709. 11 respondents filled out the questionnaire with the results being 45.44% choosing strongly agree, 44.56% choosing agree, and 10% being neutral.

Keywords: *C4.5, Database, Data Mining, Information Technology, Stroke*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Latar Belakang

Stroke adalah masalah kesehatan di seluruh dunia yang terjadi secara signifikan, penyakit ini menempati urutan kedua sebagai penyebab utama dari kematian dan kecacatan yang ada di seluruh dunia (World Health Organization, 2018). Data (Kemenkes RI, 2019) juga menunjukkan bahwa *stroke* menjadi masalah kesehatan serius di Indonesia. Kondisi ini ditandai dengan kerusakan otak akibat gangguan aliran darah non-traumatik yang terjadi secara mendadak atau bertahap, mengakibatkan berbagai manifestasi klinis seperti defisit sensorik dan motorik, gangguan kognitif, serta gangguan psikologis seperti stres, kecemasan, dan depresi (Munce et al., 2017). Oleh karena itu, manajemen stres yang efektif menjadi krusial dalam upaya pencegahan dan pengelolaan *stroke*, mengingat dampaknya yang dapat mengancam jiwa.

Berdasarkan data Organisasi Kesehatan Dunia (WHO), insiden *stroke* baru mencapai 13,7 juta kasus per tahun dengan angka kematian sebesar 5,5 juta jiwa. Secara global, penderita *stroke* diperkirakan mencapai 50 juta jiwa dengan 9 juta di antaranya mengalami kecacatan berat. Tingkat mortalitas kasus *stroke* mencapai 10%. Secara geografis, 70% dari kasus *stroke* dan 87% tingkat kematian serta kecacatan yang diakibatkan oleh *stroke* kebanyakan terjadi di negara-negara yang memiliki penghasilan yang rendah dan tingkat menengah. (Turana et al., 2021)

Pada tahun 2019, *stroke* tercatat sebagai penyebab kematian utama di Indonesia dengan tingkat mortalitas mencapai 15,4%, diikuti oleh penyakit tidak menular lainnya seperti hipertensi, diabetes, kanker, dan PPOK. Prevalensi *stroke* yang ada pada penduduk yang berumur lebih dari 15 tahun diperkirakan berkisar antara 10,9% hingga 11,3%, dengan

angka mencapai 11,4% di Jawa Tengah. Data tahun 2018 juga menunjukkan prevalensi *stroke* sebesar 7% pada populasi usia ≥ 15 tahun. (Kemenkes RI, 2019)

Seiring berkembangnya waktu ilmu kesehatan sekarang tidak hanya mengandalkan teknik analisis yang berbasis konvensional saja, melainkan sudah menggabungkan perkembangan teknologi yang ada di dalamnya. Salah satunya yang menggabungkan dunia teknologi dengan dunia kesehatan adalah dengan cara mendeteksi penyakit *stroke*. Berbagai cara yang banyak digunakan untuk mendeteksi penyakit *stroke*, yang salah satunya dengan menggunakan metode penggalian data atau *data mining*. *Data mining* adalah proses dari pengumpulan, pengolahan, dan analisis data untuk memperoleh informasi yang penting. *Data mining* menggunakan teknik-teknik analisis data dan algoritma untuk memproses data yang berukuran besar. Sedangkan menurut (Indah Werdiningsih et al., 2020) mengatakan bahwa *data mining* adalah ilmu yang dapat menggunakan teknik statistik, *machine learning* (pembelajaran mesin), visualisasi data, pengenalan pola, dan *database* untuk menangani masalah-masalah dari penghasilan informasi dari *database* yang banyak.

Untuk dapat melakukan *data mining* diperlukanlah suatu algoritma yang membantu proses tersebut. Algoritma adalah serangkaian petunjuk atau perhitungan yang dirancang untuk memecahkan suatu masalah dengan cara yang sistematis, logis, dan terstruktur. Algoritma ini dapat memproses data dan dapat menghasilkan model atau solusi yang berguna. Menurut (Kani, 2020) algoritma adalah serangkaian langkah yang logis dan terstruktur yang dirancang untuk memecahkan masalah dan dapat menghasilkan solusi tertentu. Dalam konteks *data mining*, terdapat berbagai jenis algoritma, dan penelitian ini akan berfokus pada penggunaan algoritma C4.5.

C4.5 atau *Decision Tree*, yang dalam bahasa Indonesia disebut pohon keputusan, merupakan algoritma yang paling umum digunakan dalam *data mining* untuk membangun pohon keputusan berdasarkan data yang telah diberi label. Pohon keputusan ini selanjutnya

dapat diubah menjadi aturan klasifikasi untuk dapat memprediksi kelas atau kategori dari data baru. Menurut (Azahari & Nursobah, 2021), algoritma C4.5 dapat menghasilkan pohon keputusan, sebuah model yang mudah dipahami untuk memprediksi atau mengklasifikasikan data. Algoritma C4.5 beroperasi dengan mengidentifikasikan atribut yang dapat memberikan informasi paling signifikan (*gain* tertinggi). Atribut ini kemudian dipilih sebagai simpul utama (*node*) dalam pohon keputusan. Proses ini diulang secara rekursif untuk setiap cabang pohon hingga tidak ada lagi atribut yang dapat menghasilkan simpul baru.

Berdasarkan penjelasan sebelumnya, penelitian ini bertujuan untuk mengevaluasi keakuratan algoritma C4.5 dalam mendeteksi penyakit *stroke*. Dengan itu, penelitian ini berjudul "**PENERAPAN ALGORITMA C4.5 DALAM *DATA MINING* UNTUK IDENTIFIKASI FAKTOR RISIKO *STROKE* PADA *DATASET* MEDIS**".

1.2 Identifikasi Masalah

Masalah yang bisa diangkat dari penelitian ini adalah:

1. Saat ini, mengolah data kesehatan untuk mengetahui risiko *stroke* itu sulit dan lama.
2. Masyarakat sulit mendapat informasi tentang *stroke* yang benar dan terpercaya.

1.3 Pertanyaan Penelitian

Masalah-masalah yang telah diidentifikasi sebelumnya mengarahkan pada rumusan masalah penelitian sebagai berikut:

1. Apakah penggunaan algoritma C4.5 dalam analisis data dapat membantu mengidentifikasi faktor-faktor risiko *stroke* dari *dataset* medis?
2. Apakah perancangan aplikasi berbasis *Java* berguna untuk membantu seseorang melihat hasil diagnosa *stroke*?

1.4 Tujuan dan Manfaat Penelitian

Penelitian ini bertujuan untuk menggali, menganalisis, dan menyelesaikan permasalahan yang sudah ada. Secara lebih rinci, tujuan dan manfaat penelitian ini adalah sebagai berikut:

1.4.1 Tujuan Penelitian

Penelitian ini bertujuan untuk:

a. Mengembangkan model *data mining* yang dapat mampu mengklasifikasikan data medis untuk mengidentifikasi gejala-gejala dari penyakit *stroke*.
b. Merancang aplikasi berbasis model yang dapat digunakan untuk mengetahui penyebab-penyebab potensial terjadinya *stroke*.

1.4.2 Manfaat Penelitian

Penelitian ini dapat memberikan manfaat seperti:

a. Menerapkan pendekatan *data mining* dengan algoritma C4.5 guna mengidentifikasi penyebab *stroke* yang lebih efektif.
b. Mengembangkan aplikasi berbasis data mining yang mudah digunakan masyarakat untuk mengukur risiko terkena penyakit *stroke*.

1.5 Ruang Lingkup

Cakupan dari penelitian ini dapat meliputi:

a. Penelitian ini menggunakan algoritma C4.5.
b. Penelitian ini menggunakan set data (*dataset*) yang didapat dari situs web *Kaggle*.
c. Perancangan aplikasi berbasis *desktop* dan dibuat menggunakan bahasa pemrograman *Java* menggunakan perangkat lunak *NetBeans*.
d. Pengujian dilakukan dengan menggunakan perangkat lunak *RapidMiner* dan menghitung probabilitas dari algoritma tersebut.

e. Metode penerapan yang digunakan adalah klasifikasi.

1.6 Metodologi Penelitian

Penelitian ini menggunakan metode kuantitatif. Metode ini dipilih karena lebih cocok untuk menganalisis data berupa angka dan sampel populasi. Pendekatan kuantitatif memungkinkan pengujian masalah sosial berdasarkan variabel yang dapat diukur secara numerik. (Creswell, 2014)

1.6.1 Metode Penelitian

Proses dimulai dengan pemilihan topik yang cocok dengan tujuan dan manfaat yang akan dicapai. Langkah selanjutnya studi dapat ditentukan, seperti pada penelitian ini yang berjudul "Penerapan Algoritma C4.5 dalam *Data Mining* untuk Identifikasi Faktor Risiko *Stroke* pada *Dataset* Medis".

Agar dapat memahami permasalahan yang ada dan menjelaskannya secara lebih detail, langkah-langkah yang diperlukan adalah menentukan tujuan penelitian yang lebih jelas dan mendeskripsikan tujuan tersebut dengan lengkap. Selain itu, tujuan penelitian ini juga akan diatur, dengan mempertimbangkan ada kendala yang akan dihadapi dalam perancangan dan implementasi dalam penelitian ini. Adapun langkah-langkah yang harus diperhatikan:

a. Menyadari permasalahan yang sudah ada.
b. Merincikan permasalahan yang sejelas-jelasnya.
c. Menentukan tujuan yang akan dicapai dengan aplikasi yang ingin dibuat dan dirancang.

Klasifikasi merupakan metode penting di dalam *data mining* yang dapat digunakan untuk mengelompokkan data ke dalam kategori yang sudah ditentukan. Dengan memanfaatkan algoritma, klasifikasi mempelajari pola dari

data historis (data pelatihan) yang sudah diberi label untuk kemudian memprediksi kategori dari data baru yang belum diberi label.

1.6.2 Metode Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh melalui cara-cara berikut:

a. *Dataset* dari data sekunder

   Penulis menggunakan pengumpulan data sekunder dengan mengambil data yang disediakan oleh penyedia *dataset* daring dari situs web *Kaggle*.

b. Studi Pustaka

   Penulis mencari informasi-informasi yang terkait dengan penelitian ini untuk membantu penerapan algoritma data mining C4.5 dalam mendeteksi penyebab *stroke*. Pustaka yang digunakan dari berbagai media seperti internet, buku, jurnal, dan media lainnya.

1.7 Sistematika Penulisan

Untuk mempermudah pemahaman, penelitian ini terbagi dalam 5 bab, dari BAB I sampai BAB V. Berikut adalah rincian sistematika penelitian ini.

BAB I: PENDAHULUAN

Bab ini menyajikan gambaran umum penelitian, mencakup latar belakang masalah yang mendasari penelitian ini, rumusan masalah yang ingin dijawab, tujuan dan manfaat yang diharapkan dari penelitian ini, ruang lingkup dari penelitian ini, metode yang digunakan dalam penelitian, serta struktur penulisan laporan penelitian.

BAB II: LANDASAN TEORI

Bab ini menjelaskan teori-teori yang mendasari penerapan algoritma *data mining* dalam mengidentifikasi penyebab *stroke*. Teori-teori ini mencakup teori umum dari penelitian ini, teori khusus yang memengaruhi penelitian ini, teori perancangan untuk merancang aplikasi dari penelitian ini, teori pengujian dari penelitian ini, tinjauan studi yang berisi jurnal-jurnal pendukung, dan kerangka pemikiran yang mendasari penelitian ini.

BAB III: METODOLOGI PENELITIAN

Bab ini menjelaskan secara rinci metodologi penelitian, termasuk kebutuhan-kebutuhan yang mendasari penelitian ini. Hasil analisis yang disajikan dalam bab ini menjadi dasar untuk mengembangkan proses penelitian, meliputi perancangan *Activity Diagram*, analisis kebutuhan dan masalah, identifikasi kebutuhan sistem, perhitungan manual algoritma C4.5, hasil perhitungan algoritma yang menggunakan *RapidMiner*, pengumpulan kebutuhan (*requirement elicitation*), serta jadwal pelaksanaan penelitian.

BAB IV: HASIL PEMBAHASAN

Bab ini menyajikan hasil penelitian yang meliputi perancangan basis data (*database*) dan aplikasi, termasuk menu dan prototipe. Selain itu, dijelaskan pula tangkapan layar program yang telah selesai, pengujian algoritma menggunakan *Confusion Matrix* dan *AUC*, pengujian aplikasi menggunakan *Black Box* dan kuesioner penelitian.

BAB V: SIMPULAN DAN SARAN

Bab terakhir ini menyajikan kesimpulan penelitian, saran pengembangan aplikasi, dan kendala yang dihadapi selama penelitian. Pembahasan ini mencakup solusi yang potensial untuk mengatasi masalah yang ditemukan, serta rekomendasi

untuk perbaikan aplikasi dan metode penelitian yang akan dilakukan di masa mendatang.

BAB II

LANDASAN TEORI

2.1 Teori Umum

Bab ini akan memaparkan teori-teori umum yang menjadi acuan dalam menganalisis permasalahan yang ingin diteliti.

2.1.1 Data

Data adalah jembatan terpenting yang menghubungkan mesin (perangkat keras) dengan pengguna (manusia). Sebagai komponen inti dalam sistem manajemen basis data, atau yang dalam bahasa Inggris, Database Management System (DBMS), data memainkan peran sentral dalam menyimpan, mengelola, dan menyajikan informasi yang berharga bagi para penggunanya. (Connolly & Begg, 2015)

Menurut (Coronel & Morris, 2016) data merupakan fakta mentah yang belum diproses lebih lanjut. Fakta mentah ini bisa berupa pengamatan terhadap fenomena fisik atau catatan transaksi. Dengan kata lain, data adalah bentuk awal dari informasi yang belum memiliki makna atau konteks yang jelas. (Indrajani, 2018)

Dapat disimpulkan bahwa data adalah representasi dari informasi yang menggambarkan objek-objek secara lebih rinci, yang dapat disimpan dalam bentuk angka maupun karakter.

Ada dua jenis pembagian dalam data, yaitu:

a. Data Kualitatif

   Menurut (Sugiyono, 2015), data kualitatif menyajikan informasi tentang sifat atau karakteristik dari suatu objek atau subjek, yang tidak dapat diukur dengan angka melainkan dialami atau diamati. Informasi ini biasanya dapat diperoleh melalui berbagai bentuk seperti teks, gambar, suara, atau video.

Contoh data kualitatif meliputi jenis kelamin, status sosial, atau preferensi warna. Karena sifatnya yang deskriptif, data kualitatif tidak dapat dianalisis dengan menggunakan perhitungan matematika, melainkan memberikan pemahaman yang lebih mendalam tentang suatu fenomena yang telah terjadi.

b. Data Kuantitatif

Data kuantitatif adalah data penelitian yang bisa berupa angka dan dapat dianalisis menggunakan metode statistik. Data ini bersifat konkrit dan dapat diolah secara matematis.

Data primer dapat diperoleh langsung dari sumber aslinya oleh pengumpul data, sementara data sekunder bisa didapatkan dari sumber yang telah ada sebelumnya atau yang sudah diolah oleh pihak lain.

a. Data Primer

Data primer merupakan informasi yang dikumpulkan secara langsung oleh peneliti dari sumber aslinya untuk menjawab pertanyaan penelitian yang spesifik tersebut.

b. Data Sekunder

Data sekunder merupakan informasi yang didapatkan oleh peneliti yang bukan secara langsung, melainkan melalui sumber yang sudah tersedia atau dari pihak lain yang sudah mengumpulkan data tersebut sebelumnya. Peneliti hanya perlu menyalin, mengakses, atau meminta data yang sudah dikumpulkan oleh pihak lain di lapangan.

2.1.2 Informasi

Menurut (Tukino, 2020), informasi adalah hasil pengolahan data sehingga menjadi lebih bermakna dan berguna bagi orang yang menerimanya. Nilai tambah ini dapat berupa peningkatan pemahaman, pengurangan ketidakpastian, atau dukungan

dalam pengambilan keputusan yang efektif. Kualitas informasi sangat dipengaruhi oleh keakuratan, relevansi, dan ketepatan waktu penyampaiannya. Informasi yang tidak akurat atau tidak relevan dapat menyesatkan dan berdampak negatif pada proses pengambilan keputusan.

Menurut (Coronel & Morris, 2016), informasi adalah representasi terstruktur dari data yang telah melalui proses transformasi dan pengolahan, sehingga menghasilkan nilai tambah dan relevansi dalam konteks tertentu.

Informasi adalah representasi terstruktur dari data atau fakta yang sudah melalui proses pengolahan sehingga memiliki nilai tambah dan relevansi bagi penggunanya. (Anggraeni et al., 2017)

Dari berbagai definisi yang sudah dipaparkan, dapat disimpulkan bahwa informasi merupakan hasil olahan data yang terstruktur dan terdiri dari kombinasi huruf serta simbol yang bisa menyampaikan pengetahuan tentang suatu hal yang sudah terjadi.

2.1.3 Aplikasi

Aplikasi merupakan perangkat lunak yang dirancang oleh individu atau tim pengembang untuk menjalankan fungsi tertentu. Pembuatan aplikasi melibatkan penulisan kode pemrograman serta desain antarmuka pengguna yang intuitif.

Aplikasi merupakan perangkat lunak yang dirancang untuk mengeksekusi serangkaian instruksi terprogram yang berguna memenuhi kebutuhan pengguna dalam berbagai bidang. Aplikasi memanfaatkan sumber daya komputer untuk melakukan tugas-tugas spesifik, seperti pengolahan data, manajemen sistem, hiburan, edukasi, dan produktivitas. Setiap aplikasi memiliki fungsi dan fitur yang unik, disesuaikan dengan tujuan penggunaannya.

Perangkat lunak yang dikembangkan untuk dijalankan pada komputer desktop, PC, atau laptop secara khusus disebut sebagai "aplikasi *desktop*". Sementara itu, perangkat lunak yang dirancang untuk perangkat *mobile* seperti ponsel pintar dan tablet dikenal sebagai "aplikasi *mobile*" atau "aplikasi seluler".

Aplikasi merupakan perangkat lunak yang dikembangkan untuk memfasilitasi pengguna dalam menyelesaikan berbagai tugas komputasi, seperti pengolahan dokumen, manajemen sistem operasi Windows, eksekusi permainan digital, dan aktivitas komputasi lainnya. (R. S. Hakim, 2018) Pengolah kata, lembar kerja (*spreadsheet*), dan pemutar media merupakan contoh perangkat lunak aplikasi yang umum digunakan. Aplikasi-aplikasi tersebut dirancang untuk menyelesaikan tugas-tugas yang spesifik melalui pengolahan data berdasarkan aturan dan ketentuan bahasa pemrograman tertentu. Kumpulan aplikasi yang terintegrasi dalam satu paket disebut sebagai paket atau *suite* aplikasi. Pengembangan aplikasi merupakan produk dari proses perancangan sistem yang terstruktur dan sistematis. (Muljadi et al., 2020) Penerapannya meluas ke berbagai sektor, meliputi bisnis, hiburan (seperti permainan), layanan publik, periklanan, dan hampir seluruh aspek kehidupan manusia lainnya. Aplikasi juga dapat didefinisikan sebagai perangkat lunak yang digunakan oleh pengguna awam maupun pengembang untuk menyelesaikan berbagai tugas.

2.2 Teori Khusus

Bab ini akan menjabarkan teori-teori spesifik yang mendasari topik penelitian, serta menjelaskan kaitannya dengan aplikasi, teknologi, dan metode yang akan diterapkan dalam penelitian ini.

2.2.1 *Stroke*

Stroke merupakan gangguan kesehatan yang terjadi akibat terhambatnya aliran darah menuju otak. Kondisi ini menyebabkan sel-sel otak kekurangan asupan oksigen dan mengalami kerusakan, sehingga fungsi bagian otak yang terdampak ikut terganggu. (Naziyah et al., 2019)

Stroke, atau dalam istilah medis disebut Cerebro-Vascular Accident (CVA), merupakan kondisi yang ditandai dengan gangguan fungsi saraf secara mendadak akibat terganggunya suplai darah ke otak. (Hariyanti et al., 2020)

Stroke adalah penyakit otak yang ditandai dengan gangguan-gangguan fungsi saraf secara lokal atau menyeluruh. Gangguan ini muncul tiba-tiba, berkembang pesat, dan disebabkan oleh terganggunya aliran darah ke otak yang bukan dikarenakan cedera. (Siregar & Anggeria, 2019)

Penyakit *stroke* dapat disebabkan oleh:

a. *Stroke* iskemik

   Stroke iskemik muncul ketika aliran darah menuju otak terhalang. Daerah yang sudah tidak menerima aliran darah ke otak akan berhenti bekerja dan akan mati. Jika aliran darah tidak segera langsung dipulihkan, stroke iskemik dapat mengakibatkan kerusakan otak secara permanen bahkan kematian.

b. *Stroke* pendarahan

   Stroke pendarahan disebabkan karena pembuluh darah di otak sudah pecah. Pembuluh darah yang sudah pecah ini mengakibatkan terganggunya pengiriman aliran darah yang membawa oksigen ke otak. Alih-alih darah tersebut dialirkan melalui pembuluh darah, darah tersebut justru merembes ke ruang yang berada di luar pembuluh darah

(ruang ekstravaskuler), sehingga jaringan otak mengalami kekurangan oksigen dan tertekan oleh darah yang sudah memenuhi ruang tersebut.

2.2.2 Klasifikasi

Klasifikasi merupakan teknik pengelompokan data ke dalam kategori atau kelas yang sudah ditentukan. Metode ini bertujuan untuk meningkatkan keakuratan prediksi dan analisis, terutama saat menghadapi volume data yang besar. (Osman, 2019)

Tujuan klasifikasi dalam penambangan data adalah untuk mengelompokkan data ke dalam kategori yang sudah ada. Proses ini dimulai dengan menggunakan data yang sudah diberi label kategori sebagai acuan. (Indah Werdiningsih et al., 2020)

Berdasarkan definisi klasifikasi yang telah diuraikan, dapat disimpulkan bahwa klasifikasi merupakan suatu metode analisis data yang bertujuan untuk mengidentifikasi pola atau model yang melekat dalam suatu kumpulan data. Proses ini melibatkan pengelompokan data ke dalam kelas-kelas yang berbeda berdasarkan kesamaan atribut atau karakteristik yang dimiliki. Data yang memiliki kemiripan yang tinggi akan dikelompokkan ke dalam satu kelas yang sama, sedangkan data yang tidak memiliki kemiripan akan ditempatkan pada kelas yang berbeda.

2.2.3 *Data Mining*

Menurut (Indah Werdiningsih et al., 2020), *data mining* merupakan bidang ilmu yang memanfaatkan metodologi dari statistik, pembelajaran mesin, basis data, pengenalan pola, dan visualisasi data untuk menghasilkan pengetahuan berharga dan pola tersembunyi dari kumpulan data berskala besar.

Menurut (Suntoro, 2019), *data mining* adalah proses pengumpulan informasi penting dari data dalam jumlah besar. Informasi ini diubah menjadi pengetahuan baru yang berguna untuk pengambilan keputusan yang lebih baik. Proses ini menggunakan

berbagai teknik, seperti klasifikasi, klustering, asosiasi, dan regresi, untuk menemukan pola tersembunyi, hubungan, keanehan, atau tren yang sulit ditemukan secara manual. Hasil dari *data mining* bisa berupa model prediksi, aturan klasifikasi, segmentasi pelanggan, atau rekomendasi produk, yang semuanya dapat digunakan untuk membuat keputusan bisnis yang lebih strategis dan berdasarkan data.

Data mining merupakan serangkaian teknik komputer yang digunakan untuk memperoleh secara otomatis pola-pola yang sah, baru, berguna, dan dapat dipahami dari volume data yang besar secara efisien. Pola-pola ini sebelumnya tidak diketahui dan tersembunyi dalam data. Pola-pola tersebut harus dapat ditindaklanjuti secara lebih lanjut sehingga hasil tersebut dapat digunakan dalam pengambilan keputusan dalam perusahaan. (Bhatia, 2019)

(Iriadi et al., 2020) menyebutkan bahwa *data mining* adalah proses berulang untuk menganalisis *database* guna mendapatkan informasi dan pengetahuan akurat yang dapat membantu ilmuwan dalam pengambilan keputusan dan pemecahan masalah. Metode ini sangat berguna dan unggul, karena dapat secara otomatis mengidentifikasi pola-pola terkait dalam basis data.

Proses penambangan data, yang dikenal sebagai *Knowledge Discovery in Databases* (KDD), melibatkan penggunaan metode ilmiah untuk menggali informasi berharga dari kumpulan data yang besar. (Syahril et al., 2020) Tahap awal KDD melibatkan pengumpulan data dari berbagai sumber seperti basis data terstruktur, terstruktur sebagian, dan tak terstruktur. Setelah itu, data dibersihkan, digabungkan, dan diubah formatnya agar siap dianalisis. Inti dari KDD adalah *data mining*, di mana teknik analisis seperti klasifikasi, klastering, asosiasi, dan regresi dapat digunakan untuk menggali pola, hubungan, anomali, atau tren tersembunyi dalam data besar. Berikut adalah detail proses KDD:

a. *Data Selection* (Pemilihan Data)

Data selection adalah tahapan pemilahan data yang sesuai untuk analisis dalam KDD. Pada tahap ini, data dari *database* akan disederhanakan dan dikompres untuk memastikan informasi penting tetap terjaga. Data yang telah dipilih akan disimpan dalam file yang terpisah dari *database* utama.

b. *Pre-processing* (Pemrosesan Awal)

Pre-processing adalah langkah awal untuk mempersiapkan data mentah agar lebih terstruktur dan mudah dipahami sebelum diubah dan dianalisis lebih lanjut. Data mentah seringkali tidak memiliki format-format yang teratur, sehingga proses *data mining* tidak dapat memproses hal tersebut. *Data preprocessing* adalah langkah awal dalam data mining yang melibatkan pembersihan dan transformasi data agar data tersebut sesuai untuk dianalisis.

c. *Transformation* (Transformasi Data)

Data transformation dalam *data mining* adalah langkah yang melibatkan perubahan data mentah menjadi format yang sesuai untuk analisis yang diinginkan dan kompatibel dengan algoritma *data mining* yang ingin digunakan. Tahapan ini dilakukan setelah data mentah dikumpulkan dan merupakan bagian integral dari keseluruhan proses *data mining*.

d. *Data Mining* (Pengolahan Data)

Data mining merupakan langkah krusial dalam proses KDD. Pada tahap ini, akan dipilih algoritma atau metode yang paling sesuai dengan jenis informasi yang ingin digali, seperti prediksi, klasifikasi, perkiraan, pengelompokan (*clustering*), dan sebagainya.

e. *Evaluation* (Evaluasi)

Evaluation adalah tahap final dalam proses KDD. Pada tahap ini, dilakukan pengukuran dan penilaian komprehensif terhadap kinerja dan reliabilitas model yang dihasilkan oleh algoritma yang diterapkan. Evaluasi ini dilakukan untuk menguji hipotesis awal dan untuk mengetahui berapa persen data yang diperoleh dapat dipercaya. Evaluasi juga diperlukan untuk mengetahui tingkat keakuratan sebuah algoritma tersebut apakah rendah atau tinggi.

Data Mining dibagi menjadi beberapa metode algoritma yang disesuaikan dengan fungsi-fungsinya, yaitu:

a. *Association* (Asosiasi)

Dalam *data mining*, asosiasi atau *association rule learning* adalah metode untuk menemukan hubungan antar atribut dalam kumpulan data yang besar. Metode ini menghasilkan aturan-aturan yang mengukur seberapa kuat hubungan antara dua atau lebih atribut tersebut.

b. *Classification* (Klasifikasi)

Klasifikasi dalam *data mining* merupakan suatu teknik analisis prediktif yang bertujuan untuk mengelompokkan objek data ke dalam kelas atau kategori yang sudah ditentukan sebelumnya. Teknik ini memanfaatkan pembelajaran mesin untuk membangun model klasifikasi yang berdasarkan atribut atau fitur yang relevan, sehingga memungkinkan prediksi kelas dari objek data baru yang belum memiliki label.

c. *Clustering* (Klustering)

Clustering dalam konteks *data mining*, merupakan teknik untuk mengelompokkan data menjadi beberapa himpunan yang disebut sebagai *cluster*. Data-data dalam satu *cluster* memiliki kemiripan yang tinggi satu sama lain, sementara data dari *cluster* berbeda menunjukkan tingkat kemiripan yang rendah.

d. *Prediction* (Prediksi)

Prediksi, sebagai salah satu teknik dalam *data mining* yang serupa dengan klasifikasi, memiliki tujuan utama untuk memperkirakan nilai suatu atribut berdasarkan nilai-nilai atribut lainnya yang relevan. Teknik ini memanfaatkan analisis terhadap data historis untuk menghasilkan proyeksi atau perkiraan terhadap hasil atau kejadian yang mungkin terjadi di masa depan.

e. *Estimation* (Estimasi)

Dalam *data mining*, estimasi merupakan suatu proses yang melibatkan perkiraan nilai numerik yang tidak diketahui secara pasti dalam suatu *dataset*. Metode estimasi ini dimanfaatkan untuk menghitung atau memperkirakan nilai atau kuantitas data dalam periode waktu tertentu pada *dataset* tersebut.

2.2.3.1 C4.5

Algoritma C4.5, yang dikembangkan oleh Ross Quinlan, merupakan metode pembentukan pohon keputusan yang berlandaskan pada pemilihan atribut dengan nilai *gain* tertinggi berdasarkan entropi sebagai kriteria pemisahan optimal pada setiap simpul pohon. (Sukma et al., 2019)

Algoritma C4.5 bekerja dengan dua cara utama: pertama, merancang dan membangun struktur pohon keputusan berdasarkan data yang diberikan. Kedua, dari pohon keputusan tersebut, algoritma ini menghasilkan aturan-aturan (*rule model*) berupa pernyataan "jika-maka" (*if then*) yang dapat digunakan untuk mengklasifikasikan atau memprediksi hasil berdasarkan kondisi tertentu.

Proses pembuatan *decision tree* dengan algoritma C4.5 melibatkan empat langkah utama:

1. Pemilihan Atribut Akar

   Atribut dengan nilai *gain* tertinggi dipilih sebagai akar dari pohon keputusan.

2. Pembentukan Cabang

   Cabang-cabang dibuat berdasarkan nilai-nilai yang dimiliki oleh atribut akar, sesuai dengan nilai *gain* tertinggi yang telah dihitung.

3. Pembagian Kasus

   Setiap kasus data dibagi ke dalam cabang-cabang yang telah dibuat, berdasarkan nilai atribut yang relevan.

4. Pengulangan Proses

   Langkah 1 hingga 3 diulang untuk setiap cabang, hingga semua kasus dalam satu cabang memiliki kelas yang sama. Proses ini berhenti ketika tidak ada lagi atribut yang bisa digunakan untuk membagi data.

Algoritma C4.5, juga dikenal sebagai *decision tree*, bekerja dengan mengevaluasi setiap titik keputusan dan memilih opsi terbaik hingga tidak ada lagi pilihan yang bisa diambil. Para peneliti telah memanfaatkan algoritma ini untuk membantu pengambilan keputusan yang lebih cepat, akurat, dan efisien. Dalam penelitian ini, algoritma C4.5 digunakan untuk mengidentifikasi

penyebab *stroke* dengan menganalisis gejala-gejala yang dialami pasien melalui pohon keputusan yang dibangun dari data yang ada.

Secara umum, prosedur pembentukan pohon keputusan dengan algoritma C4.5 dapat diuraikan sebagai berikut:

a. Persiapan *Data Training*: Kumpulkan data historis yang sudah tergolong ke dalam kelas-kelas yang relevan.

b. Pemilihan Akar Pohon: Tentukan atribut terbaik sebagai akar pohon dengan menghitung nilai gain dari setiap atribut. Atribut dengan nilai gain tertinggi akan terpilih. Sebelumnya, hitung nilai entropy menggunakan rumus sebagai berikut:

$$Entrophy(S) = \sum_{i=1}^{n} - pi \times \log_2 pi$$

Keterangan:

S = Himpunan kasus

A = Atribut

n = Jumlah partisi S

pi = Proporsi dari Si terhadap S

c. Hitung nilai *gain* dari atribut yang dipilih menggunakan metode *information gain*:

$$Gain(S, A) = Entrophy(S) - \sum_{i=1}^{n} \frac{Si}{S} \times Entrophy$$

d. Pengulangan Langkah 2: Ulangi langkah pemilihan akar pohon hingga semua data terpartisi.

e. Penghentian Proses: Proses partisi pohon keputusan berhenti ketika memenuhi kondisi berikut: semua data dalam satu node memiliki kelas

yang sama, tidak ada atribut lagi yang bisa dipartisi, atau tidak ada data dalam cabang yang kosong.

2.2.4 Algoritma

Algoritma merupakan serangkaian langkah logis dan terstruktur yang dirancang untuk menyelesaikan masalah-masalah yang spesifik dengan menghasilkan keluaran yang diharapkan melalui proses komputasi yang terdefinisi dengan baik. (Kani, 2020)

Menurut (Munir & Leony, 2016), algoritma merupakan serangkaian langkah logis dan terstruktur yang dirancang untuk memecahkan suatu masalah atau menyelesaikan tugas tertentu. Mereka adalah dasar dari pemrograman komputer dan berperan penting dalam berbagai bidang-bidang yang terkait dalam hal tersebut, mulai dari matematika dan sains hingga keuangan dan kecerdasan buatan.

Konsep algoritma diperkenalkan oleh Abu Ja'far Muhammad Ibnu Musa Al-Khwarizmi, seorang ilmuwan terkemuka dari Timur Tengah pada era kejayaan Islam. Istilah "algoritma" sendiri diyakini oleh beberapa ahli berasal dari namanya, Al-Khwarizmi. (Putro et al., 2022) Instruksi-instruksi ini disusun secara logis dan dapat dieksekusi secara berurutan untuk mencapai hasil yang diinginkan. Secara garis besar, terdapat tiga jenis algoritma yang umum digunakan dalam pemrograman, yaitu algoritma sekuensial yang menjalankan instruksi secara berurutan, algoritma percabangan yang menjalankan instruksi berdasarkan kondisi tertentu, dan algoritma perulangan yang mengulang instruksi hingga kondisi tertentu terpenuhi.

Beberapa karakteristik unik dari algoritma pemrograman membuatnya berbeda dari metode pemrograman lain dalam mengatasi masalah. Ciri-ciri tersebut adalah:

1. Adanya keterbatasan

   Algoritma harus mempunyai jumlah langkah yang terbatas dan pasti. Artinya, proses eksekusi algoritma harus berhenti pada suatu titik dan tidak boleh berulang tanpa henti.

2. Memiliki *input*

   Algoritma dapat menerima masukan (*input*) dari pengguna atau sumber lain. Masukan ini dapat berupa nilai, data, atau parameter yang digunakan untuk proses perhitungan atau pemrosesan.

3. Ada *output* yang sesuai

   Algoritma menghasilkan keluaran (*output*) yang merupakan hasil dari proses perhitungan atau pemrosesan. Keluaran ini dapat berupa data, informasi, atau tindakan yang dilakukan.

4. Memiliki kepastian

   Setiap langkah dalam algoritma harus didefinisikan dengan jelas dan tidak ambigu agar tidak membingungkan pengguna. Artinya, tidak boleh ada instruksi yang membingungkan atau menimbulkan penafsiran ganda.

5. Adanya keefektifan

   Sebuah algoritma yang efektif harus memberikan hasil yang akurat dan sesuai dengan tujuannya, serta menggunakan sumber daya seperti waktu dan memori secara efisien.

6. Disusun secara terstruktur

   Algoritma perlu dirancang dengan struktur yang jelas dan runtut. Setiap langkah dalam algoritma harus disusun secara logis dan teratur agar proses penyelesaian masalah menjadi efisien dan tidak bertele-tele. Dengan

demikian, waktu yang diperlukan untuk menyelesaikan masalah dapat dipersingkat.

Dari penjelasan di atas, dapat disimpulkan bahwa algoritma merupakan bagian penting dalam pemrograman. Algoritma membantu programmer memecahkan masalah dan membuat program yang lebih efisien serta efektif. Penguasaan algoritma yang baik adalah kunci menjadi *programmer* yang ahli.

2.2.5 Basis Data

Menurut (Abdulloh, 2018), basis data atau *database* merupakan himpunan terstruktur dari informasi atau data yang disimpan dalam sistem komputer, yang kemudian dapat diproses dan diolah menggunakan perangkat lunak guna memperoleh informasi yang relevan.

Database adalah kumpulan data terstruktur yang saling terhubung, dirancang untuk memungkinkan akses dan penggunaan ulang informasi secara efisien. (Kawistara & Hidayatullah, 2017)

Menurut (Kurniawan & Marhamelda, 2019), basis data merupakan suatu struktur data yang terorganisasi secara logis dan sistematis, dirancang untuk disimpan dan dikelola dalam lingkungan komputasi. Struktur ini memfasilitasi penyimpanan, pengambilan, pembaruan, dan pengelolaan informasi secara efisien. *Database* berfungsi sebagai repositori informasi yang dapat diakses, diperbarui, dan dianalisis menggunakan perangkat lunak manajemen basis data untuk dapat mendukung pengambilan keputusan dan proses bisnis.

Secara ringkas, basis data adalah kumpulan informasi terstruktur yang tersimpan secara sistematis, memungkinkan akses cepat dan mudah. Basis data dapat menyimpan berbagai jenis data, mulai dari teks hingga video. Penyimpanannya bisa terpusat di satu lokasi atau tersebar di berbagai jaringan. Keamanan basis data menjadi hal krusial,

memastikan hanya pihak berwenang yang dapat mengakses atau memodifikasi informasi di dalamnya.

2.3 Teori Perancangan

Bab ini akan menjabarkan teori-teori perancangan yang relevan dengan topik penelitian, dengan fokus pada proses pengembangan aplikasi yang menjadi tujuan penelitian ini..

2.3.1 *Java*

Menurut (Mardiani et al., 2017), *Java* diluncurkan oleh Sun Microsystems pada tahun 1995, yang merupakan bahasa pemrograman berorientasi objek yang bersifat lintas platform, memungkinkan eksekusi kode pada beragam arsitektur perangkat keras, termasuk komputer desktop dan perangkat mobile. *Java* dapat dijalankan di Windows, Linux, Unix, dan DOS dan banyak digunakan dalam pembuatan dan perancangan aplikasi baik untuk *desktop*, situs web, seluler, dll.

Java, bahasa pemrograman populer yang diciptakan oleh tim Sun Microsystems yang dipimpin oleh Patrick Naughton dan James Gosling, dirancang untuk fleksibilitas dan kesederhanaan. Awalnya bernama "*Oak*", namanya kemudian diubah menjadi "*Java*" karena konflik merek dagang. *Java* memungkinkan pengembangan aplikasi lintas platform, termasuk komputer dan perangkat seluler, dan bahkan memiliki peramban sendiri yang disebut "*HotJava*". (Nofriadi, 2015)

Menurut (Firly, 2018), *Java* adalah bahasa pemrograman yang bersifat *multi platform* yang tidak menyediakan IDE secara khusus. *Programmer* dapat menggunakan IDE yang mendukung *Java*, seperti *NetBeans*, *Eclips*, dan *TexPad*. *Java* adalah bahasa pemrograman untuk berbagai tujuan-tujuan, berbasis kelas, dan berorientasi pada objek. Bahasa pemrograman *Java* telah digunakan secara luas untuk

pengodean aplikasi web dan telah menjadi pilihan yang populer di kalangan *developer* selama lebih dari dua dekade terakhir.

2.3.2 *RapidMiner*

RapidMiner adalah perangkat lunak sumber terbuka (*open source*) yang dapat digunakan untuk menganalisis *data mining*, *text mining*, dan membuat prediksi. Dengan memanfaatkan berbagai teknik deskriptif dan prediktif, *RapidMiner* memberikan wawasan berharga yang membantu pengguna dalam membuat keputusan terbaik.

Menurut (Putri et al., 2021), *RapidMiner* merupakan perangkat lunak komprehensif yang dirancang untuk melakukan analisis data mining yang kompleks. Kemampuannya didukung oleh beragam mode operasi yang memfasilitasi berbagai teknik dan metodologi dalam proses *data mining*. Perangkat lunak ini memiliki sifat bersumber terbuka sehingga dapat dipakai oleh semua orang. *RapidMiner* pada saat kali pertama dirilis, awalnya menggunakan nama Yale yang berarti *Yet another learning environment*. Perangkat lunak tersebut dirilis oleh Ralf Klinkenbert pada tahun 2001.

RapidMiner adalah perangkat lunak lengkap yang dirancang untuk memfasilitasi analisis *data mining*, *text mining*, serta prediksi. Dengan mengintegrasikan berbagai teknik deskriptif dan prediktif, perangkat lunak ini mampu menghasilkan wawasan berharga yang dapat dimanfaatkan dalam pengambilan keputusan strategis. Dilengkapi dengan lebih dari 500 operator *data mining* yang mencakup fungsi *input*, *output*, *preprocessing*, dan visualisasi, *RapidMiner* menawarkan kemudahan untuk memenuhi kebutuhan analisis data yang kompleks. (Wahyudi et al., 2019)

Berdasarkan uraian di atas, dapat disimpulkan bahwa perangkat lunak *RapidMiner* merupakan perangkat lunak sumber terbuka (*open source*) yang dapat

digunakan untuk melakukan penggalian data (*data mining*), penerapan model, dan pengoperasian model untuk *data mining* dan *text mining*.

2.3.3 *MySQL*

MySQL adalah sistem manajemen basis data relasional (*relational database management system*, RDBMS) *open-source* yang dibangun di atas bahasa pemrograman SQL (*Structured Query Language*). *MySQL* memungkinkan pengguna untuk melakukan operasi data relasional seperti pembuatan, modifikasi, pengambilan, dan penghapusan data. Beberapa keunggulan *MySQL* adalah sebagai berikut:

a. Memungkinkan analis untuk mengelola, menyimpan, mengubah, menghapus, dan menyimpan data dengan rapi
b. Mampu memproses permintaan data dengan sangat cepat, baik dalam menerima maupun mengirimkan data
c. Bisa diakses oleh beberapa *user* secara bersamaan tanpa membuat sistem tersebut *crash* atau berhenti bekerja dengan baik
d. Aman, dengan aturan hak akses yang lebih ketat dan enkripsi kata sandi tingkat paling tinggi

Menurut (Rusmawan, 2019), *MySQL* merupakan sistem manajemen basis data yang populer dan banyak digunakan di seluruh dunia. *MySQL* memungkinkan banyak pengguna untuk mengakses dan mengelola data secara bersamaan. *MySQL* menyimpan data dalam bentuk tabel yang saling terkait, dengan kolom yang merepresentasikan jenis informasi yang disimpan dalam tabel tersebut.

MySQL adalah basis data banyak pengguna yang memiliki banyak *thread*. *MySQL* dapat digunakan sebagai basis data dan sebagai perangkat lunak untuk *database server*. *MySQL* dapat berguna untuk mengatur koleksi-koleksi di dalam struktur data (*database*), baik untuk pengelolaan maupun pembuatan basis data.

Adapun beberapa perintah dasar yang dimiliki oleh *MySQL* adalah sebagai berikut:

1. *USE*

   Perintah *MySQL* ini digunakan untuk memilih dan mengaktifkan basis data yang akan digunakan.

2. *INSERT*

   Perintah *MySQL* ini digunakan untuk mengisi data ke dalam tabel yang akan digunakan.

3. *SHOW DATABASES*

   Perintah *MySQL* ini digunakan untuk menampilkan daftar basis data dalam suatu *MySQL*.

4. *UPDATE*

   Perintah *MySQL* ini memungkinkan modifikasi data yang sudah tersimpan dalam tabel basis data.

5. *DELETE*

   Perintah *MySQL* ini menghapus data dari tabel yang ditunjuk dalam basis data.

6. *CREATE DATABASE*

   Perintah *MySQL* ini berfungsi untuk membuat basis data baru dalam sistem.

7. *CREATE TABLE*

   Perintah *MySQL* ini digunakan untuk membuat tabel baru di dalam basis data yang sedang aktif.

Dari penjelasan di atas, dapat disimpulkan bahwa *MySQL* adalah salah satu sistem manajemen basis data yang paling populer dan banyak digunakan. Kemampuan *MySQL* dalam menangani basis data besar didukung oleh utilitas pemuatan

berkecepatan tinggi, cache memori khusus, dan mekanisme peningkatan kinerja lainnya. *MySQL* bersifat *open source* dan digunakan untuk membuat tabel yang menyimpan data terkait, memungkinkan analis untuk menyimpan, mengelola, menghapus, dan memodifikasi data secara efisien dan terstruktur.

2.3.4 XAMPP

Menurut (Novendri et al., 2019), XAMPP adalah perangkat lunak bebas yang kompatibel dengan berbagai sistem operasi dan mendukung pengembangan web dengan bahasa seperti HTML, *JavaScript*, CSS, PHP, dan SQL. XAMPP mencakup *Apache*, server web yang memungkinkan pembuatan *website* secara lokal (localhost). Singkatan XAMPP berasal dari X (*cross-platform*), *Apache*, *MySQL* (*database*), *PHPMyAdmin* (manajemen *database*), dan *Perl* (bahasa *scripting*).

Menurut (Roza et al., 2020), XAMPP adalah perangkat lunak bebas lintas platform yang menggabungkan beberapa komponen perangkat lunak penting dalam satu paket instalasi. Perangkat lunak ini mengintegrasikan *Apache HTTP Server*, *MySQL database server*, serta dukungan untuk bahasa pemrograman PHP (versi 4 dan 5). Dengan instalasi yang mudah dan gratis, XAMPP banyak digunakan untuk pengembangan dan pengujian aplikasi web berbasis PHP di lingkungan lokal, baik pada sistem operasi Linux maupun Windows.

XAMPP dapat memungkinkan pengguna untuk membuat *server web* lokal dengan lebih gampang, menjalankan skrip-skrip PHP, dan mengakses *database MySQL* tanpa mengakses melalui *server* internet atau hosting eksternal.

2.3.5 *NetBeans*

Menurut (Maya, 2015), *NetBeans* merupakan aplikasi *Integrated Development Environment (IDE)* yang populer di kalangan pengembang perangkat lunak, digunakan

untuk menulis, mengkompilasi, men-debug, dan mendistribusikan program. Mirip dengan *Microsoft Visual Studio*, *NetBeans* memiliki cakupan yang lebih luas dalam pengembangan aplikasi dan merupakan perangkat lunak *open-source*.

NetBeans adalah perangkat lunak pengembangan yang dirancang khusus untuk bahasa pemrograman *Java*. Sebagai lingkungan pengembangan terintegrasi (IDE) yang bersifat *open-source*, *NetBeans* tersedia secara gratis dan mendukung berbagai sistem operasi seperti *Windows*, *Linux*, *Mac*, dan *Solaris*.

Java Development Kit (JDK) adalah sekumpulan perangkat lunak yang diperlukan untuk membuat aplikasi *Java*. Di dalamnya terdapat *Java Virtual Machine (JVM)* dan *Java Runtime Environment (JRE)*, serta alat-alat lain yang membantu proses penulisan kode program *Java* atau *Kotlin*. JDK dapat diunduh dan dipasang pada komputer untuk memulai pengembangan aplikasi.

JDK merupakan perangkat lunak lengkap yang digunakan untuk membuat dan mengembangkan aplikasi *Java*. Di dalamnya sudah termasuk JRE, kompiler, serta berbagai alat bantu lainnya. Sementara itu, JRE adalah paket perangkat lunak yang lebih sederhana, hanya berisi JVM dan pustaka kelas yang dibutuhkan untuk menjalankan aplikasi *Java*.

Integrated Development Environment (IDE) adalah aplikasi yang memudahkan pengembangan perangkat lunak dengan menyediakan berbagai alat dalam satu tempat. IDE memiliki fitur-fitur seperti penyunting kode, alat bantu pengujian, pembuatan paket perangkat lunak, dan kompilasi kode menjadi program yang dapat dijalankan. Umumnya, IDE mencakup penyunting teks untuk kode sumber, alat bantu untuk mengotomatiskan proses pembangunan perangkat lunak, dan debugger untuk mencari kesalahan dalam kode.

IDE merupakan *software* aplikasi yang pada umumnya berbasis GUI. Pada IDE ini, kita bisa menuliskan setiap baris-baris kode bahasa pemrograman dan dapat menjalankan program aplikasi yang sudah dibuat. IDE dapat berfungsi untuk merampingkan seluruh proses pengembangan *software* atau aplikasi dengan menggabungkan berbagai alat dan fungsi dari aplikasi tersebut ke dalam satu antarmuka yang terpadu.

2.3.6 *Entity Relation Diagram*

Menurut (Rusmawan, 2019), *Entity Relationship Diagram* (ERD) adalah representasi visual dari model data yang menggambarkan hubungan antar entitas, batasan-batasan, serta atribut-atributnya. ERD digunakan dalam pengembangan sistem, khususnya dalam perancangan basis data relasional, untuk memetakan hubungan antar objek di dunia nyata. Komponen utama ERD terdiri dari entitas (objek), relasi (hubungan antar entitas), dan atribut (karakteristik entitas).

Dari penjelasan sebelumnya, dapat disimpulkan bahwa ERD (*Entity Relationship Diagram*) adalah model data visual yang menggunakan notasi grafis untuk merancang struktur basis data dan menggambarkan hubungan antar entitas di dalamnya. ERD juga digunakan dalam rekayasa perangkat lunak untuk merancang skema basis data dan memahami struktur ERD tersebut. ERD juga dapat membantu memvisualisasikan ide-ide dari perancangan basis data, sehingga kesalahan dan kekurangan dalam perancangan dapat diidentifikasi, ditemukan, dan dibenarkan sebelum melakukan perubahan dalam basis data tersebut. Ada 3 jenis macam hubungan dalam ERD itu sendiri, yaitu: *one-to-one*, *one-to-many*, dan *many-to-many*.

2.3.7 *Unified Modeling Language*

Unified Modeling Language (UML) adalah standar industri yang digunakan untuk memvisualisasikan, merancang, dan mendokumentasikan sistem perangkat lunak berorientasi objek. Dikembangkan oleh Grady Booch, Jim Rumbaugh, dan Ivar Jacobson pada tahun 1995, UML pertama kali dirilis oleh *Object Management Group* pada tahun 1997 dan telah menjadi alat penting dalam pengembangan perangkat lunak modern. UML memungkinkan para pengembang untuk membuat diagram dan model yang jelas dan terstruktur, memfasilitasi komunikasi dan kolaborasi yang efektif di antara tim pengembangan.

Menurut (Sukamto & Shalahuddin, 2019), UML adalah bahasa standar yang populer di industri untuk memvisualisasikan dan mengkomunikasikan model sistem. UML menggunakan diagram dan teks pendukung untuk menjelaskan persyaratan, melakukan analisis dan perancangan, serta menggambarkan arsitektur dalam pengembangan perangkat lunak berorientasi objek.

Dari penjelasan sebelumnya, dapat disimpulkan bahwa UML adalah seperangkat aturan untuk menggambarkan perangkat lunak yang akan dibuat. UML menggunakan diagram dan simbol yang memiliki arti dan fungsi masing-masing. Dengan UML, perancang program dan programmer dapat lebih mudah memahami alur kerja dalam perangkat lunak tersebut.

2.4 Teori Pengujian

Bab ini akan membahas teori-teori pengujian yang relevan dengan topik penelitian, dengan fokus pada teori-teori yang berkaitan langsung dengan proses pengujian sistem yang akan dikembangkan.

2.4.1 *Confusion Matrix*

Confusion Matrix adalah tabel yang digunakan untuk mengukur kinerja model klasifikasi dengan membandingkan hasil prediksi dengan hasil aktual. Matriks ini terdiri dari empat sel yang merepresentasikan empat kemungkinan hasil klasifikasi: *True Positive* (TP), *True Negative* (TN), *False Positive* (FP), dan *False Negative* (FN). TP dan TN menunjukkan prediksi yang sesuai dengan kenyataan, sedangkan FP dan FN menunjukkan prediksi yang tidak sesuai.

Confusion matrix atau matriks kebingungan adalah metode evaluasi yang digunakan untuk mengukur kinerja suatu sistem klasifikasi dengan cara membandingkan hasil prediksi yang dihasilkan oleh sistem tersebut dengan hasil aktual. Metode ini memberikan representasi visual dari berbagai jenis kesalahan yang mungkin dilakukan oleh sistem, seperti *false positive* dan *false negative*, yang sangat berguna dalam mengidentifikasi area di mana sistem memerlukan perbaikan. (Zhafira et al., 2021)

Confusion Matrix adalah tabel yang memperlihatkan kinerja suatu model klasifikasi dengan membandingkan label kelas yang diprediksi dengan label kelas yang sesungguhnya. (Satria et al., 2020)

2.4.2 AUC

Area Under the Curve (AUC) merupakan ukuran kinerja model klasifikasi yang memiliki dua kelas. AUC menunjukkan seberapa baik model dapat membedakan antara kelas positif dan negatif. Nilai AUC yang lebih tinggi (mendekati 1) berarti model memiliki kinerja yang lebih baik dalam memisahkan kedua kelas tersebut. AUC dapat dipahami sebagai ringkasan dari kurva ROC, yang menggambarkan hubungan antara tingkat keberhasilan klasifikasi kelas positif dan tingkat kesalahan klasifikasi kelas positif pada berbagai ambang batas keputusan.

AUC adalah ukuran umum untuk mengevaluasi kinerja model klasifikasi biner, terutama ketika proporsi kelas tidak seimbang. AUC menunjukkan secara keseluruhan seberapa baik model dapat membedakan antara kelas positif dan negatif pada berbagai batas keputusan.

AUC adalah ukuran penting untuk mengevaluasi seberapa baik suatu model klasifikasi bekerja. AUC bekerja dengan menghitung luas area di bawah kurva ROC, yang menunjukkan hubungan antara tingkat prediksi positif yang benar (*true positive rate*, TPR) dan tingkat prediksi positif yang salah (*false positive rate*, FPR) pada berbagai ambang batas. Nilai AUC yang tinggi berarti model tersebut lebih baik dalam membedakan antara kelas positif dan negatif, menjadikannya metrik yang berguna dalam memilih dan meningkatkan model klasifikasi.

2.4.3 *Black Box*

Black box testing adalah teknik pengujian perangkat lunak yang berfokus pada fungsionalitas dari perspektif pengguna. Dengan mengabaikan struktur internal perangkat lunak, pengujian ini menganalisis input dan output untuk mengidentifikasi potensi masalah yang mungkin dihadapi pengguna saat menggunakan perangkat lunak tersebut.

Black box testing merupakan metode pengujian perangkat lunak yang berfokus pada hasil keluaran (output) berdasarkan masukan (input) yang diberikan, tanpa perlu memahami detail internal perangkat lunak tersebut, termasuk kode program yang digunakan. Proses pengujian dilakukan dengan memasukkan berbagai data pada formulir yang tersedia untuk memastikan perangkat lunak berfungsi sesuai dengan kebutuhan pengguna. (Priyaungga et al., 2020)

2.4.4 Skala *Likert*

Skala *Likert* adalah alat pengukur yang digunakan untuk memahami sikap, pendapat, dan persepsi individu atau kelompok terhadap suatu peristiwa atau fenomena sosial. Peneliti menggunakan skala ini untuk mengumpulkan data kuantitatif dan kualitatif terkait fenomena sosial tersebut, berdasarkan definisi yang telah mereka tetapkan sebelumnya. Singkatnya, skala *Likert* adalah alat penelitian untuk mengukur sikap dan pendapat.

Menurut (Sugiyono, 2015), skala *Likert* adalah alat ukur yang digunakan untuk mengukur sikap, pendapat, dan persepsi individu atau kelompok terhadap fenomena sosial. Skala ini bekerja dengan cara menguraikan variabel yang akan diukur menjadi indikator-indikator yang lebih spesifik. Indikator-indikator ini kemudian digunakan sebagai dasar untuk mengembangkan item-item dalam instrumen penelitian, yang bisa berupa pernyataan atau pertanyaan.

2.5 Tinjauan Studi

Bagian ini akan memaparkan berbagai tinjauan yang menjadi landasan dan acuan dalam penelitian ini.

2.5.1 Jurnal 1

Tabel 2.1 Jurnal 1

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Perbandingan 3 Algoritma Klasifikasi Data Mining Dalam Pro-Kontra Bahaya Rokok Elektrik*
2	Jurnal	Jurnal TEKNOINFO
3	Volume, Nomor dan Halaman	Volume 16 No.1 Hal 93-99
4	Bulan dan Tahun	Januari 2022
5	Penulis	Eva Argarini Pratama, Corie Mei Hellyana, Nuzul Imam Fadlilah

6	Penerbit	Universitas Teknokrat Indonesia
7	Tujuan Penelitian	Penelitian ini bertujuan untuk membandingkan kecocokan tiga algoritma klasifikasi (*Naïve Bayes*, *Decision Tree*, dan *Logistic Regression*) dalam menganalisis data mining terkait pro dan kontra bahaya rokok elektrik.
8	Lokasi dan Subjek Penelitian	Informasi diperoleh melalui survei menggunakan pertanyaan terbuka, kemudian diproses dan disimpan dalam format Excel. Data tersebut dikumpulkan dari berbagai kota besar di Indonesia, termasuk Jakarta, Medan, Semarang, Surabaya, dan kota-kota lainnya.
9	Perancangan Sistem	a) Algoritma *Naive Bayes* b) Algoritma *Decision Tree* c) Algoritma *Logistic Regression* d) *RapidMiner* e) Penambangan data
10	Metode yang digunakan	Data dikumpulkan melalui kuesioner yang berisi pertanyaan-pertanyaan, lalu diolah dan dianalisis menggunakan perangkat lunak Microsoft Excel.
11	Hasil Penelitian	Penelitian menunjukkan bahwa algoritma *Decision Tree* memiliki tingkat akurasi tertinggi (81,00%) dalam mengklasifikasikan pro dan kontra terkait bahaya rokok elektrik. Sementara itu, algoritma *Naive Bayes* dan *Logistic Regression* masing-masing memiliki tingkat akurasi 73,33% dan 79,67%.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Penelitian ini berhasil membandingkan kinerja tiga algoritma berbeda dalam mengklasifikasikan sentimen pro dan kontra terhadap bahaya rokok elektrik. b. Penelitian ini bertujuan untuk menentukan algoritma mana yang paling akurat dalam mengklasifikasikan sentimen tersebut. c. Peneliti menggunakan pendekatan komparatif untuk mengidentifikasi metode yang paling efektif dalam klasifikasi sentimen pro dan kontra terhadap bahaya rokok elektrik.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Penelitian ini memiliki keterbatasan karena belum mengembangkan aplikasi desktop atau web yang dapat membantu mengklasifikasikan pro dan kontra terkait bahaya rokok elektrik.
14	Kesimpulan	Berdasarkan hasil analisis, algoritma *Naive Bayes* menunjukkan performa terbaik dalam mengklasifikasikan sentimen pro-kontra terhadap rokok elektrik, dengan akurasi 73,33% dan nilai AUC 0.679. Meskipun akurasi *Naive Bayes* lebih rendah dibandingkan *Decision Tree* (81,00%) dan *Logistic Regression* (79,67%), nilai AUC yang lebih tinggi mengindikasikan kemampuan Naive Bayes yang superior dalam membedakan kelas positif dan negatif, menjadikannya algoritma klasifikasi yang sangat baik dalam konteks ini.
15	Tautan Jurnal	https://ejurnal.teknokrat.ac.id/index.php/teknoinfo/article/view/1534

2.5.2 Jurnal 2

Tabel 2.2 Jurnal 2

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes*
2	Jurnal	Jurnal SAINTEKOM
3	Volume, Nomor dan Halaman	Volume 13 No.1 Hal 42-54
4	Bulan dan Tahun	31 Maret 2023
5	Penulis	Agus Fajar Riany dan Gusmelia Testiana
6	Penerbit	STMIK Palangkaraya
7	Tujuan Penelitian	Penelitian ini bertujuan mengembangkan model klasifikasi penyakit *stroke* menggunakan algoritma *Naive Bayes* dalam kerangka *data mining*. Akurasi model ini akan dievaluasi melalui perhitungan manual dan pengujian menggunakan perangkat lunak *RapidMiner*.
8	Lokasi dan Subjek Penelitian	Data didapatkan dari *dataset repository Kaggle* tentang *Brain Stroke*. Dengan subjek penelitian tentang penyakit *stroke* pada otak.
9	Perancangan Sistem	a) Penambangan data *b) RapidMiner* c) Algoritma *Naive Bayes*
10	Metode yang digunakan	Data diperoleh melalui pengumpulan data sekunder yang bersumber dari situs web *Kaggle*.
11	Hasil Penelitian	Penelitian menunjukkan bahwa algoritma *Naive Bayes* berhasil mengklasifikasikan penyakit *stroke* dengan akurasi 92,48%. Secara rinci, terdapat 5 prediksi positif yang benar, 30 prediksi positif yang salah, 917 prediksi negatif yang benar, dan 45 prediksi negatif yang salah.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Tingkat akurasi yang tinggi, mencapai 92,48%. b. Perhitungan manual algoritma yang dilakukan oleh peneliti.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Belum mengembangkan aplikasi desktop atau web untuk membantu klasifikasi penyakit *stroke*. b. Belum membandingkan metode algoritma yang digunakan dengan metode lain.
14	Kesimpulan	Berdasarkan hasil penelitian, algoritma *Naive Bayes* dalam *data mining* terbukti efektif dalam mengklasifikasikan penyakit *stroke*. Dari data yang dianalisis, 35 data terklasifikasi menderita *stroke*, sementara 962 data tidak. Algoritma ini menunjukkan tingkat akurasi yang tinggi, yaitu 92,48%, mengindikasikan bahwa metode ini sangat akurat dalam mengklasifikasikan penyakit *stroke*.
15	Tautan Jurnal	https://ojs.stmikplk.ac.id/index.php/saintekom/article/view/352

2.5.3 Jurnal 3

Tabel 2.3 Jurnal 3

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Penerapan Data Mining Untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes*
2	Jurnal	The 2nd MDP Student Conference 2023
3	Volume, Nomor dan Halaman	Volume 2 No.1 Hal 297-305
4	Bulan dan Tahun	10 April 2023
5	Penulis	Agus Fajar Riany dan Gusmelia Testiana
6	Penerbit	Universitas Multi Data Palembang
7	Tujuan Penelitian	Penelitian ini bertujuan untuk mengembangkan model klasifikasi penyakit jantung koroner menggunakan algoritma *Naive Bayes* dalam *data mining*. Akurasi model ini akan dievaluasi melalui perhitungan manual dan pengujian menggunakan perangkat lunak *RapidMiner*.
8	Lokasi dan Subjek Penelitian	*Dataset* yang digunakan dalam penelitian tentang penyakit jantung koroner ini berasal dari repositori *Kaggle*. Terdapat 4.133 data yang terdiri dari 16 atribut, dengan 15 atribut sebagai data penjelas dan 1 atribut sebagai target penelitian.
9	Perancangan Sistem	a) Penambangan data b) *RapidMiner* c) Algoritma *Naive Bayes*
10	Metode yang digunakan	Data yang digunakan dalam penelitian ini merupakan data sekunder yang diambil dari *Kaggle*.
11	Hasil Penelitian	Penelitian ini menunjukkan bahwa algoritma *Naive Bayes* dapat memprediksi penyakit jantung koroner dalam 10 tahun mendatang dengan akurasi 79,10%. Secara rinci, model ini berhasil memprediksi 42 kasus positif dengan benar dan 609 kasus negatif dengan benar. Namun, terdapat 86 kasus positif dan 86 kasus negatif yang salah diprediksi.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Penelitian ini menunjukkan tingkat akurasi yang cukup tinggi, mencapai 79,10%. b. Algoritma yang digunakan telah diverifikasi melalui perhitungan manual oleh peneliti.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Belum adanya pengembangan aplikasi desktop atau web untuk membantu memprediksi risiko penyakit jantung koroner dalam 10 tahun ke depan.

		b. Tidak dilakukannya perbandingan dengan metode algoritma lain dalam memprediksi risiko penyakit jantung koroner.
14	Kesimpulan	Secara keseluruhan, penelitian ini menunjukkan bahwa penggunaan data mining dengan algoritma *Naive Bayes* dapat secara efektif memprediksi penyakit jantung koroner dalam jangka waktu 10 tahun ke depan. Tingkat akurasi yang dihasilkan mencapai 79,10%, menunjukkan bahwa metode ini cukup akurat dan menjanjikan dalam membantu peneliti dan profesional medis untuk mengidentifikasi individu yang berisiko tinggi terkena penyakit ini.
15	Tautan Jurnal	https://jurnal.mdp.ac.id/index.php/msc/article/view/4388

2.5.4 Jurnal 4

Tabel 2.4 Jurnal 4

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Implementasi Data Mining Untuk Klasifikasi Penyakit Asam Urat Menggunakan Algoritma C4.5*
2	Jurnal	Variance: Journal of Statistics and Its Applications
3	Volume, Nomor dan Halaman	Volume 5 No.1 Hal 25-36
4	Bulan dan Tahun	April 2023
5	Penulis	Gianovita Talarima, Ferry Kondo Lembang, Norisca Lewaherilla, Johan Bruiyf Bension
6	Penerbit	Universitas Pattimura
7	Tujuan Penelitian	Penelitian ini bertujuan menerapkan model klasifikasi berbasis algoritma C4.5 dari *data mining*, serta mengevaluasi akurasinya dalam mengklasifikasikan penyakit asam urat melalui perhitungan manual.
8	Lokasi dan Subjek Penelitian	Informasi mengenai penyakit asam urat pada 277 anggota Civitas Akademika Universitas Kristen Indonesia Maluku diperoleh melalui hasil pemeriksaan yang dilakukan pada tanggal 24 Juni 2022. Penelitian ini membagi data pasien ke dalam tiga kategori berdasarkan tingkat keparahan penyakit, yaitu berat, ringan, dan tidak ada.
9	Perancangan Sistem	a) Algoritma *Decision Tree* b) Penambangan data
10	Metode yang digunakan	Data yang digunakan dalam penelitian ini adalah data sekunder yang diperoleh dari hasil *screening* Civitas Akademika Universitas Kristen Indonesia Maluku.
11	Hasil Penelitian	Penelitian terhadap 277 pasien di Civitas Akademika UKIM menggunakan algoritma C4.5 menunjukkan bahwa mayoritas penderita asam urat adalah remaja, dengan 85 pasien perempuan

		dan 70 pasien laki-laki. Algoritma ini menghasilkan 8 aturan klasifikasi dengan tingkat akurasi 78%.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Tingkat akurasi yang mencapai 79,10%. b. Perhitungan manual algoritma yang dilakukan oleh peneliti.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Aplikasi yang digunakan untuk mengimplementasikan *data mining* tidak dirancang dan dikembangkan secara khusus untuk penelitian ini, baik dalam bentuk desktop maupun web. b. Proses pengujian algoritma tidak dilakukan menggunakan perangkat lunak RapidMiner. c. Tidak dilakukan perbandingan antara metode algoritma yang digunakan dengan algoritma lain yang relevan.
14	Kesimpulan	Kesimpulannya adalah penerapan *data mining* untuk mengklasifikasikan pasien penderita asam urat membantu peneliti untuk menemukan hasil yang akurat dengan menggunakan algoritma C4.5. Penelitian ini menghasilkan bahwa sebagian besar pasien penderita asam urat berada di usia remaja, dengan total pasien perempuan sebanyak 85 data, sedangkan pasien laki-laki sebanyak 70 data, dengan tingkat akurasi sebesar 78% dan dikategorikan sebagai cukup akurat.
15	Tautan Jurnal	https://ojs3.unpatti.ac.id/index.php/variance/article/view/10152

2.5.5 Jurnal 5

Tabel 2.5 Jurnal 5

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naives Bayes dan K-Nearest Neighbor*
2	Jurnal	INTECOMS: Journal of Information Technology and Computer Science
3	Volume, Nomor dan Halaman	Volume 6 No.1 Hal 416-428
4	Bulan dan Tahun	Juni 2023
5	Penulis	Fitrokh Nur Ikhromr, Ipin Sugiyarto, Umi Faddillah, Bibit Sudarsono
6	Penerbit	Institut Penelitian Matematika, Komputer, Keperawatan, Pendidikan dan Ekonomi
7	Tujuan Penelitian	Penelitian ini bertujuan menerapkan dan membandingkan kinerja dua algoritma *data mining*, yaitu *Naive Bayes* dan *k-Nearest Neighbor*, dalam mengklasifikasikan penyakit diabetes. Akurasi

		kedua model tersebut akan dievaluasi melalui perhitungan manual dan pengujian untuk menentukan algoritma yang lebih efektif dalam prediksi penyakit diabetes.
8	Lokasi dan Subjek Penelitian	Sumber *National Institute of Diabetes and Digestive and Kidney Diseases* menyediakan 2.000 data pasien diabetes, yang terdiri dari 9 atribut prediksi dan 2 kemungkinan hasil kelas prediksi.
9	Perancangan Sistem	a) Algoritma *Naive Bayes* b) Algoritma *k-Nearest Neighbor* *c) RapidMiner* d) Penambangan data
10	Metode yang digunakan	Data sekunder dikumpulkan untuk penelitian ini, yang bersumber dari *National Institute of Diabetes and Digestive and Kidney Diseases*.
11	Hasil Penelitian	Penelitian menunjukkan bahwa algoritma *K-Nearest Neighbor* (KNN) lebih unggul dalam memprediksi diabetes dibandingkan Naive Bayes. Ketika menggunakan 2.000 data, KNN mencapai akurasi 99%, jauh melampaui *Naive Bayes* dengan akurasi 75%. Namun, ketika menggunakan 30 data uji dan data latih, akurasi KNN turun menjadi 53%, sementara *Naive Bayes* sedikit lebih baik dengan akurasi 66%. Meskipun demikian, secara keseluruhan, KNN terbukti lebih efektif dalam klasifikasi *dataset* "*Diabetes Prediction Using Logistic Regression*".
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Tingkat akurasi KNN yang mencapai 99%. b. Peneliti menggunakan 2 algoritma yang berbeda untuk memprediksi diabetes. c. Peneliti mencari tingkat akurasi mana yang paling tinggi dalam memprediksi diabetes. d. Peneliti menunjukkan hasil perhitungan manual dari kedua algoritma tersebut.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Belum mengembangkan aplikasi yang dapat membantu proses prediksi.
14	Kesimpulan	Secara ringkas, penelitian ini membandingkan metode Naive Bayes dan KNN (*K-Nearest Neighbors*) dalam memprediksi penyakit diabetes menggunakan *data mining*. Pada penelitian ini ditunjukkan klasifikasi dengan 2.000 data menunjukkan tingkat akurasi sebesar 99% untuk KNN dan 75% untuk *Naive Bayes*, sedangkan dengan data uji dan data latih sebanyak 30 data, KNN menunjukkan tingkat akurasi sebesar 53% sedangkan *Naive Bayes* sebesar 66%. Jadi metode KNN lebih baik untuk memprediksi penyakit diabetes dengan jumlah data yang lebih besar.
15	Tautan Jurnal	https://journal.ipm2kpe.or.id/index.php/INTECOM/article/view/5916

2.5.6 Jurnal 6

Tabel 2.6 Jurnal 6

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Heart Disease Prediction using Data Mining Techniques*
2	Jurnal	International Journal of Engineering Research & Technology (IJERT)
3	Volume, Nomor dan Halaman	Volume 10 No.2 Hal 281-286
4	Bulan dan Tahun	Februari 2021
5	Penulis	Pratiksha Shetgaonkar, Dr. Shailendra Aswale
6	Penerbit	IJERT
7	Tujuan Penelitian	Penelitian ini bertujuan untuk membandingkan kinerja algoritma *Neural Network*, *Decision Tree*, dan *Naive Bayes* dalam memprediksi penyakit jantung menggunakan teknik klasifikasi data mining, serta menentukan algoritma yang paling akurat untuk prediksi tersebut.
8	Lokasi dan Subjek Penelitian	Data mengenai penderita penyakit jantung diperoleh dari *dataset Heart Disease* yang tersedia di *UCI Machine Learning Repository*.
9	Perancangan Sistem	a) Algoritma *Decision Tree* b) Algoritma *Naive Bayes* c) Algoritma *Neural Network* d) Penambangan data
10	Metode yang digunakan	Metode pengumpulan data yang digunakan dalam penelitian ini adalah melalui data sekunder yang diperoleh dari repositori UCI.
11	Hasil Penelitian	Penelitian terhadap 668 catatan dengan 14 atribut menunjukkan bahwa algoritma *Decision Tree* mencapai akurasi tertinggi (98,54%) dalam memprediksi penyakit jantung, diikuti oleh *Naive Bayes* (85,01%) dan *Neural Network* (81,83%).
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Mengidentifikasi algoritma terbaik di antara tiga metode yang diujikan untuk memprediksi penyakit jantung. b. Membandingkan kinerja algoritma-algoritma tersebut secara langsung untuk menentukan prediktor penyakit jantung yang paling akurat.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Belum adanya aplikasi yang dirancang untuk mempermudah proses prediksi. b. Tidak disajikannya contoh perhitungan manual yang dilakukan dalam penelitian.
14	Kesimpulan	Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* memiliki tingkat akurasi tertinggi

		(98,54%) dalam memprediksi penyakit jantung, dibandingkan dengan *Naive Bayes* (85,01%) dan *Neural Network* (81,83%). Dengan demikian, *Decision Tree* dapat dianggap sebagai metode algoritma yang paling akurat dan efektif untuk prediksi penyakit jantung dalam konteks penelitian ini.
15	Tautan Jurnal	https://www.ijert.org/heart-disease-prediction-using-data-mining-techniques

2.5.7 Jurnal 7

Tabel 2.7 Jurnal 7

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Review on Effective Disease Prediction through Data Mining Techniques*
2	Jurnal	International Journal on Electrical Engineering and Informatics
3	Volume, Nomor dan Halaman	Volume 13 No.3 Hal 717-733
4	Bulan dan Tahun	September 2021
5	Penulis	Muhammad Nabeel, Shumaila Majeed, Mazhar Javed Awan, Hooria Muslih-ud-Din, Mashal Wasique, Rabia Nasir
6	Penerbit	Institut Teknologi Bandung
7	Tujuan Penelitian	Penelitian ini bertujuan untuk mengevaluasi berbagai metode prediksi penyakit menggunakan teknik data mining yang efektif, serta mengidentifikasi metode mana yang memberikan hasil prediksi terbaik.
8	Lokasi dan Subjek Penelitian	*UCI Machine Learning Repository* menjadi sumber data untuk penelitian penyakit jantung, diabetes, hati, dan ginjal dalam penelitian ini.
9	Perancangan Sistem	a) Algoritma *Multilayer Perceptron*. b) Algoritma SMO. c) Algoritma *Random Forest*. d) Algoritma *Vote*. e) Algoritma *Decision Table*. f) Algoritma *Naive Bayes*. g) Algoritma J48. h) Penambangan data.
10	Metode yang digunakan	Penelitian ini menggunakan data sekunder yang bersumber dari repositori UCI.
11	Hasil Penelitian	Penelitian ini menghasilkan algoritma *Random Forest* yang lebih baik untuk memprediksi penyakit jantung dengan tingkat akurasi 83,77%, algoritma SMO yang lebih baik untuk memprediksi penyakit diabetes dengan tingkat akurasi 77,34%, algoritma SMO lagi

		yang lebih baik untuk memprediksi penyakit hati dengan tingkat akurasi 76,44%, dan algoritma *Random Forest* yang lebih baik untuk memprediksi penyakit ginjal.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Penelitian ini telah menghasilkan algoritma mana yang lebih baik antara tujuh metode algoritma yang digunakan dari masing-masing penyakit. b. Penelitian menggunakan metode algoritma pembanding untuk menemukan algoritma manakah yang lebih baik dalam memprediksi berbagai macam penyakit c. Penelitian menggunakan lebih dari tiga algoritma untuk membandingkan, dalam hal ini tujuh.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Belum adanya pengembangan aplikasi untuk mempermudah perbandingan dan prediksi. b. Tidak adanya dokumentasi hasil perhitungan manual yang dilakukan.
14	Kesimpulan	Kesimpulannya adalah algoritma *Random Forest* lebih baik untuk memprediksi penyakit jantung dan ginjal dengan tingkat akurasi 83,77% untuk penyakit jantung dan sayangnya peneliti tidak menyebutkan tingkat akurasi untuk penyakit ginjal. Sedangkan algoritma SMO lebih baik untuk memprediksi penyakit diabetes dan hati dengan tingkat akurasi 77,34% untuk penyakit diabetes dan tingkat akurasi 76,44% untuk penyakit hati.
15	Tautan Jurnal	https://ijeei.org/archives-number-71.html

2.5.8 Jurnal 8

Tabel 2.8 Jurnal 8

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes*
2	Jurnal	International Journal of Advanced Networking and Applications
3	Volume, Nomor dan Halaman	Volume 12 No.1 Hal 4509-4518
4	Bulan dan Tahun	15 Agustus 2020
5	Penulis	Gajendra Sharma, Umesh Hengaju
6	Penerbit	IJANA

7	Tujuan Penelitian	Penelitian ini bertujuan untuk menentukan algoritma terbaik dalam mengidentifikasi dan mengelompokkan pasien yang menderita diabetes.
8	Lokasi dan Subjek Penelitian	*Dataset Pima Indian Diabetes* diperoleh dari situs *Kaggle*, yang bersumber dari *UCI Machine Learning Repository*.
9	Perancangan Sistem	a) *RapidMiner*. b) Algoritma *Support Vector Machines*. c) Algoritma *Random Forest*. d) Algoritma *Naïve Bayes*. e) Algoritma *K-Nearest Neighbor*. f) Algoritma *Decision Tree*. g) Penambangan data.
10	Metode yang digunakan	Sumber data sekunder yang digunakan dalam penelitian ini berasal dari repositori UCI.
11	Hasil Penelitian	Evaluasi terhadap 768 data dengan 9 atribut menunjukkan performa beragam dalam klasifikasi indikasi diabetes menggunakan lima algoritma berbeda. Algoritma *Naive Bayes* mencatat akurasi tertinggi sebesar 76,30%, diikuti oleh *Decision Tree* (73,82%), *K-Nearest Neighbor* (71,65%), *Random Forest* (68,74%), dan *Support Vector Machines* (65,10%). Hasil ini mengindikasikan potensi *Naive Bayes* sebagai model klasifikasi yang paling efektif untuk *dataset* yang digunakan dalam penelitian ini.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Keberhasilan dalam mengidentifikasi algoritma terbaik di antara tujuh metode yang diterapkan pada masing-masing penyakit. b. Penelitian menggunakan metode algoritma pembanding untuk menemukan algoritma manakah yang lebih baik dalam memprediksi berbagai macam penyakit. c. Penelitian menggunakan lebih dari tiga algoritma untuk membandingkan, dalam hal ini lima. d. Penelitian telah melakukan hitung manual dari algoritma-algoritma tersebut.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Penelitian ini belum menghasilkan aplikasi yang dapat membantu pengguna dalam membandingkan data dan membuat prediksi secara lebih efisien.
14	Kesimpulan	Berdasarkan hasil evaluasi performa algoritma klasifikasi, *Naive Bayes* menunjukkan tingkat akurasi tertinggi sebesar 76,30% dalam mengidentifikasi individu dengan indikasi penyakit diabetes. Algoritma ini secara signifikan mengungguli *Decision Tree* (73,82%), *K-Nearest Neighbor* (71,65%), *Support Vector Machines* (65,10%), dan *Random Forest*

		(68,74%). Dengan demikian, Naive Bayes dapat direkomendasikan sebagai model yang paling efektif dan akurat untuk klasifikasi risiko diabetes dalam konteks penelitian ini.
15	Tautan Jurnal	https://www.ijana.in/v12-1.php

2.5.9 Jurnal 9

Tabel 2.9 Jurnal 9

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Neural Network Based Intelligent System for Predicting Heart Disease*
2	Jurnal	International Journal of Innovative Technology and Exploring Engineering (IJITEE)
3	Volume, Nomor dan Halaman	Volume 8 No.5 Hal 484-487
4	Bulan dan Tahun	Maret 2019
5	Penulis	K. Subhadra, Vikas B
6	Penerbit	Blue Eyes Intelligence Engineering & Sciences Publication
7	Tujuan Penelitian	Penelitian ini bertujuan untuk membandingkan kinerja berbagai algoritma (*Decision Tree, Naive Bayes, Logistic Regression, Random Forest, Support Vector Machine, Generalized Linear Model, Gradient Boosted Tree, Deep Learning, dan Multilayer Perceptron*) dalam memprediksi penyakit jantung, dengan tujuan mengidentifikasi metode yang paling akurat dan efektif.
8	Lokasi dan Subjek Penelitian	*UCI Machine Learning Repository* menjadi sumber data, dengan *dataset* yang digunakan adalah *Heart Disease*.
9	Perancangan Sistem	a) Algoritma *Logistic Regression*. b) Algoritma *Decision Tree*. c) Algoritma *Generalized Linear Model*. d) Algoritma *Naive Bayes*. e) Algoritma *Random Forest*. f) Algoritma *Support Vector Machine*. g) Algoritma *Deep Learning*. h) Algoritma *Multilayer Perception*. i) Algoritma *Gradient Boosted Tree*. j) Penambangan data.
10	Metode yang digunakan	Dalam penelitian ini, pengumpulan data dilakukan melalui pemanfaatan data sekunder yang bersumber dari repositori UCI (*UCI Machine Learning Repository*).
11	Hasil Penelitian	Evaluasi terhadap delapan algoritma *data mining* pada *dataset* penyakit jantung yang terdiri dari 303 sampel

		(297 lengkap, 6 tidak lengkap) dan 14 atribut menghasilkan performa akurasi yang bervariasi. Algoritma *Naive Bayes* menunjukkan performa terbaik dengan akurasi mencapai 90,2%, diikuti oleh *Multilayer Perceptron* dengan akurasi 94%. Sementara itu, *Logistic Regression, Random Forest, Generalized Linear Model*, dan *Gradient Boosted Tree* menunjukkan performa yang setara dengan akurasi sebesar 85,2%. Algoritma *Decision Tree* mencapai akurasi 83,6%, *Deep Learning* 88,5%, sedangkan *Support Vector Machine* menunjukkan performa terendah dengan akurasi 76,57%. Hasil ini mengindikasikan bahwa pemilihan algoritma yang tepat sangat krusial dalam prediksi penyakit jantung, dan *Naive Bayes* serta *Multilayer Perceptron* merupakan kandidat yang menjanjikan untuk pengembangan model prediksi lebih lanjut.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Penelitian ini berhasil mengidentifikasi algoritma dengan performa terbaik di antara sembilan algoritma yang diuji dalam memprediksi penyakit jantung pada pasien. b. Penelitian menggunakan metode algoritma pembanding untuk menemukan algoritma manakah yang lebih baik dalam memprediksi pasien penderita penyakit jantung. c. Penelitian menggunakan lebih dari tiga algoritma untuk membandingkan, dalam hal ini sembilan.
13	Kelemahan Penelitian	Kelemahan dari penelitian ini adalah: a. Belum adanya perancangan dan pembuatan aplikasi yang dapat membantu perbandingan dan prediksi. b. Tidak dilakukannya perhitungan manual terhadap algoritma yang digunakan.
14	Kesimpulan	Evaluasi perbandingan terhadap sembilan algoritma klasifikasi dalam prediksi penyakit jantung menunjukkan bahwa *Multilayer Perceptron* (MLP) mencapai akurasi tertinggi sebesar 94%. Performa MLP secara signifikan melebihi algoritma lain seperti *Decision Tree* (83,6%), *Logistic Regression* (85,2%), *Naive Bayes* (90,2%), *Random Forest* (85,2%), *Support Vector Machine* (76,57%), *Generalized Linear Model* (85,2%), *Gradient Boosted Tree* (85,2%), dan *Deep Learning* (88,5%). Hasil ini menempatkan MLP sebagai model optimal untuk prediksi penyakit jantung dengan tingkat akurasi yang sangat memuaskan.
15	Tautan Jurnal	https://www.ijitee.org/portfolio-item/D2770028419/

2.5.10 Jurnal 10

Tabel 2.10 Jurnal 10

No.	Data Jurnal/Makalah	Keterangan
1	Judul	*Predicting Diabetes by adopting Classification Approach in Data Mining*
2	Jurnal	International Journal on Informatics Visualization
3	Volume, Nomor dan Halaman	Volume 3 No.2-2 Hal 218-221
4	Bulan dan Tahun	30 Agustus 2019
5	Penulis	Rapinder Kaur
6	Penerbit	JOIV
7	Tujuan Penelitian	Penelitian ini berfokus pada analisis perbandingan performa algoritma *Support Vector Machines* (SVM) dan *K-Nearest Neighbor* (KNN) dalam klasifikasi prediksi penyakit diabetes. Tujuannya adalah mengidentifikasi model yang memiliki tingkat akurasi dan kesesuaian yang lebih tinggi dalam prediksi penyakit diabetes.
8	Lokasi dan Subjek Penelitian	Penelitian ini menggunakan *dataset* yang berasal dari *UCI Machine Learning Repository Site*, yang terdiri dari 9 atribut yang menggambarkan data pasien penderita diabetes.
9	Perancangan Sistem	a) Algoritma *K-Nearest Neighbor*. b) Algoritma *Support Vector Machines*. c) Penambangan data.
10	Metode yang digunakan	Data sekunder diperoleh dari repositori UCI sebagai sumber data dalam penelitian ini.
11	Hasil Penelitian	Hasil penelitian mengindikasikan bahwa penerapan algoritma *Support Vector Machine* dalam prediksi risiko penyakit jantung, dengan memanfaatkan 9 atribut pasien, menghasilkan tingkat akurasi sebesar 80%. Di sisi lain, algoritma *K-Nearest Neighbor* menunjukkan performa yang lebih unggul, dengan tingkat akurasi mencapai 83,16% dalam konteks yang sama.
12	Kekuatan Penelitian	Keunggulan penelitian ini terletak pada beberapa hal: a. Menentukan algoritma terbaik dari dua pilihan yang tersedia melalui perbandingan langsung. b. Memanfaatkan algoritma pembanding untuk mengidentifikasi metode prediksi penyakit diabetes yang lebih unggul.
13	Kelemahan Penelitian	Kekurangan penelitian ini terletak pada beberapa hal: a. Tidak dikembangkan aplikasi yang dapat mempermudah proses perbandingan dan prediksi. b. Algoritma yang digunakan tidak dihitung secara manual.

14	Kesimpulan	Evaluasi performa algoritma klasifikasi dalam memprediksi diabetes menunjukkan bahwa algoritma *K-Nearest Neighbor* (k-NN) mencapai tingkat akurasi yang lebih tinggi (83,16%) dibandingkan algoritma *Support Vector Machines* (SVM) yang hanya mencapai 80%. Berdasarkan hasil tersebut, disimpulkan bahwa k-NN merupakan model yang lebih optimal dalam mendiagnosis diabetes dengan tingkat akurasi yang cukup tinggi.
15	Tautan Jurnal	https://joiv.org/index.php/joiv/article/view/229

2.5.11 Rangkuman Model Penelitian

Tabel 2.11 Perbandingan Jurnal

Peneliti	Nama Jurnal	Tahun	Institusi	Judul dan Metode yang digunakan	Kesimpulan
Eva Argarini Pratama, Corie Mei Hellyana, Nuzul Imam Fadlilah	Jurnal TEKNOINFO, Vol. 16, No. 1, ISSN: 2615-224X	2022	Universitas Teknokrat Indonesia	*Perbandingan 3 Algoritma Klasifikasi Data Mining Dalam Pro-Kontra Bahaya Rokok Elektrik* Metode yang digunakan adalah klasifikasi *Data Mining*	Berdasarkan hasil analisis, algoritma *Naive Bayes* menunjukkan performa terbaik dalam mengklasifikasikan sentimen pro-kontra terhadap rokok elektrik, dengan akurasi 73,33% dan nilai AUC 0.679. Meskipun akurasi *Naive Bayes* lebih rendah dibandingkan *Decision Tree* (81,00%) dan *Logistic Regression* (79,67%), nilai AUC yang lebih tinggi mengindikasikan kemampuan Naive Bayes yang superior dalam membedakan kelas positif dan negatif, menjadikannya algoritma klasifikasi yang sangat baik dalam konteks ini.
Agus Fajar Riany, Gusmelia Testiana	Jurnal SAINTEKOM, Vol. 13 No. 1, E-ISSN: 2503-3247 P-ISSN: 2088-1770	2023	STMIK Palangkaraya	*Penerapan Data Mining untuk Klasifikasi Penyakit Stroke Menggunakan Algoritma Naïve Bayes* Metode yang digunakan adalah	Berdasarkan hasil penelitian, algoritma *Naive Bayes* dalam *data mining* terbukti efektif dalam mengklasifikasikan penyakit *stroke*. Dari data yang dianalisis, 35 data terklasifikasi menderita *stroke*, sementara 962 data tidak. Algoritma ini menunjukkan tingkat akurasi yang tinggi, yaitu 92,48%, mengindikasikan bahwa metode ini

				klasifikasi *data mining Naive Bayes*	sangat akurat dalam mengklasifikasikan penyakit *stroke*.
Agus Fajar Riany, Gusmelia Testiana	The 2nd MDP Student Conference 2023, Vol. 2 No. 1, E-ISSN: 2985-7406	2023	Universitas Multi Data Palembang	*Penerapan Data Mining untuk Klasifikasi Penyakit Jantung Koroner Menggunakan Algoritma Naïve Bayes* Metode yang digunakan adalah metode klasifikasi *data mining Naive Bayes*	Secara keseluruhan, penelitian ini menunjukkan bahwa penggunaan data mining dengan algoritma *Naive Bayes* dapat secara efektif memprediksi penyakit jantung koroner dalam jangka waktu 10 tahun ke depan. Tingkat akurasi yang dihasilkan mencapai 79,10%, menunjukkan bahwa metode ini cukup akurat dan menjanjikan dalam membantu peneliti dan profesional medis untuk mengidentifikasi individu yang berisiko tinggi terkena penyakit ini.
Gianovita Talarima, Ferry Kondo Lembang, Norisca Lewaherilla, Johan Bruiyf Bension	Variance: Journal of Statistics and Its Applications, Vol. 5 No. 1, E-ISSN: 2685-872X, P-ISSN: 2685-8738	2023	Universitas Pattimura	*Implementasi Data Mining Untuk Klasifikasi Penyakit Asam Urat Menggunakan Algoritma C4.5* Metode yang digunakan adalah metode klasifikasi *data mining* C4.5.	Penelitian terhadap 277 pasien di Civitas Akademika UKIM menggunakan algoritma C4.5 menunjukkan bahwa mayoritas penderita asam urat adalah remaja, dengan 85 pasien perempuan dan 70 pasien laki-laki. Algoritma ini menghasilkan 8 aturan klasifikasi dengan tingkat akurasi 78%.
Fitrokh Nur Ikhromr, Ipin Sugiyarto, Umi Faddillah, Bibit Sudarsono	INTECOMS: Journal of Information Technology and Computer Science, Vol. 6 No. 1, E-ISSN: 2614-1574, P-ISSN: 2621-3249	2023	Institut Penelitian Matematika, Komputer, Keperawatan, Pendidikan dan Ekonomi	*Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naives*	Secara ringkas, penelitian ini membandingkan metode Naive Bayes dan KNN (*K-Nearest Neighbors*) dalam memprediksi penyakit diabetes menggunakan *data mining*. Pada penelitian ini ditunjukkan klasifikasi dengan 2.000 data menunjukkan tingkat

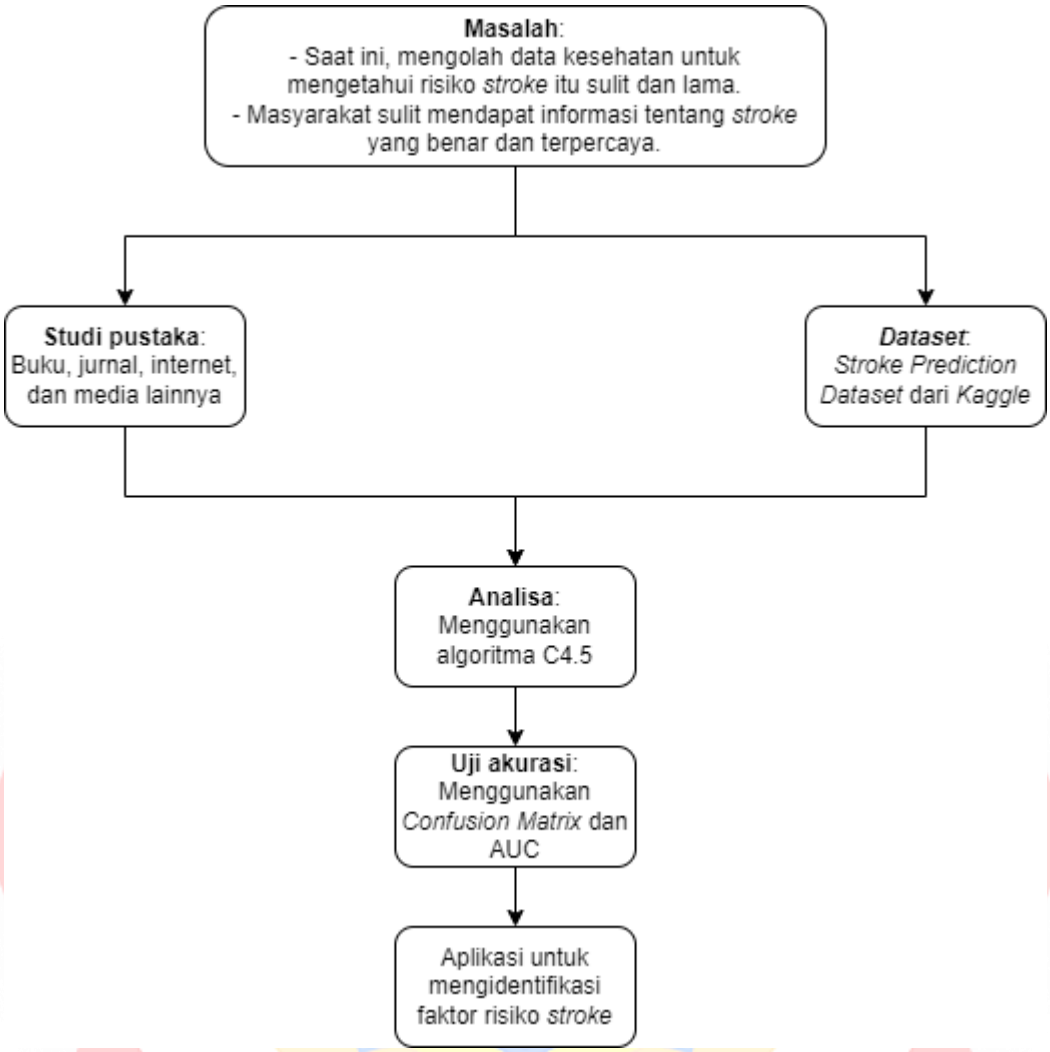
				Bayes dan K-Nearest Neighbor Metode yang digunakan adalah metode klasifikasi *Data Mining*.	akurasi sebesar 99% untuk KNN dan 75% untuk *Naive Bayes*, sedangkan dengan data uji dan data latih sebanyak 30 data, KNN menunjukkan tingkat akurasi sebesar 53% sedangkan *Naive Bayes* sebesar 66%. Jadi metode KNN lebih baik untuk memprediksi penyakit diabetes dengan jumlah data yang lebih besar.
Pratiksha Shetgaonkar, Dr. Shailendra Aswale	International Journal of Engineering Research & Technology (IJERT), Vol. 10 No. 2, ISSN: 2278-0181	2021	IJERT	*Heart Disease Prediction using Data Mining Techniques* Metode yang digunakan adalah metode klasifikasi *Data Mining*.	Hasil penelitian menunjukkan bahwa algoritma *Decision Tree* memiliki tingkat akurasi tertinggi (98,54%) dalam memprediksi penyakit jantung, dibandingkan dengan *Naive Bayes* (85,01%) dan *Neural Network* (81,83%). Dengan demikian, *Decision Tree* dapat dianggap sebagai metode algoritma yang paling akurat dan efektif untuk prediksi penyakit jantung dalam konteks penelitian ini.
Muhammad Nabeel, Shumaila Majeed, Mazhar Javed Awan, Hooria Muslih-ud-Din, Mashal Wasique, Rabia Nasir	International Journal on Electrical Engineering and Informatics, Vol. 13 No. 3, P-ISSN: 2085-6830, E-ISSN: 2087-5886	2021	Institut Teknologi Bandung	*Review on Effective Disease Prediction through Data Mining Techniques* Metode yang digunakan adalah metode klasifikasi *Data Mining*.	Kesimpulannya adalah algoritma Random Forest lebih baik untuk memprediksi penyakit jantung dan ginjal dengan tingkat akurasi 83,77% untuk penyakit jantung dan sayangnya peneliti tidak menyebutkan tingkat akurasi untuk penyakit ginjal. Sedangkan algoritma SMO lebih baik untuk memprediksi penyakit diabetes dan hati dengan tingkat akurasi 77,34% untuk penyakit diabetes

					dan tingkat akurasi 76,44% untuk penyakit hati.
Gajendra Sharma, Umesh Hengaju	International Journal of Advanced Networking and Applications, Vol. 12 No. 1, P-ISSN: 0975-0290, E-ISSN: 0975-0282	2020	IJANA	*Performance Analysis of Data Mining Classification Algorithm to Predict Diabetes* Metode yang digunakan adalah metode klasifikasi *Data Mining*.	Berdasarkan hasil evaluasi performa algoritma klasifikasi, *Naive Bayes* menunjukkan tingkat akurasi tertinggi sebesar 76,30% dalam mengidentifikasi individu dengan indikasi penyakit diabetes. Algoritma ini secara signifikan mengungguli *Decision Tree* (73,82%), *K-Nearest Neighbor* (71,65%), *Support Vector Machines* (65,10%), dan *Random Forest* (68,74%). Dengan demikian, Naive Bayes dapat direkomendasikan sebagai model yang paling efektif dan akurat untuk klasifikasi risiko diabetes dalam konteks penelitian ini.
K. Subhadra, Vikas B	International Journal of Innovative Technology and Exploring Engineering (IJITEE), Vol. 8 No. 5, ISSN: 2278-3075	2019	Blue Eyes Intelligence Engineering & Sciences Publication	*Neural Network Based Intelligent System for Predicting Heart Disease* Metode yang digunakan adalah metode klasifikasi *Data Mining*.	Evaluasi perbandingan terhadap sembilan algoritma klasifikasi dalam prediksi penyakit jantung menunjukkan bahwa *Multilayer Perceptron* (MLP) mencapai akurasi tertinggi sebesar 94%. Performa MLP secara signifikan melebihi algoritma lain seperti *Decision Tree* (83,6%), *Logistic Regression* (85,2%), *Naive Bayes* (90,2%), *Random Forest* (85,2%), *Support Vector Machine* (76,57%), *Generalized Linear Model* (85,2%), *Gradient Boosted Tree* (85,2%), dan *Deep Learning* (88,5%). Hasil ini menempatkan MLP sebagai model optimal untuk prediksi penyakit jantung

					dengan tingkat akurasi yang sangat memuaskan.
Rapinder Kaur	International Journal on Informatics Visualization, Vol.3 No. 2-2, ISSN: 2549-9610, E-ISSN: 2549-9904	2019	JOIV	*Predicting Diabetes by adopting Classification Approach in Data Mining* Metode yang digunakan adalah metode klasifikasi *Data Mining*.	Evaluasi performa algoritma klasifikasi dalam memprediksi diabetes menunjukkan bahwa algoritma *K-Nearest Neighbor* (k-NN) mencapai tingkat akurasi yang lebih tinggi (83,16%) dibandingkan algoritma *Support Vector Machines* (SVM) yang hanya mencapai 80%. Berdasarkan hasil tersebut, disimpulkan bahwa k-NN merupakan model yang lebih optimal dalam mendiagnosis diabetes dengan tingkat akurasi yang cukup tinggi.

Setelah membandingkan penelitian-penelitian sebelumnya, studi ini akan menerapkan algoritma C4.5 untuk menentukan metode terbaik dalam mendeteksi penyebab *stroke*. Meskipun banyak algoritma klasifikasi lain yang telah digunakan (seperti *Naive Bayes*, KNN, SVM, *Linear Regression*, *Random Forest*, dll.), C4.5 dipilih karena popularitas dan penggunaannya yang luas dalam klasifikasi *data mining*.

2.6 Kerangka Pemikiran

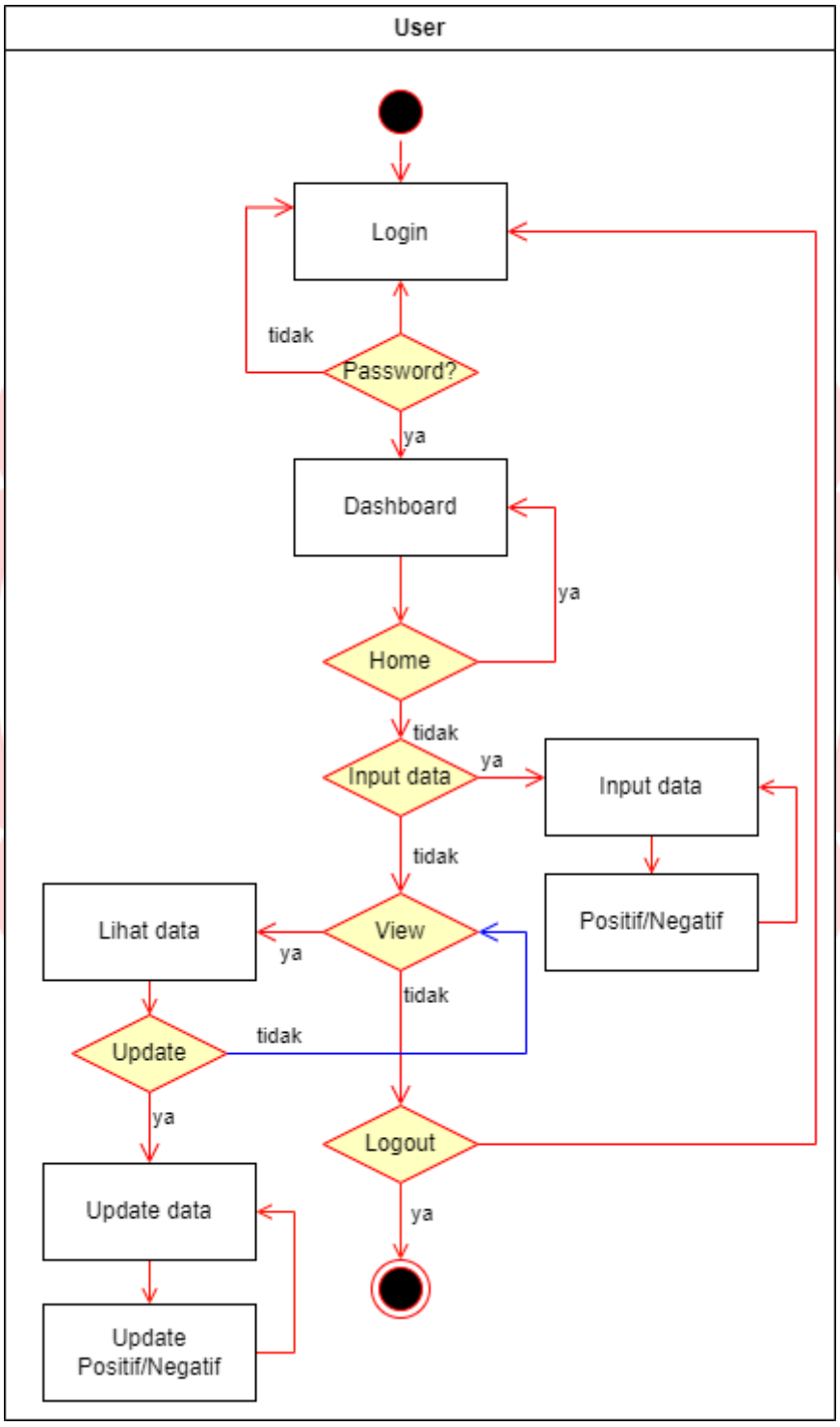


Gambar 2.1 Kerangka Pemikiran

BAB III

METODOLOGI PENELITIAN

3.1 *Activity Diagram*



Gambar 3.1 Activity Diagram

3.2 Analisa Kebutuhan

Pada tahap ini, penelitian ini mengumpulkan data mengenai penyakit *stroke* dan algoritma C4.5 untuk menguji metode algoritma tersebut dalam memprediksi faktor-faktor risiko *stroke*.

3.2.1 *Dataset Stroke Prediction*

Dataset yang akan diteliti didapat dari situs *dataset Kaggle*: https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset. *Dataset* tersebut memiliki 5110 *record* dengan 12 atribut. Atribut-atribut tersebut antara lain:

1	Id
2	Jenis Kelamin (*gender*)
3	Umur (*age*)
4	Hipertensi (*hypertension*)
5	Penyakit Jantung (*heart_disease*)
6	Pernah Menikah (*ever_married*)
7	Pekerjaan (*work_type*)
8	Tipe Tempat Tinggal (*Residence_type*)
9	Tingkat Glukosa Rata-rata (*avg_glucose_lvl*)
10	BMI (*bmi*)
11	Status Perokok (*smoking_status*)
12	*Stroke* (*stroke*)

Tabel 3.1 Sampel *Dataset*

id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
9046	Male	67	0	1	Yes	Private	Urban	228,69	36,6	formerly smoked	Positive
31112	Male	80	0	1	Yes	Private	Rural	105,92	32,5	never smoked	Positive
60182	Female	49	0	0	Yes	Private	Urban	171,23	34,4	smokes	Positive
1665	Female	79	1	0	Yes	Self-employed	Rural	174,12	24	never smoked	Positive
56669	Male	81	0	0	Yes	Private	Urban	186,21	29	formerly smoked	Positive
30669	Male	3	0	0	No	children	Rural	95,12	18	Unknown	Negative
30468	Male	58	1	0	Yes	Private	Urban	87,96	39,2	never smoked	Negative
16523	Female	8	0	0	No	Private	Urban	110,89	17,6	Unknown	Negative
56543	Female	70	0	0	Yes	Private	Rural	69,04	35,9	formerly smoked	Negative
46136	Male	14	0	0	No	Never_worked	Rural	161,28	19,1	Unknown	Negative

Tabel 3.2 Deskripsi *Dataset*

Id	Merupakan penanda pada ID pasien
Gender	Menunjukkan jenis kelamin pasien
Age	Merupakan umur dari pasien tersebut
hypertension	Merupakan hasil dari apakah pasien tersebut terkena hipertensi, dengan nilai 0 untuk negatif hipertensi dan nilai 1 untuk positif hipertensi.

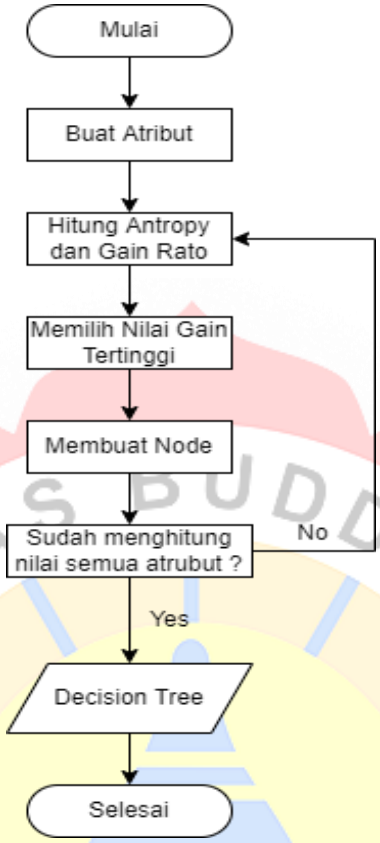
heart_disease	Merupakan hasil dari apakah pasien tersebut memiliki penyakit jantung, dengan nilai 0 untuk negatif penyakit jantung dan nilai 1 untuk positif penyakit jantung.
ever_married	Merupakan apakah pasien tersebut sudah menikah atau belum?
work_type	Merupakan pekerjaan tempat pasien tersebut berada, apakah di swasta (*private*), wiraswasta (*self-employed*), pemerintah (*govt_job*), anak-anak (*children*), atau tidak bekerja (*Never_worked*).
Residence_type	Merupakan lingkungan tempat tinggal dari pasien tersebut, apakah di perkotaan (*urban*) atau pedesaan (*rural*).
avg_glucose_lvl	Merupakan kadar tingkat glukosa dalam darah dari pasien tersebut.
Bmi	*Body Mass Index* (BMI), dalam bahasa Indonesia disebut Indeks Massa Tubuh (IMT), adalah metode untuk memperkirakan jumlah lemak tubuh berdasarkan berat dan tinggi badan seseorang. BMI digunakan untuk menentukan apakah berat badan seseorang tergolong ideal, kurang, atau berlebih. Cara menghitung BMI adalah dengan membagi berat badan (dalam kilogram) dengan kuadrat dari tinggi badan (dalam meter).
smoking_status	Merupakan apakah pasien tersebut sedang merokok (*smokes*), pernah merokok (*formerly smoked*), tidak merokok (*never smoked*), atau tidak diketahui (*Unknown*)?

Stroke	Merupakan apakah pasien tersebut terkena *stroke* atau tidak? Bila nilai 1 maka pasien tersebut adalah positif terkena *stroke* dan bila nilai 0 maka pasien tersebut adalah negatif terkena *stroke*.

3.3 Konstruksi Algoritma dan Metode

Klasifikasi merupakan suatu metode analisis data yang bertujuan untuk mengidentifikasi dan mengelompokkan pola-pola tersembunyi dalam suatu kumpulan data ke dalam kelas-kelas yang telah ditentukan sebelumnya. Proses ini menghasilkan suatu model klasifikasi (*classifier*) yang dapat digunakan untuk memprediksi kelas dari data baru yang belum diketahui labelnya. Salah satu contoh algoritma klasifikasi yang umum digunakan dalam pembelajaran mesin adalah C4.5, yang digunakan untuk membangun pohon keputusan. Proses ini melibatkan tiga tahap utama: (1) transformasi data mentah menjadi representasi pohon keputusan, (2) konversi model pohon keputusan menjadi serangkaian aturan klasifikasi, dan (3) penyederhanaan aturan-aturan tersebut untuk meningkatkan efisiensi dan interpretabilitas model.

Berikut ini adalah model untuk membuat proses klasifikasi:



Gambar 3.2 Proses Klasifikasi

Dari *dataset* penderita penyakit *stroke* yang diambil dari *Kaggle*, diolah menggunakan algoritma C4.5, kemudian setelah selesai diolah dilakukan pengujian algoritma tersebut dengan metode *Confusion Matrix* dan AUC. Hasil pengujian akurasi tersebut dijadikan sebagai bahan untuk membuat aplikasi berbasis *data mining* yang dapat digunakan untuk mendeteksi penyakit *stroke*.

Proses kerja algoritma C4.5 dalam penelitian ini dimulai dengan memilih atribut yang relevan. Selanjutnya, nilai gain tertinggi dihitung berdasarkan *entropy* dari semua atribut. Atribut dengan nilai *gain* tertinggi ditetapkan sebagai akar awal. Pohon keputusan kemudian dibangun secara rekursif dengan menerapkan proses yang sama, selalu memilih atribut dengan *gain* tertinggi pada setiap partisi. Penelitian ini menggunakan analisis deskriptif untuk memahami karakteristik data yang diteliti. Bagian ini akan membahas deskripsi dari setiap variabel yang digunakan dalam penelitian.

3.3.1 Analisis Data Mentah

Proses analisis ini bertujuan untuk memahami hubungan antara variabel dependen '*Stroke*' dan sejumlah variabel independen. Variabel independen tersebut meliputi *id*, *ever maried*, *work type*, *age*, *hypertension*, *heart disease*, *smoking status*, *resident type*, *avg glucose level*, BMI dan *Stroke*. Tabel deskripsi variabel akan memberikan gambaran rinci tentang karakteristik data mentah dari setiap variabel ini.

Tabel 3.3 Tabel Dataset

No	Variabel	Nilai	Keterangan	JumlahData
1	Id	Nominal	Kode identifikasi pasien dengan riwayat *stroke*	5110
2	Gender	Male	Jenis kelamin pasien penderita *stroke*	2115
		Female		2994
		Other		1
3	Age	Nominal	Kontinuitas usia pasien dilayani secara menyeluruh, dari balita hingga manula	5110
4	Hypertension	1	Memiliki riwayat hipertensi	498
		0	Tidak ada riwayat hipertensi	4612
5	Heart Disease	1	Memiliki riwayat penyakit jantung	276
		0	Tidak ada riwayat penyakit jantung	4834
6	Ever Maried	Yes	Menikah	3353
		No	Tidak menikah	1757
7	Work Type	Childrem	Anak-anak	687
		Govt Job	Pegawai	657

			pemerintahan	
		Never Worked	Belum pernah bekerja	22
		Private	Pekerjaan pribadi	2925
		Self Employed	Wiraswasta	819
8	Resident Type	Urban	Perkotaan	2596
		Rural	Pedesaan	2514
9	Avg Glucose Level	Nominal	Nilai rata-rata tingkat glukosa	5110
10	BMI	Nominal,N/A	Nilai indeks massa tubuh	5110
11	Smoking Status	Formerly Smoked	Pernah merokok	885
		Never Smoked	Belum pernah merokok	1892
		Smokes	Perokok	789
		Unknown	Status merokok yang tidak diketahui	1544
12	*Stroke*	1	*Stroke*	249
		0	Tidak *Stroke*	4861

3.3.2 Analisis Data Preprocessing

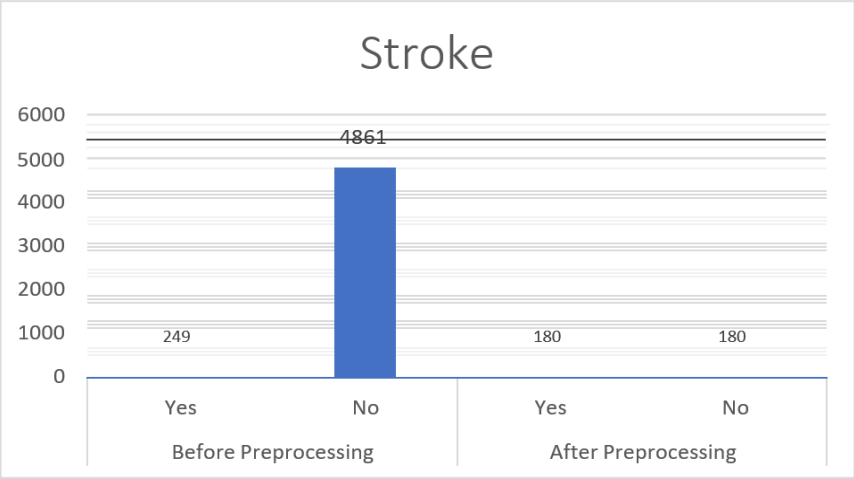
Analisis dilakukan dengan menyelidiki hubungan antara variabel target yaitu *Stroke* dengan berbagai faktor independen seperti *gender*, *age*, *hypertension*, *heart disease*, *smoking status*, *resident type*, *avg glucose level*, BMI dan *Stroke*. Sebelum menghasilkan visualisasi data *Stroke*, tahap penting praproses data dilakukan untuk mengatasi masalah potensial yang dapat mengganggu analisis. Praproses data melibatkan pembersihan dan transformasi data, termasuk menghapus data yang tidak relevan dan memastikan konsistensi format data untuk memudahkan pemrosesan oleh sistem. Langkah-langkah spesifik praproses data yang akan dilakukan akan dijelaskan lebih lanjut:

1. ***Data Cleaning***

Sebelum diproses, *dataset Kaggle* memiliki 5110 entri. Contoh data dari *Kaggle* disertakan. Selama pembersihan, 201 nilai kosong dan 1482 nilai "*unknown*" pada variabel BMI dihapus. Setelah pembersihan, tersisa 3427 entri. Karena *dataset* tidak seimbang, dilakukan penyeimbangan kelas setelah pembersihan.

$$Imbalanced\ Ratio\ (IR) = \frac{Majority\ Class}{Minority\ Class} = \frac{\text{Stroke Yes}}{\text{Stroke No}} = \frac{180}{3427}$$

Hasil perbandingan menunjukkan bahwa *dataset* yang digunakan tidak seimbang. Untuk mengatasi ketidakseimbangan ini, teknik *undersampling* diterapkan dengan mengurangi jumlah sampel dari kelas mayoritas sehingga kedua kelas memiliki jumlah data yang sama. Setelah proses *balancing*, jumlah data akhir menjadi 360. Tabel data yang telah dibersihkan memberikan gambaran perbandingan data sebelum dan sesudah proses *preprocessing*.



Gambar 3.3 Data proses *preprocessing*

2. ***Feature Selection***

Proses seleksi fitur bertujuan untuk menentukan variabel-variabel yang relevan dalam penelitian ini. Pemilihan ini didasarkan pada faktor-faktor yang dapat menyebabkan *stroke*. Variabel-variabel yang akan digunakan meliputi *gender, age, hypertension, heart disease, resident type, avg glucose level,* BMI*, smoking status*, dan kejadian *stroke* itu sendiri. Sementara itu, atribut-atribut seperti *id*, *ever married*, dan *work type* akan diabaikan. Tabel di bawah ini merangkum atribut-atribut yang akan dianalisis dalam penelitian ini.

Tabel 3.4 Tabel Proses *Selection*

No	Atribut	Keterangan	
1	Id	x	No
2	*Gender*	✓	Yes
3	*Age*	✓	Yes
4	*Hypertension*	✓	Yes
5	*Heart Disease*	✓	Yes
6	*Ever Married*	x	No
7	*Work Type*	x	No
8	*Residence Type*	✓	Yes
9	*Avg Glucose Level*	✓	Yes
10	BMI	✓	Yes
11	*Smoking Status*	✓	Yes
12	*Stroke*	✓	Label/Class

3. ***Data Transformation***

Transformasi data dilakukan agar data pada *dataset* sesuai dengan format yang dibutuhkan oleh aplikasi yang akan digunakan. Proses ini meliputi dua transformasi utama:

1. Perubahan atribut dari nominal menjadi numerik: Atribut seperti '*hypertension*', '*heart disease*', dan '*stroke*' diubah dari kategori (misalnya, "ya" atau "tidak") menjadi angka (misalnya, 1 atau 0).
2. *Discretization*: Nilai atribut yang kontinu (berkisar dalam rentang tertentu) diubah menjadi kategorik (dikelompokkan ke dalam beberapa kategori).

Tabel di bawah ini memberikan contoh proses transformasi dari nominal ke numerik.

Tabel 3.5 Data Setelah Proses *Transformation*

Atribut	Sebelum Transformasi	Setelah Transformasi
Hypertension	0	No
	1	Yes
Heart Disease	0	No
	1	Yes
Stroke	0	No
	1	Yes

Proses ini melibatkan transformasi data numerik kontinu menjadi data kategori dengan interval tertentu. Variabel yang akan diubah meliputi *age*, *avg glucose level*, dan BMI. Pengelompokan ini akan didasarkan pada kriteria-kriteria yang telah ditentukan sebelumnya.

A. Age

Menimbang berbagai faktor terkait usia, DPR RI menginisiasi perubahan terhadap Undang-Undang Nomor 13 Tahun 1998. Perubahan ini mengacu pada klasifikasi usia yang ditetapkan oleh Kementerian Kesehatan, yaitu: (L. N. Hakim, 2020)

a) 0 – 5 tahun (masa balita)

b) 6 – 11 tahun (masa kanak-kanak)

c) 12 – 16 tahun (masa remaja awal)

d) 17 – 25 tahun (masa remaja akhir)

e) 26 – 35 tahun (masa dewasa awal)

f) 36 – 45 tahun (masa dewasa akhir)

g) 46 – 55 tahun (masa lansia awal)

h) 56 – 65 tahun (masa lansia akhir)

i) Lebih dari 65 tahun (masa manula)

B. *Avg Glucose Level*

Kadar gula darah puasa normal, yang diukur dalam mg/dL, adalah sebagai berikut: (Lestari, 2024)

a) Kurang dari 101 mg/dL (Normal)

b) 101 – 125 mg/dL (Pradiabetes)

c) Lebih dari 125 mg/dL (Diabetes)

C. BMI

Indeks massa tubuh (IMT), juga dikenal sebagai BMI, adalah alat ukur umum yang dipakai untuk menilai apakah berat badan seseorang tergolong sehat atau tidak. Organisasi Kesehatan Dunia (WHO) telah menetapkan pedoman kategori berat badan berdasarkan IMT, yang berbeda untuk pria dan wanita. (Puji, 2024)

a) Kurang dari 101 mg/dL (Normal)

b) 101 – 125 mg/dL (Pradiabetes)

c) Lebih dari 125 mg/dL (Diabetes)

Proses perubahan data dari tipe kontinu menjadi kategorik secara detail dapat dilihat pada tabel berikut. Semua data telah melalui tahap *preprocessing*.

Tabel 3.6 Jenis Atribut Data Pasien *Stroke*

No	Atribut	Row Data Type	Ready Data Type
1	*Gender*	Kategorik Nominal	Kategorik Nominal
2	*Age*	Numerik Rasio	Kategorik Ordinal
3	*Hypertensi*	Kategorik Nominal	Kategorik Nominal
4	*Heart Disease*	Kategorik Nominal	Kategorik Nominal
5	*Resident Type*	Kategorik Nominal	Kategorik Nominal
6	*Avg Glucose Level*	Numerik Rasio	Kategorik Ordinal
7	*BMI*	Numerik Rasio	Kategorik Ordinal
8	*Smoking Status*	Kategorik Nominal	Kategorik Nominal
9	*Stroke*	Kategorik Nominal	Kategorik Nominal

3.4 Perhitungan Manual Metode Decision Tree C4.5

Tahap ini melibatkan pemilihan teknik dan pembuatan model. Penelitian ini menggunakan *Decision Tree* C4.5, yang menggunakan gain untuk memilih variabel yang akan menjadi node. Algoritma C4.5 menghitung gain ratio setiap atribut dan atribut dengan nilai tertinggi dipilih sebagai simpul. Sementara itu, *Entropy*(S) mengukur keragaman nilai atribut kriteria terhadap atribut keputusan dalam *dataset*. *Entropy* rendah menunjukkan keragaman rendah, sedangkan *entropy* tinggi menunjukkan keragaman tinggi.

Pembuatan model dimulai dengan menentukan node akar, diikuti dengan penentuan cabang dari setiap node. Selanjutnya, kelas-kelas dibagi pada cabang yang telah diperoleh. Proses ini diulang sampai setiap cabang memiliki kelas. Langkah awal dalam pembuatan pohon keputusan adalah menghitung nilai *entropy* total dan *entropy* dari setiap atribut. Sebagai contoh, perhitungan *entropy* total dan *entropy* dari variabel jenis kelamin dapat dilihat sebagai berikut. Menghitung *entropy* total:

$$Entropy\ total = ((-\frac{180}{360}) * \log_2(\frac{180}{360})) + ((-\frac{180}{360}) * \log_2(\frac{180}{360}))$$
$$= 1$$

Kemudian dihitung nilai *entropy* dan nilai *gain* dari variabel *Gender*

A. Male

Jumlah Kasus : 145
Ya 75
Tidak 70

$$Entropy\ total = ((-\frac{75}{145}) * \log_2(\frac{75}{145})) + ((-\frac{70}{145}) * \log_2(\frac{70}{145}))$$
$$= 0.9991$$

B. Female

Jumlah Kasus : 215
Ya 105
Tidak 110

$$Entropy\ total = ((-\frac{105}{215}) * \log_2(\frac{105}{215})) + ((-\frac{110}{215}) * \log_2(\frac{110}{215}))$$
$$= 0.9996$$

Selanjutnya, nilai *entropy* dihitung untuk setiap variabel bebas yang tersisa (*Age*, *Hypertension*, *Heart Disease*, *Smoking Status*, *Residence Type*, *Avg Glucose Level*, dan BMI). Tahap berikutnya melibatkan penghitungan nilai *gain* untuk masing-masing

variabel bebas ini. Sebagai contoh, perhitungan nilai *gain* untuk variabel Gender disajikan di bawah ini.

$$\textbf{Gain}\ (s, A) = 1 - \left(\frac{145}{360}\ \text{Entropy Male}\right) + \left(\frac{216}{360}\ \text{Entropy Female}\right)$$

$$= 0.00057$$

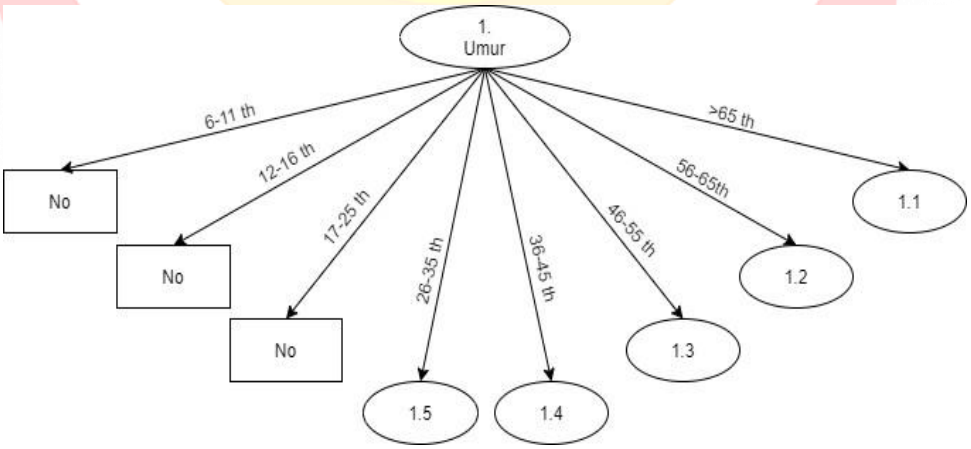
Selanjutnya, nilai gain dihitung untuk setiap variabel bebas lainnya (*Age*, *Hypertension*, *Heart Disease*, *Smoking Status*, *Residence Type*, *Avg Glucose Level*, dan BMI) menggunakan metode yang sama. Hasil perhitungan *entropy* dan *gain* untuk node akar dapat dilihat di bawah.

Tabel 3.7 Perhitungan *Node* Akar

No	Variabel	Nilai	Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Total		360	180	180	1	
1	*Gender*						0.00057
		Male	145	75	70	0.9991	
		Female	215	105	110	0.9996	
2	*Age*						0.19132
		6 - 11 th	2	0	2	0	
		12 - 16 th	2	0	2	0	
		17 - 25 th	20	0	20	0	
		26 - 35 th	16	1	15	0.3372	
		36 - 45 th	35	7	28	0.7219	
		46 - 55 th	54	24	30	0.9910	
		56 - 65 th	72	33	39	0.9949	
		Lebih dari 65 th	159	115	44	0.8509	
3	*Hypertension*						0.03073
		Yes	83	57	26	0.8968	
		No	277	123	154	0.9909	
4	*Heart Disease*						0.02886
		Yes	48	36	12	0.81128	
		No	312	144	168	0.99573	
5	*Smoking Status*						0.00772
		Formerly Smoked	104	57	47	0.9933	
		Never Smoked	181	84	97	0.9962	
		Smokes	75	39	39	0.9811	

6	*Resident Type*						0.00222
		Urban	178	94	84	0.9977	
		Rural	182	86	96	0.9978	
7	*Avg Glucose Level*						0.05424
		Kurang dari 101	187	77	110	0.9774	
		101 - 125	58	26	39	0.9039	
		Lebih dari 125	115	77	38	0.9153	
8	BMI						0.00698
		Kurang dari 18.6	5	1	4	0.7219	
		18.6 - 24.9	66	29	37	0.9893	
		25 - 29.9	120	64	56	0.9967	
		Lebih dari 29.9	169	86	83	0.9997	

Dari tabel tersebut, kita melihat bahwa '*Age*' memiliki nilai *gain* tertinggi (0.27169), menjadikannya akar pohon keputusan. '*Age*' punya 8 kategori, dan 3 kategori pertama (6-11, 12-16, 17-25 tahun) langsung menghasilkan keputusan '*No*'. Namun, 5 kategori lainnya (26-35, 36-45, 46-55, 56-65, dan 65+ tahun) perlu analisis lebih lanjut untuk menentukan keputusan akhirnya. Berdasarkan ini, kita bisa membuat gambaran awal pohon keputusan.



Gambar 3.4 Pohon Keputusan Node Akar

Setelah *node* akar dibuat, langkah berikutnya adalah menghitung ulang *node* berikutnya. Namun, data yang digunakan untuk perhitungan ini adalah data sisa dari komposisi kelas yang masuk ke *node* tersebut.

1. Perhitungan nilai variabel umur lebih dari 65 tahun

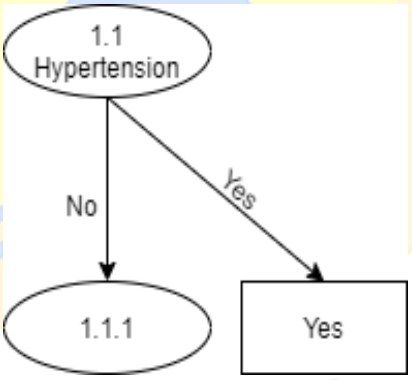
a) Node 1.1

Tabel 3.8 Perhitungan nilai variabel lebih dari 65 tahun

No	Atribut		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Age lebih dari 65 Tahun		159	115	44	0.8510	
1	*Gender*						0.0012
		Male	63	47	16	0.8175	
		Female	96	68	28	0.8709	
2	*Hypertension*						0.0180
		Yes	51	42	9	0.6723	
		No	108	73	35	0.9088	
3	*Heart Disease*						0.0024
		Yes	35	27	8	0.7755	
		No	124	88	36	0.8691	
4	*Smoking Status*						0.0037
		Formerly Smoked	48	37	11	0.7766	
		Never Smoked	90	63	27	0.8813	
		Smokes	21	15	6	0.8631	
5	*Resident Type*						0.0018
		Urban	84	59	25	0.8784	
		Rural	75	56	19	0.8165	
6	*Avg Glucose Level*						0.0160
		Kurang dari 101	71	48	23	0.9086	
		101 - 125	19	12	7	0.9495	
		Lebih dari 125	69	55	14	0.7277	

7	BMI						0.0133
		Kurang dari 18.6	2	1	1	1.0000	
		18.6 - 24.9	31	25	6	0.7088	
		25 - 29.9	54	42	12	0.7642	
		Lebih dari 29.9	72	46	26	0.9436	

Berdasarkan Tabel di atas, variabel paling berpengaruh adalah '*Hypertension*' dengan nilai gain 0.0180, sehingga menjadi cabang pertama (1.1) dalam pohon keputusan. Variabel '*Heart Disease*' memiliki dua nilai: '*Yes*' dan '*No*'. Jika nilai '*Yes*', maka keputusan langsung '*Yes*' tanpa perhitungan lanjutan. Namun, jika nilai '*No*', diperlukan perhitungan lebih lanjut. Hasil perhitungan ini menghasilkan gambaran awal pohon keputusan seperti di atas.



Gambar 3.5 *Node* 1.1

b) Node 1.1.1

Tabel 3.9 Perhitungan nilai variabel lebih dari 65 tahun

No	Atribut		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Hypertension (No)		108	73	35	0.9088	
1	*Gender*						0.0001

		Male	47	32	15	0.9035	
		Female	61	41	20	0.9127	
2	*Heart Disease*						0.0053
		Yes	24	18	6	0.8113	
		No	84	55	29	0.9297	
3	*Smoking Status*						0.0031
		Formerly Smoked	32	23	9	0.8571	
		Never Smoked	60	39	21	0.9341	
		Smokes	16	11	5	0.8960	
4	*Resident Type*						0.0028
		Urban	60	39	21	0.9341	
		Rural	48	34	14	0.8709	
5	*Avg Glucose Level*						0.0229
		Kurang dari 101	51	32	19	0.9526	
		101 - 125	16	9	6	0.9976	
		Lebih dari 125	41	32	9	0.7593	
6	BMI						0.0121
		Kurang dari 18.6	2	1	1	1	
		18.6 - 24.9	22	16	6	0.8454	
		25 - 29.9	36	27	9	0.8113	
		Lebih dari 29.9	48	28	20	0.9799	

Dari tabel di atas, kita melihat bahwa '*Avg Glucose Level*' memiliki nilai gain tertinggi (0.0229), menjadikannya cabang pertama (1.1.1) dalam pohon keputusan kita. Variabel ini memiliki tiga nilai: 'Kurang dari 101', '101 - 125', dan 'Lebih dari 125'. Nilai 'Kurang dari 101' dan 'Lebih dari 125' langsung menghasilkan keputusan '*Yes*', sehingga tidak perlu analisis lebih lanjut. Namun, nilai '101 - 125' memerlukan perhitungan tambahan untuk menentukan keputusannya. Berdasarkan hasil perhitungan ini, kita dapat membuat gambaran awal pohon keputusan seperti yang ditunjukkan.

Gambar 3.6 *Node* 1.1.1

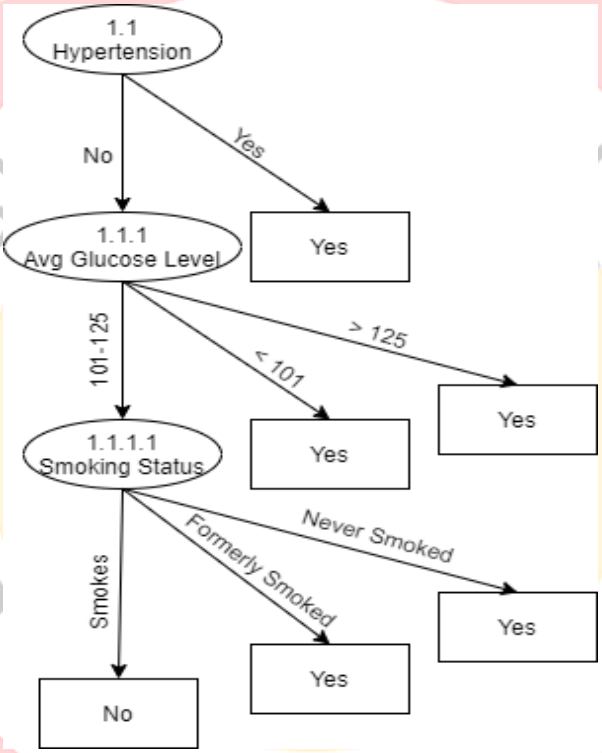
c) Node 1.1.1.1

Tabel 3.10 Perhitungan nilai variabel lebih dari 65 tahun

No	Atribut		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Avg Glucose Level (101 - 125)		16	9	7	0.9887	
1	*Gender*						0.1381
		Male	6	5	1	0.6500	
		Female	10	4	6	0.9710	
2	*Heart Disease*						0.0075
		Yes	3	2	1	0.9183	
		No	13	7	6	0.9957	
3	*Smoking*						0.2081
	Status						
		Formerly Smoked	6	3	3	1	
		Never Smoked	8	6	2	0.8113	
		Smokes	2	0	2	0	
4	*Resident Type*						0.0038
		Urban	12	7	5	0.9799	
		Rural	4	2	2	1	
5	BMI						0.1853

		Kurang 18.6	1	0	1	0	
		18.6 - 24.9	5	3	2	0.9710	
		25 - 29.9	2	2	0	0	
		Lebih 29.9	8	4	4	1	

Dari tabel tersebut, kita melihat bahwa '*Smoking Status*' memiliki nilai *gain* tertinggi (0.1875). Oleh karena itu, '*Smoking Status*' akan menjadi cabang pertama (1.1.1.1) dalam pohon keputusan kita. Variabel '*Smoking Status*' memiliki tiga kategori: '*Formerly Smoked*', '*Never Smoked*', dan '*Smokes*'.



Gambar 3.7 *Node* 1.1.1.1

2. Perhitungan nilai variabel age 56 - 65 tahun

a) Node 1.2

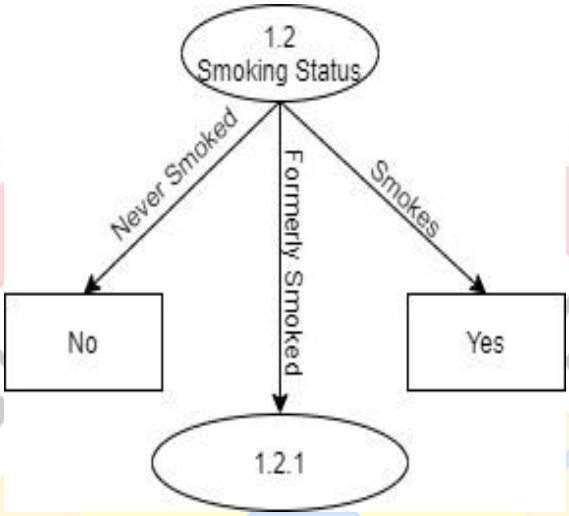
Tabel 3.11 Perhitungan nilai variabel 56-65 tahun node 1.2

No	Atribut		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Age 56-65		72	33	39	0.9949	
1	*Gender*						0.02764

		Male	36	20	16	0.9910	
		Female	36	13	23	0.9436	
2	*Hypertension*						0.01329
		Yes	20	7	13	0.9340	
		No	52	26	26	1	
3	*Heart Disease*						0.03859
		Yes	11	8	3	0.8453	
		No	61	25	36	0.9764	
4	*Smoking Status*						0.17989
		Formerly Smoked	26	14	12	0.9957	
		Never Smoked	30	6	24	0.7219	
		Smokes	16	13	3	0.6962	
5	*Resident Type*						0.01316
		Urban	34	18	16	0.9975	
		Rural	38	15	23	0.9677	
6	*Avg Glucose Level*						0.07461
		Kurang dari 101	32	9	23	0.8571	
		101 - 125	15	9	6	0.9709	
		Lebih dari 125	25	15	10	0.9709	
7	BMI						0.02772
		Kurang dari 18.6	1	0	1	0	
		18.6 - 24.9	7	2	5	0.8631	
		25 - 29.9	17	7	10	0.9774	
		Lebih dari 29.9	47	24	23	0.9996	

Dari tabel tersebut, kita melihat bahwa '*Smoking Status*' memiliki nilai *gain* tertinggi (0.179899), menjadikannya cabang pertama dalam pohon keputusan kita. '*Smoking Status*' memiliki tiga nilai: '*Formerly Smoked*', '*Never*

Smoked', dan '*Smokes*'. Kita dapat langsung mengklasifikasikan '*Never Smoked*' sebagai '*No*' dan '*Smokes*' sebagai '*Yes*'. Namun, kita perlu analisis lebih lanjut untuk '*Formerly Smoked*'. Berdasarkan perhitungan ini, berikut adalah gambaran awal pohon keputusan kita.



Gambar 3.8 *Node* 1.2

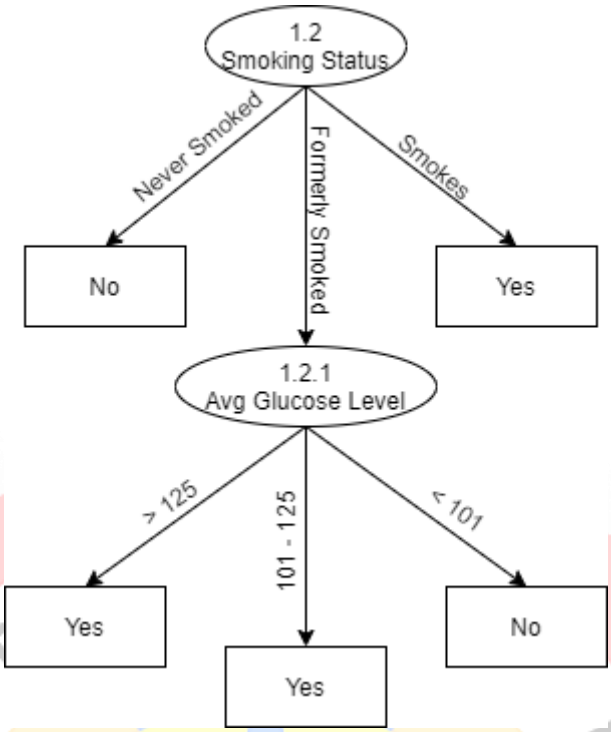
b) Node 1.2.1

Tabel 3.12 Perhitungan nilai variabel 56-65 tahun node 1.2.1

No	Atribut		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Formerly Smoked		26	14	12	0.9957	
1	*Gender*						0.0172
		Male	13	8	5	0.9612	
		Female	13	6	7	0.9957	
2	*Hypertension*						0.1789
		Yes	7	1	6	0.5916	
		No	19	13	6	0.8997	
3	*Heart Disease*						0.0063
		Yes	3	2	1	0.9183	
		No	23	12	11	0.9986	
5	*Resident Type*						0.0172

		Urban	13	8	5	0.9612	
		Rural	13	6	7	0.9957	
6	*Avg Glucose Level*						0.5792
		Kurang dari 101	15	3	12	0.7219	
		101 - 125	4	4	0	0	
		Lebih dari 125	7	7	0	0	
7	BMI						0.0245
		18.6 - 24.9	1	0	5	1	
		25 - 29.9	7	3	4	0.9852	
		Lebih dari 29.9	18	11	7	0.9640	

Dari Tabel 3.12, variabel '*Avg Glucose Level*' memiliki nilai gain tertinggi (0.5792), sehingga menjadi cabang pertama (1.2.1) dalam pohon keputusan. Variabel '*Smoking Status*' memiliki 3 nilai: 'Kurang dari 101', '101 - 125', dan 'Lebih dari 125'. Nilai 'Kurang dari 101' menghasilkan keputusan '*No*', sementara nilai '101 - 125' dan 'Lebih dari 125' menghasilkan keputusan '*Yes*'. Dengan informasi ini, pohon keputusan awal dapat dikonstruksi tanpa perhitungan tambahan.



Gambar 3.9 *Node* 1.2.1

3. Perhitungan nilai variabel age 46-55 tahun

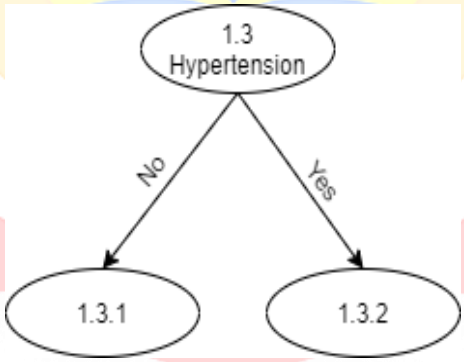
a) Node 1.3

Tabel 3.13 Perhitungan nilai variabel 46-55 tahun node 1.3

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	46-55		54	24	30	0.9910	
1	*Gender*						0.0154
		Male	20	7	13	0.9340	
		Female	34	17	17	1	
2	*Hypertension*						0.0666
		Yes	9	7	2	0.7642	
		No	45	17	28	0.9564	
3	*Heart Disease*						0.0003
		Yes	2	1	1	1	
		No	52	23	29	0.9903	
4	*Smoking Status*						0.0122
		Formerly Smoked	12	4	8	0.9183	

		Never Smoked	24	12	12	1	
		Smokes	18	8	10	0.9910	
5	*ResidentType*						0.0002
		Urban	31	14	17	0.9932	
		Rural	23	10	13	0.9876	
6	*Avg Glucose Level*						0.0026
		Kurang dari 101	30	13	17	0.9871	
		101 - 125	12	5	7	0.9798	
		Lebih dari125	12	6	6	1	
7	BMI						0.0589
		18.6 - 24.9	11	2	9	0.6840	
		25 - 29.9	18	10	8	0.9910	
		Lebih dari29.9	25	13	12	0.9988	

Dari tabel tersebut, variabel "*Hypertension*" memiliki nilai *gain* tertinggi (0.0666). Oleh karena itu, "*Hypertension*" akan menjadi cabang 1.3 dalam pohon keputusan. Variabel "*Smoking Status*" memiliki dua nilai, yaitu "*Yes*" dan "*No*". Berikut adalah gambaran sementara pohon keputusan berdasarkan perhitungan tersebut.



Gambar 3.10 *Node* 1.3

b) Node 1.3.1

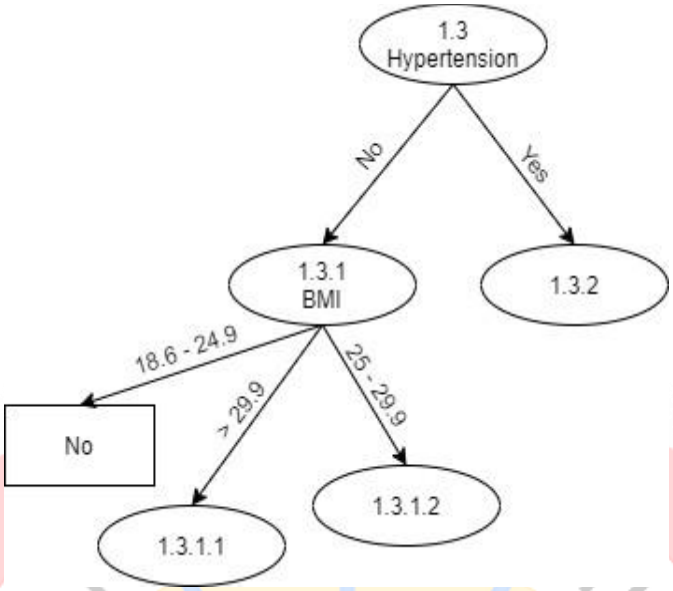
Tabel 3.14 Perhitungan nilai variabel 46-55 tahun node 1.3.1

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	HP - No		45	17	28	0.9565	

1	*Gender*						0.0012
		Male	17	6	11	0.9367	
		Female	28	11	17	0.9666	
2	*Heart Disease*						0.0154
		Yes	1	0	1	0.0000	
		No	44	17	27	0.9624	
3	*Smoking Status*						0.0198
		Formerly Smoked	9	2	7	0.7642	
		Never Smoked	19	8	11	0.9819	
		Smokes	17	7	10	0.9774	
4	*Resident Type*						0.0042
		Urban	26	9	17	0.9306	
		Rural	19	8	11	0.9819	
5	*Avg Glucose Level*						0.0050
		Kurang dari 101	28	11	17	0.9666	
		101 - 125	10	4	6	0.9710	
		Lebih dari 125	7	2	5	0.8631	
6	BMI						0.0970
		18.6 - 24.9	11	1	10	0.4395	
		25 - 29.9	16	8	8	1.0000	
		Lebih dari 29.9	18	8	10	0.9911	

Berdasarkan analisis pada tabel, diketahui bahwa BMI memiliki nilai *gain* tertinggi (0.0970), sehingga akan menjadi cabang 1.3.1 dalam pohon keputusan. Variabel BMI memiliki tiga nilai kategori: 18.6 - 24.9, 25 - 29.9, dan Lebih dari 29.9. Jika nilai BMI berada pada rentang 18.6 - 24.9, maka keputusan langsung diklasifikasikan sebagai "*No*" dan tidak memerlukan perhitungan lebih lanjut. Namun, untuk nilai BMI pada rentang 25 - 29.9 dan Lebih dari 29.9, diperlukan analisis lebih lanjut. Berdasarkan hasil

perhitungan ini, dapat digambarkan sebuah pohon keputusan sementara.



Gambar 3.11 *Node* 1.3.1

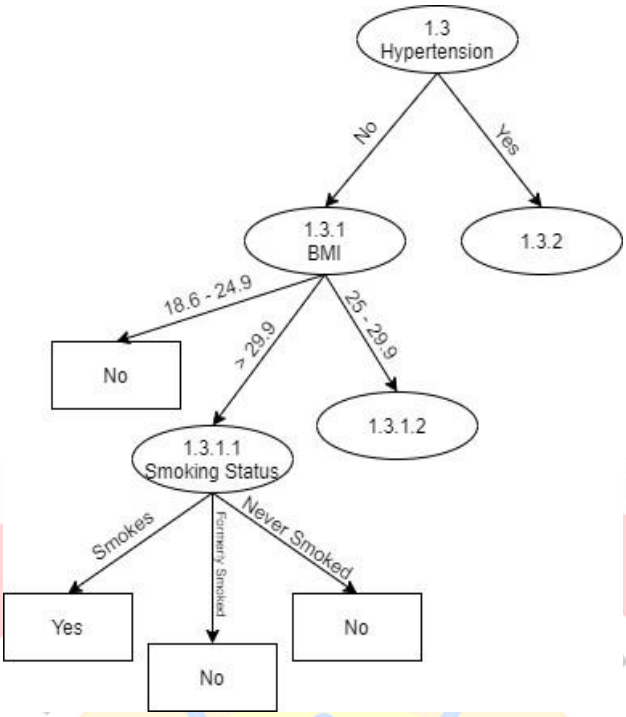
c) Node 1.3.1.1

Tabel 3.15 Perhitungan nilai variabel 46-55 tahun node 1.3.1.1

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	BMI Lebih dari 29.9		18	8	10	0.9911	
1	*Gender*						0.0045
		Male	6	3	3	1.0000	
		Female	12	5	7	0.9799	
3	*Heart Disease*						0.0490
		Yes	1	0	1	0.0000	
		No	17	8	9	0.9975	
4	*Smoking Status*						0.1488
		Formerly Smoked	5	1	4	0.7219	
		Never Smoked	6	2	4	0.9183	
		Smokes	7	5	2	0.8631	

5	*Resident Type*						0.0072
		Urban	10	4	6	0.9710	
		Rural	8	4	4	1.0000	
6	*Avg Glucos eLevel*						0.0364
		Kurangdari 101	9	4	5	0.9911	
		101 - 125	3	2	1	0.9183	
		Lebih dari 125	6	2	4	0.9183	

Dari tabel tersebut, kita dapat melihat bahwa "*Smoking Status*" memiliki nilai *gain* tertinggi (0.1488). Oleh karena itu, "*Smoking Status*" menjadi cabang pertama dalam pohon keputusan kita (1.3.1.1). Variabel ini memiliki tiga nilai: "*Formerly Smoked*", "*Never Smoked*", dan "*Smokes*". Masing-masing nilai ini langsung mengarah pada keputusan akhir: "*Formerly Smoked*" dan "*Never Smoked*" mengarah ke keputusan "*No*", sedangkan "*Smokes*" mengarah ke keputusan "*Yes*". Karena setiap nilai sudah memiliki keputusan akhir, tidak diperlukan perhitungan lebih lanjut. Dengan informasi ini, kita dapat menggambarkan pohon keputusan seperti berikut:



Gambar 3.12 *Node* 1.3.1.1

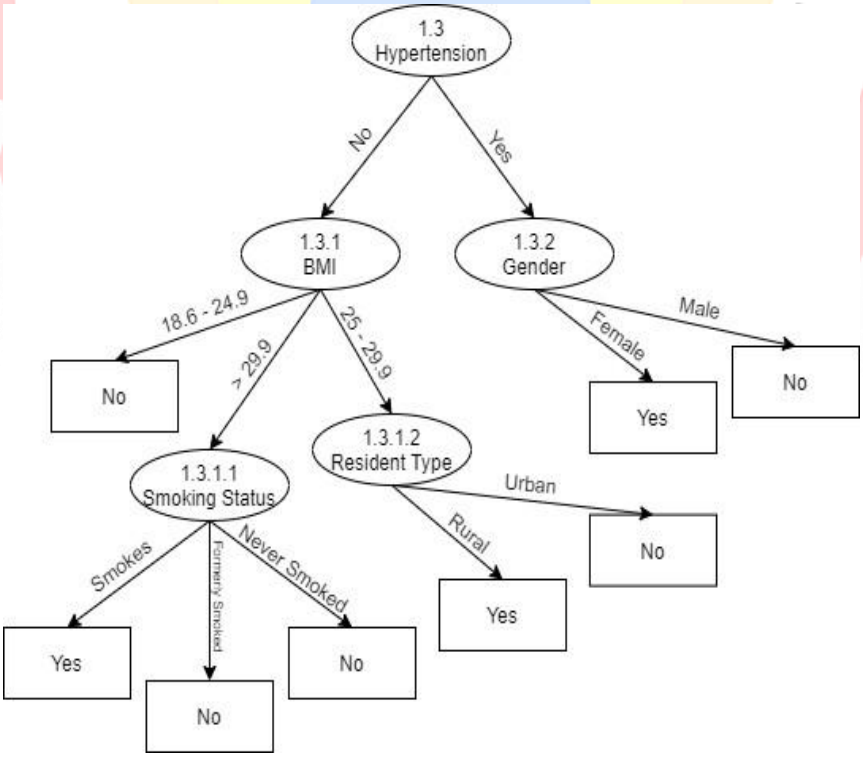
d) Node 1.3.1.2

Tabel 3.16 Perhitungan nilai variabel 46-55 tahun node 1.3.1.2

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	BMI 25-29.9		16	8	8	1	
1	*Gender*						0
		Male	6	3	3	1	
		Female	10	5	5	1	
3	*Heart Disease*						1
		No	16	16	0	0	
4	*Smoking Status*						0.0716
		Formerly Smoked	1	1	0	0	
		Never Smoked	10	5	5	1	
		Smokes	5	2	3	0.9710	
5	*Resident Type*						1
		Urban	8	4	3	1	
		Rural	8	4	5	1	

6	*Avg Glucose Level*						0.0666
		Kurangdari 101	11	6	5	0.9940	
		101 - 125	4	2	2	1	
		Lebih dari 125	1	0	1	0	

Dari data pada tabel, terlihat bahwa '*Heart Disease*' dan '*Residence Type*' memiliki nilai *gain* tertinggi. Karena ada dua variabel dengan nilai *gain* yang sama, perlu dilakukan analisis lebih lanjut. Hasilnya menunjukkan '*Residence Type*' memiliki nilai gain yang sedikit lebih tinggi, sehingga variabel ini akan menjadi cabang 1.3.1.2 dalam pohon keputusan. '*Residence Type*' memiliki dua nilai, yaitu '*Urban*' dan '*Rural*'. Nilai '*Urban*' mengarah pada keputusan '*No*', sedangkan '*Rural*' mengarah pada keputusan '*Yes*'. Dengan demikian, tidak diperlukan perhitungan lebih lanjut. Berdasarkan hasil analisis ini, pohon keputusan sementara dapat digambarkan.



Gambar 3.13 *Node* 1.3.1.2

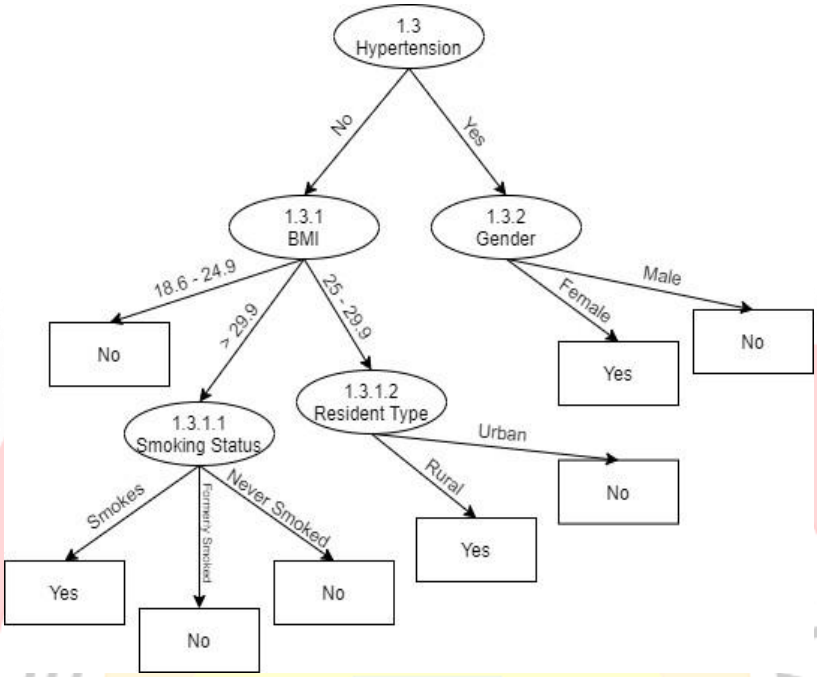
e) Node 1.3.2

Tabel 3.17 Perhitungan nilai variabel 46-55 tahun node 1.3.2

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	HP - Yes		9	7	2	0.7642	
1	*Gender*						0.4581
		Male	3	1	2	0.9183	
		Female	6	6	0	0.0000	
2	*Heart Disease*						0.0431
		Yes	1	1	0	0.0000	
		No	8	6	2	0.8113	
3	*Smoking Status*						0.0570
		Formerly Smoked	3	2	1	0.9183	
		Never Smoked	5	4	1	0.7219	
		Smokes	1	1	0	0.0000	
4	*Resident Type*						0.3198
		Urban	5	5	0	0.0000	
		Rural	4	2	2	1.0000	
5	*Avg Glucose Level*						0.1409
		Kurang dari 101	2	2	0	0.0000	
		101 - 125	2	1	1	1.0000	
		Lebih dari125	5	4	1	0.7219	
6	BMI						0.0929
		25 - 29.9	2	2	0	0.0000	
		Lebih dari29.9	7	5	2	0.8631	

Dari Tabel 3.17, kita melihat bahwa variabel '*Gender*' memiliki nilai *gain* tertinggi (0.4581), sehingga akan menjadi cabang pertama (1.3.2) dalam pohon keputusan kita. Karena variabel '*Gender*' hanya memiliki dua nilai ('*Male*' dan '*Female*'), dan setiap nilai tersebut langsung mengarah pada keputusan ('*No*' untuk '*Male*', '*Yes*' untuk '*Female*'), maka tidak diperlukan

perhitungan lebih lanjut pada cabang ini. Berdasarkan informasi ini, kita dapat membuat gambaran awal pohon keputusan:



Gambar 3.14 *Node* 1.3.2

4. Perhitungan nilai variabel *age* 36-45 tahun

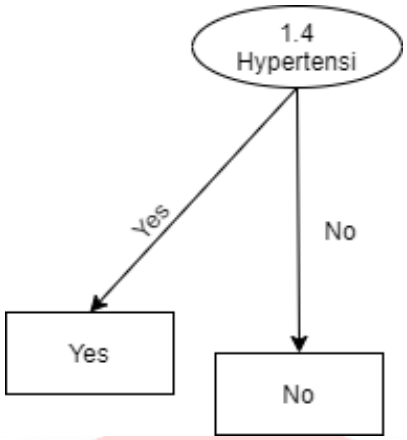
a) Node 1.4

Tabel 3.18 Perhitungan nilai variabel 36-45 tahun node 1.4

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Age 36 - 45 Tahun		35	7	28	0.7219	
1	*Gender*						0.0359
		Male	12	1	11	0.4138	
		Female	23	6	17	0.8281	
2	*Hypertension*						0.0198
		Yes	2	1	1	1	
		No	33	6	27	0.6840	
3	*Heart Disease*						0
		Yes	30	6	24	0.7219	
		No	5	1	4	0.7219	

4	*Smoking Status*						0.0008
		Formerly Smoked	11	2	9	0.6840	
		Never Smoked	14	3	11	0.7496	
		Smokes	10	2	8	0.7219	
5	*Resident Type*						0.0149
		Urban	15	4	11	0.7219	
		Rural	20	3	17	0.7219	
6	*Avg Glucose Level*						0.00397
		Kurang dari 101	27	6	21	0.7642	
		101 - 125	4	0	4	0.0000	
		Lebih dari 125	4	1	3	0.8113	
7	BMI						0.0101
		18.6 - 24.9	8	1	7	0.5436	
		25 - 29.9	15	3	12	0.7219	
		Lebih dari	12	3	9	0.8113	

Dari tabel tersebut, kita dapat melihat bahwa variabel "*Hypertension*" memiliki nilai *gain* tertinggi, yaitu 0.0198. Oleh karena itu, "*Hypertension*" akan menjadi cabang 1.5 dalam pohon keputusan kita. Variabel "*Hypertension*" memiliki dua nilai, yaitu "*Yes*" dan "*No*". Jika nilai variabelnya adalah "*Yes*", maka keputusan akhir adalah "*Yes*". Sebaliknya, jika nilai variabelnya adalah "*No*", maka keputusan akhir adalah "*No*". Dengan demikian, tidak diperlukan perhitungan lebih lanjut untuk cabang ini. Berdasarkan analisis ini, kita dapat menggambarkan pohon keputusan sementara seperti berikut:



Gambar 3.15 *Node* 1.4

5. Perhitungan nilai variabel age 26-35 tahun

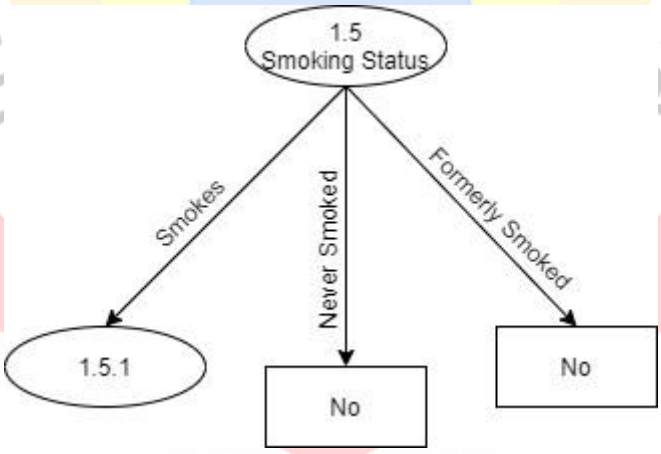
a) Node 1.5

Tabel 3.19 Perhitungan nilai variabel 26-35 tahun node 1.5

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	Age 26 - 35 Tahun		16	1	15	0.3373	
1	*Gender*						0.0269
		Male	4	0	4	0.0000	
		Female	12	1	11	0.4138	
2	*Hypertension*						0.0060
		Yes	1	0	1	0.0000	
		No	15	1	14	0.3534	
3	*Heart Disease*						0.0000
		Yes	0	0	0	0.0000	
		No	16	1	15	0.3373	
4	*Smoking Status*						0.1345
		Formerly Smoked	2	0	2	0.0000	
		Never Smoked	10	0	10	0.0000	
		Smokes	4	1	3	0.8113	
5	*Resident Type*						0.0194
		Urban	3	0	3	0.0000	
		Rural	13	1	12	0.3912	
6	*Avg Glucose Level*						0.0351

		Kurang dari 101	11	1	10	0.4395	
		101 - 125	2	0	2	0.0000	
		Lebih dari 125	3	0	3	0.0000	
7	BMI						0.1117
		Kurang dari 18.6	1	0	1	0.0000	
		18.6 - 24.9	1	0	1	0.0000	
		25 - 29.9	5	1	4	0.7219	
		Lebih dari 29.9	9	0	9	0.0000	

Dari tabel di atas, kita melihat bahwa '*Smoking Status*' memiliki nilai *gain* tertinggi (0.1345), menjadikannya cabang pertama dalam pohon keputusan kita. '*Smoking Status*' memiliki tiga nilai: '*Formerly Smoked*', '*Never Smoked*', dan '*Smokes*'. Baik '*Formerly Smoked*' maupun '*Never Smoked*' mengarah pada keputusan '*No*', jadi tidak perlu analisis lebih lanjut. Namun, nilai '*Smokes*' masih memerlukan perhitungan tambahan. Berdasarkan perhitungan ini, kita dapat membuat gambaran awal pohon keputusan seperti di bawah ini.



Gambar 3.16 *Node* 1.5

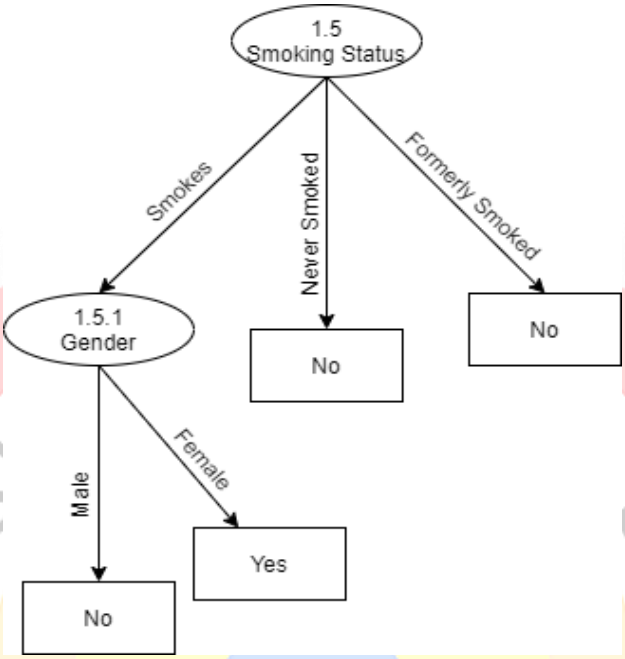
b) Node 1.5.1

Tabel 3.20 Perhitungan nilai variabel 26-35 tahun node 1.5.1

No	Variabel		Jumlah (S)	Ya (Si)	Tidak (Si)	*Entropy*	*Gain*
	ST - *Smokes*		4	1	3	0.8113	
1	*Gender*						0.3113
		Male	2	0	2	0.0000	
		Female	2	1	1	1.0000	
2	*Hypertension*						0.0000
		No	4	1	3	0.8113	
3	*Heart Disease*						0.0000
		No	4	1	3	0.8113	
5	*Resident Type*						0.0000
		Urban	0	0	0	0.0000	
		Rural	4	1	3	0.8113	
6	*Avg Glucose Level*						0.3113
		Kurang dari 101	2	1	1	1.0000	
		101 - 125	1	0	1	0.0000	
		Lebih dari 125	1	0	1	0.0000	
7	BMI						0.3113
		25 - 29.9	2	1	1	1.0000	
		Lebih dari 29.9	2	0	2	0.0000	

Dari tabel tersebut, kita melihat bahwa "*Gender*" adalah variabel paling berpengaruh dengan nilai *gain* tertinggi (0.3113). Oleh karena itu, "*Gender*" akan menjadi cabang pertama (1.5.1) dalam pohon keputusan kita. Variabel "*Gender*" memiliki dua nilai: "*Male*" dan "*Female*". Karena semua kasus dengan nilai "*Male*" mengarah ke keputusan "*No*" dan semua kasus dengan

nilai "*Female*" mengarah ke keputusan "*Yes*", kita tidak perlu melakukan perhitungan lebih lanjut untuk cabang ini. Dengan informasi ini, kita dapat membuat gambaran awal pohon keputusan.



Gambar 3.17 *Node* 1.5.1

Berdasarkan pohon keputusan yang telah dibangun secara lengkap, seluruh kasus telah berhasil diklasifikasikan. Proses ini menghasilkan 28 aturan yang memprediksi kejadian *stroke* dengan kelas "*yes*" atau "*no*".

Tabel 3.21 *Rule Tree*

Rule	**Keterangan *Rule***	**Prediksi**
1.	Jika pasien berumur 6 -11 Tahun	Tidak *Stroke*
2.	Jika pasien berumur 12 – 16 Tahun	Tidak *Stroke*
3.	Jika pasien berumur 17 – 25 Tahun	Tidak *Stroke*
4.	Jika pasien berumur 26 – 35 Tahun, pernah merokok	Tidak *Stroke*
5.	Jika pasien berumur 26 – 35 Tahun, tidak pernah merokok	Tidak *Stroke*
6.	Jika pasien berumur 26 – 35 Tahun, perokok aktif, jenis kelamin perempuan	*Stroke*
7.	Jika pasien berumur 26 – 35 Tahun, perokok, kelamin laki-laki	Tidak *Stroke*
8.	Jika pasien berumur 36 – 45 Tahun, tekanan darah normal	Tidak *Stroke*

9.	Jika pasien berumur 36 – 45 Tahun, tekanan darah tinggi	*Stroke*
10.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan normal	Tidak *Stroke*
11.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan berlebihan, dan tinggal di pedesaan	*Stroke*
12.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan berlebihan, dan tinggal di perkotaan	Tidak *Stroke*
13.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan obesitas, pernah merokok	Tidak *Stroke*
14.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan obesitas, tidak pernah merokok	Tidak *Stroke*
15.	Jika pasien berumur 46 – 55 Tahun, tekanan darah normal, berat badan obesitas, perokok aktif	*Stroke*
16.	Jika pasien berumur 46 – 55 Tahun, tekanan darah tinggi, jenis kelamin perempuan	*Stroke*
17.	Jika pasien berumur 46 – 55 Tahun, tekanan darah tinggi, jenis kelamin laki-laki	Tidak *Stroke*
18.	Jika pasien berumur 56 – 65 Tahun, pernan merokok, guna darah pradiabetes	*Stroke*
19.	Jika pasien berumur 56 – 65 Tahun, pernan merokok, guna darah normal	Tidak *Stroke*
20.	Jika pasien berumur 56 – 65 Tahun, pernan merokok, guna darah diabetes	*Stroke*
21.	Jika pasien berumur 56 – 65 Tahun, tidak pernah merokok	Tidak *Stroke*
22.	Jika pasien berumur 56 – 65 Tahun, perokok aktif	*Stroke*
23.	Jika pasien berumur di atas 65 Tahun, tekanan darah normal, gula darah paradiabetes dan pernah merokok	*Stroke*
24.	Jika pasien berumur di atas 65 Tahun, tekanan darah normal, gula darah pradiabetes dan tidak pernah merokok	*Stroke*
25.	Jika pasien berumur di atas 65 Tahun, tekanan darah normal, gula darah pradiabetes dan perokok aktif	Tidak *Stroke*
26.	Jika pasien berumur di atas 65 Tahun, tekanandarah normal, gula darah normal	*Stroke*
27.	Jika pasien berumur di atas 65 Tahun, tekanandarah normal, gula darah diabetes	*Stroke*
28.	Jika pasien berumur di atas 65 Tahun, tekanandarah tinggi	*Stroke*

Berdasarkan analisis di atas, faktor-faktor yang paling berperan dalam memprediksi *stroke* adalah: *age* pada tahap awal, kemudian *hypertension*, diikuti oleh *avg glucose level*, dan terakhir *smoking status*.

Model prediksi menunjukkan bahwa risiko *stroke* tertinggi adalah pada pasien berusia di atas 65 tahun dengan tekanan darah tinggi, dengan total 42 atribut risiko. Sebaliknya, kemungkinan terendah untuk terkena *stroke* adalah pada pasien berusia 36-45 tahun dengan tekanan darah normal, dengan total 27 atribut risiko.

Untuk melihat lebih detail aturan prediksi dengan rasio *gain* untuk *stroke*, Anda dapat merujuk pada Tabel 3.22, di mana "*Stroke* = *Yes*" menunjukkan adanya *stroke* dan "Tidak *Stroke* = *No*" menunjukkan tidak adanya *stroke*.

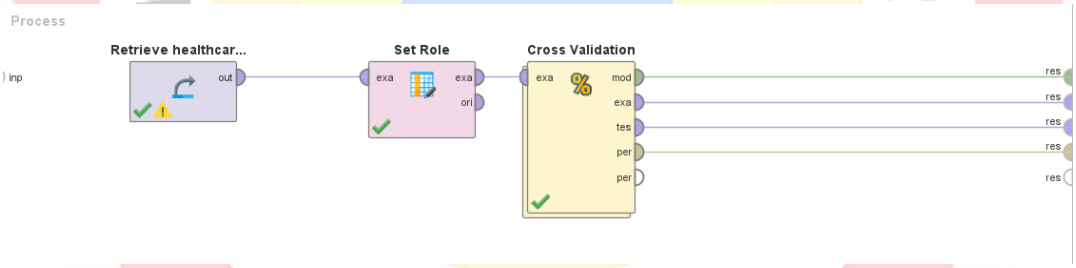
Tabel 3.22 Keterangan *Rule Text* dengan *Gain Ratio*

Rule	**Keterangan *Rule***	**Predikat *Gain Rasio***	
		Yes	**No**
1	Jika pasien berumur 6 -11 Tahunmaka Tidak *Stroke*	0	2
2	Jika pasien berumur 12 – 16 Tahun maka Tidak *Stroke*	0	2
3	Jika pasien berumur 17 – 25 Tahun maka Tidak *Stroke*	0	20
4	Jika pasien berumur 26 – 35 Tahun,pernah merokok maka Tidak *Stroke*	0	2
5	Jika pasien berumur 26 – 35 Tahun,tidak pernah merokok maka Tidak *Stroke*	0	10
6	Jika pasien berumur 26 – 35 Tahun,perokok aktif, jenis kelamin perempuan maka *Stroke*	1	1
7	Jika pasien berumur 26 – 35 Tahun,perokok, jenis kelamin laki-laki maka Tidak *Stroke*	0	2
8	Jika pasien berumur 36 – 45 Tahun,tekanan darah normal maka Tidak *Stroke*	6	27
9	Jika pasien berumur 36 – 45 Tahun,tekanan darah tinggi maka *Stroke*	1	1
10	Jika pasien berumur 46 – 55 Tahun,tekanan darah normal, berat badan normal maka Tidak *Stroke*	1	10
11	Jika pasien berumur 46 – 55 Tahun,tekanan darah normal, berat badan berlebihan, dan tinggal di pedesaanmaka *Stroke*	4	3

3.5 Pembahasan Metode dan Algoritma

Algoritma C4.5, juga dikenal sebagai *decision tree* (pohon keputusan), merupakan metode yang sudah dikenal luas dan banyak dimanfaatkan untuk mendukung pengambilan keputusan berbasis data. Metode ini termasuk dalam kategori teknik klasifikasi dan algoritma tersebut juga bekerja dengan cara mengevaluasi setiap atribut data yang berdasarkan nilai *gain* dan *entropy*. Simpul pada pohon keputusan akan ditentukan berdasarkan atribut yang memiliki nilai gain paling tinggi. Kumpulan dari simpul-simpul ini akan membentuk struktur pohon yang mudah dipahami, sehingga pengguna dapat dengan jelas melihat alur klasifikasi yang dihasilkan.

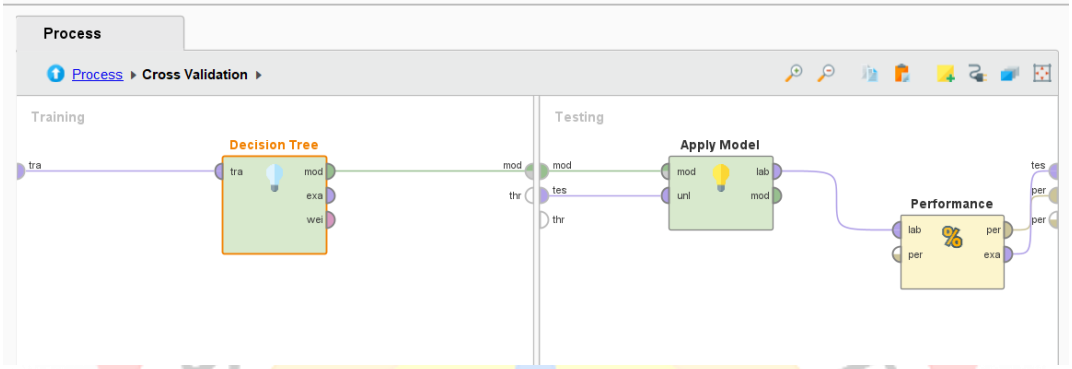
Berikut ini adalah operator yang digunakan dalam aplikasi *RapidMiner*. Pengujian dengan *RapidMiner* tersebut menggunakan 5110 data pasien yang didapatkan dari situs web *Kaggle* dengan 248 pasien positif dan 4.862 pasien positif.



Gambar 3.18 Perancangan Operator di *RapidMiner*

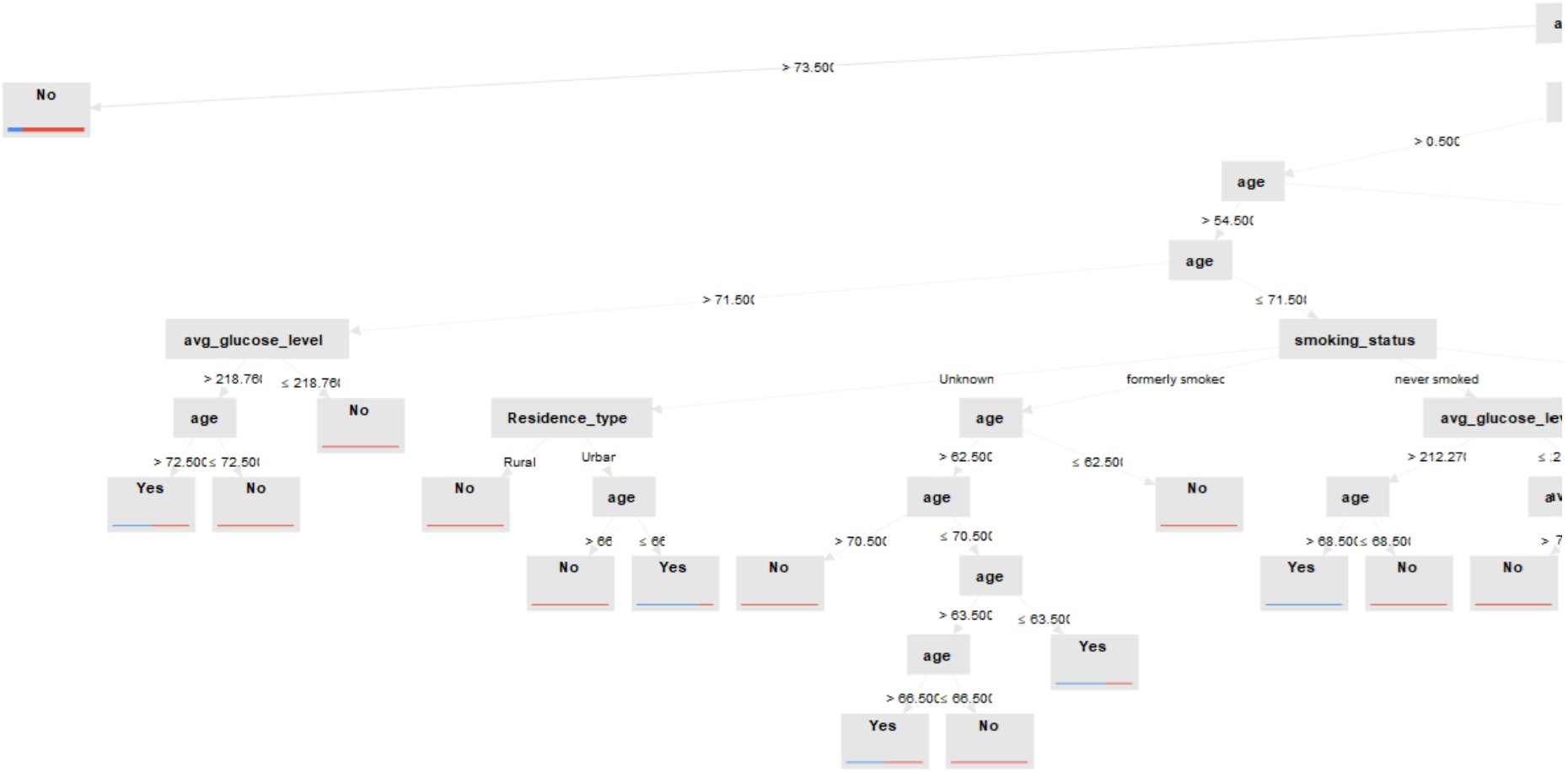
Gambar diatas merupakan tahap awal dalam menguji data menggunakan aplikasi *RapidMiner*. Pada tahap awal masukan 3 operator yaitu, operator *retrieve dataset*, *set role* dan *cross validation*. *Retrieve dataset* berguna untuk melakukan impor *dataset* yang sudah dimasukkan ke repositori lokal. *Set role* berfungsi untuk membedakan baris yang berisi nama atribut koordinat dengan baris yang berisi prediksi posisi. Baris-baris ini akan dimasukkan ke dalam kategori 'label'. Dengan demikian, saat data dikategorikan, baris-baris dalam kategori 'label' tidak akan ikut terhitung dan mempengaruhi hasil akhir. Sementara itu, *cross*

validation merupakan teknik yang membagi *dataset* menjadi beberapa bagian, dimana setiap bagian digunakan secara bergantian sebagai data pelatihan untuk membangun model dan data pengujian untuk mengevaluasi performanya. Tujuannya adalah untuk mendapatkan estimasi performa model yang lebih akurat dan menghindari *overfitting*.

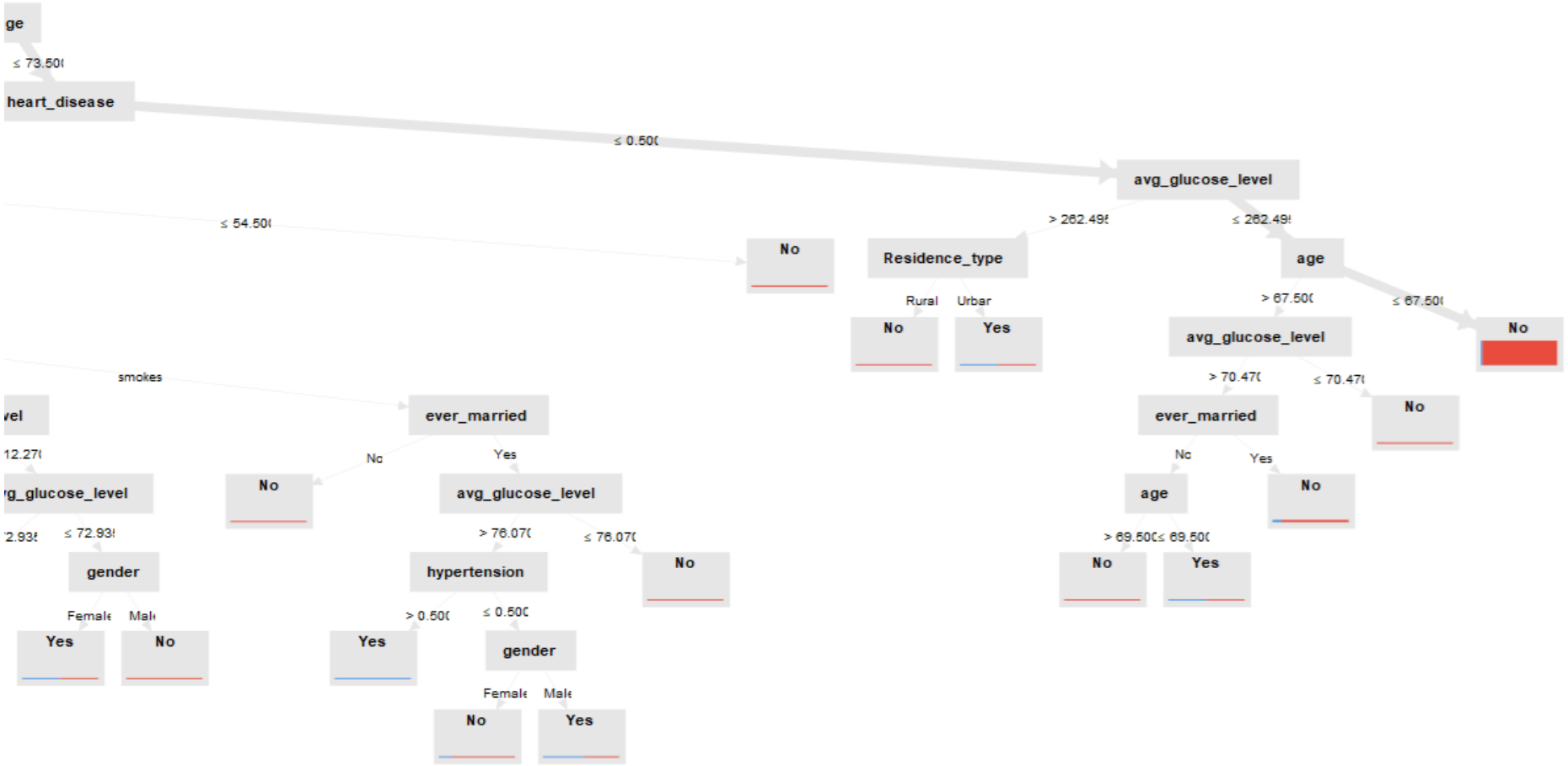


Gambar 3.19 Perancangan Operator di *RapidMiner* (*Cross Validation*)

Pada tahap berikutnya adalah dimasukkannya beberapa operator di dalam operator *cross validation* dengan cara mengeklik ganda pada operator cross validation, di dalam operator *cross validation* terdapat halaman kosong dan akan dimasukkan 3 buah operator, yaitu *decision tree*, *apply model*, dan *performance*. Algoritma *decision tree* tersebut merupakan algoritma yang digunakan untuk mengklasifikasikan seseorang yang menderita penyakit *stroke*. Dalam proses pengambilan keputusan menggunakan *decision tree*, *gain_ratio* pada kolom kriteria menjadi pertimbangan utama untuk mencapai hasil yang diinginkan. Syarat *minimal confidence* adalah 0.1, syarat *minimal gain* adalah 0.01, dan jumlah minimum data dalam setiap *leaf* adalah 2. *Apply model* berfungsi untuk menampilkan representasi awal dari data yang sedang kita olah, sedangkan *performance* digunakan untuk mengevaluasi seberapa baik suatu algoritma bekerja ketika diterapkan pada data uji. Pohon keputusan yang dihasilkan menggunakan *RapidMiner* dapat dilihat di bawah ini.



Gambar 3.20 Pohon Keputusan Bagian 1



Gambar 3.21 Pohon Keputusan Bagian 2

3.6 *Requirement Elicitation*

a. Elisitasi Tahap I

Elisitasi Tahap I dilakukan dengan cara menyebarkan RE kepada responden yang sesuai kriteria sebagai pengguna aplikasi. RE ini berisi pertanyaan tentang seluruh rancangan sistem yang dibuat. Data untuk RE ini diperoleh melalui observasi kebutuhan pengguna terhadap sistem yang saat ini belum terpenuhi.

Berikut adalah tabel yang merangkum hasil dari tahap pertama proses elisitasi kebutuhan. Tabel ini disusun berdasarkan masukan yang kami terima dari pengguna selama sesi *Requirement Elicitation*.

Tabel 3.23 *Requirement Elicitation* Tahap I

No	Kebutuhan Pemakai
1	Mempunyai tampilan yang terlihat menarik
2	Antarmuka pengguna yang gampang digunakan dan dimengerti
3	Dapat memprediksi penderita penyakit *stroke*
4	Dapat melihat hasil yang akurat dari penyakit *stroke*
5	Dapat melihat hasil diagnosa penderita penyakit *stroke*
6	Menampilkan informasi tentang penyakit *stroke*, bahaya, dan cara mencegahnya
7	Penjelasan tentang hasil diagnosa penderita penyakit *stroke*
8	Ada fitur yang mengharuskan bantuan

Dari 45 jawaban yang dikumpulkan melalui 10 responden, maka diringkas menjadi 8 jawaban yang menjadi RE di atas, 37 jawaban sisanya dibuang karena ada yang sama dan duplikat.

b. Elisitasi Tahap II

Tahap kedua elisitasi dibangun berdasarkan hasil tahap pertama, kemudian kebutuhan-kebutuhan tersebut diklasifikasikan ulang untuk diproses sesuai dengan metodologi MDI. Dalam tahap ini, terdapat kebutuhan yang bersifat opsional (I) dan kebutuhan yang wajib dieliminasi. Berikut adalah tabel dari elisitasi tahap II.

Tabel 3.24 *Requirement Elicitation* Tahap II

No	Kebutuhan Pemakai	M	D	I
1	Mempunyai tampilan yang terlihat menarik	*		
2	Antarmuka pengguna yang gampang digunakan dan dimengerti	*		
3	Dapat memprediksi penderita penyakit *stroke*		*	
4	Dapat melihat hasil yang akurat dari penyakit *stroke*		*	
5	Dapat melihat hasil diagnosa penderita penyakit *stroke*	*		
6	Menampilkan informasi tentang penyakit *stroke*, bahaya, dan cara mencegahnya	*		
7	Penjelasan tentang hasil diagnosa penderita penyakit *stroke*			*
8	Ada fitur yang mengharuskan bantuan			*

Keterangan:

M = *Mandatory* (Wajib)

D = *Desirable* (Diinginkan)

I = *Inessential* (Kurang Penting)

c. Elisitasi Tahap III

Elisitasi tahap III merupakan hasil pengembangan dari elisitasi tahap II. Pada tahap ini, kebutuhan-kebutuhan sistem yang telah dikumpulkan pada tahap sebelumnya diklasifikasikan ulang menggunakan metode TOE (*Technical*, *Operational*, *Economic*) dengan opsi penilaian HML (*High*, *Middle*, *Low*). Berikut adalah tabel elisitasi tahap III yang disusun berdasarkan tabel elisitasi tahap II.

Tabel 3.25 *Requirement Elicitation* Tahap III

Feasibility		**T**			**O**			**E**		
Risk		**H**	**M**	**L**	**H**	**M**	**L**	**H**	**M**	**L**
No	**Kebutuhan Pemakai**									
1	Mempunyai tampilan yang terlihat menarik			*		*				*
2	Antarmuka pengguna yang gampang digunakan dan dimengerti		*			*			*	
3	Dapat memprediksi penderita penyakit *stroke*		*			*			*	
4	Dapat melihat hasil yang akurat dari penyakit *stroke*	*			*				*	
5	Dapat melihat hasil diagnosa penderita penyakit *stroke*	*				*			*	
6	Menampilkan informasi tentang penyakit *stroke*,		*			*				*

	bahaya, dan cara mencegahnya									

Keterangan:

T = *Technical* (Teknikal)

O = *Operational* (Operasional)

E = *Economic* (Ekonomi)

H = *High* (Sulit Dikerjakan)

M = *Middle* (Mampu untuk Dikerjakan)

L = *Low* (Mudah Dikerjakan)

d. Elisitasi Final

Final elisitasi telah disusun berdasarkan klasifikasi metode TOE pada elisitasi tahap III, setelah melalui tiga tahap elisitasi sebelumnya. Berikut adalah hasilnya.

Tabel 3.26 *Requirement Elicitation* Final

No	Kebutuhan Pemakai
1	Mempunyai tampilan yang terlihat menarik
2	Antarmuka pengguna yang gampang digunakan dan dimengerti
3	Dapat memprediksi penderita penyakit *stroke*
4	Dapat melihat hasil yang akurat dari penyakit *stroke*
5	Dapat melihat hasil diagnosa penderita penyakit *stroke*
6	Menampilkan informasi tentang penyakit *stroke*, bahaya, dan cara mencegahnya

3.7 Jadwal Penelitian

Tabel 3.27 Jadwal Penelitian

No	Kegiatan	Bulan 2024			
		April	Mei	Juni	Juli
1.	Langkah Awal Penelitian				
	a. Perumusan dan Pengusulan Judul				
	b. Penyusunan Proposal				
2.	Langkah Pelaksanaan				
	a. Perolehan Data				
	b. Analisis Data				
3	Periode Penulisan Laporan				