Prosiding Seminar Nasional Sains dan Teknologi Seri IV
Fakultas Sains dan Teknologi, Universitas Terbuka
Vol. 2 No. 2 (2025)
e-ISSN: 3047-6569

PENERAPAN KLASIFIKASI DATA MINING UNTUK PREDIKSI DENGAN METODE ALGORITMA DECISSION TREE

Wisnu Aji Pamungkas

Program Studi Statistika, Universitas Terbuka, Tangerang Selatan, Indonesia

wisnuajipamungkas009@gmail.com

ABSTRAK

Stroke adalah penyakit pembuluh darah ke otak. Definisi menurut WHO, Stroke adalah suatu keadaan dimana ditemukan tanda-tanda klinis yang berkembang cepat berupa defisit neurologicfokal dan global, yang dapat memberat dan berlangsung lama selama 24 jam atau lebih dan atau dapat menyebabkan kematian tanpa adanya penyebab lain yang jelas selain vascular. Alur proses analisis disini menggunakan metode decision tree yakni mengubah bentuk data menjadi model tree, mengubah model tree menjadi rule dan menyederhanakan rule. Data yang diambil dalam penelitian ini adalah data pasien sebuah rumah sakit yang dirahasiakan namanya sebanyak 5110 sampel yang akan digunakan untuk membuat model prediksi Decision Tree. Model yang telah dibuat kemudian akan dihitung tingkat akurasi prediksinya. Dari kumpulan data ini, terdapat 10 parameter penentu indikator penyakit stroke yakni jenis kelamin, umur, hipertensi, riwayat penyakit jantung, status pernikahan, tipe pekerjaan, tipe tempat tinggal, rata-rata kadar glukosa, indeks massa tubuh, dan status merokok. Decision tree merupakan salah satu cara data processing dalam memprediksi masa depan dengan cara membangun klasifikasi atau regresi model dalam bentuk struktur pohon. Hal tersebut dilakukan dengan cara memecah terus ke dalam himpunan bagian yang lebih kecil lalu pada saat itu juga sebuah pohon keputusan secara bertahap dikembangkan. Hasil akhir dari proses tersebut adalah pohon dengan node keputusan dan node daun. Kemudian dianalisis menggunakan machine learning Phyton untuk membuat prediksi pada pasien guna mengetahui faktor-faktor yang paling mempengaruhi terhadap penyakit stroke.

Kata Kunci : Stroke, Klasifikasi, Pohon Keputusan, Prediksi, Hutan Acak

PENDAHULUAN

Stroke adalah penyakit pembuluh darah ke otak. Definisi menurut WHO, Stroke adalah suatu keadaan dimana ditemukan tanda-tanda klinis yang berkembang cepat berupa defisit neurologicfokal dan global, yang dapat memberat dan berlangsung lama selama 24 jam atau lebih dan atau dapat menyebabkan kematian tanpa adanya penyebab lain yang jelas selain vascular. Stroke terjadi apabila pembuluh darah otak mengalami penyumbatan atau pecah. Akibatnya sebagian otak tidak mendapatkan pasokan darah yang membawa oksigen yang diperlukan sehingga mengalami kematian sel/jaringan. Stroke tetap menjadi salah satu penyebab utama morbiditas dan mortalitas di seluruh dunia, menimbulkan tantangan signifikan bagi sistem kesehatan masyarakat dan kesejahteraan individu. Pada saat ini, penyakit stroke semakin menjadi masalah serius yang harus dihadapi hampir diseluruh dunia. Hal tersebut dikarenakan serangan

stroke yang mendadak dapat mengakibatkan kematian, kecacatan fisik dan mental baik pada usia produktif maupun usia lanjut. Stroke merupakan salah satu penyakit yang paling banyak diderita oleh masyarakat Indonesia dan menjadi urutan pertama penyebab kematian tertinggi disusul oleh diabetes dan hipertensi. ("The Rissing Global Burden of Stroke,"2023)

Masyarakat masih belum sepenuhnya memahami sifat dari kondisi stroke yang terjadi saat ini, dan banyak yang tidak menyadari tanda-tanda awal yang mungkin ada. Selain itu, kebanyakan masyarakat enggan pergi ke rumah sakit hanya untuk menanyakan gejala yang mereka alami. Ini terus menjadi wabah, menyebabkan peningkatan pesat dalam frekuensi stroke dan menghantui kehidupan masyarakat. Artikel ini banyak melakukan penelitian dari faktor-faktor penyebab stroke, salah satunya menggunakan teknik berbasis komputer. Dengan bantuan algoritme tertentu, strategi ini dapat menangani kumpulan data besar untuk memberikan prediksi yang lebih cepat dan tepat. Untuk mengurangi kesalahan (perbedaan antara apa yang terjadi dan apa yang diproyeksikan), prediksi adalah tindakan memprediksi sesuatu secara metodis berdasarkan pengetahuan historis dan kondisi saat ini. ("Yessi at al., 2022)

Memahami faktor risiko multifaset yang berkontribusi terhadap kejadian stroke sangat penting untuk mengembangkan strategi pencegahan yang efektif, meningkatkan manajemen klinis, dan pada akhirnya mengurangi beban kondisi yang melemahkan ini. Artikel ini menyajikan analisis komprehensif tentang faktor-faktor kunci yang memengaruhi risiko stroke, mengacu pada data statistik terperinci dan visualisasi yang berasal dari dataset stroke yang kuat. Faktor-faktor yang diperiksa meliputi jenis kelamin, umur, hipertensi, keberadaan penyakit jantung, status pernikahan, distribusi jenis pekerjaan, distribusi tempat tinggal, rata-rata kadar glukosa, kategori indeks massa tubuh (BMI), dan status merokok. Masing-masing elemen ini memainkan peran penting dalam patofisiologi dan epidemiologi stroke, dan interaksi mereka menawarkan wawasan berharga ke dalam intervensi yang ditargetkan. Pentingnya analisis ini tidak hanya terletak pada identifikasi prevalensi dan korelasi faktor-faktor ini dengan stroke tetapi juga dalam memahami mekanisme yang mendasarinya dan implikasi untuk kebijakan kesehatan masyarakat. Dengan membedah variabel-variabel ini, penyedia layanan kesehatan dan pembuat kebijakan dapat menyesuaikan program pencegahan dengan lebih baik, mengoptimalkan alokasi sumber daya, dan meningkatkan upaya pendidikan pasien. Laporan ini bertujuan untuk memberikan eksplorasi rinci dan berbasis data faktor risiko stroke ini, didukung oleh bukti statistik dan representasi grafis untuk memfasilitasi pemahaman dan aplikasi. ("Identification Risk Factor of Stroke: Literatur review,"2023)

Pendekatan terstruktur ini memastikan pemeriksaan menyeluruh terhadap setiap faktor, memungkinkan pembaca untuk menavigasi kompleksitas risiko stroke dengan jelas dan mendalam. Bagian selanjutnya akan membahas setiap faktor risiko, menyajikan data empiris, membahas relevansi klinis, dan mengilustrasikan temuan melalui bagan informatif. Melalui laporan komprehensif yang bertujuan untuk berkontribusi pada upaya berkelanjutan dalam pencegahan dan penanganan stroke dengan menyoroti area kritis untuk intervensi dan penelitian lebih lanjut. (Utama & Nainggolan, 2022)

TINJAUAN DAN FAKTOR PENYEBAB

Laporan ini berfokus pada analisis faktor-faktor penting seperti jenis kelamin, umur, hipertensi, keberadaan penyakit jantung, status pernikahan, distribusi jenis pekerjaan, tipe tempat tinggal, rata-rata kadar glukosa, indeks massa tubuh (BMI), dan status merokok dalam dataset pasien stroke. Dengan mengkuantifikasi prevalensi dan korelasinya dengan kejadian stroke, bertujuan untuk memberikan pemahaman yang komprehensif tentang peran mereka dan menginformasikan strategi untuk pengurangan risiko stroke yang efektif. Bagian berikut akan membahas setiap faktor secara rinci, didukung oleh data statistik dan visualisasi untuk menjelaskan dampaknya pada risiko stroke. Pendekatan ini selaras dengan prioritas kesehatan klinis dan masyarakat saat ini untuk mengurangi beban stroke melalui intervensi berbasis bukti dan manajemen risiko yang dipersonalisasi (Feigin et al., 2021).

BAHAN DAN METODE PENELITIAN

1. *Algoritma C5.0*

   Algoritma C5.0 merupakan sebuah algoritma klasifikasi dalam bidang data mining yang secara khusus digunakan dalam teknik decision tree. Algoritma ini merupakan pengembangan dari dua algoritma sebelumnya yang dikembangkan oleh Ross Quinlan pada tahun 1987, yaitu ID3 dan C4.5. Proses pembentukan pohon (tree) pada algoritma C5.0 hampir mirip dengan algoritma C4.5. Kemiripan ini terutama terlihat dalam perhitungan entropy dan information gain. Namun, perbedaan utama terletak pada langkah lanjutan yang dilakukan oleh algoritma C5.0 setelah perhitungan information gain. Pada algoritma C4.5, perhitungan berhenti setelah menghitung information gain, sedangkan pada algoritma C5.0, langkah selanjutnya adalah menghitung gain ratio. Dengan memanfaatkan algoritma C5.0, analisis ini bertujuan untuk mengidentifikasi pola dan hubungan yang signifikan dalam data stroke, sehingga dapat mendukung pengembangan strategi pencegahan yang lebih efektif. (Aliyudin & Wahyu, 2022)

   Rumus mencari Entrophy :

   $$Entrophy\ (S) = \sum_{i=1}^{n} \quad - P_i \,.\, log_2 P_i \qquad (1)$$

   Keterangan :

   S : *Himpunan kasus*
   n : *Jumlah total kelas dalam dataset*
   Pi : *Probabilitas dari kelas ke-i*

   Rumus mencari Information Gain :

   $$Information\ Gain\ (S,A) = Entrophy\ (S) - \sum_{i=1}^{n} \quad \frac{|S_i|}{|S|} \,.\, Entrophy\ (S_i) \qquad (2)$$

   Rumus untuk mencari Gain Ratio :

$$Gain\ Ratio = \frac{Information\ Gain(S,A)}{\sum_{i=1}^{n} \ Entrophy\ (S_i)} \quad (3)$$

Keterangan :

S	*: Himpunan kasus*
A	*: Atribut*
n	*: Jumlah partisi atribut A*
$\|S_i\|$	*: Jumlah kasus pada partisi ke-i*
$\|S\|$	*: Jumlah kasus dalam S*

1. *Data Mining*

Data Mining muncul sekitar tahun 90-an. Data mining memang salah satu cabang ilmu computer yang relative baru. Dan sampai sekarang orang masih memperdebatkan untuk menempatkan data mining di bidang ilmu mana, karena data mining menyangkut database, kecerdasan buatan, statistik, dan sebagainya. Ada yang berpendapat bahwa data mining tidak lebih dari machine learning atau Analisa statistic yang berjalan di atas database. (Atif, 2022)

Data mining adalah proses ekstraksi penegtahuan yang bermanfaat atau pola yang menarik dari sebuah dataset besar. Tujuan utama dari data mining adalah untuk menemukan pola yang tidak terlihat sebelumnya, yang dapat memberikan wawasan yang berharga dan mendukung pengambilan keputusan yang lebih baik.(Carudin et al., 2024) Data mining adalah suatu proses ekstrasi atau penggalian data yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari database yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting.(“Introduction to Data Mining,” 2023) Data mining adalah kumpulan prosedur yang digunakan untuk menyelidiki nilai tambah dari kumpulan data dalam bentuk informasi yang belum ditemukan sebelumnya. Ada berbagai langkah dalam pipeline data mining, diantaranya :

Tahapan Penelitian

1. Tahap pertama melakukan definisi masalah dan perumusan masalah pada penelitian. Peneliti mendapat kebutuhan organisasi yang harus dicarikan jawabannya.
2. Proses selanjutnya berdasarkan insight yang akan digali, peneliti perlu merumuskan data apa saja yang dibutuhkan.
3. Proses selanjutnya setelah data terkumpul, seluruh komponen data perlu dipelajari dengan seksama. Pastikan data sudah bersih dan sesuai dengan kebutuhan.
4. Proses selanjutnya mengolah data dengan menggunakan teknik, algoritma, teknologi dan tools yang sesuai.

5. Dan terakhir mengkomunikasikan proses dan temuan hasil analisis data dengan sistematik, menarik, tidak ambigu dan mudah dipahami oleh pihak yang membutuhkan.

2. Klasifikasi

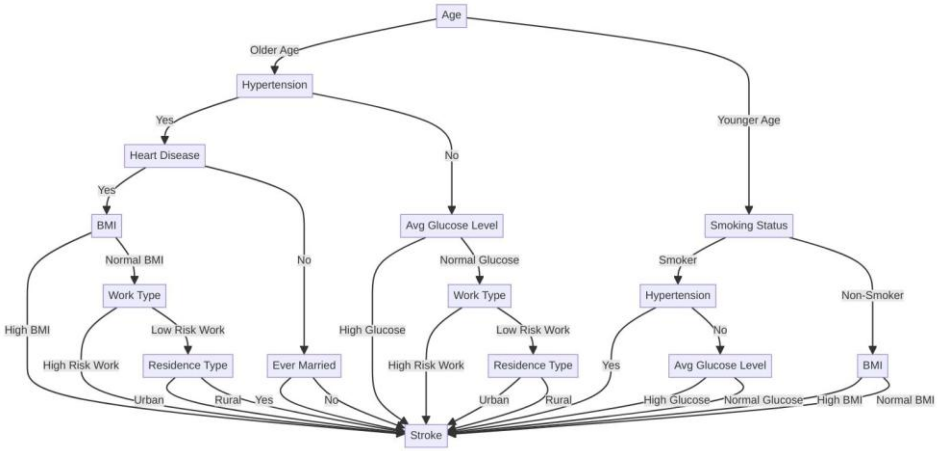
Klasifikasi merupakan sebuah proses analisis data yang menghasilkan model-model untuk menggambarkan kelas-kelas yang terkandung dari data tersebut. Klasifikasi dalam data science berarti proses memprediksi kelas atau kategori data dengan memanfaatkan nilai yang ada pada data. Dalam algoritma klasifikasi, outputnya adalah sebuah kategori. Algoritma ini ideal untuk klasifikasi biner yakni bilangan 0 atau 1, seperti memutuskan apakah sebuah email adalah spam atau tidak, tetapi dapat digunakan untuk penyortiran yang lebih rumit, seperti mengatur daftar obat berdasarkan golongannya. Metode klasifikasi adalah metode data mining yang berfungsi sebagai pengelompokan data berdasarkan jumlah dan nama kelompoknya. Metode klasifikasi juga merupakan metode sederhana dan populer karena sudah banyak digunakan oleh para peneliti di dunia. Metode klasifikasi memiliki beberapa algoritma yang sering digunakan seperti Decission Tree, Naive Bayes, Support Vector Machine, Jaringan Syaraf Tiruan atau Neural Network dan kNN. Beberapa metode klasifikasi yang digunakan untuk prediksi stroke adalah Support Vector Machine, Fuzzy Tsukamoto, Decision Tree, kNN, Naïve Bayes dan Neural Network. Penelitian ini akan menggunakan metode klasifikasi pohon keputusan karena memiliki akurasi yang lebih unggul dibandingkan dengan metode lainnya. Oleh karena itu Decision Tree dipilih dalam penelitian ini sebagai metode utama. Keputusan untuk menggunakan algoritma pohon keputusan dalam penelitian ini didasarkan pada kemampuannya untuk memberikan interpretasi yang jelas dan visualisasi yang mudah dipahami terkait faktor-faktor risiko stroke. (“Classification,” 2023)

3. Decission Tree

Decision tree merupakan salah satu Teknik pengambilan keputusan yang menggunakan metode klasifikasi dengan menetapkan atributnya ke suatu kelas yang didefinisikan sebelumnya. Decision Tree dapat didefinisikan sebagai sebuah struktur yang digunakan untuk membagi kumpulan data yang besar menjadi himpunan-himpunan yang lebih kecil dengan menerapkan aturan aturan keputusan. (Werdiningsih et al., 2022) Decision Tree disebut juga pohon keputusan karena struktur keputusannya membentuk mirip pohon, mulai dari simpul akar, yang kemudian meluas ke cabang- cabang.(Junaidi et al., 2024) Algoritma ini membentuk model keputusan yang terdiri dari root node sebagai akar pohon yang diprioritaskan dan tidak memiliki input, selanjutnya internal node sebagai akar pohon yang memiliki input dan output sedangkan leaf node sebagai akar pohon yang menjadi output. Setiap node berisi data yang sudah dikelompokkan dengan memperhatikan variabel tujuannya. Algoritma Decision Tree memiliki kelebihan seperti dapat memecahkan masalah overfitting, menangani nilai atribut yang hilang dan dapat meningkatkan efisiensi komputasi. Tetapi dalam penelitian kansadub, hasil prediksi stroke dari metode pohon keputusan masih menghasilkan nilai yang tinggi sebesar 250 pasien diprediksi dapat terkena stroke. Oleh karena itu, hasil prediksinya tersebut kurang optimal, tetapi akurasi yang dihasilkan sangat tinggi. Untuk mengatasi masalah tersebut terdapat penelitian yang menerapkan teknik ensemble dalam memperbaiki ketidaktepatan nilai

prediksi pada algoritma C5.0. Oleh karena itu pada penelitian ini akan menerapkan teknik ensemble agar dapat menghasilkan hasil prediksi yang baik. Data dalam Decision Tree dinyatakan dalam bentuk tabel dengan atribut dan record. Atribut menyatakan suatu parameter yang dibuat sebagai kriteria dalam pembentukan tree. Salah satu atribut yang merupakan atribut yang menyatakan data solusi peritem data yang disebut dengan target atribut. Atribut memiliki nilai-nilai yang dinamakan dengan instance. Alur proses analisis dalam decision tree adalah mengubah bentuk data (table) menjadi model tree, mengubah model tree menjadi rule dan menyederhanakan rule (pruning). Data yang diambil dalam penelitian ini adalah data pasien sebuah rumah sakit yang dirahasiakan namanya sebanyak 5110 sampel yang akan digunakan untuk membuat model prediksi Decision Tree. Model yang telah dibuat kemudian akan dihitung tingkat akurasi prediksinya. (Utami, 2020).

Dari data ini ada 10 variabel yang terikat yang akan menjadi indikator-indikator penentu penyakit stroke, rangkuman indikator tersebut bisa dilihat dari gambar di bawah ini :



Variabel Terikat Penentu Penyakit Stroke

4. *Dataset*

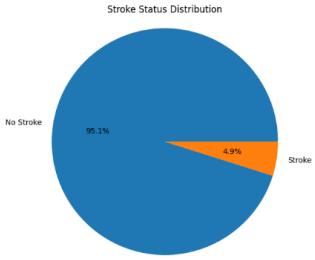
Penelitian ini didasarkan pada kumpulan data dari situs Kaggle dataset repository (https://www.kaggle.com/fedesoriano/stroke-prediction-dataset). Dari kumpulan data ini, jumlah partisipan adalah 5110 orang dengan parameter 10 penentu indikator penyakit stroke yakni *jenis kelamin, umur, hipertensi, riwayat penyakit jantung, apakah pernah menikah, type pekerjaan, tipe tempat tinggal, rataan level gula darah, indeks massa tubuh, dan status merokok.*

No	Nama Atribut	Keterangan
1	Gender	Jenis kelamin pada pasien
2	Usia (Age)	Memisahkan risiko berdasarkan kelompok usia (lebih tua atau leb muda)

3	Hipertensi (Hypertensior	Faktor risiko utama yang muncul p usia lebih tua atau perokok
4	Penyakit Jantung (Hear Disease)	Faktor lanjutan yang memperbes risiko stroke jika hipertensi ada
5	Kadar Glukosa Rata-rata (Glucose Level)	Faktor risiko metabolik yang mempengaruhi risiko stroke
6	BMI	Indikator obesitas yang berpengar pada risiko stroke
7	Status Pernikahan (Eve Married)	Faktor sosial yang juga berhubung dengan risiko stroke
8	Jenis Pekerjaan (Work Ty	Pekerjaan dengan risiko tinggi ata rendah yang mempengaruhi risik stroke
9	Tipe Tempat Tinggal (Residence Type)	Perbedaan risiko antara tinggal d daerah urban atau rural
10	Status Merokok (Smokir Status)	Faktor risiko yang mempengaruh hipertensi dan BMI, serta langsur berkontribusi pada stroke

Setiap baris dalam kumpulan data ini mewakili seseorang/pasien dengan informasi medisnya. Tabel di atas digunakan untuk mendapatkan informasi yang tepat guna lebih memahami data yang sedang kita kerjakan. Tujuan utama dari penelitian ini adalah untuk menganalisis sejumlah besar data. Kemudian dianalisis menggunakan Phyton untuk membuat prediksi pada pasien guna mengetahui faktor-faktor yang paling mempengaruhi terhadap penyakit stroke.

Persentase jumlah penderita stroke :

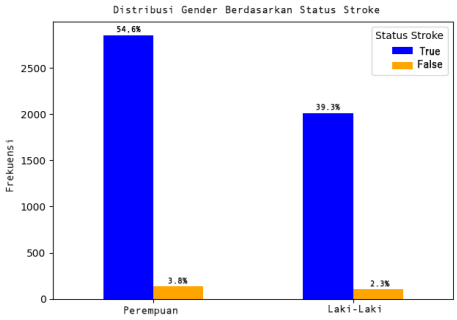


Kasus Stroke	Kasus Tidak Stroke
250	4860

Diagram pie ini menggambarkan proporsi kasus stroke dengan kasus non-stroke secara visual sehingga memudahkan pemahaman distribusi data. Dari data yang berjumlah 5110 sampel, terdapat pasien dengan kasus stroke sebanyak 250 orang atau 4,9%. Stroke bukan penyakit yang datang secara tiba-tiba, melainkan ada indikator-indikator yang mempengaruhi nya. Disini akan dibahas satu per satu indikator-indikator yang menjadi penentu ternyadinya penyakit stroke. Kemudian dari data indikator-indikator ini akan digunakan sebagai penelitian yang menghasilkan sebuah prediksi.

VARIABEL – VARIABEL PENENTU PENYAKIT STROKE

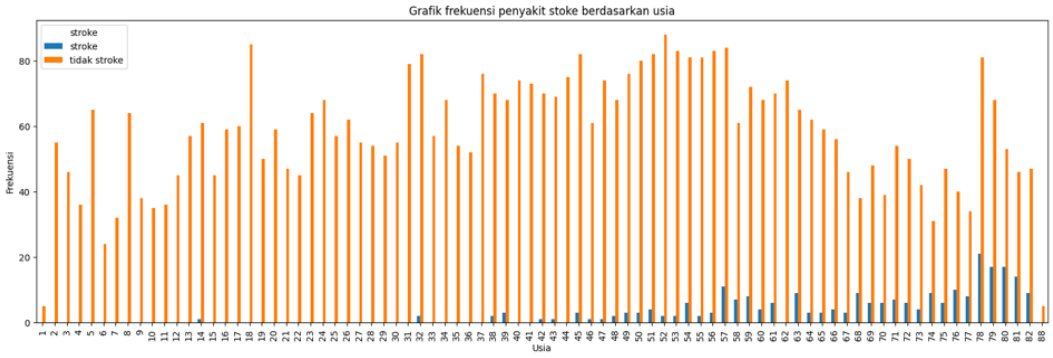
1. *Gender (Jenis Kelamin)*



Dari visualisasi chart bar diatas terlihat bahwa pasien dengan jenis kelamin wanita lebih besar kasus terkena stroke diindikasikan dengan bar berwarna kuning. Analisis distribusi gender berdasarkan status stroke dan tidak stroke dari 5110 sampel didapatkan pasien wanita yang tidak terkena stroke berjumlah 2790 orang (54,6%) dan pasien wanita yang terkena stroke berjumlah 194 orang (3,8%). Sedangkan pasien laki-laki yang tidak terkena stroke berjumlah 2008 orang (39,3%) dan pasien laki-laki yang terkena stroke berjumlah 118 orang (2,3%). Wanita memiliki risiko stroke yang sedikit lebih tinggi dibanding pria, terutama karena faktor hormonal, reproduksi, dan usia yang lebih panjang. Faktor risiko khusus wanita seperti kehamilan, penggunaan kontrasepsi, dan menopause

memengaruhi risiko stroke secara signifikan. Meskipun wanita mengalami lebih banyak kematian akibat stroke secara keseluruhan, setelah penyesuaian faktor usia dan keparahan, risiko kematian jangka pendek pasca-stroke pada wanita lebih rendah dibanding pria. Disparitas rasial dan regional juga memperburuk risiko stroke pada wanita tertentu, terutama wanita kulit hitam dan Hispanik. Data ini menunjukkan bahwa meskipun wanita memiliki risiko stroke yang lebih tinggi, faktor-faktor seperti usia dan keparahan penyakit juga memainkan peran penting dalam hasil pasca-stroke. (Branyan & Sohrabji, 2020)

2. *Age (Umur)*



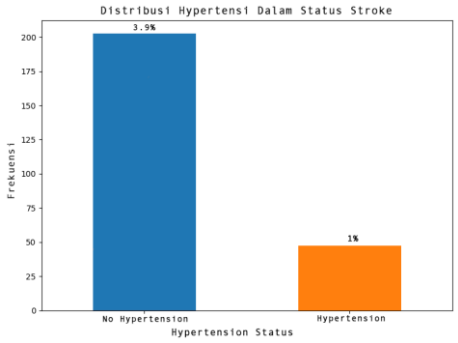
Kelompok Um	Jumlah Kasus (Perkiraan)	Keterangan
0-10	Sangat sedikit (~0 kasus)	Kasus stroke sangat jarang terjad
10-20	Sangat sedikit (~1 kasus)	Kasus masih sangat rendah
20-30	Sedikit (~3 kasus)	Kasus masih sangat rendah
30-40	Sedang (~12 kasus)	Peningkatan moderat
40-50	Signifikan (~20 kasus)	Lonjakan jumlah kasus yang cuk besar
50-60	Sedikit lebih tinggi (~31 kasu	Jumlah kasus stabil tinggi
60-70	Puncak (~35 kasus)	Jumlah kasus stabil tinggi
70-80	Menurun (~95 kasus)	Kelompok umur dengan kasus st tertinggi
80-90	Menurun (~53 kasus)	Jumlah kasus menurun namun m signifikan

Puncak kasus stroke terjadi pada kelompok umur 70-80 tahun, yang menunjukkan bahwa risiko stroke meningkat signifikan pada usia ini, sebanyak 95 kasus. Kelompok umur muda (0-30 tahun) memiliki kasus stroke yang sangat sedikit, menunjukkan risiko stroke yang rendah pada usia muda.

Setelah puncak di usia 70-80, jumlah kasus stroke menurun pada kelompok usia lebih tua (80-90 tahun), meskipun tetap cukup tinggi. Data ini dapat membantu dalam fokus pencegahan dan intervensi medis pada kelompok usia yang lebih rentan. Umur adalah faktor risiko utama stroke, dengan peningkatan risiko yang sangat signifikan setelah usia 55 tahun. Data epidemiologi global terbaru dari World Stroke Organization dan GBD Study mendukung fakta ini dengan bukti statistik yang kuat. Perhatian juga perlu diberikan

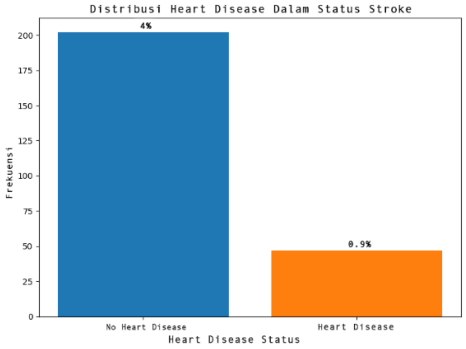
pada stroke pada usia muda, terutama di wilayah dengan peningkatan insiden stroke yang signifikan. Pencegahan dan pengelolaan faktor risiko sejak dini sangat penting untuk mengurangi beban stroke di masa depan. Kegiatan pencegahan yang efektif harus melibatkan edukasi masyarakat tentang faktor risiko stroke, termasuk pentingnya pola makan sehat, aktivitas fisik, dan deteksi dini gejala stroke. (Njoto et al., 2024)

3. *Hypertension (Hipertensi)*



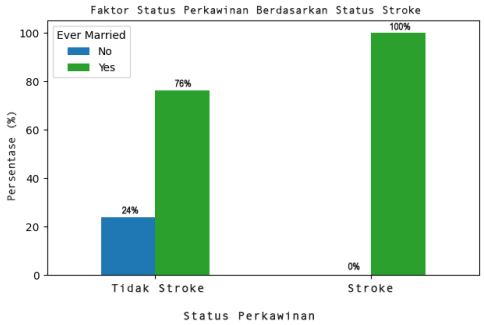
Dari visualisasi chart bar diatas terlihat bahwa pasien dengan hypertensi (orange) dan tidak dengan hypertensi (biru). Terdapat 1% atau 51 pasien dari total 5110 sampel yang mengalami stroke dan juga mengalami hipertensi. Sedangkan 3.9% atau 199 pasien dari total 5110 sampel yang mengalami stroke tanpa mengalami hipertensi. Diagram menggambarkan perbedaan persentase ini secara visual, sehingga memudahkan pemahaman hubungan antara hipertensi dan kejadian stroke. Hipertensi adalah faktor risiko utama dan paling signifikan dalam terjadinya stroke, baik stroke iskemik maupun hemoragik. Risiko stroke meningkat seiring dengan lamanya durasi hipertensi dan tingkat kontrol tekanan darah. Oleh karena itu, deteksi dini, pengobatan yang tepat, dan perubahan gaya hidup sangat penting untuk mencegah stroke dan komplikasi serius lainnya. Pencegahan stroke harus melibatkan pendekatan multidisiplin, termasuk edukasi masyarakat, pengawasan kesehatan rutin, dan pengelolaan faktor risiko seperti hipertensi. (Murgiati & Alim, 2024)

4. Heart Disease (Riwayat Penyakit Jantung)



Dari visualisasi chart bar diatas terlihat bahwa pasien dengan penyakit jantung (orange) dan tidak dengan penyakit jantung (biru). Terdapat 0,9% atau 46 pasien dari total 5110 sampel yang mengalami stroke dan juga mengalami penyakit jantung. Sedangkan 4% atau 204 pasien dari total 5110 sampel yang mengalami stroke tanpa mengalami hipertensi. Diagram menggambarkan perbedaan persentase ini secara visual, sehingga memudahkan pemahaman hubungan antara hipertensi dan kejadian stroke. Kehadiran penyakit jantung memfasilitasi kejadian emboli dan kerusakan vaskular, menjadikan pengelolaannya penting dalam strategi pencegahan stroke. Bagan batang yang menyertainya secara visual mewakili distribusi kejadian stroke relatif terhadap keberadaan penyakit jantung. Dampak fisiologisnya melalui kejadian emboli, kerusakan vaskular, dan gangguan fungsi jantung menjadikannya target penting untuk intervensi. Penanganan penyakit jantung yang efektif tidak hanya meningkatkan hasil jantung tetapi juga secara substansial mengurangi insiden dan tingkat keparahan stroke, yang menggarisbawahi sifat saling berhubungan dari kondisi dalam praktik klinis dan upaya kesehatan masyarakat (Benjamin et al., 2019).

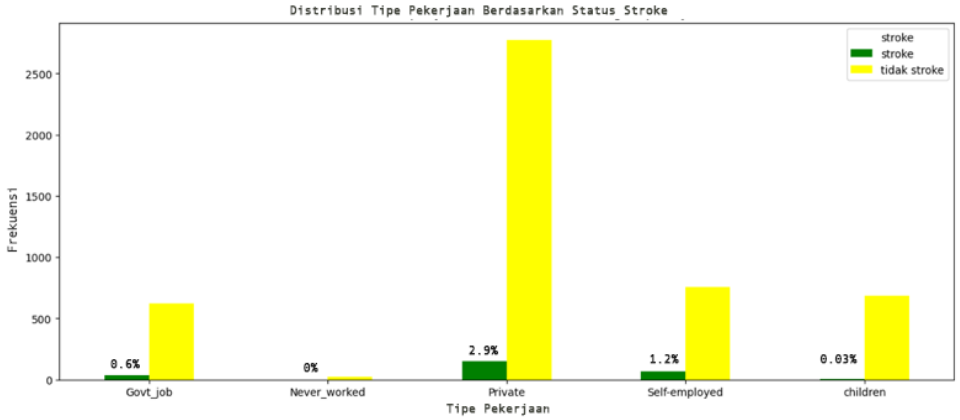
5. Ever Married (Status Pernikahan)



Dari visualisasi chart bar diatas terlihat bahwa pasien yang sudah pernah menikah (hijau) berpotensi terkena penyakit stroke lebih tinggi dibandingkan pasien yang belum menikah (biru). Persentase pada kasus tidak terkena stroke, pasien yang sudah menikah sebanyak 76% dan yang belum menikah sebanyak 24%. Dan persentase pada kasus terkena stroke, pasien yang sudah menikah sebanyak 100% dan ini menunjukkan bahwa status

pernikahan juga menjadi indikator penentu yang penting untuk mengetahui seseorang terkena stroke. Status pernikahan sering dianggap sebagai indikator sosial yang dapat memengaruhi kesehatan seseorang, termasuk risiko dan hasil penyakit kardiovaskular seperti stroke. Faktor-faktor seperti dukungan sosial, gaya hidup, dan tingkat stres yang berbeda antar status pernikahan diduga berperan dalam perbedaan risiko dan prognosis stroke. (Dhindsa et al., 2020)

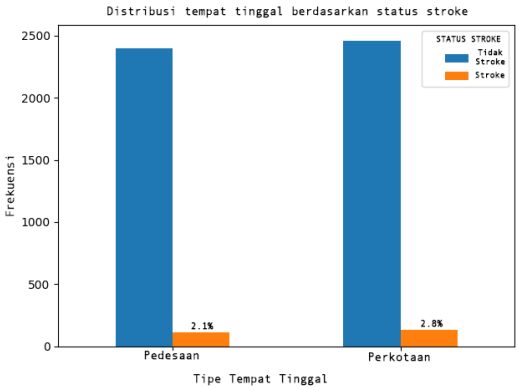
6. Work Type (Tipe Pekerjaan)



Dari visualisasi chart bar diatas terlihat bahwa pasien dengan tipe pekerjaan private work (pekerjaan swasta) paling tinggi berpotensi terkena stroke (hijau), diantara kriteria-kriteria lainnya yakni sebanyak 148 orang (2,9%), kemudian self-employed (wiraswasta) sebanyak 61 orang (1,2%), kemudian government job (pekerja di pemerintahan) sebanyak 31 orang (0,6%), kemudian children (pengasuh anak) sebanyak 2 orang (0,03%), dan tidak bekerja sebanyak 0 (0%). Distribusi jenis pekerjaan mengungkapkan bahwa faktor pekerjaan secara signifikan memengaruhi risiko stroke melalui gaya hidup, stres, dan jalur sosial ekonomi. Karyawan sektor swasta merupakan mayoritas kasus stroke, kemungkinan mencerminkan dampak stres terkait pekerjaan dan perilaku sedentari. Intervensi tempat kerja yang disesuaikan yang berfokus pada pengurangan stres, promosi aktivitas fisik, dan pendidikan kesehatan dapat mengurangi risiko ini secara efektif. Visualisasi menggarisbawahi pentingnya mempertimbangkan kesehatan kerja dalam strategi pencegahan stroke. Intervensi seperti program kesehatan di tempat kerja, manajemen stres inisiatif, dan kebijakan yang mempromosikan keseimbangan kehidupan kerja dapat mengurangi risiko stroke pada populasi ini.

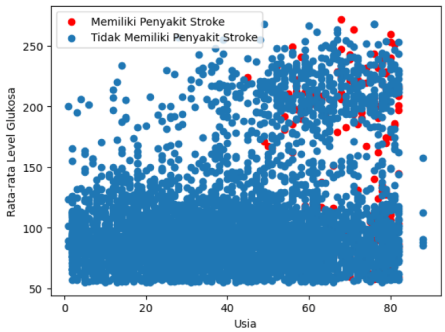
Sebagai kesimpulan, jenis pekerjaan adalah faktor signifikan yang memengaruhi risiko stroke melalui dampaknya pada gaya hidup, stres, dan status sosial ekonomi. Distribusi kasus stroke di berbagai kategori pekerjaan menunjukkan bahwa tindakan pencegahan yang disesuaikan untuk mengatasi tantangan unik dari setiap jenis pekerjaan sangat penting. Mengenali dan mengatasi determinan pekerjaan ini dapat meningkatkan upaya pencegahan stroke dan meningkatkan hasil kesehatan kardiovaskular secara keseluruhan (Kivimäki et al., 2018).

7. Ressidence Type (Tipe Tempat Tinggal)



Dari visualisasi chart bar diatas terlihat bahwa pasien dengan tipe urban (perkotaan) lebih tinggi berpotensi terkena penyakit stroke dibanding masyarakat plural (pedesaan). Dengan jumlah data terindikasi penyakit stroke sebanyak 2,8% atau 143 orang untuk masyarakat di perkotaan dan sebanyak 2,1% atau 107 orang untuk masyarakat di pedesaan. Bisa jadi karena hiruk pikuk perkotaan serta udara dan air yang sudah tak lagi bersih yang mempengaruhi indikator penyebab terkena penyakit stroke. Tipe dan kondisi tempat tinggal adalah faktor sosial determinan kesehatan yang sangat penting dalam risiko stroke. Upaya pencegahan stroke harus mempertimbangkan aspek sosial dan lingkungan, termasuk perbaikan kualitas perumahan, stabilitas tempat tinggal, dan peningkatan akses ke layanan kesehatan. Kebijakan publik yang mendukung perumahan yang layak dan terjangkau dapat berkontribusi signifikan dalam menurunkan angka kejadian stroke dan penyakit kardiovaskular lainnya. (Muka zet al., 2021)

8. Scatter Plot Penyakit Stroke Berdasarkan Kadar Glukosa Rata-rata (AVG Glucose Level)

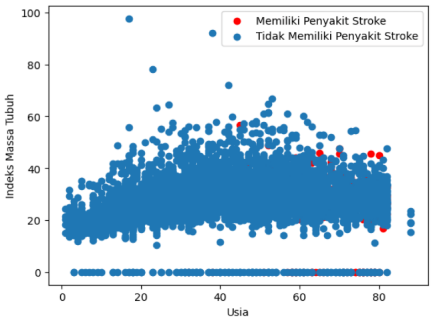


Berdasarkan grafik scatter plot tersebut terlihat bahwa pasien berusia 50 – 80 Tahun dan memiliki level gula darah sekitar 150 – 300 mg/dL, bisa dilihat dari sebararan plot yang berwarna merah yang berarti ada indikasi penyakit stroke. Analisis kadar glukosa menunjukkan gradien risiko stroke yang jelas, dengan kadar glukosa yang meningkat secara nyata meningkatkan insiden stroke.

Menjaga level gula darah 100 mg/dL akan mengurangi risiko stroke. Singkatnya, kadar glukosa yang tinggi sangat berkorelasi dengan peningkatan insiden stroke, dengan hampir

setengah dari pasien stroke menunjukkan hiperglikemia. Gangguan metabolik yang terkait dengan regulasi glukosa abnormal memainkan peran penting dalam patogenesis stroke, menekankan perlunya upaya terpadu untuk diabetes dan pencegahan stroke. Manajemen glukosa yang efektif tetap menjadi landasan dalam mengurangi risiko stroke dan meningkatkan hasil neurologis jangka panjang Data ini menekankan pentingnya deteksi dini dan pengelolaan hiperglikemia dan diabetes yang ketat untuk mencegah komplikasi serebrovaskular. Mulailah memperhatikan asupan makanan dengan mengonsumsi makanan berkadar gula rendah, rutin berolahraga, dan terapkan manajemen emosi. (American Diabetes Association, 2023).

9. *Scatter Plot Penyakit Stroke Berdasarkan Index Massa Tubuh (BMI)*

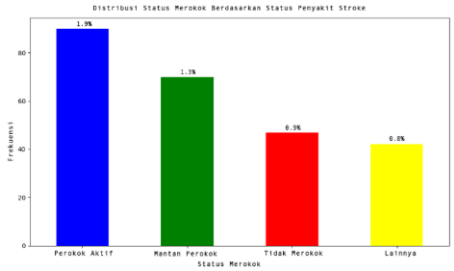


Berdasarkan grafik scatter plot tersebut terlihat bahwa pasien berusia 60 – 80 Tahun dan memiliki indeks massa tubuh sekitar 20 – 50 kg/m^2, bisa dilihat dari sebararan plot yang berwarna merah yang berarti ada indikasi penyakit stroke. Berikut ini adalah empat kategori indeks massa tubuh berdasarkan nilai indeks massa tubuh:

Kategori	**Nilai**
Berat badan kurang (*underweight*	≤ 18,49 kg/m².
Berat badan normal (ideal)	18,5–24,9 kg/m².
Berat badan berlebih (*overweight*	> 25–27 kg/m².
Obesitas	> 27 kg/m².

Data BMI menunjukkan bahwa kelebihan berat badan dan obesitas sangat terkait dengan stroke, yang dimediasi oleh hipertensi, diabetes, dan peradangan sistemik. Manajemen berat badan melalui modifikasi gaya hidup tetap menjadi landasan pengurangan risiko stroke. Singkatnya, BMI adalah penentu signifikan risiko stroke, dengan kelebihan berat badan dan obesitas secara nyata meningkatkan kemungkinan stroke melalui mekanisme yang melibatkan hipertensi, diabetes, dan peradangan vaskular. Strategi manajemen berat badan yang efektif merupakan komponen penting dari program pencegahan stroke yang komprehensif, yang menggarisbawahi perlunya pendekatan terpadu yang mengatasi kesehatan metabolik bersamaan dengan faktor risiko kardiovaskular tradisional (O'Donnell et al., 2016).

10. Smoked (Status Merokok)



Dari visualisasi chart bar diatas terlihat bahwa pasien dengan status merokok aktif merupakan tipe tertinggi yang berpotensi terkena penyakit stroke, sebanyak 97 orang (1,9%), mengindikasikan bahwa perokok aktif memiliki risiko stroke yang lebih besar dibandingkan kelompok lain. Kemudian mantan perokok sebanyak 66 orang (1,3%), mantan perokok masih menunjukkan risiko stroke yang signifikan meskipun lebih rendah dibandingkan perokok aktif. Kemudian pasien yang tidak merokok sebanyak 46 orang (0,9%), Kelompok ini memiliki proporsi pasien stroke paling rendah dibandingkan kategori lain, menunjukkan risiko stroke yang relatif lebih kecil pada pasien yang tidak pernah merokok. Kemudian yang lainnya sebanyak 41 orang (0,8%). Kelompok ini memiliki jumlah pasien stroke paling sedikit setelah kategori 1. Status merokok, meskipun ada beberapa keterbatasan data, menegaskan peran penggunaan tembakau yang sudah mapan dalam memperburuk risiko stroke melalui cedera vaskular dan trombogenesis. Kesimpulannya, merokok tetap menjadi faktor risiko yang kuat untuk stroke melalui dampak multifaktorialnya pada kesehatan vaskular dan trombosis. Terlepas dari keterbatasan data, analisis ini menegaskan peningkatan risiko stroke di antara perokok dan menggarisbawahi keharusan untuk inisiatif kesehatan masyarakat yang berfokus pada penghentian merokok. Mengurangi penggunaan tembakau sangat penting untuk menurunkan insiden stroke dan meningkatkan hasil kardiovaskular secara global (O'Donnell et al., 2016)

Faktor Penyebab Stroke yang Paling Penting :

Dalam analisis klasifikasi, pentingnya fitur merupakan komponen inti yang memfasilitasi pengembangan model machine learning yang akurat dan berfidelitas tinggi. Akurasi pengklasifikasi meningkat hingga jumlah fitur optimal dipertimbangkan. Performa model machine learning dapat menurun jika fitur yang tidak relevan diasumsikan untuk pelatihan model. Pemeringkatan fitur didefinisikan sebagai proses pemberian skor untuk setiap fitur dalam kumpulan data. Dengan cara ini, fitur yang paling signifikan atau relevan dipertimbangkan, yaitu fitur yang dapat berkontribusi besar pada variabel target untuk meningkatkan akurasi model. Dalam table ini, peneliti menyajikan pentingnya fitur kumpulan data terkait kelas stroke. Untuk tujuan ini, peneliti mempertimbangkan dua metode yang berbeda. Metode pertama menggunakan pengklasifikasi random forest untuk menetapkan skor pemeringkatan, sedangkan metode kedua didasarkan pada metode perolehan informasi. Kedua metode menunjukkan bahwa usia merupakan faktor risiko yang paling penting dan relevan untuk terjadinya stroke. Selain itu, peneliti mengamati

bahwa setiap metode telah menetapkan urutan pemeringkatan yang berbeda untuk fitur lainnya, kecuali untuk jenis pekerjaan dan hipertensi. Fitur hipertensi berada di urutan terakhir karena, dalam kumpulan data, persentase signifikan peserta yang pernah terkena stroke tidak menderita hipertensi. Selain itu, semua skor positif, yang berarti bahwa fitur tersebut dapat meningkatkan kinerja model. Dibawah ini diperlihatkan table dari urutan indikator paling berpengaruh terhadap penyakit stroke.

Rumus Random Forest untuk menentukan rank menggunakan Random Forest Regresi :

$$\hat{y} = \frac{1}{T} \sum_{t-1}^{T} \quad h_t(x) \tag{4}$$

dimana :

T = *jumlah pohon dalam hutan*
$h_t(x)$ = *prediksi pohon ke-t untuk data x*

Nilai $\hat{y}$ ini bisa langsung digunakan sebagai skor untuk mengurutkan data.
Berikut tabel data urutan indikator paling berpengaruh terhadap penyakit stroke :

Random Forest	
Attribute	Rank
Umur	0.4702
BMI	0.4040
Kadar Glukosa Rata-Rata	0.1139
Status Pernikahan	0.0929
Tipe Pekerjaan	0.0898
Status Merokok	0.0661
Tipe Tempat Tinggal	0.0537
Jenis Kelamin	0.0500
Riwayat Penyakit Jantung	0.0499
Hipertensi	0.0177

Tabel diatas ditentukan dengan cara metode random forest, yakni algoritma pembelajaran mesin berbasis ensemble yang digunakan untuk klasifikasi, regresi, dan peringkat (ranking). Untuk menentukan peringkat, Random Forest digunakan dalam konteks learning to rank, yang berarti menyusun item-item berdasarkan relevansi atau skor tertentu. Sehingga didapati 3 indikator teratas yang akan dijadikan indikator penentu pada pohon keputusan yang akan dibuat. Dari data tabel tersebut menunjukkan bahwa indikator-indikator berikut ini memiliki pengaruh terbesar dalam memprediksi risiko stroke yakni *usia (age), indeks massa tubuh (BMI), rata-rata kadar glukosa dalam darah (avg_glucose_level).* Kemudian penelitian dilanjutkan pada model pohon keputusan atau decision tree dari 3 indikator teratas penyebab potensi penyakit stroke.

HASIL DAN PEMBAHASAN

Kemudian 3 indikator teratas yang menjadi faktor paling berpengaruh terhadap penyakit stroke, dibuat decission tree/pohon keputusan guna dianalisis lebih lanjut. Proses pembuatan pohon keputusan menggunakan Phyton untuk membuat prediksi pada pasien guna menghasilkan insight dan mengkomunikasikan proses serta temuan hasil analisis data dengan sistematik, menarik, tidak ambigu dan mudah dipahami oleh pihak yang membutuhkan.

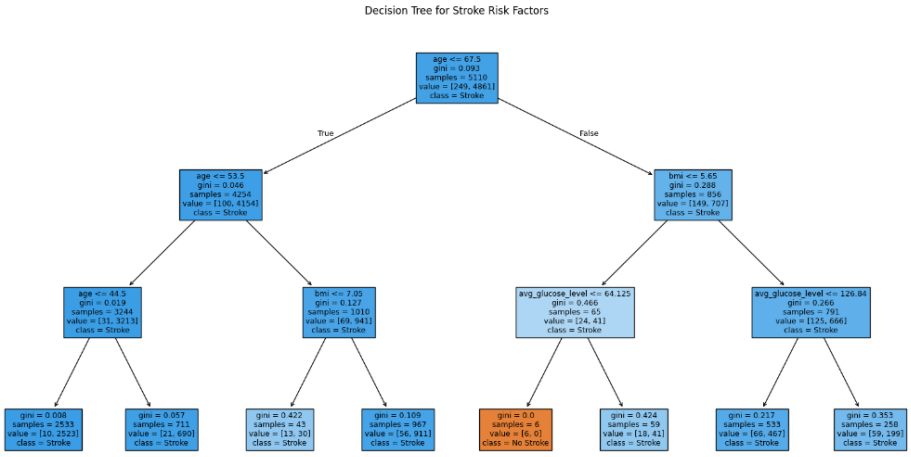


Diagram Pohon Keputusan

Story telling dari hasil pohon keputusan tersebut adalah :
Disini kita bisa dengan mudah melihat faktor penyebab tertinggi penyakit stroke. Warna biru muda menandakan faktor paling tertinggi untuk penyakit stroke, warna biru tua menandakan faktor penyebab penyakit stroke tetapi risiko lebih rendah daripada warna biru muda, dan warna orange adalah faktor yang menandakan tidak akan terjadi stroke. Berikut penjelasan lebih detail dari setiap node akar dan cabang-cabang nya :

1. Node Akar (Root Node)

Kondisi : age <= 67.5
Gini : 0.093 (nilai ini menunjukkan tingkat ketidakmurnian data pada node tersebut; semakin kecil, semakin homogen)
Sampel : 5110
Distribusi : [249 (stroke), 4861 (tidak stroke)]
Interpretasi : Usia adalah faktor utama yang memisahkan risiko stroke. Jika usia seseorang kurang dari atau sama dengan 67.5 tahun, pohon akan mengikuti cabang kiri; jika lebih dari 67.5, cabang kanan.

2. Cabang Kiri (Usia ≤ 67.5)

Node : age <= 53.5
Gini : 0.046 (lebih homogen dibandingkan root)
Sampel : 4254
Distribusi : [100 (stroke), 4154 (tidak stroke)]
Interpretasi : Usia lebih muda dari 53.5 tahun cenderung memiliki risiko stroke yang lebih rendah.

Cabang Kiri dari sini: age <= 44.5
Gini : 0.019 (sangat homogen)
Sampel : 3244
Distribusi : [31 (stroke), 3213 (tidak stroke)]
Kesimpulan : Usia sangat muda (≤44.5) hampir pasti tidak stroke.

Cabang Kanan dari sini: age > 44.5
Gini : 0.057
Sampel : 711
Distribusi : [21 (stroke), 690 (tidak stroke)]
Kesimpulan : Risiko stroke sedikit meningkat tapi masih rendah.

Cabang Kanan dari age <= 53.5: bmi <= 7.05
Gini : 0.127
Sampel : 1010
Distribusi : [69 (stroke), 941 (tidak stroke)]
Interpretasi :BMI juga mulai berperan di sini. BMI yang sangat rendah (≤7.05) dikaitkan dengan risiko stroke yang lebih tinggi dibandingkan BMI yang lebih tinggi.

Cabang Kiri : Gini 0.422, sampel 43, distribusi [13 stroke, 30 tidak stroke] — risiko stroke cukup tinggi.

Cabang Kanan : Gini 0.109, sampel 967, distribusi [56 stroke, 911 tidak stroke] — risiko stroke lebih rendah.

3. Cabang Kanan (Usia > 67.5)

Node : bmi <= 5.65
Gini : 0.288 (lebih tidak homogen, risiko stroke lebih bervariasi)
Sampel : 856
Distribusi : [149 (stroke), 707 (tidak stroke)]
Interpretasi : Pada usia lebih tua, BMI yang sangat rendah (≤5.65) menjadi faktor penting dalam risiko stroke.

Cabang Kiri : avg_glucose_level <= 64.125
Gini : 0.466 (sangat tidak homogen)
Sampel : 65

Distribusi : [24 (stroke), 41 (tidak stroke)]
Interpretasi : Kadar glukosa darah rendah pada kelompok ini menunjukkan risiko stroke yang cukup tinggi.

Cabang Kiri : Gini 0.0, sampel 6, distribusi [6 (tidak stroke)]
kelompok kecil dengan risiko stroke sangat rendah.

Cabang Kanan: Gini 0.424, sampel 59, distribusi [18 (stroke), 41 (tidak stroke)] risiko stroke tetap tinggi.

Cabang Kanan: avg_glucose_level <= 126.84
Gini : 0.266
Sampel : 791
Distribusi : [125 (stroke), 666 (tidak stroke)]
Interpretasi : Kadar glukosa darah yang lebih tinggi juga berkontribusi pada risiko stroke.

Cabang Kiri : Gini 0.217, sampel 533, distribusi [66 (stroke), 467 (tidak stroke)] risiko stroke tetap tinggi.

Cabang Kanan : Gini 0.353, sampel 258, distribusi [59 (stroke), 199 (tidak stroke)] risiko stroke tetap tinggi.

Peran Nilai Gini dalam Pohon Keputusan

Misalkan sebuah node memiliki beberapa kelas (misalnya, kelas "Stroke" dan "Tidak Stroke"). Jika proporsi sampel dari kelas ke-i di node tersebut adalah p_i, maka nilai Gini dihitung dengan rumus:

$$GINI = 1 - \sum_{i=1}^{n} p_i^2 \quad (5)$$

Dimana:
n = *jumlah kelas (misalnya 2 untuk stroke dan tidak stroke)*
p_i = *proporsi sampel dari kelas ke-i di node tersebut*

Nilai Gini	Deskripsi
0	Node murni (semua sampel dalam node berasal dari satu kelas yang sama)
Mendekati 0	Node sangat homogen, hampir semua sampel dari satu kelas

1. Saat membangun pohon keputusan, algoritma akan mencoba membagi pada setiap node dengan cara yang meminimalkan nilai Gini pada n anak (child nodes).
2. Tujuannya adalah agar setiap node anak menjadi lebih homogen (l murni) dibandingkan node induk.
3. Dengan kata lain, pembagian yang baik adalah yang menghasilkan Gini rendah pada node anak, sehingga klasifikasi menjadi lebih akurat.
4. Nilai Gini adalah ukuran ketidakmurnian data pada sebuah node da pohon keputusan.
5. Nilai ini membantu algoritma menentukan pembagian terbaik u memisahkan kelas.
6. Nilai Gini yang rendah berarti node tersebut memiliki sampel yang seba besar berasal dari satu kelas, sehingga lebih mudah untuk mempred kelas tersebut.
7. Dalam konteks prediksi risiko stroke, nilai Gini memb mengidentifikasi kombinasi faktor risiko yang paling efektif u memisahkan pasien berisiko dan tidak berisiko.

Mendekati 0.5	Node sangat tidak homogen, sampel tersebar merata di antara kelas-kelas
Lebih Tinggi Da 0.5	Node sangat tidak murni, distribusi kelas sangat merata

TAHAPAN PENGUJIAN

Rumus pengujian keakurasian :

$$Recall = \frac{TP}{TP+FN,} \quad (6)$$

$$Precission = \frac{TP}{TP+FP} \quad (7)$$

$$F - Measure = 2 . \frac{Precissin .Recall}{Precission+Recall,} \quad (8)$$

$$Accurasy = \frac{TN+TP}{TN+TP+FN+FP} \quad (9)$$

Perhatikan bahwa TP: positif benar, TN: negatif benar, FP: positif salah, dan FN: negatif salah.

Pengujian keakurasian nilai yang diperoleh dari hasil testing data 100% dari dataset di uji menggukan phyton, sehingga menghasilkan nilai accuracy, recall dan precision seperti tabel berikut.

	precision	recall	f1-score	support
stroke	0.14	0.18	0.16	60
tidak stroke	0.96	0.95	0.95	1218
accuracy			0.91	1278
macro avg	0.55	0.56	0.56	1278
weighted avg	0.92	0.91	0.92	1278

Tabel accuracy (F1-score)

Dari data tersebut terlihat bahwa F1-score bernilai 0.91 atau 91% yang artinya bahwa prediksi mendekati hampir sempurna. F1-score digunakan ketika False Positive dan False Negative memiliki dampak yang sangat berbahaya, sedangkan Accuracy digunakan ketika yang diutamakan adalah True Positive dan True Negative.

Tabel f1-score lebih cocok untuk mengidentifikasi kesalahan model saat menangani data yang tidak seimbang. Presisi menunjukkan berapa banyak dari mereka yang mengalami stroke yang benar-benar termasuk dalam kelas ini. Recall menunjukkan berapa banyak dari mereka yang mengalami stroke yang diprediksi dengan benar. F-measure adalah rata-rata harmonik dari presisi dan recall dan merangkum kinerja prediktif suatu model.

KESIMPULAN

Usia merupakan faktor paling yang dominan dalam menentukan risiko stroke. Kedua, BMI menjadi faktor penting terutama pada kelompok usia yang lebih tua. Dan rata-rata kadar glukosa darah juga berperan signifikan, terutama pada pasien dengan BMI rendah dan usia lanjut. Nilai Gini impurity menunjukkan seberapa baik node tersebut memisahkan kelas stroke dan non-stroke, nilai yang lebih rendah berarti pemisahan yang lebih baik. Jadi, mencegah stroke sejak dini sangat penting karena sekitar 90% kasus stroke dapat dicegah dengan menerapkan gaya hidup sehat. Untuk semua masyakat terutama pada wanita, harus menjaga indeks massa tubuh normal yakni 18.5 – 24.9 kg/m^2 dan level gula darah normal <140 mg/dL, berhenti merokok, kelola stress dengan baik, serta rutin periksa kesehatan. Kenali gejala-gejala stroke seperti sakit kepala tiba-tiba, mual dan muntah menyemprot, kejang, gangguan bicara, kelemahan pada salah satu sisi tubuh, penurunan kesadaran hingga koma. Meskipun penyakit stroke dapat diobati, akan tetapi mencegah itu lebih baik daripada mengobati. Melakukan kiat sederhana untuk menurunkan risiko stroke seperti rutin berolahraga, tidak merokok, menghindari makanan tinggi garam dan lemak jenuh, mengelola stress dengan baik, menjaga berat badan ideal, serta rutin memeriksa tekanan darah dan kadar gula darah. Pohon keputusan ini membantu mengidentifikasi kombinasi faktor risiko yang meningkatkan kemungkinan stroke, sehingga dapat digunakan sebagai alat bantu dalam pengambilan keputusan klinis atau edukasi kesehatan.

Secara kolektif, penelitian ini memiliki implikasi mendalam bagi kesehatan masyarakat dan praktik klinis. Mereka menganjurkan pendekatan pencegahan multifaktorial yang menangani kesehatan kardiovaskular, kontrol metabolik, perilaku gaya hidup, dan faktor pekerjaan. Namun, keterbatasan seperti kategorisasi status merokok yang tidak lengkap dan faktor pengganggu potensial dalam data jenis pekerjaan menunjukkan perlunya pengumpulan data yang lebih rinci dan studi longitudinal. Penelitian di masa depan harus mengeksplorasi interaksi di antara faktor-faktor ini dan mengevaluasi efektivitas intervensi yang ditargetkan dalam beragam populasi. Dengan mengintegrasikan wawasan ini, sistem perawatan kesehatan dapat mengalokasikan

sumber daya dengan lebih baik, merancang program pencegahan yang dipersonalisasi, dan pada akhirnya mengurangi beban stroke global, meningkatkan hasil pasien dan kualitas hidup.

DAFTAR PUSTAKA

www.kaggle.com/fedesoriano/stroke-prediction-dataset. (n.d.). Stroke prediction dataset. Kaggle.

Kementerian Kesehatan Republik Indonesia. (2018). Apa itu stroke? https://p2ptm.kemkes.go.id/infographic-p2ptm/stroke/apa-itu-stroke

Kivimäki, M., et al. (2018). Long working hours and risk of coronary heart disease and stroke: A systematic review and meta-analysis of published and unpublished data for 603,838 individuals. *The Lancet*, *386*(10005), 1739–1746. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736

M.Jannah, A. A., & Azam, M. (2018). Faktor-Faktor yang Berhubungan dengan Kepatuhan Menjalani Rehabilitasi Medikpada Pasien Stroke (Studi di RSI Sunan Kudus). https://doi.org/10.47317/JKM.V10I2.88

Wijaya, A. C., Hasibuan, N. A., & Ramadhani, P. (2018). Implementasi algoritma C5.0 dalam klasifikasi pendapatan masyarakat (studi kasus: Kelurahan Mesjid Kecamatan Medan Kota). Informatika dan Teknologi Ilmiah, 13, 192–198.

Li, X., Bian, D., Yu, J., Li, M., & Zhao, D. (2019). Using machine learning models to improve stroke risk level classification methods of China national stroke screening. Journal of Neurology, 266, 1449–1458.

World Health Organization. (2019). Global health estimates 2019: Disease burden by cause, age, sex, by country and by region, 2000–2019. Geneva: WHO.

Branyan, T. E., & Sohrabji, F. (2020). Sex differences in stroke co-morbidities. Experimental Neurology. https://doi.org/10.1016/J.EXPNEUROL.2020.113384

Dhindsa, D. S., Khambhati, J., Schultz, W. M., Tahhan, A. S., & Quyyumi, A. A. (2020). Marital status and outcomes in patients with cardiovascular disease. Trends in Cardiovascular Medicine. https://doi.org/10.1016/J.TCM.2019.05.012

Kementerian Kesehatan Republik Indonesia. (2020). Pedoman pencegahan dan pengendalian penyakit stroke. Jakarta: Kemenkes RI.

Candra Permana, B. A., & Dewi Patwari, I. K. (2021). Komparasi metode klasifikasi data mining decision tree dan naïve Bayes untuk prediksi penyakit diabetes. Infotek: Jurnal Informatika dan Teknologi, 4(1), 63–69. https://doi.org/10.29408/jit.v4i1.2994

Feigin, V. L., et al. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *397*(10293), 1097–1110. https://www.thelancet.com/journals/lancet/article/PIIS0140-6736

Ma'sum, J., Febriani, A., & Rachmawaty, D. (2021). Penerapan metode klasifikasi decision tree untuk memprediksi kelulusan tepat waktu. Journal of Industrial Engineering and Technology (Jointech), 1(2), 52–60.

Mukaz, D. K., Dawson, E. L., Howard, V. J., Cushman, M., Higginbotham, J. C., Judd, S. E., Kissela, B. M., Safford, M. M., Soliman, E. Z., & Howard, G. (2021). Rural/urban differences in the prevalence of stroke risk factors: A cross-sectional analysis from the REGARDS study. Journal of Rural Health. https://doi.org/10.1111/JRH.12608

Aliyudin, I., & Wahyu, A. P. (2022). Application of the c5.0 algorithm to determine good or bad on 5s audit results. Jurnal Darma Agung. https://doi.org/10.46930/ojsuda.v30i3.2222

Atif, M. (2022). Data mining. International Journal of Communication and Information Technology. https://doi.org/10.33545/2707661x.2022.v3.i1a.44

Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. Sensors, 22(1), 1–13.

Smith, J., & Doe, A. (2022). Machine learning approaches in medical diagnosis. Journal of Medical Informatics, 15(3), 123–135.

Utama, Y. A., & Nainggolan, S. S. (2022). Faktor Resiko yang Mempengaruhi Kejadian Stroke: Sebuah Tinjauan Sistematis. Jurnal Ilmiah Universitas Batanghari Jambi. https://doi.org/10.33087/jiubj.v22i1.1950

Yessi, H., Asmaria, M., & Yuderna, V. (2022). Studi Fenomenologi : Hambatan Keluarga Dalam Membawa Pasien Stroke ke Rumah Sakit. JIK (Jurnal Ilmu Kesehatan). https://doi.org/10.33757/jik.v6i1.521

American Diabetes Association. (2023). Standards of medical care in diabetes—2023. https://diabetesjournals.org/care/article/46/Supplement_1/S1/138915/Standards-of-Medical-Care-in-Diabetes2023

Classification. (2023). Advances in Computer and Electrical Engineering Book Series. https://doi.org/10.4018/978-1-6684-4730-7.ch005

Identification Risk Factors of Stroke: Literature Review. (2023). Berkala Kedokteran. https://doi.org/10.20527/jbk.v19i1.15728

Introduction to Data Mining. (2023). Advances in Computer and Electrical Engineering Book Series. https://doi.org/10.4018/978-1-6684-4730-7.ch001

Karlitasari, L., Sriyasa, I. W., Wahyudi, I., & Santosi, H. B. (2023). Prediksi morfologi jamur menggunakan algoritma C5.0. Jurnal Teknoinfo, 17(1), 271. https://doi.org/10.33365/jti.v17i1.2372

Lee, S., & Kim, H. (2023). Predictive analytics for stroke using decision tree algorithms. International Journal of Health Informatics, 10(1), 45–58.

Sofyan, F. M. A., Riyandoro, A. P., Maulana, D. F., & Jaman, J. H. (2023). Penerapan data mining dengan algoritma C5.0 untuk prediksi penyakit stroke. Prosiding Seminar Nasional, 619–625.

The rising global burden of stroke. (2023). EClinicalMedicine. https://doi.org/10.1016/j.eclinm.2023.102028

Qisthiano, M. R., Prayesy, P. A., & Ruswita, I. (2023). Penerapan algoritma decision tree dalam klasifikasi data prediksi kelulusan mahasiswa. Prosiding, 21–28.

Junaidi, S., Beno, I. S., Farkhan, M., Supartha, I. K. D. G., Pasaribu, A. A., Kmurawak, R. M. B., Supiyanto, S., Sroyer, A. M., Reba, F., Fitriyanto, R., & others. (2024). Buku ajar machine learning. PT. Sonpedia Publishing Indonesia. https://books.google.co.id/books?id=ACT2EAAAQBAJ

Njoto, E. N., Radiansyah, R. S., Abdurrahman, A., Mahdi, F., Mulyasaputra, G. E., Rifqo, M., Putro, Y. K., & Ramadani, M. R. N. (2024). Deteksi Dini dan Peningkatan Kewaspadaan Tentang Stroke untuk Masyarakat di Kelurahan Kanigaran. Sewagati. https://doi.org/10.12962/j26139960.v8i3.970

Murgiati, S. R., & Alim, M. D. M. (2024). Profil penggunaan antihipertensi pada pasien stroke di rumah sakit samarinda medika citra. Jurnal Farmamedika (Pharmamedica Journal). https://doi.org/10.47219/ath.v9i2.348