IMPLEMENTASI ALGORITMA *K-MEANS CLUSTERING* DAN *NAÏVE BAYES* UNTUK MENGETAHUI FAKTOR PEMICU DAN JUMLAH PENDERITA STROKE OTAK

PROPOSAL TUGAS AKHIR

Diajukan oleh :

M. Rizky Wijaya

8020190150

Untuk Persyaratan Penelitian Dan Penulisan Tugas Akhir

Sebagai Akhir Proses Studi Strata 1

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS ILMU KOMPUTER

UNIVERSITAS DINAMIKA BANGSA

2022

IDENTITAS PROPOSAL PENELITIAN

Judul Proposal : Implementasi Algoritma K-Means Clustering dan Algoritma Naive Bayes Untuk Mengetahui Faktor Pemicu Dan Jumlah Penderita Stroke Otak

Program Studi : Teknik Informatika

Jenjang Pendidikan : Strata 1 (S1)

Peneliti :

a. Nama Lengkap : M. Rizky Wijaya
b. NIM : 8020190150
c. Jenis Kelamin : Laki-laki
d. Tempat/Tgl. Lahir : Jambi/22 April 2001
e. Alamat : Jl. Pangeran Diponegoro RT.01 No.108 Kel. Tambak Sari Kec. Jambi Selatan.
f. No. Telepon : 0821-7746-0642
g. Email : m.rizkywijaya2000@gmail.com

1.1. LATAR BELAKANG

Stroke merupakan cidera yang mengancam diri manusia sehingga membutuhkan perawatan *neutokritis*. Namun *stroke* belum sepenuhnya diperiksa karena beberapa kemungkinan alasan yang terjadi, karna tidak ada definisi atau klasifikasi yang diterima secara universal,[1] pada saat ini stroke juga terjadi pada orang dibawa 40 tahun. *Stroke* merupakan penyakit yang disebabkan oleh penyumbatan darah di otak dan merupakan penyakit terbanyak ketiga setelah penyakit jantung dan kanker serta penyebab kecacatan tertinggi. Menurut *American Heart Association* (AHA) angka kematian di Amerika pertahunnya mencapai 50-100 dari 100.000 kasus penderita.

Di negara ASEAN sendiri penyakit *stroke* juga menjadi masalah serius. Angka kematian terbesar di ASEAN terjadi di Indonesia diikuti dengan Filipina, Singapura, Brunei, Malaysia, dan Thailand.[2] Hal ini membuat penyakit stroke tidak bisa di anggap remeh khususnya di Indonesia sendiri mengingan strok bisa datang secara tiba-tiba dan dapat mengancam nyawa. Maka dari itu penulis ingin mengolah data stroke dengan implementasi data dengan menggunakan dua metode yang berjudul **"Implementasi algoritma k-Means clustering dan algoritma naive bayes untuk mengetahui faktor pemicu dan jumlah penderita stroke otak"**. Metode *Naïve Bayes* adalah salah satu metode yang ada di dalam data mining untuk mengklasifikasikan data,[3] sedangkan Metode *K-Means* adalah salah satu metode dalam fungsi *clustering* atau pengelompokan, clustering mengacu pada pengelompokkan data, observasi atau kasus berdasarkan kemiripan objek yang diteliti.[4]

Penelitian ini bertujuan mengelompokkan dan mengklasifikasi data untuk mengetahui tingkat potensi terkena penyakit *stroke* dengan berbgai faktor pemicu seperti rokok, dan lain sebagainya. Penelitian ini menggunakan dua metode agar hasil dari impelmentasi data ini lebih tepat sasaran.

1.2. RUMUSAN MASALAH

Pada latar belakang yang telah dijelaskan di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

1. Mengelolah data stroke otak dengan menggunakan metode naïve bayes dan K-Means Clustering.
2. Bagaimana mengimplementasi metode K-Means Clustering pada data stoke.
3. Bagaimana mengimplementasi metode naïve bayes pada data stoke.

1.3. BATASAN MASALAH

Pembatasan masalah yang digunakan dalam sebuah pembahasan bertujuan agar dalam pembahasannya lebih terarah dan sesuai dengan tujuan yang akan dicapai. Maka penulis membatasi permasalahan seperti berikut ini:

1. Penelitian ini difokuskan pada pengelolaan data stroke otak yang terdiri dari 4981 data.
2. Objek dalam penelitian ini bersumber pada www.kaggel.com.
3. Untuk mengolah data ini, perangkat lunak yang akan digunakan adalah aplikasi WEKA dengan menggunakan metode K-Means Clustering dan Naïve Bayes.
4. Penelitian ini terdiri dari 11 atribut yaitu jenis kelamin, usia, hipertensi, penyakit jantung, tipe kerja, tipe tempat tinggal, tingkat avgglukosa, bmi arau indeks massa tubuh setatus perokok.

1.4. TUJUAN DAN MANFAAT PENELITIAN

Adapun tujuan penelitian yang dilakukan oleh penulis, yaitu:

1. Untuk dapat mengetahui seberapa besar potensi terkena stroke otak dan faktor pemicu terkena stroke.

2. Mengelolah data stroke otak yang bertujuan untuk mengetahui seberapa banyak dan seberapa besar potensi terkena stroke dengan menerapkan algoritm *K-Means Clustering* Dan *Naïve Bayes*.

1.5. LANDASAN TEORI

1.5.1. DATA MINING

Data mining ialah salah satu inti proses yang didapat dalam suatu KDD. Kebanyakan orang menganggap data mining sebagai sinonimnya KDD, Karena data mining merupakan sebagian besar yang difokuskan dalam pekerjaan KDD. Namun, langkah-langkah lainnya ialah proses-proses penting yang menjamin kesuksesan pada aplikasi KDD.[5]

Data Mining ialah proses ekstraksi informasi dari kelompok data yang dilakukan degan menggunakan algoritma dan teknik yang mencantukan bidang ilmu statistik, mesin pembelajaran, dan sistem manajemen database. Data Mining berguna untuk ringkasan informasi penting yang tersembunyi dari dataset yang besar. Dengan adanya data mining maka akan didapatkan sesuatu yang berharga berupa informasi pengetahuan di dalam kumpulan data - data yang banyak jumlahnya.[6]

Data mining merupakan analisis dari peninjauan kumpulan data untuk menemukan hubungan yang tidak diduga dan meringkas data dengan cara berbeda dengan cara yang berbeda dengan sebelumnya, yang dapat dipahami dan bermanfaat bagi pemilik data. Data mining merupakan bidang dari beberapa bidang keilmuan yang menyatukan teknik dari pembelajaran mesin, pengenalan pola, statistic, database, dan visualisasi untuk penanganan permasalahan pengambilan informasi dari database yang besar. Data Mining adalah suatu istilah yang digunakan untuk menguraikan penemuan pengetahuan di dalam database. Data mining adalah proses yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi yang bermanfaat dan pengetahuan yang terikat dari berbagai database besar. maka data mining merupakan

pengetahuan yang tersembunyi di dalam database yang di proses untuk menemukan pola dan teknik statistik matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan dari database tersebut.[7]

Dari pejelasan di atas maka, data mining merupakan sebagian besar yang difokuskan dalam pekerjaan KDD yang berguna untuk mencari informasi berharga yang tersembunyi dari data yang besar yang di proses dalam menemukan pola dan teknik statistik matematika, kecerdasan buatan, dan machine learning untuk mengekstraksi dan mengidentifikasi informasi pengetahuan dari database tersebut.

1.5.2. NAÏVE BAYES

Naïve Bayes merupakan salah satu metode yang dapat digunakan untuk mengklasifikasikan data. Bayesian classification merupakan pengklasifikasian statistik yang dapat digunakan untuk memprediski probabilitas keanggotaan suatu class.[8]

Naive bayes merupakan salah satu metode statistik untuk klasifikasi yang memungkinkan untuk menangkap ketidak pastian tentang suatu model dengan cara berprinsip pada mendefinisikan hasil probabilitas. Metode ini digunakan untuk menyelesaikan masalah diagnosa dan prediksi.[9] Teorema Bayes memiliki bentuk umum sebagai berikut:

1.5.3. $$P(H \mid X) = \frac{P(X \mid H)\, P(H)}{P(X)}$$ 1.5.4.

Dimana :

X = Data dengan class yang belum diketahui.

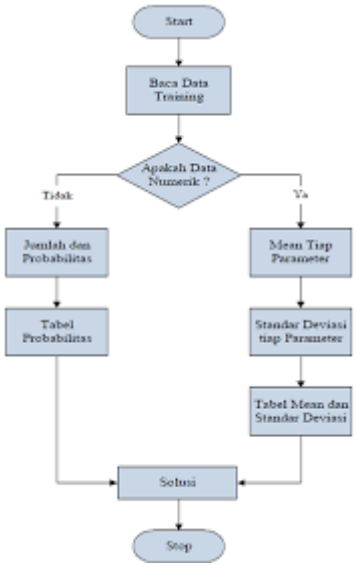
H = Hipotesis Data X merupakan suatu class spesifik

P (H | X) = probabilitas hipotesis H berdasarkan kondisi x (posteriori prob.)

P (H) = Probabilitas hipotesis H (prior prob.)

P (X | H) = probabilitas X berdasarkan kondisi tersebut

P (X) = probabilitas dari X



Gambar 1.5.2 Alur Metode Naïve Bayes

Naïve Bayes merupakan suatu kelas keputusan, dengan menggunakan perhitingan probabilitas matematika dengan syarat bahwa nilai keputusan adalah benar, berdasarkan informasi obyek.[10]

Dari ketiga penjelasan diatas dapat disimpulkan bahwa *naïve bayes* merupakan sebuah metode dalam data mining untuk menentukan klasifikasi data dengan menggunakan perhitingan probabilitas matematika yang berfungsi untuk memprediksi probabilitas keanggotaan sebuah class.

1.5.5. CLUSTERING

Menurut Widodo (2013:9) *Clustering* atau klasifikasi adalah metode yang digunakan untuk membagi rangkaian data menjadi beberapa group berdasarkan kesamaan-kesamaan yang telah ditentukan seblumnya. *Cluster* adalah sekelompok atau sekumpulan objek-objek data yang similar satu sama lain dalam *cluster* yang sama dan dissimilar terhadap objek-objek yang berbeda *cluster*.[11]

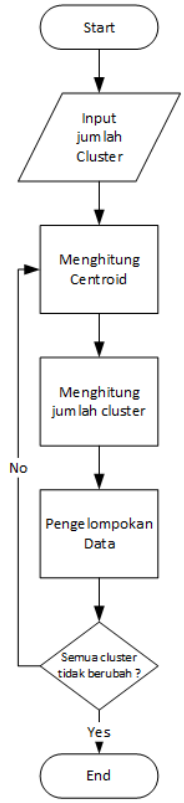
Clustering merupakan salah satu teknik dari salah satu fungsionalitas data mining, algoritma clustering merupakan algoritma pengelompokkan sejumlah data menjadi kelompok-kelompok data tertentu (cluster).[12]

Algoritma *K-Means Clustering* merupakan salah satu algoritma yang mengelompokkan data yang sama pada kelompok tertentu dan data yang berbeda pada kelompok yang lain.[13]

Dari penjelasan diatas, maka clustering merupakan sebuah algoritma yang terdapat pada data mining yang berfungsi untuk mengelompokkan sebuah data yang sama agar lebih mudah untuk mendapatkan informasi.

$$[(x, y), (a, b)] = \sqrt{(x-a)^2 + (y-b)^2}$$

persamaan *Euclidean Distance*



Gambar 1.5.5 Flowchart Clustering

1.5.6. K-MEANS

Algoritma *K-Means* merupakan metode nonheirarchial yang pada awalnya mengambil sebagian dari banyaknya komponen dari populasi untuk dijadikan pusat cluster awal. Pada step ini pusat cluster dipilih secara acak dari sekumpulan populasi data.[14]

K-Means pertama kali dipublikasikan oleh Stuart Lloyd pada tahun 1984 dan merupakan algoritma clustering yang banyak digunakan. K-Means bekerja dengan mensegmentasi objek yang ada ke dalam kelompok atau yang disebut dengan segmen sehingga objek yang berada dalam masing-masing kelompok lebih serupa satu sama lain dibandingkan dengan objek dalam kelompok yang berbeda.[15]

Algoritma K-Means merupakan salah satu algoritma dalam fungsi clustering atau pengelompokan. Clustering mengacu pada pengelompokkan data, observasi atau kasus berdasar kemiripan objek yang diteliti.[16]

Dari ke tiga penjelasan di atasdapat disimpulkan bahwa *K-Means* pertama kali dipublikasikan oleh Stuart Lloyd pada tahun 1984 yang merupakan metode *nonheirarchial* untuk memusatkan suatu cluster dari sebuah data.

1.5.7. WEKA

WEKA (*Waikato Environment for Knowledge Analysis*) merupakanperangkat lunak data mining yang dikembangkan oleh Universitas Waikato, New Zealand. Diimplementasikan pertama kali pada tahun 1997 dan mulai menjadi *open source* pada tahun 1999.[17]

WEKA merupakan salah satu sistem yang digunakan untuk melakukan mining data. Aplikasi ini berlisensi GNU *General Public License* sehingga dapat digunakan secara gratis. WEKA menggunakan *framework Java* sehingga dapat digunakan di berbagai macam sistem operasi. Penggunaan aplikasi ini bertujuan untuk memudahkan

pengguna dalam memproses data dari mulai pemrosesan awal hingga pemodelan data (Adinugroho & Sari, 2018).[18]

WEKA (*Waikato Environment for Knowledge Analysis*) merupakan aplikasi data mining *open source* berbasis *Java* menurut "Sulianta" dalam. WEKA merupakan sebuah sistem data mining yang dikembangkan oleh Universitas Waikato di Selandia Baru yang mengimplementasikan algoritma data mining. WEKA adalah sebuah koleksi mesin pembelajaran algoritma untuk tugas-tugas data mining.[19]

Dari ketiga penjelasan di atas dapat disimpulkan bahwa WEKA adalah sebuah aplikasi *open source* yang *berbasis java* yang digunakan untuk melakukan mining data yang diimplementasikan pertama kali pada tahun 1997 dan menjadi open source pada tahun 1999.

1.6. METODOLOGI PENELITIAN

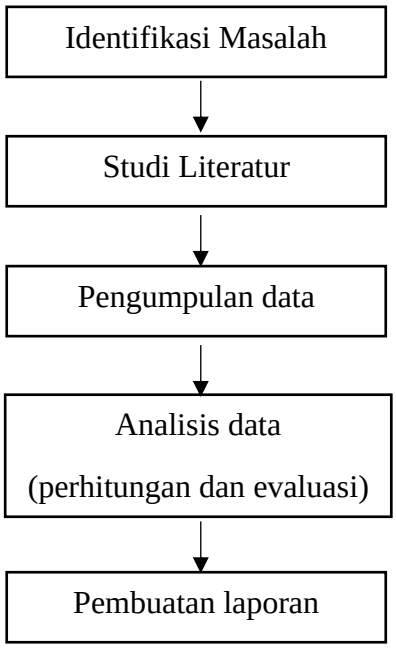
a. Alat dan Bahan

Pada latar belakang yang telah dijelaskan di atas, maka rumusan masalah dalam penelitian ini adalah sebagai berikut :

4. Perangkat Keras yang dibutuhkan, dengan spesifikasi sebagai berikut:
    - Laptop Asus X455L, dengan *processor* Intel core i3 4055U, 1.7Ghz. *RAM* 2.00 GB, *HDD,* 500 GB.
5. Perangkat lunak yang dibutuhkan dalam penelitian ini adalah sebagai berikut:
    - *Windows* 8
    - *Windows* 10
    - *Browser (Google Chrome)*
    - Apliksi WEKA
    - *Microsoft Word* dan *Excel* 2013

b. Metode Penelitian

Agar penelitian ini dapat berjalan dengan lancar, maka diperlukan sebuah kerangka kerja penelitian untuk dapat mengetahui langkah-langkah yang dibutuhkan dalam pembuatan penelitian ini agar penelitian ini dapat berjalan sesuai rencana, terstruktur, dan dapat diselesaikan tepat waktu. Adapun kerangka kerja yang digunakan adalah sebagai berikut:



Gambar 1.6. Kerangka Kerja

1. Identifikasi Masalah

   Identifikasi masalah merupakan langkah pertama yang akan dilakukan dalam pembuatan penelitian. Hal ini bertujuan agar peneliti dapat menentukan penelitan apa yang akan penulis angkat untuk diteliti. Pada tahapan ini penulis melakukan identifikasi masalah terhadap klasifikasi dan pengelompokan data penyakit stroke otak yang akan di teliti.

2. Studi Literatur

   Pada tahapan studi literatur ini, penulis melakukan kajian pustaka yang bermaksud untuk mencari, mempelajari dan mengamati refrensi-refrensi, jurnal dan lain sebagainya yang serupa atau relevan dalam penelitian yang akan penulis teliti. Studi literatur ini bertujuan untuk mendapatkan landasan teoritis mengenai permasalahan yang akan diteliti, yang bertujuan untuk dapat memahami permasalahan yang akan diteliti.

3. Pengumpulan data

   Pada tahapan ketiga ini Merupakan tahapan dalam melakukan pengumpulan data dan informasi yang akan penulis teliti dalam mengolah data stroke otak dengan metode naïve bayes dan k-means clustering, agar dapat tercapainya suatu tujuan dalam penelitian ini. penulis menggunakan beberapa cara dalam mengumpulkan data, yaitu sebagai berikut :

   a. Pengamatan (Observation)

      Observasi dilakukan dengan mengamati data yang .

   b. Study Relevansi

      Prinsip relevansi merupakan salah satu dari prinsip pengembangan kurikulum, sebagai pedoman agar kurikulum terus selaras dengan perkembangan jaman.

4. Analisis Data (Perhitungan dan Evaluasi)

   Pada tahapan ini penulis melakukan penganalisisan kepada sebuah data stroke yang telah didapat dan kemudian dilakukan pengolahan data dengan cara melakukan perhitungan dan evaluasi data dengan metode *K-Means Clustering* Dan *Naïve Bayes.*

5. Pembuatan Laporan

   Setelah semua komponen diatas terpenuhi, maka langkah terakhir yang dilakukan adalah penyusunan laporan.

1.7. JADWAL PENELITIAN

Adapun jadwal penelitian yang dilakukan untuk menyelesaikanpenelitian ini adalah :

No.	Kegiatan	Bulan															
		Oktober				November				Desember				Januari			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1.	Penyusunan Proposal																
2.	Pengumpulan Data																
3.	Analisis																
4.	Algoritma Data Produk																
5.	Evaluasi																
6.	Pengumpulan Laporan																

DAFTAR PUSTAKA

[1] S. K. Mittal and R. K. Gupta, “Heat stroke.,” *Indian Pediatr.*, vol. 23 Suppl, pp. 155–160, 1986.

[2] C. A. Dinata, Y. Syafrita, and S. Sastri, “Artiakel Penelitian,” *J. Kesehat. Andalas*, vol. 2, no. 2, 2013, [Online]. Available: http://jurnal.fk.unand.ac.id

[3] I. Riadi, R. Umar, and F. D. Aini, “Analisis Perbandingan Detection Traffic Anomaly Dengan Metode Naive Bayes Dan Support Vector Machine (Svm),” *Ilk. J. Ilm.*, vol. 11, no. 1, pp. 17–24, 2019, doi: 10.33096/ilkom.v11i1.361.17-24.

[4] E. Muningsih and S. Kiswati, “Penerapan Metode K-Means Untuk Clustering Produk Online Shop Dalam Penentuan Stok Barang,” *J. Bianglala Inform.*, vol. 3, no. 1, pp. 10–17, 2015.

[5] Suharjanti, “Prosiding Seminar Nasional Aplikasi Sains & Teknologi (SNAST) 2014 Yogyakarta, 15 November 2014 ISSN: 1979-911X,” *Snast*, no. November, pp. 211–216, 2014.

[6] R. Yanto and R. Khoiriah, “Implementasi Data Mining dengan Metode Algoritma Apriori dalam Menentukan Pola Pembelian Obat,” *Creat. Inf. Technol. J.*, vol. 2, no. 2, p. 102, 2015, doi: 10.24076/citec.2015v2i2.41.

[7] D. P. Utomo and M. Mesran, “Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung,” *J. Media Inform. Budidarma*, vol. 4, no. 2, p. 437, 2020, doi: 10.30865/mib.v4i2.2080.

[8] Y. Yuliana, P. Paradise, and K. Kusrini, “Sistem Pakar Diagnosa Penyakit Ispa Menggunakan Metode Naive Bayes Classifier Berbasis Web,” *CSRID (Computer Sci. Res. Its Dev. Journal)*, vol. 10, no. 3, p. 127, 2021, doi: 10.22303/csrid.10.3.2018.127-138.

[9] A. Supriyatna and W. P. Mustika, “Komparasi Algoritma Naive bayes dan SVM Untuk Memprediksi Keberhasilan Imunoterapi Pada Penyakit Kutil,” *J-SAKTI (Jurnal Sains Komput. dan Inform.*, vol. 2, no. 2, p. 152, 2018, doi: 10.30645/j-sakti.v2i2.78.

[10] A. Damuri, U. Riyanto, H. Rusdianto, and M. Aminudin, “Implementasi Data Mining dengan Algoritma Naïve Bayes Untuk Klasifikasi Kelayakan Penerima Bantuan Sembako,” *J. Ris. Komput.*, vol. 8, no. 6, pp. 219–225, 2021, doi: 10.30865/jurikom.v8i6.3655.

[11] Y. D. Darmi and A. Setiawan, “Penerapan Metode Clustering K-Means Dalam Pengelompokan Penjualan Produk,” *J. Media Infotama*, vol. 12, no. 2, pp. 148–157, 2017, doi: 10.37676/jmi.v12i2.418.

[12] . F., F. T. Kesuma, and S. P. Tamba, “Penerapan Data Mining Untuk Menentukan Penjualan Sparepart Toyota Dengan Metode K-Means Clustering,” *J. Sist. Inf. dan Ilmu Komput. Prima(JUSIKOM PRIMA)*, vol. 2, no. 2, pp. 67–72, 2020, doi: 10.34012/jusikom.v2i2.376.

[13] F. Nasari, C. Jhony, and M. Sianturi, “Penerapan Algoritma K-Means Clustering…,” pp. 108–119.

[14] S. Agustina, D. Yhudo, H. Santoso, and ..., “Clustering Kualitas Beras Berdasarkan Ciri Fisik Menggunakan Metode K-Means,” *Univ. Brawijaya …*, 2012, [Online]. Available: https://www.academia.edu/download/46692771/clustering-kualitas-beras-dengan-k-means.pdf

[15] A. Aditya, I. Jovian, and B. N. Sari, “Implementasi K-Means Clustering Ujian Nasional Sekolah Menengah Pertama di Indonesia Tahun 2018/2019,” *J. Media Inform. Budidarma*, vol. 4, no. 1, p. 51, 2020, doi: 10.30865/mib.v4i1.1784.

[16] A. Sani, “PENERAPAN METODE K-MEANS Related papers,” *J. Teknol.*, vol. 1, pp. 1–7, 2014.

[17] T. Mardiana and R. D. Nyoto, “Kluster Bag of Word Menggunakan Weka,” *J. Edukasi dan Penelit. Inform.*, vol. 1, no. 1, pp. 1–5, 2015, doi: 10.26418/jp.v1i1.10145.

[18] A. Pangestu, “PENERAPAN DATA MINING MENGGUNAKAN ALGORITMA K- MEANS PENGELOMPOKAN PELANGGAN BERDASARKAN KUBIKASI AIR TERJUAL MENGGUNAKAN WEKA,” vol. 11, no. 3, pp. 67–71, 2021.

[19] S. N. Arofah and F. Marisa, “Penerapan Data Mining untuk Mengetahui Minat Siswa pada Pelajaran Matematika menggunakan Metode K-Means Clustering,” *JOINTECS (Journal Inf. Technol. Comput. Sci.*, vol. 3, no. 2, pp. 85–90, 2018, doi: 10.31328/jointecs.v3i2.787.

LAMPIRAN

Informasi Atribut

- jenis kelamin: "Pria", "Wanita" atau "Lainnya".
- usia: usia pasien.
- hipertensi: 0 jika pasien tidak memiliki hipertensi, 1 jika pasien memiliki hipertensi.
- penyakit jantung: 0 jika pasien tidak memiliki penyakit jantung, 1 jika pasien memiliki penyakit jantung .
- pernah kawin: "Tidak" atau "Ya".
- tipe kerja: "anak-anak", "Pemerintah", "Belum pernah bekerja", "Swasta" atau "Wiraswasta".
- Tipe tempat tinggal: "Pedesaan" atau "Perkotaan".
- tingkat avgglukosa: kadar glukosa rata-rata dalam darah.
- bmi: indeks massa tubuh.
- smoking_status: "sebelumnya merokok", "tidak pernah merokok", "merokok" atau "Tidak diketahui".
- stroke: 1 jika pasien mengalami stroke atau 0 jika tidak.

Catatan : "Tidak diketahui" pada status_rokok berarti informasi tersebut tidak tersedia untuk pasien ini

Berikut link data Stroke otak :

https://docs.google.com/spreadsheets/d/13NZChjfFgH1XeEjrKnJG-wiakXPWU7IQ/edit?usp=drivesdk&ouid=110175368376829491277&rtpof=true&sd=true

No	Gender	Age	Hypertension	Heart Disease	Ever-Married	Worktype	Residencetype	Avgglucoselevel	bmi	Smoking_Status	Stroke
1	Male	67.0	0.0	1.0	Yes	Private	Urban	228.69	36.6	formerly smoked	1.0
2	Male	80.0	0.0	1.0	Yes	Private	Rural	105.92	32.5	never smoked	1.0
3	Female	49.0	0.0	0.0	Yes	Private	Urban	171.23	34.4	smokes	1.0
4	Female	79.0	1.0	0.0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1.0
5	Male	81.0	0.0	0.0	Yes	Private	Urban	186.21	29.0	formerly smoked	1.0
6	Male	74.0	1.0	1.0	Yes	Private	Rural	70.09	27.4	never smoked	1.0
7	Female	69.0	0.0	0.0	No	Private	Urban	94.39	22.8	never smoked	1.0
8	Female	78.0	0.0	0.0	Yes	Private	Urban	58.57	24.2	Unknown	1.0
9	Female	81.0	1.0	0.0	Yes	Private	Rural	80.43	29.7	never smoked	1.0
10	Female	61.0	0.0	1.0	Yes	Govt_job	Rural	120.46	36.8	smokes	1.0
11	Female	54.0	0.0	0.0	Yes	Private	Urban	104.51	27.3	smokes	1.0
12	Female	79.0	0.0	1.0	Yes	Private	Urban	214.09	28.2	never smoked	1.0
13	Female	50.0	1.0	0.0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1.0
14	Male	64.0	0.0	1.0	Yes	Private	Urban	191.61	37.5	smokes	1.0
15	Male	75.0	1.0	0.0	Yes	Private	Urban	221.29	25.8	smokes	1.0
16	Female	60.0	0.0	0.0	No	Private	Urban	89.22	37.8	never smoked	1.0
17	Female	71.0	0.0	0.0	Yes	Govt_job	Rural	193.94	22.4	smokes	1.0
18	Female	52.0	1.0	0.0	Yes	Self-employed	Urban	233.29	48.9	never smoked	1.0
19	Female	79.0	0.0	0.0	Yes	Self-employed	Urban	228.7	26.6	never smoked	1.0
20	Male	82.0	0.0	1.0	Yes	Private	Rural	208.3	32.5	Unknown	1.0
21	Male	71.0	0.0	0.0	Yes	Private	Urban	102.87	27.2	formerly smoked	1.0
22	Male	80.0	0.0	0.0	Yes	Self-employed	Rural	104.12	23.5	never smoked	1.0
23	Female	65.0	0.0	0.0	Yes	Private	Rural	100.98	28.2	formerly smoked	1.0

4958	Female	82.0	1.0	0.0	Yes	Private	Urban	222.52	31.8	formerly smoked	0.0
4959	Male	32.0	1.0	0.0	No	Private	Rural	74.43	31.5	Unknown	0.0
4960	Female	17.0	0.0	0.0	No	Private	Urban	92.97	26.5	formerly smoked	0.0
4961	Female	18.0	0.0	0.0	No	Private	Rural	101.12	26.4	smokes	0.0
4962	Male	59.0	1.0	0.0	Yes	Govt_job	Rural	253.93	32.1	formerly smoked	0.0
4963	Male	3.0	0.0	0.0	No	children	Rural	194.75	20.1	Unknown	0.0
4964	Female	20.0	0.0	0.0	No	Govt_job	Rural	79.53	26.5	never smoked	0.0
4965	Female	78.0	0.0	0.0	Yes	Govt_job	Urban	101.76	27.3	smokes	0.0
4966	Male	52.0	1.0	0.0	Yes	Govt_job	Rural	116.62	31.7	smokes	0.0
4967	Female	65.0	0.0	1.0	Yes	Private	Rural	57.52	29.4	formerly smoked	0.0
4968	Male	59.0	0.0	0.0	Yes	Private	Urban	223.16	33.2	Unknown	0.0
4969	Female	78.0	1.0	1.0	Yes	Private	Rural	206.53	31.2	never smoked	0.0
4970	Female	70.0	0.0	1.0	Yes	Self-employed	Rural	65.68	28.6	Unknown	0.0
4971	Female	70.0	0.0	1.0	Yes	Self-employed	Urban	240.69	30.9	smokes	0.0
4972	Male	37.0	0.0	0.0	Yes	Private	Rural	107.06	29.7	smokes	0.0
4973	Male	72.0	0.0	1.0	Yes	Private	Rural	238.27	30.7	smokes	0.0
4974	Male	1.0	0.0	0.0	No	children	Rural	107.02	18.8	Unknown	0.0
4975	Male	58.0	0.0	0.0	Yes	Govt_job	Urban	84.94	30.2	never smoked	0.0
4976	Male	31.0	0.0	0.0	No	Private	Urban	215.07	32.7	smokes	0.0
4977	Male	41.0	0.0	0.0	No	Private	Rural	70.15	29.8	formerly smoked	0.0
4978	Male	40.0	0.0	0.0	Yes	Private	Urban	191.15	31.1	smokes	0.0
4979	Female	45.0	1.0	0.0	Yes	Govt_job	Rural	95.02	31.8	smokes	0.0
4980	Male	40.0	0.0	0.0	Yes	Private	Rural	83.94	30.0	smokes	0.0
4981	Female	80.0	1.0	0.0	Yes	Private	Urban	83.75	29.1	never smoked	0.0