

SKRIPSI

**KLASIFIKASI PENDERITA STROKE BERDASARKAN
ANALISIS DATA SURVEI BRFS**



Yohanes Irfon Haryanto

NPM: 6181801006

**PROGRAM STUDI TEKNIK INFORMATIKA
FAKULTAS TEKNOLOGI INFORMASI DAN SAINS
UNIVERSITAS KATOLIK PARAHYANGAN
2022**

UNDERGRADUATE THESIS

**CLASSIFICATION OF STROKE PATIENTS BASED ON BRFSS
SURVEY DATA ANALYSIS**



Yohanes Irfon Haryanto

NPM: 6181801006

**DEPARTMENT OF INFORMATICS
FACULTY OF INFORMATION TECHNOLOGY AND SCIENCES
PARAHYANGAN CATHOLIC UNIVERSITY
2022**

LEMBAR PENGESAHAN

KLASIFIKASI PENDERITA STROKE BERDASARKAN ANALISIS DATA SURVEI BRFSS

Yohanes Irfon Haryanto

NPM: 6181801006

Bandung, 13 Januari 2022

Menyetujui,

Pembimbing

Digitally signed
by Luciana
Abednego

Luciana Abednego, M.T.

Ketua Tim Penguji

Digitally signed
by Maria V.
Claudia M.

Maria Veronica, M.T.

Anggota Tim Penguji

Digitally signed
by Husnul
Hakim

Husnul Hakim, M.T.

Mengetahui,

Ketua Program Studi

Digitally signed
by Mariskha Tri
Adithia

Mariskha Tri Adithia, P.D.Eng

PERNYATAAN

Dengan ini saya yang bertandatangan di bawah ini menyatakan bahwa skripsi dengan judul:

KLASIFIKASI PENDERITA STROKE BERDASARKAN ANALISIS DATA SURVEI BRFSS

adalah benar-benar karya saya sendiri, dan saya tidak melakukan penjiplakan atau pengutipan dengan cara-cara yang tidak sesuai dengan etika keilmuan yang berlaku dalam masyarakat keilmuan.

Atas pernyataan ini, saya siap menanggung segala risiko dan sanksi yang dijatuhkan kepada saya, apabila di kemudian hari ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya, atau jika ada tuntutan formal atau non-formal dari pihak lain berkaitan dengan keaslian karya saya ini.

Dinyatakan di Bandung,
Tanggal 13 Januari 2022



Yohanes Irfon Haryanto
NPM: 6181801006

ABSTRAK

Stroke adalah penyakit pembuluh darah otak yang ditandai dengan gangguan fungsi otak karena adanya kerusakan atau kematian jaringan otak akibat berkurang atau tersumbatnya aliran darah dan oksigen ke otak. Terdapat beberapa faktor yang berhubungan atau dapat menyebabkan stroke. Data BRFSS adalah salah satu sarana yang dapat digunakan untuk menganalisis dan mencari faktor-faktor yang berhubungan dengan stroke. BRFSS atau *Behavioral Risk Factor Surveillance System* adalah sebuah proyek yang didirikan oleh *Centers for Disease Control and Prevention* (CDC) pada tahun 1984. BRFSS merupakan sistem survei kesehatan utama negara yang mengumpulkan data negara bagian tentang penduduk AS mengenai perilaku dan kejadian berisiko terkait kesehatan mereka, kondisi kesehatan kronis, dan penggunaan layanan pencegahan.

Penelitian ini bertujuan untuk menganalisis data jawaban survei BRFSS tahun 2020 untuk mencari atribut-atribut yang berhubungan dengan stroke. Atribut yang berhubungan akan digunakan untuk proses klasifikasi penderita stroke. Eksplorasi data dilakukan agar data lebih mudah diolah dan membantu untuk menentukan teknik analisis yang tepat untuk mencapai tujuan. Setelah data sudah dieksplorasi, analisis dilakukan untuk mencari atribut-atribut yang berhubungan dengan stroke.

Metode yang digunakan untuk analisis adalah visualisasi menggunakan *bar chart* untuk melihat hubungan antara atribut stroke dengan atribut lain, dan menghitung *Chi Square* serta *Information Gain* untuk mengukur korelasi antara atribut stroke dengan atribut lain. Dari hasil analisis terdapat 21 atribut yang mempengaruhi penderita stroke yaitu kesulitan berjalan, serangan jantung, kondisi kesehatan, penyakit jantung koroner, kesulitan melakukan tugas sendirian, kesulitan berpakaian, usia, penyakit paru-paru, penyakit ginjal, kesulitan berkonsentrasi, radang sendi, diabetes, kanker lain, kanker kulit, asma, depresi, peminum berat, status merokok, pendapatan, ras, dan kategori BMI. Atribut-atribut ini digunakan sebagai fitur untuk membuat model klasifikasi untuk memprediksi penderita stroke.

Model klasifikasi yang dibangun menggunakan algoritma *Decision Tree*, *Naïve Bayes*, dan *Random Forest*. Model terbaik setiap algoritma dipilih dengan melihat nilai akurasi, presisi, *recall*, dan *f1-score*. Nilai yang diutamakan adalah nilai *recall*, karena dalam penelitian ini akan sangat besar risikonya jika model menyatakan penderita stroke sebagai orang yang sehat. Hasil penelitian menunjukkan bahwa model terbaik untuk algoritma *Decision Tree* menggunakan 7 atribut dengan nilai akurasi 75%, presisi 74%, *recall* 79%, dan *f1-score* 76%. Model terbaik untuk algoritma *Categorical Naïve Bayes* menggunakan 20 atribut dengan nilai akurasi 74%, presisi 76%, *recall* 69%, dan *f1-score* 72%. Model terbaik untuk algoritma *Random Forest* menggunakan 7 atribut dengan nilai akurasi 75%, presisi 73%, *recall* 79%, dan *f1-score* 76%.

Model-model terbaik diimplementasikan ke dalam perangkat lunak dan diuji fungsionalitasnya untuk memprediksi penderita stroke. Pengujian dilakukan dengan menjawab pertanyaan-pertanyaan yang ditampilkan oleh perangkat lunak. Dari hasil pengujian, perangkat lunak berhasil memprediksi penderita stroke dan bukan berdasarkan pertanyaan-pertanyaan yang dijawab oleh pengguna.

Kata-kata kunci: stroke, klasifikasi, atribut, hubungan, evaluasi model

ABSTRACT

Stroke is a disease of the blood vessels of the brain characterized by impaired brain function due to the presence of damage or death of brain tissue due to reduced or blocked blood and oxygen flow to the brain. There are several related factors related or can cause stroke. Data BRFSS is a means that can be used to analyze and look for factors related to stroke. BRFSS or Behavioral Risk Factor Surveillance System is a project established by the Centers for Disease Control and Prevention (CDC) in 1984. BRFSS is the nation's primary health survey system that collects state data on U.S. residents regarding their health-related risky behaviors and events, chronic health conditions, and use of preventive services. One of the questions in the survey relates to stroke.

This study aims to analyze the BRFSS survey answer data in 2020 to find attributes related to stroke. Related attributes will be used for the classification process of stroke sufferers. Data exploration is done so that data is more easily processed and helps to determine the right analysis techniques to achieve goals. After the data has been explored, the analysis is carried out to find attributes related to stroke.

The method used for analysis is a visualization using a bar chart to see the relationship between stroke attributes and other attributes, and calculate Chi Square and information gain to measure the correlation between stroke attributes and other attributes. From the results of the analysis there are 21 attributes that affect stroke sufferers, namely difficulty walking, heart attacks, health conditions, coronary heart disease, difficulty doing tasks alone, difficulty dressing, age, lung disease, kidney disease, difficulty concentrating, arthritis, diabetes, Other cancer, skin cancer, asthma, depression, heavy drinkers, smoking status, income, race, and BMI categories. These attributes are used as features to create a classification model to predict stroke sufferers.

The classification model built using the Decision Tree, Naïve Bayes, and Random Forest algorithm. The best models of each algorithm are seen from the accuracy, precision, recall, and F1-score values. But the value that takes precedence is the recall value, because in this study there would be a very large risk if the model declared stroke sufferers as healthy people. The results showed that the best model for the Decision Tree algorithm uses 7 attributes with an accuracy value of 75%, precision 74%, recall 79%, and f1-score 76%. The best models for the Categorical Naïve Bayes algorithm use 20 attributes with an accuracy value of 74%, precision 76%, recall 69%, and f1-score 72%. The best models for Random Forest algorithms use 7 attributes with an accuracy value of 75%, precision 73%, recall 79%, and f1-score 76%.

The best models are implemented into software and tested their functionality to predict stroke sufferers. Testing is done by answering the questions displayed by the software. From the test results, the software succeeded in predicting stroke sufferers and not based on the questions answered by the user.

Keywords: stroke, classification, attribute, relationship, model evaluation

KATA PENGANTAR

Puji, hormat, dan syukur kepada Tuhan yang Maha Esa atas berkat dan pendampingan-Nya, penulis dapat menyelesaikan skripsi yang berjudul “Klasifikasi Penderita Stroke Berdasarkan Analisis Data Survei BRFSS”. Skripsi ini disusun untuk memenuhi syarat kelulusan di program studi sarjana Informatika UNPAR.

Penulis menyadari bahwa dalam pengerjaan skripsi ini penulis masih mengalami kesulitan dan hambatan. Akan tetapi, banyak pihak telah membantu penulis dalam hal bimbingan, dukungan, doa, dan motivasi. Penulis mengucapkan terima kasih kepada seluruh pihak yang telah turut ambil bagian dalam proses pengerjaan skripsi ini. Secara khusus penulis mengucapkan terima kasih kepada:

- Keluarga penulis yang telah memberikan dorongan, dukungan, dan doa dalam pengerjaan skripsi ini.
- Ibu Natalia, M.Si., selaku dosen pembimbing yang telah meluangkan waktu dan tenaga untuk mendampingi penulis dengan penuh kesabaran.
- Ibu Maria dan Bapak Husnul selaku penguji yang telah memberikan kritik, saran, serta masukan untuk skripsi ini.
- Hilsong dan Noah yang sudah memberikan semangat selama menyusun skripsi.
- Syahdan dan Cevas selaku teman penulis yang sudah menyediakan tempat yang baik selama menyusun skripsi ini.
- Soetomo yang sudah membantu menyelesaikan GOW dan teman-teman moes lain yang tidak dapat penulis sebutkan satu per satu yang sudah memberikan sedikit semangat dalam mengerjakan skripsi ini.

Semoga skripsi ini dapat bermanfaat bagi pembaca dan menginspirasi untuk penelitianpenelitian berikutnya.

Bandung, Januari 2022

Penulis

DAFTAR ISI

KATA PENGANTAR	xv
DAFTAR ISI	xvii
DAFTAR GAMBAR	xix
1 PENDAHULUAN	1
1.1 Latar Belakang	1
1.2 Rumusan Masalah	2
1.3 Tujuan	2
1.4 Batasan Masalah	2
1.5 Metodologi	2
1.6 Sistematika Pembahasan	3
2 LANDASAN TEORI	5
2.1 BRFS	5
2.2 Stroke	5
2.3 Eksplorasi Data	6
2.3.1 Tipe Atribut	6
2.3.2 <i>Missing Value</i>	7
2.4 Seleksi Fitur	7
2.4.1 <i>Chi Square</i>	7
2.4.2 <i>Information Gain</i>	9
2.5 Teknik Visualisasi	10
2.5.1 <i>Bar Chart</i>	10
2.5.2 <i>Pie Chart</i>	11
2.6 Klasifikasi	11
2.6.1 <i>Decision Tree</i>	13
2.6.2 <i>Naïve Bayes</i>	15
2.6.3 <i>Random Forest</i>	17
2.6.4 Evaluasi Model	17
2.6.5 Ketidakseimbangan Kelas Data	18
2.7 <i>Tools</i> dan <i>Library</i>	19
3 ANALISIS	29
3.1 Dataset BRFS	29
3.2 Eksplorasi Dataset	29
3.3 Pemilihan Fitur dan Pembersihan Dataset	39
3.3.1 Pemilihan Fitur	39
3.3.2 Pembersihan Dataset	51
4 EKSPERIMEN	53
4.1 Eksperimen 1	54

4.2	Eksperimen 2	54
4.3	Eksperimen 3	55
4.4	Eksperimen 4	57
4.5	Eksperimen 5	58
4.6	Eksperimen 6	59
4.7	Pemilihan Model Terbaik	62
5	PERANGKAT LUNAK	67
5.1	Rancangan Perangkat Lunak	67
5.2	Hasil Perangkat Lunak	68
5.3	Pengujian Perangkat Lunak	71
6	KESIMPULAN DAN SARAN	73
6.1	Kesimpulan	73
6.2	Saran	73
	DAFTAR REFERENSI	75
	A KODE PROGRAM	77
	B HASIL EKSPERIMEN	121

DAFTAR GAMBAR

2.1	Contoh <i>Distinct Value</i>	8
2.2	Tabel Nilai Distribusi <i>Chi Square</i>	9
2.3	<i>Bar Chart</i> untuk Tabel <i>US Education Rating</i>	11
2.4	<i>Pie Chart</i> untuk Tabel <i>US Education Rating</i>	11
2.5	Gambaran Metode <i>Ensemble</i>	12
2.6	Tahapan Pertama Klasifikasi	12
2.7	Tahapan Kedua Klasifikasi	13
2.8	<i>Decision Tree</i> yang Menunjukkan Kemungkinan Pelanggan <i>AllElectronic</i> Membeli Komputer	13
2.9	Contoh Partisi dalam <i>Decision Tree</i>	14
2.10	Objek pada daerah R merupakan contoh <i>outlier</i>	14
2.11	<i>Decision Tree</i> yang Menunjukkan Kemungkinan Pelanggan <i>AllElectronic</i> Membeli Komputer	15
2.12	Aturan <i>IF-THEN</i> dari <i>Decision Tree</i> Kemungkinan Pelanggan <i>AllElectronic</i> Membeli Komputer	15
2.13	<i>Confusion Matrix</i> untuk <i>Tuple Positif dan Negatif</i>	17
2.14	Contoh Cara Kerja <i>SMOTE</i>	19
3.1	Contoh Pertanyaan dengan <i>Section Name Alcohol Consumption</i>	30
3.2	Contoh Pertanyaan Datanya Tidak Ditampilkan	30
3.3	Proporsi Penderita Stroke	30
3.4	Proporsi Jenis Kelamin	31
3.5	Proporsi Status Merokok	31
3.6	Proporsi Penderita Penyakit Jantung Koroner	31
3.7	Proporsi Penderita Serangan Jantung	32
3.8	Proporsi Penderita Diabetes	32
3.9	Proporsi Kategori BMI	32
3.10	Proporsi Kondisi Kesehatan	33
3.11	Proporsi Peminum Berat	33
3.12	Proporsi Umur	33
3.13	Proporsi Penderita Asma	34
3.14	Proporsi Penderita Kanker Kulit	34
3.15	Proporsi Pendapatan	34
3.16	Proporsi Ras	35
3.17	Proporsi Penderita Kanker Lain	35
3.18	Proporsi Penderita Penyakit Paru-Paru	35
3.19	Proporsi Penderita Penyakit Ginjal	36
3.20	Proporsi Penderita Depresi	36
3.21	Proporsi Penderita Penyakit Radang Sendi	36
3.22	Proporsi Penderita Tunanetra	37
3.23	Proporsi Penderita Tunarungu	37
3.24	Proporsi Pengguna Asuransi Kesehatan	37

3.25	Proporsi Orang yang Kesulitan Berkonsentrasi	38
3.26	Proporsi Orang yang Kesulitan Berjalan	38
3.27	Proporsi Orang yang Kesulitan Berpakaian	38
3.28	Proporsi Orang yang Kesulitan Melakukan Tugas Sendirian	39
3.29	Sebaran Penderita Serangan Jantung dengan Kelas Stroke	39
3.30	Sebaran Penderita Penyakit Jantung Koroner dengan Kelas Stroke	40
3.31	Sebaran Gender dengan Kelas Stroke	40
3.32	Sebaran Penderita yang Kesulitan Berpakaian dengan Kelas Stroke	41
3.33	Sebaran Penderita yang Kesulitan Berjalan dengan Kelas Stroke	41
3.34	Sebaran Penderita yang Kesulitan Melakukan Tugas dengan Kelas Stroke	42
3.35	Sebaran Penderita Penyakit Ginjal dengan Kelas Stroke	42
3.36	Sebaran Penderita Penyakit Paru-Paru dengan Kelas Stroke	43
3.37	Sebaran Penderita yang Kesulitan Berkonsentrasi dengan Kelas Stroke	43
3.38	Sebaran Penderita Kanker Lain dengan Kelas Stroke	44
3.39	Sebaran Penderita Kanker Kulit dengan Kelas Stroke	44
3.40	Sebaran Penderita Diabetes dengan Kelas Stroke	45
3.41	Sebaran Penderita Asma dengan Kelas Stroke	45
3.42	Sebaran Peminum Berat dengan Kelas Stroke	46
3.43	Sebaran Penderita Depresi dengan Kelas Stroke	46
3.44	Sebaran Penderita Radang Sendi dengan Kelas Stroke	47
3.45	Sebaran Kondisi Kesehatan dengan Kelas Stroke	47
3.46	Sebaran Status Merokok dengan Kelas Stroke	48
3.47	Sebaran Rentang Usia dengan Kelas Stroke	48
3.48	Sebaran Kategori BMI dengan Kelas Stroke	49
3.49	Sebaran Pendapatan dengan Kelas Stroke	50
3.50	Sebaran Ras dengan Kelas Stroke	50
4.1	<i>Confusion Matrix</i> untuk Model dengan Algoritma <i>Decision Tree</i>	63
4.2	Contoh Hasil <i>Decision Tree</i> dengan Kedalaman Pohon = 2	63
4.3	<i>Confusion Matrix</i> untuk Model dengan Algoritma <i>Categorical Naïve Bayes</i>	64
4.4	<i>Confusion Matrix</i> untuk Model dengan Algoritma <i>Random Forest</i>	64
4.5	Salah Satu Pohon Dari Model <i>Random Forest</i> dengan Kedalaman Pohon = 2	65
5.1	Tampilan Halaman Depan Perangkat Lunak	67
5.2	Tampilan Halaman <i>Decision Tree</i> , <i>Naive Bayes</i> , dan <i>Random Forest</i>	68
5.3	Tampilan Saat Perangkat Lunak Berhasil Menebak Pengguna	68
5.4	Tampilan Halaman Depan	69
5.5	Contoh Salah Satu Tampilan <i>Form Page</i>	69
5.6	Tampilan Saat Ada Pertanyaan yang Tidak Diisi	70
5.7	Tampilan Saat Pengguna Diprediksi Terkena Stroke	70
5.8	Tampilan Saat Pengguna Diprediksi Tidak Terkena Stroke	71
B.1	Hasil Prediksi Skenario Pertama Algoritma <i>Decision Tree</i> (Pertanyaan 1–3)	121
B.2	Hasil Prediksi Skenario Pertama Algoritma <i>Decision Tree</i> (Pertanyaan 4–5)	122
B.3	Hasil Prediksi Skenario Pertama Algoritma <i>Decision Tree</i> (Pertanyaan 6–7)	122
B.4	Hasil Prediksi Skenario Kedua Algoritma <i>Decision Tree</i> (Pertanyaan 1–3)	123
B.5	Hasil Prediksi Skenario Kedua Algoritma <i>Decision Tree</i> (Pertanyaan 4–5)	123
B.6	Hasil Prediksi Skenario Kedua Algoritma <i>Decision Tree</i> (Pertanyaan 6–7)	124
B.7	Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 1–2)	124
B.8	Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 3–4)	125
B.9	Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 5–8)	125
B.10	Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 9–12)	126

B.11 Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 13–16)	126
B.12 Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 17–19)	127
B.13 Hasil Prediksi Skenario Pertama Algoritma <i>Naive Bayes</i> (Pertanyaan 19–20)	127
B.14 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 1–3)	128
B.15 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 4–5)	128
B.16 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 6–9)	129
B.17 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 10–13)	129
B.18 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 14–17)	130
B.19 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 18–19)	130
B.20 Hasil Prediksi Skenario Kedua Algoritma <i>Naive Bayes</i> (Pertanyaan 19–20)	131
B.21 Hasil Prediksi Skenario Pertama Algoritma <i>Random Forest</i> (Pertanyaan 1–2)	131
B.22 Hasil Prediksi Skenario Pertama Algoritma <i>Random Forest</i> (Pertanyaan 3–4)	132
B.23 Hasil Prediksi Skenario Pertama Algoritma <i>Random Forest</i> (Pertanyaan 5–7)	132
B.24 Hasil Prediksi Skenario Kedua Algoritma <i>Random Forest</i> (Pertanyaan 1–2)	133
B.25 Hasil Prediksi Skenario Kedua Algoritma <i>Random Forest</i> (Pertanyaan 3–4)	133
B.26 Hasil Prediksi Skenario Kedua Algoritma <i>Random Forest</i> (Pertanyaan 5–7)	134

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Stroke adalah kondisi medis gawat darurat yang terjadi akibat terganggunya aliran darah menuju otak. Berdasarkan *website* CDC¹ terdapat tiga jenis stroke, yaitu stroke iskemik, stroke hemoragik, dan *transient ischemic attack* (TIA). Stroke iskemik terjadi ketika gumpalan darah menyumbat aliran darah yang menuju otak. Gumpalan darah ini paling sering disebabkan oleh penumpukan lemak pada dinding pembuluh darah. Sebagian besar stroke (87%) merupakan stroke iskemik. Stroke hemoragik terjadi ketika pembuluh darah otak pecah dan darah akan keluar ke jaringan di sekitarnya. *Transient ischemic attack* disebabkan oleh penyumbatan pembuluh darah di otak, yang bersifat sementara. Kondisi ini juga lebih dikenal dengan istilah stroke ringan. Ada beberapa faktor umum utama yang dapat menyebabkan stroke, yaitu usia, jenis kelamin, tingginya tekanan darah, merokok atau tidak, memiliki penyakit jantung atau tidak, memiliki kolestrol atau tidak, obesitas, memiliki diabetes atau tidak, dan genetik. Dari faktor-faktor umum tersebut, dapat diprediksi apakah seseorang terkena stroke atau tidak. Data BRFSS adalah salah satu sarana yang dapat digunakan untuk menganalisis dan mencari faktor-faktor yang berhubungan dengan stroke.

BRFSS² (*Behavioral Risk Factor Surveillance System*) adalah sebuah proyek yang didirikan oleh *Centers for Disease Control and Prevention* (CDC) pada tahun 1984 yang merupakan survei telepon menggunakan teknik *Random Digit Dialing* (RDD) terkait kesehatan negara yang mengumpulkan data negara bagian tentang penduduk AS yang berusia 18 tahun ke atas mengenai perilaku berisiko terkait kesehatan mereka, kondisi kesehatan kronis, dan penggunaan layanan kesehatan preventif terkait dengan penyebab utama kematian dan kecacatan di Amerika Serikat. BRFSS mengumpulkan data di seluruh 50 negara bagian AS serta Distrik Columbia dan tiga negara bagian wilayah AS. Survei BRFSS dilakukan setiap tahun dari tahun 1987–tahun ini. Data BRFSS dieksplorasi dan diolah agar dapat mempermudah proses analisis.

Eksplorasi adalah teknik yang digunakan untuk memahami data dengan cara menganalisis dan memproses data tersebut menjadi suatu informasi yang berguna seperti mengetahui tipe atribut dan mendeteksi *missing value*. Dengan adanya eksplorasi, data lebih mudah diolah dan membantu untuk menentukan teknik analisis yang tepat untuk mencapai tujuan yaitu menemukan faktor-faktor yang berhubungan dengan stroke.

Visualisasi dapat digunakan untuk membantu proses analisis. Bentuk visualisasi seperti *pie chart* dan *bar chart* dapat digunakan untuk melihat sebaran, proporsi, serta hubungan antara dua atribut yang berbeda. Untuk menentukan hubungan antara dua atribut, menggunakan visualisasi saja tidak cukup sehingga perlu menggunakan metode lain yaitu *Chi Square* dan *Information Gain*.

Chi Square adalah metode yang dapat digunakan untuk mengukur hubungan antara dua atribut kategorik. *Information Gain* adalah nilai informasi (bobot) dari suatu atribut. Kedua metode ini digunakan untuk menentukan atribut-atribut yang berhubungan dengan stroke. Semakin tinggi nilai *Chi Square* dan *Information Gain* yang didapat, maka atribut tersebut semakin berhubungan dengan stroke. Atribut-atribut yang berhubungan dengan stroke dipilih menjadi fitur untuk memprediksi

¹https://www.cdc.gov/stroke/types_of_stroke.htm (diakses 27 Maret 2022)

²https://www.cdc.gov/brfss/about/brfss_faq.htm (diakses 26 Maret 2022)

penderita stroke menggunakan teknik klasifikasi.

Klasifikasi adalah proses menemukan model yang menjelaskan dan membedakan kelas data. Klasifikasi bertujuan untuk memprediksi label kelas kategorikal. Dalam klasifikasi, dibutuhkan fitur atau atribut yang akan digunakan untuk menjadi ciri-ciri dasar dalam menggambarkan suatu objek. Dari fitur-fitur ini dapat dibuat suatu model yang dapat memprediksi suatu label kelas. Pada skripsi ini, klasifikasi digunakan untuk memprediksi apakah seseorang terkena stroke atau tidak. Ada beberapa algoritma yang dapat digunakan untuk membuat model klasifikasi seperti *Decision Tree*, *Naïve Bayes*, *Random Forest*, *Support Vector Machine*, dan lain-lain.

Beberapa *tools* dan *library* digunakan untuk mengolah dataset dan membangun model klasifikasi. *Pandas* dan *NumPy* digunakan untuk mengolah dataset agar dapat digunakan, sedangkan *Matplotlib*, *SciKit-Learn*, dan *Imbalanced Learn* digunakan untuk proses analisis dan pembangunan model klasifikasi.

Setelah membuat model klasifikasi, dilakukan evaluasi untuk melihat kinerja setiap model dengan melihat nilai metrik evaluasi yaitu akurasi, presisi, dan *recall*. Nilai evaluasi yang diutamakan adalah nilai *recall*, karena dalam penelitian ini, akan sangat besar risikonya jika model menyatakan penderita stroke sebagai orang yang sehat. Model terbaik dengan nilai *recall* tertinggi diimplementasikan ke dalam perangkat lunak dan diuji fungsionalitasnya untuk memprediksi penderita stroke. Pengujian perangkat lunak dilakukan dengan mengisi pertanyaan-pertanyaan yang ditampilkan oleh perangkat lunak, kemudian perangkat lunak akan memprediksi apakah seseorang merupakan penderita stroke atau bukan berdasarkan pertanyaan-pertanyaan yang sudah dijawab.

1.2 Rumusan Masalah

Rumusan masalah untuk skripsi ini adalah:

1. Bagaimana menentukan faktor-faktor yang berhubungan dengan stroke berdasarkan data pada survei BRFSS?
2. Bagaimana menentukan algoritma klasifikasi yang tepat untuk mengolah fitur dari dataset BRFSS?
3. Bagaimana melakukan evaluasi model untuk menentukan model terbaik?
4. Bagaimana membangun perangkat lunak untuk memprediksi penderita stroke?

1.3 Tujuan

Berhubungan dengan rumusan masalah, tujuan skripsi ini adalah:

1. Menentukan faktor-faktor yang berhubungan dengan stroke berdasarkan data pada survei BRFSS.
2. Menentukan algoritma klasifikasi yang tepat untuk mengolah fitur dari dataset BRFSS.
3. Melakukan evaluasi model untuk menentukan model terbaik.
4. Membangun perangkat lunak untuk memprediksi penderita stroke.

1.4 Batasan Masalah

Batasan masalah pada penelitian ini adalah dataset yang digunakan hanya dataset survei BRFSS tahun 2020.

1.5 Metodologi

Berikut adalah metodologi yang dilakukan dalam penelitian ini:

1. Melakukan studi literatur tentang BRFSS.
2. Melakukan studi literatur tentang penyebab dan gejala stroke.

3. Melakukan studi literatur tentang eksplorasi data.
4. Melakukan studi literatur tentang seleksi fitur.
5. Melakukan studi literatur tentang teknik visualisasi.
6. Melakukan studi literatur tentang klasifikasi.
7. Melakukan studi literatur tentang *tools* dan *library* yang digunakan untuk penelitian ini.
8. Melakukan analisis dan pengolahan dataset BRFSS tahun 2020.
9. Mempelajari dan menentukan metode klasifikasi yang paling cocok untuk digunakan.
10. Membuat model dan melakukan eksperimen terhadap model yang sudah dibuat.
11. Mengimplementasikan hasil model klasifikasi terbaik ke dalam perangkat lunak.
12. Menguji perangkat lunak yang sudah dibangun untuk proses prediksi.
13. Menulis dokumentasi hasil analisis dan implementasi dari penelitian yang sudah dilakukan.

1.6 Sistematika Pembahasan

Penelitian ini disusun berdasarkan sistematika pembahasan sebagai berikut:

1. Bab 1 Pendahuluan
Menjelaskan latar belakang, rumusan masalah, tujuan, batasan masalah, metodologi, dan sistematika pembahasan mengapa penelitian ini dilakukan.
2. Bab 2 Landasan teori
Menjelaskan teori-teori yang diperlukan untuk proses analisis dalam penelitian ini.
3. Bab 3 Analisis
Menjelaskan eksplorasi, pengolahan dataset, serta pemilihan fitur yang akan digunakan untuk eksperimen nanti.
4. Bab 4 Eksperimen
Membahas hasil eksperimen dan analisis yang sudah dilakukan menggunakan algoritma klasifikasi.
5. Bab 5 Perangkat lunak
Menjelaskan rancangan beserta hasil pengujian perangkat lunak.
6. Bab 6 Kesimpulan dan Saran
Menjelaskan kesimpulan dari penelitian yang sudah dilakukan dan saran yang dapat dilakukan untuk penelitian berikutnya.