SKRIPSI

KLASIFIKASI PENDERITA STROKE BERDASARKAN ANALISIS DATA SURVEI BRFSS

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

DAFTAR GAMBAR

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Stroke adalah kondisi medis gawat darurat yang terjadi akibat terganggunya aliran darah menuju otak. Berdasarkan *website* CDC[1] 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.

BRFSS[2] (*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

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