Implementasi Algoritma *Logistic Regression* Untuk Klasifikasi Penyakit *Stroke*

Suhliyyah[1], Hanny Hikmayanti Handayani[2], Kiki Ahmad Baihaqi[3]

[1,2,3]Universitas Buana Perjuangan Karawang
Email: [1]if19.suhliyyah@mhs.ubpkarawang.ac.id,
[2]hanny.hikmayanti@ubpkarawang.ac.id, [3]kikiahmad@ubpkarawang.ac.id

Abstrak. *Stroke* menyebabkan kerusakan di bagian otak yang muncul secara mendadak akibat dari hambatan sirkulasi darah non traumatik. Hambatan tersebut dapat menyebabkan gejala seperti kelumpuhan seisi wajah, bicara tidak jelas, bicara tidak lancar, gangguan penglihatan dan perubahan kesadaran, *Stroke* merupakan penyakit yang menjadi penyebab kematian nomor tiga tertinggi di indonesia setelah penyakit kanker dan jantung. Di indonesia, jumlah kasus dan prevalensi *stroke* tidak diketahui secara jelas. Di perkirakan 500.000 orang penduduk mengalami penyakit *stroke* setiap tahun, sekitar 2,5% atau 12.500 orang meninggal dunia dan sisanya menderita cacat ringan. Hampir setiap hari atau sekurang-kurangnya setiap tiga hari, seorang warga negara indonesia, baik tua maupun muda, meninggal dunia karena *stroke.* Penelitian ini dibuat menggunakan metode *Confusion matrix* dan pengujian menggunakan algoritma *Logistic Regression,* penelitian ini menggunakan teknik pengumpulan data dan hasil analisis untuk meningkatkan akurasi, berdasarkan variabel berpengaruh meliputi jenis kelamin, hipertensi, penyakit jantung, kadar glukosa rata-rata, berat badan dan status merokok. Berdasarkan hasil pengumpulan data yang telah dilakukan sebanyak 4981 data diperoleh hasil akurasi sebesar 94%.

Kata kunci: *Algoritma Logistic Regression; Data Mining; Penyakit Stroke.*

1 Pendahuluan

Stroke menyebabkan kerusakan di bagian otak yang muncul secara mendadak akibat dari hambatan sirkulasi darah non traumatik. Hambatan tersebut dapat menyebabkan gejala seperti kelumpuhan seisi wajah, bicara tidak jelas, bicara tidak lancar, gangguan penglihatan dan perubahan kesadaran[1], Menurut *World Health Organization* (WHO) mendefinisikan *stroke* suatu kondisi dimana tanda-tanda klinis yang berkembang dengan cepat berupa defisit neurologik fokal dan global, yang akan memberat dan terjadi selama 24 jam lebih dan dapat mengakibatkan kematian. *Stroke* merupakan penyakit yang menjadi penyebab kematian nomor tiga tertinggi di indonesia setelah penyakit kanker dan jantung. Di indonesia, jumlah kasus dan prevalensi *stroke* tidak diketahui secara jelas. Di perkirakan 500.000 orang penduduk mengalami penyakit *stroke* setiap tahun, sekitar 2,5% atau 12.500 orang meninggal dunia dan sisanya menderita cacat ringan. Hampir setiap hari atau sekurang-kurangnya setiap tiga hari, seorang

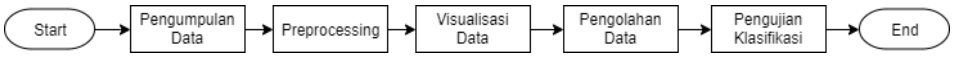
Diterima 01 Mei 2023 Direvisi 12 Mei 2023 Diterima untuk publikasi 30 Mei 2023

warga negara indonesia, baik tua maupun muda, meninggal dunia karena *stroke*[2]. Berdasarkan analisis faktor risiko *stroke* yang dapat dimodifikasi antara lain hipertensi, merokok, dislipidemia, diabetes melitus, obesitas, dan alkohol. Usia, jenis kelamin, dan riwayat keluarga merupakan faktor risiko yang tidak dapat diubah[3].

Penelitian yang dilakukan menggunakan teknik data mining di implementasikan ke dalam sistem untuk melakukan klasifikasi penyakit *stroke* dengan menggunakan metode *Support Vector Machine* (SVM), Hasil dari penelitian tersebut algoritma *Support Vector Machine* (SVM) berhasil melakukan klasifikasi penyakit *stroke* dengan akurasi yang didapatkan sebesar 76%[4]. Penelitian selanjutnya melakukan penelitian klasifikasi penyakit *stroke* dengan membandingkan algoritma *K-Nearest Neighbor* dan *Gaussian Naïve Bayes*, Hasil yang diperoleh dari penelitian tersebut pada algoritma *K-Nearest Neighbor* didapatkan dengan hasil akurasi sebesar 68,30% sedangkan pada algoritma *Gaussian Naïve Bayes* berhasil mendapatkan akurasi sebesar 74,45%[5].

2 Metode Penelitian

Pada gambar dibawah ini menunjukkan proses penelitian yang dilakukan dengan pengumpulan data, *preprocessing*, visualisasi data, pengolahan data dan pengujian klasifikasi.



Gambar 1. Rancangan Sistem

2.1 Pengumpulan Data

Penelitian ini menggunakan metode pengumpulan data penyakit *stroke* dari kaggle Jillani Soft Tech sebanyak 4981 data dengan 11 variabel yaitu jenis kelamin, usia, hipertensi, penyakit jantung, status menikah, jenis pekerjaan, jenis tempat tinggal, kadar glukosa rata-rata, *bmi*, status merokok, dan *stroke*.

2.2 Preprocessing

Preprocessing tahap proses data yang akan digunakan untuk proses analisa, *Preprocessing* berupa *encoding* dan *missing* data.

2.2.1 Encoding

Tahap *encoding* dalam penelitian ini adalah proses mengubah data jenis kelamin, umur, hipertensi, penyakit jantung, status menikah, jenis pekerjaan, tempat

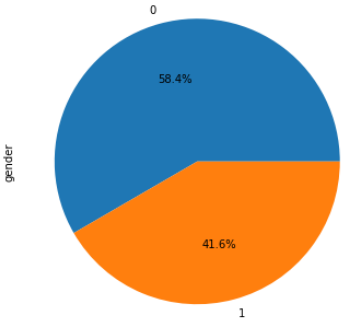
tinggal, kadar glukosa rata-rata, berat badan, status merokok, *stroke* yang belum bersifat angka dan akan diubah menjadi numerik dengan teknik kategorikal.

2.2.2 Missing Data

Tahap *missing* data dalam penelitian ini adalah proses menghilangkan data *Nan* yang berada di variabel *Bmi* pada *record* 3 dengan cara menghapus data yang terdapat *Nan* sehingga total data yang diperoleh 4980.

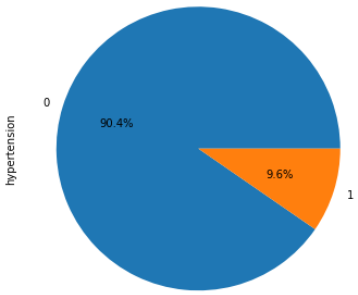
2.3 Visualisasi Data

Visualisasi data adalah tahap memproses data menggunakan elemen visual seperti diagram untuk merepresentasikan semua data yaitu jenis kelamin, hipertensi, penyakit jantung, pernah menikah, jenis pekerjaan, jenis tempat tinggal, status merokok, *stroke*.



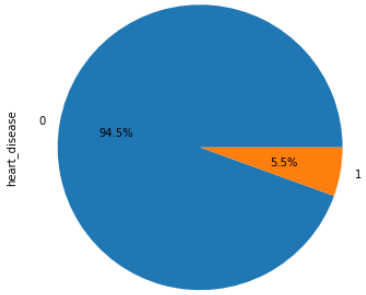
Gambar 2. Diagram jenis kelamin

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram jenis kelamin menunjukan bahwa angka 0 adalah perempuan sebesar 58,4% dan 1 adalah laki-laki sebesar 41,6%.



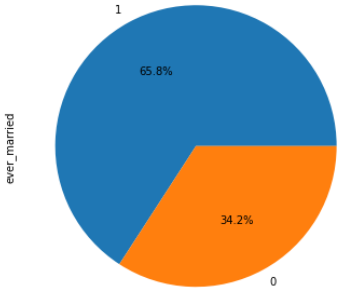
Gambar 3. Diagram hipertensi

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram hipertensi menunjukan bahwa angka 0 adalah yang tidak mempunyai hipertensi sebesar 90,4% dan 1 adalah yang mempunyai hipertensi sebesar 9,6%.



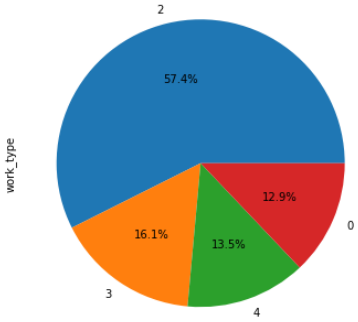
Gambar 4. Diagram penyakit jantung

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram penyakit jantung menunjukan bahwa angka 0 adalah yang tidak mempunyai penyakit jantung sebesar 94,5% dan 1 adalah yang mempunyai penyakit jantung sebesar 5,5%.



Gambar 5. Diagram status menikah

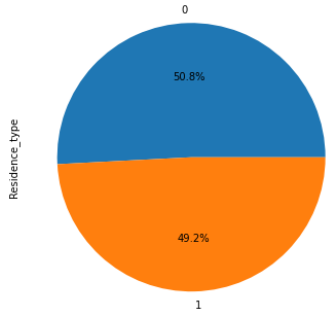
Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram status menikah menunjukan bahwa angka 0 adalah tidak menikah sebesar 34,2% dan 1 adalah menikah sebesar 65,8%.



Gambar 6. Diagram jenis pekerjaan

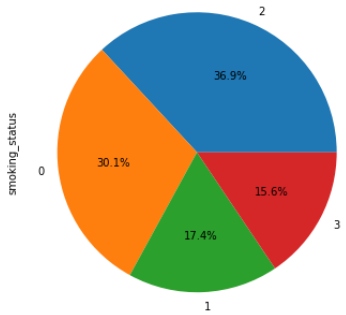
Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram jenis pekerjaan menunjukan bahwa angka 0 adalah pekerjaan pemerintah sebesar 12,9%, 2 adalah swasta sebesar 57,4%. 3 adalah wiraswasta sebesar 16,1% dan 4

adalah anak-anak sebesar 13,5%.



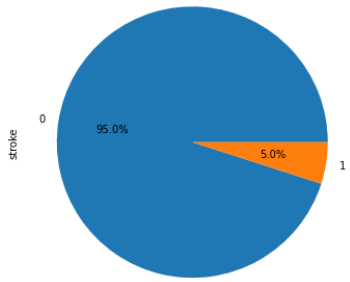
Gambar 7. Diagram tempat tinggal

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram tempat tinggal menunjukan bahwa angka 0 adalah urban sebesar 50,8% dan 1 adalah rural sebesar 49,2%.



Gambar 8. Diagram status merokok

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram status merokok menunjukan bahwa angka 0 adalah tidak diketahui sebesar 30,1%, 1 adalah sebelumnya merokok sebesar 17,4%, 2 adalah tidak pernah merokok sebesar 36,9% dan 3 adalah merokok sebesar 15,6%.



Gambar 9. Diagram *stroke*

Berdasarkan gambar diatas terdapat tahap proses visualisasi data diagram penyakit *stroke* menunjukan bahwa angka 0 adalah tidak mempunyai penyakit *stroke* sebesar 95,0% dan 1 adalah mempunyai penyakit *stroke* sebesar 5,0%.

2.4 Pengolahan Data

Algoritma *Logistic Regression* adalah menjelaskan variabel dependen dan variabel independen untuk menghubungkan satu atau lebih variabel bebas dengan variabel terkait yang berisi data berkode 0 dan 1, benar atau salah[6]. Metode ini merupakan metode jenis model klasifikasi[7], Rumus algoritma *Logistic Regression* memiliki persamaan sebagai berikut :

$$\ln(\frac{p}{1-p}) = B_0 + B_1X \quad [8]$$

Keterangan :
ln = Logaritma natural
B_0 = Konstanta
B_1 = Koefisien masing-masing variabel
x = Variabel independen
p = Probabilitas logistik yang dirumuskan sebagai berikut :

$$p = \frac{\exp{(B_0+B_1X)}}{1+\exp{(B_0+B_1X)}} = \frac{e^{B_0+B_1X}}{1+e^{B_0+B_1X}} \quad [8]$$

Keterangan :
exp atau e = Fungsi exponent

2.5 Pengujian Klasifikasi

Confusion matrix adalah penyajian data dalam bentuk *matrix* untuk mengklasifikasikan data yang terdeteksi dengan benar dan seberapa sering data diklasifikasikan sebagai data lain secara sistem. *Confusion matrix* juga berguna dalam menghitung *accuracy*, *precision*, *recall* dan *f1-score*[9]. *Accuracy* adalah pengukuran seberapa benar dari sebuah sistem yang dapat mengklasifikasikan keseluruhan data. *Precision* adalah perbandingan jumlah data positif yang telah diklasifikasikan secara akurat oleh sistem, *recall* adalah pengukuran data positif yang telah diklasifikasikan dengan benar pada sistem dan *fi-score* untuk mengoptimalkan kombinasi dari *recall* dan *precision*, *f1-score* akan menggunakan mean harmonik dari *recall* dan *precision*[10].
Berikut perhitungan nilai *Accuracy* :

$$Accuracy = \frac{932+0}{932+0+0+64} = \frac{932}{996} = 94\%$$

Berikut perhitungan nilai *Precision* :

$$Precision = \frac{932}{932+64} = \frac{932}{996} = 94\%$$

Berikut perhitungan nilai *Recall* :

$$Recall = \frac{932}{932+0} = \frac{932}{932} = 100\%$$

Berikut perhitungan nilai *f1 Score* :

$$F1\ Score = 2 * \frac{100 * 94}{100+94} = 2 * \frac{9400}{194} = 97\%$$

3 Hasil Dan Pembahasan

Data yang didapatkan dari kaggle, data penyakit *stroke* sebanyak 4980, Data yang mempunyai riwayat penyakit *stroke* terdiri dari 248 dan yang tidak mempunyai riwayat penyakit *stroke* terdiri dari 4733. Data penyakit *stroke* kemudian dibagi menjadi 80% data *training* sejumlah 3984, dan 20% data *testing* sejumlah 996. Data terdiri dari 8 kolom yaitu kolom yang berisi jenis kelamin, hipertensi, penyakit jantung, pernah menikah, jenis pekerjaan, jenis tempat tinggal, status merokok, *stroke*. Hasil penelitian ini diukur dengan *Confusion matrix*, berdasarkan gambar 10 sistem dapat mengklasifikasikan 8 kolom yaitu jenis kelamin, hipertensi, penyakit jantung, pernah menikah, jenis pekerjaan, jenis tempat tinggal, status merokok, *stroke*.

```
[[932   0]
 [ 64   0]]
```

Gambar 10. Hasil pengujian *confusion matrix*

Accuracy, precision, recall dan *f1-score* berdasarkan prediksi dapat dihitung dengan data yang berasal *Confusion matrix*.

	precision	recall	f1-score	support
0	0.94	1.00	0.97	932
1	0.00	0.00	0.00	64
accuracy			0.94	996
macro avg	0.47	0.50	0.48	996
weighted avg	0.88	0.94	0.90	996

Gambar 11. Hasil pengujian sistem

Hasil dari uji sistem yang terdapat pada gambar 11 menunjukan *accuracy* sebesar 94%, *precision* sebesar 94%, *recall* sebesar 100% dan *f1-score* sebesar 97%

4 Kesimpulan Dan Saran

Berdasarkan dari hasil pengujian pada sistem, maka dapat disimpulkan bahwa, penelitian ini menggunakan *machine learning* untuk mengklasifikasikan penyakit *stroke* dengan menggunakan algoritma *Logistic Regression.* Menggunakan 3984 data *training* dan 996 data *testing* dengan mendapatkan hasil akurasi sebesar 94%. Penerapan algoritma *Logistic Regression* memperbaiki hasil akurasi yang menggunakan algoritma *Support Vector Machine* (SVM) sebesar 76%, peningkatan hasil akurasi tersebut sebesar 18%.

Referensi

[1] Y. A. Utama and S. S. Nainggolan, "Faktor Resiko yang Mempengaruhi Kejadian Stroke: Sebuah Tinjauan Sistematis," *J. Ilm. Univ. Batanghari Jambi*, vol. 22, no. 1, p. 549, 2022, doi: 10.33087/jiubj.v22i1.1950.

[2] I. Lishania, R. Goejantoro, and Y. N. Nasution, "Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda," *J. Eksponensial*, vol. 10, no. 2, pp. 135–142, 2019, [Online]. Available: http://jurnal.fmipa.unmul.ac.id/index.php/exponensial/article/view/571

[3] D. Mutiarasari, "Ischemic Stroke: Symptoms, Risk Factors, and Prevention," *J. Ilm. Kedokt. Med. Tandulako*, vol. 1, no. 1, pp. 60–73, 2019.

[4] K. R. Sulaeman, C. Setianingsih, and R. E. Saputra, "Analisis Algoritma Support Vector Machine Dalam Klasifikasi Penyakit Stroke," *eProceedings Eng.*, vol. 9, no. 3, pp. 922–928, 2022, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/17909/17544%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/17909

[5] D. Ulfatul, M. Rachmad, H. Oktavianto, and M. Rahman, "Perbandingan Metode K-Nearest Neighbor Dan Gaussian Naive Bayes Untuk Klasifikasi Penyakit Stroke Comparison Of K-Nearest Neighbor And Gaussian Naive Bayes Methods For Stroke Disease Classification," *J. Smart Teknol.*, vol. 3, no. 4, pp. 2774–1702, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST

[6] F. D. Pramakrisna, F. D. Adhinata, N. Annisa, and F. Tanjung, "Aplikasi Klasifikasi SMS Berbasis Web Menggunakan Algoritma Logistic

Regression Web-based Classifying SMS Application Using Logistic Regression Algorithm," vol. 11, no. 2, pp. 90–97, 2022, doi: 10.34148/teknika.v11i2.466.

[7] I. F. Ramadhy and Y. Sibaroni, "Analisis Trending Topik Twitter dengan Fitur Ekspansi FastText Menggunakan Metode Logistic Regression," *J. Ris. Komputer)*, vol. 9, no. 1, pp. 2407–389, 2022, doi: 10.30865/jurikom.v9i1.3791.

[8] A. Untuk *et al.*, "APPLICATION TO PREDICT POVERTY BASED ON E-COMMERCE DATA USING LOGISTIC," vol. 6, no. 2, pp. 3109–3122, 2020.

[9] A. Farhan, "Implementasi Sentiment Analysis Cyberbullying Pada Twitter Dengan Algoritma Support Vector Machine," 2020.

[10] K. A. Shianto, K. Gunadi, and E. Setyati, "Deteksi Jenis Mobil Menggunakan Metode YOLO Dan Faster R-CNN".