Komparasi Algoritma Klasifikasi Penyakit Stroke Menggunakan *K-Nearest Neighbour* dan *Regresi Logistik*

Maheza Fiko Pratama

Prodi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Surabaya, Indonesia

maheza.22048@mhs.unesa.ac.id

Nuhaa Salsabila Shidqiyyah

Prodi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Surabaya, Indonesia

nuhaa.22455@mhs.unesa.ac.id

Azaria Syahla Fitan Adibah

Prodi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Surabaya, Indonesia

azaria.22019@mhs.unesa.ac.id

Elly Matul Immah

Prodi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Surabaya, Indonesia

ellymatul@unesa.ac.id

Riskyana Dewi Intan Puspitasari, M. Kom.

Prodi Sains Data, Fakultas Matematika dan Ilmu Pengetahuan Alam

Universitas Negeri Surabaya

Surabaya, Indonesia

riskyanapuspitasari@unesa.ac.id

Stroke merupakan penyakit neurovaskular yang menjadi penyebab kematian dan kecacatan utama di seluruh dunia. Di Indonesia, prevalensi stroke masih tinggi dan masih menjadi penyebab utama kematian dan kecacatan. Penelitian ini bertujuan mengembangkan model prediktif penyakit stroke menggunakan algoritma K-Nearest Neighbor (KNN) dan Regresi Logistik. Data yang dikumpulkan mencakup faktor risiko seperti usia, jenis kelamin, kondisi kesehatan, perilaku, dan faktor sosial ekonomi. Hasil penelitian menunjukkan bahwa algoritma KNN dapat membedakan dengan akurat antara pasien yang berisiko mengalami stroke dan yang tidak. Regresi Logistik digunakan untuk menganalisis hubungan penyebab dengan variabel respons biner atau kategorikal. Pengembangan model prediktif ini diharapkan dapat membantu mengidentifikasi risiko stroke secara efisien dan efektif, memungkinkan deteksi dini dan pengendalian faktor risiko. Manfaat sari penelitian ini berdampak dan melibatkan masyarakat umum dan tenaga medis, memungkinkan pencegahan dan pengelolaan risiko stroke secara lebih proaktif, dengan identifikasi dini pasien yang berisiko tinggi. Penelitian ini memberikan kontribusi penting dalam upaya pencegahan dan penanganan penyakit stroke, terutama di Indonesia dan negara-negara berkembang lainnya. Dengan model prediktif yang dapat diandalkan, tenaga medis dapat melakukan intervensi lebih cepat dan tepat, meningkatkan peluang penyembuhan dan mengurangi dampak kecacatan. Selain itu, kesadaran masyarakat terhadap faktor risiko yang dapat dipengaruhi juga dapat ditingkatkan melalui informasi yang diberikan oleh model ini. Implementasi model prediktif ini dapat menjadi langkah positif dalam meningkatkan kesehatan masyarakat secara keseluruhan, meminimalkan beban penyakit stroke, dan mendukung sistem kesehatan untuk memberikan pelayanan yang lebih baik dalam menangani tantangan penyakit neurovaskular ini.

Keywords— Stroke, K-Nearest Neighbour, Regresi Logistik, Model Prediktif, Klasifikasi

I. PENDAHULUAN

Stroke merupakan penyakit *neurovaskular* yang menjadi salah satu penyebab kematian di dunia. Dengan jumlah kematian sebanyak 66 juta orang dan kecacatan sebanyak 143 juta di seluruh dunia. Stroke adalah gangguan yang disebabkan oleh penyumbatan atau pecahnya pembuluh darah di otak, yang berkembang pesat dan berlangsung lebih dari 24 jam[1]. Secara global, selama empat dekade terakhir, penyakit stroke telah meningkat lebih dari 100% di negara-negara berpenghasilan rendah dan menengah[2]. Indonesia sebagai negara berkembang, menjadikan penyebab kematian pada peyakit stroke juga masih tinggi, yaitu 14,7 per mil pada 2018 dan mengalami peningkatan dari tahun 2013-2018, yaitu 2,6 per mil[3]. Prevalensi stroke pada penduduk usia di atas 15 tahun,

menurut data Riskesdas tahun 2013, mencapai 7 permil, mengalami peningkatan signifikan dibandingkan dengan tahun 2007 yang sebesar 6 permil[4]. Dalam pandangan teoritis, stroke merupakan suatu penyakit yang *multifaktorial*, di mana banyak faktor dapat menyebabkan terjadinya. Beberapa faktor yang tidak dapat dimodifikasi termasuk usia, jenis kelamin, dan faktor lainnya. Sementara itu, faktor kondisi kesehatan seperti hipertensi dan penyakit jantung juga dapat berperan. Faktor perilaku seperti kebiasaan aktivitas fisik, pola makan, dan merokok turut memengaruhi risiko stroke. Di samping itu, faktor sosial ekonomi, seperti wilayah tempat tinggal, tingkat pendidikan, dan tingkat pendapatan, juga dianggap berpotensi memainkan peran dalam terjadinya stroke[5].

Penanganan stroke ditangani oleh dokter spesialis penyakit syaraf, yang melakukan pemeriksaan dan diagnosis pada pasien. Dokter mengajukan pertanyaan untuk mengidentifikasi keluhan dan faktor-faktor pemicu stroke, sehingga dapat menyimpulkan tingkat risiko stroke pada pasien. Dengan deteksi dini dan pengendalian faktor risiko, penyakit stroke dapat dicegah. Namun, proses ini memerlukan biaya dan waktu karena melibatkan kunjungan ke rumah sakit. Untuk membantu pasien dan masyarakat umum, diperlukan suatu model prediktif penyakit stroke yang dapat membantu menghindari risiko stroke. Model ini juga berguna untuk membantu dokter atau tenaga kesehatan dalam melakukan diagnosis stroke. Pemilihan algoritma yang tepat untuk mengembangkan model prediksi penyakit stroke menjadi hal krusial karena akan memengaruhi hasil yang diperoleh[6].

Pada penelitian ini akan dilakukan perbandinga algoritma *K-Nearest Neighbor*(KNN) dan Regresi Logistik. KNN merupakan algoritma yang sederhana tapi kuat dan efektif untuk mengklasifikasikan data[7]. Algoritma KNN melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut[8], Zuriati menyatakan Hasil penelitian menunjukkan bahwa algoritma KNN mampu dengan akurat membedakan antara pasien yang berisiko mengalami stroke dan mereka yang tidak. Metode ini memanfaatkan pendekatan berbasis tetangga terdekat untuk menentukan kategori kesehatan pasien[6]. Sedangkan Regresi Logistik merupakan salah satu metode yang umum digunakan dalam menganalisis hubungan kausalitas dengan variabel respons berupa data non-metrik. Keunggulan regresi logistik terletak pada kemampuannya mengatasi skenario di mana variabel respon bersifat biner atau kategorikal[7].

Tujuan utama dari penelitian ini adalah untuk mengembangkan suatu model prediktif penyakit stroke yang dapat membantu mengidentifikasi risiko stroke pada pasien dengan lebih efisien dan efektif. Manfaat dari penelitian ini dapat dirasakan secara luas, baik oleh masyarakat umum maupun oleh tenaga medis. Dengan adanya model prediktif, diharapkan dapat membantu masyarakat dalam memahami dan mengelola risiko stroke secara lebih proaktif. Pasien yang berisiko tinggi dapat lebih dini diidentifikasi, sehingga upaya pencegahan dan pengendalian faktor risiko dapat dilakukan secara lebih tepat waktu.

KAJIAN TEORI

A. *Stroke*

Stroke adalah suatu gangguan dalam fungsi otak yang berkembang dengan cepat dan ditandai oleh gejala klinis yang terjadi selama lebih dari 24 jam, dapat berpotensi fatal. Penyebab utama stroke adalah terganggunya aliran darah ke otak. Stroke adalah penyebab utama kematian dan kecacatan kedua di seluruh dunia. Stroke bukanlah penyakit tunggal, tetapi dapat disebabkan oleh berbagai faktor risiko, proses dan mekanisme penyakit. Stroke, kadang-kadang disebut serangan otak, terjadi ketika ada sesuatu yang menghalangi suplai darah ke bagian otak atau ketika pembuluh darah di otak pecah. Dalam kedua kasus tersebut, bagian otak menjadi rusak atau mati. Stroke dapat menyebabkan kerusakan otak yang menetap, kecacatan jangka panjang, atau bahkan kematian. Selain itu, stroke adalah penyebab kematian nomor dua di dunia, yang bertanggung jawab atas sekitar 11% dari total kematian. Menurut penelitian Pusat Pengendalian dan Pencegahan Penyakit (CDC) dan Pencegahan Penyakit (CDC), setiap 40 detik seseorang mengalami stroke; setiap 3,5 menit, seseorang meninggal karena stroke[8].

B. *Logistic Regression*

Regresi logistik merupakan jenis algoritma *machine learning* yang digunakan untuk menyelesaikan masalah klasifikasi. Masalah dengan hasil biner, seperti Ya/Tidak, 0/1, Benar/Salah, disebut sebagai masalah klasifikasi. Regresi logistik adalah suatu teknik analisis statistika yang digunakan untuk menjelaskan keterkaitan antara variabel respons yang memiliki dua kategori atau lebih, dengan satu atau lebih variabel penjelas yang berskala kategori atau interval. Dalam analisis regresi logistik, variabel yang dihasilkan bersifat biner dan dikotomi. Model regresi logistik biner diterapkan ketika variabel respons menghasilkan dua kategori, yaitu 0 dan 1, sesuai dengan distribusi *Bernoulli* berikut:

$$F(Y=y)=\pi(x)^{y}\text{-}(1\text{-}\pi(x)^{1-y},\ y=0.1 \quad (1)$$

Dimana jika $y=0$ maka $P(Y=0) = 1-\pi$ dan jika $y=1$ maka $P(Y=1) = \pi$.

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_P x_p)} \quad (2)$$

Dimana:
$\pi(x)$ = peluang kejadian sukses dengan nilai probabilitas
$0 \leq \pi(x) \leq 1$
β0 = intersept (bilangan konstan),
β1, …, βp = parameter regresi logistik,
x1, …, xp = nilai peubah bebas.

C. *K-Neares Neighbours*

Algoritma KNN adalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek tersebut. Jarak antara data latih dengan data uji dihitung dengan cara mengukur jarak antara titik yang

merepresentasikan data testing dengan semua titik yang merepresentasikan data training dengan rumus *Euclidean Distance*. Persamaan atau rumus *Euclidean Distance* adalah:

$$dist = \sum_{i=1}^{p} \sqrt{(x2 - x1)\text{^}2}$$

(3)

Semakin besar nilai dist, semakin meningkat tingkat perbedaan antara dua individu, dan sebaliknya, semakin kecil nilai dist, semakin dekat tingkat kesamaan antara individu tersebut. Penentuan nilai k yang optimal untuk algoritma ini bergantung pada karakteristik data. Umumnya, nilai k yang tinggi dapat mengurangi dampak noise dalam klasifikasi, tetapi sekaligus dapat membuat batasan antar setiap klasifikasi menjadi lebih kabur[6].

II. METODE

A. Data Set

Penelitian ini menggunakan dataset publik yang tersedia pada kaggle dataset repository, yaitu dataset https://www.kaggle.com/code/sukhadadharangaonkar/stroke-prediction-dataset-knn-algorithm/notebook dengan jumlah data sebanyak 5111 baris dan terdiri dari 11 kolom dengan variabel *gender* (jenis kelamin), *age* (usia), *hypertension* (hipertensi), *heart disease* (penyakit jantung), BMI (Indeks Massa Tubuh), *ever married* (pernah menikah), *work type* (jenis pekerjaan), dan *residence type* (jenis tempat tinggal), *average glucose level* (tingkat glukosa rata-rata), *smoking status* (status merokok), dan stroke .Kolom jenis kelamin menunjukkan apakah individu tersebut adalah pria atau Wanita. Variabel usia menunjukkan rentang usia individu, memberikan informasi tentang seberapa tua atau muda mereka. Kemudian terdapat variabel hipertensi bertujuan untuk menunjukkan apakah individu tersebut menderita hipertensi atau tidak. Variabel *heart disease* untuk menunjukkan apakah individu tersebut memiliki penyakit jantung atau tidak. Selenjutnya yaitu variabel BMI untuk mengukur rasio berat badan terhadap tinggi badan, memberikan gambaran tentang proporsi tubuh seseorang. Variabel *ever married* menunjukkan apakah individu tersebut pernah menikah atau tidak. *Work type* bertujuan untuk menyajikan informasi tentang jenis pekerjaan yang dijalani oleh individu, seperti pekerjaan kantor atau pekerjaan lapangan. Variabel *residence type* menunjukkan apakah individu tinggal di daerah perkotaan atau pedesaan. Kemudian, variabel *average glucose* level untuk menunjukkan rata-rata kadar glukosa dalam darah individu. *Smoking status* merupakan variabel menyatakan apakah individu tersebut perokok atau tidak, dan terakhir terdapat kolom stroke menunjukkan apakah individu tersebut telah mengalami stroke atau tidak. Dalam konteks variabel stroke nilai 1 menunjukkan bahwa individu tersebut telah mengalami stroke, sementara nilai 0 mengindikasikan bahwa individu tersebut belum mengalami stroke. Hal ini merupakan representasi biner atau dikotomi dari kejadian stroke, di mana 1 menunjukkan keberadaan *(ya)* dan 0 menunjukkan ketiadaan *(tidak)*[9].

B. Pra Pemrosesan

Dalam penelitian ini, proses persiapan data sangat dipentingkan untuk mengoptimalkan dataset klasifikasi stroke. Dataset yang digunakan berasal dari Kaggle repository dengan 5111 baris dan 12 kolom yang mencakup berbagai variabel seperti id, jenis kelamin, usia, hipertensi, penyakit jantung, Indeks Massa Tubuh (BMI), status pernikahan, jenis pekerjaan, tipe tempat tinggal, tingkat glukosa rata-rata, status merokok, dan kolom target "stroke". Tahapan *pre-processing* melibatkan langkah-langkah berikut :

1) Data Reduction

Data Reduction adalah proses untuk mengurangi atau mereduksi sejumlah data yang tidak dibutuhkan. Data Reduction sangat berguna untuk mendapatkan atribut dan data yang akan digunakan dalam penelitian ini. Atribut yang digunakan pada penelitian ini adalah *gender, age, hypertension, heart disease, BMI, ever married, work type , residence type, average glucose level , smoking status, dan stroke* sebagai *outcome* dari yang awalnya 12 data kolom menjadi hanya 11 dimana kolom id tidak diperlukan[10].

2) Pengecekan Value

Data transformation adalah proses mengubah suatu data agar mendapatkan data yang lebih atau sesuai dengan kebutuhan. Dalam proses algoritma *K-Nearest Neighbour* dan Regresi Logistik ini data yang bisa diolah adalah data yang berupa numerik sehingga data yang berupa text seperti *gender, ever married, work type, Residence type, dan smoking status* harus diubah menjadi data yang berupa numerik agar bisa dilanjutkan ke dalam proses selanjutnya[11].

3) Oversampling

Oversampling adalah salah satu teknik yang dapat digunakan dalam pemrosesan data, terutama ketika menghadapi ketidakseimbangan data. Teknik ini melibatkan peningkatan jumlah sampel dari kelas minoritas dalam dataset. Salah satu teknik oversampling yang umum digunakan adalah *Synthetic Minority Oversampling Technique* (SMOTE). Dengan menggunakan SMOTE, kita dapat meningkatkan akurasi model pada data yang tidak seimbang[12].

4) Modeling

Algoritma klasifikasi yang digunakan adalah *K-Nearest Neighbou*r dan Regresi Logistik. Algoritma ini menggunakan atribut *gender, age, hypertension, heart disease, BMI, ever married, work type , residence type, average glucose level, dan smoking status* sebagai *independent variable* yang sudah dilakukan pembersihan serta transformasi data. Sedangkan atribut Stroke digunakan sebagai dependen variable atau kelas target yang menunjukkan jenis Stroke atau tidak yang diklasifikasikan. Seperti kasus klasifikasi pada umumnya, dataset akan dipecah menjadi 2 yaitu sebagai data training dan testing dengan pembagian dataset di rasio 20. Data training disintenis dengan SMOTE. Setelah data dipisah menjadi 2 bagian, data training digunakan untuk pembelajaran dengan menggunakan algoritma yang dipakai. Pemodelan

dibangun dengan 2 jenis yaitu pemodelan tanpa kombinasi SMOTE dan pemodelan dengan kombinasi dengan SMOTE[13].

III. PEMBAHASAN

Pada penelitian ini menggunakan Processor Intel(R) Core i5 GHz CPU, 8GB RAM, serta sistem operasi Microsoft Windows 11 64-bit. Pendekatan *k-fold cross validation* diterapkan dalam klasifikasi dengan memilih nilai $k = 5$.

A. Proses Akurasi Hasil Klasifikasi

Pada proses klasifikasi, terdapat empat algoritma klasifikasi yang digunakan di mana setiap algoritma memiliki kelebihan dan kekurangan sebagai klasifikator. Dalam klasifikasi, tidak dapat dipungkiri bahwa klasfikator tidak bekerja secara 100% benar. Oleh karena itu, perlu adanya alat ukur untuk mengukur performa klasifikasi menggunakan matriks *confusion*. Matriks *confusion* adalah tabel dengan ukuran dimana diagonal matriksnya merepresentasikan akurasi terbaik klasifikator. Nilai akurasi diperoleh dari mengetahui jumlah data yang benar terklasifikasikan. Semua algoritma klasifikasi berusaha untuk membentuk model dengan akurasi yang tinggi. Perhitungan akurasi dapat menggunakan persamaan.

$$Akurasi = \frac{TP+TN}{TP+FP+TN+FN} \quad (4)$$

Dimana TP (Positif Benar) ialah jumlah data pada kelas A yang terprediksi secara benar pada kelas A. FP (Positif Salah) ialah jumlah data pada kelas B yang terprediksi secara benar pada kelas A. TN (Negatif Benar) ialah jumlah data pada kelas B yang terprediksi secara benar pada kelas B. Sedangkan FN (Negatif Salah) ialah jumlah data pada kelas A yang terprediksi secara benar pada kelas B.

Precision atau presisi merupakan hasil pembagian antara positif benar dengan jumlah data yang tertandai sebagai positif, sedangkan perbandingan antara data positif benar (TP) dengan jumlah data yang tergolong positif disebut dengan *recall*. Persamaan 4 adalah persamaan yang digunakan untuk menghitung nilai presisi dan persamaan 5 adalah persamaan untuk menghitung *recall*. Pengaruh dari nilai presisi dalam algoritma adalah nilai dari data yang benar pada tiap kelas dari keseluruhan data yang terprediksi pada tiap kelas. Sedangkan nilai *recall* merupakan nilai dari data yang terprediksi pada tiap kelas dibandingkan keseluruhan data yang sebenarnya dalam tiap kelas[14].

$$F = 2\,X\,\frac{precision\ x\ recall}{precision + recall} \quad (5)$$

B. Hasil Akurasi Algoritma

Setiap algoritma yang digunakan terdapat beberapa parameter yang dikontrol. Pada algoritma KNN adalah mengubah nilai $k = 3$, dan 5. Untuk algoritma regresi logistik parameter kemungkinan maximum yang digunakan adalah 10. Hasil akurasi pada tiap algoritma pada Tabel 1, merupakan hasil akurasi rata-rata dari tiap-tiap parameter yang digunakan.

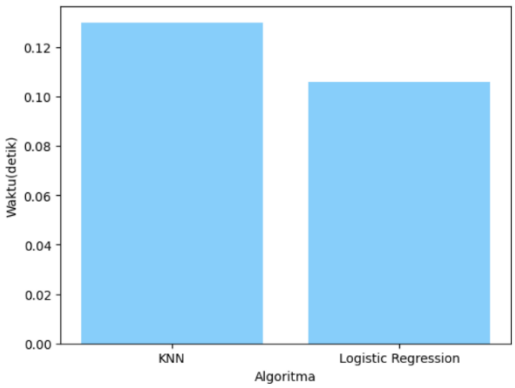
Algoritma	*Akurasi (%)*	*Precision*	*Recall*	*F1-Score*
KNN	79	97	80	88
Logistik Regression	74	98	74	85

Tabel 1. Hasil Akurasi Tiap Algoritma

Dari data yang diberikan, terdapat hasil evaluasi kinerja dua model klasifikasi, yaitu KNN dan regresi logistik, berdasarkan metrik evaluasi seperti akurasi, *precision*, *recall, dan F1-score*. Dalam hal akurasi, KNN menunjukkan performa yang lebih baik dibandingkan dengan regresi logistik, dengan nilai akurasi sebesar 79% dibandingkan dengan 74%. Hal ini mengindikasikan bahwa model KNN memiliki kemampuan yang lebih baik dalam mengklasifikasikan data dengan benar secara keseluruhan.

Namun, saat melihat metrik precision, recall, dan F1-score, kita dapat mengevaluasi lebih detail kinerja keduanya. Logistik Regression menunjukkan nilai precision yang lebih tinggi (98%) dibandingkan dengan KNN (97%), yang berarti bahwa Logistik Regression cenderung memberikan lebih sedikit false positive. Di sisi lain, KNN memiliki recall yang lebih tinggi (80%) dibandingkan dengan Logistik Regression (74%), yang menunjukkan kemampuan KNN untuk mengidentifikasi lebih banyak kasus positif sebenarnya. F1-score yang mencakup keseimbangan antara precision dan recall juga menunjukkan KNN memiliki performa yang lebih baik (88%) dibandingkan dengan Logistik Regression (85%).

Secara keseluruhan, meskipun KNN memiliki akurasi yang lebih tinggi, keputusan untuk memilih model tergantung pada tujuan spesifik dari permasalahan klasifikasi tersebut. Jika lebih diutamakan untuk menghindari *false positive*, regresi logistik mungkin menjadi pilihan yang lebih baik, sementara KNN mungkin lebih diinginkan jika identifikasi yang baik terhadap kasus positif merupakan prioritas utama.



Gambar 1. Grafik Waktu Untuk Membangun Model

Dapat diambil kesimpulan melalui Gambar 1 bahwa algoritma yang memerlukan waktu lebih lama untuk membangun model adalah algoritma regresi logistik. Tetapi, algoritma KNN tidak membangun model dari data pelatihan sehingga tidak memerlukan waktu untuk membangun model yang mengakibatkan algoritma KNN membutuhkan ruang penyimpanan lebih banyak dikarenakan setiap data *input* dibandingkan dengan seluruh data latih.

REFERENCES

[1] E. J. Benjamin *dkk.*, "Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association," *Circulation*, vol. 139, no. 10, hlm. e56–e528, Mar 2019, doi: 10.1161/CIR.0000000000000659.

[2] V. L. Feigin *dkk.*, "World Stroke Organization (WSO): Global Stroke Fact Sheet 2022," *International Journal of Stroke*, vol. 17, no. 1. SAGE Publications Inc., hlm. 18–29, 1 Januari 2022. doi: 10.1177/17474930211065917.

[3] "World Stroke Day 2023, Greater Than Stroke, Kenali dan Kendalikan Stroke."

[4] Kata, "DAFTAR ISI."

[5] J. Epidemiologi Kesehatan Indonesia *dkk.*, "Faktor-Faktor yang Berhubungan dengan Kejadian Stroke pada Penduduk Usia >15 Tahun di Provinsi Daerah Istimewa Yogyakarta (Analisis Riskesdas 2018) Factors Associated with Stroke in Population Aged >15 Years in Special Region of Yogyakarta (Analysis of Basic Health Research 2018)."

[6] Zuriati Z dan Diterima, "Klasifikasi Penyakit Stroke Menggunakan Algoritma K-Nearest Neighbor (KNN) INFORMASI ARTIKEL ABSTRAK Classification of Stroke Using the K-Nearest Neighbor (KNN) Algorithm," vol. 1, no. 1, hlm. 1–8, 2023, doi: 10.xxxxx.

[7] "287317093".

[8] S. Annas, A. Aswi, M. Abdy, dan B. Poerwanto, "Stroke Classification Model using Logistic Regression," dalam *Journal of Physics: Conference Series*, IOP Publishing Ltd, Des 2021. doi: 10.1088/1742-6596/2123/1/012016.

[9] "stroke prediction dataset - KNN algorithm."

[10] C. Haryanto, N. Rahaningsih, dan F. Muhammad Basysyar, "KOMPARASI ALGORITMA MACHINE LEARNING DALAM MEMPREDIKSI HARGA RUMAH," 2023.

[11] "5. 161240000500_BAB IV".

[12] A. E. Budiman dan A. Widjaja, "Analisis Pengaruh Teks Preprocessing Terhadap Deteksi Plagiarisme Pada Dokumen Tugas Akhir," *Jurnal Teknik Informatika dan Sistem Informasi*, vol. 6, no. 3, Des 2020, doi: 10.28932/jutisi.v6i3.2892.

[13] F. Hamami dan A. Dahlan, "KLASIFIKASI CUACA PROVINSI DKI JAKARTA MENGGUNAKAN ALGORITMA RANDOM FOREST DENGAN TEKNIK OVERSAMPLING," 2022.

[14] E. Prasetyo, "Data Mining : Mengolah Data Menjadi Informasi Menggunakan Matlab," 2015. [Daring]. Tersedia pada: https://api.semanticscholar.org/CorpusID:61559069