Jurnal Penelitian Rumpun Ilmu Teknik (JUPRIT)
Vol.2, No.4 November 2023
e-ISSN: 2963-7813; p-ISSN: 2963-8178, Hal 146-154
DOI: https://doi.org/10.55606/juprit.v2i4.3040

Penerapan K-Nearest Neighbors (KNN) dalam Memprediksi dan Menghitung Akurasi Data Penyakit Stroke

Atika Haura Siregar
Universitas Negeri Medan
E-mail: atikahaurasiregar@gmail.com

Asri Angel Tumanggor
Universitas Negeri Medan
E-mail: Asriangel161@gmail.com

Akhwan Rahmadani
Universitas Negeri Medan
E-mail: akhwan.ramadhani123@gmail.com

Jalan Willem Iskandar, Pasar V Medan Estate, Percut Sei Tuan, Deli Serdang

Abstract. *Stroke is a disease characterized by a disruption in brain function caused by a lack of oxygen and blood flow to the brain, affecting various brain functions and causing difficulties in performing activities. The classification of stroke patients is still based on medical records that are not integrated, leading to a longer time for detection. The K-Nearest Neighbors (K-NN) algorithm is a part of machine learning that can be utilized to classify cases, including the classification of stroke patients. K-NN serves as the algorithm to determine classes and incorporate new data inputted in the specified format. In this study, the researcher aims to demonstrate that the classification algorithm of K-Nearest Neighbor with Bagging optimization can be used to determine if someone is affected by stroke. The predictions from this algorithm can facilitate decision-making in the healthcare field quickly.*

Keywords: *Classification, Machine Learning, K-Nearest Neighbors (K-NN) Algorithm, Stroke Prediction*

Abstrak. Stroke merupakan penyakit yang ditandai gangguan fungsi otak yang di sebabkan kurangnya pasokan oksigen dan aliran darah ke otak sehingga mempengaruhi beberapa fungsi otak yang membuat penderita mengalami kesulitan dalam melakukan aktifitas. klasifikasi pasien stroke yang di temukan masih berupa catatan medis yang belum terintegrasi sehingga perlu waktu yang lebih lama untuk mendeteksi. Algoritma K-NN merupakan bagian dari algoritma machine learning yang dapat digunakan untuk mengklasifikasikan salah satu kasusnya yaitu klasifikasi pasien stroke. K-NN digunakan sebagai algoritma penentu kelas untuk memasukkan data baru yang diinputkan sesuai format.

Kata kunci: Klasifikasi, Machine Learning, Algoritma K-Nearest Neighbors (K-NN), Prediksi Stroke

LATAR BELAKANG

Stroke atau Cerebrovascular Accident (CVA) merupakan gangguan fungsi saraf yang disebabkan oleh gangguan aliran darah dalam otak dan menyebabkan gangguan pada aktivitas fungsional. Stroke merupakan penyebab kematian ketiga tersering di negara maju, setelah penyakit jantung dan kanker. Dengan terus bertambahnya penderita stroke tiap tahunnya, belum terdapat upaya efektif dalam menanggulangi penyakit baik

Received November 07, 2023; Revised November 22, 2023; Accepted November 30, 2023
*Atika Haura Siregar, atikahaurasiregar@gmail.com

dengan meningkatkan kesadaran masyarakat maupun pengelolaan penyakit stroke yang optimal. Data Mining adalah serangkaian proses untuk menggali nilai tambah dari suatu kumpulan data besar berupa pengetahuan yang selama ini tidak diketahui secara manual. K-Nearest Neighbor merupakan metode yang paling simple, mudah diimplementasikan hanya dengan mengatur satu parameter[1].

Stroke sendiri merupakan kondisi kesehatan yang serius dan berpotensi mengancam nyawa yang terjadi akibat gangguan aliran darah ke otak. Kurangnya pasokan oksigen dan aliran darah dapat menyebabkan kerusakan permanen pada sel-sel otak dan mengakibatkan berbagai dampak negatif pada fungsi tubuh. Deteksi dini dan penanganan yang cepat dapat meningkatkan peluang pemulihan pasien serta mencegah kerusakan yang lebih lanjut[2].

Pada saat ini, pengklasifikasian pasien stroke masih banyak mengandalkan catatan medis yang belum terintegrasi dengan baik. Proses pengumpulan dan analisis data tersebut membutuhkan waktu yang lama, sehingga menunda waktu deteksi dan pengambilan keputusan. Oleh karena itu, diperlukan suatu pendekatan yang lebih efisien dan akurat dalam mengklasifikasikan pasien stroke[3]. Stroke, atau kecelakaan serebrovaskular, melibatkan cedera pada sistem saraf pusat sebagai akibat dari penyebab vaskular, dan merupakan penyebab utama kecacatan di seluruh dunia
K-Nearest Neighbor merupakan metode yang paling simple, mudah diimplementasikan hanya dengan mengatur satu parameter [4]. Namun K Nearest Neighbor juga memiliki beberapa kelemahan utama antara lain sensitive terhadap fitur-fitur yang kurang relevan, ukuran ketetanggaan k, data berderau maupun data pencilan, kompleksitas waktu untuk mencari tetangga terdekat setiap melakukan klasifikasi, dan kompleksitas memori untuk menyimpan semua data latih.[5]

Salah satu solusi yang dapat digunakan adalah menerapkan algoritma Machine Learning, khususnya algoritma K-Nearest Neighbors (KNN). Algoritma KNN merupakan metode klasifikasi yang bekerja dengan membandingkan data yang baru dengan data yang sudah ada dalam dataset, dan menentukan kelasnya berdasarkan mayoritas kelas dari k-nearest neighbors (tetangga terdekat) di sekitarnya.

KAJIAN TEORITIS

Banyak cara telah dilakukan untuk memprediksi berbagai penyakit dengan membandingkan kinerja teknologi Data Mining prediktif. sebagai proses pemilihan fitur, algoritma analisis komponen prinsip digunakan untuk mengurangi dimensi dan mengadopsi algoritma klasifikasi dalam membangun model klasifikasi [9]. Klasifikasi adalah teknik untuk membentuk model data yang belum diklasifikasikan, maka model dapat digunakan untuk mengklasifikasikan data baru.Salah satu algoritma klasisfikasi yang cukup populer yaitu algoritman K-Nearest Neighbour(KNN). Algoritma KNNadalah sebuah metode untuk melakukan klasifikasi terhadap objek berdasarkan data pembelajaran yang jaraknya paling dekat dengan objek yang diuji[10]–[13]. Algoritma KNN akan mengelompokan hasil perhitungan dengan data latih yang mempunyai kerabat terbanyak dalamnilai jangkauan yang ditentukan. Jarak antara data latih dan data uji dihitung menggunakan persamaan Euclidean.

Penelitian ini menggali penerapan algoritma k-Nearest Neighbors (KNN) dalam konteks prediksi penyakit stroke dan penilaian akurasinya. KNN merupakan algoritma klasifikasi yang berbasis pada keberadaan tetangga terdekat dalam ruang fitur.

METODE PENELITIAN

Penelitian ini terbagi menjadi beberapa tahap yaitu tahap *pre-processing* dimana data dipersiapkan sebelum diproses untuk klasifikasi. Tahap ini hanya memiliki dua proses yaitu proses *data cleansing* dimana data duplikat dan data dengan *missing values* akan dihapus demi menghindari keabnormalan data dan proses *data transformation* dimana atribut akan diubah nilainya menjadi data berupa angka. Setelah data disiapkan, maka data akan dipisahkan menjadi dua yaitu data pelatihan dan data pengujian untuk memulai proses klasifikasi dengan menggunakan algoritma K-Nearest Neighbor dan pada akhir tahap akan dilakukan evaluasi hasil.

Bahasa pemrograman yang digunakan adalah *Python* dengan bantuan beberapa *library* yaitu *scikit-learn*, *matplotlib*, *pandas* dan *numpy*.

Data yang akan digunakan pada penelitian ini adalah data pasien penderita penyakit stroke dari kaggle. Atribut yang terdapat di data ini adalah *id, gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status* dan *stroke* dengan atribut biner dimana 0 berarti

pasien tidak mengalami stroke dan 1 berarti pasien mengalami stroke. Terdapat 5110 baris data pasien pada tabel ini.

Berikut tabel yang menunjukkan sebagian dari data yang akan digunakan.

	id	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
1	9046	Male	67	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
2	51676	Female	61	0	0	Yes	Self-employed	Rural	202.21	N/A	never smoked	1
3	31112	Male	80	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
4	60182	Female	49	0	0	Yes	Private	Urban	171.23	34.4	smokes	1
5	1665	Female	79	1	0	Yes	Self-employed	Rural	174.12	24	never smoked	1
6	56669	Male	81	0	0	Yes	Private	Urban	186.21	29	formerly smoked	1
7	53882	Male	74	1	1	Yes	Private	Rural	70.09	27.4	never smoked	1
8	10434	Female	69	0	0	No	Private	Urban	94.39	22.8	never smoked	1
9	27419	Female	59	0	0	Yes	Private	Rural	76.15	N/A	Unknown	1
10	60491	Female	78	0	0	Yes	Private	Urban	58.57	24.2	Unknown	1
11	12109	Female	81	1	0	Yes	Private	Rural	80.43	29.7	never smoked	1
12	12095	Female	61	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes	1
13	12175	Female	54	0	0	Yes	Private	Urban	104.51	27.3	smokes	1
14	8213	Male	78	0	1	Yes	Private	Urban	219.84	N/A	Unknown	1
15	5317	Female	79	0	1	Yes	Private	Urban	214.09	28.2	never smoked	1
16	58202	Female	50	1	0	Yes	Self-employed	Rural	167.41	30.9	never smoked	1
17	56112	Male	64	0	1	Yes	Private	Urban	191.61	37.5	smokes	1
18	34120	Male	75	1	0	Yes	Private	Urban	221.29	25.8	smokes	1
19	27458	Female	60	0	0	No	Private	Urban	89.22	37.8	never smoked	1
20	25226	Male	57	0	1	No	Govt_job	Urban	217.08	N/A	Unknown	1

HASIL DAN PEMBAHASAN

1. *Pre-processing Data*

a. *Data Cleansing*

Pada tahap ini dilakukan pembersihan data dengan mencari *missing value* dan *redundant data*. Setelah dilakukan pencarian, terdapat satu atribut yang memiliki banyak *missing value* yaitu atribut *bmi* dan tidak ditemukan data yang *redundant*. Dilakukan penghapusan baris yang memiliki *missing value* sehingga jumlah baris berkurang menjadi 4909 baris.

```
dataset.duplicated().sum()
```

```
0
```

```
dataset.isna().sum()
```

```
id                   0
gender               0
age                  0
hypertension         0
heart_disease        0
ever_married         0
work_type            0
Residence_type       0
avg_glucose_level    0
bmi                201
smoking_status       0
stroke               0
dtype: int64
```

```
dataset.dropna(axis=0,inplace=True)
dataset.drop(columns='id',inplace=True)
```

b. *Data Transformation*

Di dalam data ini terdapat 5 atribut yang memiliki data berbentuk kategori yaitu *gender*, *ever_married*, *work_type*, *residence_type* dan *smoking_status*. Tahap ini mengubah atribut yang memiliki data kategori menjadi numerik.

```
dataset.head(10)
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
0	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked
2	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked
3	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smokes
4	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked
5	Male	81.0	0	0	Yes	Private	Urban	186.21	29.0	formerly smoked
6	Male	74.0	1	1	Yes	Private	Rural	70.09	27.4	never smoked
7	Female	69.0	0	0	No	Private	Urban	94.39	22.8	never smoked
9	Female	78.0	0	0	Yes	Private	Urban	58.57	24.2	Unknown
10	Female	81.0	1	0	Yes	Private	Rural	80.43	29.7	never smoked
11	Female	61.0	0	1	Yes	Govt_job	Rural	120.46	36.8	smokes

```
le=LabelEncoder()
dataset.avg_glucose_level=dataset['avg_glucose_level'].astype('int64')
dataset.bmi=dataset['bmi'].astype('int64')
dataset["gender"]=le.fit_transform(dataset["gender"])
dataset["ever_married"]=le.fit_transform(dataset["ever_married"])
dataset["work_type"]=le.fit_transform(dataset["work_type"])
dataset["Residence_type"]=le.fit_transform(dataset["Residence_type"])
dataset["smoking_status"]=le.fit_transform(dataset["smoking_status"])
dataset.head(10)
```

	gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status
0	1	67.0	0	1	1	2	1	228	36	1
2	1	80.0	0	1	1	2	0	105	32	2
3	0	49.0	0	0	1	2	1	171	34	3
4	0	79.0	1	0	1	3	0	174	24	2
5	1	81.0	0	0	1	2	1	186	29	1
6	1	74.0	1	1	1	2	0	70	27	2
7	0	69.0	0	0	0	2	1	94	22	2
9	0	78.0	0	0	1	2	1	58	24	0
10	0	81.0	1	0	1	2	0	80	29	2
11	0	61.0	0	1	1	0	0	120	36	3

2. *Processing Data*

a. *Splitting Data*

Setelah proses persiapan data selesai dilakukan, data langsung dibagi menjadi dua yaitu data pengujian dan data pelatihan untuk digunakan pada model *K-Nearest Neighbor* dalam memprediksi. Dari 4909 baris data yang dipersiapkan, diambil 30% untuk dijadikan data pengujian dan 70% untuk data pelatihan sehingga didapat 1473 baris untuk data pengujian dan 3436 baris untuk data pelatihan.

```
x = dataset.drop(columns='stroke')
y = dataset.stroke
X_train, X_test, y_train, y_test = train_test_split(x, y, test_size = 0.3)
scaler = StandardScaler()
```

b. Penerapan Algoritma *K-Nearest Neighbor* dengan *Jupyter Notebook*

Pada tahap ini dilakukan pengklasifikasian *K-Nearest Neighbor* guna penghitungan dan prediksi. Pertama dilakukan penghitungan untuk mencari akurasi dari data pelatihan.

```
print("Akurasi dataset pelatihan: ",metrics.accuracy_score(y_train, knn.predict(X_train)))
```

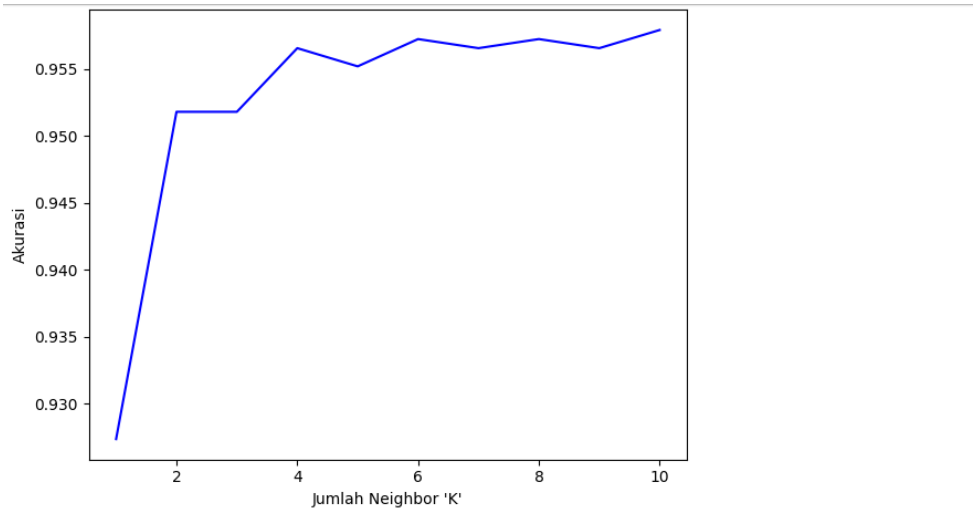
```
Akurasi dataset pelatihan:  0.9569266589057043
```

Gambar diatas menunjukkan hasil penghitungan akurasi data pelatihan yang sangat tinggi dan mendekati sempurna. Setelah mengetahui akurasi data pelathian, dilakukan pencarian nilai K dengan akurasi tertinggi pada data pengujian.

```
Kn = 11
accr_mean = np.zeros((Kn-1))
accr_std = np.zeros((Kn-1))

for i in range(1,Kn):
    kneig = KNeighborsClassifier(n_neighbors=i)
    kneig.fit(X_train, y_train)
    ypred = kneig.predict(X_test)
    accr_mean[i-1] = metrics.accuracy_score(y_test, ypred)
    accr_std[i-1] = np.std(ypred==y_test)/np.sqrt(ypred.shape[0])
```

Gambar diatas menunjukkan proses pencarian nilai K dengan akurasi paling tinggi dari rentang nilai K 1 sampai 11.




```
#hasil akurasi
print("\'K\' dengan akurasi paling tinggi adalah", accr_mean.argmax()+1, "dengan akurasi :", accr_mean.max())
```

```
'K' dengan akurasi paling tinggi adalah 10 dengan akurasi : 0.957909029192125
```

Gambar diatas berisi graf yang menunjukkan perbandingan akurasi. Dari proses tadi, didapatkan nilai K dengan akurasi terbesar yaitu K 10 dengan akurasi 95,79%.

KESIMPULAN DAN SARAN

Penelitian ini menunjukkan bahwa penerapan algoritma K-Nearest Neighbors (KNN) dapat menjadi pendekatan yang efektif dalam klasifikasi pasien stroke. Hasil eksperimen menunjukkan peningkatan signifikan dalam akurasi model, yang dapat mempercepat deteksi dan membantu tenaga medis dalam mengambil keputusan yang lebih cepat dan tepat.

Melalui penggunaan KNN, model dapat mengklasifikasikan pasien stroke dengan memanfaatkan informasi dari data pasien sebelumnya. Optimasi dengan metode Bagging memberikan keunggulan tambahan dalam meningkatkan ketahanan model terhadap variasi data dan meningkatkan kehandalan prediksi.

Dalam konteks pencegahan dan penanganan penyakit stroke, kecepatan dan akurasi deteksi sangat penting. Dengan adanya model ini, diharapkan dapat meningkatkan efisiensi sistem kesehatan dalam memberikan perawatan yang sesuai dan tepat waktu kepada pasien stroke. Penerapan teknologi Machine Learning, khususnya algoritma KNN dengan metode Bagging, menjadi langkah penting dalam mendukung pengambilan keputusan di bidang kesehatan.

Penelitian ini tidak hanya menghadirkan kontribusi terhadap pengembangan model klasifikasi pasien stroke, tetapi juga membuka peluang untuk pengembangan lebih lanjut dalam bidang Machine Learning dan kesehatan. Keberhasilan model ini membuka pintu untuk penelitian lebih lanjut dalam meningkatkan presisi dan adaptabilitas algoritma, sehingga dapat memberikan dampak positif yang lebih besar dalam upaya pencegahan dan penanganan penyakit stroke.

DAFTAR REFERENSI

Wati, A. (2020, February). Implementasi Artificial Neural Network Dalam Memprediksi Nilai Air Bersih Yang Disalurkan Di Provinsi Indonesia. In Seminar Nasional Teknologi Komputer & Sains (SAINTEKS) (Vol. 1, No. 1, pp. 182-189).

Nurmahaludin, dan Cahyano G.R. (2019, November). Klasifikasi Kualitas Air PDAM Menggunakan Algoritma KNN Dan K-Means, Prosiding SNRT (Seminar Nasional Riset Terapan), Politeknik Negeri Banjarmasin.

Vidiastanta, I.G., Hidayat, N., dan Dewi, R.K. (2020). Komparasi Metode K-Nearest Neighbor (KNN) Dengan Support Vector Machine (SVM) Untuk Klasifikasi Status Kualitas Air, Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer, No. 1, Vol .4, 312-319.

N. Reflan, A. Aflahah, Kusrini, and Juwari (2018). "Implementasi Metode K-Nearest Neighbor (Knn) Untuk Memprediksi Varietas Padi Yang Cocok Untuk Lahan Pertanian," J. Inf. Politek. Indonusa Surakarta. vol. 4, pp. 2–8.

Nugroho, S. (2020). PERBANDINGAN METODE FUZZY K-NEAREST NEIGHBOR DAN NEIGHBOR WEIGHTED K-NEAREST NEIGHBOR UNTUK DETEKSI PENYAKIT STROKE (Doctoral dissertation, University of Technology Yogyakarta).

Sutomo, F., Muaafii, D. A., Al Rasyid, D. N., Kurniawan, Y. I., Afuan, L., Cahyono, T.,& Iskandar, D. (2023). OPTIMIZATION OF THE K-NEAREST NEIGHBORS ALGORITHM USING THE ELBOW METHOD ON STROKE PREDICTION. Jurnal Teknik Informatika (Jutif), 4(1), 125-130.

Yulianto, R. A. D., Riadi, I., & Umar, R. (2023). PERANCANGAN KLASIFIKASI PASIEN STROKE DENGAN METODE K-NEAREST NEIGHBOR. *Rabit: Jurnal Teknologi dan Sistem Informasi Univrab*, *8*(2), 262-268.

Sitanggang, D., Nicholas, N., Wilson, V., Sinaga, A. R. A., & Simanjuntak, A. D. (2022). IMPLEMENTASI DATA MINING UNTUK MEMPREDIKSI PENYAKIT JANTUNG MENGGUNAKAN METODE K-NEAREST NEIGHBOR DAN LOGISTIC REGRESSION. Jurnal Tekinkom (Teknik Informasi dan Komputer), 5(2), 493-499.

Akmal, K., Faqih, A., & Dikananda, F. (2023). PERBANDINGAN METODE ALGORITMA NAÏVE BAYES DAN K-NEAREST NEIGHBORS UNTUK KLASIFIKASI PENYAKIT STROKE. JATI (Jurnal Mahasiswa Teknik Informatika), 7(1), 470-477.

HIDAYAT, R. (2021). Klasifikasi Penyakit Stroke Menggunakan Metode K-Nearest Neighbor Studi Kasus: Puskesmas Karangbinangun (Doctoral dissertation, Universitas Muhammadiyah Gresik).