Analisis Algoritma K-Nearest Neighbor untuk Prediksi Stroke pada Pasien Berdasarkan Faktor-Faktro Risiko

Shafa Auliya[1]

Jurusan Ilmu Komputer, Universitas Lampung, Bandar Lampung, Indonesia

email : shafa.auliya21@students.unila.ac.id

1. PENDAHULUAN

Penyakit *stroke* merupakan masalah kesehatan global yang serius dan memiliki dampak yang signifikan. Penyakit *stroke* mengakibatkan kerusakan di bagian otak yang disebabkan oleh penyumbatan (*ischemic)* atau pecahnya pembuluh darah (*hemorrhagic*) akibat aliran darah menuju otak mengalami gangguan. Konsekuensi paling serius yang dapat dialami oleh penderita *stroke* akibat rusaknya pembuluh darah adalah kematian (Dody & Huzaifah, 2021). Data yang ada menunjukan bahwa penyakit *stroke* saat ini menempati peringkat kedua penyebab kematian dan peringkat ketiga penyebab disabilitas didunia (Utama & Nainggolan, 2022). Di Indonesia, penyakit *stroke* juga menjadi penyebab kematian nomor tiga tertinggi setelah penyakit kanker dan jantung (Lishania dkk., 2019).

Kematian akibat *stroke* sulit untuk diperkirakan karena gejala klinisnya tidak terduga dan berkembang dengan cepat (Hartono dkk., 2019). Maka dari itu, diperlukan penggunaan teknologi *machine learning* untuk melakukan prediksi terhadap penyakit *stroke*. Salah satu metode yang dapat digunakan adalah klasifikasi. Beberapa penelitian sebelumnya telah menggunakan teknik *machine learning* untuk memprediksi penyakit *stroke*. Metode klasifikasi yang telah digunakan diantaranya adalah *Logistic Regression*, *Decision Tree*, KNN, *Support Vector Machine* dan *Naive Bayes*.

Berdasarkan penelitian terdahulu oleh Sutomo, dkk (2023) yang membahas mengenai mengoptimalkan algoritma *K-nearest neighbors (KNN)* dalam konteks prediksi stroke. Pada penelitian ini, KNN digunakan sebagai metode prediksi, tetapi penelitian ini berfokus pada optimalisasi performa KNN menggunakan metode siku (elbow method). Metode siku adalah pendekatan yang umum digunakan untuk menentukan jumlah tetangga optimal dalam KNN dengan menganalisis kecepatan perubahan kesalahan model terhadap jumlah tetangga yang berbeda. Dengan menggunakan metode ini mendapatkan nilai akurasi tertinggi sebesar 84%.

Berdasarkan dataset tersebut, maka penelitian ini menerapkan algoritma *K-nearest neighbors (KNN)* untuk melakukan prediksi terhadap penyakit *stroke*. Metode K-Nearest Neighbors (KNN) memiliki kelebihan sebagai pendekatan yang sederhana dan mudah dimengerti, non-parametrik, mampu mengatasi nonlinearitas dalam hubungan antara fitur dan label, serta adaptif terhadap perubahan dalam dataset. Selain itu, KNN tidak bergantung pada asumsi statistik tertentu dan dapat menangani outlier dengan baik. Dengan pendekatan pengelompokan berbasis ketetanggaan, KNN menjadi relevan dalam prediksi penyakit stroke, terutama pada dataset yang mungkin memiliki karakteristik yang bervariasi dan kompleks.

2. METODOLOGI PENELITIAN

2.1 Identifikasi Masalah

Identifikasi masalah dalam penelitian ini adalah bagaimana cara melakukan optimasi algoritma K-Nearest Neighbors (KNN) untuk menentukan nilai k optimum pada dataset *stroke prediction*.

2.2 Pengumpulan Data

Pada penelitian ini menggunakan dataset dari Kaggle yaitu *Stroke Prediction*. Data ini terdiri dari 5110 dataset dengan 12 kolom. Kolom ini terdiri dari *id, gender, age, hypertension, heart disease, ever married, work type, Residence type, avg glucose level, bmi, smoking status, stroke.*

Tabel 1. Deskripsi Dataset *Stroke Prediction*

Kolom	Deskripsi
ID	Pengidentifikasi unik
Gender	Jenis kelamin : Pria, Wanita atau Lainnya
Age	Usia pasien
Hypertension	0 jika pasien tidak menderita hipertensi, 1 jika pasien menderita hipertensi
Heart disease	0 bila pasien tidak mengidap penyakit jantung apa pun, 1 jika pasien mengidap penyakit jantung
Ever married	Tidak atau Ya
Work type	Jenis pekerjaan seperti : anak-anak, Pekerjaan pemerintah, Belum pernah bekerja, Swasta atau Wiraswasta
Residence type	Wilayah tempat tinggal seperti : Pedesaan atau Perkotaan
Avg glucose level	Rata-rata kadar glukosa dalam darah
BMI	Ideks massa tubuh
Smoking status	Status merokok seperti : Sebelumnya merokok, tidak pernah merokok, merokok atau tidak diketahui
Stroke	1 jika pasien terkena stroke atau 0 jika tidak terkena stroke

2.3 *Preprocessing* Data

Data Preprocessing adalah teknik penambangan data awal untuk mengubah data mentah menjadi format dan informasi yang lebih efisien dan berguna. Preprocessing data harus dilakukan dalam proses data mining, karena tidak semua data atau atribut data dalam data digunakan dalam proses data mining. Proses ini dilakukan agar data yang akan digunakan sesuai kebutuhan.

2.4 Impelementasi *K-Nearest Neighboard (KNN)*

Metode K-Nearest Neighbors atau biasa disebut KNN merupakan algoritma klasifikasi yang bekerja dengan mengambil sejumlah K data terdekat (tetangganya) sebagai acuan untuk menentukan kelas dari data baru. Algoritma ini mengklasifikasikan data berdasarkan similarity atau kemiripan atau kedekatannya terhadap data lainnya. Pada model KNN memiliki atribut yang diinisialisasi sebagai k, banyaknya nilai k adalah bilangan bulat positif, bilangan kecil dan ganjil.

Langkah-langkah dalam klasifikasi K-Nearest Neighbor (KNN) adalah:

1. Masukkan nilai gambar konvolusi yang dihitung

2. Menentukan parameter k (jumlah tetangga terdekat).
3. Menghitung proximity berdasarkan model jarak Euclidean terhadap data latih yang diberikan, dengan persamaan :

$$D(x,y) = ||x-y||_2 \sqrt{\Sigma_{j=1}^{N} |x-y|^2}$$

4. Urutkan hasil jarak yang diperoleh secara menaik (berurutan dari nilai tinggi ke rendah).
5. Hitung jumlah masing-masing kelas berdasarkan k tetangga terdekat
6. Kelas mayoritas digunakan sebagai kelas data uji.

2.5 Tahap Evaluasi

Evaluasi dataset prediksi stroke melibatkan sejumlah langkah, termasuk pembagian dataset menjadi data pelatihan dan pengujian, pra-pemrosesan data untuk normalisasi dan enkoding kategori, pemilihan metrik evaluasi seperti akurasi, presisi, *recall*, *F1-score*, *support*, serta pelatihan dan validasi model menggunakan teknik seperti validasi silang. Setelah itu, model dievaluasi pada data pengujian, hasilnya dianalisis, dan kesimpulan diambil mengenai kelayakan model untuk memprediksi *stroke.*

3. HASIL DAN PEMBAHASAN

3.1 *Preprocessing* Data

```
df.fillna({"bmi":df['bmi'].mean().round(1)}, inplace=True)
df['bmi'].head()
```

```
0    36.6
1    28.9
2    32.5
3    34.4
4    24.0
Name: bmi, dtype: float64
```

Gambar 1. Mengisi nilai NaN

Pada proses ini digunakan untuk mengisi nilai-nilai yang hilang dalam kolom 'bmi' dengan nilai rata-rata dari kolom tersebut. Hal ini adalah teknik yang umum digunakan dalam pembersihan dan persiapan data untuk memastikan bahwa data yang hilang diisi dengan nilai yang memadai sebelum dilakukan analisis lebih lanjut.

```
df = df.drop(3116)
df['gender'].value_counts()
```

```
Female    2994
Male      2115
Name: gender, dtype: int64
```

Gambar 2. Melakukan Drop Gender

Pada proses ini digunakan untuk menghapus baris dengan indeks 3116 dari kolom ‘gender’ dan setelah penghapusan tersebut kemudian menghitung frekuensi masing-masing nilai dalam kolom 'gender'.

```
df["gender"] = df["gender"].replace(["Female", "Male"], [1, 0])
df["smoking_status"] = df["smoking_status"].replace(["formerly smoked", "never smoked", "smokes", "Unknown"], [0, 1, 2, 3])
df["ever_married"] = df["ever_married"].replace(["Yes", "No"], [1, 0])
df["work_type"] = df["work_type"].replace(['Private', 'Self-employed', 'Govt_job', 'children', 'Never_worked'], [0, 1, 2, 3, 4])
df["Residence_type"] = df["Residence_type"].replace(['Urban', 'Rural'], [1, 0])
```

Gambar 3. Mengubah Tipedata *Object* menjadi *Int*

- Menggantikan nilai dalam kolom "gender" dengan nilai numerik. Misalnya, "Female" diganti dengan 1 dan "Male" diganti dengan 0.
- Menggantikan nilai dalam kolom "smoking_status" dengan nilai numerik. Misalnya, "formerly smoked" diganti dengan 0, "never smoked" diganti dengan 1, "smokes" diganti dengan 2, dan "Unknown" diganti dengan 3.
- Menggantikan nilai dalam kolom "ever_married" dengan nilai numerik. Misalnya, "Yes" diganti dengan 1 dan "No" diganti dengan 0.
- Menggantikan nilai dalam kolom "work_type" dengan nilai numerik. Misalnya, "Private" diganti dengan 0, "Self-employed" diganti dengan 1, "Govt_job" diganti dengan 2, "children" diganti dengan 3, dan "Never_worked" diganti dengan 4.
- Menggantikan nilai dalam kolom "Residence_type" dengan nilai numerik. Misalnya, "Urban" diganti dengan 1 dan "Rural" diganti dengan 0.

```
df = df[df['smoking_status'] != 3]
```

Gambar 4. Menghilangkan *Unknown* pada Kolom *Smoking* Status

Pada program ini digunakan untuk menghasilkan DataFrame baru yang tidak lagi mengandung baris dengan kategori 'Unknown' dalam kolom 'smoking_status'.

```
X = df.drop(['stroke', 'id'], axis=1).values
y = df[["stroke"]]
```

Gambar 5. Membuang Kolom Id

Pada program ini digunakan untuk menghapus kolom atau baris dari DataFrame yaitu id.

3.2 Hasil Implementasi KNN

Dataset yang telah dilakukan preprocessing data kemudian dilakukan splitting untuk data testing dan data training.

K	Akurasi
1	0.9116409537166901
2	0.94109396914446
3	0.9312762973352033
4	0.9446002805049089
5	0.9417952314165497
6	0.9467040673211781
7	0.9453015427769986
8	0.9467040673211781
9	0.9467040673211781
10	0.9460028050490884

Pada dataset prediksi stroke dengan mengguakan model K-Nearest Neighbor (KNN) yang memiliki nilai akurasi terting yaitu nilai k = 6, 8, 9 dengan nilai akurasi yang sama yaitu 0.9467040673211781.

4. KESIMPULAN

Analisis Algoritma K-Nearest Neighbor untuk prediksi stroke pada pasien berdasarkan faktor-faktor risiko menunjukkan bahwa model ini dapat memberikan prediksi yang akurat dengan mempertimbangkan variabel risiko yang signifikan. Nilai optimal parameter K, validasi model, dan interpretabilitas yang memadai menjadi aspek kunci dalam menentukan keandalan model. Hasil ini memberikan dasar yang kuat untuk penerapan potensial algoritma KNN dalam praktek klinis guna mendukung deteksi dini risiko stroke pada pasien.

DAFTAR PUSTAKA

Lishania, I., Goejantoro, R., & Nasution, Y. N. (2019). Perbandingan Klasifikasi Metode Naive Bayes dan Metode Decision Tree Algoritma (J48) pada Pasien Penderita Penyakit Stroke di RSUD Abdul Wahab Sjahranie Samarinda. *Jurnal Eksponensial*, *10*(2), 135–142.

Hartono, E., Puspitasari, M., & Adam, O. (2019). GAMBARAN TEKANAN DARAH PADA PASIEN STROKE HEMORAGIK DENGAN DIABETES MELITUS DAN NON DIABETES MELITUS DI BAGIAN SARAF RUMKITAL DR.RAMELAN SURABAYA. *Jurnal Sinaps*, *2*(1), 1–8.

Sutomo, F., Muaafii, D. A., Al Rasyid, D. N., Kurniawan, Y. I., Afuan, L., Cahyono, T., Maryanto, E., & 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. https://doi.org/10.52436/1.jutif.2023.4.1.839