https://doi.org/10.37817/Tekinfo.v26i1
P-ISSN: 1411-3635
E-ISSN: 2962-5645

Optimasi Model Algoritma Klasifikasi Data Mining Menggunakan Metode Feature Selection Untuk Prediksi Penyakit Stroke

[1]Mitha Astriani Salwa, [2]Malabay
[1]Teknik Informatika, Universitas Esa Unggul, Jakarta
[1] Teknik Informatika, Universitas Esa Unggul, Jakarta

E-mail: [1]mithasalwa08@gmail.com, [2]malabay@esaunggul.ac.id

ABSTRAK

Meningkatnya prevalensi penyakit degeneratif mendorong pemanfaatan teknologi *data mining* sebagai solusi untuk mengatasi berbagai permasalahan di bidang kesehatan. Penelitian ini bertujuan untuk memmprediksi penyakit stroke secara akurat dengan mengembangkan model algoritma klasifikasi, yaitu *K-Nearest Neighbors* yang dioptimalkan melalui metode *feature selection*, yaitu *Backward Elimination* dan *Forward Selection.* Dataset yang digunakan berasal dari *Kaggle*, terdiri dari 5.110 data dengan 12 atribut: *id, gender, age, hypertension, heart_disease, work_type, Residence_type, avg_glucose_level, bmi, smoking_status,* dan *stroke*. Teknik analisis data yang diterapkan pada penelitian ini adalah CRISP-DM dan diproses menggunakan *Jupyter Notebook* di *Visual Studio Code* dengan bahasa pemrograman Python. Hasil penelitian menunjukkan bahwa penerapan *Backward Elimination* dan *Forward Selection* mampu memberikan peningkatan signifikan terhadap akurasi model dibandingkan model algoritma tanpa fitur. Namun, *Forward Selection* menghasilkan performa lebih optimal dibandingkan *Backward Elimination*. Selain itu, akurasi yang diperoleh dari seluruh skenario dapat dikategorikan sebagai klasifikasi baik karena nilai akurasi nya mencapai lebih dari 90%. Oleh karena itu, penelitian ini memberikan kontribusi dalam mengoptimalkan penerapan algoritma klasifikasi untuk masalah prediksi medis, serta menunjukkan bahwa pemilihan metode *feature selection* yang tepat dapat berpengaruh pada kinerja model klasifikasi.

Kata kunci : ***Data Mining, Klasifikasi, K-Nearest Neighbors, Stroke, Backward Elimination, Forward Selection***

ABSTRACT

The increasing prevalence of degenerative diseases encourages the utilization of data mining technology as a solution to overcome various problems in the health sector. This research aims to accurately predict stroke disease by developing classification algorithm models, namely K-Nearest Neighbors which is optimized through feature selection methods, namely Forward Selection and Backward Elimination. The dataset used comes from Kaggle, consisting of 5,110 data with 12 attributes: id, gender, age, hypertension, heart_disease, work_type, Residence_type, avg_glucose_level, bmi, smoking_status, and stroke. The data analysis in this research was carried out following the CRISP-DM framework and implemented in Jupyter Notebook within Visual Studio Code, utilizing the Python programming language. The findings revealed that incorporating Forward Selection and Backward Elimination did not lead to a notable improvement in model accuracy compared to the algorithm without feature selection. However, Forward Selection produces more optimal performance than Backward Elimination. In addition, the accuracy obtained from all scenarios can be categorized as good classification because it reaches more than

P-ISSN: 1411-3635
E-ISSN: 2962-5645
https://doi.org/10.37817/Tekinfo.v26i1

90%. This study contributes to enhancing the application of classification algorithms in medical prediction challenges and demonstrates that the appropriate choice of feature selection methods can influence the performance of classification models.

Keyword : ***Data Mining, Classification, K-Nearest Neighbors, Stroke, Backward Elimination, Forward Selection***

1. PENDAHULUAN

Seiring meningkatnya prevalensi berbagai penyakit degenaratif salah satunya stroke (Direktorat Promosi Kesehatan dan Pemberdayaan Masyarakat, 2023), penerapan *data mining* sebagai perwujudan pemberdayaan teknologi menjadi semakin krusial. Beberapa penelitian terdahulu telah memanfaatkan pendekatan klasifikasi *data mining* sebagai instrumen untuk mendeteksi penyakit. Data mining, khususnya algoritma klasifikasi, telah banyak digunakan untuk menganalisis dan memprediksi berbagai masalah kesehatan.

Seperti pada penelitian (Halim & Anraeni, 2021), penelitian tersebut dilakukan dengan menerapkan *K-Nearest Neighbors* untuk memprediksi penyakit pneumonia dan hasilnya menunjukkan bahwa *K-Nearest Neighbors* berhasil mengklasifikasikan penyakit tersebut dengan menghasilkan akurasi sebesar 96%.

Namun, tantangan utama dalam pengolahan data kesehatan adalah banyaknya atribut atau fitur yang tersedia, yang dapat menyebabkan overfitting, meningkatkan kompleksitas komputasi, dan menurunkan akurasi model prediksi. Oleh karena itu, diperlukan pendekatan yang mampu mengidentifikasi faktor-faktor yang relevan sekaligus meningkatkan performa prediksi model. Untuk mengatasi hal ini, metode *feature selection* seperti *Backward Elimination* dan *Forward Selection* dapat diterapkan untuk memilih atribut yang paling relevan, sehingga meningkatkan akurasi dan efisiensi model.

Seperti pada penelitian (Angkasa et al., 2023), penelitian tersebut dilakukan dengan menerapkan *Naive Bayes* dengan optimasi dari *Backward Elimination* untuk memprediksi penyakit jantung dan hasilnya menunjukkan bahwa *Backward Elimination* mampu meningkatkan akurasi *Naive Bayes* dari 84,75% menjadi sebesar 98,31%.

Selain itu, telah dilakukan penelitian dengan menerapkan *Naive Bayes* dengan optimasi dari *Forward Selection* untuk memprediksi kanker payudara dan hasilnya menunjukkan bahwa *Forward Selection* mampu meningkatkan akurasi *Naive Bayes* dari 93,57% menjadi sebesar 96,49% (Astuti et al., 2021).

Berdasarkan permasalahan diatas, penulis akan menggabungkan algoritma klasifikasi *K-Nearest Neighbors* bersama metode *feature selection* yaitu *Backward Elimination* dan *Forward Selection* untuk mengidentifikasi faktor-faktor yang relevan sekaligus meningkatkan performa prediksi model. Diharapkan hasil penelitian ini dapat membantu praktisi medis dalam mengambil keputusan yang lebih baik untuk diagnosis dini serta membuka peluang untuk pengembangan sistem deteksi penyakit stroke yang lebih baik di masa depan.

2. LANDASAN TEORI

Data Mining

Data mining memanfaatkan metode dan algoritma khusus untuk menganalisis data, mengidentifikasikan pola tersembunyi, serta menghasilkan wawasan yang mendukung proses

pengambilan keputusan dan pemahaman terhadap data (Tarigan et al., 2022).

Klasifikasi

Klasifikasi adalah proses untuk mengidentifikasi karakteristik serupa pada sejumlah objek dalam basis data, kemudian mengelompokkan objek-objek tersebut ke dalam kategori yang berbeda berdasarkan model klasifikasi yang telah ditentukan (Putry, 2022).

K-Nearest Neighbors

K-Nearest Neighbors (K-NN) bekerja dengan menggunakan jarak terdekat antara data latih dan data uji (Ikhromr et al., 2023).

Algoritma K-NN memiliki sejumlah keunggulan, sifatnya yang non-parametik memungkinkan algoritma ini untuk mengatasi hubungan non-linear antara fitur dan target, serta tidak memerlukan estimasi parameter yang kompleks selama pelatihan. (Sun & Chen, 2021)

Namun, algoritma K-NN juga memiliki beberapa kekurangan yaitu pemilihan parameter K yang harus tepat, sensitivitas tinggi terhadap data *outliers*, ketidakmampuan dalam menangani data yang hilang, dan memiliki keterbatasan dalam menyimpan informasi distribusi fitur. (Nasri & AW, 2020)

Backward Elimination

Backward Elimination adalah metode seleksi fitur yang bekerja dengan menghapus atribut secara bertahap, dimulai dari fitur yang memiliki kontribusi paling kecil terhadap model. Prosesnya dimulai dengan semua fitur yang ada dalam model, kemudian secara iteratif fitur-fitur yang tidak relevan (berdasarkan uji statistik seperti nilai p dalam regresi) akan dihapus sampai tersisa fitur yang paling relevan. (Yunitasari et al., 2021)

Forward Selection

Forward Selection adalah metode yang cara kerjanya adalah dengan memilih fitur-fitur yang relevan dari sejumlah besar opsi yang ada, dan menghapus fitur yang tidak relevan untuk proses klasifikasi. (Rizal et al., 2023)

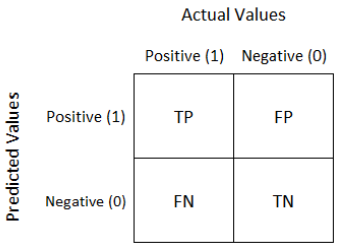
CRISP-DM

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah salah satu metode yang paling sering diterapkan untuk membantu dalam proses analisis data dan proyek *machine learning*. (Kusuma & Wicaksono, 2023) Metode CRISP-DM memiliki enam tahapan yaitu *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment.*

K-Fold Cross Validation

K-Fold Cross Validation adalah teknik evaluasi model dalam *machine learning* yang diterapkan untuk memperoleh penilaian kinerja model yang lebih tepat. Teknik ini membagi data pelatihan menjadi beberapa bagian (*fold*), lalu melatih dan menguji model beberapa kali dengan menggunakan kombinasi data yang berbeda. (Cahyanti et al., 2020)

Confusion Matrix



Gambar 2.1 *Confusion Matrix*

Confusion Matrix adalah sebuah alat untuk menilai performa model klasifikasi dalam pembelajaran mesin. Matriks ini disajikan dalam bentuk tabel yang membandingkan prediksi model dengan data aktual. (Nurul A'ayunnisa et al., 2022) Matriks ini memiliki empat komponen, yaitu:

P-ISSN: 1411-3635
E-ISSN: 2962-5645
https://doi.org/10.37817/Tekinfo.v26i1

a. *True Positive* (TP) : Jumlah data yang sebenarnya positif dan diprediksi oleh model sebagai positif.
b. False Positive (FP) : Jumlah data yang sebenarnya negatif, namun diprediksi sebagai positif oleh model (*Type I Error*).
c. False Negative (FN) : Jumlah data yang sebenarnya positif, namun diprediksi sebagai negatif oleh model (Type II Error).
d. True Negative (TN) : Jumlah data yang sebenarnya negatif dan diprediksi sebagai negatif oleh model.

Dari keempat komponen dalam *Confusion Matrix*, dapat dihitung berbagai metrik evaluasi, seperti akurasi, presisi, sensitivitas, dan *F1-Score*. Berikut ini merupakan persamaan dalam menghitung nilai akurasi, presisi, sensitivitas, dan f1-score.

a. Akurasi

$$\frac{TP+TN}{TP+TN+FP+FN}\ x\ 100\% \qquad (1)$$

b. Presisi

$$\frac{TP}{TP+FP}\ x\ 100\% \qquad (2)$$

c. Sensitivitas

$$\frac{TP}{TP+FN}\ x\ 100\% \qquad (3)$$

d. F1-Score

$$2\ x\ \frac{Precision\ x\ Recall}{Precision+Recall}\ x\ 100\% \qquad (4)$$

Python

Python salah satu bahasa pemrograman *open source* yang penulisan sintaks nya cukup sederhana dan menyediakan berbagai *library* seperti *NumPy*, *Pandas*, *Seaborn*, dan *Matplotlib* untuk memvisualisasikan data. (Nur & Cahyani, 2024)

Source code Python dapat ditulis menggunakan *Integrated Development Environment* (IDE) seperti *Visual Studio Code*, *Sublime Text*, *PyCharm*, atau IDE online seperti *Jupyter Notebook* dan *Google Colab.* (Alfarizi et al., 2023)

3. METODOLOGI

Teknik analisis data yang diterapkan penulis dalam penelitian ini adalah CRISP-DM. Seluruh tahapan dalam penelitian ini dilaksanakan dengan memanfaatkan *Jupyter Notebook* di *Visual Studio Code* dengan bahasa pemrograman Python. Berikut merupakan tahapan dalam metode CRISP-DM yaitu sebagai berikut.

Bussiness Understanding

Pada tahap ini, dilakukan untuk memahami tujuan dari penelitian dan permasalahan yang ingin diselesaikan. Pada penelitian ini, masalah yang ingin diselesaikan adalah bagaimana memprediksi penyakit stroke secara akurat berdasarkan faktor-faktor yang mempengaruhi seseorang terkena stroke.

Data Understanding

Pada tahap ini, dilakukan eksplorasi terhadap dataset yang digunakan untuk memahami struktur dan karakteristik data sebelum masuk ke proses pemrosesan data lebih lanjut. Proses yang akan dilakukan adalah membaca dataset, mengetahui informasi data, dan memvisualisasikan data.

Data Preparation

Pada tahap ini, dilakukan pengolahan data agar siap diterapkan dalam model prediksi. Tahapan yang akan dilakukan adalah *cleaning*, *transformation*, *normalization*, dan *balancing*. Tahapan ini bertujuan untuk meningkatkan kualitas data untuk proses pemodelan.

Modeling

Setelah dataset melewati berbagai tahap pembersihan dan penyesuaian format, Langkah berikutnya adalah membagi dataset menggunakan metode *10-Fold Cross Validation.* Selanjutnya, akan dilakukan pembangunan model klasifikasi dan melakukan perbandingan

https://doi.org/10.37817/Tekinfo.v26i1
P-ISSN: 1411-3635
E-ISSN: 2962-5645

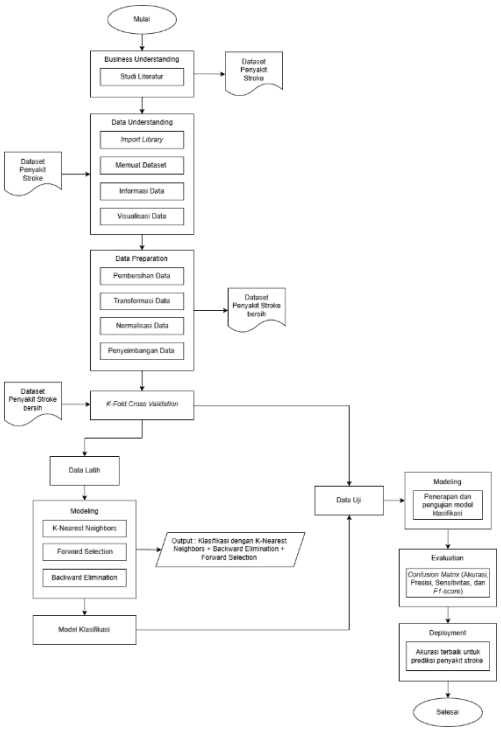
saat menerapkan *K-Nearest Neighbors* dengan dan tanpa optimasi dari *Backward Elimination* dan *Forward Selection*.

Evaluation

Pada tahap ini, dilakukan penilaian terhadap performa model klasifikasi yang dikembangkan untuk memastikan bahwa model dapat menghasilkan prediksi yang akurat dan selaras dengan tujuan penelitian. Salah satu metrik evaluasi yang digunakan adalah *Confusion Matrix*.

Deployment

Pada tahap ini, hasil akhir dari proses pemodelan algoritma klasifikasi akan dituangkan dalam laporan yang berisikan informasi sehingga dapat memberikan manfaat nyata sesuai tujuan penelitian.



Gambar 3.1 Kerangka Berpikir

4. HASIL DAN PEMBAHASAN

Bussiness Understanding

Berdasarkan permasalahan yang telah dipaparkan, penulis ingin menerapkan *data mining* menggunakan model algoritma klasifikasi *K-Nearest Neighbors* yang dikombinasikan dengan metode *feature selection*, yaitu *Backward Elimination* dan *Forward Selection* untuk memilih atribut relevan dan meningkatkan performa model algoritma.

Bussiness Understanding

Pada penelitian ini, dataset yang akan diteliti bersumber dari platform *Kaggle* oleh Fedesoriano dan berformatkan (csv). Dataset ini berisikan 5.110 data pasien stroke yang terdiri dari 249 pasien pengidap stroke dan 4.861 pasien bukan stroke dengan 12 atribut, yaitu *id, gender, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status* dan *stroke*. Berikut merupakan tahapan dari *data understanding*.

a. Import Library
   Mengimpor pustaka (*library*) yang diperlukan dalam pengolahan data, pemilihan fitur dan pembangunan model klasifikasi. Beberapa pustaka utama yang digunakan adalah *Pandas*, *NumPy*, *Scikit-learn*, *Matplotlib*, dan *Seaborn.*
b. Memuat Dataset

	id	gender	age	hypertension	heart disease	ever married	work type	Residence type	avg glucose level	bmi	smoking status	stroke
0	9046	Male	67.0	0	1	Yes	Private	Urban	228.69	36.6	formerly smoked	1
1	51676	Female	61.0	0	0	Yes	Self-employed	Rural	202.21	NaN	never smoked	1
2	31112	Male	80.0	0	1	Yes	Private	Rural	105.92	32.5	never smoked	1
3	60182	Female	49.0	0	0	Yes	Private	Urban	171.23	34.4	smoked	1
4	1665	Female	79.0	1	0	Yes	Self-employed	Rural	174.12	24.0	never smoked	1
...	...	...	...	...	...	...	...	...	...	...	...	...
5105	18234	Female	80.0	1	0	Yes	Private	Urban	83.75	NaN	never smoked	0
5106	44873	Female	81.0	0	0	Yes	Self-employed	Urban	125.20	40.0	never smoked	0
5107	19723	Female	35.0	0	0	Yes	Self-employed	Rural	82.99	30.6	never smoked	0
5108	37544	Male	51.0	0	0	Yes	Private	Rural	166.29	25.6	formerly smoked	0
5109	44679	Female	44.0	0	0	Yes	Govt_job	Urban	85.28	26.2	Unknown	0

5110 rows × 12 columns

Gambar 4.1 Memuat Dataset

Setelah library diimport, proses selanjutnya adalah memuat dataset dengan menggunakan fungsi *Pandas*.

c. Informasi Data
   Informasi data mencakup deskripsi lengkap terkait variabel-variabel yang terdapat dalam dataset, jumlah sampel (baris) dan atribut (kolom), serta tipe data dari setiap variabel, baik numerik maupun kategorik.

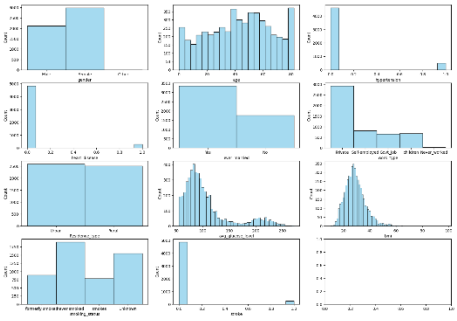
P-ISSN: 1411-3635
E-ISSN: 2962-5645
https://doi.org/10.37817/Tekinfo.v26i1

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5110 entries, 0 to 5109
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   id                 5110 non-null   int64
 1   gender             5110 non-null   object
 2   age                5110 non-null   float64
 3   hypertension       5110 non-null   int64
 4   heart_disease      5110 non-null   int64
 5   ever_married       5110 non-null   object
 6   work_type          5110 non-null   object
 7   Residence_type     5110 non-null   object
 8   avg_glucose_level  5110 non-null   float64
 9   bmi                4909 non-null   float64
 10  smoking_status     5110 non-null   object
 11  stroke             5110 non-null   int64
dtypes: float64(3), int64(4), object(5)
memory usage: 479.2+ KB
```

Gambar 4.2 Informasi Data

d. Visualisasi Data

Visualisasi data digunakan untuk memahami identifikasi distribusi data dan mendeteksi outlier, *missing value*, anomali yang dapat mempengaruhi analisis dan model klasifikasi. Berikut ini adalah diagram yang menunjukkan persebaran data pada setiap variabel.



Gambar 4.3 Visualisasi Data

Data Preparation

Proses *data preparation* pada penelitian ini mencakup beberapa serangkaian proses yaitu *cleaning*, *transformation*, *normalization*, dan *balancing*.

a. Cleaning

Proses pertama yang dilakukan saat pembersihan data adalah penghapusan variabel tidak relevan yaitu *id* karena variabel tersebut hanya berisikan kode identitas dari pasien stroke.

Setelah itu, proses selanjutnya adalah menghapus nilai kosong atau hilang (*missing value*) sebanyak 201 yang terdapat pada atribut *bmi* dengan memanfaatkan fungsi drop dari pustaka *Pandas.*

Proses selanjutnya adalah penghapusan data anomali "Other" pada variabel *gender* dengan memanfaatkan fungsi *drop* dari pustaka *Pandas*.

Kemudian proses selanjutnya adalah menghapus data *outliers* pada variabel *bmi* dan *avg_glucose_level* dalam dataset penyakit stroke menggunakan metode *Interquartile Range* (IQR). Nilai yang berada di luar rentang $Q_1 - 1.5\ x\ IQR$ atau $Q_3 + 1.5\ x\ IQR$ dianggap sebagai *outliers*.

b. Transformation

Transformasi ini bertujuan untuk mengkonversi data kategorikal menjadi format numerik menggunakan fungsi *LabelEncoder()* dari pustaka *sklearn* agar dapat diproses oleh algoritma klasifikasi seperti *K-Nearest Neighbors* (KNN) yang hanya dapat bekerja dengan data numerik.

Atribut	Sebelum Transformasi	Sesudah Transformasi
gender	Female, Male	Female = 0, Male = 1
ever_married	Yes, No	Yes = 1, No = 0
work_type	*govt_jb, never_worked, private, self-employed, children*	*govt_jb = 0, never_worked = 1, private = 2, self-employed = 3, children = 4*
Residence_type	Rural, Urban	Rural = 0, Urban = 1
smoking_status	formerly smoked, never smoked, smokes, unknown	formerly smoked = 0, never smoked = 1, smokes =2, unknown = 3

Tabel 4.1 Transformasi Data

c. Normalization

Proses normalisasi data numerik pada dataset penyakit stroke menggunakan metode *Z-Score* dan memanfaatkan fungsi *StandardScaler* dari pustaka *sklearn*. Normalisasi bertujuan untuk mengubah nilai data menjadi skala yang seragam sehingga dapat

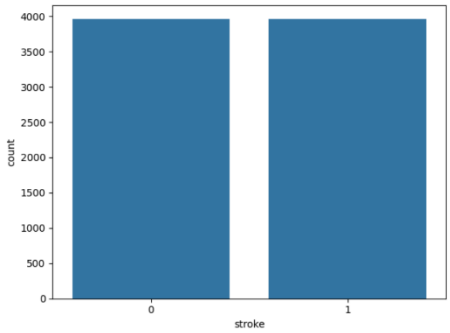
https://doi.org/10.37817/Tekinfo.v26i1
P-ISSN: 1411-3635
E-ISSN: 2962-5645

meningkatkan kinerja algoritma klasifikasi.

Gambar 4.4 Normalisasi Data

d. Balancing

Pada tahap ini, dilakukan penanganan terhadap dataset yang tidak seimbang menggunakan *Synthetic Minority Oversampling Technique* (SMOTE). Setelah dilakukan pembersihan dataset berjumlah 4.088 data yang kemudian bertambah menjadi 7.916 data setelah dilakukan *oversampling*.



Gambar 4.5 Data Seimbang

Modeling

Setelah dataset melewati berbagai tahap pembersihan dan penyesuaian format, tahap selanjutnya adalah membagi dataset menggunakan teknik *10-Fold Cross Validation*.

Setelah itu, dilakukan pembangunan model klasifikasi untuk prediksi penyakit stroke menggunakan menggunakan *K-Nearest Neighbors.* Untuk meningkatkan akurasi dan efisiensi model, diterapkan metode seleksi fitur, yaitu *Backward Elimination* dan *Forward Selection*, guna memilih fitur-fitur yang paling relevan dalam prediksi. Berikut ini merupakan *pseudocode* tiap skenario.

```
Pseudocode K-Nearest Neighbors

Input values:
D = Dataset
K = Number of neighbors
T = Target data point
Output:
Predicted label for T

Process:
1. For each data point in D:
   - Calculate the distance between T and D (using distance metrics)
2. Sort all data points in D based on the calculated distance to T
3. Select the K closest data points (neighbors) to T
4. Count the frequency of each class label among the K neighbors
5. Assign the most frequent class label from the K neighbors to T
6. Return the predicted label for T using Best_Features
```

Gambar 4.6 *Pseudocode K-Nearest Neighbors*

```
Pseudocode K-Nearest Neighbors & Backward Elimination

Input values:
D = Dataset
K = Number of neighbors
T = Target data point
Features = List of all available features in D
Output:
Predicted label for T

Process:
1. Initialize Best_Accuracy = 0
2. Initialize Best_Features = Features (start with all features)
3. For each feature f in Features:
   a. Create a subset of Features excluding f
   b. For each data point in D:
      - Calculate the distance between T and D (using the subset of Features excluding f)
   c. Sort the data points in D based on the calculated distance to T
   d. Select the K closest data points (neighbors) to T
   e. Count the frequency of each class label among the K neighbors
   f. Assign the most frequent class label from the K neighbors to T
   g. Calculate accuracy based on the predicted label for T
   h. If accuracy > Best_Accuracy:
      - Set Best_Accuracy = accuracy
      - Set Best_Features = subset of Features excluding f
4. Return the predicted label for T using Best_Features
```

Gambar 4.7 *Pseudocode K-Nearest Neighbors & Forward Selection*

P-ISSN: 1411-3635
E-ISSN: 2962-5645
https://doi.org/10.37817/Tekinfo.v26i1

```
Pseudocode K-Nearest Neighbors & Forward Selection

Input values:
D = Dataset
K = Number of neighbors
T = Target data point
Features = List of all available features in D
Output:
Predicted label for T

Process:
1. Initialize Best_Accuracy = 0
2. Initialize Best_Features = []
3. For each feature f in Features:
    a. Select a subset of Features including f
    b. For each data point in D:
        - Calculate the distance between T and D (using the subset of Features including f)
    c. Sort the data points in D based on the calculated distance to T
    d. Select the K closest data points (neighbors) to T
    e. Count the frequency of each class label among the K neighbors
    f. Assign the most frequent class label from the K neighbors to T
    g. Calculate accuracy based on the predicted label for T
    h. If accuracy > Best_Accuracy:
        - Set Best_Accuracy = accuracy
        - Set Best_Features = subset of Features including f
4. Return the predicted label for T using Best_Features
```

Gambar 4.8 *Pseudocode K-Nearest Neighbors & Backward Elimination*

Evaluation

Pada tahap ini, dilakukan evaluasi dengan menerapkan *Confusion Matrix* untuk mengukur model algoritma dalam menghasilkan nilai prediksi yang akurat.

a. Klasifikasi K-Nearest Neighbors

Dari evaluasi yang telah dilakukan, menghasilkan nilai-nilai metriks seperti pada tabel dibawah ini.

Confusion Matrix		*Nilai Sebenarnya*	
		True	False
Nilai Prediksi	True	356	44
	False	1	391

Tabel 4.2 Hasil *Confusion Matrix* Model Klasifikasi *Random Forest*

Dari hasil *Confusion Matrix* tersebut, dilakukanlah perhitungan beberapa metriks evaluasi dan dihasilkan akurasi sebesar 94,32%, presisi sebesar 89,89%, sensitivitas sebesar 99,74% dan *f1-score* sebesar 94,56%.

b. Klasifikasi K-Nearest Neighbors + Backward Elimination

Proses klasifikasi *K-Nearest Neighbors* yang dikombinasikan dengan *Backward Elimination* menghasilkan 8 atribut berpengaruh yaitu *gender*, *age*, *hypertension*, *ever_married*, *work_type*, *Residence_type*, *avg_glucose_level*, *smoking_status*. Setelah dilakukan proses evaluasi, dihasilkan nilai-nilai metriks seperti pada tabel dibawah ini.

Confusion Matrix		Nilai Sebenarnya	
		True	False
Nilai Prediksi	True	354	46
	False	2	390

Tabel 4.3 Hasil *Confusion Matrix* Model Klasifikasi *Random Forest*

Dari hasil *Confusion Matrix* tersebut, dilakukan perhitungan beberapa metriks evaluasi dan dihasilkan akurasi sebesar 93,94%, presisi sebesar 89,45%, sensitivitas sebesar 99,49% dan *f1-score* sebesar 94,20%.

c. Klasifikasi K-Nearest Neighbors + Forward Selection

Proses klasifikasi *K-Nearest Neighbors* yang dikombinasikan dengan *Forward Selection* menghasilkan 9 atribut berpengaruh

P-ISSN: 1411-3635
E-ISSN: 2962-5645

yaitu *gender, age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, smoking_status*. Dari evaluasi yang telah dilakukan, menghasilkan nilai-nilai metriks seperti pada tabel dibawah ini.

Confusion Matrix		Nilai Sebenarnya	
		True	False
Nilai Prediksi	True	6.410	523
	False	21	7.029

Tabel 4.4 Hasil *Confusion Matrix* Model Klasifikasi *Random Forest*

Dari hasil *Confusion Matrix* tersebut, dilakukan perhitungan beberapa metriks evaluasi dan dihasilkan akurasi sebesar 94,06%, presisi sebesar 90,27%, sensitivitas sebesar 99,01% dan *f1-score* sebesar 94,44%.

5. KESIMPULAN

Kesimpulan setelah dilakukan penelitian ini adalah model algoritma *K-Nearest Neighbors* tanpa fitur menghasilkan akurasi tertinggi yaitu 94,32%. Sedangkan hasil dengan kombinasi *feature selection*, yaitu *Backward Elimination* dan *Forward Selection* hanya menghasilkan nilai akurasi masing-masing sebesar 93,94% dan 94,06%. Hal ini menunjukkan bahwa kedua *feature selection* tersebut tidak meningkatkan akurasi terhadap model algoritma *K-Nearest Neighbors* secara signifikan untuk memprediksi penyakit stroke. Selain itu, dari hasil tersebut juga dapat disimpulkan bahwa *Forward Selection* mampu bekerja lebih baik dibandingkan *Backward Elimination*. Meskipun begitu, hasil akurasi dari ketiga skenario masih termasuk dalam kategori klasifikasi baik karena nilainya mencapai lebih dari 90%.

6. UCAPAN TERIMA KASIH

Penulis mengucapkan rasa terima kasih yang mendalam kepada semua pihak yang telah memberikan dukungan, arahan, dan motivasi selama proses penyusunan jurnal ilmiah ini. Pihak-pihak tersebut meliputi :

a. Malabay, S.Kom, M.Kom, selaku Dosen Pembimbing yang telah memberikan panduan, masukan, serta saran yang sangat bernilai.
b. Orang tua dan keluarga yang telah memberikan dukungan, doa, dan semangat.
c. Teman-teman seperjuangan yang telah membantu dan memberikan dukungan kepada penulis

DAFTAR PUSTAKA

Alfarizi, M. R. S., Al-farish, M. Z., Taufiqurrahman, M., Ardiansah, G., & Elgar, M. (2023). Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning. *KARIMAH TAUHID*, *2*(1), 1–6.

Angkasa, J. W., Noersasongko, E., & Purwanto. (2023). DATA PRE-PROCESSING AND FEATURE SELECTION TECHNIQUES BACKWARD ELIMINATION FOR NAIVE BAYES CLASSIFICATION ON HEART DISEASE DETECTION. *Jurnal Ekonomi Teknologi & Bisnis (JETBIS)*, *2*(4), 334–342.

Astuti, L. W., Saluza, I., Faradilla, F., & Alie, M. F. (2021). Optimalisasi Klasifikasi Kanker Payudara Menggunakan Forward Selection pada Naive Bayes. *Jurnal Ilmiah Informatika Global*, *11*(2). https://doi.org/10.36982/jiig.v11i2.1235

Cahyanti, D., Rahmayani, A., & Husniar, S. A. (2020). Analisis performa metode Knn pada Dataset pasien pengidap Kanker Payudara. *Indonesian Journal of Data and Science*, *1*(2), 39–43. https://doi.org/10.33096/ijodas.v1i2.13

Direktorat Promosi Kesehatan dan Pemberdayaan Masyarakat. (2023, October 6). *Kenali Stroke dan Penyebabnya*. Https://Ayosehat.Kemkes.Go.Id/.

Halim, A. A. D., & Anraeni, S. (2021). Analisis Klasifikasi Dataset Citra Penyakit Pneumonia menggunakan Metode K-Nearest Neighbor (KNN). *Indonesian Journal of Data and Science*, *2*(1), 01–12. https://doi.org/10.33096/ijodas.v2i1.23

Ikhromr, F. N., Sugiyarto, I., Faddillah, U., & Sudarsono, B. (2023). Implementasi Data Mining Untuk Memprediksi Penyakit Diabetes Menggunakan Algoritma Naives Bayes dan K-Nearest Neighbor. *INTECOMS: Journal of Information Technology and Computer Science*, *6*(1), 416–428. https://doi.org/10.31539/intecoms.v6i1.5916

Nasri, E., & AW, A. S. (2020). APLIKASI SELEKSI PENENTUAN NASABAH UNTUK PENJUALAN BARANG SECARA KREDIT DENGAN ALGORITMA K-NEAREST NEIGHBOR. *Jurnal Ilmiah Sains Dan Teknologi*, *4*(1), 1–11.

Nur, A., & Cahyani, P. R. (2024). EVALUASI PENGGUNA JUPYTER NOTEBOOK PADA PYTHON DALAM PEMBELAJARAN DATA SCIENCE (STUDI KASUS : KAPAL TITANIC). *Kohesi: Jurnal Sains Dan Teknologi*, *4*(10), 21–30.

Nurul A'ayunnisa, Salim, Y., & Azis, H. (2022). Analisis Performa Metode Gaussian Naïve Bayes untuk Klasifikasi Citra Tulisan Tangan Karakter Arab. *Indonesian Journal of Data and Science*, *3*(3), 115–121. https://doi.org/10.56705/ijodas.v3i3.54

Putry, N. M. (2022). KOMPARASI ALGORITMA KNN DAN NAÏVE BAYES UNTUK KLASIFIKASI DIAGNOSIS PENYAKIT DIABETES MELLITUS. *EVOLUSI : Jurnal Sains Dan Manajemen*, *10*(1). https://doi.org/10.31294/evolusi.v10i1.12514

Rizal, M., Syahaf, M. Z., Priyambodo, S. R., & Rhamdani, Y. (2023). OPTIMASI ALGORITMA NAÏVE BAYES MENGGUNAKAN FORWARD SELECTION UNTUK KLASIFIKASI PENYAKIT GINJAL KRONIS. *Naratif : Jurnal Nasional Riset, Aplikasi Dan Teknik Informatika*, *5*(1), 71–80. https://doi.org/10.53580/naratif.v5i1.200

Sun, B., & Chen, H. (2021). A Survey of *k* Nearest Neighbor Algorithms for Solving the Class Imbalanced Problem. *Wireless Communications and Mobile Computing*, *2021*(1). https://doi.org/10.1155/2021/5520990

Tarigan, P. M. S., Hardinata, J. T., Qurniawan, H., Safii, M., & Winanjaya, R. (2022). IMPLEMENTASI DATA MINING MENGGUNAKAN ALGORITMA APRIORI DALAM MENENTUKAN PERSEDIAAN BARANG (STUDI KASUS : TOKO SINAR HARAHAP). *JUST IT (Jurnal Sistem Informasi, Teknologi Informasi Dan Komputer)*, *12*(2), 51–61.

Yunitasari, Hopipah, H. S., & Mayasari, R. (2021). Optimasi Backward Elimination untuk Klasifikasi Kepuasan Pelanggan Menggunakan Algoritme k-nearest neighbor (k-NN) and Naive Bayes. *Technomedia Journal*, *6*(1), 99–110. https://doi.org/10.33050/tmj.v6i1.1531