Perbandingan Metode Naïve Bayes Dan Random Forest Pada Deteksi Penyakit Stroke Menggunakan Teknik SMOTE (Synthetic Minority Over-Sampling Technique)

Muhammad Rico Salahuddin[1], Yuni Yamasari[2]

[1,2] Jurusan Teknik Informatika/Teknik Informatika, Universitas Negeri Surabaya

[1]muhammad.19053@mhs.unesa.ac.id
[2]yuniyamasari@unesa.ac.id

***Abstrak*— Stroke disebabkan karena kurangnya pasokan darah ke otak dan terjadinya penyumbatan di dalam otak atau terjadinya pecahnya pembuluh darah di dalam otak. Selain memiliki dampak kesehatan personal, stroke juga membebani negara Indonesia dalam pembiayaan kesehatan dan masih menjadi faktor penyebab kematian terbesar di indonesai maupun di dunia. Dengan pesatnya perkembangan teknologi dan adanya model sistem deteksi penyakit stroke diharapkan dapat berkontribusi secara signifikan terhadap pencegahan dan perawatan penyakit stroke secara dini. Oleh karena itu penelitian ini dilakukan dengan menggunkan metode *naïve bayes* dan *random forest*. Hal ini untuk membandingkan kinerja kedua metode yang digunakan dalam penelitian ini dan juga pada penelitian ini juga menerapkan teknik SMOTE dikarenakan ditemukannya *imbalance class* dalam dataset yang digunakan. Setelah melalui proses tahapan alur system yang sudah ditentukan dan dapat disimpulkan bahwa metode *random forest* lebih optimal daripada metode *naïve bayes* dalam deteksi penyakit stroke yang menggunakan dataset pada penelitian ini dengan pembagian 10% data *test* dan 90% data *train* random forest mendapatkan score akurasi sebesar 95% sedangkan naïve bayes mendapatkan score sebesar 79% . Meskipun terjadi penurunan *score* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE yang tidak signifikan pada aspek akurasi pada confusion matrix akan tetapi terjadi kenaikan kinerja yang signifikan pada aspek precission, recall dan f1 score.**

***Kata Kunci*— Stroke, Deteksi, *Random Forest*, *Naive Bayes*, Teknik SMOTE.**

I. PENDAHULUAN.

Stroke merupakan penyakit yang menyerang otak, kondisi tersebut disebabkan karena kurangnya pasokan darah ke otak dan terjadinya penyumbatan di dalam otak atau terjadinya pecahnya pembuluh darah di dalam otak. Tanpa adanya pasokan darah yang membawa oksigen dan nutrisi ke otak membuat sel-sel pada sebagian area otak akan tidak berfungsi sebagaimana mestinya [2]. Ada beberapa gejala yang menunjukan orang tersebut menderita penyakit stroke seperti bentuk wajah yang tidak simetris, mengalami kesulitan dalam berbicara, keterbatasan dalam melakukan aktifitas yang membutuhkan otot tubuh seperti mengangkat kedua lengan dan lain-lain [3] . Pada usia senja penyakit stroke sering dijumpai, tak juga sebagai menjadi salah satu faktor penyebab kematian dan menjadi penyebab kecacatan pada usia produktif.

Dengan adanya pola hidup yang tidak sehat dan benyaknya pekerjaan dapat memicu beban pikiran yang dapat mengakibatkan stress, merokok, dan meminum miniuman alcohol yang akan memiliki dampat pada tingginya faktor resiko terkena penyakit stroke [6] . Selain memiliki dampak kesehatan personal, stroke juga membebani negara Indonesia dalam pembiayaan kesehatan. Berdasarkan data yang telah dikumpulkan BPJS dimana terjadi 1,7 juta kasus dengan penderita penyakit stroke telah menghabiskan dana sebesar Rp 2,1 Triliun pada tahun 2020 nilai ini akan terus bertambah seiringnya bertambahnya penderita penyakit stroke.

Pesatnya perkembangan teknologi dapat membantu manusia dalam memecahkan segala masalah yang ada dan banyaknya kebutuhan yang semakin kompleks. Dengan adanya model yang dapat melakukan deteksi penyakit stroke, diharapkan dapat berkontribusi secara signifikan terhadap pencegahan dan perawatan penyakit stroke secara dini, dengan tingginya volume data, heterogenitas dan kompleksitas dataset akan menjadi tantangan dalam melakukan deteksi penyakit stroke. Deteksi penyakit stroke bertujuan untuk mengurangi potensi kematian dan kecacatan pada manusia yang disebabkan oleh penyakit stroke.

Algoritma Naïve bayes adalah klasifikasi metode yang didasarkan pada teorama bayes. Metode pengklasifikasian menggunakan probabilitas dan statistic yang dikemukakan oleh Thomas bayes seorang ilmuwan inggris , dengan sederhananya dan efektifnya naïve bayes dapat dengan cepat untuk membangun model dan membuat model prediksi menggunakan naïve bayes [5]. Algoritma Naïve bayes cocok digunakan pada volume data yang tinggi serta data yang kosong dan dapat menangani data yang tidak sama dan adanya eror pada data tersebut.

Sedangkan algoritma random forest adalah sekumpulan decision tree/pohon keputusan yang merupakan kombinasi dari masing-masing decision tree yang kemudian dijadikan satu model. Penggunaan pohon (tree) yang semakin banyak akan membuat tingkat akurasi menjadi lebih baik [7]. Random forest dibuat dengan sejumlah pohon keputusan, dan setiap pohon memperoleh efek pengaturan posisinya dengan memanfaatkan klasifikasi yang tidak sama. Metode ini memungkinkan evaluasi dari alokasi pengambilan sampel menggunakan teknik pengambilan sampel acak [4].

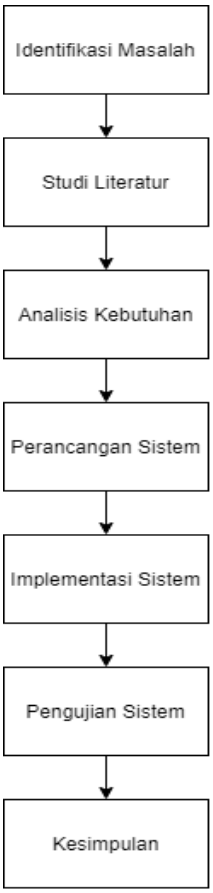
Pada penelitian sebelumnya mengenai perbandingan jaringan syaraf tiruan dan naïve bayes dalam deteksi seseorang terkena penyakit stroke hasil dari penelitiannya yaitu metode Jaringan Syaraf Tiruan mendapatkan nilai akurasi sebesar 71.11% dan naïve bayes sebesar 80.555%[6] kemudian penelitianyang menggunaan teknik SMOTE sebagai penyeimbang kelas pada klasifikasi data mining, dalam penelitian tersebut menghasilkan bahwa model klasifikasi yang telah diseimbangkan dengan metode SMOTE memiliki performa lebih baik dibandingkan jika tidak dilakukan penyeimbangan data[1].

Berdasarkan penelitian yang telah dilakukan oleh peneliti sebelumnya maka pada penelitian ini penulis ingin membandingkan 2 metode pada machine learning yaitu naïve bayes dan random forest. Naïve bayes dan random forest memiliki kelebihan dan kekurangan sendiri dalam melakukan proses prediksi atau klasifikasi, begitu pula dengan metode machine learning lainnya. Untuk meningkatkan performa model yang lebih baik penulis menggunakan teknik SMOTE

agar kelas data terdapat keseimbangan di dalamnya dan dapat menghasilkan performa yang bagus pada kedua metode tersebut Berdasarkan permasalahan dan uraian yang ada di atas maka dari itu penulis mengambil judul "Perbandingan Metode Naïve Bayes dan Random Forest pada Deteksi Penyakit Stroke Menggunakan Teknik SMOTE (Synthetic Minority Over-sampling Technique)" dalam penelitian ini.

II.METODOLOGI PENELITIAN

Metodologi penelitian adalah serangkaian kegiatan yang dilakukan untuk penelitian dalam skripsi "Perbandingan Metode Naïve Bayes dan Random Forest pada Deteksi Penyakit Stroke Menggunakan Teknik SMOTE (Synthetic Minority Over-sampling Technique)" guna mendapatkan hasil sesuai yang diharapkan dalam tujuan penelitian tentunya dengan batasan yang sudah ditentukan, dengan langkah langkahnya antara lain Identifikasi masalah, kajian pustaka, Analisis Kebutuhan, Perancangan model, Implementasi, dan pengujian model. Metode yang diterapkan pada penelitian ini adalah metode kuantitatif , berikut adalah langkah-langkah penelitian.



Gambar 1 alur penelitian

A. Identifikasi masalah

Identifikasi masalah atau menentukan masalah yang menjadi dasar dari penelitian. Dalam penelitian ini adalah mengenai perbandingan metode naïve bayes dan random forest dalam deteksi penyakit stroke, dimana peneliti akan melakukan perancangan dan pengimplementasian algoritma machine learning pada dataset yang bertujuan untuk mendeteksi apakah user menderita penyakit stroke atau tidak.

B. Studi Literatur

Pencarian terhadap sumber sumber literatur yang berhubungan dengan penelitian yang dilakukan sebagai pendukung dan juga acuan penelitian. Literatur yang dicari dan digunakan dalam penelitian ini berhubungan dengan penggunaan algoritma machine learning yang digunakan untuk mendeteksi penyakit stroke yang ada pada dataset yang didapat dari berbagai macam sumber seperti buku, artikel, jurnal internasional maupun nasional.

C. Analisis kebutuhan

Analisis kebutuhan merupakan analisis yang dibutuhkan untuk menentukan detail kebutuhan pada penelitian perbandingan metode naïve bayes dan random forest dalam deteksi penyakit stroke. Penelitian ini menggunakan metode naïve bayes dan random forest dimana kedua metode tersebut akan dibandingkan hasil kinerja kedua metode tersebut pada dataset yang digunakan dan mencari metode yang terbaik diantara kedua metode tersebut yang cocok digunakan dalam deteksi penyakit stroke menggunakan dataset yang digunakan dalam penelitian ini

C.1. Kebutuhan perangkat keras (Hardware)

Perangkat keras yang digunakan selama proses penelitian berlangsung untuk mewujudkan tujuan penelitian yaitu perangkat laptop sebagai uji coba dengan spesifikasi berikut :

Prosesor	: Intel Core i5-5200u
RAM	: 8 GB
Penyimpanan	: SSD 128 GB
Sistem Operasi	: Windows 10 Enterprise 64-bit

C.2. Kebutuhan perangkat lunak (Software)

Perangkat lunak memilliki fungsi digunakan dalam pengoperasian sistem pada penelitian perbandingan metode naïve bayes dan random forest dalam deteksi penyakit stroke ini. Pada penelitian ini digunakan adalah browser Google Chrome versi 107.0.5304.88 (Official Build) (64-bit) yang digunakan sebagai pencariaan informasi yang dibutuhkan guna melancarkan penyelesaian peneltian ini, sedangkan untuk proses coding peneliti menggunakan aplikasi anaconda navigator (anaconda3) yang didalamnya terdapat jupyter notebook 6.4.8.

D. Perancangan model

Model yang akan dirancang dan digunakan dalam penelitian ini bertujuan untuk menggunakan algoritma machine learning untuk pendeteksian penyakit stroke pada dataset. Berikut adalah gambaran proses dari model yang akan dibuat :

D.1. Input dataset

Langkah pertama yaitu mempersiapkan dataset yang digunakan. Pada penelitian ini dataset yang digunakan merupakan dataset public yang didapat dari kaggle. Untuk jumlah dataset itu sendiri ialah sebanyak 5010 data.

D.2. Preprocessing data

Sebelum data dilatih ke metode naïve bayes dan random forest data harus disiapkan terlebih dahulu agar tidak terjadi eror atau sesuatu hal yang membuat hasil evaluasi tidak maksimal dan akhirnya membuat model yang tidak maksimal untuk deteksi penyakit stroke.

D.3. Teknik SMOTE

Gambar 2 Alur perancangan model

Setelah melakukan proses preprocessing data peneliti menemukan kendala yaitu terdapat tidak seimbangnya data antara positif stroke dengan negative stroke hal ini dapat menyebabkan kurangnya tingkat akurasi model dalam melakukan deteksi data terhadap data minoritas maka peneliti menggunakan teknik SMOTE untuk mengatasi yang tidak seimbang sehingga dapat membuat model yang lebih akurat ketika mendeteksi positif stroke maupun negative stroke

D.4. Membagi data test dan data train

Setelah melakukan langkah-langkah yang terdapat diatas selanjutnya dataset dibagi menjadi 2 yaitu data test dan data train. Untuk pembagian yang dilakukan peneliti pada penelitian ini dibagi menjadi 5 scenario pembagian data test dan data train untuk melihat kinerja metode naïve bayes dan random forest pada pembagian data test dan data train mana yang terbaik pada penelitian ini.

D.5. Data test dan data train dilatih kedalam metode naïve bayes dan random forest.

Setelah dilakukannya pembagian data set ke dalam data test dan data train selanjutnya data train dan data test akan dilatih ke dalam 2 metode yang digunakan.

D.6. Evaluasi

Evaluasi yang dilakukan pada penelitian ini menggunakan confusion matrix dan k-fold cross validation dengan menggunakan 5 scenario pembagian data test dan data train berikut untuk pembagian 5 scenario.

a. Scenario 1 dengan pembagian 10% data test dan 90% data train
b. Scenario 2 dengan pembagian 20% data test dan 80% data train
c. Scenario 3 dengan pembagian 30% data test dan 70% data train
d. Scenario 4 dengan pembagian 40% data test dan 60% data train
e. Scenario 5 dengan pembagian 50% data test dan 50% data train

E. Implementasi

Hasil dari proses perancangan model akan diimplementasikan sesuai dengan alur perancangan model yang telah dibuat. Implementasi alur perancangan model dibuat dalam jupyter notebook 6.4.8 dan menggunakan bahasa pemrogaman python. Proses coding dan pemilihan algoritma disesuaikan dengan tujuan penelitian agar tetap berjalan sesuai tujuan.

F. Pengujian

Setelah model sudah diimplementasikan, model akan diuji dengan dataset yang sudah ditentukan. Parameter pengujian yang akan dilakukan adalah menguji dataset dan dioperasikan menggunakan model yang sudah dibuat, kemudian akan dinilai bagaimana kinerja metode yang digunakan pada deteksi penyakit stroke di dataset yang digunakan. Penilaian metode yang digunakan menggunakan confusion matix dan k-fold cross validation.

G. Kesimpulan

Setelah dilakukan serangkaian perencanaan, implementasi, pengujian pada sistem, maka akan didapati kesimpulan dan pemberian saran dari serangkaian penelitian yang telah dilakukan berdasarkan rumusan masalah dan batasan, hasil dari pengujian sistem dapat menjadi jawaban dari masalah yang dijelaskan serta dataset dari penelitian, yang diharapkan nantinya dapat berguna dalam sistem deteksi penyakit stroke dan dapat menjadi referensi penelitan yang selanjutnya.

III. Hasil dan Pembahasan

A. Deskripsi Data

Dataset yang digunakan dalam peneltian ini berupa data public yang didapat dari kaggle dengan total data sebanyak 5109 data dalam bentuk file excel dan memiliki 11 atribut yang diantaranya yaitu :

Id	: Identity Document
Gender	: Jenis Kelamin
Age	: Umur
Hypertension	: Tekanan Darah Tinggi
Heart Disease	: Penyakit Jantung
Ever Married	: Pernikahan
Avg Glucose Level	: Kadar Gula

BMI : Berat Tubuh
Smoking Status : Status Merokok
Residence Type : Tempat Tinggal

1. Id : unique identifier
2. *Gender* : "*Male*", "*Female*" ,dan "*Other*"
3. *Age* : Usia dari pasien yang ada di dalam dataset
4. *Hypertension* : "0" artinya pasien tidak hipertensi dan "1" artinya pasien memiliki hipertensi
5. *Heart_disease* : "0" artinya pasien tidak memiliki penyakit jantung dan "1" artinya pasien memiliki penyakit jantung
6. *Ever_married*: "*No*" artinya pasien belum menikah dan "*Yes*" artinya pasien sudah/pernah menikah.
7. *Work type* :"*children*", "Govt_jov", "*Never_worked*", "Private" or "*Self-employed*"
8. *R*esidence type: "Rural" artinya pasien tinggal di pedesaan atau "Urban" artinya pasien tinggal di perkotaan
9. *Avg glucose level* : rata-rata kadar gula dalam darah
10. Bmi: body mass index (ukuran untuk menentukan kategori berat badan seseorang)
11. *Smoking status*: "*formerly smoked*" artinya pasien mantan perokok, "*never smoked*" artinya pasien tidak merokok, "*smokes*" artinya pasien perokok or "*Unknown*".

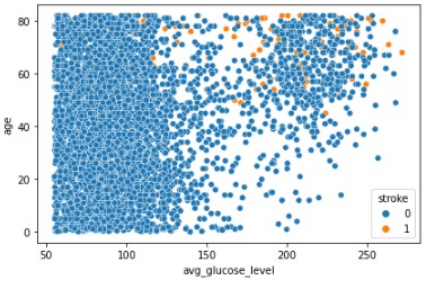
B. Implementasi

B.1. Data preprocessing

Dataset public yang digunakan masih berupa data mentah/data yang belum "bersih" (Masih terdapat missing value) jadi sebelum dilanjutkan ke tahap selanjutnya akan dilakukan pengecekan missing value pada dataset. Berdasarkan hasil pengecekan terdapat missing value pada kolom "bmi" sebanyak 201 baris. Untuk mengisi baris yang terdapat missing value peneliti menggunakan rata-rata dari total "bmi" dari label yang tidak terkena stroke dan label yang terkena stroke. setelah melakukan pembersihan data hal yang dilakukan selanjutnya yaitu mengubah data string ke numerik dan menghapus kolom yang tidak diperlukan agar dapat diproses ke langkah selanjutnya.

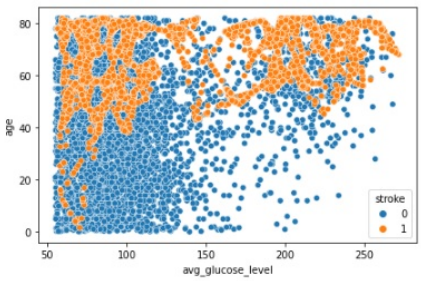
B.2. Teknik SMOTE

Setelah melalui tahap preprocessing data, tahap selanjutnya yaitu melakukan teknik SMOTE (Synthetic Minority Oversampling Technique). Dengan cara melakukan oversampling pada data minoritas dengan menggunakan konsep nearest neighbor, lalu untuk langkah-langkah dari teknik SMOTE (Synthetic Minority Over-sampling Technique) adalah sebagai berikut :



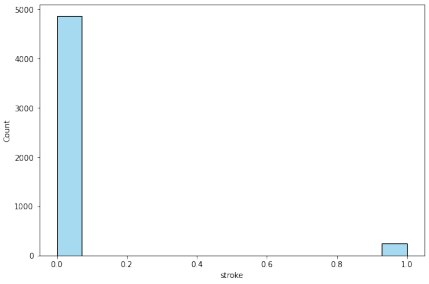
Gambar 3 Teknik SMOTE langkah 1

1. System akan melakukan analisis data mana yang termasuk mayoritas dan minoritas. Berdasarkan gambar 3 ternyata data pasien yang tidak terkena penyakit stroke lebih banyak daripada data pasien yang terkena penyakit stroke.



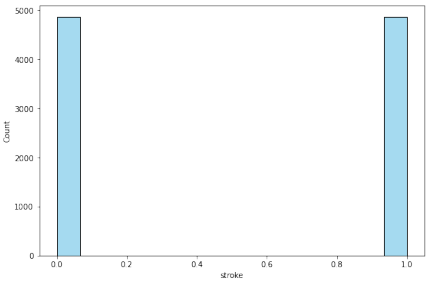
Gambar 4 Teknik SMOTE langkah 2

2. Setelah dilakukan analisis, *system* akan melakukan *oversampling* pada data minoritas hingga jumlahya sama dengan data mayoritas dengan didasarkan pada *knearest neighbour*. Berdasarkan gambar 4 setelah dilakukannya teknik SMOTE pada dataset kini data pasien yang terkena penyakit stroke dan data pasien yag tidak terkena penyakit stroke menjadi sama. Berikut adalah perbandingan jumlah data antara data sebeum teknk SMOTE dan sesudah teknik SMOTE.



Gambar 5 Sebelum teknik SMOTE

Berdasarkan gambar 5 data pasien yang terkena stroke terpaut cukup banyak dari data pasien yang tidak terkena stroke dengan rincian pasien terkena stroke berjumlah 249 pasien dan pasien yang tidak terkena stroke berjumlah 4.860 pasien terdapat selisih data sebanyak 4.611 data hal ini dapat menyebabkan kurang akuratnya model dalam melakukan deteksi pasien yang terkena stroke.



Gambar 6 Setelah teknik SMOTE

Setelah dilakukannya implementasi teknik smote pada dataset berdasarkan gambar 6 kini data pasien terkena penyakit stroke dan data pasien tidak terkena penyait stroke menjadi sama dengan rinician pasien terkena stroke berjumlah 4.860 pasien dan pasien yang tidak terkena stroke berjumlah 4.860. Kini jumlah data yang digunakan setelah teknik smote sebanyak 9.702 dari yang sebelumnya yang berjumlah 5.109.

C. Evaluasi

Setelah melalui tahap preprocessing dan teknik SMOTE selanjutnya data akan dibagi menjadi 2 yaitu data *train* dan data *test*, untuk pembagiannya akan dilakukan 5 perbedaan pembagian data sebagai berikut :

1. 10 % data *test* dan 90 % data *train*
2. 20 % data *test* dan 80% data *train*
3. 30% data *test* dan 70 % data *train*
4. 40% data *test* dan 60 % data *train*
5. 50% data *test* dan 50% data *train*

Data *train* dipergunakan untuk melatih metode yang digunakan dalam penelitian ini dan data *test* dipergunakan untuk menguji metode yang digunakan setelah melakukan *training* menggunakan data training. Berikut adalah rekap hasil evaluasi model dengan menggunakan *confusion matrix* dan *k-fold cross validation.*

C.1. Confusion Matrix

a. scenario 1

Tabel 1 *Scenario* 1 *confusion matrix naive bayes*

Data *test* 10%	Naïve Bayes		
Data *train* 90%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	87 %	79 %	- 8%
Precission	11 %	76 %	+ 65%
Recall	26 %	86 %	+ 60%
F1 Score	15 %	80 %	+ 65%

Scenario 1 dengan pembagian data *test* 10% data *train* 90% mendapatkan rata-rata delta 45.5%. Meskipun pada *scenario* 1 aspek *accuration* terdapat penurunan sebesar 8% dan menjadi penuruna yang paling banyak diantara *scenario* lainnya dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE akan tetapi jika dinilai secara keseluruhan *scenario* 1 lebih baik daripada 4 *scenario naïve bayes* lainnya.

Data *test* 30%	Naïve Bayes		
Data *train* 70%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	84 %	78 %	- 6%
Precission	14 %	74 %	+ 60%
Recall	37 %	85 %	+ 48%
F1 Score	20 %	80 %	+ 60%

Tabel 2 *Scenario* 1 *confusion matrix random forest*

Data *test* 10%	Random Forest		
Data *train* 90%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	94 %	95 %	+ 1%
Precission	11 %	95 %	+ 84%
Recall	0 %	96 %	+ 95%
F1 Score	15 %	95 %	+ 80%

Scenario 1 dengan pembagian data *test* 10% data *train* 90% mendapatkan rata-rata delta 65 %. Hal yang menonjol pada scenario 1 yaitu kenaikan *score precission* dan *recall* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling banyak diantara 4 *scenario random forest* lainnya dan kenaikan f1 *score* menjadi yang paling sedikit diantara 4 scenario lainnya.

b. scenario 2

Tabel 3 *Scenario 2 confusion matrix naive bayes*

Data *test* 20%	Naïve Bayes		
Data *train* 80%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	85 %	79 %	- 6%
Precission	14 %	75 %	+ 61%
Recall	40 %	85 %	+ 65%
F1 Score	20 %	80 %	+ 60%

Scenario 2 dengan pembagian data *test* 20% data *train* 80% mendapatkan rata-rata delta 45 %. Hal yang menonjol pada *scenario* 2 yaitu kenaikan *recall* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling banyak diantara 4 *scenario naïve bayes* lainnya yaitu sebesar 65%.

Tabel 4 *Scenario* 2 *confusion matrix random forest*

Data *test* 20%	Random Forest		
Data *train* 80%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	0.95	0.94	- 1%
Precission	0.40	0.94	+ 54%
Recall	0.40	0.95	+ 55%
F1 Score	0.10	0.95	+ 85%

Scenario 2 dengan pembagian data *test* 20% data *train* 80% mendapatkan rata-rata delta 48.25%. Hal yang menonjol pada *scenario* 2 yaitu kenaikan *score recall* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling sedikit diantara 4 *scenario random forest* lainnya.

c. scenario 3

Tabel 5 *Scenario* 3 *confusion matrix naive bayes*

Scenario 3 dengan pembagian data *test* 30% data *train* 70% mendapatkan rata-rata delta 40.5 %. Hal yang menonjol pada *scenario* 3 yaitu kenaikan *precission* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling sedikit diantara 4 *scenario naïve bayes* lainnya yaitu sebesar 60%.

Tabel 6 Scenario 3 confusion matrix random forest

Data *test* 30%	Random Forest		
Data *train* 70%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	94 %	95 %	+ 1%
Precission	75 %	94 %	+ 19%
Recall	3 %	96 %	+ 93%
F1 Score	4 %	95 %	+ 91%

Scenario 3 dengan pembagian data *test* 30% data *train* 70% mendapatkan rata-rata delta 51%. Hal yang menonjol pada scenario 3 yaitu kenaikan *score precission* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling sedikit diantara 4 *scenario random forest* lainnya.

d. scenario 4

Tabel 7 *Scenario 4 confusion matrix naive bayes*

Data *test* 40%	Naïve Bayes		
Data *train* 60%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	84 %	79 %	- 5%
Precission	12 %	75 %	+ 63%
Recall	35 %	86 %	+ 51%
F1 Score	18 %	80 %	+ 72%

Scenario 4 dengan pembagian data *test* 40% data *train* 60% mendapatkan rata-rata delta 45.25%. Hal yang menonjol pada *scenario* 4 yaitu kenaikan f1 *score* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling banyak diantara 4 *scenario naïve bayes lainnya* yaitu sebesar 72%.

Tabel 8 *Scenario* 4 *confusion matrix random forest*

Data *test* 40%	Random Forest		
Data *train* 60%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	95 %	94 %	- 1%
Precission	53 %	93 %	+ 40%
Recall	6 %	95 %	+ 89%
F1 Score	5 %	94 %	+ 91%

Scenario 4 dengan pembagian data *test* 40% data *train* 60% mendapatkan rata-rata delta 54.75 %. Sebenarnya tidak ada hal yang menonjol pada *scenario* 4 *random forest* akan tetapi kenaikan f1 *score* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE sebesar 91% itu menjadi kenaikan f1 *score* pada urutan kedua diantara 4 *scenario random forest* lainnya.

e. scenario 5

Tabel 9 *Scenario* 5 *confusion matrix naive bayes*

Data *test* 50%	Naïve Bayes		
Data *train* 50%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	85 %	79 %	- 6%
Precission	14 %	75 %	+ 61%
Recall	41 %	86 %	+ 45%
F1 Score	21 %	80 %	+ 59%

Scenario 5 dengan pembagian data *test* 50% data *train* 50% mendapatkan rata-rata delta 39.75 %. Untuk *scenari*o 5 mendapatkan kenaikan *score precission*, *recall* dan f1 *score* dari yang tidak menggunakan teknik SMOTE ke yang menggunakan teknik SMOTE adalah yang paling rendah diantara *scenario naïve bayes* lainnya.

Tabel 10 *Scenario* 5 *confusion matrix random forest*

Data *test* 50%	Random Forest		
Data *train* 50%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE	Δ
Accuration	94 %	94 %	-
Precission	37 %	93 %	+ 56%
Recall	2 %	95 %	+ 93%
F1 Score	4 %	94 %	+ 90%

Scenario 5 dengan pembagian data *test* 50% data *train* 50% mendapatkan rata-rata delta 60.75%. Hal yang menonjol pada scenario 5 yaitu kenaikan f1 *score* yang menjadi paling banyak diantara 4 *scenario random forest* lainnya dan pada *score* aspek *accuration* tidak mengalami penurunan yang terjadi seperti halnya 4 *scenario random forest* lainnya.

Urutan kinerja terbaik dari confusion matrix dengan 5 *scenario* pada metode *naïve bayes* yaitu :
1. Scenario 1 dengan rata rata delta sebesar 45.5%
2. Scenario 4 dengan rata rata delta sebesar 45.25%
3. Scenario 2 dengan rata rata delta sebesar 45%
4. Scenario 3 dengan rata rata delta sebesar 40%
5. Scenario 5 dengan rata rata delta sebesar 39.75%

Urutan kinerja terbaik dari *confusion matrix* dengan 5 *scenario* pada metode *random forest* yaitu :
1. *Scenario* 1 dengan rata rata delta sebesar 65%
2. *Scenario* 5 dengan rata rata delta sebesar 59.75%
3. *Scenario* 4 dengan rata rata delta sebesar 54.75%
4. *Scenario* 3 dengan rata rata delta sebesar 51%
5. *Scenario* 2 dengan rata rata delta sebesar 48.25%

C.2. K-Fold Cross Validation

a. scenario 1

Tabel 11 *Scenario* 1 *k-fold cross validation naive bayes*

Data *test* 10%	Naïve Bayes	
Data *train* 90%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.83	0.94
K-2	0.84	0.95
K-3	0.84	0.95

K-4	0.85	0.95
K-5	0.89	0.94
K-6	0.86	0.94
K-7	0.88	0.94
K-8	0.88	0.95
K-9	0.84	0.94
K-10	0.86	0.93
rata-rata	0.86	0.94

Berdasarkan tabel 11 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-1 dengan *score* 0.83 dan score yang paling besar terjadi pada k-5 dengan *score* 0.89, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-10 dengan *score* 0.93 dan yang paling besar terjadi pada k-2, k-3, k-4, k-8 dengan *score* 0.95.

Tabel 12 *Scenario* 1 *k-fold cross validation random forest*

Data *test* 10%	Random Forest	
Data *train* 90%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.95	0.94
K-2	0.95	0.94
K-3	0.94	0.95
K-4	0.95	0.95
K-5	0.95	0.94
K-6	0.95	0.94
K-7	0.95	0.94
K-8	0.94	0.95
K-9	0.95	0.94
K-10	0.95	0.93
rata-rata	0.95	0.94

Berdasarkan tabel 12 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-3 dan k-8 dengan score 0.94 dan *score* yang paling besar terjadi pada k-1, k-2, k-4, k-5, k-6, k-7, k-9 ,k-10 dengan *score* 0.95, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-10 dengan *score* 0.93 dan yang paling besar terjadi pada k-3, k-4, k-8 dengan *score* 0.95.

b. scenario 2

Tabel 13 *Scenario* 2 *k-fold cross validation naive bayes*

Data *test* 20%	Naïve Bayes	
Data *train* 80%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.83	0.95
K-2	0.85	0.95
K-3	0.84	0.95
K-4	0.88	0.93
K-5	0.88	0.95
K-6	0.88	0.92
K-7	0.86	0.94
K-8	0.87	0.94
K-9	0.83	0.94
K-10	0.86	0.92
rata-rata	0.86	0.94

Berdasarkan tabel 13 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-1 dan k-9 dengan *score* 0.83 dan *score* yang paling besar terjadi pada k-4, k-5, k-6 dengan *score* 0.88, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-6 dan k-10 dengan *score* 0.92 dan yang paling besar terjadi pada k-1, k-2, k-3 dengan *score* 0.95.

Tabel 14 *Scenario* 2 *k-fold cross validation random forest*

Data *test* 20%	Random Forest	
Data *train* 80%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.95	0.95
K-2	0.94	0.95
K-3	0.95	0.95
K-4	0.95	0.93
K-5	0.95	0.94
K-6	0.95	0.92
K-7	0.95	0.94
K-8	0.95	0.94
K-9	0.95	0.94
K-10	0.95	0.92
rata-rata	0.95	0.94

Berdasarkan tabel 14 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-2 dengan *score* 0.94 dan *score* yang paling besar terjadi pada k-1, k-3, k-4, k-5, k-6, k-7, k-8, k-9, k-10 dengan *score* 0.95, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-6 dan k-10 dengan *score* 0.92 dan yang paling besar terjadi pada k-1, k-2, k-3 dengan *score* 0.95.

c. scenario 3

Tabel 15 *Scenario* 3 *k-fold cross validation naive bayes*

Data *test* 30%	Naïve Bayes	
Data *train* 70%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.83	0.94
K-2	0.85	0.95
K-3	0.88	0.93
K-4	0.88	0.93
K-5	0.86	0.93
K-6	0.88	0.94
K-7	0.89	0.93
K-8	0.86	0.93
K-9	0.83	0.93
K-10	0.85	0.92
rata-rata	0.86	0.93

Berdasarkan tabel 15 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-1 dan k-9 dengan *score* 0.83 dan *score* yang paling besar terjadi pada k-7 dengan *score* 0.89, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-10 dengan *score* 0.92 dan yang paling besar terjadi pada k-2 dengan *score* 0.95.

Tabel 16 *Scenario* 3 *k-fold cross validation random forest*

Data *test* 30%	Random Forest	
Data *train* 70%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.94	0.94
K-2	0.95	0.95
K-3	0.95	0.94
K-4	0.95	0.94
K-5	0.94	0.92
K-6	0.95	0.94
K-7	0.94	0.94
K-8	0.95	0.93
K-9	0.95	0.93
K-10	0.95	0.93
rata-rata	0.95	0.94

Berdasarkan tabel 16 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-1, k-5, k-7 dengan *score* 0.94 dan *score* yang paling besar terjadi pada k-2, k-3, k-4, k-6, k-8, k-9, k-10 dengan *score* 0.95, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-5 dengan *score* 0.92 dan yang paling besar terjadi pada k-2dengan *score* 0.95.

d. scenario 4

Tabel 17 *Scenario* 4 *k-fold cross validation naive bayes*

Data test 40%	Naïve Bayes	
Data train 60%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.86	0.95
K-2	0.88	0.93
K-3	0.87	0.91
K-4	0.86	0.92
K-5	0.89	0.93
K-6	0.86	0.92
K-7	0.87	0.95
K-8	0.85	0.93
K-9	0.85	0.92
K-10	0.85	0.92
rata-rata	0.86	0.93

Berdasarkan tabel 17 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-8, k-9, k-10 dengan *score* 0.85 dan *score* yang paling besar terjadi pada k-5 dengan *score* 0.89, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-3 dengan *score* 0.91 dan yang paling besar terjadi pada k-1 dan k-7 dengan *score* 0.95.

Tabel 18 *Scenario* 4 *k-fold cross validation random forest*

Data test 40%	Random Forest	
Data train 60%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.95	0.95
K-2	0.95	0.93
K-3	0.95	0.92
K-4	0.95	0.92
K-5	0.94	0.92
K-6	0.95	0.93
K-7	0.95	0.94
K-8	0.95	0.93
K-9	0.94	0.92
K-10	0.95	0.91
rata-rata	0.95	0.93

Berdasarkan tabel 19 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-5 dan k-9 dengan *score* 0.94 dan *score* yang paling besar terjadi pada k-1, k-2, k-3, k-4, k-6, k-7, k-8, k-10 dengan *score* 0.95, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-10 dengan *score* 0.91 dan yang paling besar terjadi pada k-1dengan *score* 0.95.

e. scenario 5

Tabel 19 *Scenario* 5 *k-fold cross validation naive bayes*

Data test 50%	Naïve Bayes	
Data train 50%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.88	0.91
K-2	0.86	0.93
K-3	0.85	0.91
K-4	0.89	0.93
K-5	0.87	0.92
K-6	0.89	0.94
K-7	0.84	0.92
K-8	0.84	0.94
K-9	0.84	0.91
K-10	0.86	0.91
rata-rata	0.86	0.92

Berdasarkan tabel 19 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-7, k-8, k-9 dengan *score* 0.84 dan *score* yang paling besar terjadi pada k-6 dengan *score* 0.89, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-1, k-3, k-9, k-10 dengan *score* 0.91 dan yang paling besar terjadi pada k-6 dan k-8 dengan *score* 0.94.

Tabel 20 *Scenario* 5 *k-fold cross validation random forest*

Data test 50%	Random Forest	
Data train 50%	Tanpa Teknik SMOTE	Dengan Teknik SMOTE
K-1	0.95	0.91
K-2	0.94	0.93
K-3	0.94	0.90
K-4	0.94	0.92
K-5	0.94	0.93
K-6	0.95	0.94
K-7	0.95	0.91
K-8	0.95	0.94
K-9	0.95	0.91
K-10	0.95	0.91
rata-rata	0.95	0.92

Berdasarkan tabel 20 k-1 sampai k-10 pada tanpa teknik smote *score* yang paling kecil terjadi pada k-2, k-3, k-4, k-5 dengan *score* 0.94 dan *score* yang paling besar terjadi pada k-1, k-6, k-7, k-8, k-9, k-10 dengan *score* 0.95, sedangkan dengan yang menggunakan teknik smote untuk *score* yang paling kecil terjadi pada k-3 dengan *score* 0.90 dan yang paling besar terjadi pada k-6 dengan *score* 0.94.

Pembagian *percentage split* data *test* dan data *train* pada kinerja terbaik dari *k-fold cross validation* pada metode *naïve bayes* yaitu pada *scenario* 1 dan *scenario* 2 karena rata-rata *score* dari data yang tidak menggunakan smote ke data yang menggunakan smote mengalami kenaikan sebesar 0.8 itu menjadi kenaikan terbesar diantara 3 *scenario* lainnya. Sedangkan Pembagian *percentage split* data *test* dan data *train* pada kinerja terbaik dari *k-fold cross validation* pada metode *random forest* yaitu pada *scenario* 1, *scenario* 2 dan *scenario* 3 karena rata-rata *score* dari data yang tidak menggunakan smote ke data yang menggunakan smote mengalami penurunan sebesar 0.1 itu menjadi penurunan terkecil diantara 2 *scenario.*

D. Pengujian Model

Pada pengujian model ini data yang digunakan merupakan data berbeda dari dataset yang telah di *train* dan *test* menggunakan model serta data uji didapat dari kaggle. Untuk metode yang digunakan pada model deteksi adalah *random forest* karena dalam segi akurasi yang di ukur dari *confusion matrix* maupun *k-fold cross validation* pada *percentage split* yang terbaik nilai metode *random forest* lebih tinggi daripada metode *naïve bayes.*

Tabel 21 Data uji model

G	A1	H1	H2	E	W	R	A2	B	S1	S2	RF
0	67	0	1	1	1	0	228.69	36.6	3	1	1
0	80	0	1	1	1	1	105.92	32.5	0	1	1
1	49	0	0	1	1	0	171.23	34.4	1	1	1
1	79	1	0	1	0	1	174.12	24	0	1	1
0	81	0	0	1	1	0	186.21	29	3	1	1
1	49	0	0	1	1	1	60.22	31.5	1	0	0
0	71	0	0	1	1	0	198.21	27.3	3	0	0
1	59	0	0	1	1	0	109.82	23.7	0	0	0
1	25	0	0	1	1	0	60.84	24.5	0	0	0
1	67	0	0	1	3	1	94.61	28.4	1	0	0

G = *Gender*
A1 = *Age*
H1 = *Hypertension*
H2 = *Heart Disease*
E = *Ever Married*
W = *Work Type*
R = *Residence Type*
A2 = *Avg Glucose Level*
B = bmi
S1 = *Smoking Status*
S2 = *Stroke* label datset
RF = Hasil deteksi model menggunakan metode random forest

Berdasarkan tabel 21 terbukti dengan akurasi pada percentage spilt yang terbaik dengan *score* akurasi 95 % pada confusion matrix dan rata-rata *score* 0.94 pada *k-fold cross validation random forest* terbukti efektif dengan di tunjukannya hasil deteksi yang tepat sesuai dengan dataset yang ada.

IV. KESIMPULAN

Berdasarkan hasil dan pembahasan pada penelitian tentang "Perbandingan Metode Naïve Bayes dan Random Forest Pada Deteksi Penyakit Stroke Menggunakan Teknik SMOTE (*Synthetic Minority Oversampling Technique*).", dapat disimpulkan bahwa :

1. Perbandingan metode *naïve bayes* dan *random forest* pada dataset yang digunakan dalam penelitian ini terbukti, metode random forest memiliki akurasi lebih tinggi daripada metode naïve bayes dan metode random forest efektif dalam mendeteksi penyakit stroke dengan jumlah data yang besar.
2. Dalam mengukur kinerja metode menggunakan *k-fold cross validation* pada metode naïve bayes terdapat kenaikanan rata-rata score *k-fold cross validation* dari yang tidak menggunakan teknik SMOTE ke yang meggunakan teknik SMOTE sebaliknya metode random forest mengalami penurunan rata-rata score *k-fold cross validation* dari yang tidak menggunakan teknik SMOTE ke yang meggunakan teknik SMOTE
3. Scenario pembagian data *test* dan data *train* yang terbaik di penlitian ini adalah 10% data *test* dan 90% data *train* hal itu dapat dilihat dari *confusion matrix* yang memiliki rata-rata delta lebih besar daripada *scenario* lainnya dan *score k-fold cross validation* pada scenario 1 juga memiliki rata-rata *score* yang paling tinggi diantara *scenario* lainnya

V. SARAN

Pada hasil penelitian perbandingan metode naïve bayes dab random forest pada deteksi penyakit stroke menggunakan teknik smote diperoleh kesimpulan bahwasannya random forest lebih optimal dalam melakukan deteksi penyakit stroke. Untuk penelitian selanjutnya dapat dicoba dengan data set yang berbeda, penyakit yang berbeda, maupun metode yang berbeda agar lebih variatif atau dapat menemukan metode yang lebih baik lagi untuk mendeteksi suatu penyakit

UCAPAN TERIMA KASIH

Puji syukur penulis ucapkan kepada Allah SWT atas segala rahmat yang telah diberikan sehingga penulis dapat menyelesaikan penulisan artikel dengan judul "PERBANDINGAN METODE NAÏVE BAYES DAN RANDOM FOREST PADA DETEKSI PENYAKIT STROKE MENGGUNAKAN TEKNIK SMOTE (Synthetic Minority Over-sampling Technique)". Penulisan artikel ini dapat berjalan dengan baik karena dukungan dari beberapa pihak. Sehingga, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih kepada :

1. Kedua orang tua saya, yang selalu memberi banyak dukungan, motivasi dan semangat kepada penulis, sehingga penulis mampu menyelesaikan skripsi.
2. Dr. Yuni Yamasari S.Kom., M.Kom selaku Dosen Pembimbing yang telah mengarahkan penulis dalam menyelesaikan skripsi.
3. Ibu Anita Qoiriah, S.Kom., M.Kom dan Ibu Naim Rochmawati, S.Kom., M.T. selaku dosen penguji yang selalu memberikan masukan-masukan agar skripsi saya menjadi lebih baik.
4. Seluruh Dosen Teknik Informatika yang telah memberikan ilmunya selama penulis belajar di Program Studi S1 Teknik Informatika.
5. Semua pihak yang tidak bisa disebutkan satu persatu yang telah memberikan dukungan guna terlaksananya Skripsi ini.

REFERENSI

[1] A. A. Arifiyanti and E. D. Wahyuni, "Smote: Metode Penyeimbang Kelas Pada Klasifikasi Data Mining," *SCAN - J. Teknol. Inf. dan Komun.*, vol. 15, no. 1, pp. 34–39, 2020, doi: 10.33005/scan.v15i1.1850.

[2] D. Haryadi, D. Marini Umi Atmaja, A. Rahman Hakim, and N. Suwaryo, "Identifikasi Tingkat Resiko Penyakit Stroke Menggunakan Algoritma Regresi Linear Berganda," *Deny Haryadi, SNTEM*, vol. 1, no. November, pp. 1198–1207, 2021.

[3] F. I. Kurniadi and P. D. Larasati, "Light Gradient Boosting Machine untuk Deteksi Penyakit Stroke," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 6, no. 1, pp. 67–72, 2022, doi: 10.47970/siskom-kb.v6i1.328.

[4] A. Murugan, S. A. H. Nair, and K. P. S. Kumar, "Detection of Skin Cancer Using SVM, Random Forest and kNN Classifiers," *J. Med. Syst.*, vol. 43, no. 8, 2019, doi: 10.1007/s10916-019-1400-8.

[5] D. Nofriansyah, K. Erwansyah, and M. Ramadhan, "Penerapan Data Mining dengan Algoritma Naive Bayes Clasifier untuk Mengetahui Minat Beli Pelanggan terhadap Kartu Internet XL (Studi Kasus di CV. Sumber Utama Telekomunikasi)," *J. Saintikom*, vol. 15, no. 2, pp. 81–92, 2018.

[6] P. Biji, K. Stelechocarpus, T. Secara, I. N. Vitro, and D. A. N. Ex, "Jurnal MIPA," vol. 37, no. 2, pp. 105–114, 2014.

[7] W. Deng, Z. Huang, J. Zhang, and J. Xu, "A Data Mining Based System for Transaction Fraud Detection," *2021 IEEE Int. Conf. Consum. Electron. Comput. Eng. ICCECE 2021*, pp. 542–545, 2021, doi: 10.1109/ICCECE51280.2021.9342376.