Analisis Faktor Dominan dan Peningkatan Akurasi Model Pohon Keputusan dalam Klasifikasi Risiko Stroke pada Pasien

Dasmond Tan[1], Willsen Wijaya[2], James Andersen[3] ,Lian wira manuel maharaja[4], Gregorius Daniel Dwitama[5]

[1]*Faculty of engineering and informatics, Universitas Multimedia Nusantara,* Tangerang, Indonesia
dasmond.tan@student.umn.ac.id
[2]*Faculty of engineering and informatics, Universitas Multimedia Nusantara,* Tangerang, Indonesia
willsen.wijaya@student.umn.ac.id
[3]*Faculty of engineering and informatics, Universitas Multimedia Nusantara,* Tangerang, Indonesia
james.andersen@student.umn.ac.id
[4]*Faculty of engineering and informatics, Universitas Multimedia Nusantara,* Tangerang, Indonesia
lian.wira@student.umn.ac.id
[5]*Faculty of engineering and informatics, Universitas Multimedia Nusantara,* Tangerang, Indonesia
gregorius.daniel@student.umn.ac.id

***Abstract*—Stroke is a major health problem worldwide that often leads to death or permanent disability. Classification for early prevention can reduce the impact of stroke on a person. Key risk factors, such as BMI and average glucose level, were used in this study to classify patients at risk of stroke using the decision tree method. The decision tree model is a model with capabilities in data-based interpretation and decision-making. In addition, the CRISP-DM methodology enables the application of a structured and hierarchical data mining approach, which improves the implementation and decision-making process. There are previous studies that used the decision tree algorithm with all available features and SVM with linear and polynomial kernels as the basis for comparison. In this study, a decision tree model with feature selection of BMI and average glucose level and pruning was used to simplify the model and improve its accuracy, as well as identify the most influential features. The results of this study showed that the pruned decision tree model, with a focus on the BMI and average glucose level features, performed better than the model in the previous study, and the average glucose level became the most influential feature in classifying the potential stroke in a person. This emphasizes the importance of pruning and effective feature selection in developing effective machine learning models for stroke risk classification.**

Keywords*— *Decision Tree; SVM; CRISP-DM; Model; Classification; Stroke; Pruning.

***Abstrak*—Stroke merupakan masalah kesehatan yang utama di seluruh dunia yang sering menyebabkan kematian atau kecacatan permanen. Klasifikasi untuk pencegahan dini dapat mengurangi dampak stroke terhadap seseorang. Faktor risiko utama, seperti BMI dan kadar glukosa rata-rata, digunakan dalam studi ini untuk mengklasifikasikan pasien yang berisiko terkena stroke menggunakan metode decision tree. Model decision tree merupakan model dengan kemampuan dalam interpretasi dan pengambilan keputusan berbasis data. Selain itu, metodologi CRISP-DM memungkinkan penerapan pendekatan penggalian data yang terstruktur dan hierarkis, yang meningkatkan proses implementasi dan pengambilan keputusan. Terdapat penelitian sebelumnya yang menggunakan algoritma decision tree dengan semua fitur yang tersedia dan SVM dengan kernel linear dan polinomial menjadi dasar perbandingan. Dalam penelitian ini, digunakan model decision tree dengan pemilihan fitur yaitu BMI dan kadar glukosa rata-rata dan pruning untuk menyederhanakan model dan meningkatkan akurasinya, serta mengidentifikasi fitur yang paling berpengaruh. Hasil dari penelitian ini menunjukkan bahwa model decision tree yang di pruning, dengan fokus pada fitur BMI dan kadar glukosa rata-rata, memiliki kinerja lebih baik dibandingkan model pada penelitian sebelumnya. dan kadar glukosa rata - rata menjadi fitur paling berpengaruh dalam melakukan klasifikasi potensi stroke pada seseorang. Hal ini menekankan pentingnya penggunaan pruning dan pemilihan fitur yang efektif dalam mengembangkan model machine learning yang efektif untuk klasifikasi risiko stroke.**

Kata Kunci*— *Decision Tree; SVM; CRISP-DM; Model; Klasifikasi; Stroke; Pruning.

I. INTRODUCTION

Stroke merupakan masalah kesehatan utama di seluruh dunia, yang sering kali menyebabkan kecacatan jangka panjang atau permanen dan bahkan kematian. Stroke adalah kondisi kesehatan yang menyebabkan kerusakan dengan merobek pembuluh darah di otak [1]. Menurut World Stroke Organization (2022) secara global, lebih dari 12,2 juta atau satu dari empat orang di atas usia 25 akan mengalami stroke [2]. Klasifikasi dini dan pencegahan stroke dapat secara signifikan mengurangi dampaknya pada individu dan

sistem perawatan kesehatan. Beberapa faktor risiko berkontribusi terhadap kemungkinan terjadinya stroke, termasuk indeks massa tubuh (BMI) dan peningkatan kadar glukosa. Biasanya tingkat BMI yang lebih tinggi pada pasien diabetes dikaitkan dengan insiden stroke yang lebih tinggi [3]. Indeks massa tubuh yang tinggi dapat meningkatkan risiko pembekuan darah atau pendarahan di otak. Diabetes adalah faktor risiko independen yang diketahui untuk stroke [4], kadar glukosa yang tinggi dapat merusak pembuluh darah dari waktu ke waktu, yang berkontribusi terhadap risiko stroke. Disebabkan adanya beban perawatan kesehatan yang signifikan yang ditimbulkan oleh stroke, kebutuhan akan model klasifikasi yang efektif dapat membantu mengidentifikasi individu yang berisiko tinggi terkena stroke atau tidak. Dengan memanfaatkan data tentang faktor risiko utama, seperti BMI dan kadar glukosa rata-rata, penyedia layanan kesehatan dapat mengambil tindakan proaktif untuk mengurangi risiko stroke.

Pada penelitian ini, metode decision tree akan digunakan untuk melakukan klasifikasi terhadap pasien yang berpotensi terkena stroke atau tidak. Metode decision tree dipilih karena kemampuannya yang tinggi dalam interpretasi dan pengambilan keputusan berbasis data. Dengan menggunakan decision tree, pemahaman yang jelas tentang bagaimana kombinasi dari faktor-faktor risiko seperti BMI dan kadar glukosa dapat mempengaruhi kemungkinan terjadinya stroke dapat diperoleh. Selain itu, jenis metodologi yang digunakan adalah CRISP-DM, dimana metodologi ini menetapkan tugas dan tingkat abstraksi yang terstruktur secara hirarkis, sehingga dapat memudahkan dalam implementasi dan pengambilan keputusan dalam proyek ini dan mendukung analisis yang lebih efektif dan efisien.

Salah satu faktor yang mendorong dilakukan penelitian ini adalah terdapat beberapa penelitian yang melakukan klasifikasi risiko stroke. Terdapat penelitian terdahulu yang menggunakan algoritma decision tree dengan memanfaatkan semua fitur yang tersedia dalam dataset mereka untuk membangun model klasifikasi [5]. Selain itu, terdapat penelitian yang menggunakan algoritma SVM dengan kernel linear dan polinomial yang digunakan dalam klasifikasi risiko stroke [6]. Pada penelitian ini, algoritma decision tree digunakan dengan pemilihan fitur utama seperti BMI dan kadar glukosa rata-rata, model dapat menjadi lebih sederhana dan interpretatif, serta meningkatkan akurasi dari kinerja model. Selain itu, digunakan teknik pruning pada decision tree untuk mengurangi kompleksitas model dan meningkatkan akurasi modelnya.

Dengan demikian, penelitian ini tidak hanya akan membandingkan efektivitas model decision tree dalam mengklasifikasi potensi seseorang terkena stroke atau tidak dengan menggunakan fitur BMI dan kadar glukosa rata-rata, tetapi juga akan menilai kinerja model setelah dilakukan pruning untuk memastikan model yang dihasilkan tidak hanya akurat tetapi juga efisien dan mudah diinterpretasi. Dengan pruning, ruang lingkup pohon dipotong pendek dan hanya menyisakan simpul dan cabang yang diperlukan [7]. Selain itu, penelitian ini juga berkontribusi dalam menilai variabel yang paling berpengaruh dalam mengklasifikasikan potensi stroke pada pasien.

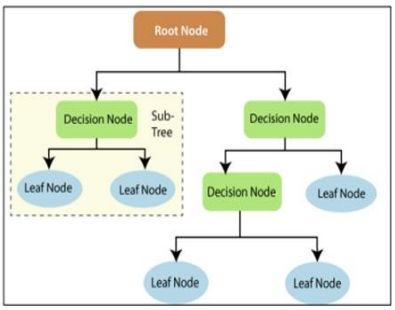
II. LITERATURE STUDY

A. *CRISP-DM*

Gambar 1. CRISP-DM

CRISP-DM adalah metodologi pengembangan proyek data mining yang menetapkan tugas dan tingkat abstraksi yang terstruktur secara hirarkis untuk memfasilitasi implementasi melalui tindakan yang membantu dalam pengambilan keputusan [8]. Kerangka kerja yang lebih fleksibel tetapi tetap terstruktur untuk menyelesaikan permasalahan bisnis menggunakan teknik data mining disediakan oleh CRISP-DM. Terdapat 6 langkah dalam CRISP-DM yang dapat dilihat dari gambar 1, yaitu: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, dan Deployment.

B. *Decision Tree*



Gambar 2. Decision Tree

Decision tree adalah teknik pembelajaran mesin yang membangun struktur pohon dari titik-titik data dataset untuk membuat prediksi yang akurat tentang data yang tidak terlihat [9]. Algoritma supervised

machine learning ini digunakan dalam klasifikasi maupun regresi oleh decision tree. Fungsi dari decision tree sendiri adalah digunakan untuk memecahkan masalah klasifikasi dan regresi dengan mengidentifikasi hubungan antar titik data dan membuat prediksi akurat tentang data yang tidak terlihat [10]. Decision Tree juga dapat digunakan untuk memecahkan masalah klasifikasi dan dievaluasi berdasarkan akurasi klasifikasi, kedalaman pohon, simpul daun, dan waktu konstruksi pohon [11]. Keuntungan menggunakan decision tree adalah visualnya yang mudah ditafsirkan dan dipahami, sehingga cocok untuk domain sensitif seperti diagnosis medis [12]. Dalam penelitian ini, klasifikasi dilakukan menggunakan decision tree. Dataset dibagi menjadi subset-subset berdasarkan fitur tertentu oleh decision tree.

C. ID3

ID3, atau Iterative Dichotomiser 3, adalah algoritma pembelajaran pohon keputusan yang digunakan untuk klasifikasi data dan analisis prediktif [13]. Algoritma ini dikembangkan oleh Ross Quinlan pada tahun 1986. Algoritma ini juga memiliki performa yang optimal dan efektif, khususnya dalam meminimalkan tingkat kesalahan [14]. ID3 memiliki keunggulan yaitu ia mampu menangani data yang lengkap dan tidak lengkap, memberikan akurasi yang baik dibandingkan metode tradisional [15]. Klasifikasi dilakukan oleh ID3, dan atribut yang paling optimal dipilih oleh algoritma ini untuk memisahkan data menjadi beberapa kelas berdasarkan konsep entropy dan information gain.

D. SVM

SVM adalah mesin vektor pendukung yang memecahkan masalah dengan sampel kecil, nonlinier, dan berdimensi tinggi dengan menggunakan berbagai metode optimasi [16]. Algoritma machine learning ini juga digunakan untuk klasifikasi dan regresi oleh SVM. Fungsi Support Vector Machines (SVM) sangat dipengaruhi oleh pilihan fungsi kernel [17]. Fungsi kernel menentukan bagaimana data diubah dan dipetakan ke dalam ruang fitur yang lebih tinggi, yang memungkinkan SVM untuk menemukan hyperplane pemisah yang optimal untuk klasifikasi atau regresi. SVM memiliki keunggulan teori yang lengkap, optimasi global, kemampuan adaptasi yang kuat, dan kemampuan generalisasi yang baik karena didasarkan pada teori pembelajaran statistik [18]. Dalam penelitian kali ini, klasifikasi stroke akan dilakukan menggunakan SVM dengan kernel sigmoid.

E. Pruning

Pruning dalam decision tree adalah proses penting yang bertujuan untuk mengurangi ukuran pohon, waktu komputasi, dan overfitting sambil mempertahankan atau meningkatkan akurasi klasifikasi. Fungsi pruning pada decision tree berfungsi untuk mengurangi ukurannya, meningkatkan waktu klasifikasi, dan meningkatkan akurasi [19]. Pada penelitian ini, digunakan parameter max_depth untuk menentukan kedalaman maksimum decision tree. Kedalaman pohon adalah jumlah level atau tingkat percabangan yang dibentuk oleh model decision tree. Dengan melakukan pembatasan terhadap kedalaman pohon, maka dapat mengontrol kompleksitas model. Semakin dalam pohon, semakin kompleks modelnya, dan semakin cenderung model akan melakukan overfitting pada data latih. Sebaliknya, semakin dangkal pohonnya, semakin sederhana modelnya, dan semakin cenderung model akan menjadi kurang mampu dalam mempelajari pola yang kompleks dari data. Oleh karena itu, pemilihan nilai yang tepat untuk max_depth harus didasarkan pada trade-off antara underfitting dan overfitting. Keunggulan dari pruning adalah meningkatkan generalisasi dengan menghapus data yang berisik dan kontradiktif dan mengurangi kompleksitas struktur pohon [20].

F. Accuracy

Accuracy merupakan sebuah rasio prediksi benar (positif dan negatif) dengan keseluruhan data. Metrik ini mengukur sejauh mana model berhasil dalam melakukan klasifikasi dengan benar. Berikut ini merupakan rumus dari accuracy.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

Keterangan :

- TP : True Positive atau jumlah pengamatan yang benar diprediksi sebagai positif.
- TN : True Negative atau jumlah pengamatan yang benar diprediksi sebagai negatif.
- FP : False Positive atau jumlah pengamatan yang seharusnya negatif, namun salah diprediksi sebagai positif.
- FN : False Negative atau jumlah pengamatan yang seharusnya positif, namun salah diprediksi sebagai negatif.

G. Penelitian sebelumnya

Terdapat dua penelitian sebelumnya yang digunakan sebagai acuan dalam penelitian ini.

1. Model decision tree dengan semua fitur.

Pada penelitian ini, digunakan model decision tree tanpa pruning dan menggunakan semua fitur yang dimiliki yaitu Gender, Age, Hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, Bmi, Smoking_status, Stroke. Hasil akurasi dari model tersebut adalah 92%.

2. Model SVM kernel linear dan polynomial

Pada penelitian ini, digunakan model SVM dengan kernel linear dan polynomial. Fitur yang digunakan adalah Gender, Age, Hypertension, Heart Disease, Ever Married, Work Type Residence Type, Avg Glucose, Level, BMI, Smoking Status, Stroke. Hasil akurasi dari model SVM dengan kernel linear adalah 76% dan kernel polynomial adalah 80%.

III. METHODOLOGY

A. Pemilihan dataset

Dalam penelitian ini, dataset yang digunakan berjudul Stroke Prediction Dataset yang didapatkan dari *Kaggle* dan digunakan untuk melakukan klasifikasi mengenai potensi terkena stroke berdasarkan atribut-atribut yang tersedia. Data ini berisi informasi yang relevan tentang setiap pasien, termasuk jenis kelamin, usia, riwayat hipertensi dan penyakit jantung, status pernikahan, jenis pekerjaan, tipe tempat tinggal, rata-rata kadar glukosa darah, indeks massa tubuh (BMI), dan status merokok. Dataset ini memiliki 5110 baris dan 12 kolom. Dataset ini juga memiliki berbagai tipe data yaitu object, int, dan float.

B. Business Understanding

Stroke merupakan masalah kesehatan utama di seluruh dunia, yang sering kali menyebabkan kecacatan jangka panjang atau bahkan kematian. Menurut Organisasi Kesehatan Dunia (WHO), stroke adalah penyebab utama kematian kedua dan penyebab utama kecacatan ketiga di seluruh dunia. Prediksi dini dan pencegahan stroke dapat secara signifikan mengurangi dampaknya pada individu dan sistem perawatan kesehatan. Beberapa faktor risiko berkontribusi terhadap kemungkinan terjadinya stroke, termasuk indeks massa tubuh (BMI) dan peningkatan kadar glukosa. Indeks massa tubuh yang tinggi dapat meningkatkan risiko pembekuan darah atau pendarahan di otak, sementara kadar glukosa yang tinggi dapat merusak pembuluh darah dari waktu ke waktu, yang berkontribusi terhadap risiko stroke. Mengingat beban perawatan kesehatan yang signifikan yang ditimbulkan oleh stroke, ada kebutuhan akan model klasifikasi yang efektif yang dapat membantu mengidentifikasi individu yang berisiko tinggi. Dengan memanfaatkan data tentang faktor risiko utama, seperti BMI dan kadar glukosa rata-rata, penyedia layanan kesehatan dapat mengambil tindakan proaktif untuk mengurangi risiko stroke.

C. Data Understanding

Tahap data understanding melibatkan pemeriksaan awal terhadap data untuk memahami karakteristik dasar, kualitas, dan potensi masalah yang ada. Berikut ini merupakan beberapa hal yang dilakukan pada tahap ini.

- Data Information

```
df.info()
<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 3. Informasi dataset

Gambar 3 di atas, ditampilkan mengenai informasi dataset, fungsi `df.info()` digunakan untuk menampilkan informasi dataset. Output dari fungsi ini memberikan informasi seperti jumlah dari total data yaitu sebanyak 5110, jumlah kolom yaitu sebanyak 12 kolom, dan tipe data setiap kolom bervariasi, di mana terdapat ‘int64’, ‘object’, dan ‘float64’.

- Data Describe

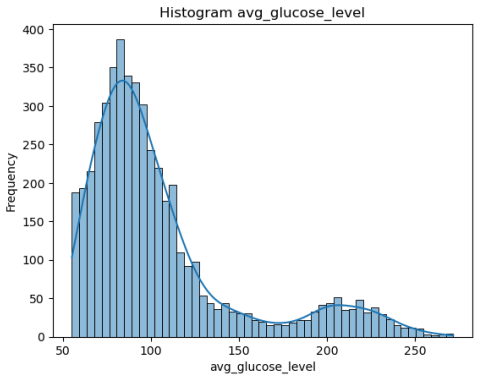
```
df.describe()
```

	id	age	hypertension	heart_disease	avg_glucose_level	bmi	stroke
count	4909.000000	4909.000000	4909.000000	4909.000000	4909.000000	4909.000000	4909.000000
mean	37064.313506	42.865374	0.091872	0.049501	105.305150	28.893237	0.042575
std	20995.098457	22.555115	0.288875	0.216934	44.424341	7.854067	0.201917
min	77.000000	0.080000	0.000000	0.000000	55.120000	10.300000	0.000000
25%	18605.000000	25.000000	0.000000	0.000000	77.070000	23.500000	0.000000
50%	37608.000000	44.000000	0.000000	0.000000	91.680000	28.100000	0.000000
75%	55220.000000	60.000000	0.000000	0.000000	113.570000	33.100000	0.000000
max	72940.000000	82.000000	1.000000	1.000000	271.740000	97.600000	1.000000

Gambar 4. Deskripsi dataset

Gambar 4 di atas, ditampilkan deskripsi dari dataset, fungsi ‘df.describe()’ digunakan untuk menampilkan deskripsi seperti perhitungan count, mean (rata-rata), standar deviasi, min (nilai minimum), 25%, 50%, 75%, dan max (nilai maximum) dari setiap kolom dalam data.

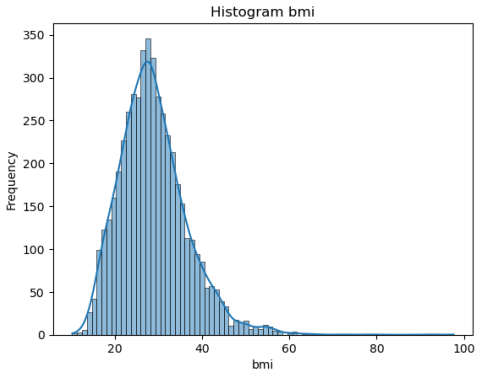
- Data Visualization



Gambar 5. Visualisasi histogram Avg_glucose_levelt

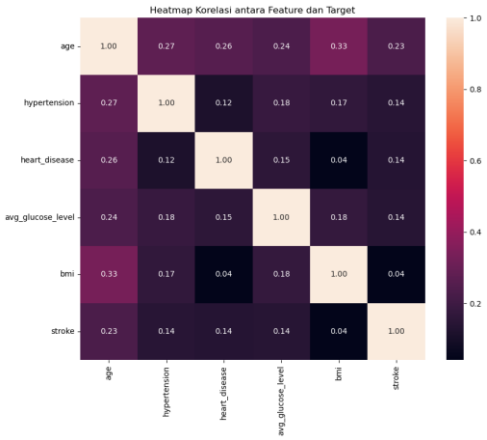
Berdasarkan hasil visualisasi histogram di atas, ditampilkan data ‘avg_glucose_level’ memiliki distribusi yang cenderung miring ke kanan dengan dua puncak utama, menunjukkan adanya kelompok data

dengan nilai glukosa yang lebih rendah dan lebih tinggi, tetapi dengan frekuensi yang berkurang saat nilai meningkat.



Gambar 6. Visualisasi histogram BMI

Histogram pada gambar 6, ditampilkan histogram distribusi Body Mass Index (BMI) pada sampel populasi. Sumbu horizontal (x) mewakili nilai BMI, dan sumbu vertikal (y) menunjukkan jumlah orang dengan nilai BMI tersebut. Mayoritas orang memiliki BMI antara 20 dan 30, dengan puncak di sekitar 25, menunjukkan bahwa kebanyakan orang berada dalam kategori berat badan normal hingga sedikit berlebih. Nilai BMI yang sangat rendah atau sangat tinggi jarang terjadi. Secara keseluruhan, grafik ini menunjukkan bahwa distribusi BMI mendekati normal tetapi sedikit condong ke kanan.



Gambar 7. Visualisasi heatmap

Heatmap di atas, ditampilkan untuk melihat seberapa kuat korelasi atau hubungan antar setiap fitur. Diantara semua fitur di atas, korelasi yang paling kuat adalah fitur ‘bmi’ dengan ‘age’ yaitu sebesar 0.33.

D. Data Preparation

- Check missing value

```
df.isnull().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
```

Gambar 8. Visualisasi boxplot outlier

Gambar diatas, merupakan code untuk menampilkan missing value dalam dataset. Terlihat bahwa, terdapat 201 missing value pada variabel BMI.

- Clear missing value

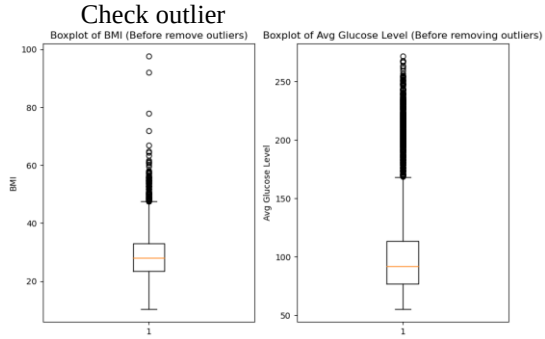
```
df.dropna(inplace=True)
df.isnull().sum()
```

```
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                  0
smoking_status       0
stroke               0
dtype: int64
```

Gambar 9. Visualisasi missing value

Gambar diatas, merupakan code untuk menampilkan missing value dalam dataset. Terlihat bahwa, terdapat 201 missing value pada variabel BMI.

- Check outlier

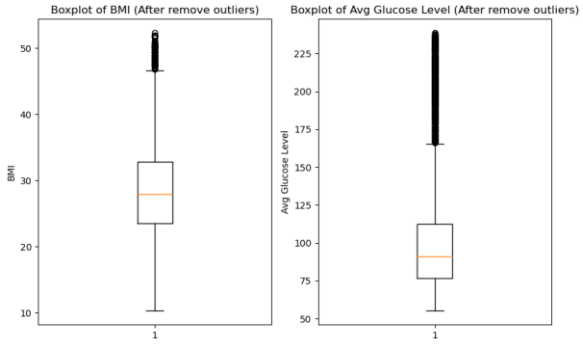


Gambar 10. Visualisasi boxplot outlier

Gambar di atas merupakan visualisasi menggunakan boxplot untuk memeriksa apakah terdapat outlier pada data numerik tersebut dan dapat dilihat terdapat outlier pada semua datanya. Mengidentifikasi keberadaan outlier dan menghilangkannya merupakan hal yang sangat penting untuk membangun dataset pelatihan yang berkualitas. Outlier memiliki dampak yang signifikan pada pembelajaran model. Outlier dapat membuat model

menjadi tidak stabil dan tidak konsisten, serta menyebabkan bias dan ketidakakuratan dalam model. Dapat dilihat, bahwa kedua variabel terdapat outlier, sehingga perlu dilakukan penghapusan terhadap nilai outlier tersebut.

- Remove outlier



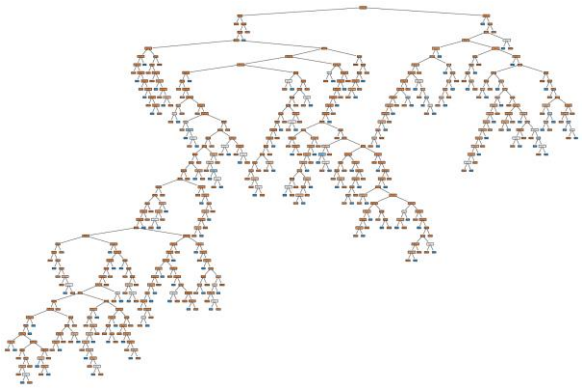
Gambar 11. Visualisasi boxplot outlier telah dihapus.

Gambar diatas merupakan visualisasi menggunakan boxplot untuk menampilkan data yang telah dihapus outliernya.

E. Modeling

- Model decision tree tanpa pruning

```
model = DecisionTreeClassifier(criterion="entropy")
model.fit(X_train, y_train)

y_pred = model.predict(X_test)
```

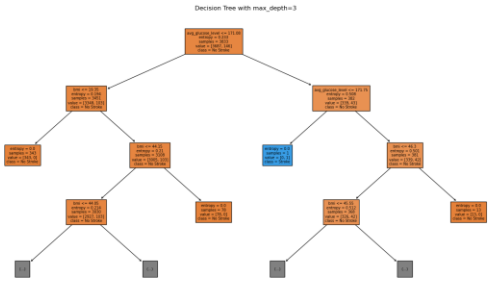


Gambar 12. Visualisasi Decision Tree sebelum pruning

Gambar di atas merupakan code dan hasil visualisasi dari model decision tree sebelum melakukan pruning. Terlihat bahwa decision tree memiliki banyak percabangan yang membentuk struktur pohon yang sangat kompleks.

- Model decision tree dengan pruning (max_depth = 3)

```
clf_p3 = DecisionTreeClassifier(criterion="entropy", max_depth=3)
clf_p3.fit(X_train, y_train)

y_pred_p = clf_p3.predict(X_test)
```

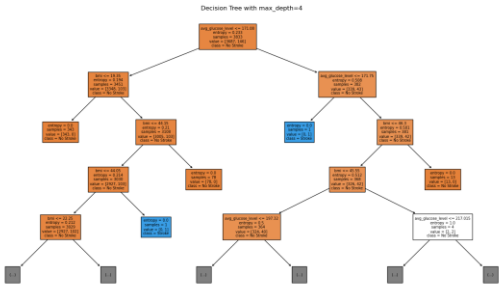


Gambar 13. Visualisasi decision tree setelah pruning (max_depth = 3)

Gambar di atas merupakan code dan hasil visualisasi dari model decision tree setelah dilakukan pruning dengan max_depth = 3. Parameter max_depth diatur pada 3 dalam pembuatan model menentukan kedalaman maksimum pohon keputusan dan dapat dilihat bahwa, hasil visualisasi tidak terlalu kompleks dan lebih sederhana dibandingkan model sebelum di pruning.

- Model decision tree dengan pruning (max_depth = 4)

```
clf_p4 = DecisionTreeClassifier(criterion="entropy", max_depth=4)
clf_p4.fit(X_train, y_train)

y_pred_p = clf_p4.predict(X_test)
```



Gambar 14. Visualisasi decision tree setelah pruning (max_depth = 4)

Gambar di atas merupakan code dan hasil visualisasi dari model decision tree setelah dilakukan pruning dengan max_depth = 4. Parameter max_depth diatur pada 4 dalam pembuatan model menentukan kedalaman maksimum pohon keputusan dan dapat dilihat bahwa, hasil visualisasi tidak terlalu kompleks dan lebih sederhana dibandingkan model sebelum di pruning.

- Model SVM dengan kernel Sigmoid

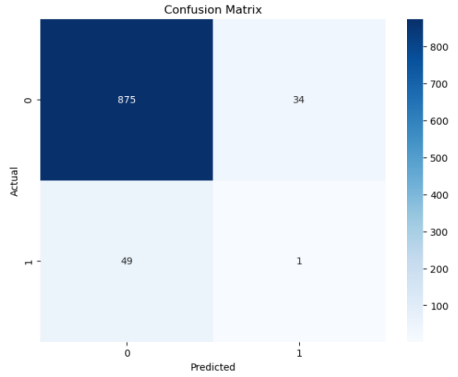
```
svm = SVC(kernel='sigmoid', random_state=42)
svm.fit(X_train, y_train)
y_pred = svm.predict(X_test)
```

Gambar 15. SVM kernel sigmoid

Code di atas merupakan code untuk membuat, melatih, dan menguji model SVM dengan kernel sigmoid untuk melakukan klasifikasi data.

F. Evaluation

- Model decision tree sebelum pruning

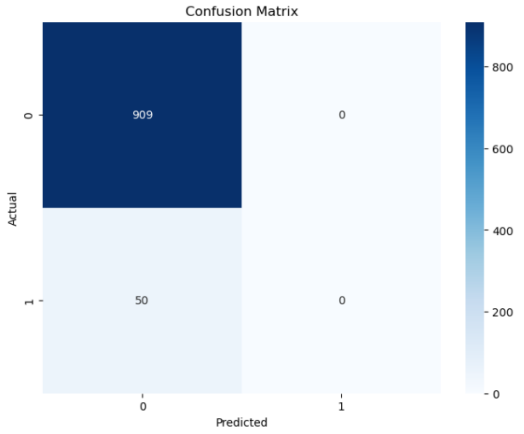



```
Accuracy: 0.913451511991658

Cross-Validation Scores:  [0.92075078 0.91762252 0.91753653 0.93319415 0.920668
06]
Mean Accuracy:  0.9219544105833973
Standard Deviation:  0.005791686145734314
```

Gambar 16. Visualisasi boxplot outlier telah dihapus

Model decision tree sebelum pruning mendapatkan akurasi sebesar 0.9134, yang menandakan bahwa model tersebut mampu mengklasifikasikan data dengan benar sebanyak 91,34% dari keseluruhan data yang diuji. Hasil dari cross-validation adalah 0.9208, 0.9176, 0.9175, 0.9332, dan 0.9207. Mean atau rata - rata akurasi dari cross-validation adalah 0.9220, menunjukkan bahwa model secara konsisten memiliki performa yang baik pada berbagai subset data. Standar deviasi sebesar 0.0058 menunjukkan bahwa variasi dalam skor akurasi cross-validation relatif kecil, mengindikasikan bahwa model memiliki kinerja yang stabil dan dapat diandalkan.

- Model decision tree dengan pruning (max_depth = 3)

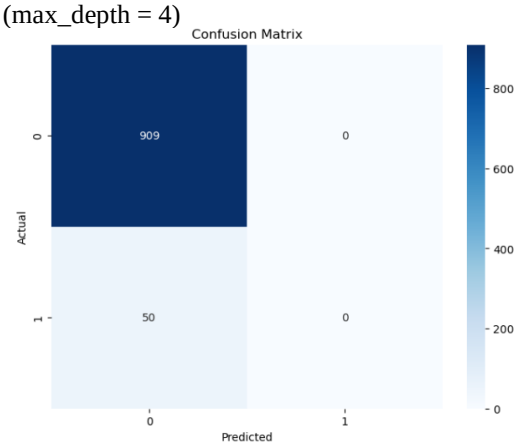



```
Accuracy: 0.9478623566214807

Cross-Validation Scores:  [0.95828989 0.95933264 0.95824635 0.95929019 0.959290
19]
Mean Accuracy:  0.9588898491600288
Standard Deviation:  0.0005080662439927236
```

Gambar 17. Visualisasi boxplot outlier telah dihapus

Model decision tree setelah pruning dengan max_depth = 3 mendapatkan akurasi sebesar 0.9479, yang menandakan bahwa model tersebut mampu mengklasifikasikan data dengan benar sebanyak 94,79% dari keseluruhan data yang diuji. Hasil dari cross-validation adalah 0.9583, 0.9593, 0.9582, 0.9593, dan 0.9593. Mean atau rata-rata akurasi dari cross-validation adalah 0.9589, menunjukkan bahwa model secara konsisten memiliki performa yang baik pada berbagai subset data. Standar deviasi sebesar 0.0005 menunjukkan bahwa variasi dalam skor akurasi cross-validation relatif kecil, mengindikasikan bahwa model memiliki kinerja yang stabil dan dapat diandalkan.

- Model decision tree dengan pruning (max_depth = 4)



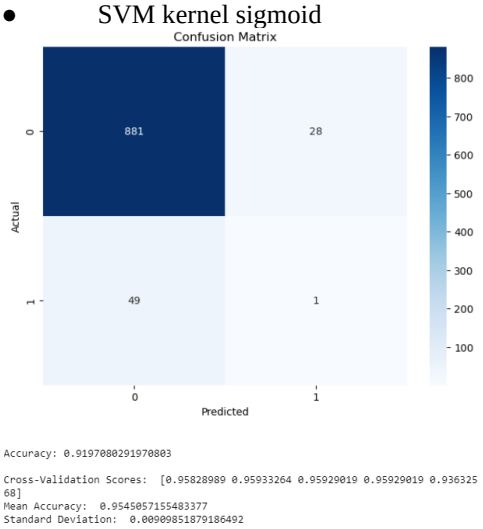

```
Accuracy: 0.9478623566214807

Cross-Validation Scores:  [0.95828989 0.95933264 0.95615866 0.95929019 0.959290
19]
Mean Accuracy:  0.9584723126255821
Standard Deviation:  0.0012218224231156933
```

Gambar 18. Visualisasi boxplot outlier telah dihapus

Model decision tree setelah pruning dengan max_depth = 4 mendapatkan sebesar 0.9479, yang

berarti model tersebut mampu mengklasifikasikan data dengan benar sebanyak 94,79% dari keseluruhan data yang diuji. Untuk mengevaluasi kinerja model lebih lanjut, dilakukan cross-validation yang menghasilkan skor akurasi sebagai berikut: 0.9583, 0.9593, 0.9562, 0.9593, dan 0.9593. Rata-rata akurasi dari cross-validation ini adalah 0.9585, menunjukkan bahwa model secara konsisten memiliki performa yang sangat baik pada berbagai subset data. Selain itu, standar deviasi sebesar 0.0012 menunjukkan bahwa variasi dalam skor akurasi cross-validation relatif kecil, mengindikasikan bahwa model memiliki kinerja yang stabil dan dapat diandalkan.

- SVM kernel sigmoid



Gambar 19. Visualisasi boxplot outlier telah dihapus

Hasil evaluasi SVM menunjukkan bahwa model memiliki akurasi sebesar 0.9197, yang berarti model tersebut mampu mengklasifikasikan data dengan benar sebanyak 91,97% dari keseluruhan data yang diuji. Melalui cross-validation, diperoleh skor akurasi sebagai berikut: 0.9583, 0.9593, 0.9593, 0.9593, dan 0.9363. Rata-rata akurasi dari cross-validation ini adalah 0.9545, menunjukkan bahwa model secara konsisten memiliki performa yang sangat baik pada berbagai subset data. Standar deviasi sebesar 0.0091 menunjukkan bahwa variasi dalam skor akurasi cross-validation relatif cukup kecil, menandakan bahwa model memiliki kinerja yang stabil dan dapat diandalkan.

G. Deployment

Pada tahap deployment, dilakukan perencanaan dengan menyusun langkah-langkah implementasi model decision tree. Selanjutnya, memastikan data disiapkan sesuai dengan kebutuhan seperti variabel yang paling berpengaruh yaitu kadar gula rata - rata untuk digunakan dalam mendiagnosis pasien. Implementasi model dilakukan dengan melakukan integrasi model decision tree ke dalam sistem rumah sakit. Selanjutnya dilakukan uji coba dan validasi model pada lingkungan operasional. Pelatihan juga dilakukan kepada staf medis tentang cara penggunaan model dan menginterpretasi hasilnya. Monitoring dan pemeliharaan model juga dilakukan untuk memastikan kinerja model tetap akurat dan relevan.

IV. RESULT AND ANALYSIS

1. Perbandingan hasil evaluasi

Berikut ini merupakan perbandingan hasil evaluasi dalam accuracy, cross-validation scores, mean accuracy, standard deviation.

- Hasil Evaluasi dari model pada penelitian ini.

	Decision Tree			SVM
	Sebelum Pruning	Setelah Pruning		
		Max_depth = 3	Max_depth = 4	
Accuracy	0.9134	0.9478	0.9478	0.9197
Cross-Validation Scores	[0.9207, 0.9176, 0.9175, 0.9331, 0.9206]	[0.9582, 0.9593, 0.9582, 0.9502, 0.9502]	[0.9582, 0.9593, 0.9561, 0.9502, 0.9592]	[0.9582, 0.9593, 0.9592, 0.9592, 0.9363]
Mean Accuracy	0.9219	0.9583	0.9584	0.9545
Standard Deviation	0.0057	0.0005	0.0012	0.0090

Tabel 1. Evaluasi penelitian sekarang

- Hasil Evaluasi dari model pada penelitian sebelumnya.

	Decision Tree	SVM (linear)	SVM (polynomial)
Accuracy	0.92	0.76	0/80

Tabel 2. Evaluasi penelitian sebelumnya

Berdasarkan hasil dari tabel kedua tabel tersebut, dapat disimpulkan bahwa, model dengan algoritma decision tree yang telah di pruning dengan fitur BMI dan avg_glucose_level mendapatkan hasil akurasi yang paling tinggi dibandingkan model lainnya. Selain itu, hasil cross-validation score, mean accuracy, dan standard deviation menunjukkan bahwa model dengan decision tree dengan pruning dan pemilihan fitur lebih baik dibandingkan model lainnya.

2. Analisis faktor paling berpengaruh dalam model decision tree.

- Model dengan max_depth = 3

```
|--- avg_glucose_level <= 171.08
|   |--- bmi <= 19.35
|   |   |--- class: 0
|   |--- bmi >  19.35
|   |   |--- bmi <= 44.15
|   |   |   |--- class: 0
|   |   |--- bmi >  44.15
|   |   |   |--- class: 0
|--- avg_glucose_level >  171.08
|   |--- avg_glucose_level <= 171.75
|   |   |--- class: 1
|   |--- avg_glucose_level >  171.75
|   |   |--- bmi <= 46.30
|   |   |   |--- class: 0
|   |   |--- bmi >  46.30
|   |   |   |--- class: 0
```

Gambar 20. Diagram decision tree dengan max depth=3

Berdasarkan hasil analisis struktur decision tree di atas avg_glucose_level adalah variabel yang paling

dominan dalam mempengaruhi klasifikasi apakah pasien berpotensi terkena stroke atau tidak karena menjadi faktor pembagi utama di root node. Selain itu, variabel bmi juga digunakan dalam beberapa percabangan keputusan, tetapi pengaruhnya tampaknya tidak signifikan dalam model ini, karena semua kelas yang dihasilkan memiliki nilai bmi yang sama (0).

- Model dengan max_depth = 4

```
|--- avg_glucose_level <= 171.08
|   |--- bmi <= 19.35
|   |   |--- class: 0
|   |--- bmi >  19.35
|   |   |--- bmi <= 44.15
|   |   |   |--- bmi <= 44.05
|   |   |   |   |--- class: 0
|   |   |   |--- bmi >  44.05
|   |   |   |   |--- class: 1
|   |   |--- bmi >  44.15
|   |   |   |--- class: 0
|--- avg_glucose_level >  171.08
|   |--- avg_glucose_level <= 171.75
|   |   |--- class: 1
|   |--- avg_glucose_level >  171.75
|   |   |--- bmi <= 46.30
|   |   |   |--- bmi <= 45.55
|   |   |   |   |--- class: 0
|   |   |   |--- bmi >  45.55
|   |   |   |   |--- class: 0
|   |   |--- bmi >  46.30
|   |   |   |--- class: 0
```

Gambar 21. Diagram decision tree dengan max depth=4

Berdasarkan hasil analisis struktur decision tree di atas, avg_glucose_level adalah variabel yang paling berpengaruh dalam menentukan klasifikasi tentang apakah seorang pasien berpotensi terkena stroke atau tidak. Variabel ini menjadi pembagi pertama di root node, menunjukkan pentingnya kadar glukosa rata-rata dalam prediksi model. Variabel bmi juga memiliki pengaruh, akan tetapi peran dan kontribusinya hanya muncul setelah nilai avg_glucose_level digunakan untuk pembagian awal.

V. CONCLUSION

Dari hasil model decision tree, terlihat bahwa rata-rata kadar glukosa (avg_glucose_level) digunakan sebagai variabel yang paling dominan dalam mempengaruhi klasifikasi potensi stroke pada pasien. Hal ini terlihat dari posisinya sebagai pembagi pertama di root node, menandakan bahwa kadar glukosa rata-rata dianggap sebagai indikator utama dalam model ini. Setelah kadar glukosa rata-rata, variabel BMI digunakan sebagai faktor pembagi tambahan, khususnya untuk pasien dengan kadar glukosa lebih rendah.

Peningkatan akurasi dicapai melalui proses pruning pada model dengan kedalaman maksimum (max_depth) 3 dan 4. Model yang di pruning menunjukkan peningkatan akurasi dibandingkan dengan model tanpa pruning. Selain akurasi, mean accuracy, cross-validation scores, dan standard deviation juga ditingkatkan setelah dilakukan pruning. Dari evaluasi, ditunjukkan bahwa strategi pruning dengan max_depth=3 dan 4 dianggap paling optimal untuk meningkatkan akurasi dan konsistensi model dalam mengklasifikasikan potensi stroke, karena hasil akurasi yang relatif sama dicapai oleh keduanya.

Dalam perbandingan performa, model decision tree yang telah di pruning dengan max_depth menunjukkan akurasi yang lebih baik dibandingkan dengan model SVM dengan kernel sigmoid. Selain akurasi, mean accuracy, cross-validation scores, dan standard deviation juga lebih baik ditunjukkan oleh model decision tree setelah pruning dengan max_depth.

Pada hasil perbandingan dengan penelitian sebelumnya dapat dilihat bahwa, hasil model decision tree setelah pruning pada penelitian ini menunjukkan akurasi sebesar 94,78%, lebih tinggi dibandingkan dengan akurasi 92% yang dicapai oleh model decision tree pada penelitian sebelumnya. Peningkatan ini mungkin disebabkan oleh penggunaan hanya fitur bmi dan avg_glucose_level dalam penelitian kami. Akurasi model decision tree kami juga lebih tinggi dibandingkan dengan akurasi 91,97% yang dicapai oleh model SVM dengan kernel sigmoid dan akurasi model SVM dari penelitian sebelumnya dengan kernel linear (76%) dan kernel polynomial (80%). Hal ini menunjukkan bahwa, performa terbaik dalam mengklasifikasikan potensi stroke diperoleh oleh model decision tree dengan pruning dengan pemilihan fitur avg_glucose_leve dan BMI. Hasil penelitian ini menekankan pentingnya teknik pruning dan pemilihan fitur yang tepat dalam pengembangan model machine learning untuk meningkatkan akurasi dan efektivitas dalam menyelesaikan masalah untuk membantu dalam klasifikasi orang yang berpotensi terkena stroke.

VI. REFERENCES

[1] S. S. .M, P. K, and P. V, “STROKE PREDICTION USING MACHINE LEARNING,” *IARJSET*, 2022, [Online]. Available: https://api.semanticscholar.org/CorpusID:237858421

[2] D. Dwilaksono, T. E. Fau, S. E. Siahaan, C. S. P. B. Siahaan, K. S. P. B. Karo, and T. Nababan, “Faktor-Faktor yang Berhubungan dengan Terjadinya Stroke Iskemik pada Penderita Rawat Inap,” *J. Penelit. Perawat Prof.*, vol. 5, no. 2, 2023, doi: 10.37287/jppp.v5i2.1433.

[3] S. Moradi *et al.*, “The Incidence of Stroke with Different BMI Ranges in Diabetic Patients,” *Int. J. Ayurvedic Med.*, vol. 14, no. 1, 2023, doi: 10.47552/ijam.v14i1.3351.

[4] X. Peng *et al.*, “Longitudinal Average Glucose Levels and Variance and Risk of Stroke: A Chinese Cohort Study,” *Int. J. Hypertens.*, vol. 2020, 2020, doi: 10.1155/2020/8953058.

[5] A. Setiawan, R. F. Waleska, M. A. Purnama, Rahmaddeni, and L. Efrizoni, “KOMPARASI ALGORITMA K-NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), DAN DECISION TREE DALAM KLASIFIKASI PENYAKIT STROKE,” *J. Inform. dan Rekayasa Elektron.*, vol. 7, no. 1, pp. 107–114, Apr. 2024, doi: 10.36595/JIRE.V7I1.1161.

[6] K. R. Sulaeman, C. Setianingsih, and R. E. Saputra, “Analisis Algoritma Support Vector Machine Dalam Klasifikasi

Penyakit Stroke," *e-Proceeding Eng.*, vol. 9, no. 3, 2022.

[7] F. M. J. Mehedi Shamrat, S. Chakraborty, M. M. Billah, P. Das, J. N. Muna, and R. Ranjan, "A comprehensive study on pre-pruning and post-pruning methods of decision tree classification algorithm," in *Proceedings of the 5th International Conference on Trends in Electronics and Informatics, ICOEI 2021*, 2021. doi: 10.1109/ICOEI51242.2021.9452898.

[8] C. E. D. Vanegas, J. C. G. Mejía, F. A. V. Agudelo, and D. E. S. Duran, "A Representation Based on Essence for the CRISP-DM Methodology," *Comput. y Sist.*, vol. 27, no. 3, 2023, doi: 10.13053/CyS-27-3-3446.

[9] T. Thomas, A. P. Vijayaraghavan, and S. Emmanuel, "Applications of Decision Trees," 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:213463984

[10] O. Njoku, "Decision Trees and Their Application for Classification and Regression Problems," *MSU Grad. Theses*, 2019.

[11] F. Aaboub, H. Chamlal, and T. Ouaderhman, "Analysis of the prediction performance of decision tree-based algorithms," in *2023 International Conference on Decision Aid Sciences and Applications, DASA 2023*, 2023. doi: 10.1109/DASA59624.2023.10286809.

[12] G. Nanfack, P. Temple, and B. Frénay, "Constraint Enforcement on Decision Trees: A Survey," *ACM Comput. Surv.*, vol. 54, no. 10, 2022, doi: 10.1145/3506734.

[13] E. E. Ogheneovo and P. A. Nlerum, "Iterative Dichotomizer 3 (ID3) Decision Tree: A Machine Learning Algorithm for Data Classification and Predictive Analysis," *Int. J. Adv. Eng. Res. Sci.*, vol. 7, no. 4, 2020, doi: 10.22161/ijaers.74.60.

[14] A. Rajeshkanna and K. Arunesh, "ID3 Decision Tree Classification: An Algorithmic Perspective based on Error rate," in *Proceedings of the International Conference on Electronics and Sustainable Communication Systems, ICESC 2020*, 2020. doi: 10.1109/ICESC48915.2020.9155578.

[15] S. Bishnoi and B. K. Hooda, "Decision Tree Algorithms and their Applicability in Agriculture for Classification," *J. Exp. Agric. Int.*, 2022, doi: 10.9734/jeai/2022/v44i730833.

[16] X. Yao, "Application of Optimized SVM in Sample Classification," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, 2022, doi: 10.14569/IJACSA.2022.0130666.

[17] L. Sunitha and M. B. Raju, "Multi-class classification for large datasets with optimized SVM by non-linear kernel function," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/2089/1/012015.

[18] Z. Jun, "The Development and Application of Support Vector Machine," in *Journal of Physics: Conference Series*, 2021. doi: 10.1088/1742-6596/1748/5/052006.

[19] Ł. Gadomer and Z. A. Sosnowski, "Pruning trees in C-fuzzy random forest," *Soft Comput.*, vol. 25, no. 3, 2021, doi: 10.1007/s00500-020-05270-3.

[20] Y. Manzali and P. M. El Far, "A new decision tree pre-pruning method based on nodes probabilities," in *2022 International Conference on Intelligent Systems and Computer Vision, ISCV 2022*, 2022. doi: 10.1109/ISCV54655.2022.9806124.

Peran anggota :

- Dasmond Tan (00000070110)

Mencari jurnal, menyusun artikel, membuat model

- Willsen Wijaya (00000070011)

Membuat model, menyusun artikel, mencari jurnal

- James Andersen (00000069612)

Mencari jurnal, melakukan preprocessing data

- Lian wira manuel maharaja (00000075938)

Mencari jurnal, menampilkan visualisasi data

- Gregorius Daniel Dwitama (00000075740)

Mencari jurnal, menampilkan visualisasi data