

BAB IV

HASIL PENELITIAN DAN PEMBAHASAN

4.1 Hasil Penelitian

Berdasarkan metodologi yang telah dirancang untuk membandingkan tingkat akurasi tertinggi dalam memprediksi kasus stunting pada balita dengan menggunakan algoritma *decision tree*, *naïve bayes*, *support vector machine*. maka hasil yang didapatkan adalah sebagai berikut :

4.1.1 Data Collection

Data yang diperoleh dari kaggle berjumlah 7573 kasus balita stunting dan tidak stunting dalam format csv.

	Gender	Age	Birth Weight	Birth Length	Body Weight	Body Length	Breastfeeding	Stunting
0	female	56	2.9	50	11.0	90.0	Yes	No
1	female	20	3.3	49	11.1	80.5	No	No
2	male	4	2.8	48	6.5	63.0	No	No
3	female	14	2.0	49	7.0	71.0	Yes	No
4	male	32	3.2	49	11.0	88.7	Yes	No

Gambar 4.1 Sampel Data

4.1.2 Expolatory Data Analysis (EDA)

a) *Duplicate*

Berikut hasil setelah dilakukan pemeriksaan, diketahui bahwa dataset tidak mengandung data duplikat dan jumlah record tetap 7573 data.



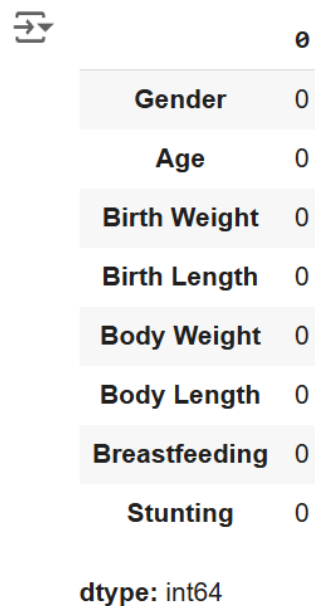
```
df.duplicated().sum()
```

0

Gambar 4.2 *Duplicate Data*

b) Missing Value

Hasil setelah dilakukan pemeriksaan missing value diketahui bahwa dataset tidak memiliki missing values sehingga data dapat diolah lebih lanjut membuat tidak ada pengurangan terhadap dataset yang tetap berjumlah 7573 data. Hasil pengecekan pada setiap variabel adalah:



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

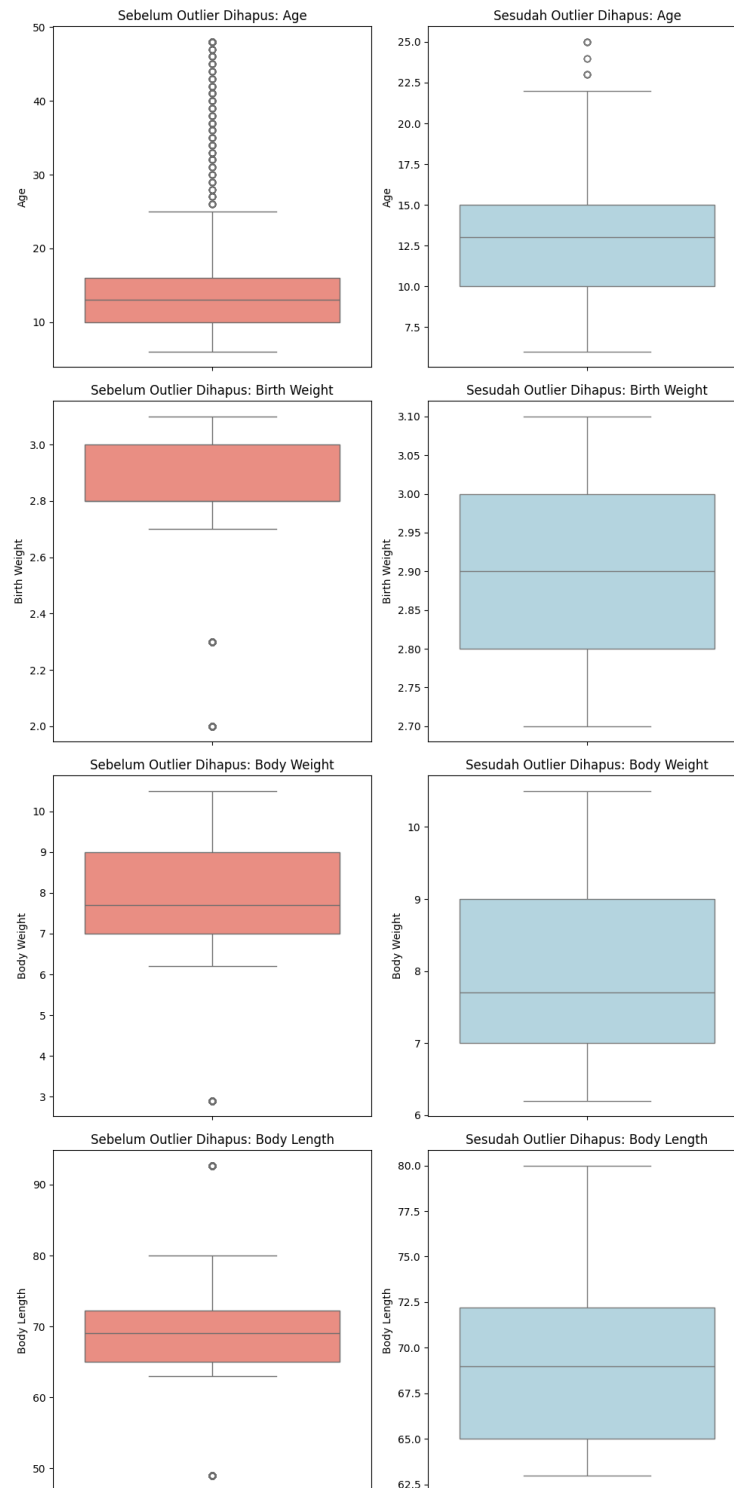
Gender	0
Age	0
Birth Weight	0
Birth Length	0
Body Weight	0
Body Length	0
Breastfeeding	0
Stunting	0

dtype: int64

Gambar 4.3 *Missing Value*

c) Outlier

Mengecek kemungkinan adanya outlier dengan mem-visualisasikannya menggunakan boxplot, setelah penghapusan outlier jumlah data menjadi 4480 record data, hasil visualisasinya adalah sebagai berikut:



Gambar 4.4 *Outlier*

```

→ Kolom: Age
Q1: 10.0
Q2 (Median): 13.0
Q3: 16.0
Ambang Batas Atas (Maksimum): 25.0
Ambang Batas Bawah (Minimum): 1.0
Jumlah Outlier Melebihi Maksimum: 690
Jumlah Outlier Kurang dari Minimum: 0
-----
Kolom: Birth Weight
Q1: 2.8
Q2 (Median): 2.8
Q3: 3.0
Ambang Batas Atas (Maksimum): 3.3000000000000003
Ambang Batas Bawah (Minimum): 2.4999999999999996
Jumlah Outlier Melebihi Maksimum: 0
Jumlah Outlier Kurang dari Minimum: 1294
-----
Kolom: Birth Length
Q1: 49.0
Q2 (Median): 49.0
Q3: 49.0
Ambang Batas Atas (Maksimum): 49.0
Ambang Batas Bawah (Minimum): 49.0
Jumlah Outlier Melebihi Maksimum: 1068
Jumlah Outlier Kurang dari Minimum: 427
-----
Kolom: Body Weight
Q1: 7.0
Q2 (Median): 7.7
Q3: 9.0
Ambang Batas Atas (Maksimum): 12.0
Ambang Batas Bawah (Minimum): 4.0
Jumlah Outlier Melebihi Maksimum: 0
Jumlah Outlier Kurang dari Minimum: 509
-----
Kolom: Body Length
Q1: 65.0
Q2 (Median): 69.0
Q3: 72.2
Ambang Batas Atas (Maksimum): 83.0
Ambang Batas Bawah (Minimum): 54.199999999999996
Jumlah Outlier Melebihi Maksimum: 489
Jumlah Outlier Kurang dari Minimum: 609
-----
Kolom: Stunting
Q1: 1.0
Q2 (Median): 1.0
Q3: 1.0
Ambang Batas Atas (Maksimum): 1.0
Ambang Batas Bawah (Minimum): 1.0
Jumlah Outlier Melebihi Maksimum: 0
Jumlah Outlier Kurang dari Minimum: 1453
-----

```

Gambar 4.5 *Upper lower*

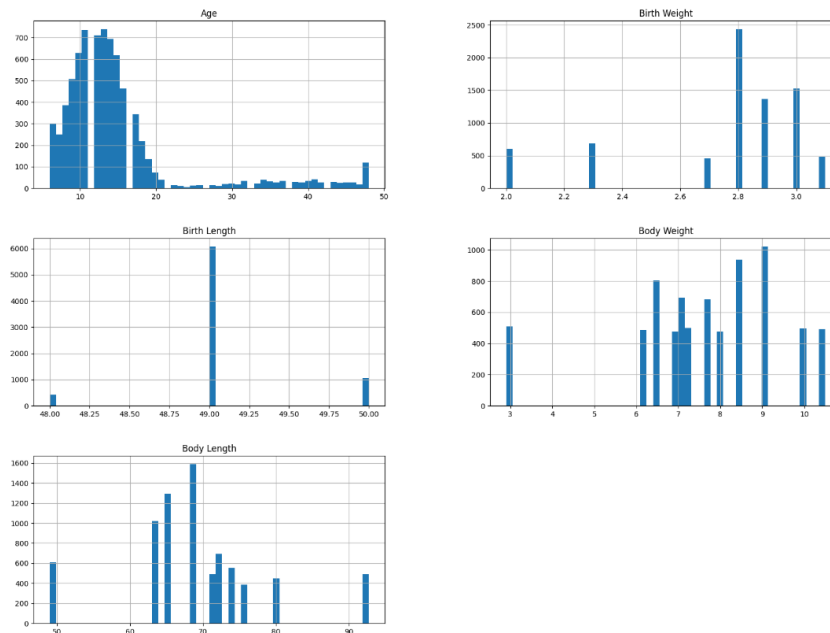
Pada penelitian ini pemeriksaan outlier menggunakan IQR (Inter Quartile Range) digunakan untuk mendeteksi outlier dalam data dengan cara menentukan batas bawah (lower) dan batas atas (upper) di luar dimana data dianggap sebagai outlier.

d) Distribusi Data

Memvisualisasikan distribusi data dengan memanfaatkan histogram agar lebih mudah dilihat dan dipahami.

	Gender	Age	Birth Weight	Birth Length	Body Weight	Body Length	Breastfeeding	Stunting
count	7573.000000	7573.000000	7573.000000	7573.000000	7573.000000	7573.000000	7573.0	7573.000000
mean	0.608874	14.886967	2.762208	49.084643	7.630648	69.120269	0.0	0.808134
std	0.488035	8.580291	0.297725	0.436202	1.762529	9.401527	0.0	0.393794
min	0.000000	6.000000	2.000000	48.000000	2.900000	49.000000	0.0	0.000000
25%	0.000000	10.000000	2.800000	49.000000	7.000000	65.000000	0.0	1.000000
50%	1.000000	13.000000	2.800000	49.000000	7.700000	69.000000	0.0	1.000000
75%	1.000000	16.000000	3.000000	49.000000	9.000000	72.200000	0.0	1.000000
max	1.000000	48.000000	3.100000	50.000000	10.500000	92.700000	0.0	1.000000

Gambar 4.0.6 Output Distribusi Data



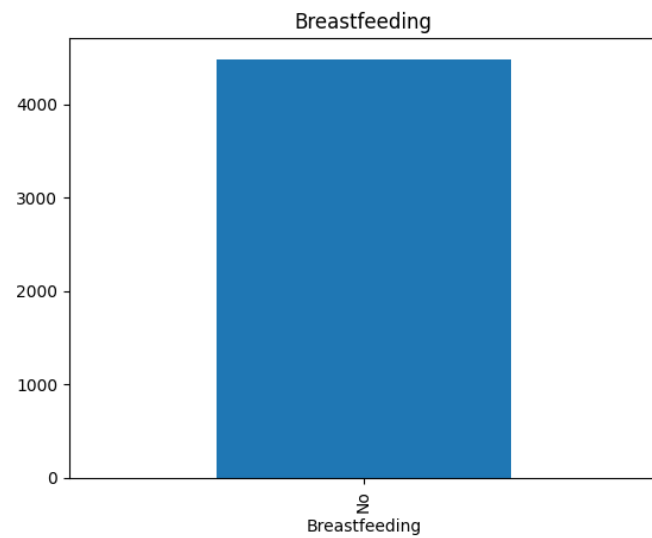
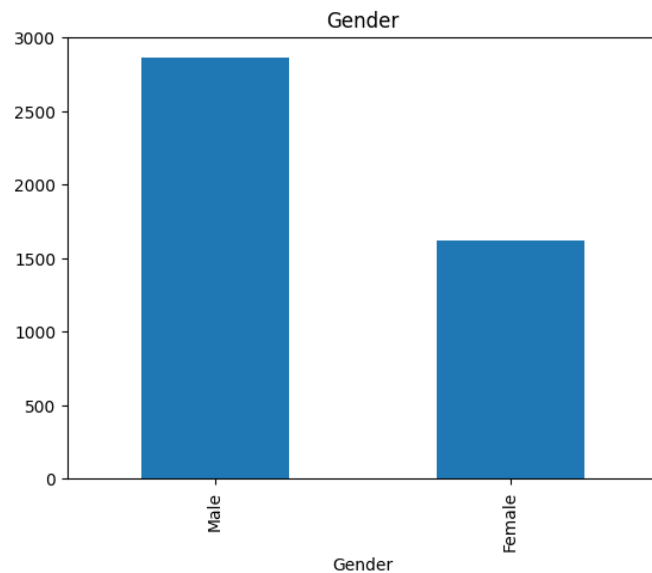
Gambar 4.7 Histogram Distribusi Data

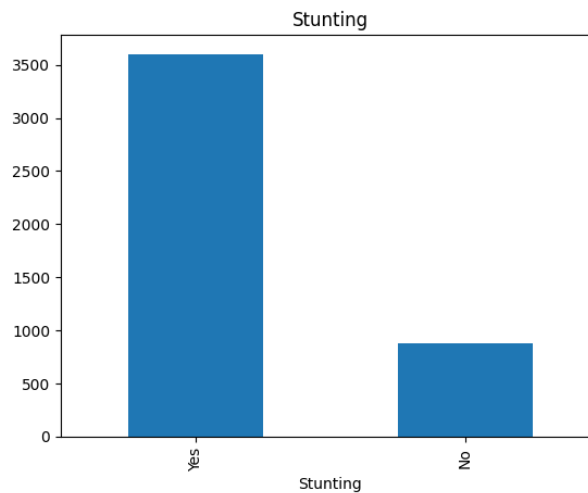
Distribusi data pada fitur-fitur yang ditampilkan menunjukkan beberapa karakteristik penting. Fitur kategori seperti Gender, Breastfeeding, dan Stunting menunjukkan distribusi yang tidak seimbang, di mana salah satu kategori memiliki jumlah yang jauh lebih dominan dibandingkan kategori lainnya. Untuk fitur numerik seperti Age, data cenderung skewed ke arah tertentu (positively skewed), dengan mayoritas individu berada pada usia yang lebih muda. Selain itu, beberapa fitur numerik seperti Birth Weight, Birth Length, dan Body Weight menunjukkan pola diskret, yang kemungkinan mencerminkan data yang terkelompok atau kurang variatif.

e) Univariate analysis

Analisis univariat dari dataset menunjukkan bahwa mayoritas sampel berjenis kelamin laki-laki sebanyak 2868 (63,8%), sedangkan perempuan sebanyak 1620 (36,2%), sehingga terdapat ketidakseimbangan yang perlu

diperhatikan dalam analisis lebih lanjut. Pada variabel Breastfeeding, seluruh sampel memiliki kategori "No" sebanyak 4488 (100%), yang menunjukkan tidak adanya variasi dalam variabel ini dan kemungkinan tidak berkontribusi dalam pemodelan. Sementara itu, pada variabel Stunting, sebagian besar sampel berada pada kategori "Yes" sebanyak 3603 (80,4%), sedangkan kategori "No" hanya mencakup 877 (19,6%).

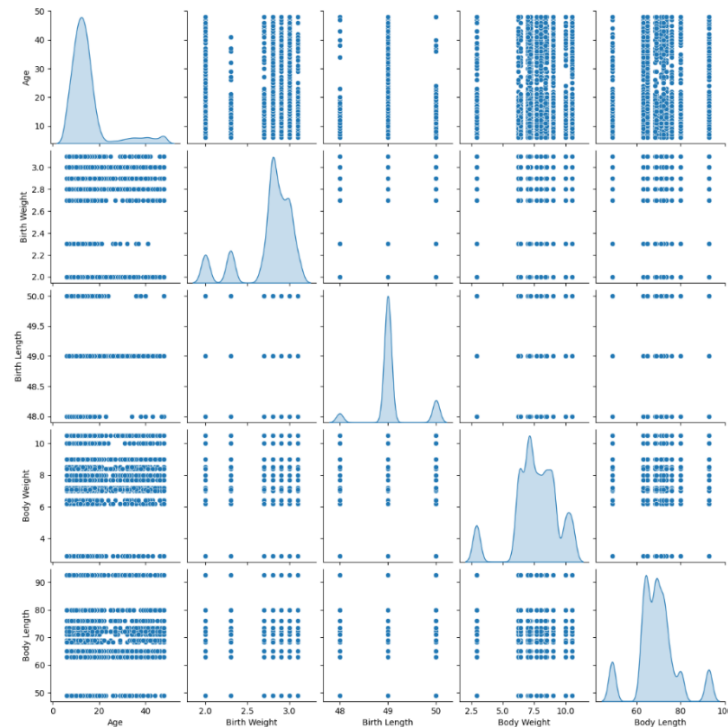




Gambar 4.8 *Univariate Analysis*

f) Multivariate Analysis

pada tahap ini saya menggunakan grafik pairplot ini digunakan untuk menganalisis hubungan multivariat antara variabel-variabel numerik dalam dataset, yaitu Age, Birth Weight, Birth Length, Body Weight, dan Body Length. Distribusi univariat pada diagonal menunjukkan pola penyebaran data yang menggambarkan karakteristik masing-masing variabel. Variabel Age cenderung memiliki distribusi yang didominasi oleh sampel pada kelompok usia muda, sementara variabel seperti Birth Weight, Birth Length, Body Weight, dan Body Length menunjukkan pola distribusi yang teratur dan terfokus. Hubungan antar variabel menunjukkan berbagai pola, misalnya Body Weight dan Body Length memiliki indikasi korelasi positif yang berarti, yang dapat menggambarkan keterkaitan yang logis antara kedua variabel tersebut. Hubungan antara Age dengan variabel lain, seperti Body Weight dan Body Length, menunjukkan pola yang bervariasi tanpa indikasi hubungan linier yang dominan.



Gambar 4.9 *Multivariate Analysis*

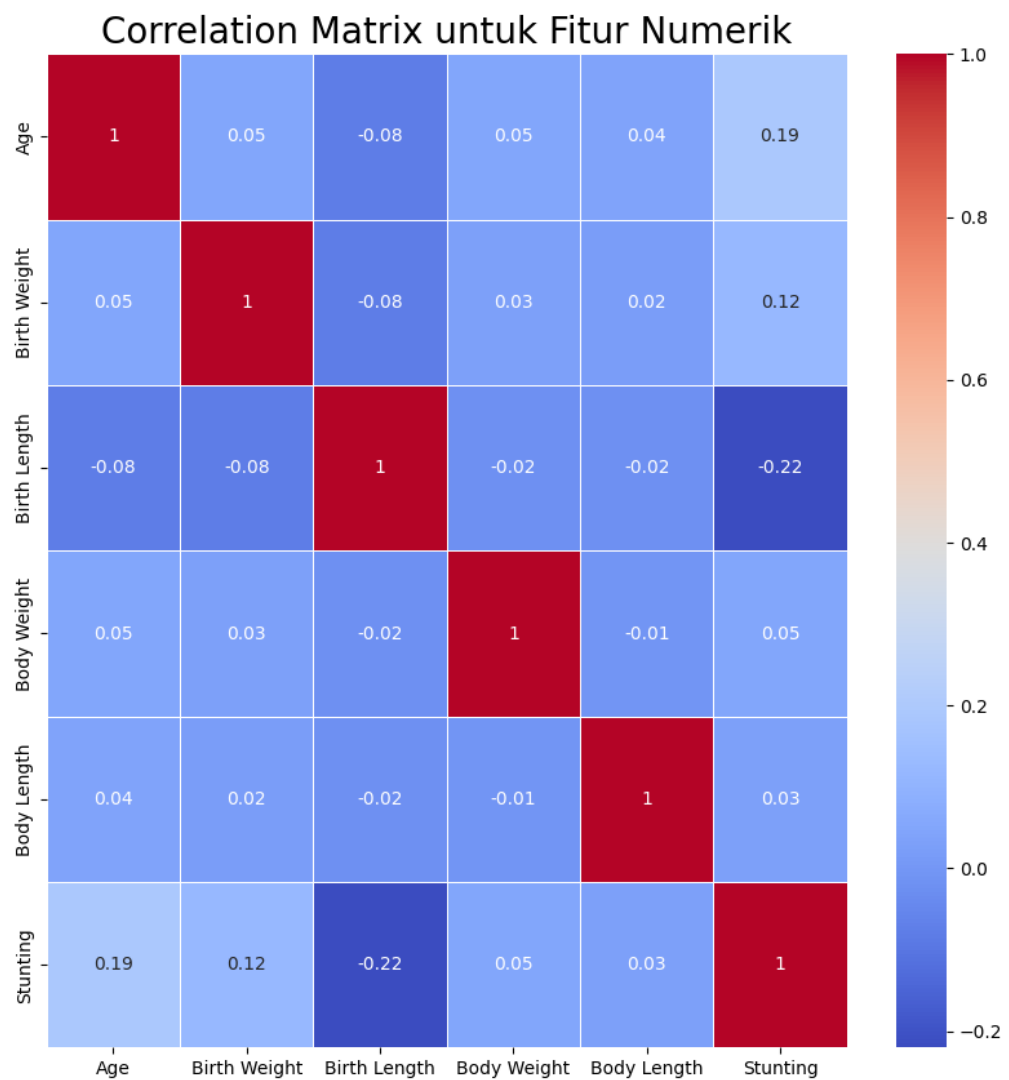
g) Matrix Colleration

Berdasarkan matriks korelasi yang ditampilkan melalui heatmap, hubungan antar fitur numerik dalam dataset ini sebagian besar sangat lemah. Korelasi antara Age (Usia) dan Stunting adalah 0.19, yang menunjukkan korelasi positif lemah, mengindikasikan bahwa semakin bertambah usia, sedikit meningkatkan kemungkinan stunting, meskipun tidak signifikan. Korelasi antara Birth Weight (Berat Badan Lahir) dan Stunting adalah 0.12, yang juga sangat lemah, menunjukkan pengaruh minimal berat badan lahir terhadap stunting.

Hubungan antara Birth Length (Panjang Badan Lahir) dan Stunting memiliki korelasi negatif lemah sebesar -0.22, yang menunjukkan bahwa semakin

panjang tubuh bayi saat lahir, kemungkinan stunting sedikit menurun. Sementara itu, hubungan Body Weight (Berat Badan) dan Stunting adalah 0.05, serta Body Length (Panjang Badan) dan Stunting adalah 0.03, keduanya menunjukkan korelasi positif yang sangat lemah, dengan pengaruh yang tidak signifikan terhadap kejadian stunting.

Korelasi antara Age dan fitur lainnya (Birth Weight, Birth Length, Body Weight, Body Length) berkisar antara -0.08 hingga 0.05, yang menunjukkan hubungan sangat lemah atau hampir tidak ada. Secara keseluruhan, fitur-fitur numerik dalam dataset ini relatif independen satu sama lain, dan risiko multikolinearitas dapat diabaikan. Meskipun terdapat korelasi positif lemah antara Age dan Stunting, serta korelasi negatif lemah antara Birth Length dan Stunting, hubungan-hubungan ini tidak cukup kuat untuk memengaruhi model secara signifikan. Oleh karena itu, fitur-fitur ini tetap dapat digunakan untuk analisis atau pemodelan lebih lanjut tanpa perlu penghapusan berdasarkan nilai korelasi.



Gambar 4.10 *Colleration Matrix*

4.1.3 Data Preprocessing

a. Normalisasi Data

Hasil output dari proses normalisasi data yang telah dilakukan :

	Gender	Age	Birth Weight	Birth Length	Body Weight
0	0.801484	0.246282	0.798750	-0.194058	1.344380
1	-1.247685	-0.453041	0.462847	-0.194058	-2.684189
2	0.801484	0.129728	0.462847	-0.194058	0.493274
3	0.801484	1.878036	0.126945	-0.194058	-0.698275
4	0.801484	0.013174	1.134652	-0.194058	1.628082
...	...	...	...	...	...
7568	0.801484	-0.453041	0.126945	-2.486722	1.628082
7569	0.801484	0.013174	0.126945	-0.194058	-2.684189
7570	-1.247685	-0.336487	0.126945	-2.486722	0.039350
7571	0.801484	0.129728	0.126945	-0.194058	0.039350
7572	-1.247685	-0.569595	0.798750	-0.194058	0.039350

	Body Length	Breastfeeding	Stunting
0	0.327599	0.0	-2.052310
1	-0.438284	0.0	0.487256
2	0.327599	0.0	0.487256
3	-0.651030	0.0	0.487256
4	-2.140248	0.0	0.487256
...	...	...	...
7568	0.465884	0.0	-2.052310
7569	0.199952	0.0	0.487256
7570	-0.651030	0.0	-2.052310
7571	-2.140248	0.0	-2.052310
7572	1.157307	0.0	0.487256

Gambar 4.11 *Normalisasi Data*

b. Data Encoding

Tahap ini mengubah atau mentransformasi data pada setiap fitur menjadi angka sehingga data dapat dilakukan proses training. Pada kolom Stunting, fitur Yes akan diubah menjadi 1 dan fitur No akan diubah menjadi 0.

	Gender	Age	Birth Weight	Birth Length	Body Weight	Body Length	Breastfeeding	Stunting
0	1	17	3.0	49	10.0	72.2	0	0
1	0	11	2.9	49	2.9	65.0	0	1
2	1	16	2.9	49	8.5	72.2	0	1
3	1	31	2.8	49	6.4	63.0	0	1
4	1	15	3.1	49	10.5	49.0	0	1

Gambar 4.12 *Data Encoding*

c. *Splitting Data*

Tahap berikutnya adalah membagi dataset menjadi dua bagian, yaitu data latih (training set) dan data uji (test set) rasio perbandingannya 80:20, proses ini penting untuk mengevaluasi performa model secara objektif. Data latih digunakan untuk membangun dan mengoptimalkan model pada penelitian ini berjumlah 3584 record, sementara data uji digunakan untuk mengevaluasi seberapa baik model dapat menggeneralisasi data baru yang tidak terlihat selama pelatihan berjumlah 896 record data.

```
# split data into train and test
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report, confusion_matrix

X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, random_state=42)

print(f"Total train data : {len(X_train)}")
print(f"Total test data : {len(X_test)}")
```

Gambar 4.13 *Splitting Data*

Output :

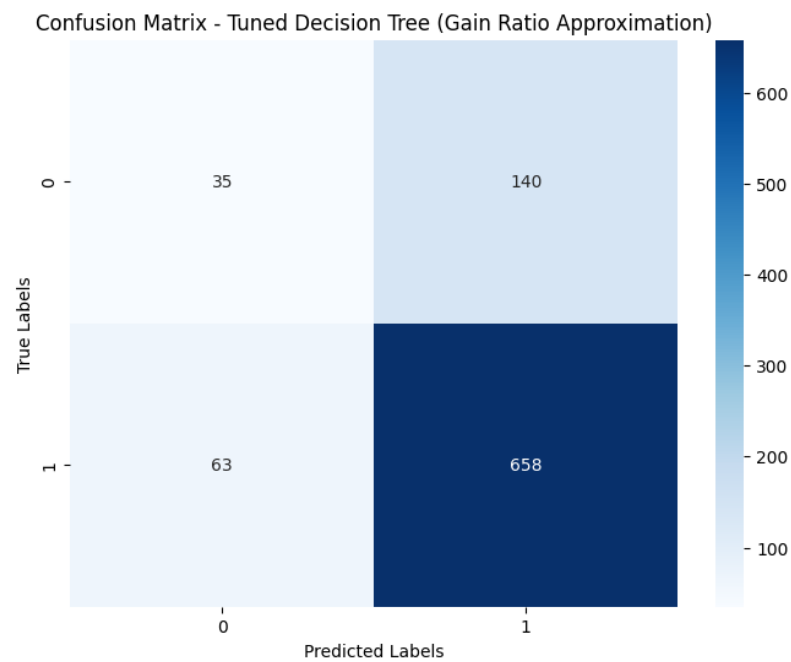
```
➡ Total train data : 3584
   Total test data : 896
```

Gambar 4.14 *Output Splitting Data*

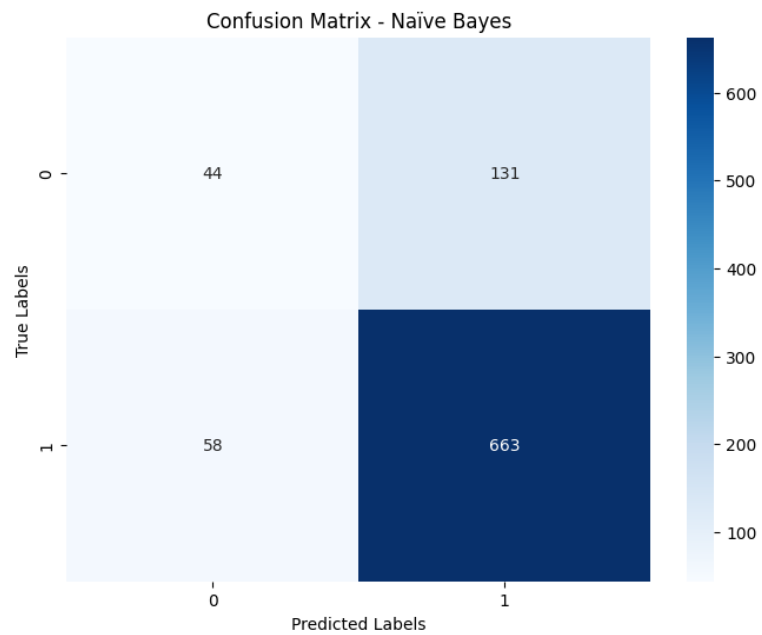
4.1.4 Model Validation

a. Confution Matrix

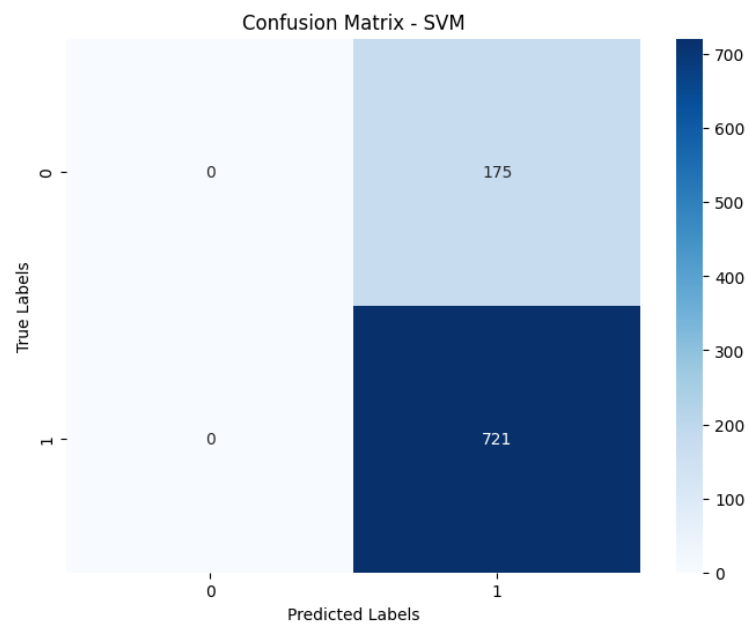
Confution matrix ini memberikan gambaran rinci mengenai hasil prediksi model dibandingkan dengan data sebenarnya, termasuk jumlah prediksi yang benar maupun salah untuk setiap kelas.



Gambar 4.15 *Confution Matrix Decision Tree*



Gambar 4.16 *Confution Matrix Naïve Bayes*

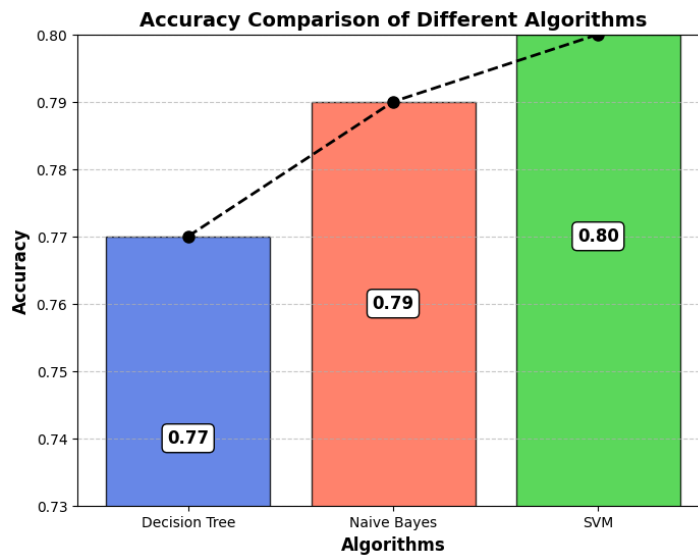


Gambar 4.17 *Confution Matrix SVM*

4.1.5 Model Evaluation

a. Akurasi

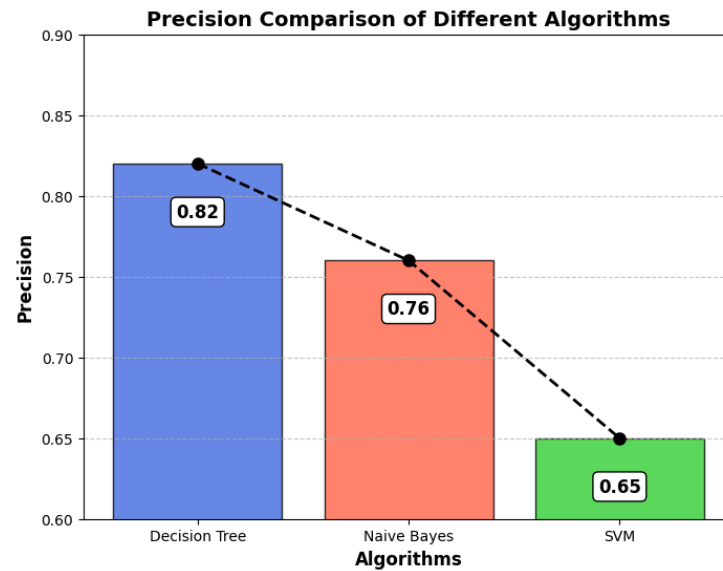
Berikut Hasil perbandingan akurasi algoritma *decision tree*, *naïve bayes* dan *support vector machine* untuk kasus stunting pada balita :



Gambar 4.18 Akurasi

b. Precision

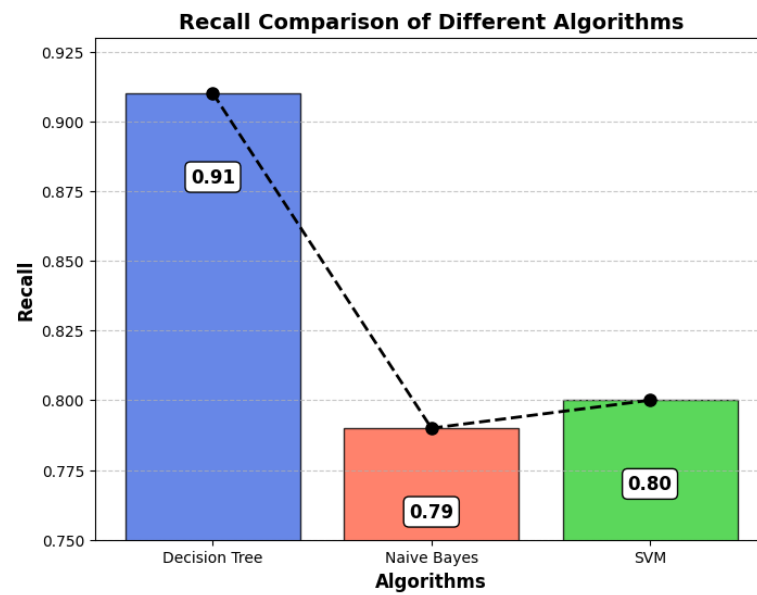
Berikut Hasil perbandingan precision algoritma *decision tree*, *naïve bayes* dan *support vector machine* untuk kasus stunting pada balita :



Gambar 4.19 *precision*

c. Recall

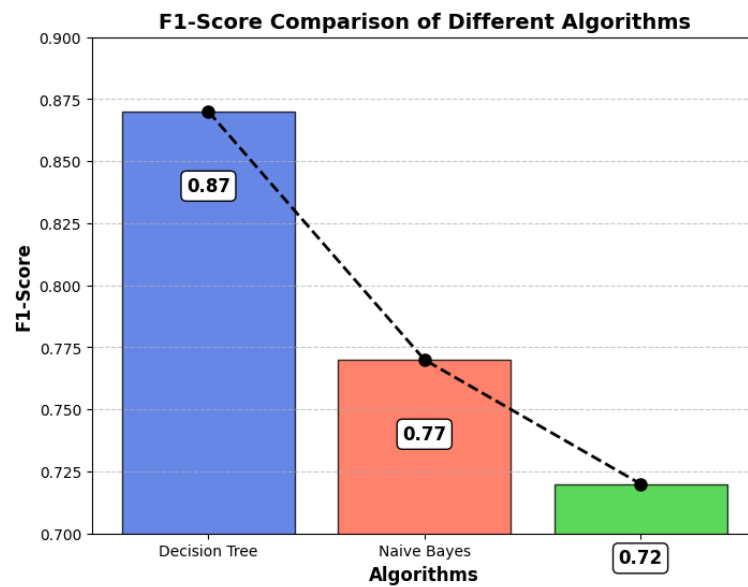
Berikut Hasil perbandingan Recall algoritma *decision tree*, *naïve bayes* dan *support vector machine* untuk kasus stunting pada balita :



Gambar 4.20 *Recall*

d. *F1-Score*

Berikut Hasil perbandingan F1-scorei algortima *decision tree*, *naïve bayes* dan *support vector machine* untuk kasus stunting pada balita :



Gambar 4.21 *F1-Score*