

Challenge Chapter Gold

Proses Analisis Perbandingan dan Analisis Deskriptif Tweet dengan menggunakan metode algoritma Decision Tree Classifier pada konten Tweet Abusive

Prabowo Nofieldi

Binar Academy – DSC 18

Background

Penelitian ini mengeksplorasi analisis konten tweet abusive dengan metode Decision Tree Classifier. Dengan meningkatnya kebencian online, deteksi otomatis menjadi krusial. Penelitian ini menggunakan algoritma Decision Tree Classifier dan membandingkannya dengan metode lain. Fokus juga diberikan pada analisis deskriptif untuk memahami karakteristik dan konteks penggunaan bahasa kasar dalam tweet. Hasilnya diharapkan dapat meningkatkan pemahaman terhadap perilaku kebencian online, memperkuat deteksi otomatis, dan menyumbang pada pengembangan alat analisis konten yang lebih canggih.



Rumusan Masalah

01.

Sejauh mana efektivitas metode Decision Tree Classifier dalam mengenali tweet yang dianggap abusive ?

02.

Apa saja ciri-ciri dan konteks penggunaan bahasa kasar atau merendahkan dalam tweet abusive yang dapat diidentifikasi melalui analisis deskriptif ?

Tujuan

01

Mengevaluasi
Efektivitas
Decision Tree
Classifier

02

Identifikasi Ciri-ciri
dan Konteks
Penggunaan
Bahasa Kasar

Metode Penelitian (1)

{Dataset}

Dalam penelitian ini terdapat beberapa dataset yang digunakan seperti;

- Dataset Tweeter (data.csv)
- Dataset Abusive (abusive.csv)
- Dataset kamus alay
(new_kamusalay.csv)

Metode Penelitian (2)

{Metode Analisa}

Analisa Deskriptif:

- Menghitung statistik deskriptif seperti rata-rata, median, modus, dan deviasi standar untuk masing-masing kolom numerik.
- Menghitung jumlah nilai unik atau frekuensi masing-masing kategori pada kolom kategori.

Analisis Perbandingan:

- Membandingkan distribusi atau statistik antar kategori tertentu, seperti, membandingkan tingkat kekerasan (HS) antara kelompok-kelompok berbeda seperti HS_Individual, HS_Group, dsb

Metode Penelitian (3)

{Detail Dataset}

Dapat dilihat isi dataset Tweeter (data.csv) terdapat table yang memiliki 13 kolom dan 13169 baris

- Tweet,
- HS_Abusive,
- HS_Individual,
- HS_Group,
- HS_Religion,
- HS_Race,
- HS_Physical,
- HS_Gender,
- HS_Other,
- HS_Weak,
- HS_Moderate,
- HS_Strong.

	Tweet	HS	Abusive	HS_Individual	HS_Group	HS_Religion	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	HS_Moderate	HS_Strong
0	- disaat semua cowok berusaha melacak perhatia...	1	1	1	0	0	0	0	0	1	1	0	0
1	RT USER: USER siapa yang telat ngasih tau elu?...	0	1	0	0	0	0	0	0	0	0	0	0
2	41. Kadang aku berfikir, kenapa aku tetap perc...	0	0	0	0	0	0	0	0	0	0	0	0
3	USER USER AKU ITU AKU\nKU TAU MATAMU SIPIT T...	0	0	0	0	0	0	0	0	0	0	0	0
4	USER USER Kaum cebong kapir udah keliatan dong...	1	1	0	1	1	0	0	0	0	0	1	0

```
• print("Shape: ", data.shape)
data.head(15)
✓ 0.0s
Shape: (13169, 13)
```

Metode Penelitian (4)

{Detail Dataset}

Dapat dilihat isi dataset kamus alay (new_kamusalay.csv) terdapat table yang memiliki 2 kolom dan 15167 baris;

- original
- placement

```
print("Shape: ", alay_dict.shape)
alay_dict.head(15)
```

Shape: (15167, 2)

	original	placement
0	anakjakartaasikasik	anak jakarta asyik asyik
1	pakcikdahtua	pak cik sudah tua
2	pakcikmudalagi	pak cik muda lagi
3	t3tapjokowi	tetap jokowi

Metode Penelitian (5)

{Detail Dataset}

Dapat dilihat isi dataset abusive (abusive.csv) terdapat table yang memiliki 1 kolom dan 124 baris, dan ini juga merupakan bagaimana cara melihat ciri-ciri Tweet yang temasuk HS(hate speech) dengan menggunakan kalimat yang terdapat pada dataset abusive.

df_abusive	
✓	0.0s
ABUSIVE	
0	alay
1	ampas
2	buta
3	keparat
4	anjing
...	...
120	rezim
121	sange
122	serbet
123	sipit
124	transgender
125 rows × 1 columns	

Hasil Penelitian (1)

{Analisa Perbandingan}

Dalam analisa ini kita membutuhkan data count dari HS dan Abusive, untuk mendapatkan data Toxic shape.

```
HS
0    7608
1    5561
Name: count, dtype: int64
```

```
Abusive
0     8126
1     5043
Name: count, dtype: int64
```

```
Toxic shape: (7309, 13)
Non toxic shape: (5860, 13)
```

Hasil Penelitian (2)

{Analisa Perbandingan}

Setelah mengelompokkan dataset yang memiliki dataset HS, Abusive, dan Toxic shape. Selanjutnya kita dapat melakukan pembuatan model Decision Tree Classifier untuk melihat jumlah data yang termasuk HS dan Non-HS. Dapat kita lihat terdapat 1118 row HS dan 1516 Non-HS.

Confusion Matrix:

```
[[1516    0]
 [  0 1118]]
```

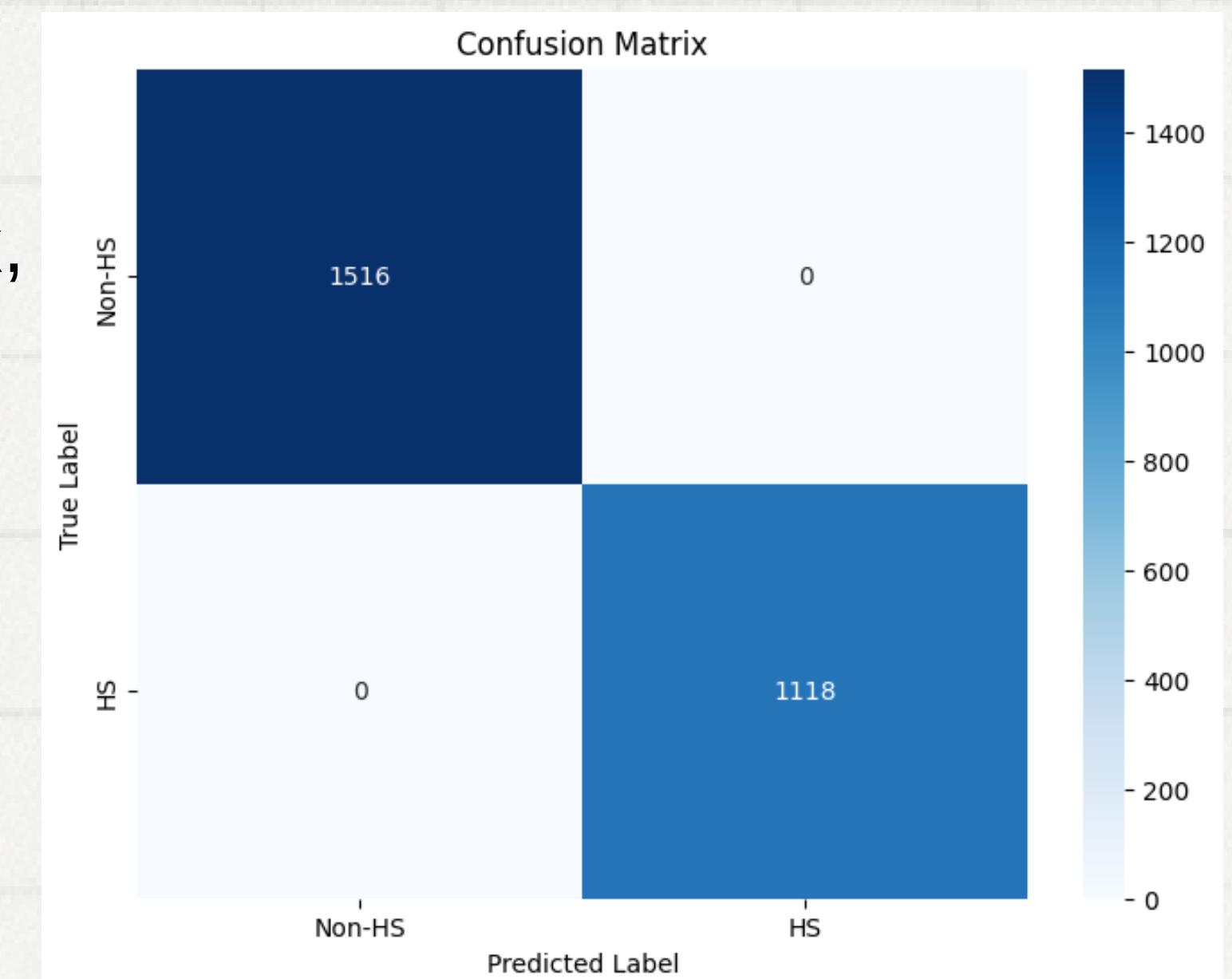
Classification Report:

	precision	recall	f1-score	support
0	1.00	1.00	1.00	1516
1	1.00	1.00	1.00	1118
accuracy			1.00	2634
macro avg	1.00	1.00	1.00	2634
weighted avg	1.00	1.00	1.00	2634

Hasil Penelitian (3)

{Analisa Perbandingan}

Visualisasi menggunakan Confusion Matrix,
pada data HS dan Non-HS.



Hasil Penelitian (4)

{Analisa Deskriptif}

Dalam analisis ini, kita menampilkan masing kelompok kolom pada table data.csv, untuk dapat melihat pembagian hasil perhitungan statistika, seperti ;

1. Min (Minimum): Nilai terkecil dalam data.
2. Q1 (Kuartil Pertama): Nilai yang membagi 25% data terendah.
3. Median (Median atau Kuartil Kedua): Nilai tengah dalam data ketika diurutkan.
4. Q3 (Kuartil Ketiga): Nilai yang membagi 75% data terendah.
5. Max (Maksimum): Nilai tertinggi dalam data.
6. Mean : Nilai rata-rata dalam data.
7. std : Deviasi Standar adalah suatu ukuran seberapa tersebar atau seberapa jauh data tersebut dari nilai rata-rata dalam sebuah kumpulan data.

	HS	Abusive	HS_Individual	HS_Group	HS_Religion	\
count	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	
mean	0.422280	0.382945	0.271471	0.150809	0.060217	
std	0.493941	0.486123	0.444735	0.357876	0.237898	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	1.000000	1.000000	1.000000	0.000000	0.000000	0.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

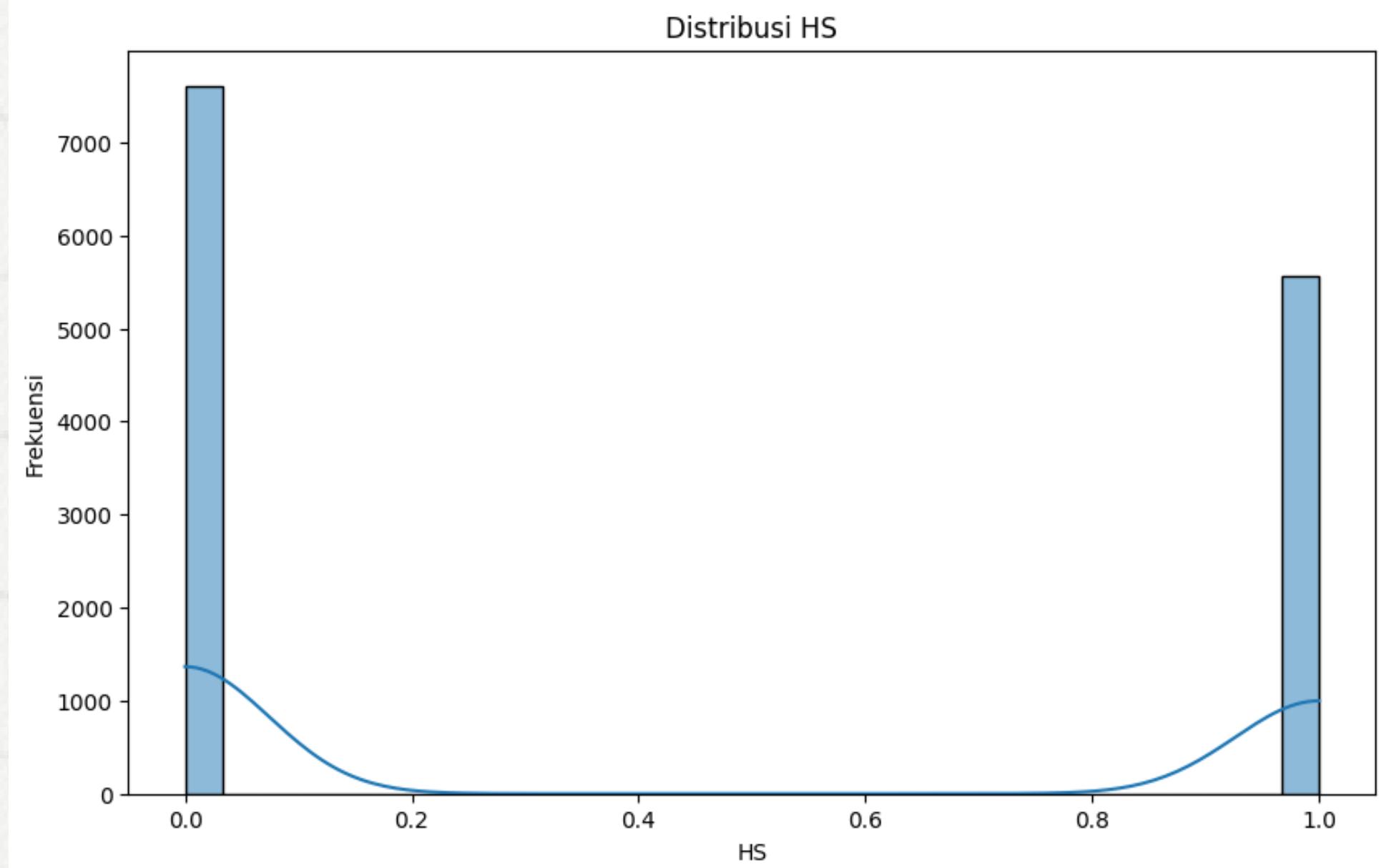
	HS_Race	HS_Physical	HS_Gender	HS_Other	HS_Weak	\
count	13169.000000	13169.000000	13169.000000	13169.000000	13169.000000	
mean	0.042980	0.024527	0.023236	0.284000	0.256891	
std	0.202819	0.154685	0.150659	0.450954	0.436935	
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	1.000000	1.000000	1.000000
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

	HS_Moderate	HS_Strong
count	13169.000000	13169.000000
mean	0.129471	0.035918
std	0.335733	0.186092
min	0.000000	0.000000
25%	0.000000	0.000000
50%	0.000000	0.000000
75%	0.000000	0.000000
max	1.000000	1.000000

Hasil Penelitian (5)

{Hasil Visualisasi}

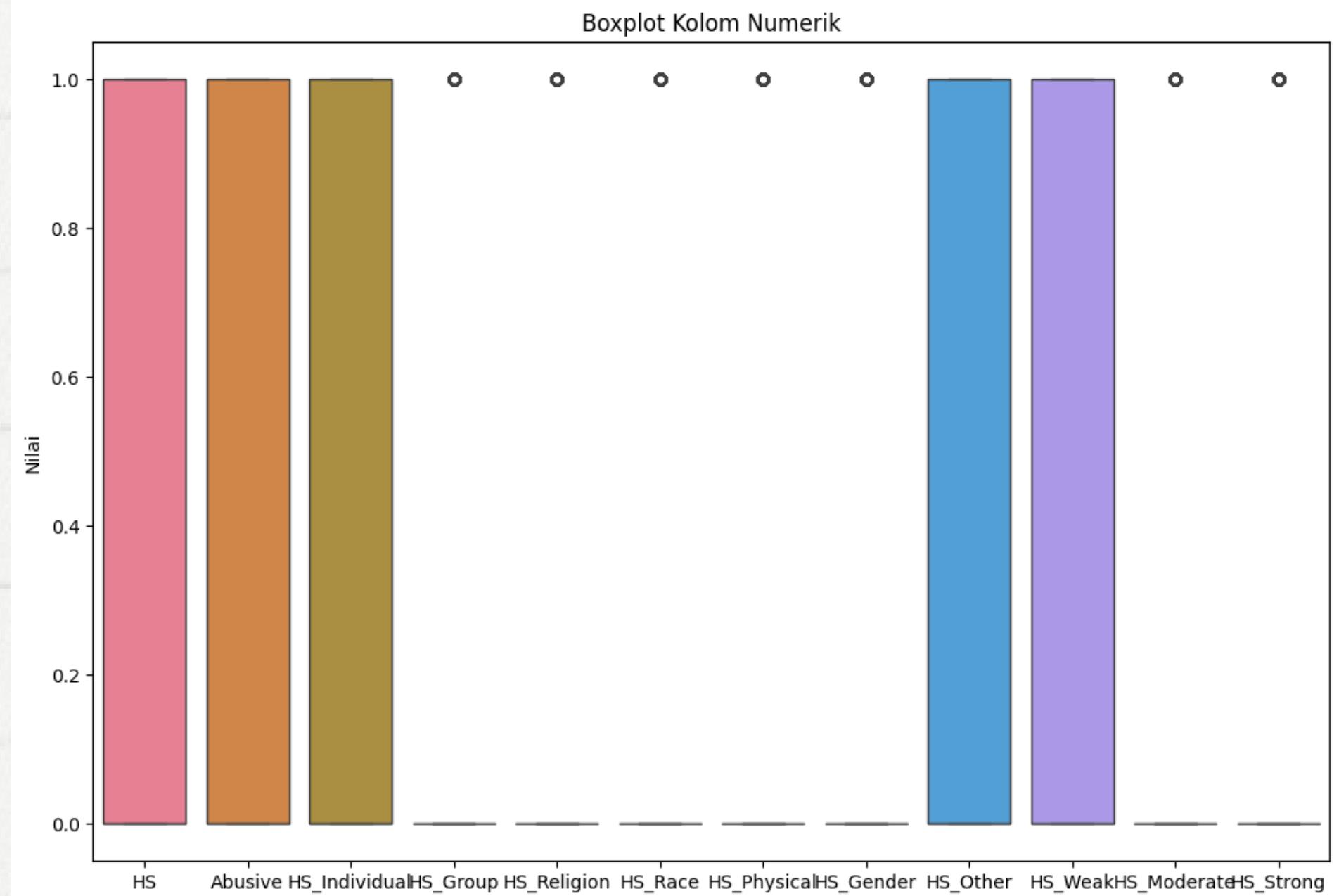
Visualisasi histogram untuk kolom HS



Hasil Penelitian (6)

{Hasil Visualisasi}

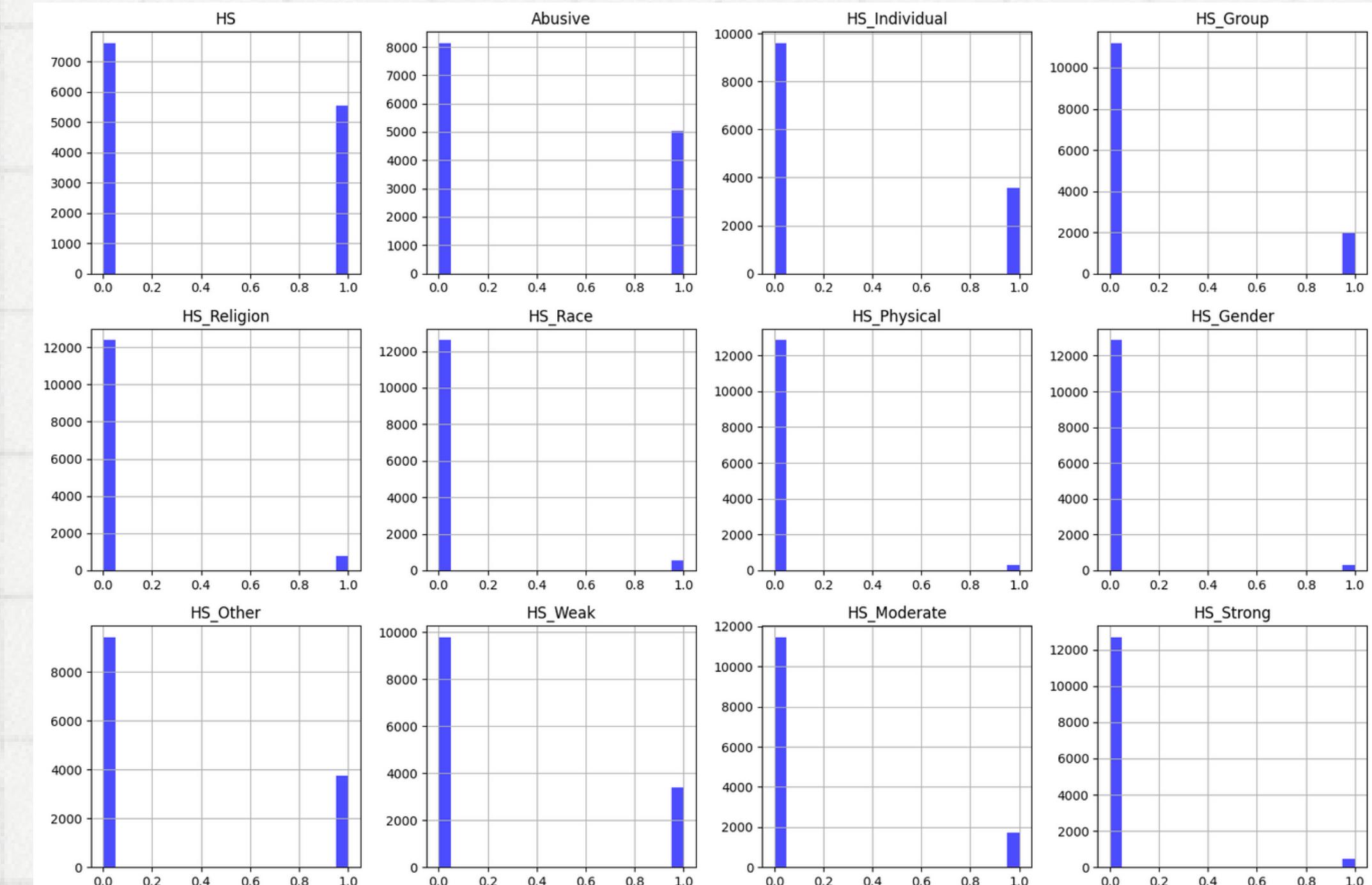
Visualisasi boxplot untuk beberapa kolom numerik



Hasil Penelitian (7)

{Hasil Visualisasi}

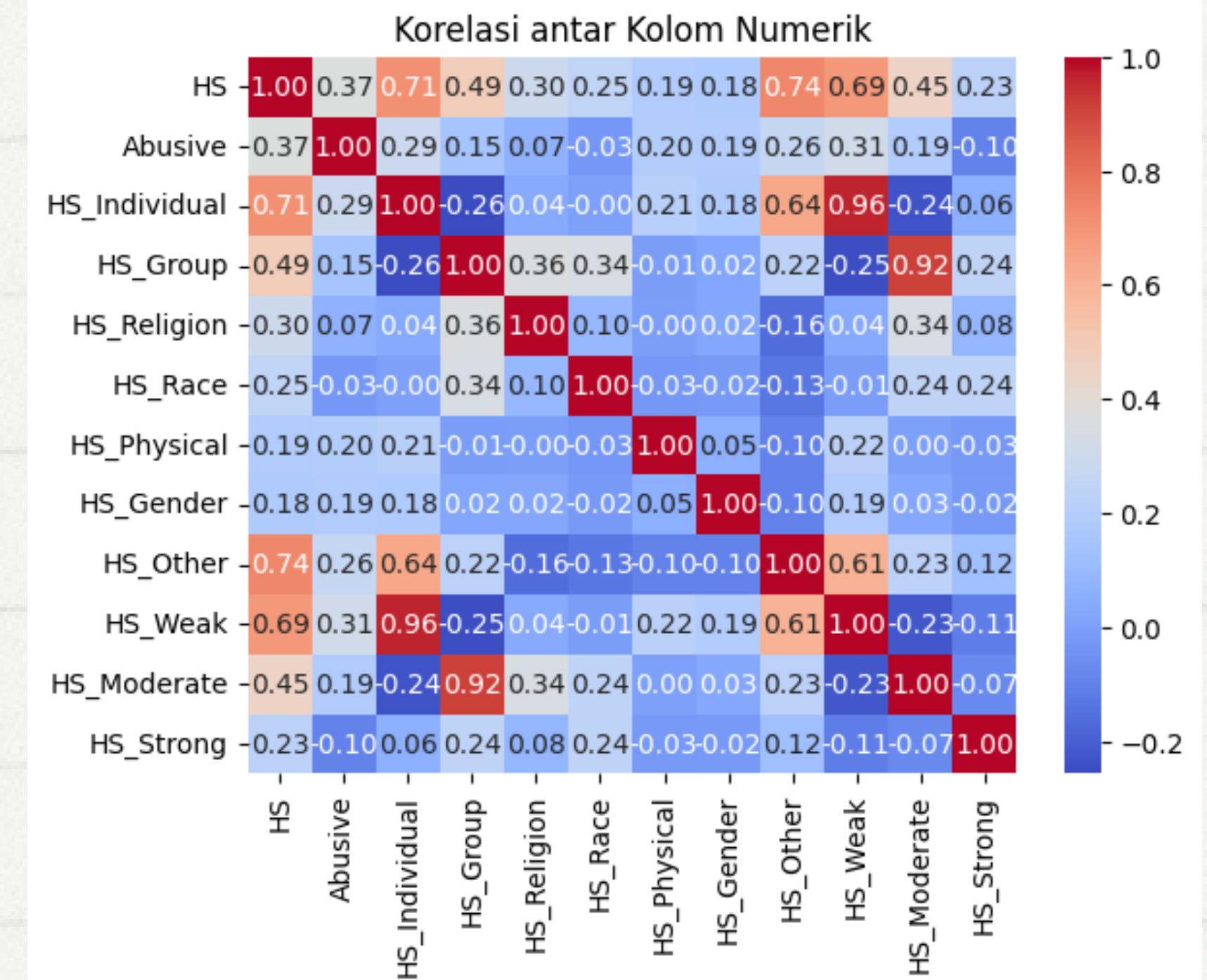
Visualisasi histogram untuk kolom numerik



Hasil Penelitian (8)

{Hasil Visualisasi}

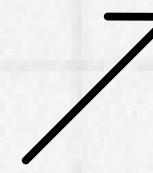
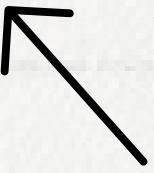
Visualisasi korelasi antar kolom numerik (Heatmap), Visualisasi ini membantu untuk memahami sejauh mana kolom-kolom numerik saling berhubungan (korelasi) dalam dataset. Nilai korelasi berkisar antara -1 hingga 1, di mana 1 menunjukkan hubungan positif sempurna, -1 menunjukkan hubungan negatif sempurna, dan 0 menunjukkan tidak adanya korelasi.



1

Performa Model:

- Melalui confusion matrix dan classification report, Anda dapat melihat seberapa baik model Decision Tree Classifier dapat memprediksi kelas "HS" dan "Non-HS".
- Perhatikan nilai seperti akurasi, presisi, recall, dan F1-score untuk mengevaluasi kinerja model.



4

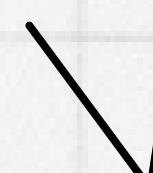
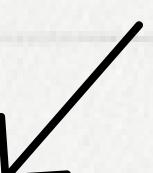
Perbaikan dan Pengembangan:

- Jika performa model kurang memuaskan, pertimbangkan untuk menyesuaikan parameter model atau mencoba algoritma klasifikasi yang berbeda.
- Analisis visualisasi dan kurva ROC, precision-recall, serta fitur penting dapat memberikan wawasan tambahan untuk perbaikan model.

2

Confusion Matrix:

- Matriks konfusi menunjukkan berapa banyak prediksi benar dan salah untuk setiap kelas.
- Diagonal utama mewakili prediksi yang benar, sementara elemen di luar diagonal utama menunjukkan kesalahan prediksi.



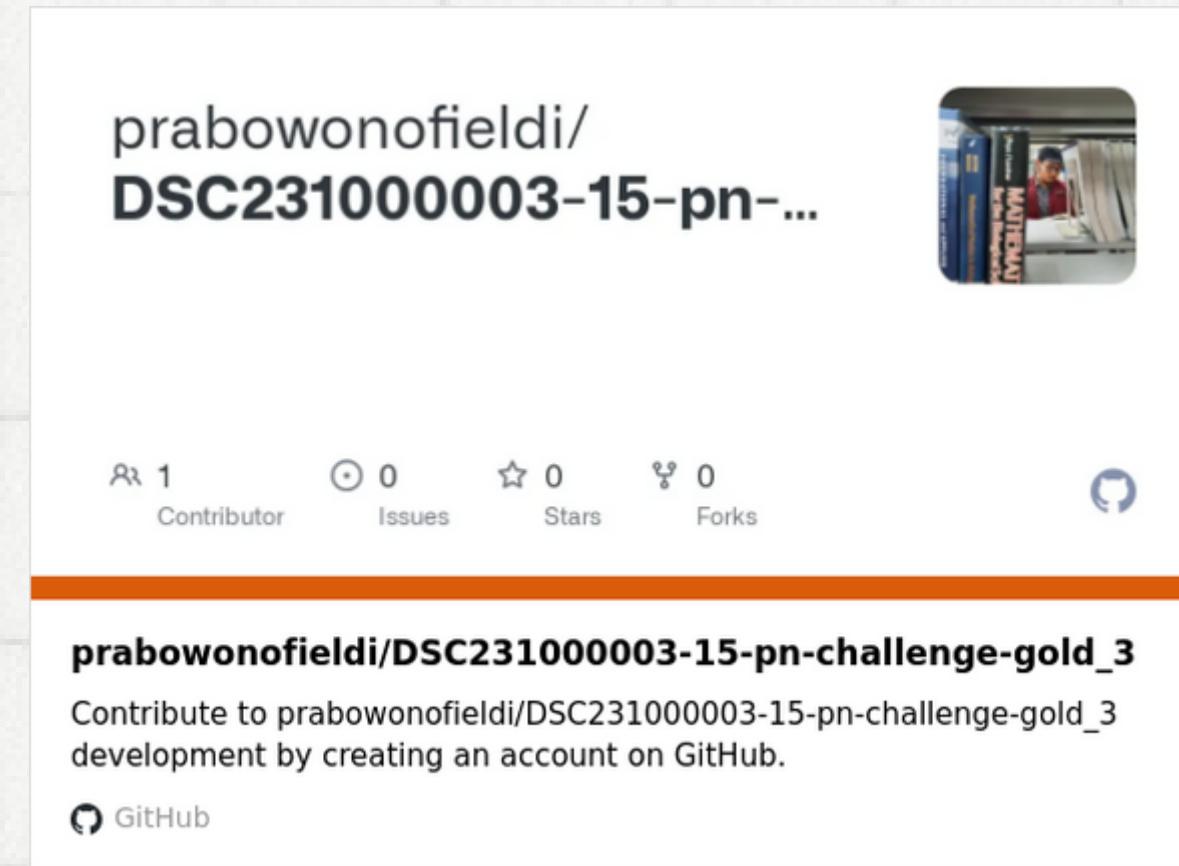
3

Visualisasi:

- Heatmap confusion matrix memberikan representasi visual tentang seberapa baik model dapat membedakan antara kelas "HS" dan "Non-HS".
- Warna biru pada diagonal utama menunjukkan prediksi yang benar, sementara warna yang berbeda menunjukkan kesalahan prediksi.

Kesimpulan

Link Github



**Thank you
very much!**