

LAPORAN PROJECT

STATISTIKA INFERENSI

KELOMPOK 7TEKNIK INFORMATIKA
KELAS B

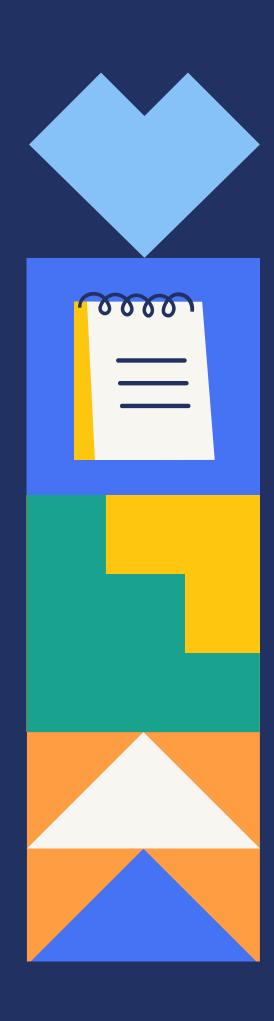


Anggota Kelompok KELOMPOK 7

- Izzat Ikhwan Hadi 225150200111010
- Muhammad Arsya Zain Yashifa 225150200111008
- Muhammad Hasan Fadhlillah 225150207111026
- Muhammad Husain Fadhlillah 225150207111027







Presentation Outline



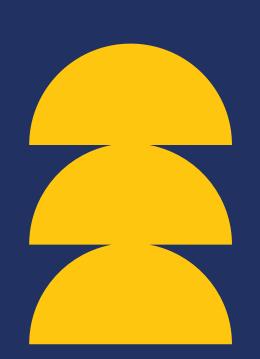


- EDA
- Uji Normalitas Data
- Transformasi Data

- Klasifikasi
- Uji Hipotesis

EXPLORATORY DATA ANALYSIS (EDA)





Informasi Dasar Dataset

Jumlah Total Sampel	Jumlah Fitur	
5656	13	

• Total Sampel:

5656 - Jumlah total sampel/observasi dalam dataset.

• Jumlah Fitur:

13 - Dataset memiliki 13 fitur/variabel.

Tipe Data Setiap Kolom

Kolom	Tipe Data	
contrast-1-0	float64	
correlation-1-0	float64	
dissimilarity-1-0	float64	
contrast-1-45	float64	
correlation-1-45	float64	
dissimilarity-1-45	float64	
contrast-1-90	float64	
correlation-1-90	float64	
dissimilarity-1-90	float64	
contrast-1-135	float64	
correlation-1-135	float64	
dissimilarity-1-135	float64	
Class	int64	



Semua kolom adalah tipe data float64 kecuali kolom 'Class' yang bertipe int64. Ini menunjukkan bahwa mayoritas fitur adalah numerik kontinu atau desimal, sedangkan 'Class' adalah fitur numerik integer atau bilangan bulat.



Missing Values

Kolom	Total Missing Values
contrast-1-0	0
correlation-1-0	0
dissimilarity-1-0	0
contrast-1-45	0
correlation-1-45	0
dissimilarity-1-45	0
contrast-1-90	0
correlation-1-90	0
dissimilarity-1-90	0
contrast-1-135	0
correlation-1-135	0
dissimilarity-1-135	0
Class	0

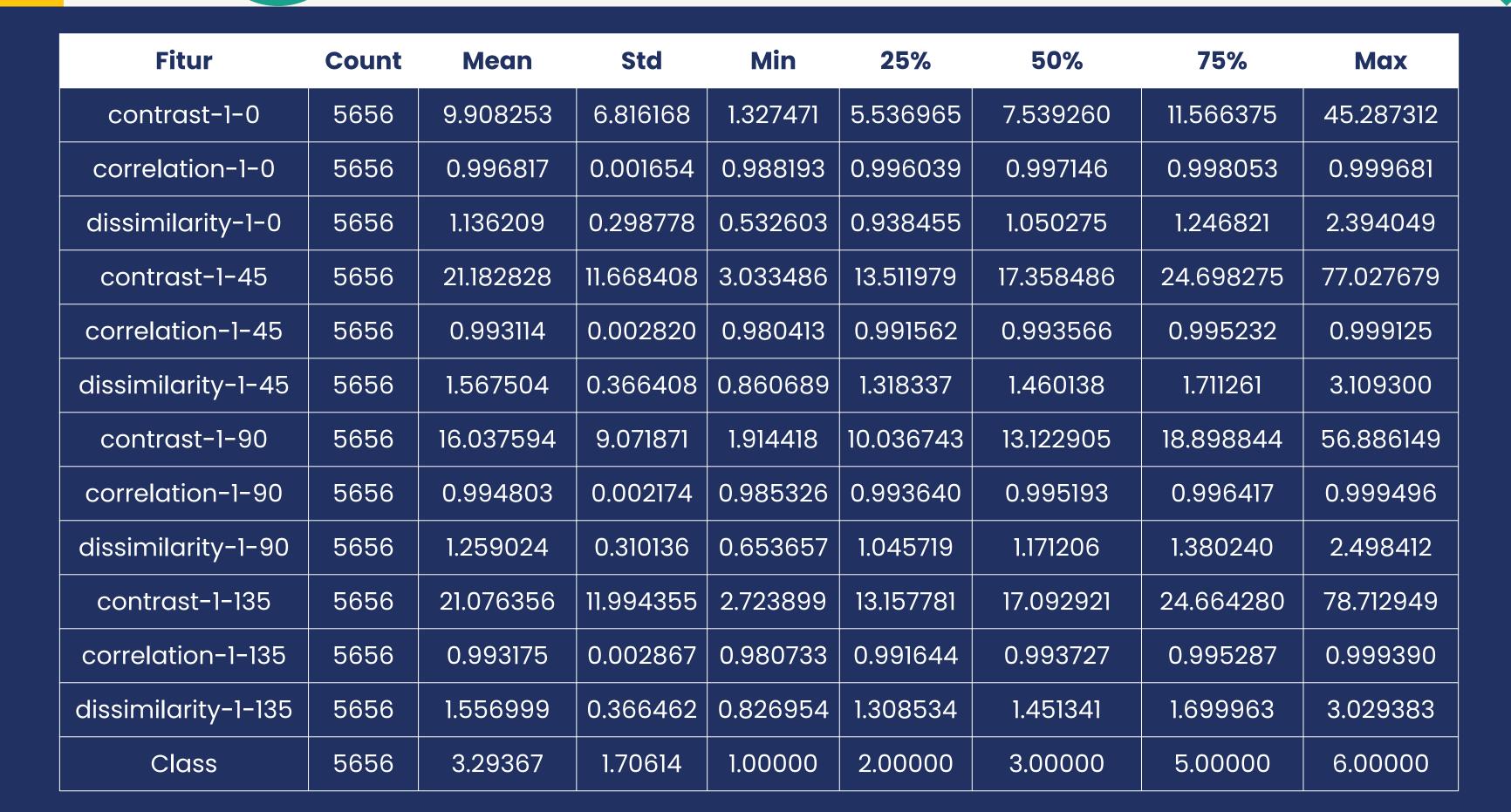


Tidak ada missing values pada dataset. Ini berarti tidak ada data yang hilang atau tidak terisi pada fitur-fitur, sehingga data dapat dianalisis secara baik.





Statistik Deskriptif Keseluruhan Data



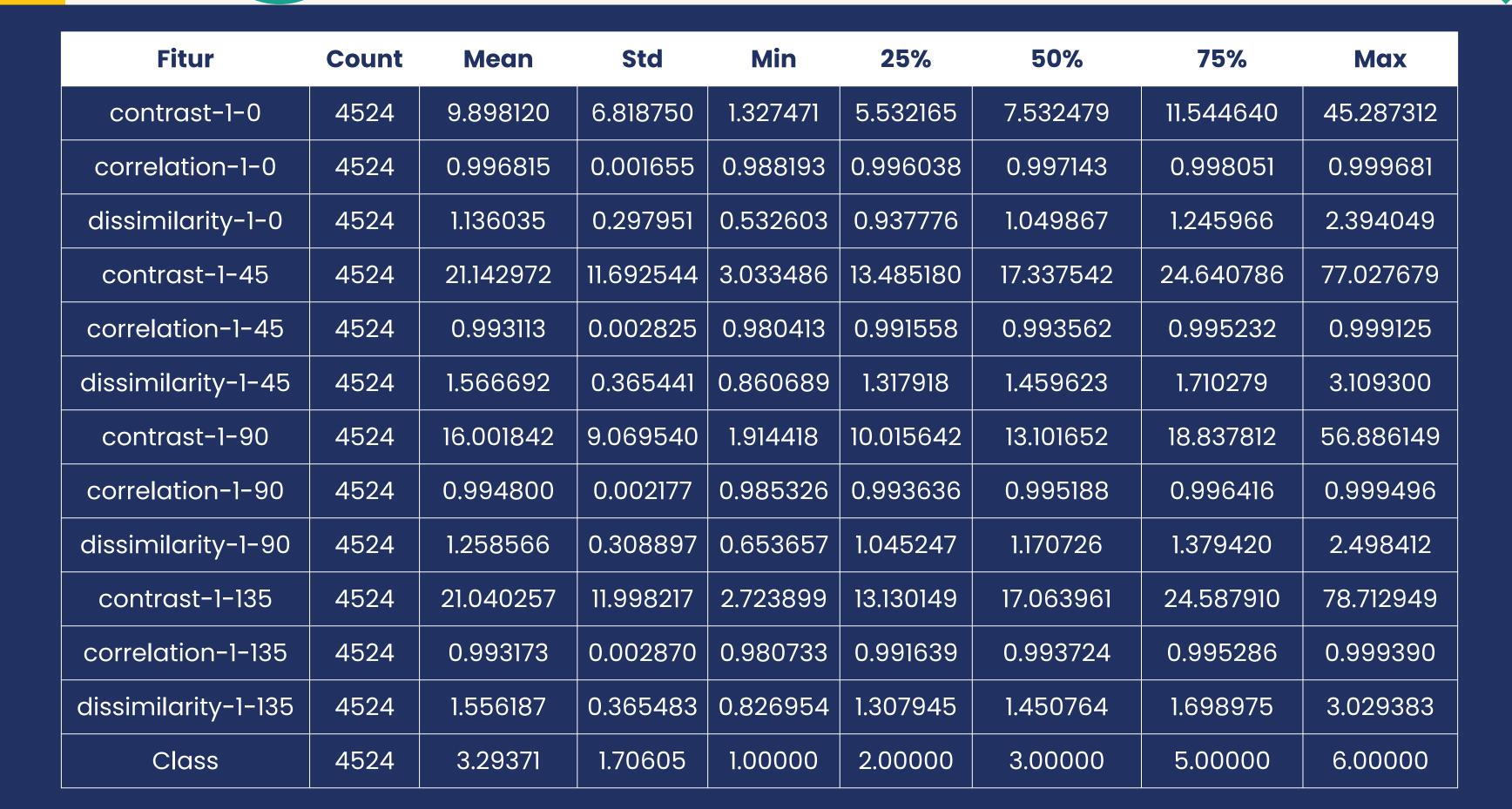
Statistik Deskriptif Keseluruhan Data

- Rentang nilai fitur yang bervariasi menunjukkan bahwa dataset memiliki keberagaman yang cukup tinggi. Ini dapat bermanfaat untuk pemodelan, tetapi juga perlu diwaspadai adanya outlier atau nilai ekstrim yang mungkin mempengaruhi hasil analisis.
- Distribusi yang cenderung simetris pada sebagian besar fitur mengindikasikan bahwa data tersebar secara normal. Namun, beberapa fitur yang memiliki distribusi condong ke kanan perlu diperhatikan karena mungkin memiliki karakteristik yang berbeda dan membutuhkan perlakuan khusus.





Statistik Deskriptif Data Latih



Statistik Deskriptif Data Latih

- Statistik deskriptif yang sama dengan keseluruhan data, tetapi hanya untuk data latih (training set) yang berjumlah 4524 sampel.
- Nilai-nilai statistik deskriptif pada data latih cenderung mirip dengan keseluruhan data, menunjukkan bahwa data latih merepresentasikan karakteristik keseluruhan dataset dengan baik.
- Ini berarti bahwa data latih dapat digunakan dengan cukup representatif untuk melatih model pembelajaran mesin, karena memiliki karakteristik yang serupa dengan keseluruhan dataset.





Statistik Deskriptif Data Uji



Statistik Deskriptif Data Uji

- Statistik deskriptif yang sama dengan keseluruhan data, tetapi hanya untuk data uji (test set) yang berjumlah 1132 sampel.
- Nilai-nilai statistik deskriptif pada data uji juga cenderung mirip dengan keseluruhan data.
- Hal ini menunjukkan bahwa data uji juga merepresentasikan karakteristik keseluruhan dataset dengan baik, sehingga dapat digunakan untuk mengevaluasi kinerja model yang dilatih dengan data latih secara akurat.

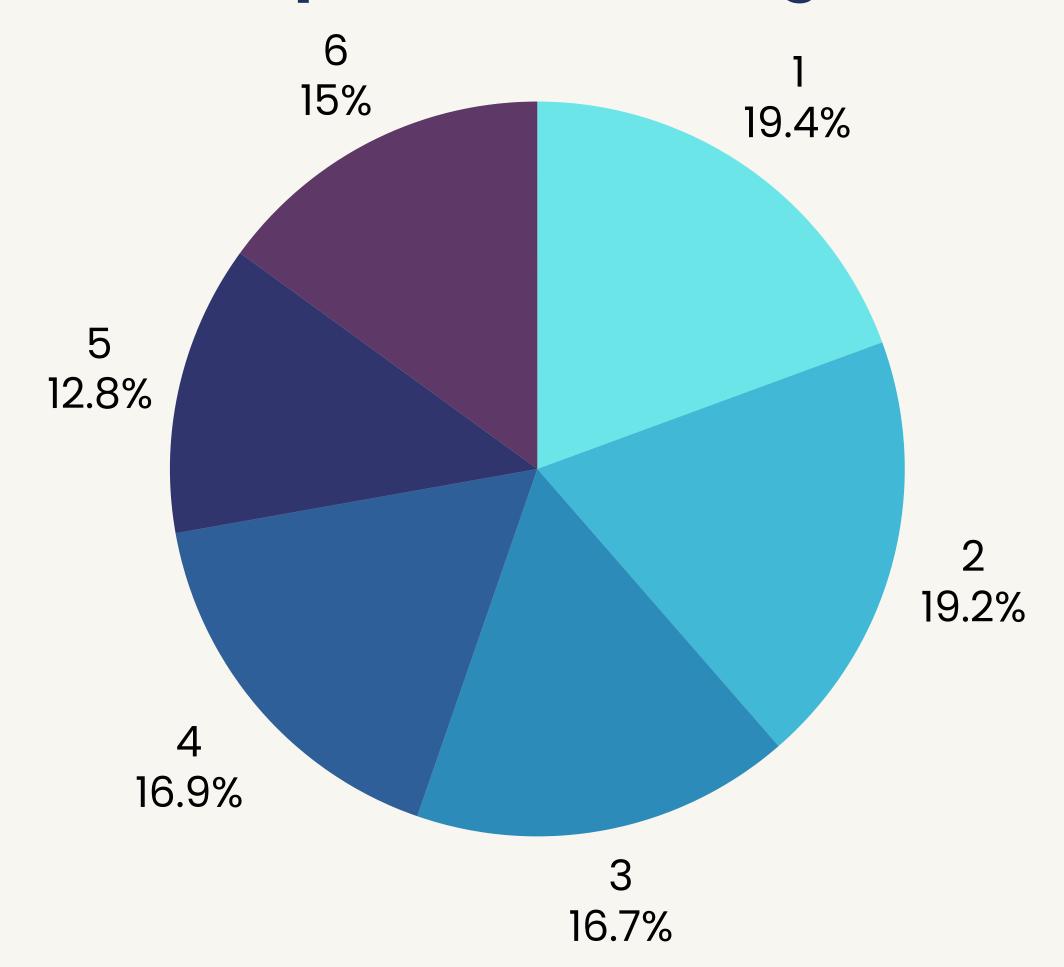




Proporsi Kelas Target

Kelas	Proporsi	
1	0.194130	
2	0.192185	
4	0.169024	
3	0.166902	
6	0.149576	
5	0.128182	

Proporsi Kelas Target

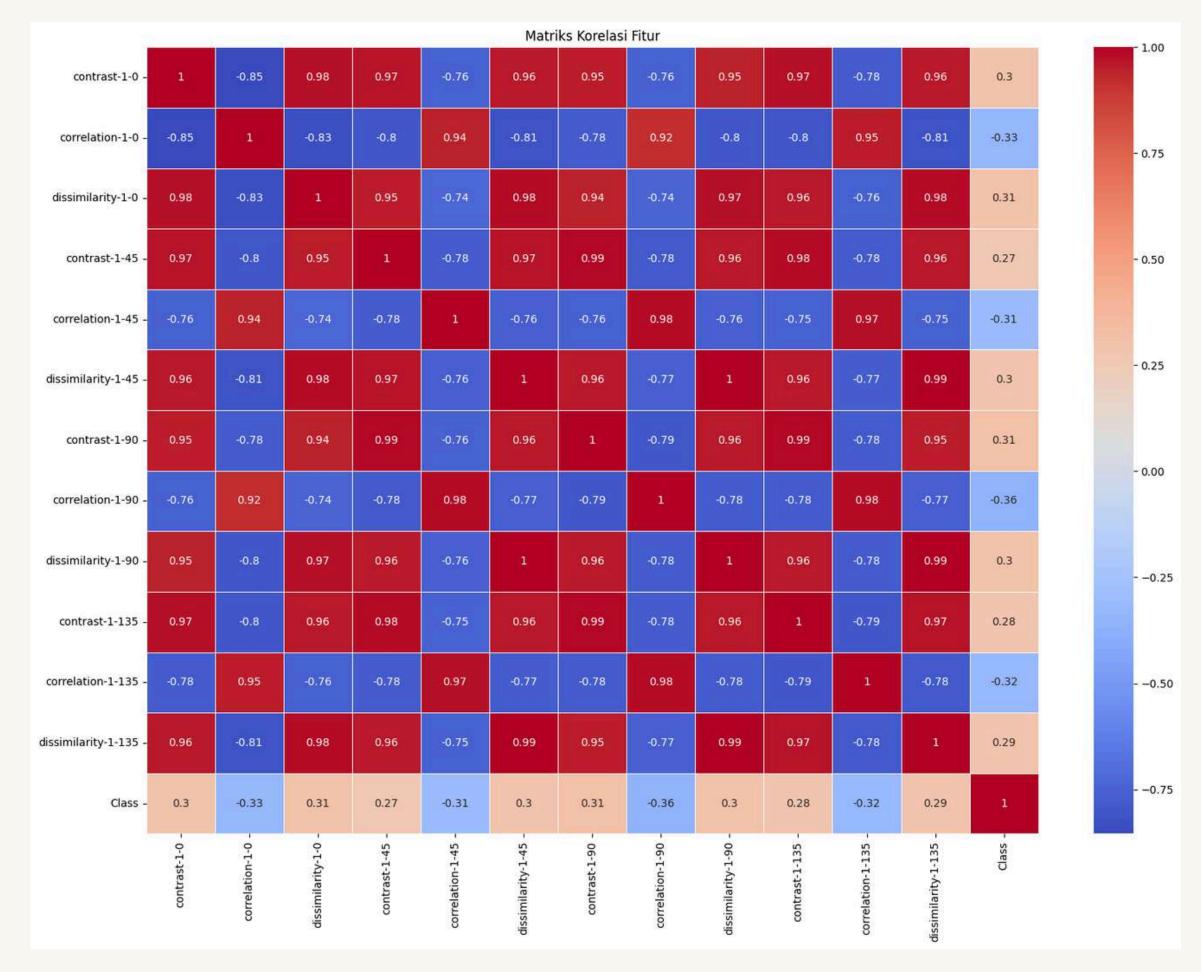




- Distribusi kelas yang cukup merata, dengan kelas 1 dan 2 memiliki proporsi terbesar (sekitar 19%) dan kelas 5 memiliki proporsi terkecil (sekitar 13%).
- Distribusi kelas yang relatif seimbang mengindikasikan bahwa dataset tidak memiliki masalah kelas yang sangat dominan atau sangat jarang. Ini akan memudahkan dalam pemodelan dan evaluasi.



Matriks Korelasi Fitur



- Mayoritas fitur memiliki korelasi yang rendah atau sedang, menunjukkan bahwa fitur-fitur tersebut memberikan informasi yang relatif independen satu sama lain.
- Beberapa pasangan fitur yang memiliki korelasi cukup tinggi, seperti 'contrast-1-0' dengan 'dissimilarity-1-0', 'contrast-1-45' dengan 'dissimilarity-1-45', dan 'contrast-1-90' dengan 'dissimilarity-1-90', perlu diperhatikan karena mungkin ada redundansi informasi.
- Diagonal utama matriks menunjukkan bahwa setiap fitur memiliki korelasi 1 dengan dirinya sendiri, yang merupakan hasil yang wajar

PENANGANAN OUTLIERS





PENANGANAN OUTLIERS

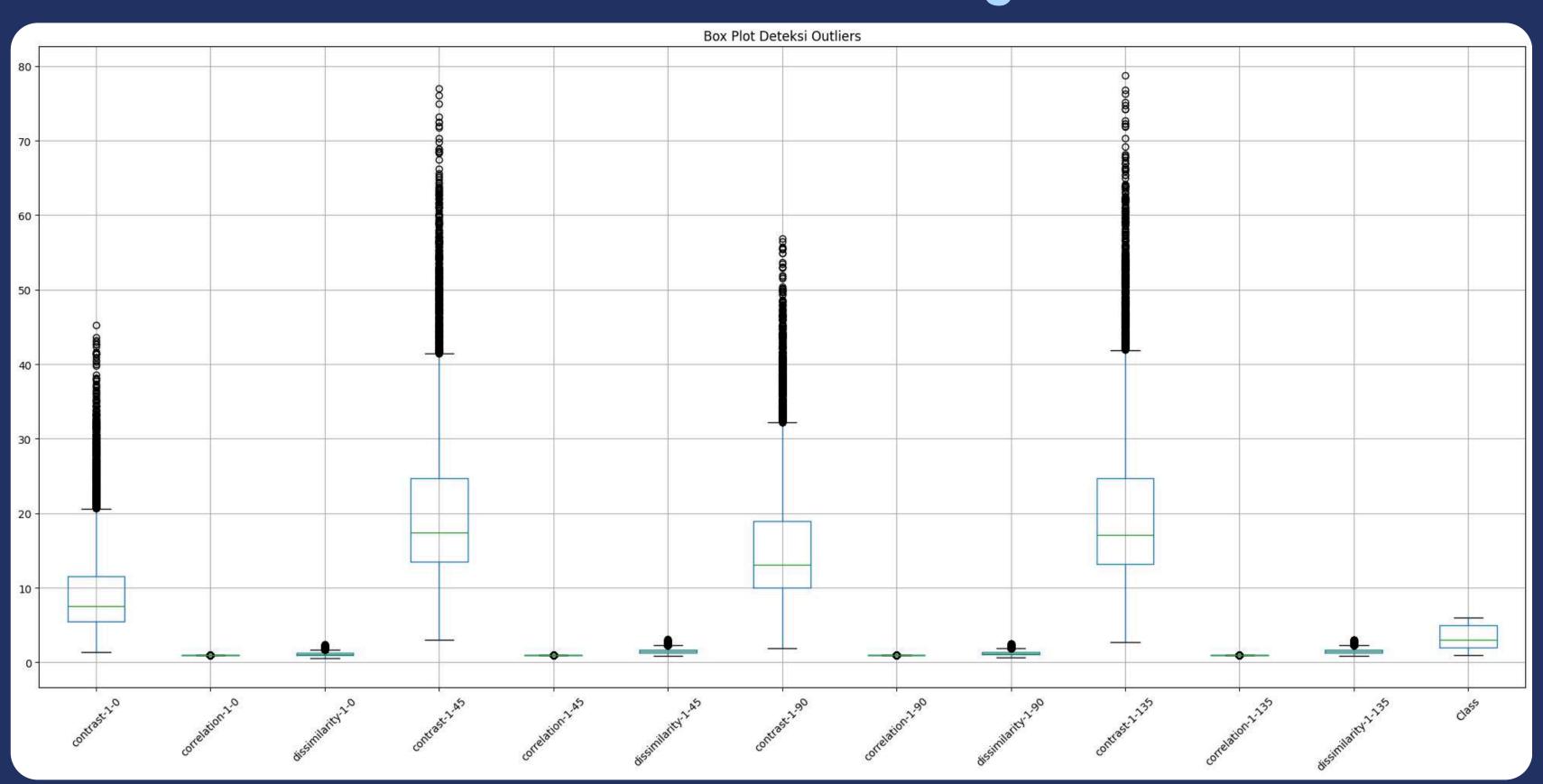


1

Deteksi Outliers Sebelum Dilakukan Handling 2

Handling
Outliers
Menggunakan
IQR

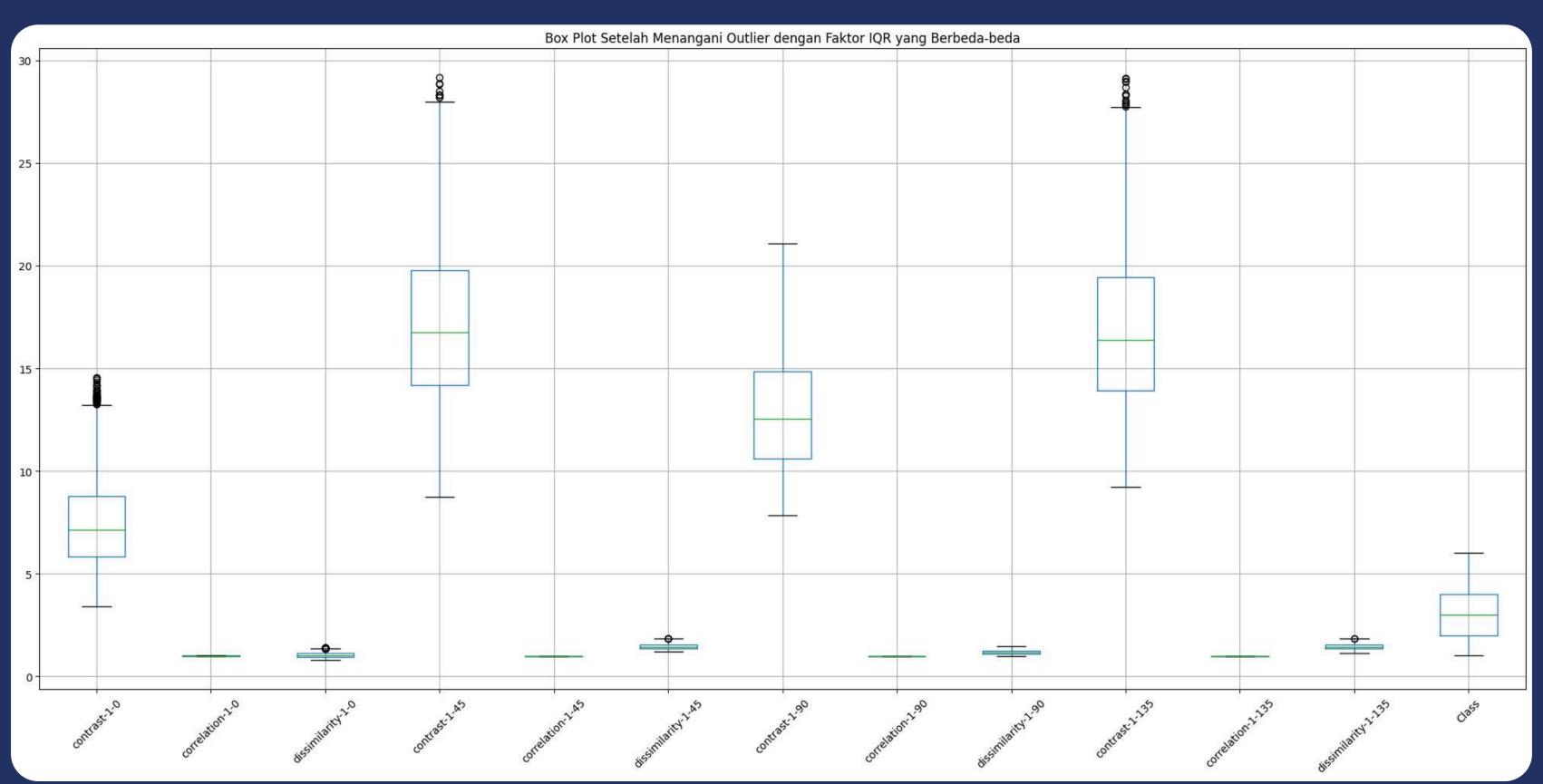
Deteksi Outliers Sebelum Dilakukan Handling



Deteksi Outliers Sebelum Dilakukan Handling

Dari box plot yang ditampilkan, kita dapat dengan jelas melihat bahwa terdapat beberapa fitur yang memiliki outliers yang teridentifikasi. Outliers tersebut ditandai dengan titik-titik yang berada jauh di luar batas atau bawah kotak (box) pada box plot. Beberapa fitur yang mengandung outliers yang cukup ekstrem antara lain adalah contrast-1-135, contrast-1-90, contrast-1-45, dan contrast-1-0.

Keberadaan outliers dalam suatu dataset dapat memberikan pengaruh yang signifikan pada hasil analisis yang akan dilakukan. Outliers dapat menyebabkan bias dalam estimasi parameter, mempengaruhi akurasi prediksi, dan mengurangi kemampuan model untuk menangkap pola-pola yang sebenarnya terdapat dalam data. Oleh karena itu, perlu dilakukan penanganan terhadap outliers sebelum melanjutkan ke tahap analisis selanjutnya.



IQR merupakan rentang nilai yang berada di antara kuartil pertama (Q1) dan kuartil ketiga (Q3). Dalam proses penanganan outliers, IQR digunakan untuk menentukan batas atas dan bawah yang wajar bagi suatu fitur. Nilai-nilai yang berada di luar batas tersebut dianggap sebagai outliers dan dihapus dari dataset.

Pada kasus ini, faktor IQR yang digunakan untuk menentukan batas atas dan bawah berbeda-beda untuk setiap fitur. Berikut penjelasan lebih rinci mengenai penggunaan faktor IQR yang berbeda-beda:

a) Fitur dengan faktor IQR 0.5:

- contrast-1-0
- correlation-1-0
- dissimilarity-1-0
- contrast-1-135
- correlation-1-135
- dissimilarity-1-135

Untuk fitur-fitur ini, batas atas dan bawah ditentukan menggunakan faktor IQR sebesar 0.5. Artinya, nilai yang berada di luar rentang Q1 -0.5IQR dan Q3 + 0.5IQR dianggap sebagai outliers.

b) Fitur dengan faktor IQR 0.75:

- contrast-1-45
- correlation-1-45
- dissimilarity-1-45

Untuk fitur-fitur ini, batas atas dan bawah ditentukan menggunakan faktor IQR sebesar 0.75. Artinya, nilai yang berada di luar rentang Q1 – 0.75IQR dan Q3 + 0.75IQR dianggap sebagai outliers.

c) Fitur dengan faktor IQR 0.25:

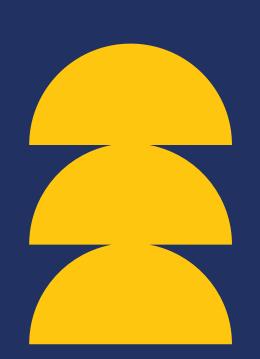
- contrast-1-90
- correlation-1-90
- dissimilarity-1-90

Untuk fitur-fitur ini, batas atas dan bawah ditentukan menggunakan faktor IQR sebesar 0.25. Artinya, nilai yang berada di luar rentang Q1 -0.25IQR dan Q3 + 0.25IQR dianggap sebagai outliers.

Penggunaan faktor IQR yang berbeda-beda ini bertujuan untuk menyesuaikan dengan karakteristik distribusi data pada masing-masing fitur, sehingga penanganan outliers dapat lebih optimal. Fitur-fitur dengan distribusi yang lebih terpusat (misalnya contrast-1-90, correlation-1-90, dissimilarity-1-90) menggunakan faktor IQR yang lebih kecil (0.25), sedangkan fitur-fitur dengan distribusi yang lebih menyebar (misalnya contrast-1-0, correlation-1-0, dissimilarity-1-0) menggunakan faktor IQR yang lebih besar (0.5).

Setelah menentukan batas atas dan bawah berdasarkan IQR, **data yang berada di luar rentang tersebut dihapus dari dataset**. Ini terlihat pada box plot kedua, di mana outliers yang sebelumnya teridentifikasi pada gambar pertama sudah tidak muncul lagi.

Dengan penanganan outliers menggunakan metode IQR yang disesuaikan untuk setiap fitur, kita dapat memperoleh **dataset yang lebih bersih** dan siap untuk dianalisis lebih lanjut.



Sampling dan Uji Distribusi

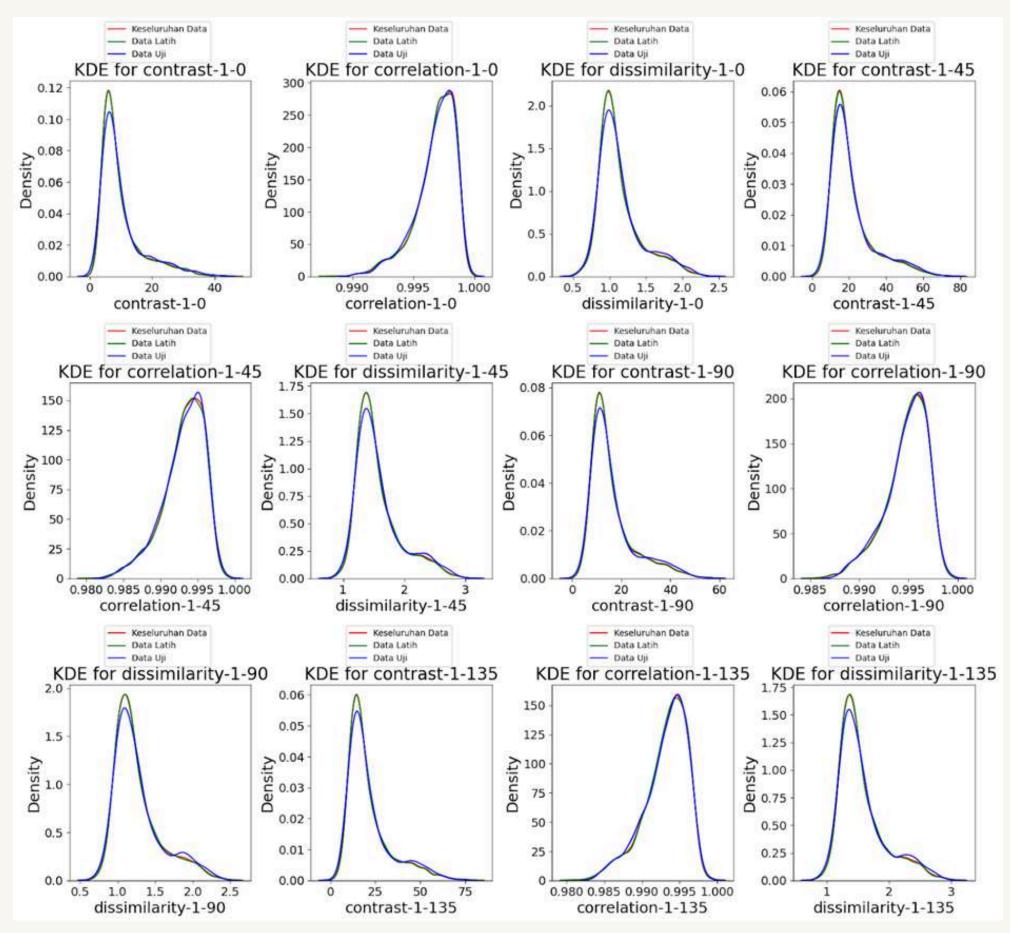
Data Latih	Data Uji	
80%	20%	

- **Sampling**: memisahkan data menjadi dua subset yaitu data latih (training data) dan data uji (testing data).
- Metode visualisasi: Kernel Density Estimate (KDE) untuk setiap fitur numerik.

SAMPLING & PENGUJIAN DISTRIBUSI DATA

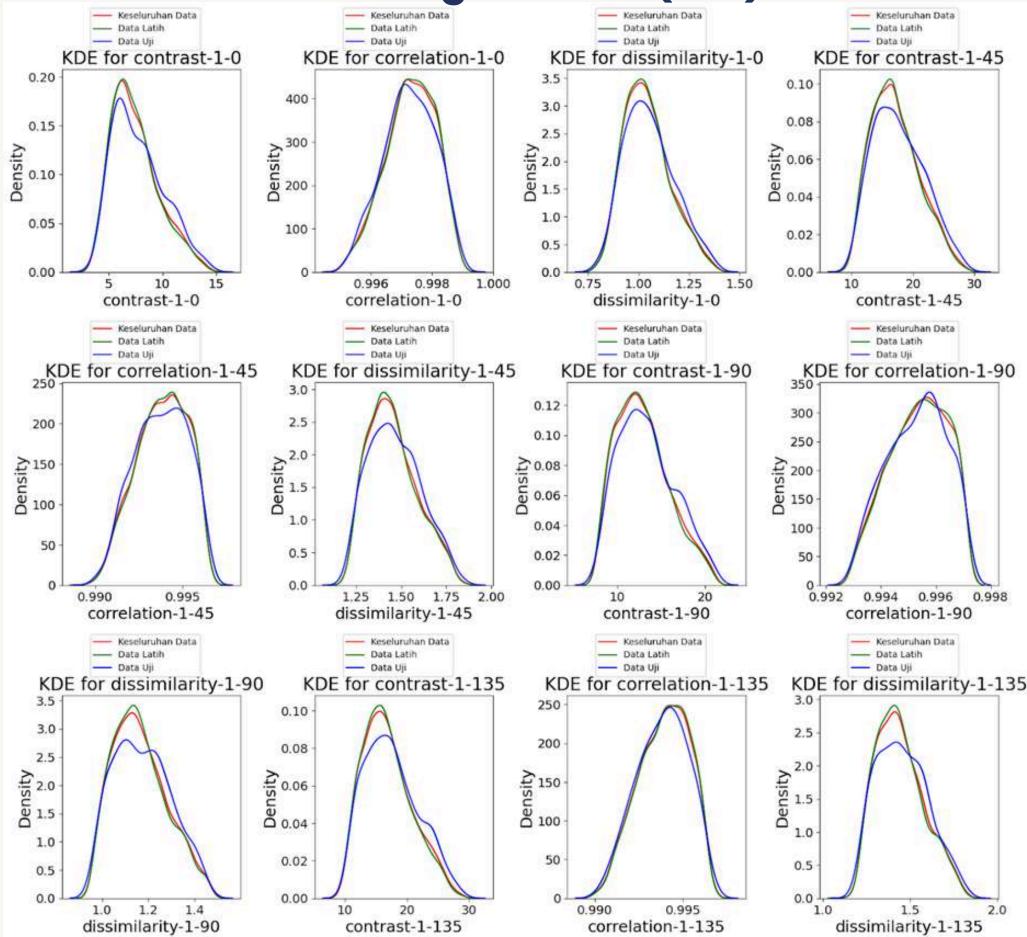


Distribusi Fitur antara Data Latih dengan Keseluruhan Data dengan KDE Plot



- Pada grafik KDE (Kernel Density Estimation) yang ditampilkan, terlihat bahwa distribusi fitur antara Keseluruhan Data, Data Latih, dan Data Uji hampir saling bertumpuk atau berhimpitan.
- Pembagian data latih dan data uji telah dilakukan secara proporsional dan tidak ada bias distribusi yang signifikan antara kedua set data tersebut yang mengindikasikan Distribusi hampir identik.
- Namun, pada beberapa fitur, terdapat outliers yang dapat teridentifikasi dari ekor-ekor distribusi yang menjulur jauh ke arah kanan atau kiri. Outliers ini perlu ditangani agar tidak memberikan pengaruh yang tidak diinginkan pada tahap analisis dan pemodelan selanjutnya.

Distribusi Fitur Sesudah Dilakukan Handling Outliers (IQR)



- Setelah penanganan outliers, distribusi fitur-fitur tersebut terlihat lebih smooth dan compact, dengan ekor-ekor distribusi yang tidak menjulur terlalu jauh. Hal ini menunjukkan bahwa outliers yang sebelumnya teridentifikasi telah berhasil dihilangkan dari dataset.
- Distribusi fitur antara Keseluruhan Data, Data Latih, dan Data Uji masih mempertahankan karakteristik yang hampir serupa setelah penanganan outliers yang mengindikasikan bahwa pembagian data telah dilakukan dengan baik dan tidak ada bias distribusi yang signifikan antara kedua set data tersebut.



Proporsi Target antara Data Latih dengan Data Keseluruhan



a. Distribusi Umum:

- Terdapat 6 kelas dengan distribusi yang relatif merata
- Range proporsi berkisar antara 13-22% untuk semua kelas
- Tidak ada kelas yang sangat dominan atau sangat minoritas

b. Perbandingan antar Kelas:

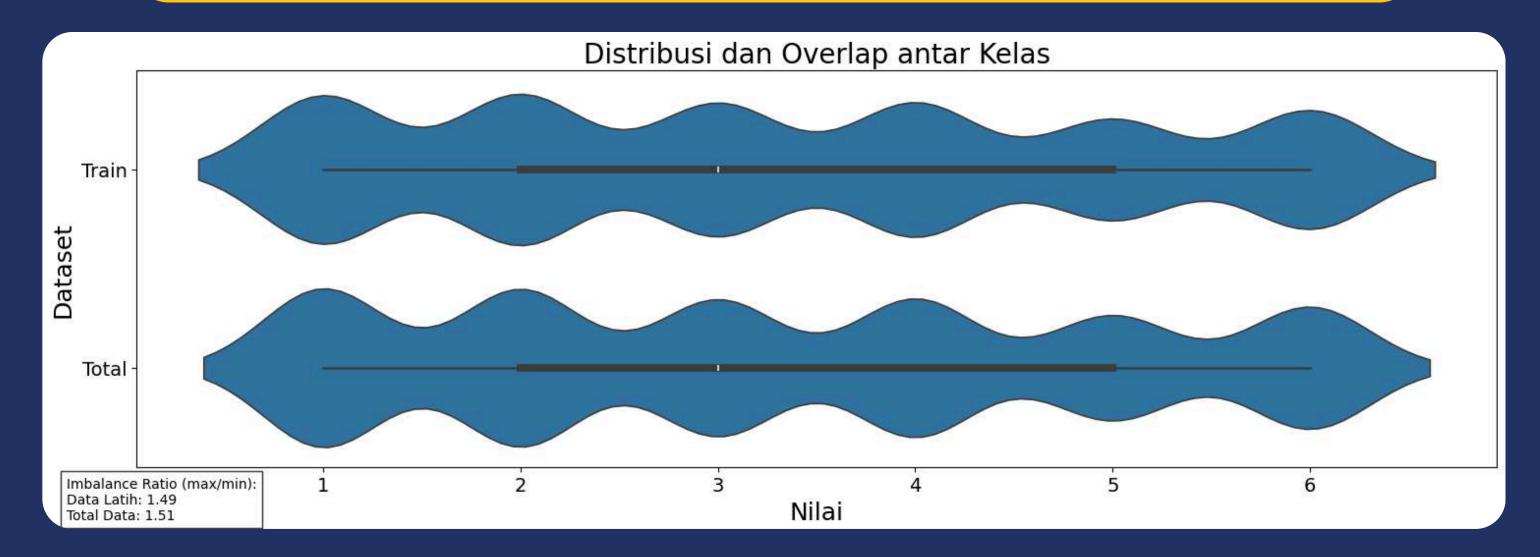
- Kelas 2 memiliki proporsi tertinggi (±22%)
- Kelas 6 memiliki proporsi terendah (±13%)
- Kelas lainnya berada di kisaran 14-19%

Perbandingan Data Latih vs Keseluruhan Data:

- Proporsi antara data latih dan keseluruhan data **sangat mirip untuk semua kelas**
- Perbedaan terbesar hanya sekitar 0.1-0.4%
- Ini menunjukkan stratifikasi sampling yang sangat baik



Proporsi Target antara Data Latih dengan Data Keseluruhan



a. Karakteristik Distribusi:

- Terlihat overlap yang signifikan antar kelas
- Bentuk distribusi antara data latih dan total data sangat mirip
- Distribusi cenderung multimodal (memiliki beberapa puncak)

b. Densitas dan Spread:

- Setiap kelas memiliki spread (sebaran) yang cukup lebar
- Densitas tertinggi terlihat di beberapa titik untuk setiap kelas
- Overlap yang tinggi menunjukkan potensi tantangan dalam klasifikasi

a. Nilai Imbalance:

- Data Latih: 1.65
- Total Data: 1.71
- Kedua nilai sangat dekat, menunjukkan konsistensi distribusi

b. Interpretasi:

- Imbalance ratio < 3 menunjukkan dataset cukup seimbang
- Tidak diperlukan teknik resampling khusus
- Perbedaan kecil antara data latih dan total (0.06) menunjukkan splitting yang baik



Uji Kesamaan Distribusi antara Data Latih dengan Keseluruhan Data menggunakan Kolmogorov-Smirnovng

Pada percobaan ini, dilakukan uji kesamaan distribusi antara data latih dan keseluruhan data menggunakan metode Kolmogorov-Smirnov (KS-Test). Uji ini **bertujuan untuk menentukan apakah distribusi data latih** serupa dengan keseluruhan data berdasarkan masing-masing fitur. Metode ini membandingkan fungsi distribusi kumulatif (CDF) dari dua dataset dan menghasilkan dua metrik utama, yaitu KS Statistic dan pvalue. KS Statistic menunjukkan perbedaan maksimum antara dua distribusi, sementara p-value mengindikasikan apakah perbedaan tersebut signifikan atau tidak. Pada percobaan ini, tingkat signifikansi yang digunakan adalah 5% (α = 0.05). Jika p-value lebih besar dari tingkat signifikansi, maka distribusi antara data latih dan keseluruhan data dianggap tidak memiliki perbedaan yang signifikan.

Hasil Distribusi Data Latih dan Keseluruhan Data menggunakan Kolmogorov-Smirnov

No.	Feature	KS Statistic	P-Value	Sama
1.	contrast-1-0	0.016875	0.836364	True
2.	correlation-1-0	0.009749	0.999468	True
3.	dissimilarity-1-0	0.013536	0.964739	True
4.	contrast-1-45	0.016967	0.831514	True
5.	correlation-1-45	0.008673	0.999948	True
6.	dissimilarity-1-45	0.016799	0.840230	True
7.	contrast-1-90	0.015941	0.882015	True
8.	correlation-1-90	0.008156	0.999988	True
9.	dissimilarity-1-90	0.017209	0.818705	True
10.	contrast-1-135	0.016945	0.832810	True
11.	correlation-1-135	0.008956	0.999896	True
12.	dissimilarity-1-135	0.016211	0.869409	True

Hasil uji pada seluruh fitur menunjukkan bahwa nilai p-value untuk semua fitur lebih besar dari 0.05, yang mengindikasikan bahwa tidak terdapat perbedaan distribusi yang signifikan antara data latih dan keseluruhan data. Selain itu, nilai KS Statistic untuk semua fitur juga relatif kecil, yang mendukung hasil bahwa perbedaan distribusi antara data latih dan keseluruhan data dapat diabaikan. Oleh karena itu, distribusi data latih dapat dianggap mewakili distribusi keseluruhan data, dan data latih ini layak digunakan sebagai subset untuk membangun model.

UJI NORMALITAS DATA DENGAN KOLMOGOROV-SMIRNOV





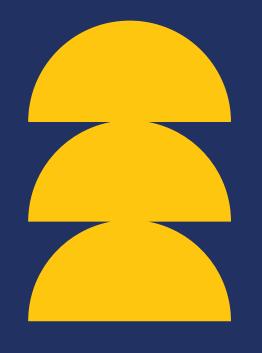
UJI NORMALITAS DATA DENGAN KOLMOGOROV-SMIRNOV



Uji normalitas Kolmogorov-Smirnov digunakan untuk menguji apakah distribusi data sampel berbeda secara signifikan dari distribusi normal. Dalam uji ini, data dibandingkan dengan distribusi normal teoretis menggunakan perbedaan maksimum kumulatif antara keduanya. Jika nilai skewness dari uji Kolmogorov Smirnov kurang dari tingkat signifikansi (0,05), maka hipotesis bahwa data berdistribusi normal ditolak.

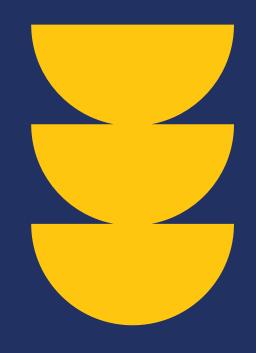






z	KS P-Value	Skewness	Normal
contrast-1-0	8.737027e-135	1.858047	TIDAK NORMAL
correlation-1-0	4.926734e-37	1.323059	TIDAK NORMAL
dissimilarity-1-0	7.946769e-87	1.323059	TIDAK NORMAL
contrast-1-45	2.788924e-105	1.623713	TIDAK NORMAL
correlation-1-45	3.854355e-20	-0.915316	TIDAK NORMAL
dissimilarity-1-45	2.165575e-78	1.305511	TIDAK NORMAL
contrast-1-90	2.168479e-106	1.597828	TIDAK NORMAL
correlation-1-90	1.571479e-22	-0.924536	TIDAK NORMAL
dissimilarity-1-90	2.131698e-75	1.279560	TIDAK NORMAL
contrast-1-135	2.547741e-109	1.651513	TIDAK NORMAL
correlation-1-135	1.797307e-24	-0.960483	TIDAK NORMAL
dissimilarity-1-135	2.586457e-77	1.298978	TIDAK NORMAL





Uji Normalitas Data

Tanpa Outlier

Fitur	KS P-Value	Skewness	Normal
contrast-1-0	6.521356e-11	0.706650	TIDAK NORMAL
correlation-1-0	1.599958e-03	-0.342511	NORMAL
dissimilarity-1-0	3.320856e-05	0.514750	TIDAK NORMAL
contrast-1-45	5.475939e-06	0.483419	NORMAL
correlation-1-45	1.263096e-04	-0.295321	NORMAL
dissimilarity-1-45	1.123786e-06	0.508232	TIDAK NORMAL
contrast-1-90	6.012455e-06	0.524750	TIDAK NORMAL
correlation-1-90	5.369402e-06	-0.270044	NORMAL
dissimilarity-1-90	1.794553e-06	0.486903	NORMAL
contrast-1-135	2.402572e-07	0.544470	TIDAK NORMAL
correlation-1-135	4.194735e-05	-0.323484	NORMAL
dissimilarity-1-135	5.639294e-06	0.500536	TIDAK NORMAL



METODE TRANSFORMASI DATA

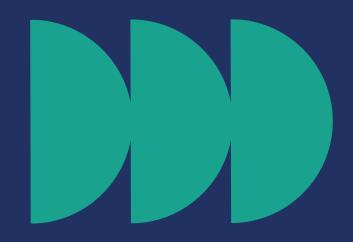


LOG TRANSFORMATION

SQUARE ROOT TRANSFORMATION

BOX-COX TRANSFORMATION

YEO-JOHNSON TRANSFORMATION



Berhasil menormalkan semua fitur

LOG TRANSFORMATION

Transformasi Log diterapkan untuk mengurangi skewness dengan mengambil logaritma dari nilai-nilai data

Fitur	KS P-Value	Skewness	Normal
contrast-1-0	0.030708	0.149660	NORMAL
correlation-1-0	0.001577	-0.343382	NORMAL
dissimilarity-1-0	0.001907	0.382994	NORMAL
contrast-1-45	0.136684	0.043778	NORMAL
correlation-1-45	0.000120	-0.296646	NORMAL
dissimilarity-1-45	0.000045	0.399173	NORMAL
contrast-1-90	0.010355	0.108069	NORMAL
correlation-1-90	0.000005	-0.270898	NORMAL
dissimilarity-1-90	0.000245	0.382732	NORMAL
contrast-1-135	0.214361	0.089350	NORMAL
correlation-1-135	0.000041	-0.324815	NORMAL
dissimilarity-1-135	0.000480	0.388436	NORMAL

SQUARE ROOT TRANSFORMATION

Transformasi Square Root mengambil akar kuadrat dari nilai data yang umumnya digunakan untuk data yang sudah cukup mendekati normal namun tetap memiliki skewness ringan

Fitur	KS P-Value	Skewness	Normal
contrast-1-0	0.000028	0.390277	NORMAL
correlation-1-0	0.001577	-0.343384	NORMAL
dissimilarity-1-0	0.001832	0.386554	NORMAL
contrast-1-45	0.020445	0.250215	NORMAL
correlation-1-45	0.000120	-0.296650	NORMAL
dissimilarity-1-45	0.000027	0.416673	NORMAL
contrast-1-90	0.004825	0.298329	NORMAL
correlation-1-90	0.000005	-0.270900	NORMAL
dissimilarity-1-90	0.000177	0.390749	NORMAL
contrast-1-135	0.003299	0.302900	NORMAL
correlation-1-135	0.000041	-0.324819	NORMAL
dissimilarity-1-135	0.000269	0.406139	NORMAL

ARC SIN TRANSFORMATION

Transformasi Arc Sin utamanya digunakan untuk data proporsi atau data yang memiliki batas antara 0 dan 1, seperti proporsi atau persentase

Fitur	KS P-Value	Skewness	Normal
contrast-1-0	-	_	TIDAK NORMAL
correlation-1-0	0.060565	-0.021328	NORMAL
dissimilarity-1-0	-	_	TIDAK NORMAL
contrast-1-45	-	_	TIDAK NORMAL
correlation-1-45	0.001316	-0.088229	NORMAL
dissimilarity-1-45	-	_	TIDAK NORMAL
contrast-1-90	_	_	TIDAK NORMAL
correlation-1-90	0.000120	-0.090688	NORMAL
dissimilarity-1-90	-	_	TIDAK NORMAL
contrast-1-135	-	_	TIDAK NORMAL
correlation-1-135	0.003620	-0.110845	NORMAL
dissimilarity-1- 135	_	_	TIDAK NORMAL

BOX-COX TRANSFORMATION

Transformasi Box Cox dirancang untuk normalisasi data yang bersifat positif dengan menggunakan parameter lambda yang dioptimalkan untuk menormalkan distribusi

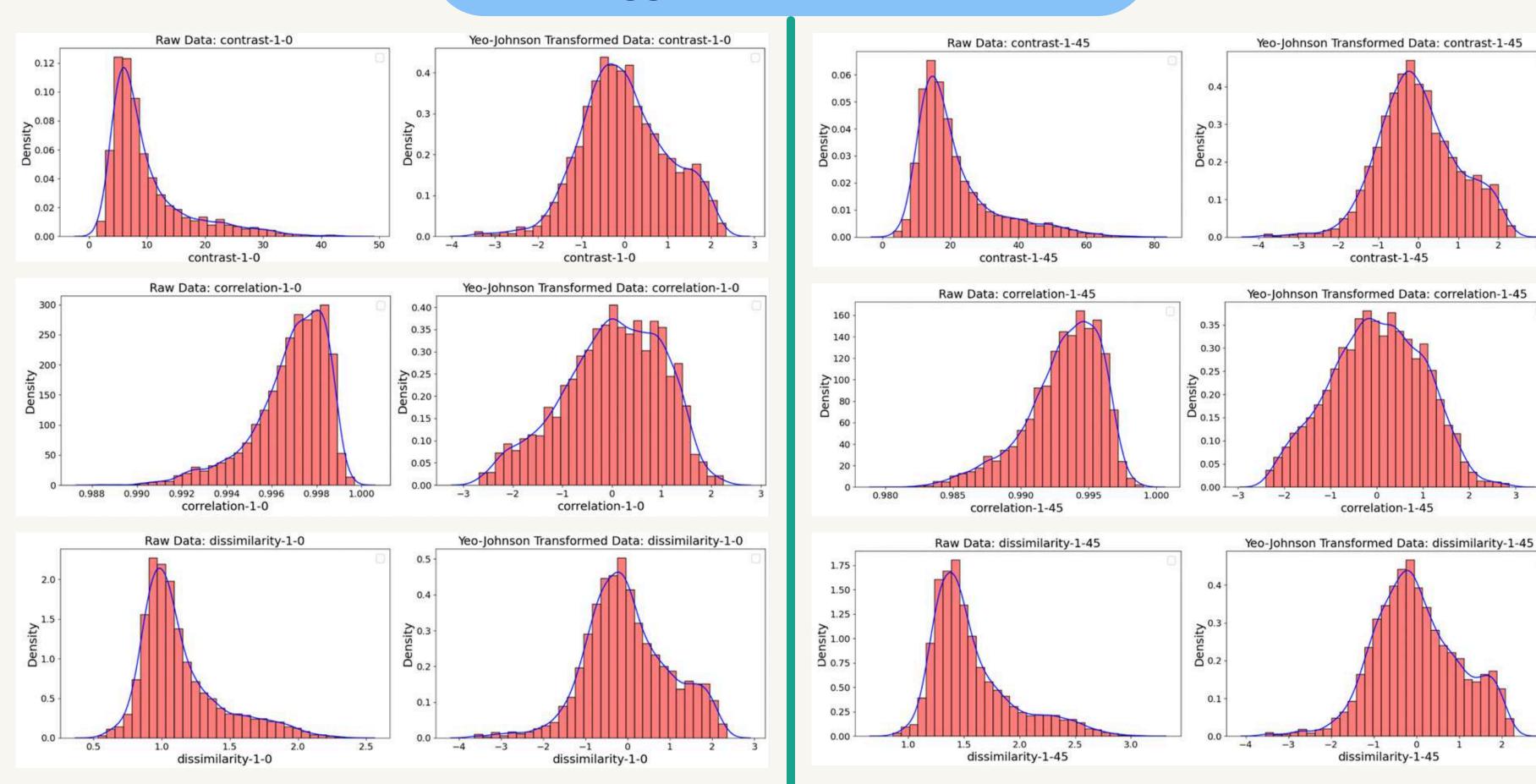
Fitur	KS P-Value	Skewness	Normal
contrast-1-0	0.449371	0.004604	NORMAL
correlation-1-0	0.029791	-0.032218	NORMAL
dissimilarity-1-0	0.390466	0.014691	NORMAL
contrast-1-45	0.181637	0.001507	NORMAL
correlation-1-45	0.000768	-0.041371	NORMAL
dissimilarity-1-45	0.057394	0.035523	NORMAL
contrast-1-90	0.009561	0.008352	NORMAL
correlation-1-90	0.000131	-0.045695	NORMAL
dissimilarity-1-90	0.070634	0.034029	NORMAL
contrast-1-135	0.247360	0.005324	NORMAL
correlation-1-135	0.004368	-0.042885	NORMAL
dissimilarity-1-135	0.105209	0.031704	NORMAL

YEO-JOHNSON TRANSFORMATION

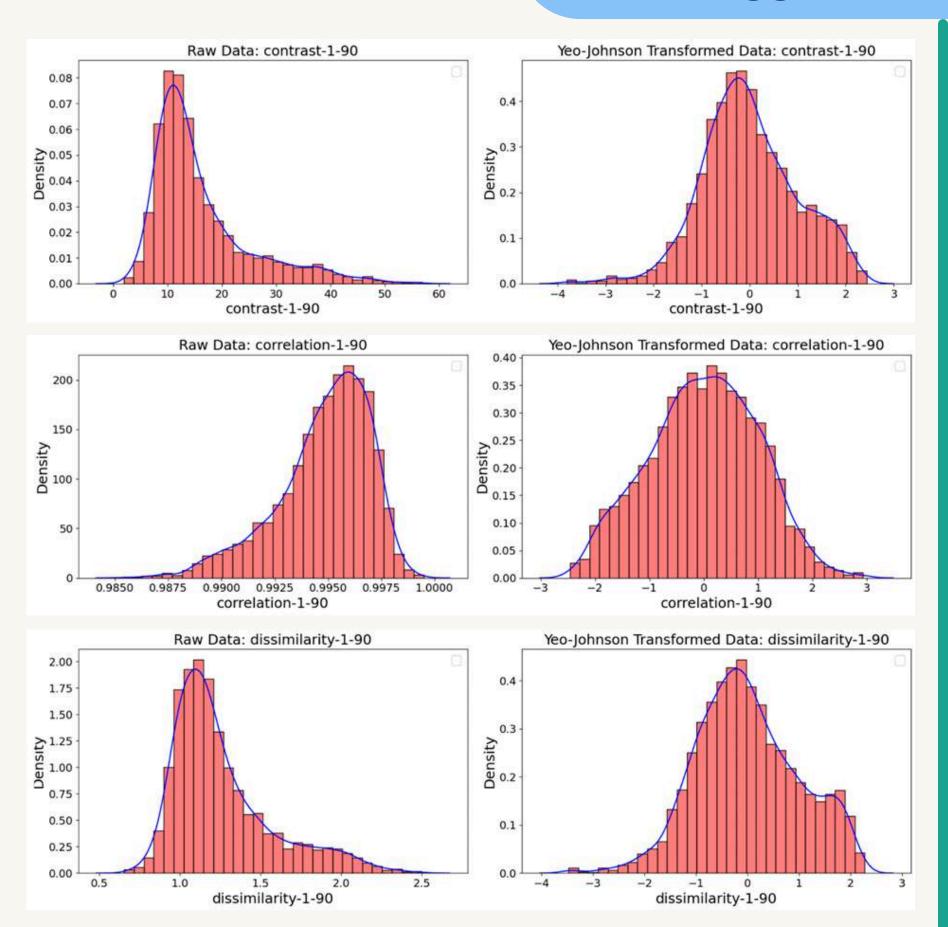
Transformasi Yeo-Johnson mirip seperti Box-Cox namun bisa digunakan untuk data yang memiliki nilai negatif dan positif

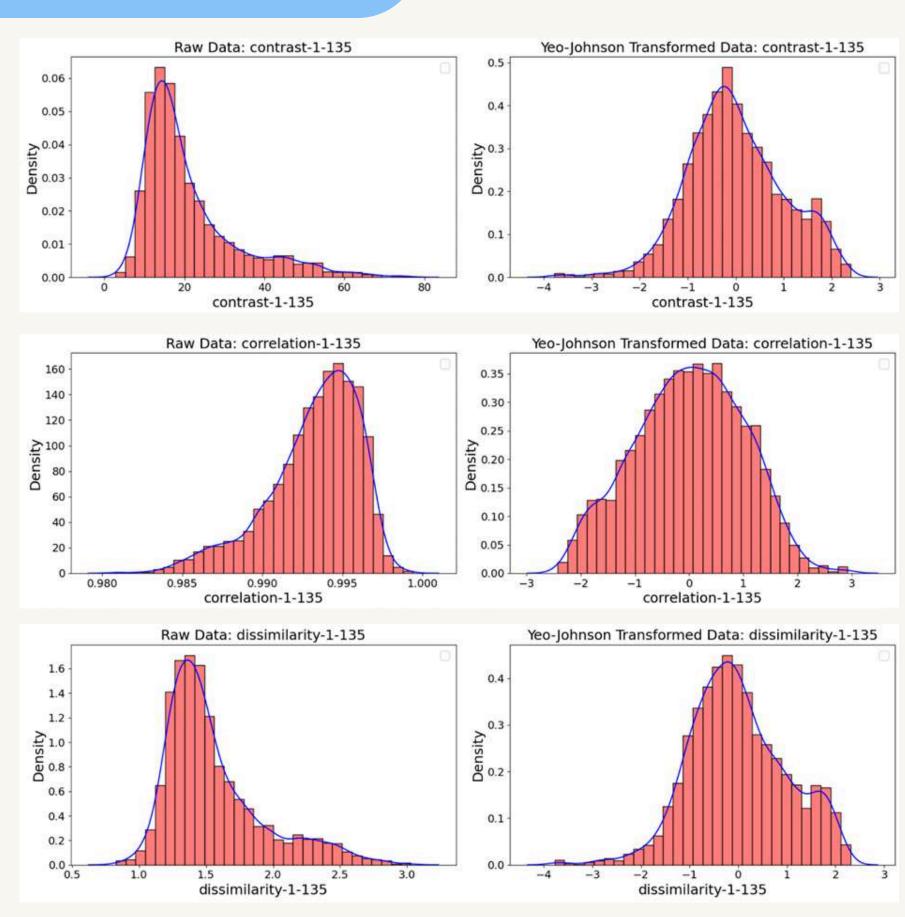
Fitur	KS P-Value	Skewness	Normal
contrast-1-0	0.398614	0.010475	NORMAL
correlation-1-0	0.007780	-0.247888	NORMAL
dissimilarity-1-0	0.322874	0.023505	NORMAL
contrast-1-45	0.171501	0.003917	NORMAL
correlation-1-45	0.000649	-0.152703	NORMAL
dissimilarity-1-45	0.047431	0.044439	NORMAL
contrast-1-90	0.008905	0.012023	NORMAL
correlation-1-90	0.000017	-0.177363	NORMAL
dissimilarity-1-90	0.060699	0.044813	NORMAL
contrast-1-135	0.238453	0.007851	NORMAL
correlation-1-135	0.000765	-0.179974	NORMAL
dissimilarity-1-135	0.094152	0.040430	NORMAL

Perbandingan Hasil Transformasi menggunakan Yeo-Johnson



Perbandingan Hasil Transformasi menggunakan Yeo-Johnson





METODE KLASIFIKASI DATA





Random Forest

Random Forest adalah metode ensemble learning yang menggabungkan multiple decision trees untuk menghasilkan prediksi yang lebih akurat dan stabil.

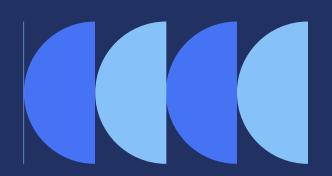
Tahapan Eksperimen dan Implementasi Model

Preprocessing Data:

- Penanganan outliers dengan IQR
- Normalisasi data menggunakan transformasi Yeo-Johnson untuk menormalkan semua fitur.

Pemilihan Model

- Menggunakan Random Forest.
- Pipeline Terbaik: Random Forest dengan FastICA untuk pengurangan dimensi sebelum klasifikasi.



Detail Konfigurasi Model

- Random Forest Classifier:
 - o Bootstrap: Diaktifkan untuk variasi data.
 - Criterion: Menggunakan entropy untuk pemisahan node.
 - Max Features: 65% dari total fitur.
 - Min Samples Leaf: 1 untuk leaf node yang detail.
 - Min Samples Split: 10 untuk mencegah overfitting.
 - N Estimators: 100 tree untuk ensemble learning.
- FastICA: Tolerance diatur ke 0.35 untuk efisiensi dan akurasi.

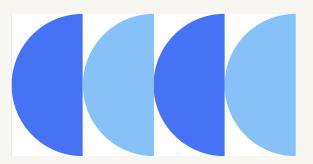
Tahapan Eksperimen dan Implementasi Model

Optimalisasi Model

- Grid Parameter yang Divariasikan:
 - o n_estimators: [50, 100, 150, 200].
 - min_samples_leaf: [1, 2, 4, 6].
- Parameter Tetap:
 - Bootstrap: Diaktifkan.
 - Criterion: Entropy.
 - Max Features: 65%.
 - Min Samples Split: 10.
 - FastICA Tolerance: 0.35.

Evaluasi Model

 Menggunakan metrik akurasi dengan cross-validation 5 kali percobaan untuk memastikan hasil yang robust.





Hasil Percobaan Random Forest



Parameter Random Forest			Akurasi Percobaan				
n_estmator	min_samples_leaf	acc_1	acc_2	acc_3	acc_4	acc_5	rata-rata
50	1	51.33%	51.94%	53.27%	51.24%	52.65%	52.08%
50	2	50.18%	53.36%	53.45%	49.12%	53.36%	51.89%
50	4	53.27%	51.59%	50.71%	51.06%	52.47%	51.82%
50	6	50.88%	52.74%	50.18%	51.33%	52.65%	51.54%
100	1	52.92%	51.33%	52.21%	51.33%	52.21%	52.00%
100	2	53.45%	51.68%	51.50%	51.86%	51.94%	52.08%
100	4	51.24%	51.50%	50.00%	52.03%	51.59%	51.27%
100	6	54.59%	51.77%	50.27%	51.33%	52.39%	52.07%



Hasil Percobaan Random Forest



Parameter Random Forest		Akurasi Percobaan					
n_estmator	min_samples_leaf	acc_1	acc_2	acc_3	acc_4	acc_5	rata-rata
150	1	52.21%	52.47%	51.59%	51.50%	49.29%	51.41%
150	2	54.06%	51.41%	53.09%	51.68%	50.97%	52.24%
150	4	54.42%	50.00%	50.71%	50.53%	51.24%	51.57%
150	6	49.73%	51.50%	50.88%	52.83%	51.59%	51.31%
200	1	51.06%	52.12%	53.80%	54.68%	51.50%	52.63%
200	2	51.41%	53.62%	49.56%	53.18%	53.00%	52.16%
200	4	53.00%	51.33%	52.21%	53.36%	52.47%	52.47%
200	6	50.07%	52.30%	53.18%	50.27%	52.03%	51.75%

UJI HIPOTESIS





Uji Hipotesis

Uji hipotesis menggunakan **ANOVA** untuk parameter 'n_estimators' dan 'min_samples_leaf'



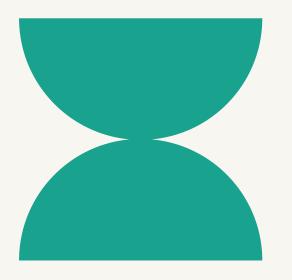


Tidak ada pengaruh signifikan dari perubahan nilai parameter terhadap akurasi





Ada pengaruh signifikan dari perubahan nilai parameter terhadap akurasi



Uji Hipotesis

ANOVA					
Parameter	F-statistic	p-value			
n_estimators	2.0376796989403028	0.1623347961084501			
min_sample_leaf	1.0188740985065572	0.4185077152369911			

Karena, nilai p-value lebih besar dari 0.05 (tingkat kepercayaan 95%), maka kita dapat menerima H0 dan menyatakan bahwa tidak ada pengaruh signifikan dalam akurasi berdasarkan faktor parameter tersebut.



TERIMA KASIH

