

**LAPORAN PROJECT  
STATISTIKA INFERENSI**

**Teknik Informatika - Kelas B**

**Dosen Pengampu Mata Kuliah:**  
Ibu Candra Dewi, S.Kom, M.Sc.



**Disusun Oleh:**

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

**PROGRAM STUDI TEKNIK INFORMATIKA  
FAKULTAS ILMU KOMPUTER  
UNIVERSITAS BRAWIJAYA  
MALANG  
2024**

## 1. EXPLORATORY DATA ANALYSIS (EDA)

### 1.1 Informasi Dasar Dataset

Jumlah Total Sampel	Jumlah Fitur
5656	13

Tabel 1. Informasi Dasar Dataset

- **Total Sampel:**  
5656 - Jumlah total sampel/observasi dalam dataset.
- **Jumlah Fitur:**  
13 - Dataset memiliki 13 fitur/variabel.

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

Tabel 2. Tipe Data pada Kolom

- **Tipe Data Setiap Kolom:**

Semua kolom adalah tipe data float64 kecuali kolom 'Class' yang bertipe int64. Ini menunjukkan bahwa mayoritas fitur adalah numerik kontinu, sedangkan 'Class' adalah fitur kategorikal.

## 1.2 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

Tabel 3. *Missing Values* Dataset

- **Missing Values:**

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.

### 1.3 Statistik Deskriptif Keseluruhan Data

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
<b>contrast-1-0</b>	5656	9.908253	6.81616	1.327471	5.53696	7.539260	11.56637	45.287312
<b>correlation-1-0</b>	5656	0.996817	0.00165	0.988193	0.99603	0.997146	0.998053	0.999681
<b>dissimilarity-1-0</b>	5656	1.136209	0.29877	0.532603	0.93845	1.050275	1.246821	2.394049
<b>contrast-1-45</b>	5656	21.18282	11.6684	3.033486	13.5119	17.35848	24.69827	77.027679
<b>correlation-1-45</b>	5656	0.993114	0.00282	0.980413	0.99156	0.993566	0.995232	0.999125
<b>dissimilarity-1-4</b>	5656	1.567504	0.36640	0.860689	1.31833	1.460138	1.711261	3.109300

<b>contrast-1-90</b>	5656	16.03759	9.07187	1.914418	10.0367	13.12290	18.89884	56.886149
<b>correlation-1-90</b>	5656	0.994803	0.00217	0.985326	0.99364	0.995193	0.996417	0.999496
<b>dissimilarity-1-9</b>	5656	1.259024	0.31013	0.653657	1.04571	1.171206	1.380240	2.498412
<b>contrast-1-135</b>	5656	21.07635	11.9943	2.723899	13.1577	17.09292	24.66428	78.712949
<b>correlation-1-13</b>	5656	0.993175	0.00286	0.980733	0.99164	0.993727	0.995287	0.999390
<b>dissimilarity-1-1</b>	5656	1.556999	0.36646	0.826954	1.30853	1.451341	1.699963	3.029383
<b>Class</b>	5656	3.29367	1.70614	1.00000	2.00000	3.00000	5.00000	6.00000

Tabel 4. Statistik Deskriptif Keseluruhan Data

- Nilai rata-rata, standar deviasi, nilai minimum, kuartil 1, median, kuartil 3, dan nilai maksimum untuk setiap fitur ditampilkan.
- 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.

#### 1.4 Statistik Deskriptif Data Latih

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
<b>contrast-1-0</b>	4524	9.898120	6.818750	1.327471	5.53216	7.53247	11.5446	45.287312
<b>correlation-1-0</b>	4524	0.996815	0.001655	0.988193	0.99603	0.99714	0.99805	0.999681
<b>dissimilarity-1-</b>	4524	1.136035	0.297951	0.532603	0.93777	1.04986	1.24596	2.394049
<b>contrast-1-45</b>	4524	21.142972	11.692544	3.033486	13.4851	17.3375	24.6407	77.027679
<b>correlation-1-4</b>	4524	0.993113	0.002825	0.980413	0.99155	0.99356	0.99523	0.999125
<b>dissimilarity-1-</b>	4524	1.566692	0.365441	0.860689	1.31791	1.45962	1.71027	3.109300
<b>contrast-1-90</b>	4524	16.001842	9.069540	1.914418	10.0156	13.1016	18.8378	56.886149

<b>correlation-1-9</b>	4524	0.994800	0.002177	0.985326	0.99363	0.99518	0.99641	0.999496
<b>dissimilarity-1-</b>	4524	1.258566	0.308897	0.653657	1.04524	1.17072	1.37942	2.498412
<b>contrast-1-135</b>	4524	21.040257	11.998217	2.723899	13.1301	17.0639	24.5879	78.712949
<b>correlation-1-1</b>	4524	0.993173	0.002870	0.980733	0.99163	0.99372	0.99528	0.999390
<b>dissimilarity-1-</b>	4524	1.556187	0.365483	0.826954	1.30794	1.45076	1.69897	3.029383
<b>Class</b>	4524	3.29371	1.70605	1.00000	2.00000	3.00000	5.00000	6.00000

Tabel 5. 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.

### 1.5 Statistik Deskriptif Data Uji

Fitur	Count	Mean	Std	Min	25%	50%	75%	Max
<b>contrast-1-0</b>	1132	9.948907	6.809551	1.410485	5.56553	7.59571	11.697	40.659629
<b>correlation-1-0</b>	1132	0.996828	0.001650	0.988779	0.99605	0.99715	0.9980	0.999561
<b>dissimilarity-1-</b>	1132	1.137787	0.302186	0.543764	0.94105	1.05215	1.2495	2.394049
<b>contrast-1-45</b>	1132	21.342778	11.590219	3.124537	13.6439	17.4972	25.007	74.633857
<b>correlation-1-4</b>	1132	0.993118	0.002797	0.981163	0.99161	0.99357	0.9952	0.998623
<b>dissimilarity-1-</b>	1132	1.570704	0.370152	0.875079	1.32013	1.46386	1.7167	3.101510

<b>contrast-1-90</b>	1132	16.173368	9.082742	2.158764	10.1827	13.2472	19.140	50.980927
<b>correlation-1-9</b>	1132	0.994818	0.002159	0.986124	0.99367	0.99522	0.9964	0.999084
<b>dissimilarity-1-</b>	1132	1.260947	0.314048	0.683701	1.04773	1.17258	1.3821	2.498412
<b>contrast-1-135</b>	1132	21.211689	11.975548	3.144687	13.2679	17.2831	24.881	75.438695
<b>correlation-1-1</b>	1132	0.993185	0.002853	0.982558	0.99167	0.99374	0.9952	0.998999
<b>dissimilarity-1-</b>	1132	1.561366	0.370269	0.863925	1.31247	1.45492	1.7063	2.978805
<b>Class</b>	1132	3.29363	1.70648	1.00000	2.00000	3.00000	5.0000	6.00000

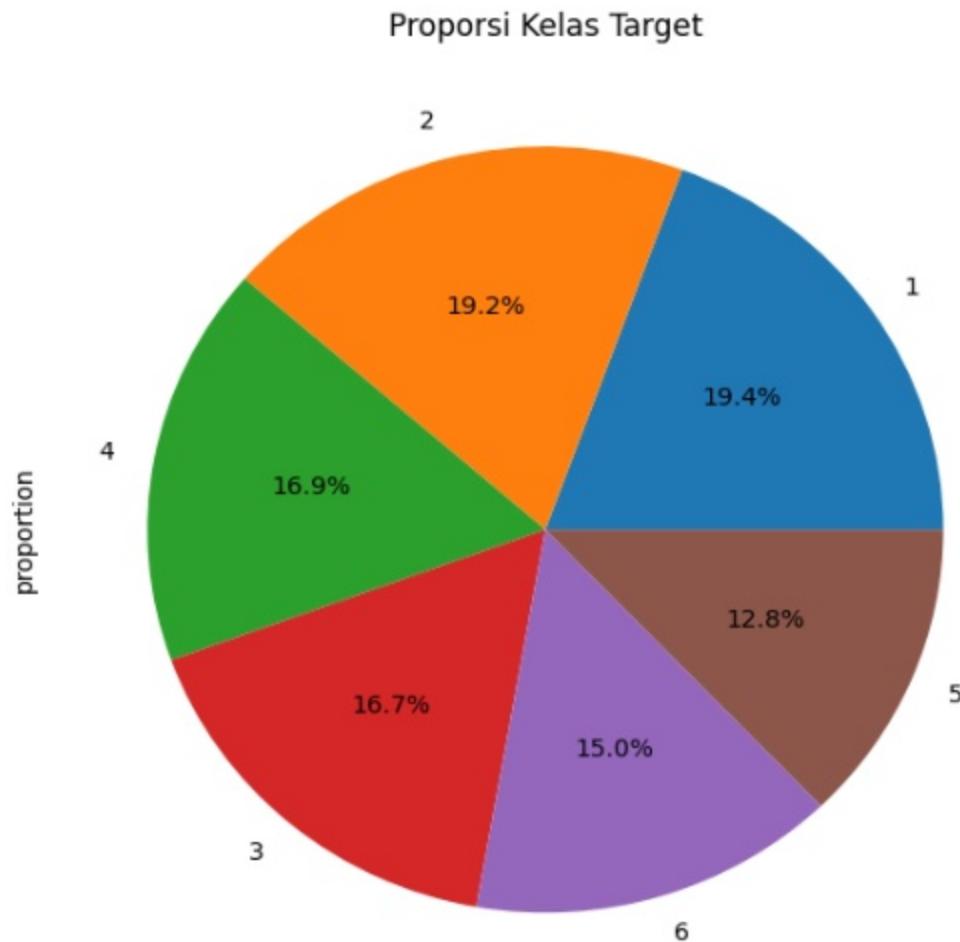
Tabel 6. Statistik Deskriptif Data Uji

- 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.

### 1.6 Proporsi Kelas Target

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

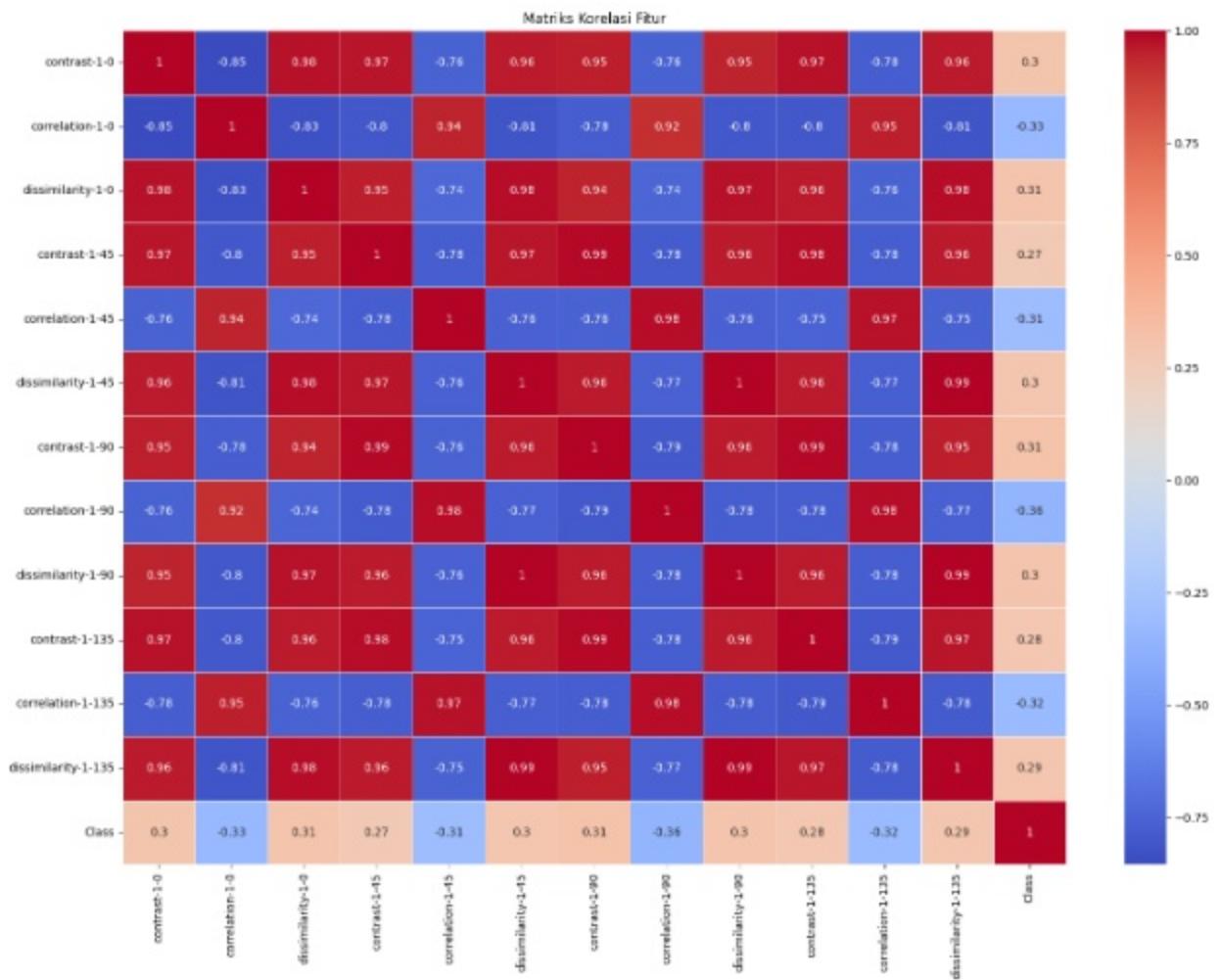
Tabel 7. Proporsi Kelas Target



Gambar 1. Diagram Proporsi Kelas Target

- Proporsi setiap kelas target ditampilkan dalam bentuk pie chart.
- 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.

## 1.7 Matriks Korelasi Fitur

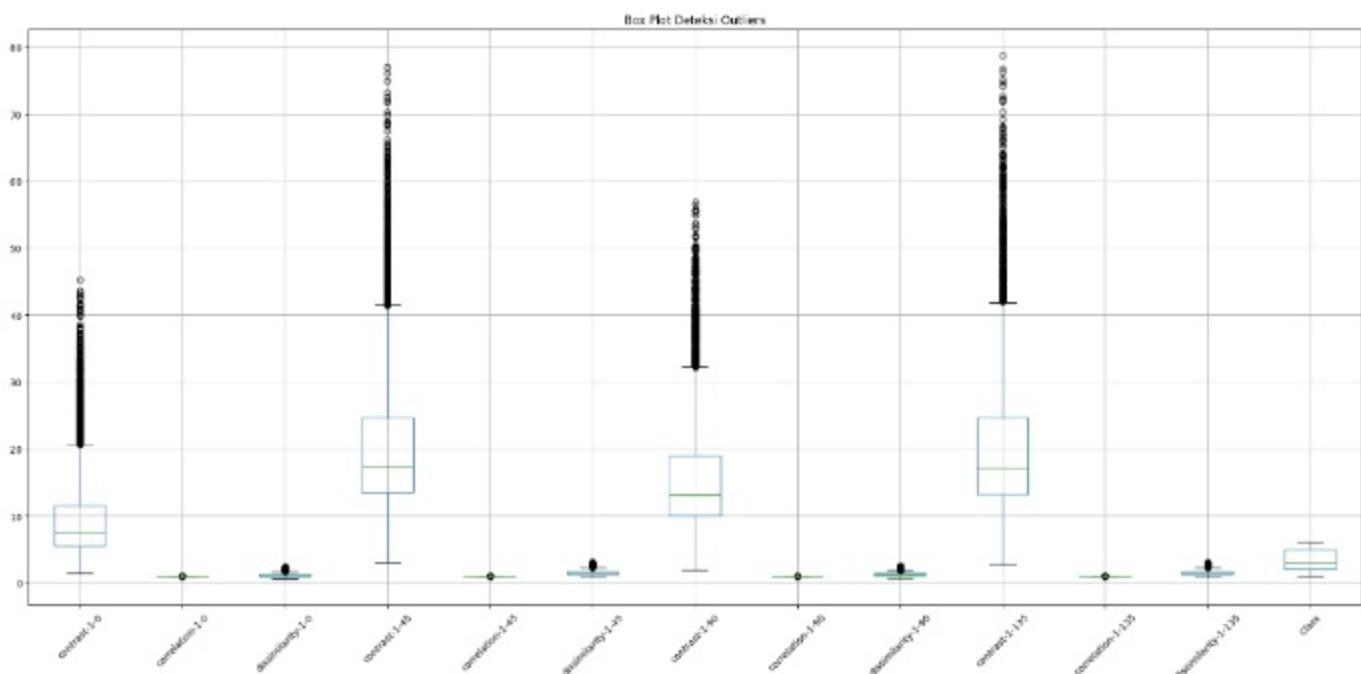


Gambar 2. Matriks Korelasi Fitur

- Matriks korelasi fitur ditampilkan dalam bentuk heatmap.
- 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.

## 2. PENANGANAN OUTLIERS

### 2.1. Deteksi Outliers Sebelum Dilakukan Handling



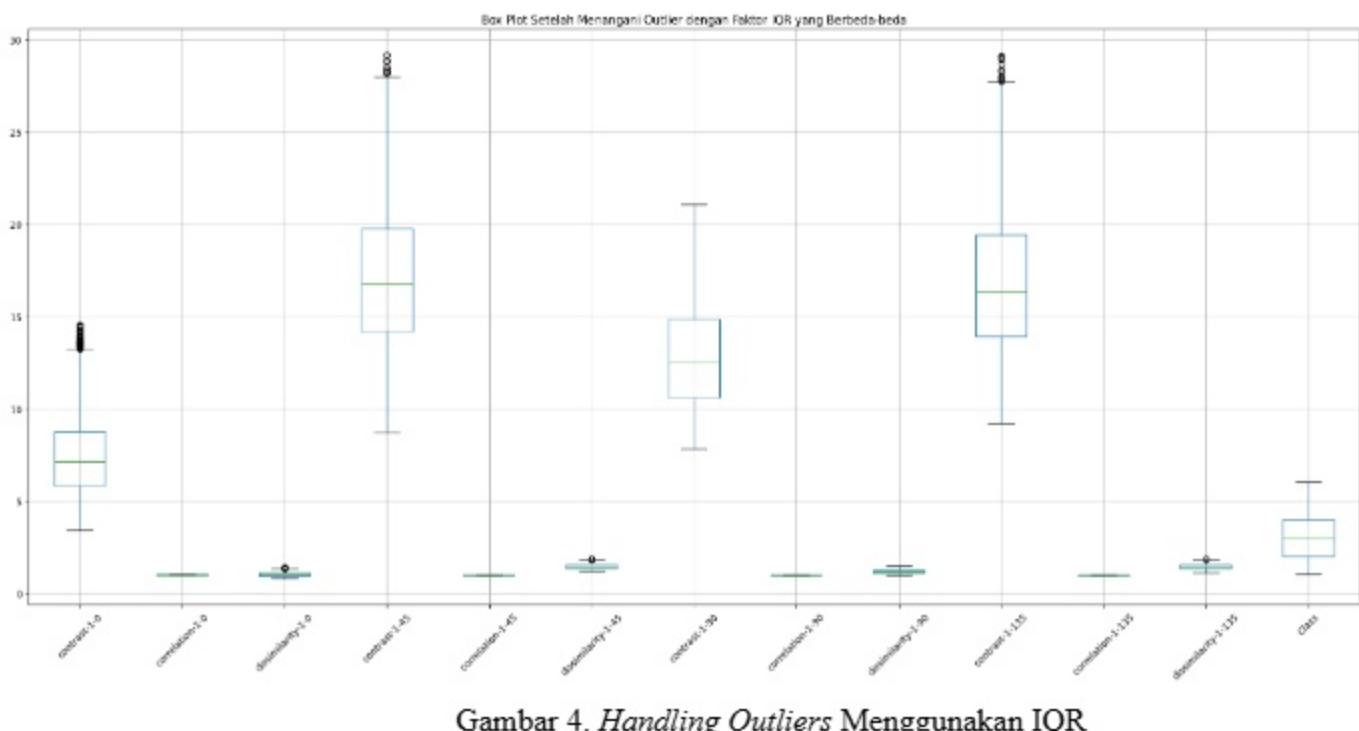
Gambar 3. *Outliers* sebelum dilakukan *Handling*

Dalam visualisasi ini, kita dapat melihat box plot yang menggambarkan distribusi data untuk berbagai fitur yang ada. Box plot merupakan salah satu alat visualisasi yang efektif untuk mengidentifikasi adanya outliers dalam suatu dataset. Outliers adalah nilai-nilai ekstrem yang berada jauh dari distribusi normal data.

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 atas atau bawah kotak (box) pada box plot. Beberapa fitur yang mengandung outliers yang cukup ekstrem antara lain adalah contrast-1-135, dissimilarity-1-90, 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.

## 2.2. Handling Outliers Menggunakan IQR



Gambar 4. *Handling Outliers Menggunakan IQR*

Setelah mengidentifikasi adanya outliers pada gambar pertama, langkah selanjutnya adalah melakukan penanganan terhadap outliers tersebut. Pada gambar kedua, kita dapat melihat box plot yang telah disesuaikan dengan penanganan outliers menggunakan metode Interquartile Range (IQR).

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

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. Hal ini penting untuk memastikan kualitas data dan keakuratan hasil analisis yang akan dilakukan pada tahap selanjutnya.

### 3. SAMPLING DAN PENGUJIAN DISTRIBUSI DATA

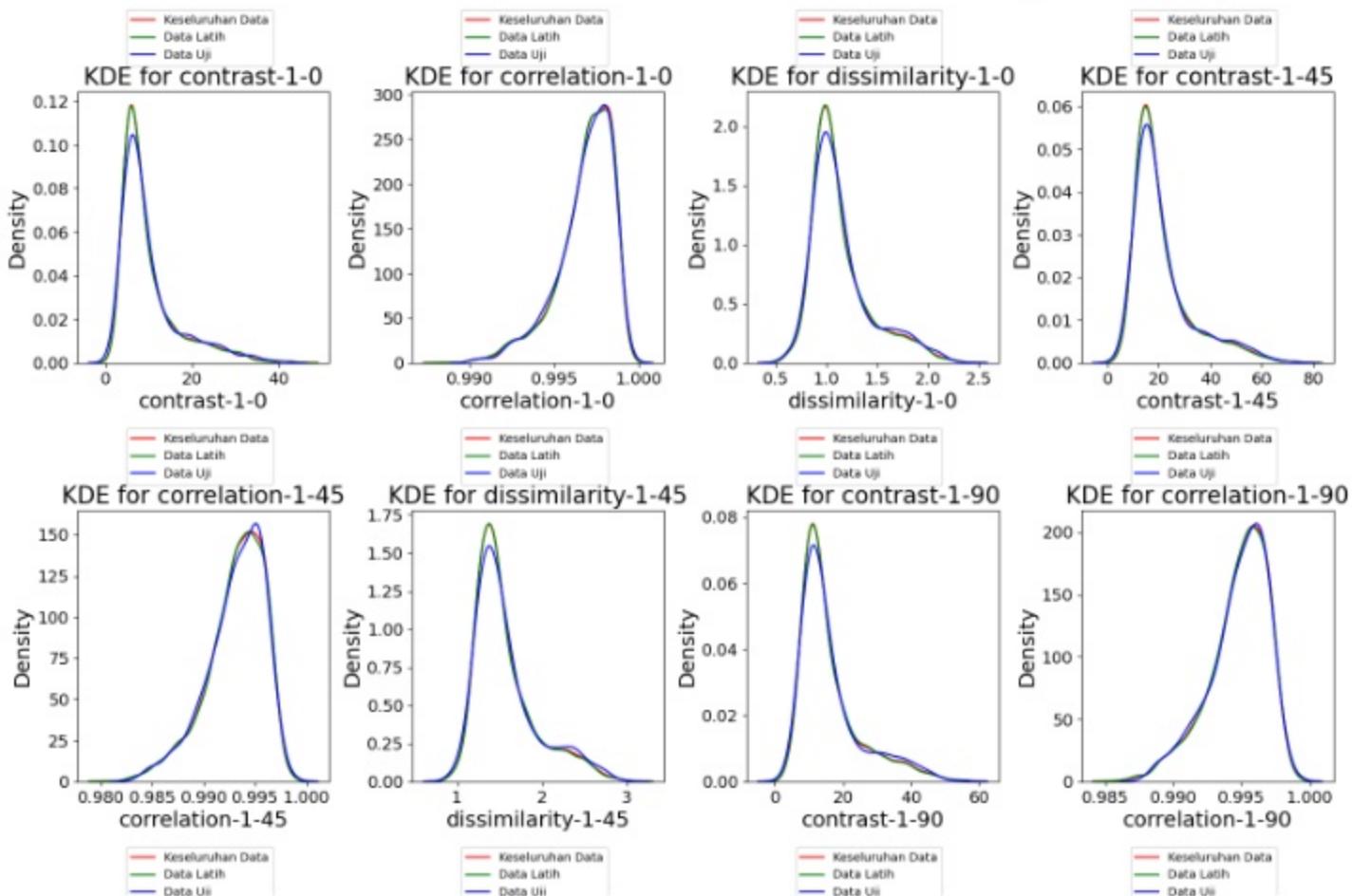
Sampling untuk memisahkan data menjadi dua subset: data latih (training data) dan data uji (testing data). Proses sampling dilakukan dengan menggunakan

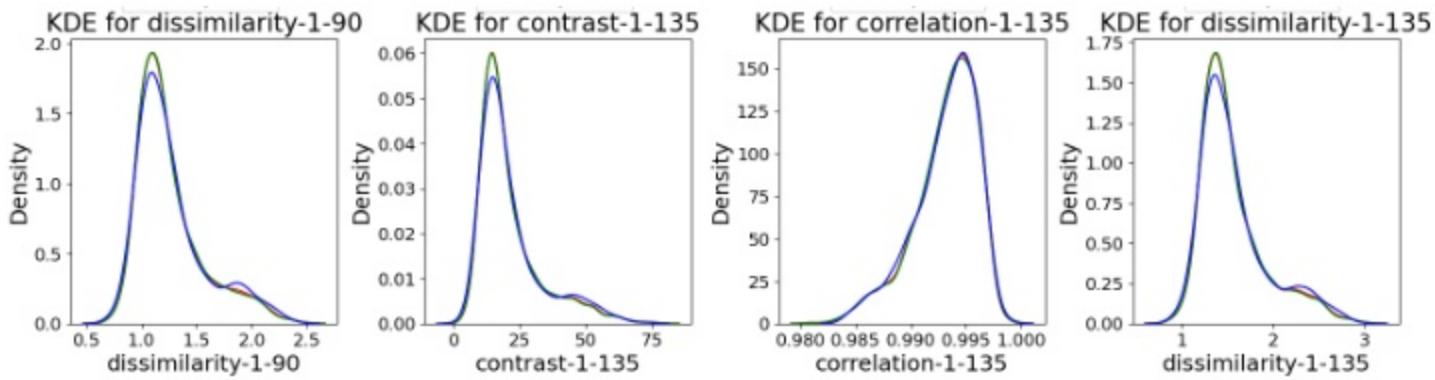
perbandingan 80:20, dimana 80% data digunakan sebagai data latih dan 20% sebagai data uji.

Metode yang digunakan adalah visualisasi Kernel Density Estimate (KDE) untuk setiap fitur numerik. Visualisasi ini bertujuan untuk memeriksa apakah data latih dan data uji memiliki pola distribusi yang serupa sehingga hasil model yang dilatih pada data latih dapat diharapkan berlaku baik pada data uji.

### 3.1. Distribusi Fitur antara Data Latih dengan Keseluruhan Data dengan KDE Plot

#### 3.1.1. Distribusi Fitur Sebelum Dilakukan Handling Outliers (IQR)





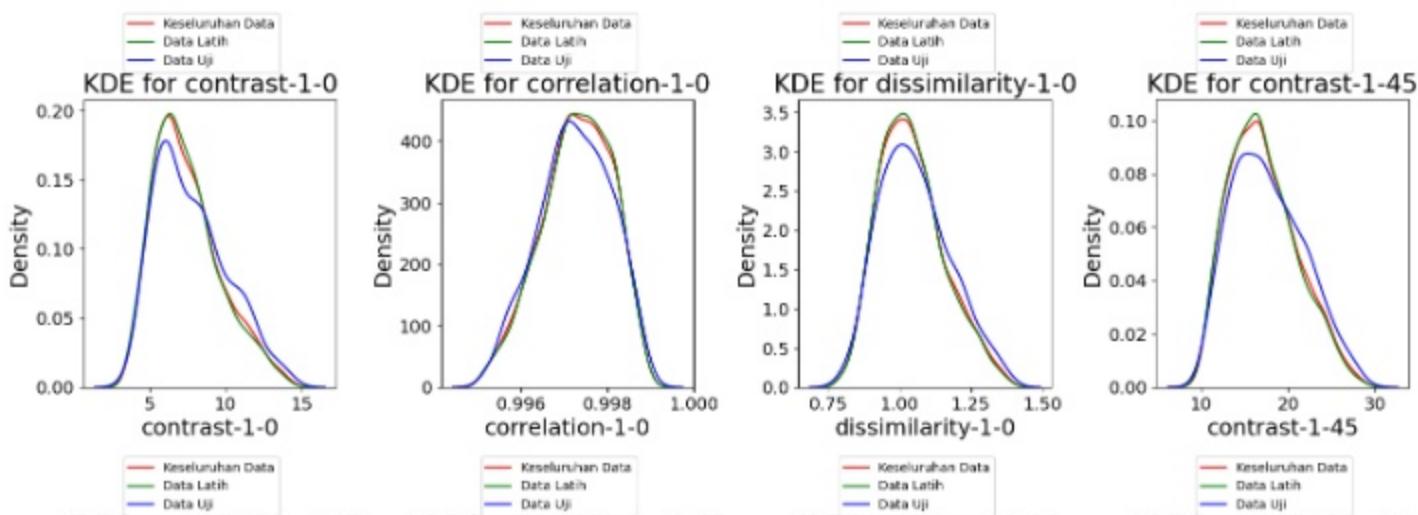
Gambar 5. Distribusi Fitur Sebelum Dilakukan Handling Outliers (IQR)

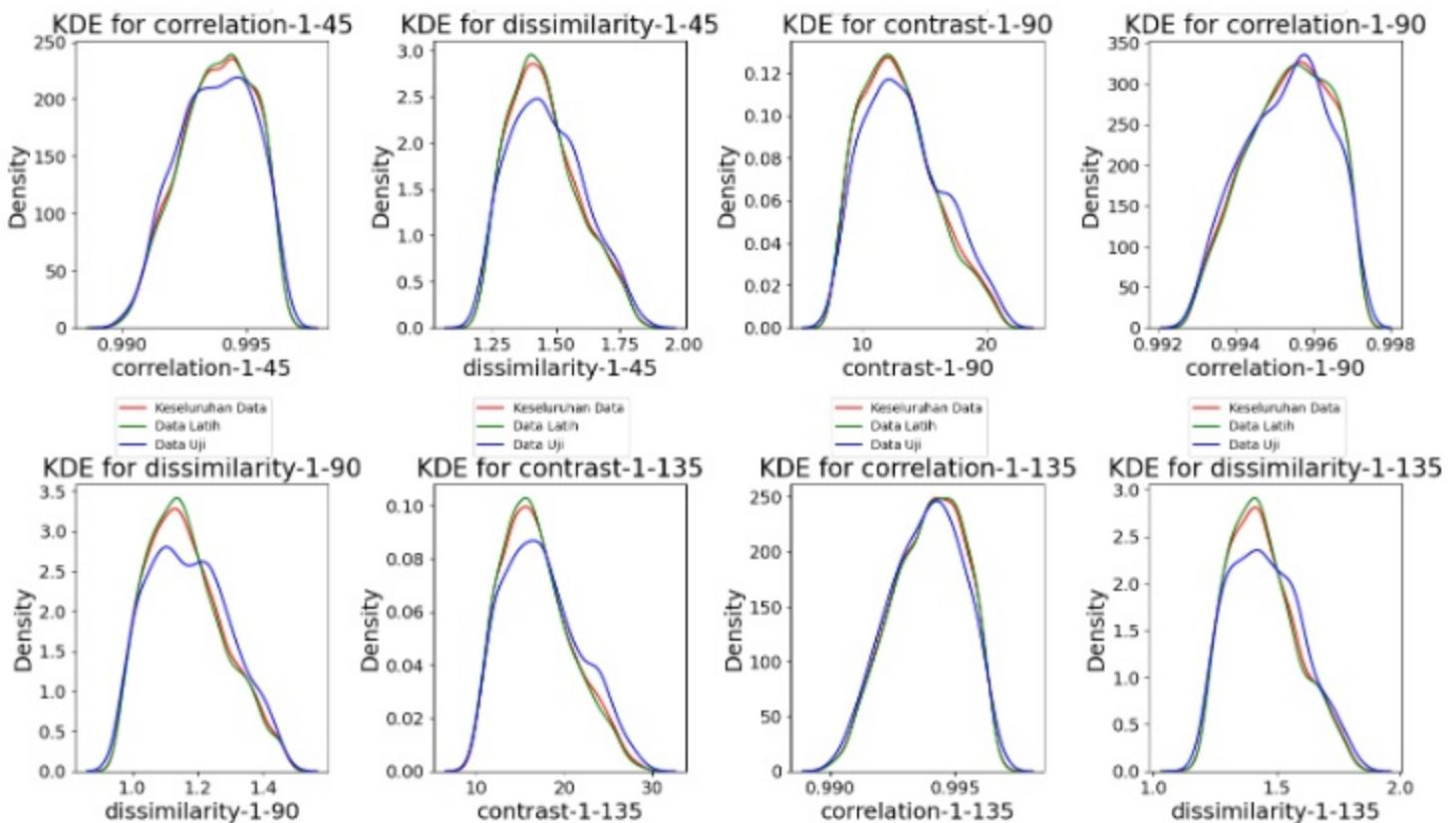
Pada grafik KDE ini, kita dapat melihat distribusi fitur-fitur yang terkait dengan contrast, correlation, dan dissimilarity pada berbagai sudut pandang ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ ,  $135^\circ$ ). Dari grafik KDE (Kernel Density Estimation) yang ditampilkan, terlihat bahwa distribusi fitur antara Keseluruhan Data, Data Latih, dan Data Uji hampir saling bertumpuk atau berhimpitan.

Hal ini mengindikasikan bahwa pembagian data latih dan data uji telah dilakukan secara proporsional dan tidak ada bias distribusi yang signifikan antara kedua set data tersebut. Distribusi yang hampir identik ini merupakan hal yang diharapkan agar model yang dibangun dapat menjadi representasi yang baik dari populasi keseluruhan.

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.

### 3.1.2. Distribusi Fitur Sesudah Dilakukan Handling Outliers (IQR)





Gambar 6. Distribusi Fitur Sesudah Dilakukan Handling Outliers (IQR)

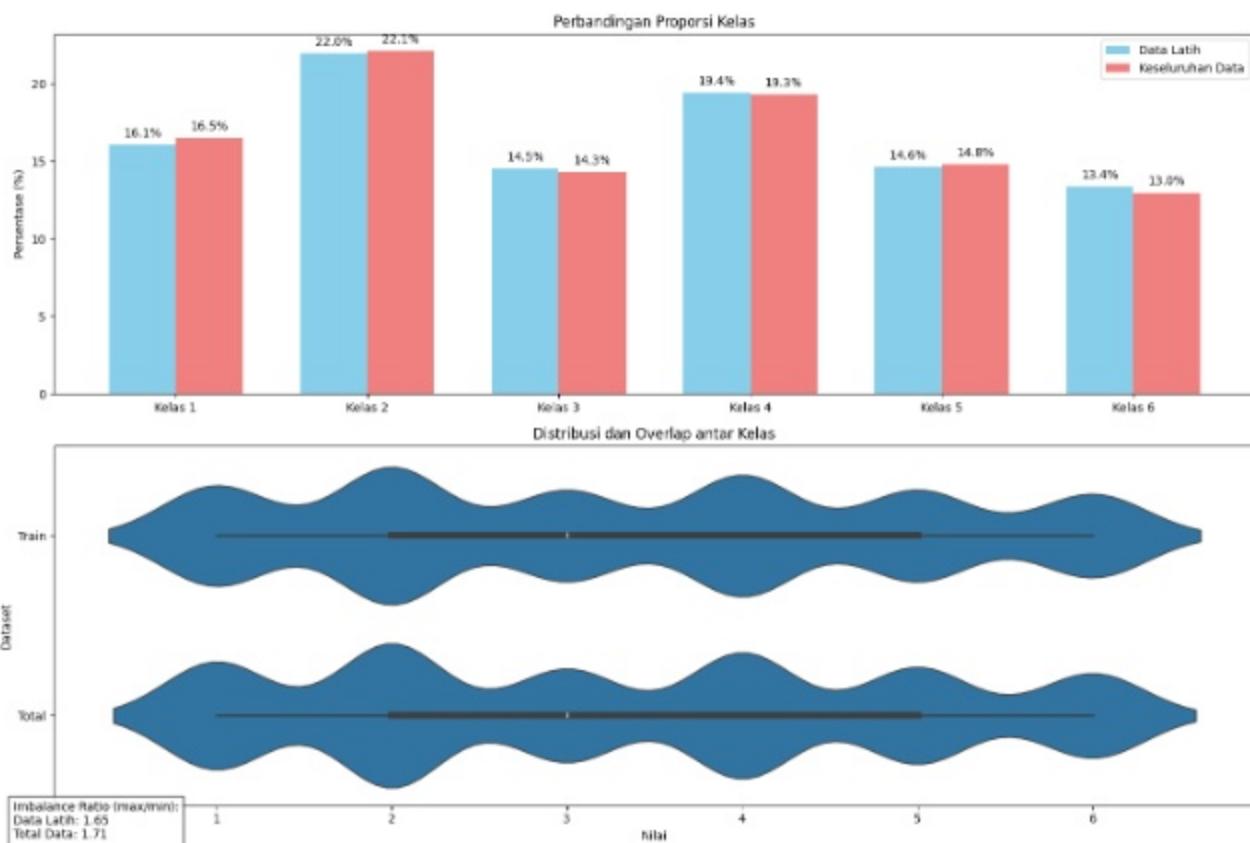
Pada grafik KDE ini, kita dapat melihat distribusi fitur-fitur yang sama setelah dilakukan penanganan outliers menggunakan metode IQR (Interquartile Range). Metode IQR digunakan untuk menentukan batas atas dan bawah yang wajar bagi setiap fitur, dengan mempertimbangkan faktor IQR yang berbeda-beda untuk setiap fitur.

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.

Dengan penanganan outliers yang disesuaikan untuk setiap fitur, kita dapat memperoleh dataset yang lebih bersih dan siap untuk dianalisis lebih lanjut. Hal ini penting untuk memastikan kualitas data dan keakuratan hasil analisis yang akan dilakukan pada tahap selanjutnya.

Secara keseluruhan, dapat disimpulkan bahwa distribusi fitur antara Keseluruhan Data, Data Latih, dan Data Uji masih mempertahankan karakteristik yang hampir serupa setelah penanganan outliers. Ini mengindikasikan bahwa pembagian data telah dilakukan dengan baik dan tidak ada bias distribusi yang signifikan antara kedua set data tersebut.

### 3.2. Proporsi Target antara Data Latih dengan Data Keseluruhan dengan Bar Plot dan Violin Plot



Gambar 7. Bar Plot dan Violin Plot Proporsi Target antara Data Latih dengan Data Keseluruhan

#### 1. Analisis Proporsi Kelas (Bar Plot)

##### 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 ( $\pm 22\%$ )
- Kelas 6 memiliki proporsi terendah ( $\pm 13\%$ )
- Kelas lainnya berada di kisaran 14-19%

##### c. 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

## **2. Analisis Distribusi dan Overlap (Violin Plot)**

### 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

## **3. Analisis Imbalance Ratio**

### 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

### **3.3. Uji Kesamaan Distribusi antara Data Latih dengan Keseluruhan Data menggunakan Kolmogorov-Smirnov**

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 p-value. 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% ( $\alpha = 0.05$ ). Jika p-value lebih besar dari tingkat signifikansi, maka distribusi antara data latih dan keseluruhan data dianggap tidak memiliki perbedaan yang signifikan.

Hasil 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.

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

Tabel 8. Hasil Distribusi Data Latih dan Keseluruhan Data menggunakan Kolmogorov-Smirnov

#### 4. 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 p-value dari uji Kolmogorov Smirnov kurang dari tingkat signifikansi (0,05), maka hipotesis bahwa data berdistribusi normal ditolak.

Tabel normalitas data dengan outlier sebelum dilakukan transformasi:

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	8.737027e-135	1.858047	False

2.	correlation-1-0	4.926734e-37	1.237148	False
3.	dissimilarity-1-0	7.946769e-87	1.323059	False
4.	contrast-1-45	2.788924e-105	1.623713	False
5.	correlation-1-45	3.854355e-20	-0.915316	False
6.	dissimilarity-1-45	2.165575e-78	1.305511	False
7.	contrast-1-90	2.168479e-106	1.597828	False
8.	correlation-1-90	1.571479e-22	-0.924536	False
9.	dissimilarity-1-90	2.131698e-75	1.279560	False
10.	contrast-1-135	2.547741e-109	1.651513	False
11.	correlation-1-135	1.797307e-24	-0.960483	False
12.	dissimilarity-1-135	2.586457e-77	1.298978	False

Tabel 9. Normalitas Data Dengan Outlier Sebelum Dilakukan Transformasi

Tabel normalitas data tanpa outlier sebelum dilakukan transformasi:

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	6.521356e-11	0.706650	False
2.	correlation-1-0	1.599958e-03	-0.342511	True
3.	dissimilarity-1-0	3.320856e-05	0.514750	False
4.	contrast-1-45	5.475939e-06	0.483419	True
5.	correlation-1-45	1.263096e-04	-0.295321	True
6.	dissimilarity-1-45	1.123786e-06	0.508232	False
7.	contrast-1-90	6.012455e-06	0.524750	False
8.	correlation-1-90	5.369402e-06	-0.270044	True
9.	dissimilarity-1-90	1.794553e-06	0.486903	True
10.	contrast-1-135	2.402572e-07	0.544470	False
11.	correlation-1-135	4.194735e-05	-0.323484	True

12.	dissimilarity-1-135	5.639294e-06	0.500536	False
-----	---------------------	--------------	----------	-------

Tabel 10. Normalitas Data Tanpa Outlier Sebelum Dilakukan Transformasi

## 5. METODE TRANSFORMASI DATA

Pada analisis ini, lima metode transformasi data digunakan untuk menormalkan fitur-fitur dalam dataset, yaitu Box-Cox, Log, Square Root (Sqrt), Arc Sin, dan Yeo-Johnson. Setiap metode memiliki pendekatan yang berbeda untuk menangani data yang tidak berdistribusi normal:

- Yeo-Johnson: Pendekatan yang fleksibel dan dapat digunakan untuk data yang memiliki nilai negatif maupun positif.
- Log: Menggunakan logaritma untuk mengurangi skewness pada data yang sangat miring, tetapi hanya dapat diterapkan pada data positif.
- Square Root (Sqrt): Mengubah data dengan mengambil akar kuadrat, sering digunakan untuk mengurangi skewness pada data yang nilainya kecil.
- Box-Cox: Metode ini digunakan untuk data positif dan mengubah distribusi menjadi lebih mendekati distribusi normal.
- Arc Sin: Metode yang sering digunakan untuk data berbentuk proporsi (antara 0 dan 1), terutama untuk distribusi berbentuk binomial.

Tabel Hasil Uji Normalitas Seluruh Metode Transformasi

No.	Metode Transformasi	Fitur Normal	Fitur Tidak Normal
1.	Yeo-Johnson	12	0
2.	Log	12	0
3.	Sqrt	12	0
4.	Box-Cox	12	0
5.	Arc Sin	4	8

Tabel 11. Hasil Uji Normalitas Seluruh Metode Transformasi

Dapat disimpulkan bahwa metode Box-cox, Log, Sqrt, dan Yeo-Johnson berhasil menormalkan seluruh fitur (12 fitur), sementara metode Arc Sin hanya menormalkan 4 fitur dari total 12 fitur. Hal ini menunjukkan bahwa metode Arc Sin memiliki keterbatasan pada jenis data dalam dataset ini.

Hasil transformasi normalisasi juga sangat dipengaruhi oleh langkah handling outlier yang telah dilakukan sebelumnya dalam tahap preprocessing menggunakan metode Interquartile Range (IQR). Penanganan outlier dengan IQR

menghilangkan nilai ekstrem yang dapat menyebabkan skewness tinggi pada data, sehingga mempermudah proses normalisasi dengan metode transformasi yang digunakan.

### 5.1. Transformasi Yeo-Johnson

Transformasi Yeo-Johnson digunakan untuk mendekati distribusi normal dan bisa digunakan untuk data yang memiliki nilai negatif dan positif. Metode ini melakukan transformasi yang mirip dengan Box-Cox namun tanpa keterbatasan pada data positif saja. Metode ini fleksibel untuk data dengan rentang nilai yang luas, jadi baik untuk menangani outlier. Namun transformasi ini menghasilkan nilai yang sulit diinterpretasikan secara langsung dan bisa jadi lambat pada dataset yang besar.

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	0.398614	0.010475	True
2.	correlation-1-0	0.007780	-0.247888	True
3.	dissimilarity-1-0	0.322874	0.023505	True
4.	contrast-1-45	0.171501	0.003917	True
5.	correlation-1-45	0.000649	-0.152703	True
6.	dissimilarity-1-45	0.047431	0.044439	True
7.	contrast-1-90	0.008905	0.012023	True
8.	correlation-1-90	0.000017	-0.177363	True
9.	dissimilarity-1-90	0.060699	0.044813	True
10.	contrast-1-135	0.238453	0.007851	True
11.	correlation-1-135	0.000765	-0.179974	True
12.	dissimilarity-1-135	0.094152	0.040430	True

Tabel 12. Hasil Transformasi Yeo-Johnson

### 5.2. Transformasi Log

Transformasi Log diterapkan untuk mengurangi skewness dengan mengambil logaritma dari nilai-nilai data, jadi sangat efektif untuk data positif

dengan rentang besar. Metode ini sangat baik dalam mengurangi skewness dan mengompres rentang data. Namun metode ini hanya berlaku untuk data dengan nilai positif, tidak dapat menangani data nol atau negatif yang tanpa penyesuaian tambahan.

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	0.030708	0.149660	True
2.	correlation-1-0	0.001577	-0.343382	True
3.	dissimilarity-1-0	0.001907	0.382994	True
4.	contrast-1-45	0.136684	0.043778	True
5.	correlation-1-45	0.000120	-0.296646	True
6.	dissimilarity-1-45	0.000045	0.399173	True
7.	contrast-1-90	0.010355	0.108069	True
8.	correlation-1-90	0.000005	-0.270898	True
9.	dissimilarity-1-90	0.000245	0.382732	True
10.	contrast-1-135	0.214361	0.089350	True
11.	correlation-1-135	0.000041	-0.324815	True
12.	dissimilarity-1-135	0.000480	0.388436	True

Tabel 13. Hasil Transformasi Log

### 5.3. Transformasi Square Root

Transformasi Square Root mengambil akar kuadrat dari nilai data yang umumnya digunakan untuk data yang sudah cukup mendekati normal namun tetap memiliki skewness ringan. Metode ini efektif untuk mengurangi skewness tanpa mengubah distribusi terlalu drastis. Sama halnya seperti metode Log, metode ini hanya cocok untuk data positif, namun kurang efektif jika skewness sangat tinggi.

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	0.000028	0.390277	True

2.	correlation-1-0	0.001577	-0.343384	True
3.	dissimilarity-1-0	0.001832	0.386554	True
4.	contrast-1-45	0.020445	0.250215	True
5.	correlation-1-45	0.000120	-0.296650	True
6.	dissimilarity-1-45	0.000027	0.416673	True
7.	contrast-1-90	0.004825	0.298329	True
8.	correlation-1-90	0.000005	-0.270900	True
9.	dissimilarity-1-90	0.000177	0.390749	True
10.	contrast-1-135	0.003299	0.302900	True
11.	correlation-1-135	0.000041	-0.324819	True
12.	dissimilarity-1-135	0.000269	0.406139	True

Tabel 14. Hasil Square Root

#### 5.4. Transformasi Box Cox

Transformasi Box Cox dirancang untuk normalisasi data yang bersifat positif dengan menggunakan parameter lambda yang dioptimalkan untuk menormalkan distribusi. Metode ini sangat fleksibel dan kuat dalam menormalkan data positif dengan beragam distribusi. Namun sama seperti metode Log, metode ini tidak dapat diaplikasikan pada data dengan nilai negatif atau nol.

No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	0.449371	0.004604	True
2.	correlation-1-0	0.029791	-0.032218	True
3.	dissimilarity-1-0	0.390466	0.014691	True
4.	contrast-1-45	0.181637	0.001507	True
5.	correlation-1-45	0.000768	-0.041371	True
6.	dissimilarity-1-45	0.057394	0.035523	True
7.	contrast-1-90	0.009561	0.008352	True

8.	correlation-1-90	0.000131	-0.045695	True
9.	dissimilarity-1-90	0.070634	0.034029	True
10.	contrast-1-135	0.247360	0.005324	True
11.	correlation-1-135	0.004368	-0.042885	True
12.	dissimilarity-1-135	0.105209	0.031704	True

Tabel 15. Hasil Box Cox

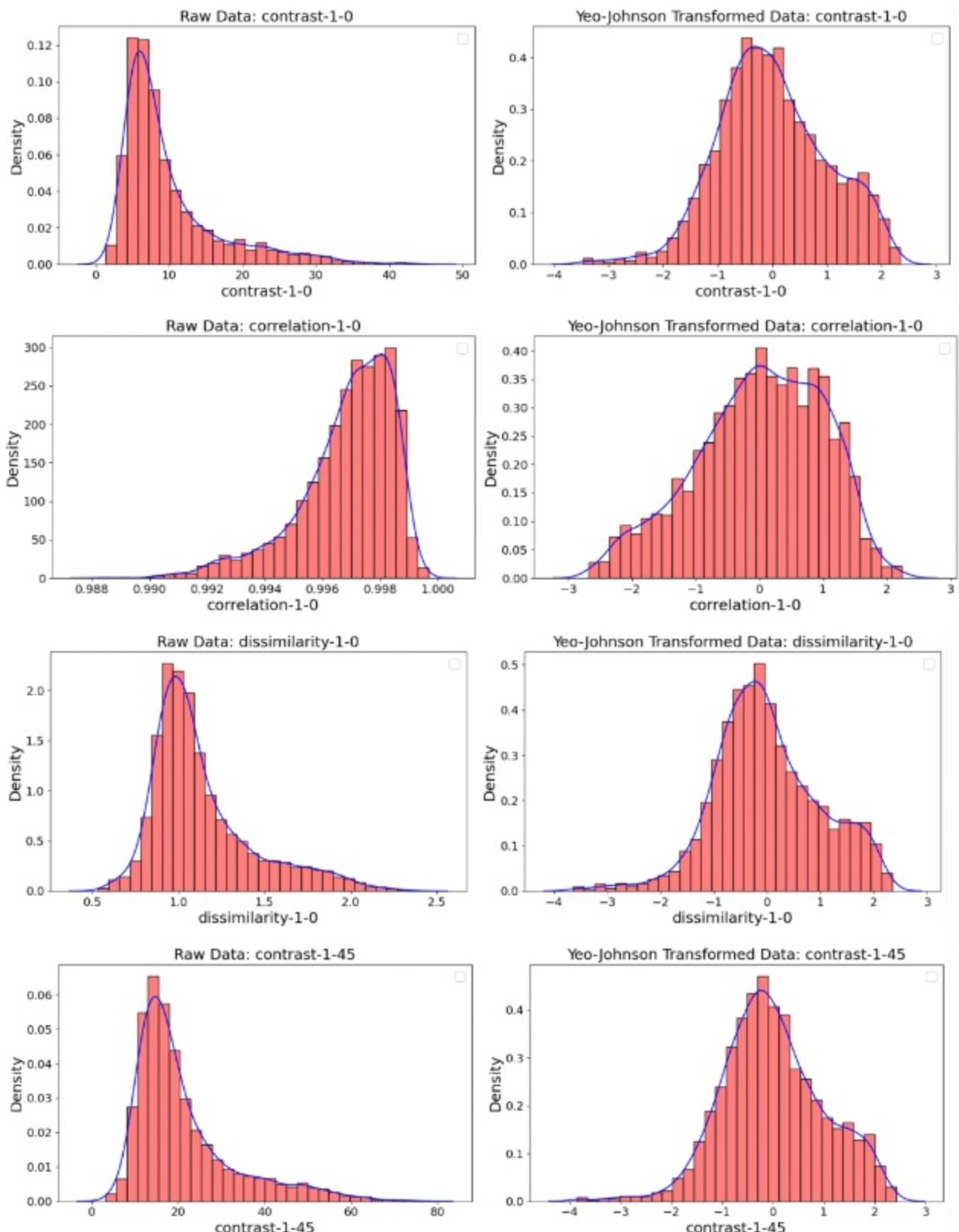
### 5.5. Transformasi Arc Sin

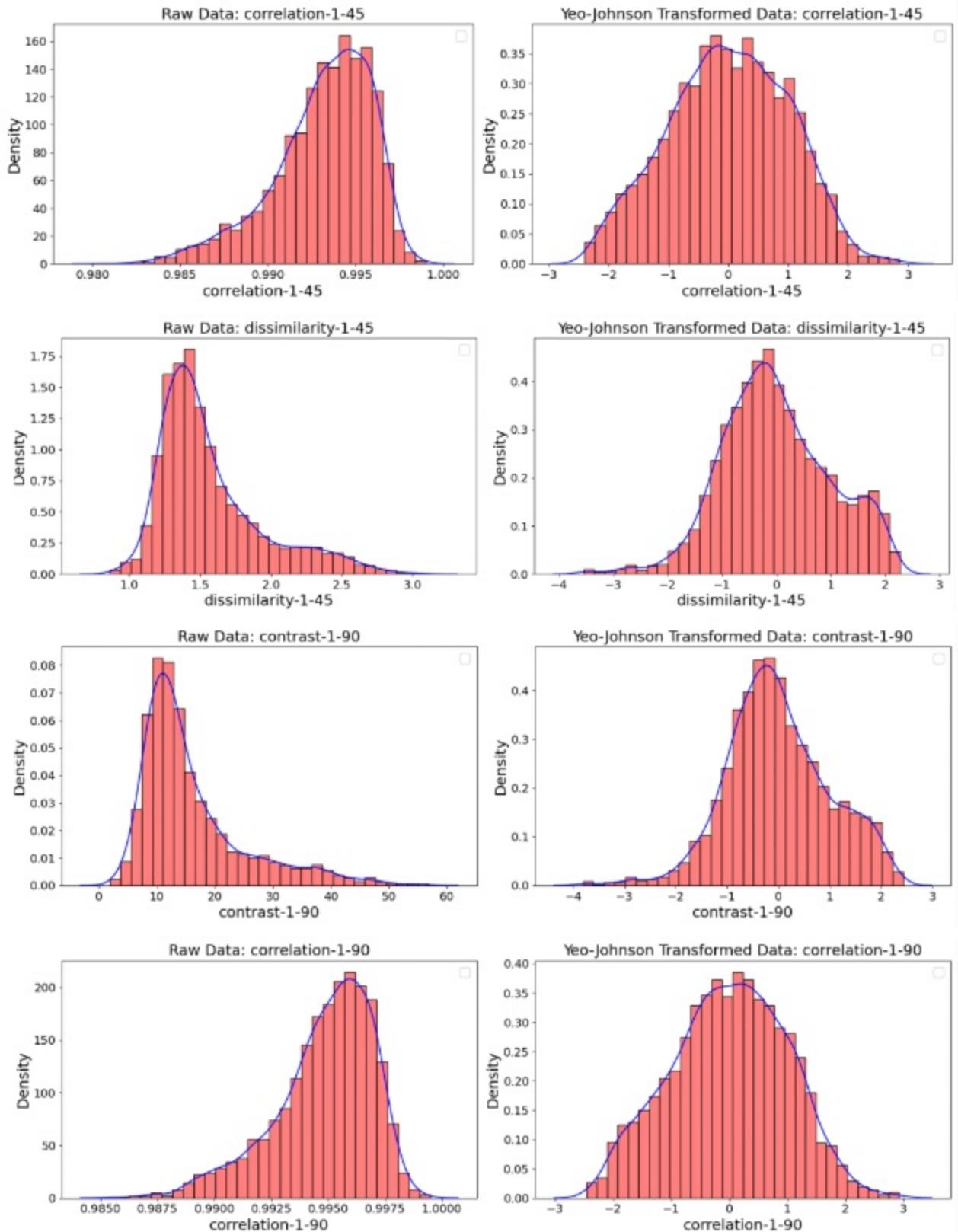
Transformasi Arc Sin utamanya digunakan untuk data proporsi atau data yang memiliki batas antara 0 dan 1, seperti proporsi atau persentase. Metode ini mampu mengurangi skewness pada data proporsi atau persentase. Namun metode ini hanya berlaku untuk data yang berada dalam rentang 0 sampai 1 sehingga memerlukan penyesuaian jika data berada di luar rentang tersebut.

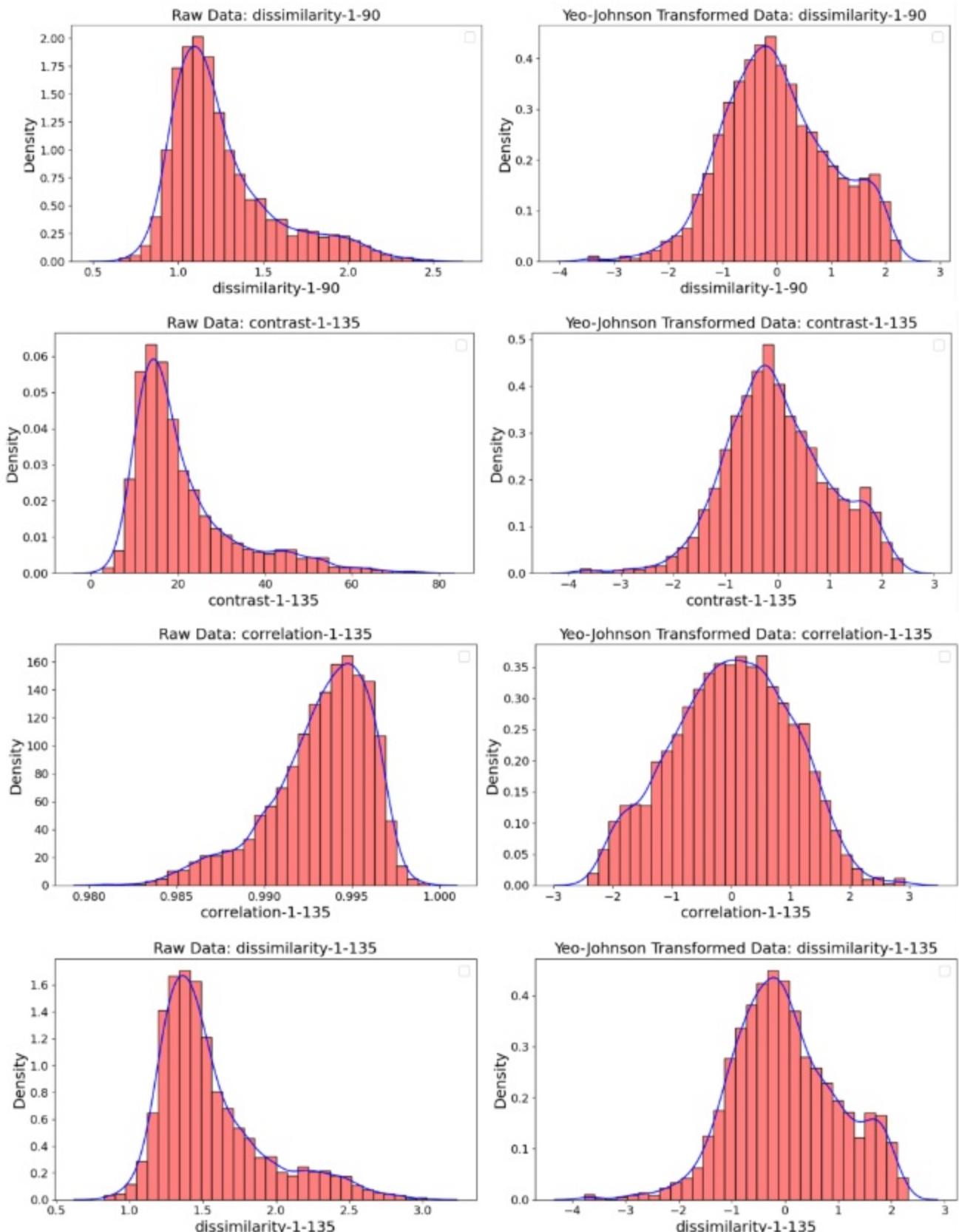
No.	Fitur	KS P-Value	Skewness	Normal
1.	contrast-1-0	-	-	False
2.	correlation-1-0	0.060565	-0.021328	True
3.	dissimilarity-1-0	-	-	False
4.	contrast-1-45	-	-	False
5.	correlation-1-45	0.001316	-0.088229	True
6.	dissimilarity-1-45	-	-	False
7.	contrast-1-90	-	-	False
8.	correlation-1-90	0.000120	-0.090688	True
9.	dissimilarity-1-90	-	-	False
10.	contrast-1-135	-	-	False
11.	correlation-1-135	0.003620	-0.110845	True
12.	dissimilarity-1-135	-	-	False

Tabel 16. Hasil Transformasi Arc Sin

## 5.6. Tabel Perbandingan Hasil Transformasi menggunakan Yeo-Johnson







Gambar 8. Perbandingan Hasil Transformasi menggunakan Yeo-Johnson

Transformasi Yeo-Johnson menghasilkan persebaran data yang lebih simetris apabila dibandingkan dengan data yang belum ditransformasi. Karena keberhasilan Yeo-Johnson dalam mengurangi skewness dan meningkatkan normalitas semua fitur, transformasi ini yang akan digunakan untuk tahapan selanjutnya, yaitu pemodelan untuk klasifikasi data.

## 6. METODE KLASIFIKASI DATA

### 6.1. Implementasi Random Forest dan FastICA untuk Klasifikasi Data

Dalam metode klasifikasi ini, kami menggunakan algoritma Random Forest yang dikombinasikan dengan FastICA untuk dimensionality reduction. Random Forest adalah metode ensemble learning yang menggabungkan multiple decision trees untuk menghasilkan prediksi yang lebih akurat dan stabil. Setiap pohon keputusan dalam Random Forest dibangun menggunakan subset data yang berbeda (bootstrap sampling) dan subset fitur yang dipilih secara acak, yang membantu mengurangi overfitting dan meningkatkan generalisasi model.

FastICA (Fast Independent Component Analysis) digunakan sebagai tahap preprocessing untuk melakukan dimensionality reduction. Metode ini bekerja dengan mencari komponen-komponen independen dari data input, yang membantu mengurangi dimensi data sambil mempertahankan informasi penting. Dengan kombinasi ini, FastICA pertama kali mentransformasi data ke dalam representasi yang lebih kompak, kemudian Random Forest melakukan klasifikasi pada data yang telah ditransformasi.

Random Forest bekerja dengan membangun sejumlah pohon keputusan secara paralel, di mana setiap pohon memberikan prediksi kelas, dan hasil akhir ditentukan melalui voting mayoritas. Setiap pohon dilatih menggunakan subset data yang berbeda melalui teknik bootstrap sampling, yang membantu meningkatkan keragaman dan mengurangi variance dalam prediksi. Pemilihan fitur secara acak pada setiap split node juga membantu menciptakan pohon-pohon yang berbeda, meningkatkan kemampuan model untuk menangkap berbagai aspek dari data.

### 6.2. Tahapan Eksperimen dan Implementasi Model

Untuk melakukan proses klasifikasi, kami melalui beberapa tahapan hingga mencapai hasil yang optimal:

1. Proses preprocessing dalam penelitian ini meliputi dua tahap utama:
  - a. Penanganan Outliers dengan IQR
    - Untuk fitur dengan faktor IQR 0.5 (contrast-1-0, correlation-1-0, dissimilarity-1-0, contrast-1-135, correlation-1-135,

- dissimilarity-1-135): Nilai di luar rentang Q1 - 0.5IQR dan Q3 + 0.5IQR dihapus
- Untuk fitur dengan faktor IQR 0.75 (contrast-1-45, correlation-1-45, dissimilarity-1-45): Nilai di luar rentang Q1 - 0.75IQR dan Q3 + 0.75IQR dihapus
- Untuk fitur dengan faktor IQR 0.25 (contrast-1-90, correlation-1-90, dissimilarity-1-90): Nilai di luar rentang Q1 - 0.25IQR dan Q3 + 0.25IQR dihapus

#### b. Normalisasi Data

Transformasi Yeo-Johnson digunakan untuk menormalisasi data karena transformasi ini mampu menormalkan semua fitur yang ada. Selanjutnya kami menerapkan AutoML dari TPOT Classifier untuk mendapatkan model classifier yang terbaik untuk menangani data yang ada. Diperoleh model Random Forest dengan FastICA sebagai bagian dari pipeline untuk melakukan dimensionality reduction sebelum dilakukan klasifikasi.

2. Algoritma model utama yang digunakan adalah Random Forest Classifier dengan FastICA sebagai komponen preprocessing. Random Forest dipilih karena kemampuannya yang unggul dalam menangani data kompleks, ketahanan terhadap overfitting, dan kemampuan untuk memberikan importance scores untuk setiap fitur. Dalam implementasinya, model ini dikonfigurasi dengan berbagai parameter yang telah dioptimalkan untuk kasus klasifikasi ini:

#### a. Random Forest:

- Bootstrap diaktifkan untuk melakukan sampling dengan penggantian, yang memungkinkan model untuk menciptakan subset data training yang berbeda-beda, meningkatkan variasi dalam ensemble dan mengurangi overfitting.
- Criterion menggunakan entropy sebagai metrik pemisahan, yang mengukur ketidakmurnian node berdasarkan distribusi kelas, membantu dalam pemilihan split yang optimal.
- Max Features dibatasi pada 65% dari total fitur, memberikan keseimbangan antara eksplorasi fitur dan efisiensi komputasi.
- Min Samples Leaf diatur ke 1, memungkinkan pembentukan leaf node yang sangat spesifik jika diperlukan.
- Min Samples Split ditetapkan pada 10, mencegah pembuatan split yang terlalu detail pada data yang terbatas.

- N Estimators diatur ke 100, menyediakan jumlah tree yang cukup untuk ensemble learning yang efektif.

b. FastICA:

- Tolerance diatur ke 0.35 untuk proses transformasi, memberikan keseimbangan antara akurasi dan efisiensi komputasi dalam proses pengurangan dimensi.

3. Proses optimalisasi model dilakukan melalui eksperimen ekstensif dengan berbagai kombinasi parameter. Pendekatan ini terbagi menjadi dua kategori utama:

a. Grid Parameter yang Divariasikan:

- **n\_estimators**: Diuji dengan nilai [50, 100, 150, 200] untuk mengevaluasi pengaruh jumlah tree terhadap performa model. Variasi ini memungkinkan kita menemukan keseimbangan optimal antara kompleksitas model dan akurasi prediksi.
- **min\_samples\_leaf**: Diuji dengan nilai [1, 2, 4, 6] untuk mengoptimalkan ukuran minimum leaf node. Parameter ini krusial untuk mengontrol tingkat detail dari pohon keputusan dan mencegah overfitting.

b. Parameter Tetap yang Dipertahankan:

- bootstrap tetap diaktifkan untuk memastikan konsistensi dalam sampling data
- criterion tetap menggunakan entropy untuk menjaga konsistensi dalam evaluasi split
- max\_features dipertahankan pada 0.65 sebagai nilai optimal dari eksperimen sebelumnya
- min\_samples\_split tetap 10 untuk menjaga stabilitas pemisahan node
- FastICA tolerance tetap 0.35 sesuai dengan hasil optimasi awal

4. Model dievaluasi menggunakan metrik akurasi, yang dihitung sebagai rasio antara prediksi benar dan total prediksi. Untuk mendapatkan hasil yang robust, eksperimen dilakukan dengan cross-validation untuk setiap kombinasi parameter.

5. Tabulasi hasil eksperimen disajikan dalam bentuk tabel yang mencakup berbagai kombinasi parameter dan metrik evaluasi yang sesuai. Setiap kombinasi parameter diuji beberapa kali untuk memastikan stabilitas hasil.

Parameter Random Forest		Akurasi Percobaan						
n_estimators	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%	
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%	51.68%	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%	

Tabel 17. Hasil Percobaan Random Forest

Percobaan menunjukkan akurasi tertinggi sebesar 54,68% pada pengaturan n\_estimators=200 dan min\_samples\_leaf=1. Hal ini menunjukkan bahwa penggunaan jumlah estimator yang lebih besar (n\_estimators) dengan pembatasan sampel minimal daun yang kecil memberikan hasil yang optimal.

Pengaturan dengan rata-rata akurasi tertinggi adalah juga pada n\_estimators=200 dan min\_samples\_leaf=1, dengan rata-rata akurasi sebesar 52,63%.

Meskipun begitu, rata-rata akurasi secara keseluruhan tidak mengalami fluktuasi yang besar antara pengaturan parameter, berkisar antara 51,31% hingga 52,63%. Hal ini menunjukkan bahwa variasi parameter mungkin memiliki pengaruh yang kecil pada performa model dalam dataset ini.

## 7. UJI HIPOTESIS

- H0: Tidak ada pengaruh signifikan dari perubahan nilai parameter terhadap akurasi.
- H1: Ada pengaruh signifikan dari perubahan nilai parameter terhadap akurasi.

Uji hipotesis menggunakan One-Way ANOVA secara terpisah untuk faktor parameter `n_estimators` dan `min_samples_leaf` untuk melihat apakah masing-masing faktor berpengaruh signifikan terhadap akurasi.

ANOVA		
Parameter	F-statistic	p-value
<code>n_estimators</code>	2.0376796989403028	0.1623347961084501
<code>min_samples_leaf</code>	1.0188740985065572	0.4185077152369911

Tabel 18. Hasil Uji Hipotesis One-Way ANOVA

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.