Simulasi dan Evaluasi Metode Imputasi Data Hilang Menggunakan Pendekatan Statistik, Pembelajaran Mesin, dan *Reinforcement Learning*

Studi percobaan metode imputasi data hilang menggunakan statistika konvensional, pembelajaran mesin, dan pembelajaran penguatan: pendekatan multi-disiplin untuk optimalisasi kualitas data ditinjau dari nilai kebaikan.

Ghardapaty G. Ghiffary & Adhiyatma Nugraha

00
Today's Outline

Today Report Outline

- ❏ Latar Belakang
- ❏ Sekilas Imputasi dan Data Hilang
- ❏ Imputasi Rata – Rata (*Mean*)
- ❏ Imputasi Nilai Tengah (*Median*)
- ❏ Imputasi *K – Nearest Neighbour*
- ❏ Imputasi *Fuzzy K – Means Clustering*
- ❏ Imputasi *Isolation Forest*
- ❏ Imputasi *Reinforcement Learning*
- ❏ Simpulan Imputasi Data Hilang

Latar Belakang

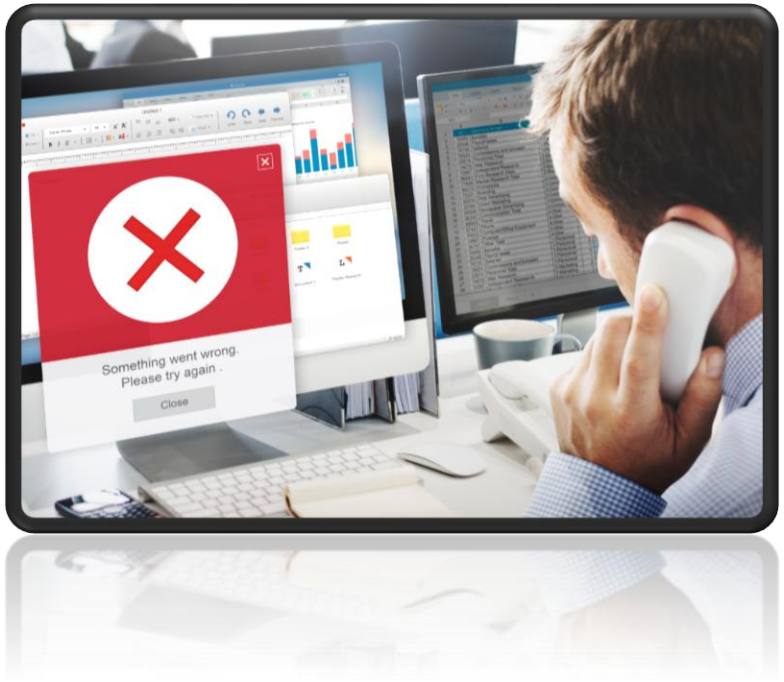
1. Menghindari Bias Analisis

Hasil analisis menjadi bias jika tidak ditangani dengan benar, karena data yang tersedia tidak mewakili populasi secara keseluruhan.

2. Pertahankan Ukuran Contoh

Imputasi tanpa menghapus baris atau kolom yang memiliki nilai hilang, sehingga menjaga informasi dan statistik.

Data hilang perlu ditangani?



3. Optimalisasi Data yang Ada

Data hilang biasa menyembunyikan informasi penting yang tersebar di peubah lain, menghapus baris atau kolom bisa fatal.

4. Mengurangi Tk. Kesalahan

Data hilang menimbulkan ketidakpastian dalam hasil analisis dan keputusan. Mengatasi ini dapat mengurangi risiko tersebut.

Sekitar 60% penelitian sosial memiliki data hilang, jika tidak diimputasi dapat mengubah kesimpulan penelitian (Little & Rubin, 2019).

Ada banyak metode dan cara dalam melakukan imputasi data hilang, tapi tunggu. Apa itu IMPUTASI DATA HILANG?

Latar Belakang

1. Imputasi Bagian 1

Ketika sampling, ditemui jawaban yang tidak dijawab Sedangkan Ketika melakukan pendugaan tidak dapat dilakukan Ketika terdapat nilai atau amatan hilang.

2. Imputasi Bagian 2

Teknik imputasi digunakan untuk menetapkan nilai pada data pengamatan yang hilang. Ada beberapa cara yang dapat dilakukan pada Teknik Imputasi ini.

3. Catatan Imputasi

Nilai hilang yang digantikan imputasi biasanya akan menghasilkan perkiraan standar error yang lebih rendah karena kesalahan yang melekat pada nilai imputasi tidak diukur.

4. Catatan Imputasi

Metodenya beragam dari menggunakan mean hingga metode kompleks berbasis pembelajaran mesin dan pendekatan baru seperti reinforcement learning.

Penggabungan metode statistika klasik dengan pembelajaran mesin dapat meningkatkan performa imputasi

Reinforcement learning, pendekatan adaptif di mana model secara iteratif belajar strategi imputasi optimal berdasarkan umpan balik.

01
Metode Imputasi

Latar Belakang

Imputasi Rata-Rata

Mengganti nilai hilang dengan rata-rata dari peubah tersebut untuk mengisi nilai hilang secara sederhana.

Imputasi Nilai Tengah

Menggunakan nilai tengah peubah untuk mengisi nilai hilang, efektif mengurangi pengaruh pencilan dibanding rata rata.

Imputasi Reinforcement Learn.

Memanfaatkan pembelajaran adaptif yang optimal untuk mengisi data hilang melalui interaksi dan umpan balik dari data yang tersedia.

Imputasi KNN

Mengisi nilai hilang berdasarkan nilai rata-rata dari tetangga terdekat yang paling mirip dalam ruang fitur atau *k.*

Imputasi Fuzzy K-Means

Menggunakan gerombol fuzzy untuk menentukan probabilitas keanggotaan data dalam gerombol dan menduga nilai hilang secara peluang.

Imputasi Isolation Forest

Mendeteksi pencilan dengan Isolation Forest guna menandai anomali data sebagai data hilang dan mengimputasi nilai yang hilang.

Data Simulasi

Data simulasi dengan 9 Peubah Prediktor (Mata Kuliah) dan 1 Peubah Respon (IPK).

Tiap peubah matkul disimulasikan dengan sebaran berbeda hingga memodelkan peubah respon yang ada*.

# Teori Statistika	# Pemrograman Statistika	# Analisis Statistika	# Analisis Deret Waktu	# Analisis Data Kategorik	# Metode Penarika
Missing: 41 (14%)	Missing: 40 (13%)	Missing: 32 (11%)	Missing: 44 (15%)	Missing: 45 (15%)	Missing:
Distinct: 80 (27%)	Distinct: 91 (30%)	Distinct: 69 (23%)	Distinct: 100 (33%)	Distinct: 61 (20%)	Distinct:
Min 2.87 Max 4.0	Min 2.5 Max 4.0	Min 3.04 Max 3.96	Min 2.5 Max 3.99	Min 3.21 Max 4.0	Min 2.99
3.61	*Missing value*	3.68	3.35	3.79	
3.18	2.96	3.61	3.44	3.83	
3.67	2.93	3.72	3.47	4.0	
3.78	3.7	3.62	3.29	3.63	
3.76	2.5	3.7	3.41	*Missing value*	
3.93	4.0	3.89	3.54	3.55	
Missing value	*Missing value*	*Missing value*	3.31	3.79	
3.48	3.05	3.8	3.63	3.33	
3.58	2.5	*Missing value*	3.25	*Missing value*	
3.46	3.21	3.94	3.15	3.81	
3.98	3.37	3.8	3.13	3.93	
3.78	2.73	3.83	3.33	3.58	
3.19	4.0	3.64	3.31	3.72	
3.52	4.0	3.85	3.32	3.21	
3.66	*Missing value*	3.66	*Missing value*	3.59	

Populasi amatan untuk data simulasi ini berjumlah 300 baris dan 10 kolom yang telah didesain mengikuti Penelitian sosial lainnya.

Tahukah Kamu ?

Pada data simulasi ini menggunakan simulasi Penelitian sosial dengan proporsi data hilang sebanyak 15 persen.

Sebaran dan visualisasi peubah – nilai hilang di lampiran

Imputasi Rata - Rata

RMSE 0.316 dan MAE 0.266 menunjukkan kesalahan yang rendah (Cocok untuk data dengan sebaran simetris dan tanpa pencilan ekstrem), karena rata rata sensitif terhadap kemencengan.

Imputasi rata rata cenderung mengurangi ragam asli data (nilai hilang diganti dengan nilai konstan) dan memengaruhi analisis statistik inferensial (misal: uji hipotesis atau interval kepercayaan).

Nilai Kebaikan Model	
RMSE	MAE
0.316	0.266

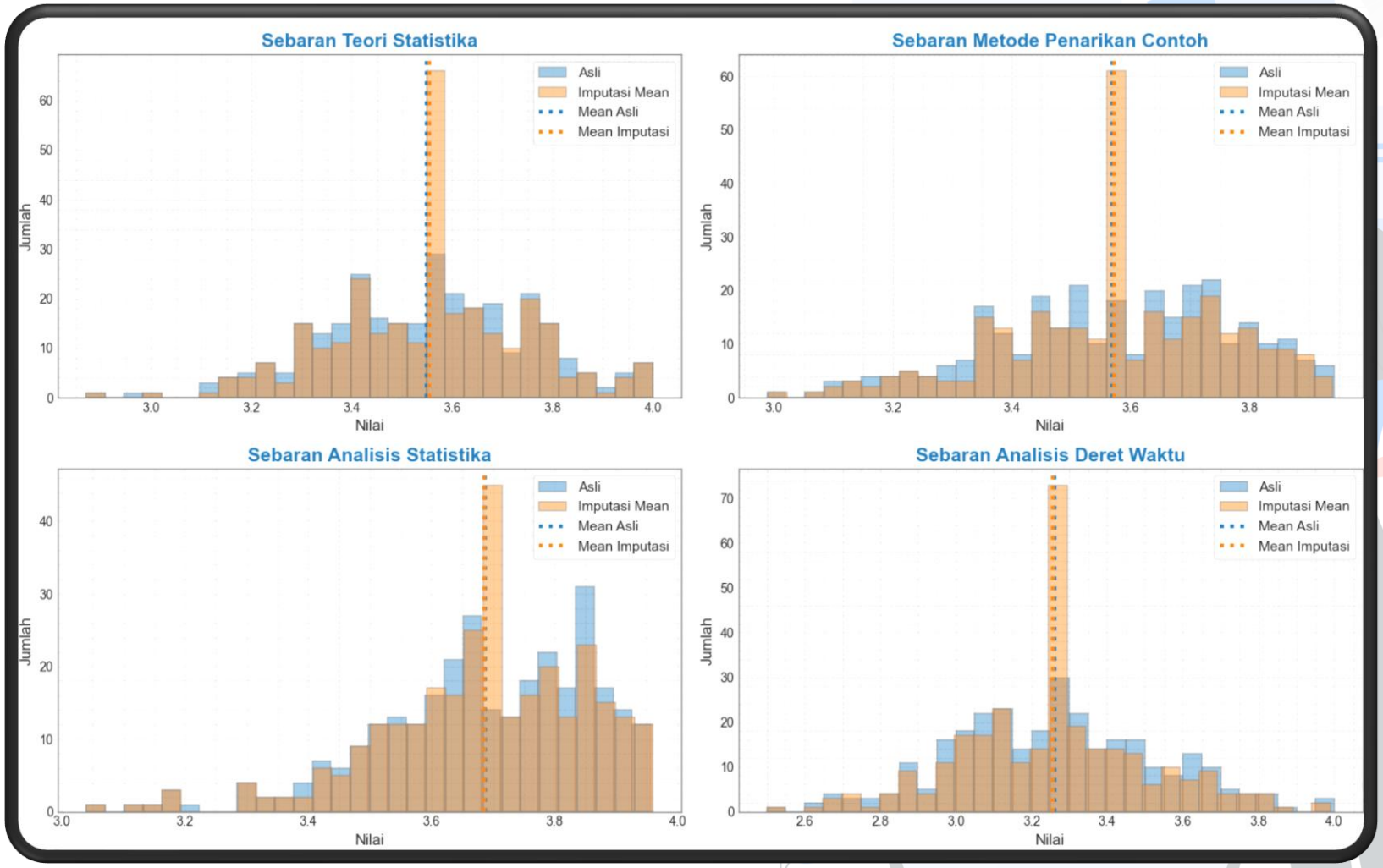
Pada Metode Penarikan Contoh (Gambar 1) menunjukkan dua puncak (Paham dan Tidak Paham), dimana Imputasi rataan mungkin memberikan nilai artifisial di antara dua kelompok (tidak merepresentasikan siapa pun).

Tahukah Kamu ?

Metode imputasi rata rata cocok untuk dengan sebaran simetris (seperti Teori Statistika), tetapi gagal menangani ragam tinggi, nilai pencilan dan kasus ketika ada 2 titik puncak.

Imputasi Rata - Rata

Gambar 1. Imputasi Rataan



Pergeseran kemencengan jika data asli menjulur. Hal ini membuat nilai pengganti menarik ekor sebaran ke tengah.

Mengabaikan korelasi antar matkul. Misal, mhs lemah di "Teori Statistika" mungkin juga lemah di "Analisis Statistika" atau lainnya.

03
Imputasi Rataan

Imputasi Nilai Tengah

RMSE 0.329 dan MAE 0.268 menunjukkan kesalahan yang rendah (Cocok untuk data dengan sebaran simetris dan tanpa pencilan ekstrem), tapi tidak lebih baik dari rata-rata.

Imputasi nilai tengah tidak terlalu mengurangi ragam asli data (nilai hilang diganti dengan nilai konstan) dan memengaruhi analisis statistik inferensial (misal: uji hipotesis atau interval kepercayaan).

Nilai Kebaikan Model	
RMSE	MAE
0.329	0.268

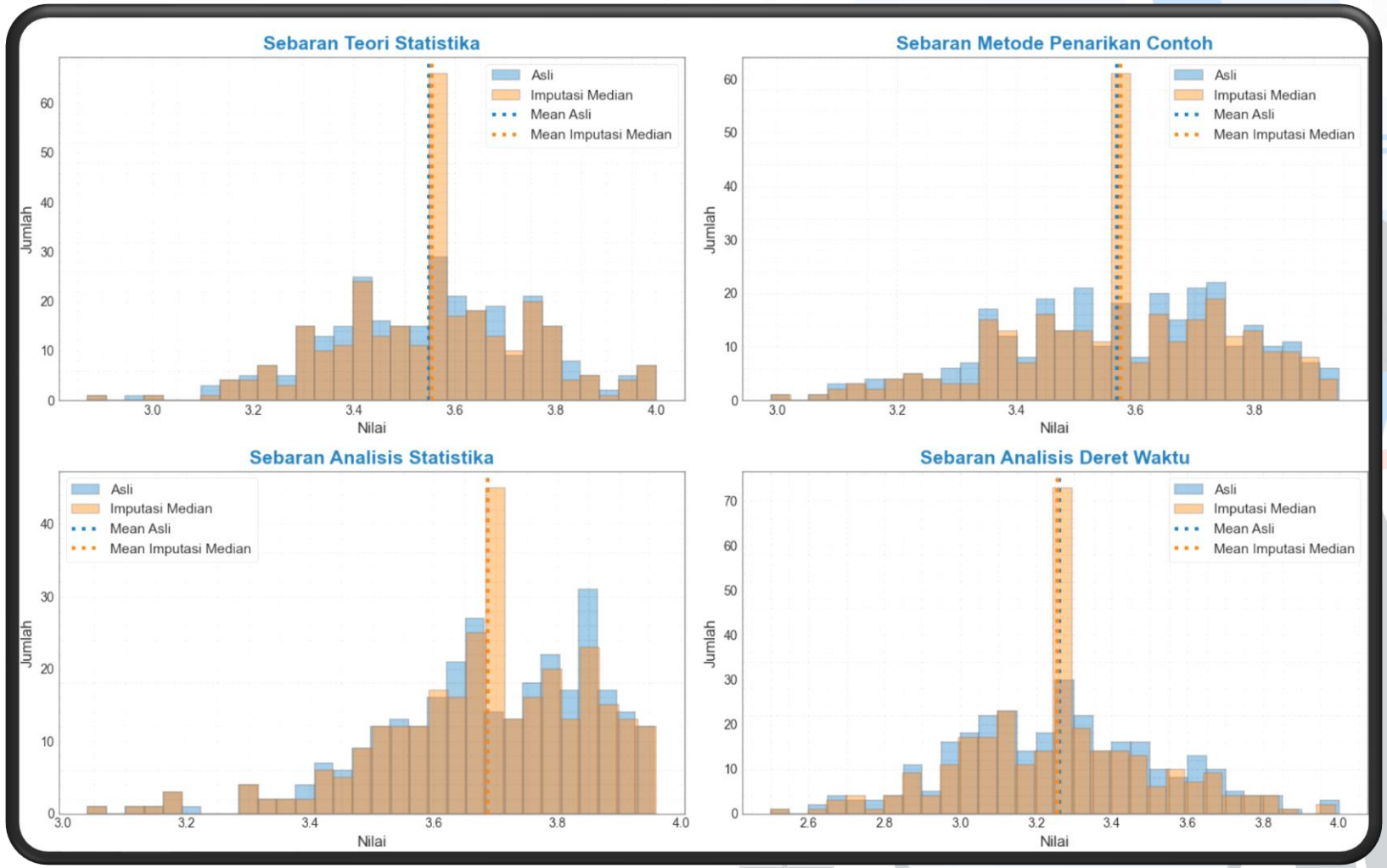
Pada Metode Penarikan Contoh (Gambar 2) menunjukkan dua puncak (Paham dan Tidak Paham), dimana Imputasi nilai tengah mungkin memberikan nilai artifisial di antara dua kelompok.

Tahukah Kamu ?

Metode imputasi nilai cocok untuk dengan sebaran simetris (seperti Teori Statistika) dan kuat untuk data pencilan serta ragam tidak terlalu tinggi tetapi gagal menangani kasus ketika ada 2 titik puncak.

Imputasi Nilai Tengah

Gambar 2. Imputasi Nilai Tengah



Tidak terlalu menggeser kemencengan jika data asli menjulur. Tapi tetap bisa merubah karakteristik sebaran asli.

Mengabaikan korelasi antar matkul. Misal, mhs lemah di "Teori Statistika" mungkin juga lemah di "Analisis Statistika" atau lainnya.

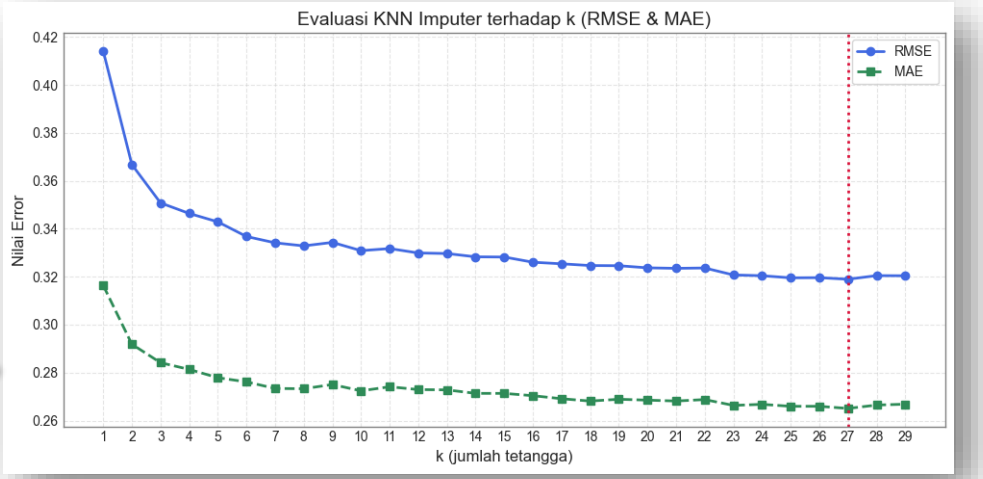
03
Imputasi Median

Imputasi K-NN

RMSE 0.319, MAE 0.265 dan *k*=27, menunjukkan kesalahan yang rendah (Cocok untuk data dengan sebaran simetris dan tanpa pencilan ekstrem) lalu lebih baik dari nilai tengah.

Nilai Kebaikan Model		
k	RMSE	MAE
27	0.319	0.265

Bentuk kurva menurun tajam di awal dan mencapai minimum lalu naik perlahan saat *k* membesar. Tidak ada fluktuasi berlebihan dan titik minimum jelas terlihat sehingga KNN cukup cocok untuk kasus ini.



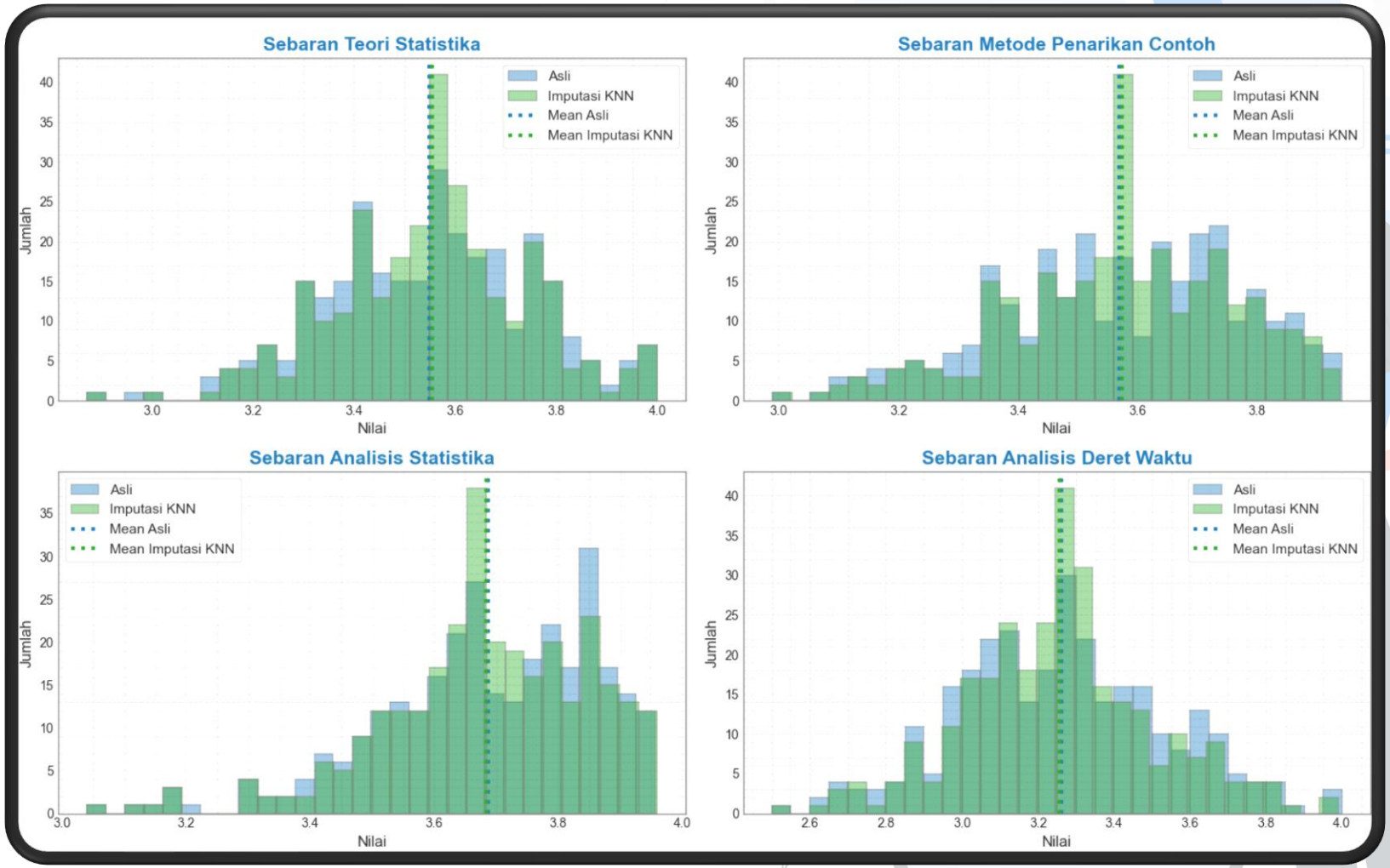
Data simulasi menunjukkan tidak terdapat gejala pencilan ekstrem sehingga penghitungan jarak lebih cocok menggunakan Euclidean dibandingkan dengan Manhattan

Tahukah Kamu ?

Karena KNN berbasis perhitungan jarak (Euclidean, Mahalanobis dan Manhattan) jadi sangat sensitif terhadap skala (perlu normalisasi).

Imputasi KNN

Gambar 3. Imputasi KNN



KNN bersifat non-parametrik, sehingga cocok pada data menjulur, banyak puncak, atau sebaran tidak normal.

KNN mempertimbangkan korelasi peubah jadi ketika ada mhs lemah di matkul A maka akan mencari kesamaan atau kemiripannya.

03
Imputasi KNN

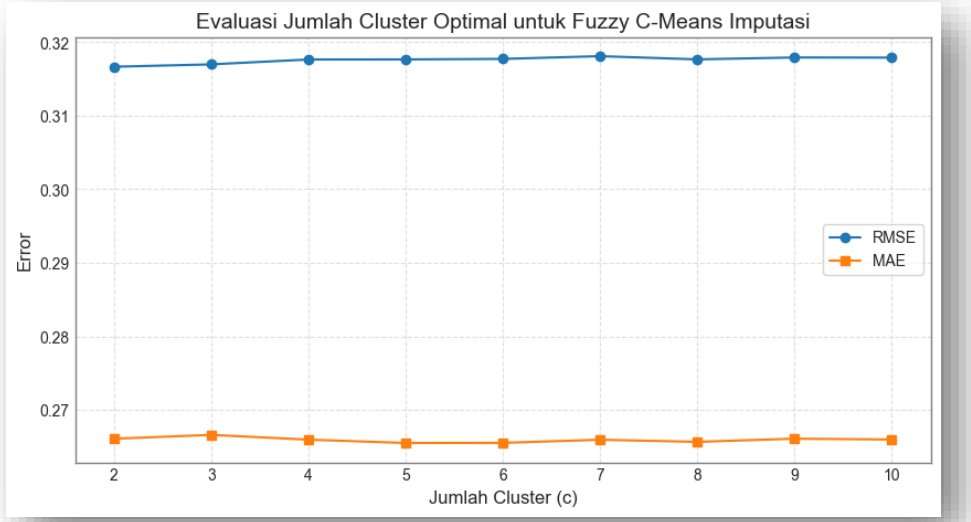
Imputasi Fuzzy K-Means

RMSE 0.316, MAE 0.266 dan k=2, menunjukkan kesalahan yang rendah (Cocok untuk data dengan sebaran simetris dan tanpa pencilan ekstrem) lalu lebih baik dari nilai tengah.

Nilai Kebaikan Model		
k	RMSE	MAE
2	0.316	0.266

2 gerombol optimum mengindikasikan bahwa sebaran data memiliki polaritas yang kuat, misalnya kelompok mahasiswa dengan nilai dominan tinggi vs rendah.

Fuzzy K-Means unggul karena mampu menangani transisi antar kelompok nilai melalui derajat keanggotaan, yang bermanfaat untuk menganalisis pola nilai yang tidak kaku.

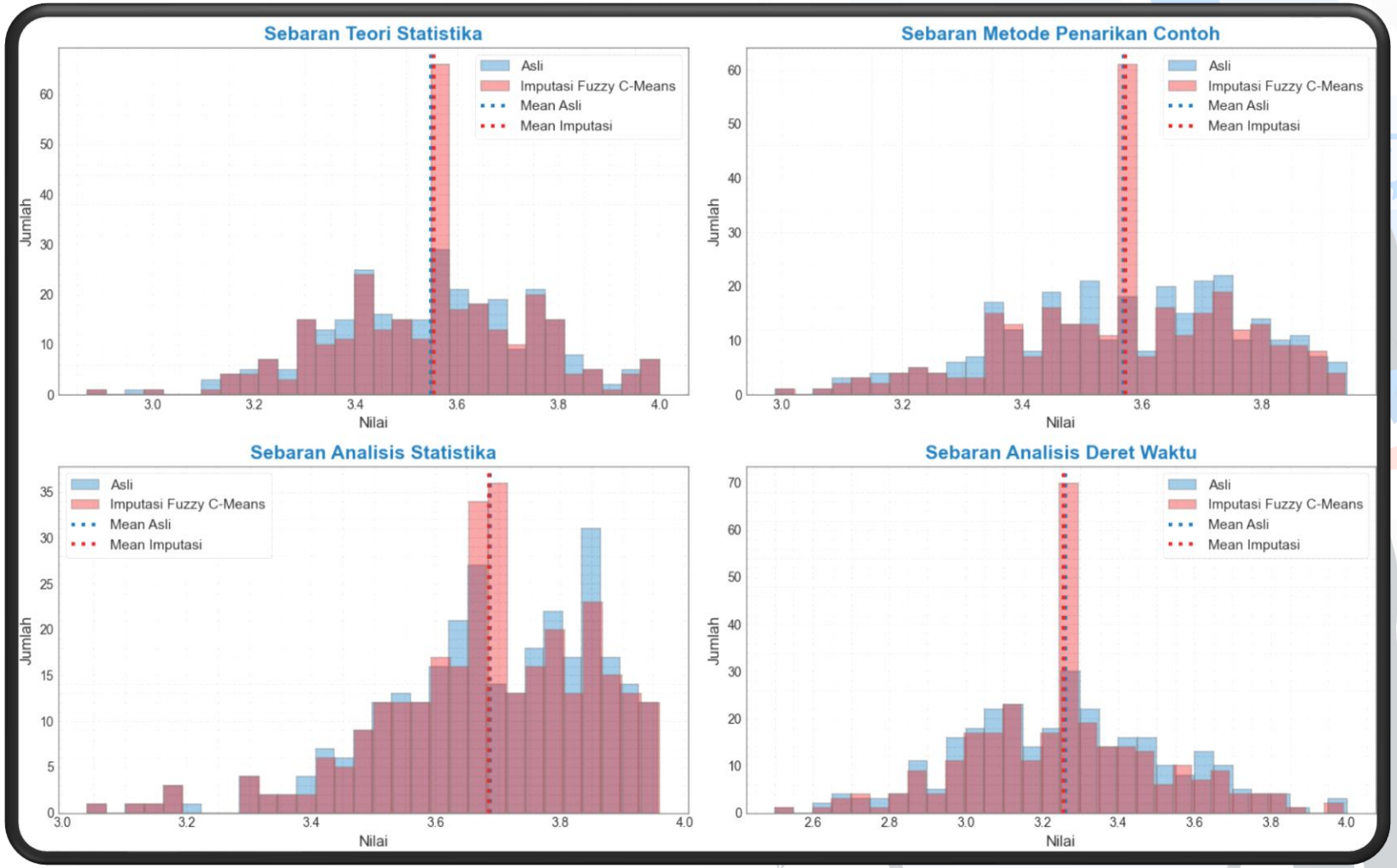


Tahukah Kamu ?

Butuh eksplorasi lebih lanjut karakteristik gerombol (misalnya rataan, ragam, atau sebaran nilai di gerombol). Memungkinkan pola “kelas homogen”.

Imputasi Fuzzy K-Means

Gambar 4. Imputasi Fuzzy



Fuzzy K-Means memungkinkan data menjadi anggota beberapa gerombol sekaligus dengan peluang berbeda.

Mengabaikan korelasi antar matkul. Misal, mhs lemah di "Teori Statistika" mungkin juga lemah di "Analisis Statistika" atau lainnya.

Imputasi Iso. Forest

RMSE 0.324 dan MAE 0.268 menunjukkan kesalahan yang rendah (Cocok untuk data dengan sebaran simetris dan tanpa pencilan ekstrem), tapi tidak cukup bagus bila dibanding dengan metode lainnya.

Isolation Forest mengidentifikasi nilai hilang yang mungkin pencilan sebelum diimputasi. Jika nilai hilang mengumpul di area pencilan, hasil imputasi akan lebih akurat.

Nilai Kebaikan Model	
RMSE	MAE
0.324	0.268

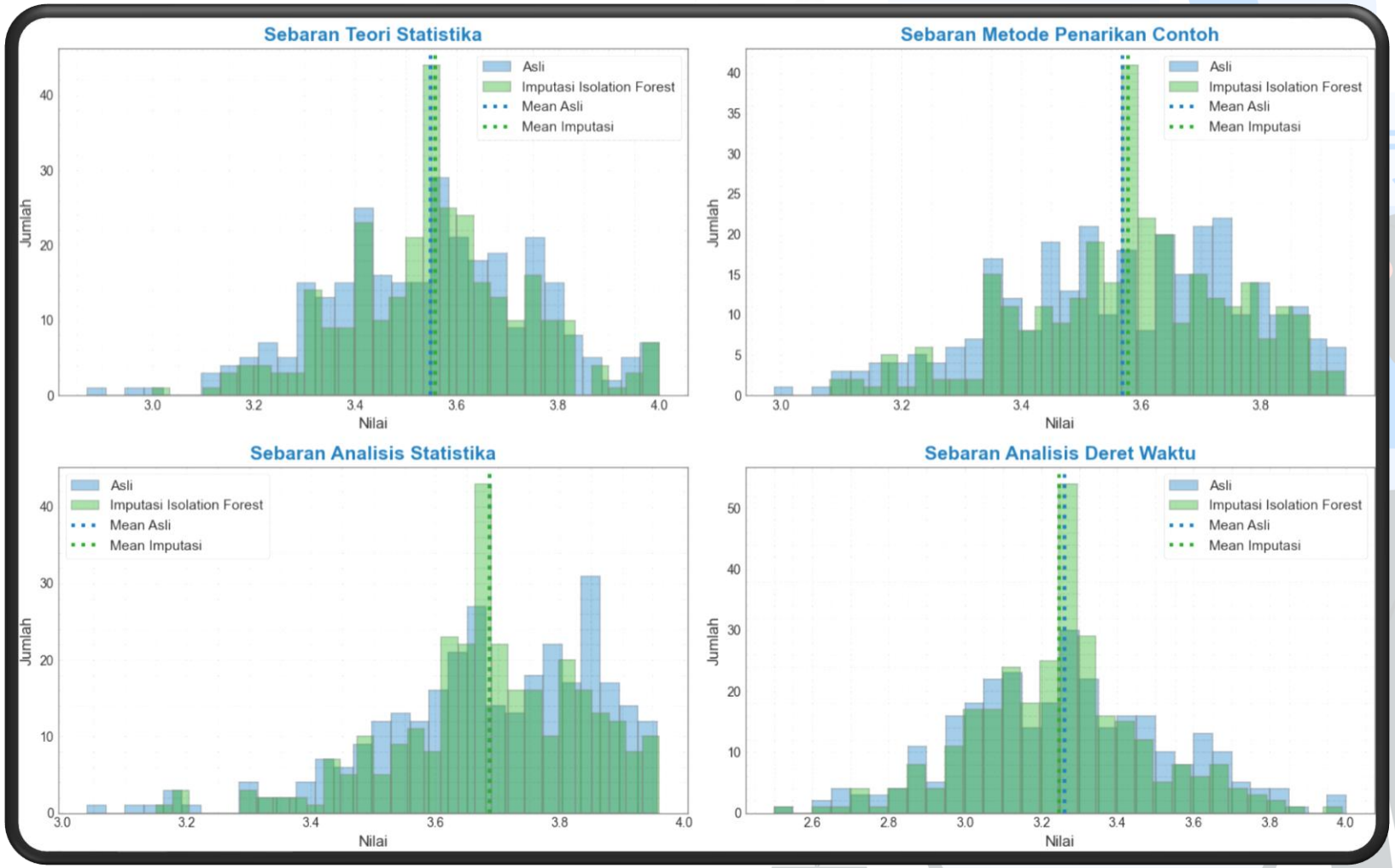
Pada Metode Penarikan Contoh (Gambar 5) menunjukkan dua puncak (Paham dan Tidak Paham), dimana Imputasi ini mungkin memberikan nilai artifisial di antara dua kelompok (tidak merepresentasikan siapa pun).

Tahukah Kamu ?

Metode imputasi rata rata cocok untuk dengan sebaran simetris (seperti Teori Statistika), tetapi gagal menangani ragam tinggi, nilai pencilan dan kasus ketika ada 2 titik puncak.

Imputasi Iso. Forest

Gambar 5. Imputasi Isolation Forest



Isolation Forest menggunakan pohon acak untuk mengisolasi pencilan secara cepat. efektif pada data berdimensi tinggi.

Mengabaikan korelasi antar matkul. Misal, mhs lemah di "Teori Statistika" mungkin juga lemah di "Analisis Statistika" atau lainnya.

Imputasi Reinforcement Learn.

MAE 0.015 dan RMSE 0.022 menunjukkan error hampir mendekati nol, artinya metode ini nyaris sempurna dalam memprediksi nilai hilang. Cocok untuk kasus dimana kesalahan kecil pun punya dampak besar.

Meski galat sangat rendah, plot sebaran pasca-imputasi tidak menunjukkan pola *overfitting* (seperti puncak artifisial), mengindikasikan model belajar pola alami data.

Nilai Kebaikan Model	
RMSE	MAE
0.022	0.015

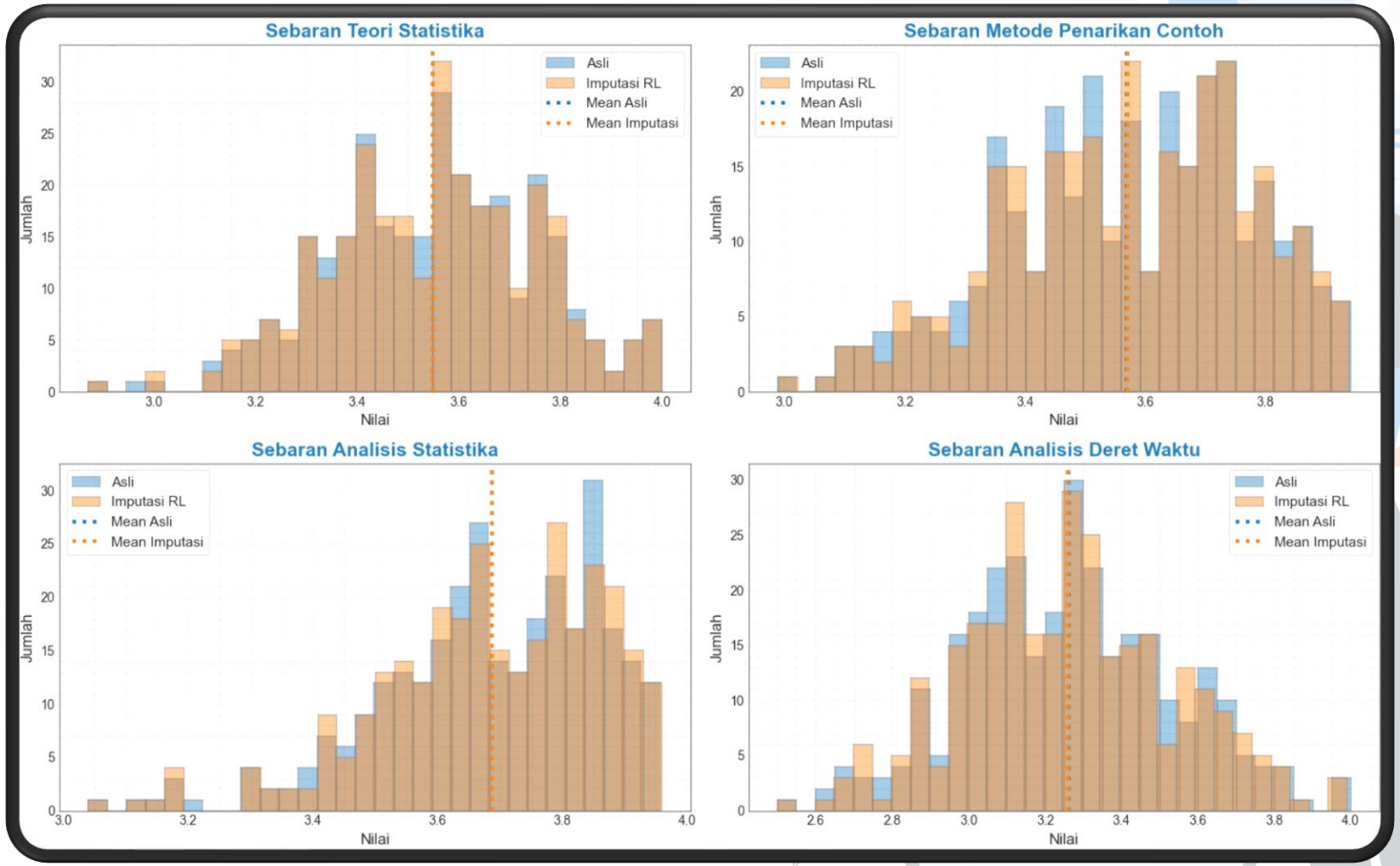
Plot sebaran memperlihatkan overlap sempurna antara data asli dan hasil imputasi. Metode ini mempertahankan bentuk sebaran, termasuk ekor dan puncak, tanpa distorsi. (Gambar 6)

Tahukah Kamu ?

Reinforcement Learning membentuk kebijakan imputasi berbasis umpan balik, mampu menangani sebaran multimodal, pencilan, dan pola unik tanpa preprocessing terpisah.

Imputasi Reinforcement Learn.

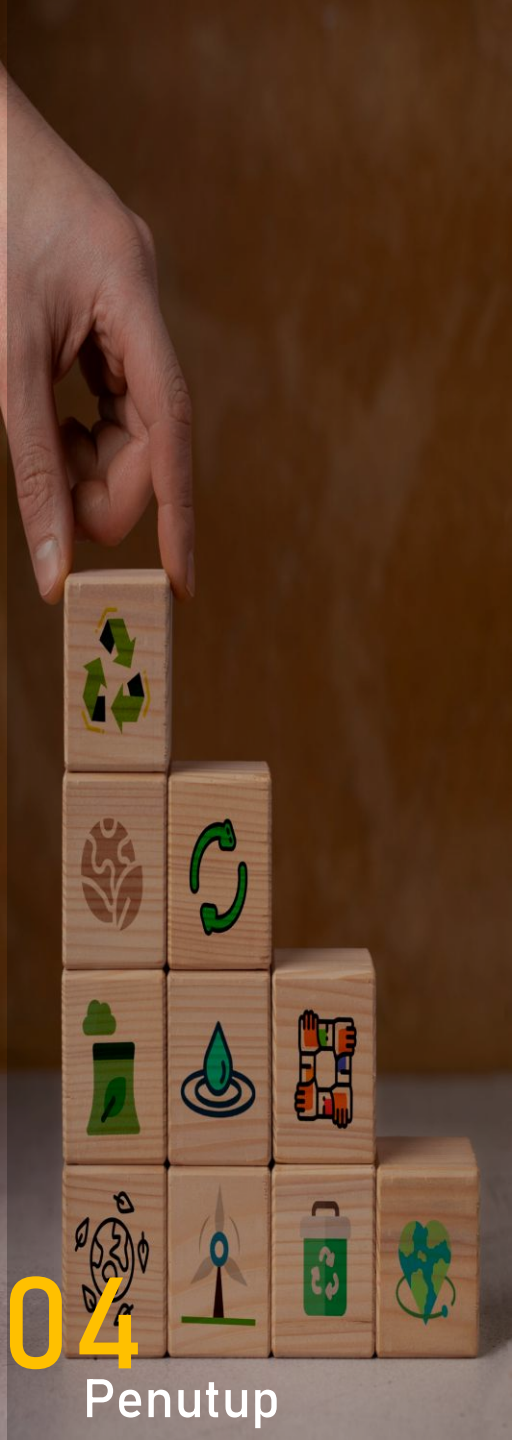
Gambar 6. Imputasi Reinfor. Learn



RL dapat diperbarui dengan data baru tanpa pelatihan ulang, cocok untuk dataset yang terus bertambah.

RL mempertimbangkan korelasi melalui umpan balik, nilai hilang diimputasi dengan mempertimbangkan nilai terkait di peubah lain

03
Imputasi R. Learn



Penutup Kajian

Simpulan Metode Imputasi

1. Pendekatan Statistika Dasar (Mean & Median), Metode ini sederhana, cepat, dan mudah diaplikasikan, namun tidak mempertimbangkan hubungan antar peubah atau pola kompleks; cocok untuk data yang relatif homogen.

2. Pendekatan Pembelajaran Mesin (KNN, Fuzzy K-Means, Isolation Forest), lebih adaptif terhadap struktur data dan pola multivariat, namun sensitif terhadap parameter (misal: jumlah tetangga, jumlah gerombol) dan bisa memerlukan preprocessing tambahan.

3. Pendekatan Penguatan Pembelajaran (*Reinforcement Learning*), memberikan hasil terbaik karena mampu menyesuaikan imputasi secara kontekstual berbasis reward – umpan balik dan sebaran data, meskipun lebih kompleks dan memerlukan proses pelatihan yang lebih lama.

Data Homogen, sederhana, ukuran tidak terlalu besar?

Gunakan imputasi statistika dan atau pembelajaran mesin.

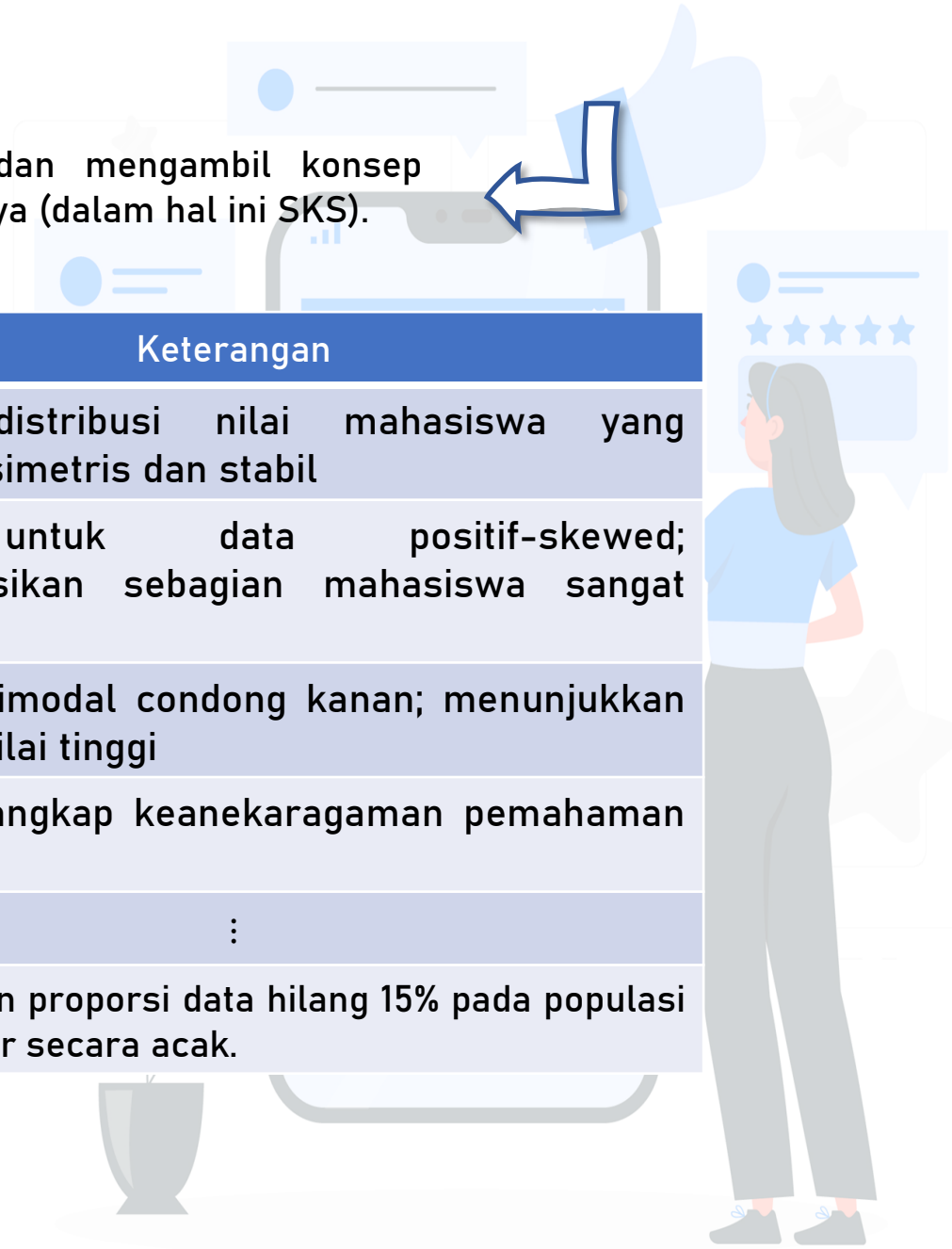
Data beragam, ukuran besar dan bersifat dinamis?

Gunakan imputasi pembelajaran mesin atau penguatan.

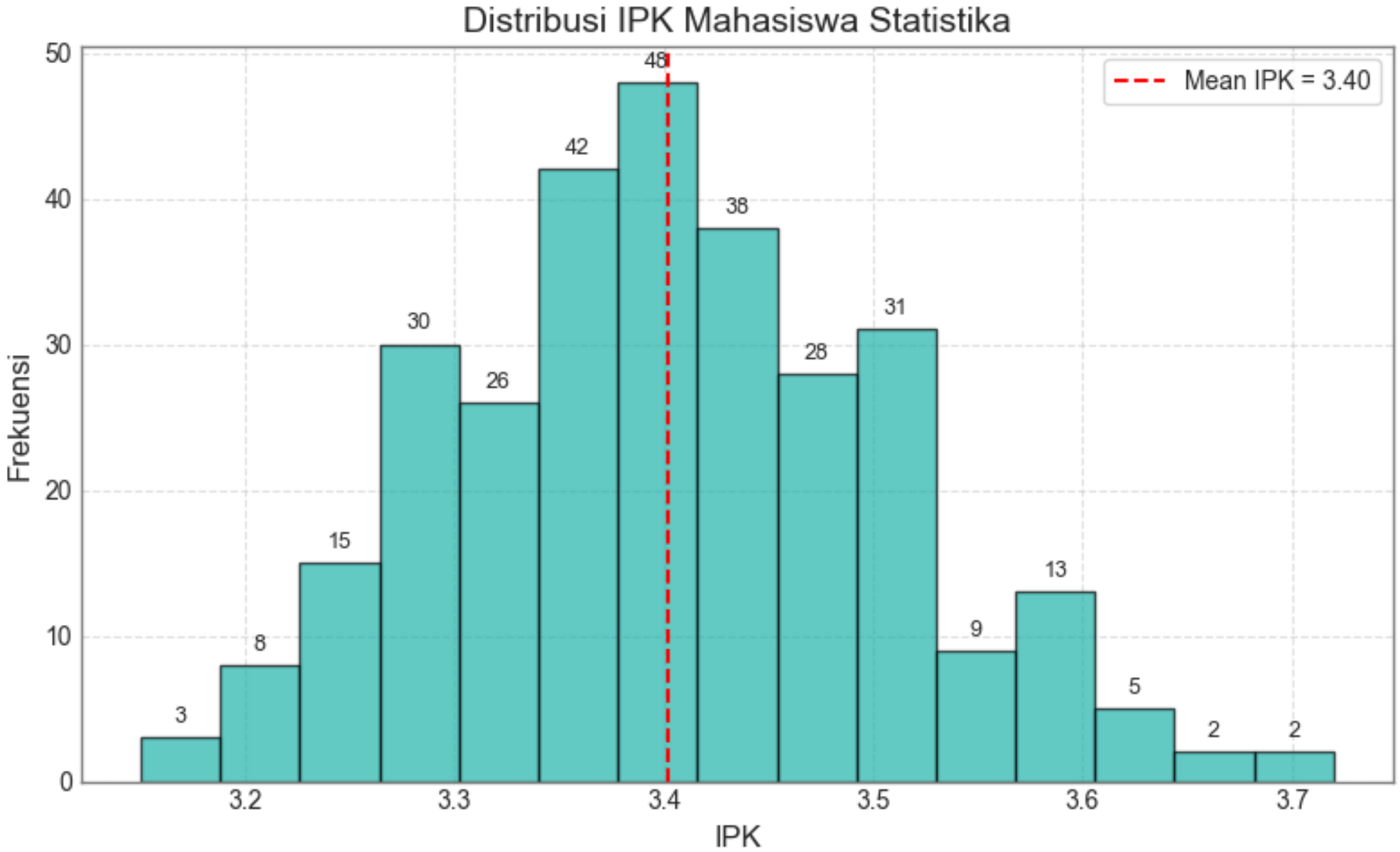
Lampiran Kajian 1

Pembangkitan data simulasi meniru suatu daerah Penelitian sosial dan mengambil konsep percontohan IPK yang dibentuk oleh nilai mata kuliah dan bobot penimbangnya (dalam hal ini SKS).

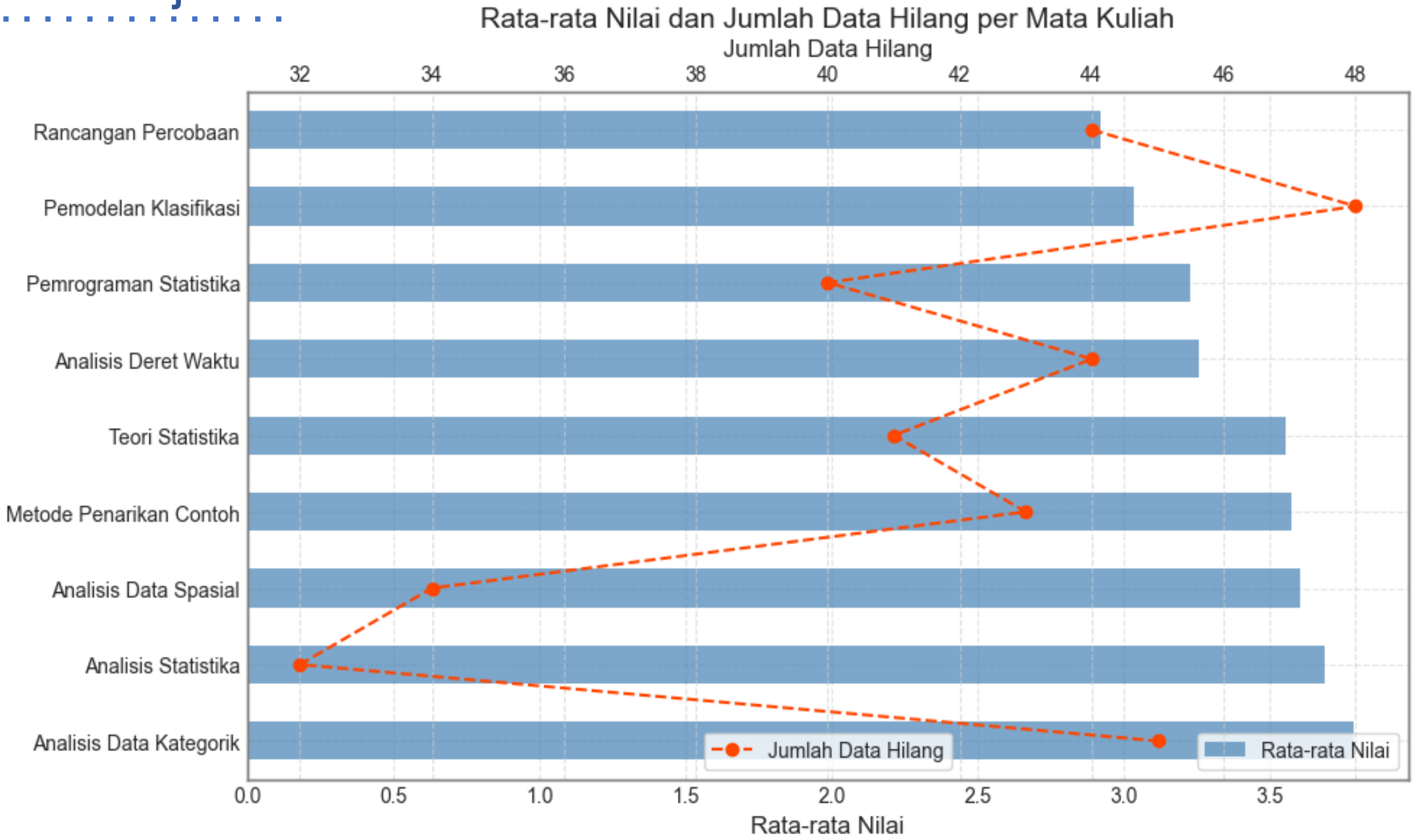
Komponen	Sebaran	Keterangan
Nilai Teori Statistika	Normal	Mewakili distribusi nilai mahasiswa yang cenderung simetris dan stabil
Nilai Pemrograman Statistika	Gamma	Cocok untuk data positif-skewed; mengindikasikan sebagian mahasiswa sangat unggul
Nilai Analisis Statistika	Beta	Sebaran unimodal condong kanan; menunjukkan mayoritas nilai tinggi
Nilai Metode Penarikan Contoh	Beta	Untuk menangkap keanekaragaman pemahaman sampling
⋮	⋮	⋮
Data Hilang	-	Menggunakan proporsi data hilang 15% pada populasi yang tersebar secara acak.

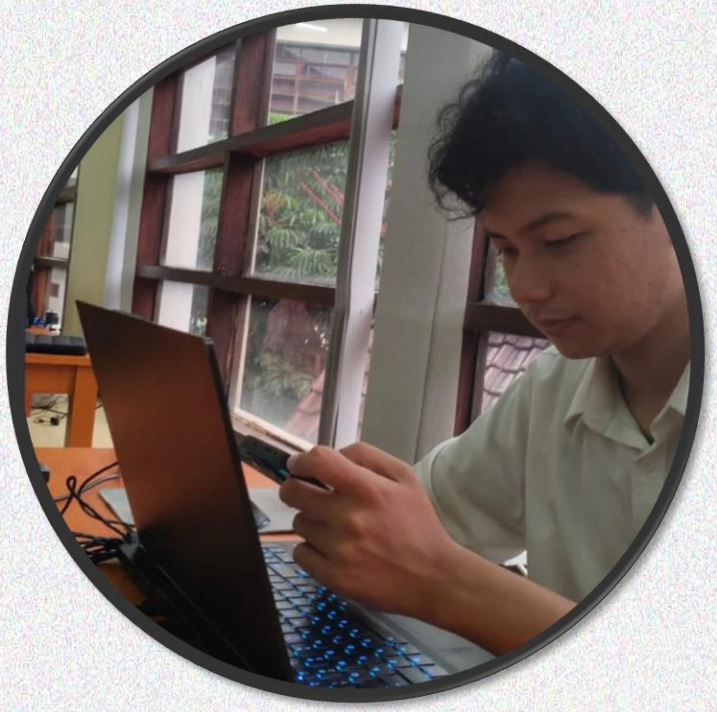


Lampiran Kajian 2



Lampiran Kajian 3





Ghardapaty Ghaly Ghiffary

Python | Excel | R | QGIS | Power BI

Ghardapaty Ghaly Ghiffary

@ghiffary_17

https://github.com/ghiffahry

ghiffaryankh@gmail.com

ghiffary_17

https://medium.com/@17611083