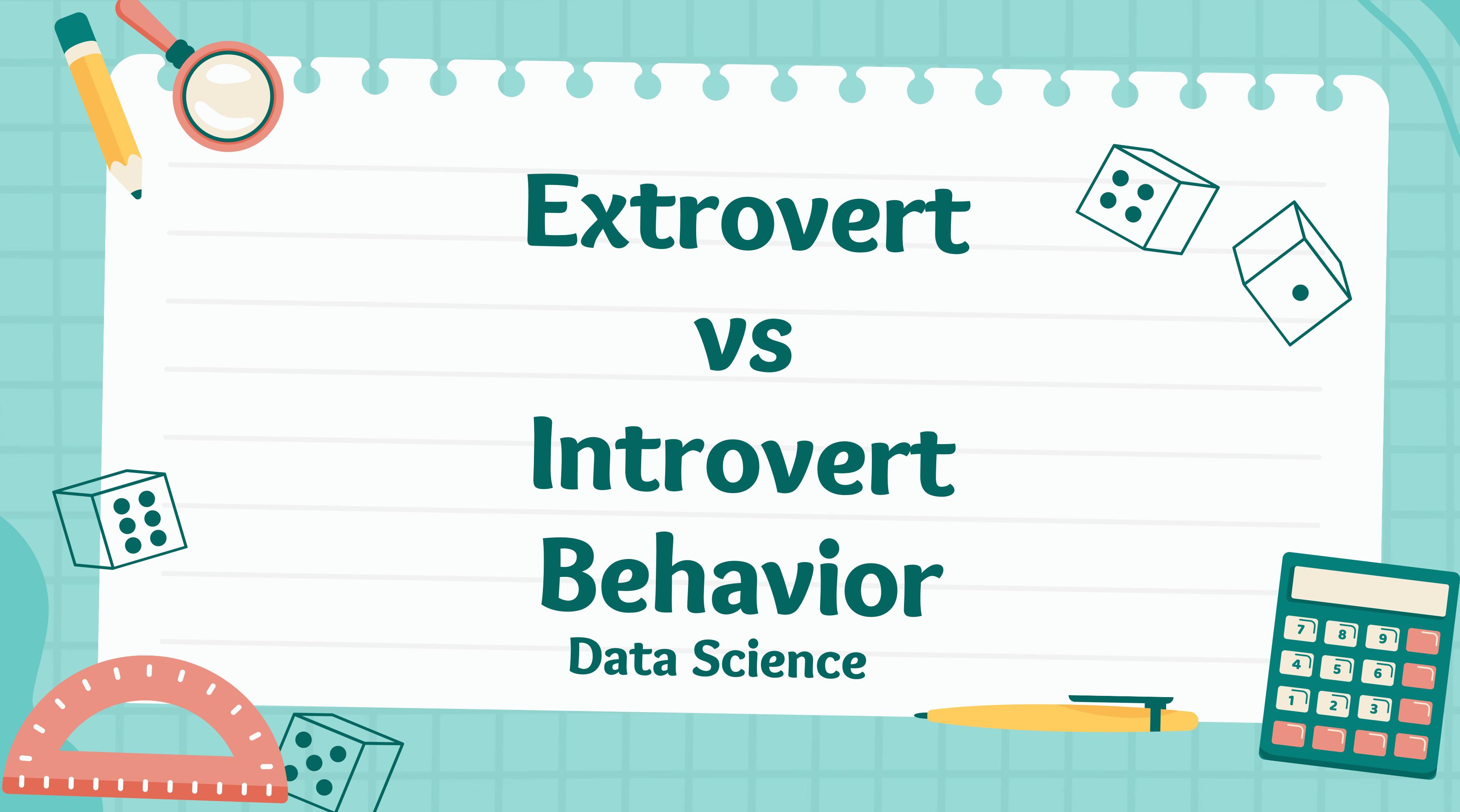


Extrovert vs Introvert Behavior Data Science





Kelompok 7



Muhammad Farkhan Ridho
41524010053



Rendy Herlandi
41524010058



Fabian Erik Rasyid Islami
41524010070

P1



Latar Belakang Masalah

Konteks industri/masyarakat

Dataset ini berada pada konteks psikologi dan sumber daya manusia (SDM), yang relevan untuk kebutuhan rekrutmen, pengembangan karyawan, konseling, dan analisis perilaku individu di masyarakat maupun industri.

Permasalahan utama

Permasalahan utamanya adalah mengidentifikasi atau memprediksi tipe/karakter kepribadian seseorang berdasarkan sejumlah fitur atau atribut (misalnya skor perilaku, kebiasaan, atau respons individu).



Latar Belakang Masalah

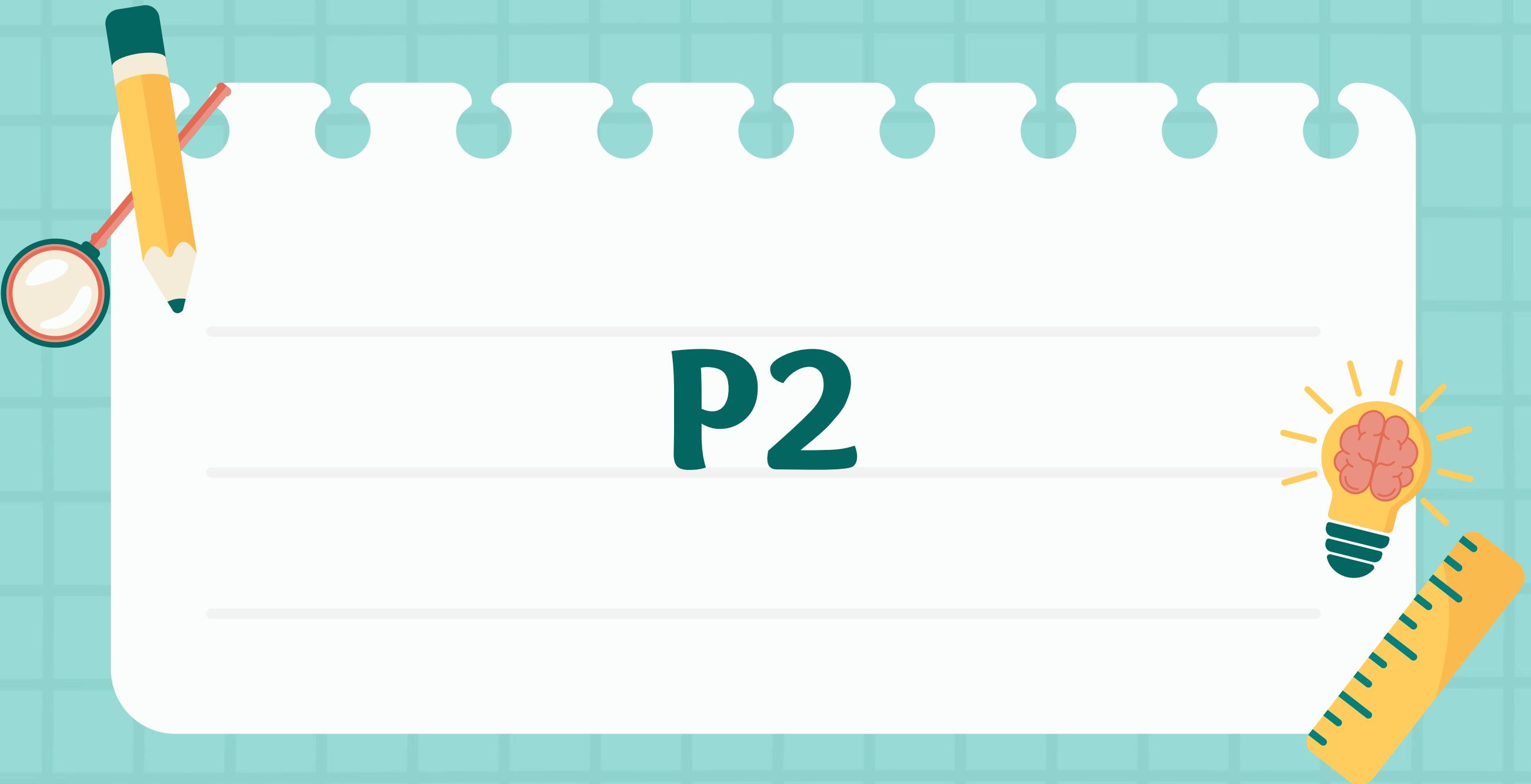
Tujuan analisis

Tujuan analisis adalah membangun model yang mampu memprediksi kepribadian individu secara akurat, sehingga dapat membantu pengambilan keputusan (misalnya penempatan kerja atau rekomendasi pengembangan diri).

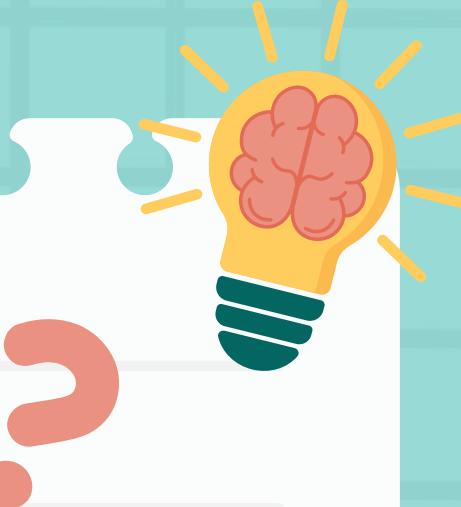
Jenis masalah

Masalah ini termasuk klasifikasi, karena variabel target berupa kategori/kelas kepribadian, bukan nilai numerik kontinu.

P2



Identifikasi dan Analisa Data



Tabel Dataset

Deskripsi Dataset Struktur Dataset dan Variabel dalam Dataset

	Time_spent_Alone	Stage_fear	Social_event_attendance	Going_outside	Drained_after_socializing	Friends_circle_size	Post_frequency	Personality
0	4.0	No	4.0	6.0	No	13.0	5.0	Extrovert
1	9.0	Yes	0.0	0.0	Yes	0.0	3.0	Introvert
2	9.0	Yes	1.0	2.0	Yes	5.0	2.0	Introvert
3	0.0	No	6.0	7.0	No	14.0	8.0	Extrovert
4	3.0	No	9.0	4.0	No	8.0	5.0	Extrovert
...
2895	3.0	No	7.0	6.0	No	6.0	6.0	Extrovert
2896	3.0	No	8.0	3.0	No	14.0	9.0	Extrovert
2897	4.0	Yes	1.0	1.0	Yes	4.0	0.0	Introvert
2898	11.0	Yes	1.0	NaN	Yes	2.0	0.0	Introvert
2899	3.0	No	6.0	6.0	No	6.0	9.0	Extrovert

2900 rows × 8 columns

Identifikasi dan Analisis Data



Before

	0	Time_spent_Alone	Social_event_attendance	Going_outside	Friends_circle_size	Post_frequency
Time_spent_Alone	63	count	2837.000000	2838.000000	2834.000000	2823.000000
Stage_fear	73	mean	4.505816	3.963354	3.000000	6.268863
Social_event_attendance	62	std	3.479192	2.903827	2.247327	4.289693
Going_outside	66	min	0.000000	0.000000	0.000000	0.000000
Drained_after_socializing	52	25%	2.000000	2.000000	1.000000	3.000000
Friends_circle_size	77	50%	4.000000	3.000000	3.000000	5.000000
Post_frequency	65	75%	8.000000	6.000000	5.000000	10.000000
Personality	0	max	11.000000	10.000000	7.000000	15.000000

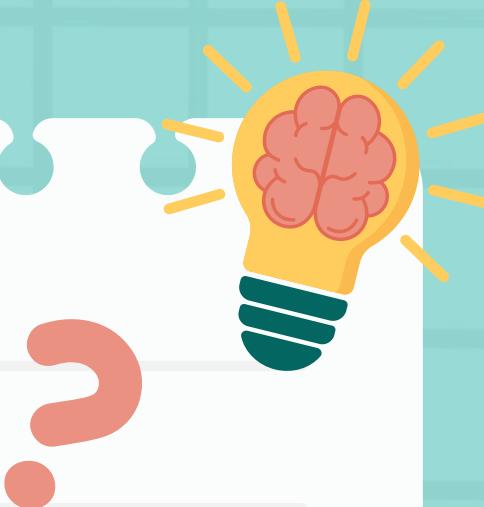
Identifikasi dan Analisis data

After

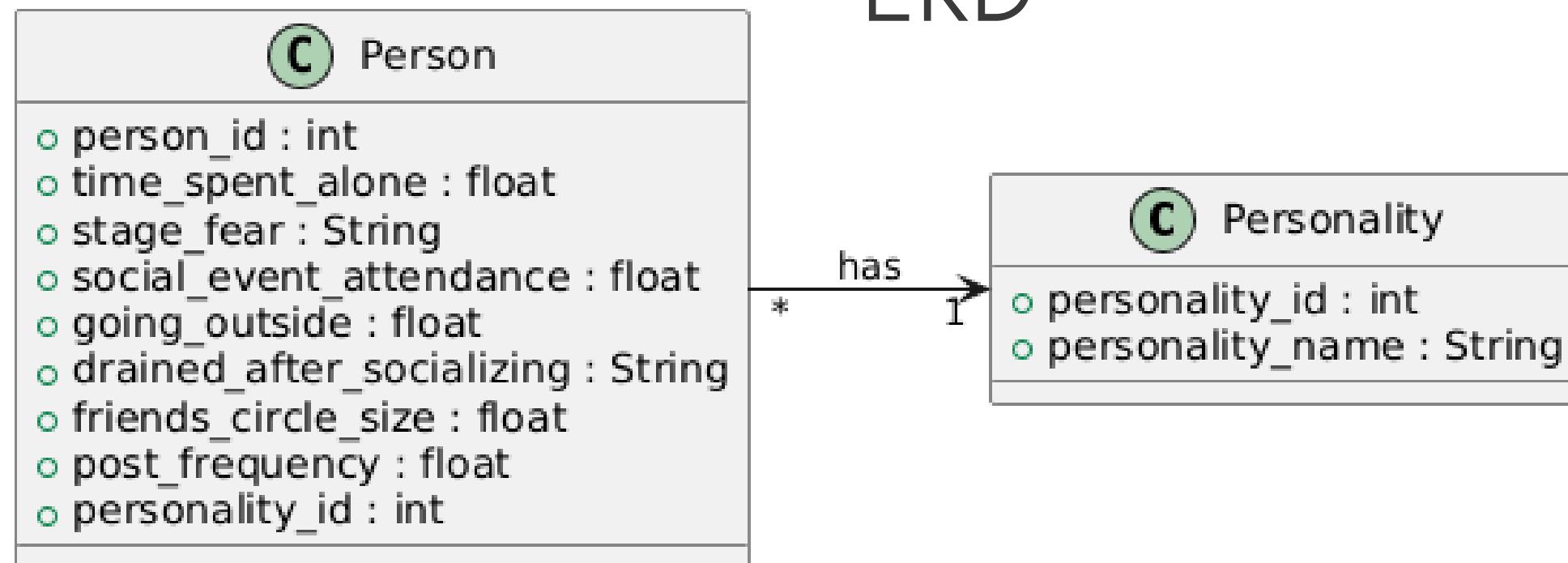


	0	Time_spent_Alone	Social_event_attendance	Going_outside	Friends_circle_size	Post_frequency
Time_spent_Alone	0	count	2900.000000	2900.000000	2900.000000	2900.000000
Stage_fear	0	mean	4.494828	3.942759	3.000000	6.235172
Social_event_attendance	0	std	3.441971	2.875987	2.221597	4.237255
Going_outside	0	min	0.000000	0.000000	0.000000	0.000000
Drained_after_socializing	0	25%	2.000000	2.000000	1.000000	3.000000
Friends_circle_size	0	50%	4.000000	3.000000	3.000000	5.000000
Post_frequency	0	75%	7.000000	6.000000	5.000000	10.000000
Personality	0	max	11.000000	10.000000	7.000000	15.000000

Identifikasi dan Analisa Data



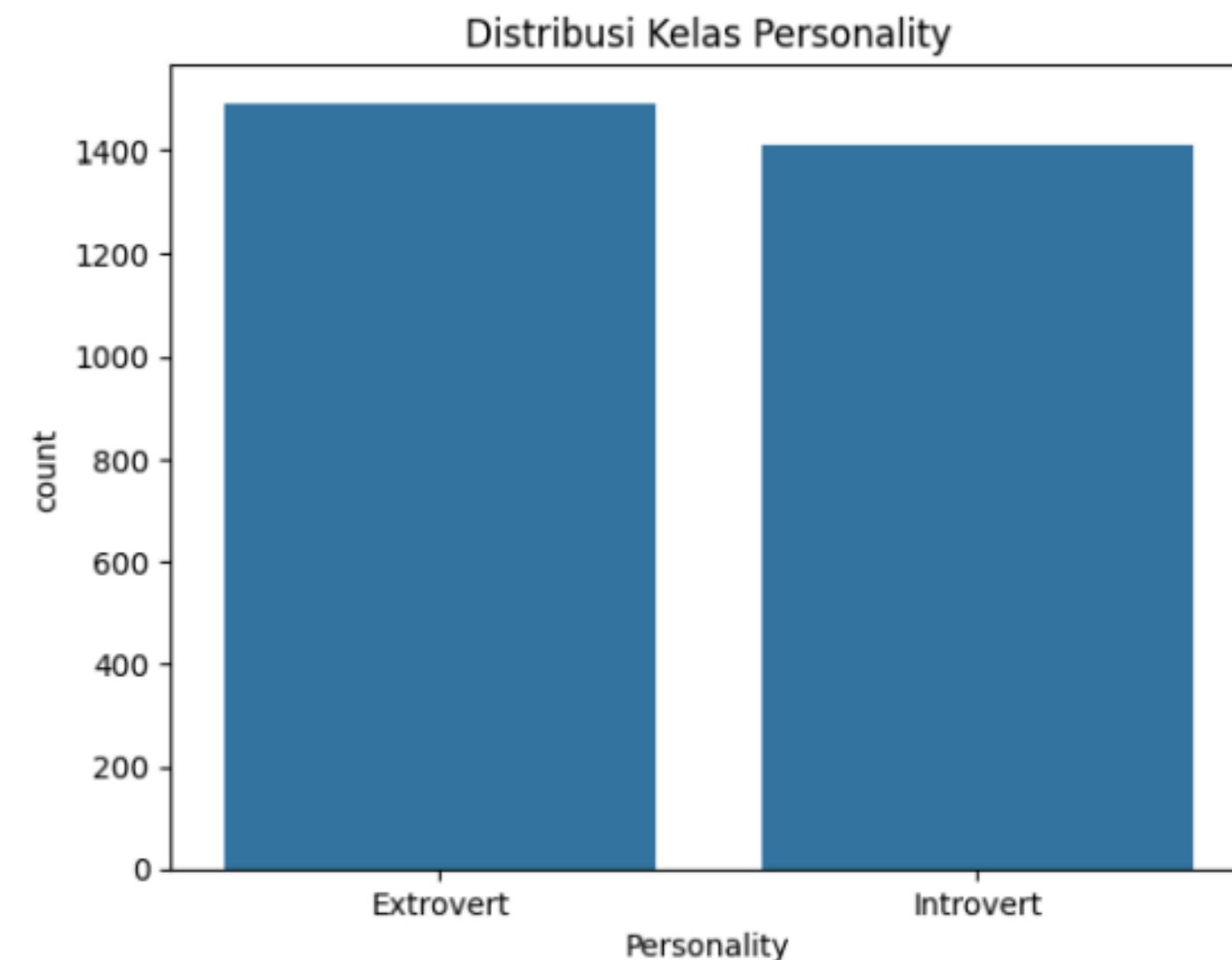
ERD



ERD ini terdiri dari dua entitas, yaitu Person dan Personality. Entitas Person menyimpan data perilaku sosial individu, sedangkan Personality menyimpan tipe kepribadian (Extrovert/Introvert). Relasinya bersifat many-to-one, di mana banyak Person memiliki satu Personality, sehingga struktur ini mendukung analisis klasifikasi kepribadian secara terorganisir.

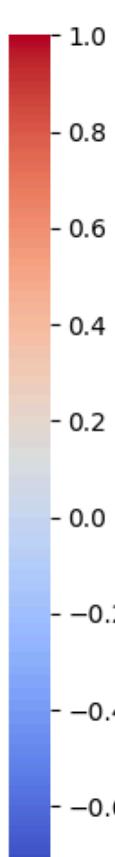
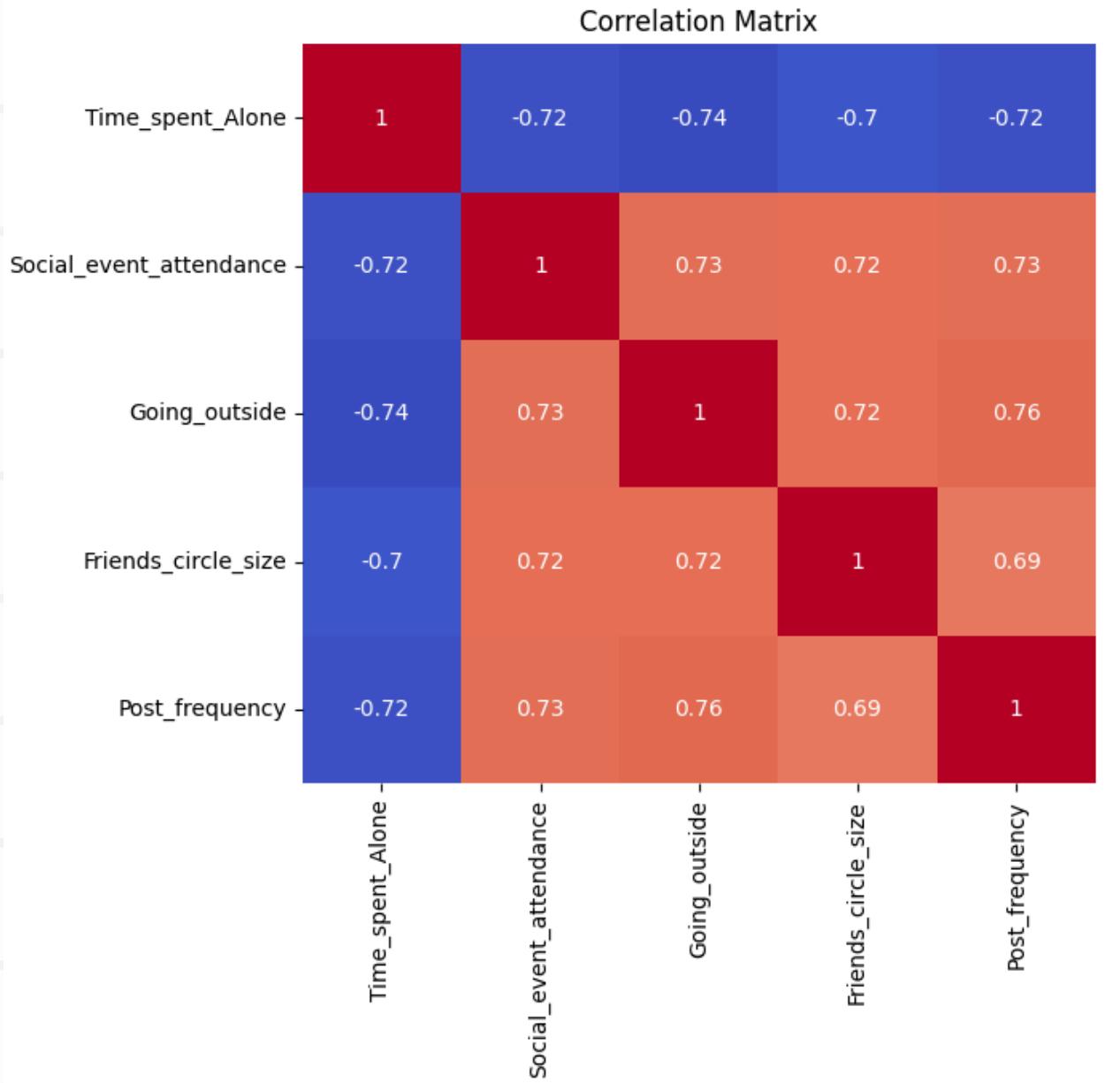
Identifikasi dan Analisa Data

Exploratory Data Analysis



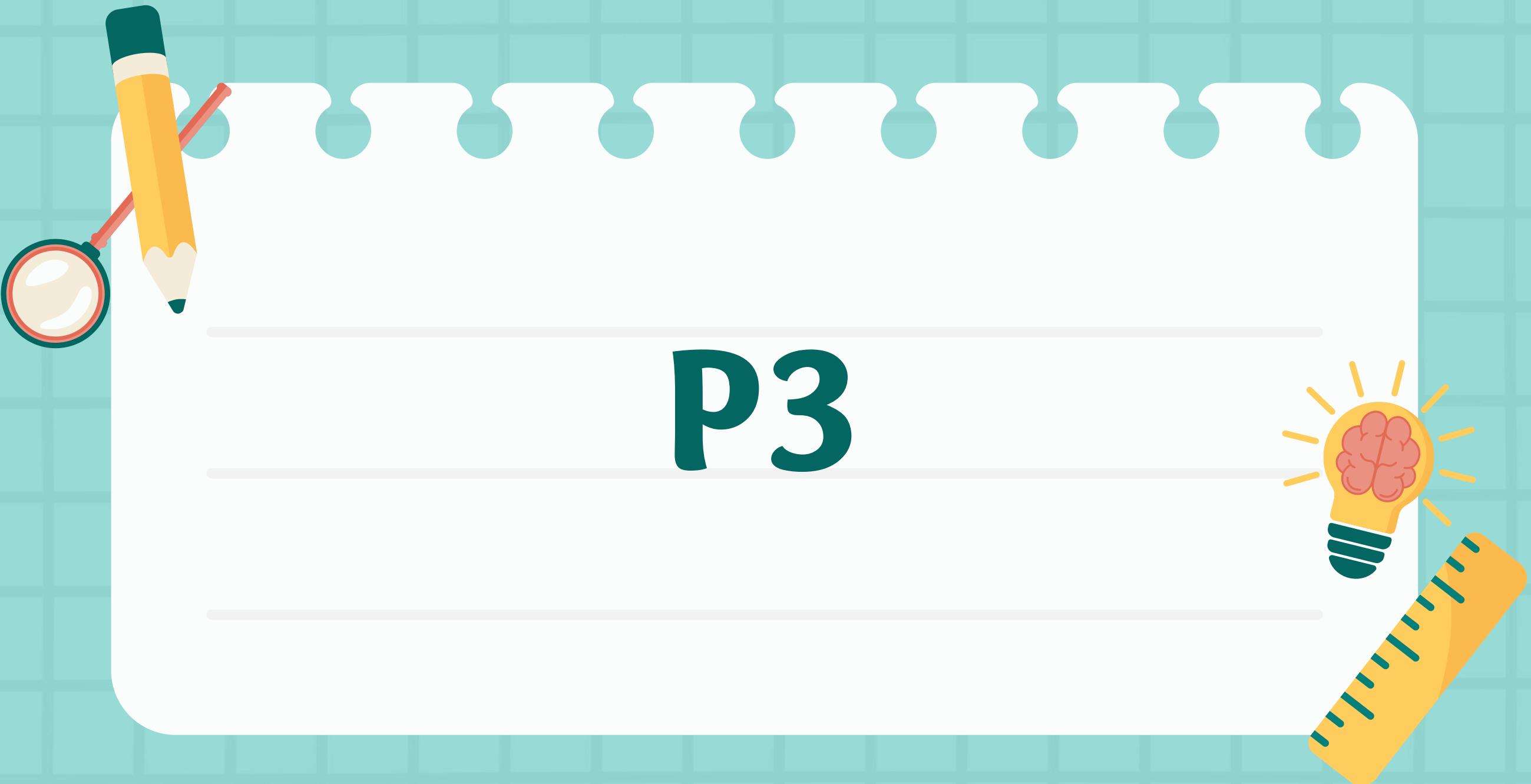
Grafik Distribusi Kelas Personality menunjukkan bahwa jumlah data Extrovert dan Introvert relatif seimbang, dengan Extrovert sedikit lebih banyak dibandingkan Introvert. Kondisi ini baik untuk analisis klasifikasi karena mengurangi risiko bias model terhadap salah satu kelas dan membantu menghasilkan performa prediksi yang lebih stabil dan akurat.

Identifikasi dan Analisa Data



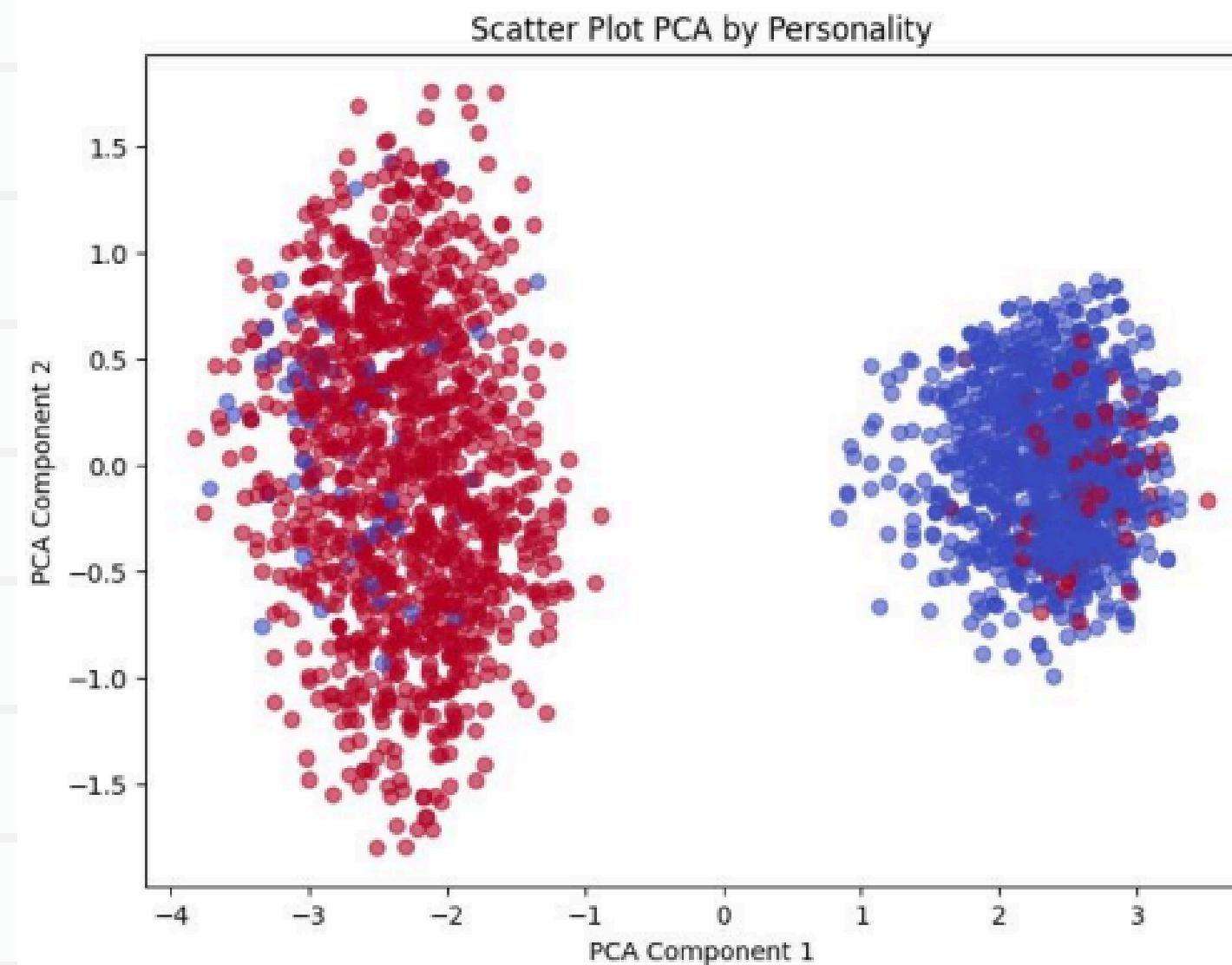
Matriks korelasi ini menunjukkan bahwa Time_spent_Alone memiliki korelasi negatif yang kuat dengan variabel aktivitas sosial lainnya, artinya semakin sering seseorang menyendiri, semakin rendah tingkat interaksi sosialnya. Sebaliknya, Social_event_attendance, Going_outside, Friends_circle_size, dan Post_frequency saling berkorelasi positif, menandakan pola perilaku sosial yang konsisten dan menjadi indikator penting dalam membedakan tipe kepribadian.

P3



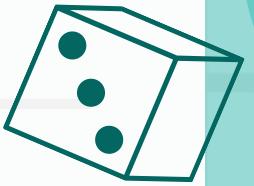


Visualisasi Model

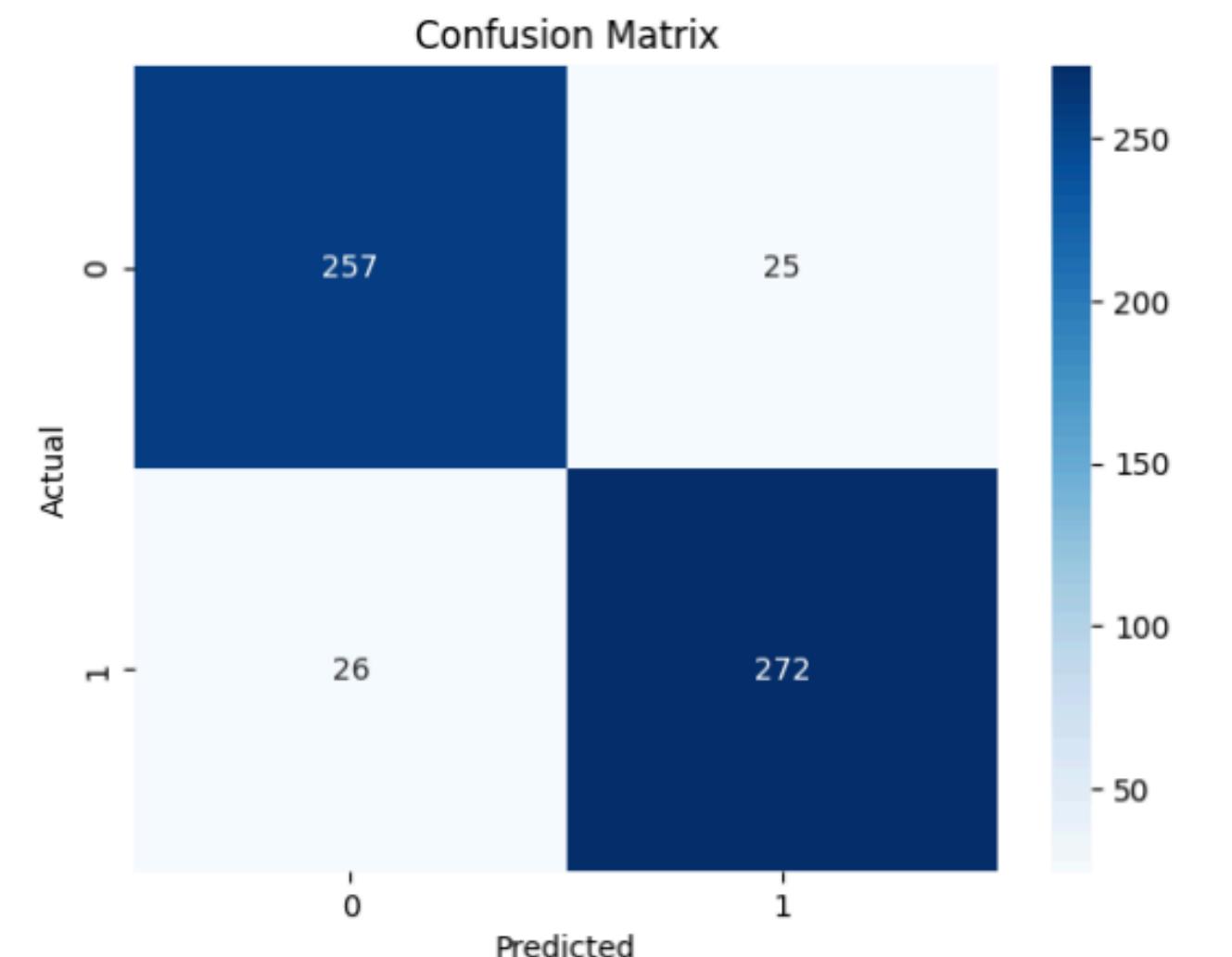


Berdasarkan visualisasi PCA pada dataset personality, terlihat dua klaster yang cukup jelas antara kelas Introvert dan Extrovert, terutama pada komponen PCA pertama. Hal ini menunjukkan bahwa fitur-fitur dalam dataset mampu membedakan kedua kelas dengan baik, sehingga mendukung performa model Regresi Logistik yang telah dibangun.





Confusion Matrix

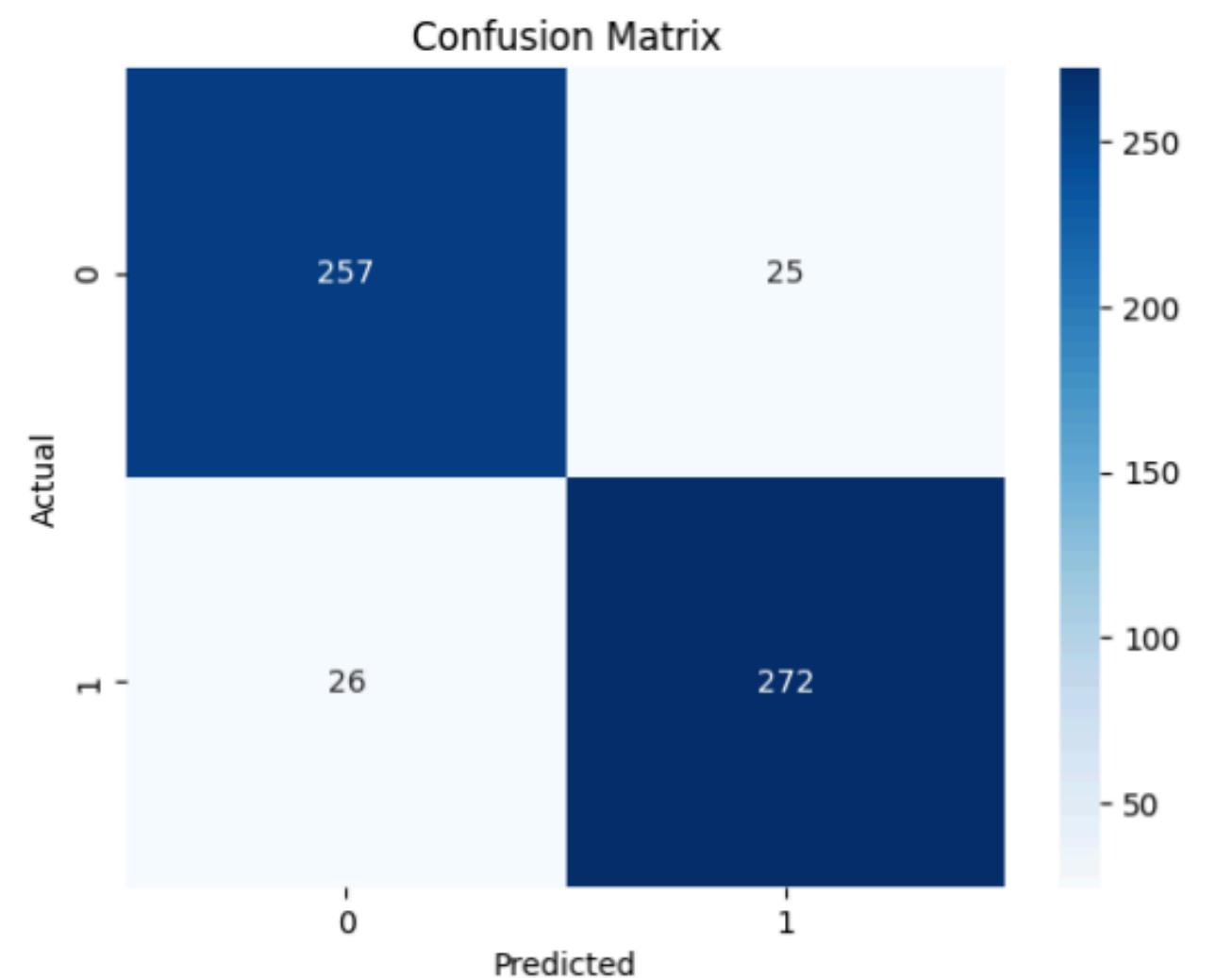
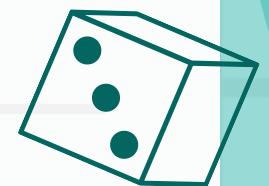


True Negative (TN): 257
False Positive (FP): 25
False Negative (FN): 26
True Positive (TP): 272





Confusion Matrix



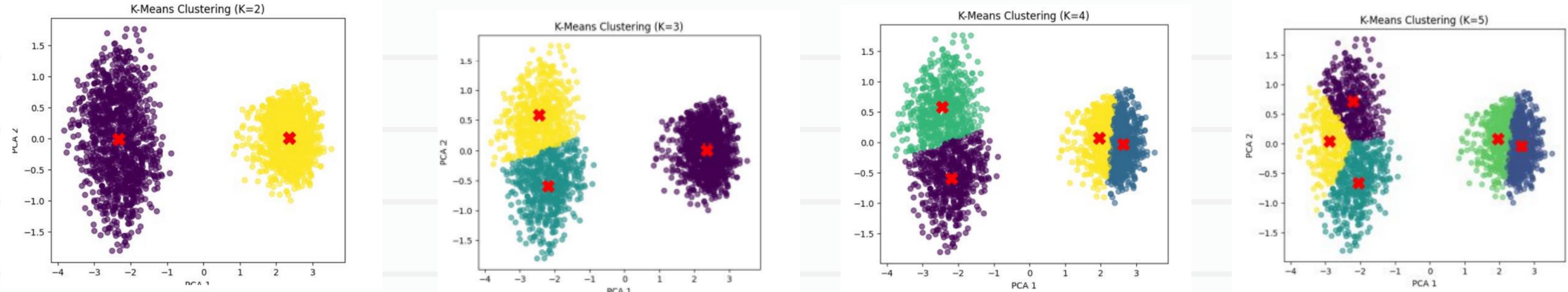
Hasil Evaluasi:

Accuracy : 0.9103448275862069
Precision: 0.9183673469387755
Recall : 0.9060402684563759
F1-score : 0.9121621621621622
AUC : 0.9474808415441001

P4

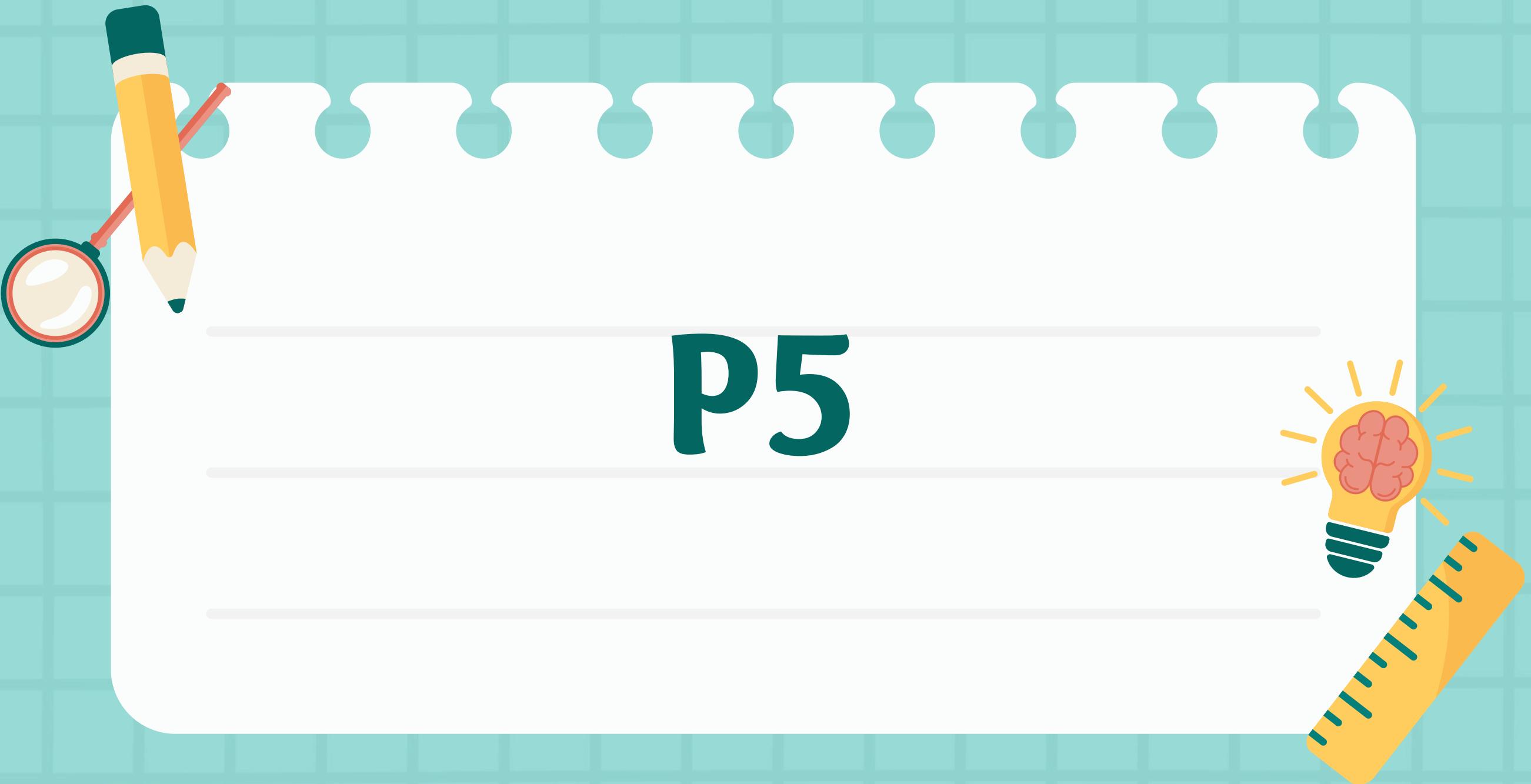


Clustering K-Means



Visualisasi K-Means ini menunjukkan pembagian data ke dalam beberapa segmen dengan tingkat detail berbeda, di mana secara bisnis $K = 4$ adalah yang paling ideal. Pada $K=2$ segmentasi masih terlalu kasar, sedangkan $K=3$ mulai memberi gambaran strategis, namun $K=4$ menghasilkan cluster yang jelas dan mudah ditindaklanjuti, misalnya membedakan pelanggan paling bernilai, pelanggan aktif, pelanggan sesekali, dan pelanggan pasif, sehingga tiap segmen bisa diberi strategi berbeda seperti loyalty, upselling, promo, atau reaktivasi. Sementara $K=5$ memang lebih detail, tetapi cenderung terlalu kompleks untuk operasional, sehingga $K=4$ menjadi pilihan paling seimbang antara insight dan kemudahan penerapan bisnis.

P5



Bandingkan Performa Tabel

	Model	Accuracy	Precision	Recall	F1-score	AUC
0	Logistic Regression	0.912069	0.915825	0.912752	0.914286	0.921242
1	Random Forest	0.910345	0.918367	0.906040	0.912162	0.947481

- Logistic Regression memiliki accuracy (91,21%), recall (91,28%), dan F1-score (91,43%) yang sedikit lebih tinggi.
- Random Forest memiliki precision (91,84%) dan AUC (94,75%) yang lebih tinggi.
- Secara keseluruhan, Logistic Regression unggul dari sisi kestabilan dan interpretabilitas, sedangkan Random Forest unggul dalam kemampuan diskriminasi kelas. Pemilihan model terbaik bergantung pada tujuan: Logistic Regression untuk analisis yang sederhana dan seimbang, dan Random Forest untuk fokus pada pemisahan kelas yang lebih kuat.





Keterbatasan dan Peluang Pengembangan

Kelebihan: Dataset yang seimbang antara kedua kelas (Introvert dan Extrovert)

Keterbatasan: Terdapat missing values pada beberapa variabel yang memerlukan imputasi Dataset.

Peluang Pengembangan Penelitian : Penambahan jumlah dan variasi data agar model lebih general, perbaikan penanganan missing values dengan metode imputasi yang lebih baik, serta pengayaan variabel biner menjadi skala yang lebih detail untuk menangkap nuansa perilaku. Selain itu, kinerja dapat ditingkatkan melalui penambahan fitur baru dan penggunaan model yang lebih canggih, serta penerapan model pada aplikasi nyata seperti personalisasi layanan dengan tetap memperhatikan aspek etika dan privasi.





Terima Kasih

