EXTROVERT VS INTROVERT BEHAVIOR DATASET

Disusun oleh kelompok 7:

1. Fabian Erik Rasyid Islami (41524010070)
2. Rendy Herlandi (41524010058)
3. Muhammad Farkhan Ridho (41524010053)

Dosen pengampu :
Inna Sabily Karima, S.Kom, M.Kom

UNIVERSITAS MERCU BUANA

FAKULTAS ILMU KOMPUTER

PROGRAM STUDI TEKNIK INFORMATIKA

2025

Pertanyaan 1 (Sub CPMK 61.1 Sub CPMK
62.1) : Jelaskan secara ringkas: konteks industri/masyarakat dataset, permasalahan utama, serta tujuan analisis dan jenis masalah (klasifikasi/regresi).

Pertanyaan 2 (Sub CPMK 43.1) : Identifikasi entitas, atribut, dan variabel target; buat ERD/skema logis; jelaskan pipeline data (akuisisi → cleaning → EDA → pemodelan → evaluasi).

Pertanyaan 3 (Sub CPMK 43.1 Sub CPMK
62.1) : Lakukan preprocessing; bangun Regresi Logistik; train–test split; evaluasi dengan Confusion Matrix, Precision, Recall, F1-score, AUC

Pertanyaan 4 (Sub CPMK 43.1) : Terapkan K-Means (K=2–5); lakukan PCA (2 komponen); visualisasikan dan jelaskan makna cluster secara bisnis/industri.

Pertanyaan 5 (Sub CPMK
62.1) : Bandingkan performa model, tentukan model terbaik beserta alasannya, dan jelaskan keterbatasan serta peluang pengembangan

Pertanyaan 6 (Sub CPMK 71.1 Sub CPMK
72.1) : Jelaskan pembagian peran tim, sertakan bukti kolaborasi (tabel/log GitHub), dan susun laporan ilmiah & slide presentasi yang ringkas dan berbasis data.

Jawaban

Nama Dataset	Sumber Data set	Jenis Data	Format Data	Jumlah Record/ Ukuran File	Daftar Atribut	Potensi Analisis
Extrovert vs Introvert Behavior	Kaggle	Klasifikasi	CSV	2900 record, ukuran file 181.4 KB++	Time_spent_Alone, Stage_fear, Social_event_attendance, Going_outside ,Drained_after_socializing, Friends_circle, _size Post_frequency,Personality	

P1:

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. Dalam era digital, pemahaman tipe kepribadian menjadi penting untuk:

- **Rekrutmen & penempatan kerja** yang sesuai dengan karakteristik tim
- **Pengembangan program pelatihan** yang dipersonalisasi
- **Konseling dan layanan kesehatan mental** berbasis data
- **Analisis perilaku konsumen** untuk strategi pemasaran

Permasalahan Utama

Permasalahan utamanya adalah **mengidentifikasi atau memprediksi tipe/karakter kepribadian seseorang** (Extrovert atau Introvert) berdasarkan sejumlah fitur atau atribut perilaku sosial, seperti:

- Waktu yang dihabiskan sendirian (Time_spent_Alone)
- Tingkat kehadiran dalam acara sosial (Social_event_attendance)
- Frekuensi aktivitas di luar ruangan (Going_outside)
- Ukuran lingkaran pertemanan (Friends_circle_size)
- Respons emosional setelah bersosialisasi (Drained_after_socializing)

Tantangannya adalah bagaimana memanfaatkan data perilaku yang terukur untuk membuat prediksi kepribadian yang akurat dan dapat diandalkan dalam pengambilan keputusan.

Tujuan Analisis

Tujuan analisis adalah **membangun model yang mampu memprediksi kepribadian individu secara akurat**, sehingga dapat membantu pengambilan keputusan dalam berbagai konteks:

- **Penempatan kerja** yang optimal berdasarkan kesesuaian kepribadian dengan kultur tim
- **Rekomendasi pengembangan diri** yang disesuaikan dengan tipe kepribadian
- **Personalisasi layanan konseling** untuk efektivitas intervensi psikologis
- **Optimasi strategi komunikasi** dalam manajemen SDM

Jenis Masalah

Masalah ini termasuk **klasifikasi biner**, karena variabel target berupa **kategori/kelas kepribadian** (Extrovert atau Introvert), bukan nilai numerik kontinu seperti pada masalah regresi. Model akan memprediksi kelas diskrit berdasarkan pola perilaku dari fitur-fitur input

P2 :

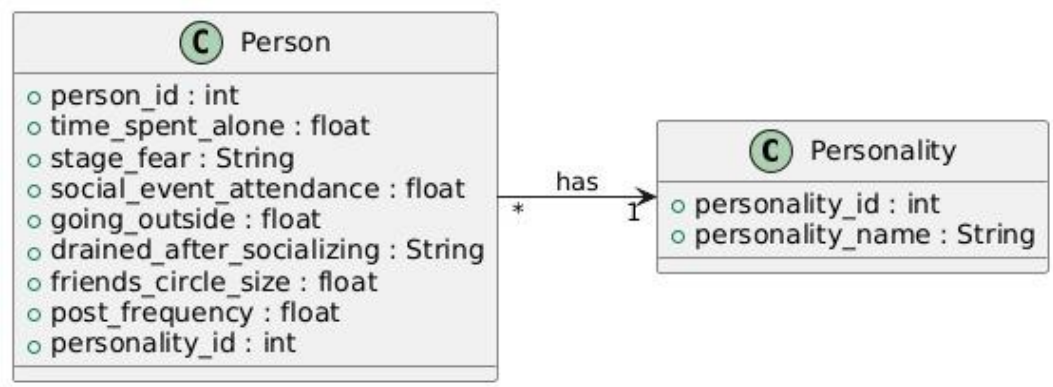
A. Identifikasi Entitas, Atribut, dan Variabel Target

Entitas Utama:
Responden
Individu yang diamati berdasarkan perilaku sosial dan kepribadian.

Atribut:

(Fitur/X):
Waktu Menyendiri: Time_spent_Alone
Kehadiran Acara Sosial: Social_event_attendance
Aktivitas di Luar Rumah: Going_outside
Ukuran Lingkar Pertemanan: Friends_circle_size
Frekuensi Posting Media Sosial: Post_frequency
Ketakutan Tampil di Publik: Stage_fear
Kelelahan Setelah Bersosialisasi: Drained_after_socializing

Variabel Target (y):
Personality
Kategori kepribadian Introvert dan Extrovert.



Pipeline Analisis Kepribadian (Extrovert vs Introvert)

1. Data Acquisition (Akuisisi Data)
Proses: Mengambil dataset perilaku sosial individu untuk analisis kepribadian.
Input: Dataset mentah berisi fitur Time_spent_Alone, Social_event_attendance, Going_outside, Friends_circle_size, Post_frequency, Stage_fear, dan Drained_after_socializing.

2. Data Cleaning & Selection (Pembersihan)
Missing Value Handling: Mengisi nilai kosong menggunakan median (numerik) dan modus (kategorikal).

Encoding: Mengubah data kategorikal (Yes/No, Introvert/Extrovert) menjadi format numerik.

3. EDA – Exploratory Data Analysis (Eksplorasi)
Korelasi: Heatmap untuk melihat hubungan antar perilaku sosial.
Analisis Distribusi: Bar chart untuk mengecek keseimbangan kelas kepribadian.
Visualisasi: Scatter plot dan PCA untuk melihat pemisahan Introvert dan Extrovert.

4. Modeling (Pemodelan)
Split Data: 80% training dan 20% testing.
Model: Logistic Regression dan Random Forest Classifier.

5. Evaluation (Evaluasi & Kesimpulan)
Confusion Matrix: Analisis kesalahan prediksi.
Metrik Evaluasi: Accuracy, Precision, Recall, F1-Score, dan AUC.
Kesimpulan: Model mampu memprediksi tipe kepribadian dengan akurasi di atas 91%.

P3

Preprocessing missing value

Before:

```
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Time_spent_Alone           2837 non-null   float64
 1   Stage_fear                 2827 non-null   object
 2   Social_event_attendance    2838 non-null   float64
 3   Going_outside              2834 non-null   float64
 4   Drained_after_socializing  2848 non-null   object
 5   Friends_circle_size        2823 non-null   float64
 6   Post_frequency             2835 non-null   float64
 7   Personality                2900 non-null   object
dtypes: float64(5), object(3)
memory usage: 181.4+ KB
```

After :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2900 entries, 0 to 2899
Data columns (total 8 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Time_spent_Alone           2900 non-null   float64
 1   Stage_fear                 2900 non-null   object
 2   Social_event_attendance    2900 non-null   float64
 3   Going_outside              2900 non-null   float64
 4   Drained_after_socializing  2900 non-null   object
 5   Friends_circle_size        2900 non-null   float64
 6   Post_frequency             2900 non-null   float64
 7   Personality                2900 non-null   object
dtypes: float64(5), object(3)
memory usage: 181.4+ KB
```

Data sangat bersih dengan 0 missing value

In [83]:
```
X = df.drop(columns=[target])
y = df[target]

y = y.map({'Introvert': 0, 'Extrovert': 1})
```

In [84]:
```
X = pd.get_dummies(X, drop_first=True)
```

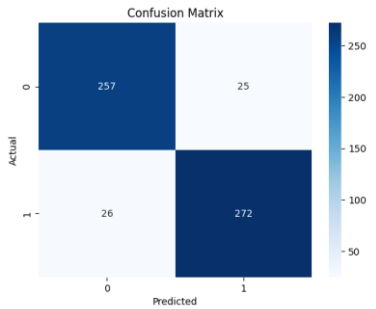
In [85]:
```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.2,
    random_state=42,
    stratify=y
)
```

In [86]:
```
from sklearn.preprocessing import StandardScaler

scaler = StandardScaler()

X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled  = scaler.transform(X_test)
```

In [87]:
```
from sklearn.linear_model import LogisticRegression

log_model = LogisticRegression(max_iter=1000)
log_model.fit(X_train_scaled, y_train)
```

Out[87]: LogisticRegression(max_iter=1000)

Berdasarkan tahapan yang dilakukan, data terlebih dahulu dipreprocessing melalui pemisahan fitur dan target, encoding data kategorikal, serta standardisasi fitur. Selanjutnya, data dibagi menjadi data latih dan data uji menggunakan metode train–test split untuk menghindari overfitting. Model Regresi Logistik kemudian dilatih menggunakan data latih yang telah distandardisasi. Tahapan ini memastikan model mampu melakukan klasifikasi secara optimal dan dapat dievaluasi secara objektif menggunakan data uji.

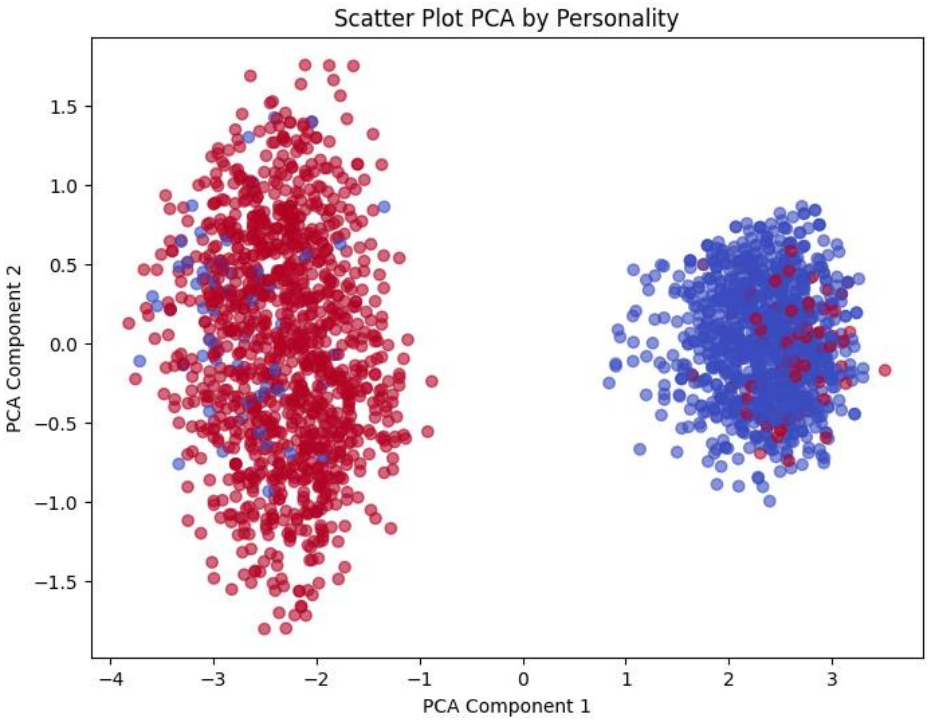


Confusion matrix menunjukkan bahwa model berhasil mengklasifikasikan sebagian besar data dengan benar, ditandai oleh nilai **True Negative (257)** dan **True Positive (272)** yang tinggi. Kesalahan prediksi relatif kecil, yaitu **False Positive (25)** dan **False Negative (26)**. Hal ini menandakan bahwa model Regresi Logistik memiliki kinerja klasifikasi yang baik dan seimbang pada kedua kelas.

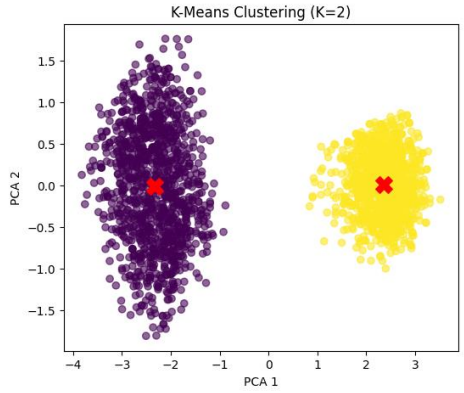
```
Accuracy : 0.9120689655172414
Precision: 0.9158249158249159
Recall   : 0.912751677852349
F1-score : 0.9142857142857143
AUC      : 0.9212420867247371
```

Berdasarkan hasil evaluasi pada dataset *personality*, model Regresi Logistik menunjukkan performa yang baik. Nilai **accuracy sebesar 91,21%** menunjukkan bahwa mayoritas data dapat diklasifikasikan dengan benar. **Precision sebesar 91,58%** menandakan ketepatan model dalam memprediksi kelas positif, sementara **recall sebesar 91,28%** menunjukkan kemampuan model dalam mengenali data positif dengan baik. Nilai **F1-score sebesar 91,43%** mengindikasikan keseimbangan yang baik antara precision dan recall. Selain itu, **AUC sebesar 92,12%** menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan kedua kelas kepribadian.

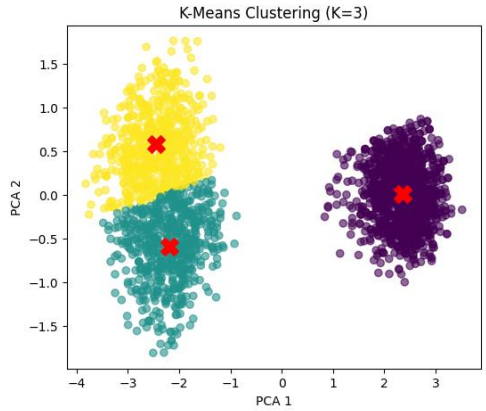
P4:



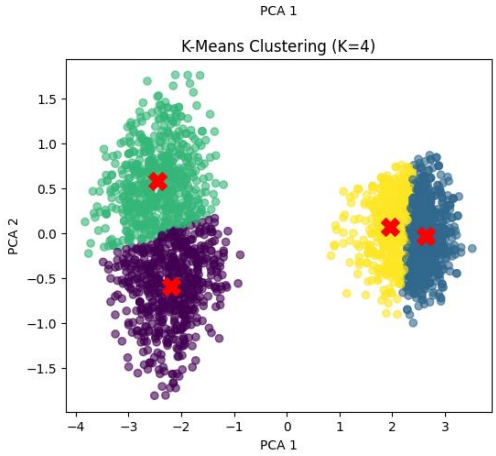
Berdasarkan visualisasi PCA pada dataset *personality*, terlihat dua klaster yang cukup jelas antara kelas *Introvert* dan *Extrovert*, terutama pada komponen PCA pertama. Hal ini menunjukkan bahwa fitur-fitur dalam dataset mampu membedakan kedua kelas dengan baik, sehingga mendukung performa model Regresi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



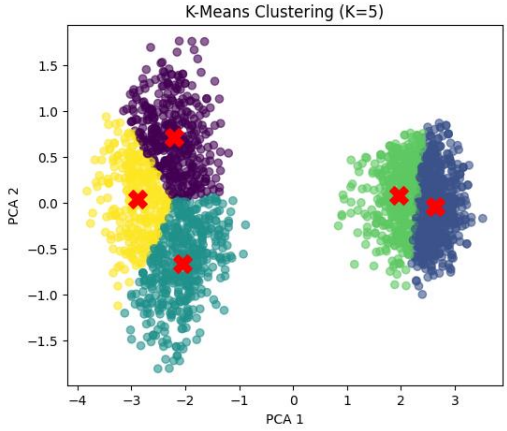
Pada K=2, data terbagi menjadi dua klaster utama yang terpisah dengan jelas pada PCA Component 1. Pembagian ini sangat sesuai dengan dua kelas kepribadian (*Introvert* dan *Extrovert*), sehingga menunjukkan bahwa struktur alami data memang membentuk dua kelompok besar.



Pada K=3, salah satu klaster utama terpecah menjadi dua sub-klaster, sementara klaster lainnya tetap relatif utuh. Hal ini menunjukkan adanya variasi karakteristik di dalam salah satu kelas kepribadian, namun secara umum pemisahan utama masih tetap jelas.



Pada K=4, data terbagi menjadi beberapa klaster yang lebih spesifik. Klaster-klaster ini mulai saling tumpang tindih, terutama pada sisi kiri grafik, yang menandakan bahwa pembagian klaster mulai terlalu detail dan tidak lagi merepresentasikan struktur kelas utama secara sederhana.



Pada K=5, klaster menjadi semakin terfragmentasi dan overlap antar klaster semakin meningkat. Kondisi ini menunjukkan bahwa nilai K terlalu besar, sehingga model membagi data secara berlebihan dan kurang merepresentasikan pola alami dataset.

Kesimpulan

Berdasarkan visualisasi, **K=2 merupakan jumlah klaster yang paling optimal**, karena memberikan pemisahan yang paling jelas dan paling sesuai dengan struktur dua kelas kepribadian pada dataset *personality*.

Makna Cluster secara Bisnis

Cluster pada dataset *personality* merepresentasikan segmentasi individu berdasarkan kepribadian. Cluster *Introvert* cocok untuk peran yang membutuhkan fokus dan analisis tinggi, seperti analis atau programmer, sedangkan cluster *Extrovert* sesuai untuk pekerjaan yang menuntut komunikasi dan interaksi sosial, seperti pemasaran dan layanan pelanggan. Pemahaman cluster ini membantu perusahaan dalam penempatan karyawan dan pengelolaan SDM yang lebih efektif.

P5:

Perbandingan Performa Model

]:

	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. Hal ini menunjukkan bahwa model ini lebih seimbang dalam mengklasifikasikan kedua kelas dan lebih baik dalam mengenali data positif secara konsisten.

Random Forest memiliki **precision (91,84%)** dan **AUC (94,75%)** yang lebih tinggi. Ini menandakan bahwa Random Forest lebih kuat dalam membedakan kelas dan lebih tepat saat memprediksi kelas positif, meskipun recall-nya sedikit lebih rendah.

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.

2. Model Terbaik dan Alasannya

Logistic Regression dapat dianggap sebagai **model terbaik** untuk studi ini karena:

Memiliki **accuracy dan recall sedikit lebih tinggi**, sehingga lebih baik dalam mengenali kedua kelas secara seimbang.

Lebih **sederhana dan interpretable**, sehingga cocok untuk analisis akademik dan pemahaman faktor-faktor yang memengaruhi kepribadian.

Performa yang stabil dan konsisten dengan hasil PCA dan clustering.

Namun, jika fokus utama adalah **kemampuan membedakan kelas (discriminative power)**, maka **Random Forest** unggul dari sisi **AUC**.

3. Keterbatasan

Dataset relatif terbatas (**2900 data**) dan berasal dari satu sumber.

Adanya **missing values** yang memerlukan imputasi, berpotensi menimbulkan bias.

Beberapa variabel bersifat **biner**, sehingga belum menangkap kompleksitas perilaku secara mendalam.

Model hanya membedakan dua tipe kepribadian (Introvert–Extrovert).

4. Peluang Pengembangan

Menambah jumlah dan variasi data untuk meningkatkan generalisasi model.

Menggunakan fitur perilaku yang lebih detail atau bersifat kontinu.

Mencoba model lain seperti **XGBoost, SVM, atau Neural Network**.

Mengembangkan klasifikasi multi-kepribadian, tidak terbatas pada dua kelas.

Menerapkan model pada sistem nyata seperti **personalisasi konten atau rekomendasi pengguna**.

PEMBAGIAN TUGAS

Nama	Pembagian tugas
Fabian Erik Rasyid islami	Membuat laporan/docx
Rendy Herlandi	Menganalisis dataset
Muhammad Farkhan Ridho	Membuat PPT