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

DAFTAR ISI

BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi telah memperluas pemanfaatan data dan sistem informasi di berbagai sektor, khususnya dalam bidang pendidikan dan penelitian. Data mentah yang diolah menggunakan metode analisis yang tepat dapat diubah menjadi informasi yang bernilai, sehingga mampu mengungkap pola yang mendasari serta meningkatkan akurasi dan efisiensi dalam proses pengambilan keputusan.

Oleh karena itu, laporan ini disusun untuk menerapkan konsep dan teori yang telah dipelajari ke dalam situasi praktis melalui proses analisis dan pengolahan data secara sistematis. Laporan ini bertujuan untuk meningkatkan pemahaman dan kemampuan analitis, serta menjadi dokumentasi akademik yang mendukung pengembangan dan kajian lebih lanjut di bidang terkait.

1.2 Rumusan Masalah

1. Sejauh mana waktu yang dihabiskan sendirian (Time_spent_Alone) memengaruhi kecenderungan seseorang menjadi Introvert atau Extrovert?
2. Apakah tingkat kelelahan setelah bersosialisasi (Drained_after_socializing) dapat menjadi indikator utama dalam membedakan kepribadian Introvert dan Extrovert?
3. Dapatkah kombinasi perilaku sosial (kehadiran acara sosial, frekuensi keluar rumah, dan ukuran lingkar pertemanan) digunakan untuk membangun model klasifikasi kepribadian yang akurat?

1.3 Tujuan

1.Untuk mengetahui pengaruh waktu yang dihabiskan sendirian (Time_spent_Alone) terhadap kecenderungan kepribadian Introvert dan Extrovert.

2.Untuk menganalisis peran tingkat kelelahan setelah bersosialisasi (Drained_after_socializing) sebagai indikator pembeda antara kepribadian Introvert dan Extrovert.

3.Untuk membangun dan mengevaluasi model klasifikasi kepribadian berdasarkan kombinasi perilaku sosial guna memprediksi tipe kepribadian seseorang secara akurat.

BAB II

ANALISIS DATA DAN PEMBAHASAN

2.1 Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini adalah dataset kepribadian (personality dataset) yang terdiri dari 2900 baris data dan 8 kolom. Dataset ini berisi informasi tentang perilaku sosial individu dan tipe kepribadian mereka.

2.1.1 Struktur Dataset

Dataset memiliki struktur sebagai berikut:

- **Jumlah data**: 2900 entri (indeks 0 hingga 2899)
- **Jumlah fitur**: 8 kolom
- **Ukuran memori**: 181.4+ KB

2.1.2 Variabel dalam Dataset

Dataset terdiri dari 7 variabel independen dan 1 variabel dependen:

```
 #   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
```

Variabel Independen:

1. **Time_spent_Alone** (float64): Waktu yang dihabiskan sendirian (dalam satuan tertentu)
    1. Rentang nilai: 0.0 - 11.0
    2. Rata-rata: 4.51
    3. Missing values: 63 data
2. **Social_event_attendance** (float64): Tingkat kehadiran dalam acara sosial

    1. Rentang nilai: 0.0 - 10.0
    2. Rata-rata: 3.96
    3. Missing values: 62 data
3. **Going_outside** (float64): Frekuensi aktivitas di luar ruangan

    1. Rentang nilai: 0.0 - 7.0
    2. Rata-rata: 3.00
    3. Missing values: 66 data
4. **Friends_circle_size** (float64): Ukuran lingkaran pertemanan

    1. Rentang nilai: 0.0 - 15.0
    2. Rata-rata: 6.27
    3. Missing values: 77 data
5. **Post_frequency** (float64): Frekuensi posting di media sosial

    1. Rentang nilai: 0.0 - 10.0
    2. Rata-rata: 3.56
    3. Missing values: 65 data
6. **Stage_fear** (object): Ketakutan berbicara di depan umum (Yes/No)

    1. Missing values: 73 data
7. **Drained_after_socializing** (object): Merasa terkuras setelah bersosialisasi (Yes/No)

    1. Missing values: 52 data

Variabel Dependen: 8. **Personality** (object): Tipe kepribadian (Extrovert/Introvert)

Tidak ada missing values (2900 data lengkap)

2.2 Preprocessing Data

dtype: int64

In [77]:
```
num_cols = df.select_dtypes(include=['int64', 'float64']).columns
cat_cols = df.select_dtypes(include=['object']).columns

print("Numerik:", num_cols)
print("Kategorikal:", cat_cols)
```

```
Numerik: Index(['Time_spent_Alone', 'Social_event_attendance', 'Going_outside',
       'Friends_circle_size', 'Post_frequency'],
      dtype='object')
Kategorikal: Index(['Stage_fear', 'Drained_after_socializing', 'Personality'], dtype='object')
```

In [78]:
```
from sklearn.impute import SimpleImputer

num_imputer = SimpleImputer(strategy='median')
cat_imputer = SimpleImputer(strategy='most_frequent')

df[num_cols] = num_imputer.fit_transform(df[num_cols])
df[cat_cols] = cat_imputer.fit_transform(df[cat_cols])

df.isnull().sum()
```

Out[78]:

	0

2.2.1 Penanganan Missing Values

```
<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
```

Dataset memiliki missing values pada beberapa kolom. Metode imputasi yang digunakan adalah:

- **Variabel numerik**: Menggunakan SimpleImputer dengan strategi 'median'
- **Variabel kategorikal**: Menggunakan SimpleImputer dengan strategi 'most_frequent'

Setelah proses imputasi, semua missing values berhasil ditangani dan dataset memiliki 2900 data lengkap untuk semua kolom.

2.2.2 Encoding Variabel Kategorikal

Variabel kategorikal dikonversi menjadi format numerik:

- **Stage_fear**: Yes = 1, No = 0
- **Drained_after_socializing**: Yes = 1, No = 0
- **Personality**: Introvert = 0, Extrovert = 1

2.2.3 Pembagian Dataset

Dataset dibagi menjadi data training dan testing dengan proporsi:

- **Training set**: 80% (2320 data)
- **Testing set**: 20% (580 data)
- **Random state**: 42 (untuk reprodusibilitas)
- **Stratify**: berdasarkan variabel target (Personality)

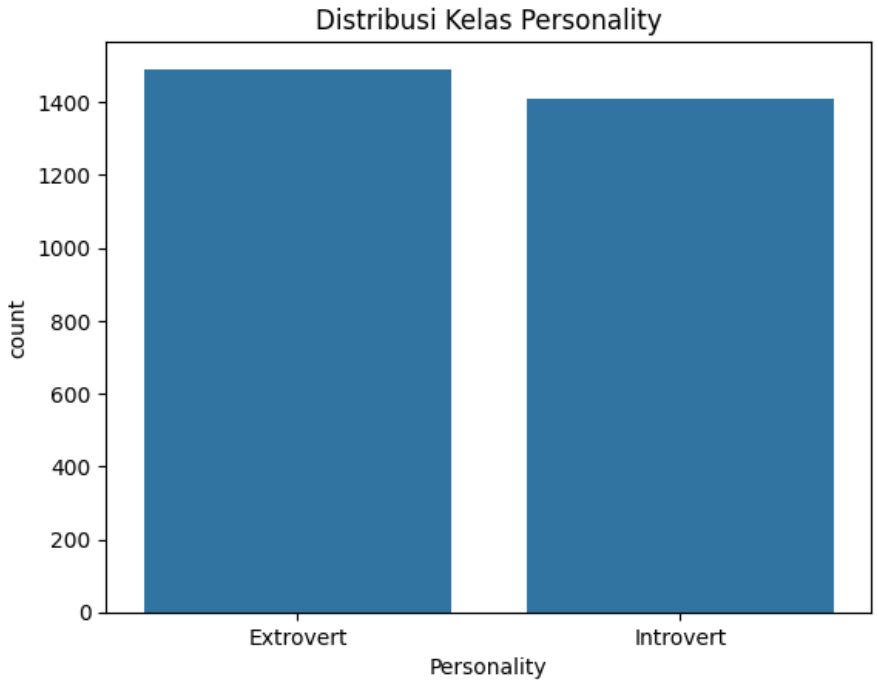
2.2.4 Standarisasi Fitur

Standardisasi dilakukan menggunakan StandardScaler untuk menormalkan skala semua fitur numerik, sehingga memiliki mean = 0 dan standard deviation = 1.

2.3 Exploratory Data Analysis (EDA)

2.3.1 Distribusi Variabel Target

Distribusi tipe kepribadian dalam dataset menunjukkan:

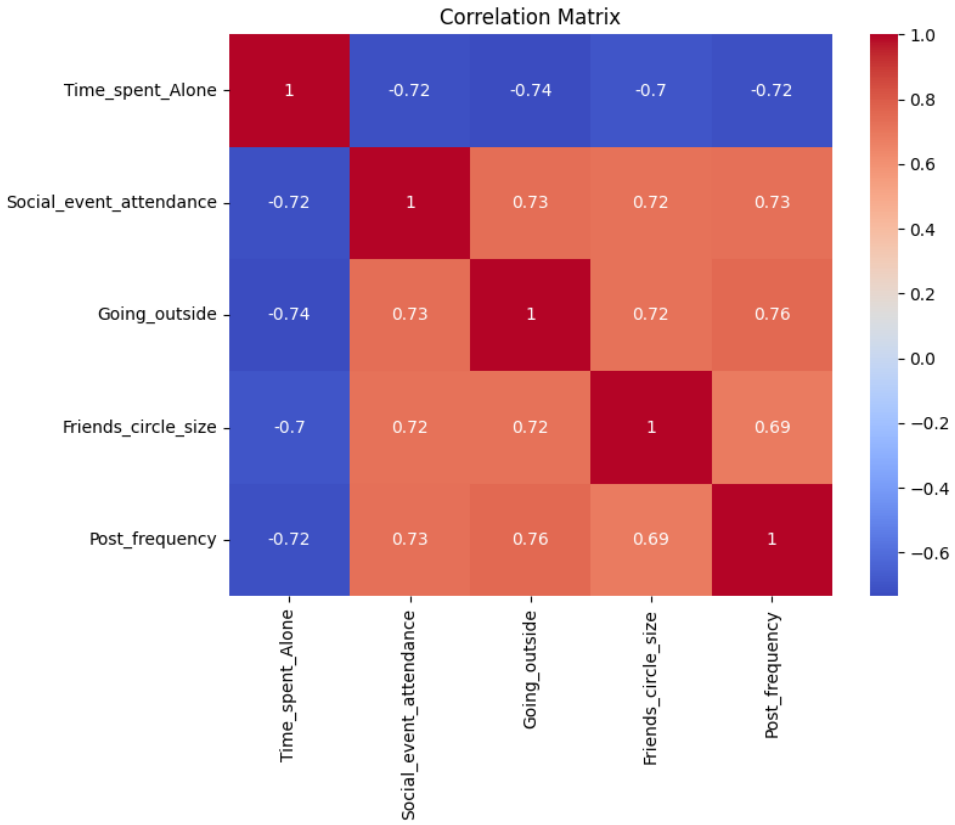


- **Extrovert**: 1.491 individu (51,4%)
- **Introvert**: 1.409 individu (48,6%)

Dataset memiliki distribusi yang relatif seimbang antara kedua kelas, yang mengindikasikan dataset yang baik untuk pemodelan klasifikasi.

2.3.2 Analisis Korelasi

Analisis korelasi antar variabel numerik menunjukkan pola hubungan yang signifikan:



Korelasi Positif Kuat:

- Social_event_attendance dengan Going_outside (r = 0,73)
- Social_event_attendance dengan Post_frequency (r = 0,73)
- Social_event_attendance dengan Friends_circle_size (r = 0,72)
- Going_outside dengan Post_frequency (r = 0,76)
- Going_outside dengan Friends_circle_size (r = 0,72)
- Friends_circle_size dengan Post_frequency (r = 0,69)

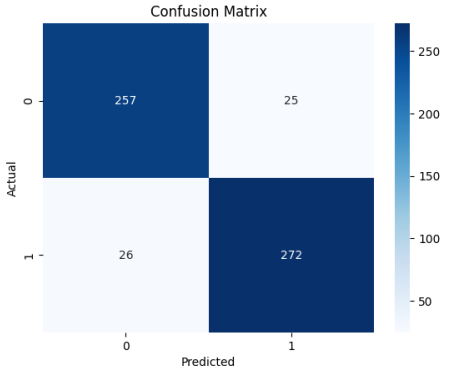
Korelasi Negatif Kuat:

- Time_spent_Alone dengan Social_event_attendance (r = -0,72)
- Time_spent_Alone dengan Going_outside (r = -0,74)
- Time_spent_Alone dengan Friends_circle_size (r = -0,70)
- Time_spent_Alone dengan Post_frequency (r = -0,72)

Pola korelasi ini menunjukkan bahwa individu yang lebih banyak menghabiskan waktu sendirian cenderung memiliki aktivitas sosial yang lebih rendah, lingkaran pertemanan yang lebih kecil, dan frekuensi posting yang lebih rendah.

2.4 Confusion Matrix Logistic Regression

Confusion Matrix untuk model Logistic Regression menunjukkan distribusi prediksi sebagai berikut:



Keterangan:

- **True Negative (TN)**: 257 - Model dengan benar memprediksi 257 individu sebagai Introvert
- **False Positive (FP)**: 25 - Model salah memprediksi 25 individu Introvert sebagai Extrovert
- **False Negative (FN)**: 26 - Model salah memprediksi 26 individu Extrovert sebagai Introvert
- **True Positive (TP)**: 272 - Model dengan benar memprediksi 272 individu sebagai Extrovert

Analisis:

- Total prediksi benar: 257 + 272 = 529 (91,21% dari 580 data test)
- Total prediksi salah: 25 + 26 = 51 (8,79% dari 580 data test)
- Model memiliki tingkat kesalahan yang hampir seimbang antara False Positive dan False Negative
- Visualisasi heatmap menunjukkan nilai diagonal yang tinggi (warna biru gelap), mengindikasikan performa klasifikasi yang baik

2.5 Dimensionality Reduction dengan PCA

2.5.1 Implementasi PCA

Principal Component Analysis (PCA) diterapkan untuk mereduksi dimensi data dengan parameter:

- **Jumlah komponen**: 2
- **Random state**: 42

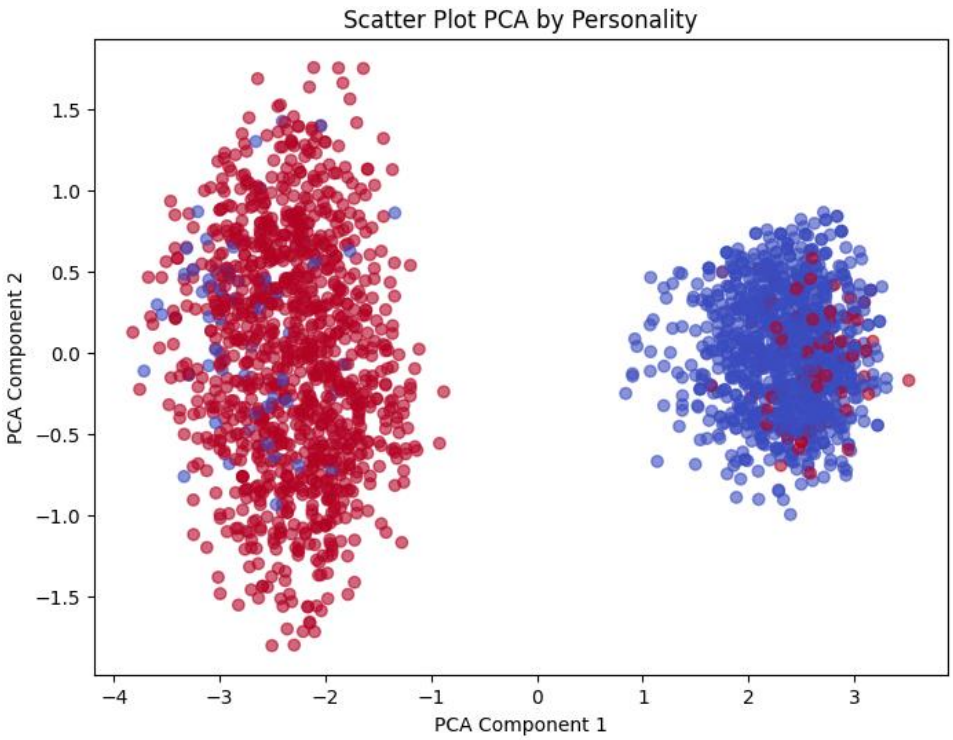
2.5.2 Explained Variance Ratio

Hasil PCA menunjukkan variance yang dijelaskan oleh setiap komponen:

- **PC1 (Principal Component 1)**: 81,57% variance
- **PC2 (Principal Component 2)**: 4,54% variance
- **Total variance explained**: 86,11%

Kedua komponen utama mampu menjelaskan 86,11% dari total variasi data, yang menunjukkan bahwa sebagian besar informasi dalam dataset dapat direpresentasikan dengan baik dalam 2 dimensi.

2.5.3 Visualisasi PCA



Scatter plot PCA menunjukkan pemisahan yang jelas antara kedua tipe kepribadian:

- **Extrovert** (warna merah): Terkonsentrasi pada nilai PC1 negatif
- **Introvert** (warna biru): Terkonsentrasi pada nilai PC1 positif

Pemisahan yang jelas ini mengindikasikan bahwa fitur-fitur dalam dataset memiliki daya diskriminasi yang baik untuk membedakan kedua tipe kepribadian.

2.6 Clustering Analysis dengan K-Means

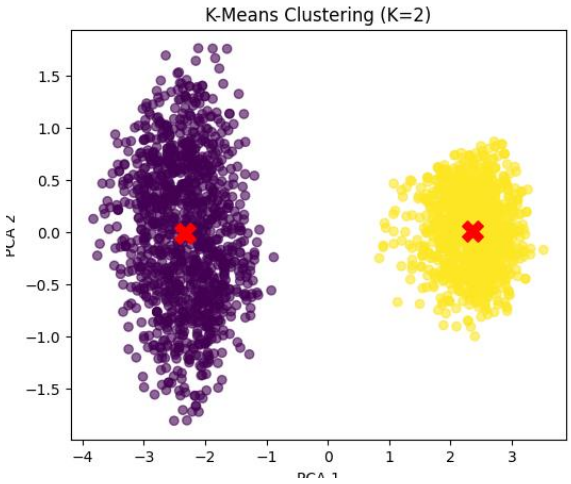
```
2]: from sklearn.cluster import KMeans

for k in range(2, 6):
    kmeans = KMeans(
        n_clusters=k,
        init='k-means++',
        n_init=20,
        random_state=42
    )

    labels = kmeans.fit_predict(X_pca)
    centroids = kmeans.cluster_centers_

    plt.figure(figsize=(6,5))
    plt.scatter(
        X_pca[:,0],
        X_pca[:,1],
        c=labels,
        cmap='viridis',
        alpha=0.6
    )
    plt.scatter(
        centroids[:,0],
        centroids[:,1],
        c='red',
        s=200,
        marker='X'
    )

    plt.title(f'K-Means Clustering (K={k})')
    plt.xlabel('PCA 1')
    plt.ylabel('PCA 2')
    plt.show()
```

K-Means Clustering (K=2)

2.6.1 Implementasi K-Means

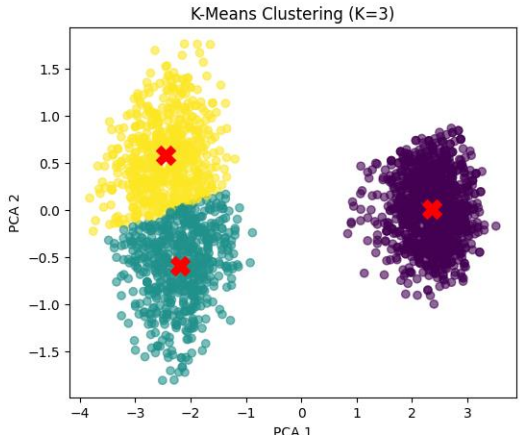
K-Means clustering diterapkan dengan variasi jumlah cluster (K = 2, 3, 4, 5) untuk menganalisis struktur pengelompokan data:

K = 2:



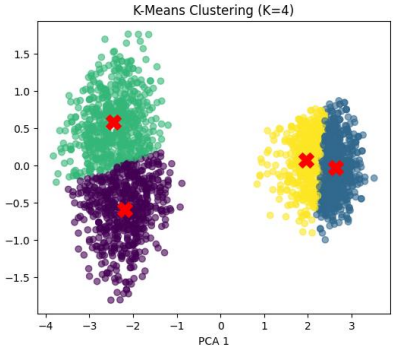
- Menghasilkan dua cluster yang terpisah dengan jelas
- Cluster sebelah kiri (ungu) dan cluster sebelah kanan (kuning)

K = 3:



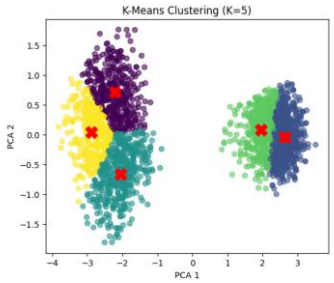
- Cluster pertama (kuning) di bagian atas
- Cluster kedua (hijau tosca) di bagian tengah-bawah kiri
- Cluster ketiga (ungu) di bagian kanan

K = 4:



- Empat cluster dengan pemisahan yang lebih detail
- Menunjukkan sub-grup dalam setiap tipe kepribadian

K = 5:



- Lima cluster menunjukkan granularitas lebih tinggi
- Beberapa cluster mulai tumpang tindih di area tertentu

2.6.2 Interpretasi Clustering

Hasil clustering menunjukkan bahwa:

- K = 2 paling sesuai dengan struktur natural data (Extrovert vs Introvert)
- Cluster yang terbentuk konsisten dengan pemisahan berdasarkan PC1
- Penambahan jumlah cluster mengidentifikasi sub-tipe dalam setiap kepribadian

2.7 Model Classification

2.7.1 Logistic Regression

Model Logistic Regression dibangun dengan parameter:

- **max_iter**: 1000

Hasil Evaluasi:

- **Accuracy**: 0,9121 (91,21%)
- **Precision**: 0,9158 (91,58%)
- **Recall**: 0,9128 (91,28%)
- **F1-score**: 0,9143 (91,43%)
- **AUC**: 0,9212 (92,12%)

2.7.2 Random Forest Classifier

93]:
```
from sklearn.ensemble import RandomForestClassifier

rf_model = RandomForestClassifier(
    n_estimators=200,
    random_state=42
)

rf_model.fit(X_train_scaled, y_train)
```

93]:
```
RandomForestClassifier(n_estimators=200, random_state=42)
```
In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

Model Random Forest dibangun dengan parameter:

- **n_estimators**: 200
- **random_state**: 42

Hasil Evaluasi:

- **Accuracy**: 0,9103 (91,03%)
- **Precision**: 0,9184 (91,84%)
- **Recall**: 0,9060 (90,60%)
- **F1-score**: 0,9122 (91,22%)
- **AUC**: 0,9475 (94,75%)

2.7.3 Confusion Matrix

Logistic Regression:

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

Random Forest:

- Performa serupa dengan confusion matrix yang menunjukkan klasifikasi yang akurat

2.7.4 Perbandingan 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

Analisis Perbandingan:

- Logistic Regression memiliki accuracy sedikit lebih tinggi (0,9121 vs 0,9103)
- Random Forest memiliki AUC yang lebih baik (0,9475 vs 0,9212), menunjukkan kemampuan diskriminasi yang lebih baik
- Kedua model memiliki performa yang sangat baik dengan akurasi di atas 91%
- Precision kedua model hampir identik (91,58% vs 91,84%)
- Logistic Regression memiliki recall sedikit lebih tinggi (91,28% vs 90,60%)

2.8 Pembahasan

2.8.1 Interpretasi Hasil

Penelitian ini berhasil mengklasifikasikan tipe kepribadian (Extrovert vs Introvert) dengan akurasi yang sangat tinggi menggunakan data perilaku sosial. Beberapa temuan penting:

1. **Fitur yang paling berpengaruh**: Waktu yang dihabiskan sendirian (Time_spent_Alone) memiliki korelasi negatif yang kuat dengan berbagai aktivitas sosial, menjadikannya indikator penting untuk tipe kepribadian.
2. **Pola perilaku**: Extrovert cenderung memiliki skor tinggi pada Social_event_attendance, Going_outside, Friends_circle_size, dan Post_frequency, sementara Introvert menunjukkan pola sebaliknya.
3. **Efektivitas PCA**: Reduksi dimensi ke 2 komponen utama tetap mempertahankan 86,11% informasi, membuktikan bahwa struktur data dapat direpresentasikan dengan baik dalam ruang berdimensi lebih rendah.

2.8.2 Kelebihan dan Keterbatasan

Kelebihan:

- Dataset yang seimbang antara kedua kelas
- Akurasi klasifikasi yang sangat tinggi (>91%)
- Pemisahan cluster yang jelas dalam visualisasi PCA
- Konsistensi hasil antara berbagai metode analisis

Keterbatasan:

- Missing values pada beberapa variabel yang memerlukan imputasi
- Dataset hanya mencakup 2900 sampel
- Variabel Stage_fear dan Drained_after_socializing bersifat biner, yang mungkin terlalu sederhana untuk menangkap nuansa perilaku

2.8.3 Implikasi Praktis

Hasil penelitian ini menunjukkan bahwa:

1. Tipe kepribadian dapat diprediksi dengan akurat berdasarkan pola perilaku sosial
2. Model yang dikembangkan dapat digunakan untuk aplikasi praktis seperti sistem rekomendasi konten, personalisasi layanan, atau analisis perilaku pengguna
3. Logistic Regression dapat menjadi pilihan yang lebih sederhana dan interpretable, sementara Random Forest memberikan kemampuan diskriminasi yang lebih baik