Kelompok 4

ASSIGNMENT MODEL DEPLOYMENT

2702352616 - Jovita Putri Aulia

2702363254 - Keyla Faritsha Rindani

2702322720 - Nabila Azwad Ambbiya

2702244284 - Renata Aqila Ridha Putri

2702264576 - Stephanie Nadya

Model Deployment

Extrovert vs. Introvert Behavior Data

| | 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 |

df.shape

(2900, 8)

Dataset ini berisi 2.900 data dengan 8 fitur yang berkaitan dengan perilaku sosial dan ciri kepribadian, yang dirancang untuk mengeksplorasi dan mengklasifikasikan individu sebagai ekstrovert atau introvert.

FEATURES

| No | Feature | Description |
|----|-------------------------|---|
| 1 | Time_spent_Alone | Hours spent alone daily (0–11). |
| 2 | Stage_fear | Presence of stage fright (Yes/No). |
| 3 | Social_event_attendance | Frequency of social events (0–10). |
| 4 | Going_outside | The occupation or profession of the person. |

FEATURES

| No | Feature | Description |
|----|---------------------------|---|
| 5 | Drained_after_socializing | Feeling drained after socializing (Yes/No). |
| 6 | Friends_circle_size | Number of close friends (0–15). |
| 7 | Post_frequency | Social media post frequency (0–10). |
| 8 | Personality | Extrovert or Introvert (target variable). |

KONDISI DATA

```
df.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2900 entries, 0 to 2899
Data columns (total 8 columns):
    Column
                               Non-Null Count Dtype
 #
                                               float64
    Time_spent_Alone
                               2837 non-null
 0
                               2827 non-null
    Stage_fear
                                               object
 1
    Social_event_attendance
                                               float64
                               2838 non-null
    Going_outside
                                               float64
                               2834 non-null
    Drained_after_socializing 2848 non-null
                                               object
    Friends_circle_size
                                               float64
                               2823 non-null
    Post_frequency
                                               float64
                               2835 non-null
    Personality
                               2900 non-null
                                               object
dtypes: float64(5), object(3)
memory usage: 181.4+ KB
```

Dari info ini bisa disimpulkan beberapa informasi:

- Data type pada tiap kolom sudah tepat.
- Ada missing values pada tiap kolom dataset kecuali kolom 'Personality'.

Berikut rincian jumlah missing value per kolom:

| Time_spent_Alone | 63 |
|---------------------------|----|
| Stage_fear | 73 |
| Social_event_attendance | 62 |
| Going_outside | 66 |
| Drained_after_socializing | 52 |
| Friends_circle_size | 77 |
| Post_frequency | 65 |
| Personality | 0 |

IMPUTASI MISSING VALUES KOLOM NUMERIK

| | Time_spent_Alone | Social_event_attendance | Going_outside | Friends_circle_size | Post_frequency |
|-------|------------------|-------------------------|---------------|---------------------|----------------|
| count | 2837.000000 | 2838.000000 | 2834.000000 | 2823.000000 | 2835.000000 |
| mean | 4.505816 | 3.963354 | 3.000000 | 6.268863 | 3.564727 |
| std | 3.479192 | 2.903827 | 2.247327 | 4.289693 | 2.926582 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 2.000000 | 2.000000 | 1.000000 | 3.000000 | 1.000000 |
| 50% | 4.000000 | 3.000000 | 3.000000 | 5.000000 | 3.000000 |
| 75% | 8.000000 | 6.000000 | 5.000000 | 10.000000 | 6.000000 |
| max | 11.000000 | 10.000000 | 7.000000 | 15.000000 | 10.000000 |

- 1. Time_spent_AloneMean > MedianImputasi dengan median
- 2. Social_event_attendance
 Mean > Median
 Imputasi dengan median

- 3. Going_outside

 Mean = Median

 Imputasi dengan mean
- 4. Friends_circle_size

 Mean > Median

 Imputasi dengan median

5. Post_frequencyMean > MedianImputasi dengan median

IMPUTASI MISSING VALUES KOLOM KATEGORIKAL

```
Stage_fear
No 1417
Yes 1410
Name: count, dtype: int64
Drained_after_socializing
No 1441
Yes 1407
Name: count, dtype: int64
```

Kolom 'Stage_fear' dan 'Drained_after_socializing' ini diimputasi dengan nilai modus masing-masing kolom.

DATA SETELAH IMPUTASI

| | Time_spent_Alone | Social_event_attendance | Going_outside | Friends_circle_size | Post_frequency |
|-------|------------------|-------------------------|---------------|---------------------|----------------|
| count | 2900.000000 | 2900.000000 | 2900.000000 | 2900.000000 | 2900.000000 |
| mean | 4.494828 | 3.942759 | 3.000000 | 6.235172 | 3.552069 |
| std | 3.441971 | 2.875987 | 2.221597 | 4.237255 | 2.894794 |
| min | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 2.000000 | 2.000000 | 1.000000 | 3.000000 | 1.000000 |
| 50% | 4.000000 | 3.000000 | 3.000000 | 5.000000 | 3.000000 |
| 75% | 7.000000 | 6.000000 | 5.000000 | 10.000000 | 6.000000 |
| max | 11.000000 | 10.000000 | 7.000000 | 15.000000 | 10.000000 |

Hasil describe setelah imputasi menunjukkan:

- Semua kolom sudah tidak ada missing values (count = 2900)
- Distribusi tidak berubah ekstrem setelah imputasi
- Tidak muncul nilai outlier baru
- Mean dan median tetap konsisten dengan kondisi sebelum imputasi

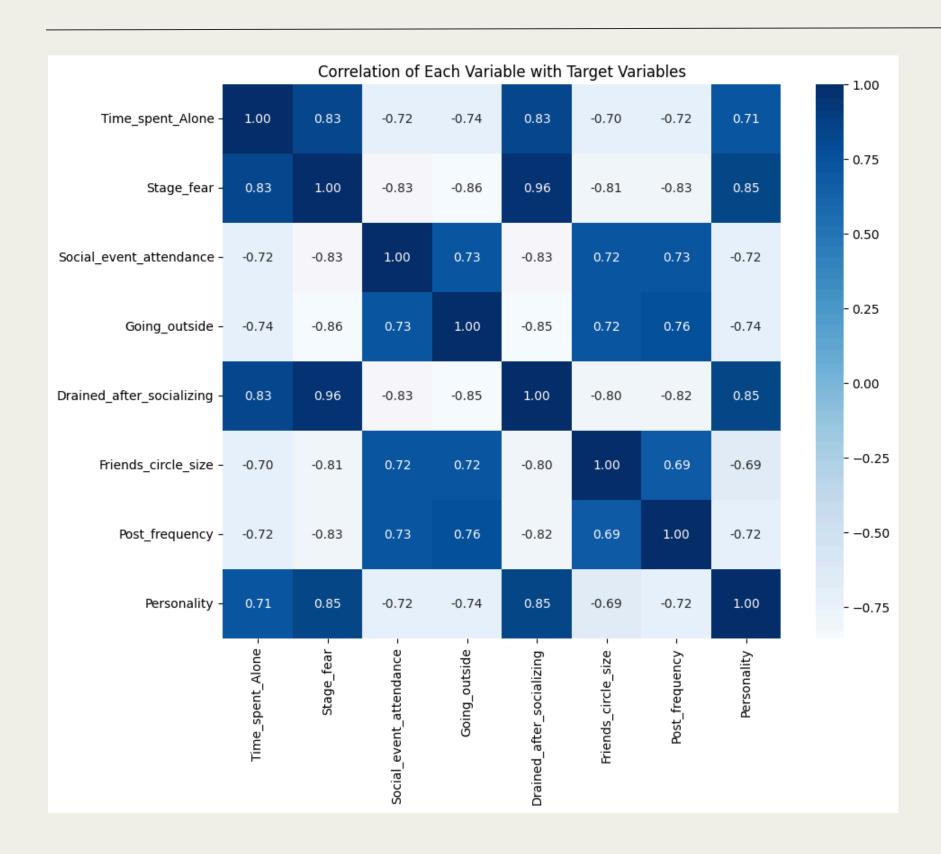
ENCODING

Terdapat 3 kolom yang perlu diencode dengan binary maupun label encoding, berikut rinciannya:

- Stage_fear dengan binary encoding
- Drained_after_socializing dengan binary encoding
- Personality dengan label encoding

Pemetaannya adalah No menjadi 0 dan Yes menjadi 1 untuk binary encoding. Sedangkan Extrovert menjadi 0 dan Introvert menjadi 1 untuk label encoding.

HEATMAP



Heatmap ini menunjukkan korelasi antar variabel. Warna biru tua menandakan korelasi yang kuat, sedangkan biru muda menandakan korelasi lemah. Beberapa poin penting:

- Time_spent_Alone berkorelasi positif dengan Stage_fear dan Drained_after_socializing.
- Stage_fear sangat berkorelasi dengan Drained_after_socializing (0.96).
- Personality berkorelasi negatif dengan variabel seperti Stage_fear dan Time_spent_Alone, menandakan kecenderungan ekstrover.

Heatmap ini membantu memahami hubungan antar fitur sebelum proses modeling

DATA SPLITTING

Pertama, kolom 'Personality' di set menjadi target variabel.

```
input_df=df.drop('Personality',axis=1)
output_df=df['Personality']
```

Lalu, data dipisah menjadi 80% training dan 20% testing sehingga pembagiannya:

```
x_train, x_test, y_train, y_test = train_test_split(input_df, output_df, test_size = 0.2, random_state = 0)

print("shape X_train: ",x_train.shape)
print("shape X_test: ",x_test.shape)
print("shape y_train: ",y_train.shape)
print("shape y_test: ",y_test.shape)

shape X_train: (2320, 7)
shape X_test: (580, 7)
shape y_train: (2320,)
shape y_test: (580,)
```

MODEL - RANDOM FOREST

| Classification Report | | | | | | |
|---------------------------------------|--------------|--------------|----------------------|-------------------|--|--|
| | precision | recall | f1-score | support | | |
| 0 | 0.89 0.93 | 0.92 0.90 | 0.91 0.92 | 270 310 | | |
| accuracy macro avg weighted avg | 0.91 0.91 | 0.91 0.91 | 0.91 0.91 0.91 | 580 580 580 | | |

Pada dataset ini, model Random Forest mampu mengklasifikasikan individu sebagai ekstrovert atau introvert dengan akurasi sebesar 91%.

Model ini menunjukkan keseimbangan performa pada kedua kelas, dengan nilai precision, recall, dan f1-score di atas 0.90. Hal ini menunjukkan bahwa model dapat mengenali pola perilaku dengan baik, tanpa cenderung bias terhadap salah satu kelas.

RANDOM FOREST (HYPERPARAMETER TUNING)

```
param_dist = {
    'n_estimators': [100, 200, 300, 400, 500],
    'max_depth': [None, 5, 10, 15, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4],
    'max_features': ['auto', 'sqrt', 'log2']
}
```

Hyperparameter tuning dilakukan menggunakan RandomizedSearchCV, yaitu metode untuk mencari kombinasi parameter terbaik secara acak. Berikut arti masing-masing parameter yang diuji:

- n_estimators: Jumlah pohon dalam model.
- max_depth: Batas kedalaman pohon.
- min_samples_split: Minimum data untuk membagi node.
- min_samples_leaf: Minimum data pada daun pohon.
- max_features: Jumlah fitur yang dipakai saat split.

Setelah dilakukan hyperparameter tuning menggunakan RandomizedSearchCV model mengalami peningkatan nilai akurasi menjadi 93%

Ini menunjukkan bahwa pemilihan parameter yang tepat berdampak signifikan terhadap performa model.

| | precision | recall | f1-score | support |
|---------------------------|--------------|--------------|--------------|------------|
| 0 | 0.91 | 0.93 | 0.92 | 270 |
| 1 | 0.94 | 0.92 | 0.93 | 310 |
| accuracy | | | 0.93 | 580 |
| macro avg weighted avg | 0.93 0.93 | 0.93 0.93 | 0.93 0.93 | 580 580 |

MODEL - XGBOOST

| Classification Report | | | | | | |
|---------------------------------------|--------------|--------------|----------------------|-------------------|--|--|
| | precision | recall | f1-score | support | | |
| 0 1 | 0.90 0.93 | 0.92 0.91 | 0.91 0.92 | 270 310 | | |
| accuracy macro avg weighted avg | 0.91 0.91 | 0.91 0.91 | 0.91 0.91 0.91 | 580 580 580 | | |

Pada dataset ini, model XGBoost memiliki nilai akurasi sebesar 91% sama seperti Random Forest

XGBoost membangun model secara bertahap dengan memanfaatkan kesalahan dari iterasi sebelumnya, sehingga mampu menghasilkan prediksi yang lebih akurat dan stabil.

XGBOOST (HYPERPARAMETER TUNING)

```
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [3, 5, 7],
    'learning_rate': [0.01, 0.1, 0.2],
    'min_child_weight': [1, 2]
}
```

Setelah dilakukan hyperparameter tuning menggunakan GridSearchCV model mengalami peningkatan nilai akurasi menjadi 93%

menunjukkan bahwa pemilihan parameter yang tepat berdampak signifikan terhadap performa model.

- n_estimators: jumlah pohon yang digunakan dalam boosting.
- max_depth: kedalaman maksimum tiap pohon (mencegah overfitting).
- learning_rate: seberapa besar kontribusi setiap pohon baru (semakin kecil, training lebih lambat tapi akurat).
- min_child_weight: jumlah minimum sampel yang dibutuhkan untuk membuat node baru (mengontrol kompleksitas model).

| Classification Report | | | | | | |
|-----------------------|-----------|--------|----------|---------|--|--|
| | precision | recall | f1-score | support | | |
| 0 | 0.91 | 0.93 | 0.92 | 270 | | |
| 1 | 0.94 | 0.92 | 0.93 | 310 | | |
| accuracy | | | 0.93 | 580 | | |
| accuracy | | | | | | |
| macro avg | 0.93 | 0.93 | 0.93 | 580 | | |
| weighted avg | 0.93 | 0.93 | 0.93 | 580 | | |

Cara Menjalankan FastAPI

1. Buat object base model

```
PreProcessing.ipynb
                       personality.py X
                                                                            streamlit_app.py
                                                          TC New Request
                                          fast_api.py
personality.py > ...
       from pydantic import BaseModel
  2
       class PersonalityBase(BaseModel):
           Time spent Alone: float
           Stage fear: int
  5
           Social_event_attendance: float
  6
           Going_outside: float
  7
           Drained after socializing: int
  8
           Friends_circle_size: float
  9
           Post frequency: float
 10
 11
```

2. Buat fast_api.py untuk bagian backend

```
PreProcessing.ipynb
                                        fast_api.py X TC New Request
                      personality.py
                                                                          streamlit_app.py
* fast_api.py > ...
  1 import uvicorn
  2 from fastapi import FastAPI
      from personality import PersonalityBase
      import pandas as pd
      import joblib
      # Load trained pipeline model
      model = joblib.load("trained_model.pkl")
      # Define Label mapping
      label_map = {0: "Extrovert", 1: "Introvert"}
 11
 12
      # Initialize app
 13
      app = FastAPI()
 15
      @app.get("/")
 16
      def read_root():
          return {"message": "Personality Prediction API is running"}
 18
 19
 20
      @app.post("/predict")
 21
      def predict(data: PersonalityBase):
          # Convert input to DataFrame
 22
          input_df = pd.DataFrame([data.dict()])
 23
 24
 25
          # Predict
 26
          prediction = int(model.predict(input_df)[0])
          label = label_map.get(prediction, "Unknown")
 27
 28
 29
 30
              "prediction": prediction,
 31
              "label": label
 32
 33
      if __name__ == '__main__':
 34
 35
          uvicorn.run(app, host='127.0.0.1', port=7860)
```

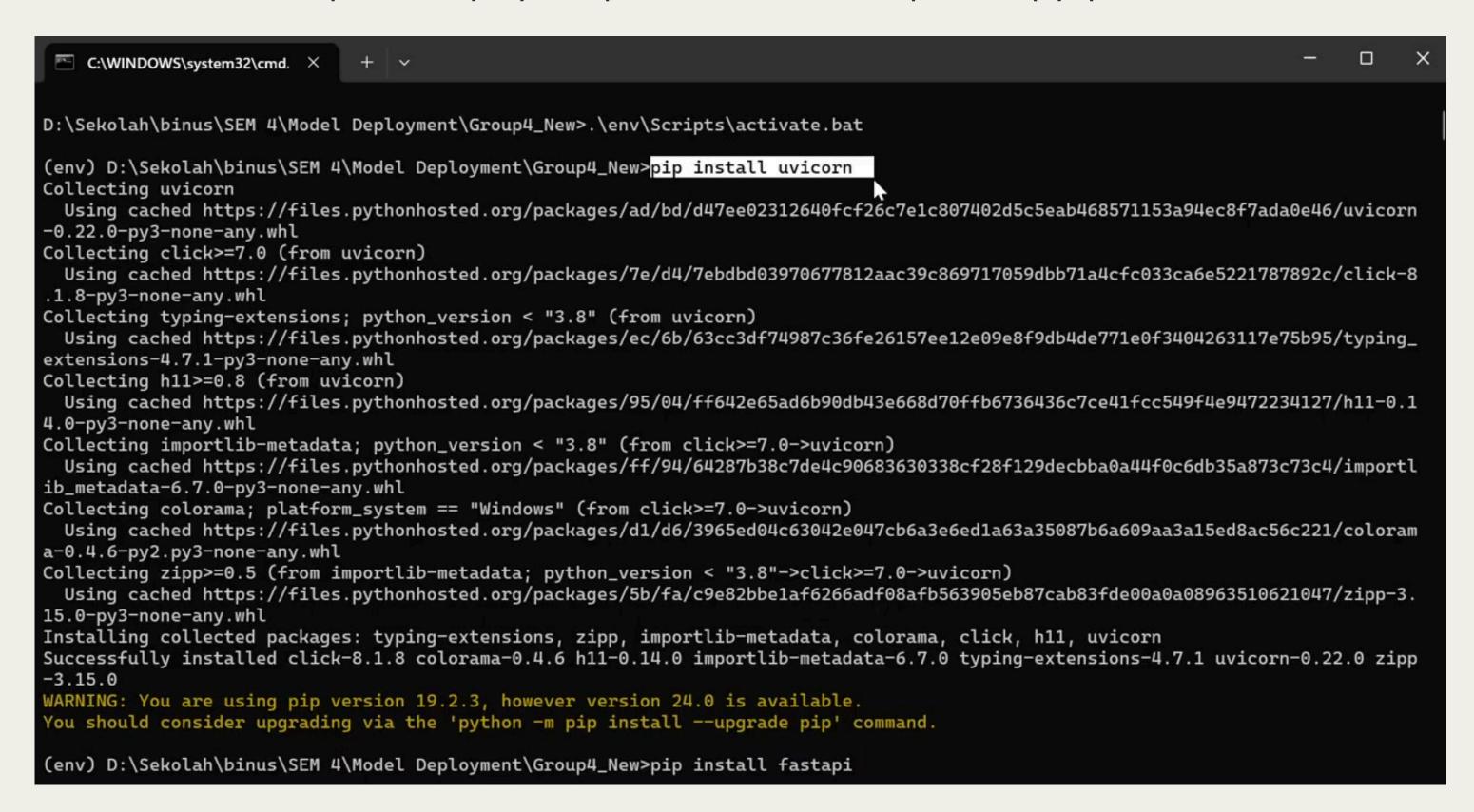
3. Pastikan sudah membuat virtual environment terlebih dahulu. Lalu open cmd dan activate si environmentnya

C:\Users\Renata Aqila>D:

D:\>cd D:\Sekolah\binus\SEM 4\Model Deployment\Group4_New

D:\Sekolah\binus\SEM 4\Model Deployment\Group4_New>.\env\Scripts\activate.bat

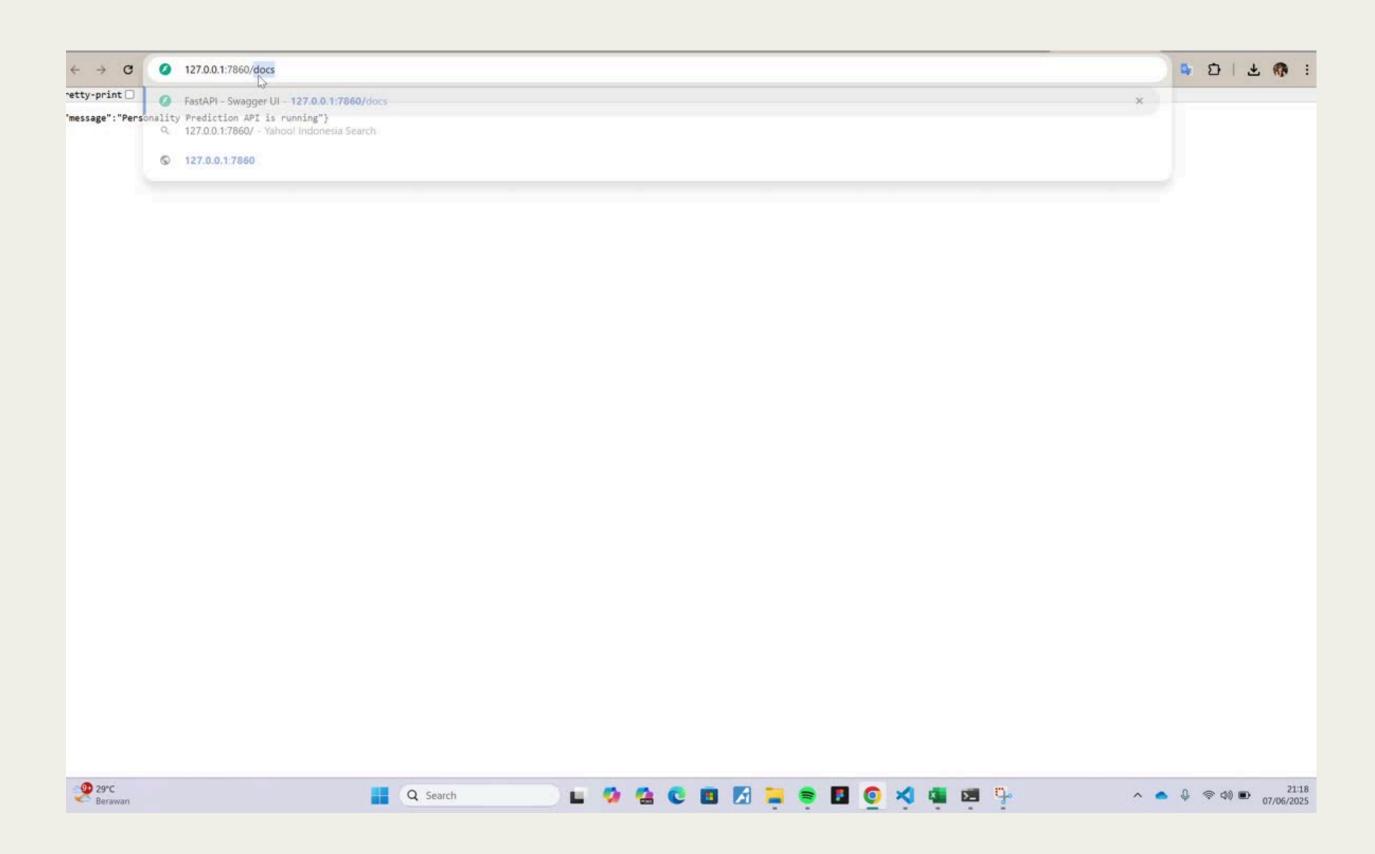
4. Lalu install beberapa library nya seperti uvicorn, fastapi, numpy, pandas, dan scikit-learn



5. Setelah itu, ketik 'uvicorn fast_api:app --reload' di cmd, dan nanti akan muncul link yang redirect ke fastAPI nya

```
(env) D:\Sekolah\binus\SEM 4\Model Deployment\Group4_New>uvicorn fast_api:app --reload --port 7860
          Will watch for changes in these directories: ['D:\\Sekolah\\binus\\SEM 4\\Model Deployment\\Group4_New']
          Uvicorn running or <a href="http://127.0.0.1:7860">http://127.0.0.1:7860</a> (Press CTRL+C to quit)
INFO:
          Started reloader process [12656] using StatReload
INFO:
          Started server process [25868]
INFO:
          Waiting for application startup.
INFO:
          Application startup complete.
INFO:
          127.0.0.1:54374 - "GET / HTTP/1.1" 200 OK
INFO:
          127.0.0.1:54375 - "GET /docs HTTP/1.1" 200 OK
INFO:
INFO:
          127.0.0.1:54375 - "GET /openapi.json HTTP/1.1" 200 OK
          127.0.0.1:54378 - "POST /predict HTTP/1.1" 200 OK
INFO:
          127.0.0.1:54378 - "POST /predict HTTP/1.1" 200 OK
INFO:
         StatReload detected changes in 'streamlit_app.py'. Reloading...
          Shutting down
INFO:
          Waiting for application shutdown.
INFO:
          Application shutdown complete.
INFO:
          Finished server process [25868]
INFO:
          Started server process [10016]
INFO:
          Waiting for application startup.
INFO:
          Application startup complete.
INFO:
```

6. Tinggal di click linknya, lalu tambahkan '/docs' di website



Testing

Di ipynb, kita menggunakan rumus codingan ini untuk mengambil salah satu data dari csv kita. Misalnya kita mengambil data di baris pertama

```
sample = df.iloc[0].to_dict()
print(sample)

     0.0s
```

{'Time_spent_Alone': 4.0, 'Stage_fear': 0.0, 'Social_event_attendance': 4.0, 'Going_outside': 6.0, 'Drained_after_socializing': 0.0, 'Friends_circle_size': 13.0, 'Post_frequency': 5.0, 'Personality': 0.0}

Testing

Setelah itu, kita input ke dalam thunder clientnya

```
JSON Content

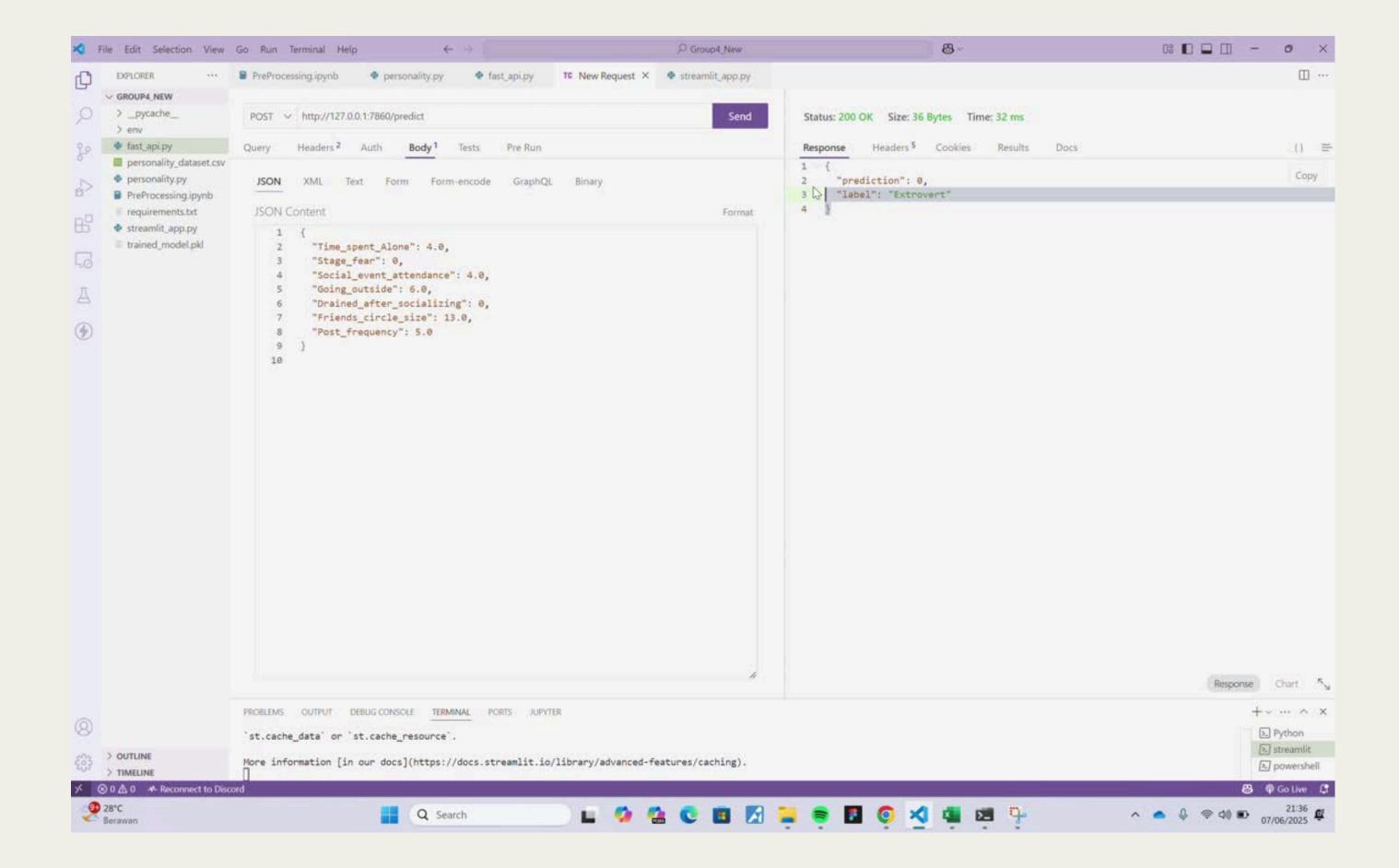
1  {
2    "Time_spent_Alone": 4.0,
3    "Stage_fear": 0,
4    "Social_event_attendance": 4.0,
5    "Going_outside": 6.0,
6    "Drained_after_socializing": 0,
7    "Friends_circle_size": 13.0,
8    "Post_frequency": 5.0
9  }
10  |
```

Testing

Seperti yang di file preprocessing, kita udah encode beberapa variable seperti Stage_fear, Drained_After_socializing dan Personality, jadi ketika input di thunder client ini, sudah menggunakan angka semua

```
JSON Content

1 {
2    "Time_spent_Alone": 4.0,
3    "Stage_fear": 0,
4    "Social_event_attendance": 4.0,
5    "Going_outside": 6.0,
6    "Drained_after_socializing": 0,
7    "Friends_circle_size": 13.0,
8    "Post_frequency": 5.0
9  }
10
```



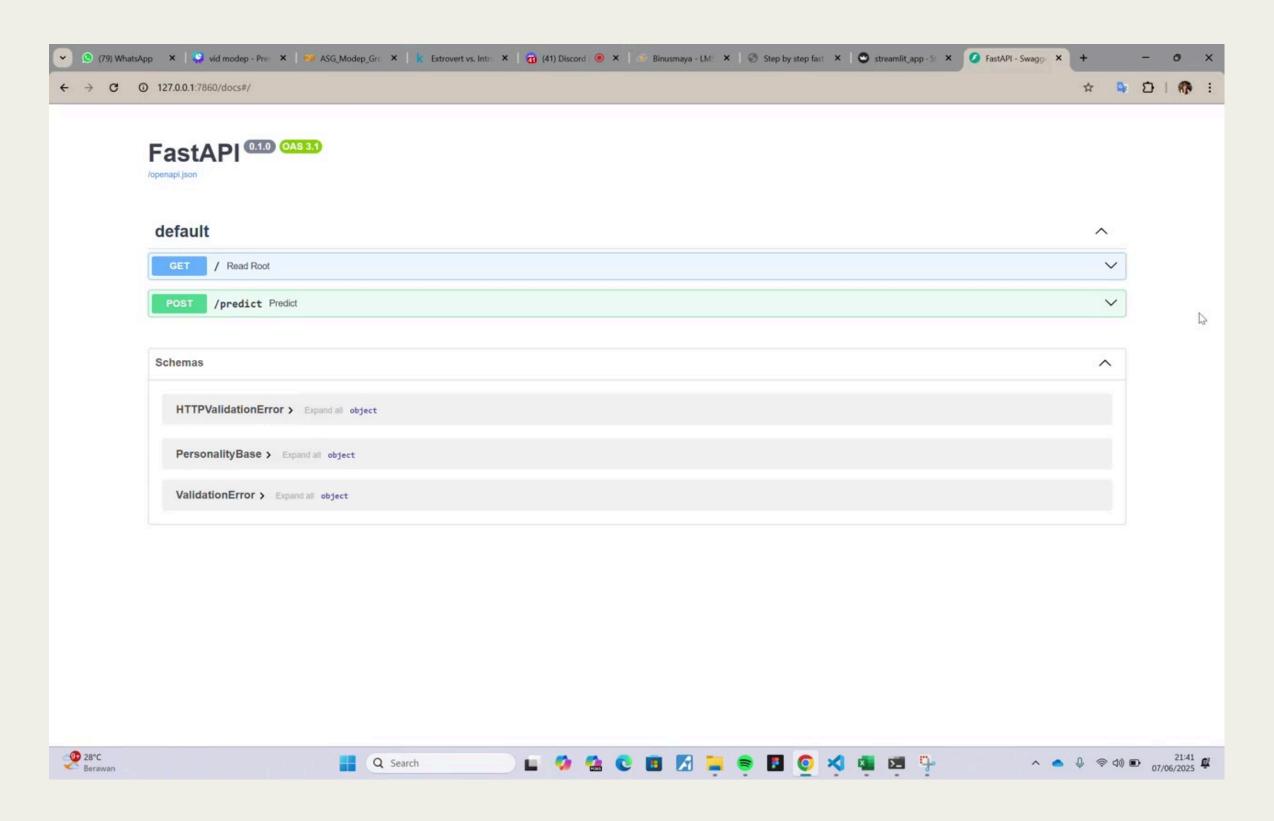
Hasil ouput nya udah sesuai dengan data yang kita punya, yaitu Extrovert

```
Response Headers <sup>5</sup> Cookies Results Docs

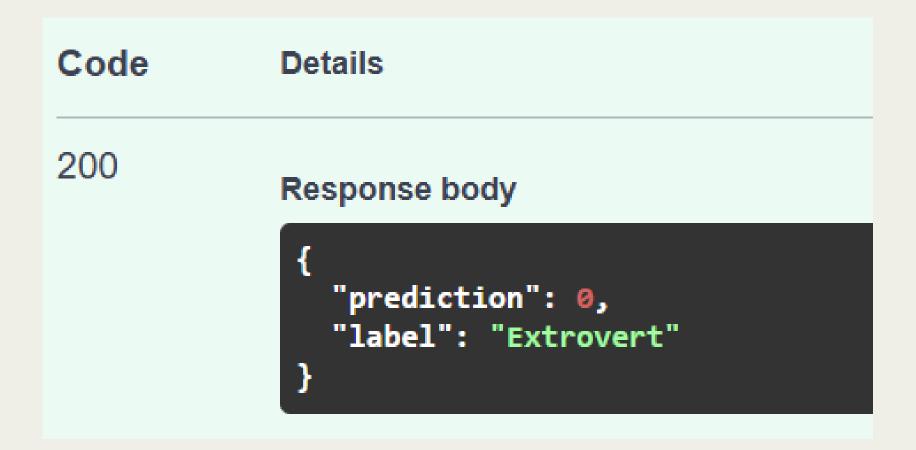
1 {
2    "prediction": 0,
3    "label": "Extrovert"
4  }

{'Time_spent_Alone': 4.0, 'Stage_fear': 0.0, 'Social_event_attendance': 4.0, 'Going_outside': 6.0, 'Drained_after_socializing': 0.0,
'Friends_circle_size': 13.0, 'Post_frequency': 5.0, 'Personality': 0.0}
```

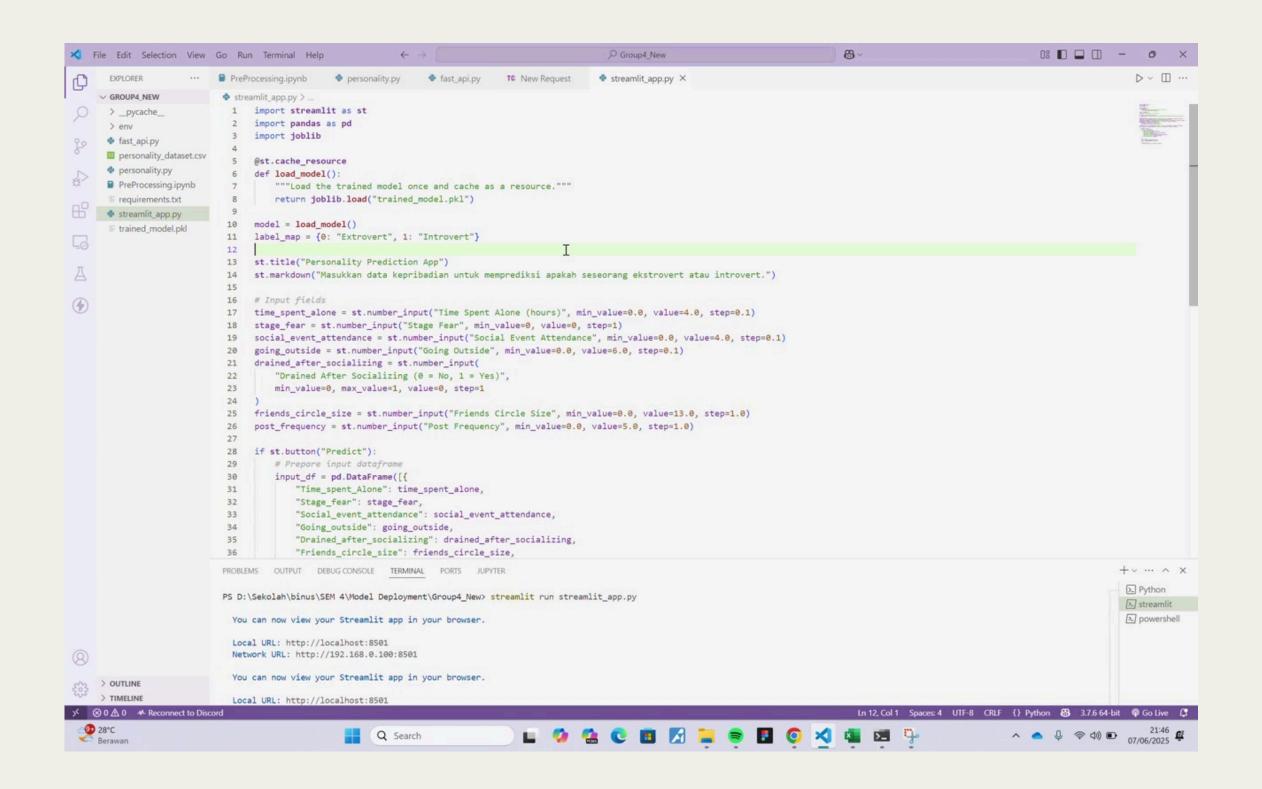
Setelah itu, kita juga coba input data di fastAPI nya, dan menunjukkan output yang sama seperti di Thunder Client



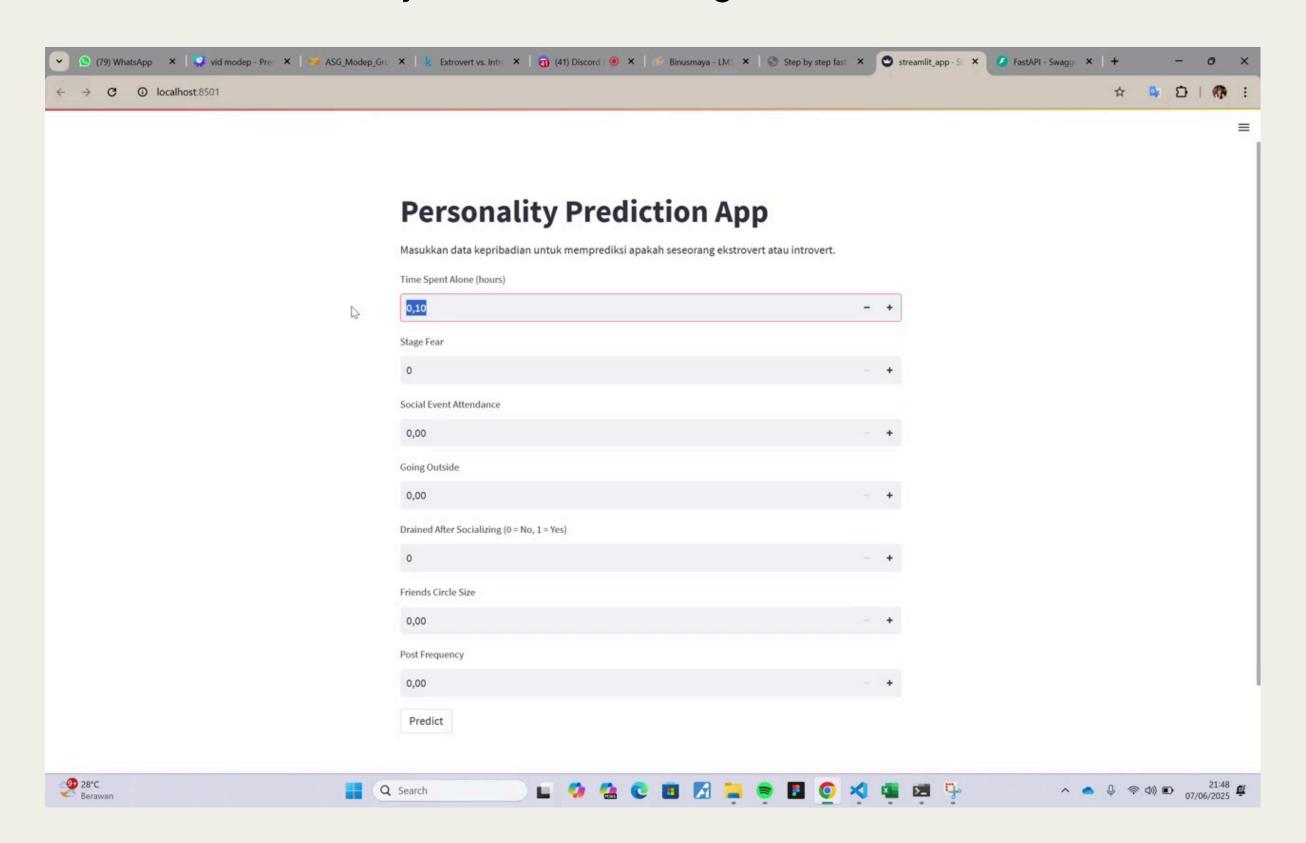
Setelah itu, kita juga coba input data di fastAPI nya, dan menunjukkan output yang sama seperti di Thunder Client



Selanjutnya, kita membuat codingan streamlitnya dan menjalankan 'streamlit run streamlit_app.py' di terminal vscode, lalu kita klik link yang tertera di terminal vscode



Redirect ke streamlit, kita langsung input sesuai dengan inputan kita di awal. Hasil output menunjukkan kalau orang adalah Extrovert



Thank You!