

DOKUMEN PROYEK

12S3205 - PENAMBANGAN DATA

CLASSIFICATION OF EXPLORING MENTAL DATA HEALTH USING LOGISTIC REGRESSION



Disusun Oleh :

12S22030	Bryan Evans Simamora
12S22049	Agnes Monica Sanjani Harefa
12S22050	Yohana Christine Sitanggang

**PROGRAM STUDI SARJANA SISTEM INFORMASI
FAKULTAS INFORMATIKA DAN TEKNIK ELEKTRO
(FITE)
INSTITUT TEKNOLOGI DEL**

DAFTAR ISI

BAB 1	3
BUSINESS UNDERSTANDING	3
1.1 Determine Business Objective	3
1.2 Determine Project Goal	3
1.3 Produce Project Plan	3
BAB 2	5
DATA UNDERSTANDING	5
2.1 Pengumpulan Data	5
2.2 Describe Data	5
2.3 Validation Data	8

BAB 1

BUSINESS UNDERSTANDING

1.1 Determine Business Objective

Kesehatan mental merupakan salah satu aspek penting dalam kesejahteraan manusia yang sering kali terabaikan. Dalam dunia modern yang penuh tekanan, terutama di lingkungan kerja atau pendidikan, semakin banyak individu yang mengalami gangguan mental seperti depresi. Namun, banyak kasus depresi tidak terdeteksi secara dini, sehingga memperburuk kondisi penderita.

Melalui proyek ini, kami bertujuan untuk menganalisis data survei kesehatan mental guna mengidentifikasi faktor-faktor risiko yang berkontribusi terhadap depresi. Harapannya, hasil dari proyek ini dapat menjadi dasar untuk melakukan pencegahan dan memberikan dukungan yang tepat kepada individu yang rentan mengalami gangguan mental.

1.2 Determine Project Goal

- Tujuan Bisnis
 - Mengidentifikasi faktor-faktor yang memengaruhi risiko gangguan mental di kalangan pekerja.
 - Memberikan insight bagi perusahaan untuk membangun lingkungan kerja yang lebih sehat.
 - Membantu pihak HR dan kebijakan dalam menyusun program dukungan mental health berdasarkan data.
 - Menyediakan visualisasi yang membantu pemangku kebijakan memahami kondisi karyawan.
- Tujuan *Data Science*
 - Membangun model klasifikasi untuk memprediksi apakah seseorang berisiko mengalami depresi.
 - Melakukan EDA (*Exploratory Data Analysis*) untuk melihat distribusi data dan pola tersembunyi.
 - Menghasilkan visualisasi untuk membantu stakeholder memahami temuan.

1.3 Produce Project Plan

- Kriteria Kesuksesan (*Success Criteria*)

Dari sisi teknis:

Model klasifikasi memiliki:

- Akurasi > 60%
- Precision > 60%
- Recall > 60
- Model berhasil di-deploy secara online dan dapat menerima input dari user.

Dari sisi bisnis:

- Model dapat digunakan sebagai alat bantu untuk deteksi dini risiko depresi.
- Hasil analisis dapat meningkatkan kesadaran dan perhatian terhadap isu kesehatan mental di lingkungan kerja maupun pendidikan.
- Model memberikan insight fitur mana yang paling memengaruhi risiko depresi.

- Kendala dan Asumsi
 - Data yang digunakan bersifat sintetis, bukan data real-world, sehingga interpretasi hasil harus dilakukan dengan hati-hati dan tidak digunakan untuk diagnosis medis.
 - Beberapa fitur mungkin memiliki nilai kosong (*missing value*) atau tidak konsisten, sehingga perlu dilakukan pembersihan dan pra-pemrosesan.
 - Dataset berasal dari *survei global* → mungkin tidak sepenuhnya mewakili konteks lokal (seperti Indonesia), namun tetap relevan untuk studi dan pengembangan model.
 - Model tidak ditujukan sebagai alat diagnosa medis, melainkan hanya sebagai tools prediksi dan analisis awal.

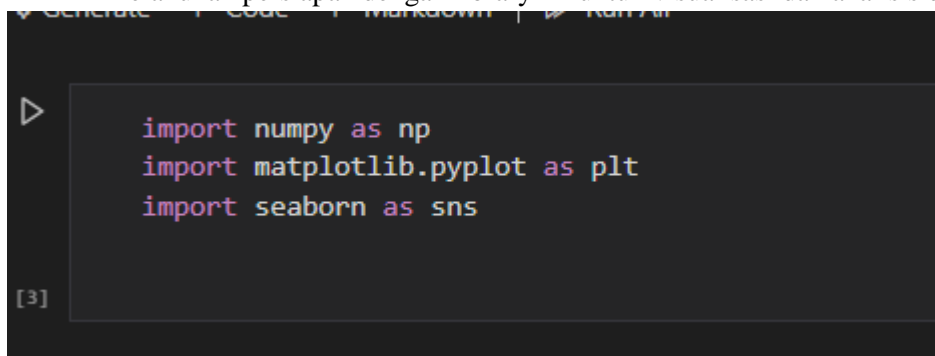
BAB 2

DATA UNDERSTANDING

2.1 Pengumpulan Data

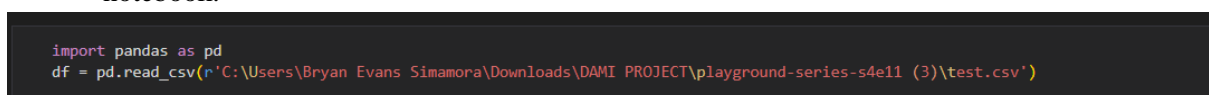
Dataset yang digunakan dalam proyek ini bersumber dari [Kaggle Playground Series S4E11](#). Dataset ini bersifat sintetis, dirancang untuk kebutuhan pembelajaran dan eksperimen dalam pengembangan model prediksi kesehatan mental.

- Format Data: CSV (Comma Separated Values)
- Jumlah Data:
 - Train Set: 140.700 baris, 20 kolom
 - Test Set: 60.300 baris (tanpa label Depression)
- Target Prediksi: Depression (0 = tidak depresi, 1 = depresi).
- Melakukan persiapan dengan library ini untuk visualisasi dan analisis data numerik



```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

- numpy → manipulasi angka dan array
- matplotlib.pyplot → visualisasi dasar
- seaborn → visualisasi lanjutan dengan tampilan menarik
- Bagian ini menjelaskan dari mana data diperoleh dan bagaimana cara memuatnya ke dalam notebook.



```
import pandas as pd
df = pd.read_csv(r'C:\Users\Bryan Evans Simamora\Downloads\DAMI PROJECT\playground-series-s4e11 (3)\test.csv')
```

Penjelasan:

- Data diambil dari file CSV bernama Book1.csv yang disimpan di Google Drive.
- File dimuat menggunakan `pandas.read_csv()`, yang merupakan cara umum untuk mengimpor data tabular ke dalam Python.

2.2 Describe Data

Data terdiri dari gabungan variabel kategorikal dan numerik, yang mencakup karakteristik demografis, kebiasaan harian, tekanan akademik/kerja, hingga aspek psikologis. Berikut beberapa fitur penting:

1. train.csv

- Jumlah data: 140.700 baris, 20 kolom
- Kolom-kolom penting:
 - Age: Rata-rata usia 40.4 tahun (18–60 tahun)

- Gender: 2 kategori → Male, Female
- City: 98 kota berbeda
- Profession: 64 jenis pekerjaan, dengan Teacher paling banyak (24.906)
- Academic Pressure, CGPA, Study Satisfaction: hanya terisi sekitar 20% data
- Depression: 0 dan 1, dengan hanya ~18% kasus depresi (klasifikasi imbalanced)

2. test.csv

- Jumlah data: 93.800 baris, 19 kolom (tanpa Depression)
- Distribusi mirip dengan train.csv
- Nilai missing di kolom seperti:
 - Academic Pressure, CGPA, Profession → banyak kosong
 - Gender, City, Degree dan lainnya → cukup lengkap

3. sample_submission.csv

- Jumlah data: 93.800 baris, 2 kolom:
 - id: Identik dengan test set
 - Depression: Default-nya semua bernilai 0 (placeholder untuk prediksi).

Tahap ini bertujuan untuk memahami struktur data, kolom, tipe data, dan melihat ringkasan statistik.

```
df.head()
```

	id	Name	Gender	Age	City	Working Professional or Student	Profession	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness
0	140700	Shivam	Male	53.0	Vizakhapatnam	Working Professional	Judge	NaN	2.0	NaN	NaN	5.0	Less than 5 hours	Moderate	LLB	No	9.0	3.0	Yes
1	140701	Sanya	Female	58.0	Kolkata	Working Professional	Educational Consultant	NaN	2.0	NaN	NaN	4.0	Less than 5 hours	Moderate	B.Ed	No	6.0	4.0	No
2	140702	Yash	Male	53.0	Japur	Working Professional	Teacher	NaN	4.0	NaN	NaN	1.0	7-8 hours	Moderate	B.Arch	Yes	12.0	4.0	No
3	140703	Nalini	Female	23.0	Rajkot	Student	NaN	5.0	NaN	6.84	1.0	NaN	More than 8 hours	Moderate	B.Sc	Yes	10.0	4.0	No
4	140704	Shourya	Male	47.0	Kalyan	Working Professional	Teacher	NaN	5.0	NaN	NaN	5.0	7-8 hours	Moderate	BCA	Yes	3.0	4.0	No

- data.head() → Menampilkan 5 baris pertama untuk melihat isi dataset secara umum.

```
df.info()
df.describe()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 93800 entries, 0 to 93799
Data columns (total 19 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   id                                    93800 non-null  int64
1   Name                                93800 non-null  object
2   Gender                              93800 non-null  object
3   Age                                  93800 non-null  float64
4   City                                93800 non-null  object
5   Working Professional or Student      93800 non-null  object
6   Profession                           69168 non-null  object
7   Academic Pressure                    18767 non-null  float64
8   Work Pressure                        75022 non-null  float64
9   CGPA                                 18766 non-null  float64
10  Study Satisfaction                   18767 non-null  float64
11  Job Satisfaction                     75026 non-null  float64
12  Sleep Duration                       93800 non-null  object
13  Dietary Habits                       93795 non-null  object
14  Degree                               93798 non-null  object
15  Have you ever had suicidal thoughts ? 93800 non-null  object
16  Work/Study Hours                     93800 non-null  float64
17  Financial Stress                      93800 non-null  float64
18  Family History of Mental Illness      93800 non-null  object
dtypes: float64(8), int64(1), object(10)
memory usage: 13.6+ MB
```

	id	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress
count	93800.000000	93800.000000	18767.000000	75022.000000	18766.000000	18767.000000	75026.000000	93800.000000	93800.000000
mean	187599.500000	40.321685	3.158576	3.011797	7.674016	2.939522	2.96092	6.247335	2.978763
std	27077.871962	12.393480	1.386666	1.403563	1.465056	1.374242	1.41071	3.858191	1.414604
min	140700.000000	18.000000	1.000000	1.000000	5.030000	1.000000	1.00000	0.000000	1.000000
25%	164149.750000	29.000000	2.000000	2.000000	6.330000	2.000000	2.00000	3.000000	2.000000
50%	187599.500000	42.000000	3.000000	3.000000	7.800000	3.000000	3.00000	6.000000	3.000000
75%	211049.250000	51.000000	4.000000	4.000000	8.940000	4.000000	4.00000	10.000000	4.000000
max	234499.000000	60.000000	5.000000	5.000000	10.000000	5.000000	5.00000	12.000000	5.000000

- `data.info()` → Memberikan informasi tentang jumlah baris, kolom, dan tipe data di setiap kolom.

Menampilkan:

- Jumlah total baris (RangeIndex: 93800 entries)
- Total kolom: 19
- Nama setiap kolom, jumlah non-null values (tidak kosong), dan tipe datanya.

Insight penting:

Beberapa kolom memiliki data kosong (null):

- Academic Pressure, CGPA, Study Satisfaction, Job Satisfaction → jumlah non-null = 18767 (hanya sekitar 20% dari total data).
- Work Pressure → 75202 data terisi (sekitar 80%).

Kolom lainnya seperti Age, Gender, City, Degree, dll memiliki data lengkap (93800).

Tipe data:

- object: data teks / kategori (misal: Gender, City, Profession)
- float64 dan int64: data numerik (misal: Age, CGPA, Financial Stress)

Kesimpulan:

Kita perlu menangani missing values pada kolom numerik seperti CGPA, Study Satisfaction, dll sebelum proses modeling. Ini bisa dengan imputasi atau drop baris tergantung konteks.

- `data.describe()` → Menampilkan statistik deskriptif (rata-rata, standar deviasi, nilai minimum/maksimum, dll) untuk kolom numerik.

Kolom	Mean	Std Dev	Min – Max	Catatan
Age	30.42	12.39	18-100	Cakupan usia sangat lebar, dari pelajar muda sampai pekerja senior.
Academic Pressure	3.15	1.38	1 – 5	Skala 1–5, rerata cukup tinggi → tekanan akademik tergolong besar.
Work Pressure	3.01	1.40	1 – 5	Mirip dengan academic pressure.
CGPA	7.67	1.46	1 – 10	Terlihat seperti skala IPK maksimal 10, tidak biasa (cek konteks).
Study Satisfaction	2.94	1.37	1 – 5	Rata-rata di bawah 3 → mungkin responden kurang puas.
Work/Study Hours	6.24	3.85	0 – 12	Variasi cukup tinggi, dari yang tidak kerja/belajar hingga 12 jam.
Financial Stress	2.98	1.41	1 – 5	Hampir 3 → kondisi finansial cukup menekan secara umum.

Kesimpulan:

Kita bisa mulai menduga pola – misalnya: apakah tekanan akademik/kerja berkorelasi dengan stres finansial atau kepuasan studi? Juga apakah usia berpengaruh terhadap tekanan kerja?

2.3 Validation Data

- Missing Value: Telah diidentifikasi dan akan ditangani dengan strategi seperti imputasi atau penghapusan.
- Outlier & Anomali: Nilai Age dalam rentang wajar (18–60), namun beberapa kategori seperti Sleep Duration dan Gender memiliki nilai non-standar.
- Data Duplikat: Belum ditemukan duplikasi signifikan pada kolom ID (unik).
- Distribusi Target: Dataset imbalanced → hanya sekitar 18% responden mengalami depresi (Depression = 1).

➤ Menampilkan jumlah nilai kosong (missing values) di setiap kolom dalam DataFrame.

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

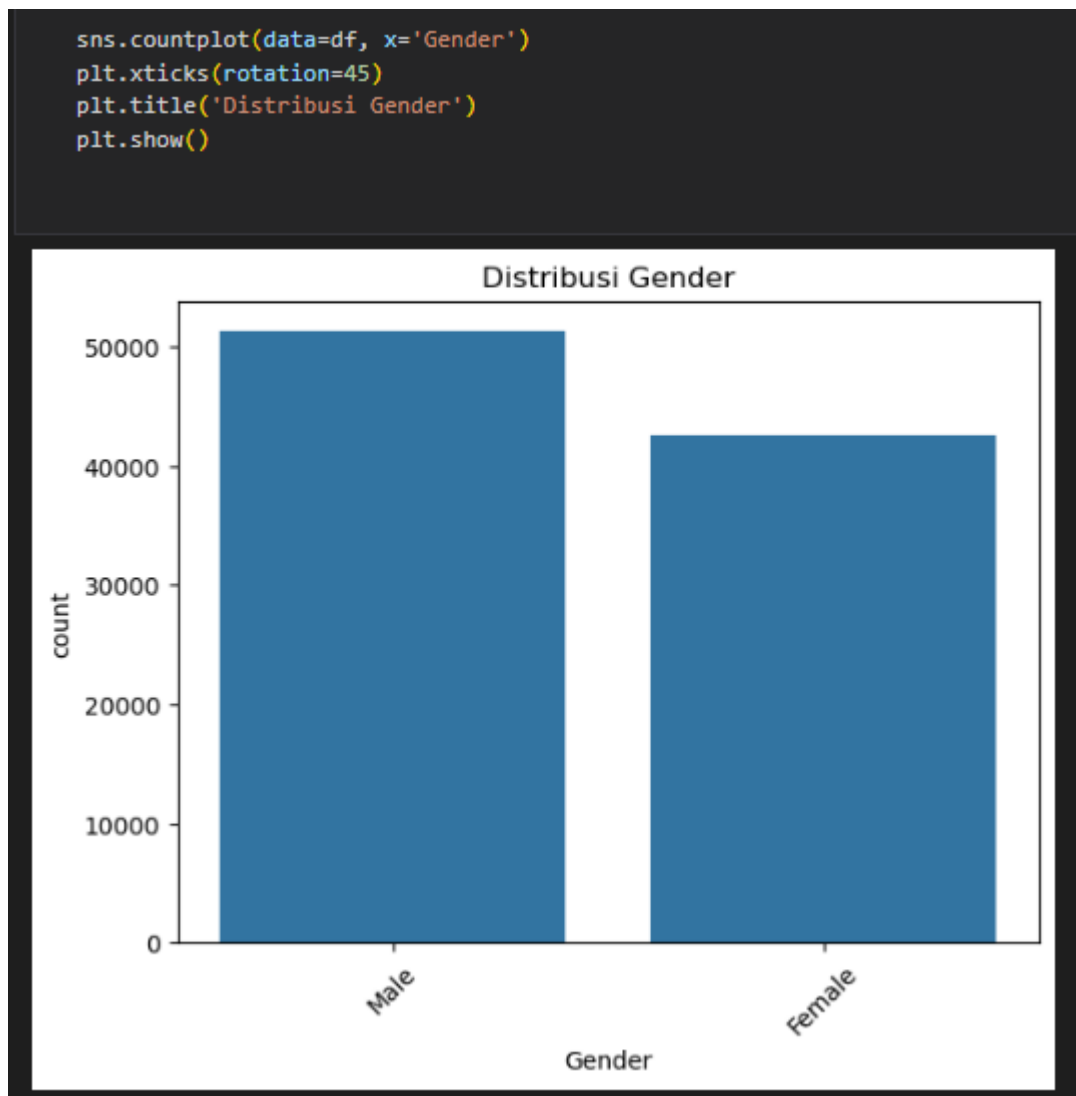
id          0
Name        0
Gender      0
Age         0
City        0
Working Professional or Student  0
Profession  24632
Academic Pressure  75033
Work Pressure  18778
CGPA        75034
Study Satisfaction  75033
Job Satisfaction  18774
Sleep Duration  0
Dietary Habits  5
Degree      2
Have you ever had suicidal thoughts ?  0
Work/Study Hours  0
Financial Stress  0
Family History of Mental Illness  0
dtype: int64
```

Kolom	Jumlah Kosong	Penjelasan
id, Name, Gender, Age, City, Work/Study Hours, Financial Stress, Family History of Mental Illness	0	Kolom ini lengkap, tidak ada nilai kosong.
Profession	24,632	Sekitar 26% data tidak punya informasi profesi. Mungkin karena banyak pelajar belum bekerja.
Academic Pressure, CGPA, Study Satisfaction	75,033–75,034	Hanya sekitar 18 ribu data yang punya nilai ini. Kemungkinan data ini hanya diisi oleh mahasiswa/pelajar.
Work Pressure, Job Satisfaction	18,778 dan 18,774	Mirip, mungkin hanya tersedia untuk yang sudah bekerja.

Kolom	Jumlah Kosong	Penjelasan
Sleep Duration, Degree, Have you ever had suicidal thoughts?	2–5	Hanya ada beberapa data kosong, bisa langsung di-drop atau diimputasi.
Dietary Habits	5	Jumlah kecil juga, tidak terlalu bermasalah.

Fungsi `df.isnull().sum()` sangat berguna dalam tahap Data Validation, karena menunjukkan data mana yang harus dibersihkan atau ditangani sebelum modeling.

- Gambar ini menunjukkan hasil visualisasi distribusi gender dalam dataset, menggunakan kode berikut:



Penjelasan Visualisasi:

Apa yang ditampilkan:

- Sumbu X (Gender): dua kategori — Male dan Female.
- Sumbu Y (count): jumlah masing-masing gender dalam dataset.
- Grafik ini adalah bar chart (diagram batang) yang menunjukkan frekuensi kemunculan setiap kategori gender.

Insight:

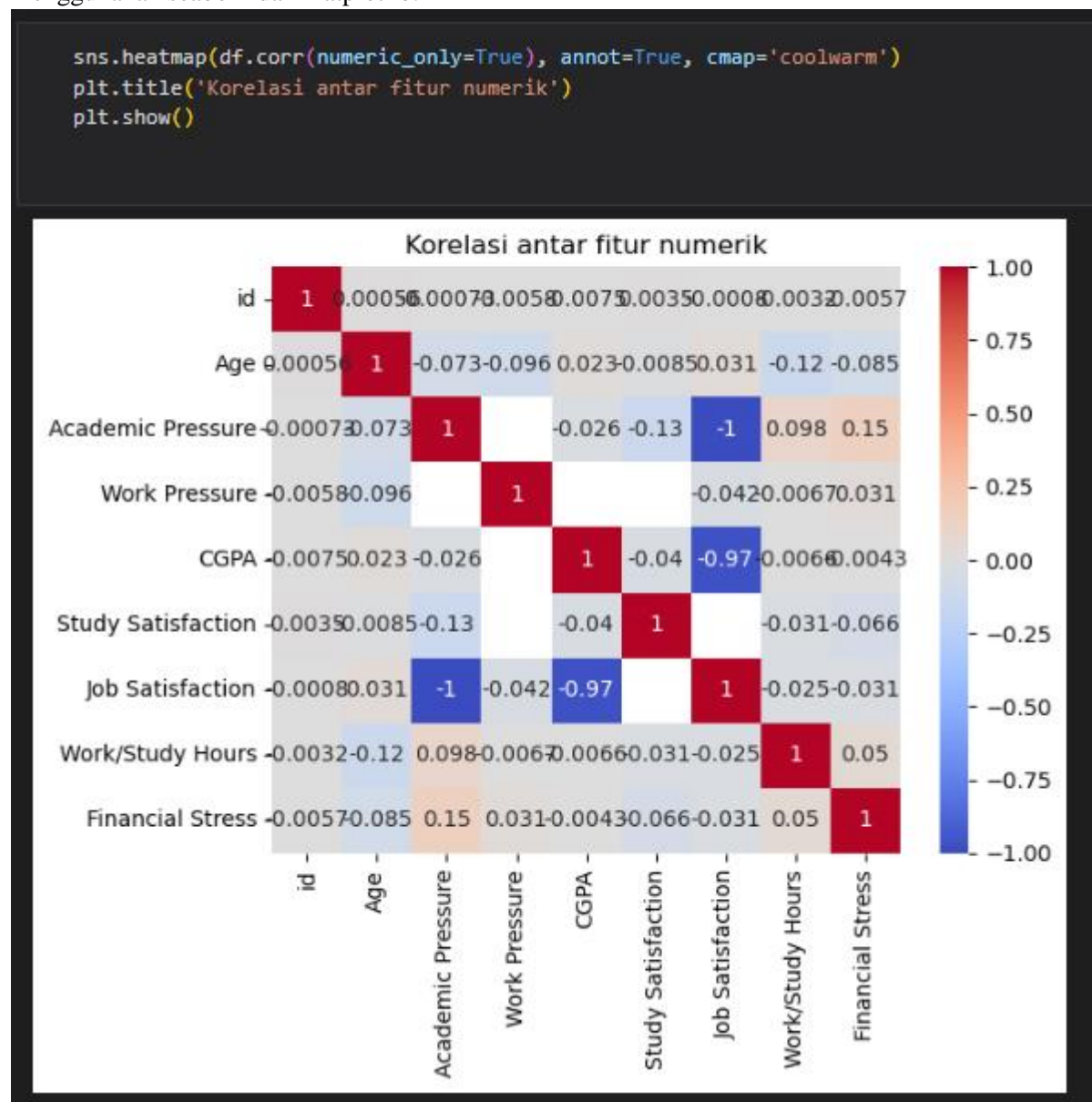
- Jumlah responden laki-laki (Male) lebih banyak dibanding perempuan (Female).
- Estimasi kasar dari grafik:
 - Male: ~52.000+
 - Female: ~43.000+
- Ini berarti komposisi gender tidak seimbang, tapi perbedaan tidak terlalu ekstrem.

Tujuan Visualisasi:

Visualisasi ini digunakan dalam tahap Describe Data untuk:

- Mengetahui komposisi kategori pada fitur Gender.
- Memastikan bahwa kolom Gender tidak memiliki nilai aneh (misal: selain Male/Female).
- Menjadi dasar jika ingin membuat analisis berdasarkan gender (misal: stres akademik antara laki-laki dan perempuan).

Gambar tersebut menunjukkan heatmap korelasi antar fitur numerik dari sebuah DataFrame menggunakan seaborn dan matplotlib.



Korelasi menunjukkan seberapa kuat dan searah dua variabel numerik saling berhubungan. Nilai korelasi (Pearson correlation coefficient) berkisar dari:

- +1 → Hubungan positif sempurna
- 0 → Tidak ada hubungan linear
- -1 → Hubungan negatif sempurna

Interpretasi Heatmap:

- Warna:
 - Merah → Korelasi positif (semakin terang, semakin kuat)
 - Biru → Korelasi negatif (semakin gelap, semakin kuat)
 - Abu-abu → Korelasi mendekati nol (hubungan lemah)
- Angka di dalam kotak menunjukkan nilai korelasi aktual.

Korelasi Penting dari Gambar:

- CGPA vs Job Satisfaction = -0.97
Korelasi negatif sangat kuat. Artinya, semakin tinggi CGPA, tingkat kepuasan kerja justru semakin rendah. Ini bisa jadi insight menarik—mungkin mahasiswa dengan nilai tinggi cenderung stres atau kurang puas dengan pekerjaan.
- Academic Pressure vs Job Satisfaction = -1.00
Korelasi negatif sempurna. Semakin tinggi tekanan akademik, semakin rendah kepuasan kerja. Sangat signifikan.
- Work Pressure vs Study Satisfaction = -0.04
Korelasi lemah negatif—tidak terlalu berarti.
- Financial Stress vs Academic Pressure = 0.15
Korelasi positif lemah. Bisa disimpulkan ada sedikit hubungan bahwa tekanan akademik meningkat saat tekanan finansial meningkat.

Fitur-fitur Tidak Relevan / Random:

- Kolom id tidak memiliki korelasi signifikan dengan fitur lain (karena memang cuma penanda identitas).
- Korelasi antar Age, Work/Study Hours, dan lainnya sebagian besar mendekati nol menunjukkan hubungan lemah.

Kesimpulan:

- Korelasi negatif kuat antara Academic Pressure, CGPA, dan Job Satisfaction adalah insight paling menonjol.
- Korelasi lemah lainnya menunjukkan banyak fitur yang mungkin independen satu sama lain.