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

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
Bertujuan untuk mengidentifikasi faktor-faktor yang memengaruhi risiko gangguan mental di kalangan pekerja. Selain itu, penelitian ini juga bertujuan untuk memberikan insight yang berguna bagi perusahaan dalam upaya membangun lingkungan kerja yang lebih sehat. Temuan dari penelitian ini diharapkan dapat membantu pihak *Human Resources* (HR) dan pembuat kebijakan dalam menyusun program dukungan kesehatan mental yang lebih efektif dan berbasis data. Lebih lanjut, penelitian ini juga akan menyediakan visualisasi yang informatif untuk membantu para pemangku kebijakan dalam memahami kondisi karyawan secara lebih komprehensif.

Tujuan *Data Science*
Penelitian ini juga bertujuan untuk membangun model klasifikasi yang dapat memprediksi apakah seseorang berisiko mengalami depresi. Untuk mendukung proses tersebut, dilakukan *Exploratory Data* Analysis (EDA) guna memahami distribusi data serta mengidentifikasi pola-pola tersembunyi yang mungkin berkaitan dengan risiko depresi. Selain itu, penelitian ini juga menghasilkan visualisasi yang dirancang untuk membantu para *stakeholder* dalam memahami temuan secara lebih jelas dan informatif.

1.3 Produce Project Plan

Dalam proyek ini, kriteria kesuksesan ditinjau dari dua aspek, yaitu teknis dan bisnis. Dari sisi teknis, model klasifikasi yang dikembangkan dianggap berhasil apabila mampu mencapai performa dengan akurasi lebih dari 60%, precision lebih dari 60%, dan recall lebih dari 60%. Selain itu, model juga harus berhasil dideploy secara online dan dapat menerima input langsung dari pengguna untuk keperluan prediksi.
Dari sisi bisnis, model diharapkan dapat dimanfaatkan sebagai alat bantu untuk deteksi dini terhadap risiko depresi, sehingga bisa berkontribusi dalam upaya pencegahan dan penanganan awal. Selain itu, hasil analisis yang diperoleh dari model juga diharapkan dapat meningkatkan kesadaran serta kepedulian terhadap isu kesehatan mental, khususnya di lingkungan kerja maupun pendidikan. Tak hanya itu, model juga harus mampu memberikan insight yang bermakna mengenai fitur-fitur apa saja yang paling berpengaruh terhadap risiko depresi, sehingga dapat menjadi bahan pertimbangan dalam pengambilan keputusan yang lebih tepat.

Namun, dalam pelaksanaan proyek ini terdapat beberapa kendala dan asumsi yang perlu diperhatikan. Salah satu kendala utama adalah bahwa data yang digunakan bersifat sintetis dan bukan berasal dari dunia nyata, sehingga interpretasi hasil harus dilakukan secara hati-hati dan tidak dapat dijadikan dasar untuk diagnosis medis. Selain itu, kemungkinan adanya nilai kosong (missing values) atau ketidakkonsistenan dalam beberapa fitur juga menjadi tantangan tersendiri yang memerlukan proses pembersihan dan pra-pemrosesan data. Dataset ini juga bersumber dari survei global yang mungkin tidak sepenuhnya mencerminkan kondisi lokal, seperti Indonesia. Meskipun demikian, dataset ini tetap relevan untuk tujuan pembelajaran dan pengembangan model. Perlu digarisbawahi bahwa model yang dibangun tidak ditujukan sebagai alat diagnostik medis, melainkan hanya sebagai tools pendukung untuk prediksi dan analisis awal dalam konteks kesehatan mental.

BAB 2

DATA UNDERSTANDING

2.1 Pengumpulan Data

Dataset yang digunakan dalam proyek ini bersumber dari [Kaggle Playground Series S4E11] dan bersifat sintetis, dirancang khusus untuk keperluan pembelajaran dan eksperimen dalam pengembangan model prediksi kesehatan mental. Dataset ini berformat CSV (*Comma Separated Values*), yang umum digunakan untuk menyimpan data dalam bentuk tabel. Terdapat dua bagian dalam dataset, yaitu *train set* dan *test set*. *Train set* terdiri dari 140.700 baris dan 20 kolom, termasuk kolom target bernama *Depression* yang menunjukkan kondisi seseorang dengan nilai 0 berarti tidak mengalami depresi dan 1 berarti mengalami depresi. Sementara itu, test set berisi 60.300 baris tanpa label *Depression*, sehingga digunakan untuk menguji kemampuan model dalam memprediksi kondisi mental pada data yang belum diketahui.

Persiapan dan Pemrosesan Awal

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

Interpretasi Line 1:

Baris kode import numpy as np, import matplotlib.pyplot as plt, dan import seaborn as sns merupakan langkah awal yang umum dalam proses analisis data dan visualisasi menggunakan Python. Library NumPy (np) digunakan untuk menangani array dan operasi matematis tingkat lanjut secara efisien. Ini sangat berguna saat kita bekerja dengan data numerik dalam skala besar. Selanjutnya, Matplotlib (plt) adalah library visualisasi dasar yang memungkinkan pengguna untuk membuat berbagai jenis grafik, seperti garis, batang, dan sebar. Sementara itu, Seaborn (sns) adalah library visualisasi berbasis Matplotlib yang menyediakan tampilan grafik yang lebih menarik secara estetika dan lebih informatif, terutama untuk visualisasi statistik seperti distribusi data dan hubungan antar variabel. Ketiga library ini sering digunakan bersama-sama untuk eksplorasi data, analisis statistik, serta pembuatan visualisasi yang menarik dan komunikatif.

Untuk melakukan eksplorasi dan visualisasi data, digunakan beberapa library Python berikut:
numpy untuk manipulasi angka dan array, matplotlib.pyplot untuk visualisasi dasar seperti histogram dan scatter plot, seaborn untuk visualisasi lanjutan dengan tampilan yang lebih menarik.

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

# Menampilkan 5 baris pertama
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 ?
0	140700	Shivam	Male	53.0	Visakhapatnam	Working Professional	Judge	NaN	2.0	NaN	NaN	5.0	Less than 5 hours	Moderate	LLB	No
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
2	140702	Yash	Male	53.0	Jaipur	Working Professional	Teacher	NaN	4.0	NaN	NaN	1.0	7-8 hours	Moderate	B.Arch	Yes
3	140703	Nalini	Female	23.0	Rajkot	Student	NaN	5.0	NaN	6.84	1.0	NaN	More than 8 hours	Moderate	BSc	Yes
4	140704	Shaurya	Male	47.0	Kalyan	Working Professional	Teacher	NaN	5.0	NaN	NaN	5.0	7-8 hours	Moderate	BCA	Yes

Dataset dimuat menggunakan fungsi pandas.read_csv() dari file Book1.csv yang sebelumnya diunggah ke Google Drive. Metode ini umum digunakan untuk mengimpor data tabular ke dalam lingkungan Python notebook seperti Google Colab atau Jupyter Notebook.

2.2 Describe Data

Tahap ini bertujuan untuk memahami struktur dan karakteristik awal dari dataset yang akan digunakan dalam pelatihan dan pengujian model. Dataset terdiri dari gabungan variabel kategorikal dan numerik, mencakup informasi demografis, kebiasaan harian, tekanan akademik atau pekerjaan, serta aspek psikologis yang berkaitan dengan kondisi mental.

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

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

Fungsi data.info() memberikan informasi lebih detail terkait jumlah baris, kolom, serta tipe data pada setiap kolom. Dari hasil ini, ditemukan bahwa beberapa kolom memiliki nilai kosong atau *missing values*, seperti *Academic Pressure*, *CGPA*, *Study Satisfaction*, dan *Job Satisfaction*, yang hanya terisi sekitar 20% dari total data. Sementara itu, kolom *Work Pressure* memiliki data yang lebih lengkap, yaitu sekitar 80%. Kolom-kolom lain seperti *Age*, *Gender*, *City*, dan *Degree* tercatat lengkap tanpa nilai yang hilang.

```
<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
```

Eksplorasi awal terhadap dataset dilakukan menggunakan beberapa fungsi dasar. Fungsi data.head() digunakan untuk menampilkan lima baris pertama dari dataset, sehingga dapat memberikan gambaran umum mengenai isi dan struktur data.

	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	75026.00000	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

Dari segi tipe data, kolom bertipe object umumnya menyimpan data kategorikal seperti *Gender*, *City*, dan *Profession*, sedangkan data numerik seperti *Age*, *CGPA*, dan *Financial Stress* direpresentasikan dalam tipe float64 atau int64. Untuk melihat ringkasan statistik dari kolom-kolom numerik, digunakan fungsi data.describe(), yang menghasilkan informasi seperti nilai rata-rata (mean), standar deviasi, serta nilai minimum dan maksimum. Informasi ini berguna untuk memahami sebaran nilai dan potensi keberadaan outlier pada data numerik yang tersedia.

2.3 Validation Data

Validasi data dilakukan untuk memastikan kualitas dan konsistensi dataset sebelum masuk ke tahap pemodelan. Beberapa aspek utama yang divalidasi meliputi nilai kosong (*missing values*), keberadaan outlier atau anomali, duplikasi data, serta distribusi variabel target.

1. Missing Values

```
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
```

Identifikasi nilai kosong dilakukan menggunakan fungsi df.isnull().sum(), yang menunjukkan jumlah nilai kosong di setiap kolom. Beberapa kolom penting ditemukan memiliki proporsi data kosong yang signifikan. Kolom seperti *id*, *Name*, *Gender*, *Age*, *City*, *Work/Study Hours*, *Financial Stress*, dan *Family*

History of Mental Illness tidak memiliki nilai kosong, sehingga tidak memerlukan penanganan khusus. Namun, kolom *Profession* memiliki sekitar 24.632 nilai kosong (sekitar 26%), kemungkinan disebabkan oleh banyaknya responden yang masih berstatus pelajar dan belum bekerja.

Kolom seperti *Academic Pressure*, *CGPA*, dan *Study Satisfaction* hanya terisi sekitar 18.000 data dari total 93.800, mengindikasikan bahwa informasi ini kemungkinan hanya relevan atau tersedia untuk sebagian responden, khususnya mahasiswa. Demikian pula, kolom *Work Pressure* dan *Job Satisfaction* memiliki sekitar 18.774–18.778 data terisi, yang dapat diasumsikan hanya diisi oleh responden yang telah bekerja. Kolom lain seperti *Sleep Duration*, *Degree*, dan *Have you ever had suicidal thoughts?* memiliki jumlah nilai kosong yang sangat kecil (2–5 data), sehingga dapat ditangani dengan imputasi sederhana atau penghapusan baris.

2. Outlier dan Anomali

Pemeriksaan terhadap outlier dilakukan terutama pada kolom numerik seperti *Age*, yang memiliki rentang nilai antara 18 hingga 60 tahun. Rentang ini masih dalam batas wajar dan tidak menunjukkan outlier ekstrem. Namun, terdapat beberapa nilai anomali dalam kolom kategorikal, seperti *Sleep Duration* dan *Gender*, yang mengandung kategori di luar nilai standar yang diharapkan (misalnya, nilai selain "Male" atau "Female").

3. Data Duplikat

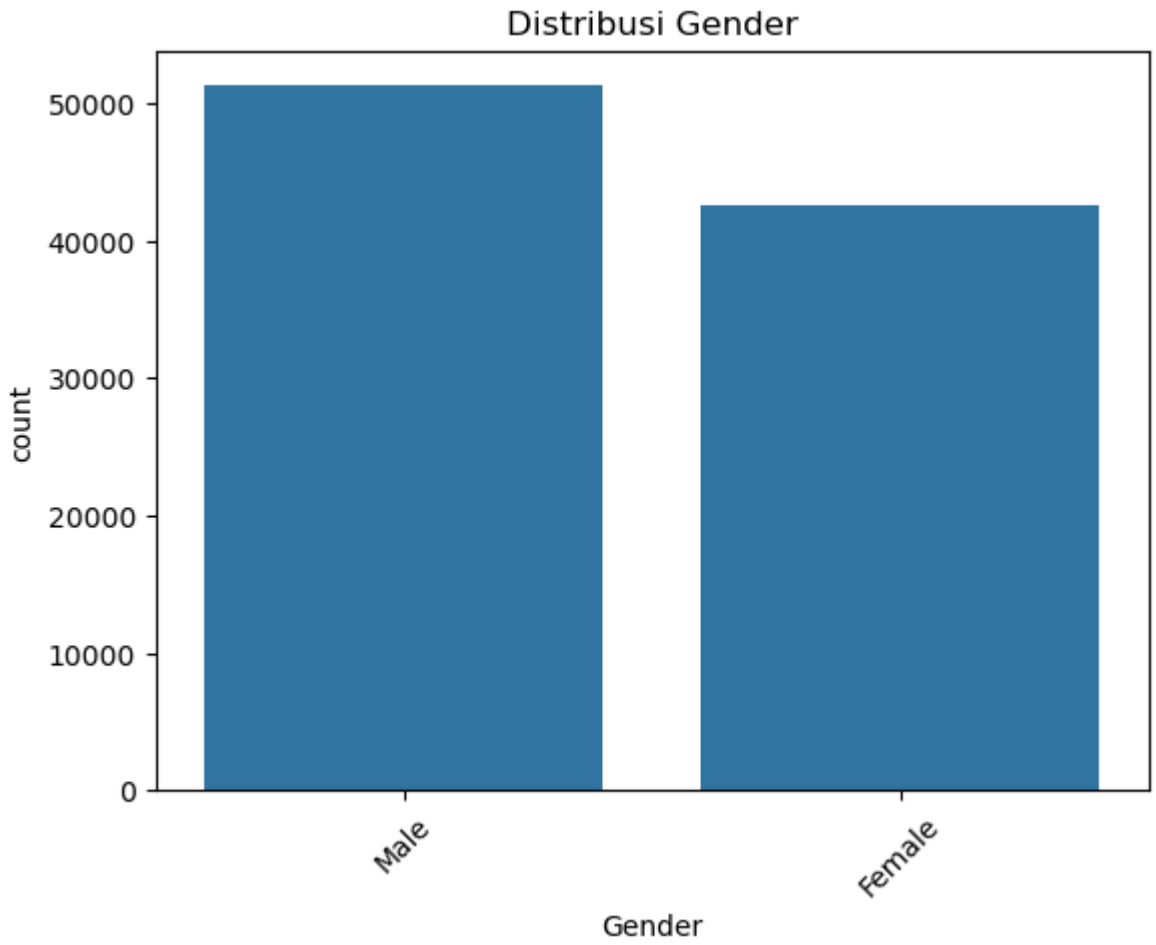
Kolom *id* digunakan sebagai identifikasi unik untuk setiap entri data. Dari hasil pemeriksaan, tidak ditemukan duplikasi signifikan pada kolom ini, sehingga setiap baris dianggap mewakili individu yang berbeda.

4. Distribusi Variabel Target

Distribusi variabel target *Depression* dalam train set menunjukkan ketidakseimbangan (*imbalanced*). Hanya sekitar 18% dari total data yang memiliki nilai 1 (mengalami depresi), sedangkan sisanya memiliki nilai 0. Ketidakseimbangan ini penting untuk diperhatikan karena dapat memengaruhi performa model klasifikasi, terutama jika metrik evaluasi tidak disesuaikan.

5. Visualisasi Distribusi Kategorikal

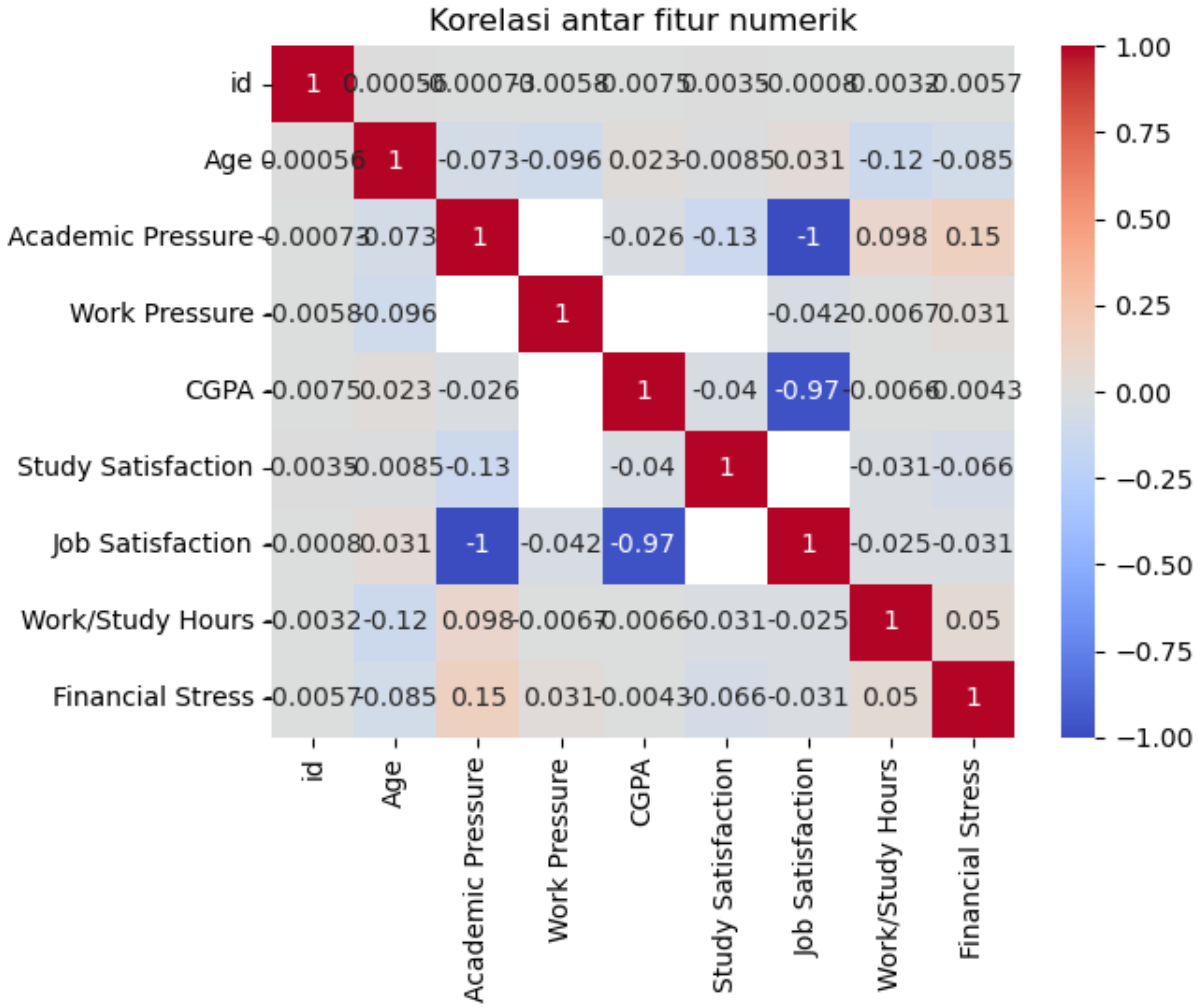
```
sns.countplot(data=df, x='Gender')
plt.xticks(rotation=45)
plt.title('Distribusi Gender')
plt.show()
```



Distribusi kategori *Gender* divisualisasikan dalam bentuk bar chart. Grafik ini menampilkan dua kategori utama yaitu Male dan Female pada sumbu X, dengan jumlah masing-masing pada sumbu Y. Hasil visualisasi menunjukkan bahwa jumlah responden laki-laki lebih banyak dibanding perempuan, dengan estimasi sekitar 52.000 laki-laki dan 43.000 perempuan. Meskipun terdapat ketidakseimbangan, perbedaan ini tidak terlalu ekstrem. Visualisasi ini juga membantu memastikan bahwa tidak ada nilai aneh dalam kolom Gender.

6. Korelasi antar Fitur Numerik

```
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Korelasi antar fitur numerik')
plt.show()
```



Untuk mengevaluasi hubungan antar fitur numerik, digunakan visualisasi heatmap korelasi yang menunjukkan nilai koefisien korelasi Pearson antara setiap pasangan fitur. Beberapa temuan penting dari heatmap tersebut antara lain:

- Korelasi negatif sangat kuat antara *CGPA* dan *Job Satisfaction* (-0.97), serta antara *Academic Pressure* dan *Job Satisfaction* (-1.00). Hal ini menunjukkan bahwa nilai akademik yang tinggi dan tekanan akademik yang besar cenderung diikuti oleh kepuasan kerja yang rendah.
- Korelasi lemah ditemukan antara *Work Pressure* dan *Study Satisfaction* (-0.04), serta *Financial Stress* dan *Academic Pressure* (0.15).
- Sebagian besar fitur lain menunjukkan korelasi rendah, seperti *Age*, *Work/Study Hours*, dan *Financial Stress*, yang mengindikasikan bahwa variabel-variabel ini cenderung berdiri sendiri atau memiliki pengaruh kecil terhadap fitur lainnya.

BAB 3

DATA PREPARATION

3.1 Data Selection

3.2 Data Cleaning

3.3 Data Construct

3.4 Labeling Data

3.5 Data Integration

BAB 4

MODELLING

4.1 Build Model

BAB 5

EVALUATION

BAB 6

DEPLOYMENT

DAFTAR PUSTAKA