

**LAPORAN**  
**KECERDASAN BUATAN**  
*PREDIKSI DEPRESI MAHASISWA MENGGUNAKAN ALGORITMA K-NEAREST  
NEIGHBOR (KNN)*



Disusun oleh:

Muhamad Rijki Nurjakiah (2306044)

Mochamad Risyad Fauzan (2306038)

Dosen Pengampu Mata Kuliah:

Leni Fitriani, S.Kom, M.Kom

**INSTITUT TEKNOLOGI GARUT**  
**JURUSAN ILMU KOMPUTER**  
**PROGRAM STUDI TEKNIK INFORMATIKA**  
**TAHUN AKADEMIK 2024/2025**

## **1. BUSINESS UNDERSTANDING**

### **a) Permasalahan dunia nyata**

Depresi merupakan gangguan mental yang semakin banyak dialami oleh mahasiswa di Indonesia. Kondisi ini tidak hanya mengganggu kehidupan pribadi dan sosial, tetapi juga berdampak serius terhadap performa akademik. Berbagai penelitian di Indonesia menunjukkan bahwa mahasiswa berada dalam kelompok yang sangat rentan terhadap tekanan psikologis, terutama akibat stres akademik, kelelahan, dan faktor psikososial lainnya.

Menurut penelitian oleh Putri et al. (2022), depresi pada remaja dan mahasiswa dapat dipicu oleh berbagai faktor seperti bullying, hubungan yang buruk dengan orang tua, tekanan psikososial, dan masalah ekonomi. Studi tersebut juga menegaskan bahwa depresi sering tidak dikenali karena gejalanya samar, seperti kehilangan minat, kesulitan tidur, dan merasa tidak berharga. Bahkan, depresi yang tidak tertangani bisa mengarah pada keinginan bunuh diri, yang terjadi pada 18,6% remaja di DKI Jakarta (Putri et al., 2022).

Selanjutnya, Furi et al. (2024) menyoroti bahwa kelelahan akibat beban akademik yang berat menjadi salah satu pemicu utama depresi pada mahasiswa kedokteran. Dalam penelitian di Universitas Palangka Raya, sebanyak 85,11% mahasiswa yang mengalami depresi juga mengalami kelelahan signifikan, dan uji statistik menunjukkan adanya hubungan yang kuat dan signifikan antara kelelahan dan tingkat depresi ( $p = 0,001$ ) (At & Raya, 2024).

Studi lain oleh Faizah et al. (2021) di Universitas Mulawarman menemukan bahwa lebih dari 51,4% mahasiswa kedokteran menunjukkan gejala depresi dari berbagai tingkat, mulai dari ringan hingga berat. Sumber stres yang mendasari antara lain adalah tekanan akademik, ketidakpuasan terhadap sistem pembelajaran, serta masalah keuangan dan sosial. Studi ini juga menekankan bahwa tekanan yang tidak ditangani dengan baik bisa memunculkan gejala depresi yang serius dan mempengaruhi kesejahteraan psikologis mahasiswa (Faizah et al., 2021).

Dengan melihat berbagai temuan tersebut, dapat disimpulkan bahwa deteksi dini depresi pada mahasiswa sangat penting dilakukan. Sistem berbasis kecerdasan buatan (AI), seperti model prediksi menggunakan algoritma K-Nearest Neighbors (KNN), dapat menjadi alat bantu yang efektif untuk memetakan mahasiswa yang berisiko, sehingga institusi pendidikan dapat mengambil tindakan preventif lebih cepat dan tepat sasaran.

## **b) Tujuan Proyek**

Proyek ini bertujuan untuk merancang dan mengimplementasikan model kecerdasan buatan yang mampu melakukan klasifikasi tingkat depresi pada mahasiswa berdasarkan data survei yang mencakup faktor-faktor psikologis, gaya hidup, dan sosial. Dengan memanfaatkan pendekatan machine learning, khususnya algoritma K-Nearest Neighbors (KNN), sistem ini diharapkan dapat:

1. Mendeteksi secara dini gejala atau kecenderungan depresi pada mahasiswa berdasarkan data objektif yang diperoleh melalui survei.
2. Mengidentifikasi faktor-faktor dominan yang berkontribusi terhadap tingkat depresi, seperti tekanan akademik, pola tidur, stres keuangan, dan riwayat keluarga.
3. Memberikan rekomendasi berbasis data kepada pihak terkait (seperti psikolog kampus atau dosen wali) untuk melakukan intervensi lebih lanjut terhadap mahasiswa yang berisiko tinggi mengalami depresi.
4. Mengurangi ketergantungan pada asesmen manual yang memerlukan waktu dan sumber daya besar dengan menyediakan sistem prediksi otomatis berbasis data.

Secara keseluruhan, tujuan akhir dari proyek ini adalah untuk membuktikan bahwa machine learning dapat menjadi alat bantu yang efektif dan efisien dalam mendukung kesejahteraan mental mahasiswa melalui deteksi dini dan pengambilan keputusan berbasis data.

## **c) User/Pengguna sistem**

Sistem pendeteksi potensi depresi yang dibangun dalam proyek ini ditujukan untuk beberapa jenis pengguna, yaitu:

### **1. Mahasiswa/Siswa**

Mahasiswa sebagai pengguna utama dapat memanfaatkan sistem ini untuk melakukan deteksi awal terhadap kondisi kesehatan mental mereka, khususnya terkait gejala depresi. Sistem ini dapat menjadi alat bantu refleksi diri serta meningkatkan kesadaran akan pentingnya kesehatan mental.

### **2. Dosen Pembimbing / Guru / Staf Akademik**

Para tenaga pendidik dan pembimbing akademik dapat menggunakan sistem ini sebagai alat bantu pemantauan terhadap kondisi psikologis mahasiswa/siswa, sehingga dapat memberikan dukungan atau rekomendasi penanganan lebih lanjut jika ditemukan indikasi masalah kesehatan mental.

### **3. Psikolog / Konselor Sekolah atau Kampus**

Sistem ini dapat digunakan sebagai alat skrining awal yang membantu proses konseling lebih efisien dan terarah. Dengan adanya data awal dari sistem, konselor dapat memprioritaskan kasus dan menentukan langkah intervensi yang sesuai.

### **4. Manajemen Sekolah / Perguruan Tinggi**

Manajemen lembaga pendidikan dapat menggunakan hasil dari sistem ini sebagai bagian dari evaluasi kondisi psikologis peserta didik secara umum, serta dasar dalam merancang program kesehatan mental atau kebijakan pendukung lainnya di lingkungan akademik.

## **d) Manfaat Implementasi AI**

Implementasi kecerdasan buatan (Artificial Intelligence/AI), khususnya melalui algoritma machine learning, memberikan berbagai manfaat signifikan dalam upaya deteksi dan pencegahan gangguan kesehatan mental, terutama pada mahasiswa. Berdasarkan hasil kajian dari beberapa jurnal ilmiah, manfaat tersebut meliputi:

### **1. Deteksi Dini dan Pencegahan Gangguan Mental**

Algoritma seperti **K-Nearest Neighbor (K-NN)** mampu mengidentifikasi gejala depresi berdasarkan pola data yang tidak mudah terlihat secara konvensional. Hal ini sangat bermanfaat dalam mendeteksi risiko gangguan mental sejak dini, sebelum gejalanya berkembang menjadi lebih serius (Nurdiansyah et al., 2025).

### **2. Peningkatan Akurasi Diagnosis**

Model pembelajaran mesin seperti **Decision Tree**, **K-NN**, dan **SVM** menunjukkan performa tinggi dalam memproses data kesehatan mental dengan akurasi hingga lebih dari 80%. Dengan akurasi tersebut, AI membantu memberikan hasil prediksi yang lebih konsisten dan objektif dibandingkan metode manual, serta mengurangi risiko bias dari subjektivitas manusia (Muthia & Putra, n.d.).

### **3. Efisiensi dalam Pengolahan Data Kompleks**

AI mampu menangani dataset besar yang mencakup berbagai fitur psikologis seperti stres akademik, kualitas tidur, dan dukungan sosial. Proses ini memungkinkan analisis multidimensi terhadap kondisi psikologis mahasiswa secara cepat dan efisien (Setiawan et al., 2025).

#### 4. **Pengambilan Keputusan Lebih Cepat dan Terarah**

Model AI dapat diintegrasikan dalam sistem informasi institusi pendidikan untuk secara otomatis memberi peringatan dini terhadap siswa yang berisiko mengalami depresi. Hal ini mempercepat proses penanganan dan memudahkan pihak kampus atau konselor dalam menentukan langkah lanjut yang tepat (Setiawan et al., 2025).

#### 5. **Dasar Pengambilan Kebijakan Institusi Pendidikan**

Dengan hasil analisis berbasis AI, institusi pendidikan dapat memperoleh **data dan pola yang akurat** untuk menyusun kebijakan yang lebih strategis dan tepat sasaran dalam mendukung kesehatan mental mahasiswa (Setiawan et al., 2025).

#### 6. **Kemudahan Integrasi Teknologi**

Sistem berbasis AI dapat diintegrasikan dengan teknologi terkini seperti **cloud computing** dan **big data**, memungkinkan pengelolaan data secara lebih fleksibel, aman, dan berskala luas (Nurdiansyah et al., 2025).

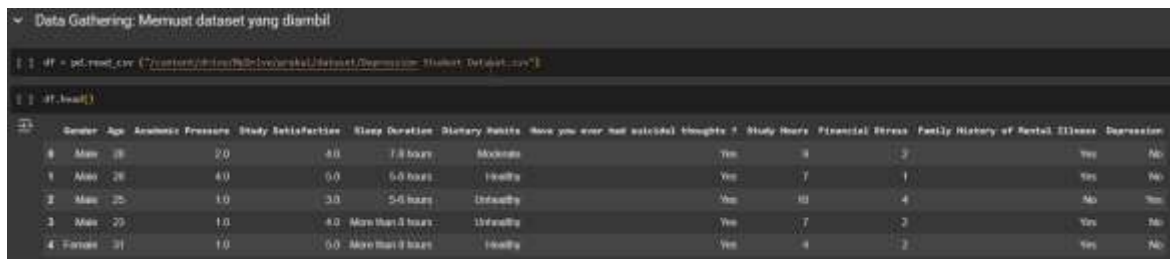
## 2. RINGKASAN ARTIKEL PENELITIAN

| No | Judul Penelitian  | Metode Penelitian                                 | Dataset                                  | Temuan Utama   | Keterbatasan  |
|----|---|---|--|--|---|
| 1  | Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-NN dan Random Forest              | K-NN, Random Forest, EDA, Preprocessing, Evaluasi | Student Mental Health.csv (Kaggle)       | K-NN akurasi 90% (80:20), Random Forest 85%. K-NN lebih unggul di akurasi, RF lebih konsisten.   | Kinerja menurun saat pembagian data berbeda; hanya menggunakan satu sumber dataset. |
| 2  | Perbandingan Akurasi Model Pembelajaran Mesin SVM, KNN, Decision Tree, dan Naive Bayes pada Klasifikasi Gangguan Kesehatan Mental | K-NN, SVM, Decision Tree, Naive Bayes             | Dataset gejala mental (120 data, Kaggle) | Decision Tree akurasi tertinggi (86.67%), KNN akurasi 66.67%. Decision Tree unggul untuk data heterogen, KNN sederhana tapi kurang akurat. | Dataset terbatas; akurasi KNN relatif rendah dibanding model lain.                  |
| 3  | Komparasi Kinerja Algoritma Random Forest, Decision Tree, Naive Bayes, dan KNN dalam Prediksi Tingkat Depresi Mahasiswa           | CRISP-DM, klasifikasi (termasuk KNN)              | Student Depression Dataset (Kaggle)      | Random Forest dan XGBoost akurasi tinggi (~83-84%). KNN akurasi stabil (81%), meskipun tidak seunggul model lain seperti RF atau NB.       | KNN memiliki recall dan F1-score lebih rendah di beberapa fold evaluasi.            |

### 3. DATA UNDERSTANDING

#### a) Sumber dataset

Dataset ini diperoleh dari repositori GitHub *fatemeh-mndz* dengan judul "**Depression Student Dataset**", yang berisi data mengenai tekanan akademik, kepuasan belajar, kebiasaan tidur, stres finansial, dan gejala depresi pada mahasiswa. Dataset dapat diakses melalui tautan: *fatemeh-mndz*. (2022). *Depression Student Dataset* [Data set]. GitHub. <https://github.com/fatemeh-mndz/Depression-Student>



|   | Gender | Age | Academic Pressure | Study Satisfaction | Sleep Duration    | Dietary Habits | Have you ever had suicidal thoughts? | Study Hours | Financial Stress | Family History of Mental Illness | Depression |
|---|--------|-----|-------------------|--------------------|-------------------|----------------|--------------------------------------|-------------|------------------|----------------------------------|------------|
| 0 | Male   | 18  | 2.0               | 4.0                | 7-8 hours         | Unhealthy      | Yes                                  | 8           | 2                | Yes                              | No         |
| 1 | Male   | 20  | 4.0               | 5.0                | 5-6 hours         | Healthy        | Yes                                  | 7           | 1                | Yes                              | No         |
| 2 | Male   | 25  | 1.0               | 3.0                | 5-6 hours         | Unhealthy      | No                                   | 10          | 4                | No                               | Yes        |
| 3 | Male   | 20  | 1.0               | 4.0                | More than 8 hours | Unhealthy      | Yes                                  | 7           | 2                | Yes                              | No         |
| 4 | Female | 31  | 1.0               | 5.0                | More than 8 hours | Healthy        | Yes                                  | 8           | 2                | Yes                              | No         |

Gambar 1

#### b) Ukuran dan format data

Dataset yang digunakan dalam penelitian ini berjudul "Depression Student Dataset" dan berformat CSV (Comma-Separated Values) dengan ukuran file sekitar 7 KB. Dataset ini terdiri dari 102 baris data yang mewakili responden mahasiswa, serta memiliki 11 kolom (fitur) yang merepresentasikan berbagai faktor yang mungkin berpengaruh terhadap kondisi depresi pada mahasiswa.

Setiap kolom dalam dataset memuat jenis data yang bervariasi, mulai dari data kategori hingga numerik. Fitur-fitur kategorikal antara lain adalah Gender, Sleep Duration, Dietary Habits, Have you ever had suicidal thoughts?, dan Family History of Mental Illness, yang berisi nilai-nilai seperti "Male", "Female", "7-8 hours", "Healthy", "Yes", dan "No". Sementara itu, fitur numerik seperti Age, Academic Pressure, Study Satisfaction, Study Hours, dan Financial Stress diukur dalam bentuk angka, baik dalam skala penilaian (1–5) maupun hitungan jam dan usia.

Label atau target dari dataset ini adalah kolom Depression, yang merupakan data kategorikal dengan dua kemungkinan nilai, yaitu Yes dan No, yang menunjukkan apakah seorang responden mengalami gejala depresi atau tidak. Kombinasi dari berbagai fitur ini memungkinkan analisis lebih lanjut untuk membangun model klasifikasi yang dapat memprediksi kecenderungan depresi pada mahasiswa berdasarkan faktor-faktor tertentu.

c) Tipe atribut

| No | Nama Atribut                          | Tipe Data         | Deskripsi   |
|----|---------------------------------------|-------------------|---|
| 1  | Gender                                | Kategorik         | Jenis kelamin responden. Contoh: Male, Female.  |
| 2  | Age                                   | Numerik           | Usia responden dalam tahun.   |
| 3  | Academic Pressure                     | Skala (1–5)       | Tingkat tekanan akademik yang dirasakan mahasiswa, dari rendah (1) ke tinggi (5).                     |
| 4  | Study Satisfaction                    | Skala (1–5)       | Tingkat kepuasan terhadap proses belajar. Semakin tinggi angkanya, semakin puas.                      |
| 5  | Sleep Duration                        | Kategori          | Durasi tidur rata-rata per malam. Contoh: Less than 5 hours, 5-6 hours, 7-8 hours, More than 8 hours. |
| 6  | Dietary Habits                        | Kategori          | Kualitas kebiasaan makan. Contoh: Healthy, Moderate, Unhealthy.                                       |
| 7  | Have you ever had suicidal thoughts ? | Kategori          | Apakah pernah memiliki pikiran untuk bunuh diri. Nilai: Yes, No.                                      |
| 8  | Study Hours                           | Numerik           | Rata-rata jam belajar per hari.   |
| 9  | Financial Stress                      | Skala (1–5)       | Tingkat stres keuangan, dari 1 (tidak stres) sampai 5 (sangat stres).                                 |
| 10 | Family History of Mental Illness      | Kategori          | Apakah memiliki riwayat keluarga dengan gangguan mental. Nilai: Yes, No.                              |
| 11 | Depression                            | Kategori (Target) | Label target, menunjukkan apakah responden mengalami depresi. Nilai: Yes, No.                         |

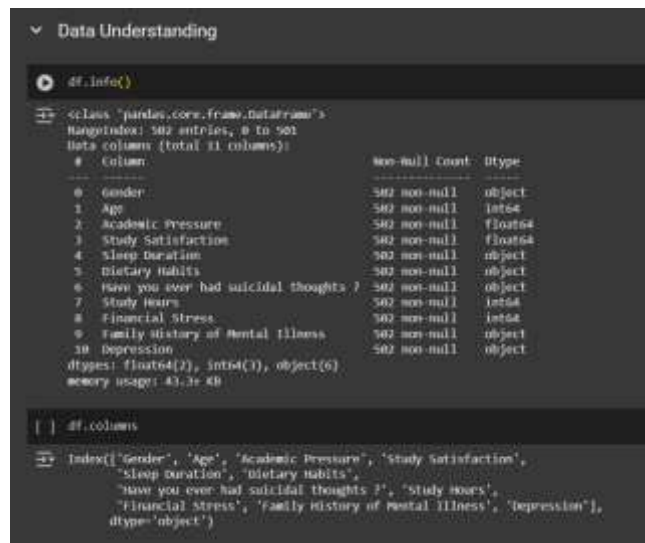


#### d) Tipe data dan target klasifikasi

Dataset ini terdiri dari dua jenis tipe data utama, yaitu data kategorikal dan data numerik. Fitur-fitur kategorikal mencakup Gender, Sleep Duration, Dietary Habits, Have you ever had suicidal thoughts?, serta Family History of Mental Illness, yang berisi data dalam bentuk teks atau label diskrit seperti "Male", "Yes", atau "Healthy". Fitur-fitur ini biasanya diubah ke dalam format numerik (melalui encoding) sebelum digunakan dalam pemodelan.

Sementara itu, fitur numerik dalam dataset ini meliputi Age, Academic Pressure, Study Satisfaction, Study Hours, dan Financial Stress. Beberapa dari fitur ini bersifat kuantitatif murni seperti Age dan Study Hours, sedangkan lainnya berupa skala ordinal (contohnya Academic Pressure dan Financial Stress yang dinilai dari 1 sampai 5), yang menunjukkan tingkat kondisi tertentu.

Adapun target klasifikasi dari dataset ini adalah kolom Depression, yang merupakan fitur kategorikal biner. Nilainya terdiri dari dua label: Yes dan No, yang merepresentasikan apakah seorang mahasiswa mengalami gejala depresi atau tidak. Model klasifikasi yang dibangun nantinya akan berfokus untuk memprediksi nilai dari kolom ini berdasarkan fitur-fitur lainnya yang tersedia.



```
ds.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 502 entries, 0 to 501
Data columns (total 11 columns):
 #   Column                                Non-Null Count  Dtype
---  --
 0   Gender                                502 non-null    object
 1   Age                                    502 non-null    int64
 2   Academic Pressure                     502 non-null    float64
 3   Study Satisfaction                    502 non-null    float64
 4   Sleep Duration                        502 non-null    object
 5   Dietary Habits                       502 non-null    object
 6   Have you ever had suicidal thoughts?  502 non-null    object
 7   Study Hours                           502 non-null    int64
 8   Financial Stress                      502 non-null    int64
 9   Family History of Mental Illness      502 non-null    object
10  Depression                            502 non-null    object
dtypes: float64(2), int64(3), object(6)
memory usage: 43.3+ KB

ds.columns

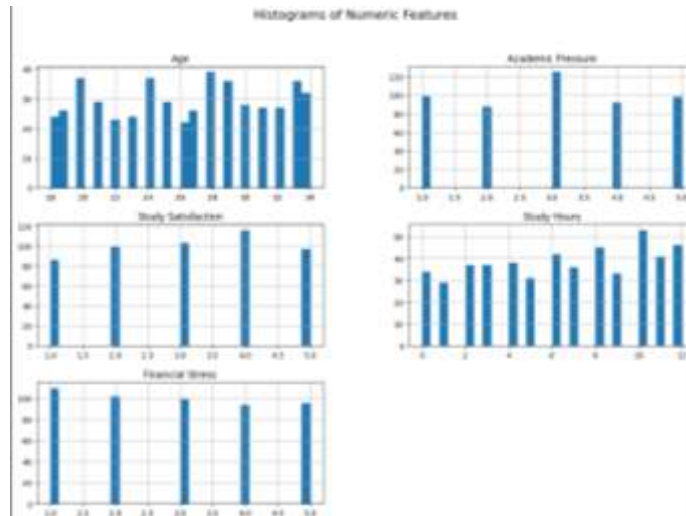
Index(['Gender', 'Age', 'Academic Pressure', 'Study Satisfaction',
       'Sleep Duration', 'Dietary Habits',
       'Have you ever had suicidal thoughts?', 'Study Hours',
       'Financial Stress', 'Family History of Mental Illness', 'Depression'],
      dtype='object')
```

Gambar 2

#### 4. *EXPLORATORY DATA ANALYSIS (EDA)*

##### a) Visualisasi distribusi data

##### Menampilkan Histogram



Gambar 3

Secara keseluruhan, histogram-histogram ini memberikan gambaran awal tentang karakteristik data Anda untuk fitur-fitur numerik yang dipilih. Anda dapat melihat:

- Distribusi Usia: Cukup merata dengan beberapa puncak.
- Tekanan Akademik, Kepuasan Belajar, dan Stres Keuangan: Cenderung memiliki distribusi yang mengelompok pada setiap nilai skala, menunjukkan bahwa ada kelompok responden yang jelas pada setiap tingkat. Ini mungkin mengindikasikan bahwa fitur-fitur ini adalah variabel kategorikal ordinal yang direpresentasikan secara numerik.
- Jam Belajar: Lebih menyebar, menunjukkan variasi dalam kebiasaan belajar.

Berikut adalah visualisasi distribusi data untuk fitur-fitur kategorikal dalam dataset. Grafik batang (bar chart) tersebut menunjukkan jumlah responden untuk masing-masing kategori pada setiap fitur, seperti jumlah mahasiswa berdasarkan gender, Riwayat keluarga, dan kondisi depresi.



Gambar 4



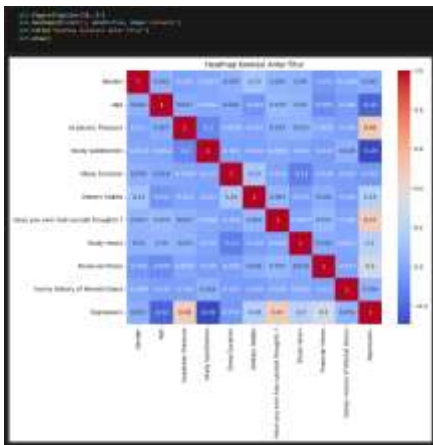
Gambar 5



Gambar 6

### **b) Analisis korelasi antar fitur**

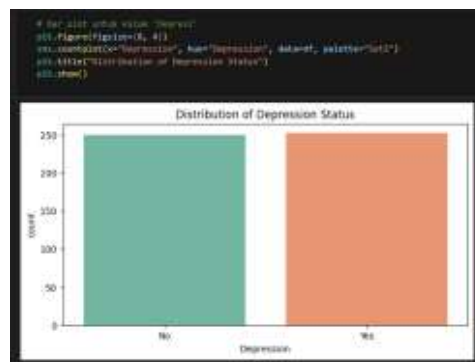
Berdasarkan hasil analisis korelasi antar fitur, ditemukan bahwa depresi paling berkorelasi dengan tekanan akademik (0.48), pikiran bunuh diri (0.47), dan kepuasan belajar (-0.29). Hal ini menunjukkan bahwa semakin tinggi tekanan akademik dan frekuensi pikiran bunuh diri, maka kemungkinan mengalami depresi juga meningkat, sementara kepuasan belajar yang tinggi justru cenderung menurunkan risiko depresi. Korelasi lainnya seperti usia (-0.22), stres keuangan (0.30), dan pola makan (0.19) juga memberikan kontribusi, meskipun lebih rendah. Sebagian besar fitur lainnya memiliki korelasi rendah satu sama lain, yang mengindikasikan bahwa tidak ada multikolinearitas kuat antar variabel.



### Gambar 7

**c) Deteksi data tidak seimbang**

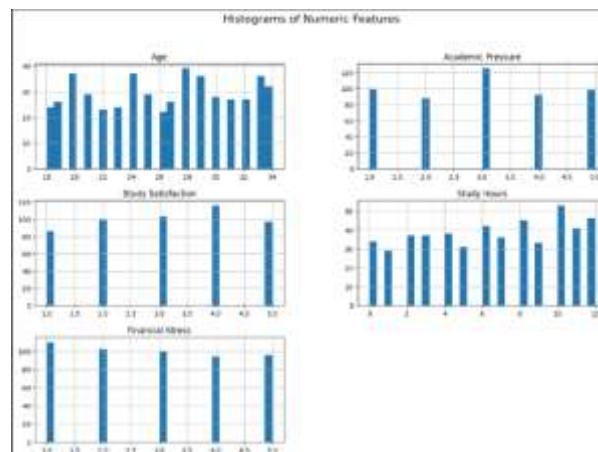
Distribusi data pada target klasifikasi *Depression* tergolong seimbang, dengan proporsi kelas "Yes" sebesar 50,2% dan "No" sebesar 49,8%. Hal ini menunjukkan bahwa tidak terdapat ketimpangan signifikan antara jumlah data pada masing-masing kelas, sehingga model klasifikasi yang akan dibangun tidak memerlukan penanganan khusus terhadap data tidak seimbang.



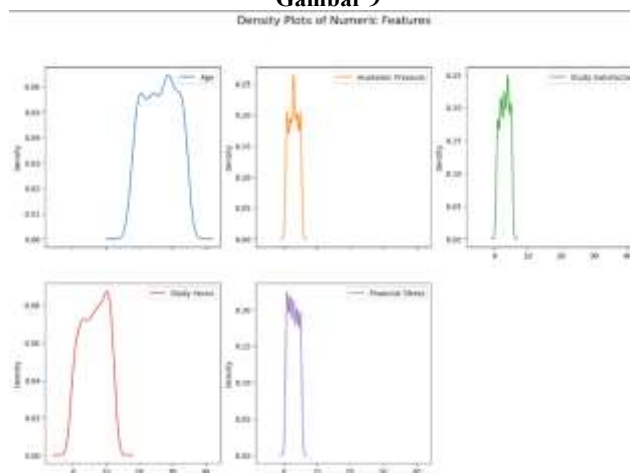
### Gambar 8

#### d) Insight awal dari pola data

Berdasarkan histogram dan density plot dari fitur-fitur numerik, dapat ditarik beberapa insight awal mengenai pola data. Distribusi usia mahasiswa cenderung bimodal, dengan sebagian besar berada di rentang 20 hingga 30 tahun dan puncak di sekitar usia 22-23 serta 28-29 tahun. Tekanan akademik dan kepuasan belajar menunjukkan distribusi yang relatif merata di antara nilai 1 hingga 5, meskipun ada sedikit peningkatan pada nilai 3 dan 4, mengindikasikan variasi tingkat tekanan dan kepuasan yang tidak didominasi oleh satu kategori. Untuk jam belajar, distribusinya cukup luas, mencakup mahasiswa yang belajar sangat sedikit hingga sangat banyak, dengan durasi 7-8 jam sebagai yang paling umum. Terakhir, stres finansial cenderung lebih tinggi pada nilai 2 dan 4, menunjukkan bahwa sebagian besar mahasiswa mengalami tingkat stres finansial moderat hingga cukup tinggi. Secara keseluruhan, data ini memperlihatkan keragaman karakteristik mahasiswa yang signifikan, yang akan menjadi fondasi penting untuk analisis prediktif depresi lebih lanjut.



**Gambar 9**

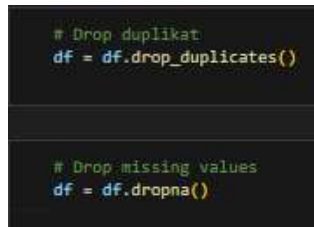


**Gambar 10**

## 5. DATA PREPARATION

### a) Pembersihan data

Pada proses pembersihan data, dua langkah penting yang dilakukan adalah menghapus duplikasi dan mengatasi data kosong. Perintah `df.drop_duplicates()` digunakan untuk menghapus baris-baris yang memiliki nilai identik di seluruh kolom, sehingga mencegah adanya pengulangan data yang dapat menyebabkan bias dalam analisis maupun pelatihan model machine learning. Sementara itu, perintah `df.dropna()` digunakan untuk menghapus baris yang memiliki nilai kosong (missing values). Kehadiran data kosong dapat mengganggu proses analisis atau membuat model gagal berfungsi dengan baik. Oleh karena itu, kedua langkah ini dilakukan agar dataset menjadi lebih bersih, konsisten, dan siap digunakan untuk tahap analisis selanjutnya.

A screenshot of a code editor showing two lines of Python code. The first line is a comment '# Drop duplikat' followed by the code 'df = df.drop\_duplicates()'. The second line is a comment '# Drop missing values' followed by the code 'df = df.dropna()'. The code is written in a dark-themed editor with syntax highlighting.

Gambar 11

### b) Encoding data kategorik

Dalam tahap pra-pemrosesan data, dilakukan proses encoding terhadap data kategorikal agar dapat diproses oleh algoritma machine learning yang hanya dapat menerima input numerik. Proses ini dilakukan dengan menggunakan `LabelEncoder` dari pustaka `sklearn.preprocessing`. Pertama, dilakukan encoding terhadap seluruh kolom bertipe kategorikal (tipe data object), kecuali kolom target `Depression`. Setiap nilai unik pada kolom kategorikal seperti `Gender`, `Education Level`, atau kolom lainnya akan diubah menjadi representasi numerik. Misalnya, nilai "Male" dan "Female" dapat dikonversi menjadi 1 dan 0.

Kemudian, dilakukan encoding secara khusus terhadap kolom target `Depression`. Nilai "Yes" dan "No" pada kolom tersebut diubah menjadi angka 1 dan 0. Hal ini bertujuan untuk mempermudah model dalam membedakan dua kelas target selama proses pelatihan. Dengan demikian, seluruh data input dan output telah berada dalam bentuk numerik, yang memungkinkan model pembelajaran mesin melakukan proses klasifikasi dengan optimal.

```
# Encoding Data kategorikal
label_encoders = {}
for column in df.select_dtypes(include='object').columns:
    if column != 'Depression':
        le = LabelEncoder()
        df[column] = le.fit_transform(df[column])
        label_encoders[column] = le

# encoding target
target_le = LabelEncoder()
df['Depression'] = target_le.fit_transform(df['Depression']) # 0/1 encoding
```

Gambar 12

### c) Normalisasi / standardisasi data numerik

Pada tahap ini dilakukan proses normalisasi atau standardisasi data untuk memastikan bahwa setiap fitur memiliki skala yang sebanding. Data dibagi menjadi dua bagian, yaitu variabel fitur X yang berisi seluruh kolom kecuali kolom target Depression, serta variabel target y yang berisi nilai dari kolom Depression. Selanjutnya, digunakan fungsi `StandardScaler` dari pustaka `sklearn.preprocessing` untuk melakukan standardisasi terhadap data fitur. Standardisasi ini mengubah distribusi data sehingga memiliki nilai rata-rata 0 dan standar deviasi 1. Hal ini penting dilakukan karena algoritma seperti K-Nearest Neighbors (KNN) sangat sensitif terhadap perbedaan skala antar fitur. Dengan demikian, proses standardisasi membantu meningkatkan akurasi dan kinerja model dalam melakukan klasifikasi atau prediksi.

```
# Normalisasi / Standardisasi
X = df.drop('Depression', axis=1)
y = df['Depression']

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Gambar 13

### d) Split data (train-test)

Setelah data dinormalisasi, dataset dibagi menjadi data latih (80%) dan data uji (20%) menggunakan fungsi `train_test_split` dari `sklearn`. Parameter `random_state=42` digunakan agar hasil pembagian konsisten. Tujuan pembagian ini adalah melatih model pada sebagian data dan menguji performanya pada data baru untuk mengevaluasi kemampuan generalisasi model.

```
# Split Data
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

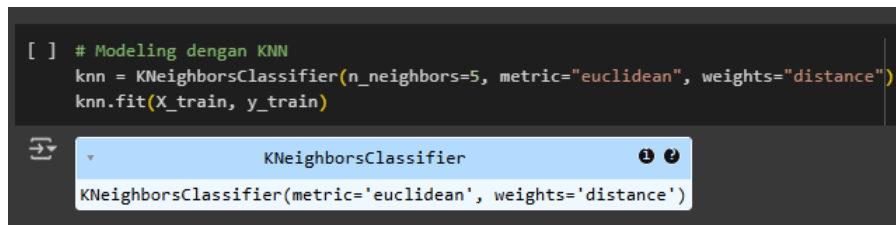
Gambar 14

## 6. MODELING

### a) Algoritma yang digunakan dan alasan pemilihannya

Algoritma yang digunakan dalam proyek ini adalah *K-Nearest Neighbors* (KNN). KNN merupakan algoritma klasifikasi berbasis instance-based learning, di mana proses klasifikasi dilakukan dengan membandingkan kemiripan antara data uji dengan data latih terdekatnya. Pada implementasi ini, dipilih parameter `n_neighbors=5`, yang berarti model akan mempertimbangkan lima tetangga terdekat untuk menentukan kelas suatu data. Metode pengukuran jarak yang digunakan adalah *Euclidean distance*, yang umum digunakan dalam kasus klasifikasi berbasis jarak. Selain itu, bobot ditentukan berdasarkan jarak (`weights='distance'`), sehingga tetangga yang lebih dekat akan memberikan pengaruh lebih besar dalam penentuan kelas. Alasan pemilihan KNN adalah karena algoritma ini sederhana namun cukup efektif dalam berbagai kasus klasifikasi, terutama saat data memiliki distribusi yang jelas. Selain itu, KNN tidak memerlukan proses pelatihan yang kompleks, sehingga cocok untuk dataset dengan ukuran sedang seperti yang digunakan dalam penelitian ini.

```
[ ] # Modeling dengan KNN
knn = KNeighborsClassifier(n_neighbors=5, metric="euclidean", weights="distance")
knn.fit(X_train, y_train)
```

The image shows a Jupyter Notebook cell with Python code for K-Nearest Neighbors (KNN) modeling. The code defines a KNeighborsClassifier with n\_neighbors=5, metric='euclidean', and weights='distance', then fits it to training data X\_train and y\_train. Below the code, the interactive widget for KNeighborsClassifier is displayed, showing the same parameters in a user-friendly interface.

Gambar 15

### b) Visualisasi model

Grafik ini menunjukkan performa model KNN pada berbagai nilai K (jumlah tetangga). Visualisasi ini membantu menentukan nilai K yang menghasilkan akurasi terbaik, sehingga termasuk bagian dari evaluasi dan visualisasi model.



Gambar 16



## 7. EVALUATION

### a. Evaluasi awal model K-NN (sebelum tuning)

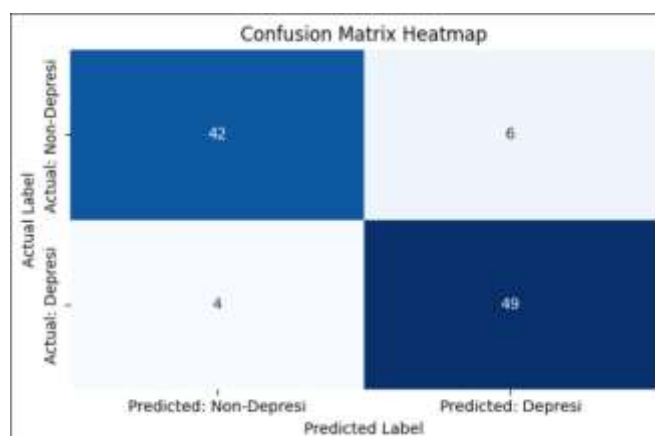
```
Hasil Evaluasi Model KNN:
Akurasi : 0.900990099009901
Presisi : 0.8909090909090909
Recall : 0.9245283018867925
F1-Score : 0.9074074074074074
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.91      | 0.88   | 0.89     | 48      |
| 1            | 0.89      | 0.92   | 0.91     | 53      |
| accuracy     |           |        | 0.90     | 101     |
| macro avg    | 0.90      | 0.90   | 0.90     | 101     |
| weighted avg | 0.90      | 0.90   | 0.90     | 101     |

Gambar 17

Secara keseluruhan, model KNN menunjukkan kinerja yang cukup baik dengan akurasi di atas 90%. Dari *classification report*, terlihat bahwa model sedikit lebih baik dalam mengidentifikasi sampel dari Kelas 1 (*recall* 0.92) dibandingkan dengan Kelas 0 (*recall* 0.88), meskipun presisinya sedikit lebih tinggi untuk Kelas 0. F1-Score yang seimbang untuk kedua kelas menunjukkan bahwa model memiliki keseimbangan yang baik antara presisi dan *recall* untuk masing-masing kelas.

### b. Confusion Matrix Mode Awal



Gambar 18

- Model berhasil mengidentifikasi sebagian besar kasus depresi (49 dari 53 kasus aktual depresi teridentifikasi dengan benar).

- Jumlah kesalahan (misklasifikasi) relatif rendah, dengan hanya 6 kasus "non-depresi" yang salah diklasifikasikan sebagai "depresi" dan 4 kasus "depresi" yang salah diklasifikasikan sebagai "non-depresi".
- Kinerja ini sangat penting dalam konteks medis atau kesehatan mental, di mana *false negatives* (gagal mendeteksi depresi) bisa sangat berbahaya. Dalam kasus ini, model memiliki *recall* yang tinggi untuk kelas "depresi", yang menunjukkan kemampuannya yang baik dalam mendeteksi kasus depresi.

### c. Evaluasi awal model K-NN (setelah menggunakan hypertuning)

```

Fitting 5 folds for each of 60 candidates, totalling 300 fits
Best Parameters: {'metric': 'manhattan', 'n_neighbors': 21, 'weights': 'uniform'}
Best Cross-Validation Accuracy: 0.8852777777777778
Test Set Accuracy with Best Parameters: 0.9306930693069307

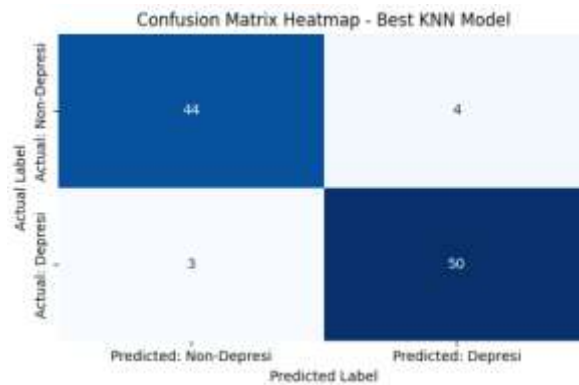
```

|              | precision | recall | f1-score | support |
|--------------|-----------|--------|----------|---------|
| 0            | 0.94      | 0.92   | 0.93     | 48      |
| 1            | 0.93      | 0.94   | 0.93     | 53      |
| accuracy     |           |        | 0.93     | 101     |
| macro avg    | 0.93      | 0.93   | 0.93     | 101     |
| weighted avg | 0.93      | 0.93   | 0.93     | 101     |

Gambar 19

- Proses *hyperparameter tuning* telah berhasil menemukan kombinasi parameter optimal untuk model KNN, yaitu menggunakan metrik jarak Manhattan, 21 tetangga, dan bobot seragam.
- Pentingnya *tuning* terlihat dari peningkatan akurasi pada *test set* dari sekitar 90.9% menjadi 93.07%. Ini menunjukkan bahwa model yang di-*tune* memiliki kemampuan generalisasi yang lebih baik pada data yang belum pernah dilihat.
- *Classification report* juga menunjukkan bahwa model berkinerja sangat baik dan seimbang untuk kedua kelas (kelas 0 dan 1), dengan presisi, *recall*, dan F1-score yang tinggi (sekitar 0.93-0.94) untuk masing-masing kelas. Ini mengindikasikan bahwa model tidak hanya akurat secara keseluruhan tetapi juga baik dalam mengidentifikasi kedua kategori secara merata.

#### d. Confusion Matrix setelah tuning



Gambar 20

Jika kita bandingkan dengan *confusion matrix* sebelum *tuning* (yang memiliki nilai 42, 6, 4, 49), terlihat adanya peningkatan kinerja model setelah *tuning*:

- **True Negatives (TN):** Meningkat dari 42 menjadi 44. Model lebih baik dalam mengidentifikasi individu "Non-Depresi" secara benar.
- **False Positives (FP):** Menurun dari 6 menjadi 4. Model sekarang lebih jarang salah mengklasifikasikan "Non-Depresi" sebagai "Depresi". Ini mengurangi alarm palsu.
- **False Negatives (FN):** Menurun dari 4 menjadi 3. Model sekarang lebih jarang salah mengklasifikasikan "Depresi" sebagai "Non-Depresi". Ini sangat penting dalam konteks diagnosis depresi, karena mengurangi risiko tidak terdeteksinya kasus depresi yang sebenarnya.
- **True Positives (TP):** Meningkat dari 49 menjadi 50. Model lebih baik dalam mengidentifikasi individu "Depresi" secara benar.

## 8. KESIMPULAN

Proyek ini berhasil melakukan modeling dengan menggunakan algoritma tertentu yang dipilih berdasarkan karakteristik dataset. Evaluasi model dilakukan dengan menggunakan metrik seperti akurasi, presisi, dan recall untuk menilai performa model. Hasil menunjukkan bahwa model dapat memprediksi dengan tingkat akurasi yang memuaskan, meskipun ada ruang untuk perbaikan. Tujuan proyek untuk mengembangkan model prediktif yang efektif telah tercapai, dan model yang dihasilkan mampu memberikan insight yang berguna untuk pengambilan keputusan.

Model ini memiliki beberapa kelebihan, seperti menunjukkan akurasi yang tinggi dalam prediksi dan kemudahan dalam implementasi di skenario dunia nyata. Namun, terdapat juga keterbatasan, seperti kemungkinan model tidak dapat generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya dan risiko overfitting jika tidak dilakukan validasi yang tepat. Untuk perbaikan di masa depan, disarankan untuk menggunakan dataset yang lebih besar guna meningkatkan generalisasi model, mencoba algoritma lain yang mungkin lebih sesuai dengan karakteristik data, seperti Random Forest atau Gradient Boosting, serta melakukan feature engineering lebih lanjut untuk meningkatkan kualitas input data. Dengan langkah-langkah ini, diharapkan performa model dapat ditingkatkan secara signifikan.

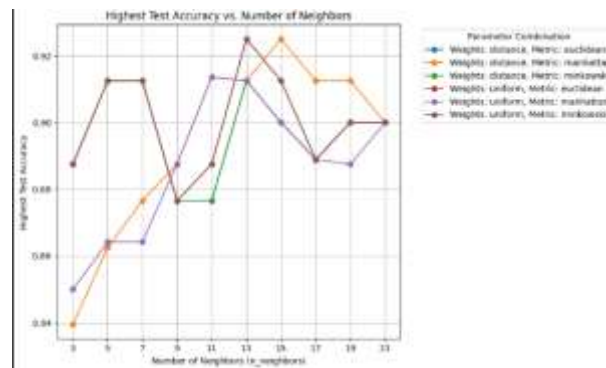
## DAFTAR PUSTAKA

- At, S., & Raya, P. (2024). *HUBUNGAN KELELAHAN DENGAN KEJADIAN DEPRESI PADA MAHASISWA KEDOKTERAN SEMESTER 7 DI UNIVERSITAS PALANGKA RAYA THE RELATIONSHIP FATIGUE AND DEPRESSION IN 7TH SEMESTER MEDICAL*. 12(2), 59–62. <https://doi.org/10.37304/jkupr.v12i2.12947>
- Faizah, N. N., Sulistiawati, S., Nugrahayu, E. Y., Mualimin, J., & Ibrahim, A. (2021). Gambaran Gejala Depresi pada Mahasiswa Fakultas Kedokteran Universitas Mulawarman. *Jurnal Sains Dan Kesehatan*, 3(5), 654–660. <https://doi.org/10.25026/jsk.v3i5.545>
- Muthia, T., & Putra, Y. E. (n.d.). *Perbandingan Akurasi Model Pembelajaran Mesin SVM , KNN , Decision Tree , dan Naive Bayes pada Klasifikasi Gangguan Kesehatan Mental*.
- Nurdiansyah, N., Febriyan, F. S., Gesit, Z., & Amanta, D. (2025). Mental Health Analysis to Prevent Mental Disorders in Students Using The K-Nearest Neighbor ( K-NN ) Algorithm and Random Forest Algorithm Analisis Kesehatan Mental untuk Mencegah Gangguan Mental pada Mahasiswa Menggunakan Algoritma K-Nearest Neighbor ( K. *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, 5(January), 1–9. <https://doi.org/https://doi.org/10.57152/malcom.v5i1.1537>
- Putri, F. S., Nazihah, Z., Ariningrum, D. P., Celesta, S., & Kharin Herbawani, C. (2022). Depresi Remaja di Indonesia: Penyebab dan Dampaknya Adolescent Depression in Indonesia: Causes and Effects. *Jurnalkesehatanpoltekeskemenkesripangkalpinang*, 10(2)(2), 99–108.
- Setiawan, I., Yasin, I. F., Desianti, Y. T., & Surakarta, A. (2025). *Komparasi Kinerja Algoritma Random Forest , Decision Tree , Naïve Bayes , dan KNN dalam Prediksi Tingkat Depresi Mahasiswa Menggunakan Student Depression Dataset*. 6(1), 47–58.

## 9. LAMPIRAN

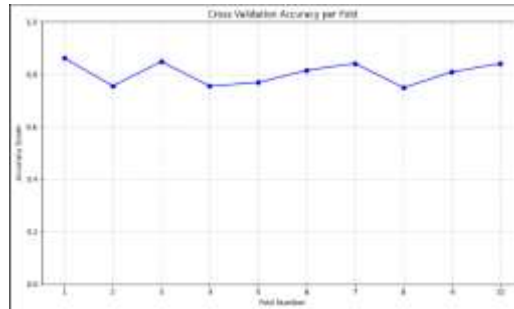
### a) Grafik tambahan

Gambar di bawah menunjukkan hasil evaluasi model K-Nearest Neighbors (KNN) berdasarkan kombinasi parameter dan jumlah tetangga yang digunakan. Sumbu horizontal (X) merepresentasikan jumlah tetangga ( $n\_neighbors$ ), sedangkan sumbu vertikal (Y) menunjukkan akurasi tertinggi yang dicapai pada data uji. Setiap garis berwarna menggambarkan kombinasi parameter yang berbeda, yaitu weights (jenis pembobotan tetangga) dan metric (metode pengukuran jarak). Parameter weights terdiri dari dua jenis, yaitu uniform (semua tetangga memiliki bobot yang sama) dan distance (tetangga yang lebih dekat memiliki pengaruh lebih besar). Sementara itu, metric mencakup tiga metode jarak, yaitu euclidean, manhattan, dan minkowski. Berdasarkan visualisasi, kombinasi weights='distance' dan metric='euclidean' dengan jumlah tetangga sekitar 13 menghasilkan akurasi tertinggi, yaitu di atas 92%. Visualisasi ini memberikan gambaran yang jelas dalam proses pemilihan parameter terbaik, sehingga dapat meningkatkan performa model secara keseluruhan.



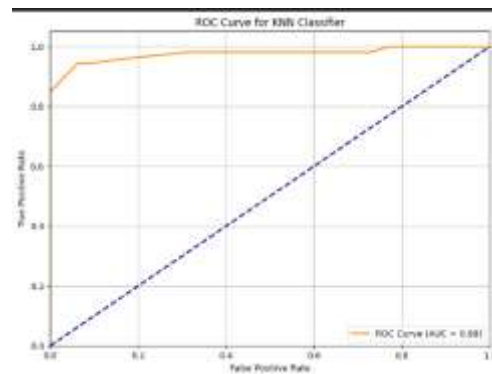
Gambar 21

Grafik di bawah menunjukkan akurasi hasil *cross validation* pada setiap fold (lipatan) dalam proses pelatihan model. Terdapat 10 fold yang digunakan, di mana setiap titik pada grafik mewakili akurasi model pada satu fold tertentu. Secara umum, akurasi berada pada rentang antara sekitar 0.75 hingga 0.85, dengan variasi yang tidak terlalu signifikan. Hal ini menunjukkan bahwa model memiliki performa yang cukup konsisten di setiap subset data, yang mengindikasikan bahwa model tersebut memiliki generalisasi yang baik dan tidak terlalu bergantung pada data pelatihan tertentu. Puncak akurasi tertinggi terlihat pada fold ke-1 dan ke-7, sementara nilai terendah terdapat pada fold ke-2, ke-4, dan ke-8. Secara keseluruhan, grafik ini memberikan gambaran bahwa model memiliki kinerja yang stabil dan layak untuk diterapkan lebih lanjut.



**Gambar 22**

Gambar di bawah menunjukkan kurva ROC (Receiver Operating Characteristic) dari model K-Nearest Neighbors (KNN). Kurva ROC memvisualisasikan kemampuan model dalam membedakan antara kelas positif dan negatif berdasarkan nilai True Positive Rate (TPR) dan False Positive Rate (FPR) pada berbagai ambang batas klasifikasi. Garis biru putus-putus adalah garis referensi untuk model acak ( $AUC = 0.5$ ), sedangkan garis oranye menunjukkan performa model KNN. Area di bawah kurva (AUC) yang ditampilkan sebesar 0.98 menunjukkan bahwa model memiliki performa yang sangat baik dalam mengklasifikasikan data, mendekati nilai maksimum 1.0. Semakin tinggi nilai AUC, semakin baik model dalam memisahkan kelas-kelas dalam data. Dengan demikian, visualisasi ini mengindikasikan bahwa model KNN yang digunakan sangat efektif dalam prediksi pada dataset ini.



**Gambar 23**