ANALISIS PERBANDINGAN ALGORITMA C4.5 DAN XGBOOST UNTUK MEMPREDIKSI KEBAHAGIAAN DARI DATA GAYA HIDUP DIGITAL

Laily Mariyatul Qibthiyah, Agustia Hananto, Elfina Novalia
Program Studi Sistem Informasi, Universitas Buana Perjuangan Karawang
lailymq6621@gmail.com

ABSTRAK

Perkembangan teknologi digital mengubah pola kehidupan masyarakat, termasuk aspek gaya hidup dan kesejahteraan psikologis. Gaya hidup digital yang mencakup penggunaan media sosial, durasi penggunaan perangkat, kualitas tidur, serta tingkat stres diperkirakan memiliki hubungan dengan tingkat kebahagiaan individu. Oleh karena itu, penelitian ini bertujuan untuk menganalisis dan membandingkan kinerja algoritma C4.5 dan XGBoost dalam memprediksi tingkat kebahagiaan berdasarkan data gaya hidup digital, serta mengidentifikasi faktor-faktor yang paling berpengaruh terhadap kebahagiaan. Penelitian ini menggunakan pendekatan kuantitatif dengan metode eksperimen komparatif. Dataset yang digunakan terdiri dari 500 data dengan 10 atribut yang mencakup variabel numerik dan kategorikal. Tahapan penelitian meliputi preprocessing data, pembagian data menggunakan metode train-test split (80:20), implementasi algoritma C4.5 dan XGBoost, serta model evaluasi menggunakan metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa kedua algoritma memiliki performa yang relatif sama dengan nilai akurasi sebesar 0,44. Nilai presisi, recall, dan F1-score pada kedua model juga tidak menunjukkan perbedaan yang signifikan. Analisis lebih lanjut menunjukkan bahwa distribusi data yang tidak buruk mempengaruhi kinerja model, di mana model cenderung lebih baik dalam memprediksi kelas mayoritas dibandingkan kelas minoritas. Berdasarkan analisis dan korelasi feature important, variabel yang paling berpengaruh terhadap tingkat kebahagiaan adalah tingkat stres, kualitas tidur, dan durasi penggunaan perangkat digital.

Kata kunci: *C4.5, XGBoost, kebahagiaan, gaya hidup digital, klasifikasi, machine learning*

1. PENDAHULUAN

Seiring dengan perkembangan teknologi digital dan penetrasi internet yang semakin luas ke dalam kehidupan sehari-hari, gaya hidup digital telah menjadi salah satu aspek penting dalam keseharian individu, termasuk kebiasaan penggunaan media sosial, platform daring, tingkat aktivitas online, hingga interaksi digital dengan lingkungan sekitar. Misalnya, penelitian menunjukkan bah spasi sebanyak satu kalia kebiasaan penggunaan internet dan digitalisasi secara umum memiliki korelasi positif dengan kesejahteraan subjektif individu, meskipun fenomena tersebut juga menghadirkan tantangan berupa tekanan sosial dan perbandingan daring yang dapat menurunkan kepuasan hidup. Oleh karena itu, memahami hubungan antara gaya hidup digital dan kebahagiaan menjadi semakin relevan dalam era industri 4.0 dan masyarakat informasi [1].

Lebih lanjut, kebahagiaan atau kesejahteraan subjektif telah diidentifikasi sebagai indikator penting dalam psikologi positif maupun penelitian sosial, karena tidak hanya mencerminkan kondisi mental individu tetapi juga berpengaruh terhadap produktivitas, kesehatan mental dan fisik, serta interaksi sosial. Sebagai contoh, studi "Lifestyle and Life Satisfaction" menunjukkan bah spasi sebanyak satu kalia pola hidup seperti konsumsi buah atau sayuran dan aktivitas fisik memiliki dampak signifikan terhadap kepuasan hidup. Dalam konteks digital, penelitian menunjukkan adanya hubungan yang kompleks antara keterampilan digital, penggunaan teknologi, dan kebahagiaan: misalnya, penggunaan keterampilan digital yang tinggi dapat meningkatkan pendapatan dan kepuasan, tetapi juga berpotensi menurunkan kebahagiaan melalui perangkap perbandingan sosial atau kecanduan digital [2].

Dengan latar belakang tersebut, maka penelitian yang berfokus pada prediksi kebahagiaan berdasarkan gaya hidup digital menggunakan teknik pembelajaran mesin (machine learning) menjadi sangat penting. Hal ini didukung oleh tren penelitian terkini yang menerapkan algoritma seperti XGBoost untuk memprediksi variabel kesejahteraan atau kepuasan hidup. Sebagai contoh, sebuah penelitian pada populasi lanjut usia di Korea menggunakan XGBoost untuk mengklasifikasikan kepuasan hidup, dan menunjukkan performa yang unggul dibandingkan metode lainnya.Di sisi lain, algoritma C4.5 (decision tree) telah banyak digunakan dalam klasifikasi data gaya hidup, kesehatan maupun perilaku, yang menunjukkan fleksibilitas dan interpretabilitasnya dalam konteks penelitian sosial-data [3].

Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis perbandingan antara algoritma C4.5 dan XGBoost dalam konteks memprediksi kebahagiaan berdasarkan data gaya hidup digital. Dengan demikian, penelitian ini diharapkan memberikan kontribusi sebagai berikut: pertama, memperkuat pemahaman tentang variabel gaya hidup digital yang berpengaruh terhadap kebahagiaan; kedua, mengevaluasi performa algoritma pembelajaran mesin yang berbeda (C4.5 vs XGBoost) dalam klasifikasi kebahagiaan; ketiga, menyediakan rekomendasi metodologis untuk penelitian selanjutnya

maupun praktik implementasi dalam aplikasi nyata yang berbasis data gaya hidup digital [4].

2. TINJAUAN PUSTAKA

2.1. Gaya Hidup Digital

Gaya hidup digital merupakan pola perilaku yang dipengaruhi oleh kemajuan teknologi informasi, di mana aktivitas sehari-hari banyak dilakukan melalui perangkat digital seperti smartphone, komputer, dan internet. Perubahan perilaku ini mencakup komunikasi, hiburan, pekerjaan, hingga transaksi ekonomi yang berbasis digital. kompetensi digital berhubungan positif dengan kesejahteraan psikologis karena individu yang mampu beradaptasi dengan teknologi cenderung merasa lebih percaya diri dan terhubung dengan lingkungannya [5]. Namun, studi lain oleh The Effect Of Digital Obsesity and Phubbing menunjukkan bahwa penggunaan perangkat digital secara berlebihan dapat menurunkan Tingkat kebahagian melalui penurunan interaksi soasial dan meningkatnya stress digital.

Fenomena ini menunjukkan bahwa gaya hidup digital memiliki dua sisi: di satu sisi meningkatkan efisiensi dan konektivitas sosial, namun di sisi lain dapat menciptakan kelelahan digital dan ketergantungan terhadap teknologi [6].

2.2. Kebahagiaan dan Kesejahteraan Subjektif

Kebahagiaan atau *subjective well-being* merupakan kondisi psikologis yang mencerminkan tingkat kepuasan seseorang terhadap kehidupannya. Menurut [7]. kebahagiaan dipengaruhi oleh faktor kesehatan, interaksi sosial, ekonomi, serta lingkungan keluarga. Dalam konteks digital, hubungan antara kebahagiaan dan penggunaan teknologi bersifat kompleks digitalisasi dapat meningkatkan rasa keterhubungan, namun juga dapat mengurangi kebahagiaan bila pengguna tidak mampu mengontrol eksposur digitalnya [8]).

Ekonomi digital memiliki dampak positif terhadap kebahagiaan individu melalui peningkatan peluang ekonomi dan efisiensi hidup. Namun, efek positif ini hanya terjadi apabila individu memiliki literasi digital yang baik dan tidak mengalami stres akibat penggunaan teknologi berlebihan. Dengan demikian, kebahagiaan di era digital bukan hanya ditentukan oleh akses terhadap teknologi, tetapi juga oleh kualitas penggunaannya [8].

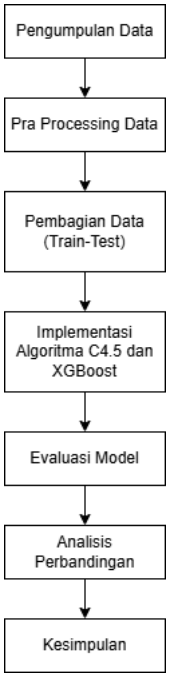
2.3. Penelitian Terdahulu

Beberapa penelitian sebelumnya telah mengkaji hubungan antara kebahagiaan dan gaya hidup digital dengan pendekatan statistik maupun machine learning. [9] meneliti hubungan antara kompetensi digital dan kesejahteraan psikologis, sementara [4] mengembangkan model jalur kebahagiaan di era digital yang melibatkan kesehatan, keluarga, dan digitalisasi. Di sisi algoritmik, penelitian [10] membandingkan C4.5 dengan metode reduksi atribut dan menemukan peningkatan akurasi signifikan, sedangkan [11] menunjukkan dominasi XGBoost pada klasifikasi data kesehatan.

Penelitian-penelitian tersebut menjadi dasar empiris bahwa integrasi antara data gaya hidup digital dan metode pembelajaran mesin mampu memberikan gambaran prediktif terhadap kebahagiaan. Namun, belum banyak studi yang secara eksplisit membandingkan C4.5 dan XGBoost dalam konteks kebahagiaan berdasarkan gaya hidup digital, sehingga penelitian ini memiliki kontribusi orisinal di bidang tersebut.

3. METODE PENELITIAN

3.1. Alur Penelitian



Ganbar 1. Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang disusun secara sistematis agar mudah dipahami dan dapat diprediksi oleh peneliti lain. Selain itu, penyusunan tahapan penelitian secara sistematis bertujuan untuk menjamin bahwa setiap proses yang dilakukan memiliki arah dan tujuan yang jelas serta dapat dipertanggungjawabkan secara ilmiah. Dengan adanya alur penelitian yang terstruktur, peneliti dapat mengontrol setiap tahap mulai dari pengumpulan data, preprocessing, pemodelan, hingga evaluasi sehingga potensi kesalahan dapat diminimalkan. Flowchart yang ditampilkan juga berfungsi sebagai panduan visual yang memperjelas hubungan antar tahap penelitian, sehingga memudahkan dalam analisis, interpretasi hasil, serta replikasi penelitian di masa mendatang oleh peneliti lain yang ingin mengembangkan atau menguji kembali metode yang digunakan. Secara umum, alur penelitian digambarkan dalam bentuk diagram blok seperti pada gambar 1.

3.2. Pengumpulan Data Menggunakan C4.5 dan XGBoost

Berdasarkan Gambar 1, penelitian ini dimulai dengan tahap Pengumpulan Data, yaitu proses memperoleh data gaya hidup digital dan tingkat kebahagiaan yang akan digunakan sebagai dasar analisis. Pada tahap ini, seluruh variabel yang diperlukan seperti durasi penggunaan media digital, kualitas tidur, tingkat stres, dan skor kebahagiaan dikumpulkan dari dataset yang tersedia. Setelah data diperoleh, peneliti meninjau struktur dataset, memastikan format file sesuai dengan kebutuhan pemodelan, dan memeriksa kelengkapan variabel agar tahap selanjutnya dapat dilaksanakan dengan baik.

3.3. Rencana Kebutuhan dan Pengumpulan Data

Sumber penelitian ini di peroleh dari Emily R selama satu tahun kebelakang dengan total 500 data. Pengumpulan data dalam penelitian ini dilakukan dengan menggunakan data sekunder dalam bentuk dataset berformat CSV. Dataset diperoleh dari data terkait kesehatan mental dan gaya hidup digital.

Setelah dataset diperoleh, dilakukan pemeriksaan awal berupa pengecekan kelengkapan data, konsistensi nilai, serta kesesuaian atribut dengan tujuan penelitian. Selanjutnya, data disiapkan untuk tahap analisis melalui proses pembersihan data (*data cleaning*) dan transformasi data agar dapat digunakan dalam pemodelan machine learning. Dataset yang telah diproses kemudian digunakan sebagai dasar dalam seluruh tahapan penelitian, mulai dari pelatihan model hingga evaluasi hasil.

Tabel 1. Data Gaya Hidup Digital

User ID	page	Gender	Daily Screen Time	Sleep Quality (1-10)	Stress Level (1-10)	Days Without Social Media	Exercise Frequency	Social Media Platform	Happiness Index (1-10)
U001	44	Male	3.1	7.0	6.0	2.0	5.0	Facebook	10.0
U002	30	Other	5.1	7.0	8.0	5.0	3.0	LinkedIn	10.0
U003	23	Other	7.4	6.0	7.0	1.0	3.0	YouTube	6.0
U004	36	Female	5.7	7.0	8.0	1.0	1.0	TikTok	8.0
U005	34	Female	7.0	4.0	7.0	5.0	1.0	X (Twitter)	8.0
U006	38	Male	6.6	5.0	7.0	4.0	3.0	LinkedIn	8.0
U007	26	Female	7.8	4.0	8.0	2.0	0.0	TikTok	7.0
U008	26	Female	7.4	5.0	6.0	1.0	4.0	Instagram	7.0
U009	39	Male	4.7	7.0	7.0	6.0	1.0	YouTube	9.0

Berdasarkan cuplikan data tersebut, Tabel 1 menampilkan contoh data responden yang digunakan dalam penelitian untuk menganalisis hubungan antara gaya hidup digital dan tingkat kebahagiaan. Data terdiri atas variabel demografis, perilaku digital, dan faktor kesehatan mental, dengan *Happiness Index* sebagai variabel target dalam pemodelan menggunakan algoritma C4.5 dan XGBoost.

Tabel 2. Atribut Data Penelitian

No	Nama Atribut	Deskripsi	Tipe Data
1.	Gender	Menunjukkan jenis kelamin responden.	Kategorikal
2.	Daily Screen Time	Rata-rata waktu penggunaan perangkat digital per hari (jam).	Numerik
3.	Sleep Quality (1-10)	Nilai kualitas tidur, semakin tinggi semakin baik.	Numerik
4.	Stress Level (1-10)	Tingkat stres individu, semakin tinggi berarti semakin stres.	Numerik
5.	Days_Without_Social_Media	Jumlah hari tanpa menggunakan media sosial.	Numerik
6.	Exercise_Frequency (week)	Frekuensi olahraga per minggu.	Numerik
7.	Social_Media_Platform	Platform media sosial yang paling sering digunakan.	Kategorikal
8.	Happiness Index (1-10)	Tingkat kebahagiaan individu dan menjadi variabel target penelitian.	Numerik

Menganalisis dan memprediksi tingkat kebahagiaan individu berdasarkan gaya hidup digital. Variabel independen terdiri atas gender, durasi penggunaan perangkat digital, kualitas tidur, tingkat stres, jumlah hari tanpa media sosial, frekuensi olahraga, serta platform media sosial. Variabel dependen adalah *Happiness Index* yang menunjukkan tingkat kebahagiaan responden dan menjadi target dalam proses pemodelan menggunakan algoritma C4.5 dan XGBoost.

3.4. Pra-Pemrosesan Data

Tahapan pra-pemrosesan data yang dilakukan sebelum penerapan algoritma C4.5 dan XGBoost. Pra-pemrosesan ini bertujuan untuk memastikan kualitas data agar siap digunakan dalam proses pemodelan dan menghasilkan prediksi yang akurat. Setiap tahap dilakukan secara sistematis mulai dari pemeriksaan data awal hingga pembagian data menjadi data latih dan data uji.

Tabel 3. Tahapan Pra-pemrosesan Data

No	Tahapan	Deskripsi Singkat
1	Pemeriksaan Data Awal	Memeriksa struktur data, tipe data, dan jumlah atribut.
2	Data Cleaning	Menghapus data tidak relevan, duplikat, dan menangani nilai kosong.
3	Encoding Data	Mengubah data kategorikal menjadi numerik.
4	Normalisasi	Menyamakan skala data numerik agar seimbang.
5	Pelabelan Target	Mengelompokkan nilai Happiness Index ke dalam kelas tertentu.
6	Data spilitting	Membagi data menjadi data latih dan data uji (80:20).

Tahapan pra-pemrosesan data pada tabel di atas bertujuan untuk menghasilkan data yang bersih, terstruktur, dan siap digunakan dalam proses pemodelan machine learning. Proses dimulai dengan pemeriksaan data awal untuk memahami struktur dataset, jumlah atribut, serta tipe data pada setiap kolom. Tahap ini membantu peneliti dalam menentukan metode pembersihan dan transformasi data yang tepat.

Selanjutnya dilakukan data *cleaning* untuk menghilangkan data yang tidak relevan, mengatasi nilai kosong, serta mendeteksi adanya data ganda atau nilai ekstrem yang berpotensi memengaruhi hasil prediksi. Tahap encoding dilakukan untuk mengubah data kategorikal seperti jenis kelamin dan platform media sosial menjadi bentuk numerik sehingga dapat diproses oleh algoritma C4.5 dan XGBoost.

Proses normalisasi bertujuan untuk menyamakan skala data numerik agar tidak terdapat perbedaan rentang nilai yang terlalu jauh antar atribut, sehingga proses pelatihan model menjadi lebih stabil dan adil. Pelabelan target dilakukan dengan mengelompokkan nilai indeks kebahagiaan ke dalam kelas tertentu sesuai kebutuhan klasifikasi.

Tahap terakhir adalah pembagian dataset menjadi data latih dan data uji dengan rasio 80:20 untuk menguji kemampuan model dalam melakukan prediksi pada data baru. Dengan tahapan ini, kualitas data dapat terjaga sehingga hasil evaluasi model menjadi lebih akurat dan dapat dipertanggungjawabkan secara ilmiah.

3.5. Pembagian Data (*Train–Test*)

Setelah data siap, proses dilanjutkan pada tahap Pembagian Data (*Train–Test).* Pada tahap ini, dataset dibagi menjadi dua bagian utama: data pelatihan (training data) sebesar 80% dari total data, dan data pengujian (testing data) sebesar 20%. Pembagian ini dilakukan secara acak agar model tidak mengalami bias dalam mempelajari pola dari data. Data pelatihan digunakan untuk membangun dan menyesuaikan model, sedangkan data pengujian digunakan untuk mengevaluasi seberapa baik model mampu memprediksi data baru yang belum pernah dilihat sebelumnya.

Teknik pembagian data ini bertujuan untuk mengukur kemampuan generalisasi model terhadap data yang tidak digunakan saat proses pelatihan. Dengan memisahkan data latih dan data uji, peneliti dapat menilai apakah model hanya menghafal pola pada data pelatihan atau benar-benar mampu mengenali pola secara umum. Pembagian data yang proporsional juga membantu menghindari *overfitting* serta memberikan gambaran yang lebih objektif terhadap kinerja algoritma C4.5 dan XGBoost dalam memprediksi tingkat kebahagiaan berdasarkan gaya hidup digital.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Penelitian

Pada bagian ini disajikan hasil penelitian yang diperoleh dari proses pengolahan data gaya hidup digital menggunakan algoritma C4.5 dan XGBoost. Tahapan yang dilakukan meliputi pengumpulan data, preprocessing, eksplorasi data, serta pemodelan dan evaluasi kinerja algoritma.

Hasil penelitian disajikan secara sistematis dimulai dari deskripsi dataset, pemeriksaan kualitas data, analisis distribusi dan korelasi antar variabel, hingga proses pembagian data sebelum dilakukan pemodelan.

4.2. Deskripsi Dataset

Dataset yang digunakan dalam penelitian ini merupakan data gaya hidup digital yang memuat informasi mengenai kebiasaan individu dalam menggunakan teknologi serta kondisi yang berkaitan dengan tingkat kebahagiaan. Dataset terdiri dari 500 data responden dengan total 10 atribut.

	User_ID	Age	Gender	Daily_Screen_Time(hrs)	Sleep_Quality(1-10)	Stress_Level(1-10)	Days_Without_Social_Media	Exercise_Frequency(week)
0	U001	44	Male	3.1	7.0	6.0	2.0	5.0
1	U002	30	Other	5.1	7.0	8.0	5.0	3.0
2	U003	23	Other	7.4	6.0	7.0	1.0	3.0
3	U004	36	Female	5.7	7.0	8.0	1.0	1.0
4	U005	34	Female	7.0	4.0	7.0	5.0	1.0

Kolom:
['User_ID', 'Age', 'Gender', 'Daily_Screen_Time(hrs)', 'Sleep_Quality(1-10)', 'Stress_Level(1-10)', 'Days_Without_Social_Media', 'Exerci
Error: Runtime no longer has a reference to this dataframe, please re-run this cell and try again.

Gambar 1. Tampilan Dataset Awal

Atribut yang terdapat dalam dataset meliputi User_ID, Age, Gender, Daily Screen Time (hrs), Sleep Quality (1–10), Stress Level (1–10), Days Without Social Media, Weighted Weight (minggu), Social Media Platform, dan Happiness Index (1–10) sebagai target variabel.

Variabel Happiness Index (1–10) digunakan sebagai indikator tingkat kebahagiaan individu dengan skala nilai dari 1 hingga 10. Sementara itu, atribut lainnya berperan sebagai variabel prediktor yang digunakan untuk membangun klasifikasi model.

Dataset ini terdiri dari kombinasi data numerik dan kategorikal, di mana atribut numerik meliputi usia, durasi penggunaan perangkat digital, kualitas tidur, tingkat stres, jumlah hari tanpa media sosial, frekuensi olahraga, serta nilai kebahagiaan. Sedangkan

kategorikal meliputi jenis kelamin dan platform media sosial yang digunakan.

Keberagaman jenis data tersebut menunjukkan bahwa dataset memiliki karakteristik yang cukup kompleks sehingga memerlukan tahap preprocessing sebelum digunakan dalam proses pemodelan menggunakan algoritma pembelajaran mesin.

4.3. Pemeriksaan Data

Pemeriksaan data dilakukan untuk mengetahui struktur dataset, tipe data pada setiap atribut, serta memastikan tidak adanya nilai yang hilang (missing value) sebelum dilakukan proses pemodelan. Tahap ini sangat penting untuk menjamin kualitas data agar hasil analisis yang diperoleh dapat dipertanggungjawabkan.

```
Info data:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   User_ID                     500 non-null    object
 1   Age                         500 non-null    int64
 2   Gender                      500 non-null    object
 3   Daily_Screen_Time(hrs)      500 non-null    float64
 4   Sleep_Quality(1-10)         500 non-null    float64
 5   Stress_Level(1-10)          500 non-null    float64
 6   Days_Without_Social_Media   500 non-null    float64
 7   Exercise_Frequency(week)    500 non-null    float64
 8   Social_Media_Platform       500 non-null    object
 9   Happiness_Index(1-10)       500 non-null    float64
dtypes: float64(6), int64(1), object(3)
memory usage: 39.2+ KB
None
```

Gambar 2. Informasi Struktur Dataset

Berdasarkan hasil pemeriksaan menggunakan fungsi info() yang ditampilkan pada Gambar 2, diketahui bahwa dataset terdiri dari 500 entri data dengan 10 atribut. Seluruh atribut memiliki jumlah data lengkap (non-null), sehingga dapat disimpulkan bahwa tidak terdapat nilai yang hilang pada dataset.

Dataset terdiri dari dua jenis tipe data, yaitu data numerik dan data kategorikal. Data numerik meliputi Usia, Durasi Layar Harian (jam), Kualitas Tidur (1–10), Tingkat Stres (1–10), Hari Tanpa Media Sosial, Frekuensi Latihan (minggu), serta Indeks Kebahagiaan (1–10). Sedangkan kategorikal data meliputi User_ID, Gender, dan Platform Media Sosial.

Hasil ini menunjukkan bahwa dataset berada dalam kondisi yang bersih dan tidak memerlukan proses penanganan data tambahan seperti imputasi, sehingga dapat langsung digunakan untuk tahap analisis selanjutnya.

4.4. Distribusi Target

Analisis target distribusi dilakukan untuk mengetahui sebaran nilai pada variabel Happiness Index (1–10) serta memastikan kualitas data sebelum dilakukan proses pemodelan. Tahap ini penting untuk memahami karakteristik dataset, khususnya dalam melihat keseimbangan antar kelas yang akan diprediksi oleh model.

```
Missing value:
User_ID                       0
Age                           0
Gender                        0
Daily_Screen_Time(hrs)        0
Sleep_Quality(1-10)           0
Stress_Level(1-10)            0
Days_Without_Social_Media     0
Exercise_Frequency(week)      0
Social_Media_Platform         0
Happiness_Index(1-10)         0
dtype: int64

Distribusi target:
Happiness_Index(1-10)
4.0       7
5.0      16
6.0      39
7.0      76
8.0     106
9.0      94
10.0    162
Name: count, dtype: int64
```

Gambar 3. Distribusi Target dan Missing Value

Berdasarkan hasil pengolahan data yang ditampilkan pada Gambar 3, diketahui bahwa seluruh atribut dalam dataset memiliki nilai missing value sebesar 0. Hal ini menunjukkan bahwa dataset berada dalam kondisi yang bersih dan tidak memerlukan proses penanganan data tambahan seperti imputasi

Selain itu, distribusi nilai pada variabel Happiness Index menunjukkan bahwa data tidak tersebar secara merata. Terlihat bahwa jumlah data pada nilai kebahagiaan tinggi, khususnya pada rentang 8 hingga 10, memiliki jumlah yang lebih dominan dibandingkan dengan nilai kebahagiaan rendah seperti 4 hingga 6.

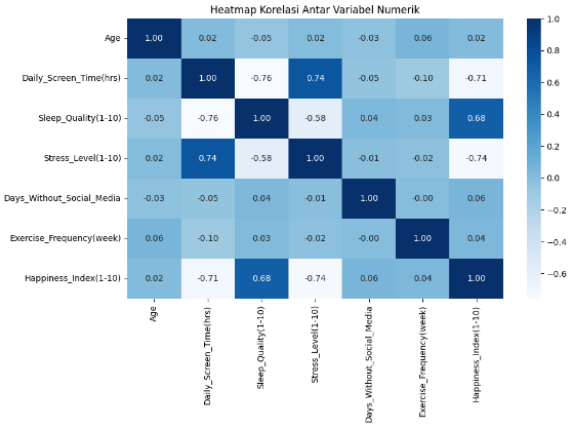
Secara rinci, nilai kebahagiaan tertinggi (nilai 10) memiliki jumlah data paling banyak, diikuti oleh nilai 8 dan 9. Sementara itu, nilai kebahagiaan rendah memiliki jumlah data yang relatif sedikit. Kondisi ini menunjukkan bahwa dataset memiliki distribusi yang tidak seimbang (imbalanced), di mana kelas mayoritas lebih mendominasi dibandingkan kelas minoritas.

Keterbatasan data ini dapat mempengaruhi kinerja model dalam proses klasifikasi. Model cenderung lebih mudah mengenali pola dari kelas mayoritas dibandingkan kelas minoritas, sehingga dapat menyebabkan bias dalam prediksi.

Oleh karena itu, pemahaman terhadap distribusi data menjadi penting dalam mentransmisikan hasil model, terutama dalam melihat kemampuan model dalam mengklasifikasikan seluruh kelas secara merata. Dengan demikian, dapat disimpulkan bahwa dataset yang digunakan dalam penelitian ini tidak memiliki missing value, namun memiliki distribusi kelas yang tidak seimbang, yang berpotensi mempengaruhi kinerja model dalam proses prediksi.

4.5. Analisis Korelasi Antar Variabel

numerik dalam dataset, khususnya hubungan antara variabel gaya hidup digital dengan tingkat kebahagiaan. Korelasi dihitung menggunakan metode Pearson Correlation dan divisualisasikan dalam bentuk heatmap.



Gambar 4. Heatmap Korelasi Atar Variabel

Berdasarkan hasil visualisasi heatmap, terdapat beberapa variabel yang memiliki hubungan signifikan terhadap Happiness Index (1–10). Variabel yang memiliki korelasi paling kuat adalah Tingkat Stres dengan nilai korelasi negatif sebesar -0.74, yang menunjukkan bahwa semakin tinggi tingkat stres, maka tingkat kebahagiaan cenderung menurun.

Selain itu, variabel Daily Screen Time juga menunjukkan korelasi negatif yang cukup kuat terhadap kebahagiaan dengan nilai sekitar -0.71. Hal ini mengindikasikan bahwa penggunaan perangkat digital yang berlebihan dapat berdampak pada penurunan tingkat kebahagiaan individu.

Sebaliknya, variabel Kualitas Tidur memiliki korelasi positif sebesar 0,68 terhadap kebahagiaan. Hal ini menunjukkan bahwa semakin baik kualitas tidur, maka tingkat kebahagiaan individu cenderung meningkat.

Variabel lainnya seperti Frekuensi Latihan dan Hari Tanpa Media Sosial menunjukkan korelasi yang relatif lemah terhadap kebahagiaan, sehingga pengaruhnya tidak terlalu signifikan dalam dataset ini.

Berdasarkan hasil tersebut, dapat disimpulkan bahwa faktor yang paling berpengaruh terhadap kebahagiaan adalah faktor psikologis dan kebiasaan digital, terutama tingkat stres, kualitas tidur, dan durasi penggunaan perangkat digital.

4.6. Preprocessing dan Pembagian Data

Tahap preprocessing dilakukan untuk mempersiapkan data sebelum digunakan dalam proses pemodelan. Proses ini bertujuan untuk memastikan bahwa data berada dalam format yang sesuai agar dapat diproses oleh algoritma pembelajaran mesin dengan optimal.

Pada tahap ini, dilakukan pemisahan antara variabel fitur (X) dan variabel target (y). Variabel target yang digunakan dalam penelitian ini adalah Happiness Index (1–10), sedangkan variabel lainnya digunakan sebagai fitur prediktor.

Selanjutnya data diklasifikasikan menjadi dua jenis, yaitu data numerik dan data kategorikal. Data numerik meliputi atribut seperti Usia, Durasi Layar Harian (jam), Kualitas Tidur (1–10), Tingkat Stres (1–10), Hari Tanpa Media Sosial, dan Frekuensi Latihan (minggu). Sedangkan kategorikal data meliputi atribut Gender dan Social Media Platform.

Untuk mengubah kategori data menjadi bentuk numerik, digunakan teknik One Hot Encoding, sehingga setiap kategori direpresentasikan dalam bentuk variabel biner. Proses ini dilakukan menggunakan ColumnTransformer, yang memungkinkan pengolahan data numerik dan kategorikal dalam satu pipeline.

Setelah tahap preprocessing selesai, dataset yang dibagi menjadi data latih dan data uji menggunakan metode train-test split dengan perbandingan 80:20. Pembagian ini dilakukan untuk melatih model pada sebagian data dan menguji performa model pada data yang belum pernah dilihat sebelumnya.

```
Fitur kategorikal: ['Gender', 'Social_Media_Platform']
Fitur numerik: ['Age', 'Daily_Screen_Time(hrs)', 'Sleep_Quality(1-10)', 'Stress_Level(1-10)', 'Days_Without_Social_Media', 'Exercise_Frequenc
Jumlah data train: 400
Jumlah data test : 100
```

Gambar 5. Pembagian Data dan Fitur

Berdasarkan hasil pembagian data yang ditampilkan pada Gambar 5, diperoleh sebanyak 400 data latih dan 100 data uji. Selain itu, proses pembagian data dilakukan menggunakan parameter stratify, sehingga distribusi kelas pada data latih dan data uji tetap seimbang.

Dengan demikian, data yang telah melalui tahap preprocessing dan pembagian data siap digunakan dalam proses pemodelan menggunakan algoritma C4.5 dan XGBoost.

4.7. Hasil Pemodelan dan Evaluasi

Pada tahap ini dilakukan proses pemodelan menggunakan dua algoritma, yaitu C4.5 dan XGBoost, untuk memprediksi tingkat kebahagiaan berdasarkan data gaya hidup digital. Selanjutnya dilakukan evaluasi terhadap kedua model menggunakan metrik akurasi, presisi, recall, dan F1-score untuk mengetahui kinerja masing-masing algoritma.

Evaluasi dilakukan menggunakan data uji (test data) yang sebelumnya tidak digunakan dalam proses pelatihan model, sehingga hasil yang diperoleh dapat menggambarkan kemampuan model dalam melakukan generalisasi terhadap data baru.

4.8. Hasil Model C4.5

```
=== HASIL C4.5 (Aproksimasi) ===
Accuracy : 0.44
Precision: 0.4676
Recall   : 0.44
F1-Score : 0.4409

Classification Report C4.5:
              precision    recall  f1-score   support

           4       0.00      0.00      0.00         1
           5       0.17      0.33      0.22         3
           6       0.50      0.12      0.20         8
           7       0.31      0.27      0.29        15
           8       0.37      0.33      0.35        21
           9       0.29      0.42      0.34        19
          10       0.74      0.70      0.72        33

    accuracy                           0.44       100
   macro avg       0.34      0.31      0.30       100
weighted avg       0.47      0.44      0.44       100
```

Gambar 6. Hasil Evaluasi Model C4.5

Berdasarkan hasil evaluasi yang ditunjukkan pada Gambar 6, model C4.5 menghasilkan nilai akurasi sebesar 0.44, dengan nilai presisi sebesar 0.4676, recall sebesar 0.44, dan F1-score sebesar 0.4409.

Hasil klasifikasi laporan menunjukkan bahwa model memiliki performa yang bervariasi pada setiap kelas. Model mampu memprediksi dengan cukup baik pada kelas dengan jumlah data yang besar, khususnya pada nilai kebahagiaan tinggi (nilai 10), yang memiliki nilai presisi dan recall yang relatif tinggi.

Namun, pada kelas dengan jumlah data yang sedikit seperti nilai 4, 5, dan 6, model menunjukkan performa yang rendah, bahkan terdapat nilai evaluasi yang mendekati nol. Hal ini menunjukkan bahwa model kesulitan dalam mengenali pola pada kelas minoritas.

Secara keseluruhan, model C4.5 menunjukkan kecenderungan bias terhadap kelas mayoritas, yang dipengaruhi oleh distribusi data yang tidak seimbang.

4.9. Hasil Model XGBoost

```
=== HASIL XGBOOST ===
Accuracy : 0.44
Precision: 0.4516
Recall   : 0.44
F1-Score : 0.4435

Classification Report XGBoost:
              precision    recall  f1-score   support

           4       0.00      0.00      0.00         1
           5       0.00      0.00      0.00         3
           6       0.14      0.12      0.13         8
           7       0.31      0.27      0.29        15
           8       0.29      0.38      0.33        21
           9       0.37      0.37      0.37        19
          10       0.80      0.73      0.76        33

    accuracy                           0.44       100
   macro avg       0.27      0.27      0.27       100
weighted avg       0.45      0.44      0.44       100
```

Gambar 7. Hasil Evaluasi Model XGBoost

Berdasarkan hasil evaluasi pada Gambar 7, model XGBoost menghasilkan nilai akurasi sebesar 0.44, dengan nilai presisi sebesar 0.4516, recall sebesar 0.44, dan F1-score sebesar 0.4435.

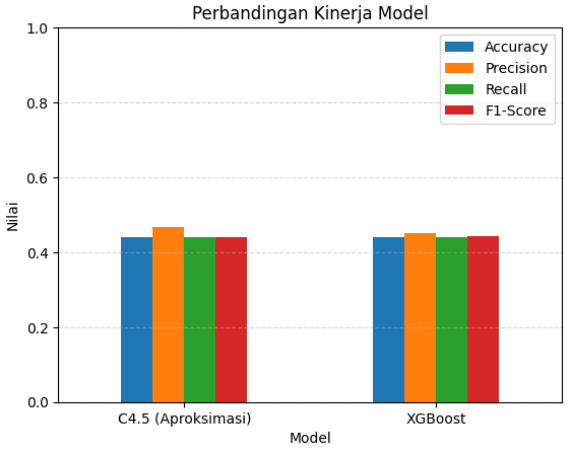
Hasil ini menunjukkan bahwa performa model XGBoost tidak jauh berbeda dengan model C4.5. Meskipun XGBoost merupakan algoritma yang lebih kompleks, model ini tetap mengalami kesulitan dalam memprediksi kelas dengan jumlah data yang sedikit.

Sama seperti model C4.5, XGBoost menunjukkan performa yang lebih baik pada kelas mayoritas, khususnya pada nilai kebahagiaan tinggi. Sebaliknya, pada kelas minoritas, model menghasilkan nilai evaluasi yang rendah, yang menunjukkan keterbatasan dalam mengenali pola pada data tersebut.

Hal ini mengindikasikan bahwa kompleksitas algoritma tidak selalu menjamin peningkatan kinerja, terutama ketika dataset memiliki distribusi yang tidak seimbang.

4.10. Perbandingan Model

Perbandingan model kinerja dilakukan untuk mengetahui perbedaan kinerja antara algoritma C4.5 dan XGBoost berdasarkan metrik evaluasi yang digunakan.



Gambar 8. Perbandingan Kinerja Model

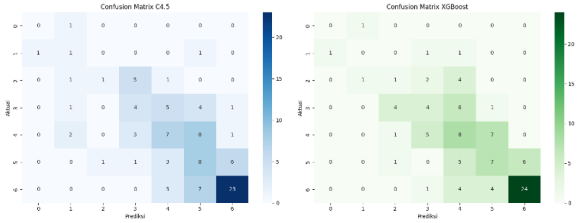
Berdasarkan hasil yang ditampilkan pada Gambar 8, terlihat bahwa kedua model memiliki nilai akurasi, presisi, recall, dan F1-score yang relatif sama. Hal ini menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara kedua algoritma dalam memprediksi tingkat kebahagiaan.

Nilai akurasi yang sama pada kedua model menunjukkan bahwa kedua algoritma memiliki kemampuan yang setara dalam mengklasifikasikan data. Perbedaan kecil pada nilai presisi dan F1-score tidak memberikan pengaruh yang signifikan terhadap performa model secara keseluruhan.

Dengan demikian, dapat disimpulkan bahwa baik C4.5 maupun XGBoost memiliki performa yang sebanding pada dataset yang digunakan dalam penelitian ini.

4.11. Analisis Confusion Matrix

Analisis Confusion Matrix dilakukan untuk mengetahui pola kesalahan klasifikasi yang terjadi pada masing-masing model. Confusion Matrix memberikan gambaran yang lebih detail mengenai jumlah prediksi yang benar dan salah pada setiap kelas.



Gambar 9. Confusion Matrix Model C4.5 dan XGBoost

Berdasarkan Gambar 9, model kedua menunjukkan kecenderungan yang sama, yaitu lebih akurat dalam memprediksi kelas dengan jumlah data yang besar (kelas mayoritas), khususnya pada nilai kebahagiaan tinggi.

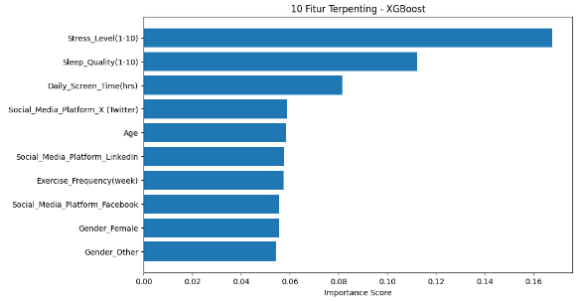
Sebaliknya, pada kelas dengan jumlah data yang sedikit, model sering mengalami kesalahan klasifikasi. Selain itu, terdapat kesalahan prediksi pada kelas yang

memiliki nilai di dekatnya, seperti antara nilai 7, 8, dan 9, yang menunjukkan adanya kesamaan karakteristik antar kelas.

Pola kesalahan ini menunjukkan bahwa model mengalami kesulitan dalam membedakan kelas minoritas serta kelas dengan karakteristik yang saling berdekatan.

4.12. Analisis Feature Importance

Analisis pentingnya fitur dilakukan untuk mengetahui variabel yang paling berpengaruh dalam proses prediksi tingkat kebahagiaan.



Gambar 1. Visualisasi Feature Importance Model XGBoost

Berdasarkan hasil yang ditampilkan pada Tabel dan Gambar 10, variabel Stress Level (1–10) memiliki tingkat kepentingan tertinggi dibandingkan variabel lainnya. Hal ini menunjukkan bahwa tingkat stres merupakan faktor utama yang mempengaruhi kebahagiaan individu.

Selain itu, variabel Sleep Quality (1–10) dan Daily Screen Time (hrs) juga memiliki pengaruh yang cukup besar terhadap model. Variabel lain seperti usia, frekuensi olahraga, serta platform media sosial memiliki kontribusi yang lebih kecil.

Hasil ini sejalan dengan analisis korelasi sebelumnya, yang menunjukkan bahwa faktor psikologis dan kebiasaan penggunaan teknologi memiliki peran penting dalam menentukan tingkat kebahagiaan.

Dengan demikian, dapat disimpulkan bahwa faktor utama yang mempengaruhi kebahagiaan dalam penelitian ini adalah tingkat stres, kualitas tidur, dan durasi penggunaan perangkat digital.

4.13. Ringkasan Hasil Penelitian

Berdasarkan seluruh tahapan penelitian yang telah dilakukan, mulai dari eksplorasi data, preprocessing, hingga pemodelan dan evaluasi, diperoleh beberapa hasil penting yang dapat dirangkum sebagai berikut.

Pertama, berdasarkan hasil analisis data awal, dataset yang digunakan dalam penelitian ini terdiri dari 500 data dengan 10 atribut, yang mencakup variabel gaya hidup digital serta tingkat kebahagiaan sebagai variabel target. Dataset memiliki kombinasi tipe data numerik dan kategorikal, serta tidak ditemukan adanya nilai yang hilang (missing value), sehingga data berada dalam kondisi yang baik untuk dianalisis lebih lanjut.

Kedua, hasil analisis target distribusi menunjukkan bahwa data memiliki distribusi yang tidak seimbang (imbalanced), di mana nilai kebahagiaan tinggi (8–10) memiliki jumlah data yang lebih dominan dibandingkan nilai kebahagiaan rendah (4–6). Kondisi ini mempengaruhi kinerja model dalam proses klasifikasi, karena model cenderung lebih mudah mengenali pola pada kelas mayoritas.

Ketiga, berdasarkan analisis korelasi antar variabel diketahui bahwa Tingkat Stres, Kualitas Tidur, dan Durasi Layar Harian merupakan variabel yang memiliki hubungan paling signifikan terhadap tingkat kebahagiaan. Tingkat Stres dan Waktu Layar Harian menunjukkan korelasi negatif, sedangkan Kualitas Tidur menunjukkan korelasi positif terhadap Indeks Kebahagiaan. Hal ini menunjukkan bahwa faktor psikologis dan kebiasaan penggunaan teknologi memiliki pengaruh utama terhadap kebahagiaan individu.

Keempat, hasil pemodelan menggunakan algoritma C4.5 dan XGBoost menunjukkan bahwa kedua model memiliki performa yang relatif sama, dengan nilai akurasi sebesar 0.44. Selain itu, nilai presisi, recall, dan F1-score pada kedua model juga berada pada kisaran yang tidak jauh berbeda. Hal ini menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara kedua algoritma dalam memprediksi tingkat kebahagiaan pada dataset yang digunakan.

Kelima, berdasarkan analisis hasil konfusi matriks, diketahui bahwa kedua model memiliki kecenderungan yang sama, yaitu lebih akurat dalam memprediksi kelas mayoritas dan kurang optimal dalam memprediksi kelas minoritas. Selain itu, model juga mengalami kesulitan dalam membedakan kelas yang memiliki nilai terdekat, khususnya pada rentang nilai kebahagiaan menengah.

Keenam, hasil analisis feature important menunjukkan bahwa variabel yang paling berpengaruh dalam memprediksi kebahagiaan adalah Stress Level, diikuti oleh Sleep Quality dan Daily Screen Time. Variabel lainnya seperti usia, frekuensi olahraga, serta platform media sosial memiliki pengaruh yang lebih kecil terhadap model.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kinerja model tidak hanya ditentukan oleh jenis algoritma yang digunakan, tetapi juga sangat dipengaruhi oleh karakteristik dataset, seperti distribusi data yang tidak seimbang serta kompleksitas hubungan antar variabel. Oleh karena itu, pemilihan metode serta strategi pengolahan data yang tepat menjadi faktor penting dalam meningkatkan kinerja model.

5. KESIMPULAN

Berdasarkan hasil penelitian yang telah dilakukan mengenai perbandingan algoritma C4.5 dan XGBoost dalam memprediksi tingkat kebahagiaan berdasarkan data gaya hidup digital, dapat disimpulkan bahwa dataset yang digunakan memiliki kualitas yang baik karena tidak ditemukan adanya nilai

yang hilang, sehingga dapat langsung digunakan dalam proses dan pemodelan. Dataset tersebut terdiri dari 500 data dengan 10 atribut yang mencerminkan berbagai aspek gaya hidup digital, seperti durasi penggunaan perangkat, kualitas tidur, tingkat stres, serta aktivitas fisik.

Hasil analisis eksplorasi data menunjukkan bahwa distribusi tingkat kebahagiaan tidak merata, dimana kelas dengan nilai kebahagiaan tinggi lebih dominan dibandingkan kelas dengan nilai rendah. Kondisi ini menyebabkan model cenderung lebih baik dalam memprediksi kelas mayoritas dibandingkan kelas minoritas. Selain itu, analisis korelasi menunjukkan bahwa variabel Stress Level, Sleep Quality, dan Daily Screen Time memiliki hubungan yang signifikan terhadap tingkat kebahagiaan, sehingga dapat dikatakan bahwa faktor psikologis dan kebiasaan penggunaan teknologi memiliki peran penting dalam menentukan kebahagiaan individu.

Berdasarkan hasil pemodelan, baik algoritma C4.5 maupun XGBoost menunjukkan performa yang relatif sama, dengan nilai akurasi sebesar 0.44 serta nilai presisi, recall, dan F1-score yang tidak jauh berbeda. Hal ini menunjukkan bahwa tidak terdapat perbedaan yang signifikan antara kedua algoritma dalam memprediksi tingkat kebahagiaan pada dataset yang digunakan. Selain itu, analisis hasil konfusi matriks menunjukkan bahwa kedua model memiliki kecenderungan yang sama, yaitu lebih akurat dalam memprediksi kelas dengan jumlah data yang besar dan kurang optimal dalam memprediksi kelas minoritas.

Lebih lanjut, hasil analisis feature important menunjukkan bahwa variabel Stress Level merupakan faktor yang paling berpengaruh dalam menentukan tingkat kebahagiaan, diikuti oleh Sleep Quality dan Daily Screen Time. Hal ini memperkuat hasil analisis sebelumnya bahwa kondisi psikologis dan kebiasaan penggunaan teknologi menjadi faktor utama dalam mempengaruhi kebahagiaan individu. Secara keseluruhan, hasil penelitian ini menunjukkan bahwa kinerja model tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga sangat dipengaruhi oleh karakteristik data yang digunakan.

DAFTAR PUSTAKA

[1] Kaur, S. (2021). Internet Usage and Adolescents' Happiness. *Research in Social Change*, *13*(1), 200-210.

[2] Gschwandtner, A., Jewell, S., & Kambhampati, U. S. (2022). Lifestyle and life satisfaction: the role of delayed gratification. *Journal of Happiness Studies*, *23*(3), 1043-1072.

[3] Yamantri, A. B., & Rifa'i, A. A. (2024). Penerapan Algoritma C4. 5 Untuk Prediksi Faktor Risiko Obesitas Pada Penduduk Dewasa. *Jurnal Komputer Antartika*, *2*(3), 118-125.

[4] Badri, M., Alkhaili, M., Aldhaheri, H., Albahar, M., & Alrashdi, A. (2023).Happiness in a digital word-the associations of health, family life, and digitalization perceived challenges-path model for Abu Dhabi. *Journal of Social Science (2720-9938)*, *4*(1).

[5] Harifuddin, H., Anriani, H. B., Asmirah, A., Azuz, F., Iskandar, A. M., & Burchanuddin, A. (2025). Digital Technology Effects on Lifestyles in Marginalized Communities in Indonesia: Post Covid-19 Disaster.

[6] Small, G. W., Lee, J., Kaufman, A., Jalil, J., Siddarth, P., Gaddipati, H., ... & Bookheimer, S. Y. (2020). Brain health consequences of digital technology use. *Dialogues in clinical neuroscience*, *22*(2), 179-187.

[7] Zheng, X., Xue, Y., Dong, F., Shi, L., Xiao, S., Zhang, J., ... & Zhang, C. (2022). The association between health-promoting-lifestyles, and socioeconomic, family relationships, social support, health-related quality of life among older adults in China: a cross sectional study. *Health and quality of life outcomes*, *20*(1), 64.

[8] He, W. W., He, S. L., & Hou, H. L. (2024). Digital economy, technological innovation, and sustainable development. *Plos one*, *19*(7), e0305520.

[9] Godard, R., & Holtzman, S. (2024). Are active and passive social media use related to mental health, wellbeing, and social support outcomes? A meta-analysis of 141 studies. *Journal of Computer-Mediated Communication*, *29*(1), zmad055.

[10] Soundararajan, A., Lim, J. X., Ngiam, N. H. W., Tey, A. J. Y., Tang, A. K. W., Lim, H. A., ... & Ng, K. Y. Y. (2023). Smartphone ownership, digital literacy, and the mediating role of social connectedness and loneliness in improving the wellbeing of community-dwelling older adults of low socio-economic status in Singapore. *Plos one*, *18*(8), e0290557.

[11] Fridayanthie, E. W. (2023). Optimization of Support Vector Machine and XGBoost Methods Using Feature Selection to Improve Classification Performance. *Journal of Informatics and Telecommunication Engineering*, *6*(2), 484-493.