Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) p-ISSN: 2355-7699
Vol. 13, No. 2, April 2026, hlm. 487-498 e-ISSN: 2528-6579

PREDIKSI RESIKO PENGGUNAAN MEDIA SOSIAL TERHADAP KESEHATAN MENTAL MENGGUNAKAN *EXPLORATORY DATA ANALYSIS (EDA)* DAN *CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING* (CRISP-DM)

Fitri Ayuning Tyas[*1], Azhar Basir[2], Amira Elistya Ardhin[3]

[1,2,3] Universitas Muhammadiyah Brebes, Kabupaten Brebes
Email: [1]tyas_fa@umbs.ac.id, [2]azhar.bs@umbs.ac.id, [3]amiraelis34@gmail.com
[*]Penulis Korespondensi

(Naskah masuk: 24 Maret 2025, diterima untuk diterbitkan: 17 April 2026)

Abstrak

Media sosial telah menjadi bagian penting dalam kehidupan masyarakat, namun peningkatan penggunaannya sering dikaitkan dengan dampak negatif terhadap kesehatan mental seperti stres, adiksi, FoMo, dan insomnia. Upaya prediksi risiko penggunaan media sosial dapat membantu menjaga kesehatan mental dengan memanfaatkan teknik data mining. Penelitian ini menggunakan metodologi CRISP-DM sebagai kerangka utama serta *Exploratory Data Analysis* (EDA) untuk mengidentifikasi tren dan anomali yang mendukung proses pemodelan. Beberapa algoritma *supervised learning* seperti C4.5, *k*-NN, dan *Naïve Bayes* diterapkan untuk memprediksi dampak negatif penggunaan media sosial terhadap kesehatan mental. Hasil eksperimen menunjukkan bahwa Naïve Bayes memberikan kinerja terbaik dengan akurasi tertinggi sebesar 92,5%, melampaui C4.5 dan k-NN. Integrasi EDA dan CRISP-DM terbukti menghasilkan model prediksi yang akurat, meskipun penerapan EDA memerlukan waktu tambahan dalam analisis. CRISP-DM berperan penting dalam menyediakan kerangka kerja yang sistematis sehingga membantu peneliti bekerja lebih terstruktur dan mengurangi risiko kesalahan. Selain itu, temuan memperlihatkan bahwa semakin lama seseorang menggunakan media sosial, semakin besar dampak negatif yang dialami, terutama bagi mereka yang menghabiskan waktu lebih dari lima jam per hari. Secara keseluruhan, hasil penelitian ini memberikan kontribusi terhadap pengembangan model prediksi berbasis data mining dan dapat menjadi landasan bagi upaya pencegahan gangguan kesehatan mental akibat penggunaan media sosial.

Kata kunci: *CRISP-DM, Exploratory Data Analysis, Media Sosial, Kesehatan Mental*

PREDICTING THE IMPACT OF SOCIAL MEDIA USE ON MENTAL HEALTH THROUGH THE USE OF EXPLORATORY DATA ANALYSIS (EDA) AND THE CROSS-INDUSTRY STANDARD DATA MINING PROCESS (CRISP-DM)

Abstract

Social media has become an integral part of modern life, enabling users to express feelings and opinions. However, its increasing use has been linked to negative impacts on mental health, such as stress, addiction, FoMo, and insomnia. Predicting the risks associated with social media use can help maintain mental well-being, and this can be achieved through data mining techniques. This study applies the CRISP-DM methodology as the main framework, complemented by Exploratory Data Analysis (EDA) to identify trends and anomalies that support the modeling process. Several supervised learning algorithms, including C4.5, k-NN, and Naïve Bayes, were employed to predict the negative impact of social media use on mental health. Experimental results show that Naïve Bayes achieved the best performance with the highest accuracy of 92.5%, outperforming both C4.5 and k-NN. The integration of EDA and CRISP-DM proved effective in producing accurate predictive models, although EDA required additional time for data analysis. CRISP-DM played a crucial role in providing a systematic framework, enabling researchers to work more structurally and minimizing the risk of errors. Furthermore, findings indicate that the longer individuals spend on social media, the greater the negative impact they experience, particularly among those using it for more than five hours per day. Overall, this study contributes to the development of predictive models based on data mining and provides insights that may support preventive efforts against mental health issues caused by excessive social media use.

Keywords: *CRISP-DM, Exploratory Data Analysis, Social Media, Mental Health*

1. PENDAHULUAN

Kemajuan teknologi internet telah merubah cara masyarakat berkomunikasi secara signifikan [1]. Media sosial digemari masyarakat untuk berkomunikasi dan berbagi informasi. Indonesia mencatatkan 191 juta pengguna aktif media sosial pada Januari 2022 dan mengalami peningkatan

signifikan sebesar 12,35% dari tahun sebelumnya [1]. Survei lain mencatat 57% remaja usia 13 tahun memiliki keinginan untuk memeriksa akun media sosial mereka paling tidak enam kali sehari meskipun tidak mengunggah sesuatu, melainkan *stalking* [2]. Partisipasi aktif (konten kreator) atau pasif (pengintai) [3] dalam komunitas media sosial merupakan teman sehari-hari bagi hampir setiap individu. Berdasarkan data tersebut disimpulkan bahwa media sosial telah menjadi bagian yang tidak terpisahkan dari masyarakat [4]. Seiring peningkatan penggunaan media sosial, sisi gelapnya dapat muncul yakni media sosial sebagai pemicu stres [4] bagi pengintai maupun konten kreator. Konten kreator mengunggah konten dengan tujuan utama mencapai jumlah *like* tertentu dan mendapatkan manfaat dari komentar, tetapi jika target *like* tidak tercapai dan muncul komentar negatif, hal itu dapat menyebabkan stres [4]. Fenomena lain yang disebabkan media sosial antara lain munculnya informasi yang berlebihan [5], perundungan siber [6], penyebaran hoax [7], perilaku adiktif [8], dan *Fear of Missing Out* (FoMO) [2]. Fenomena-fenomena tersebut mempengaruhi kesehatan mental seseorang.

Kesehatan mental, fisik, dan sosial adalah untaian kehidupan yang saling terkait [9]. Kesehatan mental yang buruk dapat berdampak fatal seperti menyebabkan perilaku bunuh diri. Bunuh diri merupakan salah satu penyebab utama kematian tidak wajar di seluruh dunia dan merupakan masalah kesehatan masyarakat [10]. Mengidentifikasi perilaku bunuh diri penting dilakukan untuk meningkatkan deteksi di masa mendatang [10]. Dalam dunia kesehatan, teknologi telah menjadi alat yang berguna untuk mengeksplorasi berbagai metode dalam upaya mendeteksi dan mengklasifikasikan jenis penyakit yang mungkin dialami seseorang [11]. Deteksi atau prediksi resiko penggunaan media sosial dapat dilakukan sebagai upaya menjaga kesehatan mental. Permasalahan prediksi dapat diatasi menggunakan teknik *data mining*. *Data mining* adalah disiplin ilmu untuk mendapatkan pengetahuan berharga dari data melalui model matematika dan analisis [12]. *Data mining* muncul sebagai proses yang memungkinkan anomali, pola, dan korelasi ditemukan dalam kumpulan data besar [9]. Beberapa faktor yang dapat mempengaruhi kesehatan mental dapat digunakan sebagai dasar penentuan prediksi resiko penggunaan media sosial terhadap kesehatan mental. Penerapan *data mining* dapat menemukan korelasi antar faktor tersebut dan menghasilkan pola prediksi.

Penerapan *data mining* di berbagai bidang kini umum dilakukan karena ketersediaan data dan manfaat yang sudah terbukti. Beberapa metodologi diusulkan untuk memandu proses *data mining*, salah satunya adalah *Cross Industry Standard Process for Data Mining* (CRISP-DM) yang telah menjadi "standar *de facto* untuk mengembangkan proyek *data mining*" [13] dan telah diterapkan ke berbagai domain sejak didefinisikan dua puluh tahun lalu [14]. Hingga saat ini CRISP-DM masih menjadi metodologi populer dalam praktik dan penelitian *data mining* [12]. CRISP-DM terdiri dari enam fase iteratif mulai dari *business understanding* hingga *deployment* [15]. *Business understanding*, *data understanding* dan *data preparation* dianggap sebagai keunggulan CRISP-DM karena membantu dalam memperoleh lebih banyak pengetahuan tentang tujuan proses bisnis dan ketersediaan data [16]. *Metode Exploratory Data Analytic* (EDA) dapat mendeteksi kesalahan, menemukan data yang sesuai, memeriksa asumsi dan menentukan korelasi di antara variabel [17]. EDA menjadi langkah awal dalam menganalisis atau memahami data sebelum melakukan pemodelan prediktif lebih lanjut. Sehingga EDA dimungkinkan dapat mendukung fase *data understanding* dan *data preparation* pada CRISP-DM.

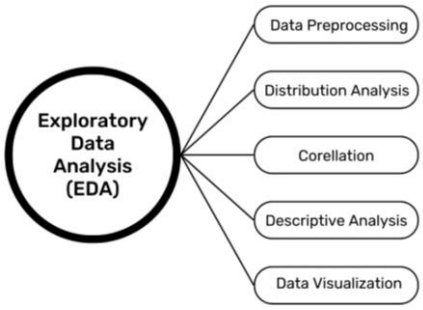
WHO (2022) melaporkan adanya peningkatan masalah kesehatan mental pada masyarakat modern, sehingga penting untuk lebih memahami bagaimana media sosial membentuk jiwa dan bagaimana individu dapat mengatasi pengaruh negatif sebagai suatu resiko. Penelitian ini menekankan pentingnya analisis data sebagai dasar sebelum pemodelan, di mana eksplorasi menyeluruh melalui EDA dapat mengungkap tren, hubungan, maupun anomali yang relevan. Dengan mengintegrasikan EDA ke dalam kerangka CRISP-DM, penelitian ini bertujuan merumuskan strategi prediksi risiko penggunaan media sosial terhadap kesehatan mental secara lebih efektif dan akurat.

2. METODE PENELITIAN

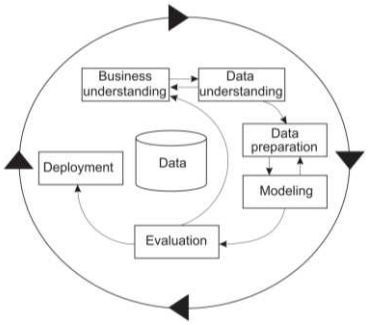
Penelitian ini mengkaji data penggunaan media sosial dan kesehatan mental untuk mengetahui pola prediksi resiko yang dihasilkan berdasarkan penerapan proses *data mining*. Metode penelitian yang digunakan dalam penelitian ini adalah metode eksperimen. Eksperimen dilakukan dengan mengintegrasi EDA dan CRISP-DM. EDA mendukung proses *data mining* CRISP-DM untuk mengeksplorasi setiap pengumpulan data, pengamatan pola dan identifikasi detail-detail penting dalam kumpulan data. Ekplorasi data yang dapat dilakukan oleh EDA secara umum digambarkan pada Gambar 1, sedangkan proses *data mining* CRISP-DM digambarkan pada Gambar 2.

Gambar 1 menunjukan beberapa teknik EDA yang umum dilakukan. *Data preprocessing* merupakan teknik prapemrosesan data yang meliputi identifikasi dan penanganan *outlier, missing value,* serta normalisasi atau standarisasi data. *Distribution analysis* digunakan untuk menganalisis setiap distribusi variabel dalam data. *Correlation* digunakan untuk mencari tingkat hubungan antara dua variabel. *Descriptive analysis* dapat menampilkan beberapa informasi penting seperti nilai rata-rata, median, modus, standar deviasi, dan variansi. Hasil dari teknik tersebut dapat divisualisasikan dalam bentuk

histogram, grafik, *heatmaps,* dan lain sebagainya pada fase *data visualization*. *Data visualization* merupakan komponen penting EDA karena menawarkan representasi grafis yang membantu dalam membedakan pola, mengidentifikasi *outlier*, dan memahami distribusi data.



Gambar 1. Exploratory Data Analysis

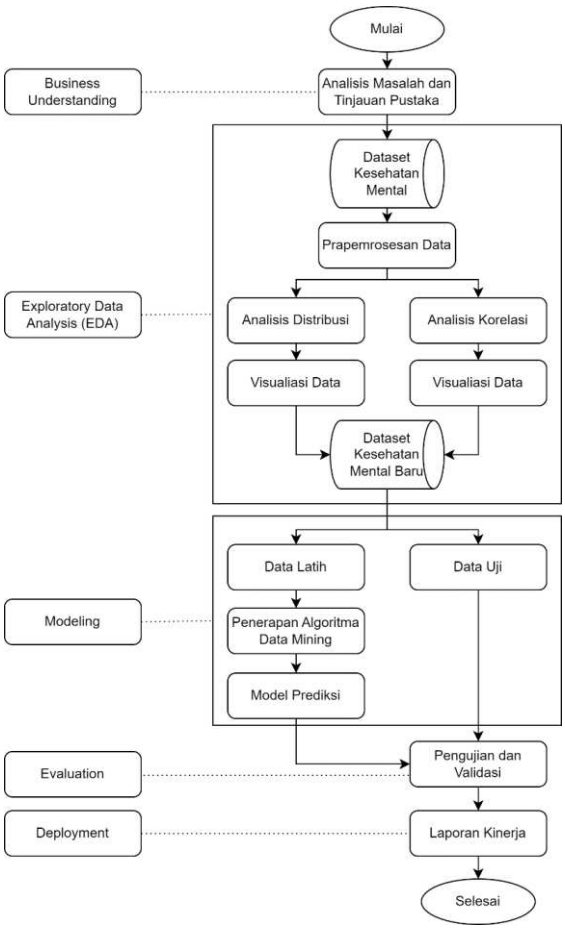


Gambar 2. CRISP-DM [13]

Gambar 2 menjelaskan tahapan proses *data mining* CRISP-DM yang memiliki enam fase iteratif yakni *business understanding, data understanding, data preparation, modeling, evaluation,* dan *deployment* [15]. Fase-fase tersebut memiliki tugas dan hubungannya masing-masing. Deskripsi masing-masing fase [15][12] dirangkum pada Tabel 1. Sedangkan tahapan eksperimen integrasi EDA dan CRISP-DM pada penelitian ini digambarkan pada Gambar 3.

Tabel 1. Deskripsi Fase CRISP-DM

No	Fase	Deskripsi
1	*Business understanding*	Fase pemahaman tujuan dan kebutuhan proyek dari sisi bisnis, kemudian mengonversi pengetahuan menjadi definisi masalah *data mining* (prediksi), menetapkan kriteria keberhasilan (kinerja algoritma), serta menyusun rencana awal untuk mencapai tujuan yang ditetapkan.
2	*Data understanding*	Fase mengumpulkan, mengeksplorasi, mendeskripsikan, dan memeriksa kualitas data, menemukan pengetahuan awal terhadap data atau mendeteksi subset menarik untuk membentuk hipotesis dari informasi yang tersembunyi.
3	*Data preparation*	Fase mempersiapkan data yang mencakup semua aktivitas untuk membangun kumpulan data akhir (data yang akan dimasukkan ke dalam alat pemodelan).
4	*Modeling*	Fase pemodelan data terdiri dari pemilihan teknik *data mining*, pembuatan data uji, dan model serta menerapkan parameter tertentu sesuai dengan masalah bisnis dan data.
5	*Evaluation*	Fase menilai atau mengevaluasi model berdasarkan kriteria evaluasi dan mengambil keputusan tentang penggunaan hasil *data mining.*
6	*Deployment*	Fase ini mencakup perencanaan penggunaan model yang dituangkan dalam panduan atau laporan akhir.



Gambar 3. Tahapan Eksperimen Integrasi EDA dan CRISP-DM

Gambar 3 menggambarkan tahapan eksperimen integrasi EDA dan CRISP-DM dalam memprediksi resiko penggunaan media sosial terhadap kesehatan mental. Tahap *data understanding* dan *data preparation* pada fase CRISP-DM digantikan dengan teknik EDA yang meliputi pramerosesan data, analisis distribusi, analisis korelasi, dan visualisasi data. Teknik EDA tersebut mengeksplorasi data secara lebih spesifik. EDA juga akan menghasilkan *dataset* berlabel yang disebut sebagai *dataset* kesehatan mental baru *(new dataset).* Pelabelan bertujuan untuk menentukan *cluster* atau kelompok dalam data. Dalam penelitian ini, algoritma *k*-means digunakan untuk mengelompokkan data, sementara metode *elbow* digunakan untuk menentukan jumlah *cluster* optimal. Menurut penelitian [18], *k*-means adalah salah satu metode yang dapat digunakan untuk mengidentifikasi kategori dalam data dengan

mengelompokkan *n* observasi ke dalam *k cluster*. Setiap observasi akan dimasukkan ke dalam *cluster* dengan rata-rata terdekat yang berfungsi sebagai prototipe *cluster*. Untuk menjelaskan serta memverifikasi konsistensi hasil pengelompokan, dapat diterapkan metode *elbow*, yang bertujuan membantu menentukan jumlah *cluster* optimal dalam data. Metode *elbow* menentukan jumlah *cluster* optimal dengan cara memperhatikan persentase hasil perbandingan antara jumlah *cluster* yang akan membentuk sudut pada titik tertentu [19].

Pada tahap *modeling* data latih *(training)* dan data uji *(testing)* akan dimodelkan menggunakan algortima prediksi yakni C4.5, *k*-NN, dan *Naïve Bayes.* Ketiga algoritma tersebut dipilih karena memiliki beberapa kelebiban. C4.5 memiliki keunggulan mudah diinterpretasikan karena memiliki struktur yang sederhana [20]. Model yang dihasilkan oleh C4.5 berupa pohon keputusan *(decision tree)*. Terkadang pohon keputusan tersebut memiliki ukuran besar dikarenakan ada cabang pohon yang tidak penting atau sering disebut dengan istilah *overfitting* [21]. Masalah *overfitting* dapat di atasi dengan teknik *pruning* untuk memotong atau menghilangkan beberapa cabang yang tidak diperlukan. *k*-NN memiliki konsep yang sederhana dan mudah untuk diterapkan, di mana algoritma ini beroperasi dengan membandingkan kesamaan antara satu data dengan data lainnya [22]. Dalam *k*-NN, pemilihan nilai *k* menjadi faktor krusial karena ditentukan secara subjektif, dan disarankan agar nilai *k* dipilih dalam bilangan ganjil [22]. Sedangkan *Naïve Bayes* merupakan salah satu metode statistika yang berguna untuk proses penentuan probabilitas keanggotaan suatu kelas atau label [23]. Tingkat akurasi terbaik antara ketiga algoritma tersebut dapat ditentukan dengan melakukan perbandingan.

Perbandingan algoritma bertujuan untuk mendapatkan kinerja algoritma terbaik yang diukur dari nilai tingkat akurasi, *precision*, *recall* dan *F1-measure*. Nilai-nilai tersebut dihasilkan menggunakan metode validasi *10-fold cross-validation* pada tahap *evaluation.* Metode validasi *10-fold cross-validation* bekerja dengan cara mempartisi himpunan *dataset* menjadi 10 *fold* yang saling bebas: f_1, f_2, …., f_{10}, sehingga masing-masing *fold* berisi 1/10 bagian *dataset*. Selanjutnya 10 himpunan dataset: D_1, D_2, …., D_{10} masing-masing berisi 9 *fold* sebagai data latih dan 1 *fold* sebagai data uji, setiap *fold* akan menjadi data uji sebanyak satu kali.

Tahap *deployment* merupakan tahap terakhir dari eksperimen integrasi EDA dan CRISP-DM. Pada tahap ini perencanaan penggunaan hasil eksperimen dapat dituangkan dalam laporan analisis dan hasil prediksi resiko penggunaan media sosial terhadap kesehatan mental. Tahap *deployment* juga bertujuan sebagai gambaran penelitian yang dapat dikembangkan pada penelitian selanjutnya.

Seluruh ekspertimen tahap integrasi EDA dan CRISP-DM pada penelitian ini dilakukan menggunakan bahasa pemrograman *Python* dengan *Google Colab* sebagai IDE.

3. HASIL DAN PEMBAHASAN

Eksperimen yang dilakukan pada penelitian ini terdiri atas integrasi EDA dan CRISP-DM untuk memprediksi resiko penggunaan media sosial terhadap kesehatan mental. Hasil eksperimen EDA berupa data baru yang siap dimodelkan melalui tahap berikutnya yakni eksperimen CRISP-DM.

3.1. *Business Understanding*

Analisis masalah dan tinjauan pustaka dilakukan untuk memahami tujuan bisnis dalam penelitian ini yakni mengidentifikasi masalah dalam memprediksi resiko penggunaan media sosial terhadap kesehatan mental. Pada tahap *business understanding*, diperoleh wawasan mengenai berbagai faktor yang memengaruhi kesehatan mental pengguna media sosial seperti usia, jenis kelamin, durasi penggunaan media sosial, jenis *platform* media sosial yang digunakan dan sebagainya. Dampak negatif yang mungkin muncul antara lain kesulitan berkonsentrasi, perasaan tertekan atau sedih, gangguan tidur, serta perilaku adiktif seperti kecemasan saat tidak menggunakan media sosial, mencari validasi di *platform* tersebut, dan perbandingan diri dengan orang lain yang dianggap ideal. Wawasan tersebut dapat dijadikan dasar untuk memprediksi resiko dampak penggunaan media sosial terhadap kesehatan mental.

Wawasan tambahan yang diperoleh pada tahap ini adalah penerapan *data mining* dapat mengungkap hubungan antara faktor-faktor yang memengaruhi kesehatan mental pengguna media sosial dan menghasilkan pola prediksi resiko dampak negatif. Pola prediksi tersebut dikategorikan kedalam resiko rendah, sedang dan tinggi. Berdasarkan pola prediksi tersebut, dapat ditarik kesimpulan mengenai faktor-faktor paling dominan dalam risiko tinggi, yang kemudian dapat digunakan sebagai dasar dalam pengambilan keputusan.

Penelitian [24] mengembangkan sistem pendukung kesehatan mental mahasiswa berbasis kecerdasan buatan yang dirancang untuk memberikan layanan personal sekaligus memprediksi potensi krisis psikologis. Hasilnya menunjukkan efektivitas sistem ini dalam mendukung kesehatan mental serta deteksi dini risiko krisis pada mahasiswa. Penelitian [25] memanfaatkan *Streamlit* untuk membangun *smart web* yang mendukung implementasi sistem prediksi kesehatan mental berbasis *machine learning*, sehingga hasil analisis dapat ditampilkan secara real-time dengan tampilan interaktif dan mudah digunakan oleh pendidik maupun tenaga kesehatan mental. Penelitian [26] mengembangkan teknik prediksi kesehatan mental dengan memanfaatkan model BERT yang dilatih menggunakan data teks berlabel dari aplikasi *Lyf Support*, sehingga mampu

mengidentifikasi percakapan terkait kesehatan mental secara akurat. Pendekatan ini terbukti efektif dalam mendeteksi indikasi gangguan psikologis melalui analisis teks. Ketiga penelitian tersebut sama-sama memanfaatkan kecerdasan buatan dan *machine learning* untuk mendukung prediksi serta deteksi dini kesehatan mental melalui analisis data teks, perilaku, maupun psikologis. Meskipun berhasil menunjukkan efektivitas sistem prediksi dengan tingkat akurasi yang tinggi, penelitian-penelitian tersebut belum secara spesifik mengungkap faktor-faktor dominan yang berperan sebagai risiko utama dalam masalah kesehatan mental.

3.2. *Exploratory Data Analysis* (EDA)

Dataset yang digunakan dalam penelitian ini adalah data *social media mental health* yang bersumber dari https://www.kaggle.com/datasets. *Dataset* tersebut merupakan data kesehatan mental pengguna media sosial dengan 483 *record* dan 20 attribut. Deskripsi attribut tersebut dirangkum pada Tabel 2.

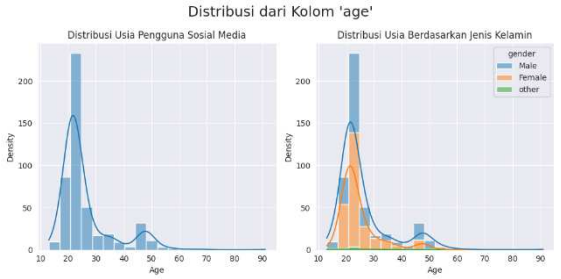
Tabel 2. *Social Media Mental Health Dataset*

No	Attribut	Deskripsi
1	*Age*	Usia
2	*gender*	Jenis kelamin
3	*relationship*	Status hubungan
4	*occupation*	Jenis pekerjaan
5	*affiliate_organization*	Organisasi afiliasi
6	*social_media_use*	Media sosial yang digunakan
7	*Platforms*	Platform media sosial
8	*avg_time_per_day*	Rata-rata waktu per hari
9	*without_purpose*	Tanpa tujuan
10	*Distracted*	Gangguan perhatian
11	*Restless*	Gelisah
12	*distracted_ease*	Mudah terganggu
13	*Worries*	Gangguan kecemasan
14	*concentration*	Gangguan konsentrasi
15	*compare_to_others*	Perbandingan social
16	*compare_feelings*	Perbandingan perasaan
17	*Validation*	Butuh validasi
18	*depressed*	Depresi
19	*daily_activity_flux*	Perubahan aktivitas harian
20	*sleeping_issues*	Gangguan tidur

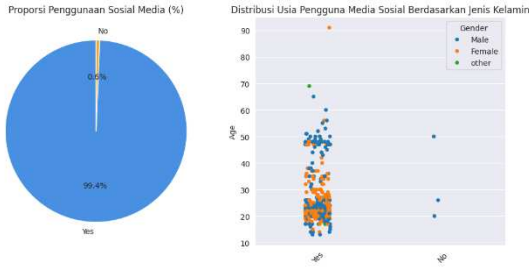
3.2.1. *Distribution Analysis* dan *Data Visualization*

Teknik EDA yang diterapkan pada penelitian ini meliputi *distribution analysis* (analisis distribusi) dan *data visualization* (visualisasi data). Analisis distribusi dilakukan untuk mengetahui persebaran data atau distribusi variable dalam *dataset* sehingga dapat membantu dalam pemilihan model prediksi. Hasil analisis distribusi disajikan dalam bentuk visualisasi data pada Gambar 4 s.d. Gambar 15.

Gambar 4 menunjukkan distribusi pengguna media sosial berdasarkan usia dan jenis kelamin. Mayoritas pengguna berusia 20-30 tahun dengan puncak di usia 22-25 tahun dan pengguna *male* (pria) lebih dominan. Secara keseluruhan, mayoritas pengguna berasal dari kelompok usia muda.



Gambar 4. Distribusi Pengguna Media Sosial Berdasarkan Usia dan Jenis Kelamin



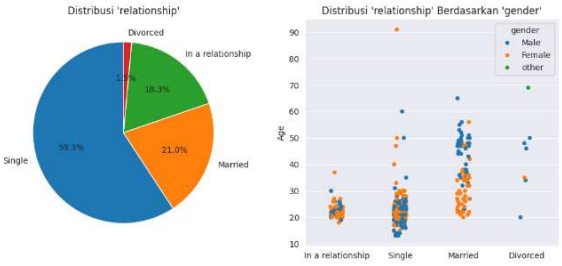
Gambar 5. Distribusi Usia Pengguna Media Sosial Berdasarkan Jenis Kelamin dan Proporsi Penggunaan Media Sosial

Gambar 5 menampilkan visualisasi data terkait persentase individu yang menggunakan atau tidak menggunakan media sosial. 99,4% individu dalam *dataset* menggunakan media sosial dengan rentang usia antara 15-40 tahun, dan 0,6% tidak menggunakan media sosial. Secara keseluruhan, visualisasi ini menunjukkan bahwa hampir semua individu dalam *dataset* menggunakan media sosial, dengan mayoritas pengguna berusia muda *male* (pria) maupun *female* (wanita).

Berdasarkan Gambar 4 dan Gambar 5 terlihat kecenderungan dominasi kelompok usia muda, khususnya 20–30 tahun, menunjukkan bahwa media sosial telah menjadi ruang interaksi utama pada fase kehidupan di mana individu sedang aktif membangun jejaring sosial, identitas diri, dan peluang karier. Dominasi pengguna pria juga dapat mengindikasikan adanya perbedaan pola akses maupun preferensi penggunaan media sosial antar gender, yang berpotensi memengaruhi jenis konten yang dikonsumsi maupun dampak psikologis yang ditimbulkan. Temuan ini penting karena mengisyaratkan bahwa analisis lebih lanjut mengenai perilaku, kebutuhan, dan kerentanan kelompok usia muda perlu dilakukan secara spesifik, mengingat mereka merupakan segmen dengan intensitas penggunaan tertinggi sekaligus paling rentan terhadap dampak negatif media sosial.

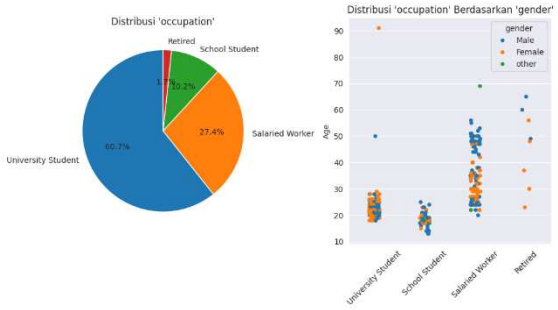
Gambar 6 menampilkan visualisasi data terkait status pengguna media sosial. Berdasarkan diagram *pie* mayoritas individu berstatus *single* (lajang) mencakup 59,3%, *married* (menikah) mencakup 21,0%, *In a relationship* (sedang dalam hubungan) mencakup 18,3%, dan *divorced* (bercerai) adalah kelompok terkecil, hanya 1,5% dari total populasi.

Hal ini menunjukkan bahwa sebagian besar individu dalam *dataset* adalah lajang.



Gambar 6. Distribusi Pengguna Media Sosial Berdasarkan Status Hubungan

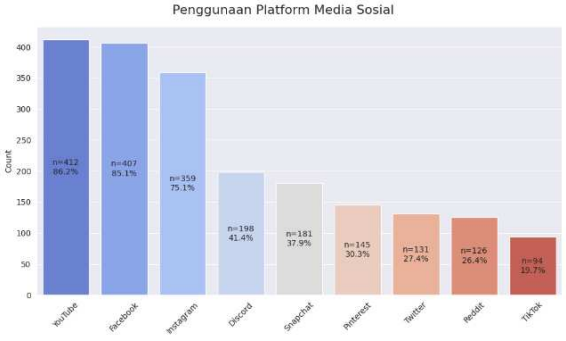
Sementara berdasarkan *scatter plot* mayoritas individu dalam kategori lajang dan sedang dalam hubungan berusia antara 20-35 tahun. Kategori menikah cenderung memiliki usia lebih tua, dengan rentang usia 30-50 tahun. Kategori bercerai memiliki jumlah yang sangat sedikit dan cenderung berusia lebih tua. Secara keseluruhan, visualisasi ini menunjukkan bahwa sebagian besar individu dalam *dataset* masih lajang, dan terdapat hubungan antara usia serta status hubungan, di mana individu yang lebih tua cenderung sudah menikah atau bercerai.



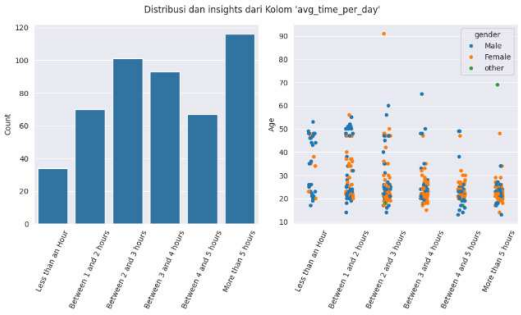
Gambar 7. Distribusi Pengguna Sosial Media Berdasarkan *Occupation*

Gambar 7 menunjukkan distribusi *occupation* (pekerjaan) pengguna media sosial. Mayoritas adalah *university student* (mahasiswa) (60,7%), diikuti *salaried worker* (pekerja bergaji) (27,4%), *school student* (pelajar sekolah) (10,2%), dan *retired* (pensiunan) (1,7%). *Scatter plot* menunjukkan mahasiswa dan pelajar sekolah berusia 10-25 tahun, pekerja bergaji 25-50 tahun, dan pensiunan 60-70 tahun. Tidak ada perbedaan signifikan antara pria dan wanita dalam tiap kategori pekerjaan. Secara keseluruhan, sebagian besar individu masih dalam dunia pendidikan, dan usia sangat berpengaruh terhadap jenis pekerjaan.

Gambar 8 menunjukkan penggunaan *platform* media sosial. *YouTube* (86,2%) dan *Facebook* (85,1%) memiliki jumlah pengguna tertinggi, diikuti *Instagram* (75,1%). *Discord* (41,4%) dan *Snapchat* (37,9%) berada di tingkat menengah, sementara *Pinterest* (30,3%), *Twitter* (27,4%), dan *Reddit* (26,4%) memiliki lebih sedikit pengguna. *TikTok* (19,7%) adalah yang paling sedikit digunakan dalam *dataset* ini, meskipun meskipun dalam tren global *platform* ini sedang berkembang pesat. Secara keseluruhan *platform* berbagi konten visual seperti *YouTube, Facebook,* dan *Instagram* adalah *platform* dominan, sementara *platform* berbasis komunitas atau diskusi seperti *Discord, Reddit,* dan *Twitter* memiliki memiliki jumlah pengguna yang terbatas.

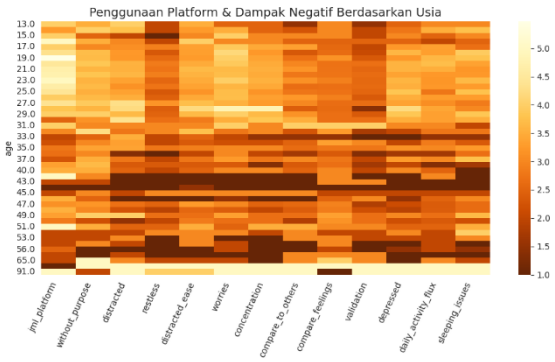


Gambar 8. Distribusi Pengguna *Platform* Media Sosial



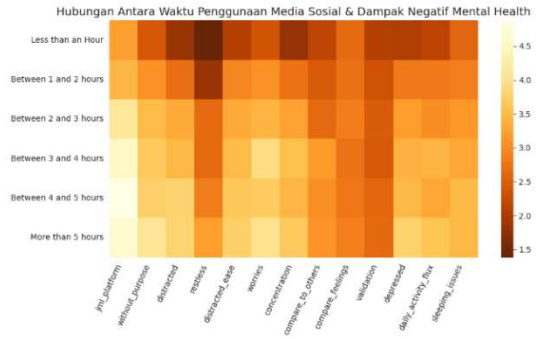
Gambar 9. Distribusi Durasi Waktu Penggunaan Media Sosial

Gambar 9 menunjukkan pola penggunaan waktu media sosial berdasarkan usia dan jenis kelamin. Kategori *more than 5 hours* (lebih dari 5 jam) memiliki jumlah tertinggi, sedangkan *less than an hour* (kurang dari 1 jam) memiliki jumlah terendah. Tren menunjukkan semakin lama waktu penggunaan, semakin banyak individu dalam kategori tersebut. *Scatter plot* menunjukkan bahwa individu muda cenderung menghabiskan lebih banyak waktu di media sosial.



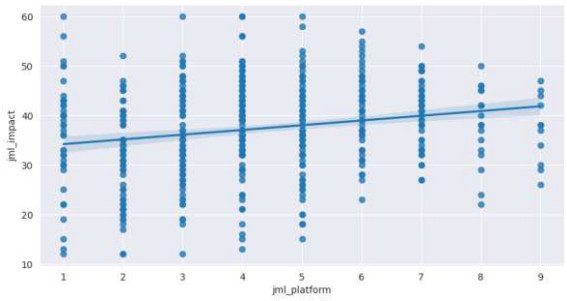
Gambar 10. Hubungan Penggunaan *Platform* Media Sosial & Dampak Negatif Berdasarkan Usia

Gambar 10 adalah *heatmap* yang menunjukkan hubungan antara usia, jumlah *platform* yang digunakan, dan dampak negatif penggunaan *platform*. Warna terang menandakan dampak lebih tinggi, sedangkan warna gelap lebih rendah. Usia 13-30 tahun lebih sering mengalami dampak seperti *distracted* (gangguan perhatian), *restless* (gelisah), *compare_to_others* (perbandingan sosial), dan *validation* (butuh validasi), sedangkan usia 40+ cenderung lebih rendah, kecuali untuk depresi dan *sleeping_issues* (gangguan tidur). Dampak seperti *daily_activity_flux* (fluktuasi/ perubahan aktivitas harian) dan *concentration* (gangguan konsentrasi) bervariasi di semua usia.



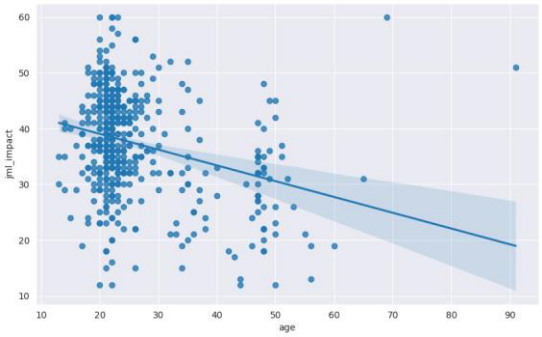
Gambar 11. Hubungan Antara Waktu Pengunaan Media Sosial dengan Dampak Negatif Terhadap Kesehatan Mental

Gambar 11 adalah *heatmap* yang menunjukkan hubungan antara durasi penggunaan media sosial dan dampak negatif terhadap kesehatan mental. Warna terang menandakan dampak lebih tinggi, sedangkan warna gelap lebih rendah. Penggunaan <1 jam per hari dikaitkan dengan dampak negatif minimal, sementara durasi lebih lama meningkatkan risiko gangguan perhatian, perbandingan sosial, dan kecemasan. Penggunaan 3-5 jam per hari menunjukkan dampak tertinggi, meskipun tidak selalu meningkat. Mengurangi waktu penggunaan dapat membantu mengurangi kecemasan, gangguan tidur, dan fluktuasi aktivitas harian.



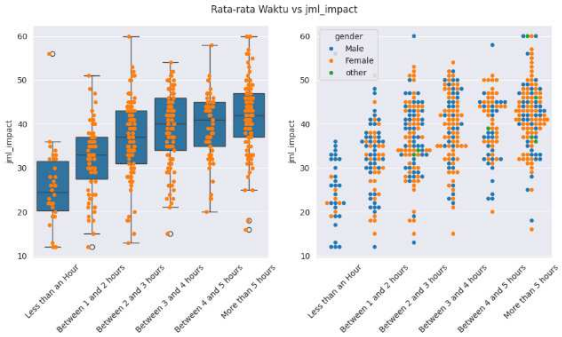
Gambar 12. Hubungan Antara Jumlah *Platform* Media Sosial dengan Jumlah Dampak Negatif

Gambar 12 adalah *scatter plot* yang menunjukkan bahwa penggunaan lebih banyak *platform* media sosial cenderung meningkatkan dampak negatif, meskipun tidak secara drastis. Terdapat variasi individu, di mana beberapa pengguna dengan sedikit *platform* mengalami dampak tinggi, dan sebaliknya. Kemiringan garis regresi yang landai menunjukkan bahwa faktor lain seperti jenis *platform* dan pola penggunaan juga berperan. Secara keseluruhan, meskipun jumlah *platform* yang lebih tinggi dikaitkan dengan lebih banyak dampak negatif, efeknya tidak mutlak dan dipengaruhi oleh kebiasaan penggunaan individu.



Gambar 13. Hubungan Antara Usia dengan Jumlah Dampak Terhadap Kesehatan Mental

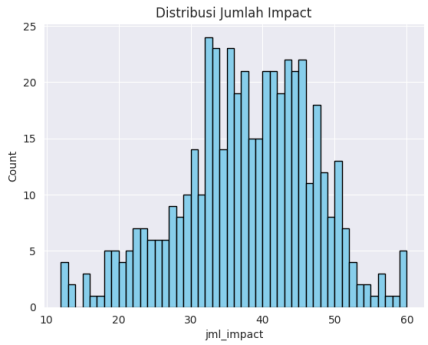
Gambar 13 menunjukkan hubungan antara usia dan jumlah dampak negatif dari media sosial. Garis regresi yang menurun menunjukkan bahwa semakin tua seseorang, semakin sedikit dampak negatif yang dialami. Mayoritas sampel berusia 18-30 tahun, dengan dampak negatif lebih tinggi dan beberapa individu mengalami dampak signifikan. Di usia muda, variasi dampak lebih besar, sedangkan di usia tua lebih stabil dan rendah. Hal ini mungkin disebabkan oleh keterlibatan yang lebih rendah atau kemampuan lebih baik dalam mengelola dampak media sosial. Remaja dan dewasa muda lebih rentan terhadap dampak negatif media sosial karena masih berada pada fase pencarian identitas dan sangat sensitif terhadap penerimaan sosial. Tingginya intensitas penggunaan membuat mereka lebih mudah terpapar perbandingan sosial, kecemasan, serta gangguan tidur. Selain itu, kemampuan regulasi emosi yang belum matang memperkuat kerentanan ini, berbeda dengan kelompok usia lebih tua yang cenderung memiliki kontrol diri lebih baik dan tingkat keterlibatan lebih rendah.



Gambar 14. Hubungan Antara Rata-Rata Waktu dengan Jumlah

Dampak Negatif Penggunaan Media Sosial Terhadap Kesehatan Mental

Gambar 14 menunjukkan bahwa semakin lama waktu penggunaan media sosial, semakin tinggi jumlah dampak negatif yang dialami. Tren ini berlaku untuk semua gender tanpa perbedaan signifikan. Pengguna yang menghabiskan lebih dari 5 jam per hari mengalami dampak negatif tertinggi, mengindikasikan bahwa durasi penggunaan berperan dalam meningkatkan resiko efek negatif.

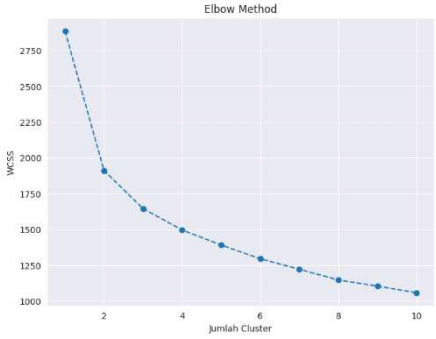


Gambar 15. Distribusi Jumlah Dampak Negatif Penggunaan Media Sosial Terhadap Kesehatan Mental

Gambar 15 menunjukkan distribusi jumlah dampak negatif dari penggunaan media sosial. Distribusi cenderung normal dengan sedikit kemiringan ke kanan. Mayoritas individu mengalami dampak negatif dalam kisaran 30-45, sementara jumlah individu dengan dampak sangat rendah (<20) atau sangat tinggi (>50) lebih sedikit. Pola ini menunjukkan bahwa dampak negatif media sosial bervariasi, tetapi sebagian besar individu berada di tengah rentang distribusi.

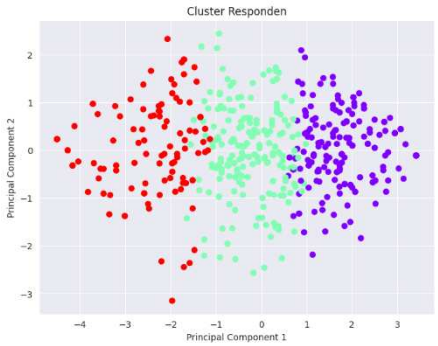
3.2.2. Pelabelan *Dataset*

Setelah menerapkan *distribution analysis* dan *data visualization*, tahap berikutnya adalah melakukan pelabelan pada *dataset*. Gambar 16 menunjukkan hasil penentuan jumlah *cluster* optimal menggunakan metode *elbow.*



Gambar 16. Hasil Penerapan Metode *Elbow*

Gambar 16 menunjukkan hasil metode *elbow*. Metode *elbow* digunakan untuk menentukan jumlah *cluster* optimal dalam algoritma *k-means*, dengan cara melihat perubahan nilai *Within Cluster Sum of Squares* (WCSS) terhadap jumlah *cluster* (*k*). WCSS sendiri mengukur seberapa rapat data dalam satu cluster, yaitu dengan menjumlahkan kuadrat jarak setiap titik ke pusat cluster-nya. Semakin kecil nilai WCSS, semakin homogen atau kompak sebuah *cluster*. Grafik menunjukkan penurunan tajam WCSS dari *k*=1 hingga *k*=2, lalu menurun lebih landai setelahnya. Penentuan jumlah *cluster* optimal berdasarkan *elbow point,* yakni titik siku (sudut) di mana WCSS tidak lagi berkurang secara signifikan. Dari grafik ini, titik *elbow* berada di *k*=3 karena setelah *k*=4 penurunan WCSS melambat, hal ini menunjukkan bahwa menambah *cluster* lebih banyak tidak memberikan keuntungan signifikan. Hasil pengelompokkan menggunakan algoritma *k*-means dan metode *elbow* digambarkan pada Gambar 17.



Gambar 17. Hasil Pengelompokkan

Gambar 17 menunjukkan hasil pengelompokkan responden menggunakan metode *k*-means. Data responden dalam *dataset* dibagi ke dalam 3 kelompok utama, sesuai dengan hasil analisis metode *elbow* yang sebelumnya menunjukkan bahwa *k*=3 adalah jumlah *cluster* yang optimal. Berdasarkan hasil terbesebut peneliti akan mengelompokkan *dataset* ke dalam 3 *cluster,* yakni *cluster* dampak negatif rendah, sedang, dan tinggi. Hasil pelabelan terdiri dari *cluster* 0 (dampak negatif rendah) dengan 172 data, *cluster* 1 (dampak negatif sedang) dengan 106 data, dan *cluster* 2 (dampak negatif tinggi) dengan 203 data.

3.3. *Modeling*

Hasil eksperimen EDA berupa *dataset* baru selanjutnya siap dimodelkan pada tahap *modeling.* Pada tahap ini pemodelan terdiri dari pemilihan algoritma *data mining* serta menerapkan parameter tertentu sesuai dengan masalah yang telah ditentukan pada tahap *business understanding* yakni menemukan pola prediksi resiko penggunaan media sosial terhadap kesehatan mental.

Eksperimen penerapan algoritma C4.5 dilakukan dengan dua parameter, yaitu dengan

pruning dan tanpa *pruning*. Pada algoritma *k*-NN, digunakan parameter dengan nilai *k*=3 dan *k*=5. Sementara itu, algoritma *Naïve Bayes* hanya menggunakan parameter standar yang dimiliki oleh algoritma tersebut. Pada setiap eksperimen penerapan algoritma tersebut pembagian data dilakukan menggunakan metode *10-fold cross-validation*.

3.4. *Evaluation*

Pada tahap *evaluation* dilakukan evaluasi terhadap hasil penilaian perbandingan kinerja algoritma C4.5, *k*-NN, dan *Naïve Bayes* yang telah dilakukan pada tahap *modelling.* Ukuran evaluasi yang digunakan adalah nilai akurasi, *precision, recall,* dan *F1-Score.* Hasil perbandingan kinerja algoritma tersebut dirangkum pada Tabel 3.

Tabel 3. Perbandingan Hasil Kinerja Algoritma

Algoritma	Akurasi	*Precision*	*Recall*	*F1-Score*
C4.5 *(pruning)*	0.8419	0.8486	0.8419	0.8406
C4.5 *(no pruning)*	0.9003	0.9054	0.9003	0.8998
k-NN (*k=3*)	0.9043	0.9087	0.9043	0.9045
k-NN (*k=5*)	0.9190	0.9250	0.9190	0.9192
Naive Bayes	0.9252	0.9308	0.9252	0.9247

Tabel x merupakan hasil evaluasi kinerja algoritma berdasarkan perbandingan metrik evaluasi.

1. Akurasi: *Naïve Bayes* memiliki nilai akurasi tertinggi (92,52%), menunjukkan bahwa algoritma ini paling andal dalam mengklasifikasikan data dengan benar secara keseluruhan. *k*-NN dengan *k*=5 berada di posisi kedua (91,90%), diikuti oleh *k*-NN dengan *k*=3 (90,43%) dan C4.5 tanpa pruning (90,03%). C4.5 dengan pruning memiliki akurasi terendah (84,19%), menunjukkan bahwa proses pruning mengurangi keakuratan model.
2. *Precision*: *Naïve Bayes* memiliki *precision* tertinggi (93,08%), menandakan bahwa ketika algoritma ini mengklasifikasikan suatu data ke dalam kategori positif, kemungkinan besar hasilnya benar. *k*-NN dengan *k*=5 berada di posisi kedua (92,50%), sedikit lebih baik dibandingkan *k*-NN dengan *k*=3 (90,87%). C4.5 tanpa *pruning* mencapai *precision* 90,54%, sementara C4.5 dengan *pruning* memiliki *precision* paling rendah (84,86%), menandakan tingkat kesalahan yang lebih tinggi dalam klasifikasi positif.
3. *Recall: Naïve Bayes* memiliki *recall* tertinggi (92,52%), menunjukkan bahwa algoritma ini paling baik dalam mendeteksi semua kasus positif yang sebenarnya. *k*-NN dengan *k*=5 memiliki *recall* sedikit lebih rendah (91,90%), diikuti oleh *k*-NN dengan *k*=3 (90,43%) dan C4.5 tanpa *pruning* (90,03%). C4.5 dengan *pruning* memiliki *recall* terendah (84,19%), yang berarti lebih banyak kasus positif yang terlewat dibandingkan metode lain.
4. *F1-score: Naïve Bayes* juga memiliki *F1-score* tertinggi (92,47%), menandakan keseimbangan optimal antara *precision* dan *recall*. *k*-NN dengan *k*=5 memiliki *F1-score* 91,92%, lebih tinggi dari kk-NN dengan *k*=3 (90,45%) dan C4.5 tanpa *pruning* (89,98%). C4.5 dengan *pruning* memiliki *F1-score* terendah (84,06%), menegaskan bahwa metode ini kurang optimal dalam menjaga keseimbangan antara *precision* dan *recall*.

Naïve Bayes menunjukkan kinerja terbaik pada seluruh metrik evaluasi, menjadikannya algoritma paling optimal untuk *dataset* ini. Keunggulannya terletak pada kemampuan mengolah data dengan fitur yang relatif independen, sehingga perhitungan probabilitas menjadi sederhana, stabil, serta efisien secara komputasi. Selain itu, *Naïve Bayes* lebih tahan terhadap variasi maupun *noise* dibandingkan algoritma lain. Temuan ini sejalan dengan hasil penelitian [27] yang membandingkan algoritma *Naïve Bayes* dan SVM dalam mendeteksi tingkat stres mahasiswa melalui *chatbot* berbasis AI. Hasilnya menunjukkan bahwa *Naïve Bayes* lebih unggul dengan akurasi 90%, sedikit lebih tinggi dibanding SVM yang mencapai 89%. Keunggulan ini disebabkan oleh kemampuan *Naïve Bayes* mengolah fitur-fitur yang relatif independen tanpa membutuhkan parameter *tuning* yang kompleks seperti SVM, sehingga model menjadi lebih stabil, efisien, dan praktis untuk diterapkan dalam sistem *real-time* seperti *chatbot* deteksi stres.

Sebagai pembanding, *k*-NN dengan k=5 menempati posisi kedua dengan performa yang cukup dekat, namun sensitivitas terhadap pemilihan parameter *k* membuat hasilnya kurang stabil. Menurut penelitian [22] kinerja *k*-NN sangat bergantung pada pemilihan nilai *k*. Jika *k* terlalu kecil, model menjadi sensitif terhadap *outlier* atau *noise*, sedangkan jika *k* terlalu besar, model cenderung lebih tahan terhadap keberadaan *outlier*.

Sementara itu, C4.5 menunjukkan performa lebih rendah, terutama setelah proses *pruning* yang justru menurunkan akurasi model. Penelitian [28] menyebutkan bahwa *pruning* pada algoritma C4.5 dilakukan untuk mengurangi risiko *overfitting*, yaitu kondisi ketika model terlalu menyesuaikan diri dengan data latih sehingga kinerjanya menurun pada data baru. Dalam tahap pembentukan pohon keputusan, C4.5 menganggap bahwa data latih bersifat reliabel atau cukup mewakili kondisi sebenarnya, sehingga pola yang muncul dari data tersebut dipercaya sebagai dasar pemilihan variabel. Namun, saat proses *pruning* dilakukan, asumsi ini berubah karena sebagian cabang pohon justru dipangkas dengan anggapan bahwa data mengandung ketidakpastian atau *noise*. Perbedaan perlakuan ini dapat menimbulkan masalah, terutama ketika dataset memang mengandung banyak *noise*, sehingga

pruning yang berlebihan justru menurunkan akurasi model alih-alih memperbaikinya.

3.5. *Deployment*

Model yang telah diuji dan divalidasi dapat diterapkan dalam sistem cerdas atau aplikasi prediksi resiko penggunaan media sosial terhadap kesehatan mental sebagai penelitian lanjutan. Model prediksi kesehatan mental dapat diterapkan dalam aplikasi kesehatan untuk memberikan rekomendasi personalisasi kepada pengguna. Selain itu hasil penerapan EDA berupa laporan analisis distribusi dan visualisasi data dapat digunakan sebagai wawasan terkait hubungan antara penggunaan media sosial dengan dampak negatif terhadap kesehatan mental.

4. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, didapatkan beberapa kesimpulan pada saat melakukan prediksi resiko penggunaan media sosial terhadap kesehatan mental menggunakan *Exploratory Data Analysis* (EDA) dan *Cross Industry Standard Process for Data Mining* (CRISP-DM). Integrasi EDA dan CRISP-DM dapat memberikan pendekatan yang tepat dalam menghasilkan model prediksi meskipun proses EDA memerlukan waktu tambahan yang tidak sedikit untuk melakukan analisis data secara detail dan menyeluruh. Keunggulan penerapan EDA adalah memastikan semua aspek penting dalam data diperhatikan tanpa terlewatkan. Hasil penerapan EDA terkait resiko penggunaan media sosial terhadap kesehatan mental menunjukkan semakin lama penggunaan media sosial, semakin besar dampak negatif yang dialami, terutama bagi mereka yang menggunakannya lebih dari 5 jam per hari. Mayoritas individu mengalami dampak negatif dalam tingkat sedang, sementara kasus dengan dampak sangat rendah atau sangat tinggi lebih jarang. Usia 13-30 tahun lebih rentan terhadap gangguan perhatian, kegelisahan, perbandingan sosial, dan kebutuhan validasi, sedangkan usia 40+ lebih sering mengalami depresi dan gangguan tidur. Beberapa dampak seperti fluktuasi aktivitas harian dan gangguan konsentrasi muncul di berbagai usia tanpa pola yang konsisten. Secara umum dapat disimpulkan bahwa durasi penggunaan media sosial menjadi faktor yang paling dominan dalam resiko tinggi. Temuan ini dapat digunakan sebagai wawasan pencegahan terganggunya kesehatan mental karena penggunaan media sosial.

Selain keunggulan EDA di atas, tahap EDA sangat krusial dalam menentukan keberhasilan tahap *modeling. Modeling* berfokus pada pembangunan model prediksi yang membutuhkan persiapan dan analisis data yang matang. Dalam penelitian ini EDA mampu menghasilkan *dataset* baru dengan label yang akan dimasukkan ke dalam pemodelan. *Dataset* berlabel merupakan ciri *dataset* yang dapat diolah menggunakan algoritma *supervised learning* seperti C4.5, *k*-NN, dan *Naïve Bayes.* Ketiga algoritma tersebut diterapkan untuk memprediksi *dataset* dampak negatif penggunaan media sosial terhadap kesehatan mental. Hasil menunjukkan bahwa *Naïve Bayes* memiliki kinerja yang lebih unggul dibandingkan dengan algoritma C4.5 dan *k*-NN. Pada penelitian selanjutnya dapat dilakukan uji statistik seperti uji *friedman* dan *nemenyi* untuk mengetahui lebih lanjut tingkat signifikansi perbedaan dari ketiga algoritma tersebut.

Dalam penelitian ini CRISP-DM mendukung penerapan EDA dengan membuat kerangka kerja yang jelas, mulai dari *business understanding* hingga *deployment*. Dengan tahapan yang terstruktur, membantu peneliti untuk bekerja secara sistematis dan mengurangi risiko kesalahan. CRISP-DM juga bersifat iteratif, sehingga memungkinkan untuk terus melakukan perbaikan pada model dan strategi berdasarkan *dataset* terbaru.

DAFTAR PUSTAKA

P. ELISA AND A. RAHMAN ISNAIN. 2024. Comparison of Random Forest, Support Vector Machine and Naive Bayes Algorithms To Analyze Sentiment Towards Mental Health Stigma. *J. Tek. Inform.*, vol. 5, no. 1, pp. 321–329, 2024, doi: https://doi.org/10.52436/1.jutif.2024.5.1.1817.

R. CHRISTINA, M. S. YUNIARDI, AND A. PRABOWO. 2019. . Hubungan Tingkat Neurotisme dengan Fear of Missing Out (FoMO) pada Remaja Pengguna Aktif Media Sosial. *Indig. J. Ilm. Psikol.*, vol. 4, no. 2, pp. 105–117, 2019, doi: 10.23917/indigenous.v4i2.8024.

B. OSATUY. 2015. Is lurking an anxiety-masking strategy on social media sites? The effects of lurking and computer anxiety on explaining information privacy concern on social media platforms. *Comput. Human Behav.*, vol. 49, no. 2015, pp. 324–332, 2015, doi: 10.1016/j.chb.2015.02.062.

M. HAUG, J. REITER, AND H. GEWALD. 2024 Content creators on Instagram—How users cope with stress on social media. *Telemat. Informatics Reports*, vol. 13, no. December 2023, p. 100111, 2024, doi: 10.1016/j.teler.2023.100111.

C. ZHANG, T. CAO, & A. ALI. 2022. Investigating the Role of Perceived Information Overload on COVID-19 Fear: A Moderation Role of Fake News Related to COVID-19. *Front. Psychol.*, vol. 13, no. June, pp. 1–14, 2022, doi: 10.3389/fpsyg.2022.930088.

R. Y. M. WONG, C. M. K. CHEUNG, B. XIAO, AND J. B. THATCHER. 2021. Standing up or standing by: Understanding bystanders' proactive reporting responses to social media

harassment," *Inf. Syst. Res.*, vol. 32, no. 2, pp. 561–581, 2021, doi: 10.1287/ISRE.2020.0983.

D. M. J. LAZER *et al.* 2018. The science of fake news. *Science*, vol. 359, no. 6380, pp. 1094–1096, Mar. 09, 2018. doi: 10.1126/science.aao2998.

Y. SUN AND Y. ZHANG. 2021. A review of theories and models applied in studies of social media addiction and implications for future research. *Addict. Behav.*, vol. 114, p. 106699, Mar. 2021, doi: 10.1016/j.addbeh.2020.106699.

C. GONÇALVES, D. FERREIRA, C. NETO, A. ABELHA, AND J. MACHADO. 2020.. Prediction of mental illness associated with unemployment using data mining," *Procedia Comput. Sci.*, vol. 177, pp. 556–561, 2020, doi: 10.1016/j.procs.2020.10.078.

J. DEL CAMPO-ÁVILA *et al.* 2024. Data mining process to detect suicidal behaviour in out-of-hospital emergency departments. *Eng. Appl. Artif. Intell.*, vol. 136, no. PA, p. 108910, 2024, doi: 10.1016/j.engappai.2024.108910.

N. ANTHIRA AND SUENDRI. 2024. Penerapan Data Mining Pada Klasifikasi Gangguan Jiwa Menggunakan Algoritma C5.0 Di RSJ. Mahoni Kota Medan. *Teknika*, vol. 18, pp. 571–582, 2024, doi: https://doi.org/10.5281/zenodo.12784435.

C. SCHRÖER, F. KRUSE, AND J. M. GÓMEZ, "A systematic literature review on applying CRISP-DM process model. *Procedia Comput. Sci.*, vol. 181, no. 2019, pp. 526–534, 2021, doi: 10.1016/j.procs.2021.01.199.

O. MARBÁN, J. SEGOVIA, E. MENASALVAS, AND C. Fernández-Baizán. 2009. Toward data mining engineering: A software engineering approach. *Inf. Syst.*, vol. 34, no. 1, pp. 87–107, 2009, doi: 10.1016/j.is.2008.04.003.

F. MARTINEZ-PLUMED *et al.* 2019. CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2021, doi: 10.1109/TKDE.2019.2962680.

P. CHAPMAN *et al.* 2000. *CRISP-DM 1.0 Step-by-step data mining guide*, vol. 78. 2000. [Online]. Available: http://www.crisp-dm.org/CRISPWP-0800.pdf

OMARI FIRAS. 2023. A combination of SEMMA & CRISP-DM models for effectively handling big data using formal concept analysis based knowledge discovery: A data mining approach. *World J. Adv. Eng. Technol. Sci.*, vol. 8, no. 1, pp. 009–014, 2023, doi: 10.30574/wjaets.2023.8.1.0147.

R. INDRAKUMARI, T. POONGODI, AND S. R. JENA. 2020. Heart Disease Prediction using Exploratory Data Analysis," *Procedia Comput. Sci.*, vol. 173, no. 2019, pp. 130–139, 2020, doi: 10.1016/j.procs.2020.06.017.

F. LIU AND Y. DENG, "Determine the Number of Unknown Targets in Open World Based on Elbow Method," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 5, pp. 986–995, May 2021, doi: 10.1109/TFUZZ.2020.2966182.

K. KUSUMANINGTYAS, M. HABIBI, I. DWIJAYANTI, AND R. SUMIYARINI, "Tweet Analysis of Mental Illness Using K-Means Clustering and Support Vector Machine," *Telematika*, vol. 20, no. 3, p. 295, Nov. 2023, doi: 10.31315/telematika.v20i3.9820.

F. A. TYAS, U. GHONI, AND S. ISMAYA, "Penentuan Rule Base pada sistem pakar identifikasi jenis kulit wajah menggunakan algoritma c4.5," in *Conference on Electrical Engineering, Informatics, Industrial Technology, and Creative Media 2023*, 2023, pp. 653–662.

D. T. LAROSE AND C. D. LAROSE, *Discovering Knowledge in Data: An Introduction to Data Mining*, Second Edi. Wiley, 2014. doi: 10.1002/9781118874059.

FITRI AYUNING TYAS, MAHDA NURAYUNI, AND HIDAYATUR RAKHMAWATI,. 2024. Optimasi Algoritma K-Nearest Neighbors Berdasarkan Perbandingan Analisis Outlier (Berbasis Jarak, Kepadatan, LOF). *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 13, no. 2, pp. 108–115, 2024, doi: 10.22146/jnteti.v13i2.9579.

Y. FINDAWATI, I. R. I. ASTUTIK, A. S. FITRONI, I. INDRAWATI, AND N. YUNIASIH. 2019. Comparative analysis of Naïve Bayes, K Nearest Neighbor and C.45 method in weather forecast. *J. Phys. Conf. Ser.*, vol. 1402, no. 6, p. 066046, Dec. 2019, doi: 10.1088/1742-6596/1402/6/066046.

Z. TIAN AND D. YI. 2024. Application of artificial intelligence based on sensor networks in student mental health support system and crisis prediction. *Meas. Sensors*, vol. 32, no. November 2023, p. 101056, 2024, doi: 10.1016/j.measen.2024.101056.

M. D. NATH, M. K. U. AHAMED, O. AHMED, T. AHMED, S. ROY, AND M. N. UDDIN. 2025. Smart web interface for student mental health prediction using machine learning with blockchain technology. *Neurosci. Informatics*, vol. 5, no. 4, p. 100236, Dec. 2025, doi: 10.1016/j.neuri.2025.100236.

M. NOUMAN, S. Y. KHOO, M. A. P. MAHMUD, AND A. Z. KOUZANI. 2025. A hybrid BERT-BiRNN framework for mental health prediction using textual data. *Nat. Lang. Process. J.*, vol. 12, no. January, p. 100165, 2025, doi: 10.1016/j.nlp.2025.100165.

E. D. S. MARIYANA, M. NOVITA, AND NUR LATIFAH DWI MUTIARA SARI. 2025. Mental Health Chatbot Application on Artificial Intelligence (AI) for Student Stress Detection Using Mobile-Based Naïve Bayes Algorithm. *Sci. J. Informatics*, vol. 12, no. 2, pp. 199–210, 2025, doi: 10.15294/sji.v12i2.24307.

S. MORAL-GARCÍA, C. J. MANTAS, J. G. CASTELLANO, AND J. ABELLÁN. 2019. Non-parametric predictive inference for solving multi-label classification. *Appl. Soft Comput. J.*, vol. 88, 2019, doi: 10.1016/j.asoc.2019.106011.