Prediksi Kepribadian Introvert dan Ekstrovert Berdasarkan Pola Perilaku Sosial Menggunakan Model Machine Learning (KNN dan Naive Bayes)

Muhammad David Rahadian

Program Studi : Sistem Informasi, Universitas Darwan Ali

Email : davidrahadian2023@gmail.com

ABSTRACT— This study aims to predict personality types, specifically introversion and extraversion, based on individuals' social behavior patterns using machine learning methods. A dataset of 2,900 samples was utilized, containing attributes such as time spent alone, stage fear, social event attendance, outdoor activity, post-socializing fatigue, friend circle size, and post frequency on social media. The research was structured in several integrated stages, starting from a literature review, data collection, preprocessing, modeling, and performance evaluation. Two classification models were built using K-Nearest Neighbors (KNN) and Naive Bayes algorithms, and evaluated using accuracy, precision, and recall metrics. The KNN model achieved an accuracy of 93.68%, a precision of 94.23%, and a recall of 92.67%, while the Naive Bayes model attained an accuracy of 93.33%, precision of 91.20%, and recall of 95.51%. The results indicate that both models effectively classify personality based on behavioral data, with KNN excelling in overall accuracy and precision, and Naive Bayes in recall. This research demonstrates that observable social behavior can be a reliable indicator of personality traits and highlights the potential of machine learning in psychological classification, system personalization, and intelligent human-computer interaction.

Keywords— personality prediction, introvert, extrovert, social behavior, KNN, Naive Bayes.

ABSTRAK— Penelitian ini bertujuan untuk memprediksi tipe kepribadian, khususnya introvert dan ekstrovert, berdasarkan pola perilaku sosial individu menggunakan metode machine learning. Dataset sebanyak 2.900 sampel digunakan, yang memuat atribut seperti waktu menyendiri, ketakutan tampil di panggung, kehadiran dalam acara sosial, aktivitas di luar rumah, kelelahan setelah interaksi sosial, ukuran lingkaran pertemanan, dan frekuensi unggahan media sosial. Penelitian ini disusun dalam beberapa tahapan yang saling terintegrasi, dimulai dari tinjauan literatur, pengumpulan data, preprocessing, pemodelan, hingga evaluasi performa. Dua model klasifikasi dibangun dengan algoritma K-Nearest Neighbors (KNN) dan Naive Bayes, serta dievaluasi menggunakan metrik akurasi, presisi, dan recall. Model KNN memperoleh akurasi 93,68%, presisi 94,23%, dan recall 92,67%, sementara model Naive Bayes mencatat akurasi 93,33%, presisi 91,20%, dan recall 95,51%. Hasil penelitian menunjukkan bahwa kedua model mampu mengklasifikasikan kepribadian dengan efektif berdasarkan data perilaku, dengan KNN unggul dalam akurasi dan presisi keseluruhan, sedangkan Naive Bayes lebih unggul dalam recall. Penelitian ini membuktikan bahwa perilaku sosial terukur dapat menjadi indikator yang andal untuk memetakan kepribadian, serta menunjukkan potensi penerapan machine learning dalam klasifikasi psikologis, personalisasi sistem, dan interaksi adaptif berbasis pengguna.

Kata kunci— prediksi kepribadian, introvert, ekstrovert, perilaku sosial, KNN, Naive Bayes.

I. PENDAHULUAN

Kepribadian merupakan salah satu aspek fundamental dalam memahami perilaku manusia secara menyeluruh. Dalam konteks psikologi modern, kepribadian tidak hanya dipahami sebagai kumpulan sifat atau karakteristik internal, tetapi juga sebagai pola yang relatif stabil dari cara berpikir, merasakan, dan berperilaku seseorang dalam berbagai situasi kehidupan [6], [11]. Kajian terhadap kepribadian memiliki peran sentral dalam berbagai disiplin ilmu karena dapat memberikan wawasan yang mendalam mengenai bagaimana individu merespons lingkungan, membuat keputusan, menjalin relasi sosial, hingga membentuk preferensi dalam aktivitas sehari-hari [12], [13].

Salah satu kerangka teoritis yang paling berpengaruh dalam memahami kepribadian adalah model Big Five Personality Traits, yang secara luas diakui sebagai pendekatan komprehensif dalam taksonomi kepribadian [4], [11]. Model ini mengidentifikasi lima dimensi dasar kepribadian, yaitu openness to experience, conscientiousness, extraversion, agreeableness, dan neuroticism. Dari kelima dimensi tersebut, ekstraversi (extraversion) merupakan salah satu aspek yang paling banyak dikaji karena berkaitan erat dengan interaksi sosial, ekspresi emosi, dan orientasi terhadap lingkungan sekitar. Individu yang cenderung ekstrovert biasanya menunjukkan sikap terbuka, aktif, dan mudah beradaptasi dalam situasi sosial. Sebaliknya, individu introvert cenderung lebih introspektif, menikmati kesendirian, dan memerlukan waktu untuk memulihkan energi setelah berinteraksi dalam kelompok besar [6].

Pemahaman terhadap dimensi kepribadian ini sangat penting, tidak hanya dalam bidang psikologi klinis, tetapi juga dalam konteks pendidikan, hubungan kerja, strategi komunikasi, hingga interaksi dalam ruang digital [3], [5]. Dalam dunia modern yang didominasi oleh interaksi daring dan pemanfaatan teknologi informasi, aspek kepribadian semakin berperan dalam menentukan bagaimana individu menyerap informasi, memilih konten, dan merespons stimulus dari lingkungan sekitarnya [14].

Oleh karena itu, pemetaan kepribadian menjadi krusial dalam merancang sistem yang mampu memberikan pengalaman yang lebih personal dan adaptif bagi pengguna [6], [14].

Di era digital saat ini, perilaku manusia dalam kehidupan sehari-hari tidak lagi terbatas pada dunia fisik. Interaksi digital melalui media sosial, forum online, platform e-learning, dan aplikasi berbasis aktivitas sosial telah menciptakan jejak perilaku yang sangat kaya. Aktivitas seperti tingkat partisipasi dalam kegiatan sosial, kenyamanan saat menyendiri, hingga intensitas membagikan informasi di media digital dapat ditangkap dalam bentuk data dan dianalisis untuk mendapatkan pemahaman tentang kecenderungan kepribadian seseorang [3], [5], [14]. Data perilaku ini, jika diproses dengan pendekatan yang tepat, memiliki potensi besar untuk menggambarkan karakter individu secara akurat [7], [11].

Salah satu pendekatan yang paling menjanjikan dalam pengolahan data perilaku sosial untuk prediksi kepribadian adalah melalui penerapan machine learning [3], [5], [14]. Machine learning, sebagai bagian dari kecerdasan buatan, memungkinkan mesin belajar dari data tanpa perlu diprogram secara eksplisit. Dengan algoritma yang tepat, model machine learning mampu mengenali pola-pola kompleks dalam data dan menghasilkan prediksi yang akurat [10], [15]. Dalam konteks prediksi kepribadian, algoritma klasifikasi menjadi sangat relevan karena memungkinkan identifikasi tipe kepribadian berdasarkan atribut-atribut perilaku tertentu [11], [13]. Beberapa penelitian terdahulu bahkan telah menunjukkan bahwa metode ini dapat mencapai tingkat akurasi yang tinggi dalam konteks prediksi non-klinis [1], [2], [4].

Penelitian ini memanfaatkan dataset yang bersumber dari platform Kaggle, yang berisi sebanyak 2.900 sampel perilaku sosial individu. Setiap sampel terdiri atas atribut-atribut penting yang merepresentasikan kecenderungan perilaku sosial, seperti Time spent Alone, Stage fear, social event attendance, Going outside, Drained after socializing, Friend circle size, dan Post frequency. Label target dalam dataset ini adalah kategori kepribadian: Introvert atau Ekstrovert. Pemilihan atribut-atribut ini tidak dilakukan secara acak, melainkan berdasarkan indikator yang telah divalidasi dalam penelitian psikologi sebelumnya mengenai keterkaitannya dengan kecenderungan interaksi sosial dan preferensi personal dalam menjalani kehidupan sehari-hari [4], [11], [13].

Permasalahan utama yang diangkat dalam penelitian ini adalah bagaimana membangun model klasifikasi yang mampu memprediksi kepribadian seseorang secara akurat berdasarkan pola perilaku sosial. Untuk menjawab permasalahan tersebut, penelitian ini mengimplementasikan dua algoritma klasifikasi yang telah banyak digunakan dalam literatur, yaitu K-Nearest Neighbors (KNN) dan Naive Bayes [1], [2], [7], [10]. Kedua algoritma ini dipilih berdasarkan efektivitasnya dalam pengolahan data numerik serta efisiensi dalam melakukan klasifikasi pada berbagai kasus prediksi perilaku [8], [9].

Penelitian-penelitian terdahulu memberikan bukti empiris mengenai keandalan algoritma-algoritma ini. Kartarina et al. [1], misalnya, dalam kajiannya mengenai prediksi kelulusan mahasiswa, menemukan bahwa KNN mampu menghasilkan akurasi hingga 96,18%, sedangkan Naive Bayes mencapai 91,94%. Studi lain oleh Azahari et al. [2] juga menunjukkan performa yang baik dari kedua metode dalam memprediksi masa studi mahasiswa. Temuan ini menjadi dasar rasional dalam pemilihan algoritma karena menunjukkan bahwa KNN unggul dalam konteks data numerik berdimensi rendah, sementara Naive Bayes lebih kompeten dalam hal efisiensi dan penanganan distribusi probabilistik yang sederhana [7], [9].

Penelitian ini tidak hanya membandingkan performa kedua algoritma dalam konteks klasifikasi kepribadian, tetapi juga bertujuan untuk mengidentifikasi atribut perilaku sosial mana yang paling berpengaruh terhadap pembentukan model prediktif. Dengan demikian, hasil yang diperoleh diharapkan dapat memberikan kontribusi tidak hanya pada aspek teoritis, tetapi juga pada penerapan praktis dalam berbagai domain, seperti sistem rekomendasi personalisasi, psikometri digital, dan layanan berbasis kecerdasan buatan yang mengadaptasi diri terhadap pengguna berdasarkan kepribadiannya [3], [4], [6], [14].

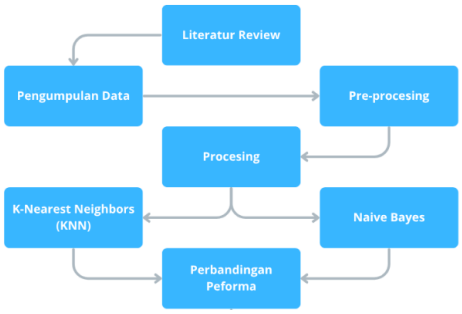
Pemilihan topik ini dilandasi oleh urgensi dan relevansi dari kebutuhan akan layanan digital yang lebih personal, adaptif, dan berbasis karakteristik unik setiap individu. Dalam lingkungan digital yang semakin dinamis, pemahaman terhadap pengguna tidak lagi cukup hanya dengan data demografis atau perilaku transaksi, tetapi juga harus mencakup aspek psikologis seperti kepribadian [6], [12], [13]. Oleh karena itu, integrasi antara ilmu komputer dan ilmu perilaku manusia menjadi kunci dalam membangun sistem yang benar-benar cerdas dan human-centered [11], [14]. Penelitian ini diharapkan menjadi salah satu kontribusi awal dalam upaya integratif tersebut, dengan pendekatan yang sistematis, valid, dan dapat direplikasi di masa depan.

II. METODOLOGI PENELITIAN

Penelitian ini menerapkan pendekatan eksperimen guna mengembangkan dan membandingkan model klasifikasi kepribadian berdasarkan pola perilaku sosial mahasiswa. Metode kuantitatif dipilih karena memungkinkan analisis numerik terhadap variabel perilaku, serta evaluasi performa model klasifikasi secara objektif menggunakan metrik seperti akurasi, presisi, dan recall. Sementara itu, pendekatan eksperimental digunakan untuk menguji dua algoritma yang berbeda, yakni K-Nearest Neighbors (KNN) dan Naive Bayes, dalam konteks yang sama dan dengan data yang identik, sehingga memungkinkan perbandingan performa yang adil dan terukur. Pilihan algoritma ini didasarkan pada penelitian sebelumnya yang menunjukkan efektivitas keduanya dalam konteks klasifikasi sosial dan psikologis [1], [2].

Untuk menjamin validitas dan keterulangan (replicability) hasil, penelitian ini disusun dalam beberapa

tahapan penelitian yang saling terintegrasi, dimulai dari tinjauan literatur, pengumpulan data, preprocessing, pemodelan, hingga evaluasi performa. Penjelasan sistematis mengenai tahapan tersebut dijabarkan sebagai berikut.



Gambar 1 Tahapan Penelitian

A. Literatur Review

Tahapan awal dalam penelitian ini diawali dengan melakukan kajian pustaka untuk memahami berbagai pendekatan klasifikasi kepribadian, metode supervised learning, serta penerapan algoritma KNN dan Naive Bayes dalam studi perilaku sosial. Kajian ini merujuk pada berbagai jurnal ilmiah terkini yang mengkaji hubungan antara perilaku sosial digital dan tipe kepribadian [1], [3]. Studi tersebut memberikan pemahaman mengenai efektivitas masing-masing algoritma, teknik pengolahan data yang digunakan, serta jenis fitur yang relevan dalam mendeteksi kepribadian. Selain itu, literatur ini juga menjadi dasar dalam menentukan variabel yang digunakan dalam penelitian, serta validasi terhadap metode yang diterapkan.

B. Pengumpulan Data

Data yang digunakan dalam penelitian ini diperoleh dari dataset terbuka yang telah digunakan dalam studi-studi terdahulu terkait perilaku sosial dan kepribadian mahasiswa. Dataset ini terdiri dari fitur kuantitatif seperti jumlah lingkaran pertemanan (Friends_circle), frekuensi kehadiran pada acara sosial (Social_event), dan intensitas memposting di media sosial (Post_frequency). Sementara itu, label target dalam klasifikasi adalah tipe kepribadian yang dikategorikan sebagai Introvert atau Ekstrovert, yang telah ditentukan berdasarkan indikator dalam data awal. Model klasifikasi kemudian mempelajari pola hubungan antara variabel perilaku tersebut dengan kategori kepribadian. Struktur data ini sejalan dengan pendekatan yang digunakan dalam penelitian [4], yang juga mengeksplorasi korelasi antara perilaku sosial dan karakteristik psikologis.

C. Preprocessing Data

Data yang diperoleh dari sumber terbuka sering kali mengandung masalah kualitas seperti nilai kosong (missing values) dan anomali data, yang dapat memengaruhi performa algoritma klasifikasi jika tidak ditangani secara tepat. Oleh karena itu, dilakukan tahapan preprocessing untuk memastikan integritas data sebelum digunakan dalam pelatihan model.

Prosedur preprocessing ini dirancang untuk meningkatkan kualitas data input dan mencegah kesalahan klasifikasi yang disebabkan oleh data yang tidak valid. Penelitian sebelumnya menunjukkan bahwa kualitas preprocessing sangat memengaruhi akurasi model klasifikasi dalam domain serupa [5].

D. Procesing Data

1. Algoritma K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) merupakan salah satu algoritma klasifikasi non-parametrik yang populer karena kesederhanaannya dan efektivitasnya dalam menangani data numerik. Prinsip kerja algoritma ini adalah dengan mengidentifikasi sejumlah tetangga terdekat (nearest neighbors) dari suatu data uji berdasarkan jarak ke data latih, kemudian menentukan kelas data uji berdasarkan mayoritas label dari tetangga tersebut. Dalam penelitian ini, digunakan nilai k sebesar 5, yang berarti klasifikasi dilakukan dengan mempertimbangkan lima data latih terdekat dalam ruang fitur.

Berbeda dari algoritma yang membutuhkan fase pelatihan eksplisit, KNN termasuk kategori lazy learner karena menyimpan seluruh data latih dan melakukan perhitungan hanya pada saat prediksi dilakukan [1], [8]. Hal ini membuat KNN sederhana untuk diimplementasikan namun cukup intensif secara komputasi saat fase prediksi, terutama jika jumlah data latih besar.

Pengukuran kedekatan antara data uji dan data latih dilakukan menggunakan rumus Euclidean Distance, yaitu:

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$$

Dalam rumus tersebut, $d(p,q)$ adalah jarak antara dua titik data p dan q dalam ruang berdimensi n, dengan p_i dan q_i masing-masing menyatakan nilai fitur ke-i dari kedua data tersebut.

Kelebihan KNN antara lain adalah fleksibilitas serta akurasi yang baik dalam menangani data berdimensi rendah hingga menengah. Namun, kelemahannya adalah sensitivitas terhadap skala atribut. Oleh karena itu, sebelum diterapkan, seluruh data dalam penelitian ini dinormalisasi terlebih dahulu guna memastikan setiap fitur memiliki kontribusi yang setara dalam proses pengukuran jarak.

2. Algoritma Naive Bayes

Naive Bayes adalah algoritma klasifikasi berbasis probabilistik yang mengacu pada Teorema Bayes, yang memungkinkan penghitungan probabilitas posterior suatu kelas berdasarkan bukti yang diamati. Model ini mengasumsikan bahwa setiap fitur dalam dataset saling independen satu sama lain, meskipun dalam praktiknya asumsi ini tidak selalu terpenuhi secara mutlak. Meski begitu, algoritma ini tetap memberikan performa yang kompetitif, terutama pada dataset berskala besar dengan kompleksitas rendah [2], [9].

Rumus umum dari Teorema Bayes yang digunakan adalah sebagai berikut:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}$$

Dalam rumus tersebut, $P(C|X)$ adalah probabilitas posterior bahwa data X termasuk dalam kelas C;$P(X|C)$ adalah probabilitas kemunculan data X jika diketahui kelasnya adalah C;$P(C)$ adalah probabilitas apriori dari kelas C; dan $P(X)$ adalah probabilitas umum dari data X.

Dalam konteks klasifikasi kepribadian pada penelitian ini, Naive Bayes menghitung probabilitas seseorang tergolong ke dalam kategori introvert atau ekstrovert berdasarkan kombinasi nilai atribut perilaku sosial yang diamati. Model kemudian memilih kelas dengan probabilitas posterior tertinggi sebagai hasil prediksi. Kelebihan utama dari algoritma ini adalah kecepatannya dalam proses pelatihan dan prediksi, serta kestabilannya meskipun fitur dalam data tidak berkorelasi secara langsung..

E. Perbandingan Performa

Untuk menilai performa dari masing-masing model yang digunakan dalam penelitian ini, diterapkan tiga metrik evaluasi utama yang umum digunakan dalam klasifikasi biner, yaitu:

1. Akurasi digunakan untuk mengukur proporsi keseluruhan prediksi yang benar dibandingkan dengan jumlah total prediksi yang dilakukan. Rumus yang digunakan untuk menghitung akurasi adalah.
$$\text{Akurasi} = \frac{TP + TN}{TP + TN + FP + FN}$$
di mana *TP (True Positive)* merupakan jumlah data introvert yang berhasil diklasifikasikan dengan benar, *TN (True Negative)* adalah jumlah data ekstrovert yang juga diklasifikasikan secara benar, *FP (False Positive)* adalah jumlah data ekstrovert yang salah diklasifikasikan sebagai introvert, dan *FN (False Negative)* adalah jumlah data introvert yang salah diklasifikasikan sebagai ekstrovert.
2. Presisi (precision) digunakan untuk mengukur seberapa besar proporsi dari seluruh prediksi positif yang benar-benar merupakan data positif. Dalam konteks ini, presisi dihitung dengan rumus:
$$\text{Presisi} = \frac{TP}{TP + FP}$$
Nilai presisi yang tinggi mengindikasikan bahwa model jarang memberikan prediksi positif yang keliru, sehingga sangat penting dalam situasi di mana kesalahan dalam mengklasifikasikan data negatif sebagai positif memiliki konsekuensi signifikan
3. Recall atau *sensitivity* berfungsi untuk mengukur sejauh mana model mampu mengenali seluruh data positif yang sebenarnya. Recall dihitung dengan rumus:
$$\text{Recall} = \frac{TP}{TP + FN}$$
Nilai recall yang tinggi menunjukkan bahwa sebagian besar data positif berhasil diidentifikasi dengan benar, yang sangat penting dalam konteks di mana kegagalan mendeteksi data positif (FN) harus dihindari.

Pemilihan ketiga metrik ini dilakukan karena masing-masing memberikan gambaran yang berbeda tentang performa model. Akurasi memberikan overview umum, sedangkan presisi dan recall memberikan insight yang lebih tajam terhadap kualitas klasifikasi, terutama dalam konteks imbalanced data atau jika satu kelas lebih penting dari kelas lain dalam aplikasi nyata.

Metodologi penelitian ini dirancang agar mampu menghasilkan evaluasi yang obyektif, sistematis, dan dapat direplikasi, sekaligus mempertimbangkan prinsip-prinsip keilmuan yang berlaku dalam analisis data berbasis machine learning. Dengan implementasi yang dilakukan pada platform RapidMiner Studio, seluruh proses mulai dari pengolahan data, pembagian dataset, pelatihan model, hingga evaluasi akhir dapat dirancang dalam bentuk workflow visual yang terstandar, efisien, dan mudah dimonitor secara komputasional.

III. DESAIN, HASIL DAN PEMBAHASAN

Penelitian ini dirancang untuk mengevaluasi kinerja dua algoritma machine learning, yaitu K-Nearest Neighbors (KNN) dan Naive Bayes, dalam mengklasifikasikan tipe kepribadian individu berdasarkan perilaku sosial. Fokus utama dari desain ini adalah membandingkan performa kedua model secara adil dan terukur melalui pendekatan eksperimen yang dilakukan menggunakan perangkat lunak RapidMiner Studio. Platform ini dipilih karena mendukung pembuatan workflow eksperimen yang visual, sistematis, serta memungkinkan integrasi berbagai proses seperti pembacaan data, pelatihan model, pengujian, dan evaluasi secara menyeluruh [10], [11].

Desain eksperimen ini menekankan prinsip validitas internal, di mana semua variabel dikendalikan agar perbandingan antar-algoritma tidak dipengaruhi oleh faktor eksternal. Untuk itu, digunakan dataset yang sama, metode pembagian data yang identik, serta parameter yang dikalibrasi agar sesuai dengan karakteristik algoritma masing-masing. Hasil eksperimen kemudian dianalisis menggunakan metrik evaluasi yang telah dijelaskan pada bab sebelumnya, yaitu akurasi, presisi, dan recall, guna mendapatkan gambaran menyeluruh terhadap kekuatan prediktif masing-masing model [12].

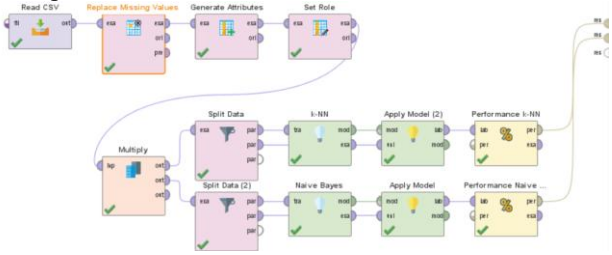
A. Desain Eksperimen

Desain eksperimen terdiri dari beberapa tahapan penting yang dijalankan dalam workflow RapidMiner, yang disusun sedemikian rupa untuk memastikan proses pelatihan dan evaluasi berjalan tanpa terjadi data leakage, serta menghasilkan perbandingan performa yang valid:

- Import Dataset: Tahap pertama adalah memuat dataset ke dalam lingkungan kerja RapidMiner dengan menggunakan operator Read CSV. Dataset yang dimuat berisi 2.900 entri, masing-masing merepresentasikan satu individu dengan atribut perilaku sosial yang telah dipersiapkan dalam format numerik. Data yang digunakan telah melalui tahap preprocessing, seperti pengisian nilai kosong, normalisasi data, dan encoding kategori menjadi

numerik, agar dapat diproses oleh algoritma machine learning secara optimal.

- Replacing misshing value: tahap penanganan nilai kosong, yang dilakukan menggunakan operator Replace Missing Values untuk mengisi entri yang hilang dengan pendekatan yang sesuai (misalnya rata-rata untuk data numerik). Langkah ini penting agar model tidak menerima data yang tidak lengkap, yang dapat menurunkan akurasi prediksi.
- Generate Atribut: normalisasi tipe data, khususnya untuk atribut yang seharusnya bertipe bilangan bulat seperti Friends_circle, Social_event, dan Post_frequency. Beberapa nilai pada atribut tersebut ditemukan dalam bentuk desimal yang tidak valid secara konseptual, sehingga dibulatkan menggunakan operator Generate Attributes dengan ekspresi round() untuk menjamin konsistensi.
- Set Role: Langkah terakhir dalam tahap preprocessing adalah penetapan atribut Personality sebagai label target, yang dilakukan menggunakan operator Set Role. Hal ini memungkinkan algoritma pembelajaran terawasi (supervised learning) mengenali atribut mana yang menjadi referensi klasifikasi selama proses training dan evaluasi.
- Data Branching: Untuk memastikan bahwa kedua algoritma diuji terhadap data yang sama tanpa saling mempengaruhi, digunakan operator Multiply untuk menggandakan dataset menjadi dua cabang. Setiap cabang digunakan secara eksklusif oleh satu algoritma, yaitu KNN dan Naive Bayes, tanpa intervensi satu sama lain. Hal ini penting untuk menjaga validitas eksperimen serta menghindari efek bias antar-model.
- Split Data: Dataset pada masing-masing jalur kemudian dibagi menjadi data latih (70%) dan data uji (30%) menggunakan operator Split Data, dengan pengaturan random seed yang konsisten untuk kedua jalur. Penggunaan seed yang sama bertujuan untuk memastikan bahwa subset data uji yang digunakan dalam evaluasi benar-benar identik untuk kedua model, sehingga perbandingan performa dapat dilakukan secara setara [5].
- Model Building:
  Pada jalur pertama, digunakan algoritma K-Nearest Neighbors dengan parameter nilai k = 5 dan menggunakan metrik jarak Euclidean Distance.Pada jalur kedua, dibangun model klasifikasi menggunakan algoritma Naive Bayes, dengan pendekatan berbasis probabilitas terhadap distribusi fitur.
- Apply Model & Evaluation: Model yang telah dilatih pada data latih kemudian diuji terhadap data uji menggunakan operator Apply Model. Hasil prediksi dari masing-masing model selanjutnya dievaluasi menggunakan operator Performance, yang menghitung metrik-metrik klasifikasi seperti akurasi, presisi, dan recall secara otomatis. Proses ini dilakukan secara independen pada masing-masing jalur, sehingga hasil evaluasi mencerminkan kinerja nyata model terhadap data yang tidak terlihat selama pelatihan.



Gambar 2 Workflow pemodelan dan evaluasi menggunakan RapidMiner

Dengan rancangan paralel seperti ini, analisis terhadap keunggulan relatif dari masing-masing algoritma dapat dilakukan secara objektif dan bebas bias. Evaluasi model pun tidak melibatkan data yang telah digunakan saat pelatihan, sehingga kemungkinan overfitting dapat diminimalisir.

B. Hasil Eksperimen

Setelah kedua model dibangun dan diuji, diperoleh hasil evaluasi performa yang ditunjukkan oleh tiga metrik utama: akurasi, presisi, dan recall. Selain itu, disajikan pula confusion matrix yang menggambarkan distribusi prediksi masing-masing kelas.

1) Model Naive Bayes

- Akurasi: 93.33%
- Presisi (positive class: Introvert): 91.20%
- Recall (positive class: Introvert): 95.51%

Confusion Matrix:

TABEL II
CONFUSION MATRIX NAIVE BAYES

	True Extrovert	True Introvert	Precision
Pred. Extrovert	408	19	95.55%
Pred. Introvert	39	404	91.20%

Recall:

- Extrovert: 91.28%
- Introvert: 95.51%

2) Model K-Nearest Neighbors (KNN)

- Akurasi: 93.68%
- Presisi (positive class: Introvert): 94.23%
- Recall (positive class: Introvert): 92.67%

Confusion Matrix:

TABEL II
CONFUSION MATRIX KNN

	True Extrovert	True Introvert	Precision
Pred. Extrovert	423	31	93.17%
Pred. Introvert	24	392	94.23%

Recall:

- Extrovert: 94.63%

- Introvert: 92.67%

C. Pembahasan

Hasil eksperimen menunjukkan bahwa kedua algoritma menghasilkan performa klasifikasi yang tinggi, dengan akurasi di atas 93%. Hal ini membuktikan bahwa atribut perilaku sosial yang digunakan dalam penelitian ini secara signifikan mencerminkan kecenderungan kepribadian individu, khususnya dalam membedakan antara introvert dan ekstrovert. Adanya perbedaan performa antar algoritma juga mengindikasikan bahwa masing-masing metode memiliki kekuatan spesifik dalam menangkap pola data.

Model Naive Bayes menunjukkan keunggulan dalam aspek recall terhadap kelas introvert, yakni sebesar 95,51% [1], [2]. Ini berarti bahwa model tersebut sangat sensitif dalam mengenali individu introvert dan jarang melewatkan sampel dari kelas tersebut. Kemampuan ini sangat penting dalam aplikasi seperti asesmen psikologis atau sistem deteksi dini, di mana identifikasi individu dengan kecenderungan introvert secara akurat menjadi prioritas. Meskipun asumsi independensi antar fitur dalam Naive Bayes jarang terpenuhi secara sempurna, hasil ini menunjukkan bahwa distribusi probabilistik dari fitur perilaku sosial cukup representatif untuk menangkap pola introversi.

Sebaliknya, model KNN memiliki keunggulan pada akurasi keseluruhan (93,68%) dan presisi kelas introvert (94,23%). Presisi yang tinggi menunjukkan bahwa model KNN sangat efektif dalam menghindari kesalahan klasifikasi positif palsu (false positives) [1], [8]. Artinya, jika model memprediksi seorang individu sebagai introvert, maka kemungkinan besar prediksi tersebut benar. Hal ini sangat bermanfaat dalam aplikasi seperti rekomendasi konten digital, pengembangan strategi komunikasi personal, atau sistem HR analytics, di mana kesalahan klasifikasi dapat menyebabkan hasil atau tindakan yang tidak relevan.

Perbedaan kinerja ini juga dapat dikaitkan dengan karakteristik dataset. Karena dataset telah dinormalisasi dan terdiri dari fitur numerik yang cukup homogen, algoritma berbasis jarak seperti KNN mampu bekerja optimal, sedangkan Naive Bayes tetap kompetitif karena jumlah fitur tidak terlalu besar dan distribusi data relatif sederhana.

Temuan ini mendukung hasil studi sebelumnya yang dilakukan oleh Kartarina et al. [1], yang menyatakan bahwa KNN unggul pada data numerik berdimensi rendah, sedangkan Naive Bayes lebih efisien secara komputasi. Penelitian ini juga memperkuat hasil riset oleh Azahari et al. [2], yang menunjukkan bahwa kedua metode mampu memberikan prediksi yang signifikan dalam domain data akademik mahasiswa, dengan akurasi tinggi pada kondisi tertentu.

Secara keseluruhan, penelitian ini menyimpulkan bahwa pemilihan algoritma klasifikasi untuk prediksi kepribadian sebaiknya didasarkan pada tujuan aplikasi:

- Jika diperlukan deteksi menyeluruh terhadap kelas tertentu (misalnya introvert), maka Naive Bayes lebih sesuai.
- Jika dibutuhkan ketepatan klasifikasi yang tinggi dengan minim kesalahan, maka KNN lebih direkomendasikan.

Dengan demikian, hasil penelitian ini membuktikan bahwa pola perilaku sosial dapat digunakan sebagai prediktor kepribadian, dan pemilihan metode klasifikasi dapat disesuaikan dengan konteks penggunaannya dalam sistem berbasis kecerdasan buatan..

IV. KESIMPULAN

Penelitian ini berhasil membuktikan bahwa pendekatan machine learning dapat digunakan secara efektif untuk memprediksi kepribadian individu, khususnya dalam membedakan antara tipe kepribadian introvert dan ekstrovert berdasarkan pola perilaku sosial. Dengan memanfaatkan atribut seperti waktu menyendiri, frekuensi kehadiran dalam kegiatan sosial, kecenderungan kelelahan setelah bersosialisasi, serta frekuensi membagikan informasi di media sosial, model yang dibangun mampu mengidentifikasi tipe kepribadian dengan tingkat akurasi tinggi.

Hasil pengujian model menunjukkan bahwa algoritma K-Nearest Neighbors (KNN) menghasilkan akurasi tertinggi sebesar 93,68%, sedangkan Naive Bayes menghasilkan akurasi sebesar 93,33% [1], [2], [8]. Performa ini menunjukkan bahwa kedua algoritma memiliki potensi kuat dalam klasifikasi kepribadian berbasis data perilaku. Lebih lanjut, Naive Bayes menunjukkan keunggulan pada aspek recall terhadap kelas introvert, yakni sebesar 95,51%, yang mengindikasikan kemampuannya dalam menangkap seluruh sampel dari kelas tersebut dengan sangat baik. Sementara itu, KNN menunjukkan presisi yang lebih tinggi, yakni 94,23%, yang berarti lebih tepat dalam memberikan label kelas introvert ketika prediksi dilakukan.

Temuan ini mendukung kesimpulan bahwa kedua algoritma mampu mengolah data perilaku sosial yang bersifat numerik menjadi dasar klasifikasi kepribadian yang valid. Performa yang ditunjukkan masing-masing model juga sejalan dengan karakteristik teknisnya: KNN lebih cocok untuk data numerik yang telah dinormalisasi dan memiliki dimensi yang tidak terlalu tinggi [1], [13], sedangkan Naive Bayes lebih efisien dan cocok untuk data dengan distribusi probabilistik yang sederhana [2], [9].

Dengan demikian, dapat disimpulkan bahwa baik KNN maupun Naive Bayes sama-sama layak diterapkan untuk tugas klasifikasi kepribadian berdasarkan data perilaku sosial. Pemilihan algoritma ideal dapat disesuaikan dengan kebutuhan aplikasi, seperti apakah lebih mengutamakan sensitivitas terhadap satu kelas (Naive Bayes) atau ketepatan klasifikasi keseluruhan (KNN). Penelitian ini juga memperkuat pemahaman bahwa data perilaku digital memiliki potensi besar untuk digunakan dalam prediksi kepribadian yang praktis dan efisien [3], [4], [6], [14].

REFERENSI

[1] N. L. P. Juniarti, N. K. Sriwinarti, dan Kartarina, "Analisis Metode K-Nearest Neighbors (K-NN) dan Naive Bayes dalam Memprediksi Kelulusan Mahasiswa," J. Teknol. Inf. dan Multimedia, vol. 3, no. 2, pp. 106–112, 2021.

[2] A. Azahari, Y. Yulindawati, D. Rosita, dan S. Mallala, "Komparasi Data Mining Naive Bayes dan Neural Network memprediksi Masa Studi Mahasiswa S1," J. Teknol. Inf. dan Ilmu Komput., vol. 7, no. 3, p. 443, 2020, doi: 10.25126/jtiik.2020732093.

[3] A. Mustofa, H. S. Nugraha, dan Y. Y. Wibowo, "Classification Algorithms to Predict Students' Extraversion-Introversion Traits," in Proc. 3rd Int. Conf. on Cybernetics and Intelligent Systems (ICORIS), pp. 1–6, 2021, doi: 10.1109/ICORIS52787.2021.9533275.

[4] S. S. Prakoso, "Big Five Personality Prediction Based in Online Behavior," J. Teknol. dan Sistem Komputer, vol. 9, no. 3, pp. 203–210, 2021.

[5] M. R. Jannah, E. Pratiwi, dan S. A. Nugroho, "Analisis Sentimen Relokasi Ibukota Nusantara Menggunakan Algoritma Naïve Bayes dan KNN," J. KomtekInfo, vol. 8, no. 1, pp. 25–30, 2022.

[6] A. K. Subramanian, "Personality Trait Prediction by Machine Learning Using Physiological Data and Driving Behavior," ICT Express, vol. 8, no. 1, pp. 63–69, 2022, doi: 10.1016/j.icte.2021.06.008.

[7] S. Anand, "Personality Prediction from Social Media Text: An Overview," Int. J. Eng. Res. Technol. (IJERT), vol. 9, no. 5, pp. 101–106, 2020.

[8] D. R. Permana dan A. Saputra, "Analysis of Classification and Naïve Bayes Algorithm K-Nearest Neighbor in Data Mining," IOP Conf. Ser.: Mater. Sci. Eng., vol. 1051, p. 012011, 2021, doi: 10.1088/1757-899X/1051/1/012011.

[9] T. R. Kusumawardani dan R. Hidayatullah, "Prediction for Diagnosing Liver Disease in Patients using KNN and Naïve Bayes Algorithms," in 2020 Int. Conf. on Smart Technology and Applications (ICoSTA), IEEE, pp. 1–5, 2020, doi: 10.1109/ICoSTA48221.2020.1570607553.

[10] F. Saputra dan R. Sari, "Personality Prediction System Based on Signatures Using Machine Learning," IOP Conf. Ser.: J. Phys.: Conf. Ser., vol. 1402, p. 066059, 2020, doi: 10.1088/1742-6596/1402/6/066059.

[11] T. Sutanto dan B. Prakoso, "Personality Prediction using Machine Learning," in Proc. 2020 IEEE Int. Conf. on Data Science and Information Technology (DSIT), pp. 27–32, doi: 10.1109/DSIT49696.2020.00013.

[12] Y. P. Rahmawati, R. Khotimah, dan M. Arifin, "Performance Comparison of SVM, Naïve Bayes, and KNN Algorithms for Analysis of Public Opinion Sentiment Against COVID-19 Vaccination on Twitter," J. Advances in Information Systems and Technology, vol. 2, no. 1, pp. 1–8, 2021.

[13] N. Styawati, "Comparison of Support Vector Machine and Naïve Bayes on Twitter Data Sentiment Analysis," J. Informatika: J. Pengembangan IT, vol. 6, no. 1, pp. 20–26, 2021.

[14] I. Surya dan D. Oktaviani, "Personality Classification from Online Text," Int. J. Eng. Res. Technol. (IJERT), vol. 9, no. 12, pp. 159–164, 2020.

[15] S. Başaran dan O. H. Ejimogu, "A Neural Network Approach for Predicting Personality From Facebook Data," in Proc. 2021 Int. Conf. on Computer Science and Engineering (UBMK), pp. 662–667, 2021, doi: 10.1109/UBMK52708.2021.9558872.