P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

Analisis Penerapan Algoritma Decision Tree, Random Forest, dan Naive Bayes pada Pengolahan Data untuk Sistem Cerdas

Judea Tirta Jordan Simamora[1], Alex Septama Sihite[2],Octav Kornelius Hutagaol[3], Marcel Alezandro Sihombing[4]

Ilmu Komputer, Fakultas Matematika dan Ilmu Pengetahuan, Universitas HKBP Nommensen Pematang Siantar, Sumatera Utara, Indonesia

Email : judeasimamora@gmail.com[1], alexsihite08@gmail.com[2], Korneliushutagaol82@gmail.com[3], Marcelalezandro25@gmail.com[4]

Email responden : judeasimamora@gmail.com

ABSTRAK

Perkembangan teknologi informasi dan transformasi digital telah menghasilkan volume data yang terus meningkat pada berbagai bidang, seperti pendidikan, kesehatan, bisnis, dan sistem informasi. Kondisi tersebut menimbulkan tantangan dalam proses pengolahan data karena metode konvensional sering kali kurang mampu menangani data dalam jumlah besar secara cepat dan akurat. Salah satu solusi yang banyak digunakan untuk mengatasi permasalahan tersebut adalah machine learning, yaitu cabang dari kecerdasan buatan (artificial intelligence) yang memungkinkan sistem mempelajari pola dari data dan menghasilkan prediksi maupun keputusan secara otomatis. Penelitian ini bertujuan untuk menganalisis penerapan machine learning dalam pengolahan data untuk sistem cerdas, mengidentifikasi algoritma yang paling sering digunakan, serta mengkaji manfaat dan tantangan yang dihadapi dalam implementasinya. Metode yang digunakan adalah literature review dengan mengumpulkan, menyeleksi, dan menganalisis berbagai sumber ilmiah berupa jurnal, buku, prosiding, dan artikel akademik yang relevan. Analisis dilakukan secara deskriptif kualitatif untuk memperoleh gambaran yang komprehensif mengenai perkembangan penerapan machine learning. Hasil kajian menunjukkan bahwa algoritma yang paling banyak digunakan meliputi Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), dan Naive Bayes. Penerapan algoritma tersebut terbukti mampu meningkatkan akurasi prediksi, mempercepat proses analisis data, serta mendukung pengambilan keputusan pada berbagai sektor. Kontribusi penelitian ini adalah memberikan sintesis literatur mengenai peran machine learning dalam pengolahan data untuk sistem cerdas sekaligus mengidentifikasi peluang pengembangan penelitian di masa mendatang. Temuan penelitian menunjukkan bahwa kualitas data, kebutuhan komputasi, dan interpretabilitas model masih menjadi tantangan utama yang perlu mendapat perhatian lebih lanjut.

Kata Kunci: Machine Learning, Pengolahan Data, Sistem Cerdas, Artificial Intelligence.

https://doi.org/10.37817/ikraith-informatika.v10i2
P-ISSN : 2580-4316
E-ISSN : 2654-8054

ABSTRACT

The rapid growth of information technology has led to a significant increase in the volume of data generated across various sectors. This condition creates challenges in data processing, particularly in obtaining accurate, efficient, and timely information to support intelligent decision-making. Machine learning, as a branch of artificial intelligence, has emerged as one of the most widely used approaches for addressing these challenges due to its ability to learn patterns from data and generate predictions automatically. However, the successful implementation of machine learning is influenced by several factors, including data quality, algorithm selection, and model evaluation processes. This study aims to analyze the application of machine learning in data processing for intelligent systems through a literature review approach. The research method was conducted by collecting, selecting, reviewing, and synthesizing relevant scientific literature from journals, books, conference proceedings, and academic publications. The analysis employed a qualitative descriptive approach to identify commonly used algorithms, implementation benefits, challenges, and recent development trends in machine learning applications. The results indicate that algorithms such as Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes are among the most frequently applied methods in intelligent data processing systems. The findings also show that machine learning contributes significantly to improving prediction accuracy, accelerating data analysis, automating decision-making processes, and enhancing system adaptability across various domains, including healthcare, education, business, finance, and information systems. Nevertheless, several challenges remain, such as low-quality datasets, imbalanced data, high computational requirements, overfitting, underfitting, and limited model interpretability. This study contributes by providing a comprehensive overview of machine learning applications, highlighting the importance of data preprocessing and appropriate algorithm selection, and identifying future research opportunities related to efficient, transparent, and explainable intelligent systems.

Keywords : Machine Learning, Data Processing, Intelligent Systems, Artificial Intelligence.

1. PENDAHULUAN

Perkembangan teknologi informasi telah menghasilkan pertumbuhan data yang sangat cepat dalam berbagai bidang kehidupan. Kondisi ini menuntut metode pengolahan data yang mampu bekerja secara cepat, akurat, dan efisien agar informasi yang dihasilkan dapat dimanfaatkan secara optimal. Salah satu pendekatan yang banyak digunakan untuk menjawab kebutuhan tersebut adalah machine learning, yaitu teknik yang memungkinkan sistem komputer mempelajari pola dari data dan menggunakan pengalaman tersebut untuk melakukan prediksi atau pengambilan keputusan secara otomatis [1];[2]. Dalam konteks yang lebih luas, *machine learning* juga dipahami sebagai bagian dari *artificial intelligence* yang berfokus pada kemampuan sistem untuk belajar dari data dan meningkatkan performanya seiring bertambahnya pengalaman, Penerapan *machine learning* dalam pengolahan data memiliki peran penting pada pengembangan sistem cerdas. Teknologi ini dapat digunakan untuk mengenali pola, mengelompokkan data, melakukan klasifikasi, serta menghasilkan rekomendasi yang mendukung proses pengambilan keputusan. Berbagai bidang seperti pendidikan, kesehatan, bisnis, keamanan, dan sistem informasi telah

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

memanfaatkan *machine learning* untuk meningkatkan kualitas layanan dan efektivitas analisis data [1];[3]. Dengan demikian, *machine learning* tidak hanya berfungsi sebagai alat analisis, tetapi juga menjadi komponen utama dalam membangun sistem yang adaptif dan berbasis data. Dalam prosesnya, pengolahan data pada *machine learning* melibatkan beberapa tahapan penting, mulai dari pengumpulan data, pembersihan data, transformasi data, pemilihan fitur, hingga evaluasi model. Tahapan-tahapan tersebut berpengaruh besar terhadap kualitas hasil yang diperoleh karena model *machine learning* sangat bergantung pada data yang digunakan untuk belajar. Data yang tidak lengkap, tidak konsisten, atau mengandung banyak *noise* dapat menurunkan akurasi model, sedangkan data yang bersih dan relevan akan membantu model menghasilkan prediksi yang lebih baik [4];[5]. Oleh karena itu, pengolahan data menjadi salah satu aspek yang sangat menentukan keberhasilan penerapan *machine learning* dalam sistem cerdas.

Secara umum, *machine learning* dibagi menjadi tiga jenis utama, yaitu *supervised learning, unsupervised learning, dan reinforcement learning* [1]. *Supervised learning* digunakan pada data berlabel untuk keperluan klasifikasi dan prediksi, sedangkan *unsupervised learning* digunakan untuk menemukan pola dalam data tanpa label, seperti *clustering*. Sementara itu, *reinforcement learning* bekerja dengan mekanisme penghargaan dan hukuman untuk menghasilkan keputusan yang optimal berdasarkan pengalaman. Ketiga pendekatan ini memiliki karakteristik masing-masing dan digunakan sesuai dengan kebutuhan permasalahan yang dihadapi. Dalam penerapan praktisnya, beberapa algoritma yang sering digunakan dalam *machine learning* antara lain *Decision Tree, Random Forest, K-Nearest Neighbor, Support Vector Machine,* dan *Naive Bayes.* Decision Tree dikenal mudah dipahami karena membentuk struktur keputusan yang jelas. *Random Forest* sering dipilih karena mampu meningkatkan akurasi dan mengurangi risiko *overfitting*. K-Nearest Neighbor digunakan untuk klasifikasi berdasarkan kemiripan data, sedangkan *Support Vector Machine* efektif untuk memisahkan kelas data yang kompleks. *Naive Bayes* banyak digunakan pada klasifikasi teks dan deteksi spam karena proses komputasinya yang relatif cepat [1]. Penelitian lain juga menunjukkan bahwa Decision Tree, Random Forest, dan Naive Bayes merupakan algoritma yang sering dibandingkan dalam kasus klasifikasi karena masing-masing memiliki karakteristik dan performa yang berbeda [6];[7].

Penelitian terdahulu menunjukkan bahwa *machine learning* mampu meningkatkan akurasi dan efisiensi dalam pengolahan data, namun tetap memiliki sejumlah tantangan. Hilmi dan Susanto menjelaskan bahwa keberhasilan model sangat dipengaruhi oleh kualitas data serta tahapan pra-pemrosesan yang dilakukan sebelum pemodelan [5]. Di sisi lain, Hastie, Tibshirani, dan Friedman menekankan bahwa masalah seperti *overfitting, underfitting,* dan bias model masih menjadi kendala yang perlu diperhatikan dalam pengembangan sistem berbasis *machine learning.* Selain itu, literatur review juga menunjukkan bahwa pemilihan metode yang tepat, pengelolaan data yang baik, serta evaluasi model yang sistematis merupakan faktor penting dalam menghasilkan kesimpulan yang valid [8];[8]. Meskipun demikian, masih terdapat beberapa *gap* dalam penelitian sebelumnya. Pertama, banyak studi hanya berfokus pada akurasi algoritma tanpa menjelaskan keterkaitan antara tahapan pengolahan data dan performa model. Kedua, penelitian yang membandingkan algoritma sering dilakukan pada kasus tertentu sehingga hasilnya belum tentu dapat digeneralisasi untuk seluruh konteks sistem cerdas. Ketiga, kajian yang secara khusus menggabungkan pembahasan *machine learning*, pengolahan data, dan perbandingan algoritma Decision Tree, Random Forest, serta Naive Bayes masih relatif terbatas [5];[3]. Oleh karena itu,

P-ISSN : 2580-4316
https://doi.org/10.37817/ikraith-informatika.v10i2 E-ISSN : 2654-8054

artikel ini disusun untuk mengisi celah tersebut melalui studi literatur yang lebih terarah [6];[7].

Berdasarkan uraian tersebut, tujuan penelitian ini adalah menganalisis penerapan *machine learning* pada pengolahan data untuk sistem cerdas dengan fokus pada algoritma Decision Tree, Random Forest, dan Naive Bayes. Penelitian ini juga bertujuan mengidentifikasi peran tahapan pra-pemrosesan data dalam mendukung performa algoritma, membandingkan kelebihan dan kekurangan tiap algoritma, serta merangkum tantangan umum dalam implementasinya. Hasil kajian diharapkan dapat memberikan gambaran yang lebih komprehensif mengenai pemanfaatan *machine learning* sebagai solusi dalam pengolahan data untuk mendukung sistem cerdas [3];[9].

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini menggunakan metode literature review untuk menganalisis penerapan machine learning pada pengolahan data untuk sistem cerdas. Metode ini dilakukan dengan mengumpulkan, menelaah, dan mensintesis berbagai sumber ilmiah yang relevan sehingga diperoleh pemahaman yang komprehensif mengenai topik penelitian [10].

Tahapan penelitian terdiri atas beberapa langkah. Pertama, menentukan topik penelitian dan merumuskan tujuan penelitian. Kedua, menentukan kata kunci pencarian seperti machine learning, data processing, intelligent systems, dan literature review. Ketiga, mengumpulkan literatur dari jurnal ilmiah, prosiding, buku, dan publikasi akademik lainnya yang relevan. Keempat, melakukan seleksi literatur berdasarkan kesesuaian topik, kualitas sumber, dan tahun publikasi. Kelima, melakukan analisis terhadap literatur yang terpilih dengan mengidentifikasi algoritma yang digunakan, manfaat penerapan machine learning, serta kendala yang ditemukan pada setiap penelitian. Tahap terakhir adalah menyusun hasil analisis dan menarik kesimpulan penelitian.

Penelitian dimulai dari tahap identifikasi topik untuk menentukan fokus kajian. Tahap berikutnya adalah pencarian literatur yang relevan dari berbagai sumber ilmiah. Literatur yang diperoleh kemudian diseleksi berdasarkan kriteria tertentu sehingga hanya sumber yang sesuai dengan tujuan penelitian yang digunakan. Setelah itu dilakukan analisis terhadap literatur terpilih untuk mengidentifikasi temuan-temuan penting terkait penerapan machine learning. Hasil analisis selanjutnya disintesis untuk memperoleh pemahaman yang menyeluruh dan digunakan sebagai dasar dalam penyusunan kesimpulan penelitian.

2.2 Kajian Pustaka Algoritma

Pada penelitian ini, algoritma yang dikaji adalah *Decision Tree, Random Forest, dan Naive Bayes.* Ketiga algoritma ini dipilih karena sering digunakan dalam penelitian klasifikasi dan pengolahan data, serta memiliki karakteristik yang berbeda sehingga cocok untuk dianalisis secara komparatif [11]. *Decision Tree* merupakan algoritma yang membentuk struktur pohon keputusan berdasarkan atribut data untuk menghasilkan aturan klasifikasi yang mudah dipahami. Kelebihan utama algoritma ini adalah *interpretabilitas* yang tinggi karena hasil keputusan dapat ditelusuri melalui cabang-cabang pohon. Namun, *Decision Tree* memiliki kelemahan berupa kecenderungan overfitting apabila data latih terlalu kompleks atau tidak seimbang [12]. *Random Forest* adalah pengembangan dari *Decision Tree* yang menggunakan banyak pohon keputusan secara bersamaan dan menggabungkan hasilnya melalui mekanisme voting atau rata-rata. Pendekatan ini meningkatkan stabilitas model dan mengurangi risiko overfitting. Dalam berbagai penelitian, *Random Forest* sering menunjukkan kinerja yang baik pada data dengan dimensi besar maupun data yang memiliki variasi tinggi. Karena itu, algoritma ini banyak dipakai dalam sistem cerdas yang

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

membutuhkan akurasi tinggi dan hasil yang relatif konsisten [11]. *Naive Bayes* merupakan algoritma klasifikasi berbasis probabilitas yang mengacu pada *Teorema Bayes*. Algoritma ini dikenal sederhana, cepat, dan efisien dalam proses komputasi. *Naive Bayes* banyak digunakan pada klasifikasi teks, deteksi spam, dan analisis sentimen karena mampu menghasilkan prediksi yang baik walaupun mengasumsikan independensi antarfitur. Walaupun sederhana, algoritma ini tetap menjadi pilihan yang relevan untuk data yang memiliki struktur tertentu dan membutuhkan proses klasifikasi yang cepat. Kajian terhadap ketiga algoritma tersebut menunjukkan bahwa tidak ada satu metode yang selalu unggul untuk semua jenis data. Pemilihan algoritma harus disesuaikan dengan karakteristik dataset, tujuan analisis, dan kebutuhan sistem yang ingin dibangun. Oleh karena itu, pembahasan dalam penelitian ini tidak hanya berfokus pada definisi algoritma, tetapi juga pada relevansi penerapannya dalam pengolahan data untuk sistem cerdas.

3. HASIL DAN PEMBAHASAN

3.1 Penerapan Machine Learning dalam Pengolahan Data untuk Sistem Cerdas

Berdasarkan hasil studi literatur yang telah dilakukan, machine learning memiliki peran yang sangat penting dalam mendukung proses pengolahan data pada sistem cerdas. Kemampuan machine learning dalam mempelajari pola dari data memungkinkan sistem melakukan klasifikasi, prediksi, dan pengambilan keputusan secara otomatis dengan tingkat akurasi yang tinggi [1];[2]. Berbeda dengan metode konvensional yang memerlukan aturan yang ditentukan secara eksplisit, machine learning memungkinkan sistem belajar dari data sehingga mampu beradaptasi terhadap perubahan kondisi dan karakteristik data yang terus berkembang [3]. Perkembangan teknologi digital telah menghasilkan volume data yang sangat besar dari berbagai sumber, seperti media sosial, sistem informasi, sensor, perangkat Internet of Things (IoT), hingga transaksi digital. Kondisi tersebut mendorong kebutuhan akan teknologi yang mampu mengolah data secara cepat dan efisien. Dalam konteks ini, machine learning menjadi salah satu solusi utama karena mampu mengubah data mentah menjadi informasi yang bernilai guna dalam mendukung pengambilan keputusan [9].

Penerapan machine learning saat ini telah berkembang pada berbagai bidang. Pada sektor kesehatan, teknologi ini digunakan untuk membantu diagnosis penyakit, analisis citra medis, dan prediksi risiko kesehatan pasien. Dalam bidang pendidikan, machine learning dimanfaatkan untuk menganalisis perilaku belajar peserta didik, mengidentifikasi tingkat keberhasilan pembelajaran, serta memberikan rekomendasi materi yang sesuai dengan kebutuhan pengguna. Pada sektor bisnis dan keuangan, machine learning banyak digunakan dalam analisis perilaku konsumen, sistem rekomendasi produk, prediksi penjualan, dan deteksi aktivitas penipuan (fraud detection) [9];[13]. Hasil kajian menunjukkan bahwa keberhasilan implementasi machine learning sangat dipengaruhi oleh kualitas data yang digunakan. Data yang lengkap, bersih, dan relevan akan menghasilkan model yang memiliki performa lebih baik dibandingkan data yang mengandung banyak kesalahan atau ketidakkonsistenan [6]. Oleh karena itu, pengolahan data menjadi salah satu faktor utama dalam keberhasilan pembangunan sistem cerdas berbasis machine learning.

3.1.1 Tahapan Penerapan Machine Learning

Berdasarkan berbagai literatur yang dikaji, penerapan machine learning dalam pengolahan data umumnya dilakukan melalui beberapa tahapan utama.

a. Pengumpulan Data

Tahap pertama adalah pengumpulan data (data collection). Data diperoleh dari berbagai sumber seperti basis data, aplikasi, sensor, dokumen digital, maupun platform

daring. Data yang dikumpulkan harus relevan dengan tujuan analisis karena kualitas data akan memengaruhi performa model yang dibangun [6].

b. Pra-pemrosesan Data

Tahap berikutnya adalah data preprocessing atau pembersihan data. Pada tahap ini dilakukan penanganan terhadap missing value, data duplikat, data tidak konsisten, serta outlier. Tahap ini bertujuan meningkatkan kualitas data sehingga model dapat belajar secara optimal [6].

c. Transformasi dan Seleksi Fitur

Setelah data dibersihkan, dilakukan transformasi data melalui normalisasi, standarisasi, maupun konversi data kategorikal ke bentuk numerik. Selanjutnya dilakukan seleksi fitur untuk memilih atribut yang paling relevan sehingga dapat meningkatkan efisiensi dan akurasi model [4].

d. Pelatihan Model

Tahap pelatihan (training) dilakukan dengan menggunakan algoritma tertentu untuk mempelajari pola dari data latih (training data). Pada tahap ini model mulai membentuk hubungan antarvariabel sehingga dapat digunakan untuk melakukan prediksi atau klasifikasi [2].

e. Pengujian dan Evaluasi Model

Tahap terakhir adalah pengujian model menggunakan testing data. Evaluasi dilakukan menggunakan indikator seperti akurasi, presisi, recall, dan F1-score untuk mengetahui tingkat keberhasilan model dalam menyelesaikan permasalahan yang dihadapi [14].

Tahapan tersebut menunjukkan bahwa keberhasilan machine learning tidak hanya bergantung pada algoritma yang digunakan, tetapi juga pada kualitas proses pengolahan data yang dilakukan sebelum proses pelatihan model.

3.1.2 Algoritma Machine Learning yang Sering Digunakan

Hasil kajian literatur menunjukkan bahwa algoritma yang paling sering digunakan dalam pengolahan data untuk sistem cerdas adalah *Decision Tree, Random Forest, K-Nearest Neighbor (KNN), Support Vector Machine (SVM),* dan *Naive Bayes.* Masing-masing algoritma memiliki mekanisme kerja, kelebihan, dan keterbatasan yang berbeda sehingga pemilihannya perlu disesuaikan dengan karakteristik data dan tujuan analisis. *Decision Tree* merupakan algoritma klasifikasi yang membangun model dalam bentuk struktur pohon keputusan yang terdiri atas root node, node internal, *branch*, dan *leaf node*. Proses pembentukan pohon diawali dengan memilih atribut terbaik sebagai root node menggunakan kriteria seperti Information Gain atau Gini Index. Selanjutnya dilakukan proses *splitting*, yaitu membagi data ke dalam beberapa subset berdasarkan atribut yang dipilih. Pembagian tersebut berlangsung secara berulang melalui mekanisme *recursive* partitioning hingga seluruh data berada pada leaf node atau memenuhi kondisi penghentian tertentu. Hasil akhir berupa aturan keputusan (if–then rules) yang mudah dipahami sehingga *Decision Tree* banyak digunakan pada sistem cerdas yang membutuhkan interpretasi hasil yang tinggi. Meskipun demikian, algoritma ini rentan mengalami overfitting apabila pohon yang terbentuk terlalu kompleks [15].Penelitian [15] menunjukkan bahwa penerapan Decision Tree dengan seleksi fitur Chi-Square mampu meningkatkan kinerja klasifikasi tingkat kelulusan mahasiswa. Temuan tersebut menunjukkan bahwa pemilihan fitur yang tepat dapat meningkatkan efektivitas proses pembentukan pohon keputusan. Selain itu, penelitian [16] juga

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

membuktikan bahwa Decision Tree mampu mengidentifikasi pola pemahaman mahasiswa pada mata kuliah pemrograman dengan menghasilkan aturan klasifikasi yang mudah diinterpretasikan.

Random Forest merupakan pengembangan dari *Decision Tree* yang menerapkan konsep *ensemble* learning. Proses algoritma diawali dengan membangun sejumlah Decision Tree menggunakan teknik *bootstrap* sampling, yaitu mengambil sampel data secara acak dengan pengembalian. Selain itu, pada setiap node hanya sebagian fitur yang dipilih secara acak sebagai kandidat pemisahan sehingga setiap pohon memiliki karakteristik yang berbeda. Setelah seluruh pohon menghasilkan prediksi, keputusan akhir diperoleh melalui mekanisme *majority* voting pada masalah klasifikasi atau rata-rata prediksi pada masalah regresi. Pendekatan ini mampu meningkatkan stabilitas model, mengurangi variansi, dan menekan risiko overfitting dibandingkan penggunaan satu Decision Tree. Oleh karena itu, Random Forest banyak digunakan pada pengolahan data berskala besar dan berdimensi tinggi yang membutuhkan tingkat akurasi tinggi [17]. [17] menunjukkan bahwa mekanisme pembentukan banyak pohon keputusan dan penggabungan hasil melalui majority voting mampu meningkatkan stabilitas model serta menghasilkan akurasi yang lebih baik dibandingkan penggunaan satu Decision Tree. Temuan tersebut menunjukkan bahwa Random Forest sesuai diterapkan pada data yang kompleks dan berdimensi tinggi.

K-Nearest Neighbor (KNN) bekerja berdasarkan prinsip kedekatan jarak antara data baru dan data latih yang telah memiliki label. Pada saat dilakukan klasifikasi, algoritma menghitung jarak antara data baru dengan seluruh data latih menggunakan ukuran jarak, seperti *Euclidean Distance,* kemudian memilih sejumlah k tetangga terdekat. Kelas yang paling banyak muncul di antara tetangga tersebut menjadi hasil klasifikasi. Algoritma ini sederhana dan mudah diterapkan, tetapi memerlukan waktu komputasi yang lebih tinggi ketika jumlah data sangat besar karena seluruh data latih harus dihitung setiap kali proses prediksi dilakukan.

Support Vector Machine (SVM) melakukan klasifikasi dengan mencari *hyperplane* optimal yang mampu memisahkan dua atau lebih kelas data dengan margin terbesar. Pada data yang tidak dapat dipisahkan secara linear, SVM memanfaatkan kernel function untuk memetakan data ke ruang berdimensi lebih tinggi sehingga batas pemisah dapat dibentuk secara lebih optimal. Pendekatan tersebut membuat SVM memiliki performa yang baik pada data berdimensi tinggi, meskipun proses pelatihannya membutuhkan komputasi yang lebih kompleks serta penentuan parameter yang tepat.

Naive Bayes merupakan algoritma klasifikasi berbasis probabilitas yang menggunakan Teorema *Bayes* dengan asumsi bahwa setiap atribut bersifat independen terhadap atribut lainnya. Proses klasifikasi diawali dengan menghitung prior probability setiap kelas, kemudian menghitung probabilitas setiap atribut terhadap masing-masing kelas (*likelihood*). Selanjutnya dihitung posterior probability, dan kelas dengan probabilitas terbesar dipilih sebagai hasil klasifikasi. Mekanisme tersebut membuat Naive Bayes memiliki proses komputasi yang cepat dan efisien sehingga banyak diterapkan pada klasifikasi teks, deteksi spam, analisis sentimen, serta berbagai permasalahan klasifikasi lainnya. Meskipun menggunakan asumsi independensi atribut yang sederhana, berbagai penelitian menunjukkan bahwa algoritma ini tetap mampu memberikan performa yang kompetitif pada berbagai jenis data [18].

3.2 Analisis Hasil Kajian Literatur

Hasil kajian terhadap berbagai sumber menunjukkan bahwa machine learning mampu memberikan kontribusi yang signifikan dalam meningkatkan kualitas pengolahan data. Teknologi ini dapat mempercepat proses analisis data, meningkatkan akurasi prediksi, serta membantu menemukan pola yang sulit dikenali secara manual [1]. Selain itu, machine learning juga memungkinkan proses pengambilan keputusan dilakukan secara lebih objektif karena keputusan yang dihasilkan didasarkan pada analisis data dan pola yang ditemukan selama proses pelatihan model [9]. Kondisi tersebut menjadikan machine learning sebagai salah satu fondasi utama dalam pengembangan sistem cerdas modern. Berdasarkan literatur yang dianalisis, faktor kualitas data menjadi aspek yang paling sering disebut sebagai penentu keberhasilan implementasi machine learning. Data yang tidak lengkap, tidak konsisten, atau mengandung banyak kesalahan dapat menyebabkan penurunan performa model secara signifikan [6]. Sebaliknya, data yang telah melalui proses preprocessing dengan baik cenderung menghasilkan model yang lebih akurat dan stabil. Hasil kajian juga menunjukkan bahwa tidak terdapat satu algoritma yang selalu unggul pada seluruh kasus. Setiap algoritma memiliki karakteristik, kelebihan, dan keterbatasan yang berbeda sehingga pemilihannya harus disesuaikan dengan kebutuhan sistem dan karakteristik data yang digunakan [4]. Temuan tersebut sejalan dengan penelitian [15] yang menunjukkan bahwa performa Decision Tree dapat ditingkatkan melalui seleksi fitur Chi-Square sehingga menghasilkan klasifikasi yang lebih optimal. Sementara itu, penelitian [16] membuktikan bahwa Decision Tree mampu mengidentifikasi pola pemahaman mahasiswa dengan tingkat interpretasi yang baik. Di sisi lain, [17] menjelaskan bahwa Random Forest menghasilkan performa yang lebih stabil dibandingkan Decision Tree karena menggunakan pendekatan ensemble learning. Selain itu, penelitian [18] menunjukkan bahwa Naive Bayes tetap mampu memberikan hasil klasifikasi yang baik dengan proses komputasi yang lebih cepat pada kasus rekomendasi pekerjaan bagi fresh graduate. Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa pemilihan algoritma perlu disesuaikan dengan karakteristik data dan tujuan analisis.

3.2.1 Tantangan dalam Penerapan Machine Learning

Hasil kajian literatur menunjukkan bahwa implementasi machine learning masih menghadapi berbagai tantangan. Salah satu tantangan utama adalah kualitas data yang rendah. Data yang tidak lengkap, tidak konsisten, maupun mengandung noise dapat menurunkan performa model sehingga menghasilkan prediksi yang kurang akurat. Selain itu, kondisi imbalanced data menyebabkan model cenderung mengenali kelas mayoritas dibandingkan kelas minoritas sehingga memengaruhi hasil klasifikasi. Tantangan lainnya adalah kebutuhan komputasi yang tinggi pada beberapa algoritma, terutama ketika diterapkan pada dataset berukuran besar. Dari sisi pemodelan, masalah overfitting dan underfitting juga masih sering dijumpai. Overfitting terjadi ketika model terlalu menyesuaikan data latih sehingga kurang mampu melakukan generalisasi terhadap data baru, sedangkan underfitting terjadi ketika model belum mampu mempelajari pola penting dari data. Selain itu, interpretabilitas model juga menjadi perhatian karena beberapa algoritma memiliki tingkat kompleksitas yang tinggi sehingga sulit dijelaskan kepada pengguna, terutama pada bidang kesehatan dan keuangan yang memerlukan transparansi dalam pengambilan keputusan [15] ; [17].

P-ISSN : 2580-4316
E-ISSN : 2654-8054 https://doi.org/10.37817/ikraith-informatika.v10i2

3.3 Pembahasan

Berdasarkan hasil kajian literatur, proses penyelesaian masalah dalam penerapan machine learning diawali dengan pengumpulan data, kemudian dilanjutkan dengan tahap preprocessing untuk membersihkan data dari missing value, noise, data duplikat, maupun ketidakkonsistenan. Selanjutnya dilakukan transformasi data dan seleksi fitur agar data lebih sesuai untuk diproses oleh algoritma. Setelah data siap, proses pelatihan model dilakukan menggunakan algoritma seperti Decision Tree, Random Forest, dan Naive Bayes, kemudian dilanjutkan dengan tahap pengujian serta evaluasi menggunakan metrik seperti akurasi, presisi, recall, dan F1-score. Rangkaian tahapan tersebut menunjukkan bahwa keberhasilan implementasi machine learning tidak hanya ditentukan oleh algoritma yang digunakan, tetapi juga oleh kualitas data dan proses pengolahan data sebelum model dibangun. Pada algoritma Decision Tree, proses klasifikasi dilakukan dengan membentuk struktur pohon keputusan yang diawali dari pemilihan atribut terbaik sebagai root node menggunakan kriteria seperti Information Gain atau Gini Index. Selanjutnya dilakukan proses splitting dan recursive partitioning hingga terbentuk leaf node yang merepresentasikan hasil klasifikasi. Mekanisme ini menghasilkan aturan keputusan (if–then rules) yang mudah dipahami sehingga sesuai digunakan pada sistem yang membutuhkan interpretasi hasil yang tinggi [17]. Sementara itu, Random Forest bekerja dengan membangun sejumlah Decision Tree menggunakan teknik bootstrap sampling dan pemilihan fitur secara acak. Hasil prediksi dari seluruh pohon kemudian digabungkan melalui mekanisme majority voting sehingga menghasilkan model yang lebih stabil dan mampu mengurangi risiko overfitting. Berbeda dengan kedua algoritma tersebut, Naive Bayes melakukan proses klasifikasi berdasarkan Teorema Bayes dengan menghitung probabilitas setiap kelas terhadap data masukan. Pendekatan probabilistik tersebut menjadikan Naive Bayes memiliki proses komputasi yang cepat dan efisien, terutama pada klasifikasi teks maupun dataset berukuran besar [18].

Berdasarkan hasil kajian literatur, ditemukan bahwa kualitas data merupakan faktor yang paling berpengaruh terhadap keberhasilan implementasi machine learning. Temuan ini sejalan dengan penelitian Hilmi dan Susanto [6] yang menyatakan bahwa proses data preprocessing memiliki pengaruh signifikan terhadap performa model prediksi. Penelitian tersebut menunjukkan bahwa model yang menggunakan data yang telah dibersihkan dan ditransformasikan mampu menghasilkan tingkat akurasi yang lebih tinggi dibandingkan model yang menggunakan data mentah. Hasil ini menunjukkan bahwa tahap preprocessing merupakan bagian penting dalam keseluruhan proses penerapan machine learning karena menentukan kualitas model yang dihasilkan. Dari sisi algoritma, hasil kajian menunjukkan bahwa Random Forest memiliki performa yang relatif stabil pada berbagai kasus klasifikasi. Temuan ini sejalan dengan penelitian [17] yang menjelaskan bahwa penggunaan pendekatan ensemble learning melalui pembentukan banyak pohon keputusan dan mekanisme majority voting mampu meningkatkan stabilitas model dibandingkan penggunaan satu Decision Tree. Selain itu, penelitian [15] menunjukkan bahwa performa Decision Tree dapat ditingkatkan melalui penerapan seleksi fitur Chi-Square, sehingga menghasilkan proses klasifikasi yang lebih optimal. Penelitian [16] juga membuktikan bahwa Decision Tree mampu mengidentifikasi pola pemahaman mahasiswa dengan menghasilkan aturan keputusan yang mudah diinterpretasikan. Di sisi lain, penelitian [18] menunjukkan bahwa

Naive Bayes tetap memberikan performa yang kompetitif dengan waktu komputasi yang relatif cepat pada proses klasifikasi. Temuan-temuan tersebut memperkuat hasil penelitian ini bahwa tidak terdapat satu algoritma yang selalu unggul untuk seluruh jenis permasalahan, melainkan pemilihannya harus disesuaikan dengan karakteristik data, tujuan analisis, serta kebutuhan sistem yang akan dibangun.

Selain itu, hasil penelitian Pratama et al. [9] menunjukkan bahwa machine learning telah banyak diterapkan pada sektor kesehatan, pendidikan, dan bisnis. Temuan tersebut sejalan dengan hasil kajian pada penelitian ini yang menunjukkan bahwa pemanfaatan machine learning semakin luas karena kemampuannya dalam mengolah data secara cepat, menemukan pola dari data dalam jumlah besar, serta mendukung proses pengambilan keputusan secara lebih objektif. Dengan demikian, penelitian ini tidak hanya menguatkan hasil penelitian terdahulu, tetapi juga memberikan sintesis yang lebih komprehensif mengenai hubungan antara kualitas data, tahapan pengolahan data, dan karakteristik algoritma dalam mendukung pengembangan sistem cerdas.

Secara keseluruhan, hasil kajian menunjukkan bahwa perkembangan machine learning saat ini tidak hanya berfokus pada peningkatan akurasi model, tetapi juga pada efisiensi komputasi, transparansi, dan kemampuan interpretasi model. Perkembangan konsep seperti Explainable Artificial Intelligence (XAI) semakin menunjukkan pentingnya menghasilkan model yang tidak hanya akurat, tetapi juga mudah dipahami oleh pengguna. Oleh karena itu, keberhasilan implementasi machine learning tidak hanya ditentukan oleh pemilihan algoritma, tetapi juga oleh kualitas data, proses pengolahan data, serta kesesuaian algoritma dengan karakteristik permasalahan yang dihadapi. Kajian ini memperkuat hasil penelitian terdahulu sekaligus memberikan gambaran yang lebih komprehensif mengenai penerapan machine learning dalam pengolahan data untuk sistem cerdas.

4. KESIMPULAN

Berdasarkan hasil studi literatur yang telah dilakukan, dapat disimpulkan bahwa keberhasilan penerapan machine learning pada sistem cerdas tidak hanya ditentukan oleh kemampuan algoritma dalam melakukan prediksi atau klasifikasi, tetapi juga oleh kualitas pengolahan data yang dilakukan sebelum proses pemodelan. Kajian ini menunjukkan bahwa setiap tahapan, mulai dari pengumpulan data, preprocessing, transformasi data, seleksi fitur, hingga evaluasi model, memiliki kontribusi yang saling berkaitan dalam menghasilkan keluaran yang akurat dan dapat diandalkan. Selain itu, hasil analisis memperlihatkan bahwa tidak terdapat satu algoritma yang dapat dianggap paling unggul untuk seluruh permasalahan karena performa model sangat dipengaruhi oleh karakteristik data, tujuan analisis, dan lingkungan implementasinya. Oleh sebab itu, pemilihan metode harus dilakukan secara tepat berdasarkan kebutuhan sistem yang akan dikembangkan. Kajian ini juga mengindikasikan bahwa arah pengembangan machine learning saat ini semakin berfokus pada peningkatan efisiensi komputasi, kemampuan adaptasi terhadap data dalam jumlah besar, serta transparansi hasil yang dihasilkan model. Temuan tersebut menunjukkan bahwa pengembangan sistem cerdas di masa mendatang perlu menyeimbangkan aspek akurasi, kecepatan pemrosesan, dan kemudahan interpretasi agar teknologi yang dibangun tidak hanya mampu memberikan hasil yang baik secara teknis, tetapi juga dapat digunakan secara efektif dan dipercaya dalam mendukung proses pengambilan keputusan pada berbagai bidang penerapan.

UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada semua pihak yang telah memberikan dukungan, bantuan, serta motivasi selama proses penyusunan artikel ini. Ucapan terima kasih juga disampaikan kepada dosen pembimbing, institusi pendidikan, serta berbagai penulis dan peneliti yang karya ilmiahnya menjadi referensi dalam penelitian ini. Dukungan dan kontribusi yang diberikan sangat membantu dalam penyelesaian artikel sehingga penelitian ini dapat disusun dengan baik. Semoga penelitian ini dapat memberikan manfaat dan menjadi referensi bagi pengembangan ilmu pengetahuan, khususnya pada bidang machine learning dan sistem cerdas.

DAFTAR PUSTAKA

[1] A. Wijoyo *et al.*, "Pembelajaran Machine Learning," vol. 3, no. 2, pp. 375–380, 2024.

[2] T. M. Mitchell, *Machine Learning*. McGraw-Hill Science/Engineering/Math, 1997.

[3] G. Tyovanny, J. Karu, V. P. Rantung, J. T. Informatika, F. Teknik, and U. N. Manado, "EduTIK: Jurnal Pendidikan Teknologi Informasi dan Komunikasi Volume 6 Nomor 1, Februari 2026," vol. 6, pp. 455–464, 2026.

[4] H. Crc and S. Marsland, *MACHINE LEARNING An Algorithmic Perspective Second Edition*. CRC Press, 2025.

[5] M. Abror, A. Hilmi, and A. Susanto, "Implementasi dan Evaluasi Model Machine Learning untuk Optimalisasi Prediksi Penjualan Produk Kue Kering," vol. 7, no. 3, pp. 1649–1661, 2025, doi: 10.47065/bits.v7i3.8657.

[6] K. H. Hanif, N. R. Muntiari, D. Harto, and D. S. Wiranata, "Perbandingan Analisis Sentimen Komentar Mahasiswa Prodi Teknik Komputer Menggunakan Algoritma Decision Tree , Support Vector Machine (SVM), dan Random Forest," vol. 12, no. 01, 2026.

[7] H. Wira Saputra, "Komparasi Algoritma Machine Learning Dalam Menganalisa Sentimen Opini Publik Terhadap Kenaikan Harga BBM Comparison of Machine Learning Algorithms in Analyzing Public Opinion Sentiments Against Fuel Price Increases," *J. Comput. Eng. Syst. Sci.*, vol. 8, no. 1, pp. 138–148, 2023, [Online]. Available: www.jurnal.unimed.ac.id

[8] J. N. Sibarani, D. R. Sirait, and S. S. Ramadhanti, "Intrusion Detection Systems pada Bot-IoT Dataset Menggunakan Algoritma Machine Learning," *J. Masy. Inform.*, vol. 14, no. 1, pp. 38–52, 2023, doi: 10.14710/jmasif.14.1.49721.

[9] D. K. S. Sekarhati, "Combating Hoax and Misinformation in Indonesia Using Machine Learning What is Missing and Future Directions," *Eng. Math. Comput. Sci. J.*, vol. 6, no. 2, pp. 143–150, 2024, doi: 10.21512/emacsjournal.v6i2.11556.

[10] F. Alfiaturrohmah and S. Sudianto, "A Segment Anything Model for Melon Pruning Based on Diameter," *Eng. Technol. Appl. Sci. Res.*, vol. 15, no. 5, pp. 26632–26639, 2025, doi: 10.48084/etasr.12207.

[11] A. Amin, S. Basri, M. F. Hassan, and M. Rehman, "Systematic Literature Review on Creativity in Software Engineering," pp. 370–396, 2016.

[12] A. R. Andhika, "Machine Learning dalam Pengembangan Perangkat Lunak," *Integr. Perspect. Soc. Sci. J.*, vol. 2, no. 01 Februari, pp. 1211–1222, 2025, [Online]. Available: https://ipssj.com/index.php/ojs/article/view/163

[13] A. A. Prayoga, M. Hasanuddin, S. Khodijah, and C. A. Rizki, "Analisis Penerapan Machine Learning dalam Sistem Prediksi dan Pengambilan Keputusan," vol. 1, no. 3, pp. 84–90, 2025.

[14] F. H. Saputra *et al.*, "Enhancing Intrusion Detection Using Random Forest and SMOTE on the NSL-KDD Dataset," *J. Syst. Comput. Eng.*, vol. 6, no. 3, pp. 240–247, 2025, doi: 10.61628/jsce.v6i3.2056.

[15] R. Wahyu Hardian and F. Ilmu Komputer, "Peningkatan Algoritma Decision Tree Dalam Mengklasifikasi Tingkat Kelulusan Mahasiswa Universitas Jambi Dengan Seleksi Fitur Chi-Square," *J. Manaj. Teknol. dan Sist. Inf.*, vol. 5, no. 2, pp. 1105–1114, 2025.

[16] S. N. Sari, P. Annisa, A. Nisa, D. Rahma, R. Prameswara Ritonga, and D. P. Utomo, "Teknik Data Mining Dengan Menggunakan Algoritma Decision Tree Untuk Mengetahui Pola Pemahaman Mahasiswa Pada Matakuliah Pemrograman," *Bull. Inf. Technol.*, vol. 6, no. 4, pp. 417–431, 2025, doi: 10.47065/bit.v5i2.2339.

[17] S. Lina, M. Sitio, S. Kom, and M. Kom, *MACHINE LEARNING BERBASIS POHON KEPUTUSAN : Penulis : Sartika Lina Mulani Sitio , S . Kom ., M . Kom*. 2026.

[18] I. N. Mulyanan, M. Yusril, H. Setyawan, and W. I. Rahayu, "Penerapan Metode Naïve Bayes Untuk Merekomendasikan," vol. 7, no. 5, pp. 3453–3460, 2023.