



Comparison of Supervised Machine Learning Classification Models in Knowledge Discovery of Pertamina Prabumulih IT Services

Perbandingan Model Klasifikasi Supervised Machine Learning dalam Knowledge Discovery Layanan TI Pertamina Prabumulih

Shafa Aurelliza Arian¹, Putri Rahel Alifia², Bagus Prihantoro³, Muhammad Iqbal Disriansyah⁴, Ken Ditha Tania⁵, Alsella Meiriza⁶, Ahmad Rifai⁷

^{1,2,3,4,5,6,7} Program Studi Sistem Informasi, Fakultas Ilmu Komputer,
Universitas Sriwijaya, Indonesia

E-Mail: ¹09031182328002@student.unsri.ac.id,
²09031182328006@student.unsri.ac.id, ³09031182328098@student.unsri.ac.id,
⁴09031282328056@student.unsri.ac.id, ⁵kenya.tania@gmail.com,
⁶allsella@unsri.ac.id, ⁷ahmadrifai@ilkom.unsri.ac.id

Received Feb 03rd 2026; Revised Mar 28th 2026; Accepted Mar 30th 2026; Available Online Apr 23th 2026

Corresponding Author: Shafa Aurelliza Arian

Copyright ©2026 by Authors, Published by Institut Riset dan Publikasi Indonesia (IRPI)

Abstract

The management data in Information Technology (IT) services at Pertamina Prabumulih requires an analytical approach to improve handling effectiveness and support data-driven decision-making. This study aims to compare the performance of several supervised machine learning classification models in IT services for the 2020–2025 period using the Knowledge discovery approach through data mining techniques on 8,627 initial records. The research stages include preprocessing, class labeling, handling data imbalance using the Synthetic Minority Over-sampling Technique (SMOTE), and data splitting scenarios of 70:30, 80:20, and 90:10. The classification process was carried out using Naïve Bayes, Support Vector Machine (SVM), and Random Forest algorithms. Model evaluation was conducted using accuracy, precision, recall, and F1-score metrics. The results indicate that the 90:10 scenario achieved the best performance, with Support Vector Machine (SVM) reaching an accuracy of 0.8806, precision of 0.8757, recall of 0.8941, and F1-score of 0.8788, outperforming the other algorithms. The desktop hardware category was identified as the most frequent case during the study period. These findings can serve as a basic strategy to prioritize service delivery, allocate resources, and strengthen Knowledge Management to improve IT services in a targeted, effective, and sustainable manner.

Keywords: Classification, Knowledge Discovery, IT Services, Supervised Machine Learning

Abstrak

Pengelolaan data layanan Teknologi Informasi (TI) di Pertamina Prabumulih memerlukan pendekatan analitik untuk meningkatkan efektivitas penanganan dan mendukung pengambilan keputusan berbasis data. Penelitian ini bertujuan membandingkan performa beberapa model klasifikasi *supervised machine learning* pada layanan TI periode 2020–2025 menggunakan pendekatan *knowledge discovery* melalui teknik *data mining* terhadap 8.627 data awal. Tahapan penelitian meliputi *preprocessing*, pelabelan kelas, penanganan ketidakseimbangan data dengan *Synthetic Minority Over-sampling Technique* (SMOTE), serta pembagian data dengan skenario 70:30, 80:20, dan 90:10. Proses klasifikasi dilakukan menggunakan algoritma *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*. Evaluasi model menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score*. Hasil menunjukkan bahwa skenario 90:10 memberikan performa terbaik, dengan *Support Vector Machine* (SVM) mencapai *accuracy* 0,8806, *precision* 0,8757, *recall* 0,8941, dan *F1-score* 0,8788, melampaui algoritma lainnya. Kategori *desktop hardware* teridentifikasi sebagai kasus terbanyak selama periode penelitian. Temuan ini dapat dimanfaatkan sebagai dasar strategis untuk prioritas penanganan layanan, alokasi sumber daya, serta penguatan *Knowledge Management* guna peningkatan layanan TI secara terarah, efektif, dan berkelanjutan.

Kata Kunci: Klasifikasi, Knowledge Discovery, Layanan TI, Supervised Machine Learning

1. PENDAHULUAN

Keandalan layanan teknologi informasi (TI) menjadi faktor penting dalam keberlangsungan operasional organisasi modern. Di Indonesia, hasil riset *Penggunaan Teknologi Digital dalam Pengembangan Bisnis* yang dilakukan oleh Litbang Kompas bekerja sama dengan Mekari menunjukkan bahwa 65% perusahaan menilai penerapan teknologi digital mampu meningkatkan efektivitas operasional bisnis. Temuan ini menegaskan bahwa layanan TI tidak lagi berperan sebagai fungsi pendukung semata, melainkan telah menjadi bagian integral dalam proses bisnis organisasi. Dalam kondisi tersebut, kemampuan organisasi dalam menangani gangguan layanan TI secara konsisten dan berkelanjutan menjadi krusial. Munculnya gangguan layanan yang berulang pada jenis layanan tertentu sering kali mengindikasikan bahwa pengetahuan hasil penanganan sebelumnya belum dimanfaatkan secara optimal, yang pada akhirnya dapat berdampak pada lamanya waktu penyelesaian gangguan, peningkatan beban kerja tim TI, serta penurunan kualitas layanan yang diterima pengguna [1].

Data tiket layanan TI yang dimiliki organisasi pada dasarnya menyimpan pengetahuan operasional yang bernilai, seperti pola gangguan, layanan yang paling sering bermasalah, serta efektivitas penyelesaian yang telah dilakukan. Penelitian yang dilakukan oleh Arientawati, Jumaryadi, dan Wibowo menegaskan bahwa pengelompokan dan analisis data gangguan layanan TI mampu mengidentifikasi kategori layanan yang dominan serta pola permasalahan yang berulang, sehingga dapat mendukung perencanaan peningkatan kualitas layanan secara lebih terarah [2]. Pada Pertamina Prabumulih, data pendukung menunjukkan bahwa layanan telekomunikasi, jaringan, dan multimedia mendominasi jumlah tiket layanan TI. Kondisi ini mengindikasikan adanya permasalahan berulang pada kategori layanan tertentu yang berpotensi dianalisis lebih lanjut untuk menghasilkan pengetahuan yang dapat dimanfaatkan secara sistematis.

Namun, keberadaan data dalam jumlah besar tidak serta-merta menghasilkan nilai tambah tanpa pengelolaan yang tepat. *Knowledge Management* (KM) berperan dalam mengelola pengetahuan agar dapat dimanfaatkan kembali secara efektif untuk meningkatkan kualitas layanan TI. Penelitian Prihartono menunjukkan bahwa penerapan KM mampu meningkatkan kualitas layanan melalui pengelolaan pengetahuan yang lebih terstruktur [3]. Selain itu, penelitian Bahri dan Tania menunjukkan bahwa pendekatan *knowledge discovery* dengan model *machine learning* seperti LSTM, *Random Forest*, dan *Support Vector Regression* dapat menghasilkan prediksi yang lebih akurat pada data historis [22].

Seiring meningkatnya volume dan kompleksitas data operasional, diperlukan pendekatan analitis untuk mengekstraksi pengetahuan dari data historis. Penelitian oleh Alifia Ayu Zahrothul Ain dan Cutifa Safitri menunjukkan bahwa penerapan *machine learning* pada pengelolaan layanan TI berbasis ITIL mampu membantu proses klasifikasi dan prioritas insiden sehingga meningkatkan efektivitas penanganan gangguan layanan [4]. Selain itu, studi oleh Fachrozi dan Tania yang membandingkan beberapa algoritma *machine learning* seperti *Naïve Bayes*, *Support Vector Machine*, *K-Nearest Neighbor*, *Decision Tree*, dan *Random Forest* menunjukkan bahwa setiap algoritma memiliki performa yang berbeda tergantung pada karakteristik data yang digunakan [23].

Kebaruan penelitian ini terletak pada tiga aspek utama. Pertama, penelitian ini menerapkan pendekatan *knowledge discovery* secara spesifik pada data tiket layanan TI di sektor industri hulu migas Indonesia yang masih jarang diteliti. Kedua, hasil klasifikasi diintegrasikan ke dalam kerangka *Knowledge Management* sebagai dasar pengambilan keputusan strategis. Ketiga, kombinasi SMOTE dengan tiga skenario pembagian data (70:30, 80:20, 90:10) memberikan evaluasi performa yang lebih komprehensif dibandingkan dengan penelitian sebelumnya yang umumnya hanya menggunakan satu skenario.

Pemilihan tiga algoritma didasarkan pada pertimbangan teoritis dan empiris yang kuat. *Naïve Bayes* dipilih karena kemampuannya menangani data teks berdimensi tinggi secara efisien dengan kompleksitas komputasi rendah, serta terbukti memberikan *baseline* yang kuat dalam klasifikasi dokumen [26]. SVM dipilih karena secara teoritis unggul dalam menemukan *hyperplane* optimal pada ruang fitur TF-IDF yang berdimensi tinggi dan *sparse* [27][28]. *Random Forest* dipilih karena mekanisme *ensemble* berbasis *bagging* yang mampu mengurangi varians dan risiko *overfitting* pada data klasifikasi multikelas yang kompleks [29]. Perbandingan ketiganya memberikan gambaran yang representatif tentang spektrum kemampuan model *machine learning* pada data layanan TI.

2. METODE PENELITIAN

2.1 Data Collection

Tahapan data collection merupakan tahapan yang menentukan kualitas serta kelengkapan data yang akan dianalisis. Proses pengumpulan data yang sistematis dan terdokumentasi dengan baik berpengaruh terhadap validitas serta keandalan hasil penelitian, karena kualitas hasil sangat ditentukan oleh ketepatan prosedur dan transparansi dalam pelaporan metode [5] [6]. Data penelitian diperoleh dari sistem manajemen tiket milik Pertamina Prabumulih. Data yang digunakan mencakup seluruh tiket layanan teknologi informasi yang tercatat selama periode 2020-2025, terdiri dari 8.627 entri, sehingga analisis dilakukan berdasarkan data historis secara menyeluruh.

2.2 Text - Preprocessing

Tahap *text-preprocessing data* merupakan proses persiapan data sebelum analisis dan pemodelan dilakukan, yang bertujuan untuk meningkatkan kualitas, konsistensi, dan kesiapan data sehingga mampu mendukung performa model serta validitas hasil penelitian [7][8]. Tahap *text-preprocessing data* dilakukan untuk menyiapkan data teks sebelum dilakukan proses analisis dan pemodelan. langkah langkah yang dilakukan diantaranya:

1. *Cleaning*
Cleaning merupakan tahap untuk mengubah seluruh huruf dalam teks menjadi huruf kecil (*lowercase*), serta menghapus karakter yang tidak relevan seperti angka, simbol, tautan, dan tanda baca.
2. *Tokenizing*
Tokenizing adalah proses pemecahan teks menjadi unit-unit yang lebih kecil berupa kata-kata, yang selanjutnya disebut sebagai token.
3. *Filtering*
Pada tahap *filtering*, dilakukan proses penyaringan terhadap kata-kata hasil tokenisasi dengan menghilangkan kata-kata yang dianggap tidak memiliki makna penting. Kata-kata umum yang sering muncul dan tidak memberikan informasi signifikan disebut sebagai *stopword*, seperti “dan”, “yang”, “serta”, dan “setelah”. Penghapusan *stopword* ini dapat membantu mengurangi ukuran data serta meningkatkan efisiensi proses pengolahan.
4. *Stemming*
Stemming merupakan proses untuk mengubah kata ke bentuk dasarnya dengan cara menghilangkan imbuhan seperti awalan (*prefix*) dan akhiran (*suffix*). Proses ini bertujuan untuk menyederhanakan variasi kata sehingga kata-kata yang memiliki akar yang sama dapat dikelompokkan dalam satu kategori, meskipun memiliki bentuk yang berbeda.

2.3 Data Labeling

Pada tahap pelabelan data, setiap tiket Layanan TI diberi label berdasarkan kategori layanan teknologi informasi menggunakan *rule-based classification* [9][10]. Label tersebut merepresentasikan jenis layanan atau permasalahan yang ditangani, seperti layanan desktop, jaringan dan internet, multimedia, telekomunikasi, serta layanan pendukung lainnya.

2.4 Synthetic Minority Oversampling Technique (SMOTE)

Untuk mengatasi ketidakseimbangan jumlah data pada setiap kategori layanan TI, penelitian ini menerapkan *Synthetic Minority Oversampling Technique* (SMOTE) sebelum proses pelatihan model. SMOTE diterapkan untuk menyeimbangkan proporsi kelas dengan menciptakan data sintesis di kelas yang kurang terwakili, sehingga dapat meningkatkan efektivitas model klasifikasi [11].

Secara teknis, SMOTE bekerja dengan cara memilih sampel dari kelas minoritas, kemudian membuat titik data sintesis baru di sepanjang garis yang menghubungkan sampel tersebut dengan k-nearest neighbor-nya. Proses ini diulang hingga setiap kelas memiliki jumlah sampel yang seimbang. Dalam penelitian ini, SMOTE diterapkan hanya pada data training untuk menghindari kebocoran data (*data leakage*) ke data testing.

2.5 Ekstraksi Fitur (TF-IDF)

Pada tahapan ini, teks yang telah melalui hasil *preprocessing* dikonversi ke dalam bentuk numerik menggunakan metode *Term Frequency–Inverse Document Frequency* (TF-IDF) [12]. Representasi TF-IDF tersebut selanjutnya digunakan sebagai input pada model *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* untuk menilai tingkat kepentingan kata dalam setiap tiket [13]. Selanjutnya, data dibagi ke dalam tiga skenario pengujian, yaitu 70:30, 80:20, dan 90:10 antara data *training* dan data *testing* guna mengevaluasi kinerja model dalam mengenali data baru.

2.6 Algoritma Klasifikasi

Pemilihan algoritma klasifikasi didasarkan pada karakteristik data teks layanan TI serta tinjauan literatur.

2.6.1 Naïve Bayes

Naïve Bayes merupakan algoritma klasifikasi probabilistik yang didasarkan pada Teorema Bayes dan mengasumsikan bahwa setiap fitur bersifat independen terhadap fitur lainnya [26]. Fitur yang digunakan berasal dari hasil ekstraksi TF-IDF pada teks tiket layanan TI. *Naïve Bayes* dipilih karena memiliki proses

komputasi yang cepat, sederhana, serta mampu menangani data teks berdimensi tinggi dengan jumlah fitur yang besar [26]. Probabilitas suatu dokumen d termasuk ke dalam kelas c dihitung menggunakan Teorema Bayes, yang ditunjukkan pada Persamaan 1.

$$P(c|d) = \frac{P(c)P(d|c)}{P(d)} \quad (1)$$

Karena nilai $P(d)$ bersifat konstan untuk seluruh kelas, maka proses klasifikasi dilakukan dengan memilih probabilitas posterior terbesar, ditunjukkan pada Persamaan 2.

$$c_{MAP} = \arg c \in C \prod_{i=1}^n P(x_i|c) \quad (2)$$

2.6.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) merupakan algoritma klasifikasi yang bekerja dengan mencari hyperplane terbaik untuk memisahkan data antarkelas secara optimal [27][28]. Hyperplane terbaik dipilih berdasarkan margin terbesar antara dua kelas, sehingga model mampu melakukan klasifikasi dengan lebih akurat [2]. Setiap tiket layanan TI direpresentasikan ke dalam bentuk vektor TF-IDF. SVM kemudian digunakan untuk memisahkan kategori layanan berdasarkan pola kata yang muncul pada masing-masing tiket [27][28]. Persamaan *hyperplane* pada SVM dituliskan pada Persamaan 3.

$$f(x) = w \cdot x + b \quad (3)$$

Untuk memperoleh hyperplane terbaik, SVM melakukan optimasi dengan meminimalkan nilai (Persamaan 4):

$$\min \frac{1}{2} ||w||^2 \quad (4)$$

dengan syarat (Persamaan 5):

$$y_i(w \cdot x_i + b) \geq 1 \quad (5)$$

SVM dipilih karena memiliki kemampuan yang baik dalam menangani data teks yang sparse dan berdimensi tinggi [2][3]. Selain itu, SVM juga mampu menghasilkan performa klasifikasi yang tinggi pada data layanan TI yang memiliki kemiripan kata antarkategori. Hal ini sesuai dengan hasil penelitian, di mana SVM memperoleh nilai *accuracy*, *precision*, *recall*, dan *F1-score* tertinggi dibandingkan dengan algoritma lainnya [2][3].

2.6.3 Random Forest

Random Forest merupakan metode *ensemble* berbasis *bagging* yang menggabungkan banyak pohon keputusan yang dibangun secara independen untuk menghasilkan prediksi yang lebih akurat dan stabil [29]. Setiap pohon dibangun menggunakan subset data dan fitur yang berbeda, sehingga mampu mengurangi varians serta meminimalkan risiko *overfitting*. Algoritma ini efektif dalam menangani data kompleks dan hubungan nonlinier, sehingga banyak digunakan dalam berbagai permasalahan klasifikasi [29].

Berdasarkan Ho et al. [29], setiap pohon dalam *Random Forest* menghasilkan prediksi terhadap data, kemudian hasil akhir diperoleh melalui agregasi berupa rata-rata dari seluruh output pohon. Persamaan prediksi *Random Forest* ditunjukkan pada Persamaan 6

$$h(X) = \frac{1}{K} \sum_{k=1}^K h_k(X) \quad (6)$$

Dimana $h(X)$ merupakan hasil prediksi akhir, K adalah jumlah pohon, dan $h_k(X)$ adalah *output* dari pohon ke- k .

2.7 Evaluasi

Tahapan evaluasi dilaksanakan untuk mengukur efektivitas dari model *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest* dalam mengelompokkan kategori layanan teknologi informasi. Penilaian kinerja model memanfaatkan *Confusion Matrix* yang menghasilkan berbagai metrik seperti *accuracy*, *precision*, *recall*, dan *F1-score* sebagai landasan untuk membandingkan keefektifan model dalam mendukung proses penemuan pengetahuan [14]. Dalam proses evaluasi tersebut, peneliti dapat memeriksa jenis kesalahan prediksi serta menghitung berbagai metrik evaluasi, seperti *accuracy*, *precision*, *recall*, dan *F1-score* [15]. Perhitungan metrik ini didasarkan pada Persamaan 7 - 10 [16].

$$Accuracy = \frac{TP + FN}{TP + FP + TN + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{10}$$

2.8 Knowledge Discovery

Tahap knowledge discovery dilakukan untuk menggali pengetahuan dari hasil klasifikasi dan evaluasi data layanan TI [17][18]. Pengetahuan yang dihasilkan berupa pola permasalahan dan kecenderungan layanan TI yang sering terjadi, yang selanjutnya dimanfaatkan sebagai bagian dari Knowledge Management untuk mendukung optimalisasi efektivitas penanganan layanan TI di Pertamina Prabumulih.

3. HASIL DAN PEMBAHASAN

Bab ini membahas proses analisis data layanan TI pada Pertamina Prabumulih yang diawali dengan tahap *preprocessing*, penyeimbangan data menggunakan SMOTE, serta ekstraksi fitur TF-IDF. Selanjutnya dilakukan pemodelan klasifikasi menggunakan *Naïve Bayes*, *Support Vector Machine (SVM)*, dan *Random Forest*, disertai perbandingan kinerja untuk menentukan model terbaik berdasarkan metrik *accuracy*, *precision*, *recall*, dan *F1-score* melalui *confusion matrix*. Hasil analisis kemudian dimanfaatkan dalam tahap *knowledge discovery* untuk mengidentifikasi pola permasalahan layanan TI guna mendukung peningkatan efektivitas penanganan layanan.

3.1. Preprocessing

Proses *preprocessing* mencakup sejumlah proses penting yang bertujuan untuk membersihkan dan menyiapkan data teks sebelum dianalisis lebih lanjut. Tahapan yang dilakukan meliputi *cleaning* untuk menghilangkan karakter tidak relevan, *tokenizing* untuk memecah teks menjadi unit kata, *filtering* untuk menghapus *stopword*, serta *stemming* untuk mengubah kata ke bentuk dasarnya. Contoh hasil setiap tahapan *preprocessing* pada sampel data tiket layanan TI disajikan pada Tabel 1.

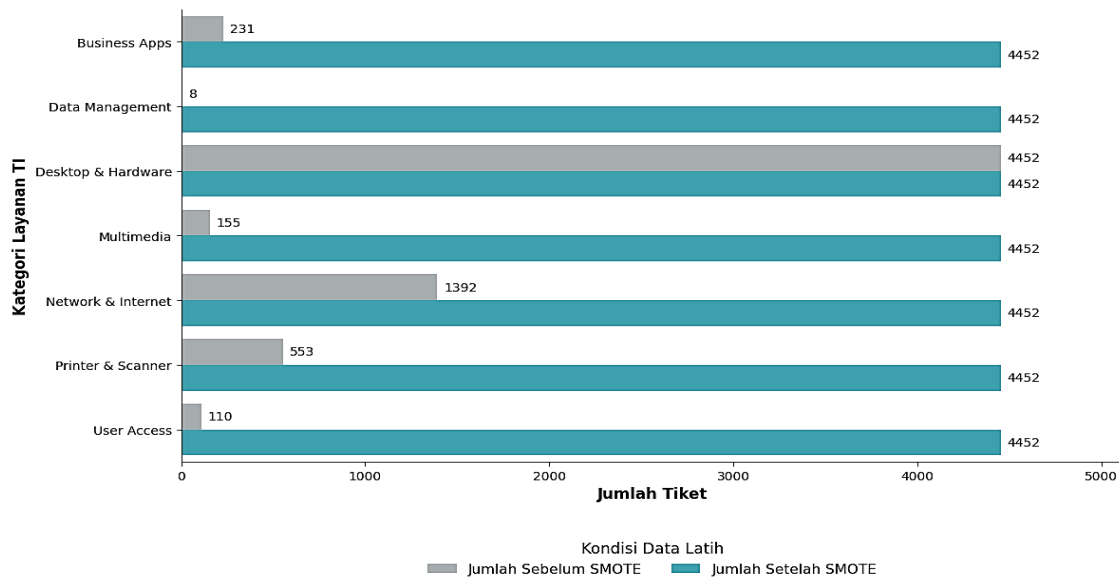
Tabel 1. Tahapan Pre-processing (Sampel Data Tiket Layanan TI Pertamina Prabumulih)

Tahapan	Hasil
Data Awal	Setting Fasilitas sound system 1 set speaker dan microphone 2 1 tv display beserta kelengkapan kabel 1 laptop beserta kelengkapan kabel 2 kabel extension panjang utk semua alat elektronik Pada hari kamis 20 November 2025 standby selama Acara berlangsung pk1 07.00 Wib
Cleaning	setting fasilitas sound system 1 set speaker dan microphone 2 1 tv display beserta kelengkapan kabel 1 laptop beserta kelengkapan kabel 2 kabel extension panjang utk semua alat elektronik pada hari kamis 20 november 2025 standby selama acara berlangsung pk1 07.00 wib
Tokenizing	['setting', 'fasilitas', 'sound', 'system', '1', 'set', 'speaker', 'microphone', '2', '1', 'tv', 'display', 'kelengkapan', 'kabel', '1', 'laptop', 'kelengkapan', 'kabel', '2', 'kabel', 'extension', 'panjang', 'alat', 'elektronik', 'kamis', '20', 'november', '2025', 'standby', 'acara', 'berlangsung', 'pukul', '07.00', 'wib']
Filtering	['aplikasi', 'bagus', 'cicilan', 'ansur', 'kecewa', 'lunas', 'minjem', 'limit', 'kunci', 'pembayaran', 'jatuh', 'tempo', 'bayar', 'jatuh', 'tempo', 'nunggak', 'cicilan', 'tolong', 'diperbaiki', 'sistim', 'alasan', 'limit', 'kunci']
Stemming	['aplikasi', 'bagus', 'cicil', 'ansur', 'kecewa', 'lunas', 'minjem', 'limit', 'kunci', 'bayar', 'jatuh', 'tempo', 'bayar', 'jatuh', 'tempo', 'nunggak', 'cicil', 'tolong', 'baik', 'sistim', 'alas', 'limit', 'kunci']

Catatan: Hasil olah data pada tabel hanya merupakan sampel dari data tiket layanan TI Pertamina Prabumulih

3.2. SMOTE

Penerapan *Synthetic Minority Over-sampling Technique (SMOTE)* membantu menyeimbangkan distribusi data latih pada setiap kategori layanan yang sebelumnya tidak merata. Sebelum dilakukan *oversampling*, jumlah data pada masing-masing kategori masih timpang, seperti *Business Apps* (231), *Data Management* (8), *Multimedia* (155), *Network & Internet* (1392), *Printer & Scanner* (553), dan *User Access* (110), sementara *Desktop & Hardware* memiliki jumlah paling tinggi yaitu 4452 data, sehingga berpotensi menimbulkan bias pada model terhadap kategori mayoritas. Setelah diterapkan SMOTE, seluruh kategori layanan memiliki jumlah data yang sama, yaitu 4452 sampel, melalui proses pembangkitan data sintetis pada kategori minoritas hingga setara dengan kategori mayoritas, sehingga model diharapkan mampu mengenali pola setiap kategori secara lebih seimbang dan meningkatkan performa klasifikasi. Gambar 1 merupakan perbandingan SMOTE.



Gambar 1. Perbandingan SMOTE

3.3. Split Data

Data yang telah melalui preprocessing dan penyeimbangan kemudian dipisahkan ke dalam *dataset* pelatihan (*training*) dan *dataset* pengujian (*testing*). Dalam penelitian ini, diterapkan tiga variasi skema pembagian data untuk menguji konsistensi performa model. Rincian jumlah data pada masing-masing skenario ditunjukkan pada Tabel 2.

Tabel 2. Hasil *split data*

Skenario	70:30	80:20	90:10
<i>Data Training</i>	6038	6901	7764
<i>Data testing</i>	2589	1726	863

3.4. Model Testing

Tahap pengujian model dilakukan untuk mengevaluasi performa klasifikasi melalui pengukuran dengan menggunakan metrik akurasi, presisi, *recall*, dan *F1-score* sebagai indikator evaluasi. Selain itu, ditetapkan berbagai skema pembagian data, yaitu 70:30, 80:20, dan 90:10, untuk mengevaluasi konsistensi dan kestabilan performa model pada proporsi data latih dan data uji yang berbeda.

3.4.1 Naïve Bayes

Tabel 3 merupakan hasil pengujian performa model dengan berbagai skenario pembagian data yang berbeda menunjukkan bahwa *Naïve Bayes* mencapai hasil terbaik pada skenario 90:10 dengan akurasi 82,97%, *precision* 71,63%, *recall* 84,86%, dan *F1-score* 75,66%, yang mengindikasikan kinerja yang lebih optimal dibandingkan dengan skenario 70:30 dan 80:20.

Tabel 3. Hasil skenario *Naïve Bayes*

Skenario	70:30	80:20	90:10
<i>Accuracy</i>	0.8227	0.8262	0.8297
<i>Precision</i>	0.6651	0.6641	0.7163
<i>Recall</i>	0.8477	0.8484	0.8486
<i>F1-score</i>	0.7186	0.7200	0.7566

3.4.2 SVM

Berdasarkan hasil pengujian pada beberapa skenario pembagian data (lihat Tabel 4), algoritma *Support Vector Machine* (SVM) menunjukkan kinerja paling optimal pada rasio 90:10. Pada skenario tersebut, model memperoleh nilai akurasi sebesar 88,06%, dengan *precision* 87,57%, *recall* 89,41%, serta *F1-score* 87,88%. Nilai-nilai ini merupakan capaian tertinggi dibandingkan dengan rasio 70:30 dan 80:20, sehingga dapat disimpulkan bahwa semakin besar proporsi data latih, semakin meningkat kinerja model *Support Vector Machine* (SVM).

Tabel 4. Hasil skenario *Support Vector Machine* (SVM)

Skenario	70:30	80:20	90:10
<i>Accuracy</i>	0.8687	0.8691	0.8806
<i>Precision</i>	0.8243	0.7865	0.8757
<i>Recall</i>	0.8639	0.8626	0.8941
<i>F1-score</i>	0.8310	0.8051	0.8788

3.4.3 Random Forest

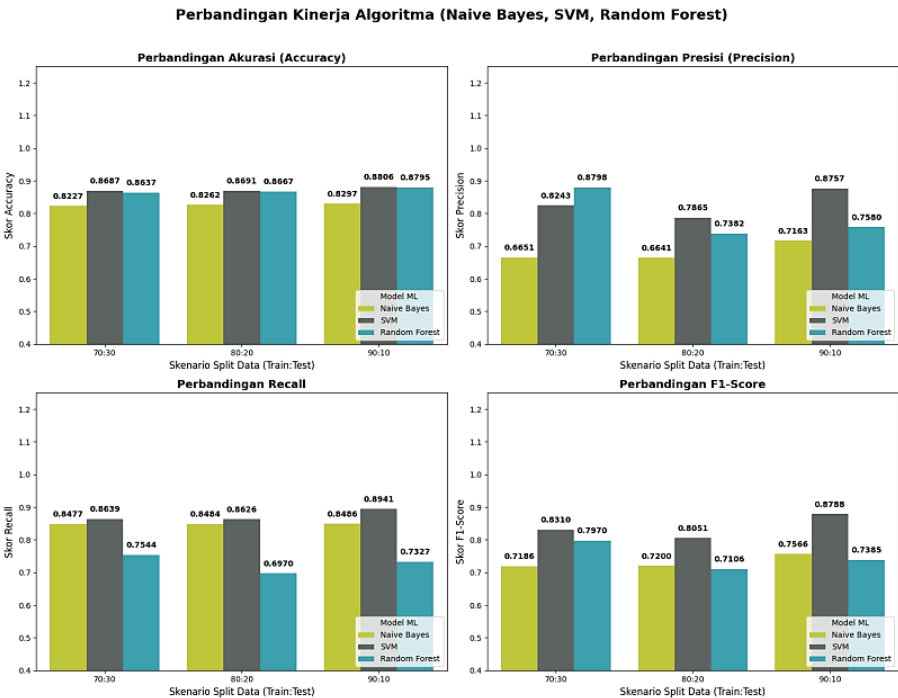
Pada Tabel 5 hasil pengujian performa model dengan beberapa skenario pembagian data menunjukkan bahwa *Random Forest* mencapai hasil terbaik pada skenario 90:10 dengan akurasi 87,95%, *precision* 75,80%, *recall* 73,27%, dan *F1-score* 73,85%, yang mengindikasikan kinerja yang lebih optimal dibandingkan dengan skenario 70:30 dan 80:20. Hal ini menunjukkan bahwa peningkatan proporsi data latih berkontribusi terhadap peningkatan performa model *Random Forest*.

Tabel 5. Hasil skenario *Random Forest*

Skenario	70:30	80:20	90:10
<i>Accuracy</i>	0.8637	0.8667	0.8795
<i>Precision</i>	0.8798	0.7382	0.7580
<i>Recall</i>	0.7544	0.6970	0.7327
<i>F1-score</i>	0.7970	0.7106	0.7385

3.5. Hasil Evaluasi

Hasil pengujian menunjukkan bahwa pembagian data dengan rasio 90:10 memberikan performa yang lebih optimal dibandingkan rasio 70:30 dan 80:20. Pada skenario ini, model *Support Vector Machine* (SVM) mencapai akurasi sebesar 88,06%, *precision* 87,57%, *recall* 89,41%, dan *F1-score* 87,88%, diikuti *Random Forest* dengan akurasi 87,95% dan *Naïve Bayes* sebesar 82,97%. Tingginya nilai *recall* SVM pada skenario 90:10 menunjukkan bahwa model mampu mengenali sebagian besar data uji dengan benar, yang merupakan indikator penting dalam konteks klasifikasi layanan TI di mana kesalahan pengenalan kategori dapat berdampak pada keterlambatan penanganan insiden. Grafik perbandingan akurasi, *recall*, *precision* dan *F1-score* dari ketiga model dapat dilihat pada Gambar 2.



Gambar 2. Grafik Perbandingan Akurasi, recall, *precision* dan *F1-score* dari Ketiga Model

3.5.1 Analisis Keunggulan SVM pada Data Teks TF-IDF

SVM unggul karena bekerja optimal pada data TF-IDF yang berdimensi tinggi dan *sparse*, di mana kompleksitas klasifikasinya bergantung pada jumlah *support vector*, bukan dimensionalitas ruang fitur [27][28]. *Naïve Bayes* memperoleh performa lebih rendah karena asumsi independensi antar fitur tidak

terpenuhi pada teks tiket yang memiliki korelasi kata kuat, seperti 'reset-password' pada kategori User Access atau 'kabel-koneksi' pada Network & Internet [26]. Adapun *Random Forest*, meskipun kompetitif, memperoleh *precision* dan *F1-score* lebih rendah karena sebagian pohon keputusan dibangun berdasarkan fitur tidak informatif pada ruang fitur TF-IDF yang sangat besar, meskipun mekanisme *majority voting* tetap membantu menstabilkan hasil akhir [29].

3.5.2 Analisis Pengaruh Skenario Pembagian Data

Skenario 90:10 menghasilkan performa terbaik karena model dilatih dengan 7.764 data, memberikan representasi pola yang lebih kaya dibandingkan skenario 70:30 (6.038 data) dan 80:20 (6.901 data). Peningkatan performa yang konsisten dari skenario 70:30 ke 90:10 juga mengindikasikan model belum mengalami *overfitting*, yang berarti data tiket Pertamina Prabumulih masih memiliki variasi yang cukup untuk terus dipelajari seiring bertambahnya proporsi data latih [29].

3.5.3 Perbandingan dengan Penelitian Sejenis

Performa SVM dalam penelitian ini kompetitif dibandingkan studi serupa. Nowak-Brzezinska dan Drabek (2025) menggunakan pendekatan *hybrid knowledge-based ML* namun tanpa variasi skenario *data splitting* [20], sementara Zangari et al. (2023) menunjukkan bahwa *contextual embedding* dapat meningkatkan akurasi klasifikasi tiket lebih lanjut [21]. Meskipun demikian, pendekatan TF-IDF tetap memberikan keseimbangan yang baik antara akurasi, interpretabilitas, dan efisiensi komputasi untuk penerapan praktis di Pertamina Prabumulih, sejalan dengan temuan bahwa model *machine learning* konvensional seperti SVM mampu memberikan performa dasar yang solid sebelum dipertimbangkan penggunaan model *transformer* yang lebih kompleks [30].

3.6. Hasil Confusion Matrix



Gambar 3. Hasil Confusion Matrix

Hasil *confusion matrix* dari ketiga model disajikan pada Gambar 3. Ketiga model menunjukkan kemampuan klasifikasi yang baik pada kategori Desktop & Hardware dan Network & Internet sebagai dua kategori dengan data terbanyak. *Naive Bayes* mengklasifikasikan 1.014 dan 200 data dengan benar, *Random Forest* 1.094 dan 193 data, sementara SVM 1.079 dan 204 data. Meskipun demikian, ketiga model menunjukkan tingkat misklasifikasi yang lebih tinggi pada kategori minoritas seperti Data Management, Multimedia, dan User Access. Hal ini disebabkan kemiripan leksikal antar kategori misalnya tiket Multimedia dan Desktop & Hardware sama-sama mengandung kata seperti 'monitor', 'speaker', dan 'display' sehingga model kesulitan membedakan keduanya secara tepat.

Di antara ketiga model, SVM menunjukkan distribusi prediksi paling konsisten lintas kategori dengan nilai *false negative* lebih rendah, memperkuat posisinya sebagai model terbaik tidak hanya dari metrik agregat tetapi juga kemampuan mengenali setiap kategori secara merata. Kategori dengan misklasifikasi tinggi seperti Data Management dan User Access perlu mendapat perhatian khusus ke depan, misalnya melalui penambahan fitur *domain-specific* atau teknik *oversampling* yang lebih adaptif.

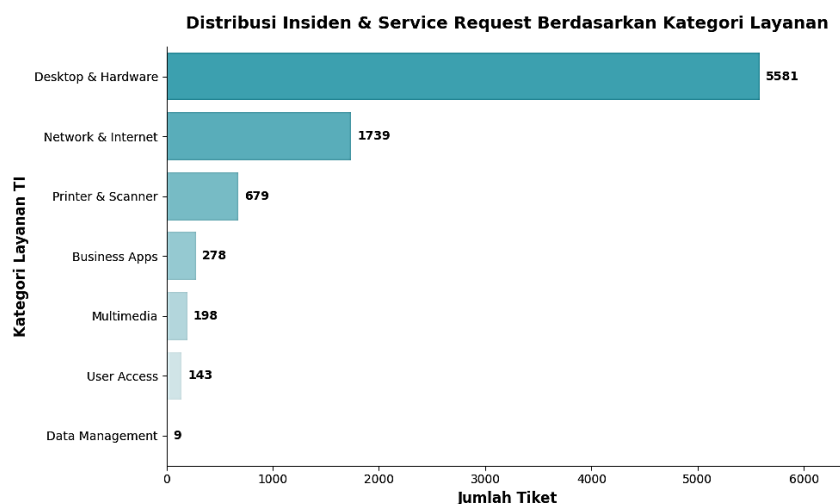
3.7. Knowledge Discovery

Hasil *knowledge discovery* pada data Layanan TI menunjukkan bahwa kategori layanan Desktop & Hardware memiliki jumlah laporan tertinggi dibandingkan kategori lainnya, diikuti oleh kategori Network & Internet serta Printer & Scanner, sedangkan kategori *Business Apps*, *Multimedia*, *User Access*, dan *Data Management* memiliki jumlah laporan yang relatif lebih rendah. Pola ini mengindikasikan bahwa permasalahan yang berkaitan dengan perangkat kerja pengguna dan infrastruktur jaringan masih menjadi sumber utama gangguan layanan TI. Temuan ini sejalan dengan penelitian yang menyatakan bahwa proses pengelompokan dan visualisasi tiket layanan mampu mengidentifikasi kategori masalah dominan secara otomatis sehingga membantu organisasi dalam menentukan prioritas perbaikan layanan [19], serta

mengungkap pola data tersembunyi dengan pendekatan *knowledge discovery* berbasis analisis teks dan *machine learning* [20]. Hal ini selaras dengan penelitian yang menunjukkan bahwa penerapan *knowledge discovery* berbasis *machine learning* yang interpretatif mampu menghasilkan wawasan yang lebih transparan dan mendukung pengambilan keputusan berbasis data secara strategis [21].

Selain itu, hasil pengujian model klasifikasi menunjukkan bahwa *Support Vector Machine* (SVM) memiliki performa terbaik dibandingkan *Naïve Bayes* dan *Random Forest*. Hal ini memperkuat temuan bahwa metode klasifikasi berbasis teks yang dikombinasikan dengan pengetahuan domain dapat meningkatkan ketepatan pengelompokan tiket layanan TI [20]. Studi lain juga menegaskan bahwa penerapan klasifikasi bertingkat dan pemanfaatan *contextual embedding* mampu meningkatkan akurasi pengenalan kategori tiket dan mempercepat proses penanganan insiden [21].

Temuan ini memiliki implikasi praktis yang signifikan bagi pengelola layanan TI Pertamina Prabumulih. Tingginya volume tiket pada kategori Desktop & Hardware mengindikasikan perlunya alokasi sumber daya teknis yang lebih besar pada kategori ini, misalnya melalui penambahan tenaga teknis, penyediaan stok suku cadang, atau penerapan jadwal perawatan preventif. Untuk kategori Network & Internet yang menduduki posisi kedua, diperlukan strategi monitoring infrastruktur jaringan yang lebih proaktif. Pengetahuan yang dihasilkan dari klasifikasi ini dapat diintegrasikan ke dalam sistem *Knowledge Management* organisasi untuk mempercepat waktu penanganan insiden di masa mendatang. Distribusi kategori layanan TI dapat dilihat pada Gambar 4.



Gambar 4. Distribusi Kategori Layanan TI

Dengan demikian, penerapan *knowledge discovery* pada data historis Layanan TI tidak hanya mampu mengungkap pola permasalahan layanan TI yang dominan, tetapi juga menyediakan dasar pengetahuan yang dapat dimanfaatkan dalam *Knowledge Management* untuk mendukung pengambilan keputusan strategis dan peningkatan kualitas layanan TI secara berkelanjutan.

4. KESIMPULAN

Penelitian ini menganalisis klasifikasi kategori layanan TI di Pertamina Prabumulih dengan mengimplementasikan tiga algoritma klasifikasi, diantaranya *Naïve Bayes*, *Support Vector Machine* (SVM), dan *Random Forest*. Analisis dilakukan terhadap data historis tiket periode 2020–2025 yang telah melalui tahapan *preprocessing*, penyeimbangan data menggunakan SMOTE, serta ekstraksi fitur *TF-IDF*. Model testing dilakukan menggunakan beberapa skenario pembagian data, dan rasio 90:10 terbukti memberikan performa paling optimal dalam menjaga akurasi dan kestabilan model.

Berdasarkan hasil evaluasi, *Support Vector Machine* (SVM) menunjukan performa paling optimal dengan akurasi sebesar 0,8806, *precision* 0,8757, *recall* 0,8941, dan *F1-score* 0,8788, dibandingkan dengan model lainnya yaitu *Naïve Bayes* dan *Random Forest*. Selain itu, analisis *confusion matrix* menunjukkan bahwa SVM memiliki kemampuan klasifikasi yang lebih konsisten dalam memprediksi kategori layanan TI secara tepat. Oleh karena itu, *Support Vector Machine* direkomendasikan sebagai model terbaik untuk klasifikasi layanan TI untuk mendukung proses pengambilan keputusan strategis dalam peningkatan kualitas layanan TI. Untuk pengembangan lebih lanjut, penelitian selanjutnya dapat difokuskan pada optimasi hiperparameter secara komprehensif, pemanfaatan representasi teks berbasis *contextual embedding* seperti IndoBERT, serta integrasi model ke dalam sistem *dashboard monitoring* atau *decision support system* berbasis real-time agar hasil *knowledge discovery* dapat dimanfaatkan secara lebih efektif dan berkelanjutan.

Hasil penelitian ini memiliki implikasi praktis yang signifikan bagi pengelola layanan TI Pertamina Prabumulih. Model SVM yang telah divalidasi dapat diintegrasikan ke dalam sistem *Knowledge Management* organisasi untuk mengotomatisasi kategorisasi tiket secara *real-time*, sehingga mempercepat penanganan insiden dan meningkatkan akurasi alokasi sumber daya teknis. Dominasi kategori Desktop & Hardware mengindikasikan perlunya penambahan tenaga teknis dan jadwal perawatan preventif, sementara kategori Network & Internet memerlukan strategi *monitoring* infrastruktur yang lebih proaktif. Pengetahuan yang dihasilkan dapat menjadi basis keputusan strategis dalam perencanaan kapasitas layanan TI serta penguatan *Knowledge Management* organisasi untuk mendukung peningkatan kualitas layanan TI secara terarah dan berkelanjutan.

REFERENSI

- [1] Kompas.com, "Riset penggunaan teknologi digital dalam pengembangan bisnis," Jan. 2025.
- [2] A. Arientawati, Y. Jumaryadi, dan A. Wibowo, "Analisis Pengelompokan Gangguan TIK Pada Sistem Pencatatan Layanan Menggunakan Algoritma K-Means dan Metode Elbow," *Techno.COM*, vol. 22, no. 2, Mei 2023, doi: 10.33633/tc.v22i2.7819.
- [3] A. Prihartono, "Aplikasi *Knowledge Management* untuk media berbagi pengetahuan layanan program studi," *J. Ticom Technol. Inf. Commun.*, vol. 14, no. 1, pp. 19–26, 2025, doi: 10.70309/ticom.v14i1.161.
- [4] A. A. Z. Ain dan C. Safitri, "Enhancing ITIL incident management: Innovative *machine learning* approaches for efficient incident prioritization and resolution," *J. Tek. Inform.*, vol. 16, no. 2, pp. 204–214, 2023, doi: 10.15408/jti.v16i2.31439.
- [5] R. M. Pulsford *et al.*, "The impact of selected methodological factors on data collection outcomes in observational studies of device-measured physical behaviour in adults: A systematic review," *International Journal of Behavioral Nutrition and Physical Activity*, vol. 20, no. 1, p. 26, 2023, doi: 10.1186/s12966-023-01423-4.
- [6] I. Bergbom dan M. Lepp, "Visual arts and drawings to communicate and explore authentic life situations, a data collection method in caring science—a hermeneutic perspective," *Scandinavian Journal of Caring Sciences*, vol. 36, no. 3, pp. 625–634, 2022, doi: 10.1111/scs.13040.
- [7] V. Werner de Vargas *et al.*, "Imbalanced data preprocessing techniques for machine learning: A systematic mapping study," *Knowledge and Information Systems*, vol. 65, no. 1, pp. 31–57, 2023, doi: 10.1007/s10115-022-01772-8.
- [8] C. C. Olisah, L. Smith, dan M. Smith, "Diabetes mellitus prediction and diagnosis from a data preprocessing and *machine learning* perspective," *Computer Methods and Programs in Biomedicine*, vol. 220, p. 106773, 2022, doi: 10.1016/j.cmpb.2022.106773.
- [9] S. K. M. Hossain, S. A. Ema, dan H. Sohn, "Rule-Based Classification Based on Ant Colony Optimization: A Comprehensive Review," 2022, doi: 10.1155/2022/2232000.
- [10] K. A. Cahyanto, M. A. Al Hilmi, dan M. Mustamiin, "Pengujian Rule-Based pada Dataset Log Server Menggunakan *Support Vector Machine* Berbasis Linear Discriminant Analysis untuk Deteksi Malicious Activity," *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, vol. 9, no. 2, pp. 245–254, Apr. 2022, doi: 10.25126/jtiik.201743299.
- [11] Ridwan, E. H. Hermaliani, dan M. Ernawati, "Penerapan Metode SMOTE Untuk Mengatasi Imbalanced Data Pada Klasifikasi Ujaran Kebencian," *Computer Science (CO-SCIENCE)*, vol. 4, no. 1, pp. 80–88, Jan. 2024, doi:10.31294/coscience.v4i1.2990
- [12] K. P. Harmandini dan K. Muslim, "Analysis of TF-IDF and TF-RF Feature Extraction on Product Review Sentiment," *Sinkron: Jurnal dan Penelitian Teknik Informatika*, vol. 8, no. 2, pp. 929–936, Apr. 2024, doi: 10.33395/v8i2.13376.
- [13] I. S. Wibowo, A. Witanti, dan I. Susilawati, "Keyword Extraction Judul Berita Online di Indonesia Menggunakan Metode TF-IDF," *JATISI*, vol. 11, no. 1, pp. 99–111, 2024, doi: 10.26418/jp.v11i3.101542.
- [14] F. R. Valerian, M. Syarief, dan D. A. Fatah, "Klasifikasi menggunakan confusion matrix sebagai metode evaluasi model," *JATI*, vol. 9, no. 2, 2025, doi: 10.36040/jati.v9i2.13062.
- [15] F. R. Valerian, M. Syarief, dan D. A. Fatah, "Klasifikasi tingkat obesitas menggunakan metode GBM dan confusion matrix," *JATI*, vol. 9, no. 2, pp. 2242–2249, Apr. 2025, doi: 10.36040/jati.v9i2.13062.
- [16] A. F. Azmi and A. Voutama, "Prediksi churn nasabah bank menggunakan klasifikasi *Random Forest* dan *Decision Tree* dengan evaluasi confusion matrix," *KOMPUTA: J. Ilm. Komput. dan Inform.*, vol. 13, no. 1, pp. 111–119, 2024, doi: 10.34010/komputa.v13i1.12639
- [17] J. Duque, F. Silva, dan A. Godinho, "Data Mining applied to *Knowledge Management*," *Procedia Computer Science*, vol. 219, pp. 455–461, 2023, doi: 10.1016/j.procs.2023.01.312.
- [18] J. Duque, F. Silva, dan A. Godinho, "Data Mining for *Knowledge Management*," *Procedia Computer Science*, vol. 239, pp. 257–264, 2024, doi: 10.1016/j.procs.2024.06.170.

- [19] R. Ishizuka, T. Ichimura, T. Maeda, dan Y. Matsumoto, "Categorization and visualization of issue tickets to support understanding of implemented features in software development projects," *Applied Sciences*, vol. 12, no. 20, 2022, doi: 10.3390/app12073222.
- [20] A. Nowak-Brzezińska dan D. Drabek, "A hybrid knowledge-based *machine learning* approach for ticket classification in the AWS cloud," *Procedia Computer Science*, vol. 223, pp. 1096–1105, 2025, doi: 10.1016/j.procs.2025.09.605.
- [21] A. Zangari, N. Angius, A. Murgia, dan R. Tonelli, "Ticket automation: An insight into current research with applications to multi-level classification scenarios," *Expert Systems with Applications*, vol. 225, 2023, doi: 10.1016/j.eswa.2023.119984.
- [22] C. A. Bahri dan K. D. Tania, "Perbandingan kinerja LSTM, *Random Forest*, dan SVR berbasis *knowledge discovery* untuk prediksi harga beras Sumatera Selatan," *JURIKOM*, vol. 12, no. 5, pp. 721–732, 2025, doi: 10.30865/jurikom.v12i5.9140.
- [23] M. Al Fachrozi dan K. D. Tania, "Comparison of Naive Bayes, SVM, K-NN, Decision Tree, and *Random Forest* in sentiment analysis," *JURTEKSI*, vol. 12, no. 1, pp. 69–80, 2025, doi: 10.36341/rabit.v11i1.7449.
- [24] M. R. Maulizidan dan K. D. Tania, "Knowledge discovery based on sentiment analysis of public perceptions about generative AI on X," *Indonesian Journal of Informatics Education*, vol. 9, no. 2, 2025, doi: 10.20961/ijie.v9i2.107758.
- [25] D. A. Ardhani dan K. D. Tania, "Knowledge discovery on e-commerce customer churn using interpretable machine learning," *Journal of Applied Informatics and Computing*, vol. 9, no. 5, pp. 2695–2702, 2025, doi: 10.30871/jaic.v9i5.10811.
- [26] N. Buslim, L. K. Oh, M. H. A. Hardy, and Y. Wijaya, "Comparative Analysis of KNN, *Naïve Bayes* and SVM Algorithms for Movie Genres Classification Based on Synopsis," *Jurnal Teknik Informatika*, vol. 15, no. 2, pp. 143–154, 2022, doi: 10.15408/jti.v15i2.29302.
- [27] I. Minatika and R. Arifudin, "Implementation of the Term Frequency-Inverse Document Frequency Method for Mental Health Classification Using Algorithm *Support Vector Machine*," *Recursive Journal of Informatics*, vol. 3, no. 2, 2024, doi: 10.15294/rji.v3i2.1921.
- [28] S. N. Cahyani and G. W. Saraswati, "Implementation of *Support Vector Machine* Method in Classifying School Library Books with Combination of TF-IDF and Word2Vec," *Jurnal Teknik Informatika (JUTIF)*, vol. 4, no. 6, pp. 1555–1566, 2023, doi: 10.52436/1.jutif.2023.4.6.1536.
- [29] W. K. O. Ho, B. S. Tang, and S. W. Wong, "Predicting property prices with *machine learning* algorithms," *Journal of Property Research*, 2021, doi: 10.1080/09599916.2020.1832558.
- [30] O. Pakkarapanit *et al.*, "AI-Based Classification of IT Support Requests in Enterprise Service Management Systems," *Systems*, vol. 14, no. 2, p. 223, 2026, doi: 10.3390/systems14020223.