Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

Perbandingan Kinerja Random forest dan SVM Pada Klasifikasi Tingkat Kekumuhan Permukiman Menggunakan SMOTE

Nurika Dwi Wahyuni, Fadhilah Syafria*, Novi Yanti, Surya Agustian

Fakultas Sains dan Teknologi, Teknik Informatika, Universitas Islam Negeri Sultan Syarif Kasim Riau, Pekanbaru, Indonesia
Email: [1]12250120360@students.uin-suska.ac.id,[2,*]fadhilah.syafria@uin-suska.ac.id, [3]novi_yanti@uin-suska.ac.id, [4]surya.agustian@uin-suska.ac.id
Email Penulis Korespondensi: fadhilah.syafria@uin-suska.ac.id
Submitted: **29/05/2026**; Accepted: **22/06/2026**; Published**: 23/06/2026**

Abstrak−Klasifikasi tingkat kekumuhan permukiman penting untuk mendukung analisis kondisi permukiman secara lebih terstruktur dan berbasis data. Penelitian ini bertujuan membandingkan kinerja *Random forest* dan *Support vector machine* (SVM) dalam klasifikasi tingkat kekumuhan permukiman di Kota Pekanbaru pada skenario tanpa SMOTE dan dengan SMOTE menggunakan data skoring indikator kekumuhan. Kontribusi penelitian ini meliputi analisis pengaruh penerapan SMOTE terhadap performa model serta evaluasi penggunaan top 10 fitur dibandingkan seluruh fitur dalam proses klasifikasi. Dataset yang digunakan berjumlah 992 data RT dari Disperkim Kota Pekanbaru tahun 2020, 2021, dan 2023 dengan 16 fitur skor indikator kekumuhan berdasarkan Permen PUPR Nomor 14 Tahun 2018, mencakup tiga kelas: Tidak Kumuh, Kumuh Ringan, dan Kumuh Sedang. Tahapan penelitian mengikuti proses KDD yang meliputi selection, *preprocessing*, *transformation*, *data mining*, evaluasi model, dan analisis hasil, dengan pembagian data 80:20 menggunakan *stratified sampling* serta evaluasi menggunakan *accuracy*, *precision*, *recall*, *F1-score*, dan *confusion matrix*. Hasil penelitian menunjukkan bahwa SVM *Linear* tanpa SMOTE memperoleh nilai sempurna pada evaluasi (*accuracy*, *precision*, *recall*, dan *F1-score* = 1,0000). Namun, hasil tersebut diinterpretasikan secara hati-hati karena label kelas diturunkan dari aturan skoring Permen PUPR Nomor 14 Tahun 2018, sehingga batas antar kelas cenderung *linear*. *Random forest* mengalami peningkatan *F1-score* setelah SMOTE dari 0,9660 menjadi 0,9700, sedangkan peningkatan terbesar terjadi pada SVM RBF dari 0,9214 menjadi 0,9779. Pengujian top 10 fitur menunjukkan model mengalami penurunan *F1-score*, sehingga penggunaan 16 fitur lebih optimal pada dataset ini.

Kata Kunci: Data Tidak Seimbang; Klasifikasi Kekumuhan; Random Forest; Skoring Indikator; SMOTE; Support Vector Machine

Abstract−Classifying slum levels is essential for a structured, data-driven analysis of settlement conditions. This study compares the performance of *Random forest* and *Support vector machine* (SVM) in classifying slum levels in Pekanbaru City across two scenarios with and without SMOTE using slum indicator scoring data. Its contributions include analyzing SMOTE's impact on model performance and evaluating the top 10 features against the full feature set. The dataset comprises 992 RT-level records from Disperkim Pekanbaru City (2020, 2021, and 2023) featuring 16 slum indicator scores based on PUPR Ministerial Regulation No. 14/2018, categorized into three classes: Non-Slum, Low Slum, and Moderate Slum. Following the KDD process (selection, preprocessing, transformation, data mining, evaluation, and analysis), the data was split 80:20 using stratified sampling and evaluated based on accuracy, precision, recall, F1-score, and confusion matrix. Results show that the Linear SVM without SMOTE achieved perfect evaluation metrics (1.0000); however, this is interpreted cautiously as the class labels derive from strict regulatory scoring rules, making class boundaries inherently linear. *Random forest* saw its F1-score rise from 0.9660 to 0.9700 after SMOTE, while the most significant improvement occurred in SVM RBF, jumping from 0.9214 to 0.9779. Testing the top 10 features led to a decreased F1-score across models, indicating that utilizing all 16 features remains optimal for this dataset.

Keywords: Imbalanced Data; Slum Classification; Random Forest; Indicator Scoring; SMOTE; Support Vector Machine

1. PENDAHULUAN

Permasalahan permukiman kumuh masih menjadi isu strategis yang belum terselesaikan secara optimal dalam pembangunan perkotaan di Indonesia. Salah satu kota yang mengalami hal tersebut adalah Kota Pekanbaru, yang mengalami pertumbuhan penduduk yang cukup pesat, sehingga berdampak pada meningkatnya kebutuhan akan hunian dan infrastruktur permukiman [1]. Kondisi tersebut menimbulkan tantangan dalam penyediaan kawasan permukiman yang layak, yang ditandai dengan penurunan kualitas lingkungan akibat keterbatasan infrastruktur, sanitasi yang kurang memadai, serta pengelolaan limbah yang belum optimal [2]. Permasalahan tersebut tidak hanya berdampak pada penurunan kualitas hidup masyarakat, tetapi juga berpotensi menimbulkan berbagai permasalahan sosial, ekonomi, dan lingkungan di kawasan perkotaan [3]. Selain itu, keberadaan permukiman kumuh juga mencerminkan adanya ketidakseimbangan antara kebutuhan hunian dengan ketersediaan sarana, prasarana, dan pelayanan dasar yang memadai di kawasan perkotaan [4]. Permasalahan ini menyebabkan munculnya kawasan permukiman kumuh yang perlu ditangani secara terencana dan terpadu [5]. Oleh karena itu, diperlukan dukungan data dan metode analisis yang mampu membantu proses identifikasi tingkat kekumuhan secara lebih sistematis, sehingga hasil penilaian dapat digunakan sebagai dasar dalam mendukung perencanaan penanganan kawasan permukiman.

Dalam praktiknya, penentuan tingkat kekumuhan di Kota Pekanbaru masih mengacu pada skoring indikator berdasarkan Peraturan Menteri PUPR Nomor 14 Tahun 2018 yang diperoleh dari Dinas Perumahan dan Kawasan Permukiman (Disperkim) [6]. Data skoring tersebut dapat dimanfaatkan sebagai dasar dalam pengembangan model klasifikasi tingkat kekumuhan permukiman menggunakan pendekatan *machine learning*. Pendekatan ini digunakan untuk membangun model klasifikasi yang mampu mengelompokkan tingkat kekumuhan berdasarkan indikator yang tersedia. Selain itu, pemanfaatan *machine learning* juga dapat menjadi alternatif pendukung dalam mengolah data

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

penilaian yang memiliki banyak indikator, sehingga proses klasifikasi dapat dilakukan secara lebih terstruktur dan konsisten. Namun, data tingkat kekumuhan yang digunakan menunjukkan distribusi kelas yang tidak seimbang, khususnya pada kelas Kumuh Sedang yang jumlah datanya jauh lebih sedikit dibandingkan kelas Tidak Kumuh dan Kumuh Ringan. Kondisi ini dapat menyebabkan model klasifikasi cenderung lebih dominan mengenali kelas mayoritas dan kurang optimal dalam mendeteksi kelas minoritas.

Beberapa penelitian sebelumnya telah menerapkan algoritma *Random forest* dan *Support vector machine* (SVM) pada berbagai kasus klasifikasi. *Random forest* dan SVM dilaporkan memiliki akurasi tinggi, yaitu lebih dari 97% pada klasifikasi data lingkungan, dengan *Random forest* memperoleh hasil lebih unggul [7]. Sementara itu, Kasahun dan Legesse melaporkan bahwa *Random forest* dan SVM mampu menghasilkan *overall accuracy* masing-masing sebesar 97% dan 96% pada klasifikasi *land use/land cover* berbasis *object-based image analysis*, dengan *Random forest* menunjukkan performa yang sedikit lebih unggul. Temuan tersebut menunjukkan bahwa kedua algoritma memiliki kemampuan yang baik dalam menyelesaikan permasalahan klasifikasi [8].Selain itu, SVM berbasis citra satelit Pleiades pernah digunakan untuk identifikasi awal kawasan kumuh dan menghasilkan akurasi sebesar 86,25% [9]. Untuk menangani data tidak seimbang, penelitian lain menerapkan *Random forest* dan SMOTE *Random forest*, dengan hasil peningkatan akurasi dari 54% menjadi 71% serta membantu mengatasi ketidakseimbangan data pada kelas minoritas [10]. Penerapan SMOTE pada algoritma SVM dan *Random forest* juga menunjukkan peningkatan performa pada kasus klasifikasi kualitas udara, yaitu *Random forest* dari 98% menjadi 99% dan SVM dari 91% menjadi 95% [11]. Penelitian lain yang menggunakan SVM dengan SMOTE dan tuning parameter juga menunjukkan peningkatan akurasi dari 98,31% menjadi 99,78% serta peningkatan *recall* kelas minoritas dari 89,19% menjadi 98,46% [12]. Hasil-hasil penelitian tersebut menunjukkan bahwa *Random forest*, SVM, dan SMOTE memiliki potensi dalam menyelesaikan permasalahan klasifikasi, khususnya pada data yang memiliki distribusi kelas tidak seimbang.

Meskipun penentuan tingkat kekumuhan telah memiliki aturan baku melalui sistem skoring berdasarkan Permen PUPR Nomor 14 Tahun 2018 [6], penelitian ini tidak ditujukan untuk menggantikan metode penilaian resmi yang telah berlaku. Penelitian ini berfokus pada evaluasi kemampuan model *machine learning* dalam mempelajari pola hubungan antar indikator kekumuhan dan mengidentifikasi tingkat kekumuhan berdasarkan data historis yang tersedia. Pengujian tersebut menjadi relevan karena distribusi kelas pada data tidak seimbang, sehingga kemampuan model dalam mengenali kelas minoritas belum tentu dapat dicapai secara optimal hanya melalui pendekatan klasifikasi sederhana. Selain itu, pendekatan *machine learning* memungkinkan analisis terhadap pengaruh ketidakseimbangan data dan efektivitas teknik penanganannya, seperti SMOTE, yang tidak menjadi fokus dalam sistem skoring konvensional.

Meskipun berbagai pendekatan *machine learning* telah diterapkan pada kasus klasifikasi lingkungan dan permukiman, belum ada kajian yang secara spesifik mengevaluasi kinerja dan batas kemampuan model *machine learning* ketika diaplikasikan pada data yang labelnya berasal dari sistem skoring deterministik seperti Permen PUPR Nomor 14 Tahun 2018 [6]. Permasalahan ketidakseimbangan kelas pada data tingkat kekumuhan Kota Pekanbaru juga menjadi tantangan dalam pengembangan model klasifikasi karena dapat memengaruhi kemampuan model dalam mengenali kelas minoritas [13]. Oleh karena itu, perlu dilakukan perbandingan kinerja algoritma klasifikasi pada data skoring tersebut, baik tanpa SMOTE maupun dengan SMOTE [14], untuk mengetahui model yang paling sesuai dalam klasifikasi tingkat kekumuhan permukiman.

Berdasarkan kondisi tersebut, penelitian ini menjadi penting karena data tingkat kekumuhan yang digunakan memiliki distribusi kelas yang tidak seimbang, sehingga model klasifikasi berpotensi lebih dominan mengenali kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas[13]. Selain itu, penggunaan data skoring resmi membuat hasil klasifikasi perlu dianalisis secara hati-hati karena kelas kekumuhan ditentukan berdasarkan akumulasi nilai indikator, sehingga performa model perlu dilihat sesuai karakteristik data yang digunakan. Selain itu, data yang digunakan merupakan data skoring resmi dari Disperkim Kota Pekanbaru yang merepresentasikan kondisi permukiman berdasarkan indikator kekumuhan pada Permen PUPR Nomor 14 Tahun 2018 [6]. Penerapan SMOTE digunakan untuk mengetahui pengaruh penyeimbangan data terhadap kinerja model, khususnya dalam mengenali kelas minoritas seperti kumuh sedang [14].

Berdasarkan permasalahan tersebut, penelitian ini memberikan tiga kontribusi utama. Pertama, penelitian ini mengevaluasi kinerja *Random forest* dan *Support vector machine* pada klasifikasi tingkat kekumuhan permukiman menggunakan data skoring resmi Kota Pekanbaru. Kedua, penelitian ini menganalisis pengaruh penerapan SMOTE terhadap performa kedua algoritma pada kondisi data yang tidak seimbang. Ketiga, penelitian ini mengevaluasi penggunaan top 10 fitur berdasarkan *feature importance* dibandingkan penggunaan seluruh fitur yang tersedia untuk mengetahui pengaruhnya terhadap hasil klasifikasi. Hasil penelitian diharapkan dapat menjadi referensi dalam pemilihan algoritma dan penggunaan teknik penyeimbangan data pada klasifikasi tingkat kekumuhan permukiman.

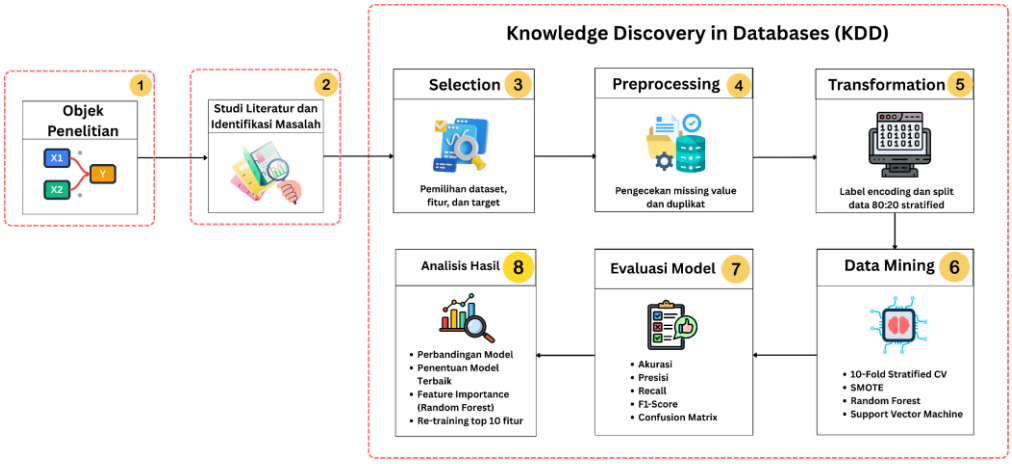
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Tahapan penelitian ini menggunakan pendekatan *Knowledge Discovery in Databases* (KDD) yang terdiri dari tahapan *selection, preprocessing, transformation, data mining,* dan *evaluation*. Tahapan tersebut digunakan untuk mengolah

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

data tingkat kekumuhan permukiman sehingga dapat menghasilkan informasi yang mendukung proses klasifikasi [15] Penelitian ini menggunakan algoritma *Random forest* dan *Support vector machine* (SVM) sebagai algoritma *supervised learning* yang umum digunakan dalam tugas klasifikasi [16], serta menerapkan dua skenario, yaitu tanpa SMOTE dan dengan SMOTE untuk menganalisis pengaruh penyeimbangan data terhadap performa model. Gambar 1 menunjukkan tahapan metodologi yang dilakukan dalam penelitian ini.



Gambar 1. Alur metodologi penelitian

2.1.1 Objek Penelitian

Objek penelitian ini adalah data tingkat kekumuhan permukiman di Kota Pekanbaru yang bersumber dari Dinas Perumahan dan Kawasan Permukiman (Disperkim) Kota Pekanbaru. Dataset terdiri dari 992 data RT yang berasal dari hasil penilaian tingkat kekumuhan tahun 2020, 2021, dan 2023 berdasarkan indikator yang mengacu pada Permen PUPR Nomor 14 Tahun 2018. Data tersebut digunakan untuk menganalisis dan mengklasifikasikan tingkat kekumuhan permukiman menggunakan algoritma *Random forest* dan *Support vector machine* (SVM).

Dataset memiliki 16 fitur yang merepresentasikan indikator kekumuhan yang mencakup aspek bangunan gedung, jalan lingkungan, penyediaan air minum, drainase lingkungan, pengelolaan air limbah, pengelolaan persampahan, dan proteksi kebakaran. Nilai pada setiap fitur merupakan skor penilaian kondisi kekumuhan Permen PUPR Nomor 14 Tahun 2018 [6], bukan skala Likert. Skor diberikan berdasarkan persentase kondisi bermasalah pada masing-masing indikator, yaitu nilai 5 untuk kondisi bermasalah sebesar 76%–100%, nilai 3 untuk 51%–75%, nilai 1 untuk 25%–50% dan nilai 0 menunjukkan kondisi bermasalah tidak ditemukan atau tidak termasuk dalam rentang penilaian.

Berdasarkan Permen PUPR Nomor 14 Tahun 2018 [6], tingkat kekumuhan permukiman diklasifikasikan ke dalam empat kategori, yaitu Tidak Kumuh, Kumuh Ringan, Kumuh Sedang, dan Kumuh Berat. Namun, berdasarkan data yang diperoleh dari Dinas Perumahan dan Kawasan Permukiman Kota Pekanbaru tahun 2020, 2021, dan 2023, tidak ditemukan data dengan kategori Kumuh Berat. Variabel target terdiri atas tiga kelas, yaitu Tidak Kumuh (602 data), Kumuh Ringan (355 data), dan Kumuh Sedang (35 data), sehingga menunjukkan adanya ketidakseimbangan kelas dengan dominasi kelas Tidak Kumuh dan jumlah paling sedikit pada kelas Kumuh Sedang. Tabel 2 menyajikan keterangan fitur yang digunakan dalam dataset.

Tabel 1. Data Tingkat Kekumuhan

No	FT1	FT2	FT3	FT4	FT5	FT6	FT7	FT8	FT9	FT10	FT11	FT12	FT13	FT14	FT15	FT16	Kelas
0	0	0	0	0	0	0	5	0	3	0	0	0	5	5	0	0	Kumuh Ringan
1	0	0	0	0	0	0	5	0	3	0	0	0	5	5	0	0	Kumuh Ringan
2	0	0	0	0	0	1	0	0	5	0	0	0	5	5	0	0	Kumuh Ringan
3	0	0	0	0	1	0	0	0	5	0	0	0	5	5	0	0	Kumuh Ringan
4	0	0	0	0	0	0	0	0	5	0	0	0	5	5	0	0	Tidak Kumuh
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
987	5	0	0	0	0	0	5	0	5	0	0	0	0	0	0	0	Tidak Kumuh
988	5	0	0	0	0	5	0	0	0	5	0	0	0	0	5	5	Kumuh Ringan

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

No	FT1	FT2	FT3	FT4	FT5	FT6	FT7	FT8	FT9	FT10	FT11	FT12	FT13	FT14	FT15	FT16	Kelas
989	5	0	0	0	0	5	0	0	0	5	0	0	0	0	5	5	Kumuh Ringan
990	1	0	0	0	1	1	0	3	0	1	0	0	0	0	5	5	Kumuh Ringan
991	1	0	0	0	1	1	0	3	0	1	0	0	0	0	5	5	Tidak Kumuh

Tabel 2 berikut menyajikan keterangan untuk setiap fitur yang terdapat dalam *dataset* yang digunakan pada penelitian ini. Penjelasan ini mencakup makna dari masing-masing fitur yang merepresentasikan indikator tingkat kekumuhan permukiman sesuai standar yang telah ditetapkan. Dengan adanya keterangan yang jelas, *dataset* menjadi lebih terstruktur dan mudah dipahami.

Tabel 2. Keterangan Fitur *Dataset*

Fitur	Keterangan
FT1	Ketidakteraturan Bangunan
FT2	Kepadatan Bangunan
FT3	Ketidaksesuaian dengan Persyaratan Teknis Bangunan
FT4	Cakupan pelayanan jalan lingkungan
FT5	Kualitas Permukaan Jalan Lingkungan
FT6	Ketersediaan akses aman air minum
FT7	Tidak Terpenuhinya kebutuhan air minum
FT8	Ketidakmampuan mengalirkan limpasan air
FT9	Ketidaktersediaan Drainase
FT10	Kualitas konstruksi drainase
FT11	Sistem pengelolaan air limbah tidak sesuai standar teknis
FT12	Prasarana dan sarana pengelolaan air limbah tidak sesuai persyaratan teknis
FT13	Prasarana dan sarana persampahan tidak sesuai dengan persyaratan teknis
FT14	Sistem pengelolaan persampahan tidak sesuai persyaratan teknis
FT15	Ketidaktersediaan prasarana proteksi kebakaran
FT16	Ketidaktersediaan sarana proteksi kebakaran

2.1.2 Studi Literatur dan Identifikasi Masalah

Tahap studi literatur dan identifikasi masalah dilakukan sebagai tahap awal dalam penelitian. Studi literatur dilakukan untuk memahami penelitian terdahulu terkait klasifikasi permukiman kumuh, *Random forest*, SVM, SMOTE, serta indikator kekumuhan berdasarkan Permen PUPR Nomor 14 Tahun 2018.[6]. Berdasarkan hasil studi literatur dan kondisi dataset yang digunakan, permasalahan utama dalam penelitian ini adalah adanya distribusi kelas yang tidak seimbang pada data tingkat kekumuhan permukiman. Kelas Kumuh Sedang memiliki jumlah data yang jauh lebih sedikit dibandingkan kelas Tidak Kumuh dan Kumuh Ringan. Kondisi tersebut dapat menyebabkan model klasifikasi lebih dominan mengenali kelas mayoritas dan kurang optimal dalam mengenali kelas minoritas [17]. Oleh karena itu, penelitian ini membandingkan kinerja algoritma *Random forest* dan *Support vector machine* pada skenario tanpa SMOTE dan dengan SMOTE untuk mengetahui model yang memberikan performa terbaik dalam klasifikasi tingkat kekumuhan permukiman.

2.1.3 Selection

Tahap selection dilakukan untuk memilih data, fitur, dan target yang digunakan dalam proses klasifikasi. Dataset yang digunakan dalam penelitian ini merupakan data tingkat kekumuhan permukiman di Kota Pekanbaru sebanyak 992 data RT yang berasal dari tahun 2020, 2021, dan 2023. Data tersebut bersumber dari Dinas Perumahan dan Kawasan Permukiman (Disperkim) Kota Pekanbaru. Data tahun 2022 tidak digunakan karena tidak tersedia pada dataset yang diperoleh.

Target klasifikasi pada penelitian ini adalah tingkat kekumuhan permukiman yang terdiri atas tiga kelas, yaitu Tidak Kumuh, Kumuh Ringan, dan Kumuh Sedang. Variabel Kelas digunakan sebagai target klasifikasi, sedangkan seluruh fitur FT1–FT16 dipertahankan dalam proses pemodelan awal. Dengan demikian, proses klasifikasi dilakukan untuk mempelajari hubungan antara indikator kekumuhan dan kategori tingkat kekumuhan pada dataset yang telah mencakup ketiga kategori tersebut.

2.1.4 *Preprocessing*

Tahap *preprocessing* dilakukan untuk memastikan dataset yang digunakan memiliki kualitas yang baik sebelum masuk ke proses transformasi dan pemodelan karena kualitas data dan fitur yang relevan berpengaruh terhadap kinerja model *machine learning* [16]. Pada tahap ini dilakukan pengecekan *missing value* dan pengecekan data duplikat. Pengecekan *missing value* dilakukan untuk mengetahui apakah terdapat data kosong pada setiap atribut, sedangkan pengecekan data duplikat dilakukan untuk memastikan tidak terdapat data yang berulang dalam dataset.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

Berdasarkan hasil pengecekan, tidak ditemukan *missing value* maupun data duplikat, sehingga seluruh data dapat digunakan dalam proses analisis. Selain itu, dilakukan penentuan kolom yang digunakan sebagai fitur dan target. Kolom identitas seperti ID_RT, Tahun, Kecamatan, Kelurahan, RT, dan RW tidak digunakan sebagai fitur dalam proses klasifikasi karena hanya berfungsi sebagai informasi administratif. Fitur yang digunakan dalam pemodelan adalah FT1 sampai FT16, sedangkan kolom Kelas digunakan sebagai target klasifikasi.

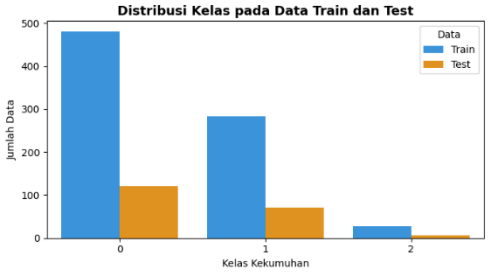
2.1.5 *Transformation*

Pada tahap ini dilakukan pengkodean variabel target atau *label encoding* untuk mengubah kelas kekumuhan yang semula berbentuk kategorikal menjadi numerik. Dalam penelitian ini, kelas Tidak Kumuh diberi label 0, Kumuh Ringan diberi label 1, dan Kumuh Sedang diberi label 2. Proses ini dilakukan agar target klasifikasi dapat diproses oleh algoritma *Random forest* dan *Support vector machine*. Hasil data setelah proses *label encoding* ditampilkan pada Tabel 3.

Tabel 3. Data Setelah *Label encoding*

No	FT1	FT2	FT3	FT4	FT5	FT6	FT7	FT8	FT9	FT10	FT11	FT12	FT13	FT14	FT15	FT16	Kelas
0	0	0	0	0	0	0	5	0	3	0	0	0	5	5	0	0	1
1	0	0	0	0	0	0	5	0	3	0	0	0	5	5	0	0	1
2	0	0	0	0	0	1	0	0	5	0	0	0	5	5	0	0	1
3	0	0	0	0	1	0	0	0	5	0	0	0	5	5	0	0	1
4	0	0	0	0	0	0	0	0	5	0	0	0	5	5	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
987	5	0	0	0	0	0	5	0	5	0	0	0	0	0	0	0	0
988	5	0	0	0	0	5	0	0	0	5	0	0	0	0	5	5	1
989	5	0	0	0	0	5	0	0	0	5	0	0	0	0	5	5	1
990	1	0	0	0	1	1	0	3	0	1	0	0	0	0	5	5	1
991	1	0	0	0	1	1	0	3	0	1	0	0	0	0	5	5	1

Setelah proses *label encoding*, dataset dibagi menjadi data latih dan data uji dengan perbandingan 80:20. Pembagian data dilakukan menggunakan metode *stratified sampling*, yaitu teknik pembagian data yang mempertahankan proporsi setiap kelas agar distribusi kelas pada data latih dan data uji tetap mewakili distribusi kelas pada dataset asli [17]. Distribusi pembagian data ditampilkan pada Gambar 2.



Gambar 2. *Split Data*

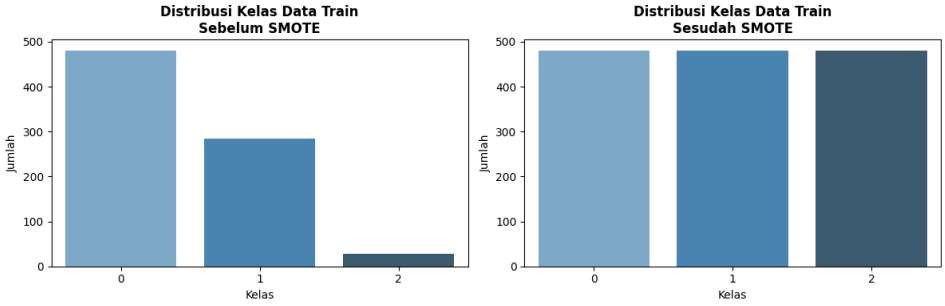
Hasil pembagian data menunjukkan bahwa data latih berjumlah 793 data, terdiri dari 481 data kelas Tidak Kumuh, 284 data kelas Kumuh Ringan, dan 28 data kelas Kumuh Sedang. Sementara itu, data uji berjumlah 199 data, terdiri dari 121 data kelas Tidak Kumuh, 71 data kelas Kumuh Ringan, dan 7 data kelas Kumuh Sedang.

2.1.6 *Data mining*

Tahap *data mining* dilakukan untuk membangun model klasifikasi tingkat kekumuhan permukiman menggunakan algoritma *Random forest* dan *Support vector machine* (SVM). Pada tahap ini, data latih dievaluasi menggunakan metode 10-Fold Stratified *Cross Validation*. Metode ini membagi data latih menjadi 10 subset dengan proporsi kelas yang tetap terjaga, kemudian proses pelatihan dan validasi dilakukan secara bergantian sebanyak 10 kali. Penggunaan *cross validation* bertujuan untuk memperoleh hasil evaluasi model yang lebih stabil dan tidak bergantung pada satu pembagian data tertentu [18]. Penggunaan Stratified *Cross Validation* dalam penelitian ini bertujuan untuk mengevaluasi konsistensi performa model pada berbagai pembagian data yang tetap mempertahankan proporsi masing-masing kelas.

Pengujian dilakukan dalam dua skenario, yaitu tanpa SMOTE dan dengan SMOTE. Pada skenario dengan SMOTE, teknik *oversampling* digunakan untuk menambahkan data sintetis pada kelas minoritas. Penerapan SMOTE dilakukan di dalam *pipeline* sehingga hanya diterapkan pada data latih di setiap fold, bukan pada data validasi maupun data uji. Hal ini dilakukan untuk menghindari kebocoran data dan menjaga objektivitas hasil evaluasi [19]. Sebaran data sebelum dan sesudah penerapan SMOTE ditampilkan pada Gambar 3. Sebelum SMOTE, jumlah data pada kelas Tidak Kumuh sebanyak 481 data, Kumuh Ringan sebanyak 284 data, dan Kumuh Sedang sebanyak 28 data. Setelah SMOTE, jumlah data pada setiap kelas menjadi seimbang, yaitu masing-masing sebanyak 481 data.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101



Gambar 3. Persebaran data sebelum dan sesudah SMOTE

2.1.7 Evaluasi Model

Evaluasi model dilakukan untuk mengukur performa prediksi dalam mengklasifikasikan tingkat kekumuhan permukiman di Kota Pekanbaru. Evaluasi dilakukan menggunakan data uji yang tidak dikenai proses SMOTE agar tetap merepresentasikan distribusi data asli. Model yang telah dilatih kemudian digunakan untuk melakukan prediksi terhadap data uji, dan hasil prediksi dibandingkan dengan label sebenarnya. Evaluasi dilakukan berdasarkan metrik-metrik berikut[16]:

a. *Accuracy*: proporsi prediksi yang benar terhadap keseluruhan data.
b. *Precision*: proporsi data yang diprediksi pada suatu kelas dan benar-benar termasuk dalam kelas tersebut.
c. *Recall*: proporsi data pada suatu kelas yang berhasil dikenali dengan benar oleh model.
d. *F1-score*: rata-rata harmonis dari *precision* dan *recall*.

2.1.8 Analisis Hasil

Tahap analisis hasil dilakukan dengan menginterpretasikan hasil evaluasi yang diperoleh dari masing-masing model. Analisis dilakukan untuk membandingkan performa algoritma *Random forest* dan *Support vector machine* pada skenario tanpa SMOTE dan dengan SMOTE. Perbandingan dilakukan berdasarkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengetahui model dengan performa terbaik dalam klasifikasi tingkat kekumuhan permukiman.

Selain itu, dilakukan analisis terhadap pengaruh penerapan SMOTE dalam meningkatkan performa model, khususnya pada kelas minoritas. Pada model *Random forest*, dilakukan analisis *feature importance* untuk mengetahui fitur yang paling berpengaruh dalam proses klasifikasi. Berdasarkan hasil *feature importance*, dilakukan pengujian lanjutan menggunakan top 10 fitur dengan kontribusi tertinggi. Model kemudian dilatih ulang menggunakan top 10 fitur tersebut dan hasilnya dibandingkan dengan performa model menggunakan seluruh 16 fitur. Tahap ini dilakukan untuk mengetahui apakah fitur dengan kontribusi tertinggi mampu mempertahankan atau meningkatkan performa model.

2.2 *Random forest*

Random forest merupakan algoritma *ensemble learning* yang dikembangkan dari *decision tree* yang bekerja dengan membangun sejumlah pohon keputusan. Algoritma ini juga memiliki kemampuan yang baik dalam menangani data yang beragam dan relatif stabil terhadap variasi data karena menggunakan teknik *bootstrap aggregating (bagging)*. Setelah seluruh *decision tree* terbentuk, prediksi akhir ditentukan melalui proses agregasi, yaitu dengan menggunakan mekanisme *voting* mayoritas pada kasus klasifikasi dan rata-rata pada kasus regresi. Adapun perumusan prediksi akhir pada *Random forest* sebagai berikut[20]:

$$\hat{y} = \text{mode}(\widehat{y_1}, \widehat{y_2}, \dots, \widehat{y_n}) \quad (1)$$

Untuk regresi, prediksi akhir diperoleh dengan menghitung nilai rata-rata dari seluruh hasil prediksi *decision tree*:

$$\hat{y} = \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i \quad (2)$$

2.3 *Support Vector Machine* (SVM)

Support vector machine (SVM) merupakan algoritma klasifikasi yang bekerja dengan mencari *hyperplane* optimal untuk memisahkan data ke dalam dua atau beberapa kelas[21]. *Support vector machine* (SVM) memiliki kemampuan yang baik dalam mengontrol kompleksitas model sehingga dapat mengurangi risiko *overfitting*. Secara matematis, fungsi keputusan pada SVM dinyatakan sebagai berikut:

$$f(x) = \text{sign}(w \cdot x + b) \quad (3)$$

Dalam penelitian ini, SVM digunakan dengan beberapa *kernel*, yaitu *linear*, *polynomial*, dan *radial basis function* (RBF), untuk menangani pola distribusi data yang berbeda.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

3. HASIL DAN PEMBAHASAN

3.1 Hasil Evaluasi Model

Perlu dijelaskan bahwa kelas tingkat kekumuhan diperoleh dari mekanisme penilaian resmi Permen PUPR Nomor 14 Tahun 2018, sehingga variabel target memiliki keterkaitan langsung dengan fitur yang digunakan. Penggunaan machine learning dalam penelitian ini tidak ditujukan untuk menggantikan mekanisme tersebut, melainkan untuk mengevaluasi kemampuan model dalam merepresentasikan pola skoring serta menganalisis pengaruh ketidakseimbangan kelas dan penerapan SMOTE terhadap performa model. Hasil evaluasi masing-masing model pada skenario tanpa SMOTE dan dengan SMOTE disajikan pada Tabel 4.

Tabel 4. Hasil Evaluasi

Algoritma	Skenario / Model	*Accuracy*	*Precision*	*Recall*	*F1-score*
Random forest	Tanpa SMOTE	0,9548	0,9714	0,9616	0,9660
	Dengan SMOTE	0,9598	0,9720	0,9683	0,9700
Support vector machine (SVM)	*Linear* Tanpa SMOTE	1,0000	1,0000	1,0000	1,0000
	Polynomial Tanpa SMOTE	0,9749	0,9823	0,9804	0,9813
	RBF Tanpa SMOTE	0,9648	0,9719	0,8871	0,9214
	Linear Dengan SMOTE	0,9749	0,9781	0,8965	0,9289
	Polynomial Dengan SMOTE	0,9648	0,9718	0,9768	0,9741
	RBF Dengan SMOTE	0,9698	0,9749	0,9815	0,9779

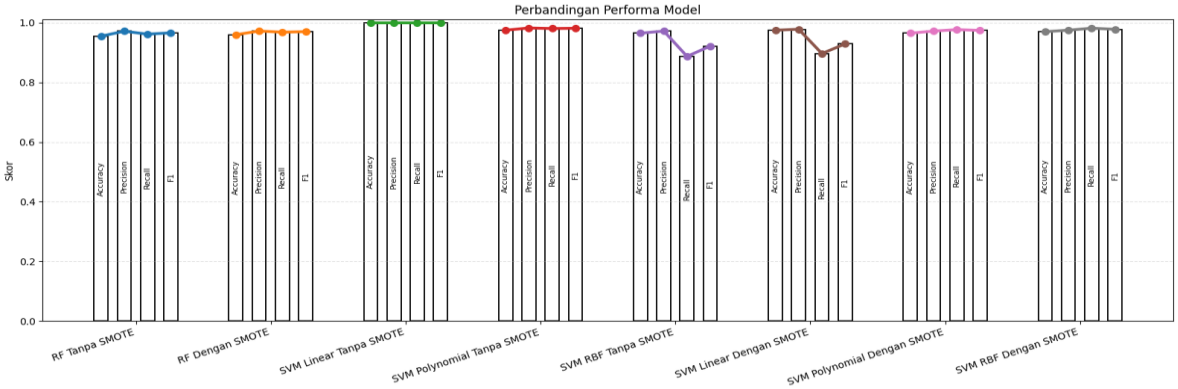
Berdasarkan Tabel 4, hasil evaluasi menunjukkan bahwa setiap model memiliki performa yang berbeda pada skenario tanpa SMOTE dan dengan SMOTE. Perbandingan dua algoritma utama menunjukkan bahwa SVM memperoleh performa tertinggi secara keseluruhan, yaitu pada SVM *Linear* tanpa SMOTE dengan *accuracy*, *precision*, *recall*, dan *F1-score* sebesar 1,0000. Sementara itu, performa terbaik pada *Random forest* diperoleh pada skenario dengan SMOTE dengan *F1-score* sebesar 0,9700.

Pada algoritma *Random forest*, penerapan SMOTE memberikan peningkatan performa meskipun tidak terlalu besar. Nilai *accuracy* meningkat dari 0,9548 menjadi 0,9598, sedangkan *F1-score* meningkat dari 0,9660 menjadi 0,9700. Hal ini menunjukkan bahwa penyeimbangan data menggunakan SMOTE memberikan pengaruh positif terhadap kinerja *Random forest*.

Pada algoritma SVM, pengaruh SMOTE berbeda pada setiap *kernel*. SVM *Linear* tanpa SMOTE memperoleh performa tertinggi dengan nilai *accuracy*, *precision*, *recall*, dan *F1-score* sebesar 1,0000. Namun, setelah penerapan SMOTE, performa SVM *Linear* menurun dengan *F1-score* sebesar 0,9289. SVM *Polynomial* juga mengalami sedikit penurunan *F1-score*, yaitu dari 0,9813 tanpa SMOTE menjadi 0,9741 dengan SMOTE. Sebaliknya, SVM RBF menunjukkan peningkatan performa yang signifikan setelah penerapan SMOTE, dengan *F1-score* meningkat dari 0,9214 menjadi 0,9779. Hal ini menunjukkan bahwa SMOTE memberikan pengaruh paling besar pada model SVM RBF.

Dengan demikian, SVM lebih unggul dari sisi nilai evaluasi tertinggi, sedangkan *Random forest* menunjukkan peningkatan performa setelah penerapan SMOTE. Selain itu, penerapan SMOTE paling efektif meningkatkan performa pada SVM RBF dibandingkan model lainnya. Untuk memberikan gambaran yang lebih jelas mengenai perbandingan performa antar model, hasil evaluasi juga disajikan dalam bentuk grafik pada Gambar 4. Grafik tersebut menunjukkan nilai *accuracy*, *precision*, *recall*, dan *F1-score* dari masing-masing model pada skenario tanpa SMOTE dan dengan SMOTE.

Berdasarkan Gambar 4, model SVM *Linear* tanpa SMOTE menunjukkan nilai tertinggi pada seluruh metrik evaluasi. Sementara itu, *Random forest* dan SVM RBF menunjukkan peningkatan performa setelah penerapan SMOTE. Secara umum, sebagian besar model menunjukkan performa yang tinggi, namun pengaruh SMOTE berbeda pada setiap algoritma dan *kernel* yang digunakan.

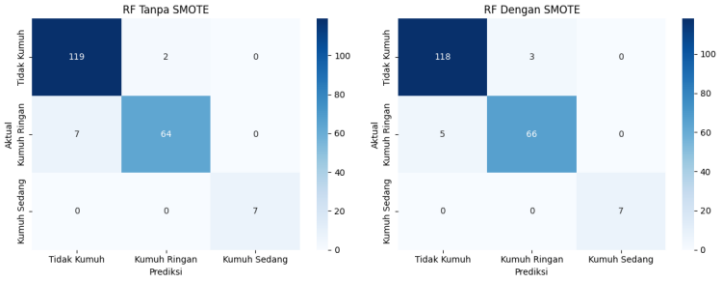


Gambar 4. Perbandingan Performa Model Berdasarkan Metrik Evaluasi

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

3.2 Analisis *Confusion Matrix*

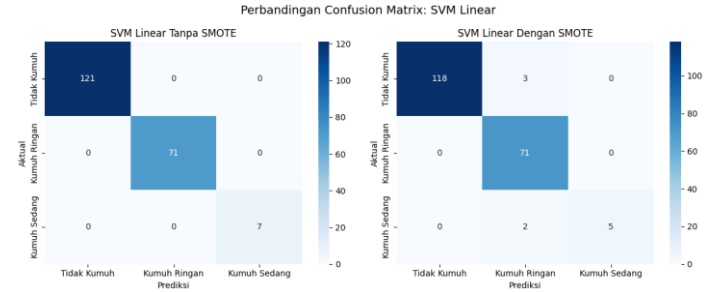
Hasil *confusion matrix Random forest* pada skenario tanpa SMOTE dan dengan SMOTE ditampilkan pada Gambar 5.



Gambar 5. Perbandingan *Confusion matrix Random forest* dengan dan tanpa SMOTE

Berdasarkan Gambar 5, pada model *Random forest* tanpa SMOTE, sebagian besar data pada kelas Tidak Kumuh berhasil diklasifikasikan dengan benar sebanyak 119 data, namun masih terdapat kesalahan klasifikasi ke kelas Kumuh Ringan sebanyak 2 data. Pada kelas Kumuh Ringan, model mampu mengklasifikasikan 64 data dengan benar, tetapi masih terdapat 7 data yang salah diprediksi sebagai Tidak Kumuh. Sementara itu, seluruh data pada kelas Kumuh Sedang berhasil diklasifikasikan dengan benar tanpa kesalahan. Setelah penerapan SMOTE, terjadi peningkatan pada kelas Kumuh Ringan, yaitu jumlah prediksi benar meningkat menjadi 66 data dan kesalahan menurun menjadi 5 data. Namun, pada kelas Tidak Kumuh terjadi sedikit penurunan prediksi benar dari 119 menjadi 118 data. Secara keseluruhan, penerapan SMOTE memberikan perbaikan yang tidak terlalu signifikan, tetapi mampu meningkatkan prediksi pada kelas Kumuh Ringan.

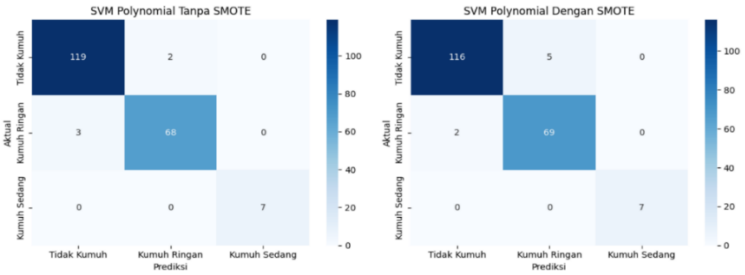
Hasil *confusion matrix* SVM *Linear* pada skenario tanpa SMOTE dan dengan SMOTE ditampilkan pada Gambar 6.



Gambar 6. Perbandingan *Confusion matrix* SVM *Linear* dengan dan tanpa SMOTE

Berdasarkan Gambar 6, model SVM *Linear* tanpa SMOTE mampu mengklasifikasikan seluruh data uji pada ketiga kelas, yaitu Tidak Kumuh, Kumuh Ringan, dan Kumuh Sedang, tanpa adanya kesalahan prediksi. Setelah penerapan SMOTE, performa model mengalami sedikit penurunan. Pada kelas Tidak Kumuh, jumlah prediksi benar menurun dari 121 menjadi 118 data, dengan 3 data salah diklasifikasikan sebagai Kumuh Ringan. Selain itu, pada kelas Kumuh Sedang terdapat 2 data yang salah diprediksi sebagai Kumuh Ringan, sehingga jumlah prediksi benar menjadi 5 data. Meskipun demikian, kelas Kumuh Ringan tetap diklasifikasikan dengan baik tanpa kesalahan.

Hasil *confusion matrix* SVM *Polynomial* pada skenario tanpa SMOTE dan dengan SMOTE ditampilkan pada Gambar 7.

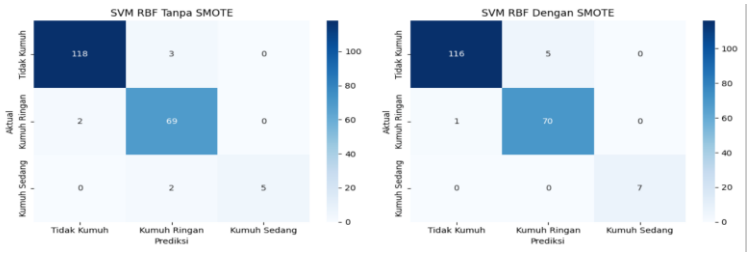


Gambar 7. Perbandingan *Confusion matrix* SVM *Polynomial* dengan dan tanpa SMOTE

Berdasarkan Gambar 7, pada model SVM *Polynomial* tanpa SMOTE, sebagian besar data Tidak Kumuh dan Kumuh Ringan berhasil diprediksi dengan benar, masing-masing sebanyak 119 dan 68 data. Masih terdapat kesalahan 2 data Tidak Kumuh menjadi Kumuh Ringan dan 3 data Kumuh Ringan menjadi Tidak Kumuh. Kelas Kumuh Sedang seluruhnya berhasil diprediksi dengan benar. Setelah SMOTE, kesalahan pada kelas Tidak Kumuh meningkat menjadi

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

5 data, sedangkan kesalahan pada kelas Kumuh Ringan menurun menjadi 2 data, dan kelas Kumuh Sedang tetap terklasifikasi dengan benar. Secara keseluruhan, SMOTE tidak memberikan perubahan signifikan pada SVM *Polynomial*, namun model tetap menunjukkan hasil klasifikasi yang baik. Hasil *confusion matrix* SVM RBF pada skenario dengan dan tanpa SMOTE ditampilkan pada Gambar 8.

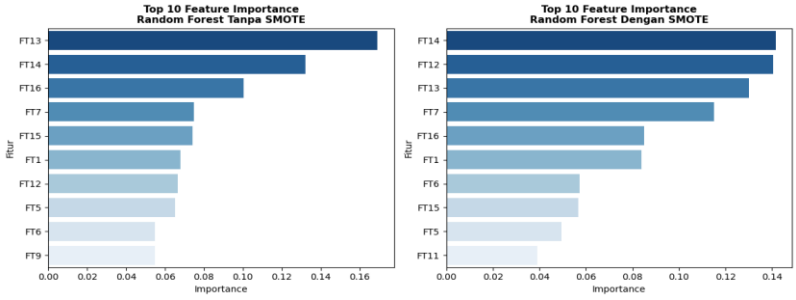


Gambar 8. Perbandingan *Confusion matrix* SVM RBF dengan dan tanpa SMOTE

Berdasarkan Gambar 8, pada model SVM RBF tanpa SMOTE, sebagian besar data Tidak Kumuh dan Kumuh Ringan berhasil diklasifikasikan dengan benar, masing-masing sebanyak 118 dan 69 data. Namun, masih terdapat kesalahan 3 data Tidak Kumuh menjadi Kumuh Ringan dan 2 data Kumuh Ringan menjadi Tidak Kumuh. Pada kelas Kumuh Sedang, hanya 5 data berhasil diklasifikasikan dengan benar dan 2 data salah diklasifikasikan sebagai Kumuh Ringan. Setelah penerapan SMOTE, seluruh data Kumuh Sedang berhasil diklasifikasikan dengan benar, kesalahan pada kelas Kumuh Ringan berkurang dari 2 menjadi 1 data, meskipun kesalahan pada kelas Tidak Kumuh meningkat dari 3 menjadi 5 data. Secara keseluruhan, SMOTE meningkatkan performa SVM RBF terutama dalam mengenali kelas Kumuh Sedang.

3.3 Analisis *Feature Importance*

Analisis feature importance dilakukan untuk mengetahui kontribusi masing-masing variabel terhadap proses klasifikasi. Hasil analisis ini digunakan untuk mengidentifikasi faktor yang paling berpengaruh dalam menentukan tingkat kekumuhan permukiman. Nilai feature importance yang lebih tinggi menunjukkan bahwa variabel tersebut memiliki pengaruh yang lebih besar dalam proses pengambilan keputusan oleh model Random Forest. Perbandingan dilakukan antara data sebelum dan sesudah penerapan SMOTE untuk melihat perubahan tingkat kontribusi setiap fitur. Seperti yang disajikan pada Gambar 9.



Gambar 9. *Top* 10 *Feature Importance Random forest* Tanpa SMOTE dan Dengan SMOTE

Berdasarkan Gambar 9, pada skenario tanpa SMOTE, fitur dengan kontribusi tertinggi adalah FT13, FT14, dan FT16. FT13 merepresentasikan prasarana dan sarana persampahan yang tidak sesuai persyaratan teknis, FT14 merepresentasikan sistem pengelolaan persampahan yang tidak sesuai standar teknis, sedangkan FT16 merepresentasikan ketidaktersediaan sarana proteksi kebakaran. Selain itu, fitur lain yang juga memiliki kontribusi cukup besar adalah FT7, FT15, FT1, FT12, FT5, FT6, dan FT9. Hasil ini menunjukkan bahwa aspek persampahan, proteksi kebakaran, kebutuhan air minum, dan kondisi bangunan menjadi faktor penting dalam menentukan tingkat kekumuhan permukiman.

Pada skenario dengan SMOTE, fitur dengan kontribusi tertinggi mengalami sedikit perubahan, yaitu FT14, FT12, dan FT13. FT14 berkaitan dengan sistem pengelolaan persampahan yang tidak sesuai standar teknis, FT12 berkaitan dengan prasarana dan sarana pengelolaan air limbah yang tidak sesuai persyaratan teknis, sedangkan FT13 berkaitan dengan prasarana dan sarana persampahan yang tidak sesuai persyaratan teknis. Fitur lain yang juga memiliki kontribusi cukup besar adalah FT7, FT16, FT1, FT6, FT15, FT5, dan FT11. Perubahan urutan ini menunjukkan bahwa penerapan SMOTE dapat memengaruhi urutan kontribusi fitur karena distribusi kelas pada data latih menjadi lebih seimbang.

3.4 Pembahasan

Tingginya performa SVM *Linear* menunjukkan bahwa data tingkat kekumuhan pada penelitian ini memiliki karakteristik yang cenderung dapat dipisahkan secara *linear*. Kondisi tersebut kemungkinan dipengaruhi oleh proses pembentukan kelas yang didasarkan pada akumulasi skor indikator sesuai Permen PUPR Nomor 14 Tahun 2018.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

Dengan batas klasifikasi yang telah ditentukan melalui rentang skor tertentu, hubungan antara fitur dan target menjadi lebih terstruktur dibandingkan permasalahan klasifikasi yang berasal dari fenomena lapangan yang lebih kompleks. Oleh karena itu, performa sempurna yang diperoleh SVM *Linear* perlu dipahami sebagai konsekuensi dari karakteristik data yang digunakan, bukan semata-mata karena keunggulan algoritma.

Pada perbandingan skenario, penerapan SMOTE memberikan pengaruh yang berbeda pada setiap model. *Random forest* mengalami peningkatan *F1-score* dari 0,9660 menjadi 0,9700, sedangkan SVM RBF mengalami peningkatan yang lebih signifikan dari 0,9214 menjadi 0,9779, terutama dalam membantu model mengenali kelas minoritas. Sebaliknya, SVM *Linear* mengalami penurunan performa setelah SMOTE, yaitu dari 1,0000 menjadi 0,9289. Penurunan performa pada SVM *Linear* setelah penerapan SMOTE menunjukkan bahwa penambahan data sintetis tidak selalu memberikan manfaat pada dataset yang telah memiliki batas kelas yang jelas. Pada penelitian ini, kelas tingkat kekumuhan dibentuk berdasarkan rentang skor yang telah ditetapkan dalam regulasi, sehingga pemisahan antar kelas relatif tegas. Penambahan sampel sintetis melalui SMOTE berpotensi menghasilkan titik data baru di sekitar batas kelas, sehingga pola pemisahan *linear* yang sebelumnya terbentuk dengan baik menjadi kurang optimal. Temuan ini menunjukkan bahwa efektivitas SMOTE sangat dipengaruhi oleh karakteristik data yang digunakan dan tidak selalu menghasilkan peningkatan performa pada seluruh model klasifikasi.

Hasil *feature importance* menunjukkan bahwa indikator persampahan, pengelolaan air limbah, proteksi kebakaran, penyediaan air minum, dan kondisi bangunan memiliki kontribusi yang relatif tinggi dalam proses klasifikasi tingkat kekumuhan permukiman. Temuan ini mengindikasikan bahwa aspek infrastruktur dasar masih menjadi faktor penting dalam membedakan tingkat kekumuhan antar wilayah permukiman. Selain merepresentasikan pola matematis dalam data, indikator-indikator tersebut juga sejalan dengan komponen penilaian yang digunakan dalam identifikasi kawasan kumuh berdasarkan Permen PUPR Nomor 14 Tahun 2018. Dengan demikian, hasil *feature importance* tidak hanya memberikan informasi mengenai kontribusi fitur terhadap model, tetapi juga dapat menjadi masukan dalam menentukan aspek yang perlu mendapat perhatian dalam upaya penanganan kawasan permukiman.

Berdasarkan hasil *feature importance*, dilakukan pengujian lanjutan dengan memilih 10 fitur yang memiliki kontribusi tertinggi terhadap proses klasifikasi. Fitur terpilih tersebut kemudian digunakan untuk melatih ulang model *Random forest* dan SVM pada skenario tanpa SMOTE dan dengan SMOTE. Proses pelatihan ulang ini dilakukan untuk mengetahui apakah penggunaan fitur dengan kontribusi tertinggi mampu mempertahankan atau meningkatkan performa model dibandingkan penggunaan seluruh 16 fitur. Hasil perbandingan *F1-score* menggunakan 16 fitur dan top 10 fitur ditampilkan pada Tabel 5.

Tabel 5. Perbandingan *F1-score* Model Menggunakan 16 Fitur dan Top 10 Fitur

Algoritma	Skenario / Model	*F1-score* 16 Fitur	*F1-score* Top 10 Fitur	Selisih	Keterangan
Random forest	Tanpa SMOTE	0.9660	0.9026	-0.0633	Menurun
	Dengan SMOTE	0.9700	0.9067	-0.0633	Menurun
Support vector machine (SVM)	*Linear* Tanpa SMOTE	1.0000	0.8164	-0.1836	Menurun
	Polynomial Tanpa SMOTE	0.9813	0.8866	-0.0948	Menurun
	RBF Tanpa SMOTE	0.9214	0.8987	-0.0227	Menurun
	Linear Dengan SMOTE	0.9289	0.8712	-0.0577	Menurun
	Polynomial Dengan SMOTE	0.9741	0.8987	-0.0754	Menurun
	RBF Dengan SMOTE	0.9779	0.9032	-0.0747	Menurun

Berdasarkan Tabel 5, seluruh model mengalami penurunan *F1-score* setelah menggunakan top 10 fitur dibandingkan penggunaan 16 fitur penuh. Penurunan terbesar terjadi pada SVM *Linear* tanpa SMOTE, yaitu sebesar 0,1836, dari *F1-score* 1,0000 menjadi 0,8164. Penurunan terkecil terjadi pada SVM RBF tanpa SMOTE, yaitu sebesar 0,0227, dari *F1-score* 0,9214 menjadi 0,8987. Hasil ini menunjukkan bahwa meskipun top 10 fitur memiliki kontribusi tertinggi berdasarkan *feature importance*, fitur lain di luar top 10 tetap memberikan informasi tambahan yang berpengaruh terhadap proses klasifikasi. Dengan demikian, penggunaan seluruh 16 fitur lebih optimal dibandingkan penggunaan top 10 fitur pada dataset ini.

Hasil penelitian menunjukkan bahwa algoritma *Random forest* memberikan performa terbaik dalam klasifikasi tingkat kekumuhan permukiman dibandingkan beberapa varian *Support vector machine* (SVM) yang diuji. Selain itu, penerapan Synthetic Minority Oversampling Technique (SMOTE) terbukti mampu meningkatkan performa pada beberapa model, terutama *Random forest* dan SVM dengan kernel RBF. Namun demikian, peningkatan tersebut tidak terjadi pada seluruh model karena pada SVM Linear dan SVM Polynomial justru ditemukan penurunan nilai F1-score setelah penerapan SMOTE. Temuan ini mengindikasikan bahwa efektivitas teknik penyeimbangan data dipengaruhi oleh karakteristik dataset dan algoritma yang digunakan.

Hasil penelitian ini sejalan dengan penelitian terdahulu pada klasifikasi kualitas udara yang menunjukkan bahwa penerapan SMOTE mampu meningkatkan performa *Random forest* maupun SVM pada data yang tidak seimbang[11]. Pada penelitian tersebut, akurasi *Random forest* meningkat dari 98% menjadi 99%, sedangkan akurasi SVM meningkat dari 91% menjadi 95% setelah penerapan SMOTE. Sejalan dengan temuan tersebut, penelitian ini juga menunjukkan adanya peningkatan performa pada *Random forest* dan SVM kernel RBF setelah penerapan SMOTE. *Random forest* mengalami peningkatan nilai F1-score dari 0,9660 menjadi 0,9700, sedangkan SVM kernel

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

RBF meningkat dari 0,9214 menjadi 0,9779. Akan tetapi, berbeda dengan penelitian terdahulu yang menunjukkan peningkatan pada seluruh model, penelitian ini menemukan bahwa SMOTE tidak selalu memberikan dampak positif karena performa SVM Linear dan SVM Polynomial justru mengalami penurunan. Hal ini menunjukkan bahwa efektivitas SMOTE sangat bergantung pada karakteristik data dan model klasifikasi yang digunakan.

Temuan penelitian ini juga mendukung penelitian sebelumnya mengenai dampak SMOTE terhadap *Random forest* pada prediksi penyakit jantung [14]. Penelitian tersebut membuktikan bahwa penerapan SMOTE mampu meningkatkan berbagai metrik evaluasi, seperti akurasi, precision, sensitivity, F1-score, specificity, G-Mean, dan Youden's Index, serta menghasilkan model yang lebih fit dan stabil. Sejalan dengan hasil tersebut, penelitian ini menunjukkan bahwa *Random forest* juga memperoleh peningkatan performa setelah penerapan SMOTE. Kesamaan hasil ini memperkuat bahwa SMOTE dapat menjadi alternatif yang efektif dalam menangani ketidakseimbangan kelas. Namun, penelitian ini memberikan temuan tambahan bahwa pengaruh SMOTE tidak bersifat universal pada seluruh algoritma. Oleh karena itu, pemilihan metode penyeimbangan data perlu mempertimbangkan karakteristik dataset serta mekanisme kerja masing-masing algoritma klasifikasi.

Jika dibandingkan dengan penelitian mengenai identifikasi permukiman kumuh menggunakan algoritma SVM dan citra satelit Pleiades di Kota Yogyakarta [9], terdapat kesamaan dalam pemanfaatan machine learning untuk permasalahan permukiman kumuh. Penelitian tersebut memperoleh tingkat akurasi sebesar 86,25% dalam mengidentifikasi kawasan kumuh dan non-kumuh. Namun, terdapat perbedaan mendasar pada pendekatan yang digunakan. Penelitian terdahulu memanfaatkan data citra penginderaan jauh dengan ekstraksi fitur spasial, tekstural, dan spektral, sedangkan penelitian ini menggunakan data skoring indikator kekumuhan berdasarkan Permen PUPR Nomor 14 Tahun 2018. Perbedaan sumber data tersebut diduga menjadi salah satu faktor yang menyebabkan *Random forest* menunjukkan performa yang lebih baik dibandingkan SVM pada penelitian ini. Dataset berbasis skoring cenderung memiliki pola yang lebih terstruktur sehingga dapat dipelajari secara lebih optimal oleh model berbasis ensemble seperti *Random forest*. Meskipun demikian, hasil penelitian ini tetap mendukung temuan sebelumnya bahwa algoritma machine learning memiliki potensi besar dalam mendukung identifikasi dan klasifikasi permukiman kumuh secara lebih objektif dan efisien.

Secara keseluruhan, hasil penelitian menunjukkan bahwa SVM Linear memperoleh performa evaluasi tertinggi pada dataset yang digunakan. Namun, performa tersebut perlu diinterpretasikan secara hati-hati karena karakteristik data tingkat kekumuhan yang dibentuk berdasarkan sistem skoring Permen PUPR Nomor 14 Tahun 2018 cenderung menghasilkan pemisahan kelas yang lebih linear. Di sisi lain, *Random forest* menunjukkan performa yang konsisten pada kedua skenario pengujian dan mengalami peningkatan setelah penerapan SMOTE. Selain itu, penelitian ini juga menunjukkan bahwa dampak SMOTE terhadap performa model tidak selalu konsisten pada seluruh algoritma. Temuan tersebut memperkaya penelitian terdahulu dengan memberikan bukti bahwa keberhasilan teknik penyeimbangan data dipengaruhi oleh karakteristik dataset, representasi fitur, serta mekanisme pembelajaran dari masing-masing algoritma machine learning.

4. KESIMPULAN

Berdasarkan hasil penelitian, *Random forest* dan *Support vector machine* (SVM) mampu merepresentasikan pola klasifikasi tingkat kekumuhan permukiman berdasarkan data skoring indikator yang mengacu pada Permen PUPR Nomor 14 Tahun 2018. Hasil evaluasi menunjukkan bahwa SVM *Linear* tanpa SMOTE memperoleh performa tertinggi dengan nilai *accuracy*, *precision*, *recall*, dan *F1-score* sebesar 1,0000. Namun, hasil tersebut perlu diinterpretasikan secara hati-hati karena kelas target berasal dari sistem skoring resmi yang memiliki keterkaitan kuat dengan fitur yang digunakan dalam proses klasifikasi, sehingga karakteristik dataset cenderung dapat dipisahkan secara linear. SMOTE memberikan dampak yang berbeda pada setiap model, di mana *Random forest* mengalami peningkatan *F1-score* dari 0,9660 menjadi 0,9700, sedangkan peningkatan terbesar terjadi pada SVM RBF dari 0,9214 menjadi 0,9779. Sebaliknya, SVM *Linear* mengalami penurunan performa setelah penerapan SMOTE, yang menunjukkan bahwa efektivitas teknik penyeimbangan data dipengaruhi oleh karakteristik dataset dan algoritma yang digunakan. Selain itu, pengujian menggunakan top 10 fitur menunjukkan bahwa seluruh model mengalami penurunan *F1-score* dibandingkan penggunaan seluruh fitur, sehingga 16 fitur yang digunakan dinilai lebih optimal pada dataset ini. Analisis *feature importance* mengidentifikasi bahwa indikator yang berkaitan dengan persampahan, pengelolaan air limbah, proteksi kebakaran, penyediaan air minum, dan kondisi bangunan merupakan faktor yang paling berpengaruh dalam klasifikasi tingkat kekumuhan, sehingga dapat menjadi bahan pertimbangan bagi Pemerintah Kota Pekanbaru dalam menentukan prioritas program penanganan kawasan kumuh. Penelitian ini menunjukkan bahwa pendekatan machine learning dapat dimanfaatkan sebagai alat bantu untuk merepresentasikan pola klasifikasi pada data skoring resmi serta mengevaluasi pengaruh ketidakseimbangan kelas terhadap performa model. Kontribusi penelitian ini terletak pada perbandingan kinerja *Random forest* dan *Support vector machine* pada klasifikasi tingkat kekumuhan permukiman menggunakan data Kota Pekanbaru, analisis pengaruh penerapan SMOTE terhadap masing-masing model, serta evaluasi penggunaan seluruh 16 fitur dibandingkan top 10 fitur berdasarkan *feature importance*. Temuan penelitian ini menunjukkan bahwa pendekatan machine learning dapat dimanfaatkan sebagai alat bantu dalam merepresentasikan pola klasifikasi pada data skoring resmi dan dapat menjadi referensi dalam pengembangan model klasifikasi tingkat kekumuhan permukiman pada data dengan karakteristik yang serupa.

Building of Informatics, Technology and Science (BITS)
Volume 8, No 1, June 2026 Page: 374−385
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v8i1.10101

REFERENCES

[1] BPS Kota Pekanbaru, "Kota Pekanbaru Dalam Angka 2025," Badan Pusat Statistik Kota Pekanbaru. Accessed: Apr. 29, 2026. [Online]. Available: https://pekanbarukota.bps.go.id/id/publication/2025/02/28/782f2589686f3095440a4005/kota-pekanbaru-dalam-angka-2025.html

[2] Z. Hasan, S. B. N. Arisaputri, F. Alicia, F. Sabila, and Z. Fuady, "Environmental Quality Assessment of Planned Residential Areas in a Peri-Urban Zone Under Urban Sprawl Pressure: A Case Study of Ingin Jaya Subdistrict, Indonesia," *Elkawnie: Journal of Islamic Science and Technology*, vol. 11, no. 2, pp. 171–188, Dec. 2025, doi: 10.22373/ekw.v11i2.31691.

[3] A. R. Sari and M. A. Ridlo, "Studi Literature : Identifikasi Faktor Penyebab Terjadinya Permukiman Kumuh Di Kawasan Perkotaan," *Jurnal Kajian Ruang*, vol. 1, no. 2, pp. 160–176, Sep. 2021, doi: https://dx.doi.org/10.30659/jkr.v1i2.20022.

[4] A. R. Nasution and S. M. Sihombing, "Evaluasi Program Kota Tanpa Kumuh (KOTAKU) dalam Penanganan Kawasan Kumuh di Kabupaten Karo," *Jurnal Manajemen dan Ilmu Administrasi Publik (JMIAP)*, vol. 6, no. 2, pp. 223–234, May 2024, doi: 10.24036/jmiap.v6i2.772.

[5] W. Z. Dela Lathifah A.R. and Z. Rusli, "Strategi Pengembangan dan Penataan Kawasan Permukiman Kumuh Kota Pekanbaru," *Journal of Comprehensive Science*, vol. 3, no. 11, pp. 4950–4968, Nov. 2024, doi: https://doi.org/10.59188/jcs.v3i11.2709.

[6] Kementerian Pekerjaan Umum dan Perumahan Rakyat, *Peraturan Menteri Pekerjaan Umum Dan Perumahan Rakyat Republik Indonesia Nomor 14/Prt/M/2018 Tentang Pencegahan Dan Peningkatan Kualitas Terhadap Perumahan Kumuh Dan Permukiman Kumuh*. 2018. Accessed: Apr. 29, 2026. [Online]. Available: https://peraturan.bpk.go.id/Details/104649/permen-pupr-no-14prtm2018-tahun-2018

[7] E. Banjarnahor, R. Belferik, W. Cendana, Y. Adi, and S. Abraham, "Analisis Implementasi Support vector machine dan Random forest untuk Prediksi Kategori Indeks Kualitas Udara Jakarta," Instek, vol. 10, no. 1, pp. 175–184, Apr. 2025, doi: https://doi.org/10.24252/instek.v10i1.56477.

[8] M. Kasahun and A. Legesse, "Machine learning for urban land use/ cover mapping: Comparison of artificial neural network, random forest and support vector machine, a case study of Dilla town," *Heliyon*, vol. 10, Oct. 2024, doi: 10.1016/j.heliyon.2024.e39146.

[9] P. Widayani, A. Fadilah, I. Z. Irawan, and K. Ghosh, "Implementing Support vector machine Algorithm for Early Slums Identification in Yogyakarta City, Indonesia Using Pleiades Images," *Forum Geografi*, vol. 37, no. 1, pp. 88–97, Jul. 2023, doi: 10.23917/forgeo.v37i1.15248.

[10] V. Oktaviani, N. Rosmawarni, and M. Panji Muslim, "Perbandingan Kinerja Random forest Dan Smote Random forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir," *IFTK Jurnal Informatik*, vol. 20, no. 1, pp. 43–49, Apr. 2024, doi: https://doi.org/10.52958/iftk.v20i1.9158.

[11] I. G. A. N. Lestari and K. A. A. Aryanto, "Peningkatan Akurasi Klasifikasi Kualitas Udara melalui Oversampling dengan Metode Support vector machine dan Random forest," *Jurnal Sistem Dan Informatika (JSI)*, vol. 18, no. 1, pp. 1–9, Nov. 2023, doi: https://doi.org/10.30864/jsi.v18i1.596.

[12] M. A. G. Muttaqin and G. Alfa Trisnapradika, "Optimasi Algoritma SVM dengan Teknik SMOTE dan Tuning Parameter pada Klasifikasi Balita Stunting," *Building of Informatics, Technology and Science (BITS)*, vol. 7, no. 3, pp. 1547–1556, Dec. 2025, doi: 10.47065/bits.v7i3.8330.

[13] D. W. Y. Rahayu, K. Umam, and M. R. Handayani, "Performance of Machine Learning Algorithms on Imbalanced Sentiment Datasets Without Balancing Techniques," *Journal of Applied Informatics and Computing (JAIC)*, vol. 9, no. 3, pp. 998–1005, Jun. 2025, doi: https://doi.org/10.30871/jaic.v9i3.9584.

[14] E. Erlin, Y. Desnelita, N. Nasution, L. Suryati, and F. Zoromi, "Dampak SMOTE terhadap Kinerja Random forest Classifier berdasarkan Data Tidak seimbang," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 21, no. 3, pp. 677–690, Jul. 2022, doi: 10.30812/matrik.v21i3.1726.

[15] M. Sofyan Alfandi and Z. Fatah, "Penerapan Data mining Menggunakan Metode K-Means Clustering Untuk Analisa Penjualan Toko Umama Hijab Kaliwates Jember," *Jurnal Riset Sistem Informasi*, vol. 1, no. 4, pp. 94–102, Dec. 2024, doi: https://doi.org/10.69714/3ty90586.

[16] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 2nd ed. O'Reilly Media, 2019.

[17] N. I. Chaerunnisa, M. Yosep, and T. Sulistyono, "Machine Learning-Based Teacher Performance Classification Using Administrative and Credit Point Assessment (PAK) Data: A Comparative Study of Decision Tree and Naive Bayes," *Journal of Applied Informatics and Computing (JAIC)*, vol. 10, no. 2, pp. 1853–1863, Apr. 2026, doi: https://doi.org/10.30871/jaic.v10i2.12377.

[18] W. Wijiyanto, A. I. Pradana, S. Sopingi, and V. Atina, "Teknik K-Fold Cross Validation untuk Mengevaluasi Kinerja Mahasiswa," *Jurnal Algoritma*, vol. 21, no. 1, pp. 239–248, May 2024, doi: 10.33364/algoritma/v.21-1.1618.

[19] B. Or, "Improving Requirements Classification with SMOTE-Tomek Preprocessing," *arXiv preprint arXiv:2501.06491*, Dec. 2025, [Online]. Available: http://arxiv.org/abs/2501.06491

[20] R. S. Andarujaya and R. R. Suryono, "Perbandingan Kinerja Algoritma Random forest, KNN, dan SVM dalam Analisis Sentimen Cryptocurrency," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 4, pp. 2288–2299, Mar. 2025, doi: 10.47065/bits.v6i4.6572.

[21] N. A. Arifuddin *et al.*, *Machine Learning*. Padang Pariaman: Lingkar Edukasi Indonesia, 2025. [Online]. Available: www.lingkaredukasiindonesia.com